(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 111201572 A
(43)申请公布日 2020.05.26

(21)申请号 201880065571.X

(22)申请日 2018.10.09

(30)优先权数据
62/570,580 2017.10.10 US
62/618,893 2018.01.18 US

(85)PCT国际申请进入国家阶段日
2020.04.08

(86)PCT国际申请的申请数据
PCT/US2018/055025 2018.10.09

(87)PCT国际申请的公布数据
WO2019/074933 EN 2019.04.18

(71)申请人 南托米克斯有限责任公司
地址 美国加利福尼亚州

(72)发明人 沙赫鲁兹·拉比扎德 查德·加纳 拉胡尔·帕鲁勒卡尔 克里斯托弗·W·赛托

(74)专利代理机构 北京柏杉松知识产权代理事务所(普通合伙) 11413
代理人 王春伟 刘继富

(51)Int.Cl.
G16B 20/00(2006.01)
G16B 30/00(2006.01)
G16H 50/20(2006.01)

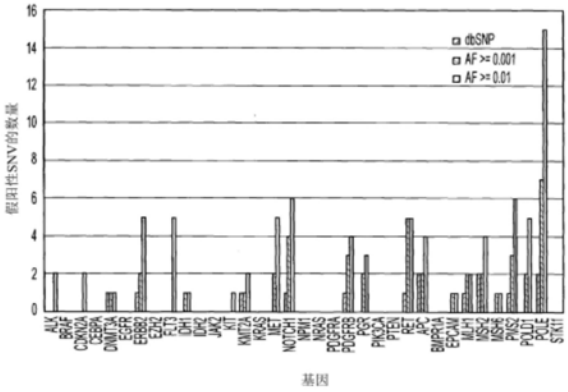
权利要求书2页 说明书25页 附图15页

(54)发明名称

用于提高精确度的癌症患者的综合基因组转录组肿瘤-正常样基因组套分析

(57)摘要

使用来自肿瘤样品和匹配的正常样品的DNA测序数据确定SNV,从而进行改进准确性的基于SNV的基因测试,并且使用来自肿瘤样品的RNA测序数据来确定如此鉴定的SNV的表达。



CN 111201572 A

1.一种以增加的准确性进行基于单核苷酸变体的癌症测试的方法,该方法包括:

从患者的肿瘤样品和匹配的正常样品获得DNA测序数据,并进一步从该肿瘤样品获得RNA测序数据;

相对于该匹配的正常样品,在该肿瘤样品中确定DNA单核苷酸变体的存在;

使用这些RNA测序数据确定这些DNA单核苷酸变体的表达;以及

基于这些单核苷酸变体的存在和表达,将至少一种DNA单核苷酸变体鉴定为与该患者的癌症状态相关。

2.如权利要求1所述的方法,其中这些DNA测序数据是全基因组DNA测序数据。

3.如权利要求1-2中任一项所述的方法,其中该肿瘤组织的DNA测序数据的读取深度为至少50x。

4.如权利要求1-3中任一项所述的方法,其中该匹配的正常组织的DNA测序数据的读取深度为至少30x。

5.如权利要求1-4中任一项所述的方法,其中使用来自该肿瘤样品和该匹配的正常样品的DNA测序数据的位置指导的同步比对来进行确定该DNA单核苷酸变体的存在的步骤。

6.如权利要求1-5中任一项所述的方法,该方法进一步包括使用这些DNA单核苷酸变体的等位基因频率过滤这些DNA单核苷酸变体。

7.如权利要求1所述的方法,其中该肿瘤组织的DNA测序数据的读取深度为至少50x。

8.如权利要求1所述的方法,其中该匹配的正常组织的DNA测序数据的读取深度为至少30x。

9.如权利要求1所述的方法,其中使用来自该肿瘤样品和该匹配的正常样品的DNA测序数据的位置指导的同步比对来进行确定该DNA单核苷酸变体的存在的步骤。

10.如权利要求1所述的方法,该方法进一步包括使用这些DNA单核苷酸变体的等位基因频率过滤这些DNA单核苷酸变体。

11.一种以增加的准确性鉴定患者的治疗选择的方法,该方法包括:

相对于该患者的匹配的正常样品,在该肿瘤样品中确定DNA单核苷酸变体的存在;

使用这些RNA测序数据确定这些DNA单核苷酸变体的表达;

鉴定如下治疗选择,该治疗选择靶向具有至少一种表达为RNA的DNA单核苷酸变体的基因。

12.如权利要求11所述的方法,其中使用来自该肿瘤样品和该匹配的正常样品的DNA测序数据的位置指导的同步比对来确定该DNA单核苷酸变体的存在。

13.如权利要求11所述的方法,其中使用具有肿瘤相关基因的多个参考序列的计算机模拟基因组套来确定该DNA单核苷酸变体的存在。

14.如权利要求11-12中任一项所述的方法,其中使用具有肿瘤相关基因的多个参考序列的计算机模拟基因组套来确定该DNA单核苷酸变体的存在。

15.如权利要求13所述的方法,其中该计算机模拟基因组套是癌症类型特异性的。

16.如权利要求13-14中任一项所述的方法,其中该计算机模拟基因组套是癌症类型特异性的。

17.如权利要求13所述的方法,其中这些肿瘤相关基因选自由以下组成的组:ABL1、EGFR、GNAS、KRAS、PTPN11、AKT1、ERBB2、GNAQ、MET、RB1、ALK、ERBB4、HNF1A、MLH1、RET、APC、

EZH2、HRAS、MPL、SMAD4、ATM、FBXW7、IDH1、NOTCH1、SMARCB1、BRAF、FGFR1、JAK2、NPM1、SMO、CDH1、FGFR2、JAK3、NRAS、SRC、CDKN2A、FGFR3、IDH2、PDGFRA、STK11、CSF1R、FLT3、KDR、PIK3CA、TP53、CTNNB1、GNA11、KIT、PTEN、VHL。

18.如权利要求13-16中任一项所述的方法，其中这些肿瘤相关基因选自由以下组成的组：ABL1、EGFR、GNAS、KRAS、PTPN11、AKT1、ERBB2、GNAQ、MET、RB1、ALK、ERBB4、HNF1A、MLH1、RET、APC、EZH2、HRAS、MPL、SMAD4、ATM、FBXW7、IDH1、NOTCH1、SMARCB1、BRAF、FGFR1、JAK2、NPM1、SMO、CDH1、FGFR2、JAK3、NRAS、SRC、CDKN2A、FGFR3、IDH2、PDGFRA、STK11、CSF1R、FLT3、KDR、PIK3CA、TP53、CTNNB1、GNA11、KIT、PTEN、VHL。

20.如权利要求11所述的方法，该方法进一步包括使用这些DNA单核苷酸变体的等位基因频率过滤这些DNA单核苷酸变体。

21.如权利要求11-18中任一项所述的方法，该方法进一步包括使用这些DNA单核苷酸变体的等位基因频率过滤这些DNA单核苷酸变体。

22.如权利要求11所述的方法，其中确定这些DNA单核苷酸变体的表达包括测量这些DNA单核苷酸变体的RNA表达水平和与预定阈值比较。

23.如权利要求11-21中任一项所述的方法，其中确定这些DNA单核苷酸变体的表达包括测量这些DNA单核苷酸变体的RNA表达水平和与预定阈值比较。

24.如权利要求22所述的方法，该方法进一步包括基于该RNA表达水平对这些DNA单核苷酸变体分等级。

25.如权利要求22-23中任一项所述的方法，该方法进一步包括基于该RNA表达水平对这些DNA单核苷酸变体分等级。

26.如权利要求22所述的方法，该方法进一步包括基于与该预定阈值的比较，将这些DNA单核苷酸变体分类为“表达组”或“未表达组”。

27.如权利要求22-25中任一项所述的方法，该方法进一步包括基于与该预定阈值的比较，将这些DNA单核苷酸变体分类为“表达组”或“未表达组”。

用于提高精确度的癌症患者的综合基因组转录组肿瘤-正常样基因组套分析

[0001] 本申请要求2017年10月10日提交的序列号为62/570,580的我们共同待决的美国临时专利申请和2018年1月18日提交的序列号为62/618,893的美国临时申请的优先权,两者均通过引用以其全文并入本文中。

技术领域

[0002] 本发明的领域是对组学数据进行谱分析,因为组学数据与癌症有关,尤其是因为其与减少因各种癌症的仅针对肿瘤的基因组套分析中的多态性所致的假阳性结果有关。

背景技术

[0003] 背景描述包括可用于理解本发明的信息。并不承认本文提供的任何信息是现有技术或与当前要求保护的发明相关,也不承认具体地或隐含地引用的任何出版物是现有技术。

[0004] 本文中的所有出版物和专利申请都通过引用并入,其程度如同每个单独的出版物或专利申请被具体地且单独地指明通过引用并入一样。在并入的参考文献中的术语的定义或用法与本文提供的该术语的定义不一致或相反的情况下,适用本文提供的该术语的定义,而不适用该术语在该参考文献中的定义。

[0005] 基于DNA测序的商购临床级基因组套测试已广泛用于临床实践中。这些基于仅针对肿瘤的分析的基于组套的测试目前是肿瘤学中用于基因组测试以提供临床决策支持的最常用方法。基于测序的方法试图鉴定驱使肿瘤生长的体细胞来源的基因组变异,并精确地将这些遗传变体与肿瘤基因组中不可避免地主要存在的遗传种系基因组变异的大背景区分开。

[0006] 2016年,美国医疗保险和医疗补助服务中心(Centers for Medicare and Medicaid Services;CMS)核准覆盖旨在为肺癌治疗提供信息的35种基因的基于仅针对肿瘤的DNA测序的测试。目前CMS批准的此测试基于靶向的基因组套的仅针对肿瘤的分析,其中特别排除了将此种分析与患者正常种系组织比较。相反地,当前批准的测试利用了参考基因组和过滤技术来区分‘真正’的体细胞变体与正常的多态性或遗传的种系变体。该测试(MolDX:L36194)被定义为“不能区分体细胞与种系改变的仅使用肿瘤组织(即,不是匹配的肿瘤和正常样)的单一测试”。然而,其他人已经报道了此仅针对肿瘤的方法增加了错误地将种系突变鉴定为体细胞来源的遗传变化和潜在的癌症驱动子突变(“假阳性”)的风险。尽管最近显示,通过分子病理学家对所有推定的体细胞变体进行的审查,可以至少在一定程度上降低与仅针对肿瘤的测序相关的假阳性率,但是此种单独的审查通常耗时并且仍然容易出错。

[0007] 因此,仍然需要用于分析来自癌症患者的组学数据的改进方法,尤其是在可能出现假阳性测试结果的情况下。

发明内容

[0008] 本发明主题涉及使用来自患者的肿瘤DNA、种系DNA和肿瘤RNA的基因组学和转录组学数据分析和/或鉴定肿瘤相关的单核苷酸变体(SNV)的各种方法,这些方法出乎意料地改进了准确性,并改进了有效治疗的机会。

[0009] 因此,在本发明主题的一方面,本发明人设想了一种以增加的准确性进行基于SNV的癌症测试的方法。此方法包括从肿瘤样品和匹配的正常样品(即,同一患者的非肿瘤样品)获得DNA测序数据的步骤,以及从该肿瘤样品获得RNA测序数据的另一个步骤。然后,该方法还包括相对于该匹配的正常样品确定该肿瘤样品中DNA单核苷酸变体的存在的步骤,以及使用这些RNA测序数据确定这些DNA单核苷酸变体的表达的步骤。在一些实施例中,使用来自该肿瘤样品和该匹配的正常样品的DNA测序数据的位置指导的同步比对来进行确定该DNA单核苷酸变体的存在的步骤。优选地,该方法还包括以下步骤:基于这些单核苷酸变体的存在和表达,将至少一个DNA单核苷酸变体鉴定为与患者的癌症状态相关。

[0010] 最典型地,这些DNA测序数据是全基因组DNA测序数据。优选地,该肿瘤组织的DNA测序数据的读取深度为至少50x,和/或该匹配的正常组织的DNA测序数据的读取深度为至少30x。在一些实施例中,该方法还包括使用这些DNA单核苷酸变体的等位基因频率过滤这些DNA单核苷酸变体的步骤。

[0011] 在本发明主题的另一方面,本发明人设想了一种以增加的准确性鉴定患者的治疗选择的方法。此方法包括相对于该患者的匹配的正常样品确定肿瘤样品中DNA单核苷酸变体的存在的步骤,以及使用RNA测序数据确定这些DNA单核苷酸变体的表达的步骤。然后,该方法还包括鉴定如下治疗选择的步骤,该治疗选择靶向具有至少一个表达为RNA的DNA单核苷酸变体的基因。

[0012] 优选地,使用来自该肿瘤样品和该匹配的正常样品的DNA测序数据的位置指导的同步比对来进行确定该DNA单核苷酸变体的存在的步骤。在一些实施例中,使用具有肿瘤相关基因的多个参考序列的计算机模拟基因组套来进行确定该DNA单核苷酸变体的存在的步骤。在此种实施例中,该计算机模拟基因组套优选是癌症类型特异性的,和/或这些肿瘤相关基因选自由以下组成的组:ABL1、EGFR、GNAS、KRAS、PTPN11、AKT1、ERBB2、GNAQ、MET、RB1、ALK、ERBB4、HNF1A、MLH1、RET、APC、EZH2、HRAS、MPL、SMAD4、ATM、FBXW7、IDH1、NOTCH1、SMARCB1、BRAF、FGFR1、JAK2、NPM1、SMO、CDH1、FGFR2、JAK3、NRAS、SRC、CDKN2A、FGFR3、IDH2、PDGFRA、STK11、CSF1R、FLT3、KDR、PIK3CA、TP53、CTNNB1、GNA11、KIT、PTEN、VHL。

[0013] 在一些实施例中,该方法还包括使用这些DNA单核苷酸变体的等位基因频率过滤这些DNA单核苷酸变体的步骤。

[0014] 在一些实施例中,确定这些DNA单核苷酸变体的表达的步骤包括测量这些DNA单核苷酸变体的RNA表达水平和与预定阈值比较。在此种实施例中,设想该方法还可以包括基于该RNA表达水平将这些DNA单核苷酸变体进行分等级的步骤和/或基于与该预定阈值的比较将这些DNA单核苷酸变体分类为“表达”或“未表达”组的步骤。

[0015] 在本发明主题的仍另一方面,本发明人设想了一种测试患者样品的方法,该方法包括从该患者的肿瘤和匹配的正常组织产生或获得DNA组学数据的步骤以及从该患者的肿瘤组织产生或获得RNA组学数据的另一个步骤。在又另一个步骤中,使用该匹配的正常组织的DNA组学数据在该肿瘤的DNA组学数据中鉴定肿瘤和患者特异性SNV,并将来自该肿瘤组

织的RNA组学数据用于确认该SNV的存在和该SNV的表达数量。

[0016] 优选地，这些DNA和/或RNA组学数据为BAM格式，并且使用递增同步比对(例如，使用可以使用这些DNA组学数据和这些RNA组学数据的BAMBAM)来进行该鉴定肿瘤和患者特异性SNV的步骤。最典型地但非必须地，这些RNA组学数据是RNAseq数据，和/或该肿瘤的DNA组学数据中的SNV在癌症驱动基因中或在遗传癌症风险基因中。例如，合适的癌症驱动基因包括ACT1、ACT2、ACT3、APC、ATM、BRAF、BRCA1、BRCA2、CHEK1、CHEK2、EGFR、ERBB2、ERBB3、ERBB4、FGFR1、FGFR2、FGFR3、HRAS、JAK3、KIT、KRAS、MET、NOTCH1、NRAS、PALB2、PDGFRA、PIC3CA、PTEN、SMO、SRC和TP53，并且合适的遗传癌症风险基因包括APC、ATM、AXIN2、BMPR1ACHD1、CHEK2、EPCAM、GREM1、MLH1、MSH2、MSH6、MUTYH、PMS2、POLD1、POLE、PTEN、SMAD4、STK11和TP53。

[0017] 在本发明主题的仍另一方面，本发明人设想了一种增加在患有肿瘤的患者中鉴定真正的体细胞单核苷酸中的准确性的方法。此方法包括以下步骤：从患者的肿瘤样品和匹配的正常样品获得DNA测序数据，并且另外从该肿瘤样品获得RNA测序数据，相对于该匹配的正常样品，确定该肿瘤样品中DNA单核苷酸变体的存在，相对于该匹配的正常样品确定该肿瘤样品中DNA单核苷酸变体的存在，以及基于这些单核苷酸变体的存在和表达将至少一个DNA单核苷酸变体鉴定为与该患者的癌症状态相关。

[0018] 最典型地，这些DNA测序数据是全基因组DNA测序数据。在一些实施例中，该肿瘤组织的DNA测序数据的读取深度为至少50x，和/或该匹配的正常组织的DNA测序数据的读取深度为至少30x。

[0019] 在一些实施例中，使用来自该肿瘤样品和该匹配的正常样品的DNA测序数据的位置指导的同步比对来进行确定该DNA单核苷酸变体的存在的步骤。在其他实施例中，该方法还可以包括使用这些DNA单核苷酸变体的等位基因频率过滤这些DNA单核苷酸变体的步骤。

[0020] 在一些实施例中，使用具有肿瘤相关基因的多个参考序列的计算机模拟基因组套来进行确定该DNA单核苷酸变体的存在的步骤。在此种实施例中，该计算机模拟基因组套优选是癌症类型特异性的，和/或这些肿瘤相关基因选自由以下组成的组：ABL1、EGFR、GNAS、KRAS、PTPN11、AKT1、ERBB2、GNAQ、MET、RB1、ALK、ERBB4、HNF1A、MLH1、RET、APC、EZH2、HRAS、MPL、SMAD4、ATM、FBXW7、IDH1、NOTCH1、SMARCB1、BRAF、FGFR1、JAK2、NPM1、SMO、CDH1、FGFR2、JAK3、NRAS、SRC、CDKN2A、FGFR3、IDH2、PDGFRA、STK11、CSF1R、FLT3、KDR、PIK3CA、TP53、CTNNB1、GNA11、KIT、PTEN、VHL。

[0021] 在一些实施例中，确定这些DNA单核苷酸变体的表达的步骤包括测量这些DNA单核苷酸变体的RNA表达水平和与预定阈值比较。在此类实施例中，还设想该方法还可以包括基于该RNA表达水平将这些DNA单核苷酸变体进行分等级，和/或基于与该预定阈值的比较将这些DNA单核苷酸变体分类为“表达组”或“未表达组”的步骤。

[0022] 本发明主题的各个目标、特征、方面和优点将根据以下关于优选实施例的详细描述以及附图而变得更清楚。

附图说明

[0023] 图1是描绘实例1中测试的45名肺癌患者中会发生的假阳性结果的数量的图。

[0024] 图2是描绘实例1中测试的所有癌症患者中会发生的假阳性结果的数量的图。

[0025] 图3是描绘实例1中测试的45名肺癌患者的真阳性和假阳性SNV的数量的图。

[0026] 图4是描绘实例1中测试的所有癌症患者的真阳性和假阳性SNV的数量的图。

[0027] 图5A至图5B是描绘实例2中针对胃肠道癌患者鉴定的体细胞和种系起源的SNV的数量的图

[0028] 图6A至图6B是描绘实例2中用等位基因频率过滤的真阳性和假阳性SNV的数量相对于基因的图。

[0029] 图7是描绘实例2中用等位基因频率过滤的真阳性和假阳性SNV的数量相对于患者的图。

[0030] 图8是描绘实例2中通过RNA表达分析鉴定的胃肠道癌患者中真阳性和假阳性SNV的数量的图。

[0031] 图9是描绘实例3中对基因组学和/或转录组学数据进行分析的肿瘤样品的数量相对于肿瘤类型的图。

[0032] 图10是描绘实例3中各种类型的癌症患者中鉴定的体细胞和种系起源的SNV的图。

[0033] 图11是描绘实例3中用等位基因频率过滤的真阳性和假阳性SNV的图。

[0034] 图12是描绘实例3中表达或未表达的错义/无义SNV的数量的图。

[0035] 图13是描绘在实例3中表达或未表达的体细胞SNV的数量的图。

具体实施方式

[0036] 本发明人出乎意料地发现，通过常规肿瘤DNA分析鉴定的单核苷酸变体(SNV)具有包括假阳性和/或假阴性的SNV的高风险，因为鉴定的大多数此类SNV是种系起源的变体。本发明人还发现，许多鉴定的体细胞SNV未表达为RNA，因此将此类未表达的体细胞SNV鉴定为肿瘤治疗的分子靶标会导致无效的癌症治疗。从不同的角度来看，本发明人现在已经发现，可以通过同时相对于匹配的正常样进行肿瘤基因组DNA的生物信息学分析以鉴定体细胞SNV，以及进行肿瘤RNA表达的生物信息学分析以鉴定表达或未表达的体细胞SNV，来显著增加基于单核苷酸变体的癌症测试的准确性。因此，本发明人设想在肿瘤中表达的此类鉴定的体细胞SNV可能与癌症状态相关，并且进一步被鉴定为肿瘤治疗的有效靶标。

[0037] 如本文所用，术语“肿瘤”是指以下项并且可与以下项互换使用：一种或多种癌细胞、癌组织、恶性肿瘤细胞或恶性肿瘤组织，它们可以位于或者发现于人体的一个或多个解剖位置中。应注意，如本文所用的术语“患者”包括诊断为患有病症(例如，癌症)的个体以及出于检测或鉴定病症的目的经历检查和/或测试的个体两者。因此，患有肿瘤的患者是指诊断为患有癌症的个体以及被怀疑患有癌症的个体二者。如本文所用，术语“提供(provide)”或“提供(providing)”是指并且包括制造、产生、放置、使得能使用、转移或使得即可使用的任何行为。

[0038] 因此，在本发明主题的一个尤其优选的方面，本发明人设想可以通过从患者的肿瘤样品和/或匹配的正常样品获得DNA和RNA数据患者以因此相对于匹配的正常样品确定肿瘤样品中的DNA单核苷酸变体并确定DNA单核苷酸变体的表达，来显著增加基于单核苷酸变体的癌症测试的准确性。设想表达为RNA的DNA单核苷酸变体可能高度准确地与患者的癌症状况相关。

[0039] 获得组学数据

[0040] 设想到了获得来自患者的肿瘤样品(肿瘤细胞或肿瘤组织)(或来自患者或健康个体的健康组织作为比较)的任何合适的方法。最典型地,可以经由活检(包括液体活检、或在手术或独立的活检程序期间经由组织切除术获得的等)获得来自患者的肿瘤样品,其可以是新鲜的或经处理的(例如,冷冻的等),直到用于从组织获取组学数据的进一步过程为止。例如,肿瘤细胞或肿瘤组织可以是新鲜的或冷冻的。又例如,肿瘤细胞或肿瘤组织可以呈细胞/组织提取物的形式。在一些实施例中,肿瘤样品可以从单个或多个不同的组织或解剖区域获得。例如,可以从患者的乳房以及其他器官(例如,肝、脑、淋巴结、血液、肺等)获得转移性乳癌组织以用作转移的乳癌组织。优选地,可以获得患者的健康组织或匹配的正常组织(例如,患者的非癌性乳房组织),或者还可以经由类似的方式获得来自健康个体(非患者)的健康组织作为比较样。

[0041] 在一些实施例中,为了确定在相关时间段内肿瘤样品的任何变化,可以在多个时间点从患者获得肿瘤样品。例如,可以在样品被确定或诊断为癌性之前和之后获得肿瘤样品(或疑似肿瘤样品)。在另一个实例中,可以在一次或一系列抗肿瘤治疗(例如,放疗、化疗、免疫疗法等)之前、期间和/或之后(例如,完成后等)获得肿瘤样品(或疑似肿瘤样品)。在仍另一个实例中,在鉴定新的转移的组织或细胞后,可以在肿瘤进展期间获得肿瘤样品(或疑似肿瘤样品)。

[0042] 从获得的肿瘤细胞或肿瘤组织中,可以将DNA(例如,基因组DNA、染色体外DNA等)、RNA(例如,mRNA、miRNA、siRNA、shRNA等)和/或蛋白质(例如,膜蛋白、胞质蛋白、核蛋白等)分离并且进一步分析以获得组学数据。替代性地和/或另外地,获得组学数据的步骤可以包括从存储一位或多位患者和/或健康个体的组学信息的数据库接收组学数据。例如,可以从患者肿瘤组织中分离的DNA、RNA和/或蛋白质获得患者肿瘤的组学数据,并且所获得的组学数据可以与患有相同类型的肿瘤或不同类型的肿瘤的其他患者的其他组学数据集一起存储在数据库(例如,云数据库、服务器等)中。从健康个体或患者的匹配的正常组织(或健康组织)获得的组学数据也可以存储在数据库中,使得在分析时可以从数据库中检索相关的数据集。同样,在获得蛋白质数据的情况下,这些数据也可以包括蛋白质活性,尤其是在蛋白质具有酶活性(例如,聚合酶、激酶、水解酶、裂解酶、连接酶、氧化还原酶等)的情况下。

[0043] 如本文所用,组学数据包括但不限于与基因组学、蛋白质组学和转录组学以及特定的基因表达或转录物分析和细胞的其他特征和生物学功能相关的信息。关于基因组学数据,合适的基因组学数据包括DNA序列分析信息,其可以通过肿瘤和匹配的正常样品的全基因组测序和/或外显子组测序(典型地在至少10x、更典型地至少20x的覆盖深度下)获得。替代性地,还可以从来自先前序列测定的已建立的序列记录(例如,SAM、BAM、FASTA、FASTQ或VCF文件)提供DNA数据。因此,数据集可以包括未处理的或已处理的数据集,并且示例性数据集包括具有BAM格式、SAM格式、FASTQ格式或FASTA格式的那些。然而,尤其优选的是,以BAM格式或作为BAMBAM diff对象提供数据集(例如,US 2012/0059670A1和US 2012/0066001A1)。组学数据可以源自全基因组测序、外显子组测序、转录组测序(例如,RNA-seq)或源自基因特异性分析(例如,PCR、qPCR、杂交、LCR等)。同样,可以按多种方式进行序列数据的计算分析。然而,在最优选的方法中,如例如在US 2012/0059670A1和US 2012/0066001A1中披露的使用BAM文件和BAM服务器通过对肿瘤和正常样品进行位置引导的同步比对在计算机中进行分析。这样的分析有利地减少假阳性新表位,并且显著地减少对存储

器和计算资源的需求。

[0044] 应当注意，应该读取针对计算机的任何语言，以包括任何合适的计算装置的组合，这些计算装置包括服务器、接口、系统、数据库、代理、端、引擎、控制器或单独或共同操作的其他类型的计算装置。应当理解，计算装置包括处理器，该处理器被配置为执行存储在有形的非暂时性计算机可读存储介质(例如，硬盘驱动器、固态驱动器、RAM、闪存、ROM等)上的软件指令。软件指令优选地配置计算装置，以提供如下文关于所披露的设备所讨论的角色、职责或其他功能。此外，所披露的技术可以体现为计算机程序产品，其包括存储软件指令的非暂时性计算机可读介质，这些软件指令使处理器执行与基于计算机的算法、过程、方法或其他指令相关联的所披露的步骤。在特别优选的实施例中，各种服务器、系统、数据库或接口交换数据使用标准化协议或算法，可能地基于HTTP、HTTPS、AES、公钥-私钥交换、web服务API、已知的金融交易协议、或其他电子信息交换方法。设备之间的数据交换可以通过以下进行：分组交换网络，即因特网、LAN、WAN、VPN或其他类型的分组交换网络；电路交换网络；信元交换网络；或其他类型的网络。

[0045] 相对于匹配的正常样品，肿瘤样品中的DNA单核苷酸变体

[0046] 设想可以通过比较从患者的肿瘤组织和匹配的正常组织(例如，患者的非肿瘤组织，包括非肿瘤血液样品的液体活检)获得的基因组DNA序列，从种系SNV区分并鉴定体细胞SNV。关于患者的肿瘤和匹配的正常组织的分析，认为许多方式适用于本文，只要此类方法将能够在肿瘤和匹配的正常序列之间产生差异序列对象或位置特异性差异的其他识别即可。示例性方法包括与外部参考序列(例如，hg18或hg19)的序列比较，或与内部参考序列(例如，匹配的正常序列)的序列比较以及对已知的常见突变模式(例如，SNV)进行的序列处理。因此，检测肿瘤与匹配的正常样、肿瘤与液体活检以及匹配的正常样与液体活检之间的突变的设想方法和程序包括iCallSV(URL:github.com/rhshah/iCallSV)、VarScan(URL:varscan.sourceforge.net)、MuTect(URL:github.com/broadinstitute/mutect)、Strelka(URL:github.com/Illumina/strelka)、Somatic Sniper(URL:gmt.genome.wustl.edu/somatic-sniper/)和BAMBAM(US 2012/0059670)。

[0047] 然而，在本发明主题的尤其优选的方面，通过第一序列数据(肿瘤样品)与第二序列数据(匹配的正常样)的增量同步比对来进行序列分析，例如使用如例如描述于Cancer Res[癌症研究]2013年10月1日;73(19):6036-45、US 2012/0059670和US 2012/0066001的算法以如此产生患者和肿瘤的特定突变数据。如将容易理解的，序列分析也可以按这样的方法进行，将来自肿瘤样品的组学数据与匹配的正常组学数据进行比较，以便如此进行如下分析，其不仅可以告知使用者对于患者体内的肿瘤而言真正的突变，还可以告知使用者在治疗期间新出现的突变(例如，经由匹配的正常和匹配的正常/肿瘤的比较、或经由肿瘤的比较)。另外，使用此类算法(尤其是BAMBAM)，可以容易地确定特定突变的等位基因频率和/或克隆群体，这可以有利地提供关于特定肿瘤细胞部分或群体的治疗成功的指示。因此，组学数据分析可以揭示错义和无义突变、拷贝数变化、杂合性丢失、缺失、插入、倒位、易位、微卫星变化等。

[0048] 此外，应当注意，数据集优选地反映了同一患者的肿瘤和匹配的正常样品，以便如此获得患者和肿瘤的具体信息。因此，可以排除不引起肿瘤的遗传种系改变(例如，沉默突变、SNP等)。当然，应该认识到，肿瘤样品可能来自初始肿瘤，来自治疗开始后的肿瘤，来自

复发的肿瘤或转移位点等。在大多数情况下，患者的匹配的正常样品可能是血液或来自与肿瘤相同的组织类型的未患病的组织。

[0049] 在一些实施例中，在将肿瘤和匹配的正常样的整个基因组或外显子组测序数据与外部参考序列进行比较的情况下，设想外部参考序列被组织为计算机模拟基因组套。优选地，计算机模拟基因组套包括多个肿瘤相关基因，包括一个或多个肿瘤驱动基因或一个或多个癌症驱动基因(例如，EGFR、KRAS、TP53、APC等)和/或药物敏感性或代谢相关基因。设想计算机模拟基因组套中基因的数量和类型可以根据患者可能患有或被诊断出的癌症类型(例如，癌症类型特异性计算机模拟基因组套)而变化，并且优选包括至少20个基因、至少30个基因、至少40个基因或至少50个基因。例如，计算机模拟基因组套可以包括以下的完整基因组序列和/或完整外显子组序列：ABL1、EGFR、GNAS、KRAS、PTPN11、AKT1、ERBB2、GNAQ、MET、RB1、ALK、ERBB4、HNF1A、MLH1、RET、APC、EZH2、HRAS、MPL、SMAD4、ATM、FBXW7、IDH1、NOTCH1、SMARCB1、BRAF、FGFR1、JAK2、NPM1、SMO、CDH1、FGFR2、JAK3、NRAS、SRC、CDKN2A、FGFR3、IDH2、PDGFRA、STK11、CSF1R、FLT3、KDR、PIK3CA、TP53、CTNNB1、GNA11、KIT、PTEN、VHL。

[0050] 另外，还设想使用DNA等位基因频率(例如，使用具有报道的群体等位基因频率的公共数据库)进一步过滤此类鉴定的DNA单核苷酸变体。在一些实施例中，可以用预定频率阈值，例如报道的≥0.01(1%)，优选≥0.005(0.5%)或更优选≥0.001(0.1%)的等位基因频率过滤DNA单核苷酸变体。

[0051] 另外，序列变化(DNA单核苷酸变体)的显著性可以通过变体识别(variant calling)来评估，其中基因组数据为BAM文件格式。由于BamBam使整个基因组中的文件对中的序列数据保持同步，所以可以轻松地实现复杂的突变模型，该模型需要来自源于两个生物样品的两个BAM文件的测序数据以及参考序列。此模型旨在最大化两个生物样品的两个序列串的联合概率。为了从两个生物样品找到两个序列串的最佳基因型，本发明人的目标是最大化由以下定义的可能性：

[0052] $P(D_g,D_t,G_g,G_t|\alpha,r)=P(L)_g|Gg)P(G_g|r)P(D_t|G_g,G_t,\alpha)P(G_t|G_g)$ (1)

[0053] $$P(D_{\downarrow g},D_{\downarrow t},G_{\downarrow g},G_{\downarrow t}|\alpha,r)=P(D_{\downarrow g}|G_{\downarrow g})P(G_{\downarrow g}|r)P(D_{\downarrow t}|G_{\downarrow g},G_{\downarrow t},\alpha)P(G_{\downarrow t}|G_{\downarrow g}$$) (1)

[0054] 其中r是观察到的参考等位基因，α是正常污染的分率，并且序列串1和2的基因型分别由$Gt=(t_1,t_2)$和$Gg=(g_1,g_2)$定义，其中t_1、t_2、g_2、$g_2\varepsilon\{A,T,C,G\}$。序列串1和2的序列数据分别被定义为读段组$D_t=\{d_t^1,d_t^2,...,d_t^m\}$和$D_g=\{d_g^1,d_g^2,...,dg^m\}$，其中观察到的碱基$d_t^i,d_g^i\varepsilon\{A,T,C,G\}$。模型中使用的所有数据必须超过用户定义的碱基和映射质量阈值。

[0055] 给定种系基因型的种系等位基因的概率被建模为以下四个核苷酸的多项式：

[0056] $$P(D_g|G_g)\approx\frac{n!}{n_A!n_T!n_G!n_C!}\prod_i^n P(d_g^i|G_g),$$

[0057] 其中，n是此位置上种系读段的总数，并且n_A、n_G、n_C、n_T是支持每个观察到的等位基因的读段。假设碱基概率$P(d_g^i|G_g)$是独立的，来自基因型G_g代表的两个亲本等位基因中的任一个，同时还掺入了测序仪的近似碱基错误率。序列串1基因型的先验概率取决于参考碱基，为：

[0058] $P(G_g|r=a)=\{\mu_{aa},\mu_{ab},\mu_{bb}\}$

[0059] 其中，μ_{aa}是该位置是纯合子参考的概率，μ_{ab}是该位置是杂合子参考的概率，并且μ_{bb}是该位置是纯合子非参考的概率。此时，序列串1先验概率不掺入关于已知遗传的SNP的任何信息。

[0060] 再次，序列2读段组的概率被定义为如下多项式：

[0061] $$P(D_t \mid D_t, G_g, \alpha) = \frac{n!}{n_A! n_T! n_G! n_C!} \prod_j^n P(d_g^j \mid G_t, G_g, \alpha)_i$$

[0062] 其中m是此位置处种系读段的总数，并且m_A、m_G、m_G、m_T是支持序列2数据集中每个观察到的等位基因的读段，并且每个序列2读段的概率是从序列2和序列1基因型得出的碱基概率的混合，受正常污染分率α控制，为

[0063] $P(d_t^i \mid G_t, G_g\alpha) = \alpha P(d_t^i \mid G_t) + (1-\alpha) P(d_t^i \mid G_g)$

[0064] 并且序列2基因型的概率由序列1基因型上的简单突变模型定义

[0065] $P(G_t \mid G_g) = \max[P(t_1 \mid g_1) P(t_2 \mid g_2), P(t_1 \mid g_2) P(t_2 \mid g_1)]$，

[0066] 其中无突变(例如，t1＝g1)的概率最大，并且转换(即，A→G，T→C)的概率可能比颠换(即，A→T，T→G)大四倍。用户可以定义多项式分布的所有模型参数α、μaa、μab、μbb和碱基概率P(di|G)。

[0067] 选择的序列2和1基因型Gt max、Gg maxi是最大化的基因型(1)，并且由以下定义的后验概率

[0068] $$\frac{P(D_g, D_t, G_g^{max}, G_t^{max} \mid \alpha, r)}{\sum_{i,j} P(D_g, D_t, G_g = i, G_t = j \mid \alpha, r)}$$

[0069] 可用于对一对推断的基因型进行置信度评分。如果序列2和序列1的基因型不同，则将报告序列2中的突变以及其相应的置信度。

[0070] 最大化序列1和2基因型中的一个或两个的可能性有助于改进两个推断的基因型的准确性，尤其是在一个或两个序列数据集对特定基因组位置的覆盖率较低的情况下。当非参考或突变等位基因的支持率较低时，分析单个测序数据集的其他突变识别算法，诸如MAQ和SNVMix更可能犯错(Li,H.等人,(2008)Mapping short DNA sequencing reads and calling variants using mapping quality scores[映射短的DNA测序读段并使用映射质量得分来识别变体],Genome Research[基因组研究],11,1851-1858;Goya,R.等人,(2010)SNVMix:predicting single nucleotide variants from next-generation sequencing of tumors[SNVMix:从下一代肿瘤测序预测单核苷酸变体],Bioinformatics[生物信息学],26,730-736)。

[0071] 除了在给定的基因组位置从所有读段收集等位基因支持率外，还会收集关于读段的信息(诸如，读段映射到正向链或反向链的哪条链、读段内等位基因的位置、等位基因的平均质量等)，并用于选择性地过滤掉假阳性识别。我们期望链和所有支持变体的等位基因的等位基因位置为随机分布，并且如果该分布与此随机分布有明显的偏差(即，发现所有变体等位基因都位于读段的尾端附近)，则此表明该变体识别是可疑的。

[0072] 还设想，还可以通过其他分析工具来进行序列变化的变体识别，这些工具包括但不限于MuTect(Nat Biotechnol.[自然生物技术]2013年3月;31(3):213-9)、MuTect2、HaploTypeCaller、Strelka2(Bioinformatics[生物信息学],第28卷,第14期,2012年7月15

日,第1811-1817页)或其他基因组假象检测工具。

[0073] DNA单核苷酸变体的表达

[0074] 另外,肿瘤和/或匹配的正常样的组学数据包含转录组数据集,该转录组数据集包括从患者获得的一种或多种RNA(优选细胞mRNA)的序列信息和表达水平(包括表达谱分析或剪接变体分析)。本领域中已知许多转录组学分析方法,并且认为所有已知方法适用于本文(例如,RNAseq、RNA杂交阵列、qPCR等)。因此,优选的材料包括mRNA和初级转录物(hnRNA),并且RNA序列信息可以从逆转录的polyA$^+$-RNA获得,该逆转录的polyA$^+$-RNA进而从同一患者的肿瘤样品和匹配的正常(健康的)样品获得。同样,应当注意,尽管通常优选polyA$^+$-RNA作为转录组的代表,但是其他形式的RNA(hn-RNA、非多聚腺苷酸化RNA、siRNA、miRNA等)也被认为适用于本文。优选的方法包括定量RNA(hnRNA或mRNA)分析和/或定量蛋白质组学分析,尤其是包括RNAseq。在其他方面,使用基于RNA-seq、qPCR和/或rtPCR的方法进行RNA定量和测序,尽管多种替代性的方法(例如,基于固相杂交的方法)也被认为是合适的。从另一个角度来看,转录组学分析(单独地或与基因组分析组合)可能适合于鉴定和定量具有癌症特异性突变和患者特异性突变的基因。

[0075] 优选地,转录组学数据集包括等位基因特异性序列信息和拷贝数信息。在此种实施例中,转录组学数据集包括基因的至少一部分的所有读段信息,优选至少10x、至少20x或至少30x。使用动态开窗方法来计算等位基因特异性拷贝数,更具体地,多数和少数拷贝数,该方法根据种系数据中的覆盖率扩大和缩小窗口的基因组学宽度,如在US 9824181中所详细描述的,该案通过引用并入本文中。如本文所用,多数等位基因是具有多数拷贝数(>总拷贝数的50%(读段支持)或最多的拷贝数)的等位基因,并且少数等位基因是具有少数拷贝数(<总拷贝数的50%(读段支持)或最少的拷贝数)的等位基因。

[0076] 本发明人设想在一些实施例中,可以通过RNA测序数据(例如,RNAseq)来确定具有一个或多个单核苷酸变体的基因(或基因的一部分)的表达。在此类实施例中,可以将一个或多个单核苷酸变体的表达评估为表达的RNA中一个或多个单核苷酸变体的存在或不存在。因此,基于RNA测序数据,可以将一个或多个单核苷酸变体分组为“表达组”或“未表达组”。在其他实施例中,具有一个或多个单核苷酸变异体的基因(或基因的一部分)的表达可以通过组合RNAseq数据与RNA定量数据(例如,使用qPCR和/或rtPCR)来确定。在此类实施例中,可以通过与预定阈值比较将一个或多个单核苷酸变体的表达水平评估为存在或不存在。设想预定阈值可以根据基因而变化。例如,预定阈值可以是健康个体的相同或类似类型的组织(例如,肝、肺等)中基因的平均RNA表达水平或患者的匹配的正常组织中基因的RNA表达水平的10%、5%或1%。替代性地,预定阈值可以根据给定的一个或多个反应中的qPCR和/或rtPCR噪声水平而变化。例如,预定阈值可以在qPCR和/或rtPCR反应的噪声水平的20%内、10%内、5%内。因此,基于RNA表达水平,可以将一个或多个单核苷酸变体分组为表达水平为预定阈值或高于预定阈值的“表达组”,或表达水平低于预定阈值的“未表达组”。

[0077] 在不希望受任何特定理论限制的情况下,本发明人设想,组合基因组学数据和转录组学数据以鉴定表达的DNA单核苷酸变体会显著降低假阳性率(错误地将种系突变鉴定为体细胞来源的癌症驱动子突变,和/或鉴定未表达为有效突变的体细胞来源的癌症驱动子突变等)和/或假阴性率(例如,排除真正的肿瘤体细胞SNV等)。在鉴定肿瘤相关基因中的DNA单核苷酸变体时,假阳性和/或假阴性率的降低进一步显著增加了鉴定与肿瘤和/或癌

症相关的基因以及鉴定具有降低的不良副作用或毒性的任何有效治疗方案的效率和准确性，因为在分析或应用的相对早期阶段，待分析和靶向的与肿瘤或癌症相关的表达的DNA单核苷酸变体的数量可能显著减少。

[0078] 因此，本发明人进一步设想，基于单核苷酸变体的存在/不存在和表达，可以将此类单核苷酸变体鉴定为与癌症相关的变体(或突变)，此类单核苷酸变体可以进一步与患者的癌症状态关联。如本文所用，术语“癌症状态”是指癌症或肿瘤的任何分子、生理、病理状况。因此，癌症状态可以包括癌症的解剖学类型(例如，胃肠道癌、肺癌、脑肿瘤等)、肿瘤的转移状态(例如，已转移、高趋势转移、未转移等)、肿瘤克隆性、肿瘤组织的免疫状态(例如，免疫遏制、免疫活化、免疫休眠等)、肿瘤的预后(例如，肿瘤的阶段、肿瘤的等级，包括肿瘤的形态发生等)。另外，癌症状态可以包括肿瘤对肿瘤治疗的敏感性或抗性(例如，对检查点抑制剂给予的抗性、对细胞因子治疗的敏感性等)、化学治疗药物的毒性(例如，由于CYP2D6酶介导的路径的成分中的突变/单个核苷酸变体等)。

[0079] 在一些实施例中，可以通过提供一个或多个显著性得分来将表达的DNA单核苷酸变体与肿瘤或癌症状态的相关性定量。例如，可以通过组合以下来确定显著性得分：DNA单核苷酸变体数量的单项得分(每一个核酸改变1分)、DNA单核苷酸变体的类型(例如，无义突变、错义突变等)、DNA单核苷酸变体的位置(例如，编码功能性结合结构域的基因的外显子3等)和生理影响(信号传导路径B的主要负因子)。同样，显著性分数可以通过包括DNA单核苷酸变体的基因的表达来确定(例如，每个未表达的DNA单核苷酸变体为-1，每个表达的DNA单核苷酸变体为+1，或基于包括DNA单核苷酸变体的基因的表达水平得出的各种增量得分，诸如包括DNA单核苷酸变体的基因的表达每增加10％，得1分等)。因此，在此类实施例中，可以基于表达(RNA中存在或不存在)或表达水平(与正常组织或健康个体相比RNA表达水平的增加或降低)来对DNA单核苷酸变体的显著性分等级。替代性地和/或另外地，包括DNA单核苷酸变体的基因的一个或多个显著性得分可以用于进一步对基因或DNA单核苷酸变体分等级。

[0080] 发明人还设想，此类鉴定和/或分等级的DNA单核苷酸变体和/或包括DNA单核苷酸变体的基因还可以用于鉴定治疗患者的癌症或肿瘤的治疗选择。例如，在确认RNA中有DNA单核苷酸变体(通过肿瘤匹配的正常样测序鉴定)之后，并且在具有一个或多个DNA单核苷酸变体的肿瘤相关基因中确认RNA表达(例如，与匹配的正常样相比至少25％，与匹配的正常样相比至少50％，与匹配的正常样相比至少75％，与匹配的正常样相比至少100％，与匹配的正常样相比至少为125％，或与匹配的正常样相比至少150％)之后，将靶向该肿瘤相关基因的药物以有效治疗该肿瘤的剂量和方案给予患者。如本文所用，靶向肿瘤相关基因的药物可以包括调节基因表达(在转录水平或翻译水平上)的药物、调节基因产物(蛋白质)的翻译后修饰的药物、调节基因产物(蛋白质)活性的药物或调节基因产物(蛋白质)降解的药物。

[0081] 如本文所用，术语“给予”药物或癌症治疗是指直接或间接给予药物或癌症治疗两者。药物或癌症治疗的直接给予典型地由卫生保健专业人员(例如，医生、护士等)进行，并且其中间接给药包括向卫生保健专业人员提供或使其可获得药物或癌症治疗以进行直接给予(例如，通过注射、口服、局部施加等)的步骤。

[0082] 实例1

[0083] 目前批准的用于肺癌的测试基于对靶向的基因组套的仅针对肿瘤的分析，其中特别排除了患者的正常种系组织。然而，如下文更详细显示，仅针对肿瘤的方法会大大增加将种系突变错误地识别为体细胞来源的癌症驱动子突变(即，假阳性)的风险，并且进一步不能告知医生潜在的可药物治疗靶标甚至在肿瘤中以有意义的量存在于何处。

[0084] 更具体地，本发明人发现在当前批准的针对肺癌患者的仅针对肿瘤的基因组套分析中发现的所有变体中有94％实际上是假阳性多态性，并且在严格过滤后仍然有48％保持假阳性。在此组套的可直接药物治疗的亚组中鉴定的真正的体细胞突变中，约18％未表达，这增加了不准确治疗决策和治疗无效的风险。在此诊断失败的背景下，显然需要改进对真正的肿瘤体细胞变体的鉴定。如下文更详细描述，此种改进的分析已经通过对肿瘤DNA、种系DNA和肿瘤RNA的协调分析完成。

[0085] 基于对仅针对肿瘤的基因组套分析的假阳性的担忧，本发明人试图证明通过以下方法提供的提高的精确度：同时对肿瘤和种系两者进行测序和分析，并改进了突变可被鉴定为疾病的潜在驱动者的置信度。如下文更详细地讨论，本发明人进行的研究证明：i)通过使用患者的正常组织作为对照进行的生物信息学分析，即肿瘤-正常样DNA测序，可以精确地多地进行出于治疗决策支持的目的的肿瘤的分子表征，并且在与RNA测序组合使用时，进一步提高了如此鉴定的真正体细胞变体的精确度，ii)来自仅针对肿瘤的序列分析的多态性的生物信息学过滤与肿瘤-正常样基因组分析的精确度不符，iii)任何真正体细胞突变在mRNA中表达的确认提供了如下的关键第二条证据，即检测的体细胞肿瘤突变可能起致癌驱动作用。

[0086] 在此实例中，使用CMS核准覆盖的35个基因组套的肿瘤和正常样种系基因组的DNA测序将源自使用仅针对肿瘤的测序方法的肿瘤体细胞变体假阳性率定量，这35个基因组套来自45名肺癌患者和621名具有33种癌症类型的所有癌症患者。还评估了这35个基因的变化的表达分析的精确度的潜在增加，该潜在增加通过RNA测序产生。

[0087] 患者和测序数据：在此实例中，本发明人集中于先前已由CMS核准进行医疗保险覆盖的35个基因的突变分析，以使临床医生能够更好地针对肺癌患者确定疗法。CMS仅在通过仅针对肿瘤的DNA测序和分析(即，不是匹配的肿瘤和正常样)鉴定基因组变体时才批准使用此基因组套。此方法不能直接区分体细胞变化与种系变化。该组套包括与体细胞肿瘤驱动子有关的25个基因(肿瘤驱动基因组套)和已知影响遗传癌症风险的10个基因(遗传风险基因组套)。肿瘤驱动基因组套由以下组成：ALK、BRAF、CDKN2A、CEBPA、DNMT3A、EGFR、ERBB2、EZH2、FLT3、IDH1、IDH2、JAK2、KIT、KMT2A、KRAS、MET、NOTCH1、NPM1、NRAS、PDGFRA、PDGFRB、PGR、PIK3CA、PTEN、RET。遗传癌症风险组套由以下组成：APC、BMPR1A、EPCAM、MLH1、MSH2、MSH6、PMS2、POLD1、POLE、STK11。

[0088] 分析了来自621名癌症患者的肿瘤DNA、肿瘤RNA和正常样DNA的全基因组测序数据，以鉴定潜在地促成癌症生长和扩增的体细胞来源的单核苷酸变体。本实例包括45名肺癌患者。所有患者对于使用此研究中描述的数据是知情同意的。从保存的组织提取DNA和RNA，并且在NantOmics临床实验室改进修正案(Clinical Laboratory Improvement Amendments;CLIA)和认证授权专业(Certified Authorization Profession;CAP)认证的测序实验室中使用Illumina平台测序。所用测试的性能特征包括以>95％的敏感性和>99％的特异性检测转录并表达为RNA的SNV。对正常种系和肿瘤基因组进行测序，分别读取大约

30×和60×的读取深度。每个肿瘤产生大约3亿个RNA测序读段。

[0089] 数据分析:用BWA比对DNA测序数据与GRCh37(www.ncbi.nlm.nih.gov/assembly/2758/),由samblaster进行重复标记,并且由GATK v2.3进行indel重新比对和碱基质量重新校准。通过bowtie比对RNA测序数据,并且通过RSEM估计RNA转录物表达。使用NantOmics Contraster分析流程进行了肿瘤与匹配的正常样的变体分析,以确定体细胞和种系SNV、插入和缺失,并鉴定肿瘤基因组的高度扩增区域。

[0090] 使用基级PhastCons保守得分、来自dbSNP的群体等位基因频率(Build142)以及其对从RefSeq数据库下载的基因转录物的预测影响(例如,DNA序列和蛋白质变化)注释小变体。

[0091] 肿瘤体细胞单核苷酸变体(SNV)的鉴别:对45名肺癌患者的肿瘤和正常样(种系)基因组进行全基因组DNA测序,从而在与肺癌的病因相关的35个相关基因的组套中鉴定了802个错义或无义的改变蛋白质的SNV。该组套包括被认为是体细胞肿瘤驱动子的25个基因(肿瘤驱动基因组套)和已知会影响遗传癌症风险的10个基因(遗传风险基因组套;表1)。在45名肺癌患者中,在147个独特的SNV位点总共存在802个SNV。所有802个变体都存在于肿瘤基因组中。肿瘤和正常样种系DNA序列的生物信息学分析显示,746个SNV中有701个(94%)起源于种系,并且剩余的45个SNV(6%)起源于体细胞组织。将同一基因组套施加于621名具有33种癌症类型的癌症患者的分析,肿瘤-正常样测序分析可鉴定10,704个错义或无义的改变蛋白质的SNV。有919个独特的SNV位点促成了鉴定的10,704个SNV。每名患者的肿瘤和正常样种系基因组分析确定10,149个(95%)SNV是种系起源的,而其余555个(5%)SNV是体细胞起源的。

[0092]

	患有所有癌症类型的患者的变体的数量			仅肺癌患者的变体的数量		
基因	独特	种系	体细胞	独特	种系	体细胞
肿瘤驱动基因组套						
ALK	32	1317 (99%)	14 (1%)	6	93 (99%)	1 (1%)
BRAF	23	5 (15%)	29 (85%)	3	0 (0%)	3 (100%)

	CDKN2A	22	35 (71%)	14 (29%)	5	2 (40%)	3 (60%)
	CEBPA	8	2 (25%)	6 (75%)	0	0	0
	DNMT3A	22	12 (52%)	11 (48%)	1	1 (100%)	0 (0%)
	EGFR	29	315 (95%)	16 (5%)	6	15 (71%)	6 (29%)
	ERBB2	38	921 (98%)	15 (2%)	7	68 (100%)	0 (0%)
	EZH2	12	117 (94%)	8 (6%)	1	3 (100%)	0 (0%)
	FLT3	25	846 (99%)	5 (1%)	6	64 (98%)	1 (2%)
	IDH1	9	85 (94%)	5 (6%)	2	2 (100%)	0 (0%)
	IDH2	10	9 (64%)	5 (36%)	0	0	0
	JAK2	18	37 (88%)	5 (12%)	0	0	0
	KIT	19	138 (93%)	10 (7%)	5	8 (62%)	5 (38%)
	KMT2A	57	72 (80%)	18 (20%)	3	2 (67%)	1 (33%)
	KRAS	16	3 (4%)	77 (96%)	4	0 (0%)	7 (100%)
	MET	28	58 (84%)	11 (16%)	5	7 (87%)	1 (13%)
	NOTCH1	59	143 (89%)	17 (11%)	8	6 (75%)	2 (25%)
[0093]	NPM1	2	1 (50%)	1 (50%)	0	0	0
	NRAS	10	1 (5%)	18 (95%)	0	0	0
	PDGFRA	24	169 (92%)	14 (8%)	2	9 (100%)	0 (0%)
	PDGFRB	28	98 (92%)	8 (8%)	8	11 (92%)	1 (8%)
	PGR	31	377 (96%)	15 (4%)	7	21 (91%)	2 (9%)
	PIK3CA	31	96 (54%)	82 (46%)	2	6 (86%)	1 (14%)
	PTEN	33	780 (97%)	24 (3%)	2	56 (100%)	0 (0%)
	RET	22	244 (96%)	9 (4%)	7	21 (100%)	0 (0%)
	总计	**608**	**5881**	**437**	**90**	**395**	**34**
	遗传风险基因组套						
	APC	85	692 (92%)	58 (8%)	7	48 (98%)	1 (2%)
	BMPR1A	5	334 (99%)	2 (1%)	1	17 (100%)	0 (0%)
	EPCAM	13	464 (100%)	0 (0%)	3	37 (100%)	0 (0%)
	MLH1	15	295 (99%)	4 (1%)	4	26 (96%)	1 (4%)
	MSH2	23	40 (89%)	5 (11%)	4	5 (100%)	0 (0%)
	MSH6	25	273 (98%)	7 (2%)	2	18 (100%)	0 (0%)
	PMS2	44	1558 (99%)	10 (1%)	13	110 (97%)	3 (3%)
	POLD1	30	208 (97%)	7 (3%)	4	11 (100%)	0 (0%)
[0094]	POLE	58	398 (96%)	18 (4%)	16	34 (92%)	3 (8%)
	STK11	13	6 (46%)	7 (54%)	3	0 (0%)	3 (100%)
	总计	**311**	**4268**	**118**	**57**	**306**	**11**

[0095] 表1

[0096] 对于肺癌患者，在肿瘤驱动基因组套和遗传风险基因组套中，分别只有7％和3％的SNV是体细胞起源的。在所有癌症患者中，在肿瘤驱动基因组套和遗传风险基因组套中，代表体细胞变化的SNV百分比对于肿瘤驱动基因组和遗传风险基因组中的基因分别为6％和3％。在已知具有体细胞癌症驱动子突变的25个基因中，预期会观察到更大比例的体细胞变体。每个基因中观察到的SNV数量存在显著变化。独特的SNV位点的数量与基因蛋白编码序列的大小密切相关(对于所有癌症类型，p值<10-9，R2＝0.70)。但是，种系、体细胞或总变体的数量与基因大小之间没有相关性(所有p值>0.40)。每个基因与癌症结果之间的关联程度可能决定了在基因之间观察到的SNV计数的变化以及每个基因中存在的自然群体遗传变异。此外，患者中富含特定的癌症驱动SNV。

[0097] 与总变体相比，少量独特变体说明了在癌症患者研究群体的许多基因组中观察到的常见SNV的存在。在621名癌症患者的样品中，有21个等位基因频率>0.02的变体，其中17个是常见的种系SNP，并且其中4个是常见的体细胞驱动子突变(KRAS中2个，PIK3CA中2个)。所有21个常见变体都存储在遗传多态性的单核苷酸多态性数据库(dbSNP)中。在所有患者中，仅观察到919个总独特变体中的645个(70％)一次。三个SNV均是种系和体细胞起源的。

[0098] 肺癌患者的仅肿瘤基因组测序(不与正常样种系基因组比较)将鉴定746个错义和无义的改变蛋白质的SNV(表1)。在肿瘤分子谱分析的情况下，被归类为体细胞起源的种系起源的任何SNV构成了假阳性结果。在不对推定的种系变体进行任何过滤的情况下，鉴于表1中呈现的数据，预期假阳性率大约为94％。图1显示了45名肺癌患者中将发生的假阳性结果的数量，并且图2描绘了在如下三种不同SNV过滤标准下，所有621名癌症患者的每个基因的相同结果：1)去除在dbSNP数据库中找到的所有SNV；2)去除报道的群体等位基因频率≥0.01(1％)的所有SNV；以及3)去除报道的群体等位基因频率≥0.001(0.1％)的所有SNV。(还去除了未报道群体等位基因频率，但在癌症患者中是常见种系SNV且存在于dbSNP中的另外三个SNV)。使用0.01的等位基因频率阈值产生最大数量的假阳性结果。通过将等位基因频率过滤阈值降至0.001，在大多数基因中假阳性的数量可以减少一半。大多数公共可用的群体等位基因频率估计值的精确度不超过0.0001，因此，群体等位基因频率阈值的进一步降低对假阳性SNV的数量具有标称作用。

[0099] 排除dbSNP数据库中存在的所有SNP使得假阳性SNV的数量达最低。然而，改进的假阳性率是以假阴性率增加为代价的，因为排除了许多真正的肿瘤体细胞SNV。排除存在于dbSNP中的所有SNV，使得在45名肺癌患者中观察到的45个真正肿瘤体细胞变体中有17个假阴性(38％)，并且在肺癌患者中的555个真正体细胞变体中有245个假阴性(44％)。使用0.001等位基因频率阈值过滤器，肺癌患者中有41个假阳性结果(观察到的746个SNV中的5％，并且过滤后剩余的86个SNV中的48％)和零个假阴性结果。相同的过滤阈值在所有621名癌症患者中产生554个假阳性结果(观察到的10,704个总SNV中的5％，并且过滤后剩余的1,107个SNV中的50％)和零个假阴性结果。

[0100] 仅针对肿瘤的测序方法的后果：过滤去除群体等位基因频率≥0.001的所有SNV之后，在35个基因的组套中，45名肺癌患者中有37名，并且在621名所有癌症患者中有472名具有至少一个错义或无义的改变蛋白质的SNV。过滤后没有SNV的7名肺癌患者和总共149名患者没有任何真正的体细胞变体，显示群体等位基因频率过滤器未产生假阴性结果。图3显示

了肺癌的真阳性的数量(即,肿瘤体细胞SNV的数量)和假阳性SNV的数量(即,遗传种系SNV的数量),并且图4显示了过滤后剩下至少一个SNV的所有患者的相同结果。肺癌和所有癌症患者的平均SNV数量分别为1.91个和1.84个。出于呈现的目的,从图2b排除了一名具有39个体细胞SNV的患者。在肺癌患者中,45名患者中有29名(65%)具有至少一个假阳性SNV,并且15名患者仅具有假阳性SNV(33%)而没有任何真正的阳性结果。尽管在以0.001的群体等位基因频率过滤后,在肺癌患者中发现的总SNV中只有5%为假阳性(发现的802个总SNV中有41个假阳性),但这些SNV分布在65%的患者中。发现的802个SNV中的大多数是常见变体,已通过过滤排除。这些结果突出了稀有种系突变对假阳性发现率的影响。在整个研究群体中,621名患者中有365名(59%)具有至少一个假阳性SNV,从而每名患者平均产生0.91个假阳性。在621名患者中有193名(31%)仅存在假阳性SNV,而没有真正的阳性结果。

[0101] 假阳性SNV可能会对患者的护理产生直接的不利影响。表2显示了12个可药物治疗的基因,在体细胞突变后针对每种基因的特定药物以及在每种基因中观察到至少1个假阳性SNV的患者人数。此外,显示了与每种药物相关的成本和可能的不良健康影响,以说明基于假阳性结果开药的财务和临床影响。仅针对肿瘤的序列分析可使患者处于不必要的严重药物不良反应的风险以及开具可能无效的药物治疗的负面影响中。

[0102]

药物	药物靶向基因	每个SNV过滤器之后具有至少一个假阳性变体的患者的数量						每名患者的近似药物成本[a]	警告和注意事项(FDA标签)
		没有过滤器		AF >= 0.01		AF >= 0.001			
		所有	LC	所有	LC	所有	LC		
克唑替尼(Crizotinib)	ALK	621	45	50	2	16	0	$ 18,349.50	肺炎,肝异常,QT延长
艾乐替尼(Alectinib)								$ 15,976.33	肝毒性,ILD/肺炎,心动过缓,肌痛,CPK升高,EFT
色瑞替尼(Ceritinib)								$ 18,964.13	GI毒性,肝毒性,ILD/肺炎,QT延长,高血糖,心动过缓,胰腺炎,EFT
布加替尼(Brigatinib)								$ 15,960.00	ILD/肺炎,HTN,心动过缓,视力障碍,CPK升高,胰酶

[0103]

									升高，高血糖，EFT
维拉非尼（Vemurafenib）	BRAF	5	0	5	0	2	0	$ 13,020.94	过敏，皮肤反应，QT 延长，肝毒性，眼科反应，肾衰竭，EFT
达拉非尼（Dabrafenib）								$ 11,412.43	发热性药物反应，高血糖，葡萄膜炎和虹膜炎，G6PD 缺乏，EFT
柯美替尼（Cobimetinib）								$ 7,856.04[a]	出血，心肌病，皮肤病学反应，视网膜病变和 RVO，肝毒性，横纹肌溶解，光过敏，EFT
曲美替尼（Trametinib）								$ 12,450.00	心肌病，RPED，RVO，ILD，皮肤毒性，EFT
阿扎胞苷（Azacitidine）	DNMT3A	12	1	12	1	11	1	$ 2,221.81[c]	血球减少，肝毒性，肾异常，EFT
地西他滨（Decitabine）								$ 3,967.37[c]	血球减少，EFT
厄洛替尼（Erlotinib）	EGFR	303	15	16	0	14	0	$ 9,390.44	ILD，肾衰竭，肝毒性，GI 穿孔，大疱和皮肤障碍，CVA，MAHA，眼障碍，EFT
阿法替尼（Afatinib）								$ 9,060.85	腹泻，大疱和皮肤障碍，ILD，肝毒性，角膜

[0104]

									炎，EFT
吉非替尼（Gefitinib）								$ 9,117.36	腹泻，大疱和皮肤障碍，ILD，肝毒性，角膜炎，EFT，GI 穿孔
奈拉替尼（Neratinib）	ERBB2	544	37	43	5	24	2	$ 12,600.00	腹泻，肝毒性，EFT
拉皮替尼（Lapitinib）								$ 6,314.31	LVEF 降低，肝毒性，腹泻，ILD 和肺炎，QT 间隔延长，EFT
鲁索替尼（Ruxolitinib）	JAK2	37	0	23	0	19	0	$ 12,932.64	血球减少，感染
伊马替尼（Imatinib）	KIT	135	8	13	1	11	0	$ 23,152.39	水肿，血球减少，CHF 和 LV 功能障碍，肝毒性，出血，GI 穿孔，心源性休克，大疱，甲状腺功能减退，EFT
达沙替尼（Dasatinib）								$ 16,084.02	骨髓抑制，血小板减少，液体潴留，QT 延长，CHF，LV 功能障碍，MI，EFT
雷戈非尼（Regorafenib）								$ 17,857.80[d]	出血，皮肤病学毒性，HTN，心脏缺血和梗塞，RPLS，GI 穿孔，伤口愈合并发症，EFT

[0105]

克唑替尼（Crizotinib）	MET	58	7	41	5	20	2	$ 18,349.50	肺炎，肝实验室异常，QT 间隔延长，EFT
卡波替尼（Cabozantinib）								$ 18,191.26	出血，GI 穿孔，血栓形成事件，HTN，腹泻，PPES，RPLS，EFT
阿昔替尼（Axitinib）	PDGFRA	160	9	36	0	13	0	$ 16,416.28	出血，GI 穿孔，血栓形成事件，HTN，甲状腺功能减退，RPLS，EFT
雷戈非尼（Regorafenib）								$ 17,857.80[d]	出血，皮肤病学毒性，HTN，心脏缺血和梗塞，RPLS，GI 穿孔，伤口愈合并发症，EFT
阿昔替尼（Axitinib）	PDGFRB	89	9	42	4	18	3	$ 16,416.28	出血，GI 穿孔，血栓形成事件，HTN，甲状腺功能减退，RPLS，EFT
雷戈非尼（Regorafenib）								$ 17,857.80[d]	出血，皮肤病学毒性，HTN，心脏缺血和梗塞，RPLS，GI 穿孔，伤口愈合并发症，EFT
艾代拉里斯（Idelalisib）	PIK3CA	96	6	0	0	0	0	$ 5,721.26[e]	皮肤反应，过敏反应，中性粒细胞减少症，

[0106]

									EFT
依维莫司（Everolimus）								$ 17,013.54	肺炎，感染，口腔溃疡，EFT
卡波替尼（Cabozantinib）	RET	217	18	22	5	19	5	$ 18,191.26	出血，GI穿孔，血栓形成事件，HTN，腹泻，PRES，RPLS，EFT
万地替尼（Vandetinib）								$ 15,445.43	QT延长，皮肤反应，ILD，缺血性脑血管事件，出血，腹泻，HTN，RPLS，EFT
具有FP SNV的独特患者的总数	621 (100%)	45 (100 %)	303 (49 %)	23 (51 %)	167 (27 %)	13 (29 %)			

[0107] 表2

[0108] AF＝群体等位基因频率;所有＝患有所有30种癌症的患者;LC＝仅肺癌患者;ILD＝间质性肺疾病;EFT＝胚胎毒性;RVO＝视网膜静脉阻塞;RPED＝视网膜色素上皮营养不良;CVA＝脑血管意外;MAHA＝微血管病性溶血性贫血;GI＝胃肠道;LVEF＝左心室射血分率;MI＝心肌梗塞;RPLS＝可逆性后脑白质脑病综合征;PRES＝后可逆性脑病综合征;

[0109] HTN＝高血压(包括高血压危机);

[0110] [a]除非另外说明,否则为30天的平均批发价格。

[0111] [b]药物不连续给予。

[0112] [c]基于2.02的身体表面积的单个循环。

[0113] [d]基于治疗21天和休息7天。

[0114] [e]基于治疗14天和休息14天。

[0115] 体细胞单核苷酸变体的表达:可从26名肺癌患者和所有患者中的378名获得RNA测序数据,这些RNA测序数据可评估肿瘤体细胞SNV的表达。表3显示了评估的体细胞SNV的总数、未表达的体细胞SNV的数量以及具有未表达的体细胞SNV的患者的数量。大百分比的SNV没有表达:对于肺癌患者,为18％(39个SNV中的7个),对于所有癌症患者,为15％(517个SNV中的75个)。基因之间表达的肿瘤体细胞变体的百分比存在很大变化。在所有癌症患者中,FLT3、PDGFRA、PGR和RET中约80％或更多的SNV未表达。在该研究群体中,肺癌患者中的9％(具有肿瘤RNA测序数据的所有26名患者中的6名)和所有癌症患者中的13％(具有肿瘤RNA测序数据的378名所有癌症患者中的51名)具有至少一个在信使RNA中未表达的真正的肿瘤体细胞SNV。4名肺癌患者中有4个肿瘤体细胞SNV未在作为表2中显示的特定药物的靶标的十二个基因中表达。所有癌症患者中有33名的肿瘤体细胞SNV未在RNA中表达。因此,仅基于DNA分析的治疗决策可能导致无效疗法的给予。

[0116]

	所有癌症类型			仅肺癌		
基因	体细胞SNV	未表达的体细胞SNV（%）	具有未表达的SNV的患者	体细胞SNV	未表达的体细胞SNV（%）	具有未表达的SNV的患者
ALK	13	10 (76%)	9	0	0	0
BRAF	24	0 (0%)	0	2	0 (0%)	0
CDKN2A	13	2 (15%)	2	3	0 (0%)	0
CEBPA	5	1 (20%)	1	0	0	0
DNMT3A	11	1 (9%)	1	0	0	0
EGFR	16	1 (6%)	1	6	0 (0%)	0
ERBB2	14	1 (7%)	1	0	0	0
EZH2	8	0 (0%)	0	0	0	0
FLT3	5	4 (80%)	4	1	1 (100%)	1
IDH1	5	0 (0%)	0	0	0	0
IDH2	5	0 (0%)	0	0	0	0
JAK2	5	1 (20%)	1	0	0	0
KIT	8	5 (63%)	5	4	2 (50%)	2
KMT2A	18	2 (11%)	2	1	0 (0%)	0
KRAS	70	2 (3%)	2	6	1 (17%)	1
MET	11	3 (27%)	3	1	1 (100%)	1
NOTCH1	16	1 (6%)	1	2	0 (0%)	0

[0117]

NPM1	1	0 (0%)	0	0	0	0
NRAS	15	0 (0%)	0	0	0	0
PDGFRA	14	11 (79%)	8	0	0	0
PDGFRB	8	3 (38%)	3	1	1 (100%)	1
PGR	14	13 (93%)	11	1	1 (100%)	1
PIK3CA	75	0 (0%)	0	1	0 (0%)	0
PTEN	23	1 (4%)	1	0	0	0
RET	9	7 (78%)	6	0	0	0
APC	54	4 (7%)	4	1	0 (0%)	0
BMPR1A	1	0 (0%)	0	0	0	0
EPCAM	0	0	0	0	0	0
MLH1	4	0 (0%)	0	1	0 (0%)	0
MSH2	5	0 (0%)	0	0	0	0
MSH6	7	1 (14%)	1	0	0	0
PMS2	10	0 (0%)	0	3	0 (0%)	0
POLD1	7	0 (0%)	0	0	0	0
POLE	16	1 (6%)	1	2	0 (0%)	0
STK11	7	0 (0%)	0	3	0 (0%)	0
总计	517	75 (15%)	51名独特患者	39	7 (18%)	6名独特患者

[0118] 表3

[0119] 当前，有两种基于测序的方法可用于鉴定患者的肿瘤体细胞变异。在第一种方法中，对代表靶向的基因组套、外显子组或整个基因组的肿瘤DNA进行测序，并基于参考基因组和在肿瘤中发现的单个基因组变体的特征过滤推定的种系变异（称为仅针对肿瘤的分析）。在群体遗传数据库中以可估计的等位基因频率鉴定基因组变异是用于判定变体是否

为遗传种系起源的常用过滤标准。如本文所示的第二种且更精确的方法是使用患者自己的种系基因组作为精确对照(而不是用于过滤的参考基因组)来区分遗传种系变体与体细胞来源的变体(称为肿瘤-正常样分析)。当前CMS批准的用于为肺癌治疗提供信息的测试基于前一种方法,并且特别排除了在确定体细胞变体中使用正常组织(种系信息)。

[0120] 对比该两种方法,本发明人分析了来自45名肺癌患者和621名所有癌症患者的肿瘤和正常样DNA测序数据与CMS核准覆盖的仅针对肿瘤的基因组套。该研究证明,在使用仅针对肿瘤的测序来鉴定体细胞变体时,假阳性率为94%(对于所有癌症,为95%)。即使在利用多种方法从推定的体细胞突变进行多态性的生物信息学过滤后,假阳性率仍在38%-94%范围内。根据使用的方法,过严格的过滤会导致潜在的假阴性。当集中于FDA批准的药物所靶向的12个基因的子集时,其中体细胞突变的鉴定可以为治疗决策提供信息,根据所用的多态性过滤方法,受假阳性识别影响的肺癌患者的百分比在29%-51%范围内。假阳性结果的其他风险源于鉴定从体细胞组织鉴定的变体,即在诸如BRCA1、BRCA2和ATM的基因中将真正的体细胞突变错误鉴定为是有害的(遗传)种系变体。在与家族性疾病的种系风险相关的10个基因(遗传风险基因组套)中,当使用仅针对肿瘤的测序方法时,在10名肺癌患者(11个变体)和101名总患者(118个变体)中发现了种系基因的真正体细胞突变。

[0121] 对来自患者正常样种系基因组和肿瘤基因组的数据进行的测序和分析消除了与仅对肿瘤基因组序列数据进行分析相关的假阳性结果。肿瘤体细胞SNV有效地为患者治疗提供信息的可能性取决于DNA变体表达为信使RNA,然后翻译成蛋白质。肿瘤的RNA测序提供了关于癌症驱动基因的相对表达水平和特定肿瘤体细胞变体的基因表达的有价值的信息。此研究中的RNA表达分析显示,从肺癌患者的肿瘤/正常样测序鉴定的真正体细胞突变中的18%以及所有癌症患者中的15%在信使RNA水平上未表达。在该研究群体中,这些结果可能影响针对肺癌患者中的9%和所有癌症患者中的13%作出的临床决策。本文提供的结果进一步证明了与从肿瘤/正常样DNA测序加RNA测序产生的针对药物靶向的分子分析的精确度提高相关的优势。

[0122] 鉴于以上所述,因此应当理解,同时对正常样种系基因组和肿瘤基因组两者的DNA进行测序和生物信息学分析对于准确鉴定用于癌症疗法的分子靶标是必需的。仅分析肿瘤基因组会导致SNV鉴定的假阳性率很高。同时进行肿瘤-正常样DNA和RNA测序分析可获得甚至更高的精确度。基于仅针对肿瘤的DNA分析或在不存在RNA分析下进行的治疗决策可能导致给予无效疗法,同时也增加了药物相关不良副作用的风险。当用于指导临床决策时,仅针对肿瘤的基因组套分析的方法可能会增加患者的风险,引起潜在的长期不良健康后果,并增加医疗成本。

[0123] 实例2

[0124] 在本实例中,本发明人包括204名患有11种胃肠道(GI)癌症类型的癌症患者,并且对肿瘤和正常样基因组进行了全基因组测序。在如下所示的45个基因组套中对错义和无义单核苷酸变体(SNV)的真阳性(真正体细胞变体)和假阳性(估计为体细胞变体的真正种系变体)进行了测量。该45个基因的组套包括26个已知的体细胞驱动基因、14个遗传癌症风险基因,并且这些基因中的5个既可以充当体细胞肿瘤驱动子又可以充当遗传风险基因。RNA测序可用于204名患者中的139名。使用公认且公开的生物信息学方法进行序列比对和SNV变体识别。在优选的方法中,使用BAMBAM来使用DNA和RNA序列同步且增量地比对并鉴定

SNV。

[0125] 结果:从仅肿瘤基因组测序鉴定的SNV中有92%是种系起源的,并且具有潜在的假阳性而不是真正的体细胞变体(体细胞=真正的体细胞变体;种系=真正的种系变体)。参见图5A和图5B。值得注意的是,使用报道的群体等位基因频率≥0.001的公共数据库过滤所有SNV仍然导致41%的假阳性率(体细胞=真正的体细胞变体;种系=真正的种系变体)。参见图6A和图6B。如图7所示,GI患者中的71%在等位基因频率过滤后具有至少一个假阳性SNV(种系)(体细胞=真正的体细胞变体;种系=真正的种系变体)。此外,RNA分析显示真实体细胞变体中的10%未表达,并且患者中的17%具有至少一个未表达的真实体细胞变体,如图8所示。

[0126] 因此,应当理解,对肿瘤基因组进行测序鉴定了遗传种系起源和肿瘤体细胞起源的所有SNV,其中大部分是种系起源的。虽然群体等位基因频率和其他参数可用于过滤SNV数据并估计体细胞与种系的起源,但这种过滤对于临床使用而言不够准确。此外,应当理解,同时对正常样种系基因组和肿瘤基因组两者的DNA进行测序和生物信息学分析对于准确鉴定分子靶标是必需的。单独分析肿瘤基因组会导致假阳性结果。同时进行肿瘤-正常样DNA和肿瘤RNA测序分析可获得更高的精确度。基于仅针对肿瘤的DNA分析或在不存在RNA下进行的治疗决策可能导致给予无效疗法,同时也增加了药物相关不良副作用的风险。

[0127] 实例3

[0128] 在本实例中,本发明人旨在比较用50个基因的常用热点组套和仅分析肿瘤组织与同时用正常样种系DNA和肿瘤RNA分析肿瘤DNA进行的肿瘤体细胞识别的准确性和精确性。具体地,在本实例中,获得了来自1879名具有42种癌症类型的癌症患者的肿瘤样品和匹配的正常样品,并且产生了这些组织的全基因组测序数据或全外显子组测序数据。下表4中显示了群组的人口统计概况,并且图9中显示了按不同癌症类型测序的分析物数量(DNA和/或RNA测序的样品数量)。表4中N<10的癌症(或图9中的其他癌症类型)包括皮肤癌(非黑色素瘤)、间皮瘤、睾丸癌、胆管癌(肝外)、肛门癌、法特氏壶腹(ampulla of vater)癌、白血病、阴道癌、骨髓瘤、小肠癌、外阴癌、阴茎癌、尿道癌。

[0129]

癌症类型	患者数量	男性数量	女性数量	最小年龄	最大年龄	中位年龄
乳腺癌	336	2	327	20	86	56
结肠	180	83	93	17	87	58
肺癌	149	67	78	9	90	65
骨和软组织癌（包括肉瘤）	139	72	62	0	82	49
胰腺癌	123	69	48	3	87	63
卵巢癌	103	0	96	25	86	58
脑癌	93	52	37	0	79	49
未知癌症类型	75	38	29	6	91	59
其他癌症	71	39	31	1	83	62
N < 10 的癌症*	52	29	20	0	87	65.5
前列腺癌	51	48	0	40	83	65
胃癌	45	26	19	15	85	61
头颈癌	41	31	8	19	86	64
肾癌	38	25	11	0	72	62
肝癌	37	25	11	9	77	63
黑色素瘤	37	24	12	29	87	64
口腔和喉癌（包括甲状腺癌）	35	21	13	42	83	63
食管癌	35	24	10	46	86	64
直肠癌	31	21	10	28	80	57
膀胱癌	30	17	12	49	92	72
未知原发性癌症	29	11	18	29	83	57
子宫癌（子宫内膜癌）	29	0	28	34	89	66
软组织癌	22	15	7	2	80	18
胆囊癌	20	7	13	39	87	65.5
胸腺癌	17	9	8	24	73	59

[0130]

宫颈癌	16	0	16	27	75	49
肾上腺癌	13	8	4	1	74	48
淋巴瘤	12	8	3	18	81	66
肾盂癌和输尿管癌	10	5	5	8	71	42
胆道癌（肝内）	10	5	4	46	78	61

[0131] 表4

[0132] 从肿瘤组织的基因组测序数据，本发明人确定所有患者具有至少一个种系单核苷酸变体(总共30955个单核苷酸变体)。然后，本发明人定量了从比较肿瘤和匹配的正常样的基因组测序数据鉴定的所有单核苷酸变体(包括种系起源的单核苷酸变体和肿瘤体细胞起源的单核苷酸变体)的数量。1879名患者中有1127名(65％)具有至少1个体细胞单核苷酸变体(总共308721个)。分析物进行了配对DNA/RNA分析的1135名患者中有741名(65％)具有至少1个体细胞单核苷酸变体(总共198844个)，从而在配对DNA/RNA分析的患者中产生了1775个独特的单核苷酸变体。如图10所示，从仅对肿瘤基因组测序鉴定的单核苷酸变体中的92％是种系起源的，指示从仅对肿瘤基因组测序鉴定的大多数单核苷酸变体可能是假阳性，而不是真正的体细胞变体。

[0133] 本发明人进一步使用群体等位基因频率和其他参数(例如，已知种系变体、gnomAD)过滤了来自仅对肿瘤基因组测序的鉴定的单核苷酸变体，以确定单核苷酸变体的比率(种系起源与肿瘤体细胞起源)。如图11所示，使用报道的等位基因频率≥0.001的gnomAD过滤从仅对肿瘤基因组测序鉴定的所有单核苷酸变体。本发明人发现过滤后的假阳

性率降低到34％。然而，本发明人设想此种假阳性率对于此类数据的任何临床使用还不够准确。

[0134] 此外，本发明人发现并非肿瘤体细胞来源的所有单核苷酸变体都在RNA中表达，指示必须使用RNA表达分析进一步过滤以在所有鉴定的单核苷酸变体中获得真正的体细胞单核苷酸变体。如图12和图13所示，错义/无义体细胞单核苷酸变体中的15％(如图12所示)和所有体细胞单核苷酸变体(错义/无义/同义)中的17％未表达。另外，本发明人发现本实例中23％的癌症患者具有至少一个不表达的体细胞单核苷酸变体(无义/错义)。从此类数据，本发明人设想，同时对DNA，即正常样种系基因组和肿瘤基因组两者进行测序和生物信息学分析对于准确鉴定分子靶标是必需的，因为仅分析肿瘤基因组会导致高假阳性体细胞变体，并且因为在使用鉴定的单核苷酸变体或具有单核苷酸变体的基因作为分子靶标时，RNA表达的缺乏可能对临床的贡献不足。从不同的角度来看，通过同时对DNA，即正常样种系基因组和肿瘤基因组两者进行测序和生物信息学分析，可以实现基因中肿瘤治疗和/或药物靶标的更高精确度鉴定和/或改进的肿瘤状态测试算法。

[0135] 如在此的说明书和贯穿随后的整个权利要求书中所使用，“一个”、“一种”以及“该”的含义包括复数参照物，除非上下文清楚地另外指明。而且，如在此的说明书中所使用，“在……中”的含义包括“在……中”和“在……上”，除非上下文清楚地另外指明。除非上下文相反地指示，否则本文阐述的所有范围应解释为包括其端点，并且开放式范围应被解释为包括商业实用值。类似地，除非上下文指出相反的情况，否则应将所有值的列表视为包含中间值。

[0136] 此外，除非本文另外指示或另外与上下文明显矛盾，否则本文所述的所有方法均能够以任何合适的顺序进行。关于本文某些实施例提供的任何和所有实例或示例性语言(例如，“诸如”)的使用仅旨在更好地说明本发明，而不对另外要求保护的本发明范围进行限制。说明书中的语言不应当被解释为指示任何未要求保护的要素为实践本发明所必需的。

[0137] 本文披露的本发明的替代要素或实施例的组不应解释为限制。每个组成员可以单独或以与组中其他成员或本文发现的其他要素的任何组合被提及或要求保护。出于方便和/或专利性的原因，组中的一个或多个成员可以包括在组中或从组中删除。在本文中，当发生任何此种包括或删除时，认为本说明书含有经修改从而满足所附权利要求书中使用的所有马库什基团(Markush group)的书面描述的基团。

[0138] 对于本领域技术人员应当清楚的是，在不脱离本文的发明构思的情况下，除了已经描述的那些之外的更多修改是可能的。因此，除了在所附权利要求的范围中之外，本发明主题不受限制。此外，在解释说明书和权利要求书时，所有术语应当以与上下文一致的尽可能广泛的方式解释。特别地，术语“包含(comprises)”和“包含(comprising)”应当被解释为以非排他性方式指代元素、组分或步骤，从而指示所提及的元素、组分或步骤可以与未明确提及的其他元素、组分或步骤一起存在、或使用、或组合。如在此的说明书和贯穿随后的整个权利要求书中所使用，“一个”、“一种”以及“该”的含义包括复数参照物，除非上下文清楚地另外指明。而且，如在此的说明书中所使用，“在……中”的含义包括“在……中”和“在……上”，除非上下文清楚地另外指明。当本说明书的权利要求书提及选自由A、B、C……和N组成的组的某物中的至少一种时，该文本应解释为仅需要组中的一种要素，而不是A加N

或B加N等。

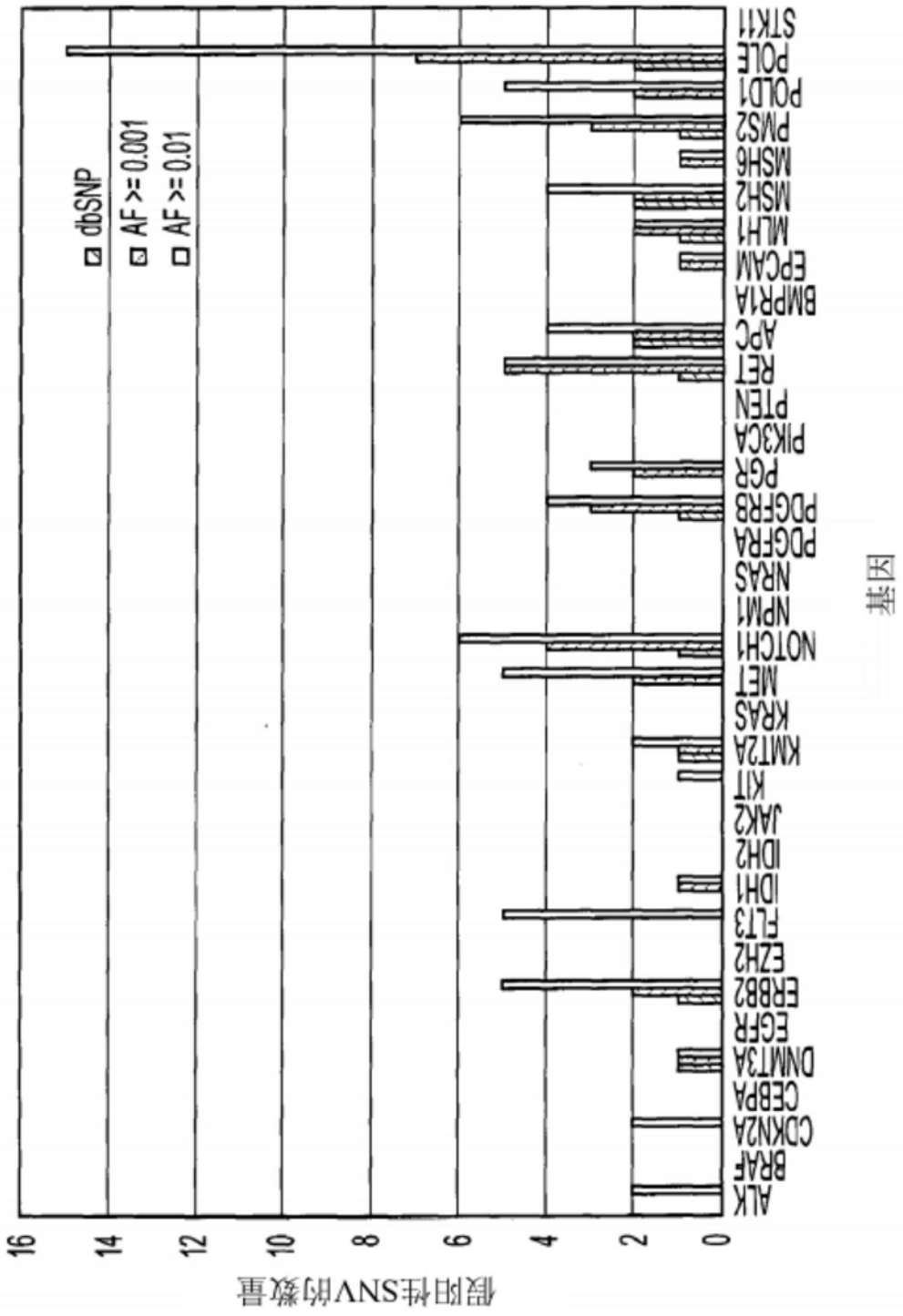

dbSNP
AF >= 0.001
AF >= 0.01
假阳性SNV的数量
0
2
4
6
8
10
12
14
16
基因
ALK
BRAF
CDKN2A
CEBPA
DNMT3A
EGFR
ERBB2
EZH2
FLT3
IDH1
IDH2
JAK2
KIT
KMT2A
KRAS
MET
NOTCH1
NPM1
NRAS
PDGFRA
PDGFRB
PGR
PIK3CA
PTEN
RET
APC
BMPR1A
EPCAM
MLH1
MSH2
MSH6
PMS2
POLD1
POLE
STK11

图1

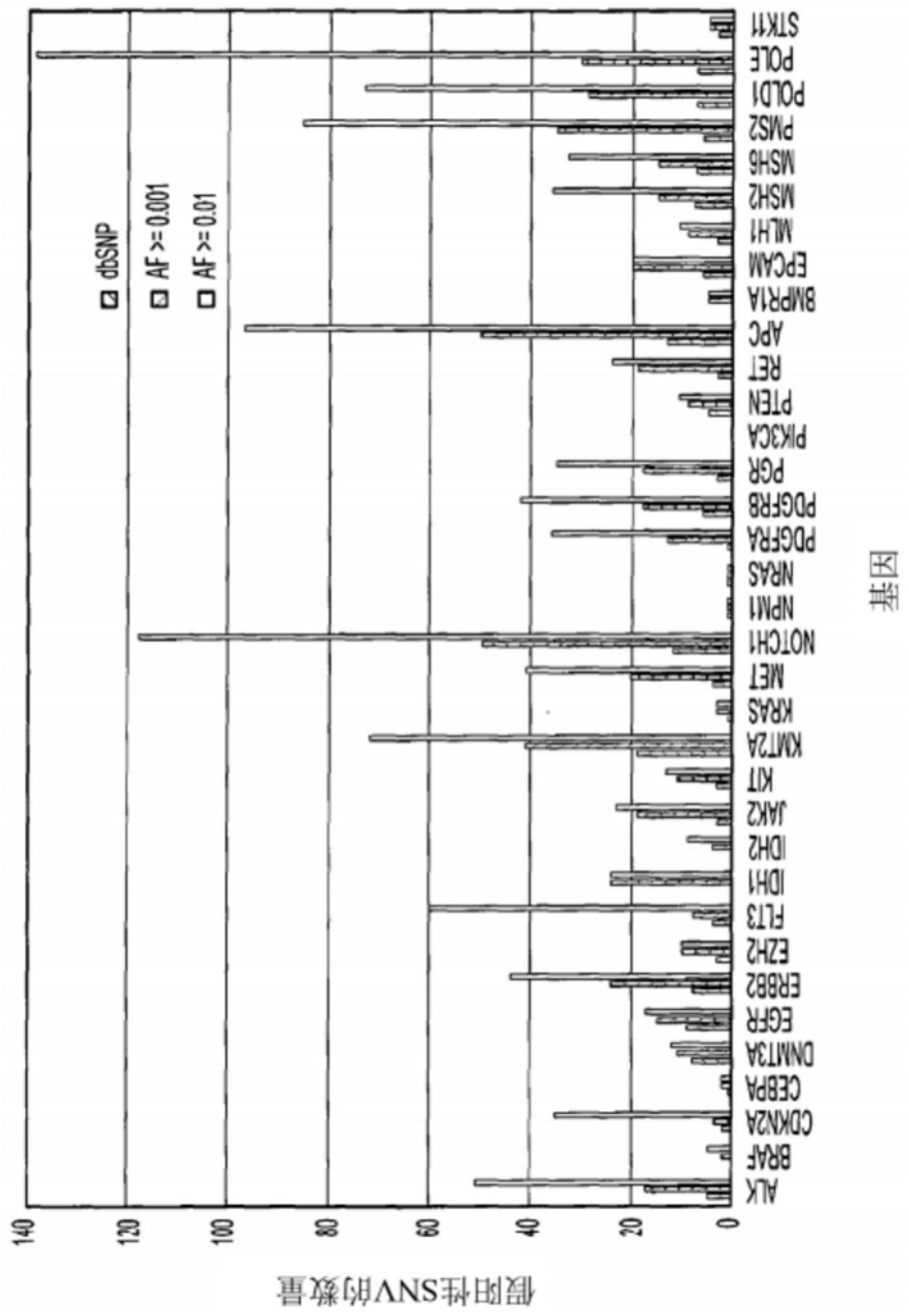

dbSNP
AF >= 0.001
AF >= 0.01
假阳性SNV的数量
基因
0
20
40
60
80
100
120
140
ALK
BRAF
CDKN2A
CEBPA
DNMT3A
EGFR
ERBB2
EZH2
FLT3
IDH1
IDH2
JAK2
KIT
KMT2A
KRAS
MET
NOTCH1
NPM1
NRAS
PDGFRA
PDGFRB
PGR
PIK3CA
PTEN
RET
APC
BMPR1A
EPCAM
MLH1
MSH2
MSH6
PMS2
POLD1
POLE
STK11

图2

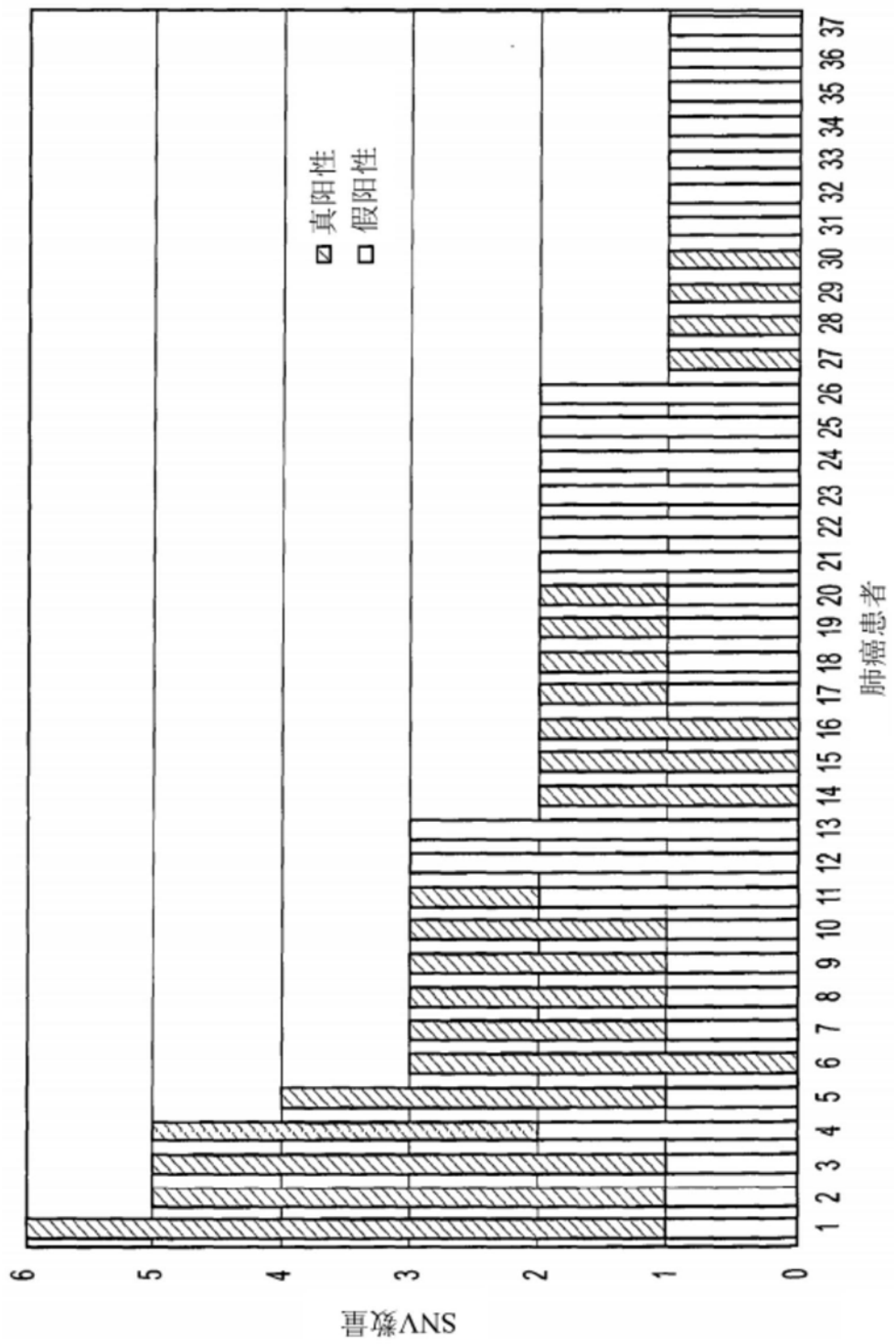

真阳性
假阳性
肺癌患者
SNV数量
0
1
2
3
4
5
6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37

图3

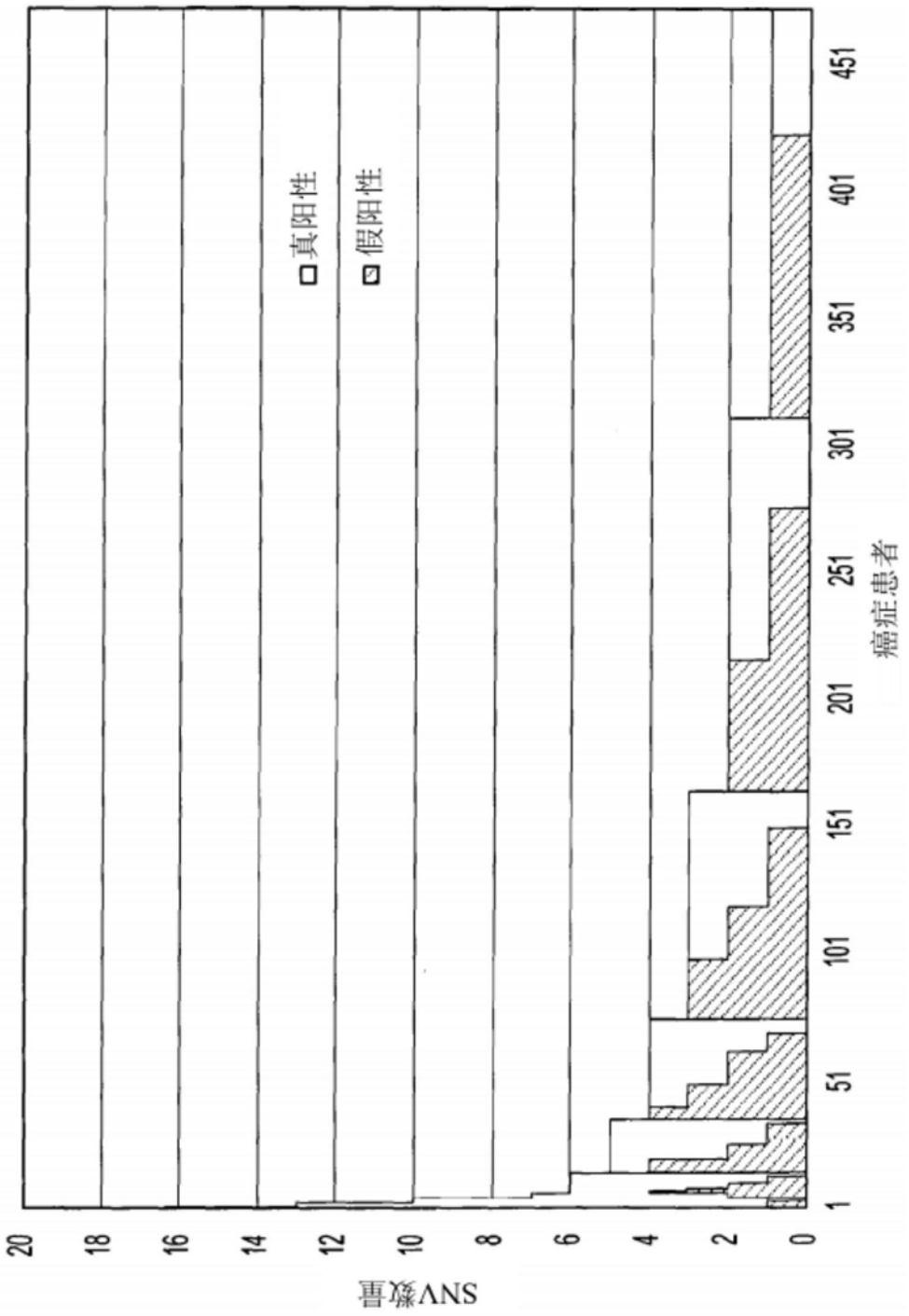

真阳性
假阳性
SNV数量
癌症患者
0
2
4
6
8
10
12
14
16
18
20
1
51
101
151
201
251
301
351
401
451

图4

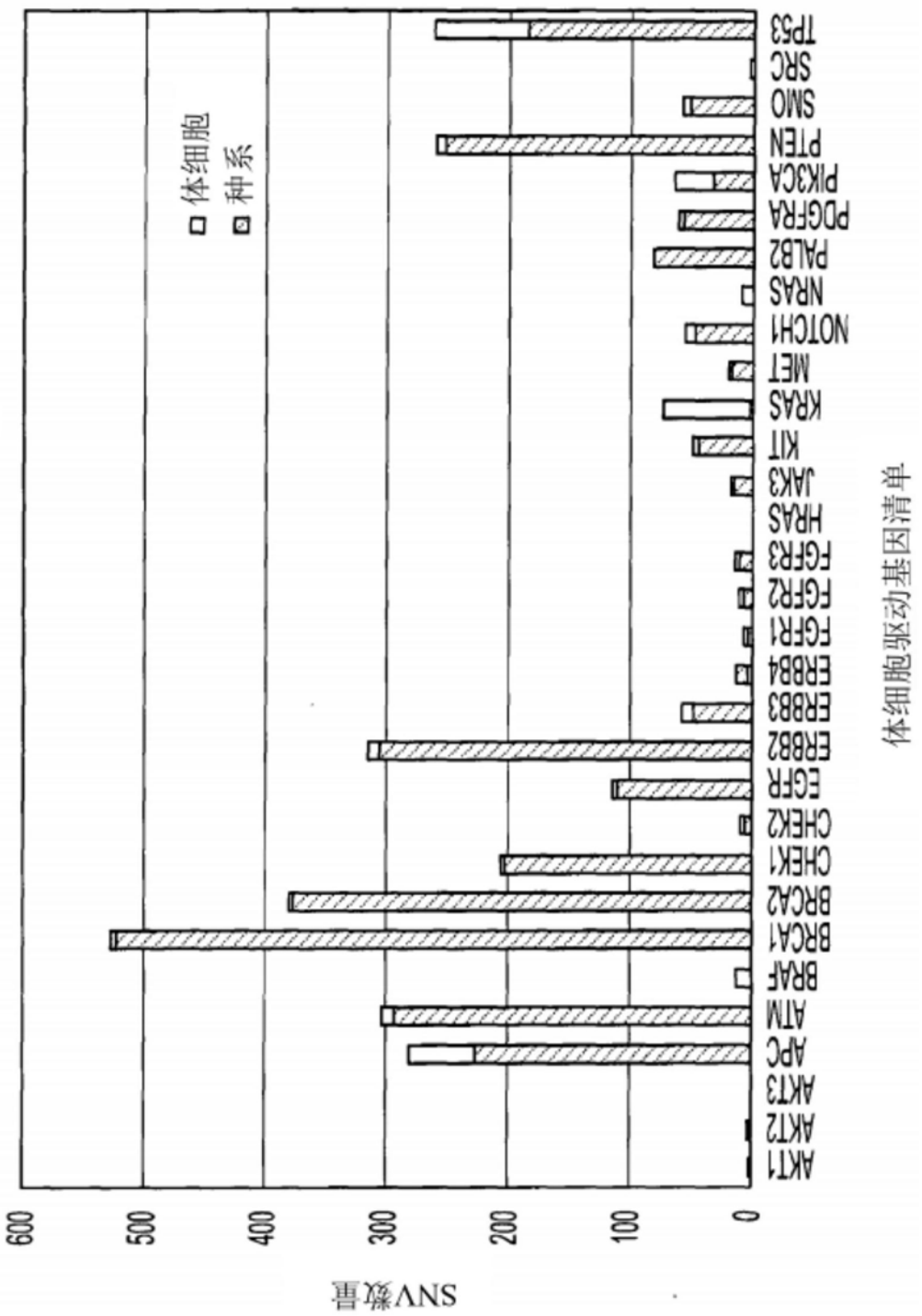

体细胞
种系
SNV数量
600
500
400
300
200
100
0
AKT1
AKT2
AKT3
APC
ATM
BRAF
BRCA1
BRCA2
CHEK1
CHEK2
EGFR
ERBB2
ERBB3
ERBB4
FGFR1
FGFR2
FGFR3
HRAS
JAK3
KIT
KRAS
MET
NOTCH1
NRAS
PALB2
PDGFRA
PIK3CA
PTEN
SMO
SRC
TP53
体细胞驱动基因清单

图5A

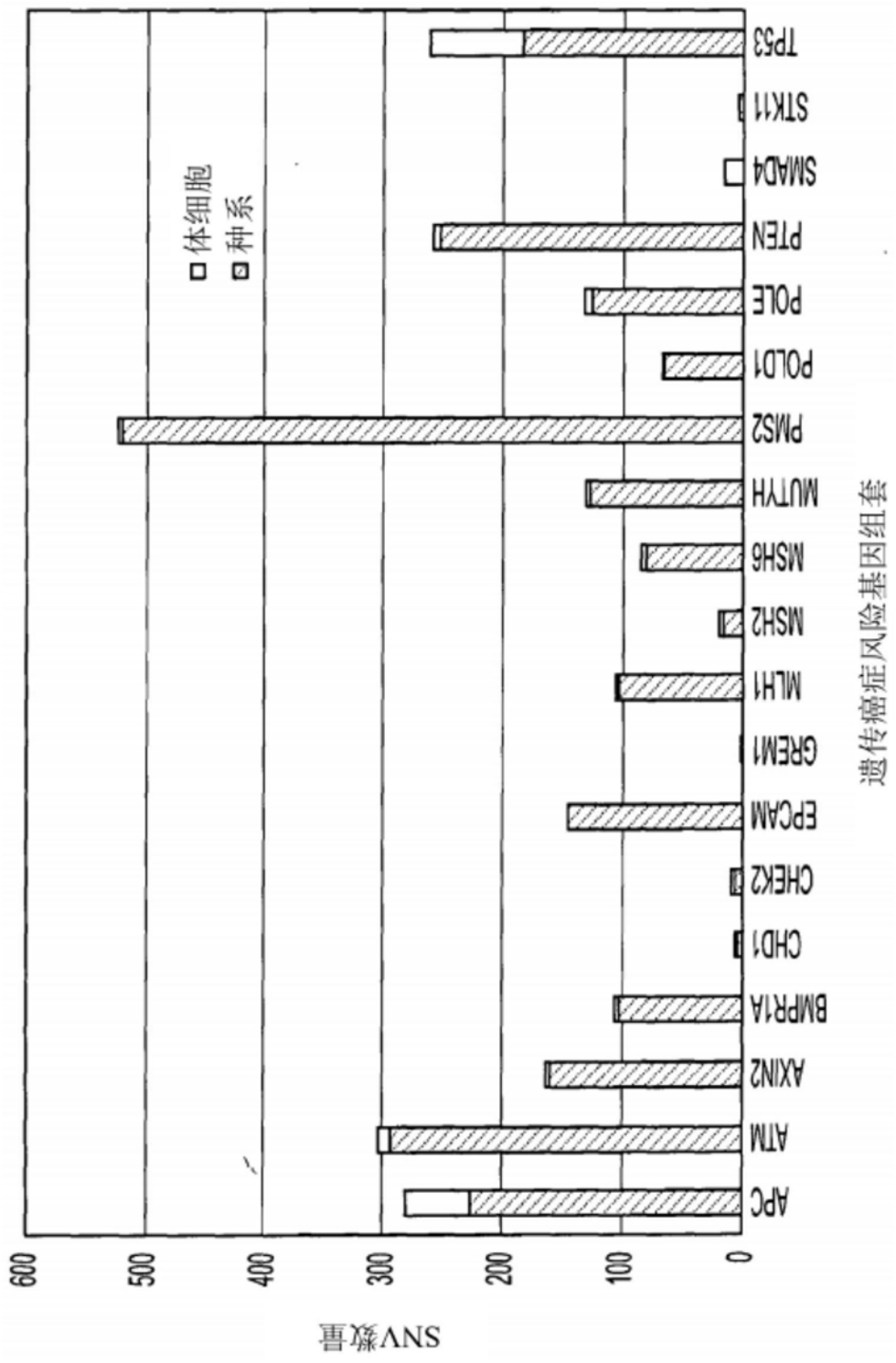

体细胞
种系
TP53
STK11
SMAD4
PTEN
POLE
POLD1
PMS2
MUTYH
MSH6
MSH2
MLH1
GREM1
EPCAM
CHEK2
CHD1
BMPR1A
AXIN2
ATM
APC
600
500
400
300
200
100
0
SNV数量
遗传癌症风险基因组套

图5B

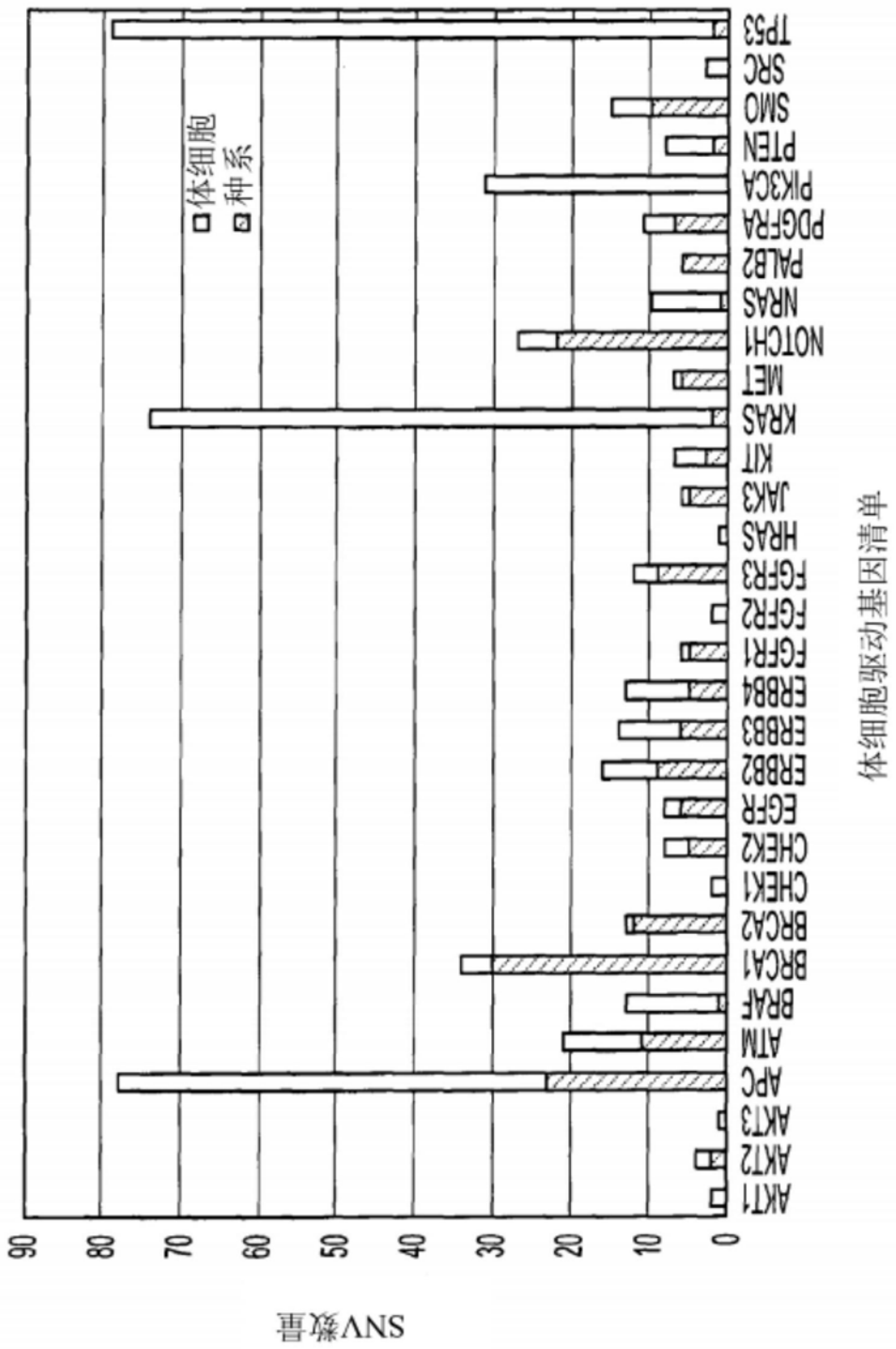

体细胞
种系
TP53
SRC
SMO
PTEN
PIK3CA
PDGFRA
PALB2
NRAS
NOTCH1
MET
KRAS
KIT
JAK3
HRAS
FGFR3
FGFR2
FGFR1
ERBB4
ERBB3
ERBB2
EGFR
CHEK2
CHEK1
BRCA2
BRCA1
BRAF
ATM
APC
AKT3
AKT2
AKT1
体细胞驱动基因清单
90
80
70
60
50
40
30
20
10
0
SNV数量

图6A

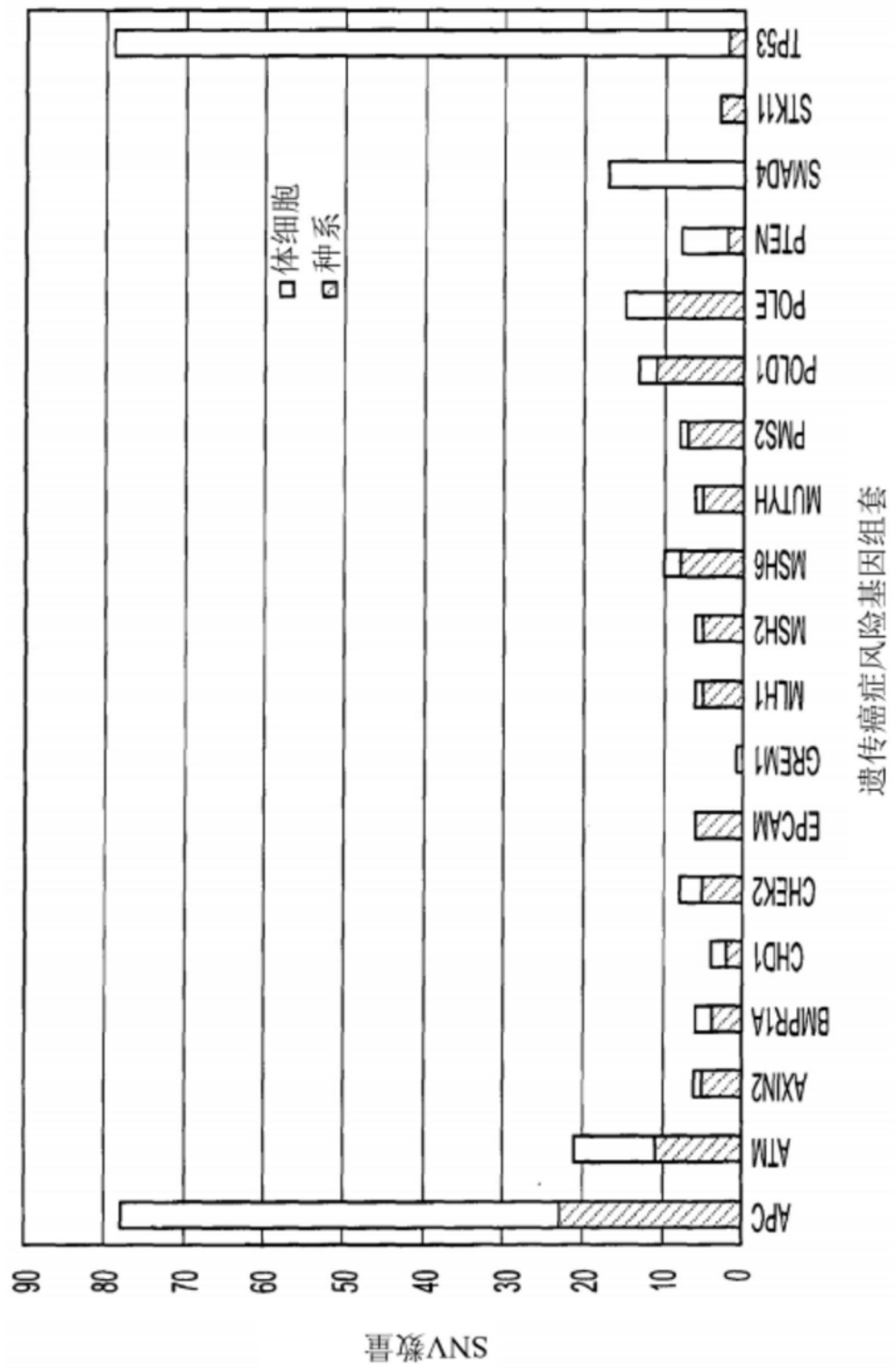

TP53
STK11
SMAD4
PTEN
POLE
POLD1
PMS2
MUTYH
MSH6
MSH2
MLH1
GREM1
EPCAM
CHEK2
CHD1
BMPR1A
AXIN2
ATM
APC
遗传癌症风险基因组套
□体细胞
■种系
90
80
70
60
50
40
30
20
10
0
SNV数量

图6B

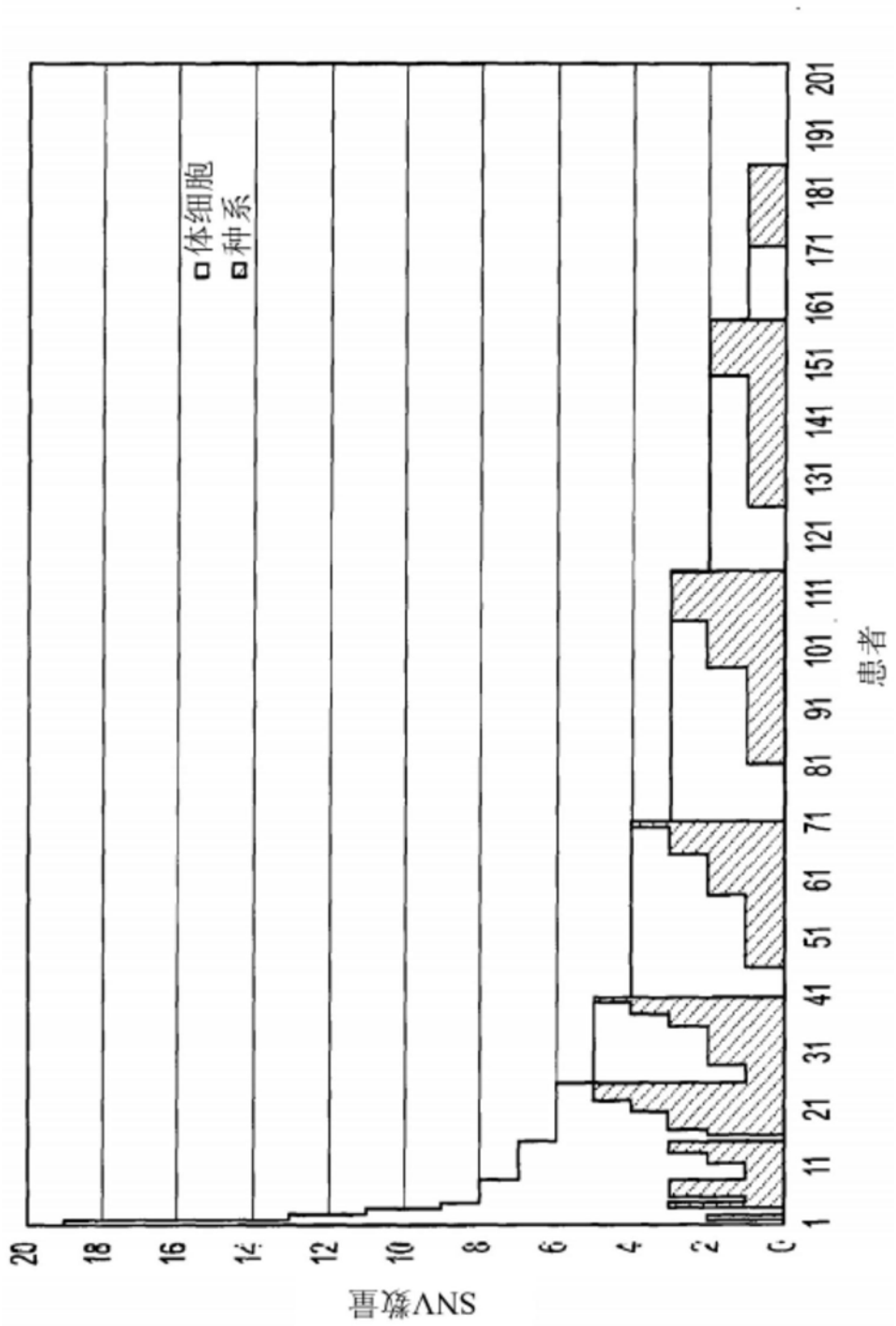


图7

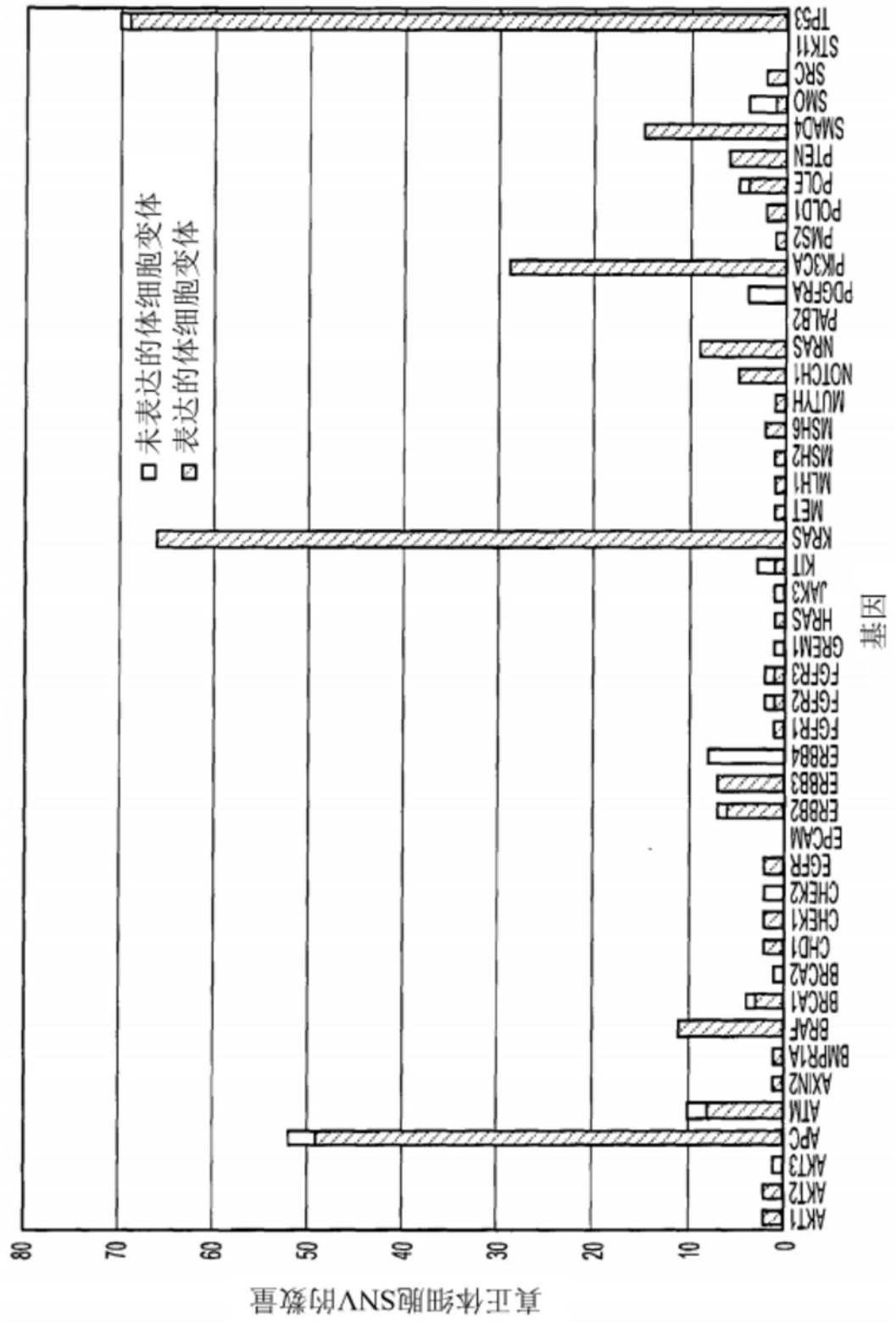

未表达的体细胞变体
表达的体细胞变体
基因
真正体细胞SNV的数量
0
10
20
30
40
50
60
70
80
AKT1
AKT2
AKT3
APC
ATM
AXIN2
BMPR1A
BRAF
BRCA1
BRCA2
CHD1
CHEK1
CHEK2
EGFR
EPCAM
ERBB2
ERBB3
ERBB4
FGFR1
FGFR2
FGFR3
GREM1
HRAS
JAK3
KIT
KRAS
MET
MLH1
MSH2
MSH6
MUTYH
NOTCH1
NRAS
PALB2
PDGFRA
PIK3CA
PMS2
POLD1
POLE
PTEN
SMAD4
SMO
SRC
STK11
TP53

图8

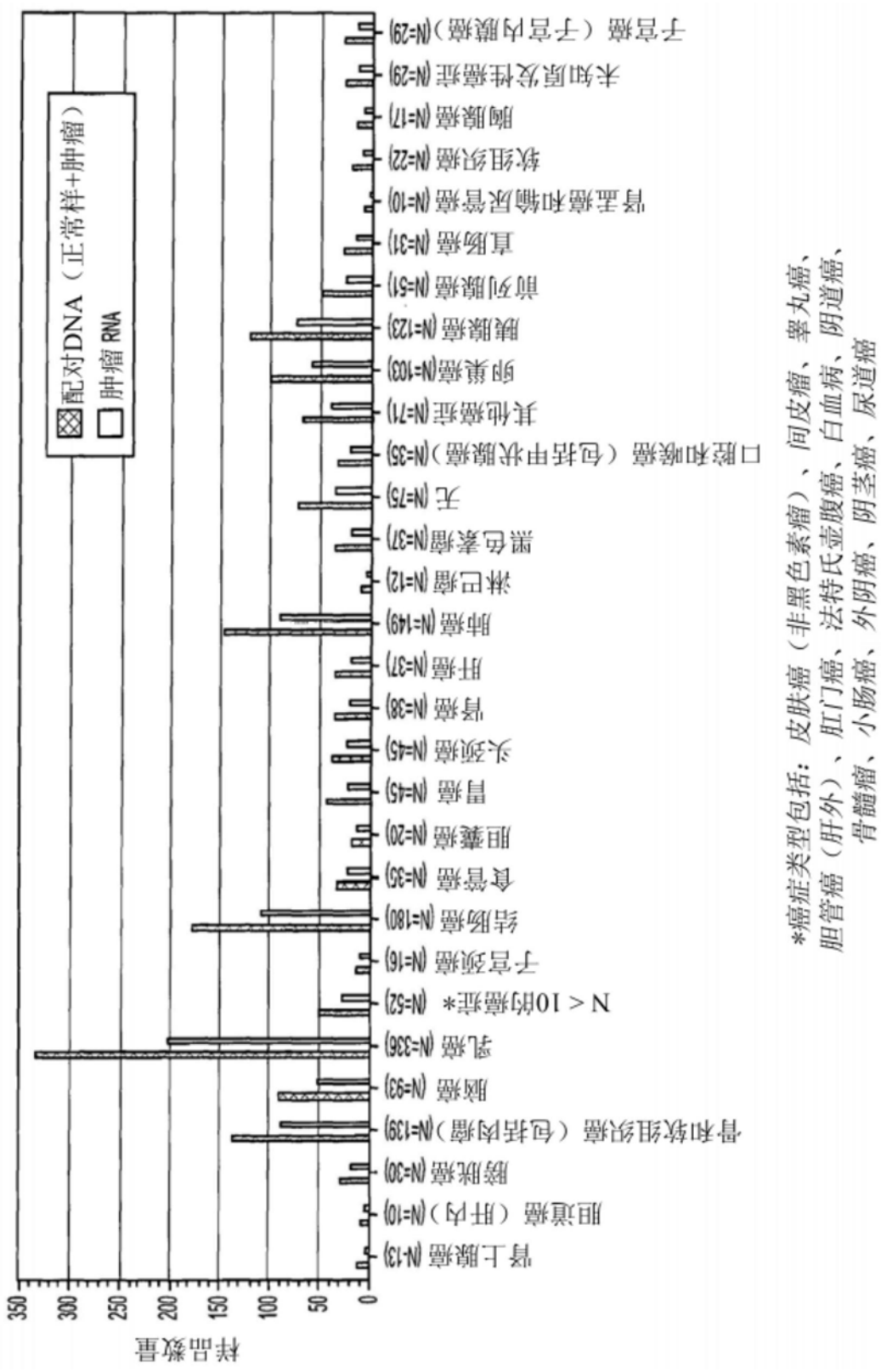

配对DNA（正常样+肿瘤）
肿瘤 RNA
样品数量
0
50
100
150
200
250
300
350
肾上腺癌 (N=13)
胆道癌 （肝内）(N=10)
膀胱癌 (N=30)
骨和软组织癌 （包括肉瘤）(N=139)
脑癌 (N=93)
乳癌 (N=336)
N < 10的癌症* (N=52)
子宫颈癌 (N=16)
结肠癌 (N=180)
食管癌 (N=35)
胆囊癌 (N=20)
胃癌 (N=45)
头颈癌 (N=45)
肾癌 (N=38)
肝癌 (N=37)
肺癌 (N=149)
淋巴瘤 (N=12)
黑色素瘤(N=37)
无 (N=75)
口腔和喉癌 （包括甲状腺癌）(N=35)
其他癌症 (N=71)
卵巢癌(N=103)
胰腺癌 (N=123)
前列腺癌 (N=51)
直肠癌 (N=31)
肾盂癌和输尿管癌 (N=10)
软组织癌 (N=22)
胸腺癌 (N=17)
未知原发性癌症 (N=29)
子宫癌 （子宫内膜癌）(N=29)
*癌症类型包括：皮肤癌（非黑色素瘤）、间皮瘤、睾丸癌、
胆管癌（肝外）、肛门癌、法特氏壶腹癌、白血病、阴道癌、
骨髓瘤、小肠癌、外阴癌、阴茎癌、尿道癌

图9

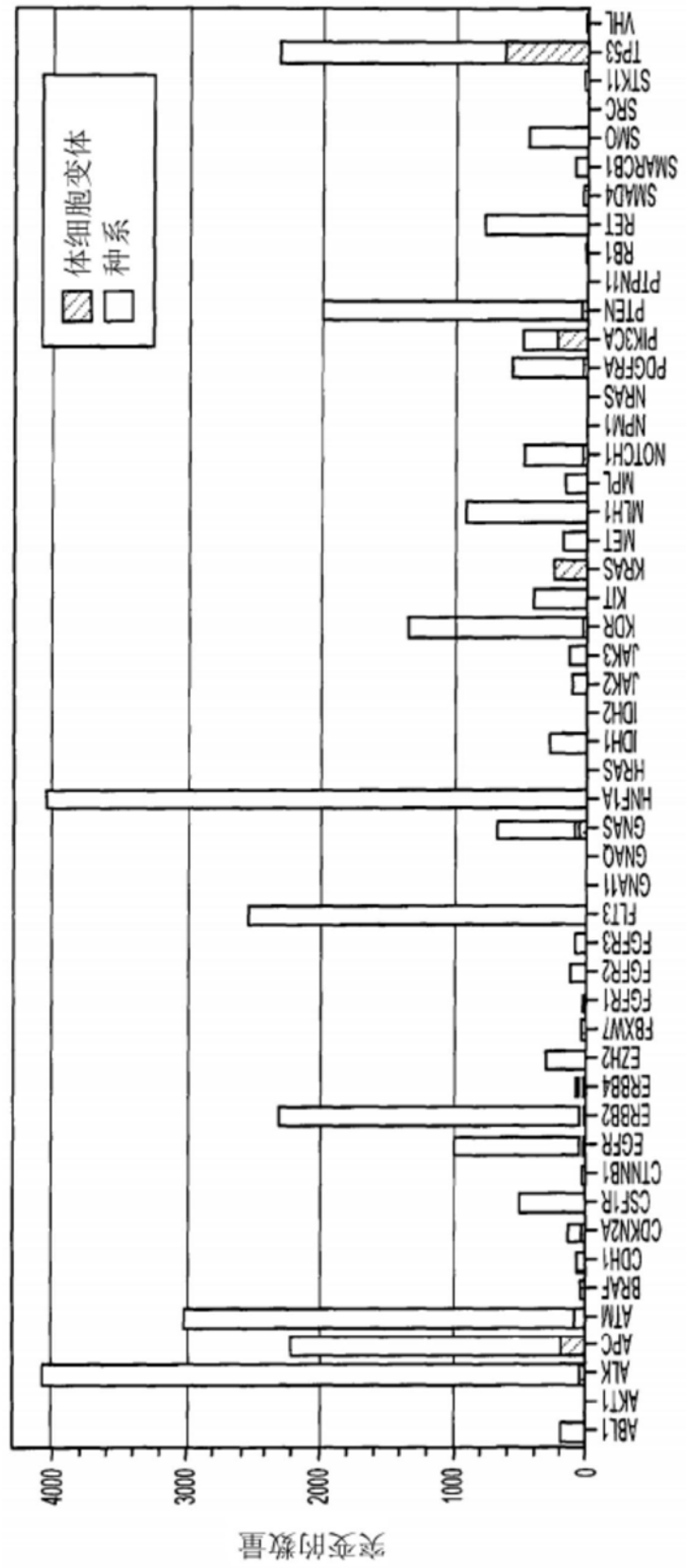

体细胞变体
种系
VHL
TP53
STK11
SRC
SMO
SMARCB1
SMAD4
RET
RB1
PTPN11
PTEN
PIK3CA
PDGFRA
NRAS
NPM1
NOTCH1
MPL
MLH1
MET
KRAS
KIT
KDR
JAK3
JAK2
IDH2
IDH1
HRAS
HNF1A
GNAS
GNAQ
GNA11
FLT3
FGFR3
FGFR2
FGFR1
FBXW7
EZH2
ERBB4
ERBB2
EGFR
CTNNB1
CSF1R
CDKN2A
CDH1
BRAF
ATM
APC
ALK
AKT1
ABL1
4000
3000
2000
1000
0
突变的数量

图10

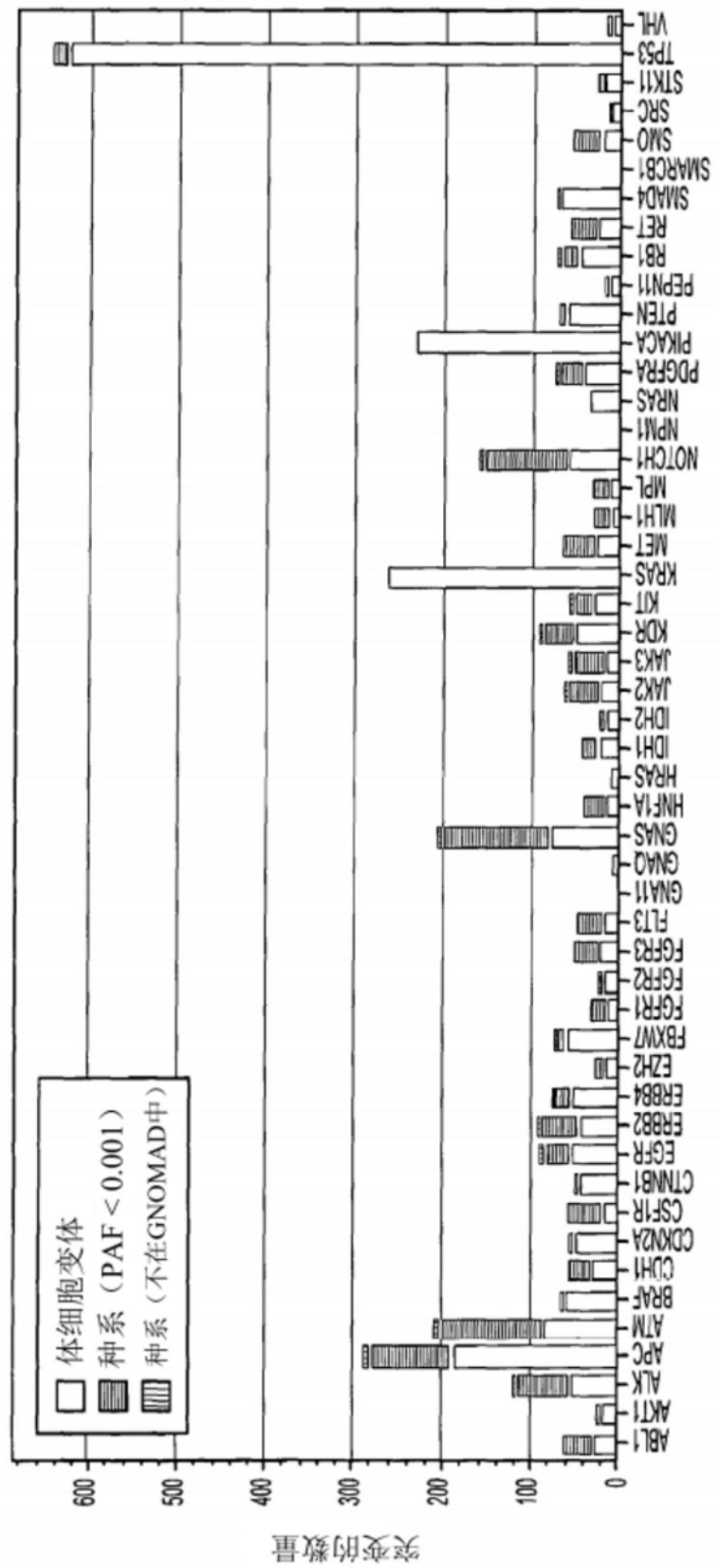

体细胞变体
种系（PAF < 0.001）
种系（不在GNOMAD中）
突变的数量
0
100
200
300
400
500
600
VHL
TP53
STK11
SRC
SMO
SMARCB1
SMAD4
RET
RB1
PEPN11
PTEN
PIKACA
PDGFRA
NRAS
NPM1
NOTCH1
MPL
MLH1
MET
KRAS
KIT
KDR
JAK3
JAK2
IDH2
IDH1
HRAS
HNF1A
GNAS
GNAQ
GNA11
FLT3
FGFR3
FGFR2
FGFR1
FBXW7
EZH2
ERBB4
ERBB2
EGFR
CTNNB1
CSF1R
CDKN2A
CDH1
BRAF
ATM
APC
ALK
AKT1
ABL1

图11

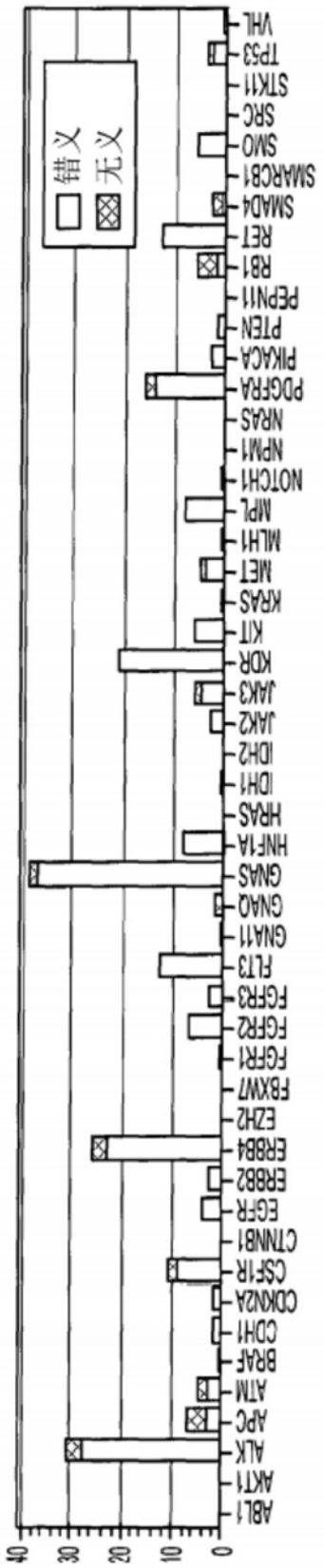

错义
无义
VHL
TP53
STK11
SRC
SMO
SMARCB1
SMAD4
RET
RB1
PEPN11
PTEN
PIKACA
PDGFRA
NRAS
NPM1
NOTCH1
MPL
MLH1
MET
KRAS
KIT
KDR
JAK3
JAK2
IDH2
IDH1
HRAS
HNF1A
GNAS
GNAQ
GNA11
FLT3
FGFR3
FGFR2
FGFR1
FBXW7
EZH2
ERBB4
ERBB2
EGFR
CTNNB1
CSF1R
CDKN2A
CDH1
BRAF
ATM
APC
ALK
AKT1
ABL1
40
30
20
10
0

图12

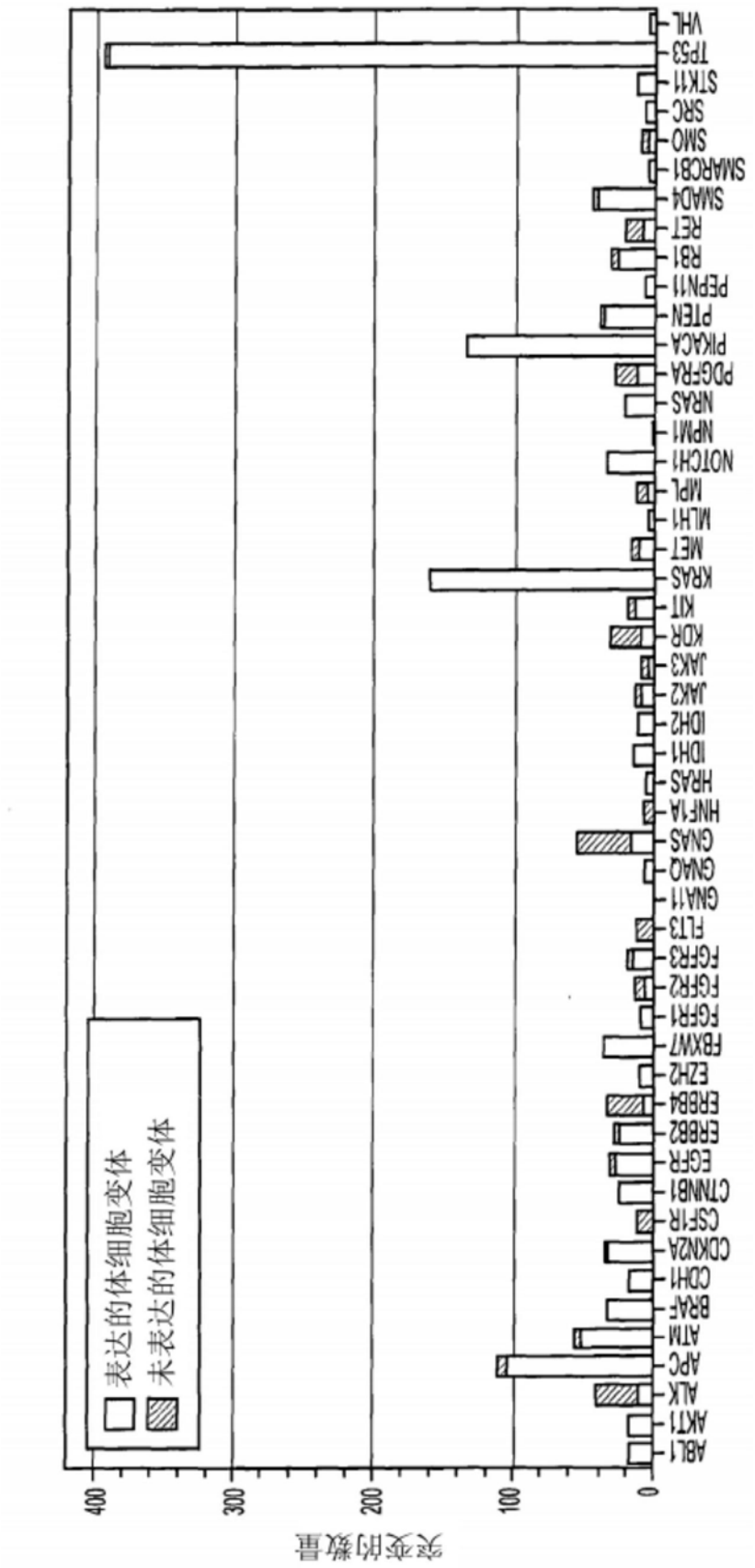

表达的体细胞变体
未表达的体细胞变体
突变的数量
0
100
200
300
400
ABL1
AKT1
ALK
APC
ATM
BRAF
CDH1
CDKN2A
CSF1R
CTNNB1
EGFR
ERBB2
ERBB4
EZH2
FBXW7
FGFR1
FGFR2
FGFR3
FLT3
GNA11
GNAQ
GNAS
HNF1A
HRAS
IDH1
IDH2
JAK2
JAK3
KDR
KIT
KRAS
MET
MLH1
MPL
NOTCH1
NPM1
NRAS
PDGFRA
PIK3CA
PTEN
PEFN11
RB1
RET
SMAD4
SMARCB1
SMO
SRC
STK11
TP53
VHL

图13