METHODS FOR CHARACTERIZING CELLS USING GENE EXPRESSION AND CHROMATIN ACCESSIBILITY

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Provisional Patent Application No. 62/976,270, filed February 13, 2020, U.S. Provisional Patent Application No. 62/979,986, filed February 21, 2020, U.S. Provisional Patent Application No. 63/114,378, filed November 16, 2020, and U.S. Provisional Patent Application No. 63/125,331, filed December 14, 2020, each of which applications is herein incorporated by reference in its entirety for all purposes.

BACKGROUND

[0002] While robust high-throughput systems for assaying either transcription or chromatin accessibility (e.g., using ATAC-seq) at single cell resolution are now widespread, researchers and clinicians typically must split cell samples and analyze each modality separately and computationally infer linkages between the gene expression and chromatin accessibility data.

SUMMARY

[0003] In an aspect, the present disclosure provides a condition of a sample comprising: a) generating a plurality of barcoded nucleic acid molecules, wherein said plurality of barcoded nucleic acid molecules comprises (i) a first subset comprising sequences corresponding to regions of accessible chromatin of a plurality of deoxyribonucleic acid (DNA) molecules of a cell or cell nucleus of said sample and (ii) a second subset comprising sequences corresponding to a plurality of ribonucleic acid (RNA) molecules of said cell or cell nucleus; b) (i) sequencing said first subset of said plurality of barcoded nucleic acid molecules to generate a first data set comprising sequencing information corresponding to said regions of accessible chromatin, (ii) sequencing said second subset of said plurality of barcoded nucleic acid molecules to generate a second data set comprising sequencing information corresponding to said plurality of RNA molecules, and (iii) generating a linked signature of said cell or cell nucleus using said first data set and said second data set; c) using said linked signature to identify a gene regulatory network indicative of said condition.

[0003a] Disclosed herein, in some embodiments, are methods for characterizing cells or cell nuclei. In an example, the present disclosure provides a method for characterizing cells or cell nuclei, comprising: providing a plurality of partitions comprising a plurality of cells or cell nuclei and a plurality of particles, wherein a partition of the plurality of partitions comprises a

cell or cell nucleus of the plurality of cells or cell nuclei and a particle of the plurality of particles, wherein (i) the plurality of cells or cell nuclei comprises a plurality of nucleic acid molecules, wherein the plurality of nucleic acid molecules comprises a plurality of ribonucleic acid (RNA) molecules and a plurality of deoxyribonucleic acid (DNA) molecules; and (ii) the plurality of particles comprises a plurality of nucleic acid barcode molecules coupled thereto, wherein a nucleic acid barcode molecule of the plurality of nucleic acid barcode molecules comprises a nucleic acid barcode sequence of a plurality of nucleic acid barcode sequences, and wherein the particle comprises a unique nucleic acid barcode sequence of the plurality of nucleic acid barcode sequences; within the plurality of partitions, using nucleic acid barcode molecules of the plurality of nucleic acid barcode molecules and nucleic acid molecules of the plurality of nucleic acid molecules to generate a plurality of barcoded nucleic acid molecules, wherein the plurality of barcoded nucleic acid molecules comprises (i) a first subset comprising sequences corresponding to RNA molecules of the plurality of RNA molecules and (ii) a second subset

of nucleic acid barcode molecules are releasable from the particles of the plurality of particles upon application of a stimulus.

[0007] In some embodiments, the stimulus is a chemical stimulus. In some embodiments, the stimulus comprises a reducing agent. In some embodiments, the plurality of nucleic acid barcode molecules are coupled to the plurality of particles via a plurality of labile moieties. In some embodiments, the method further comprises generating the plurality of partitions using a microfluidic device. In some embodiments, the method further comprises recovering the plurality of barcoded nucleic acid molecules from the at least the subset of the plurality of partitions. In some embodiments, the method further comprises, prior to (b), lysing or permeabilizing the plurality of cells or cell nuclei to provide access to the plurality of nucleic acid molecules therein. In some embodiments, the method further comprises, prior to (a), processing open chromatin structures of the plurality of cells or cell nuclei with a transposase to provide the plurality of DNA molecules. In some embodiments, within the at least the subset of the plurality of partitions, reverse transcribing the plurality of RNA molecules to provide a plurality of complementary DNA (cDNA) molecules.

[0008] Disclosed herein, in some embodiments, are methods for identifying a genetic feature. In an example, the present disclosure provides a method for identifying a genetic feature, comprising

[0009] A method for identifying a genetic feature, comprising: (a) providing a first data set corresponding to regions of accessible chromatin of a plurality of deoxyribonucleic acid (DNA) molecules of a plurality of cells or cell nuclei and a second data set corresponding to a plurality of ribonucleic acid (RNA) molecules of the plurality of cells or cell nuclei, wherein the first data set comprises a first plurality of sequencing reads corresponding to sequences of the regions of accessible chromatin and a plurality of nucleic acid barcode sequences, and wherein the second data set comprises a second plurality of sequencing reads corresponding to sequences of the plurality of RNA molecules and the plurality of nucleic acid barcode sequences, wherein a cell or cell nucleus of the plurality of cells or cell nuclei corresponds to a nucleic acid barcode sequence of the plurality of nucleic acid barcode sequences; (b) using the plurality of nucleic acid barcode sequences of the first data set and the second data set to identify first sequencing reads of the first plurality of sequencing reads and second sequencing reads of the second plurality of sequencing reads as corresponding to cells or cell nuclei of the plurality of cells or cell nuclei, thereby generating a third data set comprising sequence information corresponding to regions of accessible chromatin and RNA molecules associated with cells or cell nuclei of the plurality of cells or cell nuclei; (c) using the sequence information to identify cell types of the

some embodiments, (c) comprises (i) using the sequence information to cluster cells or cell nuclei of the plurality of cells or cell nuclei by regions of accessible chromatin signatures, (ii) using the sequence information to cluster cells or cell nuclei of the plurality of cells or cell nuclei by gene expression signatures, and (iii) using the sequence information and the cells or cell nuclei clustered by the gene expression signatures to further characterize the cells or cell nuclei clustered by the regions of accessible chromatin. In some embodiments, (c) comprises (i) using the sequence information to cluster cells or cell nuclei of the plurality of cells or cell nuclei by regions of accessible chromatin signatures, (ii) using the sequence information to cluster cells or cell nuclei of the plurality of cells or cell nuclei by gene expression signatures, and (iii) using the sequence information and the cells or cell nuclei clustered by the regions of accessible chromatin signatures to further characterize the cells or cell nuclei clustered by the gene expression signatures.

[0012] In some embodiments, the plurality of cells or cell nuclei is derived from a sample comprising a tumor or suspected of comprising a tumor. In some embodiments, the sample is derived from a bodily fluid. In some embodiments, the sample is derived from a biopsy. In some embodiments, the tumor is a B cell lymphoma tumor. In some embodiments, the method further comprises using the sequence information to identify a presence of a tumor cell or cell nucleus in the sample. In some embodiments, the method further comprises (f) using the sequence information to identify a cell type, a cell state, a tumor-specific gene expression pattern, or a tumor-specific differentially accessible region of chromatin in the sample. In some embodiments, the method further comprises, based at least in part on (f), identifying a therapeutic regimen for treatment of a subject from which the sample derives. In some embodiments, the therapeutic regimen comprises administration of a therapeutically effective amount of an agent targeting one or more targets identified in the tumor-specific gene expression pattern or the tumor-specific differentially accessible region of chromatin.

[0013] Disclosed herein, in some embodiments, are systems for identifying a genetic feature. In an example, the present disclosure provides a system for identifying a genetic feature, comprising: one or more databases comprising a first data set corresponding to regions of accessible chromatin of a plurality of deoxyribonucleic acid (DNA) molecules of a plurality of cells or cell nuclei and a second data set corresponding to a plurality of ribonucleic acid (RNA) molecules of the plurality of cells or cell nuclei, wherein the first data set comprises a first plurality of sequencing reads corresponding to sequences of the regions of accessible chromatin and a plurality of nucleic acid barcode sequences, and wherein the second data set comprises a second plurality of sequencing reads corresponding to sequences of the plurality of RNA

information to identify a presence of a tumor cell or cell nucleus in the sample. In some embodiments, the one or more computer processors are individually or collectively further programmed to compare the sequence information to sequence information from a control sample.

[0018] In some embodiments, the one or more computer processors are individually or collectively further programmed to use the sequence information to identify a cell type, a cell state, a tumor-specific gene expression pattern, or a tumor-specific differentially accessible region of chromatin in the sample. In some embodiments, the one or more computer processors are individually or collectively further programmed to, based at least in part on the use of the sequence information to identify the cell type, the cell state, the tumor-specific gene expression pattern, or the tumor-specific differentially accessible region of chromatin in the sample, identify a therapeutic regimen for treatment of a subject from which the sample derives. In some embodiments, the therapeutic regimen comprises administration of a therapeutically effective amount of an agent targeting one or more targets identified in the tumor-specific gene expression pattern or the tumor-specific differentially accessible region of chromatin. In some embodiments, the system is used to monitor efficacy of the therapeutic regimen for treatment.

[0019] Disclosed herein, in some embodiments, are methods for determining a condition of a sample. In an example, the present disclosure provides a method for determining a condition of a sample, comprising: generating (i) a first data set comprising sequencing information corresponding to regions of accessible chromatin of a plurality of deoxyribonucleic acid (DNA) molecules of a cell or cell nucleus of the sample, (ii) a second data set comprising sequencing information corresponding to a plurality of ribonucleic acid (RNA) molecules of the cell or cell nucleus, and (iii) a linked signature of the cell or cell nucleus using the first data set and the second data set; using the linked signature of the cell or cell nucleus and a control linked signature of a control cell or cell nucleus of a control sample to determine one or more regions of accessible chromatin of the plurality of DNA molecules or one or more genes expressed from the plurality of RNA molecules indicative of the condition.

[0020] In some embodiments, the method further comprises c) determining a level of the one or more regions of accessible chromatin and/or the one or more genes expressed determined in b) indicative of the condition in one or more samples of an individual suspected of having the condition. In some embodiments, the method further comprises providing a diagnostic evaluation of the condition, a prognostic evaluation of the condition, monitoring of the condition, and/or management of the condition. In some embodiments, a gene associated with the one or more regions of accessible chromatin and/or one or more genes expressed determined

b) indicative of the condition in the individual. In some embodiments, the method further comprises, prior to step a), generating a plurality of tagged DNA fragments. In some embodiments, the method further comprises, prior to a), generating a plurality of barcoded nucleic acid molecules, wherein the plurality of barcoded nucleic acid molecules comprises (i) a first subset comprising sequences corresponding to regions of accessible chromatin of the plurality of deoxyribonucleic acid (DNA) molecules of the cell or cell nucleus of the sample and (ii) a second subset comprising sequences corresponding to the ribonucleic acid (RNA) molecules of the cell or cell nucleus.

[0024] In some embodiments, the generating is performed within a plurality of partitions. In some embodiments, the method further comprises sequencing the plurality of barcoded nucleic acid molecules. In some embodiments, the first data set is generated by sequencing a first plurality of barcoded nucleic acid molecules comprising sequences corresponding to regions of accessible chromatin of the plurality of deoxyribonucleic acid (DNA) molecules of the cell or cell nucleus of the sample; and the second data set is generated by sequencing a second plurality of barcoded nucleic acid molecules comprising sequences of the ribonucleic acid (RNA) molecules of the cell or cell nucleus.

[0025] Disclosed herein, in some embodiments, are *in vitro* methods of preparing a biological sample. In an example, the present disclosure provides an in vitro method of preparing a biological sample, comprising: (a) processing open chromatin structures of T-cells and/or B-cells from the biological sample with a transposase to provide a plurality of DNA molecules; (b) generating a first plurality of barcoded nucleic acid molecules comprising the plurality of DNA molecules processed in (a); (c) generating a second plurality of barcoded nucleic acid molecules comprising a plurality of nucleic acids comprising mRNA sequences or derivatives thereof from the T-cells and/or B-cells from the biological sample; and (d) generating a first and second sequencing library from the first and second plurality of barcoded nucleic acid molecules, respectively, to determine a linked signature of a cell of the T-cells and/or B-cells.

[0026] In some embodiments, wherein step (b) and/or step (c) is performed within a plurality of partitions. In some embodiments, the method further comprises determining a significance level for the linked signature determined in step (d). In some embodiments, wherein step (c) comprises reverse transcribing the plurality of mRNA sequences from the T-cells and/or B-cells from the biological sample to provide a plurality of complementary DNA (cDNA) molecules, and the second plurality of barcoded nucleic acid molecules comprises the cDNA molecules. In

some embodiments, wherein step (c) comprises barcoding 3’ ends of the mRNA were barcoded. In some embodiments, the method further comprises encapsulating single nuclei of the T-cells

and/or B-cells in droplets prior to step (b). In some embodiments, the method further comprises (e) determining from the first and second sequencing libraries a presence, absence, and/or level of the one or more linked signatures correlated with a condition. In some embodiments, the condition is a tumor, cancer, malignancy, neoplasm, or other proliferative disease or disorder. In some embodiments, the condition is a B cell malignancy. In some embodiments, the B cell malignancy is B cell lymphoma. In some embodiments, the linked signature of a cell or cell nucleus of the T-cells and/or B-cells is compared to a control linked signature of a control cell or cell nucleus of a control sample. In some embodiments, the method comprises providing the biological sample isolated and obtained from an individual. In some aspects, the method further comprises obtaining the biological sample from the individual. In some aspects, the method may, but need not, comprise the further step of obtaining the biological sample from the individual. In some embodiments, the method is performed *ex vivo.* In some embodiments, the method further comprises providing the biological sample isolated and obtained from an individual prior to step (a).

[0027] Disclosed herein, in some embodiments, are systems for characterizing cells. In an example, the present disclosure provides a system for characterizing cells, comprising: a plurality of partitions comprising a plurality of cells or cell nuclei and a plurality of particles, wherein a partition of the plurality of partitions comprises a cell or cell nucleus of the plurality of cells or cell nuclei and a particle of the plurality of particles, wherein (i) the plurality of cells or cell nuclei comprises a plurality of nucleic acid molecules, wherein the plurality of nucleic acid molecules comprises a plurality of RNA molecules and a plurality of DNA molecules; and (ii) the plurality of particles comprises a plurality of nucleic acid barcode molecules coupled thereto, wherein a nucleic acid barcode molecule of the plurality of nucleic acid barcode molecules comprises a nucleic acid barcode sequence of a plurality of nucleic acid barcode sequences, and wherein the particle comprises a unique nucleic acid barcode sequence of the plurality of nucleic acid barcode sequences; and one or more computer processors, individually or collectively, programmed to; (a) process a plurality of barcoded nucleic acid molecules, generated in the plurality of partitions using the plurality of nucleic acid barcode molecules and the plurality of nucleic acid molecules, or derivatives thereof, to generate sequence information corresponding to the RNA molecules and the DNA molecules; and (b) use the sequence information to identify characteristics of the plurality of cells or cell nuclei.

[0028] In some embodiments, the characteristics of the plurality of cells or cell nuclei comprise cell types. In some embodiments, the cell types are selected from the group consisting of monocytes, natural killer cells, B cells, T cells, granulocytes, dendritic cells, and stromal cells.

plurality of nucleic acid barcode molecules is coupled to the plurality of particles via a plurality of labile moieties. In some embodiments, the system further comprises a microfluidic device that generates the plurality of partitions.

[0032] Disclosed herein, in some embodiments, are systems for determining a condition of a sample. In an example, a system for determining a condition of a sample, comprises: one or more databases comprising (i) a first data set comprising sequencing information corresponding to regions of accessible chromatin of a plurality of deoxyribonucleic acid (DNA) molecules of a cell or cell nucleus of the sample, (ii) a second data set comprising sequencing information corresponding to a plurality of ribonucleic acid (RNA) molecules of the cell or cell nucleus, and (iii) a linked signature of the cell or cell nucleus using the first data set and the second data set; one or more computer processors operably coupled to the one or more databases, wherein the one or more computer processors are individually or collectively programmed to use the linked signature of the cell or cell nucleus and a control linked signature of a control cell or cell nucleus of a control sample to determine one or more regions of accessible chromatin of the plurality of DNA molecules or one or more genes expressed from the plurality of RNA molecules indicative of the condition.

[0033] In some embodiments, the one or more computer processors are individually or collectively programmed to determine a level of the one or more regions of accessible chromatin and/or the one or more genes expressed indicative of the condition in one or more samples of an individual suspected of having the condition. In some embodiments, the one or more computer processors are individually or collectively programmed to generate an output that relates to providing a diagnostic evaluation of the condition, a prognostic evaluation of the condition, monitoring of the condition, and/or management of the condition. In some embodiments, the one or more computer processors are individually or collectively configured to identify a gene associated with the one or more regions of accessible chromatin and/or the one or more genes expressed as a target of a therapeutic regimen for treatment of the condition.

[0034] In some embodiments, the one or more computer processors are individually or collectively programmed to generate an output that relates to determining an administration regimen of a therapeutically effective amount of an agent to a subject targeting the target, wherein the sample is derived from the subject. In some embodiments, the one or more computer processors are individually or collectively programmed to generate an output that relates to determining an efficacy of an agent targeting the target when administered to a subject, wherein the sample is derived from the subject. In some embodiments, the one or more computer processors are individually or collectively programmed to generate an output that

the second data set comprising sequencing information corresponding to a plurality of RNA molecules of the cell or cell nucleus are barcoded with barcoded nucleic acid sequences. In some embodiments, the first data set comprising sequencing information corresponding to regions of accessible chromatin of a plurality of DNA molecules of a cell or cell nucleus of the sample and the second data set comprising sequencing information corresponding to a plurality of RNA molecules of the cell or cell nucleus are barcoded with barcoded nucleic acid sequences within a plurality of partitions.

[0039] In some embodiments, the system further comprises a device or a sequencer configured to sequence the plurality of barcoded nucleic acid sequences. In some embodiments, the first data set is generated by sequencing a first plurality of barcoded nucleic acid sequences comprising sequences corresponding to regions of accessible chromatin of the plurality of deoxyribonucleic acid (DNA) molecules of the cell or cell nucleus of the sample; and the second data set is generated by sequencing a second plurality of barcoded nucleic acid sequences comprising sequences of the ribonucleic acid (RNA) molecules of the cell or cell nucleus.

[0040] In some embodiments, the one or more computer processors are individually or collectively programmed to process the first data set and/or the second data set to generate a filtered first data set and/or a filtered second data set. In some embodiments, the filtered first data set is filtered using motif enrichment. In some embodiments, the filtered second data set is filtered using differential expression analysis. In some embodiments, the one or more computer processors are individually or collectively programmed to process the first data set and/or the second data set to generate a linkage significance. In some embodiments, the one or more computer processors are individually or collectively programmed to process the first filtered data set and/or the second filtered data set to generate an enrichment score. In some embodiments, the filtered first data set and the filtered second data set is used to generate a transcription factor-target gene network. In some embodiments, wherein a gene from the transcription factor-target gene network is identified as a target of a therapeutic regimen for treatment of the condition. In some embodiments, the target is a transcription factor.

[0041] Another example 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.

[0042] Another example 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.

CLAIMS

WHAT IS CLAIMED IS:

1. A method for determining a condition of a sample comprising:
    a) generating a plurality of barcoded nucleic acid molecules, wherein said plurality of barcoded nucleic acid molecules comprises (i) a first subset comprising sequences corresponding to regions of accessible chromatin of a plurality of deoxyribonucleic acid (DNA) molecules of a cell or cell nucleus of said sample and (ii) a second subset comprising sequences corresponding to a plurality of ribonucleic acid (RNA) molecules of said cell or cell nucleus;
    b) (i) sequencing said first subset of said plurality of barcoded nucleic acid molecules to generate a first data set comprising sequencing information corresponding to said regions of accessible chromatin, (ii) sequencing said second subset of said plurality of barcoded nucleic acid molecules to generate a second data set comprising sequencing information corresponding to said plurality of RNA molecules , and (iii) generating a linked signature of said cell or cell nucleus using said first data set and said second data set;
    c) using said linked signature to identify a gene regulatory network indicative of said condition.
2. The method of claim 1, wherein c) comprises determining one or more regions of accessible chromatin of said plurality of DNA molecules or one or more genes expressed from said plurality of RNA molecules indicative of said condition.
3. The method of claim 1, further comprising providing a diagnostic evaluation of said condition, a prognostic evaluation of said condition, monitoring of said condition, or management of said condition.
4. The method of claim 2, wherein a gene associated with said one or more regions of accessible chromatin or one or more genes expressed indicative of said condition is identified as a target of a therapeutic regimen for treatment of said condition.
5. The method of claim 1, wherein c) comprises identifying a transcription factor in said gene regulatory network.
6. The method of claim 1, wherein said sample is from a subject having a tumor or suspected of having a tumor.
7. The method of claim 1, wherein said sequencing information of said first data set and said second data set corresponds to said cell or cell nucleus via a nucleic acid barcode sequence.

8. The method of claim 1, further comprising, prior to c), clustering a plurality of cells or cell nuclei of said sample by respective regions of accessible chromatin signatures, by respective plurality of RNA molecules, or by respective linked signatures of said plurality of cells or cell nuclei
9. The method of claim 8, wherein said plurality of cells or cell nuclei are clustered by cell types selected from the group consisting of monocytes, natural killer cells, B cells, T cells, granulocytes, dendritic cells, and stromal cells.
10. The method of claim 1, wherein c) comprises comparing said sequencing information corresponding to regions of accessible chromatin, said sequencing information corresponding to said plurality of RNA molecules, or said linked signature to sequencing information generated from a different cell or cell nucleus from the same sample or a different sample.
11. The method of claim 2, further comprising d) monitoring a level of said one or more regions of accessible chromatin or said one or more genes expressed indicative of said condition in one or more samples of an individual having said condition or suspected of having said condition.
12. The method of claim 1, wherein said generating is performed within a plurality of partitions.
13. The method of claim 1, further comprising, prior to a), processing open chromatin structures of said cell or cell nucleus with a transposase to provide said plurality of DNA molecules.
14. The method of claim 1, wherein said cell comprises a T cell or a B cell, and wherein said cell nucleus comprises a T cell nucleus or a B cell nucleus.
15. The method of claim 1, further comprising processing said first data set or said second data set to generate a filtered first data set or a filtered second data set.
16. The method of claim 15, wherein said filtered first data set is filtered using motif enrichment or differential expression analysis.
17. The method of claim 1, further comprising processing said first data set or said second data set to generate a linkage significance.
18. The method of claim 1, wherein said condition is a tumor, cancer, malignancy, neoplasm, or other proliferative disease or disorder.
19. The method of claim 18, wherein said condition is a B cell malignancy.
20. The method of claim 19, wherein said B cell malignancy is B cell lymphoma.