

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 a memory and operable to receive identities of a plurality of single nucleotide polymorphisms (SNPs) at known locations on a genome of a subject; receive a plurality of genomic sequences as a genomic directed acyclic graph (DAG) that represents at least two alternative sequences per position at multiple positions, the DAG comprising a plurality of nodes and edges, the nodes representing nucleotide sequences and edges connecting pairs of the nodes; identify a plurality of candidate paths through the genomic DAG by filtering paths through the DAG that are not congruent with the identities of the plurality of SNPs, wherein the filtering further comprises excluding nodes by determining which nodes in the DAG are traversed by the identities of the plurality of SNPs and using conditional information to deduce the relevance of other nodes given the traversed nodes; receive sequence reads from a sample from the subject; and map the sequence reads to the plurality of candidate paths, 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 subject's genetic information, the method comprising obtaining information about the identities of a plurality of SNPs in the subject's genome; identifying, within a genomic directed acyclic graph (DAG) that represents at least two alternative sequences per position at multiple positions, the DAG comprising a plurality of nodes and edges, the nodes representing genetic sequences and edges connecting pairs of the nodes, a plurality of candidate paths through the DAG that are not congruent with the identities of the plurality of SNPs, wherein the plurality of candidate paths are identified by excluding nodes by determining which nodes in the DAG are traversed by the identities of the plurality of SNPs and using conditional information to deduce the relevance of other nodes given the traversed nodes; obtaining a plurality of sequence reads from the subject's genome; and mapping the plurality of sequence reads to the candidate paths. In yet 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 single nucleotide polymorphisms

(SNPs) at known locations on a genome of a subject; receive a plurality of genomic sequences as a genomic directed acyclic graph (DAG) that represents at least two alternative sequences per position at multiple positions, the DAG stored in the computer system and comprising a plurality of node objects and edge objects, the node objects representing nucleotide sequences and edge objects connecting pairs of the node objects; identifying a plurality of candidate paths through the genomic DAG by filtering paths through the DAG that are not congruent with the identities of the plurality of SNPs, wherein the filtering further comprises excluding nodes by determining which nodes in the DAG are traversed by the identities of the plurality of SNPs and using conditional information to deduce the relevance of other nodes given the traversed nodes; receiving sequence reads from a sample from the subject; and mapping the sequence reads to the plurality of candidate paths, thereby determining a sequence of at least a portion of the genome. In one example, there is disclosed 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 example, there is disclosed 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 examples, 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 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

What is claimed is:

1. A system for determining a genomic sequence, the system comprising:
 - a computer system comprising a processor coupled to a memory and operable to:
 - receive identities of a plurality of single nucleotide polymorphisms (SNPs) at known locations on a genome of a subject;
 - receive a plurality of genomic sequences as a genomic directed acyclic graph (DAG) that represents at least two alternative sequences per position at multiple positions, the DAG comprising a plurality of nodes and edges, the nodes representing nucleotide sequences and edges connecting pairs of the nodes;
 - identify a plurality of candidate paths through the genomic DAG by filtering paths through the DAG that are not congruent with the identities of the plurality of SNPs, wherein the filtering further comprises excluding nodes by determining which nodes in the DAG are traversed by the identities of the plurality of SNPs and using conditional information to deduce the relevance of other nodes given the traversed nodes;
 - receive sequence reads from a sample from the subject; and
 - map the sequence reads to the plurality of candidate paths, thereby identifying a sequence of at least a portion of the genome.
2. The system of claim 1, wherein mapping the sequence reads comprises finding optimal alignments between each sequence read and each selected path.
3. The system of claim 2, wherein finding an optimal alignment comprises finding a highest-scoring trace through a multi-dimensional matrix.
4. The system of claim 2, further operable to:
 - obtain probabilities for identities of additional SNPs based on the identities of the SNPs at the known locations; and
 - use the obtained probabilities in finding the optimal alignments.

5. The system of claim 4, wherein the probabilities are obtained from measures of linkage disequilibrium between ones of the additional SNPs and ones of the plurality of SNPs at the known locations.
6. The system of claim 1, wherein the plurality of node and edge objects each store a list of pointers to locations in memory of adjacent ones of the node and edge objects, and wherein the system is operable to map the sequence reads by finding optimal alignments between the sequence reads and the selected paths.
7. The system of claim 6, wherein finding an optimal alignment comprises 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;
 - 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.
8. The system of claim 1, wherein the DAG represents at least one locus homologous across multiple genomes with a single object for each allele.
9. The system of claim 1, wherein the DAG is annotated to list non-compossible node-pairs, and wherein the system is further 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.
10. A method for determining a subject's genetic information, the method comprising:
 - obtaining information about the identities of a plurality of SNPs in the subject's genome;

identifying, within a genomic directed acyclic graph (DAG) that represents at least two alternative sequences per position at multiple positions, the DAG comprising a plurality of nodes and edges, the nodes representing genetic sequences and edges connecting pairs of the nodes, a plurality of candidate paths through the DAG that are not congruent with the identities of the plurality of SNPs, wherein the plurality of candidate paths are identified by excluding nodes by determining which nodes in the DAG are traversed by the identities of the plurality of SNPs and using conditional information to deduce the relevance of other nodes given the traversed nodes; obtaining a plurality of sequence reads from the subject's genome; and mapping the plurality of sequence reads to the candidate paths.

11. The method of claim 10, wherein the DAG represents at least one locus homologous across multiple genomes with a single node for each allele.

12. The method of claim 10, wherein filtering paths through the DAG that are not congruent with the identities of the plurality of SNPs comprises:

identifying a node in a 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.

13. A method for determining a genomic sequence, the method comprising:

receiving, at a computer system, identities of a plurality of single nucleotide polymorphisms (SNPs) at known locations on a genome of a subject;

receive a plurality of genomic sequences as a genomic directed acyclic graph (DAG) that represents at least two alternative sequences per position at multiple positions, the DAG stored in the computer system and comprising a plurality of node objects and edge objects, the node objects representing nucleotide sequences and edge objects connecting pairs of the node objects;

identifying a plurality of candidate paths through the genomic DAG by filtering paths through the DAG that are not congruent with the identities of the plurality of SNPs, wherein the filtering further comprises excluding nodes by determining which nodes in the DAG are

traversed by the identities of the plurality of SNPs and using conditional information to deduce the relevance of other nodes given the traversed nodes;

receiving sequence reads from a sample from the subject; and

mapping the sequence reads to the plurality of candidate paths, thereby determining a sequence of at least a portion of the genome.

14. The method of claim 13, wherein the plurality of node and edge objects each stores a list of objects that it is adjacent to.

15. The method of claim 13, wherein each node and edge object stores a list of one or more adjacent objects, and wherein mapping the sequence reads comprises finding optimal alignments between the sequence reads and the candidate paths.

16. The method of claim 15, wherein each list comprises pointers to locations in memory of the adjacent objects, and wherein finding an optimal alignment comprises finding a highest-scoring trace through a multi-dimensional matrix.

17. The method of claim 16, wherein finding the highest-scoring trace comprises using the computer system to calculate match scores between the reads and at least some of the node and edge objects in the DAG, and looking backwards to predecessor objects in the DAG to identify a back-trace with an optimal score.

18. The method of claim 15, further comprising:

obtaining probabilities for identities of additional SNPs based on the identities of the plurality of SNPs at the known locations, wherein the probabilities are obtained from measures of linkage disequilibrium between ones of the additional SNPs and ones of the plurality of SNPs at the known locations; and

using the obtained probabilities in finding the optimal alignments.

19. The method of claim 14, wherein at least one path through the DAG gives a substantially entire sequence of at least one human chromosome.

20. The method of claim 13, further comprising assembling the mapped sequence reads.

21. The method of claim 13, wherein the identities of the plurality of SNPs at the known locations are received as results from a microarray assay.