A transcriptome-Based Deep Neural Network Classifier for Identifying the Site of Origin in Mucinous Cancer

Taejin Ahn¹, Kidong Kim², Hyojin Kim³, Sarah Kim¹, Sangick Park¹ and Kyoungbun Lee⁴

¹Department of Life Science, Handong Global University, Pohang, Republic of Korea. ²Department of Obstetrics and Gynecology, Seoul National University Bundang Hospital, Seongnam, Republic of Korea. ³Department of Pathology, Seoul National University Bundang Hospital, Seongnam, Republic of Korea. ⁴Department of Pathology, Seoul National University Hospital, Seoul, Republic of Korea.

ABSTRACT

PURPOSE: There is a lack of tools for identifying the site of origin in mucinous cancer. This study aimed to evaluate the performance of a transcriptome-based classifier for identifying the site of origin in mucinous cancer.

MATERIALS AND METHODS: Transcriptomic data of 1878 non-mucinous and 82 mucinous cancer specimens, with 7 sites of origin, namely, the uterine cervix (CESC), colon (COAD), pancreas (PAAD), stomach (STAD), uterine endometrium (UCEC), uterine carcinosarcoma (UCS), and ovary (OV), obtained from The Cancer Genome Atlas, were used as the training and validation sets, respectively. Transcriptomic data of 14 mucinous cancer specimens from a tissue archive were used as the test set. For identifying the site of origin, a set of 100 differentially expressed genes for each site of origin was selected. After removing multiple iterations of the same gene, 427 genes were chosen, and their RNA expression profiles, at each site of origin, were used to train the deep neural network classifier. The performance of the classifier was estimated using the training, validation, and test sets.

RESULTS: The accuracy of the model in the training set was 0.998, while that in the validation set was 0.939 (77/82). In the test set which is newly sequenced from a tissue archive, the model showed an accuracy of 0.857 (12/14). t-SNE analysis revealed that samples in the test set were part of the clusters obtained for the training set.

CONCLUSION: Although limited by small sample size, we showed that a transcriptome-based classifier could correctly identify the site of origin of mucinous cancer.

KEYWORDS: Mucinous adenocarcinoma, transcriptome, computer neural networks, unknown primary neoplasms, ovarian neoplasms

Introduction

Mucinous cancer can originate in various abdominal organs including the stomach, colorectum, appendix, pancreas, ovary, endometrium and cervix, and the treatment of mucinous cancer differs according to its site of origin. For example, the standard chemotherapy regimen for patients of mucinous cancer originating from the ovary involves the administration of paclitaxel and carboplatin, whereas, that for mucinous cancer originating from the colorectum involves the administration of 5-fluorouracil. Therefore, the identification of the site of origin is necessary to determine the course of treatment.

Identification of the site of origin in mucinous cancer is difficult because the cell morphology can be similar between tumors originating from different sites. For example, in a trial involving patients of mucinous ovarian cancer, only 45% of the cases were confirmed to be of ovarian origin by the central pathology review. The misdiagnosed site of origin may misguide the course of treatment and result in a poor prognosis.

Currently, clinicopathologic characteristics, such as the bilaterality of ovarian tumors and CK7-, CK20-, and CDX2-immunoreactivity, are being used to identify the site of origin of mucinous cancer. However, there exists a huge overlap in the clinicopathologic characteristics among tumors of different origins. Several studies have identified the site of origin using machine learning methods. For example, a study showed that a deep learning-based algorithm can identify the site of origin. A few studies have reported that transcriptome-based analyses using the machine learning method can be used to identify the site of origin in metastatic carcinomas of unknown...
origin.\(^6\)\(^\text{–}^8\) In addition, gene expression profiling has been used to train a multiclass classifier based on a support vector machine algorithm,\(^6\)\(^\text{–}^8\) and an unsupervised cluster analysis method has been applied to evaluate the diagnostic power of a set of genes.\(^7\) However, these studies did not examine the performance of the transcriptome-based classifier in histology-specific cohorts. We believe that a transcriptome-based classifier will be more useful for mucinous cancers than for non-mucinous cancers because the identification of the site of origin is more difficult in mucinous cancers than in non-mucinous cancers.\(^10\)

We hypothesized that cancers have a distinct transcriptomic pattern according to site of origin, regardless of its histologic type (mucinous vs non-mucinous), and thus, transcriptomic analysis may be used to identify the site of origin of mucinous cancer. The objective of this study was to evaluate the performance of transcriptome-based deep neural network (DNN) classifier for identifying the site of origin of mucinous cancer specimens.

### Materials and Methods

#### Data

This study was conducted at a university hospital in the Republic of Korea from 2018 to 2020. The study protocol was approved and the need to acquire informed consent was waived by the Seoul National University Bundang Hospital institutional review board (B-1806-475-303, 2018-05-29). The study was conducted according to the tenets of the Declaration of Helsinki and its later amendments.

We downloaded the transcriptomic data of cancers from 7 sites of origin, namely, the uterine cervix (CESC), colon (COAD), pancreas (PAAD), stomach (STAD), uterine endometrium (UCEC), uterine carcinosarcoma (UCS), and ovary (OV), from The Cancer Genome Atlas (TCGA) database. The data comprised of RNA expression profiles of 60,498 genes from 1878 non-mucinous and 82 mucinous cancer specimens; these were used as the training and validation sets, respectively.

Additionally, for the test set, we collected 11 formalin-fixed paraffin-embedded (FFPE) tissue samples of deceased patients with mucinous cancer and 4 pairs of FFPE and fresh tissue samples of mucinous cancer from the biospecimen repository of the Seoul National University Hospital. Two FFPE tissues of appendix cancer were excluded because training set (TCGA database) did not include appendix as a site of origin. We performed RNA sequencing of the remaining 17 samples (9 FFPE and 4 pairs of FFPE and fresh tissues); however, we failed to complete the RNA sequencing of 1 FFPE tissue sample and 1 pair of FFPE and fresh tissue samples because of their low quality. Therefore, transcriptome of 14 mucinous cancer specimens (8 FFPE tissue samples and 3 pairs of FFPE and fresh tissue samples) were included in this study.

The proportion of each site of origin in the training, validation, and test sets is summarized in Table 1. Due to the small sample size, the distribution of the sites is not even among the training, validation, and test sets. In the training set, 7 sites of origin were included (CESC, COAD, PAAD, STAD, UCEC, UCS, OV). However, 4 sites of origin (CESC, COAD, PAAD, STAD) and 3 sites of origin (CESC, COAD, OV) were included in the validation and test sets, respectively.

### RNA sequencing

A cDNA library, consisting of 151 bp long fragments, was constructed for paired-end sequencing using the TruSeq stranded mRNA Sample Preparation Kit (Illumina, CA, USA), according to the manufacturer’s instructions. Briefly, mRNA was purified from 1 μg of total RNA, which was extracted from the tissue samples, and fragmented using oligo (dT) magnetic beads. Single-stranded cDNA was synthesized from the fragmented mRNA through random hexamer priming. The single-stranded cDNA is then used as a template for the synthesis of the second strand in order to construct double-stranded cDNA molecules. After end repair, A-tailing, and adapter ligation, cDNA libraries were amplified using Polymerase Chain Reaction (PCR). The quality of the cDNA libraries was evaluated with the 2100 BioAnalyzer (Agilent, CA, USA) and quantified with a library quantification kit (Kapa Biosystems, MA, USA) according to the manufacturer’s protocol. Following cluster amplification of the denatured templates, paired-end sequencing was performed (2 × 151 bp) using the Illumina NovaSeq6000 (Illumina, CA, USA).

The adapter sequences and pair-end reads with a Phred quality score of less than 20 were trimmed, and, simultaneously, reads shorter than 50 bp were removed by using Cutadapt v.2.8.\(^11\) Filtered reads were mapped to the reference genome of the same species using STAR v.2.7.1a,\(^12\) after estimation the transcriptome expression levels using the ENCODE standard option “–quantMode TranscriptomeSAM” (refer to “Alignment” of “Help” section in the html report).

The estimation of gene expression was performed by RSEM v.1.3.1,\(^13\) taking into consideration the direction of the reads.

### Table 1. Distribution of primary sites in the training, validation, and test set.

|          | TRAINING | VALIDATION | TEST |
|----------|----------|------------|------|
| CESC     | 292      | 17         | 4    |
| COAD     | 291      | 40         | 7    |
| PAAD     | 179      | 4          | 0    |
| STAD     | 429      | 21         | 0    |
| UCEC     | 204      | 0          | 0    |
| UCS      | 57       | 0          | 0    |
| OV       | 426      | 0          | 3    |
| Subtotal | 1878 (100)| 82 (100)   | 14 (100) |

Carcinomas arising from uterine cervix (CESC), colon (COAD), pancreas (PAAD), stomach (STAD), uterine endometrium (UCEC), uterine carcinosarcoma (UCS), and ovary (OV).
which is done using option –strandedness. To improve the accuracy of the measurement, the “–estimate-rspd” option was applied, and all other options were set to default. To normalize sequencing depth among the samples, fragments per kilobase million mapped reads (FPKM) and transcripts per kilobase million mapped reads (TPM) values were calculated.

**Data pre-processing**

The workflow for data pre-processing and gene selection is depicted in Figure 1. Among the 60,498 transcripts, 20,498 transcripts were common to all sequencing data sets and were selected to build the prediction model. We applied a log2 transformation after adding 0.001 to original FPKM value of the selected genes, and standardized the value by z-scaling, which involves subtracting the mean from a sample’s expression value and dividing it by the standard deviation of all the transcripts in the sample. We applied the same scaling strategy to the validation and test sets. The combined batch and quantile normalization methods for gene expression microarray data were applied in an N + 1 manner to eliminate batch effects between the TCGA (training and validation set) and test sets and to equalize gene expression distributions. N + 1 combined batch and quantile normalization were performed on each of the 14 test samples used for prediction.14,15 We used “Combat” function of “sva” (v 3.38.0) R package for removing batch effect, and “normalizeBetweenArrays” function of “limma” (3.46.0) for quantile normalization between samples.

**Gene selection**

We calculated the R² of gene expressions value of the 3 pairs of FFPE and fresh tissue samples and selected genes with similar expression patterns (R² ≥ 0.8; Supplemental Figure 1), obtaining a total of 6667 genes. After performing equal variance analysis on the 6667 genes for each tissue, we performed t-tests for genes with equal variance and Welch’s test for genes with unequal variances to select genes that significantly differed from each other at each identified site of origin. Among the genes that significantly differed for each identified site of origin, we selected 100 differentially expressed genes in the tissues (Supplemental Tables 1 and 2). Finally, we used 427 genes as the training set after removing redundant genes from the different sites of origin of non-mucinous cancer specimens (Figure 2).
Functional analysis

We used the web version of g:Profiler to analyze the function of the selected gene. Gene function analysis was performed on the gene ontology and KEGG pathway provided by g:Profiler. We selected only functions that satisfy the adjusted $P$-value $<.05$ (Supplemental Table 3).

Prediction model

We built a DNN classifier and trained the model using the RNA expression profiles of 427 genes from 1878 non-mucinous cancer specimens as the training set. To optimize the performance, we evaluated 8100 hyper-parameter experimental combinations, the details of which are provided in Supplemental Table 4. We trained and tested 8100 SVM, XGBoost, Random Forest, and multinomial regression models for performance comparison with the DNN models.

Visualization and statistical analysis

We used Rtsne ($\geq 0.15$) R package for t-SNE analysis, using a principle component analysis of FALSE and a theta of 0.0.
Hierarchical clustering by Euclidean distance and visualization were conducted using the heatmap3 (v 1.1.6) R package.

**Results**

The accuracy of model for the training set was 0.998 (Supplemental Table 5) and that for the validation set was 0.939 (77/82) (Table 2). In the validation set, the 5 misclassifications were for specimens that had originated from CESC and COAD. Specifically, 4 cases of CESC were misclassified as COAD (n = 2) or UCEC (n = 2). One case of COAD was misclassified as STAD.

The average classification accuracy for the validation set of the DNN model was 0.909. The average classification accuracy of the machine learning models SVM, multinomial regression, Random Forest, and XGBoost was 0.898, 0.872, 0.844, and 0.823, respectively; among the 5 classification models, DNN had the highest average classification accuracy (Figure 3). We selected the model with the highest classification accuracy for the validation set among the DNN models for classification of the test set. We also performed classification on the test set using the model with the highest performance on the validation set in the machine learning model. The classification test accuracy of the machine learning models SVM, multinomial regression, Random forest, and XGBoost was 0.5, 0.143, 0.286, and 0.357, respectively.

In the test set, the accuracy of the model was 0.857 (12/14). Fresh and FFPE paired samples (paired C) that originated from CESC were misclassified as OV and STAD, respectively (Table 3). t-SNE analysis positioned samples of the validation and test set in clusters of the training set (Figure 4a). t-SNE analysis positioned samples of the test set in clusters of the training set (Figure 4b).

**Discussion**

This is the first study to demonstrate that the site of origin of mucinous cancer can be identified in a transcriptome-based analysis, and identification of the correct site of origin may aid in the optimization of the treatment strategy and improve the prognosis of the patient with mucinous cancer.
Previous studies involving transcriptome-based classifiers did not examine the performance in terms of classification of minor histology types, especially mucinous carcinoma of our interest. In this study, we trained our model based on the transcriptomes of non-mucinous cancer specimens and successfully classified the sites of origin of mucinous cancer specimens. This finding indicates that mucinous and non-mucinous cancers share common transcriptomes based on the site of origin. However, it is worthy to mention that our finding is limited if the direct origin of mucinous carcinoma is primary cancer as we have not obtained mucinous carcinoma cases originated from metastatic cancer. There is a concern that low tumor purity can dilute the tissue specific gene expression signal. This may limit the performance gene expression based classifier. TCGA samples used training in our study are expected more than 50% of purity, thus sequencing of fresh tissue more than 50% of purity may provide strong tissue specific signal to be applied to our classifier.

The transcriptome of paraffin-fixed tissues can be different from that of fresh tissues. For example, in a study examining matched fresh and FFPE tissue samples, more than half of genes were differentially expressed between the 2 tissue types. Therefore, a model trained using transcriptomic data from fresh tissues may perform poorly when transcriptomic data from FFPE tissues is used as the input. In our study, we trained and validated our model using transcriptomic data from fresh tissues (TCGA dataset). However, the test set comprised of data from both paraffin-fixed and fresh tissues. We selected genes similarly expressed between paraffin-fixed and fresh tissues to minimize the effect of tissue type. The accuracy of identifying the site of origin for the paraffin-fixed tissue samples was 0.909 (10/11), while that for fresh tissue samples was 0.667 (2/3) in the test set. These numbers suggest that difference in tissue type (paraffin-fixed vs fresh) can be overcome by the careful selection of genes for the classifier.

This study has several limitations. First, the small sample size of test set which is 14 (11 paraffin-fixed, 3 fresh tissue) newly sequenced mucinous carcinoma cases may provide biased result. Additional validation with larger sample size of mucinous carcinoma of fresh tissue is required. Second, due to rarity of mucinous cancer, cancers from several sites (UCEC, UCS, OV) were not included in the validation set, and thus, there is a possibility that the model was under-trained for these cancers. Third, we did not include some primary sites in the test set (PAAD, STAD, UCEC, UCS), and therefore, we do not know the performance of our model in these cancers.

In conclusion, although limited by the small sample size, we showed that our transcriptome-based analysis correctly identifies the site of origin of various mucinous cancers.

Acknowledgements
The biospecimens for this study were provided by the Seoul National University Hospital (SNUH) Cancer Tissue Bank. All samples derived from the Cancer Tissue Bank of SNUH were obtained upon receiving informed consent, according to institutional review board-approved protocols. We would like to thank Editage (www.editage.co.kr) for English language editing.

Author Contributions
Taejin Ahn: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – Original Draft, Writing – Review & Editing
Kidong Kim: Conceptualization Methodology, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing, Supervision, Project administration, Funding acquisition
Hyojin Kim: Methodology, Resources, Writing – Review & Editing
Sarah Kim: Software, Validation, Formal analysis, Investigation, Writing – Original Draft, Writing – Review & Editing, Visualization
Sangick Park: Software, Validation, Formal analysis, Investigation, Writing – Original Draft, Writing – Review & Editing, Visualization
Kyoungbun Lee: Conceptualization Methodology, Resources, Writing – Review & Editing

Availability of Data and Materials
The datasets used and/or analyzed during the current study are available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE163126.

Consent for Publication
Waived by the institutional review board

Ethical Statements and Informed Consent Information
The study protocol was approved and the need to acquire informed consent was waived by the Seoul National University Bundang Hospital institutional review board (B-1806-475-303, 2018-05-29). The study was conducted according to the tenets of the Declaration of Helsinki and its later amendments.

Supplemental Material
Supplemental material for this article is available online.

REFERENCES
1. Gore M, Hackshaw A, Brady WE, et al. An international, phase III randomized trial in patients with mucinous epithelial ovarian cancer (mEOC/GOG 0241) with long-term follow-up: and experience of conducting a clinical trial in a rare gynecological tumor. Gynecol Oncol. 2019;153:541-548.
2. Simons M, Ezendam N, Bulten J, Nagtegaal I, Massuger L. Survival of patients with mucinous ovarian carcinoma and ovarian metastases: a population-based cancer registry study. Int J Gynecol Cancer. 2015;25:1208-1215.
3. Lee KR, Young RH. The distinction between primary and metastatic mucinous carcinomas of the ovary: gross and histologic findings in 50 cases. Am J Surg Pathol. 2003;27:281-292.
4. Kelemen LE, Köbel M. Mucinous carcinomas of the ovary and colorectum: different origin, same dilemma. Lancet Oncol. 2011;12:1071-1080.
5. Meagher NS, Schuster K, Voss A, et al. Does the primary site really matter? Profiling mucinous ovarian cancers of uncertain primary origin (MO-CUP) to personalise treatment and inform the design of clinical trials. Gynecol Oncol. 2018;150:527-533.
6. Ramaswamy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A. 2001;98:15149-15154.
7. Buckhaults P, Zhang Z, Chen YC, et al. Identifying tumor origin using a gene expression-based classification map. Cancer Res. 2003;63:4144-4149.
8. Tornhill RW, Kowalczyk A, Rischin D, et al. An expression-based site of origin diagnostic method designed for clinical application to cancer of unknown origin. Cancer Res. 2005;65:4031-4040.
9. Lu MY, Chen TY, Williamson DFK, et al. AI-based pathology predicts origins for cancers of unknown primary. Nature. 2021;594:106-110.
10. McCluggage WG. Immunohistochemistry in the distinction between primary and metastatic ovarian mucinous neoplasms. J Clin Pathol. 2012;65:596-600.
11. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBO Rep. 2011;12:10-12.
12. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15-21.
13. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.
14. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:429-441.
15. Ritchie ME, Phipson B, Wu D, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.
16. Radtke U, Kolberg L, Kuzmin I, et al. GProfiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 2019;47:W191-W198.
17. Marczyk M, Fu C, Liu R, et al. The impact of RNA extraction method on accurate RNA sequencing from formalin-fixed paraffin-embedded tissues. BMC Cancer. 2019;19:1189.