Poster 4349

Predicting gene expression from plasma cell-free DNA using both the fragment length and fragment position

John St John[1], Erik Gafni[1], Brandon White[1], Ajay Kannan[1], Loren Hansen[1], Artur Jaroszewicz[1], Anshul Kundaje[2], Nathan Boley[1]

[1]Freenome; [2]Stanford University

BACKGROUND

- The ability to use a blood sample to determine the transcriptional state of cells that are releasing DNA into the bloodstream of a patient may be helpful in a variety of clinical applications, including the early detection of cancer
- Cell-free DNA (cfDNA) contains epigenetic signatures of the cells from which it was produced. As a result, cfDNA can be used to predict the gene expression state of cfDNA-producing cells. To date, the two published approaches used to predict gene expression largely ignore cfDNA fragment size[1,2]
- V-plots are a powerful visualization technique originally applied to MNase-seq data. These plots show the density of fragments of each length at a particular genomic location, and can provide single base pair resolution of nucleosomes as well as other proteins that protect DNA from digestion[3]
- cfDNA closely approximates MNase-seq data[2,4,5]; we therefore used V-plots as an information-rich input to our gene expression prediction model

OBJECTIVES

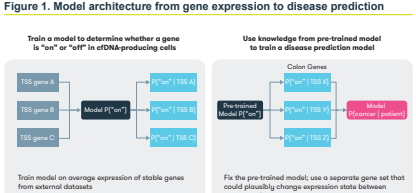
- To develop a gene expression prediction model that uses cfDNA fragment coverage and length to predict which genes are highly or lowly expressed in cfDNA-producing cells
- To apply this model of gene expression prediction to a set of colon-specific genes in order to detect colon cancer and adjacent colon-derived cfDNA, which is expected to be present in patients with advanced colorectal cancer (CRC)

METHODS

Sample collection

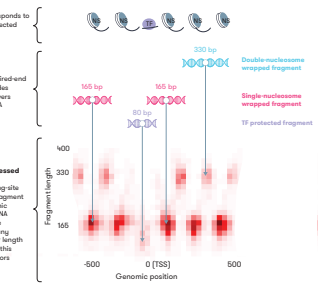
- De-identified plasma samples from patients with CRC (n=532) and non-cancer controls (n=234) were obtained from academic medical centers and commercial biobanks. CRC stage information was as follows: stage I (n=169), stage II (n=256), stage III (n=97), stage IV (n=6) and unknown stage information (n=4)

Figure 1. Model architecture from gene expression to disease prediction



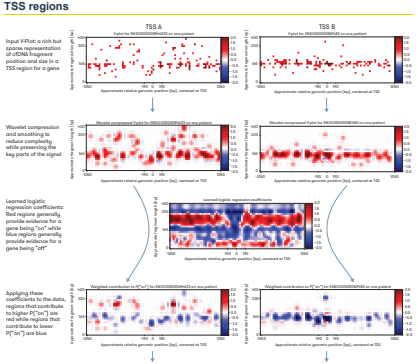
TSS - transcriptional start site (chose one per gene); p(X | Y) = probability of X, given Y

Figure 2. V-plots made from cfDNA capture DNA-protein associations and reflect transcriptional state



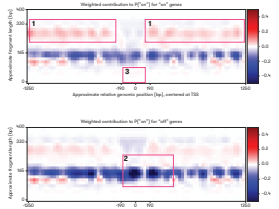
TF = Transcription Factor (small region protected)
NS = Nucleosome (large region protected)

Figure 3. Predicting gene expression from cfDNA-derived V-plots around TSS regions



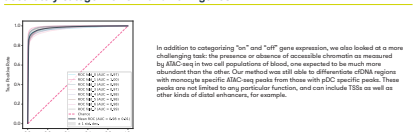
RESULTS

Figure 4. Interpretation of what the gene expression model learned



- Strongest drivers of predicting "on" are dinucleosome peaks flanking the TSS both up and downstream (**Figure 4** Region 1) and a relatively weak mononucleosome band
- Strongest drivers predicting "off" are mononucleosome positions (especially **Figure 4** Region 2) and a relatively weak dinucleosome band
- Although not always present, short sub-mononucleosome fragments at the TSS support an "on" prediction (**Figure 4** Region 3) (see the logistic regression coefficients in **Figure 3** for more evidence of this)

Figure 5. Classifiers using representations of fragment length and position accurately categorize "on" and "off" genes



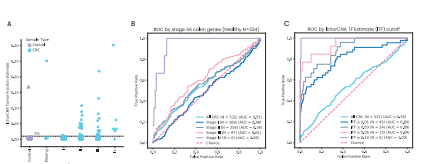
In addition to categorizing "on" and "off" gene expression, we also looked at a more challenging task: the presence or absence of accessible chromatin as measured by ATAC-seq in two cell populations of blood, one expected to be much more abundant than the other. Our method was still able to differentiate cfDNA regions with monocyte specific ATAC-seq peaks from those with pDC specific peaks. These peaks are not limited to any particular function, and can include TSSs as well as other kinds of distal enhancers, for example.

Easier task → Harder task

Method AUC (±SD)	Blood constitutive TSS <0.1FPKM vs >2SFPKM[1,3]	Blood constitutive TSS <0.1FPKM vs >1FPKM[1,3]	Monocytes (<6% of WBCs) vs pDCs (<1% of WBCs) specific ATAC-seq peaks[4]
2D Wavelet	0.98 ± 0.01	0.95 ± 0.02	0.75 ± 0.03
V-plot CNN	0.98 ± 0.01	0.95 ± 0.02	0.71 ±0.04
2D Wavelet (downsampled)	0.97 ± 0.01	0.93 ± 0.01	0.72 ± 0.02
Normalized TSS coverage	0.95 ± 0.02	0.91 ± 0.02	0.66 ± 0.05

Normalized TSS coverage only uses normalized fragment counts in "on" vs "off" genes to predict expression. This works because "on" genes have lower coverage (are less protected by nucleosomes) than "off" genes.[1]
FPKM - a normalized RNA-seq measurement of relative expression **F**ragments **P**er **K**ilobase of transcript per **M**illion mapped reads.
pDC - plasmacytoid dendritic cell
ROC - receiver operating characteristic
AUC - area under the receiver operating characteristic curve

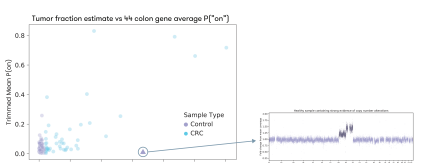
Figure 6. Tumor-targeted gene set enables classification of cancer samples as a function of cfDNA tumor fraction rather than stage



NOTE: "All" includes samples with missing stage information.

- For this approach we used 44 genes expressed in colon and not in blood cells as measured in the Roadmap Epigenomics Project.[7] We expect that colon genes will be expressed in colon cancer, as well as adjacent healthy colon tissue, which does not contribute substantial quantities of material to cfDNA in healthy individuals[10]
- IchorCNA-based TF estimates (ITF) increase with stage but most stage I-III CRC have low estimated ITF (<1%) (**Figure 6A**)
- Classification performance increases more strongly by tumor fraction than stage (**Figure 6B** & **6C**)

Figure 7. Average gene expression prediction can augment CNV-based tumor fraction estimation



- A high ITF non-cancer control displayed a low average probability of expression P("on") of the 44 colon genes, differentiating it from high ITF CRC samples (**Figure 7**). These copy number changes may either be germline or somatic but do not originate from colon/CRC DNA. Plausible sources include DNA from blood-cells or from a non-CRC tumor

CONCLUSIONS

- 2D representations of fragment length and location can be used to accurately predict extremes in gene expression (**Figure 5**)
- The method presented here can accurately predict whether a patient has cancer with high fractions of tumor-derived cfDNA, which are typically observed in later stages but can be observed at any stage of disease (**Figure 6**)
- Despite limited sensitivity in patients with low tumor fractions, one practical use for this method is in identifying cases where observed CNVs in cfDNA do not originate from the cancer of interest (**Figure 7**)

NEXT STEPS

- This approach could be used with different cell-type-specific gene sets to predict the tissue of origin of a cancer
- We are in the process of evaluating and verifying this approach on immune-derived signals, as well as combining with other analytes, for early cancer detection

ACKNOWLEDGEMENT

The authors gratefully acknowledge Dr. Andrew Godwin and the University of Kansas Cancer Center's Biospecimen Repository Core Facility staff funded in part by the National Cancer Institute Cancer Center Support Grant [P30 CA168524], National Health Services Research Scotland, Tayside Biorepository, Geneticist Inc., iSpecimen Inc., and Indivumed for support of this research by providing de-identified plasma samples. We also thank Signe Fransen, Girish Putcha and David Weinberg for their extensive suggestions, feedback, and editorial support.

REFERENCES

1. Ulz et al. *Nature Genetics*, 2016.
2. Snyder et al. *Cell*, 2016.
3. Henikoff et al. *PNAS*, 2011.
4. Ivanov et al. *BMC Genomics*, 2015.
5. Zhang et al. *Clinical Chemistry*, 2017.
6. Calderon et al. *bioRxiv*, 2018.
7. Roadmap Epigenomics Consortium. *Nature*, 2015.
8. ENCODE Project Consortium. *Nature*, 2012.
9. Adalsteinsson et al. *Nat. Commun*, 2017.
10. Moss et al. *Nat. Commun*, 2018.

Presented at the American Association for Cancer Research (AACR) Annual Meeting; March 29–April 3, 2019; Atlanta, GA