

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

In one aspect, the present invention provides 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 genomic directed acyclic graph (DAG) that represents a complete genome and known variation in that genome, wherein the DAG further 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 that include the plurality of SNPs 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, wherein the conditional information comprises the non-compossibility of nodes in the genomic DAG; 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 provides 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 a complete genome and known variation in that genome, and further 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 include the plurality of SNPs 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, wherein the conditional information comprises the non-compossibility of nodes in the genomic DAG; 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 provides 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 genomic directed acyclic graph (DAG) that represents a complete genome and known variation in that genome, wherein the DAG further 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 that include the plurality of SNPs 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, wherein the conditional information comprises the non-compossibility of nodes in the genomic DAG; 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.

The 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 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.

Disclosed is 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. Also disclosed is 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. Also disclosed is a method for determining a genomic sequence, the method comprising receiving, at a computer system, identities of a plurality of single nucleotide polymorphisms

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 genomic directed acyclic graph (DAG) that represents a complete genome and known variation in that genome, wherein the DAG further 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 that include the plurality of SNPs 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, wherein the conditional information comprises the non-compossibility of nodes in the genomic DAG;
 - 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 a complete genome and known variation in that genome, and further 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 include the plurality of SNPs 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, wherein the conditional information comprises the non-compossibility of nodes in the genomic DAG;

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 genomic directed acyclic graph (DAG) that represents a complete genome and known variation in that genome, wherein the DAG further 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 that include the plurality of SNPs 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, wherein the conditional information comprises the non-compossibility of nodes in the genomic DAG;

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.