Methods for transcriptome analysis based on next-generation sequencing (NGS) technologies have been reported. See, e.g., Wang, et al., 2009, RNA-Seq: a revolutionary tool for transcriptomics, Nat Rev Genet 10(1):57-63 2009. The RNA-Seq (RNA sequencing) method promises to rapidly generate high volumes of transcriptome data.

However, RNA-Seq faces informatics challenges, such as storing and analyzing large amounts of data, which must be overcome to make good use of the reads. Aligning and analyzing RNA-Seq reads presents problems not only in the nature of the information involved but also in the volume. Read mapping with large data sets is computationally challenging and analytical methods for differential expression are only beginning to emerge. See Garber, et al., 2011, Computational methods for transcriptome annotation and quantification using RNA-Seq, Nat Meth 8(6):469-477. The computational demands of mapping the large number of reads from RNA-Seq are greatly multiplied by the large amount of reference data that are coming available.

For example, the GENCODE Consortium seeks to identify all gene features in the human genome that have been reported. Harrow, et al., 2012, GENCODE: The reference human genome annotation for The ENCODE Project, Genome Res 22:1760-1774. The volume of material encompassed by the GENCODE project is formidable. Not only do new protein-coding loci continue to be added, the number of alternative splicing transcripts annotated steadily increases. The GENCODE 7 release includes more than 20,000 protein-coding and almost 10,000 long noncoding RNA loci (lncRNA), as well as more than 33,000 coding transcripts not represented in other sources. GENCODE also includes other features such as untranslated regions (UTRs), long intergenic noncoding RNA (lincRNA) genes, short noncoding RNAs, and alternative splice patterns. Even if an RNA-Seq study started with only a limited amount of new data, the volume of potential reference data from a source such as GENCODE is so large that making sense of the new data is computationally challenging.

Summary

The ~~invention~~ disclosure generally provides systems and methods for analysis of paired-end reads using a reference organized as a directed acyclic graph (DAG) or similar data structure. Features in the reference, such as exons and introns, provide nodes in the DAG and those features are linked to one another in their canonical genomic order by edges. The DAG can scale to any size and can, in fact, be populated in the first instance by imported sequences from an extrinsic

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.

~~Aspects of the invention~~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 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 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 certain ~~aspects~~examples, the ~~invention provides~~ 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

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 rends to the candidate paths to determine an optimal-scoring alignment. The method may include excluding any paths that are not candidate paths from any alignment calculations involving the pair of paired-end reads.

The method is suited for any number of paired-end reads. The node and the downstream node may represent a pair of exons from which the pair of paired-end reads were obtained. The method may include identifying an isoform based on the determined optimal-scoring alignment. In some ~~embodiments~~examples, the identified isoform is a novel isoform. The directed acyclic data structure may be updated to represent the novel isoform (e.g., by adding at least one new node or edge).

~~Related aspects of the invention provide~~The disclosure describes a computer system for transcriptomics. The system has a processor coupled to a non-transitory memory and is used to store a directed acyclic data structure having nodes representing RNA sequences and edges connecting pairs of the nodes. The system is operable to obtain a pair of paired-end reads from a transcriptome, find an alignment with an optimal score between a first read of the pair and a node in the data structure, identify 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 align the paired-end rends to the candidate paths to determine an optimal-scoring alignment. Preferably, the system excludes all non-candidate paths from subsequent alignment attempts. In certain ~~embodiments~~examples, the system is further operable to obtain a plurality of paired-end reads from cDNA fragments. Preferably the system is operable to determine optimal-scoring alignments for each of the plurality of pairs of paired-end reads. The cDNA fragments may represent a single transcriptome. The system may identify an isoform, such as a novel isoform, based on the determined optimal-scoring alignment. The system may be operable to update the directed acyclic data structure to represent the novel isoform. In some ~~embodiments~~examples, updating the directed acyclic data structure to represent the novel isoform comprises adding at least one new node. A node and a downstream node may represent a pair of exons from which the pair of paired-end reads were obtained. The system may be operable to exclude any paths that are not candidate paths from any alignment calculations involving the pair of paired-end reads.

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

In certain ~~aspects~~examples, ~~the invention provides~~ 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 ~~embodiments~~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 ~~embodiments~~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 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 data structure, the data structure comprising nodes

representing RNA sequences and edges connecting pairs of the nodes, 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 data structure; identify 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 paired end reads; and align the paired-end rends to the candidate paths to determine an optimal-scoring alignment. In related ~~aspects~~examples, ~~the invention provides~~ disclosed herein is a computer system of inferring isoform frequency in a transcriptome. The system includes a processor coupled to a non-transitory memory and stores a reference data structure comprising a plurality of features from a genome each represented as a node wherein pairs of the plurality of features are connected at their

proximal ends by an edge. The system can be used to obtain or determine distances between a plurality of pairs of paired-end reads from a transcriptome and the frequencies of the distances between the pairs. A set of isoform paths and isoform frequencies can then be 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. The features may include exons and the system is further operable to align a plurality of pairs of paired-end reads to the reference data structure. Each path of the determined set of isoform paths may represent a transcript isoform present in the organism. Determining the set of isoform path may include discovering a novel isoform path. The system may update the reference data structure to include the novel isoform path. The distances between the aligned pairs may 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. In some ~~embodiments~~examples, determining the set of isoform paths comprises solving a set of linear equations. In certain ~~embodiments~~examples, determining the set of isoform paths comprises obtaining one of the isoform frequencies from biological data prior to solving the set of linear equations. The system may be operable to normalize the frequencies of the distances between the aligned pairs prior to determining the set of isoform paths and isoform frequencies. The data structure may include a node representing each annotated feature in a remote genetic feature database.

Brief Description of the Drawings

FIG. 1 shows a pictorial representation of the read being compared to the DAG.

FIG. 2 shows the actual matrices that correspond to the comparison.

FIG. 3 provides a diagram of methods of the invention.

FIG. 4 diagrams a hypothetical DAG and pair of paired-end reads.

FIG. 5 shows candidate paths through the hypothetical DAG.

FIG. 6 shows the candidate paths from the hypothetical DAG.

FIG. 7 presents three alternative histories involving sequences.

FIG. 8 depicts an alignment according to the invention.

FIG. 9 shows an isoform frequency distribution.

FIG. 10 illustrates a system for implementing methods of the invention.