Latent Representations from Factor Analysis and CNNs of Genomic and Proteomic Data Reveal Immune Pathways in Colorectal Cancer

Tzu-Yu Liu, Francesco Vallania, Michael Dzamba, Mitch Bailey, Charles Roberts, Barbara Engelhardt*, C. Jimmy Lin*

Freenome Inc., South San Francisco, CA

* corresponding authors (authors@freenome.com)

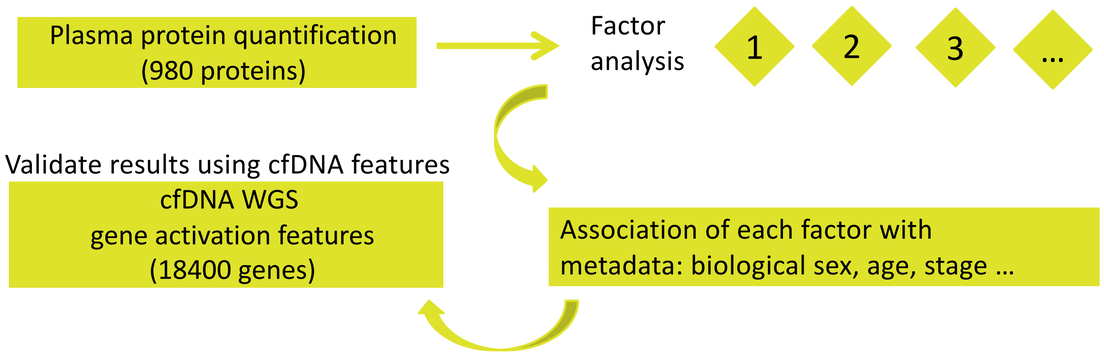
INTRODUCTION

- Plasma proteins and cell-free DNA (cfDNA) are important classes of cancer biomarkers
- Our goal is to identify markers of CRC from this high-dimensional space
- Characterization of their biological functions and interrelationships is ongoing

METHODS

We performed whole-genome sequencing (WGS) of plasma cfDNA and targeted proteomic analysis (980 plasma proteins) in a cohort of 455 human subjects (colorectal cancer (CRC) cases n=36; colonoscopy-confirmed CRC-negative controls (NEG) n=419) and applied a framework to learn latent representations of biological knowledge (Figure 1).

Figure 1. Experimental design and analytical approach



SFAmix: A latent factor model with a mixture of sparse and dense factors

Factor analysis is an efficient method to identify covarying signals by decomposing the measurements into a set of loadings and factors. By imposing priors that encourage sparsity, one may identify proteins that covary and may be used to jointly distinguish cases and controls. We applied SFAmix to the targeted proteomic data to identify relevant pathways.

- Bayesian factor analysis:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\Lambda} + \epsilon$$
$$\boldsymbol{X}_i \sim \mathcal{N}(0, \boldsymbol{I}_K)$$
$$\epsilon_j \sim \mathcal{N}(0, \Psi_j)$$

$\boldsymbol{Y}$: the matrix of observed variables, $\mathrm{n} \times p$, (e.g., plasma proteins)
$\boldsymbol{X}$: the factor matrix with K factors, $\mathrm{n} \times K$
$\boldsymbol{\Lambda}$: the loading matrix, $K \times p$
ϵ : the residual error matrix, $\mathrm{n} \times p$
$\boldsymbol{I}_K$: the $\mathrm{K} \times K$ identity matrix
$\Psi = \mathrm{diag}(\psi_1, \ldots, \psi_p)$

The observed data can be viewed as being generated from:

$$Y_{i,j} | \Lambda_{k,j}, X_{i,k}, \psi_j \sim \mathcal{N}\left(\sum_{k=1}^{K} X_{i,k}\, \Lambda_{k,j}, \psi_j\right)$$

- Mixture of sparse and dense factors:

$$\pi | \alpha, \beta \sim Beta(\alpha, \beta)$$
$$Z_k \sim Bern(\pi), \mathrm{k} = \{1, 2, \ldots, \mathrm{K}\}$$
$$\Lambda_{k,j} \mid Z_k \sim \begin{cases} p\left(\Lambda_{k,j} | \theta_{k,j}, \delta_{k,j}, \phi_k\right), & if \;\; Z_k = 1 \\ p\left(\Lambda_{k,j} | \phi_k\right), & if \;\; Z_k = 0 \end{cases}$$
$$X_{i,k} \sim \mathcal{N}(0,1)$$
$$Y_{i,j} | \Lambda_{k,j}, X_{i,k}, \psi_j \sim \mathcal{N}\left(\sum_{k=1}^{K} X_{i,k}\, \Lambda_{k,j}, \psi_j\right)$$

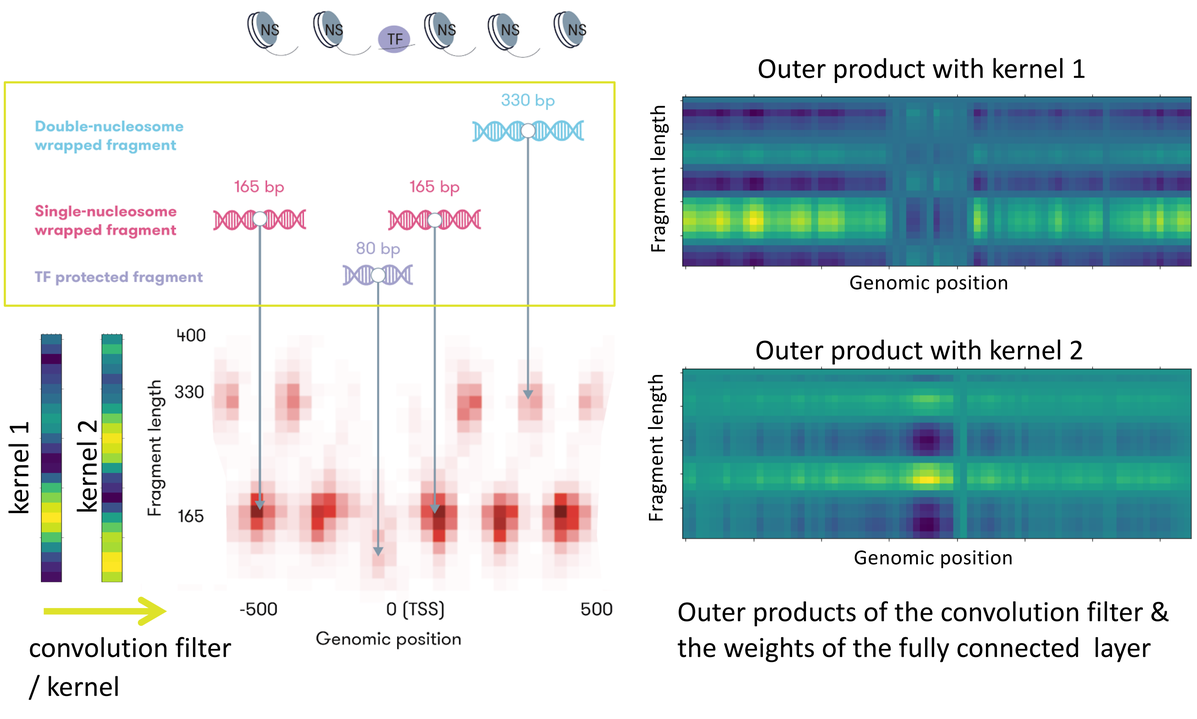
Z_k : Latent vector indicating whether a factor is sparse or dense (1: sparse; 0: dense)

$\theta_{k,j}, \delta_{k,j}, \phi_k$: Parameters that control local shrinkage for each element or factor-specific shrinkage, and can be modeled using a three parameter beta (*TPB*) distribution

Convolutional neural network (CNN) was applied to chromatin accessibility metrics at transcription start sites (TSS) to differentiate "on" vs "off" genes.

cfDNA fragment length and positioning around TSS are both markers of gene activation. We first trained a CNN as a predictor of gene activation using average expression of stable genes from external datasets and the chromatin accessibility metrics of fragment length and positioning (Figure 2). We then applied the CNN to other genes to estimate gene activation probability. These predictions were compared to the plasma protein biomarker abundances, since they may originate from the same cell types (Figure 5).

Figure 2. CNN model to predict gene activation using chromatin accessibility

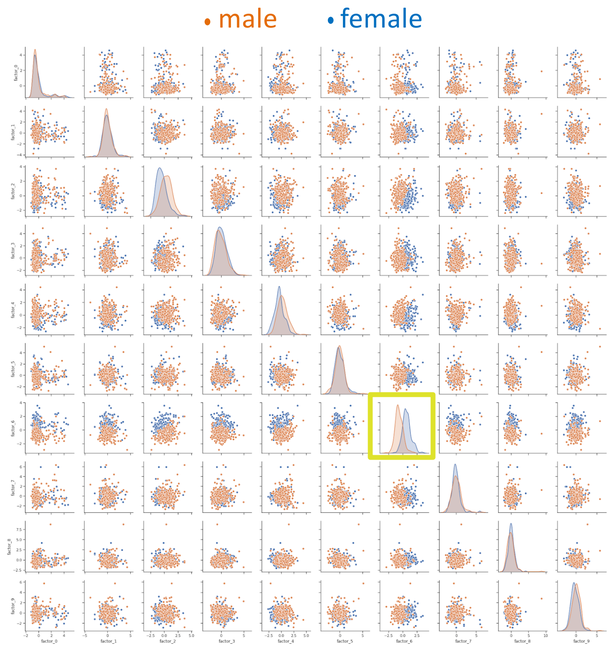


RESULTS

SFAmix applied to plasma protein abundances revealed differences in immune activation based on biological sex.

- We found significant enrichment of immune pathways within a single factor, especially those related to immune activation
- The scores from the same factor are bimodal, distinguishing females from males (Figure 3)
- This sex-specific immune activation recapitulates known differences in immune response based on biological sex (T test of male vs female: *p*-value < 2.2e-16)

Figure 3. SFAmix factor scores



pathway	padj	factor
GO_DEFENSE_RESPONSE	0.026	6
GO_REGULATION_OF_IMMUNE_SYSTEM_PROCESS	0.026	6
GO_IMMUNE_EFFECTOR_PROCESS	0.026	6
GO_POSITIVE_REGULATION_OF_IMMUNE_SYSTEM_PROCESS	0.026	6
GO_REGULATION_OF_IMMUNE_RESPONSE	0.026	6
GO_INNATE_IMMUNE_RESPONSE	0.026	6
GO_CELLULAR_RESPONSE_TO_OXYGEN_CONTAINING_COMPOUND	0.026	6
MODULE_64	0.026	6
GO_CELLULAR_RESPONSE_TO_LIPID	0.026	6
GO_CELL_ACTIVATION	0.040	6
GO_REGULATION_OF_IMMUNE_EFFECTOR_PROCESS	0.042	6
KORKOLA_CHORIOCARCINOMA	0.048	6
GO_LEUKOCYTE_MEDIATED_IMMUNITY	0.050	6

SFAmix vs PCA: PCA has difficulty finding pathways that differentiate biological sex.

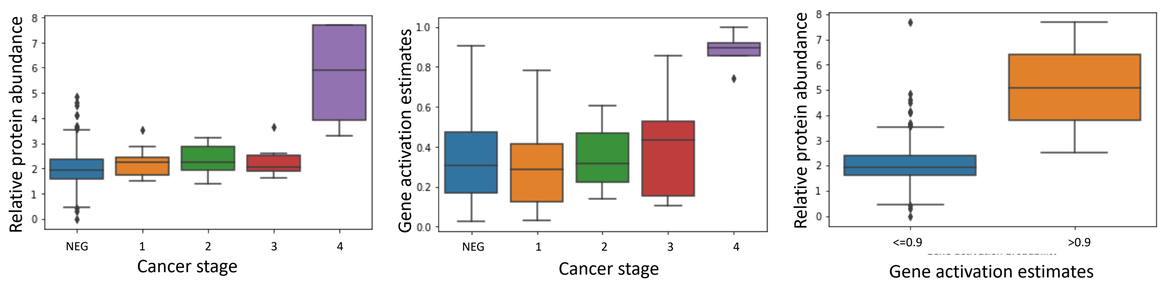
- Similar analysis with principal component analysis (PCA) replacing SFAmix identified factors that differentiated biological sex (Figure 4)
- None of these PCA-derived factors were significantly associated with known biological pathways
- These results may be caused by the orthogonality constraints or lack of dense and sparse components that jointly restrict interpretability of the corresponding factors

Figure 4. PCA factor scores

ML applied to cfDNA WGS recapitulated observed differences in protein biomarker abundances.

- CEA, one of the proteins identified using SFAmix, demonstrated high concordance between protein abundance and estimated gene activation (Figure 5)
- Elevated levels of CEA protein and estimated gene activation were observed in late-stage CRC (Figure 5). (T test of stage 4 vs NEG: relative protein abundance *p*-value ≤ 5.65e-04; gene activation *p*-value ≤ 9.80e-04.)

Figure 5. Protein abundance and gene activation estimates of CEACAM5



- CEA is a tumor-associated antigen involved in cell adhesion, migration, tumor invasion and metastasis, which is a hallmark of late-stage cancer
- These findings were validated in an independent cohort (n=1392, including 612 CRCs and 780 non-CRC controls). No protein data was available; however, elevated gene activation in cfDNA was again observed in late-stage CRC. (T test of stage 4 vs NEG: *p*-value ≤ 1.61e-02.)

CONCLUSIONS

- Analysis using a latent factor model, SFAmix, on plasma protein abundances revealed differences based on biological sex in immune activation, providing a meaningful representation of the underlying biological processes within our CRC cohort
- The estimated gene activation using a CNN applied to sequenced cfDNA demonstrated how cfDNA latent representations can recapitulate protein data, with elevated levels of both CEA protein and estimated gene activation in late-stage CRC
- These methods can be applied to other cancer types to learn biologically meaningful latent representations and to characterize immune processes involved in cancer biology

REFERENCES

John St John, Erik Gafni, Brandon White, Ajay Kannan, Loren Hansen, Artur Jaroszewicz, Anshul Kundaje, and Nathan Boley. "Predicting gene expression from plasma cell-free DNA using both the fragment length and fragment position." AACR (2019).

Chuan Gao, Christopher D. Brown, and Barbara E. Engelhardt. "A latent factor model with a mixture of sparse and dense factors to model gene expression data with confounding effects." arXiv preprint arXiv:1310.4792 (2013).

Artin Armagan, Merlise Clyde, and David B. Dunson. "Generalized beta mixtures of Gaussians." Advances in neural information processing systems (2011).