

DELETION_PENALTY: -5

The relationship between the gap penalties (INSERTION_PENALTY, OPENING_PENALTY) above help limit the number of gap openings, i.e., favor grouping gaps together, by setting the gap insertion penalty higher than the gap opening cost. Of course, alternative relationships between MISMATCH_PENALTY, MATCH_BONUS, INSERTION_PENALTY, OPENING_PENALTY and DELETION_PENALTY are possible.

While the alignment methods, described above, have been useful for assembling reads produced with next-generation sequencing techniques, they are complex and time-consuming. Additionally, these techniques are ill-suited for identifying the important nuances between diseased cells of varying disease states because the uncertainty due to aligning the reads to a common reference often drowns out small changes in the genome.

SUMMARY

The disclosure provides improved methods and systems for identifying mutations that are induced by, or associated with, disease, especially cancer. The methods allow specific changes associated with advanced stages of the disease to be easily differentiated from lesser diseased cells, thus providing insight into the relationship between the size and location of the mutations and progression of the disease. This insight can be used to identify disease progression in other patients, and the relationship also provides for faster and more accurate typing of samples later collected from the same patient to monitor for disease progression or recurrence. In one aspect, the present invention refers to a method of identifying cancer-induced genetic mutations, comprising obtaining a first nucleic acid sequence corresponding to a nucleic acid in a non-cancerous sample from a subject; identifying differences between the first nucleic acid sequence and a selected non- reference sequence; representing the identified differences between the first nucleic acid sequence and the selected reference sequence as two or more alternative paths in a first reference directed acyclic graph (DAG) comprising nodes, wherein each alternative path is placed at a position in the first reference DAG where there is a difference between the first nucleic acid sequence and the reference sequence; aligning one or more sequence reads from a second sequence corresponding to a cancerous sample from the subject to the first reference DAG , wherein the aligning considers the two or more alternative paths by looking backward to any prior nodes on the first reference DAG to find a maximum score for the one or more sequence reads; and identifying, based on the aligned one or more sequence reads to the first reference DAG, differences between the second sequence and the

first reference DAG as new mutations correlated with the cancer. In another aspect, the present invention refers to a method of identifying mutations due to an advanced stage of cancer in a subject, comprising obtaining a first sequence corresponding to a non-cancerous sample from the subject and a second sequence corresponding to a cancerous sample from the subject; identifying differences between the first sequence and the second sequence; representing the identified differences between the first sequence and the second sequence as two or more alternative paths in a reference directed acyclic graph (DAG) comprising nodes, wherein each alternative path is placed at a position in the reference DAG where there is a difference between the first sequence and the second sequence; aligning a sequence read corresponding to an advanced cancerous sample from the subject to the reference DAG, wherein the aligning considers the two or more alternative paths by looking backward to any prior nodes on the reference DAG to find a maximum score for the sequence read; and identifying, based on the aligned sequence read to the reference DAG, differences between the sequence read and the reference DAG as new mutations correlated with an advanced stage of the cancer. In yet another aspect, the present invention refers to a method for assessing progression of cancer in a subject, comprising representing variations between the genetic sequence of non-cancerous cells of a subject and the genetic sequence of cancerous cells of the subject as two or more alternative paths in a directed acyclic graph comprising nodes, wherein each alternative path is placed at positions in the directed acyclic graph where there is a variation; aligning a first sequence read corresponding to a genetic sample later collected from the subject to the directed acyclic graph; determining, based on the aligned plurality of sequence reads, differences between the first sequence read and the directed acyclic graph, wherein differences are correlated with progression of the disease. In a further example, a method for determining a genetic marker for metastatic cancer, comprising representing variations between the genetic sequence of non-cancerous cells of a subject and the genetic sequence of non-metastatic cancerous cells of the subject as two or more alternative paths in a directed acyclic graph comprising nodes, wherein each alternative path is placed at positions in the directed acyclic graph where there is a variation; aligning a plurality of sequence reads corresponding to a genetic sample from a metastatic cell of the subject to the directed acyclic graph; and determining, based on the aligned plurality of sequence reads, differences between the sequence reads and the directed acyclic graph, thereby determining at least one genetic marker correlated with the metastatic clone. In one example, there is disclosed a method of identifying disease-induced genetic mutations, comprising obtaining a first nucleic acid sequence corresponding to a nucleic acid in a non-diseased sample from an organism; identifying differences between the first sequence and a selected non-diseased reference sequence; preparing a first reference sequence

construct representing the differences between the first sequence and the selected reference sequence as two or more alternative paths at positions in the first reference sequence construct where there is a difference between the first sequence and the reference sequence; aligning one or more reads from a second sequence corresponding to a diseased sample from the organism to the first reference sequence construct; and identifying differences between the second sequence and the first reference sequence construct as mutations due to a disease. In another example, there is disclosed a method of identifying mutations due to an advanced stage of disease in an organism, comprising: obtaining a first sequence corresponding to a non-diseased sample from an organism and a second sequence corresponding to a diseased sample from the organism; identifying differences between the first sequence and the second sequence; preparing a reference sequence construct representing the differences between the first sequence and the second sequence as two or more alternative paths at positions in the reference sequence construct where there is a difference between the first sequence and the second sequence; aligning a sequence read from the organism to the reference sequence construct; and identifying differences between the sequence read and the reference sequence construct as mutations due to an advanced stage of the disease. In yet another example, there is disclosed a method for assessing progression of a known genetic disease in a subject, comprising: creating a directed acyclic graph representing the genetic sequence of non-diseased cells of a subject and the genetic sequence of diseased cells of the subject; aligning a first sequence read corresponding to a genetic sample from the subject to the directed acyclic graph, determining differences between the first sequence read and the directed acyclic graph, wherein differences indicate progression of the disease. In a further aspect, the present invention refers to a method for determining a genetic marker for metastatic cancer, comprising creating a directed acyclic graph representing variations between the genetic sequence of non-cancerous cells of a subject and the genetic sequence of non-metastatic cancerous cells of the subject; aligning a plurality of sequence reads corresponding to a genetic sample from a metastatic cell of the subject to the directed acyclic graph, determining differences between the sequence reads and the directed acyclic graph, thereby determining a genetic marker for the metastatic clone.

The disclosure employs multi-dimensional reference sequence constructs and alignment algorithms that allow new sequence samples to be simultaneously compared to multiple sequences relevant to the disease, thereby providing increased speed and accuracy in disease identification and typing. Furthermore, the reference sequence constructs of the disclosure accommodate structural variations, deletions, insertions, and polymorphisms between samples in a straightforward way,

allowing a single construct to be assembled spanning an entire chromosome or the whole genome of a patient. Using a “look-back” type analysis, the described algorithms can also be used to align new reads in a multi-dimensional space including elements from sequences of various states of disease progression to provide more accurate alignment of sequence reads, while achieving lower error rates. Alternatively, constructs of the disclosure can be used to identify and/or study variations between individuals or cohorts having similar disease

CLAIMS

1. A method of identifying cancer-induced genetic mutations, comprising:
 - obtaining a first nucleic acid sequence corresponding to a nucleic acid in a non-cancerous sample from a subject;
 - identifying differences between the first nucleic acid sequence and a selected ~~non~~-reference sequence;
 - representing the identified differences between the first nucleic acid sequence and the selected reference sequence as two or more alternative paths in a first reference directed acyclic graph (DAG) comprising nodes, wherein each alternative path is placed at a position in the first reference DAG where there is a difference between the first nucleic acid sequence and the reference sequence;
 - aligning one or more sequence reads from a second sequence corresponding to a cancerous sample from the subject to the first reference DAG, wherein the aligning considers the two or more alternative paths by looking backward to any prior nodes on the first reference DAG to find a maximum score for the one or more sequence reads; and
 - identifying, based on the aligned one or more sequence reads to the first reference DAG, differences between the second sequence and the first reference DAG as new mutations correlated with the cancer.
2. The method of claim 1, further comprising:
 - preparing a second reference DAG representing the differences between the first sequence, the second sequence, and the reference sequence as two or more alternative paths at positions in the second reference DAG where there is a difference between the first sequence and the reference sequence or where there is a difference between the second sequence and the first sequence;
 - aligning one or more reads from a third sequence corresponding to an advanced cancerous sample from the organism to the second reference DAG; and
 - identifying differences between the third sequence and the second reference DAG as mutations due to the advanced cancer.

3. The method of claim 1, wherein the second sequence represents a major genetic clone due to the cancer.
4. The method of claim 2, wherein the third sequence represents a minor genetic clone due to the cancer.
5. The method of any of claims 1 to 4, wherein the cancer is selected from breast, lung, skin, prostate, thyroid, pancreatic, bladder, or ovarian cancer.
6. The method of claims 1 to 4, wherein the cancer is leukemia or lymphoma.
7. The method of claim 1, wherein the sequence reads are at least about 50 bp in length.
8. The method of claim 7, wherein the sequence reads are at least about 100 bp in length.
9. The method of claim 1, wherein the differences between the second sequence and the first reference DAG comprise insertions, deletions, polymorphisms or structural variants.
10. The method of claim 1, wherein the first reference DAG is at least about 1,000,000 bp in length.
11. The method of claim 1, wherein the first reference DAG represents a chromosome.
12. The method of claim 1, wherein the first reference DAG represents a genome.
13. A method of identifying mutations due to an advanced stage of cancer in a subject, comprising:
 - obtaining a first sequence corresponding to a non-cancerous sample from the subject and a second sequence corresponding to a cancerous sample from the subject;
 - identifying differences between the first sequence and the second sequence;

representing the identified differences between the first sequence and the second sequence as two or more alternative paths in a reference directed acyclic graph (DAG) comprising nodes, wherein each alternative path is placed at a position in the reference DAG where there is a difference between the first sequence and the second sequence;

aligning a sequence read corresponding to an advanced cancerous sample from the subject to the reference DAG, wherein the aligning considers the two or more alternative paths by looking backward to any prior nodes on the reference DAG to find a maximum score for the sequence read; and

identifying, based on the aligned sequence read to the reference DAG, differences between the sequence read and the reference DAG as new mutations correlated with an advanced stage of the cancer.

14. The method of claim 13, further comprising diagnosing the subject as having an advanced stage of the cancer.

15. The method of claim 13 or 14, wherein the cancer is selected from breast, lung, skin, prostate, thyroid, pancreatic, bladder, or ovarian cancer.

16. The method of claim 13 or 14, wherein the cancer is leukemia or lymphoma.

17. The method of claim 13 or 14, wherein the advanced stage of the cancer is metastatic cancer.

18. A method for assessing progression of cancer in a subject, comprising:

representing variations between the genetic sequence of non-cancerous cells of a subject and the genetic sequence of cancerous cells of the subject as two or more alternative paths in a directed acyclic graph comprising nodes, wherein each alternative path is placed at positions in the directed acyclic graph where there is a variation;

aligning a first sequence read corresponding to a genetic sample later collected from the subject to the directed acyclic graph;

determining, based on the aligned plurality of sequence reads, differences between the first sequence read and the directed acyclic graph, wherein differences are correlated with progression of the disease.

19. The method of claim 18, further comprising:
 - modifying the directed acyclic graph to incorporate the differences between the first sequence read and the original directed acyclic graph;
 - aligning a second sequence read corresponding to a second genetic sample later collected from the subject to the modified directed acyclic graph; and
 - determining differences between the second sequence read and the modified directed acyclic graph, wherein differences are correlated with further progression of the disease.
20. The method of claim 18 or 19, wherein the cancer is selected from breast, lung, skin, prostate, thyroid, pancreatic, bladder, or ovarian cancer.
21. The method of claim 18 or 19, wherein the cancer is leukemia or lymphoma.
22. The method of claim 18 or 19, wherein progression of the disease correlates with metastatic cancer.
23. The method of claim 18 or 19, wherein the disease is an autoimmune disease.
24. A method for determining a genetic marker for metastatic cancer, comprising:
 - representing variations between the genetic sequence of non-cancerous cells of a subject and the genetic sequence of non-metastatic cancerous cells of the subject as two or more alternative paths in a directed acyclic graph comprising nodes, wherein each alternative path is placed at positions in the directed acyclic graph where there is a variation;
 - aligning a plurality of sequence reads corresponding to a genetic sample from a metastatic cell of the subject to the directed acyclic graph; and

determining, based on the aligned plurality of sequence reads, differences between the sequence reads and the directed acyclic graph, thereby determining at least one genetic marker correlated with the metastatic clone.