Validation of the systematic scoring of immunohistochemically stained tumour tissue microarrays using QuPath digital image analysis

Loughrey, M. B., Bankhead, P., Coleman, H. G., Hagan, R. S., Craig, S., McCorry, A. M. B., Gray, R. T., McQuaid, S., Dunne, P. D., Hamilton, P. W., James, J. A., & Salto-Tellez, M. (2018). Validation of the systematic scoring of immunohistochemically stained tumour tissue microarrays using QuPath digital image analysis. *Histopathology*. Advance online publication. https://doi.org/10.1111/his.13516

Published in:
Histopathology

Document Version:
Peer reviewed version

Validation of the systematic scoring of immunohistochemically-stained tumour tissue microarrays using *QuPath* digital image analysis

Loughrey Maurice B.[1,2,*], Bankhead Peter[3,*], Coleman Helen G.[3,4], Hagan Ryan S.[3], Craig Stephanie[1], McCorry Amy[3], Gray Ronan T.[4], McQuaid Stephen[1,2], Dunne Philip D.[3], Hamilton Peter W.[3,5], James Jacqueline A.[1,2,**], , Salto-Tellez Manuel[1,2,**].

* joint first authors
** joint senior authors

Affiliations:
[1]Northern Ireland Molecular Pathology Laboratory, Centre for Cancer Research and Cell Biology, Queen's University Belfast, Belfast, Northern Ireland, UK
[2]Cellular Pathology, Belfast Health and Social Care Trust, Belfast, Northern Ireland, UK
[3]Centre for Cancer Research and Cell Biology, Queen's University Belfast, Belfast
[4]Centre for Public Health, Queen's University Belfast, Belfast, Northern Ireland, UK
[5]Philips Digital Pathology Solutions Belfast, Northern Ireland, UK.

Disclosure/Duality of Interest

Maurice B. Loughrey and Manuel Salto-Tellez are advisors to PathXL Ltd/Philips.

Peter W Hamilton is Founder and Director in PathXL Ltd. and Business Lead for Philips Digital Pathology Solutions Belfast, Northern Ireland, UK. The opinions expressed in this presentation are solely those of the author or presenters, and do not necessarily reflect those of Philips. The information presented herein is not specific to any product of Philips or their intended uses. The information contained herein does not constitute, and should not be construed as, any promotion of Philips products or company policies.

The remaining authors declare no conflict of interest.

Running title*: QuPath* digital image analysis for tissue microarrays

Corresponding author:
Dr. Maurice Loughrey
Consultant Histopathologist
2nd floor Institute of Pathology,
Royal Victoria Hospital
Belfast Trust
Grosvenor Road
Belfast BT12 6BA
United Kingdom
maurice.loughrey@belfasttrust.hscni.net

Abstract

Precision medicine requires a robust, standardised and reproducible assessment of predictive and prognostic biomarkers, to replace current laborious and inaccurate manual scoring approaches. Output from common tissue-based biomarker studies involving immunohistochemistry applied to tissue microarrays (TMA) is limited by the lack of an efficient and reproducible scoring methodology. In this study, we examine the functionality and reproducibility of biomarker scoring using the new, open source, digital image analysis software, *QuPath*.

Three different reviewers, with varying experience of digital pathology and image analysis, applied an agreed *QuPath* scoring methodology to CD3 and p53 immunohistochemically stained TMAs from a colon cancer cohort (n=661). Manual assessment was conducted by one reviewer for CD3. Survival analyses were conducted and intra- and inter-observer reproducibility assessed.

Median raw scores for CD3 and p53 differed significantly between reviewers but this had little impact on subsequent analyses. Lower CD3 scores were detected in cases who died from colorectal cancer, compared to control cases, and this finding was significant for all three reviewers (p-value range 0.002-0.02). Higher median p53 scores were generated amongst cases who died from colorectal cancer compared with controls, but this finding was borderline or non-significant for all three reviewers (p-value range 0.04-0.12). The ability to dichomotise cases into high versus low expression of CD3 and p53 showed excellent agreement between all three reviewers (Kappa score range 0.82-0.93). All three reviewers produced dichotomised expression scores that resulted in very similar hazard ratios and 95% confidence intervals for colorectal cancer-specific survival for each biomarker. Results from manual and *QuPath* methods of CD3 scoring were comparable, but *QuPath* scoring revealed stronger prognostic stratification compared to manual scores, suggesting greater scoring accuracy compared to the manual method.

Scoring of immunohistochemically stained tumour TMAs using *QuPath* is functional and reproducible, even amongst users of limited experience in digital pathology and image analysis, and more accurate than manual scoring. Digital image analysis, such as that performed here, has the potential to alleviate a major bottleneck in biomarker discovery and validation studies.

Introduction

Tissue-based biomarker studies represent a cornerstone of current cancer research strategies. Many such studies involve immunohistochemistry (IHC) applied to tissue microarrays (TMA), to allow high throughput analysis of multiple tumour samples. [1] Output from such research is limited by the lack of an accurate, robust and efficient automated method of scoring IHC-stained markers on TMAs, to replace current laborious and expensive manual approaches, usually conducted by pathologists. This represents a major bottleneck to biomarker discovery and validation studies.

Developments in digital image analysis afford the potential to overcome some of the major drawbacks associated with manual scoring. [2] Until recently, the field of digital pathology lacked an open and accessible bioimage analysis software platform designed to meet the unique challenges involved in analysing ultra-large 2D whole slide images (up to 40GB uncompressed data per slide), the mainstay of digital pathology.

QuPath (https://qupath.github.io) is new, comprehensive digital pathology image analysis software, developed at Queen's University Belfast to address these needs, offering an open-source bioimage analysis platform that improves the speed, objectivity and reproducibility of digital pathology analysis and is capable of handling whole slide images. [3] Its functionality also permits training and subsequent classification of tumour and non-tumour cells using automated digital algorithms, with instant mark-up available for continuous review of the training process, in a manner amenable to visual inspection and quality control by a pathologist. Evaluation of IHC-stained TMAs is just one of many potential applications of *QuPath*, but one we see as most valuable to biomarker discovery research. As with any new methodology, validation by demonstration of inter-observer reproducibility and by intra-observer comparison with the current gold standard, manual assessment in this case, is an essential step towards acceptance into common usage.

Recent studies have evaluated the application of *QuPath* to the scoring of a variety of different biomarkers in breast and colon cancer tissue cohorts, mainly in TMA format. [3-6] In this study, we aim to demonstrate the functionality and reproducibility of biomarker scoring using *QuPath* image analysis involving users from different backgrounds (pathology, computer science, biochemistry) and with varying levels of

experience of digital image analysis. Two well-established and biologically distinct tissue biomarkers, CD3 and p53, were evaluated in IHC-stained TMAs from a large cohort of colon cancers by the three study reviewers using *QuPath*, and inter-observer reproducibility of raw scores and downstream survival analyses assessed. Manual scoring was conducted by one participant and manual versus *QuPath* method intra-observer reproducibility assessed. Comparison of prognostic stratification applying manual and *QuPath* scores for the established prognostic marker CD3, provides an external benchmark for assessment, related to patient outcomes.

Materials and Methods

Patient Cohort

The study utilised an established cohort of 661 stage II and III colon cancer cases, whose surgical resection specimens were retrieved as part of the creation of the Northern Ireland Colon Cancer Tissue and Data resource. The methods for the creation of this cohort have been described elsewhere.[5, 6]In brief, this resource was created using a population-based study design whereby the cases retrieved were representative of all patients with stage II or III colon cancer who were diagnosed and underwent surgical resection between 2004 and 2008. The 661 study cases represent 46% of all stage II and III colon cancer cases diagnosed in Northern Ireland during this time period. By the end of 2013, 212 individuals had died from a colorectal cancer-specific cause, with 449 individuals either still alive or having died from another cause during this timeframe (hereafter referred to as controls).

Tissue Microarray Creation

Representative tumour blocks were selected following review of all original glass slides from the surgical resection specimens and recut haematoxylin and eosin-stained sections annotated for TMA construction. Three 1mm cores from the centre region of the tumour in each block were targeted, providing triplicate sampling of all 661 tumours, totalling n=1983 tumour cores for scoring, represented in a total of 21 TMA blocks.

Immunohistochemistry

4μm sections from each TMA block were immunohistochemically stained for CD3 and p53. p53 IHC was performed on a Leica BondMax using the DO-7 antibody clone to p53 (Dako UK Ltd, Ely, UK: ER1 30mins polymer kit detection) and CD3

IHC was performed on a Ventana BenchMark using the antibody clone 2GV6 (Ventana; CC1 32minutes, Optiview detection), as previously described.[6]

These biomarkers were chosen as both are well characterised and antibodies are in routine clinical use in most pathology laboratories. CD3 is well established as a prognostic marker in colorectal cancer, high CD3-positive T-cell tumour infiltrates being associated with more favourable clinical outcomes. [7, 8] The prognostic value of p53 expression in colorectal cancer is less well established. [9, 10] More importantly, for the purposes of this study in assessing *QuPath* versatility, these biomarkers demonstrate markedly different immunohistochemical staining patterns. CD3 exclusively stains T-lymphocytes, with uniform intensity, and with no staining of tumour cells or other non-lymphoid tissue. p53 stains tumour and non-tumour cell populations (in particular reactive lymphoid cells) and staining is of variable intensity ranging from negative to strongly positive. These different staining patterns and cell populations require different approaches to scoring, whether adopting manual or digital image analysis approaches.

Image Analysis

Stained TMA sections were scanned using the Aperio ScanScope CS whole slide scanner at x40 magnification. Full details of *QuPath* including source code, documentation, links to the software download and illustrative video supplements are available at https://qupath.github.io. The scanned images were dearrayed within *QuPath* and all cores examined during the scoring process to manually exclude those with either no tumour represented or with artefacts (tissue folding, for example) precluding assessment.

Many different scoring methods are possible within *QuPath* and careful consideration is required to select the scoring method most appropriate to the biomarker and tissue in question. Different approaches were selected for CD3 and p53 (Figure 1). The single target cell population and uniform intensity staining pattern of CD3 lends itself to assessment by a simple density method within

QuPath (Figure 1A). This involved two steps: firstly, the total tissue area (tumour and non-tumour) in each core was detected using the *Simple tissue detection* command; secondly, numbers of cells staining positively for CD3 were counted using the *Fast cell counts* command. From these data, CD3 density was expressed as numbers of CD3-positive cells per mm^2 of tissue. A corresponding script was then generated to run on all imported TMA images, thereby automating the detection and export steps across slides.

The more complex staining pattern of p53 required a different approach (Figure 1B). Firstly, as staining is evident in tumour and non-tumour cell populations, a classification step was required, using training and automated algorithms in *QuPath* to distinguish the tumour epithelial cell population of interest from background inflammatory and stromal cell populations. This entailed application of the *Cell detection* command to identify all cells in all cores based upon nuclear staining, then, using selected measurements of intensity and morphology for all cells, applying a two-way random trees classifier to interactively train *QuPath* to distinguish tumour from non-tumour cells. Secondly, to allow for variation in staining intensity, the "H-score" method of scoring was applied. [11] This requires manual calibration of negative, weak, moderate and strong immunostaining thresholds based upon mean nuclear DAB optical densities, then the H-score is calculated from the extent and intensity of nuclear staining, whereby H-score = 3 x % of strongly staining cells + 2 x % of moderately staining cells + 1 x % of weakly staining cells, giving a score range of 0 to 300. In contrast to manual estimation of H-scores, *QuPath* provides a precise cell-by-cell total in this calculation.

Using these agreed *QuPath* approaches, but without any further consultation on which cores to include/exclude from scoring or which parameters within the software to apply, CD3 and p53-stained TMAs were scored independently by three reviewers, with differing backgrounds (pathology, biochemistry and computer science, reviewers 1-3 respectively) and differing levels of experience with *QuPath*, the computer scientist (reviewer 3) having written the *QuPath* programme, the pathologist (reviewer 1) having two years' experience in using *QuPath* and the biochemist (reviewer 2) having received only two weeks training in *QuPath* prior

to this study. The data from reviewer 3, using this scoring method, has been utilised and published in one prior study. [3] Manual assessment of stained TMA sections was conducted by one reviewer (reviewer 2) for CD3 only, using a 1 (low), 2 (moderate) and 3 (high) global assessment of numbers of positively staining cells.

Statistical Analysis

QuPath data output includes total numbers of cells (staining positive or negative) counted within each core. To help ensure robustness of data, any cores with fewer than 200 cells detected were removed from the statistical analysis. The median score from the remainder of the triplicate cores for each case was then derived for each independent reviewer, and for each marker. For the purposes of intra-observer comparison in this study, analysis was then further restricted to cases for which a score was available for all three reviewers. If any one reviewer excluded all three triplicate cores in a case, all reviewers' scores for that case were excluded. This resulted in a small number of exclusions, with 648 and 657 cases remaining for CD3 and p53 analysis, respectively.

Median expression scores for each biomarker according to each independent reviewer were compared between colorectal cancer deaths and controls using the Wilcoxon-rank sum test. For each reviewer, the median biomarker score within controls was then used to derive dichotomous categories for high and low CD3 and p53 expression. Kappa values were then derived to compare the inter-observer ability to dichotomise individuals into high or low expression of each biomarker. Unadjusted Cox proportional hazards models were then applied to evaluate the association between high versus low biomarker expression and colorectal-specific survival, according to the scores created by each of the three independent reviewers. Kaplan-Meier curves were created to visualise survival analysis results. Finally, comparative analysis of manual and automated scoring of CD3 by one reviewer was also conducted. *QuPath* does have internal capabilities for deriving cut-offs and producing survival curves for statistical analysis but, in this study,

scores were exported into Stata version 14.2 (StataCorp, College Station, TX, USA) for statistical analysis by a further member of the study team independent from the three reviewers.

Results

Median raw scores for CD3 and p53 were highly comparable between reviewers 1 and 3 but differed significantly for reviewer 2 (Table 1). However, this had little impact on subsequent analyses, with data from all three reviewers yielding similar results. Lower CD3 scores were detected in cases who died from colorectal cancer, compared to control cases, and this finding was significant for all three reviewers (p-value range 0.002-0.02). Higher median p53 scores were generated amongst cases who died from colorectal cancer compared with controls, but this finding was borderline or non-significant for all three reviewers (p-value range 0.04-0.12).

Similarly, all three reviewers produced dichotomised expression scores that resulted in very similar hazard ratios and 95% confidence intervals for colorectal cancer-specific survival for each of the biomarkers studied (Table 2). There were some small differences in the magnitude of hazard ratios detected. For example, the reduced risk of colorectal cancer-specific death varied from a 36% reduced risk (HR 0.64) to a 28% reduced risk (HR 0.72) for high compared with low CD3 expression, however all three reviewers detected significant reductions in risk of death (p-value range 0.002-0.02). The variation in magnitude of hazard ratios was smaller for p53 expression, ranging from 29-33% increased risk of death for high compared with low p53 H-scores. These analyses all produced borderline significant associations (p-value range 0.04-0.07). Kaplan-Meier survival curves, based on data from each reviewer, demonstrated highly similar patterns of curve separation for each biomarker (Figures 2 & 3).

Despite variations in raw scores, the ability to dichomotise cases into high versus low expression of CD3 and p53 showed excellent agreement between all three reviewers (Kappa score range 0.82-0.93, Table 3). Similar results were found when Kappa values were generated to compare scoring in only the colorectal cancer deaths, or only the control group (data not shown).

Results from manual and *QuPath* methods of CD3 scoring, conducted by the same reviewer, were highly comparable (Table 4). The highest category of CD3 staining was significantly associated with a reduced risk of colorectal cancer-specific death when both manual and *QuPath* scoring methods were applied. However, *QuPath* scoring had greater ability to distinguish low and medium CD3 scoring, resulting in a stronger dose-response association with colorectal cancer death than manual scoring (p-value for trend 0.003 compared with 0.06).

Discussion

Precision medicine requires robust, standardised and reproducible assessment of predictive and prognostic biomarkers. The lack of agreement on such biomarker scoring has been a longstanding issue both in research and clinical domains. Even amongst those progressing to commercial production and clinical use, many biomarkers have been poorly validated and there is uncertainty about the interpretation of immunohistochemical test results, in particular. [12, 13]

With increasing availability and decreasing costs of whole slide scanning facilities, digital images of stained sections are readily available for analysis and digital image analysis lends itself to reproducibility. There is a need, therefore, for digital image analysis software which is accessible, versatile, transparent, applicable to all common image file types and can handle large file sizes, to meet the challenges of standardising biomarker assessments in research and clinical settings.[14] In the research context, the development of open source software is essential to provide the scientific community with freely-available tools that can be utilised, interrogated and customised for both established and novel applications; however, the user-friendliness of such open source tools can often be lacking, which limits their widespread use. [15] *QuPath* has been created to address this need for whole slide image analysis, and several recent studies have illustrated some of its potential for TMA biomarker scoring. [3, 5, 6] In this study, we have examined some important issues around the use and reproducibility of *QuPath* for biomarker scoring, with a view to bringing this new software application to the attention of a broader pathology audience.

Three different reviewers, with varying experience of pathology and digital image analysis, ranging from minimal to many years, applied the same agreed *QuPath* scoring methodology to a large colon cancer cohort presented in TMA form, immunohistochemically stained for two well-characterised and biologically diverse biomarkers, CD3 and p53. We found considerable variation in raw *QuPath* scores, expressed as positive cell density for CD3 and H-score for p53. This can be attributed to the different ways in which the software was used; nevertheless, across the full cohort these differences had little impact on subsequent survival analysis for either biomarker, which demonstrated findings consistent with existent literature. Excellent inter-observer reproducibility was achieved in dichotomising CD3 and p53 scores. Comparing intra-observer manual and *QuPath* scoring for CD3 revealed considerably stronger prognostic stratification for *QuPath* scores, suggesting greater scoring accuracy compared to the manual method.

The observed variation in raw scores was explored further by review of annotated *QuPath* images for each of the three study participants. This variation was explained by (a) selection of different cores for scoring within the triplicate cores available for each case or (b) selection of different analysis parameters (e.g. thresholds) within *QuPath* settings for cell detection (CD3 and p53) and for staining intensity calibration (p53 only). Accurate selection of appropriate parameters requires experience of evaluating pathology images and IHC in particular. Selection of different annotated regions for cell classification during p53 scoring may have influenced the accuracy of the tumour/non-tumour classification and contributed to differential raw scores. As a result of this additional cell classification step, one may have expected greater inter-observer variation in p53 scoring than in CD3 scoring, but the converse was observed (Table 3). This may be explained partly by the relatively higher sophistication of the p53 analysis methodology, being more robust and less subject to tissue artefacts, and also less heterogeneity in p53 expression being present within the triplicate cores. Scrutiny of outlier cases with discordant CD3 scoring suggested that exclusion of different cores from scoring by each reviewer, from the triplicate set available for each case, contributed to this discordance.

The reviewer with little experience of pathology images produced scores which diverged considerably from those of the other two reviewers. However, this had little impact on stratification of scores for that reviewer, as any error was consistently applied across all images, hence the lack of any significant detriment to comparative survival analyses. Nevertheless, while somewhat reassuring regarding the robustness of the digital analysis, it would be remiss to underestimate the importance of specialist knowledge and experience in the application of digital biomarker scoring. For greatest scoring accuracy, it is highly recommended that input from a suitably trained pathologist is sought, particularly for critical points such as cell classification and selection of suitable thresholds, for example relating to tissue detection and in calibration of mild, moderate and strong staining intensity. This input may alternatively be provided by a laboratory scientist sufficiently experienced in viewing IHC slides. Identifying artefacted cores and judging whether or not sufficient artefact is present to exclude such cores from the study also benefits from pathology experience. Additionally, an understanding of the basic principles of image processing and analysis is beneficial in discerning how the software may be used to attain the best results. Consequently, we would suggest that the preferred arrangement is for a pathologist and bioimage analyst to work together in collaboration to design and verify the analysis methodology employed.

These results indicate that image analysis is not a panacea in terms of standardising biomarker scoring, and results may differ even when the same software is applied to the analysis of the same data. In this study all three reviewers employed an agreed general approach, but this nevertheless afforded room for interpretation. This should serve as a warning against standardising cut-offs based on absolute values produced through image analysis, such as the density of CD3 positive cells or H-score for p53, because these values depend upon how the software is used. Consequently, the reporting of raw *QuPath* scores should ideally be accompanied by an indication of precise methodology applied, in addition to chosen thresholds, especially if these differ from default *QuPath* settings. This would be particularly problematic for researchers wishing to use ROC-defined cut-offs to determine high or low expression categories of a biomarker. To assist with this, key parameters are automatically logged by *QuPath* during analysis, and may be exported to create batch processing scripts. Nevertheless, it must be kept in mind that the final results will clearly depend not only on the analysis methodology and parameters used, but also on the images themselves – and hence on the laboratory protocols (in particular tissue section thickness, IHC procedures) and scanner involved in generating the digital whole slide images. Consequently, running precisely the same image analysis approach (using any software) on images or data generated in different laboratories does not guarantee true comparability, and careful quality control is required at all stages to ensure the results are valid and meaningful. The ability within *QuPath* to review instantly any changes to the image mark-up following changes to the analysis method is an invaluable tool, especially for pathologists for whom "seeing is believing" as far as image data verification is concerned. Of note, *QuPath* analysis in this study has been performed on images generated from two different automated IHC platforms (Leica and Ventana) and proved itself capable of handling any associated variation.

Immunoexpression data for p53 has been presented and analysed in this study in a continuous scale, consistent with the approach taken in previous studies. [10, 16, 17] This suits the focus of this study, which is comparability of digital image analysis scoring by *QuPath,* amongst different reviewers. However, given recent developments in the understanding of p53 biology, and the relationship between different p53 mutations and immunoexpression patterns, it is now considered more appropriate to analyse p53 staining by comparing normal or "wild type" staining with aberrant extremes of staining ("mutation type"). [18-20] The latter approach to p53 analysis has been applied to this colon cancer cohort and reported separately, finding that aberrant extremes of p53 immunoexpression

(diffuse intense or completely absent) was associated with significantly poorer unadjusted disease-specific survival when compared with "wild-type" expression (p=0.003). [4]

We have demonstrated the feasibility and inter-observer reproducibility of CD3 and p53 immunoscoring using *QuPath* in the TMA setting, even with limited experience of digital pathology images and minimal *QuPath* training. This may be extrapolated to equivalent tissue samples and to other markers showing the similar patterns of staining. *QuPath* has also been utilised in recently published studies to score biomarkers demonstrating cytoplasmic (cyclooxygenase-2, 3-hydroxy-3-methylglutaryl coenzyme-A reductase) and membranous (HER2) tumour cell staining. [4-6] However, biomarkers with more complex patterns of staining, such as mismatch repair proteins or immune checkpoint inhibitors, for example PD-L1, are much more challenging for digital scoring, as multiple cell populations are stained (tumour and inflammatory) and distinction between tumour and inflammatory cell staining can be difficult when the pattern of staining is patchy. Recently, convolutional neural networks have shown great promise for pathology image analysis, and currently represent the state-of-the-art whenever such a rigorous cell classification is required. [21] However, to date the application of such 'deep learning' analysis has mostly focussed on haematoxylin and eosin staining. [22, 23]

In summary, we demonstrate the functionality, even amongst inexperienced users, and inter-observer reproducibility, of *QuPath* biomarker scoring in the setting of immunohistochemically stained tumour TMAs. This represents just one facet of this powerful, new, open source digital image analysis software, which may help to alleviate a critical bottleneck in tissue biomarker discovery research.

References

1 Ilyas M, Grabsch H, Ellis IO, et al. Guidelines and considerations for conducting experiments using tissue microarrays. *Histopathology* 2013;**62**:827-39.

2 Hamilton PW, Bankhead P, Wang Y, et al. Digital pathology and image analysis in tissue biomarker research. *Methods* 2014;**70**:59-73.

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

4 Bankhead P, Fernandez JA, McArt DG, et al. Integrated tumor identification and automated scoring minimizes pathologist involvement and provides new insights to key biomarkers in breast cancer. *Lab Invest* 2018;**98**:15-26.

5 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.

6 Gray RT, Loughrey MB, Bankhead P, et al. Statin use, candidate mevalonate pathway biomarkers, and colon cancer survival in a population-based cohort study. *Br J Cancer* 2017;**116**:1652-9.

7 Roxburgh CS, McMillan DC. The role of the in situ local inflammatory response in predicting recurrence and survival in patients with primary operable colorectal cancer. *Cancer Treat Rev* 2012;**38**:451-66.

8 Vayrynen JP, Tuomisto A, Klintrup K, et al. Detailed analysis of inflammatory cell infiltration in colorectal cancer. *Br J Cancer* 2013;**109**:1839-47.

9 Theodoropoulos GE, Karafoka E, Papailiou JG, et al. P53 and EGFR expression in colorectal cancer: a reappraisal of 'old' tissue markers in patients with long follow-up. *Anticancer Res* 2009;**29**:785-91.

10 Munro AJ, Lain S, Lane DP. P53 abnormalities and outcomes in colorectal cancer: a systematic review. *Br J Cancer* 2005;**92**:434-44.

11 McCarty KS,Jr, Szabo E, Flowers JL, et al. Use of a monoclonal anti-estrogen receptor antibody in the immunohistochemical evaluation of human tumors. *Cancer Res* 1986;**46**:4244s-8s.

12 de Gramont A, Watson S, Ellis LM, et al. Pragmatic issues in biomarker evaluation for targeted therapies in cancer. *Nat Rev Clin Oncol* 2015;**12**:197-212.

13 Schmidt C. How do you tell whether a breast cancer is HER2 positive? Ongoing studies keep debate in high gear. *J Natl Cancer Inst* 2011;**103**:87-9.

14 Deroulers C, Ameisen D, Badoual M, et al. Analyzing huge pathology images with open source software. *Diagn Pathol* 2013;**8**:92.

15 Carpenter AE, Kamentsky L, Eliceiri KW. A call for bioimaging software usability. *Nat Methods* 2012;**9**:666-70.

16 Kruschewski M, Mueller K, Lipka S, et al. The Prognostic Impact of p53 Expression on Sporadic Colorectal Cancer Is Dependent on p21 Status. *Cancers (Basel)* 2011;**3**:1274-84.

17 Melincovici CS, Mihu CM, Marginean M, et al. The prognostic significance of p53, Bax, Bcl-2 and cyclin E protein overexpression in colon cancer - an immunohistochemical study using the tissue microarray technique. *Rom J Morphol Embryol* 2016;**57**:81-9.

18 Boyle DP, McArt DG, Irwin G, et al. The prognostic significance of the aberrant extremes of p53 immunophenotypes in breast cancer. *Histopathology* 2014;**65**:340-52.

19 Kaye PV, Haider SA, James PD, et al. Novel staining pattern of p53 in Barrett's dysplasia--the absent pattern. *Histopathology* 2010;**57**:933-5.

20 McCluggage WG, Soslow RA, Gilks CB. Patterns of p53 immunoreactivity in endometrial carcinomas: 'all or nothing' staining is of importance. *Histopathology* 2011;**59**:786-8.

21 Sirinukunwattana K, Ahmed Raza SE, Yee-Wah T, et al. Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images. *IEEE Trans Med Imaging* 2016;**35**:1196-206.

22 Djuric,Ugljesa, Zadeh,Gelareh, Aldape,Kenneth, et al. Precision histology: how deep learning is poised to revitalize histomorphology for personalized cancer care. *npj Precision Oncology* 2017;**1**, Article number: 22.

23 Litjens G, Sanchez CI, Timofeeva N, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Sci Rep* 2016;**6**:26286.

Table 1. Median biomarker expression scores for CD3 (positive cells/mm^2) and p53 (H-score) for three independent reviewers.

Biomarker and Reviewer	Number of cases*	Median expression score**	Interquartile range	p-value***
CD3				
Reviewer 1	Controls n=440	567	303-984	
	CRC deaths n=208	450	235-835	0.02
Reviewer 2	Controls n=440	862	468-1437	
	CRC deaths n=208	670	333-1249	0.001
Reviewer 3	Controls n=440	554	308-933	
	CRC deaths n=208	429	231-750	0.002
p53				
Reviewer 1	Controls n=447	89	27-242	
	CRC deaths n=210	136	35-256	0.09
Reviewer 2	Controls n=447	19	3-98	
	CRC deaths n=210	39	3-117	0.12
Reviewer 3	Controls n=447	102	38-249	
	CRC deaths n=210	161	44-266	0.04

*Only cases were included where scores were available from all three reviewers.
**Positive cells/mm^2 of tissue for CD3 and H-score for p53.
***Comparing differences in median expression scores between CRC deaths and controls.
CRC, colorectal cancer.

Table 2. Unadjusted Hazard ratios and 95% confidence intervals comparing survival of patients with colon cancer according to CD3 and p53 biomarker expression, as scored by three independent reviewers.

Biomarker and Reviewer	Expression score	Number of controls/ CRC deaths	Unadjusted hazard ratio and 95% confidence intervals	p-value for trend
CD3				
Reviewer 1	Low (<567)	220/123	1.00	
	High (≥567)	220/85	0.72 (0.55-0.96)	0.02
Reviewer 2	Low (<862)	220/130	1.00	
	High (≥862)	220/78	0.64 (0.49-0.85)	0.002
Reviewer 3	Low (<554)	220/127	1.00	
	High (≥554)	220/81	0.68 (0.51-0.90)	0.006
p53				
Reviewer 1	Low (<89)	223/87	1.00	
	High (≥89)	224/123	1.32 (1.00-1.73)	0.05
Reviewer 2	Low (<19)	224/88	1.00	
	High (≥19)	223/122	1.29 (0.98-1.70)	0.07
Reviewer 3	Low (<102)	224/86	1.00	
	High (≥102)	223/124	1.33 (1.01-1.75)	0.04

CRC, colorectal cancer.

Table 3. Kappa values reflecting inter-observer variation in scoring dichomotous categories of median immunohistochemical biomarker expression for CD3 and p53.

	Reviewer	1	2	**Inter-observer agreement**
CD3 (High v. low)	2	0.82		Excellent
	3	0.85	0.86	
p53 (High v. low)	2	0.93		Excellent
	3	0.93	0.91	

Table 4. Unadjusted Hazard ratios and 95% confidence intervals comparing manual and *QuPath* scoring of CD3 immunoexpression, as scored by one reviewer.

Scoring method	**Expression score**	**Number of controls/ CRC deaths**	**Unadjusted hazard ratio and 95% confidence intervals**	**p-value for trend**
Manual	1	154/81	1.00	
	2	221/110	0.92 (0.69-1.23)	
	3	71/21	0.60 (0.37-0.97)	0.06
QuPath	Low (<971)	153/98	1.00	
	Medium (971-<2775)	222/92	0.71 (0.53-0.95)	
	High (≥2775)	71/21	0.54 (0.34-0.87)	0.003

Acknowledgements

The samples used in this research were received from the Northern Ireland Biobank, which has received funds from Health and Social Care Research and Development Division of the Public Health Agency in Northern Ireland, Cancer Research UK and the Friends of the Cancer Centre. The Northern Ireland Molecular Pathology Laboratory, which was responsible for construction of tissue microarrays, slide staining and scanning, has received funding from Cancer Research UK, the Experimental Cancer Medicine Centre Network, the Health and Social Care Research and Development Division of the Public Health Agency in Northern Ireland, the Sean Crummey Memorial Fund, the Tom Simms Memorial Fund and the Friends of the Cancer Centre. The research leading to these results has also received funding from Invest Northern Ireland.

We thank Mr. Ken Arthur for the construction of the original tissue microarrays used in this study, Ms. Victoria Bingham for her work in staining and scanning the slides, and Dr. Roisin O'Neill and the Northern Ireland Cancer Registry for their contributions to the clinical data collation.

Legends

Figure 1. Key steps in *QuPath* methods for scoring CD3 (A) and p53 (B).
Figure 2. Kaplan Meier curves for CD3 expression and colorectal cancer-specific survival.
Figure 3. Kaplan Meier curves for p53 expression and colorectal cancer-specific survival.