

(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
2013 年 10 月 10 日 (10.10.2013) WIPO | PCT



(10) 国际公布号
WO 2013/149385 A1

- (51) 国际专利分类号:
G06F 19/22 (2011.01) C12Q 1/68 (2006.01)
- (21) 国际申请号: PCT/CN2012/073545
- (22) 国际申请日: 2012 年 4 月 5 日 (05.04.2012)
- (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): 深圳华大基因健康科技有限公司 (BGI HEALTH SERVICE CO., LTD) [CN/CN]; 中国广东省深圳市盐田区北山道 146 号北山工业区 11 栋二楼、三楼, Guangdong 518083 (CN)。
- (72) 发明人: 及
- (75) 发明人/申请人 (仅对美国): 李旭超 (LI, Xuchao) [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 陈盛培 (CHEN, Shengpei) [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 陈芳 (CHEN, Fang) [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 谢伟

伟 (XIE, Weiwei) [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)。 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)。 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)。 张秀清 (ZHANG, Xiuqing) [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: 中国国际贸易促进委员会专利商标事务所 (CCPIT PATENT AND TRADEMARK LAW OFFICE); 中国北京市西城区阜成门外大街 2 号万通新世界广场 8 层, Beijing 100037 (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, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, 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,

[见续页]

(54) Title: METHOD AND SYSTEM FOR DETECTING COPY NUMBER VARIATION

(54) 发明名称: 一种拷贝数变异检测方法和系统

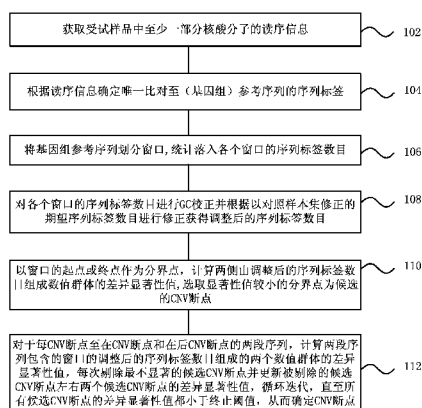


FIG. 1

102 ACQUIRE READ SEQUENCE INFORMATION OF AT LEAST ONE PART OF NUCLEIC ACID MOLECULES OF A SAMPLE UNDER TEST
104 DETERMINE A SEQUENCE LABEL FOR BEING UNIQUELY COMPARED WITH A (GENOME) REFERENCE SEQUENCE ACCORDING TO THE READ SEQUENCE INFORMATION
106 DIVIDE THE GENOME REFERENCE SEQUENCE INTO WINDOWS AND COUNTING THE NUMBER OF SEQUENCE LABELS FALLING INTO EACH WINDOW
108 PERFORM GC CORRECTION ON THE SEQUENCE LABEL NUMBER OF EACH WINDOW AND PERFORM AMENDMENT ACCORDING TO AN EXPECTED SEQUENCE LABEL NUMBER AMENDED WITH A CONTRAST SAMPLE SET TO OBTAIN AN ADJUSTED SEQUENCE LABEL NUMBER
110 TAKE A STARTING POINT OR AN END POINT OF A WINDOW AS A DEMARCATION POINT, CALCULATE A DIFFERENCE SIGNIFICANCE VALUE OF A VALUE GROUP FORMED OF ADJUSTED SEQUENCE LABEL NUMBERS AT TWO SIDES, AND SELECT A DEMARCATION POINT WITH A SMALL SIGNIFICANCE VALUE AS A CANDIDATE CNV BREAKING POINT
112 FOR TWO SEQUENCES FROM EVERY CNV BREAKING POINT TO A FORMER CNV BREAKING POINT AND A LATTER CNV BREAKING POINT, CALCULATE DIFFERENCE SIGNIFICANCE VALUES OF TWO VALUE GROUPS FORMED OF ADJUSTED SEQUENCE LABEL NUMBERS OF WINDOWS CONTAINED IN THE TWO SEQUENCES, REJECT THE LEAST SIGNIFICANT CANDIDATE CNV BREAKING POINT AT EVERY TURN, UPDATE DIFFERENCE SIGNIFICANCE VALUES OF TWO CANDIDATE CNV BREAKING POINTS ON THE LEFT AND RIGHT OF THE REJECTED CANDIDATE CNV BREAKING POINT, AND PERFORM CYCLIC ITERATION UNTIL DIFFERENCE SIGNIFICANCE VALUES OF ALL CANDIDATE CNV BREAKING POINTS ARE LESS THAN A TERMINATION THRESHOLD VALUE, THEREBY DETERMINING A CNV BREAKING POINT

(57) Abstract: Disclosed are a method and a system for detecting genome copy number variation, which relates to the technical field of bioinformatics. The method comprises obtaining a read sequence; determining sequence labels according to the read sequence; counting the number of sequence labels falling into each window; performing GC correction on the sequence label number of each window and performing amendment according to an expected sequence label number amended with a contrast sample set to obtain an adjusted sequence label number; selecting a demarcation point with a small significance value as a candidate CNV breaking point; rejecting the least significant candidate CNV breaking point at every turn, updating difference significance values of two candidate CNV breaking points on the left and right of the rejected candidate CNV breaking point and performing cyclic iteration until difference significance values of all candidate CNV breaking points are smaller than a termination threshold value, thereby determining a CNV breaking point. The method and the system of the present invention have clinical feasibility, and can precisely detect a micro-deletion/micro-repetition area of 0.5 M under the situation of using data of about 50 M.

(57) 摘要:

[见续页]



WO 2013/149385 A1



LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, 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。

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)。

(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, MD, RU, TJ, TM), 欧洲 (AL, AT,

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

本发明公开一种基因组拷贝数变异检测方法和系统, 涉及生物信息学技术领域。该方法包括: 获得读序; 根据读序确定序列标签; 统计落入各个窗口的序列标签数目; 对各个窗口的序列标签数目进行 GC 校正并根据以对照样本集修正的期望序列标签数目进行修正获得调整后的序列标签数目; 选取显著性值较小的分界点为候选的 CNV 断点; 每次剔除最不显著的候选 CNV 断点并更新被剔除的候选 CNV 断点左右两个候选 CNV 断点的差异显著性值, 循环迭代, 直至所有候选 CNV 断点的差异显著性值都小于终止阈值, 从而确定 CNV 断点。本发明的方法和系统, 具有临床可行性, 在使用 50M 左右的数据情况下, 可精确检测到 0.5M 的微缺失/微重复区域。