CLAIMS

What is claimed is:

1. A system for predicting a prevalence of loss of heterozygosity (LOH) in a target population of cells, the system comprising:

at least one memory storing computer-executable instructions; and

at least one processor in communication with the at least one memory, wherein the at least one processor is configured to execute the computer-executable instructions to:

receive genetic data for a first reference population of cells;

sequence the genetic data for the first reference population of cells to obtain first reference data, the first reference data comprising chromosome identification, nucleotide coordinates, nucleotide composition, or combinations thereof;

identify and remove heterozygous variant positions having imbalanced allelic expression in the first reference data based on single cell RNA sequencing data generated from a second reference population of cells to generate second reference data;

establish a set of homozygous positions that are a predetermined number of nucleotide distances away from each variant position of the second reference data to generate third reference data;

map identifiers for each cell of the target population of cells to the second reference data to generate first mapped identifiers;

map identifiers for each cell of the second reference population of cells to the second reference data and/or the third reference data to generate second mapped identifiers;

apply one or more inputs to a supervised machine learning model, the one or more inputs comprising the first mapped identifiers for each cell of the target population of cells and the second mapped identifiers for each cell of the second reference population of cells, the model being previously trained using historical data, the historical data comprising mapped identifiers for each cell of the target population of cells and their corresponding LOH; and

receive one or more outputs from the model, at least one of the one or more outputs including an LOH for the target population of cells,

thereby predicting the prevalence of LOH in the target population of cells;

update the historical data to include the genetic data for the target population of cells and the corresponding one or more outputs; and

re-train the model using the updated historical data.

2. The system of claim 1, wherein the target population of cells comprise genome-edited cells.

3. The system of claim 2, wherein the genome edited cells comprise CRISPR-edited cells.

4. The system of any one of claims 1-3, wherein the target population of cells comprise cancer cells.

5. The system of any one of claims 1-4, wherein the first reference population of cells comprise the same genotype as wild-type cells of the first reference population of cells.

6. The system of any one of claims 1-5, wherein the at least one processor is configured to execute the computer-executable instructions to sequence the genetic data using bulk DNA sequencing.

7. The system of claim 1-6, wherein the second reference population of cells comprise the cells untreated with genome-editing tools.

8. The system of any one of claims 1-7, wherein the identifiers comprise UMIs.

9. The system of any one of claims 1-8, wherein the model comprises a logistic regression model.

10. A computer-implemented method for predicting a prevalence of loss of heterozygosity (LOH) in a target population of cells, the method comprising:

at least one memory storing computer-executable instructions; and

at least one processor in communication with the at least one memory, wherein the at least one processor is configured to execute the computer-executable instructions to:

receiving genetic data for a first reference population of cells;

sequencing the genetic data for the first reference population of cells to obtain first reference data, the first reference data comprising chromosome identification, nucleotide coordinates, nucleotide composition, or combinations thereof;

identifying and removing heterozygous variant positions having imbalanced allelic expression in the first reference data based on single cell RNA sequencing data generated from a second reference population of cells to generate second reference data;

establishing a set of homozygous positions that are a predetermined number of nucleotide distances away from each variant position of the second reference data to generate third reference data;

mapping identifiers for each cell of the target population of cells to the second reference data to generate first mapped identifiers;

mapping identifiers for each cell of the second reference population of cells to the second reference data and/or the third reference data to generate second mapped identifiers;

applying one or more inputs to a supervised machine learning model, the one or more inputs comprising the first mapped identifiers for each cell of the target population f cells and the second mapped identifiers for each cell of the second reference population of cells, the model being previously trained using historical data, the historical data comprising mapped identifiers for each cell of the target population of cells and their corresponding LOH; and

receiving one or more outputs from the model, at least one of the one or more outputs including an LOH for the target population of cells,

thereby predicting the prevalence of LOH in the target population of cells;

updating the historical data to include the genetic data for the target population of cells and the corresponding one or more outputs; and

re-training the model using the updated historical data.

11. The computer-implemented method of claim 10, wherein the target population of cells comprise genome-edited cells.

12. The computer-implemented method of claim 11, wherein the genome-edited cells comprise CRISPR-edited cells.

13. The computer-implemented method of any one of claims 10-12, wherein the target population of cells comprise cancer cells.

14. The computer-implemented method of any one of claims 10-13, wherein the first reference population of cells comprise the same genotype as wild-type cells of the first reference population of cells.

15. The computer-implemented method of any one of claims 10-14, the method comprises sequencing the genetic data using bulk DNA sequencing.

16. The computer-implemented method of claim 10-15, wherein the second reference population of cells comprise cells untreated with genome-editing tools.

17. The computer-implemented method of any one of claims 10-16, wherein the identifiers comprise UMIs.

18. The computer-implemented method of any one of claims 10-17, wherein the model comprises a logistic regression model.

19. At least one non-transitory computer-readable storage media having computer-executable instructions embodied thereon, wherein when executed by at least one processor, the computer-executable instructions cause the at least one processor to:

receive genetic data for a first reference population of cells;

sequence the genetic data for the first reference population of cells to obtain first reference data, the first reference data comprising chromosome identification, nucleotide coordinates, nucleotide composition, or combinations thereof;

identify and remove heterozygous variant positions having imbalanced allelic expression in the first reference data to generate second reference data;

establish a set of homozygous positions that are a predetermined number of nucleotide distances away from each variant position of the second reference data to generate third reference data;

map identifiers for each cell of a target population of cells to the second reference data to generate first mapped identifiers;

map identifiers for each cell of the second reference population of cells to the second reference data and/or the third reference data to generate second mapped identifiers;

apply one or more inputs to a supervised machine learning model, the one or more inputs comprising the first mapped identifiers for each cell of the target population of cells and the second mapped identifiers for each cell of the second reference population of cells, the model being previously trained using historical data, the historical data comprising mapped identifiers for each cell of the target population of cells and their corresponding LOH; and

receive one or more outputs from the model, at least one of the one or more outputs including an LOH for the target population of cells,

thereby predicting the prevalence of LOH in the target population of cells;

update the historical data to include the genetic data for the target population of cells and the corresponding one or more outputs; and

re-train the model using the updated historical data.