(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111128297 A
(43)申请公布日 2020.05.08

(21)申请号 201911310662.8

(22)申请日 2019.12.18

(71)申请人 中国科学院生物物理研究所
地址 100101 北京市朝阳区大屯路15号
申请人 中国科学院大学

(72)发明人 徐涛 周凯欣 王友 何顺民
陈飞 王静

(74)专利代理机构 北京元周律知识产权代理有限公司 11540
代理人 史冬梅

(51)Int.Cl.
G16B 20/20(2019.01)

权利要求书1页 说明书5页 附图1页

(54)发明名称

一种基因芯片的制备方法

(57)摘要

本申请公开了一种基因芯片的制备方法,包括:获取包含指定人群基因变异位点的第一数据集;利用次等位基因频率从所述第一数据集中提取候选基因变异位点,组成候选数据集;获取多个验证数据集,每个所述验证数据集包含一种类型的经过功能验证的基因变异位点;基于所述候选数据集和所述多个验证数据集确定多种芯片变异位点,多种所述芯片变异位点构成基因芯片。本发明所制备的基因芯片中包含了多种类型的基因变异位点,实现了一种芯片多种用途,提高了芯片的适用性。

CN 111128297 A

1.一种基因芯片的制备方法,其特征在于,包括:

获取包含指定人群基因变异位点的第一数据集;

利用次等位基因频率从所述第一数据集中提取候选基因变异位点,组成候选数据集;

获取多个验证数据集,每个所述验证数据集包含一种类型的经过功能验证的基因变异位点;

基于所述候选数据集和所述多个验证数据集确定多种芯片变异位点,多种所述芯片变异位点构成基因芯片。

2.根据权利要求1所述的基因芯片的制备方法,其特征在于,基于所述候选数据集和所述多个验证数据集确定多种芯片变异位点,具体为:

筛选出所述候选数据集与每个所述验证数据集中相同的基因变异位点,将筛选出的多种基因变异位点作为芯片变异位点。

3.根据权利要求1所述的基因芯片的制备方法,其特征在于,从所述第一数据集中提取候选基因变异位点,组成候选数据集,具体为:

设定次等位基因频率的第一预设阈值;

计算所述第一数据集中每一基因变异位点的次等位基因频率;

提取所述第一数据集中次等位基因频率大于第一预设阈值的编码区基因变异位点,记为第一编码区基因变异位点,将第一编码区基因变异位点加入候选数据集中。

4.根据权利要求1所述的基因芯片的制备方法,其特征在于,从所述第一数据集中提取候选基因变异位点,组成候选数据集,具体为:

设定次等位基因频率的第二预设阈值;

提取所述第一数据集中次等位基因频率大于第二预设阈值的基因变异位点,记为第一GWAS基因变异位点,将第一GWAS基因变异位点加入候选数据集中。

5.根据权利要求3或4所述的基因芯片的制备方法,其特征在于,所述验证数据集包括人类白细胞抗原变异位点验证数据子集。

6.根据权利要求5所述的基因芯片的制备方法,其特征在于,所述验证数据集还包括药物动力学变异位点验证数据子集。

7.根据权利要求6所述的基因芯片的制备方法,其特征在于,所述验证数据集还包括族源变异位点验证数据子集。

8.根据权利要求7所述的基因芯片的制备方法,其特征在于,还包括:

利用湿测试法从第一编码区基因变异位点中筛选出编码区芯片变异位点。

9.根据权利要求8所述的基因芯片的制备方法,其特征在于,还包括:

利用基因补充方法从第一GWAS基因变异位点中筛选出GWAS芯片变异位点。

10.根据权利要求9所述的基因芯片的制备方法,其特征在于,还包括:

获取线粒体变异位点数据集;

将线粒体变异位点数据集中的线粒体基因变异位点作为芯片变异位点。

一种基因芯片的制备方法

技术领域

[0001] 本申请涉及一种基因芯片的制备方法，属于生物医学技术领域。

背景技术

[0002] 随着人类基因组计划的顺利完成，开启了人类健康与生命科学研究的新时代。生物样本库的不断发展及技术的日趋成熟，更是为人类疾病尤其是重大慢性疾病的研究提供了丰富的样本资源及临床数据支撑。采用基因芯片技术对样本进行基因分型，通过队列基因数据的生物信息学分析去寻找特定的生物标志物，成为人类攻克一系列复杂疾病的强有力的技术手段。通过基因芯片技术获取基因分型数据，其宝贵价值也日益得到人们的理解与重视，世界各国政府及科研单位更是投入大量资源针对特定国家及地区的特定人群队列进行了诸多人群队列的基因分型工作。

[0003] 由于不同国家和地区的人群在基因型上有很大区别，所以在对样本进行基因分型时，所使用的基因芯片是有针对性的，其针对的是特定的国家和人群。现有技术中，并没有针对亚洲人群的基因芯片，同时，现有的基因芯片所覆盖的基因变异位点也较少，一种芯片只有一个用途，仅可用于检测具体的疾病，导致芯片的适用性差。

发明内容

[0004] 本发明的目的在于，提供一种覆盖基因变异位点较多的基因芯片的制备方法，以解决现有技术中，基因芯片适用性差的技术问题。

[0005] 本发明提供了一种基因芯片的制备方法，包括：

[0006] 获取包含指定人群基因变异位点的第一数据集；

[0007] 利用次等位基因频率从所述第一数据集中提取候选基因变异位点，组成候选数据集；

[0008] 获取多个验证数据集，每个所述验证数据集包含一种类型的经过功能验证的基因变异位点；

[0009] 基于所述候选数据集和所述多个验证数据集确定多种芯片变异位点，多种所述芯片变异位点构成基因芯片。

[0010] 其中，包含指定人群基因变异位点的第一数据集为基于指定人群的全基因组测序数据获得的基因变异位点组成的数据集。

[0011] 优选地，基于所述候选数据集和所述多个验证数据集确定多种芯片变异位点，具体为：

[0012] 筛选出所述候选数据集与每个所述验证数据集中相同的基因变异位点，将筛选出的多种基因变异位点作为芯片变异位点。

[0013] 优选地，从所述第一数据集中提取候选基因变异位点，组成候选数据集，具体为：

[0014] 设定次等位基因频率的第一预设阈值；

[0015] 计算所述第一数据集中每一基因变异位点的次等位基因频率；

[0016] 提取所述第一数据集中次等位基因频率大于第一预设阈值的编码区基因变异位点，记为第一编码区基因变异位点，将第一编码区基因变异位点加入候选数据集中。

[0017] 优选地，从所述第一数据集中提取候选基因变异位点，组成候选数据集，具体为：

[0018] 设定次等位基因频率的第二预设阈值；

[0019] 提取所述第一数据集中次等位基因频率大于第二预设阈值的基因变异位点，记为第一GWAS基因变异位点，将第一GWAS基因变异位点加入候选数据集中。

[0020] 优选地，验证数据集包括人类白细胞抗原变异位点验证数据子集。

[0021] 优选地，所述验证数据集还包括药物动力学变异位点验证数据子集。

[0022] 优选地，所述验证数据集还包括族源变异位点验证数据子集。

[0023] 优选地，还包括：

[0024] 利用湿测试法从第一编码区基因变异位点中筛选出编码区芯片变异位点。

[0025] 优选地，还包括：

[0026] 利用基因补充方法从第一GWAS基因变异位点中筛选出GWAS芯片变异位点。

[0027] 优选地，还包括：

[0028] 获取线粒体变异位点验证数据集；

[0029] 将线粒体变异位点验证数据集中包含的线粒体基因变异位点作为芯片变异位点。

[0030] 优选地，所述指定人群为中国人。

[0031] 本发明的基因芯片的制备方法，相较于现有技术，具有如下有益效果：

[0032] 本发明是针对中国人群特有的基因变异位点设计的一款基因芯片，它包含了编码区基因变异位点、GWAS基因变异位点、HLA基因变异位点、ADME基因变异位点，族源基因变异位点和线粒体变异位点。这些变异位点都是使用中国人群的全基因组测序数据筛选出来的。本发明使用全基因组测序数据作为基础数据集，可以获得整个基因组的数据，避免基因不全影响所制备的基因芯片的精确性，同时，由于全基因组测序数据为高分辨率数据，便于从中获取大型、小型全面的变异位点。

[0033] 本发明基因芯片中的变异位点包含了大量中国人群在编码区的变异位点，对编码区的变异位点的覆盖度达到了88％。

[0034] 本发明基因芯片中的变异位点包含了大量的中国人群GWAS基因变异位点，对GWAS基因变异位点中次等位基因频率在5％以上的变异位点的覆盖率达到了96％以上。

[0035] 本发明的基因芯片中包含的HLA基因变异位点是人体免疫系统疾病相关的变异位点，使用该基因变异位点可以很好的研究1型糖尿病等自身免疫型疾病。

[0036] 本发明的基因芯片中包含的ADME基因变异位点是与药物转运相关的变异位点，使用这些变异位点的信息可以研究新药的药效，指导新药研究和开发。

[0037] 本发明的基因芯片所包含的族源基因变异位点，可以利用这些变异位点将中国南方人和中国北方人区分开，并在此基础上研究南北方的饮食差异，南北方的进化等问题。

[0038] 本发明的基因芯片中包含的线粒体变异位点，可以用于研究线粒体相关的疾病。

附图说明

[0039] 图1为本发明一种基因芯片的制备方法的流程图。

具体实施方式

[0040] 本发明的基因芯片的流程图参见图1,其具体实施过程为:

[0041] 本实施例是以包含2641个中国人的30倍测序深度的全基因组测序数据为基础数据集。使用全基因组测序数据作为基础数据集,可以获得整个基因组的数据,避免基因不全影响后续制备的基因芯片的精确性,同时,由于全基因组测序数据为高分辨率数据,便于从中获取大型、小型全面的变异位点。本实施例使用中国人的全基因组测序数据,以便利用中国人的基因变异位点,制备针对中国人的基因芯片。

[0042] 首先,利用GATK工具从基础数据集中提取基因变异位点,得到原始数据集,本实施例中获得的原始数据集中总共有1亿个基因变异位点。GATK工具会对原始数据集中的基因变异位点进行标记,将各种变异位点进行区分,同时标记出满足标准的基因变异位点标记为PASS。筛选出标记为PASS的基因变异位点,该基因变异位点中包括单核苷酸多态性变异位点(SNP变异位点)和插入缺失变异位点,总量为七千七百万个。然后,再将标记为插入缺失变异位点的位点删除,保留单核苷酸多态性变异位点,获得七千五百万个单核苷酸多态性变异位点(SNP变异位点)。

[0043] 由于所筛选获得的SNP变异位点较多,需要对其进行质量控制。本实施例利用最大丢失率、次等位基因频率和最小质量值对SNP变异位点进行质量控制。最大丢失率阈值设定为0.5,次等位基因频率阈值设定为3,最小质量阈值设定为30。该步的筛选过程为:判断SNP变异位点中每一个变异位点的碱基丢失率,当丢失率大于最大丢失率阈值时,去除该变异位点,否则保留该变异位点。将SNP变异位点中每一个变异位点的碱基的次等位基因频率与设定次等位基因频率阈值进行比较,当某一个变异位点小于设定次等位基因频率阈值时,去除该变异位点,否则保留。判断SNP变异位点中每一个变异位点的碱基质量值参数,当质量值参数小于最小质量值阈值时,去除该变异位点,否则保留该变异位点。

[0044] 进一步地,设定最小测序深度阈值为3,去除SNP变异位点中小于最小测序深度阈值的变异位点;设定个体缺失率阈值为0.05,将2641个中国人的30倍测序深度的全基因组测序数据中每一个个体的基因缺失率与设定的缺失率阈值进行比较,当某一个个体的基因缺失率大于设定的缺失率阈值,去除该个体的基因数据,否则保留该个体的基因数据。执行该步骤的目的在于,进一步去除样本中缺失率较大,影响所获得的SNP变异位点精度的个体数据;设定哈德温伯格平衡参数阈值为0.000001,将SNP变异位点中每一个变异位点的哈德温伯格平衡参数与设定的哈德温伯格平衡参数阈值进行比较,当某一个变异位点的哈德温伯格平衡参数小于设定的哈德温伯格平衡参数阈值,去除该变异位点,否则保留。利用上述质量控制方法对2641个中国人的基因变异位点进行处理,最终获得一千八百万个SNP变异位点。所获得的SNP变异位点即为经过质量控制的高质量的基因变异位点。

[0045] 接下来,需要从高质量的SNP基因变异位点中筛选出制备基因芯片所需要使用的芯片变异位点,高质量的SNP基因变异位点组成第一数据集。

[0046] 首先,设定次等位基因频率的第一预设阈值为0.1%,设定次等位基因频率的第二预设阈值为1%。然后从高质量的SNP基因变异位点中,提取次等位基因频率大于等于0.1%的编码区基因变异位点,同时提取次等位基因频率大于等于1%的GWAS基因变异位点;得到700万的候选变异位点,组成候选数据集。

[0047] 然后使用affymetrix提供的湿测试方法从编码区基因变异位点中筛选出108269

个第一编码区变异位点作为芯片变异位点;使用affymetrix提供的基因补充方法(imputation)从GWAS基因变异位中筛选出405952第一GWAS变异位点作为芯片变异位点,共得到514221个芯片变异位点。

[0048] 由于affymetrix软件利用设定的条件所筛选出的基因变异位点具有局限性,所筛选出的基因变异位点覆盖并不全面,所以将affymetrix软件筛选后的剩余基因变异作为第二数据集,对该数据集进行进一步的筛选,以筛选出更为全面的基因变异位点。进一步筛选的具体步骤为:

[0049] 获取第二数据集中次等位基因频率大于等于3%的基因变异位点,组成聚类数据集;优选的,所选取的基因变异位点的次等位基因频率为5%以上。限定基因变异位点的次等位基因的目的在于,位于限定范围内的次等位基因,其包含的信息量更多,更利于制备基因芯片。如不限定次等位基因频率,则会导致数据集较大,增加处理时间及处理繁琐度。然后,获取所述聚类数据集中基因变异位点的连锁不平衡值,计算连锁不平衡值的过程为:

[0050] 获取所述聚类数据集中,每个基因变异位点与所述聚类数据集中的其他基因变异位点之间的皮尔逊相关系数r_{ij},其中,$0<i,j\leqslant N$,N为所述聚类数据集中基因变异位点的数量;然后根据所述皮尔逊相关系数r_{ij}确定连锁不平衡值r_{ij}^2。

[0051] 基于所获取连锁不平衡值,以$r_{ij}^2=0.6$作为阈值,对所述聚类数据集中的基因变异位点进行聚类,获得多簇基因变异位点。

[0052] 进一步,判断每簇中的每个基因变异位点是否包含于验证数据集中;如果簇中基因变异位点包含于验证数据集中,则基于验证数据集中的基因变异位点所使用的探针数量对簇中基因变异位点进行评分,所使用探针数量最少的基因变异位点评分最高,本实施例筛选出的基因变异位点为评分最高的基因变异位点。当然,也可以预设阈值,从而在每个簇中筛选出评分大于阈值的多个基因变异位点。本实施例中验证数据集为affymetrix提供的经过湿测试的基因变异位点数据集。该数据集中包含了很多在基因芯片上表现较好的基因变异位点。利用本发明的方法,经过上述步骤共获得了104866个GWAS基因变异位点。设计利用affymetrix软件筛选出的514221个基因变异位点的探针和利用本发明的方法筛选出的104866个基因变异位点的探针,将两种探针固定于同一个基片上,得到基因芯片。

[0053] 经过上述步骤所制备的基因芯片,其用途有限,不能用于检测自身免疫性疾病、线粒体相关疾病等,适用性较差,因此,需要在所制得的基因芯片上增加相关基因变异位点的探针,具体步骤为:

[0054] 使用affymetrix提供的验证过HLA(human leukocyte antigen,人类白细胞抗原)变异位点验证数据集与候选数据集中的6号染色体变异位点求交集,获取交集中包含的HLA基因变异位点作为芯片变异位点,设计该变异位点的探针,然后固定于上述基因芯片上。

[0055] 使用affymetrix与illumina提供的验证过的药物动力学变异位点,即ADME(absorption、distribution、metabolism、excretion)变异位点验证数据集与候选数据集求交集,获取交集中包含的ADME基因变异位点作为芯片变异位点,设计该变异位点的探针,然后固定于上述基因芯片上。

[0056] 使用illumina提供的验证过的族源变异位点验证数据集与候选数据集求交集,获取交集中包含的族源变异位点作为芯片变异位点,设计该变异位点的探针,然后固定于上

述基因芯片上。

[0057] 使用affymetrix提供的验证过的线粒体变异位点数据集作为芯片变异位点。设计该变异位点的探针，然后固定于上述基因芯片上。最终获得了包含了编码区基因变异位点、GWAS基因变异位点、HLA基因变异位点、ADME基因变异位点，族源基因变异位点和线粒体变异位点的基因芯片。

[0058] 本发明利用多种变异位点验证数据集分别与候选数据集求交集，获取交集中包含的多种基因变异位点作为芯片变异位点，所获得的芯片变异位点更加有代表性，且获得的芯片变异位点覆盖更加全面。

[0059] 本发明是针对中国人群特有的基因变异位点设计的一款基因芯片，它包含了编码区基因变异位点、GWAS基因变异位点、HLA基因变异位点、ADME基因变异位点，族源基因变异位点和线粒体变异位点。这些变异位点都是使用中国人群的全基因组测序数据筛选出来的。本发明使用全基因组测序数据作为基础数据集，可以获得整个基因组的数据，避免基因不全影响所制备的基因芯片的精确性，同时，由于全基因组测序数据为高分辨率数据，便于从中获取大型、小型全面的变异位点。

[0060] 本发明基因芯片中的变异位点包含了大量中国人群在编码区特有的变异位点，对编码区的变异位点的覆盖度达到了88％。

[0061] 本发明基因芯片中的变异位点包含了大量的中国人群特有的GWAS基因变异位点，对GWAS基因变异变异位点中次等位基因频率在5％以上的变异位点的覆盖率达到了96％以上。

[0062] 本发明的基因芯片中包含的HLA基因变异位点是人体免疫系统疾病相关的变异位点，使用该基因变异位点可以很好的研究1型糖尿病等自身免疫型疾病。

[0063] 本发明的基因芯片中包含的ADME基因变异位点是与药物转运相关的变异位点，使用这些变异位点的信息可以研究新药的药效，指导新药研究和开发。

[0064] 本发明的基因芯片所包含的族源基因变异位点，可以利用这些变异位点将中国南方人和中国北方人区分开，并在此基础上研究南北方的饮食差异，南北方的进化等问题。

[0065] 本发明的基因芯片中包含的线粒体变异位点，可以用于研究线粒体相关的疾病。

[0066] 以上所述，仅是本申请的实施例，并非对本申请做任何形式的限制，虽然本申请以较佳实施例揭示如上，然而并非用以限制本申请，任何熟悉本专业的技术人员，在不脱离本申请技术方案的范围内，利用上述揭示的技术内容做出些许的变动或修饰均等同于等效实施案例，均属于技术方案范围内。

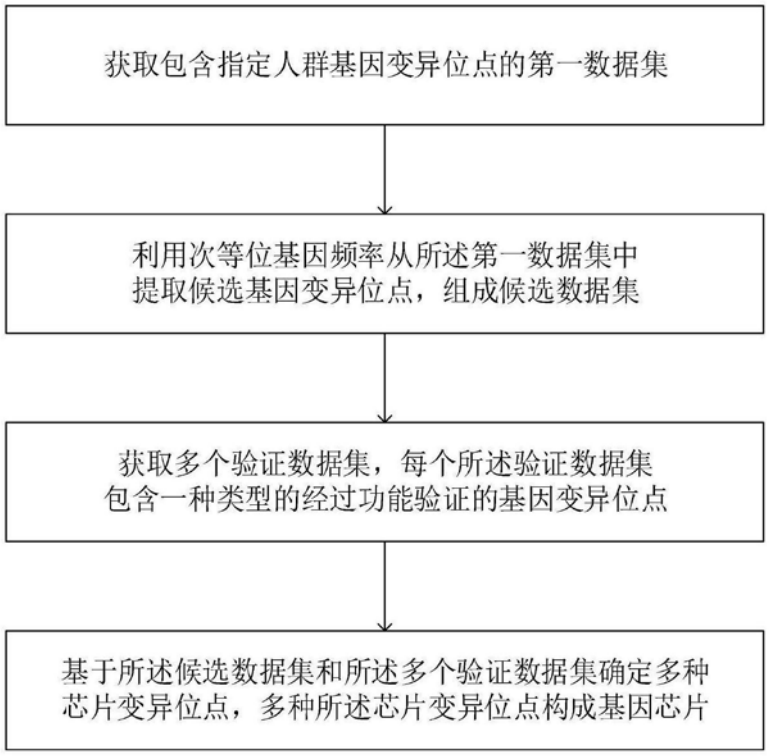
获取包含指定人群基因变异位点的第一数据集
利用次等位基因频率从所述第一数据集中
提取候选基因变异位点，组成候选数据集
获取多个验证数据集，每个所述验证数据集
包含一种类型的经过功能验证的基因变异位点
基于所述候选数据集和所述多个验证数据集确定多种
芯片变异位点，多种所述芯片变异位点构成基因芯片

图1