annotated reference. Where the new data includes paired end reads, each read of a pair may potentially align to any location along a genome of length L and in a complex DAG, there may be many more locations (i.e., if a DAG is, on average, n-plex across any loci, aligning a pair of paired-ends reads amounts to choosing from (nL)^2 options). However, insert length information provides a tool for constraining the potential alignments within a DAG—not just in distance away within a path, but also by excluding paths, limiting the number of options to be linear with respect to L. Paired-end reads may provide other information, as well. For example, paired-end reads can provide a basis for modifying a DAG, can be used to infer isoform frequency, or others. In fact, sets of paired-end reads with insert length information carry qualities of information that are well-suited to focusing DAG-based alignments and to updating DAGs to represent actually-occurring natural isoforms that may not otherwise have been discovered or noticed. Thus paired-end reads are particularly well suited for RNA-Seq projects, allowing very large sets of RNA-Seq data to be analyzed against very large reference sets. Using paired-end RNA-Seq reads to discover all isoforms of a transcriptome is not computationally intractable and transcriptomics can move at a pace and throughput not previously possible. One or any number of transcriptomes can be studied from one or any number of organisms to increase understanding of organismal function, development, and disease.

The disclosure relates to creation of a DAG that includes features (e.g., exons, introns, etc.) of a reference connected by their proximal edges in genomic order. This DAG data structure has applications in transcriptomics including, for example, identifying isoforms. Such a DAG reference is used in transcriptomics involving RNA-Seq data that includes paired-reads. RNA-Seq data may be analyzed using the DAG reference that includes features as nodes (e.g., created from an annotated source genome or transcriptome). Systems and methods of the invention provide for improved alignments using paired-end reads, building the DAG using paired-end read information, inferring isoform probabilities or frequencies, or combinations thereof.

In one example, there is disclosed a method of analyzing a transcriptome, the method comprising obtaining a pair of paired-end reads from a transcriptome; finding an alignment with an optimal score between a first read of the pair and a node in a directed acyclic data structure, the data structure comprising nodes representing RNA sequences and edges connecting pairs of the nodes; identifying candidate paths that include the node connected to a

downstream node by a path having a length substantially similar to an insert length of the pair of paired-end reads; and aligning the paired-end reads to the candidate paths to determine an optimal-scoring alignment. In one aspect, the present invention refers to a method of analyzing a transcriptome, the method comprising obtaining a pair of paired-end reads from a transcriptome; finding an alignment with an optimal score between a first read of the pair and a node representing an exon in a directed acyclic graph (DAG), the DAG comprising nodes representing a plurality of exons and introns from features in an annotated transcriptome database, and edges connecting pairs of the nodes in their canonical genomic order; identifying candidate paths within the DAG that include the node connected to a downstream node by a path having a length substantially similar to an insert length of the pair of paired-end reads; aligning the paired-end reads to the candidate paths to determine an optimal-scoring alignment by excluding any paths that are not candidate paths from any alignment calculations involving the pair of paired-end reads; and identifying an isoform from the transcriptome using the optimal-scoring alignment of the paired-end reads. In another aspect, the present invention refers to a system for transcriptomics, the system comprising a computer system comprising a processor coupled to a non-transitory memory having stored therein a directed acyclic graph (DAG), the DAG comprising nodes representing a plurality of exons and introns from features in an annotated transcriptome database, and edges connecting pairs of the nodes in their canonical genomic order, wherein the system is operable to obtain a pair of paired-end reads from a cDNA fragment; find an alignment with an optimal score between a first read of the pair and a node in the DAG; identify candidate paths within the DAG that include the node connected to a downstream node by a path having a length substantially similar to an insert length of the paired end reads; align the paired-end rends to the candidate paths to determine an optimal-scoring alignment by excluding any paths that are not candidate paths from any alignment calculations involving the pair of paired-end nodes; and identify an isoform from the transcriptome using the optimal scoring-alignment of the paired-end reads. In yet another aspect, the present invention refers to a method of inferring isoform frequency in a transcriptome, the method comprising aligning a plurality of pairs of paired-end reads from cDNA to a reference directed acyclic graph (DAG) comprising a plurality of nodes representing exons and introns from annotated features in a genetic feature database, wherein pairs of the plurality of nodes are connected in their canonical genomic order at their proximal ends by an edge, wherein the aligning for a pair of paired-end reads comprises identifying an exon at which a first member of a pair aligns, determining a set of distances that accounts for an expected insert-size value of the pair, using the determined set of distances to select downstream nodes as potential alignment sites for a second member of the pair, identifying paths that begin with the upstream node and end with one of the downstream nodes as candidate paths, and excluding any paths that are not candidate paths from any alignment calculations involving the pair of paired-end reads; determining distances between the aligned pairs of the paired-end reads and the

frequencies of the distances between the aligned pairs; and determining a set of isoform paths and isoform frequencies such that representing the isoform paths through the DAG at the isoform frequencies results in pairs of features being included in the isoform paths at the frequencies of the distances between the aligned pairs. In a further aspect, the present invention refers to a system of inferring isoform frequency in a transcriptome, the system comprising a computer system comprising a processor coupled to a non-transitory memory having stored therein a reference DAG comprising a plurality of nodes representing exons and introns from features in a genetic feature database, wherein pairs of the plurality of nodes are connected in their canonical genomic order at their proximal ends by an edge, wherein the system is operable to align a plurality of pairs of paired-end reads to the reference DAG, wherein the aligning for a pair of paired-end reads comprises identifying an exon at which a first member of a pair aligns, determining a set of distances that accounts for an expected insert-size value of the pair, using the determined set of distances to select downstream nodes as potential alignment sites for a second member of the pair, identifying paths that begin with the upstream node and end with one of the downstream nodes as candidate paths, and excluding any paths that are not candidate paths from any alignment calculations involving the pair of paired-end reads; determine distances between the plurality of pairs of paired-end reads from cDNA and the frequencies of the distances between the pairs; and determine a set of isoform paths and isoform frequencies such that representing the isoform paths through the DAG at the isoform frequencies results in pairs of features being included in the isoform paths at the frequencies of the distances between the aligned pairs. In certain examples, the disclosure describes methods of analyzing a transcriptome that involves obtaining at least one pair of paired-end reads from a transcriptome from an organism, finding an alignment with an optimal score between a first read of the pair and a node in a directed acyclic data structure (the data structure has nodes representing RNA sequences such as exons or transcripts and edges connecting pairs of nodes), identifying candidate paths that

Other examples of the invention provide for expression analysis through, e.g., systems and methods useful for inferring isoform frequency.

In certain examples, disclosed herein are methods of inferring isoform frequency in a transcriptome. The method includes obtaining a plurality of pairs of paired-end reads from a transcriptome of an organism and optionally aligning those reads to a reference data structure. The data structure includes a plurality of features from a genome each represented as a node. Pairs of the nodes, or features, are connected at their proximal ends by an edge (i.e., within pairs, the two pair-members are connected to one another at their proximal ends on the genome by an edge; since any pair-member may also be a member of other pairs, either or both ends of a feature may be connected by one or any number of edges to other features). Distances between the aligned pairs of the paired-end reads and the frequencies of the distances between the aligned pairs are determined. A set of isoform paths and isoform frequencies are determined such that representing the isoform paths through the structure at the isoform frequencies results in pairs of features being included in the isoform paths at the frequencies of the distances the aligned pairs. Each path of the determined set of isoform paths may represent a transcript isoform present in the organism. The distances between the aligned pairs may be path lengths from an upstream end of a first aligned member of each pair to a downstream end of a second aligned member of that pair. In certain examples, determining the set of isoform path includes discovering a novel isoform path. The reference data structure may be updated to include the novel isoform path.

Determining the set of isoform paths can be performed algebraically, e.g., by solving a set of linear equations. In some examples, one of the isoform frequencies is obtained from biological data prior to solving the set of linear equations (e.g., so remaining variables in the linear equations can be determined). Additionally or alternatively, the frequencies of the distances between the aligned pairs may be normalized prior to determining a set of isoform paths and isoform frequencies.

The data structure may be built from an existing genetic feature database. For example, each node may represent a feature from such a database.

In one example, there is disclosed a system for transcriptomics, the system comprising a computer system comprising a processor coupled to a non-transitory memory having stored therein a directed acyclic data structure, the data structure comprising nodes

Claims

1. A method of analyzing a transcriptome, the method comprising:

obtaining a pair of paired-end reads from a transcriptome;

finding an alignment with an optimal score between a first read of the pair and a node representing an exon in a directed acyclic graph (DAG), the DAG comprising nodes representing a plurality of exons and introns from features in an annotated transcriptome database, and edges connecting pairs of the nodes in their canonical genomic order;

identifying candidate paths within the DAG that include the node connected to a downstream node by a path having a length substantially similar to an insert length of the pair of paired-end reads;

aligning the paired-end reads to the candidate paths to determine an optimal-scoring alignment by excluding any paths that are not candidate paths from any alignment calculations involving the pair of paired-end reads; and

identifying an isoform from the transcriptome using the optimal-scoring alignment of the paired-end reads.

2. The method of claim 1, further comprising obtaining a plurality of paired-end reads from the transcriptome.

3. The method of claim 2, further comprising determining optimal-scoring alignments for each of the plurality of pairs of paired-end reads.

4. The method of claim 3, wherein the plurality of paired-end reads are obtained from cDNA fragments from the transcriptome.

5. The method of claim 1, wherein the identified isoform is a novel isoform.

6. The method of claim 5, further comprising updating the DAG to represent the novel isoform.

7. The method of claim 6, wherein updating the DAG to represent the novel isoform comprises adding at least one new node.

8. The method of claim 7, wherein the node and the downstream node represent a pair of exons from which the pair of paired-end reads were obtained.

9. The method of claim 1, further comprising:

aligning a plurality of pairs of paired-end reads to the DAG;

determining distances between the aligned pairs of the plurality of pairs of paired-end reads and the frequencies of the distances between the aligned pairs; and

determining a set of isoform paths and isoform frequencies such that representing the isoform paths through the DAG at the isoform frequencies results in pairs of features being included in the isoform paths at the frequencies of the distances between the aligned pairs.

10. The method of claim 9, wherein the features include exons.

11. The method of claim 10, wherein each path of the determined set of isoform paths represents a transcript isoform present in the organism.

12. The method of claim 9, wherein the distances between the aligned pairs comprise path lengths from an upstream end of a first aligned member of each pair to a downstream end of a second aligned member of that pair.

13. A system for transcriptomics, the system comprising:

a computer system comprising a processor coupled to a non-transitory memory having stored therein a directed acyclic graph (DAG), the DAG comprising nodes representing a plurality of exons and introns from features in an annotated transcriptome database, and edges connecting pairs of the nodes in their canonical genomic order, wherein the system is operable to:

obtain a pair of paired-end reads from a cDNA fragment;

find an alignment with an optimal score between a first read of the pair and a node in the DAG;

identify candidate paths within the DAG that include the node connected to a downstream node by a path having a length substantially similar to an insert length of the paired end reads;

align the paired-end rends to the candidate paths to determine an optimal-scoring alignment by excluding any paths that are not candidate paths from any alignment calculations involving the pair of paired-end nodes; and

identify an isoform from the transcriptome using the optimal scoring-alignment of the paired-end reads.

14. The system of claim 13, further operable to obtain a plurality of paired-end reads from cDNA fragments.

15. The system of claim 13, further operable to update the DAG to represent the novel isoform.

16. The system of claim 15, wherein updating the DAG to represent the novel isoform comprises adding at least one new node.

17. A method of inferring isoform frequency in a transcriptome, the method comprising:

aligning a plurality of pairs of paired-end reads from cDNA to a reference directed acyclic graph (DAG) comprising a plurality of nodes representing exons and introns from annotated features in a genetic feature database, wherein pairs of the plurality of nodes are connected in their canonical genomic order at their proximal ends by an edge, wherein the aligning for a pair of paired-end reads comprises identifying an exon at which a first member of a pair aligns, determining a set of distances that accounts for an expected insert-size value of the pair, using the determined set of distances to select downstream nodes as potential alignment sites for a second member of the pair, identifying paths that begin with the upstream node and end with one of the downstream nodes as candidate paths, and excluding any paths that are not candidate paths from any alignment calculations involving the pair of paired-end reads;

determining distances between the aligned pairs of the paired-end reads and the frequencies of the distances between the aligned pairs; and

determining a set of isoform paths and isoform frequencies such that representing the isoform paths through the DAG at the isoform frequencies results in pairs of features being included in the isoform paths at the frequencies of the distances between the aligned pairs.

18. The method of claim 17, wherein the cDNA represents substantially an entire transcriptome of an organism.

19. The method of claim 18, wherein each path of the determined set of isoform paths represents a transcript isoform present in the organism.

20. The method of claim 19, wherein determining the set of isoform paths includes discovering a novel isoform path.

21. The method of claim 20, further comprising updating the reference DAG to include the novel isoform path.

22. The method of claim 17, wherein the distances between the aligned pairs comprise path lengths from an upstream end of a first aligned member of each pair to a downstream end of a second aligned member of that pair.

23. The method of claim 17, further comprising normalizing the frequencies of the distances between the aligned pairs prior to determining a set of isoform paths and isoform frequencies.

24. A system of inferring isoform frequency in a transcriptome, the system comprising:

a computer system comprising a processor coupled to a non-transitory memory having stored therein a reference DAG comprising a plurality of nodes representing exons and introns from features in a genetic feature database, wherein pairs of the plurality of nodes are connected in their canonical genomic order at their proximal ends by an edge, wherein the system is operable to:

align a plurality of pairs of paired-end reads to the reference DAG, wherein the aligning for a pair of paired-end reads comprises identifying an exon at which a first member of a pair aligns, determining a set of distances that accounts for an expected insert-size value of the pair, using the determined set of distances to select downstream nodes as potential alignment sites for a second member of the pair, identifying paths that begin with the upstream node and end with one of the downstream nodes as candidate paths, and excluding any paths that are not candidate paths from any alignment calculations involving the pair of paired-end reads;

determine distances between the plurality of pairs of paired-end reads from cDNA and the frequencies of the distances between the pairs; and

determine a set of isoform paths and isoform frequencies such that representing the isoform paths through the DAG at the isoform frequencies results in pairs of features being included in the isoform paths at the frequencies of the distances between the aligned pairs.

25. The system of claim 24, wherein the cDNA represents substantially an entire transcriptome of an organism.

26. The system of claim 25, wherein each path of the determined set of isoform paths represents a transcript isoform present in the organism.

27. The system of claim 26, wherein determining the set of isoform paths includes discovering a novel isoform path.

28. The system of claim 27, further operable to update the reference DAG to include the novel isoform path.

29. The system of claim 24, wherein the distances between the aligned pairs comprise path lengths from an upstream end of a first aligned member of each pair to a downstream end of a second aligned member of that pair.

30. The system of claim 24, further operable to normalize the frequencies of the distances between the aligned pairs prior to determining the set of isoform paths and isoform frequencies.