

that it is unclear whether a particular base in the sequence is a result of a mutation in the structural variation or just a misread of the “normal” structural variation.

An additional problem is presented when a mutation or variant, e.g., a small indel or polymorphism, is located in the read in close proximity to a structural variation. The difficulty of aligning the structural variation may cause reads containing both the mutation and the structural variation to be discarded as “unalignable.” The more such reads are discarded, the more likely it becomes that the mutation gets missed entirely. As a result, meaningful rare variants or mutations (used interchangeably here) close to structural variations can be rejected due to the low alignment score associated with the structural variation. These overlooked variants may have (undiscovered) roles in regulating disease.

Because failing to detect rare variants in proximity to structural variations significantly limits the quality of genetic analysis, there is a need for sequence alignment techniques that can account for structural variations, resulting in better alignment of rare mutations or variants.

SUMMARY

The invention generally provides algorithms and methods for their implementation that 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. In one aspect, the present invention refers to a method for identifying a mutation in proximity to a structural variation in a sequence, the method comprising the steps of: obtaining a plurality of nucleic acid sequence reads, wherein at least one nucleic acid read comprises a mutation; comparing said reads to a reference sequence construct, wherein said reference sequence construct is stored in computer memory as a directed acyclic graph comprising at least two alternative sequences at a position in the reference sequence construct, one of which is a structural variation, scoring sequence overlaps for each nucleic acid read against the reference sequence construct; aligning each read to a location on the construct such that the score for each read is maximized; and identifying the mutation as being aligned within 100 bp or fewer of the structural variation. In certain aspects, examples, the invention-disclosure is implemented on assembled reference constructs that can account for the variability in genetic sequences within a species, population, or even among different cells in a single organism. A Particular benefit of reference constructs and alignment algorithms of the invention is the ability to align sequence reads that match to different sequences at a certain position of a reference sequence. That is, a reference construct allows a sequence read to align against one of at least two different sequence paths at a certain position—e.g., a path that follows a sequence common to a reference sequence or another path that follows sequence uncommon to that reference sequence (e.g. mutations, polymorphisms, structural variants).

Algorithms of the invention as described herein provide for a “look-back” type analysis of sequence information (as in Smith-Waterman), however, in contrast to known linear methods, the look back of the invention is conducted through a multi-dimensional space that includes multiple pathways and multiple nodes in order to provide more accurate alignment of complex and

lengthy sequence reads, while achieving lower overall rates of mismatches, deletions, and insertions.

In practice, the ~~invention~~ method described herein is implemented by aligning sequence reads to a series of directed, acyclic sequences spanning branch points that account for all, or nearly-all, of the possible sequence variation in the alignment, including insertions, deletions, and substitutions. Such constructs, often represented as directed acyclic graphs (DAGs) can be easily assembled from available sequence databases, including “accepted” reference sequences and variant call format (VCF) entries. When combined with DAGs, or other constructs that take into account known variation, the disclosed algorithm thus provides a multi-dimensional approach to sequence alignment that greatly improves alignment accuracy and allows for sequence resolution not possible with conventional algorithms. The techniques can be used with any sequence information, in fact, however they are most useful for aligning nucleic acid sequences and amino acid sequences, as discussed herein.

The ~~invention~~ disclosure additionally provides methods to make specific base calls at specific loci using a reference sequence construct, e.g., a DAG that represents known variants at each locus of the genome. Because the sequence reads are aligned to the DAG during alignment, the subsequent step of comparing a mutation, vis-à-vis the reference genome, to a table of known mutations can be eliminated. Using the disclosed methods, it is merely a matter of identifying a nucleic acid read as being located at a known mutation represented on the DAG and calling that mutation. Alternatively, when a mutation is not known (i.e., not represented in the reference sequence construct), an alignment will be found and the variant identified as a new mutation. The method also makes it possible to associate additional information, such as specific disease risk or disease progression, with known mutations that are incorporated into the reference sequence construct. Furthermore, in addition to having the potential to find all genetically relevant results during alignment, the disclosed methods reduce the computational resources required to make the alignments while allowing for simultaneous comparison to multiple reference sequences.

The ~~invention~~ disclosure additionally includes methods for constructing a directed acyclic graph data structure (DAG) that represents known variants at positions within the sequence of an organism. The DAG may include multiple sequences at thousands of positions, and may include multiple variants at each position, including deletions, insertions, translations, inversions, and

single-nucleotide polymorphisms (SNPs). It is also possible to tag each variant in the DAG with correlated diagnostic information, such as “breast cancer,” thereby reducing the steps needed to identify the risks to a patient providing a sample. In some ~~embodiment~~examples, the variants will be scored, weighted, or correlated with other variants to reflect the prevalence of that variant as a marker for disease.

The ~~invention~~disclosure additionally includes systems for executing methods of the ~~invention~~described herein. In one ~~embodiment~~example, a system comprises a distributed network of processors and storage capable of comparing a plurality of sequences (i.e., nucleic acid sequences, amino acid sequences) to a reference sequence construct (e.g., a DAG) representing observed variation in a genome or a region of a genome. The system is additionally capable of aligning the nucleic acid reads to produce a continuous sequence using an efficient alignment algorithm. Because the reference sequence construct compresses a great deal of redundant information, and because the alignment algorithm is so efficient, the reads can be tagged and assembled on an entire genome using commercially-available resources. The system comprises a plurality of processors that simultaneously execute a plurality of comparisons between a plurality of reads and the reference sequence construct. The comparison data may be accumulated and provided to a health care provider. Because the comparisons are computationally tractable, analyzing sequence reads will no longer represent a bottleneck between NGS sequencing and a meaningful discussion of a patient’s genetic risks.

According to certain ~~aspects~~examples, the ~~invention~~disclosure provides methods and systems for aligning a nucleic acid sequence to a reference sequence construct that accounts for different structural variations within a sequence to be aligned. With the structural variations represented in the reference sequence construct, sequence reads which include part of the structural variation are aligned to the reference sequence with a higher alignment score than possible with linear reference sequences that do not account for the variation. As a result, the different paths of the reference construct advantageously allow sequence reads containing large known sequence variations or portions thereof to successfully align to the construct for subsequent analysis. With a structural variant of a sequence read properly aligned, other sequence data that are part of the sequence read, especially those in close proximity to the structural variant likewise align to the reference construct with a higher alignment score than possible with linear reference sequence.

The consistent, reliable alignment of reads containing a structural variant allows capture of sequence data that otherwise might be discarded as unaligned sequence. Because of this, algorithms and reference constructs ~~of the invention~~described herein are used to capture rare variants located in proximity to a structural variation. A rare variant is one that occurs in the sample or population at a frequency of less than about 5% and preferably less than about 1%. For the purposes of the ~~invention~~description, rare variants may also include “unrepresented” variants, i.e., variants that are not reflected in the reference being used, be it a linear reference sequence (in pre-existing methods) or the reference constructs ~~of the invention~~described herein. For example, a rare variant that is close to a structural variant (such that a sequence read includes at least portions of the structural variant and the rare variant) will align to a reference construct ~~of the invention~~described herein along with the structural variant. As a direct result of the structural variant’s alignment, the rare variant close to the structural variant will be present in a large number of otherwise well-aligned and reliable reads because of the proper alignment of the structural variant in the sequence read to the reference construct. The consistent presence of the rare variant causes it to be recognized as a legitimate genetic variant rather than sequencing error. The use of a multi-path alignment approach as described herein allows resolution of rare sequence variations that would likely be lost using a linear reference alignment approach.

BRIEF DESCRIPTION OF THE FIGURES

FIGS. 1A and 1B depict the construction of a directed acyclic graph (DAG) representing genetic variation in a reference sequence. FIG. 1A shows the starting reference sequence and the addition of a deletion. FIG. 1B shows the addition of an insertion and a SNP, thus arriving at the Final DAG used for alignment;

FIG. 2 depicts three variant call format (VCF) entries represented as directed acyclic graphs;

FIG. 3A shows a pictorial representation of aligning a nucleic acid sequence read against a construct that accounts for an insertion event as well as the reference sequence;

FIG. 3B shows the matrices and the backtrack used to identify the proper location of the nucleic acid sequence read “ATCGAA”;

FIG. 4 depicts an associative computing model for parallel processing;

FIG. 5 depicts an architecture for parallel computation;

CLAIMS

1. A method for identifying a mutation in proximity to a structural variation in a sequence, the method comprising the steps of:

obtaining a plurality of nucleic acid sequence reads, wherein at least one nucleic acid read comprises a mutation;

comparing said reads to a reference sequence construct, wherein said reference sequence construct is stored in computer memory as a directed acyclic graph comprising at least two alternative sequences at a position in the reference sequence construct, one of which is a structural variation,

scoring sequence overlaps for each nucleic acid read against the reference sequence construct;

aligning each read to a location on the construct such that the score for each read is maximized; and

identifying the mutation as being aligned within 100 bp or fewer of the structural variation.

2. The method of claim 1, further comprising assembling the nucleic acid reads to each other based upon the alignment of the nucleic acid reads with respect to the reference sequence construct.

3. The method of claim 1, wherein the structural variation is at least 100 bp long.

4. The method of claim 1, wherein the reference sequence construct further comprises at least two additional alternative sequences at a second position in the reference construct, and one of the additional alternative sequences comprises a sequence matching the mutation.

5. The method of claim 4, wherein the first and second positions are separated by 100 bp or fewer.

6. The method of claim 1, wherein the reference sequence construct further comprises at least two additional alternative sequences at a second position in the reference construct, and neither of the additional alternative sequences comprises a sequence matching the mutation.

7. The method of claim 6, wherein the first and second positions are separated by 100 bp or fewer.
8. The method of claim 1, wherein the structural variation is 1 kilobase to 3 megabases in length.
9. The method of claim 1, wherein the mutation was not previously identified in a variant call format (VCF) file.
10. The method of claim 1, wherein the mutation was previously identified in a variant call format (VCF) file.
11. The method of claim 1, wherein the structural variation is selected from the group consisting of deletions, duplications, copy-number variations, insertions, inversions, and translocations.
12. The method of claim 1, wherein the mutation is selected from the group consisting of a deletion, a duplication, an inversion, an insertion, and a single nucleotide polymorphism.
13. The method of claim 1, wherein the mutation does not comprise a sequence matching the reference construct.
14. The method of claim 1, wherein the reference sequence construct comprises a genome of an organism.
15. The method of claim 1, wherein the reference sequence comprises a chromosome of an organism.

CLAIMS

1. A method for aligning a plurality of nucleic acid reads, the method comprising:
obtaining a plurality of nucleic acid reads, wherein at least one nucleic acid read comprises a sequence that corresponds to at least a portion of a structural variation;
providing a reference sequence construct comprising at least two alternative sequences at a first position, wherein one of the alternative sequences comprises the structural variation;
scoring sequence overlaps for each nucleic acid read against the reference sequence construct; and
aligning each nucleic acid read to a location in the construct such that the score for each nucleic acid read is maximized, thereby aligning the at least one read to the alternative sequence comprising the structural variation.
2. The method of claim 1, further comprising assembling the nucleic acid reads to each other based upon the alignment of the nucleic acid reads with respect to the construct.
3. The method of claim 1, wherein the structural variation is at least 100 bp long.
4. The method of claim 1, wherein at least one nucleic acid read comprises a mutation in proximity to the structural variation.
5. The method of claim 4, wherein the mutation is separated from the structural variation by 100 bp or fewer.
6. The method of claim 4, wherein the aligning step further comprises identifying the presence of the mutation.
7. The method of claim 6, wherein the reference sequence construct further comprises at least two additional alternative sequences at a second position in the reference construct, and one of the additional alternative sequences represents a mutation with respect to the reference sequence.

8. The method of claim 7, wherein the first and second positions are separated by 100 bp or fewer.
9. The method of claim 1, wherein the structural variation is about 1 kilobase to about 3 megabases in length.
10. The method of claim 1, wherein the known structural variation is selected from the group consisting of deletions, duplications, copy-number variations, insertions, inversions, and translocations.
11. The method of claim 4, wherein the mutation is a rare variant.
12. The method of claim 4, wherein the mutation comprises a minor allele with frequency of 5% or less among a species.
13. The method of claim 4, wherein the mutation is selected from the group consisting of a deletion, a duplication, an inversion, an insertion, and a single nucleotide polymorphism.
14. The method of claim 4, wherein the mutation is not represented in the reference construct.
15. The method of claim 1, wherein a path through the reference sequence construct represents a genome of an organism.
16. The method of claim 1, wherein a path through the reference sequence represents a chromosome, or portion thereof, of an organism.
17. The method of claim 1, wherein a path through the reference sequence construct is acyclic.
18. A system for aligning a plurality of nucleic acid reads comprising a processor and memory, wherein the memory comprises instructions that, when executed, cause the processor to:

obtain a plurality of nucleic acid reads as strings of symbols, wherein at least one string of symbols corresponds to at least a portion of a known structural variation in a genetic sequence;

provide a reference sequence construct comprising at least two different strings of symbols at a first position in the construct, wherein one of the at least two different strings of symbols represents the known structural variation;

score overlaps between each string of symbols corresponding to a nucleic acid read and the reference sequence construct, wherein a higher score corresponds to a greater likelihood of alignment between the read and reference;

identify the overlap corresponding to the highest score for each read; and

align the strings of symbols corresponding to nucleic acid reads to a location on the construct corresponding to the highest score, thereby aligning the string of symbols corresponding to at least a portion of the known structural variation to the known structural variation.

19. The system of claim 18, wherein the memory additionally includes instructions that cause the processor to assemble strings of symbols corresponding to nucleic acid reads to each other based upon the alignment of the strings of symbols with respect to the construct.

20. The system of claim 18, wherein the known structural variation is at least 100 base pairs in length.

21. The system of claim 18, wherein the string of symbols corresponding to at least a portion of a known structural variation additionally comprises symbols representing a rare variant.

22. The system of claim 21, wherein the rare variant is separated from the portion of the structural variation by symbols representing 100 bp or fewer.

23. The system of claim 21, wherein the reference sequence construct additionally comprises at least two alternative strings of symbols at a second position in the reference construct, and the

string of symbols corresponding to the rare variant aligns to one of the alternative strings of symbols at the second position.

24. The system of claim 23, wherein the first and second positions are separated by 100 bp or fewer.

25. The system of claim 18, wherein the known structural variation is between about 1 kilobase and 3 megabases in length.

26. The system of claim 18, wherein the known structural variation represents a variation selected from the group consisting of deletions, duplications, copy-number variations, insertions, inversions, and translocations.

27. A method for identifying a rare variant in proximity to a genomic structural variation, the method comprising the steps of:

- applying a plurality of nucleic acid sequence reads to a reference sequence construct comprising at least two alternative sequences at a position in the construct, wherein one of the alternative sequences comprises a structural variation;

- scoring sequence overlaps for each nucleic acid read against the reference sequence construct;

- aligning each read to a location in the construct such that the score for each read is maximized; and

- identifying a rare variant in proximity to the structural variation as a sequence in proximity of the structural variation that was not previously in the reference construct.

28. A method for identifying a rare variant in proximity to a genomic structural variation, the method comprising the steps of:

- obtaining a plurality of nucleic acid sequence reads;

- applying said reads to a reference sequence construct comprising at least two alternative sequences at a position in the construct, one of which is a structural variation;

- scoring sequence overlaps for each nucleic acid read against the reference construct;

aligning each read to a location on the construct such that the score for each read is maximized; and

identifying a rare variant in proximity to a structural variation that is aligned on the reference construct.

29. The method of claim 28, further comprising assembling the nucleic acid reads to each other based upon the alignment of the nucleic acid reads with respect to the construct.

30. The method of claim 28, wherein the structural variation is at least 100 bp long.

31. The method of claim 28, wherein at least one nucleic acid read comprises the rare variant.

32. The method of claim 31, wherein the rare variant is separated from the genomic structural variation by 100 bp or fewer.

33. The method of claim 28, wherein the reference sequence construct further comprises at least two additional alternative sequences at a second position in the reference construct, and one of the additional alternative sequences represents the rare variant.

34. The method of claim 28, wherein the reference sequence construct further comprises at least two additional alternative sequences at a second position in the reference construct, and neither of the additional alternative sequences represents the rare variant.

35. The method of claim 33, wherein the first and second positions are separated by 100 bp or fewer.

36. The method of claim 35, wherein the structural variation is about 1 kilobase to about 3 megabases in length.

37. The method of claim 28, wherein the rare variant was not previously identified in a variant call format (VCF) file.

38. The method of claim 28, wherein the known structural variation is selected from the group consisting of deletions, duplications, copy-number variations, insertions, inversions, and translocations.
39. The method of claim 28, wherein the rare variant comprises a minor allele with frequency of 5% or less among a species.
40. The method of claim 28, wherein the rare variant is selected from the group consisting of a deletion, a duplication, an inversion, an insertion, and a single nucleotide polymorphism.
41. The method of claim 28, wherein the rare variant is not represented in the reference construct.
42. The method of claim 28, wherein a path through the reference sequence construct represents a genome of an organism.
43. The method of claim 28, wherein a path through the reference sequence represents a chromosome, or portion thereof, of an organism.
44. The method of claim 28, wherein a path through the reference sequence construct is acyclic.
45. A system for identifying a rare variant in proximity to a genomic structural variation, comprising a processor and memory, wherein the memory comprises instructions that, when executed, cause the processor to:
- obtain a plurality of nucleic acid reads as strings of symbols, wherein at least one string of symbols corresponds to at least a portion of a known structural variation in a genetic sequence;
 - provide a reference sequence construct comprising at least two different strings of symbols at a first position in the construct, wherein one of the at least two different strings of symbols represents the known structural variation;

score overlaps between each string of symbols corresponding to a nucleic acid read and the reference sequence construct, wherein a higher score corresponds to a greater likelihood of alignment between the read and reference;

identify the overlap corresponding to the highest score for each read;

align the strings of symbols corresponding to nucleic acid reads to a location on the construct corresponding to the highest score, thereby aligning the string of symbols corresponding to at least a portion of the known structural variation to the known structural variation; and

identify a rare variant in at least one string of symbols.

46. The system of claim 45, wherein the memory additionally includes instructions that cause the processor to assemble strings of symbols corresponding to nucleic acid reads to each other based upon the alignment of the strings of symbols with respect to the construct.

47. The system of claim 45, wherein the known structural variation is at least 100 base pairs in length.

48. The system of claim 45, wherein the rare variant is separated from the portion of the structural variation by symbols representing 100 bp or fewer.

49. The system of claim 45, wherein the reference sequence construct additionally comprises at least two alternative strings of symbols at a second position in the reference construct, and the string of symbols corresponding to the rare variant aligns to one of the alternative strings of symbols at the second position.

50. The system of claim 49, wherein the first and second positions are separated by 100 bp or fewer.

51. The system of claim 45, wherein the known structural variation is between about 1 kilobase and 3 megabases in length.

52. The system of claim 45, wherein the known structural variation represents a variation selected from the group consisting of deletions, duplications, copy-number variations, insertions, inversions, and translocations.
53. The system of claim 45, wherein the rare variant comprises a minor allele with frequency of 5% or less among a species.
54. The system of claim 45, wherein the rare variant is selected from the group consisting of a deletion, a duplication, an inversion, an insertion, and a single nucleotide polymorphism.
55. The system of claim 45, wherein the rare variant is not represented in the reference construct.