A deep learning framework for predicting gene expression from cell-free DNA

Robert Patton[1,2], Alexander Netzley[1,2], Akira Nair[1,2], Thomas Persse[1,2], Mohamed Adil[1], Ilsa Coleman[2], Michael Yang[1,2], Patricia Galipeau[1,2], Joseph Hiatt[2], David MacPherson[1,2], Michael Haffner[2,3,4], Pete Nelson[2,3,5,6,7], Gavin Ha[1,2,6,8]

[1] Division of Public Health Sciences, Fred Hutchinson Cancer Research Center

[2] Division of Human Biology, Fred Hutchinson Cancer Research Center

[3] Division of Clinical Research, Fred Hutchinson Cancer Research Center

[4] Department of Laboratory Medicine and Pathology, University of Washington

[5] Department of Urology, University of Washington

[6] Brotman Baty Institute for Precision Medicine

[7] Division of Oncology, Department of Medicine, University of Washington

[8] Department of Genome Sciences, University of Washington

Background

Cell-free DNA (cfDNA) is increasingly utilized as a minimally invasive alternative to traditional biopsies, addressing diverse clinical needs ranging from early cancer detection to tumor phenotyping and monitoring treatment response and resistance. Current methods in the field have looked beyond traditional genomics analysis, leveraging epigenetic information inferred through cfDNA coverage and fragmentation patterns which reflect underlying chromatin structure and transcription factor binding events. Despite these advances, epigenetic analysis from cfDNA is still nascent and accurate, individual gene expression prediction from cfDNA analogous to RNA-seq remains a challenge. To overcome this we developed *Proteus*, a deep-learning framework designed to predict single gene expression levels using standard whole-genome cfDNA sequencing.

Methods

We first developed the tool *Triton* in order to comprehensively characterize the cfDNA feature landscape in regions of interest. *Triton* reports fragmentation and inferred nucleosome positioning both at the region level and at bp-resolution, providing an extensive view of the cfDNA epigenetic landscape. We then integrated these features into *Proteus*, a multi-modal convolutional neural network architecture designed to jointly analyze signals and region-level features from individual genes and predict an expression value comparable to that of RNA-seq of tissue. The model was first trained on pure circulating tumor DNA (ctDNA) bioinformatically purified from LuCaP and small-cell lung cancer patient-derived xenograft (PDX) models with matched tumor RNA-seq to develop foundational learning and to evaluate feature dependencies. We then applied transfer learning using PDX ctDNA mixed with healthy cfDNA in silico to tune the model to account for variable ctDNA fractions and to evaluate the lower limit of ctDNA needed for accurate expression prediction.

Results

Proteus was initially validated using 5-fold cross-validation with hold-out on PDX samples, and achieved R^2 greater than 0.9 at the full transcriptome level, comparable to conventional tissue RNA-seq technical

replicates. In patient samples, *Proteus* was accurate to the individual gene level, showing significant correlations with RNA-seq from metastases in phenotype-defining genes, therapeutic targets, and in gene set scores even at low tumor fractions. Differential gene expression analysis also showed high concordance between RNA-seq and *Proteus*-predicted expression values, and we successfully re-identified phenotype-defining genes directly from cfDNA.

Conclusions

Proteus represents a powerful tool for precision oncology as a means to monitor genome-wide epigenetic changes throughout the body using minimally invasive liquid biopsies, providing a direct analog to RNA-seq for predicting individual gene expression levels from standard cfDNA sequencing. Its ability to integrate seamlessly with existing RNA-seq-based tools and workflows further enhances its potential clinical utility, readily enabling the ongoing monitoring of therapeutic targets, molecular subtyping, and de novo epigenetic discovery.

Funding Acknowledgements

The Prostate Cancer Foundation

NIH/NCI R01 CA280056, DP2 CA280624

Conflict of Interest Statement

No disclosures to report.