GC Genome

Permanent Abstract Number: 6371

Deep Learning Algorithm for Multi-cancer Detection and Classification using cf-WGS

Tae-Rim Lee[1], Jin Mo Ahn[1], Joo Hyuk Sohn[2], Sook Ryun Park[3], Min Hwan Kim[2], Gun Min Kim[2], Ki-Byung Song[4], Eunsung Jun[4], Dongryul Oh[5], Jeong-Won Lee[6], Joseph J Noh[6], Young Sik Park[7], Sun-Young Kong[8], Sang Myung Woo[9], Bo Hyun Kim[9], Eui Kyu Chie[10], Hyun-Cheol Kang[10], Youn Jin Choi[11], Ki-Won Song[12], Jeong-Sik Byeon[13], Junnam Lee[1], Dasom Kim[1], Chang-Seok Ki[1], Eun-Hae Cho[1]*.

1 Genome Research Center, GC Genome, Yongin, Korea
2 Division of Medical Oncology, Department of Internal Medicine, Yonsei Cancer Center, Yonsei University College of Medicine, Seoul, Korea
3 Department of Oncology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea
4 Department of Surgery, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea
5 Department of Radiation Oncology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea
6 Department of Obstetrics and Gynecology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea
7 Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, Seoul National University Hospital, Seoul, Korea
8 Department of Laboratory Medicine, National Cancer Center, Goyang, Korea
9 Center for Liver and Pancreatobiliary Cancer, National Cancer Center, Goyang, Korea
10 Department of Radiation Oncology, Seoul National University College of Medicine, Seoul, Korea
11 Department of Obstetrics and Gynecology, Seoul St. Mary's Hospital, The Catholic University College of Medicine, Seoul, Korea
12 Division of Hepatopancreatobiliary Surgery and Liver Transplantation, Department of Surgery Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea
13 Department of Gastroenterology, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Korea.

Purpose

Several cell-free DNA (cf-DNA) features, such as genome-wide coverage, fragment size, and fragment end motif frequency, have shown their potentials for cancer detection. In this study, we developed two independent models, GC (gross chromatin), and FEMS (fragment end motif frequency and size). Each model uses images generated from genome-wide normalized sequencing coverage and cf-DNA fragment end motif frequencies according to the different cf-DNA size profiles. Then we integrated them into a single ensemble model to improve cancer detection and multi-cancer type classification accuracy.

Method

Low depth cfDNA-WGS data was generated from 1,396 patients (stage I: 14.9%, stage II: 35.6%, stage III: 24.9%, stage IV: 24.2%, unknown: 0.4%) with breast (n=702), liver (n=213), esophageal (n=155), ovarian (n=151), pancreatic (n=85), lung (n=53), head and neck (n=16), biliary tract (n=15), and colon cancer (n=6) and 417 healthy individuals. Samples were randomly split into training, validation, and test dataset stratified by cancer types and stages. Cancer types with a small number of samples (<20) were excluded for multi-cancer type classification. Each model was trained using a convolutional neural network, then integrated into a single ensemble model by averaging the predicted probabilities calculated from each model.

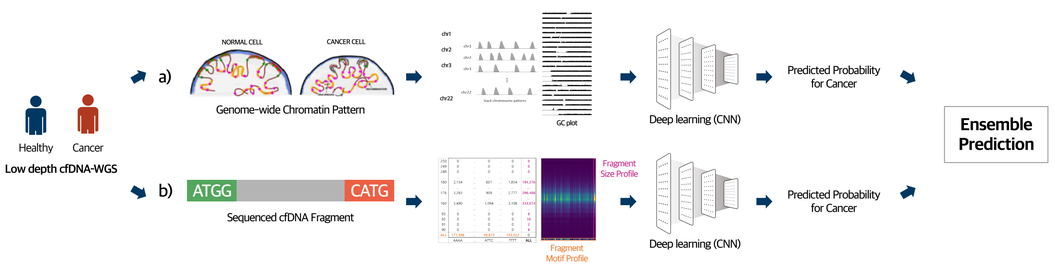

Figure 1: Schematic Overview. a) GC model. Gross chromatin (GC) model uses sequencing coverage pattern of cfDNA which mimics nucleosome positioning in genome. Coverage pattern is plotted per chromosome and stacked to create a single GC plot per sample. **b) FEMS model.** Fragment end motif frequencies and size (FEMS) information was used to create FEMS plot which represents fragmentomic profile of plasma cfDNA. X and y axes represent motif frequency and size profile respectively. Two independent CNN models were trained using GC and FEMS plot and then the predicted probabilities calculated from each model were averaged to make a final ensemble prediction.

Cancer Type	Train	Validation	Test	Total
Breast	420	140	142	702
Liver	126	43	44	213
Esophageal	92	30	33	155
Ovarian	89	31	31	151
Pancreatic	49	17	19	85
Lung	30	11	12	53
Head and Neck	9	3	4	16
Biliary Tract	8	3	4	15
Colon	2	2	2	6
Healthy	250	83	84	417
Total	1,075	363	375	1,813

Table 1: Data Used. Train, validation, and test dataset were used for model training, hyper-parameter tuning, and final performance evaluation respectively.

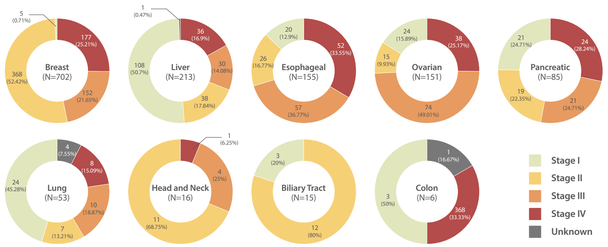

Figure 2: Cancer Stage Distribution

Result

For cancer detection, the ensemble model achieved sensitivities of 85.2% [95% confidence interval (CI): 71.8% to 94.5%], 74.9% (CI: 68.0% to 88.0%), 73.2% (CI: 66.7% to 85.9%) at a specificity of 95%, 98% and 99% and the AUC value of 0.97(CI: 0.95-0.99) in the test dataset. By the cancer stages, sensitivity was 62.8% (CI: 48.8% to 83.7%) in stage I, 66.3% (CI: 57.7% to 82.7%) in stage II, 85.9% (CI: 77.5% to 94.4%) in stage III, and 76.1% (CI: 63.4% to 87.3%) in stage IV at 99% specificity. For multi-cancer classification, the overall accuracy of 85.1% (CI: 80.4% to 89.3%) was achieved including 6 cancer types.

Model	Accuracy (95% CI)	AUC (95% CI)	95% Specificity Sensitivity (95% CI)	98% Specificity Sensitivity (95% CI)	99% Specificity Sensitivity (95% CI)
GC_FEMS Ensemble Model	92.8% (90.4% to 95.2%)	97.2% (95.4% to 98.6%)	85.2% (71.8% to 94.5%)	74.9% (68.0% to 88.0%)	73.2% (66.7% to 85.9%)

Table 2: Cancer Detection Performance in Test Dataset

Figure 3: Cancer Detection Sensitivity by Cancer Types

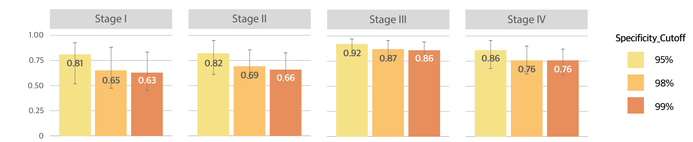

Figure 4: Cancer Detection Sensitivity by Cancer Stages

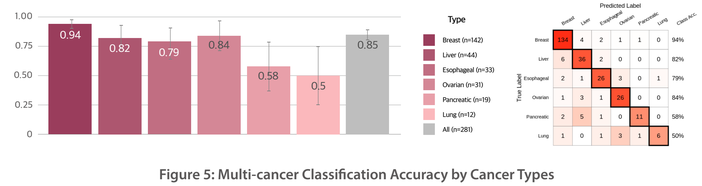

Figure 5: Multi-cancer Classification Accuracy by Cancer Types

Conclusion

Highly sensitive and accurate deep learning model for cancer detection and multi-cancer classification was generated by combining different types of cf-DNA features. This result provides the opportunity for general population multi-cancer screening using various cf-DNA features.

www.gcgenome.com | **Contact information.** ehcho@gccorp.com