

(19)



(11)

EP 3 347 466 B9

(12)

CORRECTED EUROPEAN PATENT SPECIFICATION

(15) Correction information:

Corrected version no 1 (W1 B1)
Corrections, see
Claims FR 11

(51) International Patent Classification (IPC):

C12N 15/10^(2006.01) C12P 19/34^(2006.01)

(52) Cooperative Patent Classification (CPC):
(C-Sets available)

G16B 30/10; C12N 15/1093; G16B 30/00 (Cont.)

(48) Corrigendum issued on:

26.06.2024 Bulletin 2024/26

(86) International application number:

PCT/US2016/050750

(45) Date of publication and mention
of the grant of the patent:

03.01.2024 Bulletin 2024/01

(87) International publication number:

WO 2017/044609 (16.03.2017 Gazette 2017/11)

(21) Application number: **16845040.1**

(22) Date of filing: **08.09.2016**

(54) **GENETIC COPY NUMBER DETERMINATION USING HIGH THROUGHPUT MULTIPLEX SEQUENCING OF SMASHED NUCLEOTIDES**

BESTIMMUNG DER ANZAHL GENETISCHER KOPIEN MIT
HOCHDURCHSATZ-MULTIPLEX-SEQUENZIERUNG VON ZERTRÜMMERTEN NUKLEOTIDEN
DÉTERMINATION DU NOMBRE DE COPIES GÉNÉTIQUES AU MOYEN D'UN SÉQUENÇAGE
MULTIPLEX À HAUT DÉBIT DE NUCLÉOTIDES SMASH

(84) Designated Contracting States:

**AL AT BE BG CH CY CZ DE DK EE ES FI FR GB
GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO
PL PT RO RS SE SI SK SM TR**

(74) Representative: **ABG Intellectual Property Law,
S.L.**

**Avenida de Burgos, 16D
Edificio Euromor
28036 Madrid (ES)**

(30) Priority: **08.09.2015 US 201562215540 P
03.11.2015 US 201562250405 P
05.02.2016 US 201662292151 P**

(56) References cited:

**WO-A2-2012/054873 WO-A2-2014/149134
US-A1- 2005 260 655**

(43) Date of publication of application:
18.07.2018 Bulletin 2018/29

- **JOACHIM WEISCHENFELDT ET AL: "Phenotypic impact of genomic structural variation: insights from and for human disease", NATURE REVIEWS GENETICS, vol. 14, no. 2, 1 February 2013 (2013-02-01), pages 125-138, XP055553462, GB ISSN: 1471-0056, DOI: 10.1038/nrg3373**

(73) Proprietor: **Cold Spring Harbor Laboratory
Cold Spring Harbor, NY 11724 (US)**

(72) Inventors:

- **WIGLER, Michael, H.
Lloyd Harbor, NY 11724 (US)**
- **LEVY, Dan
Merrick, NY 114566 (US)**
- **WANG, Zihua
East Northport, NY 11731 (US)**

EP 3 347 466 B9

Note: Within nine months of the publication of the mention of the grant of the European patent in the European Patent Bulletin, any person may give notice to the European Patent Office of opposition to that patent, in accordance with the Implementing Regulations. Notice of opposition shall not be deemed to have been filed until the opposition fee has been paid. (Art. 99(1) European Patent Convention).

- DOROTHY WARBURTON ET AL: "The contribution of de novo and rare inherited copy number changes to congenital heart disease in an unselected sample of children with conotruncal defects or hypoplastic left heart disease", HUMAN GENETICS, vol. 133, no. 1, 25 August 2013 (2013-08-25), pages 11-27, XP055553459, BERLIN, DE ISSN: 0340-6717, DOI: 10.1007/s00439-013-1353-9
 - DHEERAJ MALHOTRA ET AL: "CNVs: Harbingers of a Rare Variant Revolution in Psychiatric Genetics", CELL, vol. 148, no. 6, 16 March 2012 (2012-03-16), pages 1223-1241, XP028471679, ISSN: 0092-8674, DOI: 10.1016/J.CELL.2012.02.039 [retrieved on 2012-03-01]
 - ANSORGE ET AL: "Next-generation DNA sequencing techniques", NEW BIOTECHNOLOGY, ELSEVIER BV, NL, vol. 25, no. 4, 1 April 2009 (2009-04-01), pages 195-203, XP026059210, ISSN: 1871-6784, DOI: 10.1016/J.NBT.2008.12.009 [retrieved on 2009-02-03]
 - ZIHUA WANG ET AL: "SMASH, a fragmentation and sequencing method for genomic copy number analysis", GENOME RESEARCH, vol. 26, no. 6, 14 April 2016 (2016-04-14), pages 844-851, XP055552932, US ISSN: 1088-9051, DOI: 10.1101/gr.201491.115
 - PRODROMOU ET AL: "DNA fragmentation-based combinatorial approaches to soluble protein expression", DRUG DISCOVERY TODAY, ELSEVIER, AMSTERDAM, NL, vol. 12, no. 21-22, 23 October 2007 (2007-10-23), pages 931-938, XP022338222, ISSN: 1359-6446, DOI: 10.1016/J.DRUDIS.2007.08.012
- (52) Cooperative Patent Classification (CPC): (Cont.)
- C-Sets
C12N 15/1093, C12Q 2525/191, C12Q 2535/122,
C12Q 2563/185

Description

[0001] Throughout this application, various publications are referenced, including referenced in parenthesis. Full citations for publications referenced in parenthesis may be found listed at the end of the specification immediately preceding the claims.

Background of Invention

[0002] Analysis of copy number variants (CNVs) on a genomic scale is useful for assessing cancer progression and identifying congenital genetic abnormalities. CNVs are typically identified by microarray hybridization, but can also be detected by next-generation sequencing (NGS) (Alkan et al., 2009; Sudmant et al., 2010). This is generally done using algorithms that measure the number of sequence reads mapping to specific regions. Consequently, the resolution of sequence-based copy number methods depends largely on the number of independent mappings.

[0003] The current trend in next generation sequencing technologies is to increase the number of bases read per unit cost. This is accomplished by increasing the total number of sequence reads per lane of a flow cell, as well as increasing the number of bases within each read. Because the accuracy of copy number determination methods is driven by the quantity of independent reads, increased length of sequence reads does not improve the resolution of copy number analysis. Most of the genome is mapped well by short reads, on the order of 25-30 base pairs (bp). At the moment, high throughput sequencers are generating read lengths of ~150 bp, well in excess of what would suffice for unique mapping.

[0004] WO 2014/149134 A2 discloses a method for detecting copy number variations and association thereof with different diseases that includes sequencing.

[0005] Several prior art documents have disclosed genomic copy number information and its influence/impact on various diseases (Weischenfeldt et al., 2013; Warburton et al., 2013; Malhotra et al., 2012). Other documents disclosing nucleic sequence information are WO 2012/054873 A2, Ansorge et al., 2009 and Wang et al., 2016.

[0006] US 2005/260655 A1 discloses libraries of evolved proteins with different fragments ligated to each other which are between 50 and 2000 nucleotides.

Summary of the Invention

[0007] To take advantage of increasing read lengths, SMASH (Short Multiply Aggregated Sequence Homologies) was developed as a technique optimized for packing multiple independent mappings into every read. This is accomplished by breaking genomic DNA into small but still mappable segments, with a mean length of ~40 bp. These small segments are combined into chimeric fragments of DNA of lengths suitable for creating NGS libraries (300-700 bp).

[0008] The chimeric sequence reads generated by SMASH are processed using a time-efficient, memory-intensive mapping algorithm that performs a conservative partition of the long fragment read into constituent segment maps. The segment maps are utilized in the same manner as read maps in downstream copy number analysis. For 150-bp paired-end reads, the most cost-efficient sequencing platform so far, whole genome sequencing (WGS) averages less than one map per read pair, whereas SMASH averages >4. The quality of SMASH maps, i.e. the nonuniformities introduced by the sample preparation, sequencer and mapping bias, are of the same order as those seen with WGS mapping. Using correction and testing protocols most favorable to WGS data, map-for-map SMASH was shown to generate nearly equivalent quality copy number data as WGS at a fraction of the cost.

[0009] The invention is set out in the appended set of claims.

Brief Description of the Drawings

[0010]

Figure 1. Schematic of the SMASH method and size analysis.

A) Three representative genomic DNA molecules, shown in black, white and checkered boxes, originate from different chromosomes or distant regions of the same chromosome. B) By sonication and restriction enzyme cleavage, these molecules are fragmented into short double-stranded DNA segments with average length of 40-50 bp, as shown in the bioanalyzer result at right. C) These short DNA segments are then partially end-repaired and combined into longer fragments of DNA with lengths ranging from 50 bp to 7 kb. Hence, each resulting chimeric DNA fragment contains short DNA segments from different locations (shown by the varying box styles described above). D) These DNA fragments are ligated to sequencing adaptors containing sample barcodes, shown in dotted and vertically striped boxes, with the "barcode" box designating the sample barcodes. E) Size selection is carried out to enrich for DNA fragments in the size range of 250-700 bp, which is confirmed in the bioanalyzer. F) After final PCR, libraries are ready for sequencing.

Figure 2. SMASH informatics pipeline.

Panel A shows the decomposition of a read pair into a set of maximal uniquely mappable segments. In contrast to the map indicated by the arrow, the other maps satisfy the "20,4" rule (see text) and are considered countable maps. Panel B shows a stretch of chromosome 5 with bin boundaries selected so that each bin has the same number of exact matches from all 50-mers from the reference genome. Excluding duplicate reads, the number of "20,4" mappable segments present in each bin is counted in panel C. LOESS normalization is used to adjust bin counts for sample-specific GC bias (panel D). Lastly, in panel E, the data is segmented using circular binary segmentation (CBS) of the GC normalized data.

Figure 3. SMASH and WGS copy number profiles for an SSC quad.

Panel A shows the whole genome view (autosome and X chromosomes) for the four members of a family. The dots show the reference and GC normalized ratio values for WGS and SMASH. Similarly, the overlapping lines show the copy number segmentation by CBS (circular binary segmentation) for both WGS and SMASH. The black box highlights a deletion on chromosome 5 that is expanded in panel B. The deletion, identified by both methods, occurs in the father and is transmitted to the sibling in the family. Panel C illustrates the bin for bin comparison of the normalized ratio values of the father from WGS and SMASH. The dark and light points show increasingly sparse subsamples of the data points.

Figure 4. SMASH and WGS copy number profiles for SKBR3.

The SKBR3 breast cancer cell line has a complex copy number pattern. Panel A shows the whole genome view with copy number on a log scale. The dots show the GC-normalized ratio values for WGS and SMASH, while the overlapping lines show the copy number segmentation for both WGS and SMASH. Panel B expands on chromosome 14 on a linear scale. There is strong agreement between WGS and SMASH in the integer copy number segmentations and dispersion about the segment mean. Panel C illustrates the bin for bin comparison of the normalized ratio values from WGS and SMASH. The dark and light points show increasingly sparse subsamples of the data points to illustrate density.

Figure 5. Bioanalyzer results of SMASH protocols on independent samples.

Following Figure 1, right panel, we show bioanalyzer results of SMASH protocols on independent samples. Lower (35 bp) and upper markers (10.38kb) are indicated by arrows. In each panel, two of the ten profiles (in blue and dark green) show results for bad quality DNA samples. The remaining curves are of good quality. (A) Size distribution of DNA molecules after DNA fragmentation. Blue and dark green curves show a wider length range and longer average length of DNA segments than the remaining samples. (B) After random ligation of DNA segments, curves from good samples show a wide length range of DNA concatemers. (C) For the final DNA library, curves from good samples show the length range from 250bp-700bp, ideal for sequencing. The failed libraries show mainly sequencing adaptor dimers, highlighted with a star.

Figure 6. Schematic of alternative SMASH method (left panel) and bioanalyzer results (right panel).

In bioanalyzer results, x-axis represents the length of DNA segments. (A) Three genomic DNA molecules, shown in black, white and checkered boxes, are from different chromosomes or different locations of the same chromosome. (B) By dsDNA fragmentase cutting, these DNA molecules are fragmented into short double-stranded segments with average length around 35bp, as shown in bioanalyzer result on right panel. (C) Then these short DNA segments are partially end-repaired and randomly concatenated into longer fragments of DNA with length range from 50bp to 7kb. Hence, each DNA fragment contains several short DNA segments that are from different locations/chromosomes shown in different box styles as described above. (D) These DNA fragments are ligated with sequencing adaptors containing sample barcodes, shown in dotted and vertically striped boxes linked with an open box labeled "barcode". (E) Size selection is carried out to make DNA fragments in the proper size range from 250bp to 700bp, which is confirmed in the bioanalyzer result of the final DNA library. (F) After final PCR by sequencing adaptors, libraries are ready for sequencing.

Figure 7. SMASH2 compared to WGS and SMASH on SKBR3.

Similar to Figure 4, panels A and B, the agreement of the newer SMASH protocol (SMASH2) with both WGS and the previous SMASH protocol is shown. There is excellent agreement between the three methods.

Detailed Description of the Invention

[0011] SMASH reduces genomic DNA to small but still uniquely mappable segments, and randomly ligates them into chimeric stretches of DNA of lengths suitable for creating next-generation sequencing (NGS) libraries (400-500 bp).

Sequencing of these libraries results in a paradigm in which CNVs can be detected through template analysis (Levy and Wigler, 2014). The crux of its significance lies in its efficiency: SMASH can be run on average NGS instruments and yield ~6 times or more as many maps as 'standard' whole genome sequencing (WGS). On a machine that generates 300 million 150-bp paired-end reads, SMASH can obtain 60 million maps per sample at a resolution of ~10 kb.

[0012] Specifically, genomic DNA is cleaved ('smashed') into small but mappable segments by sonication and/or enzymatic activity, with a mean length of ~40 bp, then ligated into longer chimeric fragments of DNA. A second fragmentation step eliminates long (>1 kb) chimeric molecules, and fragments suitable for creating NGS libraries are purified (e.g. 400-500 bp). Barcoded sequencing adaptors are added to create libraries that can be multiplexed on a single sequencing lane, significantly reducing cost/patient. To obtain mapping information from the chimeric reads, we apply an algorithm and a set of heuristics. Suffix arrays adapted from sparseMEM (Khan et al., 2009) are used to determine 'maximal almost-unique matches' (MAMs) between a NGS read and the reference genome. The mappings within a read pair provide a unique signature for each read, allowing identification and removal of PCR duplicates. CNV detection is based on map-counting methods, employing bins of expected uniform density (Navin et al., 2011). For each sample, we count the number of maps within each bin, then adjust bin counts for GC bias by LOESS normalization. Template analysis (Levy and Wigler, 2014) is utilized to overcome distinct patterns of systematic noise that extend beyond the gross-scale corrections of GC adjustment, which is inherent in both WGS and SMASH reads. The result of these measurements is an ability to detect CNV on par with WGS.

[0013] The present invention provides a sequencing library composition comprising a first mixture of different chimeric genomic nucleic acid fragments, wherein the mixture of different chimeric genomic nucleic acid fragments contains at least 100,000 different fragments, wherein the chimeric genomic nucleic acid fragments are 250 ($\pm 10\%$) to less than 1000 ($\pm 10\%$) base pairs in length, wherein each different fragment in the mixture comprises randomly ligated DNA segments, wherein each DNA segment in the fragment is a nucleic acid molecule at least 27 base pairs in length resulting from random fragmentation of a single genome for which a reference genome is available, wherein at least 50% of the segments in the at least 100,000 different fragments are 30 to 50 base pairs in length ($\pm 10\%$), further comprising sequence adaptors ligated to the termini of the chimeric genomic nucleic acid fragments, wherein the sequence adaptors comprise a barcode identifying the sample origin of each fragment.

[0014] In some embodiments, the segments are ligated directly to each other to form a fragment.

[0015] In some embodiments, the DNA segments are 30 to 50 base pairs in length ($\pm 10\%$).

[0016] In some embodiments, the mixture of different chimeric genomic nucleic acid fragments is enriched for chimeric genomic nucleic acid fragments 250($\pm 10\%$) to 700 ($\pm 10\%$) base pairs in length, preferably 400-500 base pairs.

[0017] In some embodiments, at least 50% of the chimeric genomic nucleic acid fragments in the mixture are 250 ($\pm 10\%$) to 700 ($\pm 10\%$) base pairs in length, preferably 400-500 base pairs.

[0018] In some embodiments, the mixture of different chimeric genomic nucleic acid fragments contains fragments composed of an odd number of segments.

[0019] In some embodiments, the mixture of chimeric genomic nucleic acid fragments contains ligated segments whose two ligation points form a sequence other than a restriction enzyme recognition site.

[0020] In some embodiments, a sequence adaptor ligated to the termini of the chimeric genomic nucleic acid fragments comprises a barcode identifying the genomic source of each fragment.

[0021] In some embodiments, a sequence adaptor ligated to the termini of the chimeric genomic nucleic acid fragments comprises a primer binding site for amplification.

[0022] In some embodiments, the mixture of different chimeric genomic nucleic acid fragments is enriched for sequence adaptor-ligated chimeric genomic nucleic acid fragments 250 ($\pm 10\%$) to 700 ($\pm 10\%$) base pairs in length, preferably 400-500 base pairs.

[0023] In some embodiments, the sequencing library composition comprises amplified sequence adaptor-ligated chimeric genomic nucleic acid fragments. Such amplification may be accomplished by methods such as PCR. Primer binding for accomplishing this amplification step may be located on the ligated sequencing adaptor.

[0024] In some embodiments, the sequencing library composition further comprises a second mixture of different chimeric genomic nucleic acid fragments, wherein the second mixture of fragments is obtained from a different genome than the first mixture.

[0025] In some embodiments, the sequencing library composition further comprises a collection of multiple mixtures of different chimeric genomic nucleic acid fragments, wherein each mixture of fragments in the collection is obtained from a different genome than any other mixture in the collection.

[0026] In some embodiments, each mixture of chimeric genomic nucleic acid fragments contains fragments having a sequencing adaptor containing a unique barcode ligated onto only fragments within the mixture, such that the collection of mixtures can be multiplexed.

[0027] In some embodiments, the genomic nucleic acids are extracted from a cell, a tissue, a tumor, a cell line or from blood.

[0028] In another aspect the invention relates to a method for obtaining the composition according to the invention,

comprising

- 5 i) randomly fractionating the single genome to obtain random segments from the genome, preferably size selecting a subpopulation of segments 30 to 50 base pairs in length ($\pm 10\%$) prior to ligation, and/or wherein the subpopulation of segments is selected using bead purification;
- ii) subjecting the segments from step (i) to ligation to generate different chimeric genomic nucleic acid fragments, and
- iii) ligating sequencing adaptors to the chimeric genomic nucleic acid fragments,

thereby obtaining the mixture of different genomic nucleic acid fragments from the single genome.

10 **[0029]** In some embodiments, in step (i) the genomic nucleic acids are mechanically sheared to obtain the randomly fragmented DNA segments.

[0030] In some embodiments, the mechanical shearing is by sonication.

[0031] In some embodiments, the method further comprises subjecting the segments of genomic nucleic acids to enzymatic digestion.

15 **[0032]** In some embodiments, the enzymatic digestion of the segments of genomic nucleic acids is by the restriction enzymes CviKI-1 and NlaIII.

[0033] In some embodiments, in step (i) genomic nucleic acids are enzymatically fragmented, by

20 a) generating random DNA nicks in the genome; and

b) cutting the DNA strand opposite the nick,

thereby producing dsDNA breaks in the genomic nucleic acids resulting in DNA segments.

[0034] In some embodiments, the resulting DNA segments are end-repaired directly after genomic fragmentation.

25 **[0035]** In some embodiments, chimeric genomic nucleic acid fragments are end-repaired after their formation by random segment ligation.

[0036] In some embodiments, further comprising reducing the size of the chimeric genomic nucleic acid fragments.

[0037] In some embodiments, the method further comprises selecting for fragments about 250 to about 700 base pairs in length.

30 **[0038]** In some embodiments, the method further comprises purifying the chimeric genomic nucleic acid fragments, optionally by bead purification.

[0039] In some embodiments, the method further comprises adenylating the 3' termini of the chimeric genomic nucleic acid fragments prior to step iii).

[0040] In some embodiments, the method further comprises purifying the sequence adaptor-ligated genomic nucleic acid fragments, optionally by purification.

35 **[0041]** In some embodiments, the method further comprises selecting for sequence adaptor-ligated genomic nucleic acid fragments about 250 to about 700 base pairs in length.

[0042] In some embodiments, the method further comprises amplifying the size-selected sequence adaptor-ligated genomic nucleic acid fragments.

40 **[0043]** In some embodiments, the method further comprises ligating a unique adaptor barcode to a mixture of chimeric genomic nucleic acid fragments from the same genome, such that multiplex sequencing can be performed upon pooling of multiple mixtures from different genomes.

[0044] In some embodiments, the initial amount of genomic nucleic acids is 200 ng, 500ng, or 1 μ g ($\pm 10\%$).

[0045] In some embodiments, the genomic nucleic acids are extracted from a cell, a tissue, a tumor, a cell line or from blood.

45 **[0046]** In some embodiments, sequences are obtained from a mixture of chimeric genomic nucleic acid fragments using a next-generation sequencing platform.

[0047] In another aspect, the invention relates to a process of obtaining the nucleic acid sequence of the different chimeric genomic nucleic acid fragments of the composition described above, or produced by the method described above, comprising (i) obtaining the fragments, and (ii) sequencing the fragments, so as to obtain the nucleic acid sequence of the different chimeric genomic nucleic acid fragments.

50 **[0048]** In another aspect, the invention relates to a process for obtaining genomic copy number information from a genome, comprising

- 55 i) obtaining the nucleic acid sequence of the different chimeric genomic nucleic acid fragments of (a) a sequencing library composition comprising a first mixture of different chimeric genomic nucleic acid fragments, wherein the mixture of different chimeric genomic nucleic acid fragments contains at least 100,000 different fragments, wherein the chimeric genomic nucleic acid fragments are 250 ($\pm 10\%$) to less than 1000 ($\pm 10\%$) base pairs in length, wherein each different fragment in the mixture comprises randomly ligated DNA segments, wherein each DNA segment in

the fragment is a nucleic acid molecule at least 27 base pairs in length resulting from random fragmentation of a single genome for which a reference genome is available, wherein at least 50% of the segments in the at least 100,000 different fragments are 30 to 50 base pairs in length ($\pm 10\%$), (b) the composition of the invention described above, or (c) the composition produced by the method of the invention described above;

ii) identifying and mapping to a genome each Maximal Almost-unique Match (MAM) within a sequenced chimeric genomic nucleic acid fragment; and

iii) counting the number of mapped MAMs within a binned genome, thereby obtaining genomic copy number information.

[0049] In some embodiments, in step (ii) MAMs are identified using a longMEM software package.

[0050] In some embodiments, step (ii) further comprises filtering MAMs by discarding MAMs less than twenty base pairs and not at least four base pairs longer than required for uniqueness.

[0051] In some embodiments, step (ii) further comprises filtering MAMs by discarding MAMs in a read-pair map that are within 10,000 base pairs of one another.

[0052] In some embodiments, in step (iii) the number of mapped reads are counted in genome bin sizes that yield uniform map counts for the reference sample.

[0053] In some embodiments, in step (iii) the number of mapped reads are counted in empirically determined genome bins of uniform observation of a reference.

[0054] In some embodiments, in step (iii) the number of mapped reads are counted in genome bins of expected uniform density.

[0055] In some embodiments, in step (iii) the number of mapped reads in each bin is adjusted for GC bias by LOESS normalization.

[0056] In some embodiments, in step (iii) template analysis is utilized to reduce systematic noise in GC adjusted bin count data.

[0057] In some embodiments, in step (iii) a reference normalization is applied to bin count data by dividing GC-adjusted bin ratios by a standard sample bin ratio.

[0058] In some embodiments, in step (iii), reference normalized GC-adjusted bin count data is analyzed by circular binary segmentation.

[0059] In some embodiments, in step (iii) the total number of reference maps is matched to the total number of sample maps.

[0060] In another aspect, the invention relates to a method of diagnosing, predicting likelihood of displaying or determining the probability of inheriting a prenatal disorder, a pediatric disorder, a developmental disorder, a psychological disorder, an autoimmune disorder, cancer, congenital heart disease, schizophrenia, Autism Spectrum Disorders or a patient's response to a therapy, comprising obtaining the patient's genomic copy number information by the process for obtaining genomic copy number information from a genome of the invention.

Terms

[0061] Unless otherwise defined, all technical and scientific terms used herein have the same meaning as commonly understood by a person of ordinary skill in the art to which this invention belongs.

[0062] As used herein, and unless stated otherwise or required otherwise by context, each of the following terms shall have the definition set forth below.

[0063] As used herein, "about" in the context of a numerical value or range means $\pm 10\%$ of the numerical value or range recited or claimed, unless the context requires a more limited range.

[0064] The terms "nucleic acid molecule" and "sequence" are not used interchangeably herein. A "sequence" refers to the sequence information of a "nucleic acid molecule".

[0065] The terms "template", "nucleic acid", and "nucleic acid molecule", are used interchangeably herein, and each refers to a polymer of deoxyribonucleotides and/or ribonucleotides. "Nucleic acid" shall mean any nucleic acid, including, without limitation, DNA, RNA and hybrids thereof. The nucleic acid bases that form nucleic acid molecules can be the bases A, C, G, T and U, as well as derivatives thereof. "Genomic nucleic acid" refers to DNA derived from a genome, which can be extracted from, for example, a cell, a tissue, a tumor or blood.

[0066] As used herein, the term "chimeric" refers to being comprised of nucleic acid molecules taken from random loci within a genome that are reconnected in a random order. In SMASH, a fragment is considered to be chimeric because it is a composed of randomly ligated segments of a genome.

[0067] As used herein, the term "fragmentation" refers to the breaking up of large nucleic acids e.g. genomic DNA into smaller stretches of nucleotides. Fragmentation can be accomplished by multiple methods including but not limited to, sonication and enzymatic activity.

[0068] As used herein "contig" and "contiguous" refers to a set of overlapping sequence or sequence reads.

[0069] As used herein, the term "amplifying" refers to the process of synthesizing nucleic acid molecules that are complementary to one or both strands of a template nucleic acid. Amplifying a nucleic acid molecule typically includes denaturing the template nucleic acid, annealing primers to the template nucleic acid at a temperature that is below the melting temperatures of the primers, and enzymatically elongating from the primers to generate an amplification product.

5 The denaturing, annealing and elongating steps each can be performed once. Generally, however, the denaturing, annealing and elongating steps are performed multiple times (e.g., polymerase chain reaction (PCR)) such that the amount of amplification product is increasing, often times exponentially, although exponential amplification is not required by the present methods. Amplification typically requires the presence of deoxyribonucleoside triphosphates, a DNA polymerase enzyme and an appropriate buffer and/or co-factors for optimal activity of the polymerase enzyme. The term

10 "amplified nucleic acid molecule" refers to the nucleic acid molecules, which are produced from the amplifying process. **[0070]** As used herein, the term "mapping" refers to identifying a unique location on a genome or cDNA library that has a sequence which is substantially identical to or substantially fully complementary to the query sequence. A nucleic acid molecule containing a sequence that is capable of being mapped is considered "mappable." The nucleic acid molecule may be, but is not limited to the following: a segment of genomic material, a cDNA, a mRNA, or a segment of

15 a cDNA. **[0071]** As used herein, the term "read" or "sequence read" refers to the nucleotide or base sequence information of a nucleic acid that has been generated by any sequencing method. A read therefore corresponds to the sequence information obtained from one strand of a nucleic acid fragment. For example, a DNA fragment where sequence has been generated from one strand in a single reaction will result in a single read. However, multiple reads for the same

20 DNA strand can be generated where multiple copies of that DNA fragment exist in a sequencing project or where the strand has been sequenced multiple times. A read therefore corresponds to the purine or pyrimidine base calls or sequence determinations of a particular sequencing reaction. **[0072]** As used herein, the terms "sequencing", "obtaining a sequence" or "obtaining sequences" refer to nucleotide sequence information that is sufficient to identify or characterize the nucleic acid molecule, and could be the full length

25 or only partial sequence information for the nucleic acid molecule. **[0073]** As used herein, the term "reference genome" refers to a genome of the same species as that being analyzed for which genome the sequence information is known.

[0074] As used herein, the term "region of the genome" refers to a continuous genomic sequence comprising multiple discrete locations.

30 **[0075]** As used herein, the term "sample tag" refers to a nucleic acid having a sequence no greater than 1000 nucleotides and no less than two that may be covalently attached to each member of a plurality of tagged nucleic acid molecules or tagged reagent molecules. A "sample tag" may comprise part of a "tag."

[0076] As used herein, the term "segment" of genomic material refers to the mappable nucleic acid molecules resulting from random fragmentation of genomic DNA. A segment in a SMASH fragment are about 30 to 50 base pairs in length, and may for example have a length of 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47,

35 48, 49 or 50 base pairs. **[0077]** As used herein, the term "fragment" refers to a chimeric DNA molecule resulting from the ligation of multiple DNA segments. Thus, as used herein, a "fragment" contains at least one and usually more than one "segment," preferably 2, 3, 4, 5, 6, 7, 8, 9 or 10 segments. Although methods described herein provide segments of highly uniform length, a fragment may contain segments having lengths outside of the preferred size range of 30 to 50 base pairs.

40 **[0078]** As used herein the term "sequencing library" refers to a mixture of DNA fragments comprising the total genomic DNA from a single organism for use in sequencing. Next-generation sequencing libraries are generally size-selected and ligated to sequencing adaptors prior to sequencing. Steps in next-generation sequencing library preparation may include fragmentation, end-repairing, adenylation, sequencing adaptor ligation and PCR enrichment. A number of purification and size-selection steps may also be performed throughout the next-generation sequencing library preparation. Specifically, a "SMASH library" refers to a type of sequencing library which is composed of a mixture of fragments of genomic DNA from a single organism, wherein the fragments are chimeric nucleic acid molecules made up of smaller, yet mappable, randomly ligated segments of the genomic DNA.

45 **[0079]** As used herein the term "ligation" refers to the enzymatic joining of two nucleic acid molecules. Specifically, SMASH fragments are composed of randomly ligated DNA segments. Random ligation in this instance implies that any segment has an equal probability of being directly ligated to any other segment.

50 **[0080]** As used herein, the term "sequencing adaptor" refers to oligos bound to the 5' and 3' end of each DNA fragment in a sequencing library. Adaptors contain platform-dependent sequences that allow amplification of the fragment as well as sequences for priming the sequencing reaction. Adaptors also contain unique sequences, known as barcodes or indexes, which are used to identify the sample origin of each fragment. The adaptor may contain regions which are used as primer binding sites for other enzymatic reactions, such as amplification by PCR.

55 **[0081]** As used herein, the term "barcode", also known as an "index," refers to a unique DNA sequence within a sequencing adaptor used to identify the sample of origin for each fragment.

[0082] As used herein, the term "multiplex" refers to assigning a barcode to each mixture of fragments from a single genomic source, pooling or otherwise mixing multiple mixtures of fragments, sequencing the entire collection of mixtures in a single sequencing run and subsequently sorting and identifying the genomic origin of each read by its barcode sequence.

[0083] As used herein, "substantially the same" sequences have at least about 80% sequence identity or complementarity, respectively, to a nucleotide sequence. Substantially the same sequences or may have at least about 95%, 96%, 97%, 98%, 99% or 100% sequence identity or complementarity, respectively.

[0084] As used herein, the term "substantially unique primers" refers to a plurality of primers, wherein each primer comprises a tag, and wherein at least 50% of the tags of the plurality of primers are unique. Preferably, the tags are at least 60%, 70%, 80%, 90%, or 100% unique tags.

[0085] As used herein, the term "substantially unique tags" refers to tags in a plurality of tags, wherein at least 50% of the tags of the plurality are unique to the plurality of tags. Preferably, substantially unique tags will be at least 60%, 70%, 80%, 90%, or 1000 unique tags.

[0086] As used herein, the term "tag" refers to a nucleic acid having a sequence no greater than 1000 nucleotides and no less than two that may be covalently attached to a nucleic acid molecule or reagent molecule. A tag may comprise a part of an adaptor or a primer.

[0087] As used herein, a "tagged nucleic acid molecule" refers to a nucleic acid molecule which is covalently attached to a "tag."

[0088] Where a range of values is provided, it is understood that each intervening value, to the tenth of the unit of the lower limit unless the context clearly dictates otherwise, between the upper and lower limit of that range, and any other stated or intervening value in that stated range, is encompassed within the invention. The upper and lower limits of these smaller ranges may independently be included in the smaller ranges, and are also encompassed within the invention, subject to any specifically excluded limit in the stated range. Where the stated range includes one or both of the limits, ranges excluding either or both of those included limits are also included in the invention.

[0089] Publications and references cited herein are not admitted to be prior art.

[0090] This invention will be better understood by reference to the Experimental Details which follow, but those skilled in the art will readily appreciate that the specific experiments detailed are only illustrative of the invention as defined in the claims which follow thereafter.

Experimental Details

[0091] Examples are provided below to facilitate a more complete understanding of the invention. The following examples illustrate the exemplary modes of making and practicing the invention. However, the scope of the invention is not limited to specific embodiments disclosed in these Examples, which are for purposes of illustration only.

Methods - DNA materials

[0092] DNA samples used in this example were from two sources. One source of the genomic DNA was extracted from SKBR3, a human breast cancer cell line. The other was extracted from blood from two families, which are from the Simons Simplex Collection (SSC) with samples and data from the mother, the father, the proband, and an unaffected sibling (Fischbach and Lord, 2010).

Methods - SMASH protocol

[0093] The amount of genomic DNA required for SMASH is flexible. Three different genomic DNA inputs - 200 ng, 500 ng and 1 μ g - were tested and successfully constructed high quality libraries for all three conditions. In this example, 1 μ g of DNA was used as starting material from all the samples. DNA was diluted in 1X Tris buffer (10 mM Tris-Cl, pH 8.5) to a final volume of 75 μ l, and transferred to microtubes (Covaris). The Covaris E210 AFA instrument (Covaris) was used to shear the genomic DNA into segments with average length of 100 bp according to the manufacturer's manual. DNA segments were further cut by Cvikl-1 (NEB) and NlaIII (NEB) in 1X CutSmart buffer in a final volume of 90 μ l, which was incubated at 37°C for 1 hr. After enzyme digestion, the volume of solution was reduced to about 30 μ l by Savant SpeedVac (Thermo Scientific). DNA segments longer than 100 bp were removed as follows: adding 2.5X volume of AMPure XP beads (Beckman Coulter), mixing well, incubating at room temperature (RT) for 5 min, and collecting supernatant. The supernatant was purified by QIAquick nucleotide removal kit (Qiagen) following manufacturer's instructions. DNA segments were eluted in 30 μ l H₂O. The average length of DNA segments was 40-50 bp as determined by the Bioanalyzer 2100 (Agilent Technologies). These DNA segments were end-repaired by T4 DNA polymerase (NEB), DNA polymerase I (large Klenow fragment, NEB) and T4 Polynucleotide Kinase (NEB) at RT for 30 min. The polished DNA segments were purified by QIAquick nucleotide removal kit (Qiagen) with 30 μ l H₂O elution. The short DNA segments

5 were randomly ligated to form longer fragments of chimeric DNA with the quick ligation kit (NEB) at RT for 15 min. The long DNA chimeric fragments were purified using 1.6X AMPure XP beads, and end-repaired as earlier. A single 'A' nucleotide was added to the 3' ends of the polished DNA fragments by Klenow fragment (3'->5' exo, NEB) at 37°C for 30 min. After purification by 1.6x AMPure XP beads, barcoded sequencing adapters [Iossifov et al. 2012, Neuron] were ligated to the DNA fragments by quick ligation. This allowed for multiplex samples on sequencing lanes. DNA fragments were again purified by 1.6X AMPure XP beads, and eluted in 50 μ l H₂O. This size selection step was carried out to enrich for DNA fragments within the ideal Illumina sequencing length range of 300-700 bp. First, 0.6x (30 μ l) AMPure XP beads was added into 50 μ l of purified DNA. After incubation at RT for 5 min, supernatant was collected. 8 μ l (0.16X the original 50 μ l) of AMPure XP beads was added, and mixed well with the supernatant. This mixture was incubated at RT for 5 min. After 2 washes with 180 μ l of 80% ethanol, DNA fragments were eluted in 30 μ l H₂O. The final 8 cycles of PCR amplification were carried out on this DNA using Illumina sequencing adapters in 1X Phusion® High-Fidelity PCR Master Mix with HF Buffer (NEB). DNA libraries were quantitated on the Bioanalyzer and diluted to a concentration of 10 nM. Sequencing was performed on the HiSeq 2000 (paired-end 100 bp, Illumina) for libraries prepared from SSC families and the NextSeq 500 (paired-end 150 bp, Illumina) for libraries prepared from the SKBR3 cell line.

15 Methods - Determining maps

20 **[0094]** WGS and SMASH data were mapped to the GATK b37 genome. For WGS, read 1 was clipped to 76 bp, mapped using Bowtie1, and duplicates were then filtered using Samtools. For SMASH (after the mapping procedure described below), the multiple-MAM signature of each read pair was used to filter duplicates. For both methods, only unique mappings to chromosomes 1-22, X and Y only were bin-counted.

25 **[0095]** To prepare for mapping SMASH data, the sparseMEM package (Khan et al., 2009) was modified to increase the maximum genome size from 2.147×10^9 bases to an essentially unlimited value, and the sparse functionality was removed to increase program speed and decrease complexity. Features were added to 1) save the various suffix array index structures to disk; 2) to read them in for subsequent runs using memory-mapping; 3) to distribute reads to the parallel query threads to avoid multiple parsing of the input; and 4) to read several query files in parallel. Options were also added to read input data from FASTQ and SAM files, to output mappings and non-mapping reads in SAM and custom binary formats, and to simultaneously map to the genome and its reverse complement to avoid a Maximal Exact Match (MEM) pruning step. The resulting software package is called longMEM for its ability to handle longer genomes.

30 **[0096]** Using longMEM, we searched for Maximal Almost-unique Matches (MAMs), which are maximally extended subsequences in query reads that match uniquely within the reference and its reverse complement, but may be repeated in the query. For query reads of length Q and a reference of length R, we find all MAMs in the query in $O(Q*(Q + \log(R)))$ time using the reference, the suffix array, its inverse and an LCP (Longest Common Prefix) table.

35 **[0097]** Most segments composing SMASH reads result in MAMs that are suitable for copy number analysis. The exceptions are segments that are not present in the reference due to blocking read errors or mutation, and those that are too short to be uniquely mapped to their origin. In addition to acceptable MAMs, junctions between adjacent segments in SMASH sometimes result in one or more MEMs being found. If unique in the reference, these are reported as spurious MAMs.

40 **[0098]** MAMs were filtered by discarding MAMs less than 20 bp and not at least 4 bases longer than required for uniqueness. Assuming a random genome and ignoring the usage of restriction enzymes, this naively reduced spurious MAM contamination by a factor of 4^4 . Because the mode for minimum mappable length in the genome is 18 bp, the average is 29 bp and segments are typically 40 bp in length, it is believed that the filter did not greatly reduce the number of reported legitimate MAMs. An additional filter turns our MAMs into MUMs by ensuring that no retained MAMs in a read pair map within 10,000 bp of another, which avoids double-counting of segments containing indels or SNPs as well as MAMs read from both ends in short chimeric fragments.

45 Methods - Binning, Normalization, and Copy Number

50 **[0099]** Chromosomes 1-22, the X and the Y were divided into 50,000, 100,000 and 500,000 WGS-optimized bins by mapping every 50-mer in the reference with Bowtie1 and adjusting bin boundaries so that each bin had the same number of uniquely mapped reads assigned to it (± 1).

55 **[0100]** An equal number of mappings were assigned from SSC WGS and SMASH data to bins and added one count to each total. Counts were normalized to set the mean of all autosome bins to 1, then LOESS was performed on the normalized autosome to correct for GC site density. After bin-wise summation across samples, bad bins were selected based on upward copy number deviation from the chromosome median exceeding a MAD-based limit using a Bonferroni-corrected p value of 0.05.

[0101] SSC and SKBR3 mappings were sampled at 20, 50, 100 and up to 1000 (if available) mappings per bin and assigned them to bins, in this instance excluding bins marked as bad. Sample counts were divided at low maps per bin

on a bin-wise basis by a non-related male reference sample, using the highest maps per bin. The ratio data was normalized and GC-corrected, then segmented using CBS with the minimum segment length and alpha parameters set to 3 and 0.02, respectively. Segmented profiles were adjusted by varying the overall scale and offset within expected bounds to find the best quantal fit.

5

Methods - WGS and SMASH quantification and comparison

[0102] SSC sample signal to noise was defined for SMASH and WGS as the autosome minus the X chromosome median un-quantized ratio, divided by its measured MAD-based noise for male samples using a female reference sample (when performing reference normalization). We also counted the quantized and rounded segmented autosome bin values different than 2 to place an upper bound on deviation from the SSC diploid expectation. WGS and SMASH concordance were assessed for SSC and SKBR3 data by plotting the lengths of bin runs on histograms for un-quantized segmented ratios that differed by more than 0.2.

10

15

Example 1. Overview of SMASH.

[0103] The protocol for SMASH (see also "*Methods - Smash protocol*," above) is illustrated in Figure 1. To obtain SMASH tags, first genomic DNA was mechanically sheared by sonication, then cut with two restriction endonucleases. The ideal size fraction is obtained using bead purification (see also "*Methods - Smash protocol*," above) to enrich for the target size range of 40 bp (Figure 1). To generate the long chimeric DNAs, the SMASH tags were end-repaired and then ligated. A second fragmentation step may optionally be performed to eliminate long (>1 kb) chimeric molecules, and DNA fragments in the proper size range (300-700 bp) are purified. Barcoded sequencing adaptors are then attached to the molecules, creating libraries that can be multiplexed on a single sequencing lane. Alternatively, long chimeric DNAs can be formed by ligation of end-repaired SMASH segments, followed by attachment of barcoded sequencing adaptors to the fragments and finally selection of DNA fragments in the optimal size range for sequencing (300-700 bp) by bead purification. The protocol is robust and reproducible, typically generating libraries with nearly identical distributions of segment and fragment lengths (Figure 5). While the SMASH library may contain a low amount of segments and fragments outside of the desired size range, these contaminants are inconsequential and do not affect the copy number variation determination in any way.

20

25

30

[0104] To obtain mapping information from the chimeric reads, an algorithm and a set of heuristics was applied, described briefly here (see Figure 2 and Methods for additional details). sparseMEM (Khan et al., 2009), a program that uses suffix arrays to quickly determine all maximal almost-unique matches (or MAMs) between a NGS read and the reference genome was adapted. The mappings of a read pair provide a unique signature for each SMASH read, allowing easy identification as well as removal of PCR duplicates. A heuristic was used that identifies distinct unambiguous matches (or 'maps') spanned by the read pair.

35

[0105] The parameters of the heuristic have been calibrated to maximize quality of the copy number data by balancing the number of maps per read against the quality of the map assignment.

[0106] The copy number detection protocol of the present invention is based on map-counting methods, and it requires that bin boundaries were first determined to partition the genome. 'Bins of expected uniform density' first used for single cell genome copy number determination (Navin et al., 2011), are employed. Boundaries are chosen such that each bin contains the same expected number of maps when sequencing the reference genome with exhaustive coverage and perfect reads. SMASH and WGS have different distributions of expected map densities due to variation in map lengths. Bin boundaries were chosen suitable for WGS, and map the WGS reads in single-end mode using the first 76 bp. For each sample, the number of maps that fall within each bin was counted and bin counts were adjusted for GC bias by LOESS normalization.

45

[0107] Both WGS and SMASH have distinct patterns of systematic noise that extend beyond the gross-scale corrections of GC adjustment. This is evidenced by strong correlation between independent samples. Moreover, this systematic noise is trendy, leading to high autocorrelation, and so is likely to trigger false-positive copy number events. This error was corrected by choosing one sample as a reference, then dividing all remaining sample data by that reference. The resulting copy number segmentation typically results in segment means that are low integer fractions, reflecting copy number in the sample. With sufficient samples (and using multiple reference samples), it is possible to determine absolute copy number. For analysis of bin count data, the standard method of circular binary segmentation was used (Olshen et al., 2004).

50

55

Example 2. Optimizing pipeline parameters.

[0108] To measure performance precisely and choose parameters for pipeline processing, the signal in bins was compared on the X chromosome to those on autosomes in male subjects. Also calculated are 1) the median average

EP 3 347 466 B9

deviation (MAD) of bins to measure the magnitude of the noise, and 2) the autocorrelation as a measure of trendiness in the data, an important risk factor for segmentation error. Signal to noise ("S/N") was calculated as the difference in the medians of the autosome and X-chromosome, divided by the square root of the sum of the squares of the MADs. These statistics were used to evaluate reference normalization and mapping algorithms, and then to compare WGS to SMASH (Table 1).

[0109] First, the utility of applying reference normalization ("ref norm," Table 1) was considered. Dividing the GC-adjusted bin ratios by a standard sample bin ratio greatly improved performance for both WGS and SMASH (rows 1 through 4). Namely, reference normalization decreases "autocorrelation" up to ten fold while increasing "signal to noise".

5
10
15
20
25
30
35
40
45
50
55

Table 1

rule	type	ref norm	number of bins	maps per bin	auto correlation	autosome median	x chrom median	autosome MAD	x chrom MAD	signal to noise
-	wgs	yes	100000	50	0.012	2.008	1.032	0.194	0.138	4.102
-	wgs	no	100000	50	0.075	2.012	1.040	0.202	0.139	3.959
20,4	smash	yes	100000	50	0.011	2.010	1.071	0.196	0.146	3.833
20,4	smash	no	100000	50	0.109	2.015	1.055	0.212	0.148	3.718
20,0	smash	yes	100000	117.28	0.010	2.010	1.419	0.137	0.129	3.148
20,4	smash	yes	100000	63.98	0.012	2.006	1.062	0.176	0.129	4.333
20,8	smash	yes	100000	53.09	0.013	2.008	1.034	0.192	0.140	4.094

Table 1. Reference normalization and mapping rules.

[0110] In Table 1 auto-correlation, medians and median absolute deviation (MADs) for the autosome and X chromosomes in males, and the resultant signal-to-noise, is computed. The first four entries compare WGS and SMASH for the same bin resolution (100,000) and the same average number of maps per bin (50). Results with and without normalizing by a reference sample are shown. SMASH and WGS have similar performance and both methods reduce autocorrelation by reference normalization while maintaining signal-to-noise. The lower three entries compare SMASH performance using different rules for selecting valid maps (see text). Each SMASH instance operates on the same number of reads with the most lax rule (20,0) generating 117 maps per bin and the strictest rule (20,8) generating 53 maps per bin. The best signal-to-noise is obtained with the 20,4 rule.

[0111] Next we established a two-part, two parameter (L,K) rule for accepting the map of a substring from a SMASH read to the reference genome (see Figure 2, panel A). First, all substrings in a read were found that occur just once in the reference genome and such that the match cannot be extended. These are called "MAMs," for maximal almost-unique matches (see also *Methods - Determining maps*). A minimum match length, L, as the first parameter is required. For the data shown here, L is 20 bp. To avoid false maps that arise by chimerism, a second rule is required, namely a MAM of length M contains a substring of length M-K that maps uniquely to the genome. Many combinations of L and K were examined, and their performance was measured on an identical set of SMASH reads, with fixed bin boundaries. Only the results for rules 20:0, 20:4 and 20:8 (Table 1 rows 5-7) are shown. Despite having far fewer maps ("maps per bin"), the 20:4 rule is superior to the 20:0 rule as judged by "signal to noise". Many of the 20:0 maps must be false. This false mapping can be attributed to chimerism at fragment boundaries. On the other hand, the 20:4 rule is superior to the 20:8 rule as judged by a slightly degraded "signal to noise" that can be attributed to increased sampling error due to reduced coverage. Therefore, the 20:4 rule is employed throughout.

Example 3. Comparing WGS to SMASH profiles under optimized pipeline parameters.

[0112] The performance of WGS and SMASH was compared using autosomes and X-chromosomes as described above. Different total numbers of bins (from 50,000 to 500,000), different mean numbers of maps per bin (20, 50 and 100), collecting statistics for signal-to-noise and autocorrelation were considered, among other factors. The two methods have very similar performance characteristics (Table 2). WGS, map for map, slightly outperforms SMASH. When bin boundaries were chosen such that the reference sample has the same number of maps in each bin, the signal-to-noise ratio improves for both SMASH and WGS, and the difference between them narrows substantially (Supplementary Table 1).

Table 2

	type	# of bins	maps per bin	autosome auto-correlation	autosome median	X-chrom median	autosome MAD	X-chrom MAD	signal to noise
5	smash	50000	20	-0.002	2.032	1.111	0.297	0.218	2.497
	wgs	50000	20	0.000	2.031	1.072	0.295	0.208	2.659
	smash	50000	50	0.006	2.009	1.068	0.194	0.140	3.933
10	wgs	50000	50	0.000	2.007	1.032	0.191	0.135	4.173
	smash	50000	100	0.009	2.002	1.056	0.141	0.100	5.487
	wgs	50000	100	0.008	2.002	1.019	0.138	0.095	5.861
	smash	100000	20	0.004	2.033	1.108	0.298	0.224	2.481
15	wgs	100000	20	0.003	2.031	1.070	0.297	0.212	2.633
	smash	100000	50	0.011	2.010	1.071	0.196	0.146	3.833
	wgs	100000	50	0.012	2.008	1.032	0.194	0.138	4.102
	smash	100000	100	0.019	2.003	1.056	0.145	0.105	5.289
20	wgs	100000	100	0.019	2.002	1.021	0.143	0.099	5.633
	smash	500000	20	0.008	2.033	1.109	0.318	0.233	2.342
	wgs	500000	20	0.010	2.033	1.075	0.315	0.221	2.492
	smash	500000	50	0.016	2.013	1.073	0.225	0.159	3.410
25	wgs	500000	50	0.019	2.011	1.037	0.220	0.150	3.656
	smash	500000	100	0.024	2.004	1.059	0.181	0.122	4.319
	wgs	500000	100	0.029	2.003	1.023	0.177	0.114	4.649

30 Table 2. WGS and SMASH by number of bins and maps.

[0113] The same performance statistics as in Table 1, comparing SMASH and WGS over a range of resolutions (50K, 100K, and 500K) and coverage (20, 50, and 100 maps per bin) are computed in Table 2.

35

40

45

50

55

Supplementary Table 1

	type	# of bins	maps per bin	autosome auto-correlation	autosome median	X-chrom median	autosome MAD	X-chrom MAD	signal to noise	S/N from Table 2
5	smash	50000	20	0.006	2.020	1.120	0.295	0.161	2.679	2.497
	wgs	50000	20	0.002	2.021	1.079	0.296	0.160	2.800	2.659
10	smash	50000	50	0.009	2.005	1.074	0.193	0.106	4.234	3.933
	wgs	50000	50	0.008	2.005	1.036	0.192	0.099	4.480	4.173
	smash	50000	100	0.015	2.002	1.060	0.140	0.077	5.886	5.487
	wgs	50000	100	0.013	2.002	1.026	0.140	0.074	6.149	5.861
15	smash	100000	20	0.003	2.020	1.116	0.298	0.159	2.681	2.481
	wgs	100000	20	0.005	2.020	1.078	0.299	0.157	2.788	2.633
	smash	100000	50	0.016	2.006	1.074	0.197	0.108	4.148	3.833
	wgs	100000	50	0.008	2.005	1.039	0.197	0.102	4.368	4.102
20	smash	100000	100	0.024	2.002	1.061	0.146	0.078	5.697	5.289
	wgs	100000	100	0.017	2.002	1.024	0.145	0.077	5.979	5.633
	smash	500000	20	0.009	2.021	1.113	0.317	0.170	2.524	2.342
	wgs	500000	20	0.008	2.020	1.078	0.315	0.163	2.657	2.492
25	smash	500000	50	0.018	2.006	1.075	0.225	0.116	3.677	3.410
	wgs	500000	50	0.017	2.005	1.037	0.222	0.109	3.906	3.656
	smash	500000	100	0.028	2.001	1.060	0.182	0.090	4.622	4.319
	wgs	500000	100	0.026	2.001	1.025	0.179	0.084	4.939	4.649

30

Supplementary Table 1. Empirical bin boundaries.

[0114] The computations of Table 2 are repeated, but instead of bins of uniform expectation, bins of uniform observation of a reference are used. The bin boundaries are defined empirically: establishing bins with the same number of maps as determined empirically. The signal-to-noise is improved over the results in Table 2 ("S/N from Table 2"), with little change to the autocorrelation.

[0115] Note that as the number of bins increases, the signal-to-noise ratio diminishes: from 5.6 at 50K bins to 4.0 at 500K bins for SMASH. Similar degradation of signal occurs for WGS. It was hypothesized that this was the result of using the same total number of reference maps for normalization, independent of the number of bins. Therefore, as the number of bins increases, the number of reference maps per bin diminishes, increasing the variance of the normalized ratio. To test if this was the cause, reference normalization was performed-this time matching the total number of reference maps to the total number of sample maps. There was virtually no degradation of signal-to-noise ratio as the bin number increased (Supplementary Table 2).

45

50

55

Supplementary Table 2

	type	# of bins	maps per bin	autosome auto- correlation	autosome median	X- chrom median	autosome MAD	X- chrom MAD	signal to noise
5	smash	50000	20	0.002	2.027	1.112	0.416	0.272	1.840
	smash	100000	20	0.004	2.028	1.122	0.415	0.279	1.812
	smash	500000	20	0.003	2.025	1.112	0.417	0.279	1.819
10	smash	50000	50	0.004	2.019	1.075	0.270	0.176	2.928
	smash	100000	50	0.005	2.018	1.077	0.269	0.180	2.907
	smash	500000	50	0.011	2.019	1.076	0.271	0.178	2.908
	smash	50000	100	0.005	2.003	1.058	0.190	0.124	4.160
15	smash	100000	100	0.009	2.005	1.061	0.190	0.126	4.130
	smash	500000	100	0.021	2.005	1.059	0.194	0.128	4.073
	wgs	50000	20	0.004	2.034	1.080	0.413	0.261	1.953
	wgs	100000	20	0.001	2.032	1.075	0.414	0.261	1.955
20	wgs	500000	20	0.003	2.030	1.078	0.413	0.263	1.944
	wgs	50000	50	0.003	2.017	1.036	0.267	0.164	3.132
	wgs	100000	50	0.006	2.016	1.039	0.266	0.168	3.111
	wgs	500000	50	0.012	2.016	1.040	0.267	0.168	3.095
25	wgs	50000	100	0.008	2.003	1.017	0.189	0.114	4.452
	wgs	100000	100	0.011	2.003	1.022	0.189	0.118	4.400
	wgs	500000	100	0.025	2.003	1.024	0.189	0.119	4.385

Supplementary Table 2. Matching reference and sample coverage.

[0116] Performance statistics as in Table 2 are computed. In this table, however, the same number of maps for both the sample and the reference are used for each choice of bin resolution (50K, 100K, 500K) and for each map coverage (20, 50 and 100 reads per bin). When the number of maps are equalized between sample and reference, the signal to noise is largely insensitive to the bin resolution and depends strongly on the map coverage for both WGS and SMASH, indicating that only the depth of coverage limits resolution.

[0117] Finally, the actual profiles of samples using SMASH and WGS were compared. Bins optimized for WGS and the map selection rules discussed above were used. Genomic DNAs from two families using reference normalization (Fig. 3) and one cancer cell line without reference normalization (Fig. 4) were analyzed. For comparison, both WGS and SMASH were down-sampled to an equal number of maps. Across all scales of genome resolution - whether looking at normalized bin counts or segmented data - the profiles from the two methods look very similar. In both figures, 10 million maps distributed into 100,000 bins are shown. Parental transmission patterns appeared largely Mendelian (Fig. 3A). This is illustrated clearly in Fig. 3B, which zooms to show the transmission of a deletion from the father to an unaffected sibling. While the global segmentation patterns generated by SMASH and WGS are not completely identical, much of the variation has to do with segmentation itself. When considering bin concordance, WGS and SMASH are exceedingly similar (Fig. 3C).

[0118] Both WGS and SMASH yielded approximately the same integer-valued copy number profile for the cancer cell line SKBR3 (Fig. 4A). The copy number profiles are well matched to integer states. To illustrate the concordance between the data, a chromosome with extensive genomic copy number variation is shown in greater detail (Fig. 4B). Again, the bin-for-bin LOESS adjusted ratios are largely concordant (Fig. 4C).

Example 4. An alternate fractionation protocol for SMASH.

[0119] All of the above data derive from a version of SMASH that combines sonication and restriction endonuclease (RE) cleavage. A version that did not depend on either of those methods for genomic fragmentation, and that might be more amenable to ideal segment length distribution and randomness of SMASH maps was desirable. For this purpose NEBNext dsDNA Fragmentase (NEB) was used. NEBNext dsDNA Fragmentase (NEB) is a combination of enzymes that randomly generates nicks on dsDNA, then cuts the DNA strand opposite the nick to produce dsDNA breaks. Using

EP 3 347 466 B9

recommended conditions, segment lengths with a tighter size distribution and somewhat shorter than those obtained by sonication and RE cleavage were readily obtained. Ligation of the segments and size-selection of the fragments to an optimal length for sequencing was readily accomplished (Fig. 6). This method was then compared to our initial protocol on genomic DNA from the cancer cell line SKBR3, without normalization. The copy number profiles generated by the two methods were virtually identical (Fig. 7). The average number of maps per read increases from greater than four to more than six with the fragmentase method. The improvement is likely due to more precise sizing in this protocol. The detailed SMASH library preparation using the alternative protocol is outlined below:

Step 1 - dsDNA Fragmentation.

[0120] Set up the fragmentation reaction as follows:

Component	stock conc.	unit	vol. (μl)
Genomic DNA (200 ng - 1 μg)	varies	ng/μl	x
Fragmentase reaction Buffer v2	10	x	1
MgCl₂	200	mM	0.5
dsDNA Fragmentase (NEB, M0348L)			1
H₂O			Y
Total			10

[0121] Incubate tubes in a thermal cycler for 10 minutes at 37°C, then put the tubes on ice.

Step 2 - End-repair.

[0122] Add the following reagents into the same tube(s) as step 1:

Component	stock conc.	unit	vol. (μl)
ATP (NEB, P0756L)	10	mM	2
dNTPs (Roche 11814362001)	10	mM	1
T4 DNA Polymerase (NEB M0203L)	3	U/ul	1
Klenow Polymerase, large fragment (NEB M0210L)	5	U/μl	0.5
T4 PNK (NEB M0201L)	10	U/μl	1
H₂O			4.5
Fragmented DNA	25	ng/ul	10
total			20

[0123] Incubate the sample in a thermal cycler for 30 minutes at 20°C. Size select with AMPure XP beads (2.5X), mix well, incubate at RT for 5 min, collect supernatant, purify by nucleotide removal kit (Qiagen), and elute with 30 μl H₂O. Take 1 μl aliquot for Bioanalyzer.

Step 3 - Self random ligation.

[0124] Prepare the following reaction mix in a new 0.2 ml PCR tube:

Component	stock conc.	unit	vol. (μl)
DNA Quick Ligase Buffer	2	x	29
Quick DNA Ligase (NEB, M2200L)			1.5
Eluted DNA from step 2			27.5
Total			58

EP 3 347 466 B9

[0125] Incubate in a thermal cycler at 25°C for 15 min. Purify by AMPure XP bead (1.6X, 92.8 ul bead), wash twice with 180 ul 80% ethanol, air dry, elute by 25 ul H₂O, add to new PCR tube. Take 1 µl aliquot for Bioanalyzer.

Step 4 - Second end-repair.

[0126] Prepare the following reaction mix in a new 0.2 ml nuclease-free PCR tube:

Component	vol. (µl)
T4 DNA lig buffer w/10 mM ATP (w/DTT, B0202) 10X	3
dNTPs (Roche, 11814362001, or 04638956001) 10 mM	1
T4 DNA Polymerase (NEB M0203L) 3 U/µl	1
T4 PNK (NEB M0201L) 10 U/µl	1
Klenow Polymerase, large fragment (NEB M0210L) 5 U/µl	0.5
Size-selected DNA from step 3	23.5
Total	30

[0127] Incubate the sample on a thermal cycler for 30 minutes at 20°C. Purify with AMPure XP beads (1.6X, 48 µl), RT for 10 min, wash twice with 180 µl of 80% ethanol, elute by 21 ul H₂O.

Step 5 - Adenylate 3' ends.

[0128] Prepare the following reaction mix in a new 0.2 ml nuclease-free PCR tube:

Component	vol. (µl)
Eluted DNA from step 4	20
NEBuffer #2 10X	2.5
dATP (Roche, 100 mM, 11934511001) 2 mM	1
Klenow fragment 3'-5' exo (NEB M0212L) 5 U/µl	1.5
Total	25

[0129] Incubate the sample in a thermal cycler for 30 minutes at 37°C. Purify with AMPure XP bead (1.6X, 40 µl), incubate at RT for 10 min, wash twice with 180 µl of 80% ethanol x2, elute with 14 µl H₂O.

Step 6 - Ligate with adapters and size select with AMPure XP beads.

[0130] Prepare the following reaction mix in a new 0.2 ml nuclease-free PCR tube:

Component	stock conc	unit	Vol. (µl)
Product from step 5			13
DNA Quick Ligase Buffer	2	x	15
Barcoded adapters	10	uM	1
Quick DNA Ligase (NEB, M2200L)		U/ul	1
Total			30

[0131] Incubate at 25°C for 10 min. Purify by AMPure bead (1.6X, 48 µl), wash twice with 80% ethanol, elute with 50 µl H₂O. Size select with AMPure beads (0.6X, 30 ul), mix well and incubate at RT for 10 min, collect supernatant, add AMPure beads (0.16X, 8 µl), mix well and incubate at RT for 10 min, wash twice with 180 µl 80% ethanol, and elute with 16 µl H₂O.

Step 7 - enrichment PCR.

[0132] Set up PCR reaction as follows:

Component	stock conc.	unit	vol. (μ l)
Phusion mm (M0531L)	2	x	20
DNA from step 6			15
PE5 & PE7 primers	5	μ M (ea.)	2
H ₂ O			3
Total			40

[0133] Amplify under following conditions: denature at 98°C for 30 sec; perform 8 cycles of denaturing at 98°C for 5 sec, primer annealing at 65°C for seconds, and template extension at 72°C for 30 sec; final extension at 72°C for 10 min. Purify by AMPure beads (0.9X, 36 μ l), wash twice with 180 μ l 80% ethanol, elute with 20 μ l H₂O. Measure concentration by Nanodrop, take aliquot and dilute to 10 ng/ μ l for Bioanalyzer. The SMASH DNA library is now ready for sequencing.

[0134] Thus, the two steps of sonication and the restriction enzyme digestion in the general protocol have been replaced by one step of fragmentation with dsDNA Fragmentase (NEB) in the alternative protocol. Accordingly, the first end-repair reaction is right after the fragmentation step - there is no longer any need for purification between these two steps. Additionally, all enzyme heat-killing steps have been eliminated in the alternative protocol because enzymes are adequately removed by bead purification. Ultimately, the overall time requirement for the SMASH library preparation has been reduced by approximately one hour using the alternative protocol.

Discussion

[0135] Copy number variants (CNVs) underlie a significant amount of genetic diversity and disease. For example, Autism Spectrum Disorders (ASD) are highly influenced by genetic factors (Muhle et al., 2004; Rosenberg et al., 2009), and CNVs underlie a significant fraction of those diagnoses. Beyond ASD, copy number variants have been shown to play a role in multiple diseases, including congenital heart disease (Warburton et al., 2014), cancer (Stadler et al., 2012; Lockwood et al., 2007; Lu et al., 2011; Shlien and Malkin, 2009), schizophrenia (Szatkiewicz et al., 2014; Rees et al., 2014) and even in patients, responses to certain therapies (Willyard, 2015). CNVs can be detected by a number of means, including chromosomal microarray analysis (CMA) and whole genome sequencing (WGS), but these approaches suffer from either limited resolution (CMA) or are highly expensive for routine screening (both CMA and WGS).

[0136] In obtaining copy number information from high throughput sequencing, SMASH has a clear advantage over standard WGS. Each read is packed with multiple independent mappings, increasing the information density per read and thereby lowering cost per sample. Map for map, SMASH is comparable in quality to WGS with respect to copy number profiling. There is, of course, an enormous amount of additional structural information present in WGS data that is missing in SMASH, such as breakpoints of copy number events, small scale indels, or inversions, as a consequence of the longer reads. However, discovery of such structural events by WGS typically requires much higher coverage than what is needed for copy number determination. For detecting CNVs several kb and larger, the choice should be driven by cost.

[0137] Significant effort was invested in optimizing the design of the SMASH protocol and algorithms. These include choice of restriction enzymes and sonication conditions, heuristics for selecting maps from SMASH reads and reference sample normalization. The result is a robust method that performs at parity with WGS on a map-for-map basis. Additional changes could further increase the number of useful SMASH maps per read - the fragmentation protocol is currently set for a median of ~40 bp segments, which is optimal using the existing mapping algorithm. However, variation in segment lengths is problematic, and this variation could be reduced by adjusting the fragmentation conditions and performing more stringent size selection. To this end, the use of DNases to create random fragments with a mean of 35bp has been explored to address the issue of segment length variation. With this somewhat simplified protocol, more maps per read with comparable resolution on a map-for-map basis were obtained in preliminary experiments.

[0138] For most of the analysis of maps, bin boundaries determined for WGS were used so that SMASH could be directly compared to WGS. However, the optimal bin boundaries were shown to be those derived empirically to yield uniform map counts (Supplementary Table 2). Furthermore, it is clear that increasing the reference coverage will improve signal to noise for all samples. A lower limit to the resolution that can be obtained has not yet been determined.

[0139] Advances in sequencing technology that reduce unit cost per base pair will likely be driven by increasing read

lengths. For copy number inference from whole genome sequencing, this means a continued decline in the number of maps per base. However, SMASH, even with existing sequencers, can yield 4-6 times as many maps as standard WGS. On a machine that generates 300 million 150-bp paired-end reads for \$1500, 60 million maps per sample for 30 samples at unit cost of \$50 per sample and a resolution of ~10 kb can be obtained, not including the preparation costs for the libraries. However, using the same SMASH library, resolution and cost will be roughly linear to number of reads. Thus, SMASH can reduce the costs of testing in prenatal, pediatric and cancer genetics, allowing more patients to be tested at a lower cost and the resultant savings passed along to researchers and caregivers.

[0140] Ultimately, genomic copy number information can be used to test for prenatal, pediatric, developmental, psychological and autoimmune disorders, as well as susceptibility to disease. Examples of disorders and diseases which can be tested for using genomic copy number information include, but are not limited to, Autism Spectrum Disorders, schizophrenia, cancer and congenital heart disease. In addition to testing and diagnosis, copy number information may also be utilized to predict the likelihood of displaying or probability of inheriting a disease, syndrome or disorder. Finally, outside of the clinic SMASH may also prove to be a valuable tool for determining copy number variation in agriculturally important plants and crops.

References

[0141]

1. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, Sahinalp SC, Gibbs RA, Eichler EE. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature genetics*. 2009;41(10):1061-7. doi: 10.1038/ng.437. PubMed PMID: 19718026; PubMed Central PMCID: PMC2875196.
2. Fishbach GD, Lord C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron*. 2010; 68:192-195.
3. Khan Z, Bloom JS, Kruglyak L, Singh M. A practical algorithm for finding maximal exact matches in large sequence datasets using sparse suffix arrays. *Bioinformatics*. 2009;25(13):1609-16. doi: 10.1093/bioinformatics/btp275. PubMed PMID: 19389736; PubMed Central PMCID: PMC2732316.
4. Levy D, Wigler M. Facilitated sequence counting and assembly by template mutagenesis. *Proceedings of the National Academy of Sciences of the United States of America*. 2014;111(43):E4632-7. doi: 10.1073/pnas.1416204111. PubMed PMID: 25313059; PubMed Central PMCID: PMC4217440.
5. Lockwood WW, Coe BP, Williams AC, MacAulay C, Lam WL. Whole genome tiling path array CGH analysis of segmental copy number alterations in cervical cancer cell lines. *International journal of cancer Journal international du cancer*. 2007;120(2):436-43. doi: 10.1002/ijc.22335. PubMed PMID: 17096350.
6. Lu TP, Lai LC, Tsai MH, Chen PC, Hsu CP, Lee JM, Hsiao CK, Chuang EY. Integrated analyses of copy number variations and gene expression in lung adenocarcinoma. *PloS one*. 2011;6(9):e24829. doi: 10.1371/journal.pone.0024829. PubMed PMID: 21935476; PubMed Central PMCID: PMC3173487.
7. Muhle R, Trentacoste SV, Rapin I. The genetics of autism. *Pediatrics*. 2004;113(5):e472-86. PubMed PMID: 15121991.
8. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, Muthuswamy L, Krasnitz A, McCombie WR, Hicks J, Wigler M. Tumour evolution inferred by single-cell sequencing. *Nature*. 2011;472(7341):90-4. doi: 10.1038/nature09807. PubMed PMID: 21399628; PubMed Central PMCID: PMC4504184.
9. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*. 2004;5:557-572.
10. Rees E, Walters JT, Georgieva L, Isles AR, Chambert KD, Richards AL, Mahoney-Davies G, Legge SE, Moran JL, McCarroll SA, O'Donovan MC, Owen MJ, Kirov G. Analysis of copy number variations at 15 schizophrenia-associated loci. *The British journal of psychiatry: the journal of mental science*. 2014;204(2):108-14. doi: 10.1192/bjp.bp.113.131052. PubMed PMID: 24311552; PubMed Central PMCID: PMC3909838.

11. Rosenberg RE, Law JK, Yenokyan G, McGready J, Kaufmann WE, Law PA. Characteristics and concordance of autism spectrum disorders among 277 twin pairs. *Archives of pediatrics & adolescent medicine*. 2009;163(10):907-14. doi: 10.1001/archpediatrics.2009.98. PubMed PMID: 19805709.

5 12. Shlien A and Malkin D. Copy number variations and cancer. *Genome Medicine*. 2009;1(6):62. doi: 10.1186/gm62. PMID: 19566914. PMCID: PMC2703871.

13. Stadler ZK, Esposito D, Shah S, Vijai J, Yamrom B, Levy D, Lee YH, Kendall J, Leotta A, Ronemus M, Hansen N, Sarrel K, Rau-Murthy R, Schrader K, Kauff N, Klein RJ, Lipkin SM, Murali R, Robson M, Sheinfeld J, Feldman D, Bosl G, Norton L, Wigler M, Offit K. Rare de novo germline copy-number variation in testicular cancer. *American journal of human genetics*. 2012;91(2):379-83. doi: 10.1016/j.ajhg.2012.06.019. PubMed PMID: 22863192; PubMed Central PMCID: PMC3415553.

14. Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, Genomes P, Eichler EE. Diversity of human copy number variation and multicopy genes. *Science*. 2010;330(6004):641-6. doi: 10.1126/science.1197005. PubMed PMID: 21030649; PubMed Central PMCID: PMC3020103.

15 15. Szatkiewicz JP, O'Dushlaine C, Chen G, Chambert K, Moran JL, Neale BM, Fromer M, Ruderfer D, Akterin S, Bergen SE, Kahler A, Magnusson PK, Kim Y, Crowley JJ, Rees E, Kirov G, O'Donovan MC, Owen MJ, Walters J, Scolnick E, Sklar P, Purcell S, Hultman CM, McCarroll SA, Sullivan PF. Copy number variation in schizophrenia in Sweden. *Molecular psychiatry*. 2014;19(7):762-73. doi: 10.1038/mp.2014.40. PubMed PMID: 24776740; PubMed Central PMCID: PMC4271733.

16. Warburton D, Ronemus M, Kline J, Jobanputra V, Williams I, Anyane-Yeboah K, Chung W, Yu L, Wong N, Awad D, Yu CY, Leotta A, Kendall J, Yamrom B, Lee YH, Wigler M, Levy D. The contribution of de novo and rare inherited copy number changes to congenital heart disease in an unselected sample of children with conotruncal defects or hypoplastic left heart disease. *Human genetics*. 2014;133(1):11-27. doi: 10.1007/s00439-013-1353-9. PubMed PMID: 23979609; PubMed Central PMCID: PMC3880624.

17. Willyard C. Copy number variations' effect on drug response still overlooked. *Nature medicine*. 2015;21(3):206. doi: 10.1038/nm0315-206. PubMed PMID: 25742449.

18. Weischenfeldt J. et al. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nature Reviews Genetics*. 2013; 14(2):125-138.

19. Malhotra D. et al. CNVs: Harbingers of a rare variant revolution in psychiatric genetics. 2012. *Cell*, 148(6):1223-1241.

20. Anson et al. Next-generation DNA sequencing techniques. 2009. *New Biotechnology*, 25(4):195-203.

21. Wang Z. et al. SMASH, a fragmentation and sequencing method for genomic copy number analysis. 2016. *Genome Research*, 26(6):844-851.

45 Claims

1. A sequencing library composition comprising a first mixture of different chimeric genomic nucleic acid fragments, wherein the mixture of different chimeric genomic nucleic acid fragments contains at least 100,000 different fragments, wherein the chimeric genomic nucleic acid fragments are 250 ($\pm 10\%$) to less than 1000 ($\pm 10\%$) base pairs in length, wherein each different fragment in the mixture comprises randomly ligated DNA segments, wherein each DNA segment in the fragment is a nucleic acid molecule at least 27 base pairs in length resulting from random fragmentation of a single genome for which a reference genome is available, wherein at least 50% of the segments in the at least 100,000 different fragments are 30 to 50 base pairs in length ($\pm 10\%$), further comprising sequence adaptors ligated to the termini of the chimeric genomic nucleic acid fragments, wherein the sequence adaptors comprise a barcode identifying the sample origin of each fragment.

2. The composition of claim 1, wherein the sequence adaptors comprise a barcode identifying the genomic source of each fragment, and/or comprise a primer binding site for amplification.

3. The composition of claim 1 or 2:

wherein the segments are ligated directly to each other to form a fragment;
 wherein the DNA segments are 30 to 50 base pairs in length ($\pm 10\%$);
 wherein the mixture of different chimeric genomic nucleic acid fragments contains fragments composed of an
 odd number of segments; and/or
 wherein the mixture of chimeric genomic nucleic acid fragments contain ligated segments whose two ligation
 points form a sequence other than a restriction enzyme recognition site.

4. The composition of any one of claims 1-3, enriched for chimeric genomic nucleic acid fragments 250 ($\pm 10\%$) to 700 ($\pm 10\%$) base pairs in length, preferably 400-500 base pairs, and/or wherein at least 50% of the chimeric genomic nucleic acid fragments in the mixture are 250 ($\pm 10\%$) to 700 ($\pm 10\%$) base pairs in length, preferably 400-500 base pairs.

5. The composition of any one of claims 1-4, further comprising a second mixture of different chimeric genomic nucleic acid fragments, wherein the second mixture of fragments is obtained from a different genome than the first mixture, optionally comprising a collection of multiple mixtures of different chimeric genomic nucleic acid fragments, wherein each mixture of fragments in the collection is obtained from a different genome than any other mixture in the collection, preferably wherein each mixture of chimeric genomic nucleic acid fragments contains fragments having a sequencing adaptor containing a unique barcode ligated onto only fragments within the mixture, such that the collection of mixtures can be multiplexed.

6. A method for obtaining the composition according to any one of claims 1-5, comprising

- i) randomly fractionating the single genome to obtain random segments from the genome, preferably size selecting a subpopulation of segments 30 to 50 base pairs in length ($\pm 10\%$) prior to ligation, and/or wherein the subpopulation of segments is selected using bead purification;
- ii) subjecting the segments from step (i) to ligation to generate different chimeric genomic nucleic acid fragments; and
- iii) ligating sequencing adaptors to the chimeric genomic nucleic acid fragments,

thereby obtaining the mixture of different genomic nucleic acid fragments from the single genome.

7. The method of claim 6 further comprising adenylating the 3' termini of the chimeric genomic nucleic acid fragments prior to step iii).

8. The method of claims 6 or 7 further comprising selecting for and including in the mixture sequence adaptor-ligated genomic nucleic acid fragments about 250 base pairs in length to about 1000 base pairs in length.

9. The method of any of claims 6-8, wherein the sequence adaptor ligated to the termini of the chimeric genomic nucleic acid fragments comprises a primer binding site for amplification, wherein the method preferably further comprises amplifying the size-selected sequence adaptor-ligated genomic nucleic acid fragments.

10. The method of any of claims 6-9, wherein in step (i) the genomic nucleic acids are mechanically sheared to obtain the randomly fragmented DNA segments, preferably wherein the mechanical shearing is by sonication, and/or further comprising subjecting the segments of genomic nucleic acids to enzymatic digestion, which is preferably by the restriction enzymes CviKI-1 and NlaIII; or

wherein in step (i) genomic nucleic acids are enzymatically fragmented, by

- a) generating random DNA nicks in the genome; and
- b) cutting the DNA strand opposite the nick,

thereby producing dsDNA breaks in the genomic nucleic acids resulting in DNA segments; and/or wherein the resulting DNA segments are end-repaired directly after genomic fragmentation, and/or wherein chimeric genomic nucleic acid fragments are end-repaired after their formation by random segment ligation.

11. The method of any of claims 6-10, wherein the initial amount of genomic nucleic acids is 200 ng, 500ng, or 1 μ g

($\pm 10\%$);

and/or wherein the genomic nucleic acids are obtained from a cell, a tissue, a tumor, a cell line or from blood.

5 12. A process of obtaining the nucleic acid sequence of the different chimeric genomic nucleic acid fragments of the composition of any one of claims 1-5, or produced by the method of claims 6-11, comprising (i) obtaining the fragments, and (ii) sequencing the fragments, preferably using a next-generation sequencing platform, so as to obtain the nucleic acid sequence of the different chimeric genomic nucleic acid fragments.

10 13. A process for obtaining genomic copy number information from a genome, comprising

i) obtaining the nucleic acid sequence of the different chimeric genomic nucleic acid fragments of (a) a sequencing library composition comprising a first mixture of different chimeric genomic nucleic acid fragments, wherein the mixture of different chimeric genomic nucleic acid fragments contains at least 100,000 different fragments, wherein the chimeric genomic nucleic acid fragments are 250 ($\pm 10\%$) to less than 1000 ($\pm 10\%$) base pairs in length, wherein each different fragment in the mixture comprises randomly ligated DNA segments, wherein each DNA segment in the fragment is a nucleic acid molecule at least 27 base pairs in length resulting from random fragmentation of a single genome for which a reference genome is available, wherein at least 50% of the segments in the at least 100,000 different fragments are 30 to 50 base pairs in length ($\pm 10\%$), (b) the composition of any one of claims 1-5, or (c) the composition produced by the method of claims 6-11;

15 16 ii) identifying and mapping to a genome each Maximal Almost-unique Match (MAM) within a sequenced chimeric genomic nucleic acid fragment, preferable MAMs are identified using a longMEM software package, further preferably MAMs are filtered by discarding MAMs less than twenty base pairs and not at least four base pairs longer than required for uniqueness, and/or MAMs are filtered by discarding MAMs in a read-pair map that are within 10,000 base pairs of one another; and

20 21 iii) counting the number of mapped MAMs within a binned genome,

25 thereby obtaining genomic copy number information.

30 14. The process of claim 13, wherein in step (iii) the number of mapped reads are counted in genome bin sizes that yield uniform map counts for the reference sample, wherein in step (iii) the number of mapped reads are counted in empirically determined genome bins of uniform observation of a reference, wherein in step (iii) the number of mapped reads are counted in genome bins of expected uniform density, wherein in step (iii) the number of mapped reads in each bin is adjusted for GC bias by LOESS normalization, wherein in step (iii) template analysis is utilized to reduce systematic noise in GC adjusted bin count data, wherein in step (iii) a reference normalization is applied to bin count data by dividing GC-adjusted bin ratios by a standard sample bin ratio, wherein in step (iii), reference normalized GC-adjusted bin count data is analyzed by circular binary segmentation, and/or wherein in step (iii) the total number of reference maps is matched to the total number of sample maps.

35 40 15. A method of diagnosing, predicting likelihood of displaying or determining the probability of inheriting a prenatal disorder, a pediatric disorder, a developmental disorder, a psychological disorder, an autoimmune disorder, cancer, congenital heart disease, schizophrenia, Autism Spectrum Disorders or a patient's response to a therapy, comprising obtaining the patient's genomic copy number information by the process of claim 13 or 14.

45 Patentansprüche

1. Sequenzierungsbibliotheks-Zusammensetzung, die eine erste Mischung verschiedener chimärer genomischer Nukleinsäurefragmente umfasst, wobei die Mischung verschiedener chimärer genomischer Nukleinsäurefragmente mindestens 100.000 verschiedene Fragmente enthält, wobei die chimären genomischen Nukleinsäurefragmente eine Länge von 250 ($\pm 10\%$) bis weniger als 1000 ($\pm 10\%$) Basenpaaren aufweisen, wobei jedes verschiedene Fragment in der Mischung zufällig ligierte DNA-Segmente umfasst, wobei jedes DNA-Segment im Fragment ein Nukleinsäuremolekül mit einer Länge von mindestens 27 Basenpaaren ist, das aus einer zufälligen Fragmentierung eines einzelnen Genoms resultiert, für das ein Referenzgenom verfügbar ist, wobei mindestens 50 % der Segmente in den mindestens 100.000 verschiedenen Fragmenten eine Länge von 30 bis 50 Basenpaaren ($\pm 10\%$) aufweisen, ferner umfassend Sequenzadapter, die an die Enden der chimären genomischen Nukleinsäurefragmente ligiert sind, wobei die Sequenzadapter einen Barcode umfassen, der den Probenursprung jedes Fragments identifiziert.

2. Zusammensetzung nach Anspruch 1, wobei die Sequenzadapter einen Barcode umfassen, der die genomische

Quelle jedes Fragments identifiziert, und/oder eine Primerbindungsstelle für die Amplifikation umfassen.

3. Zusammensetzung nach Anspruch 1 oder 2:

wobei die Segmente direkt miteinander ligiert werden, um ein Fragment zu bilden;
wobei die DNA-Segmente 30 bis 50 Basenpaare ($\pm 10\%$) lang sind;
wobei die Mischung verschiedener chimärer genomischer Nukleinsäurefragmente Fragmente enthält, die aus einer ungeraden Anzahl von Segmenten bestehen; und/oder
wobei die Mischung chimärer genomischer Nukleinsäurefragmente ligierte Segmente enthält, deren zwei Ligationenpunkte eine andere Sequenz als eine Restriktionsenzym-Erkennungsstelle bilden.

4. Zusammensetzung nach einem der Ansprüche 1-3, die bezüglich chimärer genomischer Nukleinsäurefragmente mit einer Länge von 250 ($\pm 10\%$) bis 700 ($\pm 10\%$) Basenpaaren angereichert ist, vorzugsweise 400-500 Basenpaaren, und/oder wobei mindestens 50% der chimären genomischen Nukleinsäurefragmente in der Mischung eine Länge von 250 ($\pm 10\%$) bis 700 ($\pm 10\%$) Basenpaaren haben, vorzugsweise 400-500 Basenpaare.

5. Zusammensetzung nach einem der Ansprüche 1-4, weiterhin umfassend eine zweite Mischung verschiedener chimärer genomischer Nukleinsäurefragmente, wobei die zweite Mischung von Fragmenten aus einem anderen Genom als der erste Mischung erhalten wird, optional umfassend eine Sammlung mehrerer Mischungen aus verschiedenen chimären genomischen Nukleinsäurefragmenten, wobei jede Mischung von Fragmenten in der Sammlung aus einem anderen Genom als jede andere Mischung in der Sammlung erhalten wird, wobei vorzugsweise jede Mischung aus chimären genomischen Nukleinsäurefragmenten Fragmente mit einem Sequenzierungsadapter enthält, der einen eindeutigen Barcode enthält, der nur an Fragmente innerhalb der Mischung ligiert ist, so dass die Sammlung von Mischungen gemultiplext werden kann.

6. Verfahren zur Gewinnung der Zusammensetzung nach einem der Ansprüche 1-5, umfassend

- i) zufällige Fraktionierung des einzelnen Genoms, um zufällige Segmente aus dem Genom zu erhalten, wobei vor der Ligation vorzugsweise eine Subpopulation von Segmenten mit einer Länge von 30 bis 50 Basenpaaren ($\pm 10\%$) nach Größe selektiert wird und/oder wobei die Subpopulation von Segmenten mittels Perlen-Reinigung selektiert wird;
- ii) Unterwerfen der Segmente aus Schritt (i) einer Ligation, um verschiedene chimäre genomische Nukleinsäurefragmente zu erzeugen; und
- iii) Ligieren von Sequenzierungsadaptern mit den chimären genomischen Nukleinsäurefragmenten,

so dass dadurch die Mischung verschiedener genomischer Nukleinsäurefragmente aus dem einzelnen Genom erhalten wird.

7. Verfahren nach Anspruch 6, das außerdem die Adenylierung der 3'-Termini der chimären genomischen Nukleinsäurefragmente vor Schritt iii) umfasst.

8. Verfahren nach Anspruch 6 oder 7, das weiterhin das Auswählen für und das Einbeziehen in die Mischung von Sequenzadapter-ligierten genomischen Nukleinsäurefragmenten mit einer Länge von etwa 250 Basenpaaren bis etwa 1000 Basenpaaren umfasst.

9. Verfahren nach einem der Ansprüche 6-8, wobei der an die Termini der chimären genomischen Nukleinsäurefragmente ligierte Sequenzadapter eine Primerbindungsstelle zur Amplifikation umfasst, wobei das Verfahren vorzugsweise weiterhin die Amplifikation der gröÙenselektierten Sequenzadapter ligierten genomische Nukleinsäurefragmente umfasst.

10. Verfahren nach einem der Ansprüche 6-9, wobei in Schritt (i) die genomischen Nukleinsäuren mechanisch geschert werden, um die zufällig fragmentierten DNA-Segmente zu erhalten, wobei die mechanische Scherung vorzugsweise durch Ultraschall erfolgt und/oder weiter das Behandeln der Segmente von genomischen Nukleinsäuren durch enzymatische Verdauung umfasst, die vorzugsweise durch die Restriktionsenzyme CviKI-1 und NlaIII erfolgt; oder wobei in Schritt (i) genomische Nukleinsäuren enzymatisch fragmentiert werden, durch

- a) Erzeugen zufälliger DNA-Einschnitte im Genom; und
- b) Schneiden des DNA-Strangs gegenüber dem Einschnitt,

so dass dadurch dsDNA-Brüche in den genomischen Nukleinsäuren entstehen, die zu DNA-Segmenten führen; und/oder wobei die resultierenden DNA-Segmente direkt nach der genomischen Fragmentierung endrepariert werden und/oder wobei chimäre genomische Nukleinsäurefragmente nach ihrer Bildung durch zufällige Segmentligation endrepariert werden.

11. Verfahren nach einem der Ansprüche 6-10, wobei die Anfangsmenge an genomischen Nukleinsäuren 200 ng, 500 ng oder 1 μ g ($\pm 10\%$) beträgt; und/oder wobei die genomischen Nukleinsäuren aus einer Zelle, einem Gewebe, einem Tumor, einer Zelllinie oder aus Blut gewonnen werden.

12. Verfahren zum Erhalten der Nukleinsäuresequenz der verschiedenen chimären genomischen Nukleinsäurefragmente der Zusammensetzung nach einem der Ansprüche 1-5 oder hergestellt durch das Verfahren nach den Ansprüchen 6-11, umfassend (i) das Erhalten der Fragmente und (ii) Sequenzierung der Fragmente, vorzugsweise unter Verwendung einer Sequenzierungsplattform der nächsten Generation, um die Nukleinsäuresequenz der verschiedenen chimären genomischen Nukleinsäurefragmente zu erhalten.

13. Verfahren zum Erhalten von Informationen zur genomischen Kopienzahl aus einem Genom, umfassend:

- i) Erhalten der Nukleinsäuresequenz der verschiedenen chimären genomischen Nukleinsäurefragmente aus (a) einer Sequenzierungsbibliotheks-Zusammensetzung, die eine erste Mischung verschiedener chimärer genomischer Nukleinsäurefragmente umfasst, wobei die Mischung verschiedener chimärer genomischer Nukleinsäurefragmente mindestens 100.000 verschiedene Fragmente enthält, wobei die chimären genomischen Nukleinsäurefragmente eine Länge von 250 ($\pm 10\%$) bis weniger als 1000 ($\pm 10\%$) Basenpaaren aufweisen, wobei jedes verschiedene Fragment in der Mischung zufällig ligierte DNA-Segmente umfasst, wobei jedes DNA-Segment im Fragment ein Nukleinsäuremolekül mit einer Länge von mindestens 27 Basenpaaren ist, das aus einer zufälligen Fragmentierung eines einzelnen Genoms resultiert, für das ein Referenzgenom verfügbar ist, wobei mindestens 50 % der Segmente in den mindestens 100.000 verschiedenen Fragmenten eine Länge von 30 bis 50 Basenpaaren ($\pm 10\%$) aufweisen, (b) der Zusammensetzung nach einem der Ansprüche 1-5, oder (c) der Zusammensetzung, hergestellt nach dem Verfahren der Ansprüche 6-11;
- ii) Identifizieren und Zuordnen jedes "Maximum Almost-Unique Match" (MAM; maximale beinahe einzigartige Übereinstimmung) innerhalb eines sequenzierten chimären genomischen Nukleinsäurefragments zu einem Genom, wobei vorzugsweise MAMs unter Verwendung eines LongMEM-Softwarepakets identifiziert werden, weiter vorzugsweise MAMs durch Verwerfen von MAMs mit weniger als zwanzig Basenpaaren und nicht mindestens vier Basenpaare länger als für die Eindeutigkeit erforderlich gefiltert werden, und/oder MAMs gefiltert werden, indem MAMs in einer Lesepaarkarte verworfen werden, die innerhalb von 10.000 Basenpaare voneinander sind; und
- iii) Zählen der Anzahl der kartierten (mapped) MAMs innerhalb eines gruppierten ("binned") Genoms,

so dass dadurch Informationen zur genomischen Kopienzahl erhalten werden.

14. Verfahren nach Anspruch 13, wobei in Schritt (iii) die Anzahl der kartierten Lesevorgänge (mapped reads) in Genom-Bin-Größen gezählt wird, die einheitliche Kartenzählungen (uniform map counts) für die Referenzprobe ergeben, wobei in Schritt (iii) die Anzahl der kartierten Lesevorgänge (mapped reads) in empirisch bestimmten Genom-Bins einer einheitlichen Beobachtung einer Referenz gezählt wird, wobei in Schritt (iii) die Anzahl der kartierten Lesevorgänge (mapped reads) in Genom-Bins mit erwarteter einheitlicher Dichte gezählt wird, wobei in Schritt (iii) die Anzahl der kartierten Lesevorgänge (mapped reads) in jedem Bin für die GC-Verzerrung (GC-Bias) durch LOESS-Normalisierung angepasst wird, wobei in Schritt (iii) eine Vorlagenanalyse verwendet wird, um systematisches Rauschen in GCbereinigten Bin-Zählungen zu reduzieren, wobei in Schritt (iii) eine Referenznormalisierung auf Bin-Zählungen angewendet wird, indem GCbereinigte Bin-Verhältnisse durch ein Standard Proben-Bin-Verhältnis dividiert werden, wobei in Schritt (iii) referenznormalisierte GC-angepasste Bin-Zählungen (bin count data) durch zirkuläre binäre Segmentierung analysiert werden, und/oder wobei in Schritt (iii) die Gesamtzahl der Referenzkarten (reference maps) mit der Gesamtzahl der Probenkarten (sample maps) abgeglichen wird.

15. Methode zur Diagnose, Vorhersage der Anzeigewahrscheinlichkeit oder Bestimmung der Wahrscheinlichkeit, eine pränatale Störung, eine pädiatrische Störung, eine Entwicklungsstörung, eine psychologische Störung, eine Autoimmunerkrankung, Krebs, eine angeborene Herzkrankheit, Schizophrenie, Autismus-Spektrum-Störungen oder die Reaktion eines Patienten auf eine Therapie (a patient's response to a therapy) zu erben, umfassend das Erhalten

der Informationen zur genomischen Kopienzahl des Patienten durch das Verfahren nach Anspruch 13 oder 14.

Revendications

5

1. Composition de bibliothèque de séquençage comprenant un premier mélange de différents fragments d'acides nucléiques génomiques chimériques, dans laquelle le mélange de différents fragments d'acides nucléiques génomiques chimériques contient au moins 100 000 fragments différents, dans laquelle les fragments d'acides nucléiques génomiques chimériques ont une longueur de 250 ($\pm 10\%$) à moins de 1 000 ($\pm 10\%$) paires de bases, dans laquelle chaque fragment différent du mélange comprend des segments d'ADN ligaturés de manière aléatoire, dans laquelle chaque segment d'ADN du fragment est une molécule d'acide nucléique d'une longueur d'au moins 27 paires de bases obtenue à la suite d'une fragmentation aléatoire d'un seul génome pour lequel un génome de référence est disponible, dans laquelle au moins 50 % des segments des au moins 100 000 fragments différents ont une longueur de 30 à 50 ($\pm 10\%$) paires de bases, comprenant en outre des adaptateurs de séquence ligaturés aux extrémités des fragments d'acides nucléiques génomiques chimériques, dans laquelle les adaptateurs de séquence comprennent un code à barres identifiant l'échantillon d'origine de chaque fragment.

10

15

2. Composition selon la revendication 1, dans laquelle les adaptateurs de séquence comprennent un code à barres identifiant la source génomique de chaque fragment, et/ou comprennent un site de liaison d'amorce d'amplification.

20

3. Composition selon la revendication 1 ou la revendication 2 :

dans laquelle les segments sont ligaturés directement les uns aux autres pour former un fragment ;

dans laquelle les segments d'ADN ont une longueur de 30 à 50 ($\pm 10\%$) paires de bases ;

dans laquelle le mélange de différents fragments d'acides nucléiques génomiques chimériques contient des fragments composés d'un nombre impair de segments ; et/ou

dans laquelle le mélange de fragments d'acides nucléiques génomiques chimériques contient des segments ligaturés dont deux points de ligature forment une séquence autre qu'un site de reconnaissance d'enzyme de restriction.

25

30

4. Composition selon l'une quelconque des revendications 1 à 3, enrichie de fragments d'acides nucléiques génomiques chimériques d'une longueur de 250 ($\pm 10\%$) à 700 ($\pm 10\%$) paires de bases, de préférence de 400 à 500 paires de bases, et/ou dans laquelle au moins 50 % des fragments d'acides nucléiques génomiques chimériques du mélange ont une longueur de 250 ($\pm 10\%$) à 700 ($\pm 10\%$) paires de bases, de préférence de 400 à 500 paires de bases.

35

5. Composition selon l'une quelconque des revendications 1 à 4, comprenant en outre un second mélange de différents fragments d'acides nucléiques génomiques chimériques, dans laquelle le second mélange de fragments est obtenu à partir d'un génome différent de celui du premier mélange, comprenant éventuellement une collection de multiples mélanges de différents fragments d'acides nucléiques génomiques chimériques, dans laquelle chaque mélange de fragments de la collection est obtenu à partir d'un génome différent de celui de tout autre mélange de la collection, dans laquelle, de préférence, chaque mélange de fragments d'acides nucléiques génomiques chimériques contient des fragments comportant un adaptateur de séquençage contenant un unique code à barres qui n'est ligaturé qu'à des fragments compris dans le mélange, de sorte que la collection de mélanges puisse être multiplexée.

40

45

6. Méthode pour obtenir la composition selon l'une quelconque des revendications 1 à 5, comprenant les étapes consistant à

i) fractionner de manière aléatoire l'unique génome pour obtenir des segments aléatoires à partir du génome, de préférence sélectionner une taille d'une sous-population de segments d'une longueur de 30 à 50 ($\pm 10\%$) paires de bases avant ligature, et/ou dans laquelle la sous-population de segments est sélectionnée au moyen d'une purification par billes ;

ii) soumettre les segments obtenus à l'étape (i) à une ligature pour générer différents fragments d'acides nucléiques génomiques chimériques ; et

iii) ligaturer des adaptateurs de séquençage aux fragments d'acides nucléiques génomiques chimériques,

50

55

ce qui permet d'obtenir le mélange de différents fragments d'acides nucléiques génomiques chimériques à partir de l'unique génome.

EP 3 347 466 B9

7. Méthode selon la revendication 6, comprenant en outre l'étape consistant à adényler les extrémités 3' des fragments d'acides nucléiques génomiques chimériques avant l'étape iii).
8. Méthode selon la revendication 6 ou 7, comprenant en outre l'étape consistant à sélectionner et à inclure, dans le mélange, des fragments d'acides nucléiques génomiques ligaturés à des adaptateurs de séquence, d'une longueur d'environ 250 paires de bases à environ 1 000 paires de bases.
9. Méthode selon l'une quelconque des revendications 6 à 8, dans laquelle l'adaptateur de séquence ligaturé aux extrémités des fragments d'acides nucléiques génomiques chimériques comprend un site de liaison d'amorce d'amplification, dans laquelle la méthode comprend de préférence une étape consistant à amplifier les fragments d'acides nucléiques génomiques ligaturés à des adaptateurs de séquence de taille sélectionnée.
10. Méthode selon l'une quelconque des revendications 6 à 9, dans laquelle, à l'étape (i), les acides nucléiques génomiques sont découpés mécaniquement pour obtenir les segments d'ADN fragmentés de manière aléatoire, dans laquelle, de préférence, le découpage mécanique est effectué par sonication, et/ou comprenant en outre l'étape consistant à soumettre les segments d'acides nucléiques génomiques à une digestion enzymatique, qui est de préférence effectuée par les enzymes de restriction CviKI-1 et NlaIII ; ou
- dans laquelle, à l'étape (i), les acides nucléiques génomiques sont fragmentés de manière enzymatique, par
- a) une génération de coupures simple brin aléatoires d'ADN dans le génome ; et
- b) une coupure du brin d'ADN opposée à la coupure simple brin
- ce qui permet de produire des cassures d'ADNdb dans les acides nucléiques génomiques donnant des segments d'ADN ; et/ou
- dans laquelle les segments d'ADN ainsi obtenus sont réparés au niveau des extrémités, directement après une fragmentation génomique, et/ou dans laquelle les fragments d'acides nucléiques génomiques chimériques sont réparés au niveau des extrémités après leur formation par une ligature de segments aléatoires.
11. Méthode selon l'une quelconque des revendications 6 à 10, dans laquelle la quantité initiale d'acides nucléiques génomiques est de 200 ng, de 500 ng ou de 1 µg ($\pm 10\%$) ; et/ou dans laquelle les acides nucléiques génomiques sont obtenus à partir d'une cellule, d'un tissu, d'une tumeur, d'une lignée cellulaire ou de sang.
12. Procédé pour obtenir la séquence d'acides nucléiques des différents fragments d'acides nucléiques génomiques chimériques de la composition selon l'une quelconque des revendications 1 à 5, ou produite par la méthode selon les revendications 6 à 11, comprenant les étapes consistant à (i) obtenir les fragments et à (ii) séquencer les fragments, en utilisant idéalement une plateforme de séquençage de nouvelle génération, de façon à obtenir la séquence d'acides nucléiques des différents fragments d'acides nucléiques génomiques chimériques.
13. Procédé pour obtenir des informations de nombre de copies génomiques à partir d'un génome, comprenant les étapes consistant à
- i) obtenir la séquence d'acides nucléiques des différents fragments d'acides nucléiques génomiques chimériques de (a) une composition de bibliothèque de séquençage comprenant un premier mélange de différents fragments d'acides nucléiques génomiques chimériques, dans lequel le mélange de différents fragments d'acides nucléiques génomiques chimériques contient au moins 100 000 fragments différents, dans lequel les fragments d'acides nucléiques génomiques chimériques ont une longueur de 250 ($\pm 10\%$) à moins de 1 000 ($\pm 10\%$) paires de bases, dans lequel chaque fragment différent du mélange comprend des segments d'ADN ligaturés de manière aléatoire, dans lequel chaque segment d'ADN du fragment est une molécule d'acide nucléique d'une longueur d'au moins 27 paires de bases obtenue à la suite d'une fragmentation aléatoire d'un seul génome pour lequel un génome de référence est disponible, dans lequel au moins 50 % des segments des au moins 100 000 fragments différents ont une longueur de 30 à 50 ($\pm 10\%$) paires de bases, (b) la composition selon l'une quelconque des revendications 1 à 5, ou (c) la composition produite par la méthode selon les revendications 6 à 11 ;
- ii) identifier et cartographier, pour un génome, chaque correspondance maximale quasi-unique (MAM, pour *Maximal Almost-unique Match*) à l'intérieur d'un fragment d'acide nucléique génomique chimérique séquencé, les MAM préférables étant identifiées au moyen d'un progiciel longMEM, les MAM étant en outre de préférence

filtrées par une élimination de MAM inférieures à vingt paires de bases et de pas plus d'une longueur d'au moins quatre paires de bases de plus que nécessaire pour l'unicité, et/ou les MAM étant filtrées par une élimination mutuelle de MAM dans une carte de paires de lecture qui se trouvent dans 10 000 paires de bases ; et
iii) compter le nombre de MAM cartographiées à l'intérieur d'un génome compartimenté, ce qui permet d'obtenir des informations de nombre de copies génomiques.

5

14. Procédé selon la revendication 13, dans lequel, à l'étape (iii), le nombre de lectures cartographiées est compté dans des tailles de compartiments génomiques qui donnent des comptes de carte uniformes pour l'échantillon de référence, dans lequel, à l'étape (iii), le nombre de lectures cartographiées est compté dans des compartiments génomiques d'observation uniforme d'une référence, déterminés de manière empirique, dans lequel, à l'étape (iii), le nombre de lectures cartographiées est compté dans des compartiments génomiques de densité uniforme souhaitée, dans lequel, à l'étape (iii), le nombre de lectures cartographiées dans chaque compartiment est ajusté par rapport à un biais GC par une normalisation LOESS, dans lequel, à l'étape (iii), une analyse de modèle est utilisée pour réduire le bruit systématique dans des données de comptes de compartiments à GC ajusté, dans lequel, à l'étape (iii), une normalisation de référence est appliquée à des données de comptes de compartiments par une division de taux de compartiments à GC ajusté par un taux de compartiment d'échantillon standard, dans lequel, à l'étape (iii), des données de comptes de compartiments à GC ajusté normalisés de référence sont analysées par une segmentation binaire circulaire, et/ou dans lequel, à l'étape (iii), le nombre total de cartes de référence est mis en correspondance avec le nombre total de cartes d'échantillons.

10

15

20

15. Méthode pour diagnostiquer, prédire un risque de présenter, ou déterminer la probabilité d'hériter d'un trouble prénatal, un trouble pédiatrique, un trouble développemental, un trouble psychologique, un trouble auto-immun, un cancer, une maladie cardiaque congénitale, la schizophrénie, des troubles du spectre autistique ou une réponse d'un patient à une thérapie, consistant à obtenir les informations de nombre de copies génomiques d'un patient par le procédé selon la revendication 13 ou 14.

25

30

35

40

45

50

55

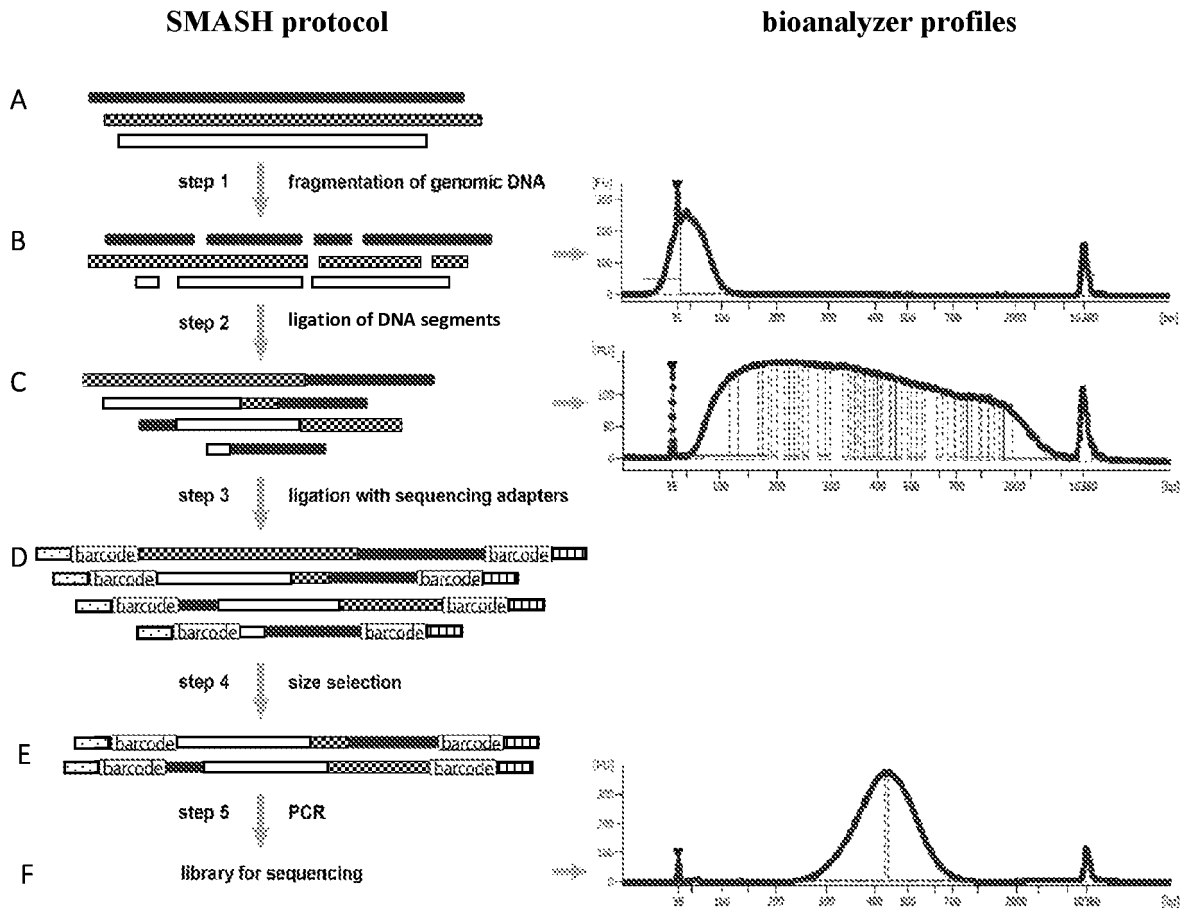


Figure 1

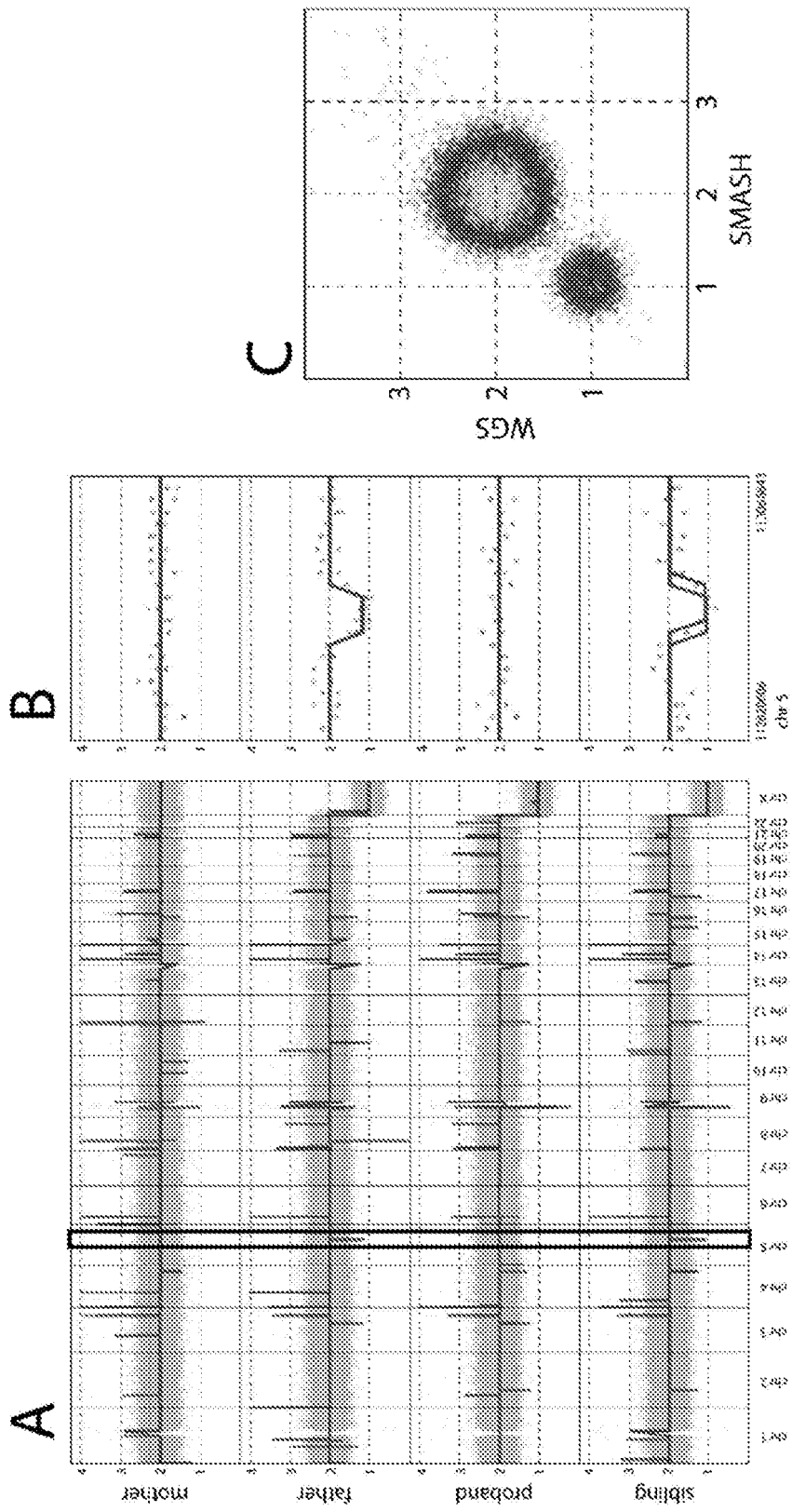


Figure 3

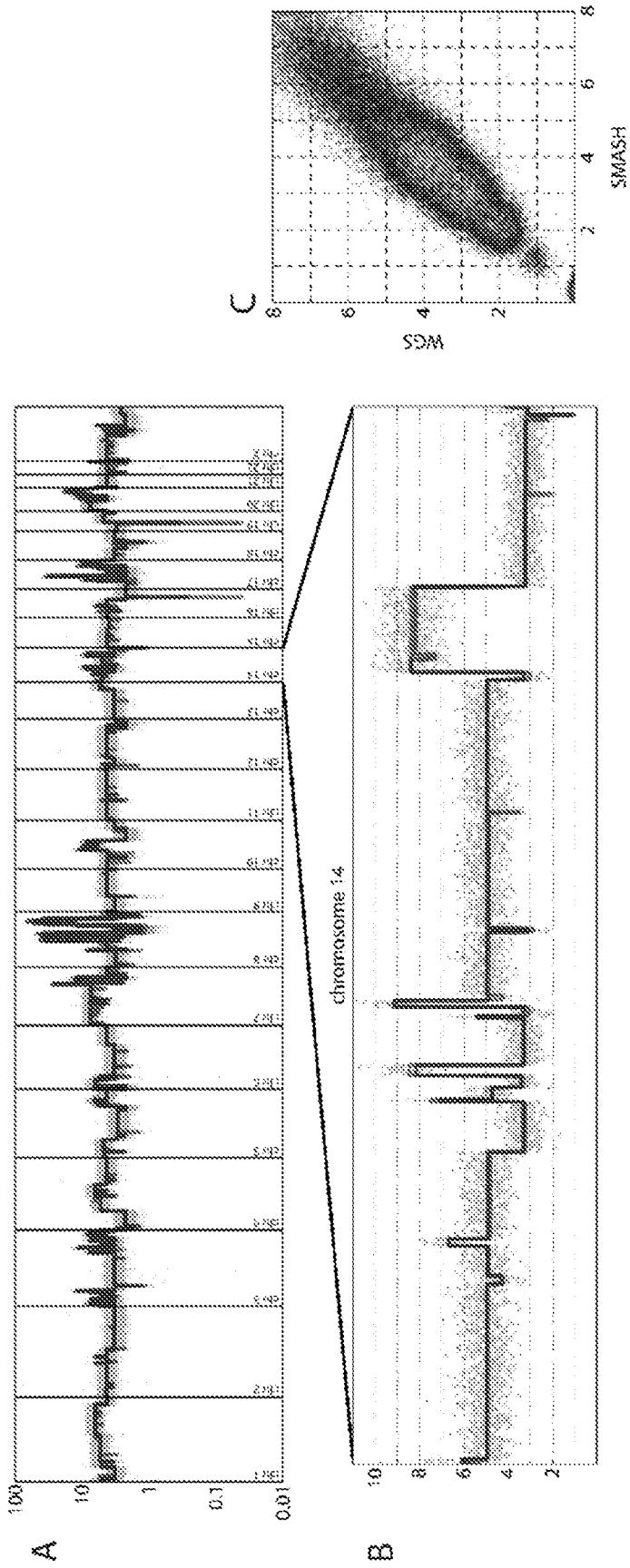


Figure 4

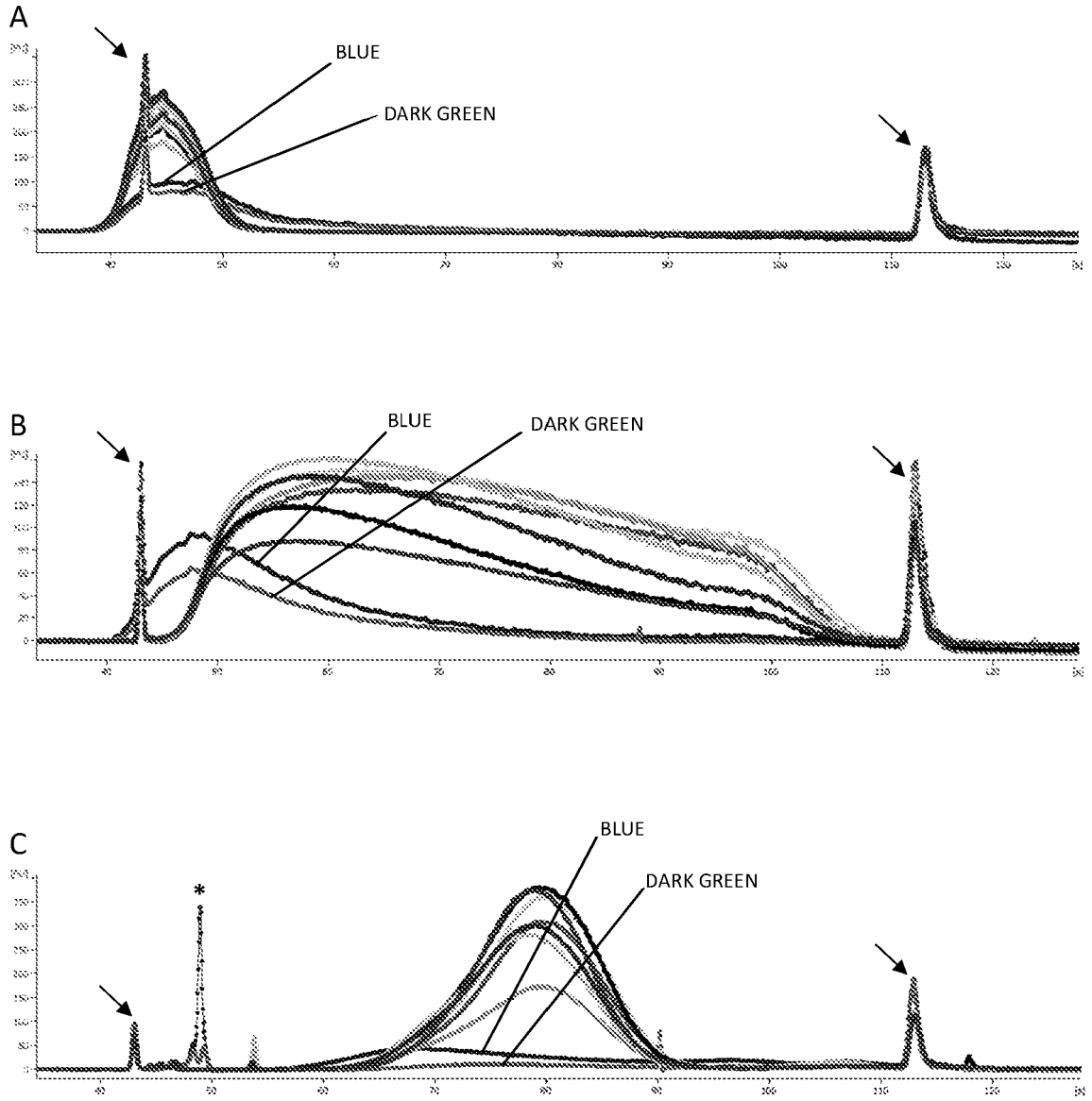


Figure 5

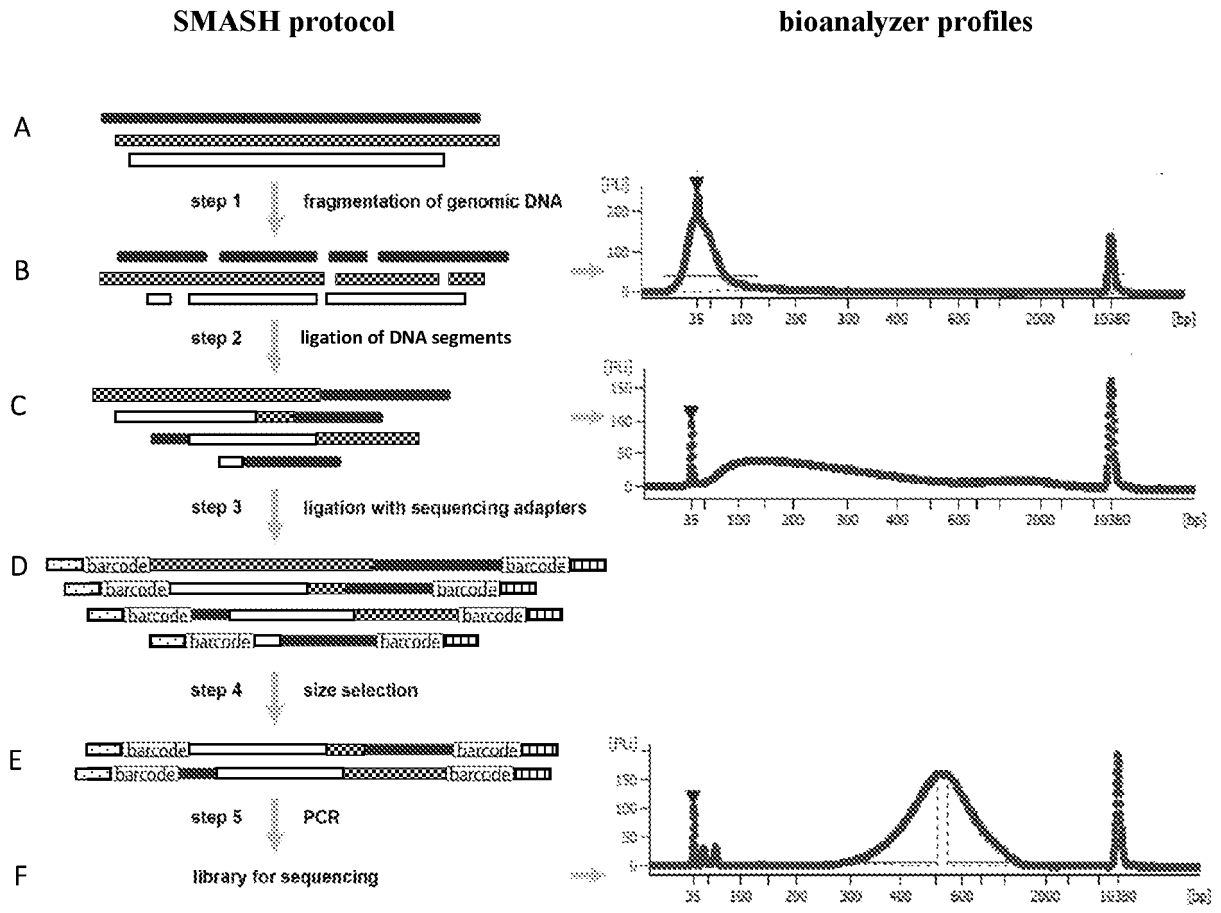


Figure 6

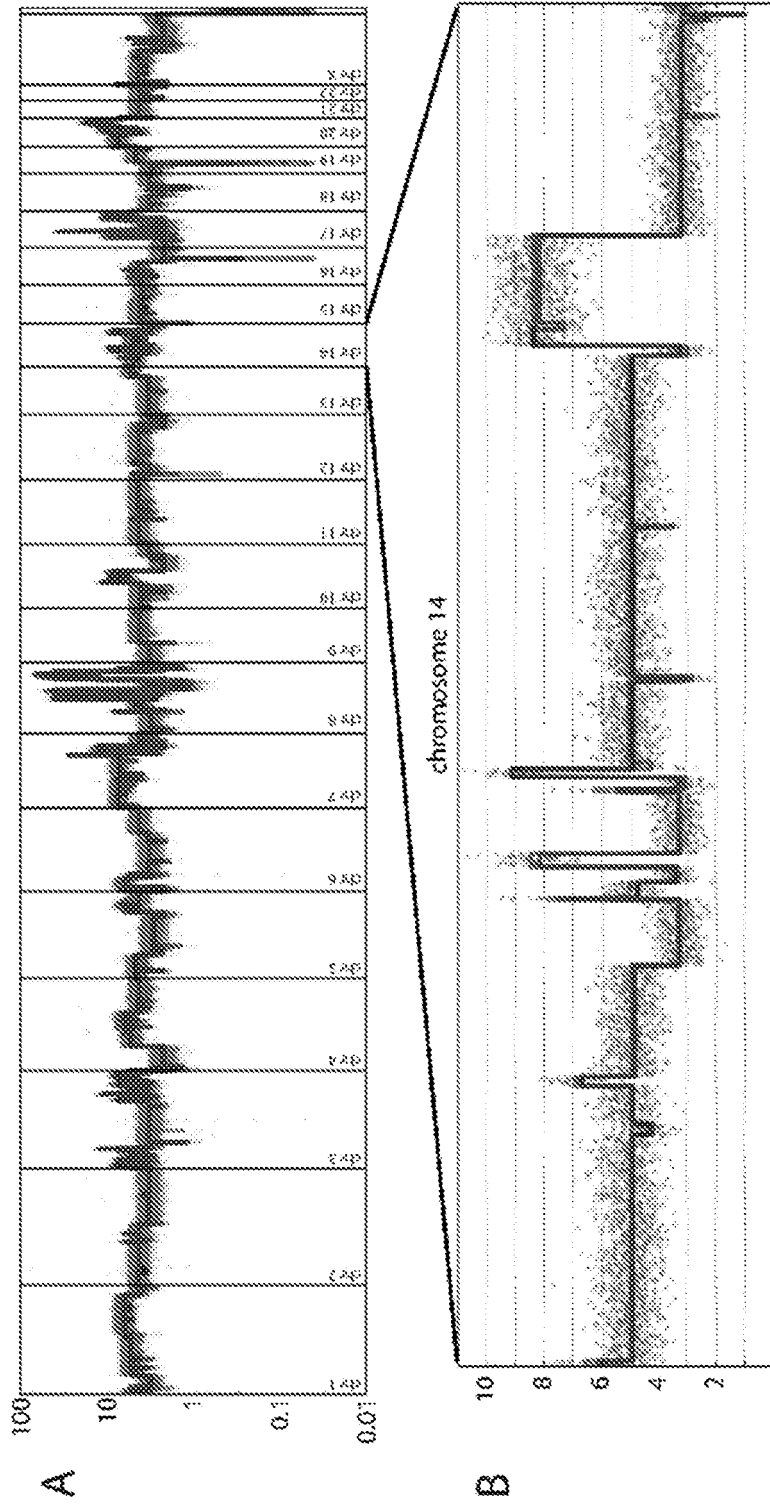


Figure 7

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Patent documents cited in the description

- WO 2014149134 A2 [0004]
- WO 2012054873 A2 [0005]
- US 2005260655 A1 [0006]

Non-patent literature cited in the description

- **ALKAN C ; KIDD JM ; MARQUES-BONET T ; AK-SAY G ; ANTONACCI F ; HORMOZDIARI F ; KITZMAN JO ; BAKER C ; MALIG M ; MUTLU O.** Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature genetics.*, 2009, vol. 41 (10), 1061-7 [0141]
- **FISHBACH GD ; LORD C.** The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron.*, 2010, vol. 68, 192-195 [0141]
- **KHAN Z ; BLOOM JS ; KRUGLYAK L ; SINGH M.** A practical algorithm for finding maximal exact matches in large sequence datasets using sparse suffix arrays. *Bioinformatics.*, 2009, vol. 25 (13), 1609-16 [0141]
- **LEVY D ; WIGLER M.** Facilitated sequence counting and assembly by template mutagenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 2014, vol. 111 (43), E4632-7 [0141]
- **LOCKWOOD WW ; COE BP ; WILLIAMS AC ; MACAULAY C ; LAM WL.** Whole genome tiling path array CGH analysis of segmental copy number alterations in cervical cancer cell lines. *International journal of cancer Journal international du cancer.*, 2007, vol. 120 (2), 436-43 [0141]
- **LU TP ; LAI LC ; TSAI MH ; CHEN PC ; HSU CP ; LEE JM ; HSIAO CK ; CHUANG EY.** Integrated analyses of copy number variations and gene expression in lung adenocarcinoma. *PloS one.*, 2011, vol. 6 (9), e24829 [0141]
- **MUHLE R ; TRENTACOSTE SV ; RAPIN I.** The genetics of autism. *Pediatrics.*, 2004, vol. 113 (5), e472-86 [0141]
- **NAVIN N ; KENDALL J ; TROGE J ; ANDREWS P ; RODGERS L ; MCINDOO J ; COOK K ; STEPANSKY A ; LEVY D ; ESPOSITO D.** Tumour evolution inferred by single-cell sequencing. *Nature.*, 2011, vol. 472 (7341), 90-4 [0141]
- **OLSHEN AB ; VENKATRAMAN ES ; LUCITO R ; WIGLER M.** Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics.*, 2004, vol. 5, 557-572 [0141]
- **REES E ; WALTERS JT ; GEORGIEVA L ; ISLES AR ; CHAMBERT KD ; RICHARDS AL ; MAHONEY-DAVIES G ; LEGGE SE ; MORAN JL ; MCCARROLL SA.** Analysis of copy number variations at 15 schizophrenia-associated loci. *The British journal of psychiatry : the journal of mental science.*, 2014, vol. 204 (2), 108-14 [0141]
- **ROSENBERG RE ; LAW JK ; YENOKYAN G ; MCGREADY J ; KAUFMANN WE ; LAW PA.** Characteristics and concordance of autism spectrum disorders among 277 twin pairs. *Archives of pediatrics & adolescent medicine.*, 2009, vol. 163 (10), 907-14 [0141]
- **SHLIEN A ; MALKIN D.** Copy number variations and cancer. *Genome Medicine.*, 2009, vol. 1 (6), 62 [0141]
- **STADLER ZK ; ESPOSITO D ; SHAH S ; VIJAI J ; YAMROM B ; LEVY D ; LEE YH ; KENDALL J ; LETOTTA A ; RONEMUS M.** Rare de novo germline copy-number variation in testicular cancer. *American journal of human genetics.*, 2012, vol. 91 (2), 379-83 [0141]
- **SUDMANT PH ; KITZMAN JO ; ANTONACCI F ; ALKAN C ; MALIG M ; TSALENKO A ; SAMPASN ; BRUHN L ; SHENDURE J ; GENOMES P.** Diversity of human copy number variation and multicopy genes. *Science*, 2010, vol. 330 (6004), 641-6 [0141]
- **SZATKIEWICZ JP ; O'DUSHLAINE C ; CHEN G ; CHAMBERT K ; MORAN JL ; NEALE BM ; FROMER M ; RUDERFER D ; AKTERIN S ; BERGEN SE.** Copy number variation in schizophrenia in Sweden. *Molecular psychiatry.*, 2014, vol. 19 (7), 762-73 [0141]
- **WARBURTON D ; RONEMUS M ; KLINE J ; JOBANPUTRA V ; WILLIAMS I ; ANYANE-YEBOA K ; CHUNG W ; YU L ; WONG N ; AWAD D.** The contribution of de novo and rare inherited copy number changes to congenital heart disease in an unselected sample of children with conotruncal defects or hypoplastic left heart disease. *Human genetics.*, 2014, vol. 133 (1), 11-27 [0141]

EP 3 347 466 B9

- **WILLYARD C.** Copy number variations' effect on drug response still overlooked. *Nature medicine.*, 2015, vol. 21 (3), 206 [0141]
- **WEISCHENFELDT J. et al.** Phenotypic impact of genomic structural variation: insights from and for human disease. *Nature Reviews Genetics.*, 2013, vol. 14 (2), 125-138 [0141]
- **MALHOTRA D. et al.** CNVs: Harbingers of a rare variant revolution in psychiatric genetics. *Cell*, 2012, vol. 148 (6), 1223-1241 [0141]
- **ANSORGE et al.** Next-generation DNA sequencing techniques. *New Biotechnology*, 2009, vol. 25 (4), 195-203 [0141]
- **WANG Z. et al.** SMASH, a fragmentation and sequencing method for genomic copy number analysis. *Genome Research*, 2016, vol. 26 (6), 844-851 [0141]