Recent research into the clinical **potential of cell-free DNA.**

Review

Inferring expressed genes by whole-genome sequencing of plasma DNA.
P Ulz, GG Thallinger, M Auer, et al. 2016. Nature Genetics 48(10):1273-8

Key findings:

- Machine learning analysis of cell-free DNA (cfDNA) sequencing coverage can be used to infer gene-expression patterns from epigenetic, nucleosome “footprints”
- This technique represents a new and noninvasive way to detect and monitor tumor transcriptome dynamics over time, with a high degree of sensitivity and accuracy

Continuing research into the origin and structure of cfDNA may lead to further breakthroughs in early-stage disease detection, disease monitoring, and therapeutic response prediction.

freenome

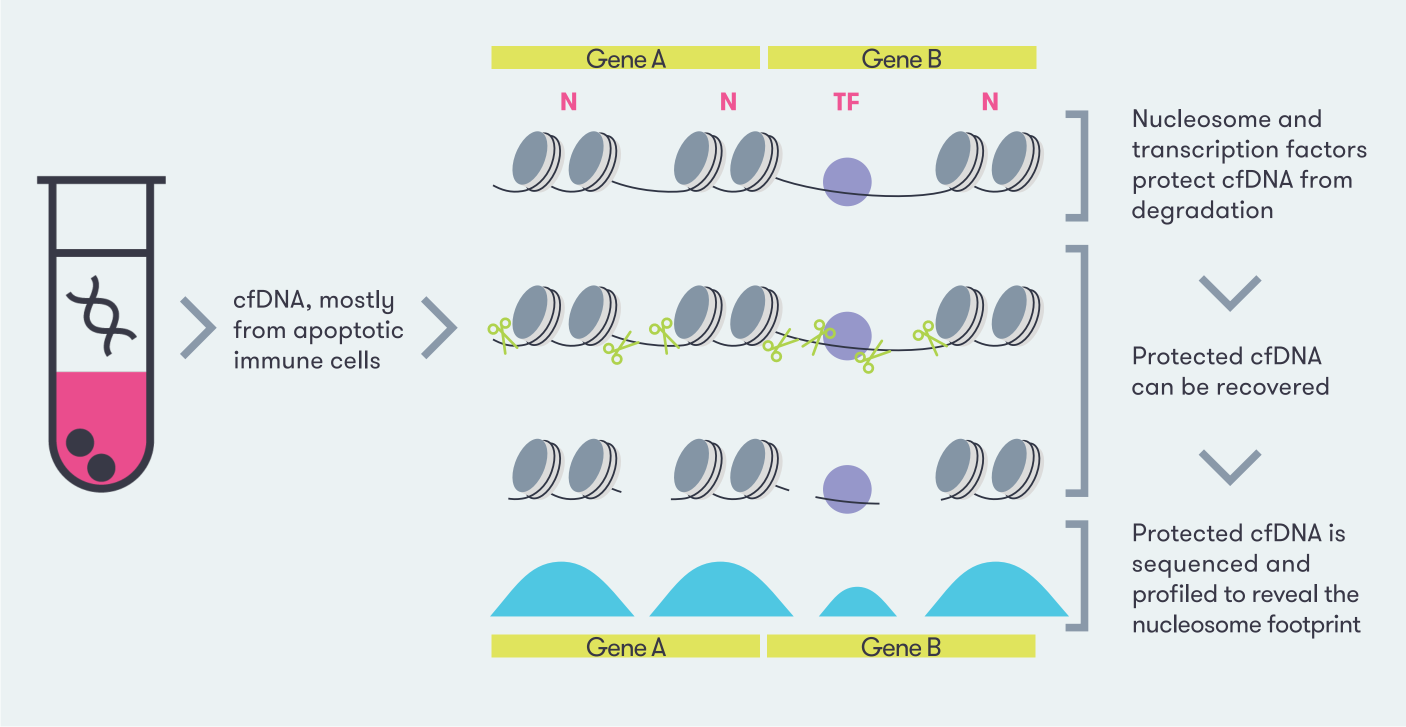
Background

Cell-free DNA is composed of fragments of circulating DNA derived from apoptotic cells. cfDNA that resists degradation long enough for analysis consists primarily of sequences that were bound within, and protected by, nucleosomes—the DNA protein complexes within which DNA is organized.*

Epigenetic patterns of nucleosome positioning, or "footprints," at a particular gene vary depending on whether the gene is actively expressed or not. Actively expressed genes are not tightly packaged within nucleosomes, allowing transcription to occur more readily. Given their lack of protection, actively expressed sequences are expected to be underrepresented in cfDNA.

Nucleosome footprints and read depth

Recovered cfDNA fragments average 166 base pairs, approximately equal to one nucleosome "wrap". In the graphic below, DNA sequences that are bound within nucleosomes (grey discs) or by transcription factors (purple spheres) are protected from degradation by enyzmes called nucleases (green scissors), leading to higher representation in sequencing coverage, or, read depth (blue).



Nucleosome and transcription factors protect cfDNA from degradation

Protected cfDNA can be recovered

Protected cfDNA is sequenced and profiled to reveal the nucleosome footprint

*Transcription factors (TF) can also help to preserve cfDNA. See Useful Terms on p.3 for definitions of scientific terms.

Methods

To investigate whether cfDNA nucleosome footprinting can help predict gene expression, the authors of the study:

1. Compared differences in cfDNA sequencing coverage between transcriptionally silent and actively expressed genes
2. Assessed the sensitivity and accuracy of gene expression predictions based on cfDNA sequencing coverage analysis
3. Determined whether blood samples from patients with cancer were informative for expressed cancer driver genes, as predicted

Useful Terms

Epigenetics
The study of genomic changes that do not involve changes to the underlying DNA sequence.

Gene expression
The process of copying (transcribing) DNA into RNA and, usually, subsequent translation into protein.

Gene promoters
Sequences of DNA near the transcription start site that do not code for proteins but, instead, initiate the expression of a gene, e.g., through binding of transcription factors.

Nucleosome
Multi-protein complexes that contain DNA, organizing it into tightly bound coils and protecting bound sequences from easy access by RNA polymerase for transcription.

Sequencing coverage/Read depth
Number of sequencing counts for a specified area of DNA within a given sample.

Transcription
The first step in gene expression: RNA polymerase and other factors generate an RNA copy (messenger RNA) based on the DNA sequence in the genome.

Transcription factor (TF)
Transcription factors are proteins involved in the process of converting DNA into RNA. The action of transcription factors allows for unique expression of each gene.

Transcription start site (TSS)
A specific nucleotide at the start of a gene sequence where transcription begins.

Results

1. cfDNA sequencing coverage at active promoters reflected reduced nucleosome binding

cfDNA sequencing coverage, or read depth, at transcription start sites (TSS) reveals a nucleosome footprint pattern associated with actively expressed genes (e.g. housekeeping genes) similar to those reported in previous research.[1]

As shown in Ulz et al. Figure 2, active genes show lower relative sequencing coverage at the TSS, and a wave of alternating high and low coverage in the 2,000 base pairs (2kb) around the TSS.

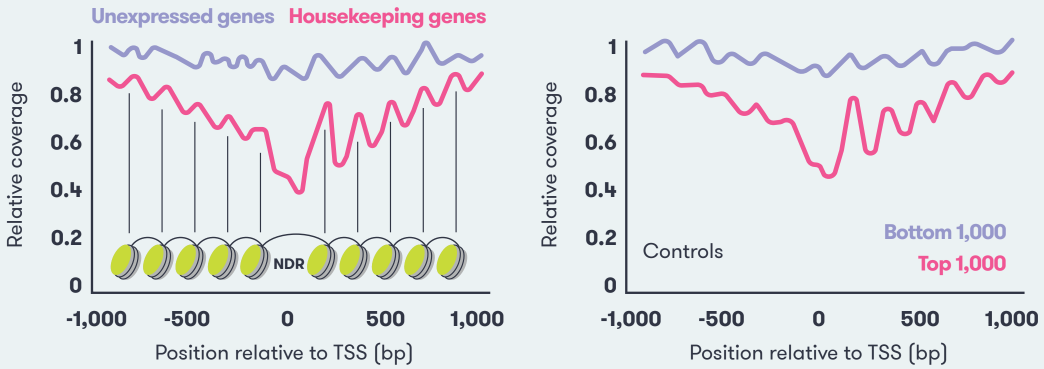


Figure 2. Sequencing coverage results for plasma samples from 104 healthy donors. **Left**: Relative sequencing coverage at promoters of housekeeping genes (pink, n=3,804) and unexpressed genes (purple, n=670). **Right**: Relative sequencing coverage for the 1,000 most expressed genes (pink) and 1,000 least expressed genes (purple) in these samples.

cfDNA sequencing coverage for actively expressed genes followed a characteristic pattern associated with reduced nucleosome binding

[1] ENCODE Project Consortium, 2012.

Results continued

2. cfDNA sequencing coverage sensitively and accurately predicted gene expression

Ulz et al. used machine learning to classify gene expression as high or low based on sequencing coverage at the nucleosome-depleted region (NDR) near the TSS and in the surrounding 2kb (2K-TSS).

The resulting predictions were both sensitive and accurate:

- For the top and bottom 1000 genes, sensitivity = 0.81 and accuracy = 0.83
- For the top and bottom 100 genes, both sensitivity and accuracy = 0.91

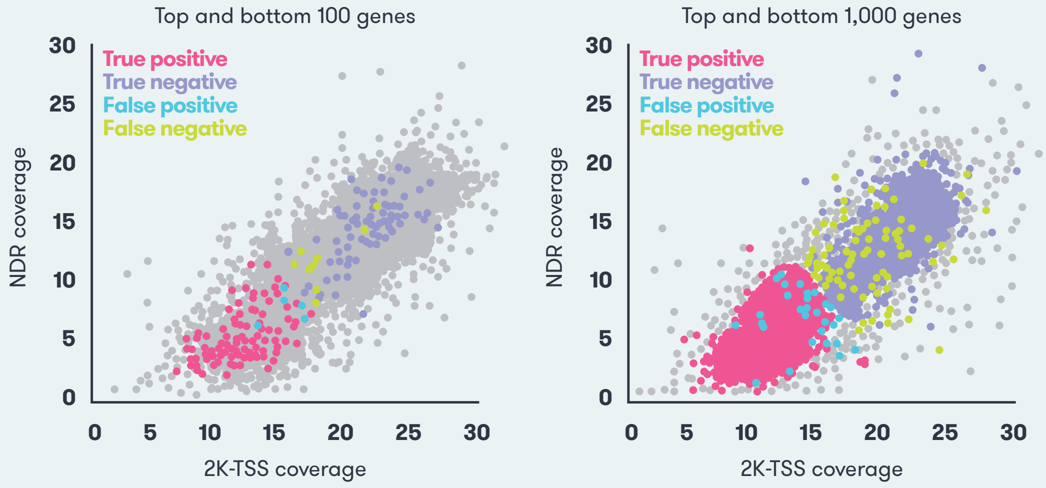


Figure 3b. Normalized read-depth for the 2kb of DNA near the TSS (2K-TSS coverage) versus coverage of the nucleosome-depleted region (NDR) at the TSS for the top and bottom most expressed 100 genes **(left)** and 1000 genes **(right)**, as predicted using support vector machines (SVMs). True positives (pink) and true negatives (purple), false-positives (blue), and false-negatives (green) are indicated. Ten-fold cross-validation of all results yielded a total accuracy of 85.8%.

cfDNA nucleosome footprint analysis differentiated high vs. low gene expression with up to 91% accuracy and 91% sensitivity

Results continued

3. cfDNA nucleosome footprint analysis identified highly-expressed genes in patients with cancer

cfDNA nucleosome-footprint analysis of blood samples from 2 patients with breast cancer accurately predicted 86-88% of the 100 most highly-expressed genes in the primary tumor, as determined by RNA sequencing of tissue biopsy material. Highly-expressed genes included known cancer-driver genes, including *ERBB2*.

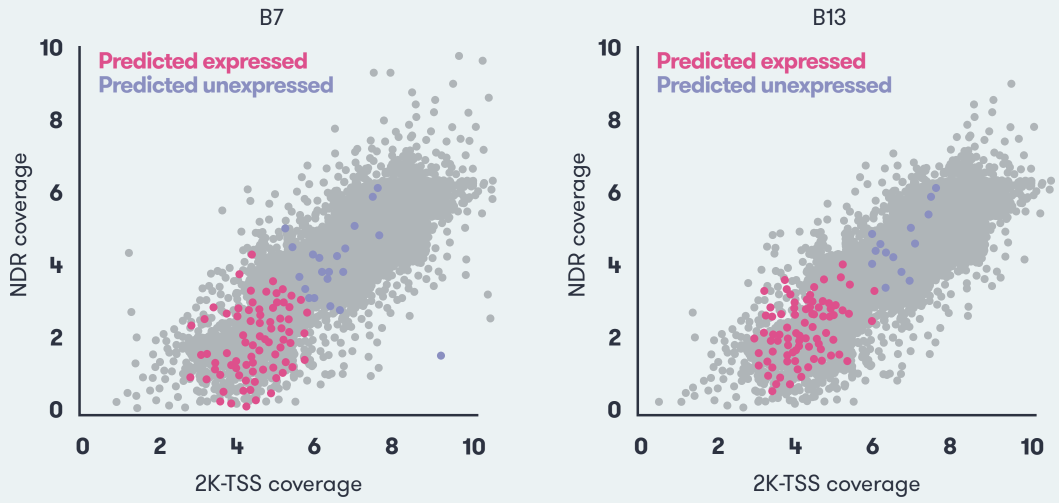


Figure 5b. Identification of expressed genes in cancer from patient blood samples. Classification accuracy for cfDNA coverage analysis for the 100 most expressed genes on chromosomes 1q (patient B7, left) and 8p11-qter (patient B13, right), as assessed by RNA sequencing of patient's primary tumor. True positives (pink) and false positive results (purple) are indicated.

Gene expression predictions based on cfDNA nucleosome-footprint analysis were well correlated with primary tumor results and included known cancer driver genes

Discussion

Peter Ulz and the other study authors show that nucleosome footprints—inferred from cfDNA sequencing coverage and analyzed through machine learning techniques—can be used to develop classifiers to sensitively and accurately predict gene expression in individuals with cancer.

Rather than rely on detecting mutations in circulating tumor cells—a needle-in-the-haystack approach—this initial investigation into the epigenetic dynamics of nucleosome footprinting suggests that more sensitive, accurate, and holistic options for cfDNA analysis may soon be available.

While the present analysis focused on patients with a relatively high tumor fraction, prior research has demonstrated that, in those with early-stage disease, most circulating cfDNA is derived from immune cells.[2] Though nucleosome footprints are known to vary by cell type,[3] further research is needed to assess whether the techniques outlined in this paper may be used to similarly infer epigenetic changes in immune cells and provide clinicians with valuable insights into cancer's interaction with the rest of the body.

Clinical Implications

Freenome, working independently and with collaborators, is pioneering the use of artificial intelligence to identify predictive, genome-wide patterns in circulating, cell-free biomarkers, including cfDNA.

In addition to enabling earlier detection of cancer, a more holistic approach to cf biomarker analysis has the potential to reveal new pathways for drug development and response prediction, helping clinicians optimize diagnosis and treatment for patients with a wide variety of tumor types.

Sign up for news and research updates at Freenome.com

[2] Lui, Y.Y. et al. Clin Chem. 2002 Mar;48(3):421-7.
[3] Valouev, A. et al. Nature. 2011 May 22;474(7352):516-20. doi: 10.1038/nature10002.