Pathology (January 2015) **47(1)**, pp. 7–12

ANATOMICAL PATHOLOGY

Development and validation of a gene expression tumour classifier for cancer of unknown primary

Richard W. Tothill[1], Fan Shi[2], Lisa Paiman[3], Justin Bedo[2], Adam Kowalczyk[2], Linda Mileshkin[1], Evangeline Buela[3], Robert Klupacs[4], David Bowtell[1,5,6,7] and Keith Byron[3]

[1]Peter MacCallum Cancer Centre, East Melbourne, [2]National (ICT) Australia, The University of Melbourne, Parkville, [3]Healthscope Pathology, Clayton, [4]Circadian Technologies Limited, Toorak, [5]The Sir Peter MacCallum Department of Oncology, University of Melbourne, Parkville, [6]The Department of Pathology, University of Melbourne, Parkville, and [7]The Department of Biochemistry, University of Melbourne, Parkville, Vic, Australia

Summary

Accurate identification of the primary tumour in cancer of unknown primary (CUP) is required for effective treatment selection and improved patient outcomes. The aim of this study was to develop and validate a gene expression tumour classifier and integrate it with histopathology to identify the likely site of origin in CUP. RNA was extracted from 450 formalin fixed, paraffin embedded samples of known origin comprising 18 tumour groups. Whole genome expression analysis was performed using a bead-based array. Classification of the tumours made use of a binary support vector machine, together with recursive feature elimination. A hierarchical tumour classifier was developed and incorporated with conventional histopathology to identify the origins of metastatic tumours. The classifier demonstrated an accuracy of 88% for correctly predicting the tumour type on a validation set of known tumours ($n = 94$). For CUP samples ($n = 49$) having a final clinical diagnosis, the classifier improved the accuracy of histology alone for both single and multiple predictions. Furthermore, where histology alone could not suggest any specific diagnosis, the classifier was able to correctly predict the primary site of origin. We demonstrate the integration of gene expression profiling with conventional histopathology to aid the investigation of CUP.

Key words: Carcinoma, classifier, CUP, gene expression, metastasis.

Received 12 May, revised 31 August, accepted 4 September 2014

INTRODUCTION

Cancers of unknown primary (CUP) account for 3–5% of all malignancies[1–4] and are among the ten most frequent cancers worldwide. The prognosis for patients with CUP is generally poor, with median survival in the order of 9 months.[5–7] Although extensive clinical and diagnostic investigations are typical in patients presenting with CUP, the origin of the tumour still remains unknown in approximately 20–50% of cases. Furthermore, reported rates of discovering a primary tumour at autopsy range from 55–85%.[8–10] In light of the poor prognosis, morbidity and patient anxiety, oncologists often must consider how far to pursue identification of the primary tumour in patients with CUP. However, when a primary tumour has been identified and specific treatment initiated, improved response rates and overall survival has been shown in some cases.[11,12] Therefore, making an accurate diagnosis of the primary site in CUP patients in a timely and cost-effective manner is imperative.

Previous studies have shown that CUP tumours are predominantly either metastatic adenocarcinomas that are well-to-moderately differentiated (50%), poorly differentiated adenocarcinomas/carcinomas (30%), or squamous cell carcinomas (15%), with the remainder being undifferentiated tumours, sarcomas, melanomas and neuroendocrine neoplasms.[3,13,14] Histopathology and immunohistochemistry (IHC) are important in the investigation of CUP; however, their accuracy in identifying the primary site of metastatic tumours has been shown to be moderate, even when extensive antibody panels are used,[15] and falls further in the diagnostic work-up of CUP cases.[16–19] Particularly low rates of identification are found in poorly differentiated or undifferentiated tumours where morphology and IHC staining may be non-specific or inconclusive. An atypical pattern of immunoreactivity in some undifferentiated tumours may also hamper classification.

Metastatic tumours often retain gene expression profiles associated with their parent organ, allowing the development of classifiers to predict the site of origin for metastatic deposits.[19–21] Such assays have classification accuracies of 75–87% for known tumours.[22–27] To a large extent, these classifiers have been used as a standalone diagnostic, rather than attempting to integrate their data with traditional histopathology. Here we describe the development and validation of a gene expression tumour classifier and its integration with routine histopathology to aid in the identification of the primary tumour in CUP.

MATERIALS AND METHODS

Samples

Routine diagnostic formalin fixed, paraffin embedded (FFPE) tumour samples were obtained from collaborating institutions, together with details of histopathology findings. All samples were obtained according to a protocol approved by the Peter MacCallum Cancer Centre Ethics Review Committee and consistent with Australian National Health and Medical Research Council human subject research guidelines. Fifty-eight tumour samples from patients with metastatic disease with no clear clinical primary tumour were also referred to the study by treating oncologists (CUP samples). The time from initial clinical presentation and routine diagnostic workup to the receipt of the sample for gene expression analysis varied for each sample. Each oncologist was contacted

Print ISSN 0031-3025/Online ISSN 1465-3931
DOI: 10.1097/PAT.0000000000000194

6 months after gene expression analysis to obtain a final clinical diagnosis for each of the samples referred. The diagnosis of the primary site in each case was based upon histopathology, imaging, clinical suspicion, or any combination of these.

RNA extraction

A haematoxylin and eosin (H&E) stained section was used to evaluate tumour cellularity, and to guide macrodissection from unstained sections. Dissected tissue was de-paraffinised by xylene and RNA extracted using the RNeasy FFPE Kit (Qiagen, Germany) according to the manufacturer's instructions. RNA recovery was measured using the Quant-iT Ribo Green kit (Life Technologies, USA) and a NanoDrop 3300 fluorimeter (Thermo Scientific, USA). RNA integrity was assessed in duplicate by a quantitative real-time PCR amplification of a 90 bp fragment of the RPL13a gene in the presence of SYBR Green and using a 7900 real-time thermal cycler (Life Technologies). A delta CT of <7 compared to intact RNA was used as a cut-off for sample inclusion.

Microarray gene expression

HumanHT-12 v4 BeadChip arrays (Illumina, USA) containing 29,285 gene-specific oligonucleotide probes were used to profile gene expression of FFPE samples. Two hundred nanograms of RNA were used for labelling and hybridisation according to the manufacturer's instructions. Initial data analysis was performed with GenomeStudio (Illumina) to obtain a quality report. Outliers were discarded if preset criteria for quality metrics, including hybridisation, stringency, gene intensity and background noise, were not met.

Bioinformatics

Gene expression data were normalised with the average normalisation algorithm[28] (GenomeStudio). Arrays detecting less than 16,000 genes were not included in further analyses. To train the tumour classifier we employed a support vector machine (SVM)[29] with recursive feature elimination as a gene selection method.[30] The SVM projected the gene-expression data into multi-dimensional (n gene) space using a kernel function. A classification boundary (hyperplane) was set in training that maximally separates the two classes (e.g., class A from class B). Classification of an unknown sample was made by projecting that sample into the same expression space enabling generation of a classification score (distance from hyperplane). For multi-class classification a one-versus-all (OVA) approach was used whereby multiple binary classifiers are firstly derived for each tumour type (e.g., colorectal versus all other tumour types). Each test sample is run through each tumour type classifier generating a prediction score for each algorithm (further explanatory notes in Supplementary Data Section 1, http://links.lww.com/PAT/A25). The tumour type SVM classification with the highest prediction score is deemed to be the first prediction, the second highest the second prediction and so forth. A leave-one-out cross validation (LOOCV) protocol[31] was initially employed to assess classification accuracy. Importantly, genes were reselected during each round of cross-validation. The final classifier used for testing on an independent test set of tumours of known origin and CUP samples was generalised from the entire training set of 450 samples using the optimal number of genes identified during LOOCV.

Histology-guided classification

Histology-guided tumour type classification was dependent upon a pathologist reviewing the tumour's morphological appearance under light microscopy, as well as IHC findings, including staining for cytokeratins, epithelial membrane antigens, vimentin, S-100 and melan-A. Based on these parameters the pathologist would then direct the use of a 'carcinoma-type' or 'non carcinoma-type' classifier (a case example of analysis can be found in Supplementary Data Section 2, http://links.lww.com/PAT/A25). In cases of undifferentiated tumours a third algorithm would be used which was trained to predict epithelial-like gene expression profiles within a tumour (epi versus non-epi algorithm).

RESULTS

Classifier training

To establish a reference dataset of gene expression profiles associated with a large number of known tumours that are typically associated with CUP, we obtained 544 FFPE tumour samples of known origin. By contrast with other studies,[23,25] we particularly enriched for metastatic rather than primary tumour samples where possible, since these reflect the clinical circumstances of CUP. Tumour groups included in the classifier were selected from reported autopsy-proven CUP primaries as well as from previous gene expression findings.[2] Lung, pancreas, liver, kidney, colorectal, stomach, prostate and ovary are the most common primary tumour sites found in CUP, with rarer types being thyroid, neuroendocrine and urinary bladder. Squamous cell carcinoma (SCC) has also been found in significant number of CUP cases at autopsy.

After filtering based on RNA quality and yield, whole genome expression data was obtained for 450 FFPE samples encompassing 18 tumour types and comprising the training set (Table 1). Many of the genes selected for the classifier (Supplementary Table 1, http://links.lww.com/PAT/A25) are not well characterised for each of the tumour classes within the database. However, many of the highly informative genes selected for each class have either a recognisable specific function for each tissue type or play a significant role in cancer development or cell cycle progression.

Initially the classifier was trained to predict tumour classes for all tumour groups tested. Using OVA analysis, we demonstrated the prediction accuracy of this to be 77.3%. In an attempt to improve the accuracy, the classification problem was decomposed into three independent classification algorithms. The first algorithm can broadly predict epithelial versus non-epithelial tumour types with an accuracy of 92% by LOOCV. In a clinical setting this classifier would be reserved for undifferentiated tumours where histopathology could not resolve whether the cancer had epithelial origins. The second and third classification algorithms were a carcinoma-type (14-class) algorithm and a non-carcinoma type (4-class) algorithm, with accuracies of 81% and 86%, respectively, by LOOCV. The combined accuracy of the histology-guided classifiers is 82%, which is superior to the reported OVA 18-class classification. As such, the histology guided classifier would be locked and utilised for the remainder of the study (Fig. 1).

Validation set

To validate the classifier, an independent set of 94 known tumours were used that had a similar distribution of tumour types to that used in the training set (Supplementary Table 2, http://links.lww.com/PAT/A25). The histology-guided gene

Table 1 Tumours used to train the classifier

Tumour class and type	Number	Metastatic tumours
Carcinoma		
Urinary bladder	16	44%
Breast	55	85%
Cholangiocarcinoma	10	40%
Colorectal	54	74%
Gastric	28	61%
Kidney	15	80%
Liver	13	15%
Lung	31	90%
Neuroendocrine	26	31%
Ovary	22	64%
Pancreas	17	59%
Prostate	17	94%
Squamous cell	32	34%
Thyroid	21	62%
Non-carcinoma		
Melanoma	50	42%
Mesothelioma	5	40%
Sarcoma	27	4%
Testicular	11	18%