Development of a semi-automated method for tumour budding assessment in colorectal cancer and comparison with manual methods

Natalie C Fisher,1 Maurice B Loughrey,1,2,3 Helen G Coleman,1,3 Melvin D Gelbard,5 Peter Bankhead4,5,6 & Philip D Dunne1

1The Patrick G Johnston Centre for Cancer Research, Queen’s University Belfast, 2Department of Cellular Pathology, Belfast Health and Social Care Trust, 3Centre for Public Health, Queen’s University Belfast, Belfast, 4Edinburgh Pathology, 5Centre for Genomic & Experimental Medicine, and 6Cancer Research UK Edinburgh Centre, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK

Date of submission 23 July 2021
Accepted for publication 25 September 2021
Published online Article Accepted 28 September 2021

Fisher N C, Loughrey M B, Coleman H G, Gelbard M D, Bankhead P & Dunne P D (2022) Histopathology 80, 485–500. https://doi.org/10.1111/his.14574

Development of a semi-automated method for tumour budding assessment in colorectal cancer and comparison with manual methods

Aims: Tumour budding (TB) is an established prognostic feature in multiple cancers but is not routinely assessed in pathology practice. Efforts to standardise and automate assessment have shifted from haematoxylin and eosin (H&E)-stained images towards cytokeratin immunohistochemistry. The aim of this study was to compare manual H&E and cytokeratin assessment methods with a semi-automated approach built within QuPath open-source software.

Methods and results: TB was assessed in cores from the advancing tumour edge in a cohort of stage II/III colon cancers (n = 186). The total numbers of buds detected with each method were as follows: manual H&E, n = 503; manual cytokeratin, n = 2290; and semi-automated, n = 5138. More than four times the number of buds were identified manually with cytokeratin assessment than with H&E assessment. One thousand seven hundred and thirty-four individual buds were identified with both manual and semi-automated assessments applied to cytokeratin images, representing 75.7% of the buds identified manually (n = 2290) and 33.7% of the buds detected with the semi-automated method (n = 5138). Higher semi-automated TB scores were due to any discrete area of cytokeratin immunopositivity within an accepted area range being identified as a bud, regardless of shape or crispness of definition, and to the inclusion of tumour cell clusters within glandular lumina (‘luminal pseudobuds’). Although absolute numbers differed, semi-automated and manual bud counts were strongly correlated across cores (ρ = 0.81, P < 0.0001). All methods of TB assessment demonstrated poorer survival associated with higher TB scores.

Conclusions: We present a new QuPath-based approach to TB assessment, which compares favourably with established methods and offers a freely available, rapid and transparent tool that is also applicable to whole slide images.

Keywords: colorectal cancer, cytokeratin immunohistochemistry, machine learning, QuPath, tumour budding

Introduction

Tumour budding (TB) is the histological manifestation of local tumour cell dissemination, and is usually most evident at the invasive front region of a tumour
mass. TB is an established prognostic factor in a number of solid tumours,\(^7\) although it has been most extensively studied in colorectal cancer (CRC). In pT1 CRC, the presence and extent of TB are predictive for nodal metastatic disease, and can thus be used as a clinical tool for identifying patients who are most likely to benefit from surgical resection.\(^2\) TB has also been shown to have prognostic value in all other stages of CRC, with most evidence having been reported for stage II disease.\(^1,3,4\)

Despite the potential clinical utility of TB, inconsistent qualitative criteria and definitions and non-standardised reporting have proven an obstacle to routine implementation in pathology practice, and TB generally remains a ‘non-core’ item in CRC reporting datasets.\(^5–7\) In an attempt to address this issue in 2016, the International Tumour Budding Consensus Conference (ITBCC) established a consensus definition of a tumour bud, namely a single tumour cell or tumour cell cluster of up to four cells, and an agreed histopathological method of assessment.\(^8\) Although encouraging data were emerging at that time regarding TB assessment with cytokeratin (CK) immunohistochemistry (IHC), most of the established evidence was based on haematoxylin and eosin (H&E) assessment. The consensus preference from the ITBCC was for H&E staining in conjunction with a three-tier scoring system within a ‘hotspot’ field area normalised to 0.785 mm\(^2\).

Since the emergence of the consensus TB definition from the ITBCC, there has been an increased focus on standardisation, reproducibility, and automation, with a view to clinical implementation. This was the subject of a recent comprehensive review, which summarised 12 publications describing different semi-automated approaches to TB assessment, almost all applied to CRC.\(^9\) Most used commercially available software, but two used open-source software (ImageJ), and some used a form of machine learning. Importantly, almost all were applied to CK IHC images, with only one method being proposed for H&E. Other groups pursuing manual rather than semi-automated assessment of TB have also advocated for a CK IHC-based approach.\(^10\) However, a recent expert Delphi consensus process addressing TB concluded that more evidence was required before the incorporation of IHC into TB scoring.\(^11\)

One advantage of CK IHC over H&E assessment is the potential for greater reproducibility in overall TB grade,\(^12\) which addresses a limiting step in progressing TB towards clinical implementation. Most studies have compared only overall TB grade, and very few studies have examined TB assessment at the individual bud level, which is probably where most discordance lies. Recently, Bokhorst et al. compared evaluation by a panel of seven ITBCC experts of 3000 candidate buds from CK-stained sections representing 46 patients with CRC, and found only moderate agreement.\(^13\) Consensus classification was not reached on 41% of the candidate buds. Agreement was slightly better in this study for H&E assessment of individual buds than for CK IHC, but far fewer H&E candidate buds were presented for evaluation.

In the current study, we compared manual H&E and CK assessment methods with a new semi-automated approach to TB assessment performed on digital images from a cohort of stage II and III colon cancers. Manual and semi-automated annotation of individual candidate buds on the same CK IHC images allowed scrutiny of discordance at the individual bud level and consideration of the optimal definition of a tumour bud for these methods of assessment. Results were analysed for all methods against impact on survival, as a measure of relative performance and a comparison of potential clinical utility.

**Materials and methods**

### Study cohort

The study used an established Northern Ireland population-based resource of 661 stage II and III colon cancers, the creation of which has been fully described previously (Northern Ireland Biobank ethical approval references NIB13-0069/87/88 and NIB20-0334).\(^14\) The resource includes tissue microarrays (TMAs) generated from representative tumour blocks containing the tumour advancing edge, with one 1 mm-diameter core per tumour taken from a random area along the advancing edge. Although this does not reflect clinical practice, in which TB grade is based on the ‘hotspot’ area from within a representative whole tumour section, the use of TMAs in this study allowed high throughput and representation of the full morphological spectrum of colon cancer.

Three micrometre sections from each TMA were stained with H&E and with an anti-CK antibody [Cam5.2; Ventana Medical Systems; mouse monoclonal; Cell Conditioning 1 for 8 min; 3,3′-diaminobenzidine (DAB) chromogen] on a BenchMark ULTRA (Ventana Medical Systems) automated slide stainer. Glass slides were scanned on an Aperio AT2 Scanner (Leica Biosystems, Newcastle upon Tyne, UK) at ×40 and imported into the open-source software QuPath (v0.2.3)\(^15\) for evaluation. The scanned
TMA images are available from the Northern Ireland Biobank upon application.

The suitability of individual CK-stained cores for inclusion was determined by manual visual assessment of the scanned images, after application of the QuPath TMA dearraying tool. Note that TMA sampling from the advancing tumour edge is likely to generate a significant number of ‘misses’, with only peritumoral tissue being sampled. Of the 486 cores with sufficient tumour present and matched clinicopathological data, individual cores were also excluded if: (i) only mucinous or signet ring cell carcinoma was present \(n = 26\); (ii) there were large areas of tumour necrosis \(n = 26\); (iii) the tumour present showed weak, patchy or negative immunostaining \(n = 82\); (iv) there was significant stromal CK immunopositivity \(n = 25\); or (v) tissue folding, fragmentation or any other technical artefacts precluded assessment \(n = 72\) (Figure S1). After the above exclusions, 255 cores remained for CK IHC assessment. Manual H&E assessment for inclusion was performed after CK IHC assessment, and a further 61 cores were excluded, because of either a lack of tumour or tissue artefacts as described above, precluding H&E assessment. A further eight cases with <1 month of follow-up time were also excluded from the analysis. This left 186 cases for analysis, having comparative TB data for all four methods of assessment, as detailed below, and clinicopathological data available, including sufficient follow-up.

**MANUAL BUDDING ASSESSMENT**

Buds were manually assessed on H&E and CK IHC images by an expert gastrointestinal pathologist (M.B.L.). This process is shown in Figure 1A–E. Within QuPath, after dearraying, individual cores were shrunk by 30 \(\mu\)m to correlate with semi-automated assessment in excluding candidate buds touching the periphery of the core. Each individual bud was manually marked on all images by use of the point tool within QuPath, enabling quick and accurate quantification per core and the ability to review each individual bud counted. The ITBCC recommendations for H&E TB assessment were followed, with the only exception being that the TMA cores did not represent the budding ‘hotspots’ for each tumour. However, each 1-mm-diameter core approximates to the ITBCC-recommended 0.785 mm\(^2\) area for TB assessment. Furthermore, because of the use of random cores from the advancing edge, our analyses were tested in a wide range of morphological conditions. Predetermination of the tumour region for assessment with the TMA approach allowed an inter-method comparison of individual buds. ‘Pseudobuds’ within areas of heavy acute inflammation were excluded, as recommended.

For initial manual assessment of CK-stained cores, the aim was to annotate as buds clusters of up to four tumour cells, as with H&E assessment, accepting that visualising and counting tumour cell nuclei is more difficult with CK IHC than with H&E (Figure 1C). Regions of irregular or ill-defined immunohistochemical staining were excluded, as some were considered to probably represent cellular fragments rather than viable buds. After this initial assessment was complete, annotated buds (‘CK all’) were reassessed by the same observer to apply the recently suggested additional criterion of nuclear pallor in defining a bud. Those single cells or clusters lacking an identifiable region of nuclear pallor were removed to generate an additional budding dataset (‘CK palor’) which excluded objects lacking this potentially important feature (Figure 1E).

**SEMI-AUTOMATED BUDDING ASSESSMENT**

The semi-automated method was based on a binary (immunopositive/immunonegative) threshold classifier built within QuPath (v0.2.3) and applied to the CK IHC images to identify tumour epithelium. This process is shown in Figure 1F–J. As before, following dearraying, individual cores were shrunk by 30 \(\mu\)m to exclude candidate buds touching the periphery of the core. All lumina completely encapsulated by positive staining were filled in, to prevent the detection of luminal tumour cells or cellular fragments mimicking buds (‘luminal pseudobuds’) (Figure 1G). A pixel classifier was created in QuPath to identify connective discrete areas of immunopositivity by combining image down-sampling, stain separation using colour deconvolution, Gaussian smoothing and global thresholding within a single step (resolution, 1.86 \(\mu\)m/pixel; channel, DAB; prefilter, Gaussian; smoothing sigma, 1.0; threshold, 0.4). With this method, buds were defined not by the number of tumour nuclei, but by the area of CK immunopositivity. An acceptable range of bud area was derived from analysis of the range of areas of the manually annotated CK buds (described in detail below). Those objects with areas outside this range were excluded as buds (Figure 1H–J).

**STATISTICAL ANALYSIS**

Cox proportional hazards (PH) analysis was conducted in Stata version 16 (Timberlake Consultants, © 2021 The Authors. Histopathology published by John Wiley & Sons Ltd. Histopathology, 80, 485–500.
StataCorp, College Station, TX, USA). All other analyses were conducted in R 4.0.3 (R Foundation for Statistical Computing, Vienna, Austria).\textsuperscript{18} Statistical differences between the clinicopathological characteristics of the subset of patients used in this study and those of the overall cohort were determined. The Wilcoxon rank-sum test was applied to those groups with two levels, and Pearson’s chi-squared test without continuity correction or Fisher’s exact test was applied to categorical variables where appropriate. The Kruskal–Wallis rank-sum test was used for the continuous variable.

Descriptive statistical analyses were performed on the number of tumour buds detected per tissue core with each of the scoring methods. Spearman’s correlation coefficient was used to determine the strength of the linear relationships between the different scoring methods.

\textsuperscript{©} 2021 The Authors. Histopathology published by John Wiley & Sons Ltd, Histopathology, 80, 485–500.

\textbf{Figure 1.} Methods of assessment of tumour budding. A–E, Manual methods. F–J, Semi-automated method. Tumour budding was manually scored on haematoxylin and eosin (H&E)-stained (A, B) and cytokeratin (CK)-stained (C–J) tissue microarrays. After dearraying, in all cores the tissue border was shrunk by 30 \( \mu \text{m} \) to exclude candidate buds touching the periphery of the cores. Buds were annotated manually within QuPath (yellow dots in B, D, and E). Initial CK-identified buds (D, CK all) were revisited to exclude those lacking a region of nuclear pallor (E, CK pallor) and generate a second dataset with application of this criterion. A semi-automated workflow was developed in QuPath (F–J). A binary classifier identified discrete CK-positive regions (red). Lumina encapsulated by positive staining were filled in to exclude ‘luminal pseudobuds’ (G). Buds were defined on the basis of the area of CK immunopositivity; the acceptable range (40–700 \( \mu \text{m}^2 \)) was derived from analysis of the range of areas of the manually annotated CK buds. Objects with areas outside this range were excluded, leaving the buds highlighted (H–J).
Univariable and multivariable analyses with the Cox PH regression model were performed to calculate hazard ratios (HRs) and 95% confidence intervals (CIs) for overall survival according to TB. Adjusted models were tested for a family history of CRC, tumour grade (differentiation) and microsatellite instability status, but these factors were excluded because they did not influence the model. Multivariable adjustments were made for age (<50, 50 to <60, 60 to <70, 70 to <80 and ≥80 years), sex (male or female), adjuvant chemotherapy receipt (yes or no), stage (II or III), and Eastern Cooperative Oncology Group performance status (0–1, 2, 3–4, and unknown). As the TMA cores in this study represented random cores from the tumour advancing edge, rather than TB hotspots, the ITBCC three-category cut-offs are not strictly applicable. Therefore, survival analysis was conducted in two ways: (i) on the basis of continuous bud counts to maximise statistical power, with per-increment increases for each method based on the relative ratios of total bud counts between methods; and (ii) with the application of modified ITBCC cut-offs to mimic categorisation of scores for clinical decision-making, and to generate Kaplan–Meier curves for prognostication. The ITBCC three-category cut-offs were used for H&E scores (≤4, 5–9 and ≥10 buds), and cut-offs for the other methods were scaled up according to the TB score distribution for each method.

**Results**

Of the original cohort, 186 individual cases were included in the study analysis. The overall clinicopathological characteristics are summarised in Table 1, which demonstrates that the subset of patient samples used in the current study show no meaningful differences from the overall stage II/III population-based cohort, and can be considered to be a representative subset for analysis.

**Deriving the Bud Area Range for the Semi-Automated Method**

To obtain semi-automated bud counts, a definition of an acceptable range of bud area, derived from analysis of the range of areas of the manually annotated CK buds, was first required. The semi-automated method initially identified all discrete areas of CK immunopositivity. Immunopositive areas, representing candidate buds, were initially captured over a wide size range (5–3000 μm²). Extremely small areas represented either tiny immunopositive tumour fragments, often in the context of gland rupture (Figure 2A,B), or non-specific immunostaining of an uncertain nature (Figure 2C,D). Large tumour areas were also annotated. By mapping of the manual CK annotations to the semi-automated annotations, the areas of all manually annotated CK buds (CK all) could be measured within QuPath (Figure 2E,F) and exported for analysis. The median CK bud area of the manually annotated CK buds (CK all), as measured by QuPath, was 225 μm² (Figure 3A; interquartile range, 133–388 μm²). The images, including manual and semi-automated annotations, of outliers at the low and high ends of the area scale were reviewed, to explain implausibly small and large areas for some manually annotated buds. In some single-cell buds, the semi-automated method excluded from the area measurement a prominent region of central nuclear pallor, thereby underestimating the true bud area (Figure 2G,H). For some closely adjacent buds, QuPath failed to resolve these as separate buds, and considered their total combined area as a single immunopositive region, resulting in an apparent manually detected bud with a large area (Figure 2J). Taking these erroneous extreme values into consideration, we chose a range of 40–700 μm² as acceptable in this study for defining a bud on the basis of the area of CK immunopositivity. After application of this definition, Figure 3B demonstrates, with a histogram, the resultant areas and frequencies of the buds detected with the semi-automated method, which gave a lower modal bud area than that given with the manual CK (all) method.

**Total Bud Count Comparisons**

The total numbers of buds detected with the different methods (Figure 4A), over the 186 TMA cores, were as follows: manual H&E, n = 503; CK all, n = 2290; CK pallor, n = 1825; and semi-automated, n = 5138. These findings indicate that more than four times as many buds were detected with the CK method (CK all) than with the H&E method, and more than three times as many if the count was restricted to those buds with central pallor (CK pallor). The semi-automated method detected >10 times more buds than the H&E method and over twice as many buds as the CK method (CK all). A comparison of bud totals and frequencies for each method showed progressively increasing numbers of cases with higher numbers of buds moving from H&E to CK to semi-automated assessments (Figure 4B). A
| Characteristics          | Entire cohort | Subset for analysis | P-value |
|--------------------------|---------------|---------------------|---------|
| **Sex** (N = 661/186), n (%) |               |                     |         |
| Male                     | 358 (54.2)    | 92 (49.5)           | 0.26    |
| Female                   | 303 (45.8)    | 94 (50.5)           |         |
| **Age (years)** (N = 661/186) | 73 (IQR 64–79) | 74 (IQR 65–80) | 0.18    |
| **ECOG performance status** (N = 410/113), n (%) |       |                     |         |
| 0–1                      | 338 (82.4)    | 92 (81.4)           | 0.86    |
| 2                        | 42 (10.2)     | 11 (9.7)            |         |
| 3–4                      | 30 (7.3)      | 10 (8.8)            |         |
| **Tumour stage** (N = 661/186), n (%) |       |                     |         |
| II                       | 394 (59.6)    | 104 (55.9)          | 0.37    |
| III                      | 267 (40.4)    | 82 (44.1)           |         |
| **Tumour location** (N = 661/186), n (%) |       |                     |         |
| Proximal                 | 375 (56.7)    | 116 (62.4)          | 0.42    |
| Distal                   | 280 (42.4)    | 69 (37.1)           |         |
| Colon unspecified        | 6 (0.9)       | 1 (0.5)             |         |
| **Tumour differentiation** (N = 657/183), n (%) |       |                     |         |
| Poor                     | 90 (13.7)     | 20 (10.9)           | 0.33    |
| Well differentiated/moderately differentiated | 567 (86.3) | 163 (89.1) | |
| **Extramural venous invasion** (N = 610/163), n (%) |       |                     |         |
| Yes                      | 165 (27.0)    | 44 (27.0)           | 0.99    |
| No                       | 445 (73.0)    | 119 (73.0)          |         |
| **MSI status** (N = 593/175), n (%) |       |                     |         |
| MSI-high                  | 136 (22.9)    | 47 (26.9)           | 0.30    |
| MSI-low                   | 10 (1.7)      | 5 (2.79)            |         |
| Microsatellite-stable     | 447 (75.4)    | 123 (70.3)          |         |
| **Adjuvant chemotherapy** (N = 661/186), n (%) |       |                     |         |
| Yes                      | 186 (28.1)    | 51 (27.4)           | 0.85    |
| No                       | 475 (71.9)    | 135 (72.6)          |         |
| **Overall survival** (N = 661/186), n (%) |       |                     |         |
| Alive                    | 354 (53.6)    | 96 (51.6)           | 0.64    |
| Dead                     | 307 (46.4)    | 90 (48.4)           |         |
A comparison of total bud numbers between the H&E and CK methods showed a moderate correlation (Figure 4C: $\rho = 0.60$, $P < 0.0001$), whereas a strong correlation was observed between the CK all and semi-automated methods (Figure 4D: $\rho = 0.81$, $P < 0.0001$).

**Table 1. (Continued)**

| Characteristics ($N =$ entire cohort/subset) | Entire cohort | Subset for analysis | $P$-value |
|---------------------------------------------|--------------|--------------------|-----------|
| CRC-specific survival ($N =$ 566/156), n (%) |              |                    |           |
| Alive                                      | 354 (62.5)   | 96 (61.5)          | 0.82      |
| Dead                                       | 212 (37.5)   | 60 (38.5)          |           |

CRC, colorectal cancer; ECOG, Eastern Cooperative Oncology Group; IQR, interquartile range; MSI, microsatellite instability.

* A distal colonic location was defined as the splenic flexure and beyond. Rectal cancers were excluded from the study.

**Figure 2.** Deriving the bud area range for the semi-automated method. A.C.E.G.I, Unannotated images. B.D.F.H.J, The corresponding annotated images. Red: initial QuPath annotations of cytokeratin (CK) immunopositivity within the broad area range of 5–3,000 $\mu$m$^2$. Yellow: manual bud annotations. A,B, Tumour gland rupture, generating multiple tiny immunopositive fragments. C,D, Tiny immunopositive fragments of an uncertain nature (arrows) detected alongside two true tumour buds (arrowheads). E,F, Six manually annotated (CK all) tumour buds, with areas measured by QuPath (range, 107–384 $\mu$m$^2$). G,H, A manually annotated single tumour cell bud with prominent nuclear pallor resulting in underestimation of the bud area by QuPath (measured as 10 $\mu$m$^2$). I,J, Two closely adjacent buds annotated manually (arrows), but considered by QuPath to be one large immunopositive area (measured as 820 $\mu$m$^2$).
BUD-BY-BUD COMPARISONS

As both manual CK assessments and the semi-automated assessment were performed on the same set of images, bud-by-bud comparison was possible for these methods. A total of 1734 individual buds were identified both with manual assessment (CK all) and semi-automated assessment, representing 75.7% of the total number of buds identified with manual assessment \( (n = 2290) \) and 33.7% of the total number of buds identified with semi-automated assessment \( (n = 5138) \) (Figure 5). If we accept the manual CK method as the relevant gold standard, these equate to the sensitivity and positive predictive value, respectively, of the semi-automated method for detection of CK (CK all) buds.

BUD DISCORDANCE BETWEEN METHODS

Many tumour areas showed excellent concordance, with buds being detected with both manual CK assessment and semi-automated assessment after application of the specified area range for the semi-automated method (Figure 6A,B). However, elsewhere, concordance between these methods was poor. This was in large part due to the semi-automated method accepting as a bud any discrete area of CK immunopositivity within the accepted area range, regardless of shape or crispness of definition, features that would typically be considered in the manual assessment of a bud (Figure 6C,D). The other main explanation for much greater numbers of buds being detected with the semi-automated method relates to ‘luminal pseudobuds’. Manual assessment discounts as buds those tumour cells or clusters lying within glandular lumina. When these were surrounded by circumferential staining, QUPath was able to fill in the glandular lumina, to avoid counting such mimics as buds (Figures 1F,G and 6E,F). However, when staining was not circumferential, QUPath counted these luminal immunopositive fragments as buds (Figure 6G,H). This was a particular problem at the core peripheries, where the complete gland circumference was not captured within the core (Figure 6I,J). The inclusion of the more stringent nuclear pallor criterion to define a CK bud by manual assessment had a minor additional impact on the discordance in bud numbers between manual CK and semi-automated assessments (Figure 5).

A smaller number of buds identified with manual assessment CK (CK all and CK pallor) were not detected with the semi-automated method. This can be explained by erroneous bud area measurement, as described above. Incorrect assessment of the true bud area, because of exclusion of a region of nuclear

---

**Figure 3.** A, A histogram of cytokeratin (CK)-positive bud areas detected manually, as measured by QUPath. This was the basis of selecting a suitable area range for defining a bud with application of the semi-automated assessment method. B, A histogram of areas of buds detected with the semi-automated method, with application of a range of 40–700 \( \mu \text{m}^2 \).
Method | Bud definition | Total number of tumour buds
--- | --- | ---
Manual H&E | Single or a cluster of up to four tumour cells on H&E. | 503
Manual CK all | Single or a cluster of up to four tumour cells on CK. | 2290
Manual CK pallor | Single or a cluster of up to four tumour cells with a visible region of nuclear pallor on CK. | 1825
Semi-automated | A region of CK immunopositivity with an area between 40-700 µm². | 5138

**Figure 4.** A total bud count comparison across all scoring methods. **A,** The total number of tumour buds detected with each of the four methods of assessing tumour budding in all study cases (n = 186). **B,** Total buds per core and the frequencies of each number within the study group, for cores with up to 20 buds detected. **C,** Correlation of total bud counts per core assessed manually with haematoxylin and eosin (H&E) staining and cytokeratin (CK) staining (CK all). **D,** Correlation of total bud counts per core assessed manually with CK staining (CK all) and with the semi-automated method.
pallor (Figure 2G,H) or failure to resolve closely adjacent buds (Figure 2I,J), generated areas below or above the accepted range, and thereby failure to identify these manually detected buds with the semi-automated method.

SURVIVAL ANALYSIS
Of the 186 patients included in the analysis, by the end of follow-up (mean \( \pm \) standard deviation, 5.5 \( \pm \) 3.0 years; range, 0.12–10 years; interquartile range, 2.89–8.19 years), 90 had died, 60 were from a CRC-related cause. All four methods of TB assessment demonstrated reduced survival associated with higher TB scores (Table 2). HRs were similar for both of the CK methods and for the semi-automated method in the univariable model (manual CK all—HR 1.09, 95% CI 1.05–1.14; manual CK pallor—HR 1.11, 95% CI 1.06–1.18; semi-automated—HR 1.09, 95% CI 1.04–1.14) and the multivariable model (manual CK all—HR 1.06, 95% CI 1.02–1.11; manual CK pallor—HR 1.08, 95% CI 1.02–1.14; semi-automated—HR 1.06, 95% CI 1.01–1.11), and slightly lower for the H&E method in both the univariable model (HR 1.03, 95% CI 1.01–1.05) and the multivariable model (HR 1.02, 95% CI 1.00–1.04). All findings were statistically significant except for H&E findings in the multivariable model.

Kaplan–Meier survival analysis showed that patients with higher TB grades had reduced overall 5-year survival, when assessed with any of the four methods presented (Figure 7). Stratification was not significant for H&E assessment (\( P = 0.14 \)) but was significant for the other three methods, all of which showed comparable stratification (\( P = 0.00016, P = 0.00014, \) and \( P = 0.0011 \)). Introduction of nuclear pallor to the manual CK assessment did not meaningfully impact on stratification.

Discussion
TB is well established as an adverse prognostic feature in CRC in several clinical settings.\(^1\) Despite considerable existing evidence in this regard, assessment of TB has not yet been incorporated into routine clinical practice. In large part, this is because of uncertainty regarding the most appropriate method of assessment, specifically the most appropriate stain for counting buds and whether to persist with manual assessment or adopt some form of semi-automated approach. In this study, we used QuPath to develop a new digital pathology-based semi-automated TB assessment tool for CK-stained sections, which we then compared with established methods of TB assessment in a cohort of colon cancers by using a TMA approach. As the study included TMA cores from the tumour advancing edge of stage II/III colon cancers, rather than the budding hotspot advocated for clinical use, the primary focus of this study was a bud-by-bud comparison of manual CK assessment and our semi-automated assessment method, rather than to provide further evidence of adverse prognostic significance of TB.

Our data indicate that CK IHC detected over four times more buds than H&E-based assessment of parallel sections, which is consistent with previous studies observing three to six times more buds with CK IHC than with H&E staining.\(^12\) Although not examined in this study, it is postulated that CK IHC is particularly valuable in highlighting single-cell buds and distinguishing these from epithelioid stromal or histiocytic cells by indicating their epithelial cell lineage, which is less readily apparent on H&E assessment. Bokhorst \textit{et al.} have hypothesised that interobserver variability on H&E assessment may be more problematic for single-cell buds than for two- to four-cell buds.\(^13\) H&E assessment allows better evaluation of the microenvironment surrounding buds, and so it is possible that a further reason contributing to fewer buds being identified with H&E assessment relates to greater exclusion of so-called pseudobuds at sites of active inflammation, often related to gland rupture.\(^1\) The inflammatory environment is less readily appreciated in CK IHC.
preparations, meaning that pseudobuds may be less identifiable and therefore less likely to be excluded.

The threshold semi-automated approach identified approximately 2.5 times more buds than manual CK assessment. Higher bud counts have been observed previously when a semi-automated method has been compared with a manual CK assessment method, but without quantification.19 We found that a bud-by-bud comparison revealed only moderate agreement between these two assessment methods for individual buds. Some of the discrepancy might be explained by the tendency of any human observer to err slightly on the side of undercounting, either by occasionally missing a possible true bud or by making a conservative judgement in an ambiguous case. In contrast, one can expect a threshold-based approach, calibrated to identify true buds on the basis of CK immunopositivity, to err definitively on the side of overestimation, because it will consistently include more irregular or ill-defined ambiguous tumour cell clusters of a defined size. It is possible that incorporating further criteria into the bud definition may improve agreement between semi-automated and manual assessments, such as a measure of circularity.20 However, given that there is no *a priori* reason to suppose that buds are circular, this could introduce
Table 2. Univariable and multivariable Cox proportional hazards overall survival analysis comparing four methods of tumour budding assessment

| Assessment method          | Overall survival [hazard ratio (95% CI)] | Unadjusted | Multivariable adjusted model |
|----------------------------|-----------------------------------------|------------|-----------------------------|
| Manual H&E                 |                                         |            |                             |
| Per one-bud increase       | 1.03 (1.01–1.05)                        | 1.02 (1.00–1.04) | 0.001 0.08 |
| P-value                    | <0.001                                  | 0.004      |                             |
| Manual CK all              |                                         |            |                             |
| Per five-bud increase      | 1.09 (1.05–1.14)                        | 1.06 (1.02–1.11) | <0.001 0.005 |
| Manual CK pallor           |                                         |            |                             |
| Per five-bud increase      | 1.11 (1.06–1.18)                        | 1.08 (1.02–1.14) | <0.001 0.005 |
| Semi-automated             |                                         |            |                             |
| Per 10-bud increase        | 1.09 (1.04–1.14)                        | 1.06 (1.01–1.11) | <0.001 0.01 |

CI, confidence interval; CK, cytokeratin; H&E, haematoxylin and eosin.

Furthermore, subjectivity. In this study, we aimed to minimise the adjustable parameters, relying primarily upon a staining threshold and area filter to achieve a replicable baseline of quantitative assessment. The area range that we selected to define a tumour bud (40–700 μm²) was based on the corresponding area range of buds identified with manual CK assessment, which is wider than that chosen by Takamatsu et al. (100–480 μm²) but narrower than that chosen by Bokhorst et al. (25–1000 μm²).2 This already indicates the lack of accepted parameters in defining bud characteristics by the use of image analysis, although such parameters will inevitably have a profound influence on the absolute numbers of buds detected. Interestingly, we found that, despite the substantial differences in absolute bud counts between the assessment methods, correlation remained high, suggesting that the signal remains high amid the noise.

As there is evidence to support a high TB score as an adverse prognostic factor across all stages of CRC,1,3,4 survival analysis was conducted with the four methods of TB assessment, as a measure of comparative performance. Despite the limitations of random core sampling, TB assessed with all four methods was, as expected, significantly associated with reduced overall survival at 5 years of follow-up. This association was weakest for H&E assessment, and non-significant in the multivariable model, but it is likely that H&E assessment, with the lowest bud counts in general, will have been impacted more by the random core approach in our study than the other methods yielding much higher bud counts. Nevertheless, the other three methods all stratified patients better than H&E assessment with respect to survival, and achieved almost identical HRs based on evaluation of continuous bud counts. Importantly, despite its simplicity and only moderate agreement with manual CK assessment for individual buds, the semi-automated threshold approach in QuPath provided an association between higher grades of TB and worse overall patient survival, even when applied to random tumour cores.

A recent modified Delphi process conducted among an international group of expert gastrointestinal pathologists supported ongoing assessment of TB with H&E-stained slides, with more evidence being required to move to IHC, but also suggested that digital image analysis was likely to facilitate implementation in clinical practice.11 As almost all TB algorithms published to date rely on CK-stained rather than H&E-stained images, it seems likely that the optimal approach will ultimately be one based on evaluation of the most representative tumour section, stained for CK. With increasing developments in digital pathology and growing access to digital whole slide images in routine practice, some form of semi-automated approach is attractive for reasons of efficiency, cost, and reproducibility. Such semi-automated methods can be easily applied over a much larger tumour area to accurately identify the TB density over any agreed area denominator. The consensus 0.785 mm² area applicable to microscopy is less relevant to whole slide image analysis. Nevertheless, most current evidence for TB significance is based on this hotspot area, and correlation with microscopy assessment of TB will be important for the foreseeable future.

It is likely that the semi-automated approach to TB assessment described in this study is overly simplistic for clinical use, as it is unable to detect some of the more subtle morphological features of tumour buds, such as nuclear pallor, or exclude mimics such as pseudobuds. Future clinical implementation will require more refined methodologies, probably involving deep learning,9,21; however, as yet no such method is widely available to the TB community. The
Semi-automated QuPath approaches developed and applied in this study will be of potential benefit to ongoing translational TB research in retrospective cohorts as a much cheaper, more efficient and readily customisable open-source method than commercial software solutions. Such tools can be utilised either as a stand-alone TB assessment method or as an adjunct to the development of more sophisticated methods. © 2021 The Authors. Histopathology published by John Wiley & Sons Ltd, Histopathology, 80, 485–500.

Figure 7. Kaplan–Meier estimates demonstrating overall survival differences in patients with stage II/III colon cancer according to low-grade, moderate-grade and high-grade tumour budding as assessed with four different methods. International Tumour Budding Consensus Conference three-category cut-offs were used for haematoxylin and eosin (H&E) scores (≤4, 5-9 and ≥10 buds), and cut-offs for the other methods were scaled up according to the total budding score distribution for each method (Figure 4A).
methods, e.g. by identifying large numbers of candidate buds for consensus expert evaluation, classification and application to training of deep learning algorithms.

Assessment of TB with CK IHC has been shown by some studies to improve interobserver reproducibility, which is an important requirement when the incorporation of any new parameter into routine pathology practice is being considered. However, a recent study employing CK IHC for TB assessment examined interobserver agreement at the individual bud level and found only moderate agreement, which was no better than that for H&E assessment. The authors considered two reasons for this: first, that individual tumour nuclei within immunopositive clusters are sometimes difficult to discern, and therefore count, on CK-stained sections; and second, that the surrounding inflammatory environment is more difficult to assess on CK-stained sections than on H&E-stained sections, making evaluation of potential ‘pseudobudding’ more challenging. Less evidence is available on the reproducibility of semi-automated methods but it is intuitive that more automation implies greater reproducibility. Takamatsu et al. found significantly better reproducibility among three pathologists when their semi-automated method was used (kappa coefficient of 0.781) than when manual assessment was used (kappa coefficient of 0.463). Nevertheless, some degree of manual oversight remains important while new methods are developed and tested.

Introducing the additional criterion of nuclear pallor into the manual CK assessment method made no meaningful alteration to the resultant HR (CK pallor — HR 1.11, 95% CI 1.06–1.18; CK all—HR 1.09, 95% CI 1.05–1.14) or the Kaplan–Meier survival stratification, providing no real evidence from this study for inclusion of this criterion. Previously suggested by Bokhorst et al. to help exclude CK-positive non-viable tumour cell fragments from consideration as buds, this feature should be the focus of future studies based on hotspot TB assessment on whole tumour sections from appropriate CRC cohorts, in order to ascertain the potential impact of this morphological criterion on the clinical relevance of TB and to inform future discussions on bud definition.

This study is limited by the random nature of the tumour core samples, limiting analysis of the clinical significance of TB scores with respect to survival analyses, and by the single-pathologist manual assessments, without any ability to assess reproducibility. However, a detailed comparison of different TB assessment methods is described, applied to a wide morphological spectrum of colon cancers, with bud-by-bud comparison between methods.

Although our CK thresholding approach resembles methods applied in previous TB studies, to our knowledge the current study is the first to describe an interactive tool for TB assessment that is freely available, open-source, and can be readily applied to whole

Figure 8. Tumour budding assessment applied to a whole slide cytokeratin (CK)-stained image of colorectal cancer. A high-grade budding case has been chosen for illustration. A, After manual annotation of the advancing edge (red line) with the QuPath line tool, the expand annotation tool is used to expand the annotation 1 mm inwards and outwards, delineating the tumour advancing edge region of interest (within the yellow outline) for budding assessment. Manually identified (yellow circles) and independently detected QuPath (red shapes) buds are shown (magnified in the inset for ‘hotspot’ area). B, A bud density heat map based on manual bud annotations. C, A bud density heat map based on QuPath bud annotations. Density colourmaps are normalised independently for each image according to the range of bud density within the image. The 0.785 mm² ‘hotspot’ is highlighted (black circle) in each image.

© 2021 The Authors. Histopathology published by John Wiley & Sons Ltd, Histopathology, 80, 485–500.
slide images as part of a full analysis workflow. This is possible because of the extensive additional functionality within QuPath, including the ability to precisely define regions of interest (e.g., a 1 mm boundary delineating the tumour advancing edge), identify hotspots, and export quantitative metrics. These features are illustrated in Figure 8, showing the application of the methods adopted in this study to a whole slide image from a sample CRC case rich in tumour buds. Manually derived and semi-automated budding density ‘heat maps’ are almost identical. In contrast to assessment approaches driven entirely by machine learning, which can be confounded by even subtle variations in staining or scanning,24,25 our comparatively simple thresholding method can be readily adapted to new images by adjusting a small number of intuitive parameters, making it immediately accessible to any laboratory wishing to apply the technique. Nevertheless, it is clearly desirable to achieve better discrimination of true buds from false positives. In this regard, QuPath’s generic support for machine learning, previously described for cell classification,15 can be incorporated into a more elaborate analysis workflow. Having established in this study the first open and replicable end-to-end analysis protocol for TB assessment suitable for whole slide images, we aim to collaborate with other groups to develop a refined, open-source bud identification algorithm based on a more diverse training dataset across multiple centres.

In conclusion, we present a new, semi-automated, QuPath-based approach to TB assessment. This shows moderate agreement with manual CK-based assessment at a bud-by-bud level and comparable ability to stratify a cohort of patients with stage II/III colon cancer for overall survival. More importantly, it shows QuPath’s potential as a freely available, rapid and transparent tool for TB assessment, applicable to whole slide images, that can be used in translational research as a stand-alone method or as an aid in developing future approaches suitable for clinical implementation.

Acknowledgements

The samples used in this research were received from the Northern Ireland Biobank, which has received funds from HSC Research and Development Division of the Public Health Agency in Northern Ireland and the Friends of the Cancer Centre. The Belfast Health and Social Care Trust Department of Cellular Pathology is acknowledged for assisting with immunohistochemistry.

Conflicts of interest

The authors have no conflicts of interest to declare.

Funding

The study cohort creation was enabled by funding from Cancer Research UK (ref. C37703/A15333 and C50104/A17592) and a Northern Ireland HSC R&D Doctoral Research Fellowship (ref. EAT/4905/13). This work was supported by the Queens University Belfast Foundation (P. D. Dunne and N. C. Fisher; Musgrave scholarship), a Cancer Research UK early detection grant (P. D. Dunne; A29834), a Cancer Research UK Career Establishment Award (H. G. Coleman; C37703/A25820), and the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation (awarded to P. Bankhead; M. D. Gelbard was funded by grant number 2019-207148).

Author contributions

Study concept and design: N. C. Fisher, M. B. Loughrey, P. Bankhead, and P. D. Dunne. Development of methodology: N. C. Fisher, M. B. Loughrey, P. Bankhead, and P. D. Dunne. QuPath software development: M. D. Gelbard and P. Bankhead. Acquisition, analysis and interpretation of data, and statistical analysis: N. C. Fisher, M. B. Loughrey, H. G. Coleman, and P. D. Dunne. Writing of the initial draft: N. C. Fisher and M. B. Loughrey. Review and revision of subsequent drafts: N. C. Fisher, M. B. Loughrey, H. G. Coleman, M. D. Gelbard, P. Bankhead, and P. D. Dunne. All authors read and approved the final manuscript.

Data Availability Statement

The QuPath scripts used in the current study are publicly available at https://github.com/petebankhead/qupath-budding-scripts. The images and datasets used in the current study are available on application to the Northern Ireland Biobank.

References

1. Lugli A, Zlobec I, Berger MD, Kirsch R, Nagtegaal ID. Tumour budding in solid cancers. Nat. Rev. Clin. Oncol. 2021; 18: 101–115.
2. Bosch S, Teerenstra S, de Wilt J, Cunningham C, Nagtegaal I. Predicting lymph node metastasis in pT1 colorectal cancer: a systematic review of risk factors providing rationale for therapy decisions. Endoscopy 2013; 45: 827–834.
3. Landau MA, Zhu B, Akwuole FN, Pai RK. Histopathological predictors of recurrence in stage III colon cancer: reappraisal of tumor deposits and tumor budding using AJCC8 criteria. *Int. J. Surg. Pathol.* 2019; 27: 147–158.

4. Van Wyk HC, Park JH, Edwards J, Horgan PG, McMillan DC, Going JJ. The relationship between tumour budding, the tumour microenvironment and survival in patients with primary operable colorectal cancer. *Br. J. Cancer* 2016; 115: 156–163.

5. Brown I, Kumarasinghe P, Rosty C et al. Colorectal cancer structured reporting protocol. 4th ed. Sydney: The Royal College of Pathologists of Australasia. 2020.

6. Loughrey M, Quirke P, Shepherd NA. Dataset for histopathological reporting of colorectal cancer. 4th ed. London: Royal College of Pathologists, 2018.

7. Burgart LJ, Kakar S, Shi C et al. Protocol for the examination of resection specimens from patients with primary carcinoma of the colon and rectum. Northfield, IL, USA: 4.1.0.0. College of American Pathologists. 2020.

8. Lugli A, Kirsch R, Ajioka Y et al. Recommendations for reporting tumor budding in colorectal cancer based on the International Tumor Budding Consensus Conference (ITBCC) 2016. *Mod. Pathol.* 2017; 30: 1299–1311.

9. Studer L, Blank A, Bokhorst J-M et al. Taking tumour budding to the next frontier—a post International Tumour Budding Consensus Conference (ITBCC) 2016 review. *Histopathology* 2021; 78: 476–484.

10. Koelzer VH, Assarazadegan N, Dawson H et al. Cytokeratin-based assessment of tumour budding in colorectal cancer: analysis in stage II patients and prospective diagnostic experience. *J. Pathol. Clin. Res.* 2017; 3: 171–178.

11. Haddad TS, Lugli A, Aherne S et al. Improving tumor budding reporting in colorectal cancer: a Delphi consensus study. *Virchows Arch.* 2021; 479: 459–469.

12. Koelzer VH, Zlobec I, Berger MD et al. Tumor budding in colorectal cancer revisited: results of a multicenter interobserver study. *Virchows Arch.* 2015; 466: 485–493.

13. Bokhorst JM, Blank A, Lugli A et al. Assessment of individual tumor buds using keratin immunohistochemistry: moderate interobserver agreement suggests a role for machine learning. *Mod. Pathol.* 2020; 33: 825–833.

14. Gray RT, Cantwell MM, Coleman HG et al. Evaluation of PTGS2 expression, PIK3CA mutation, aspirin use and colon cancer survival in a population-based cohort study. *Clin. Transl. Gastroenterol.* 2017; 8: e91.

15. Bankhead P, Loughrey MB, Fernández JA et al. QuPath: open source software for digital pathology image analysis. *Sci. Rep.* 2017; 7: 16878.

16. Lewis C, McQuaid S, Clark P et al. The Northern Ireland biobank: a cancer focused repository of science. *Open J. Bioresour.* 2018; 5: 1–6.

17. Ruifrok AC, Johnston DA. Quantification of histochemical staining by color deconvolution. *Anal. Quant. Cytol. Histol.* 2001; 23: 291–299.

18. R Core Team. *R: a language and environment for statistical computing*. 4th ed. 2020. Available at: https://www.R-project.org/

19. Jepsen RK, Klarskov LI, Lippert MF et al. Digital image analysis of pan-cytokeratin stained tumor slides for evaluation of tumor budding in pT1/pT2 colorectal cancer: results of a feasibility study. *Pathol. Res. Pract.* 2018; 214: 1273–1281.

20. Takamatsu M, Kawachi H, Yamamoto N et al. Immunohistochemical evaluation of tumor budding for stratifying T1 colorectal cancer: optimal cut-off value and a novel computer-assisted semiautomatic method. *Mod. Pathol.* 2019; 32: 675–683.

21. Pai RK, Hartman D, Schaeffer BF et al. Development and initial validation of a deep learning algorithm to quantify histological features in colorectal carcinoma including tumour budding/poorly differentiated clusters. *Histopathology* 2021; 79: 391–405.

22. Kai K, Aishima S, Aoki S et al. Cytokeratin immunohistochemistry improves interobserver variability between unskilled pathologists in the evaluation of tumor budding in T1 colorectal cancer. *Pathol. Int.* 2016; 66: 75–82.

23. Vranes V, Rajković N, Li X et al. Size and shape filtering of malignant cell clusters within breast tumors identifies scattered individual epithelial cells as the most valuable histomorphological clue in the prognosis of distant metastasis risk. *Cancers* 2019; 11: 1–16.

24. Cui M, Zhang DY. Artificial intelligence and computational pathology. *Lab. Invest.* 2021; 101: 412–422.

25. Janowczyk A, Zuo R, Gilmore H, Feldman M, Madabhushi A. HistoQC: an open-source quality control tool for digital pathology slides. *JCO Clin. Cancer Inform.* 2019; 3: 1–7.

### Supporting Information

Additional Supporting Information may be found in the online version of this article:

**Figure S1.** Representative images of cores that did not meet the inclusion criteria.