

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

Robert D. Patton^{1,2}, Alexander Netzley^{1,2}, Thomas Persse^{1,2}, Akira Nair^{1,2}, Ilsa Coleman², Patricia Galipeau^{1,2}, Pushpa Itagi^{1,2}, Michael Haffner^{2,3,4}, Joseph Hiatt², Andrew Hsieh^{2,8}, David MacPherson^{1,2}, Peter S. Nelson^{2,3,4,5,6,7,8}, 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, single gene expression prediction from cfDNA analogous to RNA-seq remains a challenge. Current liquid-biopsy based methods also require complex, multi-component assays or ultra-deep sequencing from targeted panels. To overcome this we developed *Proteus*, a deep-learning framework designed to predict individual 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 deep learning 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. In order to develop foundational learning and to evaluate feature dependencies, the model was first trained and validated 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, along with PDX ctDNA mixed with healthy cfDNA *in silico* to replicate variable tumor content with known proportions. We then applied transfer learning using cfDNA from patients with matched tissue RNA-seq from multiple metastases to tune the model to real world conditions and variability.

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 and pathway analyses also showed high concordance between RNA-seq and *Proteus*-predicted expression values, and we successfully re-identified phenotype-defining genes and pathways directly from cfDNA.

Conclusions

Proteus represents a powerful tool for precision oncology as a means to monitor genome-wide transcriptional 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* discovery.

Funding Acknowledgements

The Prostate Cancer Foundation

NIH/NCI R01 CA280056, DP2 CA280624

Conflict of Interest Statement

No disclosures to report.