Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Identification of the Most Specific Markers to Differentiate Primary Pulmonary Carcinoma From Metastatic Gastrointestinal Carcinoma to the Lung.

Bachar Alabdullah (bacharalabdullah@gmail.com)
NSW Health: New South Wales Ministry of Health https://orcid.org/0000-0002-3372-1694
Amir Hadji-Ashrafy
NSW Health: New South Wales Ministry of Health

Research

Keywords: Cancer, Lung, Gastrointestinal, Napsin-A, CK7, CK20, CDX2, TTF1, SATB2, Biomarkers

Posted Date: July 20th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-728801/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published at Diagnostic Pathology on January 14th, 2022. See the published version at https://doi.org/10.1186/s13000-021-01184-2.

Identification of the most Specific Markers to Differentiate Primary Pulmonary Carcinoma from Metastatic Gastrointestinal Carcinoma to the Lung.

Authors:

Bachar Alabdullah, Honours Bachelor of Science and Doctor of Medicine (MD)[1]

University of Sydney, Nepean Hospital

Department of Anatomical Pathology

C: +61 434 263 379

Department of Anatomical Pathology, Nepean Hospital

Derby Street, Kingswood

NSW, Australia, 2747

E: **bacharalabdullah@gmail.com**

Amir Hadji-Ashrafy, Bachelor of Medicine and Bachelor of Surgery (MBBS)

University of Sydney, Nepean Hospital

Department of Anatomical Pathology

C: +61 402 531 693

Department of Anatomical Pathology, Nepean Hospital

Derby Street, Kingswood

NSW, Australia, 2747

E: **amir.hadjiashrafy@health.nsw.gov.au**

[1] Bachar Alabdullah is currently employed at Tamworth Rural Referral Hospital as a Resident Medical Officer and is the corresponding author.

Abstract

Background: A number of biomarkers have the potential of differentiating between primary lung tumours and secondary lung tumours from the gastrointestinal tract, however, a standardised panel for that purpose does not exist yet. We aimed to identify the smallest panel that is most sensitive and specific at differentiating between primary lung tumours and secondary lung tumours from the gastrointestinal tract.

Methods: A total of 170 samples were collected, including 140 primary and 30 non-primary lung tumours and staining for CK7, Napsin-A, TTF1, CK20, CDX2, and SATB2 was performed via tissue microarray. The data was then analysed using univariate regression models and a combination of multivariate regression models and Receiver Operating Characteristic (ROC) curves.

Results: Univariate regression models confirmed the 6 biomarkers' ability to independently predict the primary outcome ($p<0.001$). Multivariate models of 2-biomarker combinations identified 11 combinations with statistically significant ORs ($p<0.05$), of which TTF1/CDX2 had the highest area under the curve (AUC) (0.983, 0.960-1.000 95% CI). The sensitivity, specificity, PPV, and NPV were 75.7%, 100%, 100%, and 37.5% respectively. Multivariate models of 3-biomarker combinations identified 4 combinations with statistically significant ORs ($p<0.05$), of which CK7/CK20/SATB2 had the highest AUC (0.965, 0.930-1.000 95% CI). The sensitivity, specificity, PPV, and NPV were 85.1%, 100%, 100%, and 41.7% respectively. Multivariate models of 4-biomarker combinations did not identify any combinations with statistically significant ORs ($p<0.05$).

Conclusions: The analysis identified the combination of CK7/CK20/SATB2 to be the smallest panel with the highest sensitivity (85.1%) and specificity (100%) for predicting tumour origin with an ROC AUC of 0.965 ($p<0.001$; SE: 0.018, 0.930-1.000 95% CI).

Keywords

Cancer. Lung. Gastrointestinal. Napsin-A. CK7. CK20. CDX2. TTF1. SATB2. Biomarkers.

Background

Lung cancer is the second most prevalent cancer in both men and women [1] and remains the leading cause of cancer-related deaths in both at 55.9 and 36.6 deaths per 100,000 respectively [1,2]. The 2015 WHO Classification of Lung Tumors divides lung tumours into 6 main types and 77 subtypes based on histological appearance. The 6 main types are epithelial, neuroendocrine, mesenchymal, lymphohistocytic, tumours of ectopic origin, and metastatic tumours, while the most important subtypes include adenocarcinoma and squamous, small, and large-cell carcinomas [3]. Another method of classification looks at the neoplastic cells' site of origin and classifies tumours as either primary (arising directly from the lungs), or secondary (metastasising to the lung from a distant site). The combination of the histological pattern and site of origin offers clinicians insight into the tumour's staging, prognosis, and management options. Ensuring accurate identification and classification is of paramount importance in the ever growing age of immunotherapy, offering patients hope instead of the current grim outlook. Identification is traditionally done by correlating clinical evidence with radiological and pathological findings. Accurate classification falls on the shoulders of the anatomical pathologist and depends on a number of factors including the quality of the biopsy, the experience of the pathologist, and the extent of tumour differentiation. Well differentiated tumours have a clear histopathological pattern and are easy to classify. Poorly differentiated tumours on the other hand do not have a clear histopathological pattern and their classification was traditionally highly dependent on the pathologist's level of expertise. This introduced inconsistencies in diagnosis, classification, and ultimately patient management. One emerging tool to assist with accurate identification and classification is the detection of specific biological markers (biomarkers) in tumour samples using immunohistochemical (IHC) staining techniques. Biomarkers are defined as "any substance, structure, or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease" [4]. Some biomarkers are preferentially expressed in certain types of tissue but not in others, thus offering a method of objectively identifying the histological subtype of a tumour even if it exhibits poor differentiation.

Since their introduction, the use of biomarkers has become part of the routine diagnostic workup. Biomarkers traditionally used for the diagnosis of lung tumours include TTF1, Napsin-A, and CK7. Diagnosis of gastrointestinal tract (GIT) tumours requires a different set of biomarkers, the most useful of which include CK20, CDX2, and most recently SATB2. Despite recommendations for their use in aiding diagnosis, which biomarkers are used highly depends on the individual practices of institutions. TTF1 [5-7] and Napsin-A [5,7] exhibit high sensitivity and specificity for well differentiated primary lung adenocarcinomas, but there are a number of reported cases of less differentiated samples that exhibit little to no staining with these two markers [7]. Moreover, there have been occasional cases of colonic adenocarcinomas that show strong and diffuse staining with TTF1 [7]. Similarly, there are reported cases of CK7 expression in primary GIT tumours [8]. CK20 is highly sensitive for the detection of colorectal carcinoma but has a low specificity [9]. CDX2 is preferentially expressed in the intestinal epithelium, thus is very sensitive for detecting colorectal adenocarcinoma [10,11], but there have been reported cases of primary lung adenocarcinomas that express CDX2 [10,11]. SATB2 has recently been shown to be highly sensitive for colorectal adenocarcinomas with sensitives that range between 80-97% [7,9,12-16]. Similar to its counterparts, SATB2 may still show low level expression in primary lung adenocarcinomas [12]. With all the shortcomings of the biomarkers currently used in practice, SATB2 has the potential of adding value in differentiating between primary and secondary lung tumours. Additionally, with the overabundance of markers that can be used for that purpose,

there are no studies to date that attempt to identify the smallest, most optimal panel for differentiating primary and secondary lung tumours; this study attempts to do address that gap.

The study aims to investigate the differential expression of the biomarkers CK7, Napsin-A, TTF1, CK20, CDX2, and SATB2 in samples of primary lung, secondary lung, and primary GIT tumours to identify the smallest panel with the highest sensitivity and specificity for differentiating between primary lung tumours and metastatic GIT tumours to the lungs. Our main hypothesis states that the expression of CK7, Napsin-A, TTF1, CK20, CDX2, and SATB2 in lung tumours is dependent on the tumour's site of origin. If true, we speculate that the use of these biomarkers will be significant in differentiating between primary and secondary lung tumours, with CK7, Napsin-A, and TTF1 being positive if primary, and CK20, CDX2, and SATB2 being positive if secondary. We also predict that a panel consisting of Napsin-A, TTF1, and SATB2 will be the most sensitive and specific at differentiating between primary and secondary lung tumours.

Methods

A low-and-negligible risk (LNR) ethics application was prepared and approval from the Nepean and Blue Mountains Local Health District Human Research Ethics Committee was received on November 27th, 2017. Subsequently, a search was performed using a central directory of tumour biopsies and resections done at Nepean Hospital (NSW, Australia). The search looked at all lung resections and biopsies, and all GIT tumour resections done in the last 60 months. The search yielded 200 lung resections, 199 core lung biopsies, 7 wedge lung biopsies, and 29 colon tumour resections. Samples to be included in the study needed to be of a pattern common to both lung and GIT tumours (such as adenocarcinomas) to allow meaningful investigation of their origin. Hence, primary squamous and small-cell lung carcinomas were excluded. Secondary tumours not originating from the GIT (e.g. metastatic pancreatic adenocarcinomas) were also excluded as their biomarker staining pattern is different to that of GIT tumours. Subsequently, pathology reports were inspected and the case number, medical record number (MRN), name, age, sex, reported diagnosis, site of origin, specimen type, tumour block numbers, and previously done biomarker stain results were collected. The samples were then filtered and any duplicates (patients with more than one biopsy from the same site) were removed. The initial working set of 179 samples was then established and included 146 primary lung, 10 secondary lung, 21 primary GIT, and 2 undifferentiated tumours. The 2 undifferentiated tumours were excluded from the analysis as they did not have an official classification. Of the 177 samples identified, 160 paraffin blocks were retrieved from storage and 17 were not found. These 17 were still included as their pathology reports showed previous biomarker stains that can be included in the analysis.

Each sample was then assigned a random number to facilitate blinding when interpreting results. This was done in excel by creating a column with a list of numbers from 1 to 177. The corresponding cells in the adjacent column were filled with randomly generated numbers using excel's RAND function. Both columns were subsequently selected and the column containing the random numbers was sorted in ascending order, causing the first column to randomise, giving each sample a unique ID.

Staining of biomarkers was done using the tissue microarray method. This method creates individual slides with multiple, small tumour samples as opposed to the traditional method of having 1 tumour sample per slide. This allows for more rapid and cost efficient staining but

increases the risk of unsuccessful stains. Of the 160 blocks retrieved, 34 did not have sufficient volume to undergo tissue microarray. The 128 blocks left were melted and the appropriate number and volume of tissue was retrieved and re-paraffinised into new biomarker-specific blocks. Each new block included a control that is known to stain positive for the respective biomarker. In total, 13 CK7, 12 Napsin-A, 8 TTF1, 15 CK20, 15 CDX2, and 15 SATB2 blocks were created. The new blocks were then used to make immunoperoxidase (IPX)-compatible slides for IHC staining using a Ventana machine. Following staining, the slides were interpreted independently by the chief and second investigators, and the results were recorded as either positive, negative, or unsuccessful. Any conflicts in interpretation were settled by consensus. Finally, the results of staining were added to the initial data set and samples with all 6 biomarkers missing were removed, leaving a total of 170 samples.

Logically, to obtain the smallest, most sensitive and specific panel of biomarkers for differentiating between primary and non-primary tumours (dichotomous outcome), binary logistic regression models that use 'tumour origin' as the dependent variable and the biomarkers as the covariates need to be created, starting with univariate models and moving on to multivariate models of 2 to 6 covariates. Firstly, the data was coded into SPSS (tumour origin: 0=non-primary, 1=primary; stains: 0=negative, 1=positive, 2=missing). For our analysis, 'non-primary' included both secondary lung tumours and primary GIT tumours as both have the same staining pattern. The missing data points were computed into SPSS as to be excluded from the analysis. A univariate binary logistic regression analysis was performed with each of the 6 biomarkers as covariates [Table 1]. Subsequently, a multivariate binary logistic regression analysis was performed using 15 different 2-biomarker combinations [Table 2], 20 different 3-biomarker combinations [Table 4], and 15 different 4-biomarker combinations, and the predicted probability of each outcome was recorded as a new variable. The combinations with non-significant odds ratios (ORs) ($p \geq 0.05$) were excluded from further analysis. Receiver Operating Characteristic (ROC) curves of the statistically significant ($p < 0.05$) combinations were constructed using the outcome's 'predicted probability' as the test variable and the outcome (tumour origin) as the state variable. The area under the curve (AUC) was then used to compare models' ability to predict the outcome [Tables 3 and 5].

Results

Positive controls were observed on all slides indicating successful overall staining using the Ventana machine. However, staining of some individual samples within each slide was unsuccessful. Table 6 summarises the proportion of successful stains for each biomarker.

Of the 143 samples stained for CK7, 119 (83.2%) were primary and 24 (16.8%) were non-primary. Of the 119 primary tumours, 111 (93.3%) stained correctly as positive and 8 (6.7%) stained incorrectly as negative. Of the 24 non-primary tumours, 14 (58.3%) stained correctly as negative and 10 (41.7%) stained incorrectly as positive [Table 7].

Of the 146 samples stained for Napsin-A, 125 (85.6%) were primary and 21 (14.4%) were non-primary. Of the 125 primary tumours, 95 (76.0%) stained correctly as positive and 30 (24.0%) stained incorrectly as negative. Of the 21 non-primary tumours, 19 (90.5%) stained correctly as negative and 2 (9.5%) stained incorrectly as positive [Table 8].

Of the 166 samples stained for TTF1, 139 (83.7%) were primary and 27 (16.3%) were non-primary. Of the 139 primary tumours, 113 (81.3%) stained correctly as positive and 26 (18.7%)

stained incorrectly as negative. Of the 27 non-primary tumours, 25 (92.6%) stained correctly as negative and 2 (7.4%) stained incorrectly as positive [Table 9].

Of the 123 samples stained for CK20, 109 (88.6%) were primary and 14 (11.4%) were non-primary. Of the 109 primary tumours, 105 (96.3%) stained correctly as negative and 4 (3.7%) stained incorrectly as positive. Of the 14 non-primary tumours, 8 (57.1%) stained correctly as positive and 6 (42.9%) stained incorrectly as negative [Table 10].

Of the 121 samples stained for CDX2, 104 (86.0%) were primary and 17 (14.0%) were non-primary. Of the 104 primary tumours, 100 (96.2%) stained correctly as negative and 4 (3.8%) stained incorrectly as positive. Of the 17 non-primary tumours, 16 (94.1%) stained correctly as positive and 1 (5.9%) stained incorrectly as negative [Table 11].

Of the 119 samples stained for SATB2, 99 (83.2%) were primary and 20 (16.8%) were non-primary. Of the 119 primary tumours, 90 (90.9%) stained correctly as negative and 9 (9.1%) stained incorrectly as positive. Of the 20 non-primary tumours, 15 (75.0%) stained correctly as positive and 5 (25.0%) stained incorrectly as negative [Table 12].

Univariate binary logistic regression models showed a significant OR ($p<0.05$) of 19.4, 30.1, 54.3, 0.029, 0.003, and 0.033 for CK7, Napsin-A, TTF1, CK20, CDX2, and SATB2 respectively [Table 1]. Table 13 summarises the sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of each marker.

Multivariate analysis was performed for each of the 15 2-biomarker combinations [Table 2]. Of the 15, 11 had statistically significant ORs ($p<0.05$). ROC curves of the 11 combinations showed all 11 to have statistically significant ($p<0.05$) AUC, with the combination of TTF1/CDX2 having the highest (0.983, 0.960-1.000 95% CI) [Table 3]. The sensitivity, specificity, PPV, and NPV of the TTF1/CDX2 panel were 75.7%, 100%, 100%, and 37.5% respectively.

Multivariate analysis was performed for each of the 20 3-biomarker combinations [Table 4]. Of the 20, only 4 had statistically significant ORs ($p<0.05$). ROC curves of the 4 combinations showed all 4 to have statistically significant ($p<0.05$) AUC, with the combination of CK7/CK20/SATB2 having the highest (0.965, 0.930-1.000 95% CI) [Table 5]. The sensitivity, specificity, PPV, and NPV of the CK7/CK20/SATB2 panel were 85.1%, 100%, 100%, and 41.7% respectively.

Multivariate analysis was performed for each of the 15 4-biomarker combinations, however, none of the combinations had statistically significant ORs ($p<0.05$).

Discussion

The results obtained from our investigation supports our main hypothesis of the dependency of biomarker expression on the tumour's site of origin. Univariate analysis clearly showed that all 6 biomarkers were statistically significant at predicting the outcome [Table 1]. CK7, Napsin-A, and TTF1 all had ORs >1, signifying that they are highly predictive of the outcome coded as 1 (primary origin). Of these 3, TTF1 had the largest OR at 54.3, which means that if a given lung cancer sample is positive for TTF1, the odds of it being primary are 54.3 times more than the odds of it being non-primary. CK20, CDX2, and SATB2 had ORs <1, signifying that they are highly predictive of the outcome coded as 0 (non-primary origin). Of these 3, CDX2 had the smallest OR for outcome 1 at 0.003, thus reciprocally having the highest OR for outcome 0 at 333.3. This means that if a given lung cancer sample is positive for CDX2, the odds of it being non-primary are 333.3 times more than the odds of it being primary. With such high predictive

abilities, the results appear promising at first, however, the flaws of each biomarker are appreciated upon inspection of the sensitivity, specificity, PPV, and NPV of each [Table 13]. For example, CK7 has a high sensitivity (93.3%) but a low specificity (58.3%). The high sensitivity indicates a low probability of a false negative result, hence a negative stain is likely to represent a true negative and can rule out primary origin. The low specificity signifies a high false positive rate, hence a positive stain is very likely to be false, hence cannot rule in primary origin. Napsin-A and TTF1 have the opposite problem of having a high specificity but a low sensitivity, thus are good at ruling in primary origin when positive but cannot rule out primary origin if negative. Similarly, CK20 and SATB2 have a high specificity but low sensitivity, thus are good at ruling in non-primary origin when positive (i.e. rule out primary origin), but cannot rule out non-primary origin (i.e. cannot rule in primary origin) when negative. CDX2 is the only biomarker that appears to have both a high sensitivity and specificity, however, it is also the marker associated with the highest standard error (SE) for its OR (reciprocal OR: 333.3 , SE 1.150, 4.2-3816.8 95% CI).

Unlike univariate models, multivariate models have at least 2 ORs (1 per covariate), thus their ability to predict the outcome cannot be judged just by comparing the values of their ORs. Consequently, ROC curves were utilised for that purpose by plotting the model's sensitivity on the y-axis and 1-specficity on the x-axis then computing the AUC. A model with a high sensitivity and a high specificity (i.e. low 1-specificity) will have an ROC curve that hugs the upper left corner, hence maximising the AUC. Of all the statistically significant multivariate models [Tables 3 and 5], the combinations of TTF1/CDX2 and CK7/CK20/SATB2 had the highest AUC (0.983 and 0.965) amongst their peers. This is contrary to our initial prediction of TTF1/Napsin-A/SATB2 being the optimal panel. The TTF1/CDX2 panel had a moderate sensitivity of 75.7%, very high specificity of 100%, and a PPV and NPV of 100% and 37.5% respectively when the positive test is defined as TTF1+/CDX2-. When compared to TTF1 or CDX2 individually, despite a slight improvement in specificity and PPV, the sensitivity and NPV decrease when compared to both (more drastically when compared to CDX2). Based on that, it is difficult to recommend the TTF1/CDX2 panel over either of its constituents individually. The CK7/CK20/SATB2 panel improves on TTF1/CDX2 slightly with 85.1% sensitivity, 100% specificity, and a PPV and NPV of 100% and 41.7% respectively when a positive test is defined as CK7+/CK20-/SATB2-. When compared to its individual constituents, it appears that the 3-biomarker panel overcomes some of the individual flaws discussed earlier. The reasonably high sensitivity overcomes the low to moderate sensitivities of CK20 and SATB2, and the high specificity overcomes the low specificity of CK7. The NPV of the panel, however, is much lower than that of its constituents, but despite that, it still appears that the CK7/CK20/SATB2 panel is superior to both the individual constituents as well as the TTF1/CDX2 panel, hence can be recommended over them. None of the 4-biomarker combinations were statistically significant, hence analysis stopped at 3-biomakers.

Conclusion

The data appears to confirm that panels with more than 3 biomarkers do not offer any additional value in differentiating between primary and non-primary tumour origin. Specifically, a panel made up of CK7, CK20, and SATB2 was shown to be the most sensitive and specific while overcoming the shortcomings of its individual constituents. We believe standardising testing to only include CK7, CK20, and SATB2 will not only provide the best quality of evidence in aiding

diagnosis, but will also contribute to significant reduction in sample processing times and associated cost, hence should be adapted in all laboratories.

Declarations

Ethics Approval and Consent to Participate

The study was performed in accordance with the Declaration of Helsinki. An ethics approval was received on Monday 27th of November 2017 from the Nepean Blue Mountains Local Health District Human Research Ethics Committee, NSW, Australia (Project Number LNR/17/NEPEAN/122).

Consent for Publication

All contributing authors are aware of the study submission and consent for publication.

Availability of Data and Materials

The data used and analysed are available from the corresponding author on request via email at bacharalabdullah@gmail.com.

Competing Interests

Both authors have no competing interests, financial or otherwise, relating to this study.

Funding

Funding was provided by research-specific departmental funding (Anatomical Pathology Trust Fund) provided by the Anatomical Pathology Department at Nepean Hospital, NSW, Australia.

Author Contributions

B.A. and A.H. designed the study concept and protocol. A.H. identified study subjects. B.A. acquired ethics approval and performed data collection. B.A. and A.H. analysed the results and drew appropriate conclusions. B.A. performed statistical analysis and wrote the manuscript.

Acknowledgements

Department of Anatomical Pathology, Nepean Hospital, NSW, Australia; University of Sydney Nepean Clinical School, NSW, Australia.

References

1. Key Statistics for Lung Cancer. [Web article]. 2016; **https://www.cancer.org/cancer/non-small-cell-lung-cancer/about/key-statistics.html**. Accessed July 17, 2017, 2017.

2. Jemal A, Ward EM, Johnson CJ, et al. Annual Report to the Nation on the Status of Cancer, 1975–2014, Featuring Survival. *JNCI: Journal of the National Cancer Institute.* 2017;109(9):djx030-djx030.

3. Travis WD, Brambilla E, Nicholson AG, et al. The 2015 World Health Organization Classification of Lung Tumors. *Journal of Thoracic Oncology.* 2015;10(9):1243-1260.

4. WHO International Programme on Chemical Safety Biomarkers in Risk Assessment: Validity and Validation. 2001; **http://www.inchem.org/documents/ehc/ehc/ehc222.htm**. Accessed July 17, 2017.

5. Boggaram V. Thyroid transcription factor-1 (TTF-1/Nkx2.1/TITF1) gene regulation in the lung. *Clin Sci (Colch).* 2009;116(1):27-35.

6. Li HC, Schmidt L, Greenson JK, Chang AC, Myers JL. Primary pulmonary adenocarcinoma with intestinal differentiation mimicking metastatic colorectal carcinoma: case report and review of literature. *Am J Clin Pathol.* 2009;131(1):129-133.

7. Ye J, Findeis-Hosey JJ, Yang Q, et al. Combination of napsin A and TTF-1 immunohistochemistry helps in differentiating primary lung adenocarcinoma from metastatic carcinoma in the lung. *Appl Immunohistochem Molecul Morphol.* 2011;19(4):313-317.

8. Chu P, Wu E, Weiss LM. Cytokeratin 7 and cytokeratin 20 expression in epithelial neoplasms: a survey of 435 cases. *Mod Pathol.* 2000;13(9):962-972.

9. Magnusson K, de Wit M, Brennan DJ, et al. SATB2 in combination with cytokeratin 20 identifies over 95% of all colorectal carcinomas. *Am J Surg Pathol.* 2011;35(7):937-948.

10. Yatabe Y, Koga T, Mitsudomi T, Takahashi T. CK20 expression, CDX2 expression, K-ras mutation, and goblet cell morphology in a subset of lung adenocarcinomas. *J Pathol.* 2004;203(2):645-652.

11. Yousem SA. Pulmonary intestinal-type adenocarcinoma does not show enteric differentiation by immunohistochemical study. *Mod Pathol.* 2005;18(6):816-821.

12. Conner JR, Hornick JL. Metastatic carcinoma of unknown primary: diagnostic approach using immunohistochemistry. *Adv Anat Pathol.* 2015;22(3):149-167.

13. Dragomir A, de Wit M, Johansson C, Uhlen M, Ponten F. The role of SATB2 as a diagnostic marker for tumors of colorectal origin: Results of a pathology-based clinical prospective study. *Am J Clin Pathol.* 2014;141(5):630-638.

14. Lin F, Shi J, Zhu S, et al. Cadherin-17 and SATB2 are sensitive and specific immunomarkers for medullary carcinoma of the large intestine. *Arch Pathol Lab Med.* 2014;138(8):1015-1026.

15. Ordonez NG. SATB2 is a novel marker of osteoblastic differentiation and colorectal adenocarcinoma. *Adv Anat Pathol.* 2014;21(1):63-67.

16. Ordonez NG. Cadherin 17 is a novel diagnostic marker for adenocarcinomas of the digestive system. *Adv Anat Pathol.* 2014;21(2):131-137.

Table 1: Outcomes of 6 different univariate binary logistic regression models for CK7, Napsin-A, TTF1, CK20, CDX2, and SATB2 showing the odds ratio (OR), p-values, 95% confidence intervals (95% CI), and standard error (SE). The OR here signifies the odds of belonging to the 'primary lung tumour' group of the 'tumour origin' outcome.

#	Marker	OR	p-value	95% CI Lower Limit	95% CI Upper Limit	SE
1	CK7	19.425	<0.001	6.576	57.384	0.553
2	Napsin-A	30.083	<0.001	6.621	136.689	0.772
3	TTF1	54.327	<0.001	12.097	243.974	0.766
4	CK20	0.029	<0.001	0.007	0.122	0.742
5	CDX2	0.003	<0.001	0.000	0.024	1.150
6	SATB2	0.033	<0.001	0.010	0.113	0.624

Table 2: Outcomes of binary logistic regression models of 15 different 2-biomarker combinations showing the odds ratio (OR), p-values, 95% confidence intervals (95% CI), and standard error (SE) of each biomarker when the other in the combination is controlled. The OR here signifies the odds of belonging to the 'primary lung tumour' group of the 'tumour origin' outcome.

#	Markers	OR	p-value	95% CI Lower Limit	95% CI Upper Limit	SE
1	CK7	1.6162×10^{1}	<0.001	3.891	66.822	0.725
	Napsin-A	1.5424×10^{1}	0.001	3.023	78.681	0.831
2	CK7	6.8660	0.005	1.813	26.000	0.679
	TTF1	2.6690×10^{1}	<0.001	5.572	127.845	0.799
3	CK7	8.4970	0.012	1.600	45.135	0.825
	CK20	3.5000×10^{-2}	<0.001	0.007	0.172	0.817
4	CK7	3.0468×10^{8}	0.996	0.000	.	4076.604
	CDX2	2.7351×10^{-10}	0.996	0.000	.	4076.604
5	CK7	1.8114×10^{1}	<0.001	3.983	82.388	0.773
	SATB2	7.9000×10^{-2}	0.001	0.019	0.335	0.734
6	Napsin-A	3.7190	0.133	0.670	20.630	0.874
	TTF1	6.1509×10^{8}	0.996	0.000	.	3920.509
7	Napsin-A	9.2810	0.016	1.522	56.608	0.923
	CK20	5.1000×10^{-2}	0.006	0.006	0.419	1.078
8	Napsin-A	1.6645×10^{1}	0.030	1.307	212.0	1.298
	CDX2	4.0000×10^{-3}	<0.001	0.000	0.055	1.370
9	Napsin-A	2.5106×10^{1}	0.001	3.498	180.201	1.006
	SATB2	2.3000×10^{-2}	<0.001	0.004	0.143	0.931
10	TTF1	3.5732×10^{1}	0.002	3.911	326.459	1.129
	CK20	1.7000×10^{-2}	<0.001	0.002	0.162	1.162
11	TTF1	2.3587×10^{1}	0.015	1.853	300.223	1.298
	CDX2	3.0000×10^{-3}	<0.001	0.000	0.043	1.373

Table 3: Outcomes of binary logistic regression models of 15 different 2-biomarker combinations showing the odds ratio (OR), p-values, 95% confidence intervals (95% CI), and standard error (SE) of each biomarker when the other in the combination is controlled. The OR here signifies the odds of belonging to the 'primary lung tumour' group of the 'tumour origin' outcome (cont.)

#	Markers	OR	p-value	95% CI Lower Limit	95% CI Upper Limit	SE
12	TTF1	3.7699×10^{1}	<0.001	5.695	249.542	0.964
	SATB2	3.4000×10^{-2}	<0.001	0.006	0.186	0.869
13	CK20	2.3810×10^{-1}	0.268	0.019	3.011	1.295
	CDX2	3.5179×10^{-10}	0.996	0.000	.	4003.691
14	CK20	7.0000×10^{-3}	<0.001	0.001	0.089	1.316
	SATB2	6.4000×10^{-2}	0.004	0.010	0.417	0.958
15	CDX2	2.0760×10^{-10}	0.996	0.000	.	4061.585
	SATB2	4.6710×10^{-9}	0.996	0.000	.	4061.585

Table 4: Comparison of the area under the curve (AUC), p-values, standard error (SE), and 95% confidence interval (95% CI) generated from the Receiver Operating Characteristic (ROC) curves for each of the 11 statistically significant 2-biomarker combinations.

#	Markers	ROC AUC	p-value	95% CI LB	95% CI UB	SE
1	TTF1 / CDX2	0.983	<0.001	0.960	1.000	0.012
2	Napsin-A / CDX2	0.978	<0.001	0.945	1.000	0.017
3	Napsin-A / SATB2	0.943	<0.001	0.892	0.993	0.026
4	TTF1 / CK20	0.935	<0.001	0.886	0.984	0.025
5	TTF1 / SATB2	0.912	<0.001	0.813	1.000	0.051
6	CK7 / TTF1	0.893	<0.001	0.808	0.978	0.043
7	CK7 / SATB2	0.883	<0.001	0.774	0.991	0.055
8	CK20 / SATB2	0.883	<0.001	0.743	1.000	0.072
9	CK7 / Napsin-A	0.877	<0.001	0.777	0.978	0.051
10	CK7 / CK20	0.848	<0.001	0.708	0.988	0.071
11	Napsin-A / CK20	0.785	<0.001	0.590	0.979	0.099

Table 5: Outcomes of binary logistic regression models of 20 different 3-biomarker combinations showing the odds ratio (OR), p-values, 95% confidence intervals (95% CI), and standard error (SE) of each biomarker when the others in the combination are controlled. The OR here signifies the odds of belonging to the 'primary lung tumour' group of the 'tumour origin' outcome.

#	Markers	OR	p-value	95% CI Lower Limit	95% CI Upper Limit	SE
1	TTF1	4.8484×10^{15}	0.994	0.000	.	5178.889287
	CK20	1.0856×10^{-17}	0.994	0.000	.	5178.889485
	SATB2	2.5186×10^{-9}	0.996	0.000	.	3732.495899
2	TTF1	6.6786×10^{7}	0.994	0.000	.	2311.469426
	CK20	3.7433×10^{-8}	0.994	0.000	.	2311.469465
	CDX2	9.0702×10^{-16}	0.993	0.000	.	3687.397976
3	TTF1	1.0066×10^{15}	0.994	0.000	.	4604.708716
	CDX2	4.7399×10^{-24}	0.992	0.000	.	5677.505089
	SATB2	4.2070×10^{-16}	0.994	0.000	.	4729.925235
4	Napsin-A	9.5063×10^{6}	0.994	0.000	.	2330.698567
	TTF1	4.2801×10^{21}	0.992	0.000	.	5162.613603
	SATB2	1.2550×10^{-15}	0.993	0.000	.	3988.728307
5	Napsin-A	7.5000×10^{-1}	0.799	0.101	5.576	1.023533
	TTF1	2.2365×10^{15}	0.995	0.000	.	5225.305260
	CK20	5.2321×10^{-9}	0.996	0.000	.	3694.849146
6	Napsin-A	1.4559	0.823	0.054	39.478	1.683770
	TTF1	3.7603×10^{8}	0.996	0.000	.	4019.542737
	CDX2	5.7080×10^{-3}	0.000447	0.000	0.102	1.471546
7	Napsin-A	6.6000	0.078	0.810	55.247	1.077199
	CK20	0.0236	0.026	0.001	0.637	1.682107
	SATB2	0.0716	0.020	0.008	0.664	1.136644

Table 6: Outcomes of binary logistic regression models of 20 different 3-biomarker combinations showing the odds ratio (OR), p-values, 95% confidence intervals (95% CI), and standard error (SE) of each biomarker when the others in the combination are controlled. The OR here signifies the odds of belonging to the 'primary lung tumour' group of the 'tumour origin' outcome (cont.)

#	Markers	OR	p-value	95% CI Lower Limit	95% CI Upper Limit	SE
8	Napsin-A	7.0029	0.189	0.384	127.805	1.481755
	CK20	0.7798	0.875	0.036	17.114	1.575845
	CDX2	4.3559×10^{-10}	0.996	0.000	.	3931.584378
9	Napsin-A	3.4919×10^{14}	0.993	0.000	.	4031.440053
	CDX2	8.186×10^{-23}	0.992	0.000	.	4999.842575
	SATB2	5.3664×10^{-15}	0.993	0.000	.	3751.338776
10	CK7	8.2054	0.027	1.271	52.988	0.951690
	TTF1	2.0773×10^{1}	0.002	3.075	140.336	0.974708
	SATB2	7.0791×10^{-2}	0.005	0.011	0.442	0.934354
11	CK7	3.6169	0.201	0.505	25.916	1.004739
	TTF1	2.5458×10^{1}	0.004	2.754	235.319	1.134664
	CK20	1.8967×10^{-2}	0.001	0.002	0.201	1.203378
12	CK7	1.5920×10^{8}	0.996	0.000	.	3900.679912
	TTF1	1.1861×10^{1}	0.063	0.878	160.214	1.328229
	CDX2	3.6135×10^{-10}	0.996	0.000	.	3900.679841
13	CK7	2.3990×10^{1}	0.006	2.534	227.096	1.146828
	Napsin-A	2.0456	0.495	0.262	15.985	1.048985
	TTF1	8.2481×10^{8}	0.996	0.000	.	3900.816921

Table 7: Outcomes of binary logistic regression models of 20 different 3-biomarker combinations showing the odds ratio (OR), p-values, 95% confidence intervals (95% CI), and standard error (SE) of each biomarker when the others in the combination are controlled. The OR here signifies the odds of belonging to the 'primary lung tumour' group of the 'tumour origin' outcome (cont.)

#	Markers	OR	p-value	95% CI Lower Limit	95% CI Upper Limit	SE
14	CK7	1.5328×10^{1}	0.013	1.760	133.461	1.104172
	Napsin-A	8.9019	0.031	1.214	65.268	1.016464
	SATB2	4.5187×10^{-2}	0.002	0.006	0.331	1.015673
15	CK7	1.1446×10^{1}	0.024	1.388	94.392	1.076468
	Napsin-A	7.5571	0.034	1.169	48.836	0.952051
	CK20	4.2770×10^{-2}	0.007	0.004	0.430	1.177368
16	CK7	2.9564×10^{8}	0.996	0.000	.	3992.659253
	Napsin-A	8.3096	0.115	0.596	115.758	1.343947
	CDX2	4.4641×10^{-10}	0.996	0.000	.	3992.659188
17	CK7	3.3543×10^{1}	0.005	2.827	397.963	1.262024
	CK20	2.8180×10^{-3}	0.000	0.000	0.063	1.581727
	SATB2	5.5332×10^{-2}	0.021	0.005	0.641	1.249947
18	CK7	5.4692×10^{7}	0.996	0.000	.	3411.229945
	CK20	1.3333×10^{-1}	0.158	0.008	2.181	1.425950
	CDX2	1.9131×10^{-16}	0.994	0.000	.	4806.835288
19	CK7	1.1369×10^{8}	0.996	0.000	.	3490.069237
	CDX2	8.8464×10^{-17}	0.994	0.000	.	4960.807981
	SATB2	1.5086×10^{-8}	0.996	0.000	.	3525.484291
20	CK20	7.634×10^{-9}	0.997	0.000	.	5322.839407
	CDX2	1.5376×10^{-16}	0.994	0.000	.	4826.238351
	SATB2	2.1357×10^{-8}	0.996	0.000	.	3694.448736

Table 8: Comparison of the area under the curve (AUC), p-values, standard error (SE), and 95% confidence interval (95% CI) generated from the Receiver Operating Characteristic (ROC) curves for each of the 4 statistically significant 3-biomarker combinations.

#	Markers	ROC AUC	p-value	95% CI LB	95% CI UB	SE
1	CK7 / CK20 / SATB2	0.965	<0.001	0.930	1.000	0.018
2	CK7 / Napsin-A / SATB2	0.964	<0.001	0.922	1.000	0.022
3	CK7 / TTF1 / SATB2	0.910	<0.001	0.797	1.000	0.057
4	CK7 / Napsin-A / CK20	0.808	0.004	0.601	1.000	0.106

Table 9: Summary of the proportion of successfully stained samples in relation to the samples that were unsuccessful and the total number of samples in the set.

Case Processing Summary						
Markers	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
CK7	143	84.1%	27	15.9%	170	100.0%
Napsin-A	146	85.9%	24	14.1%	170	100.0%
TTF1	166	97.6%	4	2.4%	170	100.0%
CK20	123	72.4%	47	27.6%	170	100.0%
CDX2	121	71.2%	49	28.8%	170	100.0%
SATB2	119	70.0%	51	30.0%	170	100.0%

Table 10: Crosstabulation of CK7 vs. Tumour origin

			Tumour origin		Total
			Non-primary	Primary	
CK7	Negative	Count	14	8	22
		% within CK7	63.6%	36.4%	100.0%
		% within Tumour origin	58.3%	6.7%	15.4%
		% of Total	9.8%	5.6%	15.4%
	Positive	Count	10	111	121
		% within CK7	8.3%	91.7%	100.0%
		% within Tumour origin	41.7%	93.3%	84.6%
		% of Total	7.0%	77.6%	84.6%
Total		Count	24	119	143
		% within CK7	16.8%	83.2%	100.0%
		% within Tumour origin	100.0%	100.0%	100.0%
		% of Total	16.8%	83.2%	100.0%

Table 11: Crosstabulation of Napsin-A vs. Tumour origin

			Tumour origin		Total
			Non-primary	Primary	
Napsin-A	Negative	Count	19	30	49
		% within Napsin-A	38.8%	61.2%	100.0%
		% within Tumour origin	90.5%	24.0%	33.6%
		% of Total	13.0%	20.5%	33.6%
	Positive	Count	2	95	97
		% within Napsin-A	2.1%	97.9%	100.0%
		% within Tumour origin	9.5%	76.0%	66.4%
		% of Total	1.4%	65.1%	66.4%
Total		Count	21	125	146
		% within Napsin-A	14.4%	85.6%	100.0%
		% within Tumour origin	100.0%	100.0%	100.0%
		% of Total	14.4%	85.6%	100.0%

Table 12: Crosstabulation of TTF1 vs. Tumour origin

			Tumour origin		Total
			Non-primary	Primary	
TTF1	Negative	Count	25	26	51
		% within TTF1	49.0%	51.0%	100.0%
		% within Tumour origin	92.6%	18.7%	30.7%
		% of Total	15.1%	15.7%	30.7%
	Positive	Count	2	113	115
		% within TTF1	1.7%	98.3%	100.0%
		% within Tumour origin	7.4%	81.3%	69.3%
		% of Total	1.2%	68.1%	69.3%
Total		Count	27	139	166
		% within TTF1	16.3%	83.7%	100.0%
		% within Tumour origin	100.0%	100.0%	100.0%
		% of Total	16.3%	83.7%	100.0%

Table 13: Crosstabulation of CK20 vs. Tumour origin

			Tumour origin		Total
			Non-primary	Primary	
CK20	Negative	Count	6	105	111
		% within CK20	5.4%	94.6%	100.0%
		% within Tumour origin	42.9%	96.3%	90.2%
		% of Total	4.9%	85.4%	90.2%
	Positive	Count	8	4	12
		% within CK20	66.7%	33.3%	100.0%
		% within Tumour origin	57.1%	3.7%	9.8%
		% of Total	6.5%	3.3%	9.8%
Total		Count	14	109	123
		% within CK20	11.4%	88.6%	100.0%
		% within Tumour origin	100.0%	100.0%	100.0%
		% of Total	11.4%	88.6%	100.0%

Table 14: Crosstabulation of CDX2 vs. Tumour origin

			Tumour origin		Total
			Non-primary	Primary	
CDX2	Negative	Count	1	100	101
		% within CDX2	1.0%	99.0%	100.0%
		% within Tumour origin	5.9%	96.2%	83.5%
		% of Total	0.8%	82.6%	83.5%
	Positive	Count	16	4	20
		% within CDX2	80.0%	20.0%	100.0%
		% within Tumour origin	94.1%	3.8%	16.5%
		% of Total	13.2%	3.3%	16.5%
Total		Count	17	104	121
		% within CDX2	14.0%	86.0%	100.0%
		% within Tumour origin	100.0%	100.0%	100.0%
		% of Total	14.0%	86.0%	100.0%

Table 15: Crosstabulation of SATB2 vs. Tumour origin

			Tumour origin		Total
			Non-primary	Primary	
SATB2	Negative	Count	5	90	95
		% within SATB2	5.3%	94.7%	100.0%
		% within Tumour origin	25.0%	90.9%	79.8%
		% of Total	4.2%	75.6%	79.8%
	Positive	Count	15	9	24
		% within SATB2	62.5%	37.5%	100.0%
		% within Tumour origin	75.0%	9.1%	20.2%
		% of Total	12.6%	7.6%	20.2%
Total		Count	20	99	119
		% within SATB2	16.8%	83.2%	100.0%
		% within Tumour origin	100.0%	100.0%	100.0%
		% of Total	16.8%	83.2%	100.0%

Table 13: The sensitivity, specificity, positive predictive value, and negative predictive value of each biomarker when considered individually.

#	Marker	Sensitivity	Specificity	PPV	NPV
1	CK7	93.3%	58.3%	91.7%	63.6%
2	Napsin-A	76.0%	90.5%	97.9%	38.8%
3	TTF1	81.3%	92.6%	98.3%	49.0%
4	CK20	57.1%	96.3%	66.7%	94.6%
5	CDX2	94.1%	96.2%	80.0%	99.0%
6	SATB2	75.0%	90.9%	62.5%	94.7%