Office de la Propriété Intellectuelle du Canada
Un organisme d'Industrie Canada

Canadian Intellectual Property Office
An agency of Industry Canada

CA 2952503 A1 2015/12/30

(21) 2 952 503

(12) DEMANDE DE BREVET CANADIEN
CANADIAN PATENT APPLICATION

(13) A1

(86) Date de dépôt PCT/PCT Filing Date: 2015/06/26
(87) Date publication PCT/PCT Publication Date: 2015/12/30
(85) Entrée phase nationale/National Entry: 2016/12/14
(86) N° demande PCT/PCT Application No.: US 2015/038175
(87) N° publication PCT/PCT Publication No.: 2015/200891
(30) Priorité/Priority: 2014/06/26 (US62/017,589)

(51) Cl.Int./Int.Cl. *G06F 19/22* (2011.01)

(71) Demandeur/Applicant:
10X GENOMICS, INC., US

(72) Inventeurs/Inventors:
SCHNALL-LEVIN, MICHAEL, US;
MACCALLUM, IAIN, US

(74) Agent: FASKEN MARTINEAU DUMOULIN LLP

(54) Titre : PROCEDES ET SYSTEMES POUR L'ASSEMBLAGE DE SEQUENCES D'ACIDE NUCLEIQUE
(54) Title: PROCESSES AND SYSTEMS FOR NUCLEIC ACID SEQUENCE ASSEMBLY

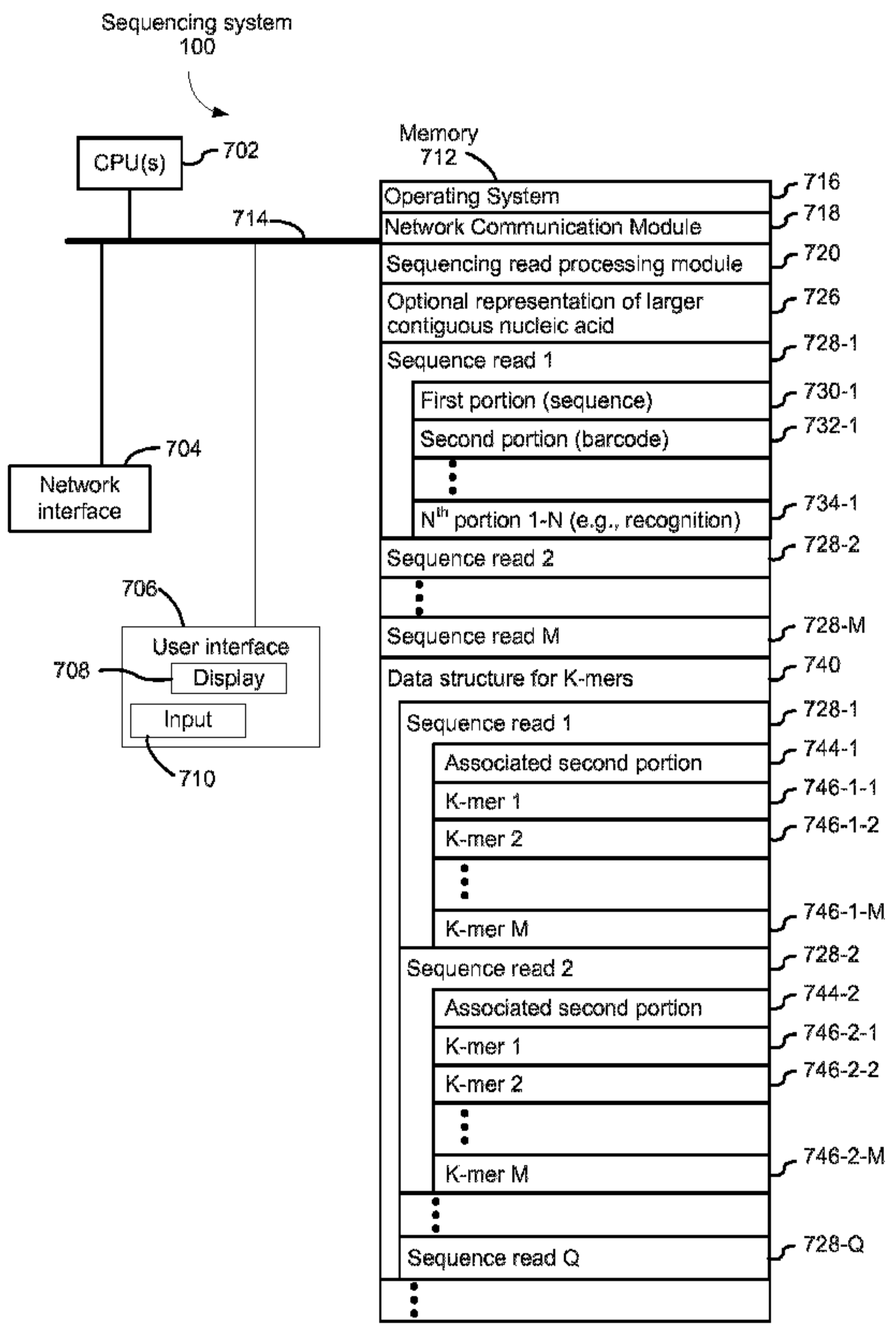


Figure 7

(57) Abrégé/Abstract:
Methods, processes, and particularly computer implemented processes and computer program products are provided for use in the analysis of genetic sequence data. The processes and products are employed in the assembly of shorter nucleic acid sequence data into longer linked and preferably contiguous genetic constructs, including large contigs, chromosomes and whole genomes.

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
30 December 2015 (30.12.2015)

WIPO | PCT

(10) International Publication Number
WO 2015/200891 A1

(51) International Patent Classification:
G06F 19/22 (2011.01)

(21) International Application Number:
PCT/US2015/038175

(22) International Filing Date:
26 June 2015 (26.06.2015)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/017,589 26 June 2014 (26.06.2014) US

(71) Applicant: **10X TECHNOLOGIES, INC.** [US/US]; 7068 Koll Center Parkway, Suite 401, Pleasanton, CA 94566 (US).

(72) Inventors: **SCHNALL-LEVIN, Michael**; 7068 Koll Center Parkway, Suite 401, Pleasanton, CA 94566 (US). **MACCALLUM, Iain**; 7068 Koll Center Parkway, Suite 401, Pleasanton, CA 94566 (US).

(74) Agents: **LOVEJOY, Brett A.** et al.; Morgan, Lewis & Bockius LLP, One Market, Spear Street Tower, San Francisco, CA 94105 (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, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, 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).

Declarations under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

[Continued on next page]

(54) Title: PROCESSES AND SYSTEMS FOR NUCLEIC ACID SEQUENCE ASSEMBLY

(57) Abstract: Methods, processes, and particularly computer implemented processes and computer program products are provided for use in the analysis of genetic sequence data. The processes and products are employed in the assembly of shorter nucleic acid sequence data into longer linked and preferably contiguous genetic constructs, including large contigs, chromosomes and whole genomes.

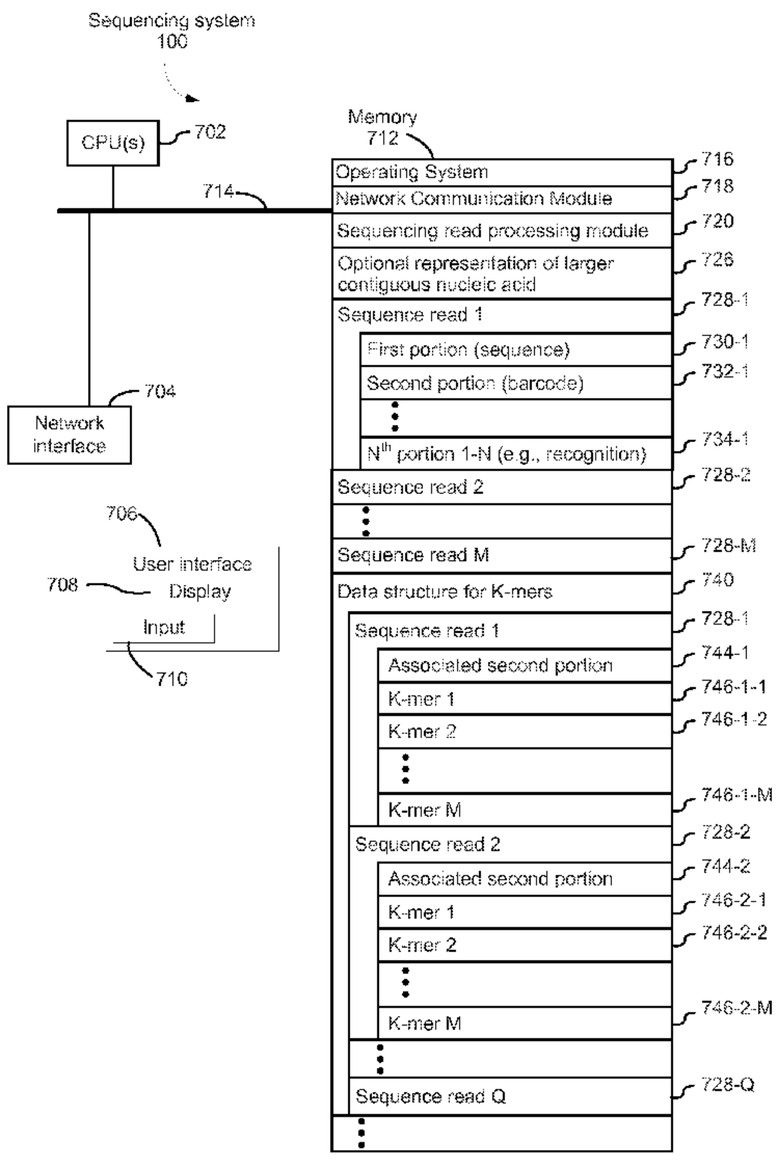


Figure 7

WO 2015/200891 A1

WO 2015/200891 A1

— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

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

PROCESSES AND SYSTEMS FOR NUCLEIC ACID SEQUENCE ASSEMBLY

CROSS-REFERENCE TO RELATED APPLICATION

[0001] This Application claims priority to United States Patent Application Number 62/017,589, filed June 26, 2014, entitled "Processes and Systems for Nucleic Acids Sequence Assembly," which is hereby incorporated by reference.

BACKGROUND OF THE INVENTION

[0002] One of the significant challenges associated with high throughput next generation sequencing technologies is in assembling relatively short sequence reads into much longer contiguous sequences. Genomic sequence assembly processes are often analogized to having a novel that is cut into small pieces, which one must then be reassembled into the complete novel. In sequence assembly, this is generally accomplished by piecing together overlapping sub-segments of the whole sequence that is to be assembled. As will be appreciated, that assembly process becomes easier when the pieces are larger and less ambiguous. For example, it would be easier to assemble a novel based upon full paragraphs, pages or chapters than it would from individual sentences, sentence fragments or random words or parts of words. Likewise, with nucleic acid sequencing, the shorter the individual sequence read, the more difficult it becomes to assemble multiple reads into a longer contiguous sequence.

[0003] While next generation sequencing technologies are capable of generating vast amounts of sequence data, e.g., generating 1 terabase of sequence data within a single run, they are plagued by the difficulties of producing only short sequence read lengths. In particular, these sequencing technologies generally acquire sequence data in reads of 100 contiguous bases or less (or when sequenced as paired end sequencing, up to approximately 200 bases). These reads then must be assembled into longer contiguous sequences. While certain sequencing technologies provide longer sequence reads of 800, 1000, 5000, or even 20,000 bases in length, this often comes at the price of system throughput, yielding only hundreds of megabases of sequence data per run.

[0004] A number of processes and algorithms have been employed in assembling relatively short sequence reads into longer contiguous sequences. In general, these short sequence reads

typically overlap with a number of other short sequence reads to provide redundant coverage over different parts of the overall longer sequence (referred to as "coverage" or "fold coverage"). At a high level, the overlapping sequences from numerous reads are used to piece together the longer sequence information. In many cases, existing known reference sequences, e.g., from the same species, are available. These reference sequences may be used as a scaffold for mapping the shorter sequence reads onto a larger genome or genome fragment. The multifold sequence coverage is then layered on top of the scaffold to provide a relatively high confidence sequence.

[0005] In many cases, however, reference sequences may not be available, requiring de novo assembly of the genome or genomic fragment being sequenced. Moreover, in many cases, such reference sequences are not helpful because the target of the sequencing efforts it to detect mutations in the target sequence. For instance, in the case where the target sequence is from a cancer, it is import to ascertain which mutations, insertions, and deletions are in the target sequence as a way of diagnosing cancer stage or some other attribute of the cancer. As such, making use of references sequences is not always useful in sequencing efforts.

[0006] Accordingly, described herein are improved methods, processes and systems for the assembly of sequence reads into larger genome or genome fragment sequences, and particularly, for de novo assembly of sequence information into larger sequence contexts, including for example, whole chromosome or even whole genome contexts.

SUMMARY

[0007] Described herein are improved genetic sequence assembly processes used for assembly of sequence reads from a variety of different sequencing systems, including e.g., both short and long read sequencing systems. The processes described herein may be used in both de novo assembly processes and in re-sequencing assembly processes or assembly processes against known reference sequences.

[0008] The methods, processes and systems described herein employ sequence reads with associated barcode sequences to assist in the overall assembly process at one or more different assembly steps.

[0009] In certain aspects, provided are methods of assembling nucleic acid sequence reads into larger contiguous sequences. The methods comprise identifying a first subset of sequence reads that comprise both overlapping sequences and a common barcode sequence in a computer

implemented system comprising a first data structure comprising a plurality of sequence reads derived from a larger contiguous nucleic acid. Sequence reads derived from a common fragment of the larger contiguous nucleic acid comprise a common barcode sequence. The first subset of sequence reads is aligned to provide a contiguous linear nucleic acid sequence

[0010] Also provided are methods of assembling nucleic acid sequence reads into larger contiguous sequences. Such methods identify a first subsequence from a set of overlapping sequence reads in a computer implemented system comprising a first data structure comprising a plurality of sequence reads derived from a larger contiguous nucleic acid. The first subsequence is then extended to one or more adjacent or overlapping sequences based upon the presence of a barcode sequence on the adjacent sequence that is common to the first subsequence. A linear nucleic acid sequence is then provided that comprises the first subsequence and the one or more adjacent sequences.

[0011] One aspect of the present disclosure provides a sequencing method comprising, at a computer system having one or more processors, and memory storing one or more programs for execution by the one or more processors, implementing a method in which a plurality of sequence reads is obtained. The plurality of sequence reads comprises a plurality of sets of sequence reads. Each respective sequence read in a set of sequence reads includes (i) a first portion that corresponds to a subset of a larger contiguous nucleic acid and (ii) a common second portion that forms an identifier that is independent of the sequence of the larger contiguous nucleic acid and that identifies a partition, in a plurality of partitions, in which the respective sequence read was formed. Each respective set of sequence reads in the plurality of sets of sequence reads is formed in a partition in the plurality of partitions and each partition includes one or more fragments (e.g., two or more, three or more, ten or more) of the larger contiguous nucleic acid that is used as the template for each respective sequence read in the partition.

[0012] In the method, a respective set of k-mers is created for each sequence read in the plurality of sequence reads. The sets of k-mers collectively comprise a plurality of k-mers. The identifier of the sequence read for each k-mer in the plurality of k-mers is retained. In many instances, there are multiple such sequencing read identifiers for at least some of the k-mers. The value *k* is less than the average length of the sequence reads in the plurality of sequence reads. Each respective set of k-mers includes some (e.g., at least eighty percent) of the possible k-mers of length k of the first portion of the corresponding sequence read.

[0013] In the method, there is tracked, for each respective k-mer in the plurality of k-mers, an identity of each sequence read in the plurality of sequence reads that contains the respective k-mer and the identifier of the set of sequence reads that contains the sequence read.

[0014] In the method, all or a portion (e.g., at least 1 percent, at least 5 percent, at least fifty percent) of the plurality k-mers are graphed as a graph comprising a plurality of nodes connected by a plurality of directed arcs. Each node comprises an uninterrupted set of k-mers in the plurality of k-mers of length k with k-1 overlap. Each arc connects an origin node to a destination node in the plurality of nodes. A final k-mer of an origin node has k-1 overlap with an initial k-mer of a destination node. A first origin node has a first directed arc with both a first destination node and a second destination node in the plurality of nodes.

[0015] In the method, a determination is made as to whether to merge the origin node with the first destination node or the second destination node in order to derive a contig sequence that is more likely to be representative of a portion of the larger contiguous nucleic acid. The contig sequence comprises (i) the origin node and (ii) one of the first destination node and the second destination node. The determining uses at least the identifiers of the sequence reads for k-mers in the first origin node, the first destination node, and the second destination node.

[0016] Another aspect provides a computing system, comprising one or more processors, memory storing one or more programs to be executed by the one or more processors, the one or more programs comprising instructions for executing the method described above.

[0017] Another aspect of the present disclosure provides a sequencing method comprising, at a computer system having one or more processors, and memory storing one or more programs for execution by the one or more processors, obtaining a plurality of sequence reads. The plurality of sequence reads comprises a plurality of sets of sequence reads. Each respective sequence read in a set of sequence reads includes (i) a unique first portion that corresponds to a subset of a larger contiguous nucleic acid and (ii) a common second portion that forms an identifier that is independent of the sequence of the larger contiguous nucleic acid and that identifies a partition, in a plurality of partitions, in which the respective sequence read was formed. Each respective set of sequence reads in the plurality of sets of sequence reads is formed in a partition in the plurality of partitions. Each such partition includes one or more fragments of the larger contiguous nucleic acid that is used as the template for each respective sequence read in the partition.

[0018] In the method, a respective set of k-mers is created for each sequence read in the plurality of sequence reads. The sets of k-mers collectively comprise a plurality of k-mers. The identifier of the sequence read for each k-mer in the plurality of k-mers is retained. The value *k* is less than the average length of the sequence reads in the plurality of sequence reads. Each respective set of k-mers includes at least some of (e.g., at least eighty percent of) the possible k-mers of the first portion of the corresponding sequence read.

[0019] In the method, there is tracked, for each respective k-mer in the plurality of k-mers, an identity of the each sequence read in the plurality of sequence reads that contains the respective k-mer. A first path is identified. The first path comprises a first set of k-mers with k-1 overlap in the plurality of k-mers. A second path is identified. The second path comprises a second set of k-mers with k-1 overlap in the plurality of k-mers. The first path intersects the second path (e.g., as illustrated in Figure 15), thereby forming the set of branch segments comprising: a left portion of the first path, a right portion of the first path, a left portion of the second path and a right portion of the second path. The identifiers associated with the k-mers of the set of branch segments are used to verify the connectivity of the branch segments. In some embodiments, this comprises evaluating a number of identifiers shared between each possible branch segment pair in the set of branch segments.

[0020] In some embodiments, a partition in the plurality of partitions comprises at least 1000 molecules with the common second portion, and each molecule in the at least 1000 molecules includes a primer sequence complementary to at least a portion of the larger contiguous nucleic acid.

[0021] In some embodiments, the identifier in the second portion of each respective sequence read in the set of sequence reads encodes a common value selected from the set {1, …, 1024}, the set {1, …, 4096}, the set {1, …, 16384}, the set {1, …, 65536}, the set {1, …, 262144}, the set {1, …, 1048576}, the set {1, …, 4194304}, the set {1, …, 16777216}, the set {1, …, 67108864}, or the set {1, …, 1×10^{12}}.

[0022] In some embodiments, the identifier is an N-mer, where N is an integer selected from the set {4, …, 20}.

[0023] In some embodiments, an average sequence read length of the plurality of sequence reads is between 40 bases and 200 bases or between 60 bases and 140 bases.

[0024] In some embodiments, the plurality of sequence reads collectively provide at least 15X coverage for the larger contiguous nucleic acid, more than ten percent of the k-mers in the plurality of k-mers are from more than one source sequence read in the plurality sequence reads, and the identifier of each such source sequence read for each k-mer represented by more than one source sequence read is retained.

[0025] In some embodiments, the plurality of sequence reads collectively provide at least 25X coverage for the larger contiguous nucleic acid, more than thirty percent of the plurality of k-mers are from more than one source sequence read in the plurality sequence reads, and the identifier of each such source sequence read for each k-mer represented by more than one source sequence read is retained.

[0026] In some embodiments, the sequence reads in the plurality of sequence reads encode between 75 and 125 bases of the larger contiguous nucleic acid and k is an odd integer between 5 and 73.

[0027] In some embodiments, a set of sequence reads in the plurality of sequence reads comprises more than 100 sequence reads, each with the same common second portion.

[0028] In some embodiments, the larger contiguous nucleic acid is a chromosome and/or is greater than 40 million base pairs in length.

BRIEF DESCRIPTION OF THE DRAWINGS

[0029] Figure 1A provides a schematic illustration of alignment of overlapping sequence reads.

[0030] Figure 1B shows a sample de Bruijn initial assembly graph based upon the sequence reads, with primary path identification.

[0031] Figure 2 shows a high level linkage map connecting primary paths in an overall assembly process in accordance with some embodiments.

[0032] Figure 3 schematically illustrates a barcode assisted assembly process in accordance with some embodiments.

[0033] Figure 4 schematically illustrates a process for genome assembly in accordance with some embodiments.

[0034] Figure 5 schematically illustrates aspects of the assembly processes in accordance with some embodiments.

[0035] Figures 6A, 6B, 6C, and 6D show de novo assembly of *E. coli* sequence data using the barcoded sequencing methods described herein.

[0036] Figure 7 is an example block diagram illustrating a computing device in accordance with some embodiments.

[0037] Figure 8 illustrates the relationship between the larger contiguous nucleic acid, the different fragments of the larger contiguous nucleic acid, and sequence reads of fragments in accordance with some embodiments.

[0038] Figure 9 illustrates the possible values of $h(W)$ for the sixty-four possible k-mers of length three in accordance with some embodiments.

[0039] Figure 10 illustrates the possible k-mers of length three generated from the sequence GATCCATCTT in accordance with some embodiments.

[0040] Figures 11A, 11B, 11C, 11D, and 11E collectively illustrate process flows for sequencing in accordance with some embodiments.

[0041] Figures 12A, 12B, 12C, 12D, 12E, 12F, and 12H collectively illustrate a method for obtaining a plurality of sequence reads, where the plurality of sequence reads comprises a plurality of sets of sequence reads, where each respective sequence read in a set of sequence reads includes (i) a first portion that corresponds to a subset of a larger contiguous nucleic acid and (ii) a common second portion that forms an identifier that is independent of the sequence of the larger contiguous nucleic acid and that identifies a partition, in a plurality of partitions, in which the respective sequence read was formed, in accordance with some embodiments.

[0042] Figure 13 illustrates a method for obtaining a plurality of sequence reads and pooling them together in accordance with some embodiments.

[0043] Figure 14 illustrates a de Bruijn graph in accordance with some embodiments.

[0044] Figure 15 illustrates another de Bruijn graph in accordance with some embodiments.

[0045] Like reference numerals refer to corresponding parts throughout the present disclosure.

DETAILED DESCRIPTION

[0046] The present disclosure generally provides methods, processes, and particularly computer implemented processes and non-transistory computer program products for use in the analysis of genetic sequence data. The processes and products described herein are especially useful in the assembly of shorter nucleic acid sequence data into longer linked and preferably contiguous genetic constructs, including large contigs, chromosomes and whole genomes.

[0047] The invention and various specific aspects and embodiments will be understood with reference to the following drawings and detailed descriptions. In some of the drawings and detailed descriptions below, the present invention is described in terms of an important independent embodiment of a system operating on a logic processing device, such as a computer system. This should not be taken to limit the invention, which, using the teachings provided herein, can be applied to any number of logic processors working together, whether incorporated into a stand alone computer system, an instrument system, or other information enabled devices or logic components incorporated into laboratory or diagnostic equipment or in functional communication therewith. For purposes of clarity, this discussion refers to devices, methods, and concepts in terms of specific examples. However, the invention and aspects thereof may have applications to a variety of types of devices and systems. It is therefore intended that the invention not be limited except as provided in the attached claims.

[0048] Furthermore, it is well known in the art that logic systems and methods such as described herein can include a variety of different components and different functions in a modular fashion. Different embodiments of the invention can include different mixtures of elements and functions and may group various functions as parts of various elements. For purposes of clarity, the invention is described in terms of systems that include many different innovative components and innovative combinations of innovative components and known components. No inference should be taken to limit the invention to combinations containing all of the innovative components listed in any illustrative embodiment in this specification. The functional aspects of the invention that are implemented on a computer or other logic processing systems or circuits, as will be understood from the teachings herein, may be implemented or accomplished using any appropriate implementation environment or programming language, such as C, C++, Cobol, Pascal, Java, Java-script, HTML, XML, dHTML, assembly or machine code programming, RTL, Python, etc. All

references, publications, patents, and patent applications cited herein are hereby incorporated by reference in their entirety for all purposes.

I. General

[0049] In general, assembly of nucleic acid sequence information into longer stretches of contiguous sequence typically involves the identification of overlapping sequences in different reads, and using those overlapping sequences to map these sequences against each other. Because genomic sequences are very large, the identification of overlapping sequences only allows for an inference of linkage between the two sequences. This inference is substantially strengthened with larger numbers of overlapping sequences in a given region of the overall sequence, e.g., higher coverage.

[0050] In some embodiments, this inference may be further strengthened when the sequences are mapped against a reference sequence. In particular, for a number of organisms, a reference sequence, e.g., a complete or nearly complete sequence of one individual's genome or a consensus of the organism's genome from multiple individuals, has been elucidated that provides a general framework of that organism's genomic sequence. Groups of overlapping sequences may then generally be mapped against a reference sequence for the given organism from which the DNA is being sequenced in order to provide greater confidence in the identified sequence linkage, as well as place the different sequence segments into the larger genomic context. However, in some embodiments of the present disclosure, reference sequences are not used.

[0051] In assembling the larger overall sequence, areas in the sequence that have insufficient information (e.g., insufficient coverage, lack of an unambiguous match to a reference sequence, etc.) can result in breaks in the overall inferred sequence. These breaks result in a number of separate contiguous sequence segments (or contigs) that can require substantial additional processing before they are eventually linked in a single contiguous sequence. These challenges, while significant in sequencing processes where reference sequences are available, are even more significant in de novo sequence assembly. By definition, de novo assemblies do not have the benefit of the secondary level of confirmation of sequence linkage that comes from a reference sequence. Instead, these sequences must be assembled from the bottom up, e.g., de novo.

[0052] In an example process, an aggregation of short sequence reads are obtained from any of a variety of different sequencing processes. The sequence reads are optionally stored in an

appropriate data structure for further evaluation. A computer algorithm analyzes these sequences to identify common sequence segments (k-mers) among subsets of the different sequence reads. Identification of k-mer subsequences is typically based upon the level of redundancy or coverage of the k-mer sequence of a desired length, which may generally be dependent upon the complexity of the genome being sequenced, the depth of sequencing coverage, and tolerances for varying error types. In many cases, k-mer length may be selected that are anywhere from about 7 to about 31 bases in length, or from about 15 to about 25 bases in length. For example, where a given k-mer is only represented less than 3, 4, 5, 10, 20 or more times, in the aggregated sequence database, it may be deemed insufficiently covered to meaningfully be used, depending upon the complexity of the sequence being analyzed, etc.

[0053] Sequence fragments that include an identified k-mer sequence are then aligned along that sequence of overlap between the sequences. Additional k-mer sequence or sequences may then be identified in these overlapping sequences to compile an aggregation of additional sequence segments to extend the contiguous sequence in an initial assembly graph using an additional assembly algorithm that extends from this k-mer to an overlapping k-mer to produce a full contig sequence, e.g., using a DeBruijn graph assembler such as Velvet, Euler or the like. In particular, the first set of reads having a first k-mer sequence may be used as an anchor for a repeated alignment exercise in which a second k-mer sequence common to one or more members of the first set, is used to extend the inferred sequence, i.e., to establish a connection between the first set of short sequences and a second set of short sequences. This is repeated so long as the sequence is able to be extended, unambiguously.

[0054] In genomic samples, however, k-mer sequences may be common to multiple disparate portions of the overall sequence, such that the identification of a subsequent k-mer sequence can implicate two or more different sets of sequence reads as extensions of a given sequence, where those two sequences are not actually related, e.g., one or more of those sequences is not the correct extension of the sequence. This results in a branch in the inferred connection between sequences, e.g., one of two or more sequences may constitute the continued sequence at this branch point. The maximal unbranched inferred sequence is generally termed a 'primary path'. Multiple primary paths are identified for a short read sequence dataset. Different primary paths may then be linked to each other through other processes, e.g., using paired end sequence data, overlaying long read sequence data, or the like.

[0055] Figures 1A and 1B provide a schematic illustration of the de novo assembly process described above. As shown, sequence fragments 102 are aligned according to included k-mer sequences that are represented at sufficiently high levels of coverage, e.g., k-mers 104, 106 and 108, as compared to the other k-mers that are represented at lower coverage, e.g., k-mers in alignments 120 and 122 k-mers. For example, in some cases, k-mers representing coverages of at least 3-4X would be used for alignment. Although illustrated as k-mers that are only eight bases in length, this is done for ease of illustration, and actual k-mer lengths may vary and may be longer, e.g., 10 mer or longer, 20mer or longer, 30mer or longer, 40mer or longer, 50mer or longer, 60mer or longer, 70mer or longer, 80mer or longer, 90mer or longer, 100mer or longer.

[0056] A series of overlapping k-mer sequences is identified, e.g., k-mers 104 and 106, to walk along the aggregate overlapping fragments and assemble a longer contiguous sequence. As shown in Figure 1B, where a given step leads to two or more overlapping k-mers, e.g., k-mers 110 and 112, it creates a branch point 114 in the contiguous sequence. Element 116 of Figure 1B illustrates another break point in the contiguous sequence The stretch of unbranched sequence between the two branch points is defined as a primary path 118. Although illustrated as a relatively short primary path, it will be appreciated that longer primary paths are generally generated, e.g., comprising 500 bases, 1 kb, 2 kb, 5 kb, or more in length.

[0057] Seed primary paths are then identified from the data. Seed primary paths, also referred to as seed paths, typically represent primary paths that are more likely to have a low copy number within the genome or other large sample nucleic acid (e.g., n approaching 1). This copy number may be inferred from the length of the sequence in the primary path, or from the overall read coverage of the primary paths, or both. From each seed path, neighboring or linked primary paths are then identified, either through an unambiguous overlap with the seed path, or through the use of other linkage information, e.g., long read sequencing data, or paired end sequence reads that can bridge two primary paths, e.g., one is represented in one primary path while the other end is represented in a different primary path. The linkage between two primary paths may be expressly determined, e.g., through sequencing of the intervening sequence, or it may be an inferred structural linkage where the relative position and connection of the two primary paths is inferred, e.g., from paired end sequence reads.

[0058] This is shown in Figure 2, which schematically illustrates a number of separate primary paths 202-210, where a structural linkage (shown as linkages 212) is determined as between the primary paths. As shown, in some cases, linkages will not be determinable.

[0059] The longest assembled contiguous sequence segments are then compiled (shown in Figure 2 as contigs 216 and 218). Typically, lengthy genome finishing processes, e.g., using laborious Sanger sequencing or other long read sequencing processes, e.g., SMRT® sequencing from Pacific Biosciences), are then used to close remaining gaps in the overall sequence, e.g., the gap between contigs 206 and 208.

[0060] As is clear from the above example, merely assembling a reasonable finished genome or reasonable number of contigs for a given genome, can involve multiple different sequencing processes, e.g., high throughput, short-read shotgun sequencing, paired-end long fragment sequencing and/or low throughput long-read DNA sequencing processes. Each of these processes can be costly and time consuming.

[0061] As described herein, however, an improved sequencing and assembly process allows one to obtain substantially more complete assembly from fewer or even a single, more efficient sequencing processes. In particular, in accordance with the processes described herein, structurally linked oligonucleotides that are sequencable using, e.g., short read sequencing techniques, are provided with common unique identifiers, e.g., barcodes, that may be used in the assembly process to infer linkage between different sequence segments and/or different primary paths, in the assembly process.

II. Linked-Read Sequencing

[0062] Barcoding of structurally linked short sequences and resulting sequence reads may be carried out by a number of methods. An exemplary process is described in detail in, for example, U.S. provisional Patent Application No. 61/940,318, filed February 7, 2014, U.S. provisional Patent Application No. 61/991,018, filed May 9, 2014, U.S. Patent Application No. 14/316,383, filed on June 26, 2014, as well as U.S. Patent Application No. 14/175,935, filed February 7, 2014, the full disclosures of which is hereby incorporated by reference in their entireties.

[0063] In brief, sequencing libraries with incorporated barcode sequences are prepared from one or more large nucleic acid sequences, e.g., a genome, chromosome or fragment thereof. This

large sample nucleic acid(s) is segmented into first fragments that will typically be on the order of 1 to about 100 kb or more in length. These long first fragments are then partitioned into separate reaction volumes. While these reaction volumes may be any of variety of different types of reaction vessels, e.g., tubes, wells, microwells, nanowells, or the like, in preferred aspects, the reaction volumes (also referred to as partitions), are comprised of droplets in an emulsion, e.g., a water-in-oil emulsion).

[0064] This partitioning may be carried out so as to ensure that at most, only a single first fragment is contained in any partition. However, such processes tend to have reduced throughput and higher costs, as the partitions are inefficiently used where most partitions would remain empty. As such, in preferred aspects, the partitioning is carried out under conditions in which there are multiple first fragments contained in individual partitions, but where there is a high probability that while multiple first fragments may be present within a given partition, such fragments will be structurally unique, e.g., representing structurally discrete and disparate regions of the larger sample nucleic acid. Typically, this is accomplished by providing the first fragments at a concentration that, based upon the size of the sample nucleic acid, the size of the first fragments, and the reaction volume, will statistically allocate structurally unique first fragments into each separate reaction volume. In particular, such partitioning will result in fewer than 10% of the partitions having structurally linked separate first fragments (overlapping, non-overlapping but immediately adjacent, or adjacent within 0.1 to 5 kb), fewer than 5% of the first fragments in a given partition being structurally linked, fewer than 1% of the first fragments in a given partition being structurally linked, and in some cases, fewer than 0.5%, 0.2%, or even fewer than 0.1% of the first fragments in a given partition being structurally linked.

[0065] Co-partitioned with the first fragments, e.g., within the same droplet, are collections of oligonucleotides that comprise a common barcode sequence within a given partition. The collection of oligonucleotides is taken from a vast and diverse library of different barcode sequences, such that different partitions have different barcode sequences contained therein. In preferred aspects, the oligonucleotides comprising the barcode sequences are provided into the individual partitions releasably attached to a bead or microcapsule, where the barcode sequences included within the oligonucleotides attached to a given bead are all substantially identical. Further, each bead will typically comprises more than 100,000, more than 1,000,000, more than 10,000,000 and in many cases, more than 50,000,000 individual oligonucleotide molecules. The

oligonucleotides may be drawn from a barcode library that includes greater than 1000, greater than 10,000, greater than 100,000, greater than 500,000, greater than 1 million, or even greater than 2 million different barcode sequences.

[0066] In preferred aspects, these oligonucleotides also include functional sequences, e.g., amplification primer sequences (which may be targeted to specific sequences, universal priming sequences or random priming sequences), or primer annealing sites, sequencing primer sequences or primer annealing sites, or the like. In preferred cases, universal or random primer sequences are included as at least one of the functional sequences within the oligonucleotides. These and other functional sequences are described in, for example, U.S. provisional Patent Application No. 61/940,318, filed February 7, 2014, U.S. provisional Patent Application No. 61/991,018, filed May 9, 2014, U.S. Patent Application No. 14/316,383, filed on June 26, 2014, as well as U.S. Patent Application No. 14/175,935, filed February 7, 2014, the full disclosures of which is hereby incorporated by reference in their entireties.

[0067] A population of second, sequencable shorter fragments, that constitute overlapping fragments of the first fragment, are then generated within the reaction volume, where the second short fragments include the common barcode sequence segment. In one exemplary process, this is carried out by releasing the oligonucleotides comprising the barcode sequences from the beads into the partitions. While the release of the oligonucleotides may be carried out using a variety of mechanisms to release the oligonucleotides from the beads, e.g., thermal cleavage, chemical cleavage, or photo-induced cleavage, in many cases, the release will be carried out by introducing a chemical cleaving agent into the partitions during the co-partitioning step that will either cleave a linkage between the oligonucleotide and the bead, or degrade the bead such that the oligonucleotide is released, or both.

[0068] Also co-partitioned with the first fragments and oligonucleotides are reagents for carrying out the amplification of different portions of the first fragments, including, for example, DNA polymerase enzyme, nucleoside triphosphates, divalent metal ions, e.g., Mn^{2+} or Mg^{2+}. Portions of the first fragments in a given partition are then replicated using the oligonucleotides, e.g., including a universal or random primer sequence, as primers for that replication, such that different portions of the first fragment are primed and replicated. In some cases, these first replicate fragments may be additionally primed by the oligonucleotides and replicated. As a result of these

replications, a collection of overlapping fragments of the first fragments in a given partition is created, where each includes a barcode sequence common to that partition.

[0069] In different partitions, different second fragments are generated from different first fragments in a similar fashion, with different barcode sequences appended to those second fragments. Across the collection of different partitions is then created a substantial sequencing library, where elements of that library that are created in the same partition, and thus may be derived from the same first fragment, will include the same barcode sequence.

[0070] The different partitions, representing the sequencing library, are then pooled and the fragments included therein are sequenced to identify the short second fragment sequences along with their appended barcode sequences. Because second fragments derived from a given first fragment will include a common barcode sequence, the presence of that barcode sequence in a given read, or set of reads provides additional linkage inferences that may be applied throughout the assembly process.

III. Application to Assembly

A. Generally

[0071] As described previously, a number of different parameters are used in assembly of shorter sequence reads into longer contiguous linked sequences. These include the presence of overlapping sequence segments between and among sequence reads that provide an inference that the sequence reads represent overlapping segments of a larger sequence context. The ability to provide longer linked shorter sequences, as well as the ability to provide linkage inferences as between larger sequence fragments from a given nucleic acid, e.g., the ability to provide fewer contigs or even a single unified sequence for the larger molecule, will depend to some extent upon the level and uniformity of sequencing coverage for the entire target nucleic acid, where regions with lower levels of sequencing coverage may result in gaps between contigs. As a result, sequencing processes that have more low coverage areas often result in more discontinuous sequence data, e.g., a larger number of contigs. While some of the coverage deficiencies can be overcome by brute force, e.g., sequencing a genome to much greater depth, that solution comes at a potentially large cost, in terms of sequencing process costs. However, in many cases, coverage deficiencies are systematic, e.g., resulting in part from the sequence itself.

[0072] By contrast, the assembly of barcoded sequencing libraries, as described herein, can provide enhanced linkage information both at the short sequence read level as well as in the higher level assembly, without necessarily requiring significant increases in coverage, or the use of ancillary additional sequencing processes.

[0073] In particular, in typical short sequence read assembly processes, short sequence reads that include overlapping sequence segments are aligned with each other to provide a larger contiguous sequence context, or contig. These contigs may then be subjected to further assembly with other contigs de novo, or by being mapped against a reference sequence. However, because these sequence reads are extremely short, e.g., 100-200 bases in length, in the context of genomes that are in the millions to billions of bases, the potential for any given read to map uniquely to a single locus within the genome is relatively low, and this can introduce significant ambiguity in the assembly process. This ambiguity may be lessened by sequencing large numbers of overlapping fragments to provide greater sequencing coverage at a give locus. However, low coverage areas of a genome will still provide difficulties in any assembly process, often leaving gaps between contigs. While lower coverage areas may be bridged by other methods as described elsewhere herein, e.g., long read sequencing, such methods add to the cost and inefficiency of the processes.

[0074] In the context of the methods described herein, however, short read sequences are provided with barcode sequences that provide additional inferences of structural relationship between two or more sequence reads without the requirement of as much sequencing coverage. In particular, aligned short read sequences may be assigned to an aligned contig based upon both an overlapping sequence with other sequences, as well as based upon their inclusion of a common barcode sequence. Using the additional structural linkage information of an associated common barcode sequence results in the identification of much longer contigs.

[0075] Likewise, bridging between two contigs may additionally be aided by the presence of barcode sequences. For example, where adjacent identified contigs include sequence reads that share overlapping barcode sequences, but are otherwise non-overlapping or insufficiently overlapping to align together, one may infer a structural linkage as between those two contigs. Briefly, the inclusion in two sequence reads provides a reasonably strong inference that the two reads were generated from fragments created within a single partition. When combined with additional linkage information, e.g., alignment of the sequence read to two or more contigs, or

where two barcodes straddle the same two contigs, it gives a significant indication of the linkage of those two contigs.

[0076] In terms of an exemplary computer implemented process, a collection of short read sequences that include associated barcode sequences, are provided within a data structure. Subsets of the sequences are aligned with each other based upon overlapping sequences within those sequences, and also based upon the association of such sequences with the same barcode sequence. Alignable sequences that do not include the same barcode sequence, or sequences that include the same barcode sequences but that are not alignable are not included, and may be aligned against different subsets of sequences. Based upon a set of aligned, commonly barcoded sequences, a linear contiguous sequence, or contig is generated. This is repeated with multiple different subsets of sequences, to provide multiple different contigs. As will be appreciated, the contigs may be longer than what would be achieved based upon sequence alignment alone, or may be provided with higher levels of confidence, despite being based upon lower sequencing coverage.

[0077] The generated contigs are then processed further to provide relative orientation and genomic context with respect to each other to establish a scaffold for the overall genomic (or other genomic component) sequence. In some cases, this may be achieved through mapping of the discrete contigs against a reference sequence. In many cases, however, mapping against a reference may not be unambiguous, e.g., based upon the presence of multiple duplicate sequence regions within the genome, or because a suitable reference sequence does not exist.

[0078] In such cases, barcoded sequences may additionally be used to provide linkage information as between contigs. For example, as noted above, in some cases, sequences that align to a first subset of sequences, but that do not share a common first barcode sequence, e.g., they include a second barcode sequence, would not be used in creating the first contig. However, such sequences may be aligned in creating a second contig. Likewise, sequences that include a first barcode, but that align to the second contig sequences may not have been used to generate the second contig. However, based upon such sequences being present in the data structure, one can ascertain a linkage between the first and second contig. In the context of the processes, contig linkage may be identified based upon the presence of a threshold level of shared barcodes as between sequences that align to different contigs, and/or based upon the presence of a threshold level of sequences that that align to the first contig but include barcodes associated with the second contig. As used herein, such threshold level may include a single sequence or, 2 sequences or more,

3 sequences or more, 4 sequences or more, 5 sequences or more, 10 sequences or more, or 20 sequences or more.

B. Application to de novo Assembly Processes

[0079] The barcoded sequencing libraries described above have particular advantages in de novo assembly. In particular, as noted above, the presence of a common barcode in two sequence reads can provide an inference of structural linkage between the two sequences, as there is an increased probability that these sequences were created within the same partition, and from the same first fragment. While this may not be definitive, as the same barcode sequence may be present in other partitions and in fragments of structurally unrelated first fragments, when combined with other indicators of structural relation, e.g., overlapping sequences, it becomes a powerful inference of structural linkage. By way of example, and with reference to the assembly process described above, Figure 3 schematically illustrates a potential impact of barcoded sequences on the assembly processes.

[0080] In one aspect, the barcodes included within the sequence reads may be used to provide a structural relationship between and among paths in an initial assembly in order to provide high level scaffolding of a sequence, e.g., providing a high confidence primary path scaffold in which the relative order of primary paths, and preferably, high confidence primary paths is identified and mapped, and against which final genomic assembly may be mapped. As will be appreciated, a high confidence primary path will generally refer to a primary path that, based upon its length and complexity, will be expected to have a low probability of duplication or multiple duplication within a genome, allowing a higher level of confidence that sequences or primary paths that connect to that primary path are, in fact, actually linked in the genome, as opposed to being linked to primary paths having a duplicate sequence. In preferred aspects, the high confidence primary paths are primary paths that will have a predicted copy number in the relevant genome approaching 1, e.g., less then 2, or equal to about 1.

[0081] In brief, a set of primary paths will have been identified from the initial assembly graph as described above, including high confidence primary paths (also referred to herein as "single copy number", "low copy number", or "CN1" paths). Each sequence read included within a given primary path will include an associated barcode sequence, as also described above. A list of barcodes associated with each primary path may then be generated for each primary path. Using the associated barcoded sequences, an initial sequence path may be plotted between adjacent CN1

primary paths, using those primary paths (both CN1 and otherwise) that share common barcode sequences. The path is then extended to the next adjacent CN1 path, providing ordering of the CN1 paths relative to each other. As will be appreciated, in many cases, duplicate barcode sequences may exist within the overall sequencing process and be attached to structurally distinct portions of a larger nucleic acid, e.g., chromosomal fragment, or genome fragment. However, because the assembly process focuses on low copy number or CN1 paths, utilizes highly diverse barcode libraries (e.g., greater than 1000, 5000, 10,000, 100,000 or even 1 million different barcode sequences), and is addressing very large genetic samples, it would be expected that the probability of the same barcode sequence being attached to similar or overlapping sequences that derive from structurally distinct, e.g., not structurally linked, parts of the genome or genomic fragment, would be extremely low, e.g., in many cases, less than 0.0001%.

[0082] This connection process is repeated among the CN1 paths until the order of all of the CN1 paths is established. The CN1 paths may then be used as a scaffold for completing assembly of the sequence between the paths. This is generally accomplished in a similar fashion, by using the sequence reads that include barcodes common to the CN1 paths to create a local sequence assembly graph between adjacent CN1 paths.

[0083] In particular, sequence reads including the same barcode sequences as a given primary path are used to stitch together a local sequence assembly graph between two adjacent primary paths in the scaffold, in the same fashion as described above for assembly of the initial assembly graph. This assembly map provides the sequence for the space between the two primary paths, closing the sequence gap. By utilizing barcode information associated with each sequence in which the k-mer exists, in addition to the overlapping sequence data in the k-mers, one can unambiguously identify the sequence extending between two primary paths. This is schematically illustrated in Figure 3. As shown, a series of k-mer sequences 302-316, similar to those schematically illustrated for the global assembly graph in Figure 1A are illustrated, except where each k-mer is coupled to an incorporated barcode sequence 318, e.g., barcode sequence “1”, barcode sequence “2” or barcode sequence “n”. As with the initial assembly graph, the presence of diverging kmer sequences at a given point or points could give rise to branch points within the sequence, e.g., branch points 320 and 322. However, by relying in part on the barcode information associated with each k-mer sequence, one can effectively identify the correct sequence path, eliminating the branch point. In particular, with reference to branch point 322, while both of k-mers

314 and 316 represent possible next sequence steps, because only the path associated with k-mer 314 includes the common barcode "1", it is identified as the next sequence step.

C. Exemplary Embodiments

[0084] Figure 7 is a block diagram illustrating a sequencing system 100 in accordance with some implementations. The device 100 in some implementations includes one or more processing units CPU(s) 702 (also referred to as processors), one or more network interfaces 704, a user interface 706, a memory 712, and one or more communication buses 714 for interconnecting these components. The communication buses 714 optionally include circuitry (sometimes called a chipset) that interconnects and controls communications between system components. The memory 712 typically includes high-speed random access memory, such as DRAM, SRAM, DDR RAM, ROM, EEPROM, flash memory, CD-ROM, digital versatile disks (DVD) or other optical storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, other random access solid state memory devices, or any other medium which can be used to store desired information; and optionally includes non-volatile memory, such as one or more magnetic disk storage devices, optical disk storage devices, flash memory devices, or other non-volatile solid state storage devices. The memory 712 optionally includes one or more storage devices remotely located from the CPU(s) 702. The memory 712, or alternatively the non-volatile memory device(s) within the memory 712, comprises a non-transitory computer readable storage medium. In some implementations, the memory 712 or alternatively the non-transitory computer readable storage medium stores the following programs, modules and data structures, or a subset thereof:

- an operating system 716, which includes procedures for handling various basic system services and for performing hardware dependent tasks;
- a network communication module (or instructions) 718 for connecting the device 700 with other devices, or a communication network;
- a sequence read processing module 720 for processing sequence reads;
- an optional representation of a larger contiguous nucleic acid 726 obtained from a biological sample;
- a plurality of sequence reads 728, each respective sequence read in the plurality of sequence reads comprising at least a first portion 730 that corresponds to a subset of the larger contiguous nucleic acid and a common second portion 732 that forms an identifier that is

independent of the sequence of the larger contiguous nucleic acid and that identifies a partition, in a plurality of partitions, in which the respective sequence read was formed; and

- a hash table 740 comprising, for each respective sequence read 728, the associated second portion 744 (bar code) and k-mers obtained by a hash of the respective sequence read 728 according to a predetermined k-mer length L

[0085] In some implementations, the user interface 706 includes an input device (e.g., a keyboard, a mouse, a touchpad, a track pad, and/or a touch screen) 710 for a user to interact with the system 700 and a display 708.

[0086] In some implementations, one or more of the above identified elements are stored in one or more of the previously mentioned memory devices, and correspond to a set of instructions for performing a function described above. The above identified modules or programs (e.g., sets of instructions) need not be implemented as separate software programs, procedures or modules, and thus various subsets of these modules may be combined or otherwise re-arranged in various implementations. In some implementations, the memory 712 optionally stores a subset of the modules and data structures identified above. Furthermore, in some embodiments, the memory stores additional modules and data structures not described above.

[0087] Although Figure 7 shows a "sequence system 700," Figure 7 is intended more as functional description of the various features which may be present in computer systems than as a structural schematic of the implementations described herein. In practice, and as recognized by those of ordinary skill in the art, items shown separately could be combined and some items could be separated.

[0088] Figure 8 illustrates the relationship between the larger contiguous nucleic acid 802, the different fragments 804 of the larger contiguous nucleic acid, and sequence reads 728 of fragments. Typically, between 1 and 250 fragments 804, between 5 and 500 fragments 804 or between 10 and 1000 fragments 804 are partitioned into a separate partition. In any event, sufficiently few of the fragments 804 are partitioned into the same partition such that the chance that the fragments 804 in a single partition have any appreciable overlapping sequences is unlikely. Sequence reads 728 of each fragment 804 are made. In typical embodiments, sequence reads 728 are short in length (e.g., less than 1000 bases) so that they can be sequenced in automated sequencers. Each sequence read 728 in a partition includes a common second portion 732 that

forms an identifier that is independent of the sequence of the larger contiguous nucleic 802 acid nucleic acid and that identifies the partition, in a plurality of partitions, in which the respective sequence read was formed.

[0089] In some embodiments, a respective set of k-mers for each sequence read 728 in the plurality of sequence reads is constructed. In such embodiments the identifier (barcode 732) of the set of sequence reads that includes the sequence read is retained for each k-mer in the plurality of k-mers. K-mers are shorter than sequence reads. In some embodiments a sequence read is between 50 and 250 bases whereas a k-mer has a value of k (meaning that it has a nucleic acid length of k) that is between 7 and 51. Each respective set of k-mers includes at least eighty percent, at least eight-five percent, at least ninety percent, at least ninety-five percent of the possible k-mers of the corresponding sequence read.

[0090] Figure 11 is a flow chart illustrating a method (1100) of sequencing a larger contiguous nucleic acid (e.g., obtained from a biological sample) (1102), in accordance with some embodiments. The method is performed at a computer system having one or more processors, and memory storing one or more programs for execution by the one or more processors (1104).

[0091] *Obtaining a plurality of sequence reads.* In accordance with the disclosed systems and methods, a plurality of sequence reads 728 is obtained (1106). The plurality of sequence reads comprises a plurality of sets of sequence reads. Each respective sequence read 728 in a set of sequence reads comprises a first portion (130) that corresponds to a subset of the larger contiguous nucleic acid and a common second portion (732) that forms an identifier that is independent of the sequence of the larger contiguous nucleic acid and that identifies a partition, in a plurality of partitions, in which the respective sequence read was formed. In other words, the identifier is not derived from, or a function of the sequencing data of the larger contiguous nucleic acid.

[0092] In some embodiments, as illustrated in Figure 8, to obtain the plurality of sequence reads 728, a larger contiguous nucleic acid 802 is fragmented to form fragments 804 and these fragments are compartmentalized, or partitioned into discrete compartments or partitions (referred to interchangeably herein as partitions). In some embodiments, more than 10, more than 100, more than 1000, more than 10,000, more than 100,000, more than 1×10^6, or more than 5×10^6 sets of sequence reads are obtained, corresponding more than 10, more than 100, more than 1000, more than 10,000, more than 100,000, more than 1×10^6, or more than 5×10^6 partitions.

[0093] Each partition maintains separation of its own contents from the contents of other partitions. As used herein, the partitions refer to containers or vessels that may include a variety of different forms, e.g., wells, tubes, micro or nanowells, through holes, or the like. In preferred aspects, however, the partitions are flowable within fluid streams. In some embodiments, these vessels are comprised of, e.g., microcapsules or micro-vesicles that have an outer barrier surrounding an inner fluid center or core, or have a porous matrix that is capable of entraining and/or retaining materials within its matrix. In a preferred aspect, however, these partitions comprise droplets of aqueous fluid within a non-aqueous continuous phase, e.g., an oil phase. A variety of different suitable vessels are described in, for example, U.S. Patent Application No. 13/966,150, filed August 13, 2013, which is hereby incorporated by reference herein in its entirety. Likewise, emulsion systems for creating stable droplets in non-aqueous or oil continuous phases are described in detail in, e.g., Published U.S. Patent Application No. 2010-0105112, which is hereby incorporated by reference herein in its entirety. In certain embodiments, microfluidic channel networks are particularly suited for generating partitions. Examples of such microfluidic devices include those described in detail in Provisional U.S. Patent Application No. 61/977,804, filed April 4, 2014, the full disclosure of which is incorporated herein by reference in its entirety for all purposes. Alternative mechanisms may also be employed in the partitioning of individual cells, including porous membranes through which aqueous mixtures of cells are extruded into non-aqueous fluids. Such systems are generally available from, e.g., Nanomi, Inc.

[0094] In the case of droplets in an emulsion, partitioning the fragments 802 into discrete partitions may generally be accomplished by flowing an aqueous, sample containing stream, into a junction into which is also flowing a non-aqueous stream of partitioning fluid, e.g., a fluorinated oil, such that aqueous droplets are created within the flowing stream partitioning fluid, where such droplets include the sample materials. As described below, the partitions, e.g., droplets, also typically include co-partitioned barcode oligonucleotides.

[0095] The relative amount of sample materials within any particular partition may be adjusted by controlling a variety of different parameters of the system, including, for example, the concentration of fragments 804 in the aqueous stream, the flow rate of the aqueous stream and/or the non-aqueous stream, and the like. The partitions described herein are often characterized by having overall volumes that are less than 1000 pL, less than 900 pL, less than 800 pL, less than 700 pL, less than 600 pL, less than 500 pL, less than 400pL, less than 300 pL, less than 200 pL, less

than 100pL, less than 50 pL, less than 20 pL, less than 10 pL, or even less than 1 pL. Where co-partitioned with beads, it will be appreciated that the sample fluid volume within the partitions may be less than 90% of the above described volumes, less than 80%, less than 70%, less than 60%, less than 50%, less than 40%, less than 30%, less than 20%, or even less than 10% of the above described volumes. In some cases, the use of low reaction volume partitions is particularly advantageous in performing reactions with small amounts of starting reagents, e.g., input larger contiguous nucleic acid fragments. Methods and systems for analyzing samples with low input nucleic acids are presented in U.S. Provisional Patent Application No. 62/017,580 June 26, 2014, the full disclosure of which is hereby incorporated by reference in its entirety.

[0096] Once the fragments 804 are introduced into their respective partitions, the fragments 804 within partitions are generally provided with unique identifiers such that, upon characterization of those fragments 804, they may be attributed as having been derived from their respective partitions. In some embodiments, such unique identifiers are previously, subsequently or concurrently delivered to the partitions that hold the compartmentalized or partitioned fragments 804, in order to allow for the later attribution of the characteristics, e.g., nucleic acid sequence information, to the sample nucleic acids included within a particular compartment, and particularly to relatively long stretches of contiguous sample nucleic acids that may be originally deposited into the partitions.

[0097] Accordingly, the fragments 804 are typically co-partitioned with the unique identifiers 732 (e.g., barcode sequences). In particularly preferred aspects, the unique identifiers 732 are provided in the form of oligonucleotides that comprise nucleic acid barcode sequences 732. The oligonucleotides 732 are partitioned such that as between oligonucleotides in a given partition, the nucleic acid barcode sequences 732 contained therein are the same, but as between different partitions, the oligonucleotides can, and preferably have differing barcode sequences. In preferred embodiments, only one nucleic acid barcode sequence 732 is associated with a given partition, although in some embodiments, two or more different barcode sequences are present in a given partition.

[0098] The larger contiguous nucleic acid 802 is typically fragmented and partitioned such that the fragments 804 of the larger contiguous nucleic acid 802 in the partitions are long fragments or stretches of contiguous nucleic acid molecules. As illustrated in Figure 8, these fragments 804

typically represent a number of overlapping fragments of the overall larger contiguous nucleic acid to be analyzed, e.g., an entire chromosome, exome, or other large genomic fragment.

[0099] The larger contiguous nucleic acid 802 is also typically fragmented and partitioned at a level whereby a given partition has a very low probability of including two overlapping fragments 804 of the starting larger contiguous nucleic acid 802. This is typically accomplished by providing the larger contiguous nucleic acid 802 at a low input amount and/or concentration during the partitioning process. As a result, in preferred cases, a given partition includes a number of long, but non-overlapping fragments 804 of the starting larger contiguous nucleic acid. The nucleic acid fragments in the different partitions are then associated with unique identifiers 732 where, for any given partition, fragments 804 contained therein possess the same unique identifier, but where different partitions include different unique identifiers. Moreover, because the partitioning step allocates the sample components into very small volume partitions or droplets, it will be appreciated that in order to achieve the desired allocation as set forth above, one need not conduct substantial dilution of the sample, as would be required in higher volume processes, e.g., in tubes, or wells of a multiwell plate. Further, because the systems described herein employ such high levels of barcode diversity, one can allocate diverse barcodes among higher numbers of genomic equivalents, as provided above. In some embodiments, in excess of 10,000, 100,000, 500,000, etc. diverse barcode types are used to achieve genome:(barcode type) ratios that are on the order of 1:50 or less, 1:100 or less, 1:1000 or less, or even smaller ratios, while also allowing for loading higher numbers of genomes (e.g., on the order of greater than 100 genomes per assay, greater than 500 genomes per assay, 1000 genomes per assay, or even more) while still providing for far improved barcode diversity per genome. Here, each such genome is an example of a larger contiguous nucleic acid.

[00100] Referring to Figure 12A, panel B, often the above-described partitioning is performed by combining the sample containing the larger contiguous nucleic acid 802 with a set of oligonucleotide tags 1202 (containing the barcodes 732) that are releasably-attached to beads prior to the partitioning step. The oligonucleotides 1202 may comprise at least a primer region 1210 and a barcode 732 region. Between oligonucleotides 1202 within a given partition, the barcode region 732 is substantially the same barcode sequence, but as between different partitions, the barcode region 732 in most cases is a different barcode sequence. In some embodiments, the primer region includes (or further comprises) an N-mer 1216 (either a random N-mer or an N-mer designed to target a particular sequence) that is used to prime the fragments 804 within the partitions. In some

cases, where the N-mer is designed to target a particular nucleic acid sequence, the primer region 1210 is designed to target a particular chromosome (e.g., human chromosome 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, X, or Y), or region of a chromosome, e.g., an exome or other targeted region. In some cases, the N-mer is designed to target a particular gene or genetic region, such as a gene or region associated with a disease or disorder (e.g., cancer). In some cases, the N-mer is designed to target a particular structural variation. Within the partitions, an amplification reaction is conducted using the primer sequence 1210/1216 (e.g. N-mer) to prime the fragments 804 at different places along their lengths. As a result of the amplification, each partition contains amplified products of the fragments 804, termed sequence reads 728, that comprise an identical or near-identical (common) barcode 732, and that represent overlapping, smaller fragments of the fragments 804 in each partition. The barcode 732 therefore serves as a marker that signifies that a set of sequence reads 728 originated from the same partition, and thus potentially also originated from the same fragment 804. It will be appreciated that there are typically several fragments 804 in any given partition. Nevertheless, in typical embodiments, fragments 804 that are in the same partition typically do not have any significant overlap and so it is possible to localize the amplified sequence reads 728 to the correct fragment 804 in any given partition. Following amplification, the sequence reads 728 are pooled, sequenced, and aligned using a sequencing algorithm. Because shorter sequence reads 728 may, by virtue of their associated barcode sequences 732, be aligned and attributed to a single, long fragment 804 of the larger contiguous nucleic acid 802, all of the identified variants on that sequence can be attributed to a single originating fragment 802 and single originating chromosome of the larger contiguous nucleic acid. Further, by aligning multiple co-located variants across multiple fragments 804, one can further characterize that chromosomal contribution. Accordingly, conclusions regarding the phasing of particular genetic variants may then be drawn. Such information may be useful for identifying haplotypes, which are generally a specified set of genetic variants that reside on the same nucleic acid strand or on different nucleic acid strands. Moreover, additionally or alternatively, structural variants are identified.

[00101] In some embodiments, referring to Figure 12A, the co-partitioned oligonucleotides 1202 also comprise functional sequences in addition to the barcode sequence 732 and the primer region sequence 1210/1216. For instance, in some embodiments, the co-partitioned oligonucleotides 1202 also comprise other functional sequences useful in the processing of the partitioned nucleic acids such as targeted or random/universal amplification primer sequences for

amplifying the fragments 804 within the partitions while attaching the associated barcode sequences, sequencing primers, hybridization or probing sequences, e.g., for identification of presence of the sequences, or for pulling down barcoded nucleic acids, or any of a number of other potential functional sequences. See, for example, the disclosure on co-partitioning of oligonucleotides and associated barcodes and other functional sequences, along with sample materials as described in, for example, U.S. Patent Application Nos. 61/940,318, filed February 7, 2014, 61/991,018, Filed May 9, 2014, and U.S. Patent Application No. 14/316,383, filed on June 26, 2014, as well as U.S. Patent Application No. 14/175,935, filed February 7, 2014, the full disclosures of which is hereby incorporated by reference in their entireties.

[00102] In one exemplary process, beads are provided that each includes large numbers of the above described oligonucleotides 1202 releasably attached to the beads, where all of the oligonucleotides attached to a particular bead include the same nucleic acid barcode sequence 732, but where a large number of diverse barcode sequences are represented across the population of beads used. Typically, the population of beads provides a diverse barcode sequence library that includes at least 1000 different barcode sequences, at least 10,000 different barcode sequences, at least 100,000 different barcode sequences, or in some cases, at least 1,000,000 different barcode sequences. Additionally, each bead typically is provided with large numbers of oligonucleotide molecules 1202 attached. In particular, the number of molecules of oligonucleotides 1202 including the barcode sequence 732 on an individual bead may be at least about 10,000 oligonucleotides, at least 100,000 oligonucleotide molecules, at least 1,000,000 oligonucleotide molecules, at least 100,000,000 oligonucleotide molecules, and in some cases at least 1 billion oligonucleotide molecules.

[00103] In some embodiments, the oligonucleotides are releasable from the beads upon the application of a particular stimulus to the beads. In some cases, the stimulus is a photo-stimulus, e.g., through cleavage of a photo-labile linkage that may release the oligonucleotides. In some cases, a thermal stimulus is used, where elevation of the temperature of the beads environment results in cleavage of a linkage or other release of the oligonucleotides form the beads. In some cases, a chemical stimulus is used that cleaves a linkage of the oligonucleotides to the beads, or otherwise results in release of the oligonucleotides from the beads.

[00104] In some embodiments, the beads including the attached oligonucleotides 1202 are co-partitioned with the individual samples, such that a single bead and a single sample are contained

within an individual partition. In some cases, where single bead partitions are desired, it may be desirable to control the relative flow rates of the fluids such that, on average, the partitions contain less than one bead per partition, in order to ensure that those partitions that are occupied, are primarily singly occupied. Likewise, in some embodiments, the flow rate is controlled to provide that a higher percentage of partitions are occupied, e.g., allowing for only a small percentage of unoccupied partitions. In preferred aspects, the flows and channel architectures are controlled as to ensure a desired number of singly occupied partitions, less than a certain level of unoccupied partitions and less than a certain level of multiply occupied partitions.

[00105] Figure 3 of United States Patent Application No. 62/072,214, filed October 29, 2014, entitled "Analysis of Nucleic Acid Sequences," which is hereby incorporated by reference and the portions of the specification describing Figure 3 provide a detailed example of one method for barcoding and subsequently sequencing a larger contiguous nucleic acid (referred to in the reference as a "sample nucleic acid") in accordance with one embodiment of the present disclosure. As noted above, while single bead occupancy may be the most desired state, it will be appreciated that multiply occupied partitions, or unoccupied partitions may often be present. Figure 4 of United States Patent Application No. 62/072,214, filed October 29, 2014, entitled "Analysis of Nucleic Acid Sequences," which is hereby incorporated by reference, and the portions of the specification describing Figure 4 provide a detailed example of a microfluidic channel structure for co-partitioning samples and beads comprising barcode oligonucleotides in accordance with one embodiment of the present disclosure.

[00106] Once co-partitioned, the oligonucleotides 1202 disposed upon the bead are used to barcode and amplify the partitioned samples. One process for use of these barcoded oligonucleotides 1202 in amplifying and barcoding samples is described in detail in U.S. Patent Application Nos. 61/940,318, filed February 7, 2014, 61/991,018, filed May 9, 2014, and 14/316,383, (Attorney Docket No. 43487-708.201) filed on June 26, 2014, the full disclosures of which are hereby incorporated by reference in their entireties. Briefly, in one aspect, the oligonucleotides present on the beads that are co-partitioned with the samples are released from their beads into the partition with the samples. The oligonucleotides typically include, along with the barcode sequence 732, a primer sequence at its 5' end 1216. In some embodiments, this primer sequence is a random oligonucleotide sequence intended to randomly prime numerous different

regions of the samples. In some embodiments the primer sequence 1216 is a specific primer sequence targeted to prime upstream of a specific targeted region of the sample.

[00107] Once released, the primer portion 1216 of the oligonucleotide 1202 anneals to a complementary region of fragments 804 in the partition. Extension reaction reagents, e.g., DNA polymerase, nucleoside triphosphates, co-factors (e.g., Mg^{2+} or Mn^{2+} etc.), that are also co-partitioned with the fragments 804 and beads, extend the primer sequence using the fragments 804 as a template, to produce a complementary sequence to the strand of the fragment 804 to which the primer annealed, and this complementary sequence includes the oligonucleotide 1202 and its associated barcode sequence 732. Annealing and extension of multiple primers to different portions of the fragments 804 in the partition may result in a large pool of overlapping complementary portions of the fragments 804, each possessing its own barcode sequence 732 indicative of the partition in which it was created. In some cases, these complementary portions may themselves be used as a template primed by the oligonucleotides 1202 present in the partition to produce a complement of the complement that again, includes the barcode sequence 732. In some cases, this replication process is configured such that when the first complement is duplicated, it produces two complementary sequences at or near its termini, to allow the formation of a hairpin structure or partial hairpin structure that reduces the ability of the molecule to be the basis for producing further iterative copies. A schematic illustration of one example of this is shown in Figure 12.

[00108] As Figure 12A shows, oligonucleotides 1202 that include a barcode sequence 732 are co-partitioned in, e.g., a droplet 1204 in an emulsion, along with a sample fragment 306. In some embodiments, the oligonucleotides 302 are provided on a bead 1208 that is co-partitioned with the larger contiguous nucleic acid fragment 804. The oligonucleotides 1202 are preferably releasable from the bead 1208, as shown in Figure 12A, panel (A). As shown in Figure 12A panel (B), the oligonucleotides 1202 includes a barcode sequence 732, in addition to one or more functional sequences, e.g., sequences 1212, 732, 1210, and 1216. For example, the oligonucleotide 1202 is shown as further comprising sequence 1212 that may function as an attachment or immobilization sequence for a given sequencing system, e.g., a P5 sequence used for attachment in flow cells of an ILLUMINA, HISEQ or MISEQ system. In other words, attachment sequence 1212 is used to reversibly attach oligonucleotide 1202 to a bead 1208 in some embodiments. As shown in Figure 12A, panel B, the oligonucleotide 1202 also includes a primer sequence 1210, which may include (or further comprise) a random or targeted N-mer 1216 (discussed above) for priming

replication of portions of the fragment 804. Also included within exemplary oligonucleotide 1202 of Figure 12A, panel B, is a sequence 1210 which may provide a sequencing priming region, such as a "read1" or R1 priming region, that is used to prime polymerase mediated, template directed sequencing by synthesis reactions in sequencing systems. In many cases, the barcode sequence 732, and immobilization sequence 1212 may be common to all of the oligonucleotides 1202 attached to a given bead. The primer sequence 1210/1216 may vary for random N-mer primers, or may be common to the oligonucleotides on a given bead for certain targeted applications.

[00109] Referring to Figure 12B, based upon the presence of primer sequence 1216, the oligonucleotides 1202a and 1202b are able to prime the fragment 804, which allows for extension of the oligonucleotides 1202a and 1202b using polymerase enzymes and other extension reagents also co-portioned with the bead 1208 and fragment 804.

[00110] As shown in Figure 12C, following extension of the oligonucleotides that, for random N-mer primers, would anneal to multiple different regions of the fragment 804; multiple overlapping complements of the fragment 804 are created, e.g., sequence reads 728-1 and 728-2. As such, Figure 12C illustrates (A) obtaining a plurality of sequence reads, where each respective sequence read 728 in the plurality of sequence reads comprises a first portion 730 that corresponds to a subset of the larger contiguous nucleic acid 802 and a common second portion 1250 that that forms an identifier 732 that is independent of the sequence of the larger contiguous nucleic acid and that identifies a partition, in a plurality of partitions, in which the respective sequence read 728 was formed (e.g., bar code 732).

[00111] The barcoded sequence reads 728 of Figure 12B may then be subjected to characterization, e.g., through sequence analysis, or they may be further amplified in the process, as shown in Figure 12D. For example, additional oligonucleotides, e.g., oligonucleotide 1202c, also released from bead 1208, may prime the fragment 1202b. In particular, again, based upon the presence of the random N-mer primer 1216 in oligonucleotide 1202c (which in many cases will be different from other random N-mers in a given partition) the oligonucleotide anneals with the fragment 1202b, and is extended to create a complement 728-3 to at least a portion of fragment 1202b which comprises a duplicate of a portion of the larger contiguous nucleic acid sequence. Extension of the oligonucleotide 1202b continues until it has replicated through the oligonucleotide portion 730 of fragment 1202b. As illustrated in Figure 12D, the oligonucleotides may be configured to promptly stop in the replication by the polymerase at a desired point, e.g., after

replicating through sequences 1216 and 1210 of oligonucleotide 1202b. As described herein, this may be accomplished by different methods, including, for example, the incorporation of different nucleotides and/or nucleotide analogues that are not capable of being processed by the polymerase enzyme used. For example, this may include the inclusion of uracil containing nucleotides within the sequence region 1210 to prevent a non-uracil tolerant polymerase to cease replication of that region. As a result, referring to Figure 12E, a sequence read 728-3 is created that includes the full-length oligonucleotide 1202 at one end, including the barcode sequence 732, the attachment sequence 1212, the primer region 1210, and the random N-mer sequence 1216. At the other end of the sequence is included the complement 1216' to the random N-mer of the first oligonucleotide 1202, as well as a complement to all or a portion of the primer sequence, shown as sequence 1210'. The R1 sequence 1210 and its complement 1210' are then able to hybridize together to form a partial hairpin structure 1260. As will be appreciated, because the random N-mers differ among different oligonucleotides, these sequences and their complements would not be expected to participate in hairpin formation, e.g., sequence 1216', which is the complement to random N-mer 1216, would not be expected to be complementary to random N-mer sequence 1216b. This would not be the case for other applications, e.g., targeted primers, where the N-mers would be common among oligonucleotides within a given partition.

[00112] By forming these partial hairpin structures, it allows for the removal of first level duplicates of the sample sequence from further replication, e.g., preventing iterative copying of copies. The partial hairpin structure also provides a useful structure for subsequent processing of the created fragments, e.g., fragment 730-3.

[00113] All of the sequence reads 728 from multiple different partitions may then be pooled for sequencing on high throughput sequencers as described herein. Because each sequence read 728 is coded as to its partition of origin, the sequence of that sequence read 728 may be attributed back to its origin based upon the presence of the barcode 732.

[00114] This is schematically illustrated in Figure 13. As shown in one example, a fragment 804-1 and a fragment 804-2 are each partitioned along with their own sets of barcode oligonucleotides 1202 as described above. Within each partition, each fragment (804-1 and 804-2) is then processed to separately to provide overlapping sequence reads 728 of the fragments 804-1 and 804-2, to form set of sequence reads 1302-1 and 1302-2. This processing provides sequence reads 728 with a barcode sequence 732 that is the same for each of the sequence reads 728 derived

from a particular first fragment 804. As shown, the set of sequence reads 1302-1 is denoted by "1" while the set of sequence reads 1302-2 is denoted by "2". A diverse library of barcodes may be used to differentially barcode large numbers of different sets of sequence reads. However, it is not necessary for every sequence read in a given partition to be barcoded with different barcode sequence. In fact, in many cases, multiple different fragments 804 may be processed concurrently in the same partition to include the same barcode sequence 732. Diverse barcode libraries are described in detail elsewhere herein.

[00115] The sets of sequence reads may then be pooled for sequencing using, for example, sequence by synthesis technologies available from Illumina or Ion Torrent division of Thermo Fisher, Inc. Once sequenced, the sequence reads 728 can be attributed to their respective fragment 804 set, e.g., as shown in aggregated reads 1314 and 1316, at least in part based upon the included barcodes, and optionally, and preferably, in part based upon the sequence of the fragment itself. The attributed sequence reads for each fragment set are then assembled to provide the assembled sequence for each sample fragment, e.g., sequences 1318 and 1320, which in turn, may be further attributed back to their respective original fragments 804. Methods and systems for assembling genomic sequences are described in, for example, U.S. Provisional Patent Application No. 62/017,589 (Attorney Docket No. 43487-729.101), filed June 26, 2014, the full disclosure of which is hereby incorporated by reference in its entirety.

[00116] Referring to Figure 11A (1108), in some embodiments, a partition in the plurality of partitions comprises at least 10 molecules, at least 100 molecules, at least 1000 molecules, at least 1 x 10^4 molecules, at least 1 x 10^5 molecules, at least 1 x 10^6 molecules, at least 1 x 10^7 molecules, or at least 1 x 10^8 molecules with the common second portion 732. In other words, the partition includes at least at least 10 molecules, at least 100 molecules, at least 1000 molecules, at least 1 x 10^4 molecules, at least 1 x 10^5 molecules, at least 1 x 10^6 molecules, at least 1 x 10^7 molecules, or at least 1 x 10^8 molecules 1202 (Figure 12A), and each such molecule 1202 further comprises a primer sequence 1210 complementary to at least a portion of the larger contiguous nucleic acid 802.

[00117] Referring to Figure 11A (1110), in some embodiments, a partition in the plurality of partitions comprises at least 10 molecules, at least 100 molecules, at least 1000 molecules, at least 1 x 10^4 molecules, at least 1 x 10^5 molecules, at least 1 x 10^6 molecules, at least 1 x 10^7 molecules, or at least 1 x 10^8 molecules with the common second portion 732, and each such molecule 1202

further comprises a primer site 1210 and a semi-random N-mer priming sequence 1216 that is complementary to part of the larger contiguous nucleic acid (1110).

[00118] Referring to Figure 11A (1112), in some embodiments, each fragment 804 in the one or more fragments of the larger contiguous nucleic acid 802 in a partition in the plurality of partitions is greater than 1 kilobase in length, is greater than 5 kilobases in length, is greater than 50 kilobases in length, is greater than 100 kilobases in length, is greater than 200 kilobases in length, is greater than 250 kilobases in length, is greater than 300 kilobases in length, or is greater than 350 kilobases in length. Typically, the fragments 804 of the larger contiguous nucleic acid 802 that are partitioned into partitions are longer than 1 kb, longer than 5 kb, longer than 10 kb, longer than 15 kb, longer than 20 kb, longer than 30 kb, longer than 40 kb, longer than 50 kb, longer than 60 kb, longer than 70 kb, longer than 80 kb, longer than 90 kb or even longer than 100 kb. In some embodiments the one or fragments in a partition in the plurality of partitions have an N50 (where the sum of the fragment 804 lengths that are greater than the stated N50 number is 50% of the sum of all fragment 804 lengths) of at least about 10kb, at least about 20kb, at least about 50kb, at least about 100kb, at least about 150kb, at least about 200kb, at least about 250kb, at least about 300 kb, at least about 350 kb, at least about 400 kb, at least about 500 kb, in excess of 1 Mb, or even in excess of 2 Mb. In some embodiments the one or fragments in each partition across the plurality of partitions have an overall N50 (where the sum of the fragment 804 lengths that are greater than the stated N50 number is 50% of the sum of all fragment 804 lengths) of at least about 10kb, at least about 20kb, at least about 50kb, at least about 100kb, at least about 150kb, at least about 200kb, at least about 250kb, at least about 300 kb, at least about 350 kb, at least about 400 kb, at least about 500 kb, in excess of 1 Mb, or even in excess of 2 Mb. In some embodiments, the one or more fragments 804 of the larger contiguous nucleic acid 802 in a partition in the plurality of partitions are between 20 kilobases and 200 kilobases in length (1114). In some embodiments, the one or more fragments of the larger contiguous nucleic acid in a partition in the plurality of partitions are between 20 kilobases and 200 kilobases in length (1116).

[00119] Referring to Figure 11B (1118), in some embodiments, the one or more fragments 804 of the larger contiguous nucleic acid in a partition in the plurality of partitions consists of between 1 and 500 different fragments, between 10 and 1000 different fragments, between 100 and 1000 different fragments, between 25 and 150 different fragments, between 5 and 500 different fragments, between 25 and 400 different fragments, between 100 and 10,000 different fragments,

between 50 and 250 different fragments, or between 2 and 125 different fragments of the larger contiguous nucleic acid (1118). In some embodiments, the one or more fragments 804 of the larger contiguous nucleic acid in a partition in the plurality of partitions comprises more than 2 different fragments 804, more than 10 different fragments 804, more than 20 different fragments 804, more than 50 different fragments 804, more than 100 different fragments 804, more than 200 different fragments 804, more than 300 different fragments 804, more than 400 different fragments 804, or more than 500 different fragments 804 of the larger contiguous nucleic acid (1118). In a specific embodiment, the one or more fragments 804 of the larger contiguous nucleic acid 802 in a partition in the plurality of partitions consists of between 5 and 100 fragments 804 of the larger contiguous nucleic acid 802 (1120).

[00120] Referring to Figure 11B (1122), in some embodiments, the plurality of sequence reads is obtained from less than 5 nanograms of nucleic or ribonucleic acid. That is, the total amount of nucleic acid 804 needed for the plurality of partitions is advantageously less than 5 nanograms in some embodiments. In varying embodiments in accordance with the present dislcosure, the plurality of sequence reads is obtained from less than 500 nanograms of nucleic or ribonucleic acid, less than 250 nanograms of nucleic or ribonucleic acid, less than 50 nanograms of nucleic or ribonucleic acid, less than 5 nanograms of nucleic or ribonucleic acid, or less than 2 nanograms of nucleic or ribonucleic acid.

[00121] Referring to Figure 11B (1124), in some embodiments, the identifier in the second portion 732 of each respective sequence read 728 in the set of sequence reads encodes a common value selected from the set {1, …, 1024}, the set {1, …, 4096}, the set {1, …, 16384}, the set {1, …, 65536}, the set {1, …, 262144}, the set {1, …, 1048576}, the set {1, …, 4194304}, the set {1, …, 16777216}, the set {1, …, 67108864}, or the set {1, …, 1×10^{12}}. For instance, consider the case in which the identifier is represented by a set of five nucleotide positions. In this instance, each nucleotide position contributes four possibilities (A, T, C or G), giving rise, when all five positions are considered, to 4 x 4 x 4 x 4 x 4 = 1024 possibilities. As such, the five nucleotide positions form the basis of the set {1,…, 1024}. Thus, when the barcode sequence is a 5-mer, the second portion 732 of each sequencing read 728 encodes a unique predetermined value selected from the set {1,…, 1024}. Likewise, when the barcode sequence is represented by a set of six nucleotide positions, the six nucleotide positions collectively contribute 4 x 4 x 4 x 4 x 4 x 4 = 4096 possibilities. As such, the six nucleotide positions form the basis of the set {1,…, 4096}. In other

words, when the barcode sequence is a 6-mer, the second portion 732 of each sequencing read 128 encodes a unique predetermined value selected from the set {1,…, 4096}. In some embodiments, the identifier is an N-mer, and N is an integer selected from the set {4, …, 20} (1126).

[00122] In some embodiments, the identifier 732 of a sequencing read 728 in the plurality of sequencing reads is localized to a contiguous set of oligonucleotides within the sequencing read. In one such exemplary embodiment, the contiguous set of oligonucleotides is an N-mer, where N is an integer selected from the set {4, …, 20} (214). In other words, in some embodiments, the barcode 732 in, for instance Figure 12A, panel B, is a contiguous set of nucleotide positions (e.g., 4 contiguous nucleotide positions, 5 contiguous nucleotide positions, 6 contiguous nucleotide positions, 7 contiguous nucleotide positions, 8 contiguous nucleotide positions, 9 contiguous nucleotide positions, 10 contiguous nucleotide positions, 11 contiguous nucleotide positions, 12 contiguous nucleotide positions, 13 contiguous nucleotide positions, 14 contiguous nucleotide positions, 15 contiguous nucleotide positions, 16 contiguous nucleotide positions, 17 contiguous nucleotide positions, 18 contiguous nucleotide positions, 19 contiguous nucleotide positions, or 20 contiguous nucleotide positions) within oligonucleotide tag 1202 which ultimately becomes second portion 732 upon transcription of the larger contiguous nucleic acid.

[00123] By contrast, in some embodiments, the identifier in the second portion of a sequencing read in the plurality of sequencing reads is localized to a noncontiguous set of oligonucleotides within the sequencing read. In one such exemplary embodiment, the predetermined noncontiguous set of nucleotides collectively consists of N nucleotides, where N is an integer in the set {4, …, 20}. As an example, in some embodiments, referring to Figure 12A, panel B, barcode sequence 732 comprises a first set of contiguous nucleotide positions at a first position in oligonucleotide tag 1202 and a second set of contiguous nucleotide positions at a second position in oligonucleotide tag 120, that is displaced from the first set of contiguous nucleotide positions by a spacer. In one specific example, the barcode sequence 732 comprises $(X1)_nY_z(X2)_m$, where X1 is n contiguous nucleotide positions, Y is a constant predetermined set of z contiguous nucleotide positions, and X2 is m contiguous nucleotide positions. In this example, the identifier in the second portion of the sequencing read 728 produced by a schema invoking this exemplary barcode is localized to a noncontiguous set of oligonucleotides, namely $(X1)_n$ and $(X2)_m$. This is just one of many examples of noncontiguous formats for barcode sequence within the scope of the present disclosure.

[00124] The nucleic acid barcode sequences 732 will typically include from 6 to about 20 or more nucleotides within the sequence of the oligonucleotides. In some embodiments, these nucleotides are completely contiguous, e.g., in a single stretch of adjacent nucleotides. In alternative embodiments, they are separated into two or more separate subsequences that are separated by one or more nucleotides. Typically, separated subsequences are separated by about 4 to about 16 intervening nucleotides.

[00125] Referring to Figure 11B (1128 and 1130), in some embodiments, an average sequence read length of the plurality of sequence reads is between 40 bases and 200 bases (1128), between 40 bases and 200 bases, between 60 bases and 140 bases, or between 30 bases and 130 bases. In some embodiments the sequence reads have a sequence length that is compatible with the flow cells of ILLUMINA, HISEQ, MISEQ or related systems.

[00126] In some embodiments, the plurality of sequence reads, in the aggregate, have an N50 (where the sum of the sequence read lengths that are greater than the stated N50 number is 50% of the sum of all sequence read lengths) of at least about 10kb, at least about 20kb, or at least about 50kb. In some aspects, sequence read lengths having an N50 of at least about 100kb, at least about 150kb, at least about 200kb, and in many cases, at least about 250kb, at least about 300 kb, at least about 350 kb, at least about 400 kb, and in some cases, or at least about 500 kb or more, are attained. In still other cases, maximum sequence read lengths in excess of 200 kb, in excess of 300 kb, in excess of 400 kb, in excess of 500 kb, in excess of 1 Mb, or even in excess of 2 Mb are obtained in accordance with the present disclosure.

[00127] Referring to Figure 11C (1128 and 1130), in some embodiments, a set of sequence reads in the plurality of sequence reads comprises more than 100 sequence reads, and each sequence read 728 of the more than 100 sequence reads includes the same common second portion (1132). That is, a given partition in the plurality of partitions yields 100 sequence reads and each such sequence read 728 has the same identifier 732. More generally, in some embodiments, a set of sequence reads in the plurality of sequence reads comprises more than 10 sequence reads, more than 200 sequence reads, more than 500 sequence reads, more than 1000 sequence reads, more than 2500 sequence reads, or more than 5000 sequence reads and each such sequence read 728 includes the same common second portion (the same identifier 732).

[00128] Referring to Figure 11C (1134, 1136), in some embodiments, the larger contiguous nucleic acid 802 is a chromosome (1134). In some embodiments, the larger contiguous nucleic acid

802 is the genome of a multi-chromosomal organism such as a human. In some embodiments, the larger contiguous nucleic acid 802 includes whole genomes, individual chromosomes, exomes, amplicons, or any of a variety of different nucleic acids of interest. In some embodiments, the larger contiguous nucleic acid 802 is greater than 40 million base pairs in length (1136). In some embodiments, the larger contiguous nucleic acid 802 is greater than 100,000 base pairs in length, is greater than 1 million base pairs in length, is greater than 5 million base pairs in length, is greater than 10 million base pairs in length, is greater than 20 million base pairs in length, is greater than 30 million base pairs in length, or is greater than 50 million base pairs in length.

[00129] *Create a set of k-mers.* Referring to Figure 11C (1138), once the plurality of sequence reads is obtained, a respective set of k-mers is created for each sequence read 728 in the plurality of sequence reads. Taken together, the sets of k-mers collectively comprise a plurality of k-mers. That is, the sets of k-mers from the plurality of sequence reads are pooled to form a plurality of k-mers. Importantly, the identifier 732 of the source sequence read 728 for each k-mer 746 in the plurality of k-mers is retained. For instance, a data structure that lists, for each respective k-mer 746 observed, the barcodes (second portions) 732 of the sequence reads 728 that contain the respective k-mer is used in some embodiments of the present disclosure. In general, any technique or data structure that tracks the barcodes 732 associated with the sequence reads 728 that contain a respective k-mer 746 are within the scope of the present disclosure.

[00130] The value k is necessarily less than the average length of the sequence reads 728 in the plurality of sequence reads. In some embodiments, k have a value of 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, or 43. In some embodiments k is an even integer between 6 and 100. In some embodiments k is an odd integer between 5 and 99. In some embodiments, each respective set of k-mers includes at least eighty percent of the possible k-mers of length k of the corresponding sequence read 728 (1138). For example, referring to Figure 10, if the sequence read has the sequence GATCCATCTT and k is set to 3, the corresponding set of k-mers includes at least 80 percent of the 8 possible k-mers, e.g., at least 6 of the 8 possible k-mers.

[00131] In some embodiments, each respective set of k-mers includes at least sixty percent, at least seventy percent, at least eighty percent, at least ninety percent, at least ninety-five percent, or all of the possible k-mers of length k of the corresponding sequence read 728 (1138). In some embodiments, a respective set of k-mers includes at least sixty percent, at least seventy percent, at least eighty percent, at least ninety percent, at least ninety-five percent, or all of the possible k-mers

of length k of the first portion of the corresponding sequence read 728 (1138). In some embodiments, two or more sets of k-mers includes at least sixty percent, at least seventy percent, at least eighty percent, at least ninety percent, at least ninety-five percent, or all of the possible k-mers of length k of the corresponding sequence read 728 (1138).

[00132] Referring to Figure 11C (1140), in some embodiments, the plurality of sequence reads collectively provide at least 15X or at least 25X coverage for the larger contiguous nucleic acid 802, more than ten percent or more than thirty percent of k-mers 746 in the plurality of k-mers are from more than one sequence read 728 in the plurality sequence reads, and the identifier 732 of each such sequence read 728 for each k-mer 746 represented by more than one sequence read is retained.

[00133] In some embodiments, more than ten percent, more than twenty percent, more than thirty percent, more than forty percent, more than fifty percent, more than sixty percent, more than seventy percent, or more than eighty percent of the k-mers in the plurality of k-mers are from more than one sequence read 728 in the plurality sequence reads. In instances where a respective k-mer 746 is found in more than one sequence read 728, the barcode (second portion) 732 of each sequence read 728 is associated with the respective k-mer 746.

[00134] In some embodiments, the plurality of sequence reads collectively provide at least 5X, 10X, 15X, 20X, 25X, or at least 30X coverage for the larger contiguous nucleic acid 802.

[00135] Referring to Figure 11C (1140), in some embodiments, the sequence reads 728 in the plurality of sequence reads encode between 75 and 125 bases of the larger contiguous nucleic acid 802 and the value k is an odd integer between 5 and 73 (1142). In some embodiments, the sequence reads 728 in the plurality of sequence reads encode between 80 and 120 bases of the larger contiguous nucleic acid 802 and the value k is an odd integer between 11 and 51. In some embodiments, the sequence reads 728 in the plurality of sequence reads encode between 90 and 110 bases of the larger contiguous nucleic acid 802 and the value k is an odd integer between 13 and 45.

[00136] Referring to Figure 11C (1144), in some embodiments, the respective set of k-mers is created by hashing the first portion 730 each sequence read 728 in the plurality of sequence reads according to a predetermined k-mer length k thereby creating a respective set of k-mers for each sequence read 728 in the plurality of sequence reads. One such method of generating k-mers from a sequence read involves computing a hash function at each position of a sliding window of

nucleotide length k overlayed on the first portion 730 of sequence read 728, in order to form a set of k-mers corresponding to the first portion 730 of the sequence read. In this method, each possible k-mer of length k is hashed using a simple function, called the hash function. One possible hash function h(W) for nucleic acid sequences maps a k-mer of the form $W = w_1 w_1 \cdots w_k$ from the first portion 730 of the sequence read 728 according to:

$$h(W) = f(w_k) + f(w_{k-1}) \times 4 + \cdots + f(w_1) \times 4^{k-1}$$

where $f(w_i) = 0, 1, 2, 3$ for $w_i = \mathrm{A},\ \mathrm{C},\ \mathrm{G},\ \mathrm{T}$. Other hash functions are possible. In general any function that assigns a k-mer of the sequence $W = w_1 w_1 \cdots w_k$ from a sequence read to a unique number that can be used to reconstruct the actual nucleic acid sequence of the k-mer can be used. Figure 9 shows the possible values of $h(W)$ for the 64 possible k-mers of length three. Accordingly, in some embodiments, the value $h(W)$ for each k-mer 746 is stored in hash table 740 or some other form of data structure. In some embodiments, as illustrated in Figure 7, for each sequence read 728, the associated second portion 732 (e.g., bar code) of the partition from which the sequence read 728 is retained. Further, in some embodiments, the data structure for k-mers 740 further includes the position of each k-mer 746 in the corresponding sequence read 728. In typical embodiments, each possible k-mer in the first portion 730 of sequence read 728 is converted to a k-mer 746. For instance, in an example where the k-mer length is three and $W = w_1 w_1 \cdots w_k$ (e.g., the second portion 732 of the sequence read 728) is GATCCATCTT, the k-mers of Figure 10 would be generated and hash values, e.g., $h(W)$ values, for each of these k-mers would be generated.

[00137] *Track the k-mers.* Referring to Figure 11D (1146), for each respective k-mer 746 in the plurality of k-mers, an identity of each sequence read 728 in the plurality of sequence reads that contains the respective k-mer 746 and the identifier 732 of the set of sequence reads that contains the sequence read is tracked. For example, referring to Figure 7, one way to accomplish this is to track all the k-mers 746 generated from a sequence read 728 along with the barcode 732. Alternative ways are possible. For instance, a data structure, or collection of data structures, that track all the sequence reads 728 that contain a respective k-mer 746, along with the barcodes 732 of these sequence reads 728, may be implemented. This data structure, or collection of data structures, would further track the sequence reads 728 that contain all other respective k-mer 746, along with the barcodes 732 of these sequence reads 728.

[00138] *Graph the k-mers.* Referring to Figure 11D (1148), the plurality k-mers are graphed as a graph comprising a plurality of nodes connected by a plurality of directed arcs. Each node

comprises an uninterrupted set of k-mers in the plurality of k-mers of length k with k-1 overlap. Each arc connects an origin node to a destination node in the plurality of nodes. A final k-mer of an origin node has k-1 overlap with an initial k-mer of a destination node. A first origin node has a first directed arc with both a first destination node and a second destination node in the plurality of nodes (1148).

[00139] Figure 14 illustrates one such graph 1400, which is known in the art as a de Bruijn graph. See Zerbino and Birney, "Velvet: Algorithms for de novo short read assembly using de Bruijn graphs," Genome Reach 2008, 18:821-829, which is hereby incorporated by reference in its entirety. Each node 1402 represents a series of overlapping k-mers. In the case of Figure 14, k = 5. Adjacent k-mers in a node 1402 overlap by k-1 nucleotides. The marginal information contained by a k-mer is its last (e.g., last 3') nucleotide. The sequence of those final nucleotides is called the sequence of the node 1402b, or s(N). In Figure 14, and in optional embodiments, each node 1402 is attached to a twin node 1402b, which represents the reverse series of reverse complement k-mers. This ensures that overlaps between reads from opposite strands are taken into account. Note that the sequences attached to a node and its twin do not need to be reverse complements of each other. The union of a node 1402 and its twin 1402b is referred to a "block." Nodes 1402 are connected by a directed "arc" 1404. In that case, the last k-mer of an arc's origin node overlaps with the first of its destination node. Because of the symmetry of the blocks, if an arc goes from node 1402-1 to node 1402-2, a symmetric arc goes from node 1402b-2 to 1402b-1. On these nodes 1402 and arcs 1404, reads are mapped as "paths" traversing the graph. Extracting the nucleotide sequence from a path is taken from the initial k-mer of the first node and the sequences of all the nodes in the path.

[00140] Figure 14 illustrates a situation in which a first origin node has a first directed arc with both a first destination node and a second destination node in the plurality of nodes (1148). In particular node 1402-2 is an origin node with a first directed arc 1404-2 to node 1402-3 and a second directed art 1404-3 to node 1402-4. This is because there is k-1 overlap between last k-mer of node 1402-2 (GATTG) and the first k-mer of node 1402-3 (ATTGA), i.e., an overlap of ATTG, as well as an overlap between last k-mer of node 1402-2 (GATTG) and the first k-mer of node 1402-4 (ATTGC), i.e., an overlap of ATTG. Thus, there are two paths in graph 1400, the path that comprises 1402-1 → 1404-1 → 1402-2 → 1404-2 → 1402-3 and the path that comprises 1402-1 → 1404-1 → 1402-2 → 1404-3 → 1402-4. The disclosed methods are useful in determining which path more likely represents the corresponding sequence of the larger nucleic acid 802.

[00141] *Determine which nodes to merge.* Referring to Figure 11D (1150), a determination as to whether to merge the origin node with the first destination node or the second destination node is made in order to derive a contig sequence comprising (i) the origin node and (ii) one of the first destination node and the second destination node. The contig sequence that is more likely to be representative of the corresponding portion of the larger contiguous nucleic acid 802 is determined and selected. For example, referring to Figure 14, a determination is made as to whether to merge node 1402-2 (origin node) and 1402-3 (first destination node) or to merge 1402-2 (origin node) and 1402-4 (second destination node). This determining 1150 advantageously uses at least the identifiers 732 of the sequence reads 728 for k-mers 740 in the first origin node, the first destination node, and the second destination node to identify the contig that is more likely to be representative of the corresponding portion of the larger contiguous nucleic acid 802.

[00142] In some embodiments, step 1150 is accomplished, at least in part, by evaluating the barcodes 732 associated with each k-mer in two or more paths. For instance, in graph 1400, the barcodes 732 of the k-mers in path 1402-1 → 1404-1 → 1402-2 → 1404-2 → 1402-3 and the barcodes 732 of the k-mers in path 1402-1 → 1404-1 → 1402-2 → 1404-3 → 1402-4 are evaluated. Since the source nucleic acid 802 typically includes multiple copies of the target nucleic acid (e.g., multiple copies of the same target chromosome or other nucleic acid that is to be sequenced), there are typically multiple barcodes 732 for each k-mer in a node. Nevertheless, the correct path, that is the path that represents a true sequence found in the larger contiguous nucleic acid 802, will have more barcodes in common than paths that do not represent the true sequence found in the larger contiguous nucleic acid 802. The actual method by which barcodes are used to identify the correct path is dependent upon a number of variables and so may vary.

[00143] In some embodiments, the more likely path, that is the better contig sequence comprising (i) the origin node and (ii) one of the first destination node and the second destination node, is selected by imposing a cut off, for instance, requiring at least 30 percent of the barcodes representing any of the k-mers in the origin node and any of the k-mers in the selected destination node be common between the two nodes.

[00144] In some embodiments, the more likely path, that is the better contig sequence comprising (i) the origin node and (ii) one of the first destination node and the second destination node, is selected by imposing a cut off, for instance, requiring at least 10 percent of the barcodes representing, at least 20 percent of the barcodes representing, at least 30 percent of the barcodes

representing, at least 40 percent of the barcodes representing, at least 50 percent of the barcodes representing or at least 60 percent of the barcodes representing any of the k-mers in the origin node and any of the k-mers in the selected destination node be common between the two nodes.

[00145] In some embodiments, the more likely path, that is the better contig sequence comprising (i) the origin node and (ii) one of the first destination node and the second destination node, is selected by identifying the node pair that has the highest percentage of shared barcodes.

[00146] Referring to Figure 11D (1152), in some embodiments, the first origin node and the first destination node are part of a first path in the graph that includes one or more additional nodes other than the origin node and the first destination node. The first origin node and the second destination node are part of a second path in the graph that includes one or more additional nodes other than the origin node and the second destination node. This is illustrated in Figure 14 as path 1402-1 → 1404-1 → 1402-2 → 1404-2 → 1402-3 and path 1402-1 → 1404-1 → 1402-2 → 1404-3 → 1402-4 where node 1402 is an additional node in both instances. In some such embodiments, the determining of 1150 comprises determining whether the first path is more likely representative of the larger contiguous nucleic acid than the second path by evaluating a number of identifiers 728 shared between the k-mers of the nodes of a first portion of the first path and the k-mers of the nodes of a second portion of the first path versus a number of identifiers shared between the k-mers of the nodes of a first portion of the second path and the k-mers of the nodes of a second portion of the second path (1152). In some embodiments, this is done on a relative basis. For instance, of the total number of identifiers 746 associated with any of the k-mers in a path, what percentage of them are shared between a first portion of the path and a second portion of the path. For example, referring to Figure 15, consider an instance where path 1 comprises nodes 1502, 1504, 1506, 1508, 1510, and path 2 comprises nodes 1512, 1514, 1506, 1516, and 1518. Path 1 can be divided into a first portion consisting of nodes 1502, 1504, and 1506 and a second portion consisting of nodes 1508 and 1510. The identifiers 732 associated with any of the k-mers of nodes 1502, 1504, and 1506 are determined. The identifiers 732 associated with any of the k-mers of nodes 1508 and 1510 are determined. This is repeated for the second path. That is, path 2 is divided into a first portion consisting of nodes 1512, 1514, and 1506 and a second portion consisting of nodes 1516 and 1518. The identifiers 732 associated with any of the k-mers of nodes 1512, 1514, and 1506 are determined. The identifiers 732 associated with any of the k-mers of nodes 1516 and 1518 are determined. The extent to which there is overlap between the identifiers 732 of the first and second

portions of the first path is compared to the extent to which there is overlap between the identifiers 732 of the first and second portions of the second path. For instance, consider the case in which thirty percent of the total number of identifiers found in either the first or second portion of the first path is found in both the first and second portion of the first path. Further, forty-five percent of the total number of identifiers found in either the first or second portion of the second path is found in both the first and second portion of the second path. In this instance, the second path will be deemed to be more likely representative of a corresponding part of the larger contiguous nucleic acid 802. Other techniques for comparing the barcodes 732 of k-mers in paths may be used, such as Bayesian analysis or Pearson's chi-squired test. The instance discovery is the use of the barcodes 732 to break ambiguity in the paths.

[00147] Referring to Figure 11E (1154), in some embodiments, the first origin node and the first destination node are part of a first path in the graph that includes one or more additional nodes other than the origin node and the first destination node. The first origin node and the second destination node are part of a second path in the graph that includes one or more additional nodes other than the origin node and the second destination node. The determining up-weights the first path relative to the second path when the first path has higher average coverage than the second path (1154). Referring again to Figure 15 to illustrate and again considering path 1 (1502, 1504, 1506, 1508, 1510), and path 2 (nodes 1512, 1514, 1506, 1516, and 1518), if there is higher coverage by the k-mers of the nodes of path 1 relative to the k-mers of the nodes of path 2, this may be used as a tie breaker in instances where process 1150 or 1152 of Figure 11 was not able to identify a clear winner based on the sequence identifiers alone.

[00148] Referring to Figure 11E (1156), in some embodiments, the first origin node and the first destination node are part of a first path in the graph that includes one or more additional nodes other than the origin node and the first destination node. The first origin node and the second destination node are part of a second path in the graph that includes one or more additional nodes other than the origin node and the second destination node. The determining up-weights the first path relative to the second path when the first path represents a longer contiguous portion of the larger contiguous nucleic acid sequence than the second path (1156). In some embodiments, process 1156 is used as a tie breaker in instances where process 1150, 1152, or 1154 of Figure 11 is not able to identify a clear winner based on the sequence identifiers alone. In some embodiments,

process 1156 is used in conjunction with processes 1150, 1152, and/or 1154 of Figure 11 to identify a winner using multiple sources of information.

[00149] Referring to Figure 11E (1158), in some embodiments, a first k-mer 746 in the first node is present in a sub-plurality of the plurality of sequence reads 728 and the identity of each sequence read 728 in the sub-plurality of sequence reads is retained for the first k-mer 728 and used by the determining 1150 to determine whether the first path is more likely representative of the larger contiguous nucleic acid sequence than the second path (1158). This arises in many instances where the sample of the larger contiguous nucleic acid 802 that is partitioned into fragements 804 in fact constitutes multiple copies of a target nucleic acid (e.g., multiple copies of the same target chromosome or other nucleic acid that is to be sequenced). This results in multiple barcodes 732 for each k-mer 746 in a node. As such, any of the processes 1150, 1152, 1154, 1156 and/or 1158 are capable of taking this into account.

[00150] Another aspect of the present disclosure provides a sequencing method comprising, at a computer system having one or more processors, and memory storing one or more programs for execution by the one or more processors, obtaining a plurality of sequence reads as disclosed above in relation to 1106 of Figure 11A. The plurality of sequence reads comprises a plurality of sets of sequence reads. Each respective sequence read in a set of sequence reads includes (i) a unique first portion that corresponds to a subset of a larger contiguous nucleic acid and (ii) a common second portion that forms an identifier that is independent of the sequence of the larger contiguous nucleic acid and that identifies a partition, in a plurality of partitions, in which the respective sequence read was formed. Each respective set of sequence reads in the plurality of sets of sequence reads is formed in a partition in the plurality of partitions. Each such partition includes one or more fragments of the larger contiguous nucleic acid that is used as the template for each respective sequence read in the partition.

[00151] In the method, a respective set of k-mers is created for each sequence read in the plurality of sequence reads as discussed above in relation to element 1138 of Figure 11C. The sets of k-mers collectively comprise a plurality of k-mers. The identifiers 732 of the sequence read 728 for each k-mer 746 in the plurality of k-mers is retained. The value k is less than the average length of the sequence reads in the plurality of sequence reads. Each respective set of k-mers includes at least some of (e.g., at least eighty percent of) the possible k-mers of the first portion of the corresponding sequence read.

[00152] In the method, there is tracked, for each respective k-mer in the plurality of k-mers, an identity of each of the sequence reads in the plurality of sequence reads that contains the respective k-mer. A first path is identified. The first path comprises a first set of k-mers with k-1 overlap in the plurality of k-mers (e.g., nodes 1502, 1504, 1506, 1508, and 1510 of Figure 15). A second path is identified. The second path comprises a second set of k-mers with k-1 overlap in the plurality of k-mers (e.g., nodes 1512, 1514, 1506, 1516, and 1518 of Figure 15). The first path intersects the second path (e.g., as illustrated in Figure 15), thereby forming the set of branch segments comprising: a left portion of the first path (nodes 1502 and 1504), a right portion of the first path (nodes 1508 and 1510), a left portion of the second path (nodes 1512 and 154) and a right portion of the second path (1516 and 1518) where the choice of what do with the bisecting node 1506 (e.g, not use it, put it in paths, or put it in one portion of each path) is application dependent. The identifiers associated with the k-mers of the set of branch segments are used to verify the connectivity of the branch segments. In some embodiments, this comprises evaluating a number of identifiers shared between each possible branch segment pairs in the set of branch segments. In some instances, this may result in rerouting a path. For example, in evaluating the graph of Figure 15, four possible paths are possible, as set forth in the following Table. As the table indicates, path 1520 to 1522 (nodes 1502, 1504, 1506, 1508, and 1510) and path 1524 to 1526 (1512, 1514, 1506, 1516, and 1518 are verified as correct because they each have more overlapping barcodes, on a relative basis, than path 1520 to 1526 and 1524 to 1522.

Left Path branch segment	Left Path branch segment	Number of Sequence Identifiers 732 in Left Branch Segment	Number of Sequence Identifiers 732 in Right Branch Segment	Number of Sequence Identifiers 732 that overlap left and right branch
1520	1522	100	110	65
1520	1526	100	105	3
1524	1522	90	110	8
1524	1526	90	105	50

[00153] In some embodiments, additional information such as path lengths and path coverage is used to verify these connections. It will be appreciated that any form of statistical analysis may be applied to the information such as that in the above table to verify the paths.

III. Exemplary Assembly Process

[00154] The following is an example of an overall de novo assembly process as described above, is set forth below, and schematically illustrated in the process flow chart shown in Figure 4. As shown, following sequencing and read generation, the short read sequences are aggregated in a data structure and subjected to pre-assembly processing at step 402. In these reads, sequence data associated with functional components of the sequenced fragments, e.g., adapter sequences, read primer sequences, etc., are removed from the sequence reads. Duplicate reads may be eliminated by identifying those that are duplicated both in the first two k-mers of the sense read and the partner or antisense read. Where the initial k-mers are common among duplicate reads and their complements, all but one can be removed as a duplicate using, e.g., A FastUniq algorithm, or other similar algorithms. In some cases, duplicate reads may get through the pre-assembly process, where one or more of those duplicates include a sequencing error. Because these first few k-mers occur at the beginning of a sequencing run where errors are fewer, it would be expected that the probability of this occurring would be low. In addition to duplicate reads, low complexity reads, e.g., reads where a single base makes up 90% or more of the read, or reads that are entirely or substantially made up of 2, 3, or 4 base repeat units.

[00155] The reads are then subjected to an error correction algorithm at step 404, as uncorrected errors in sequence reads result in fragmented assemblies and add computational complexity. Different error correction algorithms are generally available (see, e.g., Tahir et al., "Review of Genome Sequence Short Read Error Correction Algorithms", Am. J. Bioinformatics Res. 2013; 3(1) 1-9, which is hereby incorporated herein by reference) and include, for example the QUAKE, and standalone ALLPATHS-LG error correction algorithm.

[00156] Following the preassembly processing and error correction, using a k-mer size of, e.g., k=20 or k=24, a frequency spectrum of k-mers is generated across all of the aggregated reads at step 406. Assuming a typical genome profile, sequenced without significant bias, the expectation would be for a Gaussian distribution of k-mer frequency, where the center of the peak indicates the mean k-mer coverage of the genome. The size of the k-mer used for calculation of the frequency

spectrum may generally vary depending upon the complexity of the genome being analyzed, the expected level of sequencing coverage, and the like.

[00157] An initial assembly graph is then assembled at step 408 from the error corrected reads, using the k-mer assembly process described with reference to Figure 1B. In some cases, longer k-mer sequences may be used, e.g., for sequence reads of 150 bases, k-mers of 40 or more, 50 or more, or 60 or more bases may be used. From the graph, unbranched paths are collapsed into individual primary paths.

[00158] Hanging ends that represent relatively short branches, e.g., of 1, 2 or 3 k-mers, are generally prevalent in the sequence and can result from a number of origins, including from uncorrected reads. These short paths are easily recognized and may be removed from the assembly graph, either before or after identifying the primary paths. In some cases, short branches may appear, but may be derived from missing sequence in the error corrected reads, which can result from low coverage.

[00159] The barcode sequences coupled with the error corrected reads are then associated with their respective primary paths, at step 410 generating a list of the barcodes associated with each primary path. For each primary path, a copy number of that primary path within the genome is estimated. As noted, the copy number of a given primary path will generally depend upon the complexity of the genome and the length and complexity of the primary path. Copy number can generally be estimated from the number of reads that touch a primary path, and comparing it to the mean k-mer coverage derived from the k-mer spectrum calculated at step 406. Short primary paths may generally be ignored, as determining their copy number is more difficult since a much larger proportion of reads that touch on those primary paths may only partially overlap them. From this, low copy number primary paths, e.g., with a copy number approaching 1, also referred to as CN1 paths, as noted above, may be identified at step 412.

[00160] The CN1 primary paths are then assembled in a scaffold with the aid of the associated barcode sequences. In particular, the CN1 primary path that is associated with the largest number of barcodes is selected as a seed path at step 414. At step 416, the seed path is then extended to all other overlapping primary paths containing the common barcode sequence, considering each primary path only once, and ignoring primary paths that have a barcode frequency below a desired level, e.g., less than half the barcode frequency in the seed path. Figure 5 illustrates the process in greater detail. In particular, the first seed path is identified at step 502. That first

seed path is ten extended as described for step 416, above. In particular, the extension is continued through primary paths and additional CN1 paths at step 504, until the end of a given branch is reached. The path is then trimmed back to the last CN1 path through which the path traveled at step 506, such that the assembled sequence will always begin an end on a CN1 path, and be anchored around the seed path.

[00161] The paths through the CN1 paths are then condensed by removing all of the non-CN1 paths, then merging together to form a single scaffold that defines the order to the CN1 paths around the first seed path at step 508. These CN1 paths are then designated as used.

[00162] The process is then repeated at step 512 using the CN1 path, from the remaining CN1 paths, that represents the highest remaining barcode frequency as the next seed path, and scaffolding against the remaining sets of CN1 paths. This is continued until all CN1 paths are designated as used. The set of condensed primary paths is then further merged to provide a final path or scaffold that gives the relative order of all of the CN1 paths in the genome at step 514. This forms the anchor scaffold which provides the basis for final assembly.

[00163] Once the CN1 path scaffold is assembled, local assembly is carried out to fill in the scaffold gaps between CN1 paths at step 420. For each gap in the scaffold, the adjacent primary paths are selected and the barcodes associated with them are identified. The group of reads that includes these barcodes are then used in conjunction with the adjacent CN1 paths to assemble a local primary path assembly graph in the same manner as used to assemble the global primary path graph. From each CN1 path, all paths that connect to the other primary path are identified. In many cases, only a single path will be identified allowing the use of that path as the patch to the gap. Where more than one path is identified between two adjacent CN1 primary paths, the path with the strongest read support is used to patch together the adjacent CN1 primary paths and close the gap. In cases where read support does not allow selection as between two paths, both paths may be included in the graph, or one can allow the gap to remain open.

[00164] From the full assembly, a linear graph of the entire genome is created that, ideally, includes no remaining gaps or doubled paths. However, in many cases, the linear graph may include some locations of dual paths or remaining gaps, but the overall order of the sequence is known. For remaining gaps or double paths, one may further examine uncorrected sequence data, or discarded sequence data to determine if any gaps or doubles may be informed by those sequences that were removed from the process early on. Alternatively, other disambiguation techniques may

be used, such as using copy number information to inform ambiguous assemblies, or using read pair information to evaluate relative likelihood of multiple solutions for ambiguous assemblies.

[00165] In certain aspects, the methods provided herein are computer-implemented methods, wherein at least one or more steps of the method are carried out by a computer program. In some embodiments, the methods provided herein are implemented in a computer program stored on computer-readable media, such as the hard drive of a standard computer. For example, a computer program for determining at least one consensus sequence from replicate sequence reads can include one or more of the following: code for providing or receiving the sequence reads, code for identifying regions of sequence overlap between the sequence reads, code for aligning the sequence reads to generate a layout, contig, or scaffold, code for consensus sequence determination, code for converting or displaying the assembly on a computer monitor, code for applying various algorithms described herein, and a computer-readable storage medium comprising the codes.

[00166] In some embodiments, a system (e.g., a data processing system) that determines at least one assembly from a set of replicate sequences includes a processor, a computer-readable medium operatively coupled to the processor for storing memory, wherein the memory has instructions for execution by the processor, the instructions including one or more of the following: instructions for receiving input of sequence reads, instructions for overlap detection between the sequence reads, instructions that align the sequence reads to generate a layout, contig, or scaffold, instructions that apply a consensus sequence algorithm to generate at least one consensus sequence (e.g., a "best" consensus sequence, and optionally one or more additional consensus sequences), instructions that compute/store information related to various steps of the method, and instructions that record the results of the method.

[00167] In certain embodiments, various steps of the processes described herein utilize information and/or programs and generate results that are stored on computer-readable media (e.g., hard drive, auxiliary memory, external memory, server, database, portable memory device and the like. For example, information used for and results generated by the methods that can be stored on non-transitory computer-readable media include but are not limited to input sequence read information, set of pair-wise overlaps, newly generated consensus sequences, quality information, technology information, and homologous or reference sequence information.

[00168] In some aspects, the invention includes an article of manufacture for determining at least one sequence assembly that includes a machine-readable medium containing one or more programs which when executed implement the steps of the invention as described herein.

[00169] As will be understood to practitioners in the art from the teachings provided herein, the invention can be implemented in hardware and/or software. In some embodiments of the invention, different aspects of the invention can be implemented in either client-side logic or server-side logic. As will be understood in the art, the invention or components thereof may be embodied in a fixed media program component containing logic instructions and/or data that when loaded into an appropriately configured computing device cause that device to perform according to the invention. As will be understood in the art, a fixed media containing logic instructions may be delivered to a viewer on a fixed media for physically loading into a viewer's computer or a fixed media containing logic instructions may reside on a remote server that a viewer accesses through a communication medium in order to download a program component, or to which the user uploads data to be processed using the above described hardware and/or software, at the remote location or locations, in accordance with the processes set forth herein.

[00170] Information appliances (or digital devices) that are generally used in conjunction with the processes describe herein include conventional logic systems, such as conventional computer systems, including both individual or personal computers, networked computers, computer mainframes, and the like. Interaction with these processes may not be limited to such conventional computing systems, but may additionally or alternatively include any information appliance for interacting with a remote data application, and could include such devices as a digitally enabled television, cell phone/smart phone, personal digital assistant, tablet, wearable electronic devices, e.g., Google Glass, smart watch, etc. Programming for carrying out the processes described herein on any of the foregoing types of devices may reside within the memory of the device itself (either main memory or auxiliary memory), or it may be accessed from other locations and devices, e.g., computers through, e.g., an API on the device, and an appropriate communications network, e.g., LAN, WAN, Wi-Fi, cellular, Bluetooth or other communications system.

[00171] The *E. coli* genome (strain K12_DH10B) was used to model the de-novo assembly processes described herein. Sixty physical copies of the genome were sheared to an average size of 50 kilobases (standard deviation 5 kilobases) and partitioned at random into 5000 partitions, each

containing a unique barcode. The process of the creation of a barcode-sequencing library was simulated within each partition. This was followed by simulation of sequencing of this library utilizing 150bp paired-end reads to a mean depth of 30X. The simulation of the sequencing process incorporated an empirical error rate model derived from real sequencing data. Figure 6 shows a de novo assembly plot for the above-described modeled data. As shown, the assembly resulted in N50 contigs of 750 kb as compared to N50 contigs of 60 kb for data based upon short read sequence alone, e.g., without barcodes.

[00172] While the foregoing invention has been described in some detail for purposes of clarity and understanding, it will be clear to one skilled in the art from a reading of this disclosure that various changes in form and detail can be made without departing from the true scope of the invention. For example, all the techniques and apparatus described above can be used in various combinations. For example, particle delivery can be practiced with array well sizing methods as described. All publications, patents, patent applications, and/or other documents cited in this application are incorporated by reference in their entirety for all purposes to the same extent as if each individual publication, patent, patent application, and/or other document were individually and separately indicated to be incorporated by reference for all purposes.

What is claimed is:

1. A sequencing method of assembling nucleic acid sequence reads comprising:

at a computer system having one or more processors, and memory storing one or more programs for execution by the one or more processors:

obtaining a plurality of sequence reads derived from a larger contiguous nucleic acid, wherein two or more sequence reads derived from a common fragment of the larger contiguous nucleic acid comprise a common barcode sequence,

identifying a first subset of sequence reads in the plurality of sequence reads that comprise both overlapping sequences and a common barcode sequence; and

aligning the first subset of sequence reads to provide a contiguous linear nucleic acid sequence.

2. The method of claim 1, further comprising repeating the identifying and aligning steps with a plurality of different subsets of sequence reads to provide a plurality of contiguous linear nucleic acid sequences.

3. The method of claim 2, further comprising ordering the plurality of different contiguous linear nucleic acid sequences in a sequence context within the larger contiguous nucleic acid.

4. The method of claim 3, wherein the ordering comprises mapping the plurality of different contiguous linear nucleic acid sequences against a reference sequence.

5. The method of claim 3, wherein the ordering comprises:

identifying one or more sequence reads that comprise a barcode sequence common to a first contiguous linear nucleic acid sequence, but include overlapping sequences with a second contiguous linear nucleic acid sequence; and

identifying the first and second contiguous linear nucleic acid as structurally linked.

6. A method of assembling nucleic acid sequence reads into larger contiguous sequences, comprising:

at a computer system having one or more processors, and memory storing one or more programs for execution by the one or more processors:

obtaining a plurality of sequence reads derived from a larger contiguous nucleic acid,

identifying a first subsequence from a set of overlapping sequence reads in the plurality of sequence reads;

extending the first subsequence to one or more adjacent or overlapping sequences based upon the presence of a barcode sequence on the adjacent sequence that is common to the first subsequence; and

providing a linear nucleic acid sequence that comprises the first subsequence and the one or more adjacent sequences.

7. A sequencing method comprising, at a computer system having one or more processors, and memory storing one or more programs for execution by the one or more processors:

(A) obtaining a plurality of sequence reads, wherein

the plurality of sequence reads comprises a plurality of sets of sequence reads,

each respective sequence read in a set of sequence reads includes (i) a first portion that corresponds to a subset of a larger contiguous nucleic acid and (ii) a common second portion that forms an identifier that is independent of the sequence of the larger contiguous nucleic acid and that identifies a partition, in a plurality of partitions, in which the respective sequence read was formed, and

each respective set of sequence reads in the plurality of sets of sequence reads is formed in a partition in the plurality of partitions and each partition includes one or more fragments of the larger contiguous nucleic acid that is used as the template for each respective sequence read in the partition;

(B) creating a respective set of k-mers for each sequence read in the plurality of sequence reads, wherein

the sets of k-mers collectively comprise a plurality of k-mers,

the identifiers of the sequence reads for each k-mer in the plurality of k-mers is retained,

k is less than the average length of the sequence reads in the plurality of sequence reads, and

each respective set of k-mers includes at least eighty percent of the possible k-mers of length k of the first portion of the corresponding sequence read;

(C) tracking, for each respective k-mer in the plurality of k-mers, an identity of each sequence read in the plurality of sequence reads that contains the respective k-mer and the identifier of the set of sequence reads that contains the sequence read;

(D) graphing the plurality k-mers as a graph comprising a plurality of nodes connected by a plurality of directed arcs, wherein

each node comprises an uninterrupted set of k-mers in the plurality of k-mers of length k with k-1 overlap,

each arc connects an origin node to a destination node in the plurality of nodes,

a final k-mer of an origin node has k-1 overlap with an initial k-mer of a destination node, and

a first origin node has a first directed arc with both a first destination node and a second destination node in the plurality of nodes; and

(E) determining whether to merge the origin node with the first destination node or the second destination node in order to derive a contig sequence that is more likely to be representative of a portion of the larger contiguous nucleic acid, wherein the contig sequence comprises (i) the origin node and (ii) one of the first destination node and the second destination node, wherein the determining uses at least the identifiers of the sequence reads for k-mers in the first origin node, the first destination node, and the second destination node.

8. The sequencing method of claim 7, wherein,

the first origin node and the first destination node are part of a first path in the graph that includes one or more additional nodes other than the origin node and the first destination node,

the first origin node and the second destination node are part of a second path in the graph that includes one or more additional nodes other than the origin node and the second destination node, and

the determining (E) comprises determining whether the first path is more likely representative of the larger contiguous nucleic acid than the second path by evaluating a number of identifiers shared between the k-mers of the nodes of a first portion of the first path and the k-mers of the nodes of a second portion of the first path versus a number of identifiers shared between the

k-mers of the nodes of a first portion of the second path and the k-mers of the nodes of a second portion of the second path.

9. The sequencing method of claim 7, wherein

the first origin node and the first destination node are part of a first path in the graph that includes one or more additional nodes other than the origin node and the first destination node,

the first origin node and the second destination node are part of a second path in the graph that includes one or more additional nodes other than the origin node and the second destination node, and

the determining (E) up-weights the first path relative to the second path when the first path has higher average coverage than the second path.

10. The sequencing method of claim 7, wherein the

the first origin node and the first destination node are part of a first path in the graph that includes one or more additional nodes other than the origin node and the first destination node,

the first origin node and the second destination node are part of a second path in the graph that includes one or more additional nodes other than the origin node and the second destination node, and

the determining (E) up-weights the first path relative to the second path when the first path represents a longer contiguous portion of the larger contiguous nucleic acid sequence than the second path.

11. The sequencing method of any one of claims 7-10, wherein a first k-mer in the first node is present in a sub-plurality of the plurality of sequence reads and the identity of each sequence read in the sub-plurality of sequence reads is retained for the first k-mer and used by the determining (E) to determine whether the first path is more likely representative of the larger contiguous nucleic acid sequence than the second path.

12. The sequencing method of any one of claims 7-11, wherein

a partition in the plurality of partitions comprises at least 1000 molecules with the common second portion, and

each molecule in the at least 1000 molecules further comprises a primer sequence complementary to at least a portion of the larger contiguous nucleic acid.

13. The sequencing method of any one of claims 7-11, wherein

a partition in the plurality of partitions comprises at least 1000 molecules with the common second portion, and

each molecule in the at least 1000 molecules further comprises a primer site and a semi-random N-mer priming sequence that is complementary to part of the larger contiguous nucleic acid.

14. The sequencing method of any one of claims 7-13, wherein the one or more fragments of the larger contiguous nucleic acid in a partition in the plurality of partitions are greater than 50 kilobases in length.

15. The sequencing method of any one of claims 7-13, wherein the one or more fragments of the larger contiguous nucleic acid in a partition in the plurality of partitions are between 20 kilobases and 200 kilobases in length.

16. The sequencing method of any one of claims 7-15, wherein the one or more fragments of the larger contiguous nucleic acid in a partition in the plurality of partitions consists of between 1 and 500 different fragments of the larger contiguous nucleic acid.

17. The sequencing method of any one of claims 7-15, wherein the one or more fragments of the larger contiguous nucleic acid in a partition in the plurality of partitions consists of between 5 and 100 fragments of the larger contiguous nucleic acid.

18. The sequencing method of any one of claims 7-17, wherein the plurality of sequence reads is obtained from less than 5 nanograms of nucleic or ribonucleic acid.

19. The sequencing method of any one of claims 7-18, wherein the identifier in the second portion of each respective sequence read in the set of sequence reads encodes a common value selected from the set {1, …, 1024}, the set {1, …, 4096}, the set {1, …, 16384}, the set {1, …, 65536}, the

set {1, …, 262144}, the set {1, …, 1048576}, the set {1, …, 4194304}, the set {1, …, 16777216}, the set {1, …, 67108864}, or the set {1, …, 1 x 10^{12}}.

20. The sequencing method of any one of claims 7-18, wherein the identifier is an N-mer, and N is an integer selected from the set {4, …, 20}.

21. The sequencing method of any one of claims 7-20, wherein an average sequence read length of the plurality of sequence reads is between 40 bases and 200 bases.

22. The sequencing method of any one of claims 7-20, wherein an average sequence read length of the plurality of sequence reads is between 60 bases and 140 bases.

23. The sequencing method of any one of claims 7-22, wherein

the plurality of sequence reads collectively provide at least 15X coverage for the larger contiguous nucleic acid,

more than ten percent of the k-mers in the plurality of k-mers are from more than one sequence read in the plurality sequence reads, and

the identifier of each sequence read for each k-mer represented by more than one sequence read is retained.

24. The sequencing method of any one of claims 7-22, wherein

the plurality of sequence reads collectively provide at least 25X coverage for the larger contiguous nucleic acid,

more than thirty percent of the plurality of k-mers are from more than one sequence read in the plurality sequence reads, and

the identifier of each sequence read for each k-mer represented by more than one source sequence read is retained.

25. The sequencing method of any one of claims 7-24, wherein the sequence reads in the plurality of sequence reads encode between 75 and 125 bases of the larger contiguous nucleic acid and the value k is an odd integer between 5 and 73.

26. The sequencing method of any one of claims 7-25, wherein a set of sequence reads in the plurality of sequence reads comprises more than 100 sequence reads, and each sequence read of the more than 100 sequence reads includes the same common second portion.

27. The sequencing method of any one of claims 7-26, wherein the larger contiguous nucleic acid is a chromosome.

28. The sequencing method of any one of claims 7-27, wherein the larger contiguous nucleic acid is greater than 40 million base pairs in length.

29. The sequencing method of any one of claims 7-28, wherein creating the respective set of k-mers comprises hashing each sequence read in the plurality of sequence reads according to a predetermined k-mer length thereby creating a respective set of k-mers for each sequence read in the plurality of sequence reads.

30. A computing system, comprising:

one or more processors;

memory storing one or more programs to be executed by the one or more processors;

the one or more programs comprising instructions for:

(A) obtaining a plurality of sequence reads, wherein

the plurality of sequence reads comprises a plurality of sets of sequence reads,

each respective sequence read in a set of sequence reads includes (i) a unique first portion that corresponds to a subset of a larger contiguous nucleic acid and (ii) a common second portion that forms an identifier that is independent of the sequence of the larger contiguous nucleic acid and that identifies a partition, in a plurality of partitions, in which the respective sequence read was formed, and

each respective set of sequence reads in the plurality of sets of sequence reads is formed in a partition in the plurality of partitions and each partition includes one or more fragments of the larger contiguous nucleic acid that is used as the template for each respective sequence read in the partition;

(B) creating a respective set of k-mers for each sequence read in the plurality of sequence reads, wherein

the sets of k-mers collectively comprise a plurality of k-mers,

the identifier of the sequence read for each k-mer in the plurality of k-mers is retained,

k is less than the average length of the sequence reads in the plurality of sequence reads, and

each respective set of k-mers includes at least eighty percent of the possible k-mers of length k of the first portion of the corresponding sequence read;

(C) tracking, for each respective k-mer in the plurality of k-mers, an identity of each sequence read in the plurality of sequence reads that contains the respective k-mer and the identifier of the set of sequence reads that contains the sequence read;

(D) graphing the plurality k-mers as a graph comprising a plurality of nodes connected by a plurality of directed arcs, wherein

each node comprises an uninterrupted set of k-mers in the plurality of k-mers of length k with k-1 overlap,

each arc connects an origin node to a destination node in the plurality of nodes,

a final k-mer of an origin node has k-1 overlap with an initial k-mer of a destination node, and

a first origin node has a first directed arc with both a first destination node and a second destination node in the plurality of nodes; and

(E) determining whether to merge the origin node with the first destination node or the second destination node in order to derive a contig sequence that is more likely to be representative of a portion of the larger contiguous nucleic acid, wherein the contig sequence comprises (i) the origin node and (ii) one of the first destination node and the second destination node, wherein the determining uses at least the identifiers of the sequence reads for k-mers in the first origin node, the first destination node, and the second destination node.

31. A sequencing method comprising:

at a computer system having one or more processors, and memory storing one or more programs for execution by the one or more processors:

(A) obtaining a plurality of sequence reads, wherein

the plurality of sequence reads comprises a plurality of sets of sequence reads,

each respective sequence read in a set of sequence reads includes (i) a unique first portion that corresponds to a subset of a larger contiguous nucleic acid and (ii) a common second portion that forms an identifier that is independent of the sequence of the larger contiguous nucleic acid and that identifies a partition, in a plurality of partitions, in which the respective sequence read was formed,

each respective set of sequence reads in the plurality of sets of sequence reads is formed in a partition in the plurality of partitions and each partition includes one or more fragments of the larger contiguous nucleic acid that is used as the template for each respective sequence read in the partition;

(B) creating a respective set of k-mers for each sequence read in the plurality of sequence reads, wherein

the sets of k-mers collectively comprise a plurality of k-mers,

the identifier of the sequence read for each k-mer in the plurality of k-mers is retained,

k is less than the average length of the sequence reads in the plurality of sequence reads, and

each respective set of k-mers includes at least eighty percent of the possible k-mers of the first portion of the corresponding sequence read;

(C) tracking, for each respective k-mer in the plurality of k-mers, an identity of each sequence read in the plurality of sequence reads that contains the respective k-mer;

(D) identifying a first path, wherein the first path comprises a first set of k-mers with k-1 overlap in the plurality of k-mers;

(E) identifying a second path, wherein the second path comprises a second set of k-mers with k-1 overlap in the plurality of k-mers and wherein the first path intersects the second path, thereby forming the set of branch segments comprising: a left portion of the first path, a right portion of the first path, a left portion of the second path and a right portion of the second path; and

(F) using the identifiers associated with the k-mers of the set of branch segments, verify the connectivity of the branch segments.

32. The sequencing method of claim 31, wherein the using (F) comprises evaluating a number of identifiers shared between each possible branch segment pair in the set of branch segments.

33. The sequencing method of claim 31 or 32, wherein

a partition in the plurality of partitions comprises at least 1000 molecules with the common second portion, and

each molecule in the at least 1000 molecules includes a primer sequence complementary to at least a portion of the larger contiguous nucleic acid.

34. The sequencing method of any one of claims 31-33, wherein the identifier in the second portion of each respective sequence read in the set of sequence reads encodes a common value selected from the set {1, …, 1024}, the set {1, …, 4096}, the set {1, …, 16384}, the set {1, …, 65536}, the set {1, …, 262144}, the set {1, …, 1048576}, the set {1, …, 4194304}, the set {1, …, 16777216}, the set {1, …, 67108864}, or the set {1, …, 1×10^{12}}.

35. The sequencing method of any one of claims 31-34, wherein the identifier is an N-mer, wherein N is an integer selected from the set {4, …, 20}.

36. The sequencing method of any one of claims 31-35, wherein an average sequence read length of the plurality of sequence reads is between 40 bases and 200 bases.

37. The sequencing method of any one of claims 31-35, wherein an average sequence read length of the plurality of sequence reads is between 60 bases and 140 bases.

38. The sequencing method of any one of claims 31-37, wherein

the plurality of sequence reads collectively provide at least 15X coverage for the larger contiguous nucleic acid,

more than ten percent of the k-mers in the plurality of k-mers are from more than one source sequence read in the plurality sequence reads, and

the identifier of each such source sequence read for each k-mer represented by more than one source sequence read is retained.

39. The sequencing method of any one of claims 31-37, wherein

the plurality of sequence reads collectively provide at least 25X coverage for the larger contiguous nucleic acid,

more than thirty percent of the plurality of k-mers are from more than one source sequence read in the plurality sequence reads, and

the identifier of each such source sequence read for each k-mer represented by more than one source sequence read is retained.

40. The sequencing method of any one of claims 31-39, wherein the sequence reads in the plurality of sequence reads encode between 75 and 125 bases of the larger contiguous nucleic acid and k is an odd integer between 5 and 73.

41. The sequencing method of any one of claims 31-40, wherein a set of sequence reads in the plurality of sequence reads comprises more than 100 sequence reads, each with the same common second portion.

42. The sequencing method of any one of claims 31-41, wherein the larger contiguous nucleic acid is a chromosome.

43. The sequencing method of any one of claims 31-42, wherein the larger contiguous nucleic acid is greater than 40 million base pairs in length.

44. A computing system, comprising:

one or more processors;

memory storing one or more programs to be executed by the one or more processors;

the one or more programs comprising instructions for:

(A) obtaining a plurality of sequence reads, wherein

the plurality of sequence reads comprises a plurality of sets of sequence reads,

each respective sequence read in a set of sequence reads includes (i) a unique first portion that corresponds to a subset of a larger contiguous nucleic acid and (ii) a common second portion that forms an identifier that is independent of the sequence of the larger contiguous nucleic acid and that identifies a partition, in a plurality of partitions, in which the respective sequence read was formed,

each respective set of sequence reads in the plurality of sets of sequence reads is formed in a partition in the plurality of partitions and each partition includes one or more fragments

of the larger contiguous nucleic acid that is used as the template for each respective sequence read in the partition;

(B) creating a respective set of k-mers for each sequence read in the plurality of sequence reads, wherein

the sets of k-mers collectively comprise a plurality of k-mers,

the identifiers of the sequence reads for each k-mer in the plurality of k-mers is retained,

k is less than the average length of the sequence reads in the plurality of sequence reads, and

each respective set of k-mers includes at least eighty percent of the possible k-mers of the first portion of the corresponding sequence read;

(C) tracking, for each respective k-mer in the plurality of k-mers, an identity of each sequence read in the plurality of sequence reads that contains the respective k-mer;

(D) identifying a first path, wherein the first path comprises a first set of k-mers with k-1 overlap in the plurality of k-mers;

(E) identifying a second path, wherein the second path comprises a second set of k-mers with k-1 overlap in the plurality of k-mers and wherein the first path intersects the second path, thereby forming the set of branch segments comprising: a left portion of the first path, a right portion of the first path, a left portion of the second path and a right portion of the second path; and

(F) using the identifiers associated with the k-mers of the set of branch segments, verify the connectivity of the branch segments.

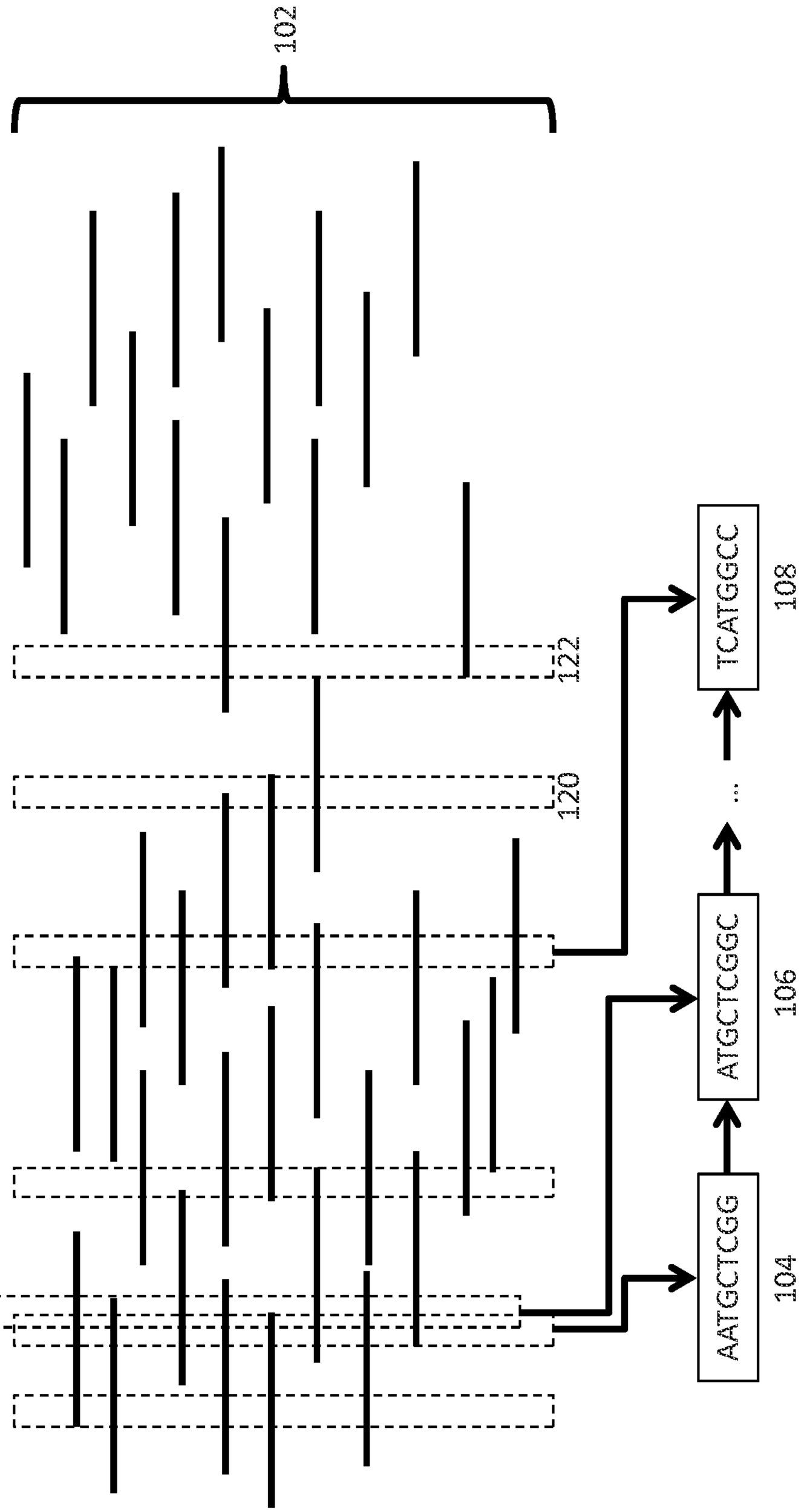


Figure 1A

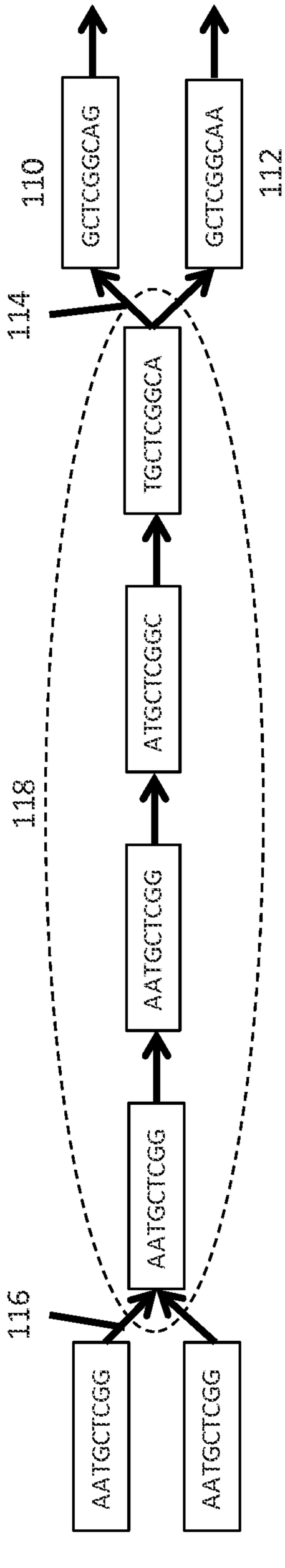


Figure 1B

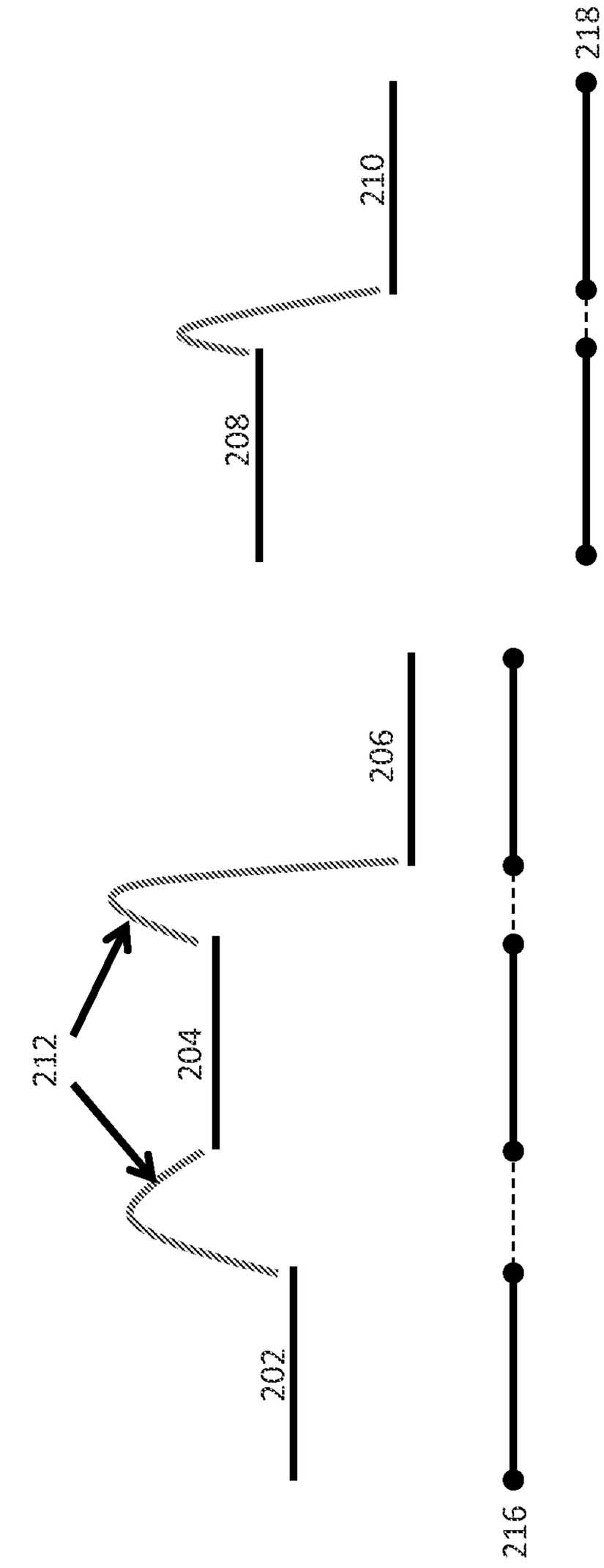


Figure 2

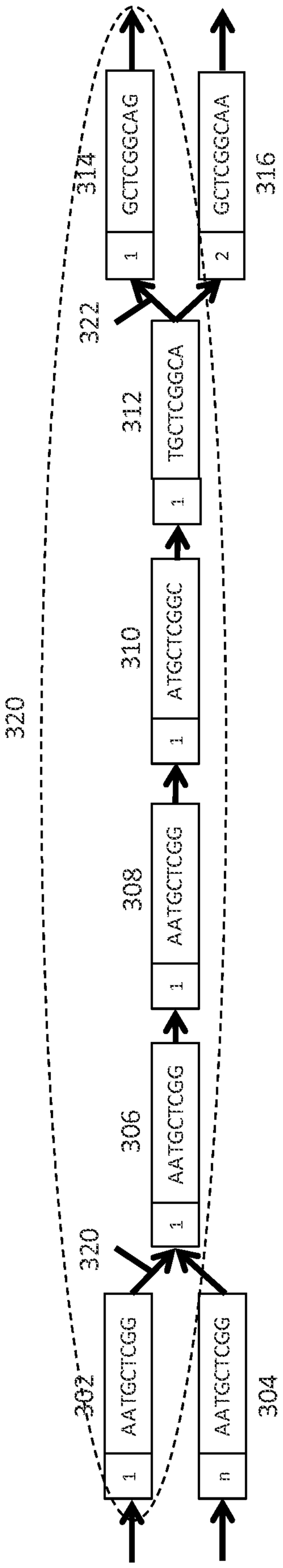

314
GCTCGGCAG
1
316
GCTCGGCAA
2
322
320
310
ATGCTCGGC
1
312
TGCTCGGCA
1
308
AATGCTCGG
1
306
AATGCTCGG
1
320
302
AATGCTCGG
1
AATGCTCGG
n
304

Figure 3

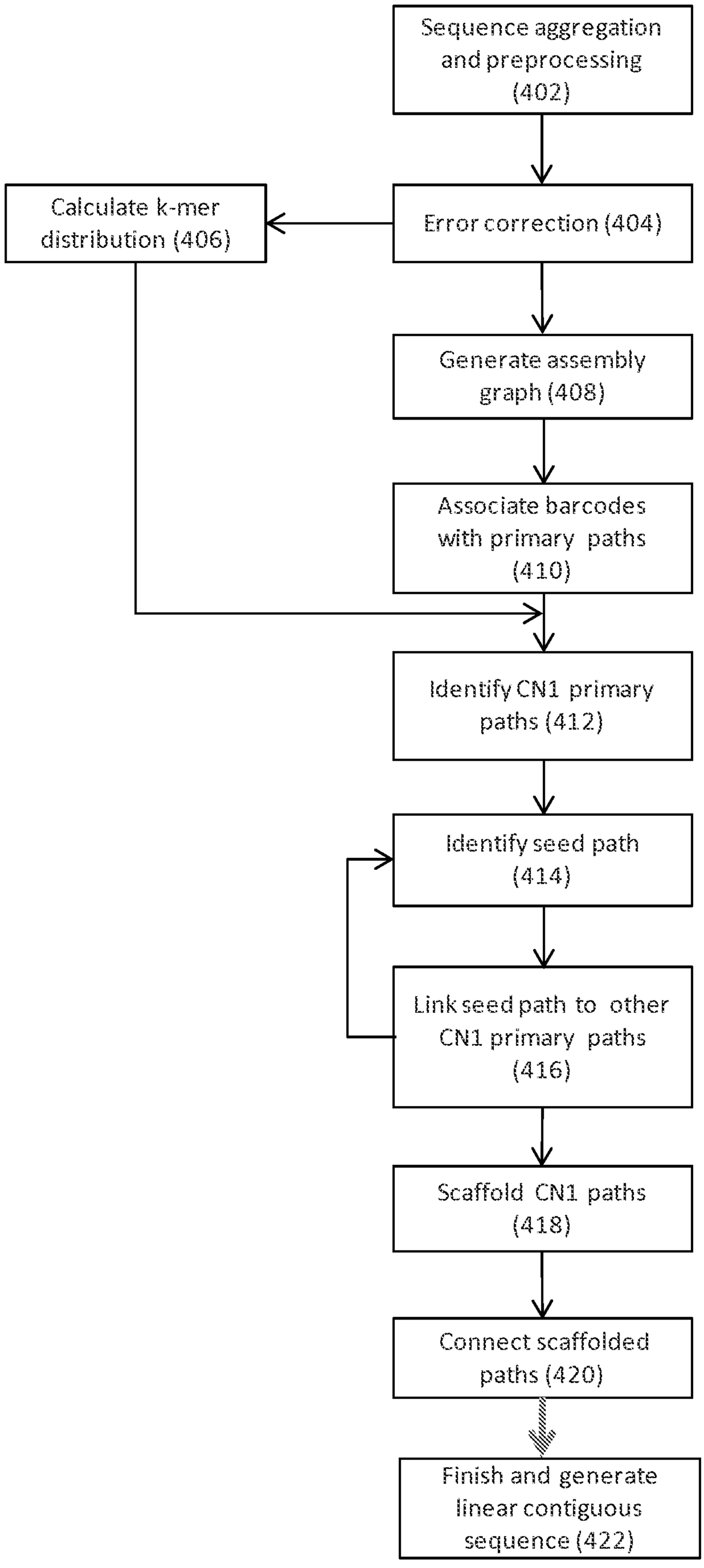

Sequence aggregation and preprocessing (402)
Error correction (404)
Calculate k-mer distribution (406)
Generate assembly graph (408)
Associate barcodes with primary paths (410)
Identify CN1 primary paths (412)
Identify seed path (414)
Link seed path to other CN1 primary paths (416)
Scaffold CN1 paths (418)
Connect scaffolded paths (420)
Finish and generate linear contiguous sequence (422)

Figure 4

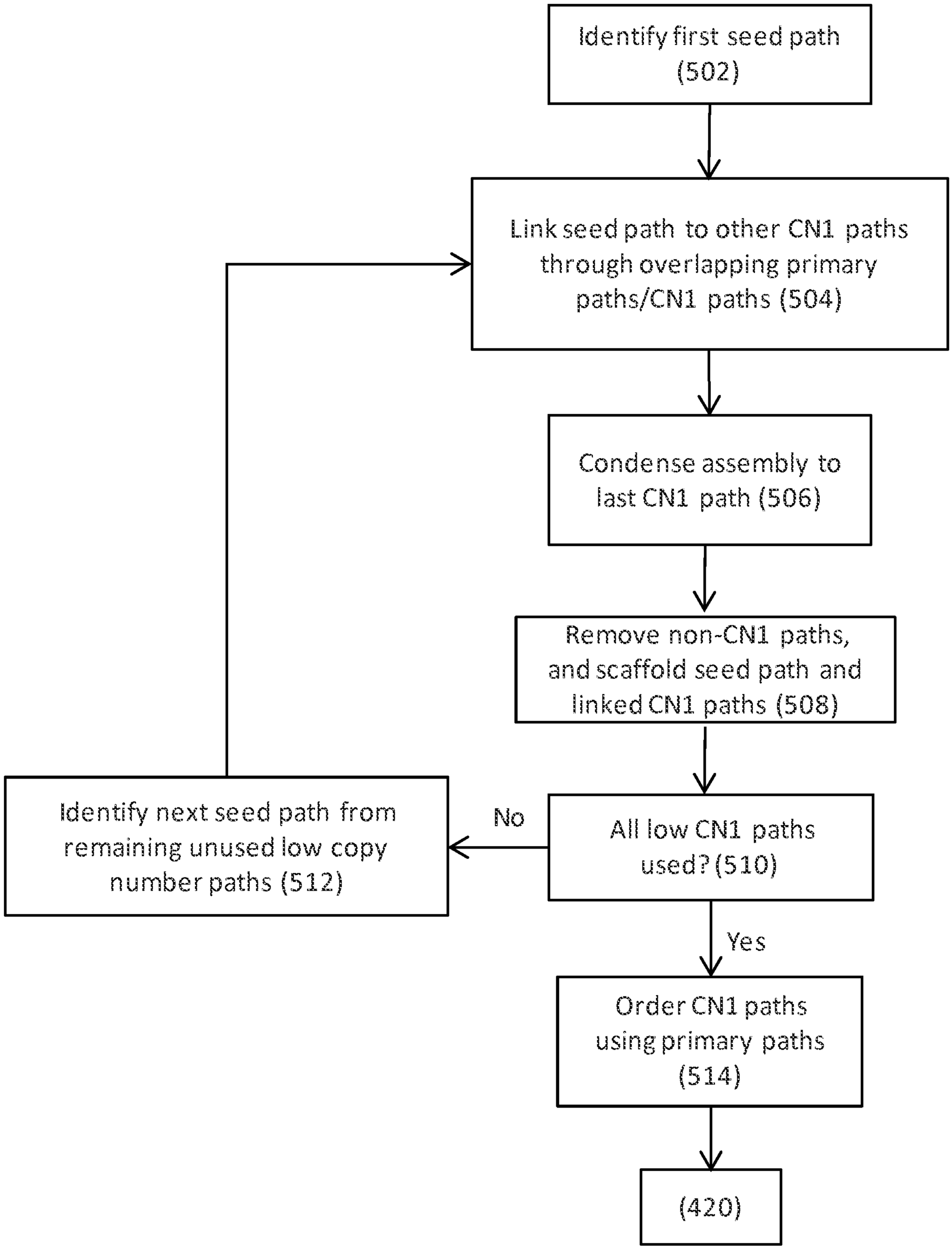

Identify first seed path (502)
Link seed path to other CN1 paths through overlapping primary paths/CN1 paths (504)
Condense assembly to last CN1 path (506)
Remove non-CN1 paths, and scaffold seed path and linked CN1 paths (508)
All low CN1 paths used? (510)
No
Identify next seed path from remaining unused low copy number paths (512)
Yes
Order CN1 paths using primary paths (514)
(420)

Figure 5

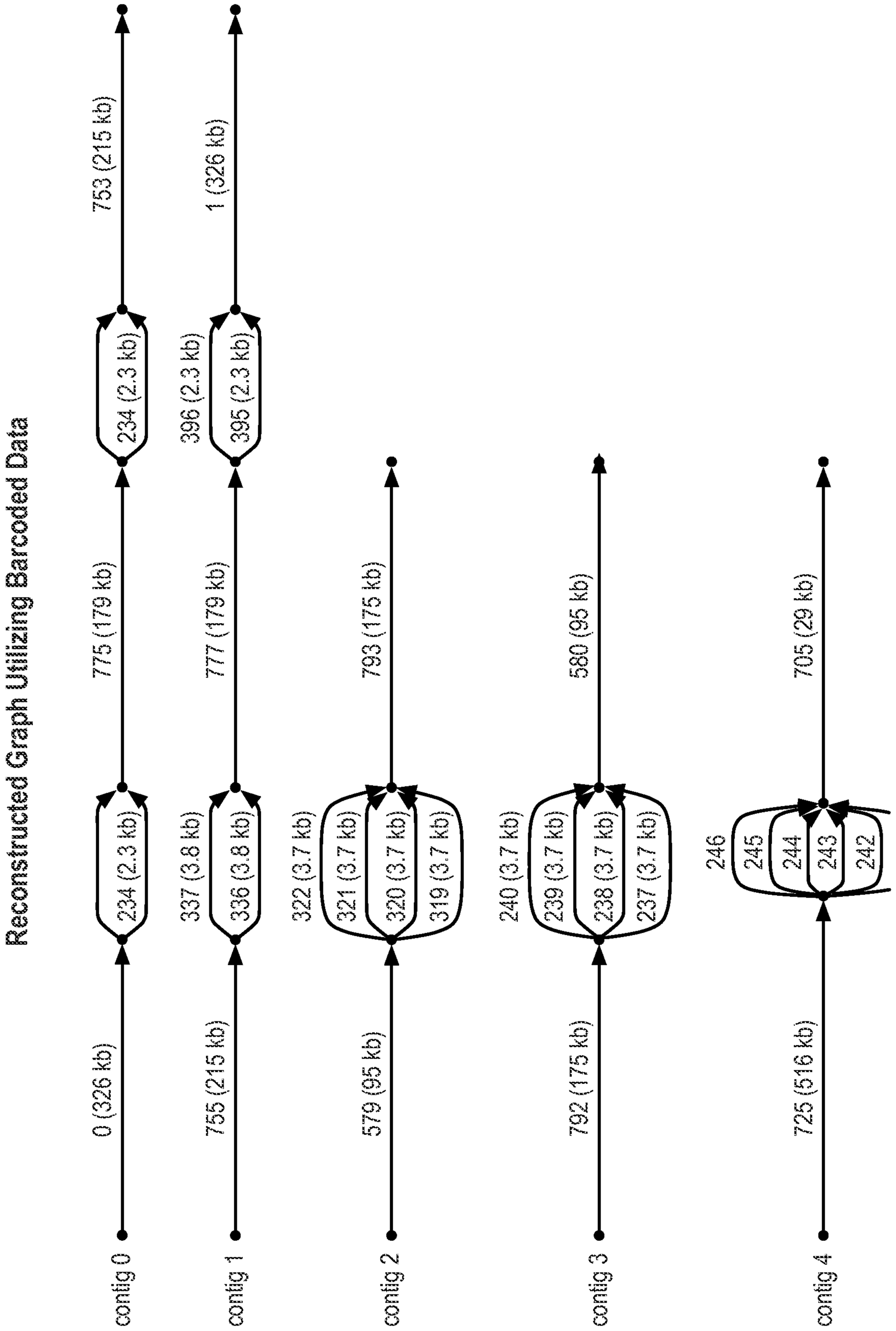

Reconstructed Graph Utilizing Barcoded Data
0 (326 kb)
775 (179 kb)
234 (2.3 kb)
753 (215 kb)
contig 0
755 (215 kb)
777 (179 kb)
396 (2.3 kb)
395 (2.3 kb)
1 (326 kb)
337 (3.8 kb)
336 (3.8 kb)
contig 1
579 (95 kb)
793 (175 kb)
322 (3.7 kb)
321 (3.7 kb)
320 (3.7 kb)
319 (3.7 kb)
contig 2
792 (175 kb)
580 (95 kb)
240 (3.7 kb)
239 (3.7 kb)
238 (3.7 kb)
237 (3.7 kb)
contig 3
725 (516 kb)
705 (29 kb)
246
245
244
243
242
contig 4

Figure 6A

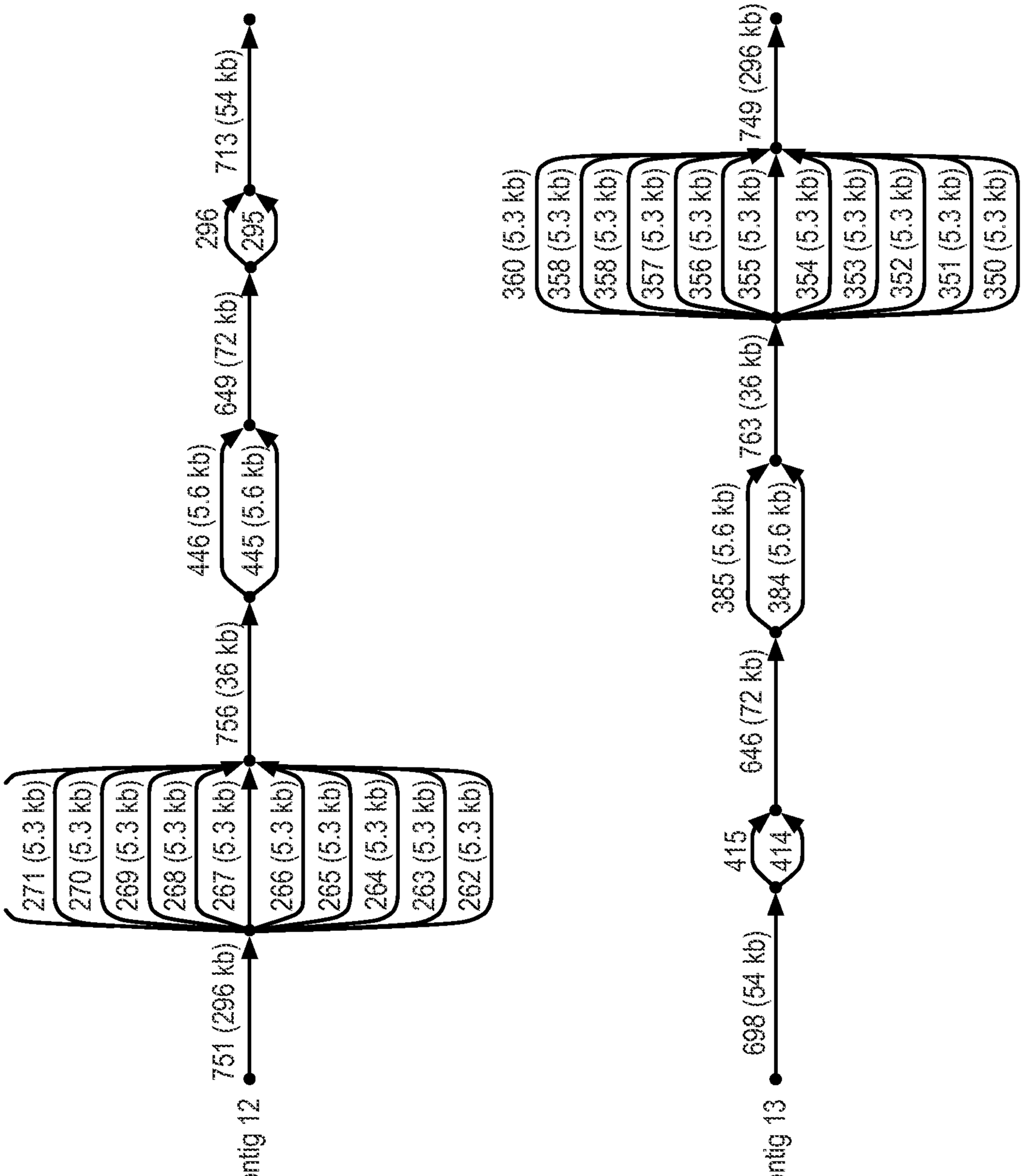

Reconstructed Graph Utilizing Barcoded Data (Continued)
contig 12
751 (296 kb)
271 (5.3 kb)
270 (5.3 kb)
269 (5.3 kb)
268 (5.3 kb)
267 (5.3 kb)
266 (5.3 kb)
265 (5.3 kb)
264 (5.3 kb)
263 (5.3 kb)
262 (5.3 kb)
756 (36 kb)
446 (5.6 kb)
445 (5.6 kb)
649 (72 kb)
296
295
713 (54 kb)
contig 13
698 (54 kb)
415
414
646 (72 kb)
385 (5.6 kb)
384 (5.6 kb)
763 (36 kb)
360 (5.3 kb)
358 (5.3 kb)
358 (5.3 kb)
357 (5.3 kb)
356 (5.3 kb)
355 (5.3 kb)
354 (5.3 kb)
353 (5.3 kb)
352 (5.3 kb)
351 (5.3 kb)
350 (5.3 kb)
749 (296 kb)

Figure 6B

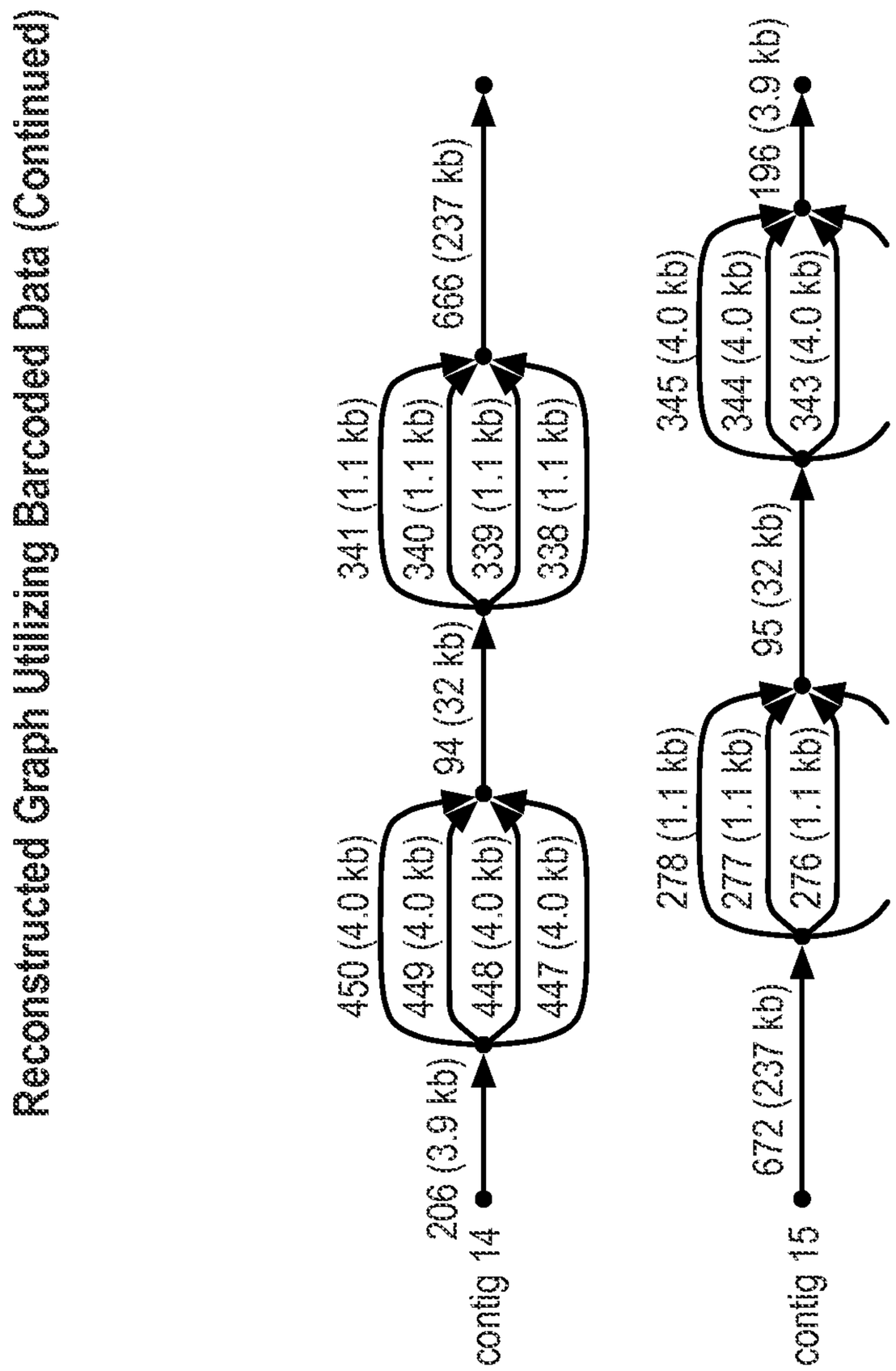

Reconstructed Graph Utilizing Barcoded Data (Continued)
206 (3.9 kb)
450 (4.0 kb)
449 (4.0 kb)
448 (4.0 kb)
447 (4.0 kb)
94 (32 kb)
341 (1.1 kb)
340 (1.1 kb)
339 (1.1 kb)
338 (1.1 kb)
666 (237 kb)
672 (237 kb)
278 (1.1 kb)
277 (1.1 kb)
276 (1.1 kb)
95 (32 kb)
345 (4.0 kb)
344 (4.0 kb)
343 (4.0 kb)
196 (3.9 kb)
Fig 14 continued
Fig 15 continued

Figure 6C

Reconstructed Graph Utilizing Barcoded Data (Continued)

Contig N50
570 Kilobases

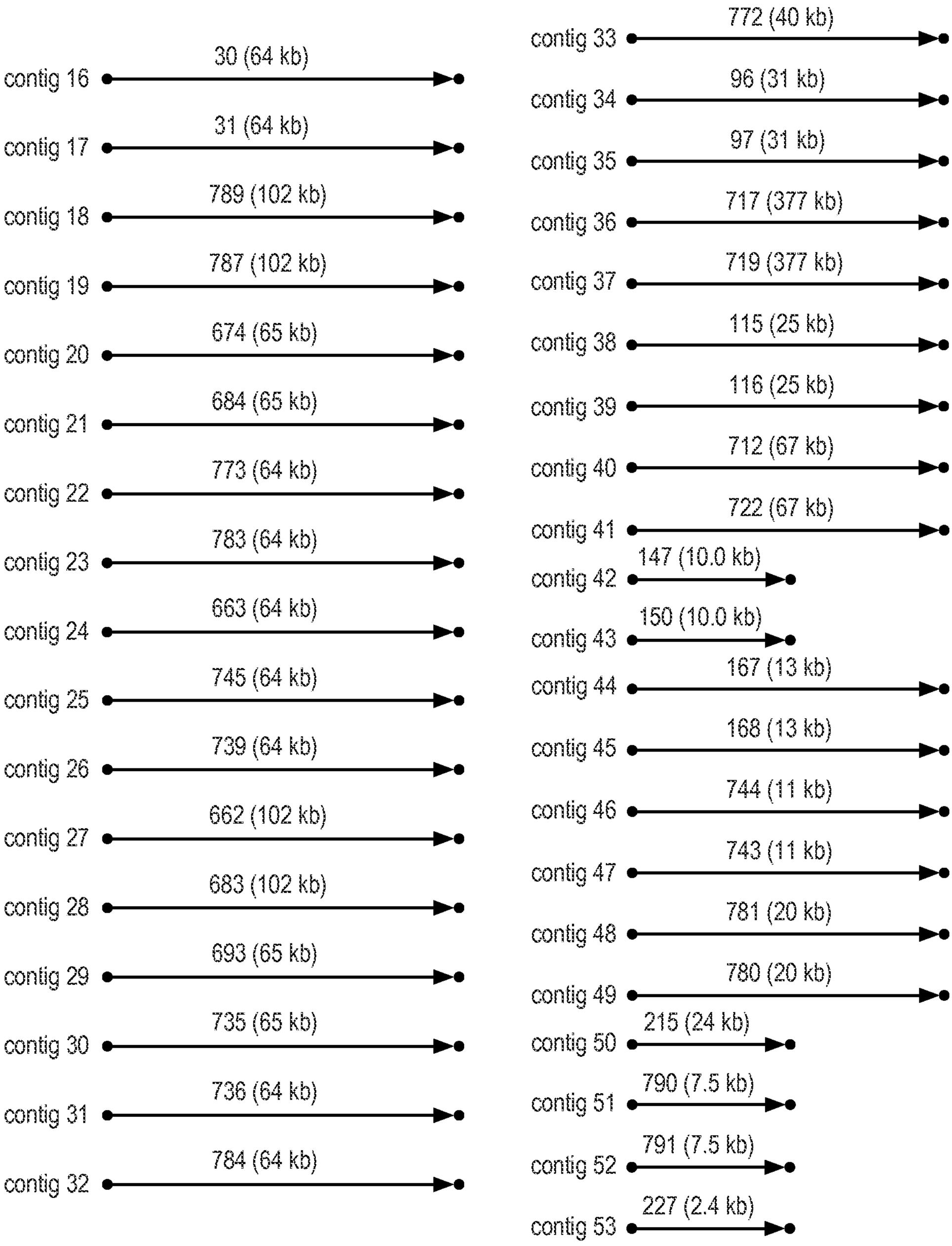


Figure 6D

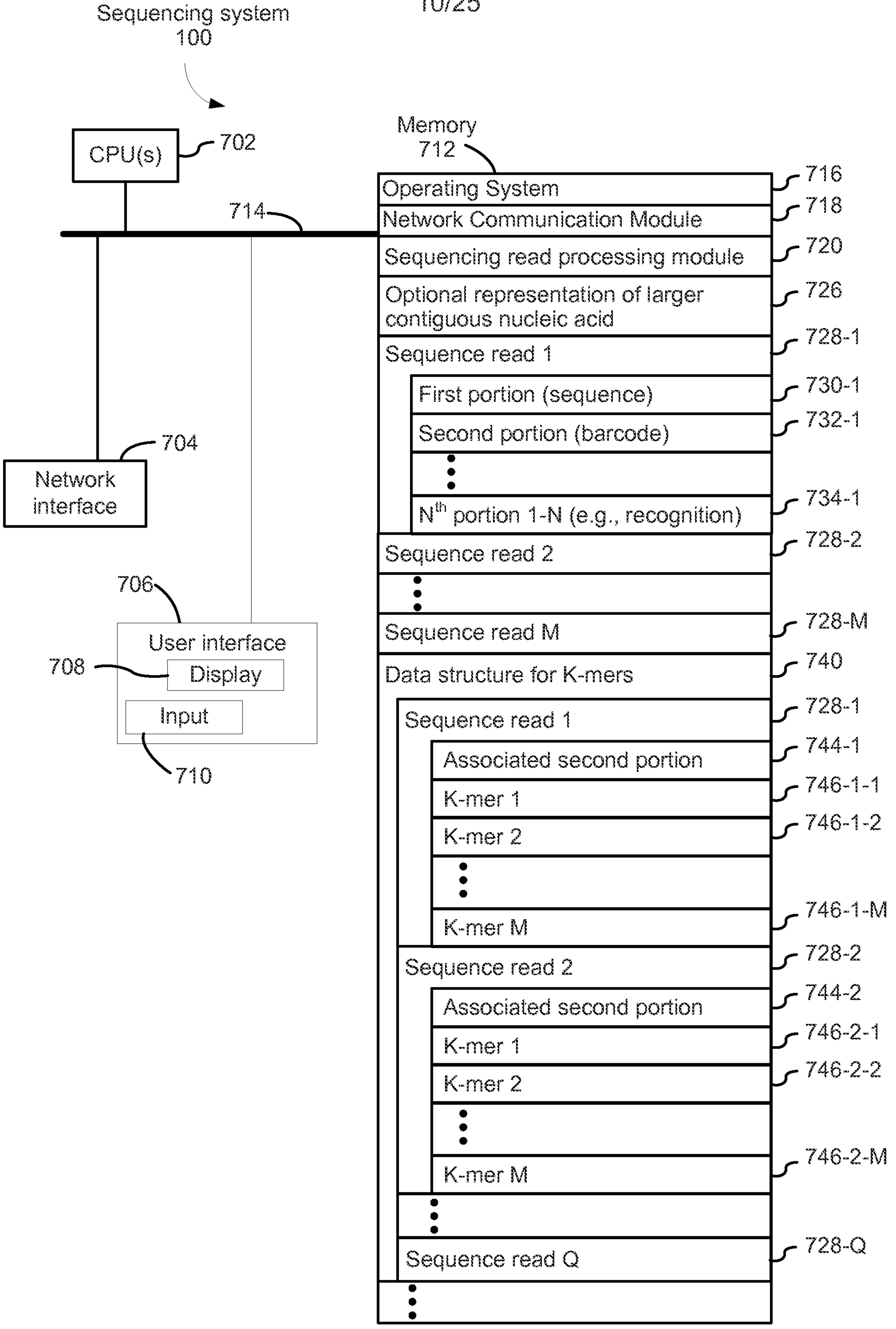

Sequencing system
100
CPU(s)
702
Memory
712
714
Operating System
716
Network Communication Module
718
Sequencing read processing module
720
Optional representation of larger contiguous nucleic acid
726
Sequence read 1
728-1
First portion (sequence)
730-1
Second portion (barcode)
732-1
Nth portion 1-N (e.g., recognition)
734-1
Sequence read 2
728-2
Sequence read M
728-M
Data structure for K-mers
740
Sequence read 1
728-1
Associated second portion
744-1
K-mer 1
746-1-1
K-mer 2
746-1-2
K-mer M
746-1-M
Sequence read 2
728-2
Associated second portion
744-2
K-mer 1
746-2-1
K-mer 2
746-2-2
K-mer M
746-2-M
Sequence read Q
728-Q
Network interface
704
User interface
706
Display
708
Input
710

Figure 7

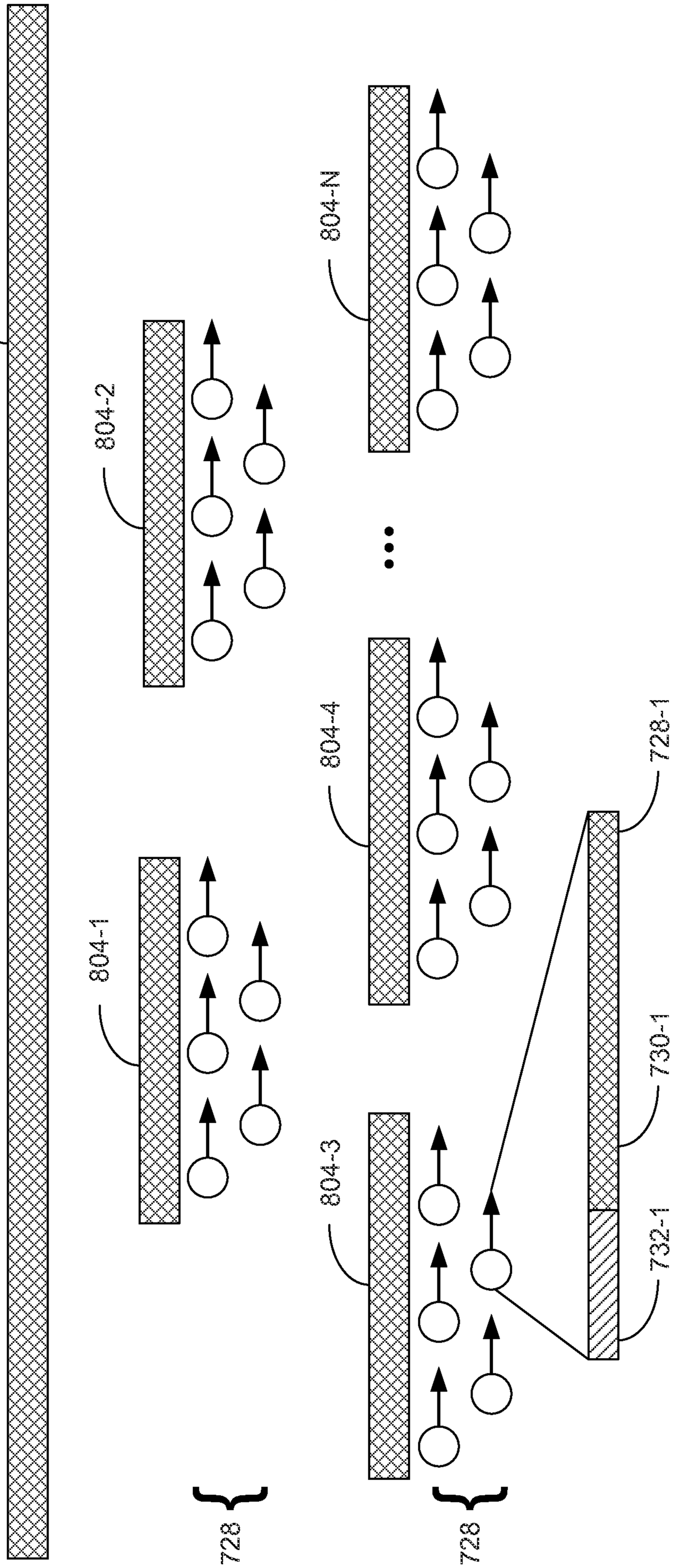


Figure 8

3-mer	lt(W)	W						
0	AAA	16	CAA	32	GAA	48	TAA
1	AAC	17	CAC	33	GAC	49	TAC
2	AAG	18	CAG	34	GAG	50	TAG
3	AAT	19	CAT	35	GAT	51	TAT
4	ACA	20	CCA	36	GCA	52	TCA
5	ACC	21	CCC	37	GCC	53	TCC
6	ACG	22	CCG	38	GCG	54	TCG
7	ACT	23	CCT	39	GCT	55	TCT
8	AGA	24	CGA	40	GGA	56	TGA
9	AGC	25	CGC	41	GGC	57	TGC
10	AGG	26	CGG	42	GGG	58	TGG
11	AGT	27	CGT	43	GGT	59	TGT
12	ATA	28	CTA	44	GTA	60	TTA
13	ATC	29	CTC	45	GTC	61	TTC
14	ATG	30	CTG	46	GTG	62	TTG
15	ATT	31	CTT	47	GTT	63	TTT

Figure 9

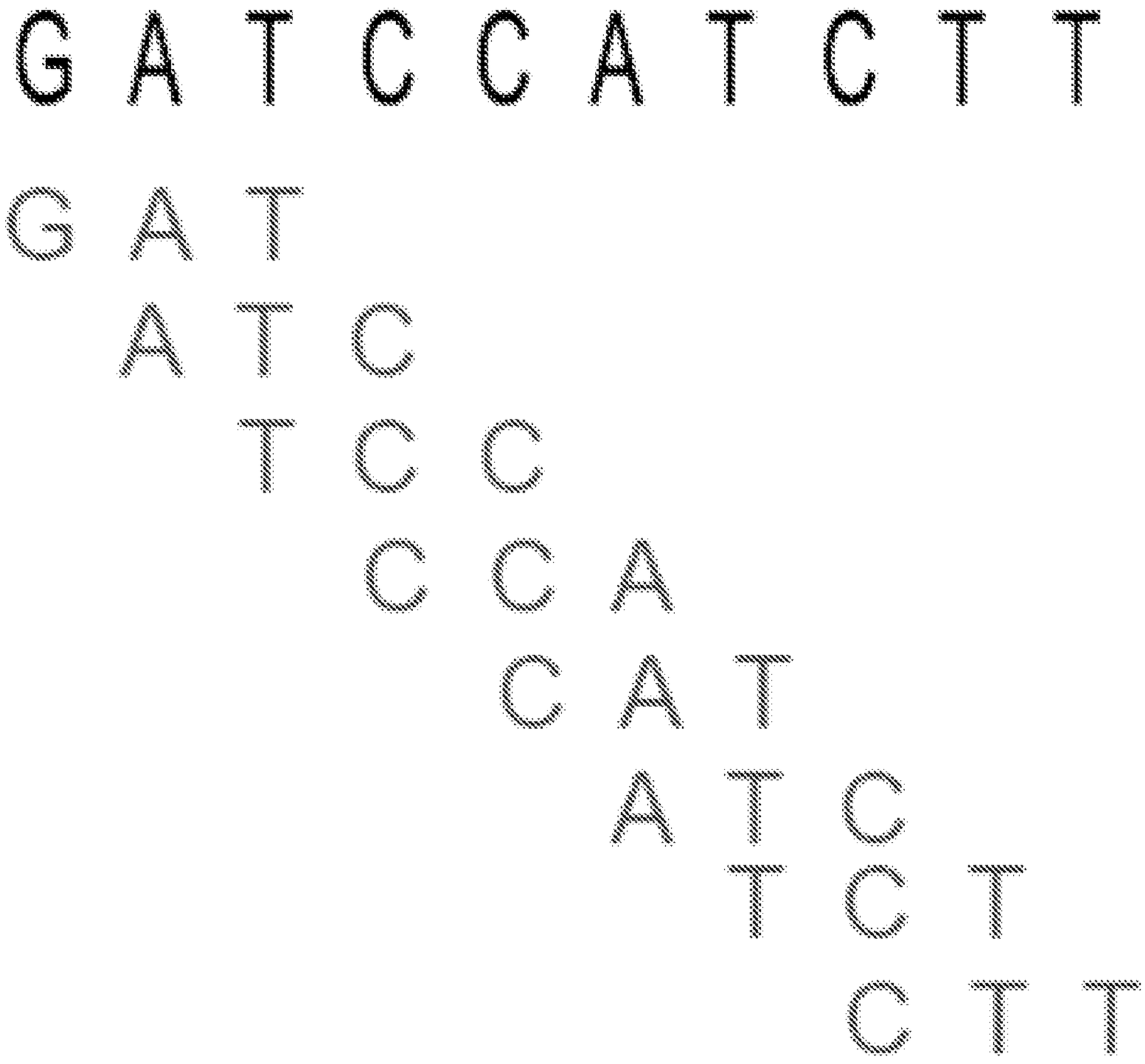


Figure 10

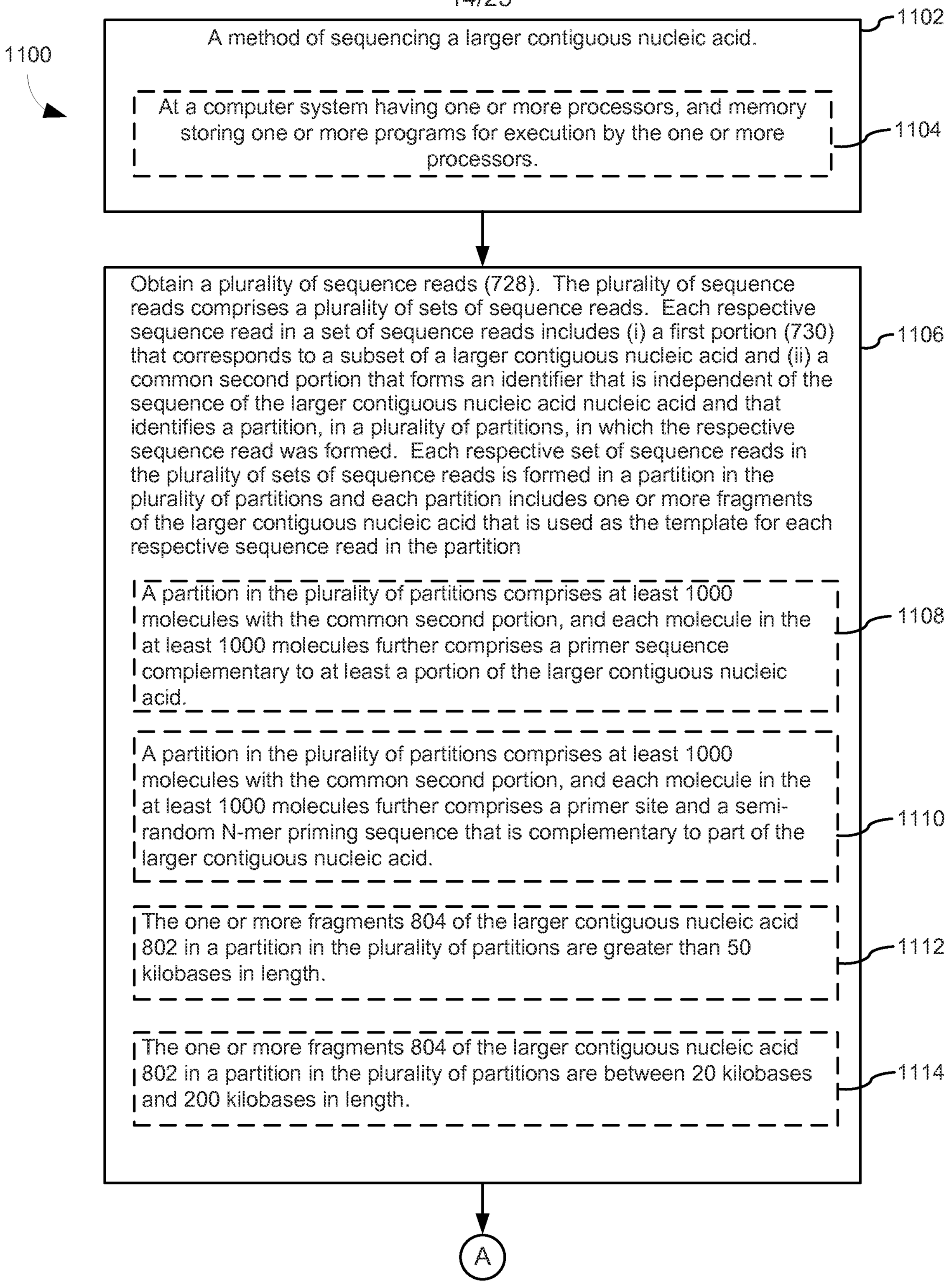

1100
A method of sequencing a larger contiguous nucleic acid.
1102
At a computer system having one or more processors, and memory storing one or more programs for execution by the one or more processors.
1104
Obtain a plurality of sequence reads (728). The plurality of sequence reads comprises a plurality of sets of sequence reads. Each respective sequence read in a set of sequence reads includes (i) a first portion (730) that corresponds to a subset of a larger contiguous nucleic acid and (ii) a common second portion that forms an identifier that is independent of the sequence of the larger contiguous nucleic acid nucleic acid and that identifies a partition, in a plurality of partitions, in which the respective sequence read was formed. Each respective set of sequence reads in the plurality of sets of sequence reads is formed in a partition in the plurality of partitions and each partition includes one or more fragments of the larger contiguous nucleic acid that is used as the template for each respective sequence read in the partition
1106
A partition in the plurality of partitions comprises at least 1000 molecules with the common second portion, and each molecule in the at least 1000 molecules further comprises a primer sequence complementary to at least a portion of the larger contiguous nucleic acid.
1108
A partition in the plurality of partitions comprises at least 1000 molecules with the common second portion, and each molecule in the at least 1000 molecules further comprises a primer site and a semi-random N-mer priming sequence that is complementary to part of the larger contiguous nucleic acid.
1110
The one or more fragments 804 of the larger contiguous nucleic acid 802 in a partition in the plurality of partitions are greater than 50 kilobases in length.
1112
The one or more fragments 804 of the larger contiguous nucleic acid 802 in a partition in the plurality of partitions are between 20 kilobases and 200 kilobases in length.
1114
A

Figure 11A

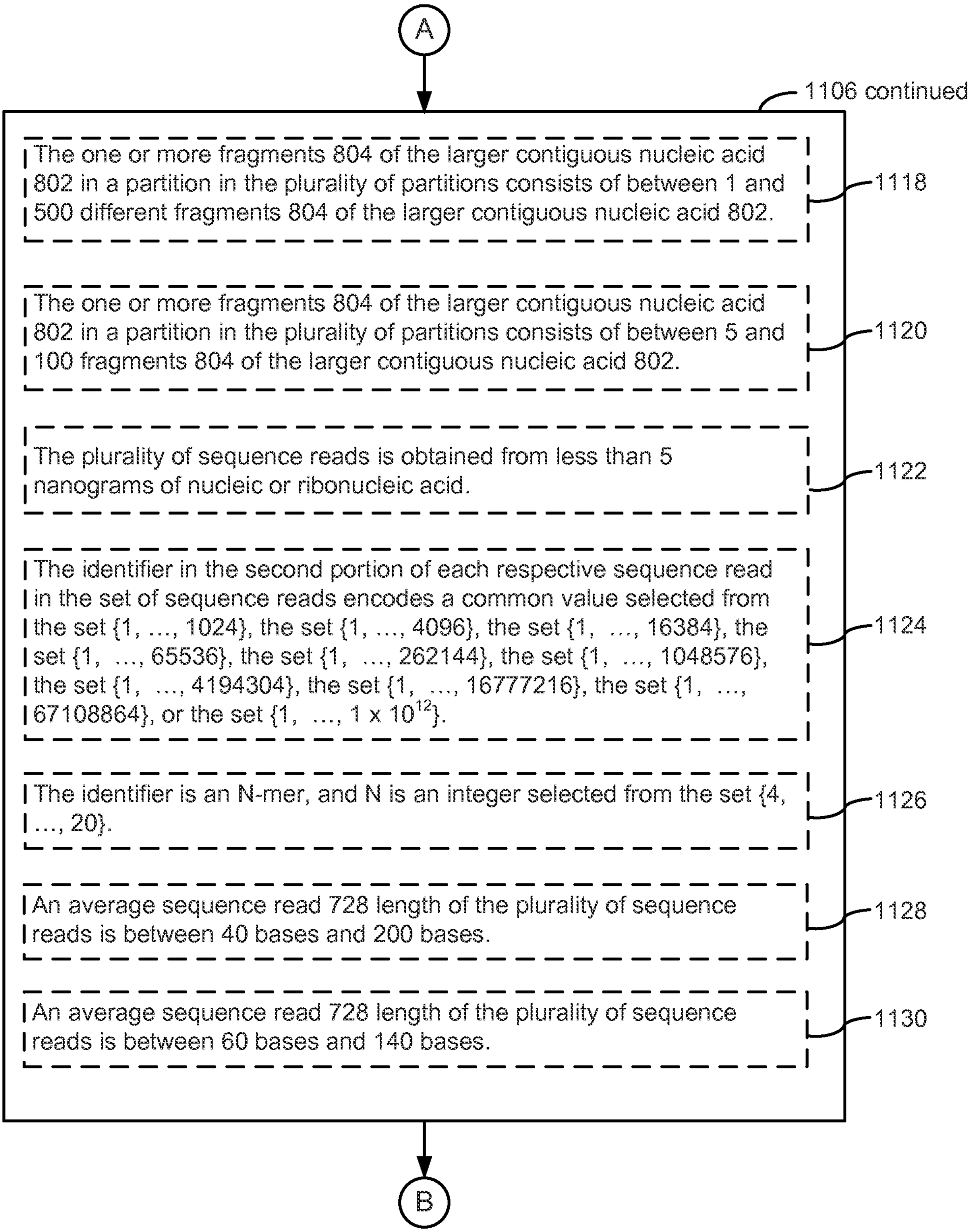

A
1106 continued
The one or more fragments 804 of the larger contiguous nucleic acid 802 in a partition in the plurality of partitions consists of between 1 and 500 different fragments 804 of the larger contiguous nucleic acid 802.
1118
The one or more fragments 804 of the larger contiguous nucleic acid 802 in a partition in the plurality of partitions consists of between 5 and 100 fragments 804 of the larger contiguous nucleic acid 802.
1120
The plurality of sequence reads is obtained from less than 5 nanograms of nucleic or ribonucleic acid.
1122
The identifier in the second portion of each respective sequence read in the set of sequence reads encodes a common value selected from the set {1, …, 1024}, the set {1, …, 4096}, the set {1, …, 16384}, the set {1, …, 65536}, the set {1, …, 262144}, the set {1, …, 1048576}, the set {1, …, 4194304}, the set {1, …, 16777216}, the set {1, …, 67108864}, or the set {1, …, 1 x 10^12}.
1124
The identifier is an N-mer, and N is an integer selected from the set {4, …, 20}.
1126
An average sequence read 728 length of the plurality of sequence reads is between 40 bases and 200 bases.
1128
An average sequence read 728 length of the plurality of sequence reads is between 60 bases and 140 bases.
1130
B

Figure 11B

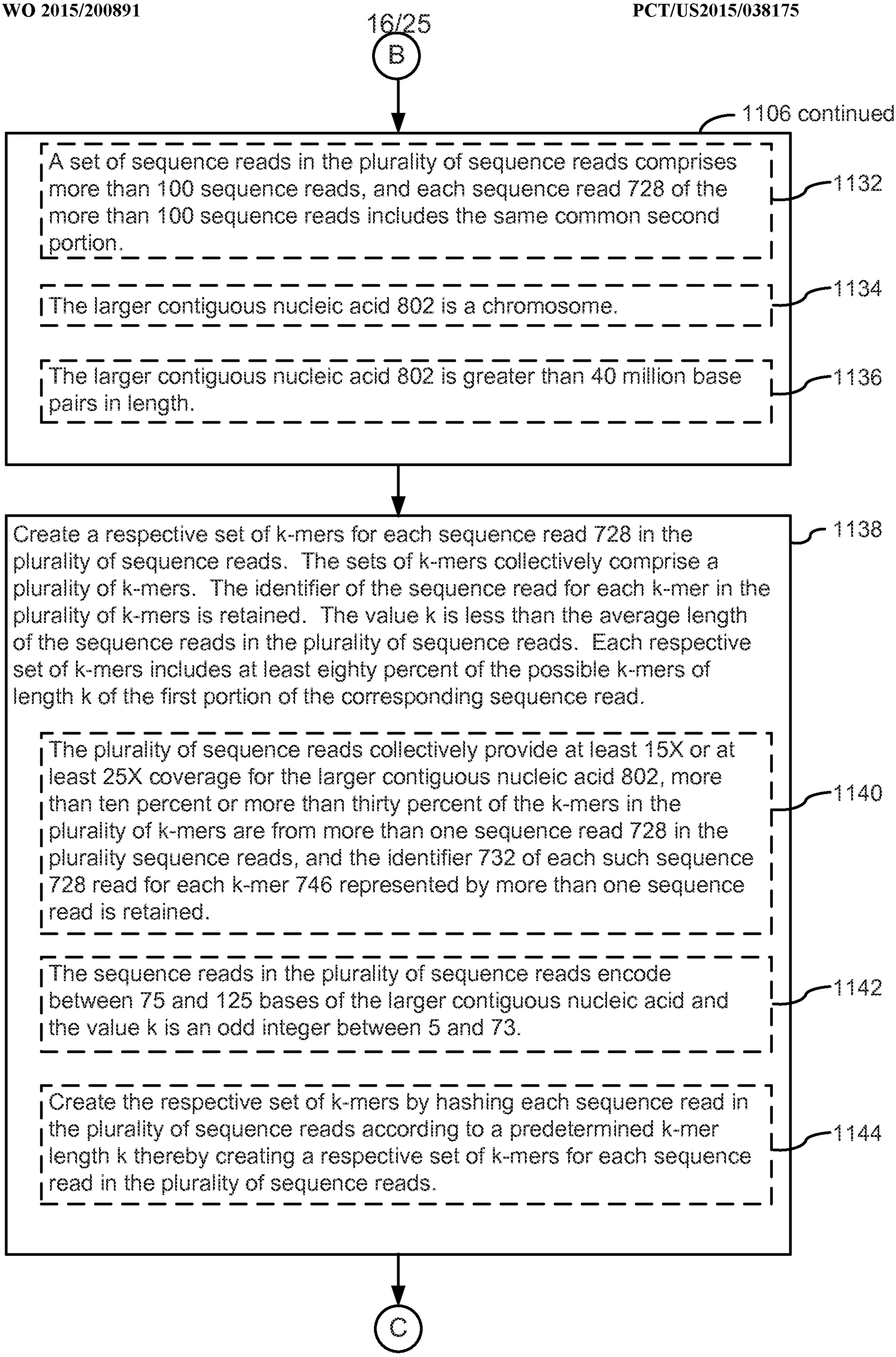

B
1106 continued
A set of sequence reads in the plurality of sequence reads comprises more than 100 sequence reads, and each sequence read 728 of the more than 100 sequence reads includes the same common second portion.
1132
The larger contiguous nucleic acid 802 is a chromosome.
1134
The larger contiguous nucleic acid 802 is greater than 40 million base pairs in length.
1136
Create a respective set of k-mers for each sequence read 728 in the plurality of sequence reads. The sets of k-mers collectively comprise a plurality of k-mers. The identifier of the sequence read for each k-mer in the plurality of k-mers is retained. The value k is less than the average length of the sequence reads in the plurality of sequence reads. Each respective set of k-mers includes at least eighty percent of the possible k-mers of length k of the first portion of the corresponding sequence read.
1138
The plurality of sequence reads collectively provide at least 15X or at least 25X coverage for the larger contiguous nucleic acid 802, more than ten percent or more than thirty percent of the k-mers in the plurality of k-mers are from more than one sequence read 728 in the plurality sequence reads, and the identifier 732 of each such sequence 728 read for each k-mer 746 represented by more than one sequence read is retained.
1140
The sequence reads in the plurality of sequence reads encode between 75 and 125 bases of the larger contiguous nucleic acid and the value k is an odd integer between 5 and 73.
1142
Create the respective set of k-mers by hashing each sequence read in the plurality of sequence reads according to a predetermined k-mer length k thereby creating a respective set of k-mers for each sequence read in the plurality of sequence reads.
1144
C

Figure 11C

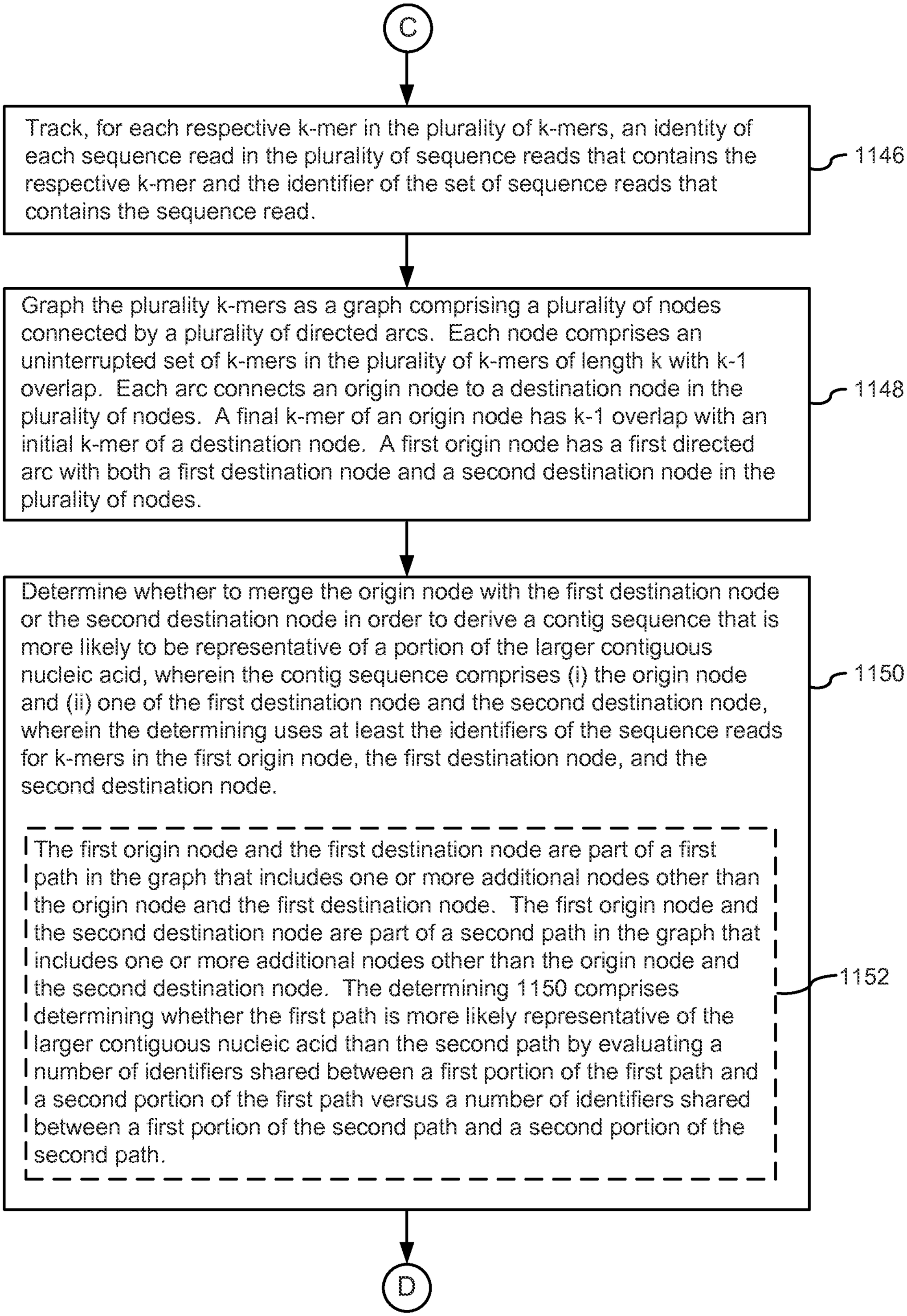

C
Track, for each respective k-mer in the plurality of k-mers, an identity of each sequence read in the plurality of sequence reads that contains the respective k-mer and the identifier of the set of sequence reads that contains the sequence read.
1146
Graph the plurality k-mers as a graph comprising a plurality of nodes connected by a plurality of directed arcs. Each node comprises an uninterrupted set of k-mers in the plurality of k-mers of length k with k-1 overlap. Each arc connects an origin node to a destination node in the plurality of nodes. A final k-mer of an origin node has k-1 overlap with an initial k-mer of a destination node. A first origin node has a first directed arc with both a first destination node and a second destination node in the plurality of nodes.
1148
Determine whether to merge the origin node with the first destination node or the second destination node in order to derive a contig sequence that is more likely to be representative of a portion of the larger contiguous nucleic acid, wherein the contig sequence comprises (i) the origin node and (ii) one of the first destination node and the second destination node, wherein the determining uses at least the identifiers of the sequence reads for k-mers in the first origin node, the first destination node, and the second destination node.
1150
The first origin node and the first destination node are part of a first path in the graph that includes one or more additional nodes other than the origin node and the first destination node. The first origin node and the second destination node are part of a second path in the graph that includes one or more additional nodes other than the origin node and the second destination node. The determining 1150 comprises determining whether the first path is more likely representative of the larger contiguous nucleic acid than the second path by evaluating a number of identifiers shared between a first portion of the first path and a second portion of the first path versus a number of identifiers shared between a first portion of the second path and a second portion of the second path.
1152
D

Figure 11D

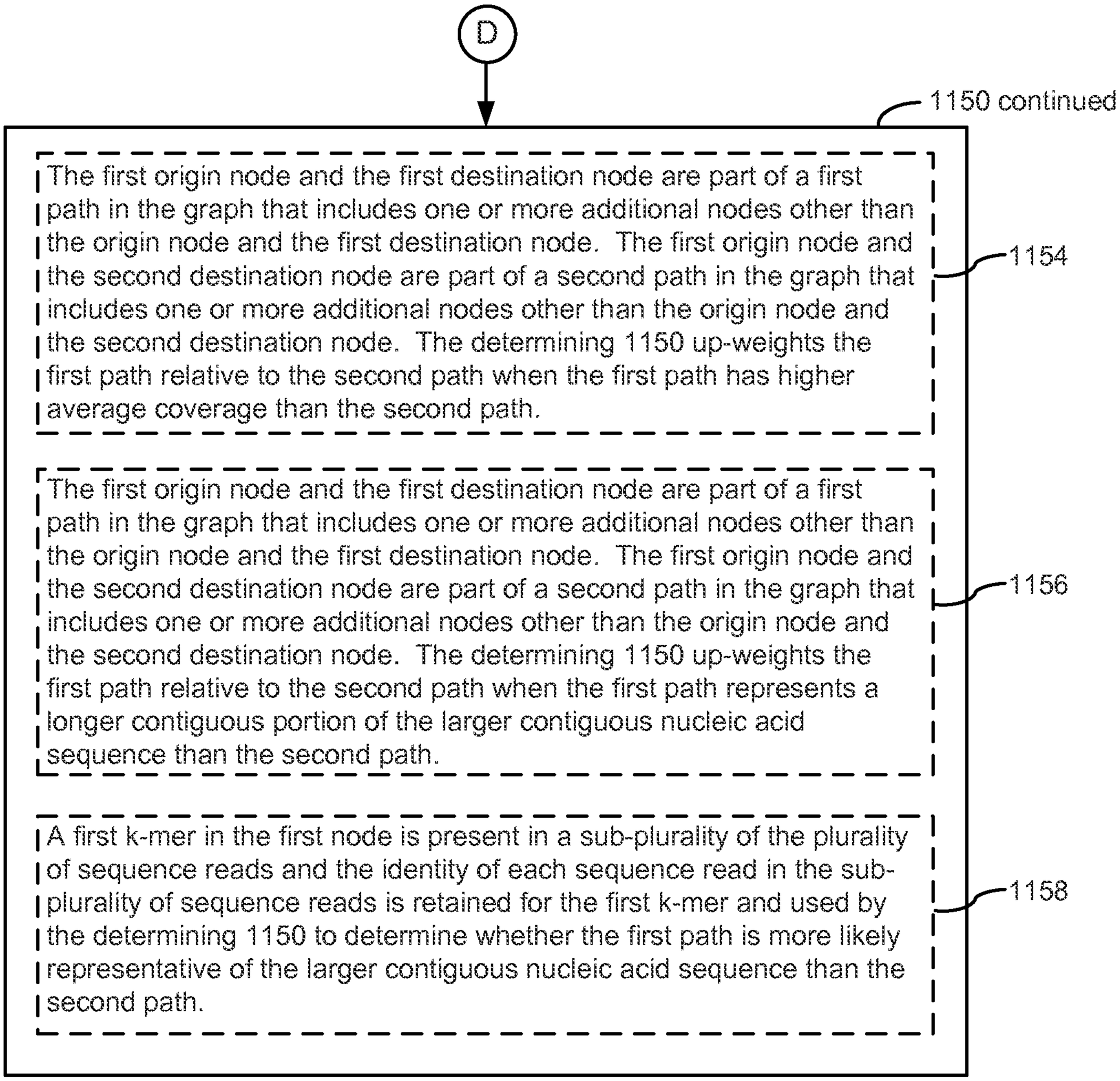

D
1150 continued
The first origin node and the first destination node are part of a first path in the graph that includes one or more additional nodes other than the origin node and the first destination node. The first origin node and the second destination node are part of a second path in the graph that includes one or more additional nodes other than the origin node and the second destination node. The determining 1150 up-weights the first path relative to the second path when the first path has higher average coverage than the second path.
1154
The first origin node and the first destination node are part of a first path in the graph that includes one or more additional nodes other than the origin node and the first destination node. The first origin node and the second destination node are part of a second path in the graph that includes one or more additional nodes other than the origin node and the second destination node. The determining 1150 up-weights the first path relative to the second path when the first path represents a longer contiguous portion of the larger contiguous nucleic acid sequence than the second path.
1156
A first k-mer in the first node is present in a sub-plurality of the plurality of sequence reads and the identity of each sequence read in the sub-plurality of sequence reads is retained for the first k-mer and used by the determining 1150 to determine whether the first path is more likely representative of the larger contiguous nucleic acid sequence than the second path.
1158

Figure 11E

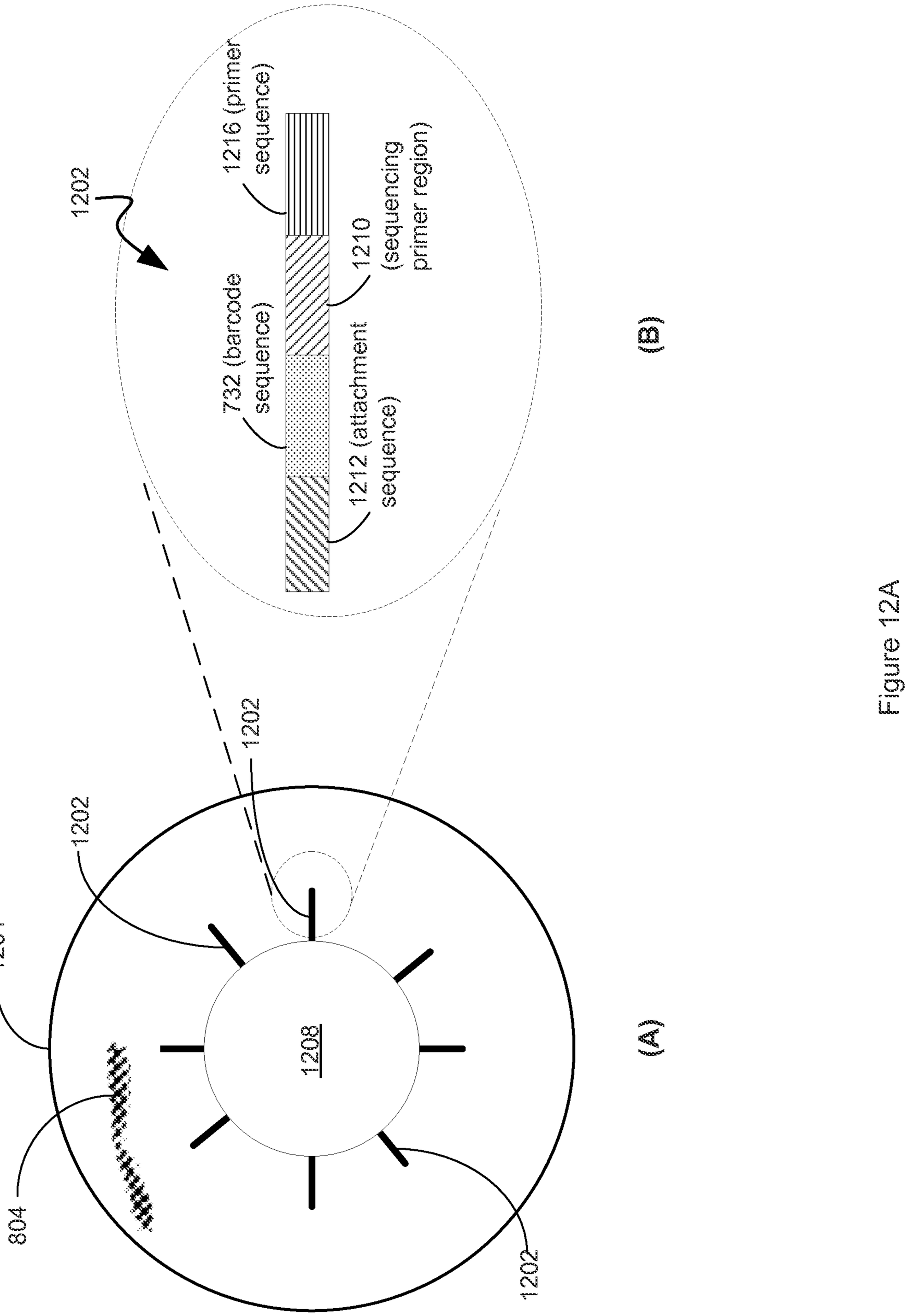

1202
1216 (primer sequence)
1210 (sequencing primer region)
732 (barcode sequence)
1212 (attachment sequence)
(B)
1202
1202
1204
1208
804
1202
(A)

Figure 12A

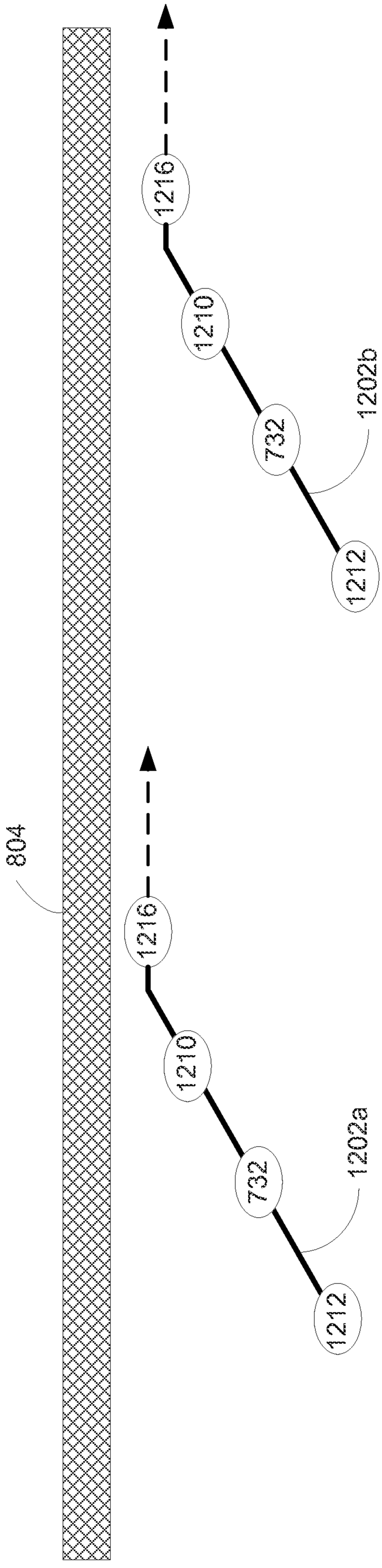


Figure 12B

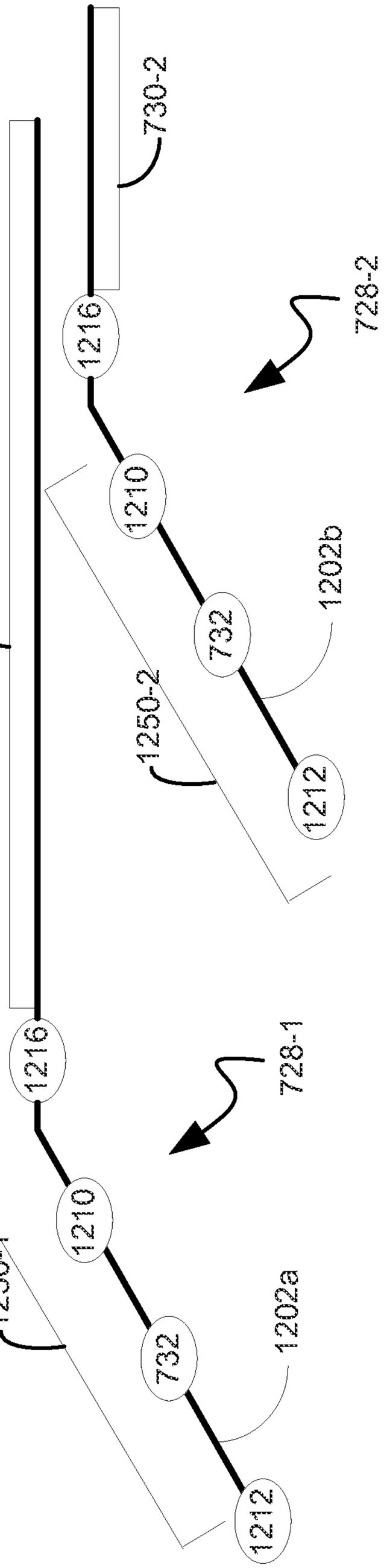


Figure 12C

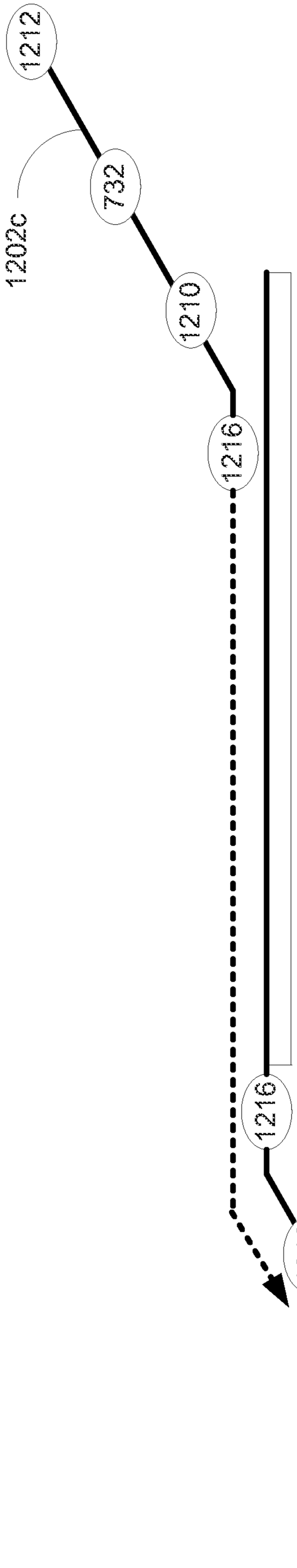


Figure 12D

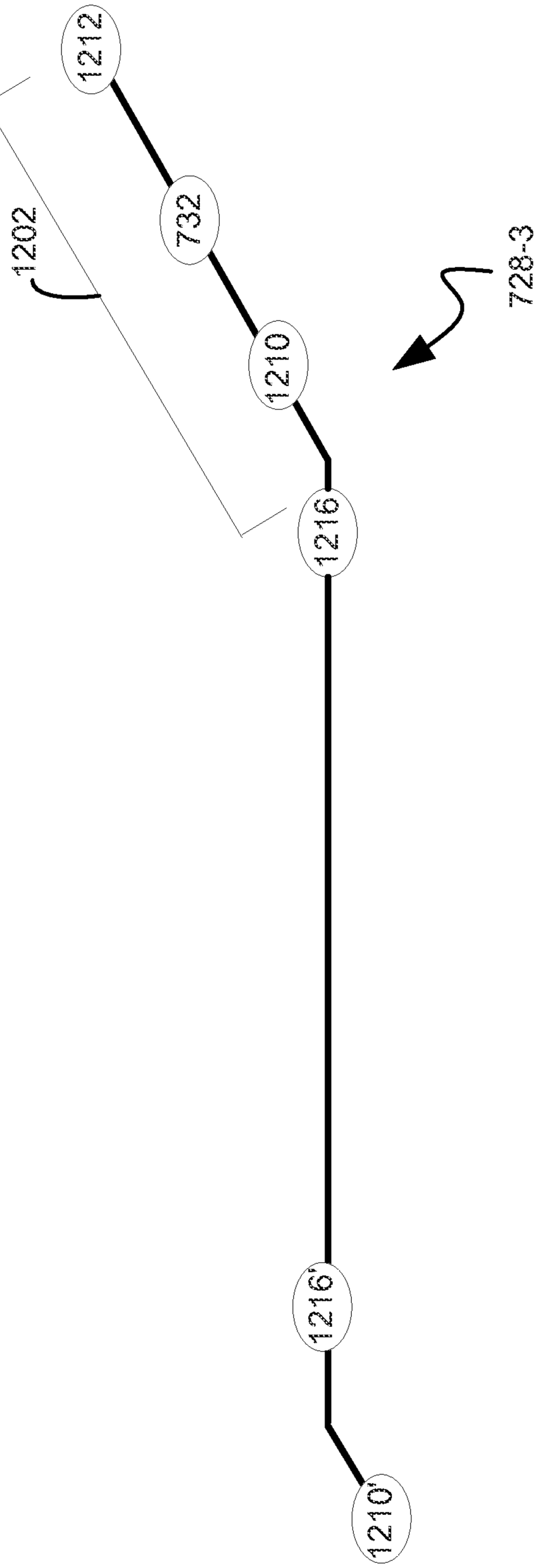
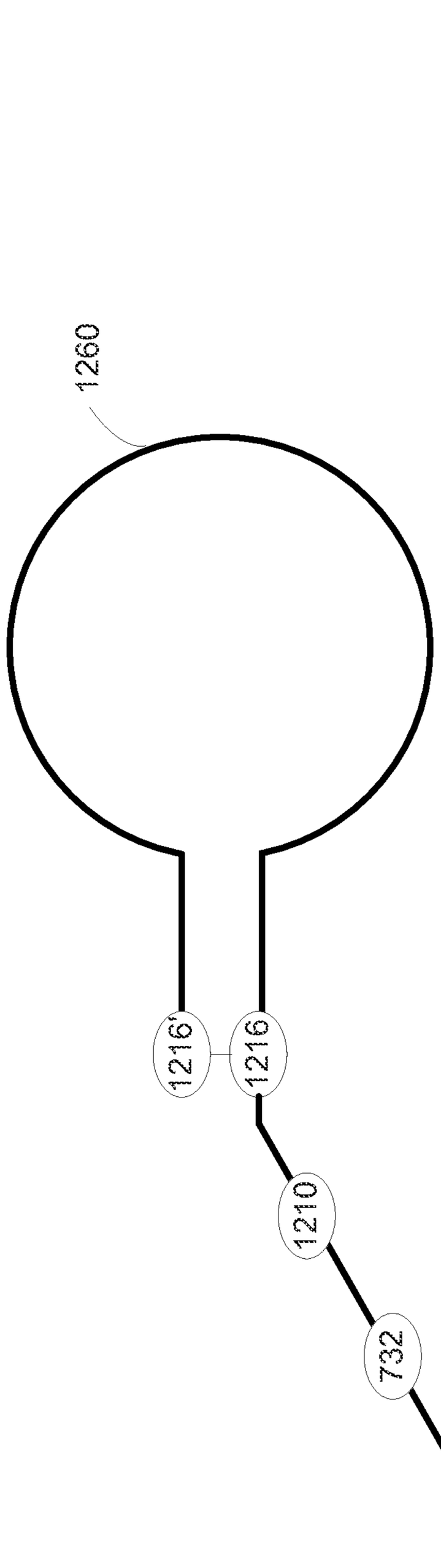


Figure 12E



Figures 12F

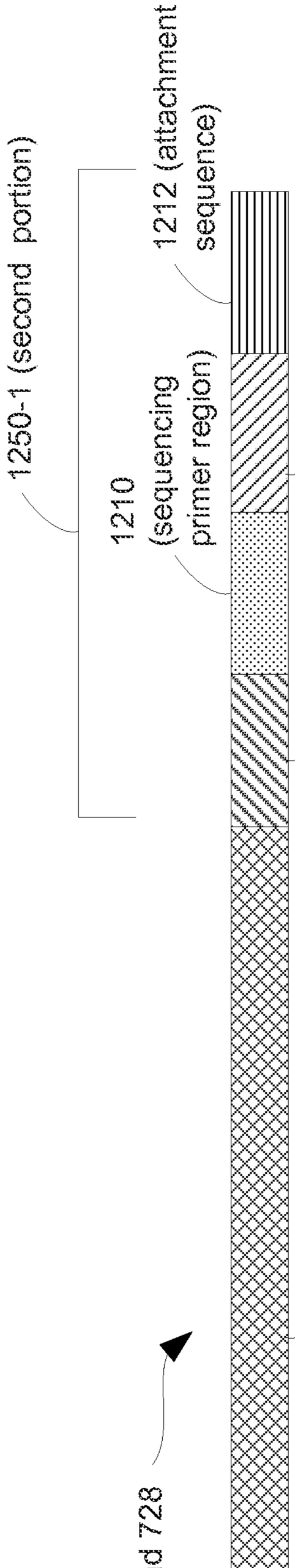


Figure 12G

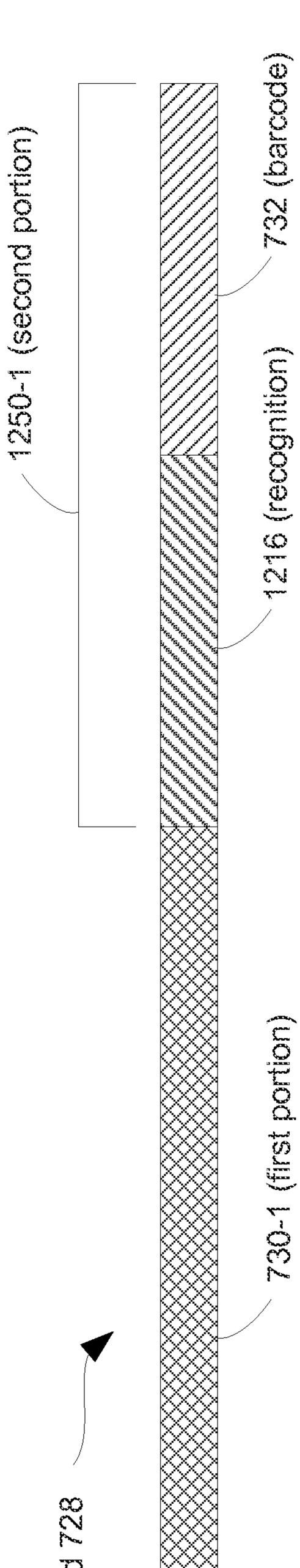


Figure 12H

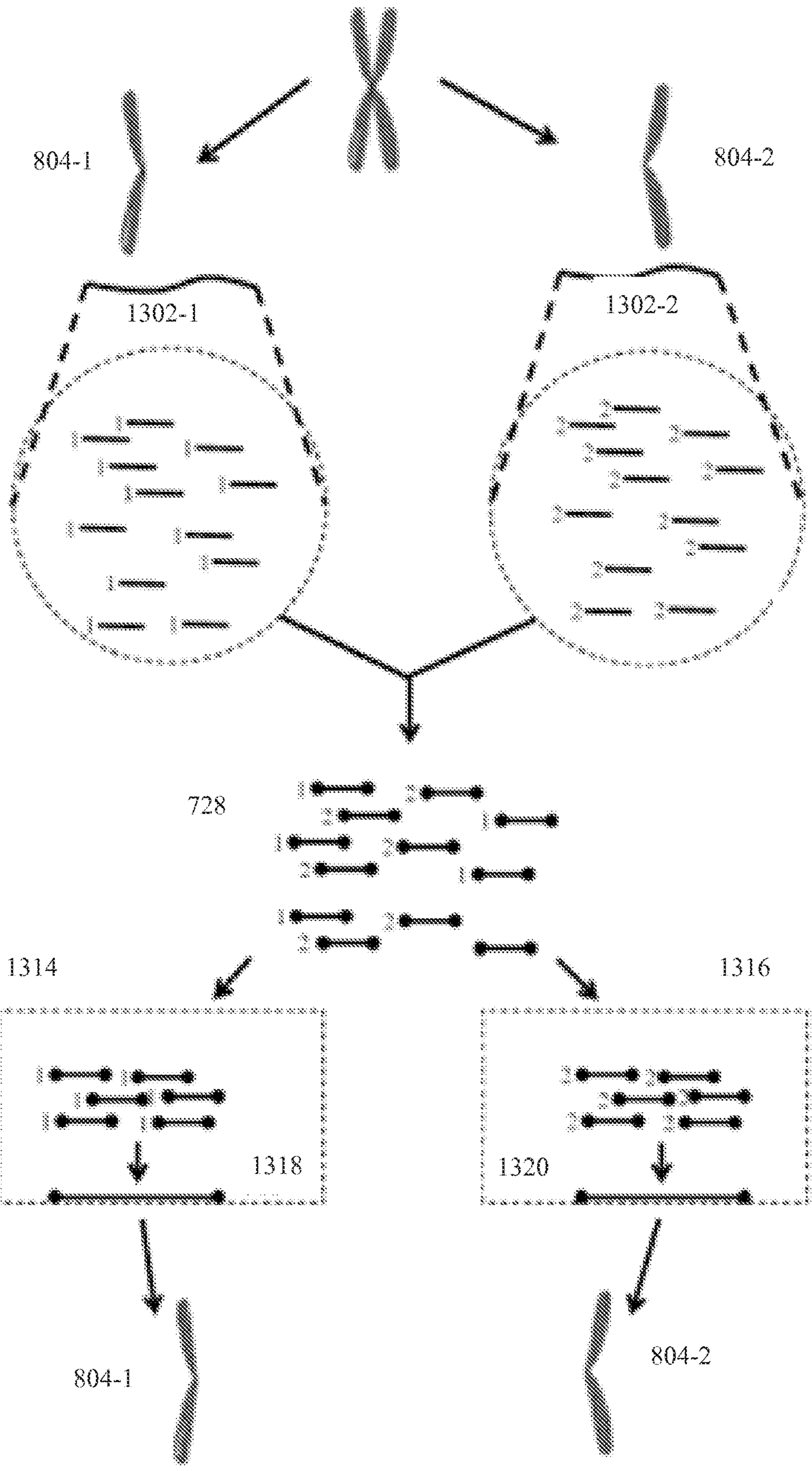


Figure 13

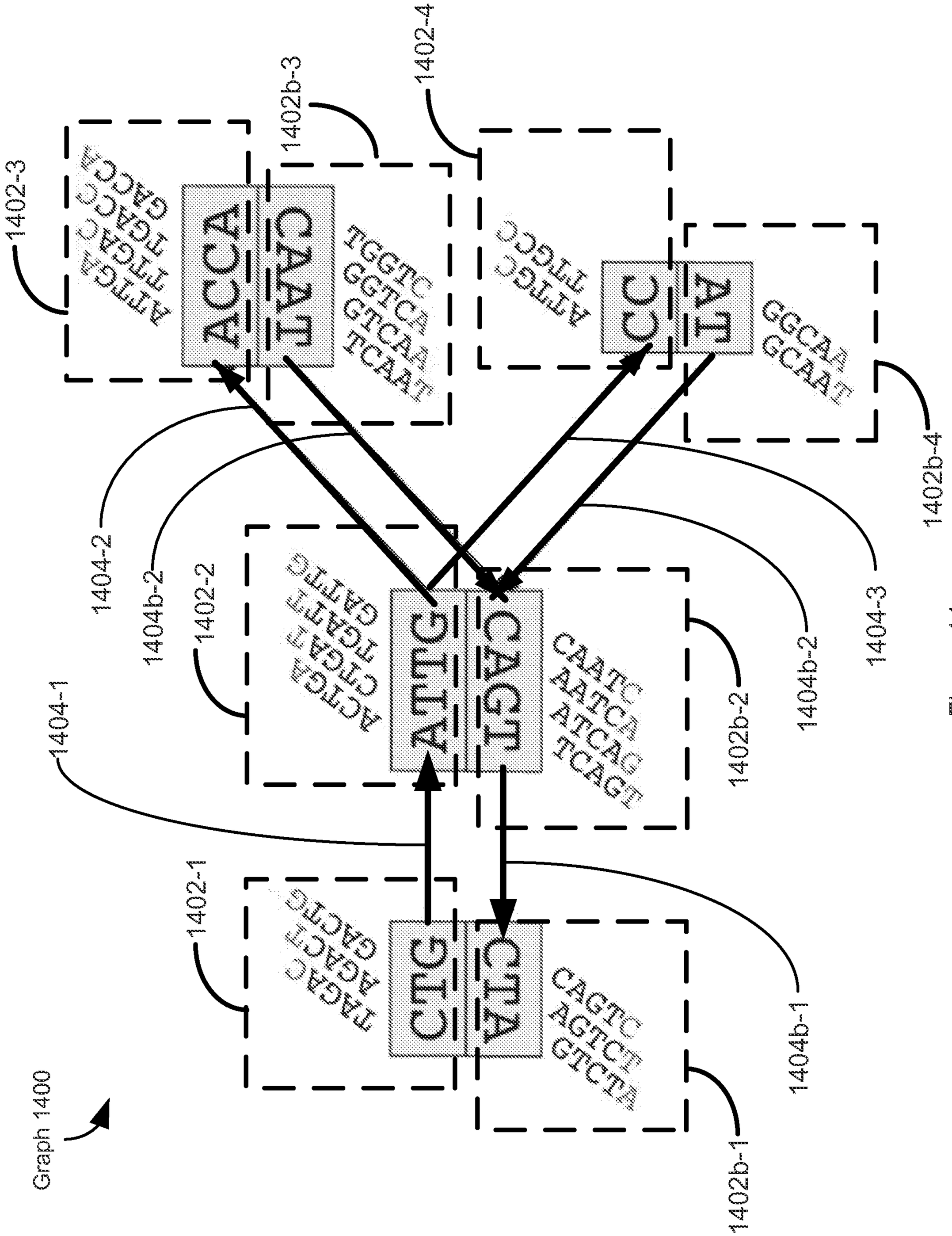

Graph 1400
1402-1
1402-2
1402-3
1402-4
1402b-1
1402b-2
1402b-3
1402b-4
1404-1
1404-2
1404-3
1404b-1
1404b-2
1404b-3
1404b-4

Figure 14

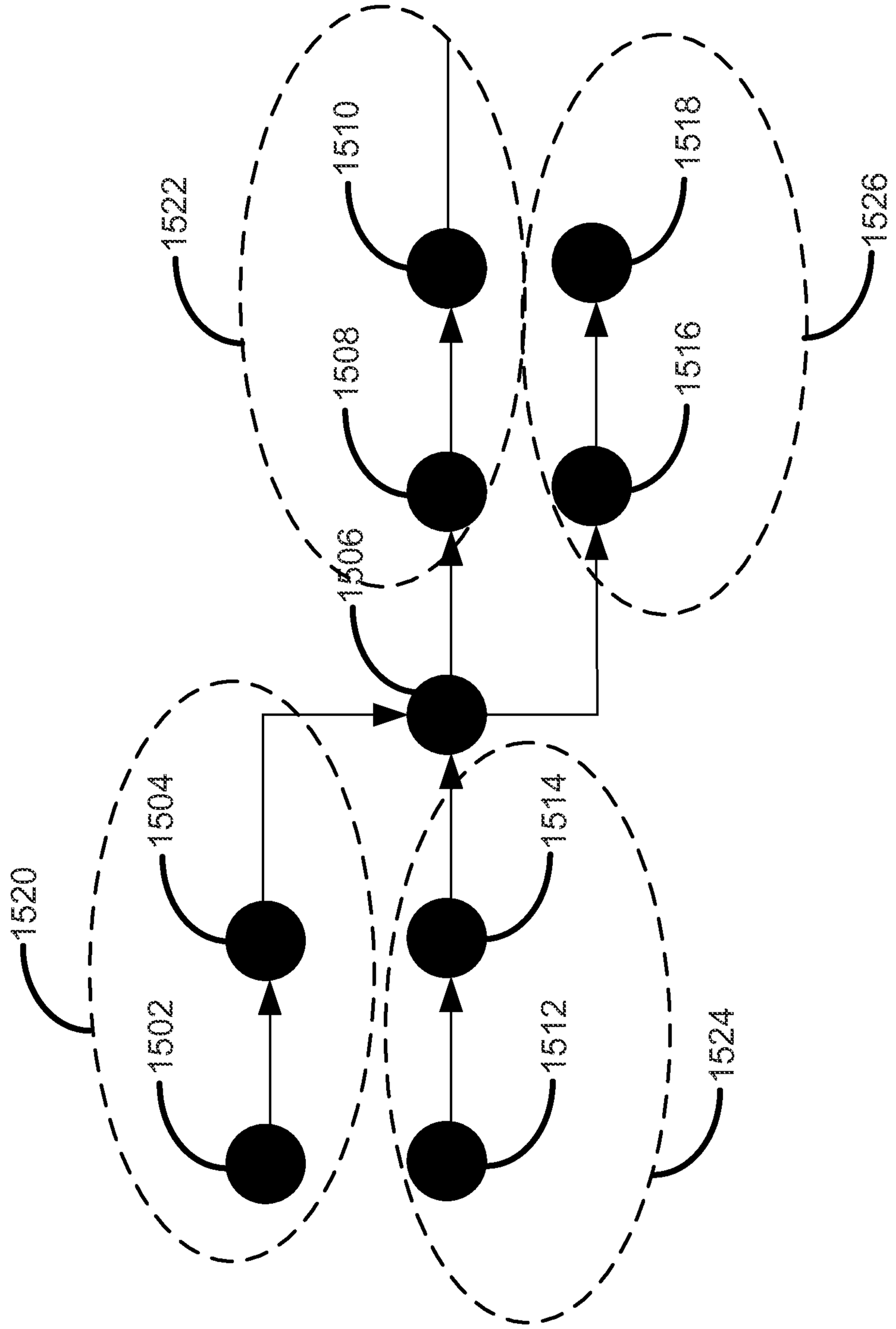

1522
1510
1518
1526
1508
1516
1506
1504
1514
1520
1502
1512
1524

Figure 15

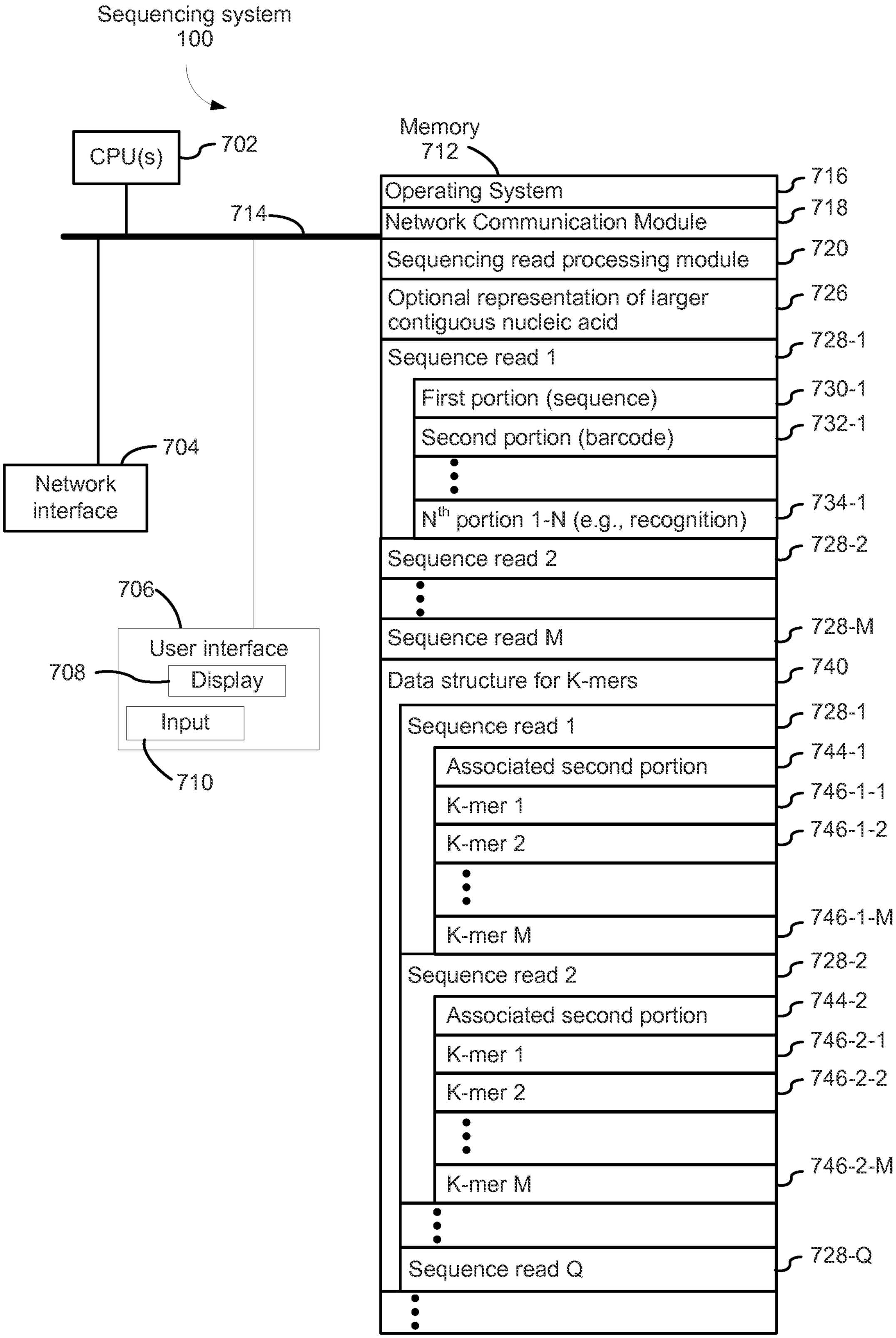

Sequencing system
100
CPU(s)
702
Memory
712
714
Operating System
716
Network Communication Module
718
Sequencing read processing module
720
Optional representation of larger contiguous nucleic acid
726
Sequence read 1
728-1
First portion (sequence)
730-1
Second portion (barcode)
732-1
N^th portion 1-N (e.g., recognition)
734-1
Sequence read 2
728-2
Sequence read M
728-M
Data structure for K-mers
740
Sequence read 1
728-1
Associated second portion
744-1
K-mer 1
746-1-1
K-mer 2
746-1-2
K-mer M
746-1-M
Sequence read 2
728-2
Associated second portion
744-2
K-mer 1
746-2-1
K-mer 2
746-2-2
K-mer M
746-2-M
Sequence read Q
728-Q
Network interface
704
User interface
706
Display
708
Input
710

Figure 7