CLAIMS

1. A method comprising:

generating, based on genotype data and phenotype data for a plurality of subjects, one or more of a genotype matrix, a quantitative trait matrix, or a binary trait matrix;

generating, based the genotype matrix, the quantitative trait matrix, and the binary trait matrix, an *n*-tuple data structure;

determining, based on the *n*-tuple data structure, one or more of a sparse vector-based genotype matrix, a sparse vector-based quantitative trait matrix, or a sparse vector-based binary trait matrix; and

processing one or more queries against one or more of the sparse vector-based genotype matrix, sparse vector-based quantitative trait matrix, or the sparse vector-based binary trait matrix.

2. The method of claim 1, wherein the genotype matrix is based on the genotype data, wherein the genotype matrix comprises a column for each of the plurality of subjects and a plurality of rows for each of a plurality of variants.

3. The method of claim 1, wherein the quantitative trait matrix is based on the phenotype data, wherein the quantitative trait matrix comprises a column for each of a plurality of quantitative traits and a plurality of rows for each of the plurality of subjects.

4. The method of claim 1, wherein the binary trait matrix is based on the phenotype data, wherein the binary trait matrix comprises a column for each of a plurality of binary traits and a plurality of rows for each of the plurality of subjects

5. The method of claim 1, further comprising appending at least a portion of a metadata matrix to one or more of the genotype matrix, the quantitative matrix, and the binary trait matrix.

6. The method of claim 1, wherein generating, based on the genotype matrix, the quantitative trait matrix, and the binary trait matrix, the *n*-tuple data structure is further based on an identifier manager.

7. The method of claim 1, wherein determining, based on the *n*-tuple data structure, one or more of a sparse vector-based genotype matrix, a sparse vector-based quantitative trait matrix, or a sparse vector-based binary trait matrix is further based on an identifier manager.

8. The method of claim 1, wherein the *n*-tuple data structure comprises a row identifier for a row, a column identifier for a column, and a value occurring at the intersection of the row and the column.

9. The method of claim 8, wherein the row identifier comprises chromosome:position:reference:alternate or chromosome:range:reference:alternate and wherein the column identifier comprises a cohort identifier.

10. The method of claim 1, further comprising assigning by an identifier manager, a global identifier and a cohort identifier to each of the plurality of subjects, wherein an subject can be assigned more than one cohort identifier and only one global identifier.

11. The method of claim 10, wherein the sparse vector-based genotype matrix comprises a column for each of the plurality of subjects and a plurality of rows for each of the plurality of genotypes, wherein at least one column comprises a sparse vector representing one or more values of the genotype matrix.

12. The method of claim 11, wherein the sparse vector-based quantitative trait matrix comprises a column for each of the plurality of subjects and a plurality of rows for each of the plurality of genotypes, wherein at least one column comprises a sparse vector representing one or more values of the quantitative trait matrix.

13. The method of claim 12, wherein the sparse vector-based binary trait matrix comprises a column for each of the plurality of subjects and a plurality of rows for each of the plurality of genotypes, wherein at least one column comprises a sparse vector representing one or more values of the binary trait matrix.

14. The method of claim 13, further comprising aligning, according to column, the sparse vector-based genotype matrix, the sparse vector-based quantitative trait matrix, and the sparse vector-based binary trait matrix.

15. The method of claim 14, wherein the aligning, according to column, the sparse vector-based genotype matrix, the sparse vector-based quantitative trait matrix, and the sparse vector-based binary trait matrix is based on one or more of the global identifiers or the cohort identifiers.

16. The method of claim 11, wherein the sparse vector representing one or more values of the genotype matrix comprises a data structure having a column for each cohort identifier associated with an subject having a non-zero value in a row of the genotype matrix.

17. The method of claim 11, wherein the sparse vector representing one or more values of the genotype matrix comprises a homozygous reference.

18. The method of claim 12, wherein the sparse vector representing one or more values of the quantitative trait matrix comprises a data structure having a column for each cohort identifier associated with an subject having a non-NULL value in a column of the quantitative trait matrix.

19. The method of claim 13, wherein the sparse vector representing one or more values of the binary trait matrix comprises a data structure having a column for each cohort

identifier associated with an subject having a non-zero value in a column of the binary trait matrix.

20. The method of claim 13, wherein the sparse vector representing one or more values of the genotype matrix or the quantitative trait matrix are configured to discard values of 0 (zero), wherein the sparse vector representing one or more values of the quantitative trait matrix is configured to allow a 0 (zero) value and to discard NULL values, wherein the sparse vector representing one or more values of the quantitative trait matrix comprises an undefined value, and wherein the sparse vector representing one or more values of the binary trait matrix comprises an undefined value.

21. The method of claim 1, further comprising:

receiving additional genotype data and additional phenotype data for an additional plurality of subjects;

assigning, by the identifier manager, a cohort identifier to each subject in common between the plurality of subjects and the additional plurality of subjects; and

assigning, by the identifier manager, a global identifier and a cohort identifier to each of the subjects not in common between the plurality of subjects and the additional plurality of subjects, wherein an subject can be assigned more than one cohort identifier and only one global identifier.

22. The method of claim 21, further comprising:

adding at least a portion of the additional genotype data to the genotype matrix;

adding at least a portion of the additional phenotype data to the quantitative trait matrix;

adding at least a portion of the additional phenotype data to the quantitative trait matrix; and

appending at least a portion of the metadata matrix to each of the genotype matrix, the quantitative trait matrix, and the binary trait matrix.

23. The method of claim 1, further comprising generating, based on one or more of the genotype matrix, the quantitative trait matrix, or the binary trait matrix, an association results matrix.

24. The method of claim 23, further comprising partitioning the association results matrix.

25. The method of claim 24, wherein partitioning the association results matrix comprises:

generating a folder data structure for each of a plurality of chromosomes;
dividing the association results matrix into a plurality of files according to genomic range; and
storing, based on the genomic range and the plurality of chromosomes, the plurality of files in the folder data structures.

26. The method of claim 1, wherein processing one or more queries against one or more of the sparse vector-based genotype matrix, sparse vector-based quantitative trait matrix, or the sparse vector-based binary trait matrix comprises:

receiving a request to perform a data comparison, wherein the request identifies one or more traits of a sparse vector-based trait matrix (TM) to compare to one or more genotypes of the sparse vector-based genotype matrix (GM), wherein the sparse vector-based trait matrix comprises at least a portion of the sparse vector-based quantitative trait matrix and at least a portion of the sparse vector-based binary trait matrix.

27. The method of claim 26, wherein a result of the data comparison comprises one or more trait - genotype associations.

28. The method of claim 26, wherein a result of the data comparison comprises one or more counts of subjects possessing both a trait and a genotype.

29. The method of claim 28, wherein the one or more counts of subjects comprises a count of subjects possessing a reference allele - reference allele (RR) genotype, a reference allele - alternate allele (RA) genotype, an alternate allele - alternate allele (AA) genotype, or a no call (NC) genotype.

30. The method of claim 29, further comprising generating, based on the one or more counts of subjects, a contingency table for each of the identified one or more traits.

31. The method of claim 30, wherein the contingency table comprises a row for case subjects and a row for control subjects, and a column for the RR genotype, the RA genotype, the AA genotype, and the NC genotype, wherein an intersection of a row and a column indicates a count of subjects representative of the row and the column.

32. The method of claim 31, further comprising evaluating, based on the contingency table, a summary statistic.

33. The method of claim 32, wherein the summary statistic comprises Fischer's exact test.

34. The method of claim 29, further comprising:

determining a genotype identifier (GID) for each of the one or more genotypes associated with the identified one or more traits;

determining a trait identifier (TID) for each of the identified one or more traits; and

generating a scaffold data structure, comprising a plurality of rows and a plurality of columns, wherein the plurality of columns comprises a genotype identifier column, a trait identifier of an associated trait column, a contingency table for the associated trait column, and a summary statistic column.

35. The method of claim 34, further comprising:

querying the scaffold data structure to identify a plurality of candidate trait - genotype associations; and

querying the plurality of TM partitions to determine TM partitions comprising a trait from the plurality of candidate trait - genotype associations.

36. The method of claim 35, wherein querying the scaffold data structure to identify a plurality of candidate trait - genotype associations, is based on the summary statistic column, the one or more counts of subjects, or both.

37. The method of claim 26, wherein the sparse vector-based genotype matrix comprises an aggregate sparse vector-based genotype matrix.

38. The method of claim 37, further comprising:

querying a source sparse vector-based genotype matrix based on a plurality of genes using one or more Boolean operators; and

generating, based on the results of querying the source sparse vector-based genotype matrix, the aggregate sparse vector-based genotype matrix.

39. The method of claim 1, wherein processing one or more queries against one or more of the sparse vector-based genotype matrix, sparse vector-based quantitative trait matrix, or the sparse vector-based binary trait matrix comprises:

generating, based on at least a portion of a sparse vector-based trait matrix (TM) and at least a portion of the sparse vector-based genotype matrix (GM), a scaffold data structure, comprising a plurality of rows and a plurality of columns, wherein the plurality of columns comprises a genotype identifier column, a trait identifier of an associated trait column, a contingency table for the associated trait column, and a summary statistic column, wherein the sparse vector-based trait matrix comprises at least a portion of the sparse vector-based quantitative trait matrix and at least a portion of the sparse vector-based binary trait matrix;

querying the scaffold data structure to identify a plurality of candidate trait - genotype associations; and

querying a plurality of TM partitions of the sparse vector-based trait matrix to determine TM partitions comprising a trait from the plurality of candidate trait - genotype associations.

40. The method of claim 39, wherein querying the scaffold data structure to identify a plurality of candidate trait - genotype associations, is based on the summary statistic column, the one or more counts of subjects, or both.

41. An apparatus comprising:

one or more processors; and

a memory storing processor executable instructions that, when executed by the one or more processors, cause the apparatus to perform the methods of any of claims 1-40.

42. A computer-readable medium storing processor executable instructions that, when executed by one or more processors, cause the one or more processors to perform the methods of any of claims 1-40.

43. A system comprising at least one computing device configured to perform the methods of any of claims 1-40.