(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 105528532 B
(45)授权公告日 2019.08.16

(21)申请号 201410525810.9

(22)申请日 2014.09.30

(65)同一申请的已公布的文献号
申请公布号 CN 105528532 A

(43)申请公布日 2016.04.27

(73)专利权人 深圳华大基因科技有限公司
地址 518083 广东省深圳市盐田区北山路146号北山工业区综合楼11F-3

(72)发明人 李欣玥 刘栋兵 熊恒

(51)Int.Cl.
G16B 25/00(2019.01)
G16B 50/00(2019.01)

(56)对比文件
CN 101281561 A,2008.10.08,
US 2014/0143188 A1,2014.05.22,
CN 103266076 A,2013.08.28,
冯桂海.基于支持向量机的A-to-I RNA编辑的计算机识别及组织特异性研究.《中国优秀硕士学位论文全文数据库基础科学辑(月刊)》.2011,第15-18页.
王端青 等.基于转录组测序数据识别黑猩猩RNA编辑位点.《生物化学与生物物理进展》.2012,第39卷(第3期),第282-293页.

审查员 贾云杰

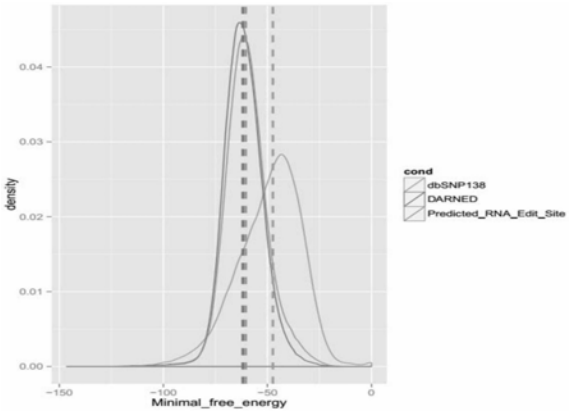
权利要求书1页 说明书10页 附图4页

(54)发明名称

一种RNA编辑位点的特征分析方法

(57)摘要

本发明提供了一种RNA编辑位点的特征分析方法,包括步骤:对待分析样品进行测序,获得DNA和RNA数据;分析获得的数据,得到RNA编辑位点数据集;统计获得所述RNA编辑位点数据集中RNA编辑位点上下游序列的RNA二级结构自由能分布曲线。该方法能够方便、快速地对RNA编辑位点数据的基本特征进行分析。



CN 105528532 B

1.一种RNA编辑位点的特征分析方法,其特征在于,包括步骤:

(1)对待分析样本进行测序,获得DNA和RNA数据;

(2)分析步骤(1)中获得的数据,得到RNA编辑位点数据集;

(3)统计获得所述RNA编辑位点数据集中RNA编辑位点上下游序列的RNA二级结构自由能分布曲线A;

(4)统计获得对照RNA编辑位点数据库中的RNA编辑位点上下游序列的RNA二级结构自由能分布曲线B,并将曲线A和曲线B进行比对,如果曲线A和曲线B大致重合,说明步骤(2)中所获得的RNA编辑位点数据集较为可靠;

所述RNA二级结构自由能分布曲线的中位数位于-55～-70kcal/mol。

2.如权利要求1所述的方法,其特征在于,所述“上下游序列”的长度为50bp－200bp。

3.如权利要求2所述的方法,其特征在于,所述“上下游序列”的长度为100bp。

4.如权利要求1所述的方法,其特征在于,所述RNA二级结构自由能分布曲线的中位数位于-60～-65kcal/mol。

5.如权利要求1所述的方法,其特征在于,所述方法还包括步骤:

(a)统计RNA编辑位点数据集中单编辑位点的编辑频率,选取差异显著的位点进行FDR矫正,获得具有显著差异的位点作为后续分析的候选位点;

(b)对RNA编辑位点数据集进行两类样本单个基因编辑位点统计,并以该统计获取两类样本之间编辑位点数差异在0.5倍以上的基因及两类样本各自独有的发生编辑的基因,供后续进行目的基因的筛选;

在所述步骤(a)中,对RNA编辑位点进行两类样本单编辑位点编辑频率的统计,并以该频率进行成对t检验,获取每个位点的差异显著性P值,选取差异显著的位点进行FDR过滤,获得在两类样本中具有显著差异的位点,作为后续分析的候选位点;

其中所述差异显著的位点是指P<0.05的位点,并且进行FDR过滤时设置P<0.05。

6.如权利要求1所述的方法,其特征在于,所述方法还包括步骤:

统计所有样本检出的编辑位点上下游各10bp位置的各碱基出现频率。

7.如权利要求5所述的方法,其特征在于,所述步骤(b)中所述两类样本为肿瘤样本和对应正常样本。

8.如权利要求1所述的方法,其特征在于,所述步骤(3)中使用的统计工具为RNAfold软件。

9.如权利要求1所述的方法,其特征在于,所述步骤(1)中待分析样本为群体样本,所述群体样本中样本数量≥50个,合并测得的DNA和RNA数据进行步骤(2)。

10.如权利要求1所述的方法,其特征在于,所述步骤(1)中待分析样本包括正常组织和/或肿瘤组织。

11.如权利要求5所述的方法,其特征在于,所述两类样本是癌症样本和对应正常样本。

一种RNA编辑位点的特征分析方法

技术领域

[0001] 本发明属于生物技术领域，具体地说，本发明涉及一种RNA编辑位点的特征分析方法。

背景技术

[0002] RNA编辑是指DNA转录之后、翻译之前在RNA水平上发生的碱基的缺失、插入或置换。在高等生物中，最主要的RNA编辑是碱基A到I(次黄嘌呤核苷)的修饰，这种修饰，通常是被ADAR蛋白酶催化产生的。由于在翻译水平上，次黄嘌呤核苷酸(I)被识别为鸟核苷酸(G)，因此在该位点的这种编辑，实际上是A到G的转换。这种改变可能导致相关蛋白质结构功能的改变，也可能改变生物体内起调控作用的RNA的结构功能的改变。根据相关文献报道表明，RNA编辑现象与癌症的发生有密切联系，因而成为了当前癌症方面研究的一个新的研究思路及研究热点。

[0003] 由于实验技术需求较多的资源的投入，当前的RNA编辑方面的研究着重于以信息学的方式进行RNA编辑位点的鉴定的发掘，并在此基础上进行了RNA编辑位点的特征统计(基因组上分布，序列模体等)及后续的一些分析工作。当前癌症方面RNA编辑的分析工作主要集中在分析基因编码区，特别是外显子区的非同义突变的研究。这主要是因为这种方式的编辑可以比较直观的反映到对基因表达产物的影响。但从现有文献中已经鉴定出的RNA编辑位点分布情况来看，这种发生在基因编码区的RNA编辑，只占总体RNA编辑位点中极少比例的一部分，更多的RNA编辑位点发生在基因的内含子区及被称为Alu的SINE(Short Interspersed Nuclear Element)区域。

[0004] 上述情况表明，RNA编辑的真正调控作用应该以以上两种区域的作用及特点密不可分。这将是之后对RNA编辑方面研究的重点思路之一。与其他的DNA变异不同(如snp，indel鉴定等)，RNA编辑的生物学鉴定及分析尚处于起步阶段，因此，缺乏统一的分析思路及相关的软硬件支持，这导致大量的精力被投入到了重复性的工作中。

[0005] 因此，对RNA编辑方面的分析，迫切需要一些较为完善的技术方案对RNA编辑位点数据进行基本特征分析，使得为RNA编辑方面的研究更为方便、快速、准确。

发明内容

[0006] 本发明的目的在于提供一种RNA编辑位点的特征分析方法。

[0007] 本发明的第一方面，提供了一种RNA编辑位点的特征分析方法，包括步骤：

[0008] (1)对待分析样本进行测序，获得DNA和RNA数据；

[0009] (2)分析步骤(1)中获得的数据，得到RNA编辑位点数据集；

[0010] (3)统计获得所述RNA编辑位点数据集中RNA编辑位点上下游序列的RNA二级结构自由能分布曲线A；优选地，所述“上下游序列”的长度为50bp－200bp；更优选地为100bp。

[0011] 在另一优选例中，所述RNA二级结构自由能分布曲线的中位数位于-55～-70；优选地位于-60～-65。

[0012] 在另一优选例中，所述方法还包括步骤：

[0013] (4)统计获得对照RNA编辑位点数据库中的RNA编辑位点上下游序列的RNA二级结构自由能分布曲线B，并将曲线A和曲线B进行比对。如果曲线A和曲线B大致重合，说明步骤(2)中所获得的RNA编辑位点数据集较为可靠。

[0014] 在另一优选例中，所述方法还包括步骤：

[0015] (a)统计RNA编辑位点数据集中单编辑位点的编辑频率，选取差异显著的位点进行FDR矫正，获得具显著差异的位点作为后续分析的候选位点；

[0016] (b)对RNA编辑位点数据集进行两类样本单个基因编辑位点统计，并以该统计获取两类样本编辑位点数变化差异在较大(较佳地，差异变化在0.5倍以上)的及两类样本各自独有的发生编辑的基因，供后续进行目的基因的筛选。

[0017] 在另一优选例中，所述方法还包括步骤：

[0018] 统计所有样本检出的编辑位点上下游各10bp位置的各碱基出现频率。

[0019] 在另一优选例中，所述步骤(b)中所述两类样本为肿瘤样本和对应正常样本。

[0020] 在另一优选例中，所述步骤(3)中使用的统计工具为RNAfold软件。

[0021] 在另一优选例中，所述步骤(1)中待分析样本为群体样本，所述群体样本中样本数量≥50个，合并测得的DNA和RNA数据进行步骤(2)。

[0022] 在另一优选例中，所述步骤(1)中待分析样本包括正常组织和/或肿瘤组织。

[0023] 在另一优选例中，所述样本选自：正常人或癌症患者。

[0024] 在另一优选例中，所述步骤(a)中，对RNA编辑位点进行两类样本(比如，癌症样本和对应正常样本)单编辑位点编辑频率的统计，并以该频率进行成对t检验，获取每个位点的差异显著性值(P值)，选取差异显著的点(如$P<0.05$)进行FDR过滤(设置$P<0.05$)，获得在两类样本中具显著差异的位点，作为后续分析的候选位点。

[0025] 在另一优选例中，所述方法包括步骤：

[0026] 进行两类样本及DARNED数据库的RNA编辑位点绘制维恩图。

[0027] 应理解，在本发明范围内中，本发明的上述各技术特征和在下文(如实施例)中具体描述的各技术特征之间都可以互相组合，从而构成新的或优选的技术方案。限于篇幅，在此不再一一累述。

附图说明

[0028] 图1显示了实施例1中数据库预测RNA编辑位点，snp位点二级结构最小自由能分布图(虚线为中位数)。

[0029] 图2显示了实施例1中RNA编辑位点上下游各10bp特征情况图。

[0030] 图3显示了实施例1中正常样本，肿瘤样本，DARNED数据库编辑位点韦恩图。

[0031] 图4显示了实施例2中数据库预测RNA编辑位点，snp位点二级结构最小自由能分布图(虚线为中位数)。

[0032] 图5显示了实施例2中RNA编辑位点上下游各10bp特征情况图。

[0033] 图6显示了实施例2中正常样本，肿瘤样本，DARNED数据库编辑位点韦恩图。

具体实施方式

[0034] 本发明人通过广泛而深入的研究，获得一种RNA编辑位点的特征分析方法，实验结果表明，所述方法能够方便、快速地对RNA编辑位点数据的基本特征进行分析，并得出准确的结果。

[0035] 测序

[0036] 在本发明中，可用常规的测序技术和平台进行测序。优选的测序方法包括：Life Technologies的proton或PGM，Illumina HiSeq，ABI SOLiD，Roche 454等测序仪器。

[0037] 在本发明中，特别适合对本发明构建的PCR－free文库进行测序的方法是Ion Proton法。在一优选例中，将符合上机测序标准的文库片段，使用The Ion Proton™System进行测序。

[0038] 数据处理

[0039] 在本发明的优选例中，数据处理通常包括以下步骤：以NCBI数据库中公布的人基因组为参考标准。将测序的reads转换为fastq格式，并与人基因组序列比对，确定匹配的读序(即比对上的读序)。

[0040] 数据处理可以用本领域采用的方法或软件进行，包括市售的软件、公开的软件(尤其是全部开源的软件)进行。

[0041] RNA编辑位点样本的获得

[0042] 目前公开的RNA编辑位点数据库包括：DARNED数据库(网址：http://darned.ucc.ie/)、RADAR数据库(网址：http://rnaedit.com/)，可以作为对照数据库。通过上述的数据库也可以获得本发明所涉及的待分析的RNA编辑位点数据。

[0043] 此外，对于群体样本RNA编辑位点数据的获得，可以采用如下方法。

[0044] 针对Illumina测序平台生产的高通量测序数据，所包括的RNA编辑位点检测方法，步骤如下：

[0045] (1)比对

[0046] (1.1)获得原始测序数据，所述原始测序数据为群体样本的测序数据；

[0047] 在本发明的一个较佳实施方式中，所述原始测序数据包括正常DNA、肿瘤DNA、正常RNA、肿瘤RNA的高通量测序数据；

[0048] (1.2)原始数据过滤，目的是过滤掉一些含有接头或者质量值比较低的片段，获得“干净的”数据；主要内容有：

[0049] (i)去除含接头的片段；当片段被接头污染时，可能测到接头序列，所以要

[0050] 除接头；

[0051] (ii)去除N的比例较高(优选地比例≥10％)的片段，N含量过高会引起比

[0052] 对错误；

[0053] (iii)去除低质量片段，测序时存在测错的概率，低质量的片段有可能存在

[0054] 测错的碱基。

[0055] (1.3)比对，利用基于Bowtie的RNA-seq比对工具tophat将测序数据比对到参考基因组上，生成bam格式的文件。

[0056] (1.4)使用GATK(Genome Analysis Toolkit)对比对结果的碱基质量值校正。illumina测序结果在给定每个碱基质量值的时候存在偏差，需要根据整个文库所有测序

reads的质量值分布进行校正。

[0057] (1.5)利用Picard工具包去除比对结果中存在的PCR重复序列。

[0058] (1.6)使用GATK分割比对结果中存在剪切的序列(含N的片段)。

[0059] (2)使用GATK的UnifiedGenotyper工具检测突变,分别对正常RNA、肿瘤RNA、正常DNA和肿瘤DNA四组bam文件进行突变检测,得到正常RNA、肿瘤RNA、正常DNA和肿瘤DNA一共4个vcf文件,作为原始RNA编辑位点数据(原始SNP)。

[0060] (3)过滤突变

[0061] (3.1)使用GATK对检测出来的SNP做VQSR(Variant quality score recalibration),对vcf(Variant Call Format)文件中的一些高质量的位点作为可信位点构建高斯混合模型(Gaussian mixture model),并对所有位点进行评估,从而过滤其中的假阳性位点,具体操作可以参考软件说明。

[0062] (3.2)分别移除DNA和RNA、RNA和dbSNP数据库共有的位点,因这些位点并不是在转录过程中发生的突变,不属于RNA编辑事件,需要排除。

[0063] (3.3)移除RNA检测的indel(插入或缺失)位点左右各30bp(base pair)内的位点,由于indel附近容易发生比对错误,造成较高的假阳性,因此将INDEL附近的位点排除掉。

[0064] (3.4)以深度大于2且突变支持数大于1作为一个可信的发生编辑的样本,若该组样本的编辑样本支持数少于2个,则视为假编辑位点滤掉。

[0065] (3.5)过滤掉FS(Phred-scaled p-value using Fisher's exact test to detect strand bias)大于20的位点。

[0066] (3.6)移除基因间区以及处在剪切位点左右2bp内的的位点,由于处在这些区域的位点的突变并不会对基因表达产物产生直接影响,因此也需要过滤掉。

[0067] 最终得到高质量的在基因区的RNA编辑位点。

[0068] 其中,步骤1.3)中基于公开的开源Bowtie的RNA-seq比对工具tophat(下载地址如:http://ccb.jhu.edu/software/tophat/index.shtml),进行比对,命令行如下:

[0069] tophat--solexa1.3-quals--read-mismatches 2--read-gap-length 3--read-edit-dist 3--library-type fr-unstranded-p 6-r 30--b2-fast--rg-center bgi--rg-platform illumina--no-novel-juncs--no-novel-indels-o dir reference sequence.fq1 sequence.fq2

[0070] 步骤1.4)中,使用公开的开源GATK(Genome Analysis Toolkit)软件(下载地址如:https://www.broadinstitute.org/gatk/),校正参数为-knownSites-nct-U-BQSR,GATK的软件使用可以参考产品使用说明。

[0071] 步骤1.5)中,公开的开源Picard工具包(下载地址如:http://picard.sourceforge.net/)去除比对结果中存在的PCR重复序列,设置如下:

[0072] java-Xmx4g-jar MarkDuplicates.jar INPUT＝in.bam OUTPUT＝out.bam METRICS_FILE＝rmdup.met REMOVE_DUPLICATES＝true VALIDATION_STRINGENCY＝SILENT ASSUME_SORTED＝true CREATE_INDEX＝true。

[0073] 步骤1.6)中,GATK的设置如下:

[0074] java-Xmx512M-jar FilterBadCigar.jar in.bam out.bam java-Xmx6g-jar GenomeAnalysisTK.jar-T SplitNCigarReads-I in.bam-o out.bam-U ALL-R

reference.fa。

[0075] 步骤(1.2)中,GATK的UnifiedGenotyper工具的设置如下:

[0076] java-Xmx6g-jar-Djava.io.tmpdir＝tmp GenomeAnalysisTK.jar-T UnifiedGenotyper-l INFO-I bam.list-R reference.fa--dbsnp dbsnp_138-stand_call_conf 30-stand_emit_conf 4-dcov 200-G Standard-nt 6-glm BOTH-U ALLOW_N_CIGAR_READS-L chr-metrics metrics-o chr.vcf

[0077] 步骤3.1)中,VQSR(Variant quality score recalibration)是指:对vcf(Variant Call Format)文件中的一些高质量的位点作为可信位点构建高斯混合模型(Gaussian mixture model),并对所有位点进行评估,从而过滤其中的假阳性位点;

[0078] 主要步骤:(i)对vcf(Variant Call Format)文件中的一些高质量的位点作为可信位点构建高斯混合模型(Gaussian mixture model),并对所有位点进行评估;(ii)将建立的高斯混合模型参数应用到输入的VCF文件,对每一个变异位点进行VQSLOD值的注释,从而过滤其中的假阳性位点。

[0079] VQSR通过机器学习的方法根据已知的变异位点训练出一组变异位点集,并会给每个位点赋一个VQSLOD值,变异位点越接近集合的中心其值就会越高;然后根据模型在对新检测出的变异位点进行打分,如果分值在训练集合内就认为是一个质量高的变异位点,否则认为是一个假阳性位点。

[0080] 步骤(3.5),FS(Phred-scaled p-value using Fisher's exact test to detect strand bias)使用Fish检验的方法,检测比对在某一位点片段是否存在链的偏好性。

[0081] 另外,现有技术已经公开了许多常规的获得RNA编辑位点的方法,例如文献Accurate identification of A-to-I RNA editingin human by transcriptome sequencing、RNA editing in the human ENCODE RNA-seq data、High levels of RNA-editing site conservation amongst 15 laboratory mouse strains中报道的方法,具体请见附录的参考文献。

[0082] RNA编辑位点的特征分析

[0083] 本发明对RNA编辑位点的特征进行分析,包括:

[0084] 1)对群体RNA编辑位点进行两类样本单编辑位点编辑频率的统计,并以该频率进行成对t检验,获取每个位点的差异显著性值(P值),选取差异显著的点(参数可修改,默认P<0.05)进行FDR过滤(参数可修改,默认P<0.05),获得较为可靠的在两类样本(比如,癌症样本和对应正常样本)中具显著差异的位点。这些位点可作为后续分析的候选位点。

[0085] 2)对群体RNA编辑位点进行两类样本单个基因编辑位点统计,并以该统计获取两类样本编辑位点数变化差异在较大(参数可修改,默认差异变化在0.5倍以上)的及两类样本各自独有的发生编辑的基因,供后续进行目的基因的筛选。

[0086] 3)统计所有样本检出的编辑位点上下游各10bp位置的各碱基出现频率,并绘图,可直观看到RNA编辑位点模体(motif)特征。

[0087] 4)统计所有样本检出的编辑位点上下游各100bp位置序列的RNA二级结构自由能分布,并进行绘图,同时也对dbsnp138及DARNED数据库的位点上下游各200bp位置序列的RNA二级结构自由能分布进行。

[0088] 5)进行两类样本及DARNED数据库的编辑位点维恩图的绘制。

[0089] 本发明的主要优点在于：

[0090] (1)首次披露了一种RNA编辑位点的特征分析方法，该方法能够方便、快速地对RNA编辑位点数据的基本特征进行分析；

[0091] (2)采用本发明的方法对种RNA编辑位点数据进行分析，分析结论准确、可靠。

[0092] (3)使用本方法可以方便的判别分析获得RNA编辑位点数据的准确性，并鉴别出RNA编辑位点数据和SNP位点数据。

[0093] 实施例1

[0094] 1.样本/数据来源

[0095] 1.1 65例前列腺癌患者，对每个患者的正常DNA、肿瘤DNA、正常RNA、肿瘤RNA分别进行高通量测序，读长为90bp，分析获取群体的RNA编辑位点数据和SNP位点数据，得到VCF格式的RNA编辑位点及相应注释信息。

[0096] 1.2 Darned数据库(网址:http://darned.ucc.ie/)

[0097] 2.分析处理RNA编辑位点数据

[0098] 本实施例中使用RNAfold软件对RNA编辑位点进行特征分析，RNAfold软件为开源软件，下载地址如:http://www.tbi.univie.ac.at/RNA/index.html#download。

[0099] 为了便于说明，表1中列出了本实施例中的生成文件及说明。

[0100] 表1本实施例中生成文件及说明

[0101]

所在目录	文件名	说明
frequency	normal.fre.overlap.txt	正常样本与肿瘤样本 RNA 编辑重叠位点编辑频率（正常样本部分）
	tumor.fre.overlap.txt	正常样本与肿瘤样本 RNA 编辑重叠位点编辑频率（肿瘤样本部分）
	normal.frequency.txt	正常样本 RNA 编辑位点编辑频率

[0102]

	tumor.frequency.txt	肿瘤样本 RNA 编辑位点编辑频率
	normal.fre.uniq.txt	正常样本特有 RNA 编辑位点编辑频率
	tumor.fre.uniq.txt	肿瘤样本特有 RNA 编辑位点编辑频率
	raw_P_value_result.txt	正常样本与肿瘤样本 RNA 编辑重叠位点编辑频率成对 t 检验 P 值列表
	P_value_result_filtered.txt	t 检验 P 值过滤正常样本与肿瘤样本 RNA 编辑重叠位点编辑频率成对 t 检验 P 值列表
	P_value_result_filtered_sorted.txt	t 检验 P 值过滤正常样本与肿瘤样本 RNA 编辑重叠位点编辑频率成对 t 检验 P 值列表（已排序）
	P_value_result_filtered_sorted_fdr_filtered.txt	fdr 矫正 P 值过滤正常样本与肿瘤样本 RNA 编辑重叠位点编辑频率（可作为候选研究位点）
GeneEdit	normal.gene.edit.txt	正常样本各基因编辑位点数统计
	tumor.gene.edit.txt	肿瘤样本各基因编辑位点数统计
	Normal.uniq.edit.gene.txt	正常样本各特有编辑基因编辑位点数统计
	Tumor.uniq.edit.gene.txt	肿瘤样本各特有编辑基因编辑位点数统计
	Tumor.Normal.overlap.edit.gene.txt	正常样本肿瘤样本重叠各基因编辑位点数统计
	Tumor.Normal.overlap.edit.gene.filtered.txt	过滤后正常样本肿瘤样本重叠各基因编辑位点数统计（可作为候选研究基因）
sequence_feature	tumor.sequence.txt	肿瘤样本 RNA 编辑位点前后各 10bp 序列
	normal.sequence.txt	正常样本 RNA 编辑位点前后各 10bp 序列
	whole.site.seuqence.txt	所有 RNA 编辑位点前后各 10bp 序列
	log.txt	未提取出序列的位点信息
	sequence.matrix.txt	提取出序列各位置各碱基出现频率统计
	sequence_feature.pdf	RNA 编辑位点上下游各 10bp 特征情况图

[0103]

sequence_structure	normal.sequence.txt	正常样本 RNA 编辑位点前后各 100bp 序列
	tumor.sequence.txt	肿瘤样本 RNA 编辑位点前后各 100bp 序列
	whole.site.seuqence.txt	所有 RNA 编辑位点前后各 100bp 序列
	log.txt	未提取出序列的位点信息
	predicted.database.site.txt	DARNED 数据库 RNA 编辑位点前后 100bp 序列二级结构最小自由能
	predicted.edit.site.mfe.txt	所有预测 RNA 编辑位点前后 100bp 序列二级结构最小自由能
	predicted.snp.site.txt	dbsnp138 数据库 snp 位点前后 100bp 序列二级结构最小自由能
	output_file_true_site.txt	所有预测 RNA 编辑位点前后 100bp 序列二级结构结果原始文件
	mfe.pdf	数据库，预测 RNA 编辑位点，snp 位点二级结构最小自由能分布图
Vienne_picture	NORMAL	正常样本编辑位点
	TUMOR	肿瘤样本编辑位点
	darned_extracted.txt	DARNED 数据库编辑位点
	vienne.png	正常样本，肿瘤样本，DARNED 数据库编辑位点韦恩图
shell	frequency.sh	NA
	GeneEdit.sh	NA
	sequence_feature.sh	NA
	sequence_structure.sh	NA
	Vienne_picture.sh	NA
	run.sh	用于提交运算的脚本

[0104] 2.1 分析RNA编辑位点、SNP位点的二级结构最小自由能分布

[0105] 获得候选RNA编辑位点及数据库SNP位点后，将位点上下游各100bp的序列提取出来存入fasta格式的文档，将该文档直接以参数形式输入RNAfold软件，获取结果文件，从结果文件中提取每个位点的最小自由能数据，以R语言绘出最小自由能分布曲线。

[0106] 2.2 RNA编辑位点上下游序列特征分析

[0107] 获得候选RNA编辑位点后，将其上下游各10bp的序列提取出来并以每个位置为单位，统计不同碱基出现频率，以R语言绘出SequenceLogo图。

[0108] 2.3 绘制正常样本，肿瘤样本，DARNED数据库RNA编辑位点的韦恩图。

[0109] 3.结果

[0110] 3.1 RNA编辑位点、SNP位点的二级结构最小自由能分布的分析结果如图1所示，从

图中可以看出，本实施例中从65名前列腺癌患者核酸数据中预测的RNA编辑位点，其二级结构最小自由能分布曲线与DARNED数据库RNA编辑位点的二级结构最小自由能分布曲线相一致。而与SNP(dbSNP138，寡核苷酸多态性数据库)的二级结构最小自由能分布曲线有显著差异。说明本发明的方法能够有效的鉴别出RNA编辑位点数据和SNP位点数据。

[0111] 3.2 RNA编辑位点上下游序列特征分析的结果如图2所示，从图中可以看出编辑位点腺嘌呤(A，对应于表2中的碱基位置为11)出现频率最高，在编辑位点上游的-1位(对应于表2中的碱基位置为10)，鸟嘌呤(G)出现频率极低，可认为是该种碱基在-1位缺失，而在编辑位点下游+1位(对应于表2中的碱基位置为12)，鸟嘌呤(G)呈现较高的频率。这些特征与之前文献(Accurate identification of A-to-I RNA editing in human by transcriptome sequencing)报道一致。

[0112] 表2碱基频率分析

[0113]

碱基位置	1	2	3	4	5	6	7	8	9	10
A频率	0.22	0.21	0.24	0.18	0.23	0.26	0.21	0.23	0.16	0.27
C频率	0.27	0.28	0.24	0.26	0.23	0.25	0.24	0.28	0.32	0.33
G频率	0.23	0.28	0.26	0.26	0.30	0.29	0.25	0.26	0.22	0.06
T频率	0.28	0.23	0.26	0.31	0.24	0.21	0.31	0.24	0.30	0.34

[0114] 表2(续)

[0115]

碱基位置	11	12	13	14	15	16	17	18	19	20	21
A频率	0.95	0.19	0.23	0.22	0.21	0.23	0.26	0.21	0.24	0.25	0.23
C频率	0.00	0.20	0.28	0.28	0.26	0.27	0.23	0.31	0.25	0.27	0.28
G频率	0.01	0.47	0.26	0.23	0.25	0.25	0.29	0.23	0.28	0.24	0.27

[0116]

T频率	0.04	0.14	0.24	0.27	0.28	0.25	0.22	0.25	0.22	0.24	0.22

[0117] 3.3 正常样本，肿瘤样本，DARNED数据库RNA编辑位点的韦恩图如图3所示，从图中可以看出三类数据相互之间的重复率并不高，这表明，在不考虑假阳性的情况下，有许多的位点都可能是新发现的RNA编辑位点。

[0118] 实施例2

[0119] 重复实施例1中的步骤，不同点在于，用以下样本替换实施例1中65例前列腺癌患者，从而分别获得RNA编辑位点数据集，并进行特征分析：

[0120] 样本：24例肺癌患者。

[0121] 结果：

[0122] 实验结果如图4、5、6所示，本实施例中从肺癌患者样本中预测的RNA编辑位点，其二级结构最小自由能分布曲线与DARNED数据库RNA编辑位点的二级结构最小自由能分布曲线相一致，而与对照SNP(dbsnp138，寡核苷酸多态性数据库)的二级结构最小自由能分布曲线有显著差异。

[0123] 在本发明提及的所有文献都在本申请中引用作为参考，就如同每一篇文献被单独引用作为参考那样。此外应理解，在阅读了本发明的上述讲授内容之后，本领域技术人员可以对本发明作各种改动或修改，这些等价形式同样落于本申请所附权利要求书所限定的范围。

[0124] 参考文献:

[0125] 1.Ramaswami G,Lin W,Piskol R,et al.Accurate identification of human Alu and non-Alu RNA editing sites[J].Nature methods,2012,9(6):579-581.

[0126] 2.Peng Z,Cheng Y,Tan B C M,et al.Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome[J].Nature biotechnology,2012,30(3):253-260.

[0127] 3.Jae Hoon Bahn,Jae-Hyung Lee et al.Accurate identification of A-to-I RNA editingin human by transcriptome sequencing.Genome Research,2012,22:142-150

[0128] 4.Eddie Park,Brian Williams,Barbara J.Wold,et al.RNA editing in the human ENCODE RNA-seq data.Genome Research,201222:1626-1633

[0129] 5.Danecek et al.High levels of RNA-editing site conservation amongst 15 laboratory mouse strains.Genome Biology 2012,13:26.

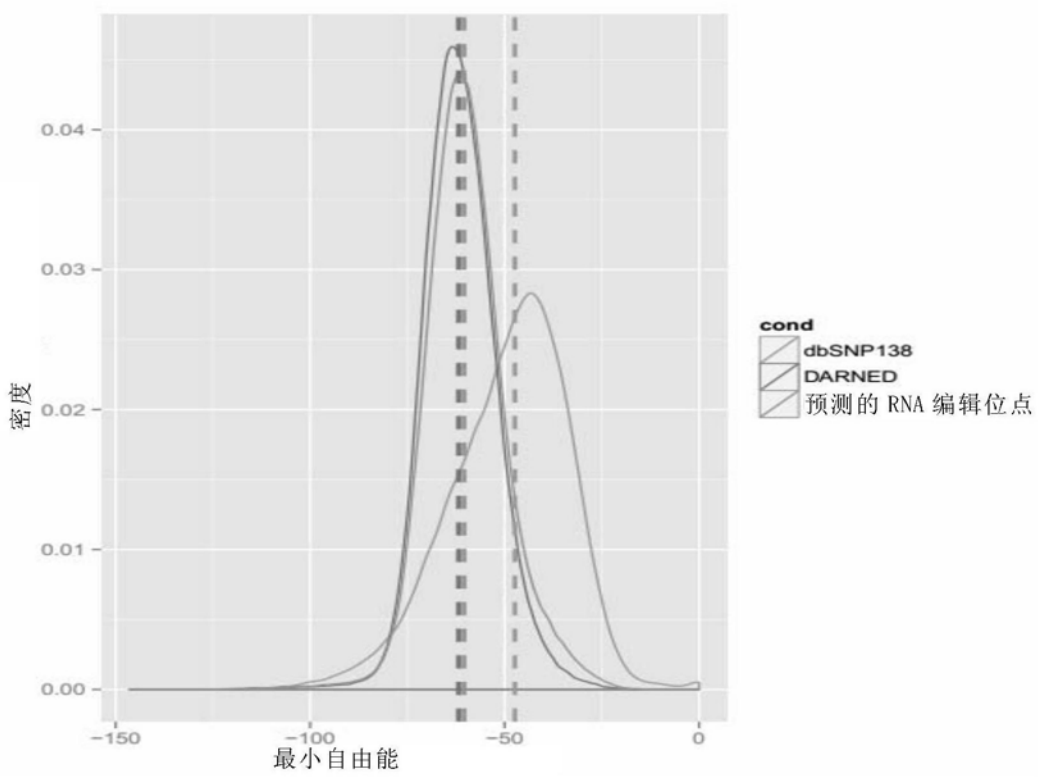

0.04
0.03
0.02
0.01
0.00
密度
−150
−100
−50
0
最小自由能
cond
dbSNP138
DARNED
预测的 RNA 编辑位点

图1

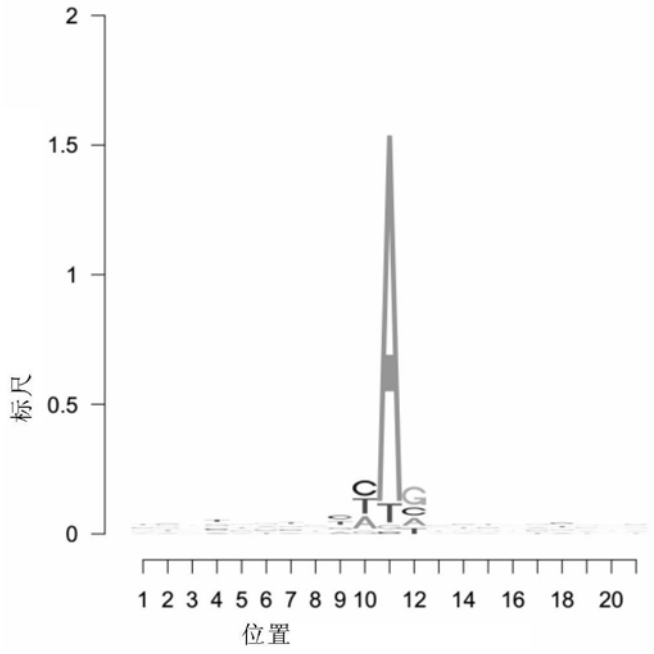

2
1.5
1
0.5
0
标尺
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20
位置

图2

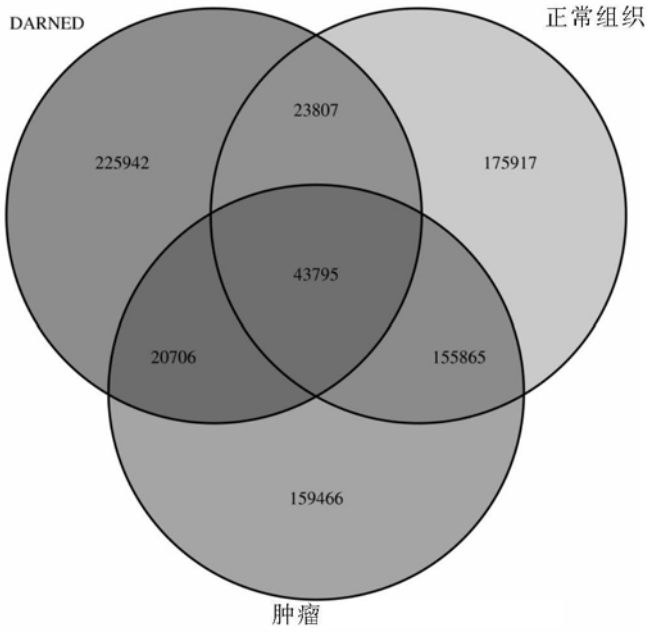

DARNED
正常组织
23807
225942
175917
43795
20706
155865
159466
肿瘤

图3

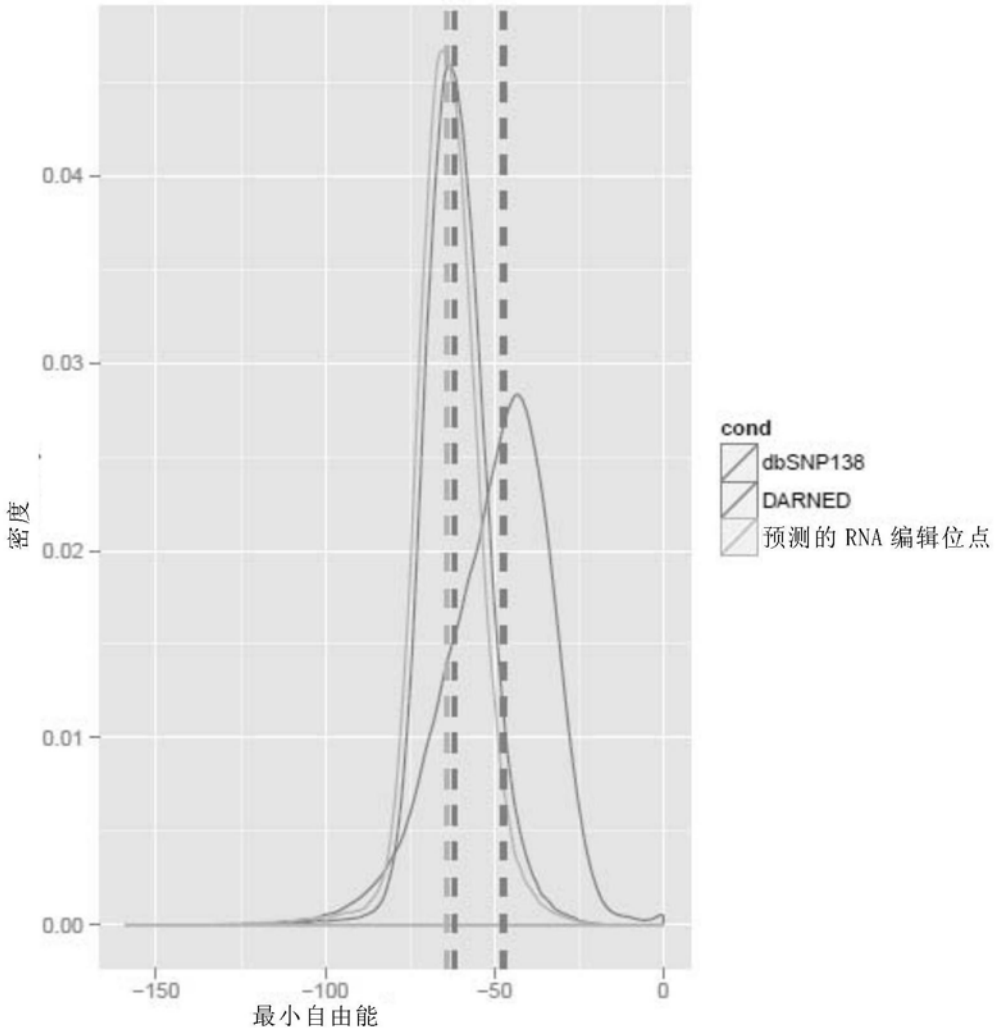

0.04
0.03
0.02
0.01
0.00
密度
−150
−100
−50
0
最小自由能
cond
dbSNP138
DARNED
预测的 RNA 编辑位点

图4

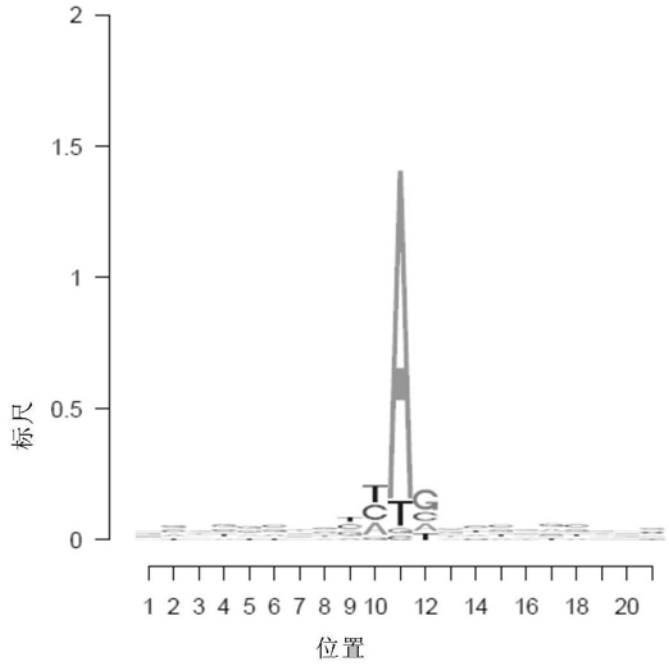

2
1.5
1
0.5
0
标尺
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20
位置

图5

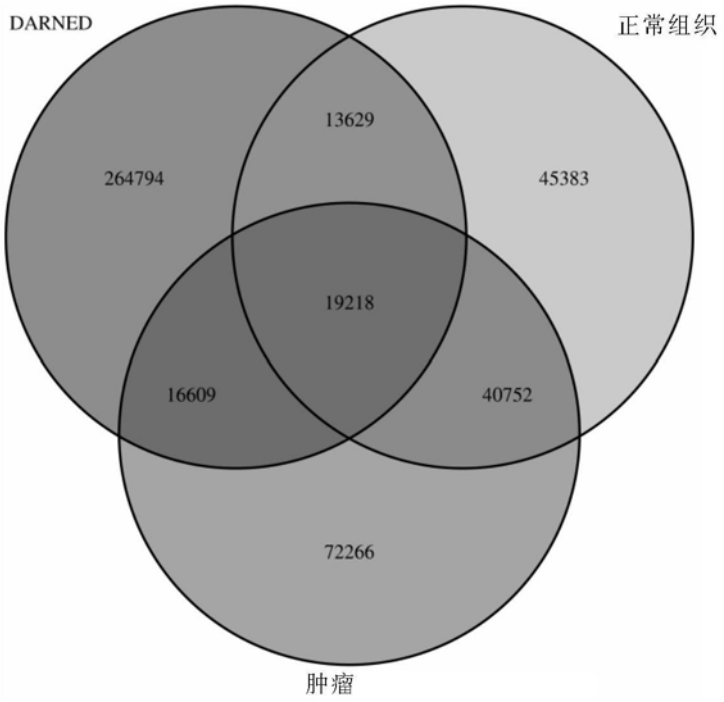

DARNED
正常组织
肿瘤
264794
13629
45383
19218
16609
40752
72266

图6