(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 107451419 B
(45)授权公告日 2020.01.24

(21)申请号 201710576797.3

(22)申请日 2017.07.14

(65)同一申请的已公布的文献号
申请公布号 CN 107451419 A

(43)申请公布日 2017.12.08

(73)专利权人 浙江大学
地址 310058 浙江省杭州市西湖区余杭塘路866号

(72)发明人 陆燕 孙喜伟 刘鹏渊 周莉媛

(74)专利代理机构 杭州求是专利事务所有限公司 33200
代理人 刘静 邱启旺

(51)Int.Cl.
G16B 30/00(2019.01)
G16B 5/00(2019.01)

(56)对比文件
CN 103088433 A,2013.05.08,
CN 103555856 A,2014.02.05,
CN 102796808 A,2012.11.28,
CN 101665835 A,2010.03.10,
CN 102648292 A,2012.08.22,
US 2012149593 A1,2012.06.14,
US 2015087529 A1,2015.03.26,
陈勋.基于简化基因组测序的油菜高通量SNP分析及白菜基因组DNA甲基化解析.《中国博士学位论文全文数据库 农业科技辑》.2014,(第9期),第D047-21页.
Allan Lazarovici.et..Probing DNA shape and methylation state on a genomic scale with DNase I.《PNAS》.2013,第110卷(第16期),第6376-6381页.
Michelle R. Lacey*.et..Modeling, simulation and analysis of methylation profiles from reduced representation bisulfite sequencing experiments.《Statistical Applications in Genetics and Molecular Biology 2013》.2013,第12卷(第6期),第1-20页.

审查员 葛晓倩

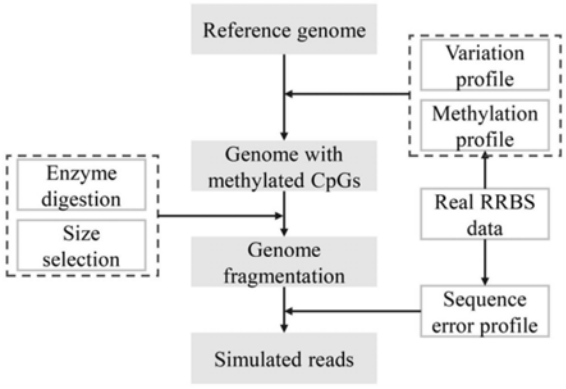
权利要求书1页 说明书3页 附图2页

(54)发明名称

通过计算机程序模拟产生简化DNA甲基化测序数据的方法

(57)摘要

本发明公开了一种通过计算机程序模拟产生简化DNA甲基化测序数据的方法,可以用来评估不同简化基因组甲基化(RRBS)测序数据比对软件的效率以及相应数据分析平台的可靠性,以此来确定最优的比对方法及相应的最优参数。本发明通过计算机程序模拟RRBS文库构建和测序过程,根据CpGs甲基化水平的分布,产生与真实RRBS测序数据非常相近的仿真数据。该仿真数据除了模拟单个碱基水平的甲基化水平外,还模拟了真实数据的插入、缺失、单核苷酸变异和结构变异等其他特性以增加真实性。此外,本发明在模拟RRBS测序过程中,通过引入经验误差模型来模拟测序过程中出现的误差,进一步增加了仿真数据的真实性。



CN 107451419 B

1.一种通过计算机程序模拟产生简化DNA甲基化测序数据的方法,其特征在于,具体包括以下四个步骤:

(1)模拟产生包含单碱基插入、缺失、单核苷酸变异和结构变异特性的参考基因组序列,变异参数由用户设定;

(2)在步骤(1)得到的参考基因组序列上模拟CpG二核苷酸位点上的甲基化水平:使用Beta统计模型来产生CpG位点的甲基化数值;考虑到真实数据中邻近CpG位点的强相关性,对100bp距离内的CpG位点的甲基化水平进行基于最大似然统计模型的修正;

(3)对步骤(2)得到的基因组序列进行计算机模拟生物体内的限制性酶切过程,记录相应的切割位置得到切割片段fragments,然后根据服从泊松分布的覆盖度参数,模拟产生单端或者双端的测序读长reads;通过内部选择过程,得到定向或者非定向文库的测序片段;

(4)模拟产生测序数据的质量文件:在Illumina测序中,碱基的测序质量值和它在读长中所处位置有关,越靠后质量越低,使用大量真实数据训练集来拟合Illumina测序的碱基质量分布,得到碱基的经验误差分布,并且用于产生数据的质量值情况。

2.根据权利要求1所述的通过计算机程序模拟产生简化DNA甲基化测序数据的方法,其特征在于:所述参考基因组包括人类各个版本参考基因组。

3.根据权利要求1所述的通过计算机程序模拟产生简化DNA甲基化测序数据的方法,其特征在于:所述计算机模拟的限制性酶切包括所有限制性内切酶。

通过计算机程序模拟产生简化DNA甲基化测序数据的方法

技术领域

[0001] 本发明属于计算机技术模拟产生简化DNA甲基化测序数据(生物信息学) 领域,具体涉及使用计算机程序模拟产生高仿真的简化DNA甲基化测序数据的方法。

背景技术

[0002] DNA甲基化是指在DNA序列不改变情况下,通过DNA化学修饰影响生物过程或者改变遗传表型。近年来,随着研究的深入,研究者发现DNA甲基化作为一种重要的表观遗传学修饰,其在肿瘤的形成发生发展过程中扮演重要角色。除此之外,研究也表明DNA甲基化还可能参与X染色体沉默,基因组印记,转座子沉默和干细胞分化等重要生物过程。因此,准确检测基因组上差异甲基化区域(DMR)对研究包括肿瘤在内的一些复杂疾病有非常重要的作用。

[0003] 随着二代测序迅猛发展以及测序成本的降低,基于高通量测序和亚硫酸盐处理的来研究甲基化的技术越来受到关注。其中最常见的是简化DNA甲基化测序技术(RRBS)。与此同时,各种基于RRBS测序数据的后期比对工具也随之发展,层出不穷;如何系统地评估这些比对工具的功效变得日趋重要。通常,由于真实RRBS测序数据的甲基化信息的实际情况无从得知,从而难以以此去评估相应的比对工具的可靠性;然而,使用计算机模拟的RRBS测序数据却可以通过自主设定参数从而得知模拟数据的“真实”甲基化信息,提供基础比较标准,从而能便捷有效地评估这些工具的表现,以此来确定最优的比对方法及相应的最优参数。

[0004] 目前已知的甲基化测序数据模拟工具大多是基于先前的全基因组甲基化测序(WGBS)设计的,并不适合目前广泛使用的RRBS测序技术数据,而且这些工具除了模拟甲基化信息外没有模拟真实数据的其它特性。近来,也有一些基于RRBS测序的数据模拟工具,但是他们只是通过统计模型产生一些DNA甲基化数据,并非模拟实际测序得到的读长(reads)数据,这些数据自然也不能用来评估相应拼接工具的可靠性。

发明内容

[0005] 本发明的目的在于针对现有技术的不足,提供一种通过计算机程序模拟产生简化DNA甲基化测序数据的方法,通过计算机程序模拟RRBS文库构建和测序过程,根据基因组上的CpGs甲基化水平的分布,产生与真实RRBS测序数据非常相近的仿真数据。其中,该过程引入了测序经验误差模型模拟测序出现的误差以增加仿真数据的真实性。该高仿真数据可以用来测试目前各种甲基化测序数据比对软件的功效,以此来确定最优的拼接方法及相应的最优参数,也可用于后续新的比对工具的辅助开发。

[0006] 本发明的目的是通过以下技术方案来实现的:一种通过计算机程序模拟产生简化DNA甲基化测序数据的方法,具体包括以下四个步骤:

[0007] (1)模拟产生包含单碱基插入、缺失、单核苷酸变异和结构变异特性(变异参数可由用户设定)的参考基因组序列,比如hg19;

[0008] (2)在步骤(1)得到的参考基因组上模拟CpG二核苷酸位点上的甲基化水平:由于人类基因组上的CpG位点的甲基化水平通常服从Beta分布,所以使用Beta统计模型来产生CpG位点的甲基化数值;此外,考虑到真实数据中邻近CpG位点的强相关性,对100bp距离内的CpG位点的甲基化水平进行基于最大似然统计模型的修正;

[0009] (3)对步骤(2)得到的基因组序列进行计算机模拟生物体内的限制性酶切过程(比如MspI限制性内切酶),记录相应的切割位置得到切割片段 (fragments),然后根据服从泊松分布的覆盖度参数,模拟产生单端或者双端的测序读长(reads);可以通过内部选择过程,得到定向或者非定向文库的测序片段。

[0010] (4)模拟产生测序数据的质量文件:在Illumina测序中,碱基的测序质量值和它在读长中所处位置有关(越靠后质量越低),因此使用大量真实数据训练集来拟合Illumina测序的碱基质量分布,得到碱基的经验误差分布,并且用于产生数据的质量值情况。

[0011] 进一步地,所述参考基因组包括人类各个版本参考基因组。

[0012] 进一步地,所述计算机模拟的限制性酶切包括所有限制性内切酶。

[0013] 本发明与背景技术相比具有的有益效果是:本发明提供了一种全面的 RRBS测序数据模拟方法,它能够提供模拟实际测序得到的读长(reads)数据,其中不仅包含了单个CpG位点的甲基化水平值,也有如插入、缺失、单核苷酸变异和结构变异等实际数据含有的其它特性值,从而能够更加全面有效地评估现有拼接工具的表现,也能够用于辅助开发新的比对工具。

附图说明

[0014] 图1为本发明的一个实施例流程图;

[0015] 图2为本发明产生的仿真数据和真实数据间的FastQC报告比较,A为碱基质量得分分布,B为碱基平均质量得分分布;

[0016] 图3为本发明方法模拟产生的和Illumina测序仪产生的测序数据对比图,A测序深度,B测序片段读长,C甲基化水平。

具体实施方式

[0017] 下面参照附图用本发明的示例性实施例对本发明进行更全面的描述及说明,但这并不意味着本发明仅限与此。

[0018] 实施例1:本发明提供的通过计算机程序(Python程序语言)模拟产生简化 DNA甲基化测序数据的方法,首先根据图1所示模拟产生简化DNA甲基化测序数据:

[0019] (1)模拟产生包含单碱基插入、缺失、单核苷酸变异和结构变异等特性 (这些变异参数可由用户自行给定)的参考基因组序列,比如hg19。

[0020] (2)在步骤(1)得到的参考基因组上模拟CpG二核苷酸位点上的甲基化水平。由于基因组上的CpG位点的甲基化水平通常服从Beta分布,所以我们使用Beta模型来产生CpG位点上的甲基化水平值。此外,考虑到真实数据中邻近CpG位点的强相关性,我们对100bp距离内的CpG位点的甲基化水平进行基于最大似然统计模型的修正。

[0021] (3)对步骤(2)得到的基因组序列进行计算机模拟生物体内的限制性酶切过程(比如MspI限制性内切酶),记录相应的切割位置得到切割片段 (fragments),然后根据服从泊

松分布的覆盖度参数，模拟产生单端或者双端的测序读长(reads)。该发明可以通过内部选择过程，实现得到定向或者非定向文库的测序片段。

[0022] 接着如图2所示模拟产生测序质量误差数据：模拟产生测序数据的质量文件。在Illumina测序中，碱基的测序质量值和它在读长中所处位置有关(越靠后质量越低)，因此我们使用大量的真实数据训练集来拟合Illumina测序的碱基质量分布，得到碱基的经验误差分布，并且用于产生数据的质量值情况。

[0023] 本发明方法性能的评估：从测序数据的碱基质量得分分布(图2A)、基因组上测序深度(图3A)、Msp1片段大小(图3B)和甲基化水平的分布(图 3C)等角度，本发明方法产生的RRBS数据和真实数据非常相似。所以，根据本发明方法产生的RRBS数据，我们可以用来测试目前各种甲基化测序数据拼接软件的功效，以此来确定最优的拼接方法及相应的最优参数，也可用于后续新的拼接工具的辅助开发。

[0024] 应当说明的是：以上实施例仅用以说明本发明的技术流程而不是对其限制，尽管参照上述实施例对本发明进行了详细的说明，所属领域的普通技术人员应当理解：依然可以对本发明的具体实施方式进行修改或者等同替换，而未脱离本发明精神和范围的任何修改或者等同替换，其均应该涵盖在本发明的权利要求范围当中。

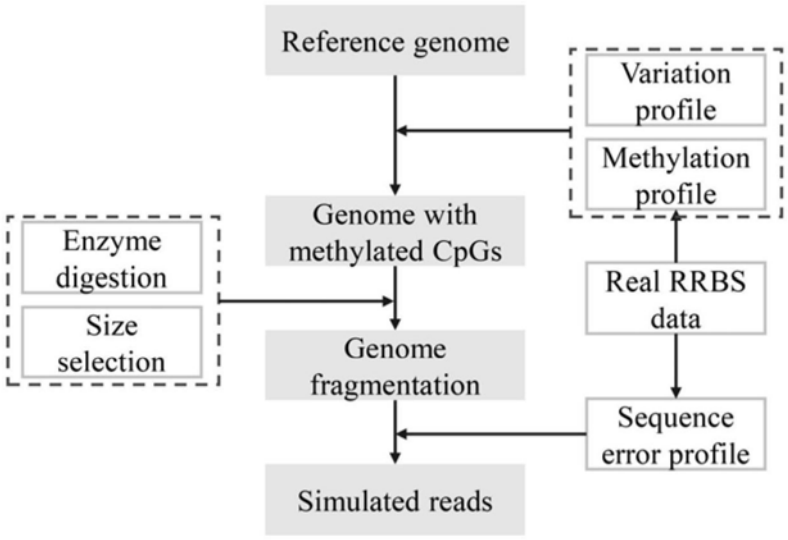

Reference genome
Variation profile
Methylation profile
Genome with methylated CpGs
Enzyme digestion
Size selection
Real RRBS data
Genome fragmentation
Sequence error profile
Simulated reads

图1

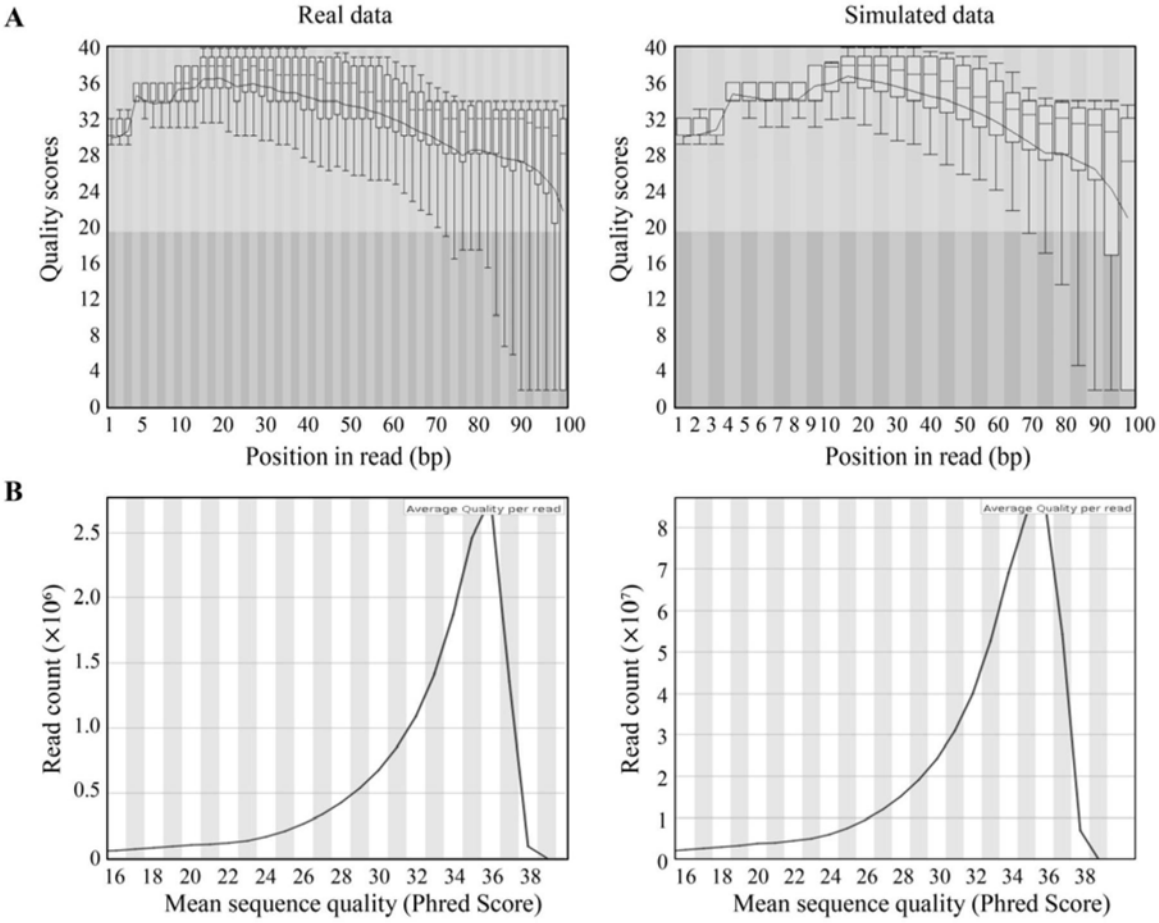

A
Real data
Simulated data
Quality scores
Position in read (bp)
B
Average Quality per read
Read count (×10⁶)
Read count (×10⁷)
Mean sequence quality (Phred Score)

图2

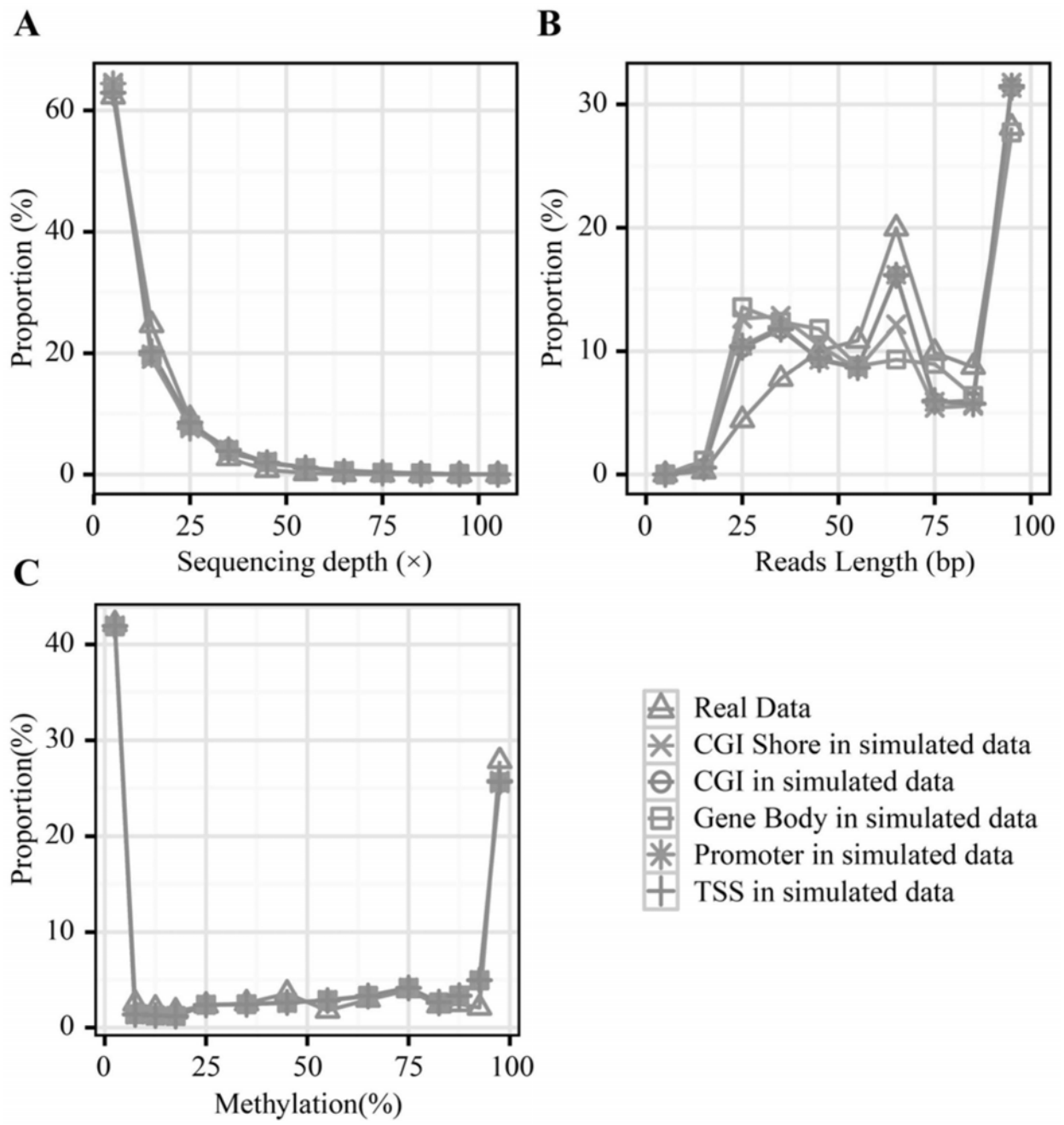

A
Proportion (%)
Sequencing depth (×)
B
Proportion (%)
Reads Length (bp)
C
Proportion(%)
Methylation(%)
Real Data
CGI Shore in simulated data
CGI in simulated data
Gene Body in simulated data
Promoter in simulated data
TSS in simulated data

图3