

CLAIMS

What is claimed:

1. A system for providing structural variation and phasing information, the system comprising one or more microprocessors, a persistent memory and a non-persistent memory that collectively store one or more programs that use the one or more microprocessors to perform a method of:

obtaining a request for structural variation and phasing information in a nucleic acid sequencing dataset, wherein

the nucleic acid sequencing dataset represents at least one target nucleic acid in a sample associated with a genome of at least one species,

the nucleic acid sequencing dataset comprises (i) a header, (ii) a synopsis, and (iii) a data section,

the data section comprises a plurality of sequencing reads and is stored in persistent memory,

the synopsis of the nucleic acid sequencing dataset comprises an index to the nucleic acid sequencing dataset,

each respective sequencing read in the plurality of sequencing reads comprises a nucleic acid sequence comprising a first portion that corresponds to a subset of a target nucleic acid in the at least one target nucleic acid and a second portion that encodes a respective identifier for the respective sequencing read in a plurality of identifiers,

each respective identifier is independent of each sequence of the at least one target nucleic acid,

the plurality of sequencing reads collectively include the plurality of identifiers, and

the nucleic acid sequence dataset is 1 gigabyte or greater in size;

responsive to obtaining the request, automatically parsing the request by:

(i) loading the header and the synopsis of the nucleic acid sequencing dataset into the non-persistent memory if not already loaded into the non-persistent memory while retaining the data section in persistent memory,

(ii) comparing the request to the synopsis of the nucleic acid sequencing dataset using the index to the nucleic acid sequencing dataset to identify one or more portions of the data section of the nucleic acid sequencing dataset,

(iii) loading the one or more identified portions of the data section into non-persistent memory, wherein the loading loads less than the entirety of the data section, and
(iv) formatting structural variation and phasing information for display using the one or more identified portions of the data section of the nucleic acid sequencing dataset.

2. The system of claim 2, wherein the header delineates a plurality of components in the nucleic acid sequencing dataset.

3. The system of claim 2, wherein the plurality of components comprises two or more components selected from the group consisting of a summary, an index to variant call data, a phase block track, a refseq index track, a gene track, an exon track, an index to read data, a structural variant dataset track, an index to a target dataset, and an index to a fragment dataset.

4. The system of claim 3, wherein the plurality of components comprises the summary and wherein the summary comprises two or more items in the group consisting of:

a percentage of known SNPs phased in the respective nucleic acid sequencing dataset,
a longest phase block in the respective nucleic acid sequencing dataset,
a number of unique barcodes used in the respective nucleic acid sequencing dataset,
an average fragment length in the respective nucleic acid sequencing dataset,
a mean of the average fragment length in the respective nucleic acid sequencing dataset,
a percentage of fragments greater than a lower threshold in the respective nucleic acid sequencing dataset,

a fragment length histogram in the respective nucleic acid sequencing dataset,
an N50 phase block size in the respective nucleic acid sequencing dataset,
a phase block histogram in the respective nucleic acid sequencing dataset,
a number of sequence reads represented by respective the nucleic acid sequencing dataset,

a median insert size in the respective nucleic acid sequencing dataset,
a median depth in the respective nucleic acid sequencing dataset,
a percent of the target genome with zero coverage in the respective nucleic acid sequencing dataset,

a mapped reads percentage for the respective nucleic acid sequencing dataset,
a PCR duplication percentage for the respective nucleic acid sequencing dataset,

a coverage histogram for the in the respective nucleic acid sequencing dataset,
an identity of a test nucleic acid that forms the basis for the respective nucleic acid sequencing dataset,
a genome source for the respective nucleic acid sequencing dataset,
a sex of an organism that originated the at least one test nucleic acid of the respective nucleic acid sequencing dataset,
a sex of the organism that originate the respective sample of the in the respective nucleic acid sequencing dataset,
a dataset file format version of the in the respective nucleic acid sequencing dataset,
and
a pointer to a plurality of structural variant calls made for the respective nucleic acid sequencing dataset.

5. (New) The system of claim 3, wherein the plurality of components comprises the index to variant call data that provides a correspondence between respective ranges of a genome of the at least one species to offsets in the data section where variant call data for the respective ranges is found.

6. The system of claim 3, wherein the plurality of components comprises the phase block track and wherein the phase block track comprises (i) a dictionary and (ii) a track data section comprising phase information for one or more chromosomes in a genome of the at least one species.

7. The system of claim 3, wherein the plurality of components comprises the refseq index track, wherein the refseq index track comprises an index of a plurality of molecular variation identifiers that are called in the sample.

8. The system of claim 3, wherein the plurality of components comprises the gene track and wherein the gene track comprises (i) a gene track dictionary and (ii) a gene track data section.

9. The system of claim 3, wherein the plurality of components comprises the index to read data wherein the index to read data comprises a lookup table between a respective identifier in the plurality of identifiers and a shortened version of the respective identifier.

10. The system of claim 3, wherein

the plurality of components comprises the structural variant dataset track, and
the structural variant dataset track comprises (i) a dictionary and (ii) a track data section comprising structural variant call information identified in the plurality of sequencing reads.

11. The system of claim 10, wherein the dictionary comprises a plurality of names, and for each respective name in the plurality of names, an offset into the track data where records for the corresponding name are found.

12. The system of claim 11, wherein

the track data section comprises a plurality of structural variant records, and
each structural variant record in the plurality of structural variant records represents a structural variant call made in the at least one target nucleic acid in the sample.

13. The system of claim 12, wherein

each respective structural variant record in the plurality of structural variant records is represented by a node in a plurality of nodes in a respective interval tree in a plurality of interval trees, and

each interval tree in the plurality of interval trees represents a chromosome in a plurality of chromosomes for the species.

14. The system of claim 3, wherein

the plurality of components comprises the index to the target dataset,
the target dataset comprises the regions of the at least one target nucleic acid in the sample that were selected for sequencing in the respective nucleic acid sequencing dataset,
the target dataset is indexed by a target dataset index stored in the synopsis, and
the target dataset is stored in the data section.

15. The system of claim 3, wherein

the plurality of components comprises the index to the fragment dataset,
the fragment dataset comprises a length, chromosomal position, identifier, and phase of each fragment of the at least one target nucleic acid in the sample,

the fragment dataset is indexed by a fragment dataset index stored in the synopsis, and the fragment dataset is stored in the data section.

16. The system of claim 3, wherein the request is for phasing information in a region of a genome and the formatted phasing information includes a graphic representation comprising:

a first haplotype track corresponding to a first parental haplotype of a first species in the at least one species in the region of the genome for the dataset,

a second haplotype track, corresponding to a second parental haplotype of the first species in the region of the genome for the nucleic acid sequencing dataset,

an indeterminate track corresponding to regions of the at least one nucleic acid sample that have not been assigned a parental haplotype in the region of the genome for the nucleic acid sequencing dataset.

17. The system of claim 16, wherein the graphic representation further comprises a graphic representation of each gene that is in the region of the genome.

18. The system of claim 16, wherein the graphic representation further comprises a coverage track for the region of the genome, wherein the coverage track comprises a plurality of vertical bars, and wherein each respective vertical bar in the plurality of vertical bars indicates an average coverage-per-base in the first dataset for a corresponding portion of the genome under the bar.

19. The system of claim 1, wherein the request is converted, without human intervention, to genomic coordinates by comparison of the request against one or more lookup tables that match alphanumeric entries of genes to genomic coordinates.

20. The system of claim 1, wherein the respective sample is associated with a genome of a plurality of species and includes at least a portion of the genome of a first species and a portion of the genome of the second species.

21. The system of claim 1, wherein

the request for structural variation and phasing information in a nucleic acid sequencing dataset is in the form $X_1:N_1-N_2$,

X_1 is an identity of a selected chromosome or a selected first contig sequence,

N_1 is a selected start position within the first chromosome or the selected first contig sequence, and

N_2 is a selected end position within the first chromosome or the selected first contig sequence.

22. The system of claim 1, wherein

the request for structural variation and phasing information in a nucleic acid sequencing dataset is in the form $X_1:N_1$,

X_1 is an identity of a selected first chromosome or a selected first contig sequence, and N_1 is a number of nucleotides, beginning at the origin of the first chromosome or the selected first contig sequence.

23. The system of claim 1, wherein the request for structural variation and phasing information in a nucleic acid sequencing dataset is in the form $Y_1, Y_2, \dots, Y_N$, wherein

each Y_i in $Y_1, Y_2, \dots, Y_N$ is either an alphanumeric identification of a selected gene, a selection of a chromosomal region, or selection of a region of a contig sequence

24. The system of claim 23, wherein

a first Y_i in $Y_1, Y_2, \dots, Y_N$ is an identity of a first chromosome or a first contig sequence having a syntax $X_1:N_1-N_2$,

X_1 is an identity of the first chromosome or the first contig sequence,

N_1 is a selected start position within the first chromosome or the first contig sequence,

N_2 is a selected end position within the first chromosome or the first contig sequence,

and

a second Y_i in $Y_1, Y_2, \dots, Y_N$ is an alphanumeric identification of a selected gene.

25. A method for providing structural variation and phasing information, the method comprising:

obtaining a request for structural variation and phasing information in a nucleic acid sequencing dataset, wherein

the nucleic acid sequencing dataset represents at least one target nucleic acid in a sample associated with a genome of at least one species,

the nucleic acid sequencing dataset comprises (i) a header, (ii) a synopsis, and (iii) a data section,

the data section comprises a plurality of sequencing reads and is stored in persistent memory,

the synopsis of the nucleic acid sequencing dataset comprises an index to the nucleic acid sequencing dataset,

each respective sequencing read in the plurality of sequencing reads comprises a nucleic acid sequence comprising a first portion that corresponds to a subset of a target nucleic acid in the at least one target nucleic acid and a second portion that encodes a respective identifier for the respective sequencing read in a plurality of identifiers,

each respective identifier is independent of each sequence of the at least one target nucleic acid,

the plurality of sequencing reads collectively include the plurality of identifiers, and

the nucleic acid sequence dataset is 1 gigabyte or greater in size;

responsive to obtaining the request, automatically parsing the request at a computer system comprising a processor, non-persistent memory, and persistent memory, by:

(i) loading the header and the synopsis of the nucleic acid sequencing dataset into the non-persistent memory if not already loaded into the non-persistent memory while retaining the data section in persistent memory,

(ii) comparing the request to the synopsis of the nucleic acid sequencing dataset using the index to the nucleic acid sequencing dataset to identify one or more portions of the data section of the nucleic acid sequencing dataset,

(iii) loading the one or more identified portions of the data section into non-persistent memory, wherein the loading loads less than the entirety of the data section, and

(iv) formatting structural variation and phasing information for display using the one or more identified portions of the data section of the nucleic acid sequencing dataset.

26. A non-transitory computer readable storage medium for providing structural variation and phasing information, wherein the non-transitory computer readable storage medium stores instructions, which when executed by a computer system comprising non-persistent memory and persistent memory, cause the computer system to perform a method comprising:

obtaining a request for structural variation and phasing information in a nucleic acid sequencing dataset, wherein

the nucleic acid sequencing dataset represents at least one target nucleic acid in a sample associated with a genome of at least one species,

the nucleic acid sequencing dataset comprises (i) a header, (ii) a synopsis, and (iii) a data section,

the data section comprises a plurality of sequencing reads and is stored in persistent memory,

the synopsis of the nucleic acid sequencing dataset comprises an index to the nucleic acid sequencing dataset,

each respective sequencing read in the plurality of sequencing reads comprises a nucleic acid sequence comprising a first portion that corresponds to a subset of a target nucleic acid in the at least one target nucleic acid and a second portion that encodes a respective identifier for the respective sequencing read in a plurality of identifiers,

each respective identifier is independent of each sequence of the at least one target nucleic acid,

the plurality of sequencing reads collectively include the plurality of identifiers, and

the nucleic acid sequence dataset is 1 gigabyte or greater in size;

responsive to obtaining the request, automatically parsing the request by:

(i) loading the header and the synopsis of the nucleic acid sequencing dataset into the non-persistent memory if not already loaded into the non-persistent memory while retaining the data section in persistent memory,

(ii) comparing the request to the synopsis of the nucleic acid sequencing dataset using the index to the nucleic acid sequencing dataset to identify one or more portions of the data section of the nucleic acid sequencing dataset,

(iii) loading the one or more identified portions of the data section into non-persistent memory, wherein the loading loads less than the entirety of the data section, and

(iv) formatting structural variation and phasing information for display using the one or more identified portions of the data section of the nucleic acid sequencing dataset.