



(51) International Patent Classification:

Not classified

(21) International Application Number:

PCT/US2024/035567

(22) International Filing Date:

26 June 2024 (26.06.2024)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

63/511,564 30 June 2023 (30.06.2023) US
63/517,160 02 August 2023 (02.08.2023) US

(71) Applicant: **ILLUMINA, INC.** [US/US]; 5200 Illumina Way, San Diego, California 92122 (US).

(72) Inventors: **FUHRMANN, Alexander**; 5200 Illumina Way, San Diego, California 92122 (US). **SANGIORGIO, Paul**; 5200 Illumina Way, San Diego, California 92122 (US). **COREY, Victoria**; 5200 Illumina Way, San Diego, California 92122 (US).

(74) Agent: **FUNK, Christopher C.** et al.; Keller Preece PLLC, 1010 North 500 East, Suite 210, North Salt Lake, Utah 84054 (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, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, 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, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, 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, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— without international search report and to be republished upon receipt of that report (Rule 48.2(g))

(54) Title: MODIFYING SEQUENCING CYCLES OR IMAGING DURING A SEQUENCING RUN TO MEET CUSTOMIZED COVERAGE ESTIMATION

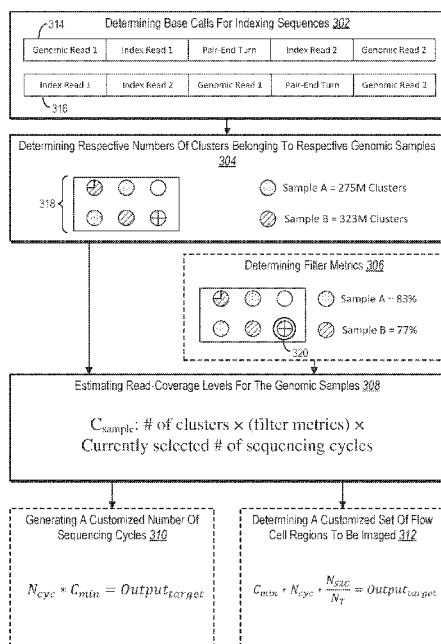


Fig. 3

(57) Abstract: This disclosure describes methods, non-transitory-computer readable media, and systems that can modify sequencing runs to ensure all genomic samples meet a target read-coverage level. The disclosed system can estimate read coverage for each genomic sample in a genomic pool based on (i) clusters belonging to each sample derived from indexing sequences and/or (ii) filter metrics corresponding to each sample within a flow-cell pool. The disclosed systems can modify a sequencing run based on the estimated read coverage and a target read coverage. For example, the disclosed systems can adjust a number of sequencing cycles within a sequencing run to ensure that all genomic samples meet the target read coverage. Additionally, or alternatively, the disclosed systems can determine a set of flow cell tiles to be imaged to ensure that all genomic samples meet the target read coverage.

MODIFYING SEQUENCING CYCLES OR IMAGING DURING A SEQUENCING RUN TO
MEET CUSTOMIZED COVERAGE ESTIMATION

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority to and the benefit of U.S. Provisional Patent Application No. 63/511,564, entitled “MODIFYING SEQUENCING CYCLES OR IMAGING DURING A SEQUENCING RUN TO MEET CUSTOMIZED COVERAGE ESTIMATION,” filed on June 30, 2023, and U.S. Provisional Patent Application No. 63/517,160, entitled “MODIFYING SEQUENCING CYCLES OR IMAGING DURING A SEQUENCING RUN TO MEET CUSTOMIZED COVERAGE ESTIMATION,” filed on August 2, 2023. Each of the aforementioned applications is hereby incorporated by reference in its entirety.

BACKGROUND

[0002] In recent years, biotechnology firms and research institutions have improved hardware and software for sequencing nucleotides and determining nucleobase calls for genomic samples. For instance, some existing sequencing machines and sequencing-data-analysis software (together “existing sequencing systems”) predict individual nucleobases within sequences by using conventional Sanger sequencing or sequencing-by-synthesis (SBS) methods. When using SBS, existing sequencing systems can monitor many thousands to billions of oligonucleotides being synthesized in parallel from templates to predict nucleobase calls for growing nucleotide reads. During a sequencing run in many existing sequencing systems, a camera captures images of irradiated fluorescent tags incorporated into oligonucleotides. After capturing such images, some existing sequencing systems determine nucleobase calls for nucleotide reads corresponding to respective clusters of oligonucleotides on a flow cell or other nucleotide-sample substrate for a given sequencing run. For example, some existing sequencing systems utilize sequencing-data-analysis software to analyze image data captured during sequencing cycles to determine nucleobase calls for given clusters of oligonucleotides and sequence such calls across sequencing cycles to determine nucleotide reads for the given clusters.

[0003] As part of such improved genomic sequencing, biotechnology firms and research institutions have also improved methods of simultaneously pooling and sequencing large numbers of genomic samples. Existing sequencing systems may pool genetic samples from different individuals to increase the number of samples analyzed in a single sequencing run. For instance, existing sequencing systems may utilize sample multiplexing (or multiplex sequencing) to add individual “barcode” or indexing sequences to each deoxyribonucleic acid (DNA) fragment during library preparation. The indexing sequences correspond to individual genomic samples within the sample pool. After the indexing sequences have been identified, existing sequencing systems may

perform demultiplexing to identify which indexing sequences—and which clusters of oligonucleotides on a flow cell—correspond with which genomic samples.

[0004] Despite recent advances in multiplexing and per-cycle image analysis, existing sequencing systems cannot accurately determine nucleotide-read coverage for a given genomic sample until after concluding a sequencing run and face other technical shortcomings that vary the level of nucleotide-read coverage for samples provided by read data from a given sequencing run. In multiplexed sequencing, for example, the number of nucleotide fragments from each genomic sample in clusters may not be evenly distributed, leading to variations in nucleotide-read depth or coverage. This uneven representation sometimes results in a sequencing device executing an insufficient number of sequencing cycles or images (or otherwise under-sequencing) for a sequencing run to generate the requisite numbers or length of nucleotide reads that satisfy a target level of coverage for a given sample. While sequencing devices can under-sequence DNA fragments extracted from some samples, sequencing devices can sometimes execute an excessive number of sequencing cycles or images (or otherwise over-sequence) for a sequencing run to generate the requisite numbers or length of nucleotide reads to satisfy the target coverage level.

[0005] Due to the uncertainty and variation of the read data coverage for a given sample produced by a given sequencing run, existing sequencing systems often inefficiently consume an inordinate amount of computing time, memory, and consumable materials to compensate for run-to-run variations. Some existing sequencing systems inefficiently consume an inordinate amount of computing time and memory to address under-sequenced samples. For instance, existing sequencing systems often perform additional sequencing cycles during a sequencing run to avoid under-sequencing some samples. The additional sequencing cycles require an excessive amount of computing time, memory, and reagents. As a result of performing additional sequencing cycles within a sequencing run, existing sequencing systems often over-sequence samples within a sample pool. While the addition of sequencing cycles results in fewer under-sequenced samples, existing sequencing systems cannot wholly eliminate under-sequenced samples. Thus, in addition to devoting excessive computing time and memory to over-sequencing samples, existing systems must also expend additional computing time to perform one or more additional sequencing runs to compensate for the under-sequenced samples of a previous sequencing run.

[0006] Likewise due to the uncertainty and variation of the read data coverage—and in addition to wasting processing time and memory—existing sequencing systems often inefficiently consume and waste excessive amounts of reagents, processing materials, and sample material during additional sequencing cycles or runs. By extending sequencing cycles to compensate for coverage uncertainty and sometimes performing additional sequencing runs to compensate for under-sequenced samples, existing sequencing systems consume extravagant amounts of

processing materials including sequencing reagents, library preparation kits, cluster amplification materials, flow cells or other nucleotide-sample substrates, scarce real estate on such flow cells, and other materials. In addition to consuming such materials, existing sequencing systems sometimes require re-extracting genomic material from an individual and re-performing library preparation necessary to seed oligonucleotide clusters on an additional flow cell to perform an additional sequencing run to compensate for a previous sequencing run that failed to produce a target nucleotide-read coverage for variant calling (or other secondary analysis) of the individual. For many existing systems, the relationship between number of cycles and processing materials consumed is a linear function. Thus, many existing sequencing systems consume inordinate amounts of processing materials and sample materials to compensate for the coverage uncertainty and variation outlined above.

[0007] Despite inefficiently utilizing computing resources and sequencing materials, some existing sequencing systems have theorized models to compensate for the uncertainty and variation of read-data coverage by modelling or attempting to implement a sequence-to-answer workflow. In theory, a sequence-to-answer workflow includes mapping and aligning read data for a genomic sample during a sequencing run including oligonucleotide clusters for the same genomic sample to determine nucleotide-read coverage for the sample in real time and to stop a sequencing run when the determined coverage satisfies a target. Such a sequence-to-answer workflow would require existing sequencing systems to transform raw sequencing data into meaningful nucleotide-read-coverage determinations through secondary analysis before the sequencing run concludes. In practice, however, such a sequence-to-answer workflow has yet to succeed at a commercial scale or with substantial improvements to reduced sequencing cycles or imaging. For example, existing sequencing systems have not developed computing models or hardware that enable such systems to generate data using sequencing devices, preprocess the data, transfer the data from sequencing devices to server devices to complete secondary analysis with sufficient speed or accurate nucleotide-read-coverage determinations to obtain the coverage answer before concluding a sequencing run. In other words, existing sequencing systems have yet to perform mapping and alignment of nucleotide reads (or other types of secondary analysis) accurately and concurrently during a corresponding sequencing run on a sequencing device in time enough to adjust the sequencing run before conclusion.

[0008] These, along with additional problems and issues exist in existing sequencing systems.

SUMMARY

[0009] This disclosure describes one or more embodiments of systems, methods, and non-transitory computer readable storage media that solve one or more of the problems described above or provide other advantages over the art. For example, the disclosed systems estimate read

coverage of genomic samples in a pool and adjusts the number of sequencing cycles to meet a target coverage based on the estimated read coverage. Additionally, or alternatively, the disclosed systems can determine a customized set of flow cell regions to be imaged from a flow cell to meet the target coverage. As part of generating an estimated read coverage, the disclosed systems may estimate variation arising from sample pooling and pass-filter variation.

[0010] To illustrate, in some embodiments, the disclosed systems perform indexing cycles to efficiently estimate respective numbers of clusters among samples within the pool. The disclosed system may also estimate pass-filter variation by generating a pass filter map comprising indications of whether oligonucleotide clusters for a sample pass a chastity filter (or other filters) for initial cycles of a sequencing run. Based on the respective numbers of clusters belonging to respective samples and the estimated numbers of clusters that pass filter, the disclosed systems can estimate read-coverage levels for individual genomic samples. The disclosed systems may further determine a customized number of sequencing cycles for the sequencing run sufficient to generate nucleotide reads satisfying a target read-coverage level for each genomic sample based on the estimated read-coverage levels. In some implementations, the disclosed systems determine a customized set of flow cell regions to be imaged from a flow cell sufficient to generate nucleotide reads satisfying a target read-coverage level. The disclosed systems further execute the sequencing run on the sequencing device by (i) finishing the customized number of sequencing cycles and/or (ii) capturing images of the customized set of flow cell regions (e.g., flow-cell tiles) during sequencing cycles of the sequencing run.

[0011] Additional features and advantages of one or more embodiments of the present disclosure will be set forth in the description which follows, and in part will be obvious from the description, or may be learned by the practice of such example embodiments.

BRIEF DESCRIPTION OF THE DRAWINGS

[0012] The detailed description refers to the drawings briefly described below.

[0013] FIG. 1 illustrates a computing system in which a sequencing device and a corresponding sequence-to-coverage system can operate in accordance with one or more embodiments of the present disclosure.

[0014] FIGS. 2A-2B illustrate potential read-coverage-level failures or other technical sequencing limitations arising from various sources of variation during sequencing runs.

[0015] FIG. 3 illustrates an overview of the sequence-to-coverage system modifying a number of sequencing cycles in a sequencing run or a number of images of flow cell regions in a sequencing run to meet a target read-coverage level in accordance with one or more embodiments of the present disclosure.

[0016] FIG. 4 illustrates the sequence-to-coverage system performing a subset of sequencing cycles with indexing cycles performed before genomic sequencing cycles in accordance with one or more embodiments of the present disclosure.

[0017] FIG. 5 illustrates the sequence-to-coverage system determining respective numbers of clusters of oligonucleotides belonging to respective genomic samples in accordance with one or more embodiments of the present disclosure.

[0018] FIG. 6 illustrates the sequence-to-coverage system determining filter metrics in accordance with one or more implementations of the present disclosure.

[0019] FIGS. 7A-7B illustrate the sequence-to-coverage system generating a customized number of sequencing cycles to meet a target read-coverage level and executing the sequencing run until finishing the customized number of sequencing cycles in accordance with one or more embodiments of the present disclosure.

[0020] FIG. 8 illustrates the sequence-to-coverage system determining a customized set of flow cell regions to be imaged during a sequencing run and executing the sequencing run by capturing images of the customized set of flow cell regions during sequencing cycles of the sequencing run in accordance with one or more embodiments of the present disclosure.

[0021] FIGS. 9A-9B illustrate improvements in sequencing efficiency resulting from execution of a customized number of sequencing cycles in accordance with one or more embodiments of the present disclosure.

[0022] FIG. 10 illustrates improvements in sequencing efficiency resulting from imaging a customized set of flow cell regions during sequencing cycles in accordance with one or more embodiments of the present disclosure.

[0023] FIG. 11 illustrates a schematic view of an example of a system that may be used to provide biological or chemical analysis in accordance with one or more embodiments of the present disclosure.

[0024] FIG. 12 illustrates a schematic view of an example of a set of components that may cooperate to provide a fluid path in the system of FIG. 11 in accordance with one or more embodiments of the present disclosure.

[0025] FIG. 13A illustrates a flowchart of a series of acts for executing a sequencing run until finishing a customized number of sequencing cycles in accordance with one or more embodiments of the present disclosure.

[0026] FIG. 13B illustrates a flowchart of a series of acts for executing a sequencing run by capturing images of a customized set of flow cell regions in accordance with one or more embodiments of the present disclosure.

[0027] FIG. 14 illustrates a block diagram of an example computing device in accordance with one or more embodiments of the present disclosure.

DETAILED DESCRIPTION

[0028] This disclosure describes one or more embodiments of a sequence-to-coverage system that can efficiently modify and execute a sequencing run to meet a target read-coverage level for genomic samples within a pool of genomic samples. For instance, the sequence-to-coverage system can determine, from a subset of sequencing cycles of a sequencing run for genomic samples, base calls for indexing sequences within clusters of oligonucleotides. The sequence-to-coverage system may further determine, based on the indexing sequences, respective numbers of clusters of oligonucleotides belonging to respective genomic samples of the genomic samples. Based on the respective numbers of clusters of oligonucleotides and a currently selected number (e.g., a preset number) of sequencing cycles for the sequencing run, the sequence-to-coverage system may estimate read-coverage levels for the genomic samples. The sequence-to-coverage system may further generate a customized number of sequencing cycles sufficient to generate nucleotide reads satisfying a target read-coverage level for each genomic sample of the genomic samples in the sequencing run. Additionally, or alternatively, the sequence-to-coverage system determines, based on the estimated read-coverage level, a customized set of flow cell regions of a flow cell (e.g., flow-cell tiles) to be imaged sufficient to generate nucleotide reads satisfying the target read-coverage level for each genomic sample of the genomic samples. The sequence-to-coverage system may execute the sequencing run on the sequencing device (i) until finishing the customized number of sequencing cycles and/or (ii) by capturing images of the customized set of flow cell regions during sequencing cycles of the sequencing run.

[0029] As just noted, the sequence-to-coverage system can determine, from a subset of sequencing cycles of a sequencing run for genomic samples, base calls for indexing sequences within clusters of oligonucleotides. In some cases, the sequence-to-coverage system expedites determining numbers of clusters of oligonucleotides belonging to respective genomic samples within a flow-cell pool (or other nucleotide-sample-substrate pool) by base calling the indexing sequences for both read pairs before base calling the genomic sequences in library templates for each sample.

[0030] Having determined base calls for the indexing sequences, the sequence-to-coverage system can determine respective numbers of clusters of oligonucleotides belonging to respective genomic samples. By demultiplexing the indexed reads to determine which indexing sequences belong to which genomic samples, the sequence-to-coverage system can quickly and efficiently estimate respective numbers of clusters corresponding to individual genomic samples within a pool. In some embodiments, the sequence-to-coverage system determines base calls for indexing

sequences (e.g., in both mates of paired-end reads) before determining base calls for genomic sequences of the nucleotide reads. In some embodiments, however, the sequence-to-coverage system determines a customized number of sequencing cycles or a customized set of flow cell regions to be imaged without finishing base calls for indexing sequences for each read before genomic sequences of each read.

[0031] As mentioned, the sequence-to-coverage system can estimate read-coverage levels based on (i) respective numbers of clusters of oligonucleotides belonging to respective genomic samples of the genomic samples in a sequencing run and (ii) a currently selected number of sequencing cycles for the sequencing run. Generally, the sequence-to-coverage system may utilize the respective numbers of clusters of oligonucleotides belonging to the respective genomic samples to estimate variation arising from imbalanced sample pooling. As explained further below, in some cases, the sequence-to-coverage system estimates an average number of nucleotide reads from a sequencing run sufficient to cover genomic regions of the individual genomic samples.

[0032] In some implementations, the sequence-to-coverage system further estimates read-coverage levels based on determined filter metrics. During a sequencing run, the sequence-to-coverage system can determine which clusters pass a chastity filter or otherwise determine other filter metrics that indicate subsets of clusters of oligonucleotides satisfying a filtering threshold for signals of the clusters of oligonucleotides. Based on determining such filter metrics, the sequence-to-coverage system can account for variations between genomic samples originating from low-quality or poor signal data.

[0033] Having estimated read-coverage levels, the sequence-to-coverage system can determine a customized number of sequencing cycles for a sequencing run sufficient to generate nucleotide reads that satisfy a target read-coverage level for each genomic sample of the genomic samples. For example, the sequence-to-coverage system can adjust a number of sequencing cycles during a sequencing run by increasing or decreasing a preset number of sequencing cycles for the sequencing run—before the sequencing run concludes. By generating the customized number of sequencing cycles, the sequence-to-coverage system can efficiently eliminate under-sequenced genomic samples and thereby avoid performing additional and unnecessary sequencing runs.

[0034] In combination with, or independent of, determining the customized number of sequencing cycles, the sequence-to-coverage system can also determine a customized set of flow cell regions to be imaged from a flow cell. More specifically, the sequence-to-coverage system can determine a customized set of flow cell regions to be imaged sufficient to generate nucleotide reads that satisfy a target read-coverage level for each genomic sample of the genomic samples. For instance, by demultiplexing nucleotide reads according to indexing sequence and determining

clusters that pass filter within a flow cell, the sequence-to-coverage system can estimate how many flow cell regions need to be imaged to satisfy a target read-coverage level.

[0035] Based on one or both of the customized number of sequencing cycles and the customized set of flow cell regions to be imaged, the sequence-to-coverage system can execute a sequencing run on a sequencing device to conclusion. For instance, the sequence-to-coverage system may execute the sequencing run on the sequencing device until finishing the customized number of sequencing cycles. Additionally, or alternatively, the sequence-to-coverage system may capture images of the customized set of flow cell regions during sequencing cycles of the sequencing run. By customizing the number of sequencing cycles and/or customizing the set of flow cell regions to be imaged, the sequence-to-coverage system can reduce consumable materials, sequencing-run time, and computing resources required to meet target read-coverage levels for each genomic sample.

[0036] As indicated above, the sequence-to-coverage system provides several technical advantages relative to existing sequencing systems by, for example, improving resource, sequence-run time, and computational efficiency relative to existing sequencing systems. In some implementations, for instance, the sequence-to-coverage system conserves sequencing cycles, imaging, consumables, and other physical resources—and reduces overuse of fluidics devices and other hardware within a sequencing device—relative to existing sequencing systems. To compensate for the read-data-coverage uncertainty and variant and otherwise satisfy target read-coverage levels for multiplexed samples described above, existing sequencing systems often duplicate sequencing cycles and sometimes perform additional sequencing runs. Such excessive sequencing cycles or runs can require additional run time and consume sequencing reagents, processing materials, and sample materials.

[0037] In contrast to such existing sequencing systems, the sequence-to-coverage system can efficiently generate a customized number of sequencing cycles and/or determine a customized set of flow cell regions to image before a sequencing run concludes and thereby execute the sequencing run according to the customized sequencing cycles or flow cell regions. In some examples, the sequence-to-coverage system can reduce one or both (i) the number of sequencing cycles and (ii) the number of flow cell regions imaged in a given sequencing run to satisfy a target read-coverage level. By tailoring parameters of a sequencing run based on a target read-coverage level, the sequence-to-coverage system can reduce the run time and the consumed physical resources (e.g., reagents) to achieve a target read-coverage level. Furthermore, by customizing the number of cycles and/or the number of flow cell regions imaged, the sequence-to-coverage system can avoid unnecessary wear and tear on the physical components of a sequencing device.

[0038] In addition to reducing the run time and consumed resources to achieve a target read-coverage level, the sequence-to-coverage system reduces the amount of compute time and consumed memory on a sequencing device for a given sequencing run to reach target read-coverage levels relative to existing sequencing systems. By estimating read-coverage levels before finishing a sequencing run, the sequence-to-coverage system can accurately execute a number of sequencing cycles required for each genomic sample to reach a target read-coverage level. Additionally, or alternatively, the sequence-to-coverage system can accurately estimate a set of flow cell regions that, when imaged during sequencing cycles, promotes a sequencing run that produces sufficient nucleotide reads for each genomic sample to reach the target read-coverage level. Relative to existing sequencing systems operating on existing sequencing devices, the sequence-to-coverage system can execute a lower number of sequencing cycles and/or image fewer flow cell regions that consume less processing and memory as a result of reduced sequencing-run time—while still achieving acceptable read-coverage levels for each genomic sample. Because of the intelligently reduced sequencing-run time, the sequence-to-coverage system can also reduce the amount of compute time required to perform a sequencing run that satisfies a target read-coverage level for genomic samples.

[0039] In addition to conserved run time, physical resources, and computing resources on a sequencing device for a given sequencing run, in some embodiments, the sequence-to-coverage system also improves computing efficiency and real-time flexibility relative to existing sequencing systems by determining real-time coverage estimates exclusively or primarily based on data generated by the sequencing device and not based on data (or based on relatively less data) from secondary analysis performed by another computing device. As mentioned, some existing sequencing systems have attempted to implement sequence-to-answer workflows that require secondary analysis and sometimes separate computing devices from a sequencing device to determine nucleotide-read coverage for individual samples before sequencing run concludes. But such sequence-to-answer workflows have failed to succeed at commercial scale or with substantial improvements to efficient sequencing runs (e.g., intelligently adjusting/reducing sequencing cycles or flow cell regions to be imaged, saving reagents or computer processing or memory). In contrast to unsuccessful sequence-to-answer workflows, the sequence-to-coverage system utilizes data obtained from primary analysis on a sequencing device to make customized determinations. For example, the sequence-to-coverage system can estimate read-coverage levels for individual genomic samples based on data available during primary analysis on a sequencing device. By determining base calls for indexing sequences and determining cluster numbers that pass filter for individual genomic samples as a basis for read-coverage estimates, the sequence-to-coverage system efficiently and extemporaneously customizes a sequencing run on a sequencing device to

avoid unnecessary sequencing cycles and/or unnecessary flow cell-region-image capture. In relying on data obtainable through primary analysis, the sequence-to-coverage system can obviate the need for further processing and exchanging data that has slowed and proved unsuccessful by existing sequencing systems that attempt a sequence-to-answer workflow.

[0040] As illustrated by the foregoing discussion, the present disclosure utilizes a variety of terms to describe features and advantages of the sequence-to-coverage system. As used herein, for example, the term “sequencing run” refers to an iterative process on a sequencing device to determine a primary structure of nucleotide sequences from a sample (e.g., genomic sample). In particular, a sequencing run includes cycles of sequencing chemistry and imaging performed by a sequencing device that incorporate nucleobases into growing oligonucleotides to determine nucleotide reads from nucleotide sequences extracted from a sample (or other sequences within a library fragment) and seeded throughout a flow cell. In some cases, a sequencing run includes replicating oligonucleotides derived or extracted from one or more genomic samples seeded in clusters throughout a flow cell. Upon completing a sequencing run, a sequencing device can generate base-call data in a file, such as a binary base call (BCL) sequence file or a fast-all quality (FASTQ) file.

[0041] Relatedly, as used herein, for example, the term “sequencing cycle” refers to an iteration of adding or incorporating one or more nucleobases to one or more oligonucleotides representing or corresponding to sample’s sequence (e.g., a genomic or transcriptomic sequence from a sample) or a corresponding adapter sequence. In some cases, a sequencing cycle includes an iteration of both incorporating nucleobases into clusters of oligonucleotides using sequencing chemistry and capturing images of such clusters attached to a flow cell. A sequencing cycle can include one or both of an indexing cycle and a genomic sequencing cycle. For instance, one cluster of oligonucleotides or a set of clusters of oligonucleotides may be undergoing a genomic sequencing cycle in which nucleobases corresponding to a sample genomic sequence are incorporated and another cluster of oligonucleotides or another set of clusters of oligonucleotides may be concurrently undergoing an indexing cycle in which nucleobases corresponding to an indexing sequence for a nucleotide read are incorporated.

[0042] As further used herein, the term “genomic sequencing cycle” refers to an iteration of adding or incorporating one or more nucleobases to one or more oligonucleotides representing or corresponding to a sample genomic sequence (or cDNA sequence). In particular, a genomic sequencing cycle can include an iteration of capturing and analyzing one or more images with data indicating individual nucleobases added or incorporated into an oligonucleotide or to oligonucleotides (in parallel) representing or corresponding to one or more sample genomic sequences. For example, in one or more embodiments, each genomic sequencing cycle involves

capturing and analyzing images to determine either single reads of DNA (or RNA) strands representing part of a genomic sample (or transcribed sequence from a genomic sample). As suggested above, however, a genomic sequencing cycle, in some cases, is specific to a cluster of oligonucleotides or a set of clusters of oligonucleotides.

[0043] By contrast, the term “indexing cycle” refers to an iteration of adding or incorporating one or more nucleobases to one or more oligonucleotides representing or corresponding to one or more indexing sequences. In particular, an indexing cycle can include an iteration of capturing and analyzing one or more images of clusters of oligonucleotides indicating one or more nucleobases added or incorporated into an oligonucleotide or to oligonucleotides (in parallel) representing or corresponding to one or more indexing sequences. An indexing cycle differs from a genomic sequencing cycle in that an indexing cycle includes sequencing of at least a nucleobase (or a majority of nucleobases) from one or more indexing sequences that identify or encode one or more sample library fragments. Because genomic sequencing cycles may be specific to a cluster or clusters of oligonucleotides, an indexing cycle for one cluster of oligonucleotides may be performed at a same time as a genomic sequencing cycle for another cluster of oligonucleotides.

[0044] Relatedly, the term “currently selected number of sequencing cycles” refers to an adjustable value that represents a number of sequencing cycles to be performed during a sequencing run. In particular, a currently selected number of sequencing cycles can be automatically determined, determined based on user selection, or preset according to a default number. For instance, the sequence-to-coverage system can determine a currently selected number of sequencing cycles equaling 150 sequencing cycles. The sequence-to-coverage system can adjust the number of sequencing cycles by increasing the number of sequencing cycles or reducing the number of sequencing cycles.

[0045] As used herein, the term “genomic sample” refers to a target genome or portion of a genome undergoing an assay or sequencing. For example, a genomic sample includes one or more sequences of nucleotides isolated or extracted from a sample organism (or a copy of such an isolated or extracted sequence). In particular, a genomic sample includes a full genome that is isolated or extracted (in whole or in part) from a sample organism and composed of nitrogenous heterocyclic bases. A genomic sample can include a segment of deoxyribonucleic acid (DNA), ribonucleic acid (RNA), or other polymeric forms of nucleic acids or chimeric or hybrid forms of nucleic acids noted below. In some cases, the genomic sample is found in a sample prepared or isolated by a kit and received by a sequencing device.

[0046] As used herein, the term “nucleobase call” (or simply “base call”) refers to a determination or prediction of a particular nucleobase (or nucleobase pair) for an oligonucleotide (e.g., read) during a sequencing cycle. In particular, a nucleobase call can indicate a determination

or prediction of the type of nucleobase that has been incorporated within an oligonucleotide on a flow cell (e.g., read-based nucleobase calls). In some cases, for a nucleotide read, a nucleobase call includes a determination or a prediction of a nucleobase based on intensity values resulting from fluorescent-tagged nucleotides added to one or more oligonucleotides of a flow cell (e.g., in a cluster of a flow cell). Alternatively, a nucleobase call includes a determination or a prediction of a nucleobase from chromatogram peaks or electrical current changes resulting from nucleotides passing through a nanopore of a flow cell. As suggested above, a single nucleobase call can be an adenine (A) call, a cytosine (C) call, a guanine (G) call, a thymine (T) call, or a uracil (U) call.

[0047] Further, as used herein, the term “sample library fragment” refers to a sample genomic sequence (or cDNA sequence) that is ligated to include one or more adapter sequences or primer sequences that facilitate detection or isolation of the sample genomic sequence or cDNA sequence. For instance, a sample library fragment can include, but is not limited to, a sample genomic sequence (or cDNA sequence) that is extracted from a sample and ligated to bond directly or indirectly with one or more of a binding adapter sequence, an indexing sequence, or a read priming sequence.

[0048] As used herein, the term “sample genomic sequence” refers to a nucleotide sequence extracted from, copied from, or complementary to a sample’s chromosome. For example, a sample genomic sequence includes a nucleotide sequence that has been separated or copied from chromosomal DNA of a sample or has been sequenced to be complementary to an extracted or copied nucleotide sequence. Accordingly, a sample genomic sequence includes genomic DNA (gDNA) for a particular unknown sample. Accordingly, as described herein, in some embodiments, the sequence-to-coverage system can use a sample complementary sequence comprising cDNA rather than a sample genomic sequence comprising gDNA in a sample library fragment or wherever suitable cDNA may replace gDNA as understood by a skilled artisan. Indeed, any embodiment or nucleotide read in this disclosure that uses or includes a sample genomic sequence can also use or include a cDNA sequence corresponding to a genomic sample.

[0049] As used herein, the term “indexing sequence” refers to a unique and artificial nucleotide sequence that identifies nucleotide reads for a sample and that is ligated to a sample’s nucleotide sequence (e.g., a gDNA fragment or cDNA fragment) or to another sequence within a sample library fragment. As indicated above, an indexing sequence can be part of a sample library fragment. Similarly, an indexing sequence can be used to sort nucleotide reads by sample or into different files, among other things, such as part of a de-multiplexing process. In some cases, a sample library fragment includes an indexing primer sequence that differs from a read priming sequence and that indicates a starting point or starting nucleobase for determining nucleobases of an indexing sequence.

[0050] As used herein, the term “cluster of oligonucleotides” refers to a localized collection of DNA or RNA molecules immobilized on a solid surface. In particular, a cluster of oligonucleotides can refer to a collection of fragment nucleotide sequences immobilized on a flow cell region of a flow cell. For example, a cluster of oligonucleotides can refer to a collection of nucleotide fragments originating from a genomic sample. A cluster of oligonucleotides can be imaged utilizing one or more light signals. For instance, an oligonucleotide-cluster image may be captured by a camera during a sequencing cycle of light emitted by irradiated fluorescent tags incorporated into oligonucleotides from one or more clusters on a flow cell.

[0051] As used herein, the term “nucleotide read” refers to an inferred sequence of one or more nucleobases (or nucleobase pairs) from all or part of a sample nucleotide sequence (e.g., a sample genomic sequence, complementary DNA). In particular, a nucleotide read includes a determined or predicted sequence of nucleobase calls for a nucleotide sequence (or group of monoclonal nucleotide sequences) from a sample library fragment corresponding to a genomic sample. For example, in some cases, a sequencing device determines a nucleotide read by generating nucleobase calls for nucleobases passed through a nanopore of a flow cell, determined via fluorescent tagging, or determined from a cluster in a flow cell.

[0052] As used herein, the term “read-coverage level” refers to a measure or value that indicates a depth or redundancy of nucleotide-sequence information for a particular genomic coordinate or genomic region of a sample. In particular, read-coverage level refers to a number of times a specific genomic coordinate or genomic region for a sample is covered or spanned by nucleotide reads. Read-coverage level can be relevant when describing the depth of sequencing data obtained for a particular genomic region of interest or a particular genomic sample. For example, read-coverage level may comprise a numeric value (e.g., 10x, 30x, 45x) indicating an average number of unique nucleotide reads for a genomic sample that span or cover genomic coordinates or regions of a human genomic sample. In some cases, read-coverage level is limited to an average number of unique nucleotide reads across a non-N portion of a human genome (e.g., non-N portion of a PAR-masked human genome).

[0053] As used herein, the term “target read-coverage level” refers to a desired or intended depth of sequencing coverage for a specific genomic coordinate or genomic region within a genomic sample. In particular, a target read-coverage level represents a minimum number of times a position within a genomic sample should be sequenced to achieve a desired level of confidence in the accuracy of the obtained sequence data. For example, a target-read-coverage level can comprise a numeric value (e.g., 40) indicating a desired read-coverage level for a given position within a genomic sample.

[0054] As further used herein, the term “genomic coordinate” (or sometimes simply “coordinate”) refers to a particular location or position of a nucleobase within a genome (e.g., an organism’s genome or a reference genome). In some cases, a genomic coordinate includes an identifier for a particular chromosome of a genome and an identifier for a position of a nucleobase within the particular chromosome. For instance, a genomic coordinate or coordinates may include a number, name, or other identifier for a chromosome (e.g., chr1, chrX, chrM) and a particular position or positions, such as numbered positions following the identifier for a chromosome (e.g., chr1:1234570 or chr1:1234570-1234870). In some cases, a genomic coordinate refers to a genomic coordinate on a sex chromosome (e.g., chrX or chrY) or mitochondrial DNA (e.g., chrM). Further, in certain implementations, a genomic coordinate refers to a source of a reference genome (e.g., mt for a mitochondrial DNA reference genome or SARS-CoV-2 for a reference genome for the SARS-CoV-2 virus) and a position of a nucleobase within the source for the reference genome (e.g., mt:16568 or SARS-CoV-2:29001). By contrast, in certain cases, a genomic coordinate refers to a position of a nucleobase within a reference genome without reference to a chromosome or source (e.g., 29727).

[0055] As used herein, a “genomic region” refers to a range of genomic coordinates. Like genomic coordinates, in certain implementations, a genomic region may be identified by an identifier for a chromosome and a particular position or positions, such as numbered positions following the identifier for a chromosome (e.g., chr1:1234570-1234870). In various implementations, a genomic coordinate includes a position within a reference genome. In some cases, a genomic coordinate is specific to a particular reference genome.

[0056] As used herein, the term “sequencing device” refers to an instrument or platform used to perform a sequencing process. In particular, a sequencing device refers to an instrument or platform used to perform a sequencing process based on sequencing by synthesis (SBS) technology, single-molecule real-time sequencing (SMRT) technology using magnetic beads or nanopores or other suitable medium. For example, a sequencing device may comprise components including, but not limited to, -flow cell receptacle, fluidics systems, lasers, imaging systems, and computational capabilities for acquiring, processing, and analyzing image data during a sequencing run.

[0057] As used herein, the term “filter metric” refers to a measure indicating a quality and reliability of sequencing data from clusters of oligonucleotides. In particular, a filter metric may comprise a value indicating the quality and/or brightness of sequencing data that has passed a certain filtering criterion. A filter metric may indicate a subset of imaged clusters of oligonucleotides that satisfy a filtering threshold for signals of the clusters of oligonucleotides. For

example, a filter metric may comprise a percent passing filter (%PF) that represents the percentage of clusters of oligonucleotides that pass a chastity filter.

[0058] As used herein, the term “filtering threshold” refers to a predetermined value or range of values used to determine whether a parameter meets a filtering standard. In particular, a filtering threshold may comprise a numerical value above (or below) which filtering metrics indicate an acceptable quality. Clusters having filtering values over the filtering threshold can be considered to pass filter. For example, a filtering threshold may comprise a threshold chastity value. A chastity value may comprise the ratio of the brightest base intensity divided by the sum of the brightest and second brightest base intensities within a cluster of oligonucleotides. The sequence-to-coverage system may determine that clusters of oligonucleotides having chastity values below the filtering threshold do not pass filter and remove them from image analysis results. To illustrate, a cluster may pass the filtering threshold if no more than 1 base call has a chastity value below 0.6.

[0059] As used herein, the term “nucleotide-sample substrate” refers to a plate or substrate, such as a flow cell, comprising oligonucleotides for sequencing nucleotide sequences from genomic samples or other sample nucleic-acid polymers. In particular, a flow cell can refer to a substrate containing fluidic channels through which reagents and buffers can travel as part of sequencing. For example, in one or more embodiments, the flow cell (e.g., a patterned flow cell or non-patterned flow cell) may comprise small fluidic channels and oligonucleotide samples that can be bound to adapter sequences on the substrate. In other implementations, a flow cell can be an open substrate with one or more regions for oligonucleotide samples to be analyzed and the oligonucleotide samples may be positioned using charged pads or other means. In yet another implementation, the nucleotide-sample substrate can be a membrane having a nanopore through which one or more oligonucleotide samples may pass. As indicated above, a flow cell can include tiles and wells (e.g., nanowells) comprising clusters of oligonucleotides.

[0060] As suggested above, a flow cell or other nucleotide-sample substrate can (i) include a device having a lid extending over a reaction structure to form a flow channel therebetween that is in communication with a plurality of reaction sites of the reaction structure and (ii) include a detection device that is configured to detect designated reactions that occur at or proximate to the reaction sites. A flow cell or other nucleotide-sample substrate may include a solid-state light detection or imaging device, such as a Charge-Coupled Device (CCD) or Complementary Metal-Oxide Semiconductor (CMOS) (light) detection device. As one specific example, a flow cell may be configured to fluidically and electrically couple to a cartridge (having an integrated pump), which may be configured to fluidically and/or electrically couple to a bioassay system. A cartridge and/or bioassay system may deliver a reaction solution to reaction sites of a flow cell according to a predetermined protocol (e.g., sequencing-by-synthesis), and perform a plurality of imaging

events. For example, a cartridge and/or bioassay system may direct one or more reaction solutions through the flow channel of the flow cell, and thereby along the reaction sites. At least one of the reaction solutions may include four types of nucleotides having the same or different fluorescent labels. The nucleotides may bind to the reaction sites of the flow cell, such as to corresponding oligonucleotides at the reaction sites. The cartridge and/or bioassay system may then illuminate the reaction sites using an excitation light source (e.g., solid-state light sources, such as light-emitting diodes (LEDs)). The excitation light may provide emission signals (e.g., light of a wavelength or wavelengths that differ from the excitation light and, potentially, each other) that may be detected by the light sensors of the flow cell.

[0061] As used herein, the term “flow cell region” refers to a region of a nucleotide-sample substrate. In particular, a flow cell region refers to an area or section of a flow cell that contains one or more clusters of oligonucleotides. For example, a flow cell region may refer to a tile of a flow cell. More specifically, flow cell regions may be organized in a grid-like pattern across a nucleotide-sample substrate, and each flow cell region corresponds to a specific position on the surface of the nucleotide-sample substrate. Flow cell regions may further contain wells (e.g., nanowells) comprising individual compartments where clusters of oligonucleotides are amplified, denatured, and subjected to sequencing.

[0062] As further used herein, the term “nucleotide read” (or simply “read”) refers to an inferred sequence of one or more nucleobases (or nucleobase pairs) from all or part of a sample nucleotide sequence (e.g., a sample genomic sequence, cDNA). In particular, a nucleotide read includes a determined or predicted sequence of nucleobase calls for a nucleotide sequence (or group of monoclonal nucleotide sequences) from a sample library fragment corresponding to a genomic sample. For example, in some cases, a sequencing device determines a nucleotide read by generating nucleobase calls for nucleobases determined via fluorescent tagging, passed through a nanopore of a nucleotide-sample substrate, or determined from a cluster in a flow cell.

[0063] As used herein, the term “nucleobase” refers to a nitrogenous base. In particular, nucleobases comprise components of nucleotides. For example, a nucleobase may be an adenine (A), cytosine (C), guanine (G), thymine (T), or uracil (U).

[0064] The following paragraphs describe a sequence-to-coverage system with respect to illustrative figures that portray example embodiments and implementations. For example, FIG. 1 illustrates a schematic diagram of a computing system 100 in which a sequence-to-coverage system 106 operates in accordance with one or more embodiments. As illustrated, the computing system 100 includes a local server device 102 connected to one or more server device(s) 110, a sequencing device 108, and a client device 114 via a network 112. While FIG. 1 shows an embodiment of the

sequence-to-coverage system 106, this disclosure describes alternative embodiments and configurations below.

[0065] As shown in FIG. 1, the local server device 102, the sequencing device 108, the server device(s) 110, and the client device 114 can communicate with each other via the network 112. The network 112 comprises any suitable network over which computing devices can communicate. Example networks are discussed in additional detail below with respect to FIG. 14.

[0066] As indicated by FIG. 1, the sequencing device 108 comprises a device for sequencing a genomic sample or other nucleic-acid polymer. In some embodiments, the sequencing device 108 analyzes nucleic-acid segments or oligonucleotides extracted from genomic samples to generate nucleotide reads or other data utilizing computer implemented methods and systems (described herein) either directly or indirectly on the sequencing device 108. More particularly, the sequencing device 108 receives nucleotide-sample substrates (e.g., flow cells) comprising nucleotide fragments extracted from samples and then copies and determines the nucleotide-base sequence of such extracted nucleotide fragments. In one or more embodiments, the sequencing device 108 utilizes SBS to sequence nucleic-acid polymers into nucleotide reads. Additionally, the sequencing device 108 can determine base calls for indexing sequences. In addition, or in the alternative to communicating across the network 112, in some embodiments, the sequencing device 108 bypasses the network 112 and communicates directly with the local server device 102 or the client device 114.

[0067] As further indicated by FIG. 1, the local server device 102 is located at or near a same physical location of the sequencing device 108. Indeed, in some embodiments, the local server device 102 and the sequencing device 108 are integrated into a same computing device, as indicated by dotted lines 122. The local server device 102 may run a sequencing system 104 to generate, receive, analyze, store, and transmit digital data, such as by receiving base-call data or determining indexing sequence data or filter metric data based on analyzing such base-call data. As shown in FIG. 1, the sequencing device 108 may send (and the local server device 102 may receive) base-call data generated during a sequencing run of the sequencing device 108. By executing software from the form of the sequencing system 104, the local server device 102 may estimate read-coverage levels for genomic samples in a pool of genomic samples. The local server device 102 may also communicate with the client device 114. In particular, the local server device 102 can send data to the client device 114, including read-coverage information for genomic samples, filter metric data, estimated read-coverage levels, a variant call file (VCF), or other information indicating nucleobase calls, genotype calls, sequencing metrics, error data, or other metrics.

[0068] As further indicated by FIG. 1, the server device(s) 110 are located remotely from the local server device 102 and the sequencing device 108. The sequencing device 108 may send (and

the server device(s) 110 may receive) base-call data from the sequencing device 108. The server device(s) 110 may also communicate with the client device 114. In particular, the server device(s) 110 can send data to the client device 114, including estimated read-coverage levels for genomic samples, VCFs, or other sequencing related information.

[0069] In some embodiments, the server device(s) 110 comprise a distributed collection of servers where the server device(s) 110 include a number of server devices distributed across the network 112 and located in the same or different physical locations. Further, the server device(s) 110 can comprise a content server, an application server, a communication server, a web-hosting server, or another type of server.

[0070] As further illustrated and indicated in FIG. 1, the client device 114 can generate, store, receive, and send digital data. In particular, the client device 114 can receive read-coverage data from the local server device 102 or receive sequencing metrics from the sequencing device 108. Furthermore, the client device 114 may communicate with the local server device 102 or the server device(s) 110 to receive a VCF comprising variant or genotype calls and/or other metrics, such as a base-call-quality metrics or pass-filter metrics. The client device 114 can accordingly present or display information pertaining to variant calls or other genotype calls within a graphical user interface to a user associated with the client device 114. For example, the client device 114 can present a target read-coverage interface comprising elements indicating potential target read-coverage levels for genomic samples.

[0071] Although FIG. 1 depicts the client device 114 as a desktop or laptop computer, the client device 114 may comprise various types of client devices. For example, in some embodiments, the client device 114 includes non-mobile devices, such as desktop computers or servers, or other types of client devices. In yet other embodiments, the client device 114 includes mobile devices, such as laptops, tablets, mobile telephones, or smartphones. Additional details regarding the client device 114 are discussed below with respect to FIG. 6.

[0072] As further illustrated in FIG. 1, the client device 114 includes a sequencing application 116. The sequencing application 116 may be a web application or a native application stored and executed on the client device 114 (e.g., a mobile application, desktop application). The sequencing application 116 can include instructions that (when executed) cause the client device 114 to receive data from the sequence-to-coverage system 106 and present, for display at the client device 114, data concerning read-coverage data for a sequencing run, data from a VCF, or other information. Furthermore, the sequencing application 116 can instruct the client device 114 to display graphical user interfaces for receiving input indicating a target read-coverage level.

[0073] As further illustrated in FIG. 1, a version of the sequence-to-coverage system 106 may be located on the client device 114 as part of the sequencing application 116. Accordingly, in some

embodiments, the sequence-to-coverage system 106 is implemented by (e.g., located entirely or in part) on the client device 114. In yet other embodiments, the sequence-to-coverage system 106 is implemented by one or more other components of the computing system 100, such as the server device(s) 110. In particular, the sequence-to-coverage system 106 can be implemented in a variety of different ways across local server device 102, the sequencing device 108, the client device 114, and the server device(s) 110. For example, the sequence-to-coverage system 106 can be downloaded from the server device(s) 110 to the local server device 102 and/or the client device 114 where all or part of the functionality of the sequence-to-coverage system 106 is performed at each respective device within the computing system 100.

[0074] As mentioned previously, sample multiplexing offers several advantages but is also associated with some potential technical challenges. FIGS. 2A-2B illustrate read-coverage-level failures or other technical sequencing limitations arising from various sources of variation during sequencing runs. FIG. 2A illustrates a chart portraying various sources of variation within sequencing runs. FIG. 2B illustrates how existing sequencing systems can both over- and under-sequence genomic samples due to sources of variation.

[0075] FIG. 2A illustrates a chart 200 portraying various sources of variation within sequencing runs. The chart 200 comprises a sector 202, a sector 204, and a sector 206. As shown by the sector 202, the most significant source of variation within a sequencing run arises from sample pooling. Sample pooling refers to the practice of combining multiple individual genomic samples into a single genomic pool before performing sequencing reactions. As described previously, sample pooling improves sequencing and computing efficiency by sequencing a plurality of genomic samples during a single sequencing run. However, some variations that may arise from sample pooling comprise unequal representation where genomic samples are unequally represented. Furthermore, sample pooling may introduce additional contamination to a sequencing run. For instance, genetic material from one genomic sample may inadvertently cross-contaminate other samples in the pool. As a result, sample pooling accounts for most of the total variation within sequencing runs having a coefficient of variation (CV) of approximately 10-15%.

[0076] As further shown by the sector 206 illustrated in FIG. 2A, pass-filter failures account for a significant portion of variation in sequencing runs. As shown, clusters of oligonucleotides that fail to pass filter account for approximately 2-5% of total variation between sequencing runs. The sector 206 represents variation arising from sources relating to the quality of sample preparation. In particular, pass-filter variation can arise due to factors including read quality (e.g., base quality scores, read-alignment scores, etc.) arising from sequencing chemistry, cycle-specific biases, or differences in the quality of input genomic samples. Pass-filter variation can further arise from different sequencing platforms. For example, different sequencing platforms may use unique

sequencing chemistries that result in variations in data quality. Furthermore, pass-filter variation may arise from varying experimental conditions such as reagent lots, laboratory protocols, and differences in the performance of quality of sequencing reagents, equipment, or other environmental factors. Additionally, pass-filter variation may be affected by sample heterogeneity where individual genomic samples within a pool of genomic samples may have varying quality, or sequencing complexity, which impacts observed pass filter metrics.

[0077] FIG. 2A also illustrates the sector 204 within the chart 200. As shown in FIG. 2A, the sector 204 comprises bioinformatic efficiency. Generally, bioinformatics efficiency refers to the ability to perform accurate secondary analysis of sequencing data. In particular, bioinformatic efficiency involves employing efficient algorithms, optimized computational resources, and streamlined processes to interpret sequencing data in a cost-effective manner. For example, issues in aligning nucleotide reads with a reference genome may result in bioinformatics efficiency variation. In some cases, bioinformatics efficiency is measured by (i) unique, aligned nucleotide reads corresponding to one or more genomic samples divided by (ii) a total number nucleotide reads from clusters that pass filter for the one or more genomic samples. As shown in FIG. 2A, bioinformatic efficiency accounts for approximately 2-5% of total variation between sequencing runs. In some examples, bioinformatics efficiency improves with (slightly lower) % pass filter values and thus can compensate, to some extent, for lower filter metrics.

[0078] Existing sequencing systems often attempt to compensate for variations in sequencing data by performing additional sequencing cycles, reducing a number of genomic samples in a pool of genomic samples, and/or (worse yet) performing additional sequencing runs. FIG. 2B illustrates a graph of sequencing data generated by existing sequencing systems. FIG. 2B illustrates a graph 208 portraying how existing systems often both over- and under-sequence genomic samples within a pool of genomic samples. The graph 208 comprises a bar graph portraying a distribution of a number of sequencing runs corresponding with read-coverage levels for the worst performing sample in the sequencing runs. As shown, the x-axis comprises unique aligned reads in gigabases (Gb) for the worst performing sample in each of the sequencing runs.

[0079] To avoid insufficient sequencing coverage for certain genomic samples, existing sequencing systems tend to over-sequence most samples by performing additional sequencing cycles. As shown in FIG. 2B, the target read-coverage level equals 40x, which corresponds to about 120 Gb. As further shown in FIG. 2B, existing systems typically sequence the worst-performing genomic samples around 15% higher than the target read-coverage level, which results in about 138 Gb. More particularly, approximately 95% of sequencing runs using existing system yield more than the target 40x read-coverage level.

[0080] As further shown in FIG. 2B, while the majority of sequencing runs are over-sequenced, about 5% of sequencing runs still yield an under-performing genomic sample. Existing systems must typically perform additional sequencing runs to re-sequence the under-performing genomic samples. Accordingly, existing systems not only utilize an excess of resources to over-sequenced most genomic samples, existing sequencing systems often need to perform additional sequencing runs to correct under-sequenced samples.

[0081] As mentioned previously, the sequence-to-coverage system 106 may customize a sequencing run by increasing or decreasing sequencing cycles or by increasing or decreasing flow cell regions to be imaged, thereby efficiently meeting a target read-coverage level before conclusion of the sequencing run. FIG. 3 illustrates an overview of the sequence-to-coverage system 106 modifying a sequencing run to meet a target read-coverage level in accordance with one or more embodiments of the present disclosure. FIG. 3 illustrates a series of acts 300 comprising an act 302 of determining base calls for indexing sequences, an act 304 of determining respective numbers of clusters belonging to respective genomic samples, an act 306 of determining filter metrics, an act 308 of estimating read-coverage levels for the genomic samples, an act 310 of generating a customized number of sequencing cycles, and an act 312 of determining a customized set of flow cell regions to be imaged.

[0082] As shown in FIG. 3, the sequence-to-coverage system 106 performs the act 302 of determining base calls for indexing sequences. The sequence-to-coverage system 106 determines base calls for indexing sequences within clusters of oligonucleotides. By determining base calls for indexing sequences, the sequence-to-coverage system 106 can accurately assign nucleotide reads to their respective genomic samples in multiplexed sequencing. The sequence-to-coverage system 106 may determine base calls for indexing sequences at different times relative to determining base calls for nucleotide reads of a genomic sample. FIG. 3 illustrates a non-indexing-first workflow 314 and an indexing-first workflow 316 comprising different order of indexing cycles and genomic sequencing cycles.

[0083] As further shown in FIG. 3, the sequence-to-coverage system 106 may perform the act 302 according to an order of indexing cycles between genomic sequencing cycles. The sequence-to-coverage system 106 may perform sequencing cycles according to an order of the non-indexing-first workflow 314. In the non-indexing-first workflow 314, the sequence-to-coverage system 106 determines base calls in the following order for paired-end reads: (i) a first nucleotide read corresponding to a first portion of the sample genomic sequence, (ii) a first indexing sequence appended to the sample genomic sequence, (iii) a second indexing sequence appended to the sample genomic sequence, and (iv) a second nucleotide read corresponding to a second portion of the sample genomic sequence. During the sequencing process, the sequence-to-coverage system 106

performs a pair-end turn between determining base calls for the first indexing sequence and the second indexing sequence. In the non-indexing-first workflow 314, the sequence-to-coverage system 106 does not complete calling the first and second indexing sequences until after determining base calls for at least one portion of the sample genomic sequence. Thus, the sequence-to-coverage system 106 does not obtain indexing sequence data until relatively further into the run index.

[0084] By contrast, in some embodiments, the sequence-to-coverage system 106 utilizes an indexing first workflow that enables the sequence-to-coverage system 106 to identify a genomic sample to which a nucleotide read corresponds before sequencing the read. The indexing-first workflow 316 illustrated in FIG. 3 portrays the sequence-to-coverage system 106 performing indexing cycles before genomic sequencing cycles. As shown, the sequence-to-coverage system 106 determines base calls in the following order for a paired-end read: (i) a first indexing sequence appended to the sample genomic sequence, (ii) a second indexing sequence appended to the sample genomic sequence, (iii) a first nucleotide read corresponding to a first portion of the sample genomic sequence, and (iv) a second nucleotide read corresponding to a second portion of the sample genomic sequence. In the indexing-first workflow 316, the sequence-to-coverage system 106 performs a pair-end turn between determining base calls for the first nucleotide read and the second nucleotide read. By utilizing the indexing-first workflow 316, the sequence-to-coverage system 106 can determine, relatively early within a sequencing run, which nucleotide reads originate from which genomic samples. FIG. 4 and the corresponding discussion further detail the sequence-to-coverage system 106 utilizing the indexing-first workflow 316 in accordance with one or more embodiments.

[0085] As shown in FIG. 3, the sequence-to-coverage system 106 performs the act 304 of determining respective numbers of clusters belonging to respective genomic samples. Generally, the sequence-to-coverage system 106 determines a balance between genomic samples within a pool of genomic samples. More particularly, based on the indexing sequences, the sequence-to-coverage system 106 determines respective numbers of clusters of oligonucleotides belonging to respective genomic samples of the genomic samples. In some embodiments, the sequence-to-coverage system 106 compares the index sequences of nucleotide reads in clusters of oligonucleotides 318 to a reference of known indexes to determine the genomic sample origin of each nucleotide read. The sequence-to-coverage system 106 may then sort the clusters of oligonucleotides 318 based on the originating samples. As further illustrated in FIG. 3, the sequence-to-coverage system 106 determines numbers of clusters belonging to each of the genomic samples in the genomic pool. FIG. 5 and the corresponding discussion further detail the sequence-

to-coverage system 106 determining respective numbers of clusters of oligonucleotides belonging to respective genomic samples in accordance with one or more embodiments.

[0086] As further illustrated in FIG. 3, the series of acts 300 optionally includes the act 306 of determining filter metrics. By determining filter metrics, the sequence-to-coverage system 106 estimates variation in sequencing data arising from pass filter issues. In some implementations, the sequence-to-coverage system 106 determines filter metrics indicating subsets of clusters of oligonucleotides satisfying a filtering threshold for signals of the clusters of oligonucleotides. The sequence-to-coverage system 106 evaluates clusters of oligonucleotides to identify filter-passing clusters of oligonucleotides. For instance, the sequence-to-coverage system 106 may evaluate empty wells or clusters that are dim, low quality, or polyclonal as filter-failing clusters of oligonucleotides. As shown in FIG. 3, the sequence-to-coverage system 106 determines that the cluster 320 does not satisfy a filtering threshold. The sequence-to-coverage system 106 may aggregate filter data for the clusters of oligonucleotides to estimate subsets of clusters of oligonucleotides originating from each genomic sample that satisfy a filtering threshold. FIG. 6 and the corresponding discussion further detail how the sequence-to-coverage system 106 determines filter metrics indicating subsets of clusters of oligonucleotides that satisfy a filtering threshold. In some embodiments, the act 306 is an optional act.

[0087] The series of acts 300 further comprises the act 308 of estimating read-coverage levels for the genomic samples. By determining respective numbers of clusters belonging to the respective genomic samples, the sequence-to-coverage system 106 can estimate in part variation arising from sample pooling. The sequence-to-coverage system 106 may more accurately estimate read-coverage levels for the genomic samples based on the respective numbers of clusters of oligonucleotides belonging to respective genomic samples and a currently selected number of sequencing cycles for the sequencing run. Additionally, in some embodiments, the sequence-to-coverage system 106 estimates the read-coverage levels based on the filter metrics. As shown, the sequence-to-coverage system 106 may generate an estimated read-coverage level for a genomic sample (C_{sample}) by multiplying the number of clusters belonging to the genomic sample and the currently selected number of sequencing cycles. As described, the currently selected number of sequencing cycles comprises a number of sequencing cycles to be performed during a sequencing run.

[0088] As further illustrated in FIG. 3, the sequence-to-coverage system 106 may determine the estimated read-coverage level for the genomic sample (C_{sample}) based on the filter metrics. In some implementations, the sequence-to-coverage system 106 access filter metrics for clusters corresponding to the particular genomic sample. In some examples, the sequence-to-coverage system 106 determines the estimated read-coverage level for the genomic sample by multiplying

the number of clusters belonging to the genomic sample by the filter metrics for the genomic sample, and the currently selected number of sequencing cycles.

[0089] Based on the estimated read-coverage levels for the genomic sample, the sequence-to-coverage system 106 modifies the sequencing process to meet a target read-coverage level. As illustrated in FIG. 3, the sequence-to-coverage system 106 performs the act 310 of generating a customized number of sequencing cycles. In particular, the sequence-to-coverage system 106 generates a customized number of sequencing cycles sufficient to generate nucleotide reads satisfying a target read-coverage level for each genomic sample of the genomic samples. Generally, the sequence-to-coverage system 106 can generate a customized number of sequencing cycles by increasing or decreasing a currently selected number of sequencing cycles. For example, the sequence-to-coverage system 106 may utilize the following equation to determine the customized number of sequencing cycles (N_{cyc})

$$N_{cyc} \times C_{min} = Output_{target}$$

where N_{cyc} represents the customized number of sequencing cycles, C_{min} represents the read-coverage level of the genomic sample with the least amount of coverage, and $Output_{target}$ represents the target read-coverage level. FIG. 7 and the corresponding discussion further detail the sequence-to-coverage system 106 generating the customized number of sequencing cycles and executing the sequencing run in accordance with one or more embodiments.

[0090] The series of acts 300 illustrated in FIG. 3 further comprises the act 312 of determining a customized set of flow cell regions to be imaged. In addition, or in the alternative, to generating a customized number of sequencing cycles, the sequence-to-coverage system 106 may also determine, from a flow cell, a customized set of flow cell regions to be imaged sufficient to generate nucleotide reads satisfying a target read-coverage level for each genomic sample of the genomic samples. In one example, the sequence-to-coverage system 106 utilizes the following equation to determine the customized set of flow cell regions to be imaged:

$$C_{min} \times N_{cyc} \times \frac{N_{S2C}}{N_T} = Output_{target}$$

where C_{min} represents the read-coverage level of the genomic sample with the least amount of coverage, N_{cyc} represents the customized number of sequencing cycles, N_{S2C} represents the customized set of flow cell regions to be imaged, N_T represents the total number of flow cell regions in the nucleotide-sample flow cell, and $Output_{target}$ represents the target read-coverage level. FIG. 8 and the corresponding paragraphs illustrate the sequence-to-coverage system 106 determining a customized set of flow cell regions to be imaged in accordance with one or more embodiments of the disclosure.

[0091] As mentioned, in some implementations, the sequence-to-coverage system 106 performs a subset of sequencing cycles according to an order of indexing cycles before genomic sequencing cycles. FIG. 4 illustrates the sequence-to-coverage system 106 performing a subset of sequencing cycles in an order of indexing cycles before genomic sequencing cycles in accordance with one or more embodiments of the present disclosure. FIG. 4 illustrates a series of acts 400 comprising an act 402 of determining base calls for a first indexing sequence, an act 404 of determining base calls for a second indexing sequence, an act 406 of determining base calls for a first nucleotide read, and an act 408 of determining base calls for a second nucleotide read.

[0092] In some implementations, the sequence-to-coverage system 106 utilizes an indexing-first workflow to determine a balance of genomic samples within a pool of genomic samples relatively early within a sequencing run. By performing indexing cycles before genomic sequencing cycles, the sequence-to-coverage system 106 determines which nucleotide reads belong to which genomic samples and a relative balance of genomic samples. In a non-indexing-first workflow, indexing sequence data from both indexing sequences appended to a sample genomic sequence is available only after the pair-end turn is complete. In contrast, the sequence-to-coverage system 106 can improve efficiency by obtaining indexing data before performing genomic sequencing cycles. Accordingly, in some implementations, the sequence-to-coverage system 106 may adjust genomic sequencing cycles in a dynamic manner based on indexing sequence information.

[0093] FIG. 4 illustrates the series of acts 400 comprising the act 402 of determining base calls for a first indexing sequence. A first index primer 412 is annealed to the primer binding site appended to the sample genomic sequence 410. After the first index primer 412 is annealed, the sequence-to-coverage system 106 determines base calls for the first indexing sequence 416. As shown in FIG. 4, the first indexing sequence 416 is appended to a sample genomic sequence 410 of a genomic sample.

[0094] As further illustrated in FIG. 4, the sequence-to-coverage system 106 performs the act 404 of determining base calls for a second indexing sequence. The sequence-to-coverage system 106 anneals a second index primer 418 to the primer binding site appended to the sample genomic sequence 410. The sequence-to-coverage system 106 determines base calls for the second indexing sequence 420. As further shown in FIG. 4, the second indexing sequence 420 is appended to the 5' end of the sample genomic sequence 410 while the first indexing sequence 416 is appended to the 7' end of the sample genomic sequence 410.

[0095] After determining base calls for the first indexing sequence 416 and the second indexing sequence 420, the sequence-to-coverage system 106 performs the act 406 of determining base calls for a first nucleotide read. More specifically, the sequence-to-coverage system 106

determines base calls for a first nucleotide read corresponding to a first portion of the sample genomic sequence 410. More specifically, in a paired-end sequencing run, the sample genomic sequence 410 is sequenced from both ends, providing complementary information about the sample genomic sequence 410. As part of performing the act 406, the sequence-to-coverage system 106 anneals a first nucleotide read primer 422 to a read primer binding site, and the sequence-to-coverage system 106 sequences the first portion of the sample genomic sequence 410.

[0096] In some embodiments, after performing the act 406, the sequence-to-coverage system 106 performs a pair-end turn. Generally, during the pair-end turn, the P7 region is cleaved and all fragments are attached by the P5 region. Prior to the pair-end turn, the P7 region is annealed to the surface of the flow cell. After the pair-end turn, the P5 region is attached to the flow cell.

[0097] Following the pair-end turn, the sequence-to-coverage system 106 performs the act 408 of determining base calls for a second nucleotide read. The sequence-to-coverage system 106 anneals the second nucleotide read primer 424 to a second read primer binding site, and the sequence-to-coverage system 106 sequences the second portion of the sample genomic sequence 410. In some embodiments, the sequence-to-coverage system 106 utilizes specialized reagents as part of the indexing-first workflow.

[0098] As mentioned, the sequence-to-coverage system 106 can determine respective numbers of clusters of oligonucleotides belonging to respective genomic samples based on the indexing sequences. FIG. 5 illustrates the sequence-to-coverage system 106 determining respective numbers of clusters of oligonucleotides belonging to respective genomic samples of the genomic samples in accordance with one or more embodiments of the present disclosure.

[0099] After determining base calls for the indexing sequences, the sequence-to-coverage system 106 determines which clusters of oligonucleotides correspond to each genomic sample in the pool of genomic samples. The sequence-to-coverage system 106 may accomplish this through a process called demultiplexing. After determining base calls for the indexing sequences, the sequence-to-coverage system 106 analyzes the raw sequencing data and uses index barcodes to assign each read to its corresponding genomic sample.

[0100] As illustrated in FIG. 5, the sequence-to-coverage system 106 accesses raw sequencing data comprising indexing sequences 504 associated with a sample genomic sequence 518, indexing sequences 506 associated with a sample genomic sequence 520, and indexing sequences 508 associated with a sample genomic sequence 522. The indexing sequences 504-508 comprise barcodes that act as unique identifiers for each genomic sample, allowing for differentiation and sorting of the reads during demultiplexing. For example, the indexing sequences 504 indicate that the sample genomic sequence 518 comes from genomic sample 1. The indexing sequences 506 indicate that the sample genomic sequence 520 originates from genomic sample 2.

[0101] In one or more implementations, the sequence-to-coverage system 106 demultiplexes nucleotide reads by utilizing a reference of known indexes. FIG. 5 illustrates a reference of registered indexes 514. The sequence-to-coverage system 106 compares indexing sequences with known indexing sequences in the reference of registered indexes 514. The reference of registered indexes 514 associates each index barcode or sequence with its respective genomic sample. For example, and as illustrated, the reference of registered indexes 514 stores indexing sequences with their corresponding genomic samples. As shown, genomic samples may correspond with one or more unique barcodes.

[0102] In some implementations, the sequence-to-coverage system 106 can identify and differentiate between assigned indexing sequences and unassigned indexing sequences. Assigned indexing sequences match indexing sequences registered for the particular run. Unassigned indexing sequences (e.g., the indexing sequences 508) do not match indexing sequences registered for the sequencing run. The sequence-to-coverage system 106 may identify unassigned indexing sequences based on determining that a given indexing sequence is absent from the reference of registered indexes 514. To illustrate, the sequence-to-coverage system 106 may compare the indexing sequences 508 with the registered indexes in the reference of registered indexes 514 and determine that the indexing sequences 508 are not in the reference of registered indexes 514. In one or more embodiments, the sequence-to-coverage system 106 identifies unassigned indexing sequences and removes, from data for the sequencing run, a subset of clusters of oligonucleotides corresponding to the unassigned indexing sequences.

[0103] As mentioned, the sequence-to-coverage system 106 identifies respective numbers of clusters of oligonucleotides belonging to respective genomic samples of the genomic samples. FIG. 5 illustrates a flow cell 502 comprising a flow cell. The flow cell 502 comprises a lane 510, which contains a flow cell region 512. The flow cell region 512 can represent a tile of the flow cell. As shown in FIG. 5, the flow cell region 512 comprises several clusters of oligonucleotides. Each cluster contains multiple copies of the same sample genomic sequence. The sequence-to-coverage system 106 identifies clusters corresponding to the genomic sample 1, and the genomic sample 2. The sequence-to-coverage system 106 further identifies clusters having the unassigned indexing sequence. As illustrated, the sequence-to-coverage system 106 determines that the clusters 524 correspond with unassigned indexing sequences that do not match the indexing sequences registered for the sequencing run.

[0104] Upon assigning clusters to genomic samples within the pool of genomic samples, the sequence-to-coverage system 106 determines a number of clusters of oligonucleotides belonging to each genomic sample. More particularly, the sequence-to-coverage system 106 counts a number of clusters belonging to each genomic sample corresponding to an assigned indexing sequence. As

illustrated in table 516 in FIG. 5, the sequence-to-coverage system 106 determines that genomic sample 1 corresponds with 275M clusters, and genomic sample 2 corresponds with 373M clusters.

[0105] Additionally, in some embodiments, the sequence-to-coverage system 106 generates and stores a genomic sample map indicating the locations of clusters corresponding with each genomic sample. As illustrated in FIG. 5, the sequence-to-coverage system 106 generates a genomic sample map 526 indicating locations of clusters corresponding to each of the genomic samples. Furthermore, as shown, the sequence-to-coverage system 106 excludes, from the genomic sample map 526 data corresponding with unassigned indexing sequences. For example, the sequence-to-coverage system 106 removes the clusters 524 from the genomic sample map 526.

[0106] As described, the sequence-to-coverage system 106 may estimate the read-coverage levels for genomic samples based on filter metrics. FIG. 6 illustrates the sequence-to-coverage system 106 determining filter metrics in accordance with one or more implementations of the present disclosure. The sequence-to-coverage system 106 determines filter metrics indicating subsets of clusters of oligonucleotides satisfying a filtering threshold for signals of the clusters of oligonucleotides. Generally, filter metrics indicate a quality and reliability of sequencing reads generated during a sequencing run.

[0107] As shown in FIG. 6, the sequence-to-coverage system 106 determines base-call-quality metrics 602. More specifically, the sequence-to-coverage system 106 determines the base-call-quality metrics 602 for a subset of sequencing cycles. To illustrate, during each sequencing cycle, the sequence-to-coverage system 106 images clusters within a flow cell region 612 (e.g., a tile of a flow cell). The sequence-to-coverage system 106 evaluates the signals emitted from the clusters of oligonucleotides to determine the base-call-quality metrics 602.

[0108] In some embodiments, the base-call-quality metrics 602 comprise a chastity value. The term “chastity value” refers to a quality metric used to assess the confidence or purity of a called nucleobase from a sequencing cycle. In particular, the chastity value is a measure of the confidence of the called base at each position within a sequencing read. For example, the chastity value may be calculated based on the intensity of the fluorescent signals emitted from the clusters of oligonucleotides. The sequence-to-coverage system 106 measures the intensity of each of the four nucleotide-specific fluorescent signals. The sequence-to-coverage system 106 may determine the chastity value by determining a ratio of the brightest base intensity divided by the sum of the brightest and second brightest base intensities. In some examples, the sequence-to-coverage system 106 can report the chastity value as a percent value ranging from 0%-100%.

[0109] As illustrated in FIG. 6, the sequence-to-coverage system 106 utilizes the base-call-quality metrics 602 and a filter threshold to determine filter-passing clusters of oligonucleotides. In particular, the sequence-to-coverage system 106 compares a quality metric for a cluster with a

filter threshold to determine whether the cluster is a filter-passing cluster. For example, the sequence-to-coverage system 106 compares quality metrics for each of the clusters within the flow cell region 612 with a filter threshold. In some examples, the filter threshold comprises a chastity threshold value (e.g., 80%). The sequence-to-coverage system 106 determines that clusters having chastity values meeting the chastity threshold value qualify as filter-passing clusters. As shown, the sequence-to-coverage system 106 determines that the clusters 614a, 614b, and 614c all have quality metrics not satisfying a filter threshold. More specifically, the chastity values for the clusters 614a-614c do not meet the chastity threshold value. Accordingly, the sequence-to-coverage system 106 determines that the clusters 614a-614c are not filter-passing clusters. The sequence-to-coverage system 106 determines that the clusters 616a-616c comprise filter-passing clusters.

[0110] As mentioned, the sequence-to-coverage system 106 determines the base-call-quality metrics 602 for a subset of sequencing cycles. To improve efficiency, the sequence-to-coverage system 106 utilizes images from early sequencing cycles to evaluate the reliability and accuracy of base calling within each cluster of oligonucleotides. As shown in FIG. 6, the sequence-to-coverage system 106 determines base-call-quality metrics 602 for the flow cell region 612 within a subset of sequencing cycles. For example, the subset of sequencing cycles may comprise the first 25 sequencing cycles of a sequencing run. The sequence-to-coverage system 106 determines the base-call-quality metrics 602 for each sequencing cycle within the subset of sequencing cycles. Furthermore, the sequence-to-coverage system 106 determines filter-passing clusters within each sequencing cycle. For example, while the sequence-to-coverage system 106 determines that the cluster 616b is a filter-passing cluster in a first sequencing cycle, the sequence-to-coverage system 106 may determine that the cluster 616b is not a filter-passing cluster in a second sequencing cycle.

[0111] As shown in FIG. 6, the sequence-to-coverage system 106 determines base-call-quality metrics 602 for clusters originating from each genomic sample by utilizing a genomic sample map 608. The genomic sample map 608 indicates locations of clusters corresponding to each of the genomic samples. For example, the genomic sample map 608 for the flow cell region 612 indicates that the clusters 614a-614b originate from genomic sample 1, and the cluster 616a and the cluster 616c originate from genomic sample 2. As shown, the genomic sample map 608 also indicates that the cluster 616b and the cluster 614c arise from unregistered genomic samples. The sequence-to-coverage system 106 may generate the genomic sample map 608 utilizing processes described above with respect to FIG. 5.

[0112] By utilizing information from the genomic sample map 608, the sequence-to-coverage system 106 can identify a number of filter-passing clusters of oligonucleotides for each genomic sample that satisfy the filtering threshold. More specifically, the sequence-to-coverage system 106

utilizes the genomic sample map 608 to determine the base-call-quality metrics for clusters of oligonucleotides for each genomic sample. By comparing the base-call-quality metrics with a filtering threshold, the sequence-to-coverage system 106 may count a number of clusters of oligonucleotides for each genomic sample that qualify as filter-passing clusters of oligonucleotides.

[0113] As illustrated in FIG. 6, the sequence-to-coverage system 106 generates a pass filter map 604. The sequence-to-coverage system 106 aggregates the base-call-quality metrics 602 across the subset of sequencing cycles to generate the pass filter map 604. Generally, the pass filter map 604 provides information about the outcome of quality filtering applied to the clusters of oligonucleotides for the subset of sequencing cycles. The pass filter map 604 indicates a percentage of clusters at a location that satisfy a filtering threshold over the subset of sequencing cycles. For example, the sequence-to-coverage system 106 determines a percent of filter-passing clusters for each cluster in the flow cell region 612. For example, the sequence-to-coverage system 106 that across the subset of sequencing cycles, 20% of the cluster 614a comprise filter-passing clusters. The sequence-to-coverage system 106 performs this determination for the remaining clusters within the flow cell region 612. In some implementations, the sequence-to-coverage system 106 further indicates, within the pass filter map 604 the genomic sample corresponding with each cluster.

[0114] The sequence-to-coverage system 106 further aggregates information for each genomic sample to generate the filter metrics 610. The filter metrics 610 indicate a subset of clusters of oligonucleotides that satisfy a filtering threshold for signals of the clusters of oligonucleotides. As shown in FIG. 6, the filter metrics comprise a percent of clusters for a genomic sample that satisfy a filtering threshold. For example, and as illustrated in FIG. 6, the sequence-to-coverage system 106 determines that 83% of clusters corresponding with genomic sample 1 satisfy the filtering threshold. In some implementations, the sequence-to-coverage system 106 determines the filter metrics by combining the percent of filter-passing clusters for clusters corresponding to the genomic sample. For example, the sequence-to-coverage system 106 may average the percent of filter-passing clusters corresponding to the genomic sample.

[0115] Furthermore, in some embodiments, the sequence-to-coverage system 106 may utilize the filter metrics 610 and the respective numbers of clusters of oligonucleotides belonging to respective genomic samples to determine a number of filter-passing clusters of oligonucleotides for each genomic sample. For example, the sequence-to-coverage system 106 may determine a number of clusters corresponding to a given genomic sample utilizing the processes described with respect to FIG. 5. The sequence-to-coverage system 106 multiplies the number of clusters for the given genomic sample by the percent of clusters for the given genomic samples that satisfy the filtering threshold. As illustrated in FIG. 6, the sequence-to-coverage system 106 determines that

275M clusters correspond with genomic sample 1. Based on determining, utilizing the pass filter map 606, that 83% of clusters for genomic sample 1 pass filter, the sequence-to-coverage system 106 determines that a number of filter-passing clusters for genomic sample 1 equals $.83 \times 275\text{M}$ or 228M.

[0116] As mentioned, the sequence-to-coverage system 106 may generate a customized number of sequencing cycles sufficient to generate nucleotide reads satisfying a target read-coverage level for each genomic sample of the genomic samples. FIGS. 7A-7B illustrate the sequence-to-coverage system 106 generating a customized number of sequencing cycles to meet a target read-coverage level and executing the sequencing run in accordance with one or more embodiments of the present disclosure. By estimating read-coverage levels for the genomic samples, the sequence-to-coverage system 106 can adjust the number of sequencing cycles to ensure that all genomic samples receive at least a target read-coverage level. FIGS. 7A-7B illustrate a series of acts 700 comprising an act 702 of starting a sequencing run, an act 704 of determining base calls for indexing sequences, an act 706 of determining filter metrics, an act 710 of generating a customized number of sequencing cycles, and an act 712 of executing the sequencing run until finishing the customized number of sequencing cycles.

[0117] The series of acts 700 illustrated in FIG. 7A includes the act 702 of starting the sequencing run. In some implementations, as part of starting the sequencing run, the sequence-to-coverage system 106 determines a target read-coverage level. For example, the sequence-to-coverage system 106 may provide, for display via a client device (e.g., the client device 114), a target-read-coverage level selection element. The sequence-to-coverage system 106 may receive user input indicating the target-read-coverage level. In some implementations, the sequence-to-coverage system 106 automatically determines the target-read-coverage level. In some implementations, the sequence-to-coverage system 106 determines a target-read-coverage level of 40x.

[0118] As further shown in FIG. 7A, the sequence-to-coverage system 106 performs the act 704 of determining base calls for indexing sequences. As described previously, by determining base calls for indexing sequences, the sequence-to-coverage system 106 determines sample-to-sample variability relatively early within the sequencing run. The sequence-to-coverage system 106 may utilize a non-indexing first workflow and an indexing-first workflow early on within a sequencing cycle. More specifically, the sequence-to-coverage system 106 may boost efficiency of sequencing runs by utilizing an indexing-first workflow. As mentioned, the sequence-to-coverage system 106 may determine base calls for indexing sequences for a subset of sequencing cycles. For example, the sequence-to-coverage system 106 may determine base calls for indexing sequences for the first 5, 10, 25, etc. sequencing cycles of the sequencing run.

[0119] FIG. 7A also illustrates the sequence-to-coverage system 106 performing the act 706 of determining filter metrics. As described previously with respect to FIG. 6, the sequence-to-coverage system 106 determines filter metrics that indicate subsets of clusters of oligonucleotides that satisfy a filtering threshold for signals of the clusters of oligonucleotides. As with determining base calls for indexing sequences, the sequence-to-coverage system 106 determines filter metrics for a subset of sequencing cycles. In some implementations, the sequence-to-coverage system 106 determines the filter metrics for a second subset of sequencing cycles that differs from a first set of subset of sequencing cycles used to perform indexing cycles before genomic sequencing cycles. For instance, the sequence-to-coverage system 106 may determine filter metrics for clusters of oligonucleotides in the first 10, 15, 20, 25, etc. sequencing cycles of the sequencing run. In some embodiments, the act 706 comprises an optional act.

[0120] In some implementations, the sequence-to-coverage system 106 also determines PhiX loss early in the sequencing run. PhiX refers to a standard control library used in sequencing runs to monitor the sequencing process and assess the performance of a sequencing platform. In some implementations, the PhiX control library is spiked into the sequencing run as a control sample. The amount of PhiX can be a small percent (e.g., 1-2%) of the input samples. During PhiX alignment, the sequence-to-coverage system 106 maps nucleotide reads to the PhiX genome to determine an amount of PhiX loss. More specifically, PhiX loss occurs when the proportion of nucleotide reads derived from the PhiX control library is significantly reduced compared to the expected or intended amount. Greater PhiX loss can indicate issues in various parameters such as cluster density, signal intensity, and base-calling accuracy. The sequence-to-coverage system 106 can determine PhiX loss early in the sequencing run by utilizing indexing and filter metrics data. Examples of determining PhiX loss are also described in U.S. Pat. No. 9,574,226 B2, the disclosure of which is incorporated herein by reference in its entirety.

[0121] As illustrated in FIG. 7A, the sequence-to-coverage system 106 performs the act 708 of estimating read-coverage levels based on determining the base calls for indexing sequences. For example, the sequence-to-coverage system 106 utilizes the following equation to generate an estimated-read-coverage level for a given sample:

$$C_{sample} = \# \text{ of clusters} \times \text{Currently selected \# of sequencing cycles}$$

Where C_{sample} represents an estimated-read-coverage level for a given sample, “# of clusters” represents a number of clusters originating from the given sample, and “currently selected # of sequencing cycles” refers to the anticipated number of sequencing cycles within the sequencing run.

[0122] In some implementations, as part of performing the act 708, the sequence-to-coverage system 106 further utilizes filter metrics determined as part of the act 706. The sequence-to-

coverage system 106 can utilize the following equation to generate an estimated-read-coverage level for a given sample

$$C_{sample} = \# \text{ of clusters} \times \text{filter metrics} \times \text{Currently selected \# of sequencing cycles}$$

Where C_{sample} represents an estimated-read-coverage level for a given sample, “# of clusters” represents a number of clusters originating from the given sample, “filter metrics” refers to a proportion or percentage of clusters arising from the given sample that satisfy a filtering threshold, and “currently selected # of sequencing cycles” refers to the anticipated number of sequencing cycles within the sequencing run. By utilizing indexing data and filter metrics data, the sequence-to-coverage system 106 can capture about 80% of yield variation.

[0123] Based on the estimated read-coverage levels, the sequence-to-coverage system 106 performs the act 710 of generating a customized number of sequencing cycles. The sequence-to-coverage system 106 can adjust a total number of sequencing cycles within a sequencing run. For example, the sequence-to-coverage system 106 can increase the number of sequencing cycles relative to the currently selected number of sequencing cycles if one genomic sample has poor coverage. Alternatively, the sequence-to-coverage system 106 can lower the total number of sequencing cycles relative to the currently selected number of sequencing cycles if the sequencing run is likely to produce excess data.

[0124] In one or more embodiments, the sequence-to-coverage system 106 generates the customized number of sequencing cycles utilizing the following equation:

$$N_{cyc} \times C_{min} = Output_{target}$$

Where N_{cyc} represents the customized number of sequencing cycles, C_{min} represents the read-coverage level of the genomic sample with the lowest estimated read-coverage level, and $Output_{target}$ represents the target read-coverage level. In some implementations, the sequence-to-coverage system 106 generates the customized number of sequencing cycles for the sequencing run by increasing or decreasing a preset number of sequencing cycles for the sequencing run.

[0125] As shown, the sequence-to-coverage system 106 determines the customized number of sequencing cycles utilizing data determined during primary analysis. More specifically, the sequence-to-coverage system 106 determines the customized number of sequencing cycles before completing the sequencing run. In some implementations, the sequence-to-coverage system 106 may determine the customized number of sequencing cycles at a sequencing device (e.g., the sequencing device 108) or a local server device (e.g., the local server device 102). More specifically, the sequence-to-coverage system 106 can determine the customized number of sequencing cycles during primary and not secondary analysis, which often occurs at a server device (e.g., server device(s) 110). By utilizing data obtained during early stages of a sequencing run, the

sequence-to-coverage system 106 can efficiently determine the customized number of sequencing cycles.

[0126] FIG. 7B illustrates the sequence-to-coverage system 106 performing the act 712 of executing the sequencing run until finishing the customized number of sequencing cycles. More specifically, the sequence-to-coverage system 106 causes a sequencing device to execute the customized number of sequencing cycles. For example, the sequence-to-coverage system 106 can cause a fluidic device to perform additional sequencing cycles or fewer sequencing cycles based on the customized number of sequencing cycles.

[0127] FIG. 7B illustrates a chart 718 depicting over-sequenced results generated by existing systems and a chart 720 depicting results generated by the sequence-to-coverage system 106 using a customized number of sequencing cycles. The x-axes of the chart 718 and the chart 720 represents a number of sequencing cycles. The y-axes of the chart 718 and the chart 720 represent a percent of genomic samples reaching a target read-coverage level (40x).

[0128] As shown in the chart 718, about 95% of genomic samples have been sequenced to a target read-coverage level at 2x150 sequencing cycles. As shown, the majority of genomic samples are over sequenced at 2x150 sequencing cycles. Furthermore, and as previously mentioned, about 5% of genomic samples remain under-sequenced and have not yet met the target read-coverage level at 2x150 sequencing cycles.

[0129] As illustrated in FIG. 7B, the sequence-to-coverage system 106 can adjust parameters of the sequencing run to improve efficiency. More specifically, the sequence-to-coverage system 106 does not only consider average read-coverage level, the sequence-to-coverage system 106 also ensures that hard-to-map genomic regions (e.g., repeat regions) are not negatively affected by reducing the number of sequencing cycles. Accordingly, in some implementations, the sequence-to-coverage system 106 evaluates a minimum number of sequencing cycles and a maximum number of sequencing cycles are before relevant metrics for hard-to-map regions begin to decline. Furthermore, the sequence-to-coverage system 106 may design flow cell (FC) capacity and the number of genomic samples within a pool such that a maximum success rate is enabled with a minimal number of default sequencing cycles.

[0130] As shown in FIG. 7B, the sequence-to-coverage system 106 decreases the size of the flow cell or increases the number of genomic samples in the pool of genomic samples. The sequence-to-coverage system 106 may decrease the size of the flow cell by reducing a number of clusters per nucleotide-sample substrate (e.g., flow cell). For example, the sequence-to-coverage system 106 may reduce a number of nanowells per flow cell. In some implementations, the sequence-to-coverage system 106 decreases the size of the flow cell by determining a reduced set of flow cell regions to be imaged. As a result, about 50% of genomic samples are sequenced to the

target read-coverage level at 2x150 sequencing cycles. The sequence-to-coverage system 106 may determine a customized number of sequencing cycles that falls between the minimum number of sequencing cycles (2x135c) and the maximum number of sequencing cycles (2x185c). In some implementations, the sequence-to-coverage system 106 increases or decreases a preset number of sequencing cycles (e.g., the default number of sequencing cycles) by a preset number of sequencing cycles within the minimum number of sequencing cycles and the maximum number of sequencing cycles. For instance, the sequence-to-coverage system 106 can decrease or increase the preset number of sequencing cycles by 15, 35, etc.

[0131] As further illustrated in FIG. 7B, the sequence-to-coverage system 106 can determine a minimum number of sequencing cycles and a maximum number of sequencing cycles. More particularly, the sequence-to-coverage system 106 determines the minimum number of sequencing cycles and a maximum number of sequencing cycles based on a flow cell size and/or number of multiplexed genomic samples. In some cases, the sequence-to-coverage system 106 automatically determines the minimum number of sequencing cycles and the maximum number of sequencing cycles. For example, the sequence-to-coverage system 106 can determine that the minimum number of sequencing cycles and the maximum number of sequencing cycles are a relatively symmetrical number of sequencing cycles below and above a default number of sequencing cycles, respectively. In some embodiments, the sequence-to-coverage system 106 determines a minimum number of sequencing cycles 15, 35, etc. cycles below a preset number of sequencing cycles. The sequence-to-coverage system 106 may also determine a maximum number of sequencing cycles 15, 35, etc. cycles above a preset number of sequencing cycles. In some examples, the sequence-to-coverage system 106 determines a minimum number of sequencing cycles and a maximum number of sequencing cycles within a preset range of sequencing cycles. For example, the sequence-to-coverage system 106 can determine a maximum number of sequencing cycles and a minimum number of sequencing cycles that are within 50 sequencing cycles of each other. In some examples, the sequence-to-coverage system 106 determines the minimum and maximum numbers of sequencing cycles based on user input.

[0132] Generally, the sequence-to-coverage system 106 can determine the minimum number of sequencing cycles to ensure a baseline coverage of all the genomic samples. The sequence-to-coverage system 106 can determine the minimum number of sequencing cycles based on the workflow or purpose of the sequencing run. The sequence-to-coverage system 106 can determine different minimum numbers of sequencing cycles for different assays. For example, some assays such as enrichment assays, require lower read-coverage levels. Other sequencing assays for sequencing hard-to-map genomic regions may require higher read-coverage levels.

[0133] By generating a customized number of sequencing cycles, the sequence-to-coverage system 106 improves the efficiency of sequencing runs. The sequence-to-coverage system 106 reduces the number of sequencing cycles required to meet a target read-coverage level relative to existing systems. For instance, existing systems require an average of 316 sequencing cycles. In contrast, the sequence-to-coverage system 106 can achieve 120Gb coverage in 226 sequencing cycles, which is 28% less than sequencing cycles by existing systems. The reduction in sequencing cycles also reduces the amount of sequencing reagents required for a sequencing run. More specifically, the sequence-to-coverage system 106 can execute sequencing runs requiring 28% less reagents than existing systems. The sequence-to-coverage system 106 further also executes sequencing runs requiring 11.5% less total materials than existing systems. Total materials may comprise, in addition to sequencing reagents, library preparation kits, flow cells, cluster amplification materials, and other processing materials. Additionally, the total runtime of a sequencing run executed by the sequence-to-coverage system 106 is, on average, 90 minutes shorter than existing sequencing runs on state-of-the-art sequencing devices. The runtime savings, however, depends on a sequencing device's time-per-cycle and, therefore, the runtime savings may be greater for sequencing devices with longer time-per-cycle metrics and lesser for sequencing devices with shorter time-per-cycle metrics. Furthermore, the sequence-to-coverage system 106 improves the genomic sample success rate from 96% to 99% by executing a sequencing run having the customized number of sequencing cycles.

[0134] In addition, or in the alternative, to generating a customized number of sequencing cycles, the sequence-to-coverage system 106 can determine a customized set of flow cell regions to be imaged sufficient to generate nucleotide reads satisfying a target-read coverage level for each genomic sample. FIG. 8 illustrates the sequence-to-coverage system 106 determining a customized set of flow cell regions to be imaged in accordance with one or more embodiments of the present disclosure. FIG. 8 illustrates a series of acts 800 comprising an act 802 of starting a sequencing run, an act 804 of determining base calls for indexing sequences, an act 806 of determining filter metrics, an act 808 of estimating read-coverage levels, an act 810 of determining a customized set of flow cell regions to be imaged, and an act 812 of executing the sequencing run by capturing images of the customized set of flow cell regions.

[0135] The sequence-to-coverage system 106 may determine to utilize one or both of generating a customized number of sequencing cycles and determining a customized set of flow cell regions to be imaged. In some applications, the sequence-to-coverage system 106 may determine to keep the number of sequencing cycles constant (e.g., 2x150c) and instead improve efficiency by adjusting the set of flow cell regions to be imaged during a sequencing run. In other applications, the sequence-to-coverage system 106 executes sequencing runs having a customized

number of sequencing cycles without adjusting the flow cell regions to be imaged during those sequencing cycles. In some applications, the sequence-to-coverage system 106 determines to utilize both tools to improve efficiency of a sequencing run. More specifically, the sequence-to-coverage system 106 may both adjust the number of sequencing cycles and the number of flow cell regions imaged within the same sequencing run.

[0136] The sequence-to-coverage system 106 may determine the customized set of flow cell regions to be imaged based on the type of imaging processes utilized by various sequencing devices. For example, the sequence-to-coverage system 106 can generate a customized number of sequencing cycles for sequencing devices with fast imaging processes. Some sequencing devices utilizes fast scanning processes, and efficiency is best improved by lowering the number of sequencing cycles. Some sequencing devices utilize slower imaging processes. For instance, sequencing devices that utilize stop-and-shoot imaging systems require more time in imaging steps than do sequencing devices that rely on scanning. Accordingly, the sequence-to-coverage system 106 may improve turnaround time by adjusting the set of flow cell regions that need to be imaged.

[0137] The acts 802-808 are like the acts 702-708 described above in reference to FIG. 7A. As with the acts 702-708 illustrated in FIG. 7, the sequence-to-coverage system 106 estimates read-coverage levels for each genomic sample within a pool of genomic samples. The following paragraphs detail variations between the acts 802-808 and the acts 702-708.

[0138] As shown in FIG. 8, the sequence-to-coverage system 106 performs the act 804 of determining base calls for indexing sequences. In some implementations, the sequence-to-coverage system 106 determines respective numbers of clusters belonging to respective genomic samples for each flow cell region. More specifically, the sequence-to-coverage system 106 determines, for each flow cell region, respective clusters of oligonucleotides belonging to respective genomic samples. The sequence-to-coverage system 106 stores a balance of genomic samples within each flow cell region. For example, the sequence-to-coverage system 106 can store flow cell region data in a genomic sample map. The sequence-to-coverage system 106 may utilize indexing data to identify flow cell regions that, when imaged, may compensate for imbalances in genomic sample representation.

[0139] The sequence-to-coverage system 106 further performs the act 806 of determining filter metrics. Generally, the sequence-to-coverage system 106 stores filter metric data for each flow cell region. For example, in some implementations, the sequence-to-coverage system 106 stores a percent passing filter metric for each flow cell region. The sequence-to-coverage system 106 may utilize the stored filter metric data for each flow cell region to apply different weights to different flow cell regions. For example, the sequence-to-coverage system 106 may determine to image flow cell regions corresponding with higher %PF than flow cell regions with lower %PF.

[0140] FIG. 8 illustrates the sequence-to-coverage system 106 performing the act 810 of determining a customized set of flow cell regions to be imaged. In some examples, the sequence-to-coverage system 106 utilizes the following equation to determine the customized set of flow cell regions to be imaged during sequencing cycles:

$$C_{min} \times N_{cyc} \times \frac{N_{s2c}}{N_T} = Output_{target}$$

Where C_{min} represents the read-coverage level of the genomic sample with the lowest expected-read-coverage level, N_{cyc} represents the customized number of sequencing cycles, N_{s2c} represents the number of flow cell regions in the customized set of flow cell regions to be imaged, N_T represents the total number of flow cell regions in the flow cell, and $Output_{target}$ represents the target read-coverage level. In implementations where the sequence-to-coverage system 106 executes a constant number of sequencing cycles, N_{cyc} represents the constant number of sequencing cycles (e.g., 2x150c).

[0141] In addition to determining a raw number of flow cell regions to be imaged, the sequence-to-coverage system 106 can identify specific flow cell regions within the flow cell to image during sequencing cycles. The sequence-to-coverage system 106 can leverage region-to-region variation to improve read-coverage levels for specific genomic samples and/or select the best-performing flow cell regions. For example, the sequence-to-coverage system 106 can image flow cell regions with more clusters belonging to a given genomic sample to improve the read-coverage level for the given genomic sample. Additionally, or alternatively, the sequence-to-coverage system 106 can image flow cell regions with higher filter metrics and/or stop imaging flow cell regions with lower filter metrics.

[0142] The series of acts 800 includes the act 812 of executing the sequencing run by capturing images of the customized set of flow cell regions. FIG. 8 illustrates a flow cell 816 comprising lanes made up of flow cell regions 818. In some implementations, a flow cell region comprises a tile of a flow cell. As shown, the sequence-to-coverage system 106 captures images of a customized set of flow cell regions 814 during sequencing cycles of the sequencing run. In other embodiments, the customized set of flow cell regions comprises a number of flow cell regions to be imaged within a lane 820. Some flow cells comprise addressable lanes where specific genomic samples are assigned to specific lanes of the flow cell. The sequence-to-coverage system 106 may generally determine to image a customized number of flow cell regions within the lane 820 to improve read-coverage levels of the genomic sample corresponding with the lane 820.

[0143] Imaging a customized set of flow cell regions yields several improvements relative to existing systems. By imaging a customized set of flow cell regions, the sequence-to-coverage system 106 can reduce the number of flow cell regions sequenced from 324 to 233—a 28%

reduction. The sequence-to-coverage system 106 further reduces the time required to complete a sequencing run relative to existing systems. For example, the sequence-to-coverage system 106 reduces runtime from 19 hours to 16 hours. The sequence-to-coverage system 106 improves efficiency at zero to very small compute costs.

[0144] The sequence-to-coverage system 106 improves efficiency of sequencing runs by executing a sequencing run until finishing a customized number of sequencing cycles. FIGS. 9A-9B illustrate improvements in sequencing efficiency resulting from execution of a customized number of sequencing cycles in accordance with one or more embodiments of the present disclosure. In particular, FIG. 9A illustrates improvements in efficiency given poor sample pooling, and FIG. 9B illustrates improvements in efficiency given optimal sequence performance. The charts in FIGS. 9A-9B portray simulated data.

[0145] FIG. 9A illustrates a chart 902 portraying read-coverage levels for genomic samples sequenced by existing systems and a chart 904 portraying read-coverage levels for genomic samples sequenced by the sequence-to-coverage system 106 under poor sample pooling conditions. The chart 902 portrays read-coverage levels for genomic samples 906a, 906b, and 906c. The chart 904 portrays read-coverage levels for genomic samples 908a, 908b, and 908c.

[0146] As shown in FIG. 9A, the genomic sample 906a fails to meet a target read-coverage level of 40x at 2x150 sequencing cycles. Under-sequencing the genomic sample 906a may require existing systems to perform an additional sequencing run to obtain sufficient data for the genomic sample 906a. By contrast, because the sequence-to-coverage system 106 determines and executes the customized number of sequencing cycles, the sequence-to-coverage system 106, the sequence-to-coverage system 106 does not under-sequence any genomic samples. For example, the sequence-to-coverage system 106 executes a sequencing run until finishing a customized number of 2x160 sequencing cycles. By increasing the number of sequencing cycles, the sequence-to-coverage system 106 ensures that the samples 908a-908c are all sequenced to a target read-coverage level.

[0147] FIG. 9B illustrates a chart 910 portraying read-coverage levels for genomic samples sequenced by existing systems and a chart 912 portraying read-coverage levels for genomic samples sequenced by the sequence-to-coverage system 106 with optimal sequence performance. For example, variation is reduced because the genomic samples may be more balanced, and 80% or more of the clusters pass filter. The chart 910 portrays read-coverage levels for genomic samples 914a, 914b, and 914c. The chart 912 portrays read-coverage levels for genomic samples 916a, 916b, and 916c. As illustrated, the genomic samples 914a-914c and the genomic samples 916a-916c demonstrate minimal variation in read-coverage level.

[0148] As shown in FIG. 9B, the existing system over-sequences the genomic samples 914a-914c. For example, in comparison to the 40x target read-coverage level, the existing system sequences the genomic samples 914a-914c to about a 72x read-coverage level when executing 2x150 sequencing cycles. In contrast, and as shown in the chart 912, the sequence-to-coverage system 106 determines a customized number of 2x120 sequencing cycles, which is fewer cycles than the default 2x150 sequencing cycles. By reducing the number of sequencing cycles, the sequence-to-coverage system 106 sequences the genomic samples 916a-916c to just meet and barely exceed the 40x target read-coverage level.

[0149] The sequence-to-coverage system 106 also improves efficiency of sequencing runs by imaging a customized set of flow cell regions during sequencing cycles. FIG. 10 illustrates improvements in sequencing efficiency resulting from imaging a customized set of flow cell regions during sequencing cycles in accordance with one or more embodiments of the present disclosure. The charts illustrated in FIG. 10 portray simulated data.

[0150] FIG. 10 illustrates a chart 1002 portraying read-coverage levels for genomic samples sequenced by existing systems and a chart 1004 portraying read-coverage levels for genomic samples sequenced by the sequence-to-coverage system 106. The chart 1002 portrays read-coverage levels for genomic samples 1006a, 1006b, and 1006c after 2x150 sequencing cycles imaging 100 flow cell regions. The chart 1004 portrays read-coverage levels for genomic samples 1008a, 1008b, and 1008c after 2x150 cycles imaging 70 flow cell regions.

[0151] As shown in FIG. 10, the existing system over-sequences the genomic samples 914a-914c. For example, in comparison to the 40x target read-coverage level, the existing system sequences the genomic samples 1006a-1006c to about a 72x read-coverage level when imaging 100 flow cell regions during sequencing cycles. In contrast, and as shown in the chart 1004, the sequence-to-coverage system 106 images a customized set of 70 flow cell regions during sequencing cycles. By reducing the number of imaged flow cell regions, the sequence-to-coverage system 106 sequences the genomic samples 916a-916c to just meet and barely exceed the 40x target read-coverage level.

[0152] Aspects of the present disclosure relate generally to devices, systems, and methods providing biological or chemical analysis. Various protocols in biological or chemical research involve performing a large number of controlled reactions on local support surfaces or within predefined reaction chambers. The designated reactions may then be observed or detected, and subsequent analysis may help identify or reveal properties of chemicals involved in the reaction. For example, in some multiplex assays, an unknown analyte having an identifiable label (e.g., fluorescent label) may be exposed to thousands of known probes under controlled conditions. Each known probe may be deposited into a corresponding well of a flow cell channel. Observing any

chemical reactions that occur between the known probes and the unknown analyte within the wells may help identify or reveal properties of the analyte. Other examples of such protocols include known DNA sequencing processes, such as sequencing-by-synthesis (SBS) or cyclic-array sequencing.

[0153] While a variety of devices, systems, and methods have been made and used to perform biological or chemical analysis, it is believed that no one prior to the inventor(s) has made or used the devices and techniques described herein.

[0154] FIG. 11 illustrates a schematic diagram of an example of a system (1100) that may be used to perform an analysis on one or more samples of interest. In some implementations, the sample may include one or more clusters of nucleotides (e.g., DNA) that have been linearized to form a single stranded DNA (sstDNA). In the implementation shown, system (1100) is configured to receive a flow cell cartridge assembly (1102) including a flow cell assembly (1103) and a sample cartridge (1104). System (1100) includes a flow cell receptacle (1122) that receives flow cell cartridge assembly (1102), a vacuum chuck (1124) that supports flow cell assembly (1103), and a flow cell interface (1126) that is used to establish a fluidic coupling between system (1100) and flow cell assembly (1103). Flow cell interface (1126) may include one or more manifolds. System (1100) further includes a sipper manifold assembly (1106), a sample loading manifold assembly (1108), and a pump manifold assembly (1110). System (1100) also includes a drive assembly (1112), a controller (1114), an imaging system (1116), and a waste reservoir (1118). Controller (1114) is electrically and/or communicatively coupled to drive assembly (1112) and to imaging system (1116); and is configured to cause drive assembly (1112) and/or the imaging system (1116) to perform various functions as disclosed herein.

[0155] In the present example, flow cell assembly (1103) includes a flow cell (1128) having a channel (1130) and defining a plurality of first openings (1132), which are fluidically coupled to the channel (1130) and arranged on a first side (1134) of the channel (1130). Flow cell (1128) further includes a plurality of second openings (1136) fluidically coupled to the channel (1130) and arranged on a second side (1138) of the channel (1130). Fluid may thus flow through flow cell (1128) via channel. While the flow cell (1128) is shown including one channel (1130), flow cell (1128) may include two or more channels (1130). Flow cell assembly (1103) also includes a flow cell manifold assembly (1140) coupled to flow cell (1128) and having a first manifold fluidic line (1142) and a second manifold fluidic line (1144). Flow cell manifold assembly (1140) may be in the form of a laminate including a plurality of layers as discussed in more detail below.

[0156] In the implementation shown, first manifold fluidic line (1142) has a first fluidic line opening (1146) and is fluidically coupled to each of the plurality of first openings (1132) of flow cell (1128); and second manifold fluidic line (1144) has a second fluidic line opening (1148) and

is fluidically coupled to each of the second openings (1136). As shown, flow cell assembly (1103) includes gaskets (1150) coupled to flow cell manifold assembly (1140) and fluidically coupled to fluidic line openings (1146, 1148). In some implementations where flow cell (1128) includes a plurality of channels (1130), flow cell manifold assembly (1140) may include additional fluidic lines (1152) that couple first fluidic line openings (1146) to a single manifold port (1154). In such implementations, a single gasket (1150) may be coupled to flow cell manifold assembly (1140) that surrounds the manifold port (1154) and is in fluidic communication with a plurality of channels (1130). In operation, flow cell interface (1126) engages with corresponding gaskets (1150) to establish a fluidic coupling between system (1100) and flow cell (1128). The engagement between flow cell interface (1126) and gaskets (1150) reduces or eliminates fluid leakage between flow cell interface (1126) and flow cell (1128).

[0157] In the implementation shown, first manifold fluidic line (1142) has a portion (1156) that is substantially parallel to a longitudinal axis (1158) of channel (1130); and second manifold fluidic line (1144) has a portion (1160) that is substantially parallel to longitudinal axis (1158) of channel (1130). Additionally, first manifold fluidic line (1142) is shown being at least partially adjacent a first end (1162) of flow cell (1128) and spaced from a second end (1164) of flow cell (1128); and second manifold fluidic line (1144) is shown being at least partially adjacent second end (1164) of flow cell (1128) and spaced from first end (1162). Other arrangements of manifold fluidic lines (1142, 1144) may prove suitable, however.

[0158] In the implementation shown, system (1100) includes a sample cartridge receptacle (1166) that receives sample cartridge (1104) that carries one or more samples of interest (e.g., an analyte). System (1100) also includes a sample cartridge interface (1168) that establishes a fluidic connection with sample cartridge (1104). Sample loading manifold assembly (1108) includes one or more sample valves (1170). Pump manifold assembly (1110) includes one or more pumps (1172), one or more pump valves (1174), and a cache (1176). Valves (1170, 1174) and pumps (1172) may take any suitable form. Cache (1176) may include a serpentine cache and may temporarily store one or more reaction components during, for example, bypass manipulations of the system (1100). While cache (1176) is shown being included in pump manifold assembly (1110), cache (1176) may alternatively be located elsewhere (e.g., in sipper manifold assembly (1106) or in another manifold downstream of a bypass fluidic line (1178), etc.).

[0159] Sample loading manifold assembly (1108) and pump manifold assembly (1110) flow one or more samples of interest from sample cartridge (1104) through a fluidic line (1180) toward flow cell cartridge assembly (1102). In some implementations, sample loading manifold assembly (1108) may individually load or address each channel (1130) of flow cell (1128) with a respective sample of interest. The process of loading channel (1130) with a sample of interest may occur

automatically using system (1100). As shown in FIG. 11, sample cartridge (1104) and sample loading manifold assembly (1108) are positioned downstream of flow cell cartridge assembly (1102). In the implementation shown, sample loading manifold assembly (1108) is coupled between flow cell cartridge assembly (1102) and pump manifold assembly (1110). To draw a sample of interest from sample cartridge (1104) and toward pump manifold assembly (1110), sample valves (1170), pump valves (1174), and/or pumps (1172) may be selectively actuated to urge the sample of interest toward pump manifold assembly (1110). Sample cartridge (1104) may include a plurality of sample reservoirs that are selectively fluidically accessible via the corresponding sample valves (1170). To individually flow the sample of interest toward channel (1130) of flow cell (1128) and away from pump manifold assembly (1110), sample valves (1170), pump valves (1174), and/or pumps (1172) may be selectively actuated to urge the sample of interest toward flow cell cartridge assembly (1102) and into respective channels (1130) of flow cell (1128).

[0160] Drive assembly (1112) interfaces with sipper manifold assembly (1106) and pump manifold assembly (1110) to flow one or more reagents that interact with the sample within flow cell (1128). In some scenarios, a reversible terminator is attached to the reagent to allow a single nucleotide to be incorporated onto a growing DNA strand. In some such implementations, one or more of the nucleotides has a unique fluorescent label that emits a color when excited. The color (or absence thereof) is used to detect the corresponding nucleotide. In the implementation shown, imaging system (1116) excites one or more of the identifiable labels (e.g., a fluorescent label) and thereafter obtains image data for the identifiable labels. The labels may be excited by incident light and/or a laser and the image data may include one or more colors emitted by the respective labels in response to the excitation. The image data (e.g., detection data) may be analyzed by system (1100). Examples of features and functionalities that may be incorporated into imaging system (1116) will be described in greater detail below.

[0161] After the image data is obtained, drive assembly (1112) interfaces with sipper manifold assembly (1106) and pump manifold assembly (1110) to flow another reaction component (e.g., a reagent) through flow cell (1128) that is thereafter received by waste reservoir (1118) via a primary waste fluidic line (1182) and/or otherwise exhausted by system (1100). Some reaction components may perform a flushing operation that chemically cleaves the fluorescent label and the reversible terminator from the ssDNA. The ssDNA may then be ready for another cycle.

[0162] The primary waste fluidic line (1182) is coupled between pump manifold assembly (1110) and waste reservoir (1118). In some implementations, pumps (1172) and/or pump valves (1174) of pump manifold assembly (1110) selectively flow the reaction components from flow cell cartridge assembly (1102), through fluidic line (1180) and sample loading manifold assembly (1108) to primary waste fluidic line (1182). Flow cell cartridge assembly (1102) is coupled to a

central valve (1184) via flow cell interface (1126). Central valve (1184) is coupled with flow cell interface (1126) via a fluidic line (1185). An auxiliary waste fluidic line (1186) is coupled to central valve (1184) and to waste reservoir (1118). In some implementations, auxiliary waste fluidic line (1186) receives excess fluid of a sample of interest from flow cell cartridge assembly (1102), via central valve (1184), and flows the excess fluid of the sample of interest to waste reservoir (1118) when back loading the sample of interest into flow cell (1128), as described herein.

[0163] Sipper manifold assembly (1106) includes a shared line valve (1188) and a bypass valve (1190). Shared line valve (1188) may be referred to as a reagent selector valve. Central valve (1184) and the valves (1188, 1190) of sipper manifold assembly (1106) may be selectively actuated to control the flow of fluid through fluidic lines (1192, 1194, 1196). Sipper manifold assembly (1106) may be coupled to a corresponding number of reagent reservoirs (1198) via reagent sippers (1200). Reagent reservoirs (1198) may contain fluid (e.g., reagent and/or another reaction component). In some implementations, sipper manifold assembly (1106) includes a plurality of ports. Each port of sipper manifold assembly (1106) may receive one of the reagent sippers (1200). Reagent sippers (1200) may be referred to as fluidic lines. Some forms of reagent sippers (1200) may include an array of sipper tubes extending downwardly along the z-dimension from ports in the body of sipper manifold assembly (1106). Reagent reservoirs (1198) may be provided in a cartridge, and the tubes of reagent sippers (1200) may be configured to be inserted into corresponding reagent reservoirs (1198) in the reagent cartridge so that liquid reagent may be drawn from each reagent reservoir (1198) into the sipper manifold assembly (1106).

[0164] Shared line valve (1188) of sipper manifold assembly (1106) is coupled to central valve (1184) via shared reagent fluidic line (1196). Different reagents may flow through shared reagent fluidic line (1196) at different times. In some versions, when performing a flushing operation before changing between one reagent and another, pump manifold assembly (1110) may draw wash buffer through shared reagent fluidic line (1196), central valve (1184), and flow cell cartridge assembly (1102).

[0165] Bypass valve (1190) of sipper manifold assembly (1106) is coupled to central valve (1184) via dedicated reagent fluidic lines (1194, 1196). Each of the dedicated reagent fluidic lines (1194, 1196) may be associated with a single reagent. The fluids that may flow through dedicated reagent fluidic lines (1194, 1196) may be used during sequencing operations and may include a cleave reagent, an incorporation reagent, a scan reagent, a cleave wash, and/or a wash buffer.

[0166] Bypass valve (1190) is also coupled to cache (1176) of pump manifold assembly (1110) via bypass fluidic line (1178). One or more reagent priming operations, hydration operations, mixing operations, and/or transfer operations may be performed using bypass fluidic line (1178). The priming operations, the hydration operations, the mixing operations, and/or the transfer

operations may be performed independent of flow cell cartridge assembly (1102). Thus, the operations using bypass fluidic line (1178) may occur during, for example, incubation of one or more samples of interest within flow cell cartridge assembly (1102). That is, shared line valve (1188) may be utilized independently of bypass valve (1190) such that bypass valve (1190) may utilize bypass fluidic line (1178) and/or cache (1176) to perform one or more operations while shared line valve (1188) and/or central valve (1184) simultaneously, substantially simultaneously, or offset synchronously perform other operations.

[0167] Drive assembly (1112) includes a pump drive assembly (1202) and a valve drive assembly (1204). Pump drive assembly (1202) may be adapted to interface with one or more pumps (1172) to pump fluid through flow cell (1128) and/or to load one or more samples of interest into flow cell (1128). Valve drive assembly (1204) may be adapted to interface with one or more of the valves (1170, 1174, 1184, 1188, 1190) to control the position of the corresponding valves (1170, 1174, 1184, 1188, 1190).

[0168] FIG. 12 shows an example of a fluidic arrangement (1220) that may be incorporated into a variation of system (1100). Fluidic arrangement (1220) of this example includes a pump manifold assembly (1222), which may operate similar to pump manifold assembly (1110) described above; a sample loading manifold assembly (1228), which may operate similar to sample loading manifold assembly (1108) described above; a flow cell interface (1240), which may operate similar to flow cell interface (1126) described above; a sipper manifold assembly (1250), which may operate similar to sipper manifold assembly (1106) described above; and a waste reservoir (1270), which may operate similar to waste reservoir (1118) described above. Pump manifold assembly (1222) is coupled with a port assembly (1258) of sipper manifold assembly (1250) via a fluidic line (1224), which may be similar to fluidic line (1178); and with sample loading manifold assembly (1228) via a fluidic line (1226). Sample loading manifold assembly (1228) is coupled with flow cell interface (1240) via fluidic line (1230), which may be similar to fluidic line (1180); and with port assembly (1258) via fluidic lines (1232, 1234). Flow cell interface (1240) is coupled with sipper manifold assembly (1250) via fluidic line (1242), which may be similar to fluidic line (1185). Sipper manifold assembly (1250) includes a manifold body (1252) and a common output port (1256), which provides fluid communication via fluidic line (1185). A valve assembly (1254) controls fluid flow through common output port (1256) and may operate similar to central valve (1184). Port assembly (1258) of sipper manifold assembly (1250) is coupled with waste reservoir (1270) via fluidic line (1272), which may be similar to fluidic line (1186).

[0169] A plurality of reagent sippers (1260) extend from manifold body (1252) and are fluidically coupled with valve assembly (1254) via respective fluid channels (1262) in manifold body (1252). Reagent sippers (1260) may operate similar to reagent sippers (1200). Valve

assembly (1254) is operable to selectively couple fluid channels (1262) with flow cell interface (1240) via common output port (1256) and fluidic line (1230), to thereby selectively provide various reagents to flow cell interface (1240). In other words, when each reagent sipper (1260) is disposed in a different respective reagent (e.g., in a respective reagent reservoir (1198)), a flow cell (e.g., like flow cell (1128)) that is coupled with flow cell interface (1240) may selectively receive those different reagents based on control of valve assembly (1254).

[0170] A plurality of reagent sippers (1260) extend from manifold body (1252) and are fluidically coupled with valve assembly (1254) via respective fluid channels (1262) in manifold body (1252). Reagent sippers (1260) may operate similar to reagent sippers (1200). Valve assembly (1254) is operable to selectively couple fluid channels (1262) with flow cell interface (1240) via common output port (1256) and fluidic line (1230), to thereby selectively provide various reagents to flow cell interface (1240). In other words, when each reagent sipper (1260) is disposed in a different respective reagent (e.g., in a respective reagent reservoir (1198)), a flow cell (e.g., like flow cell (1128)) that is coupled with flow cell interface (1240) may selectively receive those different reagents based on control of valve assembly (1254).

[0171] Referring back to FIG. 11, controller (1114) of the present example includes a user interface (1206), a communication interface (1208), one or more processors (1210), and a memory (1212) storing instructions executable by the one or more processors (1210) to perform various functions including the disclosed implementations. User interface (1206), communication interface (1133), and memory (1212) are electrically and/or communicatively coupled to the one or more processors (1210). User interface (1206) may be adapted to receive input from a user and to provide information to the user associated with the operation of system (1100) and/or an analysis taking place. User interface (1206) may include a touch screen, a display, a keyboard, a speaker(s), a mouse, a track ball, and/or a voice recognition system.

[0172] Communication interface (1208) is adapted to enable communication between system (1100) and a remote system(s) (e.g., computers) via a network(s) (e.g., the Internet, an intranet, a local-area network (LAN), a wide-area network (WAN), a coaxial-cable network, a wireless network, a wired network, a satellite network, a digital subscriber line (DSL) network, a cellular network, a Bluetooth connection, a near field communication (NFC) connection, etc.). Some of the communications provided to the remote system may be associated with analysis results, imaging data, etc. generated or otherwise obtained by system (1100). Some of the communications provided to system (1100) may be associated with a fluidics analysis operation, patient records, and/or a protocol(s) to be executed by system (1100).

[0173] The one or more processors (1210) and/or system (1100) may include one or more of a processor-based system(s) or a microprocessor-based system(s). In some implementations, the

one or more processors (1210) and/or system (1100) includes one or more of a programmable processor, a programmable controller, a microprocessor, a microcontroller, a graphics processing unit (GPU), a digital signal processor (DSP), a reduced-instruction set computer (RISC), an application specific integrated circuit (ASIC), a field programmable gate array (FPGA), a field programmable logic device (FPLD), a logic circuit, and/or another logic-based device executing various functions including the ones described herein.

[0174] Memory (1212) may include one or more of a semiconductor memory, a magnetically readable memory, an optical memory, a hard disk drive (HDD), an optical storage drive, a solid-state storage device, a solid-state drive (SSD), a flash memory, a read-only memory (ROM), erasable programmable read-only memory (EPROM), electrically erasable programmable read-only memory (EEPROM), a random-access memory (RAM), a non-volatile RAM (NVRAM) memory, a compact disc (CD), a compact disc read-only memory (CD-ROM), a digital versatile disk (DVD), a Blu-ray disk, a redundant array of independent disks (RAID) system, a cache and/or any other storage device or storage disk in which information is stored for any duration (e.g., permanently, temporarily, for extended periods of time, for buffering, for caching).

[0175] FIGS. 1-12, the corresponding text, and the examples provide a number of different methods, systems, devices, and non-transitory computer-readable media of the sequence-to-coverage system 106. In addition to the foregoing, one or more implementations can also be described in terms of flowcharts comprising acts for accomplishing a particular result, as shown in FIGS. 13A-13B. FIG. 13A illustrates a flowchart of a series of acts 1300 for executing a sequencing run until finishing a customized number of sequencing cycles in accordance with one or more embodiments of the present disclosure. FIG. 13B illustrates a flowchart of a series of acts 1362 for executing a sequencing run by capturing images of a customized set of flow cell regions in accordance with one or more embodiments of the present disclosure. While FIGS. 13A-13B illustrate acts according to one embodiment, alternative embodiments may omit, add to, reorder, and/or modify any of the acts shown in FIGS. 13A-13B. The acts of FIGS. 13A-13B can be performed as part of a method. Alternatively, a non-transitory computer readable storage medium can comprise instructions that, when executed by one or more processors, cause a computing device or a system to perform the acts depicted in FIGS. 13A-13B. In still further embodiments, a system comprising an imaging system, a fluidic system, and a computer comprising: at least one processor; and a non-transitory computer readable medium comprising instructions that, when executed by one or more processors, cause the system to perform the acts of FIGS. 13A-13B.

[0176] As shown in FIG. 13A, the series of acts 1300 includes an act 1310 of determining base calls for indexing sequences, an act 1320 of determining respective numbers of clusters belonging to genomic samples, an act 1330 of estimating read-coverage levels, an act 1340 of generating a

customized number of sequencing cycles, and an act 1350 of executing the sequencing run. For example, the series of acts 1300 can include acts to perform any of the operations described in the following clauses:

CLAUSE 1. A method comprising:

determining, from a subset of sequencing cycles of a sequencing run for genomic samples, base calls for indexing sequences within clusters of oligonucleotides;

determining, based on the indexing sequences, respective numbers of clusters of oligonucleotides belonging to respective genomic samples of the genomic samples;

estimating read-coverage levels for the genomic samples based on the respective numbers of clusters of oligonucleotides belonging to respective genomic samples and a currently selected number of sequencing cycles for the sequencing run;

generating, for the sequencing run and based on the estimated read-coverage levels, a customized number of sequencing cycles sufficient to generate nucleotide reads satisfying a target read-coverage level for each genomic sample of the genomic samples; and

executing the sequencing run until finishing the customized number of sequencing cycles.

CLAUSE 2. The method of clause 1, further comprising estimating the read-coverage levels by:

determining filter metrics indicating subsets of clusters of oligonucleotides satisfying a filtering threshold for signals of the clusters of oligonucleotides; and

estimating the read-coverage levels for the genomic samples based on the filter metrics and the respective numbers of clusters of oligonucleotides belonging to respective genomic samples.

CLAUSE 3. The method of clause 2, further comprising determining the filter metrics by determining, in a pass filter map, a percentage of clusters belonging to each genomic sample that satisfy a chastity filter for signals emitted from the clusters of oligonucleotides.

CLAUSE 4. The method of clause 2, further comprising estimating the read-coverage levels for the genomic samples by:

determining, based on the filter metrics and the respective numbers of clusters of oligonucleotides belonging to respective genomic samples, a number of filter-passing clusters of oligonucleotides for each genomic sample of the genomic samples that satisfy the filtering threshold; and

estimating a minimum number of nucleotide reads covering genomic regions of each genomic sample based on the number of filter-passing clusters of oligonucleotides.

CLAUSE 5. The method of clause 1, further comprising:

determining, based on the estimated read-coverage levels, a customized set of flow cell regions to be imaged from a flow cell sufficient to generate the nucleotide reads satisfying the target read-coverage level for each genomic sample of the genomic samples; and

executing the sequencing run by capturing images of the customized set of flow cell regions for the customized number of sequencing cycles using the imaging system.

CLAUSE 6. The method of clause 1, further comprising performing the subset of sequencing cycles according to an order of indexing cycles before genomic sequencing cycles by:

determining base calls for a first indexing sequence appended to a sample genomic sequence of a genomic sample;

determining base calls for a second indexing sequence appended to the sample genomic sequence of the genomic sample; and

after determining the base calls for the first indexing sequence and the second indexing sequence, determining base calls for a first nucleotide read corresponding to a first portion of the sample genomic sequence and determining base calls for a second nucleotide read corresponding to a second portion of the sample genomic sequence.

CLAUSE 7. The method of clause 1, further comprising determining the respective numbers of clusters of oligonucleotides belonging to the respective genomic samples by:

identifying, from among the indexing sequences, assigned indexing sequences matching indexing sequences registered for the sequencing run and unassigned indexing sequences that do not match the indexing sequences registered for the sequencing run;

removing, from data for the sequencing run, a subset of clusters of oligonucleotides corresponding to the unassigned indexing sequences;

determining respective subsets of assigned indexing sequences that correspond to the respective genomic samples; and

determining, from among the respective subsets of assigned indexing sequences, a number of clusters of oligonucleotides belonging to each genomic sample.

CLAUSE 8. The method of clause 1, further comprising generate the customized number of sequencing cycles for the sequencing run by increasing or decreasing a preset number of sequencing cycles for the sequencing run.

CLAUSE 9. The method of clause 1, further comprising generating the customized number of sequencing cycles for the sequencing run by:

identifying a minimum number of sequencing cycles and a maximum number of sequencing cycles for the sequencing run; and

increasing or decreasing a preset number of sequencing cycles for the sequencing run to the customized number of sequencing cycles within the minimum number of sequencing cycles and the maximum number of sequencing cycles.

CLAUSE 10. The method of clause 1, further comprising estimating the read-coverage levels by:

determining, from the sequencing run, a number of unique nucleotide reads aligned with a reference genome;

determining, from the sequencing run, a number of filter-passing nucleotide reads from filter-passing cluster of oligonucleotides with signals that satisfy a filtering threshold;

determining a bioinformatics efficiency metric by dividing the number of unique nucleotide reads by the number of filter-passing nucleotide reads; and

estimating the read-coverage levels for the genomic samples based on the bioinformatics efficiency metric and the respective numbers of clusters of oligonucleotides belonging to respective genomic samples.

CLAUSE 11. The method of clause 1, further comprising detecting a reagent volume of a reagent cartridge in fluid communication with the fluidic system and operating the fluidic system to perform one or more additional sequencing cycles relative to the currently selected number of sequencing cycles until finishing the customized number of sequencing cycles by aspirating one or more reagents from the reagent cartridge.

CLAUSE 12. The method of clause 1, further comprising terminating operation of the fluidic system from performing one or more sequencing cycles of the currently selected number of sequencing cycles to finish the sequencing run after performing the customized number of sequencing cycles.

[0177] As shown in FIG. 13B, the series of acts 1362 includes an act 1360 of determining base calls for indexing sequences, an act 1370 of determining respective numbers of clusters belonging to genomic samples, an act 1380 of estimating read-coverage levels, an act 1390 of determining a customized set of flow cell regions to be imaged, and an act 1392 of executing the sequencing run. For example, the series of acts 1362 can include acts to perform any of the operations described in the following clauses:

CLAUSE 13. A method comprising:

determining, from a subset of sequencing cycles of a sequencing run for genomic samples, base calls for indexing sequences within clusters of oligonucleotides;

determining, based on the indexing sequences, respective numbers of clusters of oligonucleotides belonging to respective genomic samples of the genomic samples;

estimating read-coverage levels for the genomic samples based on the respective numbers of clusters of oligonucleotides belonging to respective genomic samples and a currently selected number of sequencing cycles for the sequencing run;

determining, from a flow cell and based on the estimated read-coverage level, a customized set of flow cell regions to be imaged sufficient to generate nucleotide reads satisfying a target read-coverage level for each genomic sample of the genomic samples; and

executing the sequencing run by capturing images of the customized set of flow cell regions during sequencing cycles of the sequencing run.

CLAUSE 14. The method of clause 13, further comprising determining the customized set of flow cell regions by determining a customized number of flow cell regions to be imaged sufficient to generate the nucleotide reads satisfying the target read-coverage level for each genomic sample.

CLAUSE 15. The method of clause 13, further comprising determining the customized set of flow cell regions by determining, from a flow cell, a set of tiles to be imaged sufficient to generate the nucleotide reads satisfying the target read-coverage level for each genomic sample.

CLAUSE 16. The method of clause 13, further comprising capturing the images of the customized set of flow cell regions without adjusting a currently selected number of sequencing cycles for the sequencing run.

CLAUSE 17. The method of clause 13, further comprising determining the customized set of flow cell regions by increasing or decreasing a number of flow cell regions from an initial set of flow cell regions selected for the sequencing run.

CLAUSE 18. The method of clause 13, further comprising estimating the read-coverage levels by:

determining filter metrics indicating subsets of clusters of oligonucleotides satisfying a filtering threshold for signals of the clusters of oligonucleotides; and

estimating the read-coverage levels for the genomic samples based on the filter metrics and the respective numbers of clusters of oligonucleotides belonging to respective genomic samples.

CLAUSE 19. The method of clause 18, further comprising determining the filter metrics by determining, in a pass filter map, a percentage of clusters belonging to each genomic sample that satisfy a chastity filter for signals emitted from the clusters of oligonucleotides.

CLAUSE 20. The method of clause 18, further comprising estimating the read-coverage levels for the genomic samples by:

determining, based on the filter metrics and the respective numbers of clusters of oligonucleotides belonging to respective genomic samples, a number of filter-passing clusters of

oligonucleotides for each genomic sample of the genomic samples that satisfy the filtering threshold; and

estimating a minimum number of nucleotide reads covering genomic regions of each genomic sample based on the number of filter-passing clusters of oligonucleotides.

[0178] The methods described herein can be used in conjunction with a variety of nucleic acid sequencing techniques. Particularly applicable techniques are those wherein nucleic acids are attached at fixed locations in an array such that their relative positions do not change and wherein the array is repeatedly imaged. Embodiments in which images are obtained in different color channels, for example, coinciding with different labels used to distinguish one nucleobase type from another are particularly applicable. In some embodiments, the process to determine the nucleotide sequence of a target nucleic acid (i.e., a nucleic-acid polymer) can be an automated process. Preferred embodiments include sequencing-by-synthesis (SBS) techniques.

[0179] SBS techniques generally involve the enzymatic extension of a nascent nucleic acid strand through the iterative addition of nucleotides against a template strand. In traditional methods of SBS, a single nucleotide monomer may be provided to a target nucleotide in the presence of a polymerase in each delivery. However, in the methods described herein, more than one type of nucleotide monomer can be provided to a target nucleic acid in the presence of a polymerase in a delivery.

[0180] SBS can utilize nucleotide monomers that have a terminator moiety or those that lack any terminator moieties. Methods utilizing nucleotide monomers lacking terminators include, for example, pyrosequencing and sequencing using γ -phosphate-labeled nucleotides, as set forth in further detail below. In methods using nucleotide monomers lacking terminators, the number of nucleotides added in each cycle is generally variable and dependent upon the template sequence and the mode of nucleotide delivery. For SBS techniques that utilize nucleotide monomers having a terminator moiety, the terminator can be effectively irreversible under the sequencing conditions used as is the case for traditional Sanger sequencing which utilizes dideoxynucleotides, or the terminator can be reversible as is the case for sequencing methods developed by Solexa (now Illumina, Inc.).

[0181] SBS techniques can utilize nucleotide monomers that have a label moiety or those that lack a label moiety. Accordingly, incorporation events can be detected based on a characteristic of the label, such as fluorescence of the label; a characteristic of the nucleotide monomer such as molecular weight or charge; a byproduct of incorporation of the nucleotide, such as release of pyrophosphate; or the like. In embodiments, where two or more different nucleotides are present in a sequencing reagent, the different nucleotides can be distinguishable from each other, or alternatively, the two or more different labels can be the indistinguishable under the detection

techniques being used. For example, the different nucleotides present in a sequencing reagent can have different labels and they can be distinguished using appropriate optics as exemplified by the sequencing methods developed by Solexa (now Illumina, Inc.).

[0182] Preferred embodiments include pyrosequencing techniques. Pyrosequencing detects the release of inorganic pyrophosphate (PPi) as particular nucleotides are incorporated into the nascent strand (Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. and Nyren, P. (1996) "Real-time DNA sequencing using detection of pyrophosphate release." *Analytical Biochemistry* 242(1), 84-9; Ronaghi, M. (2001) "Pyrosequencing sheds light on DNA sequencing." *Genome Res.* 11(1), 3-11; Ronaghi, M., Uhlen, M. and Nyren, P. (1998) "A sequencing method based on real-time pyrophosphate." *Science* 281(5375), 363; U.S. Pat. No. 6,210,891; U.S. Pat. No. 6,258,568 and U.S. Pat. No. 6,274,320, the disclosures of which are incorporated herein by reference in their entireties). In pyrosequencing, released PPi can be detected by being immediately converted to adenosine triphosphate (ATP) by ATP sulfurylase, and the level of ATP generated is detected via luciferase-produced photons. The nucleic acids to be sequenced can be attached to features in an array and the array can be imaged to capture the chemiluminescent signals that are produced due to incorporation of a nucleotides at the features of the array. An image can be obtained after the array is treated with a particular nucleotide type (e.g., A, T, C or G). Images obtained after addition of each nucleotide type will differ with regard to which features in the array are detected. These differences in the image reflect the different sequence content of the features on the array. However, the relative locations of each feature will remain unchanged in the images. The images can be stored, processed and analyzed using the methods set forth herein. For example, images obtained after treatment of the array with each different nucleotide type can be handled in the same way as exemplified herein for images obtained from different detection channels for reversible terminator-based sequencing methods.

[0183] In another exemplary type of SBS, cycle sequencing is accomplished by stepwise addition of reversible terminator nucleotides containing, for example, a cleavable or photobleachable dye label as described, for example, in WO 04/018497 and U.S. Pat. No. 7,057,026, the disclosures of which are incorporated herein by reference. This approach is being commercialized by Solexa (now Illumina Inc.), and is also described in WO 91/06678 and WO 07/123,744, each of which is incorporated herein by reference. The availability of fluorescently-labeled terminators in which both the termination can be reversed and the fluorescent label cleaved facilitates efficient cyclic reversible termination (CRT) sequencing. Polymerases can also be co-engineered to efficiently incorporate and extend from these modified nucleotides.

[0184] Preferably in reversible terminator-based sequencing embodiments, the labels do not substantially inhibit extension under SBS reaction conditions. However, the detection labels can

be removable, for example, by cleavage or degradation. Images can be captured following incorporation of labels into arrayed nucleic acid features. In particular embodiments, each cycle involves simultaneous delivery of four different nucleotide types to the array and each nucleotide type has a spectrally distinct label. Four images can then be obtained, each using a detection channel that is selective for one of the four different labels. Alternatively, different nucleotide types can be added sequentially and an image of the array can be obtained between each addition step. In such embodiments, each image will show nucleic acid features that have incorporated nucleotides of a particular type. Different features are present or absent in the different images due to the different sequence content of each feature. However, the relative position of the features will remain unchanged in the images. Images obtained from such reversible terminator-SBS methods can be stored, processed and analyzed as set forth herein. Following the image capture step, labels can be removed and reversible terminator moieties can be removed for subsequent cycles of nucleotide addition and detection. Removal of the labels after they have been detected in a particular cycle and prior to a subsequent cycle can provide the advantage of reducing background signal and crosstalk between cycles. Examples of useful labels and removal methods are set forth below.

[0185] In particular embodiments some or all of the nucleotide monomers can include reversible terminators. In such embodiments, reversible terminators/cleavable fluors can include fluor linked to the ribose moiety via a 3' ester linkage (Metzker, *Genome Res.* 15:1767-1776 (2005), which is incorporated herein by reference). Other approaches have separated the terminator chemistry from the cleavage of the fluorescence label (Ruparel et al., *Proc Natl Acad Sci USA* 102: 5932-7 (2005), which is incorporated herein by reference in its entirety). Ruparel et al described the development of reversible terminators that used a small 3' allyl group to block extension, but could easily be deblocked by a short treatment with a palladium catalyst. The fluorophore was attached to the base via a photocleavable linker that could easily be cleaved by a 30 second exposure to long wavelength UV light. Thus, either disulfide reduction or photocleavage can be used as a cleavable linker. Another approach to reversible termination is the use of natural termination that ensues after placement of a bulky dye on a dNTP. The presence of a charged bulky dye on the dNTP can act as an effective terminator through steric and/or electrostatic hindrance. The presence of one incorporation event prevents further incorporations unless the dye is removed. Cleavage of the dye removes the fluor and effectively reverses the termination. Examples of modified nucleotides are also described in U.S. Pat. No. 7,427,673, and U.S. Pat. No. 7,057,026, the disclosures of which are incorporated herein by reference in their entireties.

[0186] Additional exemplary SBS systems and methods which can be utilized with the methods and systems described herein are described in U.S. Patent Application Publication No.

2007/0166705, U.S. Patent Application Publication No. 2006/0188901, U.S. Pat. No. 7,057,026, U.S. Patent Application Publication No. 2006/0240439, U.S. Patent Application Publication No. 2006/0281109, PCT Publication No. WO 05/065814, U.S. Patent Application Publication No. 2005/0100900, PCT Publication No. WO 06/064199, PCT Publication No. WO 07/010,251, U.S. Patent Application Publication No. 2012/0270305 and U.S. Patent Application Publication No. 2013/0260372, the disclosures of which are incorporated herein by reference in their entireties.

[0187] Some embodiments can utilize detection of four different nucleotides using fewer than four different labels. For example, SBS can be performed utilizing methods and systems described in the incorporated materials of U.S. Patent Application Publication No. 2013/0079232. As a first example, a pair of nucleotide types can be detected at the same wavelength, but distinguished based on a difference in intensity for one member of the pair compared to the other, or based on a change to one member of the pair (e.g. via chemical modification, photochemical modification or physical modification) that causes apparent signal to appear or disappear compared to the signal detected for the other member of the pair. As a second example, three of four different nucleotide types can be detected under particular conditions while a fourth nucleotide type lacks a label that is detectable under those conditions, or is minimally detected under those conditions (e.g., minimal detection due to background fluorescence, etc.). Incorporation of the first three nucleotide types into a nucleic acid can be determined based on presence of their respective signals and incorporation of the fourth nucleotide type into the nucleic acid can be determined based on absence or minimal detection of any signal. As a third example, one nucleotide type can include label(s) that are detected in two different channels, whereas other nucleotide types are detected in no more than one of the channels. The aforementioned three exemplary configurations are not considered mutually exclusive and can be used in various combinations. An exemplary embodiment that combines all three examples, is a fluorescent-based SBS method that uses a first nucleotide type that is detected in a first channel (e.g. dATP having a label that is detected in the first channel when excited by a first excitation wavelength), a second nucleotide type that is detected in a second channel (e.g. dCTP having a label that is detected in the second channel when excited by a second excitation wavelength), a third nucleotide type that is detected in both the first and the second channel (e.g. dTTP having at least one label that is detected in both channels when excited by the first and/or second excitation wavelength) and a fourth nucleotide type that lacks a label that is not, or minimally, detected in either channel (e.g. dGTP having no label).

[0188] Further, as described in the incorporated materials of U.S. Patent Application Publication No. 2013/0079232, sequencing data can be obtained using a single channel. In such so-called one-dye sequencing approaches, the first nucleotide type is labeled but the label is removed after the first image is generated, and the second nucleotide type is labeled only after a

first image is generated. The third nucleotide type retains its label in both the first and second images, and the fourth nucleotide type remains unlabeled in both images.

[0189] Some embodiments can utilize sequencing by ligation techniques. Such techniques utilize DNA ligase to incorporate oligonucleotides and identify the incorporation of such oligonucleotides. The oligonucleotides typically have different labels that are correlated with the identity of a particular nucleotide in a sequence to which the oligonucleotides hybridize. As with other SBS methods, images can be obtained following treatment of an array of nucleic acid features with the labeled sequencing reagents. Each image will show nucleic acid features that have incorporated labels of a particular type. Different features are present or absent in the different images due the different sequence content of each feature, but the relative position of the features will remain unchanged in the images. Images obtained from ligation-based sequencing methods can be stored, processed and analyzed as set forth herein. Exemplary SBS systems and methods which can be utilized with the methods and systems described herein are described in U.S. Pat. No. 6,969,488, U.S. Pat. No. 6,172,218, and U.S. Pat. No. 6,306,597, the disclosures of which are incorporated herein by reference in their entireties.

[0190] Some embodiments can utilize nanopore sequencing (Deamer, D. W. & Akesson, M. "Nanopores and nucleic acids: prospects for ultrarapid sequencing." Trends Biotechnol. 18, 147-151 (2000); Deamer, D. and D. Branton, "Characterization of nucleic acids by nanopore analysis". Acc. Chem. Res. 35:817-825 (2002); Li, J., M. Gershow, D. Stein, E. Brandin, and J. A. Golovchenko, "DNA molecules and configurations in a solid-state nanopore microscope" Nat. Mater. 2:611-615 (2003), the disclosures of which are incorporated herein by reference in their entireties). In such embodiments, the target nucleic acid passes through a nanopore. The nanopore can be a synthetic pore or biological membrane protein, such as α -hemolysin. As the target nucleic acid passes through the nanopore, each base-pair can be identified by measuring fluctuations in the electrical conductance of the pore. (U.S. Pat. No. 7,001,792; Soni, G. V. & Meller, "A. Progress toward ultrafast DNA sequencing using solid-state nanopores." Clin. Chem. 53, 1996-2001 (2007); Healy, K. "Nanopore-based single-molecule DNA analysis." Nanomed. 2, 459-481 (2007); Cockroft, S. L., Chu, J., Amorin, M. & Ghadiri, M. R. "A single-molecule nanopore device detects DNA polymerase activity with single-nucleotide resolution." J. Am. Chem. Soc. 130, 818-820 (2008), the disclosures of which are incorporated herein by reference in their entireties). Data obtained from nanopore sequencing can be stored, processed and analyzed as set forth herein. In particular, the data can be treated as an image in accordance with the exemplary treatment of optical images and other images that is set forth herein.

[0191] Some embodiments can utilize methods involving the real-time monitoring of DNA polymerase activity. Nucleotide incorporations can be detected through fluorescence resonance

energy transfer (FRET) interactions between a fluorophore-bearing polymerase and γ -phosphate-labeled nucleotides as described, for example, in U.S. Pat. No. 7,329,492 and U.S. Pat. No. 7,211,414 (each of which is incorporated herein by reference) or nucleotide incorporations can be detected with zero-mode waveguides as described, for example, in U.S. Pat. No. 7,315,019 (which is incorporated herein by reference) and using fluorescent nucleotide analogs and engineered polymerases as described, for example, in U.S. Pat. No. 7,405,281 and U.S. Patent Application Publication No. 2008/0108082 (each of which is incorporated herein by reference). The illumination can be restricted to a zeptoliter-scale volume around a surface-tethered polymerase such that incorporation of fluorescently labeled nucleotides can be observed with low background (Levene, M. J. et al. "Zero-mode waveguides for single-molecule analysis at high concentrations." *Science* 299, 682-686 (2003); Lundquist, P. M. et al. "Parallel confocal detection of single molecules in real time." *Opt. Lett.* 33, 1026-1028 (2008); Korlach, J. et al. "Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nano structures." *Proc. Natl. Acad. Sci. USA* 105, 1176-1181 (2008), the disclosures of which are incorporated herein by reference in their entireties). Images obtained from such methods can be stored, processed and analyzed as set forth herein.

[0192] Some SBS embodiments include detection of a proton released upon incorporation of a nucleotide into an extension product. For example, sequencing based on detection of released protons can use an electrical detector and associated techniques that are commercially available from Ion Torrent (Guilford, CT, a Life Technologies subsidiary) or sequencing methods and systems described in US 2009/0026082 A1; US 2009/0127589 A1; US 2010/0137143 A1; or US 2010/0282617 A1, each of which is incorporated herein by reference. Methods set forth herein for amplifying target nucleic acids using kinetic exclusion can be readily applied to substrates used for detecting protons. More specifically, methods set forth herein can be used to produce clonal populations of amplicons that are used to detect protons.

[0193] The above SBS methods can be advantageously carried out in multiplex formats such that multiple different target nucleic acids are manipulated simultaneously. In particular embodiments, different target nucleic acids can be treated in a common reaction vessel or on a surface of a particular substrate. This allows convenient delivery of sequencing reagents, removal of unreacted reagents and detection of incorporation events in a multiplex manner. In embodiments using surface-bound target nucleic acids, the target nucleic acids can be in an array format. In an array format, the target nucleic acids can be typically bound to a surface in a spatially distinguishable manner. The target nucleic acids can be bound by direct covalent attachment, attachment to a bead or other particle or binding to a polymerase or other molecule that is attached to the surface. The array can include a single copy of a target nucleic acid at each site (also referred

to as a feature) or multiple copies having the same sequence can be present at each site or feature. Multiple copies can be produced by amplification methods such as, bridge amplification or emulsion PCR as described in further detail below.

[0194] The methods set forth herein can use arrays having features at any of a variety of densities including, for example, at least about 10 features/cm², 100 features/cm², 500 features/cm², 1,000 features/cm², 5,000 features/cm², 10,000 features/cm², 50,000 features/cm², 100,000 features/cm², 1,000,000 features/cm², 5,000,000 features/cm², or higher.

[0195] An advantage of the methods set forth herein is that they provide for rapid and efficient detection of a plurality of target nucleic acid in parallel. Accordingly the present disclosure provides integrated systems capable of preparing and detecting nucleic acids using techniques known in the art such as those exemplified above. Thus, an integrated system of the present disclosure can include fluidic components capable of delivering amplification reagents and/or sequencing reagents to one or more immobilized DNA fragments, the system comprising components such as pumps, valves, reservoirs, fluidic lines and the like. A flow cell can be configured and/or used in an integrated system for detection of target nucleic acids. Exemplary flow cells are described, for example, in US 2010/0111768 A1 and US Ser. No. 13/273,666, each of which is incorporated herein by reference. As exemplified for flow cells, one or more of the fluidic components of an integrated system can be used for an amplification method and for a detection method. Taking a nucleic acid sequencing embodiment as an example, one or more of the fluidic components of an integrated system can be used for an amplification method set forth herein and for the delivery of sequencing reagents in a sequencing method such as those exemplified above. Alternatively, an integrated system can include separate fluidic systems to carry out amplification methods and to carry out detection methods. Examples of integrated sequencing systems that are capable of creating amplified nucleic acids and also determining the sequence of the nucleic acids include, without limitation, the MiSeq™ platform (Illumina, Inc., San Diego, CA) and devices described in US Ser. No. 13/273,666, which is incorporated herein by reference.

[0196] The sequencing system described above sequences nucleic-acid polymers present in samples received by a sequencing device. As defined herein, “sample” and its derivatives, is used in its broadest sense and includes any specimen, culture and the like that is suspected of including a target. In some embodiments, the sample comprises DNA, RNA, PNA, LNA, chimeric or hybrid forms of nucleic acids. The sample can include any biological, clinical, surgical, agricultural, atmospheric or aquatic-based specimen containing one or more nucleic acids. The term also includes any isolated nucleic acid sample such a genomic DNA, fresh-frozen or formalin-fixed paraffin-embedded nucleic acid specimen. It is also envisioned that the sample can be from a single

individual, a collection of nucleic acid samples from genetically related members, nucleic acid samples from genetically unrelated members, nucleic acid samples (matched) from a single individual such as a tumor sample and normal tissue sample, or sample from a single source that contains two distinct forms of genetic material such as maternal and fetal DNA obtained from a maternal subject, or the presence of contaminating bacterial DNA in a sample that contains plant or animal DNA. In some embodiments, the source of nucleic acid material can include nucleic acids obtained from a newborn, for example as typically used for newborn screening.

[0197] The nucleic acid sample can include high molecular weight material such as genomic DNA (gDNA). The sample can include low molecular weight material such as nucleic acid molecules obtained from FFPE or archived DNA samples. In another embodiment, low molecular weight material includes enzymatically or mechanically fragmented DNA. The sample can include cell-free circulating DNA. In some embodiments, the sample can include nucleic acid molecules obtained from biopsies, tumors, scrapings, swabs, blood, mucus, urine, plasma, semen, hair, laser capture micro-dissections, surgical resections, and other clinical or laboratory obtained samples. In some embodiments, the sample can be an epidemiological, agricultural, forensic or pathogenic sample. In some embodiments, the sample can include nucleic acid molecules obtained from an animal such as a human or mammalian source. In another embodiment, the sample can include nucleic acid molecules obtained from a non-mammalian source such as a plant, bacteria, virus or fungus. In some embodiments, the source of the nucleic acid molecules may be an archived or extinct sample or species.

[0198] Further, the methods and compositions disclosed herein may be useful to amplify a nucleic acid sample having low-quality nucleic acid molecules, such as degraded and/or fragmented genomic DNA from a forensic sample. In one embodiment, forensic samples can include nucleic acids obtained from a crime scene, nucleic acids obtained from a missing persons DNA database, nucleic acids obtained from a laboratory associated with a forensic investigation or include forensic samples obtained by law enforcement agencies, one or more military services or any such personnel. The nucleic acid sample may be a purified sample or a crude DNA containing lysate, for example derived from a buccal swab, paper, fabric or other substrate that may be impregnated with saliva, blood, or other bodily fluids. As such, in some embodiments, the nucleic acid sample may comprise low amounts of, or fragmented portions of DNA, such as genomic DNA. In some embodiments, target sequences can be present in one or more bodily fluids including but not limited to, blood, sputum, plasma, semen, urine and serum. In some embodiments, target sequences can be obtained from hair, skin, tissue samples, autopsy or remains of a victim. In some embodiments, nucleic acids including one or more target sequences can be obtained from a deceased animal or human. In some embodiments, target sequences can include nucleic acids

obtained from non-human DNA such a microbial, plant or entomological DNA. In some embodiments, target sequences or amplified target sequences are directed to purposes of human identification. In some embodiments, the disclosure relates generally to methods for identifying characteristics of a forensic sample. In some embodiments, the disclosure relates generally to human identification methods using one or more target specific primers disclosed herein or one or more target specific primers designed using the primer design criteria outlined herein. In one embodiment, a forensic or human identification sample containing at least one target sequence can be amplified using any one or more of the target-specific primers disclosed herein or using the primer criteria outlined herein.

[0199] The components of the sequence-to-coverage system 106 can include software, hardware, or both. For example, the components of the sequence-to-coverage system 106 can include one or more instructions stored on a computer-readable storage medium and executable by processors of one or more computing devices (e.g., the local server device 102). When executed by the one or more processors, the computer-executable instructions of the sequence-to-coverage system 106 can cause the computing devices to perform the bubble detection methods described herein. Alternatively, the components of the sequence-to-coverage system 106 can comprise hardware, such as special purpose processing devices to perform a certain function or group of functions. Additionally, or alternatively, the components of the sequence-to-coverage system 106 can include a combination of computer-executable instructions and hardware.

[0200] Furthermore, the components of the sequence-to-coverage system 106 performing the functions described herein with respect to the sequence-to-coverage system 106 may, for example, be implemented as part of a stand-alone application, as a module of an application, as a plug-in for applications, as a library function or functions that may be called by other applications, and/or as a cloud-computing model. Thus, components of the sequence-to-coverage system 106 may be implemented as part of a stand-alone application on a personal computing device or a mobile device. Additionally, or alternatively, the components of the sequence-to-coverage system 106 may be implemented in any application that provides sequencing services including, but not limited to Illumina, BaseSpace, Illumina MiSeq, Illumina NovaSeq, Illumina NextSeq, Illumina TruSeq, or Illumina TruSight software. “Illumina,” “BaseSpace,” “MiSeq,” “NovaSeq,” “NextSeq,” “TruSeq,” and “TruSight,” are either registered trademarks or trademarks of Illumina, Inc. in the United States and/or other countries.

[0201] Embodiments of the present disclosure may comprise or utilize a special purpose or general-purpose computer including computer hardware, such as, for example, one or more processors and system memory, as discussed in greater detail below. Embodiments within the scope of the present disclosure also include physical and other computer-readable media for

carrying or storing computer-executable instructions and/or data structures. In particular, one or more of the processes described herein may be implemented at least in part as instructions embodied in a non-transitory computer-readable medium and executable by one or more computing devices (e.g., any of the media content access devices described herein). In general, a processor (e.g., a microprocessor) receives instructions, from a non-transitory computer-readable medium, (e.g., a memory, etc.), and executes those instructions, thereby performing one or more processes, including one or more of the processes described herein.

[0202] Computer-readable media can be any available media that can be accessed by a general purpose or special purpose computer system. Computer-readable media that store computer-executable instructions are non-transitory computer-readable storage media (devices). Computer-readable media that carry computer-executable instructions are transmission media. Thus, by way of example, and not limitation, embodiments of the disclosure can comprise at least two distinctly different kinds of computer-readable media: non-transitory computer-readable storage media (devices) and transmission media.

[0203] Non-transitory computer-readable storage media (devices) includes RAM, ROM, EEPROM, CD-ROM, solid state drives (SSDs) (e.g., based on RAM), Flash memory, phase-change memory (PCM), other types of memory, other optical disk storage, magnetic disk storage or other magnetic storage devices, or any other medium which can be used to store desired program code means in the form of computer-executable instructions or data structures and which can be accessed by a general purpose or special purpose computer.

[0204] A “network” is defined as one or more data links that enable the transport of electronic data between computer systems and/or modules and/or other electronic devices. When information is transferred or provided over a network or another communications connection (either hardwired, wireless, or a combination of hardwired or wireless) to a computer, the computer properly views the connection as a transmission medium. Transmissions media can include a network and/or data links which can be used to carry desired program code means in the form of computer-executable instructions or data structures and which can be accessed by a general purpose or special purpose computer. Combinations of the above should also be included within the scope of computer-readable media.

[0205] Further, upon reaching various computer system components, program code means in the form of computer-executable instructions or data structures can be transferred automatically from transmission media to non-transitory computer-readable storage media (devices) (or vice versa). For example, computer-executable instructions or data structures received over a network or data link can be buffered in RAM within a network interface module (e.g., a NIC), and then eventually transferred to computer system RAM and/or to less volatile computer storage media

(devices) at a computer system. Thus, it should be understood that non-transitory computer-readable storage media (devices) can be included in computer system components that also (or even primarily) utilize transmission media.

[0206] Computer-executable instructions comprise, for example, instructions and data which, when executed at a processor, cause a general-purpose computer, special purpose computer, or special purpose processing device to perform a certain function or group of functions. In some embodiments, computer-executable instructions are executed on a general-purpose computer to turn the general-purpose computer into a special purpose computer implementing elements of the disclosure. The computer executable instructions may be, for example, binaries, intermediate format instructions such as assembly language, or even source code. Although the subject matter has been described in language specific to structural features and/or methodological acts, it is to be understood that the subject matter defined in the appended claims is not necessarily limited to the described features or acts described above. Rather, the described features and acts are disclosed as example forms of implementing the claims.

[0207] Those skilled in the art will appreciate that the disclosure may be practiced in network computing environments with many types of computer system configurations, including, personal computers, desktop computers, laptop computers, message processors, hand-held devices, multi-processor systems, microprocessor-based or programmable consumer electronics, network PCs, minicomputers, mainframe computers, mobile telephones, PDAs, tablets, pagers, routers, switches, and the like. The disclosure may also be practiced in distributed system environments where local and remote computer systems, which are linked (either by hardwired data links, wireless data links, or by a combination of hardwired and wireless data links) through a network, both perform tasks. In a distributed system environment, program modules may be located in both local and remote memory storage devices.

[0208] Embodiments of the present disclosure can also be implemented in cloud computing environments. In this description, “cloud computing” is defined as a model for enabling on-demand network access to a shared pool of configurable computing resources. For example, cloud computing can be employed in the marketplace to offer ubiquitous and convenient on-demand access to the shared pool of configurable computing resources. The shared pool of configurable computing resources can be rapidly provisioned via virtualization and released with low management effort or service provider interaction, and then scaled accordingly.

[0209] A cloud-computing model can be composed of various characteristics such as, for example, on-demand self-service, broad network access, resource pooling, rapid elasticity, measured service, and so forth. A cloud-computing model can also expose various service models, such as, for example, Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure

as a Service (IaaS). A cloud-computing model can also be deployed using different deployment models such as private cloud, community cloud, public cloud, hybrid cloud, and so forth. In this description and in the claims, a “cloud-computing environment” is an environment in which cloud computing is employed.

[0210] FIG. 14 illustrates a block diagram of a computing device 1400 that may be configured to perform one or more of the processes described above. One will appreciate that one or more computing devices such as the computing device 1400 may implement the sequence-to-coverage system 106. As shown by FIG. 14, the computing device 1400 can comprise a processor 1402, a memory 1404, a storage device 1406, an I/O interface 1408, and a communication interface 1410, which may be communicatively coupled by way of a communication infrastructure 1412. In certain embodiments, the computing device 1400 can include fewer or more components than those shown in FIG. 14. The following paragraphs describe components of the computing device 1400 shown in FIG. 14 in additional detail.

[0211] In one or more embodiments, the processor 1402 includes hardware for executing instructions, such as those making up a computer program. As an example, and not by way of limitation, to execute instructions for dynamically modifying workflows, the processor 1402 may retrieve (or fetch) the instructions from an internal register, an internal cache, the memory 1404, or the storage device 1406 and decode and execute them. The memory 1404 may be a volatile or non-volatile memory used for storing data, metadata, and programs for execution by the processor(s). The storage device 1406 includes storage, such as a hard disk, flash disk drive, or other digital storage device, for storing data or instructions for performing the methods described herein.

[0212] The I/O interface 1408 allows a user to provide input to, receive output from, and otherwise transfer data to and receive data from computing device 1400. The I/O interface 1408 may include a mouse, a keypad or a keyboard, a touch screen, a camera, an optical scanner, network interface, modem, other known I/O devices or a combination of such I/O interfaces. The I/O interface 1408 may include one or more devices for presenting output to a user, including, but not limited to, a graphics engine, a display (e.g., a display screen), one or more output drivers (e.g., display drivers), one or more audio speakers, and one or more audio drivers. In certain embodiments, the I/O interface 1408 is configured to provide graphical data to a display for presentation to a user. The graphical data may be representative of one or more graphical user interfaces and/or any other graphical content as may serve a particular implementation.

[0213] The communication interface 1410 can include hardware, software, or both. In any event, the communication interface 1410 can provide one or more interfaces for communication (such as, for example, packet-based communication) between the computing device 1400 and one or more other computing devices or networks. As an example, and not by way of limitation, the

communication interface 1410 may include a network interface controller (NIC) or network adapter for communicating with an Ethernet or other wire-based network or a wireless NIC (WNIC) or wireless adapter for communicating with a wireless network, such as a WI-FI.

[0214] Additionally, the communication interface 1410 may facilitate communications with various types of wired or wireless networks. The communication interface 1410 may also facilitate communications using various communication protocols. The communication infrastructure 1412 may also include hardware, software, or both that couples components of the computing device 1400 to each other. For example, the communication interface 1410 may use one or more networks and/or protocols to enable a plurality of computing devices connected by a particular infrastructure to communicate with each other to perform one or more aspects of the processes described herein. To illustrate, the sequencing process can allow a plurality of devices (e.g., a client device, sequencing device, and server device(s)) to exchange information such as sequencing data and error notifications.

[0215] In the foregoing specification, the present disclosure has been described with reference to specific exemplary embodiments thereof. Various embodiments and aspects of the present disclosure(s) are described with reference to details discussed herein, and the accompanying drawings illustrate the various embodiments. The description above and drawings are illustrative of the disclosure and are not to be construed as limiting the disclosure. Numerous specific details are described to provide a thorough understanding of various embodiments of the present disclosure.

[0216] The present disclosure may be embodied in other specific forms without departing from its spirit or essential characteristics. The described embodiments are to be considered in all respects only as illustrative and not restrictive. For example, the methods described herein may be performed with less or more steps/acts or the steps/acts may be performed in differing orders. Additionally, the steps/acts described herein may be repeated or performed in parallel with one another or in parallel with different instances of the same or similar steps/acts. The scope of the present application is, therefore, indicated by the appended claims rather than by the foregoing description. All changes that come within the meaning and range of equivalency of the claims are to be embraced within their scope.

CLAIMS

We Claim:

1. A system comprising:
an imaging system;
a fluidic system; and
a computing engine comprising:
at least one processor; and
a non-transitory computer readable medium comprising instructions that, when executed by the at least one processor, cause the system to:
determine, from a subset of sequencing cycles of a sequencing run for genomic samples, base calls for indexing sequences within clusters of oligonucleotides;
determine, based on the indexing sequences, respective numbers of clusters of oligonucleotides belonging to respective genomic samples of the genomic samples;
estimate read-coverage levels for the genomic samples based on the respective numbers of clusters of oligonucleotides belonging to respective genomic samples and a currently selected number of sequencing cycles for the sequencing run;
generate, for the sequencing run and based on the estimated read-coverage levels, a customized number of sequencing cycles sufficient to generate nucleotide reads satisfying a target read-coverage level for each genomic sample of the genomic samples; and
execute the sequencing run until finishing the customized number of sequencing cycles.
2. The system of claim 1, further comprising instructions that, when executed by the at least one processor, cause the system to estimate the read-coverage levels by:
determining filter metrics indicating subsets of clusters of oligonucleotides satisfying a filtering threshold for signals of the clusters of oligonucleotides; and
estimating the read-coverage levels for the genomic samples based on the filter metrics and the respective numbers of clusters of oligonucleotides belonging to respective genomic samples.
3. The system of claim 2, further comprising instructions that, when executed by the at least one processor, cause the system to determine the filter metrics by determining, in a pass filter map, a percentage of clusters belonging to each genomic sample that satisfy a chastity filter for signals emitted from the clusters of oligonucleotides.

4. The system of claim 2, further comprising instructions that, when executed by the at least one processor, cause the system to estimate the read-coverage levels for the genomic samples by:

determining, based on the filter metrics and the respective numbers of clusters of oligonucleotides belonging to respective genomic samples, a number of filter-passing clusters of oligonucleotides for each genomic sample of the genomic samples that satisfy the filtering threshold; and

estimating a minimum number of nucleotide reads covering genomic regions of each genomic sample based on the number of filter-passing clusters of oligonucleotides.

5. The system of claim 1, further comprising instructions that, when executed by the at least one processor, cause the system to:

determine, based on the estimated read-coverage levels, a customized set of flow cell regions to be imaged from a flow cell sufficient to generate the nucleotide reads satisfying the target read-coverage level for each genomic sample of the genomic samples; and

execute the sequencing run by capturing images of the customized set of flow cell regions for the customized number of sequencing cycles using the imaging system.

6. The system of claim 1, further comprising instructions that, when executed by the at least one processor, cause the system to perform the subset of sequencing cycles according to an order of indexing cycles before genomic sequencing cycles by:

determining base calls for a first indexing sequence appended to a sample genomic sequence of a genomic sample;

determining base calls for a second indexing sequence appended to the sample genomic sequence of the genomic sample; and

after determining the base calls for the first indexing sequence and the second indexing sequence, determining base calls for a first nucleotide read corresponding to a first portion of the sample genomic sequence and determining base calls for a second nucleotide read corresponding to a second portion of the sample genomic sequence.

7. The system of claim 1, further comprising instructions that, when executed by the at least one processor, cause the system to determine the respective numbers of clusters of oligonucleotides belonging to the respective genomic samples by:

identifying, from among the indexing sequences, assigned indexing sequences matching indexing sequences registered for the sequencing run and unassigned indexing sequences that do not match the indexing sequences registered for the sequencing run;

removing, from data for the sequencing run, a subset of clusters of oligonucleotides corresponding to the unassigned indexing sequences;

determining respective subsets of assigned indexing sequences that correspond to the respective genomic samples; and

determining, from among the respective subsets of assigned indexing sequences, a number of clusters of oligonucleotides belonging to each genomic sample.

8. The system of claim 1, further comprising instructions that, when executed by the at least one processor, cause the system to generate the customized number of sequencing cycles for the sequencing run by increasing or decreasing a preset number of sequencing cycles for the sequencing run.

9. The system of claim 1, further comprising instructions that, when executed by the at least one processor, cause the system to generate the customized number of sequencing cycles for the sequencing run by:

identifying a minimum number of sequencing cycles and a maximum number of sequencing cycles for the sequencing run; and

increasing or decreasing a preset number of sequencing cycles for the sequencing run to the customized number of sequencing cycles within the minimum number of sequencing cycles and the maximum number of sequencing cycles.

10. The system of claim 1, further comprising instructions that, when executed by the at least one processor, cause the system to estimate the read-coverage levels by:

determining, from the sequencing run, a number of unique nucleotide reads aligned with a reference genome;

determining, from the sequencing run, a number of filter-passing nucleotide reads from filter-passing cluster of oligonucleotides with signals that satisfy a filtering threshold;

determining a bioinformatics efficiency metric by dividing the number of unique nucleotide reads by the number of filter-passing nucleotide reads; and

estimating the read-coverage levels for the genomic samples based on the bioinformatics efficiency metric and the respective numbers of clusters of oligonucleotides belonging to respective genomic samples.

11. The system of claim 1, further comprising instructions that, when executed by the at least one processor, cause the system to detect a reagent volume of a reagent cartridge in fluid communication with the fluidic system and operate the fluidic system to perform one or more additional sequencing cycles relative to the currently selected number of sequencing cycles until finishing the customized number of sequencing cycles by aspirating one or more reagents from the reagent cartridge.

12. The system of claim 1, further comprising instructions that, when executed by the at least one processor, cause the system to terminate operation of the fluidic system from performing one or more sequencing cycles of the currently selected number of sequencing cycles to finish the sequencing run after performing the customized number of sequencing cycles.

13. A system comprising:

an imaging system;

a fluidic system and

a computing engine comprising:

at least one processor; and

a non-transitory computer readable medium comprising instructions that, when executed by the at least one processor, cause the system to:

determine, from a subset of sequencing cycles of a sequencing run for genomic samples, base calls for indexing sequences within clusters of oligonucleotides;

determine, based on the indexing sequences, respective numbers of clusters of oligonucleotides belonging to respective genomic samples of the genomic samples;

estimate read-coverage levels for the genomic samples based on the respective numbers of clusters of oligonucleotides belonging to respective genomic samples;

determine, from a flow cell and based on the estimated read-coverage level, a customized set of flow cell regions to be imaged sufficient to generate nucleotide reads satisfying a target read-coverage level for each genomic sample of the genomic samples; and

execute the sequencing run by capturing images of the customized set of flow cell regions during sequencing cycles of the sequencing run.

14. The system of claim 13, further comprising instructions that, when executed by the at least one processor, cause the system to determine the customized set of flow cell regions by determining a customized number of flow cell regions to be imaged sufficient to generate the nucleotide reads satisfying the target read-coverage level for each genomic sample.

15. The system of claim 13, further comprising instructions that, when executed by the at least one processor, cause the system to determine the customized set of flow cell regions by determining, from a flow cell, a set of tiles to be imaged sufficient to generate the nucleotide reads satisfying the target read-coverage level for each genomic sample.

16. The system of claim 13, further comprising instructions that, when executed by the at least one processor, cause the system to capture the images of the customized set of flow cell regions without adjusting a currently selected number of sequencing cycles for the sequencing run.

17. The system of claim 13, further comprising instructions that, when executed by the at least one processor, cause the system to determine the customized set of flow cell regions by increasing or decreasing a number of flow cell regions from an initial set of flow cell regions selected for the sequencing run.

18. The system of claim 13, further comprising instructions that, when executed by the at least one processor, cause the system to estimate the read-coverage levels by:

determining filter metrics indicating subsets of clusters of oligonucleotides satisfying a filtering threshold for signals of the clusters of oligonucleotides; and

estimating the read-coverage levels for the genomic samples based on the filter metrics and the respective numbers of clusters of oligonucleotides belonging to respective genomic samples.

19. The system of claim 18, further comprising instructions that, when executed by the at least one processor, cause the system to determine the filter metrics by determining, in a pass filter map, a percentage of clusters belonging to each genomic sample that satisfy a chastity filter for signals emitted from the clusters of oligonucleotides.

20. The system of claim 18, further comprising instructions that, when executed by the at least one processor, cause the system to estimate the read-coverage levels for the genomic samples by:

determining, based on the filter metrics and the respective numbers of clusters of oligonucleotides belonging to respective genomic samples, a number of filter-passing clusters of oligonucleotides for each genomic sample of the genomic samples that satisfy the filtering threshold; and

estimating a minimum number of nucleotide reads covering genomic regions of each genomic sample based on the number of filter-passing clusters of oligonucleotides.

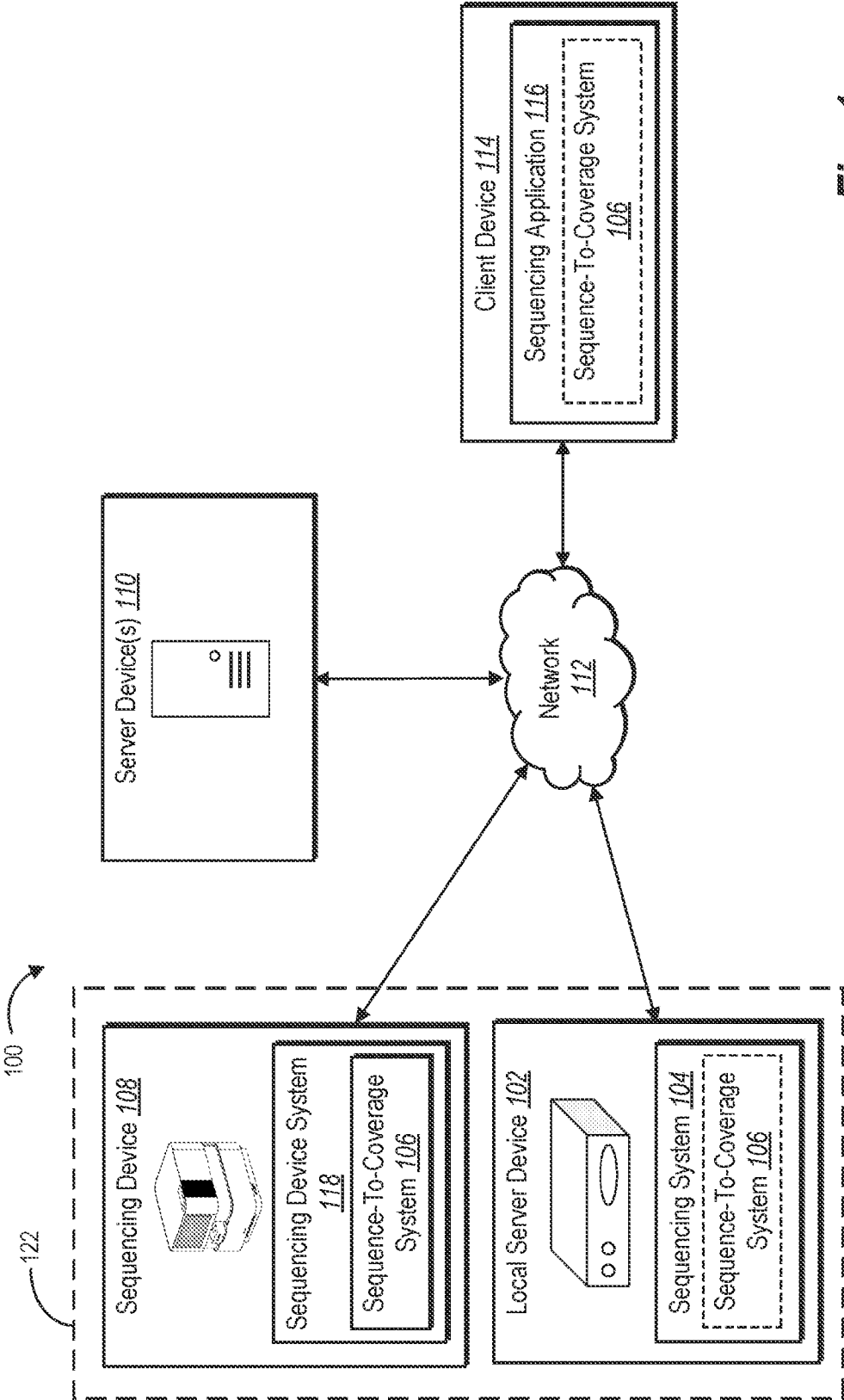


Fig. 1

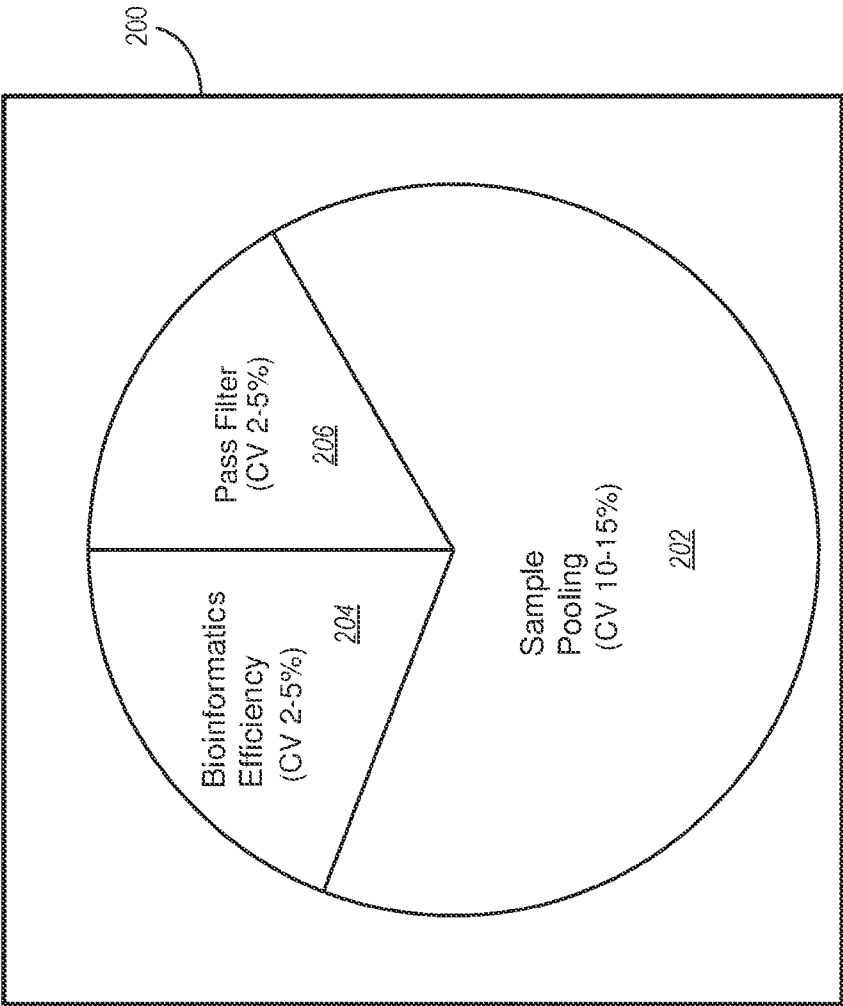
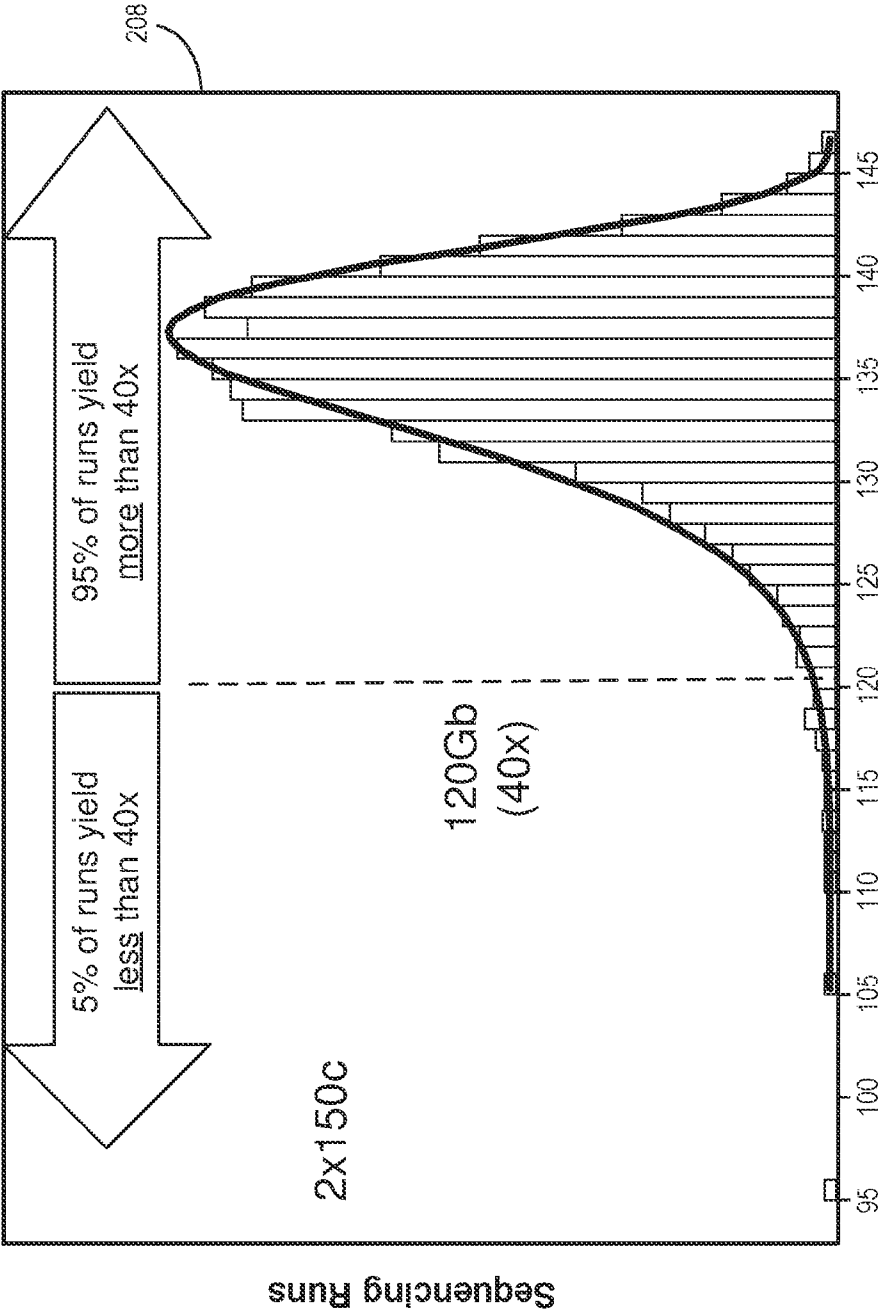


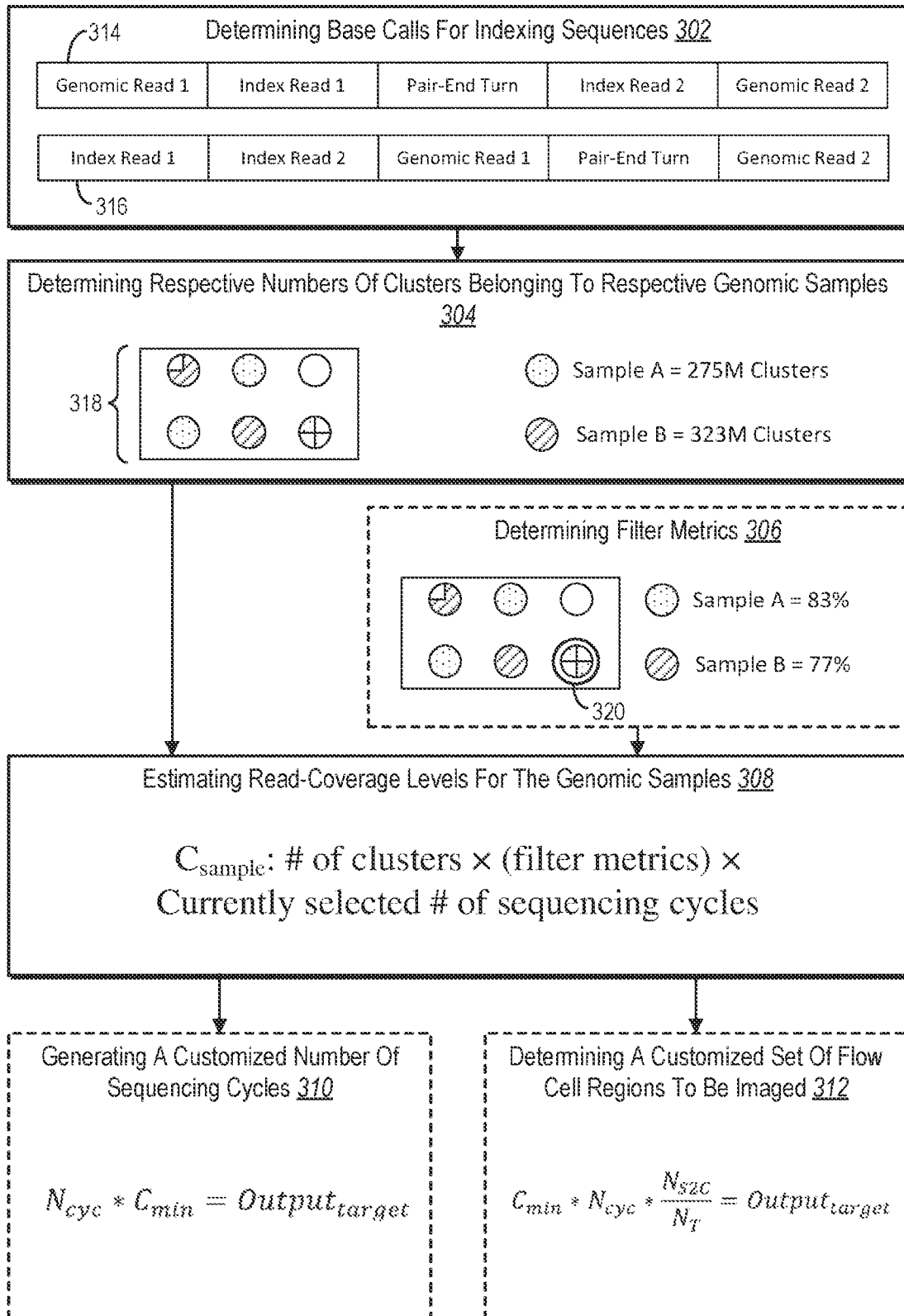
Fig. 2A



Unique Aligned Reads in Gb For Worst Performing Sample

Fig. 2B

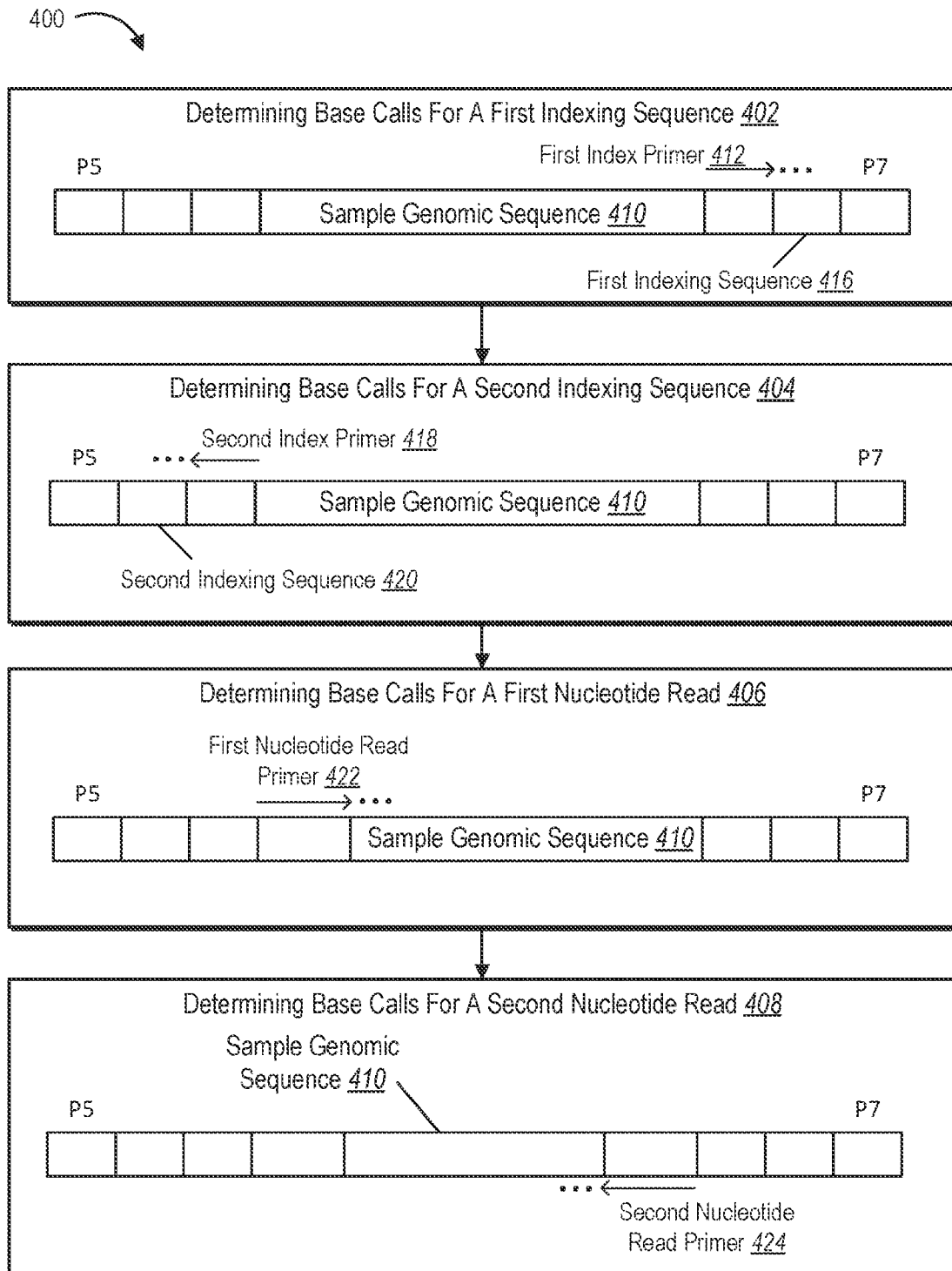
4 / 18



300

Fig. 3

5 / 18

**Fig. 4**

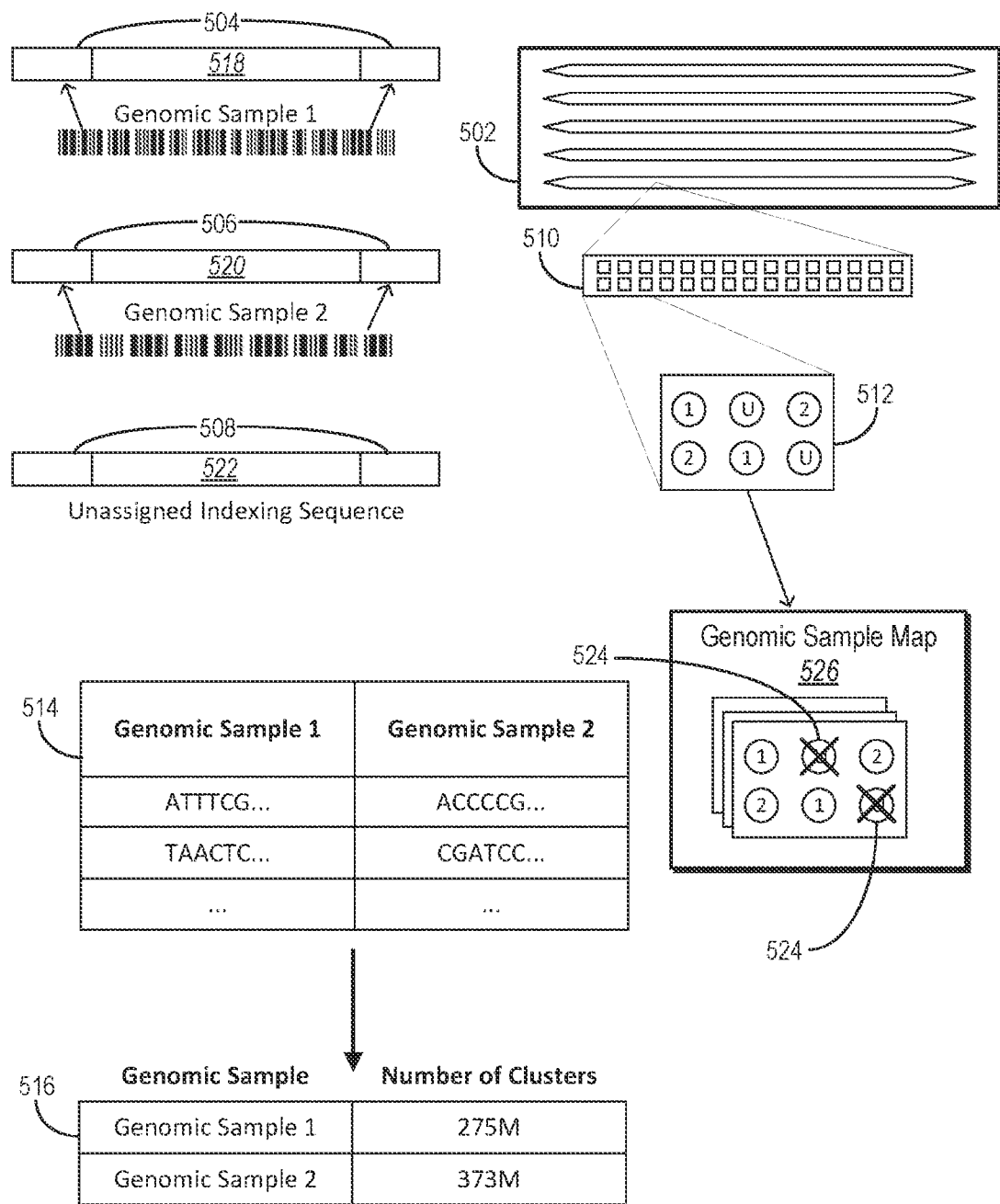
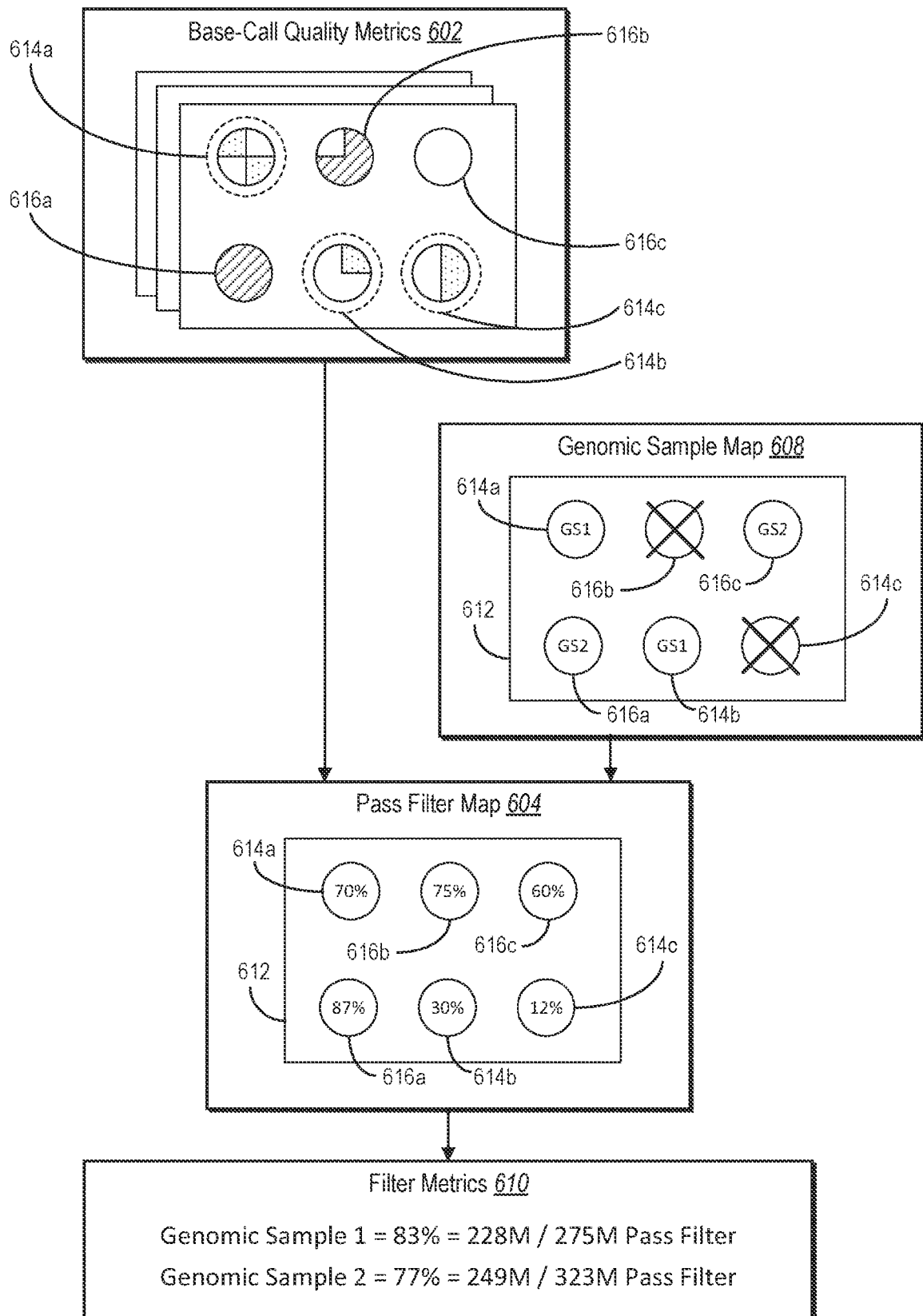
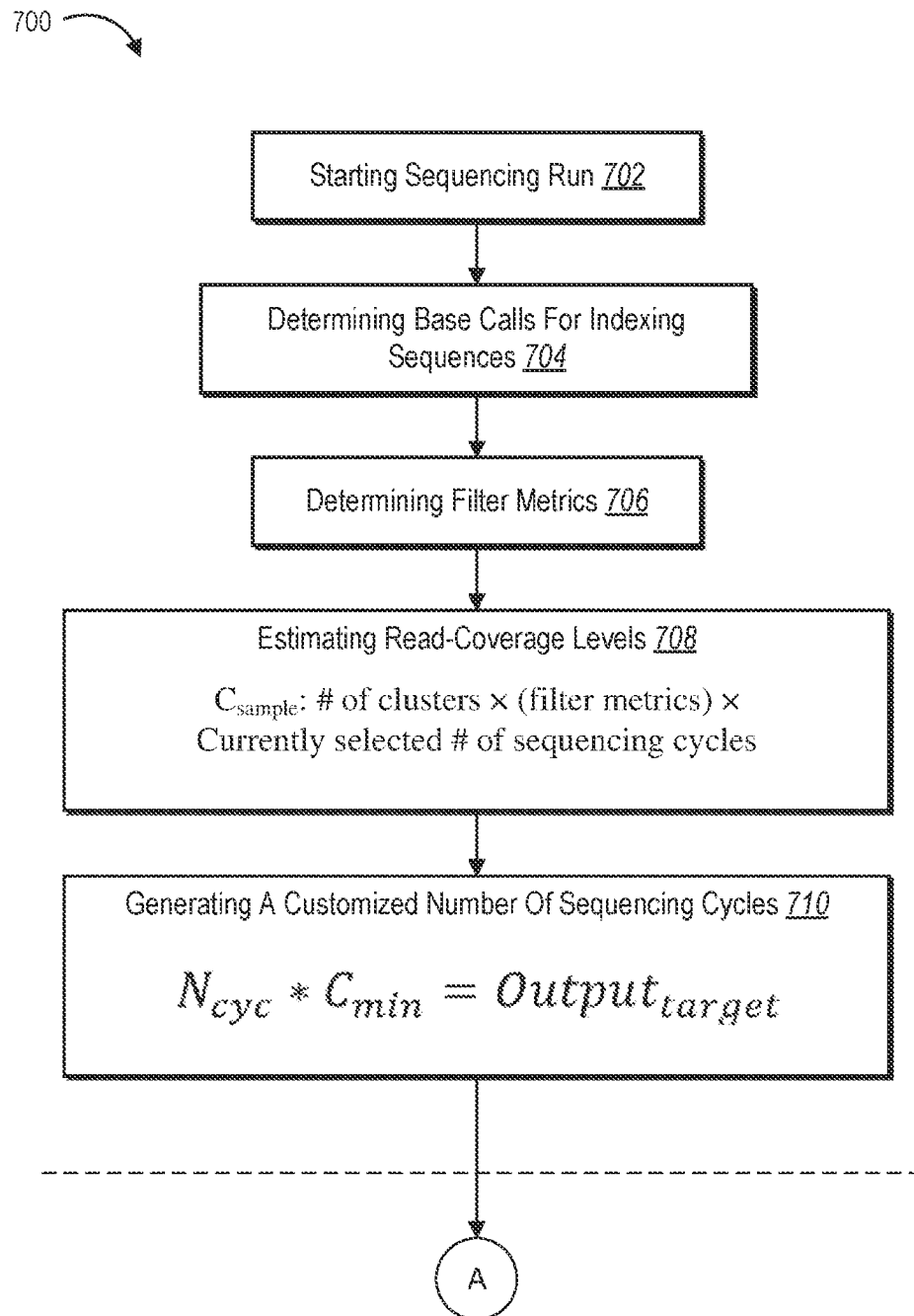


Fig. 5

7 / 18

**Fig. 6**

8 / 18

**Fig. 7A**

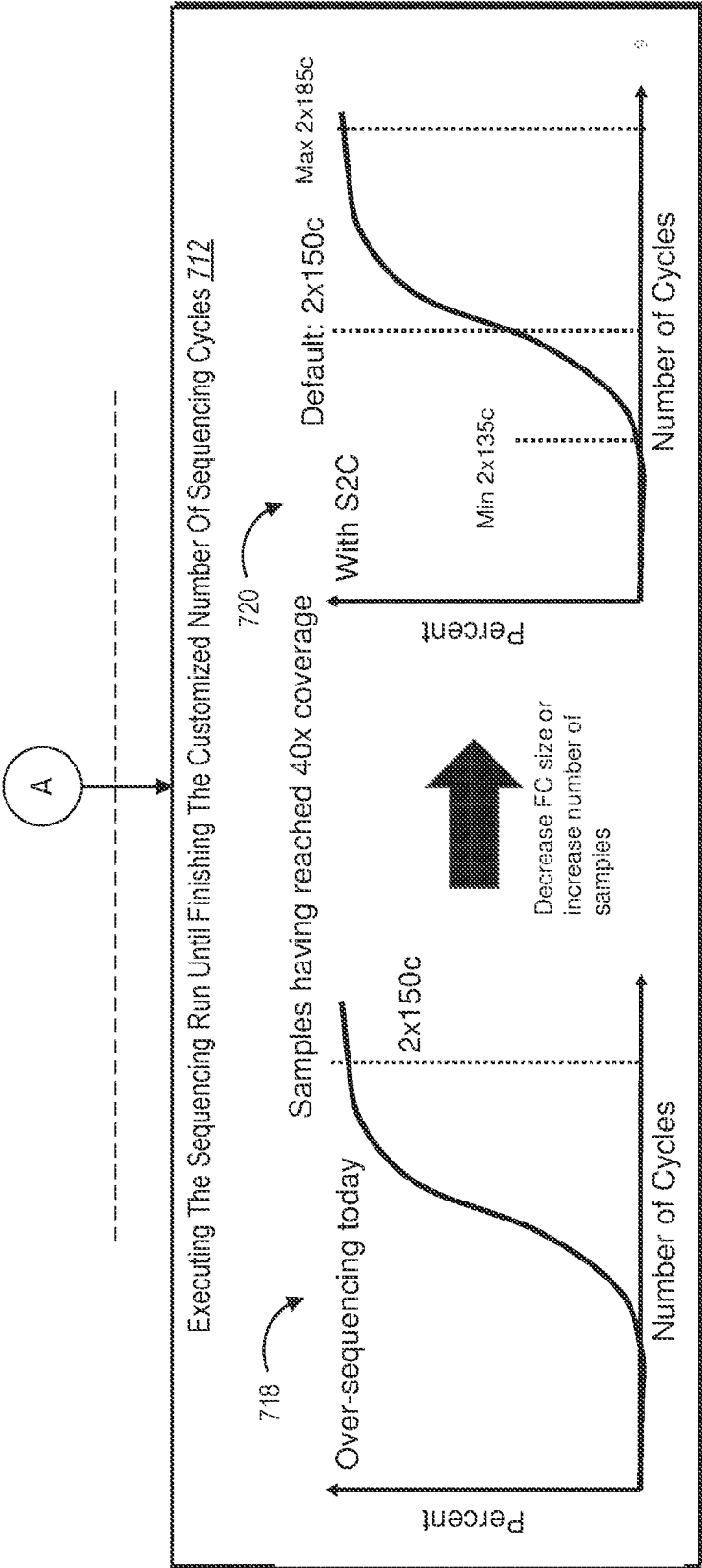
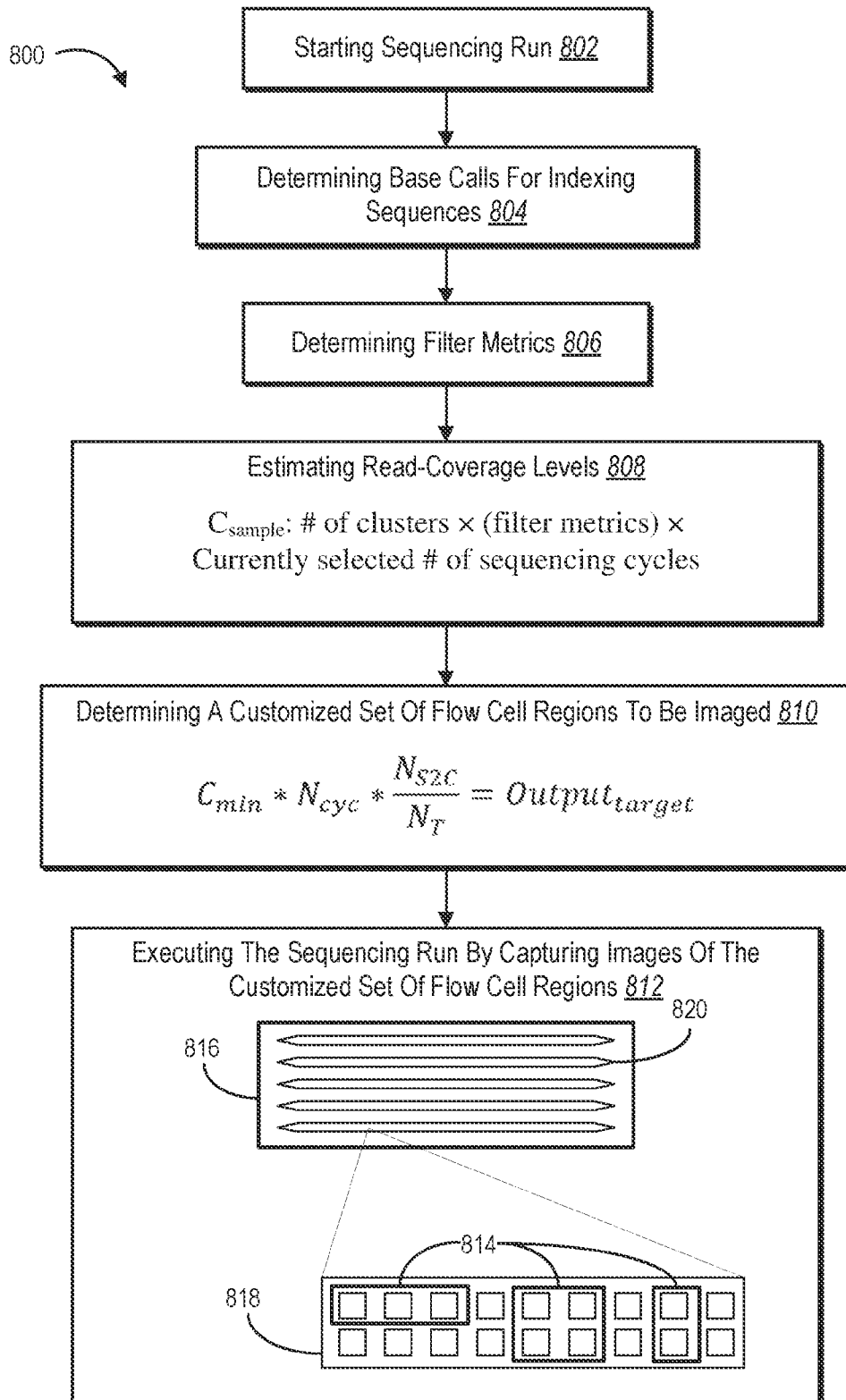


Fig. 7B

10 / 18

**Fig. 8**

Poor Sample Pooling

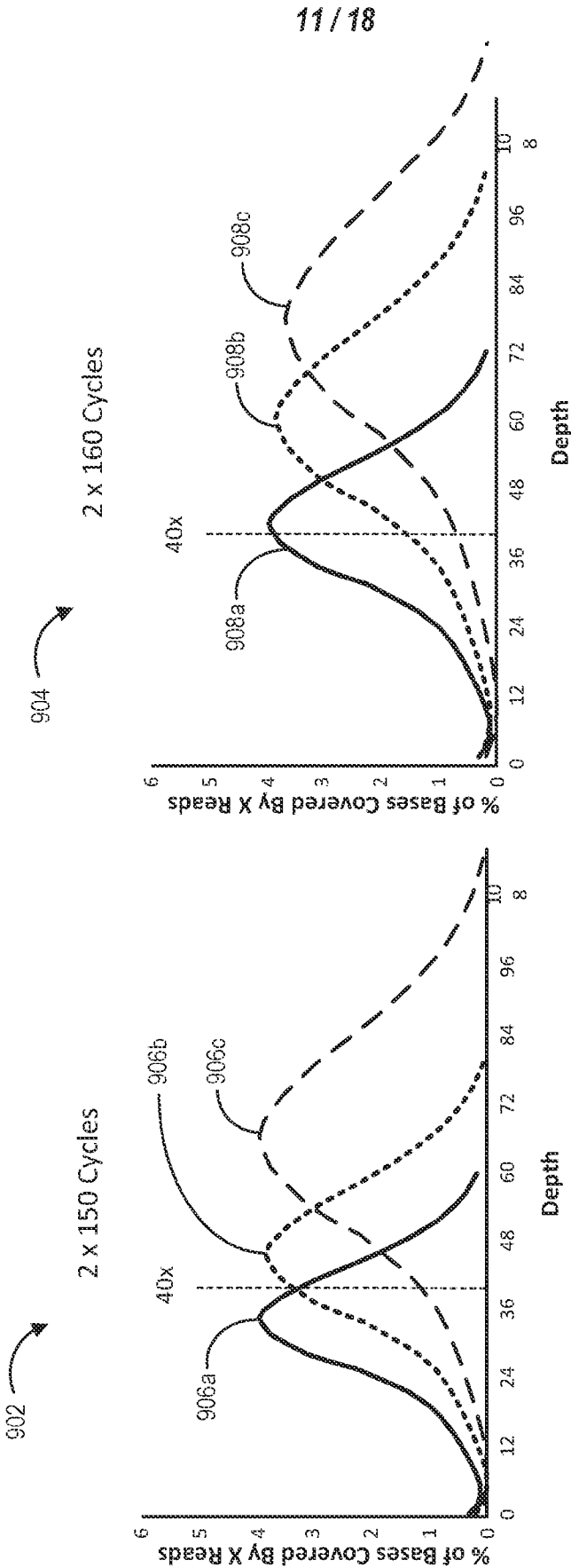


Fig. 9A

Improved Sequence
Performance

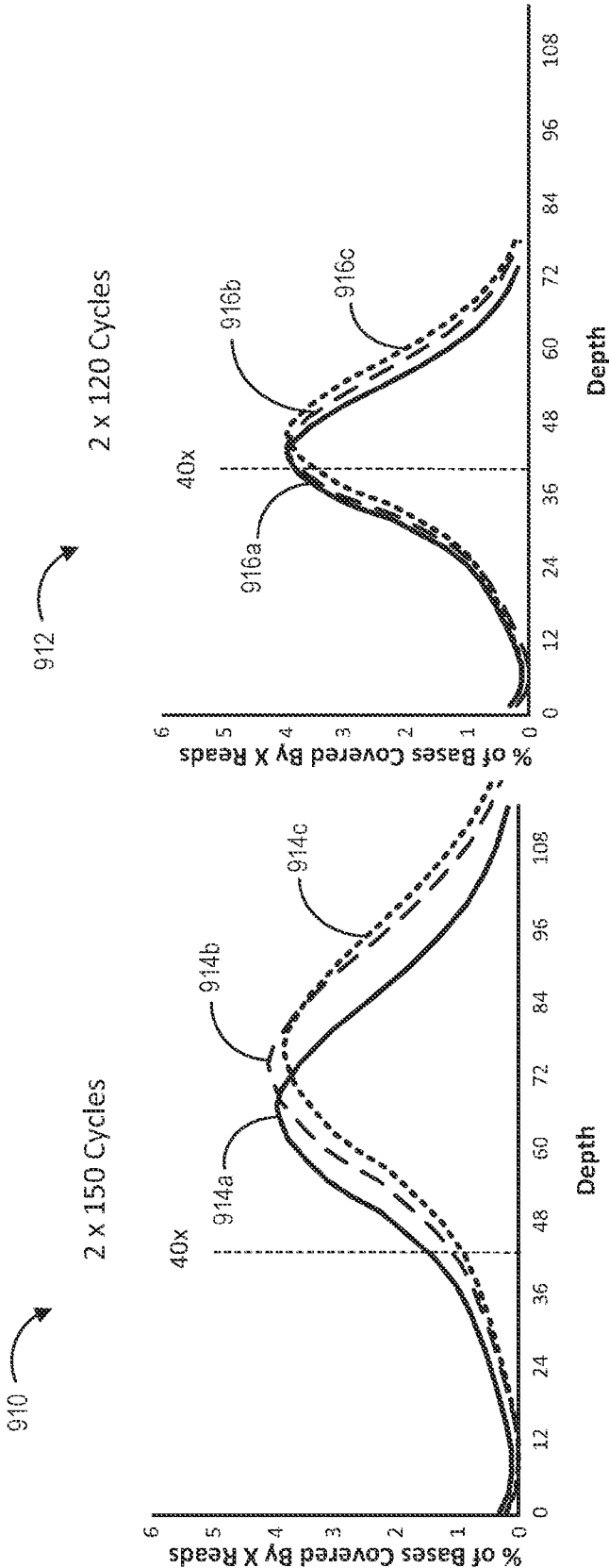


Fig. 9B

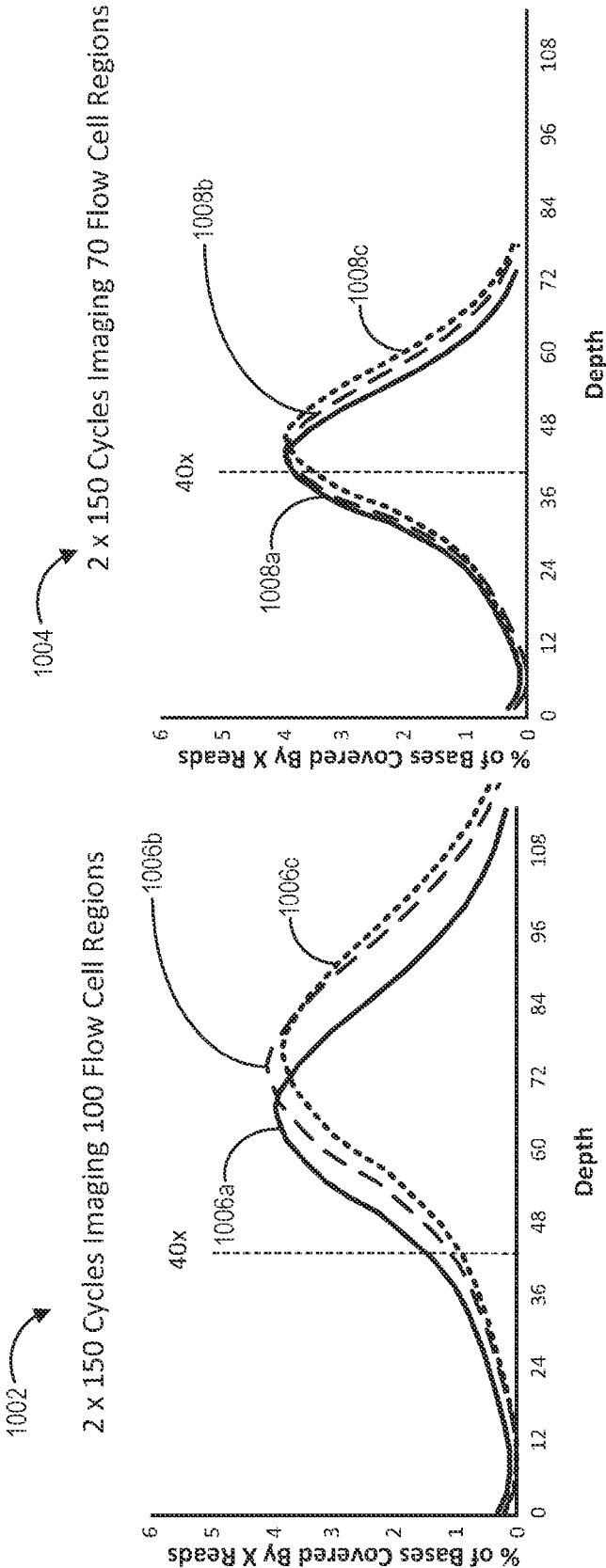


Fig. 10

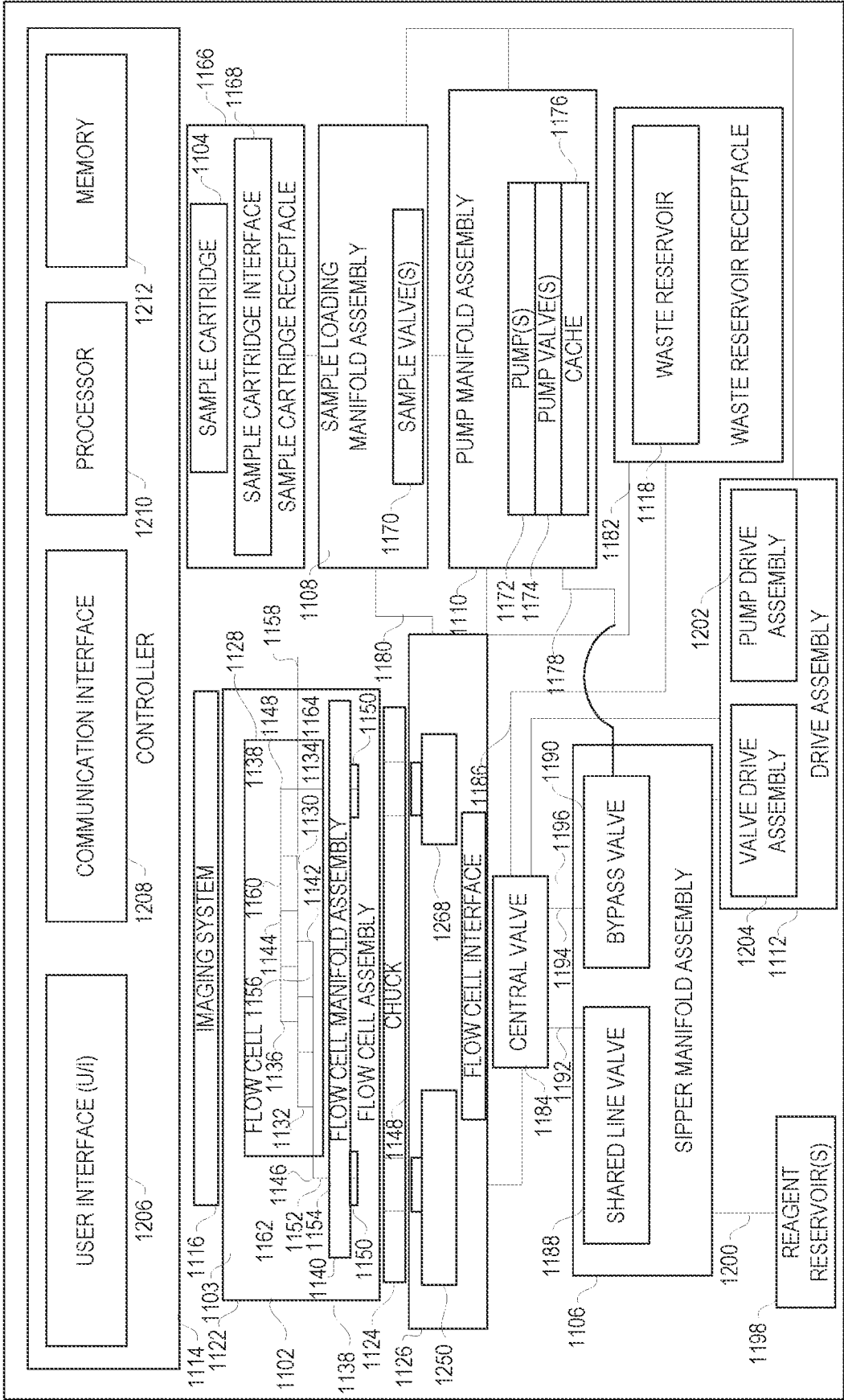


Fig. 11

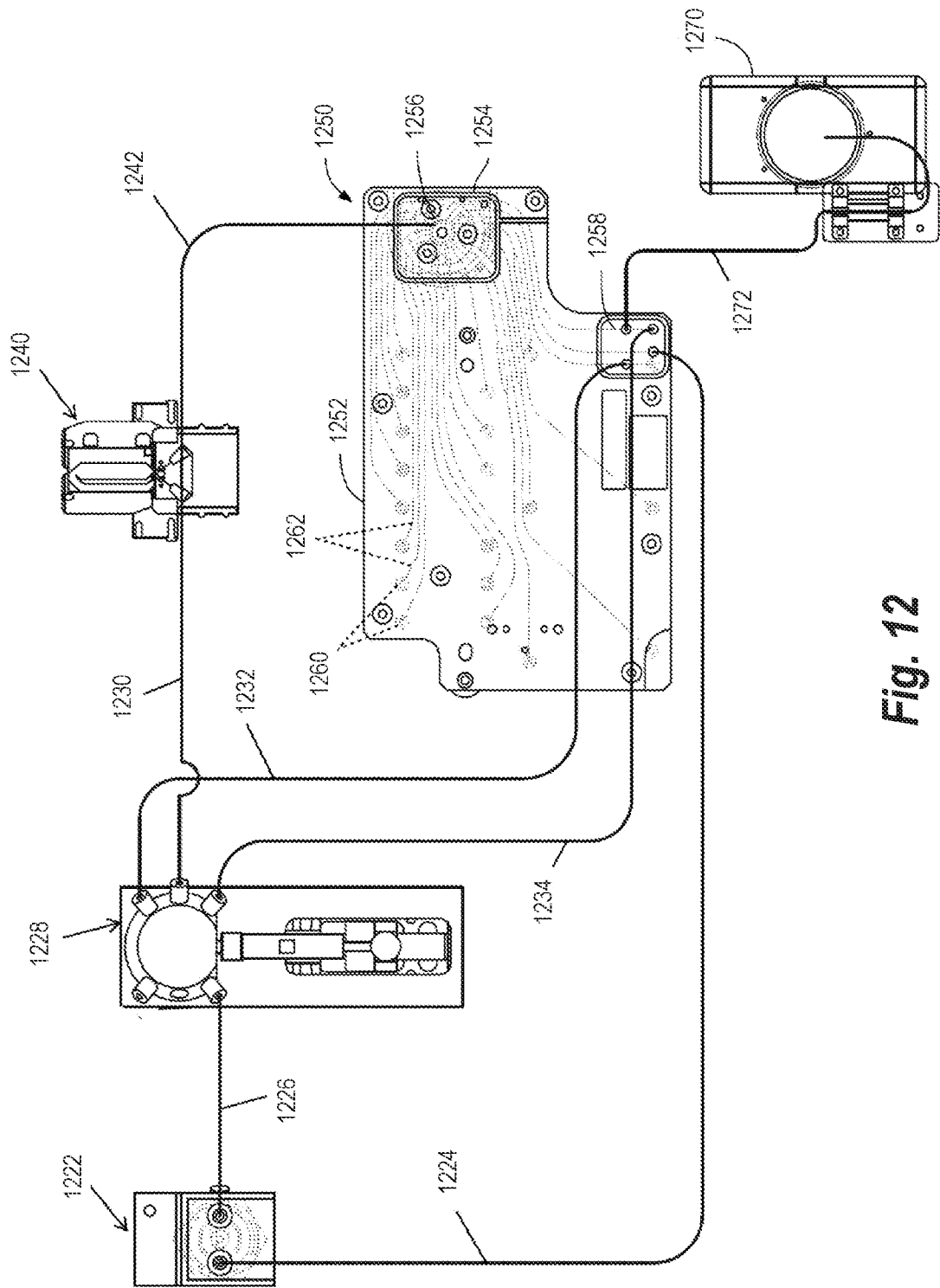
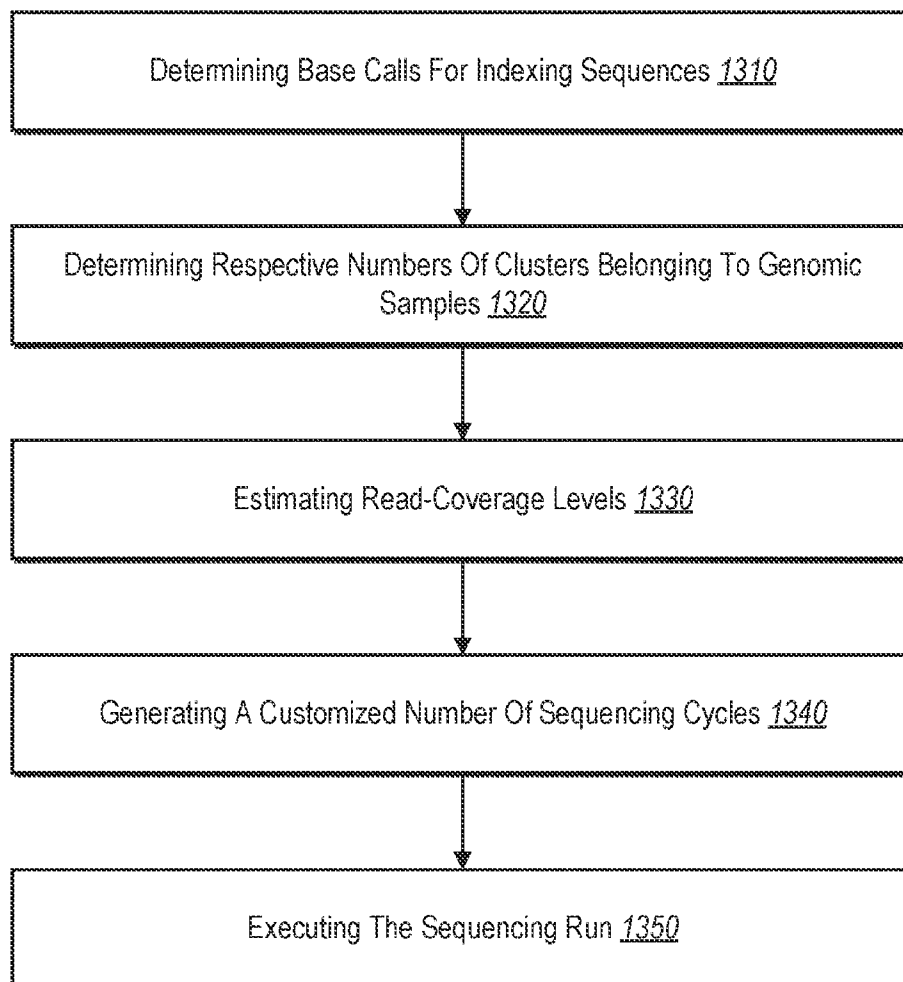
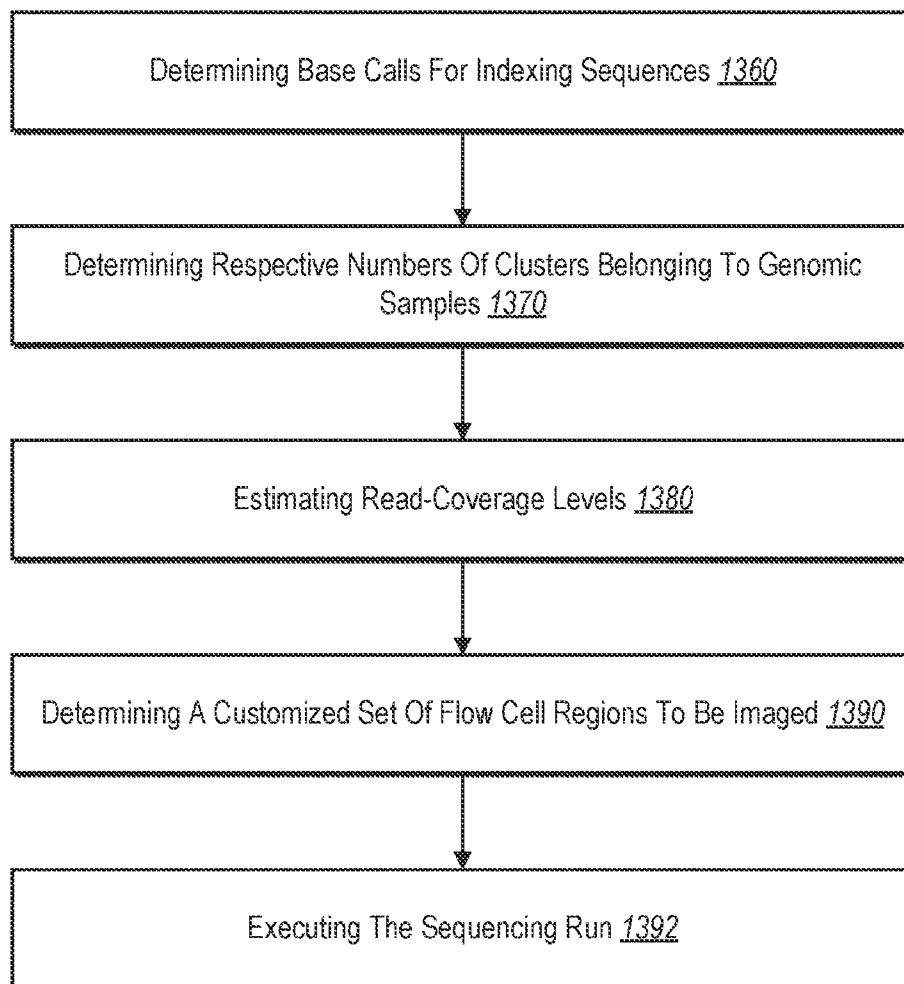


Fig. 12

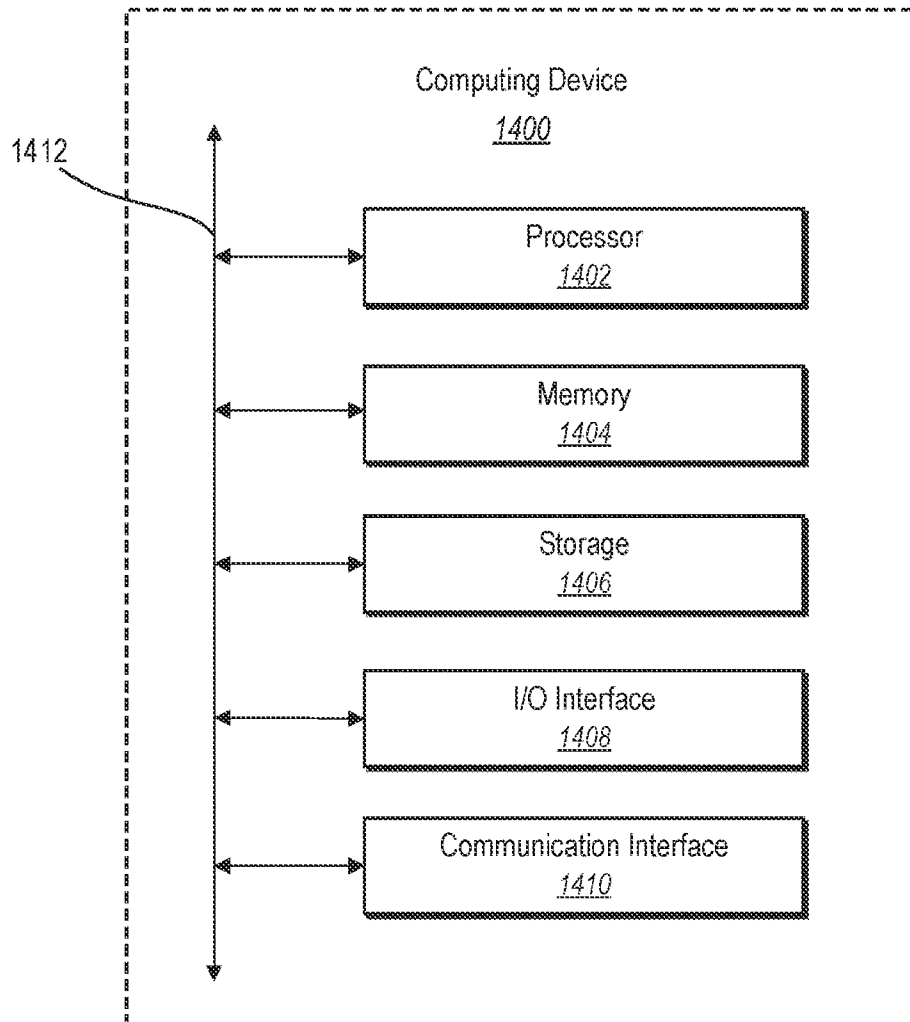
16 / 18

1300 **Fig. 13A**

17 / 18

1362 **Fig. 13B**

18 / 18

**Fig. 14**