mutations, and they lack genomic context. The results from a SNP chip are akin to being told the first letter that appears on each page of a book—the data points may reveal certain mistakes or patterns, but they do not give the whole story.

Substantial portions of a genome can be sequenced by next generation sequencing (NGS) technologies to produce very large numbers of short sequence reads. Analyzing those reads is a substantial challenge. Existing methods typically involving mapping the reads to a reference genome or performing a de novo assembly. Due to the volume of data, alignments and assemblies are performed with heuristic algorithms that must take mathematical shortcuts to get the job done. However, those imperfections are tolerated since mathematically thorough algorithms are computationally prohibitive. Not only is mathematical precision sacrificed in analyzing NGS reads, current methods make limited use of available information. NGS reads are often mapped to a somewhat arbitrary reference and called as either matching that reference or not—wholly begging the question that the reference is a useful standard.

Summary

The ~~invention provides~~disclosure describes methods of using identified nucleotides at known places in a genome to guide the analysis of sequence reads from that genome by excluding potential mappings or assemblies that are not congruent with the identified nucleotides. Where reads are mapped to a directed acyclic graph (DAG) that represents multiple alternative sequences per position at multiple positions, paths through the DAG that do not include the identified nucleotides can be excluded from read mapping with the result that computer processing time need not be spent fruitlessly comparing reads to irrelevant regions of a reference. That savings in computational resources can be applied to mapping and assembly algorithms allowing, for example, a read to be aligned to paths in the DAG using mathematically robust algorithms such as Smith-Waterman-based alignment algorithms that find the best alignment. The ~~provided~~ described methods allow for a very large number of NGS reads to be mapped or assembled, so that genome sequencing projects can proceed at a pace, depth-of-coverage, and throughput not previously possible.

The reference DAG can represent any or all known or speculated genotypes at multiple loci. As a result, sequence reads are compared to more than just a limited, linear reference. Sequence reads are compared to a variety of known variants, avoiding the misleading inferences

made by aligning to a single reference. A plurality of known variants can be documented within the DAG and read mapping can include discovering a match to a known variant in ways not otherwise possible. The DAG may include paths that are naturally sensible references for a particular subject, such as sequences from other members of a same population or ethnicity. Mapping to a DAG makes read assembly more successful as paths through the DAG will be a better fit than a linear reference and also makes results more meaningful as a subject's genome can be understood as it relates to numerous other members of the population.

Limiting the DAG search space, that is, filtering the DAG, using identified nucleotides in a subject's genome as provided by, for example, an SNP chip, makes read mapping rapid and thorough with better results and high throughput. As a result, personal genomic studies are more accessible and the medically significant information such studies can reveal will be brought to bear in improving people's lives.

Additionally, use of identified nucleotides (e.g., from an SNP chip) can improve the results from an alignment. Data can be obtained relating identified nucleotides to other nucleotides in the genome. For example, many genetic loci are in linkage disequilibrium with one another. Where a locus in a genome is identified (e.g., by a SNP chip), another locus that is linkage disequilibrium with the identified locus may then have a certain probability of a certain allele in that genome. Mapping reads to the linked locus can include incorporating the probability into an alignment algorithm, thus improving the resulting alignment by increasing the probability that the resulting alignment shows the natural relationship between the subject's genome and the other genomes represented within the DAG.

In one aspect, the present invention refers to a system for determining a genomic sequence, the system comprising a computer system comprising a processor coupled to memory and operable to receive identities of a plurality of nucleotides at known locations on a genome of a subject; select one or more genomic sequences from a plurality of genomic sequences, wherein the selected sequences include the nucleotides at the known locations; receive sequence reads from a sample from the subject; and map the sequence reads to the one or more selected genomic sequences, thereby identifying a sequence of at least a portion of the genome. In another aspect, the present invention refers to a method for determining a genomic sequence, the method comprising receiving, at a computer system, identities of a plurality of nucleotides at known locations on a genome of a subject; selecting one or more genomic sequences from a plurality of genomic sequences stored in the computer system, wherein the selected sequences include the nucleotides at the known locations; receiving sequence reads from a sample from the subject; and mapping the sequence reads to the selected genomic sequences, thereby determining a sequence of at least a portion of the genome In certain ~~aspects~~examples, ~~the invention provides~~there is disclosed a computer system for determining a genomic sequence. The computer system uses a processor coupled to memory to receive identities of a plurality of nucleotides at known locations on a genome of a subject, select one or more genomic sequences from a plurality of genomic sequences—wherein the selected sequences include the nucleotides at the known locations, and receive sequence reads from a sample from the subject. The system can map the sequence reads to the one or more selected genomic sequences, thereby identifying a sequence of at least a portion of the genome. In a preferred ~~embodiment~~example, the plurality of genomic sequences is stored in memory as a DAG comprising a plurality of node and edge objects that each stores a list of pointers to locations in memory of adjacent ones of the node and edge objects. Each of the selected genomic sequences defines a selected path through

the DAG and the system maps the sequence reads by finding optimal alignments between the sequence reads and the selected paths. Finding an optimal alignment may be done by finding a highest-scoring trace through the DAG. This highest scoring trace may be found by calculating match scores between the reads and at least some of the node and edge objects in the DAG and dereferencing at least some of the pointers to read predecessor objects in the DAG from their referenced locations in memory. A path through predecessor objects with a maximal sum of match scores is the highest-scoring trace through the DAG.

In some ~~embodiments~~examples, the system maps the sequence reads by finding optimal alignments between each sequence read and each selected path. Finding an optimal alignment may include finding a highest-scoring trace through a multi-dimensional matrix. They system may be operated to obtain probabilities for identities of additional nucleotides based on the identities of the plurality of nucleotides at the known locations and use the obtained probabilities in finding the optimal alignments. The probabilities could be obtained from measures of linkage disequilibrium between ones of the additional nucleotides and ones of the plurality of nucleotides at the known locations.

In certain ~~embodiment~~examples, the plurality of genomic sequences is stored in memory as a DAG comprising a plurality of node and edge objects that each stores a list of pointers to locations in memory of adjacent ones of the node and edge objects, wherein each of the selected genomic sequences defines a selected path through the DAG and wherein the system is operable to map the sequence reads by finding optimal alignments between the sequence reads and the selected paths. Finding an optimal alignment may be done by finding a highest-scoring trace through the DAG by calculating match scores between the reads and at least some of the node and edge objects in the DAG and dereferencing at least some of the pointers to read predecessor objects in the DAG from their referenced locations in memory, wherein a path through predecessor objects with a maximal sum of match scores is the highest-scoring trace through the DAG.

In certain ~~embodiment~~examples, the plurality of genomic sequences are stored in the memory as a DAG comprising a plurality of node and edge objects, wherein the identities of a plurality of nucleotides comprise a plurality of SNPs in the subject's genome and the system selects the one or more genomic sequences that include the SNPs by identifying candidate paths through the DAG that include the plurality of SNPs. In some ~~embodiment~~examples, the DAG represents at least one

locus homologous across multiple genomes with a single object for each allele. The DAG may be annotated to list non-compossible node-pairs, and the system may further be operable to identify a node in the list of non-compossible node pairs from one of the plurality of SNPs, identify a second node paired to the identified node in the list of non-compossible node pairs, and exclude paths containing the second node when mapping the sequence reads.

In the system, the data representing the plurality of genomic sequences may include nodes representing nucleotide sequences and edges connecting pairs of the nodes wherein the nodes and edges define a directed acyclic graph, further wherein each of the selected genomic sequences defines a selected path through the directed acyclic graph. Preferably, the directed acyclic graph represents at least two alternative sequences per position at multiple positions.

In some ~~aspects~~examples, ~~the invention provides~~disclosed herein is a method for determining a subject's genetic information by obtaining information about a plurality of SNPs in the subject's genome and identifying, within a directed acyclic data structure (or DAG) comprising nodes representing genetic sequences and edges connecting pairs of the nodes, candidate paths that include the plurality of SNPs. Sequence reads are obtained from the subject's genome and mapped to the candidate paths. The DAG may represent at least one locus homologous across multiple genomes with a single node for each allele. The DAG may be annotated to list non-compossible node-pairs and the method can include identifying a node in the list of non-compossible node pairs from one of the plurality of SNPs, identifying a second node paired to the identified node in the list of non-compossible node pairs, and excluding paths containing the second node from the mapping step.

~~Aspects~~ Examples of the invention ~~provide~~ disclose a method for determining a genomic sequence by receiving, at a computer system, identities of a plurality of nucleotides at known locations on a genome of a subject (e.g., from a microarray or SNP chip) and selecting one or more genomic sequences from a plurality of genomic sequences stored in the computer system, wherein the selected sequences include the nucleotides at the known locations. The method includes receiving sequence reads from a sample from the subject and mapping the sequence reads to the selected genomic sequences, thereby determining a sequence of at least a portion of the genome.

The data representing the plurality of genomic sequences can be a DAG with nodes representing nucleotide sequences and edges connecting pairs of the nodes. Each of the selected

genomic sequences defines a selected path through the DAG. Mapping the sequence reads may be done by finding optimal alignments between the sequence reads and the selected paths.

The method can further include obtaining probabilities for identities of additional nucleotides based on the identities of the plurality of nucleotides at the known locations and using the obtained probabilities in finding the optimal alignments. The probabilities can be obtained from measures of linkage disequilibrium between ones of the additional nucleotides and ones of the plurality of nucleotides at the known locations.

The DAG may be of any suitable scale including genome-scale or scaled to represent genes or regions. In some ~~embodiment~~examples, a DAG may be a genome-scale DAG and may include at least one path through the DAG that gives a substantially entire sequence of at least one human chromosome. In certain ~~embodiment~~examples, a DAG is of a scale to represent regions smaller than a chromosome, e.g., a gene; a region surrounding a gene; an operon; a chromosome from an organelle, bacterium, virus or plasmid; a large SV element; etc. The DAG represents at least two alternative sequences per position at multiple positions. Once mapped to the DAG (or through the mapping), the reads can be assembled.

In related ~~aspects~~examples, the ~~invention provides~~ disclosure describes a system for determining a genomic sequence. The system includes a processor coupled to memory and is operable to receive identities of a plurality of nucleotides at known locations on a genome of a subject and select one or more genomic sequences from a plurality of genomic sequences (e.g., from within a DAG), wherein the selected sequences include the nucleotides at the known locations. The system can be used to receive sequence reads from a sample from the subject and map the sequence reads to the one or more selected genomic sequences, thereby identifying a sequence of at least a portion of the genome. The system may further be operable to obtain probabilities for identities of additional nucleotides based on the identities of the plurality of nucleotides at the known locations and use the obtained probabilities in finding the optimal alignments. The probabilities may be obtained from measures of linkage disequilibrium between ones of the additional nucleotides and ones of the plurality of nucleotides at the known locations.

Brief Description of the Drawings

FIG. 1 illustrates using a DAG to represent and manipulate bioinformatic data.