FULL PAPER

BJC

British Journal of Cancer (2017) 116, 237–245 | doi: 10.1038/bjc.2016.404

BJC OPEN

Keywords: cancer; immunohistochemistry; tissue microarray; crowdsourcing; biomarker; pathology

Crowdsourcing for translational research: analysis of biomarker expression using cancer microarrays

Jonathan Lawson[1,6], Rupesh J Robinson-Vyas[1,6], Janette P McQuillan[1,6], Andy Paterson[1], Sarah Christie[1], Matthew Kidza-Griffiths[1], Leigh-Anne McDuffus[2], Karwan A Moutasim[3], Emily C Shaw[1,3], Anne E Kiltie[4], William J Howat[2], Andrew M Hanby[5], Gareth J Thomas[3] and Peter Smittenaar*[,1]

[1]Cancer Research UK, 407 St John Street, London EC1V 4AD, UK; [2]Cancer Research UK Cambridge Institute, University of Cambridge, Robinson Way, Cambridge CB2 0RE, UK; [3]Southampton CRUK Centre, University of Southampton Faculty of Medicine, Tremona Road, Southampton SO16 6YD, UK; [4]CRUK/MRC Oxford Institute for Radiation Oncology, University of Oxford, Roosevelt Drive, Oxford OX3 7DQ, UK and [5]Cancer Research UK Leeds Centre, University of Leeds, Beckett Street, Leeds LS9 7TF, UK

Background: Academic pathology suffers from an acute and growing lack of workforce resource. This especially impacts on translational elements of clinical trials, which can require detailed analysis of thousands of tissue samples. We tested whether crowdsourcing – enlisting help from the public – is a sufficiently accurate method to score such samples.

Methods: We developed a novel online interface to train and test lay participants on cancer detection and immunohistochemistry scoring in tissue microarrays. Lay participants initially performed cancer detection on lung cancer images stained for CD8, and we measured how extending a basic tutorial by annotated example images and feedback-based training affected cancer detection accuracy. We then applied this tutorial to additional cancer types and immunohistochemistry markers – bladder/ki67, lung/EGFR, and oesophageal/CD8 – to establish accuracy compared with experts. Using this optimised tutorial, we then tested lay participants' accuracy on immunohistochemistry scoring of lung/EGFR and bladder/p53 samples.

Results: We observed that for cancer detection, annotated example images and feedback-based training both improved accuracy compared with a basic tutorial only. Using this optimised tutorial, we demonstrate highly accurate (>0.90 area under curve) detection of cancer in samples stained with nuclear, cytoplasmic and membrane cell markers. We also observed high Spearman correlations between lay participants and experts for immunohistochemistry scoring (0.91 (0.78, 0.96) and 0.97 (0.91, 0.99) for lung/EGFR and bladder/p53 samples, respectively).

Conclusions: These results establish crowdsourcing as a promising method to screen large data sets for biomarkers in cancer pathology research across a range of cancers and immunohistochemical stains.

Personalised medicine is reliant on the determination of markers and genetic profiles that facilitate targeting of therapies to those who will benefit the most. Achieving this aim depends on translational studies from clinical trials whereby success of the new agent, modality or regime is correlated with profiles observed in the target tissues. By their nature these studies generate large tissue sets. Progress therefore depends on pathologists having sufficient time for research, which is becoming increasingly

*Correspondence: Dr P Smittenaar; E-mail: petersmittenaar@gmail.com
[6]These authors contributed equally to this work.

Received 3 September 2016; revised 7 November 2016; accepted 9 November 2016; published online 13 December 2016

difficult in an environment of increasing workload and severe financial constraints on healthcare and research across the world. In this context, the future of medical research is critically dependent upon innovation to improve productivity and increase efficiency (UK Accelerated Access Review). We hypothesised that contributions from the general public – also known as 'crowdsourcing' – can have a role in accelerating biomedical research. Here we explore its application in the field of immunohistochemistry (IHC) scoring in human cancer tissue samples.

Histopathologists have a key role in both medical diagnostics and translational research. While demand for histopathologists has never been higher, in most part due to increases in cancer cases (+30% in the UK since the late 1970s; Cancer Research UK, 'Cancer incidence for all cancers combined', 2013), there has been a precipitous decline in the academic histopathology workforce. In the US, the proportion of pathologists in the population is predicted to drop by 35% between 2010 and 2030 (Robboy *et al*, 2013), whereas the UK has seen a 60% drop in academic pathologists between 2000 and 2012 (Wilkins, 2015). Many of the solutions proposed to address this deficit can only be realised in the long-term, whereas more resource is required immediately to ensure an ongoing contribution of tissue sample interrogation to translational research. Machine learning promises to automate many routine evaluations (Bolton *et al*, 2010; Wilbur, 2014; Bouzin *et al*, 2015; Howat *et al*, 2015), but commonly requires large, validated data sets for its development. Crowdsourcing can provide such data sets in addition to solving an immediate need for analytical resource.

Crowdsourcing (or citizen science) is the provision of services by distributed members of the general public. Such services take many forms, including problem solving, nature surveys, environmental monitoring, and data processing (Ranard *et al*, 2014). Crowdsourcing has existed for close to two centuries but experienced a surge in popularity over the past decade, particularly facilitated by internet and mobile technologies. Current scientific applications include the classification of images of distant galaxies (Lintott *et al*, 2008), puzzle games designed to create a three-dimensional visual representation of the brain (Seung and Burnes, 2012), discovering tertiary structures of proteins (Cooper *et al*, 2010), as well as bug hunting and genome sequence analysis (Kawrykow *et al*, 2012; Good and Su, 2013; Rallapalli *et al*, 2015).

Here we crowdsourced the analysis of tumour samples prepared as tissue microarrays (TMAs). Tissue microarrays facilitate high-throughput molecular analysis of tissue samples to investigate associations between tumour-specific protein expression and clinical outcomes (Giltnane and Rimm, 2004). Although automated analysis of TMAs has proven to be effective for specific screening protocols, particularly in breast cancer (Turbin *et al*, 2008; Bolton *et al*, 2010; Konsti *et al*, 2011; Howat *et al*, 2015), it was also observed that algorithms underperform on less well-established markers such as cytokeratin (CK) 5/6 and epidermal growth factor receptor 1 (EGFR/HER1; Howat *et al*, 2015). In the same study, 20–25% of samples had to be manually excluded from the analysis. This suggests a synergy between crowdsourcing and automated analysis, whereby manual exclusion and scoring could precede the training of an automated algorithm. A key feature in this approach is that crowdsourcing can compensate for slight deficits in accuracy through the sheer volume of data it can process.

We previously developed Cell Slider (www.cellslider.net) to invite members of the public to score breast cancer TMA cores for oestrogen receptor (ER) staining (Candido Dos Reis *et al*, 2015). We observed that users tended to overestimate the number of cancer cells in an image, compromising the accuracy of IHC scores. This lack of specificity in Cell Slider was most likely due to a minimal level of instruction provided prior to scoring the samples, as well as a restrictive interface showing only a small portion of a TMA sample, preventing users access to an overview of the tissue. Here we present a novel crowdsourcing interface developed to improve upon Cell Slider. First we set out to test the effects of feedback-based training and provision of annotated example images on the ability of scorers to detect cancer in a sample. We then used this improved tutorial to assess performance in cancer detection in four sample types selected as being of interest to academic pathologists. Finally, we examined the accuracy of IHC scoring in a lung cancer sample with membrane expression of EGFR, and bladder cancer with nuclear expression of p53.

MATERIALS AND METHODS

Participant recruitment and ethics. Participants were recruited through e-mails to individuals registered for non-pathology Cancer Research UK crowdsourcing projects. Newsletters and advertising were used to recruit new volunteers specifically for Trailblazer, and additional paid testers were recruited via the Prolific Academic platform (http://www.prolific.ac/, £7.50 per hour). We combined results from volunteers and paid participants as although paid testers are considerably faster than volunteers, the performance of the groups is not significantly different (data not shown). All participants provided informed consent to participation and storage of their data. The Health Research Authority approved this study (14/NW/1033). All participants that completed the test samples were included in the results reported here. None of the participants expressed any professional experience with pathology, but otherwise no demographic or data on educational achievement was collected. No participant participated more than once in any of the experiments.

Tissue microarray samples. Samples from oesophageal and lung tissues were prepared as TMAs, immunohistochemically stained and imaged by the research groups of GJT and WH as previously described (Ward *et al*, 2014). The AK lab prepared the bladder cancer samples with p53 IHC as described previously (Cazier *et al*, 2014), and the bladder samples were stained with Ki67 using a clone MIB-1 (Dako, Agilent Technologies) at 1 in 1000 dilution on a Leica Bond machine, with Epitope Retrieval 1 buffer for 20 min. For all samples, patients consented to the use of their tissue for research (bladder cancer samples ethical approval 13/LO/0540; lung and oesophageal cancer samples ethical approval REC no. 10/H0504/32)

Expert scoring. We obtained expert scores for the cancer detection task from 3 experts for all samples but lung/CD8, for which we had two additional experts. Three experts scored the IHC lung/EGFR sample and three experts scored the IHC bladder/p53 sample. The expert scores were provided through the same web interface used by the participants, except for the bladder samples which were scored as digital images in the lab. Pathologists entered their ratings independently from one another. Final expert consensus values, used to rate non-specialist participants were calculated as the majority vote (for cancer detection tasks) or the median value across experts (for IHC scoring of biomarker proportion and intensity).

Online platform. All Trailblazer releases were developed using Pybossa, an open-source framework specifically developed for online crowdsourcing (https://github.com/PyBossa/). The stack consisted of Python, Django, Postgres, Javascript and jQuery. The platform was hosted on Amazon Web Services. Our code – available under a GNU Affero license – can be found at https://citizenscience.github.io.

Detecting cancer cells. Participants were presented with a sequence of images and asked to identify regions where cancer was present. Ten images were overlaid by a 6×6 grid for a total of

360 squares (Figure 1). Participants then marked each square as containing no cancer (green), one or more cancer cells (red) or no tissue (blank; Figure 1B). A scrollable gallery of reference images illustrating a variety of cancer and non-cancer cells were included to aid correct analysis. The same ten images of lung cancer stained by CD8 (lung/CD8) were used throughout the testing of different tutorial mechanics. The ten images for each experiment were confirmed by consultant pathologists to be representative of the variety of possible tissue morphologies and biomarker staining patterns. The images were presented to each participant in a random order. We used a full factorial design (Figure 2A) to assess the effect of annotated images and feedback-based training in tutorials. The basic tutorial consisted of an ~10- to 15-min, passive, text- and image-based set of instructions, developed based on interviews and training sessions with pathologists. Whilst all participants viewed the basic tutorial, they were randomly assigned to one of 4 groups in the factorial design. The tests investigated two additional tutorial elements. Firstly, the addition of 5 annotated images, shown to participants during the tutorial. Secondly, feedback-based training presented with 5 training images before the test images. For two images they were provided immediate feedback on each answer. For the remaining 3 images feedback was provided only after scoring the majority of the image. This was designed to mimic the learning experience of other successful crowdsourcing experiments (e.g., in EyeWire; Kim *et al*, 2014). The same five example images were used for both annotated images and feedback, and no images from the tutorial were used for testing. In addition to lung/CD8 a further three data sets were tested to confirm the accuracy of the tutorial including annotated images and feedback-based training.

IHC biomarker scoring. Cancer detection is only the first step in TMA scoring; the next step is to score the percentage of cancer cells that are stained and the intensity of such staining. We therefore set out to test how accurately participants could score cancer staining, given the improved tutorial for cancer detection. We selected 21 lung/EGFR cytoplasmic stain samples and 30 bladder/p53 nuclear stain samples representative of the majority of clinical samples to test this, whereby each participant scored a random set of 10 images. These images were separate to the images used initially for cancer detection; no images from the tutorial were used as a test image. Participants indicated proportion of staining as a percentage, in increments of 5%. Where proportion was above 0%, i.e. stained cancer cells were present, the participant was asked to score staining intensity as 1 (weak), 2 (moderate), or 3 (strong). The product of these two, that is, a score between 0 and 300, is called the McCarty 'H' score and is commonly used to relate IHC to patient outcomes and treatment response (McCarty *et al*, 1986). The tutorial for cancer detection was extended to explain IHC scoring, and users practiced IHC scoring through feedback-based training prior to scoring the TMA scores on which their performance was assessed. This extended tutorial, consisting of both a cancer detection and IHC scoring tutorial, took between 20 and 30 min to complete.

Statistical analysis. Analyses presented in this paper are either at the level of individual participants or at the level of consensus

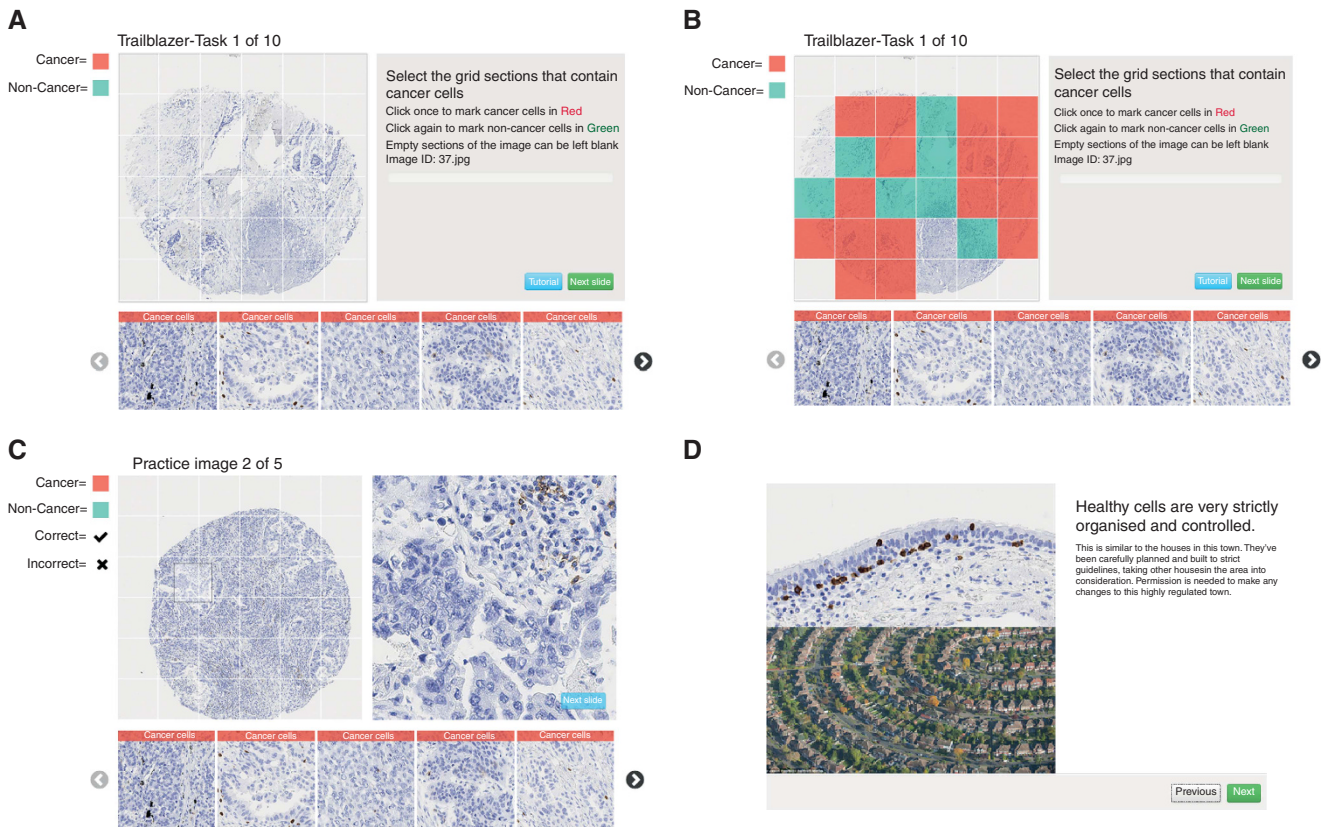

Figure 1. The 'Trailblazer' interface for viewing, annotating and scoring tissue microarray (TMA) cores. (**A**) Participants evaluated squares on a 6x6 grid overlaid on a TMA for the presence of cancer cells. (**B**) They were asked to mark squares with cancer as red, cells without cancer as green and completely empty squares as blank. (**C**) To aid in cancer detection and IHC scoring, the participant could move their cursor over the core to reveal a high magnification view of the area under the cursor. Furthermore, a scrollable gallery of high magnification example images of cancer tissue and healthy tissue was available at the bottom of the screen. (**D**) Prior to starting the task each participant completed a ~10-minute tutorial explaining the type of sample and how to distinguish cancer cells from non-cancer cells, of which a screenshot is shown here. In the first experiment we tested the effect of feedback-based training and/or annotated images provided in addition to this baseline tutorial.

ratings based on the aggregation of multiple participants. Whereas the former informs us about the effect of tutorial changes on individual performance, aggregated data underlies the power of crowdsourcing and is therefore the metric of interest when assessing the usefulness of this approach. During tutorial development for cancer detection, each participant provided 360 ratings (36 grid squares in 10 images). In the analysis we equated 'blank' and 'no cancer' responses such that each rating was binomial (positive or negative for cancer). We furthermore excluded 53 squares which contained no tissue whatsoever, as these would artificially boost the specificity. Each participant rating was then compared with the expert consensus on the basis of the presence or absence of cancer cells in each square. These comparisons were used to identify true positive (TP), true negative (TN), false positive (FP) and false negative (FN) responses from which sensitivity ($TP/[TP+FN]$), specificity ($TN/[TN+FP]$) and F1-score ($2\times TP/[2\times TP+FP+FN]$) were calculated (Figure 2). The general linear model was used to obtain coefficients and *P*-values on the main effects of feedback-based training and annotated images, and on their interaction. We computed Cohen's Kappa for each participant against the expert consensus, between pairs of experts, and for the participant consensus against the expert consensus.

One pertinent question in crowdsourcing is how many participants are required to provide accurate analyses for each image, with the underlying assumption of diminishing returns in group performance as more participants are added. We explored this question for both cancer detection and IHC scoring by bootstrapping various group sizes. We used the AUC described by the receiver operating characteristic – a common classification measure for a binary classifier – to assess group performance. Bootstrapping was used to estimate the accuracy of hypothetical groups between 3 and 40 participants in size. For a group size *n*, we sampled *n* participants from the complete population of participants with replacement, 500 times. Similarly, IHC scoring accuracy was assessed on the basis of Spearman *r* between the median expert score and bootstrapped groups of participants. For each image, we took the median of all responses for that image to calculate the aggregate H-score. IHC bootstrapping was performed using 10 000 samples.

All analyses were performed in Python using SciPy (Jones *et al*, 2001), scikit-learn (Pedregosa *et al*, 2011), scikits-bootstrap (https://github.com/cgevans/scikits-bootstrap), Pandas (McKinney, 2010) and NumPy (van der Walt *et al*, 2011). Graphs were created using Matplotlib (Hunter, 2007).

RESULTS

Identification of cancer cells. In our first experiment we tested the efficacy of two tutorial elements such that participants could better distinguish cancer from non-cancer tissue. In the basic tutorial without annotated images or feedback, *individual* participants (as opposed to the aggregate of multiple responses which is more commonly used in crowdsourcing) achieved an average sensitivity of 0.74 ± 0.04 (95% CI of the mean), specificity of 0.66 ± 0.04, and F1-score of 0.70 ± 0.03 (Figure 2B). We calculated main effects and interactions for the two factors using linear regression (see Table 1 for statistics). We found both annotated images and feedback-based training had statistically significant positive effects on the F1-score, with no interaction between the factors. In our experiment, adding both factors improved the F1-score by $\sim$0.05. Both tutorial components were therefore used in follow-up experiments. It is worth noting that the sensitivity-specificity trade-off was shifted strongly in favour of sensitivity in response to feedback-based training, whereas annotated images had no such effect (Table 1). In other words, feedback-based training lowers the threshold to indicate a square contains cancer.

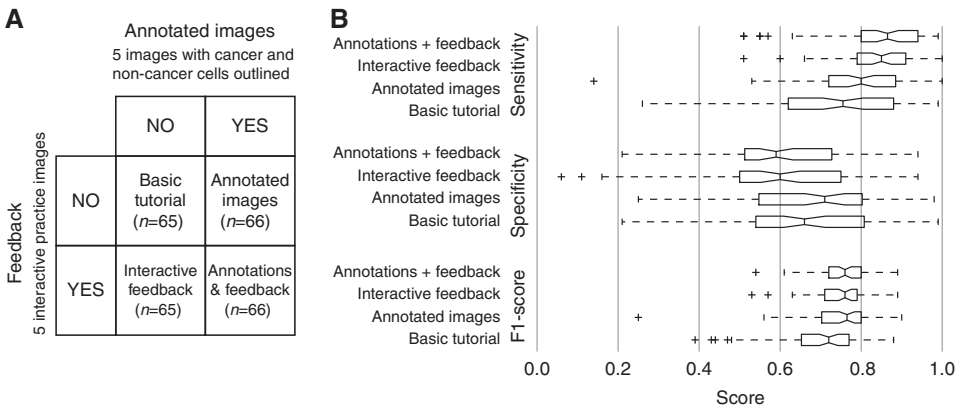

Figure 2. Full factorial design to test the effect of annotated images and feedback-based training on cancer detection performance of individual participants. (**A**) Experimental design and number of participants in each cell. (**B**) Box-plot graph showing performance in cancer detection across individuals in each of the four groups, expressed as F1-score, specificity and sensitivity. Statistics for main effects and interactions are shown in Table 1.

Table 1. Main effects of annotated images and feedback-based training and their interaction

Factor	F1-score	Specificity	Sensitivity
Annotated images	**β = 2.11, (0.03, 4.20) *P* = 0.047**	β = 1.18 (−3.11, 5.48) *P* = 0.59	β = 2.69 (−0.64, 6.02) *P* = 0.11
Feedback-based training	**β = 3.14 (1.06, 5.23) *P* = 0.003**	**β = −7.59 (−11.89, −3.30) *P* = 0.001**	**β = 8.77 (5.44, 12.10) *P* < 0.001**
Interaction	β = −1.81 (−3.89, 0.27) *P* = 0.09	β = −0.71 (−5.00, 3.58) *P* = 0.75	β = −1.95 (−5.28, 1.38) *P* = 0.25

All regression coefficients represent estimated change in performance when adding the factor, multiplied by 100. For example, adding annotated images is estimated to improve the F1-score by 0.0211. Values in brackets represent 95% confidence interval of the coefficient. Cells in bold are significant at $P<0.05$ uncorrected for multiple comparisons.

Cancer detection in different cancers and biomarkers. We used the improved tutorial to test three additional data sets: a further set of lung samples stained for EGFR ($N = 76$ participants), oesophageal samples stained for CD8 ($N = 49$ participants), and bladder samples stained for Ki67 ($N = 49$ participants). Critically, we now looked at both individual and aggregate performance, the latter by combining multiple cancer/no cancer responses for each individual square in the image. We first calculated Cohen's kappa for each participant with the expert consensus, revealing large differences between participants (Figure 3A). We then aggregated participants by calculating a majority consensus score for each square, which yielded 'moderate' to 'substantial' agreement (Landis and Koch, 1977) in each of the 4 sample types (Figure 3A). We also calculated the pairwise agreement between each of the experts and the average of those pairwise agreements. Strikingly, in 3 out of 4 sample types the majority consensus of participants was in better agreement with the expert consensus than experts among one another (Figure 3A).

A second way of quantifying performance of the aggregate group is to use the area under the receiver operating characteristic curve (AUC). Specifically, we were interested in the relationship between the number of participants evaluating a sample and the accuracy as measured by AUC. For each of the 4 sample types, we bootstrapped 500 samples for a number of participant population sizes between 3 and 40. In all cases the average AUC approached a maximum of ~ 0.95 asymptotically as the number of participants per sample increased (Figure 3B). Altogether, in the majority of samples, a relatively small group of lay participants was able to approach levels of accuracy that would be expected from any one trained expert relative to another.

Immunohistochemistry scoring. Having established tutorial elements that improve participant performance in the detection of cancer in TMAs and demonstrated that these permit high levels of agreement with experts in several different sample types, a key question remained: would the new interface yield reliable scoring of immunohistochemical staining in TMA samples? To answer this question, we tested IHC accuracy in the membrane/cytoplasmic marker EGFR in lung cancer ($N = 35$ participants, each scoring 10 images) and for the nuclear marker p53 in bladder cancer ($N = 45$ participants, each scoring 10 images). In the lung/EGFR data we observed a Spearman correlation of 0.91 (bootstrapped 95% CI = (0.78, 0.96)) between the median participant response and median expert score (Figure 4A). In the bladder/p53 sample, this same correlation was 0.97 (95% CI = (0.91, 0.99); Figure 4B). We also calculated how accuracy improved as we increased the number of participants evaluating each image (Figure 5). As was the case in cancer detection, having more than 5–10 participants rate each image did not yield substantial increases in group performance.

DISCUSSION

In this paper we addressed the hypothesis that crowdsourcing – distributing work to members of the general public – can be used to accurately analyse cancer TMA samples, using an online platform specifically developed for the clear presentation of samples. We initially examined the ability to distinguish cancer tissue from non-cancer tissue, a critical first step in IHC analysis, and found that annotated images and feedback-based training positively impacted on performance in lung/CD8 samples. We then applied this training method to three more sample types – lung/EGFR, oesophageal/CD8, and bladder/Ki67 – finding that aggregated responses from participants showed agreement with experts at a similar level as experts with one another, with AUCs between 0.90 and 0.95. Finally, we tested our improved tutorial for its usefulness in IHC scoring itself, finding strong correlations based on H-score between crowdsourced scores and experts. Altogether, these results provide evidence that the public can accurately analyse TMA samples, and suggest crowdsourcing as a potential additional resource to meet the growing demand for analysis resource in pathology research.

Our previous work in the analysis of breast cancer samples stained for oestrogen receptor showed an AUC of 0.95 for cancer detection at the whole core level, as well as strong correlations for IHC scoring with expert ratings (Cell Slider; Candido Dos Reis *et al*, 2015). However, this proof of principle was performed in the most common cancer and marker available, which can be analysed accurately using automated methods (e.g., Turbin *et al*, 2008; Bouzin *et al*, 2015; Howat *et al*, 2015). Here, we tested analytically challenging cancer types as well as immunohistochemical stains for which algorithms are either scarce or require considerable involvement from experts. By testing the crowdsourcing approach across a breadth of samples, we have shown this method to be flexible and widely applicable, including in sample types where algorithms struggle (Howat *et al*, 2015). Although both sample types we used for IHC scoring achieved high correlations with experts, the higher level of accuracy for bladder/p53 samples compared with lung/EGFR is most likely caused by the fact that the former is a nuclear marker whereas the latter is membranous. To our knowledge crowdsourcing has only seen limited investigation in cancer research. One study in renal cell carcinoma compared pathologists, research fellows, members of the public, and a fully automated algorithm on nucleus detection and segmentation (Irshad *et al*, 2014). They observed that members of the public performed similarly to research fellows, and either similarly to or better than the algorithm depending on the task.

Algorithms trained on large amounts of labelled data perform extremely well in many computer vision challenges (e.g., ImageNet; Russakovsky *et al*, 2015) including in cancer pathology (e.g., Walton *et al*, 2009; Beck *et al*, 2011). However, with over 200 cancer types and dozens of available immunohistochemical markers labelling different cellular components (nucleus, cytoplasm, and cell membrane) separately, obtaining sufficient training data for even a proportion of sample types is a considerable challenge. Crowdsourcing can provide a solution by scoring large data sets of samples for which no algorithms are available, and by subsequently making these data publicly available for researchers and commercial entities to develop automated methods. It is common practice for algorithms to supersede manual analysis in this way, as exemplified by the development of galaxy classifiers based on Galaxy Zoo data (Banerji *et al*, 2010), automated rather than crowdsourced analysis of electron microscopy data (Lee *et al*, 2015), and across the field of genomics. Our findings suggest such successes may be achieved on a large scale in pathology, where crowdsourcing can accelerate research by processing large volumes of samples currently being collected in clinical trials, as well as the vast amounts of tissue stored from past trials and routine archival material where patient consent is in place. Although crowdsourcing is not necessarily more resource-efficient than expert scoring – as it still requires ~ 10 lay people to score each image to achieve accurate results – the sheer size of the general public and therefore the number of people that could potentially contribute to analysis provides a unique opportunity to accelerate research.

We set out to test two tutorial elements that might improve performance on the cancer detection task, and observed both annotated images and feedback-based training boosted overall accuracy. It has previously been observed that crowdsourcing can be improved by various means, including self-censoring of submissions when a user is uncertain of a response (Shah and Zhou, 2015), using videos rather than only text- or image-based instruction (Starr *et al*, 2014), having mini-breaks especially for complicated tasks (Rzeszotarski *et al*, 2013), presenting context-sensitive help (Andersen *et al*, 2012), and financial punishment for disagreement with other users (Shaw *et al*, 2011). Most research in

crowdsourcing accuracy has been on paid workers, for example recruited through Amazon Turk. In the case of unpaid citizen science, however, users participate for non-financial reasons, primarily a desire to contribute to research (Raddick *et al*, 2013; Wright *et al*, 2015; Land-Zandstra *et al*, 2016) and to learn about science (e.g., Rotman *et al*, 2012). In such cases, offering financial

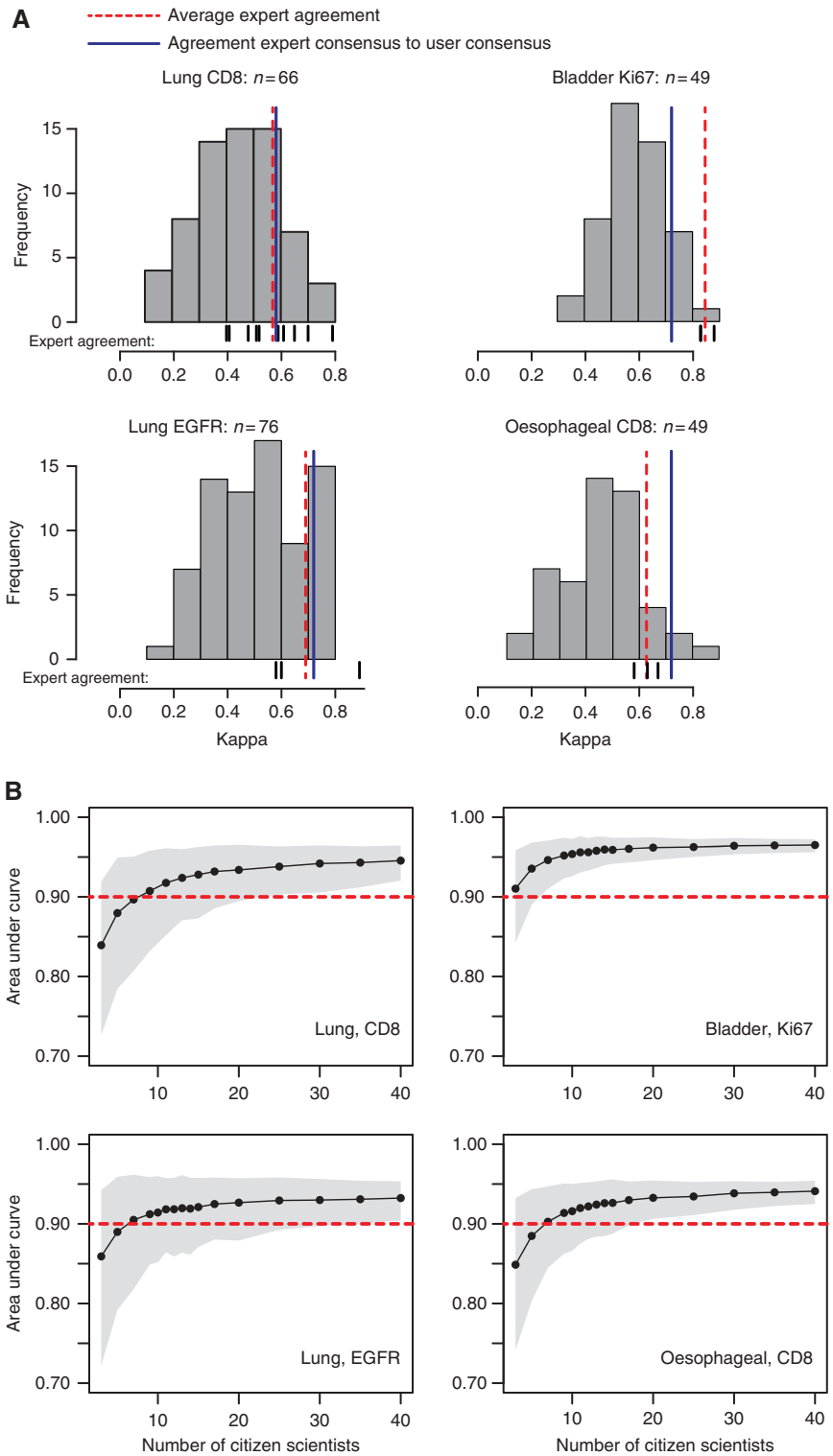

Figure 3. Accuracy of aggregated responses across four sample types. (**A**) We used Cohen's kappa to calculate correspondence between raters. The histogram indicates the distribution of kappas of each individual participant with the expert consensus. The solid blue line indicates the agreement between the majority consensus of all participants compared with the expert consensus, showing the majority outperforms the average individual. The pairwise kappas between experts are indicated as small black lines underneath the histogram; the average of the pairwise kappas is indicated in the dashed red line. (**B**) A second method to compare the participant consensus with expert consensus is the area under the receiver operating characteristic curve (AUC). Here we examined how the AUC changed as we varied the number of participants included in the consensus between 3 and 40. The red dotted line indicates an AUC of 0.90. Shaded areas indicate the bootstrapped 95th percentile CI.

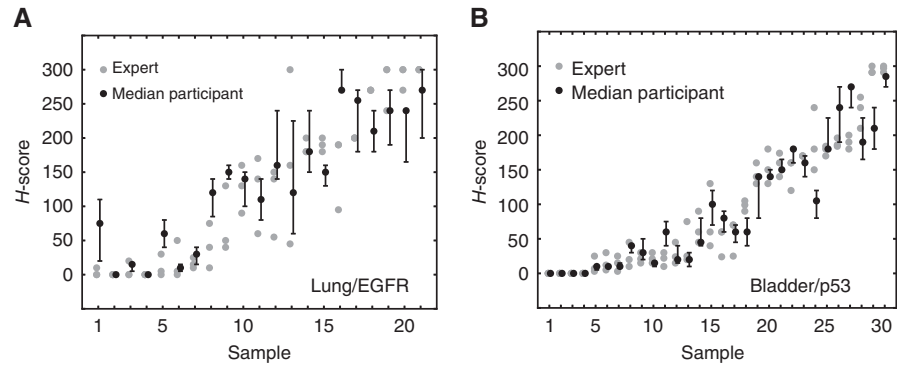

Figure 4. Comparison of expert and aggregated participant H-scores for each image. (**A**) Lung/EGFR sample. Grey dots indicate the three individual expert scores per sample, black dots indicate median H-score based on all participants who evaluated the image, error bars indicate the bootstrapped 95th percentile confidence interval of the median. The images have been sorted along the *x* axis by median expert score. (**B**) Bladder/p53 sample. For details described under (**A**).

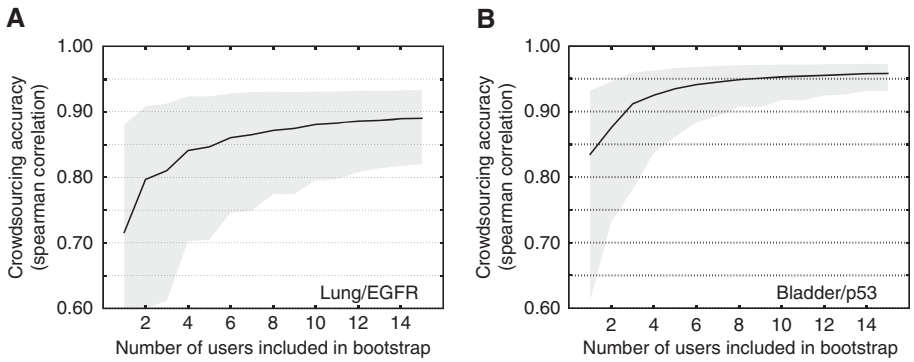

Figure 5. *H*-score accuracy as a function of number of participants evaluating each image. (**A**) In lung/EGFR we observed that the Spearman correlation between participants and experts strongly increased as we included more participants in the aggregate score. The black line represents the median of the bootstrapped samples, and the shaded area represents the bootstrapped 95th percentile confidence interval of the median. (**B**) Bladder/p53, legend as in subplot a.

incentives to improve accuracy would seem undesirable. Others have focused on improving the user experience to coax users to dedicate more time to the project, as experienced users are on average more productive than new users (Sauermann and Franzoni, 2015). All such tools, including our findings on tutorial optimisation, may be combined to establish crowdsourcing as an accurate tool for data analysis.

From this series of experiments, we conclude that crowdsourcing is an accurate and reliable analysis tool in TMA scoring – a major bottleneck in current clinical cancer research. We hope these results will encourage others in not only histopathology but cancer research more broadly, to take up crowdsourcing as a viable tool to analyse their data especially when the initial investment to set up crowdsourcing is outweighed by the ability to scale analysis (e.g., to segment 3D tissue samples; Booth *et al*, 2015). For those doing so, our open-source software can be used freely. Crowdsourcing in biomedicine is becoming more widespread (see for example https://citscibio.org/), and cancer research in particular stands to benefit a great deal from further investment given a combination of research need and strong public support.

ACKNOWLEDGEMENTS

We thank all the volunteers that contributed to the experiments presented here. We are grateful to the patients for making their biopsies available for research. We also thank the developers at Crowdcrafting/PyBossa and the Digital and IT teams at Cancer Research UK for technical support. We acknowledge support from Amazon.com, Inc. in providing free hosting credits for the project. We thank the rest of the Cancer Research UK Citizen science team: Amy Garcia, Leslie Harris, Jess Vasiliou, Josh Lee and Hannah Keartland. Dryden Williams and Steven Harris contributed to the codebase for the platform. Mike Thompson contributed to UX design, as did Chiara Garratini and Jonny Hancox from Intel. We thank Alexandra Walker, University of Oxford, for her p53 and Ki67 staining of the bladder TMAs. We acknowledge funding from Cancer Research UK, Cancer Research UK programme grant C5255/A15935 (AEK), and Cancer Research UK Accelerator Grant C11512/A20256 (GJT).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Andersen E, O'Rourke E, Liu Y-E, Snider R, Lowdermilk J, Truong D, Cooper S, Popovic Z (2012) The impact of tutorials on games of varying complexity. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*. p 59ACM Press: New York, USA.

Banerji M, Lahav O, Lintott CJ, Abdalla FB, Schawinski K, Bamford SP, Andreescu D, Murray P, Raddick MJ, Slosar A (2010) Galaxy Zoo: reproducing galaxy morphologies via machine learning. *Mon Not R Astron Soc* **406**(1): 342–353.

Beck AH, Sangoi AR, Leung S, Marinelli RJ, Nielsen TO, van de Vijver MJ, West RB, van de Rijn M, Koller D (2011) Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci Transl Med* **3**(108): 108ra113.

Bolton KL, Garcia-Closas M, Pfeiffer RM, Duggan MA, Howat WJ, Hewitt SM, Yang XR, Cornelison R, Anzick SL, Meltzer P, Davis S, Lenz P, Figueroa JD, Pharoah PDP, Sherman ME (2010) Assessment of automated image analysis of breast cancer tissue microarrays for epidemiologic studies. *Cancer Epidemiol Biomarkers Prevent* **19**: 992–999.

Booth ME, Treanor D, Roberts N, Magee DR, Speirs V, Hanby AM (2015) Three–dimensional reconstruction of ductal carcinoma in situ with virtual slides. *Histopathology* **66**(7): 966–973.

Bouzin C, Lamba Saini M, Khaing K-K, Ambroise J, Marbaix E, Grégoire V, Bol V (2015) Digital pathology: elementary, rapid and reliable automated image analysis. *Histopathology* **68**(6): 888–896.

Candido Dos Reis FJ, Lynn S, Ali HR, Eccles D, Hanby A, Provenzano E, Caldas C, Howat WJ, McDuffus L-A, Liu B, Daley F, Coulson P, Vyas RJ, Harris LM, Owens JM, Carton AFM, McQuillan JP, Paterson AM, Hirji Z, Christie SK, Holmes AR, Schmidt MK, Garcia-Closas M, Easton DF, Bolla MK, Wang Q, Benitez J, Milne RL, Mannermaa A, Couch F, Devilee P, RAEM Tollenaar, Seynaeve C, Cox A, Cross SS, Blows FM, Sanders J, de Groot R, Figueroa J, Sherman M, Hooning M, Brenner H, Holleczek B, Stegmaier C, Lintott C, Pharoah PDP (2015) Crowdsourcing the General Public for Large Scale Molecular Pathology Studies in Cancer. *EBioMedicine* **2**: 681–689.

Cazier JB, Rao SR, McLean CM, Walker AK, Wright BJ, Jaeger EEM, Kartsonaki C, Marsden L, Yau C, Camps C, Kaisaki P, The Oxford-Illumina WGSC, Taylor J, Catto JW, Tomlinson IPM, Kiltie AE, Hamdy FC (2014) Whole-genome sequencing of bladder cancers reveals somatic CDKN1A mutations and clinicopathological associations with mutation burden. *Nat Commun* **5**, e-pub ahead of print 29 April 2014; doi:10.1038/ncomms4756.

Cooper S, Khatib F, Treuille A, Barbero J, Lee J, Beenen M, Leaver-Fay A, Baker D, Popović Z (2010) Predicting protein structures with a multiplayer online game. *Nature* **466**(7307): 756–760.

Giltnane JM, Rimm DL (2004) Technology insight: Identification of biomarkers with tissue microarray technology. *Nat Clin Pract Oncol* **1**(2): 104–111.

Good B, Su A (2013) Crowdsourcing for bioinformatics. *Bioinformatics*; e-pub ahead of print 19 June 2013; doi:10.1093/bioinformatics/btt333.

Howat WJ, Blows FM, Provenzano E, Brook MN, Morris L, Gazinska P, Johnson N, McDuffus LA, Miller J, Sawyer EJ (2015) Performance of automated scoring of ER, PR, HER2, CK5/6 and EGFR in breast cancer tissue microarrays in the Breast Cancer Association Consortium. *J Pathol Clin Res* **1**(1): 18–32.

Hunter JD (2007) Matplotlib: A 2D Graphics Environment. *Comput Sci Eng* **9**: 90–95.

Irshad H, Montaser-Kouhsari L, Waltz G, Bucur O, Nowak J, Dong F, Knoblauch NW, Beck AH (2014) Crowdsourcing image annotation for nucleus detection and segmentation in computational pathology: evaluating experts, automated methods, and the crowd. *Pac Symp Biocomput* 294–305.

Jones E, Oliphant T, Peterson P (2001) {SciPy}: Open source scientific tools for {Python}.

Kawrykow A, Roumanis G, Kam A, Kwak D (2012) Phylo: a citizen science approach for improving multiple sequence alignment. *PLoS One* **7**: e31362.

Kim JS, Greene MJ, Zlateski A, Lee K, Richardson M, Turaga SC, Purcaro M, Balkam M, Robinson A, Behabadi BF, Campos M, Denk W, Seung HS (2014) Space-time wiring specificity supports direction selectivity in the retina. *Nature* **509**: 331–336.

Konsti J, Lundin M, Joensuu H, Lehtimäki T, Sihto H, Holli K, Turpeenniemi-Hujanen T, Kataja V, Sailas L, Isola J, Lundin J (2011) Development and evaluation of a virtual microscopy application for automated assessment of Ki-67 expression in breast cancer. *BMC Clin Pathol* **11**: 3.

Land-Zandstra AM, Devilee JLA, Snik F, Buurmeijer F, van den Broek JM (2016) Citizen science on a smartphone: Participants' motivations and learning. *Public Underst Sci* **25**: 45–60.

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* **33**(1): 159–174.

Lee K, Zlateski A, Vishwanathan A, Seung HS (2015) Recursive Training of 2D-3D Convolutional Networks for Neuronal Boundary Detection. arXiv preprint arXiv:150804843.

Lintott CJ, Schawinski K, Slosar A, Land K, Bamford S, Thomas D, Raddick MJ, Nichol RC, Szalay A, Andreescu D, Murray P, Vandenberg J (2008) Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Mon Not R Astron Soc* **389**(3): 1179–1189.

McCarty Jr KS, Szabo E, Flowers JL, Cox EB, Leight GS, Miller L, Konrath J, Soper JT, Budwit DA, Creasman WT, Seigler HF, McCarty Sr KS (1986) Use of a monoclonal anti-estrogen receptor antibody in the immunohistochemical evaluation of human tumors. *Cancer Res* **46**: 4244s–4248s.

McKinney W (2010) Data Structures for Statistical Computing in Python. In: Varoquaux G, van der Walt S, Millman J (eds). *Proceedings of the 9th Python in Science Conference*. SciPy: Pasadena, CA, USA, pp 51-56.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine Learning in Python. *J Mach Learn Res* **12**: 2825–2830.

Raddick MJ, Bracey G, Gay PL, Lintott CJ, Cardamone C, Murray P, Schawinski K, Szalay AS, Vandenberg J (2013) Galaxy Zoo: Motivations of citizen scientists. arXiv preprint arXiv:13036886.

Rallapalli G, Saunders DG, Yoshida K, Edwards A, Lugo CA, Collin S, Clavijo B, Corpas M, Swarbreck D, Clark M, Downie JA, Kamoun S, MacLean D (2015) Lessons from Fraxinus, a crowd-sourced citizen science game in genomics. *Elife* **4**: e07460.

Ranard BL, Ha YP, Meisel ZF, Asch DA, Hill SS, Becker LB, Seymour AK, Merchant RM (2014) Crowdsourcing—harnessing the masses to advance health and medicine, a systematic review. *J Gen Intern Med* **29**(1): 187–203.

Robboy SJ, Weintraub S, Horvath AE, Jensen BW, Bruce Alexander C, Fody EP, Crawford JM, Clark JR, Cantor-Weinberg J, Joshi MG, Cohen MB, Prystowsky MB, Bean SM, Gupta S, Powell SZ, Speights Jr VO, Gross DJ, Stephen Black-Schaffer W (2013) Pathologist workforce in the united states i. development of a predictive model to examine factors influencing supply. *Arch Pathol Lab Med* **137**: 1723–1732.

Rotman D, Preece J, Hammock J, Procita K, Hansen D, Parr C, Lewis D, Jacobs D (2012) Dynamic changes in motivation in collaborative citizen-science projects. In: *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12*. p 217, ACM Press: New York, USA.

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) ImageNet Large Scale Visual Recognition Challenge. *Int J Comput Vision* **115**: 211–252.

Rzeszotarski JM, Chi E, Paritosh P, Dai P (2013) Inserting micro-breaks into crowdsourcing workflows. *First AAAI Conference on Human Computation and Crowdsourcing*. Association for the Advancement of Artificial Intelligence: Palm Springs, CA, USA.

Sauermann H, Franzoni C (2015) Crowd science user contribution patterns and their implications. *Proc Natl Acad Sci USA* **112**(3): 679–684.

Seung H, Burnes L (2012) Eyewire. Available at http\\:eyewire.org.

Shah NB, Zhou D (2015) Double or nothing: multiplicative incentive mechanisms for crowdsourcing. *Adv Neural Inf Process Syst* **28**: 1–9.

Shaw AD, Horton JJ, Chen DL (2011) Designing incentives for inexpert human raters. *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. ACM New York: Hangzhou, China, pp 275-284.

Starr J, Schweik CM, Bush N, Fletcher L, Finn J, Fish J, Bargeron CT (2014) Lights, camera…citizen science: assessing the effectiveness of smartphone-based video training in invasive plant identification. *PLoS One* **9**: e111433.

Turbin DA, Leung S, Cheang MCU, Kennecke HA, Montgomery KD, McKinney S, Treaba DO, Boyd N, Goldstein LC, Badve S, Gown AM, van de Rijn M, Nielsen TO, Gilks CB, Huntsman DG (2008) Automated quantitative analysis of estrogen receptor expression in breast carcinoma does not differ from expert pathologist scoring: a tissue microarray study of 3484 cases. *Breast Cancer Res Treat* **110**(3): 417–426.

van der Walt SF, Colbert SC, Varoquaux Gl (2011) The NumPy array: a structure for efficient numerical computation. *Comput Sci Eng* **13**: 22–30.

Walton NA, Brenton JD, Caldas C, Irwin MJ, Akram A, Gonzalez-Solares E, Lewis JR, MacCullum P, Morris LJ, Rixon GT (2009) PathGrid: The Transfer of Astronomical Image Algorithms to the Analysis of Medical Microscopy Data. *Astronomical Data Analysis Software and Systems XVIII*

ASP Conference Series. Vol 411, Astornomical Society of the Pacific: Quebec City, Canada.

Ward M, Thirdborough S, Mellows T, Riley C, Harris S, Suchak K, Webb A, Hampton C, Patel N, Randall C (2014) Tumour-infiltrating lymphocytes predict for outcome in HPV-positive oropharyngeal cancer. *Br J Cancer* **110**(2): 489–500.

Wilbur DC (2014) Digital pathology: get on board—the train is leaving the station. *Cancer Cytopathol* **122**(11): 791–795.

Wilkins BS (2015) Pathology in Cancer Research. National Cancer Research Institute [Online]. http://www.ncri.org.uk/initiatives/pathology (accessed on 2 December 2016).

Wright DR, Underhill LG, Keene M, Knight AT (2015) Understanding the Motivations and Satisfactions of Volunteers to Improve the Effectiveness of Citizen Science Programs. *Soc Nat Resour* **28**: 1013–1029.