(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 119785886 A
(43)申请公布日 2025.04.08

(21)申请号 202411978853.2

(22)申请日 2024.12.31

(71)申请人 合肥综合性国家科学中心大健康研究院
地址 230601 安徽省合肥市经济技术开发区宿松路4090号1号楼202

(72)发明人 吴增丁 费才溢 徐实

(74)专利代理机构 上海一平知识产权代理有限公司 31266
专利代理师 徐迅 崔佳佳

(51)Int.Cl.
G16B 30/10(2019.01)
G16B 40/00(2019.01)

权利要求书2页 说明书14页
序列表（电子公布） 附图5页

(54)发明名称

基于全长和短片段转录组高通量测序技术的新生蛋白检测方法与系统

(57)摘要

本发明开发了一种样本新生蛋白检测分析方法,该方法提供了一键式检测分析流程,可以全面分析全长和短片段的RNA-seq数据,涵盖了所有可能诱导潜在新蛋白的变异,包括大规模变异和小规模变异,并基于上述多种变异识别新的转录本。最后,通过将所有类型的新转录本整合并转换为fasta格式的蛋白质序列,以便用于蛋白质后续分析,如蛋白质结构预测、新抗原免疫表位分析等。

CN 119785886 A

1.一种新蛋白质的检测方法,其特征在于,所述方法包括步骤:

(A)提供原始数据,构建文库,所述数据包括二代(NGS)转录组测序数据和三代转录组测序数据,所述三代转录组测序数据为HiFi数据;

(B)对原始数据进行预处理和质量控制,其中:

所述预处理包括步骤:对所述三代转录组测序数据进行抛光;将所述二代转录组测序数据比对到参考基因组上,获得第一个BAM文件;对所述三代转录组测序数据进行修剪,获得全长非嵌合序列(FLNC),将所述FLNC比对到所述参考基因组上,获得第二个BAM文件;两个所述BAM文件用于小规模变异分析;对所述FLNC进行聚类,获得第三个BAM文件,所述BAM文件用于大规模变异分析;

所述质量控制包括:计算零模波导孔(ZMWs)的数量;计算与所述参考基因组中唯一位置比对上的读段的数量;计算与所述参考基因组中多个位置比对上的读段的数量;计算无法与所述参考基因组比对上的读段的数量;使用所述BAM文件计算核糖体相关RNA碱基、内含子碱基、基因间区碱基和UTR区碱基的数量和比例;计算5'端和3'端的测序深度覆盖;计算比对到所述参考基因组正链和负链的读段比例;

(C)检测新变异亚型(NVI);

(D)对检测结果进行分层过滤和注释,获得过滤的NVI,其中,所述过滤的NVI包括核苷酸序列格式的NVI和转录本格式的NVI;

(E)利用过滤的NVI生成新蛋白。

2.如权利要求1所述的方法,其特征在于,在步骤(C)中,包括步骤:

(C1)使用cDNA_cupcake的collapse_isoforms_by_sam.py去除所述HiFi数据中的冗余并检测NVI;使用cDNA_cupcake的fusion_finder.py检测基因融合;使用PBsv软件检测NVI;去除从多个软件获得的结果,获得HiFi数据的检测结果;

(C2)使用Vardict软件检测所述NGS数据中的NVI,获得NGS数据的检测结果的VCF文件。

3.如权利要求1所述的方法,其特征在于,在步骤(D)中,所述分层过滤包括:

(D1)对异构体/融合进行过滤;

(D2)对SNV/INDEL进行过滤;

其中,在步骤(D1)中,包括步骤:

(d1.1)去除丰度低于阈值的NVI;

(d1.2)去除缺少5'外显子的NVI;

(d1.3)去除噪音NVI;

(d1.4)采用数据库对其余NVI进行注释,所述数据库为chimerDB 4.0数据库;

在步骤(D2)中,包括:

(d2.1)去除测序深度低于阈值的NVI,所述测序深度的阈值的计算方式如下:

$$(VD)VariantDepth = \frac{TPM \times total\_RPK}{1000000 * 100} \times mean\_read\_length$$

,

其中,RPK指每千碱基读段数;TPM指每百万转录本;mean_read_length指读段长度的平均值;

(d2.2)去除VCF文件FILTER列中带有噪音标志的NVI;

(d2.3)去除来自SNP或种系突变的NVI;

(d2.4)采用数据库对其余NVI进行注释,所述数据库包括COSMIC数据库。

4.如权利要求1所述的方法,其特征在于,在步骤(E)中,包括步骤:

(E1)将所述核苷酸序列格式的NVI转化为转录本格式的NVI;

(E2)合并两个转录本格式的NVI并去除冗余,将去除冗余的NVI储存在fasta文件中;

(E3)利用GeneMarkS-T软件将所述去除冗余的NVI转换为蛋白质序列;

(E4)利用BLAST软件将所述蛋白质序列与UniProt蛋白质数据库比较,从而判断所述蛋白质序列是否是新蛋白。

5.一种新生蛋白鉴定的软件或系统,其特征在于,所述软件或系统包括:

输入单元,所述输入单元被配置为输入序列数据,所述的输入序列数据为NGS转录组测序数据和三代转录组测序数据;

检测模块,所述检测模块被配置为对所述输入序列数据进行预定的检测,从而获得分析结果,并且所述检测模块包括:

(Z1)预处理与质量控制模块,所述预处理与质量控制模块被配置为执行权利要求1所述方法中步骤(B),从而获得预处理与质量控制的结果;

(Z2)检测模块,所述检测模块被配置为执行权利要求2所述的方法,从而获得检测结果;

(Z4)过滤与注释模块,所述预处理与质量控制模块被配置为执行权利要求3所述的方法,从而获得过滤与注释结果;

输出单元,所述输出单元被配置为执行权利要求4所述的方法,从而输出检测模块的检测结果。

6.由权利要求1-4任一项所述方法、或权利要求5所述软件或系统生成的蛋白质,其特征在于,所述蛋白质由如SEQ ID NO:1-4所示核苷酸序列编码而来。

7.一种多核苷酸,其特征在于,所述多核苷酸包含如SEQ ID NO:1-4所示核苷酸序列。

8.一种组合物,其特征在于,所述组合物包含权利要求6所述的蛋白质或权利要求7所述多核苷酸。

9.一种权利要求6所述蛋白质、权利要求7所述多核苷酸或权利要求8所述组合物的用途,其特征在于,用于制备治疗肿瘤的试剂或药物。

10.一种的权利要求1-4任一项所述方法、或权利要求5所述软件或系统用途,其特征在于,用于蛋白质分析,所述分析包括蛋白质结构预测、新抗原免疫表位分析。

基于全长和短片段转录组高通量测序技术的新生蛋白检测方法与系统

技术领域

[0001] 本发明涉及生物信息学与新生蛋白鉴定与检测领域,具体地,涉及一种基于全长和短片段转录组高通量测序技术的新生蛋白检测方法与系统。

背景技术

[0002] 人类疾病领域的许多基本问题,包括癌症,都与基因和转录水平上的碱基替换、缺失、插入、移码、内含子保留、选择性剪接以及新的未注释开放阅读框(nuORFs)翻译所产生的新蛋白密切相关。由于新蛋白不存在于正常细胞中,它们不会有免疫耐受,因此很容易被识别为外来抗原。新生蛋白可被消化成短肽,宿主系统将其视为T细胞抗原,导致激发细胞免疫反应,进而消灭表达新型蛋白质的细胞。基于这一原理,制药公司或研究机构已经开发了基于新抗原的免疫疗法,以作为一种有前景的癌症治疗方法。此外,新蛋白还与其他复杂疾病有关,例如阿尔茨海默病、II型糖尿病、肝病、肾病、肥胖和心血管疾病。因此,有效识别新蛋白对于理解人类疾病和开发相应的治疗方法至关重要。

[0003] 识别和表征蛋白质的传统方法是结合串联质谱(MS)和质谱后的数据库搜索。MS依赖于包含规范蛋白组中常规序列的参考数据库,所以几乎不可能识别规范蛋白组之外的蛋白质,特别是那些由患者特异性个体变异产生的蛋白质。MS的肽组学方法还有其他技术缺陷:富含Cys的肽的检测率显著降低,低分子量(<500Da)或高分子量(>3kDa)肽的检测灵敏度低。显然,由技术问题导致的偏差限制了MS在识别新蛋白中的应用。在过去的十年里,下一代测序(NGS)取得了重大进展,实现了样本特异性全基因组测序或总RNA测序,可以结合MS和NGS两种方法进行新蛋白分析。但仍存在以下不足之处:(1)由于下一代测序读长短,因而在检测大规模变异方面存在固有的技术缺陷;(2)规范蛋白质的鉴定协议通常是冗余的,需要在蛋白质水平上通过MS和在RNA水平上通过NGS进行检测,对输入样本要求较高;(3)现有基于核苷酸测序的生物信息分析工作流程,大多数只分析一般的常规转录本或主要关注单一或少数变异类别,因而无法涵盖所有样本特异性变异。

[0004] 综上所述,本领域迫切需要一种能覆盖多种变异类型、高效且准确的新蛋白质检测的方法和系统。

发明内容

[0005] 本发明的目的是提供一种基于全长和短片段转录组测序数据的新蛋白质鉴定方法和系统,具体地,涉及利用NGS和HiFi的RNA-seq数据,通过质量控制、变异检测、注释和分层过滤流程,最终生成新蛋白序列。

[0006] 本发明的第一方面,提供了一种新蛋白质的检测方法,所述方法包括步骤:

[0007] (A)提供原始数据,构建文库,所述数据包括二代(NGS)转录组测序数据和三代转录组测序数据,所述三代转录组测序数据为HiFi数据;

[0008] (B)对原始数据进行预处理和质量控制,其中:

[0009] 所述预处理包括步骤:对所述三代转录组测序数据进行抛光;将所述二代转录组测序数据比对到参考基因组上,获得第一个BAM文件;对所述三代转录组测序数据进行修剪,获得全长非嵌合序列(FLNC),将所述FLNC比对到所述参考基因组上,获得第二个BAM文件;两个所述BAM文件用于小规模变异分析;对所述FLNC进行聚类,获得第三个BAM文件,所述BAM文件用于大规模变异分析;

[0010] 所述质量控制包括:计算零模波导孔(ZMWs)的数量;计算与所述参考基因组中唯一位置比对上的读段的数量;计算与所述参考基因组中多个位置比对上的读段的数量;计算无法与所述参考基因组比对上的读段的数量;使用所述BAM文件计算核糖体相关RNA碱基、内含子碱基、基因间区碱基和UTR区碱基的数量和比例;计算5'端和3'端的测序深度覆盖;计算比对到所述参考基因组正链和负链的读段比例;

[0011] (C)检测新变异亚型(NVI);

[0012] (D)对检测结果进行分层过滤和注释,获得过滤的NVI,其中,所述过滤的NVI包括核苷酸序列格式的NVI和转录本格式的NVI;

[0013] (E)利用过滤的NVI生成新蛋白。

[0014] 在另一优选例中,所述二代转录组测序数据包括Illumina测序数据。

[0015] 在另一优选例中,所述原始数据包括选自下组的一个或多个:掺入2% SIRV-Set 4的人脑参考RNA样本、肿瘤异基因移植物RNA样本。

[0016] 在另一优选例中,所述肿瘤异基因移植物RNA样本的获得方式包括:将MC38细胞系皮下注射到C57BL/6小鼠中。

[0017] 在另一优选例中,通过环化一致性测序(CCS)对所述三代转录组测序数据进行抛光。

[0018] 在另一优选例中,通过选自下组的方法对所述FLNC进行聚类:层次对齐、迭代聚类、或其组合。

[0019] 在另一优选例中,通过选自下组的方法对所述三代转录组测序数据进行修剪:

[0020] (a)用Lima软件修剪所述三代转录组测序数据条形码;

[0021] (b)用Iso-Seq3软件修剪所述三代转录组测序数据的ployA尾和接头序列;

[0022] 或其组合。

[0023] 在另一优选例中,通过计算核糖体相关RNA碱基的数量和比例确定所述文库是否有核糖体相关RNA残余。

[0024] 在另一优选例中,所述核糖体相关RNA碱基的数量和比例为0%。

[0025] 在另一优选例中,通过计算5'和3'端的测序深度覆盖确定所述文库是否发生降解。

[0026] 在另一优选例中,所述5'端的测序深度覆盖≥0.7。

[0027] 在另一优选例中,所述3'端的测序深度覆盖≥0.7,较佳地≥0.9。

[0028] 在另一优选例中,通过计算比对到所述参考基因组正链和负链的读段比例确定所述文库性质。

[0029] 在另一优选例中,所述文库性质包括非链特异性文库和链特异性文库。

[0030] 在另一优选例中,所述非链特异性文库的所述比对到参考基因组正链和负链的读段比例保持平衡,所述"保持平衡"是指所述比对到参考基因组的正链的读段数目N1,与所

述比对到参考基因组负链的读段数目N2相比,满足N1/N2＝100±5%。

[0031] 在另一优选例中,所述链特异性文库的所述比对到参考基因组正链和负链的读段比例显著不平衡,所述“显著不平衡”是指所述比对到参考基因组的正链的读段数目N1,与所述比对到参考基因组负链的读段数目N2相比,满足N1/N2≥1900%,或N2/N1≥1900%。

[0032] 在另一优选例中,在步骤(C)中,包括步骤:

[0033] (C1)使用cDNA_cupcake的collapse_isoforms_by_sam.py去除所述HiFi数据中的冗余并检测NVI;使用cDNA_cupcake的fusion_finder.py检测基因融合;使用PBsv软件检测NVI;去除从多个软件获得的结果,获得HiFi数据的检测结果;

[0034] (C2)使用Vardict软件检测所述NGS数据中的NVI,获得NGS数据的检测结果的VCF文件。

[0035] 在另一优选例中,在步骤(C1)中,通过将相同的转录本异构体折叠成唯一的转录本异构体,从而去除冗余。

[0036] 在另一优选例中,在步骤(C1)中,对所述唯一的转录本异构体的全长读段数量进行量化,获得量化值用于步骤(D)。

[0037] 在另一优选例中,在步骤(D)中,所述分层过滤包括:

[0038] (D1)对异构体/融合进行过滤;

[0039] (D2)对SNV/INDEL进行过滤;

[0040] 其中,在步骤(D1)中,包括步骤:

[0041] (d1.1)去除丰度低于阈值的NVI;

[0042] (d1.2)去除缺少5'外显子的NVI;

[0043] (d1.3)去除噪音NVI;

[0044] (d1.4)采用数据库对其余NVI进行注释,所述数据库为chimerDB 4.0数据库;

[0045] 在步骤(D2)中,包括:

[0046] (d2.1)去除测序深度低于阈值的NVI,所述测序深度的阈值的计算方式如下:

[0047] $$(VD)VariantDepth = \frac{TPM \times total\_RPK}{1000000 * 100} \times mean\_read\_length$$,

[0048] 其中,RPK指每千碱基读段数;TPM指每百万转录本;mean_read_length指读段长度的平均值;

[0049] (d2.2)去除VCF文件FILTER列中带有噪音标志的NVI;

[0050] (d2.3)去除来自SNP或种系突变的NVI;

[0051] (d2.4)采用数据库对其余NVI进行注释,所述数据库包括COSMIC数据库。

[0052] 在另一优选例中,所述丰度的阈值为2。

[0053] 在另一优选例中,所述TPM＝35。

[0054] 在另一优选例中,所述噪音标志包括P8、Q10、bias。

[0055] 在另一优选例中,根据dbSNP注释信息和等位基因频率(AF)去除来自SNP或种系突变的NVI。

[0056] 在另一优选例中,在步骤(E)中,包括步骤:

[0057] (E1)将所述核苷酸序列格式的NVI转化为转录本格式的NVI;

[0058] (E2)合并两个转录本格式的NVI并去除冗余,将去除冗余的NVI储存在fasta文件

中;

[0059] (E3)利用GeneMarkS-T软件将所述去除冗余的NVI转换为蛋白质序列;

[0060] (E4)利用BLAST软件将所述蛋白质序列与UniProt蛋白质数据库比较,从而判断所述蛋白质序列是否是新蛋白。

[0061] 本发明的第二方面,提供了一种新生蛋白鉴定的软件或系统,所述软件或系统包括:

[0062] 输入单元,所述输入单元被配置为输入序列数据,所述的输入序列数据为NGS转录组测序数据和三代转录组测序数据;

[0063] 检测模块,所述检测模块被配置为对所述输入序列数据进行预定的检测,从而获得分析结果,并且所述检测模块包括:

[0064] (Z1)预处理与质量控制模块,所述预处理与质量控制模块被配置为执行本发明第一方面所述方法中步骤(B),从而获得预处理与质量控制的结果;

[0065] (Z2)检测模块,所述检测模块被配置为执行本发明第一方面所述方法中步骤(C),从而获得检测结果;

[0066] (Z4)过滤与注释模块,所述预处理与质量控制模块被配置为执行本发明第一方面所述方法中步骤(D),从而获得过滤与注释结果;

[0067] 输出单元,所述输出单元被配置为执行本发明第一方面所述方法中步骤(E),从而输出检测模块的检测结果。

[0068] 本发明的第三方面,提供了由本发明第一方面所述方法、或本发明第二方面所述软件或系统生成的蛋白质,所述蛋白质由如SEQ ID NO:1-4所示核苷酸序列编码而来。

[0069] 本发明的第四方面,提供了一种多核苷酸,所述多核苷酸包含如SEQ ID NO:1-4所示核苷酸序列。

[0070] 本发明的第五方面,提供了一种组合物,所述组合物包含本发明第四方面所述的蛋白质。

[0071] 在另一优选例中,所述组合物还包含药学上可接受的载体、稀释剂或赋形剂。

[0072] 本发明的第六方面,提供了一种本发明第三方面所述的蛋白质、本发明第四方面所述的多核苷酸或本发明第五方面所述组合物的用途,用于制备治疗肿瘤的试剂或药物。

[0073] 本发明的第七方面,提供了一种本发明第一方面所述的方法或本发明第二方面所述的软件或系统用途,用于蛋白质分析,所述分析包括蛋白质结构预测、新抗原免疫表位分析。

[0074] 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

[0075] 图1显示了本发明中样本新生蛋白检测分析流程。通过综合分析NGS和HiFi测序数据,一体化检测所有新变异和异构体,包括单核苷酸变异(SNV)、插入-缺失突变(INDELs)、基因融合(FUSION)、内含子保留和选择性剪接的ISOFORM。分析的中间步骤包括多重严格过滤,最终获得低背景噪声的结果。最后,将转换的氨基酸序列与蛋白质数据库进行比较,以

选择新的或经典的蛋白质。

[0076] 图2显示了新变异异构体(NVI)检测的类别。对于小规模变异,包括(A)点突变和(C)基因融合,使用短读长测序读段进行分析。对于大规模变异,包括(B)RNA选择性剪接、(C)基因融合和(D)NuORFs,使用长读长测序数据进行分析。RNA融合使用长读长和短读长数据进行分析。

[0077] 图3中(A)说明了长读长序列的折叠过程。因为序列a和b,序列c和d的内含子完全相同,所以它们分别折叠得到唯一的潜在转录本候选,而序列a和c由于内含子的差异无法进一步折叠。(B)表明许多序列在5'端无法对齐,但在3'端可以对齐。这样的序列在生物学上可能是同一原始转录本的5'降解,因此它们被折叠成一个唯一的转录本异构体。

[0078] 图4显示了所有新变异异构体(NVI)初步检测结果的分层过滤。倒三角形代表过滤的逻辑。每一层代表一个步骤的过滤操作。

[0079] 图5显示了MC38分析结果。(A)分析结果显示了Deliverables部分下研究的变体的核苷酸和相应的蛋白质序列。(B)使用PacBio BAM文件在IGV中展示的PB1596.5变体结果。(C)整合分析得到的最终新蛋白序列。(D)使用BLAST比较PB1596.5变体的蛋白序列与Q1KYM0。(E)PB1596.5的电泳结果,从左到右依次为Marker组、MC38细胞、正常组织对照组和MC38异基因移植物。

具体实施方式

[0080] 本发明人经过广泛而深入的研究,创造性地开发了一种样本新生蛋白检测分析方法,该方法提供了一键式检测分析流程,可以全面分析全长和短片段的RNA-seq数据,涵盖了所有可能诱导潜在新蛋白的变异,如包括大型基因融合、转录本外显子跳跃、内含子保留、RNA可变剪接和非编码区异常表达等在内的大规模变异,以及包括插入、缺失、重复、替换和小规模基因融合在内的小规模变异,并基于上述多种变异识别新的转录本。最后,通过将所有类型的新转录本整合并转换为fasta格式的蛋白质序列,以便用于蛋白质后续分析,如蛋白质结构预测、新抗原免疫表位分析等。

[0081] 应当理解,以下以各种详细程度描述本发明的具体方法和实验条件、是用于提供对本发明的实质理解。下面提供了本说明书中使用的某些术语的定义。除非另外定义,否则本文中所用的全部技术与科学术语均具有如本发明所属领域的普通技术人员通常理解的相同含义。

[0082] 术语

[0083] 在提供数值范围的情况下,除非上下文另外清楚地指出,否则应当理解,该值20的每个中间整数、该值的每个中间整数的每十分之一、在该范围的上限与下限之间和在该规定范围中的任何其他中间值都包括在本发明内。这些较小范围的上限和下限可以独立地包括在较小范围内,并且也涵盖在本发明内,但须遵守规定范围内的任何明确排除的限制。例如,“1至50”包括“2至25”、“5至20”、“25至50”、“1至10”等。

[0084] 如本文所用,术语“含有”或“包括(包含)”可以是开放式、半封闭式和封闭式的。换言之,所述术语也包括“基本上由……构成”、或“由……构成”。

[0085] 如本文所用,术语“新生蛋白”、“新蛋白”与“新型蛋白质”可以互换使用,是指非天然存在的蛋白质。在本发明中,新生蛋白可以由本发明提供的检测方法从长度长数据和短

读长数据中获得。

[0086] 如本文所用,术语“比对”与“对齐”可以互换使用,是指将两个或多个序列排列在一起,比较相似之处。本发明将转录组数据与参考基因组进行比较,获得BAM文件。

[0087] 如本文所用,术语“条形码”是指混合样本中单个样本的唯一标识,用于区分不同样本。由于测试仪器的测序能力远大于测试样本量,因此为避免浪费,常用同一通道(lane)测定多个样本。为不同样本添加特定标识,从而在后续数据分析时将不同样品数据进行区分。

[0088] 如本文所用,术语“基因(Gene)”是指生物学中位于DNA或RNA中,能够编码合成包括RNA和蛋白质在内的基因产物的核苷酸序列。基因通常被视为基本遗传单位,即一段有功能性的DNA或RNA。

[0089] 如本文所用,术语“转录本(Transcipt)”是指由一个基因通过转录形成的一种或多种成熟的RNA分子。基因转录形成的RNA可通过不同加工过程形成不同的转录本。所述不同的转录本即为“转录本异构体”。

[0090] 如本文所用,术语“新变异异构体”与“NVI”可以互换使用,是指利用本发明的检测方法或检测系统,基于遗传变异的数据获得的结果。

[0091] 如本文所用,术语“单核苷酸变异(SNV)”是指单一核苷酸发生的变异,其变异频率没有限制。

[0092] 如本文所用,术语“插入-缺失突变(INDELs)”是指在基因组序列中发生的小片段的插入或缺失,其长度在1-50bp之间。

[0093] 如本文所用,术语“短片段”与“短读长”可以互换使用,是指运用二代测序技术获得的序列。二代测序技术的测序成本低,测序速度高,且具有较高的准确性。二代测序的读段长度通常不超过500bp。由于读段长度较短,短读长数据通常包含小规模变异。在本发明中,利用短片段数据进行小规模变异分析。

[0094] 如本文所用,术语“全长”与“长读长”可以互换使用,是指运用三代测序技术获得的序列。三代测序的读段长度通常可以达到几千个甚至碱基对,因此可以包含大规模变异的信息。在本发明中,三代测序的数据主要用于大规模变异分析,也可用于小规模变异分析。

[0095] 样本新生蛋白检测分析方法

[0096] 样本新生蛋白检测分析方法(后简称:检测方法)采用Snakemake作为其工作流管理系统,实现在高性能计算环境中的受控和可扩展的流程执行。Snakemake是一个高效且通用的工作流管理工具,为科学数据分析提供了诸多优势,包括卓越的内存管理、强大的可移植性、模块化设计和可重复的结果。本发明的检测方法的整体流程见图1。

[0097] 作为一个综合性的生物信息学流程,该检测方法与各种数据类型兼容,包括NGS和HiFi RNA测序数据。它提供了一系列分析步骤,包括质量控制、变异检测、注释和分层过滤,以及新蛋白序列生成。该检测方法提供了广泛的变异检测类别,如插入、缺失、移码、基因融合、内含子保留和选择性剪接,以实现对转录组数据的全面和详细分析。这些变异类别有助于识别新的转录本或蛋白质,在此将其称为新变异异构体(NVI)。

[0098] 要使用该检测方法流程,需要一个由HiFi测序生成的subreads.bam文件的原始数据,以及一个包含可选参数和必要的参考和注释文件路径的配置文件。如果有的话,也可以

包括由NGS平台生成的原始数据。用户可以使用默认参数,或在初始测试运行之前根据特定需求调整参数。在本地部署后,用户只需为后续分析指定样本特定的原始数据路径。该检测方法流程包含二十多个步骤,可以分为四个不同的部分(如图1):

[0099] 1)原始数据的质量控制和预处理,

[0100] 2)新变异异构体(NVI)的检测,

[0101] 3)检测结果的分层过滤和注释,

[0102] 4)以fasta格式从过滤后的NVI生成蛋白质序列,并识别不存在于规范蛋白组中的新蛋白。

[0103] 环化一致性测序

[0104] 如本文所用,术语“环化一致性测序”、“环化测序”、“CCS”可以互换使用,通过环化一致性测序可以获得高精度的单分子一致性HiFi读段。CCS过程包括切割交替出现在在接头和插入序列之间的“聚合酶读段”,以去除接头,从而获得了多个被称为“子读段”的插入序列。这些子读段序列是通过插入序列的环化测序生成的。来自同一源插入序列的子读段在很大程度上是相同的,但测序错误会在这些来自相同插入序列的子读段中引入了随机错误。为了解决这个问题,将这些子读段相互对齐并进行抛光处理。一般来说,只要进行三次或以上的环化测序,即获得三个或更多的子读段序列,CCS抛光就可以产生准确度高达99.9%的HiFi读段。CCS过程是本发明中最耗时的步骤。

[0105] 遗传变异

[0106] 本发明中的遗传变异是指所有可能生成新蛋白的基因上的变异,包括小规模变异和大规模变异。所述小规模变异包括短插入、缺失、重复、替换和小规模基因融合;所述大规模变异包括大型基因融合、转录本外显子跳跃、内含子保留、RNA可变剪接和非编码区异常表达,以及长插入、缺失和重复。

[0107] 本发明的主要优点包括:

[0108] (1)本发明的检测方法通过整合长读长和短读长测序技术,结合创新的生物信息学分析流程,能够全面检测各种尺度的遗传变异,包括点突变、插入缺失、基因融合、内含子保留等,从而识别更多潜在的新蛋白,解决了检测大规模变异和小规模变异,以及分析复杂样本时的局限性。

[0109] (2)本发明的检测方法采用严格的质量控制和过滤步骤,降低假阳性率,提高检测结果的可靠性与准确性。

[0110] (3)本发明的检测方法通过优化算法和并行计算,加速分析过程,提高检测效率。

[0111] 下面结合具体实施例,进一步阐述本发明。应理解,这些实施例仅用于说明本发明而不用于限制本发明的范围。下列实施例中未注明具体条件的实验方法,通常按照常规条件,例如Sambrook等人,分子克隆:实验室手册(New York:Cold Spring Harbor Laboratory Press,1989)中所述的条件,或按照制造厂商所建议的条件。除非另外说明,否则百分比和份数是重量百分比和重量份数。

[0112] 实施例1:原始数据的质量控制和预处理

[0113] 本实施例涉及对原始HiFi测序数据进行质量控制和预处理。

[0114] 本实施例使用PacBio提供的开源软件对数据进行预处理。具体地,使用环化一致性测序(Circular consensus sequencing,CCS)软件获得高精度的单分子一致性HiFi读

段。这个过程包括切割交替出现在在接头和插入序列之间的“聚合酶读段”,以去除接头,从而获得了多个被称为“子读段”的插入序列。这些子读段序列是通过插入序列的环化测序生成的。来自同一源插入序列的子读段在很大程度上是相同的,但测序错误会在这些来自相同插入序列的子读段中引入了随机错误。为了解决这个问题,将这些子读段相互对齐并进行抛光处理。一般来说,只要进行三次或以上的环化测序,即获得三个或更多的子读段序列,CCS抛光就可以产生准确度高达99.9%的HiFi读段。其中,CCS步骤是整个流程中最耗时的步骤。

[0115] 生成HiFi读段后,使用Lima软件修剪条形码,以便区分在同一文库池中构建的多个样本。如果文库只包含一个样本,则可以省略此步骤。随后,使用Iso-Seq3软件的refine模块修剪HiFi读段两头的polyA尾和接头序列,从而获得了全长非嵌合(Full-length non-chimeric,FLNC)序列,同时也去除了残余的串联重复。将获得的FLNC序列与参考基因组进行比对,生成BAM文件。同时,将NGS数据比对到同一参考基因组上,生成另一个BAM文件。这两个BAM文件将进一步用于小规模变异分析。此外,使用Iso-Seq3软件对FLNC读段进行聚类,采用层次n*log(n)对齐和迭代聚类合并,从而识别其他大规模变异。

[0116] 数据预处理涉及多个处理步骤和分析模块,这些步骤和模块可能会带来异常,并影响后续分析。因此,在每个步骤实施多维质量控制以确保下游分析和关于新蛋白的最终结果。例如,单分子测序所用的孔数,称为零模波导孔(Zero-Mode Waveguide,ZMWs),代表数据量。此外,经过抛光的HiFi读段以及去除polyA尾后剩余的FLNC读段必须达到所需的准确度。另外,还利用“uniq_mapped_reads(与参考基因组中唯一位置比对上的读段)”、“NonUniq_mapped_reads(与多个位置比对上的读段)”和“Unmapped_reads(无法比对的读段)”等指标来评估测序数据质量。这些质量控制指标可用于每个步骤的分析,并提供分析问题的反馈。比如,若ZMWs的数量太低,表明测序的输入文库不够,提示研究人员增加样本输入以提高数据量,优选地,ZMWs≥1M(百万);若Unmapped_reads比例过高,可能表明有其它物种的污染。

[0117] 使用BAM文件计算RIBOSOMAL_BASES、INTRONIC_BASES、INTERGENIC_BASES和UTR_BASES的数量和比例以评估文库质量。INTERGENIC_BASES比例高可能表明DNA污染,而RIBOSOMAL_BASES比例高则表明核糖体RNA去除存在问题,会导致检测敏感性降低。使用RSeQC软件中的geneBody_coverage.py模块分析基因5’和3’端的测序深度覆盖,获得MEDIAN_5PRIME_BIAS和MEDIAN_3PRIME_BIAS。MEDIAN_5PRIME_BIAS值低表明转录本5’端的覆盖深度显著降低,可能表明严重的文库降解。结合评估RNA质量的RIN值,上述分析有助于全面确定是否需要重新测序。

[0118] 此外,使用RSeQC软件中的infer_experiment.py模块确定比对到参考基因组上的“+”和“-”链的读段比例。比例平衡表明非链特异性文库构建,而比例显著不平衡表明链特异性文库构建。链特异性文库可用于分析重叠的两个基因的基因表达和异构体检测。

[0119] 实施例2:新变异异构体(NVI)的检测

[0120] 本实施例涉及利用HiFi数据以及NGS数据进行大规模变异和小规模变异分析以检测新变异异构体。

[0121] 对原始数据进行实施例1中预处理后,进行新变异异构体(NVI)的检测。将所有可能生成新蛋白的遗传变异分为两类:小规模变异(如短插入、缺失、重复、替换和小规模基因

融合)和大规模变异(如大型基因融合、转录本外显子跳跃、内含子保留、RNA可变剪接和非编码区异常表达,以及长插入、缺失和重复),类型见图2。

[0122] 为了能够同时分析不同规模的变异,本实施例将各种变异分析软件整合到该检测方法中,以满足各类变异的分析要求并提高整体分析性能。本实施例利用数据的特点优化了分析方法。由于具有长读长序列的优势,HiFi读段主要用于分析大规模变异;由于在单碱基水平上具有更高准确性且更容易获得高深度测序数据的特点,NGS读段主要用于分析小规模变异。

[0123] 对于HiFi读段,使用cDNA_cupcake的collapse_isoforms_by_sam.py模块将相同的转录本异构体折叠成唯一的转录本异构体。对不同唯一异构体对应的全长读段数量进行量化,以在后续过滤步骤中用作参数。使用cDNA_cupcake的fusion_finder.py检测基因融合,因其可作为新蛋白或嵌合蛋白的重要来源。HiFi读段也可用于分析小规模变异,例如,使用PBsv分析各种结构变异,包括约10bp的缺失,这些变异也可能出现在NGS变异分析结果中。去除从多个软件获得的变异结果中的冗余。

[0124] 对于NGS读段,使用Vardict软件分析小规模变异,并生成存储所有可能变异的VCF文件。过滤不产生新蛋白的变异,如非编码区的单碱基替换或同义、无义或沉默突变。

[0125] 将基于NGS数据分析的大规模变异与HiFi数据分析的结果进行比较并整合,以进一步提高结果的可靠性。

[0126] 实施例3:分层过滤

[0127] 3.1异构体/融合的过滤

[0128] 由于融合也被视为一种特定类型的异构体,其过滤过程和步骤与异构体基本相同。过滤检测的异构体通常涉及4个步骤(如图4):

[0129] (S1)根据丰度去除低于预设阈值(默认为2,高质量异构体应该至少由2个FLNC读段支持)的异构体。通常,每个检测异构体的丰度可以直接从PacBio HiFi测序数据中量化。然而,如果可以从同一样本的另一份等分样品中获得额外的NGS测序数据,就能更准确地量化异构体的丰度。

[0130] (S2)去除一部分异构体(见图3),这些异构体与最长的异构体在3'端具有相同的外显子,但缺少一些5'外显子,说明这些5'截短的转录本可能是由在RNA在文库构建过程中5'端降解造成的。文库降解越严重,由于5'截短而被过滤掉的转录本数量就越多。

[0131] (S3)去除由PCR嵌合、限制性酶连接或测序错误等引入的“人工”异构体。这些人工异构体是从测序数据中检测的真实信号,但对样本来说是噪音,因而必须被清除。由于单一特征很难识别导致人工异构体的随机干扰,本实施例采用了基于随机森林的机器学习方法构建的SQANTI3工具来解决这一问题。SQANTI3提取多达47个特征,包括覆盖度、外显子坐标、支持读段数等,并基于这些特征有效实施过滤。

[0132] (S4)与数据库进行注释。这一步可以识别与临床疾病相关的有意义的异构体/融合,以及潜在的治疗靶点。基因注释结果的详细程度和准确性高度依赖于注释数据库的质量和数量。对于人类融合基因注释,采用chimerDB 4.0数据库,该数据库总结了癌症基因组图谱(TCGA)项目中患者的近10万个融合候选基因,包括数万个具有实验证据支持的融合基因。

[0133] 3.2SNV/INDEL的过滤

[0134] SNV/INDEL的过滤涉及4个步骤(见图4):

[0135] (X1)去除测序深度低于置信阈值的变异。使用以下公式计算深度阈值:

[0136] $$(VD)VariantDepth = \frac{TPM \times total\_RPK}{1000000 * 100} \times mean\_read\_length$$

[0137] 其中,RPK指每千碱基读段数,通过将读段计数除以每个基因的千碱基长度计算得出;TPM指每百万转录本,通过将PRK值除以“每百万”缩放因子计算得出;mean_read_length指读段长度的平均值。

[0138] 由于TPM高于35的变异转录本翻译的肽段会被认为具有免疫原性,因此将TPM设置为35,可以使用公式计算变异深度的阈值。

[0139] (X2)去除背景噪音,主要使用VCF文件中的FILTER列来去除带有P8、Q10、bias等标志的变异。

[0140] (X3)通过综合分析dbSNP注释信息和等位基因频率(AF)去除来自SNP或种系突变的变异。

[0141] (X4)利用COSMIC等数据库注释和识别与临床疾病相关的突变。

[0142] 实施例4:新蛋白生成

[0143] 本实施例涉及综合不同分析软件的结果,转换文件格式并去除冗余以获得最终新蛋白生成结果。

[0144] 来自不同分析软件的所有种类变异结果经过分层过滤以获得各自的最终结果。为了便于后续分析,首先将变异的核苷酸序列转换为转录本格式。转录本异构体和融合分析结果已经是核苷酸序列格式,不需要进一步转换。然而,通过Vardict获得的NGS数据的VCF格式小规模变异和通过PBsv获得的HiFi读段的结构变异需要转换为转录本格式。整合结果变异文件并去除冗余变异后,将整合的变异转录本存储在fasta文件中。

[0145] 在转录本分析之后,使用GeneMarkS-T软件检查表达的蛋白质序列,将转录本序列转换为向相应蛋白质序列。由于存在多个开放阅读框,较长的转录本会产生多个蛋白质序列。通过这种分析获得的变异通常构成新蛋白,包括肿瘤特异性蛋白。使用BLAST软件将上述蛋白与从UniProt下载的蛋白质数据库进行比较,以确定变异产生的蛋白是否已经被研究和发表,便于在特定的研究背景下人工判断选择最终结果。已发表的蛋白序列并不会被随意消除,而是在fasta文件中注释可能的同源蛋白ID和一致性值。此外,某些已发表的蛋白可能在肿瘤组织中特异性表达,因此也被认为与研究目的相关。

[0146] 实施例5:模型性能的验证

[0147] 本实施例涉及验证本发明的检测方法的工作原理,包括使用两种数据或样本对其进行验证。

[0148] 5.1数据集

[0149] 第一个样本是掺入2% SIRV-Set 4的人脑参考RNA样本。SIRV-Set 4是一种包含人工异构体集的生物学参考标准。从NCBI获取了该RNA样本的PacBio测序数据,格式为FLNC.read.bam,表明它已经过预处理。SIRV-Set 4包含69个人工转录本变体,模拟了7个人类模型基因座的剪接特征,全面反映了选择性剪接、选择性转录起始和终止位点、重叠基因和反义转录本的多样性。此外,SIRV-Set 4还包括五个长度类别(分别为4kb、6kb、8kb、10kb和12kb)的15个人工转录本变体,以模拟人类转录本的长度复杂性。

[0150] 第二个样本是通过将MC38细胞系皮下注射到C57BL/6小鼠中形成肿瘤异基因移植物而获得的。在处死小鼠后,收集肿瘤异基因移植物和配对的正常小肠组织,并将其分割成约300mg的小块。这些组织块立即用1xPBS洗涤,放入单独的离心管中,在低温下运送到GeneDenovo有限公司进行RNA提取、PacBio HiFi长读段测序和Illumina短读段测序。获得测序数据后,使用本发明的检测方法进行分析。

[0151] 由于肿瘤异基因移植物和正常组织在转录组和蛋白质组方面存在显著差异,因此验证的目的是使用该检测方法识别肿瘤异基因移植物特有的突变和异常转录本表达,并获得相应的新型蛋白质。

[0152] 5.2验证结果

[0153] 本发明的检测方法利用PacBio HiFi长读长数据对人脑参考RNA样本进行分析,以及利用两种读长数据对MC38异基因移植样本进行分析。

[0154] 由于人脑参考RNA样本已经进行了预处理,包括CCS抛光步骤,因而本次验证流程从使用MINIMAP2对上述预抛光的HiFi读段进行比对开始。由于该验证数据集缺乏短读长数据,因此本次验证仅专注于分析大规模变异,特别是掺入SIRV-Set 4中的不同异构体。而MC38异基因移植样本同时包含PacBio HiFi长读和Illumina短读数据,因此需要分析大规模和小规模突变类型。

[0155] 由于两个数据集都是原始数据,MC38异基因移植的分析从数据预处理和质量控制步骤开始,并执行了新生蛋白检测的所有分析模块。新生蛋白检测在配备Intel(R)Xeon(R)Platinum处理器(8269CY CPU,3.10GHz)和Ubuntu 20.04.4LTS Linux操作系统的高性能计算机集群上运行。大约1小时后,本发明检测方法完成了对人脑参考RNA样本数据的分析;而MC38异基因移植数据的分析花费了近5小时。通过统计分析每个模块的运行时间,发现CCS步骤是最耗时的。不需要进行CCS步骤的人脑参考RNA样本节省了大量时间。

[0156] 对于人脑参考RNA样本,首先评估了质量控制指标,获得的HiFi读段总数为190万,几乎占据了SMRT Cell芯片数据容量的一半。PCT_RIBOSOMAL_BASES指标为0%,表明tRNA的存在可以忽略不计。MEDIAN_5PRIME_BIAS和MEDIAN_3PRIME_BIAS值分别为0.73和0.99,表明RNA文库相对完整,没有发生明显降解。“++--”和“+-,-+”值分别为0.859和0.0042,表明文库构建是链特异性的。上述质量控制指标共同表明数据的数量和质量符合所需标准。

[0157] 由于掺入人脑参考RNA样本的SIRV-Set 4是一种具有异构体精确信息的生物学参考标准,本实施例确认了所有真实异构体的检测结果,计算检测率以确定本发明的新生蛋白检测方法是否会错误地遗漏真阳性。对于SIRV-set 4的12个长转录本,本发明的检测方法达到了100%(12/12)的检测率(见表1和表2)。然而,对于69个短SIRV转录本的检测,本发明的检测方法仅达到了89.9%(62/69)的检测率,遗漏了7个异构体。之后,对未被检测出的异构体进行探究,发现未被检测出的异构体是由于它们的内含子或外显子长度较短,导致在全局比对中得到高映射分数,随后被错误分类为软裁剪(Soft clip)或缺失。人类RefSeq转录组的统计数据表明,绝大多数外显子和内含子的长度都超过20bp。这三个错误分类的异构体的外显子和内含子分别只有9bp、31bp和20bp,代表了极端情况。另外在原始数据中缺失的四个异构体表明它们在文库构建或核苷酸测序过程中丢失。

[0158] 表1检测方法正确检测出短SIRV的异构体的情况

[0159]

转录本 ID	转录本总数	正确检测数量	正确检测率	没有检测到的转录本
SIRV1	8	7	87.50%	SIRV105
SIRV2	6	6	100.00%	/
SIRV3	11	10	90.90%	SIRV311
SIRV4	7	6	85.70%	SIRV404
SIRV5	12	10	83.30%	SIRV503，SIRV512
SIRV6	18	17	94.40%	SIRV618
SIRV7	7	6	85.70%	SIRV708

[0160]

合计	69	62	89.90%	/

[0161] 表2检测方法正确检测出长SIRV的异构体的情况

转录本 ID	转录本总数	正确检测数量	正确检测率	没有检测到的转录本
SIRV10001	1	1	100.00%	/
SIRV10002	1	1	100.00%	/
SIRV10003	1	1	100.00%	/
SIRV12001	1	1	100.00%	/
SIRV12002	1	1	100.00%	/
SIRV12003	1	1	100.00%	/
SIRV4001	1	1	100.00%	/
SIRV4002	1	1	100.00%	/
SIRV4003	1	1	100.00%	/
SIRV6001	1	1	100.00%	/
SIRV6002	1	1	100.00%	/
SIRV6003	1	1	100.00%	/
SIRV8001	1	1	100.00%	/
SIRV8002	1	1	100.00%	/
SIRV8003	1	1	100.00%	/
合计	15	15	100.00%	/

[0162]

[0163] 对于MC38肿瘤异基因移植物,本实施例首先对其进行了质量控制评估。结果表明,数据的数量和质量足以进行后续分析。使用Vardict和PBsv在短读数据上检测到了大量高置信度的小规模变异;而大规模变异,如内含子保留、外显子跳跃和非编码区域的异常表达,主要通过使用cDNA_cupcake分析长读长HiFi数据来检测,包括通过fusion_finder识别的基因融合。这些大规模变异在SQANTI和分层过滤后与小规模变异整合,并存储在单独的fasta文件中(如图5A)。关于大规模遗传变异,未发现任何明确的基因融合或结构改变的证据,这与先前对MC38细胞系的观察一致。然而,本实施例确认了一个特定转录本异构体存在五个转录本变异,这些变异得到了强有力的支持证据。其中四个变异,分别命名为PB.349.2、PB.458.1、PB.1517.1和PB.1528.1,分别位于小鼠基因组的Cep131、F2RL1、

Gm5345和KLHL26基因座。这四个变异都在内含子区域表达了一个额外的片段作为新的外显子。剩下的转录本变异PB.1596.5位于小鼠基因组的基因间区域,作为一个新的转录本异常表达(如图5B)。所有这些变异最初都是使用HiFi长读段识别的,随后在各自的剪接位点通过NGS短读段进行验证。这些变异都没有在相应的配对正常组织中检测到,表明它们是肿瘤特异的。

[0164] 此外,本实施例使用GMST为这些变异生成了相应的蛋白质序列(如图5C),并通过BLAST与UniProt数据库tr_mouse_canon_isoform.fasta进行同源性分析。结果显示,前四个变异(包括PB.349.2、PB.458.1、PB.1517.1和PB.1528.1)最高同源蛋白的一致性值低于60%,说明它们具有新颖性。然而,PB.1596.5表现出不同的模式,比对结果显示其与最同源的蛋白质Q1KYM0的一致性值为99%(667/669)(如图5D)。已有报道Q1KYM0是一种在Neuro-2a肿瘤细胞中表达的内源性反转录病毒(ERV)蛋白,可导致免疫逃逸。用siRNA敲低表达Q1KYM0的基因可以延缓肿瘤生长并增加小鼠存活率。这些发现加强了支持PB.1596.5肿瘤特异性的证据,并强调了结合大规模和小规模变异检测以及随后进行新蛋白分析的重要性。

[0165] 最后,为进一步确认所识别变异结果的可靠性,设计了针对PB.1596.5的引物并提取RNA进行PCR验证。结果显示,在MC38细胞和MC38异基因移植物中都有显著的过表达,而对照正常组织则没有表达(如图5E)。所有结果表明新生蛋白检测方法可以分析全方位变异,并在识别长异构体方面表现出高性能。

[0166] 讨论

[0167] 在癌症免疫治疗的新时代,“新抗原”的概念已经广泛流行。新抗原可以被定义为源自肿瘤细胞中异常遗传变化的免疫原性蛋白质/表位。由于正常细胞不存在这种异常遗传变化,新抗原可以被视为肿瘤细胞的“绝对”特异性生物标志物。因此,近年来生物技术公司和研究机构一直在积极开发针对新抗原的免疫疗法。如今,如何从临床样本中有效识别新抗原已成为转化研究需要解决的一个重要问题。

[0168] 识别新抗原的“经典”方法是基于在规范的开放阅读框(ORFs)中搜索非同义突变。这种“经典”方法易于理解,但其局限性也很明显:只有具有高突变负荷的肿瘤在ORFs内部才有足够的非同义突变来诱导新蛋白;此外,具有免疫原性的新抗原只是新蛋白的一个很小的子集。由于许多癌症类型具有低肿瘤突变负荷(<5个突变/Mb),例如前列腺癌、胶质母细胞瘤、急性髓系白血病等。对于这些癌症类型,在规范ORFs中难以找到足够的新抗原用于免疫治疗。

[0169] 然而,由于异常的遗传变化,大多数类型的癌细胞涉及正常细胞中不存在的替代RNA剪接。对癌症基因组图谱数据的RNA-seq分析提供了令人信服的证据,证明了肿瘤特异性替代剪接事件的普遍性,这些事件产生的新抗原预计比错义突变更具免疫原性。此外,一些研究表明,RNA替代剪接提供了另一种新蛋白的来源,这些蛋白是从正常细胞中从未翻译的RNA序列中翻译而来的。因此,识别基于大规模突变的新蛋白和新抗原(如RNA替代剪接)将是突破“经典”方法限制的一种有效新方法,从而大大扩展针对新抗原的免疫疗法的应用。

[0170] 作为研究肽组或蛋白质组的强大工具,质谱(MS)理论上可以识别源自RNA替代剪接的新蛋白。事实上,MS的性能在很大程度上依赖于参考数据库的特征。正如在引言部分所

讨论的,规范的人类蛋白质组对于识别完全不属于规范人类蛋白质组的新蛋白来说是一个不利的参考数据库。为了解决上述逻辑悖论,普遍使用的HiFi和NGS技术提供了一种科学合理的方法来建立包含假设蛋白序列的样本特异性参考数据库,这些序列源自RNA翻译。

[0171] 在分析RNA测序数据时,全面检测所有可能导致蛋白质改变的变异是一个关键需求。在这方面,包括融合、单核苷酸多态性(SNP)和插入/缺失(INDEL)在内的多种机制都会导致新RNA转录本的产生,以及随后新蛋白的诱导。考虑到任务的复杂性,需要一个强大的生物信息学工具来处理RNA测序数据并覆盖所有类型的RNA转录本变异。

[0172] 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。

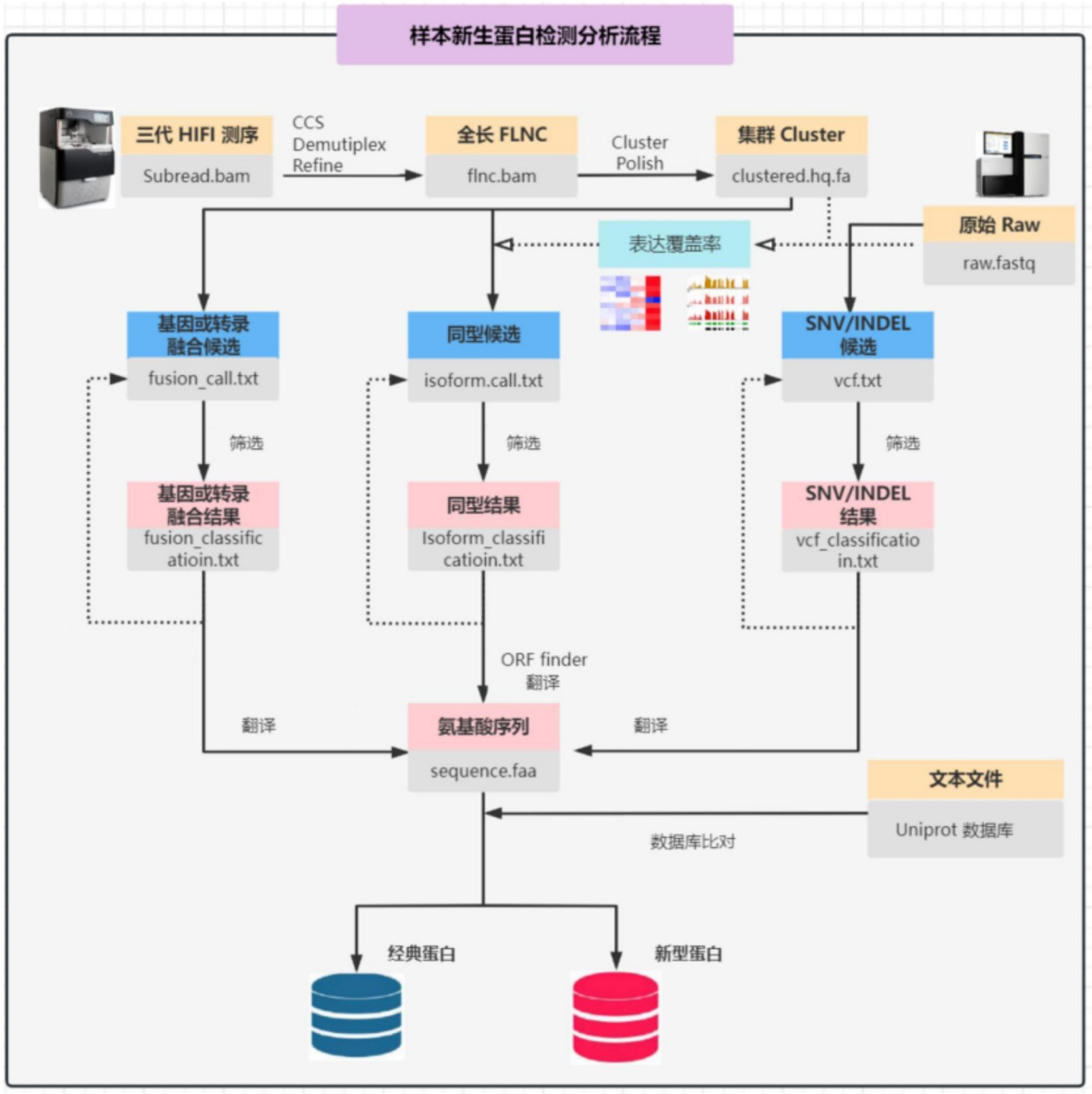
样本新生蛋白检测分析流程
三代 HIFI 测序
Subread.bam
CCS
Demutiplex
Refine
全长 FLNC
flnc.bam
Cluster
Polish
集群 Cluster
clustered.hq.fa
原始 Raw
raw.fastq
表达覆盖率
基因或转录
融合候选
fusion_call.txt
同型候选
isoform.call.txt
SNV/INDEL
候选
vcf.txt
筛选
筛选
筛选
基因或转录
融合结果
fusion_classific
atioin.txt
同型结果
Isoform_classifi
catioin.txt
SNV/INDEL
结果
vcf_classificatio
in.txt
ORF finder
翻译
翻译
翻译
氨基酸序列
sequence.faa
文本文件
Uniprot 数据库
数据库比对
经典蛋白
新型蛋白

图1

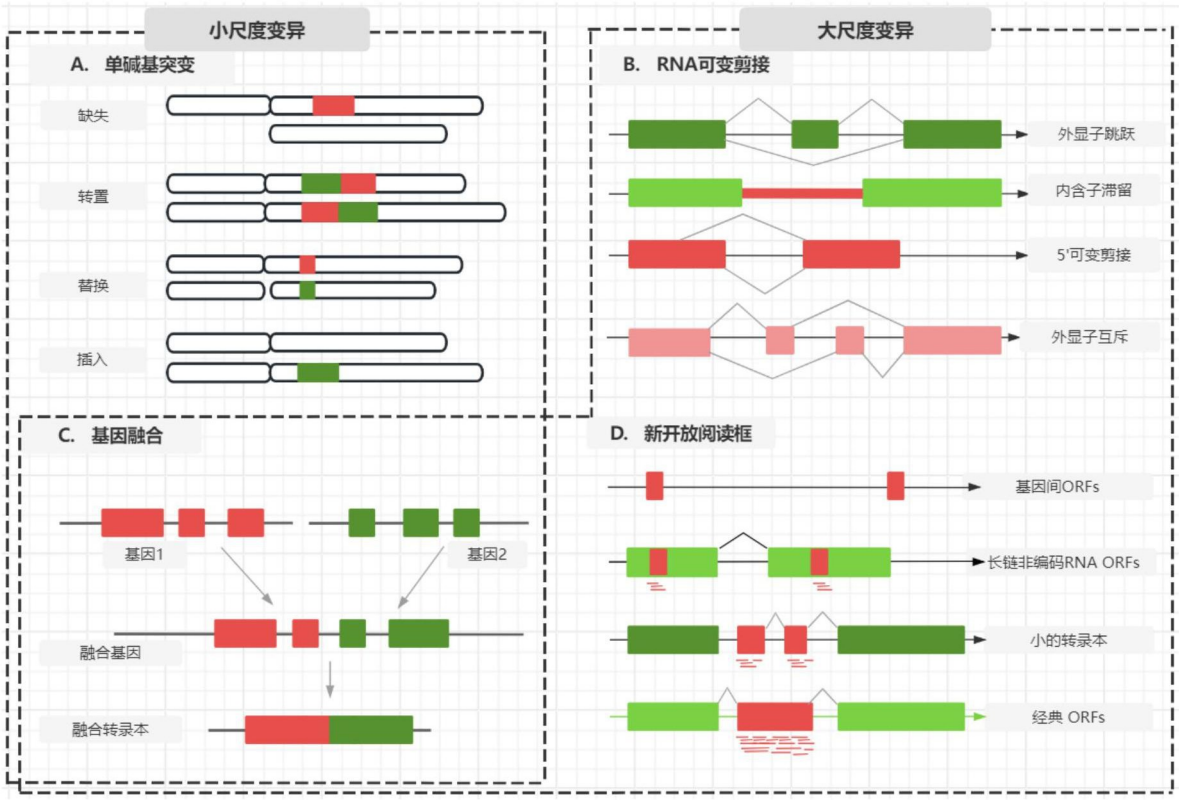
小尺度变异
A. 单碱基突变
缺失
转置
替换
插入
C. 基因融合
基因1
基因2
融合基因
融合转录本
大尺度变异
B. RNA可变剪接
外显子跳跃
内含子滞留
5'可变剪接
外显子互斥
D. 新开放阅读框
基因间ORFs
长链非编码RNA ORFs
小的转录本
经典 ORFs

图2

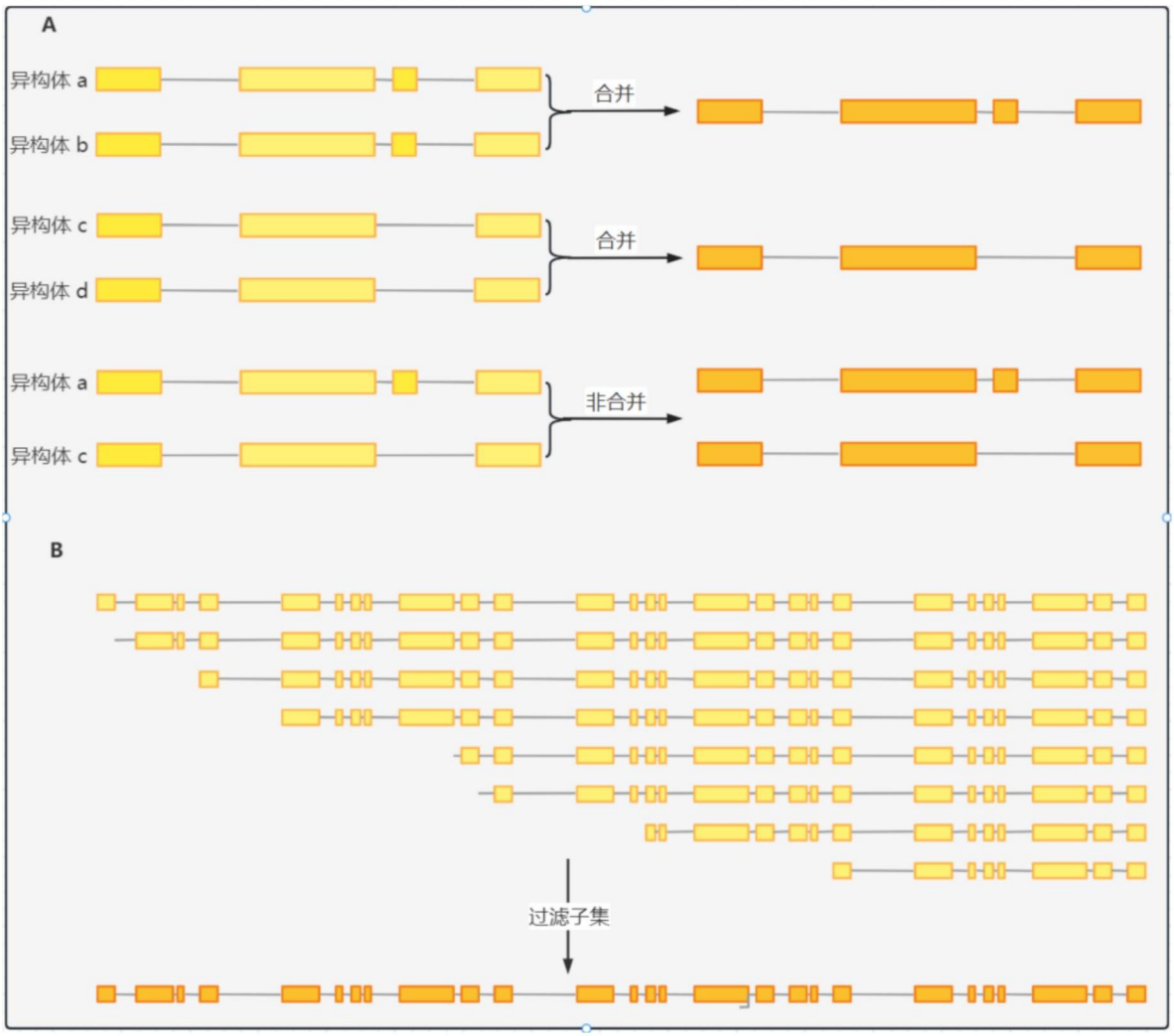

图3

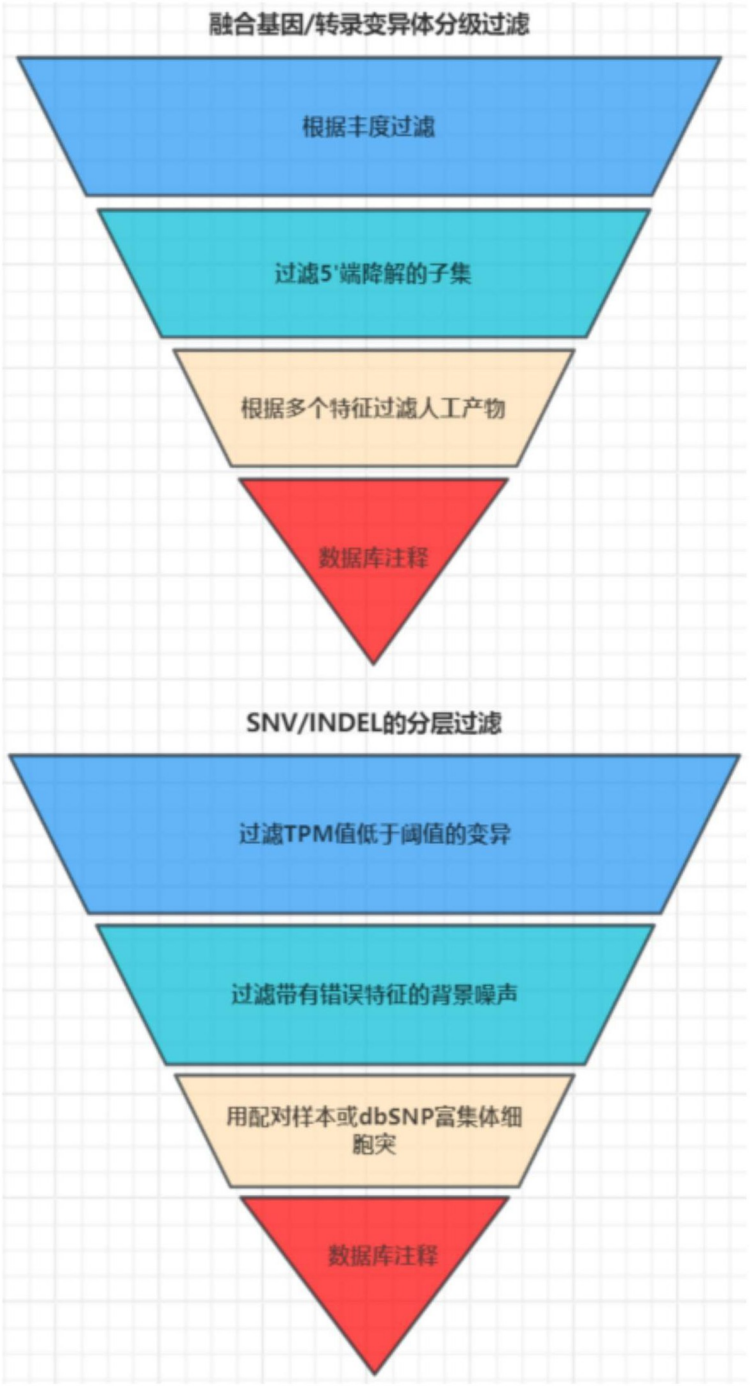
融合基因/转录变异体分级过滤
根据丰度过滤
过滤5'端降解的子集
根据多个特征过滤人工产物
数据库注释
SNV/INDEL的分层过滤
过滤TPM值低于阈值的变异
过滤带有错误特征的背景噪声
用配对样本或dbSNP富集体细胞突
数据库注释

图4

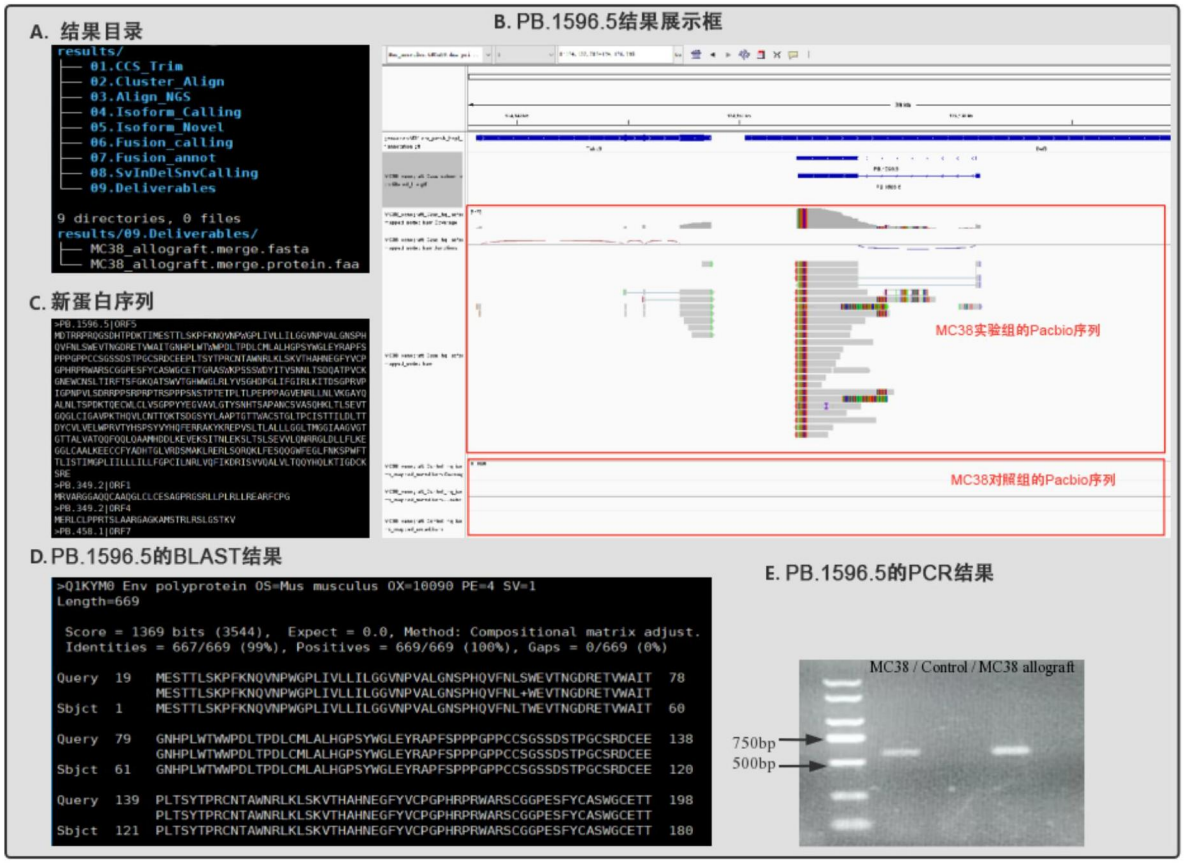
A. 结果目录
results/
01.CCS_Trim
02.Cluster_Align
03.Align_NGS
04.Isoform_Calling
05.Isoform_Novel
06.Fusion_calling
07.Fusion_annot
08.SvIndelSnvCalling
09.Deliverables
9 directories, 0 files
results/09.Deliverables/
MC38_allograft.merge.fasta
MC38_allograft.merge.protein.faa
B. PB.1596.5结果展示框
MC38实验组的Pacbio序列
MC38对照组的Pacbio序列
C. 新蛋白序列
>PB.1596.5|ORF5
>PB.349.2|ORF1
>PB.349.2|ORF4
>PB.458.1|ORF7
D. PB.1596.5的BLAST结果
>Q1KYM0 Env polyprotein OS=Mus musculus OX=10090 PE=4 SV=1
Length=669
Score = 1369 bits (3544), Expect = 0.0, Method: Compositional matrix adjust.
Identities = 667/669 (99%), Positives = 669/669 (100%), Gaps = 0/669 (0%)
Query 19 MESTTLSKPFKNQVNPWGPLIVLLILGGVNPVALGNSPHQVFNLSWEVTNGDRETVWAIT 78
MESTTLSKPFKNQVNPWGPLIVLLILGGVNPVALGNSPHQVFNL+WEVTNGDRETVWAIT
Sbjct 1 MESTTLSKPFKNQVNPWGPLIVLLILGGVNPVALGNSPHQVFNLTWEVTNGDRETVWAIT 60
Query 79 GNHPLWTWWPDLTPDLCMLALHGPSYWGLEYRAPFSPPPGPPCCSGSSDSTPGCSRDCEE 138
GNHPLWTWWPDLTPDLCMLALHGPSYWGLEYRAPFSPPPGPPCCSGSSDSTPGCSRDCEE
Sbjct 61 GNHPLWTWWPDLTPDLCMLALHGPSYWGLEYRAPFSPPPGPPCCSGSSDSTPGCSRDCEE 120
Query 139 PLTSYTPRCNTAWNRLKLSKVTHAHNEGFYVCPGPHRPRWARSCGGPESFYCASWGCETT 198
PLTSYTPRCNTAWNRLKLSKVTHAHNEGFYVCPGPHRPRWARSCGGPESFYCASWGCETT
Sbjct 121 PLTSYTPRCNTAWNRLKLSKVTHAHNEGFYVCPGPHRPRWARSCGGPESFYCASWGCETT 180
E. PB.1596.5的PCR结果
MC38 / Control / MC38 allograft
750bp
500bp

图5