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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
19 September 2013 (19.09.2013)

WIPO | PCT

(10) International Publication Number
WO 2013/138604 A1

(51) International Patent Classification: *C12Q 1/68* (2006.01)

(21) International Application Number: PCT/US2013/031429

(22) International Filing Date: 14 March 2013 (14.03.2013)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/611,993 16 March 2012 (16.03.2012) US

(71) Applicant: THE BROAD INSTITUTE, INC. [US/US]; 7 Cambridge Center #7034, Cambridge, MA 02142 (US).

(72) Inventors: CARNEIRO, Mauricio; 30 Caldwell Street, #420, Charlestown, MA 02129 (US). DEPRISTO, Mark; 333 Harvard Street, Apt. 7, Cambridge, MA 02139 (US).

(74) Agents: ELRIFI, Ivor, R. et al.; Mintz Levin Cohn Ferris Glovsky And Popeo, P.C., One Financial Center, Boston, MA 02111 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, 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, 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, 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, ML, MR, NE, SN, TD, TG).

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

(54) **Title**: SYSTEMS AND METHODS FOR REDUCING REPRESENTATIONS OF GENOME SEQUENCING DATA

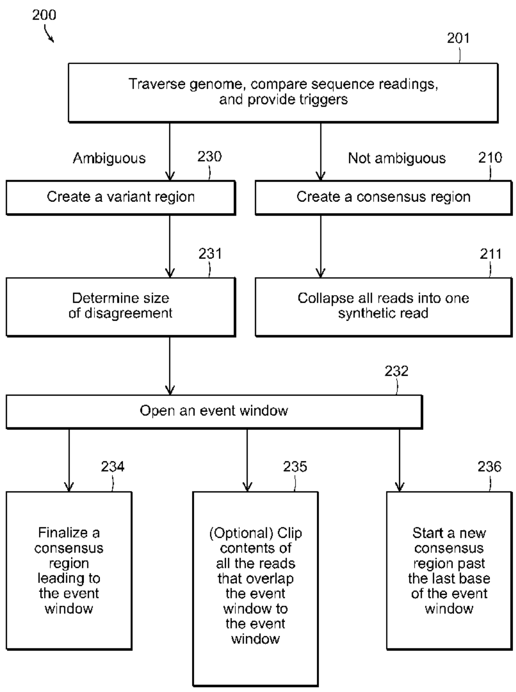

FIG. 2

(57) **Abstract**: Systems and methods are provided for reducing (e.g. compressing) representations of DNA sequencing data. These can include providing a plurality of DNA sequence reads, selecting a region of the reads and comparing the reads in the region to determine an ambiguity, and creating a consensus region and collapsing the selected region into a synthetic read if the ambiguity does not exceed a threshold.

WO 2013/138604 A1

SYSTEMS AND METHODS FOR REDUCING REPRESENTATIONS OF GENOME SEQUENCING DATA

CROSS-REFERENCE TO RELATED APPLICATION

[0001] This application claims priority to US Provisional Patent Application No. 61/611,993, filed on March 16, 2012, the contents of which are incorporated herein by reference in their entireties.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH

[0002] The present disclosure was made with government support under U54HG003067 awarded by the National Institutes of Health. The government has certain rights in the present disclosure.

FIELD OF THE DISCLOSURE

[0003] The present disclosure relates generally to the field of genome sequencing. More particularly, the disclosure relates to methods and systems for reducing the amount of data of sequence reads without substantially losing accuracy and/or statistical power.

BACKGROUND OF THE DISCLOSURE

[0004] As genome sequencing becomes cheaper and more efficient, the amount of sequence data available has grown exponentially. In fact, the rate of data growth has outpaced Moore's law's expected rate of growth for computing hardware. Having more data produced by sequencing can be very useful as this, for example, highly empowers medical studies to discover rare variations that could be associated with diseases.

[0005] Currently DNA is sequenced read by read. A read is a continuous sequence of DNA outputted by a sequencing instrument. Each read is then aligned to a reference genome using any alignment tool and the alignment information is then written out to a generic data analysis format (e.g. BAM file). This file contains all reads emitted by the instrument, and each read contains the position in the reference genome to which that read corresponds along with other additional information produced by the alignment and the instrument (e.g. base qualities, mate alignment information, DNA strand the read came from, etc.) that are also very important for subsequent analysis. The resulting file is typically approximately 10 gigabytes large for a human whole exome sequencing of one sample and approximately 300 gigabytes for a typical whole genome sequencing of one sample. These numbers are highly dependent on the sequencing depth and efficiency of the sequencing center.

[0006] Because of sequencing errors and the imprecision associated with the process of DNA sequencing, data analysis requires multiple observations of the same locus in the genome to ascertain its reading. A typical whole genome sequencing project will aim to cover every loci in the genome with at least 30 distinct reads. This overlap generates an excess of information that is necessary for quality control and statistical power to distinguish between accurate readings and instrument artifacts but overwhelms the downstream analysis tools.

[0007] Since sequencing data files are typically very large, they cannot be easily transferred from one storage means to another. In fact, even simple analyses of these files can take a very long time. Furthermore, loading multiple sequencing data files in computer memory is sometimes impossible (e.g. due to computing capacity limitations). One work-around for such large-scale analysis is batching (i.e. a process in which the collection of files is analyzed in smaller groups of files separately and later have the results combined through various statistical methods). Batching is undesirable as it creates inaccurate genotyping likelihoods (causing loss of rare mutations) and significantly reduces the statistical power to distinguish true variation from sequencing artifacts in the different batches.

[0008] Current analysis systems and methods, which are limited by the computing capacity to process the volume of data, are unable to keep pace with the data growth. For example, one major issue facing organizations that are required to store publication data, such as NCBI and EBI, is the enormous storage required to store all the data and the difficulty associated with transferring these large files to analysts. As a result, there has been investigation into better compression algorithms, and even loss-of-data as a solution. Another issue relates to the analysis of the vast amount of DNA data, which can take a very long time even for a simple analysis.

SUMMARY OF THE DISCLOSURE

[0009] In view of the foregoing, there is a need to provide a tool, which addresses the limitations of current systems and methods for DNA data analysis.

[0010] Embodiments of the present disclosure provide a solution, including computer systems and methods for reducing (e.g. compressing) representations of DNA sequencing data.

[0011] In accordance with the present disclosure, a computer-implemented method is provided for compressing DNA sequence reads. The method may include providing a plurality of DNA sequence reads, selecting a region of the reads and comparing the reads in the region to determine an ambiguity, and creating a consensus region and collapsing the selected region into a synthetic read if the ambiguity does not exceed a threshold. In some embodiments, the synthetic read amalgamates bases of the selected region of the reads.

[0012] In some embodiments, the method may include one or more of: tagging the synthetic read with a base-counts annotation indicating a number of the reads that agree with the synthetic read; tagging the synthetic read with a base-qualities annotation indicating an estimated accuracy of an instrument in detecting bases of the selected region; tagging the synthetic read with a mapping-quality annotation representing an agreement of the reads; creating a variant region if the ambiguity exceeds the threshold.

[0013] In some embodiments, the ambiguity is based on at least one trigger. The at least one trigger may be selected from the group consisting of SNP Trigger, Deletion Trigger, Insertion Trigger, Soft-clip Trigger, Coverage Trigger, and Mate Trigger.

[0014] In some embodiments, if the ambiguity exceeds the threshold, the method further includes determining a size of disagreement among the reads, opening an event window having the size of disagreement about the selected region, and clipping contents of the reads that overlap the event window to the event window. In some embodiments, if the ambiguity exceeds the threshold, the method further includes finalizing a consensus region leading to the event window. In some embodiments, if the ambiguity exceeds the threshold, the method further includes starting a new consensus region past a last base of the event window.

[0015] In some embodiments, the size of disagreement corresponds to a number of contiguous loci where the reads have different bases. In some embodiments, the threshold is a maximum threshold of variation among the reads or for a number of deletions and a window size. In some embodiments, the method can further include discarding some of the reads to meet a maximum requirement.

[0016] According to some embodiments of the present disclosure, a system, method, and non-transitory computer-readable medium are provided for reducing (or compressing) DNA sequences (readings). Computer memory (e.g. one or more databases) is provided that stores DNA sequences. A computer system (e.g. including one or more processors) in communication with the computer memory is also provided. The computer system is configured to provide a graphical user interface for displaying, for example, user options and data to a user.

[0017] Computer program products are also described that comprise non-transitory computer readable media storing instructions, which when executed one or more data processor of one or more computing systems, causes at least one data processor to perform operations herein. Similarly, computer systems are also described that may include one or more data processors and a memory coupled to the one or more data processors. The memory may temporarily or permanently store instructions that cause at least one processor to perform one or more of the operations described herein. In addition, methods can be implemented by one or more data processors either within a single computing system or distributed among two or more computing systems. Such computing systems can be connected and can exchange data and/or commands or other instructions or the like via one or more connections, including but not limited to a connection over a network (e.g. the Internet, a wireless wide area network, a local area network, a wide area network, a wired network, or the like), via a direct connection (wired or peer-to-peer wireless) between one or more of the computing systems, etc.

[0018] The details of one or more variations of the subject matter described herein are set forth in the accompanying drawings and the description below. Other features and advantages of the subject matter described herein will be apparent from the description and drawings, and from the claims.

BRIEF DESCRIPTION OF THE DRAWINGS

[0019] For a better understanding of the present disclosure, reference is made to the following description, taken in conjunction with the accompanying drawings, in which like reference characters refer to like parts throughout, and in which:

[0020] FIG. 1 is a block diagram of a system according to some embodiments of the present disclosure;

[0021] FIG. 2 is a flow chart showing various aspects of some embodiments of the present disclosure;

[0022] FIGS. 3-7 contain graphical illustrations of data in accordance with exemplary embodiments of the present disclosure; and

[0023] FIG. 8 is graph of the actual and expected file size and average coverage in accordance with exemplary embodiments of the present disclosure.

DETAILED DESCRIPTION OF THE DISCLOSURE

[0024] Reference will now be made to FIG. 1 showing a new system according to some embodiments of the present disclosure. As shown, System 10 includes a DNA Analysis Application 101, a Database 102 for storing data, and a Compression Module 103 for applying a unique sequence reduction (compression) method as will be discussed below. In some embodiments of the present disclosure, the System 10 provides a user input/output interface 104 for inputting data and/or commands, and displaying information. In some embodiments of the present disclosure, DNA Analysis Application 101, Database 102, and/or Compression Module 103 may reside on one or more computers and executed by one or more processors. In some embodiments of the present disclosure, interface 104 includes suitable equipment for providing outputs from system 10 to the user. Such interface(s) may include one or more display devices (e.g., liquid crystal display (LCD) device of a personal or home computer, or a mobile phone display), and/or any other suitable output device(s).

[0025] Generally, Compression Module 103 compresses (reduces) sequence reads by selectively compressing all the information that unambiguously agrees across all reads that cover the same location in the genome and writing it in a format that still gives programs sufficient statistical power to perform data analysis. As will be explained in more detail below, Compression Module 103 creates consensus regions and variant regions in the genome and treats them separately. Consensus regions are contiguous stretches of the genome where all the reads unambiguously (i.e. meeting one or more predefined parameters) agree on what bases have been read, whereas variant regions are windows in the genome where the read data does not fully agree and further analysis is necessary to determine whether there is a real event occurring or if the disagreement stems from sequencing or alignment inaccuracy. Ambiguity is defined by trigger action (*see* Triggers below for more information). An ambiguous region is any contiguous stretch of DNA that triggers at least one active trigger and an unambiguous region is any contiguous stretch of DNA that does not trigger any of the active triggers.

[0026] References will now be made to FIG. 2, showing various steps implemented by the System 10, according to some embodiments of the present disclosure. It should be noted while some embodiments of the present disclosure execute the steps in the order presented, other embodiments may execute some of the steps in different orders (e.g. some steps may be taken before or simultaneously as another).

[0027] As shown, DNA Analysis Application 101 is configured to allow System 10 to traverse genome, compare sequence readings and provide one or more triggers at step 201 (details of the triggers will be discussed below). The sequence reads may be already stored in Database 102, or may be entered/uploaded by the user. If a selected region meets (or does not meet) certain criteria and is deemed to be not ambiguous, Compression Module 103 creates a consensus region at the selected region (or designates the selected region as a consensus region) at step 210, and collapses all the reads into one synthetic (i.e. compressed) read at step 211.

[0028] If a selected region meets (or does not meet) certain criteria, and is deemed to be ambiguous, Compression Module 103 creates a variant region at the selected region (or designates the selected region as a variant region) at 230. If a variant region is created or designated, Compression Module 103 determines the size of disagreement (or variation between the sequences) at 231 and opens an event window at 232. Compression Module 103 then finalizes the region (e.g. a consensus region) leading to the event window at 234; optionally clip the contents of all the reads that overlap the event window to the event window at 235; and starts a new selected region (e.g. a new consensus region) past the last base of the event window at 236. Compression Module 103 may then select a new region and repeat the process iteratively to the end of the sequences.

[0029] Some of the key features discussed above will now be discussed in more detail below.

Consensus Region

[0030] According to some embodiments of the present disclosure, in a consensus region, all reads are collapsed into one synthetic read that amalgamates the information contained in the agreement. The read is synthetic because it is a representation of the unambiguous agreement, and not a direct observation reported by the sequencing instrument. The synthetic read however is no different than any read. It has all the bases representative of the agreement and all the typical annotations of a read plus an extra field that carries the number of observations for each base in the synthetic read. These annotations are summarized below:

Base Counts

[0031] For example if 30 reads agree that an 'A' was observed on a given location, then the synthetic read will contain an 'A' and a corresponding annotation in the extra field of 30 denoting the number of reads that agreed for that base. This array is represented under the additional tag RR. This field is represented with a delta representation for extra compression. That is, for example, if a synthetic read has base counts of 30, 30, 31, 32, 30, 29 and 27, the array will carry the first count unchanged, and the difference from the first base to every other base in the read. Therefore the counts above would be represented by the following array: 30, 0, 1, 2, 0, -1 and -3.

Base Qualities

[0032] Every read may have an array of quality scores associated with each base in the read. The quality score may, for example, represent the estimated accuracy of the instrument in detecting that base. In some embodiments of the present disclosure, in the synthetic read, each base is a representation of an agreement of multiple reads and the base quality of each base is the average of all the base qualities of that base in all the reads that contributed to that base in the synthetic read.

Mapping Quality

[0033] Every read may have a measure of how well it aligns with the reference genome in the chosen position. In some embodiments of the present disclosure, the mapping quality of a synthetic read is the average of the root mean square of the mapping qualities of every read that contributed to each base in the synthetic read.

Variant Region

[0034] In some embodiments of the present disclosure, when a variant region is triggered (see triggers for more information), Compression Module 103 first determines the size of the disagreement (i.e. the number of contiguous loci where the overlapping reads report different bases). Secondly it opens a window (or event window) around the event big enough to capture all the information leading to and after the event. Finally it finalizes the consensus read leading to the event up until one base before the first base of the window, and starts a new consensus region one base past the last base of the window. All the reads that overlap the event window can optionally have their contents clipped to the event window (i.e. all bases outside the window can be 'thrown away' because they have already been represented in the consensus region), alternatively they can be unclipped and their representation will be discounted from the overlapping consensus read accordingly. The result is a narrow window of overlapping reads that includes the event and all necessary extended information for the event to be interpreted later by other analysis tools. In the case where multiple events close together create overlapping windows, they can be combined into one single event window that covers from the start of the event window of the first event to the end of the event window of the last event.

[0035] When processing multiple samples together, these samples can be processed jointly or independently. In an independent process, each sample can be reduced following the trigger rules. In a joint process, all samples are processed at the same time, and if any trigger in any sample triggers a variant region, then all samples can maintain the representation of the variant window. This has applications when multiple samples need to be analyzed jointly by downstream tools (e.g. in cancer studies, tumor and normal samples must be reduced jointly).

[0036] In some embodiments, when the disagreement is not clearly a sequencing artifact disambiguated by the overlapping reads, a representation of the full data plus necessary information can be preserved for downstream analysis. In some embodiments of the present disclosure, it is not the goal of Compression Module 103 to determine what event is responsible for the disagreement, instead, the focus is in assuring that analysis tools will have all the necessary information to understand the event in the data.

Triggers

[0037] A trigger is a programatic way to determine whether the data has any ambiguity at a given region. In some embodiments of the present disclosure, Compression Module 103 can operate with any number of triggers. As the genome is traversed, every locus is analyzed by all triggers provided. Each trigger will determine, according to its specifications (e.g. one or more parameters) whether or not the region is ambiguous. If any trigger decides that the region is ambiguous, a variant region is created. The trigger determines how big the event window should be in order to preserve enough data for further statistical analysis of this event by other analytical tools. Different triggers may require different sizes of event window.

[0038] In some embodiments of the present disclosure, as Compression Module 103 traverses the genome, following the order given by the provided reference genome index file, it evaluates each locus with all the triggers to determine whether or not this locus contains an event (i.e. when a trigger is activated) and, given the size of the largest event window of all active triggers, what is the last locus before the event window that can be turned into a consensus read.

SNP Trigger

[0039] In some embodiments of the present disclosure, Compression Module 103 is provided with a single nucleotide polymorphism (SNP) trigger. In every locus, the SNP triggers looks for a maximum threshold of variation t_s (a user-defined parameter). If the variation among all reads covering that locus exceeds the given threshold, the SNP trigger immediately triggers a variant region with an event window of size s_s (user-defined parameter). In each iteration, the SNP trigger marks the locus s_s+1 positions before the current locus as ready for consensus (and consequently all loci before it).

Deletion Trigger

[0040] In some embodiments of the present disclosure, Compression Module 103 is provided with a deletions trigger. The deletions trigger works in a similar fashion as the SNP trigger using a maximum threshold for the number of deletions t_d and a window size s_d (both user-defined parameters), except that the events (the deletions) can happen in multiple contiguous loci. Therefore the deletions trigger will accumulate information until the last locus of the event is processed to trigger and size the event window. In every iteration where no deletion is found, the behavior is similar to the SNP trigger, marking the locus s_d+1 positions before the current locus as ready for consensus (and consequently all loci before it), but once a locus has enough deletions that it exceeds the threshold t_d then the deletions trigger will continuously mark the same locus that was s_d+1 positions away from the first locus that exceed the threshold t as ready for consensus while calculating the number of deletions for the subsequent loci until it finds one that does not exceed t and then trigger a variant region with a window that spans s loci before the first locus that exceeded the threshold t_d, and s_d loci after the last locus that exceeded the threshold t.

Insertion Trigger

[0041] In some embodiments of the present disclosure, Compression Module 103 is provided with a insertions trigger. For each locus, the insertions trigger looks for a maximum threshold of insertions to the right of the locus t_i (a user-defined parameter). If the number of insertions to the right of the evaluated locus among all reads exceeds the given threshold t_i, the insertions trigger immediately triggers a variant region with an event window of size s_i (e.g. a user-defined parameter). In each iteration, the insertions trigger marks the locus s_i+1 positions before the current locus as ready for consensus (and consequently all loci before it).

Soft-clip Trigger

[0042] In some embodiments of the present disclosure, Compression Module 103 is provided with a soft-clip trigger. For each locus, the soft-clip trigger looks for a maximum threshold of soft-clipped bases t_{sc} (a user-defined parameter). If the number of soft-clipped bases among all reads covering the evaluated locus exceeds the given threshold t_{sc}, the soft-clip trigger immediately triggers a variant region with an event window of size s_{sc} (user-defined parameter). In every iteration, the soft-clip trigger marks the locus s_{sc}+1 positions before the current locus as ready for consensus (and consequently all loci before it).

Coverage Trigger

[0043] In some embodiments of the present disclosure, Compression Module 103 is provided with a coverage trigger. For each locus, the coverage trigger looks for abnormal changes in coverage among reads with a permissive threshold t_c (a user-defined parameter). If the difference in coverage among all reads covering the evaluated locus exceeds the given threshold t_c, the coverage trigger immediately triggers a variant region with an event window of size s_c (user-defined parameter). In every iteration, the coverage trigger marks the locus s_c+1 positions before the current locus as ready for consensus (and consequently all loci before it). The coverage trigger will extend the event size by comparing each subsequent locus after the locus that triggered with the running average coverage before the locus that triggered. While the difference exceeds the threshold t_c, the event window is increased to represent the entire region with the relative drop or excess in coverage.

Mate Trigger

[0044] In some embodiments of the present disclosure, Compression Module 103 is provided with a mate trigger. For each locus, the mate trigger looks for a maximum threshold of reads with incorrectly mapped mates t_m (a user provided parameter). If the number of reads with incorrectly mapped mates exceeds the given threshold t_m, the mate trigger immediately triggers a variant region with an event window of size s_m (user-defined parameter). The event is represented by the first base aligned by the leftmost read in the locus that has an incorrectly mapped mate. If the incorrectly mapped mate is aligned to the left of the current mate, then the first aligned base of the leftmost mate will be used. In each iteration, the insertions trigger marks the locus s_m+1 positions before the current locus as ready for consensus (and consequently all loci before it). An incorrect alignment can be defined as any of the following inconsistencies: mate inexistent, mate mapped to a different chromosome or mate mapped significantly further than the expected insert size (threshold for significance v_m defined by the user)

Default and Additional Triggers

[0045] In some embodiments of the present disclosure, Compression Module 103 is provided with several default triggers. Additional triggers can be developed by the user and provided as a plugin to the Compression Module 103.

Read Name Compression

[0046] In some embodiments of the present disclosure, all reads outputted by the Compression Module have their names compressed for better use of the data. For example, all reads may be numbered from 1 to N and synthetic reads may have the letter “C” in front of the number. Reads that had the same name before applying Compression Module 103 will have the same compressed name after applying Compression Module 103, guaranteeing the relationship between paired reads.

Original Alignment

[0047] In some embodiments of the present disclosure, reads that form the variant region will carry the original start and/or end locations (prior to clipping)- if the process of clipping makes it impossible to retrieve either location. Due to the clipping contract of Compression Module 103, the hard clip operator will be used to denote each reference location clipped (e.g. aligned bases or deletions, but not insertions and soft-clipped bases), therefore, adding all the leading hard clip operators will provide the original start of the read, similarly adding all the tail hard clip operators will provide the original end of the read provided that Compression Module 103 does not hard-clip soft-clipped bases. In these cases (and only in these cases) the read will carry an extra tag OM and OE marking, respectively, the shift to their original alignment start and the shift to their original alignment end. In some embodiments of the present disclosure, the true original alignment can be retrieved by calculating the original alignment as described above, and then adding the OM to the original alignment start and subtracting the OE from the original alignment end. If any of the tags do not exist, it is the same as adding or subtracting zero to the original alignment.

Heterozygous compression

[0048] In some embodiments, a special trigger can be used for sites where the variation is clear and falls within the acceptable threshold for the number of chromosomes in the organism. For example, in humans, a hetorozygous site is one where approximately 50% of the data shows evidence for one allele and the other 50% shows evidence for the other allele. In these cases, four consensus reads are created in the event window. Two consensus reads for each chromosome, one representing the data originating from the positive strand of the DNA, and the other for the data originating from the negative strand of the DNA. The trigger has one parameter "e", to define the threshold to trigger the heterozygous compression, if and only if, the data shows 50% +/- e support for each allele (in the diploid case). This trigger is generic and can also be applied for triploid organisms (e.g. seedless watermelons) and tetraploid organisms (e.g. salmon fish) with expected frequencies of 33% and 25% respectively, plus or minus the parameter e.

Known sites trigger

[0049] In some embodiments, given a list of sites the algorithm can preemptively create a variant window around those sites regardless of the data. The variant region will then follow the rules of heterozygous compression or standard compression based on the data, but the variant region itself will be enforced even if there is no variation. The only parameter to this mode is the list of sites. An example use of this mode is when given a list of previously known mutated sites in the population that need to be further analyzed across samples regardless of variation status.

Additional Options

[0050] In some embodiments of the present disclosure, Compression Module 103 includes several additional ways to compress the data even more using the following options:

Downsampling

[0051] Frequently, analysis tools do not need more than a certain amount of data to determine the event happening in the loci covered by the event window. Some regions of the genome in sequencing projects can be covered by as many as 1000 reads. By defining what is the maximum depth necessary to evaluate a variant region, in some embodiments of the present disclosure, Compression Module 103 discards reads to try and satisfy the maximum requirements provided. Several downsampling algorithms may be implemented by Compression Module 103. For example, one default down-sampler may maintain the shape of the coverage distribution by keeping the proportion of coverage while guaranteeing the locus with the minimum coverage will be covered to the maximum required coverage and the other loci will maintain the proportions by keeping a proportional number of extra reads on top of the maximum. In other embodiments, Compression Module 103 includes a simpler down-sampler that tries to flatten the distribution by bringing every loci as close as possible to the maximum distribution. In other embodiments, Compression Module 103 includes a position biased down-sampler that prefers to maintain a distribution of reads that have different alignments, avoiding piles of reads that represent the same genomic region. The downsampling incorporated by Compression Module 103 may include any procedure that randomly discards reads to achieve the maximum depth necessary.

Minimum Base Quality Threshold

[0052] Base qualities are indications from the instruments of how likely it is that the base is an error. Low base qualities are very likely to be errors and can optionally be hidden from the triggers. In some embodiments of the present disclosure, if a base does not meet the minimum base quality threshold, the triggers will not see it and will make their decisions on whether or not to trigger a variant region without regards to what base the low quality base reports.

Minimum Mapping Quality Threshold

[0053] Similarly, in some embodiments of the present disclosure, reads with mapping quality below a given threshold may be kept hidden from the triggers, not contributing to the decision of creating variant regions. Mapping quality is a measure of how likely it is that the alignment of the read is correct. Unlikely alignments are usually full of artifactual mismatches that create unnecessary variant regions.

Minimum Tail Quality

[0054] A common error mode of some DNA sequencing instruments is to lose sensitivity either at the start or the end of the read (tails). When this happens, the tails have significantly lower base quality scores than the middle of the read. In some embodiments of the present disclosure, Compression Module 103 can hard-clip (remove) the tails of these reads when a low quality tail is detected. The user can specify either a threshold base quality to determine a low quality tail, or the maximum drop in quality acceptable within the tails of the reads and Compression Module 103 may programmatically determine whether or not the quality is significantly lower on the tails and clip them accordingly.

Discard non-standard read tags

[0055] In some embodiments of the present disclosure, upstream tools ran on the data may write out special tags for each read that carry specific information corresponding to that tool (for example, Compression Module 103 may add RR, OM and OE tags). Compression Module 103 may give the user the option to erase all previous tags (that are not part of the alignment) added by these tools. In some embodiments, the user is provided with the ability to remove all tags or select tags to be kept.

Discard Soft-Clipped Bases

[0056] In some embodiments of the present disclosure, bases are soft-clipped when the aligner fails to align and chose to soft-clip around the tail-ends of the read. The user may have the option to get rid of these tails or keep them for further analysis. Soft-clipped bases may only be seen by triggers that request to see them.

Discard Adaptor Sequences

[0057] Sometimes the sequences have chunks of the adaptor sequence appended to the reads. More often than not these bases are soft clipped by the aligner, but sometimes their sequence matches the alignment and they will be represented as matches. In some embodiments of the present disclosure, Compression Module 103 can estimate the fragment size by comparing the alignments of the read and its mate (the read that starts on the opposite end of the fragment) and discard all bases that fall outside the inferred fragment.

Discard off-target sequences

[0058] For targeted sequencing projects, such as whole exome or re-sequencing validation projects, there will always be several reads that fall outside of the desired targeted region. In some embodiments of the present disclosure, Compression Module 103 offers the option of discarding all reads that are fully off-target. It also discards all bases from on-target reads that extend off-target, keeping only the on-target bases.

[0059] References will now be made to FIGS. 3-7, which contain graphical illustrations of what the reduced reads data could look like in accordance with exemplary embodiments of the present disclosure. FIG. 3 is a graphical illustration of the original data (BAM) versus the reduced reads data (BAM) in accordance with exemplary embodiments of the present disclosure. As can be seen, the reduced data include consensus reads (regions), variable regions, and homozygous variants. FIG. 4 includes two graphical illustrations. On the left side is a graphical illustration of the original data (BAM) versus the reduced reads data (BAM) showing multiple variants merging variant region(s). On the right side is a graphical illustration of the original data (BAM) versus the reduced reads data (BAM) showing the long deletion.

[0060] As shown in FIG. 5, in some embodiments of the present disclosure, the reduced reads data (BAM) can allow tools to perform the same or similar annotations as the original data (BAMs). To this end, in some embodiments, one or more modifications to the annotation engine are made.

[0061] FIGS. 6 and 7 show additional examples of what the reduced reads data in accordance with exemplary embodiments of the present disclosure could look like. As shown in FIG. 6, the diversity among the reads themselves can result in variable regions where select reads are retained, and that a conserved region can produce reduced reads. As shown in FIG. 7, the region of diversity among the reads (there's a het SNP, and a partial realigned het indel) can result in a variable region. Reads spanning the region (e.g. sites of diversity +/- 40 bp) can be clipped to the region, and emitted. In some embodiments, reads can be downsampled to 50x average coverage across the region.

[0062] FIG. 8 is a graph of the actual file size and the average coverage (in accordance with some embodiments of the present disclosure) versus the expected file size and average coverage. As can be seen, in accordance with some embodiments of the present disclosure, there is only a marginal increase in file size with coverage rather than a linear relationship between the file size and the average coverage.

Example Results

Type 2 Diabetes Project

[0063] The T2D project (confidential) has sampled 2269 individuals and sequenced their whole exomes. We have compared standard analysis results using the standard data (original BAM files) and reduce representation of the BAM files. The analysis we chose for this study was identifying single nucleotide polymorphisms.

[0064] The original BAM file was called using batches of 100 samples which is the standard procedure because of the limitation in the amount of data that can be loaded into memory for the analysis. These batches are then merged following the institute's best practices. The original dataset will henceforth be called batches and the reduced representation dataset reduced. The reduced dataset was called in one pass, without merging. We restricted our analysis to Chromosome 1. We used the Genome Analysis Toolkit (GATK) for all analysis. Calls were made using the GATK's Unified Genotyper and variant filtration was performed using the GATK's Variant Quality Score Recalibration tool.

[0065] We reduced the BAM files using the SNP, insertion and deletion triggers with an event window of size 30 and triggering thresholds of 95%, 99% and 99% respectively. We also used minimum mapping quality of 20 (PHRED-scaled), minimum base quality of 20

(PHRED-scaled), downsampling of 100x, minimum tail quality of 2 (PHRED-scaled) and discarded all original read tags, off-target reads and bases and soft-clipped bases.

[0066] The reduced dataset had on average 130Mb per sample, while the batches had approximately 13Gb per sample. The reduced dataset ran through the analysis pipeline in under 19 hours while the batches ran for a little over one week.

[0067] The reduced dataset provided 84,742 SNPs and the batches dataset provided 84,140. These numbers are well within the acceptable variation of different calls. The big difference is in rare variants and multi-allelic SNPs (where different people have a mutation on the same locus, but the 'alternate' allele is not the same -- for example on a given locus, the reference genome has an 'A', but one individual has a 'T' and another individual has a 'G') because those are extremely rare and the likelihoods associated with such calls get severed by the batching scheme.

[0068] The reduce dataset recovered all 770 multi-allelic SNPs while the batches only identified approximately 498 correct multi-allelic SNPs and incorrectly identified 612 loci as multi-allelics when they were actually bi-allelic.

[0069] Another big difference is in the filtering power. When using all samples together, the filtering machinery is much empowered and does a much better job at choosing the correct variants to filter. In the batches dataset, we had 4,265 variants filtered incorrectly while the reduced dataset had none.

[0070] Overall the reduced dataset allowed for a much smaller footprint in file sizes (reducing it by approximately 100 fold with the options chosen). The runtime of the analysis tools was significantly improved (in the order of 20-40 fold). The combination of the smaller footprint and faster processing times allows the reduced representation datasets to be analyzed at the same time providing, in the end, more useful data to the analysis tools and keeping them from processing regions of the genome that are in unambiguous concordance. With that power, the result is faster analyses with a smaller footprint and higher quality results.

[0071] One or more aspects or features of the subject matter described herein may be realized in digital electronic circuitry, integrated circuitry, specially designed ASICs (application specific integrated circuits), computer hardware, firmware, software, and/or combinations thereof. These various implementations may include implementation in one or more computer programs that are executable and/or interpretable on a programmable system including at least one programmable processor, which may be special or general purpose, coupled to receive data and instructions from, and to transmit data and instructions to, a storage system, at least one input device (e.g., mouse, touch screen, etc.), and at least one output device.

[0072] These computer programs, which can also be referred to programs, software, software applications, applications, components, or code, include machine instructions for a programmable processor, and can be implemented in a high-level procedural and/or object-oriented programming language, and/or in assembly/machine language. As used herein, the term “machine-readable medium” refers to any computer program product, apparatus and/or device, such as for example magnetic discs, optical disks, memory, and Programmable Logic Devices (PLDs), used to provide machine instructions and/or data to a programmable processor, including a machine-readable medium that receives machine instructions as a machine-readable signal. The term “machine-readable signal” refers to any signal used to provide machine instructions and/or data to a programmable processor. The machine-readable medium can store such machine instructions non-transitorily, such as for example as would a non-transient solid state memory or a magnetic hard drive or any equivalent storage medium. The machine-readable medium can alternatively or additionally store such machine instructions in a transient manner, such as for example as would a processor cache or other random access memory associated with one or more physical processor cores.

[0073] These computer programs, which can also be referred to programs, software, software applications, applications, components, or code, include machine instructions for a programmable processor, and can be implemented in a high-level procedural language, an object-oriented programming language, a functional programming language, a logical programming language, and/or in assembly/machine language. As used herein, the term “machine-readable medium” refers to any computer program product, apparatus and/or device, such as for example magnetic discs, optical disks, memory, and Programmable Logic Devices (PLDs), used to provide machine instructions and/or data to a programmable processor, including a machine-readable medium that receives machine instructions as a machine-readable signal. The term “machine-readable signal” refers to any signal used to provide machine instructions and/or data to a programmable processor. The machine-readable medium can store such machine instructions non-transitorily, such as for example as would a non-transient solid state memory or a magnetic hard drive or any equivalent storage medium. The machine-readable medium can alternatively or additionally store such machine instructions in a transient manner, such as for example as would a processor cache or other random access memory associated with one or more physical processor cores.

[0074] With certain aspects, to provide for interaction with a user, the subject matter described herein can be implemented on a computer having a display device, such as for example a cathode ray tube (CRT) or a liquid crystal display (LCD) monitor for displaying information to the user and a keyboard and a pointing device, such as for example a mouse or a trackball, by which the user may provide input to the computer. Other kinds of devices can be used to provide for interaction with a user as well. For example, feedback provided to the user can be any form of sensory feedback, such as for example visual feedback, auditory feedback, or tactile feedback; and input from the user may be received in any form, including, but not limited to, acoustic, speech, or tactile input. Other possible input devices include, but are not limited to, touch screens or other touch-sensitive devices such as single or multi-point resistive or capacitive trackpads, voice recognition hardware and software, optical scanners, optical pointers, digital image capture devices and associated interpretation software, and the like.

[0075] The subject matter described herein may be implemented in a computing system that includes a back-end component (e.g., as a data server), or that includes a middleware component (e.g., an application server), or that includes a front-end component (e.g., a client computer having a graphical user interface or a Web browser through which a user may interact with an implementation of the subject matter described herein), or any combination of such back-end, middleware, or front-end components. The components of the system may be interconnected by any form or medium of digital data communication (e.g., a communication network). Examples of communication networks include a local area network ("LAN"), a wide area network ("WAN"), and the Internet.

[0076] The computing system may include clients and servers. A client and server are generally remote from each other and typically interact through a communication network. The relationship of client and server arises by virtue of computer programs running on the respective computers and having a client-server relationship to each other.

[0077] The subject matter described herein can be embodied in systems, apparatus, methods, and/or articles depending on the desired configuration. The implementations set forth in the foregoing description do not represent all implementations consistent with the subject matter described herein. Instead, they are merely some examples consistent with aspects related to the described subject matter. Although a few variations have been described in detail above, other modifications or additions are possible. In particular, further features and/or variations can be provided in addition to those set forth herein. For example, the implementations described above can be directed to various combinations and subcombinations of the disclosed features and/or combinations and subcombinations of several further features disclosed above. In addition, the logic flow(s) depicted in the accompanying figures and/or described herein do not necessarily require the particular order shown, or sequential order, to achieve desirable results. Other implementations may be within the scope of the following claims.

CLAIMS:

1. A computer-implemented method for compressing DNA sequence reads, the method comprising:

providing a plurality of DNA sequence reads;

selecting a region of the reads and comparing the reads in the region to determine an ambiguity; and

creating a consensus region and collapsing the selected region into a synthetic read if the ambiguity does not exceed a threshold.

2. The method of claim 1, wherein the synthetic read amalgamates bases of the selected region of the reads.

3. The method of claim 1, further comprising tagging the synthetic read with a base-counts annotation indicating a number of the reads that agree with the synthetic read.

4. The method of claim 1, further comprising tagging the synthetic read with a base-qualities annotation indicating an estimated accuracy of an instrument in detecting bases of the selected region.

5. The method of claim 1, further comprising tagging the synthetic read with a mapping-quality annotation representing an agreement of the reads.

6. The method of claim 1, further comprising creating a variant region if the ambiguity exceeds the threshold.

7. The method of claim 1, wherein the ambiguity is based on at least one trigger.

8. The method of claim 7, wherein the at least one trigger is selected from the group consisting of SNP Trigger, Deletion Trigger, Insertion Trigger, Soft-clip Trigger, Coverage Trigger, and Mate Trigger.

9. The method of claim 5, wherein if the ambiguity exceeds the threshold, the method further comprises:

determining a size of disagreement among the reads;

opening an event window having the size of disagreement about the selected region; and

clipping contents of the reads that overlap the event window to the event window.

10. The method of claim 9, wherein if the ambiguity exceeds the threshold, the method further comprises finalizing a consensus region leading to the event window.

11. The method of claim 9, wherein if the ambiguity exceeds the threshold, the method further comprises starting a new consensus region past a last base of the event window.

12. The method of claim 9, wherein the size of disagreement corresponds to a number of contiguous loci where the reads have different bases.

13. The method of claim 1, wherein the threshold is a maximum threshold of variation among the reads.

14. The method of claim 1, wherein the threshold includes a maximum threshold for a number of deletions and a window size.

15. The method of claim 1, further comprising discarding some of the reads to meet a maximum requirement.

16. A system for compressing DNA sequence reads, the system comprising:

means for providing a plurality of DNA sequence reads;

means for selecting a region of the reads and comparing the reads in the region to determine an ambiguity; and

means for creating a consensus region and collapsing the selected region into a synthetic read if the ambiguity does not exceed a threshold.

17. A non-transitory computer readable medium comprising computer-executable instructions recorded thereon for causing a computer to perform the method according to claim 1.

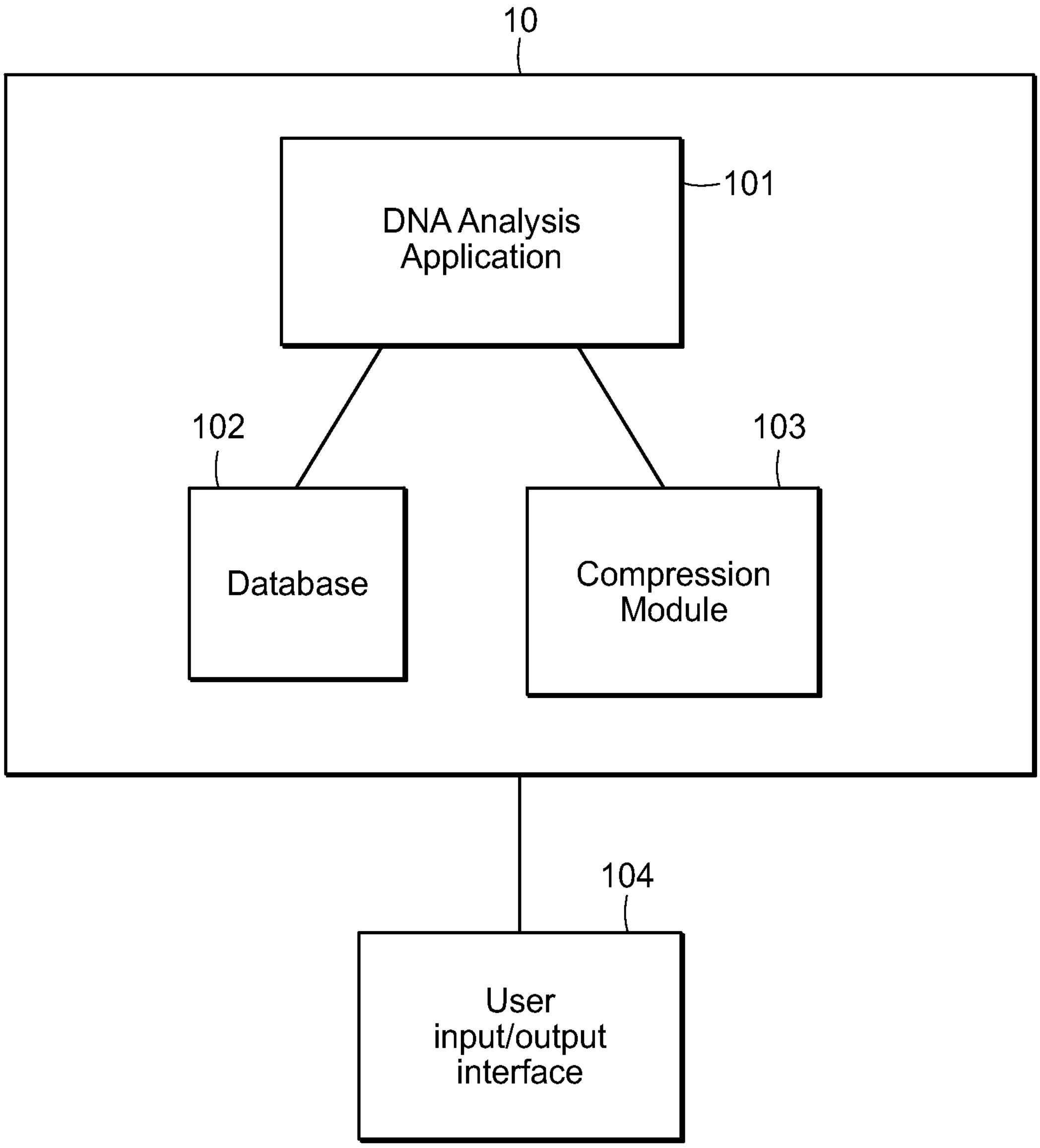
10
101
DNA Analysis Application
102
Database
103
Compression Module
104
User input/output interface

FIG. 1

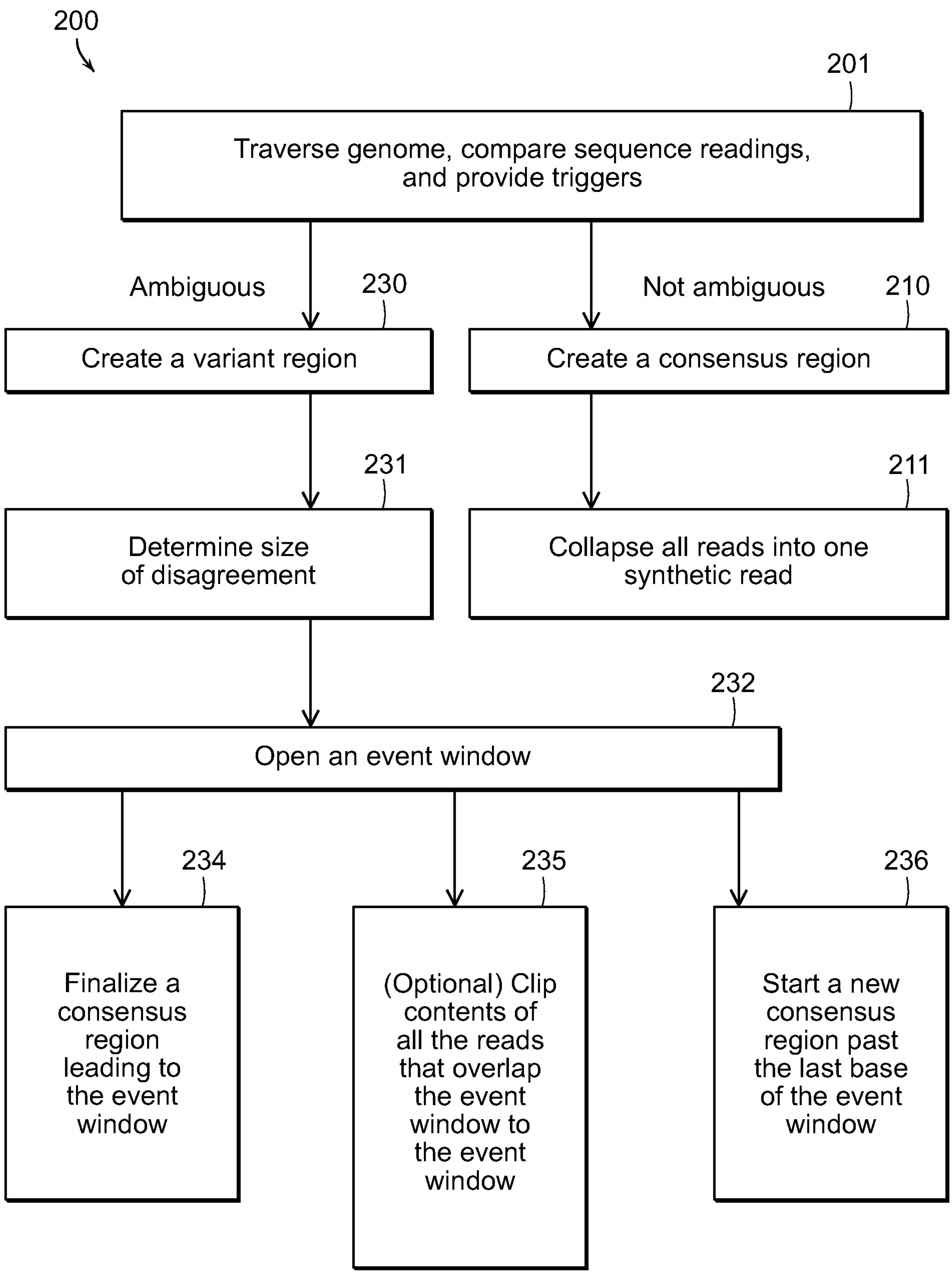
200
201
Traverse genome, compare sequence readings, and provide triggers
Ambiguous
230
Not ambiguous
210
Create a variant region
Create a consensus region
231
211
Determine size of disagreement
Collapse all reads into one synthetic read
232
Open an event window
234
235
236
Finalize a consensus region leading to the event window
(Optional) Clip contents of all the reads that overlap the event window to the event window
Start a new consensus region past the last base of the event window

FIG. 2

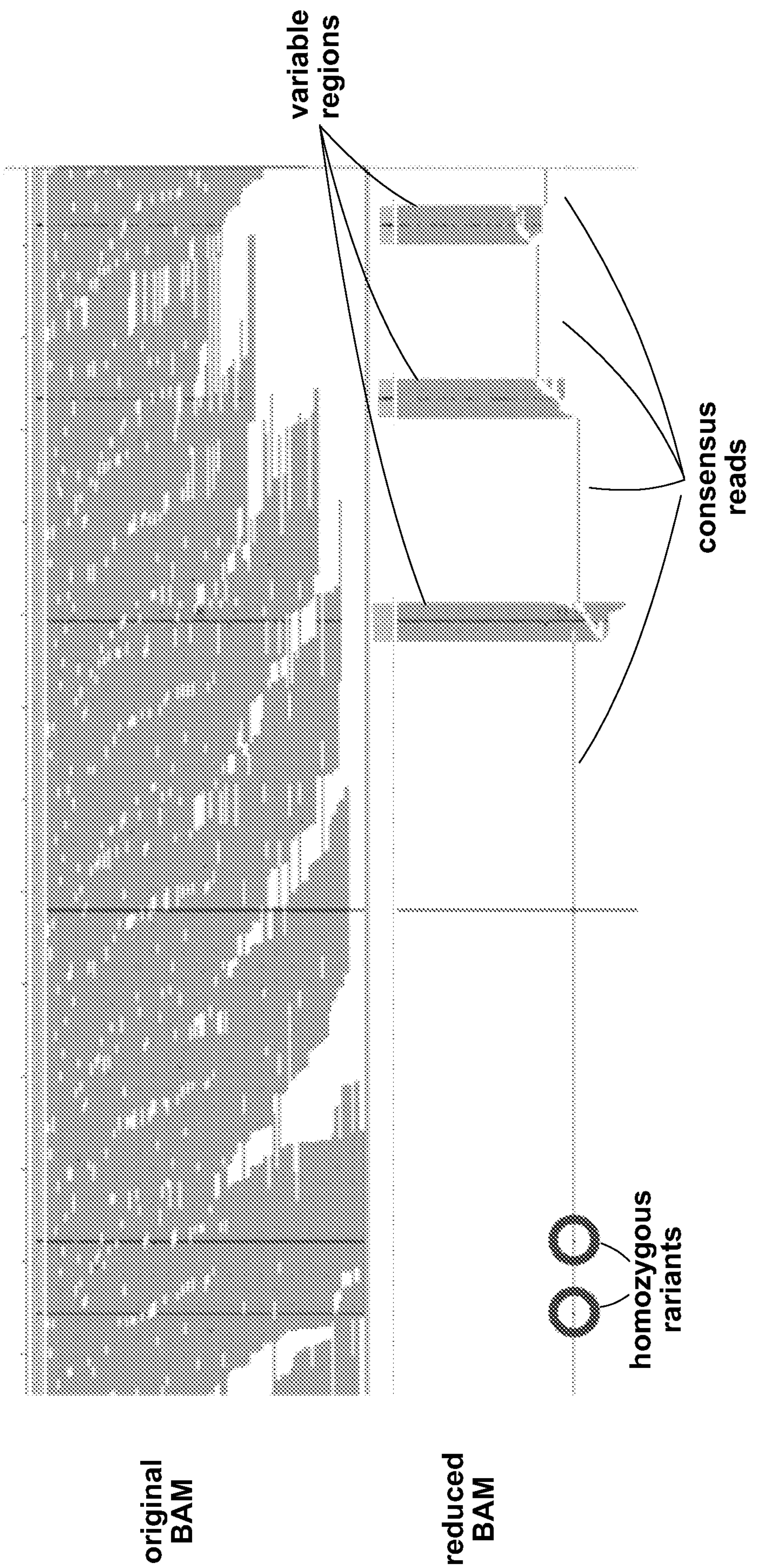
variable regions
consensus reads
homozygous rariants
original BAM
reduced BAM

FIG. 3

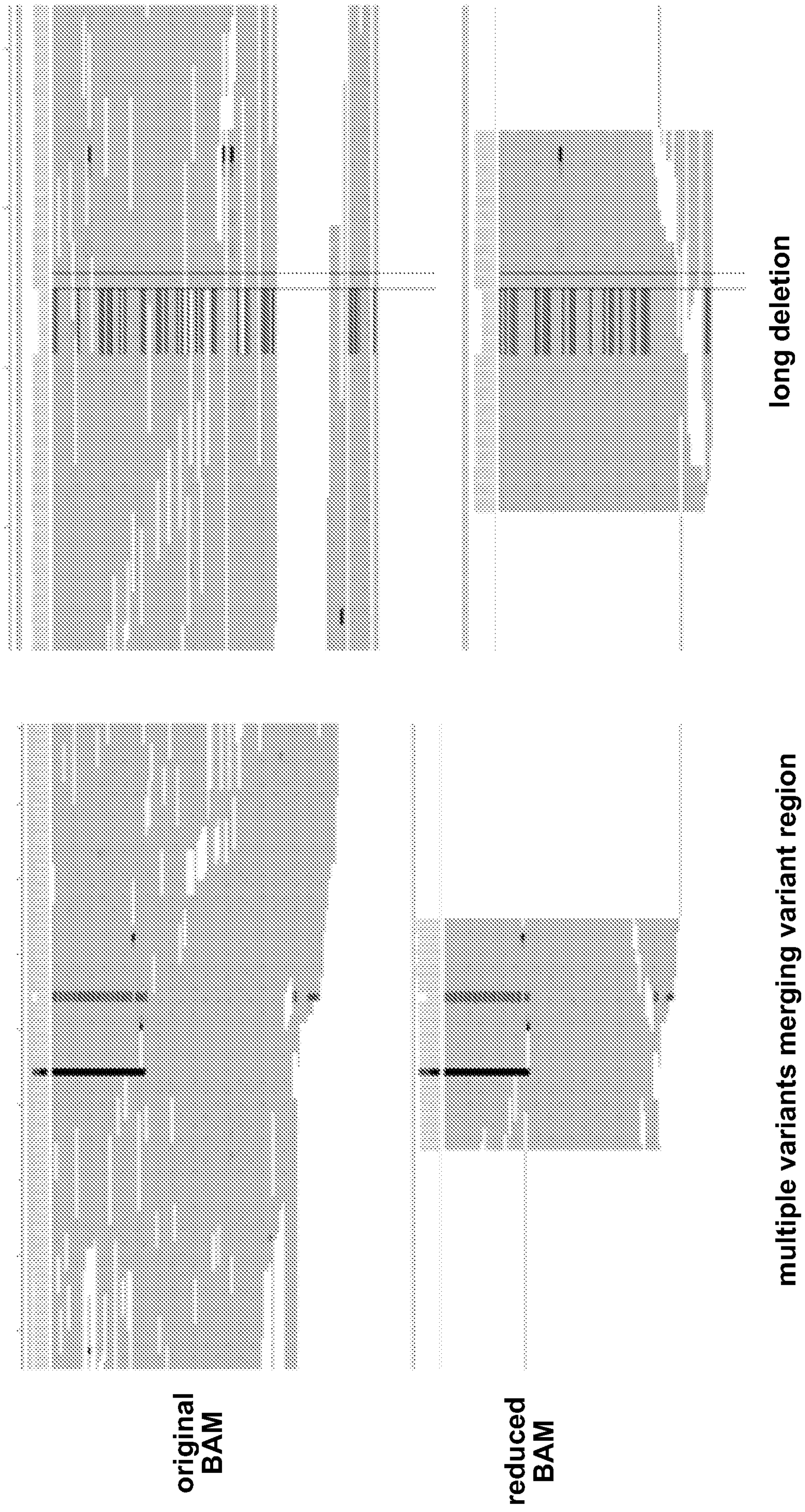
original BAM
reduced BAM
long deletion
multiple variants merging variant region

FIG. 4

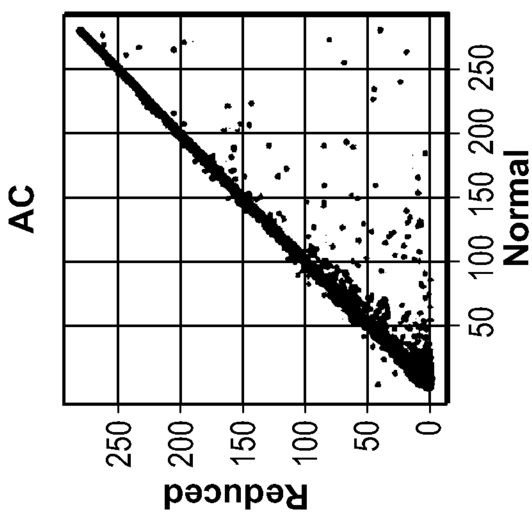

FIG. 5A

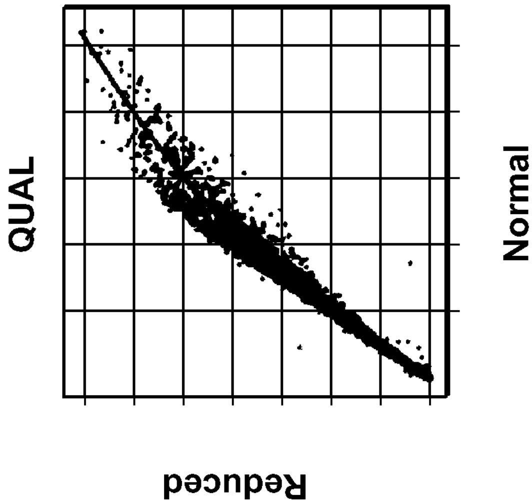

FIG. 5C

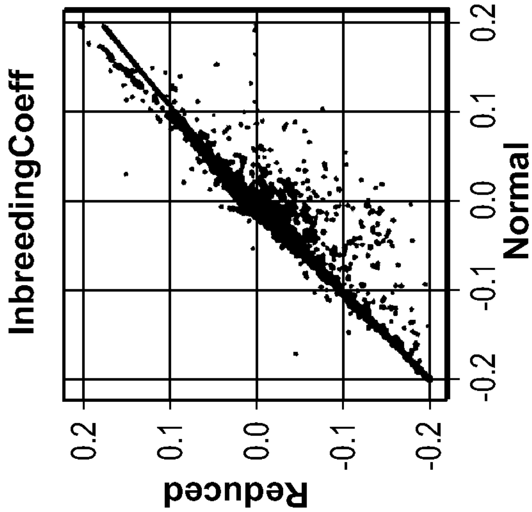

FIG. 5E

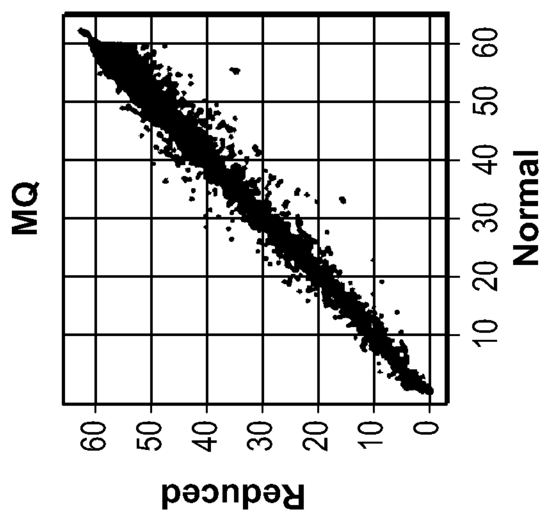

FIG. 5B

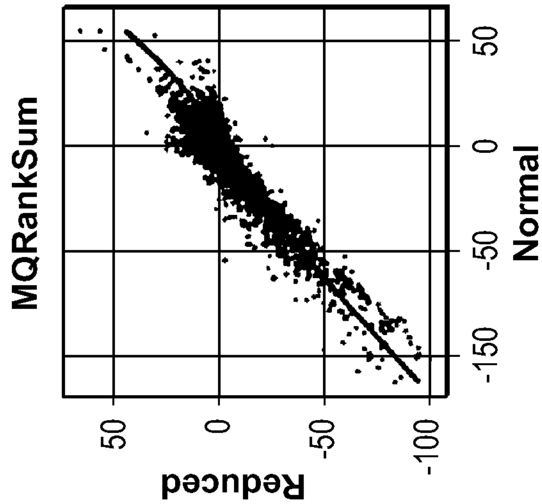

FIG. 5D

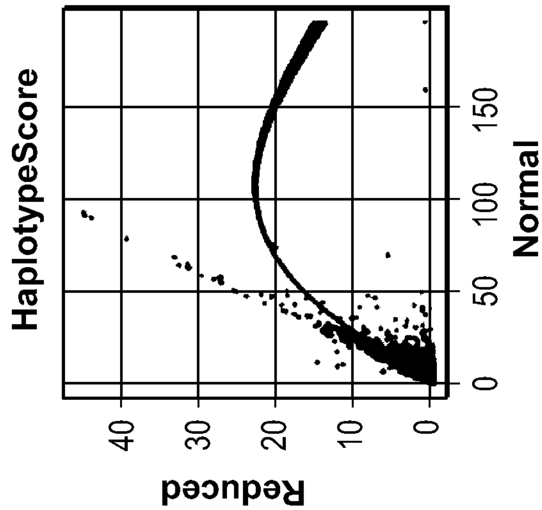

FIG. 5F

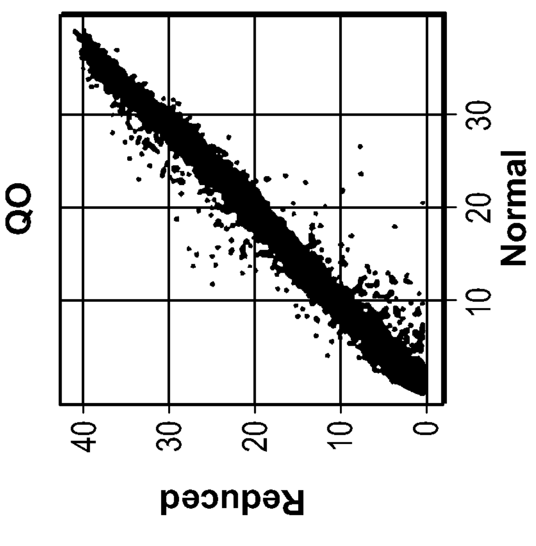

FIG. 5G

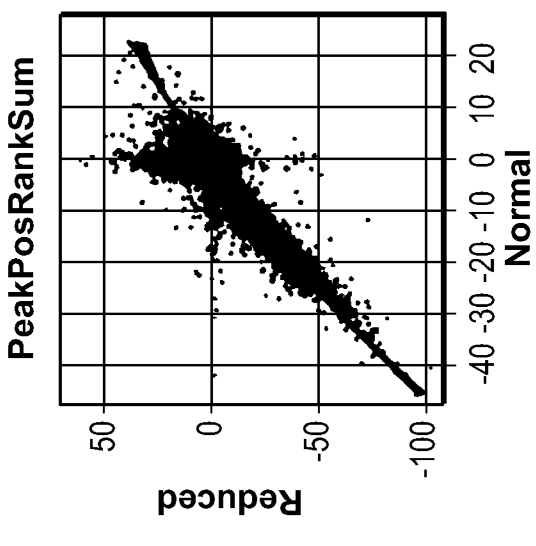

FIG. 5H

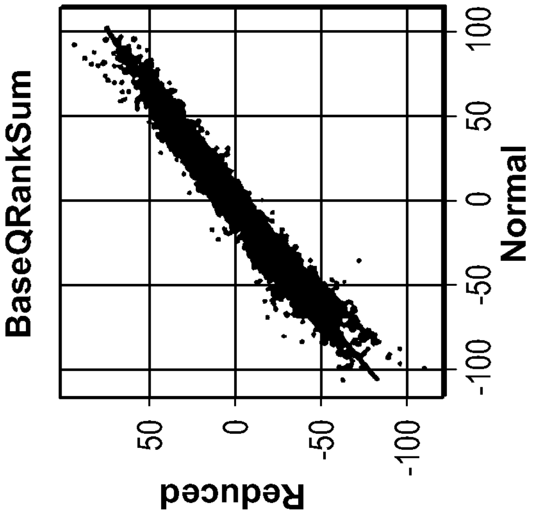

FIG. 5I

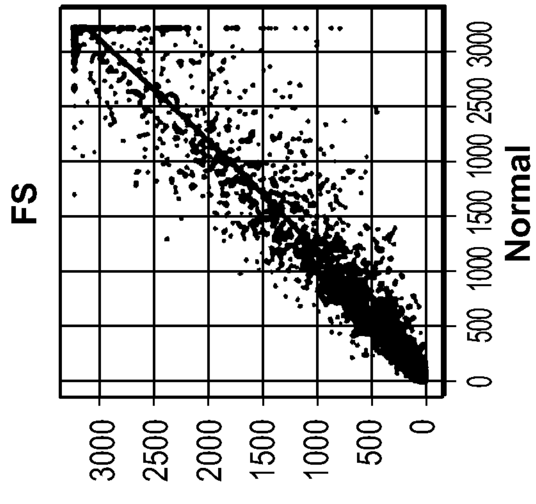

FIG. 5J

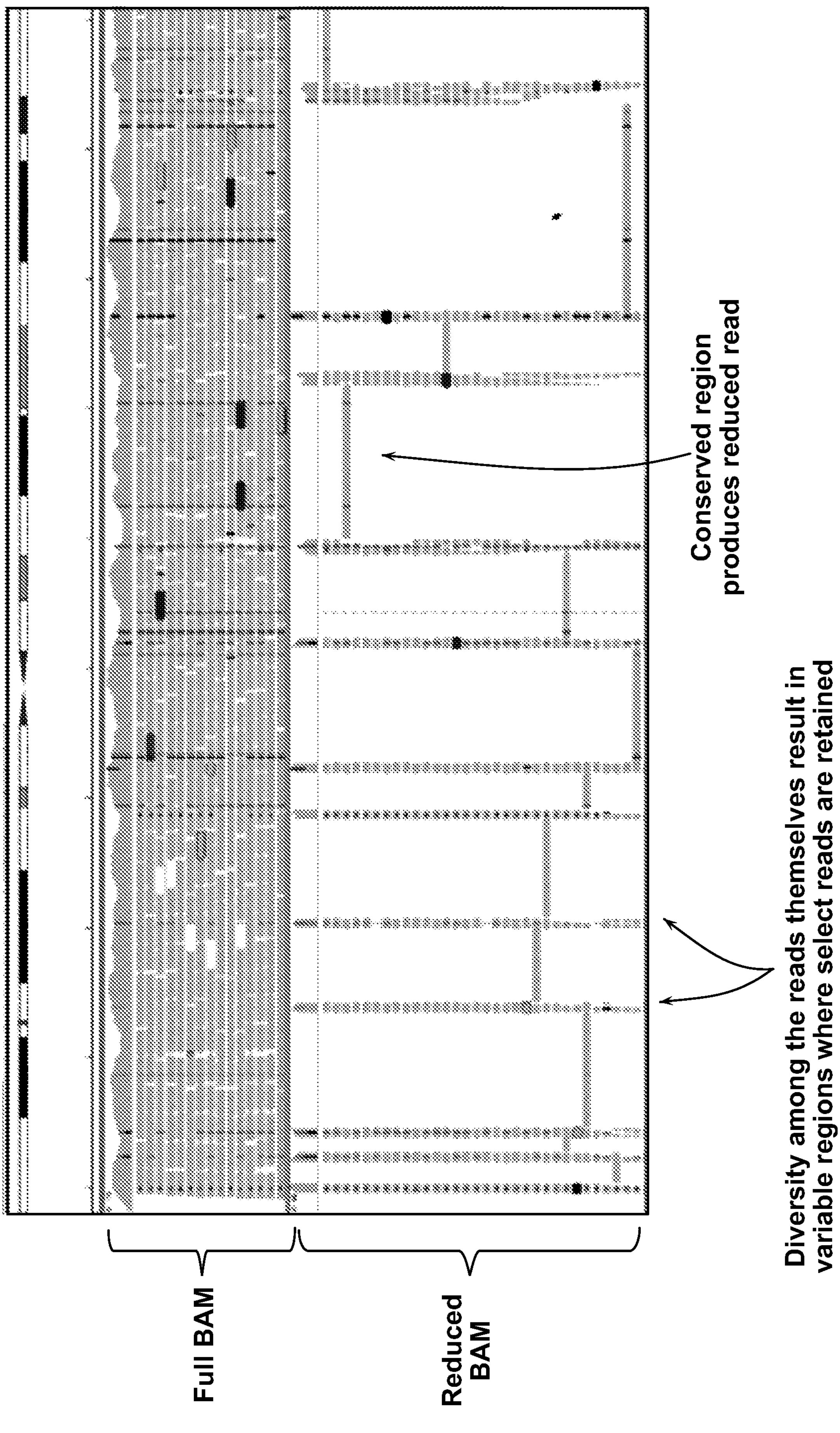
Full BAM
Reduced BAM
Conserved region produces reduced read
Diversity among the reads themselves result in variable regions where select reads are retained

FIG. 6

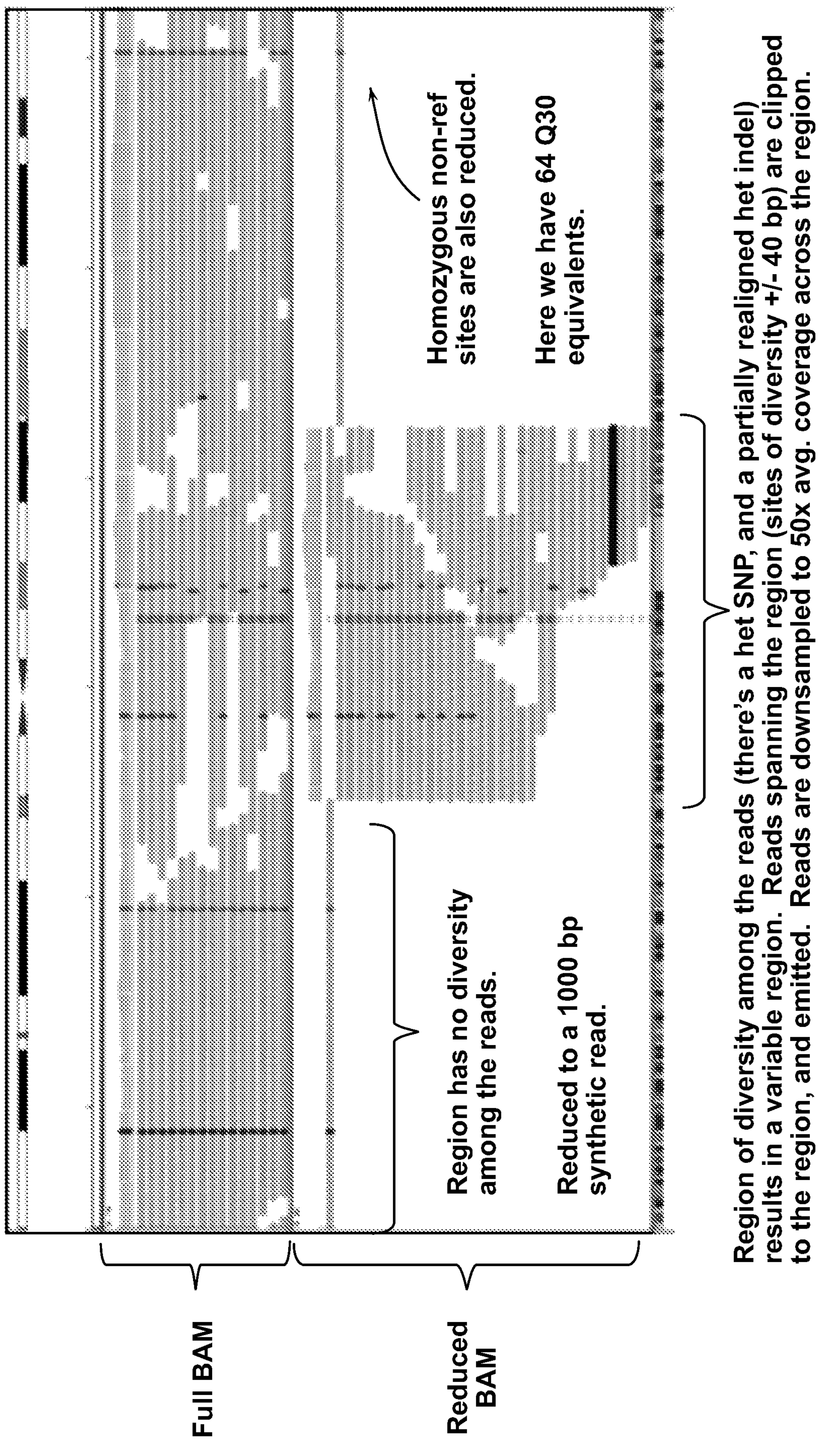
Full BAM
Reduced BAM
Region has no diversity among the reads.
Reduced to a 1000 bp synthetic read.
Homozygous non-ref sites are also reduced.
Here we have 64 Q30 equivalents.
Region of diversity among the reads (there's a het SNP, and a partially realigned het indel) results in a variable region. Reads spanning the region (sites of diversity +/- 40 bp) are clipped to the region, and emitted. Reads are downsampled to 50x avg. coverage across the region.

FIG. 7

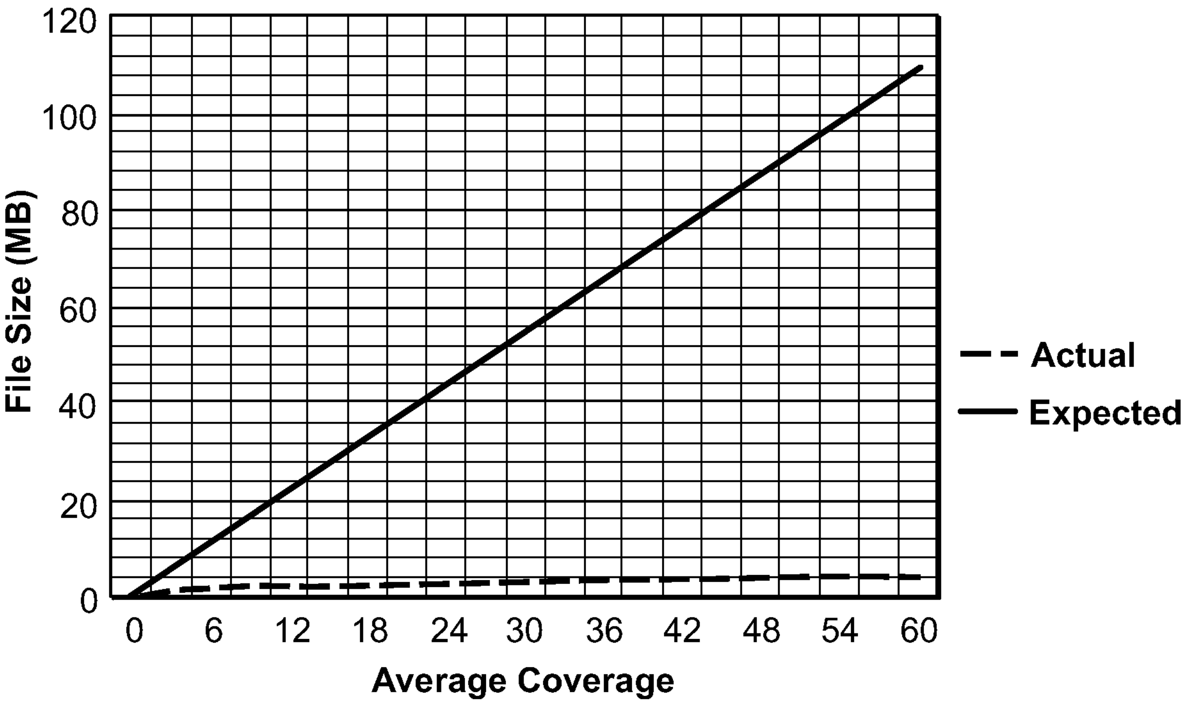
120
100
80
60
40
20
0
File Size (MB)
0
6
12
18
24
30
36
42
48
54
60
Average Coverage
Actual
Expected

FIG. 8

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2013/031429

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - C12Q 1/68 (2013.01)
USPC - 435/6.1
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC(8) - C07H 21/00, 21/04; 15/09, 15/10; C12P 19/34; C12Q 1/04, 1/68; C40B 30/00, 30/04; G01N 33/53, 33/543, 37/00 (2013.01)
USPC - 435/6.1, 6.11, 6.18, 91.2; 506/9; 536/23.1

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
CPC - C12Q 1/6806, 1/6809, 1/6827, 1/6834, 1/6855, 1/6858, 1/6869; C12N 15/10, 15/1093 (2013.01)

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
Patbase, Google Patent, PubMed

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	LI et al. "ANDES: Statistical tools for the Analyses of Deep Sequencing," BMC Research Notes, 15 July 2010 (15.07.2010), Vol. 3, Pgs. 199-210. entire document	1-17
A	US 2009/0233291 A1 (CHEN et al) 17 September 2009 (17.09.2009) entire document	1-17
A	WO 2004/029298 A2 (STOCKWELL et al) 08 April 2004 (08.04.2004) entire document	1-17
P, A	CARNEIRO et al. "Pacific biosciences sequencing technology for genotyping and variation discovery in human data," BMC Genomics, 05 August 2012 (05.08.2012), Vol. 13, Pgs. 375-381. entire document	1-17

☐ Further documents are listed in the continuation of Box C. ☐

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
06 June 2013	19 JUN 2013

Name and mailing address of the ISA/US	Authorized officer:
Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201	Blaine R. Copenheaver PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Box No. I **Nucleotide and/or amino acid sequence(s) (Continuation of item 1.c of the first sheet)**

1. With regard to any nucleotide and/or amino acid sequence disclosed in the international application, the international search was carried out on the basis of a sequence listing filed or furnished:

   a. (means)

      ☐ on paper

      ☐ in electronic form

   b. (time)

      ☐ in the international application as filed

      ☐ together with the international application in electronic form

      ☐ subsequently to this Authority for the purposes of search

2. ☐ In addition, in the case that more than one version or copy of a sequence listing has been filed or furnished, the required statements that the information in the subsequent or additional copies is identical to that in the application as filed or does not go beyond the application as filed, as appropriate, were furnished.

3. Additional comments:

ISA/225 mailed on 22 March 2013. No electronic sequence listing was submitted in response to the ISA/225.