

(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
2016 年 1 月 21 日 (21.01.2016)



(10) 国际公布号
WO 2016/008146 A1

- (51) 国际专利分类号:
C12Q 1/68 (2006.01) G06F 19/00 (2011.01)
- (21) 国际申请号: PCT/CN2014/082462
- (22) 国际申请日: 2014 年 7 月 18 日 (18.07.2014)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (71) 申请人: 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: 曹红志 (CAO, Hongzhi); 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。刘爽 (LIU, Shuang); 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。张涛 (ZHANG, Tao); 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。庄雪寒 (ZHUANG, Xuehan); 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: 深圳鼎合诚知识产权代理有限公司 (DHC IP ATTORNEYS); 中国广东省深圳市福田区金田路与福华路交汇处现代商务大厦 2201, Guangdong 518048 (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, IR, IS, JP, KE, KG, 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, SA, 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, KM, ML, MR, NE, SN, TD, TG)。

[见续页]

(54) Title: GENDER IDENTIFICATION METHOD AND APPARATUS FOR SAMPLES

(54) 发明名称: 一种样本性别鉴定方法及装置

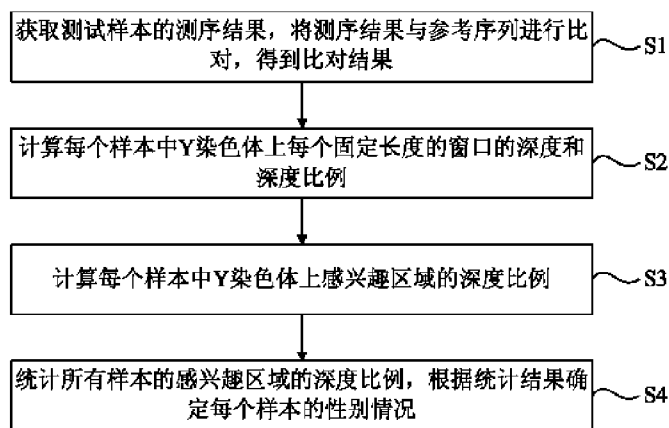


图 1 / FIG.1

- S1 ACQUIRE A SEQUENCING RESULT OF TEST SAMPLES, COMPARE THE SEQUENCING RESULT WITH A REFERENCE SEQUENCE, AND OBTAIN A COMPARISON RESULT
- S2 CALCULATE A DEPTH AND A DEPTH PROPORTION OF EACH WINDOW WITH FIXED LENGTH ON A Y-CHROMOSOME IN EACH SAMPLE
- S3 CALCULATE A DEPTH PROPORTION OF A REGION OF INTEREST ON THE Y-CHROMOSOME IN EACH SAMPLE
- S4 COLLECT STATISTICS ON DEPTH PROPORTIONS OF THE REGIONS OF INTEREST OF ALL THE SAMPLES, AND DETERMINE THE GENDER OF EACH SAMPLE ACCORDING TO A STATISTICAL RESULT

(57) Abstract: A gender identification method and apparatus for samples. The method comprises: acquiring a sequencing result of test samples, comparing the sequencing result with a reference sequence, and obtaining a comparison result; calculating a depth and a depth proportion of each window with fixed length on a Y-chromosome in each sample on the basis of the comparison result; calculating a depth proportion of a region of interest on the Y-chromosome in each sample; and collecting statistics on depth proportions of the regions of interest of all the samples, and determining the gender of each sample according to the collected statistical depth proportions.

(57) 摘要: 一种样本性别鉴定方法及装置, 其中方法包括: 获取测试样本的测序结果, 将所述测序结果与参考序列进行比对, 得到比对结果; 基于所述比对结果, 计算每个样本中 Y 染色体上每个固定长度的窗口的深度和深度比例; 计算每个样本中 Y 染色体感兴趣区域的深度比例; 统计所有样本的感兴趣区域的深度比例, 根据统计的深度比例确定每个样本的性别。

WO 2016/008146 A1

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

一种样本性别鉴定方法及装置

技术领域

本发明涉及基因组学及生物信息学技术领域，具体涉及一种样本性别鉴定方法及装置。

背景技术

在全基因组关联分析中，样本性别信息是针对个体的质量控制（quality control）的关键部分，可以用来补全表型信息。如果样本数据量不足，将可能导致无法进行性别信息的判断，使得缺少重要信息，从而不能对样本的准确性进行评估，或者需要进行额外的实验来判断性别。

通常样本的性别判断方法是基于 X 染色体上 SNP（单核苷酸多态性，Single Nucleotide Polymorphism）的杂合率（heterozygosity rates）来进行的，因为男性在 X 染色体上的 SNP 杂合率远低于女性在 X 染色体上的 SNP 杂合率。这种性别判断方法需要在 X 染色体上设计捕获探针，当 X 染色体上没有设计捕获探针时，准确度较低，无法准确判断大部分样本的性别。

发明内容

根据本发明的第一方面，本发明提供一种样本性别鉴定方法，包括：获取测试样本的测序结果，将所述测序结果与参考序列进行比对，得到比对结果；基于所述比对结果，计算每个样本中 Y 染色体上每个固定长度的窗口的深度和深度比例；计算每个样本中 Y 染色体上感兴趣区域的深度比例；统计所有样本的感兴趣区域的深度比例，根据统计结果确定每个样本的性别。

依据本发明的另一方面提供一种样本性别鉴定装置，包括：数据输入单元，用于输入数据；数据输出单元，用于输出数据；存储单元，用于存储数据，其中包括可执行的程序；处理器，与数据输入单元、数据输出单元及存储单元数据连接，用于执行存储单元中存储的可执行的程序，该程序的执行包括完成上述样本性别鉴定方法。

依据本发明的再一方面提供一种计算机可读存储介质，用于存储供计算机执行的程序，本领域普通技术人员可以理解，在执行该程序时，通过指令相关硬件可完成上述染色体非整倍性检测方法的全部或部分步骤。所称存储介质可以包括：只读存储器、随机存储器、磁盘或光盘等。

依据本发明的方法通过 Y 染色体上的数据分布来判断性别，相比现

有在 X 染色体上设计捕获探针的方法, 本发明的方法不需要特别设计 Y 染色体捕获位点, 节约了成本, 且相比现有 PLINK 方法节省了 snp calling 步骤, 由此可以节约时间且步骤的减少可以减少错误的产生。

附图说明

为了更清楚地说明本发明实施例, 下面将对实施例中所需要使用的附图作简单地介绍, 其中:

图 1 是依据本发明的一种实施方式的性别鉴定方法的流程示意图;

图 2 是依据本发明的另一种实施方式的性别鉴定方法的流程示意图;

图 3 是依据本发明的一种实施方式的性别鉴定方法的前期准备工作的示意图;

图 4 是依据本发明的一种实施方式的性别鉴定方法的整体数据处理的流程示意图;

图 5 是采用依据本发明的一种实施方式的性别鉴定方法的示例中 RA 数据的第一 XY 非同源区域的箱线图;

图 6 是采用依据本发明的一种实施方式的性别鉴定方法的示例中 307 数据的第一 XY 非同源区域的箱线图;

图 7 是采用依据本发明的一种实施方式的性别鉴定方法的示例中 EXON 数据的第二 XY 非同源区域的箱线图;

图 8 是采用依据本发明的一种实施方式的性别鉴定方法的示例中 EXON 数据的 Pars 区域的箱线图;

图 9 是采用依据本发明的一种实施方式的性别鉴定方法的示例中 307 数据的 Pars 区域的箱线图;

图 10 是采用依据本发明的一种实施方式的性别鉴定方法的示例中 RA 数据的 Pars 区域的箱线图。

具体实施方式

不同于传统的根据 X 染色体的 SNP 杂合度的判断方法, 本发明进行样本性别判断的思路是, 基于样本的芯片捕获数据, 统计样本在 Y 染色体上感兴趣区域的深度比例, 据此区分性别。此外, 本发明还结合性别判断情况来确定样本是否为污染样本。

下面通过具体实施方式结合附图对本发明作进一步详细说明。除非另有说明, 本文所用的所有技术和科学术语具有与本发明所述的技术领域的普通技术人员通常理解的不同含义。

实施例 1:

依据本发明的一种实施方式, 提供一种样本性别鉴定方法, 如图 1 所示, 包括如下步骤 S1~S4。

步骤 S1. 获取测试样本的测序结果, 将测序结果与参考序列进行比对, 得到比对结果。

测序样本为利用目标区域捕获芯片捕获指定物种的基因片段, 实施例中测序样本为基于芯片捕获的人类基因组 DNA 样品。

依据本发明的实施方式, 对芯片捕获的方法和设备不作限定, 并且对从样本中提取用于测序的核酸的方法和设备也不作限定, 可以采用各种已有的手段进行, 例如商品化的核酸提取试剂盒, 另外, 对样本的测序方法和设备也没有特殊依赖。

本实施方式测序过程得到的原始数据以 fastq 格式 (又称 fq 格式) 文件存储, 其存储着读长序列 (即 reads, 也称读段) 以及 reads 的测序质量等。在得到原始的 fq 数据后, 对其进行去接头、去低质量 reads 等基本处理, 该基本处理可以是例如第二代测序数据采用的处理方法, 本实施方式不作限制, 基本处理后得到干净的序列 (称为 clean reads), 为测序结果。

所使用的参考序列是已知序列, 可以是预先获得的目标个体所属生物类别中的任意的参考模板。例如, 若目标个体是人类, 参考序列可选择美国国家生物技术信息中心 (NCBI, national center for biotechnology information) 数据库中的人类基因组参考序列。本实施方式中, 参考序列选择为 NCBI 数据库中版本 37.3 (hg19; NCBI Build 37.3) 的人类基因组参考序列。

在将测序结果比对到参考序列时, 可使用各种比对软件, 例如 Tmap, BWA (Burrows-Wheeler Aligner), SOAP (Short Oligonucleotide Analysis Package) 等, 本实施方式对此不作限定。例如可以采用 BWA 软件, BWA 是为二代测序的短序列比对参考序列而开发的比对软件。比对参数可以是例如: `bwa aln -o 1 -e 50`。aln 是 BWA 提供的功能之一, 目的是将输入目标序列与参考序列进行比对。比对后得到 .bam 文件, 其中 bam 文件采用的 bam 格式是用于存储对参考序列的片段的一个通用比对格式。

步骤 S2. 计算每个样本中 Y 染色体上每个固定长度的窗口的深度和深度比例。

本实施方式中，测序结果中的 Y 染色体按照固定长度划分窗口，固定长度优选为 100bp（碱基对，base pair）。

简明起见，将每个固定长度的窗口的深度记为 D_i ，将 Y 染色体总深度

记为 $\sum_{i=0}^n D_i$ ，将每个固定长度的窗口的深度比例（即每个固定长度的

窗口的深度占 Y 染色体总深度的比例）记为 R_i ，显然 $R_i = D_i / \sum_{i=0}^n D_i$ ，

且 $\sum_{i=0}^n R_i = 1$ ，其中 i 为窗口的编号， $0 \leq i \leq n$ ， n 为窗口总数。

步骤 S3. 计算每个样本中 Y 染色体上感兴趣区域的深度。

本实施方式中感兴趣区域为 Y 染色体上的拟常染色体区域（PARs, pseudo autosomal regions）。

Y 染色体的基本数据情况为：总碱基数为 59373566，N 含量（即尚未测出的碱基）为 56.79%，GC 含量为 39.97%。其中，Y 染色体有两部分是 PARs，为与 X 染色体发生重组的 Y 染色体部分，其位置为：chrY（Y 染色体基因组序列）：10001-2649520 和 chrY：59034050-59363566，以及 chrX（X 染色体基因组序列）：60001-2699520 和 chrX：154931044-155260560。

经研究发现，女性样本与男性样本在 PARs 区域的深度相近似，而由于男性样本的 Y 染色体总深度大于女性样本的 Y 染色体总深度，所以女性样本在此区域的深度比例显著大于男性在此区域的深度比例，因此，可以利用 PARs 区域的这一特点来区分男女性别。这里，深度是指 Y 染色体每个位点的深度，深度比例是指 Y 染色体一段区域的深度占 Y 染色体总深度的比例。

简明起见，将感兴趣区域（本实施方式中为 Y 染色体的 PARs）记为区域 A，类似步骤 S2，区域 A 按照步骤 S2 中的固定长度划分窗口，固定长度优选为 100bp，将得到的区域 A 的窗口数目记为 m ，将一个样

本中 Y 染色体上区域 A 的深度（即这 m 个窗口的深度比例的求和）记

为 DA ，显然 $DA = \sum_{l=0}^m R_l$ ，其中 R_l 中的 l 为区域 A 中的第一个固定长度

的窗口。

步骤 S4. 统计所有样本的感兴趣区域的深度比例，确定出每个样本的性别。

简明起见，将某个样本 j 中 Y 染色体上区域 A 的深度比例记为 DA_j ，其中 $0 \leq j \leq N$ ， N 为所有样本的数量。

根据统计所有 DA_j ，可以利用模式识别中常用的分类算法区分出每个样本的性别，例如根据 SVM（支持向量机，Support Vector Machine，一种广泛使用的数理统计学习方法）的分析原理对样本的性别进行判断。

本实施例中通过箱线图的方式区分出样本的性别。箱线图 (Boxplot) 也称箱须图 (Box-whisker Plot)，是利用数据中的五个统计量：最小值、第一四分位数、中位数、第三四分位数与最大值来描述数据的一种方法，它也可以粗略地看出数据是否具有对称性、分布的分散程度等信息，特别可以用于对几个样本的比较。在本实施例中，画出所有 DA_j 的箱线图后，男女样本将分开至不同区域范围，至此可以区分性别信息。其它实施例中也可以不采用箱线图的方式，而是采用其它统计图。

基于本实施例，还可以进一步判断出样本是否被污染。例如，在步骤 S4 中，如果存在某样本的 DA_j 不在男性范围和女性范围内的样本，无法准确判断出性别，需要进一步检查样本信息时，一般可以将该样本划归为污染样本。

基于本实施例，还可以进一步确定错误率，错误率是指错误样本（性别信息不匹配）占总样本的比例，也即采用本实施例判断错误的样本性别信息与实际样本性别信息的比值，假设总样本数为 N ，错误样本数为 M ，则错误率为 $e = M/N$ 。

通过本实施例，首先，使用捕获芯片数据进行性别判断，不需要特别设计 Y 染色体的捕获位点，一定程度上节约了成本；其次，样本性别鉴定可从比对结果 bam 文件判断性别，对比现有采用 PLINK 方法，节省了 snp calling 的步骤，可以节约时间，而且步骤的减少也进一步减少了错误的产生；此外，由于 Y 染色体数据是随机捕获，对数据分布具有

特异性的样本（即感兴趣区域的深度比例不落入男性范围和女性范围，可以判断为污染样本，从而可以在后续的数据分析中排除。

实施例 2:

为对样本的性别情况进行鉴定，本实施方式的整体思路是寻找 Y 染色体上男女差异最大的区域。

一般地，无论 Y 染色体上是否设计有探针和无探针，Y 染色体的数据都能或多或少被捕获，因为所有探针的特异性不会为 100%，也就意味着有其他区域的信息可以被捕获；基于此，本文假设所涉及的比对在 Y 染色体上的数据是准确的且实验中没有污染，则女性 Y 染色体和男性 Y 染色体的数据来源为如下：

对于 Y 染色体上有探针的数据：

1) 女性 Y 染色体数据来源包括三种，即①探针区域与 Y 同源区域，②X 染色体与 Y 同源区域，以及③其它突变导致与 Y 同源。显然，此时女性 Y 染色体数据的主要来源为 X 染色体与 Y 同源区域，该区域又称 XY 的同源区域。

2) 男性 Y 染色体数据来源包括四种，即①探针捕获区域、②X 染色体与 Y 同源区域、③实验中未洗脱的 Y 染色体（即随机残留的 Y 染色体信息）、以及④其它突变导致与 Y 同源。显然，此时男性 Y 染色体数据的主要来源为探针捕获区域。

对于 Y 染色体上无探针的数据：

3) 女性 Y 染色体数据来源包括三种，即①X 染色体与 Y 同源区域、②非 Y 染色体的探针区域与 Y 同源区域、以及③其它突变导致与 Y 同源。显然，此时女性 Y 染色体数据的主要来源为 X 染色体与 Y 同源区域（即 XY 的同源区域）。

4) 男性 Y 染色体数据来源包括四种，即①非 Y 染色体的探针区域与 Y 同源区域、②实验中未洗脱的 Y 染色体、③X 染色体与 Y 同源区域、以及④其它突变导致与 Y 同源。显然，此时男性 Y 染色体数据的主要来源为实验中未洗脱的 Y 染色体（即随机捕获的区域）。

可见，女性 Y 染色体的数据来源主要是 XY 的同源区域，男性 Y 染色体的数据来源在 Y 染色体有探针的情况下主要是探针捕获区域，在 Y 染色体无探针的情况下主要是随机捕获的区域。其中，针对 Y 染色体上有探针的数据，由于 Y 染色体上的数据来源比较明确，男女之间差距比

较大，容易判断性别；而针对 Y 染色体上无探针的数据，猜测男女 Y 染色体上的数据来源，统计数据判断猜测是否正确，找到男女之间的差距，据此来判断性别。

也就是说，对于 Y 染色体有探针的数据，只要从男性 Y 染色体数据量集中的探针区域中去掉女性 Y 染色体主要数据区域（即 XY 的同源区域），则可以得到男女差异最大的区域（即 XY 的非同源区域，本文称为第二 XY 非同源区域）；而对于 Y 染色体无探针的数据，男性 Y 染色体的数据是随机分布的，女性 Y 染色体数据来源是 XY 同源区域，从 Y 染色体的非 N 区（N 区又称 gap 区，即还没有测出碱基序列的区域）去掉 XY 同源区域，得到的即为男女差异最大的区域（即 XY 的非同源区域，本文称为第一 XY 非同源区域）。显然，无论是否设计有探针，男女差异最大的区域在 XY 的非同源区域。

基于上述分析，依据本发明的一种实施方式，提供一种样本性别鉴定方法，基本步骤与实施例 1 相同，区别在于，实施例 1 中感兴趣区域为 Y 染色体上的 PARs 区域，且比对软件允许出现 repeat 比对的数据，而本实施例中感兴趣区域为 XY 非同源区域，且比对得到的是 unique 比对的数据，也就是说，本实施例在进行类似实施例 1 的步骤 S3 时，需要针对 Y 染色体上有无探针分情况处理，此时图 1 修改为图 2 所示。

实施例 3:

依据本发明的一种实施方式，提供一种样本性别鉴定方法，其实际上是综合了实施例 1 和实施例 2 两种处理过程。

首先进行前期的区域准备，以 100bp 的长度为单位，如图 3 所示，以便为后续统计样本各感兴趣区域的深度比例打下基础，由图 3 可见，最终确定由三种感兴趣区域来判断性别，分别为：第一 XY 非同源区域、第二 XY 非同源区域和 PARs 区域。对所有类型的数据，都可以通过 PARs 区域来判断性别；对 Y 染色体上有探针的数据，可以由第二 XY 非同源区域判断性别；对 Y 染色体上无探针的数据，可以由第一 XY 非同源区域判断性别。

然后进行数据处理流程，图 4 示意性地展示了本实施例的整个数据处理的流程，其中的基本数据处理包括了去接头、比对、去重复等步骤。由于 PARs 区域多为 repeat 比对，因此单独统计其深度比例。

本领域普通技术人员可以理解，上述实施方式中各种方法的全部或

部分步骤可以通过程序来指令相关硬件完成，该程序可以存储于一计算机可读存储介质中，存储介质可以包括：只读存储器、随机存储器、磁盘或光盘等。

依据本发明的另一方面还提供一种样本性别鉴定装置，包括：数据输入单元，用于输入数据；数据输出单元，用于输出数据；存储单元，用于存储数据，其中包括可执行的程序；处理器，与上述数据输入单元、数据输出单元及存储单元数据连接，用于执行存储单元中存储的可执行的程序，该程序的执行包括完成上述实施方式中各种方法的全部或部分步骤。

以下结合具体实验对依据本发明的样本性别鉴定方法的运行进行详细的描述。

数据来源如下：

第一组数据：EXON 数据，为全外显子捕获数据，Y 染色体上存在探针区域，145 个男性样本和 25 个女性样本；

第二组数据：RA（类风湿性关节炎，Rheumatoid Arthritis）数据，为 MHC（主要组织相容性复合体，Major Histocompatibility Complex）芯片捕获数据，Y 染色体上不存在探针区域，70 个男性样本和 290 个女性样本；

第三组数据：307 数据，为 MHC 芯片捕获数据，Y 染色体上不存在探针区域，88 个男性样本和 80 个女性样本。

下面将讨论三种区域，分别为：第一 XY 非同源区域、第二 XY 非同源区域和 PARs 区域。

1、对于第一 XY 非同源区域

适用数据类型：Y 染色体上没有设计捕获探针的数据。

区域获得方法：对 hg19 用 blast 找到 Y 染色体上 X 和 Y 染色体的同源区域。转化成 100bp 为单位的区域格式，记为区域 TY（与 Y 染色体的同源性最高的染色体是 X 染色体，因此只考虑 XY 的同源区域）。用 Y 染色体的非 N 区去掉区域 TY，得到 Y 染色体上的第一 XY 非同源区域。

实施方法：对 Y 染色上的 unique 比对的 reads 按照前述数据处理步骤进行统计，得到每个样本的 DA 值，而后确认性别信息。

实验一：采用上述第二组数据，即 RA 数据，其中样本为 70（男性）

VS 290 (女性), 采用本发明的方法得到的错误率为 1.38%。

从图 5 可以看出, 利用第一 XY 非同源区域区分男女是可行的, 其中正常值的男性样本的比值范围为 62.34338-80.8513, 正常女性样本的范围为 16.884035-24.253415。由此标准, 有两个男性样本落入女性区域中间, 三个女性样本落入男性区域中间。判断为样本信息错误, 性别录入有误。另外, 女性位于异常值范围内的样本较多, 初步认为是实验污染所致, 具体原因需要进一步研究。

实验二:: 采用上述第二组数据, 即 307 重组研究, 其中样本为 88 (男性) VS 80 (女性), 采用本发明的方法得到的错误率为 0.5%。

从图 6 同样可以看出, 利用第一 XY 非同源区域区分男女是可行的, 其中正常值的男性样本的比值范围为 62.30353-85.99705, 正常女性样本的范围为 23.557455-31.589075。与 RA 项目相比, 男性的比值范围基本稳定, 由此标准, 有一个男性样本落入女性区域中间, 无女性样本落入男性区域中间。判断为样本信息错误, 性别录入有误。此数据中有 3 个样本为异常值, 同样初步认为是实验污染所致。

2、对于第二 XY 非同源区域

适用数据类型: Y 染色体上设计捕获探针的数据。

区域获得方法: 提供的数据包含在 Y 染色体上设计的探针, 处理 Y 染色体上的探针区域为 100bp 的格式, 记为区域 TZ。用区域 TZ 减去区域 TY 得到探针区域中的 XY 非同源区域, 记为第二 XY 非同源区域。

实施方法: 对 Y 染色上的 unique 比对的 reads 按照前述数据处理步骤进行统计, 得到每个样本的 DA 值, 而后确认性别信息。

实验三: 采用上述第一组数据, 即 EXON (全基因组外显子捕获数据), 其中样本为 145 (男性) VS 25 (女性), 采用本发明的方法得到的错误率为 1.17%。

由图 7 可以看出, 第二 XY 非同源区域同样可以区分性别。其中正常男性范围为 40.49876-49.71196 正常女性样本范围为 0.0-0.9673012。无男性样本落入女性范围中, 两个女性样本落入男性范围中。部分男性样本为异常值, 初步判断为实验污染样本。

3、对于 PARs 区域

适用数据类型: 比对软件允许出现 repeat 比对的数据, 且对 Y 染色体上是否有探针不作要求。

区域获得方法：提取 hg19 中 Y 染色体的 pseudo autosomal regions, 处理成 100bp 为单位的格式，记为 PARs 区域。

实施方法：对 Y 染色上的所有 reads（包括 repeat 比对）按照前述数据处理步骤进行统计，得到每个样本的 DA 值，而后确认性别信息。

实验四：采用上述第一组数据，即 EXON（全基因组外显子捕获数据），其中样本为 145（男性）VS 25（女性），采用本发明的方法得到的错误率为 1.17%。

图 8 与图 7 应用的都是 EXON 研究的数据，得到同样的结果，有两个女性样本落入男性区域范围内。图中的正常值的男性区域样本范围是 10.362235-14.012735，正常值女性样本范围是 76.56507-81.11711。

实验五：采用上述第三组数据，即 307 重组研究，其中样本为 88（男性）VS 80（女性），采用本发明的方法得到的错误率为 0.5%。

图 9 与图 6 应用的都是 307 数据，得到的结果也相同。一个男性样本落入女性样本区域范围中。其中正常值的男性样本的比值范围为 2.0461375-2.9630875，正常女性样本的范围为 24.340505-34.821805。

实验六：采用上述第二组数据，即 RA 数据，其中样本为 70（男性）VS 290（女性），采用本发明的方法得到的错误率为 1.1%。

图 10 与图 5 同为 RA 项目的数据。其中正常值的男性样本的比值范围为 1.8148895-2.9568075，正常女性样本的范围为 29.943245-36.244065。性别判断的结果与图 1 稍有不同。图 1 中有三个女性样本落入男性区域中，而在图 10 则只有两个样本落入男性范围中，将图 5 比图 10 多出的样本记为样本 A。图 5 中的样本 A 处在正常男性比值的边缘区，可以说这个样本偏向男性，具体的性别还需继续确认。

以上应用了具体个例对本发明进行阐述，只是用于帮助理解本发明并不用以限制本发明。对于本领域的一般技术人员，依据本发明的思想，可以对上述具体实施方式的变化。

权 利 要 求

1. 一种样本性别鉴定方法，其特征在于，包括：

获取测试样本的测序结果，将所述测序结果与参考序列进行比对，得到比对结果；

基于所述比对结果，计算每个样本中 Y 染色体上每个固定长度的窗口的深度和深度比例；

计算每个样本中 Y 染色体上感兴趣区域的深度比例；

统计所有样本的感兴趣区域的深度比例，根据统计的深度比例确定每个样本的性别。

2. 如权利要求 1 所述的方法，其特征在于，所述感兴趣区域为 Y 染色体上的拟常染色体区域。

3. 如权利要求 1 所述的方法，其特征在于，所述感兴趣区域包括第一 XY 非同源区域和第二 XY 非同源区域，所述第一 XY 非同源区域为 Y 染色体无探针时 Y 染色体的非 N 区去除 XY 同源区域后的区域，所述第二 XY 非同源区域为 Y 染色体有探针时 Y 染色体去除 XY 同源区域后的区域。

4. 如权利要求 1 所述的方法，其特征在于，所述根据统计结果确定每个样本的性别这一步骤包括：

根据统计的所有样本的感兴趣区域的深度比例，确定出男性范围和女性范围，根据每个样本的 Y 染色体上感兴趣区域的深度比例落在所述男性范围和女性范围的结果，确定每个样本的性别。

5. 如权利要求 1 所述的方法，其特征在于，还包括：

如果一个样本的 Y 染色体上感兴趣区域的深度比例不落在所述男性范围和女性范围，则确定该样本为污染样本。

6. 如权利要求 1 所述的方法，其特征在于，还包括：

根据确定出的所有样本的性别，计算错误率，所述错误率为性别鉴定错误的样本数与总样本数的比值。

7. 如权利要求 1-6 任一项所述的方法，其特征在于：所述测试样本为基于芯片捕获的人类基因组 DNA 样品。

8. 如权利要求 1-6 任一项所述的方法，其特征在于，所述固定长度为 100bp。

9. 如权利要求 1-6 任一项所述的方法，其特征在于：所述深度是指 Y 染色体每个位点的深度，所述深度比例是指 Y 染色体在一段区域的深度占 Y 染色体总深度的比值。

10. 一种样本性别鉴定装置，其特征在于，包括：

数据输入单元，用于输入数据；

数据输出单元，用于输出数据；

存储单元，用于存储数据，其中包括可执行的程序；

处理器，与所述数据输入单元、数据输出单元及存储单元数据连接，用于执行所述可执行的程序，所述程序的执行包括完成如权利要求 1-9 任意一项所述的方法。

11. 一种计算机可读存储介质，其特征在于，用于存储供计算机执行的程序，所述程序的执行包括完成如权利要求 1-9 任意一项所述的方法。

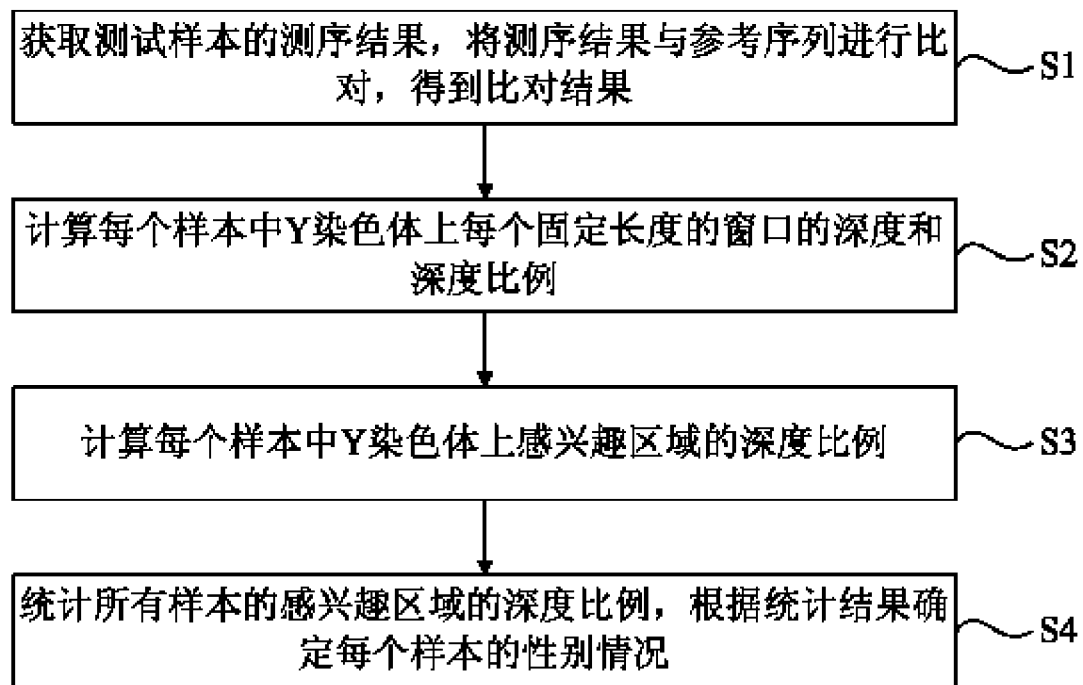


图 1

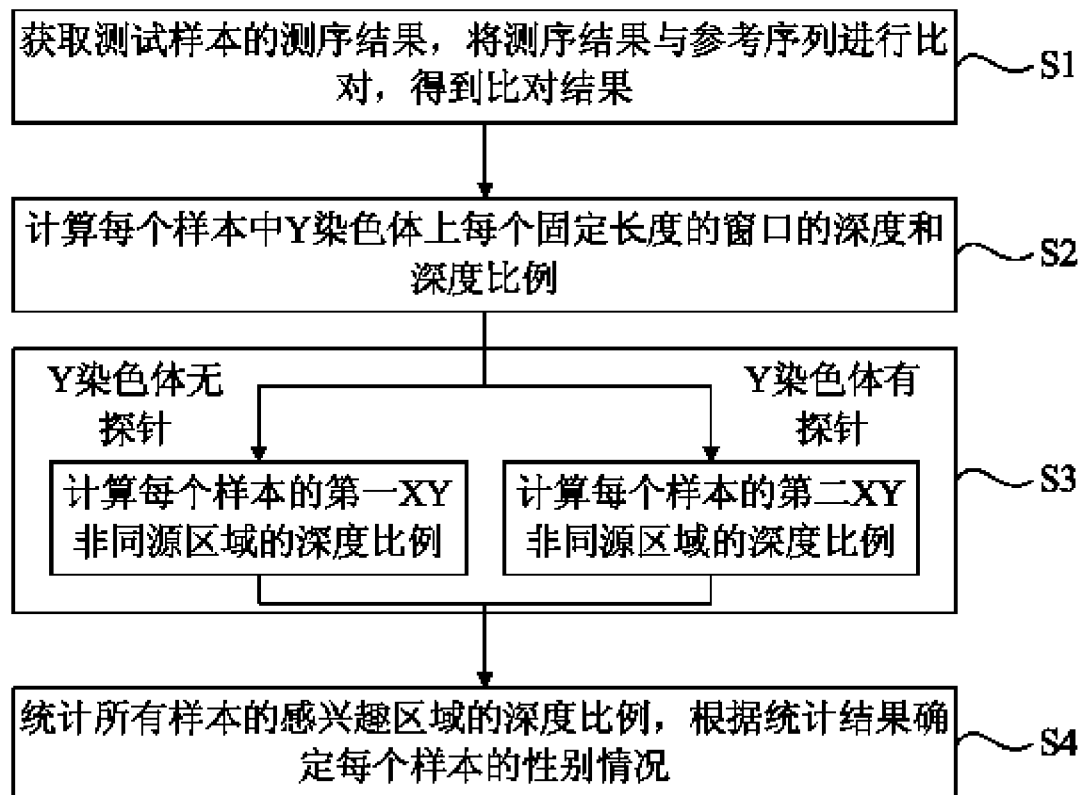


图 2

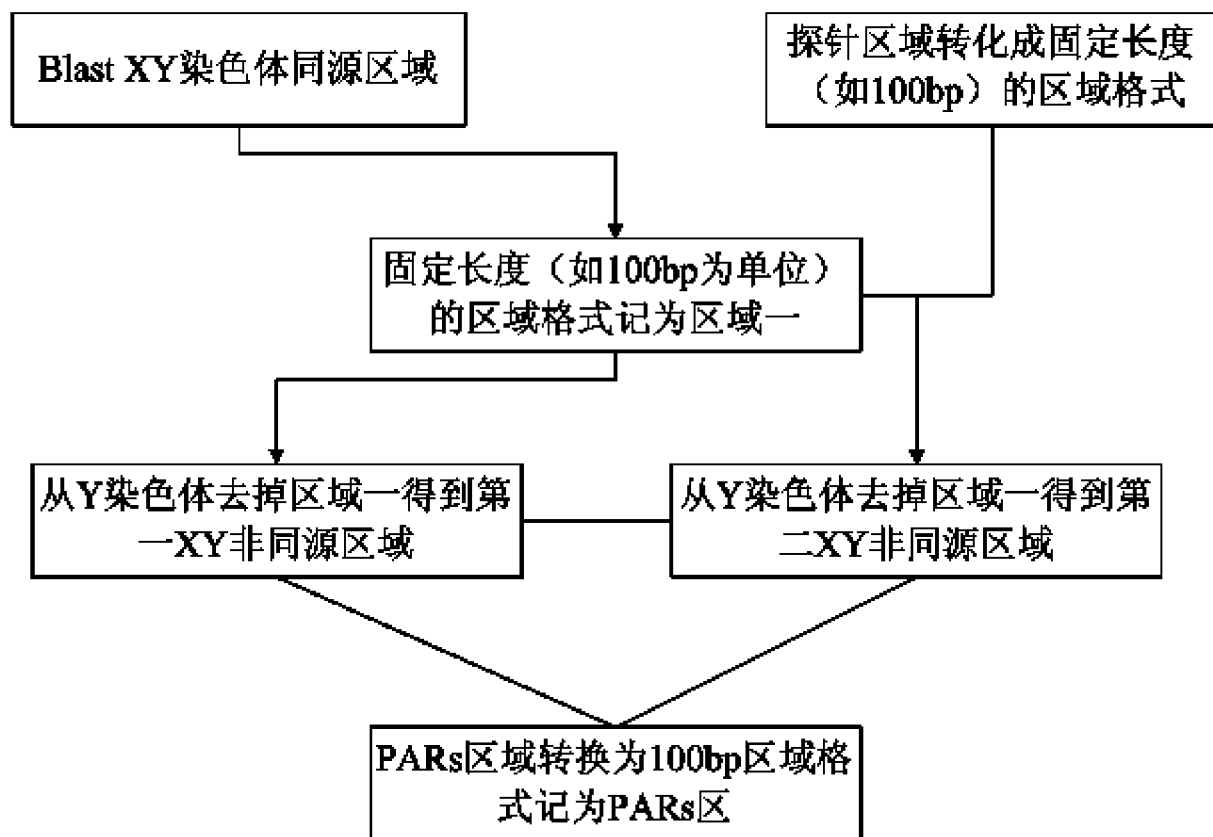


图 3

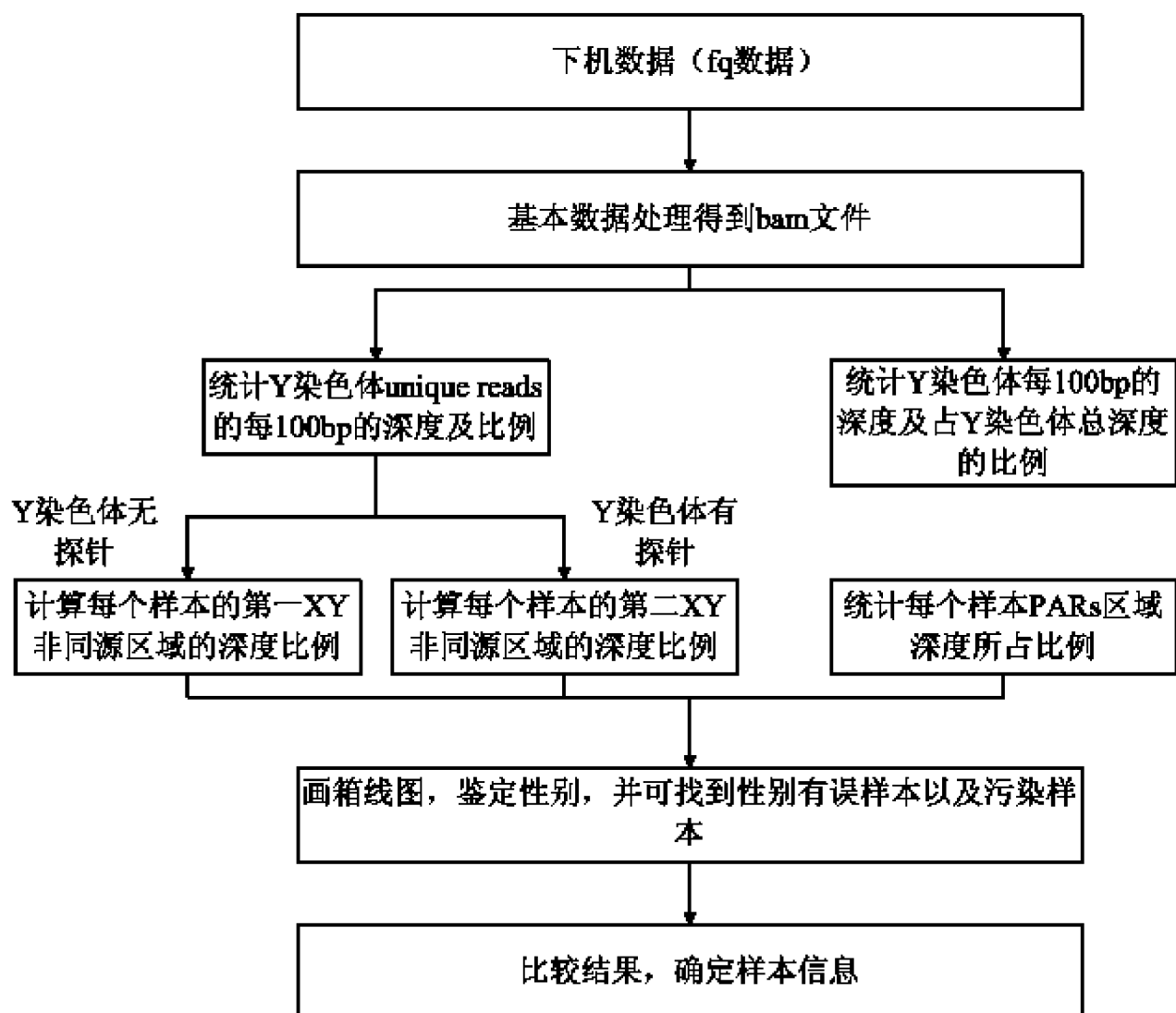


图 4

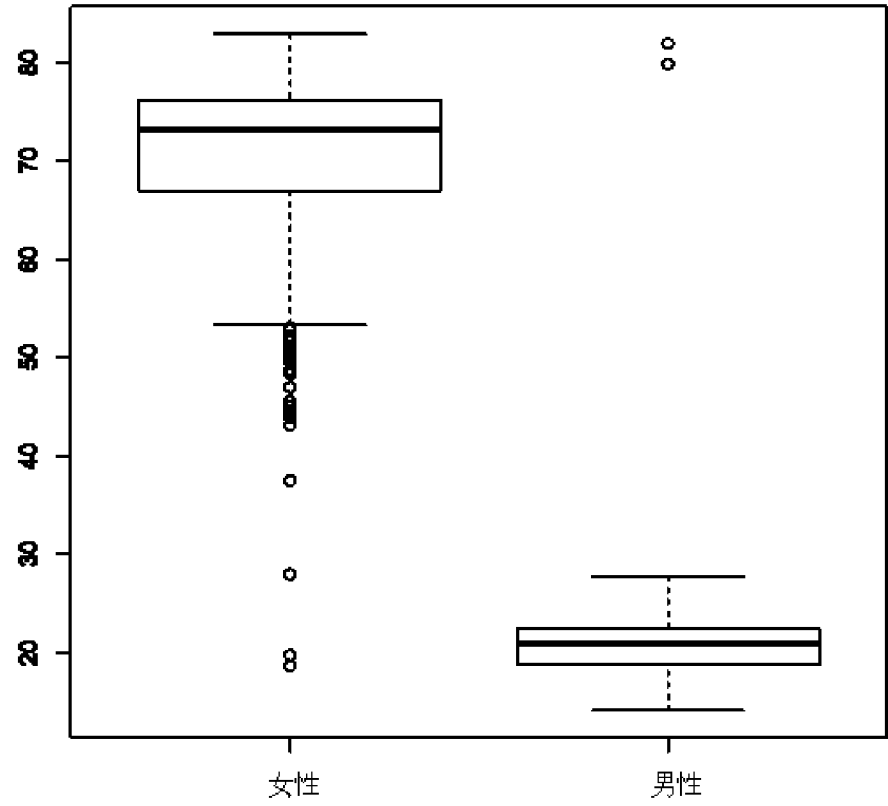


图 5

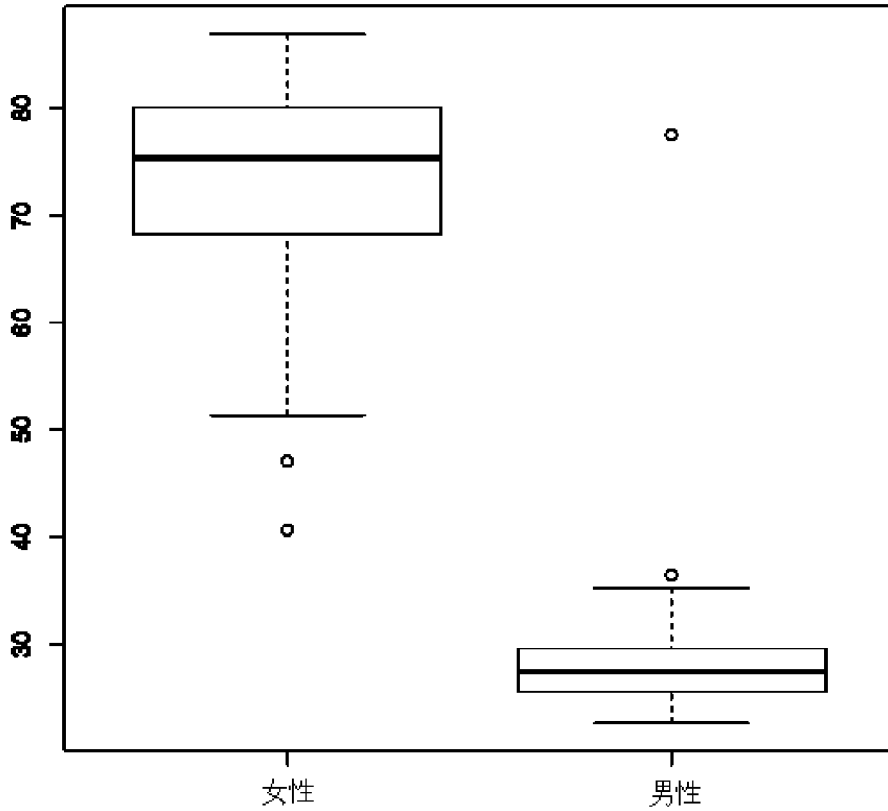


图 6

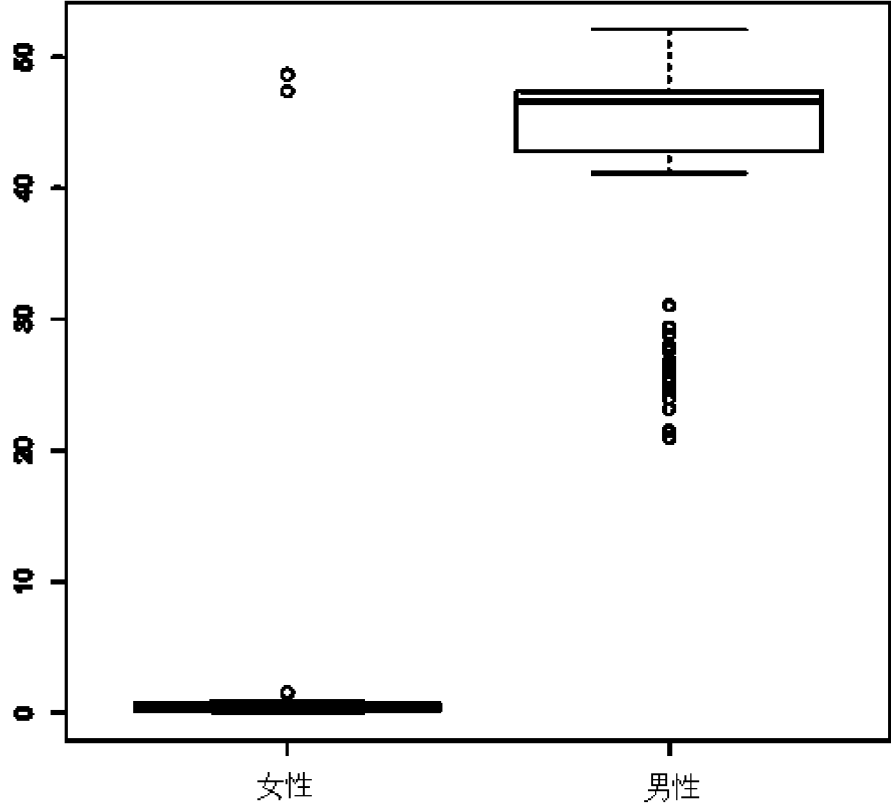


图 7

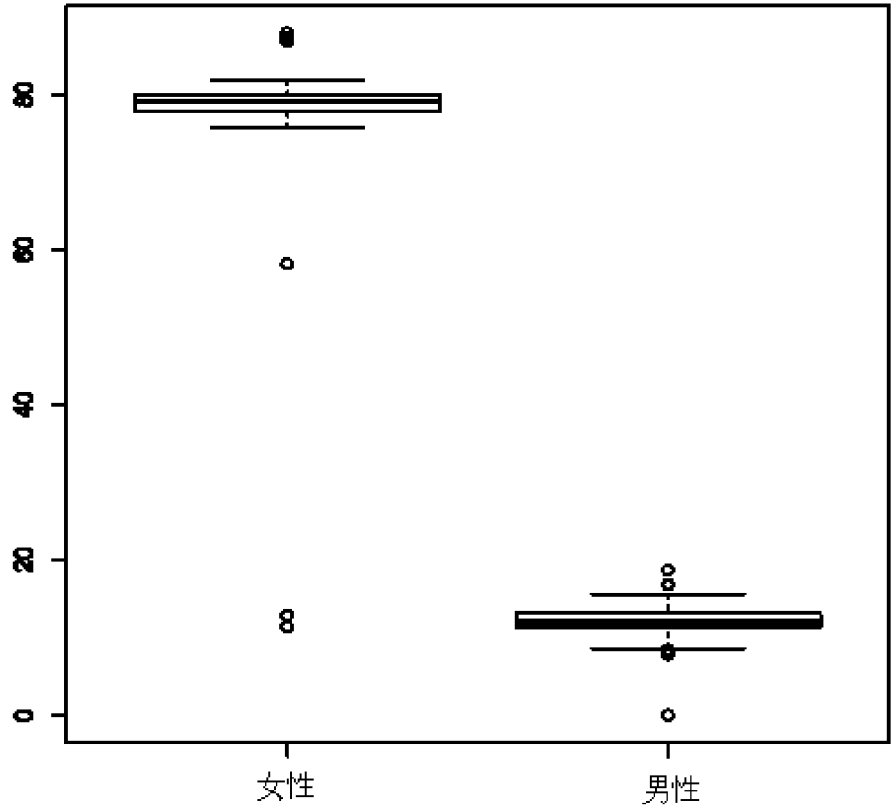


图 8

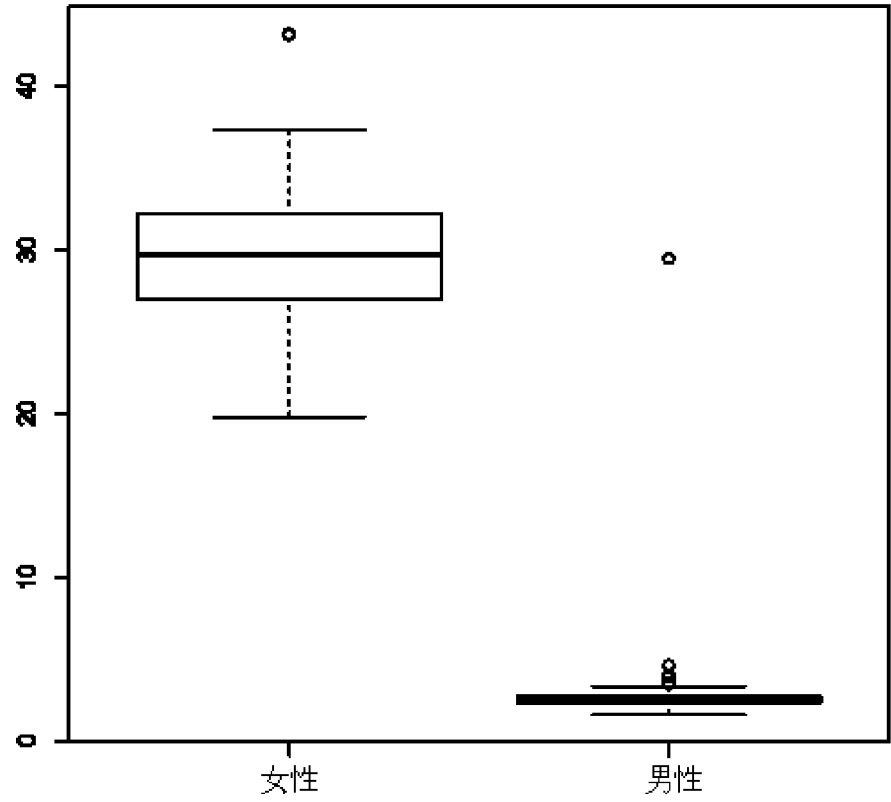


图 9

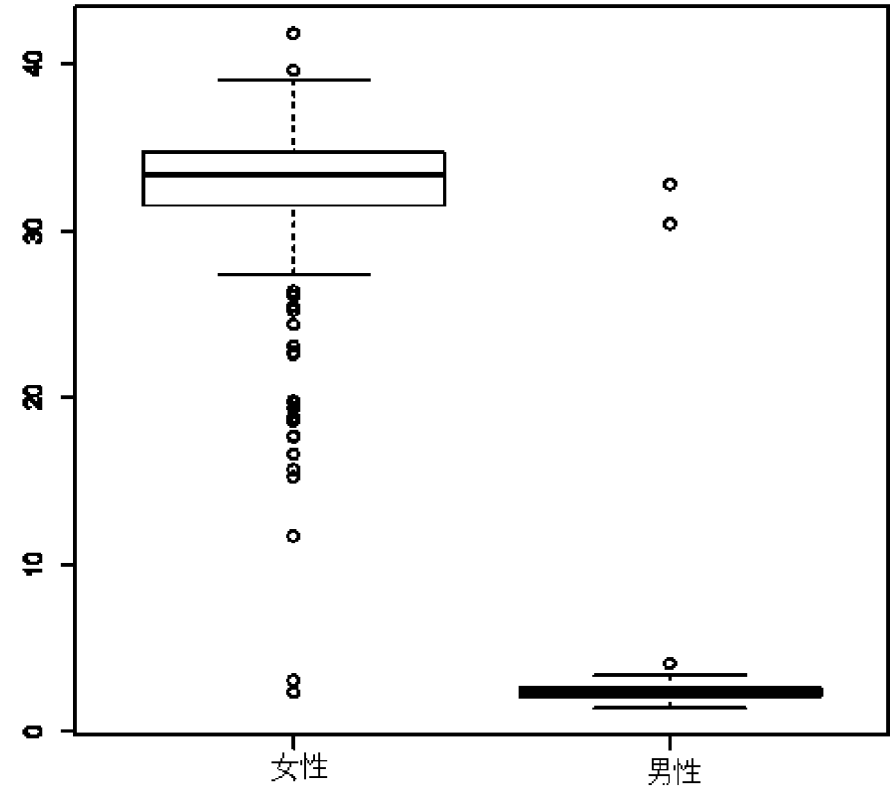


图 10

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2014/082462

A. CLASSIFICATION OF SUBJECT MATTER

C12Q 1/68 (2006.01) i; G06F 19/00 (2011.01) i

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)

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)

WPI, EPODOC, CNPAT, CNKI, ISI Web of knowledge: gender, sex, identification, detection, sequencing, compare, alignment, depth, coverage, reads, PAR, pseudoautosomal, non-homologous, window, input, output, storage, processor

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	Songlin Chen et al., "Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle", Nature Genetics, volume 46, number 3, 02 February 2014 (02.02.2014), the part "online methods"	1-10
A	David R. Bentley et al., "Accurate whole human genome sequencing using reversible terminator chemistry:", Nature, volume 456, number 7218, 06 November 2008 (06.11.2008), pages 53-59	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 10 June 2015	Date of mailing of the international search report 02 July 2015
Name and mailing address of the ISA 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 LI, Yang Telephone No. (86-10) 62411031

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2014/082462

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☒ Claims Nos.: 11
because they relate to subject matter not required to be searched by this Authority, namely:
Claim 11 relates to a computer readable storage medium for storing computer executable programs. The subject matter thereof is mere presentations of information, the situation set out in PCT Rule 39. 1 (v).
2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on protest

- ☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
- ☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- ☐ No protest accompanied the payment of additional search fees.

国际检索报告

国际申请号

PCT/CN2014/082462

A. 主题的分类

C12Q 1/68(2006.01)i; G06F 19/00(2011.01)i

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)

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))

WPI, EPODOC, CNPAT, CNKI, ISI Web of knowledge: 性别, 鉴定, 检测, 测序, 比较, 比对, 深度, PAR, 拟常染色体区, 窗口, 读段, 覆盖, 非同源区, 数据输入, 数据输出, 存储器, 存储单元, 处理器, gender, sex, identification, detection, compare, alignment, comparison, depth, coverage, reads, pseudoautosomal, non-homologous, input, output, storage, processor

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
X	Songlin Chen等. "Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle" Nature Genetics, 第46卷卷, 第3期期, 2014年 2月 2日 (2014 - 02 - 02), "online methods" 部分	1-10
A	David R. Bentley等. "Accurate whole human genome sequencing using reversible terminator chemistry" Nature, 第456卷卷, 第7218期期, 2008年 11月 6日 (2008 - 11 - 06), 第53-59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

2015年 6月 10日

国际检索报告邮寄日期

2015年 7月 2日

ISA/CN的名称和邮寄地址

中华人民共和国国家知识产权局(ISA/CN)
北京市海淀区蓟门桥西土城路6号
100088 中国

传真号 (86-10)62019451

授权官员

李洋

电话号码 (86-10)62411031

第II栏 某些权利要求被认为是不能检索的意见(续第1页第2项)

根据条约第17条(2)(a)，对某些权利要求未做国际检索报告的理由如下：

1. ☒ 权利要求： 11
因为它们涉及不要求本单位进行检索的主题，即：
[1] 权利要求11涉及一种计算机可读存储介质，用于存储供计算机执行的程序。其主题是单纯的信息表述，属于PCT细则39.1 (v)所列情形。
2. ☐ 权利要求：
因为它们涉及国际申请中不符合规定的要求的部分，以致不能进行任何有意义的国际检索， 具体地说：
3. ☐ 权利要求：
因为它们是从属权利要求，并且没有按照细则6.4(a)第2句和第3句的要求撰写。