

(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2月7日 (07.02.2013)



(10) 国际公布号
WO 2013/016864 A1

(51) 国际专利分类号:
C12Q 1/68 (2006.01)

(21) 国际申请号: PCT/CN2011/077861

(22) 国际申请日: 2011年8月1日 (01.08.2011)

(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): **朱昱其 (ZHU, Yuqi)** [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, Guangdong 518083 (CN)。 **金鑫 (JIN, Xin)** [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, Guangdong 518083 (CN)。 **李莹 (LI, Ying)** [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, Guangdong 518083 (CN)。 **李英睿 (LI, Yingrui)** [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, 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 OF-

FICE); 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, 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, PE, PG, PH, PL, PT, RO, RS, RU, 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。

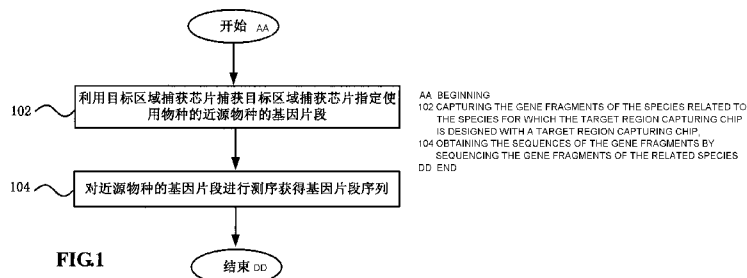
(84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (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 CAPTURING TARGET REGION AND METHOD AND SYSTEM FOR PROCESSING BIOINFORMATICS THEREOF

(54) 发明名称: 目标区域捕获方法及其生物信息处理方法和系统



(57) Abstract: The invention discloses a method for capturing target region, and a method and a system for processing bioinformatics thereof. Said method for capturing target region includes: capturing the gene fragments of the species related to the species for which the target region capturing chip is designed with a target region capturing chip, and obtaining the sequences of the gene fragments by sequencing the gene fragments of the related species. The methods and system of the invention help to screen the genome-wide data of the related species by using the gene chip designed for a single species in the target region capturing of other related species.

(57) 摘要: 本发明公开了一种目标区域捕获方法及其生物信息处理方法和系统。所述目标区域捕获方法包括利用目标区域捕获芯片捕获目标区域捕获芯片指定使用物种的近源物种的基因片段, 并对所述近源物种的基因片段进行测序获得基因片段序列。本发明的方法和系统, 将为单一物种设计的基因芯片用于其它近源物种的目标区域捕获, 从而帮助筛选近源物种的全基因组数据。



WO 2013/016864 A1

目标区域捕获方法及其生物信息处理方法和系统

技术领域

本发明涉及生物信息学技术领域，尤其涉及一种目标区域捕获方法及其生物信息处理方法和系统。

背景技术

新一代测序技术的发展，推动了模式物种的测序。全基因组测序成本一降再降，越来越多物种的基因组序列被破解，随之而来的大量基因信息急需大规模、高通量的分析方法和手段对其进行进一步研究。

20 世纪 90 年代初发展起来的基因芯片技术是大规模研究基因结构和功能的一种重要技术，配合高通量测序技术可以实现对某一物种特定目标区域，尤其是外显子进行捕获并测序，进而寻找功能基因并对其进行后续生物信息学分析。这种方法捕获效率高，成本相比全基因组测序较低，准确性也非常好。目前，国外学者利用目标区域捕获技术，使用外显子捕获芯片对人进行外显子测序，进而开展疾病相关研究，已取得一系列进展。

但是，由于目前各大芯片公司所推出的芯片都是针对模式物种的基因组来设计的，有许多已经有全基因组数据的物种并没有相应的芯片出售，因此，这些物种功能基因的获得仍然需要依赖传统的 PCR（聚合酶链式反应），经过大量的引物设计、合成、实验及优化，费时费力，工作量极大，全基因组数据也没有被很好的利用起来。

发明内容

本公开的一个方面要解决的一个技术问题是提供一种目标区

域捕获方法，能够提高物种的目标区域捕获的效率。

本公开的一个方面提供一种目标区域捕获方法，包括：

利用目标区域捕获芯片捕获目标区域捕获芯片指定使用物种的近源物种的基因片段；

对近源物种的基因片段进行测序获得基因片段序列。

根据本公开的目标区域捕获方法的一个实施例，该方法还包括：获取目标区域捕获芯片指定使用物种参考基因组上的捕获区域序列；根据捕获区域序列获取近源物种参考基因组上的近源区域。

根据本公开的目标区域捕获方法的一个实施例，该方法还包括：对基因片段序列进行过滤以去除不合格的基因片段序列。

本公开提供的目标区域捕获方法，通过将目标区域捕获芯片应用于该芯片指定使用物种的近源物种的目标区域的捕获，提高了没有目标区域捕获芯片的物种的目标区域捕获的效率。

本公开的另一个方面要解决的技术问题是提供一种生物信息处理方法和系统，能够提高处理效率。

本公开的另一个方面提供一种生物信息处理方法，包括：

利用目标区域捕获芯片捕获目标区域捕获芯片指定使用物种的近源物种的基因片段；

对近源物种的基因片段进行测序获得基因片段序列；

获取目标区域捕获芯片指定使用物种参考基因组上的捕获区域序列；

根据捕获区域序列获取近源物种参考基因组上的近源区域；

根据近源物种参考基因组和近源物种的近源区域，对基因片段序列进行生物信息分析。

根据本公开的生物信息处理方法的一个实施例，生物信息分析包括：组装、单核苷酸多态性位点寻找、拷贝数变异检测、插

入/缺失检测、染色体结构变异检测、SNP 位点与疾病关联分析、或 SNP 位点与药物作用相关分析。

本公开的又一个方面提供一种生物信息处理系统，包括：

捕获区域序列获取装置，用于获取目标区域捕获芯片指定使用物种参考基因组上的捕获区域序列；

近源区域获取装置，用于根据捕获区域序列获取指定使用物种的近源物种参考基因组上的近源区域；

基因片段捕获装置，用于利用目标区域捕获芯片捕获近源物种的基因片段；

测序装置，用于对近源物种的基因片段进行测序获得基因片段序列。

根据本公开的生物信息处理系统的一个实施例，该系统还包括过滤装置，用于对测序装置获得的基因片段序列进行过滤以去除不合格的基因片段序列。

根据本公开的生物信息处理系统的一个实施例，该系统还包括信息分析装置，用于根据近源物种的参考基因组和近源物种的近源区域对基因片段序列进行生物信息分析。

本公开提供的生物信息处理方法和系统，通过目标区域捕获芯片实现近源物种的目标区域捕获，根据芯片的区域捕获文件获得近源物种参考基因组上的近源区域，从而有助于后续的生物信息分析，提高了生物信息分析的效率。

附图说明

图 1 示出根据本发明目标区域捕获方法的一个实施例的流程图；

图 2 示出根据本发明目标区域捕获方法的另一个实施例的流程图；

图 3 示出根据本发明生物信息处理方法的一个实施例的流程图；

图 4 示出根据本发明的人基因组芯片捕获区域同源到猴子基因组上的分布情况图；

图 5 示出根据本发明的生物信息处理系统的一个实施例的结构图；

图 6 示出根据本发明的生物信息处理系统的另一个实施例的结构图。

具体实施方式

下面参照附图对本发明进行更全面的描述，其中说明本发明的示例性实施例。

图 1 示出根据本发明基因组目标区域捕获方法的一个实施例的流程图。

如图 1 所示，在步骤 102，利用目标区域捕获芯片捕获目标区域捕获芯片指定使用物种的近源物种的基因片段。近源物种之间的遗传距离较近，遗传物质的相似度高（特别是在外显子这样的高度保守区域上），近源物种之间功能类似的基因外显子中总会出现序列相似度极高甚至完全相同的保守区域，从而使现有的目标区域捕获芯片捕获其他近源物种的基因片段存在可能。利用指定物种的目标区域捕获芯片，按照芯片操作手册对要研究的近源物种（目标物种）的 DNA 样品进行目标区域捕获，捕获近源物种的基因片段。

在步骤 104，对近源物种的基因片段进行测序获得基因片段序列。例如，对捕获到近源物种的基因片段进行高通量测序获得基因片段序列，高通量测序技术可以采用 Illumina GA 测序技术，或者其他高通量测序技术。

上述实施例中，通过将目标区域捕获芯片应用于近源物种的目标区域捕获，实现了近源物种的目标区域捕获，将物种种类较少的芯片更好的利用起来，有助于获得更多物种个体功能基因及其后续生物信息分析。

图 2 示出根据本发明基因组目标区域捕获方法的另一个实施例的流程图。

在步骤 202，获取目标区域捕获芯片指定使用物种参考基因组上的捕获区域序列。例如，根据芯片厂商提供的目标区域捕获芯片的捕获区域文件，在该芯片的指定使用物种的参考基因组上截取对应的捕获区域序列。

在步骤 204，根据捕获区域序列获取近源物种参考基因组上的近源区域。进一步，根据近源物种的近源区域可以生成该目标区域捕获芯片对该近源物种的捕获区域文件。例如，将得到的芯片指定使用物种捕获区域序列，利用比对软件（例如，适合于局部比对的软件 Blast）与要研究的近源物种的参考基因组进行同源性比对，根据比对结果获取近源物种参考基因组上与捕获区域序列对应的近源区域。同源性比对过程允许出现错配，可以允许出现的错配率例如为 1% ~ 20%、或者 1%~10%，或者 5%左右。可根据不同的需求降低或升高错配率。例如，根据芯片指定使用物种与目标的近源物种之间的遗传距离决定错配率的范围，对于遗传距离大的物种之间接受较大的错配率，对于遗传距离小的物种之间接受较小的错配率。比对结果的内容例如包括：每一段捕获区域序列与近源物种的参考基因组的哪一段同源、同源性大小、这一片段在近源物种参考基因组上的具体坐标；根据坐标信息，得到该芯片在近源物种参考基因组中的可能与捕获区域序列对应的近源区域。进一步，可以根据近源物种的近源区域生成该芯片对近源物种的捕获区域文件。

在步骤 206，利用目标区域捕获芯片捕获目标区域捕获芯片指定使用物种的近源物种的基因片段。

在步骤 208，对近源物种的基因片段进行测序获得基因片段序列。

需要指出，步骤 202 和步骤 204 可以也可以在步骤 208 后执行，或者与步骤 206 和步骤 208 并行执行。

生物在进化的过程中，由于编码蛋白质的基因序列通常处于选择压力之下，因此，不同物种的基因中都有基因序列非常保守的区域，使得原本是为单一物种设计的基因芯片用于其它近源物种目标区域的捕获提供了可能。但是，目标区域捕获芯片仅适用于指定使用物种的观念，限制了普通技术人员去仔细分析为单一物种设计的基因芯片用于其它近源物种目标区域的捕获的可行性，以及具体实现的技术方案。目前为止，尚未发现公开发表用该思路做出来的研究成果。而且，上述可能性并不能使本领域的技术人员自然地将为单一物种设计的基因芯片用于其它近源物种目标区域的捕获，还需要克服一些技术问题以应用于生物信息分析。上述实施例中，通过目标区域捕获芯片提供的捕获区域文件帮助筛选近源物种的全基因组数据，从而获得近源物种的近源区域，进而获得目标区域捕获芯片对于该近源物种的捕获区域序列，大大减少了后续比对工作的工作量。

上述实施例中，利用同源性比对找到近源物种的捕获区域，从而可以精细定位捕获下来的序列片段在近源物种的参考基因组上的位置，方便后续分析，克服了通过目标区域捕获芯片获得近源物种的测序序列后无法进行后续分析的障碍。

根据本发明的一个实施例，在对近源物种的基因片段进行测序获得基因片段序列后，还对基因片段序列进行过滤以去除不合格的基因片段序列。不合格的基因片段序列例如包括：测序质量

低于低质量阈值的碱基个数超过整条序列碱基个数预定百分比（例如，40%以上，50%以上）的序列，低质量阈值由具体测序技术及测序环境而定；序列中测序结果不确定的碱基（如 Illumina GA 测序结果中的 N）个数超过整条序列碱基个数预定比例（例如 10%，20%）的序列；存在外源序列的基因片段序列，与其它实验引入的外源序列比对，如各种接头序列，若序列中存在外源序列则认为是不合格序列。将剩余合格的短片段序列中的接头序列去除。

在获得近源物种的基因片段序列后，可以根据近源物种的参考基因组和近源物种的近源区域，对基因片段序列进行生物信息分析。

图 3 示出根据本发明生物信息处理方法的一个实施例的流程图。

如图 3 所示，在步骤 302，利用指定物种的目标区域捕获芯片，按照芯片操作手册对近源物种的 DNA 样品进行目标区域捕获，获得近源物种的基因片段。

在步骤 304，对捕获到的基因片段进行高通量测序获得基因片段序列。其中，高通量测序技术可以为 Illumina GA 测序技术，也可以为现有的其他高通量测序技术。

在步骤 306，接收测序数据（基因片段序列），进行测序数据预处理。例如对基因片段序列进行过滤。

在步骤 308，根据芯片厂商提供的目标区域捕获芯片的捕获区域文件，在该芯片的指定使用物种的参考基因组上截取对应的捕获区域序列。

在步骤 310，将上述步骤中得到的芯片指定使用物种捕获区域序列，利用适合于局部比对的软件如 Blast，与想要研究的近源物种的参考基因组进行同源性比对，得到比对结果。

在步骤 312, 生成该芯片对近源物种的捕获区域文件。

在步骤 314, 根据目标物种的参考基因组及获得的目标物种的捕获区域文件, 对测序序列进行后续的生物信息分析。生物信息分析包括但不限于组装、单核苷酸多态性 (SNP) 位点寻找、拷贝数变异 (CNV) 检测、插入/缺失 (Indel) 检测、染色体结构变异 (SV) 检测、SNP 位点与疾病的关联分析、SNP 位点与药物作用的相关分析。

下面介绍本发明生物信息处理方法的一个应用例。在该应用例中:

近源物种 (目标物种) 样本: 2 个食蟹猴 (crab-eating macaque, 学名: *Macaca fascicularis*) 的 DNA 样本。

目标区域捕获芯片: 安捷伦公司 (38M) 人外显子芯片 (该芯片的商品名: Agilent 2100 Bioanalyzer, 货号/型号: G2938A)

芯片指定使用物种的设计对象: 人基因组 (Hg18, 即 NCBI BUILD36), 可以访问网页 <http://www.ncbi.nlm.nih.gov/sites/genome/>, 进入下面的 ftp 下载地址下载 hg18 数据: ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/ARCHIVE/

近源物种参考基因组: 食蟹猴 (CE) 参考基因组 (数据下载地址: http://climb.genomics.cn/Macaca_fascicularis)

实施例具体操作流程:

(1) 跨物种捕获并测序

利用安捷伦公司 (38M) 人外显子芯片, 对食蟹猴的 DNA 样品进行目标区域捕获及 Illumina GA 测序, 接收测序序列, 即基因片段序列;

(2) 过滤测序序列

接收到高通量测序序列后，对测序序列进行过滤，去除不合格的序列。不合格序列包括：测序质量值低于 20 的碱基个数超过整条序列碱基个数的 50%则认为是不合格序列；序列中测序结果不确定的碱基（Illumina GA 测序结果中的 N）个数超过整条序列碱基个数的 10%则认为是不合格序列；除样本接头序列外，与其它实验引入的外源序列比对，如各种接头序列。若序列中存在外源序列则认为是不合格序列。

（3）获取芯片捕获区域序列

根据芯片厂商安捷伦公司提供的该目标区域捕获芯片的捕获区域文件，在人的参考基因组（Hg18）上截取对应的捕获区域序列，将其写入到一个 fastq 文件 exoncapture.fa 中，其格式为：

```
>chr1.txt:20138:20258
```

```
CCAAAGTCCAGCAGTTGTCCCTCCTGGAATCCGTTG  
GCTTGCCTCCGGCATT TTTTGGCCCTTGCCTTTTAGGGTTG  
CCAGATTAAAAGACAGGATGCCCAGCTAGTTTGAATTTTA  
GATAA
```

```
>chr1.txt:58932:59892
```

```
AACGAGTGAAACGAATAACTCTATGGTGACTGAATTC  
ATTTTCTGGGTCTCTCTGATTCTCAGGAACTCCAGACCT  
TCCTATTTATGTTGTTTTTTGTATTCTATGGAGGAATCGTG  
TTTGGAACCTTCTTATTGTCATAACAGTGGTATCTGACT  
CCCACCTTCACTCTCCCAT
```

其中>所在的一行代表该捕获区域序列在 Hg18 上处于哪条染色体，以及它的起始、终止坐标，下面的一行是这一段捕获区域序列的具体基因型。

（4）芯片捕获区域序列与近源物种参考基因组的同源性比对

将得到的上述捕获区域序列文件 exoncapture.fa, 利用 Blast 比对软件, 与食蟹猴的参考基因组 CE.fa 进行同源性比对, 比对结果生成文件 out.psl, 其主要内容如表 1 所示:

match	114	
mismatch	5	
rep.match	0	
N's	0	
Q gap count	1	
Q gap bases	2	
T gap count	0	
T gap bases	0	
strand	-	
Q name	chr1.txt:20138:20258	
Q size	121	
Q start	0	
Q end	121	
T name	chr13	
T size	137686314	
T start	113205694	
T end	113205813	
... ..		

表 1

其中第一、二行表示同源性比对时候比对上与未比对上位点的个数, 第十列表示捕获区域序列名称, 第十四至十七列表示与该捕获区域同源的片段在食蟹猴参考基因组 CE.fa 上的具体染色体以及该染色体长度、同源片段起始、终止坐标。总结起来, 表

1 所代表的意义即为：芯片设计的在人参考基因组的 1 号染色体的 20138 bp 到 20258 bp 这一段区域，与食蟹猴参考基因组 13 号染色体上 113205694 bp 到 113205813 bp 这一段区域同源，食蟹猴参考基因组的 13 号染色体的总长为 137686314 bp。

根据芯片的捕获区域，截取 hg18 基因组片段，得到理论上的芯片捕获片段，将其与食蟹猴参考基因组做同源分析。结果发现 97.7% 的片段可以在食蟹猴参考基因组上找到同源区域，这种同源性表现为整段外显子片段的同源。剩下的 2.3% 找不到同源区域，估计可能是人类在进化中产生的特有片段。

图 4 示出人基因组芯片捕获区域同源到猴子基因组上的分布情况，直观地展示了芯片能抓到食蟹猴的基因片段位于猴子染色体的哪些地方（其中，能抓到的地方用灰色表示）。很明显灰色区域分布在猴子所有染色体上，而且比较均匀，说明人基因组外显子芯片是可以捕获到猴子各条染色体上的大量外显子片段。

(5) 生成芯片对近源物种的捕获区域文件

根据同源性比对结果文件 out.psl，得到安捷伦 38M 芯片在食蟹猴的参考基因组上的捕获区域，生成安捷伦 38M 芯片对食蟹猴的参考基因组的捕获区域文件，并按照染色体分开，生成文件 capture_region_chr*.txt，具体格式为：

染色体号	起始坐标	终止坐标
chr10	15525	16380
chr10	21092	21333
chr10	23712	23932
chr10	24665	24786
chr10	28859	29220
chr10	50626	50875
chr10	51197	51438

chr10	53645	53766
-------	-------	-------

表 2

其中，第一列表示染色体号，第二列表示起始坐标，第三列表示终止坐标。

(6) 后续生物信息分析

将第一步得到的测序序列 fastq 文件作为输入文件，以食蟹猴的参考基因组 CE.fa 为参考序列，以第 5 步得到的 capture_region_chr*.txt 文件作为捕获区域，利用外显子捕获流程 ECP (Exon Capture Pipeline) 用短序列比对软件 SOAP 进行比对，得到结果文件 pe.soap。

根据跑完 ECP 流程生成的 sample.report 文件，得到这只猴子的测序数据质量如下：

Mean Depth (均值深度)	Mismatch Rate (错配率)	Coverage (覆盖度)
25×	0.84%	0.85

表 3

上表中 coverage 是指实际捕捉的片段能覆盖理论上芯片捕获区域的百分数，即用人基因组芯片捕获到食蟹猴基因组上的片段能覆盖 85% 的芯片设计捕获区域。因为跑 ECP 流程时全部用的是人的参数，条件要求很高，只有比对得非常好而且比对上的次数足够多的才算成是覆盖到了，85% 的覆盖度，说明以现在的实验条件，有 85% 的芯片设计区域是可以很好的捕获下来的食蟹猴的相关基因区域。

综上分析，可以确定用针对人基因组的外显子芯片可以高效的获取大量食蟹猴的外显子片段，测序质量也较为稳定。

根据得到的结果与食蟹猴的参考基因组比对进行后续各种生物信息学分析。

上述应用例中，通过跨物种序列同源性比对，得到了人的目标区域捕获芯片对于其他物种的捕获区域，实现了一种跨物种目标区域捕获的方法，对人的近源物种进行目标区域捕获且捕获质量很高，从而将低成本、高准确性的目标区域捕获的研究方法扩展到了跨物种的层次。

近源物种分析一般限于人-黑猩猩这样遗传距离非常近的物种间的研究，但本发明方法的适用范围要宽得多，上述应用例中用到的人-食蟹猴的遗传距离就相对较远，甚至还可以将本发明的方法应用到人-小鼠的研究，是对本领域常规做法的突破。

图 5 示出根据本发明的生物信息处理系统的一个实施例的结构图。如图 5 所示，该生物信息处理系统包括：捕获区域序列获取装置 51，用于获取目标区域捕获芯片指定使用物种参考基因组上的捕获区域序列；近源区域获取装置 52，根据捕获区域序列获取装置 51 获取的捕获区域序列获取指定使用物种的近源物种参考基因组上的近源区域；基因片段捕获装置 53，利用目标区域捕获芯片捕获近源物种的基因片段；测序装置 54 对近源物种的基因片段进行测序获得基因片段序列。

图 6 示出根据本发明的生物信息处理系统的另一个实施例的结构图。和图 5 相比，该实施例中的生物信息处理系统除了包括捕获区域序列获取装置 51、近源区域获取装置 52、基因片段捕获装置 53，和测序装置 54，还包括信息分析装置 66。信息分析装置 66 根据近源物种的参考基因组和近源区域获取装置 52 获取的近源物种的近源区域对测序装置 54 获得的基因片段序列进行生物信息分析。生物信息分析例如包括：组装、单核苷酸多态性位点寻找、拷贝数变异检测、插入/缺失检测、染色体结构变异检测、SNP 位点与疾病关联分析、或 SNP 位点与药物作用相关分析。可选地，该生物信息处理系统还可以包括过滤装置 65，对测

序装置 54 获得基因片段序列进行过滤以去除不合格的基因片段序列，将过滤后的基因片段序列发送给信息分析装置 66 进行后续处理。

本发明的方法和系统是目前已有全基因组数据，但没有相应基因芯片的物种样品实现大规模目标区域捕获的优选技术方案。

对于图 5 至图 6 中各个装置或单元的功能，可以参考上文中关于本发明方法的实施例中对应部分的说明，为简洁起见，在此不再详述。

本领域的技术人员应当理解，对于图 5 至图 6 中的各个装置，可以通过单独的处理或计算设备实现，或者将其集成为一个独立的设备实现。在图 5 至图 6 中用框示出以说明它们的功能。部分功能块可以用硬件、软件、固件、中间件、微代码、硬件描述语音或者它们的任意组合来实现。举例来说，一个或者两个功能块都可以利用运行在微处理器、数字信号处理器（DSP）或任何其他适当计算设备上的代码实现。代码可以表示过程、功能、子程序、程序、例行程序、子例行程序、模块或者指令、数据结构或程序语句的任意组合。代码可以位于计算机可读介质中。计算机可读介质可以包括一个或者多个存储设备，例如，包括 RAM 存储器、闪存存储器、ROM 存储器、EPROM 存储器、EEPROM 存储器、寄存器、硬盘、移动硬盘、CD-ROM 或本领域公知的其他任何形式的存储介质。计算机可读介质还可以包括编码数据信号的载波。

本领域技术人员将意识到硬件、固件和软件配置在这些情况下的可替换性，以及如何最好地实现每个特定应用地该功能。

本发明的描述是为了示例和描述起见而给出的，而并不是无遗漏的或者将本发明限于所公开的形式。很多修改和变化对于本领域的普通技术人员而言是显然的。选择和描述实施例是为了更

好说明本发明的原理和实际应用，并且使本领域的普通技术人员能够理解本发明从而设计适于特定用途的带有各种修改的各种实施例。

权 利 要 求

1. 一种目标区域捕获方法，其特征在于，包括：

利用目标区域捕获芯片捕获所述目标区域捕获芯片指定使用物种的近源物种的基因片段；

对所述近源物种的基因片段进行测序获得基因片段序列。

2. 根据权利要求 1 所述的基因组目标区域捕获方法，其特征在于，还包括：

获取所述目标区域捕获芯片指定使用物种参考基因组上的捕获区域序列；

根据所述捕获区域序列获取所述近源物种参考基因组上的近源区域。

3. 根据权利要求 2 所述的方法，其特征在于，还包括：

对所述基因片段序列进行过滤以去除不合格的基因片段序列。

4. 根据权利要求 3 所述的方法，其特征在于，所述对所述基因片段序列进行以去除不合格的基因片段序列包括：

去除测序质量低于低质量阈值的碱基个数超过基因片段序列全部碱基个数预定百分比的基因片段序列；

和/或

去除测序结果不确定的碱基个数超过基因片段序列全部碱基个数预定比例的基因片段序列；

和/或

去除存在外源序列的基因片段序列。

5. 根据权利要求 2 所述的方法，其特征在于，所述获取目标区域捕获芯片指定使用物种参考基因组上的捕获区域序列包括：

根据所述目标区域捕获芯片的捕获区域文件在所述目标区域

捕获芯片指定使用物种的参考基因组上获取所述捕获区域序列。

6. 根据权利要求 2 所述的方法，其特征在于，所述根据所述捕获区域序列获取所述近源物种参考基因组上的近源区域包括：

将所述捕获区域序列与所述近源物种的参考基因组进行同源比对，获得所述近源物种参考基因组上的近源区域。

7. 根据权利要求 2 所述的方法，其特征在于，还包括：

根据所述近源物种参考基因组上的近源区域生成所述目标区域捕获芯片对所述近源物种的捕获区域文件。

8. 根据权利要求 1 所述的方法，其特征在于，所述利用目标区域捕获芯片捕获所述目标区域捕获芯片指定使用物种的近源物种的基因片段包括：

按照所述目标区域捕获芯片的操作手册对所述近源物种的 DNA 样品进行目标区域捕获，获得所述近源物种的基因片段。

9. 一种基于目标区域捕获的生物信息处理方法，其特征在于，包括权利要求 2 至 8 中任意一项所述的方法，还包括：

根据所述近源物种参考基因组和所述近源物种的近源区域，对所述基因片段序列进行生物信息分析。

10. 根据权利要求 9 所述的生物信息处理方法，其特征在于，所述生物信息分析包括：组装、单核苷酸多态性位点寻找、拷贝数变异检测、插入/缺失检测、染色体结构变异检测、SNP 位点与疾病关联分析、或 SNP 位点与药物作用相关分析。

11. 一种生物信息处理系统，其特征在于，包括：

捕获区域序列获取装置，用于获取所述目标区域捕获芯片指定使用物种参考基因组上的捕获区域序列；

近源区域获取装置，用于根据所述捕获区域序列获取所述指定使用物种的近源物种参考基因组上的近源区域；

基因片段捕获装置，用于利用目标区域捕获芯片捕获所述近

源物种的基因片段;

测序装置, 用于对所述近源物种的基因片段进行测序获得基因片段序列。

12. 根据权利要求 11 所述的系统, 其特征在于, 还包括:

过滤装置, 用于对所述测序装置获得所述基因片段序列进行过滤以去除不合格的基因片段序列。

13. 根据权利要求 11 所述的系统, 其特征在于, 还包括:

信息分析装置, 用于根据所述近源物种的参考基因组和所述近源物种的近源区域对所述基因片段序列进行生物信息分析。

14. 根据权利要求 13 所述的系统, 其特征在于, 所述生物信息分析包括: 组装、单核苷酸多态性位点寻找、拷贝数变异检测、插入/缺失检测、染色体结构变异检测、SNP 位点与疾病关联分析、或 SNP 位点与药物作用相关分析。

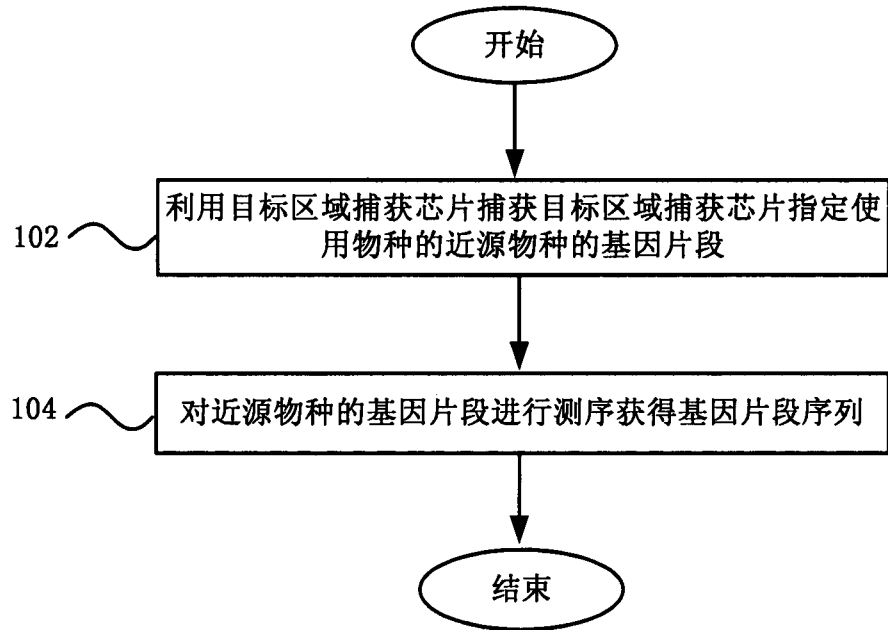


FIG.1

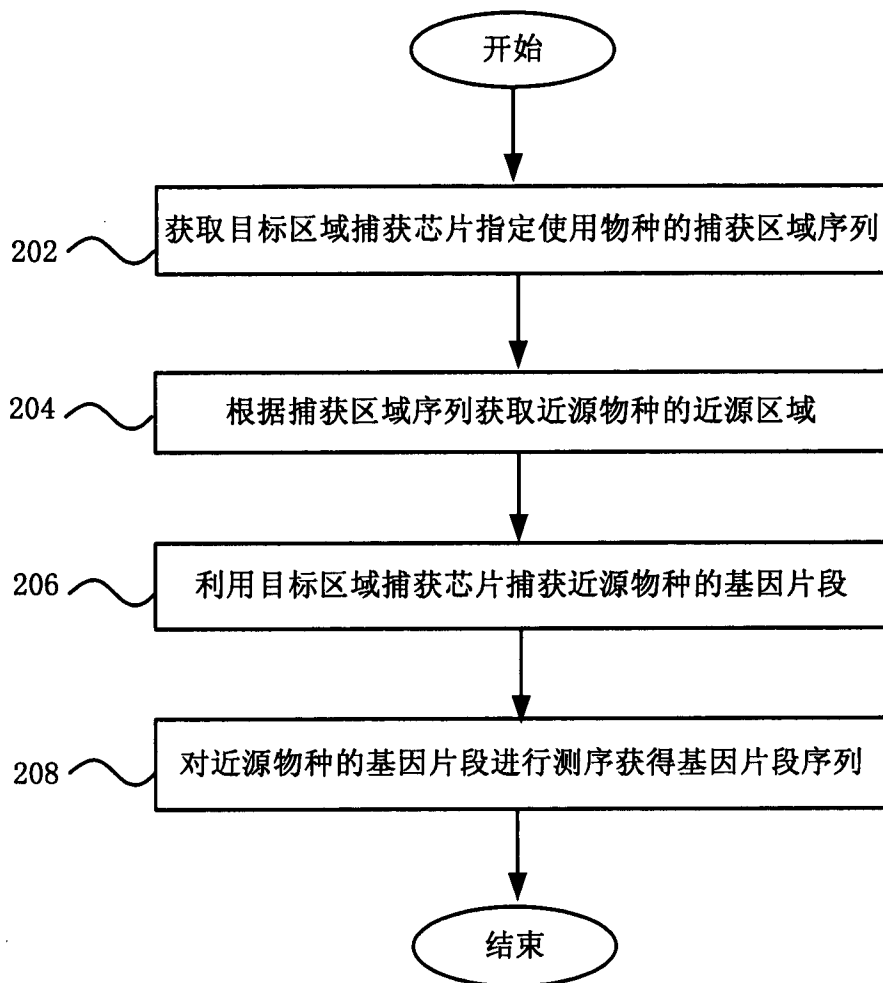


FIG.2

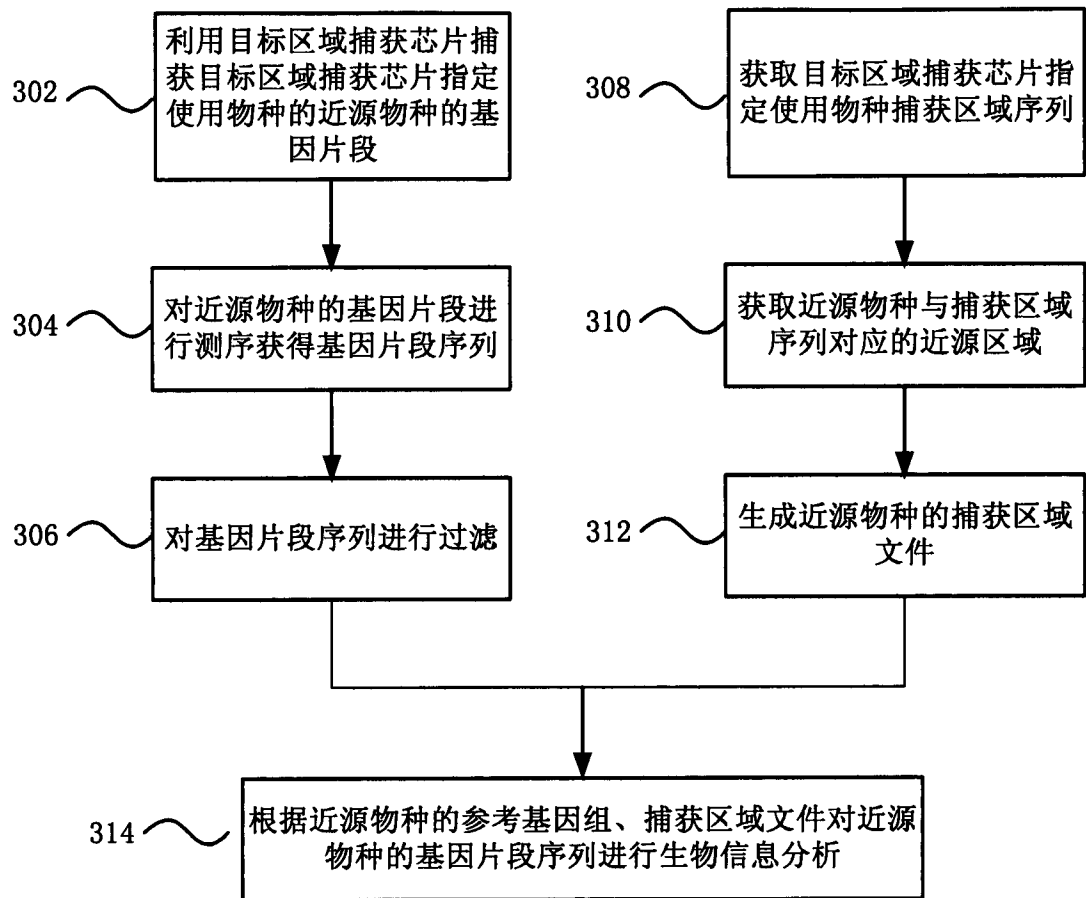
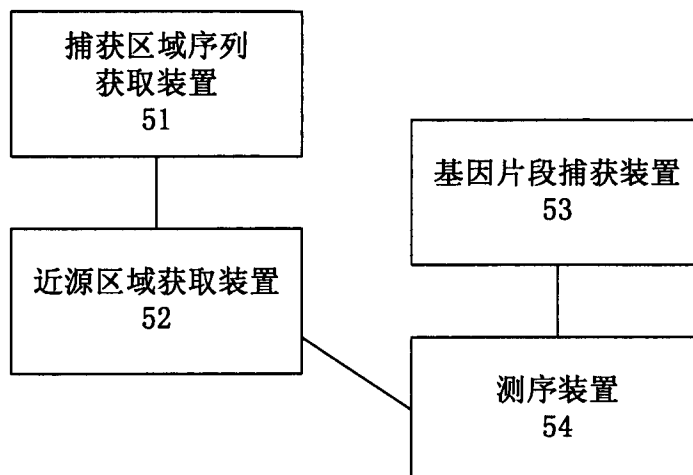
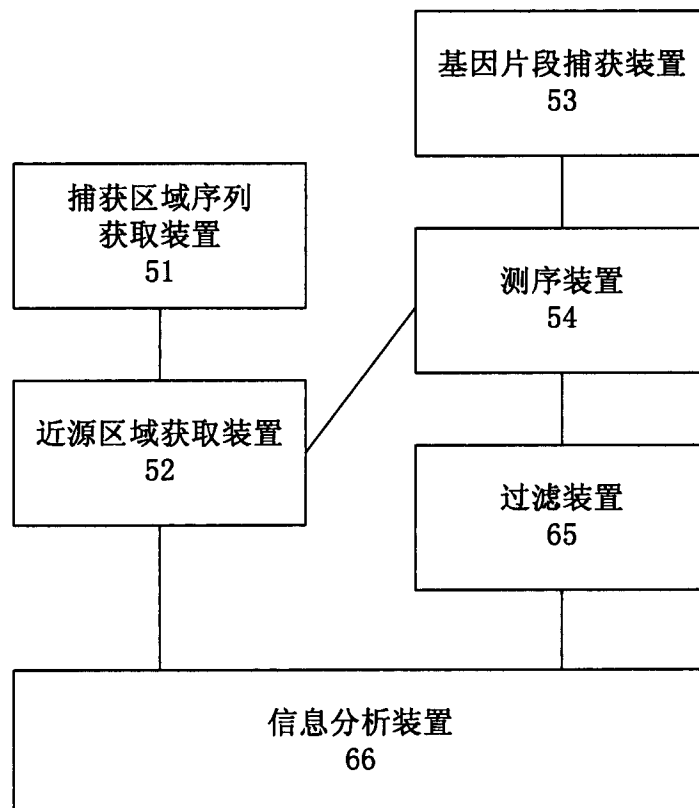


FIG.3



FIG.4

**FIG.5****FIG.6**

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2011/077861**A. CLASSIFICATION OF SUBJECT MATTER**

C12Q 1/68 (2006.01) 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)

IPC: C12Q

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)

VEN; CNABS; CPRS; CNKI; NCBI; ISI-WEB OF KNOWLEDGE; GOOGLE SCHOLAR: reference sequence, near source, monkey, species, cross-species, trans, capture, genome, genomic, reference, array, microarray, chip, human, target, sequencing, probe, relative

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	GEORGE, R.D. et al. Trans genomic capture and sequencing of primate exomes reveals new targets of positive selection. GENOME RES. July 2011, vol. 21, no. 10, pages 1686-1694, see the whole document, particularly abstract, figure 4, and pages 1688-1691	1-14
X	BURBANO, H.A. et al. Targeted investigation of the Neandertal genome by array-based sequence capture. SCIENCE. May 2010, vol. 328, no. 5979, pages 723-725, see the whole document, particularly abstract, and pages 723-724	1

☒ 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
25 April 2012 (25.04.2012)Date of mailing of the international search report
10 May 2012 (10.05.2012)Name and mailing address of the ISA/CN:
State Intellectual Property Office of the P. R. China
No. 6, Xitucheng Road, Jimengqiao
Haidian District, Beijing 100088, China
Facsimile No.: (86-10) 62019451

Authorized officer

TIAN, YuanTelephone No.: (86-10) **62411047**

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2011/077861

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	ALKAN, C. et al. Personalized copy number and segmental duplication maps using next-generation sequencing. NAT GENET. October 2009, vol. 41, no. 10, pages 1061-1067, see the whole document	1-14

A. 主题的分类

C12Q 1/68 (2006.01) 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)

IPC: C12Q

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))

VEN; CNABS; CPRS; CNKI; NCBI; ISI-WEB OF KNOWLEDGE; GOOGLE SCHOLAR: 物种, 跨物种, 捕获, 基因组, 参考序列, 芯片, 探针, 测序, 近源, 人, 猴, 目标, 靶, species, cross-species, trans, capture, genome, genomic, reference, array, microarray, chip, human, target, sequencing, probe, relative

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	Renee D. George 等. Trans genomic capture and sequencing of primate exomes reveals new targets of positive selection. Genome Res. 7 月 2011, 第 21 卷, 第 10 期, 第 1686-1694 页 参见全文, 特别是摘要, 图 4, 第 1688-1691 页	1-14
X	Hernán A. Burbano 等. Targeted investigation of the Neandertal genome by array-based sequence capture. Science. 5 月 2010, 第 328 卷, 第 5979 期, 第 723-725 页 参见全文, 特别是摘要, 第 723-724 页	1

☒ 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
25. 4 月 2012 (25. 04. 2012)

国际检索报告邮寄日期
10.5 月 2012 (10.05.2012)

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
田园
电话号码: (86-10) 62411047

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	Alkan C等.Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet.10月2009, 第41卷, 第10期, 第1061-1067页 参见全文	1-14