Deep learning algorithm for cancer detection using multimodal characteristics of whole methylome sequencing of cf-DNA

Abstract Presentation Number: 6697

Jun Tae Park[1], Minjung Kim[1], Sook Ryun Park[2], Ki-Byung Song[3], Eunsung Jun[3], Dongryul Oh[4], Jeong-Won Lee[5], Young Sik Park[6], Ki-Won Song[7], Jeong-Sik Byeon[8], Bo Hyun Kim[9], Chang-Seok Ki[1], Eun-Hae Cho[1]

1 Genome Research Center, GC Genome, Yongin, Korea
2 Department of Oncology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea
3 Department of Surgery, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea
4 Department of Radiation Oncology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea
5 Department of Obstetrics and Gynecology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea
6 Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, Seoul National University Hospital, Seoul, Korea
7 Division of Hepatopancreatobiliary Surgery and Liver Transplantation, Department of Surgery Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea
8 Department of Gastroenterology, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Korea
9 Center for Liver and Pancreatobiliary Cancer, National Cancer Center, Goyang, Korea

Background

Various cell-free DNA (cfDNA) features including methylation and genomic profiles have been investigated for their potential use in early cancer detection. We developed deep learning models based the data generated by the enzymatic conversion based whole methylome sequencing of cfDNA.

Method

Cell-free whole genome Enzymatic Methyl sequencing (cfWEMseq) data were generated from 198 cancer patients (stage I: 11%, II: 17%, III: 22%, IV: 20%, unknown: 31%) and 69 healthy controls. The cancer types were consisted of breast (n=31), liver (n=24), esophageal (n=38), pancreatic (n=30), colon (n=34), ovarian (n=18), and lung (n=23). Sequence data was produced on average of 200 million reads using Novaseq 6000 (Illumina). For model training and evaluation, data partitioning was stratified by cancer type, and 5-fold cross validation was used. Coverage and methylation beta values were calculated by binning at fixed size of 100K, 1M, and 5M base and variable size from Topologically Associated Domains (TAD). Genome Coverage (GC), Genome Methylation Beta values (GMB), and Mutation Signature (MS) features were trained using a one-dimensional convolutional neural network (1D-CNN). The performance of the model was evaluated by measuring the average value of the results measured in each test set of 5 fold.

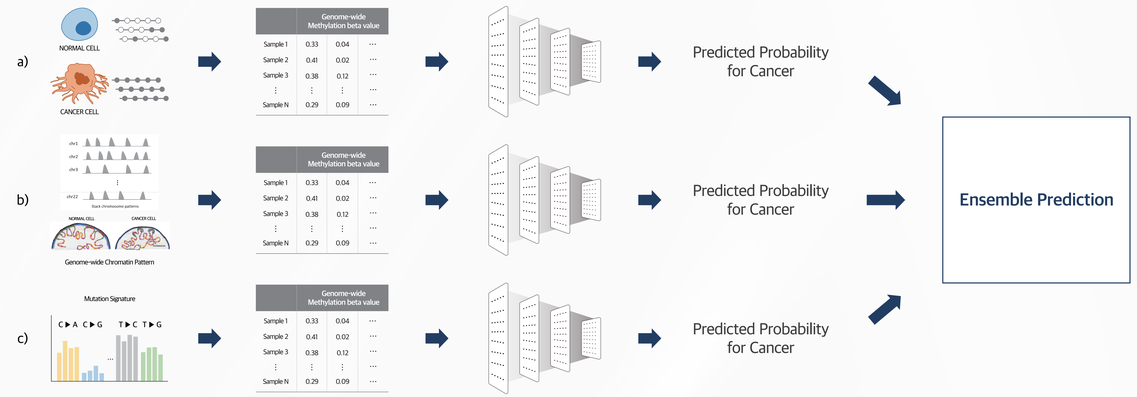

Figure 1. a) Methylation Beta Value model, b) Genome Coverage model, c) Mutation Signature model. Each model was averaged to make a final ensemble prediction.

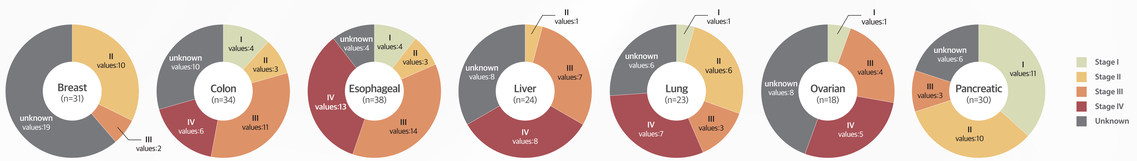

Figure 2: Stage Distribution by Cancer Types

Result

We tested the cancer detection performance of various feature combinations using all data from cfWEMseq (n=267). Regardless of the bin size, the GMB single model achieved higher performance than the GC single model. The best-performing model is the ensemble model of GMB (100k bin) and MS. The cancer detection performance of this ensemble model reached an accuracy 96% (CI: 93.6% to 98.1%), AUC 0.99 (CI: 0.97 to 1.0) and sensitivity 98.0% (CI: 92.4% to 99.5%) with a specificity of 90%.

Model	Accuracy (95% CI)	AUC (95% CI)	90% Specificity Sensitivity (95% CI)	95% Specificity Sensitivity (95% CI)
MS_GMB(100K) Ensemble Model	95.9% (93.6% to 98.1%)	0.99 (0.97 to 1.00)	98.0% (92.4% to 99.5%)	93.4% (82.3% to 99.0%)

Table 1: Cancer Detection Performance in Test Dataset

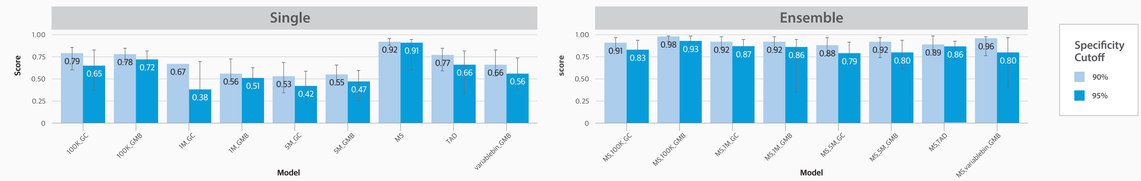

Figure 3: Cancer Detection Sensitivity by Model Types

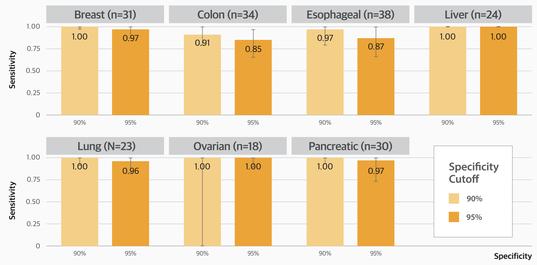

Figure 4: Cancer Detection Sensitivity by Cancer Types

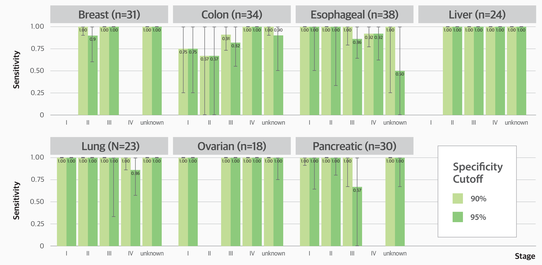

Figure 5: Cancer Detection Sensitivity by Cancer Stage

Conclusion

These results provide an opportunity for higher accuracies by integrating methylation information and genomic data using cfWEMseq.

The authors have no financial conflicts of interest to disclose concerning the presentation. | **Contact information:** parkjt@gccorp.com