

SUMMARY

[0006] The present disclosure provides methods and systems for assessing (e.g., modeling) transcription factor (TF) binding sites (TFBSs) and using TFBS information to detect, assess, diagnose, and analyze disease states and identify treatment responsiveness.

[0007] Next-generation sequencing-based genome-wide assays may be used to provide TF-binding patterns and the associated chromatin architecture. As nucleosomes and sequence-specific TFs bind regulatory deoxyribonucleic acid (DNA) regions in a mutually exclusive fashion, TFs either compete or interact with nucleosomes, which affects both their occupancy and positioning. In a given population of cells, nucleosome occupancy refers to the average number of nucleosomes measured within a specified genomic region, whereas nucleosome positioning indicates the probability of a reference point on a nucleosome (usually the dyad, e.g., the midpoint of a canonical nucleosome) existing at a specific genomic coordinate.

[0008] Cell-free circulating nucleic acid, such as cell-free DNA (cfDNA), may provide an easily-accessible source of nucleic acid for TFBS analysis. Such cfDNA may be the product of a digestion process that preferentially degrades DNA that is not protected by proteins, such as the histone complex. Cell-free DNA coverage patterns may reflect nucleosome positioning and occupancy caused by transcription factors actively binding the genome. These nucleosome occupancy patterns measured through cfDNA may then be used to infer the activity of TFs in the normal and tumor genomes.

[0009] TFs may bind preferentially within open chromatin, which may affect nucleosome positioning. Circulating cell-free DNA from blood plasma may represent mono-nucleosomal DNA, and nucleosome plasma footprints may be informative regarding TFBS.

[0010] The present disclosure provides methods and systems for charting of nucleosome positions from cfDNA to provide information about TFs for applications relating to disease identification, prediction, staging, and/or identifying treatment responsiveness. Methods and systems are described herein for using transcription factor information determined from nucleosome footprints in nucleic acid molecules (e.g., cfDNA). Information from nucleosome footprints in nucleic acid molecules may be used to evaluate, assess, detect, and diagnose diseases such as cancers. In some examples, the information may be featurized and used as inputs into machine learning models useful in many of these applications such as disease identification, prediction, staging and identifying treatment responsiveness.

[0011] In an embodiment, the present disclosure provides a computer-implemented method to determine a transcription factor binding profile in a nucleic acid sample from a subject, the

method comprising: (a) providing a set of sequence reads from deoxyribonucleic acid (DNA) extracted from the subject; (b) using said set of sequence reads to generate a coverage pattern for a transcription factor; (c) processing the coverage pattern to provide a signal; and (d) processing the signal with a reference signal, wherein the signal and the reference signal have different frequencies, thereby determining a transcription factor binding profile in the sample.

[0012] In some examples, the DNA is cell-free DNA.

[0013] In some examples, (c) comprises using a low-pass filter. In some examples, (c) comprises using a Savitzky-Golay filter.

[0014] In some examples, the subject is a human.

[0015] In a first aspect, the present disclosure provides a computer-implemented method for detecting a presence or absence of a disease in a subject, the method comprising: (a) providing a set of sequence reads from deoxyribonucleic acid (DNA) extracted from the subject; (b) using said set of sequence reads to generate a coverage pattern for a transcription factor; (c) processing the coverage pattern to provide a signal; and (d) processing the signal with a reference signal, wherein the signal and the reference signal have different frequencies, thereby detecting said presence or absence of said disease in said subject.

[0016] In some examples, the DNA is cell-free DNA.

[0017] In some examples, the disease is cancer.

[0018] In some examples, (b) comprises aligning the set of sequence reads to a reference sequence to provide an aligned sequence pattern, selecting regions of the aligned sequence pattern that correspond to binding sites of the transcription factor, and normalizing the aligned sequence pattern in the regions. In some examples, (d) comprises calculating an accessibility score for each of the binding sites of the transcription factor.

[0019] In some examples, (c) comprises using a low-pass filter. In some examples, (c) comprises using a Savitzky-Golay filter.

[0020] In some examples, the subject is a human.

[0021] In some examples, the transcription factor is a cancer-specific transcription factor. In some examples, the transcription factor is selected from the group consisting of GRH-L2, ASH-2, HOX-B13, EVX2, PU.1, Lyl-1, Spi-B, FOXA1, HNF-1A, HNF-4A, HNF-4G, and DLX-2.

[0022] In some examples, the accessibility scores for at least 2, or at least 5, or at least 10, or at least 15, or at least 20, or at least 25 transcription factor binding sites are determined and inputted into a machine learning model to train a classifier capable of distinguishing between healthy subjects and cancer patients, between disease progressors and non-progressors, between a

examples, the trained classifier comprises a trained machine learning classifier. In some examples, the trained machine learning classifier comprises a supervised machine learning algorithm. In some examples, the supervised machine learning algorithm comprises one or more of: a regression, a support vector machine, a tree-based method, a neural network, and a random forest.

[0039] In another embodiment, the present disclosure provides methods to allow classification of patients by tumor type, including, for example, tumor subtypes (e.g., subtypes of prostate cancer, colorectal cancer, breast cancer, lung cancer), or tumor stage, which may have important clinical implications for patient management including treatment planning and responsiveness. Accordingly, the methods provided herein for mapping tumor-specific transcription factor binding *in vivo* based on patient samples (e.g., blood, plasma, or serum samples), thereby making a key part of the noncoding genome amenable for clinical analysis.

[0040] In some examples, the method comprises distinguishing subtypes of disease.

[0041] In some examples, the method comprises distinguishing subtypes of cancer.

[0042] In some examples, the method comprises distinguishing subtypes of prostate cancer, colorectal cancer, breast cancer, and lung cancer.

[0043] In some examples, the method comprises distinguishing prostate cancer subtype, e.g., among patients having prostate adenocarcinoma or small-cell neuroendocrine prostate cancer.

[0044] In some examples, the method comprises distinguishing stage of cancer (e.g., among stage I, II, III, and IV cancers).

[0045] In some examples, the method comprises distinguishing stage I and II cancers from stage III and IV cancers.

[0046] In some examples, transcription factors GRHL2, FOXA1, HOXB13, AR, and NKX3-1 are associated with increased accessibility scores or open chromatin accessibility in patients with prostate adenocarcinoma.

[0047] In some examples, an open accessibility of transcription factors GRHL2, FOXA1, HOXB13, AR, and NKX3-1 is indicative of prostate adenocarcinoma.

[0048] In some examples, transcription factors REST, GRHL2, FOXA1, HOXB13, AR, and NKX3-1 is associated with decreased or closed chromatin accessibility in patients with small-cell neuroendocrine prostate cancer.

[0049] In some examples, a decreased accessibility of transcription factors REST, GRHL2, GRHL3, FOXA1, FOXA2, GATA2, GATA3, HOXB13, AR, and NKX3-1 is indicative of small-cell neuroendocrine prostate cancer.

[0050] In some examples, an increased accessibility of transcription factors GLIS1, SOX2, and SOX11 are indicative of small-cell neuroendocrine prostate cancer.

[0051] In a second aspect, the present disclosure provides a system comprising a computing device comprising at least one computer processor, an operating system configured to perform executable instructions, a memory, and a computer program including instructions executable by the computing device to provide a computer application for detecting a presence or absence of a disease in a subject, the computer application comprising: a sequence module programmed to obtain a set of sequence reads from deoxyribonucleic acid (DNA) extracted from the subject; a coverage module programmed to use the set of sequence reads to generate a coverage pattern for a transcription factor; a signal module programmed to process the coverage pattern to provide a signal; a detection module programmed to process the signal with a reference signal, wherein the signal and the reference signal have different frequencies, thereby detecting the presence or absence of the disease in the subject.

[0052] In a third aspect, the present disclosure provides a non-transitory computer-readable medium comprising machine-executable code that, upon execution by one or more computer processors, implements a method for detecting a presence or absence of a disease in a subject, the method comprising: (a) providing a set of sequence reads from deoxyribonucleic acid (DNA) extracted from the subject; (b) using said set of sequence reads to generate a coverage pattern for a transcription factor; (c) processing the coverage pattern to provide a signal; and (d) processing the signal with a reference signal, wherein the signal and the reference signal have different frequencies, thereby detecting said presence or absence of said disease in said subject.

[0053] In a fourth aspect, the present disclosure provides a system for detecting a presence or absence of a disease in a subject, the system comprising: a database comprising a set of sequence reads from deoxyribonucleic acid (DNA) extracted from the subject; and one or more computer processors operatively coupled to the database, wherein the one or more computer processors are individually or collectively programmed to: (a) use the set of sequence reads to generate a coverage pattern for a transcription factor; (b) process the coverage pattern to provide a signal; (c) process the signal with a reference signal, wherein the signal and the reference signal have different frequencies, thereby detecting the presence or absence of the disease in the subject.

[0054] In a fifth aspect, the present disclosure provides a computer-implemented method for monitoring a progression or regression of a disease in a subject, the method comprising: (a) providing a first set of sequence reads from deoxyribonucleic acid (DNA) extracted from the subject at a first time and a second set of sequence reads from DNA extracted from the subject at

[0064] In some examples, the transcription factor is a cancer-specific transcription factor.

[0065] In some examples, the transcription factor is selected from the group consisting of GRHL2, ASH-2, HOXB13, EVX2, PU.1, Lyl-1, Spi-B, and FOXA1.

[0066] In some examples, the transcription factor is selected from the group consisting of HNF-1a, HNF-4a, HNF-4g, EVX-2 and DLX-2.

[0067] In some examples, the method further comprises, based on (f), adjusting a therapeutic regimen for the disease in the subject.

[0068] In a sixth aspect, the present disclosure provides a system comprising a computing device comprising at least one computer processor, an operating system configured to perform executable instructions, a memory, and a computer program including instructions executable by the computing device to provide a computer application for monitoring a progression or regression of a disease in a subject, the computer application comprising: a sequence module programmed to obtain a first set of sequence reads from deoxyribonucleic acid (DNA) extracted from the subject at a first time and a second set of sequence reads from DNA extracted from the subject at a second time that is later than the first time; a coverage module programmed to use the first set of sequence reads to generate a first coverage pattern for a transcription factor and use the second set of sequence reads to generate a second coverage pattern for the transcription factor; a signal module programmed to process the first coverage pattern to provide a first signal, and process the second coverage pattern to provide a second signal; a first processing module programmed to process the first signal with a reference signal, wherein the first signal and the reference signal have different frequencies; a second processing module programmed to process the second signal with the reference signal, wherein the second signal and the reference signal have different frequencies; and a monitoring module programmed to, based on the processing of the first signal and the second signal with the reference signal, monitor the progression or regression of the disease in the subject.

[0069] In a seventh aspect, the present disclosure provides a non-transitory computer-readable medium comprising machine-executable code that, upon execution by one or more computer processors, implements a method for monitoring a progression or regression of a disease in a subject, the method comprising: (a) providing a first set of sequence reads from deoxyribonucleic acid (DNA) extracted from the subject at a first time and a second set of sequence reads from DNA extracted from the subject at a second time that is later than the first time; (b) using the first set of sequence reads to generate a first coverage pattern for a transcription factor and using the second set of sequence reads to generate a second coverage pattern for the transcription factor;

(c) processing the first coverage pattern to provide a first signal and processing the second coverage pattern to provide a second signal; (d) processing the first signal with a reference signal, wherein the first signal and the reference signal have different frequencies; (e) processing the second signal with the reference signal, wherein the second signal and the reference signal have different frequencies; and (f) based on the processing of the first signal and the second signal with the reference signal, monitoring the progression or regression of the disease in the subject.

[0070] In an eighth aspect, the present disclosure provides a system for monitoring a progression or regression of a disease in a subject, the system comprising: a database comprising a first set of sequence reads from deoxyribonucleic acid (DNA) extracted from the subject at a first time and a second set of sequence reads from DNA extracted from the subject at a second time that is later than the first time; and one or more computer processors operatively coupled to the database, wherein the one or more computer processors are individually or collectively programmed to: (a) use the first set of sequence reads to generate a first coverage pattern for a transcription factor and use the second set of sequence reads to generate a second coverage pattern for the transcription factor; (b) process the first coverage pattern to provide a first signal, and process the second coverage pattern to provide a second signal; (c) process the first signal with a reference signal, wherein the first signal and the reference signal have different frequencies; (d) process the second signal with the reference signal, wherein the second signal and the reference signal have different frequencies; and (e) based on the processing of the first signal and the second signal with the reference signal, monitor the progression or regression of the disease in the subject.

[0071] In another embodiment, the present disclosure provides a system to determine a transcription factor binding profile in a nucleic acid sample from a subject, the system comprising a processor configured to: (a) analyze a set of sequence reads from deoxyribonucleic acid (DNA) extracted from the subject; (b) using the set of sequence reads to generate a coverage pattern for a transcription factor; (c) processing the coverage pattern to provide a signal; and (d) processing the signal with a reference signal, wherein the signal and the reference signal have different frequencies, thereby determining a transcription factor binding profile.

[0072] In some examples, the DNA is cell-free DNA.

[0073] In some examples, (c) comprises using a low-pass filter. In some examples, (c) comprises using a Savitzky-Golay filter.

[0074] In some examples, the subject is a human.

[0075] In another aspect, the present disclosure provides a system for detecting a presence or absence of a disease in a subject, comprising a processor configured to: (i) use a set of sequence

reads from deoxyribonucleic acid (DNA) extracted from the subject to generate a coverage pattern for a transcription factor; (ii) process the coverage pattern to provide a signal, wherein the signal has a different frequency than a reference signal; and (iii) processing the signal with the reference signal, thereby detecting the presence or absence of the disease in the subject.

[0076] In some examples, the present disclosure provides a system for classifying a tumor by tumor subtype or tumor stage, comprising a processor configured to: (i) use a first set of sequence reads from deoxyribonucleic acid (DNA) extracted from the subject at a first time and a second set of sequence reads extracted from DNA from the subject at a second time that is later than the first time to generate a first coverage pattern for a transcription factor and a second coverage pattern for the transcription factor; (ii) process the first coverage pattern to provide a first signal and process the second coverage pattern to provide a second signal, wherein the first signal and the second signal have different frequencies than a reference signal; and (iii) processing the first signal with the reference signal and processing the second signal with the reference signal, to monitor the progression or regression of the disease in the subject.

[0077] In another embodiment, the present disclosure provides a system for monitoring progression or regression of a disease in a subject, comprising a processor configured to: (i) use a first set of sequence reads from deoxyribonucleic acid (DNA) extracted from the subject at a first time and a second set of sequence reads extracted from DNA from the subject at a second time that is later than the first time to generate a first coverage pattern for a transcription factor and a second coverage pattern for the transcription factor; (ii) process the first coverage pattern to provide a first signal, and process the second coverage pattern to provide a second signal, wherein the first signal and the second signal have different frequencies than a reference signal; and (iii) processing the first signal with the reference signal and processing the second signal with the reference signal, to monitor the progression or regression of the disease in the subject.

[0078] Another embodiment of the present disclosure provides a non-transitory computer readable medium comprising machine executable code that, upon execution by one or more computer processors, implements any of the methods above or elsewhere herein.

[0079] Another embodiment of the present disclosure provides a system comprising one or more computer processors and computer memory coupled thereto. The computer memory comprises machine executable code that, upon execution by the one or more computer processors, implements any of the methods above or elsewhere herein.

[0080] In a ninth aspect, the present disclosure provides a computer-implemented method for determining a tumor-specific transcription factor binding site (TFBS) pattern, the method comprising: (a) providing a first set of sequence reads from

deoxyribonucleic acid (DNA) extracted from the subject at a first time and a second set of sequence reads from DNA extracted from the subject at a second time that is later than the first time; (b) using the first set of sequence reads to generate a first coverage pattern for a transcription factor and using the second set of sequence reads to generate a second coverage pattern for the transcription factor; (c) processing the first coverage pattern to provide a first signal and processing the second coverage pattern to provide a second signal; (d) processing the first signal with a reference signal, wherein the first signal and the reference signal have different frequencies; (e) processing the second signal with the reference signal, wherein the second signal and the reference signal have different frequencies; and (f) based on the processing of the first signal and the second signal with the reference signal, determining the tumor-specific TFBS pattern.

[0081] In a tenth aspect, the present disclosure provides a system comprising a computing device comprising at least one computer processor, an operating system configured to perform executable instructions, a memory, and a computer program including instructions executable by the computing device to provide a computer application for determining a tumor-specific TFBS pattern, the computer application comprising: a sequence module programmed to obtain a first set of sequence reads from deoxyribonucleic acid (DNA) extracted from the subject at a first time and a second set of sequence reads from DNA extracted from the subject at a second time that is later than the first time; a coverage module programmed to use the first set of sequence reads to generate a first coverage pattern for a transcription factor and use the second set of sequence reads to generate a second coverage pattern for the transcription factor; a signal module programmed to process the first coverage pattern to provide a first signal, and process the second coverage pattern to provide a second signal; a first processing module programmed to process the first signal with a reference signal, wherein the first signal and the reference signal have different frequencies; a second processing module programmed to process the second signal with the reference signal, wherein the second signal and the reference signal have different frequencies; and a detection module programmed to, based on the processing of the first signal and the second signal with the reference signal, determine the tumor-specific TFBS pattern.

[0082] In an eleventh aspect, the present disclosure provides a non-transitory computer-readable medium comprising machine-executable code that, upon execution by one or more computer processors, implements a method for determining a tumor-specific TFBS pattern, the method comprising: (a) providing a first set of sequence reads from deoxyribonucleic acid (DNA) extracted from the subject at a first time and a second set of sequence reads from DNA extracted

from the subject at a second time that is later than the first time; (b) using the first set of sequence reads to generate a first coverage pattern for a transcription factor and using the second set of sequence reads to generate a second coverage pattern sequence reads for the transcription factor; (c) processing the first coverage pattern to provide a first signal and processing the second coverage pattern to provide a second signal; (d) processing the first signal with a reference signal, wherein the first signal and the reference signal have different frequencies; (e) processing the second signal with the reference signal, wherein the second signal and the reference signal have different frequencies; and (f) based on the processing of the first signal and the second signal with the reference signal, determining the tumor-specific TFBS pattern.

[0083] In a twelfth aspect, the present disclosure provides a system for determining a tumor-specific TFBS pattern, the system comprising: a database comprising a first set of sequence reads from deoxyribonucleic acid (DNA) extracted from the subject at a first time and a second set of sequence reads from DNA extracted from the subject at a second time that is later than the first time; and one or more computer processors operatively coupled to the database, wherein the one or more computer processors are individually or collectively programmed to: (a) use the first set of sequence reads to generate a first coverage pattern for a transcription factor and use the second set of sequence reads to generate a second coverage pattern for the transcription factor; (b) process the first coverage pattern to provide a first signal, and process the second coverage pattern to provide a second signal; (c) process the first signal with a reference signal, wherein the first signal and the reference signal have different frequencies; (d) process the second signal with the reference signal, wherein the second signal and the reference signal have different frequencies; and (e) based on the processing of the first signal and the second signal with the reference signal, determine the tumor-specific TFBS pattern.

[0083a] In a thirteenth aspect, the present disclosure provides a system for detecting a presence or absence of a disease in a subject, comprising a processor configured to: (i) use sequence reads from deoxyribonucleic acid (DNA) extracted from said subject to generate a coverage pattern for a transcription factor; (ii) process said coverage pattern to provide a signal, wherein said signal has a different frequency than a reference signal; and (iii) process said signal with said reference signal, to provide a diagnosis of said disease in said subject.

[0083b] In a fourteenth aspect, the present disclosure provides a system for monitoring a progression or regression of a disease in a subject, comprising a processor configured to: (i) use a first set of sequence reads from deoxyribonucleic acid (DNA) extracted from said subject at a first time and a second set of sequence reads extracted from DNA from said subject at a second

time that is later than said first time to generate a first coverage pattern for a transcription factor and a second coverage pattern for said transcription factor; (ii) process said first coverage pattern to provide a first signal, and process said second coverage pattern to provide a second signal, wherein said first signal and said second signal have different frequencies than a reference signal; and (ii) process said first signal with said reference signal and process said second signal with said reference signal, to monitor said progression or regression of said disease in said subject.

[0084] Another aspect of the present disclosure provides a non-transitory computer readable medium comprising machine executable code that, upon execution by one or more computer processors, implements any of the methods above or elsewhere herein.

[0085] Another aspect of the present disclosure provides a system comprising one or more computer processors and computer memory coupled thereto. The computer memory comprises machine executable code that, upon execution by the one or more computer processors, implements any of the methods above or elsewhere herein.

Additional aspects and advantages of the present disclosure will become readily apparent to those skilled in this art from the following detailed description, wherein only illustrative examples of the present disclosure are shown and described. As will be realized, the present