(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
20 July 2017 (20.07.2017)

WIPO | PCT

(10) International Publication Number
WO 2017/123864 A1

(51) International Patent Classification:
C12Q 1/68 (2006.01) *G01N 33/574* (2006.01)
G01N 33/50 (2006.01) *G06F 19/22* (2011.01)

(21) International Application Number:
PCT/US2017/013329

(22) International Filing Date:
13 January 2017 (13.01.2017)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
14/994,385 13 January 2016 (13.01.2016) US

(71) Applicant: **SEVEN BRIDGES GENOMICS INC.** [US/US]; 1 Main St, 5th Floor, Suite 500, Cambridge, Massachusetts 02142 (US).

(72) Inventors: **LEE, Wan-Ping**; 1 Main St, 5th Floor, Suite 500, Cambridge, Massachusetts 02142 (US). **LOCKE, Devin**; 1 Main St, 5th Floor, Suite 500, Cambridge, Massachusetts 02142 (US).

(74) Agents: **MEYERS, Thomas C.** et al.; Brown Rudnick LLP, One Financial Center, Boston, Massachusetts 02111 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT,

[Continued on next page]

(54) Title: SYSTEMS AND METHODS FOR ANALYZING CIRCULATING TUMOR DNA

(57) **Abstract**: The invention provides oncogenomic methods for detecting tumors by identifying circulating tumor DNA. A patient-specific reference directed acyclic graph (DAG) represents known human genomic sequences and non-tumor DNA from the patient as well as known tumor-associated mutations. Sequence reads from cell-free plasma DNA from the patient are mapped to the patient-specific genomic reference graph. Any of the known tumor-associated mutations found in the reads and any *de novo* mutations found in the reads are reported as the patient's tumor mutation burden.

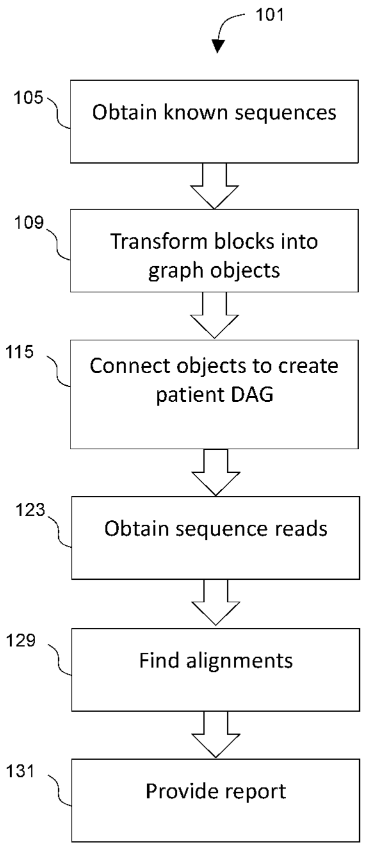


FIG. 1

HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

SYSTEMS AND METHODS FOR ANALYZING CIRCULATING TUMOR DNA

Related Application

The present application claims the benefit of and priority to U.S. Application Serial No. 14/994,385, filed January 13, 2016, the content of which is incorporated by reference herein in its entirety.

Technical Field

The invention relates to oncogenomics and to the detection, monitoring, and treatment of tumors.

Background

About 8 million people die of cancer each year according to the World Health Organization. Cancer is the uncontrolled growth of cells and can lead to malignant tumors that resist treatment, spread through the body, and potentially recur after removal. Early detection and vigilant monitoring are key to minimizing cancer's harm.

A lump in a patient may be biopsied to determine if it is a malignant tumor, but this only occurs after the lump has appeared. Some tumors, such as lung tumors, require difficult and painful biopsies in which a fine needle is inserted to the site of the tumor. A minimally-invasive procedure known as "liquid biopsy", which involves sequencing DNA from a patient's blood, promises to provide early detection.

Unfortunately, there are limits to the insights offered by a liquid biopsy. Sequence reads are generally analyzed by mapping them to a reference genome. Any difference between a sequence read and the reference is called as a mutation in the patient without regard to which should be considered the healthy genotype. The physician is left to search the literature to interpret the mutations found in the patient.

Summary

Circulating tumor DNA is identified using sequence reads from a patient's cell-free plasma DNA. The sequence reads are mapped to a genomic graph that represents a diversity of human genotypes including the patient’s own healthy, non-tumor genotype. The genomic graph is annotated to identify known tumor-associated mutations and thus provides a successful match, or “hit”, on any sequence reads that include those known tumor-associated mutations. The patient’s tumor-related mutation population is reported and includes those known tumor-associated mutations to which the sequence reads mapped. Where any of the sequence reads include *de novo* mutations relative to the patient’s non-tumor genotype, those *de novo* mutations are included in the reported patient’s tumor-related mutation population. Thus by mapping sequence reads from a patient’s cell-free plasma DNA to an annotated reference graph that includes the patient’s non-tumor genotype, systems and methods of the invention may be used to provide a report of the patient’s tumor-related mutation population, which includes both known tumor-associated mutations and *de novo* mutations found in ctDNA in the patient.

The methods described herein may be used with patients known or suspected to have cancer, as a method of identifying causal variants in order to better target treatments and, with tests over time, for determining whether treatments have worked and what new mutations may have been selected for in treatment. Because DNA from all of a patient’s tumors and tumor clones circulates in the bloodstream, ctDNA allows the clinician to obtain information on the population of tumor clones present in a patient, rather than just one particular clone or tumor. Additionally, systems and methods of the invention may be used to analyze sequence reads from a patient’s cell-free plasma DNA to detect ctDNA as a means of screening for cancer by determining whether ctDNA (and thus, presumably, tumors) are present and tracking them over time.

At the core of the detection of ctDNA from sequence reads from a patient’s cell-free plasma DNA is the use of the reference graph, such as a patient-specific reference graph or a patient-specific directed acyclic graph (DAG). The patient-specific genomic reference graph is used for mapping sequence reads. The patient-specific genomic reference graph contains a variety of genomic references including, for example, one or a plurality of human genomes as well as sequences from the patient’s non-tumor tissue. Use of a DAG supports very fast analysis of very large numbers of sequence reads. Without the use of the patient-specific genomic reference graph, to make the number of pairwise alignment comparisons required to compare each of the potentially millions of sequence reads from an NGS instrument run to each segment of every included reference would be intractable in many settings. The patient-specific genomic reference graph is created by aligning the reference sequences to identify portions that match when aligned. Redundancy is eliminated by storing those portions as single objects in memory. The complete sequence of any one of the reference sequences is represented by connecting those objects in their genomic order and those connections may be implemented using pointers to physical locations in computer memory of adjacent objects. This allows operations such as searches, lookups, or alignments to be performed without having to refer to an extrinsic index as would be required in a traditional sequence database. That is, the index-free adjacency of the patient-specific genomic reference graph is particularly well-suited for the rapid and efficient detection of ctDNA from NGS sequence reads.

In certain aspects, the invention provides a method for analyzing tumor DNA. The method includes obtaining a patient-specific genomic reference graph that represents known human genomic sequences as well as non-tumor DNA from the patient. The patient-specific genomic reference graph includes objects in a tangible memory subsystem of a computer system, wherein matching homologous portions of the sequences are each represented by a single object. Preferably, the patient-specific genomic reference graph includes one or more known tumor-associated mutations and one or more patient mutations. The method includes aligning sequence reads—obtained by sequencing a sample containing cell-free plasma DNA from the patient—to the patient-specific genomic reference graph and providing a report that includes at least one mutation in the sequence reads relative to the patient-specific genomic reference graph revealed in the aligning step. The sequence reads may be obtained by sequencing the sample containing the cell-free plasma DNA from the patient. The sequence reads may be aligned to the patient-specific genomic reference graph by using a processor of the computer system to perform a multi-dimensional look-back operation to find a highest-scoring trace through a multi-dimensional matrix.

The report can identify a patient's tumor-related mutation population, e.g., any of the known tumor-associated mutations as well as any *de novo* mutations found in the sequence reads. The report may also describe what proportion of the sequence reads align to mutation-associated branches in the patient-specific genomic reference graph. Additionally or alternatively, the report may identify distinct clones present in a tumor in the patient. Similarly, the report may describe a novel driver mutation present in the cell-free plasma DNA from the patient. Methods of the invention may include adding the novel driver mutation to the one or more known tumor-associated mutations in the patient-specific genomic reference graph for subsequent uses.

In some embodiments, methods include building the patient-specific genomic reference graph by obtaining the known human genomic sequences, obtaining non-tumor sequences for the non-tumor DNA from the patient, and finding and deleting redundancies among matching, homologous portions of the known human genomic sequences and the non-tumor sequences, leaving the matching homologus portions. An object is created in the tangible memory subsystem for each of the matching homologous portions and connections are created between pairs of the objects to create the patient-specific genomic reference graph.

Methods of the invention may include aligning a second set of sequence reads to the patient-specific genomic reference graph and producing a second report that identifies a patient's second tumor-related mutation population at a time different from an initial time associated with the patient's tumor-related mutation population. The second report may compare the patient's second tumor-related mutation population to the patient's initial tumor-related mutation population.

In some embodiments, the objects of the patient-specific genomic reference graph include pointers to adjacent ones of the objects such that the objects are linked into paths to represent the plurality of known human genomic sequences, wherein each pointer identifies a physical location in the memory subsystem at which the adjacent object is stored. In certain embodiments, objects of the reference DAG comprise vertex objects connected by edge objects and an adjacency list for each vertex object and edge object, wherein the adjacency list for a vertex object or edge object lists the edge objects or vertex objects to which that vertex object or edge object is adjacent. In certain embodiments, the objects each include an adjacency list, wherein each entry in an adjacency list is a pointer to an adjacent object, wherein each pointer identifies a physical location in the memory subsystem at which the adjacent object is stored. In some embodiments, the patient-specific genomic reference graph uses index-free adjacency to link the objects into paths to represent the plurality of known human genomic sequences.

In related aspects, the invention provides a system for analyzing tumor DNA. The system includes one or more processors coupled to a memory subsystem. Stored in the memory subsystem is a patient-specific genomic reference graph representing known human genomic sequences as well as non-tumor DNA from a patient, wherein matching homologous portions of the sequences are each represented by a single one of a plurality of objects stored in the memory subsystem. The system also includes instructions executable by the processor to cause the system to align sequence reads from cell-free plasma DNA of the patient to the patient-specific genomic reference graph and provide a report that includes at least one mutation in the sequence reads relative to the patient-specific genomic reference graph revealed by the aligning step. Preferably, the patient-specific genomic reference graph includes one or more known tumor-associated mutations and one or more patient mutations. The report identifies a patient’s tumor-related mutation population, e.g., any *de novo* mutations and any of the known tumor-associated mutations found in the sequence reads. Optionally, the report further may identify any of: what portion of the sequence reads align to mutation-associated branches in the patient-specific genomic reference graph; a first clone and a second clone present in a tumor in the patient; and a novel driver mutation present in the cell-free plasma DNA from the patient. In some embodiments, the system is operable to build the patient-specific genomic reference graph by obtaining the known human genomic sequences; obtaining non-tumor sequences for the non-tumor DNA from the patient; finding and deleting redundancies among homologous portions of the known human genomic sequences and the non-tumor sequences, leaving the matching homologous portions; creating one of the objects in the tangible memory subsystem for each of the matching homologous portions; and creating connections between pairs of the objects to create the patient-specific genomic reference graph. In some embodiments, the objects of the patient-specific genomic reference graph include pointers to adjacent ones of the objects such that the objects are linked into paths to represent the plurality of known human genomic sequences, wherein each pointer identifies a physical location in the memory subsystem at which the adjacent object is stored. In certain embodiments, objects of the reference DAG comprise vertex objects connected by edge objects and an adjacency list for each vertex object and edge object, wherein the adjacency list for a vertex object or edge object lists the edge objects or vertex objects to which that vertex object or edge object is adjacent. In certain embodiments, the objects each include an adjacency list, wherein each entry in an adjacency list is a pointer to an adjacent object, wherein each pointer identifies a physical location in the memory subsystem at which the adjacent object is stored. In some embodiments, the patient-specific genomic reference graph uses index-free adjacency to link the objects into paths to represent the plurality of known human genomic sequences.

Brief Description of the Drawings

FIG. 1 diagrams a method for analyzing tumor DNA.

FIG. 2 shows the creation of a patient-specific genomic reference graph.

FIG. 3 shows known tumor mutations and patient mutations on the patient-specific genomic reference graph.

FIG. 4 diagrams a system of the invention.

FIG. 5 shows the use of adjacency lists.

FIG. 6 shows adjacency lists where vertices and edges are stored as objects.

FIG. 7 illustrates obtaining sequence reads from a sample.

FIG. 8 illustrates aligning sequence reads to a patient-specific genomic reference graph.

FIG. 9 shows the matrices that represent the comparison.

FIG. 10 illustrates a report of tumor-related mutations for a patient.

FIG. 11 shows a report of what portion of the sequence reads align to DAG branches.

FIG. 12 shows a report describing tumor clones in a patient.

Detailed Description

Systems and methods of the invention avoid the hassle and expense of biopsying tumors by instead operating with sequences from cell-free fragments of tumor DNA circulating in the bloodstream, known as circulating tumor DNA, or ctDNA. Genomic studies show that virtually all cancer tumors are associated with somatic genetic mutations, some of which rarely occur in healthy cells and thus provide specific tumor biomarkers. Tumor biomarkers can be assessed by looking at ctDNA. Liquid biopsies can provide temporal measurements of the total tumor burden as well as identify specific mutations that arise during therapy. Cancer may be detected or monitored by analyzing sequence reads obtained, for example, by sequencing a patient sample that includes the ctDNA. The sequence reads are analyzed by mapping them to a genomic reference that includes normal, non-tumor genetic sequence from the patient. Moreover, the reference is a patient-specific reference graph, such as a patient directed acyclic graph (DAG) that also includes any other known human reference sequences of interest that are available for inclusion. For example, the patient-specific genomic reference graph may include one or more published human genomes (e.g., hg18) and sequences obtained by sequencing the patient's healthy, non-tumor tissue.

The patient-specific genomic reference graph efficiently stores the reference sequences because matching, homologous portions of the reference sequences are represented using only a single object to eliminate redundancy. The patient-specific genomic reference graph supports very fast analysis commensurate with the pace at which data is generated by next-generation sequencing (NGS) technologies, because the objects in the DAG are stored and accessed using spatial memory addressing. Use of the patient-specific genomic reference graph provides results that are more accurate than other methods because each of the sequence reads is analyzed by simultaneously comparing it to the full range of known genetic variation, which avoids misleading inferences that arise when mapping a sequence read to a linear reference. Due to those benefits, mapping the ctDNA sequence reads to the patient-specific genomic reference graph allows tumor mutations to be identified rapidly and accurately.

In addition to allowing for non-invasive sampling, because DNA from all of a patient's tumors and tumor clones is circulating in the bloodstream, ctDNA allows the clinician to obtain information on the population of tumor clones present in a patient, rather than just one particular clone or tumor. Besides describing tumor clones in a patient, methods of the invention may be used in the sequencing of cell-free plasma DNA to detect ctDNA as a means of screening for cancer. See Newman et al., 2014, An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage, Nat Med. 20(5):548-554, incorporated by reference.

Systems and methods of the invention address the challenge of separating ctDNA from other cell-free plasma DNA (DNA circulating in the blood plasma and not contained within cells). With prior art approaches, ctDNA is identified as such when known cancer-associated mutations are identified while ctDNA not containing any known cancer-associated mutations may be missed.

Systems and methods of the invention may be used with patients known to have cancer, for identifying causal variants in order to better target treatments and, with tests over time, for determining whether treatments have worked and what new mutations may have been selected for in treatment. Systems and methods of the invention may also be used with patients who are simply being screened for cancer, e.g., for determining whether ctDNA (and thus, presumably, tumors) are present and tracking them over time.

FIG. 1 diagrams a method 101 for analyzing tumor DNA. The method includes obtaining 105 known human genomic sequences as well as sequences from non-tumor DNA from the patient. Portions of the sequences that match each other when aligned are inferred to be homologous and identified as blocks that are transformed 109 into objects that are stored in a tangible memory device. The objects are connected 115 by edges to create a patient-specific genomic reference graph such that there is a path through the patient-specific genomic reference graph for each of the original known human genomic sequences as well as a path or paths for the non-tumor DNA from the patient.

Method 101 further includes obtaining 123 sequence reads from a sample containing cell-free plasma DNA from the patient and finding 129 alignments between the sequence reads and the patient-specific genomic reference graph. A report is provided 131 that includes at least one mutation in the sequence reads relative to the patient's non-tumor DNA revealed in the aligning step. The report may identify any mutation found in the sequence reads (i.e., novel to that patient) and any of the known tumor-associated mutations that are found in the sequence reads as the patient's tumor-related mutation population.

FIG. 2 shows the creation of a patient-specific genomic reference graphpatient reference graph 331 from known human genomic sequences as well as non-tumor sequences from the patient. The known sequences 303—which include the known human genomic sequences as well as non-tumor DNA from the patient—are transformed into a graph 331 that includes vertex objects 305 and edge objects 309. Each of the sequences 303 are aligned to the others and in some embodiments, a multiple sequence alignment is performed. Portions of the sequences that match each other when aligned are identified as blocks and those blocks are transformed 109 into objects 205 that are stored in a tangible memory device. This results in finding and deleting redundancies among homologous portions of the known human genomic sequences and the non-tumor sequences, leaving the matching homologous portions. Transforming those blocks into objects is done by creating one of the objects in the tangible memory subsystem for each of the blocks. The computer is used to create connections between pairs of the objects to create the patient-specific genomic reference graph.

In the fragments of sequence represented in FIG. 2, it can be seen that bases 2-5 of first sequence align to, and match, bases 2-5 of the second sequence. Thus those segments of those two sequences are identified as a block and systems of the invention create an object 305 to represent that AGTT string. It is noted that this object could potentially be stored using one byte of information. For example, if A = 00, C = 01, G = 10, and T = 11, then this block contains 00101111 (one byte). Where the original sequences 303 contain a very large number of reference sequences, the described methods provide a consider improvement to the operation of the computer system in comparison to a prior art method that stores an entire multiple sequence alignment.

The objects 305 are connected 115 to create paths such that there is a path for each of the original known sequences 303. The paths are directed and preferably in the sense that the direction of each path corresponds to the 5' to 3' directionality of the sequences. Further, the graph is acyclic in that no cycles are allowed due to the directionality of the paths. However, in certain embodiments, the paths may lack an associated directionality component, and some paths may be acyclic. The connections creating the paths can themselves be implemented as objects so that the blocks are represented by vertex objects 305 and the connections are represented by edge objects 309. Thus the directed graph comprises vertex and edge objects stored in the tangible memory device. The directed graph 331 represents the plurality of known sequences 303 in that each one of the original sequences can be retrieved by reading a path in the direction of that path. However, the directed graph 331 is a different article that the original sequences 303, at least in that portions of the sequences that match each other when aligned have been transformed into single objects 303. Thus if the original article includes hundreds or more complete human genomes, then billions of characters of information from the original article are transformed into a graph that uses much less disk space than if stored as, e.g., a multiple sequence alignment. It should be noted that in the figures, the patient-specific genomic reference graph is represented using only a few nodes and edges and a few nucleotide characters whereas in actual implementation, the DAG will be much larger than what is depicted here and too large to be represented in such figures. The patient-specific genomic reference graphpatient reference graph 331 is preferably annotated to include one or more known tumor-associated mutations and one or more patient mutations.

One approach to building the patient-specific genomic reference graph is to first build a human reference DAG incorporating known human reference sequences and variants. Then, sequence reads from a sample of the patient's normal (non-tumor) DNA are aligned to that human reference DAG. The patient-specific ctDNA-detection reference DAG, or patient-specific genomic reference graph, is then built by incorporating any *de novo* mutations found using that first alignment step along with known tumor-associated mutations into the reference DAG.

FIG. 3 shows known tumor mutations and patient mutations on the patient-specific genomic reference graph. Sequence reads from a patient sample containing ctDNA are aligned to the patient-specific genomic reference graph to report all known tumor-related mutations (if any) that those sequence reads successfully align to, along with any new de novo mutations identified, together with a proportion of reads aligned to each mutation branch preferably normalized for total number of reads aligning at that position. This may be taken as the patient’s tumor-related mutation population at t0, e.g., as an initial assessment of clonality.

It may be desirable to, at a later point in time (e.g., in the treatment scenario after treatment has had time to take effect; in the screening scenario after a reasonable time has passed such that screening again makes sense), align a new sets of reads from a newly-sequenced sample of the patient’s cell-free plasma DNA to the patient-specific genomic reference graph. Then systems and methods of the invention may be used to report all tumor-related mutations that reads are aligned to together with proportion aligned, e.g., as the patient’s tumor-related mutation population at t1, along with a comparison of this population of mutations to the previous round of testing.

Preferably, the alignment and reporting steps are performed by one or more processors of a computer system wherein the patient-specific genomic reference graph is stored within a non-transitory, tangible memory subsystem.

FIG. 4 diagrams a system of the invention. The system 401 may be used to perform methods of the invention. The system 401 includes at least one computer 433. Optionally, the system 401 may further include one or more of a server computer 409 and a sequencer 455, which may be coupled to a sequencer computer 451. Each computer in the system 401 includes a processor coupled to a memory device and at least one input/output device. Thus the system 401 includes at least one processor coupled to a memory subsystem (e.g., a memory device or collection of memory devices 475). Using those mechanical components, the system 401 is operable to obtain a sequence generated by sequencing nucleic acid from a genome of a patient. The system uses the processor to transform the sequence 303 into the graph 331.

Processor refers to any device or system of devices that performs processing operations. A processor will generally include a chip, such as a single core or multi-core chip, to provide a central processing unit (CPU). A processor may be provided by a chip from Intel or AMD. A processor may be any suitable processor such as the microprocessor sold under the trademark XEON E7 by Intel (Santa Clara, CA) or the microprocessor sold under the trademark OPTERON 6200 by AMD (Sunnyvale, CA).

The memory subsystem 475 contains one or any combination of memory devices. A memory device is a mechanical device that stores data or instructions in a machine-readable format. Memory may include one or more sets of instructions (e.g., software) which, when executed by one or more of the processors of the disclosed computers can accomplish some or all of the methods or functions described herein. Preferably, each computer includes a non-transitory memory device such as a solid state drive, flash drive, disk drive, hard drive, subscriber identity module (SIM) card, secure digital card (SD card), micro SD card, or solid-state drive (SSD), optical and magnetic media, others, or a combination thereof.

Using the described components, the system 401 is operable to produce a report and provide the report to a user via an input/output device. An input/output device is a mechanism or system for transferring data into or out of a computer. Exemplary input/output devices include a video display unit (e.g., a liquid crystal display (LCD) or a cathode ray tube (CRT)), a printer, an alphanumeric input device (e.g., a keyboard), a cursor control device (e.g., a mouse), a disk drive unit, a speaker, a touchscreen, an accelerometer, a microphone, a cellular radio frequency antenna, and a network interface device, which can be, for example, a network interface card (NIC), Wi-Fi card, or cellular modem.

Preferably the graph is stored in the memory subsystem using adjacency lists, which may include pointers to identify a physical location in the memory subsystem 475 where each vertex is stored. In a preferred embodiment, the graph is stored in the memory subsystem 475 using adjacency lists. In some embodiments, there is an adjacency list for each vertex. For discussion of implementations see 'Chapter 4, Graphs' at pages 515-693 of Sedgewick and Wayne, 2011, Algorithms, 4th Ed., Pearson Education, Inc., Upper Saddle River NJ, 955 pages, the contents of which are incorporated by reference and within which pages 524-527 illustrate adjacency lists.

FIG. 5 shows the use of adjacency lists. An adjacency list 501 is used for each vertex 305. The system 401 uses a processor to create a graph 331 that includes vertex objects 305 and edge objects 309 through the use of adjacency, i.e., adjacency lists or index free adjacency. Thus, the processor may create the graph 331 using index-free adjacency wherein a vertex 305 includes a pointer to another vertex 305 to which it is connected and the pointer identifies a physical location in on a memory device 475 where the connected vertex is stored. The graph 331 may be implemented using adjacency lists such that each vertex or edge stores a list of such objects that it is adjacent to. Each adjacency list comprises pointers to specific physical locations within a memory device for the adjacent objects.

In the top part of FIG. 5, the graph 331 is illustrated in a cartoon-like visual-friendly format. The graph 331 will typically be stored on a physical device of memory subsystem 475 in a fashion that provide for very rapid traversals. In that sense, the bottom portion of FIG. 5 is not cartoon-like and represents that objects are stored at specific physical locations on a tangible part of the memory subsystem 475. Each node 305 is stored at a physical location, the location of which is referenced by a pointer in any adjacency list 501 that references that node. Each node 305 has an adjacency list 501 that includes every adjacent node in the graph 331. The entries in the list 501 are pointers to the adjacent nodes.

In certain embodiments, there is an adjacency list for each vertex and edge and the adjacency list for a vertex or edge lists the edges or vertices to which that vertex or edge is adjacent.

FIG. 6 shows adjacency lists where vertices and edges are stored as objects. An adjacency list 601 is used for each vertex 305 and edge 309. As shown in FIG. 6, system 401 creates the graph 331 using an adjacency list 601 for each vertex and edge, wherein the adjacency list 601 for a vertex 305 or edge 309 lists the edges or vertices to which that vertex or edge is adjacent. Each entry in adjacency list 601 is a pointer to the adjacent vertex or edge.

Preferably, each pointer identifies a physical location in the memory subsystem at which the adjacent object is stored. In the preferred embodiments, the pointer or native pointer is manipulatable as a memory address in that it points to a physical location on the memory but also dereferencing the pointer accesses intended data. That is, a pointer is a reference to a datum stored somewhere in memory; to obtain that datum is to dereference the pointer. The feature that separates pointers from other kinds of reference is that a pointer's value is interpreted as a memory address, at a low-level or hardware level. The speed and efficiency of the described graph genome engine allows a sequence to be queried against a large-scale genomic reference graph 331 representing millions or billions of bases, using a computer system 401. Such a graph representation provides means for fast random access, modification, and data retrieval.

In some embodiments, fast random access is supported and graph object storage are implemented with index-free adjacency in that every element contains a direct pointer to its adjacent elements (e.g., as described in U.S. Pub. 2014/0280360 and U.S. Pub. 2014/0278590, incorporated by reference), which obviates the need for index look-ups, allowing traversals (e.g., as done in the modified SW alignment algorithm described herein) to be very rapid. Index-free adjacency is another example of low-level, or hardware-level, memory referencing for data retrieval (as required in alignment and as particularly pays off in terms of speed gains in the modified, multi-dimensional Smith-Waterman alignment described below). Specifically, index-free adjacency can be implemented such that the pointers contained within elements are in-fact references to a physical location in memory.

Since a technological implementation that uses physical memory addressing such as native pointers can access and use data in such a lightweight fashion without the requirement of separate index tables or other intervening lookup steps, the capabilities of a given computer, e.g., any modern consumer-grade desktop computer, are extended to allow for full operation of a genomic-scale graph (i.e., a graph 331 that represents all loci in a substantial portion of the subject's genome). Thus storing graph elements (e.g., nodes and edges) using a library of objects with native pointers or other implementation that provides index-free adjacency—i.e., embodiments in which data is retrieved by dereferencing a pointer to a physical location in memory—actually improves the ability of the technology to provide storage, retrieval, and alignment for genomic information since it uses the physical memory of a computer in a particular way.

While no specific format is required for storage of a patient-specific genomic reference graph, FIGS. 5 and 6 are presented to illustrate useful formats. With reference back to FIG. 1, it is noted that methods of the invention use the patient-specific genomic reference graph with tumor sequence reads that are obtained from the patient. In some embodiments, sequence reads are obtained as an electronic article, e.g., uploaded, emailed, or FTP transferred from a lab to system 401. In certain embodiments, sequence reads are obtained by sequencing.

FIG. 7 illustrates obtaining sequence reads 705 from a sample 703. The illustrated steps may be applicable to either or both of obtaining sequences of the non-tumor DNA from the patient or obtaining sequence reads from a cell-free plasma DNA from a sample from the patient. For non-tumor DNA reads, a tissue sample may be obtained. For cell-free plasma DNA reads, a blood sample may be obtained from the patient and cell-free plasma DNA (cfDNA) may be isolated. Any suitable method may be used to isolate cfDNA and it may be preferable to use a commercially-available kit such as the circulating nucleic acid kit sold under the trademark QIAAMP by Qiagen (Venlo, Netherlands) or the plasma/serum cell-free circulating DNA purification mini kit sold by Norgen Biotek Corp. (Ontario, Canada). After isolation, sample 703 includes cell-free plasma DNA in a form amenable to sequencing, e.g., by next-generation sequencing (NGS) instruments.

In certain embodiments, sequence reads are obtained by performing sequencing 713 on a sample 703 from a subject. Sequencing may be by any method known in the art. See, generally, Quail, et al., 2012, A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers, BMC Genomics 13:341. DNA sequencing techniques include classic dideoxy sequencing reactions (Sanger method) using labeled terminators or primers and gel separation in slab or capillary, sequencing by synthesis using reversibly terminated labeled nucleotides, pyrosequencing, 454 sequencing, Illumina/Solexa sequencing, allele specific hybridization to a library of labeled oligonucleotide probes, sequencing by synthesis using allele specific hybridization to a library of labeled clones that is followed by ligation, real time monitoring of the incorporation of labeled nucleotides during a polymerization step, polony sequencing, and SOLiD sequencing.

A sequencing technique that can be used includes, for example, use of sequencing-by-synthesis systems sold under the trademarks GS JUNIOR, GS FLX+ and 454 SEQUENCING by 454 Life Sciences, a Roche company (Branford, CT), and described by Margulies, M. et al., Genome sequencing in micro-fabricated high-density picotiter reactors, Nature, 437:376-380 (2005); U.S. Pat. 5,583,024; U.S. Pat. 5,674,713; and U.S. Pat. 5,700,673, the contents of which are incorporated by reference herein in their entirety. 454 sequencing involves two steps. In the first step of those systems, DNA is sheared into blunt-end fragments attached to DNA capture beads and then amplified in droplets. In the second step, pyrosequencing is performed on each DNA fragment in parallel. Addition of one or more nucleotides generates a light signal that is recorded by a CCD camera in a sequencing instrument.

Another example of a DNA sequencing technique that can be used is SOLiD technology by Applied Biosystems from Life Technologies Corporation (Carlsbad, CA). In SOLiD sequencing, genomic DNA is sheared into fragments, and adaptors are attached to generate a fragment library. Clonal bead populations are prepared in microreactors containing beads, primers, template, and PCR components. Following PCR, the templates are denatured and enriched and the sequence is determined by a process that includes sequential hybridization and ligation of fluorescently labeled oligonucleotides.

Another example of a DNA sequencing technique that can be used is ion semiconductor sequencing using, for example, a system sold under the trademark ION TORRENT by Ion Torrent by Life Technologies (South San Francisco, CA). Ion semiconductor sequencing is described, for example, in Rothberg, et al., An integrated semiconductor device enabling non-optical genome sequencing, Nature 475:348-352 (2011); U.S. Pubs. 2009/0026082, 2009/0127589, 2010/0035252, 2010/0137143, 2010/0188073, 2010/0197507, 2010/0282617, 2010/0300559, 2010/0300895, 2010/0301398, and 2010/0304982, each incorporated by reference. DNA is fragmented and given amplification and sequencing adapter oligos. The fragments can be attached to a surface. Addition of one or more nucleotides releases a proton (H+), which signal is detected and recorded in a sequencing instrument.

Another example of a sequencing technology that can be used is Illumina sequencing. Illumina sequencing is based on the amplification of DNA on a solid surface using fold-back PCR and anchored primers. Genomic DNA is fragmented and attached to the surface of flow cell channels. Four fluorophore-labeled, reversibly terminating nucleotides are used to perform sequential sequencing. After nucleotide incorporation, a laser is used to excite the fluorophores, and an image is captured and the identity of the first base is recorded. Sequencing according to this technology is described in U.S. Pub. 2011/0009278, U.S. Pub. 2007/0114362, U.S. Pub. 2006/0024681, U.S. Pub. 2006/0292611, U.S. Pat. 7,960,120, U.S. Pat. 7,835,871, U.S. Pat. 7,232,656, U.S. Pat. 7,598,035, U.S. Pat. 6,306,597, U.S. Pat. 6,210,891, U.S. Pat. 6,828,100, U.S. Pat. 6,833,246, and U.S. Pat. 6,911,345, each incorporated by reference.

Other examples of a sequencing technology that can be used include the single molecule, real-time (SMRT) technology of Pacific Biosciences (Menlo Park, CA) and nanopore sequencing as described in Soni and Meller, 2007 Clin Chem 53:1996-2001.

As shown in FIG. 7, sequencing 713 generates a plurality of reads 705. Reads according to the invention generally include sequences of nucleotide data anywhere from tens to thousands of bases in length. Reads may be stored in any suitable format such as, for example, FASTA or FASTQ format. FASTA is originally a computer program for searching sequence databases and the name FASTA has come to also refer to a standard file format. See Pearson & Lipman, 1988, Improved tools for biological sequence comparison, PNAS 85:2444-2448. A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than (">") symbol in the first column. FASTQ files are similar to FASTA but further include a line of quality scores. Typically, sequence reads will be obtained 105 in a format such as FASTA, FASTQ, or similar.

The sequence reads from the cell-free plasma DNA are aligned to the patient-specific genomic reference graph. Since the patient-specific genomic reference graph includes non-tumor sequences from the patient, any difference between the sequence reads from the cell-free plasma DNA and the patient-specific genomic reference graph, or *de novo* mutation, is presumptively a tumor-associated mutation.

FIG. 8 illustrates aligning sequence reads to a patient-specific genomic reference graph. The computer system may be used to find 129 alignments 801 between a sequence read 709 and the graph 331. Using alignment operations of the invention, reads can be rapidly mapped to a graph despite their large numbers or short lengths. Numerous benefits obtain by using a graph as a reference. For example, aligning against a graph is more accurate than aligning against a linear reference and then attempting to adjust one's results in light of other extrinsic information. This is primarily because the latter approach enforces an unnatural asymmetry between the sequence used in the initial alignment and other information. Aligning against an object that potentially represents all the relevant physical possibilities is much more computationally efficient than attempting to align against a linear sequence for each physical possibility (the number of such possibilities will generally be exponential in the number of junctions). A modified Smith-Waterman operation for comparing a sequence to a reference graph is provided here as an extension of pairwise alignment methods.

Pairwise alignment generally involves placing one sequence along part of target, introducing gaps according to an algorithm, scoring how well the two sequences match, and preferably repeating for various positions along the reference. The best-scoring match is deemed to be the alignment and represents an inference of homology between alignment portions of the sequences. In some embodiments, scoring an alignment of a pair of nucleic acid sequences involves setting values for the scores of substitutions and indels. When individual bases are aligned, a match or mismatch contributes to the alignment score by a substitution probability, which could be, for example, 1 for a match and -0.33 for a mismatch. An indel deducts from an alignment score by a gap penalty, which could be, for example, -1. Gap penalties and substitution probabilities can be based on empirical knowledge or a priori assumptions about how sequences evolve. Their values affect the resulting alignment. Particularly, the relationship between the gap penalties and substitution probabilities influences whether substitutions or indels will be favored in the resulting alignment.

Stated formally, an alignment represents an inferred relationship between two sequences, x and y. For example, in some embodiments, an alignment A of sequences x and y maps x and y respectively to another two strings x' and y' that may contain spaces such that: (i) $|x'|=|y'|$; (ii) removing spaces from x' and y' should get back x and y, respectively; and (iii) for any i, x'[i] and y'[i] cannot be both spaces.

A gap is a maximal substring of contiguous spaces in either x' or y'. An alignment A can include the following three kinds of regions: (i) matched pair (e.g., x'[i]=y'[i]; (ii) mismatched pair, (e.g., $x'[i]\neq y'[i]$ and both are not spaces); or (iii) gap (e.g., either x'[i..j] or y'[i..j] is a gap). In certain embodiments, only a matched pair has a high positive score a. In some embodiments, a mismatched pair generally has a negative score b and a gap of length r also has a negative score g+rs where g, s<0. For DNA, one common scoring scheme (e.g. used by BLAST) makes score a=1, score b=-3, g=-5 and s=-2. The score of the alignment A is the sum of the scores for all matched pairs, mismatched pairs and gaps. The alignment score of x and y can be defined as the maximum score among all possible alignments of x and y.

Any pair may have a score a defined by a 4×4 matrix B of substitution probabilities. For example, B(i,i)=1 and $0 < B(i,j)[\text{for } i\neq j] < 1$ is one possible scoring system. For instance, where a transition is thought to be more biologically probable than a transversion, matrix B could include B(C,T)=.7 and B(A,T)=.3, or other values desired or determined by methods known in the art.

A pairwise alignment, generally, involves—for sequence Q (query) having m characters and a reference genome T (target) of n characters—finding and evaluating possible local alignments between Q and T. For any $1\leq i\leq n$ and $1\leq j\leq m$, the largest possible alignment score of T[h..i] and Q[k..j], where $h\leq i$ and $k\leq j$, is computed (i.e. the best alignment score of any substring of T ending at position i and any substring of Q ending at position j). This can include examining all substrings with cm characters, where c is a constant depending on a similarity model, and aligning each substring separately with Q. Each alignment is scored, and the alignment with the preferred score is accepted as the alignment. One of skill in the art will appreciate that there are exact and approximate algorithms for sequence alignment. Exact algorithms will find the highest scoring alignment, but can be computationally expensive. Two well-known exact algorithms are Needleman-Wunsch (J Mol Biol, 48(3):443-453, 1970) and Smith-Waterman (J Mol Biol, 147(1):195-197, 1981; Adv. in Math. 20(3), 367-387, 1976). A further improvement to Smith-Waterman by Gotoh (J Mol Biol, 162(3), 705-708, 1982) reduces the calculation time from O(m^2n) to O(mn) where m and n are the sequence sizes being compared and is more amendable to parallel processing. In the field of bioinformatics, it is Gotoh's modified algorithm that is often referred to as the Smith-Waterman algorithm. Smith-Waterman approaches are being used to align larger sequence sets against larger reference sequences as parallel computing resources become more widely and cheaply available. See, e.g., Amazon's cloud computing resources. All of the journal articles referenced herein are incorporated by reference in their entireties.

The original Smith-Waterman (SW) algorithm aligns linear sequences by rewarding overlap between bases in the sequences, and penalizing gaps between the sequences. Smith-Waterman also differs from Needleman-Wunsch, in that SW does not require the shorter sequence to span the string of letters describing the longer sequence. That is, SW does not assume that one sequence is a read of the entirety of the other sequence. Furthermore, because SW is not obligated to find an alignment that stretches across the entire length of the strings, a local alignment can begin and end anywhere within the two sequences.

The original SW algorithm is expressed for an n×m matrix H, representing the two strings of length n and m, in terms of equation (1):

$$H_{k0}=H_{0l}=0 \ (\text{for } 0\le k\le n \text{ and } 0\le l\le m) \qquad (1)$$

$$H_{ij}=\max\{H_{(i-1,j-1)}+s(a_i,b_j),H_{(i-1,j)}-W_{in},H_{(i,j-1)}-W_{del},0\}$$

$$(\text{for } 1\le i\le n \text{ and } 1\le j\le m)$$

In the equations above, s(ai,bj) represents either a match bonus (when ai = bj) or a mismatch penalty (when ai ≠ bj), and insertions and deletions are given the penalties Win and Wdel, respectively. In most instances, the resulting matrix has many elements that are zero. This representation makes it easier to backtrace from high-to-low, right-to-left in the matrix, thus identifying the alignment.

Once the matrix has been fully populated with scores, the SW algorithm performs a backtrack to determine the alignment. Starting with the maximum value in the matrix, the algorithm will backtrack based on which of the three values ($H_{i-1,j-1}$, $H_{i-1,j}$, or $H_{i,j-1}$) was used to compute the final maximum value for each cell. The backtracking stops when a zero is reached. The optimal-scoring alignment may contain greater than the minimum possible number of insertions and deletions, while containing far fewer than the maximum possible number of substitutions.

SW or SW-Gotoh may be implemented using dynamic programming to perform local sequence alignment of the two strings, S and A, of sizes m and n, respectively. This dynamic programming employs tables or matrices to preserve match scores and avoid re-computation for successive cells. Each element of the string can be indexed with respect to a letter of the sequence, that is, if S is the string ATCGAA, S[1] = A.

Instead of representing the optimum alignment as $H_{i,j}$ (above), the optimum alignment can be represented as B[j,k] in equation (2) below:

$$B[j, k] = \max(p[j, k], i[j, k], d[j, k], 0) \qquad (\text{for } 0 < j \leq m, 0 < k \leq n) \qquad (2)$$

The arguments of the maximum function, B[j,k], are outlined in equations (3)-(5) below, wherein MISMATCH_PEN, MATCH_BONUS, INSERTION_PEN, DELETION_PEN, and OPENING_PEN are all constants, and all negative except for MATCH_BONUS (PEN is short for PENALTY). The match argument, p[j,k], is given by equation (3), below:

$$p[j,k] = \max(p[j-1,k-1], i[j-1,k-1], d[j-1,k-1]) + \text{MISMATCH\_PEN}, \text{ if } S[j] \neq A[k] \quad (3)$$
$$= \max(p[j-1,k-1], i[j-1,k-1], d[j-1,k-1]) + \text{MATCH\_BONUS}, \text{ if } S[j] = A[k]$$

the insertion argument i[j,k], is given by equation (4), below:

$$i[j,k] = \max(p[j-1,k] + \text{OPENING\_PEN}, i[j-1,k], d[j-1,k] + \quad (4)$$
$$\text{OPENING\_PEN}) + \text{INSERTION\_PEN}$$

and the deletion argument d[j,k], is given by equation (5), below:

$$d[j,k] = \max(p[j,k-1] + \text{OPENING\_PEN}, i[j,k-1] + \qquad (5)$$
$$\text{OPENING\_PEN}, d[j,k-1]) + \text{DELETION\_PEN}$$

For all three arguments, the [0,0] element is set to zero to assure that the backtrack goes to completion, i.e., p[0,0] = i[0,0] = d[0,0] = 0.

The scoring parameters are somewhat arbitrary, and can be adjusted to achieve the behavior of the computations. One example of the scoring parameter settings (Huang, Chapter 3: Bio-Sequence Comparison and Alignment, ser. Curr Top Comp Mol Biol. Cambridge, Mass.: The MIT Press, 2002) for DNA would be:

MATCH_BONUS: 10

MISMATCH_PEN: −20

INSERTION_PEN: −40

OPENING_PEN: −10

DELETION_PEN: −5

The relationship between the gap penalties (INSERTION_PEN, OPENING_PEN) 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_PEN, MATCH_BONUS, INSERTION_PEN, OPENING_PEN and DELETION_PEN are possible.

In some embodiments, the methods and systems of the invention use a modified Smith-Waterman operation that involves a multi-dimensional look-back through the graph 331. Multi-dimensional operations of the invention provide for a “look-back” type analysis of sequence information (as in Smith-Waterman), wherein the look back is conducted through a multi-dimensional space that includes multiple pathways and multiple nodes. The multi-dimensional algorithm can be used to align sequence reads against the graph-type reference. That alignment algorithm identifies the maximum value for Ci,j by identifying the maximum score with respect to each sequence contained at a position on the graph. In fact, by looking “backwards” at the preceding positions, it is possible to identify the optimum alignment across a plurality of possible paths.

The modified Smith-Waterman operation described here, aka the multi-dimensional alignment, provides exceptional speed when performed in a genomic graph system that employs physical memory addressing (e.g., through the use of native pointers or index free adjacency as discussed above). The combination of multi-dimensional alignment to a graph 331 with the use of spatial memory addresses (e.g., native pointers or index-free adjacency) improves what the computer system is capable of, facilitating whole genomic scale analysis and epigenetic profiling to be performed using the methods described herein.

The operation includes aligning a sequence, or string, to a graph. For the purpose of defining the algorithm, let S be the string being aligned, and let D be the directed graph to which S is being aligned. The elements of the string, S, are bracketed with indices beginning at 1. Thus, if S is the string ATCGAA, S[1] = A, S[4] = G, etc.

In certain embodiments, for the graph, each letter of the sequence of a node will be represented as a separate element, d. In a preferred embodiment, node or edge objects contain the sequences and the sequences are stored as the longest-possible string in each object. A predecessor of d is defined as:

(i) If d is not the first letter of the sequence of its node, the letter preceding d in its node is its (only) predecessor;

(ii) If d is the first letter of the sequence of its node, the last letter of the sequence of any node (e.g., all exons upstream in the genome) that is a parent of d's node is a predecessor of d.

The set of all predecessors is, in turn, represented as P[d].

In order to find the "best" alignment, the algorithm seeks the value of M[j,d], the score of the optimal alignment of the first j elements of S with the portion of the graph preceding (and including) d. This step is similar to finding Hi,j in equation 1 above. Specifically, determining M[j,d] involves finding the maximum of a, i, e, and 0, as defined below:

$$M[j, d] = \max\{a, i, e, 0\} \qquad (6)$$

where

$e = \max\{M[j, p^*] + \text{DELETE\_PEN}\}$ for p^* in $P[d]$

$i = M[j-1, d] + \text{INSERT\_PEN}$

$a = \quad \max\{M[j-1, p^*] + \text{MATCH\_SCORE}\}$ for p^* in $P[d]$, if $S[j] = d$;

$\max\{M[j-1, p^*] + \text{MISMATCH\_PEN}\}$ for p^* in $P[d]$, if $S[j] \neq d$

As described above, e is the highest of the alignments of the first j characters of S with the portions of the graph up to, but not including, d, plus an additional DELETE_PEN. Accordingly, if d is not the first letter of the sequence of the node, then there is only one predecessor, p, and the alignment score of the first j characters of S with the graph (up-to-and-including p) is equivalent to M[j,p] + DELETE_PEN. In the instance where d is the first letter of the sequence of its node, there can be multiple possible predecessors, and because the DELETE_PEN is constant, maximizing [M[j, p*] + DELETE_PEN] is the same as choosing the predecessor with the highest alignment score with the first j characters of S.

In equation (6), i is the alignment of the first j-1 characters of the string S with the graph up-to-and-including d, plus an INSERT_PEN, which is similar to the definition of the insertion argument in SW (see equation 1).

Additionally, a is the highest of the alignments of the first j characters of S with the portions of the graph up to, but not including d, plus either a MATCH_SCORE (if the jth character of S is the same as the character d) or a MISMATCH_PEN (if the jth character of S is not the same as the character d). As with e, this means that if d is not the first letter of the sequence of its node, then there is only one predecessor, i.e., p. That means a is the alignment score of the first j-1 characters of S with the graph (up-to-and-including p), i.e., M[j-1,p], with either a MISMATCH_PEN or MATCH_SCORE added, depending upon whether d and the jth character of S match. In the instance where d is the first letter of the sequence of its node, there can be multiple possible predecessors. In this case, maximizing {M[j, p*] + MISMATCH_PEN or MATCH_SCORE} is the same as choosing the predecessor with the highest alignment score with the first j-1 characters of S (i.e., the highest of the candidate M[j-1,p*] arguments) and adding either a MISMATCH_PEN or a MATCH_SCORE depending on whether d and the jth character of S match.

Again, as in the SW algorithm, the penalties, e.g., DELETE_PEN, INSERT_PEN, MATCH_SCORE and MISMATCH_PEN, can be adjusted to encourage alignment with fewer gaps, etc.

As described in the equations above, the operation finds the optimal (e.g., maximum) value for the sequence 709 by calculating not only the insertion, deletion, and match scores for that element, but looking backward (against the direction of the graph) to any prior nodes on the graph to find a maximum score.

FIG. 9 shows the matrices that represent the comparison. The modified Smith-Waterman operation of the invention identifies the highest score and performs a backtrack to identify the proper alignment of the sequence. See, e.g., U.S. Pub. 2015/0057946 and U.S. Pub. 2015/0056613, both incorporated by reference. Systems and methods of the invention can be used to provide a report that identifies a modified base at the position within the genome of the subject. Other information may be found in Kehr et al., 2014, Genome alignment with graph data structures: a comparison, BMC Bioinformatics 15:99, incorporated by reference.

FIG. 10 illustrates a report of tumor-related mutations for a patient. The report 1001 preferably includes all known tumor-related mutations (if any) that reads are aligned to, along with any new de novo mutations identified. The report may show what proportion of reads aligned to each mutation branch, which counts may be normalized for total number of reads aligning at that position. The normalization step is recommended to correct for variation in coverage. Report 1001 is the patient's tumor-related mutation population at t0. The population of mutations can be understood as an initial assessment of clonality. Method of the invention may be used to monitor a tumor over time. For example, methods may be used to align a second set of sequence reads to the patient-specific genomic reference graph and produce a second report that identifies a patient's second tumor-related mutation population at a time different from an initial time associated with the patient's tumor-related mutation population.

The second report may include a comparison of the patient's second tumor-related mutation population to the patient's tumor-related mutation population.

FIG. 11 shows a report 1101 of patient mutations at a later point in time which may be deemed to be time = t1 (e.g., in the treatment scenario after treatment has had time to take effect; in the screening scenario after a reasonable time has passed such that screening again makes sense). A new set of reads from a newly-sequenced sample of the patient's cell-free plasma DNA is aligned to the patient-specific genomic reference graph. The alignment is used to report all tumor-related mutations that reads are aligned to together with proportion aligned. Thus report 1101 shows the patient's tumor-related mutation population at time = t1 along with a comparison of this population of mutations to the previous round of testing at time = t0. Any of the reports may identify a novel driver mutation present in the cell-free plasma DNA from the patient. Systems and methods of the invention may be used to add the novel driver mutation to the one or more known tumor-associated mutations in the patient-specific genomic reference graph for subsequent uses.

FIG. 12 shows a report describing the population of tumor clones present in a patient. It is understood that transformation and metastases are likely clonal in nature in that they derive from single cells. Indeed, identifying variants from a tumor provides the genotype of the founder cell. Numerous, tumor-associated driver mutations are known from sequencing and screening. A tumor genotype may be characterized by thousands of mutations, many of which are neutral. Data suggest that each tumor has an individually unique genomic profile. But prior attempts to profile tumor genotypes have provided only snapshots from a single time point. Sampling over time promises to allow tracking the evolution of a tumor. It has been suggested that patterns of segregation of mutations within sub-clones is lost when DNA is extracted from the total cell population. Systems and methods of the invention may allow for sub-clones to be properly reconstructed by mapping the cell-free plasma DNA reads to the patient-specific genomic reference graph and in essence haplotyping the results (e.g., as described in co-owned U.S. Provisional Application No. 62/165,403, filed May 22, 2015, and any patent or publication of that provisional, the contents of which are incorporated by reference in their entirety). Systems and methods of the invention may be used to provide reports that describe the clonal evolution of tumors, as well as specifically describing selectively advantageous or ‘driver’ lesions, selectively neutral or ‘passenger’ lesions, and deleterious lesions, as well as lesions that increase the rate of other genetic changes (‘mutator’ lesions). For additional background, see Greaves & Maley, 2012, Clonal evolution in cancer, Nature 481(7381):306-313, the contents of which are incorporated by reference for all purposes.

Discussed above are methods for analyzing tumor DNA using a patient-specific genomic reference graph. The invention also provides systems and methods for analyzing tumor DNA without using a patient-specific reference. The alternative non-patient-specific reference method may be preferable in some instances for being simpler in that it doesn’t require whole-genome sequencing and more invasive sampling. The method for analyzing tumor DNA using a human reference DAG comprises the following steps:

1. building a human reference DAG comprising known reference genomes and known tumor-associated mutations;

2. obtaining sequence reads from a sequenced sample of a patient’s cell-free plasma DNA and aligning the sequence reads to the human reference DAG; and

3. reporting one or more tumor-related mutations (preferably all) that the sequence reads align to, optionally with a proportion of reads aligned to that mutation branch (this may be taken as the patient’s tumor-related mutation population at t0). The method may additionally include:

4. aligning—at a later point in time (e.g., in the treatment scenario after treatment has had time to take effect; in the screening scenario after a reasonable time has passed such that screening again makes sense)—a second set of sequence reads from a newly-sequenced sample of the patient's cell-free plasma DNA to the human reference DAG; and

5. reporting one or more (preferably all) known tumor-related mutations (if any) that sequence reads align to, preferably along with any *de novo* mutations identified, optionally with proportion aligned as in step 3 (which may be taken as the patient's tumor-related mutation population at t1) along with a comparison of this population of mutations to the previous round of testing.

Optionally, systems and methods of the invention may include variant calling 807 to describe *de novo* mutations from the ctDNA. The variant calling can include aligning sequence reads to the graph and reporting SNP alleles in a format such as a Sequence Alignment Map (SAM) or a Variant Call Format (VCF) file. Some background may be found in Li & Durbin, 2009, Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics 25:1754-60 and McKenna et al., 2010, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res 20(9):1297-1303, the contents of each of which are incorporated by reference. Variant calling 831 produces results ("variant calls") that may be stored in a format similar to a sequence alignment map (SAM) or binary alignment map (BAM) file—comprising an alignment string (the SAM format is described, e.g., in Li, et al., The Sequence Alignment/Map format and SAMtools, Bioinformatics, 2009, 25(16):2078-9). Additionally or alternatively, output from the variant calling may be provided in a format similar to a variant call format (VCF) file, e.g., in report 1201. A typical VCF file will include a header section and a data section. The header contains an arbitrary number of meta-information lines, each starting with characters '##', and a TAB delimited field definition line starting with a single '#' character. The field definition line names eight mandatory columns and the body section contains lines of data populating the columns defined by the field definition line. The VCF format is described in Danecek et al., 2011, The variant call format and VCFtools, Bioinformatics 27(15):2156-2158. Further discussion may be found in U.S. Pub. 2013/0073214; U.S. Pub. 2013/0345066; U.S. Pub. 2013/0311106; U.S. Pub. 2013/0059740; U.S. Pub. 2012/0157322; U.S. Pub. 2015/0057946 and U.S. Pub. 2015/0056613, each incorporated by reference. Systems and methods of the invention may be used to describe and report a tumor-related mutation population or specific driver mutations as well as to profile changes over time.

Thus use of systems and methods of the invention provide a product that facilitates oncogenomics and patient counseling. A physician may use a report 1201 provided by the system to determine a medical course of action or counsel a patient on health and wellness issues.

Incorporation by Reference

References and citations to other documents, such as patents, patent applications, patent publications, journals, books, papers, web contents, have been made throughout this disclosure. All such documents are hereby incorporated herein by reference in their entirety for all purposes.

Equivalents

Various modifications of the invention and many further embodiments thereof, in addition to those shown and described herein, will become apparent to those skilled in the art from the full contents of this document, including references to the scientific and patent literature cited herein. The subject matter herein contains important information, exemplification and guidance that can be adapted to the practice of this invention in its various embodiments and equivalents thereof.

What is claimed is:

1. A method for analyzing tumor DNA, the method comprising:

obtaining a patient-specific genomic reference graph that represents known human genomic sequences as well as non-tumor DNA from a patient, the patient-specific genomic reference graph comprising objects in a tangible memory subsystem of a computer system, wherein matching homologous portions of the sequences are each represented by a single object, and further wherein the patient-specific genomic reference graph is annotated to describe one or more known tumor-associated mutations;

aligning sequence reads, obtained by sequencing a sample containing cell-free plasma DNA from the patient, to the patient-specific genomic reference graph to find at least one mutation in the cell-free plasma DNA relative to the non-tumor DNA from the patient or at least one known tumor-associated mutation; and

providing a report that circulating-tumor DNA (ctDNA) in the patient includes the at least one mutation found in the cell-free plasma DNA or at least one known tumor-associated mutation.

2. The method of claim 1, wherein the report identifies a patient's tumor-related mutation population, wherein the tumor-related mutation population includes all mutations in the cell-free plasma DNA relative to the non-tumor DNA from the patient and any of the known tumor-associated mutations found in the sequence reads.

3. The method of claim 2, wherein the report further includes what portion of the sequence reads align to mutation-associated branches in the patient-specific genomic reference graph.

4. The method of claim 2, wherein the report identifies a first clone and a second clone present in a tumor in the patient.

5. The method of claim 2, wherein the report identifies a novel driver mutation present in the cell-free plasma DNA from the patient.

6. The method of claim 5, further comprising adding the novel driver mutation to the one or more known tumor-associated mutations in the patient-specific genomic reference graph for subsequent uses.

7. The method of claim 2, further comprising building the patient-specific genomic reference graph by

obtaining the known human genomic sequences;

obtaining non-tumor sequences for the non-tumor DNA from the patient;

finding matching homologous portions among the known human genomic sequences and the non-tumor sequences;

creating one of the objects in the tangible memory subsystem for each of the matching homologous portions; and

creating connections between pairs of the objects to create the patient-specific genomic reference graph.

8. The method of claim 7, further comprising aligning a second set of sequence reads to the patient-specific genomic reference graph and producing a second report that identifies a patient's second tumor-related mutation population at a time different from an initial time associated with the patient's tumor-related mutation population, wherein the second report includes a comparison of the patient's second tumor-related mutation population to the patient's tumor-related mutation population.

9. The method of claim 7, wherein the objects of the patient-specific genomic reference graph include pointers to adjacent ones of the objects such that the objects are linked into paths to represent the plurality of known human genomic sequences, wherein each pointer identifies a physical location in the memory subsystem at which the adjacent object is stored.

10. The method of claim 9, further comprising obtaining the sequence reads by sequencing the sample containing the cell-free plasma DNA from the patient.

11. A system for analyzing tumor DNA, the system comprising a processor coupled to a memory subsystem having stored therein:

a patient-specific genomic reference graph representing known human genomic sequences as well as non-tumor DNA from a patient, wherein matching homologous portions of the sequences are each represented by a single one of a plurality of objects stored in the memory subsystem; and

instructions executable by the processor to cause the system to

align sequence reads from cell-free plasma DNA of the patient to the patient-specific genomic reference graph to find at least one mutation in the cell-free plasma DNA relative to the non-tumor DNA from the patient; and

provide a report that circulating-tumor DNA (ctDNA) in the patient includes the at least one mutation found in the cell-free plasma DNA.

12. The system of claim 11, wherein the report identifies a patient's tumor-related mutation population, wherein the tumor-related mutation population includes all mutations in the cell-free plasma DNA relative to the non-tumor DNA from the patient and any of the known tumor-associated mutations found in the sequence reads.

13. The system of claim 12, wherein the report further includes what portion of the sequence reads align to mutation-associated branches in the patient-specific genomic reference graph.

14. The system of claim 12, wherein the report identifies a first clone and a second clone present in a tumor in the patient.

15. The system of claim 12, wherein the report identifies a novel driver mutation present in the cell-free plasma DNA from the patient.

16. The system of claim 15, further operable to add the novel driver mutation to the one or more known tumor-associated mutations in the patient-specific genomic reference graph for subsequent uses.

17. The system of claim 12, further operable to build the patient-specific genomic reference graph by

obtaining the known human genomic sequences;

obtaining non-tumor sequences for the non-tumor DNA from the patient;

finding matching homologous portions among the known human genomic sequences and the non-tumor sequences;

creating one of the objects in the tangible memory subsystem for each of the matching homologous portions; and

creating connections between pairs of the objects to create the patient-specific genomic reference graph.

18. The system of claim 17, further operable to align a second set of sequence reads to the patient-specific genomic reference graph and produce a second report that identifies a patient's tumor-related mutation population at a time different from an initial time associated with the patient's tumor-related mutation population, wherein the second report includes a comparison of the patient's second tumor-related mutation population to the patient's tumor-related mutation population.

19. The system of claim 17, wherein the objects of the patient-specific genomic reference graph include pointers to adjacent ones of the objects such that the objects are linked into paths to represent the plurality of known human genomic sequences, wherein each pointer identifies a physical location in the memory subsystem at which the adjacent object is stored.

20. The system of claim 17, further comprising a nucleic acid sequencing instrument, the system further operable to obtain the sequence reads by sequencing the sample containing the cell-free plasma DNA from the patient using the nucleic acid sequencing instrument.

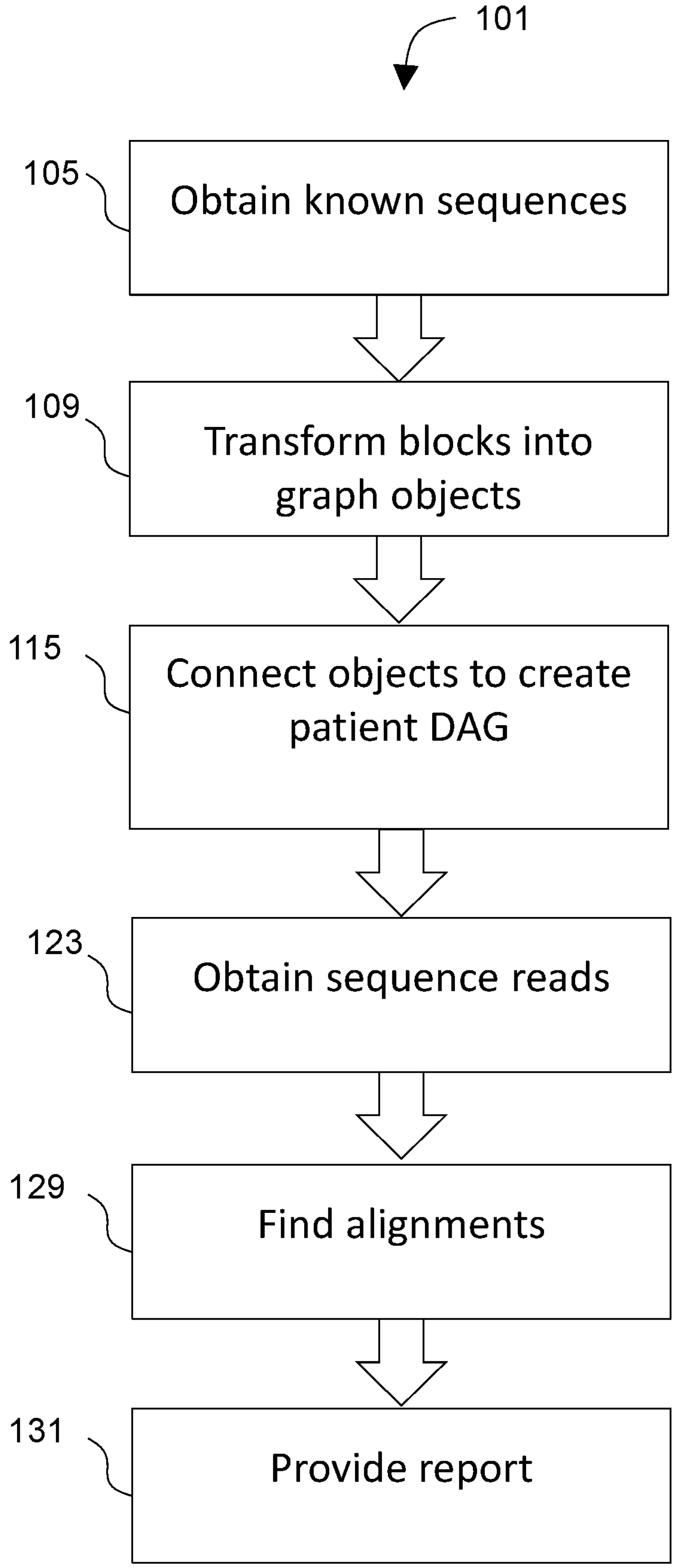

101
105
Obtain known sequences
109
Transform blocks into graph objects
115
Connect objects to create patient DAG
123
Obtain sequence reads
129
Find alignments
131
Provide report

FIG. 1

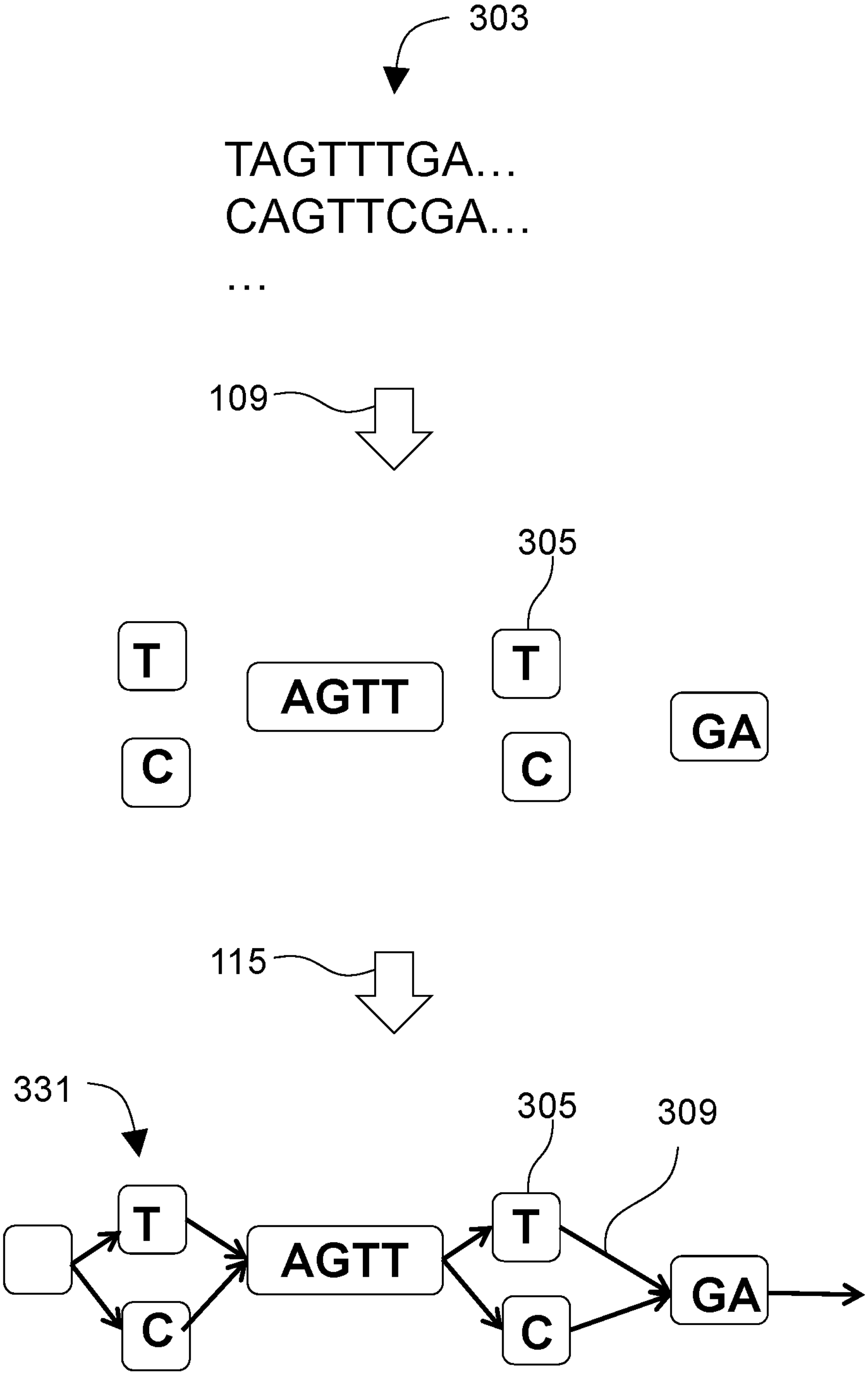

303
TAGTTTGA…
CAGTTCGA…
…
109
T
C
AGTT
305
T
C
GA
115
331
T
C
AGTT
305
T
C
309
GA

FIG. 2

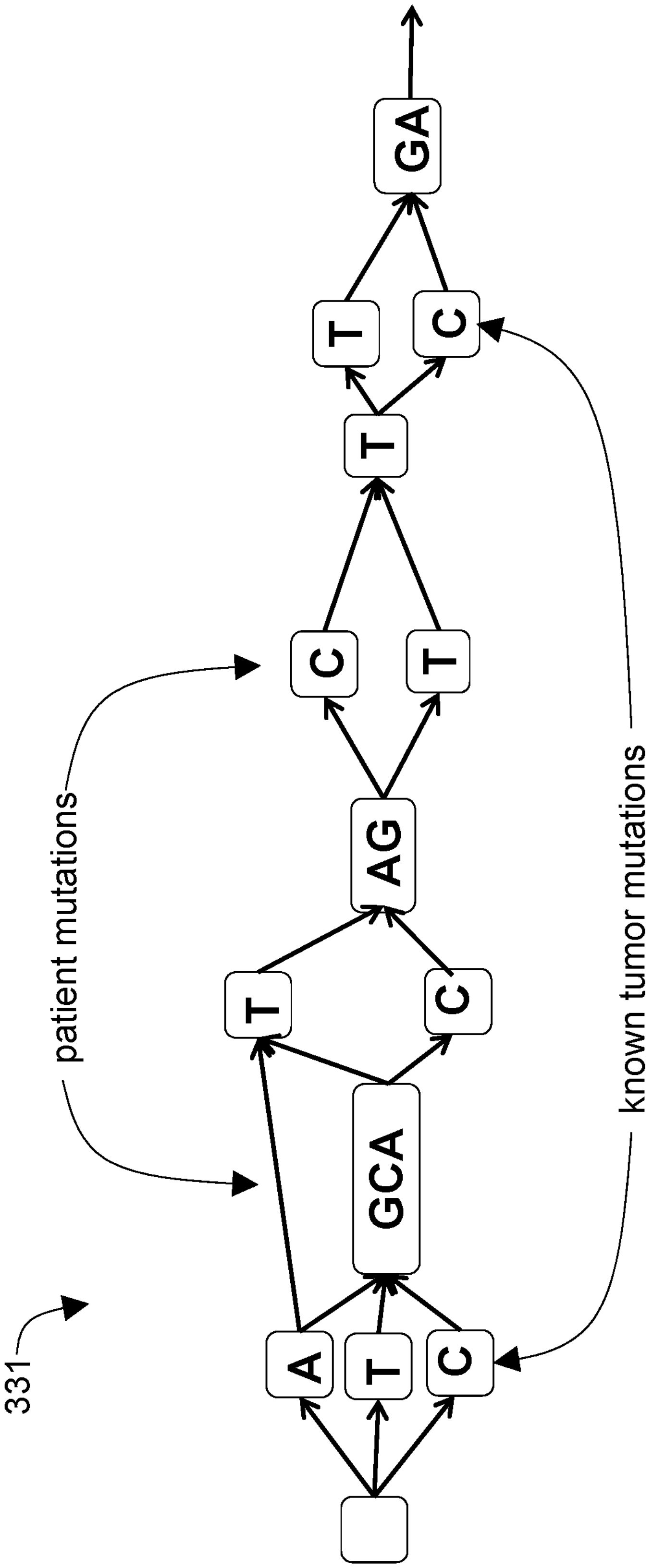

331
A
T
C
GCA
T
C
AG
C
T
T
T
C
GA
patient mutations
known tumor mutations

FIG. 3

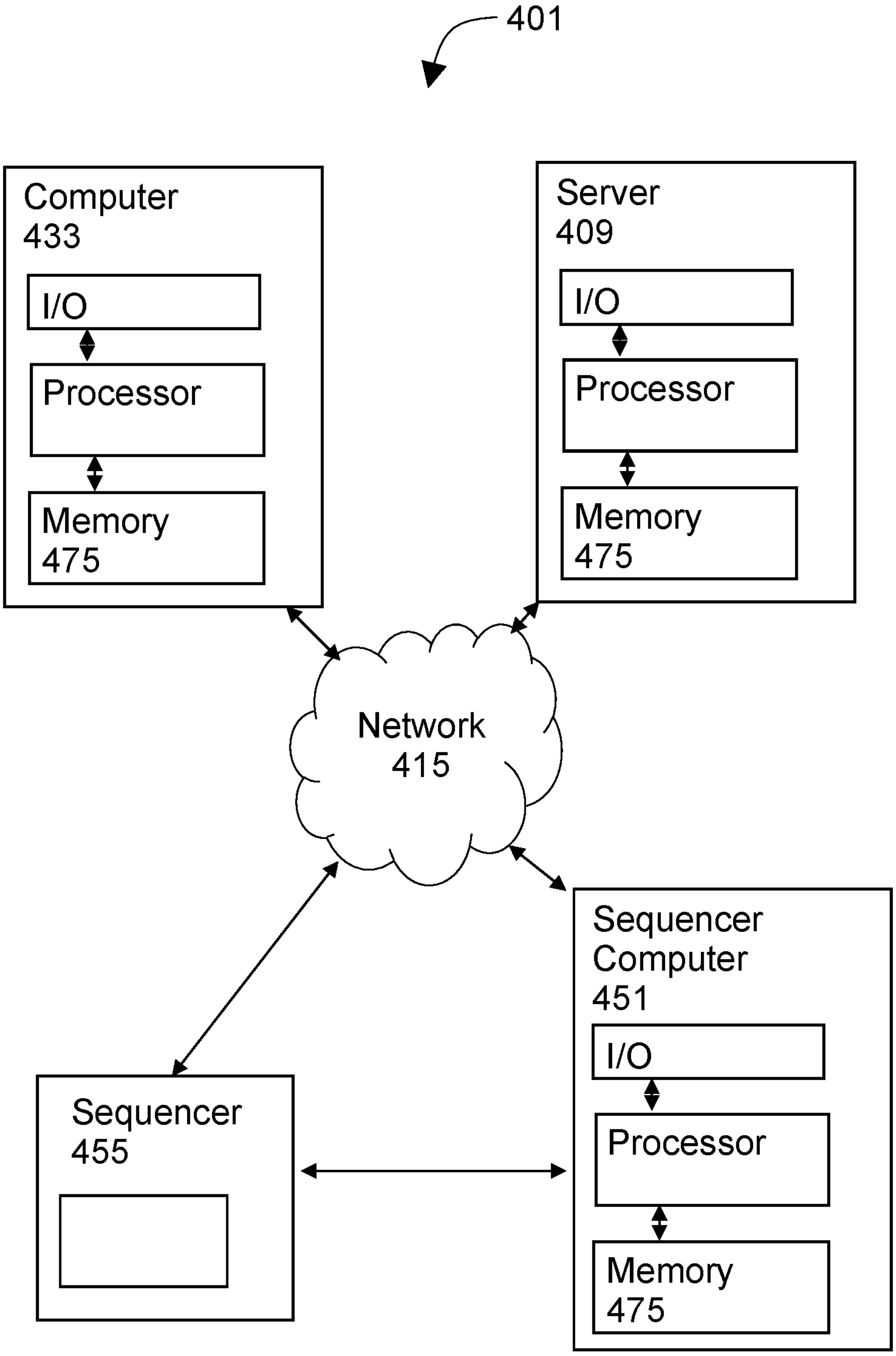

401
Computer
433
I/O
Processor
Memory
475
Server
409
I/O
Processor
Memory
475
Network
415
Sequencer
455
Sequencer
Computer
451
I/O
Processor
Memory
475

FIG. 4

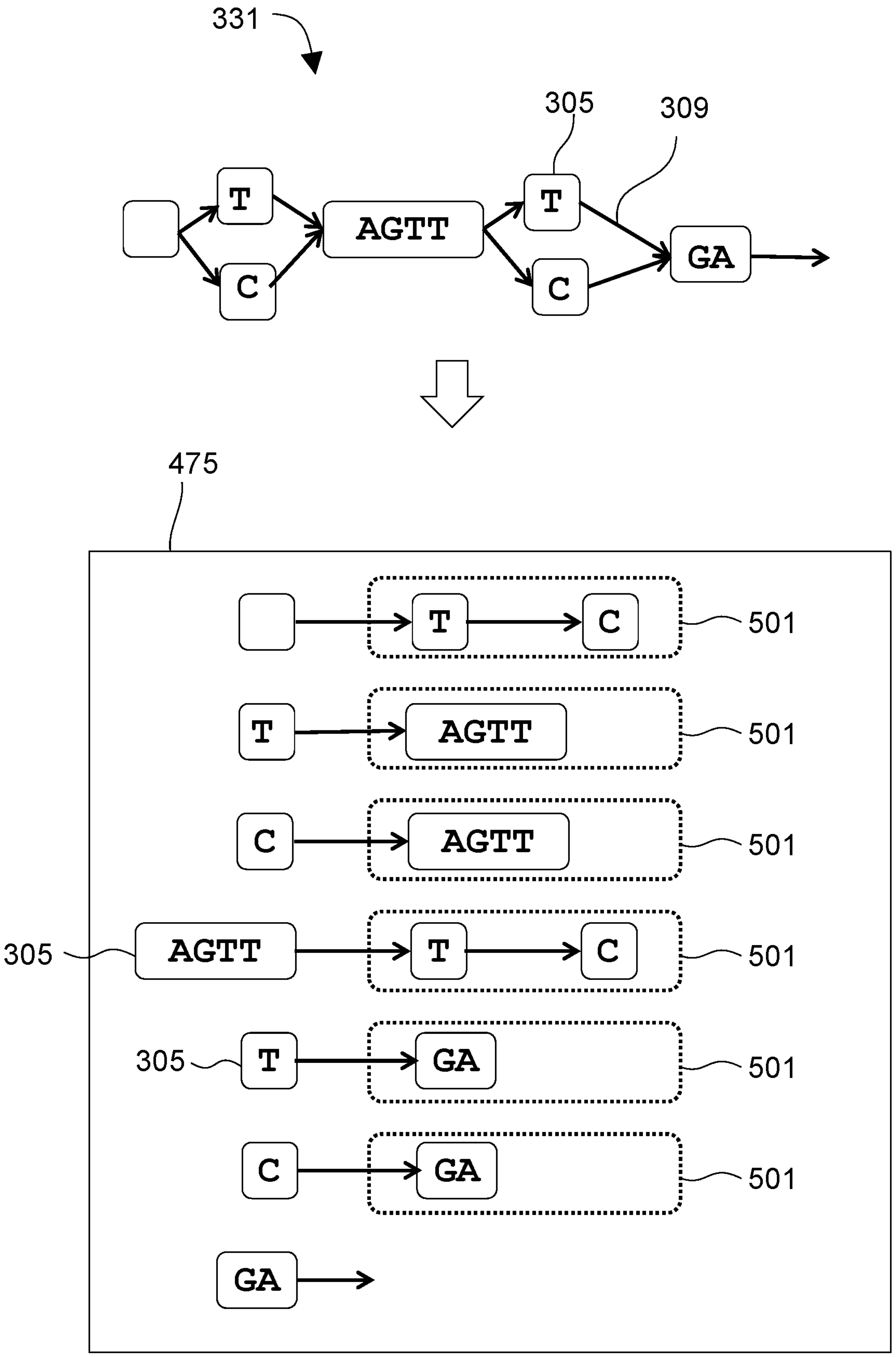

331
305
309
T
C
AGTT
T
C
GA
475
T
C
501
T
AGTT
501
C
AGTT
501
305
AGTT
T
C
501
305
T
GA
501
C
GA
501
GA

FIG. 5

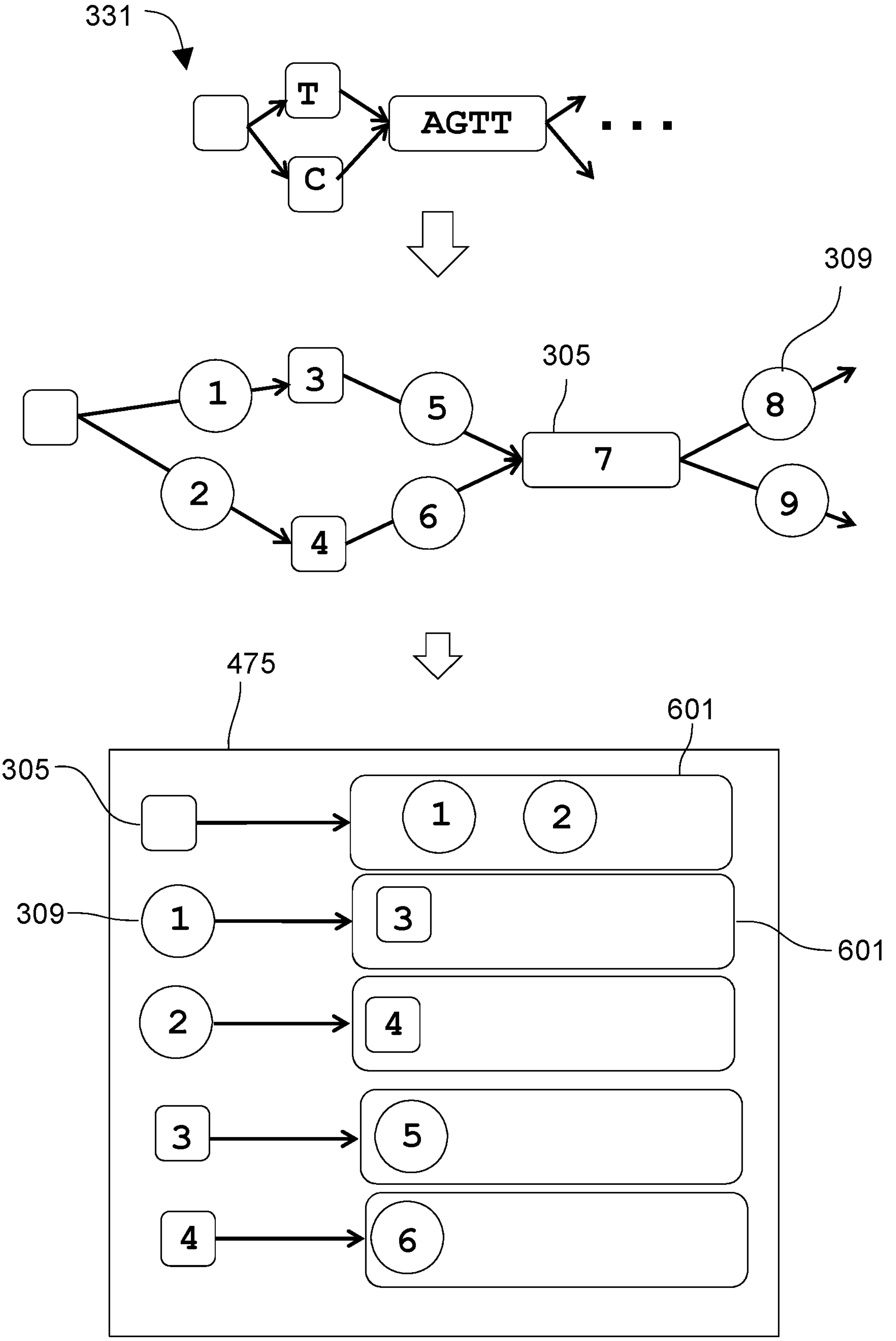
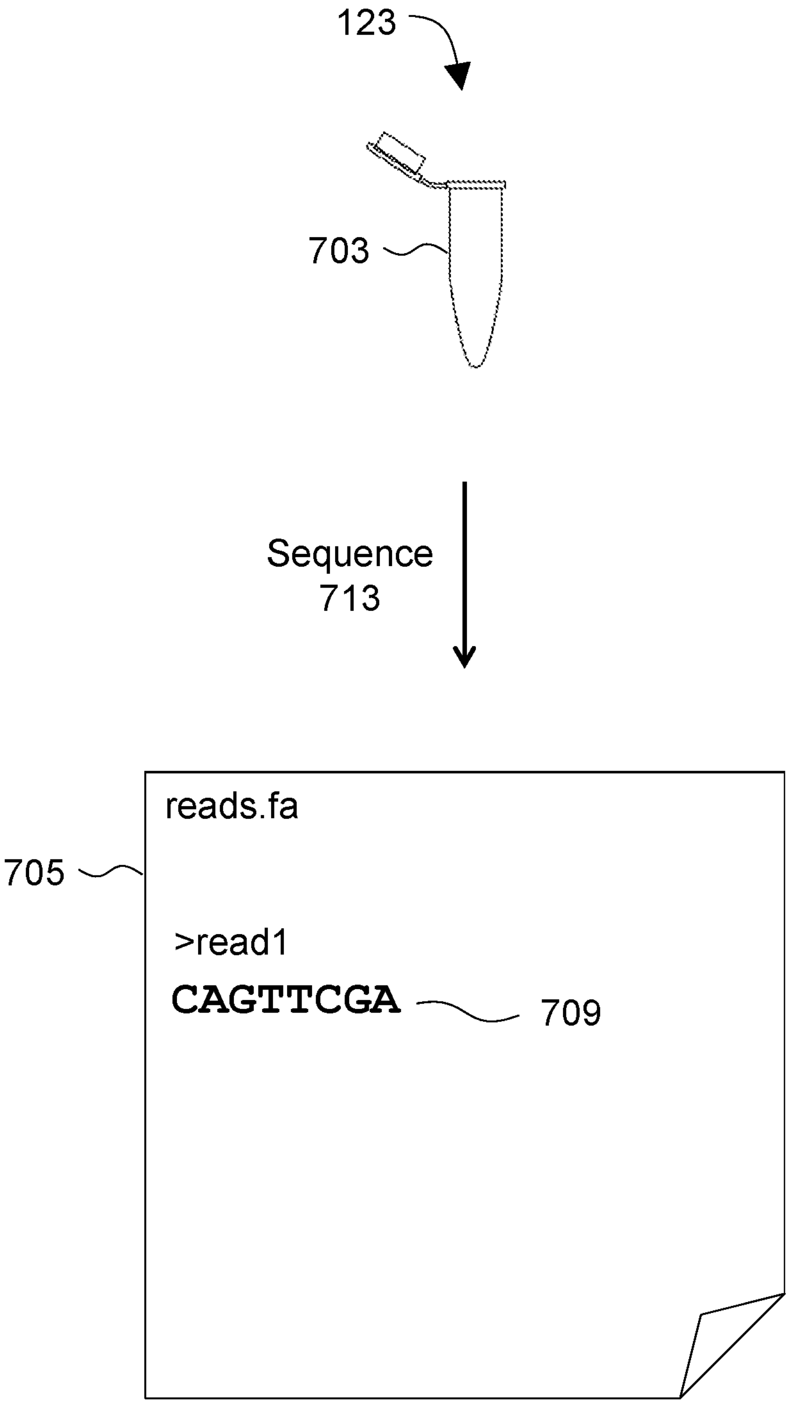


FIG. 6



```
>read1
CAGTTCGA
```

FIG. 7

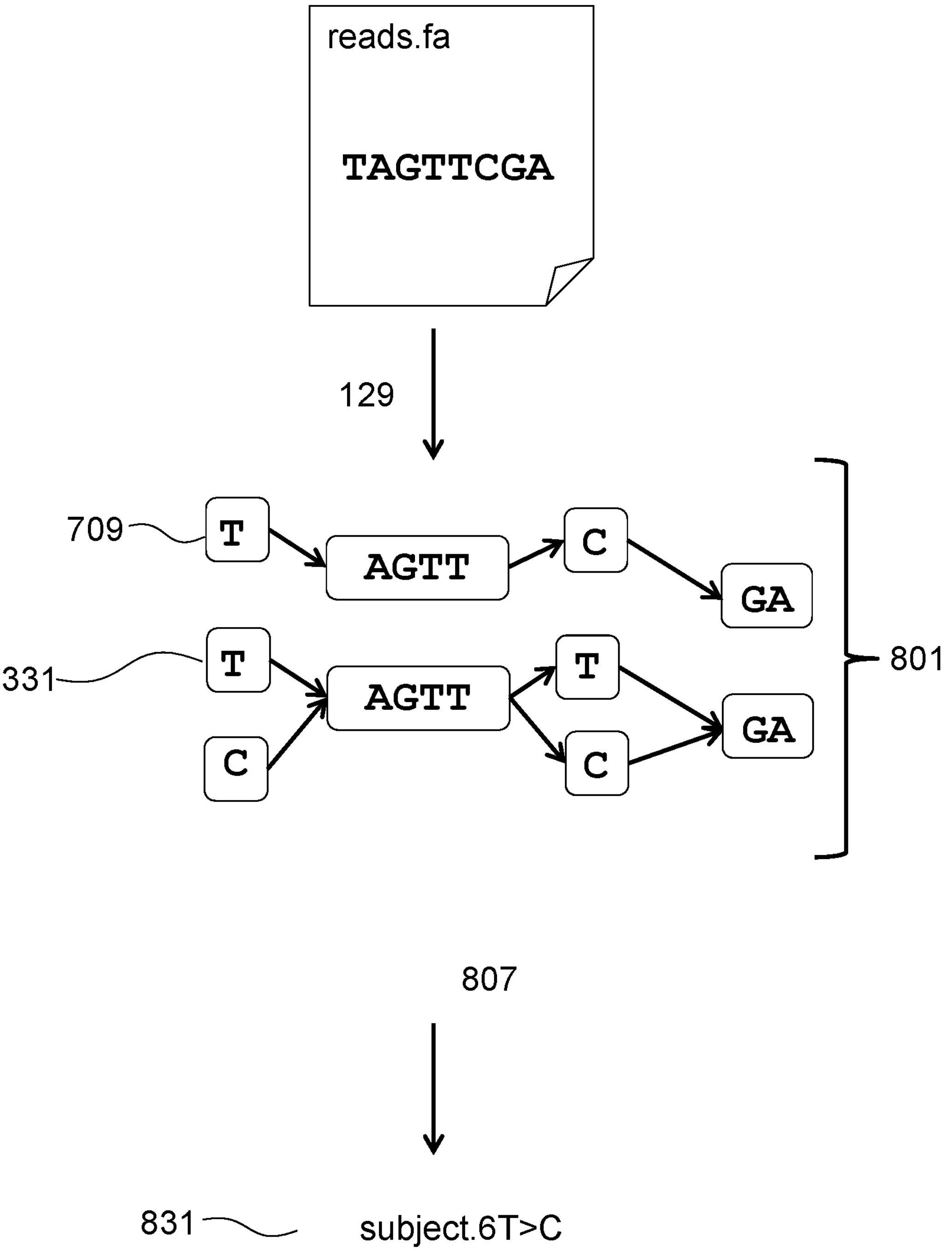
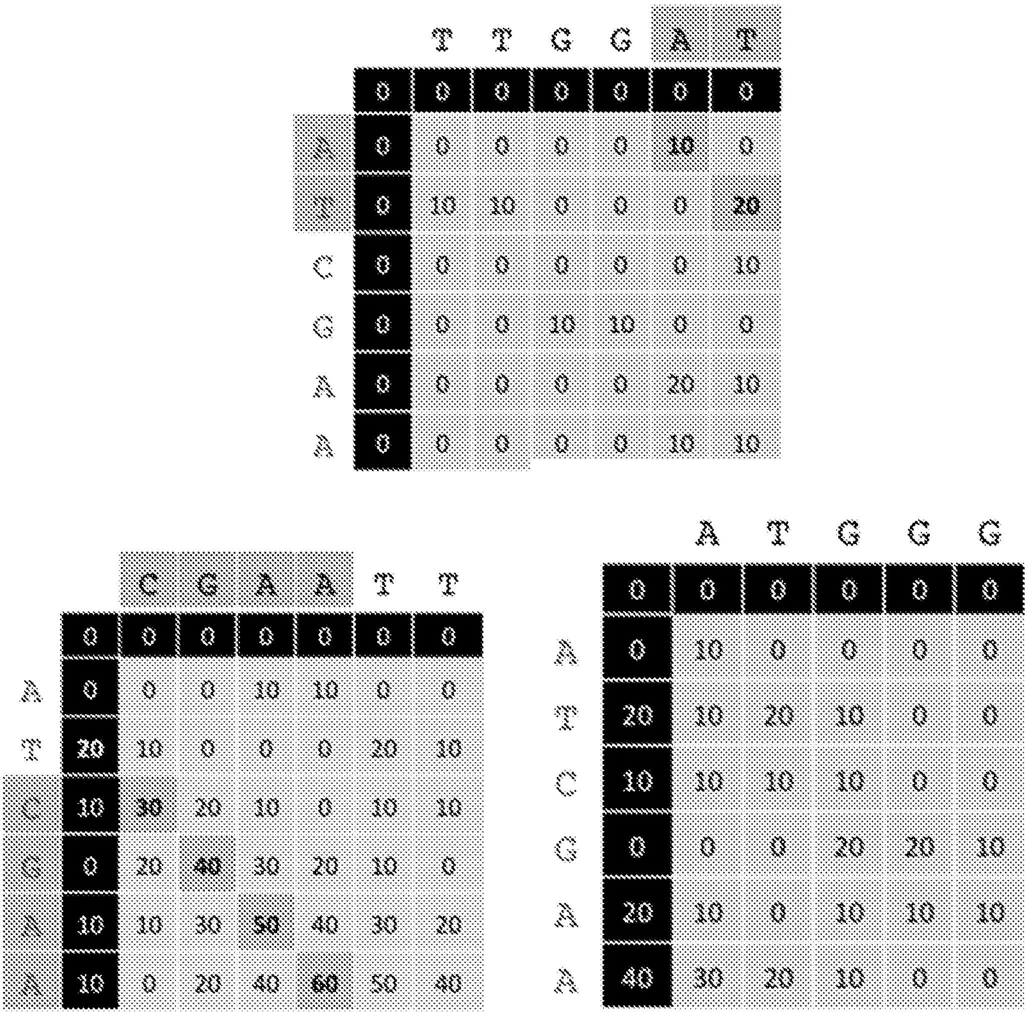

reads.fa
TAGTTCGA
129
709
T
AGTT
C
GA
331
T
C
AGTT
T
C
GA
801
807
831
subject.6T>C

FIG. 8

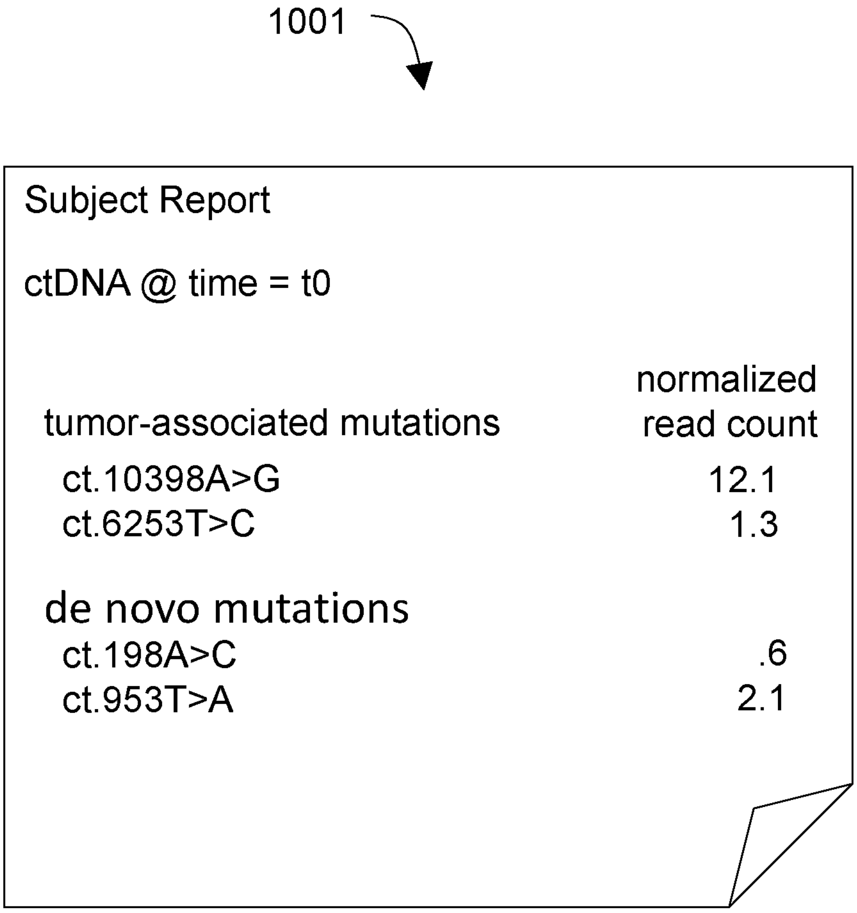


		T	T	G	G	A	T
	0	0	0	0	0	0	0
A	0	0	0	0	0	10	0
T	0	10	10	0	0	0	20
C	0	0	0	0	0	0	10
G	0	0	0	10	10	0	0
A	0	0	0	0	0	20	10
A	0	0	0	0	0	10	10

		C	G	A	A	T	T
	0	0	0	0	0	0	0
A	0	0	0	10	10	0	0
T	20	10	0	0	0	20	10
C	10	30	20	10	0	10	10
G	0	20	40	30	20	10	0
A	10	10	30	50	40	30	20
A	10	0	20	40	60	50	40

		A	T	G	G	G
	0	0	0	0	0	0
A	0	10	0	0	0	0
T	20	10	20	10	0	0
C	10	10	10	10	0	0
G	0	0	0	20	20	10
A	20	10	0	10	10	10
A	40	30	20	10	0	0

FIG. 9

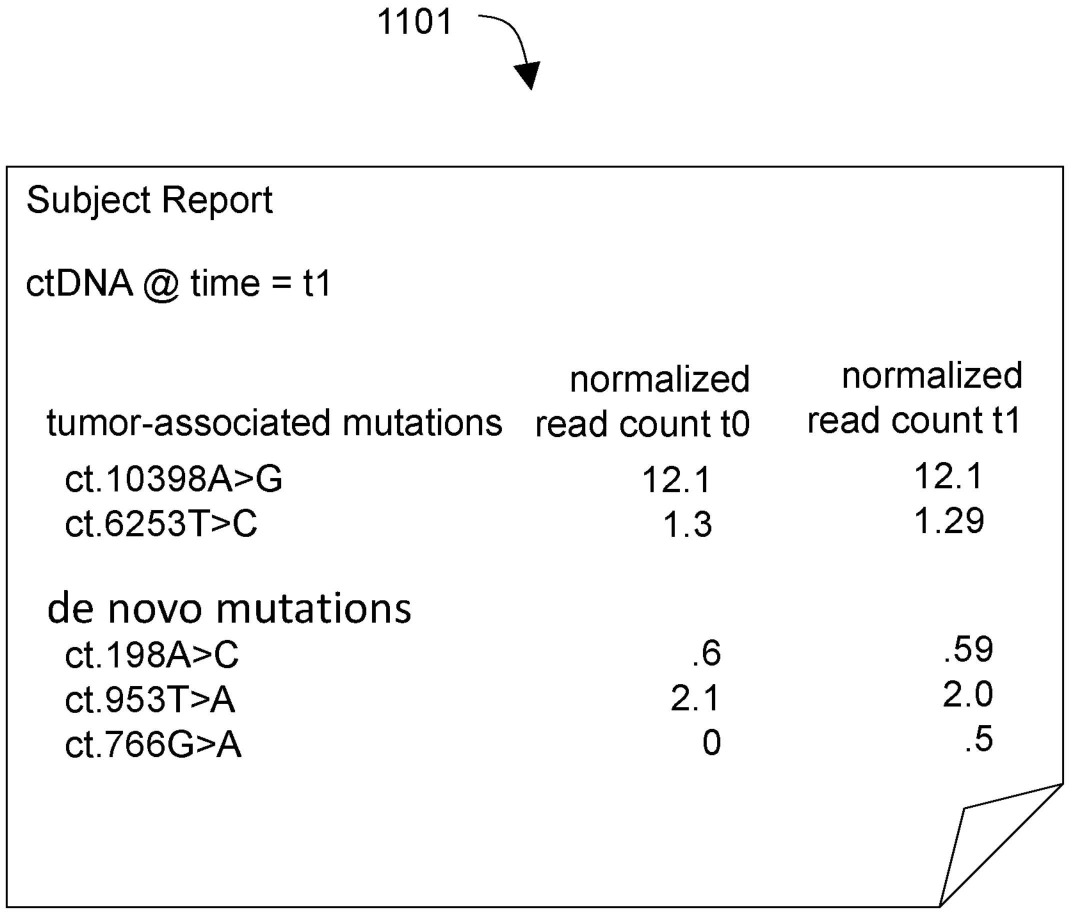


Subject Report

ctDNA @ time = t0

tumor-associated mutations	normalized read count
ct.10398A>G	12.1
ct.6253T>C	1.3
de novo mutations	
ct.198A>C	.6
ct.953T>A	2.1

FIG. 10



Subject Report

ctDNA @ time = t1

tumor-associated mutations	normalized read count t0	normalized read count t1
ct.10398A>G	12.1	12.1
ct.6253T>C	1.3	1.29
de novo mutations		
ct.198A>C	.6	.59
ct.953T>A	2.1	2.0
ct.766G>A	0	.5

FIG. 11

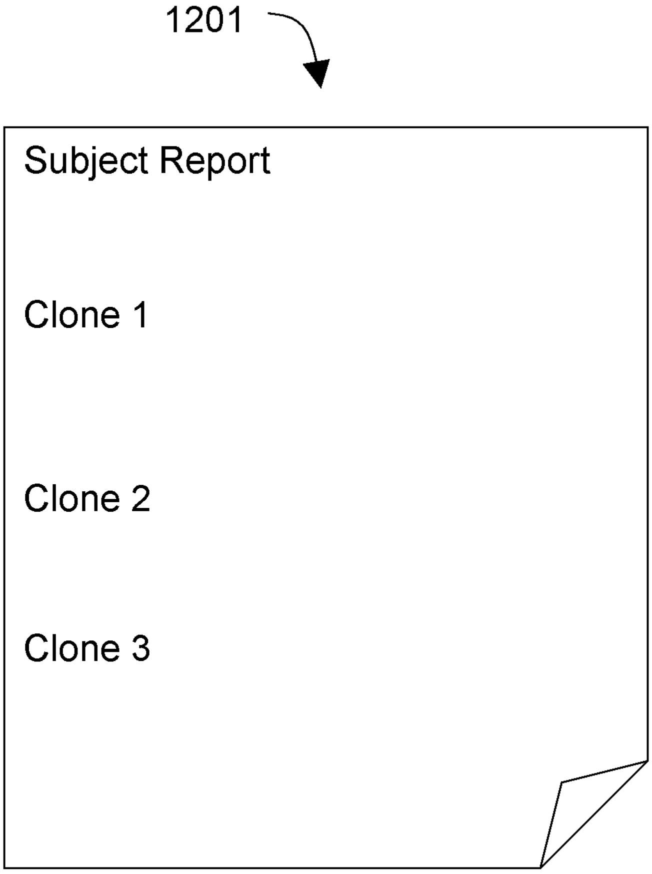


FIG. 12

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US17/13329

A. CLASSIFICATION OF SUBJECT MATTER

IPC - C12Q 1/68; G01N 33/50, 33/574; G06F 19/22 (2017.01)

CPC - C12Q 1/68, 1/6883, 1/6886, 1/6869; G01N 33/5017, 33/574, 33/57415; G06F 19/22

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
See Search History document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
See Search History document

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 2015/027050 A1 (SEVEN BRIDGES GENOMICS INC.) 26 February 2015; abstract; page 6, paragraphs 1-2; page 9, paragraph 1; page 16, paragraph 1	1-20
A	US 2015/0344970 A1 (THE JOHN HOPKINS UNIVERSITY.) 03 December 2015; paragraphs [0006]-[0007], [0023], [0049]	1-20
A	US 2010/0041048 A1 (DIEHL, F et al.) 18 February 2010; abstract; paragraphs [0053], [0081], [0095]; figure 1	1-20
A	(NEWMAN, AM et al.) An Ultrasensitive Method for Quantitating Circulating Tumor DNA with Broad Patient Coverage. Nature Medicine. 06 April 2014; Vol. 20, No. 5; pages 1-22; abstract; page 4, paragraph 4; page 8, paragraphs 2-4; page 9, paragraph 3; DOI: 10.1038/nm.3519.	1-20
A	(OLSSON, E et al.) Serial Monitoring of Circulating Tumor DNA in Patients with Primary Breast Cancer for Detection of Occult Metastatic Disease. EMBO Molecular Medicine. 18 May 2015; Vol. 7, No. 8; pages 1034-1047; page 1037, column 1, paragraph 2; page 1042, column 1, paragraph 2; page 1044, column 1, paragraph 2; figures 4-5; DOI: 10.15252/emmm.201404913.	1-20

□ Further documents are listed in the continuation of Box C. □ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search
12 March 2017 (12.03.2017)

Date of mailing of the international search report
07 APR 2017

Name and mailing address of the ISA/
Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, Virginia 22313-1450
Facsimile No. 571-273-8300

Authorized officer
Shane Thomas
PCT Helpdesk: 571-272-4300
PCT OSP: 571-272-7774