

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2014 年 3 月 20 日 (20.03.2014)



(10) 国际公布号
WO 2014/040206 A1

- (51) 国际专利分类号:
C12Q 1/68 (2006.01) *G06F 19/22* (2011.01)
- (21) 国际申请号: PCT/CN2012/001261
- (22) 国际申请日: 2012 年 9 月 12 日 (12.09.2012)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (71) 申请人 (对除美国外的所有指定国): **深圳华大基因研究院 (BGI SHENZHEN)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **深圳华大基因科技有限公司 (BGI SHENZHEN CO., LIMITED)** [CN/CN]; 中国广东省深圳市盐田区北山路 146 号北山工业区综合楼 11F-3, Guangdong 518083 (CN)。
- (72) 发明人: 及
- (75) 发明人/申请人 (仅对美国): **张帆 (ZHANG, Fan)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **罗锐邦 (LUO, Ruibang)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **李娜 (LI, Na)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **李英睿 (LI, Yingrui)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **王俊 (WANG, Jun)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **汪建 (WANG, Jian)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **杨焕明**

(YANG, Huanming) [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。

(74) 代理人: **北京北翔知识产权代理有限公司 (PEK-SUNG INTELLECTUAL PROPERTY LTD.)**; 中国北京市海淀区学院路 35 号世宁大厦 908 室, Beijing 100191 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

(54) Title: METHOD FOR DETECTING COPY NUMBER VARIATIONS BY GENOME SEQUENCING FRAGMENTS

(54) 发明名称: 利用基因组测序片段检测拷贝数变异的方法

(57) Abstract: The invention provides a method for detecting copy number variations. The method arrives at an assessment of copy number variations by sequencing a target sample and analyzing the depth of coverage of sequencing fragments.

(57) 摘要: 本发明提供了一种检测拷贝数变异的方法, 该方法通过对目标样本进行测序、并对测序片段的覆盖深度进行分析来实现对拷贝数变异的评估。



WO 2014/040206 A1

利用基因组测序片段检测拷贝数变异的方法

技术领域

本发明涉及基因组序列分析技术领域，更具体而言涉及利用基因组
5 测序片段检测拷贝数变异的方法。

背景技术

拷贝数变异 (Copy Number Variations, CNV) 是指基因组上长度
超过 1 kb 的 DNA 片段，与参考基因组相比存在拷贝数的差异。这种差
异可以表现为拷贝数的增加——包括插入和重复，或者表现为拷贝数
10 的减少——包括缺失和零基因型 (null genotype)。CNV 在人类基因
组中的分布普遍，占基因组中超过 10% 的序列。

目前 CNV 检测主要采用比较基因组杂交 (comparative genomic
hybridization) 技术，该技术通过将试验样品和参照样品基因组 DNA
同时与微阵列芯片上 DNA 探针杂交，直观地得到试验样品中基因组
15 DNA 发生变异的位点信息及拷贝数量变化信息，该技术成本高，分辨
率低，该技术对于 10-25 kb 的 CNV 灵敏度低。另外，用于 CNV 检测的技
术还有以下几种：荧光定量 PCR 技术，其不足之处是一个反应只能测
定一种 CNV，需要进行多次重复；荧光原位杂交，其缺点是探针不稳
定、操作繁琐且不能 100% 的杂交；直接测序，该技术可以检测插入、
20 重排、断点 (breakpoint)，但其缺点是效率低，覆盖度小；多重连接
探针扩增技术，该技术可以在一个反应中同时测定多个 CNV，但其缺
点是覆盖范围小，CNV 本身大小有局限。利用这些技术检测 CNV 还有
一个共同的缺点是成本都较高。

目前基于高通量测序结果的 CNV 检测方法主要是基于双末端测序比
25 对 (paired-end read mapping, PEM) 结果。但是 PEM 的局限性在于，
多种类型的 CNV——包括复杂基因组区域中的大片段的插入和变异——
难以被检测到，且对于超过平均插入文库的插入检测有局限性。

发明内容

为了克服上述 CNV 检测灵敏度不高、检测长度限制、操作繁琐、成本高等局限性，本发明提供了检测拷贝数变异的方法。该方法通过分析基因组序列及其相对应位置的测序片段的覆盖深度，对局部拷贝数进行

5 评估。

这种基于测序片段的覆盖深度检测拷贝数变异的方法包括步骤：

1) 对目标样本进行测序，得到测序片段；

2) 将以上得到的测序片段与参考基因组序列比对，优选去除上述
10 比对结果中的重复和冗余，得到参考序列每个位点的覆盖深度，即比
对结果中涵盖该位点的测序片段数目；

3) 根据以上每个位点的覆盖深度，将所有位点的覆盖深度取平均
得出序列的平均覆盖深度，以同样的方式计算参考序列上具有预设片
段长度的所有窗口的覆盖深度，这些窗口被定义为以下四种类型之一：

- 15 a. 正常窗口：覆盖深度与序列平均覆盖深度相同的窗口，
b. 重复窗口：覆盖深度明显大于平均覆盖深度的窗口，
c. 缺失窗口：覆盖深度明显小于平均覆盖深度的窗口，
d. N 窗口：基本没有覆盖深度的窗口；

4) 将满足以下条件任一项的三以上连续窗口合并，并判断合并后
的区域是否属于缺失和重复这两种 CNV：

20 i. 连续重复窗口或连续缺失窗口、

ii. 缺失窗口之间有 N 窗口间隔，例如缺失窗口+N 窗口+缺失窗
口，其中 N 窗口不能连续出现一个以上、

iii. 重复窗口之间有 N 窗口间隔，例如重复窗口+N 窗口+重复窗
口，其中 N 窗口不能连续出现一个以上、

25 iv. 缺失窗口之间有正常窗口间隔，例如缺失窗口+正常窗口+缺失
窗口，其中正常窗口的覆盖深度应满足减去 3 倍 SD 后落在缺失窗口的
覆盖深度范围内，并且其中正常窗口不能连续出现一个以上、

v. 重复窗口之间有正常窗口间隔，例如重复窗口+正常窗口+重复
窗口，其中正常窗口的覆盖深度应满足加上 3 倍 SD 后落在重复窗口覆
30 盖深度范围内，并且其中正常窗口不能连续出现一个以上、

其中，SD 为所有位点的覆盖深度取平均得出序列的平均覆盖深度

标准差；

5) 对以上判断属于缺失和重复这两种 CNV 的合并区域降低假阳性，所得到的区域就是发生了插入或缺失的 CNV 区域。

在上述步骤 1) 和 2) 之间，优选包括步骤 1')：评估测序结果是否合格，如果不合格则需要重新测序，并且如果在测序过程中引入接头序列，则去除这些接头序列。

本发明的方法可以灵敏地检测到落在那些富含结构变异的复杂区域中的 CNV，同时具有无检测长度限制、操作简单，以及从而带来的成本低等优点，这些有点是本领域中其他技术达不到的。

10 具体实施方式

下面对本发明进行更全面的描述，详细描述了本发明的示例性实施例。

本发明的目的是提供检测拷贝数变异的方法，旨在利用测序数据，通过生物信息学方法检测目标样本与参照基因组间发生拷贝数变异的区域。

本发明检测拷贝数变异的方法可以包括以下步骤：

第一步，在对目标样品进行测序，并优选评估测序结果是否合格，如果不合格则需要重新测序；如果测序中使用了接头，还要去除引入的接头。

20 测序方法很多，可提供本发明的测序数据的测序法例如有 454 测序、Illumina 测序。测序片段长度一般是 90 bp 或 100 bp，例如 Illumina 测序法得到测序片段的长度一般是 90 bp，采用双末端测序。本发明方法使用的测序片段的长度可以是 100 bp，优选是 90 bp。

25 对于本发明而言，测序深度可以是 10×，优选是 20×，最优选是 30× 以上。例如，可以使用 35× 的测序深度。

本领域已知对测序片段进行评估的方法，例如，对测序结果的评估可以主要包括以下两方面：互补碱基含量比例是否均衡，例如 G/C 碱基即 G 碱基的比例和 C 碱基比例是否接近等，例如本领域中通常使用均值上下 3 倍的 G/C 比，如果差异落在该范围外则说明测序结果不合格；碱基的质量和 N（测序结果中碱基不确定）的含量，如果低质量的碱基，说明测序结果不合格。

第二步，将以上得到的测序片段与参考基因组序列比对，优选去除上述比对结果中的重复和冗余，统计位点的序列信息和覆盖深度信息，即比对结果中涵盖该位点的测序片段数目。

本文中“重复”在用于比对结果的情况下时是指测序片段本来只应该测一次，结果由于 PCR 的原因，被测了多次，是多个测序片段表现出完全一致的基因内容。

本文中“冗余”在用于比对结果的情况下时是指在测序过程中，实验技术需求添加在真实测序片段两端的人工冗余序列。

参照序列通常选取其序列已被确定的序列，例如可以来自公共数据库，或者可以来自商业数据库。例如，对于人的样品来说，参照序列可以是人基因组 hg18 或 hg19 的序列。目前 hg19 的相关数据库相对较多且 hg19 测出来的碱基量比 hg18 要多，即样品比对率会相对较高，所以优先选择 hg19。

位点的序列信息是比对结果中包含该位点的测序片段，位点的覆盖深度信息是比对结果中包含该位点的测序片段数目。

序列比对可以通过任何一种序列比对程序，例如短寡核苷酸分析包（Short Oligonucleotide Analysis Package, SOAP）和 BWA 比对（Burrows-Wheeler Aligner）进行，将测序片段与参考基因组序列比对，得到测序片段在参考基因组上的位置。进行序列比对可以使用程序提供的默认参数进行，或者由本领域技术人员根据需要对参数进行选择。

另外，还可以对比对结果进行筛选，例如去除比对结果落在多个位置的序列，因为这些序列无法提供唯一的比对位置；去除重复出现的序列，因为这些序列可能是由于前期实验的误差引入，例如由测序错误引起，去除这种序列可使检测结果更加精准。

在比对后，可以以本领域中已知的任何方法计算位点的覆盖深度，计算位点的覆盖深度是基于覆盖该位点的测序后的序列数目。例如，利用短寡核苷酸分析包中的覆盖深度计算程序（SOAP coverage）计算参考基因组每个位点的覆盖深度。

第三步，根据以上每个位点的覆盖深度，将所有位点的覆盖深度取平均得出序列的平均覆盖深度，以同样方式计算参考序列的具有预设片段长度的所有窗口的覆盖深度，这些窗口被定义为以下四种类型

之一:

- a. 正常窗口: 覆盖深度与序列平均深度相同的窗口
- b. 重复窗口: 覆盖深度明显大于平均深度的窗口
- c. 缺失窗口: 覆盖深度明显小于平均深度的窗口
- 5 d. N窗口: 基本没有覆盖深度的窗口,

在本发明中, 窗口可以为 70 bp-100 bp、100 bp、100 bp-200 bp 或 50 bp-300 bp, 优选 50 bp-150 bp, 最优选约 100 bp。大窗口(比如 1000 bp)不能提供 CNV 断点的精确位置, 并且大窗口不能准确的检测短片
10 段 CNV。对于一般使用 30×的样品测序量, 100 bp 窗口测序片段的数目的分布很接近正态分布, 这样使可以假设数据计算也是符合正态分布的。小窗口大小的测序片段分布却不符合正态分布。

在本发明中, 窗口的测序片段覆盖深度的计算是统计每个位点覆盖的测序片段数之和除以该窗口的位点数。

在本发明中, 覆盖深度与序列平均覆盖深度相同的窗口, 是指二者基本相同的窗口。例如, 在一些实施方案中, 这样的两个覆盖深度在统计学上差异不显著。或者可以依据其他方式定义, 例如, 在一些
15 实施方案中, 这样的两个覆盖深度的差异在 1 倍以内, 75%以内, 50%以内, 或 20%以内, 例如 10%以内或 5%以内。

在本发明中, 覆盖深度明显大于平均覆盖深度是指前者是后者的至少 1.2 倍, 至少 1.5 倍, 至少 2 倍、至少 4 倍或至少 8 倍。
20

在本发明中, 覆盖深度明显小于平均覆盖深度是指后者是前者的至少 1.2 倍, 至少 1.5 倍, 至少 2 倍、至少 4 倍或至少 8 倍。

在本发明中, N窗口是基本没有覆盖深度的窗口, 优选是没有覆盖深度的窗口。基本没有覆盖深度, 是指覆盖深度是平均深度的小于
25 50%, 小于 20%, 小于 10%, 小于 5%或小于 2%。特别地, 对于双末端测序, 组装出来的重叠群序列之间的相对位置是可以确定的, 但是具体序列未知, 可以把两条已知相对位置的重叠群序列连在一起, 中间未知的碱基用 N 表示, 在这个区域上的窗口即为 N窗口;

在一些实施方案中, 正常窗口可以被定义变异窗口(重复窗口和
30 缺失窗口)和 N窗口之外的窗口。

在本发明中, 优选通过设置上述标准, 将所有窗口分配到上述四种窗口类型之一并且仅之一。

在一个实施方案中，该步骤具体如下：为了评估覆盖深度，以预设片段长度例如 100 bp 为窗口在参考序列上滑行取窗口，统计与这些窗口的比对上的测序片段数目。窗口的覆盖深度的计算是统计每个位点覆盖的测序片段数之和除以该窗口的位点数，例如 100。

- 5 另外，对于 Illumina 的测序片段而言，测序覆盖深度会受到 GC 含量的影响，所以优选根据在具体 GC 比下观察到的覆盖深度的偏差来调整窗口的测序片段数目。例如，调整公式是 $\sim ri = ri * m / m_{GC}$ ，其中 $\sim ri$ 是公式校正后的测序片段数目， ri 是第 i 个窗口的测序片段数目， m_{GC} 是与第 i 个窗口有相同 G+C 比例的所有窗口的测序片段数目的中值，
10 m 是所有窗口的总体中值。对于 Illumina 的测序片段而言，后续分析可以基于这些 GC 调整后的数据进行。

第四步，将满足以下条件任一项的三以上连续窗口合并，并判断合并后的区域是否属于缺失和重复这两种 CNV：

- 15 i. 连续重复窗口或连续缺失窗口、
 ii. 缺失窗口之间有 N 窗口间隔，例如缺失窗口 + N 窗口 + 缺失窗口，其中 N 窗口不能连续出现一个以上、
 iii. 重复窗口之间有 N 窗口间隔，例如重复窗口 + N 窗口 + 重复窗口，其中 N 窗口不能连续出现一个以上、
20 iv. 缺失窗口之间有正常窗口间隔，例如缺失窗口 + 正常窗口 + 缺失窗口，其中正常窗口的覆盖深度应满足减去 3 倍 SD 后落在缺失窗口的覆盖深度范围内，并且其中正常窗口不能连续出现一个以上、
 v. 重复窗口之间有正常窗口间隔，例如重复窗口 + 正常窗口 + 重复窗口，其中正常窗口的覆盖深度应满足加上 3 倍 SD 后落在重复窗口覆盖深度范围内，并且其中正常窗口不能连续出现一个以上；
25 SD 为所有位点的覆盖深度取平均得出序列的平均覆盖深度标准差。

- 30 对于合并后的窗口是否属于缺失和重复这两种 CNV，可以利用本领域中已有的任何方法进行检测。例如，可以利用 Event-wise testing 算法（Seungtae Yoon, Zhenyu Xuan, Vladimir Makarov, Kenny Ye, Jonathan Sebat, Sensitive and accurate detection of copy number variants

using read depth of coverage. Genome Res. 2009 September; 19(9):

1586–1592.) 进行 CNV 检测。Event-wise testing 是一种基于测序片段数据的新 CNV 检测算法，它主要基于显著性检验。在该算法中，拷贝数的增加和减少可以通过窗口覆盖深度的增加和减少反映出来。利用该算法可以找出测序片段显著增加或减少的预设片段长度例如 100 bp 的窗口。

在一个具体实施方案中，利用以下方式窗口是否属于缺失和重复这两种 CNV。

首先，将第 i 个窗口的测序片段数目转换成 Z 值 (Z_i , i 是自然数)，转换方法是以起始位点落入该窗口的测序片段数目减去所有窗口的测序片段数目平均值（即所有测序片段总数和除以窗口数目），将相减的结果除以窗口测序片段数的标准差，得到 Z 值。再将 Z 值进一步转换成可能性上限值 $piUpper = P(Z > z_i)$ 和可能性下限值 $piLower = P(Z < z_i)$ ，即概率的上限和下限， $P(Z > z_i)$ 是 Z 大于 Z_i 的概率和， $P(Z < z_i)$ 是 Z 小于 Z_i 的概率和。对于有 l 个连续窗口的区域 A ，如果它符合 $\max\{piUpper | i \in A\} < (FPR/(L/l))^{(1/l)}$ ，则认为它发生了重复变异；如果它符合 $\max\{piLower | i \in A\} < (FPR/(L/l))^{(1/l)}$ ，则认为它发生了缺失，其中 FPR （错误率）是指对于参考基因组的重复或缺失的假阳性值之和， I 型错误是 FPR （假阳性率）——即检测出的 cnv 里面不是真实的 CNV 占检测出的所有 CNV 的比率， II 型错误是 FNR （假阴性率）——即没有被检测出的真实 cnv 占真实 cnv 总数的比率； L 是指参考基因组的总窗口数目， l 是指 A 区域的窗口数目，大于等于 1 且小于等于 L 。

很明显，如果 A 的所有探针是来自于正常状态，即 A 是 CNV 的可能性小于 $FPR/(L/l)$ ，该窗口被定义为正常窗口。

这里需要注意的是，没有以 FPR 除以所有的窗口数目减 1 ($L-1$)，而是除以 l ，因为前者在控制 I 型错误上面过于保守，这里考虑了重叠窗口的数据检验是非独立的。

另外，对于缺失和重复这两种 CNV 可以分开进行检测。对于重复和缺失的检测均用上述公式，但前者的阈值是前述 $piUpper$ ，后者的是 $piLower$ 。对于它们每一种的检测，首先寻找两个存在 CNV 的窗口，然后每增加 1 个窗口就重复增加后窗口是否也存在该 CNV。要注意的

是, $FPR/(L/l) \wedge (1/l)$ 的阈值是随着 l 的增加而增加的。当 $FPR/(L/l) \wedge (1/l)$ 超过 0.5 时在 $N-1$ 处停止重复上述步骤。

第五步, 对以上判断合属于缺失和重复这两种 CNV 的合并区域降低假阳性, 所得到的区域就是发生了插入或缺失的 CNV 区域。

降低假阳性可以通过一定的过滤条件实现。例如, 将那些测序片段覆盖深度中值在平均覆盖深度值 0.75 倍和 1.25 倍之间的变异区域过滤掉。

另外, 还任选通过 Z 值检验来检验每个变异区域的显著性, 可以用显著性水平 10^{-6} 作为阈值来过滤合并区域。其中, 显著性水平 10^{-6} 和用 $FPR0.05$ 来检测 2 个变异区域的显著水平是接近相符的; 另外, 根据很多 CNV 在所有显著水平上的人工期望, 10^{-6} 阈值被认定是合理的。

15 实施例

下面结合实施例, 对本发明进行详细说明。但应理解, 以下实施例仅是对本发明实施方式的举例说明, 而非是对本发明的范围限定。

在本实施例中, 利用本发明的方法检测拷贝数变异计划中样品 NA19238 的测序数据的拷贝数变异。拷贝数变异计划是由 sanger 研究院发起的, 旨在研究拷贝数变异对人类健康的影响。本实施例分析对象是该项目其中一个样品 NA19238 (尼日利亚约鲁巴人), 其 CNV 已知结果通过针对全基因组的阵列比较基因组杂交获得。(数据可获自 ftp://ftp.sanger.ac.uk.pub.cnv_project/)

第一步, 对 NA19238 数据下载得到已经完成比对的高深度测序数据, 比对结果格式为 sam 格式(参见

<http://samtools.sourceforge.net/SAM1.pdf>), 包括 24 条染色体的比对结果 (22 条常染色体和 XY 染色体)。数据比对采用比对软件 BWA (<http://bio-bwa.sourceforge.net/bwa.shtml>), 比对使用的参考基因组为 hg19 (下载地址:

30 <http://hgdownload.cse.ucsc.edu/downloads.html#human>)。

第二步, 本步骤利用 SOAP 程序完成, 详细操作步骤依据该程序的操作说明 (Short Oligonucleotide Analysis Package,

<http://soap.genomics.org.cn>)

在序列比对结果的基础上，对参考基因组上每个位点的被覆盖的深度情况进行统计，并存放到覆盖深度文件中。覆盖深度统计使用 SOAPcoverage 程序 (Version: 2.7.7, 下载地址:

- 5 <http://soap.genomics.org.cn/down/soap.coverage.tar.gz>), 得到每个位点的覆盖深度情况。

第三步，用本发明的方法对该样品的拷贝数变异进行检测。需要第二步获得的覆盖深度文件和参考基因组 hg19 的序列文件作为输入文件，运行后得到的各窗口拷贝数情况以及是否为重复或缺失的拷贝数变异发生窗口的结果存储文件，文件结果每列说明如下。

10

列	描述
1	序列 ID 序列长度 序列覆盖情况
2	事件
3	起始窗口
4	窗口数
5	局部平均覆盖度
6	局部覆盖度中数
7	局部拷贝数
8	全局拷贝数
9	局部方差
10	合并事件数

程序实现原理是将参考基因组的序列文件和位点覆盖深度文件分别存为哈希序列表和哈希覆盖深度表。

- 15 统计对应窗口的覆盖深度和显著性。对局部窗口的覆盖深度进行变异检测，找出窗口平均覆盖深度与全基因组平均覆盖深度水平明显差异的窗口。

- 20 该步骤具体如下：为了评估覆盖深度，以预设片段长度 100 bp 为窗口在参考序列上滑行取窗口，统计与这些窗口的比对上的测序片段数目。窗口的覆盖深度的计算是统计每个位点覆盖的测序片段数之和除以该窗口的位点数。

第四步，对于连续的相同变异的或是满足以下合并条件的窗口，进行窗口合并，最终将片段信息汇总打印：对于 CNV，我们定义 4 种状态：缺失窗口、重复窗口、正常窗口和 N 窗口。以每三个相邻窗口考察合并情况：连续重复窗口或连续缺失窗口；缺失窗口之间有 N 窗口间隔，例如缺失窗口+N 窗口+缺失窗口，其中 N 窗口不能连续出现一个以上；重复窗口之间有 N 窗口间隔，例如重复窗口+N 窗口+重复窗口，其中 N 窗口不能连续出现一个以上；缺失窗口之间有正常窗口间隔，例如缺失窗口+正常窗口+缺失窗口，其中正常窗口的覆盖深度应满足减去 3 倍 SD 后落在缺失窗口的覆盖深度范围内，并且其中正常窗口不能连续出现一个以上；重复窗口之间有正常窗口间隔，例如重复窗口+正常窗口+重复窗口，其中正常窗口的覆盖深度应满足加上 3 倍 SD 后落在重复窗口覆盖深度范围内，并且其中正常窗口不能连续出现一个以上。

然后，对于合并后的窗口是否属于缺失区域和重复区域这两种 CNV。

首先，将第 i 个窗口的测序片段数目转换成 Z 值 (Z_i , i 是自然数)，转换方法是以起始位点落入该窗口的测序片段数目减去所有窗口的测序片段数目平均值（即所有测序片段总数和除以窗口数目），将相减的结果除以窗口测序片段数的标准差，得到 Z 值。再将 Z 值进一步转换成可能性上限值 $piUpper = P(Z > z_i)$ 和可能性下限值 $piLower = P(Z < z_i)$ ，即概率的上限和下限， $P(Z > z_i)$ 是 Z 大于 Z_i 的概率和， $P(Z < z_i)$ 是 Z 小于 Z_i 的概率和。对于有 1 个连续窗口的区域 A ，如果它符合 $\max\{piUpper | i \in A\} < (FPR/(L/l))^{(1/l)}$ ，则认为它发生了重复变异；如果它符合 $\max\{piLower | i \in A\} < (FPR/(L/l))^{(1/l)}$ ，则认为它发生了缺失，其中 FPR （错误率）是指对于参考基因组的重复或缺失的假阳性值之和，I 型错误是 FPR （假阳性率）——即检测出的 cnv 里面不是真实的 CNV 占检测出的所有 CNV 的比率，II 型错误是 FNR （假阴性率）——即没有被检测出的真实 cnv 占真实 cnv 总数的比率； L 是指参考基因组的总窗口数目， l 是指 A 区域的窗口数目，大于等于 1 且小于等于 L 。

很明显，如果 A 的所有探针是来自于正常状态，即 A 是 CNV 的可能性小于 $FPR/(L/l)$ ，该窗口属于正常窗口。

另外，对于缺失和重复这两种 CNV 可以分开进行检测。对于重复和缺失的检测均用上述公式，但前者的阈值是前述 π_{Upper} ，后者的是 π_{Lower} 。对于它们每一种的检测，首先寻找两个存在 CNV 的窗口，然后每增加 1 个窗口就重复增加后窗口是否也存在该 CNV。要注意的是， $FPR/(L/l) \wedge (1/l)$ 的阈值是随着 l 的增加而增加的。当 $FPR/(L/l) \wedge (1/l)$ 超过 0.5 时在 $N-1$ 处停止重复上述步骤。

第五步：对第四步得到的拷贝数变异结果进行过滤统计

对以上判断合属于缺失和重复这两种 CNV 的合并区域降低假阳性。降低假阳性可以通过一定的过滤条件实现。

10 过滤标准：1、发生拷贝数变异事件的窗口数要大于 10，即拷贝数变异区域要大于 1kb；2、发生拷贝数变异事件的覆盖度中位数倍数占全局覆盖度中位数的比例不在 0.75 和 1.25 之间。对本发明得到的结果过滤后，按拷贝事件（重复和缺失），对拷贝事件发生的数目和总长度进行统计，统计结果如下：

	重复	缺失	总和
拷贝数事件发生数目	10892	24007	34899
拷贝数变异长度 (bp)	176381800	447025700	623407500
拷贝数变异平均长度	16194	18621	17863

15

第六步，将用本发明检测得到的拷贝数变异结果与该样品芯片拷贝数变异扫描结果进行比较。利用发生的变异事件在基因组上的起始位点和终止位点信息，得到本发明检测结果和芯片检测结果一致的拷贝数变异情况，最终一致长度及比例如下表格。由表格结果看出，本发明检测出的拷贝数变异具有真实性。

20

	总共	重复	缺失
芯片检测拷贝数变异事件总长	9295697	7100675	2195022
本发明检测与芯片检测结果一致的总长	8050496	6171784	1878712
一致比例	86.60%	86.92%	85.59%

权 利 要 求 书

1. 一种检测拷贝数变异的方法，包括步骤：

1) 对目标样本进行测序，得到测序片段；

2) 将以上得到的测序片段与参考基因组序列比对，得到参考序列每个位点的覆盖深度，即比对结果中涵盖该位点的测序片段数目；

3) 根据以上每个位点的覆盖深度，将所有位点的覆盖深度取平均得出序列的平均覆盖深度，以同样的方式计算参考序列上具有预设片段长度的所有窗口的覆盖深度信息，这些窗口被定义为正常窗口、重复窗口、缺失窗口和 N 窗口四种类型之一，其中正常窗口为覆盖深度与序列平均覆盖深度相同的窗口，重复窗口为覆盖深度明显大于平均覆盖深度的窗口，缺失窗口为覆盖深度明显小于平均覆盖深度的窗口，N 窗口为基本没有覆盖深度的窗口；

4) 将满足以下条件任一项的三以上连续窗口合并，并判断合并后的区域是否属于缺失和重复这两种 CNV：

i. 连续重复窗口或连续缺失窗口、

ii. 缺失窗口之间有 N 窗口间隔，例如缺失窗口+N 窗口+缺失窗口，其中 N 窗口不能连续出现一个以上、

iii. 重复窗口之间有 N 窗口间隔，例如重复窗口+N 窗口+重复窗口，其中 N 窗口不能连续出现一个以上、

iv. 缺失窗口之间有正常窗口间隔，例如缺失窗口+正常窗口+缺失窗口，其中正常窗口的覆盖深度应满足减去 3 倍 SD 后落在缺失窗口的覆盖深度范围内，并且其中正常窗口不能连续出现一个以上、

v. 重复窗口之间有正常窗口间隔，例如重复窗口+正常窗口+重复窗口，其中正常窗口的覆盖深度应满足加上 3 倍 SD 后落在重复窗口覆盖深度范围内，并且其中正常窗口不能连续出现一个以上

其中，SD 为所有位点的覆盖深度取平均得出序列的平均覆盖深度标准差；

5) 对以上合并区域降低假阳性，所得到的区域就是发生了插入或缺失的拷贝数变异区域。

2. 根据权利要求 1 所述的方法，在步骤 1) 和 2) 之间，还包括步骤 1')：评估测序结果是否合格，如果不合格则需要重新测序，并且

如果在测序过程中引入接头序列，则去除这些接头序列。

3. 根据权利要求1或2的方法，其中步骤2)中在测序片段与参考基因组序列比对后包括步骤：去除比对结果中的重复和冗余。

4. 上述权利要求任一项的方法，其中步骤1)中的测序是454测序或Illumina测序。

5. 上述权利要求任一项的方法，其中步骤1)中测序的测序深度是10×、20×、30×或35×。

6. 上述权利要求任一项的方法，其中步骤3)中的预设片段长度为70 bp-100 bp、100 bp、100 bp-200 bp、50 bp-300 bp或50 bp-150 bp。

7. 上述权利要求任一项的方法，其中在步骤3)中覆盖深度与序列平均覆盖深度相同的窗口是差异20%、10%或5%的窗口。

8. 上述权利要求任一项的方法，其中在步骤3)中覆盖深度明显大于平均覆盖深度的窗口是指覆盖深度大于平均覆盖深度的2倍、4倍或8倍，覆盖深度明显小于平均覆盖深度是指覆盖深度小于平均覆盖深度的2倍、4倍或8倍。

9. 上述权利要求任一项的方法，其中步骤5)中的降低假阳性通过过滤进行。

10. 权利要求9的方法，其中所述过滤的过滤条件是测序片段中值在总体中值0.75倍和1.25倍之间，或者显著性水平 10^{-6} 。

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2012/001261

A. CLASSIFICATION OF SUBJECT MATTER

See the extra sheet

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC: C12Q, G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

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

DWPI, CNABS, CPRSABS, CNMED, SIPOABS, CNKI, ISI Web of Knowledge, Google Scholar: copy number mutation; copy number variation, CNV, sequencing, align, compare, depth, window, merge

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 101914628 A (BGI SHENZHEN CO., LTD.), 15 December 2010 (15.12.2010), see the whole document, and particularly claims 1-5	1-10
A	YOON, S. et al., Sensitive and accurate detection of copy number variants using read depth of coverage, GENOME RESEARCH, 05 August 2009, vol. 19, pages 1586-1592, and particularly pages 1586-1588.	1-10
A	NORD, A.S. et al., Accurate and exact CNV identification from targeted high-throughput sequence data, BMC GENOMICS, 12 April 2011, vol. 12, no. 184, pages 1-10, and particularly pages 7-9.	1-10

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

☒ See patent family annex.

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

Date of the actual completion of the international search
05 June 2013 (05.06.2013)

Date of mailing of the international search report
20 June 2013 (20.06.2013)

Name and mailing address of the ISA/CN:
State Intellectual Property Office of the P. R. China
No. 6, Xitucheng Road, Jimenqiao
Haidian District, Beijing 100088, China
Facsimile No.: (86-10) 62019451

Authorized officer

TIAN, Tian

Telephone No.: (86-10) **62412276**

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2012/001261

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CHIANG, D.Y. et al., High-resolution mapping of copy-number alterations with massively parallel sequencing, NATURAL METHODS, 30 November 2008, vol. 6, no. 1, pages 99-103.	1-10

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2012/001261

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 101914628 A	15.12.2010	HK 1147528 A0	12.08.2011
		WO 2012027958 A1	08.03.2012
		CN 101914628 BB	09.01.2013

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2012/001261

A. CLASSIFICATION OF SUBJECT MATTER

C12Q 1/68 (2006.01) i

G06F 19/22 (2011.01) i

国际检索报告

国际申请号
PCT/CN2012/001261

A. 主题的分类

见附加页

按照国际专利分类(IPC)或者同时按照国家分类和 IPC 两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

IPC: C12Q, G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

DWPI, CNABS, CPRSABS, CNMED, SIPOABS, CNKI, ISI Web of Knowledge, Google Scholar: 拷贝数变异、拷贝数突变、测序、比对、深度、窗口、合并; copy number variation、CNV、sequencing、align、compare、depth、window、merge。

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
A	CN 101914628 A (深圳华大基因科技有限公司) 15. 12 月 2010 (15.12.2010) 参见全文, 特别是权利要求 1-5	1-10
A	Yoon S 等, Sensitive and accurate detection of copy number variants using read depth of coverage, Genome Research, 05.8 月 2009, 第 19 卷, 第 1586-1592 页, 特别是第 1586-1588 页。	1-10
A	Nord AS 等, Accurate and exact CNV identification from targeted high-throughput sequence data, BMC Genomics, 12.4 月 2011, 第 12 卷, 第 184 期, 第 1-10 页, 特别是第 7-9 页。	1-10

☒ 其余文件在 C 栏的续页中列出。

☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期
05.6 月 2013 (05.06.2013)

国际检索报告邮寄日期
20.6 月 2013 (20.06.2013)

ISA/CN 的名称和邮寄地址:
中华人民共和国国家知识产权局
中国北京市海淀区蓟门桥西土城路 6 号 100088
传真号: (86-10)62019451

受权官员

田甜
电话号码: (86-10) **62412276**

C(续). 相关文件

类 型	引用文件，必要时，指明相关段落	相关的权利要求
A	Chiang DY等, High-resolution mapping of copy-number alterations with massively parallel sequencing, Natural methods, 30.11月2008, 第6卷, 第1期, 第99-103页。	1-10

关于同族专利的信息

PCT/CN2012/001261

PCT/ISA/210 表(同族专利附件) (2009 年 7 月)

A. 主题的分类

C12Q 1/68 (2006.01) i

G06F 19/22 (2011.01) i