yet a different genotype. In other instances, the sequences of interest have moved (relative to the "normal" position) because of their proximity to a structural variation.

Incorporating structural variations into the state-of-the art genotyping methods is very difficult, however, because there are many known variants from which any particular read might originate. For each N structural variations, there are approximately 2^N different references that have to be compared to the assembled sequence in order to genotype the sequence. In other words, to accommodate one structural variant, the assembled sequence must be compared to at least 2 separate reference sequences to genotype the sample, but accommodating 20 possible structural variations requires comparing the sequence to roughly one million different references. Even using state-of—the-art methods with parallel computing, this is a very expensive proposition. Furthermore, this combinatorial explosion makes it impossible to feasibly genotype longer sequences that include hundreds of possible structural variations. Thus, approximations are made in the name of reducing computational time. In other words, current methods do not fairly represent the structural variations that are commonly found in many genomes.

SUMMARY

The disclosure provides methods and systems for efficiently genotyping sequence reads by aligning the reads directly to a reference sequence construct that simultaneously accounts for multiple alleles at multiple loci in the genome of the organism. Furthermore, the methods and systems of the disclosure make it possible to deal with structural variations in an efficient way, greatly reducing the computing power necessary to genotype genetic samples using next generation sequencing (N.G.S.). Additionally, because the reference sequence construct accounts for the various possible alleles within the construct, it is possible to directly genotype the sample by merely aligning the reads of the sample to the construct. Particular patterns of alignment are only possible for particular genotypes, thus it is not necessary to compare the assembled sequence to a reference sequence and then compare the variations to mutation files associated with that reference. Disclosed herein is a method of genotyping a genetic sample, comprising: obtaining a plurality of sequence reads corresponding to a genetic sample; aligning the plurality of sequence reads to a reference sequence construct that represents different genotypes as alternative sequences at a first position in the construct; and genotyping the sample based upon the alignment of the plurality of sequence reads against the reference sequence construct. Also disclosed herein is a system for aligning a plurality of reads comprising a processor and memory, wherein the memory comprises instructions that, when executed, cause the processor to obtain a plurality of sequence reads corresponding to a genetic sample; align the plurality of sequence reads to a reference sequence construct that represents different genotypes as alternative sequences at a first position in the construct; and genotype the sample based upon the alignment of the plurality of sequence reads against the reference

sequence construct. Disclosed herein is also a computer-implemented method of genotyping a genetic sample, comprising: providing a plurality of sequence reads corresponding to a genetic sample; aligning the plurality of sequence reads to a reference sequence construct that represents different genotypes as alternative sequences at a first position in the construct using a computer processor; and genotyping the sample based upon the alignment of the plurality of sequence reads against the reference sequence construct. Thus, in one aspect, the present invention refers to a method of genotyping a genetic sample, comprising obtaining a plurality of sequence reads corresponding to a genetic sample; aligning the plurality of sequence reads to a reference directed acyclic graph (DAG) comprising a first path corresponding to a first allele and a second path corresponding to a second allele at a position in the reference DAG, wherein the first allele comprises a genetic structural variation; and determining a genotype for the genetic sample based upon the number of sequence reads aligning to the first path and the second path. In another aspect, the present invention refers to a computer-implemented method of genotyping a genetic sample, comprising providing a plurality of sequence reads corresponding to a genetic sample; aligning the plurality of sequence reads to a reference directed acyclic graph (DAG) comprising nodes, wherein the reference DAG comprises a first path corresponding to a first allele and a second path corresponding to a second allele at a position in the reference DAG using a computer processor, wherein the first allele comprises a genetic structural variation; and determining a genotype for the genetic sample based upon the number of sequence reads aligning to the first path and the second path. In a further aspect, the present invention refers to a system for aligning a plurality of reads comprising a processor and memory, wherein the memory comprises instructions that, when executed, cause the processor to obtain a plurality of sequence reads corresponding to a genetic sample; align the plurality of sequence reads to a reference directed acyclic graph (DAG), wherein the reference DAG comprises a first path corresponding to a first allele and a second path corresponding to a second allele t at a position in the reference DAG, wherein the first allele comprises a genetic structural variation; and determine a genotype for the genetic sample based upon the number of sequence reads aligning to the first path and the second path.

The methods and systems of the invention transform linear, local sequence alignment processes such as, for example, Smith-Waterman-Gotoh, into multi-dimensional alignment algorithms that provide increased parallelization, increased speed, increased accuracy, and the ability to align reads through an entire genome. Algorithms of the invention provide for a "look-

CLAIMS

1. A method of genotyping a genetic sample, comprising:

obtaining a plurality of sequence reads corresponding to a genetic sample;

aligning the plurality of sequence reads to a reference directed acyclic graph (DAG) comprising a first path corresponding to a first allele and a second path corresponding to a second allele at a position in the reference DAG, wherein the first allele comprises a genetic structural variation; and

determining a genotype for the genetic sample based upon the number of sequence reads aligning to the first path and the second path .

2. The method of claim 1, wherein the method does not include comparing the sequence reads to a variant call format (VCF) file or a single nucleotide polymorphism database (dbSNP).

3. The method of claim 1, wherein the reference DAG further comprises a third path corresponding to a third allele at the position in the reference DAG.

4. The method of claim 1, wherein the alignment of two or more related reads is used to genotype the sample.

5. The method of claim 1, further comprising determining an allele frequency for the sample based upon the number of sequence reads that align to the first path and the second path .

6. The method of claim 1, further comprising determining a confidence value for the genotype based upon the number of overlapping base pairs between a sequence read and one of the first path and the second path, wherein more overlap correlates with greater confidence.

7. The method of claim 1, wherein the reference DAG represents a chromosome.

8. The method of claim 1, wherein the reference DAG represents a genome.

9. The method of claim 1, wherein the first path and the second path differ from each other by a base deletion, a base insertion, or a polymorphism.

10. The method of claim 1, wherein the sequence reads are obtained by sequencing a genetic sample of a subject with a method selected from Sanger sequencing, pyrosequencing, ion semiconductor sequencing, sequencing by synthesis, sequencing by ligation, and single-molecule real-time sequencing.

11. The method of claim 1, wherein the reference DAG construct comprises greater than about 1,000 base pairs.

12. The method of clam 11, wherein the reference DAG construct comprises greater than about 1,000,000 symbols.

13. The method of claim 1, wherein the plurality of sequence reads comprises greater than about 1000 sequence reads.

14. The method of claim 1, wherein the majority of the plurality of sequence reads are greater than about 50 base pairs in length.

15. The method of claim 14, wherein the majority of the plurality of sequence reads are greater than about 100 base pairs in length.

16. The method of claim 1, wherein the genetic sample originates in blood, urine, saliva, sputum, stool, nipple aspirate, sweat, hair follicle, buccal swab, or tissue.

17. The method of claim 16, further comprising isolating a plurality of nucleic acids from the sample.

18. The method of claim 1, further comprising diagnosing a disease based upon the genotype of the sample.

19. A computer-implemented method of genotyping a genetic sample, comprising:
providing a plurality of sequence reads corresponding to a genetic sample;

aligning the plurality of sequence reads to a reference directed acyclic graph (DAG) comprising nodes, wherein the reference DAG comprises a first path corresponding to a first allele and a second path corresponding to a second allele at a position in the reference DAG using a computer processor, wherein the first allele comprises a genetic structural variation; and

determining a genotype for the genetic sample based upon the number of sequence reads aligning to the first path and the second path .

20. The method of claim 19, wherein the plurality of sequence reads are provided in a non-transitory computer-readable medium.

21. The method of claim 19, wherein the reference DAG is provided in in a non-transitory computer-readable medium.

22. A system for aligning a plurality of reads comprising a processor and memory, wherein the memory comprises instructions that, when executed, cause the processor to:

obtain a plurality of sequence reads corresponding to a genetic sample;

align the plurality of sequence reads to a reference directed acyclic graph (DAG), wherein the reference DAG comprises a first path corresponding to a first allele and a second path corresponding to a second allele t at a position in the reference DAG, wherein the first allele comprises a genetic structural variation ; and

determine a genotype for the genetic sample based upon the number of sequence reads aligning to the first path and the second path.

23. The system of claim 22, wherein aligning comprises:

simultaneously comparing a string of symbols corresponding to a sequence read to a plurality of different genotypes at the position,

scoring overlaps between the string of symbols and each of the first path and the second path, wherein a higher score corresponds to a greater amount of overlap; and

identifying the overlaps corresponding to the highest score for the sequence read.

24. The system of claim 23, further comprising writing a file to memory corresponding to the identified genotype.

25. The system of claim 22, wherein the plurality of sequence reads are provided in a non-transitory computer-readable medium.

26. The system of claim 22, wherein the reference DAG is provided in in a non-transitory computer-readable medium.

27. The system of claim 22, wherein the system comprises a plurality of processors, and wherein each processor is configured to align a portion of the plurality of sequence reads to the reference DAG.

28. The system of claim 22, wherein the instructions additionally cause the processor to output the determined genotype for the sample.