(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 106951732 B
(45)授权公告日 2020.03.10

(21)申请号 201611137830.4

(22)申请日 2011.05.25

(65)同一申请的已公布的文献号
申请公布号 CN 106951732 A

(43)申请公布日 2017.07.14

(30)优先权数据
61/396,356 2010.05.25 US

(62)分案原申请数据
201180025750.9 2011.05.25

(73)专利权人 加利福尼亚大学董事会
地址 美国加利福尼亚州

(72)发明人 J·Z·森波 D·豪斯勒

(74)专利代理机构 北京纪凯知识产权代理有限公司 11245
代理人 王永伟 颜芳

(51)Int.Cl.
G16B 30/00(2019.01)
G16H 50/20(2018.01)

(56)对比文件

CN 101539967 A,2009.09.23,

Tobias Sjoblom et al..The consensus coding sequences of human breast and colorectal cancers.《Science》.2006,第314卷(第5797期),

Quinlan et al..BEDTools: a flexible suite of utilities for comparing genomic features.《BIOINFORMATICS》.2010,第26卷(第6期),

何华.DNA序列比对最大似然度进化模型.《中国优秀硕士学位论文全文数据库信息科技辑(月刊)》.2009,(第7期),

MATTHIAS H. KRAUS et al..A position 12-activated H-ras oncogene in all HS578T mammary carcinosarcoma cells but not normal mammary cells of the same patient.《Biochemistry》.1984,第18卷

审查员 杜锦锦

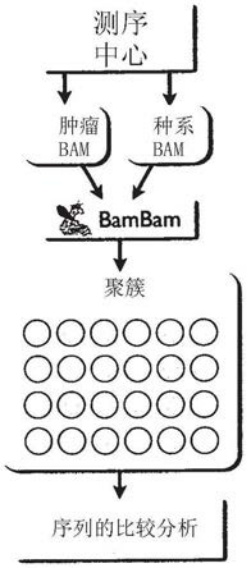
权利要求书2页 说明书20页 附图3页

(54)发明名称

基于计算机的基因组序列分析系统

(57)摘要

本发明涉及评价和/或预测临床状况如癌症、转移、AIDS、孤独症、阿尔茨海默症和/或帕金森症的结果的方法。本方法还可用于监测和追踪临床治疗方案过程中和之后患者DNA和/或RNA的变化。本方法还可用于评价与这种临床状况相关的蛋白质和/或代谢物水平。本方法还用于确定患者特定预后的概率结果。



CN 106951732 B

1.基于计算机的基因组序列分析系统,包括:

存储设备,其存储至少两个基因组序列数据组,所述至少两个基因组序列数据组包括:

肿瘤序列数据组,包括患者的肿瘤组织样本的肿瘤序列串,所述肿瘤序列串包括肿瘤阅读组;和

匹配的正常数据组,包括同一患者的正常组织样本的匹配的正常序列串,所述匹配的正常序列串包括种系阅读组;和

序列分析引擎,其与所述存储设备连接并且被配置以:

鉴定所述肿瘤序列串与所述匹配的正常序列串之间的共同基因组位置,以获得基因组阅读的累积对;

同时地和同步地从所述肿瘤序列数据组检索所述肿瘤序列串的肿瘤阅读和从所述匹配的正常数据组检索与所述共同基因组位置重叠的所述匹配的正常序列串的种系阅读;

将检索的所述肿瘤阅读和所述种系阅读存储在计算机的内存中,其中基于给定基因组位置,所述肿瘤序列串的肿瘤阅读与所述匹配的正常序列串的种系阅读在所述计算机的所述内存中递增地同步,并且其中所述递增地同步包括从所述内存舍弃不与下一个共同基因组位置重叠的累积对的基因组阅读;

根据从所述肿瘤序列串的序列阅读和所述匹配的正常序列串的序列阅读得到的概率,鉴定与所述给定基因组位置相关的基因组改变,其中通过对于所述内存中的累积对,将所述内存中的所述肿瘤阅读与内存中的所述种系阅读进行比较,来鉴定所述基因组改变;和

在设备内存中存储所述基因组改变。

2.根据权利要求1所述的系统,其中所述基因组改变包括基因组变体。

3.根据权利要求2所述的系统,其中所述基因组变体包括体细胞变体。

4.根据权利要求2所述的系统,其中所述基因组变体包括种系变体。

5.根据权利要求1所述的系统,其中所述基因组改变包括单核苷酸多态性。

6.根据权利要求1所述的系统,其中所述基因组改变包括选自下列的改变:等位基因特异的拷贝数、杂合性丢失、结构重排、染色体融合和断点。

7.根据权利要求1所述的系统,其中所述肿瘤序列数据组包括肿瘤BAM文件。

8.根据权利要求1所述的系统,其中所述匹配的正常数据组包括正常BAM文件。

9.根据权利要求1所述的系统,其中所述基因组序列数据组包括多于两个数据组。

10.根据权利要求9所述的系统,其中所述多于两个数据组包括相关测序数据组。

11.根据权利要求9所述的系统,其中所述多于两个数据组包括至少一个复发数据组。

12.根据权利要求1所述的系统,其中序列分析引擎进一步被配置以与所述肿瘤序列数据组和匹配的正常数据组同时地读取和同步第三数据组。

13.根据权利要求1所述的系统,进一步包括基因组浏览器,所述基因组浏览器被配置以显示与所述肿瘤序列数据组和所述匹配的正常数据组相关的基因组改变。

14.根据权利要求1所述的系统,其中所述存储设备被配置以将所述至少两个基因组序列数据组作为文件存储在文件系统中。

15.根据权利要求1所述的系统,其中所述肿瘤序列串和所述正常序列串的给定基因组位置基于参考基因组。

16.根据权利要求1所述的系统,其中所述概率通过最大化肿瘤和种系基因型的联合似

然来确定。

17.根据权利要求16所述的系统，其中最大化所述联合似然包括从患者数据得到如下限定的概率：

$$P(D_g,D_t,G_g,G_t|\alpha,r)=P(D_g|G_g)P(G_g|r)P(D_t|G_g,G_t,\alpha)P(G_t|G_g) \quad (1)$$

其中r是观察的参考等位基因，α是正常污染的分数，其中肿瘤和种系基因型由$Gt=(t_1,t_2)$和$Gg=(g_1,g_2)$限定，其中$t_1,t_2,g_1,g_2\in\{A,T,C,G\}$，并且其中肿瘤和种系序列数据分别被限定为下列阅读组：$D_t=\{d_t^1,d_t^2\cdots,d_t^n\}$和$D_g=\{d_g^1,d_g^2\cdots,d_g^n\}$，其中观察到的碱基$d_t^i,d_g^i\in\{A,T,C,G\}$，其中给定种系基因型的种系等位基因的概率被模拟为四种核苷酸的多项式：

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

其中n是基因组位置的种系阅读的总数，并且n_A、n_G、n_C、n_t是支持各观察到的等位基因的阅读，并且给定肿瘤基因型的肿瘤等位基因的概率被模拟为四种核苷酸的多项式：

$$P(D_t|G_t,G_g,\alpha)=\frac{n!}{n_A!n_T!n_G!n_C!}\prod_i^n P(d_t^i|G_t,G_g,\alpha),$$

其中n是基因组位置的肿瘤阅读的总数，并且n_A、n_G、n_C、n_T是支持各观察到的等位基因的患者数据阅读。

18.根据权利要求1所述的系统，其中所述肿瘤组织样本或所述正常组织样本中的至少一种是血液。

基于计算机的基因组序列分析系统

[0001] 本申请是分案申请，原申请的申请日为2011年5月25日，中国申请号为201180025750.9，国际申请号为PCT/US2011/000939，发明名称为“BAMBAM:高通量测序数据的平行比较分析”。

[0002] 与其他申请的关系

[0003] 本申请涉及2010年5月25日提交的名为“Bambam:高通量测序数据的平行比较分析”的美国临时专利申请序号61/396,356，并且要求其优先权，在此将其全部内容引入作为参考。

[0004] 本发明部分利用下列美国联邦机构的资金进行：国家癌症研究所编号1U24CA143858-01。美国联邦政府对本发明拥有一定权利。

技术领域

[0005] 本发明涉及处理个体或对象生物途径的数据和鉴定其组分从而确定个体或对象是否具有病症或疾病危险的方法。本方法可用作利用SAM/BAM格式的文件中存储的短阅读比对(short-read alignment)对个体或对象的肿瘤和种系测序数据进行比较分析的工具。数据处理方法计算总拷贝数和等位基因特异性拷贝数，使等位基因失衡区域的种系序列分阶(phase)，发现体细胞和种系序列变体，和推断体细胞和种系的结构变化区域。本发明还涉及利用本方法诊断对象是否易患癌症、自身免疫性疾病、细胞周期疾病或其他疾病。

背景技术

[0006] 现代癌症治疗的核心前提是，患者诊断、预后、危险评估和治疗响应预期可通过癌症分类(stratification)得到提高，癌症分类基于肿瘤基因组、转录和外因基因组特征，同时还有诊断时收集的相关临床信息(例如，患者病史、肿瘤组织学及阶段)以及随后的临床后续数据(例如，治疗方案和疾病复发事件)。

[0007] 随着诸如癌症基因组图谱(TCGA)的项目发布多发性肿瘤和匹配的正常全基因组序列，极其需要可由这些大数据组(TCGA,2008)提取尽可能多的基因组信息的计算有效的工具。考虑到高覆盖(>30X)下单个患者的全基因组序列的压缩形式可能是数以百计的千兆字节，比较成对的这些大数据组的分析缓慢且难以管理，但对于发现各个患者肿瘤中存在的多种基因组变化绝对是有必要的。

[0008] 乳腺癌在临床上和基因组方面是异质的，由几种病理和分子方面不同的亚型组成。在各亚型中，患者对常规和目标治疗剂的响应不同，推动了标记物引导的治疗策略的发展。乳腺癌细胞系的集合反映出多种在肿瘤中发现的分子亚型和途径，表明用候选治疗性化合物治疗细胞系可导致分子亚型、途径和药物响应之间的关联得到确定。在77种治疗性化合物的测试中，几乎全部药物在这些细胞系中显示差异响应，约一半显示亚型、途径和/或基因组异常-特异性响应。这些观察结果暗示了可指示临床药物调配的响应和抗性机制以及有效组合药物的尝试。

[0009] 目前需要提供可用于表征、诊断、治疗和确定疾病和病症结果的方法。

发明内容

[0010] 本发明提供了生成可用于确定个体危险的数据库的方法，该个体危险具体是，例如，但不限于，个体易患疾病、病症或状况的危险；个体工作地点、住所、学校或类似地点的危险；个体暴露于毒素、致癌物质、突变剂及类似物的危险；以及个体饮食习惯的危险。此外，本发明提供了可用于鉴定具体个体、动物、植物或微生物的方法。

[0011] 在一个实施方式中，本发明提供了得到差异遗传序列对象(目标)的方法，该方法包括：提供对遗传数据库的访问，该遗传数据库存储(a)表示第一组织的第一遗传序列串(sequence string)和(b)表示第二组织的第二遗传序列串，其中第一和第二序列串具有多个相应的子串(sub-string)；提供对与遗传数据库连接的序列分析引擎(engine)的访问；通过利用多个相应子串中至少一个的已知位置递增地同步第一和第二序列串，利用序列分析引擎形成局部比对；通过序列分析引擎，利用局部比对生成局部比对中第一与第二序列串之间的局部差异串；和通过序列分析引擎，利用局部差异串更新差异序列数据库中的差异遗传序列对象。在优选实施方式中，第一和第二遗传序列串分别表示第一和第二组织至少10％的基因组、转录组或蛋白质组。在可选的优选实施方式中，第一和第二遗传序列串分别表示第一和第二组织至少50％的基因组、转录组或蛋白质组。在另一可选的优选实施方式中，第一和第二遗传序列串分别表示第一和第二组织的基本上整个基因组、转录组或蛋白质组。在另一优选的实施方式中，相应的子串包括纯合等位基因。在可选的优选实施方式中，相应的子串包括杂合等位基因。在另一更优选的实施方式中，遗传序列对象包括文件。在还更优选的实施方式中，文件符合标准化格式。在最优选的实施方式中，文件符合SAM/BAM格式。

[0012] 在优选实施方式中，同步步骤包括，基于第一串中的先验已知位置比对多个子串中的至少一个。在可选的优选实施方式中，同步步骤包括，基于已知参考串——包括多个子串中至少一个的已知位置——比对多个子串中的至少一个。在更优选的实施方式中，已知参考串是共有序列。

[0013] 在另一优选的实施方式中，同步步骤包括，比对窗口中多个子串中的至少一个，该窗口的长度小于多个子串中至少一个的长度。

[0014] 在另一优选的实施方式中，差异遗传序列对象表示至少一条染色体的多个局部差异串。

[0015] 在另一优选的实施方式中，差异遗传序列对象表示第一组织的基本上整个基因组的多个局部差异串。

[0016] 还有其他优选实施方式中，差异遗传序列对象包括这样的特征：包括描述差异遗传序列对象的元数据。在更优选的实施方式中，特征包括第一和第二组织的至少一种的状态。在还更优选的实施方式中，状态包括第一和第二组织中至少一种的生理状态。在最优选的实施方式中，生理状态包括选自肿瘤生长、凋亡、分化状态、组织年龄和治疗响应性的状态。

[0017] 在可选的更优选的实施方式中，状态包括遗传状况。在最优选的实施方式中，遗传状况包括选自至少一种倍性、基因拷贝数、重复拷贝数、倒位、缺失、病毒基因插入、体细胞突变、种系突变、结构重排、易位和杂合性丢失的状况。

[0018] 在可选的更优选的实施方式中，状态包括组织中与信号传导途径相关的途径模型

信息。在最优选的实施方式中，信号传导途径选自生长因子信号传导途径、转录因子信号传导途径、凋亡途径、细胞周期途径和激素响应途径。

[0019] 在可选的实施方式中，第一和第二组织源自相同的生物实体，生物实体选自患者、健康个体、细胞系、干细胞、实验动物模型、重组细菌细胞和病毒。在可选的实施方式中，第一组织是健康组织，并且其中第二组织是患病组织。在更优选的实施方式中，患病组织包括肿瘤组织。

[0020] 本发明还提供了如本文公开的方法，其中该方法进一步包括如下步骤：在第一序列串全长中，迭代地递增地同步化第一和第二序列串。

[0021] 本发明还提供了提供健康护理服务的方法，该方法包括：提供对与医疗记录存储设备在信息上连接的分析引擎的访问，其中存储设备存储患者的差异遗传序列对象；利用患者差异遗传序列对象中存在多个局部差异串的局部差异串或丛(constellation)，通过分析引擎产生患者特异的数据组；和基于患者特异的数据组，通过分析引擎产生患者特异的指示。在优选实施方式中，医疗记录存储设备被配置为智能卡，并由患者携带。在另一优选的实施方式中，医疗记录存储设备被健康护理人员远程访问。还有其它优选实施方式中，患者的差异遗传序列对象包括至少两条染色体的多个局部差异串。在更进一步优选的实施方式中，患者的差异遗传序列对象包括基本上患者整个基因组的多个局部差异串。在另一优选的实施方式中，患者的差异遗传序列对象包括表示至少两种组织类型或相同组织的至少两个时间间隔结果的多个局部差异串。在更优选的实施方式中，相同组织的至少两个时间间隔结果得自治疗开始之前和之后。在最优选的实施方式中，相同组织的至少两个时间间隔结果得自治疗开始之前和之后。

[0022] 在另一可选的优选实施方式中，本文公开的患者特异的指示选自诊断、预后、治疗结果预期、治疗策略建议和处方。

[0023] 本发明还提供了分析群体的方法，该方法包括：在群体医疗记录数据库中获得和存储多个差异遗传序列对象，其中该记录数据库与分析引擎在信息上连接；通过分析引擎鉴定多个差异遗传序列对象中的多个局部差异串丛，从而产生丛记录；和通过分析引擎利用丛记录生成群体分析记录。在优选实施方式中，群体包括多个血亲。在可选的优选实施方式中，群体包括特征在于共享至少一个共同特征的多个成员，该共同特征选自暴露于病原、暴露于毒性剂、健康史、治疗史、治疗成功、性别、物种和年龄。在另一可选的优选实施方式中，群体包括特征在于共享至少一个共同特征的多个成员，该共同特征选自地理位置、种族和职业。在更进一步可选的优选实施方式中，群体分析记录包括父子关系或母子关系的确定。

[0024] 在可选的实施方式中，本文公开的方法进一步包括将个体患者的丛记录与群体分析记录进行比较的步骤。在优选实施方式中，将个体患者的丛记录与群体分析记录进行比较的步骤生成患者特异的记录。在更优选的实施方式中，患者特异的记录包括危险评估或鉴定患者属于指定群体。在可选的更优选的实施方式中，患者特异的记录包括诊断、预后、治疗结果预期、治疗策略建议和处方。

[0025] 本发明进一步提供了分析个人的差异遗传序列对象的方法，该方法包括：在与分析引擎在信息上连接的医疗记录数据库中存储参考差异遗传序列对象；通过分析引擎计算个人差异遗传序列对象中的多个局部差异串与参考差异遗传序列对象中的多个局部差异

串之间的偏差，产生偏差记录；通过分析引擎利用偏差记录生成个人特异性偏差概况。在优选实施方式中，参考差异遗传序列对象由个人的多个局部差异串计算得到。在另一优选的实施方式中，参考差异遗传序列对象由个人的多个局部差异串计算得到。

[0026] 关于本文公开的各种方法，在优选实施方式中，患者或个人选自诊断患有状况的患者或个人，该状况选自疾病和病症。在更优选的实施方式中，状况选自获得性免疫缺陷综合征(AIDS)、阿狄森病、成人呼吸窘迫综合征、过敏症、强直性脊柱炎、淀粉样变性病、贫血、哮喘、动脉粥样硬化、自身免疫性溶血性贫血、自身免疫性甲状腺炎、良性前列腺增生症、支气管炎、切东二氏综合征、胆囊炎、克罗恩病、特应性皮炎、皮肌炎、糖尿病、气肿、胎儿红细胞增多症、结节性红斑、萎缩性胃炎、肾小球性肾炎、古德帕斯彻综合征、痛风、慢性肉芽肿性疾病、格雷夫斯病、桥本甲状腺炎、嗜伊红细胞增多症、肠易激综合征、多发性硬化症、重症肌无力、心肌或心包炎症、骨关节炎、骨质疏松、胰腺炎、多囊卵巢综合征、多发性肌炎、银屑病、莱特尔综合征、类风湿性关节炎、硬皮病、重度联合免疫缺陷病(SCID)、斯耶格伦综合征、全身过敏、全身性红斑狼疮、系统性硬化症、血小板减少性紫癜、溃疡性结肠炎、葡萄膜炎、维尔纳综合征、癌症并发症、血液透析和体外循环、病毒、细菌、真菌、寄生虫、原生动物和蠕虫感染；和腺癌、白血病、淋巴瘤、黑素瘤、骨髓瘤、肉瘤、畸胎癌和具体地，肾上腺癌、膀胱癌、骨癌、骨髓癌、脑癌、乳腺癌、子宫颈癌、胆囊癌、神经节癌、胃肠道癌、心脏癌、肾癌、肝癌、肺癌、肌癌、卵巢癌、胰腺癌、甲状旁腺癌、阴茎癌、前列腺癌、唾液腺癌、皮肤癌、脾癌、睾丸癌、胸腺癌、甲状腺癌和子宫癌，静坐不能、阿尔茨海默症、健忘症、肌萎缩性侧索硬化(ALS)、共济失调、双极性疾病、紧张症、大脑性麻痹、脑血管疾病、克-雅二氏病、痴呆、抑郁、唐氏综合征、迟发性运动障碍、张力障碍、癫痫、亨廷顿病、多发性硬化症、肌肉萎缩症、神经痛、神经纤维瘤、神经病、帕金森病、皮克病、色素性视网膜炎、精神分裂症、季节性情绪疾病、老年痴呆、中风、图雷特综合征和包括腺癌、黑素瘤和畸胎癌在内的癌症，特别是脑癌。

[0027] 在另一优选的实施方式中，状况选自癌症，如腺癌、白血病、淋巴瘤、黑素瘤、骨髓瘤、肉瘤、畸胎癌，和具体地，肾上腺癌、膀胱癌、骨癌、骨髓癌、脑癌、乳腺癌、子宫颈癌、胆囊癌、神经节癌、胃肠道癌、心脏癌、肾癌、肝癌、肺癌、肌癌、卵巢癌、胰腺癌、甲状旁腺癌、阴茎癌、前列腺癌、唾液腺癌、皮肤癌、脾癌、睾丸癌、胸腺癌、甲状腺癌和子宫癌；免疫疾病，如获得性免疫缺陷综合征(AIDS)、阿狄森病、成人呼吸窘迫综合征、过敏症、强直性脊柱炎、淀粉样变性病、贫血、哮喘、动脉粥样硬化、自身免疫性溶血性贫血、自身免疫性甲状腺炎、支气管炎、胆囊炎、接触性皮炎、克罗恩病、特应性皮炎、皮肌炎、糖尿病、气肿、淋巴细胞毒素的发作性淋巴细胞减少症、胎儿红细胞增多症、结节性红斑、萎缩性胃炎、肾小球性肾炎、古德帕斯彻综合征、痛风、格雷夫斯病、桥本甲状腺炎、嗜伊红细胞增多症、肠易激综合征、多发性硬化症、重症肌无力、心肌或心包炎症、骨关节炎、骨质疏松、胰腺炎、多发性肌炎、银屑病、莱特尔综合征、类风湿性关节炎、硬皮病、斯耶格伦综合征、全身过敏、全身性红斑狼疮、系统性硬化症、血小板减少性紫癜、溃疡性结肠炎、葡萄膜炎、维尔纳综合征、癌症并发症、血液透析和体外循环、病毒、细菌、真菌、寄生虫、原生动物和蠕虫感染、创伤、布鲁顿X-连锁无丙种球蛋白血症、常见变异型免疫缺陷(CV1)、迪乔治综合征(胸腺发育不全)、胸腺发育不良、隔离IgA缺乏症(isolated IgA deficiency)、重度联合免疫缺陷病(SCID)、血小板减少症和湿疹的免疫缺陷(威-奥氏综合征)、切东二氏综合征、慢性肉芽肿性疾病、遗传性血管神经性水肿和库兴病相关的免疫缺陷；和发育疾病，如肾小管酸中毒、贫血、库兴综合征、

软骨发育不全性侏儒(achondroplastic dwarfism)、杜兴和贝克尔肌肉萎缩症、癫痫、性腺发育不全、WAGR综合征(威尔姆斯瘤、无虹膜、泌尿生殖系统异常和精神发育迟滞)、史密斯-马盖尼斯综合征、骨髓增生异常综合征、遗传性粘膜上皮异常增生、遗传性皮肤角化病、遗传性神经病如夏-马-图病和神经纤维瘤、甲状腺功能减退症、脑积水、癫痫病如Syndenham舞蹈病和大脑性麻痹、脊柱裂、无脑畸形、颅脊柱裂、先天性青光眼、白内障、感觉神经性听力损失;以及与细胞生长和分化、胚胎发生和形态发生相关的任何疾病,包括对象的任何组织、器官或系统,例如、脑、肾上腺、肾、骨骼或生殖系统。

[0028] 在更进一步可选的优选实施方式中,状况选自内分泌疾病,如与垂体功能减退相关的疾病,包括性腺功能减退、席汉综合征、尿崩症、卡尔曼病、汉-许-克三氏病、累-赛二氏病、结节病、空蝶鞍综合征和侏儒症;垂体功能亢进,包括肢端肥大症、巨人症和抗利尿激素(ADH)分泌异常综合征(SIADH);和与甲状腺功能减退相关的疾病,包括甲状腺肿、粘液性水肿、与细菌感染相关的急性甲状腺炎、与病毒感染相关的亚急性甲状腺炎、自身免疫性甲状腺炎(桥本病)和呆小症;与甲状腺功能亢进相关的疾病,包括甲状腺毒症及其各种形式、格雷夫斯病、胫骨前粘液性水肿、毒性多结节性甲状腺肿、甲状腺癌和普鲁麦病;和与甲状旁腺功能亢进相关的疾病,包括康恩病(慢性高血钙);呼吸系统疾病,如过敏、哮喘、急性和慢性炎性肺病、ARDS、气肿、肺充血和水肿、COPD、间质性肺病和肺癌;癌症,如腺癌、白血病、淋巴瘤、黑素瘤、骨髓瘤、肉瘤、畸胎癌,和具体地,肾上腺癌、膀胱癌、骨癌、骨髓癌、脑癌、乳腺癌、子宫颈癌、胆囊癌、神经节癌、胃肠道癌、心脏癌、肾癌、肝癌、肺癌、肌癌、卵巢癌、胰腺癌、甲状旁腺癌、阴茎癌、前列腺癌、唾液腺癌、皮肤癌、脾癌、睾丸癌、胸腺癌、甲状腺癌和子宫癌;和免疫学疾病,如获得性免疫缺陷综合征(AIDS)、阿狄森病、成人呼吸窘迫综合征、过敏症、强直性脊柱炎、淀粉样变性病、贫血、哮喘、动脉粥样硬化、自身免疫性溶血性贫血、自身免疫性甲状腺炎、支气管炎、胆囊炎、接触性皮炎、克罗恩病、特应性皮炎、皮肌炎、糖尿病、气肿、淋巴细胞毒素的发作性淋巴细胞减少症、胎儿红细胞增多症、结节性红斑、萎缩性胃炎、肾小球性肾炎、古德帕斯彻综合征、痛风、格雷夫斯病、桥本甲状腺炎、嗜伊红细胞增多症、肠易激综合征、多发性硬化症、重症肌无力、心肌或心包炎症、骨关节炎、骨质疏松、胰腺炎、多发性肌炎、银屑病、莱特尔综合征、类风湿性关节炎、硬皮病、斯耶格伦综合征、全身过敏、全身性红斑狼疮、系统性硬化症、血小板减少性紫癜、溃疡性结肠炎、葡萄膜炎、维尔纳综合征、癌症并发症、血液透析和体外循环、病毒、细菌、真菌、寄生虫、原生动物和蠕虫感染以及创伤。

[0029] 本发明进一步提供了得到差异遗传序列对象的方法,该方法包括:提供对遗传数据库的访问,该遗传数据库存储(a)表示第一组织的第一遗传序列串和(b)表示第二组织的第二遗传序列串,其中第一和第二序列串具有多个相应的子串;提供对与遗传数据库连接的序列分析引擎的访问;通过利用多个相应子串中至少一个的已知位置递增地同步第一和第二序列串,利用序列分析引擎形成局部比对;通过序列分析引擎,利用局部比对生成局部比对中第一与第二序列串之间的局部差异串;和通过序列分析引擎,利用局部差异串生成差异序列数据库中的差异遗传序列对象,从而得到差异序列对象。

[0030] 本发明进一步提供了生成差异遗传序列对象的转化方法,差异遗传序列对象表示第一遗传序列与第二序列之间的临床相关差异,该方法包括步骤:(i)提供对遗传数据库的访问,该遗传数据库存储(a)表示第一组织的第一遗传序列串和(b)表示第二组织的第二遗

传序列串,其中第一和第二序列串具有多个相应的子串;(ii)提供对与遗传数据库连接的序列分析引擎的访问;(iii)通过利用多个相应子串中至少一个的已知位置递增地同步第一和第二序列串,利用序列分析引擎形成局部比对;(iv)通过序列分析引擎,利用局部比对生成局部比对中第一与第二序列串之间的局部差异串;和(v)通过序列分析引擎,利用局部差异串生成差异序列数据库中的差异遗传序列对象,从而得到差异序列对象,其中差异序列对象向用户提供目标信息。

[0031] 在优选实施方式中,目标信息选自遗传相关信息、代谢相关信息、毒理相关信息、临床相关信息、时间相关信息、地理相关信息、职业危险相关信息、生活史相关信息及类似信息。

附图说明

[0032] 图1示例了"BamBam"数据流的示意图。

[0033] 图2示例了等位基因特异性拷贝数计算的概括视图。

[0034] 图3示例了结构变化呼叫的概括视图。

[0035] 图4示例了鉴定基因组中发生结构重排的位置的示例性方法。

[0036] 图5示例了示例性肿瘤特异性基因组浏览器。

具体实施方式

[0037] 本文公开的实施方式是说明性和示例性的,并非意为限制本发明。可应用其他实施方式,并且可进行结构变化,而没有脱离本发明权利要求的范围。

[0038] 如本文和所附权利要求所用,单数形式"一(a)"、"一(an)"和"该(所述,the)"包括复数指代,除非上下文明确另外表示。因此,例如,"一等位基因(或等位基因)"的指代包括多个这种等位基因,"一簇(簇)"的指代是指代一个或多个簇及其等同形式,等等。

[0039] 如本文所用,术语"管理的(curated)"意为根据科学和/或临床原理利用本领域的公知方法测试、分析和鉴定生物分子组和/或非生物分子组之间的关系,本领域的公知方法如分子生物学、生物化学、生理学、解剖学、基因组学、转录组学、蛋白质组学、代谢组学、ADME和生物信息学技术及类似技术。该关系可以是生物化学性的,如生物化学途径、遗传途径、代谢途径、基因调控途径、基因转录途径、基因翻译途径、miRNA调控途径、假基因调控途径及类似途径。

[0040] 高通量数据提供对癌组织中分子变化的全面观察。新技术允许对肿瘤样本和癌细胞系的基因组拷贝数变化、基因表达、DNA甲基化和外遗传的状态进行同时基因组范围分析(genome wide assay)。

[0041] 计划在不久的将来对多种肿瘤进行研究,如癌症基因组图谱(TCGA)、抗癌(Stand Up To Cancer,SU2C)和更多研究。当前数据组的分析发现,患者之间的遗传改变可不同,但通常涉及共同的途径。因此鉴定癌症进程涉及的相关途径和检测其在不同患者中如何改变是非常重要的。

[0042] 在诸如癌症基因组图谱(TCGA)的项目发布了多种完全测序的肿瘤及匹配的正常基因组的情况下,非常需要能够有效分析这些大量数据组的工具。

[0043] 为此目的,我们开发了BamBam,其是利用SAM/BAM-格式的文件(SAMtools

library;Li H,Handsaker B,Wysoker A,Fennell T,Ruan J,Homer N,Marth G,Abecasis G,Durbin R;1000Genome Project Data Processing Subgroup.The Sequence Alignment/Map format and SAMtools.Bioinformatics.2009Aug 15;25(16):2078-9.Epub 2009Jun 8)中包含的比对短阅读据同时分析患者肿瘤和种系基因组的各基因组位置的工具。BamBam连接SAMtools库,利用SAM/BAM-格式文件中的短阅读比对同时分析患者的肿瘤和种系基因组。在本公开中,BamBam工具可以是序列分析引擎,其用于比较序列——包含信息串的序列。在一个实施方式中,信息串包含生物学信息,例如,多核苷酸序列或多肽序列。在另一实施方式中,生物学信息可包括表达数据,例如mRNA转录子或rRNA或tRNA或肽或多肽或蛋白质的相对浓度水平。在另一实施方式中,生物学信息可以是蛋白质修修饰的相对量,该修饰如例如,但不限于,磷酸化、硫酸化、乙酰化、甲基化、糖基化、唾液酸化、用糖基磷脂酰肌醇修饰或用蛋白多糖修饰。

[0044] 本处理方法使BamBam能够有效计算全部拷贝数和推断肿瘤和种系基因组中的结构变化(例如,染色体易位)区域;有效计算全部和等位基因特异性拷贝数;推断呈现杂合性丢失(LOH)的区域;和发现体细胞和种系序列变体(例如,点突变)和结构重排(例如,染色体融合)。此外,通过同时比较两个基因组序列,BamBam还可直接区分体细胞与种系序列变体,计算肿瘤基因组中的等位基因特异性拷贝数变化,和使种系单倍型在肿瘤基因组中等位基因比例改变的染色体区域中分阶。通过将这些分析全部一起引入单个工具,研究人员可利用BamBam发现患者肿瘤基因组中存在的多种类型的基因组改变,通常是特定基因等位基因,其有助于鉴定肿瘤发生的潜在驱动因子。

[0045] 为确定发现的变体是体细胞(即,仅在肿瘤中发现的变体序列)还是种系(即,遗传的或可遗传的变体序列)变体,需要以某种方式比较肿瘤与匹配的正常基因组。这可通过如下相继进行:总结肿瘤和种系的每个基因组位置的数据,然后组合结果用于分析。不幸地是,由于全基因组BAM文件其压缩形式为数百个千兆字节(未压缩是1-2百万兆字节),需要存储用于后续分析的中间结果将是极其巨大的,并且合并和分析极其缓慢。

[0046] 为避免这个问题,BamBam同时读取两个文件,恒定地保持各BAM文件彼此同步,并累积两文件之间每个共同基因组位置重叠的基因组阅读。对于每一对累积(pileup),BamBam运行一系列上述分析,然后舍弃累积,并移至下一个共同基因组位置。通过用本方法处理这些大批量BAM文件,计算机RAM被最低限度地使用,并且处理速度主要受限于文件系统可读取两文件的速度。这使得BamBam能够快速处理大批量数据,同时其灵活性足以在单个计算机上或在整个计算机组中运行。用BamBam处理这些文件的另一重要益处是其输出相当小,仅由各文件中发现的重要差异组成。这产生基本上是患者肿瘤与种系基因组之间的全基因组差异,需要的磁盘存储器远远小于若各文件的全基因组信息均单独存储所占用的磁盘存储器。

[0047] BamBam是计算有效的方法,用于测量大测序数据组,以产生一组高质量基因组事件,该高质量基因组事件存在于相对于其种系的各肿瘤中。这些结果提供对肿瘤染色体动态的扫视,提高我们对肿瘤最终状态及导致其事件的理解。BamBam数据流的示例性方案显示在图1中。

[0048] 本发明的一个具体的示例性实施方式是生成和应用差异遗传序列对象。如本文所用,该对象代表由BamBam技术示例的数字对象,并反映出参考序列(例如,第一序列)与分析

序列(例如,第二序列)之间的差异。对象可被认为是多个不同市场的阻碍。从市场的角度来看,人们可能认为下列因素与该对象的应用和管理有关:

[0049] ○对象可以是关于参数向量(例如,时间、地理区域、遗传树、物种等)的动态的和变化。

[0050] ○对象可被认为相对于对象或参考序列彼此具有"距离"。该距离可根据相关尺寸进行测量。例如,该距离可以是与假设的正常值相距的偏差或相对于时间的趋势。

[0051] ○对象可以指示危险:发生疾病、暴露易感性的危险、在一个地点的工作危险等。

[0052] ○对象可被管理,用于呈现于利益相关者:健康护理人员、保险公司、患者等。

[0053] ■可显示为图形对象

[0054] ■可显示为统计学形式:单个人、群体、标准化人等。

[0055] ○参考序列可由对象生成,形成标准化序列。标准化序列可基于得自所测对象的共有序列而构建。

[0056] ○对象表示为大型亚基因组或基因组信息,而非单个基因比对,并且被注释/包含标准软件可读的元数据。

[0057] ○对象可具有可检测到的内部样式(pattern)或结构:一个点的突变组可与状况相关的另一个点的第二组突变有关;差异样式丛可能是热点;利用多变量分析或其它AI技术来鉴定相关性;检测热点(例如,存在、不存在等)的显著性。

[0058] ○与单个人相关的对象可被用作安全密钥。

[0059] 更新差异序列对象:更新包括生成、修饰、改变、缺失等;

[0060] ○可基于模板。

[0061] ○可以是重新(de novo)对象。

[0062] ○可以是已存在的对象。

[0063] 在可选的示例性实施方式中,本方法可用于确定和预期患者对治疗的响应性:预期的、假设的、预测的、实际的,及类似的。

[0064] 在可选的示例性实施方式中,本方法可用于提供患者特异的指示:处方、建议、预后及类似指示。

[0065] 在一个实施方式中,本方法可用于提供临床信息,该临床信息可用于多种诊断和治疗应用,如检测癌症组织、对癌症组织分期、检测转移组织及类似应用;检测神经疾病,如但不限于,阿尔茨海默症、肌萎缩性侧索硬化(ALS)、帕金森病、精神分裂症、癫痫、及其并发症;发育疾病,如DiGeorge综合征、孤独症;自身免疫性疾病,如多发性硬化症、糖尿病、及类似疾病;治疗感染,如但不限于,病毒感染、细菌感染、真菌感染、利什曼原虫病、血吸虫病、疟疾、绦虫病、象皮病、线虫感染、nematines及类似疾病。

[0066] 在一个实施方式中,本方法可用于提供临床信息,以检测和定量与基因或蛋白质表达改变相关的状况的改变的基因结构、基因突变、基因生物化学修饰,包括信使RNA(mRNA)、核糖体RNA(rRNA)、转移RNA(tRNA)、微RNA(miRNA)、反义RNA(asRNA)及类似物的改变和/或修饰。与表达改变相关的状况、疾病或病症包括获得性免疫缺陷综合征(AIDS)、阿狄森病、成人呼吸窘迫综合征、过敏症、强直性脊柱炎、淀粉样变性病、贫血、哮喘、动脉粥样硬化、自身免疫性溶血性贫血、自身免疫性甲状腺炎、良性前列腺增生症、支气管炎、切东二氏综合征、胆囊炎、克罗恩病、特应性皮炎、皮肌炎、糖尿病、气肿、胎儿红细胞增多症、结节

性红斑、萎缩性胃炎、肾小球性肾炎、古德帕斯彻综合征、痛风、慢性肉芽肿性疾病、格雷夫斯病、桥本甲状腺炎、嗜伊红细胞增多症、肠易激综合征、多发性硬化症、重症肌无力、心肌或心包炎症、骨关节炎、骨质疏松、胰腺炎、多囊卵巢综合征、多发性肌炎、银屑病、莱特尔综合征、类风湿性关节炎、硬皮病、重度联合免疫缺陷病(SCID)、斯耶格伦综合征、全身过敏、全身性红斑狼疮、系统性硬化症、血小板减少性紫癜、溃疡性结肠炎、葡萄膜炎、维尔纳综合征、癌症并发症、血液透析和体外循环、病毒、细菌、真菌、寄生虫、原生动物和蠕虫感染;和腺癌、白血病、淋巴瘤、黑素瘤、骨髓瘤、肉瘤、畸胎癌,和具体地,肾上腺癌、膀胱癌、骨癌、骨髓癌、脑癌、乳腺癌、子宫颈癌、胆囊癌、神经节癌、胃肠道癌、心脏癌、肾癌、肝癌、肺癌、肌癌、卵巢癌、胰腺癌、甲状旁腺癌、阴茎癌、前列腺癌、唾液腺癌、皮肤癌、脾癌、睾丸癌、胸腺癌、甲状腺癌和子宫癌。诊断分析可利用杂交或扩增技术来比较患者生物样本与标准样本中的基因表达,从而检测改变的基因表达。这种比较的定性或定量方法在本领域是公知的。

[0067] 在另一实施方式中,本方法可用于提供临床信息以检测和定量与基因或蛋白质表达改变相关疾病的改变的基因结构、基因突变、基因生物化学修饰,包括信使RNA(mRNA)、核糖体RNA(rRNA)、转移RNA(tRNA)、微RNA(miRNA)、反义RNA(asRNA)及类似物的改变和/或修饰。与表达改变相关的疾病包括静坐不能、阿尔茨海默症、健忘症、肌萎缩性侧索硬化(ALS)、共济失调、双极性疾病、紧张症、大脑性麻痹、脑血管疾病、克-雅二氏病、痴呆、抑郁、唐氏综合征、迟发性运动障碍、张力障碍、癫痫、亨廷顿病、多发性硬化症、肌肉萎缩症、神经痛、神经纤维瘤、神经病、帕金森病、皮克病、色素性视网膜炎、精神分裂症、季节性情绪疾病、老年痴呆、中风、图雷特综合征和癌症——包括腺癌、黑素瘤和畸胎癌,特别是脑癌。

[0068] 在一个实施方式中,本方法可用于提供与哺乳动物蛋白质表达或活性改变相关的状况的临床信息。这种状况的实例包括但不限于,获得性免疫缺陷综合征(AIDS)、阿狄森病、成人呼吸窘迫综合征、过敏症、强直性脊柱炎、淀粉样变性病、贫血、哮喘、动脉粥样硬化、自身免疫性溶血性贫血、自身免疫性甲状腺炎、良性前列腺增生症、支气管炎、切东二氏综合征、胆囊炎、克罗恩病、特应性皮炎、皮肌炎、糖尿病、气肿、胎儿红细胞增多症、结节性红斑、萎缩性胃炎、肾小球性肾炎、古德帕斯彻综合征、痛风、慢性肉芽肿性疾病、格雷夫斯病、桥本甲状腺炎、嗜伊红细胞增多症、肠易激综合征、多发性硬化症、重症肌无力、心肌或心包炎症、骨关节炎、骨质疏松、胰腺炎、多囊卵巢综合征、多发性肌炎、银屑病、莱特尔综合征、类风湿性关节炎、硬皮病、重度联合免疫缺陷病(SCID)、斯耶格伦综合征、全身过敏、全身性红斑狼疮、系统性硬化症、血小板减少性紫癜、溃疡性结肠炎、葡萄膜炎、维尔纳综合征、癌症并发症、血液透析和体外循环、病毒、细菌、真菌、寄生虫、原生动物和蠕虫感染;和腺癌、白血病、淋巴瘤、黑素瘤、骨髓瘤、肉瘤、畸胎癌、和具体地、肾上腺癌、膀胱癌、骨癌、骨髓癌、脑癌、乳腺癌、子宫颈癌、胆囊癌、神经节癌、胃肠道癌、心脏癌、肾癌、肝癌、肺癌、肌癌、卵巢癌、胰腺癌、甲状旁腺癌、阴茎癌、前列腺癌、唾液腺癌、皮肤癌、脾癌、睾丸癌、胸腺癌、甲状腺癌和子癌、静坐不能、阿尔茨海默症、健忘症、肌萎缩性侧索硬化、共济失调、双极性疾病、紧张症、大脑性麻痹、脑血管疾病、克-雅二氏病、痴呆、抑郁、唐氏综合征、迟发性运动障碍、张力障碍、癫痫、亨廷顿病、多发性硬化症、肌肉萎缩症、神经痛、神经纤维瘤、神经病、帕金森病、皮克病、色素性视网膜炎、精神分裂症、季节性情绪疾病、老年痴呆、中风、图雷特综合征和癌症——包括腺癌、黑素瘤和畸胎癌,特别是脑癌。

[0069] 在又一实施方式中,本方法可用于提供临床信息以检测和定量与基因或蛋白质表

达改变相关疾病的改变的基因结构、基因突变、基因生物化学修饰，包括信使RNA(mRNA)、核糖体RNA(rRNA)、转移RNA(tRNA)、微RNA(miRNA)、反义RNA(asRNA)及类似物的改变和/或修饰。这种疾病的实例包括，但不限于，癌症，如腺癌、白血病、淋巴瘤、黑素瘤、骨髓瘤、肉瘤、畸胎癌，和具体地，肾上腺癌、膀胱癌、骨癌、骨髓癌、脑癌、乳腺癌、子宫颈癌、胆囊癌、神经节癌、胃肠道癌、心脏癌、肾癌、肝癌、肺癌、肌癌、卵巢癌、胰腺癌、甲状旁腺癌、阴茎癌、前列腺癌、唾液腺癌、皮肤癌、脾癌、睾丸癌、胸腺癌、甲状腺癌和子宫癌；免疫疾病，如获得性免疫缺陷综合征(AIDS)、阿狄森病、成人呼吸窘迫综合征、过敏症、强直性脊柱炎、淀粉样变性病、贫血、哮喘、动脉粥样硬化、自身免疫性溶血性贫血、自身免疫性甲状腺炎、支气管炎、胆囊炎、接触性皮炎、克罗恩病、特应性皮炎、皮肌炎、糖尿病、气肿、淋巴细胞毒素的发作性淋巴细胞减少症、胎儿红细胞增多症、结节性红斑、萎缩性胃炎、肾小球性肾炎、古德帕斯彻综合征、痛风、格雷夫斯病、桥本甲状腺炎、嗜伊红细胞增多症、肠易激综合征、多发性硬化症、重症肌无力、心肌或心包炎症、骨关节炎、骨质疏松、胰腺炎、多发性肌炎、银屑病、莱特尔综合征、类风湿性关节炎、硬皮病、斯耶格伦综合征、全身过敏、全身性红斑狼疮、系统性硬化症、血小板减少性紫癜、溃疡性结肠炎、葡萄膜炎、维尔纳综合征、癌症并发症、血液透析和体外循环、病毒、细菌、真菌、寄生虫、原生动物和蠕虫感染、创伤、布鲁顿X-连锁无丙种球蛋白血症、常见变异型免疫缺陷(CVI)、迪乔治综合征(胸腺发育不全)、胸腺发育不良、隔离IgA缺乏症、重度联合免疫缺陷病(SCID)、血小板减少症和湿疹的免疫缺陷(威-奥氏综合征)、切东二氏综合征、慢性肉芽肿性疾病、遗传性血管神经性水肿和与库兴病相关的免疫缺陷；和发育疾病，如肾小管酸中毒、贫血、库兴综合征、软骨发育不全性侏儒、杜兴和贝克尔肌肉萎缩症、癫痫、性腺发育不全、WAGR综合征(威尔姆斯瘤、无虹膜、泌尿生殖系统异常和精神发育迟滞)、史密斯-马盖尼斯综合征、骨髓增生异常综合征、遗传性粘膜上皮异常增生、遗传性皮肤角化病、遗传性神经病如夏-马-图病和神经纤维瘤、甲状腺功能减退症、脑积水、癫痫病如Syndenham舞蹈病和大脑性麻痹、脊柱裂、无脑畸形、颅脊柱裂、先天性青光眼、白内障、感觉神经性听力损失以及与细胞生长和分化、胚胎发生和形态发生相关的任何疾病——涉及对象的任何组织、器官或系统，例如，脑、肾上腺、肾、骨骼或生殖系统。

[0070] 在另一实施方式中，本方法可用于提供临床信息以检测和定量基因或蛋白质表达改变相关疾病的改变的基因结构、基因突变、基因生物化学修饰，包括信使RNA(mRNA)、核糖体RNA(rRNA)、转移RNA(tRNA)、微RNA(miRNA)、反义RNA(asRNA)及类似物的改变和/或修饰。这种疾病的实例包括，但不限于，内分泌疾病，如与垂体功能减退相关的疾病，包括性腺功能减退、席汉综合征、尿崩症、卡尔曼病、汉-许-克三氏病、累-赛二氏病、结节病、空蝶鞍综合征和侏儒症；垂体功能亢进，包括肢端肥大症、巨人症和抗利尿激素(ADH)分泌异常综合征(SIADH)；和与甲状腺功能减退相关的疾病，包括甲状腺肿、粘液性水肿、与细菌感染相关的急性甲状腺炎、与病毒感染相关的亚急性甲状腺炎、自身免疫性甲状腺炎(桥本病)和呆小症；与甲状腺功能亢进相关的疾病，包括甲状腺毒症及其各种形式、格雷夫斯病、胫骨前粘液性水肿、毒性多结节性甲状腺肿、甲状腺癌和普鲁麦病；和与甲状旁腺功能亢进相关的疾病，包括康恩病(慢性高血钙)；呼吸系统疾病，如过敏、哮喘、急性和慢性炎性肺病、ARDS、气肿、肺充血和水肿、COPD、间质性肺病和肺癌；癌症，如腺癌、白血病、淋巴瘤、黑素瘤、骨髓瘤、肉瘤、畸胎癌，和具体地，肾上腺癌、膀胱癌、骨癌、骨髓癌、脑癌、乳腺癌、子宫颈癌、胆囊癌、神经节癌、胃肠道癌、心脏癌、肾癌、肝癌、肺癌、肌癌、卵巢癌、胰腺癌、甲状旁腺癌、阴茎

癌、前列腺癌、唾液腺癌、皮肤癌、脾癌、睾丸癌、胸腺癌、甲状腺癌和子宫癌;和免疫学疾病,如获得性免疫缺陷综合征(AIDS)、阿狄森病、成人呼吸窘迫综合征、过敏症、强直性脊柱炎、淀粉样变性病、贫血、哮喘、动脉粥样硬化、自身免疫性溶血性贫血、自身免疫性甲状腺炎、支气管炎、胆囊炎、接触性皮炎、克罗恩病、特应性皮炎、皮肌炎、糖尿病、气肿、淋巴细胞毒素的发作性淋巴细胞减少症、胎儿红细胞增多症、结节性红斑、萎缩性胃炎、肾小球性肾炎、古德帕斯彻综合征、痛风、格雷夫斯病、桥本甲状腺炎、嗜伊红细胞增多症、肠易激综合征、多发性硬化症、重症肌无力、心肌或心包炎症、骨关节炎、骨质疏松、胰腺炎、多发性肌炎、银屑病、莱特尔综合征、类风湿性关节炎、硬皮病、斯耶格伦综合征、全身过敏、全身性红斑狼疮、系统性硬化症、血小板减少性紫癜、溃疡性结肠炎、葡萄膜炎、维尔纳综合征、癌症并发症、血液透析和体外循环、病毒、细菌、真菌、寄生虫、原生动物和蠕虫感染和创伤。多核苷酸序列可用于DNA印迹分析或RNA印迹分析,点印迹或其他基于膜的技术;PCR技术;浸渍,点触(pin)和ELISA分析;和微阵列——其利用患者的流体或组织检测改变的核酸序列表达。这种定性或定量方法在本领域是公知的。

[0071] 发明特征和最佳实施方式

[0072] “BamBam”是计算有效的方法,用于测量大测序数据组,以产生一组高质量基因组事件,该高质量基因组事件存在于相对于其种系的各肿瘤中。这些结果提供对肿瘤染色体动态的扫视,提高对肿瘤最终状态及导致其事件的理解。

[0073] 诊断

[0074] 本文描述的方法可用于检测和定量与基因或蛋白质表达改变相关的状况、疾病或病症的改变的基因结构、基因突变、基因生物化学修饰,包括信使RNA(mRNA)、核糖体RNA(rRNA)、转移RNA(tRNA)、微RNA(miRNA)、反义RNA(asRNA)及类似物的改变和/或修饰。本文描述的方法还可用于检测和定量改变的基因表达、mRNA表达的不存在/存在相对于过度、或用于在治疗干预过程中监测mRNA水平。与表达改变相关的状况、疾病或病症包括特发性肺动脉高压、继发性肺动脉高压、细胞增殖性疾病,特别是间变性少突神经胶质瘤、星形细胞瘤、少突星形细胞瘤、成胶质细胞瘤、脑膜瘤、神经节细胞瘤、神经元肿瘤、多发性硬化症、亨廷顿病、乳腺癌、前列腺癌、胃腺癌、转移性神经内分泌癌、非增殖性纤维囊性和增殖性纤维囊性乳腺疾病、胆囊炎和胆石病、骨关节炎和类风湿性关节炎;获得性免疫缺陷综合征(AIDS)、阿狄森病、成人呼吸窘迫综合征、过敏症、强直性脊柱炎、淀粉样变性病、贫血、哮喘、动脉粥样硬化、自身免疫性溶血性贫血、自身免疫性甲状腺炎、良性前列腺增生症、支气管炎、切东二氏综合征、胆囊炎、克罗恩病、特应性皮炎、皮肌炎、糖尿病、气肿、胎儿红细胞增多症、结节性红斑、萎缩性胃炎、肾小球性肾炎、古德帕斯彻综合征、痛风、慢性肉芽肿性疾病、格雷夫斯病、桥本甲状腺炎、嗜伊红细胞增多症、肠易激综合征、多发性硬化症、重症肌无力、心肌或心包炎症、骨关节炎、骨质疏松、胰腺炎、多囊卵巢综合征、多发性肌炎、银屑病、莱特尔综合征、类风湿性关节炎、硬皮病、重度联合免疫缺陷病(SCID)、斯耶格伦综合征、全身过敏、全身性红斑狼疮、系统性硬化症、血小板减少性紫癜、溃疡性结肠炎、葡萄膜炎、维尔纳综合征、血液透析、体外循环、病毒、细菌、真菌、寄生虫、原生动物和蠕虫感染;催乳素生成疾病、不育,包括输卵管疾病、排卵缺陷和子宫内膜异位、动情周期中断、月经周期中断、多囊卵巢综合征、卵巢过度刺激综合征、子宫内膜或卵巢肿瘤、子宫肌瘤、自身免疫性疾病、子宫外孕和畸形发生;乳腺癌、纤维囊性乳腺疾病和乳溢;精子发生中断、精子生理异

常、良性前列腺增生症、前列腺炎、Peyronie病、性无能、男子女性型乳房;光化性角膜炎、动脉硬化、滑囊炎、硬变、肝炎、混合性结缔组织疾病(MCTD)、骨髓纤维化、阵发性睡眠性血红蛋白尿、真性红细胞增多症、原发性血小板增多症、癌症并发症、癌症——包括腺癌、白血病、淋巴瘤、黑素瘤、骨髓瘤、肉瘤、畸胎癌,和具体地,肾上腺癌、膀胱癌、骨癌、骨髓癌、脑癌、乳腺癌、子宫颈癌、胆囊癌、神经节癌、胃肠道癌、心脏癌、肾癌、肝癌、肺癌、肌癌、卵巢癌、胰腺癌、甲状旁腺癌、阴茎癌、前列腺癌、唾液腺癌、皮肤癌、脾癌、睾丸癌、胸腺癌、甲状腺癌和子宫癌。另一方面,本发明的核酸。

[0075] 本文描述的方法可用于检测和定量与基因或蛋白质表达改变相关的疾病的改变的基因结构、基因突变、基因生物化学修饰,包括信使RNA(mRNA)、核糖体RNA(rRNA)、转移RNA(tRNA)、微RNA(miRNA)、反义RNA(asRNA)及类似物的改变和/或修饰。本文描述的方法还可用于检测和定量改变的基因表达;mRNA的表达不存在、存在或过度;或用于在治疗干预过程中监测mRNA水平。与表达改变相关的疾病包括静坐不能、阿尔茨海默症、健忘症、肌萎缩性侧索硬化、共济失调、双极性疾病、紧张症、大脑性麻痹、脑血管疾病、克-雅二氏病、痴呆、抑郁、唐氏综合征、迟发性运动障碍、张力障碍、癫痫、亨廷顿病、多发性硬化症、肌肉萎缩症、神经痛、神经纤维瘤、神经病、帕金森病、皮克病、色素性视网膜炎、精神分裂症、季节性情绪疾病、老年痴呆、中风、图雷特综合征和癌症——包括腺癌、黑素瘤和畸胎癌,特别是脑癌。

[0076] 为提供与基因表达相关的状况、疾病或病症的诊断依据,建立了正常或标准表达概况。这可通过在杂交或扩增条件下将用探针从正常对象——动物或人类提取生物样本组合而实现。标准的杂交可通过将利用正常对象获得的值与实验值进行比较而被定量,该实验采用已知量的基本上纯化的目标序列。可将这种方式下获得的标准值与得自症状显示为特定状况、疾病或病症的患者的样本的值进行比较。利用标准值与特定状况相关的值的偏差来诊断该状况。

[0077] 这种分析还可用于评价动物研究和临床试验中具体治疗性处理方案的效力,或监测个体患者的治疗。在状况的存在确立并且治疗方案启动后,可定期反复进行诊断分析,以确定患者体内的表达水平是否开始接近正常对象中观察的水平。该分析还可用于检测、定量或测量指示和/或鉴定肿瘤存在、肿瘤不存在或进行临床处理或治疗的个体的缓解状态的基因结构、基因突变、基因生物化学修饰,包括对信使RNA(mRNA)、核糖体RNA(rRNA)、转移RNA(tRNA)、微RNA(miRNA)、反义RNA(asRNA)及类似物的改变和/或修饰。由连续的分析获得的结果可用于显示数天至数月范围时间的治疗效力。

[0078] 本文公开的方法还可用于检测、定量和关联之前未被鉴定或关联于特定临床疾病、病症或状况的基因结构、基因突变、基因生物化学修饰的变化,包括信使RNA(mRNA)、核糖体RNA(rRNA)、转移RNA(tRNA)、微RNA(miRNA)、反义RNA(asRNA)及类似物的改变和/或修饰。在可选方案中,本文公开的方法可用于鉴定新的临床疾病、病症或状况。然后,可将基因结构、基因突变和基因生物化学修饰的新变化与核酸序列或蛋白质序列的已知化学和生物化学性质进行比较,并可利用与临床疾病、病症或病症相关的上述改变生成关于细胞代谢的新数据库和认识,用于临床应用。

[0079] 模型系统

[0080] 动物模型可被用作生物分析,此时其呈现与人类类似的毒性响应,并且其中暴露

条件是与人类暴露相关的。哺乳动物是最常见的模型，并且大多数毒性研究是对啮齿动物如大鼠或小鼠进行的，这是因为低成本、可用性和充足的参考毒理学。啮齿动物近交品系提供用于研究目的基因表达不足或过表达的生理结果和发展疾病诊断和治疗方法的便利模型。过表达特定基因(例如，分泌在乳汁中)的哺乳动物近交品系还可充当由该基因表达的蛋白质的便利来源。

[0081] 毒理学

[0082] 毒理学是试剂对活系统的影响的研究。多数毒性研究对大鼠或小鼠进行，以有助于预期这些试剂对人类健康的影响。生理、行为、稳态过程和致死性的定性和定量变化的观察被用于生成毒性概况和评估在暴露于试剂后对人类健康的影响。

[0083] 遗传毒理学鉴定和分析产生遗传突变的试剂的能力。遗传毒性试剂通常具有有助于与核酸相互作用的共同化学或物理性质，并且在染色体异常传给后代时最为有害。毒理学研究可鉴定增加后代结构或功能异常性频率的试剂——如果在受孕前给予任一亲代、在妊娠期间给予母体或给予发育生物体。小鼠和大鼠最常用于这些测试，因为其繁殖周期短，产生符合统计学要求所需的生物体数量。

[0084] 急性毒性测试基于将试剂单次给予对象以确定试剂的症状学或致死性。进行三个实验：(a)初始剂量范围调查实验、(b)缩窄有效剂量范围的实验和(c)建立剂量响应曲线的最终实验。

[0085] 长期毒性测试基于反复给予试剂。大鼠和狗常用于这些研究，以提供不同种家族物种的数据。除致癌作用外，相当多的证据证明以高剂量浓度每日给予试剂三至四个月的时间将揭示成年动物毒性的大多数形式。

[0086] 利用持续一年或更长时间的慢性毒性测试来证明试剂不存在毒性或具有致癌可能性。在对大鼠进行研究时，应用最少三个测试组加一个对照组，并且在实验开始和在整个实验过程中每隔一段时间检查和监测动物。

[0087] 转基因动物模型

[0088] 过表达目的基因或目的基因表达不足的转基因啮齿动物可以是近交的，并用于模拟人类疾病或测试治疗剂或毒性剂。(参见美国专利号4,736,866;5,175,383;和5,767,337;引入本文作为参考。)在一些情况下，引入的基因可在胎儿发育或出生后在特异组织类型中、于特异时间被活化。转基因的表达通过如下得到监测：在用实验药物治疗进行挑战之前、之中和之后分析转基因动物的表型或组织特异性mRNA表达。

[0089] 胚胎干细胞

[0090] 从啮齿动物胚胎分离的胚胎干细胞(ES)保持形成胚胎的潜力。当将ES细胞置于载体胚胎中时，其恢复正常发育，并有助于活的出生动物的全部组织。ES细胞是用于生成实验敲除和敲入啮齿动物品系的优选细胞。小鼠ES细胞，如小鼠129/SvJ细胞系，得自早期小鼠胚胎，并在本领域公知的培养条件下生长。敲除品系的载体包含候选疾病基因，该疾病基因候选体被修饰以包括标记基因，该标记基因中断体内转录和/或翻译。载体通过转化方法如电穿孔、脂质体递送、微注射及本领域公知的类似方法被引入ES细胞。内源啮齿动物基因在细胞分裂期间通过同源重组和整合被中断的疾病基因取代。转化的ES细胞被鉴定，并优选被微注入小鼠细胞胚泡，如来自C57BL/6小鼠品系的小鼠细胞胚泡。胚泡被外科转移至假孕雌亲，并将所得的嵌合后代进行基因分型和繁殖，以生成杂合或纯合品系。

[0091] ES细胞还被用于研究各种细胞类型和组织的体外分化，如神经细胞、造血谱系和心肌细胞(Bain et al.(1995)Dev.Biol.168:342-357;Wiles and Keller(1991) Development 11 1:259-267;和Klug et al.(1996)J.Clin.Invest.98:216-224)。近期的发展证明，得自人胚细胞的ES细胞还可在体外操作，分化成为八个单独的细胞系，包括内胚层、中胚层和外胚层细胞类型(Thomson(1998)Science 282:1145-1147)。

[0092] 敲除分析

[0093] 在基因敲除分析中，候选人类疾病基因区域经酶修饰包括非哺乳动物基因，如新霉素磷酸转移酶基因(neo;参见，例如，Capecchi(1989)Science 244:1288-1292)。插入的编码序列中断目标基因的转录和翻译，并阻止疾病候选蛋白质的生物化学合成。修饰的基因被转化到培养的胚胎干细胞(上文所述)中，转化的细胞被注入啮齿动物囊胚，并且囊胚被植入假孕雌亲。转基因后代经杂交获得纯合近交系。

[0094] 敲入分析

[0095] 全能ES细胞，存在于胚胎发育早期，可被用于生成人类疾病的敲入型人源化动物模型(猪)或转基因动物模型(小鼠或大鼠)。在敲入技术下，人类基因区域被注入动物ES细胞，并且人类序列通过重组整合到动物细胞基因组中。包含整合的人类基因的全能ES细胞被如上所述处理。对近交动物进行研究和处理，以获得关于类似的人类状况的信息。这些方法已被用于模拟数种人类疾病。(参见，例如，Lee et al.(1998)Proc.Natl.Acad.Sci.95: 11371-11376;Baudoin et al.(1998)Genes Dev.12:1202-1216;和Zhuang et al.(1998) Mol.Cell Biol.18:3340-3349)。

[0096] 非人类灵长类动物模型

[0097] 动物测试领域处理基础科学如生理学、遗传学、化学、药理学和统计学的数据和方法。这些数据在评价治疗剂对非人类灵长类动物作用中至关重要，因为其可能与人类的健康相关联。在疫苗和药物评价中猴被用作人类替代品，并且其响应与人类暴露在类似情况下是相关的。猕猴(食蟹猕猴(Macaca fascicularis)、恒河猕猴(Macaca mulata))和普通狨猴(Callithrix jacchus)是用于这些研究的最常见的非人类灵长类动物(NHP)。由于高成本与建立和维持NHP群体有关，早期研究和毒理学研究通常在啮齿动物模型中进行。在应用行为测量如药物成瘾的研究中，NHP是测试动物第一选择。此外，NHP和个人对多种药物和毒素呈现不同的敏感性，并且可被分成这些试剂的“广代谢体”和“弱代谢体”。

[0098] 发明的示例性应用

[0099] 个性化药物保证向最可能获益的那些患者递送特定治疗(一种或多种)。我们已显示，约一半的治疗性化合物在一种或多种临床相关的转录或基因组乳腺癌亚型中优先有效。这些发现支持确定响应相关分子亚型在乳腺癌治疗中的重要性。我们还显示，关于细胞系的转录和基因组数据的途径整合揭示了子网络，其为观察到的亚型特异性响应提供机理解释。细胞系与肿瘤之间子网络活性的比较分析显示，多数亚型特异性子网络在细胞系与肿瘤之间保留。这些分析支持如下观点：临床前在充分表征的细胞系小组中筛选实验化合物能够鉴定候选的响应相关分子标志，该候选的响应相关分子标志能够用于早期临床试验中的敏感性富集。我们提出，这种体外评估方法将增加在化合物临床开发开始前鉴定到响应性肿瘤亚型的似然(likelihood)，从而降低成本，增加最终FDA批准的可能性，和有可能避免与治疗不可能响应的患者相关的毒性。在本研究中，我们仅已评估限定转录亚型的分

子标志和所选的再现基因组拷贝数异常(CNA)。我们预期,本方法的能力和精确性将随着分析中包括额外的分子特征如遗传突变、甲基化和可选的剪接而增加。同样,增加细胞系小组的大小将增加评估小组内较不常见的分子样式的能力和增加代表人类乳腺癌中存在的多样性的更完整范围的概率。

[0100] 在此,我公开了新的软件工具,我们称其为BamBam,其能够快速比较肿瘤(体细胞)与种系匹配的测序数据组。BamBam输出的结果不同,产生各患者样本包含的体细胞和种系变体的详尽目录。该目录为研究人员提供了快速发现肿瘤发展过程中发生的重要变化的能力,还提供了患者种系中存在的可指示疾病易患性的高质量变体。BamBam的进一步改进将由具体搜索相同的基因组区域中存在的可指出肿瘤发生的驱动因子的多种类型的变体(例如,基因的一个等位基因缺失,另一等位基因包含断点的截短突变)的方法组成。我们还计划扩展BamBam管线(pipeline)的能力。

[0101] 在另外的实施方式中,多核苷酸核酸可用于待开发的任何分子生物技术,只要新技术依赖于当前已知的核酸分子的性质,包括但不限于,诸如三倍体遗传密码和具体碱基对的相互作用的性质。

[0102] 参考下面的实施例,本发明将更容易被理解,该实施例被包括在内仅为示例本发明的某些方面和实施方式,而非限制。

[0103] 实施例

[0104] 实施例I:通过参考基因组进行的数据组同步化

[0105] 将全部短阅读均与相同的参考基因组进行比对,使参考基因组成为由多个相关的样本组织序列数据的自然方式。BamBam接收两个短阅读测序数据组——一个来自肿瘤,另一个是来自相同患者的匹配正常基因(“种系”)和参考基因组,并读取这些数据组,使得两数据组中重叠相同基因组位置的全部序列可用于同时处理。这是处理这种数据的最有效方法,同时还能够进行复杂分析,该分析将难以或不能以顺序方式实现,在此各数据组被单独处理,结果仅在之后组合。

[0106] 这种方法容易被扩展至两个以上的相关测序数据组。例如,如果将三个样本——匹配的正常样本、肿瘤样本和复发样本——测序,则本方法可用于搜索针对肿瘤&复发样本特异的变化和仅针对复发特异的变化,这表明复发肿瘤已由其据推测衍生来源的原肿瘤略微发生变化。而且,可应用这种相同的方法确定儿童基因组的遗传部分,假设测序样本来自儿童、父亲和母亲。

[0107] 实施例II:体细胞和种系变体呼叫

[0108] 由于BamBam保持整个同步基因组中的序列数据同时处于成对文件中,可容易实施需要来自肿瘤和种系BAM文件以及人类参考的测序数据的复杂突变模型。该模型旨在最大化种系基因型(假设种系阅读和参考核苷酸)和肿瘤基因型(假设种系基因型、简单突变模型、肿瘤样本中污染正常组织的分数的评估和肿瘤序列数据)的联合概率(joint probability)。

[0109] 为找到最佳的肿瘤和种系基因型,我们旨在最大化如下限定的似然

[0110] $P(D_g, D_t, G_g, G_t | \alpha, r)$

[0111] $= P(D_g | G_g) P(G_g | r) P(D_t | G_g, G_t, \alpha) P(G_t | G_g)$

[0112] 其中r是观察的参考等位基因,α是正常污染的分数,并且肿瘤和种系基因型由Gt

＝(t1,t2)和Gg＝(g1,g2)限定,其中t1,t2,g1,g2∈{A,T,C,G}。肿瘤和种系序列数据分别由如下阅读组限定：$D_t = \{d_t^1, d_t^2, \ldots, d_t^m\}$和$D_g = \{d_g^1, d_g^2, \ldots, d_g^n\}$,并且观察到的碱基$d_t^i, d_g^i \in \{A, T, C, G\}$。模型中所用的全部数据均必需超过用户限定的碱基,并映射质量阈值(mapping quality threshold)。

[0113] 种系等位基因——假设种系基因型——的概率被模拟为基于四种核苷酸的多项式：

[0114] $$P(D_g|G_g) = \frac{n!}{n_A!n_T!n_G!n_C!}\prod_i^n P(d_g^i|G_g),$$

[0115] 其中n是该位置种系阅读的总数,n_A,n_G,n_c,n_T是支持各观察到的等位基因的阅读。碱基概率$P(d_g^i|G_g)$被假设是独立的,来自由基因型G_g表示的两种亲代等位基因中任一种,同时还包括测序仪的近似碱基错误率。关于种系基因型的先验(prior)基于参考碱基被条件化为

[0116] $P(G_g|r=a) = \{\mu_{aa}:\mu_{ab},\mu_{bb}\}$,

[0117] 其中μ_{aa}是该位置作为纯合参考的概率,μ_{ab}是杂合参考,并且μ_{bb}是纯合非参考。此时,种系先验不包括关于已知的遗传SNP的任何信息。

[0118] 肿瘤阅读组的概率再次被限定为多项式

[0119] $$P(D_t|G_t, G_g, \alpha) = \frac{n!}{n_A!n_T!n_G!n_C!}\prod_i^n P(d_t^i|G_t, G_g, \alpha),$$

[0120] 其中m是该位置种系阅读的总数,m_A,m_G,m_c,m_T是支持肿瘤数据组中各观察到的等位基因的阅读,并且各肿瘤阅读的概率是得自肿瘤和种系基因型的碱基概率的组合,其受控于正常污染的分数α,为

[0121] $$P(d_t^i|G_t, G_g, \alpha) = \alpha P(d_t^i|G_t) + (1-\alpha)P(d_t^i|G_g)$$

[0122] 并且肿瘤基因型的概率由种系基因型的简单突变模型限定

[0123] $P(G_t|G_g) = \max[P(t_1|g_1)P(t_2|g_2), P(t_1|g_2)P(t_2|g_1)]$,

[0124] 其中无突变的概率(例如,t1＝g1)是最大的,并且转换(即,A→G.T→c)的概率比颠换(即,A→T,T→G)的概率可能高四倍。多项分布的所有模型参数α,μ_{aa},μ_{ab},μ_{bb}和碱基概率,$P(d^i|G)$可由用户设定。

[0125] 选定的肿瘤和种系基因型G_t^{max},G_g^{max}是最大化(1)的肿瘤和种系基因型,并且后验概率——由如下限定：

[0126] $$\frac{P(D_g, D_t, G_g^{max}, G_t^{max}|\alpha, r)}{\sum_{i,j} P(D_g, D_t, G_g = i, G_t = j|\alpha, r)}$$

[0127] 可用于评定成对推断基因型的可信度。如果肿瘤和种系基因型不同,则推定的体细胞突变(一个或多个)将会连同其各自的可信度被报告。

[0128] 最大化肿瘤和种系基因型的联合似然(joint likelihood)有助于提高推断基因型的准确性,特别是在一个或两个序列数据组具有低覆盖的特定基因组位置的情况下。其

他突变呼叫算法,如分析单个测序数据组的MAQ和SNVMix,在非参考或突变体等位基因具有低支持时更有可能产生错误(Li,H.,et al.(2008)Mapping short DNA sequencing reads and calling variants using mapping quality scores,Genome Research,11,1851-1858;Goya,R.et al.(2010)SNVMix:predicting single nucleotide variants from next-generation sequencing of tumors,Bioinformatics,26,730-736)。

[0129] 除由给定基因组位置处的全部阅读收集等位基因支持外,还收集关于阅读的信息(如其使读取图滞留于、前进至或倒退至阅读中的等位基因位置,等位基因平均质量,等),并将其用于选择性滤出假阳性呼叫。我们预期,支持变体的所有等位基因的链和等位基因位置随机分配,并且如果分配显著偏离此随机分配(即,发现所有变体等位基因接近阅读尾部),则这表明变体呼叫是可疑的。

[0130] 实施例III:全部拷贝数和等位基因特异的拷贝数

[0131] 利用动态窗口显示方法计算全部体细胞拷贝数,该动态窗口显示方法根据肿瘤或种系数据的覆盖扩大或缩小窗口基因组宽度。该方法以零宽度的窗口初始。肿瘤或种系序列数据的每个单独阅读将记录为肿瘤计数Nt或种系计数Ng。各阅读的开始和终止位置将限定窗口区域,该窗口区域在新阅读超过现有窗口的界限时扩大。在肿瘤或种系计数超过用户限定阈时,记录窗口的尺寸和位置,以及Nt、Ng和相对覆盖度Nt。根据局部读取覆盖调整Ng窗口尺寸将产生低覆盖区域(例如,重复区域)大窗口或显示体细胞扩增区域的小窗口,从而增加扩增子的基因组分辨率和增加我们限定扩增界限的能力。

[0132] 类似地计算等位基因特异性拷贝数——除仅包括认为是种系杂合的位置外,如示(参见图2)。杂合性被限定为在种系中被认为具有两个不同的等位基因的位置,每个亲代贡献一个等位基因。利用相同的动态窗口显示技术——上文所述用于全部拷贝数,计算多数和少数拷贝数,从而汇集相同基因组附近的数据。杂合位点的多数等位基因在本文中被限定为这样的等位基因:其在肿瘤数据组中具有最大数量的重叠该基因组位置的支持阅读,而少数等位基因是具有最少支持的等位基因。肿瘤和种系数据中归因于多数等位基因的所有计数均将进行多数拷贝数计算,少数等位基因同样也是。然后通过种系数据Ng中两种等位基因的计数,标准化多数和少数等位基因的计数,从而计算多数和少数拷贝数。

[0133] 利用等位基因特异性拷贝数鉴定显示杂合性丢失(拷贝中性和拷贝损失)以及单个等位基因特异性扩增或缺失的基因组区域。最后这点对于帮助将引起疾病的等位基因可能地区分为在肿瘤序列数据中扩增或未缺失的等位基因尤为重要。此外,经受半合损失的区域(例如,一个亲代染色体臂)可用于直接评估测序肿瘤样本中正常污染物量,其可用于提高上述种系和肿瘤基因型的模拟。

[0134] 图2显示等位基因特异的拷贝数计算的概括。利用种系和肿瘤测序数据确定杂合基因型的位置,如通过种系变体呼叫算法确定。收集所有重叠这些位置的阅读,并且在肿瘤和种系中发现杂合基因型两个等位基因中每一个的阅读支持。多数等位基因被确定为具有最高支持的等位基因,并且通过由种系中该位置的阅读总数标准化该计数来计算多数拷贝数。

[0135] 实施例IV:基因型分阶

[0136] BamBam试图通过利用肿瘤中大规模的基因组扩增或缺失所引起的等位基因失衡,使在种系中发现的所有杂合位置分阶。在肿瘤序列数据的每个位置选择多数投票基础呼叫

(vote base call),从而构建肿瘤中存在的分阶的单倍型。多数投票选择短阅读库中所观察到的数量最多的等位基因,其应选择在缺失事件后仍在肿瘤中的等位基因或扩增事件的复制等位基因。还鉴定各个位置上种系的等位基因状态,在此如果仅存在一个具有所需阅读支持的等位基因,则认为是纯合位置,如果至少两个等位基因具有所需阅读支持则认为是杂合位置。假设肿瘤的单倍型代表两个亲代单倍型其中之一,在此得到第二亲代单倍型作为不属于肿瘤单倍型的种系等位基因序列。此程序在基因组范围被应用,而与肿瘤中的等位基因比例无关,因此我们预期将在多数和少数等位基因之间同样平衡的区域中基本上随机的基因型的单倍型分配。种系序列的准确分解将仅存在于这样的区域:显示一致的等位基因失衡,该等位基因失衡是由于肿瘤中的单个基因组事件(例如,区域扩增或缺失)。

[0137] 肿瘤衍生的单倍型的确定可通过比较肿瘤衍生的单倍型与得自HapMap项目(International HapMap Consortium(2007),Nature,7164:851-861)的分阶的基因型实现。

[0138] 实施例V:利用成对末端聚簇推断结构变化

[0139] 为鉴定推定的染色体内和染色体间重排,BamBam搜索不一致的成对阅读,在此,配对中的各阅读映射参考序列的离散区域。染色体内不一致的配对是具有异常大插入尺寸的配对(即,参考序列上分隔成对阅读的基因组距离超过用户限定阈值)或以不正确定向映射(即倒位)的配对。染色体间不一致的配对由映射不同染色体的成对阅读限定。与其他配对比对相同位置的所有不一致的成对末端阅读被去除,以避免仅由源自短阅读库制备中PCR扩增步骤的大量阅读支持的呼叫重排。该过程的概括显示在图3中。

[0140] 所有不一致的成对末端阅读按照其基因组位置进行聚簇,以限定近似的基因组区域,其中断点被认为存在于此。聚集过程由如下组成:将与推定的断点两侧的其他阅读重叠的单独阅读分组在一起。所有重叠阅读的链定向还必须匹配配对簇或不被包括在配对簇中。当簇中重叠的不一致配对的数超过用户限定阈时,限定描述重排的断点。如果重排存在于种系和肿瘤数据组的相同位置时,则如下将其进行比较。种系重排要求,肿瘤和种系数据组支持相同的重排,这是因为在种系中观察到的结构变化在肿瘤中以某种方式被逆转从而精确地符合参考序列,是非常不可能的。另一方面,体细胞重排必须仅在肿瘤测序数据中被观察到,并且基本上不存在于种系数据组中。满足这些要求的重排被存储用于后处理分析和可视化,而不满足这些要求的重排被舍弃,作为测序仪器、样本制备(如全基因组扩增)或所用短阅读映像算法的系统性偏差造成的人造重排。

[0141] 图3显示结构变化呼叫的概括。推定的结构变体的最初鉴定是通过BamBam利用不一致映射的阅读对确定的,在此两阅读完全映射参考基因组,但是以异常的非参考方式。然后通过被称为bridget的程序、利用任何可用的拆分阅读,完善由BamBam发现的推定的断点。

[0142] 实施例VI:利用拆分阅读(split read)完善结构变化

[0143] BamBam最初发现的断点是近似的,这是因为其采用完全映射阅读,完全映射阅读其本质上不能重叠断点的实际接合处,因为其表示参考序列(或种系数据组,在体细胞重排的情况下)中不存在的序列。为完善我们对断点位置的了解,开发了被称为Bridget的程序,其被概述在图4中。

[0144] Bridget被给予由BamBam发现的近似断点,并通过完全映射配对(mate)搜索锚定

在推定的断点附近的所有未比对的阅读。这些未映射的阅读中的每一个均具有成为“拆分阅读”的潜力，该“拆分阅读”与重排断点接合处重叠。断点两侧周围局部化的基因组序列被拆分成一组独特的片段(tile)(目前片段尺寸＝16bp)，并且建立片段序列及其在参考基因组中的位置的片段数据库。通过将阅读拆分为相同尺寸的片段和在阅读中标注其位置，对每个未比对的阅读构建类似的片段数据库。将参考片段数据库与未比对片段数据库进行比较，确定各未比对片段在参考中的基因组位置。通过确定在参考阅读和未比对阅读——断点一侧一个——中连续的的最大组片段，计算这些位置的“双生成集(Dual spanning set)”。

[0145] 参考坐标中“双生成集”的最小和最大基因组位置精确地确定了断点的位置以及序列的定向(或链型(strandedness))。在具有描述断点左侧边限和右侧边限的信息的情况下，重排的序列被完全限定，即，左侧被(染色体＝chr1，位置＝1000bp，链＝正向)限定，右侧被(染色体＝chr5，位置＝500,000bp，链＝反向)限定。断点的序列同源性(即，短序列，如“CA”被观察到在断点两个边限上是一致的，但仅在两序列的接合处比对的阅读中被观察到一次)也由这些双生成集确定。

[0146] 对于每个未比对的阅读，双生成集确定可能的断点位置。由于各未比对的阅读可确定略微不同的断点位置(因为断点附近的序列错误、重复参考等)，利用所有由双生成集确定的断点位置来生成可能的接合序列。将所有未映射的阅读与这些可能的接合序列中的每一个重新比对，并且相对于阅读如何与原序列完美比对来测量其比对的总体提高。导致比对分数最大提高的接合序列被评为真重排的最佳候选。如果此最佳接合序列导致比对分数极少至无提高，则此接合序列被舍弃，因为其不能表示真重排。在这种情况下，还可确定，拆分阅读确认的不存在是证据，证明由BamBam发现的原始结构重排可能是人造的。

[0147] 图4显示精确地鉴定基因组中发生结构重排的位置的示例性方法。确定可能的拆分阅读和参考基因组的片段(或kmers)。确定双生成集(表示为该图底部的深红色和紫色框)，其完全限定如何构建重排序列。双生成集对于拆分阅读中序列错误或SNP是强力的。

[0148] 实施例VII：肿瘤特异性基因组浏览器

[0149] 为可视化BamBam输出的所有结果，开发了肿瘤基因组浏览器，其同时显示在单个肿瘤样本中发现的所有基因组变体——相对于其匹配的正常基因组，如图5所示。其能够显示全部&等位基因特异的拷贝数、染色体内和染色体间重排和突变以及小插入/缺失。其以线性和环形图显示数据，后者远明显更适于显示染色体间重排。

[0150] 通过在单个图像中一起显示数据，用户可快速浏览单个样本的数据，并了解拷贝数变化和结构变化之间的关系。例如，大型的染色体内缺失类型的重排在断点之间的区域应具有一致的拷贝数下降。而且，用拷贝数数据显示突变数据使用户能够了解体细胞突变是否随后被扩增或野生型等位基因是否在肿瘤中缺失，两种重要的数据点均表明在此样本的肿瘤发生中基因组位点的重要性。

[0151] 图5显示示例性肿瘤特异性基因组浏览器。该浏览器在单个图像中显示通过BamBam发现的所有高水平体细胞差异，使得能够合成多个不同的数据组以给出肿瘤基因组的全部图片。该浏览器能够快速放大和缩小基因组区域，如上所示，仅以若干次点击由完整的基因组视图变成单碱基分辨率。

[0152] 实施例VIII：计算要求

[0153] BamBam和Bridget均以C编写，仅需要标准C库和最新的SAM工具源代码(可得自http://samtools.sourceforge.net)。其可作为单个过程运行，或在整个簇中拆分成一系列工作(job)(例如，每条染色体一个工作)。处理各包含数十亿个100bp阅读的成对250GB BAM文件，BamBam将在约5小时内以单个过程完成其全基因组分析，或在约30分钟内基于适度的簇(24个节点)完成其全基因组分析。BamBam的计算要求可被忽略，仅需要足够的RAM以存储与单个基因组位置重叠的阅读数据和足够的盘空间以存储在肿瘤或种系基因组中发现的被充分支持的变体。

[0154] Bridget也具有非常适度的计算要求。在单个机器上的运行时间一般小于1秒，其包括集合参考序列和断点附近任何潜在的拆分阅读、建立参考和拆分阅读的片段数据库、确定所有双生成集、构建潜在的接合序列、将所有拆分阅读与参考序列和各接合序列重新比对、以及确定最佳接合序列所必需的时间。高度扩增的或具有大量未映射阅读的区域增加Bridget的运行时间，但这可通过Bridget的易于平行性被缓解。

[0155] 实施例IX:基因组DNA的分离

[0156] 从患者收集血液或其他组织样本(2-3ml)，并将其在-80℃下储存在含EDTA的管中，备用。按照制造商的说明书(PUREGENE,Gentra Systems,Minneapolis MN)，利用DNA分离试剂盒，从血液样本提取基因组DNA。测量DNA纯度，为260和280nm下的吸光比(1cm光路；A_{260}/A_{280})，用Beckman分光光度计测量。

[0157] 实施例X:SNP的鉴定

[0158] 通过PCR，利用为该区域特异设计的引物扩增患者DNA样本的基因区域。利用本领域技术人员公知的方法测序PCR产物，如上所述。利用Phred/Phrap/Consed软件验证在序列轨迹中鉴定的SNP，并将其与NCBI SNP数据库中存储的已知SNP进行比较。

[0159] 实施例XI:统计学分析

[0160] 值被表示为平均值±SD.x^2分析(Web Chi平方计算器,Georgetown Linguistics,Georgetown University,Washington DC)被用于评估正常对象与疾病患者的基因型频率之间的差异。如所示地，进行兼带事后分析(post-hoc analysis)的单向ANOVA，以比较不同患者组之间的血液动力学。

[0161] 本领域技术人员将理解，上述实施方式的多种改动和修改可被配置而不脱离本发明的范围和精神。本领域已知的其他适当的技术和方法可以多种具体方式被本领域技术人员根据本文所述的本发明描述而应用。因此，要理解的是，本发明可除本文具体描述以外进行实践。上文描述意为示例性的，而非限制性的。基于阅读上文描述，多种其他实施方式将对本领域技术人员而言是显而易见的。因此，本发明的范围应参考所附权利要求以及该权利要求应得的全部等同范围而确定。

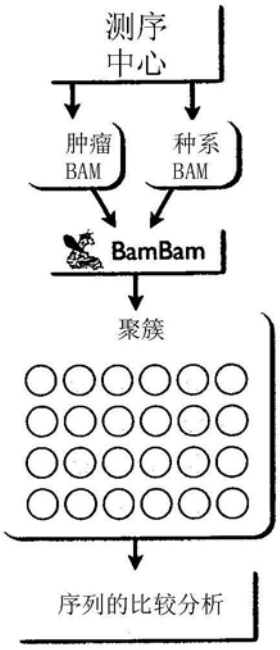

测序
中心
肿瘤
BAM
种系
BAM
BamBam
聚簇
序列的比较分析

图1

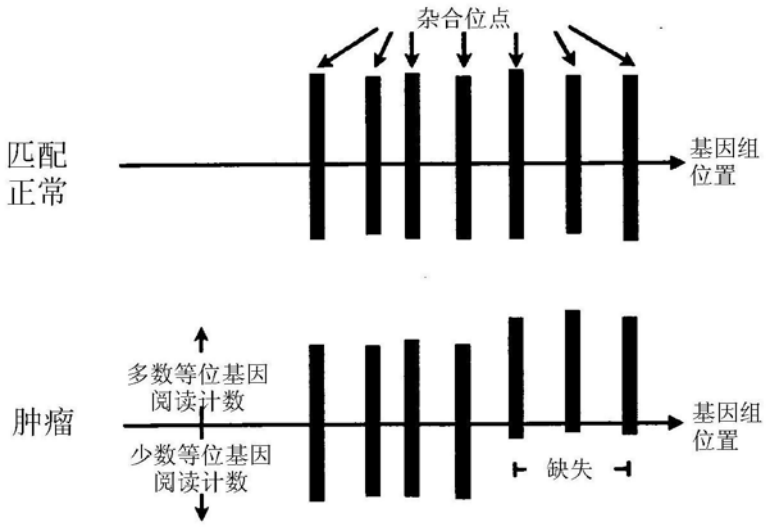

杂合位点
匹配
正常
基因组
位置
肿瘤
多数等位基因
阅读计数
少数等位基因
阅读计数
缺失
基因组
位置

图2

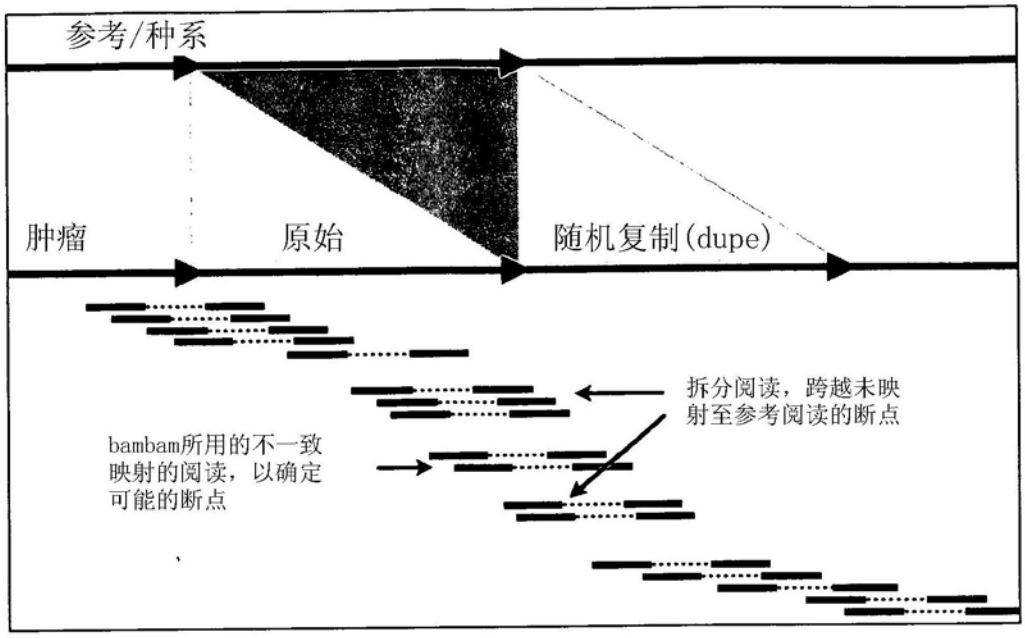

参考/种系
肿瘤
原始
随机复制(dupe)
拆分阅读，跨越未映射至参考阅读的断点
bambam所用的不一致映射的阅读，以确定可能的断点

图3

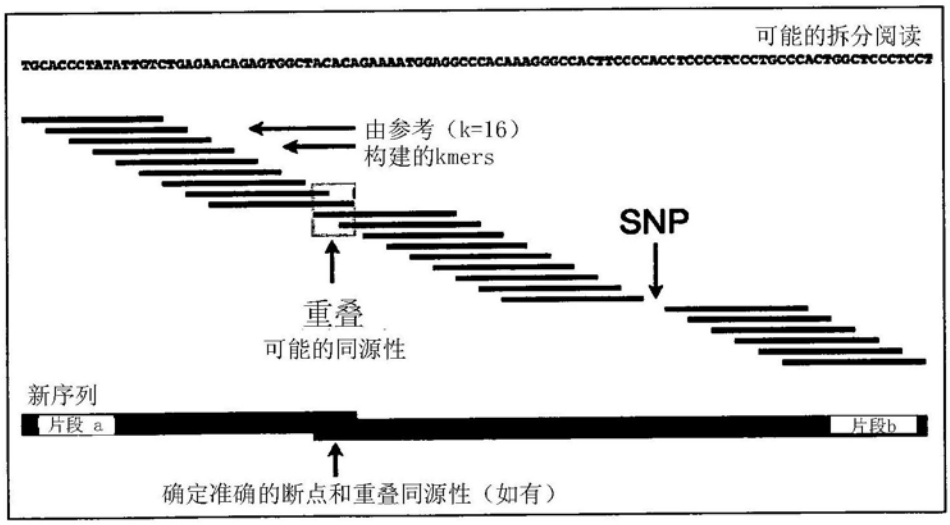

可能的拆分阅读
由参考（k=16）构建的kmers
SNP
重叠
可能的同源性
新序列
片段 a
片段b
确定准确的断点和重叠同源性（如有）

图4

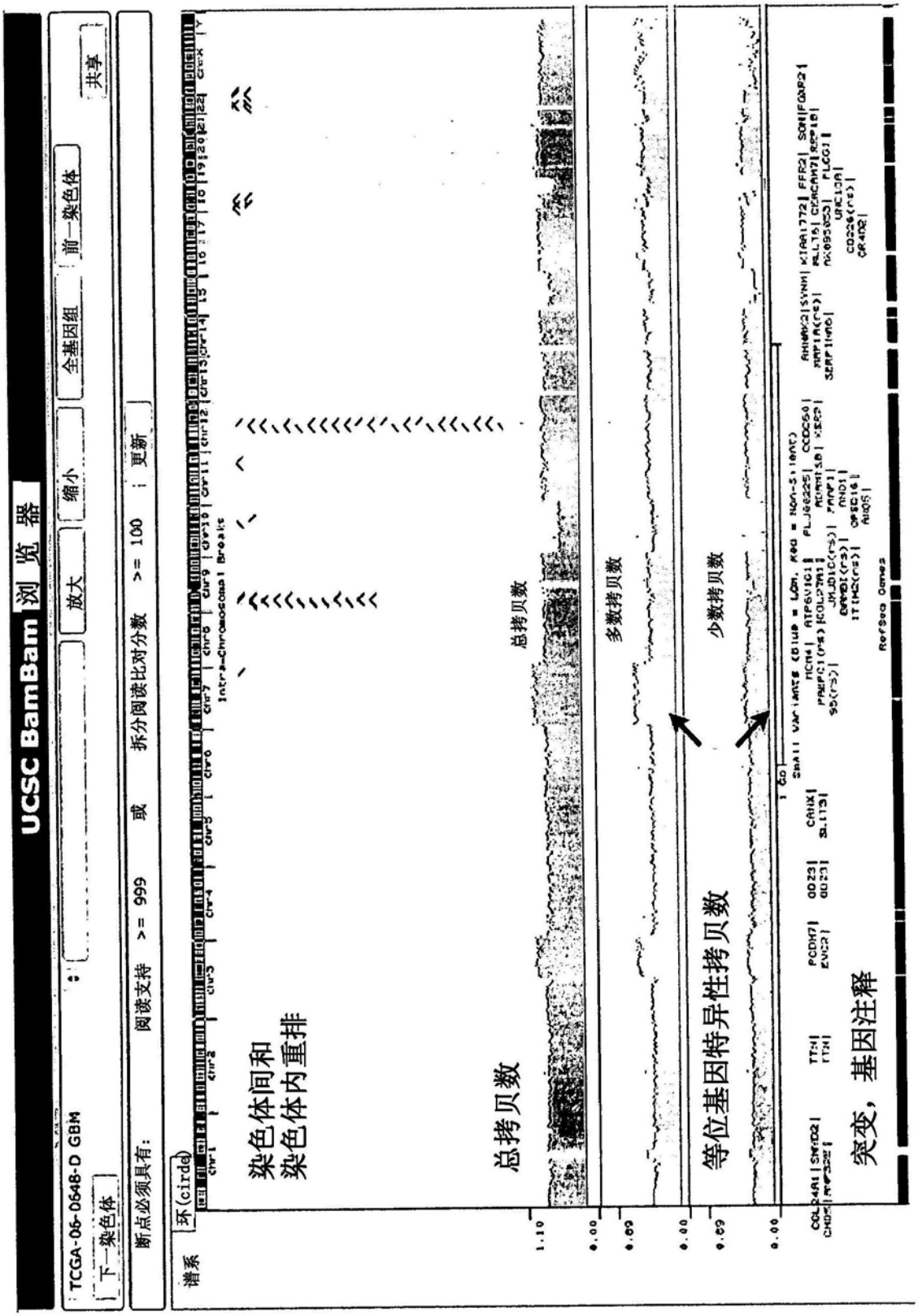

UCSC BamBam 浏览器
TCGA-06-0648-D GBM
放大
缩小
全基因组
前一染色体
下一染色体
共享
断点必须具有:
阅读支持 >= 999
或
拆分阅读比对分数 >= 100
更新
谱系
环(circle)
染色体间和
染色体内重排
总拷贝数
多数拷贝数
少数拷贝数
等位基因特异性拷贝数
突变，基因注释

图5