METHOD AND SYSTEM FOR DETECTING COPY NUMBER VARIATION

TECHNICAL FIELD

Embodiments of the present disclosure generally relate to a field of bioinformatics, more particularly to a method of detecting a copy number variation (CNV) and a system thereof.

BACKGROUND

CNV is one of genomic structure variations, with a narrow definition as a presence of a copy number change of a DNA fragment in a chromosome. Types and reasons of the genome structure variation may comprise: 1. deletion (terminal deletion, interstitial deletion); 2. translocation (reciprocal translocation, Robertsonian translocation); 3. inversion; 4. ring chromosome; 5. dicentric chromosome; 6. insert, and etc. A broader definition of CNV also includes: for example, structure variations such as chromosome aneuploidy and partial aneuploidy.

Presently available methods for detecting the copy number variation, mainly include high resolution chromosome karyotype analysis, FISH (fluorescence in situ hybridization), Array CGH (array comparative genomic hybridization), MLPA (Multiplex Ligation-Dependent Probe Amplification), PCR (Polymerase Chain Reaction), and etc, among which FISH detection is regarded as a gold standard for genetic diagnosis, which may effectively be used to detect most of known chromosome deletions or duplications. However, these methods generally suffer from low efficiencies, particularly when used for whole-genome scans, which can consume larger amounts of resources or may not be able to detect unknown CNV.

Thus, there is an urgent need for a new method of detecting a copy number variation to characterize a known site and explore an unknown site in the genome.

SUMMARY

One technical problem to be solved by the present disclosure is to provide a method of detecting a copy number variation and a system thereof, which may accurately detect a copy number variation even including micro-deletion and micro-duplication.

Embodiments of a first broad aspect of the present disclosure provide a method of detecting a copy number variation. According to embodiments of the present disclosure, the method may comprise following steps:

obtaining reads from at least one part of a nucleic acid molecule of a sample,

determining uniquely-mapped reads aligned to a (genomic) reference sequence based on the obtained reads,

dividing the genomic reference sequence into a plurality of windows, and calculating the number of uniquely-mapped reads falling into each of the plurality of windows,

subjecting the number of uniquely-mapped reads falling into each of the plurality of windows to a GC correction, and to a correction based on the expected number of uniquely-mapped reads corrected by a control set to obtain the corrected number of uniquely-mapped reads,

calculating a significance of value population consisting of the corrected number of uniquely-mapped reads between two sides of a demarcation point, with the demarcation point being selected as a starting point or a terminal point of each of the plurality of windows, to select the demarcation point having a smaller value of the significance as a candidate of a CNV breakpoint;

calculating a significance of two value populations consisting of the corrected number of uniquely-mapped reads respectively contained in two sequences, with one sequence ranging from each of the CNV breakpoint to an adjacent upstream CNV breakpoint, and the other sequence ranging from each of the CNV breakpoint to an adjacent downstream CNV breakpoint, and

removing the candidate of the CNV breakpoint having the least significance for each time and recalculating the significance of two candidates of the CNV breakpoint adjacent to the removed candidate of the CNV breakpoint, cycling and iterating until the significances of all candidates of the CNV breakpoints are less than a final cut-off, to determine the CNV breakpoint.

Optionally, the method may further comprise a step of: subjecting the at least one part of the nucleic acid molecule of the sample to sequencing, to obtain the reads.

Optionally, each of the plurality of windows may contain the same number of reference unique reads, or each of the plurality of windows may have a same length.

Optionally, the final cut-off may be obtained based on the control set consisting of normal samples.

Optionally, the step of subjecting the number of uniquely-mapped reads falling into each of the plurality of windows to the GC correction and to a correction based on the expected number of uniquely-mapped reads corrected by a control set to obtain the corrected number of uniquely-mapped reads may further comprise: grouping the plurality of windows based on a GC content, and obtaining a correction coefficient based on a mean value of the number of uniquely-mapped reads within one group and a mean value

of the number of uniquely-mapped reads for all of the plurality of windows, and then subjecting the number of uniquely-mapped reads falling into each of the plurality of windows to the correction, to obtain the GC-corrected number of uniquely-mapped reads.

Optionally, the expected number of uniquely-mapped reads corrected by the control set may be obtained by following steps:

calculating a ratio of the GC-corrected number of uniquely-mapped reads falling into each of the plurality of windows in the control set to the total number of uniquely-mapped reads;

obtaining a mean value of the ratios for all windows corresponding to the control set; and

calculating the expect number of uniquely-mapped reads for each of the plurality of windows in the sample based on obtained mean value of the ratios and the total number of uniquely-mapped reads in the sample.

Optionally, after the CNV breakpoint is determined, the method may further comprise:

subjecting the sequence between two CNV breakpoints to a confidence selection, wherein the confidence selection comprises steps of:

determining a confidence interval of the normal corrected number of uniquely-mapped reads using the control set, based on a distribution of the corrected number of uniquely-mapped reads; and

determining an abnormality presenting in the sequence between the two CNV breakpoints, if the mean value of the corrected number of uniquely-mapped reads within the sequence is outside the confidence interval.

Optionally, the corrected number of uniquely-mapped reads may fit a normal distribution, and the confidence interval may be 95%.

Optionally, a cyclization with a single chromosome or a whole genome may be performed when selecting the candidate of the CNV breakpoint.

Optionally, the method may further comprise:

obtaining the sample from a human, with the sample comprising amniotic fluid obtained by amniocentesis, villus obtained by chorionic villi sampling, umbilical cord blood obtained by percutaneous umbilical blood sampling, spontaneous miscarrying fetus tissue or human peripheral blood; and/or obtaining a genomic DNA of the sample by a DNA extraction method such as a salting-out method, a column chromatography method, a beads method, or a SDS method; and/or subjecting the genome DNA of the sample to random fragmenting by enzyme digestion, pulverization, ultrasound, or HydroShear method, to obtain DNA fragments; and/or subjecting the DNA fragments to single-end

sequencing or pair-end sequencing, to obtain the reads of the DNA fragments.

Optionally, the method may further comprise adding different indexes to each of the DNA fragments of the samples, to distinguish different samples.

Embodiments of a second broad aspect of the present disclosure provide a system for determining a copy number variation. According to embodiments of the present disclosure, the system may comprise:

a reads obtaining unit, for obtaining reads from at least one part of a nucleic acid molecule of a sample;

a uniquely-mapped reads determining unit, for determining uniquely-mapped reads aligned to a genome reference sequence based on obtained reads;

the number of uniquely-mapped reads calculating unit, for dividing the genome reference sequence into a plurality of windows, and calculating the number of uniquely-mapped reads falling into each of the plurality of windows

the number of uniquely-mapped reads correcting unit, for subjecting the number of uniquely-mapped reads falling into each of the plurality of windows to a GC correction, and to a correction based on the expected number of uniquely-mapped reads corrected by a control set to obtain the corrected number of uniquely-mapped reads;

a candidate of breakpoint selecting unit, for calculating a significance of value population consisting of the corrected number of uniquely-mapped reads between two sides of a demarcation point, with the demarcation point being selected as a starting point or a terminal point of each of the plurality of windows, to select the demarcation point having a smaller value of the significance as a candidate of a CNV breakpoint;

a breakpoint determining unit, for calculating a significance of two value populations consisting of the corrected number of uniquely-mapped reads respectively contained in two sequences, with one sequence ranging from each of the CNV breakpoint to an adjacent upstream CNV breakpoint, and the other sequence ranging from each of the CNV breakpoint to an adjacent downstream CNV breakpoint, and removing the candidate of the CNV breakpoint having the least significance for each time and recalculating the significance of two candidates of the CNV breakpoint adjacent to the removed candidate of the CNV breakpoint, cycling and iterating until the significances of all candidates of the CNV breakpoint are less than a final cut-off, to determine the CNV breakpoint.

Optionally, each of the plurality of windows may contain the same number of reference unique reads, or each of the plurality of windows may have a same length.

Optionally, the final cut-off may be obtained based on the control set consisting of normal samples.

Optionally, the number of uniquely-mapped reads correcting unit comprises:

a GC correction module, for grouping the plurality of windows based on a GC content, and obtaining a correction coefficient based on a mean value of the number of uniquely-mapped reads within one group and a mean value of the number of uniquely-mapped reads for all of the plurality of windows, and then subjecting the number of uniquely-mapped reads falling into each of the plurality of windows to the correction, to obtain the GC-corrected number of uniquely-mapped reads;

a window correction module, for calculating a ratio of the GC-corrected number of uniquely-mapped reads falling into each of the plurality of windows in the control set to the total number of uniquely-mapped reads; obtaining a mean value of the ratios for all windows corresponding to the control set; and calculating the expect number of uniquely-mapped reads for each of the plurality of windows in the sample based on obtained mean value of the ratios and the total number of uniquely-mapped reads in the sample.

Optionally, after the CNV breakpoint is determined by the breakpoint determining unit, the system may further comprise:

a breakpoint filtering unit, for determining a confidence interval of the normal corrected number of uniquely-mapped reads using the control set, based on a distribution of the corrected number of uniquely-mapped reads; and determining an abnormality presenting in the sequence between the two CNV breakpoints, if the mean value of the corrected number of uniquely-mapped reads within the sequence is outside the confidence interval.

Optionally, the corrected number of uniquely-mapped reads may fit a normal distribution, and the confidence interval may be 95%.

Optionally, in the candidate of breakpoint selecting unit, a cyclization with a single chromosome or a whole genome is performed when selecting the candidate of the CNV breakpoint.

Optionally, the system may further comprise:

means for obtaining the sample from a human, with the sample comprising amniotic fluid obtained by amniocentesis, villus obtained by chorionic villi sampling, umbilical cord blood obtained by percutaneous umbilical blood sampling, spontaneous miscarrying fetus tissue or human peripheral blood;

and/or

means for obtaining a genomic DNA of the sample by a DNA extraction method such as a salting-out method, a column chromatography method, a beads method, or a SDS

method;

and/or

means for subjecting the genome DNA of the sample to random fragmenting by enzyme digestion, pulverization, ultrasound, or HydroShear method, to obtain DNA fragments;

and/or

means for subjecting the DNA fragments to single-end sequencing or pair-end sequencing, to obtain the reads of the DNA fragments.

Optionally, different samples are distinguished by adding different indexes to each of the DNA fragments of the samples.

One of the advantages of the method of detecting a copy number variation and a system thereof according to embodiments of the present disclosure lies in having clinical feasibility, which may accurately detect the copy number variation comprising micro-deletion and micro-duplication.

BRIEF DESCRIPTION OF THE DRAWINGS

These and other aspects and advantages of embodiments of the present disclosure will become apparent and more readily appreciated from the following descriptions made with reference the accompanying drawings, in which:

Fig.1 is a flow chart showing a method of detecting a copy number variation according to an embodiment of the present disclosure;

Fig.2 is a flow chart showing a method of detecting a copy number variation according to another embodiment of the present disclosure;

Fig.3 is a flow chart showing a method of detecting a copy number variation according to a further embodiment of the present disclosure;

Fig.4 is a brief flow chart showing a method of chromosomal CNV analysis according to an embodiment of the present disclosure;

Fig.5 is a schematic diagram showing a system of detecting a copy number variation according to an embodiment of the present disclosure;

Fig.6 is a schematic diagram showing a system of detecting a copy number variation according to anther embodiment of the present disclosure;

Fig.7A-7H show detection results of eight samples respectively in an example of the present disclosure.

DETAILED DESCRIPTION

Terms used herein are explained as below:

Copy number variation (CNV): refers to a change of copy number having a length more than 1 kb of nucleic acid molecule resulting from comparing a nucleic acid sequence of a sample to be tested with a nucleic acid sequence of a normal sample. Cases and reasons of the copy number variation may comprise: deletion, such as micro-deletion; insertion, such as micro-insertion, micro-duplication, duplication, inversion, transposition and complicated multi-site variation.

Aneuploidy: may refer to an addition or a reduction of the chromosome number presenting in a genetic material as compared comparing with a normal sample, and may further include an addition or reduction of an entire or partial chromosome. The copy number variation involved in the present disclosure may also include a case of aneuploidy.

Sequencing: a process of obtaining information of a nucleic acid sequence of a sample. The sequencing may be performed by various methods, including but not limited to, dideoxy chain termination; preferably a high-throughput sequencing method, including but not limited to, a Next-Generation sequencing technique or a single molecule sequencing technique.

The Next-Generation sequencing platform (Metzker ML. Sequencing technologies-the next generation. Nat Rev Genet. 2010 Jan; 11(1):31-46) includes, but are not limited to Illumina-Solexa (GATM, HiSeq2000TM, etc), ABI-Solid and Roche-454 (pryosequencing) sequencing platform; the single molecule sequencing platform (technique) includes, but are not limited to True Single Molecule DNA sequencing of Helicos Company, single molecule real-time (SMRTTM) of Pacific Biosciences Company, and nanopore sequencing technique of Oxford Nanopore Technologies Company (Rusk, Nicole (2009-04-01). Cheap Third-Generation Sequencing. Nature Methods 6 (4): 2446 (4)), etc.

A type of sequencing may be single-end sequencing and pair-end sequencing, a sequencing length may be 50 bp, 90 bp or 100 bp. In an embodiment of the present disclosure, the sequencing platform is Illumina/Solexa, the type of sequencing is pair-end sequencing, to obtain DNA sequence molecule having a length of 100 bp with a relationship of a bi-directional position.

In an embodiment of the present disclosure, a sequencing depth may be determined based on a length of variation fragment in a chromosome of a sample to be test. The higher sequencing depth, the higher sensitivity of the detection, i.e., the smaller length of deletion fragment and duplication fragment which can be detected. The sequencing depth may be 0.1-30×, i.e., the total amount is 0.1-30 folds in relative to a length of human genome, for example, in an embodiment of the present disclosure, the sequencing depth

is 0.1× (2.5×10^8 bp).

Reads: a nucleic acid sequence having a certain length (generally longer than 20 bp), for example a sequencing result of the sequence generated by sequencer, which may be aligned to a specific region or location of a reference sequence by a sequence alignment method.

Sequence alignment (aligning): refers to a process of comparing one or more nucleic acid sequences to a reference sequence. Specifically, a nucleic acid sequence having a relative shorter length (such as a read) is aligned to a reference genome sequence, to determine a location of the nucleic acid sequence having the relative shorter length in the reference genome. When using a computer to perform sequence alignment, the sequence alignment may be performed by any one of sequence alignment procedures, such as ELAND (efficient local alignment of nucleotide data), SOAP (Short Oligonucleotide Analysis Package) and BWA (Burrows-Wheeler Aligner), etc. A standard for recognizing a successful alignment is classified into a non-fault-tolerant alignment (100% of matches) and partial-fault-tolerant alignment (less than 100% of matches).

uniquely-mapped reads: refer to reads which may be aligned to a unique position of a reference sequence (for example a reference genome sequence).

Reference unique reads: refer to sequences having a fixed length and a uniquely-mapped position aligned to a reference sequence (generally a reference genome). A process of obtaining the reference unique reads, for example includes, dividing a reference genome into a plurality of sequences having a fixed length, aligning the plurality of sequences back to the reference genome, selecting a sequence uniquely-aligned to the reference genome as a reference uniquely-mapped sequence. The fixed length is determined based on a sequence length of a sequencing result obtained by a sequencer, which may specifically refer to a mean length. Different sequencer may obtain sequencing results of different sequence length. Specifically to each time of sequencing, the sequence length of the sequencing result may also be different, in which a certain subjective and experience factors may exist in the selection thereof.

Index: may refer to a nucleic acid sequence having a specific length and playing a function as a marker. When DNA molecules to be tested derive from a plurality of samples to be tested, each of the plurality of samples may be added with different indexes, for distinguishing the plurality of samples during sequencing (Micah Hamady, Jeffrey J Walker, J Kirk Harris et al. Error-correcting barcoded primers forpyrosequencing hundreds of samples in multiplex. Nature Methods, 2008, March, Vol.5 No.3), to achieve subjecting the

plurality of samples to simultaneous sequencing. The index is for distinguishing different sequences, without affecting other functions of the DNA molecule added with the index.

GC correction: since a certain GC bias exists among batches or within one batch, which may cause copy number bias presenting in a region having a high GC content or a low GC content of a genome. The CG correction is performed with sequencing data based on a control set, to obtain corrected relative number of reads in each window, by which such bias may be eliminated, and an accuracy of detecting the copy number variation may be improved.

Mean value: the mean value used herein generally refers to an arithmetic mean value or a median.

The number of uniquely-mapped reads: the number of uniquely-mapped reads may be statistic number obtained by calculating based on the initial number, or may be corrected value obtained by subjecting the number of uniquely-mapped reads to a correction with a correction coefficient, for example may be a ratio, in some cases may be interchangeable with "copy rate".

Sample to be tested: may also be called as a test sample in some cases, which refers to a sample including a nucleic acid molecule which is suspected having variation. A type of the nucleic acid is not subjected to special restriction, which may be desoxyribonucleic acid (DNA), or may be ribonucleic acid (RNA), preferably is DNA. RNA may be converted to DNA having a corresponding sequence by conventional means, for subsequent detections and analysis.

Control sample: is relative to the sample to be tested, which is regarded as a normal sample. Generally, normal means having a normal phenotype.

Control sample set (Control set): refers to a set consisting of the control samples, in an embodiment of the present disclosure, the number of the control samples in the control set is required to be more than 30.

Reference will be made in detail to embodiments of the present disclosure. The embodiments described herein with reference to drawings are explanatory, illustrative, and used to generally understand the present disclosure.

With continuous development of high-throughput sequencing technology and gradually decrease of sequencing cost, the sequencing technology has been increasingly and extensively applied in detection of chromosome variation.

Sequencing technology to detect chromosomal aberrations in terms of gaining more widely used.

To improve the technology of detecting a copy number variation in clinic, the present

disclosure designs a technical solution for screening the copy number variation in a whole genome level based on high-throughput sequencing technology, which has advantages of being high throughput, high specific and high accuracy of alignment. A detection result may be obtained by obtaining a sample from a subject; extracting DNA; high throughput sequencing, subjecting obtained data to analysis.

Fig.1 is a flow chart showing a method of detecting a copy number variation according to an embodiment of the present disclosure.

As shown in Fig.1, in step 102, reads are obtained from at least one part of a nucleic acid molecule of a sample. At least one part of nucleic acid molecule or an entire nucleic acid molecule in a test sample may be subjected to sequencing to obtain reads. Reads of the part of nucleic acid molecule of the test sample may be obtained, or reads of the entire nucleic acid molecule may be obtained. For example, a genomic DNA molecule from a test sample is randomly fragmented to obtain DNA fragments, which are then subjected to sequencing to obtain reads having a certain length. The length of obtained reads may be within a certain range, and the reads having a fixed length may be obtained by truncating. The DNA fragments may have a length of 50 bp~1500 bp, such as, 50 bp~150 bp, 150 bp~350 bp, 350 bp~500 bp, 500 bp~700 bp, 700 bp~1000 bp or 1000 bp~1500 bp. For example, the DNA fragments may have a length of 50 bp, 90 bp, 100 bp, 150 bp, 300 bp, 350 bp, 500 bp, 700 bp, 1000 bp, 1500 bp. In an example, 300 bp~700 bp is preferred, 350 bp~500 bp is more preferred. The length of the reads may having a large difference due to different sequencer, for example, general sequence length of equipments such as illumina-solexa and life technologies-solid is within a range of 300 bp, while sequence length obtained by Roche-454, conventional Sanger sequencing, ultramodern single molecule sequencing system may be approximately or exceed 1000 bp. To meet an requirement for unique alignment, when selecting uniquely-mapped reads, sequences having a length of 20 bp or more are generally selected, preferably, selected sequences have a length of 26 bp or more.

Step 104, uniquely-mapped reads aligned to a genome reference sequence are determined based on the obtained reads. For example, entire or partial sequences of the reads are aligned to the genome reference sequence, to obtain site information of the reads in the genome, and further obtain site information of the reads on a specific chromosome. For a sample deriving from a human subject, the human genome reference sequence may be a human genome reference sequence in NCBI database. In an example of the present disclosure, the human genome reference sequence is a human genome reference sequence of Build 36 in NCBI database (hg18; NCBI Build 36), alignment

software used is SOAPaligner/soap2. Those DNA fragment reads uniquely-mapped to the genome reference sequence are selected, namely, those reads uniquely-mapped to the genome reference sequence only one time, i.e., uniquely-mapped reads aligned to the (genome) reference sequence.

Step 106, the genome reference sequence is divided into a plurality of windows, and the number of uniquely-mapped reads falling into each of the plurality of windows is calculated. A method for detecting the plurality of windows may comprise: fragmenting a reference genome into fragments having a sequencing length being same as that of the sample to be tested, which are subjected to alignment by same alignment software with same parameter, to screen out uniquely-mapped positions on a chromosome; determining one of the plurality of windows by every interval having a certain length of the uniquely-mapped positions. A cross-sliding between the pluralities of the windows may or may not exist depending on selection. The number of uniquely-mapped sites which may be included in the plurality of windows relates to sequencing data volume of the sample to be tested. Generally, the expected reads number of the sample to be tested falling into each of the plurality of windows is 300 or more, so as to guarantee the reads number falling into the plurality of windows fitting Poisson distribution. For example, assuming that the number of uniquely-mapped sites of a genome is N, effective reads number of the sample to be tested is n, and the expected number of reads falling into each of the plurality of windows in E, then each of the plurality of windows of the reference genome may include $\frac{N}{n} \times E$ of the uniquely-mapped sites.

Step 108, the number of uniquely-mapped reads falling into each of the plurality of windows is subjected to a GC correction, and to a correction based on the expected number of uniquely-mapped reads corrected by a control set to obtain the corrected number of uniquely-mapped reads. In an example, the step of subjecting the number of uniquely-mapped reads falling into each of the plurality of windows to the GC correction and to a correction based on the expected number of uniquely-mapped reads corrected by a control set to obtain the corrected number of uniquely-mapped reads further comprises:

grouping the plurality of windows based on a GC content, and obtaining a correction coefficient based on a mean value of the number of uniquely-mapped reads within one group and a mean value of the number of uniquely-mapped reads for all of the plurality of windows, and then subjecting the number of uniquely-mapped reads falling into each of the plurality of windows to the correction, to obtain the GC-corrected number of uniquely-mapped reads;

the expected number of uniquely-mapped reads corrected by the control set is obtained by following steps:

calculating a ratio of the GC-corrected number of uniquely-mapped reads falling into each of the plurality of windows in the control set to the total number of uniquely-mapped reads;

obtaining a mean value of the ratios for all windows corresponding to the control set; and

calculating the expect number of uniquely-mapped reads for each of the plurality of windows in the sample based on obtained mean value of the ratios and the total number of uniquely-mapped reads in the sample.

Step 110, a significance of value population consisting of the corrected number of uniquely-mapped reads between two sides of a demarcation point is calculated, with the demarcation point being selected as a starting point or a terminal point of each of the plurality of windows, to select the demarcation point having a smaller value of the significance as a candidate of a CNV breakpoint. For example, the predetermined number of windows is selected as a candidate of a CNV breakpoint within a whole genome range based on a p value presenting a significance level of copy number variation between two sides of each of the plurality of windows, to obtain a significance value of the candidate of each of the CNV breakpoints, namely, a p value.

Step 112, a significance of two value populations consisting of the corrected number of uniquely-mapped reads respectively contained in two sequences is calculated, with one sequence ranging from each of the CNV breakpoint to an adjacent upstream CNV breakpoint, and the other sequence ranging from each of the CNV breakpoint to an adjacent downstream CNV breakpoint, and

the candidate of the CNV breakpoint having the least significance is removed for each time and the significance of two candidates of the CNV breakpoint adjacent to the removed candidate of the CNV breakpoint is recalculated, which are subjected cycling and iterating until the significances of all candidates of the CNV breakpoint are less than a final cut-off, to determine the CNV breakpoint. The final cut-off generally is preset. For example, the final cut-off is obtained by subjecting the control set consisting of normal samples to analysis and process.

In the above examples, selection of the CNV breakpoints is performed by: aligning obtained reads to the genome reference sequence, calculating the number of uniquely-mapped reads falling into each of the plurality of windows, subjecting the number of uniquely-mapped reads falling into each of the plurality of windows to a GC correction

and to a correction based on a control set, and subjecting a significance difference to cycling and iterating, so as to achieve CNV detection, by which may accurately determine relative small copy number variation comprising micro-deletion/micro-duplication.

For the sample deriving from a human, the subject sample may be a genomic DNA obtained from amniotic fluid obtained by amniocentesis, villus obtained by chorionic villi sampling, umbilical cord blood obtained by percutaneous umbilical blood sampling, spontaneous miscarrying fetus tissue or human peripheral blood. The genomic DNA may be extracted by conventional extraction methods such as a salting-out method, a column chromatography method, a beads method, or a SDS method. In an example, the column chromatography method is preferred, such column chromatography method may comprise: obtaining exposed DNA molecules by subjecting blood, tissue and cells to cell lysis buffer and proteinase K, and combining the DNA molecules to a silicone membrane under a condition of high salinity, and then eluting the DNA molecules from the silicone membrane under a condition of low salinity and high pH value. A specific principal and method may refer to specification of Tiangen TIANamp Micro DNA Kit (DP36).

If DNA fragments to be detected derive from a plurality of subject samples, the DNA fragments of each subject sample may be added with different indexes having a length of 4 bp to 12 bp, for distinguishing different sample during sequencing (Micah Hamady, Jeffrey J Walker, J Kirk Harris et al. Error-correcting barcoded primers forpyrosequencing hundreds of samples in multiplex. Nature Methods, 2008, March, Vol.5 No.3). By such, a plurality of subject samples may be subjected to simultaneous detection, which may improve efficiency and decrease detecting cost.

Fig.2 is a flow chart showing a method of detecting a copy number variation according to another embodiment of the present disclosure.

Step 202, a genomic DNA molecule of a subject sample is randomly fragmented to obtain DNA fragments, which may be performed by enzyme digestion, pulverization, ultrasound, or HydroShear method. The ultrasound method is preferably used, such as S-series from Covaris Company (based on AFA technique, when sound energy/mechanical energy released from a sensor passing through a DNA sample, gas is dissolved and forms a bubble. After the energy is removed, the bubble is ruptured to generate a capability for fragmenting the DAN fragments. By setting conditions such as a certain energy intensity and time interval, the DNA molecule may be fragmented into fragments having a certain size. A specific principal and method may refer to specification for S-series of Covaris Company), is used for fragmenting the DNA molecule into fragments having a certain concentrated size.

Step 204, the DNA fragments are subjected to sequencing, to obtain sequencing sequences of the DNA fragments, namely, reads. The reads obtained from sequencing may have a certain length within a range. The reads having a fixed length may be obtained by subjecting reads of the DNA fragments to truncation. The reads of the DNA fragments used in the present example hereafter refer to reads having a fixed length. Methods used for sequencing may be high-throughput sequencing methods such as Illumina/Hiseq2000, ABI/SOLiD, Roche/454. Sequencing type may be Single-end sequencing or Pair-end sequencing, and sequencing length may be 50 bp to 1500 bp. In an example of the present disclosure, the used sequencing platform is Illumina/Hiseq2000, the sequencing type is Pair-end sequencing, and obtained DNA sequence molecules with a length of 100 bp has a relationship of a bi-directional position. The sequencing depth may be determined based on the length of variation fragment in a chromosome. The higher sequencing depth, the higher sensitivity of the detection, i.e., the smaller length of deletion fragment and duplication fragment which can be detected. In an example of the present disclosure, the sequencing volume of the human subject sample ranges of $2\sim900\times10^8$ of reads.

Step 206, the reads are aligned to a genome reference sequence, to obtain site information of the reads in a genome.

Step 208, uniquely-mapped reads aligned to a genome reference sequence is selected.

Step 210, the number of uniquely-mapped reads falling into each of the plurality of windows is determined. For each subject sample, the number of uniquely-mapped reads falling into each of the plurality of windows is calculated (recorded as $n_{i,j}$, subscripts I and j respectively represent No. of the plurality of window and No. of the subject samples, for distinguishing which are omitted for brevity).

Step 212, an average GC content for each of the plurality of window in the genome is determined to determine a correction coefficient for each of the plurality of windows; and the corrected number of uniquely-mapped reads for each of the plurality of windows is obtained based on the correction coefficient. Such step is mainly to subject the number of uniquely-mapped reads falling into each of the plurality of windows to the correction based on the GC content for each of the plurality of windows, which may be called as batches correction or GC correction.

The average GC content (recorded as $GC_{i,j}$) of the uniquely-mapped reads falling into each of the plurality of windows is calculated. The calculation of the average GC

content ($GC_{i,j}$) is calculating average GC content of all uniquely-mapped reads falling into each of the plurality of windows. During calculation, the total number of bases G and C of all uniquely-mapped reads is recorded as N_{gc}, a total length of all uniquely-mapped reads is recorded as L, then $GC_{i,j} = \frac{N_{gc}}{L}$.

To correct a difference of data volume of the subject sample, each of the plurality of windows are grouped based on $GC_{i,j}$, i.e., windows having same $GC_{i,j}$ are grouped into one group, a median or an arithmetic mean value $m_{g,j}$ for the number of uniquely-mapped reads in each group is determined, which is divided by a median or an arithmetic mean value m_j for the number of uniquely-mapped reads in the whole genome level, to obtain the correction coefficient $c_{g,j}$ ($c_{g,j} = m_{g,j}/m_j$), in which the subscript g represents GC contents of different groups. The original number of uniquely-mapped reads falling into each of the plurality of windows $n_{i,j}$ is multiplied by the correction coefficient $c_{g,j}$ to obtain a corrected value of the number of uniquely-mapped reads falling into each of the plurality of windows (recorded as $n_{i,j}'$).

Step 214, the corrected number of uniquely-mapped reads falling into each of the plurality of windows is divided by the expected number for the corresponding each of the plurality of windows, to obtain corrected number of uniquely-mapped reads for each of the plurality of windows, i.e., copy rate. The expected number for the corresponding each of the plurality of windows is obtained by a control set consisting of normal samples. Such steps is mainly to subject the number of uniquely-mapped reads for each of the plurality of windows to correction based on data of normal samples, which may be called as window correction.

In the sample of control set, a percentage of relative number of uniquely-mapped reads ($P_{i,j}$) is defined as a ratio of the number of uniquely-mapped reads within a window ($n_{i,j}'$) to the total number of uniquely-mapped reads for the whole genome (N_j), i.e., $P_{i,j} = \frac{n_{i,j}'}{N_j}$; and then an average percentage of each of the plurality of windows in the

control set $\overline{P_i}$ is calculated, i.e., $\overline{P_i} = \frac{1}{n}\sum_{j=1}^{n} P_{i,j}$. In the subject sample, the copy rate of each of the plurality of windows $r_{i,j}$ is obtained by dividing the corrected number of uniquely-mapped reads $n_{i,j}$ by the expected number within the window (multiplying the total number of uniquely-mapped reads for the whole genome with the percentage of the number of uniquely-mapped reads falling into each of the plurality of windows), i.e., $r_{i,j} = \frac{n_{i,j}}{\overline{P_i} \times N_j}$.

For selecting the control set, a method of library constructing, sequencing reagent and sequencing type should be consistent with the sample to be tested as much as possible, so as to improve effect of correcting the sample to be tested by the control sample. The sample in the control set should be a normal sample with a sample volume being 30 or more.

Step 216, the predetermined number of windows is selected as a candidate of a CNV breakpoint within a whole genome range based on a p value presenting a significance level of copy number variation between two sides of each of the plurality of windows, to obtain a significance value of the candidate of each of the CNV breakpoints, namely, a p value.

1) Selection of the candidate of CNV breakpoints may comprise: for all windows of the whole genome, calculating a difference of copy number changes for a certain number of windows between two sides of each of the plurality of windows (the number of windows is normally more than 30 or satisfies a restriction of the lowest sample amount for a detection model, to make the detection model having significance difference); and selecting a certain number (for example, 1% of the total number of windows) of sites (corresponding to the plurality of windows) as the candidate of the CNV breakpoints (namely, a demarcation point of each CNV fragments), based on the significant difference level in the whole genome range (from small value to large value of the p value).

2) Initialization may comprise: a set of all ranked breakpoints is recorded as $B_c = \{b_1, b_2, ...b_k .., b_s\}$, then two adjacent breakpoints k+1 and k-1 exist at two sides of each breakpoint k. By calculating a significance difference between a set of copy number from k-1 to k and a set of copy number from k to k+1, a p value representing a significance difference between two sides of each breakpoint is obtained. For example, in an example, a Run test in non-parameter detection is selected, which evaluates the significant different

of two populations by homogeneous status of distribution with mixed elements of the two populations, specifically referring to Wald, A. & Wolfowitz, J. On a test whether two samples are from the same population. The Annals of Mathematical Statistics 11, 147-162 (1940).

Step 218, the candidate of the CNV breakpoint having the least significance is removed for each time, and the significance of two candidates of the CNV breakpoint adjacent to the removed candidate of the CNV breakpoint is recalculated, which is subjected to cycling and iterating until the significances of all candidates of the CNV breakpoint are less than a final p value (i.e., a final cut-off), in which the final cut-off is obtained based on the control set.

Iteration merge may comprise: by continuous cycling and iterating, the candidate of the breakpoints having the least significance is removed for each time, and the p values for the two adjacent breakpoints are recalculated until all of the p values are less than the final p value.

The final p value is obtained by: for example, subjecting the control sample to the above operation of iteration merger; recording the largest p value for each iteration merger; performing the above operation of iteration merger until merging into one fragment. By then, according to a change trend of the largest p value, the largest p value for one time of iteration merger corresponding to one site having the most dramatic change of the p value (i.e., selecting one site having the most obvious change of slope in a curve of the p value change (a site having the largest curvature)) or that of the previous time of iteration merger is taken as the final cut-off.

The above steps 216 and 218 may also called as fragmentation. Notice: when selecting the windows and the breakpoints in 1) and 2) of the step 214, a cyclization with a single chromosome or a whole genome may be considered to be performed. The cyclization with the single chromosome refers to: when calculating the plurality of windows near the starting point of a chromosome, if the effective window number on the left side is insufficient for statistical detection, then sufficient number of windows for calculation is obtained reversely from the end point of such chromosome; in a similar way, for those positions, of which the sufficient effective number of windows near the right side of the end point is not able to be obtained, are obtained from the front point of the chromosome. Such operation makes those windows located at the front point and the end point still can be calculated. The cyclization with the whole genome is: indexing the end point of the former chromosome when the effective number of windows located at the front point of each chromosome, while indexing the front point of the latter chromosome when the effective

number of windows located at the end point is insufficient, while chromosome 1 connects to chromosome Y.

After the step 218, the method further comprises subjecting the sequence between two CNV breakpoints to a confidence selection, in which the confidence selection comprises steps of: determining a confidence interval of normal copy rate using the control set, based on a distribution of the copy rate; and determining an abnormality presenting in the sequence between the two CNV breakpoints, if the mean value of the copy rate within the sequence is outside the confidence interval. In an example, the copy rate the copy rate fits a normal distribution, and the confidence interval is 95%. The fragmenting result is subjected to a filtration by such step, to obtain a reliable result. If the mean value of $r_{i,j}'$ is smaller than a lower cut-off value or larger than an upper cut-off value, then a corresponding result is output as a positive result.

Selection of cut-off value may comprise: calculating a distribution of copy rate for a window in each of the control samples, which is based on central-limit theorem, in which reads in the window are random, then the copy rate r fits a normal distribution, and quantiles of the left and the right with a significance level being as 0.05 are selected. A mean value thereof is calculated in the control set, which are taken as the upper and lower cut-off value for screening copy number variation.

In the above examples, accuracy of detection result is improved by the batches correction and the window correction. By introducing the control set, accuracy may also be improved by enlarging the control set, which may reduce requirement for initial DNA amount.

Fig.3 is a flow chart showing a method of detecting a copy number variation according to a further embodiment of the present disclosure. Fig.3 includes a flow chart showing a method of handling a control set consisting of normal samples (3A) and a flow chart showing a method of handling a subject sample (3B). The control set is mainly used for obtaining data for correcting the subject sample, and for the final cut-off value being as a terminating condition of iterating the subject sample.

As shown in Fig.3, the flow chart 3A comprises:

Step 301A, extracting DNA molecule from the control sample;

Step 311A, randomly fragmenting the DNA molecule of the control sample into DNA fragments, which are then subjected to sequencing, to obtain sequencing sequence data of the DNA molecule of the control sample, i.e., reads;

Step 312A, aligning the reads of the control sample to a reference genome sequence;

Step 313A, calculating the number of uniquely-mapped reads falling into each of the

plurality of windows, i.e., the number of uniquely-mapped reads;

Step 314A, subjecting the control sample to the batches correction;

Step 315A, obtaining the expected number of the plurality of windows by the control sample, for subjecting the control sample to the window correction.

Step 316A, selecting breakpoints and fragmenting. The step of selecting a candidate of CNV breakpoints comprises: removing the candidate of the CNV breakpoint having the least significance for each time and recalculating the p value of two candidates of the CNV breakpoint adjacent to the removed candidate of the CNV breakpoint, cycling and iterating until the number of rest fragments equals to a predetermined value (such as 24).

Step 317A, determining a final cut-off value. By calculating a mean value of current final p value, the final p value may be effectively obtained, as the final cut-off value of the condition for terminating iteration with the subject sample.

Step 3B comprises:

Step 310B, extracting DNA molecule of a subject sample;

Step 311B, randomly fragmenting the DNA molecule of the control sample into DNA fragments, which are then subjected to sequencing, to obtain reads of the DNA molecule of the control sample;

Step 312B, aligning the reads of the DNA molecule of the control sample to a reference genome sequence;

Step 313B, calculating the number of uniquely-mapped reads falling into each of the plurality of windows, i.e., the number of uniquely-mapped reads.

Step 314B, subjecting the subject sample to batches correction.

Step 315B, subjecting the subject sample to window correction based on the expected number of uniquely-mapped reads falling into each of the plurality of windows for the control sample;

Step 316B, selecting breakpoints and fragmenting;

Step 317B, subjecting obtained result to filtering.

For selecting the control set, a method of library constructing, sequencing reagent and sequencing type should be consistent with the sample to be tested as much as possible, so as to improve effect of correcting the sample to be tested by the control sample. The sample in the control set should be a normal sample with a sample volume being 30 or more.

Fig.4 is a brief flow chart showing a method of chromosomal CNV analysis according to an embodiment of the present disclosure.

As shown in Fig.4, step 401 is DNA extraction and sequencing: after genomic DNA is

extracted in accordance with specification of Tiangen DP327-02 Kit, a library is constructed in accordance to a process of library construction based on Illumina/Hiseq2000 standard. During such process, those DNA molecules concentrating in 500 bp are ligated with adaptors at both ends thereof for sequencing, in which each sample is also added with different indexes respectively, by which data from the plurality of samples may be distinguished in a result of sequencing for one time.

Step 402, sequence alignment: A sequencing method of Illumina/Hiseq2000 is used for sequencing (other sequencing methods such as ABI/SOLiD may also achieve a same or similar effect), each sample may obtain reads of DNA fragments having a certain length, which is aligned to a standard human genome reference sequence in NCBI database using SOAP2, to obtain information of a corresponding position where the reads locate in the genome. To avoid interference of repetitive sequence to CNV analysis, only those reads uniquely mapped to the human genome reference sequence are selected, i.e., those reads can be mapped to the human genome reference sequence only once, also called as the number of uniquely-mapped reads, as the effective data for subsequent CNV analysis.

Step 403, PSCC analysis. A serious of bioinformatics methods of detecting a copy number variation in a whole genome copy which are independently developed by the inventor of the present disclosure are utilized, which comprises: subjecting the subject sample to batches correction, subjecting the subject sample to window correction based on a control set, normalization and fragmentation (segmentation).

Step 404, the fragments having the determined copy number in the step 403 are subjected to CNV analysis, a copy rate≤0.7 and a copy rate ≥1.3 of the subject sample are taken as a cut-off value for detecting deletion and duplication of the fragments, by which the fragments having the copy number variation in a whole genome level is obtained by analysis, and then obtained result is subjected to visualization.

In the above examples, used software algorithm is a series procedure of detecting a copy number variation for a whole genome developed by Shenzhen BGI, which are called as PSCC. It can generate data by the Next-Generation sequencing technology, subjecting the subject sample to batches correction, and performing data correction with a control set, normalization and segmentation, to estimate degree and size of copy number variation of the subject sample. Under a condition of a lower sequencing depth (50 M of sequencing short-sequences), about 0.5 Mb of fragments having a single copy number variation (CNV) may be detected.

Fig.5 is a schematic diagram showing a system of detecting a copy number variation

according to an embodiment of the present disclosure. As shown in Fig.5, the system may comprise: a reads obtaining unit 51, for obtaining reads from at least one part of a nucleic acid molecule of a sample; a uniquely-mapped reads determining unit 52, for determining uniquely-mapped reads aligned to a genome reference sequence based on obtained reads; the number of uniquely-mapped reads calculating unit 53, for dividing the genome reference sequence into a plurality of windows, and calculating the number of uniquely-mapped reads falling into each of the plurality of windows; the number of uniquely-mapped reads correcting unit 54, for subjecting the number of uniquely-mapped reads falling into each of the plurality of windows to a GC correction, and to a correction based on the expected number of uniquely-mapped reads corrected by a control set to obtain the corrected number of uniquely-mapped reads; a candidate of breakpoint selecting unit 55, for calculating a significance of value population consisting of the corrected number of uniquely-mapped reads between two sides of a demarcation point, with the demarcation point being selected as a starting point or a terminal point of each of the plurality of windows, to select the demarcation point having a smaller value of the significance as a candidate of a CNV breakpoint; a breakpoint determining unit 56, for calculating a significance of two value populations consisting of the corrected number of uniquely-mapped reads respectively contained in two sequences, with one sequence ranging from each of the CNV breakpoint to an adjacent upstream CNV breakpoint, and the other sequence ranging from each of the CNV breakpoint to an adjacent downstream CNV breakpoint, and removing the candidate of the CNV breakpoint having the least significance for each time and recalculating the significance of two candidates of the CNV breakpoint adjacent to the removed candidate of the CNV breakpoint, cycling and iterating until the significances of all candidates of the CNV breakpoint are less than a final cut-off, to determine the CNV breakpoint; the final cut-off is obtained based on the control set consisting of normal samples. When the number of uniquely-mapped reads calculating unit 53 divides the plurality of windows, each of the plurality of windows may contain the same number of reference unique reads, or each of the plurality of windows may have a same length. In an example, in the candidate of breakpoint selecting unit 55, a cyclization with a single chromosome or a whole genome is performed when selecting the candidate of the CNV breakpoint.

In the above examples, the uniquely-mapped reads determining unit determines uniquely-mapped reads which can be uniquely aligned to the (genomic) reference sequence based on obtained reads, the number of uniquely-mapped reads correcting unit subjects the number of uniquely-mapped reads falling into each of the plurality of windows

to a correction, and the candidate of breakpoint selecting unit and the breakpoint determining unit perform cycling and iterating of gene significant difference for selecting the CNV breakpoints, by which CNV detection can be completed, which may accurately detect a region comprising the copy number variation including smaller micro-deletion/micro-duplication.

Fig.6 is a schematic diagram showing a system of detecting a copy number variation according to anther embodiment of the present disclosure. As shown in Fig.6, the system comprises: a reads obtaining unit 51, a uniquely-mapped reads determining unit 52, the number of uniquely-mapped reads calculating unit 53, the number of uniquely-mapped reads correcting unit 64, a candidate of breakpoint selecting unit 55 and a breakpoint determining unit 56. The reads obtaining unit 51, the uniquely-mapped reads determining unit 52, the number of uniquely-mapped reads calculating unit 53, the candidate of breakpoint selecting unit 55 and the breakpoint determining unit 56 may refer to specific description in Fig.5, which is omitted for brevity. The number of uniquely-mapped reads correcting unit 64 further includes a GC correction module 641 and a window correction module 642. In which, the GC correction module is used for grouping the plurality of windows based on a GC content, and obtaining a correction coefficient based on a mean value of the number of uniquely-mapped reads within one group and a mean value of the number of uniquely-mapped reads for all of the plurality of windows, and then subjecting the number of uniquely-mapped reads falling into each of the plurality of windows to the correction, to obtain the GC-corrected number of uniquely-mapped reads; the window correction module 642 is used for calculating a ratio of the GC-corrected number of uniquely-mapped reads falling into each of the plurality of windows in the control set to the total number of uniquely-mapped reads; obtaining a mean value of the ratios for all windows corresponding to the control set; and calculating the expect number of uniquely-mapped reads for each of the plurality of windows in the sample based on obtained mean value of the ratios and the total number of uniquely-mapped reads in the sample, also called as copy rate.

In an example of the present disclosure, the system further comprises: a breakpoint filtering unit 67, for determining a confidence interval of the normal corrected number of uniquely-mapped reads using the control set, based on a distribution of the corrected number of uniquely-mapped reads; and determining an abnormality presenting in the sequence between the two CNV breakpoints, if the mean value of the corrected number of uniquely-mapped reads within the sequence is outside the confidence interval. In an example, the corrected number of uniquely-mapped reads fits a normal distribution, and

the confidence interval is 95%.

In an example, the system further comprises: means for obtaining the sample from a human, with the sample comprising amniotic fluid obtained by amniocentesis, villus obtained by chorionic villi sampling, umbilical cord blood obtained by percutaneous umbilical blood sampling, spontaneous miscarrying fetus tissue or human peripheral blood; and/or means for obtaining a genomic DNA of the sample by a DNA extraction method such as a salting-out method, a column chromatography method, a beads method, or a SDS method; and/or means for subjecting the genome DNA of the sample to random fragmenting by enzyme digestion, pulverization, ultrasound, or HydroShear method, to obtain DNA fragments; and/or means for subjecting the DNA fragments to single-end sequencing or pair-end sequencing, to obtain the reads of the DNA fragments. In an example, different samples are distinguished by adding different indexes to each of the DNA fragments of the samples.

Functions of each unit in Fig.5 and Fig.6, may refer to specification of the corresponding part above according to the method in examples of the present disclosure, which are omitted for brevity.

Those skilled in the art will recognize hardware, firmware and software configurations in these cases are interchangeable, to achieve the function of each specific application as much as possible. .

Reference will be made in detail to examples of the present disclosure. It would be appreciated by those skilled in the art that the following examples are explanatory, and cannot be construed to limit the scope of the present disclosure. If the specific technology or conditions are not specified in the examples, a step will be performed in accordance with the techniques or conditions described in the literature in the art or in accordance with the product instructions. If the manufacturers of reagents or instruments are not specified, the reagents or instruments may be commercially available. Descriptions in following brackets respectively illustrate catalog No. of different manufacturers for various reagents or kits. The adaptor and index used for sequencing derive from Multiplexing Sample Preparation Oligonutide Kit of Illumina Company.

Application example 1. Detection with 2 samples having chromosome number variation and 6 samples having micro-deletion

1. DNA extraction

8 samples (hereinafter referred as Sample1, Sample2, Sampel3, ..., Sample8) were subjected to DNA extraction in accordance with specification of TIANamp Micro DNA Kit

(DP316) of Tiangen. The extracted DNA was subjected to library construction in accordance with revised specification of Illumina/Hiseq2000. Those DNA molecules concentrating in a length of 500 bp were ligated with an adaptor at both ends thereof for sequencing. Each sample was added with different indexes respectively, which were then subjected to hybridization with a complementary adaptor on Flowcell surface, by which nucleic acid molecules clustering grew under a certain condition, which were then subjected to Pair-end sequencing on Illumina Hiseq2000, to obtain sequences of DNA fragments having a length of 100 bp.

In details, about 100 ng of DNA obtained from the above amniotic fluid sample (Quant-IT dsDNA HS Assay kit), were subjected to library construction in accordance with revised specification of Illumina/Hiseq2000, of which detailed procedure could refer to *prior art (specification of standard library construction of Illumina/Solexa provided in* http://www.illumina.com/). The obtained DNA library and insert fragments were determined having a length of 500 bp using 2100Bioanalyzer (Agilent), which were subjected to sequencing on computer after accurate quantification by QPCR.

2. Sequencing: in the current example, above DNA obtained from the 8 samples were subjected to sequencing on computer in accordance with specifications of ClusterStation and Hiseq2000 (PEsequencing) officially published by Illumina/Solexa, to obtain about 5G of data volume for each sample, which were distinguished by ligated indexes. Using alignment software SOAP2, DNA sequence obtained by sequencing was aligned to a human genome reference sequence Build 36 in NCBI database (hg18; NCBIBuild36), to obtain information of a corresponding position where the reads locate in the genome.

3. Data analysis

a) basic statistics: calculating the number of uniquely-mapped reads falling into each of the plurality of windows (i.e., the number of uniquely-mapped reads, recorded as $n_{i,j}$, subscripts I and j respectively represented No. of the plurality of window and No. of the subject samples, for distinguishing which are omitted for brevity), and an average GC content (recorded as $GC_{i,j}$).

b) batches correction: to correct a difference of sample data volume, the plurality of windows were grouped based on the average GC content of reads falling into thereof. A median or an arithmetic average for each group was divided by a median or an arithmetic average for the whole genome, to obtain a correction coefficient C_g, in which the subscript g represented GC contents of different groups. The original number of uniquely-mapped

reads falling into each of the plurality of windows $n_{i,j}$ was multiplied by the correction coefficient c_g to obtain a corrected value of the number of uniquely-mapped reads falling into each of the plurality of windows (recorded as $n_{i,j}'$).

c) data correction: sequencing data obtained from 90 samples in YH population were selected as the control set. In the sample of control set, a percentage of relative number of uniquely-mapped reads ($P_{i,j}$) was defined as a ratio of the correct number of uniquely-mapped reads within a window ($n_{i,j}'$) to the total number of uniquely-mapped reads for the whole genome (N_j), i.e., $P_{i,j} = \frac{n_{i,j}'}{N_j}$; and then an average percentage of each of the plurality of windows in the control set $\overline{P_i}$ was calculated, i.e., $\overline{P_i} = \frac{1}{n}\sum_{j=1}^{n} P_{i,j}$. In the subject sample, the copy rate of each of the plurality of windows $r_{i,j}$ was obtained by dividing the corrected number of uniquely-mapped reads $n_{i,j}'$ by the expected number within the window (multiplying the total number of uniquely-mapped reads for the whole genome with the percentage of the number of uniquely-mapped reads falling into each of the plurality of windows), i.e., $r_{i,j} = \frac{n_{i,j}'}{\overline{P_i} \times N_j}$.

d) fragmentation (segmentation)

① selection of a candidate of breakpoints: for all $r_{i,j}$ n the whole genome, calculating a difference of copy number changes for 100 of windows between two sides thereof; and selecting 10000 of points as the candidate of the CNV breakpoints, based on the significant difference level in the whole genome range (from small value to large value of the p value).

② initialization: a set of all ranked breakpoints was recorded as $B_c = \{b_1, b_2, ...b_k .., b_s\}$, then two adjacent breakpoints k+1 and k-1 existed at two sides of each breakpoint k. By calculating a significance difference between a set of copy number from k-1 to k and a set of copy number from k to k+1, a p value representing a significance difference between two sides of each breakpoint was obtained (a Run test of non-parameter detection was used in the current example).

③ Iteration merger included: by continuous cycling and iterating, the candidate of the breakpoints having the least significance is removed for each time, and the p values of two adjacent breakpoints are recalculated until all of the p values are less than $1E^{-50}$.

Notice: when selecting the windows and the breakpoints in steps ① and ②, a cyclization with a whole genome was performed, to make those windows located at the front point and the end point still can be calculated.

e) cut-off value and filtration: a result obtained after fragmentation was subjected to filtering. If the mean value of $r_{i,j}^{'}$ was smaller than 0.7 or larger than 1.3, a corresponding result was output as a positive result.

f) visualization of the positive result

4. Statistical calculation

For the 8 samples, the detection result and verification result were shown in details in Table below.

In which the verification result was obtained by CGH chip (comparative genomic hybridization chip). Human Genome CGH Microarray Kit (Agilent Technologies Inc.) was used in the experiments, which was in accordance with specification provided by manufacturer.

Table 1. CNV results of 8 samples

Sample No.	Detection result	length	Verification result	Determination
Sample1	T7	159.14M	T7	consistent
Sample2	XYY	59.37M	XYY	consistent
Sample3	chr5:1.. 36877903 deletion chr18: 39024931.. 76038278 addition	36.88M 37.01M	5p15.3~p13.2 (183931~36816731)×1; 18p12.3~q23 (39086755~76067279)×3	consistent
Sample4	chr5:1..17710089 deletion	17.71M	5p15.33~p15.1×1	consistent
Sample5	chr15:21236149..26219186 deletion	4.98M	15q11.2~q13.1×1	consistent
Sample6	chr1:1..5065299 addition	5.07M	1p36.33~p36.32×3	consistent

Sample7	chr5:105562861..106156933 deletion	0.59M	5q21.3×1	consistent
Sample8	chr9:6851755..7248416 deletion	0.40M	9p24.1×1	consistent

In which, chr represented chromosome, T7 represented trisomy of chromosome 7, and XYY represented trisomy variation of sex chromosome.

Fig.7A-7H show detection results of eight samples respectively in an example of the present disclosure.

It could be seen from the above Table 1 and Fig. 7A-7H that, small to 0.4M of fragments having micro-deletion and large to entire chromosome number variation could be accurately detected and located by the method of the present disclosure, which proved that efficiency and accuracy thereof are both excellent.

Comparing with those analysis methods of detecting the copy number variation which has been reported so far, advantages of the present disclosure mainly comprises:

1) resolution: by using about 50 M of data, 0.5 M of region having micro-deletion can be accurately detected.

2) expandability: in addition to increasing sequencing volume, the accuracy may also be improved by enlarging the control set, which may reduce requirement for initial DNA amount.

3) more stable and more comprehensive: there is few detailed description in reported articles, while the present disclosure involves various aspect such as batches correction and population correction with data, as well as optimization with condition for fragmentation, etc.

The method of the present disclosure may be used to subject those target patients to whole genome copy number variation detection, which is benefit to provide genetic counseling and basis for clinical decision, by which may make an accurate pathology determination for patients suffering micro-deletion syndrome. Those target patients of the present disclosure may be patients suffering from micro-deletion or potential carriers. The target patients are used for illustrating the present disclosure, but not to limit the scope of the present disclosure.

Although explanatory embodiments have been shown and described, it would be appreciated by those skilled in the art that the above embodiments cannot be construed to limit the present disclosure, and changes, alternatives, and modifications can be made in the embodiments without departing from spirit, principles and scope of the present disclosure.

WHAT IS CLAIMED IS:

1. A method of detecting a copy number variation comprising following steps:

obtaining reads from at least one part of a nucleic acid molecule of a sample,

determining uniquely-mapped reads aligned to a genomic reference sequence based on the obtained reads,

dividing the genomic reference sequence into a plurality of windows, and calculating the number of uniquely-mapped reads falling into each of the plurality of windows,

subjecting the number of uniquely-mapped reads falling into each of the plurality of windows to a GC correction, and to a correction based on the expected number of uniquely-mapped reads corrected by a control set to obtain the corrected number of uniquely-mapped reads,

calculating a significance of value population consisting of the corrected number of uniquely-mapped reads between two sides of a demarcation point, with the demarcation point being selected as a starting point or a terminal point of each of the plurality of windows, to select the demarcation point having a smaller value of the significance as a candidate of a CNV breakpoint;

calculating a significance of two value populations consisting of the corrected number of uniquely-mapped reads respectively contained in two sequences, with one sequence ranging from each of the CNV breakpoint to an adjacent upstream CNV breakpoint, and the other sequence ranging from each of the CNV breakpoint to an adjacent downstream CNV breakpoint, and

removing the candidate of the CNV breakpoint having the least significance for each time and recalculating the significance of two candidates of the CNV breakpoint adjacent to the removed candidate of the CNV breakpoint, cycling and iterating until the significances of all candidates of the CNV breakpoint are less than a final cut-off, to determine the CNV breakpoint.

2. The method of claim 1, further comprising a step of:

subjecting the at least one part of the nucleic acid molecule of the sample to sequencing, to obtain the reads.

3. The method of claim 1, wherein each of the plurality of windows contains the same number of reference unique reads, or each of the plurality of windows has a same length.

4. The method of claim 1 or 2, wherein the final cut-off is obtained based on the control set consisting of normal samples.

5. The method of claim 1, wherein the step of subjecting the number of uniquely-mapped reads falling into each of the plurality of windows to the GC correction and to a correction based on the expected number of uniquely-mapped reads corrected by a control set to obtain the corrected number of uniquely-mapped reads further comprises:

grouping the plurality of windows based on a GC content, and obtaining a correction coefficient based on a mean value of the number of uniquely-mapped reads within one group and a mean value of the number of uniquely-mapped reads for all of the plurality of windows, and then subjecting the number of uniquely-mapped reads falling into each of the plurality of windows to the correction, to obtain the GC-corrected number of uniquely-mapped reads;

and/or

the expected number of uniquely-mapped reads corrected by the control set is obtained by following steps:

calculating a ratio of the GC-corrected number of uniquely-mapped reads falling into each of the plurality of windows in the control set to the total number of uniquely-mapped reads;

obtaining a mean value of the ratios for all windows corresponding to the control set; and

calculating the expect number of uniquely-mapped reads for each of the plurality of windows in the sample based on obtained mean value of the ratios and the total number of uniquely-mapped reads in the sample.

6. The method of claim 5, wherein a cyclization with a single chromosome or a whole genome is performed when selecting the candidate of the CNV breakpoint.

7. The method of claim 1or 2, wherein after the CNV breakpoint is determined, the method further comprises:

subjecting the sequence between two CNV breakpoints to a confidence selection, wherein the confidence selection comprises steps of:

determining a confidence interval of the normal corrected number of uniquely-mapped reads using the control set, based on a distribution of the corrected number of uniquely-mapped reads; and

determining an abnormality presenting in the sequence between the two CNV breakpoints, if the mean value of the corrected number of uniquely-mapped reads within the sequence is outside the confidence interval.

8. The method of claim 7, wherein the corrected number of uniquely-mapped reads fits a normal distribution, and the confidence interval is 95%.

9. The method of any one of claims 1 to 8, wherein the method further comprises:

obtaining the sample from a human, with the sample comprising amniotic fluid obtained by amniocentesis, villus obtained by chorionic villi sampling, umbilical cord blood obtained by percutaneous umbilical blood sampling, spontaneous miscarrying fetus tissue or human peripheral blood;

and/or

obtaining a genomic DNA of the sample by a DNA extraction method such as a salting-out method, a column chromatography method, a beads method, or a SDS method;

and/or

subjecting the genome DNA of the sample to random fragmenting by enzyme digestion, pulverization, ultrasound, or HydroShear method, to obtain DNA fragments;

and/or

subjecting the DNA fragments to single-end sequencing or pair-end sequencing, to obtain the reads of the DNA fragments.

10. The method of claim 1, wherein the method further comprises:

adding different indexes to each of the DNA fragments of the samples, to distinguish different samples.

11. A system for detecting a copy number variation comprising:

a reads obtaining unit, for obtaining reads from at least one part of a nucleic acid molecule of a sample;

a uniquely-mapped reads determining unit, for determining uniquely-mapped reads aligned to a genome reference sequence based on obtained reads;

the number of uniquely-mapped reads calculating unit, for dividing the genome reference sequence into a plurality of windows, and calculating the number of uniquely-mapped reads falling into each of the plurality of windows;

the number of uniquely-mapped reads correcting unit, for subjecting the number of uniquely-mapped reads falling into each of the plurality of windows to a GC correction, and to a correction based on the expected number of uniquely-mapped reads corrected by a control set to obtain the corrected number of uniquely-mapped reads;

a candidate of breakpoint selecting unit, for calculating a significance of value population consisting of the corrected number of uniquely-mapped reads between two sides of a demarcation point, with the demarcation point being selected as a starting point or a terminal point of each of the plurality of windows, to select the demarcation point having a smaller value of the significance as a candidate of a CNV breakpoint;

a breakpoint determining unit, for calculating a significance of two value populations consisting of the corrected number of uniquely-mapped reads respectively contained in two sequences, with one sequence ranging from each of the CNV breakpoint to an adjacent upstream CNV breakpoint, and the other sequence ranging from each of the CNV breakpoint to an adjacent downstream CNV breakpoint, and removing the candidate of the CNV breakpoint having the least significance for each time and recalculating the significance of two candidates of the CNV breakpoint adjacent to the removed candidate of the CNV breakpoint, cycling and iterating until the significances of all candidates of the CNV breakpoint are less than a final cut-off, to determine the CNV breakpoint.

12. The system of claim 11, wherein each of the plurality of windows contains the same number of reference unique reads, or each of the plurality of windows has a same length.

13. The system of claim 11, wherein the final cut-off is obtained based on the control set consisting of normal samples.

14. The system of claim 11, wherein the number of uniquely-mapped reads correcting unit comprises:

a GC correction module, for grouping the plurality of windows based on a GC content, and obtaining a correction coefficient based on a mean value of the number of uniquely-mapped reads within one group and a mean value of the number of uniquely-mapped reads for all of the plurality of windows, and then subjecting the number of uniquely-mapped reads falling into each of the plurality of windows to the correction, to obtain the GC-corrected number of uniquely-mapped reads;

a window correction module, for calculating a ratio of the GC-corrected number of

uniquely-mapped reads falling into each of the plurality of windows in the control set to the total number of uniquely-mapped reads; obtaining a mean value of the ratios for all windows corresponding to the control set; and calculating the expect number of uniquely-mapped reads for each of the plurality of windows in the sample based on obtained mean value of the ratios and the total number of uniquely-mapped reads in the sample.

15. The system of claim 11, wherein after the CNV breakpoint is determined by the breakpoint determining unit, the system further comprises:

a breakpoint filtering unit, for determining a confidence interval of the normal corrected number of uniquely-mapped reads using the control set, based on a distribution of the corrected number of uniquely-mapped reads; and determining an abnormality presenting in the sequence between the two CNV breakpoints, if the mean value of the corrected number of uniquely-mapped reads within the sequence is outside the confidence interval.

16. The system of claim 15, wherein the corrected number of uniquely-mapped reads fits a normal distribution, and the confidence interval is 95%.

17. The system of any one of claims 11 to 16, wherein the system further comprises:

means for obtaining the sample from a human, with the sample comprising amniotic fluid obtained by amniocentesis, villus obtained by chorionic villi sampling, umbilical cord blood obtained by percutaneous umbilical blood sampling, spontaneous miscarrying fetus tissue or human peripheral blood;

and/or

means for obtaining a genomic DNA of the sample by a DNA extraction method such as a salting-out method, a column chromatography method, a beads method, or a SDS method;

and/or

means for subjecting the genome DNA of the sample to random fragmenting by enzyme digestion, pulverization, ultrasound, or HydroShear method, to obtain DNA fragments;

and/or

means for subjecting the DNA fragments to single-end sequencing or pair-end sequencing, to obtain the reads of the DNA fragments.

18. The system of claim 11, wherein different samples are distinguished by adding different indexes to each of the DNA fragments of the samples.

19. The system of claim 11, wherein in the candidate of breakpoint selecting unit, a cyclization with a single chromosome or a whole genome is performed when selecting the candidate of the CNV breakpoint.