

METHODS AND SYSTEMS FOR ANALYSIS OF SEQUENCING DATA

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application is a nonprovisional application claiming priority to U.S. Provisional Application Serial No. 61/118,395, filed November 26, 2008, the disclosure of which is incorporated herein by reference in its entirety.

FIELD OF THE INVENTION

[0002] The present technology relates to molecular sciences, such as genomics. More particularly, the present technology relates to methods and systems for the analysis of sequencing data.

BACKGROUND

[0003] The detection of specific nucleic acid sequences present in a biological sample can be used as a method for identifying and classifying microorganisms, diagnosing infectious diseases, detecting and characterizing genetic abnormalities, identifying genetic changes associated with cancer, studying genetic susceptibility to disease, and measuring response to various types of treatment. A common technique for detecting specific nucleic acid sequences in a biological sample is nucleic acid sequencing.

[0004] Nucleic acid sequencing methodology has evolved significantly from the chemical degradation methods used by Maxam and Gilbert and the strand elongation methods used by Sanger. Today several sequencing methodologies are in use which allow for the parallel processing of thousands of nucleic acids all in a single sequencing run. As such, the information generated from a single sequencing run can be enormous.

SUMMARY

[0005] Embodiments of the present invention technology relates to the analysis of sequencing data as it is being generated. In some embodiments of the present invention, such analysis permits the identification of the source of a target nucleic acid that is being sequenced prior to obtaining the complete sequence of the target nucleic acid or prior to the end of a sequencing run. In some embodiments, sequencing runs can be terminated prior to

completion. In some of these and other embodiments, sequencing is terminated based on analyzing the data (e.g. the quantity of data, the quality of data, the information provided by the data – such as the ability of the information contained in the data to answer a question being asked, etc.). In some embodiments, sequencing is terminated based on analysis of the data which results in a determination that sufficient data has been obtained (e.g. sufficient data to identify a species, sufficient data to complete sequencing, sufficient data to identify all markers of interest, etc.). Sufficient data may include the minimum amount of data necessary to perform a certain analysis (e.g. identify a species, make a diagnosis, obtain a full sequence, etc.), or may include obtaining data with sufficient redundancy to add a desired confidence when performing the analysis. Terminating sequencing may include immediately terminating sequencing, performing a specified (e.g. fixed or variable) amount of further sequencing, and/or initiating a termination procedure (such as flushing reagents, sending a notice, etc.). In addition to, or as an alternative to terminating sequencing upon meeting the specified criteria, other actions could also be taken. When the specified data is obtained (e.g. the amount needed to identify a species, to make a diagnosis, etc.) a notice can be sent to a user (e.g. an electronic message can be sent, an indicator may illuminate or vibrate, etc.), a different system may be activated (e.g. to run another test, to take corrective action based on the diagnosis or species identified, etc.), and/or some other action may be taken in response to the determination that sufficient and/or specified data has been obtained.

[0006] In some embodiments of the present invention, methods and systems are described for characterizing a target nucleic acid while determining a portion of the nucleotide sequence of the target nucleic acid. Certain embodiments include methods and systems for identifying the source of a target nucleic acid by comparing the accumulating nucleotide sequence of a portion of a target nucleic acid, or the accumulating sequences of portions of a plurality of target nucleic acids, to a population of reference nucleotide sequences.

[0007] According to a first aspect of the present invention, there is provided a method~~Some embodiments described herein include methods~~ for identifying the source of a target nucleic acid.~~Such methods can include, said method comprising~~ the steps of (a) initiating a sequencing process to determine the nucleotide sequence of a target nucleic acid or a fragment thereof, thereby generating a nucleotide sequence of at least a portion of the target nucleic acid; (b) prior to terminating the sequencing process, comparing the nucleotide sequence of the at least a portion of the target nucleic acid to a population of

identifying the organism from which the genome was derived after only a fraction of the whole genome has been sequenced.

[0012] EmbodimentsAspects of the presently described methods are particularly relevant to the identification of target nucleic acids obtained from metagenomic samples. As such, in preferred embodiments, the methods described herein relate to the identification of the source of a target nucleic acid that is obtained from one or more metagenomic samples. In particular embodiments, the methods can be used to identify the source sufficiently to distinguish and/or identify the species (e.g. to uniquely identify the species, to uniquely identify a sub-species, to identify a set of species and/or subspecies, etc.) of organism from other candidate species. In some embodiments, identification of a set of species is sufficient (e.g. where each species and/or sub-species within the set is corrected or accounted for in a common way – such as treating using the same medicine, eradicated using the same technique, etc.) for the “identification of a species.” In other embodiments, identification of a species involves uniquely identifying the species.

[0013] In some embodiments of the methods described herein, the reference nucleotide sequences within the population of reference nucleotide sequences are indexed in a database by association with a particular species of the specified organism. In other embodiments, the reference nucleotide sequences are further indexed in the database by association with a particular subspecies of the specified organism. In still other methods for identifying the source of a target nucleic acid, the reference nucleotide sequences within the population of reference nucleotide sequences are indexed in a database by association with one or more groups of organisms. In some embodiments, the reference nucleotide sequences within the population of reference nucleotide sequences can be indexed in a database by a hierarchical association with a plurality of groups of organisms. In yet other embodiments, the plurality of groups of organisms can be phylogenetically related. In some embodiments of the methods described herein, the target nucleic acid contains at least a portion of a nucleic acid encoding one or more genes for which phylogenetic relationships are known. Such genes can be useful for identifying organisms of interest or relationships between organisms. Exemplary genes for which phylogenetic relationships are well established include, but are not limited to, RuBisCo, NifH, sulfite reductase, a mitochondrial nucleic acid or 16S rRNA. In some embodiments, the mitochondrial nucleic acid comprises cytochrome c oxidase subunit I.

[0014] In some embodiments of the presently described methods, the sequencing process comprises array-based sequencing. In this and other embodiments, the sequencing process can comprise a process selected from the group consisting of sequencing by hybridization, sequencing by synthesis and sequencing by ligation. In additional embodiments, other methods of sequencing can be employed with the methods described herein.

[0015] Some of the methods described herein also include comparing the nucleotide sequence of at least a portion of a target nucleic acid to a population of reference nucleotide sequences using a heuristic algorithm. In such embodiments, the algorithm can comprise, for example, a BLAST algorithm or a FASTA algorithm.

[0016] In some of the methods described herein the threshold for determining whether a subpopulation of reference nucleotide sequences match the nucleotide sequence of the at least a portion of the target nucleic acid comprises a user specified threshold. In some embodiments, the threshold can be determined using one or more parameters. In some embodiments, the one or more parameters can comprise percent nucleotide sequence identity.

[0017] In some embodiments for identifying the source of a target nucleic acid, the subpopulation of reference nucleotide sequences permits sufficient identification of the source of the target nucleic acid if at least a specified percentage of the reference nucleotide sequences within the subpopulation are from the same genus of organism. In still other embodiments, the subpopulation of reference nucleotide sequences permits sufficient identification of the source of the target nucleic acid if at least a specified percentage of the reference nucleotide sequences within the subpopulation are from the same species of organism. In yet other embodiments, the subpopulation of reference nucleotide sequences permits sufficient identification of the source of the target nucleic acid if at least a specified percentage of the reference nucleotide sequences within the subpopulation are from the same subspecies of organism.

[0018] According to a second aspect, there is provided a system~~In addition to the methods described herein, systems for identifying the source of a target nucleic acid, said system comprising~~are described. Such systems can include: a computer containing a memory, the computer interfaced with a database containing a population of reference nucleotide sequences from specified organisms; a nucleic acid sequencer configured to perform a sequencing process to determine the nucleotide sequence of a target nucleic acid or a fragment thereof, thereby generating in said memory a nucleotide