Inferring gene expression from cell-free DNA fragmentation profiles

Mohammad Shahrokh Esfahani, Emily G. Hamilton, Mahya Mehrmohamadi, Barzin Y. Nabet, Stefan K. Alig, Daniel A. King, Chloe B. Steen, Charles W. Macaulay, Andre Schultz, Monica C. Nesselbush, Joanne Soo, Joseph G. Schroers-Martin, Binbin Chen, Michael S. Binkley, Henning Stehr, Jacob J. Chabon, Brian J. Sworder, Angela Hui, Matthew J. Frank, Everett J. Moding, Chih Long Liu, Aaron M. Newman, James M. Isbell, Charles M. Rudin, Bob T. Li, David M. Kurtz, Maximilian Diehn#, Ash A. Alizadeh#

Table of Contents

Supplementary Notes:

Supplementary Notes.

1. Cell-free DNA features are correlated with gene expression.

Furthermore, TSS regions were distinguished from exonic and intronic by having the highest representation of subnucleosomal fragments (*P*<0.0001, **Extended Data Fig. 1c**).

We also tested whether the partially protected, subnucleosomal cfDNA fragments that are 100-150 bases long could derive from tumor tissues. As previously described, in patients with non-small cell lung cancers (NSCLC) [1], we observed cfDNA molecules harboring tumor mutations to have significantly higher representation of subnucleosomal fragments than their wild-type counterparts (*P*<6E-08, **Extended Data Fig. 1d**). Therefore, the prevalence of subnucleosomal fragments observed in cfDNA correlate with expression levels and can derive from solid tumor origin.

We also examined whether the distance from the TSS impacts correlations between cfDNA fragmentomic features and gene expression. When considering the ~20kb region flanking each promoter, we observed the peak correlation between cfDNA PFE and gene expression to be centered at the TSS. However, in comparison to NDR, correlation of PFE with gene expression had broader dispersion and extended into regions flanking the TSS (**Fig. 1e**).

2. Subnucleosomal cfDNA fragments.

We tested whether the partially protected, subnucleosomal cfDNA fragments that are 100-150 bases long could derive from tumor tissues. As previously described, in patients with non-small cell lung cancers (NSCLC) [1], we observed cfDNA molecules harboring tumor mutations to have significantly higher representation of subnucleosomal fragments than their wild-type counterparts (*P*<6E-08, **Extended Data Fig. 1d**). Therefore, the prevalence of subnucleosomal fragments observed in cfDNA correlate with expression levels and can derive from solid tumor origin.

3. Validation of gene expression inference model.

We also examined the robustness of our gene expression inference model by considering its performance on cfDNA data from different subjects, and various independent ground truth transcriptome data sources obtained by RNA-Seq. We therefore profiled two additional cfDNA samples from two healthy adults by deep whole genome sequencing. As ground truth, we also profiled the matched leukocytes of these two individuals by RNA sequencing. In both cases, we found expression inferences from cfDNA WGS using our model to be strongly and significantly correlated across the transcriptome as measured by RNA-Seq TPM (*r*=0.86, and *r*=0.91 with *P*<2.2E-16, **Extended Data Figs. 2c-d**). Therefore, the generalized linear model described here appears robust for estimating gene expression levels from cfDNA and is not substantially impacted by the source of cfDNA, or by the ground-truth transcriptome data employed for training.

To validate the performance of our model in healthy controls versus patients with cancer, we next re-analyzed genome-wide cfDNA profiling data from 40 healthy adults and 46 patients with early-stage lung cancers that were previously profiled by WGS at ~20-40x coverage[2]. We observed similar performance for predicting leukocyte gene expression levels when considering the average cfDNA meta-profile across the genome in the 40 healthy subjects (**Extended Data Figs. 2e-f**). When considering groups of 10 genes across the transcriptome, Pearson correlations between model predicted expression and expected RNA expression levels from PBMCs remained ~0.85. Nevertheless, gene expression levels inferred from plasma cfDNA fragmentomic profiles of lung cancer patients were lower compared to PBMC transcriptomes (*P*=0.018; **Extended Data Fig. 2g**). Hypothesizing that the lower correlation in lung cancer may be driven by an increased contribution of lung cancer-derived fragments, we used tumor fraction estimates by ichorCNA[3] and observed a significant negative correlation with inferred leukocyte expression levels (*r*=-0.69, *P*= 0.0005, **Extended Data Fig. 2h)**. This experiment demonstrates that tumor-derived cfDNA can substantially reduce the contribution of the leukocyte compartment to the cell-free nucleic acid pool, and this contribution can be measured by inferring tissue-specific gene expression from cfDNA when tumor burden is high.

4. EPIC-seq selector design.

We then identified subtype-specific genes by evaluating those differentially expressed in NSCLC adenocarcinoma (LUAD) versus squamous cell carcinoma (LUSC) and DLBCL germinal center B- (GCB) versus activated B-cell (ABC) like subtypes. Specifically, we identified 69 differentially expressed genes (DEGs) when stratifying 1,156 NSCLC tumors by histological subtype from The Cancer Genome Atlas (TCGA; *n*=601 LUAD[4] vs *n*=555 LUSC[5], **Fig. 3b, Supplementary Table 3**). We separately identified 44 DEGs when stratifying 381 DLBCL tumors by molecular cell-of-origin (COO) subtype from prior publications (n=138 GCB vs n=243 ABC tumors[6], **Fig. 3c, Supplementary Table 3**). In addition to these 113 genes for classification of lung cancers and lymphoma subtypes, we also included 50 genes that are differentially expressed in leukocyte subsets[7] as well as 16 genes as additional controls (**Methods**).

For each gene of interest, we designed probes to capture the ~2kb region flanking the TSS, then profiled plasma cfDNA from by deep sequencing of the targeted regions to a median ~2,000x unique depth of coverage as previously described[1, 8]. In cfDNA fragmentomic profiles captured by WGS, we observed marginal gains in transcriptome wide correlations beyond ~500x nominal coverage depth (**Fig. 1h)**. Nevertheless, for our EPIC-Seq experiments and our modestly sized panel, we targeted ~2000x unique depth (~4-fold excess) for three reasons: (1) to guarantee saturation of the correlation plateau, (2) to avoid any gene-to-gene variability in accuracy of EPIC-

Seq predictions of expression levels that might otherwise be attributable to spurious differences in depth variability due to non-uniform hybrid capture of the TSS regions of genes of interest, and (3) to address the lower partial concentration of cfDNA from non-hematopoietic tissues in circulation.

5. EPIC-seq lung cancer classifier evaluation.

Epigenetic signals in cfDNA captured by our EPIC-Seq lung cancer classifier were significantly correlated with total metabolic tumor volumes (MTV), as measured by ^{18}Fluorodeoxyglucose (FDG) uptake in combined positron emission tomography and computed tomography studies (PET/CT; ρ=0.67; *P*=0.04; **Extended Data Fig. 4a**), consistent with higher ctDNA concentrations in patients with larger tumor burdens[1, 9]. We also compared lung cancer epigenetic signals from EPIC-Seq in cfDNA with corresponding lung tumor-derived mutation signals from ctDNA separately measured by CAPP-Seq[8]. Here again, EPIC-Seq lung signals in cfDNA seemed to capture tumor burden, as we observed significant correlation with the mean allelic fractions (AF) of tumor-derived somatic mutations measured by CAPP-Seq on the same specimens (ρ=0.5, *P*=3E-5; **Extended Data Fig. 4b**).

6. EPIC-Seq DLBCL classifier evaluation

For patients with available PET/CT scans, we also observed a significant trend for scores from the epigenetic classifier in distinguishing patients with high versus low tumor burden[10] as measured by total MTV (Wilcoxon *P*=0.015; **Extended Data Fig. 5a**). This same trend was also observed in the validation set (**Extended Data Fig. 5b**).

To further evaluate how EPIC-Seq scores reflect tumor burden in cfDNA, we compared them with the mean allele fractions (AFs) of mutations previously measured by CAPP-Seq on the same blood specimens[11, 12]. Notably, DLBCL epigenetic scores determined by EPIC-Seq were strongly correlated with the mean mutant AFs determined by CAPP-Seq (ρ=0.66, *P*<2E-16; **Extended Data Fig. 5c**).

7. Concordance of EPIC-seq inferred expression with tumor in the context of DLBCL prognosis.

Therefore, we wished to evaluate the utility of EPIC-Seq for noninvasively measuring expression of genes with prognostic associations in DLBCL. Using univariate Cox proportional hazard regression models, we tested the prognostic value of individual genes using pre-treatment blood plasma from 69 patients and used Z-scores to measure the relative strength of these associations. We first assessed the prognostic concordance of our results in blood plasma against primary tumor specimens by examining the correlation between our EPIC-Seq results with those described in 3 recent tumor expression profiling studies that relied on surgical DLBCL tissue specimens[6, 13, 14]. When comparing the prognostic value of genes profiled in this manner, we observed a significant correlation of Z-scores from our study using plasma cfDNA with prior studies using tumor RNA (*P*=0.026; **Extended Data Fig. 5i**).

8. Pre-analytical Factors

Importantly, we did not observe a significant impact of several pre-analytical factors on cfDNA fragmentation entropy measurements, including blood collection tube types, the time between phlebotomy and plasma isolation, and the number of PCR cycles (**Extended Data Fig. 6**). Separately, we observed relatively modest impact of several factors that might confound accuracy of expression estimates derived from cfDNA entropy measurements, including corrections for GC fraction and presence of somatic copy number aberrations (**Extended Data Fig. 6**). Finally, we developed a mechanistic framework for how cfDNA fragmentation mirrors activity level of expressed genes in human tissues (**Extended Data Fig. 7a-c**). Using this model framework, we used simulations to explore the parameters influencing the likelihood of detection of expression of a given gene of interest within cfDNA as a function of tumor burden (**Extended Data Fig. 7d**). Nevertheless, future studies will need to address the relative strengths and weaknesses of epigenetic signals that can be gleaned from cfDNA and how this approach compares to direct cfRNA measurements.

Supplementary Table Legends.

Supplementary Table 1. List of samples analyzed or profiled in this study. Whole-genome (n=116) and whole-exome (n=39) sequencing of cell-free DNA samples were used for the discovery of PFE, training the gene expression inference model and its validation. The WGS data were either profiled in this study (n=30) or downloaded from Zviran et al. (EGA accession number EGAS00001004406). The WES data were profiled in this study (n=39). Cell-free DNA from 288 subjects were profiled using EPIC-Seq.

Supplementary Table 2. Gene groups- average expression values of genes in each group in PBMC, normalized PFE, NDR, OCF, WPS, and MDS in the deep WGS sample. Each row in this table corresponds to one group of genes (10 genes per group, n=1,748 groups), when they are sorted according to the expression level in PBMC. Five columns, corresponding to different fragmentomic features (PFE [this study], NDR, OCF, WPS, and MDS) are shown in this table.

Supplementary Table 3. TSSs in the EPIC-Seq selector. Each row corresponds to one TSS in the EPIC-seq sequencing panel (‘selector’).

Supplementary Table 4. EPIC-Seq samples’ clinical characteristics and scores corresponding to different classifiers. EPIC-Seq was applied to 373 samples, of which 329 passed the QC steps and were used to show the utility of the inferred gene expression in different applications: detection, subtype classification, and patient response to treatment prediction.

References

1. Chabon, J.J. et al. Integrating genomic features for non-invasive early lung cancer detection. *Nature* **580**, 245-251 (2020).
2. Zviran, A. et al. Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring. *Nat Med* **26**, 1114-1124 (2020).
3. Adalsteinsson, V.A. et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat Commun* **8**, 1324 (2017).
4. Cancer Genome Atlas Research, N. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543-550 (2014).
5. Cancer Genome Atlas Research, N. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519-525 (2012).
6. Schmitz, R. et al. Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. *N Engl J Med* **378**, 1396-1407 (2018).
7. Newman, A.M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* **12**, 453-457 (2015).

8. Newman, A.M. et al. Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat Biotechnol* **34**, 547-555 (2016).
9. Newman, A.M. et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat Med* **20**, 548-554 (2014).
10. Cottereau, A.S. et al. Molecular Profile and FDG-PET/CT Total Metabolic Tumor Volume Improve Risk Classification at Diagnosis for Patients with Diffuse Large B-Cell Lymphoma. *Clin Cancer Res* **22**, 3801-3809 (2016).
11. Scherer, F. et al. Distinct biological subtypes and patterns of genome evolution in lymphoma revealed by circulating tumor DNA. *Sci Transl Med* **8**, 364ra155 (2016).
12. Kurtz, D.M. et al. Circulating Tumor DNA Measurements As Early Outcome Predictors in Diffuse Large B-Cell Lymphoma. *J Clin Oncol* **36**, 2845-2853 (2018).
13. Chapuy, B. et al. Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nat Med* **24**, 679-690 (2018).
14. Ennishi, D. et al. Double-Hit Gene Expression Signature Defines a Distinct Subgroup of Germinal Center B-Cell-Like Diffuse Large B-Cell Lymphoma. *J Clin Oncol* **37**, 190-201 (2019).