SYSTEM, METHOD, AND APPARATUS FOR COPY-NUMBER ESTIMATION OF GENETIC SEQUENCES FOR GENERATION OF RECOMBINANT PROTEINS FOR USE IN THERAPEUTIC APPLICATIONS

CROSS-REFERENCE TO RELATED APPLICATION

[0001] This application claims the benefit of and priority to US Provisional Patent Application having serial number 63/551,348 filed on February 8, 2024. The entire contents of which is incorporated herein by reference.

BACKGROUND

Relevant Field

[0002] The present disclosure relates to copy-number estimation of a genetic sequence. More particularly, the present disclosure relates to a system, method and apparatus for copy-number estimation of a genetic sequences for generation of recombinant proteins for use in therapeutic applications.

Description of Related Art

[0003] The production of recombinant proteins, especially for therapeutic use, has heavily relied on Chinese hamster ovary (CHO) cells. These cells are preferred for their ability to execute complex protein modifications after translation. However, the long-term cultivation of CHO cells presents certain challenges. A concern is the reduction in recombinant protein expression levels. There can also be changes in the proteins' primary structure, impacting their functionality and safety. One issue is the decrease in expression levels of the recombinant proteins. Additionally, there can be alterations in the primary structure of these proteins, which can have profound effects on their functionality and safety. Thus, characterizing genetic translation sequences is typically done given their therapeutic use in the generation of recombination proteins from transgenes.

[0004] Given the therapeutic nature of these proteins, the regulatory submission process for these therapeutic proteins involves a safety assessment of CHO-derived recombinant clones. An aspect of this assessment is the molecular characterization at the DNA level. This characterization process includes determining the copy number, integrity, and stability of the transgene, as well as elucidating the nature of its integration within the host genome.

[0005] Historically, the process of molecular characterization of these clones has primarily relied on Southern blot analysis in conjunction with Sanger sequencing. More contemporary methods have included quantitative PCR (qPCR), and more recently, digital droplet PCR (ddPCR). While these methods are robust and have been the mainstay in clone characterization, they are not without limitations. They tend to be both time-intensive and resource-consuming, which can be significant drawbacks in a field where efficiency and cost-effectiveness are increasingly valued.

[0006] The advent of next-generation sequencing (NGS) technologies marked a significant advancement in the field of molecular characterization. NGS offers a highly sensitive alternative to traditional methods like Southern blot analysis, and it is both more cost-effective and labor-efficient. However, the application of NGS in this context is not without its challenges. The current NGS methods, which were primarily developed for the characterization of genetically modified crops, generally depend on PCR-based library preparation workflows. These methods, along with the simplistic approaches used for copy number estimation, can lead to inaccuracies. Specifically, they may cause a skewed determination of transgene copy number, necessitating the employment of additional ddPCR methodologies for precise estimation.

SUMMARY

[0007] Embodiments disclosed herein include using a single PCR-free whole genome NGS library preparation and sequencing workflow with advanced bioinformatics analysis. This new method compares data against a reference sequence, enabling accurate determination of copy number, variant calling, detection of vector structural variants, and annotation of vector integration. This approach represents a significant advancement over existing methods, providing more accurate and efficient characterization of CHO-derived recombinant clones.

[0008] Different known techniques may be utilized, such as an NGS method known as Targeted Locus Amplification (TLA) and long-read sequencing techniques, notably those based on Oxford Nanopore technology. Some embodiments of the present disclosure can build upon and enhance these existing methodologies, offering a more comprehensive and efficient solution for the molecular characterization of CHO-derived recombinant clones.

[0009] An embodiment of the method of determining genomic-region copy number involves several steps. The method may sample of genomic nucleic acids is isolated and purified. Then, a shotgun library is created from the sample of genomic nucleic

acids. The shotgun library is sequenced, generating reads. These reads are aligned to a reference sequence, which may include a reference genome and a vector sequence, using a processor. The reference sequence may also include one or more auxiliary sequences such as Escherichia coli genome sequence. The reference genome is partitioned into multiple sizes using the processor.

[0010] The processor determines the mean read coverage for each size of the partitioned reference genome. This determination continues until a predetermined number of consecutive mean-read coverage peaks are formed, with an equal coverage distance between each peak. The processor then determines a respective copy number associated with each of these consecutive mean-read coverage peaks.

[0011] The processor may then establish a model that is configured to estimate the copy number of a target sequence. In summary, this method utilizes multiple steps, including isolation and purification of genomic nucleic acids, creation of a shotgun library, sequencing, alignment, partitioning of the reference genome, determination of mean read coverage and copy numbers, and the establishment of a model for copy number estimation.

[0012] A method of determining genomic-region copy number involves several steps. The method may include isolating and purifying a sample of genomic nucleic acids. Then, a shotgun library is created from the sample of genomic nucleic acids. The shotgun library is sequenced, generating reads. These reads are aligned to a reference sequence, which includes of a reference genome and a vector sequence, using a processor. The reference genome is partitioned into multiple sizes using the processor.

[0013] The processor determines the mean read coverage for each size of the partitioned reference genome. This determination continues until a predetermined number of consecutive mean-read coverage peaks are formed, with an equal coverage distance between each peak. The processor then determines a respective copy number associated with each of these consecutive mean-read coverage peaks.

[0014] The processor may then establish a model that is configured to estimate the copy number of a target sequence. In summary, this method utilizes multiple steps, including isolation and purification of genomic nucleic acids, creation of a shotgun library, sequencing, alignment, partitioning of the reference genome, determination of mean read coverage and copy numbers, and the establishment of a model for copy number estimation.

[0015] The method may estimate the copy number of a target sequence that is within the vector sequence. Specifically, after determining the mean read coverage and consecutive mean-read coverage peaks of the partitioned reference genome, the processor is able to establish a model to estimate the copy number of a target sequence. This target sequence refers to a genetic sequence located within the vector sequence that was used to construct the reference sequence, along with the reference genome. By analyzing the correlation between read coverage and copy number across the reference genome fragments, the model can then precisely calculate the predicted copy number of such a target sequence found within the vector based on its observed coverage from the aligned sequencing reads.

[0016] The method may involve determining the copy number of the vector sequence as the target sequence. Specifically, after isolating and purifying genomic nucleic acids from a sample, creating a shotgun library from the sample, sequencing the library to generate reads, and aligning the reads to a reference sequence including a reference genome and vector sequence, the processor partitions the reference genome into multiple sizes. The processor then determines the mean read coverage for each size of the partitioned reference genome until a predetermined number of consecutive mean-read coverage peaks are formed with an equal coverage distance between each peak. Furthermore, the processor determines the respective copy number associated with each of these consecutive peaks. Finally, the processor establishes a model configured to estimate the copy number of the vector sequence, which serves as the target sequence in this case.

[0017] The method may determine the copy number of a specific predetermined gene that is located within the vector sequence. After isolating and purifying genomic nucleic acids from a sample, a shotgun library is created and sequenced to generate reads. These reads are then aligned to a reference sequence including a reference genome and the vector sequence using a processor. The processor partitions the reference genome into multiple sizes and determines the mean read coverage for each size until a number of consecutive coverage peaks are observed with an equal distance between them. From this, the processor can determine the respective copy numbers associated with each coverage peak. Finally, the processor establishes a model to estimate the copy number of a target sequence. If the target sequence is a predetermined gene within the vector, the model can specifically determine the copy number of that gene based on its observed read coverage.

[0018] Additionally, the model developed by the processor to estimate the copy number of a target sequence may be configured to perform this estimation without requiring a genomic location parameter for the target sequence. Specifically, the model does not need prior knowledge of where the target sequence is located in the reference genome in order to calculate an estimated copy number, in some embodiments. Rather, the model may rely solely on the observed read coverage of the target sequence as determined from the aligned sequencing reads. This enables the model to provide an unbiased copy number prediction for any target sequence based only on its measured coverage, without needing to know the precise genomic location of the target sequence, in this specific embodiment.

[0019] The reference sequence used in this method may be a DNA genome of a Chinese hamster ovary (CHO) cell line. CHO cell lines are commonly used for the production of recombinant proteins through the insertion of transgenes via vectors. Determining the copy number of these inserted transgenes is used for assessing CHO cell lines being considered for therapeutic protein production. Using the CHO genome as the reference sequence allows for the method to estimate the copy number of transgenes integrated into the CHO genome through this vector-mediated insertion approach. The resultant copy number estimations provide information about the expression and stability of the transgenes in candidate CHO cell lines for biotherapeutic applications.

[0020] The processor may analyze a reference sequence that includes the genome of a mammalian cancer-cell line. Specifically, the reference sequence used in the alignment step can be formed from the DNA genome of a mammalian cancer-cell line. By utilizing the genome of a mammalian cancer-cell line as the reference sequence, this method can determine genomic-region copy numbers for cancer samples. The copy number estimation of cancer genomic regions could provide insights into alterations present in the cancer genome compared to a normal genome. This may help further the understanding of cancer biology and development of new cancer therapeutics.

[0021] The method may involve creating a shotgun library from the sample of genomic nucleic acids. To prepare the library, next-generation sequencing or NGS techniques are utilized. Specifically, the shotgun library is created using a PCR-free workflow. By employing a polymerase chain reaction or PCR-free method, this helps reduce potential biases and errors that could be introduced through PCR amplification. Thus, the act of creating the shotgun library in the method uses a PCR-

free workflow when constructing the library with next-generation sequencing or NGS techniques. Utilizing a PCR-free approach for the library preparation may aid in generating an unbiased representation of the genomic sample for subsequent analysis.

[0022] One approach for creating the shotgun library involves utilizing sequencing-by-synthesis. With this method, the shotgun library undergoes sequencing using sequencing-by-synthesis technology. Specifically, sequencing-by-synthesis is performed using reversible dye terminators, where fluorescently labeled nucleotides are added one by one during repeated cycles. Imaging is performed after each nucleotide incorporation to identify the added base and determine the sequence of bases in each DNA fragment. This cycle is repeated to generate reads representing short segments of the genetic material present in the shotgun library. Overall, sequencing-by-synthesis allows the parallel sequencing of millions of DNA fragments from the shotgun library.

[0023] One approach for sequencing the shotgun library is to utilize Illumina sequencing. Illumina sequencing is a type of high-throughput next generation sequencing (NGS) that can generate large numbers of DNA reads in a parallel manner. This sequencing method involves first preparing a shotgun library from the sample of genomic nucleic acids, as described previously. The shotgun library then undergoes Illumina sequencing, which employs a sequencing by synthesis approach to determine the sequence of nucleotide bases in the fragmented DNA pieces. This generates millions or billions of short sequencing reads representing segments of the genomic material present in the original sample. The reads obtained from Illumina sequencing are then available for downstream alignment and analysis, as discussed further in the summary.

[0024] One method that can be used in creating the shotgun library involves utilizing DNABSEQ sequencing. DNABSEQ sequencing is a tagmentation-based approach offered by Illumina. With DNABSEQ sequencing, the shotgun library undergoes an enzymatic fragmentation and tag attachment process called tagmentation. This enables simultaneous fragmentation and tagging of the DNA fragments prior to clonal amplification and sequencing. Utilizing DNABSEQ sequencing for creating the shotgun library provides an alternative to other sequencing methods that can involve PCR amplification steps which may introduce biases and errors.

[0025] The reference genome is partitioned into a plurality of sizes ranging from 100 base pairs (bp) to 51,200 bp. Specifically, the processor partitions the reference genome into fragment sizes between 100 bp up to 51,200 bp. This range of fragment sizes allows the analysis of read coverage across different lengths of the reference genome. By partitioning the reference genome into fragments of varying lengths within this size range, the processor can determine trends and relationships between the observed read coverage and the genomic copy number for the target sequence. This size range for partitioning the reference genome facilitates the robust estimation of copy numbers.

[0026] In partitioning the reference genome into multiple sizes, the processor analyzes fragment sizes ranging from 50 base pairs (bp) to 51,200 bp, including partition sizes of 50 bp, 100 bp, 200 bp, 400 bp, 800 bp, 1,600 bp, 3,200 bp, 6,400 bp, 12,800 bp, 25,600 bp, and 51,200 bp. By determining the mean read coverage for each of these partition sizes, the processor can observe patterns in the coverage values across the varying fragment sizes. This adaptive partitioning allows the processor to tune the fragment size to optimize peak resolution when analyzing the distribution of read coverages.

[0027] The method partitions the reference genome into a plurality of sizes ranging from 50 base pairs (bp) to 51,200 bp. These sizes include 50 bp, 100 bp, 200 bp, 400 bp, 800 bp, 1,600 bp, 3,200 bp, 6,400 bp, 12,800 bp, 25,600 bp, and 51,200 bp. Specifically, the processor utilizes the partition sizes of 800 bp, 1,600 bp, 3,200 bp, 6,400 bp, 12,800 bp, 25,600 bp, and 51,200 bp to estimate the copy number of the target sequence. The processor determines the mean read coverage for each of these partition sizes and analyzes the coverage peaks and distances between peaks. Based on this analysis, the processor is then able to estimate the copy number associated with the target sequence.

[0028] To estimate a standard deviation of the determined copy numbers, the processor may utilize partition sizes ranging from 50 base pairs (bp) to 12,800 bp. Specifically, the processor analyzes the mean read coverages determined for partition sizes of 50 bp, 100 bp, 200 bp, 400 bp, 800 bp, 1,600 bp, 3,200 bp, 6,400 bp, and 12,800 bp. By leveraging the coverage data obtained across these varying partition sizes, the processor can assess the variability and noise present in the coverage measurements. This variability analysis allows the processor to estimate a standard deviation for the copy numbers, which provides a measure of the reliability and accuracy of the determined copy number values.

[0029] When partitioning the reference genome, the processor may ensure that all partitions of the partitioned reference genome are non-overlapping. Specifically, each segment of the reference genome that is analyzed is only included in a single partition and does not overlap with any other partition, in one specific embodiment. This non-overlapping approach prevents the same region of the reference genome from being double counted when determining the mean read coverage for each partition size. Only analyzing non-overlapping segments allows for an accurate assessment of coverage across the reference genome for each tested partition size. The non-overlapping partitions provide reliable coverage data that the processor can then use to determine the mean read coverage for each size until the predetermined number of consecutive mean-read coverage peaks are identified.

[0030] The method may involve partitioning the reference genome into sizes where at least two of the partitions overlap with each other. By allowing some level of overlap between the partitions of the reference genome, it enables a more comprehensive analysis of read coverage across the reference genome. For example, coverage information from an overlapping region between a first partition of 800 base pairs and a second partition of 1,600 base pairs can be leveraged to provide additional data points for determining the relationship between observed read coverage and copy number across the reference genome. The overlapping partitions can thereby facilitate a more accurate estimation of genomic copy numbers from the aligned sequence read data.

[0031] In determining the genomic-region copy number, the processor partitions the reference genome into multiple sizes. In specific embodiments, all partitions of the partitioned reference genome may be sequential partitions. That is, when partitioning the reference genome, the processor may not create random or non-sequential partitions in this embodiment. Rather, it may partition the reference genome into consecutive sections or fragments of varying lengths. This allows the processor to systematically analyze the read coverage across the sequentially partitioned sizes of the reference genome. The sequential partitioning enables the processor to methodically determine the mean read coverage for each size and identify consecutive mean-read coverage peaks that indicate distinct copy numbers. The consecutive nature of the partitions facilitates accurate identification of the relationship between coverage and copy number.

[0032] The method determines the number of consecutive mean-read coverage peaks that will be analyzed. The processor is configured to observe a predetermined

number of consecutive mean-read coverage peaks with an equal coverage distance between each peak. In one example implementation of the method, the predetermined number of consecutive mean-read coverage peaks that is analyzed is three peaks. By setting the number of peaks to three, the processor will determine the mean read coverage for each size of the partitioned reference genome until it observes three consecutive peaks with an equal coverage distance between each peak. This allows the method to leverage patterns in the sequencing data represented by three distinct coverage peaks in order to accurately estimate the copy numbers of target sequences.

[0033] To determine the copy number associated with each peak, the processor compares the equal coverage distance observed between the peaks to the value of the first peak. This first peak is one of the predetermined number of consecutive mean-read coverage peaks identified during the method. By comparing the equal coverage distance to the value of the first peak, the processor can establish a first respective copy number that corresponds to the first peak of the consecutive mean-read coverage peaks. Specifically, the processor analyzes the relationship between the equal coverage distance and the value of the first peak to deduce the appropriate copy number to assign to that first peak. This process allows the processor to systematically determine a respective copy number for each of the peaks based on their relative values and the observed equal coverage distance between peaks.

[0034] The processor analyzes the equal coverage distance between the consecutive mean read coverage peaks. For the first peak of the predetermined number of consecutive mean read coverage peaks, if the first peak's value is approximately equal to the equal coverage distance, the processor determines that the first respective copy number corresponding to that first peak is one. By comparing the value of the first peak to the established equal coverage distance between peaks, the processor can deduce that if the first peak's value is roughly equal to the distance, it represents a copy number of one. This process enables the processor to systematically assign a respective copy number to each peak based on its value relative to the equal spacing between peaks observed in the coverage data.

[0035] In determining the respective copy numbers associated with each of the consecutive mean-read coverage peaks, if there are two peaks with an equal coverage distance between them, the processor may determine that the first peak corresponds to a copy number of one. Additionally, the processor may determine that the second peak of the two consecutive peaks corresponds to a second respective

copy number of two, whereby the coverage value of the second peak would be approximately twice the equal coverage distance between the two peaks. By comparing the coverage values of the peaks to the equal coverage distance, the processor is able to deduce the respective copy numbers of each peak in the sequence. This process enables the processor to analyze the read coverage data and assign copy numbers to represent the copy state reflected by each distinct coverage peak.

[0036] One step in the method may involve determining the respective copy number associated with consecutive mean-read coverage peaks. The mean-read coverage peaks formed have an equal coverage distance between them. To determine the respective copy numbers, the processor compares the value of the first peak to the equal coverage distance. If the value of the first peak is about twice the equal coverage distance, the processor will determine the first respective copy number to be two. This indicates that when the coverage value of the first peak is approximately double the distance between peaks, the copy number assigned to that first peak will be two.

[0037] The method determines the copy numbers associated with the consecutive mean-read coverage peaks formed with an equal coverage distance between them. If the first peak is determined to have a copy number of two based on its coverage value being approximately twice the equal coverage distance, then the second peak of the predetermined number of consecutive mean-read coverage peaks will correspond to a second respective copy number of three. This is because if the coverage value of the first peak represents a copy number of two, and the peaks have an equal coverage distance between them, then the coverage value of the second peak would be approximately three times the equal coverage distance. Therefore, the second peak is assigned a copy number of three.

[0038] The method may also utilize a regression model to correlate coverage numbers with copy numbers. Specifically, the processor establishes a regression model that mathematically represents the relationship between observed read coverage and genomic copy number. By determining the coverage numbers across the partitioned reference genome and the copy numbers assigned to the consecutive mean coverage peaks, the processor can generate a regression model. This regression model correlates the coverage numbers to the copy numbers using a statistical regression technique. The resulting regression equation encapsulates how coverage and copy number vary together based on the analyzed data. This

regression model then serves as the established model for the processor to estimate copy numbers of target sequences based on their coverage numbers determined from the aligned reads.

[0039] The method may involve partitioning the reference genome into multiple sizes using the processor. The partitioned reference genome may correspond to at least one gene within the genome. By partitioning the reference genome and analyzing the read coverage across the different partition sizes, the processor is able to determine trends and patterns related to the correlation between read coverage and copy number for the genomic region containing the gene or genes of interest. This gene-level analysis allows close examination of the coverage profiles for specific genes to facilitate more targeted copy number estimation for the genes present in the partitioned reference genome segments.

[0040] Additionally, the method may further comprise normalizing the read coverages by a reference coverage value obtained from a control sample. To perform this normalization, the processor first obtains read coverage data from a separate control sample. This control sample coverage represents a baseline reference value against which the coverages from the sample can be standardized. The processor then divides each of the read coverages from the sample by the reference coverage value derived from the control. Normalizing the coverages in this way accounts for any technical variations between samples, such as differences in the total number of sequenced reads, and allows for more accurate comparison of coverage levels between genomic regions. This normalization step enhances the ability of the method to reliably estimate copy numbers by controlling for technical biases in the sequencing data.

[0041] The method may further involve estimating a standard deviation of the estimated copy numbers based on the read coverages and a size of the target sequence. To estimate the standard deviation, the processor analyzes the read coverages across different sizes of the partitioned reference genome. Using the varying read coverages from the different partition sizes and the estimated copy numbers, the processor calculates a standard deviation for each estimated copy number. This standard deviation accounts for variability in the read coverages based on the size of the target sequence, such as a gene, which was used to estimate the copy number. The estimated standard deviation provides a measure of reliability for each copy number prediction based on the observed variability in read coverages associated with sequences of varying lengths.

[0042] In addition to determining the copy numbers, the method may further involve determining a confidence interval for the estimated copy numbers. This determination is based on the standard deviation of the copy number estimates and the estimated copy numbers themselves. Specifically, the processor first estimates the standard deviation of the copy numbers based on the read coverages and the size of the target sequence, as described in an earlier step. This standard deviation reflects the variability in the copy number estimates. The processor then uses both this standard deviation and the actual estimated copy numbers to calculate a confidence interval for each copy number prediction. The confidence interval provides a measure of reliability or precision for each estimated copy number value.

[0043] The alignment of the sequencing reads to the reference sequence is performed using an accelerated version of the BWA aligner. Specifically, the processor aligns the generated reads to the reference sequence, which includes the reference genome concatenated with the vector sequence. This alignment is conducted using an accelerated version of the BWA aligner algorithm. BWA, which stands for Burrows-Wheeler Aligner, is commonly used to map low-divergent sequences against a large reference genome. Using an accelerated version of this algorithm allows the reads to be aligned to the reference sequence in an optimized and efficient manner. Employing an accelerated aligner helps improve the speed of the read alignment process, thus enabling alignment of the hundreds of millions of reads within a reasonable timeframe for downstream analyses.

[0044] In determining the mean read coverage for each size of the partitioned reference genome, the processor may calculate a median coverage value for each size. Specifically in this embodiment, the processor does not simply calculate the average coverage across all reads mapped to a partitioned region of a given size in some embodiments. Rather, it calculates the median coverage value for each partitioned size to minimize the impact of outliers in the coverage data. By taking the median coverage rather than the straight average, the influence of very high or very low coverage reads is reduced. This may lead to a more accurate characterization of the central tendency of coverage across each genomic partition size. Using the median coverage value for each partition size enables the processor to robustly determine the relationship between coverage amounts and copy numbers, facilitating the downstream generation of an effective model for estimating genomic copy number.

[0045] The method further involves training the model using the mean base-level coverage of each partitioned reference genome sequence and known copy-numbers within the reference sequence. To train the model, the processor first determines the mean base-level coverage for the partitions of different sizes that have known copy-numbers. These partitions with known copy-numbers allow the processor to establish the relationship between the measured coverage and the actual copy-numbers. The coverage measurements and corresponding known copy-numbers are then used by the processor to develop and optimize the model. The trained model captures the correlation between coverage levels and copy-numbers observed within the reference sequence itself. This process enables the model to more accurately estimate the copy-number of target sequences based on their coverage levels.

[0046] The method may further involve performing a quality control analysis on the read coverages to identify potential sequencing errors and filter out false-positive copy number calls. By analyzing the read coverages, the processor can identify errors that may have occurred during the sequencing process. This allows the processor to remove read coverages impacted by sequencing errors from downstream analyses. Additionally, the processor filters out any copy number calls that were falsely identified as true positive due to sequencing errors or other technical artifacts. Performing this quality control analysis may enhance the accuracy of the copy number determination by excluding low quality or erroneous read coverage data and copy number calls that are not truly representative of the genomic sample, in some specific embodiments.

[0047] The method further comprises visualizing the estimated copy numbers and their associated confidence intervals on a Circos plot. A Circos plot is a circular graphical representation that can be utilized to analyze and depict various genomic features derived from the sequence data. In particular, the Circos plot generated by this method displays tracks showing the estimated copy numbers of target genomic regions, read coverage information, and confidence intervals corresponding to the copy number predictions. This visual depiction aids in the interpretation and examination of the results from the copy number analysis. It provides insights into the estimated abundance and variability of target sequences within the sample in an intuitive format.

[0048] The method can determine the copy number of one or more specific genes of interest within the vector sequence. By analyzing the reads that are generated from sequencing and subsequently aligned to the reference sequence, which

includes both the reference genome and vector sequence, the processor can estimate the coverage of the genes of interest located within the vector. Using this observed coverage of the genes of interest and the model that correlates coverage to copy number, the processor then has the ability to predict the precise copy number of the one or more genes of interest situated in the vector sequence. This provides valuable information about the abundance and representation of the genes of interest contained within the sample's genomic material.

[0049] The method may further involve determining the copy number of one or more genes of interest (GOIs) located within the vector sequence. To estimate the copy number of these GOIs, the processor first obtains the observed read coverage for each GOI from the aligned sequencing reads. The processor then applies the model that correlates read coverage to copy number. By inputting the measured read coverage of a GOI into the model, which incorporates the relationship between coverage and copy number, the processor can calculate the estimated copy number for that particular GOI based on its observed read coverage data and the model parameters. This allows the method to precisely determine the copy number for specific genes within the vector, providing useful information about the abundance and representation of the target GOI sequences integrated into the genome.

[0050] The method may analyze a sample containing genetically modified cells with a vector sequence that has been integrated into the genomic DNA. This vector sequence can be a full-size vector or a partial-size vector, depending on the specific genetic modification applied to the cells. A full-size vector may represent the complete, unaltered vector sequence used to modify the cells. A partial-size vector could indicate rearrangements or deletions in the integrated vector sequence compared to the original vector. By determining whether a full-size or partial-size vector has integrated into the genomic DNA, the method provides insights into the structure and characteristics of the transgene modification present in the analyzed sample. This information supports comprehensive molecular characterization of the cells for purposes such as selection of desired biotherapeutic cell lines or assessment of genetic engineering approaches.

[0051] The method can further involve identifying and reporting any vector-vector concatenation events based on split reads that map to different parts of the vector sequence. To detect vector concatenation events, the sequencing reads aligned to the vector reference sequence are extracted from the global alignment file. These extracted reads are then re-aligned to the vector reference sequence using

commonly used read alignment software. This re-alignment process aims to identify vector chimeric reads, which are reads aligning to different non-homologous parts of the vector sequence. A junction position is considered valid if it is supported by multiple different chimeric reads. By analyzing these split reads that map to different parts of the vector sequence, the method can detect whether any vector-vector concatenation events have occurred. This provides comprehensive characterization of the vector structure within the genomic sample.

[0052] The method can also involve detecting and reporting any sequence variants in an integrated vector. To do this, the processor compares the generated sequencing reads to the known vector sequence. By analyzing differences or mismatches between the reads and the vector sequence, the processor is able to identify the presence of any single nucleotide mutations, small insertions, or deletions that may exist in the integrated vector. These potential sequence variants are then evaluated, as true variants could represent real mutations that have occurred over time in the integrated vector. However, some observed variants may simply be sequencing errors introduced during the library preparation or sequencing processes. The processor applies various screening steps and filtering criteria to determine high-confidence sequence variants and eliminate false positives. Any verified sequence variants identified in this manner are then output and reported by the processor. This allows for characterization of any changes that may have developed in the integrated vector sequence compared to the original reference sequence.

[0053] The method can further involve detecting vector integration points in the reference genome. This is done by analyzing chimeric reads from the aligned reads. Chimeric reads are reads that partially map to the vector sequence and partially map to non-vector regions. Additionally, paired reads are examined where one of the paired reads maps to the vector sequence and the other paired read maps to a non-vector region. These chimeric reads and paired reads are then assembled de novo to generate longer contigs. The contigs are aligned to the reference genome, which includes both the host genome and vector sequences. This alignment process can identify the boundaries between the donor vector DNA and the host genome, thereby determining the precise locations where the vector has inserted within the host genome.

[0054] The method can be utilized for the molecular characterization of Chinese hamster ovary (CHO) cells that are commonly used for producing recombinant proteins. CHO cells are frequently the preferred cell type used for recombinant

protein production due to their ability to execute complex protein post-translational modifications. However, long-term cultivation of CHO cells can present challenges, such as a decrease in recombinant protein expression levels over time or changes in the protein structures that impact functionality and safety. The method allows for detailed molecular characterization of CHO cells at the genomic level, helping to evaluate copy number stability, integrity of transgene sequences, and any alterations within the host CHO genome. This type of comprehensive molecular analysis supports the selection and validation of CHO cell lines for the safe and effective production of recombinant therapeutic proteins.

[0055] This method of determining genomic-region copy number can be used to characterize the reference genome at a molecular level. Specifically, after isolating and purifying genomic nucleic acids from a sample, a shotgun library is created and sequenced to generate reads. These reads are then aligned to a reference sequence using a processor, where the reference sequence includes both a reference genome and a vector sequence. The reference genome is partitioned into multiple sizes by the processor and mean read coverage is determined for each size. Once consecutive peaks in mean read coverage are identified with equal distances between peaks, the processor determines the respective copy number associated with each peak. Finally, a model is established by the processor to estimate the copy number of target sequences. This process allows molecular characterization of the reference genome by analyzing read coverage and determining copy numbers across the genome.

[0056] The method can be used for the molecular characterization of mammalian cancer cell lines. Specifically, the method allows characterization of the copy number and genomic content of integrated vectors or transgenes within cancer cells. By applying this method to cancer cell samples, researchers can gain valuable insights into genetic alterations in cancer. For example, the method may be used to analyze transgene integration patterns and copy number changes in human cancer cell lines used in research. This molecular characterization of transgenic cancer cell lines can provide important information to further scientific understanding of cancer genetics.

[0057] The method further involves calculating a coverage variability for each size of the partitioned reference genome. This is done to estimate a reliability parameter of the estimated copy numbers. Specifically, after partitioning the reference genome into different sizes and determining the mean read coverage for each size, the processor calculates the variability in the read coverages across the different sizes.

This coverage variability provides an indication of the reliability of the estimated copy numbers, with lower variability corresponding to more reliable estimates. By factoring in the variability, the processor can account for noise and uncertainty in the coverage measurements when determining the confidence and accuracy of the copy number predictions.

[0058] In determining the mean read coverage for each size of the partitioned reference genome, the processor calculates the mean base-level coverage of each partition by taking into account the size and copy number of each fragment. Specifically, the processor considers both the length of each partitioned reference genome fragment, as well as the expected copy number based on prior analysis, when computing the average base-level coverage across a given partition. This allows the processor to more accurately determine the relationship between coverage and copy number by normalizing for differences in fragment size and predicted copy state during the calculation of mean coverage values. Taking the size and putative copy number into account provides a more refined evaluation of coverage trends and correlations relative to copy number.

[0059] The method may also involve determining coverage anomalies within the genomic regions. To determine any coverage anomalies, the processor analyzes the mean read coverage across the partitioned sizes of the reference genome. By examining the coverage values at each partition size, the processor can identify any irregular or abnormal coverage readings that deviate from expected patterns. Such anomalous coverage results could be caused by local sequencing errors or mismapping of reads due to repeats or structural variations in the genome. Identifying any coverage anomalies enables the processor to filter out errant coverage data from the determination of mean read coverage values and copy number estimates. In some specific embodiments, only coverage patterns deemed regular are utilized to establish the relationship between coverage and copy number for generating an accurate copy number estimation model.

[0060] The method can further involve identifying regions of homologous sequences within the vector sequence. To do this, after sequencing the shotgun library and aligning the generated reads to the reference sequence, the processor analyzes the vector sequence portion of the reference sequence. The processor may search for sequences within the vector that are identical or very similar to each other. These homologous sequences represent regions of the vector that share a high degree of similarity in their nucleotide composition. By locating homologous regions

in the vector, the method provides additional characterization of the vector's internal structure and composition beyond just the copy number determination. This added step enhances the molecular characterization of the integrated vector sequence.

[0061] The method may also involve comparing the estimated copy number of the target sequence with a predetermined threshold in order to estimate the copy number stability and integrity of the target sequence. By comparing the estimated copy number of the target sequence, such as a gene of interest, to a predetermined threshold value, insights can be gained regarding the stability and integrity of the copy number. If the estimated copy number is within the threshold range, it suggests the copy number is stable and intact. However, if the estimated copy number falls outside the threshold range, it could indicate the copy number is unstable or compromised in integrity. Therefore, comparing the estimated copy number to a predetermined threshold provides a means for evaluating the stability and consistency of the target sequence copy number.

[0062] The method further involves validating the accuracy of the copy number determination using a molecular characterization technique. This molecular characterization can come from one of several established methods, including Southern blot analysis, digital droplet PCR (ddPCR), or quantitative PCR (qPCR). Southern blot analysis is a traditional technique that uses gel electrophoresis and DNA hybridization to visualize DNA fragments and validate estimated copy numbers. Additionally, ddPCR and qPCR are highly sensitive methods that enable accurate validation through direct measurement of DNA copy counts. By comparing the estimated copy numbers from the partitioning and model-based approach to results from Southern blot, ddPCR or qPCR analysis, the method is able to validate the accuracy of the copy number determination. This validation step helps to confirm the reliability of the copy number estimates calculated through this novel genomic partitioning and modeling workflow.

[0063] To improve the accuracy of variant detection and integration site annotation, the method may generate the reads using long-read sequencing technology. Long-read sequencing is able to produce much longer reads compared to short-read technologies. These longer reads provide additional sequence information that can enhance the precision of identifying variations present in the vector sequence when compared to the reference. They also allow more precise determination of vector integration points by facilitating assembly of split reads and paired reads that span vector-host junctions. Therefore, employing long-read sequencing for read

generation within the method can benefit the downstream analyses for detecting variants and annotation of vector integration sites.

[0064] Additional analysis can be performed on the sequenced reads to detect any rearrangements or structural variations that may exist within the reference sequence. The method may analyze the alignment pattern of the sequenced reads that were aligned to the reference sequence, which includes the reference genome and vector sequence. By examining the alignment pattern of the reads, the processor can identify any potential rearrangements or structural variations that differ from the reference sequence. This could involve analyzing aspects of the read alignments such as alignment location, orientation, and pairing information. Detecting any rearrangements or structural variations provides further characterization of the genomic sequence of the sample beyond just copy number determination. This additional analysis enhances the comprehensive molecular profiling of the sample achieved through this method.

[0065] Additionally, the method can involve generating a molecular characterization report. This report may summarize results of the analysis. It contains information such as the estimated copy number of the target sequence as determined by the model. The report also includes details regarding any detected variants in the sequence. Specifically, it outlines the variant profile, providing data on identified single nucleotide variants and small insertions or deletions. Furthermore, the report specifies the integration sites where the target sequence has incorporated into the reference genome. Locating these integration sites is achieved by analyzing chimeric reads and paired reads mapped to vector and non-vector regions. In summary, the molecular characterization report consolidates findings about the target sequence's estimated abundance, variants, and genomic placement within the reference.

[0066] The method may further involve determining a ploidy distribution of the reference genome to refine the estimated copy number of the target sequence. To determine the ploidy distribution, the processor analyzes the observed coverages of fragments with varying sizes that were partitioned from the reference genome. By examining the coverages across different fragment sizes, the processor can identify patterns in the data related to the ploidy, or copy number state, of regions in the reference genome. The location and width of peaks in the ploidy distribution provide insights regarding the true copy number profile. For example, the dominant peak often corresponds to the primary copy number. Taking into account the ploidy

distribution can help improve the accuracy of copy number estimates made for the target sequence.

[0067] An accuracy of the estimated copy number determined by the method is further validated using at least one of Fish analysis, ddPCR analysis, or qPCR. Experimental validation techniques, such as fluorescence in situ hybridization (FISH), digital droplet PCR (ddPCR), or quantitative PCR (qPCR), can be applied to the biological sample to obtain independent copy number measurements for comparison against the estimates calculated using the method. Comparing the method's estimated copy numbers to values obtained from well-established validation methods helps to confirm the reliability and accuracy of the copy number determinations achieved through the computational pipeline.

[0068] To further refine the copy number estimates, the processor may fetch coverages of different fragments with varying sizes in the reference genome. These fragment sizes include 50 base pairs (bp), 100 bp, 200 bp, 400 bp, 800 bp, 1,600 bp, 3,200 bp, 6,400 bp, and 12,800 bp. The processor extracts the coverage information for each of these fragment sizes from the aligned reads. Analyzing the coverages across different fragment sizes allows the processor to better understand the relationship between coverage and copy number. This relationship may be used for developing an accurate copy number estimation model. Smaller fragment sizes help assess coverage variability and noise, while larger fragments are more useful for discerning differences in copy numbers. Therefore, fetching coverages of fragments with varying sizes, from 50 bp to 12,800 bp, facilitates generating a precise model for estimating genomic copy numbers.

[0069] Additionally, the method may involve normalizing the read coverages. The processor determines the largest peak among the consecutive mean-read coverage peaks identified during the determination of mean read coverage for each size of the partitioned reference genome. The processor then normalizes the coverage at each point by dividing the coverage at that point by the coverage of the largest peak. This process of normalization provides normalized coverage values that can be used in downstream analysis by the processor. Normalizing the coverage based on the largest peak helps account for differences in total read numbers that may occur between samples or runs, thereby allowing for a more effective comparison of coverage data across multiple experiments.

[0070] The method may further involve normalizing the occurrence values determined from the analysis. Specifically, the processor may normalize an

occurrence by a total occurrence of all coverages. This is done by dividing the occurrence at each data point by the sum of occurrences across all data points, in some embodiments. This calculation provides normalized occurrence values that allow for comparison of the frequency of different coverage amounts. Normalizing in this manner helps account for differences in the total number of occurrences that may exist between samples or experiments. It rectifies the values to make them directly comparable, which can improve the accuracy and reliability of subsequent analysis steps like determining the copy numbers or detecting anomalies. Overall, normalization of the occurrence values enables more robust statistical analysis of the sequencing data.

[0071] To further refine the copy number estimates, the processor may extract the mean read coverage for each of the plurality of sizes of the partitioned reference genome where the copy numbers are known. The processor then computes the standard deviation for each size of the partitioned reference genome. Computing the standard deviation assesses the variation in coverage between fragments of different sizes. Using the standard deviation, the processor can gain a better understanding of the reliability of the estimated copy numbers based on the amount of variation in coverage observed between different sized fragments. Assessing the variation in fragment coverage using the standard deviation helps to refine the estimated copy numbers determined through the method.

[0072] The method can additionally involve determining concatenation events of the vector sequence. To do this, the processor analyzes the aligned reads to identify potential vector concatenation junction reads. A vector concatenation junction read is a read that contains two segments that align to different parts of the vector sequence, joining the vector sequence segments together within the read. By identifying these junction reads, the processor can determine if any concatenation events have occurred among the vector sequences integrated into the host genome. Detecting vector concatenation events provides further insight into the structural characteristics and organization of the vector sequences present in the sample.

[0073] The method may further involve determining junction reads in order to detect concatenation events of the vector sequence. A junction read refers to a read that contains two segments aligning to different parts of the vector sequence. Specifically, the method analyzes the aligned sequencing reads to identify junction reads that have a first segment mapping to one region of the vector sequence and a second segment mapping to another non-homologous region of the vector sequence.

These junction reads indicate that the two vector sequence regions are joined together within a single read. Detecting junction reads in this manner allows the method to determine the presence of vector concatenation events, where separate parts of the vector sequence have fused together. This provides insights into the structural characteristics and integrity of the integrated vector.

[0074] The method may further involve identifying chimeric reads from the aligned reads. Chimeric reads refer to sequencing reads that partially map to the vector sequence and partially map to non-vector regions of the reference genome. By examining these chimeric reads, which contain segments that align to both vector and non-vector genomic regions, the method can gain insights into the genomic location of vector integration. Analyzing the chimeric reads can help determine where the vector sequence has integrated within the host genome, providing information on the insertion points and boundaries between the vector and non-vector genomic regions. This additional analysis of the chimeric reads enhances the characterization of the vector integration and sequence structure obtained through the copy number determination process.

[0075] To detect vector concatenation events and identify the precise location of vector insertion in the host CHO genome, the method analyzes the sequencing reads for junction reads. Junction reads are reads that contain two segments aligning to different parts of the vector sequence. The method extracts sequencing reads aligned to the vector reference sequence from the global alignment file. These extracted reads are then re-aligned to the vector reference sequence using commonly used read aligning software, such as consed/phrap software. This re-alignment process aims to identify vector chimeric reads, which are reads aligning to different non-homologous parts of the vector sequence. A junction position is considered valid if it is supported by multiple different chimeric reads. By determining vector-vector junctions based upon the identified chimeric reads, the method can thereby determine the precise location of vector insertion within the reference genome, which includes both the host genome and vector sequences.

[0076] The method further involves performing de novo assembly of the identified chimeric reads and their respective mates to generate longer contiguous sequences, also known as contigs. These contigs represent continuous stretches of the DNA sequence. The de novo assembly process joins together the sequenced reads, which only provide short segments of the full sequence, to reconstruct longer fragments. By assembling the short reads that partially map to the vector sequence and partially

map to non-vector regions, as well as the reads where one paired read maps to the vector and the other maps to non-vector regions, the method aims to generate longer contigs spanning the boundaries between the vector insertion and the host genomic sequence. These assembled contigs provide information to help identify the precise junction points where the vector has integrated within the reference genome.

[0077] The method further involves aligning any assembled contigs to the reference genome, which consists of both the host genome and vector sequences. This alignment process serves to identify boundaries between the reference genome and vector sequences. Specifically, the assembled contigs that were generated from chimeric reads are aligned to the reference genome. This reference genome includes the sequences of the host genome as well as the vector sequences. By aligning the contigs to this combined reference genome, the locations where the reference genome sequences and the vector sequences join together can be identified. This process thereby delineates the boundaries between the genomic regions of the reference genome and the vector sequences. Moreover, an alignment file is created from this alignment of the assembled contigs to the reference genome. This newly formed alignment file captures the defined boundaries between the reference genome and vector sequences.

[0078] The method may utilize a system for determining copy number. This system comprises a processor that is configured to perform the method. The processor is able to complete any of the steps described in the method claims. Specifically, the processor can isolate and purify the genomic sample, create the shotgun library, sequence the library to generate reads, align the reads to the reference sequence, partition the reference genome into multiple sizes, determine the mean read coverage for each partition size, determine the respective copy number for consecutive coverage peaks, and establish a model to estimate the copy number of a target sequence. The processor is used to carry out the various acts of the method, including aligning sequences, partitioning genomes, calculating coverage statistics, assigning copy numbers, and developing a copy number estimation model. By utilizing this system with a configured processor, the method is able to determine genomic copy numbers in an automated, computational manner.

[0079] The method may be implemented by an apparatus that contains an operative set of processor-executable instructions. This operative set of processor-executable instructions is configured to cause a processor to perform one or more acts of the method described here. Specifically, the processor-executable

instructions allow a processor to align the reads to a reference sequence, partition the reference genome into multiple sizes, determine the mean read coverage for each size of the partitioned reference genome until a predetermined number of consecutive mean-read coverage peaks are formed with an equal coverage distance between each peak, determine a respective copy number associated with each of these consecutive peaks, and establish a model configured to estimate the copy number of a target sequence. By utilizing processor-executable instructions, the method can be implemented using an apparatus that incorporates a processor to carry out the various steps for determining genomic-region copy number in an automated fashion.

[0080] The method may involve using a computer-readable medium storing instructions that, when executed by a processor, causes the processor to perform multiple steps. The processor is first instructed to align the generated reads from sequencing to a reference sequence. This reference sequence consists of a reference genome combined with a vector sequence. Next, the processor is instructed to partition the reference genome into a plurality of different sizes. The processor then determines the mean read coverage for each size of the partitioned reference genome. It continues this determination until a predetermined number of consecutive mean-read coverage peaks are formed, with an equal coverage distance between each peak. Following this, the processor determines a respective copy number associated with each of the consecutive mean-read coverage peaks. Finally, the processor is instructed to establish a model that is configured to estimate the copy number of a target sequence. In this way, a computer-readable medium can store instructions that direct a processor to perform the method of aligning reads, partitioning the genome, determining coverage, assigning copy numbers, and creating a copy number estimation model.

[0081] The method may also utilize a computer-readable medium that is configured to cause a processor to perform additional steps of the method. Specifically, the computer-readable medium could direct the processor to carry out various aspects of the method as described herein. These additional embodiments may include isolating and purifying the sample, creating the shotgun library using different sequencing techniques, partitioning the reference genome into a range of sizes including 50 base pairs to 51,200 base pairs, determining the copy numbers by comparing the peak distances or using regression analysis, developing the model to estimate copy numbers without genomic location parameters, and applying the method to characterize Chinese hamster ovary cells, mammalian cancer cell lines, or specific

genes of interest. The computer-readable medium thus allows the processor to flexibility perform any combination of these additional steps to determine genomic copy numbers.

[0082] One embodiment of the computer-readable medium used in the method is that is non-transitory. Additionally, the computer-readable medium may cause the processor to partition the reference genome into multiple sizes, determine the mean read coverage for each size until a number of consecutive coverage peaks are formed with an equal distance between them, and determine a respective copy number for each peak. This computer-readable medium storing the instructions may be non-transitory, meaning it is not transient and is a physical medium, such as a hard drive, USB drive, solid state drive, or other fixed storage medium, as opposed to a transitory medium like signals or carrier waves.

[0083] The method further involves extracting genomic DNA (gDNA) from a mammalian cell line to obtain the sample of genomic nucleic acids. This gDNA sample has a mass within the range of 300 ng to 2000 ng. This gDNA sample then undergoes tagmenting to form tagged gDNA fragments. The sample is first cleaned up using sample preparation beads. Indexes are then ligated to the tagged gDNA fragments from the tagmenting. Finally, the sample is cleaned up a second time using sample preparation beads, resulting in the library that is used for next generation sequencing (NGS).

[0084] The method can be utilized for the molecular characterization of Chinese hamster ovary (CHO) cells or Homo sapiens cell lines that are commonly used for producing recombinant proteins. CHO cells and certain Homo sapiens cell lines are frequently preferred cell types used for recombinant protein production due to their ability to execute complex protein post-translational modifications. However, long-term cultivation of these cell types can present challenges, such as a decrease in recombinant protein expression levels over time or changes in the protein structures that impact functionality and safety. The method allows for molecular characterization of CHO cells or relevant Homo sapiens cell lines at the genomic level, helping to evaluate copy number stability, integrity of transgene sequences, and any alterations within the host genome. This type of molecular analysis supports the selection and validation of optimal cell lines for the safe and effective production of recombinant therapeutic proteins.

[0085] The method may involve isolating and purifying a sample of genomic nucleic acids from the biological sample. This step provides the raw genomic material for

subsequent analysis. The sample may contain a homogeneous cell type or a heterogeneous mixture of various cell types. To isolate the nucleic acids, the sample first undergoes cell lysis to break open the cell membranes and release the contents. This allows for the extraction of the genomic DNA. It is preferable that the extracted genomic DNA has high molecular weight in order to avoid shearing and fragmentation during subsequent processing steps. Genomic DNA of high molecular weight helps maintain the integrity of the DNA, which may be used for creating an unbiased library representative of the genomic sample and obtaining accurate sequencing results.

[0086] The method utilizes a PCR-free library preparation workflow to reduce potential biases and errors that could be introduced through PCR amplification. Specifically, when creating the shotgun library from the purified sample of genomic nucleic acids, a tagmentation-based approach is used that does not include a polymerase chain reaction (PCR) amplification step. By employing a PCR-free method, this helps decrease the number of mutations, insertions, or deletions occurring compared to library preparation methods that use a PCR amplification step. Therefore, the act of creating the shotgun library in the method uses a PCR-free workflow when constructing the library with next-generation sequencing techniques. Utilizing a PCR-free approach for the library preparation aids in generating an unbiased representation of the genomic sample for subsequent analysis.

[0087] The method may further involve adjusting the volume of the sample preparation beads utilized during the first cleanup step or the second cleanup step in order to modify the median size of the fragments within the library. By varying the volume of the sample preparation beads used in either the initial cleanup stage or subsequent cleanup stage, the median length of the fragments that comprise the sequencing library can be altered. This allows for optimization of the fragment size distribution based on the intended sequencing application or downstream analyses.

[0088] The method modifies the median library fragment size by adjusting the volume of the sample preparation beads during the first clean-up step or the second clean-up step. In one embodiment, the method adjusts the volumes such that the median library fragment size is greater than 350 base pairs. Adjusting the volumes of the sample preparation beads used in the clean-up steps enables tuning of the final median fragment size of DNA inserts within the library. A median fragment size larger than 350 base pairs may allow for more robust detection of integration sites compared to a smaller median fragment size.

[0089] The method may involve performing a first cleaning step after tagmenting the sample of genomic DNA and before ligating indexes. This first cleaning step utilizes sample preparation beads to clean up the sample. Specifically, this first cleaning step comprises using sample preparation beads in a volume that is selected within the range of 30 microliters to 50 microliters. By selecting the volume of sample preparation beads between 30 and 50 microliters for the first cleaning step, the method is able to modify parameters such as the median library fragment size resulting from the downstream library preparation process.

[0090] The method may involve adjusting parameters in the library preparation steps to modify characteristics of the resulting DNA library. Specifically, the second cleaning step in the process of preparing the shotgun library from the sample of purified genomic nucleic acids includes utilizing sample preparation beads in a volume that is selected within the range of 10 microliters to 20 microliters. By selecting the volume of sample preparation beads used during this second cleaning step from within this specified range, it allows for tuning characteristics such as the size distribution profile of fragments in the resultant sequencing library. The size distribution of DNA fragments can influence downstream steps like clustering of template molecules and sequencing accuracy, thus modifying this Parameter provides a means to optimize the quality of the sequencing data output.

[0091] The method may involve isolating and purifying a sample of genomic nucleic acids. A shotgun library is then created from the sample. The shotgun library is sequenced, generating reads, which are aligned to a reference sequence includes a reference genome and a vector sequence using a processor. The reference genome is partitioned into multiple sizes by the processor. The processor determines the mean read coverage for each size of the partitioned reference genome and continues this determination until a predetermined number of consecutive mean-read coverage peaks are formed, with an equal coverage distance between each peak. The processor then determines a respective copy number associated with each of these consecutive peaks. The processor may then establish a model configured to estimate the copy number of a target sequence.

[0092] Additionally, in the step involving adjusting the volume of the sample preparation beads during the first and second cleaning steps to modify the median library fragment size, if the method utilizes a volume of 33 microliters of sample preparation beads in the first cleaning step and a volume of 15 microliters of sample

preparation beads in the second cleaning step, it can further modify the median library fragment size.

[0093] The method may involve adjusting the volume of sample preparation beads used during the first and second cleaning steps in order to modify the median library fragment size. Specifically, when the volume of sample preparation beads used in the first cleaning step is 45 microliters and the volume of sample preparation beads used in the second cleaning step is 12 microliters, it allows the researcher to control the resulting size of fragments in the DNA library created for next generation sequencing. Controlling the fragment size in this manner helps the researcher to optimize the library for downstream sequencing applications and data analysis.

[0094] The method can also involve reducing sequencing errors that occur during high-throughput next generation sequencing (NGS). This is accomplished by performing high-throughput NGS on the DNA library that was prepared according to the steps involving: extracting genomic DNA from a mammalian cell line to obtain the sample, tagmenting the sample to form tagged DNA fragments, cleaning up the sample a first time using sample preparation beads, ligating indexes to the tagged fragments, and cleaning up the sample a second time using sample preparation beads to result in the final library. Sequencing this library produced with fewer mutations, insertions, or deletions compared to methods using PCR amplification helps minimize errors introduced during the sequencing process.

[0095] The method described herein may also involve the preparation of a shotgun library with a mean fragment length of approximately 1500 base pairs, utilizing Nanopore sequencing technology for sequencing, in some specific embodiments. The genomic nucleic acids are isolated from a Chinese Hamster Ovary (CHO)-derived clonal cell line, specifically comprising genomic DNA (gDNA). The gDNA is fragmented into sizes ranging from about 0.3 to 2 kilobases, followed by the conversion of the fragmented DNA into a sequencing-compatible library. Additionally, the resulting DNA library may optionally be quantified using a Qubit fluorometer to ensure target sequencing performance.

BRIEF DESCRIPTION OF THE DRAWINGS

[0096] These and other aspects will become more apparent from the following detailed description of the various embodiments of the present disclosure with reference to the drawings wherein:

[0097] Fig. 1 is a block diagram illustration of a system configured to estimate copy number of a target sequence by generating and applying a model in accordance with an embodiment of the present disclosure;

[0098] Fig. 2 is a block diagram illustration of a computing device configured to estimate copy number of a target sequence by generating and applying a model in accordance with an embodiment of the present disclosure;

[0099] Fig. 3 is a diagram illustrating alignment of sequencing reads to a reference sequence comprising a reference genome and vector sequence in accordance with an embodiment of the present disclosure;

[00100] Fig. 4 is a diagram illustrating partitioning of the reference genome into fragments of varying sizes and determining coverage of the fragments in accordance with an embodiment of the present disclosure;

[00101] Fig. 5 is a graph illustrating a correlation between coverage and copy number used to generate a copy number estimation model in accordance with an embodiment of the present disclosure;

[00102] Fig. 6 is a graph illustrating gene density and read coverage across -CHO genes of known copy number used to create correlation curve between coverage and copy number in accordance with an embodiment of the present disclosure;

[00103] Fig. 7 is a graph illustrating ploidy distribution based on coverage of genomic regions of different size to refine copy number estimates in accordance with an embodiment of the present disclosure;

[00104] Fig. 8 shows the standard deviation as a function of a target fragment size and its copy number estimate in accordance with an embodiment of the present disclosure; and

[00105] Fig. 9 is a graph illustrating normalization in accordance with an embodiment of the present disclosure;

[00106] Fig. 10 shows a flow chart diagram of a method 1000 for estimating the copy number of genetic sequences, such as integrated transgenes in CHO cell lines used for recombinant protein production, in accordance with an embodiment of the present disclosure;

[00107] Fig. 11 provides a comparison of insert size distributions generated by some embodiments of different library preparation protocols for use with the system of Fig.

1 to compare insert length distributions across several PCR-free library preparation protocols in accordance with an embodiment of the present disclosure;

[00108] Fig. 12 shows the coverage uniformity produced by using Illumina® DNA PCR-Free library preparation workflow to generate the NGS library for use with the system of Fig. 1 in accordance with an embodiment of the present disclosure;

[00109] Fig. 13 is a Bioanalyzer trace illustrating the fragment size distribution of sheared DNA used for Nanopore library preparation, confirming the effectiveness of the fragmentation process in accordance with an embodiment of the present disclosure;

[00110] Figs. 14A-14C provide a comparative analysis of sequencing coverage between Nanopore and Illumina sequencing methods, illustrating the correlation and consistency of coverage across the vector sequence in accordance with an embodiment of the present disclosure;

[00111] Fig. 15 presents a comparison of Gene of Interest (GOI) copy number predictions derived from Nanopore and Illumina sequencing in accordance with an embodiment of the present disclosure; and

[00112] Fig. 16 displays an Integrative Genomics Viewer (IGV) screenshot showing comprehensive coverage of the GOI by Nanopore sequencing in accordance with an embodiment of the present disclosure.

DETAILED DESCRIPTION

[00113] Fig. 1 shows a block diagram illustration of a cloud-based system 100 to estimate copy number of a target sequence (e.g., a target genetic sequence) by creating and utilizing a model in accordance with embodiment of the present disclosure. The system 100 includes a cloud-service provider 102, one or more personal computers 104, a genetic sequencer 103, and a mobile device 106. The system 100 also includes a genetic sequence analyzer component 112. The genetic sequence analyzer component 112 can be used generate a model which may be used to estimate a copy number of a target sequence as described below in more detail. The copy number estimate of the genetic material from the sample 101 can be used to accept or reject the use of a cell line for use in a therapeutic purpose (e.g., using Chinese Hamster Ovary ("CHO") cells for recombinant protein synthesis using a vector to insert a transgene into the CHO genome). In other embodiments, the sample may be one of the following: CHO DG44, CHO-S, C0101, F435, CHO

DXB11, CHO K1 SF, CHO K1 PF, CHO K1 ATCC, CHO K1 ECACC. CHOZn, Human Embryonic Kidney 293 (HEK-293), Per.C6 cells, CAP-T, NS0, Sf9 and Sf21-insect.

[00114] The cloud-service provider 102 may be configured to facilitate the analysis of genomic data and the generation of a model by providing an on-demand high-performance computing capability. In some embodiments of the present disclosure, the cloud service provider 102 may be a hosted service such as a company that offers cloud computing services to businesses and individuals such that the cloud service provider 102 provides the infrastructure, software, and platforms required to host, manage, and deliver cloud-based services. In some embodiments, the cloud service provider 102 may provide infrastructure as a service, platform as a service, and/or software as a service. The cloud service provider 102 may be configured to scale up or down its computing resources based upon demand from users at a given moment.

[00115] Referring generally to the system 100, the personal computers 104, the genetic sequencer 103, and the mobile device 106 may communicate with each other via a network 108. The network 108 may be Wi-Fi, ethernet, Bluetooth, etc. and may utilize the internet and associated protocols, such as TCP/IP. The network 108 may be a local area network, a wide-area network, a physical bus (such as a Universal Serial Bus), the internet, or some combination thereof. The computers 104 and/or mobile device 106 may be used to communicate, control, or interface with the sequencer 103 and/or the genetic sequence analyzer component 112 via the network 108.

[00116] A sample 101 is a biological sample containing DNA that needs to be sequenced to obtain detailed genetic information. The sample 101 may include a clonal cell line with a specific genomic region containing integrated vector(s) carrying on transgenes . These transgenes may be designed to serve various purposes, such as the production of recombinant proteins for therapeutic purposes or genetic modification of the cells for specific research objectives. The vector with transgenes can vary in size, complexity, and diversity, depending on the specific goals of the genetic modification. Typically, vector carries on transgene(s), which includes a DNA sequence encoding the protein or genetic trait of interest, along with regulatory elements such as promoters, enhancers, or terminators that ensure proper expression and regulation of the transgene. Thus, the sample 101 may include

Chinese Hamster Ovary cells with a transgene inserted by a vector to create recombinant proteins for therapies.

[00117] The transgenes can be integrated into the host genome through a variety of methods, including plasmid-based transfection, viral vector-mediated gene delivery, or other gene-editing techniques such as CRISPR/Cas9. These techniques allow targeted insertion of transgenes into specific genomic loci or random integration into the genome. Once integrated, the transgenes become a permanent part of the host genome and are passed on to daughter cells during cell division. The number of transgene copies can vary in different cells within the clonal cell line or specific genomic region.

[00118] To sequence the sample 101, a genetic sequencer 103 is used. The genetic sequencer 103 is a high-throughput sequencing instrument that can generate large numbers of DNA reads 105 in a parallel manner. The sequencing process can start by preparing a PCR-free shotgun sequencing library from the sample 101. The library preparation workflow may follow the PCR-free method, which helps reduce potential PCR amplification biases and errors. Various commercial kits such as Illumina® DNA PCR-Free Prep, Tagmentation kit, Truseq DNA PCR-free kit, Perkin Elmer NextFlex Rapid XT Kit, and Qiagen QIAseq FX PCR Free DNA Library Kit can be used for preparing PCR-free libraries.

[00119] Once the library is prepared, it is loaded onto the genetic sequencer 103, which performs the sequencing process. The genetic sequencer 103 uses high-throughput next-generation sequencing technologies such as sequencing by synthesis (SBS) to generate millions or billions of short reads 105 from the DNA fragments in the library. The resulting reads 105 represent short segments of the genetic material present in the sample 101. These reads 105 are typically of a fixed length, i.e., 150 bases but may also be in the range from 50 to 600 bases or longer, depending on the specific sequencing platform and technology used.

[00120] The read data 105 is transmitted over the network 108 to the genetic sequence analyzer component 112. Upon reaching the genetic sequence analyzer component 112, the read data 105 is received and processed by the sequence read ingestor 123. The sequence read ingestor 123 performs the initial step of ingesting the read data 123 into the analysis pipeline of the genetic sequence analyzer component 112. The read ingestor 123 performs various tasks to process and prepare the read data for subsequent analysis steps. This may include data validation, quality control checks, and formatting to ensure compatibility with the

downstream analysis processes. The reads are then stored as read sequences 151 of the database 132.

[00121] The genetic sequence analyzer component 112 also includes an alignment component 114, which aligns the read data 105 to a reference sequence that may include a reference genome (such as CHO), a vector sequence (as shown in Fig. 3), and other sequences. The alignment component 114 utilizes algorithms and bioinformatics techniques to identify the corresponding positions of the read sequences within the reference genome.

[00122] Referring again to Figs. 1 and 3 the alignment component 114 processes the incoming reads from the read sequences 151 and performs sequence alignment. In the alignment process, the alignment component 114 compares each read sequence of the read sequences 151 to the reference sequence 142 using alignment algorithms, such as the Burrows-Wheeler Aligner (BWA) or other industry-standard solutions. These algorithms efficiently match the many reads 105 generated by the genetic sequencer 103 to the reference sequence and/or identify the best-fit alignment positions.

[00123] The alignment component 114 not only aligns the reads to the reference sequence 142, but it also considers specific parameters and variations related to the PCR-free library preparation workflow. This specialized alignment strategy ensures accurate alignment and reduces biases or errors that may arise from the library preparation process.

[00124] The alignment component 114 produces output data that includes aligned read positions, alignment quality scores, and other relevant information. This aligned data serves as the basis for subsequent copy number estimation, sequence variant detection, and structural variant analysis.

[00125] Following the alignment step, the copy-number estimator 124 in the genetic sequence analyzer component 112 generates model parameters 146 to create a model that is used to estimate the copy number of the target sequence. The copy-number estimator 124 utilizes the aligned read positions and the alignment quality scores from the alignment component 114 as inputs.

[00126] In generating the model, the copy-number estimator 124 employs various statistical and machine learning techniques. It analyzes the distribution of aligned read coverage across the reference sequence to identify patterns and correlations related to copy number variations. In some embodiments, the copy-number

estimator 124 assume that the actual copy of any genomic region in our sample is proportional to the observed coverage. Since the vector or gene-of-interest (GOI) may be sequenced together with the host genome (as shown in Fig. 3), the copy-number estimator can leverage the correlation between coverage and copy from the latter to estimate the copy of the former. To this end, the coverage for both host genes and vector bases may be estimated as shown in Fig. 4)

[00127] Through data processing and analysis, the copy-number estimator 124 derives model parameters that effectively capture the relationship between read coverage and copy number. Thus, the model parameters 146 may include regression coefficients, intercepts, and other factors that characterize the relationship between read coverage and copy number. These model parameters 146 may be derived through optimization algorithms that minimize the errors or differences between the predicted copy numbers and the known copy numbers obtained from validation methods. The optimization process aims to create a model that provides accurate and reliable estimates of copy number based on read coverage. Fig. 5 shows one such model where the median coverages corresponding to copy 1, 2, and 3, which were derived from analysis of coverage of thousands of genes of known copy numbers (Fig. 6), is correlated to the copy number using linear regression.

[00128] Referring to Figs. 1 and 4, the copy-number estimator 124 begins by partitioning parts of the reference sequence 142 into partitioned sequences. These partitioned sequences 142 are then aligned to the reference sequence 142 using the alignment component 114, which data is then stored in the partitioned sequences 144 of the database 132. The copy-number estimator 124 then fetches coverages of different fragments with varying sizes from the partitioned sequences 144, ranging from 50bp to 12800bp. The fragments are utilized to draw the ploidy distribution as shown in Fig. 7, creating a scatter and line plot with coverage on the x-axis and occurrence on the y-axis. Fig. 7 is a graph illustrating ploidy distribution based on coverage of genomic regions of different size to refine copy number estimates (e.g., 12800 bp) The fragment size is fine-tuned to optimize peak visibility and resolution as shown in Fig. 8.

[00129] In some additional embodiments, some cell lines may use a customized fragment size choice in order to obtain clear peaks (Fig. 9). Specifically, we will check the peak visibility and resolution under different fragment sizes from 1kb to 100kb (Fig. 7). For each fragment sizes, partitions will be sequentially fetched across the host genome, so that maximum data points will be obtained and used. It is worth

mentioning that using such approach partitions of fragments with different sizes may overlap with each other (e.g., the 800bp partition chr1:1-800 is overlapped with the 1600bp partition chr1:1-1600).

[00130] In one specific embodiment, after deciding the optimal fragment size, we will locate the dominant peak in the ploidy distribution and scan its left and right flanking region with sliding windows (20-200 data points each window). In each window, we will use polynomial regression (degree under 10) to fit the trendline and also calculate the derivatives for each position. Eventually, the average derivatives will be calculated for each position after collecting data from all windows. Starting from the dominant peak, the first blocks of negative average derivatives (with block size above 5 data points) will indicate the sub-peak position.

[00131] In some additional embodiments, a Convolutional Neural Network (CNN) is utilized to predict the cell line based on the ploidy distribution. The coverage on the x-axis is normalized by the dominant peak position (see Fig. 9), while the occurrence on the y-axis is normalized by the total occurrence of all coverages. The CNN is trained using the normalized values as a 1D signal for prediction, taking into account parameters such as the ploidy of the dominant peak and the peaks used for peak calling.

[00132] The copy-number estimator 124 then determines the correlation between coverage and copy under the optimized fragment size to generate the model parameters 146 (see Figs. 5 & 9 for examples). The median coverage of each qualified peak is considered the expected value. The gaps between the peaks of dominant and sub-dominant ploidy define the 1x coverage. The algorithm employs a regression method, such as the Least Square Method, to evaluate the correlation and aims for a high Pearson correlation coefficient close to 0.999.

[00133] Based on the trend line derived from the correlation analysis as shown in Fig. 5, (and hence, the model parameters 146), the copy-number estimator 124 of Fig. 1 can estimate the most likely copies for a target sequence using a coverage of the target sequence (such as a gene of interest) as found the in the read sequences 151. Thus, the copy-number 124 estimator can estimate a copy number specific to one or more genes of interest.

[00134] To calculate a 95% Confidence Interval for the Gene of Interest (GOI) copy prediction, the following steps may be implemented. Firstly, each main host chromosome is scanned and large segments of different ploidies are obtained, ensuring continuity and consistent ploidy with, in some specific embodiments, 20

units or 5 Mb (see Fig. 4). Next, the median coverage of each segment is divided by the 1x coverage to calculate the expected copies for each segment. For segments with known copies, the coverages for each of the nine fragment sizes are extracted, and the standard deviation (std or sigma) is computed for each size (see Fig. 6 and 8).

[00135] An equation is generated to represent the negative linear correlation between log2(fragment size) and sigma, allowing for the prediction of sigma for fragment sizes among the nine standard fragment sizes with a decimal precision of two for the log2(fragment size). (see Fig. 8) Subsequently, a data table is created, containing predicted sigma values for all segments of the main chromosomes in a specific embodiment. A positive correlation is then identified between copies and sigma, specifically a linear positive correlation between log2(copies) and exp2(sigma).

[00136] Utilizing the equation and data table, sigma for fragment sizes is predicted based on existing copies, with a decimal precision of two for the log2(copies). This prediction is facilitated by a two-dimensional lookup table, enabling the estimation of sigma for any given log2(fragment size) (with decimal precision of two) and log2(copies) (with decimal precision of two) through table lookup. The +/- 95% Confidence Interval for each GOI prediction is calculated by referencing the sigma of coverage in the table, using the log2(copies of GOI) and log2(GOI size).

[00137] The sigma of coverage is then converted to the sigma of copies using the trend line established in step six, which represents the linear correlation between coverage and copies. If the predicted copy of the GOI is less than one, the log2(one copy of GOI) and log2(GOI size) are utilized for the prediction. Otherwise, the log2(predicted copy of GOI) and log2(GOI size) are used for the prediction.

[00138] Referring again to Fig. 1, once the model parameters 146 are generated, the copy-number estimator 124 applies the model to estimate the copy number of the target sequence based on the read coverage stored the read sequences. It utilizes the alignment data obtained from the alignment component 114, including the aligned read positions and coverage information, to predict the copy number of the target sequence. The model parameters are used to transform the observed read coverage into an estimate of the copy number, taking into account the specific characteristics of the PCR-free library preparation workflow. The estimation of the copy number involves applying the model equation, which incorporates the regression coefficients, intercepts, and other relevant factors determined during the model generation

process. By inputting the read coverage from the aligned data into the model equation, the copy-number estimator 124 calculates the estimated copy number of the target sequence.

[00139] The resulting copy number estimation provides valuable information about the abundance and representation of the target sequence within the sample. It facilitates the characterization of the genomic content, such as the presence of integrated transgenes or specific genetic regions. The copy number estimates obtained from the model can be utilized for various purposes, including therapeutic cell line selection, gene expression analysis (e.g., gene expression or transcriptomic may be correlated with DNA copy number and/or the availability of a regulatory sequence may be correlated to the level of gene expression), and genetic engineering applications.

[00140] It will be appreciated that the specific implementation of the copy-number estimator 124 and the modeling techniques described herein may vary. Different statistical, mathematical, or machine learning approaches can be employed to generate the model parameters 146 and estimate the copy number based on aligned read coverage. The configuration, algorithms, and methods utilized by the copy-number estimator 124 may be adapted or customized based on the specific application, advances in technology, and improvements in data analysis techniques.

[00141] In addition to copy number prediction, other analyses are performed on the sequenced reads 105. The reads 105 are analyzed to detect vector integration positions in the host genome. Furthermore, the method can determine the precise location of vector insertion within the host CHO genome by analyzing the global alignment BAM file. Partially mapped reads, referred to as chimeric reads, and their mates are examined, along with reads where one of the paired reads is mapped to the vector and the other read is mapped to the non-vector region. These reads can be *de novo* assembled to generate longer contigs. The assembled contigs are then aligned to the reference genome, which includes both the host genome and vector sequences and to other sequences of interest, to create a BAM file. The host-vector boundaries can be identified using the "SA" tags in the BAM file. In some embodiments, the vector part of assembled contigs is masked to create a second set of contigs (masked set). Both the unmasked and masked contig sets are then mapped to the reference genome. The integration position in the host genome is considered valid if both unmasked and masked contig versions align to the same host genome location.

[00142] Chimeric reads that partially map to the vector and partially map to non-vector regions, and reads where one of the paired reads maps to the vector and the other reads map to non-vector regions, are examined. *De novo* assembly is performed on these reads to generate longer contigs, which are then aligned to the reference genome to identify donor DNA-host boundaries and insertion points within the host genome. Repeat masker annotation is used to improve the accuracy of identifying insertion points in repeat regions. The method can involve detecting vector concatenation events and identifying the precise location of vector insertion in the host CHO genome.

[00143] That is, to detect vector concatenation events, sequencing data is analyzed for junction reads, which are reads that contain two segments aligning to different parts of the vector. In some embodiments, sequencing reads aligned to the vector reference sequence are extracted from the global alignment file. These extracted reads are then re-aligned to the vector reference sequence using consed/phrap software or other commonly used read aligning software. This re-alignment process aims to identify vector chimeric reads, which are reads aligning to different non-homologous parts of the vector sequence. A junction position is considered valid if it is supported by four or more different chimeric reads.

[00144] Sequence variants in the integrated vector are also detected by analyzing the sequencing reads 105. The pysam package is utilized to collect variants at the read level, and a series of screening steps are applied to filter out low-quality variants. Filtering criteria consider variant frequency, quality score, read support from both plus and minus strands, and distance from potential junctions. This filtering process helps eliminate false-positive variant calls and ensure high-quality variant identification. Variations in the NGS reads generated during the sequencing process, such as single nucleotide mutations, small insertions, or deletions, may exist compared to the reference vector sequence. These variations could be real mutations or sequencing errors introduced during library preparation, particularly if PCR amplification is involved. To identify these variants, the method utilizes the pysam package to collect variants at the sequencing read level. Subsequently, a series of screening steps is implemented to identify true sequence variants. First, reads with about 5% of bases as variants or more are excluded. Then, in this specific embodiment, only sequence variants with a Phred quality score greater than 15 are considered, ensuring a high level of confidence in the variant call. Additionally, the final variants may be supported by reads from both the plus and minus strands, originate from different read pairs, and have about 8 or more base pairs from the

nearest junction or structural variant. These filtering criteria, tailored for the PCR-free library preparation workflow, help eliminate false-positive variant calls and enhance variant calling accuracy.

[00145] Finally, a Circos plot can be generated to visualize and analyze various features of the integrated vector for display on a GUI component 118. The plot includes tracks for gene-of-interest (GOI) positions, read coverages, repeat regions, split read junctions indicating potential vector concatenation events, and variants. This visual representation aids in the interpretation and analysis of the sequence data, providing insights into the structure and characteristics of the integrated vector and its relationship with the GOIs. For visualizing and analyzing various features of integrated donor DNA, such as read coverage anomalies, vector repeat regions, concatenated vector sequence regions, vector sequence variants, and host genome integration locations, a circus plot is employed. In one embodiment, the circus plot is created using Python and the R scripts, enabling comprehensive analysis and visualization of the integrated vector sequence within the CHO genome.

[00146] In summary, the process of sequencing the sample 101 involves preparing a PCR-free shotgun sequencing library, performing high-throughput sequencing using a genetic sequencer 103, and generating short reads 105 representing the genetic material in the sample. These reads 105 are then subjected to bioinformatics analyses, including alignment to the reference genome, determination of copy numbers, detection of vector integration positions, identification of sequence variants, and generation of visual representations like the Circos plot. These analyses provide detailed information about the integrated transgene clones, their copy numbers, structural variations, and sequence variants, contributing to a comprehensive understanding of the genetic characteristics of the sample.

[00147] Referring again to Fig. 1 in some embodiments, a specialized application for interfacing with a genetic sequence analyzer component 112 may be used, such as a mobile application on the mobile device 106 or a desktop application on the personal computer 104. The communications may include transmitting data in HTML, XML, JSON, YAML, or any data format. The genetic sequence analyzer component 112 may provide user-level accounts to individuals through a typical login mechanism. The genetic sequence analyzer component 112 may be a web application, a webserver, a web service, etc. and may utilize one or more protocols to communicate data.

[00148] The cloud-service provider 102 may provide the genetic sequence analyzer component 112 as a webpage, a webapp, a program for download and execution on the computer 104 or the mobile device 106. The genetic sequence analyzer component 112 includes the copy-number estimator 124, a communications component 116, a GUI component 118, and a process model executer 120.

[00149] As previously mentioned, in yet additional embodiments, the copy-number estimator 124 may assign a confidence score using Monte Carlo simulation data. In one embodiment, a frequentist confidence score may be derived using frequentist statistics. For example, a confidence score may use sample data of a distribution, hypothesis testing, p-values, significance testing, confidence intervals etc. In additional embodiments, a confidence score is calculated for each (or a set of) using posterior probabilities in a Bayesian estimate, which represent the updated belief about the copy number value estimates. Thus, the confidence score, for example, may be a credible interval of a posterior distribution or of a Bayesian estimator.

[00150] The process model executer 120 may execute one or more of the stored models to determine the copy number for one or more subcomponents and/or execute the alignment component 114. The process model executer 120 may be executable code configured to execute, interpret, or utilize the models, for example using a virtual processor 124, in a manner to report the copy numbers to the genetic sequence analyzer component 112.

[00151] The genetic sequence analyzer component 112 also includes the communications component 116. The communications component 116 may facilitate seamless communication and data exchange between multiple software applications, devices, and systems. That is, the communications component 116 may include protocol handling, message formatting, data serializing, encryption, and authentication to facilitate the communication with the computers 104 and/or the mobile device 106. The communications component 116 may utilize a message formatting mechanism to format the messages into formats, such as XML, JSON, binary formats, and/or proprietary message formats. The communications component 116 may utilize various encryption algorithms, such as RSA, AES, ECC, symmetric encryption, asymmetric encryption etc. to enable secure communications between the genetic sequence analyzer component 112 and the computers 104 and/or the mobile device 106.

[00152] The genetic sequence analyzer component 112 also includes a real-time data ingestor 123. The real-time data ingestor 123 may be a real-time or near real-

time ingestor configured to collect and collate read sequences 151 received via the sequencer 103. The read sequences 114 may be stored in the database 132. A user account 150 may be associated with the reference sequence 142, the partitioned sequences, the model parameters, and/or the read sequences 114.

[00153] The GUI component 118 can render a display for use by the computer 104 and/or the mobile device 106. The GUI component 118 may be a webpage-based provider, such as flask, an HTML server, a web framework, etc. The GUI component 118 may provide widgets, information, buttons, options, and menus to thereby facilitate a user's interaction with the genetic sequence analyzer component 112.

[00154] The GUI component 118 can be used to log into user accounts 150 so that a user can create, save, or retrieve data from the database 132 or otherwise interface with any account features. Additionally or alternatively, the GUI component 118 can save favorites, select default parameters, or adjust plotting parameters. The GUI component 118 can direct other components to execute instructions based upon a workflow initiated by a user. That is, the GUI component 118 may receive events, such as a mouse click, button press, or GUI widget interaction to initiate a routine, series of steps, or series of acts. For example, the GUI component 118 may guide a user step-by-step on how to set up and work with the process models 144 within the database 132.

[00155] The GUI component 118 may also be used to visualize the results of the copy-number estimator 124, in aggregate, in simulation, and/or may provide various visualization tools to analyze the data. Thus, each user can log into a user account 150 to visual the results of their analysis.

[00156] A resource dispatcher 110 may dispatch requests to perform an action to one or more virtual servers 122, each of which has a virtual processor 124, a virtual memory 126, and a virtual disk space 128. The virtual servers 122 can be executed on one or more servers 121 on a server farm 119 as dispatched and activated by the resource dispatcher 110.

[00157] Fig. 2 show a block diagram illustration of a computing device 200 to to estimate copy number of a target sequence (e.g., a target genetic sequence) by creating and utilizing a model in accordance with embodiment of the present disclosure. The computing device 200 of Fig. 2 may be the computer 104 or mobile device 106 of Fig. 1. The computing device 200 includes an I/O interface 210 to communicate therewithin. The computing device 200 includes a data store 204, a processor 206, a network interface 208, a memory 225, and user I/O devices 226.

The data store 204 stores data and may be a hard drive, flash drive, thumb drive, volatile memory, non-volatile memory, semi-volatile memory etc. The processor 206 can execute one or more processor-executable instructions 212, which may be stored in the data store 204 and/or the memory 225. For example, the processor 206 can execute processor-executable instructions 212 stored in memory 225 that was retrieved from the data store 204. The memory 225 also includes program data 214 that may include information related to the processor-executable instructions 212. The computing device 200 may include user I/O devices 226, such as a cursor device 230 (e.g., touchscreen or mouse), a keyboard 232 (virtual or physical), and/or a monitor 228 (which may be a touchscreen). The computing device 200 communicates with the network 202 via a network interface 208.

[00158] Although the computing device 200 of Fig. 2 may be used as part of the system 100 of Fig. 1, in some embodiments, the environmental impact minimization functionality may reside wholly within the computing device 200 of Fig. 2. For example, the genetic sequence analyzer component 112 of Fig. 1 may reside within the processor-executable instructions 212 of Fig. 2 as the genetic sequence analyzer component 242. For example, the alignment component 234, the communications component 236, the GUI component 238, the process model executer 240, the sequence read ingestor 241, and the copy-number estimator 243 of Fig. 2 may have the same or similar functionality as the alignment component 114, the communications component 116, the GUI component 118, the process model executer 120, the sequence read ingestor 123, and the copy-number estimator 124 of Fig. 1, respectively. The database 244 may be similar to the database 132 of Fig. 1. The database 244 may, for example, be an SQLite 3 database embedded on the computing device 200.

[00159] Thus, in some embodiments the genetic sequence analyzer component 242 may reside wholly on a local device (such as on the computers 104, the mobile device 106, etc.) may be partially within a cloud service provider 102, and/or may be organized in a hybrid local and cloud configuration. In some embodiments, the genetic sequence analyzer component 242 may be an application, may be executed on the computers 104, the mobile device 106, the cloud service provider 102, the genetic sequencer 103, the computing device 200, etc. or some combination thereof.

[00160] The database 244 of Fig. 2 may be like or identical to the database 132 of Fig. 1. That is, the reference sequence 246, the partitioned sequence 248, the model parameters 250, the user accounts 254, and the read sequences 252 may be similar

or identical to the reference sequence 142, the partitioned sequence 144, the model parameters 146, the user accounts 150, and the read sequences 151 of Fig. 1, respectively.

[00161] Fig.10 shows a flow chart diagram of a method 1000 for estimating the copy number of genetic sequences, such as integrated transgenes in CHO cell lines used for recombinant protein production, in accordance with an embodiment of the present disclosure. The method 1000 begins by isolating and purifying a genomic sample (Acts 1002-1004). A PCR-free shotgun library is then created from the sample (Act 1006) and sequenced to generate reads (Act 1008). The reads are aligned to a reference genome composed of the host reference sequence (e.g. CHO) and any vector sequences (Act 1010). The host genome is partitioned into fragments of varying sizes, such as 50bp to 12800bp (Act 1012). The coverage of each fragment size is analyzed to identify consecutive coverage peaks with equal distances between them, corresponding to ploidy levels (Act 1014). Respective copy numbers are assigned to each peak based on the coverage distances (Act 1016). Finally, a model is generated to correlate coverage and copy number (Act 1018). This model can then estimate the copy number of any target sequence, such as transgenes, based on its observed coverage. The method offers several advantages over existing approaches. The PCR-free library preparation reduces biases. The combination of genome partitioning and ploidy analysis enables accurate copy number prediction without prior knowledge of the target location in some specific embodiments. The model provides sequence-independent estimates. Overall, the approach delivers rapid, accurate, and cost-effective copy number and molecular characterization of cell lines used for biopharmaceutical production.

[00162] Variations may utilize different fragment sizes, peak numbers, regression models, confidence interval calculations, quality control analyses, and visualizations. The method facilitates precise copy number determination of integrated vectors and transgenes, detection of rearrangements, concatenation events, variants, and integration sites. It represents a significant improvement over traditional Southern blot and dPCR methods for cell line characterization.

[00163] Refer to Fig. 10 for additional embodiments of the method 1000 as follows. Act 1002 involves isolating a sample of genomic nucleic acids from the biological sample to be characterized. This provides the raw genomic material for subsequent analysis. The sample may contain homogenous cell types or a heterogeneous mixture of various cell types, e.g., such as a clonal cell line expressing recombinant

proteins. To isolate the nucleic acids, the sample first undergoes cell lysis to break open the cell membranes and release the contents. Cell lysis can be achieved through various methods, including chemical lysis, enzymatic lysis, mechanical lysis via bead-beating, or thermal lysis.

[00164] Once the cells are lysed, the crude lysate contains a complex assortment of biomolecules, including genomic DNA, RNA, proteins, lipids, and metabolites. To isolate the genomic DNA specifically, a series of extraction and purification steps are employed. Potential extraction methods include phenol-chloroform extraction, silica column-based extraction, magnetic bead-based extraction, and extraction kits utilizing proprietary reagents.

[00165] Preferably, the extraction protocol may maximize DNA yield while minimizing shearing. The protocol may involve steps to separate out proteins and other contaminants, employing enzymes like proteinase K or RNase A to digest proteins and RNA. Repeated ethanol precipitation cycles concentrate the DNA while washing away impurities.

[00166] The isolated genomic DNA may be examined for quality and quantity. Quality assessment via gel electrophoresis or fluorometric methods may be used to determine DNA integrity, size distribution, and potential RNA/protein contamination. DNA quantification may utilize spectrophotometric methods like UV absorption at 260nm or fluorometric assays using dyes that selectively bind DNA.

[00167] The quantity of starting genomic DNA depends on downstream library preparation and sequencing methods. For PCR-free shotgun library preparation, hundreds of nanograms to micrograms of DNA may be used.

[00168] If the initial isolation produces low DNA yield or poor quality, optimized protocols may be used to improve output. Potential optimizations include adjusting cell number, lysis buffers, purification reagents, precipitation parameters, and extraction kits.

[00169] Once sufficient quantity and quality of genomic DNA is obtained, it undergoes further preparation and quality control before shotgun library construction. This can include additional purification, size selection via gel extraction or SPRI beads to remove small DNA fragments, and examination of DNA size distribution. The properly isolated genomic DNA is used for the input for Act 1004.

[00170] Act 1004 involves purifying the sample of genomic nucleic acids that was isolated in Act 1002. This purification step can remove contaminants and enzyme inhibitors that could interfere with subsequent library preparation and sequencing.

[00171] The purification is performed using various methods such as silica-based columns, magnetic beads, filter plates, or precipitation techniques. Commercial kits utilizing these technologies are available, such as Beckman Coulter's Agencourt AMPure XP kit, New England Biolabs' Monarch Genomic DNA Purification kit, Zymo Research's DNA Clean and Concentrator kit, among others.

[00172] The purified DNA sample may undergo additional quality control steps like quantification, purity checks using spectrophotometry or fluorometry, and integrity analysis using agarose gel electrophoresis. Quantification determines DNA concentration while purity checks assess contaminants like proteins, RNA, or salts. Gel electrophoresis verifies DNA integrity by ensuring high molecular weight fragments are present.

[00173] During purification, steps are taken to avoid shearing and fragmenting the DNA. Gentle handling, wide-bore tips, and avoidance of excessive vortexing help maintain high molecular weight DNA. The purification parameters like buffer conditions and bead-to-sample ratios are optimized to maximize DNA recovery.

[00174] Purified DNA is eluted in low salt buffers like 10mM Tris-HCl or TE buffer. Elution volumes are adjusted to achieve desired DNA concentrations for the next library preparation steps.

[00175] For PCR-free library preparation as described herein, in some embodiments, PCR inhibitors like phenol, ethanol, salts, or metal ions may be avoided. Additional wash steps may be used to thoroughly remove any co-purified inhibitors.

[00176] The purified DNA sample may be quantified by fluorescent dsDNA binding dyes like Qubit or PicoGreen. It is quality checked using spectrophotometric ratios like A260/A280 and A260/A230 to verify protein and solvent contamination are within acceptable limits. DNA integrity is checked by running aliquots on agarose gels to confirm high molecular weight DNA is present.

[00177] Act 1006 involves creating a shotgun library from the purified sample of genomic nucleic acids. This act prepares the nucleic acid sample for sequencing in the subsequent steps.

[00178] To create the shotgun library, the genomic DNA sample first undergoes fragmentation into smaller pieces through mechanical shearing or enzymatic digestion. The size of the DNA fragments is optimized based on the sequencing technology to be used. For Illumina sequencing, fragment sizes of 200-500 bp may be used.

[00179] The fragmented DNA is then subjected to end-repair to generate blunt ends. The 5'-phosphorylated blunt ends may be used for adaptor ligation in the next acts. Optionally, size-selection of the end-repaired DNA fragments may be performed using agarose gel extraction to obtain fragments of the predetermined size range.

[00180] Adapter oligonucleotides are then ligated to both ends of the size-selected DNA fragments. The adapters provide priming sites for amplification and sequencing of the library fragments. Different adapter sequences with distinct molecular barcodes can optionally be used to allow sample multiplexing.

[00181] Following adapter ligation, the ligation products may go through several clean-up and size-selection steps to remove unligated adapters, enzymatic reagents, and small DNA fragments. This improves the quality of the sequencing library.

[00182] In some protocols, the adapter-ligated fragments may be amplified through PCR to enrich for fragments with adapters on both ends and increase library yield. However, PCR amplification can introduce biases and errors, so some PCR-free methods omit this step.

[00183] Overall, this shotgun library preparation workflow randomly fragments the genome and adds platform-specific adapters to generate a complex library representing the genome or target regions. This unbiased library is compatible with downstream high-throughput sequencing, enabling the parallel sequencing of millions of DNA fragments.

[00184] The specific reagents and kits used for fragmentation, end-repair, size selection, adapter ligation, and clean-up can vary based on the sample type and sequencing platform. Both automated and manual options for library prep workflows are available. Overall, Act 1006 can create a sequencing-ready shotgun library from the genomic DNA sample.

[00185] Act 1008 involves sequencing the shotgun library created in Act 1006 using the genetic sequencer 103 to generate sequencing reads 105 representing the sample's genomic content. This act may utilize high-throughput next generation

sequencing (NGS) technologies to parallel sequence millions or billions of DNA fragments from the shotgun library.

[00186] In one embodiment, Act 1008 performs sequencing by synthesis (SBS) using reversible dye terminators. The shotgun library fragments are amplified clonally on the surface of a flow cell, forming separate clusters. Fluorescently labeled nucleotides are then added one by one, with imaging performed after each incorporation to identify the added base. This cycle is repeated to determine the sequence of bases in each fragment, generating reads of lengths such as 100bp or 150bp.

[00187] Illumina sequencing platforms such as NovaSeq, NextSeq, MiSeq etc. may be used in Act 1008 to perform the high-throughput sequencing. These instruments offer ultra-high throughput with outputs ranging from megabytes to terabases of data, ensuring comprehensive sampling and coverage of the shotgun library. Kits and reagents optimized for PCR-free library preparation workflows may be utilized for sequencing.

[00188] In another embodiment, Act 1008 may use DNABSEQ sequencing, a tagmentation-based approach offered by Illumina. Here, the shotgun library undergoes an enzymatic fragmentation and tag attachment process called tagmentation. This enables simultaneous fragmentation and tagging of the DNA fragments prior to clonal amplification and sequencing.

[00189] Act 1008 results in the generation of many sequencing reads 105 that act as a representation of the sample's 101 genomic makeup. The reads 105 may be paired-end reads with lengths of 100-250 bp, offering sequence information from both ends of each library fragment. The depth of sequencing is tuned based on the size and complexity of the reference genome to ensure optimal coverage across regions of interest.

[00190] The raw read data 105 from Act 1008 may be assessed for quality metrics such as base quality scores, GC content, kmer diversity etc. to ensure compatibility with downstream analysis. In some embodiments, it may assumed that the sequencer output generates qualified data given that each sample will have a large number of reads. Additionally or alternatively, in some embodiments, no fastqc is performed before the bioinformatics pipeline. In yet additional embodiments, mapping metrics in the final result may be utilized and/or an unqualified sample may be examined. Overall, Act 1008 provides sequencing data that forms the foundation for

subsequent bioinformatics workflows to determine copy numbers, identify variants, or characterize integrated vectors.

[00191] Act 1010 involves aligning, using a processor, the sequencing reads generated in Act 1008 to a reference sequence formed from a reference genome and a vector sequence.

[00192] The alignment is performed using an optimized aligner algorithm, such as BWA, BWA-MEM, Sentieon BWA, Bowtie2, etc. to efficiently map the short sequencing reads to the reference genome. Special considerations may be taken into account for the PCR-free library preparation workflow used to generate the sequencing data in Act 1006, in some embodiments.

[00193] Specifically, in one embodiment, the alignment process involves the following detailed steps:1. The raw sequencing reads from Act 1008 are retrieved and preprocessed. This includes filtering low quality reads, trimming adapter sequences, and removing duplicate reads. 2. The preprocessed reads are then aligned to the combined reference sequence using BWA, BWA-MEM, Sentieon BWA, Bowtie2, etc. algorithm with optimized parameters. The reference sequence comprises the host reference genome (e.g. CHO genome) concatenated with the full vector sequence. 3. Multiple threading and GPU acceleration techniques are employed to significantly improve the speed of the read alignment process. This enables alignment of hundreds of millions of reads within a reasonable timeframe. 4. Proper paired-end alignment is enforced for reads generated from longer DNA fragments, requiring both ends to align concordantly. This mitigates the occurrences of spurious alignments. 5. Alignment quality scores are recalibrated to correct for biases and errors that may have been introduced during sequencing. This helps improve the accuracy of downstream analysis. 6. PCR duplicate reads may be collapsed to avoid overestimation of coverage. This may be used, in some embodiments, to account for duplication artifacts inherent in PCR amplification. 7. Reads may be realigned around indels using the RealignerTargetCreator and IndelRealigner tools from GATK to clean up misalignments. 8. The resulting BAM file contains the aligned positions of all reads against the reference genome, along with metadata like mapping quality and flags. This serves as input for downstream copy number analysis. 9. Quality control checks may be performed to ensure satisfactory alignment rate, proper pairing rates, expected insert sizes based on library prep, reasonable coverage distribution, and lack of contamination. 10. For improved structural variation detection, longer reads generated from technologies like Oxford

Nanopore can be incorporated into the alignment pipeline. In summary, Act 1010 performs alignment of sequencing reads to the combined reference genome and vector sequence. The output provides the foundation for accurate characterization of copy number, variants, and integration sites.

[00194] Act 1012 involves partitioning, using the processor, the reference genome into a plurality of sizes. This act is performed by the copy-number estimator component after the read alignment step.

[00195] The partitioning allows the analysis of read coverage across the reference genome in fragments of different sizes. This enables the determination of the relationship between read coverage and copy number, which may be used developing an accurate copy number estimation model.

[00196] Specifically, the reference genome is divided into non-overlapping partitions of varying sizes, ranging from 50 base pairs (bp) to 51,200 bp. In one embodiment, the partition sizes include 50 bp, 100 bp, 200 bp, 400 bp, 800 bp, 1,600 bp, 3,200 bp, 6,400 bp, 12,800 bp, 25,600 bp, and 51,200 bp.

[00197] The copy-number estimator fetches coverage information for each partition size by extracting the relevant data from the alignment file containing the mapped sequencing reads. The coverage represents the number of sequenced reads that align to each partition.

[00198] By analyzing the coverages across different partition sizes, the copy-number estimator can determine trends and patterns related to the correlation between coverage and copy number. Smaller partition sizes help assess coverage variability and noise, while larger partitions are more useful for discerning copy number differences.

[00199] In one implementation, the partition sizes of 800 bp, 1,600 bp, 3,200 bp, 6,400 bp, 12,800 bp, 25,600 bp, and 51,200 bp are utilized by the copy-number estimator to estimate the copy number based on the coverage peaks. The smaller 50 bp to 400 bp partitions are leveraged to estimate coverage variability and standard deviation for determining confidence intervals.

[00200] The partitioning enables tuning the fragment size to optimize peak resolution in the coverage-based ploidy distribution analysis. It also allows focused analysis of coverage for specific genes or regions of interest.

[00201] Overall, the act of partitioning facilitates determination of the relationship between coverage and copy number by enabling fragmentation of the reference

genome into segments of different sizes and evaluation of coverage across the various partition sizes. This supports the development of an accurate model for estimation of the copy number of target sequences.

[00202] Act 1014 involves determining, using the processor, a mean read coverage for each size of the partitioned reference genome until a predetermined number of consecutive mean-read coverage peaks form with an equal coverage distance between the consecutive mean-read coverage peaks.

[00203] In more detail, the processor partitions the reference genome into a plurality of fragment sizes ranging from 50bp to 51,200bp, including 50bp, 100bp, 200bp, 400bp, 800bp, 1600bp, 3200bp, 6400bp, 12800bp, 25600bp, and 51200bp. The processor fetches the aligned read coverages for each fragment size from the alignment output stored in memory or disk.

[00204] The processor calculates the median coverage value for each fragment size to determine the mean read coverage. The median, rather than mean, may be used to minimize the effect of outliers.

[00205] The processor analyzes the mean read coverage across the spectrum of fragment sizes. It identifies consecutive peaks in the coverage where the distance between the peaks is approximately equal. For example, three consecutive peaks may be detected at coverages of 100X, 200X, and 300X, indicating an equal coverage distance of 100X between peaks.

[00206] The number of consecutive equal-distance peaks is predetermined, for example set to three peaks. The processor continues partitioning the reference genome into progressively larger sizes, determining the mean read coverage for each size, until the predetermined number of equal-distance coverage peaks is achieved.

[00207] Once the predetermined number of peaks with equal coverage distance has been identified, the associated fragment size is used for subsequent copy number determination. For example, if three equal-distance peaks are found at a fragment size of 12800bp, then 12800bp is utilized for estimating copy numbers.

[00208] In some embodiments, the 12800bp size represents the optimal balance between capturing integral copy numbers with distinct peaks and minimizing noise or variability in coverage. Smaller fragment sizes exhibit more variability while larger sizes do not resolve distinct copy numbers well. In some embodiments, some cell line may be unique and thus use a different fragment size, e.g., such as 51200bp.

Another factor is the sequencing depth - lowly sequenced sample may use larger fragment size. However, in some embodiments, the assay can generate relative consistent sequencing depth (such as 300-500million reads per sample); thus, cell line specificity may be considered with or without the sequencing depth.

[00209] In some embodiments, the processor may optimize or tune the fragment size to improve peak resolution in the coverage spectrum. For example, the processor may evaluate fragment sizes at 10bp intervals surrounding 12800bp, such as 12790bp, 12800bp, 12810bp, and determine which size produces the most distinct equal-distance peaks for improved copy number estimation.

[00210] The resulting optimized fragment size, number of observed peaks, and equal peak coverage distance are then used by the processor to estimate the copy number and generate a model correlating coverage to copy number, as described in Act 1016 and Act 1018.

[00211] This detailed multi-step process of adaptive fragment sizing, mean coverage calculation, equal-distance peak identification, and fragment size tuning enables robust and accurate estimation of genomic copy numbers from the aligned sequence read data. The algorithmic approach provides a data-driven method for selecting optimal parameters tailored to the specific sample sequencing results.

[00212] Act 1016 involves determining, using the processor, a respective copy number associated with each of the predetermined number of consecutive mean-read coverage peaks.

[00213] In one embodiment, this determination is performed by analyzing the equal coverage distance between the consecutive mean-read coverage peaks and comparing it to the coverage value of each peak. For example, if there are three consecutive peaks with equal coverage distances between them, the first peak is considered to represent a copy number of 1x. This is done by comparing the coverage value of the first peak to the equal coverage distance - if they are approximately equal, then the first peak is assigned a copy number of 1x.

[00214] The second peak is then assigned a copy number of 2x, since its coverage value is approximately twice the equal coverage distance. Similarly, the third peak is assigned a copy number of 3x. In this way, the respective copy number of each peak can be deduced based on its coverage value relative to the equal coverage distance between peaks.

[00215] In another embodiment, the determination of copy numbers utilizes a linear regression model that correlates coverage values to copy numbers. The model is trained using the mean coverage values of partitioned reference genome fragments with known copy numbers. This allows the generation of a trendline that captures the relationship between coverage and copy number. Using this trendline equation, the copy number corresponding to each peak's coverage value can be estimated.

[00216] Additionally, quality control analyses may be performed on the read coverage data to remove potential sequencing biases and errors. Thresholds can be set to filter out peaks arising from coverage anomalies. The remaining high-confidence peaks are then used for copy number estimation.

[00217] The copy number analysis may also incorporate factors such as the variability in coverage across different fragment sizes and the estimated reliability of coverage measurements. Statistical techniques can determine confidence intervals for each copy number prediction based on the fragment size and observed coverage distribution.

[00218] To further refine the copy number estimates, the ploidy distribution of the genome may be analyzed by plotting coverage against occurrence for different fragment sizes. The location and width of ploidy peaks in the distribution can provide insights into the true copy state. For example, the dominant peak often represents the main copy number.

[00219] In summary, Act 1016 analyzes the read coverage data to deduce the respective copy number of consecutive coverage peaks. This is achieved through comparison to equal coverage distances, application of trained regression models, quality control filtering, confidence interval calculation, and ploidy distribution analysis. The resulting copy number profile facilitates accurate estimation of the copy state for target gene sequences.

[00220] Act 1018 involves determining, using the processor, a model configured to estimate a copy number of a target sequence. This model is generated by leveraging the correlation between read coverage and copy number observed in the host genome.

[00221] Specifically, the processor fetches coverage information for partitioned fragments of varying sizes across the host genome, ranging from 50bp to 51,200bp in exponential increments. The processor calculates the median coverage for each fragment size and plots the coverage distribution. A series of evenly spaced

consecutive peaks in the coverage plot are identified, indicating the dominant ploidy states of the host genome.

[00222] The distance between the peaks corresponds to the expected 1x coverage. By comparing the peak positions to the 1x coverage, the processor assigns putative copy numbers to each peak. This establishes a correlation between observed coverage and copy number for the host genome.

[00223] The processor then performs regression analysis, such as least squares regression, to fit a trendline representing the mathematical relationship between coverage (independent variable) and copy number (dependent variable). The resulting regression formula serves as the model for predicting copy numbers based on observed coverage.

[00224] To estimate the copy number of a target sequence, such as a gene-of-interest (GOI) within the vector, the processor determines its coverage from the aligned reads. This measured coverage is input into the regression model equation to calculate the predicted copy number of the target sequence.

[00225] The model is trained and optimized using the coverage and known copy number of host genome fragments. In some specific embodiments, it does not require the genomic location of the target sequence. This enables unbiased copy number prediction for any target sequence based solely on its observed coverage.

[00226] Additionally, the processor may estimate confidence intervals for the copy number prediction using the variability in coverage across different fragment sizes of the host genome. The fragment coverage deviations can be correlated with copy number to calculate the prediction error margins at a predetermined confidence level.

[00227] The coverage-to-copy number model may be further refined by determining the ploidy distribution of the host genome. The processor plots the frequency of observed coverages to reveal distinct ploidy peaks. It then adjusts the initial coverage-to-copy ratios based on the dominant peak positions. This improves the accuracy of the model for subsequent copy number estimations.

[00228] In summary, Act 1018 involves developing a mathematical model representing the relationship between observed sequence coverage and copy number. This model is applied to estimate the copy number of any target sequence based solely on its measured coverage derived from the aligned sequencing reads. In some specific embodiments, the model does not require prior knowledge of the target sequence location in some embodiments of the present disclosure.

EXAMPLES

[00229] Refer to Figs. 1, 11, and 12 for an example of using the system 100 of Fig. 1 as follows: In some embodiments, a DNA library of fragments is prepared using a shotgun PCR-free DNA library preparation workflow. In some embodiments, PCR-free DNA libraries are prepared using a kit selected from the group consisting of: Illumina® DNA PCR-Free Prep, Tagmentation kit; TruSeq® DNA PCR-free kit; PerkinElmer NEXTFLEX® Rapid XT Kit; and Qiagen QIAseq® FX DNA Library Kit

[00230] In some embodiments, the input test material for generating a next generation sequencing (NGS) library comprised genomic DNA (gDNA) isolated from immunoglobulin-producing Chinese hamster ovary (CHO) clonal cell lines. Alternatively, in some embodiments, the gDNA is isolated from Homo sapiens cell lines engineered to produce a therapeutic protein. In some embodiments, the engineering of CHO clonal cell lines had the following process: plasmid transfection, plate mini-pools (5000 cells/well, Gln- medium), static assay in 96-well plates, fed-batch productivity assay, clone by limiting dilution (0.5 cells/well), static assay in 96-well plates, fed-batch productivity assay, or stability study and feed evaluation. In some embodiments, a gDNA sample of 300 to 2000 nanograms (300 ng-2000 ng) was used in the library construction. In some embodiments, library preparation was based on a 300 ng sample of gDNA.

[00231] In some embodiments, the method of library preparation includes: 1. Tagmenting genomic DNA (Hands-on: 5 minutes, Total: 10 minutes); 2. Post tagmentation clean-up (Hands-on: 10 minutes, Total: 20 minutes); 3. Ligate indexes (Hands-on: 5 minutes, Total: 45 minutes); 4. Clean-up libraries (Hands-on: 15 minutes, Total: 45 minutes); 5. Quantify and pool libraries; 6. Dilute to starting concentration (Reagents: RSB); and 7. Sequencing set-up.

[00232] In some embodiments, modifications to step 4 (second clean-up step) of the method enabled adjustment of the library insert size. In some embodiments, final cleanup (step 4) and size selection impact the resultant library. In some embodiments, double-sided solid-phase reversible immobilization (SPRI) cleanup and size selection discarded low and high molecular weight fragments, reducing yield but tightening the size distribution. In some embodiments, adjusting the volume of sample purification beads (SPB) during double-sided SPRI cleanup allows for fine-tuning the final library insert size. Is some embodiments, fragment size distribution within the library is sensitive to the sample preparation beads (SPB) volume,

especially during the first cleanup. Increasing the SPB volume in one cleanup step decreases the median fragment size. While an insert size of around 350 bp is optimal for many applications, longer insert sizes was observed to allow more robust integration site detection.

[00233] In some embodiments, the DNA library is sequenced by NGS. In some embodiments, the NGS uses sequencing by synthesis (SBS) technology. In some embodiments, the DNA library is sequenced using the Illumina® NextSeq™ 2000 and DNABSEQ™ G400 instruments (MGI Tech). In some embodiments, sequencing produces about 400 to 2,000 million reads from the DNA library.

[00234] In some embodiments, reads are aligned to the genome of a control organism for genetic stability analysis. In some embodiments, the control organism is selected from CHO or Homo sapiens reference genomes and to the inserted vector sequence.

[00235] In some embodiments, the NGS reads generated during the sequencing process may not perfectly match the reference vector sequence. In some embodiments, small variants, such as single nucleotide mutations, small insertions, or deletions, are present in the sequencing reads. In some embodiments, the variants are mutations or sequencing errors introduced during the library preparation if PCR amplification is used. In some embodiments, the PCR-free library preparation workflow used herein allows to apply filtering algorithms that reduce possible sequencing errors.

[00236] In some embodiments, the NGS library is loaded onto a NextSeq® 2000 instrument, which performs sequencing and primary data processing to generate Fastq reads. The resulting Fastq files are then transferred to the SB platform for further analysis using a bioinformatics pipeline. In some embodiments, the output of the pipeline includes at least one selected from the group consisting of the copy number(s) of the gene(s) of interest, vector integration coordinates, and vector sequence variants.

Example 1. Preparing the input materials

[00237] The input test material for generating a next generation sequencing (NGS) library comprised genomic DNA (gDNA) isolated from immunoglobulin-producing Chinese hamster ovary (CHO) clonal cell lines. The engineering of CHO clonal cell lines had the following process: plasmid transfection à plate mini-pools (5000 cells/well, Gln- medium) à static assay in 96-well plates à fed-batch productivity

assay à clone by limiting dilution (0.5 cells/well) à static assay in 96-well plates à fed-batch productivity assay à stability study and feed evaluation.

[00238] All test samples used in this assay development had been well characterized and had varying numbers of vector copy integration determined by droplet digital PCR (ddPCR). The plasmid sequences used for developing individual clonal cell lines are provided in Table 1 below.

Table 1

CHO Clone	Vector Name	Length, bp
Clone27	CLE209	11489
Clone89	CLE209	11489
NIST6 P0	pCGS3.2-NIST	11034
NIST6 P20	pCGS3.2-NIST	11034
NIST2 P0	pCGS3.2-NIST	11034
NIST2 P20	pCGS3.2-NIST	11034
NIST15 P0	pCGS3.2-NIST	11034
NIST15 P20	pCGS3.2-NIST	11034
NIST21 P0	pCGS3.2-NIST	11034
NIST21 P20	pCGS3.2-NIST	11034
NIST31 P0	pCGS3.2-NIST	11034
NIST31 P20	pCGS3.2-NIST	11034
R14 P0	CLE426	15803
R14 P20	CLE426	15803
R18 P0	CLE426	15803
R18 P20	CLE426	15803
P41 P0	CLE426	15803
R41 P20	CLE426	15803
S46 P0	CLE459	16961
S46 P20	CLE459	16961
S49 P0	CLE459	16961
S49 P20	CLE459	16961
H124 P0	UCOE-Her	15732
H124 P20	UCOE-Her	15732

[00239] DNA was extracted from engineered CHO cell lines using the QIAGEN® EZ1 DNA Tissue Kit according to the procedures outlined in OPRD1046. No exogenous controls were added to the test samples prior to extraction. The purified DNA was eluted in the buffer provided with the kit, and the quantity was assessed by Qubit Fluorimeter (ThermoFisher), following the manufacturer's instructions. A minimum of 300 to 2000 nanograms (300 ng-2000 ng) of DNA from each test sample, per library, was required to perform the library construction using the Illumina® DNA PCR-Free Library Prep, Tagmentation Kit in this example. As shown in Table 2, sufficient material was extracted from each test sample to proceed with DNA library construction.

Table 2

SAMPLE ID	gDNA Stock Conc. [ng/µL]	Library Yield [ng/µL]	Library Yield [nM]	Library Batch	Library Kit Used
P20NIST31	116	1.766	6.35	Lib030221	Illumina DNA PCR Free
P20NIST15	66	0.89	1.53	Lib031021	Illumina DNA PCR Free
P20NIST13	104	0.89	3.2	Lib031921	Illumina DNA PCR Free
P20NIST6	202	3.42	10.942	Lib032321	Illumina DNA PCR Free
P20NIST21	94	2.566	8.99	Lib032521	Illumina DNA PCR Free
Clone27	94.61	2.906	10.451	Lib032521	Illumina DNA PCR Free
P0NIST21	85	0.68	2.2	Lib011321	Illumina DNA PCR Free
P20NIST21	87	0.714	2.4	Lib011321	
S46 P0	NA	0.9	3.5	Lib091622	Illumina DNA PCR Free
S46 P20	NA	1.4	4.6	Lib091622	
R41P0	58.8	1.2	3.4	Lib063022	Illumina DNA PCR Free
R41P20	52.8	1.22	4.1		
S49P0	66.6	NA	NA		Illumina DNA PCR Free
S49P20	60	NA	NA		
R18P0	66.6	1.14	3.9	Lib061022	Illumina DNA PCR Free
R18P20	79	0.912	3.1	Lib061022	
R14P0	81	NA	NA	Lib121522	Illumina DNA PCR Free
R14P20	77.1	NA	NA	Lib121522	
H124P0	NA	0.624	2.1	Lib110422	Illumina DNA PCR Free
H124P20	NA	0.62	2.1	Lib110422	
S252	63.9	0.401	1.4	Lib111021	Illumina DNA PCR Free
P20NIST21	61.6	2.86	NA	Lib112221	TruSeq PCR Free Library Kit
S252	58.8	3.22	NA		
Clone 27	NA	NA	NA	LIB022122	NextFlex Rapid XT Kit (Perkin Elmer)
H124P0	23.8	0.56	1.9	LIB24Aug22	Illumina DNA PCR-FREE
H124P20	81.6	0.52	1.8		

Example 2. Evaluating Library Preparation Workflows

[00240] PCR-free library preparation workflows from different commercial suppliers were compared for their consistency and reproducibility including the distribution of DNA insert lengths, the uniformity of coverage across the entire genome, the percentage of mapped reads, and the percentage of Q30 bases for construction of a DNA library for NGS.

[00241] FIG. 11 provides a comparison of insert size distributions generated by different library preparation protocols to compare insert length distributions across several PCR-free library preparation protocols from Illumina, Qiagen, and Perkin Elmer. Each library preparation workflow was repeated multiple times to demonstrate the consistency of produced insert sizes. Libraries were prepared from 300 ng of gDNA isolated from different engineered CHO-K1 clones, sequenced on a NextSeq™ 2000, and data analysis performed using Dynamic Read Analysis for GENomics (DRAGEN) Germline analysis workflow v. 3.8.4.

[00242] Illumina® DNA PCR-Free Library Prep, On-Bead Tagmentation was observed to have the most consistency of produced insert sizes. This method is detailed in the draft Standard Operating Procedure (SOP) BPRD 2489. In summary, the method includes: 1. Tagmenting genomic DNA (Hands-on: 5 minutes, Total: 10 minutes); 2. Post tagmentation clean-up (Hands-on: 10 minutes, Total: 20 minutes); 3. Ligate indexes (Hands-on: 5 minutes, Total: 45 minutes); 4. Clean-up libraries (Hands-on: 15 minutes, Total: 45 minutes); 5. Quantify and pool libraries; 6. Dilute to starting concentration (Reagents: RSB); 7. Sequencing set-up.

[00243] Modifications to step 4 of the standard workflow enabled adjustment of the library insert size. Final cleanup (step 4) and size selection impact the resultant library. Double-sided solid-phase reversible immobilization (SPRI) cleanup and size selection discard low and high molecular weight fragments, reducing yield but tightening the size distribution. Adjusting the volume of sample purification beads (SPB) during double-sided SPRI cleanup allows for fine-tuning the final library insert size. Fragment size distribution within the library is sensitive to the sample preparation beads (SPB) volume, especially during the first cleanup. Increasing the SPB volume in one cleanup step decreases the median fragment size. While an insert size of around 350 bp is optimal for many applications, longer insert sizes was observed to allow more robust integration site detection.

[00244] Library preparation was based on a 300 ng sample of gDNA. Single strand fragments were double stranded by four cycles of PCR for the purpose of fragment size distribution characterization. The median library fragment size and yield were modified by varying the volume of the Illumina® Sample Preparation Beads (SPB) used during the first or second step of cleanup. ND, not determined. Results are shown in Table 3.

Table 3

1st step - SPB [µl]	22	33	45	56	67	45	45	45	45	45	45	45
2nd step - SPB [µl]	15	15	15	15	15	8	10	12	18	18	21	24
Median insert size [bp]	ND	459	334	214	ND	ND	ND	373	334	296	ND	ND

[00245] The other analyzed PCR-free library protocols were Perkin Elmer NEXTFLEX® Rapid DNA-seq Kit 2.0, Illumina® TruSeq™ Library Preparation Kit, and Qiagen's QIAseq® FX protocol Qiagen's QIAseq® FX protocol was observed consistently to produce shorter inserts compared to the other analyzed library types.

[00246] Fig. 12 shows the Illumina DNA PCR-Free coverage uniformity. Uniform coverage facilitates accurately detecting genetic variants, in some specific embodiments, in sequencing data because it enables the identification of variants that are far from the mean depth. To evaluate coverage across the genome at different GC content levels, normalized coverage data from Illumina® DNA PCR-Free, Tagmentation Library Preparation Kit were plotted against the GC content of the CHO-K1 cell genome. Since 20% to 70% of the genome of CHO-K1 consists of GC sequences, this range was specifically analyzed. The results in Fig. 12 shows that the PCR-free kit provides even coverage levels across a wide range of GC content in whole-genome sequencing data.

Example 3. Reproducibility

[00247] The main parameters of libraries prepared manually to those prepared with different gDNA amounts of CHO Clone27 by the automated Perkin Elmer Sciclone® G3 NGSx Workstation were compared. Both read quality and mapping quality are in agreement, although the resulting library insert lengths are shorter in the automated procedure, they remained in an acceptable range.

[00248] The data quality and reproducibility of the library resulting from Illumina® DNA PCR-Free Prep, Tagmentation protocol described herein was tested using eight independently prepared NGS libraries with input of 300 ng of gDNA from CHO-K1 Clone27 as described above. Libraries were sequenced on a NextSeq® 2000, and data analysis was performed using Illumina® DRAGEN Germline analysis workflow v. 3.8.4. The results demonstrated reproducibility of library parameters in seven out of eights repeats.

INTEGRATION OF NANOPORE SEQUENCING FOR COPY NUMBER DETERMINATION

[00249] Refer to Figs. 13, 14, 15, and 16 for an example of using the system 100 of Fig. 1 as follows. In addition to the previously described methods and systems for determining genomic-region copy number, embodiments disclosed herein also include the utilization of Nanopore sequencing technology as an alternative approach for generating the shotgun library and performing sequencing. This method allows for the analysis of longer DNA fragments, providing capabilities for confirming gene integrity, detecting structural variations, and estimating copy numbers.

Sample Preparation and Genomic DNA Isolation

[00250] An embodiment of the method involves isolating and purifying genomic DNA (gDNA) from a biological sample. Specifically, gDNA is extracted from Chinese Hamster Ovary (CHO)-derived clonal cell lines that have been engineered to express recombinant proteins via integrated expression constructs. The extraction of high-quality, high-molecular-weight gDNA may be used in subsequent fragmentation and library preparation steps.

[00251] Genomic DNA extraction may be performed using protocols optimized for purity, integrity, and/or other parameters. In this embodiment, the Qiagen EZ1 DNA Tissue Kit is utilized to extract gDNA from CHO cells with vector integration. An appropriate amount of gDNA, ranging from 500 nanograms (ng) to 1 microgram (µg), may be isolated for downstream processing. The extraction process ensures that the gDNA is suitably free from contaminants and is suitable for subsequent applications. The purity and concentration of the extracted gDNA can be assessed using fluorescence-based detection methods, such as the Qubit fluorometer, which provides quantification of double-stranded DNA.

Fragmentation of Genomic DNA

[00252] Once the gDNA is isolated, can undergo controlled fragmentation to achieve DNA fragments within the range of approximately 300 base pairs (bp) to 2 kilobases (kb), with a mean fragment length of about 1,500 bp. This fragmentation may be accomplished using mechanical shearing methods. For example, the Covaris S220 Focused-ultrasonicator may be employed, following guidelines from the "Quick Guide: DNA Shearing with S220 Focused-ultrasonicator, Covaris."

[00253] Multiple shearing cycles may be performed to ensure uniform fragmentation across the desired size range. After shearing, the fragmented gDNA may be subjected to a double-sided Ampure bead cleanup to remove excessively short or long fragments and/or to concentrate the DNA. All bead-purified material can then be pooled for subsequent steps. The quality and fragment size distribution of the sheared DNA may be evaluated using instruments such as the Agilent Bioanalyzer or TapeStation.

[00254] Referring now to Figure 13, the Bioanalyzer trace provides a detailed representation of the fragment size distribution of the sheared genomic DNA, which has been prepared specifically for Nanopore library construction in one specific embodiment.

[00255] The trace indicates a prominent peak with a mean fragment length of approximately 1,664 base pairs (bp), which is a parameter used for the downstream applications of Nanopore sequencing. This mean fragment length is indicative of the controlled and effective fragmentation achieved through the use of mechanical shearing, such as when employing the Covaris S220 Focused-ultrasonicator. The precise targeting of this fragment size range, approximately 300 bp to 2 kilobases (kb), may be adjusted to balance the need for sufficient read length coverage and the practical limitations of sequencing technology, thereby adjusting the sequencing output.

[00256] The Bioanalyzer trace further reveals the distribution of fragment sizes, which is captured through the full width at half maximum (FWHM) of the peak, providing insights into the uniformity and precision of the shearing process. The presence of any additional peaks or shoulders may suggest variability in the fragmentation process, potentially indicating the presence of unsheared genomic DNA or excessively sheared fragments, both of which could adversely affect sequencing performance in some specific embodiments. However, the trace shown in Figure 13 demonstrates a narrow distribution, supporting the conclusion that the fragmentation process has been suitably controlled.

[00257] Additionally, the trace includes quantitative metrics such as the area under the curve, which represents the total amount of DNA analyzed, and the percentage of total DNA that falls within the desired fragment size range. This information may be used to ensure that the majority of the DNA is appropriately sized for subsequent library preparation and sequencing.

[00258] The effective fragmentation, as confirmed by the mean fragment length of 1,664 bp, may be further validated by the subsequent steps involving double-sided Ampure bead cleanup, which can facilitate the removal of both excessively short and long DNA fragments. This cleanup process may contribute to the purity and concentration of the DNA, as verified by fluorescence-based quantification methods such as the Qubit fluorometer, ensuring that the DNA is suitable for high-fidelity Nanopore sequencing, in this specific exemplary embodiment.

Conversion to a Sequencing-Compatible Library

[00259] The fragmented gDNA, starting with 1 μg of fragmented and purified DNA material, is then converted into a sequencing-compatible DNA library suitable for Nanopore sequencing. This process may involve end-repair and A-tailing of the DNA

fragments to generate blunt-ended, 3'-adenylated DNA molecules. End-repair enzymes remove overhangs and fill in recessed 3' ends, while A-tailing adds a single adenine nucleotide to the 3' ends of the DNA fragments.

[00260] These prepared DNA fragments may then be ligated with sequencing adapters provided in Nanopore library preparation kits, such as the Ligation Sequencing Kit V14 (SQK-LSK114). The adapter ligation step attaches motor proteins and unique molecular identifiers to the DNA fragments. The motor proteins facilitate efficient translocation of the DNA molecules through the nanopores during sequencing, while the unique molecular identifiers enable accurate read identification and demultiplexing.

[00261] Following adapter ligation, the library undergoes further purification using Ampure bead cleanup to remove excess adapters, enzymes, and other reaction components. All bead-purified material is pooled. The purified library is then assessed for quality and concentration using the Qubit fluorometer and the Agilent Bioanalyzer to ensure it meets the requirements for optimal Nanopore sequencing performance.

Library Quantification and Quality Assessment

[00262] The prepared sequencing library may be quantified using sensitive fluorescence-based methods, such as the Qubit fluorometer, which measures DNA concentration with suitable specificity and accuracy. Additionally, the quality and fragment size distribution of the library may be evaluated using platforms like the Agilent Bioanalyzer or TapeStation. These analyses may be used to ensure that the library has appropriate fragment sizes and minimal degradation that are suitable for high-quality Nanopore sequencing.

Nanopore Sequencing Procedure

[00263] Approximately 35 femtomoles (fmol) of the quantified library can be loaded onto a Nanopore flow cell, such as the PromethION Flow Cell (R10.4.1). The sequencing run is initiated using a Nanopore sequencer, such as the PromethION 2 instrument. The number of produced sequencing reads can typically range from 40 million to over 140 million.

[00264] During sequencing, individual DNA molecules pass through nanopores driven by an applied voltage. As the DNA translocates through the pore, changes in

ionic current are measured. Each nucleotide base causes a characteristic disruption in the ionic current, and these signals are decoded in real-time to determine the nucleotide sequence of each molecule.

Data Processing and Alignment

[00265] Following sequencing, the raw Nanopore reads may undergo basecalling and preprocessing to convert the raw signal data (ionic current disruptions) into nucleotide sequences. Basecalling can be performed using software provided by the Nanopore platform, such as Guppy, which utilizes neural network algorithms to decode the signal data accurately.

[00266] The preprocessed reads may be then aligned to a composite reference sequence using alignment tools optimized for long-read data, such as minimap2 or enhanced Sentieon minimap2 alignment software with optimized parameters. The reference sequence comprises both the host genome (e.g., CHO genome) and the vector sequence, allowing for comprehensive mapping of the reads.

[00267] The alignment process accommodates the higher error rates inherent in Nanopore sequencing by allowing for gaps and mismatches during the alignment. Quality control checks are performed to ensure satisfactory alignment rates, reasonable coverage distribution, and lack of contamination. The output of the alignment is a BAM file containing the positions of the reads aligned to the reference genome, along with associated metadata such as mapping quality scores, alignment flags, and supplementary information. This aligned dataset may be used for downstream analyses, including copy number estimation, variant detection, and integration site identification.

Copy Number Estimation and Comparison

[00268] The copy number of the target sequence, such as the GOI within the vector, can be estimated using methodologies similar to those previously described but adapted for long-read data. The mean read coverage for the target sequence may be calculated by determining the average depth of reads aligning to the GOI region.

[00269] A model correlating coverage to copy number may be applied to determine the estimated copy number. This model may be derived from calibration data or from statistical analysis correlating coverage levels to known copy numbers in control samples.

[00270] In Fig. 15, a comparison of copy number predictions based on Nanopore and Illumina sequencing reads is presented. The Nanopore-derived GOI copy number values (indicated by points) are within the 95% confidence interval of the copy number values obtained using Illumina sequencing (depicted by light bars). This comparison demonstrates the reliability and accuracy of Nanopore sequencing for copy number estimation, indicating that Nanopore sequencing is a viable alternative to traditional short-read sequencing methods for this application.

Confirmation of Gene Integrity and Structural Variation Detection

[00271] Fig. 16 provides a detailed visualization of the Gene of Interest (GOI) coverage achieved through Nanopore sequencing, as displayed using the Integrative Genomics Viewer (IGV). This figure demonstrates the capability of Nanopore sequencing to generate long reads that span the entire GOI, ensuring accurate mapping and confirmation of gene integrity.

[00272] The top section of Fig. 16 features a bar, clearly marking the GOI region within the vector sequence. This visual marker spans the entire length of the GOI, providing a reference for the subsequent alignment data. The genomic coordinates are indicated along the axis, offering localization of the GOI within the broader genomic context.

[00273] Below the GOI marker, the coverage track is displayed, showing the depth and distribution of sequencing reads across the genomic region. The coverage is represented by a series of arrows and bars, with each arrow indicating a read that aligns to the reference sequence. The consistent coverage across the GOI region shows the uniformity and completeness of the sequencing data obtained from Nanopore technology.

[00274] The alignment of individual Nanopore reads is shown beneath the coverage track. Each read is depicted as a horizontal bar, with the alignment details indicating how reads span the entire GOI. The annotation "6 reads cover whole GOI" is prominently displayed, emphasizing that six independent reads fully encompass the GOI, thereby confirming the gene's complete and intact integration within the genome.

[00275] This Fig. 16 illustrates Nanopore sequencing in terms of read length and coverage continuity. The ability to generate long reads that cover entire genomic regions is may be used for verifying gene integrity, particularly in applications where

the full-length sequence is used for functional validation, such as therapeutic protein production.

[00276] One of the significant advantages of Nanopore sequencing is the ability to generate long reads that can span entire genes or large genomic regions. This capability facilitates the confirmation of the integrity of the GOI, ensuring that it is full-length and unaltered, which may be used for therapeutic protein production.

[00277] In addition, long reads may improve the detection of structural variations such as insertions, deletions, inversions, or rearrangements. These variations may not be readily detected with short-read sequencing technologies due to limitations in read length and difficulties in mapping repetitive or complex regions.

[00278] The incorporation of Nanopore sequencing in the method provides several possible benefits:

[00279] **Long-Read Capability:** Enables sequencing of longer DNA fragments, improving the resolution of repetitive or complex genomic regions and providing more contiguous coverage over the target sequences. This may enhance the ability to detect and characterize large structural variations and to assemble complete gene sequences in some specific embodiments.

[00280] **Structural Variation Detection:** Facilitates the identification of large-scale genomic alterations that may not be detected with short-read sequencing technologies. Long reads can span breakpoints or junctions, allowing for suitably precise mapping of structural variants.

[00281] **Gene Integrity Confirmation:** Facilitates the verification of the complete sequence of the GOI, ensuring that it is full-length and unaltered. This may be used for therapeutic protein production, as changes in the gene sequence can affect protein function and safety.

[00282] **Flexible and Rapid Analysis:** Nanopore sequencing may offer real-time data generation and the ability to adjust run parameters, possibly providing a flexible platform for various genomic analyses. The rapid turnaround time and scalability make it suitable for high-throughput applications in some embodiments.

[00283] The use of Nanopore sequencing also allows for direct comparison with traditional short-read sequencing methods. As demonstrated in Figs. 14 and 15, the coverage consistency and copy number estimations are highly correlated between Nanopore and Illumina sequencing, with a correlation coefficient of $r = 0.868$, validating the effectiveness of the method.

Data Validation and Quality Control

[00284] Refer now specifically to Figs. 14A-14C, which presents a detailed comparative analysis of sequencing coverage between Nanopore and Illumina sequencing methods. The figure is divided into three distinct panels, each providing insights into the coverage consistency and correlation between these sequencing technologies.

[00285] The upper panel illustrates the base-level coverage across the vector sequence, with distinct lines to differentiate between the sequencing methods. The first line represents coverage obtained from Illumina sequencing, while the second and third lines depict coverage from two separate Nanopore sequencing replicates. A box marks the location of the Gene of Interest (GOI) within the vector sequence, highlighting the region of interest for copy number analysis. The coverage patterns demonstrate the consistency and reliability of Nanopore sequencing in capturing the transgene region, with a correlation coefficient r=0.868 indicated at the bottom, signifying a strong correlation between the sequencing methods.

[00286] The middle panel provides a scatterplot analysis of the correlation between the two Nanopore replicates. This analysis confirms the reproducibility of the Nanopore sequencing method, with a correlation coefficient of r=0.868, reflecting high consistency across independent sequencing runs. The scatterplot demonstrates that variations in coverage are minimal, ensuring reliable data for downstream analyses.

[00287] The lower panel showcases a scatterplot comparing the base-level coverage between Nanopore and Illumina sequencing. Although the figure indicates a correlation coefficient of r=0.8, this suggests a robust but slightly lower correlation compared to the Nanopore replicates. This panel underscores the complementary nature of Nanopore and Illumina sequencing, where Nanopore's long-read capability may be used to provide additional insights into genomic regions that may be challenging for short-read technologies, in some specific embodiments.

[00288] Refer now specifically to Fig. 15, which provides a comparison of Gene of Interest (GOI) copy number predictions derived from Nanopore and Illumina sequencing technologies. Fig. 15 demonstrates the accuracy and reliability of Nanopore sequencing for copy number estimation, in a specific embodiment.

[00289] The central feature of the figure is a bar graph representing the current Aptegra copy number prediction by Illumina sequencing. The light bar indicates the

mean value, while the dashed lines above and below the bar denote the 95% Confidence Interval (CI). This visual representation provides a reference framework against which Nanopore predictions are compared.

[00290] Two points, labeled "Nanopore 1st prediction" and "Nanopore 2nd prediction," are plotted on the graph. Both predictions fall within the 95% CI of the Illumina-based prediction, as indicated by the red points positioned within the dashed lines. This placement underscores the validity of Nanopore sequencing as a reliable alternative for copy number determination, offering comparable accuracy to traditional Illumina methods, in some specific embodiments.

[00291] Throughout the process, quality control checks may be performed to ensure data reliability. This includes verifying satisfactory alignment rates, assessing coverage distribution for uniformity, and checking for potential contamination. These steps may be used for accurate characterization of copy number, detection of variants, and precise identification of integration sites, in some specific embodiments.

[00292] Various alternatives and modifications can be devised by those skilled in the art without departing from the disclosure. Accordingly, the present disclosure is intended to embrace all such alternatives, modifications and variances. Additionally, while several embodiments of the present disclosure have been shown in the drawings and/or discussed herein, it is not intended that the disclosure be limited thereto, as it is intended that the disclosure be as broad in scope as the art will allow and that the specification be read likewise. Therefore, the above description should not be construed as limiting, but merely as exemplifications of particular embodiments. And, those skilled in the art will envision other modifications within the scope and spirit of the claims appended hereto. Other elements, steps, methods and techniques that are insubstantially different from those described above and/or in the appended claims are also intended to be within the scope of the disclosure.

[00293] The embodiments shown in the drawings are presented only to demonstrate certain examples of the disclosure. And, the drawings described are only illustrative and are non-limiting. In the drawings, for illustrative purposes, the size of some of the elements may be exaggerated and not drawn to a particular scale. Additionally, elements shown within the drawings that have the same numbers may be identical elements or may be similar elements, depending on the context.

[00294] Where the term "comprising" is used in the present description and claims, it does not exclude other elements or steps. Where an indefinite or definite article is used when referring to a singular noun, e.g., "a," "an," or "the," this includes a plural

of that noun unless something otherwise is specifically stated. Hence, the term "comprising" should not be interpreted as being restricted to the items listed thereafter; it does not exclude other elements or steps, and so the scope of the expression "a device comprising items A and B" should not be limited to devices consisting only of components A and B. This expression signifies that, with respect to the present disclosure, the only relevant components of the device are A and B.

[00295] Furthermore, the terms "first," "second," "third," and the like, whether used in the description or in the claims, are provided for distinguishing between similar elements and not necessarily for describing a sequential or chronological order. It is to be understood that the terms so used are interchangeable under appropriate circumstances (unless clearly disclosed otherwise) and that the embodiments of the disclosure described herein are capable of operation in other sequences and/or arrangements than are described or illustrated herein.

What is claimed is:

1. A method of determining genomic-region copy number, the method comprising:

isolating a sample of genomic nucleic acids;

purifying the sample of the genomic nucleic acids;

creating a shotgun library from the sample of the genomic nucleic acids;

sequencing the shotgun library to generate reads;

aligning, using a processor, the reads to a reference sequence formed from a reference genome and a vector sequence;

partitioning, using the processor, the reference genome into a plurality of sizes;

determining, using the processor, a mean read coverage for each size of the partitioned reference genome until a predetermined number of consecutive mean-read coverage peaks form with an equal coverage distance between the consecutive mean-read coverage peaks;

determining, using the processor, a respective copy number associated with each of the predetermined number of consecutive mean-read coverage peaks; and

determining, using the processor, a model configured to estimate a copy number of a target sequence.

2. The method according to claim 1, wherein the reference sequence is also formed from at least one auxiliary sequence.

3. The method according to claim 2, wherein the at least one auxiliary sequence includes an Escherichia coli genome sequence.

4. The method according to claim 1, wherein the target sequence is a sequence within the vector sequence.

5. The method according to claim 1, wherein the target sequence is the vector sequence.

6. The method according to claim 1, wherein the target sequence is a predetermined gene within the vector sequence.

7. The method according to claim 1, wherein the model is configured to estimate the copy number without a genomic location parameter.

8. The method according to claim 1, wherein the reference sequence is a DNA genome of a Chinese hamster ovary cell line.

9. The method according to claim 1, wherein the reference sequence is a DNA genome of a mammalian cancer-cell line.

10. The method according to claim 1, wherein the shotgun library is prepared using an NGS shotgun library using a PCR-free workflow.

11. The method according to claim 1, wherein the sequencing of the shotgun library uses sequencing-by-synthesis.

12. The method according to claim 1, wherein the sequencing of the shotgun library uses illumina sequencing.

13. The method according to claim 1, wherein the sequencing of the shotgun library uses DNABSEQ sequencing.

14. The method according to claim 1, wherein the plurality of sizes of the partitioned reference genome are from 100bp to 51,200bp.

15. The method according to claim 1, wherein the plurality of sizes of the partitioned reference genome includes a 50bp, a 100bp, a 200bp, a 400bp, an

800bp, a 1600bp, a 3200bp, a 6400bp, a 12800bp, a 25600bp, and a 51200bp.

16. The method according to claim 15, wherein the 800bp, the 1600bp, the 3200bp, the 6400bp, the 12800bp, the 25600bp, and the 51200bp size are utilized by the processor to estimate the copy number.

17. The method according to claim 15, wherein the 50bp, the 100bp, the 200bp, the 400bp, the 800bp, the 1600bp, the 3200bp, the 6400bp, and the 12800bp are utilized by the processor to estimate a standard deviation of the copy number.

18. The method according to claim 1, wherein all partitions of the partitioned reference genome are non-overlapping.

19. The method according to claim 1, wherein at least two partitions of the partitioned reference genome are overlapping.

20. The method according to claim 1, wherein all partitions of the partitioned reference genome are sequential partitions.

21. The method according to claim 1, wherein the predetermined number of consecutive mean-read coverage peaks is three.

22. The method according to claim 1, wherein the act of determining, using the processor, the respective copy number associated with each of the predetermined number of consecutive mean-read coverage peaks comprises:

comparing the equal coverage distance to a first value of a first peak of the predetermined number of consecutive mean-read coverage peaks to determine a first respective copy number corresponding to the first peak of the predetermined number of consecutive mean-read coverage peaks.

23. The method according to claim 22, wherein when the first value is about equal to the equal coverage distance, the first respective copy number is one.

24. The method according to claim 22, wherein a second peak of the predetermined number of consecutive mean-read coverage peaks corresponds to a second respective copy number of two.

25. The method according to claim 22, wherein when the first value is about twice the equal coverage distance the first respective copy number is two.

26. The method according to claim 25, wherein a second peak of the predetermined number of consecutive mean-read coverage peaks corresponds to a second respective copy number of three.

27. The method according to claim 1, wherein the model is a regression model that correlates coverage numbers with copy numbers.

28. The method according to claim 1, wherein the partitioned reference genome corresponds to at least one gene.

29. The method according to claim 1, further comprising normalizing the read coverages by a reference coverage value obtained from a control sample.

30. The method according to claim 1, further comprising estimating a standard deviation of the estimated copy numbers based on the read coverages and a size of the target sequence.

31. The method according to claim 30, further comprising determining a confidence interval for the estimated copy numbers based on the standard deviation and the estimated copy numbers.

32. The method according to claim 1, wherein the alignment of the reads to the reference sequence is performed using an accelerated version of a BWA aligner.

33. The method according to claim 1, wherein the determination of the mean read coverage for each size of the partitioned reference genome comprises calculating a median coverage value for each size.

34. The method according to claim 1, wherein the model is trained using mean base-level coverage of each partition and known copy-numbers within the reference sequence.

35. The method according to claim 1, further comprising performing a quality control analysis on the read coverages to identify sequencing errors and filtering out false-positive copy number calls.

36. The method according to claim 1, further comprising visualizing the estimated copy numbers and their confidence intervals on a circos plot.

37. The method according to claim 1, wherein the target sequence is one or more genes of interest (GOIs) within the vector sequence.

38. The method according to claim 37, further comprising determining the copy number of the one or more GOIs based on their observed read coverages and the model.

39. The method according to claim 1, wherein the vector sequence is an integrated full-size or partial-size vector.

40. The method according to claim 1, further comprising identifying and reporting any vector-vector concatenation events based on split reads that map to different parts of the vector sequence.

41. The method according to claim 1, further comprising detecting and reporting sequence variants in an integrated vector by comparing the reads to the vector sequence.

42. The method according to claim 1, further comprising determining a location of at least one vector integration point in the reference genome by analyzing chimeric reads of the aligned reads and paired reads of the aligned reads mapped to vector regions and non-vector regions.

43. The method according to claim 1, wherein the method is used for molecular characterization of Chinese hamster ovary (CHO) cells used for producing recombinant proteins.

44. The method according to claim 1, wherein the method is used for molecular characterization of the reference genome.

45. The method according to claim 1, wherein the method is used for molecular characterization of a mammalian cancer-cell line.

46. The method according to claim 1, further comprising calculating a coverage variability for each size of the partitioned reference genome to estimate a reliability parameter of the estimated copy numbers.

47. The method according to claim 1, wherein the determination of the mean read coverage for each size of the partitioned reference genome includes calculating a mean base-level coverage of each partition considering the size and copy number of each fragment.

48. The method according to claim 1, further comprising determining a coverage anomaly.

49. The method according to claim 1, further comprising identifying regions of homologous sequences within the vector sequence.

50. The method according to claim 1, further comprising comparing the estimated copy number of the target sequence with a predetermined threshold to estimate copy number stability and integrity.

51. The method according to claim 1, further comprising validating an accuracy using a molecular characterization from one of a Southern blot analysis, a ddPCR, and a qPCR.

52. The method according to claim 1, wherein the reads are generated using a long-read sequencing technology to thereby improve an accuracy of variant detection and integration site annotation.

53. The method according to claim 1, further comprising detecting and reporting rearrangements or structural variations within the reference sequence by analyzing an alignment pattern of the reads.

54. The method according to claim 1, further comprising generating a molecular characterization report summarizing the estimated copy number, a variant profile, and integration sites of the target sequence.

55. The method according to claim 1, further comprising determining a ploidy distribution of the reference genome based on observed coverages of fragments with varying sizes to refine the estimated copy number of the target sequence.

56. The method according to claim 1, the method further comprising validating an accuracy of the estimated copy number using at least one of Fish analysis, ddPCR analysis, and a qPCR.

57. The method according to claim 1, further comprising fetching coverages of different fragments with varying sizes in the reference genome,

including sizes of 50, 100, 200, 400, 800, 1600, 3200, 6400, and 12800bp, by extracting coverage information from the aligned reads.

58. The method according to claim 1, further comprising normalizing a coverage based upon the largest peak of the consecutive mean-read coverage peaks, dividing the coverage at each point by the coverage of the largest peak to provide normalized coverage values.

59. The method according to claim 1, further comprising normalizing an occurrence by a total occurrence of all coverages by dividing the occurrence at each point by a sum of occurrences across all points to provide normalized occurrence values.

60. The method according to claim 1, further comprising:

extracting the mean read coverage for each of the plurality of sizes for known copies; and

computing a standard deviation for each size to assess a variation in fragment coverage.

61. The method according to claim 1, further comprising determining concatenation events of the vector sequence.

62. The method according to claim 61, further comprising determining a junction read having a first segment of the vector sequence and a second segment of the vector sequence, wherein the first and second segments are joined together within a junction read of the reads.

63. The method according to claim 1, further comprising identifying chimeric reads from the aligned reads, wherein the chimeric reads partially map to the vector sequence and partially map to non-vector regions.

64. The method according to claim 63, further comprising determining vector-vector junctions based upon the chimeric reads to thereby determine a location of a vector insertion within the reference genome.

65. The method according to claim 64, further comprising performing de novo assembly of the chimeric reads and respective mates to generate longer contigs representing continuous stretches of sequence.

66. The method according to claim 65, further comprising aligning the assembled contigs to the reference genome, which includes both the host genome and vector sequences, to identify reference genome-vector boundaries and create an alignment file.

67. The method according to claim 1, wherein the shotgun library has a mean fragment length of about 1500 base pairs, and the sequencing of the shotgun library is performed using Nanopore sequencing technology.

68. The method according to claim 67, wherein the sample of genomic nucleic acids is isolated from a Chinese Hamster Ovary (CHO)-derived clonal cell line.

69. The method according to claim 68, wherein the sample of genomic nucleic acids comprises genomic DNA (gDNA) isolated from the CHO-derived clonal cell line.

70. The method according to claim 69, further comprising fragmenting the genomic DNA into fragments having sizes within the range of about 0.3 to 2 kilobases.

71. The method according to claim 70, further comprising converting the fragmented genomic DNA into a sequencing-compatible DNA library.

72. The method according to claim 71, further comprising quantifying the sequencing-compatible DNA library using a Qubit fluorometer.

73. A system for determining copy number, the system comprising:

a processor configured to perform any method according to any one of claims 1-72.

74. An apparatus for determining copy number implemented by an operative set of processor executable instructions, the operative set of processor executable instructions configured to perform one or more acts of a method according to any one of claims 1-72.

75. A computer-readable medium storing instruction that, when executed by a processor, causes the processor to:

aligning, using the processor, the reads to a reference sequence formed from a reference genome and a vector sequence;

partitioning, using the processor, the reference genome into a plurality of sizes;

determining, using the processor, a mean read coverage for each size of the partitioned reference genome until a predetermined number of consecutive mean-read coverage peaks form with an equal coverage distance between the consecutive mean-read coverage peaks;

determining, using the processor, a respective copy number associated with each of the predetermined number of consecutive mean-read coverage peaks; and

determining, using the processor, a model configured to estimate a copy number of a target sequence.

76. The medium according to claim 75, wherein the computer-readable medium is further configured to cause the processor to perform the method of any one of claims 2-72.

77. The medium according to claim 75 or 76, wherein the computer-readable medium is a non-transitory computer readable medium.

78. The method according to claim 1, further comprising the acts of:

(i) extracting genomic DNA (gDNA) from a mammalian cell line to obtain the sample of genomic nucleic acids;

(ii) tagmenting the sample of gDNA having a mass within the range of 300 ng to 2000 ng to form tagged gDNA fragments;

(iii) cleaning up the sample a first time using sample preparation beads;

(iv) ligating indexes to the tagged gDNA fragments from the tagmenting in step (ii); and

(v) cleaning up the sample a second time using sample preparation beads resulting in the library for NGS.

79. The method of claim 78, wherein the mammalian cell line is a Chinese hamster ovary (CHO) cell line or a *Homo sapiens* cell line.

80. The method of any one of claims 78 and 79, wherein the DNA is high molecular weight.

81. The method of any one of claims 78 to 80, wherein the number of mutations, insertions, or deletions is decreased compared to the number of mutations, insertions, or deletions occurring in library preparation methods using a polymerase chain reaction (PCR) amplification step.

82. The method of any one of claims 78 to 81, further comprising adjusting the volume of the sample preparation beads during the first clean-up step or the second clean-up step to modify the median library fragment size.

83. The method of claim 82, wherein the median library fragment size is greater than 350 base pairs (bp).

84. The method of any one of claims 78 to 83, wherein the first cleaning step comprises sample preparation beads in a volume selected from the range of 30 µl to 50 µl.

85. The method of any one of claims 78 to 84, wherein the second cleaning step comprises sample preparation beads in a volume selected from the range of 10 µl to 20 µl.

86. The method of any one of claims 78 and 85, wherein the volume of sample preparation beads in the first cleaning step is 33 µl and wherein the volume of sample preparation beads in the second cleaning step is 15 µl.

87. The method of any one of claims 84 and 85, wherein the volume of sample preparation beads in the first cleaning step is 45 µl and wherein the volume of sample preparation beads in the second cleaning step is 12 µl.

88. A method of reducing sequencing errors during high-throughput next generation sequencing (NGS), the method comprising:

(i) performing high-throughput next generation sequencing (NGS) on the library prepared by the steps of any one of claims 78 to 87.