

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 ~~invention-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 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 aspect, the present invention refers to 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 aspect, the present invention refers to 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 ~~invention-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 ~~invention-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 ~~invention-disclosure~~ can be used to identify and/or study variations between individuals or cohorts having similar disease

stages. In an ~~embodiment~~example, the ~~invention-disclosure~~ is implemented by aligning sequence reads to a series of directed, acyclic sequences spanning branch points that account for sequence variation in the samples, including insertions, deletions, and substitutions. Such constructs, often represented as directed acyclic graphs (DAGs) can be assembled from existing variable sequence files or the constructs can be fabricated *de novo* using a standard reference as a starting point.

Once a sequence construct has been fabricated to account for sequence variation among differently diseased samples, those constructs can be used to identify disease risk in new samples from the same individual or, in some instances, from other individuals. In particular, because portions of the sequence construct can be tagged with secondary information, such as “metastatic,” the subsequent step of comparing a mutation, vis-à-vis the reference genome, to a table of known mutations can be eliminated. Thus, it is merely a matter of identifying a sample as being aligned to sequences in the construct that are indicative of the disease or stage. Alternatively, when a mutation is not known (i.e., not represented in the reference sequence construct), an alignment will be found, whereby the variant can be identified as a new mutation. Thus, using this iterative process, it is possible to compare and/or identify differences between major and minor cancer clones, or between pre-cancerous and cancerous samples.

The ~~invention-disclosure~~ additionally includes systems for executing the methods of the ~~invention-disclosure~~. In one ~~embodiment~~example, a system comprises a distributed network of processors and storage capable of comparing a plurality of sequences (i.e., nucleic acid sequences, amino acid sequences) to a reference sequence construct (e.g., a DAG) representing observed variation in a genome or a region of a genome. The system is additionally capable of aligning the nucleic acid reads to produce a continuous sequence using an efficient alignment algorithm. Because the reference sequence construct compresses a great deal of redundant information, and because the alignment algorithm is so efficient, the reads can be tagged and assembled on an entire genome using commercially-available resources. The system comprises a plurality of processors that simultaneously execute a plurality of comparisons between a plurality of reads and the reference sequence construct. The comparison data may be accumulated and provided to a health care provider. Because the comparisons are computationally tractable, analyzing sequence reads will no longer represent a bottleneck between NGS sequencing and a meaningful discussion of a patient’s genetic risks.