

Epithelial–mesenchymal transition biomarkers and support vector machine guided model in preoperatively predicting regional lymph node metastasis for rectal cancer

X-J Fan^{1,5}, X-B Wan^{2,5}, Y Huang^{1,5}, H-M Cai³, X-H Fu¹, Z-L Yang¹, D-K Chen¹, S-X Song¹, P-H Wu¹, Q Liu⁴, L Wang^{*,1} and J-P Wang^{*,1}

¹Gastrointestinal Institute of Sun Yat-sen University, the Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou, Guangdong 510655, China;

²Department of Medical Oncology, the Third Affiliated Hospital, Sun Yat-sen University, Guangzhou, Guangdong 510630, China; ³School of Information Science and Technology, Sun Yat-sen University, Guangzhou, Guangdong 510000, China; ⁴State Key Laboratory of Oncology in Southern China, Cancer Center, Sun Yat-sen University, Guangzhou, Guangdong 510060, China

BACKGROUND: Current imaging modalities are inadequate in preoperatively predicting regional lymph node metastasis (RLNM) status in rectal cancer (RC). Here, we designed support vector machine (SVM) model to address this issue by integrating epithelial–mesenchymal-transition (EMT)-related biomarkers along with clinicopathological variables.

METHODS: Using tissue microarrays and immunohistochemistry, the EMT-related biomarkers expression was measured in 193 RC patients. Of which, 74 patients were assigned to the training set to select the robust variables for designing SVM model. The SVM model predictive value was validated in the testing set (119 patients).

RESULTS: In training set, eight variables, including six EMT-related biomarkers and two clinicopathological variables, were selected to devise SVM model. In testing set, we identified 63 patients with high risk to RLNM and 56 patients with low risk. The sensitivity, specificity and overall accuracy of SVM in predicting RLNM were 68.3%, 81.1% and 72.3%, respectively. Importantly, multivariate logistic regression analysis showed that SVM model was indeed an independent predictor of RLNM status (odds ratio, 11.536; 95% confidence interval, 4.113–32.361; $P < 0.0001$).

CONCLUSION: Our SVM-based model displayed moderately strong predictive power in defining the RLNM status in RC patients, providing an important approach to select RLNM high-risk subgroup for neoadjuvant chemoradiotherapy.

British Journal of Cancer (2012) **106**, 1735–1741. doi:10.1038/bjc.2012.82 www.bjcancer.com

Published online 26 April 2012

© 2012 Cancer Research UK

Keywords: SVM; EMT; regional lymph node metastasis; colorectal cancer

Colorectal cancer remains one of the leading causes of cancer mortality worldwide (Jemal *et al*, 2009). For rectal cancer (RC) patients with regional lymph node metastasis (RLNM), preoperative combined-modality treatment displays favourable therapeutic outcomes, including downsizing node-stage, reduction in local recurrence and treatment-associated toxicity (Minsky *et al*, 1992; Leonard *et al*, 2005; Sebag-Montefiore *et al*, 2009). Given its pivotal role in guiding the therapeutic selection, accurate assessment of RLNM status become an urgent step prior to the administration of preoperative therapeutic regimen. Although imaging modalities, including endoscopic ultrasound (EUS), magnetic resonance imaging (MRI) and computed tomography (CT), have been widely used in clinical practice to estimate the RLNM status, the sensitivities and specificities for accurately evaluating node

involvement were, however, less than 70% and 80%, respectively (Bipat *et al*, 2004). In addition, use of size and morphological criteria by current imaging systems may lead to misdiagnose, particularly when the nodes < 5 mm in size with smooth contour would greatly lowered the predictive sensitivity, specificity and accuracy (Liang *et al*, 1999). Clearly, to develop a more accurate and clinical adaptable approach to predict RLNM status will be of great benefit for RC patients from preoperative therapy.

Epithelial–mesenchymal transition (EMT) is a process whereby epithelial cells lose cell-to-cell adhesion and gain mesenchymal phenotype characterised by repression of membrane proteins such as E-cadherin and β -catenin, and overexpression of nuclear β -catenin, N-cadherin, Snail and Twist (Brabletz *et al*, 2001; Blanco *et al*, 2002; Thiery, 2002; Peinado *et al*, 2007). EMT endows cells with migratory and invasive properties, thus promotes tumour cells to migrate to regional lymph node or distant organ. Altered expression of EMT-related biomarker p120 catenin from membrane to cytoplasm, for example, correlated with a shortened overall survival time, elevated tumour stage and lymph node metastasis in colon cancer (Bellocin *et al*, 2005). Similarly, overexpression of EMT inducer molecule Nanog upregulated Snail and Slug activity, promoted the cellular motility and predicted a

*Correspondence: Professor J-P Wang or Professor L Wang;

E-mail: wangjjpgz@yahoo.com.cn (JPW) or

leiwangyinghu@yahoo.com.cn (LW)

⁵ These authors contributed equally to this work.

Received 2 November 2011; revised 14 February 2012; accepted 20 February 2012; published online 26 April 2012

poor prognosis in colorectal cancer (Meng *et al*, 2010). Beyond the molecular markers, tumour-specific biomarkers (cancer antigen (CA) 199 and carcinoembryonic antigen (CEA)) and clinicopathological parameters had been reported to predict lymph node metastasis in colorectal cancer (Mroczo *et al*, 2007; Fang *et al*, 2009). However, it was admitted that the single marker, which reflected the cancer phenotype from only one specific aspect, was hard to predict tumour progression individually (McShane *et al*, 2005). Hence, to develop a method that can integrate the biomarkers in a more wide scope might achieve the goal of accurately predicting RLNM status in RC patients.

Recently, data mining technology, such as decision tree (Dietzel *et al*, 2010), neurofuzzy model (Catto *et al*, 2009), artificial neural networks (Anderson *et al*, 2010), logistic regressions (Vermeulen *et al*, 2009), Bayesian network (Stojadinovic *et al*, 2010) and support vector machines (SVM) (Zhu *et al*, 2009), have been applied to predict tumour progression and clinical outcome by integrating multigene and/or clinicopathological features. Support vector machine algorithms method, for example, performs a higher power in two categories classification, displaying strong predictive power in diagnosing lymph node metastasis for breast cancer and cutaneous melanoma (Mocellin *et al*, 2006; Sattlecker *et al*, 2010). Thus, multi-markers incorporated SVM algorithms might shed the bright light on the way to personalised prediction of RLNM status in RC.

In the present study, we analysed the expression levels of seven EMT signalling biomarkers by immunohistochemistry (IHC), as well as demographical, clinicopathological and serological variables in 193 RC patients. In the training set (74 cases), the robust variables that had high predictive ability in predicting RLNM status were subjected to SVM model construction. Furthermore, we validated the predictive value of our SVM model in the testing set (119 cases). Our study revealed that SVM model had powerful predictive value in refining RLNM status, providing a novel approach to tailored selection of RC patients for neoadjuvant therapy.

PATIENTS AND METHODS

Patient selection

One hundred and ninety-three pathologically confirmed RC patients with or without RLNM were recruited from April 2007 to December 2009 in Gastrointestinal Institute of Sun Yat-sen University, the Sixth Affiliated Hospital of Sun Yat-sen University. None of these patients received neo-adjuvant chemo- or chemoradiotherapy. Of which, 74 patients were randomly assigned to the training set and the rest of 119 patients were assigned to the testing set for validating the predictive power of the SVM model. Detailed, 37 patients (50.0%) in the training set and 82 patients (68.9%) in the testing set were pathologically confirmed with RLNM. The clinic-serologic variables, including gender, age, tumour stage, CEA, CA19-9 and CA125, were all obtained preoperatively. The patients were classified according to the 7th edition of the American Joint Committee on Cancer staging system (Takahashi *et al*, 2007). The detailed characteristics of the two cohorts were listed in Table 1. Our study was approved by Clinical Ethics Review Committee at the Sixth Affiliated Hospital of Sun Yat-sen University and the clinical consent was obtained from all the patients.

Tissue microarrays (TMAs)

The TMAs of 193 RC tumour specimens and additional 20 normal tissues were collected from the Tissue Bank at the Gastrointestinal Institute of Sun Yat-sen University, the Sixth Affiliated Hospital, Sun Yat-sen University. As previously reported that EMT occurred

Table 1 Patient characteristics

Variable	Training set (n = 74)	Testing set (n = 119)
Gender		
Male	40	74
Female	34	45
Age (Year)		
Mean ≥ 60.0 vs < 60.0	32 vs 42	56 vs 63
Range	18–87	21–89
Tumour stage		
T1+T2	8	12
T3	63	104
T4	3	3
RLNM status		
Negative	37	37
Positive	37	82
CEA ($ng\ ml^{-1}$)		
Mean ≥ 16.3 vs < 16.3	12 vs 62	19 vs 100
Range	0.01–1404	0.41–471.4
CA19-9 ($unit\ ml^{-1}$)		
Mean ≥ 35.8 vs < 35.8	13 vs 61	23 vs 96
Range	0.08–3508	0.8–863.6
CA125 ($unit\ ml^{-1}$)		
Mean ≥ 16.8 vs < 16.8	17 vs 57	28 vs 91
Range	0.5–633.1	0.7–909.4

Abbreviations: CA = cancer antigen; CEA = carcinoembryonic antigen; RLNM = regional lymph node metastasis.

at the invasive front of colorectal adenocarcinoma (Brabletz *et al*, 2001), we re-reviewed hematoxylin and eosin (H&E)-stained slides and selected the marginal area of tumour for TMAs design. For each case, three cores (1 mm diameter) were taken from the selected marginal region of the tumour and additional two cores (1 mm diameter) were taken from histologically confirmed normal adjacent colorectal mucosa to construct the TMAs using Tissue Array (Alphelys, MINIPORE, Plaisir, France). Briefly, the tissue cylinders were taken from the selected region of each donor tissue block and deposited into a recipient block. The recipient blocks were further performed by H&E staining to verify the adequacy of the tumour and normal tissues.

Immunohistochemistry staining

The TMAs were sectioned at 4 μm intervals, deparaffinised three times in xylene for 30 min and rehydrated with graded alcohols (100% ethyl alcohol for 5 min, 95% ethyl alcohol for 3 min and 75% ethyl alcohol for 3 min). Sections were then heated in antigen retrieval solution (sodium citrate, pH 6.0) in microwave for 15 min, incubated in H_2O_2 for 10 min and blocked with 1% bovine serum albumin (diluted in PBS buffer) for 10 min at room temperature. Further, the sections were incubated with the 100 μl primary antibodies against epithelial markers E-cadherin (rabbit monoclonal, clone 24E10, Cell Signaling, #3195, Danvers, MA, USA, diluted 1:50) and β -catenin (mouse monoclonal, clone 5H10, Millipore, MAB2081, Billerica, MA, USA, diluted 1:100), mesenchymal cell marker N-cadherin (mouse monoclonal, clone 13A9, Upstate, 05-915, Waltham, MA, USA, diluted 1:100), transcriptional markers Snail (rabbit polyclonal, clone L70G2, Abcam, ab70983, San Francisco, CA, USA, diluted 1:200) and Twist (rabbit polyclonal, clone O-20, Santa Cruz, sc-102032, CA, USA, diluted 1:200), at 4°C overnight. All of the primary antibodies were diluted with DAKO antibody diluent solution. Furthermore, the

sections were washed in PBS buffer three times for 5 min, treated with 100 μ l secondary antibody for 30 min at room temperature and stained with DAB until the brown granules appeared in membrane, cytoplasm or nucleus according to the manufacturer's instructions. The sections were then washed in PBS buffer for 10 min. Finally, the sections were counterstained with hematoxylin for 2 min at room temperature, washed in dH₂O 3 times for 5 min and mounted the coverslips. The normal epithelium in each section served as the internal negative (N-cadherin, Snail and Twist) or positive (E-cadherin and β -catenin) control. The ganglion cells of the myenteric plexus were used as positive control for N-cadherin reactivity. The RC samples that expressed strong positive staining of Snail or Twist with lowest background were used as positive control for Snail and Twist, respectively. Meanwhile, a negative control was also employed by changing the specific primary antibody with non-immune serum immunoglobulins at the 1:200 dilution.

Semi-quantitative evaluation of IHC staining

The level of the immunoreaction was semi-quantitatively scored by assessing staining intensity and extent as we previously reported (Wan *et al*, 2010). We scored the staining intensity as following: negative (score 0), bordering (score 1), weak (score 2), moderate (score 3) and strong (score 4). Staining extent was graded into five parts according to the percentage of staining cells in the field: negative (score 0), 0–25% (score 1), 26–50% (score 2), 51–75% (score 3) and 76–100% (score 4). The merged overall score (staining intensity score plus percentage of staining cells score) was subjected to further analysis. For nuclear β -catenin, a positive or negative score was given for each tissue sample when $\geq 5\%$ or $< 5\%$ of tumour cells showed nuclear staining, respectively (Baumgart *et al*, 2007). Immunohistochemistry staining was scored independently by two pathologists (Y Huang and XJ Fan) blindly to clinical data.

Selection of cutoff score for 'positive' expression of variables

The receiver operating characteristic (ROC) curve analysis was used to select cutoff scores of all candidate variables in the training sets. At each score, the sensitivity and specificity being studied was plotted to generate a ROC curve. Briefly, the score localised closest to the point with both maximum sensitivity and specificity, the point (0.0, 1.0) on the curve, was chosen as the cutoff score, leading to the greatest number of patients classified as with or without RLNM. According to the ROC curve analysis, SVM model and all the candidate variables were dichotomised: SVM (positive RLNM vs negative RLNM), gender (male vs female), age (≥ 62.5 vs < 62.5 years), tumour stage (T3 + T4 vs T1 + T2), CEA (≥ 3.90 vs < 3.90), CA19-9 (≥ 13.35 vs < 13.35), CA125 (≥ 10.00 vs < 10.00), β -catenin (nucleus, positive vs negative) and the other EMT-related biomarkers (high level vs low level).

The RLNM status prediction by SVM model

The SVM model, coded by Matlab software (MathWorks, Natick, MA, USA), was used to predict the RLNM status. Firstly, we selected the variables that had high power in predicting RLNM status, from all the candidate variables by SVM method and ROC analysis. Secondly, we designed and trained our SVM model by integrating the selected variables in the training set. After the completion of the training process, the algorithmic SVM model would be 'fixed' for further running. The detailed steps of the SVM model construction were shown in Supplementary Information. In the testing set, the 'feature' of the selected variables in each patient would be input into the SVM model. Finally, the RLNM status of each patient would be predicted and output as 0 (without

RLNM) or 1 (with RLNM) by our SVM model. The output results of each patient would be subjected to further univariate and multivariate analysis.

Statistical analysis

The correlations between expression levels of EMT-related biomarkers and RLNM status was evaluated by chi-square test. The univariate and multivariate analyses were performed by binary logistic regression model to estimate the odds ratio (OR) and 95% confidence interval (95% CI). This study was designed with 80% power (two-sided α level of 0.05) to construct the SVM prediction model. All *P*-values quoted were two-sided and *P* < 0.05 was considered statistically significant. Statistical analysis was performed using SPSS v. 17.0 (SPSS, Inc., Chicago, IL, USA).

RESULTS

EMT-related biomarkers expression in RC

We investigated the expression levels of EMT-related biomarkers in the training set (*n* = 74) and testing set (*n* = 119) by IHC analysis. In normal epithelia, epithelial markers E-cadherin and β -catenin showed strong membranous expression (Figure 1a1 and 1a2). Conversely, mesenchymal marker N-cadherin, as well as transcriptional factors Snail and Twist were failed to be detected (Figure 1a3–1a5). In RC tumour tissues with negative RLNM, membranous markers of E-cadherin and β -catenin were reduced expression in both training set and testing set (Figure 1b1 and 1b2). Meanwhile, mesenchymal marker N-cadherin, as well as transcriptional factors Snail and Twist showed weak expression (Figure 1b3–1b5). In RC tumour tissues with positive RLNM, E-cadherin and β -catenin were lost expression (Figure 1c1 and 1c2). Whereas, N-cadherin, Snail and Twist showed strong expression (Figure 1c3–1c5). Moreover, chi-square test showed that only Snail expression level was evidently varied between RLNM positive and negative subset in the testing set (*P* = 0.001, Table 2). However, the significantly different expression level was not observed in RLNM positive and negative subgroups regarding to E-cadherin (*P* = 0.327), N-cadherin (*P* = 0.424), β -catenin (membrane, *P* = 0.426), β -catenin (cytoplasm, *P* = 0.406), β -catenin (nucleus, *P* = 0.747) and Twist (*P* = 0.845, Table 2) in the testing set. Similarly, no significant relationship was found between the EMT-related biomarkers and RLNM status (E-cadherin, *P* = 0.484; N-cadherin, *P* = 1.000; membranous β -catenin, *P* = 0.484; cytoplasmic β -catenin, *P* = 1.000; nuclear β -catenin, *P* = 1.000; Snail, *P* = 0.484 and Twist, *P* = 0.243; Table 2) in the training set.

Association between EMT-related biomarkers expression, clinicopathological variables and RLNM status

The cutoff point of each variable was determined by ROC curve analysis in the training set. In the testing set, the ROC-curve generated cutoff point of each variable was subjected to univariate analysis with respect to their RLNM status. In the training set, age (OR, 2.410; 95% CI, 0.947–6.131; *P* = 0.065) showed borderline significance in predicting RLNM status. In the testing set, Snail (OR, 4.286; 95% CI, 1.692–10.858; *P* = 0.002) and SVM model (OR, 9.231; 95% CI, 3.588–23.751; *P* < 0.0001) showed significantly high power in predicting RLNM status; whereas age (OR, 1.970; 95% CI, 0.895–4.334; *P* = 0.092) and nuclear β -catenin (OR, 2.981; 95% CI, 0.847–10.486; *P* = 0.089) were of borderline significance. Evidently, ROC curve analysis showed that SVM model owned the largest AUC compared with other variables (AUC of SVM was 0.747 in the testing set, Table 3, Figure 2), suggesting the more powerful predictive value of SVM model in refining RLNM status.

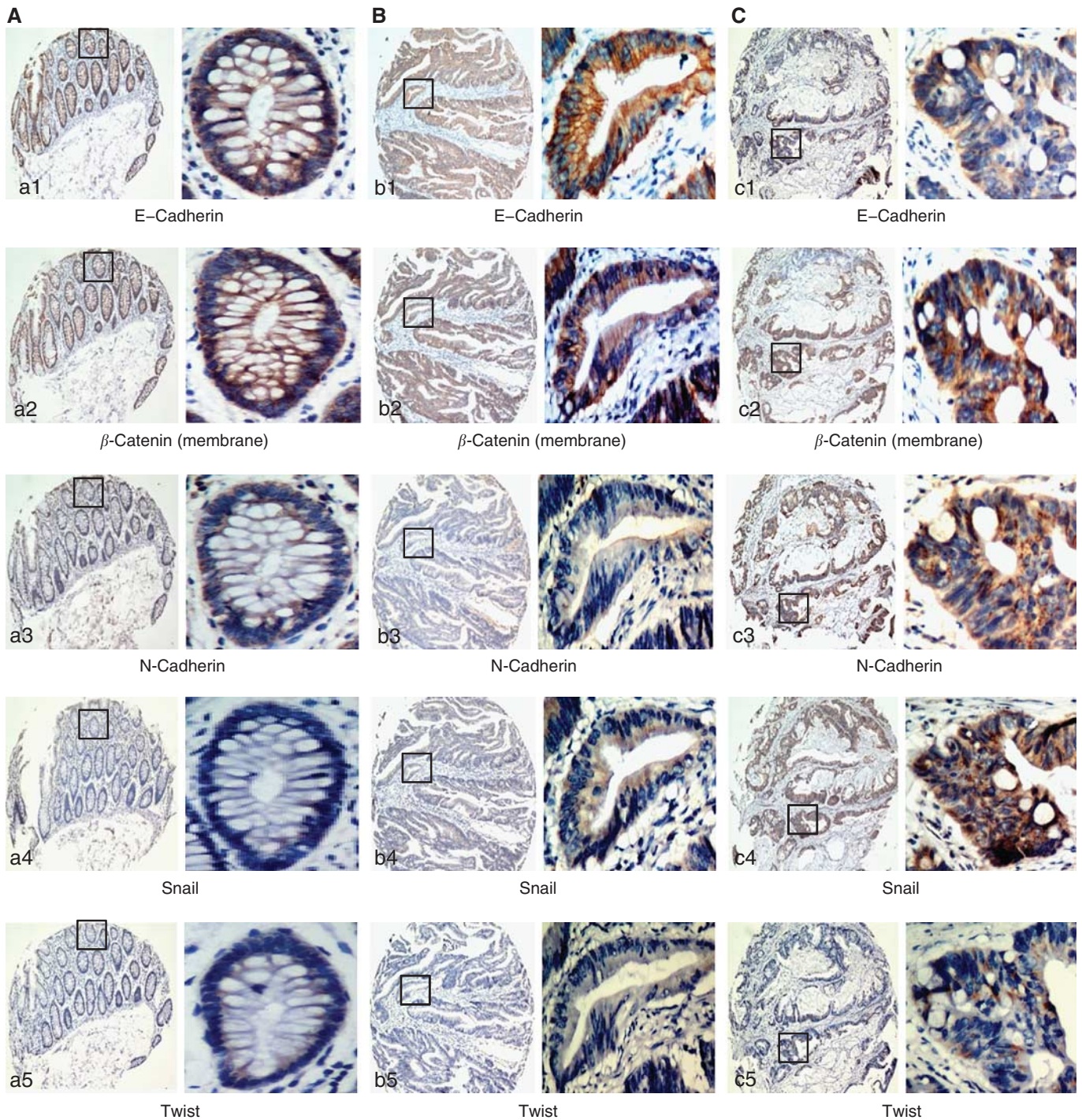


Figure 1 Immunohistochemical staining of EMT-related biomarkers in RC normal epithelium (**A**, left panels, $\times 100$, right panels, $\times 400$), tumour tissues with negative RLNM (**B**, left panels, $\times 100$, right panels, $\times 400$) and with positive RLNM (**C**, left panels, $\times 100$, right panels, $\times 400$). In each subgroup, right panels displayed representative area of left panels with enlarged view. Normal colorectal epithelium showed strong membranous expressions of E-cadherin and β -catenin (a1–2), negative expressions of N-cadherin (a3), Snail (a4) and Twist (a5). Tumour tissues with negative RLNM showed reduced membranous expressions of epithelial markers E-cadherin and β -catenin (b1–2), weak expressions of N-cadherin (b3), Snail (b4) and Twist (b5). Tumour tissues with positive RLNM lost the membranous expressions of epithelial markers E-cadherin and β -catenin (c1–2), compared with the strong expressions of N-cadherin (c3), Snail (c4) and Twist (c5). One representative staining of EMT-related biomarkers was shown.

The SVM model in defining the RLNM status

In the training set, six EMT-related biomarkers (E-cadherin, N-cadherin, cytoplasmic β -catenin, nuclear β -catenin, Snail and Twist) and two demographical variables (gender and age) were

integrated to design the SVM model. In the testing set (82 cases with RLNM and 37 cases without RLNM), we identified 63 cases with high risk to RLNM and 56 cases with low risk after educating the model in the training set. In RLNM subset, our SVM model identified 56 out of 82 cases with RLNM. Meanwhile, 30 of 37 cases

were identified as low risk in subgroup without RLNM. Detailed, the predictive value of SVM in sensitivity, specificity, positive predictive value, negative predictive value and overall accuracy were 68.3%, 81.1%, 88.9%, 53.6% and 72.3%, respectively. More importantly, our multivariate logistic regression analysis revealed that SVM model was indeed an independent predictive factor for RLNM (OR, 11.536; 95% CI, 4.113–32.361; $P < 0.0001$). Moreover, tumour stage was also displayed a significantly predictive value for

RLNM, though with the lower power (OR, 4.443; 95% CI, 1.064–18.557; $P = 0.041$) (Table 4).

DISCUSSION

Preoperative adjuvant therapy had proved to be of great survival benefit in RC patients with RLNM (Rohwer *et al*, 2010). The current imaging system was, however, inadequate to define the RLNM status, particularly for the regional lymph node with round shape and the size of < 5 mm (Low *et al*, 2008). Results from a meta-analysis demonstrated that the sensitivities and specificities of the three imaging modalities, EUS, CT and MRI, in evaluating the lymph node involvement were similar (EUS, 67 and 78%; CT, 55 and 74%; MRI, 66 and 76%) (Bipat *et al*, 2004; Engelen *et al*, 2008). In the present study, we designed a multi-markers and mathematical algorithm combined SVM approach to predict the RLNM status in RC patients. By integrating six EMT-related biomarkers and two demographical variables, we firstly educated our SVM model in the training set and further validated its predictive value in the testing set. Our internal validation showed that the SVM model owned the powerful efficacy in individually defining RLNM status (the sensitivity and specificity were 68.3% and 81.1%, respectively, in the testing set). We found that the sensitivity and specificity of our SVM model were similar to the current three imaging modalities. However, the advantage of our SVM model was marked for the metastatic regional lymph node with round shape and the size < 5 mm, which was a major limitation of the current imaging modalities. Thus, our SVM model might serve as a useful tool to complement the current imaging system in refining RLNM status and select patient tailored for preoperative adjuvant therapy in RC.

Data mining and molecular or genetic markers combined method had raised potential clinical implementation in prediction of patient outcome for solid tumour. In non-small-cell lung cancer (NSCLC), five gene signature (DUSP6, MMD, STAT1, ERBB3 and LCK)-based decision-tree analysis reached a 96% overall accuracy in predicting patient prognosis (Takahashi *et al*, 2007). Compared with other data mining algorithm, including of neural networks (artificial and fuzzy) (Catto *et al*, 2009; Anderson *et al*, 2010), clustering and decision trees (Modlin *et al*, 2009; Dietzel *et al*, 2010), SVM performed classification by constructing an N -dimensional space that optimally separates the data into two categories. This two-category feature made SVM model was superior in predicting RLNM status with two classifications

Table 2 Relationship between EMT-related biomarkers expression level and RLNM status in RC patients

Variable	Training set			Testing set		
	RLNM status		P-value	RLNM status		P-value
	Negative	Positive		Negative	Positive	
<i>E-cadherin</i>						
< 5.50	18	22	0.484	18	48	0.327
≥ 5.50	19	15		19	34	
<i>N-cadherin</i>						
< 4.50	20	19	1.000	24	46	0.424
≥ 4.50	17	18		13	36	
β -Catenin (membrane)						
< 4.50	18	22	0.484	14	39	0.426
≥ 4.50	19	15		23	43	
β -Catenin (cytoplasm)						
< 1.00	21	21	1.000	27	53	0.406
≥ 1.00	16	16		10	29	
β -Catenin (nucleus)						
Negative	35	35	1.000	31	77	0.747
Positive	2	2		6	5	
<i>Snail</i>						
< 5.50	22	18	0.484	30	41	0.001
≥ 5.50	15	19		7	41	
<i>Twist</i>						
< 4.50	17	23	0.243	19	40	0.845
≥ 4.50	20	14		18	42	

Abbreviations: EMT = epithelial–mesenchymal transition; RC = rectal cancer; RLNM = regional lymph node metastasis.

Table 3 Association between EMT-related biomarkers expression, clinicopathological variables and RLNM status in RC patients

Variable	Training set				Testing set			
	P-value	OR	95% CI	AUC	P-value	OR	95% CI	AUC
Gender, female vs male	0.352	1.548	0.617–3.885	0.446	0.686	1.182	0.527–2.650	0.481
Age, < 62.5 vs ≥ 62.5 years	0.065	2.410	0.947–6.131	0.378	0.092	1.970	0.895–4.334	0.440
Tumour stage, 1+2 vs 3+4	1.000	1.000	0.231–4.338	0.538	0.145	2.452	0.734–8.190	0.562
CEA, < 3.90 vs ≥ 3.90	0.816	1.114	0.448–2.773	0.563	0.619	1.223	0.552–2.711	0.524
CA19-9, < 13.35 vs ≥ 13.35	0.642	0.805	0.323–2.007	0.599	0.231	1.617	0.737–3.551	0.549
CA125, < 10.00 vs ≥ 10.00	1.000	1.000	0.402–2.489	0.524	0.437	1.362	0.625–2.968	0.446
<i>E-cadherin</i> , < 5.50 vs ≥ 5.50	0.352	1.548	0.617–3.885	0.419	0.316	1.490	0.683–3.251	0.430
<i>N-cadherin</i> , < 4.50 vs ≥ 4.50	0.816	0.897	0.360–2.236	0.501	0.369	0.692	0.310–1.546	0.564
β -Catenin (membrane), < 4.50 vs ≥ 4.50	0.352	1.548	0.617–3.885	0.409	0.324	1.490	0.674–3.294	0.506
β -Catenin (cytoplasm), < 1.00 vs ≥ 1.00	1.000	1.000	0.399–2.509	0.514	0.371	0.677	0.288–1.592	0.535
β -Catenin (nucleus), negative vs positive	1.000	1.000	0.133–7.502	0.500	0.089	2.981	0.847–10.486	0.449
<i>Snail</i> , < 5.50 vs ≥ 5.50	0.352	1.548	0.617–3.885	0.629	0.002	4.286	1.692–10.858	0.729
<i>Twist</i> , < 4.50 vs ≥ 4.50	0.164	1.933	0.765–4.884	0.431	0.795	0.902	0.415–1.962	0.533
SVM, 1 vs 0	< 0.0001	–	–	1.000	< 0.0001	9.231	3.588–23.751	0.747

Abbreviations: AUC = area under the ROC curve; CA = cancer antigen; CEA = carcinoembryonic antigen; CI = confidence interval; EMT = epithelial–mesenchymal transition; OR = odds ratio; RC = rectal cancer; RLNM = regional lymph node metastasis; SVM = support vector machine.

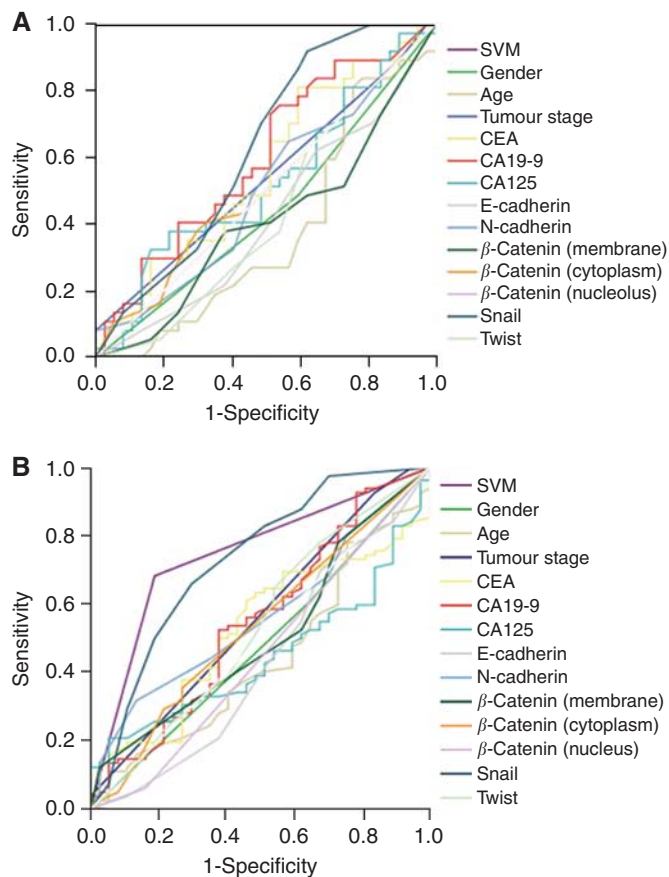


Figure 2 ROC curves plotted for positive RLNM, using EMT-related biomarkers, as well as clinicopathological, serological variables and SVM model in the training set (A) and testing set (B). At each score, the sensitivity and specificity for the RLNM status being studied were plotted, thus generating a ROC curve in the training set. The score, that closest to the point with both maximum sensitivity and specificity (0.0, 1.0), was chosen as the cutoff point for further analysis in the testing set.

Table 4 Results of multivariate logistic regression analysis in testing set

Variable	RLNM status		
	P-value	OR	95% CI
SVM	<0.0001	11.536	4.113–32.361
Tumour stage	0.041	4.443	1.064–18.557
CEA	0.913	1.055	0.403–2.762
CA19-9	0.115	2.178	0.827–5.736
CA125	0.497	1.373	0.549–3.432
β -Catenin (membrane)	0.369	1.525	0.608–3.824

Abbreviations: CA = cancer antigen; CEA = carcinoembryonic antigen; CI = confidence interval; OR = odds ratio; RLNM = regional lymph node metastasis; SVM = support vector machine.

(with RLNM vs without RLNM) for RC patients. In the present study, we applied SVM model to choose the robust markers to refine RLNM status from 13 candidate variables, including EMT-related biomarkers, as well as demographical, clinicopathological and serological biomarkers. In colorectal cancer, EMT occurred at the invasive front of tumour and acted as an important driving force for invasion and metastasis formation (Huber *et al*, 2005). Overexpression of EMT driver Twist, for example, had shown a close correlation with nodal invasion and unfavourable outcomes in patients with colorectal cancer (Valdes-Mora *et al*, 2009; Okada *et al*, 2010). Thus, we selected EMT signalling biomarkers, which

might represent the key players in the initiation and progression of RLNM for RC, to construct our SVM model. Interestingly, our univariate logistic regression analysis found that Snail, other than Twist, E-cadherin, N-cadherin and β -catenin (cytoplasm and nucleus), was the only independent predictive biomarker to RLNM status. However, we also found that, by integrating these five predictive biomarkers, which had no significant correlation with RLNM, into SVM model, the OR in refining the risk to RLNM increased two-fold when compared with Snail (OR: 9.231 vs 4.286, Table 3) alone. Taken together, our data showed that multi-markers integrated approach, other than the single one, might reflect the progression of RLNM more concisely, leading to a potential usage in tailored selection of RLNM patients to preoperative adjuvant therapy.

In colorectal cancer, gene expression signature identified 73 discriminating genes had reached to an accuracy of 88.4% in predicting the presence of RLNM (Watanabe *et al*, 2009). DNA microarray analysis showed that EMT signalling transforming growth factor β inhibitor BAMBI and β -catenin coactivator BCL9-2 might be highly expressed in RLNM patients (Watanabe *et al*, 2009). Compared with these massive gene signature-based models (Kwon *et al*, 2004; Fritzmann *et al*, 2009; Watanabe *et al*, 2009), the IHC staining was easily to be implemented and our IHC-SVM arithmetical approach might to be a useful decision-support tool in future clinical practice. By complementing with the imaging system, our SVM model raised potential clinical implications for RC patients: (i) the subset that were predicted with higher RLNM risk by our SVM model could be given the preoperative chemo- or chemoradiotherapy; (ii) the subgroup that were identified as lower RLNM risk by our SVM model should be subjected to surgery as soon as possible. Otherwise, preoperative adjuvant treatment might result in unnecessary overtreatment, lead to serious side effects and cause the patients missing the optimal opportunity for effective surgery.

Moreover, we also noticed that, compared with the 96% overall accuracy of data mining method in prediction of NSCLC prognosis and the 88.4% accuracy of gene profiling in predicting RLNM in colorectal cancer (Takahashi *et al*, 2007; Watanabe *et al*, 2009), our SVM model reached a lower (72.3%) predictive accuracy in refining RLNM status for RC patients. The underlying reason might in part due to the other potential robust variables, such as lymphovascular invasion (LVI) and perineural invasion (PNI), which denoted an increased infiltrative growth pattern (Huh *et al*, 2010), were not included in our study. Thus, LVI, PNI and EMT signalling marker integrated SVM predictive model, complemented by imaging system, might reach a higher accuracy in our ongoing prospective studies. In addition, there are limitations of this study. The absence of an independent subset of patients with round lymph nodes <5 mm in size was one limitation. Furthermore, another independent validation set with larger sample size from multicenter should be used to test the predictive value of our SVM model.

Taken together, our study showed that multi-markers, including EMT-related biomarkers and clinicopathological features, integrated SVM model owned a powerful capacity in individually predicting RLNM status in RC patients, offering a promising method to select particular patients for preoperative adjuvant therapy.

ACKNOWLEDGEMENTS

This work was supported by National Natural Science Foundation of China (No.30872488 and 81072042 to L Wang, No.81001086 to XJ Fan and No.81000934 to XB Wan).

Supplementary Information accompanies the paper on British Journal of Cancer website (<http://www.nature.com/bjc>)

REFERENCES

- Anderson B, Hardin JM, Alexander DD, Grizzle WE, Meleth S, Manne U (2010) Comparison of the predictive qualities of three prognostic models of colorectal cancer. *Front Biosci (Elite Ed)* 2: 849–856
- Baumgart E, Cohen MS, Silva Neto B, Jacobs MA, Wotkowicz C, Rieger-Christ KM, Biolo A, Zeheb R, Loda M, Libertino JA, Summerhayes IC (2007) Identification and prognostic significance of an epithelial-mesenchymal transition expression profile in human bladder tumors. *Clin Cancer Res* 13: 1685–1694
- Belovlin DI, Bates RC, Muzikansky A, Rimm DL, Mercurio AM (2005) Altered localization of p120 catenin during epithelial to mesenchymal transition of colon carcinoma is prognostic for aggressive disease. *Cancer Res* 65: 10938–10945
- Bipat S, Glas AS, Slors FJ, Zwinderman AH, Bossuyt PM, Stoker J (2004) Rectal cancer: local staging and assessment of lymph node involvement with endoluminal US, CT, and MR imaging—a meta-analysis. *Radiology* 232: 773–783
- Blanco MJ, Moreno-Bueno G, Sarrío D, Locascio A, Cano A, Palacios J, Nieto MA (2002) Correlation of Snail expression with histological grade and lymph node status in breast carcinomas. *Oncogene* 21: 3241–3246
- Brabletz T, Jung A, Reu S, Porzner M, Hlubek F, Kunz-Schughart LA, Knuechel R, Kirchner T (2001) Variable beta-catenin expression in colorectal cancers indicates tumor progression driven by the tumor environment. *Proc Natl Acad Sci USA* 98: 10356–10361
- Catto JW, Abbod MF, Linkens DA, Larre S, Rosario DJ, Hamdy FC (2009) Neurofuzzy modeling to determine recurrence risk following radical cystectomy for nonmetastatic urothelial carcinoma of the bladder. *Clin Cancer Res* 15: 3150–3155
- Dietzel M, Baltzer PA, Vag T, Groschel T, Gajda M, Camara O, Kaiser WA (2010) Application of breast MRI for prediction of lymph node metastases - systematic approach using 17 individual descriptors and a dedicated decision tree. *Acta Radiol* 51: 885–894
- Engelen SM, Beets-Tan RG, Lahaye MJ, Kessels AG, Beets GL (2008) Location of involved mesorectal and extramesorectal lymph nodes in patients with primary rectal cancer: preoperative assessment with MR imaging. *Eur J Surg Oncol* 34: 776–781
- Fang W, Fan B, Xiong B (2009) Analysis of pathological risk factors for lymph node metastasis in colorectal cancer. *Hepato-gastroenterology* 56: 663–666
- Fritzmann J, Morkel M, Besser D, Budczies J, Kosel F, Brembeck FH, Stein U, Fichtner I, Schlag PM, Birchmeier W (2009) A colorectal cancer expression profile that includes transforming growth factor beta inhibitor BAMBI predicts metastatic potential. *Gastroenterology* 137: 165–175
- Huber MA, Kraut N, Beug H (2005) Molecular requirements for epithelial-mesenchymal transition during tumor progression. *Curr Opin Cell Biol* 17: 548–558
- Huh JW, Kim HR, Kim YJ (2010) Lymphovascular or perineural invasion may predict lymph node metastasis in patients with T1 and T2 colorectal cancer. *J Gastrointest Surg* 14: 1074–1080
- Jemal A, Siegel R, Ward E, Hao Y, Xu J, Thun MJ (2009) Cancer statistics, 2009. *CA Cancer J Clin* 59: 225–249
- Kwon HC, Kim SH, Roh MS, Kim JS, Lee HS, Choi HJ, Jeong JS, Kim HJ, Hwang TH (2004) Gene expression profiling in lymph node-positive and lymph node-negative colorectal cancer. *Dis Colon Rectum* 47: 141–152
- Leonard GD, Brenner B, Kemeny NE (2005) Neoadjuvant chemotherapy before liver resection for patients with unresectable liver metastases from colorectal carcinoma. *J Clin Oncol* 23: 2038–2048
- Liang XH, Jackson S, Seaman M, Brown K, Kempkes B, Hibshoosh H, Levine B (1999) Induction of autophagy and inhibition of tumorigenesis by beclin 1. *Nature* 402: 672–676
- Low G, Tho LM, Leen E, Wiebe E, Kakumanu S, McDonald AC, Poon FW (2008) The role of imaging in the pre-operative staging and post-operative follow-up of rectal cancer. *Surgeon* 6: 222–231
- McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM (2005) Reporting recommendations for tumor marker prognostic studies (REMARK). *J Natl Cancer Inst* 97: 1180–1184
- Meng HM, Zheng P, Wang XY, Liu C, Sui HM, Wu SJ, Zhou J, Ding YQ, Li JM (2010) Overexpression of nanog predicts tumor progression and poor prognosis in colorectal cancer. *Cancer Biol Ther* 9: 295–302
- Minsky BD, Cohen AM, Kemeny N, Enker WE, Kelsen DP, Reichman B, Saltz L, Sigurdson ER, Frankel J (1992) Combined modality therapy of rectal cancer: decreased acute toxicity with the preoperative approach. *J Clin Oncol* 10: 1218–1224
- Mocellin S, Ambrosi A, Montesco MC, Foletto M, Zavagno G, Nitti D, Lise M, Rossi CR (2006) Support vector machine learning model for the prediction of sentinel node status in patients with cutaneous melanoma. *Ann Surg Oncol* 13: 1113–1122
- Modlin IM, Gustafsson BI, Drozdov I, Nadler B, Pfragner R, Kidd M (2009) Principal component analysis, hierarchical clustering, and decision tree assessment of plasma mRNA and hormone levels as an early detection strategy for small intestinal neuroendocrine (carcinoid) tumors. *Ann Surg Oncol* 16: 487–498
- Mroczo B, Groblewska M, Wereszczynska-Siemiatkowska U, Okulczyk B, Kedra B, Laszewicz W, Dabrowski A, Szmitkowski M (2007) Serum macrophage-colony stimulating factor levels in colorectal cancer patients correlate with lymph node metastasis and poor prognosis. *Clin Chim Acta* 380: 208–212
- Okada T, Suehiro Y, Ueno K, Mitomori S, Kaneko S, Nishioka M, Okayama N, Sakai K, Higaki S, Hazama S, Hirata H, Sakaida I, Oka M, Hinoda Y (2010) TWIST1 hypermethylation is observed frequently in colorectal tumors and its overexpression is associated with unfavorable outcomes in patients with colorectal cancer. *Genes Chromosomes Cancer* 49: 452–462
- Peinado H, Olmeda D, Cano A (2007) Snail, Zeb and bHLH factors in tumour progression: an alliance against the epithelial phenotype? *Nat Rev Cancer* 7: 415–428
- Rohwer N, Dame C, Haugstetter A, Wiedenmann B, Detjen K, Schmitt CA, Cramer T (2010) Hypoxia-inducible factor 1alpha determines gastric cancer chemosensitivity via modulation of p53 and NF-kappaB. *PLoS One* 5: e12038
- Sattlecker M, Bessant C, Smith J, Stone N (2010) Investigation of support vector machines and Raman spectroscopy for lymph node diagnostics. *Analyst* 135: 895–901
- Sebag-Montefiore D, Stephens RJ, Steele R, Monson J, Grieve R, Khanna S, Quirke P, Couture J, de Metz C, Myint AS, Bessell E, Griffiths G, Thompson LC, Parmar M (2009) Preoperative radiotherapy versus selective postoperative chemoradiotherapy in patients with rectal cancer (MRC CR07 and NCIC-CTG C016): a multicentre, randomised trial. *Lancet* 373: 811–820
- Stojadinovic A, Eberhardt C, Henry L, Eberhardt J, Elster EA, Peoples GE, Nissan A, Shriver CD (2010) Development of a Bayesian classifier for breast cancer risk stratification: a feasibility study. *Eplasty* 10: e25
- Takahashi Y, Coppola D, Matsushita N, Cuaing HD, Sun M, Sato Y, Liang C, Jung JU, Cheng JQ, Mule JJ, Pledger WJ, Wang HG (2007) Bif-1 interacts with Beclin 1 through UVRA9 and regulates autophagy and tumorigenesis. *Nat Cell Biol* 9: 1142–1151
- Thiery JP (2002) Epithelial-mesenchymal transitions in tumour progression. *Nat Rev Cancer* 2: 442–454
- Valdes-Mora F, Gomez del Pulgar T, Bandres E, Cejas P, Ramirez de Molina A, Perez-Palacios R, Gallego-Ortega D, Garcia-Cabezas MA, Casado E, Larrauri J, Nistal M, Gonzalez-Baron M, Garcia-Foncillas J, Lacal JC (2009) TWIST1 overexpression is associated with nodal invasion and male sex in primary colorectal cancer. *Ann Surg Oncol* 16: 78–87
- Vermeulen J, De Preter K, Naranjo A, Vercruyse L, Van Roy N, Hellemans J, Swerts K, Bravo S, Scaruffi P, Tonini GP, De Bernardi B, Noguera R, Piqueras M, Canete A, Castel V, Janoueix-Lerosey I, Delattre O, Schleiermacher G, Michon J, Combaret V, Fischer M, Oberthuer A, Ambros PF, Beiske K, Benard J, Marques B, Rubie H, Kohler J, Potschger U, Ladenstein R, Hogarty MD, McGrady P, London WB, Laureys G, Speleman F, Vandesompele J (2009) Predicting outcomes for children with neuroblastoma using a multigene-expression signature: a retrospective SIOPEX/COG/GPOH study. *Lancet Oncol* 10: 663–671
- Wan XB, Fan XJ, Chen MY, Xiang J, Huang PY, Guo L, Wu XY, Xu J, Long ZJ, Zhao Y, Zhou WH, Mai HQ, Liu Q, Hong MH (2010) Elevated Beclin 1 expression is correlated with HIF-1alpha in predicting poor prognosis of nasopharyngeal carcinoma. *Autophagy* 6: 395–404
- Watanabe T, Kobunai T, Tanaka T, Ishihara S, Matsuda K, Nagawa H (2009) Gene expression signature and the prediction of lymph node metastasis in colorectal cancer by DNA microarray. *Dis Colon Rectum* 52: 1941–1948
- Zhu ZH, Sun JB, Ma Y, Shao JY, Long H, Zhang X, Fu JH, Zhang LJ, Su XD, Wu QL, Ling P, Chen M, Xie ZM, Hu Y, Rong TH (2009) Three immunomarker support vector machines-based prognostic classifiers for stage IB non-small-cell lung cancer. *J Clin Oncol* 27: 1091–1099

This work is published under the standard license to publish agreement. After 12 months the work will become freely available and the license terms will switch to a Creative Commons Attribution-NonCommercial-Share Alike 3.0 Unported License.