(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 120019143 A
(43)申请公布日 2025.05.16

(21)申请号 202380071979.9
(22)申请日 2023.08.15
(30)优先权数据
63/398,483 2022.08.16 US
63/380,523 2022.10.21 US
(85)PCT国际申请进入国家阶段日
2025.04.09
(86)PCT国际申请的申请数据
PCT/US2023/072257 2023.08.15
(87)PCT国际申请的公布数据
WO2024/040083 EN 2024.02.22
(71)申请人 布罗德研究所股份有限公司
地址 美国马萨诸塞州
申请人 哈佛大学的校长及成员们
(72)发明人 D·R·刘 M·内格鲍尔 E·张
(74)专利代理机构 北京市柳沈律师事务所 11105
专利代理师 陈丹丹
(51)Int.Cl.
C12N 9/22(2006.01)
A61K 38/00(2006.01)
A61K 48/00(2006.01)
C12N 9/78(2006.01)
C12N 15/10(2006.01)

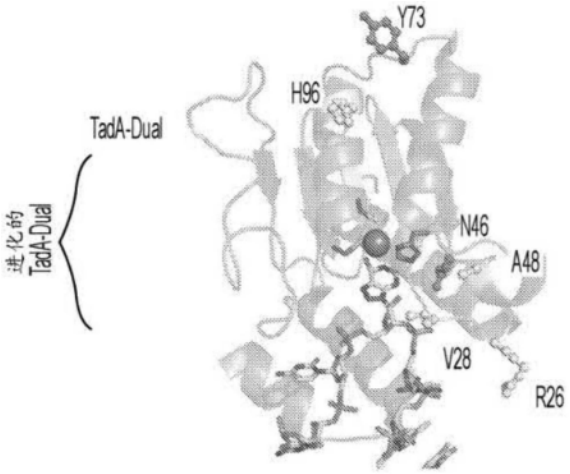
权利要求书12页 说明书197页
序列表（电子公布） 附图117页

(54)发明名称

经进化的胞苷脱氨酶及使用其编辑DNA的方法

(57)摘要

本公开总体上涉及衍生自胞苷脱氨酶的经进化的胞苷脱氨酶,以及使用其编辑DNA的方法。在一些方面,本公开描述了TadA衍生的腺苷脱氨酶(TadA-CD)的定向进化以进行胞苷脱氨。在一些实施方案中,与亲本TadA变体相比,TadA-CD包含多个突变。在一些实施方案中,TadA-CD与可编程DNA结合蛋白融合。本公开的其他方面总体上涉及包含可编程DNA结合蛋白和TadA-CD的胞嘧啶碱基编辑器(CBE)。在一些实施方案中,与天然存在的CBE相比,所公开的胞嘧啶碱基编辑器具有改进的转化效率和降低的脱靶编辑频率。还提供了用于生成和递送CBE的多核苷酸、载体和试剂盒。还提供了含有此类载体和CBE的细胞。此外,还提供了包括施用CBE的治疗方法。



CN 120019143 A

1.脱氨酶,其包含与SEQ ID NO:41的氨基酸序列至少80％、85％、90％、95％、98％、99％或99.5％相同的氨基酸序列,其中对应于SEQ ID NO:41的残基26的氨基酸是除R以外的任何氨基酸。

2.脱氨酶,其包含与SEQ ID NO:41的氨基酸序列至少80％、85％、90％、95％、98％、99％或99.5％相同的氨基酸序列,其中对应于SEQ ID NO:41的残基27的氨基酸是除E以外的任何氨基酸。

3.脱氨酶,其包含与SEQ ID NO:41的氨基酸序列至少80％、85％、90％、95％、98％、99％或99.5％相同的氨基酸序列,其中对应于SEQ ID NO:41的残基28的氨基酸是除V以外的任何氨基酸。

4.脱氨酶,其包含与SEQ ID NO:8的氨基酸序列至少80％、85％、90％、95％、98％、99％或99.5％相同的氨基酸序列,其中对应于SEQ ID NO:41的残基48的氨基酸是除R以外的任何氨基酸。

5.脱氨酶,其包含与SEQ ID NO:41的氨基酸序列至少80％、85％、90％、95％、98％、99％或99.5％相同的氨基酸序列,其中对应于SEQ ID NO:41的残基73的氨基酸是除Y以外的任何氨基酸。

6.脱氨酶,其包含与SEQ ID NO:41的氨基酸序列至少80％、85％、90％、95％、98％、99％或99.5％相同的氨基酸序列,其中对应于SEQ ID NO:41的残基96的氨基酸是除H以外的任何氨基酸。

7.脱氨酶,其在SEQ ID NO:41的氨基酸序列中包含残基E27、V28和H96处的突变,并且进一步包含选自R26、M61、Y73、I76、M151、Q154和A158的残基处的至少一个突变,或在同源腺苷脱氨酶中的相应突变。

8.根据权利要求1-7中任一项所述的脱氨酶,其中所述脱氨酶能够脱氨DNA中的胞苷。

9.根据权利要求7-8中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:41的氨基酸序列中包含选自E27A、E27K、V28G、V28A和H96N的至少一个突变,并且进一步包含选自R26G、M61I、Y73H、Y73S、Y73C、I76F、M151I、Q154R、Q154H和A158S的残基处的至少一个突变,或同源腺苷脱氨酶中的相应突变。

10.根据权利要求7-8中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:41的氨基酸序列中包含突变E27A、V28G和H96N,并且进一步包含选自R26G、M61I、Y73H、Y73S、Y73C、I76F、M151I、Q154R、Q154H和A158S的至少一个突变,或在同源腺苷脱氨酶中的相应突变。

11.根据权利要求7-8中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:41的氨基酸序列中包含突变E27K、V28G和H96N,并且进一步包含选自R26G、M61I、Y73H、Y73S、Y73C、I76F、M151I、Q154R、Q154H和A158S的至少一个突变,或在同源腺苷脱氨酶中的相应突变。

12.根据权利要求7-8中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:41的氨基酸序列中包含突变E27A、V28A和H96N,并且进一步包含选自R26G、M61I、Y73H、Y73S、Y73C、I76F、M151I、Q154R、Q154H和A158S的至少一个突变,或在同源腺苷脱氨酶中的相应突变。

13.根据权利要求7-8中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:41的氨基酸序列中包含突变E27K、V28A和H96N,并且进一步包含选自R26G、M61I、Y73H、Y73S、Y73C、I76F、M151I、Q154R、Q154H和A158S的至少一个突变,或在同源腺苷脱氨酶中的相应突变。

14.根据权利要求1-13中任一项所述的脱氨酶,其进一步包含位置V106处的突变。

15.根据权利要求1-13中任一项所述的脱氨酶,其进一步包含突变V106W。

16.根据权利要求1-15中任一项所述的脱氨酶,其包含在选自R26、M61、Y73、I76、M151、Q154和A158的残基处的至少两个突变。

17.根据权利要求1-16中任一项所述的脱氨酶,其包含在选自R26G、M61I、Y73H、I76F、M151I、Q154H、Q154R和A158S的残基处的至少两个突变。

18.根据权利要求1-17中任一项所述的脱氨酶,其包含选自E27A、V28G、I76F和M151I的至少一个突变。

19.根据权利要求1-17中任一项所述的脱氨酶,其包含选自E27A、V28G、I76F和A158S的至少一个突变。

20.根据权利要求1-17中任一项所述的脱氨酶,其包含选自E27A、V28G、I76F、Q154R和A158S的至少一个突变。

21.根据权利要求1-17中任一项所述的脱氨酶,其包含选自E27K、V28A和M61I的至少一个突变。

22.根据权利要求1-17中任一项所述的脱氨酶,其包含选自E27A、V28G、Y73H、Q154H和A158S的至少一个突变。

23.根据权利要求1-18中任一项所述的脱氨酶,其包含突变R26G、E27A、V28G、I76F、H96N和M151I。

24.根据权利要求1-17中任一项所述的脱氨酶,其包含突变R26G、E27A、V28G、I76F、H96N和A158S。

25.根据权利要求1-17中任一项所述的脱氨酶,其包含突变R26G、E27A、V28G、I76F、H96N、Q154R和A158S。

26.根据权利要求1-17中任一项所述的脱氨酶,其包含突变E27K、V28A、M61I和H96N。

27.根据权利要求1-17中任一项所述的脱氨酶,其包含突变E27A、V28G、Y73H、H96N、Q154H和A158S。

28.根据权利要求7-27中任一项所述的脱氨酶,其中所述脱氨酶的胞苷脱氨活性超过TadA-8e的胞苷脱氨活性。

29.根据权利要求7-28中任一项所述的脱氨酶,其中所述脱氨酶的胞苷脱氨活性超过所述脱氨酶的腺苷脱氨活性。

30.根据权利要求7-29中任一项所述的脱氨酶,其中所述脱氨酶的胞苷脱氨活性与腺苷脱氨活性的比率为至少约5:1、6:1、7:1、8:1、9:1、10:1、11:1、12:1、13:1、14:1、15:1、17:1、19:1、20:1、21:1、23:1、25:1、30:1或大于30:1。

31.根据权利要求7-30中任一项所述的脱氨酶,其中所述脱氨酶的胞苷脱氨活性与腺苷脱氨活性的比率为至少约10:1。

32.根据权利要求7-31中任一项所述的脱氨酶,其中所述脱氨酶的胞苷脱氨活性与腺苷脱氨活性的比率为至少约20:1。

33.根据权利要求2-32中任一项所述的脱氨酶,其中所述脱氨酶实现至少约10％、15％、20％、25％、30％、35％、40％、45％、50％、55％、60％、65％、70％、75％或80％的胞苷转化为胸腺嘧啶的效率。

34.根据权利要求2-33中任一项所述的脱氨酶,其中所述脱氨酶实现至少约75％的胞

苷转化为胸腺嘧啶的效率。

35.脱氨酶,其在SEQ ID NO:41的氨基酸序列中包含残基R26、V28、A48和Y73处的突变,或在同源腺苷脱氨酶(例如TadA-dual,SEQ ID NO:39)中的相应突变。

36.根据权利要求35所述的脱氨酶,其中所述脱氨酶能够脱氨DNA中的胞苷。

37.根据权利要求35或36所述的脱氨酶,其中所述脱氨酶进一步包含残基H96处的突变。

38.根据权利要求35-37中任一项所述的脱氨酶,其包含突变R26G、V28A、A48R、Y73S和H96N。

39.根据权利要求35-38中任一项所述的脱氨酶,其包含突变R26G、V28G、A48R和Y73C。

40.根据权利要求35-39中任一项所述的脱氨酶,其中所述脱氨酶的腺苷脱氨活性与胞苷脱氨活性的比率为至少约0.7:1、0.8:1、0.9:1、1:1、1.1:1、1.2:1、1.3:1、1.4:1或1.5:1。

41.脱氨酶,其包含与SEQ ID NO:39的氨基酸序列至少80%、85%、90%、95%、98%、99%或99.5%相同的氨基酸序列,其中对应于SEQ ID NO:39的残基46的氨基酸是除N以外的任何氨基酸。

42.脱氨酶,其在SEQ ID NO:39的氨基酸序列中包含残基N46处的突变,并进一步包含选自G26、A28、L34、N46、R48、R64、Q71、S73、N96、G105、H154和A162的残基处的至少一个突变,或在同源腺苷脱氨酶中的相应突变。

43.根据权利要求41或42所述的脱氨酶,其中所述脱氨酶能够脱氨DNA中的胞苷。

44.根据权利要求41-43中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含选自N46I、N46V、N46L和N46C的至少一个突变,并进一步包含S73P和H154Q突变,或在同源腺苷脱氨酶中的相应突变。

45.根据权利要求41-43中任一项所述的脱氨酶,其中所述脱氨酶SEQ ID NO:39的氨基酸序列中包含在位置N46V处的突变,并进一步包含在残基S73P、G105S和H154Q处的突变,或在同源腺苷脱氨酶中的相应突变。

46.根据权利要求41-43中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含位置N46L处的突变,并进一步包含残基G26R、R48P、S73P、N96H和H154Q处的突变,或在同源腺苷脱氨酶中的相应突变。

47.根据权利要求41-43中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含位置N46V处的突变,并进一步包含残基Q71H、S73P和H154Q处的突变,或在同源腺苷脱氨酶中的相应突变。

48.根据权利要求41-43中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含位置N46C处的突变,并进一步包含残基S73P、H154Q和A162V处的突变,或在同源腺苷脱氨酶中的相应突变。

49.根据权利要求41-43中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含位置N46I处的突变,并进一步包含残基H154Q处的突变,或在同源腺苷脱氨酶中的相应突变。

50.根据权利要求41-43中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含选自由N46T、N46V、N46C和N46L组成的组的突变,并进一步包含位置H154Q处的突变,或在同源腺苷脱氨酶中的相应突变。

51.根据权利要求41-43中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含N46T处的突变,或在同源腺苷脱氨酶中的相应突变。

52.根据权利要求41-43中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含选自由N46V、N46L和N46C组成的组的至少一个突变,并进一步包含选自由残基S73P、S73Y和A162V组成的组的一个或多个突变,或在同源腺苷脱氨酶中的相应突变。

53.根据权利要求41-43中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含N46C突变,并进一步包含残基S73P和A162V处的突变,或在同源腺苷脱氨酶中的相应突变。

54.根据权利要求41-43中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含N46C突变,并进一步包含残基S73P处的突变,或在同源腺苷脱氨酶中的相应突变。

55.根据权利要求41-43中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含N46C突变,并进一步包含残基S73Y和A162V处的突变,或在同源腺苷脱氨酶中的相应突变。

56.根据权利要求41-43中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含N46V突变,并进一步包含残基Q71H和S73P处的突变,或在同源腺苷脱氨酶中的相应突变。

57.根据权利要求41-43中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含N46V突变,并进一步包含残基S73P处的突变,或在同源腺苷脱氨酶中的相应突变。

58.根据权利要求41-43中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含N46L处的突变,或在同源腺苷脱氨酶中的相应突变。

59.根据权利要求41-43中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含N46V突变,并进一步包含残基S73P处的突变,或在同源腺苷脱氨酶中的相应突变。

60.根据权利要求41-43中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含N46V处的突变,或在同源腺苷脱氨酶中的相应突变。

61.根据权利要求41-43中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含N46V突变,并进一步包含残基R48P处的突变,或在同源腺苷脱氨酶中的相应突变。

62.根据权利要求41-43中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含N46C突变,并进一步包含残基S73P处的突变,或在同源腺苷脱氨酶中的相应突变。

63.根据权利要求41-43中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含N46L突变,并进一步包含残基L34M和S73P处的突变,或在同源腺苷脱氨酶中的相应突变。

64.根据权利要求41-43中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含N46L突变,并进一步包含残基S73P处的突变,或在同源腺苷脱氨酶中的相应突变。

65.根据权利要求41-43中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含N46L突变,并进一步包含残基R48P、R64K和S73P处的突变,或在同源腺苷脱氨酶中的相应突变。

66.脱氨酶,其包含与SEQ ID NO:39的氨基酸序列至少80%、85%、90%、95%、98%、99%或99.5%相同的氨基酸序列,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含Q71S和H154Q处的突变,或在同源腺苷脱氨酶中的相应突变。

67.脱氨酶,其包含与SEQ ID NO:39的氨基酸序列至少80%、85%、90%、95%、98%、99%或99.5%相同的氨基酸序列,其中对应于SEQ ID NO:79的残基46的氨基酸是除N以外的任何氨基酸。

68.脱氨酶,其在SEQ ID NO:39的氨基酸序列中包含残基T79处的突变,并进一步包含选自A28、N46、R48、S73、N96和G105的残基处的至少一个突变,或在同源腺苷脱氨酶中的相应突变。

69.根据权利要求67或68所述的脱氨酶,其中所述脱氨酶能够脱氨DNA中的胞苷。

70.根据权利要求67-69中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含选自N79T或N79P的至少一个突变,并进一步包含选自由N46L、N46V、N46I、R48、R48P、S73P、S73Y、S73H、N96H和G105S组成的组的一个或多个突变,或在同源腺苷脱氨酶中的相应突变。

71.根据权利要求67-69中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含N79T突变,并进一步包含残基N46L、S73P、N79T和N96H处的突变,或在同源腺苷脱氨酶中的相应突变。

72.根据权利要求67-69中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含N79T突变,并进一步包含残基N46L、S73P和N79T处的突变,或在同源腺苷脱氨酶中的相应突变。

73.根据权利要求67-69中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含N79T突变,并进一步包含残基R48A、S73P和N79T处的突变,或在同源腺苷脱氨酶中的相应突变。

74.根据权利要求67-69中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含N79T突变,并进一步包含残基N46V处的突变,或在同源腺苷脱氨酶中的相应突变。

75.根据权利要求67-69中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含N79T突变,并进一步包含残基N46V和S73P处的突变,或在同源腺苷脱氨酶中的相应突变。

76.根据权利要求67-69中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含N79T突变,并进一步包含残基A28V、N46L、R48A、S73Y和N96H处的突变,或在同源腺苷脱氨酶中的相应突变。

77.根据权利要求67-69中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含N79T突变,并进一步包含残基N46I和S73P处的突变,或在同源腺苷脱氨酶中的相应突变。

78.根据权利要求67-69中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨

基酸序列中包含N79T突变,并进一步包含残基N46V和S73P处的突变,或在同源腺苷脱氨酶中的相应突变。

79.根据权利要求67-69中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含N79P突变,并进一步包含残基R48P和S73H处的突变,或在同源腺苷脱氨酶中的相应突变。

80.根据权利要求67-69中任一项所述的脱氨酶,其中所述脱氨酶在SEQ ID NO:39的氨基酸序列中包含N79T突变,并进一步包含残基A28V、N46I、R48A和S73Y处的突变,或在同源腺苷脱氨酶中的相应突变。

81.脱氨酶,其在SEQ ID NO:39的氨基酸序列中包含Q71S和H154Q处的突变,或在同源腺苷脱氨酶中的相应突变。

82.根据权利要求41-81中任一项所述的脱氨酶,其中所述脱氨酶的胞苷脱氨活性超过TadA-Dual的胞苷脱氨活性。

83.根据权利要求41-82中任一项所述的脱氨酶,其中所述脱氨酶的腺苷脱氨活性与胞苷脱氨活性的比率为至少约0.001:1、0.005:1、0.007:1、0.01:1、0.05:1、0.07:1或0.1:1。

84.胞苷脱氨酶,其已经通过连续进化和/或非连续进化从腺苷脱氨酶进化。

85.根据权利要求84所述的胞苷脱氨酶,其中所述腺苷脱氨酶包含与SEQ ID NO:34-39、41-54、33、315、317-323、326、354和355中任一个的氨基酸序列至少80%、85%、90%、95%、98%、99%或99.5%相同的氨基酸序列。

86.碱基编辑器,其包含核酸可编程DNA结合蛋白(napDNAbp)结构域和包含根据权利要求1-85中任一项所述的脱氨酶的TadA-CD结构域。

87.根据权利要求86所述的碱基编辑器,其中所述napDNAbp结构域是切口酶。

88.根据权利要求86或87所述的碱基编辑器,其中所述napDNAbp结构域选自SpCas9n、dCas9、CasX、CasY、C2c1、C2c2、C2c3、GeoCas9、CjCas9、Cas12a、Cas12b、Cas12g、Cas12h、Cas12i、Cas13b、Cas13c、Cas13d、Cas14、Csn2、xCas9、Cas9-NG、LbCas12a、enAsCas12a、SaCas9、SaCas9-KKH、环状排列的Cas9、Argonaute(Ago)结构域、SmacCas9、Spy-macCas9、SpCas9-VRQR、SpCas9-NRRH、SpaCas9-NRTH、SpCas9-NRCH、eNme2Cas9、eNme2-C Cas9、enCjCas9、SauriCas9、Cas9-NG-VRQR及其变体。

89.根据权利要求86-88中任一项所述的碱基编辑器,其中所述napDNAbp结构域包含与SEQ ID NO:74-77、343-346、347、348、351-353和356-358中的任一个至少85%、90%、92.5%、95%、97%、98%或99%相同的氨基酸序列。

90.根据权利要求86-88中任一项所述的碱基编辑器,其中所述napDNAbp结构域是Cas9n或SpCas9-NG结构域。

91.根据权利要求86-90中任一项所述的碱基编辑器,其中所述napDNAbp结构域包含SEQ ID NO:77或343所示的氨基酸序列。

92.根据权利要求86-91中任一项所述的碱基编辑器,其中所述napDNAbp结构域是eNme2-C Cas9结构域。

93.根据权利要求86-92中任一项所述的碱基编辑器,其中所述napDNAbp结构域是enCjCas9结构域。

94.根据权利要求86-93中任一项所述的碱基编辑器,其中所述napDNAbp结构域是

SaCas9结构域。

95.根据权利要求91-94中任一项所述的碱基编辑器,其中所述napDNAbp结构域包含SEQ ID NO:347、348和353中任一个所示的氨基酸序列。

96.根据权利要求86-95中任一项所述的碱基编辑器,其中当与包含选自由CTT、CTC、CTA、CTG、CCT、CCC、CCA、CCG、CAT、CAC、CAA、CAG、CGT、CGC、CGA、CGG、TCT、TCC、TCG、ACT、ACC、ACA、ACG、GCT、GCC、GCA、GCG、TTC、TAC、TGC、ATC、AAC、AGC、GTC、GAC和GGC组成的组的靶序列的DNA接触时,所述碱基编辑器提供至少约10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%或80%的胞嘧啶转化为胸腺嘧啶的效率。

97.根据权利要求86-96中任一项所述的碱基编辑器,其中所述碱基编辑器引起在约0.1%至约0.35%的范围内的脱靶编辑频率。

98.根据权利要求86-97中任一项所述的碱基编辑器,其中所述碱基编辑器还包含一个或多个UGI结构域。

99.根据权利要求86-98中任一项所述的碱基编辑器,其中所述碱基编辑器还包含两个UGI结构域。

100.根据权利要求86-99中任一项所述的碱基编辑器,其中所述碱基编辑器包含一个或多个核定位序列(NLS)。

101.根据权利要求86-100中任一项所述的碱基编辑器,其中所述碱基编辑器还包含二分核定位信号(bpNLS)。

102.根据权利要求101所述的碱基编辑器,其中所述二分核定位信号包含选自由以下组成的组的氨基酸序列:KRTADGSEFEPKKKRKV(SEQ ID NO:155)、KRPAATKKAGQAKKKK(SEQ ID NO:276)、KKTELQTTNAENKTKKL(SEQ ID NO:277)、KRGINDRNFWRGENGRKTR(SEQ ID NO:278)和RKSGKIAAIVVKRPRK(SEQ ID NO:279)。

103.根据权利要求101或102所述的碱基编辑器,其中所述二分核定位信号包含SEQ ID NO:276或155所示的氨基酸序列。

104.根据权利要求86-103中任一项所述的碱基编辑器,其中所述碱基编辑器包含以下结构:NH_2-[第一核定位序列]-[TadA-CD结构域]-[napDNAbp结构域]-[第一UGI结构域]-[第二UGI结构域]-[第二核定位序列]-COOH,其中"]-["的每个实例指示任选的接头序列的存在;任选地,其中所述碱基编辑器包含以下结构:NH_2-[第一NLS]-[TadA-CD结构域]-[SaCas9n]-[UGI结构域]-[UGI结构域]-[第二NLS]-COOH;NH_2-[第一NLS]-[TadA-CD结构域]-[eNme2-C Cas9n]-[UGI结构域]-[UGI结构域]-[第二NLS]-COOH;NH_2-[第一NLS]-[TadA-CD结构域]-[CjCas9n]-[UGI结构域]-[UGI结构域]-[第二NLS]-COOH;或NH_2-[第一NLS]-[TadA-CD结构域]-[SpCas9-NG]-[UGI结构域]-[UGI结构域]-[第二NLS]-COOH。

105.根据权利要求86-104中任一项所述的碱基编辑器,其中所述TadA-CD结构域和所述napDNAbp结构域经由包含氨基酸序列SGGSSGGSSGSETPGTSESATPESSGGSSGGS(SEQ ID NO:165)的接头连接;所述napDNAbp结构域和所述第一UGI结构域经由包含SGGSGGSGGS(SEQ ID NO:166)的氨基酸序列的接头连接;所述第一UGI结构域和所述第二UGI结构域经由包含SGGSGGSGGS(SEQ ID NO:166)的氨基酸序列的接头连接;和/或所述第二UGI结构域和所述第二核定位序列经由包含SGGS(SEQ ID NO:160)的氨基酸序列的接头序列连接。

106.根据权利要求86-105中任一项所述的碱基编辑器,其中所述碱基编辑器包含与

SEQ ID NO:19-31中的任一个至少85％、90％、92.5％、95％、97％、98％或99％相同的氨基酸序列。

107.根据权利要求86-106中任一项所述的碱基编辑器,其中所述碱基编辑器包含SEQ ID NO:19-31所示的氨基酸序列中的任一个。

108.复合物,其包含根据权利要求86-107中任一项所述的碱基编辑器和与所述碱基编辑器的所述napDNAbp结构域结合的向导RNA。

109.根据权利要求108所述的复合物,其中所述向导RNA为20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50、55、60、65、70、75、80、85、90、95、100、125、150、175或200个核苷酸长。

110.根据权利要求108或109所述的复合物,其中所述向导RNA为15至80个核苷酸长,并包含与靶序列互补的至少10个、至少15个或至少20个连续核苷酸的序列。

111.根据权利要求108-110中任一项所述的复合物,其中所述向导RNA包含与靶序列互补的15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39或40个连续核苷酸的序列。

112.根据权利要求110或111所述的复合物,其中所述靶序列是DNA序列。

113.根据权利要求110-112中任一项所述的复合物,其中所述靶序列位于生物体的基因组中。

114.根据权利要求113所述的复合物,其中所述生物体是原核生物。

115.根据权利要求114所述的复合物,其中所述原核生物是细菌。

116.根据权利要求113所述的复合物,其中所述生物体是真核生物。

117.根据权利要求116所述的复合物,其中所述真核生物是植物或真菌。

118.根据权利要求116所述的复合物,其中所述真核生物是哺乳动物。

119.根据权利要求118所述的复合物,其中所述哺乳动物是啮齿动物。

120.根据权利要求118所述的复合物,其中所述哺乳动物是人。

121.根据权利要求110-120中任一项所述的复合物,其中所述靶序列位于细胞的基因组中。

122.根据权利要求121所述的复合物,其中所述细胞是植物细胞、啮齿动物细胞或人细胞。

123.根据权利要求121或122所述的复合物,其中所述细胞是T细胞或造血干细胞(HSC)。

124.多核苷酸,其编码根据权利要求86-107中任一项所述的碱基编辑器。

125.根据权利要求124所述的多核苷酸,其中所述多核苷酸针对在人细胞中的表达进行密码子优化。

126.根据权利要求124或125所述的多核苷酸,其中所述多核苷酸针对在哺乳动物细胞中的表达进行密码子优化。

127.载体,其包含根据权利要求124-126中任一项所述的多核苷酸。

128.根据权利要求127所述的载体,其中所述载体包含驱动所述多核苷酸表达的异源启动子。

129.根据权利要求127或128所述的载体,其进一步包含编码向导RNA(gRNA)的多核苷

酸。

130.根据权利要求127-129中任一项所述的载体,其中所述载体是mRNA构建体。

131.根据权利要求127-130中任一项所述的载体,其中所述载体是重组AAV载体。

132.根据权利要求129所述的载体,其中编码所述gRNA的多核苷酸的方向相对于根据权利要求124-126中任一项所述的多核苷酸是反向的。

133.重组腺相关病毒(rAAV)颗粒,其包含根据权利要求127-132中任一项所述的AAV载体。

134.细胞,其包含根据权利要求1-85中任一项所述的脱氨酶、根据权利要求86-107中任一项所述的碱基编辑器、根据权利要求108-123中任一项所述的复合物、根据权利要求124-126中任一项所述的多核苷酸、根据权利要求127-132中任一项所述的载体或根据权利要求133所述的rAAV颗粒。

135.根据权利要求134所述的细胞,其中所述细胞是T细胞。

136.根据权利要求134所述的细胞,其中所述细胞是干细胞。

137.根据权利要求134所述的细胞,其中所述细胞是人造血干细胞(HSC)。

138.根据权利要求134-137中任一项所述的细胞,其中所述细胞已经从受试者中获得,并与根据权利要求86-107中任一项所述的碱基编辑器、根据权利要求108-123中任一项所述的复合物或根据权利要求127-132中任一项所述的载体离体接触。

139.药物组合物,其包含根据权利要求86-107中任一项所述的碱基编辑器、根据权利要求108-123中任一项所述的复合物或根据权利要求127-132中任一项所述的载体。

140.根据权利要求139所述的药物组合物,其进一步包含药学上可接受的赋形剂。

141.方法,其包括将核酸与根据权利要求86-107中任一项所述的碱基编辑器或根据权利要求108-123中任一项所述的复合物接触。

142.根据权利要求141所述的方法,其中所述核酸包含细胞基因组中的靶序列。

143.根据权利要求141或142所述的方法,其中所述核酸是DNA。

144.根据权利要求141-143中任一项所述的方法,其中所述核酸是双链DNA。

145.根据权利要求142-144中任一项所述的方法,其中所述靶序列包含与疾病或病症相关的序列。

146.根据权利要求142-145中任一项所述的方法,其中所述靶序列包含BCL11A增强子中的序列。

147.根据权利要求142-146中任一项所述的方法,其中所述靶序列包含CCR5或CXCR4基因中的序列。

148.根据权利要求142-147中任一项所述的方法,其中所述靶序列包含与疾病或病症相关的点突变。

149.根据权利要求141-148中任一项所述的方法,其中所述碱基编辑器或所述复合物的活性导致所述点突变的校正。

150.根据权利要求142-149中任一项所述的方法,其中所述靶序列包含与疾病或病症相关的T至C点突变,并且其中突变体C碱基的脱氨导致与所述疾病或病症不相关的序列。

151.根据权利要求142-150中任一项所述的方法,其中所述靶序列包含与疾病或病症相关的A至G点突变,并且其中与所述A至G点突变的G碱基互补的突变体C碱基的脱氨导致与

所述疾病或病症不相关的序列。

152.根据权利要求150或151所述的方法,其中所述突变体C的脱氨导致由突变体密码子编码的氨基酸的变化。

153.根据权利要求150-152中任一项所述的方法,其中所述突变体C的脱氨导致编码野生型氨基酸的密码子。

154.根据权利要求151所述的方法,其中与所述A至G点突变的G碱基互补的C碱基的脱氨导致由突变体密码子编码的氨基酸的变化。

155.根据权利要求150-154中任一项所述的方法,其中所述脱氨导致终止密码子的引入。

156.根据权利要求150-154中任一项所述的方法,其中所述脱氨导致终止密码子的去除。

157.根据权利要求155或156所述的方法,其中所述终止密码子包含核酸序列5'-TAG-3'、5'-TAA-3'或5'-TGA-3'。

158.根据权利要求150-157中任一项所述的方法,其中所述脱氨导致剪接位点的引入。

159.根据权利要求150-157中任一项所述的方法,其中所述脱氨导致剪接位点的去除。

160.根据权利要求150-157中任一项所述的方法,其中所述脱氨导致基因启动子中突变的引入。

161.根据权利要求160所述的方法错误!未找到引用来源,其中所述突变导致与所述基因启动子可操作地连接的基因的转录的增加。

162.根据权利要求160或161所述的方法,其中所述突变导致与所述基因启动子可操作地连接的基因的转录的减少。

163.根据权利要求150-162中任一项所述的方法,其中所述脱氨导致基因阻遏物中突变的引入。

164.根据权利要求163所述的方法,其中所述突变导致与所述基因阻遏物可操作地连接的基因的转录的增加。

165.根据权利要求163或164所述的方法,其中所述突变导致与所述基因阻遏物可操作地连接的基因的转录的减少。

166.根据权利要求142-165中任一项所述的方法,其中所述靶序列编码蛋白质,并且其中所述点突变位于密码子中并导致与野生型密码子相比由所述突变体密码子编码的氨基酸的变化。

167.根据权利要求141-166中任一项所述的方法,其中所述接触步骤在受试者体内进行。

168.根据权利要求141-167中任一项所述的方法,其中所述接触步骤在体外或离体进行。

169.根据权利要求167所述的方法,其中所述受试者已被诊断患有疾病或病症。

170.根据权利要求169所述的方法,其中所述疾病或病症是HIV/AIDS或镰状细胞病。

171.根据权利要求142-170中任一项所述的方法,其中所述靶序列包含DNA序列5'-NCN-3',其中N为A、T、C或G。

172.根据权利要求171所述的方法,其中所述5'-NCN-3'序列中心的C被脱氨。

173.根据权利要求171或172所述的方法,其中所述5'-NCN-3'序列中心的C被改变为T。

174.根据权利要求141-174中任一项所述的方法,其中所述方法导致低于20%、19%、18%、16%、14%、12%、10%、8%、6%、4%、2%、1%、0.5%、0.2%或0.1%的插入缺失(indel)形成。

175.根据权利要求141-174中任一项所述的方法,其中所述方法提供至少约5:1、6:1、7:1、8:1、9:1、10:1、11:1、12:1、13:1、14:1、15:1、17:1、19:1、20:1、21:1、23:1、25:1、30:1或大于30:1的胞苷脱氨活性与腺苷脱氨活性的比率。

176.根据权利要求141-175中任一项所述的方法,其中所述脱氨酶的胞苷脱氨活性与腺苷脱氨活性的比率为至少约10:1。

177.根据权利要求141-176中任一项所述的方法,其中所述脱氨酶的胞苷脱氨活性与腺苷脱氨活性的比率为至少约20:1。

178.根据权利要求141-177中任一项所述的方法,其中所述方法提供低于1%、低于0.75%、低于0.5%、低于0.4%、低于0.35%、低于0.25%、低于0.2%、低于0.15%或低于0.1%的脱靶编辑频率。

179.根据权利要求141-178中任一项所述的方法,其中所述方法提供约0.35%或更低的脱靶编辑频率。

180.根据权利要求141-179中任一项所述的方法,其中所述方法导致约25:1、50:1、65:1、75:1、80:1、85:1、90:1、95:1、100:1、110:1、125:1或超过125:1的中靶:脱靶编辑的比率。

181.根据权利要求141-180中任一项所述的方法,其中所述方法在CXCR4或CCR5基因中导致约90:1或更高的中靶:脱靶编辑的比率。

182.根据权利要求141-181中任一项所述的方法,其中所述靶序列包含靶窗口,其中所述靶窗口包含靶核苷酸碱基对。

183.根据权利要求182所述的方法,其中所述靶窗口的长度为1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20个核苷酸。

184.方法,其包括向受试者施用根据权利要求127-132中任一项所述的载体、根据权利要求134-138中任一项所述的细胞或根据权利要求139或140所述的药物组合物。

185.根据权利要求184所述的方法,其中所述受试者是哺乳动物。

186.根据权利要求184或185所述的方法,其中所述受试者是人。

187.根据权利要求184-186中任一项所述的方法,其中所述施用步骤包括离体工程化根据权利要求134-138中任一项所述的细胞,并将所述细胞施用至所述受试者。

188.试剂盒,其包含核酸构建体,所述核酸构建体包含:

(a)编码根据权利要求86-107中任一项所述的碱基编辑器的核酸序列;

(b)编码gRNA的核酸序列;以及

(c)驱动(a)的序列和/或(b)的序列表达的一个或多个异源启动子。

189.根据权利要求188所述的试剂盒,其进一步包含编码向导RNA主链的表达构建体,其中所述构建体包含克隆位点,所述克隆位点定位于允许将与靶序列相同或互补的核酸序列克隆到所述向导RNA主链中的位置。

190.(a)根据权利要求86-107中任一项所述的碱基编辑器和(b)将(a)的所述碱基编辑器靶向至双链DNA分子中的靶标C:G核苷酸碱基对的向导RNA在DNA编辑中的用途。

191.根据权利要求86-107中任一项所述的碱基编辑器、根据权利要求108-123中任一项所述的复合物、根据权利要求134-138中任一项所述的细胞或根据权利要求130或140所述的药物组合物作为药物的用途。

192.根据权利要求86-107中任一项所述的碱基编辑器、根据权利要求108-123中任一项所述的复合物、根据权利要求134-138中任一项所述的细胞或根据权利要求130或140所述的药物组合物作为治疗镰状细胞病的药物的用途。

193.载体系统,其包括:

(i)选择质粒,其包含编码腺苷脱氨酶的分离的核酸,其按以下顺序包含:腺苷脱氨酶蛋白和编码分裂内含肽的N端部分的序列;

(ii)第一辅助质粒,其按以下顺序包含:由Lac启动子可操作地控制的编码向导RNA的序列和由T7 RNA启动子可操作地控制的编码M13噬菌体基因III(gIII)肽的序列;

(iii)第二辅助质粒,其按以下顺序包含:编码分裂内含肽的C端部分的序列和编码dCas9-UGI融合物的序列;以及

(iv)第三辅助质粒,其包含非编码链和编码链,其中所述编码链包含表达构建体,所述表达构建体按以下顺序包含:启动子、核糖体结合位点和编码T7 RNA聚合酶及降解决定子标签的序列,其中与编码T7 RNA聚合酶的序列的3'末端相对的所述非编码链包含CAA序列。

194.根据权利要求193所述的载体系统,其中所述分裂内含肽是Npu(点形念珠藻)内含肽。

195.根据权利要求193或194所述的载体系统,其中所述腺苷脱氨酶是TadA-8e。

196.根据权利要求193-195中任一项所述的载体系统,其进一步包含诱变质粒。

197.细胞,其包含根据权利要求193-196中任一项所述的载体系统。

198.使用一轮或多轮PACE或PANCE进化选择从腺苷脱氨酶进化的胞苷脱氨酶的方法。

199.根据权利要求198所述的方法,其中所述一轮或多轮PACE或PANCE进化包括:

编码与NpuN内含肽融合的突变TadA8e蛋白的选择噬菌体,

a.第一质粒,其编码与dCas9-UGI融合的NpuC内含肽,

b.第二质粒,其编码由T7或proT7启动子驱动的基因III(gIII)并编码sgRNA,以及

c.第三质粒,其编码T7 RNA聚合酶-降解决定子融合物。

200.根据权利要求199所述的方法,其中所述T7 RNA聚合酶-降解决定子融合物在所述T7 RNA聚合酶和降解决定子结构域之间的界面处含有靶序列。

201.根据权利要求198-200中任一项所述的方法,其中所述靶序列含有一个或多个胞嘧啶核苷酸,当被编辑为胸腺嘧啶时,在所述T7 RNA聚合酶-降解决定子融合物的所述T7 RNA聚合酶和降解决定子结构域之间插入终止密码子。

经进化的胞苷脱氨酶及使用其编辑DNA的方法

[0001] 相关申请

[0002] 本申请根据35 U.S.C.§119(e)要求由David R.Liu et al.于2022年8月16日提交的、标题为“Evolved Cytosine Deaminases and Methods of Editing DNA Using the Same”的美国临时申请(U.S.S.N.63/398,483);以及由David R.Liuet al.于2022年10月21日提交的、标题为“Evolved Cytosine Deaminases and Methods of Editing DNA Using the Same”的美国临时申请(U.S.S.N.63/380,523)的优先权。这两项申请通过引用以其整体并入本文。

[0003] 政府支持

[0004] 本发明是在美国国立卫生研究院授予的资助号RM1HG009490、R01EB027793、R01EB031172、R35GM118062和U01AI142756的支持下完成的。政府对本发明拥有特定权利。

[0005] 电子序列表的引用

[0006] 电子序列表(B119570170W000-SEQ-JQM.xml;大小:457,025字节;和创建日期:2023年8月15日)的内容通过引用以其整体并入本文。

背景技术

[0007] 碱基编辑器(BE)是进行体内正向遗传诱变筛选的有用工具,并具有通过精确安装靶标点突变来校正致病性点突变的潜力。BE包含Cas蛋白和碱基修饰酶(例如脱氨酶)的融合物。胞嘧啶碱基编辑器(CBE)将C·G碱基对转化为T·A碱基对,而腺嘌呤碱基编辑器(ABE)将A·T碱基对转化为G·C碱基对。CBE和ABE共同可以介导所有四种可能的转换突变(例如,C至T、A至G、T至C和G至A)。参考国际专利申请号:PCT/US2017/045381(2018年2月8日公布)、国际专利申请号:PCT/US2018/056146(其于2019年4月25日公布为WO 2019/079347)、Koblan et al.,Nat Biotechnol(2018)以及Gaudelli et al.,Nature 551,464-471(2017)。

[0008] 高度活性的天然修饰DNA的胞苷脱氨酶(如APOBEC家族酶)可以脱氨CBE的Cas蛋白结构域生成的R环之外的瞬时暴露的单链DNA片段,导致低水平但广泛的Cas非依赖性基因组修饰[13-15,19]。同样,可以有力接合RNA的高活性胞苷脱氨酶也可以介导不依赖于向导RNA杂交的非期望的RNA脱氨[21]。在使用现有CBE的编辑中观察到的显著Cas非依赖性脱靶DNA和RNA编辑可能会限制这些CBE在需要最小化脱靶编辑的应用中的使用[15]。现有的CBE包括BE3(其包含结构NH_2-[NLS]-[rAPOBEC1脱氨酶]-[Cas9切口酶(D10A)]-[UGI结构域]-[NLS]-COOH)、BE4(其包含结构NH_2-[NLS]-[rAPOBEC1脱氨酶]-[Cas9切口酶(D10A)]-[UGI结构域]-[UGI结构域]-[NLS]-COOH)以及BE4max(其为BE4的版本,其中碱基编辑器编码构建体已针对人类细胞表达进行了密码子优化)。Cas非依赖性脱靶效应源于碱基编辑器与DNA位点的随机缔合,这是由于过表达的碱基编辑器对DNA的固有亲和力。已经发现若干基于TadA*的ABE的Cas非依赖性脱靶DNA编辑检测不到或频率低得多[13],尽管低水平的RNA脱氨仍可以从一些ABE的过表达中检测到[8,9,34]。

[0009] 本领域需要在保持高中靶活性的同时表现出较低的Cas非依赖性脱靶编辑的新型

胞苷脱氨酶和胞嘧啶碱基编辑器。本领域还需要更小尺寸的CBE,例如,尺寸小到足以由单一腺相关病毒(AAV)载体编码(例如,包装容量为约4.7kb)。

发明内容

[0010] 本公开提供了脱氨酶的首次定向进化以选择性地脱氨不同碱基。本公开提供了经工程化以优先脱氨DNA中的胞苷的腺苷脱氨酶的变体。因此,本公开提供了作为腺苷脱氨酶(例如,野生型或工程化的tRNA腺苷脱氨酶(TadA))的变体的胞苷脱氨酶。本公开提供了包含优先脱氨DNA中的胞苷的脱氨酶变体结构域和核酸可编程结合蛋白(napDNAbp)结构域的胞嘧啶碱基编辑器,其中腺苷脱氨酶变体能够以与现有胞苷脱氨酶相似或相同的程度脱氨核酸分子中的胞苷。在一些方面,本公开提供了尺寸最小化的脱氨酶变体,这些变体提供与现有胞嘧啶碱基编辑器(CBE)相比具有降低的脱靶效应同时保持现有CBE的高编辑效率的碱基编辑器。在一些方面,本公开提供了利用本文提供的脱氨酶和碱基编辑器的碱基编辑器、复合物、核酸、载体、细胞、组合物、方法、试剂盒和用途。

[0011] 本公开至少部分基于以下假设:腺苷脱氨酶可以进一步进化为识别胞嘧啶作为底物,并且这种进化可能导致一类新的高选择性胞苷脱氨酶和CBE,具有高编辑效率和较低的脱靶Cas非依赖性DNA和RNA编辑(与天然存在的胞苷脱氨酶相比)。野生型TadA在进化上与胞苷脱氨酶相关。此外,已报道经进化的ABE变体中的低水平的胞苷脱氨[11,31,32]。另外,TadA7.10(TadA-7.10P48R)的诱变显示出破坏腺苷选择性并增加编辑窗口中SpCas9原间隔区位置6处(将SpCas9原间隔区相邻基序PAM计数为位置21-23)的5'-TC背景下的胞苷脱氨[32],尽管在其他背景和位置处腺苷脱氨仍然是优选的。最后,作用于RNA的腺苷脱氨酶(ADAR)已经进化为在RNA中同时进行胞苷脱氨和腺苷脱氨[33]。

[0012] 本公开总体上涉及用于基因编辑的碱基编辑器(BE)。迄今为止报道的碱基编辑器包含与脱氨酶(例如"碱基"修饰结构域)融合的可编程DNA结合蛋白结构域(例如Cas9)。在一些情况下,BE还可以包括额外的结构域,这些结构域改变细胞DNA修复过程以提高所得单核苷酸变化的效率、掺入和/或稳定性。可编程DNA结合结构域引导脱氨酶在向导RNA编程的靶位点直接将一个碱基转化为另一个碱基。迄今为止已开发了两类主要的BE:胞苷BE(CBE),其将C·G转化为T·A;以及腺嘌呤BE(ABE),其将A·T转化为G·C。共同地,CBE和ABE能够校正所有四种类型的转换突变(C至T、G至A、A至G和T至C)。由于已知的疾病相关基因变体中有一半是点突变,而转换突变占已知致病性点突变的约60%,BE被广泛用于研究和治疗各种细胞类型和生物体中的遗传疾病,包括人类遗传疾病的动物模型。

[0013] CBE和ABE可以包括本领域技术人员已知的任何可编程DNA结合结构域。CBE进一步包含配置为脱氨胞苷的脱氨酶;而ABE包含配置为脱氨腺苷的脱氨酶。不希望受任何特定理论的束缚,通常认为当前的CBE包含配置为将胞苷脱氨为尿嘧啶的天然存在的脱氨酶或其变体。相比之下,ABE包含tRNA特异性腺苷脱氨酶,其已进化(使用实验室技术如PACE和PANCE突变)为接受DNA底物,如国际专利申请号PCT/US2021/016827(2021年2月5日提交,通过引用并入本文)中描述的那些,以实现A·T至G·C编辑。迄今为止报告的所有ABE[4,8-10],包括已进入临床试验[2]或已获准进行临床试验[1]的那些,均使用TadA7.10或该脱氨酶的进化或工程化变体。TadA7.10是最先进的ABE(ABE7.10)的腺苷脱氨酶,其在国际公布号WO 2018/027078(2018年8月2日公布)中公开。TadA7.10也是ABEmax的脱氨酶结构域,ABEmax是

ABE7.10已针对人类细胞表达进行了密码子优化的变体。例如,当前一代的ABE变体ABE8e(其含有TadA-8e突变体腺苷脱氨酶)通常比现有的CBE实现更高的编辑效率,尽管野生型TadA对tRNA底物有强烈的偏好[9,11,12]。TadA-8e和ABE8e在国际公布号WO 2021/158921(2021年8月12日公布)中描述。

[0014] ABE相对于其CBE对应物具有若干优势。例如,与大多数CBE脱氨酶相比,TadA酶的加工性较低并因此通常能够实现更高的单核苷酸编辑精度[3,7,8,11]。与CBE相比,ABE还提供了更低水平的Cas非依赖性脱靶编辑[8,9,13-15]。这一优势可能源于与野生型TadA相比(针对tRNA茎的K_m=830nm),常用胞苷脱氨酶与核酸底物的更紧密的非辅助结合(APOBEC1与mRNA结合的米氏常数K_m为0.21nM)。这也可能源于野生型TadA不能加工DNA,以及TadA-8e仅以Cas依赖性方式使用TadA7.10进行进化的事实。基因组挖掘[19]和蛋白质工程化提供了具有较低Cas非依赖性DNA和RNA编辑的替代胞苷脱氨酶,但迄今为止,这些变体存在中靶编辑活性降低和/或尺寸较大的问题[15,20-24]。

[0015] 经进化的TadA腺苷脱氨酶的长度为166个氨基酸,比常用的胞苷脱氨酶诸如APOBEC1(227个氨基酸)、AID(182个氨基酸)[25]、CDA(207个氨基酸)[7]或APOBEC3A(198个氨基酸)[26]明显更小,这使得TadA衍生的碱基编辑器更容易通过尺寸受限的方法和系统(如AAV)递送到细胞中。事实上,TadA的小尺寸使得ABE而非CBE能够使用单一AAV在体内递送到动物组织中[27,28]。

[0016] 本公开的发明人假设,通过定向进化使腺苷脱氨酶进行胞苷脱氨可能产生CBE,这些CBE维持高的中靶活性但继承当前ABE的较低Cas非依赖性脱靶编辑和较小尺寸(例如,使其更容易通过尺寸受限的方法如AAV递送到细胞中)。因此,在一些实施方案中,本公开提供了包含突变腺苷脱氨酶(优先脱氨DNA中的胞苷)和napDNAbp结构域(例如Cas9切口酶)的CBE。本文所述的从TadA脱氨酶进化的胞苷脱氨酶称为“TadA-CD”,并且本文公开的含有TadA-CD的CBE在本文中称为“TadCBE”。

[0017] 因此,本公开的方面涉及CBE,其包含可编程DNA结合蛋白(例如Cas9)和优先脱氨DNA中的嘧啶特别是胞苷的经进化的脱氨酶。例如,所公开的TadA-CD脱氨酶变体表现出约10:1、15:1、20:1或超过20:1的胞苷脱氨与腺嘌呤脱氨的比率。在特定实施方案中,所公开的脱氨酶变体表现出约20:1的胞苷脱氨与腺嘌呤脱氨的比率。本文所述的一种或多种TadA-CD脱氨酶包含多个突变,这些突变位于活性位点附近的环上,对于腺苷至胞苷的转变选择性至关重要。这些突变赋予TadA-CD具有现有ABE中使用的腺苷脱氨酶(如TadA-8e)所表现出的低脱靶编辑频率同时对DNA靶区域中的胞苷具有活性的独特优势。它们还具有尺寸最小化(例如<4.7kb)的优势,这赋予了将含有这些脱氨酶变体的TadCBE以下能力:在单一AAV载体中编码,而不是通过两个内含肽介导的分裂AAV载体编码,或者使用工程化的病毒-脂质颗粒(例如,如本文所述的那些)。在一些实施方案中,TadCBE进一步包含对胞苷碱基编辑活性有用的任何napDNAbp结构域,以及尿嘧啶糖基化酶抑制剂(UGI)结构域。这些TadA-CD变体是通过连续和/或非连续进化方法生成的,包括在TadA-8e底物(或起点)上进行的PACE实验。

[0018] 本公开的其他方面涉及噬菌体辅助进化选择系统(例如PACE和/或PANCE),以增强ABE的腺苷脱氨酶结构域对胞嘧啶的底物特异性(其中ABE含有Cas9或Cas9直系同源物)。在一些实施方案中,选择技术包括用于PACE进化的载体系统,该系统包含低严格性载体和高

严格性载体。另外的方面涉及含有这些载体或所公开载体系统的细胞。例如,在一些实施方案中,高度活性的腺苷脱氨酶TadA-8e通过PACE进行进化(例如,突变)以进行胞苷脱氨。经进化的TadA-CD含有位于脱氨酶活性位点附近d环上的突变,这对于腺苷至胞苷的转变选择性至关重要。

[0019] 与最常用的天然存在的CBE(如BE4max及其变体)相比,所公开的TadCBE提供了相当或更高的中靶活性、更小的尺寸和/或显著更低的Cas非依赖性DNA和RNA脱靶编辑活性,这两者都可以通过引入V106W突变进一步抑制,而不会降低中靶编辑。这些TadCBE可以用于在哺乳动物细胞(如原代人T细胞和造血干细胞和祖细胞)中在治疗相关基因组基因座进行单一或多路复用碱基编辑,如本文所示。其他细胞也是可能的,并在本文别处公开。TadCBE的创建扩展了胞嘧啶碱基编辑器在基因编辑中的实用性。

[0020] 在一些实施方案中,经进化的TadA-CD在SEQ ID NO:41(即TadA-8e脱氨酶)的氨基酸序列中可以包含在残基E27、V28和H96处的突变,并且可以进一步包含在选自R26、M61、Y73、I75、M151、Q154和A158的残基处的至少一个突变,或同源腺苷脱氨酶中的相应突变。示例性同源脱氨酶包括衍生自金黄色葡萄球菌、伤寒沙门氏菌、腐败希瓦氏菌、流感嗜血杆菌、新月柄杆菌和枯草芽孢杆菌中的任一个的TadA脱氨酶。因此,在一些实施方案中,经进化的TadA-CD可以包含SEQ ID NO:317-323、354和355中任一个中的一个或多个突变,这些突变赋予胞苷活性。在一些实施方案中,经进化的TadA-CD可以包含SEQ ID NO:34-40、42-54、33、315和326中任一个中的一个或多个突变,这些突变赋予胞苷活性。本公开的脱氨酶可以从迄今为止报道的具有腺苷脱氨酶活性的任何腺苷脱氨酶进化。

[0021] 在一些实施方案中,所公开的TadA-CD变体包含与TadA-8e的氨基酸序列(SEQ ID NO:41)至少80%、85%、90%、95%、98%、99%或99.5%相同的氨基酸序列,其中对应于SEQ ID NO:41的残基27的氨基酸是除E以外的任何氨基酸。

[0022] 在一些实施方案中,TadA-CD变体包含与SEQ ID NO:41的氨基酸序列至少80%、85%、90%、95%、98%、99%或99.5%相同的氨基酸序列,其中对应于SEQ ID NO:41的残基28的氨基酸是除V以外的任何氨基酸。

[0023] 在其他实施方案中,TadA-CD变体包含与SEQ ID NO:41的氨基酸序列至少80%、85%、90%、95%、98%、99%或99.5%相同的氨基酸序列,其中对应于SEQ ID NO:41的残基96的氨基酸是除H以外的任何氨基酸。

[0024] 在一些实施方案中,所公开的TadA-CD变体包含与SEQ ID NO:34-40中任一个的氨基酸序列至少80%、85%、90%、95%、98%、99%或99.5%相同的氨基酸序列。在其他实施方案中,TadA-CD变体包含SEQ ID NO:34-40中任一个的氨基酸。

[0025] 所公开的TadA-CD变体可以进一步包含V106W突变。在一些实施方案中,V106W突变导致在所评估的靶标上低于或等于1.5%、低于或等于1%、低于或等于0.75%、低于或等于0.5%、低于或等于0.25%、低于或等于0.1%、低于或等于0.05%或低于或等于0.01%的腺嘌呤碱基编辑(上述编辑频率可以代表平均值或最大值)。

[0026] 本公开的其他方面涉及包含可编程DNA结合结构域(例如napDNAbp)和所公开的进化TadA-CD结构域的碱基编辑器。在一些实施方案中,碱基编辑器的napDNAbp是Cas9蛋白,如Cas9切口酶。在一些实施方案中,碱基编辑器的napDNAbp是Nme2Cas9蛋白(如eNme2Cas9切口酶)或Nme2Cas9变体。在一些实施方案中,碱基编辑器的napDNAbp是表6中列出的任何

蛋白。在一些实施方案中,碱基编辑器进一步包含UGI结构域。在一些实施方案中,碱基编辑器进一步包含核定位结构域。因此,本文提供了TadCBE。在另一个方面,本公开描述了复合物,其包含任何所公开的碱基编辑器和与碱基编辑器的napDNAbp结构域结合的向导RNA。

[0027] 在一些方面,本公开涉及TadA衍生的胞苷脱氨酶,其提供靶标胞嘧啶至胸腺嘧啶以及靶标腺嘌呤至鸟嘌呤的高效转化(本文称为"TadA-dual"脱氨酶和碱基编辑器)。TadA-dual脱氨酶能够编辑原间隔区内,并且特别是在原间隔区的编辑窗口内的C和A碱基。这些编辑器以大致相等的效率安装A至G和C至T的编辑(例如,包含TadA-dual的碱基编辑器,SEQ ID NO:39)。

[0028] 在一些实施方案中,TadA-dual脱氨酶相对于TadA-8e(SEQ ID NO.41)发生突变。在一些实施方案中,TadA-dual脱氨酶包含胞苷脱氨酶,该胞苷脱氨酶包含选自R26G、V28A、A48R、Y73S和H96N的一个、两个、三个、四个或五个突变(例如TadA-CDf,SEQ ID NO:39)。

[0029] 在一些实施方案中,TadA-dual脱氨酶相对于TadA-CDf(SEQ ID NO:39)发生突变。在一些实施方案中,TadA-dual脱氨酶在SEQ ID NO:39的氨基酸序列中包含位置N46处的突变。在一些实施方案中,TadA-dual脱氨酶包含与SEQ ID NO:39-54的序列同一性至少80%、至少85%、至少90%、至少95%、至少99%、至少99.5%或至少99.8%相同的氨基酸序列。

[0030] 在一些实施方案中,TadA-dual脱氨酶相对于腺苷对胞嘧啶具有增加的亲和力。例如,在一些实施方案中,双重(dual)编辑器以0.7:1、0.8:1、0.9:1、1:1、1.1:1、1.2:1、1.3:1、1.4:1或1.5:1的比率提供A至G和C至T的编辑。然而,在一些实施方案中,TadA-dual脱氨酶对胞嘧啶具有比对腺苷更高的特异性。

[0031] 在其他实施方案中,TadA-dual(例如SEQ ID NO:39)脱氨酶可以进一步突变(例如使用PANCE和/或PACE)以产生相对于腺苷对胞嘧啶具有增加的亲和力的胞苷脱氨酶。例如,在一些实施方案中,脱氨酶的腺苷脱氨活性与胞苷脱氨活性的比率至少为约0.001:1、0.005:1、0.007:1、0.01:1、0.05:1、0.07:1或0.1:1。

[0032] 本公开的另外方面涉及编码napDNAbp、胞苷脱氨酶及其融合蛋白的多核苷酸、载体和细胞。在一些实施方案中,本公开的碱基编辑器可以在如本文公开的多核苷酸中编码。在一些实施方案中,本公开的脱氨酶变体可以在如本文公开的多核苷酸中编码。在某些实施方案中,所公开的载体包含编码本公开的碱基编辑器中的任一者的多核苷酸。在其他实施方案中,本公开提供了包含本文所述的脱氨酶变体、碱基编辑器、复合物、核酸或载体中的任一者的细胞和组合物。此外,本文还提供了编码任何所公开的碱基编辑器和任选地向导RNA的AAV载体。

[0033] 本公开的其他方面提供了包含本文所述的胞苷脱氨酶或其变体、碱基编辑器、复合物、病毒、核酸和/或载体中的任一者的药物组合物。

[0034] 在一些方面,本公开涵盖包括将核酸分子(例如DNA)与本文所述的碱基编辑器或复合物中的任一者接触的方法。例如,在一些实施方案中,该方法包括将本文所述的BE中的任一者与sgRNA和DNA接触。这些方法中的接触可以在体内、体外或离体进行。

[0035] 其他实施方案描述了使用本文所述的碱基编辑器的方法。在一些实施方案中,该方法包括在DNA编辑中使用(a)本发明的任何碱基编辑器和(b)将(a)的碱基编辑器靶向至双链DNA分子中的C:G核苷酸碱基对的向导RNA。在其他实施方案中,该方法包括使用本发明的碱基编辑器、复合物或药物组合物作为药物。在某些实施方案中,该方法包括使用本发明

的碱基编辑器、复合物或药物组合物作为药物,以治疗疾病、病症或病况(例如镰状细胞病或HIV/AIDS)。

[0036] 在一些实施方案中,本公开提供了选择(例如进化、工程化等)胞嘧啶碱基编辑器的方法。这些方法可以包括通过几轮连续的PACE和/或PANCE进化来进化腺苷碱基编辑器。在某些实施方案中,该方法包括编码与NpuN内含肽融合的突变TadA-8e蛋白的选择噬菌体、编码与dCas9-UGI融合的NpuC内含肽的第一质粒、编码由T7或proT7启动子驱动的gIII并编码sgRNA的第二质粒以及编码T7 RNA聚合酶-降解决定子融合物的第三质粒。

[0037] 在另一个方面,本公开涵盖使用本文所述的载体中的任一者生成本文所述的碱基编辑器中的一种或多种的方法。

[0038] 本公开的进一步方面还涉及包含核酸构建体的试剂盒,该核酸构建体包含(a)编码本文所述的碱基编辑器中的任一者的核酸序列,和(b)编码向导RNA的核酸序列。在一些实施方案中,该核酸构建体进一步包含驱动(a)的序列和/或(b)的序列表达的一个或多个异源启动子。

[0039] 在某些方面,本文所述的碱基编辑器可以施用于受试者以治疗疾病或病症。因此,提供了其中将所述的TadCBE施用于受试者并编辑受试者基因组中的靶序列的方法。靶序列可以包含突变体C:G碱基对,例如与疾病或病症相关的突变体C:G碱基对。在这些方法的各种实施方案中,碱基编辑器的胞苷脱氨程度超过腺苷脱氨程度10倍、15倍、20倍或超过20倍(比率为10:1、15:1、20:1或超过20:1)。

[0040] 本公开还提供了本文所述的碱基编辑器中的任一者和将该碱基编辑器靶向至核酸分子中的靶C:G碱基对的向导RNA在制备用于核酸编辑的试剂盒或组合物中的用途,其中核酸编辑包括在适合C:G核碱基对的胞嘧啶(C)脱氨的条件下使核酸分子与碱基编辑器和向导RNA接触。本公开还提供了本文所述的碱基编辑器中的任一者和将该碱基编辑器靶向至核酸分子中的靶C:G碱基对的向导RNA在制备用于评估碱基编辑器的脱靶效应的试剂盒中的用途。

[0041] 本公开的其他优点和新颖特征将在结合附图考虑本公开的各种非限制性实施方案的详细描述时变得显而易见。

附图说明

[0042] 本公开的非限制性实施方案将通过示例的方式参考附图进行描述,这些附图是示意性的,并不旨在按比例绘制。在附图中,每个相同或几乎相同的组分通常由单个数字表示。为了清晰起见,并非每个组分都在每个图中标记,也不是每个实施方案的每个组分都在不需要图示的情况下显示,以允许本领域普通技术人员理解本公开。在附图中:

[0043] 图1A-1E.胞苷脱氨酶从TadA-8e的噬菌体辅助进化。(图1A)基于TadA的胞苷脱氨酶从tRNA脱氨酶TadA的进化轨迹。(图1B)PACE概述。选择噬菌体(紫色)编码正在进化的蛋白质。大肠杆菌宿主(灰色)含有1)用于多样化噬菌体的诱变质粒(红色)和2)调节pIII表达的质粒系统(蓝色,由gIII编码)。仅具有期望活性的变体触发pIII的产生并繁殖。不具有期望活性的噬菌体不能繁殖并被稀释出培养池(lagoon)。(图1C)胞苷脱氨的选择回路。TadA-8e变体在选择噬菌体(SP,紫色)上编码。大肠杆菌除了诱变质粒外还携带四种辅助质粒以建立选择回路:P1含有碱基编辑器的Cas9-UGI组分。噬菌体感染后,完整的碱基编辑器通过

分裂Npu内含肽系统(黄色)重建。P2编码向导RNA和gIII,其处于T7启动子的转录控制下。P3含有通过与降解决定子标签融合而失活的T7 RNA聚合酶。C·G至T·A编辑活性在T7 RNAP和降解决定子之间插入终止密码子,从而产生活性T7 RNAP,这导致gIII的转录和噬菌体繁殖。(图1D)本文所述的CBE回路的两个版本。在这两种情况下,C·G至T·A编辑在降解决定子标签之前插入终止密码子,导致活性T7 RNAP。较不严格的回路需要在非编码链上的C·G至T·A编辑(上方),并且可以耐受一个非期望A至G编辑。较严格的回路需要编码链上的C·G至T·A编辑,并且不能耐受任何非期望A·T至G·C编辑。(图1E)胞苷脱氨酶从TadA-8e的噬菌体辅助非连续进化。每次PANCE传代中使用的ProD(较强,较不严格)或ProA(较弱,较严格)启动子如所示。在每次传代中,除非另有说明,否则噬菌体被1:50稀释。在若干轮进化后,尽管传代之间的稀释率增加,但噬菌体滴度稳定,表明胞苷脱氨活性的进化。

[0044] 图2A-2D.经进化的TadA*变体催化胞苷脱氨。(图2A)本文进化并表征的TadA-8e变体的汇总。这些变体代表了在九次PANCE传代后或在159小时PACE后的保守突变。有关突变的完整列表,请参见图13-14。(图2B)评估大肠杆菌中靶质粒的碱基编辑的方法。细胞用靶质粒(蓝色)和碱基编辑器质粒(紫色)共转化。碱基编辑器表达用阿拉伯糖诱导。16小时后,收获细胞,并通过高通量测序分析靶质粒。(图2C)大肠杆菌中对选择回路靶位点匹配的原间隔区的碱基编辑。C·G至T·A编辑以蓝色显示。A·T至G·C编辑以洋红色显示。点代表单个生物学重复,并且条形代表来自四个独立生物学重复的平均值±标准差。(图2D)ABE8e的冷冻电镜结构(PDB:6VPC)[18]中经进化的突变的位置。

[0045] 图3.哺乳动物细胞中具有SpCas9结构域的经进化的TadCBE的表征。在BE4max架构或具有2xUGI的ABE8e中使用SpCas9切口酶结构域的指定碱基编辑器与靶向每个图所示原间隔区的九种向导RNA中的每种一起转染。靶胞嘧啶为蓝色,靶腺嘌呤为洋红色,并且PAM序列加下划线。C·G至T·A碱基编辑以蓝色阴影显示。A·T至G·C碱基编辑以洋红色阴影显示。点代表单个值,并且条形代表三个独立生物学重复的平均值±标准差。HEK293T位点3缩写为HEK3,并且HEK293T位点4缩写为HEK4。

[0046] 图4.哺乳动物细胞中具有经进化的eNme2-C Cas9结构域的经进化的脱氨酶的表征。在BE4max架构或具有2xUGI的ABE8e中使用eNme2-C Cas9切口酶结构域(PAM=N_4CN)的指定碱基编辑器与靶向每个图所示原间隔区的六种向导RNA中的每种一起转染。靶胞嘧啶为蓝色,靶腺嘌呤为洋红色,并且PAM序列加下划线。C·G至T·A碱基编辑以蓝色阴影显示。A·T至G·C碱基编辑以洋红色阴影显示。点代表单个值,并且条形代表三个独立生物学重复的平均值±标准差。

[0047] 图5A-5D.TadCBE的碱基编辑窗口和Cas非依赖性脱靶DNA和RNA编辑的表征。(图5A)具有2xUGI的ABE8e、TadCBEa和TadCBEa-V106W在HEK293T中九个不同靶基因组位点间的碱基编辑活性窗口。点代表在原间隔区内指定位置处含有指定碱基的所有位点间的平均编辑。用于此分析的单个数据点在图2A-2D、图14和图16A-16B中。(图5B)使用正交R环测定法测量Cas非依赖性脱靶DNA编辑的方法[15]。(图5C)在由死亡金黄色葡萄球菌Cas9生成的六个正交R环(SaR1-SaR6)内所有胞嘧啶间的平均Cas非依赖性脱靶编辑。(图5D)脱靶RNA编辑。在用指示的碱基编辑器转染后48小时从HEK293T细胞中收获RNA。在cDNA合成后,扩增CTNNB1、IP90和RSL1D1并通过高通量测序进行分析。对于图5C-5D,点代表单个生物学重复,并且条形代表三个独立生物学重复的平均值±标准差。

[0048] 图6.原发性人类T细胞和造血干细胞及祖细胞中通过TadCBE在治疗相关基因座处的碱基编辑。编码指示碱基编辑器或作为阴性对照的GFP的mRNA与在指定原间隔区处靶向(上方)CXCR4或(中部)CCR5的合成向导RNA一起电穿孔到人T细胞(n＝4个供体)中。靶胞嘧啶为蓝色,靶腺嘌呤为洋红色,并且PAM序列加下划线。3天后,从T细胞裂解物中收获基因组DNA并通过高通量测序进行分析。灰色框指示CXCR4和CCR5中终止密码子安装的期望位置。在胞嘧啶碱基编辑后产生TAG(CXCR4)和TAA(CCR5)终止密码子的靶向胞苷加下划线。下方图显示,编码指示碱基编辑器或作为阴性对照的GFP的mRNA与靶向BCL11A增强子的合成向导RNA一起电穿孔到造血干细胞及祖细胞中。3天后,从细胞裂解物中收获基因组DNA并通过高通量测序进行分析。C·G至T·A碱基编辑以蓝色阴影显示,A·T至G·C碱基编辑以洋红色阴影显示。点代表单个生物学重复,并且条形代表n＝4个供体(上方和中部)或n＝3个供体(下方)的平均值±标准差。

[0049] 图7.PACE和PANCE回路中脱氨选择性选择的基础。在回路1中,只有当碱基编辑器对A_7和A_8两者进行脱氨时,终止密码子的形成才受到阻碍。因此,回路1耐受适度水平的A脱氨。在回路2中,单一腺嘌呤A_6的脱氨将阻止终止密码子形成并阻碍回路激活和噬菌体繁殖。因此,回路2对于针对腺苷脱氨的选择更为严格。

[0050] 图8A和8B.PANCE滴度和经进化的TadA-CD基因型。(图8A)培养池1-7在PANCE期间的噬菌体滴度。通过从ProD(最强,最不严格)到ProA(最弱,最严格)增加启动子强度、增加稀释倍数以及通过从回路1切换到回路2来调制严格性。培养池1-6接种有编码TadA8e-NpuN的噬菌体,而培养池7接种有编码TadA8e A48R-NpuN的噬菌体。(图8B)各种PANCE培养池(L1-L7)在PANCE后的基因型。

[0051] 图9A-9C.PACE滴度和经进化的TadA-CD基因型。(图9A)PACE期间的噬菌体滴度和培养池流速。培养池1在t＝43小时后在S2060细胞中显示出活性非依赖性繁殖,表明噬菌体进化出选择非依赖性复制,并且未继续。(图9B)在选择非依赖性繁殖出现之前,在t＝43小时时来自培养池1的经进化的TadA*变体的基因型。(图9C)在不同时间点,来自培养池2的经进化的TadA*变体的基因型。

[0052] 图10A-10C.TadA-CDa的AlphaFold模型。(图10A)显示了ABE8e的冷冻电镜结构(PDB ID 6VPC)[1]与含有腺苷的8-azancbularine(8Az)底物模拟物的DNA结合。Val 28(洋红色)支持腺嘌呤底物相对于催化锌的正确定位。(图10B)使用Chimera软件[2]中的“Swapna”功能将8Az替换为胞苷。在所得模型中,胞嘧啶的C4(其在脱氨期间针对亲核攻击进行靶向)距离8Az的靶标碳约1 Å,因此可能需要DNA底物的移动以实现生产性催化。Val 28可以阻碍DNA底物向TadA-8e口袋深处的这种移动。(图10C)使用AlphaFold[3]以生成经进化的TadA-CDa的模型。将ABE8e结构叠加以生成具有来自6VPC的DNA底物R环的模型。经进化的酶预计与TadA8e相比不会在二级结构中采用任何明显差异。TadA-8e中Val 28至TadA-CD中发现的较小的Ala或Gly残基的进化替换可能缓解预计会阻碍胞嘧啶中靶C4相对于催化锌离子的生产性定位的空间限制。

[0053] 图11.在九个基因组靶位点处通过SpCas9变体的插入缺失和C·G至G·C编辑。在BE4max架构或具有2xUGI的ABE8e中使用SpCas9切口酶结构域的指定碱基编辑器与靶向每个图中所示原间隔区的九种向导RNA中的每种一起转染到HEK293T细胞中。C·G至G·C碱基编辑以蓝色阴影显示。插入缺失以灰色显示。点代表单个值,并且条形代表三个独立生物学

重复的平均值±标准差。相应的中靶编辑数据可见于图3。

[0054] 图12.在六个基因组靶位点处通过eNme2-C Cas9变体的插入缺失和C·G至G·C编辑。在BE4max架构或具有2xUGI的ABE8e中使用eNme2-C Cas9切口酶结构域的指定碱基编辑器与靶向每个图中所示原间隔区的六种向导RNA中的每种一起转染到HEK293T细胞中。C·G至G·C碱基编辑以蓝色阴影显示。插入缺失以灰色显示。点代表单个值,并且条形代表三个独立生物学重复的平均值±标准差。相应的中靶数据在图4中。

[0055] 图13.V106W对TadA-CD突变的接近性。TadA-CD的进化期间生成的突变以蓝色显示。残基V106以红色显示。将V106W添加到TadA-7.10、TadA-8e和TadA-8.17减少脱靶编辑活性[4-6]。将V106W添加到TadCBEa-e增加相对于腺苷对胞苷脱氨的选择性,并减少脱靶编辑活性。

[0056] 图14.在六个基因组靶位点处通过V106W变体的碱基编辑。在BE4max架构或具有2xUGI的ABE8e中使用SpCas9切口酶结构域的指定碱基编辑器与靶向每个图中所示原间隔区的六种向导RNA中的每种一起转染到HEK293T细胞中。靶胞嘧啶为蓝色,靶腺嘌呤为洋红色,并且PAM序列加下划线。C·G至T·A碱基编辑以蓝色阴影显示。A·T至G·C碱基编辑以洋红色阴影显示。点代表单个值,并且条形代表三个独立生物学重复的平均值±标准差。

[0057] 图15.在六个基因组靶位点处通过V106W变体的插入缺失和C·G至G·C编辑。在BE4max架构或具有2xUGI的ABE8e中使用SpCas9切口酶结构域的指定碱基编辑器与靶向每个图中所示原间隔区的六种向导RNA中的每种一起转染到HEK293T细胞中。C·G至G·C碱基编辑以蓝色阴影显示。插入缺失以灰色显示。点代表单个值,并且条形代表三个独立生物学重复的平均值±标准差。

[0058] 图16A-16B.TadA-CD(V106W)变体在三个额外基因组靶位点处的碱基编辑、插入缺失形成和C·G至G·C编辑。在BE4max架构或具有2xUGI的ABE8e中使用SpCas9切口酶结构域的指定碱基编辑器与靶向每个图中所示原间隔区的三种向导RNA中的每种一起转染到HEK293T细胞中。(图16A)靶胞嘧啶为蓝色,靶腺嘌呤为洋红色,并且PAM序列加下划线。C·G至T·A碱基编辑以蓝色阴影显示。A·T至G·C碱基编辑以洋红色阴影显示。(图16B)C·G至G·C碱基编辑以蓝色阴影显示。插入缺失以灰色显示。点代表单个值,并且条形代表三个独立生物学重复的平均值±标准差。

[0059] 图17.CBE在九个基因组靶位点间的碱基编辑活性窗口。点代表在原间隔区内指定位置含有指定碱基的所有位点间的平均编辑。用于此分析的单个数据点在图2A-2D、图14和图16A-16B中。

[0060] 图18.在Cas非依赖性R环编辑实验中EMX1的中靶编辑。在BE4max架构或具有2xUGI的ABE8e中使用SpCas9切口酶结构域的指定碱基编辑器与靶向EMX1的SpCas9向导RNA以及指定的SaCas9 sgRNA一起转染到HEK293T细胞中。显示了指示的碱基编辑器在EMX1中C_5和C_6的平均中靶C·G至T·A碱基编辑。点代表单个值,并且条形代表三个独立生物学重复的平均值±标准差。相应的Cas依赖性脱靶数据在图5C、图19和图20中显示。

[0061] 图19.在由SaCas9生成的六个正交R环内单个位点处的Cas非依赖性脱靶C·G至T·A编辑。正交R环测定在BE4max架构中的CBE变体上进行[7]。细胞用碱基编辑器和靶向EMX1基因座的一个SpCas9 sgRNA以及正交死亡SaCas9和对应于Sa位点1-6的一个SaCas9 sgRNA一起转染。点代表单个生物学重复,并且条形代表三个独立生物学重复的平均值±标准差。

相应的中靶数据在图18中。

[0062] 图20.在由SaCas9生成的六个正交R环内单个位点处通过TadCBEe V106W的Cas非依赖性脱靶C·G至T·A编辑。正交R环测定在BE4max架构中的CBE变体上进行。细胞用碱基编辑器和靶向EMX1基因座的一个SpCas9 sgRNA以及正交死亡SaCas9和对应于Sa位点1-6的一个SaCas9sgRNA一起转染。点代表单个生物学重复,并且条形代表三个独立生物学重复的平均值±标准差。

[0063] 图21A-21C.在六个基因组SaCas9 R环处通过TadCBEe V106W的Cas非依赖性脱靶DNA编辑。正交R环测定在BE4max架构中的CBE变体上进行。细胞用碱基编辑器和靶向EMX1基因座(中靶)的一个SpCas9 sgRNA以及正交死亡SaCas9和对应于Sa位点1-6(SaR1-SaR6)的一个SaCas9 sgRNA一起转染。图21A显示了EMX1基因座处的中靶编辑。图21B显示了图中所描绘的指示原间隔区内所有腺嘌呤间的平均C·G至T·A碱基编辑。(图21C)图中描绘了指示的原间隔区内所有腺嘌呤间的平均A·T至G·C碱基编辑。点代表单个生物学重复,并且条形代表三个独立生物学重复的平均值±标准差。

[0064] 图22.在六个基因组SaCas9 R环处的Cas非依赖性脱靶DNA编辑。正交R环测定在BE4max架构中的CBE变体上进行。细胞用碱基编辑器和靶向EMX1基因座(中靶)的一个SpCas9 sgRNA以及正交死亡SaCas9和对应于Sa位点1-6(SaR1-SaR6)的一个SaCas9 sgRNA一起转染。图中描绘了指示的原间隔区内所有腺嘌呤间的平均A·T至G·C碱基编辑。点代表单个生物学重复,并且条形代表三个独立生物学重复的平均值±标准差。

[0065] 图23A和23B.针对TadCBEe V106W在三种转录物间检验的所有胞嘧啶和腺嘌呤的Cas非依赖性脱靶RNA编辑。在用指示的碱基编辑器转染后48小时,从HEK293T细胞中收获总RNA。在cDNA合成后,扩增CTNNB1、IP90和RSL1D1并通过高通量测序进行分析。同时,从平行转染的另一板中收获基因组DNA。分析基因组DNA中EMX1的中靶编辑作为碱基编辑器活性的对照。图23A显示了与RNA编辑分析对应的样品中EMX1的中靶编辑。图23B显示了平均C至U(蓝色阴影)或A至I(洋红色阴影)。点代表单个生物学重复,并且条形代表三个独立生物学重复的平均值±标准差。

[0066] 图24.RNA脱靶编辑实验中EMX1的中靶编辑。指示的碱基编辑器在两个平行板中转染到HEK293T细胞中。在一个板中,用指示的碱基编辑器转染后48小时从HEK293T细胞中收获RNA并如图23A-23B所述进行分析。同时,从平行转染的另一板中收获基因组DNA。分析基因组DNA中EMX1的中靶编辑作为碱基编辑器活性的对照。点代表单个生物学重复,并且条形代表三个独立生物学重复的平均值±标准差。

[0067] 图25.HEK3的已知脱靶位点的Cas依赖性编辑。在BE4max架构或具有2xUGI的ABE8e中使用SpCas9切口酶结构域的指定碱基编辑器与靶向HEK293T位点3(HEK3)的向导RNA一起转染到HEK293T细胞中。转染后72小时,收获基因组DNA并使用表2A-2E中的引物扩增已知脱靶位点。C·G至T·A碱基编辑以蓝色阴影显示。A·T至G·C碱基编辑以洋红色阴影显示。点代表单个值,并且条形代表三个独立生物学重复的平均值±标准差。

[0068] 图26.HEK4的已知脱靶位点的Cas依赖性编辑。在BE4max架构或具有2xUGI的ABE8e中使用SpCas9切口酶结构域的指定碱基编辑器与靶向HEK293T位点4(HEK4)的向导RNA一起转染到HEK293T细胞中。转染后72小时,收获基因组DNA并使用表2A-2E中的引物扩增已知脱靶位点。C·G至T·A碱基编辑以蓝色阴影显示。A·T至G·C碱基编辑以洋红色阴影显示。点

代表单个值,并且条形代表三个独立生物学重复的平均值±标准差。

[0069] 图27A-27B.EMX1的已知脱靶位点的Cas依赖性编辑。在BE4max架构或具有2xUGI的ABE8e中使用SpCas9切口酶结构域的指定碱基编辑器与靶向EMX1的向导RNA一起转染到HEK293T细胞中。转染后72小时,收获基因组DNA并使用表2A-2E中的引物扩增已知脱靶位点。C·G至T·A碱基编辑以蓝色阴影显示。A·T至G·C碱基编辑以洋红色阴影显示。点代表单个值,并且条形代表三个独立生物学重复的平均值±标准差。相应的中靶数据在图35中显示。

[0070] 图28.BCL11A的已知脱靶位点的Cas依赖性编辑。在BE4max架构或具有2xUGI的ABE8e中使用SpCas9切口酶结构域的指定碱基编辑器与靶向BCL11A的向导RNA一起转染到原代人CD34阳性造血干细胞及祖细胞(n=3个供体)中。转染后72小时,收获基因组DNA并使用表2A-2E中的引物扩增已知脱靶位点。C·G至T·A碱基编辑以蓝色阴影显示。A·T至G·C碱基编辑以洋红色阴影显示。点代表单个值,并且条形代表三个独立生物学重复的平均值±标准差。

[0071] 图29A-29B.靶向CXCR4和CCR5的T细胞实验的C·G至G·C编辑和插入缺失。编码指示的碱基编辑器或作为阴性对照的GFP的mRNA与在指定原间隔区处靶向(图29A)CXCR4或(图29B)CCR5的两种合成向导RNA一起电穿孔到原代人T细胞(n=4个供体)中。3天后,从T细胞裂解物中收获基因组DNA并通过高通量测序进行分析。C·G至G·C碱基编辑以蓝色阴影显示。插入缺失以灰色显示。点代表单个值,并且条形代表三个独立生物学重复的平均值±标准差。

[0072] 图30.靶向CXCR4和CCR5的T细胞实验的Cas依赖性脱靶编辑。编码指示的碱基编辑器或作为阴性对照的GFP的mRNA与在指定原间隔区处靶向CXCR4或CCR5的两种合成向导RNA一起电穿孔到原代人T细胞(n=4个供体)中。3天后,从T细胞裂解物中收获基因组DNA并使用表2A-2E中的引物扩增已知脱靶位点。C·G至T·A碱基编辑以蓝色阴影显示。A·T至G·C碱基编辑以洋红色阴影显示。点代表单个值,并且条形代表三个独立生物学重复的平均值±标准差。

[0073] 图31A-31B.造血干细胞及祖细胞中编辑BCL11A的C·G至G·C编辑、插入缺失和Cas依赖性脱靶编辑。编码指示的碱基编辑器或作为阴性对照的GFP的mRNA与在指定原间隔区处靶向BCL11A的合成向导RNA一起电穿孔到CD34阳性人造血干细胞及祖细胞(n=3个供体)中。3天后,从细胞裂解物中收获基因组DNA并通过高通量测序进行分析。(图31A)C·G至G·C碱基编辑以蓝色阴影显示。插入缺失以灰色显示。(图31B)使用表2A-2E中列出的引物扩增已知的Cas依赖性脱靶位点。C·G至G·C碱基编辑以蓝色阴影显示。A·T至G·C碱基编辑以洋红色阴影显示。点代表单个值,并且条形代表三个独立生物学重复的平均值±标准差。

[0074] 图32A-32C.使用基因组整合mESC靶序列文库对TadCBE的表征。图32A显示了通过编辑文库分析的碱基编辑器的整体效率和选择性。数据显示了在原间隔区位置-9到20之间所有文库成员间的编辑测序读段的平均分数,其中位置21-23是PAM。图32B显示了BE4max、TadCBEa-d、TadCBEd V106W和双重碱基编辑器TadDE在10,683个基因组整合靶位点间的编辑谱。编辑窗口定义为平均编辑效率≥峰值编辑效率的30%的原间隔区位置。在文库实验中测试的所有变体的窗口图可见于图39。图32C显示了通过进行编辑效率回归确定的

TadCBEd和TadCBEd V106W对于胞嘧啶和腺嘌呤碱基编辑结果的序列基序。序列基序的不透明度与保留序列集上的测试R成正比。用于所有变体的完整的序列基序图在图41A和41B中显示。

[0075] 图33.测试TadCBE中的单个突变。大肠杆菌中与选择回路靶位点相匹配的原间隔区的碱基编辑。细胞用靶质粒和碱基编辑器质粒共转化。用阿拉伯糖诱导碱基编辑器的表达。16小时后,收获细胞,并通过高通量测序分析靶质粒。(上图)通过进化鉴定出的单个突变添加到ABE8e中不足以用于生成CBE。(中图)TadCBEa-c和TadCBEe中的突变的分析。ABE8e环区中的突变赋予了对胞苷脱氨的选择性,同时辅助突变加强活性。(下图)TadCBEd中的突变的分析。C·G至T·A编辑以蓝色显示。A·T至G·C编辑以洋红色显示。点代表单个生物学重复,并且条形代表四个独立生物学重复的平均值±标准差。

[0076] 图34.TadCBE的回归分析。大肠杆菌中与选择回路靶位点相匹配的原间隔区的碱基编辑。细胞用靶质粒和碱基编辑器质粒共转化。用阿拉伯糖诱导碱基编辑器的表达。16小时后,收获细胞并通过高通量测序分析靶质粒。(上图)显示的突变是相对于TadCBEa(灰色框)。(下图)显示的突变是相对于TadCBEe(灰色框)。C·G至T·A编辑以蓝色显示。A·T至G·C编辑以洋红色显示。点代表单个生物学重复,并且条形代表四个独立生物学重复的平均值±标准差。

[0077] 图35.EMX1的中靶编辑。在BE4max架构或具有2xUGI的ABE8e中使用SpCas9切口酶结构域的指定碱基编辑器与靶向EMX1的向导RNA一起转染到HEK293T细胞中。转染后72小时,收获基因组DNA并使用表1中的引物扩增已知脱靶位点。C·G至T·A碱基编辑以蓝色阴影显示。A·T至G·C碱基编辑以洋红色阴影显示。点代表单个值,并且条形代表三个独立生物学重复的平均值±标准差。相应的脱靶数据在图27A和27B中显示。

[0078] 图36.通过TadCBE V106W对EMX1的中靶和脱靶编辑。在BE4max架构或具有2xUGI的ABE8e中使用SpCas9切口酶结构域的指定碱基编辑器与靶向EMX1的向导RNA一起转染到HEK293T细胞中。转染后72小时,收获基因组DNA并使用表4中的引物扩增已知脱靶位点。C·G至T·A碱基编辑以蓝色阴影显示。A·T至G·C碱基编辑以洋红色阴影显示。点代表单个值,并且条形代表三个独立生物学重复的平均值±标准差。

[0079] 图37.mESC文库实验的示意图。数千对sgRNA和相应的靶位点被整合到mESC中并用碱基编辑器处理。含有碱基编辑器的细胞通过抗生素选择富集,并且将文库盒扩增用于高通量测序。

[0080] 图38.mESC文库实验中重复之间的相关性。每个重复在每个靶位点处的未校正C·G至T·A编辑效率。红色虚线是总最小二乘回归线。

[0081] 图39.mESC文库编辑实验中TadCBE V106W变体的编辑窗口。编辑窗口定义为原间隔区内的位置,其中在该位置处的转化碱基的平均分数为最大编辑位置处的平均编辑的至少30%。C·G至T·A碱基编辑以蓝色显示。A·T至G·C碱基编辑以红色显示。

[0082] 图40A和40B.mESC文库实验中V106W对峰值编辑的效应。图40A显示了使用TadCBEd(有和没有V106W置换)对在原间隔区位置6处含有胞嘧啶的每个文库成员的C·G至T·A编辑效率。红色虚线是总最小二乘回归线。图40B显示了使用TadCBEd(有和没有V106W)对在原间隔区位置6处含有腺嘌呤的每个文库成员的A·T至G·C编辑效率。红色虚线是总最小二乘回归线。

[0083] 图41A和41B.TadCBE的背景偏好序列基序。通过进行编辑效率回归得到的碱基编辑活性的序列基序。标志的不透明度与保留测试集上的R成正比。提供了C·G至T·A碱基编辑(图41A)和A·T至G·C碱基编辑(图41B)的图。

[0084] 图42.具有经进化的eNme2-C Cas9结构域的经进化的脱氨酶的表征。在BE4max架构或具有2xUGI的ABE8e中使用eNme2-C Cas9切口酶结构域(PAM=N_4CN)的指定碱基编辑器与靶向每个图所示的原间隔区的六种向导RNA中的每种一起转染到HEK293T细胞中。靶胞嘧啶为蓝色,靶腺嘌呤为洋红色,并且PAM序列加下划线。C·G至T·A碱基编辑以蓝色阴影显示。A·T至G·C碱基编辑以洋红色阴影显示。点代表单个值,并且条形代表三个独立生物学重复的平均值±标准差。

[0085] 图43.在六个基因组靶位点处通过eNme2-C Cas9变体的插入缺失和C·G至G·C编辑。在BE4max架构或具有2xUGI的ABE8e中使用eNme2-C Cas9切口酶结构域的指定碱基编辑器与靶向每个图所示的原间隔区的六种向导RNA中的每种一起转染到HEK293T细胞中。C·G至G·C碱基编辑以蓝色阴影显示。插入缺失以灰色显示。点代表单个值,并且条形代表三个独立生物学重复的平均值±标准差。相应的中靶数据在图42中显示。

[0086] 图44.具有SaCas9结构域的经进化的脱氨酶的表征。在BE4max架构或具有2xUGI的ABE8e中使用SaCas9切口酶结构域(PAM=NNGRRT)的指定碱基编辑器与靶向每个图所示的原间隔区的九种向导RNA中的每种一起转染到HEK293T细胞中。靶胞嘧啶为蓝色,靶腺嘌呤为洋红色,并且PAM序列加下划线。C·G至T·A碱基编辑以蓝色阴影显示。A·T至G·C碱基编辑以洋红色阴影显示。点代表单个值,并且条形代表三个独立生物学重复的平均值±标准差。

[0087] 图45.在九个基因组靶点处通过SaCas9变体的插入缺失和C·G至G·C编辑。在BE4max架构或具有2xUGI的ABE8e中使用SaCas9切口酶结构域的指定碱基编辑器与靶向每个图所示的原间隔区的六种向导RNA中的每种一起转染到HEK293T细胞中。C·G至G·C碱基编辑以蓝色阴影显示。插入缺失以灰色显示。点代表单个值,并且条形代表三个独立生物学重复的平均值±标准差。相应的中靶数据如图44所示。

[0088] 图46.在哺乳动物细胞中具有SpCas9的TadDE的表征。在BE4max架构或具有2xUGI的ABE8e中使用SaCas9切口酶结构域(PAM=NGG)的指定碱基编辑器与靶向每个图所示的原间隔区的九种向导RNA中的每种一起转染。靶胞嘧啶为蓝色,靶腺嘌呤为洋红色,并且PAM序列加下划线。C·G至T·A碱基编辑以蓝色阴影显示。A·T至G·C碱基编辑以洋红色阴影显示。点代表单个值,并且条形代表三个独立生物学重复的平均值±标准差。

[0089] 图47.在九个基因组靶点处通过SpCas9变体的插入缺失和C·G至G·C编辑。在BE4max架构或具有2xUGI的ABE8e中使用SaCas9切口酶结构域的指定碱基编辑器与靶向每个图所示的原间隔区的六种向导RNA中的每种一起转染到HEK293T细胞中。C·G至G·C碱基编辑以蓝色阴影显示。插入缺失以灰色显示。点代表单个值,并且条形代表三个独立生物学重复的平均值±标准差。相应的中靶数据在图46中。

[0090] 图48.靶向CXCR4和CCR5的T细胞实验中V106W变体的中靶编辑。编码指示的碱基编辑器或作为阴性对照的GFP的mRNA与在指定原间隔区处靶向(上方)CXCR4或(下方)CCR5的两个合成向导RNA一起电穿孔到人T细胞(n=4个供体)中。靶胞嘧啶为蓝色,靶腺嘌呤为洋红色,并且PAM序列加下划线。3天后,从T细胞裂解物中提取基因组DNA并通过高通量测序进

行分析。灰色框指示CXCR4和CCR5中终止密码子安装的期望位置。通过胞嘧啶碱基编辑产生TAG(CXCR4)和TAA(CCR5)终止密码子的靶向胞苷加下划线。点代表单个值,并且条形代表三个独立生物学重复的平均值±标准差。

[0091] 图49.使用TadCBEe V106W变体靶向CXCR4和CCR5的T细胞实验中的C·G至G·C编辑和插入缺失。编码指示的碱基编辑器或作为阴性对照的GFP的mRNA与在指定原间隔区处靶向(上方)CXCR4或(下方)CCR5的两个合成向导RNA一起电穿孔到原代人T细胞(n=4个供体)中。3天后,从T细胞裂解物中提取基因组DNA并通过高通量测序进行分析。C·G至G·C碱基编辑以蓝色阴影显示。插入缺失以灰色显示。点代表单个值,并且条形代表三个独立生物学重复的平均值±标准差。

[0092] 图50.使用TadCBEe V106W变体靶向CXCR4和CCR5的T细胞实验中的Cas依赖性脱靶编辑。编码指示的碱基编辑器或作为阴性对照的GFP的mRNA与在指定原间隔区处靶向(上方)CXCR4或(下方)CCR5的两个合成向导RNA一起电穿孔到原代人T细胞(n=4个供体)中。3天后,从T细胞裂解物中收获基因组DNA并使用表4中的引物扩增已知的脱靶位点。C·G至T·A碱基编辑以蓝色阴影显示。A·T至G·C碱基编辑以洋红色阴影显示。点代表单个值,并且条形代表三个独立生物学重复的平均值±标准差。

[0093] 图51A-51F.用于在疾病相关位点安装终止密码子的活性和选择性胞嘧啶碱基编辑器的预言性用途。残留的A至G编辑阻止了正确的终止密码子安装(图51A)。胞嘧啶碱基编辑器从TadA双重碱基编辑器(TadA-DE)的进化的示意图(图51B)。描述噬菌体辅助连续进化或PACE的示意图(左)和根据一些实施方案使用的选择回路(右)。在一些实施方案中,使用选择回路和诱变质粒(红色)的大肠杆菌宿主细胞的连续流被编码部分脱氨酶的选择噬菌体(SP)感染。在该选择回路的特定实施方案中,噬菌体传播与gIII(P2)的表达相关联,gIII只能在活性T7 RNA聚合酶存在下转录。在一些实施方案中,T7 RNA聚合酶(P3)与C端降解决定子融合,脱氨酶必须进行C至U编辑以在降解决定子之前安装终止密码子,从而产生活性T7 RNA聚合酶。在噬菌体感染的情况下,使用分裂内含肽系统(P1)完成完整的脱氨酶,并且可以在脱氨酶上发生突变。有益的突变导致培养池中的噬菌体繁殖和富集,而适应性较差的噬菌体无法繁殖并随后被恒定流出物冲走(图51C)。来自TadA-DE的活性和选择性胞嘧啶碱基编辑器的进化轨迹。对TadA-DE进行噬菌体辅助非连续进化(PANCE),直到噬菌体滴度增加,尽管稀释倍数和启动子强度的严格性增加,表明发生了有益的突变。所得变体在脱氨酶的N46位置处识别出保守突变,因此在N46位置处构建了NNK文库,并对这些变体进行了PANCE。为了更进一步增加严格性,对来自两个PANCE的所得变体进行PACE>100小时。稀释倍数在右侧y轴上指示(图51D)。ABE8e的冷冻电镜结构(PDB:6VPC)具有标记的新的保守突变(图51E)。

[0094] 图52A-52E.PANCE后来自PANCE培养池(L1-L2)的基因型(图52B)。使用N46处的NNK文库进行PANCE后来自PANCE培养池(L1-L3)的基因型(图52B)。使用N46处的NNK文库进行PANCE后来自PANCE培养池(L1)的基因型(图52C)。使用N46处的NNK文库进行PANCE后来自PANCE培养池(L1)的不同时间点的基因型(图52D)。使用N46处的NNK文库进行PANCE后来自PANCE培养池(L2)的不同时间点的基因型(图52E)。图52A中显示了选定的序列。

[0095] 图53.在大肠杆菌中TadCBE的活性和序列背景特异性的谱分析。条形指示CBE变体当在底物文库上测试时的平均活性,底物文库经设计以含有原间隔区位置6处的靶碱基(A

或C),其中5'和3'碱基变化为A、T、C或G。每个点代表给定序列背景中含有指定编辑的测序读段的百分比。点根据碱基的5'背景着色(A,红色;C,绿色;G,蓝色;T,黄色)。新进化突变中的突变相对于TadDE列出。TadDE＝TadA8e R26G V28A A48 Y73S H96N。虽然eTdCBEmax、CBET-1.52、TadCBEd、TadDE N46I Y73P和TadDE N46C Y73P取决于序列背景显示出较低的活性,但经进化的变体TadDE N46V Y73P和TadDE N46L Y73P无论序列背景如何都显示出超过80％的编辑。

[0096] 图54.在哺乳动物细胞中经进化的活性和选择性胞嘧啶碱基编辑器与现有胞嘧啶碱基编辑器的比较。TadDE N46变体与在BE4max架构中具有SpCas9切口酶的现有胞嘧啶碱基编辑器以及靶向三个原间隔区的向导RNA一起转染到HEK293T细胞中。TadDE N46变体显示出相当的中靶活性,且无残留的A至G编辑。点代表独立生物学重复的单个值。PAM序列加下划线。HEK293T位点2缩写为HEK2,并且HEK293T位点4缩写为HEK4。TadDE N46变体与在BE4max架构中具有eNme-Cas9切口酶的现有胞嘧啶碱基编辑器以及靶向两个原间隔区的向导RNA一起转染到HEK293T细胞中。TadDE N46变体显示出更高或相当的中靶活性,且无残留的A至G编辑。点代表独立生物学重复的单个值。PAM序列加下划线。

[0097] 图55.通过TadCBE的Cas9非依赖性和RNA脱靶编辑。由死亡金黄色葡萄球菌Cas9生成的四个正交R环(SaR1-SaR4)中所有胞嘧啶间的平均Cas9非依赖性脱靶编辑。新进化突变中的突变是相对于TadDE列出。TadDE N46变体显示出与TadCBEd相比类似的脱靶编辑。点代表独立生物学重复的单个值(图55A)。脱靶RNA编辑。TadDE N46变体显示出与TadCBEd相比类似的脱靶编辑。点代表独立生物学重复的单个值(图55B)。

[0098] 图56.在HEK293T细胞中通过TadCBE在治疗相关基因座处的终止密码子安装。TadCBE用于在PCSK9中安装终止密码子,这是一种正在探索的降低血液胆固醇的治疗策略。灰色框表示终止密码子安装的期望位置。新进化突变中的突变是相对于TadDE列出。来自TadCBEd的残留A至G编辑导致终止密码子擦除,表明TadDE N46变体中残留A至G编辑的缺乏对于终止密码子安装至关重要。点代表独立生物学重复的单个值。PAM序列加下划线。

[0099] 图57.HEK3的已知脱靶位点的中靶和Cas依赖性编辑。TadDE N46变体与在BE4max架构中具有SpCas9切口酶的现有胞嘧啶碱基编辑器以及靶向HEK3的向导RNA一起转染到HEK293T细胞中。新进化突变中的突变是相对于TadDE列出。TadDE N46变体显示出与TadCBEd相比类似的脱靶编辑。点代表独立生物学重复的单个值。

[0100] 图58.HEK4的已知脱靶位点的中靶和Cas依赖性编辑。TadDE N46变体与在BE4max架构中具有SpCas9切口酶的现有胞嘧啶碱基编辑器以及靶向HEK4的向导RNA一起转染到HEK293T细胞中。新进化突变中的突变是相对于TadDE列出。TadDE N46变体显示出与TadCBEd相比类似的脱靶编辑。点代表独立生物学重复的单个值。

[0101] 图59.EMX1的已知脱靶位点的靶向和Cas依赖性编辑。TadDE N46变体与在BE4max架构中具有SpCas9切口酶的现有胞嘧啶碱基编辑器以及靶向EMX1的向导RNA一起转染到HEK293T细胞中。新进化突变中的突变是相对于TadDE列出。TadDE N46变体显示出与TadCBEd相比类似的脱靶编辑。点代表独立生物学重复的单个值。

[0102] 图60.BCL11a的已知脱靶位点的靶向和Cas依赖性编辑。TadDE N46变体与在BE4max架构中具有SpCas9切口酶的现有胞嘧啶碱基编辑器以及靶向BCL11a的向导RNA一起转染到HEK293T细胞中。新进化突变中的突变是相对于TadDE列出。TadDE N46变体显示出与

TadCBEd相比类似的脱靶编辑。点代表独立生物学重复的单个值。

[0103] 图61.与Cas非依赖性脱靶编辑相关的在EMX1处的中靶编辑。TadDE N46变体与在BE4max架构中具有SpCas9切口酶的现有胞嘧啶碱基编辑器以及靶向EMX1的SpCas9向导RNA与SaCas9向导RNA一起转染到HEK293T细胞中。新进化突变中的突变是相对于TadDE列出。点代表独立生物学重复的单个值。

[0104] 图62.与RNA脱靶编辑相关的在EMX1处的中靶编辑。TadDE N46变体与在BE4max架构中具有SpCas9切口酶的现有胞嘧啶碱基编辑器一起转染到在两个板中的HEK293T细胞中。在一个板中,转染后48小时收获RNA,并且在另一个板中,收获基因组DNA。针对EMX1的中靶编辑分析基因组DNA。新进化突变中的突变是相对于TadDE列出。TadDE N46变体显示出与TadCBEd相比类似的脱靶编辑。点代表独立生物学重复的单个值。

[0105] 图63.图53的延续。图53中的每个图也在图63中表示;然而,图53中的每个数据点(表示为点)在图63中显示为条形。

[0106] 定义

[0107] 如本文和权利要求中所使用的,除非上下文另有明确规定,单数形式“一”、“一个”和“该”包括单数引用和复数引用。因此,例如,“一个药剂”的提及包括单个药剂和多个此类药剂。

[0108] AAV

[0109] “腺相关病毒”或“AAV”是感染人和一些其他灵长类物种的病毒。野生型AAV基因组是单链脱氧核糖核酸(ssDNA),可以是正义的,也可以是负义的。基因组包含两个反向末端重复序列(ITR),一个在DNA链的每个末端,和两个开放阅读框(ORF):ITR之间的rep和cap。rep ORF包含编码AAV生命周期所需的Rep蛋白的四个重叠基因。cap ORF包含编码衣壳蛋白:VP1、VP2和VP3的重叠基因,VP1、VP2和VP3一起相互作用以形成病毒衣壳。VP1、VP2和VP3由一个mRNA转录物翻译,其可以以两种不同的方式剪接:可以切除较长或较短的内含子,从而形成两个mRNA的同工型:约2.3kb和约2.6kb长的mRNA同工型。衣壳形成由大约60个单独的衣壳蛋白亚基组成的超分子组装体,形成能够保护AAV基因组的无包膜的T-1二十面体晶格。成熟衣壳包含比例为约1:1:10的VP1、VP2和VP3(分子量分别为约87kDa、73kDa和62kDa)。

[0110] rAAV颗粒可以包含核酸载体(例如重组基因组),其可以至少包含:(a)包含编码感兴趣的蛋白质或多肽(例如,分裂Cas9或分裂核碱基)或感兴趣的RNA(例如,gRNA)的序列的一个或多个异源核酸区域,或包含编码Rep蛋白的序列的一个或多个核酸区域;和(b)在一个或多个核酸区(例如,异源核酸区域)侧翼的包含反向末端重复(ITR)序列(例如,野生型ITR序列或工程化ITR序列)的一个或多个区。在一些实施方案中,核酸载体的大小为4kb至5kb(例如,大小为4.2至4.7kb)。在一些实施方案中,核酸载体进一步包含编码Rep蛋白的区域。在一些实施方案中,核酸载体是环状的。在一些实施方案中,核酸载体是单链的。在一些实施方案中,核酸载体是双链的。在一些实施方案中,双链核酸载体可以是例如自互补载体,其含有核酸载体的区域,该区域与核酸载体的另一区域互补,从而引发核酸载体的双链性的形成。

[0111] 脱氨酶

[0112] 术语“脱氨酶”或“脱氨酶结构域”是指催化脱氨反应的蛋白质或酶。在一些实施方

案中,脱氨酶是腺苷(或腺嘌呤)脱氨酶,其催化腺嘌呤或腺苷的水解脱氨。在一些实施方案中,腺苷脱氨酶催化脱氧核糖核酸(DNA)中的腺嘌呤或腺苷水解脱氨为肌苷。在其他实施方案中,脱氨酶是胞苷(或胞嘧啶)脱氨酶,其催化胞苷或胞嘧啶的水解脱氨。

[0113] 本文提供的脱氨酶可以来自任何生物体,例如细菌。在一些实施方案中,脱氨酶或脱氨酶结构域是来自生物体的天然存在的脱氨酶的变体。在一些实施方案中,脱氨酶或脱氨酶结构域在自然界中不存在。例如,在一些实施方案中,脱氨酶或脱氨酶结构域与天然存在的脱氨酶至少50%、至少55%、至少60%、至少65%、至少70%、至少75%、至少80%、至少85%、至少90%、至少95%、至少96%、至少97%、至少98%、至少99%或至少99.5%相同。

[0114] 如本文所用,术语“腺苷脱氨酶”或“腺苷脱氨酶结构域”是指催化腺苷(或腺嘌呤)脱氨反应的蛋白质或酶。出于本公开的目的,术语“腺苷”和“腺嘌呤”可以互换使用。例如,出于本公开的目的,提及“腺嘌呤碱基编辑器”(ABE)是指与“腺苷碱基编辑器”(ABE)相同的实体。类似地,出于本公开的目的,提及“腺嘌呤脱氨酶”是指与“腺苷脱氨酶”相同的实体。然而,本领域的普通技术人员将理解,“腺嘌呤”是指嘌呤碱基,而“腺苷”是指包括嘌呤碱基(腺嘌呤)和糖部分(例如,核糖或脱氧核糖)的较大核苷分子。在某些实施方案中,本公开提供了包含一个或多个腺苷脱氨酶结构域的碱基编辑器融合蛋白。例如,腺苷脱氨酶结构域可以包含通过接头连接的第一腺苷脱氨酶和第二脱氨酶结构域的异二聚体。本文提供的腺苷脱氨酶(例如,工程化腺苷脱氨酶或经进化的腺苷脱氨酶)是将DNA或RNA中的腺嘌呤(A)转化为肌苷(I)的酶。此类腺苷脱氨酶可导致A:T至G:C碱基对转化。在一些实施方案中,脱氨酶是来自生物体的天然存在的脱氨酶的变体。在一些实施方案中,脱氨酶在自然界中不存在。例如,在一些实施方案中,脱氨酶与天然存在的脱氨酶至少50%、至少55%、至少60%、至少65%、至少70%、至少75%、至少80%、至少85%、至少90%、至少95%、至少96%、至少97%、至少98%、至少99%或至少99.5%相同。

[0115] 在一些实施方案中,腺苷脱氨酶衍生自细菌,例如,大肠杆菌、金黄色葡萄球菌、伤寒沙门氏菌、腐败希瓦氏菌、流感嗜血杆菌或者新月柄杆菌。在一些实施方案中,腺苷脱氨酶是TadA脱氨酶。在一些实施方案中,TadA脱氨酶是大肠杆菌TadA脱氨酶(ecTadA)。在一些实施方案中,TadA脱氨酶是截短的大肠杆菌TadA脱氨酶。例如,截短的ecTadA相对于全长ecTadA可以缺失一个或多个N-端氨基酸。在一些实施方案中,截短的ecTadA相对于全长ecTadA可以缺失1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、6、17、18、19或20个N-端氨基酸残基。在一些实施方案中,截短的ecTadA相对于全长ecTadA可以缺失1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、6、17、18、19或20个C-端氨基酸残基。在一些实施方案中,ecTadA脱氨酶不包含N-端甲硫氨酸。参考2018年3月15日公开的美国专利公开号2018/0073012,其通过引用并入本文。

[0116] 如本文所用,术语“胞苷脱氨酶”或“胞苷脱氨酶结构域”是指催化胞苷或胞嘧啶脱氨反应的蛋白质或酶。术语“胞苷”和“胞嘧啶”在本公开内容中可互换使用。例如,出于本公开的目的,提及“胞嘧啶碱基编辑器”(CBE)是指与“胞嘧啶碱基编辑器”(CBE)相同的实体。类似地,出于本公开的目的,提及“胞苷脱氨酶”是指与“胞嘧啶脱氨酶”相同的实体。然而,本领域的普通技术人员将理解,“胞嘧啶”是指嘧啶碱基,而“胞苷”是指包括嘧啶碱基(胞嘧啶)和糖部分(例如,核糖或脱氧核糖)的较大核苷分子。胞苷脱氨酶由CDA基因编码,是催化从胞苷(即,与核糖环附接时的碱基胞嘧啶,即称为胞苷的核苷)至尿苷(C至U)和从胞苷至

脱氧尿苷(C至U)的胺基的去除的酶。胞苷脱氨酶的非限制性实例是APOBEC1(“载脂蛋白B mRNA编辑酶,催化多肽1”)。另一个实例是AID(“活化诱导胞苷脱氨酶”)。在标准Watson-Crick氢键配对下,胞嘧啶碱基与鸟嘌呤碱基形成氢键。当胞苷转化为尿苷(或胞苷转化为脱氧尿苷)时,尿苷(或尿苷的尿嘧啶碱基)发生与腺嘌呤碱基的氢键配对。因此,胞苷脱氨酶将“C”转化为尿苷(“U”)将导致在细胞修复和/或复制过程中插入“A”而不是“G”。由于腺嘌呤“A”与胸腺嘧啶“T”配对,胞苷脱氨酶与DNA复制协同作用,导致双链DNA分子中的C·G配对转化为T·A配对。

[0117] 反义链

[0118] 在遗传学中,双链DNA内的片段的“反义”链是模板链,并且其被认为以3’至5’方向运行。相比之下,“正义”链是双链DNA内从5’至3’运行的片段,并且其与从3’至5’运行的DNA反义链或模板链互补。在编码蛋白质的DNA片段的情况下,正义链是具有与mRNA相同序列的DNA链,其在转录期间以反义链作为其模板,并且最终经历(通常,不总是)翻译成蛋白质。因此,反义链负责随后翻译成蛋白质的RNA,而正义链具有与mRNA几乎相同的组成。注意,对于dsDNA的每个片段,可能有两组正义链和反义链,这取决于读取的方向(因为正义链和反义链与视角有关)。最终是基因产物或mRNA决定了dsDNA的一个片段的哪条链被称为正义链或反义链。

[0119] 碱基编辑

[0120] “碱基编辑”是指基因组编辑技术,其涉及在靶向基因组基因座处将特定核酸碱基转化为另一个。在某些实施方案中,这可以在不需要双链DNA断裂(DSB)或单链断裂(即,切口产生)的情况下实现。迄今为止,包括基于CRISPR的系统在内的其他基因组编辑技术都是从在感兴趣的基因座处引入DSB开始的。随后,细胞DNA修复酶修补断裂,通常导致DSB位点处的碱基的随机插入或缺失(插入缺失)。然而,当期望在靶基因座处引入或校正点突变而不是随机破坏整个基因时,这些基因组编辑技术是不合适的,因为校正率低(例如,通常0.1%至5%),同时主要基因组编辑产物是插入缺失。为了在不同时引入随机插入缺失的情况下提高基因校正的效率,本发明人先前修饰了CRISPR/Cas9系统,以将一个DNA碱基直接转化为另一个而不形成DSB。参见,Komor,A.C.,et al.,Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage.Nature 533,420-424(2016),其全部内容通过引用并入本文。

[0121] 碱基编辑器

[0122] 如本文所用,术语“碱基编辑器(BE)”是指包含能够对在核酸序列内(例如,DNA或RNA)的碱基(例如,A、T、C、G或U)产生修饰从而将一个碱基转化为另一个碱基(例如,A至G、A至C、A至T、C至T、C至G、C至A、G至A、G至C、G至T、T至A、T至C、T至G)的多肽。在一些实施方案中,碱基编辑器能够使核酸内的碱基(如DNA分子内的碱基)脱氨。在腺嘌呤碱基编辑器的情况下,碱基编辑器能够使DNA中的腺嘌呤(A)脱氨。这样的碱基编辑器可以包括与腺苷脱氨酶融合的核酸可编程DNA结合蛋白(napDNAbp)。一些碱基编辑器包括在本文所述的碱基编辑方法中使用的CRISPR介导的融合蛋白。在一些实施方案中,碱基编辑器包含与脱氨酶融合的核酸酶失活的Cas9(dCas9),该脱氨酶经由形成R环以向导RNA编程的方式结合核酸,但不切割核酸。例如,融合蛋白的dCas9结构域可以包括D10A和H840A突变(其使得Cas9能够仅切割核酸双链体的一条链),如PCT/US2016/058344中所述,其于2017年4月27日公开为WO

2017/070632,并通过引用以其整体并入本文。酿脓链球菌Cas9的DNA切割结构域包括两个亚结构域,HNH核酸酶亚结构域和RuvC1亚结构域。HNH亚结构域切割与gRNA互补的链("靶向链",或其中发生编辑或脱氨作用的链),而RuvC1亚结构域切割含有PAM序列的非互补链("非编辑链")。RuvC1突变体D10A在靶向链中生成切口,而HNH突变体H840A在非编辑链上生成切口(参见Jinek et al.,Science,337:816-821(2012);Qi et al.,Cell.28;152(5):1173-83(2013)。

[0123] 如本文所用,术语胞苷、胞嘧啶和脱氧胞苷都是同义词,是指能够使用CBE进行编辑的胞苷。同样,术语腺苷、腺嘌呤和脱氧腺嘌呤都指能够使用ABE进行编辑的腺嘌呤。此外,术语胞苷碱基编辑器、胞嘧啶碱基编辑器等是同义词。类似地,术语腺苷碱基编辑器、腺嘌呤碱基编辑器等是同义词。

[0124] 在一些实施方案中,核碱基编辑器是大分子或大分子复合物,其主要(例如,超过80%、超过85%、超过90%、超过95%、超过99%、超过99.9%或100%)导致使用1)修饰核苷酸、核苷或核碱基的酶和2)可被编程以与特定核酸序列结合的核酸结合蛋白的组合,将多核酸序列中核碱基转化成另一个核碱基(即,转换或颠换)。

[0125] 在一些实施方案中,核碱基编辑器包含DNA结合结构域(例如,可编程DNA结合结构域,如dCas9或nCas9),其将碱基编辑器引导至靶序列。在一些实施方案中,核碱基编辑器包含与可编程DNA结合结构域(例如,dCas9或nCas9)融合的核碱基修饰酶。"核碱基修饰酶"是可以修饰核碱基并将一个核碱基转化为另一个的酶(例如,脱氨酶如胞苷脱氨酶或腺苷脱氨酶)。在一些实施方案中,核碱基编辑器可以靶向核酸序列中的胞嘧啶(C)碱基,并将C转化为胸腺嘧啶(T)碱基。在一些实施方案中,C至T的编辑由脱氨酶(例如胞苷脱氨酶)进行。也考虑了可以执行其他类型的碱基转化(例如,腺苷(A)至鸟嘌呤(G)、C至G)的碱基编辑器。

[0126] 在一些实施方案中,将C转化为T的核碱基编辑器包含胞苷脱氨酶。"胞苷脱氨酶"是指催化化学反应"胞嘧啶+H_2O→尿嘧啶+NH_3"或"5-甲基-胞嘧啶+H_2O→胸腺嘧啶+NH_3"的酶。从反应式可以明显看出,此类化学反应导致C至U/T的核碱基变化。在基因的情况下,这样的核苷酸变化或突变可以反过来导致蛋白质中的氨基酸变化,这可以影响蛋白质的功能,例如,功能丧失或功能获得。在一些实施方案中,C至T核碱基编辑器包含与胞苷脱氨酶融合的dCas9或nCas9。在一些实施方案中,胞苷脱氨酶结构域融合至dCas9或nCas9的N端。在一些实施方案中,碱基编辑器进一步包含抑制尿嘧啶糖基化酶和/或核定位信号的结构域。此类核碱基编辑器已在本领域中有所描述,例如,Rees&Liu,Nat Rev Genet.2018;19(12):770-788,Rees,et al.Sci.Advances 5,eaax5717(2019),和Koblan et al.,Nat Biotechnol.2018;36(9):843-846;以及2018年3月15日公开的美国专利公开号2018/0073012,其于2018年10月30日发布为美国专利号10,113,163;2017年5月4日公开的美国专利公开号2017/0121693,其于2019年1月1日发布为美国专利号10,167,457;2017年4月27日公开的国际公开号WO 2017/070633;2015年6月18日公开的美国专利公开号2015/0166980;2017年12月12日发布的美国专利号9,840,699;2018年9月18日发布的美国专利号10,077,453;2019年1月31日公开的PCT公开号WO 2019/023680;2018年9月27日公开的国际公开号WO 2018/0176009;2019年5月23日提交的国际申请号PCT/US2019/033848;2019年8月23日提交的国际申请号PCT/US2019/47996;2020年4月17日提交的美国临时申请号62/835,490;2019年11月15日提交的国际申请号PCT/US2019/61685;2019年10月24日提交的国际申请号

PCT/US2019/57956;2019年6月7日提交的美国临时申请号62/858,958;2019年10月29日提交的国际公开号PCT/US2019/58678,其各自的内容通过引用以其整体并入本文。

[0127] 在一些实施方案中,核碱基编辑器将A转化为G。在一些实施方案中,核碱基编辑器包含腺苷脱氨酶。"腺苷脱氨酶"是参与嘌呤代谢的酶。它是来自食物的腺苷的分解和组织中核酸周转所必需的。它在人体中的主要功能是免疫系统的发展和维持。腺苷脱氨酶在DNA的背景下催化腺苷的水解脱氨(形成肌苷,其碱基配对为G)。没有已知的腺苷脱氨酶作用于DNA。相比之下,已知的腺苷脱氨酶仅作用于RNA(tRNA或mRNA)。接受DNA底物并将dA脱氨为脱氧肌苷的经进化的腺苷脱氨酶已在下文进行了描述:例如,2017年8月3日提交的PCT申请PCT/US2017/045381(其公开为WO2018/027078)和2019年5月23日提交的PCT申请号PCT/US2019/033848(其公开为WO 2019/226953),其各自的内容通过引用通过引用并入本文。

[0128] 示例性腺嘌呤碱基编辑器(ABE)(或"腺苷碱基编辑器")和胞嘧啶碱基编辑器(CBE)(或"胞嘧啶碱基编辑器")也描述于下文:Rees&Liu,Base editing:precision chemistry on the genome and transcriptome of living cells,Nat.Rev.Genet.2018;19(12):770-788;以及2018年3月15日公开的美国专利公开号2018/0073012,其于2018年10月30日发布为美国专利号10,113,163;2017年5月4日公开的美国专利公开号2017/0121693,其于2019年1月1日发布为美国专利号10,167,457;2017年4月27日公开的国际公开号WO 2017/070633;2015年6月18日公开的美国专利公开号2015/0166980;2017年12月12日发布的美国专利号9,840,699;以及2018年9月18日发布的美国专利号10,077,453,其各自的内容通过引用以其整体并入本文。

[0129] Cas9

[0130] 术语"Cas9"或"Cas9核酸酶"是指包含Cas9结构域或其片段(例如,包含Cas9的活性或失活的DNA切割结构域和/或Cas9的gRNA结合结构域的蛋白质)的RNA引导的核酸酶。如本文所用,"Cas9结构域"是包含Cas9的活性或失活的切割结构域和/或Cas9的gRNA结合结构域的蛋白质片段。"Cas9蛋白"是全长Cas9蛋白。Cas9核酸酶有时也称为casn1核酸酶或CRISPR(Clustered Regularly Interspaced Short Palindromic Repeat,成簇的规律间隔的短回文重复序列)相关核酸酶。CRISPR是适应性免疫系统,其可针对可移动遗传元件(病毒、转座元件和接合质粒)提供保护。CRISPR簇含有间隔区、与先前移动元件互补的序列和靶标入侵核酸。CRISPR簇被转录并加工成CRISPR RNA(crRNA)。在II型CRISPR系统中,前crRNA的正确加工需要反式编码的小RNA(tracrRNA)、内源性核糖核酸酶3(rnc)和Cas9结构域。tracrRNA用作核糖核酸酶3辅助加工前crRNA的向导。随后,Cas9/crRNA/tracrRNA内切核酸地切割与间隔区互补的线性或环状dsDNA靶标。首先内切核酸地切割不与crRNA互补的靶链,然后外切核酸地修剪3'-5'。在自然界中,DNA结合和切割通常需要蛋白质和两种RNA。然而,单一向导RNA("sgRNA",或简称为"gNRA")可以被工程化以将crRNA和tracrRNA两者的方面并入单个RNA种类中。参见,例如,Jinek M.,Chylinski K.,Fonfara I.,Hauer M.,Doudna J.A.,Charpentier E.Science 337:816-821(2012),其全部内容通过引用在此并入。Cas9识别CRISPR重复序列中的短基序(PAM或原间隔区相邻基序)以帮助区分自身与非自身。Cas9核酸酶序列和结构是本领域技术人员熟知的(参见,例如,"Complete genome sequence of an M1 strain of Streptococcus pyogenes."Ferretti et al.,J.J.,McShan W.M.,Ajdic D.J.,Savic D.J.,Savic G.,Lyon K.,Primeaux C.,Sezate S.,

Suvorov A.N.,Kenton S.,Lai H.S.,Lin S.P.,Qian Y.,Jia H.G.,Najar F.Z.,Ren Q.,Zhu H.,Song L.,White J.,Yuan X.,Clifton S.W.,Roe B.A.,McLaughlin R.E.,Proc.Natl.Acad.Sci.U.S.A.98:4658-4663(2001);"CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III."Deltcheva E.,Chylinski K.,Sharma C.M.,Gonzales K.,Chao Y.,Pirzada Z.A.,Eckert M.R.,Vogel J.,Charpentier E.,Nature 471:602-607(2011);和"A programmabledual-RNA-guided DNA endonuclease in adaptive bacterial immunity."Jinek M.,Chylinski K.,Fonfara I.,Hauer M.,Doudna J.A.,Charpentier E.Science 337:816-821(2012),其各自的全部内容通过引用并入本文)。Cas9直系同源物已在各种物种中被描述,包括但不限于,酿脓链球菌和嗜热链球菌。基于本公开内容,其他合适的Cas9核酸酶和序列对于本领域技术人员将是显而易见的,并且此类Cas9核酸酶和序列包括来自Chylinski,Rhun,and Charpentier,"The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems"(2013)RNA Biology 10:5,726-737中公开的生物体和基因座的Cas9序列;其全部内容通过引用并入本文。在一些实施方案中,Cas9核酸酶包含部分损害或失活DNA切割结构域的一个或多个突变。

[0131] 核酸酶失活的Cas9结构域可互换地称为"dCas9"蛋白(对于核酸酶-"死的"Cas9)。用于生成具有失活的DNA切割结构域的Cas9结构域(或其片段)的方法是已知的(参见,例如,Jinek et al.,Science.337:816-821(2012);Qi et al.,"Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression"(2013)Cell.28;152(5):1173-83,其各自的全部内容通过引用并入本文)。例如,已知Cas9的DNA切割结构域包括两个亚结构域,HNH核酸酶亚结构域和RuvC1亚结构域。HNH亚结构域切割与gRNA互补的链,而RuvC1亚结构域切割非互补链。这些亚结构域内的突变可以沉默Cas9的核酸酶活性。例如,突变D10A和H840A使酿脓链球菌Cas9的核酸酶活性完全失活(Jinek et al.,Science.337:816-821(2012);Qi et al.,Cell.28;152(5):1173-83(2013))。在一些实施方案中,提供了包含Cas9片段的蛋白质。例如,在一些实施方案中,蛋白质包含两个Cas9结构域中的一个:(1)Cas9的gRNA结合结构域;或(2)Cas9的DNA切割结构域。在一些实施方案中,包含Cas9或其片段的蛋白质被称为"Cas9变体"。Cas9变体与Cas9或其片段共享同源性。例如,Cas9变体与野生型Cas9(例如,SEQ ID NO:200的SpCas9)至少约70%相同、至少约80%相同、至少约90%相同、至少约95%相同、至少约96%相同、至少约97%相同、至少约98%相同、至少约99%相同、至少约99.5%相同、至少约99.8%相同或至少约99.9%相同。在一些实施方案中,Cas9变体与野生型Cas9(例如,SEQ ID NO:200的SpCas9)相比可具有1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、21、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50个或更多个氨基酸变化。在一些实施方案中,Cas9变体包含Cas9的片段(例如,gRNA结合结构域或DNA切割结构域),使得该片段与野生型Cas9(例如,SEQ ID NO:200的SpCas9)的相应片段至少约70%相同、至少约80%相同、至少约90%相同、至少约95%相同、至少约96%相同、至少约97%相同、至少约98%相同、至少约99%相同、至少约99.5%相同或至少约99.9%相同。在一些实施方案中,该片段是相应野生型Cas9(例如,SEQ ID NO:200的SpCas9)的氨基酸长度的至少30%、至少35%、至少40%、至少45%、至少50%、至少55%、至少60%、至少65%、至少70%、至少75%、至少80%、至少85%、至少90%、至少95%、至少96%、至少97%、至少

98％、至少99％或至少99.5％。

[0132] 如本文所用,术语“nCas9”或“Cas9切口酶”是指Cas9或其变体,其仅切割靶切割位点的一条链或使靶切割位点的一条链产生切口,从而在双链DNA分子中引入切口而不是产生双链断裂。这可以通过在野生型Cas9中引入使Cas9的两种内切核酸酶活性之一失活的适当的突变来实现。考虑了使一种Cas9内切核酸酶活性失活但使另一种保持完整的任何合适的突变,例如野生型酿脓链球菌Cas9氨基酸序列中的D10A或H840A突变之一,或野生型金黄色葡萄球菌Cas9氨基酸序列中的D10A突变,可用于形成nCas9。

[0133] cDNA

[0134] 术语“cDNA”是指从RNA模板拷贝的DNA链。cDNA与RNA模板互补。

[0135] 环状置换体

[0136] 如本文所用,术语“环状置换体”是指包含环状置换的蛋白质或多肽(例如,Cas9),环状置换是蛋白质结构构型的变化,涉及蛋白质氨基酸序列中出现的氨基酸顺序的变化。换句话说,环状置换体是与野生型对应物相比具有改变的N端和C端的蛋白质,例如,蛋白质的野生型C端半部变成新的N端半部。环状置换(或CP)本质上是蛋白质一级序列的拓扑重排,通常使用肽接头连接其N端和C端,同时在不同位置分裂其序列以产生新的相邻N端和C端。结果是蛋白质结构具有不同的连接性,但通常具有相同的整体相似的三维(3D)形状,并且可能包括改进或改变的特性,包括降低的蛋白水解敏感性、改进的催化活性、改变的底物或配体结合和/或改进的热稳定性。环状置换蛋白可以存在于自然界中(例如,伴刀豆球蛋白A和凝集素)。此外,环状置换可作为翻译后修饰的结果而发生,或者可使用重组技术进行工程改造(例如,参见Oakes et al.,“Protein Engineering of Cas9 for enhanced function,”Methods Enzymol,2014,546:491-511以及Oakes et al.,“CRISPR-Cas9 Circular Permutants as Programmable Scaffolds for Genome Modification,”Cell,January10,2019,176:254-267,每篇均通过引用并入本文)。

[0137] CRISPR

[0138] CRISPR是细菌和古细菌中的DNA序列家族(即,CRISPR簇),其代表由已入侵原核生物的病毒的先前感染的片段。DNA片段被原核细胞用于检测和破坏来自相似病毒的后续攻击的DNA,并且与CRISPR相关蛋白(包括Cas9及其同源物)和CRISPR相关RNA的阵列一起有效地组成原核免疫防御系统。在自然界中,CRISPR簇被转录并加工成CRISPR RNA(crRNA)。在某些类型的CRISPR系统中(例如,II型CRISPR系统),前crRNA的正确加工需要反式编码的小RNA(tracrRNA)、内源性核糖核酸酶3(rnc)和Cas9结构域。tracrRNA用作核糖核酸酶3辅助加工前crRNA的向导。随后,Cas9/crRNA/tracrRNA内切核酸地切割与RNA互补的线性或环状dsDNA靶标。具体地,首先内切核酸地切割不与crRNA互补的靶链,然后外切核酸地修剪3’-5’。在自然界中,DNA结合和切割通常需要蛋白质和两种RNA。然而,单一向导RNA(“sgRNA”,或简称为“gRNA”)可以被工程化以将crRNA和tracrRNA两者的方面并入单个RNA种类-向导RNA中。参见,例如,Jinek M.,Chylinski K.,Fonfara I.,Hauer M.,Doudna J.A.,Charpentier E.Science 337:816-821(2012),其全部内容通过引用并入本文。Cas9识别CRISPR重复序列中的短基序(PAM或原间隔区相邻基序)以帮助区分自身与非自身。CRISPR生物学以及Cas9核酸酶序列和结构是本领域技术人员熟知的(参见,例如,“Complete genome sequence of an M1 strain of Streptococcus pyogenes.”Ferrettiet al.,

J.J.,McShan W.M.,Ajdic D.J.,Savic D.J.,Savic G.,Lyon K.,Primeaux C.,Sezate S.,Suvorov A.N.,Kenton S.,Lai H.S.,Lin S.P.,Qian Y.,Jia H.G.,Najar F.Z.,Ren Q.,Zhu H.,Song L.,White J.,Yuan X.,Clifton S.W.,Roe B.A.,McLaughlin R.E.,Proc.Natl.Acad.Sci.U.S.A.98:4658-4663(2001);"CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III."Deltcheva E.,Chylinski K.,Sharma C.M.,Gonzales K.,Chao Y.,Pirzada Z.A.,Eckert M.R.,Vogel J.,Charpentier E.,Nature 471:602-607(2011);和"A programmabledual-RNA-guided DNA endonuclease in adaptive bacterial immunity."Jinek M.,Chylinski K.,Fonfara I.,Hauer M.,Doudna J.A.,Charpentier E.Science 337:816-821(2012),其各自的全部内容通过引用并入本文)。Cas9直系同源物已在各种物种中被描述,包括但不限于,酿脓链球菌和嗜热链球菌。基于本公开内容,其他合适的Cas9核酸酶和序列对于本领域技术人员将是显而易见的,并且此类Cas9核酸酶和序列包括来自Chylinski,Rhun,and Charpentier,"The tracrRNA and Cas9 families of type IICRISPR-Cas immunity systems"(2013)RNA Biology 10:5,726-737中公开的生物体和基因座的Cas9序列;其全部内容通过引用并入本文。

[0139] 在某些类型的CRISPR系统(例如,II型CRISPR系统)中,正确处理前crRNA需要反式编码小RNA(tracrRNA)、内源性核糖核酸酶3(rnc)和Cas9蛋白。tracrRNA充当核糖核酸酶3辅助处理前crRNA的向导。随后,Cas9/crRNA/tracrRNA内切核酸地切割与RNA互补的线性或环状核酸靶标。具体而言,首先内切核酸地切割不与crRNA互补的靶标链,然后3'-5'外切核酸地修剪。在自然界中,DNA结合和切割通常需要蛋白质和两种RNA。然而,可以对单一向导RNA("sgRNA"或简称为"gRNA")进行工程化改造,以将crRNA和tracrRNA两者的实施方案并入单一RNA种类——向导RNA。

[0140] 一般而言,"CRISPR系统"是指参与CRISPR相关("Cas")基因表达或指导Cas基因活性的转录物和其他元件的统称,包括编码Cas基因的序列、tracr(反式激活CRISPR)序列(例如tracrRNA或活性部分tracrRNA)、tracr配对序列(在内源性CRISPR系统的背景下涵盖"直接重复"和tracrRNA处理的部分直接重复)、向导序列(在内源性CRISPR系统的背景下也称为"间隔区")或来自CRISPR基因座的其他序列和转录物。该系统的tracrRNA与向导RNA上存在的tracr配对序列互补(完全地或部分地)。

[0141] 降解决定子

[0142] 术语"降解决定子"或"降解决定子结构域"是指影响、控制、指导或以其他方式调节多肽降解速率的多肽的一部分。降解决定子可以高度可变,并且可以包括短氨基酸序列、结构基序和/或暴露的氨基酸。此外,降解决定子可以位于多肽内的任何位置(例如,在N端、C端或在一级结构内的内部位置)。由降解决定子调节的多肽降解的特定机制不受限制,并且可以包括泛素依赖性降解(即,涉及基于蛋白酶体的降解的降解)或泛素非依赖性降解。例如,由SMN2基因的外显子8编码的NH3-EMLA-COOH(SEQ ID NO:384)的4氨基酸序列尾部起降解决定子的作用,从而引发SMN2的降解。

[0143] 有效量

[0144] 如本文所用,术语"有效量"是指足以引起期望生物反应的生物活性剂的量。例如,在一些实施方案中,碱基编辑器的有效量可以指足以编辑靶位点核苷酸序列(例如基因组)的碱基编辑器的量。在一些实施方案中,本文提供的碱基编辑器(例如包含Cas9切口酶结构

域和核碱基修饰结构域(例如脱氨酶结构域)的碱基编辑器)的有效量可以指足以诱导由碱基编辑器特异性结合和编辑的靶位点的编辑的碱基编辑器的量。在一些实施方案中,本文提供的碱基编辑器的有效量可以指足以诱导具有以下特征的编辑的碱基编辑器的量:>50%产物纯度、在靶序列周围区域上的<5%插入缺失和/或2-8个核苷酸的编辑窗口。在其他实施方案中,碱基编辑器的有效量可以指足以诱导以下编辑的碱基编辑器的量:>45%产物纯度、<10%插入缺失、预期点突变与插入缺失的比率至少为5:1和/或编辑窗口为2-10个核苷酸。如本领域技术人员将理解的,药剂(例如碱基编辑器、核酸酶、脱氨酶、杂交蛋白、蛋白质和多核苷酸的复合物或多核苷酸(例如gRNA))的有效量可以根据各种因素而变化,如例如根据期望的生物反应,例如根据待编辑的特定等位基因、基因组或靶位点,根据靶细胞或组织(即待编辑的细胞或组织)以及根据所使用的药剂而变化。

[0145] 脱靶编辑和中靶编辑

[0146] 如本文所用,术语“脱靶编辑”是指在典型碱基编辑器结合窗口(即从一个原间隔区位置到另一个原间隔区位置,通常为2至8个核苷酸长)之外的序列中对核苷酸(例如胞嘧啶)引入非预期的修饰(例如脱氨)。脱靶编辑可能是由于gRNA序列与靶序列的弱结合或非特异性结合所致。脱靶编辑也可能是由于碱基编辑器的核苷酸修饰结构域(例如脱氨酶结构域)与靶序列无关的基因座中的核碱基的内在缔合所致。

[0147] 术语“Cas9依赖性脱靶编辑”是指非预期修饰的引入是由Cas9-gRNA复合物(例如gRNA与碱基编辑器的Cas9结构域之间的复合物)与核酸位点的弱结合或非特异性结合引起的,该核酸位点与靶序列具有相当高(例如超过60%,或相对于靶序列具有少于6个错配)的序列同一性。相比之下,术语“Cas9非依赖性脱靶编辑”是指非预期修饰的引入是由碱基编辑器(例如核苷酸修饰结构域)与核酸位点的弱缔合引起的,该核酸位点与靶序列不具有高序列同一性(约60%或更少,或相对于靶序列具有6-8个或更多错配)。由于这些缔合不依赖于Cas9-gRNA复合物与相关核酸位点之间的任何杂交而发生,因此它们被称为“Cas9非依赖性的”。

[0148] 如本文所用,术语“中靶编辑”是指对靶序列中的核苷酸(例如胞嘧啶)引入预期的修饰(例如脱氨),例如使用本文描述的碱基编辑器。

[0149] 如本文所用,术语“中靶编辑频率”和“中靶编辑效率”是指被编辑的预期碱基对的数量或比例。例如,如果碱基编辑器编辑了其预期靶向的10%的碱基对(例如,在细胞内或细胞群内),则碱基编辑器的效率可以被描述为10%。编辑效率的某些方面包含修饰(例如,脱氨)DNA内特定的核苷酸,而不生成大量的或高百分比的插入或缺失(即,插入缺失)。普遍接受的是,当紧邻在靶序列周围的区域生成小于5%的插入缺失(如在总靶核苷酸底物上测量的)时,编辑构成高编辑效率。生成超过20%的插入缺失通常被认为是编辑效率差或低。

[0150] 如本文所用,术语“脱靶编辑效率”是指被编辑的非预期碱基对的数量或比例。中靶和脱靶编辑频率可以通过本文所述的方法和测定来测量,进一步考虑到本领域已知的技术,包括高通量测序读段。如本文所用,高通量测序涉及与刚好位于感兴趣的靶序列或脱靶序列上游或下游的核酸(例如,DNA)区域互补的核酸引物(例如,DNA引物)的杂交。因为DNA靶序列和Cas9非依赖性脱靶序列在本文公开的方法中是先天已知的,所以可以使用本领域已知的技术(例如PhusionU PCR试剂盒(Life Technologies)、Phusion HS II试剂盒(Life Technologies)和Illumina MiSeq试剂盒)设计与感兴趣的靶序列和Cas9非依赖性脱靶序

列上游或下游的区域具有足够互补性的核酸引物。由于许多Cas9依赖性脱靶位点与感兴趣的靶位点具有高序列同一性,因此同样可以使用本领域已知的技术和试剂盒设计与Cas9依赖性脱靶位点上游或下游的区域具有足够互补性的核酸引物。这些试剂盒利用聚合酶链式反应(PCR)扩增,其产生扩增子作为中间产物。靶序列和脱靶序列可以包含进一步含有原间隔区和PAM的基因组基因座。因此,如本文所用,术语"扩增子"可以指构成基因组基因座、原间隔区和PAM的聚集体的核酸分子。本文使用的高通量测序技术可以进一步包括Sanger测序和和/或全基因组测序(WGS)。所公开的碱基编辑器的脱靶效应可以使用2020年11月25日提交的国际申请号PCT/US2020/624628中公开的测定和方法来测量,其通过引用并入本文。

[0151] 功能等效物

[0152] 术语"功能等效物"是指与第一生物分子在功能上等效但在结构上不一定等效的第二生物分子。例如,"Cas9等效物"是指具有与Cas9相同或基本上相同的功能但不一定具有相同的氨基酸序列的蛋白质。在本公开的上下文中,本说明书通篇提及"蛋白X或其功能等效物"。在此上下文中,蛋白X的"功能等效物"包括具有等效功能的蛋白X的任何同源物、旁系同源物、片段、天然存在的、工程化的、环状置换的、突变的或合成的形式。

[0153] 融合蛋白

[0154] 如本文所用,术语"融合蛋白"是指包含来自至少两种不同蛋白质的蛋白质结构域的杂合多肽。一种蛋白可以位于融合蛋白的氨基端(N端)部分或羧基端(C端)蛋白,从而分别形成"氨基端融合蛋白"或"羧基端融合蛋白"。蛋白质可以包含不同的结构域,例如,核酸结合结构域(例如,指导蛋白与靶位点结合的Cas9的gRNA结合结构域)和核酸编辑蛋白的核酸切割结构域或催化结构域。另一个示例包括与腺苷脱氨酶融合的Cas9或其等效物。本文提供的任何蛋白质可以通过本领域已知的任何方法产生。例如,本文提供的蛋白可以经由重组蛋白表达和纯化产生,这尤其适用于包含肽接头的融合蛋白。重组蛋白表达和纯化的方法是众所周知的,并且包括Green and Sambrook,Molecular Cloning:A Laboratory Manual(4th ed.,Cold Spring Harbor Laboratory Press,Cold Spring Harbor,N.Y.(2012))中描述的那些,其全部内容通过引用并入本文。

[0155] 向导核酸

[0156] 术语"向导核酸"或"napDNAbp-编程核酸分子"或等效地"向导序列"是指一种或多种核酸分子,其与napDNAbp蛋白缔合并指导或以其他方式编程napDNAbp蛋白以定位至特定靶核苷酸序列(例如,基因组的基因座),该序列与缔合至该蛋白质的一种或多种核酸分子(或其部分或区域)互补,从而使napDNAbp蛋白与特定靶位点处的核苷酸序列结合。非限制性示例是CRISPR-Cas基因组编辑系统的Cas蛋白的向导RNA。

[0157] 向导RNA是特殊类型的向导核酸,其最常见地与CRISPR-Cas9的Cas蛋白缔合并且与Cas9缔合,引导Cas9蛋白至DNA分子中的特定序列,该特定序列与向导RNA的原间隔区序列互补。如本文所用,"向导RNA"是指内源性细菌crRNA和tracrRNA的合成融合物,其为Cas9核酸酶提供至靶DNA的靶向特异性以及支架和/或结合能力。这种合成融合物在自然界中不存在,并且通常也称为sgRNA。然而,该术语还包括与Cas9等效物、同源物、直系同源物或旁系同源物缔合的等效向导核酸分子,无论是天然存在的还是非天然存在的(例如,工程化的或重组的),并且其以其它方式对Cas9等效物进行编程以定位至特定靶核苷酸序列。Cas9等效物可以包括来自任何类型的CRISPR系统(例如,II型、V型、VI型)的其他napDNAbp,包括

Cpf1(V型CRISPR-Cas系统)、C2c1(V型CRISPR-Cas系统)、C2c2(VI型CRISPR-Cas系统)和C2c3(V型CRISPR-Cas系统)。进一步的Cas等效物在Makarova et al.,“C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector,”Science 2016;353(6299)中描述,其内容通过引用并入本文。本文提供了向导RNA的示例性序列和结构。此外,本文提供了用于设计合适的向导RNA序列的方法。

[0158] 向导RNA(“gRNA”)

[0159] 如本文所用,“向导RNA”是特定类型的向导核酸,其通常与CRISPR-Cas9的Cas蛋白相关,并且与Cas9结合,将Cas9蛋白引导至DNA分子中的特定序列,该特定序列包含与向导RNA的原间隔区序列的互补性。然而,该术语还涵盖与Cas9等效物、同源物、直系同源物或旁系同源物(无论是天然存在的还是非天然存在的(例如,工程化的或重组的))缔合的等效向导核酸分子,并且其以其他方式编程Cas9等效物以定位至特定靶核苷酸序列。Cas9等效物可以包括来自任何类型的CRISPR系统(例如,II、V、VI型)的其他napDNAbp,包括Cpf1(V型CRISPR-Cas系统)、C2c1(V型CRISPR-Cas系统)、C2c2(VI型CRISPR-Cas系统)和C2c3(V型CRISPR-Cas系统)。进一步的Cas等效物描述于Makarova et al.,“C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector,”Science 2016;353(6299),其内容通过引用并入本文。本文提供了向导RNA的示例性序列和结构。

[0160] 向导RNA可以包含各种结构元件,包括但不限于(a)间隔区序列—向导RNA中的序列(具有约20nt的长度),其结合靶DNA的互补链(并且具有与DNA的原间隔区相同的序列)和(b)gRNA核心(或gRNA支架或主链序列)—是指负责Cas9结合的gRNA内的序列,其不包括用于向导Cas9至靶DNA的约20bp间隔区序列。

[0161] 如本文所用,“向导RNA靶序列”是指与PAM链中的原间隔区序列互补的约20个核苷酸。靶序列是与向导RNA的间隔区序列退火或被其靶向的序列。向导RNA的间隔区序列和原间隔区具有相同的序列(除了间隔区序列是RNA,而原间隔区是DNA)。

[0162] 如本文所用,术语“向导RNA支架序列”是指gRNA内负责Cas9结合的序列,其不包括用于将Cas9引导至靶DNA的20bp间隔区/靶向序列。

[0163] 宿主细胞

[0164] 如本文所用,术语“宿主细胞”是指可以承载、复制和转移可用于本文提供的持续进化过程的噬菌体载体的细胞。在其中载体是病毒载体的实施方案中,合适的宿主细胞是可以被病毒载体感染、可以复制病毒载体并可以将病毒载体包装成可以感染新鲜宿主细胞的病毒颗粒的细胞。如果细胞支持病毒载体基因的表达、病毒基因组的复制和/或病毒颗粒的生成,则该细胞可以承载病毒载体。确定细胞是否是给定病毒载体的合适宿主细胞的一个标准是确定细胞是否可以支持病毒载体所衍生自的野生型病毒基因组的病毒生命周期。例如,如果病毒载体是经修饰的M13噬菌体基因组,如本文所述的一些实施方案中所提供的,则合适的宿主细胞将是可以支持野生型M13噬菌体生命周期的任何细胞。本领域技术人员熟知可用于持续进化过程的病毒载体的合适宿主细胞,并且本公开在这方面不受限制。在一些实施方案中,病毒载体是噬菌体并且宿主细胞是细菌细胞。在一些实施方案中,宿主细胞是大肠杆菌细胞。合适的大肠杆菌宿主株对于本领域技术人员来说是显而易见的,并且包括但不限于New England Biolabs(NEB)Turbo、Top10F’、DH12S、ER2738、ER2267和XL1-Blue MRF’。这些株名称是本领域公认的,并且这些株的基因型已被很好地表征。应理解,上

述株仅是示例性的,并且本发明在这方面不受限制。术语“新鲜”在本文中与宿主细胞上下文中的术语“非感染的”或“未感染的”互换使用,是指未被包含本文提供的连续进化过程中使用的感兴趣基因的病毒载体感染的宿主细胞。然而,新鲜的宿主细胞可以已被与待进化的载体无关的病毒载体感染,或者已被相同或相似类型但不携带感兴趣基因的载体感染。

[0165] 在一些实施方案中,宿主细胞是原核细胞,例如细菌细胞。在一些实施方案中,宿主细胞是大肠杆菌细胞。在一些实施方案中,宿主细胞是真核细胞,例如,酵母细胞、昆虫细胞或哺乳动物细胞。当然,宿主细胞的类型将取决于所采用的病毒载体,并且合适的宿主细胞/病毒载体组合对于本领域技术人员将是显而易见的。

[0166] 内含肽和分裂内含肽

[0167] 如本文所用,术语“内含肽”是指存在于所有生命领域的生物体中的自加工多肽结构域。内含肽(间插蛋白)进行独特的自加工事件,称为蛋白质剪接,其中它通过切割两个肽键将自身从较大的前体多肽中切除,并在此过程中通过形成新的肽键连接侧翼的外显肽(外部蛋白)序列。这种重排发生在翻译后(或可能与翻译同时发生),因为发现内含肽基因嵌入在其他蛋白质编码基因的框架内。此外,内含肽介导的蛋白质剪接是自发的;它不需要外部因素或能量源,只需要折叠内含肽结构域。这个过程也称为顺式蛋白质剪接,与使用“分裂内含肽”的反式蛋白质剪接的自然过程相反。

[0168] 分裂内含肽是内含肽的子类别。与更常见的连续内含肽不同,分裂内含肽被转录和翻译为两个单独的多肽,即N-内含肽和C-内含肽,每个多肽都与一个外显肽融合。翻译后,内含肽片段自发地和非共价地组装成典型的内含肽结构,以进行反式蛋白质剪接。

[0169] 内含肽和分裂内含肽是自剪接RNA内含子的蛋白质等价物(参见Perler et al., Nucleic Acids Res.22:1125-1127(1994)),其催化自身从前体蛋白中切除,并伴随融合称为外显肽的侧翼蛋白序列(综述于Perleret al.,Curr.Opin.Chem.Biol.1:292-299(1997);Perler,F.B.Cell 92(1):1-4(1998);Xu et al.,EMBO J.15(19):5146-5153(1996))。

[0170] 如本文所用,术语“蛋白质剪接”是指切除前体蛋白质(内含肽)的内部区域,并将蛋白质的侧翼区域(外显肽)连接以形成成熟蛋白质的过程。这种自然过程已在原核生物和真核生物的众多蛋白质中观察到(Perler,F.B.,Xu,M.Q.,Paulus,H.Current Opinion in Chemical Biology 1997,1,292-299;Perler,F.B.Nucleic Acids Research 1999,27,346-347)。内含肽单元含有催化蛋白质剪接所需的必要组分,并且通常含有参与内含肽移动的内切核酸酶结构域(Perler,F.B.,Davis,E.O.,Dean,G.E.,Gimble,F.S.,Jack,W.E.,Neff,N.,Noren,C.J.,Thomer,J.,Belfort,M.Nucleic Acids Research 1994,22,1127-1127)。然而,产生的蛋白质是连接的,而不是作为单独的蛋白质表达的。蛋白质剪接也可以反式进行,其中在单独的多肽上表达的分裂内含肽自发结合形成单个内含肽,然后经历蛋白质剪接过程以连接到单独的蛋白质。

[0171] 蛋白质剪接机制的阐明已导致大量基于内含肽的应用(Comb,et al.,美国专利号5,496,714;Comb,et al.,美国专利号5,834,247;Camarero and Muir,J.Amer.Chem.Soc.,121:5597-5598(1999);Chong,et al.,Gene,192:271-281(1997),Chong,et al.,Nucleic Acids Res.,26:5109-5115(1998);Chong,et al.,J.Biol.Chem.,273:10567-10577(1998);Cotton,et al.J.Am.Chem.Soc.,121:1100-1101(1999);Evans,et al.,

J.Biol.Chem.,274:18359-18363(1999);Evans,et al.,J.Biol.Chem.,274:3923-3926(1999);Evans,et al.,Protein Sci.,7:2256-2264(1998);Evans,et al.,J.Biol.Chem.,275:9091-9094(2000);Iwai and Pluckthun,FEBS Lett.459:166-172(1999);Mathys,et al.,Gene,231:1-13(1999);Mills,et al.,Proc.Natl.Acad.Sci.USA 95:3543-3548(1998);Muir,et al.,Proc.Natl.Acad.Sci.USA 95:6705-6710(1998);Otomo,et al.,Biochemistry 38:16040-16044(1999);Otomo,et al.,J.Biolmol.NMR 14:105-114(1999);Scott,et al.,Proc.Natl.Acad.Sci.USA 96:13638-13643(1999);Severinov and Muir,J.Biol.Chem.,273:16205-16209(1998);Shingledecker,et al.,Gene,207:187-195(1998);Southworth,et al.,EMBO J.17:918-926(1998);Southworth,et al.,Biotechniques,27:110-120(1999);Wood,et al.,Nat.Biotechnol.,17:889-892(1999);Wu,et al.,Proc.Natl.Acad.Sci.USA 95:9226-9231(1998a);Wu,et al.,Biochim Biophys Acta 1387:422-432(1998b);Xu,et al.,Proc.Natl.Acad.Sci.USA 96:388-393(1999);Yamazaki,et al.,J.Am.Chem.Soc.,120:5591-5592(1998))。每篇参考文献通过引用并入本文。

[0172] 接头

[0173] 如本文所用,术语“接头”是指连接两个分子或结构域的化学基团或分子,例如,dCas9和脱氨酶。通常,接头位于两个基团、分子或其他结构域之间或侧翼,并经由共价键与每一个基团、分子或其他结构域连接,从而将两者连接。在一些实施方案中,接头是一个氨基酸或多个氨基酸(例如,肽或蛋白质)。在一些实施方案中,接头是有机分子、基团、聚合物或化学结构域。化学基团包括但不限于二硫化物、腙和叠氮化物结构域。在一些实施方案中,接头的长度为5-100个氨基酸,例如,长度为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、30-35、35-40、40-45、45-50、50-60、60-70、70-80、80-90、90-100、100-150或150-200个氨基酸。还考虑了更长或更短的接头。在一些实施方案中,接头是XTEN接头。在一些实施方案中,接头是32个氨基酸的接头。在其他实施方案中,接头是30个、31个、33个或34个氨基酸的接头。

[0174] 突变

[0175] 如本文所用,术语“突变”是指序列(例如,核酸或氨基酸序列)内的残基被另一个残基的置换;序列内一个或多个残基的缺失或插入;或待校正的受试者的基因组序列内的残基的置换。本文通常通过鉴定原始残基、随后是残基在序列中的位置以及新置换的残基的身份来描述突变。用于制备本文提供的氨基酸置换(突变)的各种方法是本领域熟知的,并且由例如Green and Sambrook,Molecular Cloning:A Laboratory Manual(4th ed.,Cold Spring Harbor Laboratory Press,Cold Spring Harbor,N.Y.(2012))提供。突变可以包括多种类别,例如单碱基多态性、微重复区、插入缺失和倒位,并且不意味着以任何方式进行限制。突变可以包括“功能丧失性”突变,其是降低或消除蛋白质活性的突变。大多数功能丧失性突变是隐性的,因为在杂合子中,第二个染色体拷贝携带编码全功能蛋白质的基因的未突变版本,该蛋白质的存在补偿了突变的影响。存在一些例外,其中功能丧失性突变是显性的,一个示例是单倍体不足,其中生物体不能耐受杂合子经历的大约50%的蛋白质活性降低。这是人类一些遗传疾病的解释,包括马凡氏综合征,它是由称为原纤维蛋白的结缔组织蛋白的基因突变引起的。突变还包括“功能获得性”突变,其是赋予蛋白质或细胞

在正常条件下不存在的异常活性的突变。许多功能获得性突变位于调控序列中而不是编码区中,因此可能具有许多后果。例如,突变可能导致一个或多个基因在错误的组织中表达,这些组织获得它们通常缺乏的功能。替代性地,突变可能导致一个或多个参与控制细胞周期的基因过度表达,从而导致不受控制的细胞分裂,进而导致癌症。由于其性质,功能获得性突变通常是显性的。

[0176] napDNAbp

[0177] 代表“核酸可编程DNA结合蛋白”的术语“napDNAbp”是指可以与一个或多个核酸分子(即,其可以广泛地称为“napDNAbp编程核酸分子”,并且包括例如在Cas系统的情况下的向导RNA)缔合(例如,形成复合物)的任何蛋白质,该核酸分子指导或以其他方式编程蛋白质以定位于特定靶核苷酸序列(例如,基因组的基因座),该核苷酸序列与缔合于蛋白质的一个或多个核酸分子(或其部分或区域)互补,从而使蛋白质与特定靶位点处的核苷酸序列结合。术语napDNAbp包括CRISPR-Cas9蛋白,以及Cas9等效物、同源物、直系同源物或旁系同源物,无论是天然存在的还是非天然存在的(例如,工程化的或修饰的),并且可以包括来自任何类型的CRISPR系统的Cas9等效物(例如,II、V、VI型),包括Cpf1(V型CRISPR-Cas系统)、C2c1(V型CRISPR-Cas系统)、C2c2(VI型CRISPR-Cas系统)、C2c3(V型CRISPR-Cas系统)、dCas9、GeoCas9、CjCas9、Cas12a、Cas12b、Cas12c、Cas12d、Cas12g、Cas12h、Cas12i、Cas13d、Cas14、Argonaute和nCas9。进一步的Cas等效物在Makarova et al.,“C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector,”Science2016;353(6299)中描述,其内容通过引用并入本文。然而,可与本发明结合使用的核酸可编程DNA结合蛋白(napDNAbp)不限于CRISPR-Cas系统。本发明包括任何这样的可编程蛋白质,例如来自格氏嗜盐碱杆菌(NgAgo)的Argonaute蛋白,其也可用于DNA引导的基因组编辑。NgAgo引导的DNA系统不需要PAM序列或向导RNA分子,这意味着基因组编辑可以通过表达通用NgAgo蛋白并在任何基因组序列上引入合成寡核苷酸来简单进行。参见Gao et al.,DNA-guided genome editing using the Natronobacterium gregoryi Argonaute.Nature Biotechnology 2016;34(7):768-73,其通过引用并入本文。

[0178] 在一些实施方案中,napDNAbp是RNA可编程核酸酶,当与RNA处在复合物中时,可称为核酸酶:RNA复合物。通常,结合的RNA被称为向导RNA(gRNA)。gRNA可以作为两个或更多个RNA的复合物或作为单个RNA分子存在。作为单个RNA分子存在的gRNA可以被称为单一向导RNA(sgRNA),尽管“gRNA”可互换地用于指作为单个分子或作为两个或更多个分子的复合物存在的向导RNA。通常,作为单一RNA种类存在的gRNA包含两个结构域:(1)与靶核酸共享同源性的结构域(例如,并且指导Cas9(或等效物)复合物与靶标的结合);和(2)结合Cas9蛋白的结构域。在一些实施方案中,结构域(2)对应于称为tracrRNA的序列,并且包含茎环结构。例如,在一些实施方案中,结构域(2)与如Jinek et al.,Science 337:816-821(2012)的图1E中所描绘的tracrRNA同源,其全部内容通过引用并入本文。gRNAs的其他示例(例如,包括结构域2的那些)可以在题为“mRNA-Sensing Switchable gRNAs”的美国专利号9,340,799和2013年9月6日提交的,公布为WO 2015/035136的并且题为“Delivery System For Functional Nucleases”的国际专利申请号PCT/US2014/054247中找到,各自的全部内容通过引用并入本文。在一些实施方案中,gRNA包含结构域(1)和(2)中的两个或更多个,并且可以被称为“延伸的gRNA”。例如,延伸的gRNA将例如结合两个或更多个Cas9蛋白并在两个或

更多个不同区域结合靶核酸,如本文所述。gRNA包含与靶位点互补的核苷酸序列,其介导核酸酶/RNA复合物与所述靶位点的结合,提供核酸酶:RNA复合物的序列特异性。在一些实施方案中,RNA可编程核酸酶是(CRISPR相关系统)Cas9内切核酸酶,例如来自酿脓链球菌的Cas9(Csn1)(参见,例如,"Complete genome sequence of an M1 strain of Streptococcus pyogenes."Ferretti J.J.et al..,Proc.Natl.Acad.Sci.U.S.A.98:4658-4663(2001);"CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III."Deltcheva E.et al.,Nature 471:602-607(2011);和"A programmabledual-RNA-guided DNA endonuclease in adaptive bacterial immunity."Jinek M.et al.,Science 337:816-821(2012),其各自的全部内容通过引用并入本文。

[0179] napDNAbp核酸酶(例如,Cas9)使用RNA:DNA杂交来靶向DNA切割位点,这些蛋白质原则上能够靶向由向导RNA指定的任何序列。使用napDNAbp核酸酶(如Cas9)进行位点特异性切割(例如,以修饰基因组)的方法是本领域已知的(参见,例如,Cong,L.et al.Multiplex genome engineering using CRISPR/Cas systems.Science 339,819-823(2013);Mali,P.et al.RNA-guided human genome engineering via Cas9.Science 339,823-826(2013);Hwang,W.Y.et al.Efficient genome editing in zebrafish using a CRISPR-Cas system.Nature Biotechnology 31,227-229(2013);Jinek,M.et al.RNA-programmed genome editing in human cells.eLife 2,e00471(2013);Dicarlo,J.E.et al.,Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems.Nucleic Acid Res.(2013);Jiang,W.et al.RNA-guided editing of bacterial genomes using CRISPR-Cas systems.Nature Biotechnology 31,233-239(2013);其各自的全部内容通过引用并入本文)。

[0180] 切口酶

[0181] 术语"切口酶"是指napDNAbp,其仅具有仅切割靶DNA的一条链而不是两条链的单一核酸酶活性(例如,两个核酸酶结构域中的一个失活)。因此,切口酶型napDNAbp不会留下双链断裂。在一些实施方案中,任何所公开的碱基编辑器或载体可以包含含有D10A突变的S.pyogenes Cas9切口酶(SpCas9n或nCas9)。在一些实施方案中,任何所公开的碱基编辑器可以包含含有D16A突变的Nme2Cas9切口酶(Nme2Cas9n)。

[0182] 核定位信号

[0183] 核定位信号或序列(NLS)是标记、指定或以其他方式标识蛋白质以通过核转运导入细胞核的氨基酸序列。通常,该信号由暴露在蛋白质表面上的带正电荷的赖氨酸或精氨酸的一个或多个短序列组成。不同核定位蛋白可以共享相同的NLS。NLS具有与核输出信号(NES)相反的功能,其靶向核外的蛋白质。因此,单个核定位信号可以将与其相关联的实体引导至细胞核。这样的序列可以是任何大小和组成,例如多于25、25、15、12、10、8、7、6、5或4个氨基酸,但优选包含已知作为核定位信号(NLS)起作用的至少4至8个氨基酸序列。

[0184] 核酸分子

[0185] 如本文所用,术语"核酸分子"是指RNA以及单链和/或双链DNA。核酸分子可以是天然存在的,例如在基因组、转录物、mRNA、tRNA、rRNA、siRNA、snRNA、质粒、粘粒、染色体、染色单体或其他天然存在的核酸分子的背景下。另一方面,核酸分子可以是非天然存在的分子,例如重组DNA或RNA、人工染色体、工程化基因组或其片段,或合成DNA、RNA、DNA/RNA杂合体,

或包括非天然存在的核苷酸或核苷。此外,术语“核酸”、“DNA”、“RNA”和/或相似术语包括核酸类似物,例如,具有除磷酸二酯主链以外的类似物。核酸可以从天然来源纯化,使用重组表达系统产生,并且任选地纯化、化学合成等等。在适当情况下,例如在化学合成分子的情况下,核酸可以包含核苷类似物,例如具有化学修饰的碱基或糖和主链修饰的类似物。除非另有说明,核酸序列以5’至3’方向呈现。在一些实施方案中,核酸是或包含天然核苷(例如腺苷、胸苷、鸟苷、胞苷、尿苷、腺苷、脱氧胸苷、脱氧鸟苷和胞苷);核苷类似物(例如2-氨基腺苷、2-硫代胸苷、肌苷、吡咯并嘧啶、3-甲基腺苷、5-甲基胞苷、2-氨基腺苷、C5-溴尿苷、C5-氟尿苷、C5-碘尿苷、C5-丙炔基尿苷、C5-丙炔基胞苷、C5-甲基胞苷、2-氨基腺苷、7-脱氮腺苷、7-脱氮鸟苷、肌苷腺苷(inosinedenosine)、8-氧代鸟苷、O(6)-甲基鸟嘌呤和2-硫代胞苷);化学修饰的碱基;生物修饰的碱基(例如,甲基化碱基);插入碱基;修饰的糖(例如2’-氟核糖、核糖、2’-脱氧核糖、阿拉伯糖和己糖);和/或修饰的磷酸基团(例如硫代磷酸酯和5’-N-亚磷酰胺键)。

[0186] PACE

[0187] 如本文所用,术语“噬菌体辅助连续进化(PACE)”是指采用噬菌体作为病毒载体的连续进化。PACE技术的一般概念已经描述于下文:例如2009年9月8日提交的国际PCT申请PCT/US2009/056194,2010年3月11日公开为WO 2010/028347;2011年12月22日提交的国际PCT申请PCT/US2011/066747,2012年6月28日公开为WO 2012/088381;美国申请,2015年5月5日发布的美国专利号9,023,594,2015年1月20日提交的国际PCT申请PCT/US2015/012022,2015年9月11日公开为WO 2015/134121,以及2016年4月15日提交的国际PCT申请PCT/US2016/027795,2016年10月20日公开为WO 2016/168631,其各自的全部内容通过引用并入本文。

[0188] 启动子

[0189] 术语“启动子”是本领域公认的并且是指具有被细胞转录机制识别并且能够启动下游基因转录的序列的核酸分子。启动子可以是组成型活性的,意指启动子在给定的细胞环境中总是有活性的,或是条件活性的,意指启动子仅在特定条件存在下有活性。例如,条件型启动子可以仅在存在将与启动子中的调控元件相关的蛋白质连接到基本转录机制的特定蛋白质的情况下才有活性,或者仅在没有抑制分子的情况下才有活性。条件活性启动子的亚类是诱导型启动子,其需要小分子“诱导剂”的存在以获得活性。诱导型启动子的示例包括但不限于阿拉伯糖诱导型启动子、Tet-on启动子和他莫昔芬诱导型启动子。各种组成型、条件型和诱导型启动子是本领域技术人员熟知的,并且本领域技术人员将能够确定可用于实施本发明的各种此类启动子,而本发明并不限于此方面。在各种实施方案中,本公开提供了具有用于驱动编码融合蛋白(或其一个或多个单独组分)的核酸序列表达的合适启动子的载体。

[0190] 产物纯度

[0191] 如本文所用,术语“产物纯度”是指碱基编辑反应中期望产物占总产物的百分比。例如,CBE的产物纯度可以测量为其中靶标C被编辑为T的总编辑测序读段(其中靶标C已被转化为不同碱基的读段)占核酸感兴趣部分的百分比。产物纯度包括插入缺失的不存在,以及碱基转化的期望产物。

[0192] 术语“R环”是指三链体结构,其中双链DNA的两条链被一段核苷酸分开并由单链

RNA分子(例如,gRNA)分开。R环形成可以通过与napDNAbp蛋白或结构域(例如,Cas9)缔合的具有与DNA互补性的gRNA的杂交来诱导。当生成其形成的机制(例如,napDNAbp-gRNA复合物)彼此独立地发挥作用时,两个R环被称为“正交”。

[0193] 原间隔区

[0194] 如本文所用,术语“原间隔区”是指邻近PAM(原间隔区相邻基序)序列的DNA中的序列(约20bp)。该原间隔区与向导RNA的间隔区序列共享相同的序列。向导RNA与靶DNA(具体而言,其一条链,即靶DNA序列的“靶链”与“非靶链”)上的原间隔区序列的互补序列退火。为了使Cas9发挥作用,它还需要特定的原间隔区相邻基序(PAM),该基序因Cas9基因的细菌种类而异。最常用的Cas9核酸酶衍生自酿脓链球菌,识别基因组DNA中靶序列直接下游存在的NGG的PAM序列,位于非靶链上。技术人员将理解,现有技术文献有时将“原间隔区”称为向导RNA本身上的约20nt靶特异性向导序列,而不是将其称为“间隔区”。因此,在某些情况下,本文使用的术语“原间隔区”可与术语“间隔区”互换使用。说明书中关于“原间隔区”或“间隔区”出现的上下文将有助于告知读者该术语是指gRNA还是DNA靶标。

[0195] 原间隔区相邻基序(PAM)

[0196] 如本文所用,术语“原间隔区相邻序列”或“PAM”是指约2-6个碱基对的DNA序列,其是Cas9核酸酶的重要靶向组分。通常,PAM序列在任一链上,并且在Cas9切割位点的5’至3’方向的下游。典型PAM序列(即,与酿脓链球菌的Cas9核酸酶或SpCas9缔合的PAM序列))是5’-NGG-3’,其中“N”是后跟两个鸟嘌呤(“G”)核碱基的任何核碱基。不同的PAM序列可以与来自不同生物体的不同Cas9核酸酶或等效蛋白缔合。此外,可以修饰任何给定的Cas9核酸酶(例如SpCas9)以改变核酸酶的PAM特异性,使得核酸酶识别替代的PAM序列。

[0197] 例如,参考典型SpCas9氨基酸序列是SEQ ID NO:200,PAM序列可以通过引入一个或多个突变来修饰,包括(a)D1135V、R1335Q和T1337R“VQR变体”,其改变PAM对NGAN或NGNG的特异性,(b)D1135E、R1335Q和T1337R“EQR变体”,其改变PAM对NGAG的特异性,以及(c)D1135V、G1218R、R1335E和T1337R“VRER变体”,其改变PAM对NGCG的特异性。此外,典型SpCas9的D1135E变体仍然识别NGG,但与野生型SpCas9蛋白相比具有更高的选择性。

[0198] 还应理解,来自不同细菌物种的Cas9酶(即,Cas9直系同源物)可以具有不同的PAM特异性。例如,来自金黄色葡萄球菌的Cas9(SaCas9)识别NGRRT或NGRRN。此外,来自脑膜炎奈瑟菌的Cas9(NmCas)识别NNNNGATT。在另一个实例中,来自嗜热链球菌的Cas9(StCas9)识别NNAGAAW。在又一个实例中,来自齿垢密螺旋体的Cas9(TdCas)识别NAAAAC。这些是示例并且不意味着是限制性的。将进一步理解,非SpCas9结合多种PAM序列,这使得它们在期望的靶切割位点不存在合适的SpCas9PAM序列时是有用的。此外,非SpCas9可以具有使它们比SpCas9更有用的其他特征。例如,来自金黄色葡萄球菌的Cas9(SaCas9)比SpCas9小约1千碱基,因此它可以包装到腺相关病毒(AAV)中。可以进一步参考Shah et al.,“Protospacer recognition motifs:mixed identities and functional diversity,”RNA Biology,10(5):891-899(其通过引用并入本文)。

[0199] 正义链

[0200] 在遗传学上,“正义”链是双链DNA内从5’至3’运行的片段,并且其与从3’至5’运行的DNA反义链或模板链互补。在编码蛋白质的DNA片段的情况下,正义链是具有与mRNA相同序列的DNA链,其在转录期间以反义链作为其模板,并且最终经历(通常,不总是)翻译成蛋

白质。因此,反义链负责随后翻译成蛋白质的RNA,而正义链具有与mRNA几乎相同的组成。注意,对于dsDNA的每个片段,可能有两组正义链和反义链,这取决于读取的方向(因为正义链和反义链与视角有关)。最终是基因产物或mRNA决定了dsDNA的一个片段的哪条链被称为正义链或反义链。

[0201] 间隔区序列

[0202] 如本文所用,与向导RNA相关的术语"间隔区序列"是指向导RNA中约20个核苷酸的部分,其含有与靶DNA序列中的原间隔区序列互补的核苷酸序列。间隔区序列与原间隔区序列退火以在靶位点处形成ssRNA/ssDNA杂交结构,以及与原间隔区序列互补的内源DNA链的相应R环ssDNA结构。

[0203] 受试者

[0204] 如本文所用,术语"受试者"是指个体生物体,例如,个体哺乳动物。在一些实施方案中,受试者是人。在一些实施方案中,受试者是非人哺乳动物。在一些实施方案中,受试者是非人灵长类动物。在一些实施方案中,受试者是啮齿类动物。在一些实施方案中,受试者是绵羊、山羊、牛、猫或狗。在一些实施方案中,受试者是脊椎动物、两栖动物、爬行动物、鱼、昆虫、苍蝇或线虫。在一些实施方案中,受试者是研究动物。在一些实施方案中,受试者是基因工程化的,例如,基因工程化的非人受试者。受试者可以是任何性别,也可以处于任何发展阶段。在一些实施方案中,受试者是植物。

[0205] 靶位点

[0206] 术语"靶位点"是指由融合蛋白(例如本文提供的dCas9-脱氨酶融合蛋白)编辑的核酸分子内的序列。靶位点还指融合蛋白和gRNA的复合物结合的核酸分子内的序列。

[0207] 转录终止子

[0208] "转录终止子"是导致转录停止的核酸序列。转录终止子可以是单向的或双向的。它由参与通过RNA聚合酶的RNA转录物的特异性终止的DNA序列组成。转录终止子序列阻止上游启动子对下游核酸序列的转录激活。转录终止子可能是体内所必需的,以达到期望的表达水平或避免某些序列的转录。当转录终止子能够终止与其连接的序列的转录时,认为转录终止子与核苷酸序列"可操作地连接"。

[0209] 最常用的终止子类型是正向终止子。当正向转录终止子位于通常转录的核酸序列下游时,它将导致转录中止。在一些实施方案中,提供双向转录终止子,其通常导致转录在正向链和反向链上终止。在一些实施方案中,提供反向转录终止子,其通常仅在反向链上终止转录。

[0210] 在原核系统中,终止子通常分为两类(1)rho非依赖性终止子和(2)rho依赖性终止子。Rho非依赖性终止子通常由回文序列构成,该序列形成富含G-C碱基对的茎环,后跟若干T碱基。不受理论的束缚,转录终止的常规模型是茎环导致RNA聚合酶暂停,而poly-A尾的转录导致RNA:DNA双螺旋解旋并与RNA聚合酶解离。

[0211] 在真核系统中,终止子区可以包含特异性DNA序列,其允许新转录物的位点特异性切割以暴露聚腺苷酸化位点。这向专门的内源聚合酶发出信号以将约200个A残基的片段(polyA)添加到转录物的3'末端。用这种polyA尾修饰的RNA分子似乎更稳定并且翻译更有效。因此,在涉及真核生物的一些实施方案中,终止子可以包含用于切割RNA的信号。在一些实施方案中,终止子信号促进信息的聚腺苷酸化。终止子和/或聚腺苷酸化位点元件可用于

增强输出核酸水平和/或使核酸之间的通读最小化。

[0212] 根据本公开使用的终止子包括本文所述或本领域普通技术人员已知的任何转录终止子。终止子的示例包括但不限于基因的终止序列如例如牛生长激素终止子,和病毒终止序列如例如SV40终止子、spy、yejM、secG-leuU、thrLABC、rrnB T1、hisLGDCBHAFI、metZWV、rrnC、xapR、aspA和arcA终止子。在一些实施方案中,终止信号可以是不能转录或翻译的序列,例如由序列截短产生的那些。

[0213] 转换

[0214] 如本文所用,“转换”是指嘌呤核碱基的互换(A↔G)或嘧啶核碱基的互换(C↔T)。这类互换涉及形状相似的核碱基。本文公开的组合物和方法能够诱导靶DNA分子中的一个或多个转换。本文公开的组合物和方法还能够在同一靶DNA分子中诱导转换和颠换两者。这些变化涉及A↔G、G↔A、C↔T或T↔C。在具有Watson-Crick配对核碱基的双链DNA的背景下,颠换是指以下碱基对交换:A:T↔G:C、G:G↔A:T、C:G↔T:A或T:A↔C:G。本文公开的组合物和方法能够诱导靶DNA分子中的一个或多个转换。本文公开的组合物和方法还能够在同一靶DNA分子中诱导转换和颠换两者,以及其他核苷酸变化,包括缺失和插入。

[0215] 治疗

[0216] 术语“治疗”(treatment、treat、treating)是指旨在逆转疾病或病症或其一种或多种症状、减轻疾病或病症或其一种或多种症状、延迟疾病或病症或其一种或多种症状的发作或抑制疾病或病症或其一种或多种症状的进展的临床干预,如本文所述。如本文所用,术语“治疗”(treatment、treat、treating)是指旨在逆转疾病或病症或其一种或多种症状、减轻疾病或病症或其一种或多种症状、延迟疾病或病症或其一种或多种症状的发作或抑制疾病或病症或其一种或多种症状的进展的临床干预,如本文所述。在一些实施方案中,治疗可以在一种或多种症状已经发展之后和/或在疾病已经诊断之后施用。在其他实施方案中,治疗可以在没有症状的情况下施用,例如,以预防症状或延迟症状的发作或抑制疾病的发作或进展。例如,治疗可以在症状发作之前施用于易感个体(例如,根据症状史和/或根据遗传或其他易感性因素)。治疗也可以在症状已经消退后继续,例如,以预防或延迟其复发。

[0217] 尿嘧啶糖基化酶抑制剂

[0218] 如本文所用,术语“尿嘧啶糖基化酶抑制剂”或“UGI”是指能够抑制尿嘧啶-DNA糖基化酶碱基切除修复酶的蛋白质。在一些实施方案中,UGI结构域包含野生型UGI或如SEQ ID NO:272所示的UGI。在一些实施方案中,本文提供的UGI蛋白包括UGI片段和与UGI或UGI片段同源的蛋白质。例如,在一些实施方案中,UGI结构域包含SEQ ID NO:272所示的氨基酸序列的片段。在一些实施方案中,UGI片段包含如下氨基酸序列:包含SEQ ID NO:272所示的氨基酸序列的至少60%、至少65%、至少70%、至少75%、至少80%、至少85%、至少90%、至少95%、至少96%、至少97%、至少98%、至少99%或至少99.5%。在一些实施方案中,UGI包含与SEQ ID NO:272所示的氨基酸序列同源的氨基酸序列,或与SEQ ID NO:272所示的氨基酸序列的片段同源的氨基酸序列。在一些实施方案中,包含UGI或UGI的片段或者UGI或UGI片段的同源物的蛋白质被称为“UGI变体”。UGI变体与UGI或其片段共享同源性。例如,UGI变体与野生型UGI或如SEQ ID NO:272所示的UGI至少70%相同、至少75%相同、至少80%相

同、至少85％相同、至少90％相同、至少95％相同、至少96％相同、至少97％相同、至少98％相同、至少99％相同、至少99.5％相同或至少99.9％相同。在一些实施方案中,UGI变体包含UGI的片段,使得该片段与野生型UGI或如SEQ ID NO:272所示的UGI的相应片段至少70％相同、至少80％相同、至少90％相同、至少95％相同、至少96％相同、至少97％相同、至少98％相同、至少99％相同、至少99.5％相同或至少99.9％。在一些实施方案中,UGI包含以下氨基酸序列:

[0219] MTNLSDIIEKETGKQLVIQESILMLPEEVEEVIGNKPESDILVHTAYDESTDENVMLLTSDAPEYKPWALVIQDSNGENKIKML(SEQ ID NO:272)(P14739|UNGI_BPPB2尿嘧啶-DNA糖基化酶抑制剂)。

[0220] 变体

[0221] 如本文所用,术语“变体”是指具有偏离自然界中存在的蛋白质的特征但保留至少一种功能(即结合、相互作用或酶促能力和/或治疗特性)的蛋白质。变体与野生型蛋白至少约70％相同、至少约80％相同、至少约90％相同、至少约95％相同、至少约96％相同、至少约97％相同、至少约98％相同、至少约99％相同、至少约99.5％相同或至少约99.9％相同。例如,Cas9的变体可以包含与野生型Cas9氨基酸序列相比在氨基酸残基中具有一个或多个变化的Cas9。作为另一个示例,脱氨酶的变体可以包含与野生型脱氨酶氨基酸序列相比在氨基酸残基中具有一个或多个变化的脱氨酶,例如,在脱氨酶的祖先序列重建之后。这些变化包括化学修饰,包括不同氨基酸残基的置换、截短、共价添加(例如标签)和任何其他突变。该术语还涵盖参考序列的环状置换体、突变体、截短体或结构域,并且其表现出与参考序列相同或基本上相同的一种或多种功能活性。该术语还包括野生型蛋白质的片段。

[0222] 相对于野生型蛋白质,保留特性的水平或程度可能降低,但通常在种类上相同或相似。通常,变体总体上非常相似,并且在许多区域中与本文所述的蛋白质的氨基酸序列相同。本领域技术人员将理解如何制备和使用保持全部或至少一些功能能力或特性的变体。

[0223] 变体蛋白可以包含与例如野生型蛋白或本文提供的任何蛋白(例如SMN蛋白)的氨基酸序列至少80％、85％、90％、95％、96％、97％、98％、99％或100％相同的氨基酸序列,或替代性地由其组成。

[0224] 所谓具有与查询氨基酸序列至少例如95％“相同”的氨基酸序列的多肽,是指除了主题多肽序列可以包括查询氨基酸序列的每100个氨基酸最多至五个氨基酸改变之外,主题多肽的氨基酸序列与查询序列相同。换句话说,为了获得具有与查询氨基酸序列至少95％相同的氨基酸序列的多肽,主题序列中最多至5％的氨基酸残基可以被插入、缺失或用另一个氨基酸置换。参考序列的这些改变可以发生在参考氨基酸序列的氨基端或羧基端位置或这些末端位置之间的任何位置,单独地散布在参考序列中的残基之间或散布在参考序列内的一个或多个连续基团中。

[0225] 实际上,任何特定多肽是否与例如蛋白质如SMN蛋白的氨基酸序列至少80％、85％、90％、95％、96％、97％、98％或99％相同,可以使用已知的计算机程序按惯例确定。用于确定查询序列(本发明的序列)和主题序列之间的最佳总体匹配(也称为全局序列比对)的优选方法,可以使用基于Brutlag et al.(Comp.App.Biosci.6:237-245(1990))的算法的FASTDB计算机程序来确定。在序列比对中,查询序列和主题序列两者都是核苷酸序列或两者都是氨基酸序列。所述全局序列比对的结果表示为同一性百分比。用于FASTDB氨基酸比对的优选参数是:矩阵＝PAM 0、k-tuple＝2、错配罚分＝1、连接罚分＝20、随机化组长度

=0、截止分数=1、窗口大小=序列长度、空位罚分=5、空位大小罚分=0.05、窗口大小=500或主题氨基酸序列的长度,以较短者为准。

[0226] 如果由于N端或C端缺失,而不是由于内部缺失,主题序列比查询序列短,则必须对结果进行手动校正。这是因为FASTDB程序在计算全局同一性百分比时没有考虑主题序列的N端和C端截短。对于相对于查询序列在N端和C端被截短的主题序列,通过计算查询序列中位于主题序列N端和C端的残基数(这些残基与相应的主题残基不匹配/不对齐)占查询序列总碱基的百分比来校正百分比同一性。残基是否匹配/对齐由FASTDB序列比对的结果确定。然后从由上述FASTDB程序使用指定参数计算的百分比同一性中减去该百分比,以得到最终的百分比同一性分数。该最终百分比同一性分数用于本发明的目的。出于手动调整同一性百分比分数的目的,仅考虑与查询序列不匹配/不对齐的主题序列的N端和C端的残基。也就是说,仅查询主题序列最远的N端和C端残基之外的残基位置。

[0227] 载体

[0228] 如本文所用,术语“载体”是指可以被修饰以编码感兴趣的基因并且能够进入宿主细胞、突变并在宿主细胞内复制,然后将载体的复制形式转移到另一宿主细胞中的核酸。示例性的合适载体包括病毒载体,如逆转录病毒载体或噬菌体和丝状噬菌体,以及接合质粒。基于本公开,另外的合适的载体对于本领域技术人员将是显而易见的。

[0229] 野生型

[0230] 如本文所用,术语“野生型”是本领域技术人员所理解的术语,并且意指生物体、株、基因或特征在自然界中存在的典型形式,以区别于突变体或变体形式。

具体实施方式

[0231] 本公开提供了包含进化定向的腺苷脱氨酶结构域(例如,如本文所述优先脱氨DNA中的胞苷的腺苷脱氨酶TadA的变体)和能够结合特定核苷酸序列的napDNAbp结构域(例如,Cas9蛋白)的胞嘧啶碱基编辑器,其中腺苷脱氨酶变体为碱基编辑器(TadCBE)提供了更小的尺寸和更低的脱靶效应,同时保持了现有CBE的高编辑效率。TadCBE对胞苷的脱氨可以导致从胞嘧啶(C)至(T)的点突变(本文称为核酸编辑的过程),从而将C·G碱基对转化为T·A碱基对。此类碱基编辑器可用于靶向编辑核酸序列,如DNA分子。此类碱基编辑器尤其可用于DNA的体外靶向编辑,例如用于生成突变体细胞或动物。此类碱基编辑器可用于在活体哺乳动物细胞中引入靶向突变。此类碱基编辑器还可用于在离体细胞中引入靶向突变以校正遗传缺陷,例如在获自受试者并随后重新引入同一或另一受试者的细胞中,或用于基因组中多个基因的多路复用编辑。这些碱基编辑器可用于体内引入靶向突变,例如在受试者中校正遗传缺陷或引入疾病相关基因的失活突变,或用于基因组的多路复用编辑。本文所述的胞嘧啶碱基编辑器可以用于T至C突变的靶向编辑(例如,靶向基因组编辑)。本发明提供了利用本文提供的脱氨酶和碱基编辑器的脱氨酶、碱基编辑器、核酸、载体、细胞、组合物、方法、试剂盒和用途。

[0232] 此处,PACE和PANCE被用来改变TadA-8e的底物特异性,从而产生了一类新的选择性胞苷脱氨酶(TadA-CD)和胞嘧啶碱基编辑器(图1A)。为了实现胞苷脱氨,TadA-CD变体在活性位点附近与DNA主链相互作用的残基处获得了突变。所公开的TadA-CD胞嘧啶碱基编辑器(TadCBE)是高度活性的,并且在哺乳动物细胞中的各种位点上表现出与当前BE4max、

evoAPOBEC1-BE4max(evoA)和evoFERNY-BE4max(evoFERNY)CBE相当或更高的C·G至T·A转化效率。TadA-CD还与SpCas9(PAM=NGG)和经进化的eNme2-C Cas9(PAM=N_4CN)变体兼容,促进了广泛的靶标可及性。脱靶分析显示,TadCBE诱导比广泛使用的基于APOBEC的CBE变体更低的Cas非依赖性脱靶DNA和RNA编辑。添加V106W突变[9,34]进一步减少TadCBE的脱靶编辑,优化它们的编辑窗口,并改进C·G至T·A的选择性,同时保持了峰值中靶编辑效率。本文通过在小鼠胚胎干细胞(mESC)中使用10,638个基因组整合的、高度可变靶位点的文库对经进化的TadCBE进行了广泛表征,以确定TadCBE的选择性和序列背景偏好。TadA-CD还与SpCas9和经进化的eNme2-C Cas9变体两者兼容,促进了广泛的靶标可及性。所公开的TadCBE可用于在人细胞中在治疗相关基因座进行高效的胞嘧啶碱基编辑,包括多路复用编辑,并且特别是在原代人造血干细胞和祖细胞(HSPC)中在治疗相关位点进行胞嘧啶编辑。这些公开的TadCBE在例如原代人T细胞中的CXCR5和CCR5基因处表现出比现有CBE更精确的编辑窗口和更少的旁观者编辑。本公开提供了一类新的小型CBE,具有高中靶活性、便于精确碱基编辑的明确定义的编辑窗口以及低脱靶活性,并确立了腺苷脱氨酶进化为选择性胞苷脱氨酶的潜力。

[0233] 在一些方面,本公开涉及具有靶向胞嘧啶活性的腺苷脱氨酶(例如,TadA-CD)。在一些实施方案中,TadA-CD是从先前工程化为作用于单链DNA(而不是RNA)以用于腺苷碱基编辑应用的大肠杆菌tRNA腺苷脱氨酶(例如,TadA-8e)进化而来的。本领域技术人员将了解,PACE和PANCE方法可用于在TadA-8e结构域中引入额外的突变,从而改变酶的底物特异性,以产生TadA-CD。在一些实施方案中,TadA-CD(例如,突变TadA-8e脱氨酶)与亲本TadA-8e具有80%到99.5%的序列同源性。在一些情况下,相对于亲本TadA-8e,TadA-CD脱氨酶在E27、V28和H96处包含突变,并且进一步在选自R26、M61、Y73、I76、M151、Q154和A158的残基处包含至少一个突变。

[0234] 在一些实施方案中,相对于亲本TadA-8e变体,TadA-CD变体相对于腺苷对胞嘧啶具有增强的选择性和脱氨活性。例如,在一些实施方案中,TadA-CD脱氨酶在靶序列的C_4和C_5位置将85%到92%(取决于变体类型)的C-T碱基对转化为T-A碱基对,同时腺嘌呤的编辑少于2%;而包含TadA-8e脱氨酶的碱基编辑器在靶序列的A_6位置将约90%的A-T碱基对转化为G-C碱基对,其中C-G至T-A碱基对的编辑少于2%(参见实施例2)。这代表了TadA-CD相比TadA-8e变体在胞嘧啶与腺嘌呤碱基编辑能力上的超过3000倍的变化。

[0235] 在一些方面,本公开涉及包含与具有胞苷活性的TadA-CD脱氨酶(例如,TadCBE)融合的核酸可编程DNA结合蛋白(例如,Cas9)结构域的胞嘧啶碱基编辑器(CBE)。在一些实施方案中,napDNAbp结构域包含Cas同源物、旁系同源物、直系同源物或类似物。napDNAbp结构域可以选自Cas9、Cas9n(例如,SpCas9n)、dCas9、CasX、CasY、C2c1、C2c2、C2c3、GeoCas9、CjCas9、Cas12a、Cas12b、Cas12g、Cas12h、Cas12i、Cas13b、Cas13c、Cas13d、Cas14、Csn2、xCas9、Cas9-NG、LbCas12a、enAsCas12a、SaCas9、SaCas9-KKH、环状排列的Cas9、Argonaute(Ago)结构域、SmacCas9、Spy-macCas9、SpCas9-VRQR、SpCas9-NRRH、SpaCas9-NRTH、SpCas9-NRCH、eNme2Cas9、eNme2-C Cas9、enCjCas9、SauriCas9、Cas9-NG-VRQR和/或其变体。在某些实施方案中,napDNAbp结构域包含或是衍生自酿脓链球菌或金黄色葡萄球菌的Cas9结构域或Cas12a结构域。在一些情况下,napDNAbp结构域是衍生自脑膜炎奈瑟菌的Nme2Cas9结构域。在一些实施方案中,napDNAbp结构域包含核酸酶死亡的Cas9(dCas9)结构域、Cas9切口

酶(nCas9)结构域或核酸酶活性Cas9结构域。在一些情况下,napDNAbp结构域是CjCas9。在各种实施方案中,napDNAbp结构域是切口酶。

[0236] 所公开的CBE表现出低水平的非期望编辑,如低Cas9非依赖性脱靶编辑。所公开的CBE在用于编辑核酸中的靶序列的方法后,表现出较少的插入和/或缺失(indel)以及对RNA分子的非期望编辑。对于若干治疗相关位点和细胞类型,所公开的CBE还表现出超过最常用CBE的编辑效率。

[0237] 在一些方面,TadA-CD表现出比天然胞嘧啶碱基编辑器更窄的编辑窗口,同时保持相当或更高的最大编辑效率。综上所述,TadCBE的小尺寸、其与eNme2Cas9(和eNme2-CCas9)的兼容性、其更集中的编辑窗口以及其相对于腺嘌呤碱基编辑对胞嘧啶的高编辑效率和选择性,证明了它们适用于各种精确胞嘧啶碱基编辑应用。

[0238] 本公开的其他方面涉及包含本文所述的TadCBE和一个或多个向导RNA(例如,单一向导RNA“sgRNA”)的组合物。此外,本公开提供了编码和/或表达本文所述的TadCBE的核酸分子,以及用于表达本文所述的TadCBE和/或gRNA的表达载体或构建体(例如,AAV载体),包含所述核酸分子和表达载体的宿主细胞,以及一个或多个gRNA,以及用于递送和/或施用本文所述的基于核酸的实施方案的组合物。特别是,本公开提供了改进的递送所公开的碱基编辑器的方法,例如向受试者递送。将所公开的TadCBE变体作为RNP而不是DNA质粒递送,通常会增加中靶:脱靶DNA编辑比率。将所公开的TadCBE变体作为mRNA分子递送(例如,使用电穿孔)可能会增加编辑效率。已在国际公布号WO/2019/226953(2019年11月28日发布)和Komoret al.,Sci.Adv.2017;3:eaao4774中描述了在体内具有约50%的明显中靶编辑效率的CBE,其中每篇通过引用并入本文。所公开的CBE可以表现出对靶胞嘧啶碱基的更高中靶编辑效率。

[0239] 本文还提供了将任何所公开的TadCBE与核酸分子(例如,包含靶序列的核酸分子(例如,DNA))接触的方法。在所公开的方法的一些实施方案中,观察到低脱靶DNA和/或RNA编辑效应。在一些实施方案中,核酸分子包含DNA,例如单链DNA或双链DNA。核酸分子的靶序列可以包含含有胞嘧啶(C)的靶核碱基对。靶序列可以包含在基因组内,例如人基因组内。靶序列可以包含与疾病或病症如镰状细胞病或HIV/AIDS相关的序列,例如,具有点突变的靶序列。在其他实施方案中,靶核苷酸序列在啮齿动物(如小鼠或大鼠)的基因组中。在其他实施方案中,靶核苷酸序列在家养动物(如马、猫、狗或兔)的基因组中。在一些实施方案中,靶核苷酸序列在研究动物的基因组中。在一些实施方案中,靶核苷酸序列在基因工程化非人受试者的基因组中。在一些实施方案中,靶核苷酸序列在植物的基因组中。在一些实施方案中,靶核苷酸序列在微生物(如细菌)的基因组中。

[0240] 此外,本公开提供了生成本文所述的TadCBE的方法,以及使用碱基编辑器或编码任何这些碱基编辑器的核酸分子在包括编辑核酸分子(例如,基因组)的应用中的方法。在某些实施方案中,工程化本文提供的碱基编辑器的方法涉及噬菌体辅助连续进化(PACE)系统或非连续系统(例如,PANCE),其可用于进化碱基编辑器的一个或多个组分(例如,脱氨酶结构域)。在某些实施方案中,在成功进化碱基编辑器的一个或多个组分(例如,脱氨酶结构域)后,制造碱基编辑器的方法包括本领域技术人员已知的重组蛋白表达方法和技术。示例性碱基编辑器通过将腺苷脱氨酶结构域与本文公开的多种napDNAbp结构域(例如,Cas9结构域)的任一种融合或缔合来制造。

[0241] 不希望受任何特定理论束缚,本文所述的TadCBE通过使用TadA变体脱氨C碱基,经由尿嘧啶的形成引起C至T突变,从而在核酸底物中诱导编辑。据信,将一个或多个尿嘧啶DNA糖基化酶抑制剂与CBE的脱氨酶和napDNAbp结构域融合抑制先天DNA修复过程,当与被工程化为使未编辑DNA链(例如,含有原始C-G靶碱基对的G的链)产生切口的核酸可编程DNA结合蛋白(例如,dCas9)偶联时,导致原始C·G碱基对转化为T·A碱基对。不希望受任何特定理论束缚,据信所公开的脱氨酶中残基26-28的突变(相对于TadA8e脱氨酶)促进了DNA底物的主链的"滑动",使得该腺苷脱氨酶的结合口袋能够接受胞嘧啶。

[0242] 在一些实施方案中,本文所述的TadCBE已被工程化为表现出高度靶向和高效的编辑能力。此类TadCBE可用于例如靶向和逆转疾病相关基因中的单核苷酸多态性(SNP),如与镰状细胞病和HIV/AIDS相关的基因。然而,在一些情况下,本文所述的TadCBE可以允许将靶C至T、A和G的混合物的置换。例如,缺乏UGI结构域的TadCBE可用作靶向随机体内诱变的筛选平台。更具体地说,它们可以用作正向遗传工具,以筛选在碱基分辨率上的功能获得和/或功能丧失变体。

[0243] 脱氨酶结构域

[0244] 本公开提供了从现有腺苷碱基编辑器(ABE)的腺苷脱氨酶结构域进化的胞苷碱基编辑器(TadCBE)。本文使用的腺苷脱氨酶是使用标准方法进化的,以在哺乳动物DNA中将腺苷(A)转化为肌苷(I)。此类腺苷脱氨酶可以导致A:T至G:C碱基对转化。最先进的ABE是ABE7.10,其在国际公布号WO 2018/027078(2018年8月2日公布)中公开。最近生成的ABE是ABE8e,其含有包含称为TadA8e的单脱氨酶变体的腺苷脱氨酶结构域,如国际公布号WO 2021/158921(2021年8月12日公布)中所述。TadA8e相对于TadA7.10(ABE7.10的腺苷脱氨酶)含有九个突变。TadA7.10也是ABEmax的脱氨酶结构域,ABEmax是ABE7.10的变体,已针对人类细胞中的表达进行了密码子优化。

[0245] 在一些实施方案中,腺苷脱氨酶是已知腺苷脱氨酶TadA7.10的变体,其包含与野生型ecTadA(SEQ ID NO:325)相比的以下突变:W23R、H36L、P48A、R51L、L84F、A106V、D108N、H123Y、S146C、D147Y、R152P、E155V、I156F和K157N。在一些实施方案中,所公开的腺苷脱氨酶是衍生自除大肠杆菌以外的物种,如金黄色葡萄球菌、伤寒沙门氏菌、腐败希瓦氏菌、流感嗜血杆菌、新月柄杆菌或枯草芽孢杆菌的TadA的变体。

[0246] 本文公开的进化实验的底物是TadA-8e,其含有相对于TadA7.10的以下突变:A109S、T111R、D119N、H122N、Y147D、F149Y、T166I和D167N。关于噬菌体辅助进化实验方法的公开,请参考国际公布号WO 2018/027078;国际公布号WO 2019/079347(2019年4月25日公布);国际公布号WO 2019/226593(2019年11月28日公布);美国专利公布号2018/0073012(2018年3月15日公布),其于2018年10月30日发布为美国专利号10,113,163;美国专利公布号2017/0121693(2017年5月4日公布),其于2019年1月1日发布为美国专利号10,167,457;国际公布号WO 2020/214842(2020年10月22日公布)、和国际专利申请号PCT/US2020/033873(2020年5月20日提交)、国际公布号WO 2020/236982(2020年11月26日公布)和国际公布号WO 2021/158921,其中每篇的内容通过引用以其整体并入本文。

[0247] 本文提供了用于经进化的腺苷脱氨酶的示例性、非限制性实施方案。在一些实施方案中,任何所公开的碱基编辑器的腺苷脱氨酶结构域包含单一腺苷脱氨酶或单体。在一些实施方案中,腺苷脱氨酶结构域包含2、3、4或5个腺苷脱氨酶。在一些实施方案中,腺苷脱

氨酶结构域包含两个腺苷脱氨酶或二聚体。在一些实施方案中,脱氨酶结构域包含工程化(或进化)的脱氨酶和野生型脱氨酶(如,野生型大肠杆菌衍生的脱氨酶)的二聚体。应当理解,本文提供的突变(例如,ecTadA中的突变)可以应用于其他腺苷碱基编辑器中的腺苷脱氨酶,例如以下中提供的那些:国际公布号WO 2018/027078(2018年8月2日公布);国际公布号WO 2019/079347(2019年4月25日公布);国际申请号PCT/US2019/033848(2019年5月23日提交),其于2019年11月28日公布为国际公布号WO 2019/226593;美国专利公布号2018/0073012(2018年3月15日公布),其于2018年10月30日发布为美国专利号10,113,163;美国专利公布号2017/0121693(2017年5月4日公布),其于2019年1月1日发布为美国专利号10,167,457;国际公布号WO 2017/070633(2017年4月27日公布);美国专利公布号2015/0166980(2015年6月18日公布);美国专利号9,840,699(2017年12月12日发布);美国专利号10,077,453(2018年9月18日发布)和国际专利申请号PCT/US2020/28568(2020年4月16日提交);其中所有通过引用以其整体并入本文。

[0248] 下文公开了根据本公开可以进化为胞苷脱氨酶的示例性腺苷脱氨酶底物。提供了衍生自枯草芽孢杆菌(完整序列如SEQ ID NO:318所示)、金黄色葡萄球菌(SEQ ID NO:317)和酿脓链球菌(SEQ ID NO:354)的示例性TadA脱氨酶。显示了大肠杆菌TadA-8e中的氨基酸置换,以及枯草芽孢杆菌、金黄色葡萄球菌和酿脓链球菌TadA脱氨酶中的同源突变。因此,本领域技术人员将能够在任何天然存在的腺苷脱氨酶(例如,具有与ecTadA的同源性)中生成与本文所述的任何突变(例如,ecTadA中鉴定的任何突变)相对应的突变。在一些实施方案中,腺苷脱氨酶衍生自原核生物。在一些实施方案中,腺苷脱氨酶来自细菌。在一些实施方案中,腺苷脱氨酶来自大肠杆菌、金黄色葡萄球菌、伤寒沙门氏菌、腐败希瓦氏菌、流感嗜血杆菌、新月柄杆菌或枯草芽孢杆菌。在一些实施方案中,腺苷脱氨酶来自大肠杆菌。本领域技术人员将能够通过本领域熟知的方法(例如,通过序列比对和同源残基的确定)在任何同源蛋白中和在相应编码核酸中鉴定相应残基。

[0249] 在一些实施方案中,腺苷脱氨酶底物包含TadA9或其变体。TadA9相对于TadA-8e含有V82S和Q154R置换。(换句话说,TadA9相对于TadA7.10含有Y147R、Q154R和I76Y突变。)在一些实施方案中,腺苷脱氨酶包含与TadA9的氨基酸序列(SEQ ID NO:33)至少80%、至少85%、至少90%、至少95%、至少96%、至少97%、至少98%、至少99%或至少99.5%相同的氨基酸序列。TadA9在文献中可能被称为TadA*8.9。含有TadA9脱氨酶的ABE在本文中称为ABE9。TadA9更详细的描述于Gaudelli et al.,Nat Biotechnol.2020Jul;38(7):892-900和PCT公布号WO 2021/050571(2021年3月18日公布)中,其中每篇通过引用并入本文。

[0250] 在一些实施方案中,腺苷脱氨酶底物包含TadA20、TadA-8.17-m(TadA17)或其变体。TadA20相对于TadA7.10含有I76Y、V82S、Y123H、Y147R和Q154R置换。TadA17相对于TadA7.10含有V82S和Q154R置换。TadA20和TadA17更详细的描述于Gaudelli et al.,Nat Biotechnol.2020Jul;38(7):892-900和WO 2021/050571(2021年3月18日公布)中。TadA20在文献中可能被称为TadA*8.20。在一些实施方案中,腺苷脱氨酶包含与TadA20的氨基酸序列(SEQ ID NO:326)至少80%、至少85%、至少90%、至少95%、至少96%、至少97%、至少98%、至少99%或至少99.5%相同的氨基酸序列。含有TadA20脱氨酶的ABE在本文中称为ABE20。其在本领域中可以被称为ABE8.20、ABE8.20-d或ABE8.20-m。含有TadA17脱氨酶的ABE在本文中称为ABE17。其在本领域中可以被称为ABE8.17或ABE8.17-m。

[0251] 在一些实施方案中,腺苷脱氨酶包含与SEQ ID NO:317-323中的任何氨基酸序列至少80%、至少85%、至少90%、至少95%、至少96%、至少97%、至少98%、至少99%或至少99.5%相同的氨基酸序列。

[0252] 在某些实施方案中,腺苷脱氨酶结构域包含具有与以下中的一个具有至少80%、至少85%、至少90%、至少95%、至少98%、至少99%或至少99.5%序列同一性的序列的腺苷脱氨酶:

[0253] TadA 7.10(大肠杆菌):

[0254] MSEVEFSHEYWMRHALTLAKRARDEREVPVGAVLVLNNRVIGEGWNRAIGLHDPTAHAEIMALRQGGLVMQNYRLIDATLYVTFEPCVMCAGAMIHSRIGRVVFGVRNAKTGAAGSLMDVLHYPGMNHRVEITEGILADECAALLCYFFRMPRQVFNAQKKAQSSTD(SEQ ID NO:315)

[0255] TadA-8e(大肠杆菌):

[0256] SEVEFSHEYWMRHALTLAKRARDEREVPVGAVLVLNNRVIGEGWNRAIGLHDPTAHAEIMALRQGGLVMQNYRLIDATLYVTFEPCVMCAGAMIHSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDFYRMPRQVFNAQKKAQSSIN(SEQ ID NO:350)

[0257] Tad1:

[0258] SEVEFSHEYWMRHALTLAKRARDEGEVPVGAVLVLNNRVIGEGWNRAIGLYDPTAHAEIMALRQGGLVMQNYRLIDATLYVTFEPCVMCAGAMIHSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMDHRVEITEGILADECAALLCDF(SEQ ID NO:2)

[0259] Tad2:

[0260] SEVEFSHEYWMRHALTLAKRARDEGEVPVGAVLVLNNRVIGEGWNRAIGLHDPTAHAEIMALRQGGLVMQNYRLIDATLYVTFEPCVMCAGAMIHSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMDHRVEITEGILADECAALLCDFYRMPRQVFNAQKKAQSSIN(SEQ ID NO:2)

[0261] Tad3:

[0262] SEVEFSHEYWMRHALTLAKRARDEREVPVGAVLVLNNRVIGEGWNRAIGLHDPTAHAEIMALRQGGLVMQNYGLIDATLYVTFEPCVMCAGAIIHSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDFYRMPRQVFNAQKKAQSSIN(SEQ ID NO:3)

[0263] Tad4:

[0264] SEVEFSHEYWMRHALTLAKRARDEREVPVGAVLVLNNRVIGEGWNRAIGLHDPTAHAEIMALRQGGLVMQNYRLIDATLYVTFEPCVMCAGAMIHSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMDHRVEITEGILADECAALLCDFYRMPRQVFNAQKKAQSSIN(SEQ ID NO:4)

[0265] Tad6:

[0266] SEVEFSHEYWMRHALTLAKRARDEGEVPVGAVLVLNNRVIGEGWNRAIGLYDPTAHAEIMALRQGGLVMQNYGLIDATLYVTFEPCVMCAGAMIHSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMDHRVEITEGILADECAALLCDFYRMPRQVFNAQKKAQSSIN(SEQ ID NO:5)

[0267] Tad6-SR:

[0268] SEVEFSHEYWMRHALTLAKRARDEGEVPVGAVLVLNNRVIGEGWNRAIGLYDPTAHAEIMALRQGGLVMQNYGLIDATLYSTFEPCVMCAGAMIHSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMDHRVEITEGILADECAALLCDFYRMPRRVFNAQKKAQSSIN(SEQ ID NO:6)

[0269] TadA9:

[0270] SEVEFSHEYWMRHALTLAKRARDEGEVPVGAVLVLNNRVIGEGWNRAIGLYDPTAHAEIMALRQGGLV MQNYRLIDATLYVTFEPCVMCAGAMIHSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMDHRVEITEGILANECAALL CDFYRMPRQVFNAQKKAQSSIN(SEQ ID NO:33)

[0271] TadA20

[0272] SEVEFSHEYWMRHALTLAKRARDEREVPVGAVLVLNNRVIGEGWNRAIGLHDPTAHAEIMALRQGGLV MQNYRLYDATLYSTFEPCVMCAGAMIHSRIGRVVFGVRNAKTGAAGSLMDVLHHPGMNHRVEITEGILADECAALL CRFFRMPRRVFNAQKKAQSSTD(SEQ ID NO:326)

[0273] 枯草芽孢杆菌TadA:

[0274] MGSHMTNDIYFMTLAIEEAKKAAQLGEVPIGAIITKDDEVIARAHNLRETLQQPTAHAEHIAIERAAK VLGSWRLEGCTLYVTLEPCVMCAGTIVMSRIPRVVYGADDPKGGCSGSLMNLLQQSNFNHRAIVDKGVLKEACSTL LTTFFKNLRANKKSTN(SEQ ID NO:317)

[0275] 金黄色葡萄球菌TadA:

[0276] MTQDELYMKEAIKEAKKAEEKGEVPIGAVLVINGEIIARAHNLRETEQRSIAHAEMLVIDEACKALGT WRLEGATLYVTLEPCPMCAGAVVLSRVEKVVFGAFDPKGGCSGTLMNLLQEERFNHQAEVVSGVLEEECGGMLSAF FRELRKKKKAARKNLSE(SEQ ID NO:318)

[0277] 伤寒沙门氏菌(S.typhimurium)TadA:

[0278] MPPAFITGVTSLSDVELDHEYWMRHALTLAKRAWDEREVPVGAVLVHNHRVIGEGWNRPIGRHDPTAH AEIMALRQGGLVLQNYRLLDTTLYVTLEPCVMCAGAMVHSRIGRVVFGARDAKTGAAGSLIDVLHHPGMNHRVEII EGVLRDECATLLSDFFRMRRQEIKALKKADRAEGAGPAV(SEQ ID NO:319)

[0279] 腐败希瓦氏菌(S.putrefaciens)TadA:

[0280] MDEYWMQVAMQMAEKAEAAGEVPVGAVLVKDGQQIATGYNLSISQHDPTAHAEILCLRSAGKKLENYR LLDATLYITLEPCAMCAGAMVHSRIARVVYGARDEKTGAAGTVVNLLQHPAFNHQVEVTSGVLAEACSAQLSRFFK RRRDEKKALKLAQRAQQGIE(SEQ ID NO:320)

[0281] 流感嗜血杆菌F3031(H.influenzae)TadA:

[0282] MDAAKVRSEFDEKMMRYALELADKAEALGEIPVGAVLVDDARNIIGEGWNLSIVQSDPTAHAEIIALR NGAKNIQNYRLLNSTLYVTLEPCTMCAGAILHSRIKRLVFGASDYKTGAIGSRFHFFDDYKMNHTLEITSGVLAEE CSQKLSTFFQKRREEKKIEKALLKSLSDK(SEQ ID NO:321)

[0283] 新月柄杆菌(C.crescentus)TadA:

[0284] MRTDESEDQDHRMMRLALDAARAAAEAGETPVGAVILDPSTGEVIATAGNGPIAAHDPTAHAEIAAMR AAAAKLGNYRLTDLTLVVTLEPCAMCAGAISHARIGRVVFGADDPKGGAVVHGPKFFAQPTCHWRPEVTGGVLADE SADLLRGFFRARRKAKI(SEQ ID NO:322)

[0285] 硫还原地杆菌(G.sulfurreducens)TadA:

[0286] MSSLKKTPIRDDAYWMGKAIREAAKAAARDEVPIGAVIVRDGAVIGRGHNLREGSNDPSAHAEMIAIR QAARRSANWRLTGATLYVTLEPCLMCMGAIILARLERVVFGCYDPKGAAGSLYDLSADPRLNHQVRLSPGVCQEEC GTMLSDFFRDLRRRKKAKATPALFIDERKVPPEP(SEQ ID NO:323)

[0287] 酿脓链球菌(S.pyogenes)TadA:

[0288] MPYSLEEQTYFMQEALKEAEKSLQKAEIPIGCVIVKDGEIIGRGHNAREESNQAIMHAEIMAINEANA HEGNWRLLDTTLFVTIEPCVMCSGAIGLARIPHVIYGASNQKFGGADSLYQILTDERLNHRVQVERGLLAADCANI MQTFFRQGRERKKIAKHLIKEQSDPFD(SEQ ID NO:354)

[0289] 超嗜热菌(A.aeolicus)TadA:

[0290] MGKEYFLKVALREAKRAFEKGEVPVGAIIVKEGEIISKAHNSVEELKDPTAHAEMLAIKEACRRLNTKYLEGCELYVTLEPCIMCSYALVLSRIEKVIFSALDKKHGGVVSVFNILDEPTLNHRVKWEYYPLEEASELLSEFFKKLRNNII(SEQ ID NO:355)

[0291] 在一些实施方案中,TadA脱氨酶是全长大肠杆菌TadA脱氨酶(ecTadA)。例如,在某些实施方案中,腺苷脱氨酶结构域包含含有以下氨基酸序列的脱氨酶:

[0292] MRRAFITGVFFLSEVEFSHEYWMRHALTLAKRAWDEREVPVGAVLVHNNRVIGEGWNRPIGRHDPTAHAEIMALRQGGLVMQNYRLIDATLYVTLEPCVMCAGAMIHSRIGRVVFGARDAKTGAAGSLMDVLHHPGMNHRVEITEGILADECAALLSDFFRMRRQEIKAQKKAQSSTD(SEQ ID NO:325)

[0293] TadA衍生的胞苷脱氨酶(TadA-CD)

[0294] 本公开的某些方面涉及具有增强的胞嘧啶特异性和胞苷脱氨活性的经进化的腺苷脱氨酶。根据某些实施方案,经进化的脱氨酶能够脱氨DNA中的胞苷。在一些实施方案中,脱氨酶是使用连续和/或非连续的实验室指导方法(例如,PACE和PANCE)从亲本腺苷脱氨酶进化的。在一些实施方案中,使用PACE和/或PANCE进化的亲本腺苷脱氨酶具有胞苷脱氨酶活性。本公开的脱氨酶可以从迄今为止已报道具有腺苷脱氨酶活性的任何腺苷脱氨酶进化,如例如国际专利申请号PCT/US2017/045381(2017年8月3日提交)、国际专利申请号PCT/US2020/028568(2020年4月16日提交)、国际专利申请号PCT/US2021/016827(2021年2月5日提交)、国际专利申请号PCT/US2022/073781(2022年7月15日提交)中描述的那些,其中所有通过引用以其整体并入本文。在一些情况下,亲本脱氨酶包含大肠杆菌tRNA腺苷脱氨酶(TadA)。本申请的脱氨酶可以从先前突变的(即经进化的)亲本TadA变体进化,如例如国际专利申请号PCT/US2021/016827(2021年2月5日提交,2021年8月12日公布为WO 2021/158921)中描述的那些。例如,在一些实施方案中,亲本腺苷脱氨酶是TadA7.10。在其他实施方案中,亲本腺苷脱氨酶是TadA8e变体,其相对于TadA7.10含有额外的8个突变:A109S、T111R、D119N、H122N、Y147D、F149Y、T166I和D167N。其他亲本腺苷脱氨酶底物也是可能的。

[0295] 在一些实施方案中,本申请的TadA衍生的胞苷脱氨酶使用噬菌体辅助连续进化(PACE)和非连续进化(PANCE)的组合衍生自亲本腺苷脱氨酶(例如,TadA-8e)。根据某些实施方案,亲本腺苷脱氨酶包含与SEQ ID NO:41的氨基酸序列至少80%相同、至少85%相同、至少90%相同、至少95%、至少98%相同、至少99%相同和至少99.5%相同的氨基酸序列。在一些情况下,亲本腺苷脱氨酶包含SEQ ID NO:41的序列。

[0296] 在一些实施方案中,经进化的TadA衍生的胞苷脱氨酶是至少部分地与亲本TadA-8e变体同源的。例如,根据某些实施方案,TadA衍生的胞苷脱氨酶(例如,TadA-CD)包含与SEQ ID NO:41的氨基酸序列至少80%相同、至少85%相同、至少90%相同、至少95%相同、至少98%相同、至少99%相同和至少99.5%相同的氨基酸序列,其中SEQ ID NO:41的残基27是除E(谷氨酸)以外的任何氨基酸。具有其他序列同源性的TadA-CD也是可能的。例如,在某些实施方案中,TadA衍生的胞苷脱氨酶(例如,TadA-CD)包含与SEQ ID NO:41的氨基酸序列至少80%相同、至少85%相同、至少90%相同、至少95%相同、至少98%相同、至少99%相同和至少99.5%相同的氨基酸序列,其中SEQ ID NO:41的残基28是除V(缬氨酸)以外的任何氨基酸。在另一个示例性实施方案中,TadA衍生的胞苷脱氨酶与SEQ ID NO:41的氨基酸序列至少80%相同、至少85%相同、至少90%相同、至少95%相同、至少98%相同、至少99%

相同和至少99.5％相同,其中SEQ ID NO:41的残基96是除H(组氨酸)以外的任何氨基酸。

[0297] 本领域技术人员将理解,TadA衍生的胞苷脱氨酶(例如,TadA-CD)相对于亲本腺苷脱氨酶(例如,TadA-8e)可以包含多个突变。在一些实施方案中,本申请的脱氨酶(例如,TadA-CD)包含残基E27、V28和H96处的突变。在一些实施方案中,所公开的脱氨酶在SEQ ID NO:41的氨基酸序列中进一步包含选自R26、M61、Y73、I76、M151、Q154和A158的残基处的至少一个突变,或同源腺苷脱氨酶中的相应突变。

[0298] 在一些实施方案中,脱氨酶在SEQ ID NO:41的氨基酸序列中包含选自E27A、E27K、V28G、V28A和H96N处的至少一个突变,并进一步包含选自R26G、M61I、Y73H、Y73S、Y73C、I76F、M151I、Q154R、Q154H和A158S的残基处的至少一个突变,或同源腺苷脱氨酶中的相应突变。其他突变也是可能的。例如,在某些实施方案中,TadA-CD酶在SEQ ID NO:41的氨基酸序列中包含选自E27A、V28G和H96N的突变,并进一步包含选自R26G、M61I、Y73H、Y73S、Y73C、I76F、M151I、Q154R、Q154H和A158S的至少一个突变,或同源腺苷脱氨酶中的相应突变。

[0299] 其他示例性实施方案可以包括:(1)脱氨酶,其在SEQ ID NO:41的氨基酸序列包含E27K、V28G和H96N突变,并进一步包含中选自R26G、M61I、Y73H、Y73S、Y73C、I76F、M151I、Q154R、Q154H和A158S的至少一个突变,或同源腺苷脱氨酶中的相应突变;(2)脱氨酶,其在SEQ ID NO:41的氨基酸序列包含E27A、V28A和H96N突变,并进一步包含选自R26G、M61I、Y73H、Y73S、Y73C、I76F、M151I、Q154R、Q154H和A158S的至少一个突变,或同源腺苷脱氨酶中的相应突变;(3)脱氨酶,其在SEQ ID NO:41的氨基酸序列包含E27K、V28A和H96N突变,并进一步包含选自R26G、M61I、Y73H、Y73S、Y73C、I76F、M151I、Q154R、Q154H和A158S的至少一个突变,或同源腺苷脱氨酶中的相应突变。

[0300] 在一些实施方案中,TadA衍生的胞苷脱氨酶(TadA-CD)包含在选自R26、M61、Y73、I76、M151、Q154和A158的残基处的至少两个突变(相对于亲本脱氨酶)。在其他实施方案中,TadA-CD包含在选自R26G、M61I、Y73H、I76F、M151I、Q154H、Q154R和A158S的残基处的至少两个突变。

[0301] 在某些方面,提供了可以保留一些A至G碱基编辑活性的TadA衍生的胞苷脱氨酶。不希望受任何特定理论束缚,经由回复分析已确定TadA-8e脱氨酶(如SEQ ID NO:41所示)的残基26-28(其位于活性位点附近的环上)对于将腺苷至胞苷的转化选择性至关重要。进一步认为,底物在活性位点的定位是脱氨选择性的关键决定因素,并且序列背景可能影响靶碱基的选择性脱氨,因为TadA-CD与5’和3’之间的相互作用可能影响底物在活性位点的定位。

[0302] 再次,不希望受理论束缚,据信当腺嘌呤位于编辑窗口的中心(例如,SpCas9的原间隔区位置5或6,同时PAM为位置21-23)并且前面是T或C时,残留的A至T编辑最高。在一些实施方案中,添加V106W突变通过与C脱氨相比更大程度地抑制A脱氨来改进选择性。

[0303] 在一些方面,提供了TadA衍生的胞苷脱氨酶,其提供靶胞嘧啶至胸腺嘧啶和靶腺嘌呤至鸟嘌呤的高效转化(本文称为“TadA-Dual”脱氨酶和碱基编辑器)。TadA-Dual脱氨酶能够编辑原间隔区内的C和A碱基,并且特别是在原间隔区的编辑窗口内。这些编辑器以大致相当的效率安装A至G和C至T编辑两者。

[0304] 例如,所公开的TadA双脱氨酶以大约1.1:1的比率安装A至G和C至T编辑。在一些实施方案中,双重编辑器以0.7:1、0.8:1、0.9:1、1:1、1.1:1、1.2:1、1.3:1、1.4:1或1.5:1的比

率提供A至G和C至T编辑。其他范围也是可能的,包括超过1.5:1的比率。这些经进化的TadA脱氨酶及含有这些脱氨酶的“双”编辑器能够以与C·G至T·A几乎相同的效率编辑A·T至G·C,可能适用于筛选应用,如筛选新型Cas同源结构域和其他napDNAbp结构域针对各种靶序列的编辑活性的方法。这些脱氨酶和双重编辑器还可以用于诱变应用,如体内正向遗传诱变筛选或靶向随机诱变筛选。这些双重编辑器还可以用于多路复用编辑应用。

[0305] 双重编辑器

[0306] 在一些实施方案中,基于TadA的双重编辑器包含胞苷脱氨酶,该胞苷脱氨酶包含选自R26G、V28A、A48R、Y73S和H96N的一个、两个、三个、四个或五个突变。该双重编辑器在本文中称为TadDE,并且双重编辑脱氨酶在本文中称为TadA-CDf(例如,TadA-Dual),其具有如SEQ ID NO:39所示的氨基酸序列。

[0307] 因此,在一些实施方案中,本文提供了脱氨酶,其包含在SEQ ID NO:41的氨基酸序列中残基R26、V28、A48和Y73处的突变,或同源腺苷脱氨酶中的相应突变。本文还提供了脱氨酶,其在SEQ ID NO:41的氨基酸序列中包含残基R26、E27、V28、A48和Y73处的突变(即进一步包含E27突变)。在特定实施方案中,这些脱氨酶包含突变R26G、V28A、A48R、Y73S和H96N。在一些实施方案中,这些脱氨酶包含突变R26G、V28G、A48R和Y73C。

[0308] 如上文和本文所述,使用PACE和PANCE方法进化的优选的TadA衍生的胞苷脱氨酶可以包含一个或多个突变。例如,TadA-CD变体可以包含选自以下的至少一个突变:R26G、E27A、V28G、I76F、H96N和M151I(例如,TadA-CDa,SEQ ID NO:34);R26G、E27A、V28G、I76F、H96N和A158S(例如,TadA-CDb,SEQ ID NO:35);R26G、E27A、V28G、I76F、H96N、Q154R和A158S(例如,TadA-CDc,SEQ ID NO:36);E27A、V28G、Y73H、H96N、Q154H和A158S(例如,TadA-CDd,SEQ ID NO:37);R26G、V28A、A48R、Y73S和H96N(例如,TadA-CDe,SEQ ID NO:38);V28A、A48R和Y73S(例如,TadA-CDf,SEQ ID NO:39),以及R26G、V28G、A48R和Y73C(例如,TadA-CDg,SEQ ID NO:40)。

[0309] 在一些优选实施方案中,脱氨酶包含以下突变:R26G、E27A、V28G、I76F、H96N和A158S(例如,TadA-CDa,SEQ ID NO:34),R26G、E27A、V28G、I76F、H96N、Q154R和A158S(例如,TadA-CDb,SEQ ID NO:35),R26G、E27A、V28G、I76F、H96N和M151I(例如,TadA-CDc,SEQ ID NO:36),E27K、V28A、M61I和H96N(例如,TadA-CDd,SEQ ID NO:37),E27A、V28G、Y73H、H96N、Q154H和A158S(例如,TadA-CDe,SEQ ID NO:38),R26G、V28A、A48R、Y73S和H96N(例如,TadA-CDf,SEQ ID NO:39),以及R26G、V28G、A48R和Y73C(例如,TadA-CDg,SEQ ID NO:40)。

[0310] 本领域普通技术人员将理解,由于继承的突变类型和组合的不同,本文所述的经进化的脱氨酶可以表现出对胞嘧啶和/或腺苷碱基的不同特异性和/或脱氨活性。在一些实施方案中,TadA-CD的胞苷脱氨活性超过了TadA-8e的胞苷脱氨活性。例如,TadA-CD变体的胞苷脱氨活性可以大于或等于10倍、大于或等于20倍、大于或等于40倍、大于或等于80倍、大于或等于100倍、大于或等于200倍、大于或等于400倍、大于或等于800倍、大于或等于1000倍、大于或等于2000倍、大于或等于3000倍、大于或等于4000倍的TadA-8e的胞苷脱氨活性。在其他实施方案中,TadA-CD变体的胞苷脱氨活性低于或等于4000倍、低于或等于2000倍、低于或等于1000倍、低于或等于800倍、低于或等于400倍、低于或等于200倍、低于或等于100倍、低于或等于80倍、低于或等于40倍、低于或等于20倍或低于或等于10倍的TadA-8e的胞苷脱氨活性。

[0311] 在一些实施方案中,TadA-CD脱氨酶的腺苷脱氨活性低于TadA-8e的脱氨活性。例如,在一些情况下,TadA-CD变体的腺苷脱氨活性低于或等于4000倍、低于或等于2000倍、低于或等于1000倍、低于或等于800倍、低于或等于400倍、低于或等于200倍、低于或等于100倍、低于或等于80倍、低于或等于40倍、低于或等于20倍或低于或等于10倍的TadA-8e的腺苷脱氨活性。

[0312] 在一些实施方案中,上文和本文所述的TadA-CD变体还可以包含V106W突变。最近发现,包含V106W突变的腺苷脱氨酶TadA变体(如国际专利公布号WO 2021/214842和WO 2021/158921中描述的那些,其中每篇通过引用并入本文)具有降低的DNA和RNA的Cas非依赖性脱靶编辑同时维持高水平中靶腺苷脱氨活性。在一些实施方案中,包含V106W突变的TadA-CD变体的平均峰值编辑效率大于或等于50%、大于或等于60%、大于或等于70%、大于或等于80%和大于或等于90%。在其他实施方案中,包含V106W突变的TadA-CD变体的平均峰值编辑效率低于或等于90%、低于或等于80%、低于或等于70%、低于或等于60%或低于或等于50%。仅含有单个TadA脱氨酶结构域(而非单链二聚体)的ABE允许编辑器的大小的减少[30,31]。此外,虽然SaCas9足够小(1053个氨基酸长度,SEQ ID NO:347)以提供单一AAV兼容的碱基编辑器,但其NNGRRT PAM的稀有性极大地限制了其效用。由于碱基编辑需要合适PAM的存在以将靶核苷酸置于编辑窗口内,整体上提供广泛PAM兼容性以及简单高效体内递送的TadCBE将推动碱基编辑的体内应用。

[0313] 在一些实施方案中,表10中列出的任何一个脱氨酶还可以包含V106W突变。在一些实施方案中,TadA-CD变体包含表10中列出的任何氨基酸序列的至少80%、至少85%、至少90%、至少95%、至少98%、至少99%或至少99.5%,其中表10中列出的任何一个序列进一步包含V106W突变。

[0314] 在一些实施方案中,TadA变体包含与表10中列出的任何氨基酸序列至少80%、85%、90%、95%、98%、99%或99.5%的同一性。

[0315] 表10.

[0316]

TadA 变体	序列	SEQ ID NO:
TadA-CDa	MSEVEFSHEYWMRHALTLAKRARDEGAGPVGAVLVLNNRVIGEGWNRAIGLHDPTAHAEIMALRQGGLVMQNYRLFDATLYVTFEPCVMCAGAMINSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDFYRMPRQVFNSQKKAQSSIN	34
TadA-CDb	MSEVEFSHEYWMRHALTLAKRARDEGAGPVGAVLVLNNRVIGEGWNRAIGLHDPTAHAEIMALRQGGLVMQNYRLFDATLYVTFEPCVMCAGA	35

[0317]

	MINSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADEC AALLCDFYRMPRRVFNSQKKAQSSIN	
TadA-CDc	MSEVEFSHEYWMRHALTLAKRARDEGAGPVGAVLVLNNRVIGEGWN RAIGLHDPTAHAEIMALRQGGLVMQNYRLFDATLYVTFEPCVMCAGA MINSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADEC AALLCDFYRIPRQVFNSQKKAQSSIN	36
TadA-CDd	MSEVEFSHEYWMRHALTLAKRARDERKAPVGAVLVLNNRVIGEGWN RAIGLHDPTAHAEIIALRQGGLVMQNYRLIDATLYVTFEPCVMCAGAMI NSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAA LLCDFYRMPRQVFNAQKKAQSSIN	37
TadA-CDe	MSEVEFSHEYWMRHALTLAKRARDERAGPVGAVLVLNNRVIGEGWN RAIGLHDPTAHAEIMALRQGGLVMQNHRLIDATLYVTFEPCVMCAGA MINSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADEC AALLCDFYRMPRHVFNSQKKAQSSIN	38
TadA-CDf (例如，TadA-Dual)	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWN RRIGLHDPTAHAEIMALRQGGLVMQNSRLIDATLYVTFEPCVMCAGA MINSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADEC AALLCDFYRMPRHVFNAQKKAQSSIN	39
TadA-CDg	MSEVEFSHEYWMRHALTLAKRARDEGEGPVGAVLVLNNRVIGEGWN RRIGLHDPTAHAEIMALRQGGLVMQNCRLIDATLYVTFEPCVMCAGA MIHSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADEC AALLCDFYRMPRQVFNAQKKAQSSIN	40
TadA-8e (底物)	MSEVEFSHEYWMRHALTLAKRARDEREVPVGAVLVLNNRVIGEGWNR AIGLHDPTAHAEIMALRQGGLVMQNYRLIDATLYVTFEPCVMCAGAMI HSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAA LLCDFYRMPRQVFNAQKKAQSSIN	41

[0318] 在一些实施方案中,TadDE双重编辑器的双重编辑器脱氨酶(例如,TadA-CDf或TadA-Dual,SEQ ID NO:39)可以进一步进化,例如使用下文或本文别处进一步描述的PACE和/或PANCE测定。在一些实施方案中,TadA-Dual脱氨酶(例如,TadA-CDf,SEQ ID NO:39)进一步进化以增强对胞嘧啶碱基的特异性并减少对腺苷碱基的特异性。例如,图51E显示了列出经进化的TadA-Dual脱氨酶(例如,TadDE-1至TadDE-5)及其相对于未突变TadA-Dual脱氨酶及其亲本TadA-8e脱氨酶的突变的表格。

[0319] 在一些实施方案中,如图51C所示,使用PACE对TadA-Dual脱氨酶进行突变。在一些实施方案中,噬菌体辅助连续进化或PACE(图51C,左)与选择回路(图51C,右)结合使用。在一些实施方案中,编码部分脱氨酶(SP)的选择噬菌体感染大肠杆菌宿主细胞的连续流。本领域技术人员将理解,大肠杆菌宿主细胞还必须含有定义选择回路的质粒以及诱变质粒。

在选择回路中,噬菌体繁殖与gIII(P2)的表达相关联,gIII只能使用活性T7 RNA聚合酶转录。在一些实施方案中,T7 RNA聚合酶(P3)与C端降解决定子融合,并且脱氨酶必须进行C至U编辑以在降解决定子之前安装终止密码子,从而产生活性T7 RNA聚合酶。在噬菌体感染的情况下,使用分裂内含肽系统(P1)完成完整的脱氨酶,并且可以在脱氨酶上发生突变。有益的突变导致噬菌体在培养池中繁殖和富集,而适应性较差的噬菌体无法繁殖并随后被恒定流出物洗出。

[0320] 在一些实施方案中,如图51D所示,使用噬菌体辅助非连续进化(PANCE)对TadA-Dual脱氨酶进行突变。在一些实施方案中,对TadA-Dual脱氨酶(SEQ ID NO:39)进行PANCE直到噬菌体滴度增加,尽管稀释倍数和启动子强度的严格性更高,表明发生了有益的突变。在一些实施方案中,有益的突变包含脱氨酶中位置N46处的突变。

[0321] 在一些实施方案中,在位置N46处对NNK文库进行PANCE以进一步鉴定有益的突变。在一些实施方案中,可以作为诱变测定的组合。例如,在一些实施方案中,对来自PANCE研究的结果变体进行超过100小时的PACE。

[0322] 在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46I、A48R、Y73P和H96N突变(TadA-CD-1,图51E,PANCE,SEQ ID NO:42)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46T、A48R、Y73P和H96N突变(TadA-CD-2,图51E,PANCE,SEQ ID NO:43)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46T、A48R、Y73S和H96N突变(TadA-CD-3,图51E,PANCE,SEQ ID NO:44)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46V、A48R、Y73S和H96N突变(TadA-CD-4,在N46处对NNK文库进行PANCE,SEQ ID NO:45)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46V、A48R、Y73P和H96N突变(TadA-CD-5,在N46处对NNK文库进行PANCE,SEQ ID NO:46)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46L、A48R、Y73P和H96N突变(TadA-CD-6,在N46处对NNK文库进行PANCE,SEQ ID NO:47)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含V28A、N46L、A48P和Y73P突变(TadA-CD-7,在N46处对NNK文库进行PANCE,SEQ ID NO:48)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含V28A、N46C、A48P和Y73P突变(TadA-CD-8,在N46处对NNK文库进行PANCE,SEQ ID NO:49)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46V、A48R、Y73P和H96N突变(TadA-CD-9,图51E,PACE,SEQ ID NO:50)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46V、A48R、Q71H、Y73P和H96N突变(TadA-CD-10,图51E,PACE,SEQ ID NO:51)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46L、A48R、Y73P和H96N突变(TadA-CD-11,图51E,PACE,SEQ ID NO:52)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46C、A48R、Y73P和H96N突变(TadA-CD-12,图51E,PACE,SEQ ID NO:53)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46C、A48R、Y73P、H96N和A162V突变(TadA-CD-13,图51E,PACE,SEQ ID NO:54)。

[0323] 在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46I、A48R、Y73S和H96N突变(TadA-CD-14,图51E,PANCE,SEQ ID NO:359)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、A48R、Q71S、Y73S和H96N突变(TadA-CD-15,图51E,PANCE,SEQ ID NO:360)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46L、A48R和Y73P突变(TadA-CD-16,图51E,PANCE,SEQ ID NO:361)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46L、A48R、Y73P和H96N突变(TadA-CD-17,图51E,PANCE,SEQ ID NO:362)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、Y73P和H96N突变(TadA-CD-18,图51E,PANCE,SEQ ID NO:363)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46V、A48R、Y73S和H96N突变(TadA-CD-19,图51E,PANCE,SEQ ID NO:364)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46V、A48R、Y73P和H96N突变(TadA-CD-20,图51E,PANCE,SEQ ID NO:365)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G和N46L突变(TadA-CD-21,图51E,PANCE,SEQ ID NO:366)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46I、A48R、Y73P和H96N突变(TadA-CD-22,图51E,PANCE,SEQ ID NO:367)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46V、A48R、Y73P和H96N突变(TadA-CD-23,图51E,PANCE,SEQ ID NO:368)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、A48P、Y73H、T79P和H96N突变(TadA-CD-24,图51E,PANCE,SEQ ID NO:369)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、N46I和H96N突变(TadA-CD-25,图51E,PANCE,SEQ ID NO:370)。

[0324] 在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46V、A48R、Y73P和H96N突变(TadA-CD-26,图51E,PANCE,SEQ ID NO:371)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46L、A48R、Y73S和H96N突变(TadA-CD-27,图51E,PANCE,SEQ ID NO:372)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46C、A48R、H96N和A162V突变(TadA-CD-28,图51E,PANCE,SEQ ID NO:373)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46V、A48R、Q71H、Y73P和H96N突变(TadA-CD-29,图51E,PANCE,SEQ ID NO:374)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46C、A48R、Y73P和H96N突变(TadA-CD-30,图51E,PANCE,SEQ ID NO:375)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46C、A48R、Y73P、H96N和A162V突变(TadA-CD-31,图51E,PANCE,SEQ ID NO:376)。

[0325] 在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46V、A48R、Y73P和H96N突变(TadA-CD-32,图51E,PANCE,SEQ ID NO:377)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46V、A48R、Y73S和H96N突变(TadA-CD-33,图51E,PANCE,SEQ ID NO:378)。在

一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46V、A48P、Y73S和H96N突变(TadA-CD-34,图51E,PANCE,SEQ ID NO:379)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46C、A48R、Y73P和H96N突变(TadA-CD-35,图51E,PANCE,SEQ ID NO:380)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、L34M、N46L、A48R、Y73P和H96N突变(TadA-CD-36,图51E,PANCE,SEQ ID NO:381)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46L、A48R、Y73P和H96N突变(TadA-CD-37,图51E,PANCE,SEQ ID NO:382)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含R26G、V28A、N46L、A48P、R64K、Y73P和H96N突变(TadA-CD-38,图51E,PANCE,SEQ ID NO:383)。

[0326] 在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46I、S73P和H154Q突变(TadA-CD-1,图51E,PANCE,SEQ ID NO:42)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46T突变(TadA-CD-2,图51E,PANCE,SEQ ID NO:43)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46T和H154Q突变(TadA-CD-3,图51E,PANCE,SEQ ID NO:44)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46V和H154Q突变(TadA-CD-4,在N46处对NNK文库进行PANCE,SEQ ID NO:45)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46V、S73P、G105S和H154Q突变(TadA-CD-5,在N46处对NNK文库进行PANCE,SEQ ID NO:46)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46L、S73P和H154Q突变(TadA-CD-6,在N46处对NNK文库进行PANCE,SEQ ID NO:47)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含G26R、N46L、R48P、S73P、N96H和H154Q突变(TadA-CD-7,在N46处对NNK文库进行PANCE,SEQ ID NO:48)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46C、N96H和H154Q突变(TadA-CD-8,在N46处对NNK文库进行PANCE,SEQ ID NO:49)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46V、S73P和H154Q突变(TadA-CD-9,图51E,PACE,SEQ ID NO:50)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46V、Q71H、S73P和H154Q突变(TadA-CD-10,图51E,PACE,SEQ ID NO:51)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46L和H154Q突变(TadA-CD-11,图51E,PACE,SEQ ID NO:52)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46C、S73P和H154Q突变(TadA-CD-12,图51E,PACE,SEQ ID NO:53)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46C、S73P、H154Q和A162V突变(TadA-CD-13,图51E,PACE,SEQ ID NO:54)。

[0327] 在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46I和H154Q突变(TadA-CD-14,图51E,PACE,SEQ ID NO:359)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:41的氨基酸序列包含Q71S和H154Q突变(TadA-CD-15,图51E,PANCE,SEQ ID NO:360)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46L、S73P、N79T和N96H突变(TadA-CD-16,图

51E,PANCE,SEQ ID NO:361)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46L、S73P和N79T突变(TadA-CD-17,图51E,PANCE,SEQ ID NO:362)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含R48A、S73P和N79T突变(TadA-CD-18,图51E,PANCE,SEQ ID NO:363)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46V和N79T突变(TadA-CD-19,图51E,PANCE,SEQ ID NO:364)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46V、S73P和N79T突变(TadA-CD-20,图51E,PANCE,SEQ ID NO:365)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含A28V、N46L、R48A、S73Y、N79T和N96H突变(TadA-CD-21,图51E,PANCE,SEQ ID NO:366)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46I、S73P和N79T突变(TadA-CD-22,图51E,PANCE,SEQ ID NO:367)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46V、S73P、N79T和G106S突变(TadA-CD-23,图51E,PANCE,SEQ ID NO:368)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含R48P、S73H和N79P突变(TadA-CD-24,图51E,PANCE,SEQ ID NO:369)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含A28V、N46I、R48A、S73Y和N79T突变(TadA-CD-25,图51E,PANCE,SEQ ID NO:370)。

[0328] 在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46V和S73P突变(TadA-CD-26,图51E,PANCE,SEQ ID NO:371)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46L突变(TadA-CD-27,图51E,PANCE,SEQ ID NO:372)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46C、S73Y和A162V突变(TadA-CD-28,图51E,PANCE,SEQ ID NO:373)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46V、Q71H和S73P突变(TadA-CD-29,图51E,PANCE,SEQ ID NO:374)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46C和S73P突变(TadA-CD-30,图51E,PANCE,SEQ ID NO:375)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46C、S73P和A162V突变(TadA-CD-31,图51E,PANCE,SEQ ID NO:376)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46V和S73P突变(TadA-CD-32,图51E,PANCE,SEQ ID NO:377)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46V突变(TadA-CD-33,图51E,PANCE,SEQ ID NO:378)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46V和R48P突变(TadA-CD-34,图51E,PANCE,SEQ ID NO:379)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46CV和S73P突变(TadA-CD-35,图51E,PANCE,SEQ ID NO:380)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含L34M、N46L和S73P突变(TadA-CD-36,图51E,PANCE,SEQ ID NO:381)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46L和S73P突变(TadA-CD-37,图51E,PANCE,SEQ ID NO:382)。在一些实施方案中,经进化的TadA-Dual脱氨酶相对于SEQ ID NO:39的氨基酸序列包含N46L、r48P、R64K和S73P突变(TadA-CD-38,图51E,PANCE,SEQ ID NO:383)。

[0329] 在一些实施方案中,从TadA-Dual脱氨酶进化的TadA-CD脱氨酶具有改进的对胞嘧啶碱基特异性。在一些实施方案中,经进化的TadA-CD脱氨酶表现出与本文所述的其他进化的脱氨酶相似的胞嘧啶中靶活性。在一些实施方案中,从TadA-Dual脱氨酶进化的脱氨酶对胞嘧啶碱基具有增加的特异性并且对腺苷碱基具有降低的特异性。在一些实施方案中,从TadA-Dual脱氨酶进化的脱氨酶表现出没有残留的A至G碱基编辑(例如,TadA-CD-1至TadA-CD-38)。

[0330] 在一些实施方案中,TadA-CD-1当被并入到BE4max架构中时,表现出没有残留的A至G碱基编辑。在一些实施方案中,TadA-CD-2当被并入到BE4max架构中时,表现出没有残留的A至G碱基编辑。在一些实施方案中,TadA-CD-3当被并入到BE4max架构中时,表现出没有残留的A至G碱基编辑。在一些实施方案中,TadA-CD-4当被并入到BE4max架构中时,表现出没有残留的A至G碱基编辑。在一些实施方案中,TadA-CD-5当被并入到BE4max架构中时,表现出没有残留的A至G碱基编辑。上述描述并非旨在以任何方式进行限制,并且本文所述的进化TadA-CD脱氨酶可以与本领域技术人员已知的任何合适架构一起使用。

[0331] 在一些实施方案中,从TadA-dual进化的TadA-CD包含与表11中列出的任何氨基酸序列至少80%、85%、90%、95%、98%、99%或99.5%相同。

[0332] 在一些实施方案中,表11中列出的任何一个脱氨酶还可以包含V106W突变。在一些实施方案中,TadA-CD变体与表10中列出的任何氨基酸序列的至少80%、至少85%、至少90%、至少95%、至少98%、至少99%或至少99.5%,其中表11中列出的任何一个序列进一步包含V106W突变。

[0333] 表11.从TadA-Dual(SEQ ID NO:39)衍生的示例性突变TadA-CD相关物的列表。TadA-8e和TadA-Dual的序列作为参考提供。

[0334]

TadA-8e (底物)	MSEVEFSHEYWMRHALTLAKRARDEREVPVGAVLVLNNRVIGEGWNRA IGLHDPTAHAEIMALRQGGLVMQNYRLIDATLYVTFEPCVMCAGAMIHS RIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCD FYRMPRQVFNAQKKAQSSIN	41
TadA-dual	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWNRR IGLHDPTAHAEIMALRQGGLVMQNSRLIDATLYVTFEPCVMCAGAMINS RIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCD FYRMPRHVFNAQKKAQSSIN	39
TadA-CD-1	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWIRRI GLHDPTAHAEIMALRQGGLVMQNPRLIDATLYVTFEPCVMCAGAMINSR IGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDF YRMPRQVFNAQKKAQSSIN	42
TadA-CD-2	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWTRRI GLHDPTAHAEIMALRQGGLVMQNSRLIDATLYVTFEPCVMCAGAMINSR IGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDF YRMPRHVFNAQKKAQSSIN	43
TadA-CD-3	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWTRRI GLHDPTAHAEIMALRQGGLVMQNSRLIDATLYVTFEPCVMCAGAMINSR IGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDF YRMPRQVFNAQKKAQSSIN	44
TadA-CD-4	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWVRRI GLHDPTAHAEIMALRQGGLVMQNSRLIDATLYVTFEPCVMCAGAMINSR IGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDF YRMPRQVFNAQKKAQSSIN	45
TadA-CD-5	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWVRRI GLHDPTAHAEIMALRQGGLVMQNPRLIDATLYVTFEPCVMCAGAMINSR IGRVVFSVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDF YRMPRQVFNAQKKAQSSIN	46

[0335]

TadA-CD-6	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWLRRI GLHDPTAHAEIMALRQGGLVMQNPRLIDATLYVTFEPCVMCAGAMINSR IGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDF YRMPRQVFNAQKKAQSSIN	47
TadA-CD-7	MSEVEFSHEYWMRHALTLAKRARDEREAPVGAVLVLNNRVIGEGWLRPI GLHDPTAHAEIMALRQGGLVMQNPRLIDATLYVTFEPCVMCAGAMIHSR IGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDF YRMPRQVFNAQKKAQSSIN	48
TadA-CD-8	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWCRRI GLHDPTAHAEIMALRQGGLVMQNSRLIDATLYVTFEPCVMCAGAMIHSR IGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDF YRMPRQVFNAQKKAQSSIN	49
TadA-CD-9	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWVRRI GLHDPTAHAEIMALRQGGLVMQNPRLIDATLYVTFEPCVMCAGAMINSR IGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDF YRMPRQVFNAQKKAQSSIN	50
TadA-CD-10	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWVRRI GLHDPTAHAEIMALRQGGLVMHNPRLIDATLYVTFEPCVMCAGAMINSR IGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDF YRMPRQVFNAQKKAQSSIN	51
TadA-CD-11	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWLRRI GLHDPTAHAEIMALRQGGLVMQNSRLIDATLYVTFEPCVMCAGAMINSR IGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDF YRMPRQVFNAQKKAQSSIN	52
TadA-CD-12	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWCRRI GLHDPTAHAEIMALRQGGLVMQNPRLIDATLYVTFEPCVMCAGAMINSR IGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDF YRMPRQVFNAQKKAQSSIN	53
TadA-CD-13	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWCRRI GLHDPTAHAEIMALRQGGLVMQNPRLIDATLYVTFEPCVMCAGAMINSR IGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDF YRMPRQVFNAQKKVQSSIN	54
TadA-CD-14	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWIRRI GLHDPTAHAEIMALRQGGLVMQNSRLIDATLYVTFEPCVMCAGAMINSR IGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDF YRMPRQVFNAQKKAQSSIN	359
TadA-CD-15	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWNRR IGLHDPTAHAEIMALRQGGLVMSNSRLIDATLYVTFEPCVMCAGAMINSR IGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDF YRMPRQVFNAQKKAQSSIN	360
TadA-CD-16	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWLRRI GLHDPTAHAEIMALRQGGLVMQNPRLIDATLYVTFEPCVMCAGAMIHSR IGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDF YRMPRQVFNAQKKAQSSIN	361
TadA-CD-17	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWLRRI GLHDPTAHAEIMALRQGGLVMQNPRLIDATLYVTFEPCVMCAGAMINSR IGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDF YRMPRQVFNAQKKAQSSIN	362

[0336]

TadA-CD-18	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWNRAIGLHDPTAHAEIMALRQGGLVMQNPRLIDATLYVTFEPCVMCAGAMINSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDFYRMPRQVFNAQKKAQSSIN	363
TadA-CD-19	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWVRRIGLHDPTAHAEIMALRQGGLVMQNSRLIDATLYVTFEPCVMCAGAMINSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDFYRMPRQVFNAQKKAQSSIN	364
TadA-CD-20	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWVRRIGLHDPTAHAEIMALRQGGLVMQNPRLIDATLYVTFEPCVMCAGAMINSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDFYRMPRQVFNAQKKAQSSIN	365
TadA-CD-21	MSEVEFSHEYWMRHALTLAKRARDEGEVPVGAVLVLNNRVIGEGWLRAIGLHDPTAHAEIMALRQGGLVMQNYRLIDATLYVTFEPCVMCAGAMIHSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDFYRMPRQVFNAQKKAQSSIN	366
TadA-CD-22	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWIRRIGLHDPTAHAEIMALRQGGLVMQNPRLIDATLYVTFEPCVMCAGAMINSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDFYRMPRQVFNAQKKAQSSIN	367
TadA-CD-23	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWVRRIGLHDPTAHAEIMALRQGGLVMQNPRLIDATLYVTFEPCVMCAGAMINSRIGRVVFSVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDFYRMPRQVFNAQKKAQSSIN	368
TadA-CD-24	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWNRPIGLHDPTAHAEIMALRQGGLVMQNHRLIDAPLYVTFEPCVMCAGAMINSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDFYRMPRQVFNAQKKAQSSIN	369
TadA-CD-25	MSEVEFSHEYWMRHALTLAKRARDEGEVPVGAVLVLNNRVIGEGWIRAIGLHDPTAHAEIMALRQGGLVMQNYRLIDATLYVTFEPCVMCAGAMINSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDFYRMPRQVFNAQKKAQSSIN	370
TadA-CD-26	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWVRRIGLHDPTAHAEIMALRQGGLVMQNPRLIDATLYVTFEPCVMCAGAMINSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDFYRMPRQVFNAQKKAQSSIN	371
TadA-CD-27	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWLRRIGLHDPTAHAEIMALRQGGLVMQNSRLIDATLYVTFEPCVMCAGAMINSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDFYRMPRQVFNAQKKAQSSIN	372
TadA-CD-28	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWCRRIGLHDPTAHAEIMALRQGGLVMQNYRLIDATLYVTFEPCVMCAGAMINSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDFYRMPRQVFNAQKKVQSSIN	373
TadA-CD-29	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWVRRIGLHDPTAHAEIMALRQGGLVMHNPRLIDATLYVTFEPCVMCAGAMINSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDFYRMPRQVFNAQKKAQSSIN	374

[0337]

TadA-CD-30	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWCRRI GLHDPTAHAEIMALRQGGLVMQNPRLIDATLYVTFEPCVMCAGAMINSR IGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDF YRMPRQVFNAQKKAQSSIN	375
TadA-CD-31	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWCRRI GLHDPTAHAEIMALRQGGLVMQNPRLIDATLYVTFEPCVMCAGAMINSR IGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDF YRMPRQVFNAQKKVQSSIN	376
TadA-CD-32	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWVRRI GLHDPTAHAEIMALRQGGLVMQNPRLIDATLYVTFEPCVMCAGAMINSR IGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDF YRMPRQVFNAQKKAQSSIN	377
TadA-CD-33	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWVRRI GLHDPTAHAEIMALRQGGLVMQNSRLIDATLYVTFEPCVMCAGAMINSR IGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDF YRMPRQVFNAQKKAQSSIN	378
TadA-CD-34	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWVRPI GLHDPTAHAEIMALRQGGLVMQNSRLIDATLYVTFEPCVMCAGAMINSR IGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDF YRMPRQVFNAQKKAQSSIN	379
TadA-CD-35	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWCRRI GLHDPTAHAEIMALRQGGLVMQNPRLIDATLYVTFEPCVMCAGAMINSR IGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDF YRMPRQVFNAQKKAQSSIN	380
TadA-CD-36	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVMVLNNRVIGEGWLR RIGLHDPTAHAEIMALRQGGLVMQNPRLIDATLYVTFEPCVMCAGAMIN SRIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLC DFYRMPRQVFNAQKKAQSSIN	381
TadA-CD-37	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWLRRI GLHDPTAHAEIMALRQGGLVMQNPRLIDATLYVTFEPCVMCAGAMINSR IGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDF YRMPRQVFNAQKKAQSSIN	382
TadA-CD-38	MSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVLVLNNRVIGEGWLRPI GLHDPTAHAEIMALKQGGLVMQNPRLIDATLYVTFEPCVMCAGAMINSR IGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECAALLCDF YRMPRQVFNAQKKAQSSIN	383

[0338] napDNAbp结构域

[0339] 本文所述的碱基编辑器包含核酸可编程DNA结合(napDNAbp)结构域。napDNAbp与至少一个向导核酸(例如,向导RNA)缔合,该向导核酸将napDNAbp定位到包含与向导核酸或其部分(例如,向导RNA的原间隔区)互补的DNA链(即靶链)的DNA序列。换句话说,向导核酸“编程”napDNAbp结构域以定位并结合到靶链的互补序列。napDNAbp结构域与互补序列的结合使碱基编辑器的核碱基修饰结构域(即腺苷脱氨酶结构域)能够接近并酶促脱氨靶链中的靶碱基。

[0340] napDNAbp可以是CRISPR(成簇规律间隔短回文重复)相关核酸酶。如上所概述,CRISPR是适应性免疫系统,提供对移动遗传元件(病毒、转座元件和接合质粒)的保护。CRISPR簇含有间隔区(与先前的移动元件互补的序列),并靶向入侵的核酸。CRISPR簇被转录并加工成CRISPR RNA(crRNA)。在II型CRISPR系统中,正确加工前crRNA需要反式编码的小RNA(tracrRNA)、内源性核糖核酸酶3(rnc)和Cas9蛋白。tracrRNA作为核糖核酸酶3辅助

加工前crRNA的向导。随后,Cas9/crRNA/tracrRNA内切核酸地切割与间隔区互补的线性或环状双链DNA靶标。与crRNA不互补的靶链首先被内切核酸地切割,然后被3'-5'外切核酸地修剪。在自然界中,DNA结合和切割通常需要蛋白质和两种RNA。然而,可以工程改造单一向导RNA("sgRNA"或简称为"gRNA"),以便将crRNA和tracrRNA两者的方面并入到单个RNA种类中。参见例如Jinek et al.,Science 337:816-821(2012),其全部内容通过引用并入本文。

[0341] 对可以与所公开的腺苷脱氨酶结合使用的各种napDNAbp的描述并非旨在以任何方式进行限制。碱基编辑器可以包含典型SpCas9,或任何直系同源Cas9蛋白,或任何变体Cas9蛋白(包括任何天然存在的变体、突变体或其他工程化版本的Cas9),这些Cas9是已知的或可以通过定向进化或其他诱变过程制造或进化的。在各种实施方案中,napDNAbp具有切口酶活性,即仅切割靶DNA序列的一条链。在其他实施方案中,napDNAbp具有失活的核酸酶,例如是"死亡的"蛋白。可以使用的其他变体Cas9蛋白是相对于典型SpCas9具有较小分子量(例如,用于更容易的递送)的那些或具有修饰或重排的一级氨基酸序列(例如,环状排列形式)的那些。本文所述的碱基编辑器还可以包含Cas9等效物,包括Cas12a/Cpf1蛋白。本文使用的napDNAbp(例如,SpCas9、SaCas9或SaCas9变体或SpCas9变体)还可以还可以含有改变/增强其PAM特异性的各种修饰。本公开考虑了与本文公开的任何Cas9蛋白具有至少70%、至少75%、至少80%、至少85%、至少90%、至少91%、至少92%、至少93%、至少94%、至少95%、至少96%、至少97%、至少98%、至少99%或至少99.9%的序列同一性的任何Cas9、Cas9变体或Cas9等效物。在一些实施方案中,napDNAbp结构域包含野生型Cas9的切口酶变体。在一些实施方案中,napDNAbp结构域包含本文公开的任何Cas9切口酶。

[0342] 在一些实施方案中,napDNAbp指导在靶序列位置处的一条链或两条链的切割,如在靶序列内和/或在靶序列的互补序列内。在一些实施方案中,napDNAbp指导从靶序列的第一个或最后一个核苷酸约1、2、3、4、5、6、7、8、9、10、15、20、25、50、100、200、500或更多个碱基对内的一条或两条链的切割。例如,来自酿脓链球菌的Cas9的RuvC I催化结构域中的天冬氨酸至丙氨酸置换(D10A)将Cas9从切割两条链的核酸酶转化为切口酶(切割单链)。其他使Cas9成为切口酶的突变示例包括但不限于相对于典型SpCas9序列的H840A、N854A和N863A,以及相对于Nme2Cas9序列的H588A和D16A,以及其他Cas9变体或Cas9等效物中的等效氨基酸位置。

[0343] 如本文所用,术语"Cas蛋白"指的是从自然界中获取的全长Cas蛋白、具有与天然存在的Cas蛋白不同的序列的重组Cas蛋白、或仍然保留所公开方法所需的所有或大部分必要基本功能的任何Cas蛋白片段,即(i)具有Cas蛋白对靶DNA的核酸可编程结合,以及(ii)在一条链上使靶DNA序列产生切口的能力。本文考虑的Cas蛋白涵盖CRISPR Cas9蛋白,以及Cas9等效物、变体(例如,Cas9切口酶(nCas9)或核酸酶失活的Cas9(dCas9))同源物、直系同源物或旁系同源物,无论是天然存在的还是非天然存在的(例如,工程化或重组的),并且可以包括来自任何类型CRISPR系统(例如,II型、V型、VI型)的Cas9等效物,包括Cpf1(V型CRISPR-Cas系统)、C2c1(V型CRISPR-Cas系统)、C2c2(VI型CRISPR-Cas系统)和C2c3(V型CRISPR-Cas系统)。更多Cas等效物描述于Makarova et al.,"C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector,"Science 2016;353(6299),其内容通过引用并入本文。

[0344] 术语"Cas9"或"Cas9结构域"包括来自任何生物体的任何天然存在的Cas9、任何天

然存在的Cas9等效物或功能片段、任何Cas9同源物、直系同源物或旁系同源物、以及天然存在或工程化的Cas9的任何突变体或变体。术语Cas9并非意指特别限制,并且可以称为“Cas9或等效物”。示例性Cas9蛋白在本文中进一步描述和/或在本领域中描述,并通过引用并入本文。本公开对本公开的碱基编辑器中使用的特定napDNAbp没有限制。

[0345] 如本文所用,术语“紧凑型Cas9蛋白”、“紧凑型napDNAbp”和“[Cas蛋白的]紧凑型变体”指的是具有少于约1250个氨基酸的氨基酸长度的Cas9蛋白或变体。在一些实施方案中,紧凑型Cas9蛋白或紧凑型napDNAbp含有少于1250个氨基酸、少于1240个氨基酸、少于1230个氨基酸、少于1220个氨基酸、少于1210个氨基酸、少于1200个氨基酸、少于1190个氨基酸、少于1180个氨基酸、少于1170个氨基酸、少于1160个氨基酸、少于1150个氨基酸、少于1140个氨基酸、少于1130个氨基酸、少于1120个氨基酸、少于1110个氨基酸、少于1100个氨基酸、少于1050个氨基酸、少于1000个氨基酸、少于950个氨基酸、少于900个氨基酸、少于850个氨基酸、少于800个氨基酸、少于750个氨基酸、少于700个氨基酸、少于650个氨基酸、少于600个氨基酸、少于550个氨基酸或少于500个氨基酸的长度。这些术语还包括由具有少于约3750个核苷酸的长度的核酸序列编码的任何Cas9蛋白或变体。本公开的碱基编辑器可以包含紧凑型napDNAbp和/或紧凑型Cas9蛋白。在一些实施方案中,紧凑型Cas9蛋白比SpCas9短约350个氨基酸。在一些实施方案中,紧凑型Cas9蛋白的长度为约1000个氨基酸。在一些实施方案中,紧凑型蛋白是酿脓链球菌Cas9(SpCas9)、Cpf1、CasX、CasY、C2c1、C2c2、C2c3、Cas12a、Cas12b、Cas12g、Cas12h、Cas12i、Cas13b、Cas13c、Cas13d、Cas14、Csn2、Cas3或CasΦ的紧凑型变体。“紧凑型变体”可以指相对于野生型Cas9蛋白(如野生型SpCas9或Cpf1)具有一个或多个截短或一个或多个缺失的Cas9蛋白。

[0346] 另外的Cas9序列和结构为本领域技术人员所熟知(参见例如“Complete genome sequence of an M1 strain of Streptococcus pyogenes.”Ferretti et al.,J.J.,McShan W.M.,Ajdic D.J.,Savic D.J.,Savic G.,Lyon K.,Primeaux C.,Sezate S.,Suvorov A.N.,Kenton S.,Lai H.S.,Lin S.P.,Qian Y.,Jia H.G.,Najar F.Z.,Ren Q.,Zhu H.,Song L.,White J.,Yuan X.,Clifton S.W.,Roe B.A.,McLaughlin R.E.,Proc.Natl.Acad.Sci.U.S.A.98:4658-4663(2001);“CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III.”Deltcheva E.,Chylinski K.,Sharma C.M.,Gonzales K.,Chao Y.,Pirzada Z.A.,Eckert M.R.,Vogel J.,Charpentier E.,Nature 471:602-607(2011);以及“A programmabledual-RNA-guided DNA endonuclease in adaptive bacterial immunity.”Jinek M.,Chylinski K.,Fonfara I.,Hauer M.,Doudna J.A.,Charpentier E.Science 337:816-821(2012),其中每篇的全部内容通过引用并入本文),并也在下文中提供。

[0347] 以下提供了Cas9和Cas9等效物的实例;然而,这些具体实例并非旨在限制。本公开的碱基编辑器可以使用任何合适的napDNAbp,包括任何合适的Cas9或Cas9等效物。

[0348] 在一些实施方案中,Cas9包含或衍生自野生型SaCas9(例如,金黄色葡萄球菌,1053AA,123kDa)。在一些实施方案中,野生型SaCas9包含以下氨基酸序列:

[0349] MGKRNYILGLDIGITSVGYGIIDYETRDVIDAGVRLFKEANVENNEGRRSKRGARRLKRRRRHRIQRVKKLLFDYNLLTDHSELSGINPYEARVKGLSQKLSEEEFSAALLHLAKRRGVHNVNEVEEDTGNELSTKEQISRNSKALEEKYVAELQLERLKKDGEVRGSINRFKTSDYVKEAKQLLKVQKAYHQLDQSFIDTYIDLLETRRTYYEGPGEGS

PFGWKDIKEWYEMLMGHCTYFPEELRSVKYAYNADLYNALNDLNNLVITRDENEKLEYYEKFQIIENVFKQKKKPT
LKQIAKEILVNEEDIKGYRVTSTGKPEFTNLKVYHDIKDITARKEIIENAELLDQIAKILTIYQSSEDIQEELTNL
NSELTQEEIEQISNLKGYTGTHNLSLKAINLILDELWHTNDNQIAIFNRLKLVPKKVDLSQQKEIPTTLVDDFILS
PVVKRSFIQSIKVINAIIKKYGLPNDIIIELAREKNSKDAQKMINEMQKRNRQTNERIEEIIRTTGKENAKYLIEK
IKLHDMQEGKCLYSLEAIPLEDLLNNPFNYEVDHIIPRSVSFDNSFNNKVLVKQEENSKKGNRTPFQYLSSSDSKI
SYETFKKHILNLAKGKGRISKTKKEYLLEERDINRFSVQKDFINRNLVDTRYATRGLMNLLRSYFRVNNLDVKVKS
INGGFTSFLRRKWKFKKERNKGYKHHAEDALIIANADFIFKEWKKLDKAKKVMENQMFEEKQAESMPEIETEQEYK
EIFITPHQIKHIKDFKDYKYSHRVDKKPNRKLINDTLYSTRKDDKGNTLIVNNLNGLYDKDNDKLKKLINKSPEKL
LMYHHDPQTYQKLKLIMEQYGDEKNPLYKYYEETGNYLTKYSKKDNGPVIKKIKYYGNKLNAHLDITDDYPNSRNK
VVKLSLKPYRFDVYLDNGVYKFVTVKNLDVIKKENYYEVNSKCYEEAKKLKKISNQAEFIASFYKNDLIKINGELY
RVIGVNNDLLNRIEVNMIDITYREYLENMNDKRPPHIIKTIASKTQSIKKYSTDILGNLYEVKSKKHPQIIKK(SEQ
ID NO:347)

[0350] 在一些实施方案中,SaCas9的序列包含与SEQ ID NO:347至少至少80%、至少85%、至少90%、至少95%、至少99%、至少99.5%或至少99.8%的序列同一性。

[0351] 在一些实施方案中,Cas9包含或衍生自野生型SpCas9(例如,SpCas9,酿脓链球菌M1,SwissProt登录号Q99ZW2,野生型)。在一些实施方案中,野生型SaCas9包含以下氨基酸序列:

[0352] MDKKYSIGLDIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEATRLKRTA
RRRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHERHPIFGNIVDEVAYHEKYPTIYHLRKKLVD
STDKADLRLIYLALAHMIKFRGHFLIEGDLNPDNSDVDKLFIQLVQTYNQLFEENPINASGVDAKAILSARLSKSR
RLENLIAQLPGEKKNGLFGNLIALSLGLTPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIGDQYADLFLAAKNL
SDAILLSDILRVNTEITKAPLSASMIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSKNGYAGYIDGGASQEEF
YKFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIHLGELHAILRRQEDFYPFLKDNREKIEKILTFRI
PYYVGPLARGNSRFAWMTRKSEETITPWNFEEVVDKGASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNEL
TKVKYVTEGMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFKKIECFDSVEISGVEDRFNASLGTYHDLLKI
IKDKDFLDNEENEDILEDIVLTLTLFEDREMIEERLKTYAHLFDDKVMKQLKRRRYTGWGRLSRKLINGIRDKQSG
KTILDFLKSDGFANRNFMQLIHDDSLTFKEDIQKAQVSGQGDSLHEHIANLAGSPAIKKGILQTVKVVDELVKVMG
RHKPENIVIEMARENQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQLQNEKLYLYYLQNGRDMYVDQEL
DINRLSDYDVDHIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMKNYWRQLLNAKLITQRKFDNLTKAE
RGGLSELDKAGFIKRQLVETRQITKHVAQILDSRMNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINN
YHHAHDAYLNAVVGTALIKKYPKLESEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLANGE
IRKRPLIETNGETGEIVWDKGRDFATVRKVLSMPQVNIVKKTEVQTGGFSKESILPKRNSDKLIARKKDWDPKKYG
GFDSPTVAYSVLVVAKVEKGKSKKLKSVKELLGITIMERSSFEKNPIDFLEAKGYKEVKKDLIIKLPKYSLFELEN
GRKRMLASAGELQKGNELALPSKYVNFLYLASHYEKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILAD
ANLDKVLSAYNKHRDKPIREQAENIIHLFTLTNLGAPAAFKYFDTTIDRKRYTSTKEVLDATLIHQSITGLYETRI
DLSQLGGD(SEQ ID NO:200)

[0353] 在一些实施方案中,SpCas9的序列包含与SEQ ID NO:200至少至少80%、至少85%、至少90%、至少95%、至少99%、至少99.5%或至少99.8%的序列同一性。

[0354] Cas切口酶

[0355] 在一些实施方案中,所公开的碱基编辑器可以包含含有Cas切口酶的napDNAbp结构域。在一些实施方案中,本文所述的碱基编辑器包含Cas9切口酶。在一些实施方案中,任何所公开的碱基编辑器或载体可以包含含有D10A突变的酿脓链球菌Cas9切口酶(SpCas9n或nCas9)。在一些实施方案中,任何所公开的碱基编辑器可以包含Nme2Cas9切口酶(Nme2Cas9n)或eNme2-C Cas9切口酶(eNme2-C Cas9n),其中每个含有D16A突变。

[0356] 术语“Cas9切口酶”或“nCas9”指的是能够在双链DNA分子靶标中引入单链断裂的Cas9变体。在一些实施方案中,Cas9切口酶仅包含单个功能性核酸酶结构域。野生型Cas9(例如,典型SpCas9)包含两个独立的核酸酶结构域,即RuvC结构域(其切割非原间隔区DNA链)和HNH结构域(其切割原间隔区DNA链)。在一个实施方案中,Cas9切口酶在RuvC结构域中包含使RuvC核酸酶活性失活的突变。例如,天冬氨酸(D)10、组氨酸(H)983、天冬氨酸(D)986或谷氨酸(E)762的突变已被报道为RuvC核酸酶结构域的功能丧失突变,并创建了功能性Cas9切口酶(例如,Nishimasu et al.,“Crystal structure of Cas9 in complex with guide RNA and target DNA,”Cell 156(5),935-949,其通过引用并入本文)。因此,RuvC结构域中的切口酶突变可以包括D10X、H983X、D986X或E762X,其中X是除野生型氨基酸以外的任何氨基酸。在某些实施方案中,切口酶可以是D10A、H983A、D986A或E762A,或其组合。

[0357] 在一些实施方案中,任何所公开的碱基编辑器的napDNAbp结构域包含酿脓链球菌Cas9切口酶(SpCas9n)。在一些实施方案中,任何所公开的碱基编辑器的napDNAbp结构域包含与SEQ ID NO:343至少80%、至少85%、至少90%、至少95%或至少99%的序列同一性。在一些实施方案中,任何所公开的碱基编辑器的napDNAbp结构域包含SEQ ID NO:343的氨基酸序列。

[0358] 在一些实施方案中,任何所公开的碱基编辑器的napDNAbp结构域包含金黄色葡萄球菌Cas9切口酶(SaCas9n)。在一些实施方案中,任何所公开的碱基编辑器的napDNAbp结构域包含与SEQ ID NO:351至少80%、至少85%、至少90%、至少95%或至少99%的序列同一性。在一些实施方案中,任何所公开的碱基编辑器的napDNAbp结构域包含SEQ ID NO:351的氨基酸序列。

[0359] 在一些实施方案中,任何所公开的碱基编辑器的napDNAbp结构域包含脑膜炎奈瑟菌Cas9切口酶(Nme2Ca9n)或其变体。在一些实施方案中,任何所公开的碱基编辑器的napDNAbp结构域包含与SEQ ID NO:352或353至少80%、至少85%、至少90%、至少95%或至少99%的序列同一性。在一些实施方案中,任何所公开的碱基编辑器的napDNAbp结构域包含SEQ ID NO:352或353的氨基酸序列。在一些实施方案中,napDNAbp结构域包含SEQ ID NO:353的氨基酸序列。eNme2-C Cas9(SEQ ID NO:353)变体显示出对靶向NNNNCN(N_4CN)PAM的偏好。含有该eNme2-C变体的碱基编辑器在人细胞中的N_4CC PAM上产生了约60%或更高的碱基编辑效率,其代表了相对于含有野生型Nme2Cas9的碱基编辑器的两倍改进。

[0360] 在一些实施方案中,任何所公开的碱基编辑器的napDNAbp结构域包含野生型Nme2Cas9核酸酶(SEQ ID NO:349)。

[0361] 在各种实施方案中,Cas切口酶可以在RuvC核酸酶结构域中具有突变,并具有以下氨基酸序列之一,或其具有与其具有至少80%、至少85%、至少90%、至少95%或至少99%的序列同一性的氨基酸序列的变体。

[0362]

描述	序列	SEQ ID NO:
Cas9 切口酶 酿脓链球菌 Q99ZW2 Cas9 具有 D10**A**	MDKKYSIGLAIGTNSVGWAVITDEYKVPSKKFKVLGNTD RHSIKKNLIGALLFDSGETAEATRLKRTARRRYTRRKNRIC YLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHERHPIFGN IVDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHM IKFRGHFLIEGDLNPDNSDVDKLFIQLVQTYNQLFEENPIN ASGVDAKAILSARLSKSRRLENLIAQLPGEKKNGLFGNLIA LSLGLTPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIG DQYADLFLAAKNLSDAILLSDILRVNTEITKAPLSASMIKR YDEHHQDLTLLKALVRQQLPEKYKEIFFDQSKNGYAGYI DGGASQEEFYKFIKPILEKMDGTEELLVKLNREDLLRKQR TFDNGSIPHQIHLGELHAILRRQEDFYPFLKDNREKIEKILT FRIPYYVGPLARGNSRFAWMTRKSEETITPWNFEEVVDKG ASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELT KVKYVTEGMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQL KEDYFKKIECFDSVEISGVEDRFNASLGTYHDLLKIIKDKD FLDNEENEDILEDIVLTLTLFEDREMIEERLKTYAHLFDDK VMKQLKRRRYTGWGRLSRKLINGIRDKQSGKTILDFLKSD GFANRNFMQLIHDDSLTFKEDIQKAQVSGQGDSLHEHIAN LAGSPAIKKGILQTVKVVDELVKVMGRHKPENIVIEMARE NQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQL QNEKLYLYYLQNGRDMYVDQELDINRLSDYDVDHIVPQS FLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMKNYW RQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVE TRQITKHVAQILDSRMNTKYDENDKLIREVKVITLKSKLV SDFRKDFQFYKVREINNYHHAHDAYLNAVVGTALIKKYP KLESEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNI MNFFKTEITLANGEIRKRPLIETNGETGEIVWDKGRDFATV RKVLSMPQVNIVKKTEVQTGGFSKESILPKRNSDKLIARK KDWDPKKYGGFDSPTVAYSVLVVAKVEKGKSKKLKSVK ELLGITIMERSSFEKNPIDFLEAKGYKEVKKDLIIKLPKYSL FELENGRKRMLASAGELQKGNELALPSKYVNFLYLASHY EKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILA DANLDKVLSAYNKHRDKPIREQAENIIHLFTLTNLGAPAA FKYFDTTIDRKRYTSTKEVLDATLIHQSITGLYETRIDLSQL GGD	SEQ ID NO: 343

[0363]

描述	序列	SEQ ID NO:
Cas9 切口酶 酿脓链球菌 Q99ZW2 Cas9 具有 E762A	MDKKYSIGLDIGTNSVGWAVITDEYKVPSKKFKVLGNTD RHSIKKNLIGALLFDSGETAEATRLKRTARRRYTRRKNRIC YLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHERHPIFGN IVDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHM IKFRGHFLIEGDLNPDNSDVDKLFIQLVQTYNQLFEENPIN ASGVDAKAILSARLSKSRRLENLIAQLPGEKKNGLFGNLIA LSLGLTPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIG DQYADLFLAAKNLSDAILLSDILRVNTEITKAPLSASMIKR YDEHHQDLTLLKALVRQQLPEKYKEIFFDQSKNGYAGYI DGGASQEEFYKFIKPILEKMDGTEELLVKLNREDLLRKQR TFDNGSIPHQIHLGELHAILRRQEDFYPFLKDNREKIEKILT FRIPYYVGPLARGNSRFAWMTRKSEETITPWNFEEVVDKG ASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELT KVKYVTEGMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQL KEDYFKKIECFDSVEISGVEDRFNASLGTYHDLLKIIKDKD FLDNEENEDILEDIVLTLTLFEDREMIEERLKTYAHLFDDK VMKQLKRRRYTGWGRLSRKLINGIRDKQSGKTILDFLKSD GFANRNFMQLIHDDSLTFKEDIQKAQVSGQGDSLHEHIAN LAGSPAIKKGILQTVKVVDELVKVMGRHKPENIVIAMARE NQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQL QNEKLYLYYLQNGRDMYVDQELDINRLSDYDVDHIVPQS FLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMKNYW RQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVE TRQITKHVAQILDSRMNTKYDENDKLIREVKVITLKSKLV SDFRKDFQFYKVREINNYHHAHDAYLNAVVGTALIKKYP KLESEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNI MNFFKTEITLANGEIRKRPLIETNGETGEIVWDKGRDFATV RKVLSMPQVNIVKKTEVQTGGFSKESILPKRNSDKLIARK KDWDPKKYGGFDSPTVAYSVLVVAKVEKGKSKKLKSVK ELLGITIMERSSFEKNPIDFLEAKGYKEVKKDLIIKLPKYSL FELENGRKRMLASAGELQKGNELALPSKYVNFLYLASHY EKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILA DANLDKVLSAYNKHRDKPIREQAENIIHLFTLTNLGAPAA FKYFDTTIDRKRYTSTKEVLDATLIHQSITGLYETRIDLSQL GGD	SEQ ID NO: 344
Cas9 切口酶 酿脓链球菌 Q99ZW2 Cas9 具有 H983A	MDKKYSIGLDIGTNSVGWAVITDEYKVPSKKFKVLGNTD RHSIKKNLIGALLFDSGETAEATRLKRTARRRYTRRKNRIC YLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHERHPIFGN IVDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHM IKFRGHFLIEGDLNPDNSDVDKLFIQLVQTYNQLFEENPIN ASGVDAKAILSARLSKSRRLENLIAQLPGEKKNGLFGNLIA LSLGLTPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIG DQYADLFLAAKNLSDAILLSDILRVNTEITKAPLSASMIKR YDEHHQDLTLLKALVRQQLPEKYKEIFFDQSKNGYAGYI DGGASQEEFYKFIKPILEKMDGTEELLVKLNREDLLRKQR TFDNGSIPHQIHLGELHAILRRQEDFYPFLKDNREKIEKILT FRIPYYVGPLARGNSRFAWMTRKSEETITPWNFEEVVDKG ASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELT KVKYVTEGMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQL KEDYFKKIECFDSVEISGVEDRFNASLGTYHDLLKIIKDKD FLDNEENEDILEDIVLTLTLFEDREMIEERLKTYAHLFDDK	SEQ ID NO: 345

	描述	序列	SEQ ID NO:
		VMKQLKRRRYTGWGRLSRKLINGIRDKQSGKTILDFLKSD GFANRNFMQLIHDDSLTFKEDIQKAQVSGQGDSLHEHIAN LAGSPAIKKGILQTVKVVDELVKVMGRHKPENIVIEMARE NQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQL QNEKLYLYYLQNGRDMYVDQELDINRLSDYDVDHIVPQS FLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMKNYW RQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVE TRQITKHVAQILDSRMNTKYDENDKLIREVKVITLKSKLV SDFRKDFQFYKVREINNYHAAHDAYLNAVVGTALIKKYP KLESEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNI MNFFKTEITLANGEIRKRPLIETNGETGEIVWDKGRDFATV RKVLSMPQVNIVKKTEVQTGGFSKESILPKRNSDKLIARK KDWDPKKYGGFDSPTVAYSVLVVAKVEKGKSKKLKSVK ELLGITIMERSSFEKNPIDFLEAKGYKEVKKDLIIKLPKYSL FELENGRKRMLASAGELQKGNELALPSKYVNFLYLASHY EKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILA DANLDKVLSAYNKHRDKPIREQAENIIHLFTLTNLGAPAA FKYFDTTIDRKRYTSTKEVLDATLIHQSITGLYETRIDLSQL GGD	
[0364]	Cas9 切口酶 酿脓链球菌 Q99ZW2 Cas9 具有 D986A	MDKKYSIGLDIGTNSVGWAVITDEYKVPSKKFKVLGNTD RHSIKKNLIGALLFDSGETAEATRLKRTARRRYTRRKNRIC YLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHERHPIFGN IVDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHM IKFRGHFLIEGDLNPDNSDVDKLFIQLVQTYNQLFEENPIN ASGVDAKAILSARLSKSRRLENLIAQLPGEKKNGLFGNLIA LSLGLTPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIG DQYADLFLAAKNLSDAILLSDILRVNTEITKAPLSASMIKR YDEHHQDLTLLKALVRQQLPEKYKEIFFDQSKNGYAGYI DGGASQEEFYKFIKPILEKMDGTEELLVKLNREDLLRKQR TFDNGSIPHQIHLGELHAILRRQEDFYPFLKDNREKIEKILT FRIPYYVGPLARGNSRFAWMTRKSEETITPWNFEEVVDKG ASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELT KVKYVTEGMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQL KEDYFKKIECFDSVEISGVEDRFNASLGTYHDLLKIIKDKD FLDNEENEDILEDIVLTLTLFEDREMIEERLKTYAHLFDDK VMKQLKRRRYTGWGRLSRKLINGIRDKQSGKTILDFLKSD GFANRNFMQLIHDDSLTFKEDIQKAQVSGQGDSLHEHIAN LAGSPAIKKGILQTVKVVDELVKVMGRHKPENIVIEMARE NQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQL QNEKLYLYYLQNGRDMYVDQELDINRLSDYDVDHIVPQS FLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMKNYW RQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVE TRQITKHVAQILDSRMNTKYDENDKLIREVKVITLKSKLV SDFRKDFQFYKVREINNYHHAHAAYLNAVVGTALIKKYP KLESEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNI MNFFKTEITLANGEIRKRPLIETNGETGEIVWDKGRDFATV RKVLSMPQVNIVKKTEVQTGGFSKESILPKRNSDKLIARK KDWDPKKYGGFDSPTVAYSVLVVAKVEKGKSKKLKSVK ELLGITIMERSSFEKNPIDFLEAKGYKEVKKDLIIKLPKYSL FELENGRKRMLASAGELQKGNELALPSKYVNFLYLASHY EKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILA	SEQ ID NO: 346

[0365]

描述	序列	SEQ ID NO:
	DANLDKVLSAYNKHRDKPIREQAENIIHLFTLTNLGAPAA FKYFDTTIDRKRYTSTKEVLDATLIHQSITGLYETRIDLSQL GGD	
Cas9 切口酶金黄色葡萄球菌(SaCas9)具有 D10**A**	MGKRNYILGLAIGITSVGYGIIDYETRDVIDAGVRLFKEAN VENNEGRRSKRGARRLKRRRRHRIQRVKKLLFDYNLLTD HSELSGINPYEARVKGLSQKLSEEEFSAALLHLAKRRGVH NVNEVEEDTGNELSTKEQISRNSKALEEKYVAELQLERLK KDGEVRGSINRFKTSDYVKEAKQLLKVQKAYHQLDQSFI DTYIDLLETRRTYYEGPGEGSPFGWKDIKEWYEMLMGHC TYFPEELRSVKYAYNADLYNALNDLNNLVITRDENEKLE YYEKFQIIENVFKQKKKPTLKQIAKEILVNEEDIKGYRVTS TGKPEFTNLKVYHDIKDITARKEIIENAELLDQIAKILTIYQ SSEDIQEELTNLNSELTQEEIEQISNLKGYTGTHNLSLKAIN LILDELWHTNDNQIAIFNRLKLVPKKVDLSQQKEIPTTLVD DFILSPVVKRSFIQSIKVINAIIKKYGLPNDIIIELAREKNSKD AQKMINEMQKRNRQTNERIEEIIRTTGKENAKYLIEKIKLH DMQEGKCLYSLEAIPLEDLLNNPFNYEVDHIIPRSVSFDNS FNNKVLVKQEENSKKGNRTPFQYLSSSDSKISYETFKKHIL NLAKGKGRISKTKKEYLLEERDINRFSVQKDFINRNLVDT RYATRGLMNLLRSYFRVNNLDVKVKSINGGFTSFLRRKW KFKKERNKGYKHHAEDALIIANADFIFKEWKKLDKAKKV MENQMFEEKQAESMPEIETEQEYKEIFITPHQIKHIKDFKD YKYSHRVDKKPNRKLINDTLYSTRKDDKGNTLIVNNLNG LYDKDNDKLKKLINKSPEKLLMYHHDPQTYQKLKLIMEQ YGDEKNPLYKYYEETGNYLTKYSKKDNGPVIKKIKYYGN KLNAHLDITDDYPNSRNKVVKLSLKPYRFDVYLDNGVYK FVTVKNLDVIKKENYYEVNSKCYEEAKKLKKISNQAEFIA SFYKNDLIKINGELYRVIGVNNDLLNRIEVNMIDITYREYL ENMNDKRPPHIIKTIASKTQSIKKYSTDILGNLYEVKSKKH PQIIKK	SEQ ID NO: 351
Cas9 切口酶脑膜炎奈瑟氏球菌(Nme2Cas9)具有 D16**A**	MAAFKPNPINYILGLaIGIASVGWAMVEIDEEENPIRLIDLG VRVFERAEVPKTGDSLAMARRLARSVRRLTRRRAHRLLR ARRLLKREGVLQAADFDENGLIKSLPNTPWQLRAAALDR KLTPLEWSAVLLHLIKHRGYLSQRKNEGETADKELGALL KGVANNAHALQTGDFRTPAELALNKFEKESGHIRNQRGD YSHTFSRKDLQAELILLFEKQKEFGNPHVSGGLKEGIETLL MTQRPALSGDAVQKMLGHCTFEPAEPKAAKNTYTAERFI WLTKLNNLRILEQGSERPLTDTERATLMDEPYRKSKLTYA QARKLLGLEDTAFFKGLRYGKDNAEASTLMEMKAYHAIS RALEKEGLKDKKSPLNLSSELQDEIGTAFSLFKTDEDITGR LKDRVQPEILEALLKHISFDKFVQISLKALRRIVPLMEQGK RYDEACAEIYGDHYGKKNTEEKIYLPPIPADEIRNPVVLRA LSQARKVINGVVRRYGSPARIHIETAREVGKSFKDRKEIEK RQEENRKDREKAAAKFREYFPNFVGEPKSKDILKLRLYEQ QHGKCLYSGKEINLVRLNEKGYVEIDHALPFSRTWDDSFN NKVLVLGSENQNKGNQTPYEYFNGKDNSREWQEFKARV ETSRFPRSKKQRILLQKFDEDGFKECNLNDTRYVNRFLCQ FVADHILLTGKGKRRVFASNGQITNLLRGFWGLRKVRAE NDRHHALDAVVVACSTVAMQQKITRFVRYKEMNAFDGK TIDKETGKVLHQKTHFPQPWEFFAQEVMIRVFGKPDGKPE	SEQ ID NO: 352

[0366]

描述	序列	SEQ ID NO:
	FEEADTPEKLRTLLAEKLSSRPEAVHEYVTPLFVSRAPNRK MSGAHKDTLRSAKRFVKHNEKISVKRVWLTEIKLADLEN MVNYKNGREIELYEALKARLEAYGGNAKQAFDPKDNPF YKKGGQLVKAVRVEKTQESGVLLNKKNAYTIADNGDMV RVDVFCKVDKKGKNQYFIVPIYAWQVAENILPDIDCKGY RIDDSYTFCFSLHKYDLIAFQKDEKSKVEFAYYINCDSSNG RFYLAWHDKGSKEQQFRISTQNLVLIQKYQVNELGKEIRP CRLKKRPPVR	

[0367] Cas9等效物

[0368] 在一些实施方案中,本文所述的碱基编辑器可以包括任何Cas9等效物。如本文所用,术语“Cas9等效物”是一个广义术语,涵盖与在本发明的碱基编辑器中的Cas9发挥相同功能的任何napDNAbp,尽管其氨基酸一级序列和/或其三维结构可能不同,并且/或者从进化角度来看可能无关。因此,尽管Cas9等效物包括进化上相关的本文描述或涵盖的任何Cas9直系同源物、同源物、突变体或变体,Cas9等效物也包括可能通过趋同进化过程进化出与Cas9具有相同或相似功能的蛋白质,但这些蛋白质在氨基酸序列和/或三维结构上不一定有任何相似性。本文所述的碱基编辑器涵盖任何能够提供与Cas9相同或相似功能的任何Cas9等效物,即使该Cas9等效物可能是基于通过趋同进化产生的蛋白质。例如,如果Cas9指的是CRISPR-Cas系统的II型酶,那么Cas9等效物可以指CRISPR-Cas系统的V型或VI型酶。

[0369] 例如,Cas12e(CasX)是据报道具有与Cas9相同功能但通过趋同进化进化出的Cas9等效物。因此,Liuet al.,“CasX enzymes comprises a distinct family of RNA-guided genome editors,”Nature,2019,Vol.566:218-223描述的Cas12e(CasX)蛋白被考虑与本文所述的碱基编辑器一起使用。此外,Cas12e(CasX)的任何变体或修饰也是能够想到的,并且在本公开的范围内。

[0370] Cas9是在多种物种中进化的细菌酶。然而,本文考虑的Cas9等效物也可以是从古菌中获得的,古菌构成与细菌不同的单细胞原核微生物的域和界。

[0371] 在一些实施方案中,Cas9等效物可以指Cas12e(CasX)或Cas12d(CasY),这些已在例如Burstein et al.,“New CRISPR-Cas systems from uncultivated microbes.”Cell Res.2017Feb 21.doi:10.1038/cr.2017.21中描述,其全部内容通过引用并入本文。使用基因组解析的宏基因组学,鉴定出了许多CRISPR-Cas系统,包括古菌生命领域中首次报道的Cas9。这种不同的Cas9蛋白在研究较少的纳米古菌中被发现,作为活性CRISPR-Cas系统的一部分。在细菌中,发现了两个以前未知的系统,CRISPR-Cas12e和CRISPR-Cas12d,它们是迄今为止发现的最紧凑的系统之一。在一些实施方案中,Cas9指的是Cas12e或Cas12e的变体。在一些实施方案中,Cas9指的是Cas12d或Cas12d的变体。应当理解,其他RNA引导的DNA结合蛋白也可以用作核酸可编程DNA结合蛋白(napDNAbp),并且在本公开的范围内。另见Liu et al.,“CasX enzymes comprises a distinct family of RNA-guided genome editors,”Nature,2019,Vol.566:218-223。考虑了所有这些Cas9等效物。

[0372] 在一些实施方案中,Cas9等效物包含与天然存在的Cas12e(CasX)或Cas12d(CasY)蛋白至少85%、至少90%、至少91%、至少92%、至少93%、至少94%、至少95%、至少96%、至少97%、至少98%、至少99%或至少99.5%相同的氨基酸序列。在一些实施方案中,napDNAbp是天然存在的Cas12e(CasX)或Cas12d(CasY)蛋白。在一些实施方案中,napDNAbp

包含与野生型Cas部分或本文提供的任何Cas部分至少85％、至少90％、至少91％、至少92％、至少93％、至少94％、至少95％、至少96％、至少97％、至少98％、至少99％或至少99.5％相同的氨基酸序列。

[0373] 在各种实施方案中,核酸可编程DNA结合蛋白包括但不限于Cas9(例如dCas9和nCas9)、C2C3Cas12e(CasX)、Cas12d(CasY)、Cas12a(Cpf1)、Cas12b1(C2c1)、Cas13a(C2c2)、Cas12c(C2c3)、Argonaute。与Cas9具有不同PAM特异性的核酸可编程DNA结合蛋白的一个实例是来自普雷沃氏菌和弗朗西斯菌的成簇规律间隔短回文重复序列1(即Cas12a(Cpf1))。与Cas9类似,Cas12a(Cpf1)也是2类CRISPR效应物,但它是V型酶亚组的成员,而不是II型亚组。已经证明,Cas12a(Cpf1)介导了具有与Cas9不同特征的强大DNA干扰。Cas12a(Cpf1)是单一RNA引导的缺乏tracrRNA的内切核酸酶,并且其利用富含T的原间隔区相邻基序(TTN、TTTN或YTN)。此外,Cpf1经由交错的DNA双链断裂切割DNA。在16种Cpf1家族蛋白中,来自氨基酸球菌属和毛螺菌科的两种酶在人细胞中显示具有高效的基因组编辑活性。Cpf1蛋白在本领域是已知的,并且之前已有描述,例如Yamano et al.,“Crystal structure of Cpf1 in complex with guide RNA and target DNA.”Cell(165)2016,p.949-962;其全部内容通过引用并入本文。

[0374] 在其他实施方案中,Cas蛋白可以包括任何CRISPR相关蛋白,包括但不限于Cas12a、Cas12b1、Cas1、Cas1B、Cas2、Cas3、Cas4、Cas5、Cas6、Cas7、Cas8、Cas9(也称为Csn1和Csx12)、Cas10、Csy1、Csy2、Csy3、Cse1、Cse2、Csc1、Csc2、Csa5、Csn2、Csm2、Csm3、Csm4、Csm5、Csm6、Cmr1、Cmr3、Cmr4、Cmr5、Cmr6、Csb1、Csb2、Csb3、Csx17、Csx14、Csx10、Csx16、CsaX、Csx3、Csx1、Csx15、Csf1、Csf2、Csf3、Csf4、其同源物或其修饰版本,并且优选包含切口酶突变(例如,对应于SEQ ID NO:200的野生型Cas9多肽的D10A突变的突变)。

[0375] 在各种其他实施方案中,napDNAbp可以是以下任何一种蛋白:Cas9、C2c3Cas12a(Cpf1)、Cas12e(CasX)、Cas12d(CasY)、Cas12b1(C2c1)、Cas13a(C2c2)、Cas12c(C2c3)、GeoCas9、CjCas9、Cas12a、Cas12b、Cas12g、Cas12h、Cas12i、Cas13b、Cas13c、Cas13d、Cas14、Csn2、xCas9、SpCas9-NG、环状排列的Cas9或Argonaute(Ago)结构域,或其变体。

[0376] 具有修饰的PAM特异性的Cas9变体

[0377] 本公开的碱基编辑器还可以包含具有修饰的PAM特异性的Cas9变体。例如,本文所述的碱基编辑器可以利用任何天然存在或工程化的SpCas9变体,这些变体具有扩展和/或放宽的PAM特异性,这些变体在文献中已有描述,包括在Nishimasu et al.,“Engineered CRISPR-Cas9 nuclease with expanded targeting space,”Science,2018,361:1259-1262;Chatterjee et al.,“Robust Genome Editing of Single-Base PAM Targets with Engineered ScCas9 Variants,”BioRxiv,April 26,2019中。本公开的一些方面提供了对靶序列表现出活性的Cas9蛋白,该靶序列在其3’末端不包含典型PAM(5’-NGG-3’,其中N为A、C、G或T)。在一些实施方案中,Cas9蛋白对在其3’末端包含5’-NGG-3’PAM序列的靶序列表现出活性。在一些实施方案中,Cas9蛋白对在其3’末端包含5’-NNG-3’PAM序列的靶序列表现出活性。在一些实施方案中,Cas9蛋白对在其3’端包含5’-NNA-3’PAM序列的靶序列表现出活性。在一些实施方案中,Cas9蛋白对在其3’末端包含5’-NNC-3’PAM序列的靶序列表现出活性。在一些实施方案中,Cas9蛋白对在其3’末端包含5’-NNT-3’PAM序列的靶序列表现出活性。在一些实施方案中,Cas9蛋白对在其3’末端包含5’-NGT-3’PAM序列的靶序列表现

出活性。在一些实施方案中,Cas9蛋白对在其3’末端包含5’-NGA-3’PAM序列的靶序列表现出活性。在一些实施方案中,Cas9蛋白对在其3’末端包含5’-NGC-3’PAM序列的靶序列表现出活性。在一些实施方案中,Cas9蛋白对在其3’末端包含5’-NAA-3’PAM序列的靶序列表现出活性。在一些实施方案中,Cas9蛋白对在其3’末端包含5’-NAC-3’PAM序列的靶序列表现出活性。在一些实施方案中,Cas9蛋白对在其3’末端包含5’-NAT-3’PAM序列的靶序列表现出活性。在其他实施方案中,Cas9蛋白对在其3’末端包含5’-NAG-3’PAM序列的靶序列表现出活性。

[0378] 上述关于可以与当前公开的碱基编辑器一起使用的各种napDNAbp的描述并不旨在以任何方式进行限制。碱基编辑器可以包含典型SpCas9、或任何直系同源Cas9蛋白、或任何Cas9变体(包括任何天然存在的变体、突变体或其他工程化版本的Cas9),这些变体是已知的或者可以通过定向进化或其他诱变过程制造或进化的。在各种实施方案中,Cas9或Cas9变体具有切口酶活性,即仅切割靶DNA序列的一条链。在其他实施方案中,Cas9或Cas9变体具有失活的核酸酶,即“死亡的”Cas9蛋白。其他可以使用的Cas9变体包括那些具有小于典型SpCas9的分子量(例如,用于更容易的递送)的变体或具有修饰或重排的一级氨基酸结构(例如,环状置换体形式)的变体。本文所述的碱基编辑器还可以包含Cas9等效物,包括Cas12a/Cpf1和Cas12b蛋白,这些蛋白是趋同进化的结果。本文使用的napDNAbp(例如SpCas9、Cas9变体或Cas9等效物)也可以含有各种修饰,这些修饰改变/增强了它们的PAM特异性。最后,本申请考虑了与参考Cas9序列(如参考SpCas9典型序列或参考Cas9等效物(例如Cas12a/Cpf1))具有至少70%、至少75%、至少80%、至少85%、至少90%、至少91%、至少92%、至少93%、至少94%、至少95%、至少96%、至少97%、至少98%、至少99%或至少99.9%序列同一性的任何Cas9、Cas9变体或Cas9等效物。

[0379] 在一些实施方案中,SpCas9(H840A)包含与SEQ ID NO:480中的氨基酸序列至少80%、至少85%、至少90%、至少95%、至少99%或至少99.5%相同的序列。

[0380] SpCas9(H840A)

[0381] DKKYSIGLDIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEATRLKRTARRRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHERHPIFGNIVDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHMIKFRGHFLIEGDLNPDNSDVDKLFIQLVQTYNQLFEENPINASGVDAKAILSARLSKSRRLENLIAQLPGEKKNGLFGNLIALSLGLTPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIGDQYADLFLAAKNLSDAILLSDILRVNTEITKAPLSASMIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSKNGYAGYIDGGASQEEFYKFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIHLGELHAILRRQEDFYPFLKDNREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSEETITPWNFEEVVDKGASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVTEGMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFKKIECFDSVEISGVEDRFNASLGTYHDLLKIIKDKDFLDNEENEDILEDIVLTLTLFEDREMIEERLKTYAHLFDDKVMKQLKRRRYTGWGRLSRKLINGIRDKQSGKTILDFLKSDGFANRNFMQLIHDDSLTFKEDIQKAQVSGQGDSLHEHIANLAGSPAIKKGILQTVKVVDELVKVMGRHKPENIVIEMARENQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQLQNEKLYLYYLQNGRDMYVDQELDINRLSDYDVDAIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMKNYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKHVAQILDSRMNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNYHHAHDAYLNAVVGTALIKKYPKLESEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLANGEIRKRPLIETNGETGEIVWDKGRDFATVRKVLSMPQVNIVKKTEVQTGGFSKESILPKRNSDKLIARKKDWDPKKYGG

FDSPTVAYSVLVVAKVEKGKSKKLKSVKELLGITIMERSSFEKNPIDFLEAKGYKEVKKDLIIKLPKYSLFELENG
RKRMLASAGELQKGNELALPSKYVNFLYLASHYEKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADA
NLDKVLSAYNKHRDKPIREQAENIIHLFTLTNLGAPAAFKYFDTTIDRKRYTSTKEVLDATLIHQSITGLYETRID
LSQLGGD(SEQ ID NO:386)

[0382] 在特定实施方案中,具有扩展的PAM能力的Cas9变体是SpCas9(H840A)VRQR,具有以下氨基酸序列(其中相对于SEQ ID NO:480的SpCas9(H840A)的V、R、Q、R置换以粗体下划线显示。此外,对于SpCas9(H840A)VRQR(“SpCas9-VRQR”),SpCas9(H840)中的甲硫氨酸残基被移除。该SpCas9变体具有改变的PAM特异性,其识别5’-NGA-3’的PAM,而不是典型5’-NGG-3’的PAM:

[0383] SpCas9-VRQR

[0384] DKKYSIGLDIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEATRLKRTAR
RRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHERHPIFGNIVDEVAYHEKYPTIYHLRKKLVDS
TDKADLRLIYLALAHMIKFRGHFLIEGDLNPDNSDVDKLFIQLVQTYNQLFEENPINASGVDAKAILSARLSKSRR
LENLIAQLPGEKKNGLFGNLIALSLGLTPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIGDQYADLFLAAKNLS
DAILLSDILRVNTEITKAPLSASMIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSKNGYAGYIDGGASQEEFY
KFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIHLGELHAILRRQEDFYPFLKDNREKIEKILTFRIP
YYVGPLARGNSRFAWMTRKSEETITPWNFEEVVDKGASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELT
KVKYVTEGMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFKKIECFDSVEISGVEDRFNASLGTYHDLLKII
KDKDFLDNEENEDILEDIVLTLTLFEDREMIEERLKTYAHLFDDKVMKQLKRRRYTGWGRLSRKLINGIRDKQSGK
TILDFLKSDGFANRNFMQLIHDDSLTFKEDIQKAQVSGQGDSLHEHIANLAGSPAIKKGILQTVKVVDELVKVMGR
HKPENIVIEMARENQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQLQNEKLYLYYLQNGRDMYVDQELD
INRLSDYDVDAIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMKNYWRQLLNAKLITQRKFDNLTKAER
GGLSELDKAGFIKRQLVETRQITKHVAQILDSRMNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNY
HHAHDAYLNAVVGTALIKKYPKLESEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLANGEI
RKRPLIETNGETGEIVWDKGRDFATVRKVLSMPQVNIVKKTEVQTGGFSKESILPKRNSDKLIARKKDWDPKKYGG
FVSPTVAYSVLVVAKVEKGKSKKLKSVKELLGITIMERSSFEKNPIDFLEAKGYKEVKKDLIIKLPKYSLFELENG
RKRMLASARELQKGNELALPSKYVNFLYLASHYEKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADA
NLDKVLSAYNKHRDKPIREQAENIIHLFTLTNLGAPAAFKYFDTTIDRKQYRSTKEVLDATLIHQSITGLYETRID
LSQLGGD(SEQ ID NO:74)。

[0385] 在另一个特定实施方案中,具有扩展的PAM能力的Cas9变体是SpCas9(H840A)VQR,具有以下氨基酸序列(其中相对于SEQ ID NO:480的SpCas9(H840A)的V、Q、R置换以粗体下划线显示。此外,对于SpCas9(H840A)VRQR(“SpCas9-VQR”),SpCas9(H840)中的甲硫氨酸残基被移除。该SpCas9变体具有改变的PAM特异性,其识别5’-NGA-3’的PAM,而不是典型5’-NGG-3’的PAM:

[0386] SpCas9-VQR

[0387] DKKYSIGLDIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEATRLKRTAR
RRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHERHPIFGNIVDEVAYHEKYPTIYHLRKKLVDS
TDKADLRLIYLALAHMIKFRGHFLIEGDLNPDNSDVDKLFIQLVQTYNQLFEENPINASGVDAKAILSARLSKSRR
LENLIAQLPGEKKNGLFGNLIALSLGLTPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIGDQYADLFLAAKNLS

DAILLSDILRVNTEITKAPLSASMIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSKNGYAGYIDGGASQEEFY
KFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIHLGELHAILRRQEDFYPFLKDNREKIEKILTFRIP
YYVGPLARGNSRFAWMTRKSEETITPWNFEEVVDKGASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELT
KVKYVTEGMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFKKIECFDSVEISGVEDRFNASLGTYHDLLKII
KDKDFLDNEENEDILEDIVLTLTLFEDREMIEERLKTYAHLFDDKVMKQLKRRRYTGWGRLSRKLINGIRDKQSGK
TILDFLKSDGFANRNFMQLIHDDSLTFKEDIQKAQVSGQGDSLHEHIANLAGSPAIKKGILQTVKVVDELVKVMGR
HKPENIVIEMARENQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQLQNEKLYLYYLQNGRDMYVDQELD
INRLSDYDVDAIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMKNYWRQLLNAKLITQRKFDNLTKAER
GGLSELDKAGFIKRQLVETRQITKHVAQILDSRMNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNY
HHAHDAYLNAVVGTALIKKYPKLESEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLANGEI
RKRPLIETNGETGEIVWDKGRDFATVRKVLSMPQVNIVKKTEVQTGGFSKESILPKRNSDKLIARKKDWDPKKYGG
FVSPTVAYSVLVVAKVEKGKSKKLKSVKELLGITIMERSSFEKNPIDFLEAKGYKEVKKDLIIKLPKYSLFELENG
RKRMLASAGELQKGNELALPSKYVNFLYLASHYEKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADA
NLDKVLSAYNKHRDKPIREQAENIIHLFTLTNLGAPAAFKYFDTTIDRKQYRSTKEVLDATLIHQSITGLYETRID
LSQLGGD(SEQ ID NO:75)。

[0388] 在另一个特定实施方案中,具有扩展的PAM能力的Cas9变体是SpCas9(H840A)VRER,具有以下氨基酸序列(其中相对于SEQ ID NO:480的SpCas9(H840A)V、R、E、R置换以粗体下划线显示。此外,对于SpCas9(H840A)VRER(“SpCas9-VRER”),SpCas9(H840)中的甲硫氨酸残基被移除。该SpCas9变体具有改变的PAM特异性,其识别5’-NGCG-3’的PAM,而不是典型5’-NGG-3’的PAM:

[0389] SpCas9-VRER

[0390] DKKYSIGLDIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEATRLKRTAR
RRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHERHPIFGNIVDEVAYHEKYPTIYHLRKKLVDS
TDKADLRLIYLALAHMIKFRGHFLIEGDLNPDNSDVDKLFIQLVQTYNQLFEENPINASGVDAKAILSARLSKSRR
LENLIAQLPGEKKNGLFGNLIALSLGLTPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIGDQYADLFLAAKNLS
DAILLSDILRVNTEITKAPLSASMIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSKNGYAGYIDGGASQEEFY
KFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIHLGELHAILRRQEDFYPFLKDNREKIEKILTFRIP
YYVGPLARGNSRFAWMTRKSEETITPWNFEEVVDKGASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELT
KVKYVTEGMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFKKIECFDSVEISGVEDRFNASLGTYHDLLKII
KDKDFLDNEENEDILEDIVLTLTLFEDREMIEERLKTYAHLFDDKVMKQLKRRRYTGWGRLSRKLINGIRDKQSGK
TILDFLKSDGFANRNFMQLIHDDSLTFKEDIQKAQVSGQGDSLHEHIANLAGSPAIKKGILQTVKVVDELVKVMGR
HKPENIVIEMARENQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQLQNEKLYLYYLQNGRDMYVDQELD
INRLSDYDVDAIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMKNYWRQLLNAKLITQRKFDNLTKAER
GGLSELDKAGFIKRQLVETRQITKHVAQILDSRMNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNY
HHAHDAYLNAVVGTALIKKYPKLESEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLANGEI
RKRPLIETNGETGEIVWDKGRDFATVRKVLSMPQVNIVKKTEVQTGGFSKESILPKRNSDKLIARKKDWDPKKYGG
FVSPTVAYSVLVVAKVEKGKSKKLKSVKELLGITIMERSSFEKNPIDFLEAKGYKEVKKDLIIKLPKYSLFELENG
RKRMLASARELQKGNELALPSKYVNFLYLASHYEKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADA
NLDKVLSAYNKHRDKPIREQAENIIHLFTLTNLGAPAAFKYFDTTIDRKEYRSTKEVLDATLIHQSITGLYETRID

LSQLGGD(SEQ ID NO:76)。

[0391] 在另一个实施方案中,具有扩展的PAM能力的Cas9变体是SpCas9-NG,如Nishimasu et al.,“Engineered CRISPR-Cas9 nuclease with expanded targeting space,”Science,2018,361:1259-1262所报道,其通过引用并入本文。相对于典型SpCas9序列(SEQ ID NO:200),SpCas9-NG(VRVRFRR)具有以下氨基酸序列置换:R1335V、L1111R、D1135V、G1218R、E1219F、A1322R和T1337R。该SpCas9具有放宽的PAM特异性,即对NGH(其中H=A、T或C)的PAM具有活性。参见Nishimasu et al.,“Engineered CRISPR-Cas9 nuclease with expanded targeting space,”Science,2018,361:1259-1262,其通过引用并入本文。

[0392] SpCas9-NG

[0393] MDKKYSIGLDIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEATRLKRTARRRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHERHPIFGNIVDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHMIKFRGHFLIEGDLNPDNSDVDKLFIQLVQTYNQLFEENPINASGVDAKAILSARLSKSRRLENLIAQLPGEKKNGLFGNLIALSLGLTPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIGDQYADLFLAAKNLSDAILLSDILRVNTEITKAPLSASMIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSKNGYAGYIDGGASQEEFYKFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIHLGELHAILRRQEDFYPFLKDNREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSEETITPWNFEEVVDKGASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVTEGMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFKKIECFDSVEISGVEDRFNASLGTYHDLLKIIKDKDFLDNEENEDILEDIVLTLTLFEDREMIEERLKTYAHLFDDKVMKQLKRRRYTGWGRLSRKLINGIRDKQSGKTILDFLKSDGFANRNFMQLIHDDSLTFKEDIQKAQVSGQGDSLHEHIANLAGSPAIKKGILQTVKVVDELVKVMGRHKPENIVIEMARENQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQLQNEKLYLYYLQNGRDMYVDQELDINRLSDYDVDHIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMKNYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKHVAQILDSRMNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNYHHAHDAYLNAVVGTALIKKYPKLESEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLANGEIRKRPLIETNGETGEIVWDKGRDFATVRKVLSMPQVNIVKKTEVQTGGFSKESIRPKRNSDKLIARKKDWDPKKYGGFVSPTVAYSVLVVAKVEKGKSKKLKSVKELLGITIMERSSFEKNPIDFLEAKGYKEVKKDLIIKLPKYSLFELENGRKRMLASARFLQKGNELALPSKYVNFLYLASHYEKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADANLDKVLSAYNKHRDKPIREQAENIIHLFTLTNLGAPRAFKYFDTTIDRKVYRSTKEVLDATLIHQSITGLYETRIDLSQLGGD(SEQ ID NO:77)

[0394] 此外,可以使用任何可用的方法来获得或构建变体或突变体Cas9蛋白。如本文所用,术语“突变”指的是序列(例如核酸序列或氨基酸序列)内一个残基被另一个残基置换,或者序列内一个或多个残基的插入或删除。突变在本文中通常通过鉴定原始残基,接着是残基在序列中的位置和新置换的残基的身份来描述。本文提供的用于进行氨基酸置换(突变)的各种方法在本领域是众所周知的,并且由例如Green and Sambrook,Molecular Cloning:A Laboratory Manual(4th ed.,Cold Spring Harbor Laboratory Press,Cold Spring Harbor,N.Y.(2012))提供。突变可以包括多种类别,如单碱基多态性、微重复区、插入缺失和倒位,并不旨在以任何方式进行限制。突变可以包括“功能丧失”突变,这是减少或废除蛋白质活性的突变的正常结果。大多数功能丧失突变是隐性的,因为在杂合子中,第二条染色体拷贝携带了编码完全功能蛋白的未突变基因,其存在补偿了突变的影响。突变还包括“功能获得”突变,这是赋予蛋白质或细胞在正常情况下不具备的异常活性的突变。许

多功能获得突变位于调控序列而不是编码区,并且因此可以产生多种后果。例如,突变可以导致一个或多个基因在错误的组织中表达,这些组织获得了它们通常缺乏的功能。由于其性质,功能获得突变通常是显性的。

[0395] 可以使用定点诱变将突变引入参考Cas9蛋白中。本领域已知的较旧的定点诱变方法依赖于将待突变的序列亚克隆到载体中,如M13噬菌体载体,该载体允许分离单链DNA模板。在这些方法中,将诱变引物(即能够退火到待突变位点但在待突变位点带有一个或多个错配核苷酸的引物)退火到单链模板上,然后从诱变引物的3'末端开始聚合模板的互补链。然后将所得的双链体转化到宿主细菌中,并针对期望的突变筛选噬斑。最近,定点诱变采用了PCR方法,其具有不需要单链模板的优点。此外,还开发了不需要亚克隆的方法。在进行基于PCR的定点诱变时,必须考虑几个问题。首先,在这些方法中,期望减少PCR循环次数,以防止聚合酶引入的非期望突变的扩增。其次,必须采用选择以减少反应中持续存在的未突变亲本分子的数量。第三,优选使用延长长度的PCR方法,以便允许使用单一PCR引物组。以及第四,由于一些热稳定聚合酶的非模板依赖性末端延伸活性,通常需要在程序中并入末端抛光步骤,然后再进行PCR生成的突变体产物的平端连接。

[0396] 突变也可以通过定向进化过程引入,例如噬菌体辅助连续进化(PACE)或噬菌体辅助非连续进化(PANCE)。如本文所用,术语"噬菌体辅助连续进化(PACE)"指的是使用噬菌体作为病毒载体的连续进化。PACE技术的一般概念已描述于例如:国际PCT申请PCT/US2009/056194,2009年9月8日提交,2010年3月11日公布为WO 2010/028347;国际PCT申请PCT/US2011/066747,2011年12月22日提交,2012年6月28日公布为WO 2012/088381;美国申请,美国专利号9,023,594,2015年5月5日发布,国际PCT申请PCT/US2015/012022,2015年1月20日提交,2015年9月11日公布为WO 2015/134121,以及国际PCT申请PCT/US2016/027795,2016年4月15日提交,2016年10月20日公布为WO 2016/168631,其中每个的全部内容通过引用并入本文。变体Cas9也可以通过噬菌体辅助非连续进化(PANCE)获得,如本文使用的术语"噬菌体辅助非连续进化(PANCE)"指的是使用噬菌体作为病毒载体的非连续进化。PANCE是一种简化的技术,用于快速体内定向进化,使用含有待进化的感兴趣基因的正在进化的"选择噬菌体"(SP)在新鲜的大肠杆菌宿主细胞中连续瓶转移,从而允许宿主大肠杆菌中的基因保持不变,而SP中含有的基因不断进化。长期以来,连续瓶转移一直是实验室微生物进化的广泛可及方法,并且最近,类似的方法也被开发用于噬菌体进化。PANCE系统的特征在于严格性低于PACE系统。

[0397] 具有修饰的PAM特异性的紧凑型Cas9变体

[0398] 在一些实施方案中,napDNAbp包含紧凑型Cas蛋白,如衍生自空肠弯曲杆菌(C.jejuni)、耳葡萄球菌(S.auricularis)、脑膜炎奈瑟菌(N.meningitidis)或金黄色葡萄球菌(S.aureus)的Cas9。在示例性实施方案中,napDNAbp包含CjCas9切口酶、SauriCas9切口酶、Nme2Cas9切口酶、SaCas9切口酶或SaKKH-Cas9切口酶。在一些实施方案中,napDNAbp不是Nme2Cas9蛋白或切口酶。在一些实施方案中,napDNAbp不是SaCas9蛋白或切口酶。

[0399] 在一些实施方案中,本公开的碱基编辑器包含napDNAbp结构域,该结构域包含衍生自脑膜炎奈瑟菌(Nme或Nme2)的Cas9直系同源物。在一些实施方案中,napDNAbp结构域包含Nme2Cas9。在其他实施方案中,napDNAbp结构域是Nme2Cas9结构域。在一些实施方案中,本公开的碱基编辑器包含Nme2Cas9切口酶。Nme2Cas9识别简单的二核苷酸PAM,NNNNCC或

N_4CC(其中N是任何核苷酸),如Edraki et al.,Molecular Cell 73,714-726中所述,其通过引用并入本文。在其他实施方案中,napDNAbp结构域包含Nme2Cas9变体。Nme2Cas9的变体可以识别PAM的更广泛阵列。在一些实施方案中,本公开的Nme2Cas9变体识别单核苷酸嘧啶PAM。在一些实施方案中,Nme2Cas9变体识别序列为NYN的PAM,其中Y是任何嘧啶(即C、T或U)。在其他实施方案中,Nme2Cas9变体识别序列为NNNNCN或N_4CN的PAM。在一些实施方案中,Nme2Cas9变体是eNme2Cas9切口酶(SEQ ID NO:439)。在一些实施方案中,Nme2Cas9变体是eNme2-C Cas9切口酶(SEQ ID NO:353)。

[0400] 野生型Nme2Cas9的序列如SEQ ID NO:349所示。在一些实施方案中,本公开的碱基编辑器包含napDNAbp结构域,其具有与SEQ ID NO:349至少90%、至少95%、至少98%或至少99%相同的序列。在一些实施方案中,本公开的碱基编辑器包含napDNAbp,其包含SEQ ID NO:5。该蛋白在本文中可称为工程化Nme2Cas9或eNme2Cas9。在各种实施方案中,任何公开的TadCBE都包含Nme2Cas9或Nme2Cas9的变体。

[0401] 野生型Nme2Cas9

[0402] MAAFKPNPINYILGLDIGIASVGWAMVEIDEEENPIRLIDLGVRVFERAEVPKTGDSLAMARRLARSV
RRLTRRRAHRLLRARRLLKREGVLQAADFDENGLIKSLPNTPWQLRAAALDRKLTPLEWSAVLLHLIKHRGYLSQR
KNEGETADKELGALLKGVANNAHALQTGDFRTPAELALNKFEKESGHIRNQRGDYSHTFSRKDLQAELILLFEKQK
EFGNPHVSGGLKEGIETLLMTQRPALSGDAVQKMLGHCTFEPAEPKAAKNTYTAERFIWLTKLNNLRILEQGSERP
LTDTERATLMDEPYRKSKLTYAQARKLLGLEDTAFFKGLRYGKDNAEASTLMEMKAYHAISRALEKEGLKDKKSPL
NLSSELQDEIGTAFSLFKTDEDITGRLKDRVQPEILEALLKHISFDKFVQISLKALRRIVPLMEQGKRYDEACAEI
YGDHYGKKNTEEKIYLPPIPADEIRNPVVLRALSQARKVINGVVRRYGSPARIHIETAREVGKSFKDRKEIEKRQE
ENRKDREKAAAKFREYFPNFVGEPKSKDILKLRLYEQQHGKCLYSGKEINLVRLNEKGYVEIDHALPFSRTWDDSF
NNKVLVLGSENQNKGNQTPYEYFNGKDNSREWQEFKARVETSRFPRSKKQRILLQKFDEDGFKECNLNDTRYVNRF
LCQFVADHILLTGKGKRRVFASNGQITNLLRGFWGLRKVRAENDRHHALDAVVVACSTVAMQQKITRFVRYKEMNA
FDGKTIDKETGKVLHQKTHFPQPWEFFAQEVMIRVFGKPDGKPEFEEADTPEKLRTLLAEKLSSRPEAVHEYVTPL
FVSRAPNRKMSGAHKDTLRSAKRFVKHNEKISVKRVWLTEIKLADLENMVNYKNGREIELYEALKARLEAYGGNAK
QAFDPKDNPFYKKGGQLVKAVRVEKTQESGVLLNKKNAYTIADNGDMVRVDVFCKVDKKGKNQYFIVPIYAWQVAE
NILPDIDCKGYRIDDSYTFCFSLHKYDLIAFQKDEKSKVEFAYYINCDSSNGRFYLAWHDKGSKEQQFRISTQNLV
LIQKYQVNELGKEIRPCRLKKRPPVR(SEQ ID NO:349)

[0403] 本文所述的Nme2Cas9变体开头处的“e”表示“经进化的”Nme2变体。相对于野生型Nme2Cas9的氨基酸置换以粗体下划线表示。

[0404] eNme2-C Cas9

[0405] MAAFK**S**NPINYILGLDIGIASVGWAMVEIDEE**G**NPIRLIDLGVRVFERAEVPKTGD
SLAMARRLARSVRRLTRRRAHRLLRARRLLKREGVLQAADFDENGLI**T**SLPNTPWQL
RAAALDRKLTPLEWSAVLLHLIKHRGYLSQRKNEGETA**A**KELGALLKGVANNAHAL
QTGDFRTPAELALNKFEKESGHIRNQRGDYSHTFSRKDLQAELILLFEKQKEFGNPHVS
GGLKEGIETLLMTQRPALSGDAVQKMLGHCT**L**EP**T**EPKAAKNTYTAERFIWLTKLNN
LRILEQGSERPLTDTER**S**TLMDEPYRKSKLTYAQARKLLGLEDTAFFKGLRYGKDNAE

ASTLMEMKAYHAISRALEKEGLKDKKSPLNLSSELQDEIGTAFSLFKTDEDITGRLKDR
VQPEILEALLKHISFDKFVQISLKALRRIVPLMEQGKRYDEACAEIYGVHYGKKNTEEK
IYLPPIPADEIRNPVVLRALSQARKVINGVVRRYGSPARIHIETAREVGKSFKDRKEIAK
RQEENRKDREKAAAKFREYFPNFVGEPKSKDILKLRLYEQQHGKCLYSGKEINLVRLN
EKGYVEIDHALPFSRTWDDSFNNKVLVLGSENQNKGNQTPYEYFNGKDNSREWQEFK
ARVETSRFPSSKKQRILLQKFDEDGFKECNLNDTRYVNRFLCQFVADHILLTGKGKRR

[0406] VVASNGQITNLLRGFWRLRKVRAENDRHHALDAVVVACSTVAMQQKITRFVRYKEM
NAFDGKTVDKETGKVLYQKTHFPQPWEFFAQEVMIRVFGKPDGKPEFEEADTPEKLR
TLLAEKLSSRPEAVHEYVTPLFVSRAPNRKMSGAHKDTLRSAKRFVKHNEKISVKRV
WLTEIKLADLENMVNYKNGREIELYEALKARLEAYGGNAKQAFDPKDNPFYKKGGQ
LVKAVRVEKTQKSGVLLNKKNAYTIADNGDMVRVDVFCKVDKKGKNQYFIVPIYAW
QVAENILPDIDCKGYRIDDSYTFCFSLHKYDLIAFQKDEKSKVEFAYYINCDSSSGGFY
LAWHDKGSREQRFRISTQNLALIQKYQVNELGKEIRPCRLKKRPPVR (SEQ ID NO: 353)

[0407] 在一些实施方案中,本公开的碱基编辑器包含napDNAbp,其包含衍生自空肠弯曲杆菌(CjCas9)的紧凑型Cas9直系同源物。在一些实施方案中,napDNAbp包含CjCas9。在一些实施方案中,本公开的碱基编辑器包含CjCas9切口酶。CjCas9识别NNNNACA和NNNNACAC PAM。参见Kimet al.,Nature Communications 8(14500):1-12(2017),其通过引用并入本文。CjCas9(切口酶)的序列如SEQ ID NO:348所示。在一些实施方案中,本公开的碱基编辑器包含napDNAbp结构域,其具有与SEQ ID NO:348至少90%、至少95%、至少98%或至少99%相同的序列。在一些实施方案中,所公开的碱基编辑器包含napDNAbp,其包含SEQ ID NO:348。该蛋白的长度为984个氨基酸。该蛋白在本文中可以称为工程化CjCas9或enCjCas9。合理工程化的CjCas9变体(enCjCas9)在Nakagawa,et al.,Communications Biology,(2022)5:211中描述,其通过引用并入本文。在各种实施方案中,任何所公开的TadCBE都包含CjCas9或enCjCas9(SEQ ID NO:348)的变体。

[0408] MARILAFAIGISSIGWAFSENDELKDCGVRIFTKVENPKTGESLALPRRLARSARKRLARRKARLNHL
KHLIANEFKLNYEDYQSFDESLAKAYKGSLISPYELRFRALNELLSKQDFARVILHIAKRRGYDDIKNSDDKEKGA
ILKAIKQNEEKLANYQSVGEYLYKEYFQKFKENSKEFTNVRNKKESYERCIAQSFLKDELKLIFKKQREFGFSFSK
KFEEEVLSVAFYKRALKDFSHLVGNCSFFTDEKRAPKNSPLAFMFVALTRIINLLNNLKNTEGILYTKDDLNALLN
EVLKNGTLTYKQTKKLLGLSDDYEFKGEKGTYFIEFKKYKEFIKALGEHNLSQDDLNEIAKDITLIKDEIKLKKAL
AKYDLNQNQIDSLSKLEFKDHLNISFKALKLVTPLMLEGKKYDEACNELNLKVAINEDKKDFLPAFNETYYKDEVT
NPVVLRAIKEYRKVLNALLKKYGKVHKINIELAREVGKNHSQRAKIEKEQNENYKAKKDAELECEKLGLKINSKNI
LKLRLFKEQKEFCAYSGEKIKISDLQDEKMLEIDHIYPYSRSFDDSYMNKVLVFTKQNQEKLNQTPFEAFGNDSAK
WQKIEVLAKNLPTKKQKRILDKNYKDKEQKNFKDRNLNDTRYIARLVLNYTKDYLDFLPLSDDENTKLNDTQKGSK
VHVEAKSGMLTSALRHTWGFSAKDRNNHLHHAIDAVIIAYANNSIVKAFSDFKKEQESNSAELYAKKISELDYKNK
RKFFEPFSGFRQKVLDKIDEIFVSKPERKKPSGALHEETFRKEEEFYQSYGGKEGVLKALELGKIRKVNGKIVKNG
DMFRVDIFKHKKTNKFYAVPIYTMDFALKVLPNKAVARSKKGEIKDWILMDENYEFCFSLYKDSLILIQTKDMQEP
EFVYYNAFTSSTVSLIVSKHDNKFETLSKNQKILFKNANEKEVIAKSIGIQNLKVFEKYIVSALGEVTKAEFRQRE

DFKK(SEQ ID NO:348)

[0409] 本公开的碱基编辑器还可以包含具有修饰的PAM特异性的Cas9变体。本公开的一些方面提供了对靶序列表现出活性的Cas9蛋白,该靶序列在其3'末端不包含典型PAM(5'-NGG-3',其中N为A、C、G或T)。在一些实施方案中,Cas9蛋白对在其3'末端包含5'-NGG-3'PAM序列的靶序列表现出活性。在一些实施方案中,Cas9蛋白对在其3'末端包含5'-NNG-3'PAM序列的靶序列表现出活性。在一些实施方案中,Cas9蛋白对在其3'末端包含5'-NNA-3'PAM序列的靶序列表现出活性。在一些实施方案中,Cas9蛋白对在其3'末端包含5'-NNC-3'PAM序列的靶序列表现出活性。在一些实施方案中,Cas9蛋白对在其3'末端包含5'-NNT-3'PAM序列的靶序列表现出活性。在一些实施方案中,Cas9蛋白对在其3'末端包含5'-NGT-3'PAM序列的靶序列表现出活性。在一些实施方案中,Cas9蛋白对在其3'末端包含5'-NGA-3'PAM序列的靶序列表现出活性。在一些实施方案中,Cas9蛋白对在其3'末端包含5'-NGC-3'PAM序列的靶序列表现出活性。在一些实施方案中,Cas9蛋白对在其3'末端包含5'-NAA-3'PAM序列的靶序列表现出活性。在一些实施方案中,Cas9蛋白对在其3'末端包含5'-NAC-3'PAM序列的靶序列表现出活性。在一些实施方案中,Cas9蛋白对在其3'末端包含5'-NAT-3'PAM序列的靶序列表现出活性。在其他实施方案中,Cas9蛋白对在其3'末端包含5'-NAG-3'PAM序列的靶序列表现出活性。

[0410] 在一些实施方案中,本公开的碱基编辑器包含napDNAbp结构域,该结构域包含SpCas9-NG,其具有对应于NGN的PAM。在一些实施方案中,所公开的碱基编辑器包含napDNAbp结构域,其具有与SpCas9-NG至少90%、至少95%、至少98%或至少99%相同的序列。SpCas9-NG的序列如下所示:

[0411] MDKKYSIGLAIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEATRLKRTARRRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHERHPIFGNIVDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHMIKFRGHFLIEGDLNPDNSDVDKLFIQLVQTYNQLFEENPINASGVDAKAILSARLSKSRRLENLIAQLPGEKKNGLFGNLIALSLGLTPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIGDQYADLFLAAKNLSDAILLSDILRVNTEITKAPLSASMIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSKNGYAGYIDGGASQEEFYKFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIHLGELHAILRRQEDFYPFLKDNREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSEETITPWNFEEVVDKGASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVTEGMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFKKIECFDSVEISGVEDRFNASLGTYHDLLKIIKDKDFLDNEENEDILEDIVLTLTLFEDREMIEERLKTYAHLFDDKVMKQLKRRRYTGWGRLSRKLINGIRDKQSGKTILDFLKSDGFANRNFMQLIHDDSLTFKEDIQKAQVSGQGDSLHEHIANLAGSPAIKKGILQTVKVVDELVKVMGRHKPENIVIEMARENQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQLQNEKLYLYYLQNGRDMYVDQELDINRLSDYDVDHIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMKNYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKHVAQILDSRMNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNYHHAHDAYLNAVVGTALIKKYPKLESEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLANGEIRKRPLIETNGETGEIVWDKGRDFATVRKVLSMPQVNIVKKTEVQTGGFSKESIRPKRNSDKLIARKKDWDPKKYGGFVSPTVAYSVLVVAKVEKGKSKKLKSVKELLGITIMERSSFEKNPIDFLEAKGYKEVKKDLIIKLPKYSLFELENGRKRMLASARFLQKGNELALPSKYVNFLYLASHYEKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADANLDKVLSAYNKHRDKPIREQAENIIHLFTLTNLGAPRAFKYFDTTIDRKVYRSTKEVLDATLIHQSITGLYETRIDLSQLGGD(SEQ ID NO:356)

[0412] 在一些实施方案中,本公开的碱基编辑器包含napDNAbp结构域,该结构域包含金黄色葡萄球菌Cas9切口酶KKH,或SaCas9-KKH或SaKKH-Cas9,其具有对应于NNNRRT或NNGRRT的PAM。该Cas9变体相对于野生型SaCas9(如SEQ ID NO:347所示)含有氨基酸置换D10A、E782K、N968K和R1015H(“KKH”)。在一些实施方案中,所公开的碱基编辑器包含napDNAbp结构域,其具有与SaCas9-KKH至少90%、至少95%、至少98%或至少99%相同的序列。SaCas9(和SaKKH-Cas9)的长度为1053个氨基酸。SaCas9-KKH(切口酶)的序列如下所示:

[0413] 金黄色葡萄球菌Cas9切口酶KKH(SaCas9-KKH)

[0414] MGKRNYILGLAIGITSVGYGIIDYETRDVIDAGVRLFKEANVENNEGRRSKRGARRLKRRRRHRIQRV
KKLLFDYNLLTDHSELSGINPYEARVKGLSQKLSEEEFSAALLHLAKRRGVHNVNEVEEDTGNELSTKEQISRNSK
ALEEKYVAELQLERLKKDGEVRGSINRFKTSDYVKEAKQLLKVQKAYHQLDQSFIDTYIDLLETRRTYYEGPGEGS
PFGWKDIKEWYEMLMGHCTYFPEELRSVKYAYNADLYNALNDLNNLVITRDENEKLEYYEKFQIIENVFKQKKKPT
LKQIAKEILVNEEDIKGYRVTSTGKPEFTNLKVYHDIKDITARKEIIENAELLDQIAKILTIYQSSEDIQEELTNL
NSELTQEEIEQISNLKGYTGTHNLSLKAINLILDELWHTNDNQIAIFNRLKLVPKKVDLSQQKEIPTTLVDDFILS
PVVKRSFIQSIKVINAIIKKYGLPNDIIIELAREKNSKDAQKMINEMQKRNRQTNERIEEIIRTTGKENAKYLIEK
IKLHDMQEGKCLYSLEAIPLEDLLNNPFNYEVDHIIPRSVSFDNSFNNKVLVKQEENSKKGNRTPFQYLSSSDSKI
SYETFKKHILNLAKGKGRISKTKKEYLLEERDINRFSVQKDFINRNLVDTRYATRGLMNLLRSYFRVNNLDVKVKS
INGGFTSFLRRKWKFKKERNKGYKHHAEDALIIANADFIFKEWKKLDKAKKVMENQMFEEKQAESMPEIETEQEYK
EIFITPHQIKHIKDFKDYKYSHRVDKKPNRKLINDTLYSTRKDDKGNTLIVNNLNGLYDKDNDKLKKLINKSPEKL
LMYHHDPQTYQKLKLIMEQYGDEKNPLYKYYEETGNYLTKYSKKDNGPVIKKIKYYGNKLNAHLDITDDYPNSRNK
VVKLSLKPYRFDVYLDNGVYKFVTVKNLDVIKKENYYEVNSKCYEEAKKLKKISNQAEFIASFYKNDLIKINGELY
RVIGVNNDLLNRIEVNMIDITYREYLENMNDKRPPHIIKTIASKTQSIKKYSTDILGNLYEVKSKKHPQIIKKG
(SEQ ID NO:357)

[0415] 在一些实施方案中,所公开的碱基编辑器包含napDNAbp,其包含衍生自耳葡萄球菌的Cas9蛋白(S.auri Cas9或SauriCas9)。在一些实施方案中,所公开的碱基编辑器包含SauriCas9切口酶。SauriCas9识别NNGG和NNNGG PAM。SauriCas9(切口酶)的序列如SEQ ID NO:358所示。在一些实施方案中,所公开的碱基编辑器包含napDNAbp结构域,其具有与SEQ ID NO:358至少90%、至少95%、至少98%或至少99%相同的序列。在一些实施方案中,所公开的碱基编辑器包含napDNAbp,其包含SEQ ID NO:358。该蛋白的长度为1061个氨基酸。

[0416] MQENQQKQNYILGLAIGITSVGYGLIDSKTREVIDAGVRLFPEADSENNSNRRSKRGARRLKRRRIHR
LNRVKDLLADYQMIDLNNVPKSTDPYTIRVKGLREPLTKEEFAIALLHIAKRRGLHNISVSMGDEEQDNELSTKQQ
LQKNAQQLQDKYVCELQLERLTNINKVRGEKNRFKTEDFVKEVKQLCETQRQYHNIDDQFIQQYIDLVSTRREYFE
GPGNGSPYGWDGDLLKWYEKLMGRCTYFPEELRSVKYAYSADLFNALNDLNNLVVTRDDNPKLEYYEKYHIIENVF
KQKKNPTLKQIAKEIGVQDYDIRGYRITKSGKPQFTSFKLYHDLKNIFEQAKYLEDVEMLDEIAKILTIYQDEISI
KKALDQLPELLTESEKSQIAQLTGYTGTHRLSLKCIHIVIDELWESPENQMEIFTRLNLKPKKVEMSEIDSIPTTL
VDEFILSPVVKRAFIQSIKVINAVINRFGLPEDIIIELAREKNSKDRRKFINKLQKQNEATRKKIEQLLAKYGNTN
AKYMIEKIKLHDMQEGKCLYSLEAIPLEDLLSNPTHYEVDHIIPRSVSFDNSLNNKVLVKQSENSKKGNRTPYQYL
SSNESKISYNQFKQHILNLSKAKDRISKKKRDMLLEERDINKFEVQKEFINRNLVDTRYATRELSNLLKTYFSTHD
YAVKVKTINGGFTNHLRKVWDFKKHRNHGYKHHAEDALVIANADFLFKTHKALRRTDKILEQPGLEVNDTTVKVDT
EEKYQELFETPKQVKNIKQFRDFKYSHRVDKKPNRQLINDTLYSTREIDGETYVVQTLKDLYAKDNEKVKKLFTER

PQKILMYQHDPKTFEKLMTILNQYAEAKNPLAAYYEDKGEYVTKYAKKGNGPAIHKIKYIDKKLGSYLDVSNKYPE TQNKLVKLSLKSFRFDIYKCEQGYKMVSIGYLDVLKKDNYYYIPKDKYEAEKQKKKIKESDLFVGSFYYNDLIMYE DELFRVIGVNSDINNLVELNMVDITYKDFCEVNNVTGEKRIKKTIGKRVVLIEKYTTDILGNLYKTPLPKKPQLIF KRGEL(SEQ ID NO:358)

[0417] 在一些实施方案中,napDNAbp包含SauriCas9-KKH变体或SauriCas9-KKH切口酶变体。SauriCas9-KKH含有相应的三重KKH突变:Q788K、Y973K和R1020H。参见Hu et al.(2020) PLoS Biol.18(3):e3000686,其通过引用并入本文。

[0418] 在一些实施方案中,本公开的碱基编辑器包含napDNAbp结构域,该结构域包含酿脓链球菌Cas9切口酶KKH,或SpCas9-KKH,其具有对应于NNNRRT的PAM。

[0419] 在一些实施方案中,Cas变体是SpRY的变体,其具有赋予高保真性的突变。这种变体称为SpRY-HF或SpRY-HF1。SpRY的高保真变体或本文提供的任何Cas变体可以包含相对于SEQ ID NO:74的N497A、R661A、Q695A和/或Q926A突变中的一个或多个,或本文提供的任何Cas9中的相应突变。具有高保真性的Cas9变体在本领域是已知的,并且对本领域技术人员来说是显而易见的。例如,高保真性的Cas9结构域已在Kleinstiver,B.P.,et al."High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects."Nature 529,490-495(2016);和Slaymaker,I.M.,et al."Rationally engineered Cas9 nucleases with improved specificity."Science 351,84-88(2015)中描述,其中每个通过引用并入本文。

[0420] 在一些实施方案中,所公开的Cas变体包括衍生自猕猴链球菌(Streptococcus macacae)(例如猕猴链球菌NCTC 11558)的Cas9,或SmacCas9的变体。在一些实施方案中,Cas变体包含SmacCas9的杂交变体,该变体并入了SpCas9结构域和SmacCas9结构域,并且称为Spy-macCas9或其变体。在一些实施方案中,Cas变体包含SmacCas9的杂交变体,其并入了SpCas9(iSpy Cas9)结构域的增强核酸裂解变体,并称为iSpy-macCas9。相对于Spymac-Cas9,iSpyMac-Cas9含有两个突变,R221K和N394K,这些突变是通过对Spy Cas9的深度突变扫描鉴定的,可以提高该蛋白在大多数靶标上的修饰率。参见Jakimoet al.,bioRxiv,A Cas9 with Complete PAM Recognition for Adenine Dinucleotides(Sep 2018),其通过引用并入本文。Jakimo et al.表明,杂交体Spy-macCas9和iSpy-macCas9识别短5'-NAA-3' PAM,并识别所有评估的腺嘌呤二核苷酸PAM序列,并在人类细胞中具有强大的编辑效率。Liu et al.工程改造了含有Spy-mac Cas9的碱基编辑器,并证明了含有Spymac结构域的胞苷和腺嘌呤碱基编辑器可以在体内诱导高效C至T和A至G转化。此外,Liu et al.提出,Spy-mac Cas9的PAM范围可以是5'-TAAA-3',而不是Jakimo et al.报道的5'-NAA-3'(参见Liuet al.Cell Discovery(2019)5:58,其通过引用并入本文)。

[0421] 上述与Cas9变体或Cas9等效物相关的任何参考文献均通过引用以其整体在此并入,如果尚未如此说明。

[0422] 下表提供了本公开的Nme2Cas变体与其他Cas同源物(R=任何嘌呤,Y=任何嘧啶,N=任何核苷酸)所靶向的PAM偏好的比较。

[0423] 表6-示例性Cas同源物的PAM偏好

[0424]

Cas 同源物	PAM 偏好
酿脓链球菌 Cas9(SpCas9)	NGG [典型]

[0425]

Cas 同源物	PAM 偏好
WT Nme1Cas9	N4GAYW/N4GYTT/N4GTCT
WT Nme2Cas9	NNNNCC 或 NNNNGATT
eNme2-C Cas9	NNNNCN
SaCas9	NNGRRT/NGRRT 或 NNGRRN/NGRRN
SaCas9-KKH	NNNRRT 或 NNGRRT
SauriCas9	NNGG
CjCas9	NNNNACA、NNNNRYAC 或 NNNNACAC
StCas9	NNAGAAW
SpRY	NYN
SmacCas9	NAA/TAAA
SpCas9-VRQR	NGAN 或 NGNG
SpCas9-EQR	NGAG
SpCas9-VRER	NGCG

[0426] PAM/原间隔区序列

[0427] 对于典型的(即酿脓链球菌Cas9衍生的)碱基编辑器,碱基编辑需要在位于靶核苷酸附近约15个碱基对的位置存在原间隔相邻基序(PAM)。每个可编程DNA结合蛋白结构域识别不同的PAM序列。只有大约四分之一的致病性转换点突变具有合适定位的典型PAM“NGG”序列,该序列与酿脓链球菌Cas9(SpCas9)衍生的碱基编辑器兼容。天然存在的胞苷脱氨酶已显示出与许多Cas同源物的广泛兼容性,包括金黄色葡萄球菌Cas9(SaCas9)[98]、SaCas9-KKH[8]、Cas12a(Cpf1)[9,10]、SpCas9-NG[11]和环状排列的CP-Cas9[7],极大地扩展了它们的靶向范围。

[0428] 在一些实施方案中,napDNAbp包含具有序列和位于具有序列上游的原间隔区序列。在一些实施方案中,原间隔区序列位于具有序列TGG的PAM的上游。在其他实施方案中,原间隔区序列位于具有序列GGG的PAM的上游。在其他实施方案中,原间隔区序列位于具有序列AGG的PAM的上游。在一些实施方案中,原间隔区序列位于具有序列CGG的PAM的上游。在一些实施方案中,原间隔区序列位于具有序列AGACCC的PAM的上游。在其他实施方案中,原间隔区序列位于具有序列ACCTCA的PAM的上游。在一些实施方案中,原间隔区序列位于具有序列GGGGCG的PAM的上游。在其他实施方案中,原间隔区序列位于具有序列CAGCCG的PAM的上游。在一些实施方案中,原间隔区序列位于具有序列GCGGCT的PAM的上游。在其他实施方案中,原间隔区序列位于具有序列GGGGCA的PAM的上游。在一些实施方案中,原间隔区序列位于具有序列AAGGGT的PAM的上游。在其他实施方案中,原间隔区序列位于具有序列TCGGGT的PAM的上游。在一些实施方案中,原间隔区序列位于具有序列GAGAGT的PAM的上游。在一些

实施方案中,原间隔区序列位于具有序列CAGAAT的PAM的上游。在一些实施方案中,原间隔区序列位于具有序列CTGGGT的PAM的上游。

[0429] 在一些实施方案中,预期的编辑碱基对位于PAM位点上游1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20个核苷酸处。在一些实施方案中,预期的编辑碱基对位于PAM位点下游。在一些实施方案中,预期的编辑碱基对位于PAM位点下游1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20个核苷酸处。在一些实施方案中,该方法不需要典型(例如NGG)PAM位点。在一些实施方案中,靶区域包含靶窗口,其中靶窗口包含靶核碱基对。

[0430] 本公开的原间隔区序列可以包括但不限于以下序列:

[0431]

序列	SEQ ID NO:
GGAATCCCTTCTGCAGCACC	208
GAGTCCGAGCAGAAGAAGAA	209
TTTATCACAGGCTCCAGGAA	210
GGCCCAGACTGAGCACGTGA	211
ACTCACGCTGGATAGCCTCC	212
GTCATCTTAGTCATTACCTG	213
GGCACTGCGGCTGGAGGTGG	214
CACCTACCTAAGAACCATCC	215
TTCGTATCTGTAAAACCAAG	216
CAATGTGTCAACTCTTGACA	217
CAGTTTCAGCACATCATGGT	218
GGTCCAGGCGAACGCGAAGT	219
GGAATCCCTTCTGCAGCACC	208

[0432]

GAGTCCGAGCAGAAGAAGAA	209
TTTATCACAGGCTCCAGGAA	210
GGCCCAGACTGAGCACGTGA	211
ACTCACGCTGGATAGCCTCC	212
GTCATCTTAGTCATTACCTG	213
GCAAGAGCACAAGAGGAAGAGAG	226
GGCTACAGCAACAGGGTGGTGG	227
GGTGTGCAGACGGCAGTCACTAG	228
GTCTCCGCTTTAACCCCCACCTC	229
CCTCTTGCCTCCACTGGTTGTGC	230
GGTCCAAAGCAGGATGACAGGCA	231
CACCCAGACTGAGCACGTGC	232
GACACAGACTGGGCACGTGA	233
AGCTCAGACTGAGCAAGTGA	234
AGACCAGACTGAGCAAGAGA	235
GAGCCAGAATGAGCACGTGA	236
TGCACTGCGGCCGGAGGAGG	237
GGCTCTGCGGCTGGAGGGGG	238
GGCACGACGGCTGGAGGTGG	239
GGCATCACGGCTGGAGGTGG	240
GGCGCTGCGGCGGGAGGTGG	241
GAGTTAGAGCAGAAGAAGAA	242
GAGTCTAAGCAGAAGAAGAA	243
GAGGCCGAGCAGAAGAAAGA	244
GAGTCCTAGCAGGAGAAGAA	245
AAGTCTGAGCACAAGAAGAA	246
GAGTCCGGGAAGGAGAAGAA	247
GAGCCGGAGCAGAAGAAGGA	248
AAGTCCGAGGAGAGGAAGAA	249
GAATCCAAGCAGGAGAAGAA	250

[0433]

ACGTCTGAGCAGAAGAAGAA	251
CCTATCACTGGCTCCAGGAA	252
CTTATCATAGGCCCCAGGAA	253
TAGGGTGTCAACTCTTGACA	254
TTATCTGTCAACTCTTGACA	255
ATACCTGCCAACTCTTGACA	256
TGTTCTCAGCACATCATGGT	257
GTGGTAGACAGCATGTGTCCTA	258
ATTTACAGCCTGGCCTTTGGGG	259
GTGTCAGGTAATGTGCTAAACA	260
GGTGGAGGAGGGTGCATGGGGT	261
TCTGCTTCTCCAGCCCTGGC	262
GATGTTCCAATCAGTACGCA	263

[0434] 编辑窗口

[0435] 本公开的碱基编辑器可以具有包含靶核碱基对的靶窗口(例如编辑窗口或脱氨窗口)的可变的靶区域,其中在靶区域内安装核碱基编辑。在一些实施方案中,TadA-CD具有对应于原间隔区的原间隔区位置2-12的C至T碱基编辑窗口。在一些实施方案中,TadA-CD碱基编辑器具有对应于原间隔区的原间隔区位置2-12的C至T碱基编辑窗口。在特定实施方案中,TadA-CD碱基编辑器具有对应于原间隔区位置3至8的C至T碱基编辑窗口。本公开的碱基编辑器在原间隔区位置5至7之间的胞嘧啶上可以具有特别高的编辑活性。

[0436] 在一些实施方案中,靶窗口(例如编辑窗口)包含1-10个核苷酸。在一些实施方案中,编辑窗口的长度为1-9、1-8、1-7、1-6、1-5、1-4、1-3、1-2或1个核苷酸。在一些实施方案中,靶窗口(例如编辑窗口)的长度为1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20个核苷酸。在一些实施方案中,预期的编辑碱基对位于编辑窗口内。在一些实施方案中,编辑窗口包含预期的编辑碱基对。

[0437] 在某些情况下,TadA-CD碱基编辑窗口从原间隔区的位置2之后、位置3之后、位置4之后、位置5之后、位置6之后、位置7之后、位置8之后、位置9之后、位置10之后和位置11之后开始。在一些实施方案中,编辑窗口在原间隔区的位置12之前、位置11之前、位置10之前、位置9之前、位置8之前、位置7之前、位置6之前、位置5之前、位置4之前和位置3之前结束。

[0438] 在一些实施方案中,包含V106W突变的TadA-CD碱基编辑器相对于缺乏该突变的TadA-CD碱基编辑器具有更窄的编辑窗口。例如,TadA-CDa(SEQ ID NO:34)的碱基编辑窗口在原间隔区的约位置4和约位置9之间。在某些实施方案中,包含V106W突变的TadA-CD碱基编辑器(例如TadA-CDa V106W和TadA-CDd V106W)在原间隔区的位置3和位置9之间具有C至T碱基编辑窗口,或其任何组合。例如,编辑器可以在原间隔区的位置3、位置4、位置5、位置

6、位置7、位置8或位置9处安装C至T置换,或其任何组合。

[0439] 在一些情况下,TadA-CD V106W碱基编辑窗口从原间隔区的位置2之后、位置4之后、位置5之后、位置6之后、位置7之后、位置8之后或位置9之后开始。在一些实施方案中,编辑窗口在原间隔区的位置10之前、位置9之前、位置8之前、位置7之前、位置6之前、位置5之前、位置4之前结束。

[0440] 在一些实施方案中,TadA-CD碱基编辑器在原间隔区的位置4和位置7之间具有A至G碱基编辑窗口。在一些情况下,TadA-CD碱基编辑器在原间隔区的位置4、位置5、位置6或位置7或其任何组合处安装A至G编辑。根据一些实施方案,通过包含V106W突变,TadA-CD的A至G碱基编辑特性可以缩小到原间隔区的位置5和位置7之间。

[0441] 本领域技术人员将理解,上文和本文所述的TadA-CD碱基编辑器具有比几种现有胞苷脱氨酶(如rAPOBEC1、evoAPOBEC1(evoA)、evoFERNY和YE1)更窄的C至T碱基编辑窗口。例如,BE4max和evoABE4max显示出范围从原间隔区的位置1至位置14的C至T编辑窗口;evoFERNY-BE4显示出从位置1至约位置10的C至T编辑窗口;并且YE1-BE4显示出从原间隔区的位置3至位置9的C至T编辑窗口(见图3)。

[0442] 本领域技术人员还将理解,上文和本文所述的TadA-CD碱基编辑器与其进化来源的亲本腺苷脱氨酶(例如TadA-8e)相比,具有更窄的A至G碱基编辑窗口和更宽的C至T碱基编辑窗口。例如,TadA-8e显示出在原间隔区的位置1和位置15之间的A至G碱基编辑窗口,以及在原间隔区的位置4至位置7之间的C至T碱基编辑窗口。

[0443] 本公开的TadA-CD碱基编辑器可以在原间隔区序列内将一个或多个靶胞嘧啶转化为胸腺嘧啶。例如,在一些实施方案中,TadA-CB可以在原间隔区序列内转化2个胞嘧啶、3个胞嘧啶、4个胞嘧啶或5个胞嘧啶。

[0444] 编辑效率

[0445] 本公开的方面涉及本文所述的胞嘧啶碱基编辑器在包含靶碱基对的靶窗口的靶区域内编辑DNA靶序列的效率。在一些实施方案中,至少10%、15%、20%、25%、30%、35%、40%、45%或50%的预期碱基对被编辑。在一些实施方案中,任何所公开的碱基编辑器或使用这些碱基编辑器的方法的C至T转化的效率在所有测序读段中为至少80%。在特定实施方案中,TadCBEa实现了平均51%-60%的靶胞嘧啶的转化效率。

[0446] 在一些实施方案中,任何所公开的碱基编辑器或使用这些碱基编辑器的方法在临床相关基因(如CXCR5和CCR5基因,其与HIV/AIDS相关)中提供了平均70%的胞嘧啶转化效率。

[0447] 在一些实施方案中,所公开的脱氨酶的胞苷脱氨活性(以及因此所公开的碱基编辑器的胞嘧啶编辑活性)以显著比例超过该脱氨酶的腺苷脱氨活性。例如,所公开的Tad-CD脱氨酶的胞苷脱氨活性与腺苷脱氨活性的比率为至少约5:1、6:1、7:1、8:1、9:1、10:1、11:1、12:1、13:1、14:1、15:1、17:1、19:1、20:1、21:1、23:1、25:1、30:1或超过30:1。在一些实施方案中,脱氨酶的胞苷脱氨活性与腺苷脱氨活性的比率为至少约10:1。在一些实施方案中,该比率为至少约20:1。在一些实施方案中,该比率为约5:1-7.5:1、7.5:1-9.5:1、5:1-10:1、10:1-15:1、15:1-20:1、10-17:1、12:1-17:1、20:1-21:1、21:1-25:1、20:1-30:1、25:1-35:1、30:1-35:1、30:1-40:1、40:1-42:1、21:1-42:1、25:1-40:1、10:1-40:1、25-45:1、30:1-50:1、45:1-50:1、50:1-60:1、55:1-65:1、60:1-70:1、70:1-80:1、80:1-85:1、10:1-80:1、

40:1-80:1、20:1-60:1、20:1-80:1或75:1-85:1。

[0448] 在一些实施方案中,TadA-CD的峰值编辑效率与天然胞嘧啶碱基编辑器(例如含有APOBEC1、evoFERNY或evoA脱氨酶的BE4max编辑器)相当。在一些实施方案中,TadA-CD碱基编辑器的编辑效率相对于天然胞嘧啶碱基编辑器更高。例如,在一些实施方案中,TadA-CDa、TadA-CDb和TadA-CDc以5%和48%之间的效率在原间隔区的位置3-8编辑Nme50基因。在一些情况下,TadA-CD包含V106W置换,其在缩小碱基编辑器的编辑窗口的同时维持编辑效率。

[0449] 在一些实施方案中,公开的TadCBE和编辑方法包括将DNA与任何所公开的TadCBE接触的步骤,导致在所有测序读段中,靶核碱基对处至少约20%、21%、25%、30%、35%、40%、50%、60%、70%、80%、85%或超过85%的中靶(C至T)碱基编辑效率。接触步骤可以导致至少约40%、41%、42%、43%、44%、45%、46%、47%、48%、49%、50%、52%、55%、60%、62%、65%、70%、72%、75%、80%、82%、85%或超过85%的C至T碱基编辑效率。特别是,接触步骤导致大于75%的中靶碱基编辑效率。在某些实施方案中,可以实现99%的碱基编辑效率。

[0450] 在一些情况下,本文所述的TadA-CD碱基编辑器具有在20%和80%之间的C至T编辑效率。在一些实施方案中,C至T编辑效率大于或等于10%、20%、25%、30%、40%、45%、50%、55%、60%、65%、70%、75%或80%。在其他实施方案中,C至T编辑效率低于或等于95%、低于或等于90%、低于或等于80%、低于或等于70%、低于或等于60%、低于或等于50%、低于或等于40%、低于或等于30%、低于或等于20%、低于或等于10%、低于或等于5%或低于或等于1%。

[0451] 在一些情况下,本公开的碱基编辑器可以对给定原间隔区序列内的靶向核苷酸具有不同的碱基编辑效率(例如将C转化为T)。换句话说,本发明的TadA-CD碱基编辑器可以优先编辑原间隔区序列中的某个位置(或多个位置)。例如,在一些实施方案中,TadA-CDa变体优先编辑原间隔区序列$GC_2A_3A_4GA_6GC_8A_9C_{10}A_{11}A_{12}$GAGGAAGAGAGAGACCC(SEQ ID NO:385)的C8位置,其中PAM加下划线;而TadA-CDc变体以相似的效率编辑C8和C10位置两者。

[0452] 因此,在一些实施方案中,TadCBE在编辑窗口内的原间隔区的每个位置(例如位置1至位置15)处的编辑效率在20%和80%之间。在一些实施方案中,编辑窗口内的原间隔区的每个位置(例如位置1至位置15)处的编辑效率大于或等于10%、大于或等于20%、大于或等于30%、大于或等于40%、大于或等于50%、大于或等于60%、大于或等于70%、大于或等于80%或大于或等于85%。在其他实施方案中,编辑窗口内的原间隔区的每个位置(例如位置1至位置15)处的编辑效率低于或等于85%、低于或等于80%、低于或等于70%、低于或等于60%、低于或等于50%、低于或等于40%、低于或等于30%、低于或等于20%、低于或等于10%、低于或等于5%或低于或等于1%。

[0453] 因此,在一些实施方案中,本申请的TadCBE在与包含靶序列的DNA接触时提供了至少20%、21%、25%、30%、35%、40%、41%、42%、43%、44%、45%、46%、47%、48%、49%、50%、52%、55%、60%、62%、65%、70%、72%、75%、80%、82%、85%或超过85%的C至T碱基转化效率,靶序列选自由以下组成的组:CTT、CTC、CTA、CTG、CCT、CCC、CCA、CCG、CAT、CAC、CAA、CAG、CGT、CGC、CGA、CGG、TCT、TCC、TCG、ACT、ACC、ACA、ACG、GCT、GCC、GCA、GCG、TTC、TAC、TGC、ATC、AAC、AGC、GTC、GAC和GGC。

[0454] 所公开的TadCBE对胞嘧啶碱基具有更高的亲和力和特异性,并且因此不太容易脱氨腺苷残基。在一些情况下,本文所述的TadA-CD碱基编辑器具有较低的残留A至G编辑效率,例如在0.1%和20%之间。在一些实施方案中,A至G编辑效率大于或等于0.1%、大于或等于5%、大于或等于10%或大于或等于20%。在其他实施方案中,A至G编辑效率低于或等于95%、低于或等于90%、低于或等于80%、低于或等于70%、低于或等于60%、低于或等于50%、低于或等于40%、低于或等于30%、低于或等于20%、低于或等于10%、低于或等于5%或低于或等于0.1%。

[0455] 在一些情况下,本文所述的TadA-CD碱基编辑器具有残留的(脱靶)C至G编辑能力。在一些实施方案中,V106W变体与天然TadA-CD碱基编辑器相比具有减少的C至G编辑。在一些实施方案中,TadA-CD V016W突变体在HEK293T细胞、T细胞和HSPC中减少C至G编辑。

[0456] 向导序列

[0457] 本公开进一步提供了用于根据所公开的编辑方法使用的向导RNA。本公开提供了设计用于识别靶序列的向导RNA(gRNA)。此类gRNA可以设计为具有与靶序列内的原间隔区具有互补性的向导序列(或"间隔区")。

[0458] 还提供了用于与一个或多个所公开的腺嘌呤碱基编辑器一起使用的向导RNA,例如在所公开的编辑核酸分子的方法中。此类gRNA可以设计为具有与待编辑的靶序列内的原间隔区具有互补性的向导序列,和具有与任何所公开的碱基编辑器的napDNAbp结构域(例如所公开的碱基编辑器的Cas9切口酶结构域)特异性相互作用的主链序列。

[0459] 在各种实施方案中,碱基编辑器可以与一个或多个向导序列复合、结合或以其他方式缔合(例如经由任何类型的共价键或非共价键)。向导序列变为与碱基编辑器缔合或结合,并引导其定位到与向导序列或其部分具有互补性的特定靶序列。向导序列的具体设计实施方案将取决于基因组靶序列(即待编辑的期望位点)的核苷酸序列和碱基编辑器中存在的napDNAbp类型(例如Cas9蛋白类型),以及其他因素,如PAM序列位置、靶序列中的百分比G/C含量、微同源区域的程度、二级结构等。

[0460] 通常,向导序列是与靶多核苷酸序列具有足够互补性的任何多核苷酸序列,以与靶序列杂交并引导napDNAbp(例如Cas9或Cas9变体)与靶序列的序列特异性结合。在一些实施方案中,当使用合适的比对算法进行最佳比对时,向导序列与其相应靶序列之间的互补性程度为约或超过约50%、60%、75%、80%、85%、90%、95%、97.5%、99%或更多。最佳比对可以通过使用任何合适的序列比对算法确定,其非限制性示例包括Smith-Waterman算法、Needleman-Wunsch算法、基于Burrows-Wheeler变换的算法(例如Burrows Wheeler Aligner)、ClustalW、Clustal X、BLAT、Novoalign(Novocraft Technologies)、ELAND(Illumina,San Diego,Calif.)、SOAP(可在soap.genomics.org.cn获得)和Maq(可在maq.sourceforge.net获得)。

[0461] 在一些实施方案中,向导序列的长度为约或超过约5、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、35、40、45、50、75、80、85、90、95、100或更多个核苷酸。在其他实施方案中,向导序列的长度为约或超过约20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50、55、60、65、70、75、80、85、90、95、100、125、150、175或200个核苷酸。在一些实施方案中,每个gRNA包含与靶序列(或脱靶位点)互补的至少10个连续核苷酸(例如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、38、39或40个连续核苷酸)的向导序列。在一些实施方案中,每个gRNA包含与靶序列(或脱靶位点)互补的至少15个连续核苷酸的向导序列。在其他实施方案中,每个gRNA包含与靶序列(或脱靶位点)互补的至少20个连续核苷酸的向导序列。

[0462] 在一些实施方案中,向导序列的长度小于约75、50、45、40、35、30、25、20、15、12或更少个核苷酸。向导序列引导碱基编辑器与靶序列特异性结合的能力可以通过任何合适的测定进行评估。例如,可以将碱基编辑器的组分(包括要测试的向导序列)提供给具有相应靶序列的宿主细胞,如通过用编码本文公开的碱基编辑器组分的载体转染,然后评估靶序列内的优先切割。类似地,可以通过提供靶序列、碱基编辑器的组分,包括待测试的向导序列和与测试向导序列不同的对照向导序列来原位评估靶多核苷酸序列的切割,并比较测试向导序列反应和对照向导序列反应之间在靶序列上的结合或切割速率。其他测定也是可能的,并且本领域技术人员可以想到。

[0463] 可以选择向导序列以靶向任何靶序列。在一些实施方案中,靶序列是细胞基因组内的序列。示例性靶序列包括在靶基因组中独特的那些。

[0464] 在一些实施方案中,选择向导序列以减少向导序列内的二级结构程度。二级结构可以通过任何合适的多核苷酸折叠算法确定。一些程序基于计算最小吉布斯自由能。一个此类算法的实例是mFold,如Zuker&Stiegler(Nucleic Acids Res.9(1981),133-148)所述。另一个实例折叠算法是在线网络服务器RNAfold,由维也纳大学理论化学研究所开发,使用中心结构预测算法(参见例如A.R.Gruber et al.,2008,Cell 106(1):23-24;和PA Carr&GM Church,2009,NatureBiotechnology 27(12):1151-62)。其他算法可见于Chuai,G.et al.,DeepCRISPR:optimized CRISPR guide RNA design by deep learning,Genome Biol.19:80(2018),以及美国申请系列号61/836,080和美国专利号8,871,445(2014年10月28日发布)中,其中每个的全部内容通过引用并入本文。

[0465] gRNA的向导序列与tracr配偶(也称为"主链")序列连接,tracr配偶序列又与tracr序列杂交。tracr配偶序列包括与tracr序列具有足够互补性的任何序列,以促进以下一个或多个:(1)在含有相应tracr序列的细胞中切除由侧翼为tracr配偶序列的向导序列;和(2)在靶序列处形成复合物,其中该复合物包含与tracr序列杂交的tracr配偶序列。通常,互补性程度是指tracr配偶序列和tracr序列沿两个序列的较短序列长度的最佳比对。最佳比对可以通过任何合适的比对算法确定,并且可以进一步考虑二级结构,如tracr序列或tracr配偶序列内的自互补性。在一些实施方案中,当最佳比对时,tracr序列和tracr配偶序列沿两个序列的较短序列长度的互补性程度为约或超过约25%、30%、40%、50%、60%、70%、80%、90%、95%、97.5%、99%或更高。在一些实施方案中,tracr序列的长度为约或超过约5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、25、30、40、50或更多个核苷酸。在一些实施方案中,tracr序列和tracr配偶序列包含在单一转录物内,使得两者之间的杂交产生具有二级结构(如发夹)的转录物。用于发夹结构的优选成环序列长度为四个核苷酸,并且最优选具有序列GAAA。然而,可以使用更长或更短的环序列,也可以使用替代序列。这些序列优选包括核苷酸三联体(例如AAA)和额外的核苷酸(例如C或G)。成环序列的实例包括CAAA和AAAG。在本发明的实施方案中,转录物或转录的多核苷酸序列具有至少两个或更多个发夹。在某些实施方案中,转录物具有两个、三个、四个或五个发夹。在本发明的进一

步实施方案中,转录物具有最多五个发夹。在一些实施方案中,单一转录物还包括转录终止序列;优选这是polyT序列,例如六个T核苷酸。

[0466] 包含向导序列、tracr配偶序列和tracr序列的单一(DNA)多核苷酸的非限制性示例如下(从5'到3'列出),其中"N"表示向导序列的碱基,第一组小写字母表示tracr配偶序列,并且第二组小写字母表示tracr序列,并且最后的poly-T序列表示转录终止子:

[0467] (1)NNNNNNNNgtttttgtactctcaagatttaGAAAtaaatcttgcagaagctacaaagataaggcttcatgccgaaatcaacaccctgtcattttatggcagggtgttttcgttatttaaTTTTTT(SEQ ID NO:264);

[0468] (2)NNNNNNNNNNNNNNNNNNgtttttgtactctcaGAAAtgcagaagctacaaagataaggcttcatgccgaaatcaacaccctgtcattttatggcagggtgttttcgttatttaaTTTTTT(SEQ ID NO:265);

[0469] (3)NNNNNNNNNNNNNNNNNNNNgtttttgtactctcaGAAAtgcagaagctacaaagataaggcttcatgccgaaatcaacaccctgtcattttatggcagggtgtTTTTT(SEQ ID NO:266);

[0470] (4)NNNNNNNNNNNNNNNNNNNNgttttagagctaGAAAtagcaagttaaaataaggctagtccgttatcaacttgaaaaagtggcaccgagtcggtgcTTTTTT(SEQ ID NO:267);

[0471] (5)NNNNNNNNNNNNNNNNNNNNgttttagagctaGAAATAGcaagttaaaataaggctagtccgttatcaacttgaaaaagtgTTTTTTT(SEQ ID NO:268);以及

[0472] (6)NNNNNNNNNNNNNNNNNNNNgttttagagctagAAATAGcaagttaaaataaggctagtccgttatcaTTTTTTTT(SEQ ID NO:269)。

[0473] 在一些实施方案中,序列(1)至(3)与来自嗜热链球菌CRISPR1的Cas9结合使用。在一些实施方案中,序列(4)至(6)与来自酿脓链球菌的Cas9结合使用。在一些实施方案中,tracr序列是与包含tracr配偶序列的转录物分开的转录物。

[0474] 在一些实施方案中,根据所公开的编辑方法使用的向导RNA包含含有修饰的核糖核苷酸的合成单一向导RNA(sgRNA)。在一些实施方案中,向导RNA含有修饰,如2'-O-甲基化核苷酸和硫代磷酸酯键。在一些实施方案中,向导RNA在前三个和最后三个核苷酸中含有2'-O-甲基修饰,并在前三个和最后三个核苷酸之间含有硫代磷酸酯键。示例性修饰的合成sgRNA在Hendel A.et al.,Nat.Biotechnol.33,985-989(2015)中公开,其通过引用并入本文。

[0475] 在一些实施方案中,根据所公开的编辑方法使用的向导RNA包含被酿脓链球菌Cas9蛋白或结构域(如所公开的碱基编辑器的SpCas9结构域)识别的主链结构。被SpCas9蛋白识别的主链结构可以包含序列5'-[向导序列]-guuuuagagcuagaaauagcaaguuaaaauaaggcuaguccguuaucaacuugaaaaaguggcaccgagucggugcuuuuu-3'(SEQ ID NO:339),其中向导序列包含与靶序列的原间隔区序列互补的序列。参见美国公布号2015/0166981(2015年6月18日公布),其公开内容通过引用并入本文。向导序列通常为20个核苷酸长。

[0476] 在其他实施方案中,根据所公开的编辑方法使用的向导RNA包含被金黄色葡萄球菌Cas9蛋白识别的主链结构。被SaCas9蛋白识别的主链结构可以包含序列5'-[向导序列]-guuuuaguacucuguaaugaaaauuacagaaucuacuaaaacaaggcaaaaugccguguuuaucucgucaacuuguuggcgagauuuuuuu-3'(SEQ ID NO:78)。

[0477] 在一些实施方案中,根据所公开的编辑方法使用的向导RNA包含被脑膜炎奈瑟菌Cas9蛋白或结构域(如Nme2Cas9结构域)识别的主链结构。被Nme2Cas9蛋白识别的主链结构(或支架)可以包含以下提供的序列:5'-[向导序列]-gttgtagctccctttctcatttcggaaacga

aatgagaaccgttgctacaataaggccgtctgaaaagatgtgccgcaacgctctgccccttaaagcttctgctttaaggggcatcgttta-3'(SEQ ID NO:274)。该支架序列被Nme1Cas9、Nme2Cas9和Nme3Cas9蛋白识别。

[0478] 基于本公开,用于将公开的TadCBE靶向特定基因组位点的合适向导RNA序列对于本领域技术人员将是显而易见的。此类合适的向导RNA序列通常包含与待编辑的靶碱基对上游或下游50个核苷酸内的核酸序列互补的向导序列。本文提供了适用于将任何提供的TadCBE靶向特定靶序列的一些示例性向导RNA序列。另外的向导序列在本领域是众所周知的,并且可以与本文所述的碱基编辑器一起使用。另外的示例性向导序列公开于例如Jinek M.,et al.,Science 337:816-821(2012);Mali P,Esvelt KM&Church GM(2013)Cas9 as a versatile tool for engineering biology,Nature Methods,10,957-963;Li JF et al.,(2013)Multiplex and homologous recombination-mediated genome editing in Arabidopsis and Nicotiana benthamiana using guide RNA and Cas9,Nature Biotechnology,31,688-691;Hwang,W.Y.et al.,Efficient genome editing in zebrafish using a CRISPR-Cas system,Nature Biotechnology 31,227-229(2013);Cong L et al.,(2013)Multiplex genome engineering using CRIPSR/Cas systems,Science,339,819-823;Cho SW et al.,(2013)Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease,Nature Biotechnology,31,230-232;Jinek,M.et al.,RNA-programmed genome editing in human cells,eLife 2,e00471(2013);Dicarlo,J.E.et al.,Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems.Nucleic Acid Res.(2013);Briner AE et al.,(2014)Guide RNA functional modules direct Cas9 activity and orthogonality,Mol Cell,56,333-339,其中每个的全部内容通过引用并入本文。

[0479] 复合物

[0480] 本公开的方面涉及包含任何碱基编辑器和与任何napDNAbp结构域结合的向导RNA的复合物,如本文所述。在一些实施方案中,复合物包含长约15-100个核苷酸的向导RNA,并且包含与靶序列互补的至少10个、至少15个或至少20个连续核苷酸的序列。在一些情况下,向导RNA包含与靶序列互补的15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39或40个连续核苷酸的序列。在某些实施方案中,向导RNA为20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50、55、60、65、70、75、80、85、90、95、100、125、150、175或200个核苷酸长。

[0481] 在一些实施方案中,复合物包含含有DNA序列的靶序列。在一些情况下,靶序列位于生物体的基因组中。该生物体在一些实施方案中可以是原核生物或真核生物。该生物体在一些实施方案中可以是本领域技术人员所知的任何类型的原核生物或真核生物。例如,在一些实施方案中,原核生物是细菌,并且真核生物是植物或真菌。然而,在一些情况下,真核生物可以是脊椎动物或哺乳动物。根据某些实施方案,哺乳动物可以是例如啮齿动物或人。

[0482] 在某些实施方案中,复合物包含位于细胞基因组中的靶序列。该细胞可以来自任何来源(例如原核生物或真核生物),并且可以是任何分类(例如肌肉细胞、皮肤细胞、心脏细胞、肝细胞等),如本领域技术人员所知的。示例性实施方案包括但不限于小鼠细胞、大鼠

细胞或人细胞。在一些实施方案中,细胞是T细胞或造血干细胞。

[0483] TadA衍生的胞嘧啶碱基编辑器

[0484] 本公开的方面涉及包含与核酸可编程DNA结合蛋白(napDNAbp结构域)(例如Cas9结构域)融合的TadA衍生的胞苷脱氨酶(例如TadA-CD)的新型胞嘧啶碱基编辑器(TadCBE)。任何所公开的胞嘧啶碱基编辑器可以包含一个或多个接头。任何所公开的胞嘧啶碱基编辑器可以包含一个或多个UGI结构域,例如两个UGI结构域。任何所公开的胞嘧啶碱基编辑器可以包含一个或多个接头。任何所公开的胞嘧啶碱基编辑器可以包含一个或多个核定位序列(NLS)。

[0485] 相对于现有的碱基编辑器,所公开的新型TadCBE包含TadA衍生的胞苷脱氨酶(如TadA-CDa、TadA-CDc、TadA-CDd和TadA-CDd(V106W)脱氨酶)、napDNAbp结构域、一个或多个尿嘧啶糖基化酶抑制剂(UGI)结构域和核定位序列(NLS)结构域的新型组合。

[0486] 所公开的TadCBE还可以包含一个或多个核定位信号(NLS)和/或两个尿嘧啶糖基化酶抑制剂(UGI)结构域。因此,碱基编辑器可以包含以下结构:NH_2-[第一核定位序列]-[TadA-CD结构域]-[napDNAbp结构域]-[第一UGI结构域]-[第二UGI结构域]-[第二核定位序列]-COOH,其中每个“]-[”表示存在任选的接头序列。示例性TadCBE可以具有包含“BE4max”架构的结构,具有NH_2-[NLS]-[TadA-CD结构域]-[Cas9切口酶结构域]-[UGI结构域]-[UGI结构域]-[NLS]-COOH结构,具有优化的核定位信号,并且其中napDNAbp结构域包含Cas9切口酶,如SpCas9n。据报道,这种BE4max结构对于在人细胞中表达具有优化的密码子使用,如Koblan et al.,Nat Biotechnol.2018;36(9):843-846所报道,其通过引用并入本文。

[0487] 在其他实施方案中,示例性TadCBE可以具有包含修饰的BE4max架构的结构,该架构含有napDNAbp结构域,该结构域包含除Cas9切口酶之外的Cas9变体,如SpCas9-NG、SaCas9n、eNme2Cas9n、CjCas9n或enCjCas9n。因此,本公开的示例性TadCBE可以包含以下结构:NH_2-[TadA-CD结构域]-[SaCas9n]-[UGI结构域]-[UGI结构域]-COOH;NH_2-[TadA-CD结构域]-[eNme2Cas9n]-[UGI结构域]-[UGI结构域]-COOH;NH_2-[TadA-CD结构域]-[eNme2-C Cas9n]-[UGI结构域]-[UGI结构域]-COOH;NH_2-[TadA-CD结构域]-[CjCas9n]-[UGI结构域]-[UGI结构域]-COOH;或NH_2-[TadA-CD结构域]-[SpCas9-NG]-[UGI结构域]-[UGI结构域]-COOH,其中每个“]-[”指示存在任选的接头序列。另外的示例性TadCBE可以包含以下结构:NH_2-[NLS]-[TadA-CD结构域]-[SaCas9n]-[UGI结构域]-[UGI结构域]-[NLS]-COOH;NH_2-[NLS]-[TadA-CD结构域]-[eNme2Cas9n]-[UGI结构域]-[UGI结构域]-[NLS]-COOH;NH_2-[NLS]-[TadA-CD结构域]-[eNme2-C Cas9n]-[UGI结构域]-[UGI结构域]-[NLS]-COOH;NH_2-[NLS]-[TadA-CD结构域]-[CjCas9n]-[UGI结构域]-[UGI结构域]-[NLS]-COOH;或NH_2-[NLS]-[TadA-CD结构域]-[SpCas9-NG]-[UGI结构域]-[UGI结构域]-[NLS]-COOH。

[0488] 在一些实施方案中,napDNAbp结构域包含与SEQ ID NO:19具有至少80%、85%、90%、92.5%、95%、98%或99.5%序列同一性的氨基酸序列。在一些实施方案中,napDNAbp结构域包含SEQ ID NO:19的氨基酸序列。在一些实施方案中,napDNAbp结构域包含与SEQ ID NO:25或28具有至少80%、85%、90%、92.5%、95%、98%或99.5%序列同一性的氨基酸序列。在一些实施方案中,napDNAbp结构域包含SEQ ID NO:25或28的氨基酸序列。在一些实施方案中,napDNAbp结构域包含与SEQ ID NO:21或22中任一个具有至少80%、85%、90%、

92.5%、95%、98%或99.5%序列同一性的氨基酸序列。在一些实施方案中,napDNAbp结构域包含SEQ ID NO:21或22的氨基酸序列。

[0489] 所公开的TadCBE可以识别扩展的PAM序列,具有改进的脱氨5'-TC靶标的效率,和/或在更窄的靶窗口中进行编辑。在一些实施方案中,所公开的胞嘧啶碱基编辑器包含经进化的核酸可编程DNA结合蛋白(napDNAbp),如经进化的Cas9。

[0490] 示例性胞嘧啶碱基编辑器包含与以下氨基酸序列SEQ ID NO:19-31至少85%、至少90%、至少95%、至少98%、至少99%或至少99.5%相同的序列。在一些实施方案中,本文所述的任何腺嘌呤碱基编辑器可以包含与SEQ ID NO:19-31中任一个氨基酸序列相比有1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、25、30个或超过30个氨基酸不同的氨基酸序列。这些差异可以包括相对于参考序列插入、缺失或置换的氨基酸。在一些实施方案中,所公开的腺苷脱氨酶结构域含有与SEQ ID NO:19-31中任一个具有约50、约75、约100、约125、约150、约175、约200、约300、约400、约500个或超过500个连续氨基酸相同的片段。

[0491] 在一些实施方案中,本文所述的TadCBE包含本领域技术人员已知的任何合适的胞苷脱氨酶结构域,如表10和表11中列出的那些。在一些实施方案中,本文所述的TadCBE包含本领域技术人员已知的任何合适的napDNAbp结构域,如本文和别处公开的那些。本领域技术人员将理解,本文提供的示例仅意在用于示例目的,并不旨在以任何方式进行限制。例如,本领域技术人员将理解,本文公开的方法和组合物可用于生产包含CD(例如表10和表11中所示的那些)和napDNAbp(例如SpCas9、SaCas9、eNme2-C Cas9、CjCas等)的各种组合的TadCBE。

[0492] 以下提供了示例性TadCBE。Tad-CD结构域以粗体和下划线表示。在一些实施方案中,以下所示的任一个TadCBE还可以包含V106W突变。

[0493] TadA-CDa碱基编辑器(SpCas9n napDNAbp结构域)(TadCBEa)的氨基酸序列

[0494] **MKRTADGSEFESPKKKRKV<u>SSEVEFSHEYWMRHALTLAKRARDEGAGPVGAVL</u>**

<u>VLNNRVIGEGWNRAIGLHDPTAHAEIMALRQGGLVMQNYRLFDATLYVTFEPCV</u>

<u>MCAGAMINSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECA</u>

<u>ALLCDFYRMPRQVFNSQKKAQSSIN</u>*SGGSSGGSSGSETPGTSESATPESSGGSSGGS*DK

KYSIGLAIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEAT

RLKRTARRRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHERHPIFGN

IVDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHMIKFRGHFLIEGDLNPDNSD

VDKLFIQLVQTYNQLFEENPINASGVDAKAILSARLSKSRRLENLIAQLPGEKKNGLFG

NLIALSLGLTPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIGDQYADLFLAAKNLS

DAILLSDILRVNTEITKAPLSASMIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSK

NGYAGYIDGGASQEEFYKFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIH

LGELHAILRRQEDFYPFLKDNREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSEETITP

WNFEEVVDKGASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVTE

GMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFKKIECFDSVEISGVEDRFNAS

LGTYHDLLKIIKDKDFLDNEENEDILEDIVLTLTLFEDREMIEERLKTYAHLFDDKVMK
QLKRRRYTGWGRLSRKLINGIRDKQSGKTILDFLKSDGFANRNFMQLIHDDSLTFKEDI
QKAQVSGQGDSLHEHIANLAGSPAIKKGILQTVKVVDELVKVMGRHKPENIVIEMARE
NQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQLQNEKLYLYYLQNGRDMY
VDQELDINRLSDYDVDHIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMK
NYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKHVAQILDSR
MNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNYHHAHDAYLNAVVGT
ALIKKYPKLESEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLAN
[0495] GEIRKRPLIETNGETGEIVWDKGRDFATVRKVLSMPQVNIVKKTEVQTGGFSKESILPK
RNSDKLIARKKDWDPKKYGGFDSPTVAYSVLVVAKVEKGKSKKLKSVKELLGITIME
RSSFEKNPIDFLEAKGYKEVKKDLIIKLPKYSLFELENGRKRMLASAGELQKGNELALP
SKYVNFLYLASHYEKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADANLD
KVLSAYNKHRDKPIREQAENIIHLFTLTNLGAPAAFKYFDTTIDRKRYTSTKEVLDATLI
HQSITGLYETRIDLSQLGGD*SGGSGGSGGSTNLSDIIEKETGKQLVIQESILMLPEEVEE*
*VIGNKPESDILVHTAYDESTDENVMLLTSDAPEYKPWALVIQDSNGENKIKML*SGGS**K**
RTADGSEFEPKKKRKV* (SEQ ID NO: 19)

[0496] TadA-CDb碱基编辑器(SpCas9n napDNAbp结构域)(TadCBEb)的氨基酸序列

MKRTADGSEFESPKKKRKV<u>**SSEVEFSHEYWMRHALTLAKRARDEGAGPVGAVL**
VLNNRVIGEGWNRAIGLHDPTAHAEIMALRQGGLVMQNYRLFDATLYVTFEPCV
MCAGAMINSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECA
ALLCDFYRMPRRVFNSQKKAQSSIN</u>*SGGSSGGSSGSETPGTSESATPESSGGSSGGS*DK
KYSIGLAIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEAT
RLKRTARRRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHERHPIFGN
[0497] IVDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHMIKFRGHFLIEGDLNPDNSD
VDKLFIQLVQTYNQLFEENPINASGVDAKAILSARLSKSRRLENLIAQLPGEKKNGLFG
NLIALSLGLTPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIGDQYADLFLAAKNLS
DAILLSDILRVNTEITKAPLSASMIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSK
NGYAGYIDGGASQEEFYKFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIH
LGELHAILRRQEDFYPFLKDNREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSEETITP
WNFEEVVDKGASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVTE

GMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFKKIECFDSVEISGVEDRFNAS
LGTYHDLLKIIKDKDFLDNEENEDILEDIVLTLTLFEDREMIEERLKTYAHLFDDKVMK
QLKRRRYTGWGRLSRKLINGIRDKQSGKTILDFLKSDGFANRNFMQLIHDDSLTFKEDI
QKAQVSGQGDSLHEHIANLAGSPAIKKGILQTVKVVDELVKVMGRHKPENIVIEMARE
NQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQLQNEKLYLYYLQNGRDMY
VDQELDINRLSDYDVDHIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMK
NYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKHVAQILDSR
MNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNYHHAHDAYLNAVVGT

[0498] ALIKKYPKLESEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLAN
GEIRKRPLIETNGETGEIVWDKGRDFATVRKVLSMPQVNIVKKTEVQTGGFSKESILPK
RNSDKLIARKKDWDPKKYGGFDSPTVAYSVLVVAKVEKGKSKKLKSVKELLGITIME
RSSFEKNPIDFLEAKGYKEVKKDLIIKLPKYSLFELENGRKRMLASAGELQKGNELALP
SKYVNFLYLASHYEKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADANLD
KVLSAYNKHRDKPIREQAENIIHLFTLTNLGAPAAFKYFDTTIDRKRYTSTKEVLDATLI
HQSITGLYETRIDLSQLGGD*SGGSGGSGGS**TNLSDIIEKETGKQLVIQESILMLPEEVEE*
VIGNKPESDILVHTAYDESTDENVMLLTSDAPEYKPWALVIQDSNGENKIKML*SGGS***K*
RTADGSEFEPKKKRKV* (SEQ ID NO: 20)

[0499] TadA-CDc碱基编辑器(SpCas9n napDNAbp结构域)(TadCBEc)的氨基酸序列

MKRTADGSEFESPKKKRKV<u>**SSEVEFSHEYWMRHALTLAKRARDEGAGPVGAVL**
VLNNRVIGEGWNRAIGLHDPTAHAEIMALRQGGLVMQNYRLFDATLYVTFEPCV
MCAGAMINSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECA
ALLCDFYRIPRQVFNSQKKAQSSIN</u>*SGGSSGGSSGSETPGTSESATPESSGGSSGGS*DKK
YSIGLAIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEATR
LKRTARRRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHERHPIFGNI

[0500] VDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHMIKFRGHFLIEGDLNPDNSD
VDKLFIQLVQTYNQLFEENPINASGVDAKAILSARLSKSRRLENLIAQLPGEKKNGLFG
NLIALSLGLTPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIGDQYADLFLAAKNLS
DAILLSDILRVNTEITKAPLSASMIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSK
NGYAGYIDGGASQEEFYKFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIH
LGELHAILRRQEDFYPFLKDNREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSEETITP

WNFEEVVDKGASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVTE
GMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFKKIECFDSVEISGVEDRFNAS
LGTYHDLLKIIKDKDFLDNEENEDILEDIVLTLTLFEDREMIEERLKTYAHLFDDKVMK
QLKRRRYTGWGRLSRKLINGIRDKQSGKTILDFLKSDGFANRNFMQLIHDDSLTFKEDI
QKAQVSGQGDSLHEHIANLAGSPAIKKGILQTVKVVDELVKVMGRHKPENIVIEMARE
NQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQLQNEKLYLYYLQNGRDMY
VDQELDINRLSDYDVDHIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMK
NYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKHVAQILDSR
MNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNYHHAHDAYLNAVVGT
[0501] ALIKKYPKLESEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLAN
GEIRKRPLIETNGETGEIVWDKGRDFATVRKVLSMPQVNIVKKTEVQTGGFSKESILPK
RNSDKLIARKKDWDPKKYGGFDSPTVAYSVLVVAKVEKGKSKKLKSVKELLGITIME
RSSFEKNPIDFLEAKGYKEVKKDLIIKLPKYSLFELENGRKRMLASAGELQKGNELALP
SKYVNFLYLASHYEKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADANLD
KVLSAYNKHRDKPIREQAENIIHLFTLTNLGAPAAFKYFDTTIDRKRYTSTKEVLDATLI
HQSITGLYETRIDLSQLGGD*SGGSGGSGGSTNLSDIIEKETGKQLVIQESILMLPEEVEE*
*VIGNKPESDILVHTAYDESTDENVMLLTSDAPEYKPWALVIQDSNGENKIKML*SGGS**K**
RTADGSEFEPKKKRKV* (SEQ ID NO: 21)

[0502] TadA-CDd碱基编辑器(SpCas9n napDNAbp结构域)(TadCBEd)的氨基酸序列

MKRTADGSEFESPKKKRKV<u>**SSEVEFSHEYWMRHALTLAKRARDERKAPVGAVL**
VLNNRVIGEGWNRAIGLHDPTAHAEIIALRQGGLVMQNYRLIDATLYVTFEPCV
MCAGAMINSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECA
ALLCDFYRMPRQVFNAQKKAQSSIN</u>*SGGSSGGSSGSETPGTSESATPESSGGSSGGS*DK
KYSIGLAIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEAT
[0503] RLKRTARRRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHERHPIFGN
IVDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHMIKFRGHFLIEGDLNPDNSD
VDKLFIQLVQTYNQLFEENPINASGVDAKAILSARLSKSRRLENLIAQLPGEKKNGLFG
NLIALSLGLTPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIGDQYADLFLAAKNLS
DAILLSDILRVNTEITKAPLSASMIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSK
NGYAGYIDGGASQEEFYKFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIH

LGELHAILRRQEDFYPFLKDNREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSEETITP
WNFEEVVDKGASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVTE
GMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFKKIECFDSVEISGVEDRFNAS
LGTYHDLLKIIKDKDFLDNEENEDILEDIVLTLTLFEDREMIEERLKTYAHLFDDKVMK
QLKRRRYTGWGRLSRKLINGIRDKQSGKTILDFLKSDGFANRNFMQLIHDDSLTFKEDI
QKAQVSGQGDSLHEHIANLAGSPAIKKGILQTVKVVDELVKVMGRHKPENIVIEMARE
NQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQLQNEKLYLYYLQNGRDMY
VDQELDINRLSDYDVDHIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMK
NYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKHVAQILDSR

[0504] MNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNYHHAHDAYLNAVVGT
ALIKKYPKLESEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLAN
GEIRKRPLIETNGETGEIVWDKGRDFATVRKVLSMPQVNIVKKTEVQTGGFSKESILPK
RNSDKLIARKKDWDPKKYGGFDSPTVAYSVLVVAKVEKGKSKKLKSVKELLGITIME
RSSFEKNPIDFLEAKGYKEVKKDLIIKLPKYSLFELENGRKRMLASAGELQKGNELALP
SKYVNFLYLASHYEKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADANLD
KVLSAYNKHRDKPIREQAENIIHLFTLTNLGAPAAFKYFDTTIDRKRYTSTKEVLDATLI
HQSITGLYETRIDLSQLGGD*SGGSGGSGGS**TNLSDIIEKETGKQLVIQESILMLPEEVEE*
VIGNKPESDILVHTAYDESTDENVMLLTSDAPEYKPWALVIQDSNGENKIKML*SGGS***K*
RTADGSEFEPKKKRKV** (SEQ ID NO: 22)

[0505] TadA-CDe碱基编辑器(SpCas9n napDNAbp结构域)(TadCBEe)的氨基酸序列

MKRTADGSEFESPKKKRKV<u>**SSEVEFSHEYWMRHALTLAKRARDERAGPVGAVL**
VLNNRVIGEGWNRAIGLHDPTAHAEIMALRQGGLVMQNHRLIDATLYVTFEPCV
MCAGAMINSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECA
ALLCDFYRMPRHVFNSQKKAQSSIN</u>*SGGSSGGSSGSETPGTSESATPESSGGSSGGS*DK
KYSIGLAIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEAT

[0506] RLKRTARRRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHERHPIFGN
IVDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHMIKFRGHFLIEGDLNPDNSD
VDKLFIQLVQTYNQLFEENPINASGVDAKAILSARLSKSRRLENLIAQLPGEKKNGLFG
NLIALSLGLTPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIGDQYADLFLAAKNLS
DAILLSDILRVNTEITKAPLSASMIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSK

NGYAGYIDGGASQEEFYKFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIH
LGELHAILRRQEDFYPFLKDNREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSEETITP
WNFEEVVDKGASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVTE
GMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFKKIECFDSVEISGVEDRFNAS
LGTYHDLLKIIKDKDFLDNEENEDILEDIVLTLTLFEDREMIEERLKTYAHLFDDKVMK
QLKRRRYTGWGRLSRKLINGIRDKQSGKTILDFLKSDGFANRNFMQLIHDDSLTFKEDI
QKAQVSGQGDSLHEHIANLAGSPAIKKGILQTVKVVDELVKVMGRHKPENIVIEMARE
NQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQLQNEKLYLYYLQNGRDMY
VDQELDINRLSDYDVDHIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMK
NYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKHVAQILDSR
[0507] MNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNYHHAHDAYLNAVVGT
ALIKKYPKLESEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLAN
GEIRKRPLIETNGETGEIVWDKGRDFATVRKVLSMPQVNIVKKTEVQTGGFSKESILPK
RNSDKLIARKKDWDPKKYGGFDSPTVAYSVLVVAKVEKGKSKKLKSVKELLGITIME
RSSFEKNPIDFLEAKGYKEVKKDLIIKLPKYSLFELENGRKRMLASAGELQKGNELALP
SKYVNFLYLASHYEKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADANLD
KVLSAYNKHRDKPIREQAENIIHLFTLTNLGAPAAFKYFDTTIDRKRYTSTKEVLDATLI
HQSITGLYETRIDLSQLGGD*SGGSGGSGGSTNLSDIIEKETGKQLVIQESILMLPEEVEE*
*VIGNKPESDILVHTAYDESTDENVMLLTSDAPEYKPWALVIQDSNGENKIKML*SGGS**K**
RTADGSEFEPKKKRKV* (SEQ ID NO: 23)

[0508] TadA-CDa(V106W)碱基编辑器(SpCas9n napDNAbp结构域)(TadCBEa(V106W))的氨基酸序列

MKRTADGSEFESPKKKRKV<u>**SSEVEFSHEYWMRHALTLAKRARDEGAGPVGAVL**
VLNNRVIGEGWNRAIGLHDPTAHAEIMALRQGGLVMQNYRLFDATLYVTFEPCV
MCAGAMINSRIGRVVFGWRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECA
ALLCDFYRMPRQVFNSQKKAQSSIN</u>*SGGSSGGSSGSETPGTSESATPESSGGSSGGS*DK
[0509] KYSIGLAIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEAT
RLKRTARRRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHERHPIFGN
IVDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHMIKFRGHFLIEGDLNPDNSD
VDKLFIQLVQTYNQLFEENPINASGVDAKAILSARLSKSRRLENLIAQLPGEKKNGLFG
NLIALSLGLTPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIGDQYADLFLAAKNLS

DAILLSDILRVNTEITKAPLSASMIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSK
NGYAGYIDGGASQEEFYKFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIH
LGELHAILRRQEDFYPFLKDNREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSEETITP
WNFEEVVDKGASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVTE
GMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFKKIECFDSVEISGVEDRFNAS
LGTYHDLLKIIKDKDFLDNEENEDILEDIVLTLTLFEDREMIEERLKTYAHLFDDKVMK
QLKRRRYTGWGRLSRKLINGIRDKQSGKTILDFLKSDGFANRNFMQLIHDDSLTFKEDI
QKAQVSGQGDSLHEHIANLAGSPAIKKGILQTVKVVDELVKVMGRHKPENIVIEMARE
NQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQLQNEKLYLYYLQNGRDMY
VDQELDINRLSDYDVDHIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMK
[0510] NYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKHVAQILDSR
MNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNYHHAHDAYLNAVVGT
ALIKKYPKLESEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLAN
GEIRKRPLIETNGETGEIVWDKGRDFATVRKVLSMPQVNIVKKTEVQTGGFSKESILPK
RNSDKLIARKKDWDPKKYGGFDSPTVAYSVLVVAKVEKGKSKKLKSVKELLGITIME
RSSFEKNPIDFLEAKGYKEVKKDLIIKLPKYSLFELENGRKRMLASAGELQKGNELALP
SKYVNFLYLASHYEKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADANLD
KVLSAYNKHRDKPIREQAENIIHLFTLTNLGAPAAFKYFDTTIDRKRYTSTKEVLDATLI
HQSITGLYETRIDLSQLGGD*SGGSGGSGGS**TNLSDIIEKETGKQLVIQESILMLPEEVEE*
VIGNKPESDILVHTAYDESTDENVMLLTSDAPEYKPWALVIQDSNGENKIKML*SGGS***K*
RTADGSEFEPKKKRKV* (SEQ ID NO: 24)

[0511] TadA-CDd(V106W)碱基编辑器(SpCas9n napDNAbp结构域)(TadCBEd(V106W))的氨基酸序列

MKRTADGSEFESPKKKRKV<u>**SSEVEFSHEYWMRHALTLAKRARDERKAPVGAVL**
VLNNRVIGEGWNRAIGLHDPTAHAEIIALRQGGLVMQNYRLIDATLYVTFEPCV
MCAGAMINSRIGRVVFGWRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECA
ALLCDFYRMPRQVFNAQKKAQSSIN</u>*SGGSSGGSSGSETPGTSESATPESSGGSSGGS*DK
[0512] KYSIGLAIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEAT
RLKRTARRRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHERHPIFGN
IVDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHMIKFRGHFLIEGDLNPDNSD
VDKLFIQLVQTYNQLFEENPINASGVDAKAILSARLSKSRRLENLIAQLPGEKKNGLFG

NLIALSLGLTPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIGDQYADLFLAAKNLS
DAILLSDILRVNTEITKAPLSASMIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSK
NGYAGYIDGGASQEEFYKFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIH
LGELHAILRRQEDFYPFLKDNREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSEETITP
WNFEEVVDKGASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVTE
GMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFKKIECFDSVEISGVEDRFNAS
LGTYHDLLKIIKDKDFLDNEENEDILEDIVLTLTLFEDREMIEERLKTYAHLFDDKVMK
QLKRRRYTGWGRLSRKLINGIRDKQSGKTILDFLKSDGFANRNFMQLIHDDSLTFKEDI
QKAQVSGQGDSLHEHIANLAGSPAIKKGILQTVKVVDELVKVMGRHKPENIVIEMARE
NQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQLQNEKLYLYYLQNGRDMY
VDQELDINRLSDYDVDHIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMK
[0513] NYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKHVAQILDSR
MNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNYHHAHDAYLNAVVGT
ALIKKYPKLESEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLAN
GEIRKRPLIETNGETGEIVWDKGRDFATVRKVLSMPQVNIVKKTEVQTGGFSKESILPK
RNSDKLIARKKDWDPKKYGGFDSPTVAYSVLVVAKVEKGKSKKLKSVKELLGITIME
RSSFEKNPIDFLEAKGYKEVKKDLIIKLPKYSLFELENGRKRMLASAGELQKGNELALP
SKYVNFLYLASHYEKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADANLD
KVLSAYNKHRDKPIREQAENIIHLFTLTNLGAPAAFKYFDTTIDRKRYTSTKEVLDATLI
HQSITGLYETRIDLSQLGGD*SGGSGGSGGSTNLSDIIEKETGKQLVIQESILMLPEEVEE*
***VIGNKPESDILVHTAYDESTDENVMLLTSDAPEYKPWALVIQDSNGENKIKML*SGGS**K
RTADGSEFEPKKKRKV* (SEQ ID NO: 25)

[0514] TadA-CDf碱基编辑器(SpCas9n napDNAbp结构域)(TadCBEf)的氨基酸序列

MKRTADGSEFESPKKKRKV<u>**SSEVEFSHEYWMRHALTLAKRARDEGEAPVGAVL**
VLNNRVIGEGWNRRIGLHDPTAHAEIMALRQGGLVMQNSRLIDATLYVTFEPCV
MCAGAMINSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECA
ALLCDFYRMPRQVFNAQKKAQSSIN*SGGSSGGSSGSETPGTSESATPESSGGSSGGS*</u>DK
KYSIGLAIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEAT
RLKRTARRRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHERHPIFGN
IVDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHMIKFRGHFLIEGDLNPDNSD

VDKLFIQLVQTYNQLFEENPINASGVDAKAILSARLSKSRRLENLIAQLPGEKKNGLFG
NLIALSLGLTPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIGDQYADLFLAAKNLS
DAILLSDILRVNTEITKAPLSASMIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSK
NGYAGYIDGGASQEEFYKFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIH
LGELHAILRRQEDFYPFLKDNREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSEETITP
WNFEEVVDKGASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVTE
GMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFKKIECFDSVEISGVEDRFNAS
LGTYHDLLKIIKDKDFLDNEENEDILEDIVLTLTLFEDREMIEERLKTYAHLFDDKVMK
QLKRRRYTGWGRLSRKLINGIRDKQSGKTILDFLKSDGFANRNFMQLIHDDSLTFKEDI
QKAQVSGQGDSLHEHIANLAGSPAIKKGILQTVKVVDELVKVMGRHKPENIVIEMARE
NQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQLQNEKLYLYYLQNGRDMY
[0516] VDQELDINRLSDYDVDHIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMK
NYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKHVAQILDSR
MNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNYHHAHDAYLNAVVGT
ALIKKYPKLESEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLAN
GEIRKRPLIETNGETGEIVWDKGRDFATVRKVLSMPQVNIVKKTEVQTGGFSKESILPK
RNSDKLIARKKDWDPKKYGGFDSPTVAYSVLVVAKVEKGKSKKLKSVKELLGITIME
RSSFEKNPIDFLEAKGYKEVKKDLIIKLPKYSLFELENGRKRMLASAGELQKGNELALP
SKYVNFLYLASHYEKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADANLD
KVLSAYNKHRDKPIREQAENIIHLFTLTNLGAPAAFKYFDTTIDRKRYTSTKEVLDATLI
HQSITGLYETRIDLSQLGGD*SGGSGGSGGSTNLSDIIEKETGKQLVIQESILMLPEEVEE*
***VIGNKPESDILVHTAYDESTDENVMLLTSDAPEYKPWALVIQDSNGENKIKML*SGGS**K
RTADGSEFEPKKKRKV* (SEQ ID NO: 26)

[0517] TadA-CDg碱基编辑器(SpCas9n napDNAbp结构域)(TadCBEg)的氨基酸序列

MKRTADGSEFESPKKKRKV<u>**SSEVEFSHEYWMRHALTLAKRARDEGEGPVGAVL**</u>
<u>**VLNNRVIGEGWNRRIGLHDPTAHAEIMALRQGGLVMQNCKLIDATLYVTFEPCV**</u>
<u>**MCAGAMIHSRIGRVVFGVRNSKRGAAGSLMNVLNYPGMNHRVEITEGILADECA**</u>
[0518] <u>**ALLCDFYRMPRQVFNAQKKAQSSIN***SGGSSGGSSGSETPGTSESATPESSGGSSGGS*</u>DK
KYSIGLAIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEAT
RLKRTARRRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHERHPIFGN

[0519] IVDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIYLALAHMIKFRGHFLIEGDLNPDNSD
VDKLFIQLVQTYNQLFEENPINASGVDAKAILSARLSKSRRLENLIAQLPGEKKNGLFG
NLIALSLGLTPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIGDQYADLFLAAKNLS
DAILLSDILRVNTEITKAPLSASMIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSK
NGYAGYIDGGASQEEFYKFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQIH
LGELHAILRRQEDFYPFLKDNREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSEETITP
WNFEEVVDKGASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVTE
GMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFKKIECFDSVEISGVEDRFNAS
LGTYHDLLKIIKDKDFLDNEENEDILEDIVLTLTLFEDREMIEERLKTYAHLFDDKVMK
QLKRRRYTGWGRLSRKLINGIRDKQSGKTILDFLKSDGFANRNFMQLIHDDSLTFKEDI
QKAQVSGQGDSLHEHIANLAGSPAIKKGILQTVKVVDELVKVMGRHKPENIVIEMARE
NQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQLQNEKLYLYYLQNGRDMY
VDQELDINRLSDYDVDHIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMK
NYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKHVAQILDSR
MNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNYHHAHDAYLNAVVGT
ALIKKYPKLESEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLAN
GEIRKRPLIETNGETGEIVWDKGRDFATVRKVLSMPQVNIVKKTEVQTGGFSKESILPK
RNSDKLIARKKDWDPKKYGGFDSPTVAYSVLVVAKVEKGKSKKLKSVKELLGITIME
RSSFEKNPIDFLEAKGYKEVKKDLIIKLPKYSLFELENGRKRMLASAGELQKGNELALP
SKYVNFLYLASHYEKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADANLD
KVLSAYNKHRDKPIREQAENIIHLFTLTNLGAPAAFKYFDTTIDRKRYTSTKEVLDATLI
HQSITGLYETRIDLSQLGGD*SGGSGGSGGSTNLSDIIEKETGKQLVIQESILMLPEEVEE*
*VIGNKPESDILVHTAYDESTDENVMLLTSDAPEYKPWALVIQDSNGENKIKML*SGGS**K**
RTADGSEFEPKKKRKV* (SEQ ID NO: 27)

[0520] TadA-CDa:eNme2Cas9碱基编辑器的氨基酸序列

[0521] MKRTADGSEFESPKKKRKVSSEVEFSHEYWMRHALTLAKRARDEGAGPVGAVLVLNNRVIGEGWNRAI
GLHDPTAHAEIMALRQGGLVMQNYRLFDATLYVTFEPCVMCAGAMINSRIGRVVFGVRNSKRGAAGSLMNVLNYPG
MNHRVEITEGILADECAALLCDFYRMPRQVFNSQKKAQSSINSGGSSGGSSGSETPGTSESATPESSGGSSGGSAA
FKPNPINYILGLAIGIASVGWAMVEIDEEGNPIRLIDLGVRVFERAEVPKTGDSLAMARRLARSVRRLTRRRAHRL
LRARRLLKREGVLQAADFDENGLITSLPNTPWQLRAAALDRKLTPLEWSAVLLHLIKHRGYLSQRKNEGETAAKEL
GALLKGVANNAHALQTGDFRTPAELALNKFEKESGHIRNQRGDYSHTFSRKDLQAELILLFEKQKEFGNPHVSGGL
KEGIETLLMTQRPALSGDAVQKMLGHCTLEPTEPKAAKNTYTAERFIWLTKLNNLRILEQGSERPLTDTERSTLMD
EPYRKSKLTYAQARKLLGLEDTAFFKGLRYGKDNAEASTLMEMKAYHAISRALEKEGLKDKKSPLNLSSELQDEIG
TAFSLFKTDEDITGRLKDRVQPEILEALLKHISFDKFVQISLKALRRIVPLMEQGKRYDEACAEIYGVHYGKKNTE
EKIYLPPIPADEIRNPVVLRALSQARKVINGVVRRYGSPARIHIETAREVGKSFKDRKEIAKRQEENRKDREKAAA

KFREYFPNFVGEPKSKDILKLRLYEQQHGKCLYSGKEINLVRLNEKGYVEIDHALPFSRTWDDSFNNKVLVLGSEN
QNKGNQTPYEYFNGKDNSREWQEFKARVETSRFPSSKKQRILLQKFDEDGFKECNLNDTRYVNRFLCQFVADHILL
TGKGKRRVVASNGQITNLLRGFWRLRKVRAENDRHHALDAVVVACSTVAMQQKITRFVRYKEMNAFDGKTVDKETG
KVLYQKTHFPQPWEFFAQEVMIRVFGKPDGKPEFEEADTPEKLRTLLAEKLSSRPEAVHEYVTPLFVSRAPNRKMS
GAHKDTLRSAKRFVKHNEKISVKRVWLTEIKLADLENMVNYKNGREIELYEALKARLEAYGGNAKQAFDPKDNPFY
KKGGQLVKAVRVEKTQKSGVLLNKKNAYTIADNGDMVRVDVFCKVDKKGKNQYFIVPIYAWQVAENILPDIDCKGY
RIDDSYTFCFSLHKYDLIAFQKDEKSKVEFAYYINCDSSSGGFYLAWHDKGSREQRFRISTQNLALIQKYQVNELG
KEIRPCRLKKRPPVSGGSGGSGGSTNLSDIIEKETGKQLVIQESILMLPEEVEEVIGNKPESDILVHTAYDESTDE
NVMLLTSDAPEYKPWALVIQDSNGENKIKMLSGGSGGSGGSTNLSDIIEKETGKQLVIQESILMLPEEVEEVIGNK
PESDILVHTAYDESTDENVMLLTSDAPEYKPWALVIQDSNGENKIKMLSGGSKRTADGSEFEPKKKRKV(SEQ ID
NO:28)

[0522] TadA-CDa:SaCas9碱基编辑器的氨基酸序列

[0523] MKRTADGSEFESPKKKRKVSSEVEFSHEYWMRHALTLAKRARDEGAGPVGAVLVLNNRVIGEGWNRAI
GLHDPTAHAEIMALRQGGLVMQNYRLFDATLYVTFEPCVMCAGAMINSRIGRVVFGVRNSKRGAAGSLMNVLNYPG
MNHRVEITEGILADECAALLCDFYRMPRQVFNSQKKAQSSINSGGSSGGSSGSETPGTSESATPESSGGSSGGSGK
RNYILGLAIGITSVGYGIIDYETRDVIDAGVRLFKEANVENNEGRRSKRGARRLKRRRRHRIQRVKKLLFDYNLLT
DHSELSGINPYEARVKGLSQKLSEEEFSAALLHLAKRRGVHNVNEVEEDTGNELSTKEQISRNSKALEEKYVAELQ
LERLKKDGEVRGSINRFKTSDYVKEAKQLLKVQKAYHQLDQSFIDTYIDLLETRRTYYEGPGEGSPFGWKDIKEWY
EMLMGHCTYFPEELRSVKYAYNADLYNALNDLNNLVITRDENEKLEYYEKFQIIENVFKQKKKPTLKQIAKEILVN
EEDIKGYRVTSTGKPEFTNLKVYHDIKDITARKEIIENAELLDQIAKILTIYQSSEDIQEELTNLNSELTQEEIEQ
ISNLKGYTGTHNLSLKAINLILDELWHTNDNQIAIFNRLKLVPKKVDLSQQKEIPTTLVDDFILSPVVKRSFIQSI
KVINAIIKKYGLPNDIIIELAREKNSKDAQKMINEMQKRNRQTNERIEEIIRTTGKENAKYLIEKIKLHDMQEGKC
LYSLEAIPLEDLLNNPFNYEVDHIIPRSVSFDNSFNNKVLVKQEENSKKGNRTPFQYLSSSDSKISYETFKKHILN
LAKGKGRISKTKKEYLLEERDINRFSVQKDFINRNLVDTRYATRGLMNLLRSYFRVNNLDVKVKSINGGFTSFLRR
KWKFKKERNKGYKHHAEDALIIANADFIFKEWKKLDKAKKVMENQMFEEKQAESMPEIETEQEYKEIFITPHQIKH
IKDFKDYKYSHRVDKKPNRELINDTLYSTRKDDKGNTLIVNNLNGLYDKDNDKLKKLINKSPEKLLMYHHDPQTYQ
KLKLIMEQYGDEKNPLYKYYEETGNYLTKYSKKDNGPVIKKIKYYGNKLNAHLDITDDYPNSRNKVVKLSLKPYRF
DVYLDNGVYKFVTVKNLDVIKKENYYEVNSKCYEEAKKLKKISNQAEFIASFYNNDLIKINGELYRVIGVNNDLLN
RIEVNMIDITYREYLENMNDKRPPRIIKTIASKTQSIKKYSTDILGNLYEVKSKKHPQIIKKGSGGSGGSGGSTNL
SDIIEKETGKQLVIQESILMLPEEVEEVIGNKPESDILVHTAYDESTDENVMLLTSDAPEYKPWALVIQDSNGENK
IKMLSGGSGGSGGSTNLSDIIEKETGKQLVIQESILMLPEEVEEVIGNKPESDILVHTAYDESTDENVMLLTSDAP
EYKPWALVIQDSNGENKIKMLSGGSKRTADGSEFEPKKKRKV(SEQ ID NO:29)

[0524] TadA-CDa:SpCas9-NG碱基编辑器的氨基酸序列

[0525] MKRTADGSEFESPKKKRKVSSEVEFSHEYWMRHALTLAKRARDEGAGPVGAVLVLNNRVIGEGWNRAI
GLHDPTAHAEIMALRQGGLVMQNYRLFDATLYVTFEPCVMCAGAMINSRIGRVVFGVRNSKRGAAGSLMNVLNYPG
MNHRVEITEGILADECAALLCDFYRMPRQVFNSQKKAQSSINSGGSSGGSSGSETPGTSESATPESSGGSSGGSDK
KYSIGLAIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETAEATRLKRTARRRYTRRKNRI
CYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHERHPIFGNIVDEVAYHEKYPTIYHLRKKLVDSTDKADLRLIY
LALAHMIKFRGHFLIEGDLNPDNSDVDKLFIQLVQTYNQLFEENPINASGVDAKAILSARLSKSRRLENLIAQLPG

EKKNGLFGNLIALSLGLTPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIGDQYADLFLAAKNLSDAILLSDILR
VNTEITKAPLSASMIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQSKNGYAGYIDGGASQEEFYKFIKPILEKM
DGTEELLVKLNREDLLRKQRTFDNGSIPHQIHLGELHAILRRQEDFYPFLKDNREKIEKILTFRIPYYVGPLARGN
SRFAWMTRKSEETITPWNFEEVVDKGASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVTEGMR
KPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFKKIECFDSVEISGVEDRFNASLGTYHDLLKIIKDKDFLDNEE
NEDILEDIVLTLTLFEDREMIEERLKTYAHLFDDKVMKQLKRRRYTGWGRLSRKLINGIRDKQSGKTILDFLKSDG
FANRNFMQLIHDDSLTFKEDIQKAQVSGQGDSLHEHIANLAGSPAIKKGILQTVKVVDELVKVMGRHKPENIVIEM
ARENQTTQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQLQNEKLYLYYLQNGRDMYVDQELDINRLSDYDVD
HIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMKNYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAG
FIKRQLVETRQITKHVAQILDSRMNTKYDENDKLIREVKVITLKSKLVSDFRKDFQFYKVREINNYHHAHDAYLNA
VVGTALIKKYPKLESEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLANGEIRKRPLIETNG
ETGEIVWDKGRDFATVRKVLSMPQVNIVKKTEVQTGGFSKESIRPKRNSDKLIARKKDWDPKKYGGFVSPTVAYSV
LVVAKVEKGKSKKLKSVKELLGITIMERSSFEKNPIDFLEAKGYKEVKKDLIIKLPKYSLFELENGRKRMLASARF
LQKGNELALPSKYVNFLYLASHYEKLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADANLDKVLSAYN
KHRDKPIREQAENIIHLFTLTNLGAPRAFKYFDTTIDRKVYRSTKEVLDATLIHQSITGLYETRIDLSQLGGDSGG
SGGSGGSTNLSDIIEKETGKQLVIQESILMLPEEVEEVIGNKPESDILVHTAYDESTDENVMLLTSDAPEYKPWAL
VIQDSNGENKIKMLSGGSGGSGGSTNLSDIIEKETGKQLVIQESILMLPEEVEEVIGNKPESDILVHTAYDESTDE
NVMLLTSDAPEYKPWALVIQDSNGENKIKMLSGGSKRTADGSEFEPKKKRKV(SEQ ID NO:30)

[0526] TadA-CDa:enCjCas9碱基编辑器的氨基酸序列

[0527] MKRTADGSEFESPKKKRKVSSEVEFSHEYWMRHALTLAKRARDEGAGPVGAVLVLNNRVIGEGWNRAI
GLHDPTAHAEIMALRQGGLVMQNYRLFDATLYVTFEPCVMCAGAMINSRIGRVVFGVRNSKRGAAGSLMNVLNYPG
MNHRVEITEGILADECAALLCDFYRMPRQVFNSQKKAQSSINSGGSSGGSSGSETPGTSESATPESSGGSSGGSAR
ILAFAIGISSIGWAFSENDELKDCGVRIFTKVENPKTGESLALPRRLARSARKRYARRKARLNHLKHLIANEFKLN
YEDYQSFDESLAKAYKGSLISPYELRFRALNELLSKQDFARVILHIAKRRGYDDIKNSDDKEKGAILKAIKQNEEK
LANYQSVGEYLYKEYFQKFKENSKEFTNVRNKKESYERCIAQSFLKDELKLIFKKQREFGFSFSKKFEEEVLSVAF
YKRALKDFSHLVGNCSFFTDEKRAPKNSPLAFMFVALTRIINLLNNLKNTEGILYTKDDLNALLNEVLKNGTLTYK
QTKKLLGLSDDYEFKGEKGTYFIEFKKYKEFIKALGEHNLSQDDLNEIAKDITLIKDEIKLKKALAKYDLNQNQID
SLSKLEFKDHLNISFKALKLVTPLMLEGKKYDEACNELNLKVAINEDKKDFLPAFNETYYKDEVTNPVVLRAIKEY
RKVLNALLKKYGKVHKINIELAREVGKNHSQRAKIEKEQNENYKAKKDAELECEKLGLKINSKNILKLRLFKEQKE
FCAYSGEKIKISDLQDEKMLEIDHIYPYSRSFDDSYMNKVLVFTKQNQEKLNQTPFEAFGNDSAKWQKIEVLAKNL
PTKKQKRILDKNYKDKEQKNFKDRNLNDTRYIARLVLNYTKDYLDFLPLSDDENTKLNDTQKGSKVHVEAKSGMLT
SALRHTWGFSAKDRNNHLHHAIDAVIIAYANNSIVKAFSDFKKEQESNSAELYAKKISELDYKNKRKFFEPFSGFR
QKVLDKIDEIFVSKPERKKPSGALHEETFRKEEEFYQSYGGKEGVLKALELGKIRKVNGKIVKNGDMFRVDIFKHK
KTNKFYAVPIYTMDFALKVLPNKAVARSKKGEIKDWILMDENYEFCFSLYKDSLILIQTKKMQEPEFVYYNAFTSS
TVSLIVSKHDNKFETLSKNQKILFKNANEKEVIAKSIGIQNLKVFEKYIVSALGEVTKAEFRQREDFKKSGGSGGS
GGSTNLSDIIEKETGKQLVIQESILMLPEEVEEVIGNKPESDILVHTAYDESTDENVMLLTSDAPEYKPWALVIQD
SNGENKIKMLSGGSKRTADGSEFEPKKKRKV(SEQ ID NO:31)

[0528] 接头

[0529] 本文所述的融合蛋白可以包括一个或多个接头。如上所定义,本文所用的术语“接

头”指的是连接两个分子或部分(例如核酸酶的结合结构域和切割结构域)的化学基团或分子。在一些实施方案中,接头将RNA可编程核酸酶的gRNA结合结构域与脱氨酶(例如胞苷脱氨酶或腺苷脱氨酶)的催化结构域连接起来。在一些实施方案中,接头将Cas结构域与本文提供的脱氨酶变体连接起来。通常,接头位于两个基团、分子或其他部分之间或侧接于两个基团、分子或其他部分,并经由共价键与每个基团连接,从而将两者连接起来。在一些实施方案中,接头是一个氨基酸或多个氨基酸(例如肽或蛋白质)。在一些实施方案中,接头是有机分子、基团、聚合物或化学部分。在一些实施方案中,接头的长度为5-100个氨基酸,例如长度为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、30-35、35-40、40-45、45-50、50-60、60-70、70-80、80-90、90-100、100-150或150-200个氨基酸。也考虑了更长或更短的接头。

[0530] 接头可以像共价键一样简单,或者其可以是多原子长度的聚合物接头。在某些实施方案中,接头是基于多肽或氨基酸的。在其他实施方案中,接头不是肽样的。在某些实施方案中,接头是共价键(例如碳-碳键、二硫键、碳-杂原子键等)。在某些实施方案中,接头是酰胺键的碳-氮键。在某些实施方案中,接头是环状或非环状、取代或未取代、支链或非支链的脂肪族或杂脂肪族接头。在某些实施方案中,接头是聚合物(例如聚乙烯、聚乙二醇、聚酰胺、聚酯等)。在某些实施方案中,接头包含氨基烷酸的单体、二聚体或聚合物。在某些实施方案中,接头包含氨基烷酸(例如甘氨酸、乙酸、丙氨酸、β-丙氨酸、3-氨基丙酸、4-氨基丁酸、5-戊酸等)。在某些实施方案中,接头包含氨基已酸(Ahx)的单体、二聚体或聚合物。在某些实施方案中,接头基于碳环部分(例如环戊烷、环己烷)。在其他实施方案中,接头包含聚乙二醇部分(PEG)。在其他实施方案中,接头包含氨基酸。在某些实施方案中,接头包含肽。在某些实施方案中,接头包含芳基或杂芳基部分。在某些实施方案中,接头基于苯环。接头可以包括官能化部分,以促进来自肽的亲核体(例如硫醇、氨基)附接至接头。任何亲电试剂都可以用作接头的一部分。示例性亲电试剂包括但不限于活化酯、活化酰胺、迈克尔受体、烷基卤化物、芳基卤化物、酰基卤化物和异硫氰酸酯。

[0531] 在一些其他实施方案中,接头包含氨基酸序列(GGGGS)n(SEQ ID NO:156)、(G)n(SEQ ID NO:157)、(EAAAK)n(SEQ ID NO:158)、(GGS)n(SEQ ID NO:159)、(SGGS)n(SEQ ID NO:160)、(XP)n(SEQ ID NO:161)或其任何组合,其中n独立地为1和30之间的整数,并且其中X为任何氨基酸。在一些实施方案中,接头包含氨基酸序列(GGS)n(SEQ ID NO:159),其中n为1、3或7。在一些实施方案中,接头包含氨基酸序列SGSETPGTSESATPES(SEQ ID NO:162)。在一些实施方案中,接头包含氨基酸序列SGSETPGTSESA(SEQ ID NO:163)。在一些实施方案中,接头包含氨基酸序列SGSETPGTSESATPEGGSGGS(SEQ ID NO:164)。在一些实施方案中,接头包含氨基酸序列SGGSSGGSSGSETPGTSESATPESSGGSSGGS(SEQ ID NO:165)。在一些实施方案中,接头包含氨基酸序列SGGSGGSGGS(SEQ ID NO:166)。在一些实施方案中,接头包含氨基酸序列SGGS(SEQ ID NO:160)。在其他实施方案中,接头包含氨基酸序列SGGSSGGSSGSETPGTSESATPESAGSYPYDVPDYAGSAAPAAKKKKLDGSGSGGSSGGS(SEQ ID NO:167)。在一些实施方案中,接头包含氨基酸序列GGS(SEQ ID NO:159)、GGSGGS(SEQ ID NO:168)、GGSGGSGGS(SEQ ID NO:169)、SGGSSGGSSGSETPGTSESATPESSGGSSGGSS(SEQ ID NO:170)、SGSETPGTSESATPES(SEQ ID NO:162)或SGGSSGGSSGSETPGTSESATPESAGSYPYDVPDYAGSAAPAAKKKKLDGSGSGGSSGGS(SEQ ID NO:167)。

[0532] 在某些实施方案中,接头可以用于连接本发明的任何肽或肽结构域或部分(例如与脱氨酶结构域连接或融合的napDNAbp)。本文所述的融合蛋白的任何结构域也可以通过任何上述接头相互连接。

[0533] UGI结构域和其他碱基编辑器组分

[0534] 在一些方面,本文所述的融合蛋白(例如碱基编辑器)可以包含一个或多个尿嘧啶糖基化酶抑制剂(UGI)结构域。在一些实施方案中,融合蛋白包含两个UGI结构域。UGI结构域指的是能够抑制尿嘧啶-DNA糖基化酶碱基切除修复酶的蛋白质。

[0535] 在一些实施方案中,UGI结构域包含野生型UGI或如SEQ ID NO:272所示的UGI或其变体。在一些实施方案中,本文提供的UGI蛋白包括UGI片段和与UGI或UGI片段同源的蛋白质。例如,在一些实施方案中,UGI结构域包含SEQ ID NO:272所示的氨基酸序列的片段。在一些实施方案中,UGI片段包含与SEQ ID NO:272所示的氨基酸序列至少60%、至少65%、至少70%、至少75%、至少80%、至少85%、至少90%、至少95%、至少96%、至少97%、至少98%、至少99%或至少99.5%相同的氨基酸序列。在一些实施方案中,UGI包含与SEQ ID NO:272所示的氨基酸序列同源的氨基酸序列,或与SEQ ID NO:272所示的氨基酸序列的片段同源的氨基酸序列。在一些实施方案中,包含UGI、UGI片段或UGI同源物的蛋白质被称为“UGI变体”。UGI变体与UGI或其片段具有同源性。例如,UGI变体与野生型UGI或SEQ ID NO:272所示的UGI至少70%相同、至少75%相同、至少80%相同、至少85%相同、至少90%相同、至少95%相同、至少96%相同、至少97%相同、至少98%相同、至少99%相同、至少99.5%相同或至少99.9%相同。在一些实施方案中,UGI变体包含UGI片段,使得该片段与野生型UGI或SEQ ID NO:272所示的UGI的相应片段至少70%相同、至少80%相同、至少90%相同、至少95%相同、至少96%相同、至少97%相同、至少98%相同、至少99%相同、至少99.5%相同或至少99.9%相同。在一些实施方案中,UGI包含以下氨基酸序列:

[0536] >sp|P14739|UNGI_BPPB2尿嘧啶-DNA糖基化酶抑制剂

[0537] MTNLSDIIEKETGKQLVIQESILMLPEEVEEVIGNKPESDILVHTAYDESTDENVMLLTSDAPEYKPWALVIQDSNGENKIKML(SEQ ID NO:272)。

[0538] 本文所述的融合蛋白(例如碱基编辑器)还可以包括一个或多个附加元件。在某些实施方案中,附加元件可以包含碱基修复的效应物,如碱基修复的抑制剂。

[0539] 在一些实施方案中,本文所述的碱基编辑器可以包含一个或多个异源蛋白质结构域(例如除了碱基编辑器组分之外约或超过约1、2、3、4、5、6、7、8、9、10个或更多个结构域)。碱基编辑器可以包含任何附加的蛋白质序列,并且任选地在任何两个结构域之间包含接头序列。可能存在的其他示例性特征是定位序列如细胞质定位序列、输出序列如核输出序列、或其他定位序列以及序列标签。

[0540] 可以融合到碱基编辑器或其组分(例如napDNAbp结构域、胞苷脱氨酶结构域或NLS结构域)的蛋白质结构域的实例包括但不限于表位标签和报告基因序列。表位标签的非限制性实例包括组氨酸(His)标签、V5标签、FLAG标签、流感血凝素(HA)标签、Myc标签、VSV-G标签和硫氧还蛋白(Trx)标签。报告基因的实例包括但不限于谷胱甘肽-5-转移酶(GST)、辣根过氧化物酶(HRP)、氯霉素乙酰转移酶(CAT)、β-半乳糖苷酶、β-葡萄糖醛酸酶、萤光素酶、绿色荧光蛋白(GFP)、HcRed、DsRed、青色荧光蛋白(CFP)、黄色荧光蛋白(YFP)和自荧光蛋白包括蓝色荧光蛋白(BFP)。碱基编辑器可以融合到编码蛋白质或蛋白质片段的基因序列,该

蛋白质或蛋白质片段结合DNA分子或结合其他细胞分子,包括但不限于麦芽糖结合蛋白(MBP)、S标签、Lex A DNA结合结构域(DBD)融合物、GAL4 DNA结合结构域融合物和单纯疱疹病毒(HSV)BP16蛋白融合物。其他可能形成碱基编辑器的一部分的结构域在美国专利公布号2011/0059502(2011年3月10日发布)中描述,其全部内容通过引用并入本文。

[0541] 可以与本文公开的碱基编辑器、方法和系统一起使用的报告基因序列包括但不限于谷胱甘肽-5-转移酶(GST)、辣根过氧化物酶(HRP)、氯霉素乙酰转移酶(CAT)、β-半乳糖苷酶、β-葡萄糖醛酸酶、萤光素酶、绿色荧光蛋白(GFP)、HcRed、DsRed、青色荧光蛋白(CFP)、黄色荧光蛋白(YFP)和自荧光蛋白包括蓝色荧光蛋白(BFP)、HSV胸苷激酶、rpoB,可以引入细胞以编码基因,该基因中可以引入将在用于所述系统的生长选择测定中赋予对特定培养基的抗性的突变。

[0542] 可能存在的其他示例性特征是用于融合蛋白的溶解、纯化或检测的标签。本文提供的合适蛋白质标签包括但不限于生物素羧化酶载体蛋白(BCCP)标签、myc标签、钙调蛋白标签、FLAG标签、血凝素(HA)标签、bgh-polyA标签、多组氨酸标签(也称为组氨酸标签或His标签)、麦芽糖结合蛋白(MBP)标签、nus标签、谷胱甘肽-S-转移酶(GST)标签、绿色荧光蛋白(GFP)标签、硫氧还蛋白标签、S标签、Softag(例如Softag 1、Softag 3)、strep标签、生物素连接酶标签、FlAsH标签、V5标签和SBP标签。另外的合适的序列对本领域技术人员来说是显而易见的。在一些实施方案中,融合蛋白可以包含一个或多个His标签。

[0543] 核定位序列(NLS)

[0544] 在各种实施方案中,本文所述的Cas蛋白可以融合到一个或多个核定位序列(NLS),这有助于促进蛋白质进入细胞核的转位(translocation)。在一些实施方案中,本文所述的融合蛋白可以包含一个或多个NLS。此类序列在本领域是众所周知的并且可以包括以下实例:

[0545]

描述	序列	SEQ ID NO:
SV40 大 T-Ag 的 NLS	PKKKRKV	142
NLS	MKRTADGSEFESPKKKRKV	143
NLS	MDSLLMNRRKFLYQFKNV RWAKGRRETYLC	144
核质蛋白的 NLS	AVKRPAATKKAGQAKKKK LD	145
EGL-13 的 NLS	MSRRRKANPTKLSENAKK LAKEVEN	146
c-MYC 的 NLS	PAAKRVKLD	147
TUS-蛋白的 NLS	KLKIKRPVK	148
多瘤大 T-Ag 的 NLS	VSRKRPRP	149
丁型肝炎病毒抗原的 NLS	EGAPPAKRAR	150
鼠 p53 的 NLS	PPQPKKKPLDGE	151
PE1 和 PE2 的 NLS	SGGSKRTADGSEFEPKKKR KV	152
二分 sv40 nls	KRTADGSEFESPKKKRKV	153

[0546] 上述NLS实例是非限制性的。本文提供的融合蛋白可以包含任何已知的NLS序列，包括Cokol et al.,“Finding nuclear localization signals,”EMBO Rep.,2000,1(5):411-415;和Freitas et al.,“Mechanisms and Signals for the Nuclear Import of Proteins,”Current Genomics,2009,10(8):550-7中描述的那些的任一个，其中每篇通过引用并入本文。

[0547] 在各种实施方案中，本文公开的融合蛋白和编码融合蛋白的构建体进一步包含一个或多个、优选至少两个核定位序列。在某些实施方案中，融合蛋白包含至少两个NLS。在具有至少两个NLS的实施方案中，NLS可以是相同的NLS，或者它们可以是不同的NLS。在一些实施方案中，一个或多个NLS是二分NLS(“bpNLS”)。在某些实施方案中，所公开的融合蛋白包含两个二分NLS。在一些实施方案中，所公开的融合蛋白包含超过两个二分NLS。

[0548] NLS融合的位置可以在融合蛋白的N端、C端或序列内部(例如插入在所编码的napDNAbp组分(例如本文公开的任何Cas14a1变体)和脱氨酶结构域(例如腺苷或胞苷脱氨酶)之间)。

[0549] NLS可以是本领域的任何已知NLS序列。NLS也可以是用于核定位的任何未来发现的NLS。NLS还可以是任何天然存在的NLS，或任何非天然存在的NLS(例如具有一个或多个期望突变的NLS)。

[0550] 术语“核定位序列”或“NLS”指的是促进蛋白质到细胞核中的输入的氨基酸序列，

例如通过核运输。核定位序列在本领域是众所周知的并且对本领域技术人员将是显而易见的。例如,NLS序列描述于Plank等人的国际PCT申请PCT/EP2000/011690(2000年11月23日提交,2001年5月31日发布为WO/2001/038547)中,其全部内容通过引用并入本文。在一些实施方案中,NLS包含氨基酸序列PKKKRKV(SEQ ID NO:142)、MDSLLMNRRKFLYQFKNVRWAKGRRETYLC(SEQ ID NO:144)、KRTADGSEFESPKKKRKV(SEQ ID NO:153)或KRTADGSEFEPKKKRKV(SEQ ID NO:155)。在其他实施方案中,NLS包含氨基酸序列NLSKRPAAIKKAGQAKKKK(SEQ ID NO:204)、PAAKRVKLD(SEQ ID NO:147)、RQRRNELKRSF(SEQ ID NO:205)或NQSSNFGPMKGGNFGGRSSGPYGGGGQYFAKPRNQGGY(SEQ ID NO:206)、KRPAATKKAGQAKKKK(SEQ ID NO:276)、KKTELQTTNAENKTKKL(SEQ ID NO:277)、KRGINDRNFWRGENGRKTR(SEQ ID NO:278)或RKSGKIAAIVVKRPRK(SEQ ID NO:279)。

[0551] 在本公开的一个方面,碱基编辑器、引导编辑器或其他融合蛋白可以使用一个或多个核定位序列(NLS)、优选至少两个NLS进行修饰。在某些实施方案中,融合蛋白使用两个或更多个NLS进行修饰。本公开考虑了在本公开时本领域已知的任何核定位序列的使用,或在本文提交后鉴定或以其他方式在现有技术中可得的任何核定位序列的使用。代表性的核定位序列是指导蛋白质至表达该序列的细胞核的肽序列。核定位信号主要是碱性的,可以位于蛋白质氨基酸序列的几乎任何位置,通常包含四个氨基酸(Autieri&Agrawal,(1998)J.Biol.Chem.273:14731-37,通过引用并入本文)至八个氨基酸的短序列,并且通常富含赖氨酸和精氨酸残基(Maginet al.,(2000)Virology 274:11-16,通过引用并入本文)。核定位序列通常包含脯氨酸残基。已经鉴定出多种核定位序列,并已用于影响生物分子从细胞质到细胞核的运输。参见例如Tinland et al.,(1992)Proc.Natl.Acad.Sci.U.S.A.89:7442-46;Moede et al.,(1999)FEBS Lett.461:229-34,其通过引用并入本文。目前认为转位涉及核孔蛋白。

[0552] 大多数核定位序列(NLS)可以分为三大类:(i)单组分NLS,由SV40大T抗原NLS(PKKKRKV,SEQ ID NO:142)例示;(ii)二分NLS,由通过可变数量的间隔区氨基酸分隔的两个碱性结构域组成,由非洲爪蟾核质蛋白NLS(KRXXXXXXXXXXKKKL,SEQ ID NO:154)例示;以及(iii)非典型序列,如hnRNP A1蛋白的M9、流感病毒核蛋白NLS和酵母Gal4蛋白NLS(Dingwall和Laskey 1991)。

[0553] 核定位序列出现在蛋白质氨基酸序列的不同位置。NLS已在蛋白质的N端、C端和中心区域中得到鉴定。因此,本公开提供了可以在融合蛋白的C端和/或N端以及内部区域使用一个或多个NLS修饰的融合蛋白。应选择不作为组成NLS残基发挥功能的较长序列的残基,以免在例如张力或空间上干扰核定位信号本身。因此,尽管包含NLS的序列在组成上没有严格的限制,但在实践中,这种序列在长度和组成上可能会受到功能上的限制。

[0554] 本公开考虑了修饰融合蛋白以包含一个或多个NLS的任何合适的方法。在一个方面,融合蛋白可以被工程化以表达在其N端或C端(或两者)翻译融合至一个或多个NLS的融合蛋白,即以形成Cas蛋白-NLS融合构建体、碱基编辑器-NLS融合构建体或引导编辑器-NLS融合构建体。在其他实施方案中,融合蛋白编码的核苷酸序列可以被遗传修饰以在编码的碱基编辑器的内部区域中引入编码一个或多个NLS的阅读框。此外,NLS可以包括在融合蛋白与N端、C端或内部附接的NLS氨基酸序列之间,例如并且在蛋白质的中心区域中,编码的各种氨基酸接头或间隔区域。因此,本公开还提供了用于表达包含碱基编辑器或引导编辑

器和一个或多个NLS以及其他组分的融合蛋白的核苷酸构建体、载体和宿主细胞。

[0555] 本文所述的融合蛋白还可以包含通过一个或多个接头与融合蛋白连接的核定位序列,接头例如聚合性、氨基酸、核酸、多糖、化学或核酸接头元件。本公开考虑范围内的接头并非旨在具有任何限制,并且可以是任何合适类型的分子(例如聚合物、氨基酸、多糖、核酸、脂质或任何合成的化学接头结构域),并且可以通过任何合适的策略与融合蛋白连接,从而在融合蛋白与一个或多个NLS之间形成键(例如共价键、氢键)。

[0556] 降低的脱靶效应

[0557] 本领域技术人员将理解,在编辑过程中可能会发生脱靶编辑事件。例如,碱基编辑可能导致胞苷和/或腺嘌呤碱基的非期望RNA编辑和/或脱靶DNA编辑,以及插入和缺失(插入缺失(indel))。在一些实施方案中,本公开的碱基编辑器包含与napDNAbp融合的进化胞苷脱氨酶,将总体脱靶编辑频率降低至约0.35%或更低。

[0558] 降低的RNA编辑效应

[0559] 本文公开的进化碱基编辑器具有降低的和/或低RNA编辑效应。在一些实施方案中,碱基编辑器经进化或工程化以具有降低的RNA编辑效应。如本文所用,术语"RNA编辑效应"指的是在细胞RNA(例如信使RNA(mRNA))中引入核苷酸的修饰(例如脱氨)。DNA碱基编辑效率的重要目标是DNA内特定核苷酸的修饰(例如脱氨),而在RNA内不引入类似核苷酸的修饰。当检测到的突变以0.3%或更低的频率引入RNA分子时,RNA编辑效应被认为是"低的"或"降低的"。

[0560] 本公开进一步提供了施用所公开的TadA-CD碱基编辑器的方法,其中该方法产生降低的和/或低RNA编辑效应。本公开还提供了诱导(或产生、提供或引起)降低的和/或低RNA编辑效应的腺嘌呤碱基编辑器。在一些实施方案中,碱基编辑器在细胞mRNA转录物中提供0.3%或更低的平均胞苷(C)至胸腺嘧啶(T)(C至T)编辑频率。在一些实施方案中,碱基编辑器在RNA中提供约0.3%或更低的平均胞苷(A)至胸腺嘧啶(T)(C至T)实际和/或一致的编辑频率。碱基编辑器可以在RNA中提供以下实际或平均C至T编辑频率:约0.5%或更低、0.4%或更低、0.35%或更低、0.25%或更低、0.2%或更低、0.15%或更低、0.13%或更低、0.1%或更低、0.08%或更低或0.075%或更低。在特定实施方案中,碱基编辑器提供约0.25%的平均C至T编辑频率。在特定实施方案中,碱基编辑器TadA-CDa、TadA-CDb和TadA-CDc诱导小于或等于0.1%(检测限)的平均C至T编辑频率。其他碱基编辑器变体,例如TadA-CDd和TadA-CDe,在一些实施方案中可以分别诱导约0.3%和0.2%的平均C至T编辑频率。无论变体类型如何,掺入V106W置换可将所有TadA-CD变体的脱靶RNA编辑减少到低于或等于0.13%。

[0561] 在一些实施方案中,这些方法诱导(或提供或引起)人细胞(例如HEK293细胞)的mRNA转录组间约0.3%或更低的平均胞苷(C)至胸腺嘧啶(T)编辑频率。这些方法可以在RNA中诱导以下实际或平均C至T转录组范围的编辑频率:约0.5%或更低、0.4%或更低、0.35%或更低、0.25%或更低、0.2%或更低、0.15%或更低、0.13%或更低、0.1%或更低、0.08%或更低或0.075%或更低。在特定实施方案中,所公开的方法(例如涉及使用TadCBEd或TadCBEe编辑器编辑的方法)诱导0.3%或0.2%的人mRNA转录组范围的平均C至T编辑频率。

[0562] 降低的脱靶DNA编辑效应

[0563] 通过包括将增加DNA特异性的突变安装到碱基编辑器的Cas9组分中、在sgRNA中添

加5’鸟苷核苷酸或将碱基编辑器作为核糖核蛋白复合物(RNP)递送等策略,已经降低了向导RNA依赖性脱靶碱基编辑。向导RNA非依赖性脱靶编辑可能由于碱基编辑器的脱氨酶结构域以Cas9非依赖性方式与C或A碱基结合而产生。以下实施例证实,本文公开的进化TadA-CD变体确实表现出可检测的向导RNA非依赖性脱靶DNA突变。然而,本文提供的一些经进化的TadA-CD变体,如TadA-CDa(V106W)至TadACDe(V106W),相对于TadA-CDa至TadACDe表现出降低的Cas9非依赖性脱靶DNA突变。所公开的胞嘧啶碱基编辑器的脱靶效应可以使用正交R环测定来测量,如国际申请号PCT/US2020/624628(2020年11月25日提交)中所公开的,其通过引用并入本文。

[0564] 在一些方面,本文提供了产生(或引起)降低的脱靶效应的胞嘧啶碱基编辑器和方法,该方法通过将DNA与任何这些公开的碱基编辑器接触来编辑DNA。在各种实施方案中,设计了用于确定napDNAbp结构域非依赖性(例如Cas9非依赖性)(或gRNA非依赖性)脱靶编辑事件的脱靶编辑频率的方法。编辑事件可以包括TadCBE的脱氨事件。依赖于napDNAbp-向导RNA复合物的脱靶脱氨事件通常在与靶序列具有高序列同一性(例如大于60%序列同一性)的序列中。这些类型的事件是由于napDNAbp-向导RNA复合物与与靶序列共享同一性的序列的不完全杂交而产生的。相比之下,不依赖于napDNAbp-向导RNA复合物而发生的脱靶事件是由于碱基编辑器与DNA序列(通常是与靶序列不共享高序列同一性的序列)的随机结合而产生的,这是由于碱基编辑器的核苷酸修饰结构域(例如脱氨酶结构域)的碱基编辑器与DNA的内在亲和力。napDNAbp非依赖性(例如Cas9非依赖性)编辑事件可能特别发生在当碱基编辑器在被评估的系统(如细胞或受试者)中过表达时。

[0565] 在一些实施方案中,所公开的TadCBE表现出脱靶编辑频率(例如A>G编辑)。在一些情况下,腺嘌呤在编辑窗口内的位置可能会影响脱靶编辑频率。在一些情况下,将腺嘌呤放置在编辑窗口的中心会增加脱靶编辑频率。例如,在HEK293T细胞中编辑PDCD1靶位点导致TadA-CDb和Tad-CDc分别34%或36%的腺嘌呤碱基编辑,而TadCD-d最多11%的编辑。

[0566] 在一些实施方案中,在本文公开的碱基编辑器内包括V106W突变可以改进(例如降低)脱靶编辑频率。将V106W添加到TadA-CD中,减少TadA-CDb(V106W)的A>G编辑最多12%,并且减少TadA-CDd(V106W)的编辑最多5%(两者均在PDCD1处观察到最大值)。

[0567] 所公开的TadCBE表现出低脱靶编辑频率,特别是低Cas9非依赖性脱靶编辑频率,同时表现出高中靶编辑效率。例如,基于TadA-CDa(V106W)的变体在哺乳动物细胞中的靶序列中可以表现出0.38%或更低的平均脱靶编辑频率,同时保持约80%或更高的中靶编辑效率。作为另一个非限制性实施方案,基于TadA-CDb(V106W)的变体在哺乳动物细胞中的靶序列中可以表现出约0.62%或更低的平均脱靶编辑频率,同时保持约80%或更高的中靶编辑效率。其他示例性实施方案可以包括变体TadA-CDc、TadA-CDd和TadA-CDe,它们在哺乳动物细胞中的靶序列中可以分别表现出0.48%或更低、1.1%或更低和0.05%或更低的平均脱靶编辑频率,同时保持约80%或更高的中靶编辑效率。在其他实施方案中,基于TadA-CD-V106W的变体(例如TadA-CDa-e)在人细胞中的靶序列中可以表现出0.68%或更低的插入缺失频率和/或5%或更低的平均脱靶编辑频率,同时保持80%的中靶编辑效率。(参见图5。)这些脱靶编辑频率可能低于几种现有的胞苷脱氨酶,如rAPOBEC1、evoAPOBEC1(evoA)、evoFERNY和YE1。任何所公开的TadCBE表现出的Cas依赖性脱靶编辑可以与BE4max和EvoA-BE4max表现出的水平相似。

[0568] 在一些实施方案中,TadA-CD对胞嘧啶与腺嘌呤脱氨的选择性在超过10,000个靶位点上的平均值范围从低的11倍偏好胞嘧啶脱氨(例如对于TadA-CDb)到高的27倍(例如对于TadA-CDd)。对于V106W变体,这种选择性进一步增强,从低的20倍(例如对于TadA-CDb(V106W))到高的48倍(例如对于TadA-CDd(V106W))。这些超过10,000个靶基因组位点可以位于小鼠胚胎干细胞或人胚胎干细胞中。

[0569] 在一些实施方案中,所公开的胞苷脱氨酶在使用各种Cas同源物和其他napDNAbp时表现出低脱靶编辑频率,特别是Cas9非依赖性低脱靶编辑频率,同时表现出高中靶编辑效率。在一些实施方案中,TadA-CD脱氨酶或TadA-CDd(V106W)脱氨酶在使用各种napDNAbp(如SpCas9、SaCas9和SaKKH-Cas9)时,在哺乳动物细胞中的靶序列中可以表现出0.32%或更低的脱靶编辑频率,同时保持约80%或更高的中靶编辑效率。

[0570] 在一些实施方案中,所公开的碱基编辑器引起低于1.5%(例如低于1.25%、低于1.0%、低于0.75%或低于0.5%)的脱靶DNA编辑(例如脱靶脱氨)频率。在一些情况下,脱靶编辑频率低于1.5%、1.25%、低于1.1%、低于1%、低于0.75%、低于0.5%、低于0.4%、低于0.25%、低于0.2%、低于0.15%、低于0.1%或0.05%或更低。

[0571] 在一些实施方案中,所公开的TadCBE和包括将DNA与任何公开的TadCBE接触的步骤的编辑方法导致以下实际或平均脱靶DNA编辑频率:约2.0%或更低、1.75%或更低、1.5%或更低、1.2%或更低、1%或更低、0.9%或更低、0.8%或更低、0.75%或更低、0.7%或更低、0.65%或更低或0.6%或更低。在一些实施方案中,所公开的编辑方法导致以下实际或平均脱靶DNA编辑频率:0.5%、低于0.5%、低于0.4%、低于0.35%、低于0.3%、低于0.25%、低于0.2%或低于0.1%。在特定实施方案中,这些方法导致约0.32%至1.3%的实际或平均脱靶DNA编辑频率(例如,用于评估包含TadA-CD-V106W脱氨酶的TadCBE的脱靶频率的方法)。这些脱靶编辑频率可以在与靶序列具有任何水平序列同一性的序列中获得。如本文用于指代脱靶DNA编辑频率的修饰语“平均”指的是在给定靶碱基对以外的位点处检测到的所有编辑事件的平均值(例如通过高通量测序所检测)。

[0572] 在一些实施方案中,所公开的编辑方法进一步导致以下实际或平均Cas9非依赖性脱靶DNA编辑频率:约2.0%或更低、1.75%或更低、1.5%或更低、1.2%或更低、1%或更低、0.9%或更低、0.8%或更低、0.75%或更低、0.7%或更低、0.65%或更低或0.6%或更低。换句话说,所公开的编辑方法进一步导致在与靶序列具有60%或更低序列同一性的序列中以下实际或平均脱靶DNA编辑频率:约2.0%或更低、1.75%或更低、1.5%或更低、1.2%或更低、1%或更低、0.9%或更低、0.8%或更低、0.75%或更低、0.7%或更低、0.65%或更低或0.6%或更低。在一些实施方案中,所公开的编辑方法导致在与靶序列具有60%或更低序列同一性的序列中以下实际或平均脱靶DNA编辑频率:0.5%、低于0.5%、低于0.4%、低于0.35%、低于0.3%、低于0.25%、低于0.2%或低于0.1%。在一些实施方案中,这些编辑频率是在包含相对于靶序列的原间隔区序列有5、6、7、8、9、10个或更多个错配的原间隔区序列的序列中获得的。在特定实施方案中,这些方法导致0.4%或更低的实际或平均Cas9非依赖性脱靶DNA编辑频率。

[0573] 在各种实施方案中,所公开的编辑方法导致约25:1、50:1、65:1、75:1、80:1、85:1、90:1、95:1、100:1、110:1、125:1或超过125:1的中靶:脱靶编辑的比率。在各种实施方案中,公开的编辑方法导致约150:1、200:1、300:1、400:1、500:1、600:1、700:1、800:1、900:1、

1000:1、1100:1、1200:1、1250:1、1275:1、1300:1、1325:1、1350:1、1400:1、1500:1或超过1500:1的中靶:脱靶编辑的比率。如本文所用,中靶:脱靶编辑的比率等同于反映中靶脱氨的测序读段相对于已知或预测的脱靶位点或候选脱靶位点的脱氨的测序读段的比率。候选脱靶位点可以使用实验测定或计算算法(例如Cas-OFFinder)鉴定,从而可以测量中靶:脱靶编辑的比率。例如,候选脱靶位点可以使用实验测定(如EndoV-Seq、GUIDE-Seq或CIRCLE-Seq)鉴定。

[0574] 在一些实施方案中,所公开的编辑方法导致CXCR4或CCR5基因中约25:1、50:1、65:1、75:1、80:1、85:1、90:1、95:1、100:1、110:1、125:1或超过125:1的中靶:脱靶编辑的比率。在各种实施方案中,所公开的编辑方法导致CXCR4或CCR5基因中约150:1、200:1、300:1、400:1、500:1、600:1、700:1、800:1、900:1、1000:1、1100:1、1200:1、1250:1、1275:1、1300:1、1325:1、1350:1、1400:1、1500:1或超过1500:1的中靶:脱靶编辑的比率。在一些情况下,CXCR4或CCR5基因中的中靶:脱靶编辑的比率为约90:1或更高。

[0575] 在一些实施方案中,所公开的编辑方法导致等同于预期点突变:非预期点突变的比率的中靶:脱靶编辑的比率。在一些实施方案中,所公开的编辑方法导致至少25:1、至少30:1、至少40:1、至少50:1、至少75:1、至少90:1、至少100:1、至少150:1、至少200:1、至少250:1、至少500:1、至少1000:1、至少1100:1、至少1200:1、至少1250:1、至少1300:1、至少1350:1、至少1400:1、至少1500:1或更高的预期点突变与非预期点突变的比率。

[0576] 在一些实施方案中,所公开的编辑方法导致并且所公开的碱基编辑器生成非常低程度的旁观者编辑(即在靶碱基附近且不改变预期编辑方法的结果的核碱基处的同义脱靶点突变)。在一些实施方案中,所公开的编辑方法导致少于10个、少于9个、少于8个、少于7个、少于6个、少于5个、少于4个、少于3个、少于2个、少于1个或零个非沉默旁观者编辑。

[0577] 降低的插入缺失频率

[0578] 本公开的一些方面基于以下认识:本文提供的任何胞嘧啶碱基编辑器都能够在不生成显著比例的插入缺失(indel)的情况下修饰特定的DNA碱基。如本文所用,“插入缺失”指的是DNA底物内核苷酸碱基的插入或缺失。此类插入或缺失可以导致基因编码区内的移码突变。在一些实施方案中,期望生成高效修饰(例如突变或脱氨)DNA内的特定核苷酸,而不在核酸中生成大量插入或缺失(即插入缺失)(同时具有比现有腺嘌呤碱基编辑器更低的RNA编辑效应)的胞嘧啶碱基编辑器。

[0579] 在某些实施方案中,本文提供的任何胞嘧啶碱基编辑器能够生成比插入缺失更大比例的预期修饰(例如点突变或脱氨)。在一些实施方案中,本文提供的碱基编辑器能够生成大于1:1的预期点突变与插入缺失的比率。在一些实施方案中,本文提供的碱基编辑器能够生成至少1.5:1、至少2:1、至少2.5:1、至少3:1、至少3.5:1、至少4:1、至少4.5:1、至少5:1、至少5.5:1、至少6:1、至少6.5:1、至少7:1、至少7.5:1、至少8:1、至少10:1、至少12:1、至少15:1、至少20:1、至少25:1、至少30:1、至少40:1、至少50:1、至少100:1、至少200:1、至少300:1、至少400:1、至少500:1、至少600:1、至少700:1、至少800:1、至少900:1或至少1000:1或更高的预期点突变与插入缺失的比率。预期突变和插入缺失的数量可以使用任何合适的方法确定,例如以下实施例中使用的方法。在一些实施方案中,插入缺失频率对应于靶序列处含有插入缺失的总测序读段的百分比。因此,为了计算插入缺失频率,扫描测序读段以寻找与可能发生插入缺失的窗口两侧侧翼的对两个10-bp序列的精确匹配。如果未找到精确

匹配,则该读段被排除在分析之外。如果该插入缺失窗口的长度与参考序列精确匹配,则该读段被分类为不含有插入缺失。如果插入缺失窗口比参考序列长或短两个或更多个碱基,则该测序读段分别被分类为插入或缺失。

[0580] 在一些实施方案中,本文提供的胞嘧啶碱基编辑器能够限制DNA底物区域中插入缺失的形成。在一些实施方案中,该区域是位于碱基编辑器靶向的核苷酸处或碱基编辑器靶向的核苷酸的2、3、4、5、6、7、8、9或10个核苷酸内的区域。在一些实施方案中,本文提供的任何碱基编辑器可以以低于1%、低于1.5%、低于2%、低于2.5%、低于2.8%、低于3%、低于3.5%、低于4%、低于4.5%、低于5%、低于6%、低于7%、低于8%、低于9%、低于10%、低于12%、低于15%或低于20%的频率在核酸区域诱导插入缺失的形成。在一些实施方案中,本文提供的任何碱基编辑器在与包含靶序列的核酸接触时可以诱导或生成低于20%、19%、18%、16%、14%、12%、10%、8%、6%、4%、3%、2%、1%、0.5%、0.2%、0.1%或0.05%的插入缺失形成。在核酸区域形成的插入缺失数量可以取决于核酸(例如细胞基因组内的核酸)暴露于碱基编辑器的时间量。在一些实施方案中,插入缺失的数量或比例是在将核酸(例如细胞基因组内的核酸)暴露于胞嘧啶碱基编辑器至少1小时、至少2小时、至少6小时、至少12小时、至少24小时、至少36小时、至少48小时、至少3天、至少4天、至少5天、至少7天、至少10天或至少14天后确定的,诸如通过转染编码编辑器的载体。在一些实施方案中,插入缺失频率在3天后确定。

[0581] 在一些实施方案中,TadA-CDa-e碱基编辑器可以在包含靶序列的核酸区域以低于1.1%的频率诱导插入缺失形成(参见图29A)。在一些实施方案中,使用所公开的TadCBE的所公开的编辑方法可以导致在包含靶序列的核酸(例如DNA)中低于20%、19%、18%、16%、14%、12%、10%、8%、6%、4%、2%、1.5%、1%、0.5%、0.2%或0.1%的插入缺失形成。

[0582] 本公开的某些方面基于以下认识:本文提供的任何碱基编辑器能够有效地在DNA(例如受试者基因组内的DNA)中生成预期突变(如点突变),而不生成大量非预期突变(如非预期点突变)。在一些实施方案中,预期突变是由与gRNA结合的特定碱基编辑器生成的突变,该gRNA专门设计用于生成预期突变(例如脱氨)。在一些实施方案中,预期突变是与疾病或病症相关的突变,例如镰状细胞病或HIV/AIDS。在一些实施方案中,预期突变是与疾病或病症相关的腺嘌呤(A)至鸟嘌呤(G)点突变。在一些实施方案中,预期突变是与疾病或病症相关的胸腺嘧啶(T)至胞嘧啶(C)点突变。在一些实施方案中,预期突变是基因编码区内的腺嘌呤(A)至鸟嘌呤(G)点突变。在一些实施方案中,预期突变是基因编码区内的胸腺嘧啶(T)至胞嘧啶(C)点突变。

[0583] 在一些实施方案中,预期突变是生成终止密码子的脱氨,例如基因编码区内的提前终止密码子。在一些实施方案中,预期突变是消除终止密码子的突变。在一些实施方案中,预期突变消除了包含核酸序列5'-TAG-3'、5'-TAA-3'或5'-TGA-3'的终止密码子。

[0584] 在一些实施方案中,预期突变是改变基因调控序列(例如基因启动子或基因阻遏物)的脱氨。在一些实施方案中,预期突变是引入基因启动子的脱氨。在特定实施方案中,引入基因启动子的脱氨导致与基因启动子可操作连接的基因的转录的减少。在其他实施方案中,脱氨导致与基因启动子可操作连接的基因的转录的增加。

[0585] 在一些实施方案中,预期突变是改变基因剪接的脱氨。因此,在一些实施方案中,预期脱氨导致基因中剪接位点的引入。在其他实施方案中,预期脱氨导致剪接位点的移除。

[0586] 在一些实施方案中,本文提供的任何碱基编辑器能够生成大于1:1的预期突变与非预期突变(例如预期点突变:非预期点突变)的比率。在一些实施方案中,本文提供的任何碱基编辑器能够生成至少1.5:1、至少2:1、至少2.5:1、至少3:1、至少3.5:1、至少4:1、至少4.5:1、至少5:1、至少5.5:1、至少6:1、至少6.5:1、至少7:1、至少7.5:1、至少8:1、至少10:1、至少12:1、至少15:1、至少20:1、至少25:1、至少30:1、至少40:1、至少50:1、至少100:1、至少150:1、至少200:1、至少250:1、至少500:1或至少1000:1或更高的预期突变与非预期突变(例如预期点突变:非预期点突变)的比率。应当理解,本节和下节中描述的碱基编辑器的特征可以应用于本文提供的任何碱基编辑器或使用碱基编辑器的方法。

[0587] 用于生成胞嘧啶碱基编辑器的方法

[0588] 本发明还涉及通过各种操作模式生成公开的胞嘧啶碱基编辑器的方法,这些操作模式包括但不限于通过密码子优化在细胞中实现更高的表达水平,以及使用核定位序列(NLS)、优选至少两个NLS(例如两个二分NLS)以增加表达的碱基编辑器在细胞核中的定位。

[0589] 本文考虑的碱基编辑器可以包括例如通过密码子优化导致增加的表达的修饰。

[0590] 在一些实施方案中,编码任何公开的碱基编辑器(或其组分)的多核苷酸针对在特定细胞(如真核细胞(例如人细胞))中的表达进行密码子优化。真核细胞可以是衍生自特定生物体的那些,特定生物体如哺乳动物,包括但不限于人、小鼠、大鼠、兔子、狗或非人灵长类动物。通常,密码子优化指的是通过将天然序列中的至少一个密码子(例如约或超过约1、2、3、4、5、10、15、20、25、50个或更多个密码子)替换为该宿主细胞基因中更频繁或最频繁使用的密码子来修饰核酸序列用于感兴趣宿主细胞中增强的表达,同时保持天然氨基酸序列。各种物种对特定氨基酸的某些密码子表现出特定的偏好。密码子偏好(不同生物体之间密码子使用的差异)通常与信使RNA(mRNA)的翻译效率相关,而翻译效率据信又取决于被翻译的密码子的特性和特定转运RNA(tRNA)分子的可用性等因素。细胞中选定tRNA的优势通常反映了在肽合成中最频繁使用的密码子。因此,可以根据密码子优化为给定生物体中的最佳基因表达定制基因。密码子使用表很容易获得,例如在“密码子使用数据库(Codon Usage Database)”中,并且这些表可以通过多种方式进行调整。参见Nakamura,Y.,et al.“Codon usage tabulated from the international DNA sequence databases:status for the year 2000”Nucl.Acids Res.28:292(2000)。用于密码子优化特定序列以在特定宿主细胞中表达的计算机算法也是可用的,如Gene Forge(Aptagen;Jacobus,Pa.)也是可用的。在一些实施方案中,编码CRISPR酶序列中的一个或多个密码子(例如1、2、3、4、5、10、15、20、25、50个或更多个或所有密码子)对应于特定氨基酸的最常用密码子。

[0591] 上述描述旨在对于生成具有增加的表达的碱基编辑器是非限制性的,并且从而增加编辑效率。

[0592] 定向进化方法(例如PACE或PANCE)

[0593] 本公开的各种实施方案涉及提供用于工程化本公开的碱基编辑器或碱基编辑器结构域的定向进化方法和系统(例如适当的载体、细胞、噬菌体、流动容器等)。本公开提供了用于所公开的定向进化方法的载体系统,以工程化任何公开的碱基编辑器或碱基编辑器结构域(例如任何公开的碱基编辑器的经进化的腺苷脱氨酶结构域)。

[0594] 本文提供的定向进化载体系统和方法允许在宿主细胞流中通过多代病毒生命周期进化病毒载体中的感兴趣基因(例如碱基编辑器编码基因或腺苷脱氨酶编码基因),以获

得期望的功能或活性。

[0595] 本公开的一些实施方案提供了噬菌体辅助连续进化(PACE)方法,包括:(a)将细菌宿主细胞群体与包含待进化的感兴趣基因且缺乏感染性噬菌体生成所需的基因的噬菌体群体接触,其中(1)噬菌体允许在宿主细胞中表达感兴趣基因;(2)宿主细胞是适合噬菌体感染、复制和包装的宿主细胞;以及(3)宿主细胞包含编码生成感染性噬菌体所需基因的表达构建体,其中该基因的表达依赖于感兴趣基因的基因产物的功能。在一些实施方案中,该方法还包括(b)在允许感兴趣基因突变、感染性噬菌体产生以及宿主细胞被噬菌体感染的条件下孵育宿主细胞群体,其中受感染的细胞从宿主细胞群体中移除,并且其中宿主细胞群体被尚未被噬菌体感染的新鲜宿主细胞补充。在一些实施方案中,该方法还包括(c)从宿主细胞群体中分离编码经进化的蛋白质的突变噬菌体复制产物。

[0596] 在PACE中,选择下的基因在M13噬菌体基因组上编码。其活性通过控制基因III的表达与M13繁殖相关联,使得只有活性变体产生感染性后代噬菌体。使噬菌体持续繁殖和诱变,但突变仅在噬菌体基因组中积累,而不在宿主或其选择回路中积累,因为新鲜宿主细胞不断流入(和流出)生长容器,有效地重置了选择背景。

[0597] PANCE/PACE进化回路的开发

[0598] PACE能够通过每天多代的突变、选择和复制实现生物分子的快速连续进化。在PACE期间,宿主大肠杆菌细胞持续稀释含有感兴趣基因的噬菌体群体(选择噬菌体,SP)。感兴趣基因替换了SP上的基因III,该基因是后代噬菌体感染性所必需的。含有期望基因变体的SP触发来自辅助质粒(AP)的宿主细胞基因III表达。宿主细胞DNA质粒编码遗传回路,将SP中编码的蛋白质的期望活性与AP上基因III的表达联系起来。因此,含有期望基因变体的SP变体可以繁殖,而编码非活性变体的噬菌体不会生成感染性后代,并迅速从培养容器(或培养池)中稀释出去。阿拉伯糖诱导的诱变质粒(MP)控制噬菌体突变率。

[0599] 新PACE选择的关键是将基因III的表达与感兴趣的活性联系起来。设计了低严格性选择,其中碱基编辑激活T7 RNA聚合酶,后者转录gIII。单个编辑事件可以在转录编辑后的DNA后立即导致高输出扩增。参考国际专利公布WO 2019/023680,2019年1月31日公布;Badran,A.H.&Liu,D.R.In vivo continuous directed evolution.Curr.Opin.Chem.Biol.24,1-10(2015);Dickinson,B.C.,Packer,M.S.,Badran,A.H.&Liu,D.R.A system for the continuous directed evolution of proteases rapidly reveals drug-resistance mutations.Nat.Commun.5,5352(2014);Hubbard,B.P.et al.Continuous directed evolution of DNA-binding proteins to improve TALEN specificity.Nat.Methods 12,939-942(2015);Wang,T.,Badran,A.H.,Huang,T.P.&Liu,D.R.Continuous directed evolution of proteins with improved soluble expression.Nat.Chem.Biol.14,972-980(2018),以及Thuronyi,B.W.et al.Continuous evolution of base editors with expanded target compatibility and improved activity.Nat.Biotechnol.,1070-1079(2019),其中每篇通过引用并入本文。

[0600] 还参考了使用BE-HIVE工具生成的高通量哺乳动物DNA碱基编辑器文库,其用于评估腺嘌呤碱基编辑器的编辑活性和编辑窗口。BE-HIVE模型在国际申请号PCT/US2021/016924(其于2021年8月12日公布为公布号WO/2021/158995)中有更详细的描述。

[0601] 本公开提供了用于对胞嘧啶碱基编辑器的腺苷脱氨酶结构域进行定向进化的载

体系统。在一些实施方案中,该载体系统包含表达构建体,该构建体包含编码分裂内含肽部分(例如,分裂内含肽的N端部分或C端部分)的核酸,该核酸与编码产生感染性噬菌体颗粒所需的基因(如gIII蛋白(pIII蛋白)或其部分(例如片段))的核酸可操作地连接。在一些实施方案中,分裂内含肽部分是分裂内含肽的C端部分(例如,Npu(点形念珠藻)分裂内含肽的C端部分)。在一些实施方案中,分裂内含肽C端部分定位于编码产生感染性噬菌体颗粒所需的基因或其部分的核酸的上游(例如,相对于核酸的5')。在一些实施方案中,分裂内含肽部分是分裂内含肽的N端部分(例如,Npu分裂内含肽的N端部分)。在一些实施方案中,分裂内含肽N端部分定位于编码产生感染性噬菌体颗粒所需的基因或其部分的核酸的下游(例如,相对于核酸的3')。在一些实施方案中,所公开的载体系统表达构建体(例如,在第一辅助质粒或第二辅助质粒中)还包含编码luxAB的序列。

[0602] 在一些实施方案中,本文所述的载体系统包含:(i)选择质粒,其包含分离的核酸,该核酸包含编码经进化的腺苷脱氨酶的表达构建体,其按以下顺序包含:腺苷脱氨酶蛋白和编码分裂内含肽N端部分的序列;(ii)第一辅助质粒,其包含分离的核酸,该核酸包含按以下顺序的表达构建体:由Lac启动子可操作控制的编码向导RNA的序列和由T7 RNA启动子可操作控制的编码M13噬菌体gIII蛋白信号肽的序列,其中编码gIII蛋白信号肽的序列缺少信号肽结构域的一个或多个核酸碱基;(iii)第二辅助质粒,其包含分离的核酸,该核酸包含按以下顺序的表达构建体:编码分裂内含肽C端部分的序列和编码dCas9的序列;以及(iv)第三辅助质粒,其包含分离的核酸,该核酸包含按以下顺序的表达构建体:启动子、核糖体结合位点和编码T7 RNA聚合酶(RNAP)的序列,该RNAP包含导致可通过腺嘌呤碱基编辑校正的两个终止密码子的突变(参见图1B和18)。在一些实施方案中,分裂内含肽是Npu分裂内含肽。在一些实施方案中,这些终止密码子在位置57和58处创建。在一些实施方案中,腺嘌呤碱基编辑校正T7 RNAP编码区中位置57和58处的突变,并诱导置换回野生型Q57和R58(参见图1C)。在某些实施方案中,所公开的载体系统还包含多个第三辅助质粒,每个包含独特的核糖体结合位点或独特的启动子。可以开发具有不同的启动子和核糖体结合位点(RBS)的多达五、六、七、八、九或十个第三辅助质粒的变体,以调节PACE进化的负严格性。在某些实施方案中,载体系统还包含诱变质粒。

[0603] 在一些实施方案中,载体系统作为试剂盒的一部分提供,该试剂盒在一些实施方案中可用于执行PACE以产生腺苷脱氨酶蛋白变体。例如,在一些实施方案中,试剂盒包含容纳载体系统的选择噬菌粒的第一容器、容纳载体系统的第一辅助质粒的第二容器以及容纳载体系统的第二辅助质粒的第三容器。在一些实施方案中,试剂盒还包含诱变质粒。用于PACE的诱变质粒通常是现有技术已知的,例如描述于国际PCT申请号PCT/US2016/027795(2016年9月16日提交,公布为WO 2016/168631)中,其全部内容通过引用并入本文。在一些实施方案中,试剂盒还包含一套用于执行PACE的书面或电子说明书。

[0604] 在本文提供的定向进化方法和系统的一些实施方案中,病毒载体或噬菌体是丝状噬菌体,例如M13噬菌体,如本文其他地方详细描述的M13选择噬菌体。在一些此类实施方案中,产生感染性病毒颗粒所需的基因是M13基因III(gIII)。

[0605] 在一些实施方案中,宿主细胞的孵育持续足以进行至少10、20、30、40、50、100、200、300、400、500、600、700、800、900、1000、1250、1500、1750、2000、2500、3000、4000、5000、7500、10000个或更多个连续的病毒生命周期的时间。在某些实施方案中,病毒载体是M13噬

菌体,并且单个病毒生命周期的长度为约10-20分钟。

[0606] 在一些实施方案中,选择病毒载体/宿主细胞组合,其中病毒载体的生命周期显著短于宿主细胞的细胞分裂之间的平均时间。许多细胞类型和载体的平均细胞分裂时间和病毒载体生命周期时间在本领域中是众所周知的,使得本领域技术人员能够确定此类宿主细胞/载体组合。在某些实施方案中,宿主细胞以以下速率从与病毒载体接触的宿主细胞群体中移除:该速率导致宿主细胞在被移除之前在宿主细胞群体中停留的平均时间短于宿主细胞的细胞分裂之间的平均时间,但长于所使用的病毒载体的平均生命周期。其结果是宿主细胞平均在其位于宿主细胞群体的时间期间没有足够的时间繁殖,而病毒载体则在宿主细胞停留在细胞群体期间有足够的时间感染宿主细胞、在宿主细胞中复制并生成新的病毒颗粒。这确保了宿主细胞群体中唯一复制的核酸是病毒载体,并且宿主细胞基因组、辅助质粒或任何其他核酸构建体无法获得允许逃避施加的选择压力的突变。

[0607] 例如,在一些实施方案中,宿主细胞在宿主细胞群体中停留的平均时间为约10、约11、约12、约13、约14、约15、约16、约17、约18、约19、约20、约21、约22、约23、约24、约25、约30、约35、约40、约45、约50、约55、约60、约70、约80、约90、约100、约120、约150或约180分钟。

[0608] 在一些实施方案中,宿主细胞在宿主细胞群体中停留的平均时间取决于宿主细胞分裂有多快以及感染(或接合)所需有多长。通常,流速应快于细胞分裂所需的平均时间,但足够慢以允许病毒性(或接合性)繁殖。前者会例如因培养基类型而变化,并且可以通过添加细胞分裂抑制剂抗生素(如大肠杆菌中的FtsZ抑制剂等)来延迟。由于连续进化中的限制步骤是细胞至细胞的基因转移所需蛋白质的产生,载体被洗出的流速将取决于感兴趣基因的当前活性。在一些实施方案中,如本文所述的感染性颗粒生成所需的蛋白质的可滴定产生可以缓解这一问题。在一些实施方案中,噬菌体感染的指示器允许实时的当前活性水平的流速的计算机控制的优化。

[0609] 在一些实施方案中,新鲜宿主细胞包含病毒载体选择所需的辅助质粒,例如,包含正在进化的噬菌体所缺少的生成感染性噬菌体颗粒所需基因的辅助质粒。在一些实施方案中,宿主细胞是通过将未感染的宿主细胞与相关载体(例如辅助质粒和任选地诱变质粒)接触,并生长足够用于补充连续进化实验中的宿主细胞群体的宿主细胞的数量而生成的。将质粒和其他基因构建体引入宿主细胞的方法为本领域技术人员所熟知,并且本发明在这方面不受限制。对于细菌宿主细胞,此类方法包括但不限于感受态细胞的电穿孔和热休克。

[0610] 在一些实施方案中,辅助质粒包含选择标志物,例如抗生素抗性标志物,并且新鲜宿主细胞在相应抗生素存在下生长以确保质粒存在于宿主细胞中。当存在多个质粒时,通常使用不同的标志物。此类选择标志物及其在细胞培养中的用途为本领域技术人员所知,并且本发明在这方面不受限制。

[0611] 在特定实施方案中,第一辅助质粒包含基因III,并且第二辅助质粒包含由G至T突变失活的T7 RNAP基因,该突变导致早期终止密码子。第三辅助质粒可以包含编码dCas9的核苷酸,其N端与快速剪接内含肽的C端半部融合。示例性噬菌体质粒可以包含编码腺苷脱氨酶的核苷酸,其C端与快速剪接内含肽的N端半部融合。全长碱基编辑器由两个内含肽组分重组而成。

[0612] 在一些实施方案中,选择标志物是壮观霉素抗生素抗性标志物。在其他实施方案中,选择标志物是氯霉素或羧苄青霉素抗性标志物。细胞可以用含有失活的壮观霉素抗性

基因的选择质粒转化,该基因在活性位点具有需要通过A:T至C:G的编辑来校正的突变。未能安装壮观霉素抗性基因中的正确颠换突变的细胞将死亡,而做出校正的细胞将存活。表达靶向壮观霉素抗性基因活性位点突变的sgRNA和核苷酸修饰结构域-dCas9碱基编辑器的大肠杆菌细胞铺板于含有256μg/mL壮观霉素的2xYT琼脂上。对存活的菌落(通过CFU测量)进行测序,以找到在经进化的存活细胞中表达的碱基编辑器的共有突变。类似的筛选测定用于在腺嘌呤碱基编辑器开发过程中进化DNA中的腺苷脱氨酶活性,如Gaudelli,N.M.et al.,Programmable base editing of A•T to G•C in genomic DNAwithout DNA cleavage.Nature 551,464-471(2017)中所述,其通过引用以其整体并入本文。

[0613] 在一些实施方案中,连续进化实验中的宿主细胞群体由在平行连续培养中生长的新鲜宿主细胞补充。在一些实施方案中,与病毒载体接触的宿主细胞群体中宿主细胞的细胞密度和新鲜宿主细胞群体的密度基本相同。

[0614] 通常,从与病毒载体接触的细胞群体中移除的细胞包括用病毒载体感染的细胞和未感染的细胞。在一些实施方案中,细胞从细胞群体中连续移除,例如通过实现细胞从群体中的连续流出。在其他实施方案中,细胞从群体中半连续性地或间歇性地移除。在一些实施方案中,新鲜细胞的补充将与细胞从细胞群体的移除模式相匹配,例如,如果细胞连续移除,新鲜细胞将连续引入。然而,在一些实施方案中,补充和移除的模式可能不匹配,例如,细胞群体可能用新鲜细胞连续补充,并且细胞可能半连续或以批次方式移除。

[0615] 在一些实施方案中,新鲜宿主细胞补充的速率和/或宿主细胞移除的速率基于对细胞群体中宿主细胞的量化进行调整。例如,在一些实施方案中,监测包含宿主细胞群体的培养基的浊度,并且如果浊度低于阈值水平,则调整宿主细胞流入与宿主细胞流出的比率,以实现群体中的宿主细胞数量的增加,如增加的细胞培养浊度所呈现的。在其他实施方案中,如果浊度高于阈值水平,则调整宿主细胞流入与宿主细胞流出的比率,以实现群体中宿主细胞数量的减少,如减少的细胞培养浊度所呈现的。将宿主细胞群体中的细胞密度维持在特定密度范围内,确保足够宿主细胞可用作正在进化的病毒载体群体的宿主,并避免因营养物的耗竭而影响病毒包装以及来自于培养过度拥挤的细胞来源的毒素的积累。

[0616] 在一些实施方案中,宿主细胞群体中的细胞密度和/或流入的新鲜宿主细胞密度为约102细胞/ml至约1012细胞/ml。在一些实施方案中,宿主细胞密度为约102细胞/ml、约103细胞/ml、约104细胞/ml、约105细胞/ml、约5•105细胞/ml、约106细胞/ml、约5•106细胞/ml、约107细胞/ml、约5•107细胞/ml、约108细胞/ml、约5•108细胞/ml、约109细胞/ml、约5•109细胞/ml、约1010细胞/ml或约5•1010细胞/ml。在一些实施方案中,宿主细胞密度超过约1010细胞/ml。

[0617] 在一些实施方案中,宿主细胞群体与诱变剂接触。在一些实施方案中,与病毒载体(例如噬菌体)接触的细胞群体持续暴露于诱变剂,其浓度允许感兴趣基因的增加的突变率,但在宿主细胞暴露于诱变剂期间不会对宿主细胞产生显著毒性。在其他实施方案中,宿主细胞群体间歇性地与诱变剂接触,产生增加诱变并因此增加病毒载体的多样化的阶段。例如,在一些实施方案中,宿主细胞暴露于足以在约10%、约20%、约50%或约75%的时间内生成感兴趣基因的增加的诱变率的诱变剂浓度。

[0618] 在一些实施方案中,宿主细胞包含诱变表达构建体,例如在细菌宿主细胞的情况下,包含诱变质粒。在一些实施方案中,诱变质粒包含编码促进诱变的基因产物(例如校对

缺陷的DNA聚合酶)的基因表达盒。在其他实施方案中,诱变质粒包括参与SOS应激反应的基因(例如UmuC、UmuD'和/或RecA)。在一些实施方案中,促进诱变的基因受诱导型启动子控制。合适的诱导型启动子为本领域技术人员所熟知,并且包括例如阿拉伯糖诱导型启动子、四环素或多西环素诱导型启动子以及他莫昔芬诱导型启动子。在一些实施方案中,宿主细胞群体与诱导型启动子的诱导剂接触,其量足以实现增加的诱变率。例如,在一些实施方案中,提供细菌宿主细胞群体,其中宿主细胞包含诱变质粒,其中dnaQ926、UmuC、UmuD'和RecA表达盒受阿拉伯糖诱导型启动子控制。在一些此类实施方案中,宿主细胞群体与诱导剂(例如阿拉伯糖)接触,其量足以诱导增加的突变率。

[0619] 在一些实施方案中,通过提供不选择感兴趣基因中的功能获得突变的宿主细胞流来实现病毒载体群体的多样化,该功能获得突变用于病毒载体群体的复制、诱变和繁殖。在一些实施方案中,宿主细胞是表达生成感染性病毒颗粒所需的所有基因的宿主细胞,例如表达完整辅助噬菌体的细菌细胞,并且因此不对感兴趣基因施加选择压力。在其他实施方案中,宿主细胞包含辅助质粒,该质粒包含条件启动子,具有足以在即使没有感兴趣基因的显著功能获得突变的情况下也能支持病毒载体的繁殖的基线活性。这可以通过使用"泄漏"条件启动子、通过使用高拷贝数辅助质粒(从而放大基线泄漏)和/或通过使用感兴趣基因的初始版本实现低活性水平而期望功能获得突变实现显著更高活性的条件启动子来实现。

[0620] 使用噬菌体颗粒在宿主细胞群体中进行碱基编辑器连续定向进化的详细方法公开于以下中:国际PCT申请,PCT/US2009/056194,2009年9月8日提交,2010年3月11日公布为WO 2010/028347;国际PCT申请,PCT/US2011/066747,2011年12月22日提交,2012年6月28日公布为WO 2012/088381;美国专利号9,023,594,2015年5月5日发布;美国专利号9,771,574,2017年9月26日发布;美国专利号9,394,537,2016年7月19日发布;国际PCT申请,PCT/US2015/012022,2015年1月20日提交,2015年9月11日公布为WO 2015/134121;美国专利号10,179,911,2019年1月15日发布;国际申请号PCT/US2019/37216,2019年12月19日公布为WO 2019/241649,国际专利公布WO 2019/023680,2019年1月31日公布,国际PCT申请,PCT/US2016/027795,2016年4月15日提交,2016年10月20日公布为WO 2016/168631,以及国际公布号WO 2020/041751,2020年2月27日公布,其中每篇通过引用并入本文。

[0621] 设计适用于执行本文所述选择策略的条件启动子的方法和策略为本领域技术人员所熟知。有关示例性合适选择策略和驱动细胞间基因转移所需基因(例如基因III(gIII))表达的条件启动子设计方法的概述,参见Vidal and Legrain,Yeast n-hybrid review,Nucleic Acid Res.27,919(1999),其以其整体并入本文。

[0622] 本公开提供了用于连续进化过程的载体。在一些实施方案中,提供了用于噬菌体辅助连续进化的噬菌体载体。在一些实施方案中,提供了选择噬菌体,其包含缺失产生感染性噬菌体颗粒生成所需的至少一个基因的噬菌体基因组以及待进化的感兴趣基因。参考2019年1月31日公布的国际专利公布WO 2019/023680,其通过引用并入本文。

[0623] 例如,在一些实施方案中,选择噬菌体包含缺失感染性M13噬菌体颗粒生成所需基因(例如全长gIII)的M13噬菌体基因组。在一些实施方案中,选择噬菌体包含提供噬菌体生命周期所需的所有其他噬菌体功能的噬菌体基因组,但感染性噬菌体颗粒生成所需的基因除外。在一些此类实施方案中,提供了M13选择噬菌体,其包含gI、gII、gIV、gV、gVI、gVII、gVIII、gIX和gX基因,但不包含全长gIII。在一些实施方案中,选择噬菌体包含gIII的3'片

段,但不包含全长gIII。gIII的3'末端包含启动子并且保留该启动子活性在一些实施方案中有益于gVI(其紧邻gIII 3'启动子下游)的增加的表达,或有益于宿主细胞中噬菌体基因的表达水平的更平衡(类似于野生型噬菌体)比率,其转而可以导致更高效的噬菌体生产。在一些实施方案中,gIII基因的3'片段包含3'-gIII启动子序列。在一些实施方案中,gIII的3'片段包含gIII的最后180bp、最后150bp、最后125bp、最后100bp、最后50bp或最后25bp。在一些实施方案中,gIII的3'片段包含gIII的最后180bp。

[0624] 提供了M13选择噬菌体,其包含噬菌体基因组中(例如插入在gVIII 3'终止子下游和gIII-3'启动子上游)的感兴趣基因。在一些实施方案中,提供了M13选择噬菌体,其包含用于将感兴趣基因克隆到噬菌体基因组中的多克隆位点,例如插入在gVIII 3'终止子下游和gIII-3'启动子上游的多克隆位点(MCS)。

[0625] 本发明的一些实施方案提供了用于连续进化程序的载体系统,包括病毒载体(例如选择噬菌体)和匹配的辅助质粒。在一些实施方案中,提供了用于基于噬菌体的连续定向进化的载体系统,其包括:(a)包含待进化的感兴趣基因的选择噬菌体,其中噬菌体基因组缺失生成感染性噬菌体所需的基因;以及(b)辅助质粒,其包含由条件启动子控制的生成感染性噬菌体颗粒所需的基因,其中条件启动子由感兴趣基因编码的基因产物的功能激活。

[0626] 在一些实施方案中,选择噬菌体是本文所述的M13噬菌体。例如,在一些实施方案中,选择噬菌体包含M13基因组,包括生成噬菌体颗粒所需的所有基因,例如gI、gII、gIV、gV、gVI、gVII、gVIII、gIX和gX基因,但不包括全长gIII基因。在一些实施方案中,选择噬菌体基因组包含F1或M13复制起点。在一些实施方案中,选择噬菌体基因组包含gIII基因的3'片段。在一些实施方案中,选择噬菌体包含位于gIII 3'启动子上游和gVIII 3'终止子下游的多克隆位点。

[0627] 在示例性的PACE方法中,每个含有诱变质粒的宿主细胞被稀释到含有适当的抗生素的5mL Davis Rich Medium(DRM)中,并生长至A600为0.4-0.8。然后将细胞用于接种到恒化器(60mL)中,恒化器可以保持在用新鲜DRM以1-1.5体积/小时连续稀释下以保持细胞密度大致恒定。在接种噬菌体之前,培养池最初充满DRM,然后用恒化器培养连续稀释至少2小时。可以在噬菌体添加前1小时将阿拉伯糖的储备溶液(1M)直接泵入培养池(最终浓度为10mM)。在噬菌体接种后的前12小时,脱水四环素(3.3μg/mL)存在于储备溶液中。培养池以约107pfu/mL的起始滴度接种。稀释速率可以通过调节培养池体积(5-20mL)和/或培养流入速率(10-20mL/h)来调整。每24小时通过注射器移除培养液(500μL)对培养池进行采样。样品在13,500g下离心2分钟,移除上清液并储存在4℃。通过噬斑形成(plaquing)评估滴度。通过含有pT7-AP和不含质粒的S2060细胞上的噬斑形成监测T7 RNAP或基因III重组噬菌体的存在。通过诊断PCR从单一噬斑中评估噬菌体基因型。

[0628] 本公开的一些实施方案提供了的感兴趣基因的非连续进化的方法。在某些实施方案中,非连续进化的方法是PANCE。在其他实施方案中,非连续进化的方法是基于抗生素或板的选择方法。PANCE使用与PACE相同的遗传回路来激活噬菌体繁殖,但不同于连续稀释容器,噬菌体通过用先前传代的等分试样感染新鲜宿主细胞培养物来手动传代。因为在选择过程中几乎没有丢失弱活性噬菌体变体的风险,并且因为噬菌体稀释的有效速率要低得多,PANCE不如PACE严格。

[0629] 示例性的PANCE方法包括首先在含有0.5%葡萄糖(w/v)和适当浓度抗生素的2xYT

琼脂上生长含有大肠杆菌的诱变质粒的宿主株,直到大体积光密度达到A600＝0.5-0.6。用诱变质粒定期重新转化细胞以确保质粒未被失活。然后将期望浓度的等分试样(通常为2mL)转移到较小的烧瓶中,补充40mM诱导剂阿拉伯糖(Ara)用于诱变质粒,并用选择噬菌体(SP)感染。为了增加滴度水平,还可以提供漂移质粒,其使噬菌体能够在未通过选择的情况下繁殖。表达受诱导型启动子控制,并且可以用0-40ng/mL的脱水四环素开启。处理后的培养物可以分成期望数量的单培养管中的2mL培养物或96孔板中的500μL培养物,并用选择噬菌体感染(参见图19)。这些培养物在37℃下孵育8-12小时以促进噬菌体生长,生长通过噬菌体滴度测定确认,然后收获。噬菌体生长后,使用感染细胞的等分试样转染含有宿主大肠杆菌的后续烧瓶。含有经进化的噬菌体的上清液可以分离并储存在4℃。该过程可以持续进行,直到期望的表型进化到所需的转移次数,同时通过减少细菌感染的噬菌体的孵育时间或滴度逐步增加严格性。。在本文提供的示例性PANCE方案中,该过程在25次培养传代中迭代进行。参考Suzuki T.et al.,Crystal structures reveal an elusive functional domain of pyrrolysyl-tRNA synthetase,Nat Chem Biol.13(12):1261-1266(2017),以其整体并入本文。

[0630] 在一些实施方案中,在本文所述的非连续进化方法期间,通过惩罚不期望的活性来应用负选择。在一些实施方案中,这是通过使不期望的活性干扰pIII的产生来实现的。例如,表达与gIII RBS和/或起始密码子互补的反义RNA是应用负选择的一种方法,而表达蛋白酶(例如TEV)并将蛋白酶识别位点工程化到pIII中是另一种方法。

[0631] 用于具有期望活性的基因产物的其他非连续选择方案为本领域技术人员所熟知,或从本公开中显而易见。在某些实施方案中,在成功定向进化腺嘌呤碱基编辑器的一个或多个组分(例如Cas9结构域或腺苷脱氨酶结构域)后,制备碱基编辑器的方法包括为本领域普通技术人员所熟知的重组蛋白表达方法。

[0632] 在一些实施方案中,PACE/PANCE方法包括:(a)编码与NpunN内含肽融合的突变TadA8e蛋白的选择噬菌体,(b)编码与dCas9-UGI融合的NpuC内含肽的第一质粒,(c)编码由T7或proT7启动子驱动的gIII并编码sgRNA的第二质粒,以及(d)编码T7 RNA聚合酶-降解决定子融合物的第三质粒。在某些情况下,T7 RNA聚合酶-降解决定子融合物在T7 RNA聚合酶和降解决定子结构域之间的界面处含有靶序列。靶序列可以包含一个或多个胞嘧啶核苷酸,其在被编辑为尿嘧啶时,会在T7 RNA聚合酶-降解决定子融合物的T7 RNA聚合酶和降解决定子结构域之间插入终止密码子。

[0633] 在一些实施方案中,本文所述的启动子可以是强启动子,使进化回路的严格性更低。替代性地或另外地,在一些实施方案中,本文所述的启动子可以是弱启动子,从而使进化回路的严格性更高。

[0634] 选择回路的开发

[0635] 本公开的各种实施方案涉及提供用于调节选择严格性的定向进化方法。本公开提供了用于所公开的定向进化方法的选择回路,以工程化任何公开的碱基编辑器或碱基编辑器结构域(例如任何公开的碱基编辑器的胞苷脱氨酶结构域)。本文所述的选择回路允许调节对残留腺苷脱氨活性的耐受性。

[0636] 选择回路使用本文别处所述的PACE/PANCE方法进行。正在进化的感兴趣蛋白(例如TadA-8e)在选择噬菌体(SP)上编码,该噬菌体感染大肠杆菌宿主细胞[46]。大肠杆菌含有持

续诱变噬菌体基因组的诱变质粒(MP)以及调节基因III的表达的辅助质粒(AP),该基因III编码pIII(噬菌体复制的关键蛋白)。由于gIII已从SP基因组中移除,仅编码具期望活性的正在进化的变体的噬菌体触发大肠杆菌中pIII的产生并复制,从而导致活性基因变体(例如具有胞苷脱氨酶活性的突变体TadA-8e)的繁殖。在持续诱变和稀释(例如使用PACE)下,缺乏期望活性的噬菌体迅速从选择容器("培养池")中稀释,而进化出有益突变的噬菌体则持续存在。

[0637] 在一些实施方案中,TadA-8e脱氨酶在SP内编码,并且宿主大肠杆菌细胞含有:(i)MP,(ii)编码SpCas9的辅助质粒,(iii)与C端(3'末端)降解决定子标签融合的自失活T7 RNA聚合酶(T7 RNAP),以及(iv)由T7 RNAP转录控制的基因III。噬菌体感染后,SP编码的脱氨酶通过反式内含肽剪接与Cas9连接以重建碱基编辑器。为了激活选择回路,碱基编辑器必须进行C·G至T·A以在T7 RNAP和降解决定子之间创建终止密码子,从而产生有活性的T7 RNAP。无降解决定子的T7 RNAP然后转录gIII,导致噬菌体繁殖。

[0638] 在一些实施方案中,选择回路允许在进化早期阶段可能发生的相对于腺苷脱氨对胞苷脱氨的更低选择性。为了实现这一点,使用原间隔区序列$\underline{C_6}A_7A_8$对非模板链(非编码链)进行编辑,原间隔区序列被编辑为$\underline{T_6}A_7A_8$以在胞苷脱氨时引入终止密码子(例如引入TAA终止密码子的碱基编辑以及无降解决定子的T7 RNAP的表达)。在原间隔区位置7或位置8的脱靶腺苷脱氨不会阻止密码子的安装(例如导致TGA或TAG终止密码子),除非A_7和A_8都转化为G(导致编码Trp的密码子TGG),这使得该回路耐受适度水平的腺苷脱氨,因此更适合早期阶段的TadA8e进化(回路1)(参见图1D)。在一些实施方案中,非编码链包含原间隔区序列$T_4\underline{C_6}A_7A_8G_9$。

[0639] 因此,在一些方面,本文提供了载体系统,包括:选择质粒,其包含编码腺苷脱氨酶的分离的核酸,其按以下顺序包含:腺苷脱氨酶蛋白和编码分裂内含肽的N端部分的序列;第一辅助质粒,其按以下顺序包含:由Lac启动子可操作控制的编码向导RNA的序列和由T7 RNA启动子可操作控制的编码M13噬菌体基因III(gIII)肽的序列;第二辅助质粒,其按以下顺序包含:编码分裂内含肽C端部分的序列和编码dCas9-UGI融合物的序列;以及第三辅助质粒,其包含非编码链和编码链,其中编码链包含表达构建体,表达构建体按以下顺序包含:启动子、核糖体结合位点以及编码T7 RNA聚合酶和降解决定子的序列,其中与编码T7 RNA聚合酶的3'末端相对的非编码链包含CAA序列。在特定实施方案中,CAA序列是原间隔区序列$\underline{C_6}A_7A_8$,如原间隔区序列$T_4\underline{C_6}A_7A_8G_9$。在一些实施方案中,腺苷脱氨酶是TadA-8e。在一些实施方案中,分裂内含肽是Npu(点形念珠藻)内含肽。

[0640] 在一些实施方案中,选择回路允许相对于腺苷脱氨对胞苷脱氨的更高选择性。为了实现这一点,根据一些实施方案,模板链(编码链)包含原间隔区序列$A_6\underline{C_5}\underline{C_4}$,当在$C_5$和/或$C_4$处胞苷脱氨时产生$A_6\underline{T_5}\underline{T_4}$、$A_6\underline{T_5}\underline{C_4}$或$A_6\underline{C_5}\underline{T_4}$,并引入终止密码子(TAA、TAG、TGA),从而表达无降解决定子的T7RNAP。然而,该选择回路对甚至单一腺苷脱氨也不耐受,因为这导致$\underline{G_6}\underline{T_5}\underline{T_4}$、$\underline{G_6}\underline{T_5}\underline{C_4}$或$\underline{G_6}\underline{C_5}\underline{T_4}$的原间隔区序列,对应于非终止密码子CAA、CAG和CGA。因此,该回路对选择腺苷脱氨更为严格(回路2)(参见图1D)。因此,在一些实施方案中,提供了第三辅助质粒,其包含非编码链和编码链,其中编码链包含表达构建体,表达构建体按以下顺序包含:启动子、核糖体结合位点和编码T7 RNA聚合酶和降解决定子的序列,其中编码链在编码T7 RNA聚合酶序列3'末端包含ACC序列(例如原间隔区序列$A_6\underline{C_5}\underline{C_4}$)。

[0641] 在一些实施方案中,回路2中待进化的蛋白质是回路1的产物。换句话说,回路1可用于获得一组对腺苷碱基和胞苷碱基具有特异性和活性的经进化的TadA-8e脱氨酶。这些经进化的脱氨酶可以进一步使用回路2进行进化,以筛选出具有残留腺苷特异性和活性的变体,从而产生对胞嘧啶碱基具有高特异性的胞苷脱氨酶。

[0642] 在一些实施方案中,选择回路包含编码融合至NpuN内含肽的突变TadA8e蛋白的选择噬菌体。在某些情况下,第一质粒编码融合至dCas9-UGI的NpuC内含肽,第二质粒编码由T7或proT7启动子驱动的gIII并编码sgRNA。在某些实施方案中,第三质粒编码T7 RNA聚合酶-降解决定子融合物。在一些实施方案中,T7 RNA聚合酶-降解决定子融合物在T7 RNA聚合酶和降解决定子结构域之间的界面处含有靶序列。在一些情况下,靶序列可以含有一个或多个胞嘧啶核苷酸,其当被编辑为胸腺嘧啶时,会在T7 RNA聚合酶-降解决定子融合物的T7 RNA聚合酶和降解决定子结构域之间插入终止密码子。

[0643] 载体

[0644] 本公开的制备和使用碱基编辑器的多个方面涉及包含编码胞嘧啶碱基编辑器的一个或多个载体的载体系统(例如,包含编码胞嘧啶碱基编辑器的多核苷酸的载体)。载体可以设计用于克隆和/或表达本发明的胞嘧啶碱基编辑器。载体还可以设计用于将本公开的经进化的腺嘌呤碱基编辑器转染到一个或多个细胞中,例如用于使用本文所述的碱基编辑器系统和方法治疗的靶标病变真核细胞。在一些情况下,载体可以包含编码RNA(例如向导RNA)的多核苷酸。

[0645] 载体可以设计用于在原核或真核细胞中表达碱基编辑器转录物(例如核酸转录物、蛋白质或酶)。例如,碱基编辑器转录物可以在细菌细胞(如大肠杆菌)、昆虫细胞(使用杆状病毒表达载体)、酵母细胞、植物细胞或哺乳动物细胞中表达。合适的宿主细胞在Goeddel,Gene Expression Technology:Methods In Enzymology 185,Academic Press.San Diego,Calif.(1990)中进一步讨论。替代性地,编码本文所述的一个或多个经进化的腺嘌呤碱基编辑器的表达载体可以在体外转录和翻译,例如使用T7启动子调控序列和T7聚合酶。在一些实施方案中,载体包含驱动编码碱基编辑器的多核苷酸表达的异源启动子。编码本文提供的胞嘧啶碱基编辑器的载体可以包含以下任何鉴定的DNA载体:TadCBEa-eNme2-C-BE4max、TadCBEa-enCjCas9-BE4max、TadCBEa-SpCas9-BE4max、TadCBEa-SaCas9-BE4max、TadCBEa-SpCas9-NG-BE4max。这些载体如下提供。

[0646] 本公开的示例性载体包含任何碱基编辑器编码载体,如SEQ ID NO:100-104所示。在一些实施方案中,所公开的载体包含与SEQ ID NO:100-104中任一个具有至少80%、至少85%、至少90%、至少95%、至少98%、至少99%或至少99.5%序列同一性的核酸序列。在一些实施方案中,本文所述的任何载体可以包含相对于SEQ ID NO:100-104中任一个的序列具有1-5、5-10、10-15、15-20、20-25、25-30、30-35、35-40、40-45、45-50或超过50个核苷酸差异的核酸序列。这些差异可以包括相对于SEQ ID NO:100-104中任一个插入、缺失或置换的核苷酸。在一些实施方案中,所公开的载体含有与SEQ ID NO:100-104中任一个具有约50、约75、约100、约125、约150、约175、约200、约300、约400、约500个或超过500个连续核苷酸相同的片段。

[0647] TadCBEa-eNme2-C-BE4max载体

[0648] tagttattaatagtaatcaattacggggtcattagttcatagcccatatatggagttccgcgttacat

aacttacggtaaatggcccgcctggctgaccgcccaacgacccccgcccattgacgtcaataatgacgtatgttcc
catagtaacgccaatagggactttccattgacgtcaatgggtggagtatttacggtaaactgcccacttggcagta
catcaagtgtatcatatgccaagtacgccccctattgacgtcaatgacggtaaatggcccgcctggcattatgccc
agtacatgaccttatgggactttcctacttggcagtacatctacgtattagtcatcgctattaccatggtgatgcg
gttttggcagtacatcaatgggcgtggatagcggtttgactcacggggatttccaagtctccaccccattgacgtc
aatgggagtttgttttggcaccaaaatcaacgggactttccaaaatgtcgtaacaactccgccccattgacgcaaa
tgggcggtaggcgtgtacggtgggaggtctatataagcagagctggtttagtgaaccgtcagatccgctagagatc
cgcggccgctaatacgactcactatagggagagccgccaccatgaaacggacagccgacggaagcgagttcgagtc
accaaagaagaagcggaaagtcagttctgaggtggagttttcccacgagtactggatgagacatgccctgaccctg
gccaagagggcacgggatgaggggcgggtcctgtgggagccgtgctggtgctgaacaatagagtgatcggcgagg
gctggaacagagccatcggcctgcacgacccaacagcccatgccgaaattatggccctgagacagggcggcctggt
catgcagaactacagactgtttgacgccaccctgtacgtgacattcgagccttgcgtgatgtgcgccggcgccatg
atcaactctaggatcggccgcgtggtgtttggcgtgaggaactcaaaaagaggcgccgcaggctccctgatgaacg
tgctgaactaccccggcatgaatcaccgcgtcgaaattaccgagggaatcctggcagatgaatgtgccgccctgct
gtgcgatttctatcggatgcctagacaggtgttcaattctcagaagaaggcccagagctccatcaactctggcgga
tctagcggaggatcctctggcagcgagacaccaggaacaagcgagtcagcaacaccagagagcagtggcggcagca
gcggcggcagcgcagcattcaagcctaacccaatcaattacatcctgggactggcaatcggaatcgcatccgtggg
atgggctatggtggagatcgacgaggaggggaatcctatccggctgatcgatctgggcgtgagagtgtttgagagg
gccgaggtgccaaagaccggcgattctctggctatggcccggagactggcacggagcgtgaggcgcctgacacgga
gaagggcacacaggctgctgagggcacgccggctgctgaagagagagggcgtgctgcaggcagcagacttcgatga
gaatggcctgatcacgagcttgccaaacacccccctggcagctgagagcagccgccctggacaggaagctgacacca
ctggagtggtctgccgtgctgctgcacctgatcaagcaccgcggctacctgagccagcggaagaacgagggagaga
cagcagccaaggagctgggcgccctgctgaagggagtggccaacaatgcccacgccctgcagaccggcgatttcag
gacacctgccgagctggccctgaataagtttgagaaggagtccggccacatcagaaaccagaggggcgactatagc
cacaccttctcccgcaaggatctgcaggccgagctgatcctgctgttcgagaagcagaaggagtttggcaatccac
acgtgagcggaggcctgaaggagggaatcgagaccctgctgatgacacagaggcctgccctgtccggcgacgcagt
gcagaagatgctggggcactgcaccctcgagcctacagagccaaaggccgccaagaacacctacacagccgagcgg
tttatctggctgacaaagctgaacaatctgagaatcctggagcagggatccgagaggccactgaccgacacagaga
ggtccaccctgatggatgagccttaccggaagtctaaactgacatatgcccaggccagaaagctgctgggcctgga
ggacaccgccttctttaagggcctgagatacggcaaggataatgccgaggcctccacactgatggagatgaaggcc
tatcacgccatctctcgcgccctggagaaggagggcctgaaggacaagaagtccccccctgaacctgagctccgagc
tgcaggatgagatcggcaccgccttctctctgtttaagaccgacgaggatatcacaggccgcctgaaggacagggt
gcagcctgagatcctggaggccctgctgaagcacatctctttcgataagtttgtgcagatcagcctgaaggccctg
agaaggatcgtgccactgatggagcagggcaagcggtacgacgaggcctgcgccgagatctacggcgttcactatg
gcaagaagaacacagaggagaagatctatctgccccctatccctgccgacgagatcagaaatcctgtggtgctgag
ggccctgtcccaggcaagaaaagtgatcaacggagtggtgcgccggtacggatctccagcccggatccacatcgag
accgccagagaagtgggcaagagcttcaaggaccggaaggagatcgcgaagagacaggaggagaatcgcaaggatc
gggagaaggccgccgccaagtttagggagtacttccctaactttgtgggcgagccaaagtctaaggacatcctgaa
gctgcgcctgtacgagcagcagcacggcaagtgtctgtatagcggcaaagagatcaatctggtgcggctgaacgag

aagggctatgtggagatcgatcacgccctgcctttctccagaacctgggacgattcttttaacaataaggtgctgg
tgctgggcagcgagaaccagaataagggcaatcagacaccatacgagtatttcaatggcaaggacaactccaggga
gtggcaggagttcaaggcccgcgtggagacctctagatttcccagtagcaagaagcagcggatcctgctgcagaag
ttcgacgaggatggctttaaggagtgcaacctgaatgacaccagatacgtgaaccggttcctgtgccagtttgtgg
ccgatcacatcctgctgaccggcaagggcaagagaagggtggtcgcctctaatggccagatcacaaacctgctgag
ggggttttggagactgaggaaggtgcgggcagagaatgacagacaccacgcactggatgcagtggtggtggcatgc
agcaccgtggcaatgcagcagaagatcacaagattcgtgaggtataaggagatgaacgcctttgacggcaagaccg
tcgataaggagacaggcaaggtgctgtaccagaagacccacttcccccagccttgggagttctttgcccaggaagt
tatgatccgggtgttcggcaagccagacggcaagcctgagtttgaggaggccgataccccagagaagctgaggaca
ctgctggcagagaagctgtctagcaggccagaggcagtgcacgagtacgtgaccccgctgttcgtgtccagggcac
ccaatcggaagatgtctggcgcccacaaggacacactgagaagcgccaagaggtttgtgaagcacaacgagaagat
ctccgtgaagagagtgtggctgaccgagatcaagctggccgatctggagaacatggtgaattacaagaacggcagg
gagatcgagctgtatgaggccctgaaggcaaggctggaggcctacggaggaaatgccaagcaggccttcgacccaa
aggataacccctttttataagaagggaggacagctggtgaaggccgtgcgggtggagaagacccagaaaagcggcgt
gctgctgaataagaagaacgcctacacaatcgccgacaatggtgatatggtgagagtggacgtgttctgtaaggtg
gataagaagggcaagaatcagtactttatcgtgcctatctatgcctggcaggtggccgagaacatcctgccagaca
tcgattgcaagggctacagaatcgacgatagctatacattctgtttttccctgcacaagtatgacctgatcgcctt
ccagaaggatgagaagtccaaggtggagtttgcctactatatcaattgcgactcctctagcggcgggttctacctg
gcctggcacgataagggcagcagggagcagcggtttcgcatctccacccagaatctggcgctgatccagaagtatc
aggtgaacgagctgggcaaggagatcaggccatgtcggctgaagaagcgcccacccgtgagcggcgggagcggcgg
gagcgggggagcactaatctgagcgacatcattgagaaggagactgggaaacagctggtcattcaggagtccatc
ctgatgctgcctgaggaggtggaggaagtgatcggcaacaagccagagtctgacatcctggtgcacaccgcctacg
acgagtccacagatgagaatgtgatgctgctgacctctgacgcccccgagtataagccttgggccctggtcatcca
ggattctaacggcgagaataagatcaagatgctgagcggaggatccggaggatctggaggcagcaccaacctgtct
gacatcatcgagaaggagacaggcaagcagctggtcatccaggagagcatcctgatgctgcccgaagaagtcgaag
aagtgatcggaaacaagcctgagagcgatatcctggtccataccgcctacgacgagagtaccgacgaaaatgtgat
gctgctgacatccgacgccccagagtataagccctgggctctggtcatccaggattccaacggagagaacaaaatc
aaaatgctgtctggcggctcaaaaagaaccgccgacggcagcgaattcgagcccaagaagaagaggaaagtctaac
cggtcatcatcaccatcaccattgagtttaaacccgctgatcagcctcgactgtgccttctagttgccagccatct
gttgtttgcccctcccccgtgccttccttgaccctggaaggtgccactcccactgtcctttcctaataaaatgagg
aaattgcatcgcattgtctgagtaggtgtcattctattctgggggggtggggtggggcaggacagcaaggggaggga
ttgggaagacaatagcaggcatgctggggatgcggtgggctctatggcttctgaggcggaaagaaccagctggggc
tcgataccgtcgacctctagctagagcttggcgtaatcatggtcatagctgtttcctgtgtgaaattgttatccgc
tcacaattccacacaacatacgagccggaagcataaagtgtaaagcctagggtgcctaatgagtgagctaactcac
attaattgcgttgcgctcactgcccgctttccagtcgggaaacctgtcgtgccagctgcattaatgaatcggccaa
cgcgcggggagaggcggtttgcgtattgggcgcacttccgcttcctcgctcactgactcgctgcgctcggtcgttc
ggctgcggcgagcggtatcagctcactcaaaggcggtaatacggttatccacagaatcaggggataacgcaggaaa
gaacatgtgagcaaaaggccagcaaaaggccaggaaccgtaaaaaggccgcgttgctggcgttttccataggctc
cgccccccctgacgagcatcacaaaaatcgacgctcaagtcagaggtggcgaaacccgacaggactataaagatacc

aggcgtttccccctggaagctccctcgtgcgctctcctgttccgaccctgccgcttaccggatacctgtccgcctt
tctcccttcgggaagcgtggcgctttctcatagctcacgctgtaggtatctcagttcggtgtaggtcgttcgctcc
aagctgggctgtgtgcacgaacccccgttcagcccgaccgctgcgccttatccggtaactatcgtcttgagtcca
acccggtaagacacgacttatcgccactggcagcagccactggtaacaggattagcagagcgaggtatgtaggcgg
tgctacagagttcttgaagtggtggcctaactacggctacactagaagaacagtatttggtatctgcgctctgctg
aagccagttaccttcggaaaaagagttggtagctcttgatccggcaaacaaaccaccgctggtagcggtggttttt
ttgtttgcaagcagcagattacgcgcagaaaaaaaggatctcaagaagatcctttgatcttttctacggggtctga
cactcagtggaacgaaaactcacgttaagggattttggtcatgagattatcaaaaaggatcttcacctagatcctt
ttaaattaaaaatgaagttttaaatcaatctaaagtatatatgagtaaacttggtctgacagttaccaatgcttaa
tcagtgaggcacctatctcagcgatctgtctatttcgttcatccatagttgcctgactccccgtcgtgtagataac
tacgatacgggagggcttaccatctggccccagtgctgcaatgataccgcgggacccacgctcaccggctccagat
ttatcagcaataaaccagccagccggaagggccgagcgcagaagtggtcctgcaactttatccgcctccatccagt
ctattaattgttgccgggaagctagagtaagtagttcgccagttaatagtttgcgcaacgttgttgccattgctac
aggcatcgtggtgtcacgctcgtcgtttggtatggcttcattcagctccggttcccaacgatcaaggcgagttaca
tgatcccccatgttgtgcaaaaaagcggttagctccttcggtcctccgatcgttgtcagaagtaagttggccgcag
tgttatcactcatggttatggcagcactgcataattctcttactgtcatgccatccgtaagatgcttttctgtgac
tggtgagtactcaaccaagtcattctgagaatagtgtatgcggcgaccgagttgctcttgcccggcgtcaatacgg
gataataccgcgccacatagcagaactttaaaagtgctcatcattggaaaacgttcttcggggcgaaaactctcaa
ggatcttaccgctgttgagatccagttcgatgtaacccactcgtgcacccaactgatcttcagcatcttttacttt
caccagcgtttctgggtgagcaaaaacaggaaggcaaaatgccgcaaaaaagggaataagggcgacacggaaatgt
tgaatactcatactcttccttttcaatattattgaagcatttatcagggttattgtctcatgagcggatacatat
ttgaatgtatttagaaaaataaacaaatagggttccgcgcacatttccccgaaaagtgccacctgacgtcgacgg
atcgggagatcgatctcccgatcccctagggtcgactctcagtacaatctgctctgatgccgcatagttaagccag
tatctgctccctgcttgtgtgttggaggtcgctgagtagtgcgcgagcaaaatttaagctacaacaaggcaaggct
tgaccgacaattgcatgaagaatctgcttagggttaggcgttttgcgctgcttcgcgatgtacgggccagatatac
gcgttgacattgattattgac(SEQ ID NO:100)

[0649] TadCBEa-enCjCas9-BE4max载体

[0650] attgatgcagtaatcatagcctacgctaacaacagcatcgtgaaagccttctccgatttcaagaaaga
acaggagtctaatagcgccgagttgtacgccaagaaaatttccgaattggactataaaaataagagaaaattcttc
gaacccttctccgggtttcgccaaaaggtcttagataagatcgacgagattttcgtttccaagcccgaaagaaaaa
agccttcaggggcactgcacgaagagacattccgcaaggaagaggaattttaccaatcttacggtggtaaagaggg
agttctgaaggctctggagcttgggaagatccgcaaggtaaacgggaaaatcgtgaaaaacggggacatgttcagg
gtggatatcttcaagcacaaaaagaccaacaagttctacgcagtacccatctacactatggatttcgctttaaagg
ttctcccaaataaggcggtggctcgatcgaagaaaggagagatcaaggactggatcttaatggatgaaaattacga
gttttgcttctcgctctacaaagatagcctgattctgatccagacaaaaaagatgcaggaaccagaatttgtttat
tataacgccttcacgagcagtacagtgtccctgattgtgagcaagcatgataacaagttcgagactctgtctaaga
atcagaaaatccttttcaagaacgccaacgagaaggaggtcatcgcaaagtcaattggcatccaaaacctgaaggt
gttcgagaaatacatagtgtccgcactcggtgaagtaactaaagccgaatttcgacagcgcgaggattttaagaaa
agcggcgggagcggcgggagcgggggggagcactaatctgagcgacatcattgagaaggagactgggaaacagctgg

tcattcaggagtccatcctgatgctgcctgaggaggtggaggaagtgatcggcaacaagccagagtctgacatcct
ggtgcacaccgcctacgacgagtccacagatgagaatgtgatgctgctgacctctgacgcccccgagtataagcct
tgggccctggtcatccaggattctaacggcgagaataagatcaagatgctgagcggaggatccaaaagaaccgccg
acggcagcgaattcgagcccaagaagaagaggaaagtctaaccggtcatcatcaccatcaccattgagtttaaacc
cgctgatcagcctcgactgtgccttctagttgccagccatctgttgtttgcccctcccccgtgccttccttgaccc
tggaaggtgccactcccactgtcctttcctaataaaatgaggaaattgcatcgcattgtctgagtaggtgtcattc
tattctgggggtggggtggggcaggacagcaaggggaggattgggaagacaatagcaggcatgctggggatgcg
gtgggctctatggcttctgaggcggaaagaaccagctggggctcgataccgtcgacctctagctagagcttggcgt
aatcatggtcatagctgtttcctgtgtgaaattgttatccgctcacaattccacacaacatacgagccggaagcat
aaagtgtaaagcctagggtgcctaatgagtgagctaactcacattaattgcgttgcgctcactgcccgctttccag
tcgggaaacctgtcgtgccagctgcattaatgaatcggccaacgcgcggggagaggcggtttgcgtattgggcgca
cttccgcttcctcgctcactgactcgctgcgctcggtcgttcggctgcggcgagcggtatcagctcactcaaaggc
ggtaatacggttatccacagaatcaggggataacgcaggaaagaacatgtgagcaaaaggccagcaaaaggccagg
aaccgtaaaaaggccgcgttgctggcgtttttccataggctccgccccctgacgagcatcacaaaaatcgacgct
caagtcagaggtggcgaaacccgacaggactataaagataccaggcgtttccccctggaagctccctcgtgcgctc
tcctgttccgaccctgccgcttaccggatacctgtccgcctttctcccttcgggaagcgtggcgctttctcatagc
tcacgctgtaggtatctcagttcggtgtaggtcgttcgctccaagctgggctgtgtgcacgaacccccccgttcagc
ccgaccgctgcgccttatccggtaactatcgtcttgagtccaacccggtaagacacgacttatcgccactggcagc
agccactggtaacaggattagcagagcgaggtatgtaggcggtgctacagagttcttgaagtggtggcctaactac
ggctacactagaagaacagtatttggtatctgcgctctgctgaagccagttaccttcggaaaaagagttggtagct
cttgatccggcaaacaaaccaccgctggtagcggtggtttttttgtttgcaagcagcagattacgcgcagaaaaaa
aggatctcaagaagatcctttgatcttttctacggggtctgacactcagtggaacgaaaactcacgttaagggatt
ttggtcatgagattatcaaaaaggatcttcacctagatccttttaaattaaaaatgaagttttaaatcaatctaaa
gtatatatgagtaaacttggtctgacagttaccaatgcttaatcagtgaggcacctatctcagcgatctgtctatt
tcgttcatccatagttgcctgactccccgtcgtgtagataactacgatacgggagggcttaccatctggccccagt
gctgcaatgataccgcgggacccacgctcaccggctccagatttatcagcaataaaccagccagccggaagggccg
agcgcagaagtggtcctgcaactttatccgcctccatccagtctattaattgttgccgggaagctagagtaagtag
ttcgccagttaatagtttgcgcaacgttgttgccattgctacaggcatcgtggtgtcacgctcgtcgtttggtatg
gcttcattcagctccggttcccaacgatcaaggcgagttacatgatcccccatgttgtgcaaaaaagcggttagct
ccttcggtcctccgatcgttgtcagaagtaagttggccgcagtgttatcactcatggttatggcagcactgcataa
ttctcttactgtcatgccatccgtaagatgcttttctgtgactggtgagtactcaaccaagtcattctgagaatag
tgtatgcggcgaccgagttgctcttgcccggcgtcaatacgggataataccgcgccacatagcagaactttaaaag
tgctcatcattggaaaacgttcttcggggcgaaaactctcaaggatcttaccgctgttgagatccagttcgatgta
acccactcgtgcacccaactgatcttcagcatcttttactttcaccagcgtttctgggtgagcaaaaacaggaagg
caaaatgccgcaaaaaagggaataagggcgacacggaaatgttgaatactcatactcttccttttcaatattatt
gaagcatttatcagggttattgtctcatgagcggatacatatttgaatgtatttagaaaaataaacaaatagggGT
tccgcgcacatttccccgaaaagtgccacctgacgtcgacggatcgggagatcgatctcccgatcccctagggtcg
actctcagtacaatctgctctgatgccgcatagttaagccagtatctgctccctgcttgtgtgttggaggtcgctg
agtagtgcgcgagcaaaatttaagctacaacaaggcaaggcttgaccgacaattgcatgaagaatctgcttagggt

taggcgttttgcgctgcttcgcgatgtacgggccagatatacgcgttgacattgattattgactagttattaatag
taatcaattacggggtcattagttcatagcccatatatggagttccgcgttacataacttacggtaaatggcccgc
ctggctgaccgcccaacgacccccgcccattgacgtcaataatgacgtatgttcccatagtaacgccaatagggac
tttccattgacgtcaatgggtggagtatttacggtaaactgcccacttggcagtacatcaagtgtatcatatgcca
agtacgccccctattgacgtcaatgacggtaaatggcccgcctggcattatgcccagtacatgaccttatgggact
ttcctacttggcagtacatctacgtattagtcatcgctattaccatggtgatgcggttttggcagtacatcaatgg
gcgtggatagcggtttgactcacggggatttccaagtctccaccccattgacgtcaatgggagtttgttttggcac
caaaatcaacgggactttccaaaatgtcgtaacaactccgccccattgacgcaaatgggcggtaggcgtgtacggt
gggaggtctatataagcagagctggtttagtgaaccgtcagatccgctagagatccgcggccgctaatacgactca
ctatagggagagccgccaccatgaaacggacagccgacggaagcgagttcgagtcaccaaagaagaagcggaaagt
cagttctgaggtggagttttcccacgagtactggatgagacatgccctgaccctggccaagagggcacgggatgag
ggggcgggtcctgtgggagccgtgctggtgctgaacaatagagtgatcggcgagggctggaacagagccatcggcc
tgcacgacccaacagcccatgccgaaattatggccctgagacagggcggcctggtcatgcagaactacagactgtt
tgacgccaccctgtacgtgacattcgagccttgcgtgatgtgcgccggcgccatgatcaactctaggatcggccgc
gtggtgtttggcgtgaggaactcaaaaagaggcgccgcaggctccctgatgaacgtgctgaactaccccggcatga
atcaccgcgtcgaaattaccgagggaatcctggcagatgaatgtgccgccctgctgtgcgatttctatcggatgcc
tagacaggtgttcaattctcagaagaaggcccagagctccatcaactctggcggatctagcggaggatcctctggc
agcgagacaccaggaacaagcgagtcagcaacaccagagagcagtggcggcagcagcggcggcagcgcccgcatcc
tcgctttcgcaatcggaatctctagtatcggatgggccttctctgaaaacgacgaactgaaagactgcggcgtgag
aatcttcacaaaggttgaaaaccctaaaacaggcgagtctttagctctgccacgtaggttggcccgctccgcccga
aaaaggtatgctcggcggaaggctcgcctcaaccacttgaagcatttgatagctaatgagttcaaactgaactacg
aagattaccagtccttcgacgagtcattggcaaaagcctacaaaggcagccttatcagtccttatgagttgagatt
tcgcgcactcaacgaactgctttctaagcaagactttgctagggtcattctgcacatcgcaaaacggcgaggttat
gacgatatcaagaactccgacgataaagaaaagggagccattctcaaggcgatcaaacagaatgaggaaaaattgg
caaactaccagagtgtgggcgagtatctgtataaagagtatttccagaagtttaaggaaaacagcaaggagtttac
aaacgtcagaaataaaaaggagtcttacgagagatgcatcgcgcagtcattcctcaaagatgagctgaagctgata
tttaagaagcaacgcgaatttggtttctcattctctaagaagttcgaagaggaggttctttccgtggcgttttaca
agagggcgctcaaagacttctcccacctggttggtaactgtagtttcttcacggatgagaagcgagctcccaaaaa
ttctcccctggctttcatgtttgttgccctgactcggatcattaacctgctgaacaacctgaaaaatactgaaggg
atcttgtatacgaaggacgacctaaatgcactcctgaatgaagtgctcaaaaacggaactctaacctataaacaga
ccaagaaattactggggctctctgacgactacgagttcaagggcgagaagggtacttattttatcgaattcaaaaa
gtataaggagttcattaaagcattgggggaacacaacctcagccaggacgatctcaatgaaattgccaaggacatc
acgctgattaaagacgagataaaactgaaaaaggcactggccaagtatgacctcaaccagaaccagatcgactctc
tgtccaagctggagttcaaagaccacctaaacatatccttcaaagccctgaaactggtcacccctctaatgctcga
aggaaaaaaatacgacgaggcgtgtaatgaactgaatcttaaggtggccatcaatgaggataagaaggactttctt
ccagcctttaacgagacatattacaaagacgaggtcacaaacccggttgtgctgagggccataaaagagtatcgga
aggttctgaatgccctcctgaagaagtacggcaaagtgcacaaaataaatatcgaattggctagggaggtggggaa
gaaccattctcagcgagcaaagatcgagaaagagcagaatgagaactacaaagccaagaaagacgccgaactggag
tgcgaaaagctggggcttaaaataaacagtaaaaacatcctgaaattaagattgttcaaagagcaaaaggagtttt

gcgcctactcaggggaaaaaatcaaaatatcagacctgcaggacgagaaaatgctggagatcgaccatatctatcc
gtatagcaggtcatttgacgattcctacatgaacaaagtgcttgtgtttaccaaacagaaccaagaaaagctgaac
caaaccccctttgaggctttcggaaacgactcagccaagtggcagaaaatcgaagtcctagccaagaatctgccta
caaaaaaacaaaagaggattcttgataagaactataaggacaaggaacagaaaaactttaaagacaggaacctgaa
tgacacgaggtacattgcgcgactggttctaaactataccaaagactacctggatttcctccctctgagcgacgac
gagaatactaaactgaatgatacccagaaaggctcaaaggtccacgttgaggctaagtccgggatgctgactagcg
ccctccgccacacgtggggcttcagcgccaaagatcggaataatcatcttcatcacgct(SEQ ID NO:101)

[0651] TadCBEa-SpCas9-BE4max载体

[0652] tagttattaatagtaatcaattacggggtcattagttcatagcccatatatggagttccgcgttacat
aacttacggtaaatggcccgcctggctgaccgcccaacgacccccgcccattgacgtcaataatgacgtatgttcc
catagtaacgccaatagggactttccattgacgtcaatgggtggagtatttacggtaaactgcccacttggcagta
catcaagtgtatcatatgccaagtacgccccctattgacgtcaatgacggtaaatggcccgcctggcattatgccc
agtacatgaccttatgggactttcctacttggcagtacatctacgtattagtcatcgctattaccatggtgatgcg
gttttggcagtacatcaatgggcgtggatagcggtttgactcacggggatttccaagtctccaccccattgacgtc
aatgggagtttgttttggcaccaaaatcaacgggactttccaaaatgtcgtaacaactccgccccattgacgcaaa
tgggcggtaggcgtgtacggtgggaggtctatataagcagagctggtttagtgaaccgtcagatccgctagagatc
cgcggccgctaatacgactcactatagggagagccgccaccatgaaacggacagccgacggaagcgagttcgagtc
accaaagaagaagcggaaagtcagttctgaggtggagttttcccacgagtactggatgagacatgccctgaccctg
gccaagagggcacgggatgaggggcgggtcctgtgggagccgtgctggtgctgaacaatagagtgatcggcgagg
gctggaacagagccatcggcctgcacgacccaacagcccatgccgaaattatggccctgagacagggcggcctggt
catgcagaactacagactgtttgacgccaccctgtacgtgacattcgagccttgcgtgatgtgcgccggcgccatg
atcaactctaggatcggccgcgtggtgtttggcgtgaggaactcaaaaagaggcgccgcaggctccctgatgaacg
tgctgaactaccccggcatgaatcaccgcgtcgaaattaccgagggaatcctggcagatgaatgtgccgccctgct
gtgcgatttctatcggatgcctagacaggtgttcaattctcagaagaaggcccagagctccatcaactctggcgga
tctagcggaggatcctctggcagcgagacaccaggaacaagcgagtcagcaacaccagagagcagtggcggcagca
gcggcggcagcgacaagaagtacagcatcggcctggccatcggcaccaactctgtgggctgggccgtgatcaccga
cgagtacaaggtgcccagcaagaaattcaaggtgctgggcaacaccgaccggcacagcatcaagaagaacctgatc
ggagccctgctgttcgacagcggcgaaacagccgaggccacccggctgaagagaaccgccagaagaagatacacca
gacggaagaaccggatctgctatctgcaagagatcttcagcaacgagatggccaaggtggacgacagcttcttcca
cagactggaagagtccttcctggtggaagaggataagaagcacgagcggcaccccatcttcggcaacatcgtggac
gaggtggcctaccacgagaagtaccccaccatctaccacctgagaaagaaactggtggacagcaccgacaaggccg
acctgcggctgatctatctggccctggcccacatgatcaagttccggggccacttcctgatcgagggcgacctgaa
ccccgacaacagcgacgtggacaagctgttcatccagctggtgcagacctacaaccagctgttcgaggaaaacccc
atcaacgccagcggcgtggacgccaaggccatcctgtctgccagactgagcaagagcagacggctggaaaatctga
tcgcccagctgcccggcgagaagaagaatggcctgttcggaaacctgattgccctgagcctgggcctgacccccaa
cttcaagagcaacttcgacctggccgaggatgccaaactgcagctgagcaaggacacctacgacgacgacctggac
aacctgctggcccagatcggcgaccagtacgccgacctgtttctggccgccaagaacctgtccgacgccatcctgc
tgagcgacatcctgagagtgaacaccgagatcaccaaggcccccctgagcgcctctatgatcaagagatacgacga
gcaccaccaggacctgaccctgctgaaagctctcgtgcggcagcagctgcctgagaagtacaaagagattttcttc

gaccagagcaagaacggctacgccggctacattgacggcggagccagccaggaagagttctacaagttcatcaagc
ccatcctggaaaagatggacggcaccgaggaactgctcgtgaagctgaacagagaggacctgctgcggaagcagcg
gaccttcgacaacggcagcatcccccaccagatccacctgggagagctgcacgccattctgcggcggcaggaagat
ttttacccattcctgaaggacaaccgggaaaagatcgagaagatcctgaccttccgcatcccctactacgtgggcc
ctctggccaggggaaacagcagattcgcctggatgaccagaaagagcgaggaaaccatcacccctggaacttcga
ggaagtggtggacaagggcgcttccgcccagagcttcatcgagcggatgaccaacttcgataagaacctgcccaac
gagaaggtgctgcccaagcacagcctgctgtacgagtacttcaccgtgtataacgagctgaccaaagtgaaatacg
tgaccgagggaatgagaaagcccgccttcctgagcggcgagcagaaaaaggccatcgtggacctgctgttcaagac
caaccggaaagtgaccgtgaagcagctgaaagaggactacttcaagaaaatcgagtgcttcgactccgtggaaatc
tccggcgtggaagatcggttcaacgcctccctgggcacataccacgatctgctgaaaattatcaaggacaaggact
tcctggacaatgaggaaaacgaggacattctggaagatatcgtgctgaccctgacactgtttgaggacagagagat
gatcgaggaacggctgaaaacctatgcccacctgttcgacgacaaagtgatgaagcagctgaagcggcggagatac
accggctggggcaggctgagccggaagctgatcaacggcatccgggacaagcagtccggcaagacaatcctggatt
tcctgaagtccgacggcttcgccaacagaaacttcatgcagctgatccacgacgacagcctgacctttaaagagga
catccagaaagcccaggtgtccggccagggcgatagcctgcacgagcacattgccaatctggccggcagccccgcc
attaagaagggcatcctgcagacagtgaaggtggtggacgagctcgtgaaagtgatgggccggcacaagcccgaga
acatcgtgatcgaaatggccagagagaaccagaccacccagaagggacagaagaacagccgcgagagaatgaagcg
gatcgaagagggcatcaaagagctgggcagccagatcctgaaagaacaccccgtggaaaacacccagctgcagaac
gagaagctgtacctgtactacctgcagaatgggcgggatatgtacgtggaccaggaactggacatcaaccggctgt
ccgactacgatgtggaccatatcgtgcctcagagctttctgaaggacgactccatcgacaacaaggtgctgaccag
aagcgacaagaaccggggcaagagcgacaacgtgccctccgaagaggtcgtgaagaagatgaagaactactggcgg
cagctgctgaacgccaagctgattacccagagaaagttcgacaatctgaccaaggccgagagaggcggcctgagcg
aactggataaggccggcttcatcaagagacagctggtggaaacccggcagatcacaaagcacgtggcacagatcct
ggactcccggatgaacactaagtacgacgagaatgacaagctgatccgggaagtgaaagtgatcaccctgaagtcc
aagctggtgtccgatttccggaaggatttccagttttacaaagtgcgcgagatcaacaactaccaccacgcccacg
acgcctacctgaacgccgtcgtgggaaccgccctgatcaaaaagtaccctaagctggaaagcgagttcgtgtacgg
cgactacaaggtgtacgacgtgcggaagatgatcgccaagagcgagcaggaaatcggcaaggctaccgccaagtac
ttcttctacagcaacatcatgaacttttcaagaccgagattaccctggccaacggcgagatccggaagcggcctc
tgatcgagacaaacggcgaaaccggggagatcgtgtgggataagggccgggatttgccaccgtgcggaaagtgct
gagcatgccccaagtgaatatcgtgaaaaagaccgaggtgcagacaggcggcttcagcaaagagtctatcctgccc
aagaggaacagcgataagctgatcgccagaaagaaggactgggaccctaagaagtacggcggcttcgacagcccca
ccgtggcctattctgtgctggtggtggccaaagtggaaaagggcaagtccaagaaactgaagagtgtgaaagagct
gctggggatcaccatcatggaaagaagcagcttcgagaagaatcccatcgactttctggaagccaagggctacaaa
gaagtgaaaaaggacctgatcatcaagctgcctaagtactccctgttcgagctggaaaacggccggaagagaatgc
tggcctctgccggcgaactgcagaagggaaacgaactggccctgccctccaaatatgtgaacttcctgtacctggc
cagccactatgagaagctgaagggctcccccgaggataatgagcagaaacagctgtttgtggaacagcacaagcac
tacctggacgagatcatcgagcagatcagcgagttctccaagagagtgatcctggccgacgctaatctggacaaag
tgctgtccgcctacaacaagcaccgggataagcccatcagagagcaggccgagaatatcatccacctgtttaccct
gaccaatctgggagcccctgccgccttcaagtactttgacaccaccatcgaccggaagaggtacaccagcaccaaa

gaggtgctggacgccaccctgatccaccagagcatcaccggcctgtacgagacacggatcgacctgtctcagctgg
gaggtgacagcggcgggagcggcgggagcgggggggagcactaatctgagcgacatcattgagaaggagactgggaa
acagctggtcattcaggagtccatcctgatgctgcctgaggaggtggaggaagtgatcggcaacaagccagagtct
gacatcctggtgcacaccgcctacgacgagtccacagatgagaatgtgatgctgctgacctctgacgcccccgagt
ataagccttgggccctggtcatccaggattctaacggcgagaataagatcaagatgctgagcggaggatccggagg
atctggaggcagcaccaacctgtctgacatcatcgagaaggagacaggcaagcagctggtcatccaggagagcatc
ctgatgctgcccgaagaagtcgaagaagtgatcggaaacaagcctgagagcgatatcctggtccataccgcctacg
acgagagtaccgacgaaaatgtgatgctgctgacatccgacgccccagagtataagccctgggctctggtcatcca
ggattccaacggagagaacaaaatcaaaatgctgtctggcggctcaaaaagaaccgccgacggcagcgaattcgag
cccaagaagaagaggaaagtctaaccggtcatcatcaccatcaccattgagtttaaacccgctgatcagcctcgac
tgtgccttctagttgccagccatctgttgtttgcccctcccccgtgccttccttgaccctggaaggtgccactccc
actgtcctttcctaataaaatgaggaaattgcatcgcattgtctgagtaggtgtcattctattctgggggggtgggg
tggggcaggacagcaaggggaggattgggaagacaatagcaggcatgctggggatgcggtgggctctatggcttc
tgaggcggaaagaaccagctggggctcgataccgtcgacctctagctagagcttggcgtaatcatggtcatagctg
tttcctgtgtgaaattgttatccgctcacaattccacacaacatacgagccggaagcataaagtgtaaagcctagg
gtgcctaatgagtgagctaactcacattaattgcgttgcgctcactgcccgctttccagtcgggaaacctgtcgtg
ccagctgcattaatgaatcggccaacgcgcgggaagaggcggtttgcgtattgggcgcacttccgcttcctcgctc
actgactcgctgcgctcggtcgttcggctgcggcgagcggtatcagctcactcaaaggcggtaatacggttatcca
cagaatcaggggataacgcaggaaagaacatgtgagcaaaaggccagcaaaaggccaggaaccgtaaaaaggccgc
gttgctggcgtttttccataggctccgccccctgacgagcatcacaaaaatcgacgctcaagtcagaggtggcga
aacccgacaggactataaagataccaggcgtttccccctggaagctccctcgtgcgctctcctgttccgaccctgc
cgcttaccggatacctgtccgcctttctcccttcgggaagcgtggcgctttctcatagctcacgctgtaggtatct
cagttcggtgtaggtcgttcgctccaagctgggctgtgtgcacgaaccccccgttcagcccgaccgctgcgcctta
tccggtaactatcgtcttgagtccaacccggtaagacacgacttatcgccactggcagcagccactggtaacagga
ttagcagagcgaggtatgtaggcggtgctacagagttcttgaagtggtggcctaactacggctacactagaagaac
agtatttggtatctgcgctctgctgaagccagttaccttcggaaaaagagttggtagctcttgatccggcaaacaa
accaccgctggtagcggtggtttttttgtttgcaagcagcagattacgcgcagaaaaaaaggatctcaagaaaatc
ctttgatcttttctacggggtctgacactcagtggaacgaaaactcacgttaagggattttggtcatgagattatc
aaaaaggatcttcacctagatccttttaaattaaaaatgaagttttaaatcaatctaaagtatatatgagtaaact
tggtctgacagttaccaatgcttaatcagtgaggcacctatctcagcgatctgtctatttcgttcatccatagttg
cctgactccccgtcgtgtagataactacgatacgggagggcttaccatctggccccagtgctgcaatgataccgcg
ggacccacgctcaccggctccagatttatcagcaataaaccagccagccggaagggccgagcgcagaagtggtcct
gcaactttatccgcctccatccagtctattaattgttgccgggaagctagagtaagtagttcgccagttaatagtt
tgcgcaacgttgttgccattgctacaggcatcgtggtgtcacgctcgtcgtttggtatggcttcattcagctccgg
ttcccaacgatcaaggcgagttacatgatcccccatgttgtgcaaaaaagcggttagctccttcggtcctccgatc
gttgtcagaagtaagttggccgcagtgttatcactcatggttatggcagcactgcataattctcttactgtcatgc
catccgtaagatgcttttctgtgactggtgagtactcaaccaagtcattctgagaatagtgtatgcggcgaccgag
ttgctcttgcccggcgtcaatacgggataataccgcgccacatagcagaactttaaaagtgctcatcattggaaaa
cgttcttcggggcgaaaactctcaaggatcttaccgctgttgagatccagttcgatgtaacccactcgtgcaccca

actgatcttcagcatcttttactttcaccagcgtttctgggtgagcaaaaacaggaaggcaaaatgccgcaaaaaa
gggaataagggcgacacggaaatgttgaatactcatactcttccttttttcaatattattgaagcatttatcagggt
tattgtctcatgagcggatacatatttgaatgtatttagaaaaataaacaaatagggttccgcgcacatttcccc
gaaaagtgccacctgacgtcgacggatcgggagatcgatctcccgatcccctagggtcgactctcagtacaatctg
ctctgatgccgcatagttaagccagtatctgctccctgcttgtgtgttggaggtcgctgagtagtgcgcgagcaaa
atttaagctacaacaaggcaaggcttgaccgacaattgcatgaagaatctgcttagggttaggcgttttgcgctgc
ttcgcgatgtacgggccagatatacgcgttgacattgattattgac(SEQ ID NO:102)

[0653] TadCBEa-SaCas9-BE4max载体

[0654] tagttattaatagtaatcaattacggggtcattagttcatagcccatatatggagttccgcgttacat
aacttacggtaaatggcccgcctggctgaccgcccaacgacccccgcccattgacgtcaataatgacgtatgttcc
catagtaacgccaatagggactttccattgacgtcaatgggtggagtatttacggtaaactgcccacttggcagta
catcaagtgtatcatatgccaagtacgccccctattgacgtcaatgacggtaaatggcccgcctggcattatgccc
agtacatgaccttatgggactttcctacttggcagtacatctacgtattagtcatcgctattaccatggtgatgcg
gttttggcagtacatcaatgggcgtggatagcggtttgactcacggggatttccaagtctccaccccattgacgtc
aatgggagtttgttttggcaccaaaatcaacgggactttccaaaatgtcgtaacaactccgccccattgacgcaaa
tgggcggtaggcgtgtacggtgggaggtctatataagcagagctggtttagtgaaccgtcagatccgctagagatc
cgcggccgctaatacgactcactatagggagagccgccaccatgaaacggacagccgacggaagcgagttcgagtc
accaaagaagaagcggaaagtcagttctgaggtggagttttcccacgagtactggatgagacatgccctgaccctg
gccaagagggcacgggatgaggggcgggtcctgtgggagccgtgctggtgctgaacaatagagtgatcggcgagg
gctggaacagagccatcggcctgcacgacccaacagcccatgccgaaattatggccctgagacagggcggcctggt
catgcagaactacagactgtttgacgccaccctgtacgtgacattcgagccttgcgtgatgtgcgccggcgccatg
atcaactctaggatcggccgcgtggtgtttggcgtgaggaactcaaaaagaggcgccgcaggctccctgatgaacg
tgctgaactaccccggcatgaatcaccgcgtcgaaattaccgagggaatcctggcagatgaatgtgccgccctgct
gtgcgatttctatcggatgcctagacaggtgttcaattctcagaagaaggcccagagctccatcaactctggcgga
tctagcggaggatcctctggcagcgagacaccaggaacaagcgagtcagcaacaccagagagcagtggcggcagca
gcggcggcagcgggaagcgaaattacattctggggctggccattggcattacatcagtgggctatggcatcattga
ctacgagacaagggacgtgatcgacgccggcgtgagactgttcaaggaggccaacgtggagaacaatgagggccgg
agatccaagaggggagcaaggcgcctgaagcggagaaggcgccacagaatccagagagtgaagaagctgctgttcg
attacaacctgctgaccgaccactccgagctgtctggcatcaatccttatgaggccagagtgaagggcctgtccca
gaagctgtctgaggaggagtttagcgccgccctgctgcacctggcaaagaggagaggcgtgcacaacgtgaatgag
gtggaggaggacaccggcaacgagctgtccacaaaggagcagatcagccgcaattccaaggccctggaggagaagt
atgtggccgagctgcagctggagcggctgaagaaggatggcgaggtgaggggctccatcaatcgcttcaagacctc
tgactacgtgaaggaggccaagcagctgctgaaggtgcagaaggcctaccaccagctggatcagtcctttatcgat
acatatatcgacctgctggagacaaggcgcacatactatgagggaccaggagagggctctcccttcggctggaagg
acatcaaggagtggtacgagatgctgatgggccactgcacctattttccagaggagctgagaagcgtgaagtacgc
ctataacgccgatctgtacaacgccctgaatgacctgaacaacctggtcatcaccagggatgagaacgagaagctg
gagtactatgagaagttccagatcatcgagaacgtgttcaagcagaagaagaagcctacactgaagcagatcgcca
aggagatcctggtgaacgaggaggacatcaagggctaccgcgtgacctccacaggcaagccagagttcaccaatct
gaaggtgtatcacgatatcaaggacatcacagcccggaaggagatcatcgagaacgccgagctgctggatcagatc

gccaagatcctgaccatctatcagagctccgaggacatccaggaggagctgaccaacctgaatagcgagctgacac
aggaggagatcgagcagatcagcaatctgaagggctacaccggcacacacaacctgagcctgaaggccatcaatct
gatcctggatgagctgtggcacacaaacgacaatcagatcgccatctttaaccggctgaagctggtgccaaagaag
gtggacctgtcccagcagaaggagatcccaaccacactggtggacgatttcatcctgtctcccgtggtgaagcgga
gcttcatccagagcatcaaagtgatcaacgccatcatcaagaagtacggcctgcccaatgatatcatcatcgagct
ggccagggagaagaactccaaggacgcccagaagatgatcaatgagatgcagaagaggaaccgccagaccaatgag
cggatcgaggagatcatcagaaccacaggcaaggagaacgccaagtacctgatcgagaagatcaagctgcacgata
tgcaggagggcaagtgtctgtattctctggaggccatccctctggaggacctgctgaacaatccattcaactacga
ggtggatcacatcatcccccggagcgtgagcttcgacaattcttttaacaataaggtgctggtgaagcaggaggag
aacagcaagaagggcaataggaccccttccagtacctgtctagctccgattctaagatcagctacgagacattca
agaagcacatcctgaatctggccaagggcaagggccgcatcagcaagaccaagaaggagtacctgctggaggagcg
ggacatcaacagattctccgtgcagaaggacttcatcaaccggaatctggtggacaccagatacgccacacgcggc
ctgatgaatctgctgcggtcttatttcagagtgaacaatctggatgtgaaggtaaagagcatcaacggcggcttca
cctcctttctgcggagaaagtggaagtttaagaaggagcgcaacaagggctataagcaccacgccgaggatgccct
gatcatcgccaatgccgacttcatctttaaggagtggaagaagctggacaaggccaagaaagtgatggagaaccag
atgttcgaggagaagcaggccgagagcatgcccgagatcgagacagagcaggagtacaaggagattttcatcacac
ctcaccagatcaagcacatcaaggacttcaaggactacaagtattctcacagggtggataagaagcccaaccgcga
gctgatcaatgacaccctgtatagcacacggaaggacgataagggcaataccctgatcgtgaacaatctgaacggc
ctgtacgacaaggataatgacaagctgaagaagctgatcaacaagtctcccgagaagctgctgatgtaccaccacg
atcctcagacatatcagaagctgaagctgatcatggagcagtacggcgacgagaagaacccactgtataagtacta
tgaggagacaggcaactacctgacaaagtatagcaagaaggataatggccccgtgatcaagaagatcaagtactat
ggcaacaagctgaatgcccacctggacatcaccgacgattaccctaactctcgcaataaggtggtgaagctgagcc
tgaagccataccggttcgacgtgtacctggacaacggcgtgtataagtttgtgacagtgaagaatctggatgtgat
caagaaggagaactactatgaggtgaacagcaagtgctacgaggaggccaagaagctgaagaagatcagcaaccag
gccgagttcatcgcctctttttacaacaatgacctgatcaagatcaatggcgagctgtatagagtgatcggcgtga
acaatgatctgctgaacagaatcgaagtgaatatgatcgacatcacctacagggagtatctggagaacatgaatga
taagaggccccctcgcatcatcaagaccatcgcctctaagacacagagcatcaagaagtacagcacagacatcctg
gggaacctgtatgaagtcaagagcaagaaacatcctcagattatcaagaaaggcagcggcgggagcggcgggagcg
gggggagcactaatctgagcgacatcattgagaaggagactgggaaacagctggtcattcaggagtccatcctgat
gctgcctgaggaggtggaggaagtgatcggcaacaagccagagtctgacatcctggtgcacaccgcctacgacgag
tccacagatgagaatgtgatgctgctgacctctgacgcccccgagtataagccttgggccctggtcatccaggatt
ctaacggcgagaataagatcaagatgctgagcggaggatccggaggatctggaggcagcaccaacctgtctgacat
catcgagaaggagacaggcaagcagctggtcatccaggagagcatcctgatgctgcccgaagaagtcgaagaagtg
atcggaaacaagcctgagagcgatatcctggtccataccgcctacgacgagagtaccgacgaaaatgtgatgctgc
tgacatccgacgccccagagtataagccctgggctctggtcatccaggattccaacggagagaacaaaatcaaaat
gctgtctggcggctcaaaaagaaccgccgacggcagcgaattcgagcccaagaagaagaggaaagtctaaccggtc
atcatcaccatcaccattgagtttaaacccgctgatcagcctcgactgtgccttctagttgccagccatctgttgt
ttgcccctcccccgtgccttccttgaccctggaaggtgccactcccactgtcctttcctaataaaatgaggaaatt
gcatcgcattgtctgagtaggtgtcattctattctgggggggtggggtggggcaggacagcaaggggaggattggg

aagacaatagcaggcatgctggggatgcggtgggctctatggcttctgaggcggaaagaaccagctggggctcgat
accgtcgacctctagctagagcttggcgtaatcatggtcatagctgtttcctgtgtgaaattgttatccgctcaca
attccacacaacatacgagccggaagcataaagtgtaaagcctagggtgcctaatgagtgagctaactcacattaa
ttgcgttgcgctcactgcccgctttccagtcgggaaacctgtcgtgccagctgcattaatgaatcggccaacgcgc
gggaagaggcggtttgcgtattgggcgcacttccgcttcctcgctcactgactcgctgcgctcggtcgttcggctg
cggcgagcggtatcagctcactcaaaggcggtaatacggttatccacagaatcaggggataacgcaggaaagaaca
tgtgagcaaaaggccagcaaaaggccaggaaccgtaaaaaggccgcgttgctggcgtttttccataggctccgccc
ccctgacgagcatcacaaaaatcgacgctcaagtcagaggtggcgaaacccgacaggactataaagataccaggcg
tttccccctggaagctccctcgtgcgctctcctgttccgaccctgccgcttaccggatacctgtccgcctttctcc
cttcgggaagcgtggcgctttctcatagctcacgctgtaggtatctcagttcggtgtaggtcgttcgctccaagct
gggctgtgtgcacgaacccccgttcagcccgaccgctgcgccttatccggtaactatcgtcttgagtccaacccg
gtaagacacgacttatcgccactggcagcagccactggtaacaggattagcagagcgaggtatgtaggcggtgcta
cagagttcttgaagtggtggcctaactacggctacactagaagaacagtatttggtatctgcgctctgctgaagcc
agttaccttcggaaaaagagttggtagctcttgatccggcaaacaaaccaccgctggtagcggtggtttttttgtt
tgcaagcagcagattacgcgcagaaaaaaaggatctcaagaaaatcctttgatcttttctacggggtctgacactc
agtggaacgaaaactcacgttaagggattttggtcatgagattatcaaaaaggatcttcacctagatccttttaaa
ttaaaaatgaagttttaaatcaatctaaagtatatatgagtaaacttggtctgacagttaccaatgcttaatcagt
gaggcacctatctcagcgatctgtctatttcgttcatccatagttgcctgactccccgtcgtgtagataactacga
tacgggagggcttaccatctggccccagtgctgcaatgataccgcgggacccacgctcaccggctccagatttatc
agcaataaaccagccagccggaagggccgagcgcagaagtggtcctgcaactttatccgcctccatccagtctatt
aattgttgccgggaagctagagtaagtagttcgccagttaatagtttgcgcaacgttgttgccattgctacaggca
tcgtggtgtcacgctcgtcgtttggtatggcttcattcagctccggttcccaacgatcaaggcgagttacatgatc
ccccatgttgtgcaaaaaagcggttagctccttcggtcctccgatcgttgtcagaagtaagttggccgcagtgtta
tcactcatggttatggcagcactgcataattctcttactgtcatgccatccgtaagatgcttttctgtgactggtg
agtactcaaccaagtcattctgagaatagtgtatgcggcgaccgagttgctcttgcccggcgtcaatacgggataa
taccgcgccacatagcagaactttaaaagtgctcatcattggaaaacgttcttcggggcgaaaactctcaaggatc
ttaccgctgttgagatccagttcgatgtaacccactcgtgcacccaactgatcttcagcatcttttactttcacca
gcgtttctgggtgagcaaaaacaggaaggcaaaatgccgcaaaaaagggaataagggcgacacggaaatgttgaat
actcatactcttccttttcaatattattgaagcatttatcagggttattgtctcatgagcggatacatatttgaa
tgtatttagaaaaataaacaaatagggttccgcgcacatttccccgaaaagtgccacctgacgtcgacggatcgg
gagatcgatctcccgatcccctagggtcgactctcagtacaatctgctctgatgccgcatagttaagccagtatct
gctccctgcttgtgtgttggaggtcgctgagtagtgcgcgagcaaaatttaagctacaacaaggcaaggcttgacc
gacaattgcatgaagaatctgcttagggttaggcgttttgcgctgcttcgcgatgtacgggccagatatacgcgtt
gacattgattattgac(SEQ ID NO:103)

[0655] TadCBEa-SpCas9-NG-BE4max载体

[0656] tagttattaatagtaatcaattacggggtcattagttcatagcccatatatggagttccgcgttacat
aacttacggtaaatggcccgcctggctgaccgcccaacgacccccgcccattgacgtcaataatgacgtatgttcc
catagtaacgccaatagggactttccattgacgtcaatgggtggagtatttacggtaaactgcccacttggcagta
catcaagtgtatcatatgccaagtacgccccctattgacgtcaatgacggtaaatggcccgcctggcattatgccc

agtacatgaccttatgggactttcctacttggcagtacatctacgtattagtcatcgctattaccatggtgatgcg
gttttggcagtacatcaatgggcgtggatagcggtttgactcacggggatttccaagtctccaccccattgacgtc
aatgggagtttgttttggcaccaaaatcaacgggactttccaaaatgtcgtaacaactccgccccattgacgcaaa
tgggcggtaggcgtgtacggtgggaggtctatataagcagagctggtttagtgaaccgtcagatccgctagagatc
cgcggccgctaatacgactcactatagggagagccgccaccatgaaacggacagccgacggaagcgagttcgagtc
accaaagaagaagcggaaagtcagttctgaggtggagttttcccacgagtactggatgagacatgccctgaccctg
gccaagagggcacgggatgaggggcgggtcctgtgggagccgtgctggtgctgaacaatagagtgatcggcgagg
gctggaacagagccatcggcctgcacgacccaacagcccatgccgaaattatggccctgagacagggcggcctggt
catgcagaactacagactgtttgacgccaccctgtacgtgacattcgagccttgcgtgatgtgcgccggcgccatg
atcaactctaggatcggccgcgtggtgtttggcgtgaggaactcaaaaagaggcgccgcaggctccctgatgaacg
tgctgaactaccccggcatgaatcaccgcgtcgaaattaccgagggaatcctggcagatgaatgtgccgccctgct
gtgcgatttctatcggatgcctagacaggtgttcaattctcagaagaaggcccagagctccatcaactctggcgga
tctagcggaggatcctctggcagcgagacaccaggaacaagcgagtcagcaacaccagagagcagtggcggcagca
gcggcggcagcgacaagaagtacagcatcggcctggccatcggcaccaactctgtgggctgggccgtgatcaccga
cgagtacaaggtgcccagcaagaaattcaaggtgctgggcaacaccgaccggcacagcatcaagaagaacctgatc
ggagccctgctgttcgacagcggcgaaacagccgaggccacccggctgaagagaaccgccagaagaagatacacca
gacggaagaaccggatctgctatctgcaagagatcttcagcaacgagatggccaaggtggacgacagcttcttcca
cagactggaagagtccttcctggtggaagaggataagaagcacgagcggcaccccatcttcggcaacatcgtggac
gaggtggcctaccacgagaagtaccccaccatctaccacctgagaaagaaactggtggacagcaccgacaaggccg
acctgcggctgatctatctggccctggcccacatgatcaagttccggggccacttcctgatcgagggcgacctgaa
ccccgacaacagcgacgtggacaagctgttcatccagctggtgcagacctacaaccagctgttcgaggaaaaccccc
atcaacgccagcggcgtggacgccaaggccatcctgtctgccagactgagcaagagcagacggctggaaaatctga
tcgcccagctgcccggcgagaagaagaatggcctgttcggaaacctgattgccctgagcctgggcctgacccccaa
cttcaagagcaacttcgacctggccgaggatgccaaactgcagctgagcaaggacacctacgacgacgacctggac
aacctgctggcccagatcggcgaccagtacgccgacctgtttctggccgccaagaacctgtccgacgccatcctgc
tgagcgacatcctgagagtgaacaccgagatcaccaaggcccccctgagcgcctctatgatcaagagatacgacga
gcaccaccaggacctgaccctgctgaaagctctcgtgcggcagcagctgcctgagaagtacaaagagatttttctc
gaccagagcaagaacggctacgccggctacattgacggcggagccagccaggaagagttctacaagttcatcaagc
ccatcctggaaaagatggacggcaccgaggaactgctcgtgaagctgaacagagaggacctgctgcggaagcagcg
gaccttcgacaacggcagcatccccccaccagatccacctgggagagctgcacgccattctgcggcggcaggaagat
ttttacccattcctgaaggacaaccgggaaaagatcgagaagatcctgaccttccgcatcccctactacgtgggcc
ctctggccaggggaaacagcagattcgcctggatgaccagaaagagcgaggaaaccatcacccctggaacttcga
ggaagtggtggacaagggcgcttccgcccagagcttcatcgagcggatgaccaacttcgataagaacctgcccaac
gagaaggtgctgcccaagcacagcctgctgtacgagtacttcaccgtgtataacgagctgaccaaagtgaaatacg
tgaccgagggaatgagaaagcccgccttcctgagcggcgagcagaaaaaggccatcgtggacctgctgttcaagac
caaccggaaagtgaccgtgaagcagctgaaagaggactacttcaagaaaatcgagtgcttcgactccgtggaaatc
tccggcgtggaagatcggttcaacgcctccctgggcacataccacgatctgctgaaaattatcaaggacaaggact
tcctggacaatgaggaaaacgaggacattctggaagatatcgtgctgaccctgacactgtttgaggacagagagat
gatcgaggaacggctgaaaacctatgcccacctgttcgacgacaaagtgatgaagcagctgaagcggcggagatac

accggctggggcaggctgagccggaagctgatcaacggcatccgggacaagcagtccggcaagacaatcctggatt
tcctgaagtccgacggcttcgccaacagaaacttcatgcagctgatccacgacgacagcctgacctttaaagagga
catccagaaagcccaggtgtccggccagggcgatagcctgcacgagcacattgccaatctggccggcagccccgcc
attaagaagggcatcctgcagacagtgaaggtggtggacgagctcgtgaaagtgatgggccggcacaagcccgaga
acatcgtgatcgaaatggccagagagaaccagaccacccagaagggacagaagaacagccgcgagagaatgaagcg
gatcgaagagggcatcaaagagctgggcagccagatcctgaaagaacaccccgtggaaaacacccagctgcagaac
gagaagctgtacctgtactacctgcagaatgggcgggatatgtacgtggaccaggaactggacatcaaccggctgt
ccgactacgatgtggaccatatcgtgcctcagagctttctgaaggacgactccatcgacaacaaggtgctgaccag
aagcgacaagaaccggggcaagagcgacaacgtgccctccgaagaggtcgtgaagaagatgaagaactactggcgg
cagctgctgaacgccaagctgattacccagagaaagttcgacaatctgaccaaggccgagagaggcggcctgagcg
aactggataaggccggcttcatcaagagacagctggtggaaacccggcagatcacaaagcacgtggcacagatcct
ggactcccggatgaacactaagtacgacgagaatgacaagctgatccgggaagtgaaagtgatcaccctgaagtcc
aagctggtgtccgatttccggaaggatttccagttttacaaagtgcgcgagatcaacaactaccaccacgcccacg
acgcctacctgaacgccgtcgtgggaaccgccctgatcaaaaagtaccctaagctggaaagcgagttcgtgtacgg
cgactacaaggtgtacgacgtgcggaagatgatcgccaagagcgagcaggaaatcggcaaggctaccgccaagtac
ttcttctacagcaacatcatgaacttttcaagaccgagattaccctggccaacggcgagatccggaagcggcctc
tgatcgagacaaacggcgaaaccggggagatcgtgtgggataagggccgggattttgccaccgtgcggaaagtgct
gagcatgccccaagtgaatatcgtgaaaaagaccgaggtgcagacaggcggcttcagcaaagagtctatccggccc
aagaggaacagcgataagctgatcgccagaaagaaggactgggaccctaagaagtacggcggcttcgtgagcccca
ccgtggcctattctgtgctggtggtggccaaagtggaaaagggcaagtccaagaaactgaagagtgtgaaagagct
gctggggatcaccatcatggaaagaagcagcttcgagaagaatcccatcgactttctggaagccaagggctacaaa
gaagtgaaaaaggacctgatcatcaagctgcctaagtactccctgttcgagctggaaaacggccggaagagaatgc
tggcctctgcccggttcctgcagaagggaaacgaactggccctgccctccaaatatgtgaacttcctgtacctggc
cagccactatgagaagctgaagggctcccccgaggataatgagcagaaacagctgtttgtggaacagcacaagcac
tacctggacgagatcatcgagcagatcagcgagttctccaagagagtgatcctggccgacgctaatctggacaaag
tgctgtccgcctacaacaagcaccgggataagcccatcagagagcaggccgagaatatcatccacctgtttaccct
gaccaatctgggagcccctcgggccttcaagtactttgacaccaccatcgaccggaaggtgtaccggagcaccaaa
gaggtgctggacgccaccctgatccaccagagcatcaccggcctgtacgagacacggatcgacctgtctcagctgg
gaggtgacagcggcgggagcggcgggagcgggggggagcactaatctgagcgacatcattgagaaggagactgggaa
acagctggtcattcaggagtccatcctgatgctgcctgaggaggtggaggaagtgatcggcaacaagccagagtct
gacatcctggtgcacaccgcctacgacgagtccacagatgagaatgtgatgctgctgacctctgacgcccccgagt
ataagccttgggccctggtcatccaggattctaacggcgagaataagatcaagatgctgagcggaggatccggagg
atctggaggcagcaccaacctgtctgacatcatcgagaaggagacaggcaagcagctggtcatccaggagagcatc
ctgatgctgcccgaagaagtcgaagaagtgatcggaaacaagcctgagagcgatatcctggtccataccgcctacg
acgagagtaccgacgaaaatgtgatgctgctgacatccgacgccccagagtataagccctgggctctggtcatcca
ggattccaacggagagaacaaaatcaaaatgctgtctggcggctcaaaaagaaccgccgacggcagcgaattcgag
cccaagaagaagaggaaagtctaaccggtcatcatcaccatcaccattgagtttaaacccgctgatcagcctcgac
tgtgccttctagttgccagccatctgttgtttgcccctcccccgtgccttccttgaccctggaaggtgccactccc
actgtcctttcctaataaaatgaggaaattgcatcgcattgtctgagtaggtgtcattctattctgggggggtgggg

tggggcaggacagcaaggggaggattgggaagacaatagcaggcatgctggggatgcggtgggctctatggcttc
tgaggcggaaagaaccagctggggctcgataccgtcgacctctagctagagcttggcgtaatcatggtcatagctg
tttcctgtgtgaaattgttatccgctcacaattccacacaacatacgagccggaagcataaagtgtaaagcctagg
gtgcctaatgagtgagctaactcacattaattgcgttgcgctcactgcccgctttccagtcgggaaacctgtcgtg
ccagctgcattaatgaatcggccaacgcgcggggagaggcggtttgcgtattgggcgcacttccgcttcctcgctc
actgactcgctgcgctcggtcgttcggctgcggcgagcggtatcagctcactcaaaggcggtaatacggttatcca
cagaatcaggggataacgcaggaaagaacatgtgagcaaaaggccagcaaaaggccaggaaccgtaaaaaggccgc
gttgctggcgtttttccataggctccgccccctgacgagcatcacaaaaatcgacgctcaagtcagaggtggcga
aacccgacaggactataaagataccaggcgtttccccctggaagctccctcgtgcgctctcctgttccgaccctgc
cgcttaccggatacctgtccgcctttctcccttcgggaagcgtggcgctttctcatagctcacgctgtaggtatct
cagttcggtgtaggtcgttcgctccaagctgggctgtgtgcacgaacccccgttcagcccgaccgctgcgccttа
tccggtaactatcgtcttgagtccaacccggtaagacacgacttatcgccactggcagcagccactggtaacagga
ttagcagagcgaggtatgtaggcggtgctacagagttcttgaagtggtggcctaactacggctacactagaagaac
agtatttggtatctgcgctctgctgaagccagttaccttcggaaaaagagttggtagctcttgatccggcaaacaa
accaccgctggtagcggtggtttttttgtttgcaagcagcagattacgcgcagaaaaaaaggatctcaagaagatc
ctttgatcttttctacggggtctgacactcagtggaacgaaaactcacgttaagggattttggtcatgagattatc
aaaaaggatcttcacctagatccttttaaattaaaaatgaagttttaaatcaatctaaagtatatatgagtaaact
tggtctgacagttaccaatgcttaatcagtgaggcacctatctcagcgatctgtctatttcgttcatccatagttg
cctgactccccgtcgtgtagataactacgatacgggagggcttaccatctggccccagtgctgcaatgataccgcg
ggacccacgctcaccggctccagatttatcagcaataaaccagccagccggaagggccgagcgcagaagtggtcct
gcaactttatccgcctccatccagtctattaattgttgccgggaagctagagtaagtagttcgccagttaatagtt
tgcgcaacgttgttgccattgctacaggcatcgtggtgtcacgctcgtcgtttggtatggcttcattcagctccgg
ttcccaacgatcaaggcgagttacatgatcccccatgttgtgcaaaaaagcggttagctccttcggtcctccgatc
gttgtcagaagtaagttggccgcagtgttatcactcatggttatggcagcactgcataattctcttactgtcatgc
catccgtaagatgcttttctgtgactggtgagtactcaaccaagtcattctgagaatagtgtatgcggcgaccgag
ttgctcttgcccggcgtcaatacgggataataccgcgccacatagcagaactttaaaagtgctcatcattggaaaa
cgttcttcggggcgaaaactctcaaggatcttaccgctgttgagatccagttcgatgtaacccactcgtgcaccca
actgatcttcagcatcttttactttcaccagcgtttctgggtgagcaaaaacaggaaggcaaaatgccgcaaaaaa
gggaataagggcgacacggaaatgttgaatactcatactcttccttttcaatattattgaagcatttatcagggt
tattgtctcatgagcggatacatatttgaatgtatttagaaaaataaacaaatagggttccgcgcacatttcccc
gaaaagtgccacctgacgtcgacggatcgggagatcgatctcccgatcccctagggtcgactctcagtacaatctg
ctctgatgccgcatagttaagccagtatctgctccctgcttgtgtgttggaggtcgctgagtagtgcgcgagcaaa
atttaagctacaacaaggcaaggcttgaccgacaattgcatgaagaatctgcttagggttaggcgttttgcgctgc
ttcgcgatgtacgggccagatatacgcgttgacattgattattgac(SEQ ID NO:104)

[0657] 载体可以在原核细胞中引入并繁殖。在一些实施方案中,原核生物用于扩增待引入真核细胞的载体的副本,或作为生产待引入真核细胞的载体的中间载体(例如作为病毒载体包装系统的一部分扩增质粒)。在一些实施方案中,原核生物用于扩增载体的副本并表达一个或多个核酸,如提供用于向宿主细胞或宿主生物体递送的一种或多种蛋白质的来源。在原核生物中表达蛋白质通常在大肠杆菌中进行,其中载体含有组成型或诱导型启动

子,指导融合物或非碱基编辑器的表达。

[0658] 融合表达载体也可用于表达本公开的TadA-CD碱基编辑器。此类载体通常在其编码的蛋白质中添加多个氨基酸,如重组蛋白的氨基末端。此类融合载体可以服务于一个或多个目的,例如:(i)以增加重组蛋白的表达;(ii)以增加重组蛋白的溶解度;以及(iii)以通过作为亲和纯化中的配体帮助重组蛋白的纯化。通常在融合表达载体中,在融合部分和重组蛋白的连接处引入蛋白水解切割位点,以便在碱基编辑器纯化后将重组蛋白与融合部分分离。此类酶及其识别序列包括Factor Xa、凝血酶和肠激酶。示例性融合表达载体包括pGEX(Pharmacia Biotech Inc;Smith和Johnson,1988.Gene 67:31-40)、pMAL(New England Biolabs,Beverly,Mass.)和pRIT5(Pharmacia,Piscataway,N.J.),它们分别将谷胱甘肽S-转移酶(GST)、麦芽糖E结合蛋白或蛋白A融合到靶标重组蛋白上。

[0659] 合适的诱导型非融合大肠杆菌表达载体的实例包括pTrc(Amrann et al.,(1988)Gene 69:301-315)和pET 11d(Studier et al.,GENE EXPRESSION TECHNOLOGY:METHODS IN ENZYMOLOGY 185,Academic Press,San Diego,Calif.(1990)60-89)。

[0660] 在一些实施方案中,载体使用杆状病毒表达载体在昆虫细胞中驱动蛋白质表达。可用于在培养的昆虫细胞(例如SF9细胞)中表达蛋白质的杆状病毒载体包括pAc系列(Smith,et al.,1983.Mol.Cell.Biol.3:2156-2165)和pVL系列(Lucklow and Summers,1989.Virology 170:31-39)。

[0661] 在一些实施方案中,载体能够使用哺乳动物表达载体在哺乳动物细胞中驱动一个或多个序列的表达。哺乳动物表达载体的示例包括pCDM8(Seed,1987.Nature 329:840)和pMT2PC(Kaufman,et al.,1987.EMBO J.6:187-195)。当用于哺乳动物细胞时,表达载体的控制功能通常由一个或多个调控元件提供。例如,常用的启动子衍生自多瘤病毒、腺病毒2、巨细胞病毒、猿猴病毒40以及本文所公开的和本领域公知的其他。用于原核和真核细胞的其他合适的表达系统参见例如Sambrook,et al.,MOLECULAR CLONING:A LABORATORY MANUAL.2nd ed.,Cold Spring Harbor Laboratory,Cold Spring Harbor Laboratory Press,Cold Spring Harbor,N.Y.,1989的第16和17章。

[0662] 在一些实施方案中,重组哺乳动物表达载体能够优先在特定细胞类型中指导核酸的表达(例如,使用组织特异性调控元件以表达核酸)。组织特异性调控元件为本领域所熟知。合适的组织特异性启动子的非限制性实例包括白蛋白启动子(肝脏特异性;Pinkert,et al.,1987.Genes Dev.1:268-277)、淋巴特异性启动子(Calame and Eaton,1988.Adv.Immunol.43:235-275)特别是T细胞受体启动子(Winoto and Baltimore,1989.EMBO J.8:729-733)和免疫球蛋白启动子(Baneiji,et al.,1983.Cell 33:729-740;Queen and Baltimore,1983.Cell 33:741-748)、神经元特异性启动子(例如神经丝启动子;Byrne and Ruddle,1989.Proc.Natl.Acad.Sci.USA 86:5473-5477)、胰腺特异性启动子(Edlund,et al.,1985.Science 230:912-916)和乳腺特异性启动子(例如乳清蛋白启动子,美国专利号4,873,316和欧洲申请公布号264,166)。发育调控的启动子也涵盖在内,例如鼠hox启动子(Kessel and Gruss,1990.Science 249:374-379)和α-胎蛋白启动子(Campes and Tilghman,1989.Genes Dev.3:537-546)

[0663] 使用TadA衍生的胞嘧啶碱基编辑器的方法

[0664] 本公开的一些方面提供了使用本文所述的TadA-CD碱基编辑器的方法,如例如编

辑核酸(例如双链DNA序列的碱基对)。在一些实施方案中,该方法包括以下步骤:a)将核酸(例如双链DNA序列)的靶区域与包含碱基编辑器(例如与TadA-CD结构域融合的Cas9结构域)和向导核酸(例如gRNA)的复合物接触,其中靶区域包含靶向核碱基对。作为这些方法实施方案的结果,诱导所述靶区域的链分离,将所述靶区域的单链中所述靶核碱基对中的第一核碱基转化为第二核碱基,并且所述靶区域中不超过一条链被切割(或产生切口),其中与第一核碱基互补的第三核碱基被与第二核碱基互补的第四核碱基替换。

[0665] 在某些方面,本发明涉及一种方法,包括将核酸与本文所述的任何碱基编辑器(例如TadA-CD)或复合物接触。在一些实施方案中,核酸包含细胞基因组中的靶序列(例如DNA)。在一些情况下,核酸是DNA。DNA可以是单链的或双链的。

[0666] 根据一些实施方案,靶序列可以包含与疾病或病症相关的序列。在一些实施方案中,疾病或病症是HIV/AIDS。在一些实施方案中,疾病或病症是镰状细胞或相关的血红蛋白病。在一些情况下,靶序列可以包含靶基因序列。例如,在一些实施方案中,靶序列包含BCL11A增强子或者CCR5或CXCR4基因中的序列(例如基因内的子序列)。在一些情况下,靶序列可以包含与疾病或病症相关的点突变(例如CCR5中的突变降低HIV感染性)。

[0667] 在一些情况下,将包含含有点突变的靶序列的核酸与本文所述的一个或多个碱基编辑器接触导致点突变的校正。在一些实施方案中,靶序列包含与疾病或病症相关的T至C点突变,其可以例如通过使用本文所述的TadCBE对突变体C碱基进行脱氨来校正,导致与疾病或病症不相关的序列。在某些实施方案中,靶序列包含与疾病或病症相关的A至G点突变,并且对与A至G点突变的G碱基互补的C碱基的脱氨导致与疾病或病症无关的序列。在一些情况下,靶序列(例如编码蛋白质)包含密码子(例如突变体C密码子)中的点突变,该突变与野生型密码子相比改变了由突变体密码子编码的氨基酸。因此,例如使用所公开的任何TadCBE对突变体C密码子进行脱氨,可以用于将由突变体密码子编码的氨基酸改变为野生型氨基酸。

[0668] 在一些实施方案中,靶序列可以包含一个或多个C:T或A:G点突变。使用本文所述的TadCBE对与A至G点突变的鸟嘌呤碱基互补的胞嘧啶碱基进行脱氨,导致由突变体密码子编码的氨基酸发生变化。同样,使用TadCBE对与C至T点突变的T碱基互补的A碱基进行脱氨,导致编码野生型氨基酸的密码子。在一些实施方案中,靶序列包含DNA序列5'-NCN-3',其中N是A、T、C或G。在其他实施方案中,靶序列包含DNA序列5'-NCN-3',其中胞苷被脱氨。本领域技术人员将理解,在正常的先天修复过程后,DNA序列5'-NCN-3'中的脱氨胞苷(例如尿嘧啶)会变为T(DNA聚合酶将尿嘧啶读取为胸苷)。

[0669] 如下文更详细描述的,靶序列的脱氨可能在一些情况下导致终止密码子的移除和/或插入。不希望受理论的束缚,本领域技术人员将理解,可接受的终止密码子包含核酸序列5'-TAG-3'、5'-TAA-3'或5'-TGA-3'。

[0670] 在一些实施方案中,靶序列包含第一核碱基,第一核碱基包含胞苷。在一些实施方案中,序列包含第二核碱基,第二核碱基包含脱氨胞苷。在一些实施方案中,序列包含第三核碱基,第三核碱基包含鸟嘌呤。在一些实施方案中,靶序列包含第四核碱基,第四核碱基包含胸腺嘧啶。在一些实施方案中,第二核碱基被与第四核碱基互补的第五核碱基替换,从而生成预期的编辑碱基对(例如A:T至G:C)。在一些实施方案中,第五核碱基是腺嘌呤。在一些实施方案中,至少5%的预期碱基对被编辑。

[0671] 在一些实施方案中,TadCBE可用于将胞苷脱氨为尿嘧啶。在一些情况下,脱氨导致剪接位点的引入和/或移除。在其他情况下,脱氨导致基因启动子中的突变的引入,例如以增加和/或减少与基因启动子可操作连接的基因的转录。在一些实施方案中,脱氨导致基因阻遏物中的突变的引入,例如以增加和/或减少与基因阻遏物可操作连接的基因的转录。

[0672] 在一些实施方案中,将核酸与本文所述的任何碱基编辑器(例如TadA-CD)或复合物接触是在受试者体内进行的(例如使用载体如AAV)。在一些情况下,受试者已被诊断出与点突变相关的疾病或病症。示例性疾病或病症包括HIV或镰状细胞病。在其他实施方案中,将核酸与本文所述的任何碱基编辑器或复合物接触的步骤是在体外或离体进行的。

[0673] 在一些实施方案中,第一核碱基是胞嘧啶。在一些实施方案中,第二核碱基是脱氨胞嘧啶或肌苷。在一些实施方案中,第三核碱基是鸟嘌呤。在一些实施方案中,第四核碱基是腺嘌呤。在一些实施方案中,该方法还包括用与第四核碱基互补的第五核碱基替换第二核碱基,从而生成预期的编辑碱基对(例如C:G至T:A)。在一些实施方案中,第五核碱基是胸腺嘧啶。在一些实施方案中,至少5%的预期碱基对被编辑。在一些实施方案中,至少10%、15%、20%、25%、30%、35%、40%、45%或50%的预期碱基对被编辑。

[0674] 在一些实施方案中,预期的编辑碱基对是PAM位点上游1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20个核苷酸。在一些实施方案中,预期的编辑碱基对是PAM位点下游。在一些实施方案中,预期的编辑碱基对是PAM位点下游1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20个核苷酸。在一些实施方案中,该方法不需要典型(例如NGG)PAM位点。在一些实施方案中,碱基编辑器包含接头。在一些实施方案中,接头长度为1-25个氨基酸。在一些实施方案中,接头长度为5-20个氨基酸。在一些实施方案中,接头长度为10、11、12、13、14、15、16、17、18、19或20个氨基酸。在一些实施方案中,靶区域包含靶窗口,其中靶窗口包含靶核碱基对。在一些实施方案中,靶窗口包含1-10个核苷酸。在一些实施方案中,靶窗口长度为1-9、1-8、1-7、1-6、1-5、1-4、1-3、1-2或1个核苷酸。在一些实施方案中,靶窗口长度为1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20个核苷酸。在一些实施方案中,预期的编辑碱基对位于靶窗口内。在一些实施方案中,靶窗口包含预期的编辑碱基对。在一些实施方案中,使用本文提供的任何腺嘌呤碱基编辑器进行该方法。在一些实施方案中,靶窗口是脱氨窗口。

[0675] 在一些方面,本公开提供了具有扩展的靶窗口的改进的腺嘌呤碱基编辑器。在一些实施方案中,所公开的碱基编辑器的靶窗口对应于靶序列的原间隔区位置3-8,其中原间隔区位置0对应于向导RNA序列的第一个连续核苷酸与靶序列互补的位置,或靶基因的转录起始位点的位置。在一些实施方案中,具有更宽靶窗口的碱基编辑器包含TadCBEa(如SEQ ID NO:19所示)。在一些实施方案中,具有更宽靶窗口的碱基编辑器包含TadCBEb(SEQ ID NO:20)。在一些实施方案中,具有更宽靶窗口的碱基编辑器包含TadCBEc(SEQ ID NO:21)。原间隔区位置0也可以指距离PAM最远的核苷酸位置。在所公开的碱基编辑器的其他实施方案中,碱基编辑器具有扩展的靶窗口,其相对于靶基因的转录起始位点的位置对应于靶序列的原间隔区位置3-14。在其他实施方案中,靶窗口对应于原间隔区位置4-11。在其他实施方案中,靶窗口对应于原间隔区位置8-14。在其他实施方案中,靶窗口对应于原间隔区位置9-14。在一些实施方案中,靶窗口位于基因(例如HBG、HBB或BC11A)中。

[0676] 在一些实施方案中,靶DNA序列包含与疾病或病症相关的序列。在一些实施方案

中,靶DNA序列包含与疾病或病症相关的点突变。在一些实施方案中,碱基编辑器(例如包含经进化的胞苷脱氨酶和Cas9结构域)或复合物的活性导致点突变的校正。在一些实施方案中,靶DNA序列包含与疾病或病症相关的C→T点突变,并且其中突变体C碱基的脱氨导致与疾病或病症不相关的序列。在一些实施方案中,靶DNA序列编码蛋白质,并且点突变位于密码子中,并且导致由突变体密码子编码的氨基酸与野生型密码子相比发生变化。在一些实施方案中,突变体C的脱氨导致由突变体密码子编码的氨基酸的变化。在一些实施方案中,突变体C的脱氨导致密码子编码野生型氨基酸。在一些实施方案中,接触是在受试者体内进行的。在一些实施方案中,受试者患有或已被诊断出患有疾病或病症。

[0677] 多路复用碱基编辑应用

[0678] 在一些方面,本公开提供了使用公开的胞嘧啶碱基编辑器同时编辑两个或更多个核酸靶位点的方法。在独特基因组基因座的多路复用碱基编辑中,与不同靶序列具有互补性的多个gRNA能够在多个(例如5、10、15、20、25个或更多个)靶序列中的每一个处同时或在单个迭代或周期内形成碱基编辑器-gRNA复合物。在所公开的多路复用碱基编辑方法的一些实施方案中,所公开的TadCBE可以靶向人细胞(如原代人T细胞)中的多个基因或多个染色体。

[0679] CRISPR/Cas系统的发现和广泛实施极大地扩展了基因组工程化的工具箱,并彻底改变了基础生物学研究、活体系统中的数据存储、农业科学和医学的未来前景。与之前的方法相比,基于CRISPR/Cas的基因组编辑器的一个优势在于通过使用多个向导RNA(gRNA)进行多路复用的能力。这不仅使得在单一细胞群体中筛选向导文库,还能够同时靶向多达六个独特的基因座。然而,当与单一向导转染相比时,每个位点的编辑效率往往会降低。

[0680] 本公开提供了碱基编辑的方法,包括:将核酸分子(例如DNA)与多个复合物接触,其中每个复合物包含碱基编辑器和与该碱基编辑器的napDNAbp结构域结合的向导RNA(gRNA),其中多个复合物中的至少两个复合物各自包含独特的gRNA,该gRNA包含至少10个连续核苷酸的向导序列,该向导序列与细胞基因组DNA中的独特靶序列互补。在某些实施方案中,细胞是真核细胞,例如哺乳动物细胞。在某些实施方案中,细胞是人细胞。在某些实施方案中,多个公开的碱基编辑器-gRNA复合物在真核细胞(例如哺乳动物细胞)内的多个靶基因座处进行同时编辑(即,在单一迭代内)。

[0681] 在一些实施方案中,这些多路复用编辑方法的任何靶序列包含基因组基因座。在一些实施方案中,多个靶序列包含独特的基因组基因座。在一些实施方案中,至少一个靶序列包含HBG启动子或BCL11A增强子中的序列。在一些实施方案中,至少一个靶序列包含CXCR4或CCR5基因中的序列。因此,在一些实施方案中,多路复用碱基编辑方法用于在CXCR4和CCR5基因中同时或在单个迭代或周期内安装C至T突变(参见实施例5)。

[0682] 治疗方法

[0683] 本公开提供了治疗被诊断出患有与点突变相关或由点突变引起的疾病的受试者的方法,该点突变可以通过本文提供的融合蛋白(例如,包含本文所述的任何Nme2Cas9变体和脱氨酶的碱基编辑器融合蛋白)校正。例如,在一些实施方案中,提供了一种方法,包括向患有此类疾病(例如与点突变相关的癌症等疾病)的受试者施用有效量的碱基编辑器和与该碱基编辑器形成复合物的gRNA,该方法校正点突变或将失活突变引入疾病相关基因中。在一些实施方案中,提供了一种方法,包括向患有此类疾病(例如与点突变相关的癌症)的

受试者施用有效量的碱基编辑器-gRNA复合物,该复合物校正点突变或将失活突变引入疾病相关基因中。本文还提供了包括向受试者施用一种或多种载体的方法,该载体含有表达碱基编辑器和与该碱基编辑器形成复合物的gRNA的核苷酸序列。

[0684] 在一些实施方案中,疾病是增殖性疾病。在一些实施方案中,疾病是遗传性疾病。在一些实施方案中,疾病是肿瘤性疾病。在一些实施方案中,疾病是代谢性疾病。本领域技术人员将知晓可以通过校正点突变或将失活突变引入疾病相关基因来治疗的其他疾病,并且本公开在此方面不受限制。

[0685] 本公开提供了治疗另外的疾病或病症的方法,例如与点突变相关或由点突变引起的疾病或病症,该点突变可以通过碱基编辑校正。本文描述了一些此类疾病,基于本公开,可以使用本文提供的策略和融合蛋白(例如碱基编辑器)治疗的另外的合适疾病对于本领域技术人员将是显而易见的。示例性的合适疾病和病症列于下文。应理解,相应序列中特定位置或残基的编号取决于所使用的特定蛋白质和编号方案。编号可以不同,例如,在成熟蛋白质的前体和成熟蛋白质本身中,并且物种之间序列的差异可能会影响编号。本领域技术人员将能够通过本领域众所周知的方法(例如,通过序列比对和同源残基的确定)鉴定任何同源蛋白质和相应编码核酸中的相应残基。示例性的合适疾病和病症包括但不限于镰状细胞病、早衰症、囊性纤维化和鸟氨酸转氨甲酰酶(OTC)缺乏症。在一些实施方案中,所公开的组合物和方法可以适用于编辑镰状细胞病中的临床相关点突变,如HBB^S(Makassar等位基因)。

[0686] 使用一个或多个胞苷或腺嘌呤碱基编辑器通过校正点突变或将失活突变引入疾病相关基因来治疗疾病、病症或病况的示例性方法公开于以下国际公布号中:WO 2021/222318,2021年11月4日公布;WO 2021/183693,2021年9月16日公布;WO 2021/158999,2021年8月12日公布;WO2020/051360,2020年3月12日公布;以及WO 2019/079347,2019年4月25日公布,其中每篇通过引用并入本文。

[0687] 在一些方面,本公开提供了使用本文所述的任一种融合蛋白(例如碱基编辑器)和将该碱基编辑器靶向核酸分子中的靶C:G碱基对的向导RNA用于制备碱基编辑试剂盒的用途,其中碱基编辑包括在适合将C:G核碱基对中的胞嘧啶(C)置换为胸腺嘧啶(T)的条件下,将核酸分子与碱基编辑器和gRNA接触。在这些用途的一些实施方案中,核酸分子是双链DNA分子。在一些实施方案中,接触步骤诱导双链DNA在靶区域处的分离。在一些实施方案中,接触步骤还包括在双链DNA的一条链产生切口,其中这一条链包含未突变的链,该未突变的链包含靶C:G核碱基对中的G。

[0688] 在所描述的用途的一些实施方案中,接触步骤在体外进行。在其他实施方案中,接触步骤在体内进行。在一些实施方案中,接触步骤在受试者(例如人受试者或非人动物受试者)中进行。在一些实施方案中,接触步骤在细胞(如人细胞或非人动物细胞)中进行。

[0689] 本公开还提供了本文所述的任一种融合蛋白作为药物的用途。本公开还提供了本文所述的融合蛋白和gRNA的复合物中的任一种作为药物的用途。

[0690] 药物组合物

[0691] 本公开的一些方面涉及包含本文所述的任何腺苷至胞苷脱氨酶、碱基编辑器或碱基编辑器-gRNA复合物的药物组合物。本公开的其他方面涉及包含编码本文所述的TadA-CD脱氨酶、碱基编辑器或碱基编辑器-gRNA复合物的核酸片段的任何多核苷酸或载体的药物

组合物。本公开还提供了包含本文所述的rAAV载体、双rAAV载体和核糖核蛋白的颗粒的药物组合物。

[0692] 本文使用的术语“药物组合物”是指配制用于药物用途的组合物。在一些实施方案中,药物组合物还包含药学上可接受的载剂。在一些实施方案中,药物组合物包含另外的药剂(例如用于特异性递送、延长半衰期或其他治疗性化合物)。

[0693] 在一些实施方案中,本文所述的任何碱基编辑器、gRNA和/或复合物作为药物组合物的一部分提供。在一些实施方案中,药物组合物包含本文提供的任何碱基编辑器。在一些实施方案中,药物组合物包含本文提供的任何复合物。在一些实施方案中,药物组合物包含gRNA、碱基编辑器和药学上可接受的赋形剂。药物组合物可以任选地包含一种或多种另外的治疗活性物质。

[0694] 在一些实施方案中,本文提供的组合物被配制用于递送至受试者,例如人受试者,以在受试者体内实现靶向基因组修饰。在一些实施方案中,从受试者获取细胞并与本文提供的任何药物组合物接触。在一些实施方案中,从受试者移除并离体与药物组合物接触的细胞重新引入受试者,任选地在细胞中实现或检测到期望的基因组修饰后。递送包含核酸酶的药物组合物的方法是已知的,并且描述于例如美国专利号6,453,242;6,503,717;6,534,261;6,599,692;6,607,882;6,689,558;6,824,978;6,933,113;6,979,539;7,013,219;7,163,824,9,526,784;9,737,604;以及美国专利公布号2018/0127780(2018年5月10日公布)和2018/0236081(2018年8月23日公布),其中所有的公开内容通过引用以其整体并入本文。尽管本文提供的药物组合物的描述主要针对适合施用至人的药物组合物,但本领域技术人员将理解,此类组合物通常适合施用至各种动物或有机体。对适合施用至人的药物组合物的修饰以使该组合物适合施用至各种动物是本领域熟知的,并且普通兽医药理学家可以设计并/或执行此类修饰,仅需进行普通实验(如果有)。施用药物组合物的受试者预期但不限于人和/或其他灵长类动物;哺乳动物、家养动物、宠物和商业相关哺乳动物,如牛、猪、马、绵羊、猫、狗、小鼠和/或大鼠;和/或鸟类,包括商业相关鸟类,如鸡、鸭、鹅和/或火鸡。

[0695] 本文所述的药物组合物的制剂可以通过药理学领域中已知或未来开发的任何方法制备。通常,此类制备方法包括将活性成分与赋形剂和/或一种或多种其他辅助成分缔合的步骤,然后(如有必要和/或期望)将产品成型和/或包装成期望的单剂量或多剂量单位。

[0696] 药物制剂还可以包含药学上可接受的赋形剂,本文使用的“药学上可接受的赋形剂”包括任何和所有溶剂、分散介质、稀释剂或其他液体媒介物、分散或悬浮助剂、表面活性剂、等渗剂、增稠或乳化剂、防腐剂、固体粘合剂、润滑剂等,以适应期望的特定剂型。Remington's The Science and Practice of Pharmacy,21st Edition,A.R.Gennaro(Lippincott,Williams&Wilkins,Baltimore,MD,2006;通过引用以其整体并入本文)公开了用于配制药物组合物的各种赋形剂及用于其制备的已知技术。针对用于生产包含核酸酶的药物组合物的另外的合适方法、试剂、赋形剂和溶剂,另见2010年11月2日提交的PCT申请PCT/US2010/055131(公布号WO 2011/053982,2011年5月5日公布),通过引用以其整体并入本文。除非任何常规赋形剂介质与物质或其衍生物不相容,如通过产生任何不期望生物效应或以有害方式与药物组合物的任何其他组分相互作用,否则其使用被考虑在本公开的范围内。

[0697] 在一些方面,本公开提供了包含多个本文所述的任何碱基编辑器和gRNA的药物组合物,其中多个碱基编辑器中的至少五个各自结合独特的gRNA,以及药学上可接受的赋形剂。

[0698] 本文中使用的术语“药学上可接受的载剂”是指药学上可接受的材料、组合物或媒介物,如液体或固体填充剂、稀释剂、赋形剂、制造助剂(例如润滑剂、滑石粉、硬脂酸镁、硬脂酸钙或硬脂酸锌)或溶剂封装材料,涉及将化合物从机体的一个部位(例如递送部位)携带或运输到另一个部位(例如器官、组织或机体的一部分)。在与制剂的其他成分相容且对受试者的组织无害(例如生理相容、无菌、生理pH等)的意义上,药学上可接受的载剂是“可接受的”。可用作药学上可接受的载剂的材料的一些实例包括:(1)糖类,如乳糖、葡萄糖和蔗糖;(2)淀粉,如玉米淀粉和马铃薯淀粉;(3)纤维素及其衍生物,如羧甲基纤维素钠、甲基纤维素、乙基纤维素、微晶纤维素和醋酸纤维素;(4)粉末状黄芪胶;(5)麦芽;(6)明胶;(7)润滑剂,如硬脂酸镁、十二烷基硫酸钠和滑石粉;(8)赋形剂,如可可脂和栓剂蜡;(9)油类,如花生油、棉籽油、红花油、芝麻油、橄榄油、玉米油和大豆油;(10)二醇类,如丙二醇;(11)多元醇类,如甘油、山梨醇、甘露醇和聚乙二醇(PEG);(12)酯类,如油酸乙酯和月桂酸乙酯;(13)琼脂;(14)缓冲剂,如氢氧化镁和氢氧化铝;(15)海藻酸;(16)无热原水;(17)等渗盐水;(18)林格氏溶液;(19)乙醇;(20)pH缓冲溶液;(21)聚酯、聚碳酸酯和/或聚酐;(22)填充剂,如多肽和氨基酸;(23)血清组分,如血清白蛋白、HDL和LDL;(22)C2-C12醇类,如乙醇;以及(23)用于药物制剂中的其他无毒相容物质。湿润剂、着色剂、释放剂、包衣剂、甜味剂、调味剂、香料、防腐剂和抗氧化剂也可以存在于制剂中。术语如“赋形剂”、“载剂”、“药学上可接受的载剂”等在本文中可互换使用。

[0699] 在一些实施方案中,药物组合物被配制用于递送至受试者,例如用于基因编辑。本文所述的药物组合物的合适施用途径包括但不限于:外部、皮下、透皮、皮内、病灶内、关节内、腹膜内、膀胱内、透黏膜、牙龈、牙内、耳蜗内、经鼓膜、器官内、硬膜外、鞘内、肌肉内、静脉内、血管内、骨内、眼周、肿瘤内、脑内和脑室内施用。

[0700] 在一些实施方案中,本文所述的药物组合物局部施用于患病部位(例如肿瘤部位)。在一些实施方案中,本文所述的药物组合物通过注射、通过导管方式、通过栓剂方式或或通过植入物方式施用至受试者,植入物可以是多孔、无孔或凝胶状材料,包括膜(如硅胶膜)或纤维。

[0701] 在其他实施方案中,本文所述的药物组合物以控释系统递送。在一个实施方案中,可以使用泵(参见例如Langer,1990,Science 249:1527-1533;Sefton,1989,CRC Crit.Ref.Biomed.Eng.14:201;Buchwald et al.,1980,Surgery 88:507;Saudek et al.,1989,N.Engl.J.Med.321:574)。在另一个实施方案中,可以使用聚合材料(参见例如Medical Applications of Controlled Release(Langer and Wise eds.,CRC Press,Boca Raton,Fla.,1974);Controlled Drug Bioavailability,Drug Product Design and Performance(Smolen and Balleds.,Wiley,New York,1984);Ranger and Peppas,1983,Macromol.Sci.Rev.Macromol.Chem.23:61。另见Levy et al.,1985,Science 228:190;During et al.,1989,Ann.Neurol.25:351;Howard et al.,1989,J.Neurosurg.71:105)。其他控释系统在例如Langer(同上)中讨论。

[0702] 在一些实施方案中,药物组合物根据常规程序配制为适合静脉内或皮下施用至受

试者(例如人)的组合物。在一些实施方案中,用于通过注射施用的药物组合物是无菌等渗水缓冲液中的溶液。必要时,药物还可以包括增溶剂和局部麻醉剂(如利多卡因)以减轻注射部位的疼痛。通常,成分单独供应或以单位剂量形式混合在一起供应,例如作为在标明活性剂的量的密封容器如安瓿或小袋中的干燥的冻干粉末或无水浓缩物。如果药物通过输注施用,其可以与含有无菌药用级水或盐水的输液瓶一起提供。如果药物组合物通过注射施用,可以提供无菌注射用水或盐水的安瓿,以便成分可以在施用前混合。

[0703] 用于全身施用的药物组合物可以是液体,例如无菌盐水、乳酸林格氏液或汉克氏溶液。此外,药物组合物可以是固体形式,并在使用前立即重新溶解或悬浮。冻干形式也在考虑范围内。

[0704] 药物组合物可以包含在脂质颗粒或囊泡中,如脂质纳米颗粒(LNP)、脂质体或微晶,这也适合用于肠胃外施用。颗粒可以是任何合适的结构,如单层或多层,只要组合物包含在其中。化合物可以包埋在"稳定的质粒-脂质颗粒"(SPLP)中,SPLP含有融合脂质二油酰磷脂酰乙醇胺(DOPE)、低水平(5-10mol%)的阳离子脂质并通过聚乙二醇(PEG)涂层稳定(Zhang Y.P.et al.,Gene Ther.1999,6:1438-47)。带正电荷的脂质,如N-[1-(2,3-二油酰氧基)丙基]-N,N,N-三甲基-铵甲基硫酸盐或"DOTAP",特别优选用于此类颗粒和囊泡。此类脂质颗粒的制备是众所周知的。参见例如美国专利号4,880,635;4,906,477;4,911,928;4,917,951;4,920,016;和4,921,757;其中每篇通过引用并入本文。

[0705] 本文所述的药物组合物可以例如作为单位剂量施用或包装。术语"单位剂量"在提及本公开的药物组合物使用时,指的是适合作为用于受试者的单一剂量的物理离散单位,每个单位包含预定量的活性材料,计算为与所需稀释剂(即载剂或媒介物)缔合以产生期望的治疗效果。

[0706] 此外,药物组合物可以作为药物试剂盒提供,包括:(a)含有冻干形式的本发明化合物的容器,以及(b)含有用于注射的药学上可接受的稀释剂(例如无菌水)的第二容器。药学上可接受的稀释剂可用于重构或稀释本发明的冻干化合物。任选地,与此类容器相关联的可以是监管药品或生物产品的制造、使用或销售的政府机构规定形式的通知,该通知反映了该机构对制造、使用或销售用于人类施用的批准。

[0707] 在另一个方面,含有用于治疗上述疾病的材料的制品也包括在内。在一些实施方案中,该制品包含容器和标签。合适的容器包括例如瓶子、小瓶、注射器和试管。容器可以由多种材料制成,如玻璃或塑料。在一些实施方案中,容器内装有对治疗本文所述疾病有效的组合物,并可以具有无菌接入端口。例如,容器可以是静脉输液袋或具有可被皮下注射针穿刺的塞子的小瓶。组合物中的活性剂是本发明的化合物。在一些实施方案中,容器上或与容器相关联的标签表明该组合物用于治疗所选疾病。该制品还可以包括包含药学上可接受缓冲液(如磷酸盐缓冲盐水、林格氏溶液或右旋糖溶液)的第二容器。它还可以包括从商业和用户角度期望的其他材料,包括其他缓冲液、稀释剂、过滤器、针头、注射器和使用说明的包装插页。

[0708] 递送方法

[0709] 本公开提供了将本文所述的胞嘧啶碱基编辑器(例如,以本文所述的经进化的碱基编辑器、或编码其的载体或构建体的形式)递送到细胞中的方法。此类方法可以涉及使用多个复合物转导(例如,经由转染)细胞,每个复合物包含碱基编辑器和gRNA分子。在一些实

施方案中,gRNA与碱基编辑器的napDNAbp结构域(例如nCas9结构域)结合。在一些实施方案中,每个gRNA包含与靶序列互补的至少10个连续核苷酸(例如10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29或30个连续核苷酸)的向导序列。在某些实施方案中,这些方法涉及转染核酸构建体(例如质粒和mRNA构建体或重组mRNA构建体),这些构建体各自(或共同)编码碱基编辑器和gRNA分子的复合物的组分。在某些实施方案中,任何公开的碱基编辑器和gRNA作为蛋白质:RNA复合物(如核糖核蛋白(RNP)复合物)施用。在一些实施方案中,任何公开的碱基编辑器作为mRNA构建体与gRNA分子一起施用。在一些实施方案中,通过电穿孔或脂质体转染(例如使用Lipofectamine®)实现向细胞的施用。

[0710] 在某些实施方案中,编码碱基编辑器的核酸构建体(例如mRNA构建体)与编码gRNA分子的构建体分开转染到细胞中。在某些实施方案中,这些组分在单个构建体上编码并一起转染。在其他实施方案中,本文所公开的方法涉及将包含碱基编辑器和gRNA分子的复合物引入细胞(体内或体外),该复合物已在这些细胞外部表达和克隆。在一些实施方案中,所公开的方法涉及以100ng的量引入编码碱基编辑器的DNA构建体。

[0711] 在一些方面,本发明提供了包括将一种或多种多核苷酸(如本文所述的一种或多种载体)、其一种或多种转录物和/或从其转录的一种或多种蛋白质递送到宿主细胞的方法。在一些方面,本发明还提供了通过此类方法产生的细胞,以及包含此类细胞或由此类细胞产生的生物体(如动物、植物或真菌)。在一些实施方案中,与向导序列组合(并任选地复合)的如本文所述的碱基编辑器递送到细胞中。

[0712] 在一些实施方案中,提供的递送方法包括核转染、显微注射、生物弹道法、病毒体、脂质体、免疫脂质体、聚阳离子或脂质:核酸缀合物、裸DNA、人工病毒颗粒和增强DNA摄取的药剂。

[0713] 在另一个方面,本公开公开了包含任一种当前公开的载体的药物组合物。在某些实施方案中,药物组合物还包含药学上可接受的赋形剂。在某些实施方案中,药物组合物还包含脂质和/或聚合物。在某些实施方案中,脂质和/或聚合物是阳离子的。此类脂质颗粒的制备是众所周知的。参见例如美国专利号4,880,635;4,906,477;4,911,928;4,917,951;4,920,016;4,921,757;和9,737,604,其中每篇通过引用并入本文。

[0714] 核酸递送的示例性方法包括脂质体转染、核转染、电穿孔(例如MaxCyte电穿孔)、稳定基因组整合(例如piggybac)、显微注射、生物弹道法、病毒体、脂质体、免疫脂质体、聚阳离子或脂质:核酸缀合物、裸DNA、人工病毒颗粒和增强DNA摄取的药剂。脂质体转染描述于例如美国专利号5,049,386、4,946,787;和4,897,355),并且脂质体转染试剂商业销售(例如Transfectam™、Lipofectin™和SF Cell Line 4D-Nucleofector X Kit™(Lonza))。适用于高效受体识别脂质体转染多核苷酸的阳离子和中性脂质包括Feigner,WO 91/17424;WO 91/16024的脂质。递送可以是到细胞(例如体外或离体施用)或靶组织(例如体内施用)的。

[0715] 在所公开方法的某些实施方案中,编码碱基编辑器的构建体与编码gRNA的构建体分开转染到细胞中。在某些实施方案中,这些组分在单个构建体上编码并一起转染。在特定实施方案中,这些编码碱基编辑器和gRNA的单个构建体可以迭代地转染到细胞中,每次迭代与一个靶序列亚组相关联。在特定实施方案中,这些单个构建体可以在几天时间段内转染到细胞中。在其他实施方案中,它们可以在几小时时间段内转染到细胞中。在其他实施方

案中,它们可以在几周时间段内转染到细胞中。

[0716] 在所公开的方法中,靶细胞可以在转染后与碱基编辑器-gRNA复合物孵育两天或48小时,以实现多路复用碱基编辑。靶细胞可以在转染后孵育30小时、40小时、54小时、60小时或72小时。靶细胞可以在转染后与碱基编辑器-gRNA复合物孵育四天、五天、七天、九天、十一天、十三天或更长时间。

[0717] 脂质:核酸复合物,包括靶向脂质体如免疫脂质复合物的制备,是本领域技术人员熟知的(参见例如Crystal,Science 270:404-410(1995);Blaese et al.,Cancer Gene Ther.2:291-297(1995);Behr et al.,Bioconjugate Chem.5:382-389(1994);Remy et al.,Bioconjugate Chem.5:647-654(1994);Gao et al.,Gene Therapy 2:710-722(1995);Ahmad et al.,Cancer Res.52:4817-4820(1992);美国专利号4,186,183、4,217,344、4,235,871、4,261,975、4,485,054、4,501,728、4,774,085、4,837,028和4,946,787)。

[0718] 在其他实施方案中,本文提供的递送方法和载体是RNP复合物。碱基编辑器的RNP递送显著增加了碱基编辑的DNA特异性。RNP递送碱基编辑器导致靶向和非靶向DNA编辑的解耦。RNP递送消除了非重复位点的脱靶编辑,同时保持了与质粒递送相当的中靶编辑,并即使在高度重复的VEGFA位点2也大大降低了脱靶DNA编辑。参见H.A.et al.,Improving the DNA specificity and applicability of base editing through protein engineering and protein delivery,Nat.Commun.8,15790(2017)、美国专利号9,526,784(2016年12月27日发布)以及美国专利号9,737,604(2017年8月22日发布),其中每篇通过引用并入本文。在一些实施方案中,RNP复合物以无DNA的工程化病毒样颗粒(eVLP)递送,该颗粒高效包装和递送碱基编辑器RNP。参见Banskota et al.,Cell 185,250-265,Jan.2022,其通过引用并入本文。

[0719] 基于RNA或DNA病毒的递送系统用于递送核酸的用途利用高度进化的过程,用于将病毒靶向体内的特定细胞并将病毒有效载荷运输到细胞核。病毒载体可以直接施用于患者(体内),或者它们可以用于体外处理细胞,并且修饰后的细胞可以任选性地施用至患者(离体)。常规的基于病毒的系统可以包括逆转录病毒、慢病毒、腺病毒、腺相关病毒和单纯疱疹病毒载体用于基因转移。使用逆转录病毒、慢病毒和腺相关病毒基因转移方法整合宿主基因组是可能的,通常导致插入的转基因长期表达。此外,已在许多不同的细胞类型和靶组织中观察到高转导效率。

[0720] 通过掺入外来包膜蛋白,可以改变病毒的趋向性,从而扩大靶细胞的潜在靶标群体。慢病毒载体是能够转导或感染非分裂细胞的逆转录病毒载体,并通常产生高病毒滴度。因此,逆转录病毒基因转移系统的选择将取决于靶组织。逆转录病毒载体包含顺式作用的长末端重复序列(LTR),具有包装最长达6-10kb的外源序列的能力。最小的顺式作用LTR足以进行载体的复制和包装,然后用于将治疗性基因整合到靶细胞中以提供永久性转基因表达。广泛使用的逆转录病毒载体包括基于鼠白血病病毒(MuLV)、长臂猿白血病病毒(GaLV)、猴免疫缺陷病毒(SIV)、人类免疫缺陷病毒(HIV)及其组合的载体(参见例如Buchscher et al.,J.Virol.66:2731-2739(1992);Johann et al.,J.Virol.66:1635-1640(1992);Sommnerfelt et al.,Virol.176:58-59(1990);Wilson et al.,J.Virol.63:2374-2378(1989);Miller et al.,J.Virol.65:2220-2224(1991);PCT/US94/05700)。在优选瞬时表达的应用中,可以使用基于腺病毒的系统。基于腺病毒的载体能够在许多细胞类型中实现

非常高的转导效率,并且不需要细胞分裂。使用此类载体,已经获得了高滴度和高水平的表达。该载体可以在相对简单的系统中大量生产。腺相关病毒("AAV")载体也可用于用靶核酸转导细胞,例如在核酸和肽的体外生产中,以及用于体内和离体基因治疗程序(参见例如West et al.,Virology 160:38-47(1987);美国专利号4,797,368;WO 93/24641;Kotin,Human Gene Therapy 5:793-801(1994);Muzyczka,J.Clin.Invest.94:1351(1994)。重组AAV载体的构建在许多出版物中有所描述,包括美国专利号5,173,414;Tratschin et al.,Mol.Cell.Biol.5:3251-3260(1985);Tratschin,et al.,Mol.Cell.Biol.4:2072-2081(1984);Hermonat&Muzyczka,PNAS 81:6466-6470(1984);以及Samulski et al.,J.Virol.63:03822-3828(1989)。

[0721] 包装细胞通常用于形成能够感染宿主细胞的病毒颗粒。此类细胞包括包装腺病毒的293细胞,以及包装逆转录病毒的ψ2细胞或PA317细胞。用于基因疗法的病毒载体通常通过产生将核酸载体包装成病毒颗粒的细胞系来生成。载体通常含有包装和随后整合到宿主中所需的最小病毒序列,其他病毒序列被待表达的多核苷酸的表达盒所取代。缺失的病毒功能通常由包装细胞系以反式提供。例如,用于基因疗法的AAV载体通常仅具有来自AAV基因组的ITR序列,这些序列是包装和整合到宿主基因组中所必需的。病毒DNA在细胞系中包装,该细胞系含有编码其他AAV基因(即rep和cap)的辅助质粒,但缺乏ITR序列。细胞系也可以被作为辅助病毒的腺病毒感染。辅助病毒促进AAV载体的复制和来自辅助质粒的AAV基因的表达。由于缺乏ITR序列,辅助质粒不会被大量包装。使用腺病毒的污染可以通过例如热处理来降低,腺病毒对其比AAV更敏感。本领域技术人员已知用于核酸递送的另外的方法。参见US2003/0087817(2003年5月8日公布)、国际专利申请号WO 2016/205764(2016年12月22日公布)、国际专利申请号WO 2018/071868(2018年4月19日公布)、美国专利公布号2018/0127780(2018年5月10日公布)以及国际公布号WO2020/236982(2020年11月26日公布),其中每篇的公开内容通过引用并入本文。

[0722] 使用rAAV载体的碱基编辑器递送

[0723] 本公开的TadCBE含有经进化的胞苷脱氨酶结构域,该结构域含有单个脱氨酶,即脱氨酶单体(如TadA-CDa单体)。在一些实施方案中,TadA-CD单体长度为约166个氨基酸。

[0724] 在一些实施方案中,任何公开的尺寸减小的TadA-CD变体(例如TadA-CDa、TadA-CDb、TadA-CDc和TadA-CDd)与单AAV递送兼容,如Davis et al.,Nat Biomed Eng.2022 Jul 28中所述,其通过引用并入本文。每个变体分别含有TadA-CD腺苷至胞苷脱氨酶以及SauriCas9、SaKKH-Cas9和SaCas9的切口酶变体。在任何这些公开的碱基编辑器中,分别可以使用SauriCas9、SaKKH-Cas9、SaCas9、CjCas9和Nme2Cas9的野生型或切口酶变体。

[0725] 本公开的递送方法的方面涉及使用重组腺相关病毒载体递送任何公开的核酸分子。本公开的rAAV颗粒包含封装在病毒衣壳蛋白中的rAAV载体(即rAAV的重组基因组)。参见美国专利公布号2018/0127780(2018年5月10日公布)以及PCT公布号WO 2020/236982(2020年11月26日公布),其中每篇的公开内容通过引用并入本文。

[0726] 在一些实施方案中,AAV核酸载体是单链的。在一些实施方案中,AAV核酸载体是自互补的。在各种实施方案中,本公开的rAAV载体不含有任何内含肽。

[0727] 在一些实施方案中,促进整合的病毒序列包括反向末端重复(ITR)序列。在一些实施方案中,核酸分子在每一侧侧翼为ITR序列。在一些实施方案中,核酸载体还包含编码如

本文所述的AAV Rep蛋白的区域,该区域可以包含在侧翼为ITR的区域内或区域外。ITR序列可以衍生自任何AAV血清型(例如1、2、3、4、5、6、7、8、9或10),或可以衍生自多于一个血清型。在一些实施方案中,ITR序列衍生自AAV8或AAV9。在一些实施方案中,在包装任何公开的rAAV颗粒的方法中,提供了包含编码Rep蛋白和/或Cap(衣壳)蛋白的区域的核酸质粒如辅助质粒。

[0728] 在各种实施方案中,任何公开的碱基编辑器(或融合蛋白)构建体可以被工程化用于在一个或多个AAV载体中递送。任何公开的AAV载体可以包含5'和3'反向末端重复(ITR),这些ITR位于编码任何公开的碱基编辑器的多核苷酸(或构建体)侧翼。在一些实施方案中,任何碱基编辑器构建体可以被工程化用于在单个rAAV载体中递送。在一些实施方案中,任何公开的碱基编辑器构建体具有4.9千碱基或更少的长度,并因此可以包装到单一AAV载体中,同时侧翼为ITR。在一些实施方案中,任何公开的碱基编辑器构建体具有在5'和3'ITR之间约4.65kb、约4.70kb、约4.725kb、约4.75kb、约4.80kb、约4.825kb、约4.85kb或约4.90kb之间的长度。在一些实施方案中,任何公开的碱基编辑器构建体具有在4.7kb和4.9kb之间的长度,例如约4.8kb。

[0729] 在一些实施方案中,任何公开的碱基编辑器构建体或含有编码碱基编辑器的多核苷酸的rAAV载体包含编码碱基编辑器的第一片段,并进一步包含编码向导RNA(如单一向导RNA)的第二核酸片段。在一些实施方案中,该gRNA编码(第二)核酸片段的方向相对于编码碱基编辑器的片段的方向是反向的。在一些实施方案中,第一核酸片段由第一启动子可操作地控制,第二核酸片段由第二启动子(例如U6启动子)可操作地控制。在一些实施方案中,第一启动子与第二启动子不同。本公开提供了包含任何上述碱基编辑器构建体的单AAV载体。

[0730] 本公开提供了包含任何公开的AAV载体的重组AAV颗粒。这些rAAV颗粒可以包含AAV载体和衣壳蛋白。衣壳蛋白可以是任何血清型。

[0731] 因此,本文提供的与任何公开的用途、方法和组合物相关的rAAV颗粒可以是任何血清型,包括任何衍生物或假型(例如1、2、3、4、5、6、7、8、9、10、11、12、13、2/1、2/5、2/8、2/9、3/1、3/5、3/8或3/9)。rAAV可以包含待递送至细胞的遗传负载(即表达感兴趣基因的重组核酸载体,如由rAAV携带到细胞中的完整碱基编辑器)。rAAV可以是嵌合的。

[0732] 任何公开的碱基编辑器都可以通过单一AAV载体递送。在一些实施方案中,AAV载体包含尺寸最小化的碱基编辑器和调控组分,使载体具有在单一AAV载体的4.7kb-4.9kb包装容量内的长度。在一些实施方案中,单一AAV载体含有第一核酸片段,其包含:(i)5'ITR;(ii)第一核酸片段,其包含编码可操作地连接到第一启动子的碱基编辑器的序列,其中碱基编辑器包含核酸可编程DNA结合蛋白(napDNAbp)结构域和脱氨酶结构域;以及聚腺苷酸化(polyA)信号;(iii)第二核酸片段,其编码可操作地连接到第二启动子的向导RNA(gRNA);以及(iv)3'ITR,其中5'ITR和3'ITR之间的长度小于约4.90kb。在一些实施方案中,rAAV载体基本上由组分(i)-(iv)组成。在一些实施方案中,通过单一AAV载体递送的碱基编辑器含有作为紧凑蛋白的napDNAbp结构域,如金黄色葡萄球菌Cas9(SaCas9)、脑膜炎奈瑟菌2Cas9(Nme2Cas9)、空肠弯曲菌Cas9(CjCas9)或耳葡萄球菌(SauriCas9)结构域,或其变体。

[0733] 所公开的递送方法的一些方面涉及在单一AAV载体中编码编辑器,并进一步编码

向导RNA,用于在单个rAAV颗粒中包装。因此,在一些实施方案中,任何公开的碱基编辑器都可以在单一AAV载体中编码,而无需使用任何分裂点或内含肽。还描述了若干其他特殊的考虑因素,以解释碱基编辑的独特特征,包括优化第二切口位点靶标和正确包装碱基编辑器到病毒载体(包括慢病毒和rAAV)中。

[0734] 因此,本公开提供了包含编码任何公开的碱基编辑器的表达构建体的rAAV载体和rAAV载体颗粒。在示例性实施方案中,任何公开的碱基编辑器都以单一rAAV颗粒递送到一个或多个细胞中。

[0735] 在一些方面,本公开提供了含有多个任何公开的rAAV颗粒的组合物。在一些方面,本公开提供了含有多个任何公开的rAAV颗粒的宿主细胞。在一些实施方案中,宿主细胞是哺乳动物细胞,例如人细胞或啮齿动物细胞。在示例性实施方案中,宿主细胞是人细胞。在其他实施方案中,宿主细胞是酵母细胞、植物细胞或细菌细胞。

[0736] 在其他实施方案中,碱基编辑器可以在分裂点处分裂,并作为完整/完全碱基编辑器的两个半部提供。这两个半部可以递送到细胞中(例如作为表达的蛋白质或在单独的表达载体上),并且一旦在细胞内接触,这两个半部通过每个碱基编辑器半部上的内含肽的自剪接作用形成完整的碱基编辑器。可以将分裂内含肽序列工程化到编码的碱基编辑器的每个半部中,以促进它们在细胞内的反式剪接,并同时恢复完整的功能性TadCBE。这些基于分裂内含肽的方法可以克服体内递送的若干障碍。例如,编码一些碱基编辑器的DNA大于重组AAV(rAAV)的包装限制,因此需要不同的解决方案。一种此类解决方案是将编辑器与包装到两个单独的rAAV颗粒中的分裂内含肽对融合,当两个单独的rAAV颗粒共同递送到细胞时,重建功能性编辑器蛋白。在此类实施方案中,碱基编辑器可以在分裂点处分裂为两个半部。这两个半部可以递送到细胞中(例如作为表达的蛋白质或在单独的表达载体上),并且一旦在细胞内接触,这两个半部通过每个碱基编辑器半部上的内含肽的自剪接作用形成完整的碱基编辑器。可以将分裂内含肽序列工程化到编码的碱基编辑器的每个半部中,以促进它们在细胞内的反式剪接,以及完整的功能性TadCBE的伴随恢复。

[0737] 本领域已知将任何公开的rAAV颗粒以及包含rAAV颗粒的组合物和宿主细胞递送到靶细胞或靶组织的方法。在一些实施方案中,任何公开的rAAV颗粒、宿主细胞或组合物被递送到受试者,如哺乳动物受试者。在一些实施方案中,rAAV颗粒被递送到人受试者。

[0738] 在一些实施方案中,所公开的rAAV颗粒和组合物以单次注射如单次全身注射施用至受试者。在一些实施方案中,所公开的rAAV颗粒和组合物以多次注射施用于受试者。已知rAAV颗粒在几天内转导靶组织,但通常允许三到四周完成转导、基因组整合和从细胞中清除。因此,在一些方面,任何公开的rAAV颗粒或组合物被施用于受试者持续三周的时间段。在一些方面,任何公开的rAAV颗粒或组合物被施用于受试者持续三到四周之间的时间段。

[0739] 在一些实施方案中,任何公开的rAAV颗粒或组合物以约10^{15}、约10^{14}、约10^{13}、约10^{12}、约10^{11}或少于约10^{11}个载体基因组(vg)/kg体重的治疗有效量施用于受试者或靶组织。在一些实施方案中,rAAV颗粒以10^{15}vg/kg和10^{14}vg/kg之间、10^{14}vg/kg和10^{13}vg/kg之间、10^{13}vg/kg和10^{12}vg/kg之间、10^{12}vg/kg和10^{11}vg/kg之间或10^{12}vg/kg和10^{11}vg/kg之间的量施用。在一些实施方案中,rAAV颗粒以10^{14}vg/kg和10^{11}vg/kg之间的量施用。在一些实施方案中,任何公开的rAAV颗粒或组合物以比双AAV颗粒递送常规剂量更低的剂量施用至受试者的靶组织,如PCT公布号WO 2020/236982(2020年11月26日公布)和Levy,J.M.,et al.Nat

Biomed Eng 4,97-110(2020)中所述。

[0740] 如本文所用,rAAV颗粒的血清型是指重组病毒的衣壳蛋白的血清型。在一些实施方案中,本文公开的rAAV颗粒包含rAAV2、rAAV3、rAAV3B、rAAV4、rAAV5、rAAV6、rAAV8、rAAV9、rAAV10、rPHP.B、rPHP.eB或rAAV9颗粒,或其变体。在特定实施方案中,所公开的rAAV颗粒是rAAV8或rAAV9颗粒。

[0741] 血清型衍生物和假型的非限制性实例包括rAAV2/1、rAAV2/5、rAAV2/8、rAAV2/9、AAV2-AAV3杂交体、AAVrh.10、AAVrh.74、AAVhu.14、AAV3a/3b、AAVrh32.33、AAV-HSC15、AAV-HSC17、AAVhu.37、AAVrh.8、CHt-P6、AAV2.5、AAV6.2、AAV2i8、AAV-HSC15/17、AAVM41、AAV9.45、AAV6(Y445F/Y731F)、AAV2.5T、AAV-HAE1/2、AAV克隆32/83、AAVShH10、AAV2(Y->F)、AAV8(Y733F)、AAV2.15、AAV2.4、AAVM41和AAVr3.45。具有嵌合VP1蛋白的衍生物和假型的非限制性实例是rAAV2/5-1VP1u,其具有AAV2的基因组、AAV5的衣壳主链和AAV1的VP1u。具有嵌合VP1蛋白的衍生物和假型的其他非限制性实例包括rAAV2/5-8VP1u、rAAV2/9-1VP1u和rAAV2/9-8VP1u。

[0742] AAV衍生物/假型以及生产此类衍生物/假型的方法是本领域已知的(参见例如Mol.Ther.2012Apr;20(4):699-708.doi:10.1038/mt.2011.287.Epub 2012Jan 24.The AAV vector toolkit:poised at the clinical crossroads.Asokan A1,Schaffer DV,Samulski RJ)。生产和使用假型rAAV载体的方法是本领域已知的(参见例如Duan et al.,J.Virol.,75:7662-7671,2001;Halbert et al.,J.Virol.,74:1524-1532,2000;Zolotukhin et al.,Methods,28:158-167,2002;以及Auricchio et al.,Hum.Molec.Genet.,10:3075-3081,2001)。

[0743] ITR序列和含有ITR序列的质粒是本领域已知的并且可商业获得(参见例如可从Vector Biolabs,Philadelphia,PA;Cellbiolabs,San Diego,CA;Agilent Technologies,Santa Clara,Ca;以及Addgene,Cambridge,MA获得的产品和服务;以及Gene delivery to skeletal muscle results in sustained expression and systemic delivery of a therapeutic protein.Kessler PD,Podsakoff GM,Chen X,McQuiston SA,Colosi PC,Matelis LA,Kurtzman GJ,Byrne BJ.Proc Natl Acad Sci USA.1996Nov 26;93(24):14082-7;和Curtis A.Machida.Methods in Molecular Medicine™.Viral Vectors for Gene Therapy Methods and Protocols.10.1385/1-59259-304-6:201©Humana Press Inc.2003.Chapter 10.Targeted Integration by Adeno-Associated Virus.Matthew D.Weitzman,Samuel M.Young Jr.,Toni Cathomen and Richard Jude Samulski;美国专利号5,139,941和5,962,313,所有这些通过引用并入本文)。

[0744] 在一些实施方案中,本公开的rAAV载体包含一个或多个调控元件以控制异源核酸区域的表达(例如启动子、转录终止子和/或其他调控元件)。在一些实施方案中,第一和/或第二核苷酸序列可操作地连接到一个或多个(例如1、2、3、4、5个或更多个)转录终止子。可以根据本公开使用的转录终止子的非限制性实例包括牛生长激素基因(bGH)、人生长激素基因(hGH)、SV40、CW3、φ的转录终止子(或聚腺苷酸化信号)或其组合。在示例性实施方案中,转录终止子是SV40聚腺苷酸化信号。在示例性实施方案中,转录终止子不含有转录后响应元件,如WPRE元件。

[0745] 在一些方面,本文提供了制造(或制备或包装)任何公开的rAAV颗粒的方法。rAAV

颗粒可以根据本领域已知的任何方法制备。制造或包装rAAV颗粒的方法是本领域已知的，并且试剂可商业获得(参见例如Zolotukhin et al.Production and purification of serotype 1,2,and 5recombinant adeno-associated viral vectors.Methods 28(2002) 158-167;以及美国专利公布号US2007-0015238和US2012-0322861,其通过引用并入本文;以及可从ATCC和Cell Biolabs,Inc.获得的质粒和试剂盒)。例如,包含感兴趣基因的质粒可以与一个或多个辅助质粒组合,例如含有rep基因(例如编码Rep78、Rep68、Rep52和Rep40)和cap基因(编码VP1、VP2和VP3,包括如本文所述的修饰的VP2区域)的质粒,并转染到重组细胞中,以便rAAV颗粒可以被包装并随后纯化。

[0746] 在一些实施方案中,所公开的rAAV颗粒提供了对靶组织的转导,以实现负载或转基因(例如根据本公开的碱基编辑器)的表达和翻译,持续足够的时间以在靶细胞的基因组中安装期望的突变。在一些实施方案中,期望的突变是C至T突变。在一些实施方案中,所公开的rAAV颗粒提供了足够持续时间的碱基编辑器转基因的足够表达和翻译,以在基因组中安装期望的(中靶)突变,同时具有可耐受水平的脱靶效应,如旁观者编辑。在一些实施方案中,所公开的rAAV颗粒提供了足够持续时间的碱基编辑器转基因的足够表达和翻译,以在基因组中安装期望的突变,而没有明显的脱靶编辑。在一些实施方案中,所公开的rAAV颗粒提供了足够持续时间的碱基编辑器转基因的足够表达和翻译,以在基因组中安装期望的突变,而没有明显的旁观者编辑。

[0747] 施用所公开的rAAV颗粒组合物的合适途径包括但不限于:外部、皮下、透皮、皮内、病灶内、关节内、腹膜内、膀胱内、透黏膜、牙龈、牙内、耳蜗内、经鼓膜、器官内、硬膜外、鞘内、肌肉内、静脉内、全身、血管内、骨内、眼周、肿瘤内、脑内、肠胃外和脑室内施用。在一些实施方案中,施用途径是全身(静脉内)。在一些实施方案中,本文所述的药物组合物局部施用至患病部位。

[0748] 另外的核酸递送方法是本领域技术人员已知的。参见例如美国公布号2003/0087817,其通过引用并入本文。应当理解,任何碱基编辑器(例如本文提供的任何碱基编辑器)可以以任何合适的方式引入细胞中,无论是稳定地还是瞬时地。在一些实施方案中,碱基编辑器可以转染到细胞中。在一些实施方案中,细胞可以用编码碱基编辑器的核酸构建体转导或转染。例如,细胞可以用编码碱基编辑器或翻译的碱基编辑器的核酸转导(例如用编码碱基编辑器的病毒)或转染(例如用编码碱基编辑器的质粒)。这种转导可以是稳定的或瞬时的。在一些实施方案中,表达碱基编辑器或含有碱基编辑器的细胞可以用一个或多个gRNA分子转导或转染,例如当碱基编辑器包含Cas9(例如nCas9)结构域时。在一些实施方案中,表达碱基编辑器的质粒可以通过电穿孔、瞬时(例如脂质体转染)和稳定基因组整合(例如piggybac)以及病毒转导或本领域技术人员已知的其他方法引入细胞中。

[0749] 试剂盒和细胞

[0750] 本公开的一些方面提供了包含核酸构建体的试剂盒,该构建体包含编码能够使脱氧核糖核酸(DNA)分子中的腺苷脱氨的腺苷脱氨酶的核苷酸序列。在一些实施方案中,核苷酸序列编码本文提供的任何腺苷脱氨酶。在一些实施方案中,核苷酸序列包含驱动腺苷脱氨酶表达的异源启动子。核苷酸序列还可以包含驱动gRNA表达的异源启动子,或驱动碱基编辑器和gRNA表达的异源启动子。

[0751] 在一些实施方案中,试剂盒还包含编码向导核酸主链(例如向导RNA主链)的表达

构建体,其中该构建体包含定位于允许将与靶序列相同或互补的核酸序列克隆到向导核酸(例如向导RNA主链)中的位置的克隆位点。

[0752] 本公开还提供了包含核酸构建体的试剂盒,该构建体包含:(a)编码以下的核苷酸序列:与腺苷脱氨酶融合的napDNAbp(例如Cas9结构域),或如本文提供的包含napDNAbp(例如Cas9结构域)和腺苷脱氨酶的碱基编辑器;以及(b)驱动(a)的序列表达的异源启动子。在一些实施方案中,试剂盒还包含编码向导核酸主链(例如向导RNA主链)的表达构建体,其中该构建体包含定位于允许将与靶序列相同或互补的核酸序列克隆到向导核酸(例如向导RNA主链)中的位置的克隆位点。

[0753] 在一些实施方案中,试剂盒包含:(a)编码本发明的任一个碱基编辑器的核酸序列,(b)编码gRNA的核酸序列,以及驱动(a)的序列和/或(b)的序列表达的一个或多个异源启动子。在一些情况下,试剂盒还包含表达构建体,表达构建体编码向导RNA主链和定位于允许将与靶序列相同或互补的核酸序列克隆到向导RNA主链中的位置的克隆位点。

[0754] 本公开的一些实施方案提供了包含本文提供的任何碱基编辑器或复合物的宿主细胞。在一些实施方案中,宿主细胞包含编码本文提供的任何碱基编辑器的核苷酸构建体。在一些实施方案中,细胞包含本文提供的任何核苷酸或载体。在一些实施方案中,细胞是干细胞。在一些实施方案中,细胞是人干细胞,例如人干细胞和祖细胞(HSPC)。在一些实施方案中,细胞是动员的(例如普乐沙福(plerixafor)动员的)外周血HSPC。在某些实施方案中,细胞是T细胞,如原代人T细胞。在一些情况下,细胞是人HSC。

[0755] 在一些实施方案中,宿主细胞用本文所述的一个或多个载体瞬时或非瞬时转染。在一些实施方案中,细胞以在受试者体内自然发生的方式被转染。在一些实施方案中,被转染的细胞取自受试者。在一些实施方案中,细胞衍生自取自受试者的细胞,如细胞系。本领域已知多种用于组织培养的细胞系。在一些实施方案中,细胞已从受试者中移除,并与任何公开的碱基编辑器、复合物、载体或多核苷酸离体接触。

[0756] 在一些实施方案中,宿主细胞用本文所述的一个或多个载体瞬时或非瞬时转染。在一些实施方案中,细胞以在受试者体内自然发生的方式被转染。在一些实施方案中,被转染的细胞取自受试者。在一些实施方案中,细胞衍生自取自受试者的细胞,如细胞系。本领域已知多种用于组织培养的细胞系。细胞系的实例包括但不限于C8161、CCRF-CEM、MOLT、mIMCD-3、NHDF、HeLa-S3、Huh1、Huh4、Huh7、HUVEC、HASMC、HEKn、HEKa、MiaPaCell、Panc1、PC-3、TF1、CTLL-2、C1R、Rat6、CV1、RPTE、A10、T24、J82、A375、ARH-77、Calu1、SW480、SW620、SKOV3、SK-UT、CaCo2、P388D1、SEM-K2、WEHI-231、HB56、TIB55、Jurkat、J45.01、LRMB、Bcl-1、BC-3、IC21、DLD2、Raw264.7、NRK、NRK-52E、MRC5、MEF、Hep G2、HeLa B、HeLa T4、COS、COS-1、COS-6、COS-M6A、BS-C-1猴肾上皮、BALB/3T3小鼠胚胎成纤维细胞、3T3 Swiss、3T3-L1、132-d5人胎儿成纤维细胞;10.1小鼠成纤维细胞、293-T、3T3、721、9L、A2780、A2780ADR、A2780cis、A 172、A20、A253、A431、A-549、ALC、B16、B35、BCP-1细胞、BEAS-2B、bEnd.3、BHK-21、BR 293.BxPC3.C3H-10T1/2、C6/36、Cal-27、CHO、CHO-7、CHO-IR、CHO-K1、CHO-K2、CHO-T、CHO Dhfr-/-、COR-L23、COR-L23/CPR、COR-L23/5010、COR-L23/R23、COS-7、COV-434、CML T1、CMT、CT26、D17、DH82、DU145、DuCaP、EL4、EM2、EM3、EMT6/AR1、EMT6/AR10.0、FM3、H1299、H69、HB54、HB55、HCA2、HEK-293、HeLa、Hepa1c1c7、HL-60、HMEC、HT-29、Jurkat、JY细胞、K562细胞、Ku812、KCL22、KG1、KYO1、LNCap、Ma-Mel 1-48、MC-38、MCF-7、MCF-10A、MDA-MB-231、

MDA-MB-468、MDA-MB-435、MDCK II、MDCK 11、MOR/0.2R、MONO-MAC 6、MTD-1A、MyEnd、NCI-H69/CPR、NCI-H69/LX10、NCI-H69/LX20、NCI-H69/LX4、NIH-3T3、NALM-1、NW-145、OPCN/OPCT细胞系、Peer、PNT-1A/PNT 2、RenCa、RIN-5F、RMA/RMAS、Saos-2细胞、Sf-9、SkBr3、T2、T-47D、T84、THP1细胞系、U373、U87、U937、VCaP、Vero细胞、WM39、WT-49、X63、YAC-1、YAR及其转基因品种。细胞系可从本领域技术人员已知的多种来源获得(参见例如美国典型培养物保藏中心(ATCC)(Manassus,Va.))。在一些实施方案中,用本文所述的一个或多个载体转染的细胞用于建立包含一个或多个载体衍生序列的新细胞系。在一些实施方案中,用如本文所述的CRISPR系统组分(如通过一个或多个载体的瞬时转染或用RNA转染)瞬时转染细胞,并通过CRISPR复合物的活性进行修饰,用于建立包含含有修饰但缺乏任何其他外源序列的细胞的新细胞系。在一些实施方案中,用本文所述的一个或多个载体瞬时或非瞬时转染的细胞,或由此类细胞衍生的细胞系用于评估一种或多种测试化合物。

[0757] 在一些实施方案中,宿主细胞是已经从受试者中移除并与本文所述的任何碱基编辑器、复合物或载体离体接触的细胞。

[0758] 在一些方面,本公开提供了本文所述的任一个碱基编辑器和将该碱基编辑器靶向至核酸分子中的A:T碱基对的向导RNA在制备用于核酸编辑的试剂盒的用途,其中核酸编辑包括在适合用鸟嘌呤(G)置换A:T核碱基对中的腺嘌呤(A)的条件下,将核酸分子与碱基编辑器和向导RNA接触。在这些用途的一些实施方案中,核酸分子是双链DNA分子。在一些实施方案中,接触步骤诱导双链DNA在靶区域的分离。在一些实施方案中,接触步骤从而包括使双链DNA的一条链产生切口,其中这一条链包含未突变的链,该未突变的链包含靶A:T核碱基对中的T。

[0759] 在所描述的用途的一些实施方案中,接触步骤在体外进行。在其他实施方案中,接触步骤在体内进行。在一些实施方案中,接触步骤在受试者(例如人受试者或非人动物受试者)中进行。在一些实施方案中,接触步骤在细胞(如人细胞或非人动物细胞)中进行。

[0760] 本公开还提供了本文所述的任一个腺嘌呤碱基编辑器作为药物的用途。本公开还提供了本文所述的任一个腺嘌呤碱基编辑器和向导RNA的复合物作为药物的用途。

[0761] 应当理解,上述概念以及下面讨论的其他概念可以以任何合适的组合排列,因为本公开在这方面不受限制。此外,本公开的其他优点和新颖特征将从结合附图考虑的各种非限制性实施方案的以下详细描述中变得显而易见。

[0762] 实施例

[0763] 实施例1.针对胞苷脱氨的噬菌体辅助进化选择的开发。

[0764] 噬菌体辅助连续进化(PACE)已实现了多种蛋白质功能的快速实验室进化,包括蛋白质-蛋白质相互作用[35]、tRNA合成酶[36]、DNA结合蛋白[37-39]、蛋白酶[40,41]、聚合酶[42]、代谢酶[43-45]和碱基编辑器[9,12]。在PACE期间,正在进化的蛋白质在选择噬菌体(SP)上编码,该噬菌体感染大肠杆菌宿主细胞[46]。大肠杆菌含有持续诱变噬菌体基因组的诱变质粒(MP)以及建立调节基因III的表达的选择回路的辅助质粒(AP),该基因III编码pIII,一种用于噬菌体复制的关键蛋白质。由于gIII已从SP基因组中移除,仅编码具有期望活性的正在进化的变体的噬菌体触发大肠杆菌中pIII的产生并复制,从而导致活性基因变体的繁殖(图1B)。在持续诱变和稀释下,缺乏期望活性的噬菌体迅速从选择容器(“培养池”)中稀释,而进化出有益突变的噬菌体则持续存在。

[0765] 先前,开发了CBE-PACE选择[12],其中胞苷脱氨酶在SP内编码,并且宿主大肠杆菌细胞含有(i)MP,(ii)编码SpCas9的辅助质粒,(iii)与C端降解决定子融合的自失活T7 RNA聚合酶(T7 RNAP),以及(iv)由T7 RNAP转录控制的基因III。噬菌体感染后,SP编码的脱氨酶通过反式内含肽剪接与Cas9连接,以重建碱基编辑器。为了激活选择回路,碱基编辑器必须进行C·G至T·A编辑,以在T7 RNAP和降解决定子之间创建终止密码子,从而生成活性T7 RNAP。无降解决定子的T7 RNAP随后转录gIII,导致噬菌体繁殖[12]。

[0766] 为了开发选择通过TadA的胞苷脱氨的PACE回路,对先前的CBE选择回路进行修改以适应具有高初始腺苷脱氨活性的酶(图1C)。在原始回路中,通过模板链上CCA的C至T转化,将TGG(Trp)编辑为终止密码子(TAG、TGA或TAA)。然而,这种策略将腺嘌呤(其在所有终止密码子(TAG、TGA、TAA)中与胸腺嘧啶相对)置于靶原间隔区的位置6处。鉴于位置6被ABE8e高度编辑[9],且A_6的A至G编辑阻止终止密码子形成(因为CGG、CAG、CGA和CAA都编码氨基酸),该原始回路需要相对于腺苷脱氨对胞苷脱氨的高选择性,这在早期进化的ABE8e变体中不太可能发现。

[0767] 为了解决这个问题,开发了,改为编辑非模板链的新的选择回路(图1D)。在新回路中,$\underline{C}_6A_7A_8$被编辑为$\underline{T}_6A_7A_8$,以在胞苷脱氨时引入终止密码子。腺苷脱氨不会阻止新选择中的终止密码子安装(TAA、TGA或TAG),除非A_7和A_8都被转化为G(TGG=Trp),这使得该回路耐受中等水平的腺苷脱氨,因此更适合早期TadA8e进化(回路1)。在新回路中进行初步进化后,可以切换到原始模板链回路(回路2),以利用其针对对腺苷脱氨的固有强负选择(图7)。

[0768] 实施例2.胞苷脱氨酶进化。

[0769] 使用回路1起始了TadA-8e的噬菌体辅助非连续进化(PANCE)(图1E)。在PANCE中,含有AP和MP的大肠杆菌宿主细胞被含有感兴趣基因的噬菌体感染并过夜生长,无需连续稀释。第二天,含有噬菌体的上清液被稀释到新鲜的宿主细胞培养物中,并重复该过程以富集含有活性胞苷脱氨酶的噬菌体。与PACE相比,PANCE提供了较低的严格性,因此在早期进化活动期间是有帮助的,在早期进化活动中保留具有低初始活性的遗传多样性变体可能是关键的[9,41,43]。在用诱导的MP6诱变的四轮PANCE之后[47],噬菌体开始过夜繁殖>100倍,表明对于胞苷脱氨的改进的活性。为了增加选择的严格性,增加了传代之间的稀释倍数,并降低了T7 RNAP上游启动子的强度(表1,图8A-8C)。接下来,使用回路2进行PANCE的额外传代(图8A-8C),以针对腺苷脱氨进行选择,同时保持胞苷脱氨活性。为了进一步增加选择严格性,使用回路2对PANCE存活的噬菌体池进行了159小时的连续进化(PACE)(图9A-9C)。从PANCE和PACE的所有阶段出现的TadA-8e变体在约10^{139}倍的平均总稀释度下存活。

[0770] 分离了PANCE和PACE中存活的单个噬菌体并测序,以鉴定进化期间获得的TadA-8e突变(图2A,图8A-9C)。在所有测序的噬菌体中观察到残基26-28的突变的显著普遍性,其中R26G、E27K、E27A和V28G突变在若干单独进化的培养池中是高度代表性的。接下来,在大肠杆菌中针对碱基编辑测定了经进化的变体。将来自噬菌体的三个经进化的TadA变体从噬菌体亚克隆到低拷贝质粒上的BE4max架构[48](从N端到C端:TadA*-SpCas9-UGI-UGI),并设计了含有来自噬菌体进化所在的选择回路的序列的高拷贝靶标质粒。将编码向导RNA的碱基编辑器质粒和靶标质粒共转化到大肠杆菌细胞中,共转化允许阿拉伯糖诱导后的编辑过夜发生。随后,对靶标质粒进行了高通量测序(图2B)。

[0771] 测序结果揭示了,与起始TadA-8e变体相比,经进化的TadA变体的选择性的显著移

动。虽然含有TadA-8e的碱基编辑器在靶标质粒中A_6处产生了90％的A•T至G•C编辑,在C_4和C_5处产生了1％-2％的C•G至T•A编辑,经进化的变体反而导致在胞嘧啶的85％-92％编辑以及腺嘌呤的1％-2％编辑(图2C),表示胞嘧啶相对于腺嘌呤碱基编辑的＞3000倍的变化。这些结果表明,使用选择回路1和2的PANCE和PACE进化出TadA变体,以下称为TadA-胞苷脱氨酶(TadA-CD),其具有强胞苷脱氨活性以及相较于腺嘌呤碱基编辑对胞嘧啶的高选择性。

[0772] 从用TadA-8e A48R(含有增加TadA-7.10中混杂性的突变[32])感染的培养池,在大肠杆菌编辑测定中也鉴定了进行A•T至G•C(80％)和C•G至T•A编辑(73％)两者的变体(图2C)。因此,该变体用作基于TadA的双重编辑器(TadDE)。TadDE比先前报道的将胞苷和腺苷脱氨酶两者融合至Cas结构域的双重编辑器更小[49-53],并且可能特别适用于需要广泛诱变的应用[54],诸如遗传筛选[55,56]。

[0773] 为了鉴定经进化的突变的潜在作用,将突变定位到ABE8e的冷冻电镜结构(PDB 6VPC)[18]上。高度保守的突变预计位于活性位点附近的环上(图2D)。该环与靶编辑附近的单链DNA底物的主链相互作用,并支持碱基相对于催化锌离子的生产性定向。其他保守突变,包括A158S和Q154R,也被定位到TadA和单链DNA底物的界面上。使用AlphaFold[57]的TadA-CDa的结构预测表明,与ABE8e的冷冻电镜结构(6VPC,图10A-10C)相比,这些突变预计不会显著改变TadA的二级结构。相比之下,进化期间残基26-28从Arg-Glu-Val至较小的氨基酸(如Gly-Ala-Gly)的观察到的突变可能缓解了空间冲突,否则预计会阻碍嘧啶C_4针对亲核攻击和脱氨的正确定位(图10A-10C)。这些观察结果共同表明,经进化的突变可能改变结合的DNA底物的构象,以实现高效的胞苷脱氨并阻碍腺苷脱氨。

[0774] 接下来,进行了诱变和回复分析,以探究通过进化发现的突变的作用。单独来看,没有一个突变足以改变选择性(图33)。然而,仅在环区域添加两个突变(TadCBEa-c,e中的E27A V28G和TadCBEd中的E27K V28A)就足以改变TadCBE的选择性,使其适度偏好胞苷脱氨,尽管编辑效率较低(图33)。在PANCE或PACE期间进化的另外的突变大大提高活性并改进对C•G至T•A转化的选择性。环区域外突变的回复通常会降低活性,但不降低选择性(图34)。因此,该回复分析支持了残基26-28在调节经进化的TadA变体的脱氨选择性中的重要性。

[0775] 实施例3.哺乳动物细胞中TadA-CD的表征、TadCBE与Cas9直系同源物的兼容性以及编辑窗口。

[0776] 受到TadA-CD在细菌中特性的激励,在哺乳动物细胞中评估了经进化的TadA-CD胞嘧啶碱基编辑器(TadCBE)。将五种TadCBE变体(TadCBEa-e)克隆到BE4max架构[48]中由CMV启动子调控的哺乳动物表达载体中。这五种TadCBE变体与三种最广泛使用的工程化和经进化的CBE一起进行了测定:BE4max[48]、evoA[12]和evoFERNY[12]。HEK293T细胞中用每个碱基编辑器质粒和sgRNA质粒共转染,允许编辑发生72小时,然后对来自基因组DNA中的靶位点进行测序。在HEK293T细胞中测试的九个不同靶位点上,TadCBE变体通常产生靶标C•G至T•A编辑(TadCBEa-e在所有九个测试位点间的平均峰值编辑为51％-60％),与从典型BE4max、evoA和evoFERNY CBE(在所有九个位点间的平均峰值编辑分别为47％、55％和41％)观察到的相当或更高(图3和图11)。这些结果表明,TadCBE可以在哺乳动物细胞中进行高效的C•G至T•A编辑。

[0777] 经进化的TadCBE变体通常显示出较低残留的A•T至G•C编辑,TadCBEa-e在所有九个测试位点上腺苷间的平均编辑为1.5%-4.5%,因此相较于A•T至G•C编辑显示出对C•G至T•A编辑的优异选择性(图3)。相比之下,相同碱基编辑器架构中的ABE8e(带有2xUGI)在九个位点间平均为31% A•T至G•C编辑以及2.0%C•G至T•A编辑。对于九个测试位点中的七个,期望C•G至T•A编辑与残留A•T至G•C编辑的比率非常高,TadCBE变体a、c、d和e平均为21到42倍,并且TadCBEb平均为9.2倍(图3)。总的来说,这些观察结果表明,残留的A•T至G•C编辑在经进化的TadCBE变体间通常较低,并且主要限于一小亚组靶位点、原间隔区位置和TadCBE变体。必要时,在脱氨酶结构域中引入V106W可以进一步降低残留的A•T至G•C编辑(见下文)。

[0778] 为了测试TadCBE是否与除酿脓链球菌Cas9之外的Cas9同源物兼容,构建了TadCBE变体和来自脑膜炎奈瑟菌的Nme2Cas9的PACE进化的变体,这些变体扩展了除SpCas9的典型NGG PAM[50]之外的可及PAM的范围。进化了作为核酸酶或作为碱基编辑器接近广泛的单嘧啶PAM位点的Nme2Cas9变体[51](参见Huang,T.P.et al.Nature Biotechnology(2022),其通过引用并入本文)。生成了TadA-CD与eNme2-C切口酶(PAM=N_4CN)和两个UGI结构域的融合物,所得eNme2-C-TadCBE与向导RNA质粒共转染,并检查了HEK293T细胞中六个基因组基因座的碱基编辑。在所有测试位点间,TadCBE的峰值编辑效率与APOBEC1、evoFERNY和evoAPOBEC1相当(图4、图12)。尽管C•G至T•A编辑在某些位点上超过50%,但在测试的六个eNme2靶位点的任一个上,残留的A•T至G•C编辑从未超过5.3%。因此,TadCBE使用eNme2 Cas9表现出稳健的活性和选择性。这些观察结果表明与其他Cas蛋白的潜在兼容性,这些Cas蛋白与SpCas9和eNme2-C Cas9一起可以提供对多种PAM序列的接近,用于TadCBE的多功能靶向。

[0779] 实施例4.TadCBE和V106W变体的中靶和脱靶编辑。

[0780] TadCBE的TadA起源为最小化脱靶编辑提供了若干优势,包括可能包括发现在先前的TadA工程化努力中降低脱靶DNA或RNA编辑的突变[34,58,59]。对于ABE,将V106W添加到TadA-7.10、TadA-8e或TadA-8.17-m在所有三种情况下都降低了DNA和RNA的Cas非依赖性脱靶编辑,同时维持高水平的中靶活性[8,9,34]。测试了V106W突变在引入TadCBE中时是否可以降低脱靶DNA或RNA编辑同时维持中靶活性和选择性。由于TadA-CD中的若干经进化的突变与V106接近,尚不清楚添加V106W是否会破坏期望的TadA-CD特性(图13)。

[0781] 首先,评估了含有V106W的TadCBE的中靶活性。构建了TadCBEa-e的V106W变体,并评估了它们在HEK293T细胞中在九个靶位点处的编辑效率。TadCBE变体a至e耐受V106W的添加,并维持了高中靶胞苷脱氨活性,TadCBEa-d V106W在九个测试靶位点间平均为56%的峰值C•G至T•A靶标编辑效率,几乎与TadCBEa-d的57%平均峰值编辑效率相匹配(图5A、图14-17)。TadCBEa-e V106W变体表现出比TadCBEa-d稍窄的编辑窗口,同时维持了高峰值编辑效率(图17)。令人鼓舞的是,在这九个位点间,与相应的TadCBE变体相比,TadCBE V106W变体的胞嘧啶与腺嘌呤碱基编辑选择性平均改进了3.1倍(图17)。因此,TadCBE-V106W变体保留了高效的胞嘧啶碱基编辑,同时改进了相较于腺苷脱氨对胞苷的选择性,并优化了编辑窗口。

[0782] 接下来,使用先前建立的正交R环测定[15,19]评估了TadCBE和TadCBE-V106W变体的Cas非依赖性DNA编辑(图5B)。该测定测量了碱基编辑器在由正交、催化失活的金黄色葡萄球菌Cas9(SaCas9)生成的脱靶R环中修饰ssDNA的倾向。通过对六个不相关的脱靶SaCas9 R

环的基因组DNA进行测序,确定TadCBE与BE4max(3.6%)和evoA(3.8%)相比平均具有3.7倍更低的Cas非依赖性脱靶C·G至T·A编辑(0.84%-1.2%)(图5C、图18-21C)。evoFERNY(0.58%)和YE1(0.53%)的平均脱靶活性也较低。添加V106W进一步将TadCBE的Cas非依赖性脱靶编辑降低了平均1.9倍(对于V106W TadCBE变体a至e分别为0.38%、0.62%、0.48%、1.1%和0.05%)。与TadCBE对胞苷脱氨的选择性一致,未检测到通过任何TadCBE的显著脱靶A·T至G·C编辑(图22)。这些发现表明,经进化的TadCBE具有固有的低Cas非依赖性编辑脱靶DNA编辑,可以通过添加V106W进一步抑制,同时维持高中靶C·G至T·A编辑和低残留A·T至G·C编辑。

[0783] 还评估了通过TadCBE的脱靶RNA编辑(图5D、图23A-23B和图24)。在通过TadCBEa-e、BE4max、evoA、evoFERNY、ABE8e或ABE8e-V106W转染HEK293T细胞后,从细胞中提取RNA。在cDNA合成后,通过RT-PCR扩增了三个靶标转录物(CTNNB1、IP90和RSL1D1),这些转录物由于它们的丰度或与天然TadA tRNAArg2底物的序列相似性而先前用于测量脱靶RNA编辑[4,15,19,34],并通过高通量测序分析C至U或A至I编辑。虽然BE4max和evoA编辑了这些转录物中分析的胞嘧啶的约0.7%,但evoFERNY、YE1和TadCBEa、TadCBEb和TadCBEc都编辑了≤0.1%的胞嘧啶(检测限)(图5D、图23A-23B)。TadCBEd和TadCBEe分别在三个转录物间平均编辑了0.3%和0.2%的胞嘧啶。添加V106W将这两种情况的平均脱靶RNA编辑降低到≤0.13%(图5D、图23A-23B)。

[0784] 总的来说,这些数据表明,与BE4max和evoA相比,TadCBE提供了更低频率的Cas非依赖性脱靶DNA和RNA编辑。通过TadCBE的脱靶编辑的频率明显低于具有相当靶向活性和大小的任何其他CBE。当进一步降低脱靶编辑至关重要时,添加V106W最小化了脱靶DNA和RNA编辑,将编辑窗口聚焦到约4-5bp,并最小化了残留的腺苷脱氨,同时仅略微降低了最大中靶活性。

[0785] 最后,当碱基编辑器通过不完美的Cas9结合与类似于靶位点的非靶位点结合时,发生Cas依赖性脱靶编辑[60]。在SpCas9碱基编辑器以及靶向HEK293T位点3(以下简称HEK3)、HEK293T位点4(以下简称HEK4)、EMX1和BCL11A的sgRNA的22个已知脱靶位点处在HEK293T细胞中分析了Cas依赖性脱靶活性(图25-28、图35-36)。在多个已验证的脱靶位点间,观察到通过TadCBE的Cas依赖脱靶编辑水平与对BE4max和EvoA变体观察到的编辑的低水平大致相似(图25-28、图35-36)。YE1和evoFERNY的Cas依赖性脱靶活性仍然较低,这与这些变体的较低中靶活性一致(图25-28、图35-36)。

[0786] 总的来说,这些发现表明,与典型CBE相比,TadCBE提供改进的Cas非依赖性脱靶DNA和RNA编辑,并且Cas依赖性脱靶DNA编辑的低水平与当前使用的具有相似中靶编辑效率的CBE观察到的那些一致。已知使用接合较低脱靶基因座的高保真Cas蛋白可以降低Cas依赖性脱靶DNA碱基编辑[61],并且它们在TadCBE中的使用可能会提供同样的益处。

[0787] 实施例5.原代人T细胞中治疗相关基因座处的多路复用碱基编辑,以及人造血干细胞中治疗相关位点处的碱基编辑。

[0788] 评估了TadCBE是否能够在T细胞中进行靶基因座的多路复用编辑以支持治疗应用。T细胞中的多路复用碱基编辑可以用于修饰或破坏多个基因,而没有由多重双链断裂引起的染色体异常和细胞状态扰动的风险[58-62]。为了确定TadCBE是否能够在原代人T细胞中进行多路复用编辑,靶向CXCR4和CCR5基因座进行同步碱基编辑,以在两个HIV共受体中安装

提前终止密码子(图6)[63]。进行了TadCBE变体a、b、c、d和e的体外转录。然后,将TadCBE mRNA与靶向CXCR4和CCR5的向导RNA一起电穿孔(图6)[63]到原代人T细胞中,并在两个靶位点处分析编辑效率。

[0789] TadCBE对靶标胞嘧啶(CXCR4中的C7、CCR5中的C9)进行了高效(平均70%)和选择性编辑,导致每个基因中的提前终止密码子安装(图6)。TadCBE的编辑效率与BE4max(67%)和evoA(76%)的那些相似(图6)。所有测试的碱基编辑器观察到的插入缺失频率都相对较低(通常≤0.68%,图29A-29B)。与HEK293T细胞中的数据一致(图17),TadA-CD在原代人T细胞中的CXCR4和CCR5处表现出更精确的编辑窗口以及更少的旁观者编辑。由于TadCBE维持了高编辑效率和产物纯度,但提供了比APOBEC和evoA显著更低的Cas非依赖性脱靶DNA和RNA编辑(图5C-5D和图18-22),TadCBE为T细胞的多路复用胞嘧啶碱基编辑提供了有前景的替代方案。

[0790] 还将通过TadCBE的T细胞编辑与evoFERNY和YE1的T细胞编辑进行了比较,evoFERNY和YE1提供了与TadCBE相似的低脱靶编辑(图5C-5D、图6和图18-22)。TadCBE支持了在T细胞中比evoFERNY和YE1显著更高的编辑效率。在CXCR4处,通过TadCBE的靶标C·G至T·A编辑效率平均为evoFERNY和YE1的编辑效率的1.5倍至1.7倍,而在CCR5处,平均TadCBE编辑效率为平均高4.9倍至11倍。针对CCR5向导RNA,分析了三个已知Cas依赖性脱靶位点和CXCR4的一个已知脱靶。TadCBEa-e、evoFERNY和YE1的Cas依赖性脱靶编辑(≤0.12%)低于BE4max(0.1-0.58%)和evoA(0.1-1.0%)(图30)。接下来,在T细胞中测试了TadCBEa-d的V106W变体。相对于它们的TadCBE对应物,V106W变体展示出在CXCR4的C7处1.3倍至1.9倍更低的平均活性,以及在CCR5的C9处1.4倍至3.3倍更低的平均活性,其中C·G至G·C编辑的成比例下降(图48-50)。这些数据与V106W变体的较窄编辑窗口一致,并表明与HEK293T细胞的质粒转染相比,TadCBE的更瞬时mRNA递送可能揭示出更大的编辑活性范围。总体而言,这些发现表明,当在治疗相关性的靶位点处对原代人T细胞进行碱基编辑时,与当前使用的CBE相比,TadCBE提供了中靶和脱靶编辑特性的有利组合。

[0791] 最后,评估了TadCBE在人造血干细胞和祖细胞(HSPC)中的编辑效率。将TadCBEa-e mRNA与靶向BCL11A增强子的合成向导一起电穿孔到原代人CD34阳性细胞中。增强子处的突变可以降低BCL11A的表达,导致胎儿血红蛋白表达的诱导,作为镰状细胞病的潜在治疗方法[73,74]。为了进行比较,平行电穿孔了编码BE4max、evoAPOBEC1(evoA)、evoFERNY、YE1或GFP(作为阴性对照)的mRNA。evoFERNY和YE1分别产生了仅2.0%和2.7%平均编辑,而BE4max和evoA的平均编辑效率分别为7.0%和7.4%(图6)。所有五个测试的TadCBE支持比BE4max或evoA高2至3倍的编辑效率,平均为14%-23%(图6)。所有测试的CBE都产生了低水平的插入缺失(≤1.1%,图31A)和Cas依赖性脱靶编辑(≤0.87%,图31B)。这些结果表明,对于一些治疗相关位点和细胞类型,TadCBE的编辑效率可以超过最常用的CBE。

[0792] 讨论

[0793] TadA已在实验室中从在大肠杆菌中发现的tRNA编辑酶进化和工程化为广泛使用的腺嘌呤碱基编辑器,包括若干已经进入临床[2]或即将进入临床试验[1]。经进化的TadA变体提供了许多有益于精确基因编辑应用的特征,包括以前在胞嘧啶碱基编辑器中不存在的一些特性。在本研究中,催化高效和选择性胞苷脱氨的新TadA变体的进化使得TadCBE的开发成为可能,这是一类新的CBE,提供高中靶编辑、低脱靶Cas非依赖性和Cas依赖性DNA编辑、低

脱靶RNA编辑以及足够小的尺寸以适应单一AAV[27,28]。在HEK293T细胞中,TadCBE使用SpCas9、Nme2Cas9和SaCas9在多个位点上进行高效的C·G至T·A编辑。这些结果代表了首次定向进化脱氨酶以选择性脱氨不同的碱基,而不是简单地放宽靶标碱基特异性——这是经进化的选择性TadCBE脱氨酶的同时正负选择系统的结果。

[0794] 与常用CBE的并排比较揭示了,TadCBE提供使其非常适合典型BE4max、evoA、evoFERNY和YE1可能面临限制的应用的独特特性。当需要精确编辑时,TadCBE的窄编辑窗口是有益的。尽管具有与BE4max和evoA相当的中靶编辑效率,但TadCBE表现出显著更低的Cas非依赖性脱靶DNA和RNA编辑。evoFERNY和YE1也表现出低Cas非依赖性编辑,但在一些靶基因座处(包括T细胞中的CXCR4和CCR5以及HSPC中的BCL11A)展示出不同的编辑谱,并实现了显著更低的编辑效率。因此,TadA-CD从TadA-8e的进化扩展了TadA在基因编辑中的实用性,展示了生成碱基编辑器的新策略,并提供了具有有利编辑特性的新型CBE家族。

[0795] TadCBEd提供了TadCBE变体的最高中靶编辑和选择性用于一般胞嘧啶碱基编辑应用。当必须将脱靶DNA或RNA编辑或残留的A·T至G·C编辑保持在最低时,推荐使用TadCBEd-V106W。

[0796] 方法

[0797] 一般方法和分子克隆

[0798] 使用Gibson组装(New England BioLabs)或USER克隆(New England BioLabs)进行所有质粒构建。无核酸酶水(Qiagen)用于PCR反应和克隆。对于所有其他实验,使用MilliQ纯化系统(Millipore)纯化水。使用Phusion HiFi聚合酶或Phusion U Green Hot Start IIDNA聚合酶(ThermoFisher Scientific)进行PCR。在Gibson或USER克隆后,将克隆产物转化到Mach 1化学感受态大肠杆菌(ThemoFisher Scientific)中。选择抗生素以以下最终浓度使用:羧苄青霉素:100μg/ml;壮观霉素:50μg/ml;卡那霉素:50μg/ml;氯霉素:25μg/ml;四环素:10μg/ml。在Sanger测序(Quintara Boston)之前,使用Illustra Templiphi 100扩增试剂盒(GE Healthcare Life Sciences)扩增质粒DNA。用于细菌转化的序列确认质粒使用Miniprep试剂盒(Qiagen)纯化。用于哺乳动物转染的质粒根据制造商的说明使用Midiprep试剂盒(Qiagen)纯化。通过nanodrop定量质粒。表1给出了本工作中使用的细菌质粒的完整列表。

[0799] 噬菌体克隆

[0800] 对于噬菌体的USER组装,将0.2pmol的每个PCR片段加入20μL的最终体积中。USER组装后,将20μL USER反应转化到100μL含有pJC175e的化学感受态S2060大肠杆菌宿主细胞中[46]。对于噬菌体的Gibson组装,加入0.2pmol的每个PCR片段,构成20μL的最终体积。Gibson组装后,将20μL Gibson反应转化到100μL含有pJC175e的化学感受态S2060大肠杆菌宿主细胞中[46]。转化了pJC175e的细胞能够进行活性非依赖性的噬菌体繁殖,并在无抗生素的2×YT培养基中在37℃摇动生长5小时。然后将细菌在10,000g下离心1分钟,并如下所述进行噬斑形成以分离克隆噬菌体群体。单个噬斑在DRM培养基(由USBiological CS050H-001/CS050H-003制备)中生长6-8小时。将细菌在6,000g下离心10分钟以从上清液中去除大肠杆菌。含有噬菌体的上清液通过0.22μm PVDF Ultrafree离心过滤器(Millipore)过滤以去除残留的细菌。对于测序,使用引物AB1793(5'-TAATGGAAACTTCCTCATGAAAAAGTCTTTAG)(SEQ ID NO:270)和AB1396(5'-ACAGAGAGAATAACATAAAAACAGGGAAGC)(SEQ ID NO:271)扩增噬菌

体内的感兴趣基因,并通过Sanger测序(Quintara)对PCR产物进行测序。引物退火到噬菌体主链,侧翼是正在进化的感兴趣基因。序列确认的噬菌体储存在4℃。

[0801] 化学感受态细胞的制备和转化

[0802] 在所有噬菌体繁殖、PANCE和PACE实验中使用S2060株[75]。为了制备感受态细胞,将过夜培养物稀释250倍到补充有四环素的50ml 2×YT培养基(United States Biologicals)中,并在37℃以230RPM摇动培养至OD_{600}约0.4-0.6,然后在冰上孵育20分钟。然后在4℃以4,000g离心10分钟沉淀细胞。通过加入5ml TSS(补充有5%v/v DMSO、10%w/v PEG 3350和20mM MgCl2的LB培养基)重悬细胞沉淀。细胞悬浮物用移液器轻柔移液以完全混合,等分成100μL体积,在液氮中快速冷冻,并储存在-80℃。

[0803] 为了转化细胞,将100μL在冰上解冻的感受态细胞加入在20μL 5x KCM溶液(500mM KCl、150mM CaCl2和250mM MgCl2在水中)和80μL水中的质粒的预冷混合物(每个1-2μL;每次转化最多3个质粒)中,并通过移液器移液轻柔混合。将混合物在冰上孵育15分钟,然后在42℃热休克90秒,然后加入800μL SOC培养基(New England BioLabs)进行挽救。允许细胞在37℃下以230RPM摇动1-1.5小时以恢复,然后铺板在含有适当抗生素的2×YT培养基+1.5%琼脂(United States Biologicals)上,并在37℃孵育16-18小时。

[0804] 用于噬菌体滴度定量和克隆的噬斑测定

[0805] 噬菌体在含有pJC175e质粒的S2060大肠杆菌宿主细胞上进行噬斑形成,以实现活性非依赖性繁殖[46]。为了准备处于适当生长的噬菌体的大肠杆菌宿主细胞用于噬斑形成,将宿主细胞的过夜培养物(新鲜或储存在4℃最长达3天)在含有适当抗生素的DRM中稀释50倍。细胞在37℃生长至OD_{600}为0.8-1.0(约2小时),此时在噬菌体制备期间将它们移到冰桶中。噬菌体原液用DRM按10倍序列稀释(最多10^6倍)。为了准备用于噬斑形成的板,将熔化的2×YT培养基琼脂(1.5%琼脂,55℃)与Bluo-gal(Gold Bio,4%w/v在DCM中)混合至最终浓度为0.08% Bluo-gal。将熔化的琼脂混合物移液到四分Petri培养皿的象限中(每个象限2ml),并在室温下静置5分钟使其凝固。为了准备上层琼脂,制备2×YT培养基和熔化的2×YT培养基琼脂(1.5%,导致0.6%琼脂的最终浓度)的3:2混合物,并在55℃储存直至使用。为了进行噬斑形成,将100μL细胞与10μL噬菌体在2ml文库管(VWR International)中混合。加入900μL温热的上层琼脂到细胞和噬菌体混合物中,用移液器移液至混合,然后立即移液到Petri培养皿中的一个象限中的固体琼脂培养基上。允许上层琼脂在25℃静置2分钟。然后将板在37℃孵育过夜,不进行倒置。通过量化蓝色噬斑确定噬菌体滴度。对于高通量噬斑形成,试剂针对12孔板的孔的孔调整如下:900μL下层琼脂、450μL上层琼脂、10μL噬菌体、100μL细胞。

[0806] 噬菌体过夜繁殖测定

[0807] 如上所述制备用AP和CP质粒转化的S2060细胞,并接种到DRM中。使细胞生长过夜。第二天,将宿主细胞稀释50倍到新鲜的DRM中,并在37℃生长至OD_{600}为0.3-0.5。将宿主细胞分配到96孔板(每孔1ml,Axygen)的孔中,然后加入已知滴度的噬菌体至输入浓度为105p.f.u./ml。培养物在37℃以230rpm摇动过夜生长(14-20小时)。然后将板在4,000g下离心10分钟以去除细胞,使噬菌体留在上清液中。然后通过上述噬斑形成对上清液进行滴定。通过将输出的繁殖的噬菌体滴度除以输入的噬菌体浓度计算富集倍数。

[0808] 噬菌体辅助非连续进化(PANCE)

[0809] PANCE实验按照已发表的方案[76]进行。如上所将用AP和CP转化的S2060宿主细胞制备成化学感受态。化学感受态宿主细胞用诱变质粒(MP6)[47]转化,并铺板在含有100mM葡萄糖和适当抗生素的2×YT琼脂上。将四到八个菌落挑入含有1ml DRM和适当抗生素的96孔板的单个孔中。将菌落重悬并连续稀释10倍八次到DRM中。用多孔密封膜密封板,并在37℃以230RPM摇动生长16-18小时。将含有具有OD_{600}约0.3-0.4的稀释物的孔合并,用20mM阿拉伯糖诱导诱变,并分配到96孔板中期望数量的1ml培养物中。然后用选择噬菌体以指示的稀释度接种培养物(表3和图8)。感染的培养物在37℃生长12-18小时,并在第二天通过4,000x g离心10分钟收集。将100μL含有进化噬菌体的上清液转移到96孔PCR板中,用箔密封,并储存在4℃。分离的噬菌体然后用于感染下一次传代,并在选择期间重复该过程。噬菌体滴度通过如前所述的qPCR[76]或如上所述的噬斑测定确定。进化期间使用的启动子和核糖体结合位点的序列在表7中。

[0810] 噬菌体辅助连续进化(PACE)

[0811] PACE实验按照先前发表的方案[67]进行。如上针对PANCE所述制备含有诱变质粒的宿主细胞。将12个菌落挑入含有1ml DRM和适当抗生素的96孔板的单个孔中。将菌落重悬并以10倍八次连续稀释到DRM中。用多孔密封膜密封板,并在37℃以230RPM摇动生长16-18小时。将含有具有OD_{600}约0.3-0.4的稀释物的孔合并并用于接种含有100ml DRM的恒化器。恒化器生长至OD_{600}约0.4-0.8,然后以1-1.5恒化器体积/小时的速率用新鲜DRM连续稀释以保持细胞密度恒定。恒化器维持在80-100ml的体积。

[0812] 在选择噬菌体(SP)感染之前,用15ml来自恒化器的培养物填充培养池,并用10mM阿拉伯糖预诱导至少1小时。培养池以起始滴度为108pfu/ml的SP感染。为了增加严格性,培养池的稀释率随时间增加,如图9所示。在进化期间,从培养池废液管线以按指示时间收集SP的样品(800μL)。样品在6,000g下离心10分钟,并且上清液储存在4℃。使用转化有pJC175e的S2060细胞通过噬斑测定确定SP样品的滴度[46]。通过PCR使用AB1793/AB1396引物对确定单个噬斑的序列,如上文噬菌体克隆方法中所述。使用Mutato进行突变分析。Mutato可在hub.docker.com/r/araguram/mutato上作为Docker镜像获取。

[0813] 大肠杆菌中质粒编辑的高通量测序分析

[0814] 为了生成表达碱基编辑器的细胞,将20μL 10-beta电感受态大肠杆菌(New England BioLabs)分配到16孔Nucleocuvette条中。将靶标质粒和编辑器质粒(每个0.5μL,100-200ng/μL)加入每个孔中,并使用4D-Nucleofector系统(Lonza)用细菌程序X-5电穿孔大肠杆菌。电穿孔细胞立即在120μL SOC培养基(New England BioLabs)中恢复,在37℃以230rpm摇动1小时。将细胞铺板在适当的选择抗生素上,伴随100mM葡萄糖以抑制碱基编辑器的表达,然后在37℃孵育过夜。第二天早上,将单一菌落接种到96孔板的单个孔中的含有抗生素的300μL DRM中(每个条件n=4个重复)。用多孔密封膜密封板,并且细胞在37℃摇动生长至饱和(约8小时)。将饱和培养物1:50稀释到1ml含有抗生素的DRM中,并生长至对数中期(约1.5小时)。为了诱导碱基编辑器的表达,向培养物中加入阿拉伯糖(最终浓度为30mM),并且细胞在37℃以230RPM摇动过夜生长。16小时后,细胞通过用多通道移液器混合来重悬,并从每个孔中转移60μL到PCR板中。使用热循环仪(BioRad)在95℃煮沸8分钟裂解细胞。细胞裂解物在分析前储存在-20℃。

[0815] 对于高通量测序,使用1μL大肠杆菌裂解物作为PCR模板,用Nextera HTS引物扩增

以安装表2中所示的适体。使用Phusion HiFi聚合酶(New England BioLabs)进行扩增。条形码化和高通量测序如下文所述进行用于哺乳动物细胞实验。

[0816] 一般哺乳动物细胞培养

[0817] HEK293T(ATCC CRL-3216)细胞购自ATCC,并在补充有10％(v/v)胎牛血清(Gibco,合格)的Dulbecco改良Eagle培养基(DMEM)加GlutaMAX(ThermoFisher Scientific)中培养。细胞在37℃和5％ CO2下孵育、维持和培养。细胞系经其各自供应商认证,并检测为支原体阴性。未分化的129P2/OlaHsd mESC(雄性)系如先前所述维持[11]。简而言之,细胞在明胶包被的板上在mESC培养基(Knockout DMEM(Life Technologies)、0.55mM 2-巯基乙醇(Sigma)和1x ESGRO LIF(Millipore)、5nM GEK-3抑制剂XV和500nM UO123中维持。细胞在37℃和5％ CO2下孵育、维持和培养。细胞系经其各自供应商认证,并支原体检测呈阴性。

[0818] HEK293T细胞转染

[0819] 在转染前16-24小时,将细胞以每孔1.5x 104个细胞的密度接种到96孔板(Corning)上。转染条件如下:0.5μL Lipofectamine 2000(Thermo Fisher Scientific)、100ng编辑器质粒和40ng向导RNA质粒混合,并用Opti-MEM减血清培养基(Thermo Fisher Scientific)稀释至总体积12.5μL,并按照制造商的方案进行转染。细胞在约60％-80％汇合度时转染。

[0820] 从哺乳动物细胞培养物中分离基因组DNA

[0821] 转染后,细胞培养3天,之后去除培养基,用1x PBS溶液(100μL)清洗细胞,并经由每孔加入50μL裂解缓冲液(10mM Tris-HCl,pH 8.0,0.05％SDS,20μg/ml蛋白酶K(New England BioLabs))进行细胞裂解以收获基因组DNA。细胞裂解混合物在37℃孵育1-1.5小时,然后转移到96孔PCR板中,并在80℃下进行酶灭活30分钟。所得基因组DNA混合物储存在-20℃直至分析。

[0822] 从体外转录的碱基编辑器mRNA的生成

[0823] 碱基编辑器mRNA由PCR产物生成,该PCR产物由模板质粒扩增,该模板质粒含有如前所述克隆的感兴趣的碱基编辑器的表达载体[8]。使用正向引物IVT-F和反向引物IVT-R(表4)在200μL总反应中扩增PCR产物,使用QIAquick PCR纯化试剂盒(Qiagen)纯化,并在50μL无核酸酶水中洗脱。使用HiScribe T7-High-yield RNA Synthesis试剂盒(New England BioLabs)根据制造商的方案进行体外转录,但使用N1-甲基-假尿苷(TriLink Biotechnologies)代替尿苷的完全取代以及用CleanCap AG(TriLink Biotechnologies)的共转录加帽。通过氯化锂沉淀进行mRNA分离。简言之,对于160μL IVT反应,添加0.5体积的7.5M氯化锂(240μL最终体积)并通过移液混合。将混合物在4℃孵育20分钟后,将样品以15,000×g离心20分钟。弃去上清液,并用400μL冰冷的70％乙醇重悬沉淀。将混合物在4℃下离心15分钟,并且再次弃去上清液。将所得沉淀在室温下风干5分钟,然后重悬于100-200μL无核酸酶水中。将再悬浮液的等分试样稀释5倍以通过nanodrop进行定量。将样品归一化至2μg/μL并在80℃储存。

[0824] 将TadCBE mRNA和sgRNA电穿孔到T细胞或HSC中

[0825] 来自去识别的人供体(n＝4)的血沉棕黄层获自Memorial Blood Centers(St.Paul,MN),并使用Lymphoprep和SepMate管(STEMCELL Technologies,Vancouver,Canada)分离外周血单核细胞。从中,使用EasySep Human CD4+T细胞分离试剂盒(STEMCELL

Technologies,Vancouver,Canada)纯化CD4+细胞,然后用Dynabeads™Human T-Expander CD3/CD28珠(Thermo Fisher Scientific,Waltham,MA)激活,并在X-VIVO™ 15无血清造血细胞培养基(Lonza,Basel,Switzerland)中培养,其中含有:5% AB人血清(Valley Biomedical,Winchester,VA)、GlutaMAX(Gibco,Waltham,MA)、N-乙酰半胱氨酸(Sigma Aldrich,St.Louis,MO)、50U/ml青霉素和50μg/ml链霉素(Gibco,Waltham,MA)以及300IU/ml IL-2。72小时后,去除珠,并使用Neon电穿孔系统(ThermoFisher,Waltham,MA)用10μl吸头用2μg候选碱基编辑器mRNA和100pmol sgRNA(Synthego,Redwood City,CA)电穿孔300,000个T细胞。本实验中使用的化学合成向导RNA的序列列于表5中。

[0826] 从Fred Hutchinson癌症研究中心(Seattle,WA)的血液学卓越核心中心获得没有任何识别供体信息的CD34+细胞,并在SFEMII培养基(STEMCELL Technologies,Vancouver,Canada)中培养,其中含有:50U/ml青霉素和50μg/ml链霉素(Gibco,Waltham,MA)、100ng/ml重组人血小板生成素、干细胞因子(TPO;BioLegend,San Diego,CA)、Flt-3配体和IL-6(Peprotech,Cranbury,NJ)以及0.75μM StemRegenin1和500nM UM729(STEMCELL Technologies,Vancouver,Canada)。解冻后48小时(n=3个供体),使用Amaxa(Lonza,Basel,Switzerland)4D-Nucleofector方案将2μg编辑器mRNA和100pmol sgRNA电穿孔到200,000HSC中,该方案针对Nucleovette条中的P3原代细胞系4D Nucleofector试剂盒,程序DZ-100。本实验中使用的化学合成向导RNA的序列列于表5中。

[0827] 基因转移后72小时,使用QuickExtract™DNA提取溶液(Madison,WI)收获细胞沉淀的DNA。使用Phusion U Multiplex PCR Master Mix(Thermo Fisher Scientific,Waltham,MA)在以下条件下进行Illumina测序的PCR扩增:98℃ 30秒;98℃ 10秒、64℃ 30秒和72℃ 20秒的30-35个循环;和最后72℃ 5分钟。

[0828] 基因组DNA样品的高通量DNA测序

[0829] 哺乳动物细胞系基因组DNA的高通量测序如先前所述进行[4]。用于靶基因组位点的PCR扩增的引物列于表2A-2E中。靶扩增子的序列列于表2A-2E中。在按照制造商的方案在Illumina MiSeq仪器上进行测序之前,使用Qubit dsDNA高灵敏度测定试剂盒(Thermo Fisher Scientific)通过qPCR与KAPA DNA定量试剂盒(Roche)定量DNA浓度。

[0830] Cas非依赖性RNA编辑的分析

[0831] RNA脱靶编辑分析如先前所述进行[15]。简而言之,HEK293T细胞的平行板用250ng的编码编辑器的质粒和83ng的EMX1向导RNA质粒转染,如上所述。一个板用于评估EMX1基因座处的中靶基因组DNA编辑,如上所述。另一个板用于如下的RNA编辑分析:转染后48小时,使用RNeasy试剂盒(Qiagen)按照制造商的说明裂解细胞。简而言之,去除培养基并用PBS清洗细胞,然后在RLT Plus缓冲液(QIAGEN)中裂解。将细胞转移到DNA消除柱中。将乙醇加入流穿液中,然后转移到RNeasy旋转柱中。样品用RW1清洗,然后用RDD缓冲液(QIAGEN®)中的RNase-Free DNase进行柱上DNA消化。样品然后用RW1缓冲液清洗,随后用RPE缓冲液清洗。RNA在45μl无核酸酶水中洗脱,并向每个样品中加入2μl RNaseOUT(Thermo Fisher Scientific)。

[0832] 使用SuperScript IV第一链合成试剂盒(Thermo Fisher Scientific)按照制造商的说明生成互补DNA。OligodT引物通过在65℃加热然后冰上冷却1分钟退火到RNA上。准备逆转录反应并加入退火混合物中。无逆转录酶对照包括在内作为gDNA污染的对照。反应

在50℃下孵育10分钟,80℃下孵育10分钟,然后在冰上冷却1分钟。进行任选的RNaseH RNA降解以增加cDNA扩增效率。使用1μl的每个cDNA样品进行扩增子测序的第一次PCR;剩余的测序方案与DNA测序使用的方案相同。用于第一次PCR的引物列于表3中。

[0833] TadCBE编辑结果的文库分析

[0834] 通过将新编辑器序列克隆到先前描述的p2T-CMV-AID-BE4max-BlastR质粒[11]中构建碱基编辑器质粒。含有先前报告的10,683成员“全面12kChar”文库[11]的未分化129P2/OlaHsd mESC(雄性)系。细胞解冻并如先前所述[11]在15cm板上维持。为了将碱基编辑器质粒整合到含有整合文库的细胞系中,使用Lipofectamine 3000用Tol2转座酶质粒转染细胞,并在转染后第二天开始用杀稻瘟菌素选择4天,然后收获。在整个过程维持每个文库盒300x的平均覆盖率。每个碱基编辑器进行两个生物学重复。在抗生素选择后4天(碱基编辑器转染后5天)从细胞中收集基因组DNA。对于文库样品,每个样品使用20μg基因组DNA进行PCR1扩增和测序,并保持每个靶标的平均测序深度为2,800x。使用表8中指定的引物进行PCR1以扩增内源性基因座或文库盒。使用NEBNext Index Primer set1和2(New England Biolabs)进行PCR2以添加全长Illumina测序适体。所有PCR反应均使用NEBNext Ultra IIQ5Master Mix进行。所有PCR反应的延伸时间延长至每循环2分钟以防止PCR扩增偏差。样品通过Tape Station(Agilent)定量,合并,并使用KAPA文库定量试剂盒(Roche)定量后进行测序。文库测序在Illumina NextSeq上进行,使用配对末端读段(94个正向;56个反向)。

[0835] 数据处理和分析使用Python 3.9进行。文库样品使用bcl2fastq2(Illumina)对每个编辑器/重复进行解复用,其中所有通道合并。为了将每个配对末端读段分配给文库成员,丢弃靶位点和sgRNA间隔区序列中低于Q28的任何读段。然后使用局部敏感哈希用靶位点上的平铺6-mer提名候选靶位点。过滤掉sgRNA间隔区序列与候选靶位点不匹配的任何读段。最后,通过执行Needleman-Wunsch比对(评分参数:匹配=1,错配=-1,空位开放=-5,空位延伸=0,起始空位=0)对每个靶位点进行基因分型。

[0836] 在进一步数据分析之前,考虑了测序数据中的两个噪声源。首先,含有基因组整合文库的mESC细胞系的扩增可能导致选择后初始细胞文库中存在的误差的随机扩增(所谓的“批次效应”)。其次,Illumina系统上的下一代测序偶尔会错误分配读段。为了最小化这两个误差源,仅考虑-9至20碱基编辑窗口内的A至G、C至T、C至G和C至A突变。

[0837] 通过使用单向ANOVA比较窗口内每个位置的每个突变的频率,研究了突变数据中的潜在批次效应。令人鼓舞的是,窗口内没有超出统计噪声范围的批次特异性突变(Bonferroni校正显著性水平为0.005)。

[0838] 最后,过滤掉可能是由于Illumina测序噪声的读段。考虑到由于罕见碱基编辑结果的突变可能仍然在两个文库重复中存在,即使它们在每个重复中的存在低于传统上被认为是噪声的阈值。因此,基于使用速率参数为10^{-3}(Q30)的伯努利分布,计算了在两个重复中观察到每个位置上的每个突变在相应读段数中的可能性。所有由于测序噪声的可能性小于5%的突变都被保留。

[0839] 对于位置编辑效率分析,将含有每个突变的读段数在重复之间合并,并除以每个给定文库成员观察到的总读段数。以这种方式合并重复(而不是例如平均每个重复的频率),因为它是描述给定位置碱基编辑效率的假设伯努利分布的速率参数的最大似然估计。

[0840] 在本分析中,“平均编辑效率”定义为整个文库中含有指定编辑结果的(噪声过滤

的、批次合并的)读段的平均分数。为了定义相较于腺嘌呤脱氨对胞嘧啶的选择性,首先计算文库所有成员中≥30%编辑窗口内位置处的平均胞嘧啶编辑效率和平均腺嘌呤编辑效率。然后计算每个位置选择性的几何平均值,以获得每个编辑器"总体"选择性的保守估计。由于给定位置只能含有胞嘧啶或腺嘌呤,因此在给定场景中的真实选择性将取决于各自碱基的位置。

[0841] 为了生成这些编辑器的背景偏好的序列基序,首先用稳定化logit函数转化编辑分数:$\log\left(\frac{x+\epsilon}{1+\epsilon-x}\right)$,其中∈是一个小常数,用于稳定输入接近0或1时的函数行为。对于这些目的,使用∈=0.001,因为这是由于Illumina测序的噪声的保守估计。然后进行随机训练/测试分割(80:20),并使用$\alpha=10^{-5}$训练岭回归以生成在序列标志中可视化的权重。

[0842] 为了评估在TadCBEd中包含V106W突变后C·G至T·A和A·T至C·G转化效率的倍数变化,对每个编辑器安装指定编辑的(噪声过滤的、批次校正的)效率进行总最小二乘(TLS)回归。执行TLS而不是普通最小二乘,因为计算涉及两个测量变量之间的关系(而不是一个变量对另一个独立变量的依赖性)。平均倍数减少定义为回归权重的倒数(其中x是TadCBEd,y是TadCBEd-V106W)。

[0843] DNA测序和靶向扩增子测序的HTS数据的分析

[0844] 使用MiSeq Reporter(Illumina)对单个高通量测序数据集进行解复用。随后的解复用测序读段使用CRISPResso2[77]进行分析,如先前所述[15]。所有编辑值代表n=3个独立生物学重复,其中显示平均值±SEM。

[0845] 实施例6.TadCBE对哺乳动物细胞中10,683个基因组靶位点的广泛表征

[0846] TadCBE的活性可能因靶位点而显著变化(图3)。为了全面表征TadCBE在哺乳动物细胞中广泛范围的位点间的活性,使用先前报道的10,683个配对sgRNA和整合到小鼠胚胎干细胞系(mESC,图37)中的靶位点的"全面背景文库"[11]对TadCBE变体进行了碱基编辑结果的高通量分析。这些文库包括具有在原间隔区位置6处的底物A或C核苷酸周围所有可能的6-mer以及跨位置-1到13(将紧邻原间隔区上游的位置计为位置0)的所有可能的5-mer的靶位点,具有最小的序列偏倚[11]。优化了碱基编辑条件以允许待检测的碱基编辑器之间的差异。在整个实验过程中,每个文库成员的平均细胞覆盖率≥300x,维持每个靶标的平均测序深度≥2,800×,这使得能够以高灵敏度检测编辑结果。针对TadCBEa-e、TadCBEa-d的V106W变体和TadDE以及作为参考的BE4max每个碱基编辑器收集了两个生物学重复[11],并验证了文库测定数据在生物学重复之间具有很强的一致性(图38)。

[0847] 所得文库数据用于量化每个TadCBE的编辑活性和C·G至T·A选择性(图32A)。在10,638个整合的靶位点中,所有TadCBE和TadCBE-V106W变体相较于BE4max(21%)具有更大的平均编辑效率(平均28%-31%的读段具有任何C·G至T·A编辑)(图32A)[11]。然后表征了编辑窗口,其定义为原间隔区内平均≥峰值平均编辑效率30%的位置(图32B、图39)。TadCBE编辑通常以在原间隔区位置6周围为中心。最高活性的变体TadCBEd具有与BE4max的编辑窗口(位置3-9)相似的编辑窗口(原间隔区位置3-9),而其余TadCBE和V106W-TadCBE具有略窄的窗口(位置3-8,图32B、图39)。

[0848] TadCBE相对于腺嘌呤编辑对胞嘧啶编辑的选择性因碱基编辑器而异。在典型TadCBE(不含V106W)中,TadCBEd显示出最高的C·G至T·A选择性,其编辑窗口内每个位置的C·G至T·A与A·T至G·C编辑比率的几何平均值为26.8(表6)。值得注意的是,添加

V106W显著改进了所有TadCBE变体的C•G至T•A选择性(TadCBEd V106W选择性=47.8),同时最小程度影响了最大编辑位置的碱基编辑活性(图32A-32B、图40A-40B)。例如,在TadCBEd中添加V106W将文库靶标中的峰值编辑从35%降低到31%。

[0849] 与上文显示的离散靶位点示例一致,TadCBE的C•G至T•A选择性在全面背景文库中因靶位点而异。在全面背景文库的所有靶位点中,对于TadCBE在3.4%-6.6%的读段(平均)中以及对于V106W-TadCBE在1.0%-2.7%的读段中观察到腺嘌呤碱基编辑(图32A)。通过对编辑效率进行回归分析生成序列基序,以确定影响胞嘧啶和腺嘌呤脱氨的序列特征(图32C、图41A和41B)。TadCBE与其祖先ABE7.10具有相似的序列背景偏好,偏好编辑5'Y(Y=T/C)之前的胞嘧啶和腺嘌呤碱基,而不偏好5'A[11]。在进行旁观者腺嘌呤碱基编辑时,TadCBE保留了ABE7.10的序列背景偏好(偏好5'YAY且不偏好5'AAA)。然而,TadCBE反而略微不偏好5'ACT。3'偏好的差异可能是由于实现改变的选择性所需的底物定位的差异,因为与相邻碱基的相互作用可能会改变靶胞苷在活性位点中的位置(图10A-10C)。值得注意的是,腺嘌呤碱基编辑的靶序列依赖性(TadCBEd测试R:0.64)相较于胞嘧啶碱基编辑(TadCBEd测试R:0.36)更强。

[0850] TadDE进行非常相似水平的腺嘌呤和胞嘧啶碱基编辑(ABE:CBE比率=1.1),并且具有与TadCBE相似的序列背景依赖性(图32A和32B、图41A和41B、表6)。TadDE非常高效,在文库实验中平均编辑35%的读段(图32A)。在观察到C•G至T•A编辑的情况下,针对TadDE观察到A•T至G•C编辑的概率为0.62,相比之下最具选择性的TadCBE变体TadCBEd-V106W为0.04(表6)。TadDE的高活性、混杂性和小尺寸使其成为并行A•T至G•C和C•G至T•A编辑的有前途的工具。

[0851] 总的来说,这些数据表明TadCBE比传统的窄窗口CBE具有更大的胞嘧啶脱氨活性。此外,在脱氨酶结构域中引入V106W降低了残留的A•T至G•C编辑活性,同时最小程度影响本次实验中所有TadCBE的C•G至T•A编辑。总体而言,TadCBEd实现了最大的胞嘧啶脱氨活性和高C•G至T•A选择性,这通过添加V106W进一步改进。

[0852] 实施例7.TadCBE与Cas9直系同源物的兼容性及编辑窗口表征

[0853] 使用具有不同PAM需求的Cas9直系同源物扩展了碱基编辑器的可靶向序列空间。为了测试TadCBE是否与除酿脓链球菌Cas9之外的Cas9同源物兼容,构建了既有来自脑膜炎奈瑟菌的Nme2Cas9的PACE进化变体的TadCBE变体,这些变体扩展了除SpCas9的典型NGG PAM之外的可及PAM的范围[62]。最近进化了作为核酸酶或作为碱基编辑器接近广泛的单嘧啶PAM位点的Nme2Cas9变体[63]。使用eNme2-C变体切口酶(SEQ ID NO:353)(PAM=N_4CN)和两个UGI结构域生成TadA-CD的融合物,所得eNme2-C-TadCBE与向导RNA质粒共转染,并在HEK293T细胞中的六个基因组基因座处检查碱基编辑。在所有测试的位点中,TadCBE的峰值编辑效率与APOBEC1、evoFERNY和evoAPOBEC1的相当(图42-43)。尽管在一些位点C•G至T•A编辑超过50%,但在测试的六个eNme2靶位点的任一个处,残留的A•T至G•C编辑从未超过5.3%。因此,TadCBE使用eNme2 Cas9表现出强大的活性和选择性。

[0854] 接下来在BE4max架构中使用金黄色葡萄球菌Cas9(SaCas9)测试了TadCBE[64]。SaCas9(1053个氨基酸)比SpCas9(1368个氨基酸)小,并识别不同的PAM序列(PAM=NNGRRT)。发现使用SaCas9的TadCBEs在9个位点上具有强大的C•G至T•A编辑(4.1%-44%),且在任何位点处的A•T至G•C编辑均低于5.5%(图44-45)。这些观察结果表明与其

他Cas蛋白的潜在兼容性,这些Cas蛋白与SpCas9、eNme2-C Cas9和SaCas9一起可以提供对多种PAM序列的接近,用于TadCBE的多功能靶向。此外,还发现TadDE在哺乳动物细胞中使用SpCas9、eNme2-C Cas9和SaCas9在TadCBE具有选择性的位点上同时进行A·T至G·C和C·G至T·A编辑,表明双重编辑器也具有广泛的Cas9兼容性(图42-47)。

[0855] TadCBE表现出比BE4max、evoA和evoFERNY CBE更窄的编辑窗口,同时维持相当或更高的最大编辑效率(图42)。例如,BE4max和evoA在原间隔区位置3-18处以4.2%-47%的效率编辑Nme50,而TadCBEa、TadCBEb和TadCBEc仅在较窄的位置3-8窗口上以5%-48%的效率进行编辑(图42)。TadCBE的较窄的碱基编辑活性窗口可能源于较少的持续脱氨酶,因为APOBEC家族脱氨酶的进行性性质可以在每次DNA结合事件中催化多个水解脱氨反应[65]。虽然宽编辑窗口对于一些应用(如靶向基因破坏或碱基编辑筛选)有用,但TadCBE的较窄窗口应有利于期望仅一个靶碱基的修饰的精确编辑应用,特别是在使用支持更宽碱基编辑窗口的Cas9结构域时[62,66]。总的来说,TadCBE的小尺寸、其与eNmeCas9和SaCas9的兼容性、其更集中的编辑窗口以及其相较于腺嘌呤碱基编辑对胞嘧啶的高编辑效率和选择性,证明了其在各种精确胞嘧啶碱基编辑应用中的适用性。

[0856] 实施例8.使用噬菌体辅助进化从TadA双重碱基编辑器开发活性和选择性胞嘧啶碱基编辑器。

[0857] 在本实施例中,描述了活性和选择性胞嘧啶碱基编辑器用于在疾病相关位点安装终止密码子的用途。为此,具有选择回路和诱变质粒(红色)的大肠杆菌宿主细胞的连续流由编码部分脱氨酶的选择噬菌体(SP)感染。在选择回路中,噬菌体繁殖与gIII(P2)的表达相关联,gIII只能用活性T7 RNA聚合酶转录。T7 RNA聚合酶(P3)与C端降解决定子融合,并且脱氨酶必须执行C至U编辑以在降解决定子之前安装终止密码子,从而产生活性T7 RNA聚合酶。在噬菌体感染的事件中,使用分裂内含肽系统(P1)完成完整脱氨酶,并且脱氨酶上可以发生突变。有益突变导致噬菌体在培养池中繁殖并富集,而适应性较差的噬菌体无法繁殖并随后被恒定流出物洗出。

[0858] 所得变体在脱氨酶的位置N46处鉴定出保守突变,因此在位置N46处构建了NNK文库,并对这些变体进行了PANCE。为了进一步增加严格性,在来自两个PANCE的所得变体上进行了>100小时的PACE。稀释倍数在右侧y轴上指示。图52A-52E显示了来自PANCE和PACE的示例性突变表,其中描述了保守突变。

[0859] 实施例9.在大肠杆菌中TadCBE的活性和序列背景特异性的谱分析。

[0860] 32成员的单链DNA文库(IDT oligopools)设计为含有在原间隔区位置6处的靶碱基(A或C),其中5'和3'碱基变化为A、T、C或G。每个文库成员含有独特的分子标识符(UMI)条形码。使用引物对MN1591/MN1592和KAPA聚合酶,在200μl反应体积中使用1.5nM模板用68℃退火温度和3分钟延伸时间,扩增单链寡核苷酸三个循环。纯化PCR产物(Qiagen)并使用Gibson(NEB)将其组装到BamHI/EcoRI消化的质粒MNp553中。在使用Glyco-blue(Thermo Fisher)的纯化后,将文库转化到NEB 10-beta电感受态细胞中。将细胞稀释物立即铺板以计算文库大小,然后在羧苄青霉素中过夜生长剩余转化体以选择转化体。第二天,通过midiprep(Qiagen)纯化文库质粒。

[0861] 同时,制备含有指定的感兴趣编辑器质粒的电感受态NEB10-beta细胞,然后在DRM中生长以抑制表达。然后将40μl含有编辑器的电感受态细胞用100ng文库质粒电穿孔,在

1ml S.O.C.培养基中挽救5分钟,在10ml DRM中稀释,并使用壮观霉素、羧苄青霉素和30mM阿拉伯糖过夜生长以诱导编辑器表达。在37℃和200rpm摇动生长16小时后,通过miniprep分离质粒。使用1μl质粒作为PCR1和HTS分析的模板,如下所述。

[0862] 为了分析文库的编辑结果,使用MiSeq Reporter(Illumina)解复用测序读段,然后使用SeqKit将其分选为靶扩增子。然后使用CRISPResso2对输出进行测序。结果如图53所示。

[0863] 实验10.经进化的活性和选择性胞嘧啶碱基编辑器与现有胞嘧啶碱基编辑器在哺乳动物细胞中的比较。

[0864] 将TadDE N46变体与在BE4max架构中具有SpCas9切口酶的现有胞嘧啶碱基编辑器以及靶向三个原间隔区的向导RNA一起转染到HEK293T细胞中。TadDE N46变体显示出相当的中靶活性,且无残留的A至G编辑。点代表来自独立生物学重复的个体值。PAM序列加下划线。HEK293T位点2缩写为HEK2,HEK293T位点4缩写为HEK4。

[0865] 将TadDE N46变体与在BE4max架构中具有eNme-Cas9切口酶的现有胞嘧啶碱基编辑器以及靶向两个原间隔区的向导RNA一起转染到HEK293T细胞中。TadDE N46变体显示出更高或相当的中靶活性,且无残留的A至G编辑。点代表来自独立生物学重复的个体值。PAM序列加下划线。来自该实验的结果如图54所示。

[0866] 参考文献

[0867] 1.Eisenstein,M.Base editing marches on the clinic.Nat Biotechnol 40,623-625(2022).

[0868] 2.ISRCTN-ISRCTN15323014:CAR T cells to fight T cell leukaemia.https://www.isrctn.com/ISRCTN15323014doi:10.1186/ISRCTN15323014.

[0869] 3.Komor,A.C.,Kim,Y.B.,Packer,M.S.,Zuris,J.A.&Liu,D.R.Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage.Nature 533,420-424(2016).

[0870] 4.Gaudelli,N.M.et al.Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage.Nature 551,464-471(2017).

[0871] 5.Mok,B.Y.et al.A bacterial cytidine deaminase toxin enables CRISPR-free mitochondrial base editing.Nature 583,631-637(2020).

[0872] 6.Mok,B.Y.et al.CRISPR-free base editors with enhanced activity and expanded targeting scope in mitochondrial and nuclear DNA.Nat Biotechnol(2022)doi:10.1038/s41587-022-01256-8.

[0873] 7.Nishida,K.et al.Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems.Science 353,aaf8729(2016).

[0874] 8.Gaudelli,N.M.et al.Directed evolution of adenine base editors with increased activity and therapeutic application.Nat Biotechnol38,892-900(2020).

[0875] 9.Richter,M.F.et al.Phage-assisted evolution of an adenine base editor with improved Cas domain compatibility and activity.Nat Biotechnol 38,883-891(2020).

[0876] 10.Cho,S.-I.et al.Targeted A-to-G base editing in human mitochondrial DNA with programmabledeaminases.Cell 185,1764-1776.e12(2022).

[0877] 11.Arbab,M.et al.Determinants of Base Editing Outcomes from Target Library Analysis and Machine Learning.Cell 182,463-480.e30(2020).

[0878] 12.Thuronyi,B.W.et al.Continuous evolution of base editors with expanded target compatibility and improved activity.Nat Biotechnol 37,1070-1079(2019).

[0879] 13.Jin,S.et al.Cytosine,but not adenine,base editors induce genome-wide off-target mutations in rice.Science 364,292-295(2019).

[0880] 14.Zuo,E.et al.Cytosine base editor generates substantial off-target single-nucleotide variants in mouse embryos.Science 364,289-292(2019).

[0881] 15.Doman,J.L.,Raguram,A.,Newby,G.A.&Liu,D.R.Evaluation and minimization of Cas9-independent off-target DNA editing by cytosine base editors.Nat Biotechnol 38,620-628(2020).

[0882] 16.Chester,A.,Weinreb,V.,Carter,C.W.&Navaratnam,N.Optimization of apolipoprotein B mRNAediting by APOBEC1 apoenzyme and the role of its auxiliary factor,ACF.RNA 10,1399-1411(2004).

[0883] 17.Kim,J.et al.Structural and kinetic characterization of Escherichia coli TadA,the wobble-specific tRNA deaminase.Biochemistry 45,6407-6416(2006).

[0884] 18.Lapinaite,A.et al.DNA capture by a CRISPR-Cas9-guided adenine base editor.Science 369,566-571(2020).

[0885] 19.Yu,Y.et al.Cytosine base editors with minimized unguided DNA and RNA off-target events and high on-target activity.Nat Commun 11,2052(2020).

[0886] 20.Rees,H.A.et al.Improving the DNA specificity and applicability of base editing through protein engineering and protein delivery.Nat Commun 8,15790(2017).

[0887] 21.Grünewald,J.et al.Transcriptome-wide off-target RNA editing induced by CRISPR-guided DNA base editors.Nature 569,433-437(2019).

[0888] 22.Kim,Y.B.et al.Increasing the genome-targeting scope and precision of base editing with engineered Cas9-cytidine deaminase fusions.Nat Biotechnol35,371-376(2017).

[0889] 23.Gehrke,J.M.et al.An APOBEC3A-Cas9 base editor with minimized bystander and off-target activities.Nat Biotechnol 36,977-982(2018).

[0890] 24.Berríos,K.N.et al.Controllable genome editing with split-engineered base editors.Nat Chem Biol 17,1262-1270(2021).

[0891] 25.Qiao,Q.et al.AID Recognizes Structured DNA for Class Switch Recombination.Mol Cell 67,361-373.e4(2017).

[0892] 26.Wang,X.et al.Efficient base editing in methylated regions with a human APOBEC3A-Cas9 fusion.Nat Biotechnol 36,946-949(2018).

[0893] 27.Davis,J.R.et al.Efficient in vivo base editing via single adeno-associated viruses with size-optimized genomes encoding compact adenine base editors.Nat Biomed Eng(2022)doi:10.1038/s41551-022-00911-4.

[0894] 28.Zhang,H.et al.Adenine Base Editing in vivo with a Single Adeno-Associated Virus Vector.bioRxiv 2021.12.13.472434 Preprint at https://doi.org/10.1101/2021.12.13.472434(2022).

[0895] 29.Iyer,L.M.,Zhang,D.,Rogozin,I.B.&Aravind,L.Evolution of the deaminase fold and multiple origins of eukaryotic editing and mutagenic nucleic acid deaminases from bacterial toxin systems.Nucleic Acids Res 39,9473-9497(2011).

[0896] 30.Rubio,M.A.T.et al.An adenosine-to-inosine tRNA-editing enzyme that can perform C-to-U deamination of DNA.Proc Natl Acad Sci U S A 104,7821-7826(2007).

[0897] 31.Kim,H.S.,Jeong,Y.K.,Hur,J.K.,Kim,J.-S.&Bae,S.Adenine base editors catalyze cytosine conversions in human cells.Nat Biotechnol 37,1145-1148(2019).

[0898] 32.Jeong,Y.K.et al.Adenine base editor engineering reduces editing of bystander cytosines.Nat Biotechnol 39,1426-1433(2021).

[0899] 33.Abudayyeh,O.O.et al.A cytosine deaminase for programmable single-base RNA editing.Science 365,382-386(2019).

[0900] 34.Rees,H.A.,Wilson,C.,Doman,J.L.&Liu,D.R.Analysis and minimization of cellular RNA editing by DNA adenine base editors.Sci Adv 5,eaax5717(2019).

[0901] 35.Badran,A.H.et al.Continuous evolution of Bacillus thuringiensis toxins overcomes insect resistance.Nature 533,58-63(2016).

[0902] 36.Bryson,D.I.et al.Continuous directed evolution of aminoacyl-tRNA synthetases.Nat Chem Biol 13,1253-1260(2017).

[0903] 37.Hubbard,B.P.et al.Continuous directed evolution of DNA-binding proteins to improve TALEN specificity.Nat Methods 12,939-942(2015).

[0904] 38.Miller,S.M.et al.Continuous evolution of SpCas9 variants compatible with non-G PAMs.Nat Biotechnol 38,471-481(2020).

[0905] 39.Brödel,A.K.,Rodrigues,R.,Jaramillo,A.&Isalan,M.Accelerated evolution of a minimal 63-amino acid dual transcription factor.Sci Adv6,eaba2728(2020).

[0906] 40.Dickinson,B.C.,Packer,M.S.,Badran,A.H.&Liu,D.R.Asystem for the continuous directed evolution of proteases rapidly reveals drug-resistance mutations.Nat Commun 5,5352(2014).

[0907] 41.Blum,T.R.et al.Phage-assisted evolution of botulinum neurotoxin proteases with reprogrammed specificity.Science 371,803-810(2021).

[0908] 42.Pu,J.,Zinkus-Boltz,J.&Dickinson,B.C.Evolution of a split RNA

polymerase as a versatile biosensor platform.Nat Chem Biol 13,432-438(2017).

[0909] 43.Roth,T.B.,Woolston,B.M.,Stephanopoulos,G.&Liu,D.R.Phage-Assisted Evolution of Bacillus methanolicus Methanol Dehydrogenase 2.ACS Synth Biol 8,796-806(2019).

[0910] 44.Jones,K.A.,Snodgrass,H.M.,Belsare,K.,Dickinson,B.C.&Lewis,J.C.Phage-Assisted Continuous Evolution and Selection of Enzymes for Chemical Synthesis.ACS Cent.Sci.7,1581-1590(2021).

[0911] 45.Johnston,C.W.,Badran,A.H.&Collins,J.J.Continuous bioactivity-dependent evolution of an antibiotic biosynthetic pathway.Nat Commun 11,4202(2020).

[0912] 46.Esvelt,K.M.,Carlson,J.C.&Liu,D.R.A system for the continuous directed evolution of biomolecules.Nature 472,499-503(2011).

[0913] 47.Badran,A.H.&Liu,D.R.Development of potent in vivo mutagenesis plasmids with broad mutational spectra.Nat Commun 6,8425(2015).

[0914] 48.Koblan,L.W.et al.Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction.Nat Biotechnol 36,843-846(2018).

[0915] 49.Sakata,R.C.et al.Base editors for simultaneous introduction of C-to-T and A-to-G mutations.Nat Biotechnol 38,865-869(2020).

[0916] 50.Xie,J.et al.ACBE,a new base editor for simultaneous C-to-T and A-to-G substitutions in mammalian systems.BMC Biol 18,131(2020).

[0917] 51.Zhang,X.et al.Dual base editor catalyzes both cytosine and adenine base conversions in human cells.Nat Biotechnol 38,856-860(2020).

[0918] 52.Grünewald,J.et al.A dual-deaminase CRISPR base editor enables concurrent adenine and cytosine editing.Nat Biotechnol 38,861-864(2020).

[0919] 53.Liang,Y.et al.AGBE:a dual deaminase-mediated base editor by fusing CGBE with ABE for creating a saturated mutant population with multiple editing patterns.Nucleic Acids Res 50,5384-5399(2022).

[0920] 54.Li,C.et al.Targeted,random mutagenesis of plant genes with dual cytosine and adenine base editors.Nat Biotechnol 38,875-882(2020).

[0921] 55.Hanna,R.E.et al.Massively parallel assessment of human variants with base editor screens.Cell 184,1064-1080.e20(2021).

[0922] 56.Cuella-Martin,R.et al.Functional interrogation of DNA damage response variants with base editing screens.Cell 184,1081-1097.e19(2021).

[0923] 57.Jumper,J.et al.Highly accurate protein structure prediction with AlphaFold.Nature 596,583-589(2021).

[0924] 58.Grünewald,J.et al.CRISPR DNA base editors with reduced RNA off-target and self-editing activities.Nat Biotechnol 37,1041-1048(2019).

[0925] 59.Zhou,C.et al.Off-target RNA mutation induced by DNA base editing

and its elimination by mutagenesis.Nature 571,275-278(2019).

[0926] 60.Park,S.&Beal,P.A.Off-Target Editing by CRISPR-Guided DNA Base Editors.Biochemistry 58,3727-3734(2019).

[0927] 61.Kleinstiver,B.P.et al.High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects.Nature 529,490-495(2016).

[0928] 62.Edraki,A.et al.A Compact,High-Accuracy Cas9 with a Dinucleotide PAM for In Vivo Genome Editing.Mol Cell 73,714-726.e4(2019).

[0929] 63.Huang,T.P.et al.High-throughput continuous evolution of compact Cas9 variants targeting single-nucleotide-pyrimidine PAMs.Nature Biotechnology Under revision,.

[0930] 64.Ran,F.A.et al.In vivo genome editing using Staphylococcus aureus Cas9.Nature 520,186-191(2015).

[0931] 65.Chelico,L.,Pham,P.,Calabrese,P.&Goodman,M.F.APOBEC3G DNA deaminase acts processively 3'-->5'on single-stranded DNA.Nat Struct Mol Biol 13,392-399(2006).

[0932] 66.Anzalone,A.V.,Koblan,L.W.&Liu,D.R.Genome editing with CRISPR-Cas nucleases,base editors,transposases and prime editors.Nat Biotechnol 38,824-844(2020).

[0933] 67.Song,Y.et al.Large-Fragment Deletions Induced by Cas9 Cleavage while Not in the BEs System.Mol Ther Nucleic Acids 21,523-526(2020).

[0934] 68.Kosicki,M.,Tomberg,K.&Bradley,A.Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements.Nat Biotechnol 36,765-771(2018).

[0935] 69.Ihry,R.J.et al.p53 inhibits CRISPR-Cas9 engineering in human pluripotent stem cells.Nat Med 24,939-946(2018).

[0936] 70.Alanis-Lobato,G.et al.Frequent loss of heterozygosity in CRISPR-Cas9-edited early human embryos.Proc Natl Acad Sci U S A 118,e2004832117(2021).

[0937] 71.Enache,O.M.et al.Cas9 activates the p53 pathway and selects for p53-inactivating mutations.Nat Genet 52,662-668(2020).

[0938] 72.Knipping,F.et al.Disruption of HIV-1 co-receptors CCR5 and CXCR4in primary human T cells and hematopoietic stem and progenitor cells using base editing.Mol Ther 30,130-144(2022).

[0939] 73.Canver,M.C.et al.BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis.Nature 527,192-197(2015).

[0940] 74.Wu,Y.et al.Highly efficient therapeutic gene editing of human hematopoietic stem cells.Nat Med 25,776-783(2019).

[0941] 75.Carlson,J.C.,Badran,A.H.,Guggiana-Nilo,D.A.&Liu,D.R.Negative selection and stringency modulation in phage-assisted continuous

evolution.Nat Chem Biol 10,216-222(2014).

[0942] 76.Miller,S.M.,Wang,T.&Liu,D.R.Phage-assisted continuous and non-continuous evolution.Nat Protoc 15,4101-4127(2020).

[0943] 77.Clement,K.et al.CRISPResso2 provides accurate and rapid genome editing sequence analysis.Nat Biotechnol 37,224-226(2019).

[0944] 78.Chen,L.,Zhu,B.,Ru,G.et al.Re-engineering the adenine deaminase TadA-8e for efficient and specific CRISPR-based cytosine base editing.Nat Biotechnol 41,663-672(2023).https://doi.org/10.1038/s41587-022-01532-7

[0945] 79.Lam,D.K.,Feliciano,P.R.,Arif,A.et al.Improved cytosine base editors generated from TadA variants.Nat Biotechnol 41,686-697(2023).https://doi.org/10.1038/s41587-022-01611-9

[0946] 80.Chadwick,A.C.,Wang,X.,&Musunuru,K.In Vivo Base Editing of PCSK9 (Proprotein Convertase Subtilisin/Kexin Type 9)as a Therapeutic Alternative to Genome Editing.Arteriosclerosis,thrombosis,and vascular biology,37(9),1741-1747(2017).https://doi.org/10.1161/ATVBAHA.117.309881

[0947] 实施例1-7的补充信息

[0948] 表1:本文所述的质粒和选择噬菌体(SP)。

[0949]

名称	用途(抗性)	来源	ORF1	ORF2
pBT44c	回路 1 和 2 (KmR)	p15a	proC.SD8.intC.dCas9.UGI	
pMN404	回路 1 (CbR)	SC101-E93K	proT7.sd8.gIII.luxAB	proLac.sgRNA [回路 1]
pMN405	回路 1 (SpR)	ColE1	proA.SynRBS 0.4k.T7-RNAP- 降解决定子	
pMN406	回路 1 (SpR)	ColE1	proB.SynRBS 0.4k.T7-RNAP- 降解决定子	
pMN407	回路 1 (SpR)	ColE1	proC.SynRBS 0.4k.T7-RNAP- 降解决定子	
pMN408	回路 1 (SpR)	ColE1	proD.SynRBS 0.4k.T7-RNAP- 降解决定子	
pBT120b	回路 2 (CbR)	SC101-E93K	proT7.sd8.gIII.luxAB	proLac.sgRNA [回路 2]
pBT2dR3-ProA	回路 2 (SpR)	ColE1	proA.SynRBS 0.4k.T7-RNAP- 降解决定子	
pBT2dR3-ProB	回路 2 (SpR)	ColE1	proB.SynRBS 0.4k.T7-RNAP- 降解决定子	
pBT2dR3-ProC	回路 2 (SpR)	ColE1	proC.SynRBS 0.4k.T7-RNAP- 降解决定子	
pBT2dR3-ProD	回路 2 (SpR)	ColE1	proD.SynRBS 0.4k.T7-RNAP- 降解决定子	
pMN618	大肠杆菌编辑 (CbR)	SC101	pBAD.SD8.EvoAPOBEC1-8e.dSpCas9-UGI-UGI	proLac.sgRNA [回路 2]
pMN619	大肠杆菌编辑 (CbR)	SC101	pBAD.SD8.TadA-8e.dSpCas9-UGI-UGI	proLac.sgRNA [回路 2]
pMN620	大肠杆菌编辑 (CbR)	SC101	pBAD.SD8.TadA-CDa.dSpCas9-UGI-UGI	proLac.sgRNA [回路 2]
pMN621	大肠杆菌编辑 (CbR)	SC101	pBAD.SD8.TadA-CDb.dSpCas9-UGI-UGI	proLac.sgRNA [回路 2]
pMN622	大肠杆菌编辑 (CbR)	SC101	pBAD.SD8.TadA-CDc.dSpCas9-UGI-UGI	proLac.sgRNA [回路 2]

[0950]

pMN415	大肠杆菌编辑靶标(KmR)	RSF1030	靶标：ACTTCGCGTTCGCCTGGACC, PAM=AGG	
pMN573	BE4max IVT 模板	pUC	proT7（突变).BE4max	
pMN574	EvoFERNY IVT 模板	pUC	proT7（突变).EvoFERNY	
pMN575	EvoAPOBEC IVT 模板	pUC	proT7 (突变).EvoA	
pMN576	TadCBEa IVT 模板	pUC	proT7（突变).TadCBEa	
pMN577	TadCBEb IVT 模板	pUC	proT7（突变).TadCBEb	
pMN578	TadCBEc IVT 模板	pUC	proT7（突变).TadCBEc	
pMN579	TadCBEd IVT 模板	pUC	proT7（突变).TadCBEd	
pMN580	TadCBEe IVT 模板	pUC	proT7（突变).TadCBEe	
pMN582	YE1 IVT 模板	pUC	proT7 (突变).YE1	
pMN731	TadCBEa V106W IVT 模板	pUC	proT7（突变).TadCBEa	
pMN732	TadCBEb V106W IVT 模板	pUC	proT7（突变).TadCBEb	
pMN733	TadCBEc V106W IVT 模板	pUC	proT7（突变).TadCBEc	
pMN734	TadCBEd V106W IVT 模板	pUC	proT7（突变).TadCBEd	
pMN735	TadCBEe V106W IVT 模板	pUC	proT7（突变).TadCBEe	
pMN446	哺乳动物表达，SpCas9	pUC	pCMV.TadCBEa	
pMN447	哺乳动物表达，SpCas9	pUC	pCMV.TadCBEb	
pMN448	哺乳动物表达，SpCas9	pUC	pCMV.TadCBEc	
pMN450	哺乳动物表达，SpCas9	pUC	pCMV.TadCBEd	
pMN584	哺乳动物表达，SpCas9	pUC	pCMV.TadCBEe	
pMN554	哺乳动物表达，SpCas9	pUC	pCMV.ABE8e V106W-UGI-UGI	
pMN556	哺乳动物表达，SpCas9	pUC	pCMV.TadCBEa V106W	

[0951]

pMN557	哺乳动物表达，SpCas9	pUC	pCMV.TadCBEb V106W	
pMN558	哺乳动物表达，SpCas9	pUC	pCMV.TadCBEc V106W	
pMN560	哺乳动物表达，SpCas9	pUC	pCMV.TadCBEd V106W	
pMN585	哺乳动物表达，SpCas9	pUC	pCMV.TadCBEe V106W	
pMN597	哺乳动物表达，SpCas9	pUC	pCMV.TadDE	
pMN599	哺乳动物表达，eNme2-C	pUC	pCMV.ABE8e-UGI-UGI	
pMN607	哺乳动物表达，eNme2-C	pUC	pCMV.BE4max (eNme2-C)	
pMN608	哺乳动物表达，eNme2-C	pUC	pCMV.EvoFERNY (eNme2-C)	
pMN609	哺乳动物表达，eNme2-C	pUC	pCMV.EvoA (eNme2-C)	
pMN601	哺乳动物表达，eNme2-C	pUC	pCMV.TadCBEa (eNme2-C)	
pMN602	哺乳动物表达，eNme2-C	pUC	pCMV.TadCBEb (eNme2-C)	
pMN603	哺乳动物表达，eNme2-C	pUC	pCMV.TadCBEc (eNme2-C)	
pMN605	哺乳动物表达，eNme2-C	pUC	pCMV.TadCBEd (eNme2-C)	
pMN606	哺乳动物表达，eNme2-C	pUC	pCMV.TadCBEe (eNme2-C)	
pMN709	哺乳动物表达，eNme2-C S6P	pUC	pCMV.TadDE (eNme2-C S6P)	
pMN745	哺乳动物表达，SaCas9	pUC	pCMV.BE4max (SaCas9)	
pMN746	哺乳动物表达，SaCas9	pUC	pCMV.evoA (SaCas9)	
pMN747	哺乳动物表达，SaCas9	pUC	pCMV.evoFERNY (SaCas9)	
pMN748	哺乳动物表达，SaCas9	pUC	pCMV.ABE8e-UGI-UGI (SaCas9)	
pMN749	哺乳动物表达，SaCas9	pUC	pCMV.TadCBEa (SaCas9)	
pMN750	哺乳动物表达，SaCas9	pUC	pCMV.TadCBEb (SaCas9)	
pMN751	哺乳动物表达，SaCas9	pUC	pCMV.TadCBEc (SaCas9)	
pMN752	哺乳动物表达，SaCas9	pUC	pCMV.TadCBEd (SaCas9)	
pMN753	哺乳动物表达，SaCas9	pUC	pCMV.TadCBEe (SaCas9)	

[0952]

pMN754	哺乳动物表达，SaCas9	pUC	pCMV.TadDE (SaCas9)	
pMN589	哺乳动物表达，SpCas9	pUC	pCMV.ABE8e E59A-UGI-UGI	
pMN739	哺乳动物表达，SpCas9 文库分析	pUC	p2T-CMV.BE4max.BlastR	
pMN717	哺乳动物表达，SpCas9 文库分析	pUC	p2T-CMV.TadCBEa.BlastR	
pMN718	哺乳动物表达，SpCas9 文库分析	pUC	p2T-CMV.TadCBEb.BlastR	
pMN719	哺乳动物表达，SpCas9 文库分析	pUC	p2T-CMV.TadCBEc.BlastR	
pMN720	哺乳动物表达，SpCas9 文库分析	pUC	p2T-CMV.TadCBEd.BlastR	
pMN721	哺乳动物表达，SpCas9 文库分析	pUC	p2T-CMV.TadCBEe.BlastR	
pMN723	哺乳动物表达，SpCas9 文库分析	pUC	p2T-CMV.TadCBEa V106W.BlastR	
pMN724	哺乳动物表达，SpCas9 文库分析	pUC	p2T-CMV.TadCBEb V106W.BlastR	
pMN725	哺乳动物表达，SpCas9 文库分析	pUC	p2T-CMV.TadCBEc V106W.BlastR	
pMN726	哺乳动物表达，SpCas9 文库分析	pUC	p2T-CMV.TadCBEd V106W.BlastR	
SPMN21	PA(N)CE, ΔgIII SP	M13 f1	p_gIII.SD4.TadA-8e-NpuN	

[0953] 表2A-2E:本文所述的靶原间隔区和扩增子以及用于基因组DNA扩增的对应引物。

[0954] 表2A:SpCas9基因组基因座:

[0955]

位点名称	原间隔区	PAM	扩增子	HTS-F	HTS-R
FANCF	GGAATCCCTTCTGCAGCACC (SEQ ID NO: 208)	TGG	CATTGCAGAGAGGCGTATCATTTCGCGGATGTTCCAATCAGTACGCAGAGAGTCGCCGTCTCCAAGGTGAAAGCGGAAGTAGGGCCTTCGCGCACCTCATGGAATCCCTTCTGCAGCACCTGGATCGCTTTTCCGAGCTTCTGGC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNCATTGCAGAGAGGCGT	TGGAGTTCAGACGTGTGCTCTTCCGATCTGGGGTCCCAGGTGCTGAC

[0956]

位点名称	原间隔区	PAM	扩增子	HTS-F	HTS-R
			GGTCTCAAGCACTACCTACGTCAGCACCTGGGACCCC (SEQ ID NO: 7)	ATCA (SEQ ID NO: 32)	(SEQ ID NO: 66)
EMX1	GAGTCCGAGCAGAAGAAGAA (SEQ ID NO: 209)	GGG	CAGCTCAGCCTGAGTGTTGAGGCCCCAGTGGCTGCTCTGGGGGCCTCCTGAGTTTCTCATCTGTGCCCCTCCCTCCCTGGCCCAGGTGAAGGTGTGGTTCCAGAACCGGAGGACAAAGTACAAACGGCAGAAGCTGGAGGAGGAAGGGCCTGAGTCCGAGCAGAAGAAGAAGGGCTCCCATCACATCAACCGGTGGCGCATTGCCACGAAGCAGGCCAATGGGGAGGACATCGATGTCACCTCCAATGACTAGGGTGGGCAACCACAAACCCACGAG (SEQ ID NO: 8)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNCAGCTCAGCCTGAGTGTTGA (SEQ ID NO: 55)	TGGAGTTCAGACGTGTGCTCTTCCGATCTCTCGTGGGTTTGTGGTTGC (SEQ ID NO: 67)
BCL11A	TTTATCACAGGCTCCAGGAA (SEQ ID NO: 210)	GGG	GCCAGAAAAGAGATATGGCATCTACTCTTAGACATAACACACCAGGGTCAATACAACTTTGAAGCTAGTCTAGTGCAAGCTAACAGTTGCTTTTATCACAGGCTCCAGGAAGGGTTTGGCCTCTGATTAGGGTGGGGGCGTGGGTGGGGTAGAAGAGGACTGGCAGACCTCTCCATCGGTGGCCGTTTGCCCAGGGGGGCCTCTTTCGGAAGGCTCTCT (SEQ ID NO: 9)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNGCCAGAAAAGAGATATGGCATC (SEQ ID NO: 56)	TGGAGTTCAGACGTGTGCTCTTCCGATCTAGAGAGCCTTCCGAAAGAGG (SEQ ID NO: 68)
HEK3	GGCCCAGACTGAGCACGTGA (SEQ ID NO: 211)	TGG	ATGTGGGCTGCCTAGAAAGGCATGGATGAGAGAAGCCTGGAGACAGGGATCCCAGGGAAACGCCCATGCAATTAGTCTATTTCTGCTGCAAGTAAGCATGCATTTGTAGGCTTGATGCTTTTTTTCTGCTTCTCCAGCCCTGGCCTGGGTCAATCCTTGGGGCCCAGACTGAGCACGTGATGGCAGAGGAAAGGAAGCCCTGCTTCCTCCAGAGGGCGTCGCAGGACAGCTTTTCCTAGACAGGGGCTAGTATGTGCAGCTCCTGCACCGGGATACTGGTTGACAAGTTTGGCTGGG (SEQ ID NO: 10)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNATGTGGGCTGCCTAGAAAGG (SEQ ID NO: 57)	TGGAGTTCAGACGTGTGCTCTTCCGATCTCCAGCCAAACTTGTCAACC (SEQ ID NO: 69)

[0957]

位点名称	原间隔区	PAM	扩增子	HTS-F	HTS-R
B2M	ACTCACGCTGGATAGCCTCC (SEQ ID NO: 212)	AGG	GGCGGGCATTCCTGAAGCTGACAGCATTCGGGCCGAGATGTCTCGCTCCGTGGCCTTAGCTGTGCTCGCGCTACTCTCTCTTTCTGGCCTGGAGGCTATCCAGCGTGAGTCTCTCCTACCCTCCCGCTCTGGTCCTTCCTCTCCCGCTCTGCACCCTCTGTGGCCCTCGCTGTGCTCTCTCGCTCCGTGACTTCCCTTCTCCAAGT (SEQ ID NO: 11)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNGGCGGGCATTCCTGAAGCTGA (SEQ ID NO: 58)	TGGAGTTCAGACGTGTGCTCTTCCGATCTACTTGGAGAAGGGAAGTCACGGA (SEQ ID NO: 70)
RNF2	GTCATCTTAGTCATTACCTG (SEQ ID NO: 213)	AGG	ACGTCTCATATGCCCCTTGGCAGTCATCTTAGTCATTACCTGAGGTGTTCGTTGTAACTCATATAAACTGAGTTCCCATGTTTTGCTTAATGGTTGAGTTCCGTTTGTCTGCACAGCCTGAGACATTGCTGGAAATAAAGAAGAGAGAAAAACAATTTTAGTATTTGGAAGGGAAGTGCTATGGTCTGAATGTATGTGTCCCACCAAAATTCCTACGT (SEQ ID NO: 12)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNACGTCTCATATGCCCCTTGG (SEQ ID NO: 59)	TGGAGTTCAGACGTGTGCTCTTCCGATCTACGTAGGAATTTTGGTGGGACA (SEQ ID NO: 71)
HEK4	GGCACTGCGGCTGGAGGTGG (SEQ ID NO: 214)	GGG	GAACCCAGGTAGCCAGAGACCCGCTGGTCTTCTTTCCCCTCCCCTGCCCTCCCCTCCCTTCAAGATGGCTGACAAAGGCCGGGCTGGGTGGAAGGAAGGGAGGAAGGGCGAGGCAGAGGGTCCAAAGCAGGATGACAGGCAGGGGCACCGCGGCGCCCCGGTGGCACTGCGGCTGGAGGTGGGGGTTAAAGCGGAGACTCTGGTGCTGTGTGACTACAGTGGGGGCCCTGCCCTCTCTGAGCCCCCGCCTCCAGGCCTGTGTGTGTGTCTCCGTTCGGGTTGAAAGGA (SEQ ID NO: 13)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNGAACCCAGGTAGCCAGAGAC (SEQ ID NO: 60)	TGGAGTTCAGACGTGTGCTCTTCCGATCTTCCTTTCAACCCGAACGGAG (SEQ ID NO: 72)
PDCD1	CACCTACCTAAGAACCATCC (SEQ ID NO: 215)	TGG	GACCTGCCAGGGACTGAGGGTGGAAGGTCCCTCCAGACCCCTGGCTCTGGGACACCTGACCGCCGACCCCACCTACCTAAGAACCATCCTGGCCGCCAGCCCAGTTGTAGCACCGCCCAGACGACT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNGACCTGCCAG	TGGAGTTCAGACGTGTGCTCTTCCGATCTGAGAAGGCGGCAC

[0958]

位点名称	原间隔区	PAM	扩增子	HTS-F	HTS-R
			GGCCAGGGCGCCTGTGGGATCTGCATGCCTGGAGCAGCCCCACCAGAGTGCCGCCTTCTC (SEQ ID NO: 14)	GGACTGAGGG (SEQ ID NO: 61)	TCTGGTG (SEQ ID NO: 73)
TRAC	TTCGTATCTGTAAAACCAAG (SEQ ID NO: 216)	AGG	TCTGCCTTGGGGAAAACCGTGGGTGTGTCCTGCAGGCCATGCAGGCCTGGGACATGCAAGCCCATAACCGCTGTGGCCTCTTGGTTTTACAGATACGAACCTAAACTTTCAAAACCTGTCAGTGATTGGGTTCCGAATCCTCCTCCTGAAAGTGGCCGGGTTTAATCTGCTCATGACGCTGCG (SEQ ID NO: 15)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNTCTGCCTTGGGGAAAACCGT (SEQ ID NO: 62)	TGGAGTTCAGACGTGTGCTCTTCCGATCTCGCAGCGTCATGAGCAGATT (SEQ ID NO: 79)
R5-7	CAATGTGTCAACTCTTGACA (SEQ ID NO: 217)	GGG	GCTGGTCATCCTCATCCTGATAAACTGCAAAAGGCTGAAGAGCATGACTGACATCTACCTGCTCAACCTGGCCATCTCTGACCTGTTTTTCCTTCTTACTGTCCCCTTCTGGGCTCACTATGCTGCCGCCCAGTGGGACTTTGGAAATACAATGTGTCAACTCTTGACAGGGCTCTATTTTATAGGCTTCTTCTCTGGAATCTTCTTCATCATCCTCCTGACAATCGATAGGTACCTGGCTGTCGTCCATGCTGTGTTTGCTTTAAAAGCCAGGACGGTCACCTTTGGGGTGGTGACAAGTGTGATCACTTGGGTGGTGGCTGTGTTTGCGTCTCTCCCAGGAA (SEQ ID NO: 16)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNGCTGGTCATCCTCATCCTG (SEQ ID NO: 63)	TGGAGTTCAGACGTGTGCTCTTCCGATCTTTCCTGGGAGAGACGCAAAC (SEQ ID NO: 80)
X4-2	CAGTTTCAGCACATCATGGT (SEQ ID NO: 218)	TGG	AAGCCAGGATGAGGATGACTGTGGTCTTGAGGGCCTTGCGCTTCTGGTGGCCCTTGGAGTGTGACAGCTTGGAGATGATAATGCAATAGCAGGACAGGATGACAATACCAGGCAGGATAAGGCCAACCATGATGTGCTGAAACTGGAACACAACCACCCACAAGTCATTGGGGTAGAAGCGGTCACAGATATATCTGTCATCTGCCTCACTGACGTTGGCAAAGATGAAGTCGGGAATAGTCAGCAGGAGGGC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNAAGCCAGGATGAGGATGACTG (SEQ ID NO: 64)	TGGAGTTCAGACGTGTGCTCTTCCGATCTCCTGCCCTCCTGCTGACTAT (SEQ ID NO: 81)

[0959]

位点名称	原间隔区	PAM	扩增子	HTS-F	HTS-R
			AGG (SEQ ID NO: 17)		
回路2	GGTCCAGGCGAACGCGAAGT (SEQ ID NO: 219)	CGG	GCGCTCGACACCGATATCATAAACATTCTGTTTGCCTAAATATTTACGCGTGGCAATTTTGATTGAGCCACCTTCTGCCCTGGTCCAGGCGAACGCGAAGTCGGATCCTCCGGAATTCGGGAGATTATCCACGCGCATCAGGTCTAATTCGCGTTCAAAAATCTCATCAATGGGGAGCATCTGACCATCTACAGTCATAAATTTGTGATCCTTTGTGGCACGGATTAAAGATCCGTCTTCCAGACAATATTCGAACACTTC (SEQ ID NO: 18)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNGAAGTGTTCGAATATTGTCTGGAAGA (SEQ ID NO: 65)	TGGAGTTCAGACGTGTGCTCTTCCGATCTGCGCTCGACACCGATATC (SEQ ID NO: 82)

[0960] 表2B:eNme2-C基因组基因座:

[0961]

位点名称	原间隔区	PAM	扩增子	HTS-F	HTS-R
TS90	GCAAGAGCACAAGAGGAAGAGAG (SEQ ID NO: 226)	AGACCC	GCTCAGAAAAAGGGCCCTGACAACTCTTTTCATCTTCTAGGTATGACAACGAATTTGGCTACAGCAACAGGGTGGTGGACCTCATGGCCCACATGGCCTCCAAGGAGTAAGACCCCTGGACCACCAGCCCCAGCAAGAGCACAAGAGGAAGAGAGAGACCCTCACTGCTGGGGAGTCCCTGCCACACTCAGTCCCCCACCACACTGAATCTC (SEQ ID NO: 83)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNGCTCAGAAAAAGGGCCCTGA (SEQ ID NO: 87)	TGGAGTTCAGACGTGTGCTCTTCCGATCTGAGATTCAGTGTGGTGGGGGG (SEQ ID NO: 91)
eNme2-C位点5	GGCTACAGCAACAGGGTGGTGG (SEQ ID NO: 227)	ACCTCA	GCTCAGAAAAAGGGCCCTGACAACTCTTTTCATCTTCTAGGTATGACAACGAATTTGGCTACAGCAACAGGGTGGTGGACCTCATGGCCCACATGGCCTCCAAGGAGTAAGACCCCTGGACCACCAGCCCCAGCAAGAGCACAAGAGGAAGAGAGAGACCCTCACTGCTGGGGAGTCCCTGCCACACTCAGTCCCCCACCACACTGAATCTC (SEQ ID NO: 83)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNGCTCAGAAAAAGGGCCCTGA (SEQ ID NO: 87)	TGGAGTTCAGACGTGTGCTCTTCCGATCTGAGATTCAGTGTGGTGGGGGG (SEQ ID NO: 91)

[0962]

eNme2-C位点66	GGTGTGCAGACGGCAGTCACTAG (SEQ ID NO: 228)	GGGGCG	CATTCCCTCTTTAGCCAGAGCCGGGGTGTGCAGACGGCAGTCACTAGGGGGCGCTCGGCCACCACAGGGAAGCTGGGTGAATGGAGCGAGCAGCGTCTTCGAGAGTGAGGACGTGTGTGTCTGTGGGGTGAGTGAGTGTGTGCGTGTGGGGTTGAGGGCGTTGGAGCGGGGAGAAGGCCAGGGGTCACTCCAGGATTCCAATAGATCTG (SEQ ID NO: 84)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNCATTCCCTCTTTAGCCAGAGCCGG (SEQ ID NO: 88)	TGGAGTTCAGACGTGTGCTCTTCCGATCTCAGATCTATTGGAATCCTGGAGTGACC (SEQ ID NO: 92)
eNme2-C位点50	GTCTCCGCTTTAACCCCCACCTC (SEQ ID NO: 229)	CAGCCG	GGCGAGGCAGAGGGTCCAAAGCAGGATGACAGGCAGGGGCACCGCGGCGCCCCGGTGGCACTGCGGCTGGAGGTGGGGGTTAAAGCGGAGACTCTGGTGCTGTGTGACTACAGTGGGGGCCCTGCCCTCTCTGAGCCCCCGCCTCCAGGCCTGTGTGTGTGTCTCCGTTCGGGTTGAAAGGAGCCCGGGAAAAAGGCCCCAGAAGGAG (SEQ ID NO: 85)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNGGCGAGGCAGAGGGTCCAAA (SEQ ID NO: 89)	TGGAGTTCAGACGTGTGCTCTTCCGATCTCCTTCTGGGGCCTTTTTCCC (SEQ ID NO: 93)
eNme2-C位点39	CCTCTTGCCTCCACTGGTTGTGC (SEQ ID NO: 230)	GCGGCT	CTACCTGCGCCACATCCATCGGCGCTTTGGTCGGCATGGCCCCATTCGCACGGCTCTGGAGCGGCGGCTGCACAACCAGTGGAGGCAAGAGGGCGGCTTTGGGCGGGGTCCAGTTCCGGGATTAGCGAACTTCCAGGCCCTCGGTCACTGTGACGTCCTGCTCTCTCTGCGCCTGCTGGAGAACCGGGCCCTCGGGGATGCAGCTCGTTACCACCTGGTGCAGCAACTCTTT (SEQ ID NO: 86)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNAAAGAGTTGCTGCACCAGGT (SEQ ID NO: 90)	TGGAGTTCAGACGTGTGCTCTTCCGATCTCTACCTGCGCCACATCCATC (SEQ ID NO: 94)
eNme2-C位点11	GGTCCAAAGCAGGATGACAGGCA (SEQ ID NO: 231)	GGGGCA	GAACCCAGGTAGCCAGAGACCCGCTGGTCTTCTTTCCCCTCCCCTGCCCTCCCCTCCCTTCAAGATGGCTGACAAAGGCCGGGCTGGGTGGAAGGAAGGGAGGAAGGGCGAGGCAGAGGGTCCAAAGCAGGATGACAGGCAGGGGCACCGCGGCGCCCCGGTGGCACTGCGGCTGGAGGTGGGGGTTAAAGCGGAGACTCTGGTGCTGTGTGACT	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNGAACCCAGGTAGCCAGAGAC (SEQ ID NO: 60)	TGGAGTTCAGACGTGTGCTCTTCCGATCTTCCTTTCAACCCGAACGGAG (SEQ ID NO: 76)

[0963]

			ACAGTGGGGGCCCTGCCC TCTCTGAGCCCCCGCCTC CAGGCCTGTGTGTGTGTC TCCGTTCGGGTTGAAAGG A (SEQ ID NO: 13)		

[0964] 表2C:SpCas9 Cas依赖性脱靶位点:

[0965]

位 点 名称	原间隔区	PAM	扩增子	HTS-F	HTS-R
HEK 3OT1	CACCCAG ACTGAGC ACGTGC (SEQ ID NO: 232)	TGG	TCCCCTGTTGACCTGGAG AAGCATGAACCAGTCAAA AAGTTTAAAGACAAGAGC ATTAACTGCACCAGTGGG CAGCTCAGCTCAGACACC AGTAGCGTGGGCACCCAG ACTGAGCACGTGCTGGAG CCCAAGAAATGCAGAGAC CTGTGCACCTCTGGTCAG GGCAAGTACAGTG (SEQ ID NO: 95)	ACACTCTT TCCCTACA CGACGCT CTTCCGAT CTNNNNT CCCCTGTT GACCTGG AGAA (SEQ ID NO: 126)	TGGAGT TCAGAC GTGTGC TCTTCCG ATCTCA CTGTAC TTGCCCT GACCA (SEQ ID NO: 181)
HEK 3OT2	GACACAG ACTGGGC ACGTGA (SEQ ID NO: 233)	GGG	TTGGTGTTGACAGGGAGC AACTTCACAGTCCCAGGC ATCAGGACACAGACTGGG CACGTGAGGGAAGCCCAA GGGAGAGGACTGGTGTAA TCGAGGCTGACTCCACTT TTAATGTTTGACTGATGA TAGGTTTCAAGTCTCACT AAGTCTCCTTCCCCTTCTG CCCACATCTCAG (SEQ ID NO: 96)	ACACTCTT TCCCTACA CGACGCT CTTCCGAT CTNNNNT TGGTGTTG ACAGGGA GCAA (SEQ ID NO: 127)	TGGAGT TCAGAC GTGTGC TCTTCCG ATCTCT GAGATG TGGGCA GAAGGG (SEQ ID NO: 182)
HEK 3OT3	AGCTCAG ACTGAGC AAGTGA (SEQ ID NO: 234)	GGG	TGAGAGGGAACAGAAGG GCTAAGACTAAAAGGAAC AGAGGAGTTCATAGTGAG CGGTAAAGAGCTCAGACT GAGCAAGTGAGGGGCTCA GCCTCCCATGGAGGACAG GGGGCTGGGGCCCCTGGC TGATGTCTGGACTGAAGC CCCCACGCCCAGAGGTTC TTGGGCCTTTGGAC (SEQ ID NO: 97)	ACACTCTT TCCCTACA CGACGCT CTTCCGAT CTNNNNT GAGAGGG AACAGAA GGGCT (SEQ ID NO: 128)	TGGAGT TCAGAC GTGTGC TCTTCCG ATCTGT CCAAAG GCCCAA GAACCT (SEQ ID NO: 183)
HEK 3OT4	AGACCAG ACTGAGC AAGAGA (SEQ ID NO: 235)	GGG	CCTAGCACTTTGGAAGGT CGAAGCGGCAGGATGGCT TCAACCCAGGAGTTCGAG ACCAGACTGAGCAAGAG AGGGAGAGTGTCTGTATT AACAACAAACAAACAAA CAAAAAACTAAACTAAAA GAAACTGTGGTGTATAAT	ACACTCTT TCCCTACA CGACGCT CTTCCGAT CTNNNNT CCTAGCA CTTTGGA AGGTCG	TGGAGT TCAGAC GTGTGC TCTTCCG ATCTGC TCATCTT AATCTG CTCAGC

[0966]

位点名称	原间隔区	PAM	扩增子	HTS-F	HTS-R
			ATAAAATTCTGGCTGAGC AGATTAAGATGAGC (SEQ ID NO: 98)	(SEQ ID NO: 129)	C (SEQ ID NO: 184)
HEK 3OT5	GAGCCAG AATGAGC ACGTGA (SEQ ID NO: 236)	GGG	AAAGGAGCAGCTCTTCCT GGTGGAAATTGCGAGCAG AGGCTGCGTGAGTTCCGT AACTCGCACACAGCCTCC ATTTGGAGCCAGAATGAG CACGTGAGGGACCCCGGG CAGAGGGGCCAGTGCTGA CATTATGCTCCATGCAAC CTCCCATCCTGTTGTGGG AGATGGTGCAGAC (SEQ ID NO: 99)	ACACTCTT TCCCTACA CGACGCT CTTCCGAT CTNNNNA AAGGAGC AGCTCTTC CTGG (SEQ ID NO: 130)	TGGAGT TCAGAC GTGTGC TCTTCCG ATCTGT CTGCAC CATCTC CCACAA (SEQ ID NO: 185)
HEK 4OT1	TGCACTG CGGCCGG AGGAGG (SEQ ID NO: 237)	TGG	GGCATGGCTTCTGAGACT CATAGCTGGGGCTGAAGA TCCCTAGGGGGGCTCTGC TGGGCTCACTGCTCTCCA GAGTGGTCCAGCCCGGCT GCAGGGTGCTGCTTCCAG CTTGGTGCACTGCGGCCG GAGGAGGTGGAGGATGG AAAGTAAGATTCAAAGAC AGGGAGTGCAAGGG (SEQ ID NO: 105)	ACACTCTT TCCCTACA CGACGCT CTTCCGAT CTNNNNG GCATGGC TTCTGAG ACTCA (SEQ ID NO: 131)	TGGAGT TCAGAC GTGTGC TCTTCCG ATCTGT CTCCCTT GCACTC CCTGTCT TT (SEQ ID NO: 186)
HEK 4OT2	GGCTCTG CGGCTGG AGGGGG (SEQ ID NO: 238)	TGG	TTTGGCAATGGAGGCATT GGGCAGGGGAAGCCTGTC TTCAGGGCACATGCACGT GCGCAGGGCTCTGCGGCT GGAGGGGGTGGGGTTGCT GTTAGTGACAGGGGCCCC AGCCAGGCAGGTTTCAGG ATTGGGGAGCACTTGCTT CGGCTCCCTTGCTCTCATG GGCAGCCTCTTC (SEQ ID NO: 106)	ACACTCTT TCCCTACA CGACGCT CTTCCGAT CTNNNNT TTGGCAA TGGAGGC ATTGG (SEQ ID NO: 132)	TGGAGT TCAGAC GTGTGC TCTTCCG ATCTGA AGAGGC TGCCCA TGAGAG (SEQ ID NO: 187)
HEK 4OT3	GGCACGA CGGCTGG AGGTGG (SEQ ID NO: 239)	GGG	GGTCTGAGGCTCGAATCC TGGCAGCAGGTCCTTCAT GGCAAGGCGGGAAAAGA GAAAAGCCAACGGGTTCT CATGCTGGGAAAAGATGC CGGGCACGACGGCTGGAG GTGGGGGGTTGGGAGTGG GTGGGATGCTTGCGTGCC CTGCATGAGGTGCAGGGA TATGGAGGCCACAG (SEQ ID NO: 107)	ACACTCTT TCCCTACA CGACGCT CTTCCGAT CTNNNNG GTCTGAG GCTCGAA TCCTG (SEQ ID NO: 133)	TGGAGT TCAGAC GTGTGC TCTTCCG ATCTCT GTGGCC TCCATA TCCCTG (SEQ ID NO: 188)
HEK 4OT4	GGCATCA CGGCTGG	AGG	TTCCACCAGAACTCAGCC CAGGCTGCTGTGGGATGG	ACACTCTT TCCCTACA	TGGAGT TCAGAC

[0967]

位点名称	原间隔区	PAM	扩增子	HTS-F	HTS-R
	AGGTGG (SEQ ID NO: 240)		AATCACCTGCACCCGGATGTTCTTTCTGGGCTGGTACATACAGGCAAGGCATCACGGCTGGAGGTGGAGGGGGCCTAACCCGGGGTTGCCCAGGAAGGGGTTTGCACATGGATTCGGTGTGTTGTGGAGGAACCGAGG (SEQ ID NO: 108)	CGACGCTCTTCCGATCTNNNNTTTCCACCAGAACTCAGCCC (SEQ ID NO: 134)	GTGTGCTCTTCCGATCTCCTCGGTTCCTCCACAACAC (SEQ ID NO: 189)
HEK 4OT5	GGCGCTGCGGCGGGAGGTGG (SEQ ID NO: 241)	AGG	CACGGGAAGGACAGGAGAAGGTGCTGGACCGCCTGGACTTTGTGCTGACCAGCCTTGTGGCGCTGCGGCGGGAGGTGGAGGAGCTGAGAAGCAGCCTGCGAGGGCTTGCGGGGGAGATTGTTGGGGAGGTCCGGTGAGTAATGCGGCTTCTTCTCCTGCTTTATCCCTCCCCTGC (SEQ ID NO: 109)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNCACGGGAAGGACAGGAGAAG (SEQ ID NO: 135)	TGGAGTTCAGACGTGTGCTCTTCCGATCTGCAGGGGAGGGATAAAGCAG (SEQ ID NO: 190)
EMX 1OT1	GAGTTAGAGCAGAAGAAGAA (SEQ ID NO: 242)	AGG	TGCCCAATCATTGATGCTTTTATACCATCTTGGGGTTACAGAAAGAATAGGGGCTTATGGCATGGCAAGACAGATTGTCAGAGTTAGAGCAGAAGAAGAAAGGCATGGAGTAAAGGCAATCTTGTGCAGATGTACAGGTAGCAGCCCTCAGAAAAAATAGGTGATAGTCTATGGTAAATGTTTCT (SEQ ID NO: 110)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNTGCCCAATCATTGATGCTTTT (SEQ ID NO: 136)	TGGAGTTCAGACGTGTGCTCTTCCGATCTAGAAACATTTACCATAGACTATCACCT (SEQ ID NO: 191)
EMX 1OT2	GAGTCTAAGCAGAAGAAGAA (SEQ ID NO: 243)	GAG	GTAGCCTCTTTCTCAATGTGCTTCAACCCATCACGGCCTTTGCAAATAGAGCCCTTTATTCATAGTAGACAAGAGTCTAAGCAGAAGAAGAAGAGAGCCACTACCCAACCATCTACTCTTCTAATGGTGTTTTCCTACAAAGGCCAAGTCATGAGACTGCATCCTTGTGAAAGC (SEQ ID NO: 111)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNAGTAGCCTCTTTCTCAATGTGC (SEQ ID NO: 137)	TGGAGTTCAGACGTGTGCTCTTCCGATCTGCTTTCACAAGGATGCAGTCT (SEQ ID NO: 192)
EMX 1OT3	GAGGCCGAGCAGAAGAAAGA (SEQ ID NO: 244)	CGG	GAGCTAGACTCCGAGGGGAGGCTGCGAGCCGCAAGCGCAGGAGCCGGGTGGGAGAGAGACCCCTTCTTCTGCAAATGAGGAGGCCGAGCAGAAGAAAGACGGCGACAGATGTTGGGGGGAGGG	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNGAGCTAGACTCCGAG	TGGAGTTCAGACGTGTGCTCTTCCGATCTTCCTCGTCCTGCTCTC

[0968]

位点名称	原间隔区	PAM	扩增子	HTS-F	HTS-R
			GACGGTTTGTGAGGGATAGGGAGAGAAAGTCTAAGTGAGAGCAGGACGAGGA (SEQ ID NO: 112)	GGGA (SEQ ID NO: 138)	ACTT (SEQ ID NO: 193)
EMX1OT4	GAGTCCTAGCAGGAGAAGAA (SEQ ID NO: 245)	GAG	AGAGGCTGAAGAGGAAGACCAGACTCAGTAAAGCCTGGAGGCTGCCAGGTAGGGCTGGGGCCAGCATGACCTGAGTCCTAGCAGGAGAAGAAGAGGCAGCCTAGAGTCTTCTGTGAAGTGCACATAGAAGAGAGACTGGGGCCAAGCCACAAAAGATAGAATGCACAGCTGGGCC (SEQ ID NO: 113)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNAGAGGCTGAAGAGGAAGACCA (SEQ ID NO: 139)	TGGAGTTCAGACGTGTGCTCTTCCGATCTGGCCCAGCTGTGCATTCTAT (SEQ ID NO: 194)
EMX1OT5	AAGTCTGAGCACAAGAAGAA (SEQ ID NO: 246)	TGG	GTAGTTCTGACATTCCTCCTGAGGGAAAATAAATAAATTAATTAAAAATATATATATATGTATAATGATAAACATGCTAACAAAGTCTGAGCACAAGAAGAATGGTGAGAAGGAATACATTTTATCTAATAAATATGTAAGCCATTAATAAAATGTAAACCATTAAAACAACAAATAAACCTTTCAGATATTGACCA (SEQ ID NO: 114)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNGTAGTTCTGACATTCCTCCTGAG (SEQ ID NO: 140)	TGGAGTTCAGACGTGTGCTCTTCCGATCTTGGTCAATATCTGAAAGGTTTATTTGT (SEQ ID NO: 195)
EMX1OT6	GAGTCCGGGAAGGAGAAGAA (SEQ ID NO: 247)	AGG	CCAAGAGGGCCAAGTCCTGGCTGTCTGCCTCTGACGACGAGCAAGGTGGAGGCCCTTGGTTAGCAGGATGGGTGGTGAGGAGTCCGGGAAGGAGAAGAAAGGCTCAGCGCGGCTTGCCTGAGCCTCCCTCCTCCCAGCTCCCGGCCCCTGCTGCCGGCGGCTGTCACTCCTCGCTG (SEQ ID NO: 115)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNCCAAGAGGGCCAAGTCCTG (SEQ ID NO: 141)	TGGAGTTCAGACGTGTGCTCTTCCGATCTCAGCGAGGAGTGACAGCC (SEQ ID NO: 196)
EMX1OT7	GAGCCGGAGCAGAAGAAGGA (SEQ ID NO: 248)	GGG	CACTCCACCTGATCTCGGGGCGCTGTGCGCTGAGGAAGGCGCGGGCGAGCCGGAGCAGAAGAAGGAGGGAGGGAGCCAGCCGCTGCAGCCACCACCGCCACCATGTCCTACCAAGGCAAGAAGAACATCCCGCGGATCACGGTGAGTCCGGGCGCCGCTGCTCCCTCCCTCCTCG (SEQ ID NO: 116)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNCACTCCACCTGATCTCGGGG (SEQ ID NO: 171)	TGGAGTTCAGACGTGTGCTCTTCCGATCTCGAGGAGGGAGGGAGCAG (SEQ ID NO: 197)

[0969]

位点名称	原间隔区	PAM	扩增子	HTS-F	HTS-R
EMX1OT8	AAGTCCGAGGAGAGGAAGAA (SEQ ID NO: 249)	AGG	ACCACAAATGCCCAAGAGACATCACCACTTGGAGAGTCAGAGGTCACAAAAGAGGGGCCCAACTCCTGTAGAAGTCCGAGGAGAGGAAGAAAGGGTTCTGGAGCTCTCAGGCGTCAGGGCCAGGCCTGCACCCTTCTGTGCCCCTCCATGAATGGCTGGCCGGCCCTTGACTGTGTC (SEQ ID NO: 117)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACCACAAATGCCCAAGAGAC (SEQ ID NO: 172)	TGGAGTTCAGACGTGTGCTCTTCCGATCTGACACAGTCAAGGGCCGG (SEQ ID NO: 198)
EMX1OT9	GAATCCAAGCAGGAGAAGAA (SEQ ID NO: 250)	GGA	CCCACCTTTGAGGAGGCAAAAGGGAATAAACTTGTGCTTATTTGTTGGAAGAGCAAATATGTTTTTTTGAAACCGAATTATGGATGGGGATGTGGGGGTGGGAACTAGGCAAGGGTCTCAGGGGAATCCAAGCAGGAGAAGAAGGAGGGAAAAACCACTCTCTTCTCAGATGGAA (SEQ ID NO: 118)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNCCCACCTTTGAGGAGGCAAA (SEQ ID NO: 173)	TGGAGTTCAGACGTGTGCTCTTCCGATCTTTCCATCTGAGAAGAGAGTGGT (SEQ ID NO: 199)
EMX1OT10	ACGTCTGAGCAGAAGAAGAA (SEQ ID NO: 251)	TGG	GTCATACCTTGGCCCTTCCTCTGTACTCTATACAGAGTCCAGCTCTGGCCTGGGAAAATACTTTCAGACAAAACGTCTGAGCAGAAGAAGAATGGACAGAACTCTGAGGACATTCTTGAGGCACTGGCAGAACCTCTGCAGGAAGACGAGAGCATTGCTGGTGTGGGCCTAGGGA (SEQ ID NO: 119)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNGTCATACCTTGGCCCTTCCT (SEQ ID NO: 174)	TGGAGTTCAGACGTGTGCTCTTCCGATCTTCCCTAGGCCCACACCAG (SEQ ID NO: 201)
BCL11AOT1	CCTATCACTGGCTCCAGGAA (SEQ ID NO: 252)	GGG	ACCTGTGGGCATCCTGAGTTGCTTCTGATGTCCCACCCATCACCTTGACCTGCTCAGAGCAGAGCATTGTTCTGAAATCTGAGGCATTGTCCTGCCCACTGGCCTATCACTGGCTCCAGGAAGGGCCTAGTGTCTCTGACCAGCTCTAGATCACCTCCTCCTCCTCCTGAGCCCTGTACGTTGCCAGGCTGATGAGAGGAGTGGGGCCGTGA (SEQ ID NO: 120)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACCTGTGGGCATCCTGAGTTGC (SEQ ID NO: 175)	TGGAGTTCAGACGTGTGCTCTTCCGATCTTCACGGCCCCACTCCTCTCA (SEQ ID NO: 202)
BCL11AOT2	CTTATCATAGGCCCCAGGAA	AGG	CTTGGCGCAGTTCCTGTGTATGGATATTCTTACAGAATCGCTACTCTCCCTCTCC	ACACTCTTTCCCTACACGACGCT	TGGAGTTCAGACGTGTGC

[0970]

位点名称	原间隔区	PAM	扩增子	HTS-F	HTS-R
	(SEQ ID NO: 253)		TTTGAGCTGGCCTAGCTTTGGCTTATCATAGGCCCCAGGAAAGGCCAGGGGACTGGGGTACCGGTTAGAGGGATATAAAAGTTCATTCTGCCTTGTACGTATGTTTAATTGATTAGAACACTTCATTTTCTCACAGCATGTA (SEQ ID NO: 121)	CTTCCGATCTNNNNCTTGGCGCAGTTCCTGTGTATG (SEQ ID NO: 176)	TCTTCCGATCTTACATGCTGTGAGAAAATGAAGTGT (SEQ ID NO: 203)
CCR5OT1	TAGGGTGTCAACTCTTGACA (SEQ ID NO: 254)	CAG	GACTTCTGGTTATCAGTAGGCTGTTAGTAGTTAAGTTTTGGGGGAGTCAAAAGTTATACCTGCATTTGACTGTGTCAAGAGTTGACACCCTAACCTCCACGTTGTTTAAGAGTTAACTGTAGTTTGAACTACACTAGTTTGTACTGGTTTGCAGTTATTTATTATACTATATTCATCATTACTTTTATGTTTCAATGAACCATTGTTGAGCCATGATCCATTTTG (SEQ ID NO: 122)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNGACTTCTGGTTATCAGTAGGCTGTT (SEQ ID NO: 177)	TGGAGTTCAGACGTGTGCTCTTCCGATCTCAAAATGGATCATGGCTCAA (SEQ ID NO: 207)
CCR5OT3	TTATCTGTCAACTCTTGACA (SEQ ID NO: 255)	AAT	TGAGGACTTGGTGAATCGTGGTATCATTTGTCAAGAGTTGACAGATAAGAGGAAGGGAGTAGGTTTGTCAGGGAAAGGGTTAATTTTTTTTTTTTGGAGACTTGTTGCCCAGGCTGGAGTACAATGGCATGATCTTGGCTCACCACAACTTCCGCCTCCTGGGTTCAAGTGATTTTGCTGCCTCAGCCTCCCGAGTAGCTGGGATTACAGGCATGTGCCACCACACCCAGCTAATT (SEQ ID NO: 123)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNTGAGGACTTGGTGAATCGTGG (SEQ ID NO: 178)	TGGAGTTCAGACGTGTGCTCTTCCGATCTAATTAGCTGGGTGTGGTGGC (SEQ ID NO: 220)
CCR5OT8	ATACCTGCCAACTCTTGACA (SEQ ID NO: 256)	TGG	CAATTACAGAAAGTGGATTGGTTGTGGAATGTACAGATTAACAATGGAGAGCACATAACCCTCAGTCTGGCACACCAGTGGAGGCTAAAGAAACTGAAAGGTGTTCTGCCCAGGGTGCAGGGACTCCATGTCAAGAGTTGGCAGGTATGGAAAGCAAGAGGGTTGGGACAGAAACCAGGGTAATTAATTAAGTGACTATTTTAGGACTAACTTCAAGGGTTACATGAGCTCC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNCAATTACAGAAAGTGGATTGGTTG (SEQ ID NO: 179)	TGGAGTTCAGACGTGTGCTCTTCCGATCTTTGTCTTGAGATTTTCCTAATGATTT (SEQ ID NO: 221)

[0971]

位点名称	原间隔区	PAM	扩增子	HTS-F	HTS-R
			CTCCCCTCCAAAAACAAA TCATTAGGAAAATCTCA (SEQ ID NO: 124)		
CXCR4OT8	TGTTCTCAGCACATCATGGT (SEQ ID NO: 257)	TGT	TTGCTGAAAGTGACGACA CAGCCAGTGAAGCTGGGA CTGGAATGTCAGTTTCCA TAACTCCAGCCCCCTTTTT TAACCGTTAGGAAACATC CTTTCACCTCAGGTCCCCT CAGTCTCTAGTCCACACC ATGATGTGCTGAGAACAG GAGGCATACTATCAGCAT TTTAAAAGGACCTCTTCC TTGTTTCTCCTGTTCTCTG GCCCCAGAGAGCTCTTCT GAGACCCTGGGGTCCAGA AGGGAGCCCTGGAGAACT TAGCTCCAAAGACTTTTC CTGAGCA (SEQ ID NO: 125)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNTTGCTGAAAGTGACGACACAG (SEQ ID NO: 180)	TGGAGTTCAGACGTGTGCTCTTCCGATCTTGCTCAGGAAAAGTCTTTGGA (SEQ ID NO: 222)

[0972] 表2D:SaCas9正交R环位点:

[0973]

位点名称	原间隔区	PAM	扩增子	HTS-F	HTS-R
SaR1	GTGGTAGACAGCATGTGTCCTA (SEQ ID NO: 258)	AAGGGT	TGGTGGAGTGCTCTGTGT TTGTCTTTATAAACCCAG ATGAGAGGATGAAGGCA ACAAGCTTCTGTACCAAC ATACATGCCCCTTTGCCTC AAGTCTGGTTATTTTAGG GGGATGCTAGGTTGCTTT GGGTCTACCTTACTGAGA AAATGGCCCCAGGTCATT GTCATGTCCAGTTGTGGT AGACAGCATGTGTCCTAA AGGGTATATTCACATGCA TGTGCAAAAATACAGGGG TCCTTCTAACCCTATCACA GAGAAGCAGGAGACTGC (SEQ ID NO: 223)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNTGCAGTCTCCTGCTTCTCTG (SEQ ID NO: 275)	TGGAGTTCAGACGTGTGCTCTTCCGATTGGTGGAGTGCTCTGTGTTTG (SEQ ID NO: 283)
SaR2	ATTTACAGCCTGGCCTTTGGGG (SEQ ID NO: 259)	TCGGGT	GCTACAGAAAGGTCAGCA GCTATATTTAACCTCAGA CCAGGGTGCGGTGGGAGA TCTGGTTTCCGGAAGACG GAATGGGGAGAAGGGCA GGTTCCCCGAGGCGCCCA GACACCCAATCCTCCCGG TGACATTTACAGCCTGGC CTTTGGGGTCGGGTCAAC	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNGGACATTTCCACCGCAAAATG (SEQ ID NO:	TGGAGTTCAGACGTGTGCTCTTCCGATGCTACAGAAAGGTCAGCAGC (SEQ ID

[0974]

			GCTAGGCTGGCAGGGGAA GGGCGGGGCCGTGAGGTG AGCCGGCGCTGCAGGAAG GGGCCACCACCAGAGGG GCCATTTTGCGGTGGAAA TGTCC (SEQ ID NO: 224)	280)	NO: 284)
SaR3	GTGTCA GGTAAT GTGCTA AACA (SEQ ID NO: 260)	GAGA GT	AAGTGTTCAGCTGCTTTTC TTTCATTTATTCCACATAT AATTACTATAATTGCTAA ACATTTATTTAGTGTCAG GTAATGTGCTAAACAGAG AGTTACTGCTCAGACATG TAATAATAATAAATAACA CATCAAATAACCATACCA TTTTAAGCTGTAGTATTAT GAAGGGAAATCTGGAGC AAAGAGAATAGACTGTAG GGAAACCAGTTAAGAAAT AGGACATGGAGGCTAGGT GCAG (SEQ ID NO: 225)	ACACTCTT TCCCTACA CGACGCT CTTCCGAT CTNNNNC TGCACCTA GCCTCCAT GTC (SEQ ID NO: 281)	TGGAGT TCAGAC GTGTGC TCTTCC GATCTG CTGTGG CATCCA GAGACA T (SEQ ID NO: 285)
SaR4	GGTGGA GGAGGG TGCATG GGGT (SEQ ID NO: 261)	CAGA AT	TTTGCTTATCCAGAAAAG GGAGTGATTGCTTCCAGG GGCCTCAGGGGAATAAAT CATAGAATCCTGGACAAG GTTTGAAGGACAGGTAGG ATTTGGGTGGGTGGAGGA GGGTGCATGGGGTCAGAA TTGTAACCGAAAACTCAT TCCAGGTGGATAGAGAAA ATTTCTAGTGTTGTTGTTT TTAAACTATTTGGGGGAC TGGCACAGACCCTTTTTG AATACCTGATGGGCTCAC ATTTCTGTCGAATCCCAG (SEQ ID NO: 273)	ACACTCTT TCCCTACA CGACGCT CTTCCGAT CTNNNNG GAGGTGG AGAGAGG ATGT (SEQ ID NO: 282)	TGGAGT TCAGAC GTGTGC TCTTCC GATCTT CCTGAG GTCTAG GAACCC G (SEQ ID NO: 286)
SaR5	TCTGCTT CTCCAG CCCTGG C (SEQ ID NO: 262)	CTGG GT	ATGTGGGCTGCCTAGAAA GGCATGGATGAGAGAAG CCTGGAGACAGGGATCCC AGGGAAACGCCCATGCAA TTAGTCTATTTCTGCTGCA AGTAAGCATGCATTTGTA GGCTTGATGCTTTTTTTCT GCTTCTCCAGCCCTGGCC TGGGTCAATCCTTGGGGC CCAGACTGAGCACGTGAT GGCAGAGGAAAGGAAGC CCTGCTTCCTCCAGAGGG CGTCGCAGGACAGCTTTT CCTAGACAGGGGCTAGTA TGTGCAGCTCCTGCACCG GGATACTGGTTGACAAGT TTGGCTGGG (SEQ ID NO:	ACACTCTT TCCCTACA CGACGCT CTTCCGAT CTNNNNA TGTGGGCT GCCTAGA AAGG (SEQ ID NO: 57)	TGGAGT TCAGAC GTGTGC TCTTCC GATCTC CCAGCC AAACTT GTCAAC C (SEQ ID NO: 69)

[0975]

			10)		
SaR6	GATGTTCCAATCAGTACGCA (SEQ ID NO: 263)	GAGAGT	CATTGCAGAGAGGCGTATCATTTCGCGGATGTTCCAATCAGTACGCAGAGAGTCGCCGTCTCCAAGGTGAAAGCGGAAGTAGGGCCTTCGCGCACCTCATGGAATCCCTTCTGCAGCACCTGGATCGCTTTTCCGAGCTTCTGGCGGTCTCAAGCACTACCTACGTCAGCACCTGGGACCCC (SEQ ID NO: 7)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNCATTGCAGAGAGGCGTATCA (SEQ ID NO: 32)	TGGAGTTCAGACGTGTGCTCTTCCGATCTGGGGTCCCAGGTGCTGAC (SEQ ID NO: 66)

[0976] 表2E:SaCas9基因组基因座:

位点名称	原间隔区	PAM	扩增子	HTS-F	HTS-R
SA1	GTGGTAGACAGCATGTGTCCTA (SEQ ID NO: 258)	AAGGGT	TGGTGGAGTGCTCTGTGTTTGTCTTTATAAACCCAGATGAGAGGATGAAGGCAACAAGCTTCTGTACCAACATACATGCCCCTTTGCCTCAAGTCTGGTTATTTTAGGGGGATGCTAGGTTGCTTTGGGTCTACCTTACTGAGAAAATGGCCCCAGGTCATTGTCATGTCCAGTTGTGGTAGACAGCATGTGTCCTAAAGGGTATATTCACATGCATGTGCAAAAATACAGGGGTCCTTCTAACCCTATCACAGAGAAGCAGGAGACTGC (SEQ ID NO: 223)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNTGCAGTCTCCTGCTTCTCTG (SEQ ID NO: 275)	TGGAGTTCAGACGTGTGCTCTTCCGATTGGTGGAGTGCTCTGTGTTTG (SEQ ID NO: 283)
SA2	ATTTACAGCCTGGCCTTTGGGG (SEQ ID NO: 259)	TCGGGT	GCTACAGAAAGGTCAGCAGCTATATTTAACCTCAGACCAGGGTGCGGTGGGAGATCTGGTTTCCGGAAGACGGAATGGGGAGAAGGGCAGGTTCCCCGAGGCGCCCAGACACCCAATCCTCCCGGTGACATTTACAGCCTGGCCTTTGGGGTCGGGTCAACGCTAGGCTGGCAGGGGAAGGGCGGGGCCGTGAGGTGAGCCGGCGCTGCAGGAAGGGGCCACCACCAGAGGGGCCATTTTGCGGTGGAAATGTCC (SEQ ID NO: 224)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNGGACATTTCCACCGCAAAATG (SEQ ID NO: 280)	TGGAGTTCAGACGTGTGCTCTTCCGATGCTACAGAAAGGTCAGCAGC (SEQ ID NO: 284)
SA3	GTGTCAGGTAATGTGCTAAAC	GAGAGT	AAGTGTTCAGCTGCTTTTCTTTCATTTATTCCACATATAATTACTATAATTGCTAAACATTTATTTAGTGTCAGGTAATG	ACACTCTTTCCCTACACGACGCTCTTCC	TGGAGTTCAGACGTGTGCTCTTCC

[0978]

	A (SEQ ID NO: 260)		TGCTAAACAGAGAGTTACT GCTCAGACATGTAATAATAA TAAATAACACATCAAATAA CCATACCATTTTAAGCTGTA GTATTATGAAGGGAAATCT GGAGCAAAGAGAATAGACT GTAGGGAAACCAGTTAAGA AATAGGACATGGAGGCTAG GTGCAG (SEQ ID NO: 225)	GATCTNN NNCTGCA CCTAGCC TCCATGT C (SEQ ID NO: 281)	GATCTG CTGTGG CATCCA GAGACA T (SEQ ID NO: 66)
SA4	TCTGCTT CTCCAGC CCTGGC (SEQ ID NO: 262)	CTG GGT	ATGTGGGCTGCCTAGAAAG GCATGGATGAGAGAAGCCT GGAGACAGGGATCCCAGG GAAACGCCCATGCAATTAG TCTATTTCTGCTGCAAGTAA GCATGCATTTGTAGGCTTGA TGCTTTTTTTCTGCTTCTCC AGCCCTGGCCTGGGTCAAT CCTTGGGGCCCAGACTGAG CACGTGATGGCAGAGGAA AGGAAGCCCTGCTTCCTCC AGAGGGCGTCGCAGGACA GCTTTTCCTAGACAGGGGC TAGTATGTGCAGCTCCTGC ACCGGGATACTGGTTGACA AGTTTGGCTGGG (SEQ ID NO: 10)	ACACTCT TTCCCTA CACGACG CTCTTCC GATCTNN NNATGTG GGCTGCC TAGAAAG G (SEQ ID NO: 57)	TGGAGT TCAGAC GTGTGC TCTTCC GATCTC CCAGCC AAACTT GTCAAC C (SEQ ID NO: 69)
SA5	GATGTTC CAATCAG TACGCA (SEQ ID NO: 263)	GAG AGT	CATTGCAGAGAGGCGTATC ATTTCGCGGATGTTCCAATC AGTACGCAGAGAGTCGCCG TCTCCAAGGTGAAAGCGGA AGTAGGGCCTTCGCGCACC TCATGGAATCCCTTCTGCA GCACCTGGATCGCTTTTCC GAGCTTCTGGCGGTCTCAA GCACTACCTACGTCAGCAC CTGGGACCCC (SEQ ID NO: 7)	ACACTCT TTCCCTA CACGACG CTCTTCC GATCTNN NNCATTG CAGAGAG GCGTATC A (SEQ ID NO: 32)	TGGAGT TCAGAC GTGTGC TCTTCC GATCTG GGGTCC CAGGTG CTGAC (SEQ ID NO: 72)
SA6	GCAAGGC CCGGCGC ACGGTGG (SEQ ID NO: 291)	CGG GGT	GATCGCTTTTCCGAGCTTCT GGCGGTCTCAAGCACTACC TACGTCAGCACCTGGGACC CCGCCACCGTGCGCCGGGC CTTGCAGTGGGCGCGCTAC CTGCGCCACATCCATCGGC GCTTTGGTCGGCATGGCCC CATTCGCACGGCTCTGGAG CGGCGGCTGCACAACCAGT GGAGGCAAGAGGGCGGCT TTGGGCGGGGTCCAGTTCC GGGATTAGCGAACTTCCAG GCCCTCGGTCACTGTGACG TCCTGCTCTCTCTGCGCCTG CTGGAGAACCGGGCCCTCG	ACACTCT TTCCCTA CACGACG CTCTTCC GATCTNN NNGATCG CTTTTCC GAGCTTC TGGC (SEQ ID NO: 304)	TGGAGT TCAGAC GTGTGC TCTTCC GATCTG CACCAG GTGGTA ACGAGC TGCATC (SEQ ID NO: 287)

[0979]

			GGGATGCAGCTCGTTACCA CCTGGTGC (SEQ ID NO: 299)		
SA7	ACGTGCT CAGTCTG GGCCCC (SEQ ID NO: 292)	AAG GAT	ATGTGGGCTGCCTAGAAAG GCATGGATGAGAGAAGCCT GGAGACAGGGATCCCAGG GAAACGCCCATGCAATTAG TCTATTTCTGCTGCAAGTAA GCATGCATTTGTAGGCTTGA TGCTTTTTTTCTGCTTCTCC AGCCCTGGCCTGGGTCAAT CCTTGGGGCCCAGACTGAG CACGTGATGGCAGAGGAA AGGAAGCCCTGCTTCCTCC AGAGGGCGTCGCAGGACA GCTTTTCCTAGACAGGGGC TAGTATGTGCAGCTCCTGC ACCGGGATACTGGTTGACA AGTTTGGCTGGG (SEQ ID NO: 10)	ACACTCT TTCCCTA CACGACG CTCTTCC GATCTNN NNATGTG GGCTGCC TAGAAAG G (SEQ ID NO: 57)	TGGAGT TCAGAC GTGTGC TCTTCC GATCTC CCAGCC AAACTT GTCAAC C (SEQ ID NO: 285)
SA8	GCGAGGC AGAGGGT CCAAAG (SEQ ID NO: 293)	CAG GAT	GAACCCAGGTAGCCAGAG ACCCGCTGGTCTTCTTTCCC CTCCCCTGCCCTCCCCTCCC TTCAAGATGGCTGACAAAG GCCGGGCTGGGTGGAAGG AAGGGAGGAAGGGCGAGG CAGAGGGTCCAAAGCAGG ATGACAGGCAGGGGCACC GCGGCGCCCCGGTGGCACT GCGGCTGGAGGTGGGGGTT AAAGCGGAGACTCTGGTGC TGTGTGACTACAGTGGGGG CCCTGCCCTCTCTGAGCCC CCGCCTCCAGGCCTGTGTG TGTGTCTCCGTTCGGGTTG AAAGGA (SEQ ID NO: 13)	ACACTCT TTCCCTA CACGACG CTCTTCC GATCTNN NNGAACC CAGGTAG CCAGAGA C (SEQ ID NO: 60)	TGGAGT TCAGAC GTGTGC TCTTCC GATCTT CCTTTC AACCCG AACGGA G (SEQ ID NO: 72)
SA9	CCACTGT AGTCACA CAGCAC (SEQ ID NO: 294)	CAG AGT	GAACCCAGGTAGCCAGAG ACCCGCTGGTCTTCTTTCCC CTCCCCTGCCCTCCCCTCCC TTCAAGATGGCTGACAAAG GCCGGGCTGGGTGGAAGG AAGGGAGGAAGGGCGAGG CAGAGGGTCCAAAGCAGG ATGACAGGCAGGGGCACC GCGGCGCCCCGGTGGCACT GCGGCTGGAGGTGGGGGTT AAAGCGGAGACTCTGGTGC TGTGTGACTACAGTGGGGG CCCTGCCCTCTCTGAGCCC CCGCCTCCAGGCCTGTGTG TGTGTCTCCGTTCGGGTTG AAAGGA (SEQ ID NO: 13)	ACACTCT TTCCCTA CACGACG CTCTTCC GATCTNN NNGAACC CAGGTAG CCAGAGA C (SEQ ID NO: 60)	TGGAGT TCAGAC GTGTGC TCTTCC GATCTT CCTTTC AACCCG AACGGA G (SEQ ID NO: 72)

[0980] 表3.用于RNA脱靶分析的cDNA扩增子序列和引物。

[0981]

名称	扩增子	HTS-F	HTS-R
RSL1D1	TTGGCTTTCCAAATCAGTGGGTCTGACTTGAGGTCTGTGATGTGACCCTTTTCCTCACCTGCTCAACCATTATTCACATGGACTCCATCATATTCATTTGTAGTCATTCCCAGAGTGGCCCAGTGAGGGTCTCGCTGTATGAGAGTCGGCTACGGAATTTAGGAGAAACAGAAGTTTCTTGGCTTTCATGCTGAGCTTGTTGGTCTAAGCTTATGAG (SEQ ID NO: 288)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNTGGCTTTCCAAATCAGTGGGTC (SEQ ID NO: 295)	TGGAGTTCAGACGTGTGCTCTTCCGATCTCTCATAAGCTTAGACCAACAAGC (SEQ ID NO: 298)
CTNNB1	TTTGATGGAGTTGGACATGGCCATGGAACCAGACAGAAAAGCGGCTGTTAGTCACTGGCAGCAACAGTCTTACCTGGACTCTGGAATCCATTCTGGTGCCACTACCACAGCTCCTTCTCTGAGTGGTAAAGGCAATCCTGAGGAAGAGGATGTGGATACCTCCCAAGTCCTGTATGAGTGGGAACAGGGATTTTCTCAGTCCTTCACTCAAGAACAAGTAGCTGG (SEQ ID NO: 289)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNATTTGATGGAGTTGGACATGGCC (SEQ ID NO: 296)	TGGAGTTCAGACGTGTGCTCTCCAGCTACTTGTTCTTGAGTGAAGG (SEQ ID NO: 300)
IP90	CTGGTTGACCAATCTGTGGTGAATAGTGGAAATCTGCTCAATGACATGACTCCTCCTGTAAATCCTTCACGTGAAATTGAGGACCCAGAAGACCGGAAGCCCGAGGATTGGGATGAAAGACCAAAAATCCCAGATCCAGAAGCTGTCAAGCCAGATGACTGGGATGAAGATGCCCCTGCTAAGATTCCAGATGAAGAGGCCACAAAACCCGAAGGCTGGTTAGATGATGAGCCTGAGTACGTAC (SEQ ID NO: 290)	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNCTGGTTGACCAATCTGTGGTG (SEQ ID NO: 297)	TGGAGTTCAGACGTGTGCTCTCTGCGTCTGGATCAGGTACG (SEQ ID NO: 301)

[0982] 表4.用于生成针对IVT的碱基编辑器扩增子的引物。

[0983]

名称	序列(5'-3')
IVT-F	TCGAGCTCGGTACCTAATACGACTCACTATAAGG (SEQ ID NO: 302)
IVT-R	TTTTT

30％窗口内的碱基((每个位置的平均CBE编辑)与(每个位置的平均ABE编辑)的比率)的几何平均值。P(ABE|CBE)是在观察到C•G至G•C编辑的情况下在读段中观察到A•T至G•C编辑的平均概率。

[0989]

变体	选择性	P (ABE|CBE) rep1	P (ABE|CBE) rep2
BE4max	>100	0.01	0.00
TadCBEa	18.19	0.10	0.11
TadCBEb	10.98	0.17	0.18
TadCBEc	13.84	0.12	0.14
TadCBEd	26.87	0.09	0.09
TadCBEe	15.72	0.12	0.13
TadCBEa V106W	31.62	0.06	0.05
TadCBEb V106W	19.91	0.08	0.09
TadCBEc V106W	28.80	0.06	0.06
TadCBEd V106W	47.80	0.04	0.05
TadDE	1.00	0.62	0.63

[0990] 表7.启动子和RBS序列

[0991]

名称	启动子/RBS	描述	序列
SP	PgIII	针对噬菌体上的 TadA*-NpuN 的启动子	AATTCACCTCGAAAGCAAGTTGATAAACTGATACAATTAAAGGCTCCT (SEQ ID NO: 305)
SP	RBS	针对噬菌体上的 TadA*-NpuN 表达的 RBS	AAGGAGGAAAA (SEQ ID NO: 306)
P1	ProC	针对 NpuC-dCas9-ugi 的启动子	CACAGCTAACACCACGTCGTCCCTATCTGCTGCCCTAGGTCTATGAGTGGTTGCTGGATAACTTTACGGGCATGCATAAGGCTCGTATGATATATTCAGGGAGACCACAACGGTTTCCCTCTACAAATAATTTTGTTTAACTTTTACTAGAGTGGGACCCTACCTGCAGGTGCAGT (SEQ ID NO: 307)

[0992]

P1	SD8	针对 NpuC-dCas9-ugi 的表达的 RBS	AAGGAGGAAAAAAAA (SEQ ID NO: 308)
P2	pLac	针对 sgRNA 的启动子	GGCTTTACACTTTATGCTTCCGGCTC GTATGTTGTGTGG (SEQ ID NO: 309)
P2	proT7	T7 启动子控制 gIII 转录	TAATACGACTCACTATAGGGAGA (SEQ ID NO: 310)
P2	sd8	针对 gIII 的表达的 RBS	AAGGAAAAAAAAA (SEQ ID NO: 311)
P3	ProA	针对 T7RNAP 的启动子。因严格性而异(A 最弱，最严格)	CACAGCTAACACCACGTCGTCCCTA TCTGCTGCCCTAGGTCTATGAGTGGT TGCTGGATAACTTTACGGGCATGCA TAAGGCTCGTAGGCTATATTCAGGG AGACCACAACGGTTTCCCTCTACAA ATAATTTTGTTTAACTTTTACTAGAG TGGGACCCTACCTGCAGGTGCAGT (SEQ ID NO: 312)
P3	ProB	针对 T7RNAP 的启动子。因严格性而异。	CACAGCTAACACCACGTCGTCCCTA TCTGCTGCCCTAGGTCTATGAGTGGT TGCTGGATAACTTTACGGGCATGCA TAAGGCTCGTAATATATATTCAGGG AGACCACAACGGTTTCCCTCTACAA ATAATTTTGTTTAACTTTTACTAGAG TGGGACCCTACCTGCAGGTGCAGT (SEQ ID NO: 313)
P3	ProC	针对 T7RNAP 的启动子。因严格性而异。	CACAGCTAACACCACGTCGTCCCTA TCTGCTGCCCTAGGTCTATGAGTGGT TGCTGGATAACTTTACGGGCATGCA TAAGGCTCGTATGATATATTCAGGG AGACCACAACGGTTTCCCTCTACAA ATAATTTTGTTTAACTTTTACTAGAG TGGGACCCTACCTGCAGGTGCAGT (SEQ ID NO: 307)
P3	ProD	针对 T7RNAP 的启动子。因严格性而异(D 最强，最不严格)	CACAGCTAACACCACGTCGTCCCTA TCTGCTGCCCTAGGTCTATGAGTGGT TGCTGGATAACTTTACGGGCATGCA TAAGGCTCGTATAATATATTCAGGG AGACCACAACGGTTTCCCTCTACAA ATAATTTTGTTTAACTTTTACTAGAG TGGGACCCTACCTGCAGGTGCAGT (SEQ ID NO: 314)
P3	RBSR3	针对 T7 RNAP 的表达的 RBS	ACTACATCATCAGGC (SEQ ID NO: 316)

[0993] 表8.用于碱基编辑器的mESC 12kChar文库分析的PCR1和PCR2的引物。用于文库分析的样品(11个样品,每个样品有两个生物学重复,共22个样品)被扩增并条形码化以进行分析。进行了4次NextSeq运行,每次最多运行7个样品。每次运行中分配给样品的引物是唯一的。运行1含有7个样品,使用FWD PCR1A和REV PCR2-1至PCR2-7进行条形码化。运行2含有7个样品,使用FWD PCR1B和REV PCR2-1至PCR2-7进行条形码化。运行3含有7个样品,使用FWD PCR1C和REV PCR2-1至PCR2-7进行条形码化。运行4含有1个样品,使用FWD PCR1D和REV PCR2-1进行条形码化。

[0994]

名称	目的	序列
Fwd PCR1 A	文库 Fwd PCR1	AATGATACGGCGACCACCGAGATCTACACATTACT CGACACTCTTTCCCTACACGAC (SEQ ID NO: 324)
Fwd PCR1 B	文库 Fwd PCR1	AATGATACGGCGACCACCGAGATCTACACTCCGGA GAACACTCTTTCCCTACACGAC (SEQ ID NO: 327)
Fwd PCR1 C	文库 Fwd PCR1	AATGATACGGCGACCACCGAGATCTACACCGCTCA TTACACTCTTTCCCTACACGAC (SEQ ID NO: 328)
Fwd PCR1 D	文库 Fwd PCR1	AATGATACGGCGACCACCGAGATCTACACGAGATT CCACACTCTTTCCCTACACGAC (SEQ ID NO: 329)
Rev PCR1 Lib	文库 Rev PCR1	GTGACTGGAGTTCAGACGTGTGCTCTTC CGATCTGTGGAAAGGACGAAACACCG (SEQ ID NO: 330)
Fwd PCR2 Lib	文库 Fwd PCR2	AATGATACGGCGACCACCGAGATCTACAC (SEQ ID NO: 331)
Rev PCR2-1	文库 Rev PCR2	CAAGCAGAAGACGGCATACGAGATATCGTGATGTG ACTGGAGTTCAGACGTGTGCT (SEQ ID NO: 332)
Rev PCR2-2	文库 Rev PCR2	CAAGCAGAAGACGGCATACGAGATATACATCGGTG ACTGGAGTTCAGACGTGTGCT (SEQ ID NO: 333)
Rev PCR2-3	文库 Rev PCR2	CAAGCAGAAGACGGCATACGAGATATGCCTAAGTG ACTGGAGTTCAGACGTGTGCT (SEQ ID NO: 334)
Rev PCR2-4	文库 Rev PCR2	CAAGCAGAAGACGGCATACGAGATATTGGTCAGTG ACTGGAGTTCAGACGTGTGCT (SEQ ID NO: 335)
Rev PCR2-5	文库 Rev PCR2	CAAGCAGAAGACGGCATACGAGATATCACTGTGTG ACTGGAGTTCAGACGTGTGCT (SEQ ID NO: 336)
Rev PCR2-6	文库 Rev PCR2	CAAGCAGAAGACGGCATACGAGATATATTGGCGTG ACTGGAGTTCAGACGTGTGCT (SEQ ID NO: 337)
Rev PCR2-7	文库 Rev PCR2	CAAGCAGAAGACGGCATACGAGATATGATCTGGTG ACTGGAGTTCAGACGTGTGCT (SEQ ID NO: 338)

[0995] 表9.本文所述的碱基编辑器的Cas9组分的序列

[0996]

名称	序列
SpCas9 D10A	DKKYSIGLAIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIG ALLFDSGETAEATRLKRTARRRYTRRKNRICYLQEIFSNEMAKVDDSF FHRLEESFLVEEDKKHERHPIFGNIVDEVAYHEKYPTIYHLRKKLVDS TDKADLRLIYLALAHMIKFRGHFLIEGDLNPDNSDVDKLFIQLVQTYN QLFEENPINASGVDAKAILSARLSKSRRLENLIAQLPGEKKNGLFGNLI ALSLGLTPNFKSNFDLAEDAKLQLSKDTYDDDLDNLLAQIGDQYADL FLAAKNLSDAILLSDILRVNTEITKAPLSASMIKRYDEHHQDLTLLKAL VRQQLPEKYKEIFFDQSKNGYAGYIDGGASQEEFYKFIKPILEKMDGT EELLVKLNREDLLRKQRTFDNGSIPHQIHLGELHAILRRQEDFYPFLKD NREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSEETITPWNFEEVVD KGASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYV TEGMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFKKIECFDS VEISGVEDRFNASLGTYHDLLKIIKDKDFLDNEENEDILEDIVLTLTLFE DREMIEERLKTYAHLFDDKVMKQLKRRRYTGWGRLSRKLINGIRDK QSGKTILDFLKSDGFANRNFMQLIHDDSLTFKEDIQKAQVSGQGDSLH EHIANLAGSPAIKKGILQTVKVVDELVKVMGRHKPENIVIEMARENQT TQKGQKNSRERMKRIEEGIKELGSQILKEHPVENTQLQNEKLYLYYLQ NGRDMYVDQELDINRLSDYDVDHIVPQSFLKDDSIDNKVLTRSDKNR GKSDNVPSEEVVKKMKNYWRQLLNAKLITQRKFDNLTKAERGGLSE LDKAGFIKRQLVETRQITKHVAQILDSRMNTKYDENDKLIREVKVITL KSKLVSDFRKDFQFYKVREINNYHHAHDAYLNAVVGTALIKKYPKLE SEFVYGDYKVYDVRKMIAKSEQEIGKATAKYFFYSNIMNFFKTEITLA NGEIRKRPLIETNGETGEIVWDKGRDFATVRKVLSMPQVNIVKKTEV QTGGFSKESILPKRNSDKLIARKKDWDPKKYGGFDSPTVAYSVLVVA KVEKGKSKKLKSVKELLGITIMERSSFEKNPIDFLEAKGYKEVKKDLII KLPKYSLFELENGRKRMLASAGELQKGNELALPSKYVNFLYLASHYE KLKGSPEDNEQKQLFVEQHKHYLDEIIEQISEFSKRVILADANLDKVLS AYNKHRDKPIREQAENIIHLFTLTNLGAPAAFKYFDTTIDRKRYTSTKE VLDATLIHQSITGLYETRIDLSQLGGD (SEQ ID NO: 340)
eNme2-C Cas9 *(相对于报告的eNme2-C 序列的S6P 突变)*	AAFK**P**NPINYILGLAIGIASVGWAMVEIDEEENPIRLIDLGVRVFERAE VPKTGDSLAMARRLARSVRRLTRRRAHRLLRARRLLKREGVLQAAD FDENGLIKSLPNTPWQLRAAALDRKLTPLEWSAVLLHLIKHRGYLSQ RKNEGETADKELGALLKGVANNAHALQTGDFRTPAELALNKFEKES GHIRNQRGDYSHTFSRKDLQAELILLFEKQKEFGNPHVSGGLKEGIET LLMTQRPALSGDAVQKMLGHCTFEPAEPKAAKNTYTAERFIWLTKL NNLRILEQGSERPLTDTERATLMDEPYRKSKLTYAQARKLLGLEDTAF FKGLRYGKDNAEASTLMEMKAYHAISRALEKEGLKDKKSPLNLSSEL QDEIGTAFSLFKTDEDITGRLKDRVQPEILEALLKHISFDKFVQISLKAL RRIVPLMEQGKRYDEACAEIYGDHYGKKNTEEKIYLPPIPADEIRNPV VLRALSQARKVINGVVRRYGSPARIHIETAREVGKSFKDRKEIEKRQE ENRKDREKAAAKFREYFPNFVGEPKSKDILKLRLYEQQHGKCLYSGK EINLVRLNEKGYVEIDAALPFSRTWDDSFNNKVLVLGSENQNKGNQT PYEYFNGKDNSREWQEFKARVETSRFPRSKKQRILLQKFDEDGFKEC NLNDTRYVNRFLCQFVADHILLTGKGKRRVFASNGQITNLLRGFWGL RKVRAENDRHHALDAVVVACSTVAMQQKITRFVRYKEMNAFDGKTI DKETGKVLHQKTHFPQPWEFFAQEVMIRVFGKPDGKPEFEEADTPEK LRTLLAEKLSSRPEAVHEYVTPLFVSRAPNRKMSGAHKDTLRSAKRF

[0997]

	VKHNEKISVKRVWLTEIKLADLENMVNYKNGREIELYEALKARLEAY GGNAKQAFDPKDNPFYKKGGQLVKAVRVEKTQESGVLLNKKNAYTI ADNGDMVRVDVFCKVDKKGKNQYFIVPIYAWQVAENILPDIDCKGY RIDDSYTFCFSLHKYDLIAFQKDEKSKVEFAYYINCDSSNGRFYLAWH DKGSKEQQFRISTQNLVLIQKYQVNELGKEIRPCRLKKRPPV (SEQ ID NO: 341)
SaCas9 D10A	GKRNYILGLAIGITSVGYGIIDYETRDVIDAGVRLFKEANVENNEGRRS KRGARRLKRRRRHRIQRVKKLLFDYNLLTDHSELSGINPYEARVKGL SQKLSEEEFSAALLHLAKRRGVHNVNEVEEDTGNELSTKEQISRNSKA LEEKYVAELQLERLKKDGEVRGSINRFKTSDYVKEAKQLLKVQKAY HQLDQSFIDTYIDLLETRRTYYEGPGEGSPFGWKDIKEWYEMLMGHC TYFPEELRSVKYAYNADLYNALNDLNNLVITRDENEKLEYYEKFQIIE NVFKQKKKPTLKQIAKEILVNEEDIKGYRVTSTGKPEFTNLKVYHDIK DITARKEIIENAELLDQIAKILTIYQSSEDIQEELTNLNSELTQEEIEQISN LKGYTGTHNLSLKAINLILDELWHTNDNQIAIFNRLKLVPKKVDLSQQ KEIPTTLVDDFILSPVVKRSFIQSIKVINAIIKKYGLPNDIIIELAREKNSK DAQKMINEMQKRNRQTNERIEEIIRTTGKENAKYLIEKIKLHDMQEGK CLYSLEAIPLEDLLNNPFNYEVDHIIPRSVSFDNSFNNKVLVKQEENSK KGNRTPFQYLSSSDSKISYETFKKHILNLAKGKGRISKTKKEYLLEERD INRFSVQKDFINRNLVDTRYATRGLMNLLRSYFRVNNLDVKVKSING GFTSFLRRKWKFKKERNKGYKHHAEDALIIANADFIFKEWKKLDKAK KVMENQMFEEKQAESMPEIETEQEYKEIFITPHQIKHIKDFKDYKYSH RVDKKPNRELINDTLYSTRKDDKGNTLIVNNLNGLYDKDNDKLKKLI NKSPEKLLMYHHDPQTYQKLKLIMEQYGDEKNPLYKYYEETGNYLT KYSKKDNGPVIKKIKYYGNKLNAHLDITDDYPNSRNKVVKLSLKPYR FDVYLDNGVYKFVTVKNLDVIKKENYYEVNSKCYEEAKKLKKISNQ AEFIASFYNNDLIKINGELYRVIGVNNDLLNRIEVNMIDITYREYLENM NDKRPPRIIKTIASKTQSIKKYSTDILGNLYEVKSKKHPQIIKKG (SEQ ID NO: 342)

[0998] 等效物和范围

[0999] 本领域技术人员将认识到或能够仅使用常规实验确定本文所述的本发明的具体实施方案的许多等效物。本发明的范围不旨在限于上述描述,而是如所附权利要求中所阐述的。

[1000] 在权利要求书中,例如“一”、“一个”和“该”可以表示一个或多于一个,除非有相反的指示或从上下文中显而易见。对于在组的一个或多个成员之间包括“或”的权利要求或描述,如果组成员中的一个、多于一个或所有成员存在于给定产品或方法中、用于给定产品或方法中或以其他方式与给定产品或方法相关,则除非有相反的指示或从上下文中显而易见,否则认为该权利要求或描述是满足的。本发明包括其中恰好一个组成员存在于给定产品或方法中、用于给定产品或方法中或以其他方式与给定产品或方法相关的实施方案。本发明还包括其中多于一个或所有组成员存在于给定产品或方法中、用于给定产品或方法中或以其他方式与给定产品或方法相关的实施方案。

[1001] 此外,应当理解,本发明涵盖其中来自一个或多个权利要求或来自说明书的相关部分的一个或多个限制、元素、条款、描述性术语等被引入到另一个权利要求中的所有变化、组合和排列。例如,从属于另一权利要求的任何权利要求可以被修改以包括在从属于同一基础权利要求的任何其他权利要求中发现的一个或多个限制。此外,在权利要求叙述组合物的情况下,应理解,包括使用组合物用于本文公开的任何目的的方法,并且包括根据本文公开的任何制备方法或本领域已知的其他方法制备组合物的方法,除非另有指示或除非本领域普通技术人员显然认为会出现矛盾或不一致。

[1002] 在元素以列表形式呈现的情况下,例如,以马库什组格式,应当理解,还公开了元

素的每个子组，并且可以从该组中移除任何元素。还应注意，术语“包含”旨在是开放的，并且允许包括另外的元素或步骤。应当理解，通常，在本发明或本发明的方面被称为包含特定元素、特征、步骤等的情况下，本发明或本发明的方面的某些实施方案由或基本上由这些元素、特征、步骤等组成。为了简单起见，本文中没有具体阐述那些实施方案。因此，对于包含一个或多个元素、特征、步骤等的本发明的每个实施方案，本发明还提供了由或基本上由那些元素、特征、步骤等组成的实施方案。

[1003] 在给出范围的情况下，包括端点。此外，应当理解，除非另有指示或从上下文和/或本领域普通技术人员的理解中以其他方式显而易见，否则表示为范围的值可以假设在本发明的不同实施方案中所述范围内的任何特定值，至范围下限单位的十分之一，除非上下文另有明确规定。还应理解，除非另有指示或从上下文和/或本领域普通技术人员的理解中以其他方式显而易见，否则表示为范围的值可以假设给定范围内的任何子范围，其中子范围的端点以与该范围下限的单位的十分之一相同的准确度表示。

[1004] 另外，应当理解，当前发明的任何特定实施方案可以被明确地排除在任何一个或多个权利要求之外。在给出范围的情况下，该范围内的任何值可以明确地从任何一个或多个权利要求中排除。本发明的组合物和/或方法的任何实施方案、元素、特征、应用或方面可以从任何一个或多个权利要求中排除。为了简洁起见，本文未明确阐述其中排除一个或多个元素、特征、目的或方面的所有实施方案。

[1005] 本文提及的所有出版物、专利和序列数据库条目(包括上面列出的那些项目)在此通过引用以其整体并入，如同每个单独的出版物或专利被具体地和单独地指示通过引用并入。在冲突的情况下，以本申请(包括本文中的任何定义)为准。

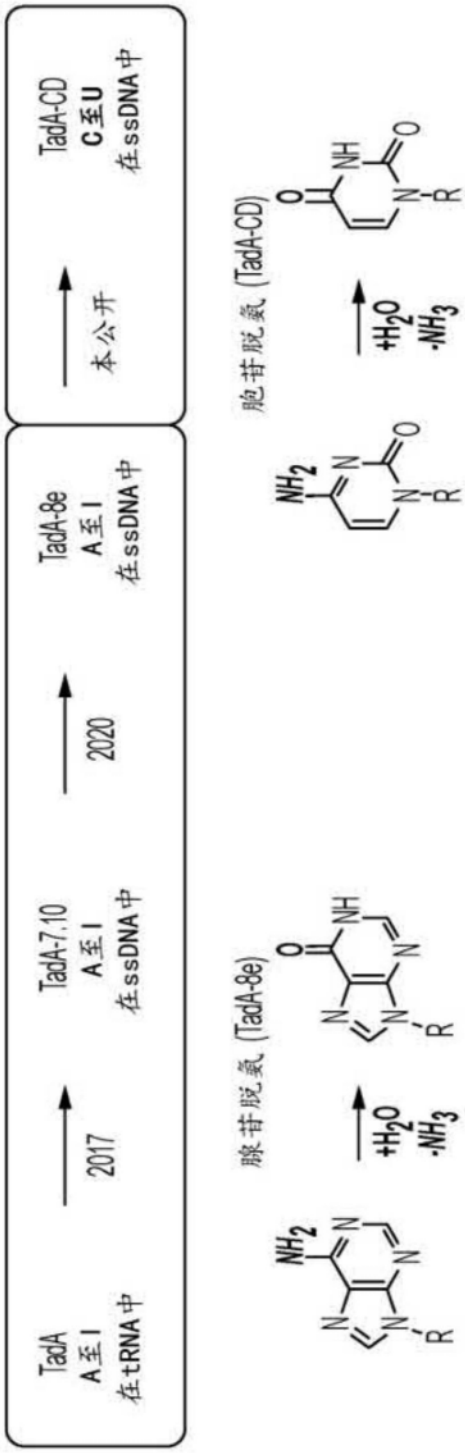

TadA
A至I
在tRNA中
2017
TadA-7.10
A至I
在ssDNA中
2020
TadA-8e
A至I
在ssDNA中
本公开
TadA-CD
C至U
在ssDNA中
腺苷脱氨 (TadA-8e)
+H2O
-NH3
胞苷脱氨 (TadA-CD)
+H2O
-NH3

图1A

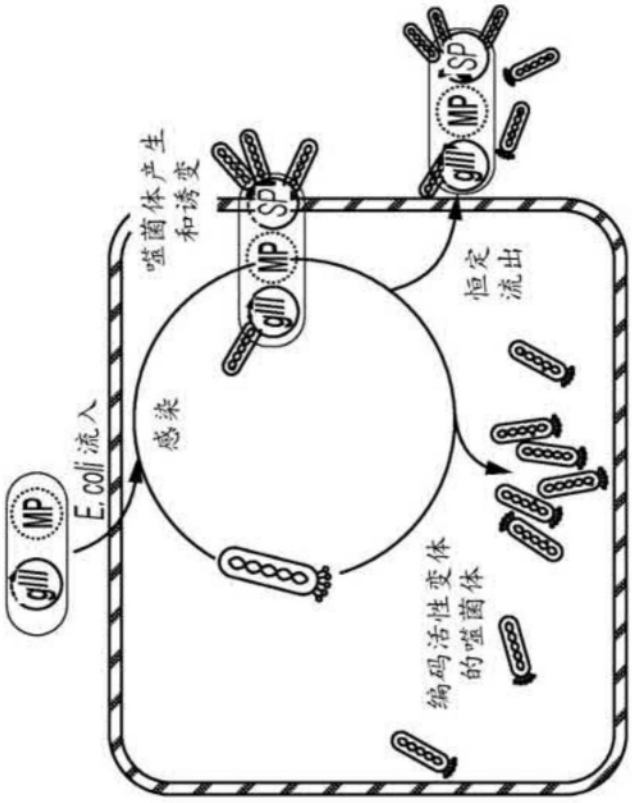

gIII
MP
E. coli 流入
感染
噬菌体产生
和诱变
SP
gIII
MP
SP
恒定
流出
编码活性变体
的噬菌体

图1B

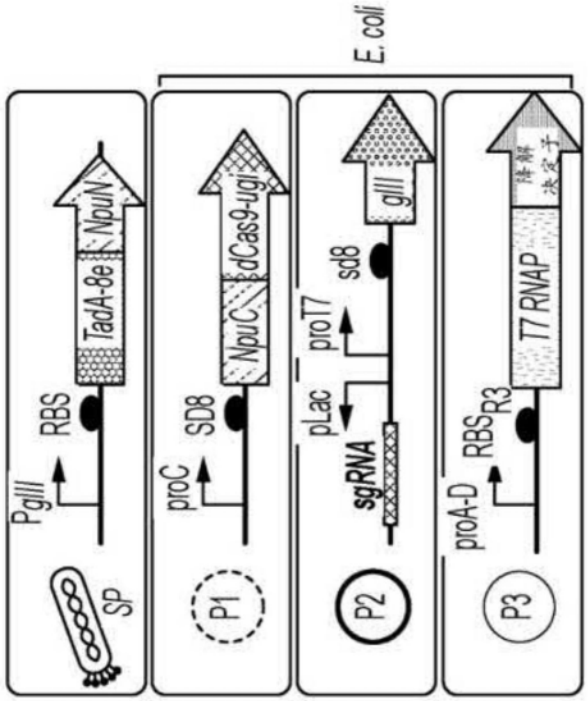

E. coli
PgIII
RBS
TadA-8e
NpuN
SP
P1
proC
SD8
NpuC
dCas9-ugi
P2
pLac
sgRNA
proT7
sd8
gIII
proA-D
RBS
R3
T7 RNAP
降解
决定子
P3

图1C

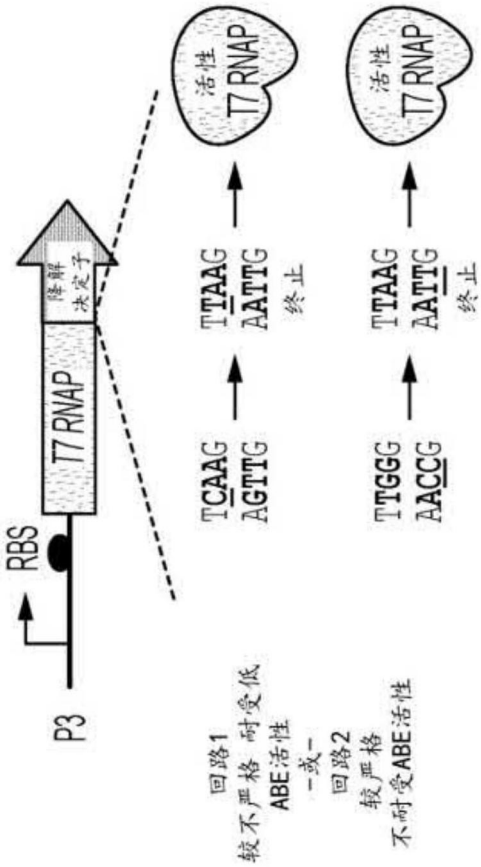

P3
RBS
T7 RNAP
降解决定子
TCAAG
AGTTG
TTAAG
AATTG
终止
活性
T7 RNAP
TTGGG
AACCG
TTAAG
AATTG
终止
活性
T7 RNAP
回路1
较不严格 耐受低
ABE活性
-或-
回路2
较严格
不耐受ABE活性

图1D

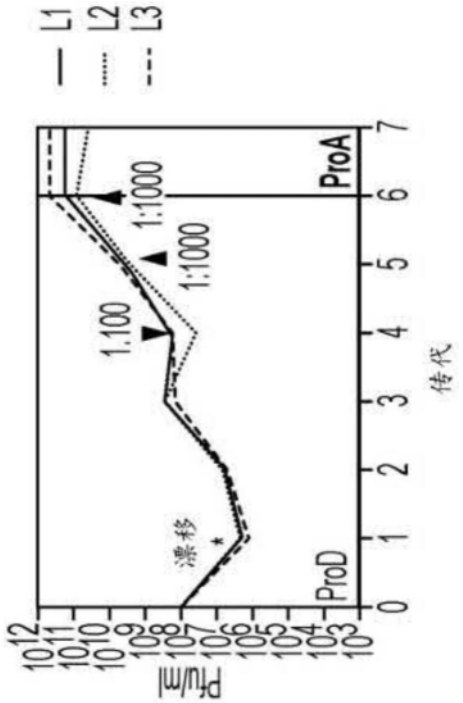

L1
L2
L3
ProA
ProD
1:1000
1:1000
1:100
漂移
传代
PFU/ml

图1E

TadA-8e	R26	E27	V28	A48	M61	Y73	I76	H96	M151	Q154	A158	
TadA-CDa	G	A	G				F	N			S	PACE
TadA-CDb	G	A	G				F	N		R	S	
TadA-CDc	G	A	G				F	N	I			
TadA-CDd		K	A		I			N				PANCE
TadA-CDe		A	G			H		N		H	S	
TadA-Dual	G		A	R		S		N				

图2A

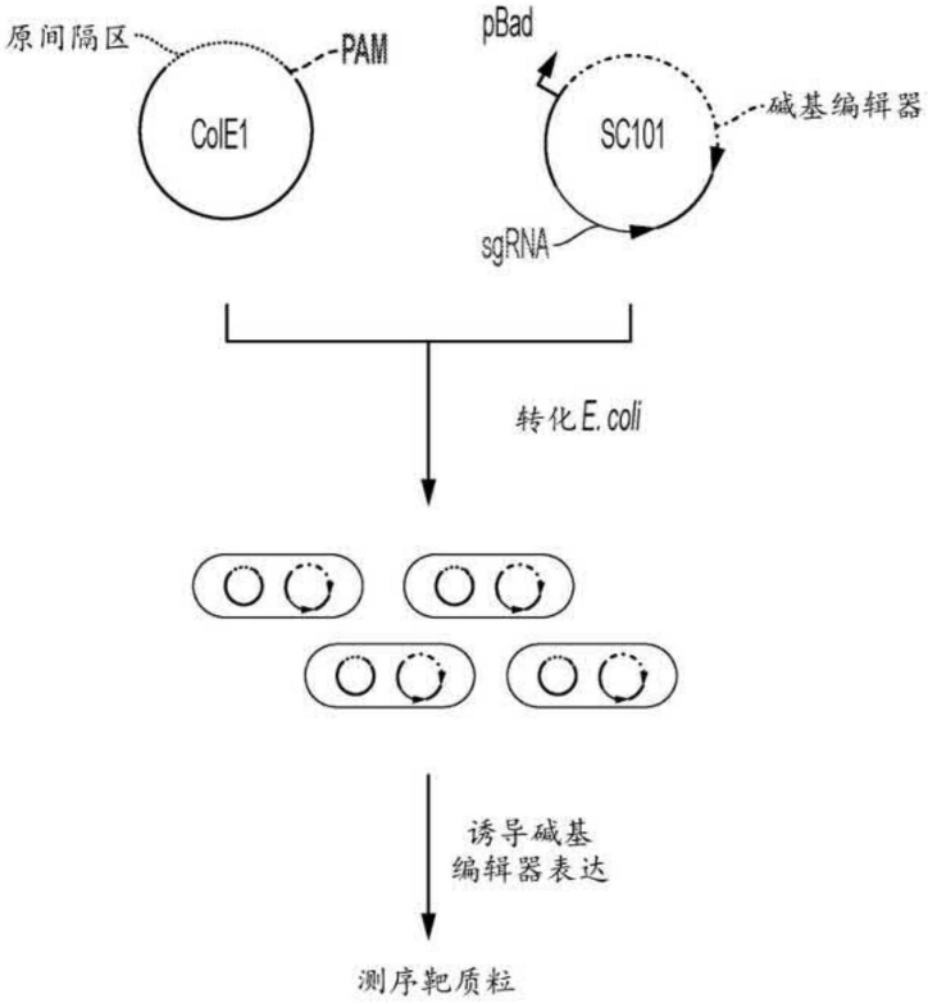


图2B

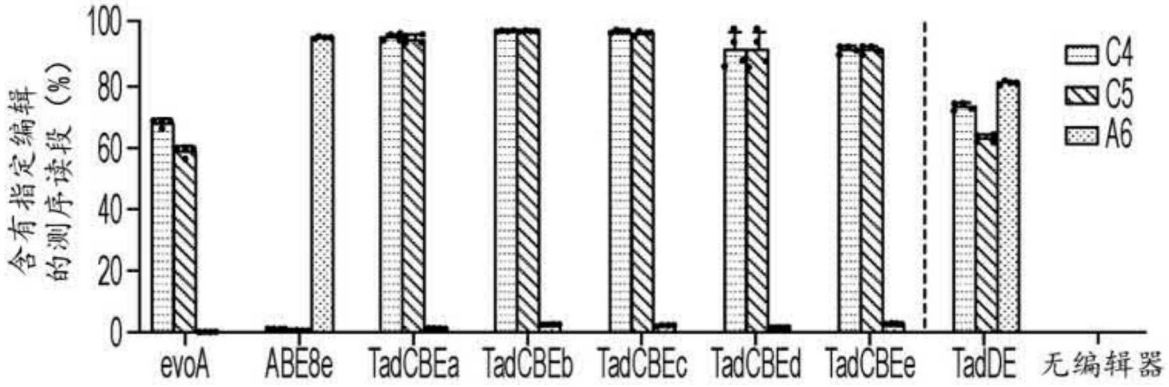

含有指定编辑的测序读段（%）
100
80
60
40
20
0
C4
C5
A6
evoA
ABE8e
TadCBEa
TadCBEb
TadCBEc
TadCBEd
TadCBEe
TadDE
无编辑器

图2C

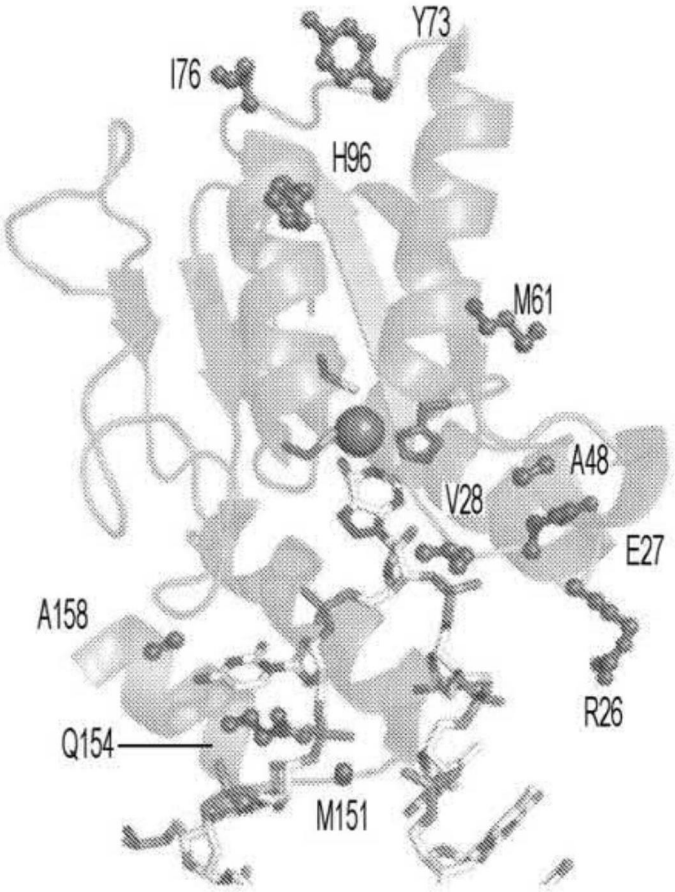

Y73
I76
H96
M61
A48
V28
E27
A158
R26
Q154
M151

图2D

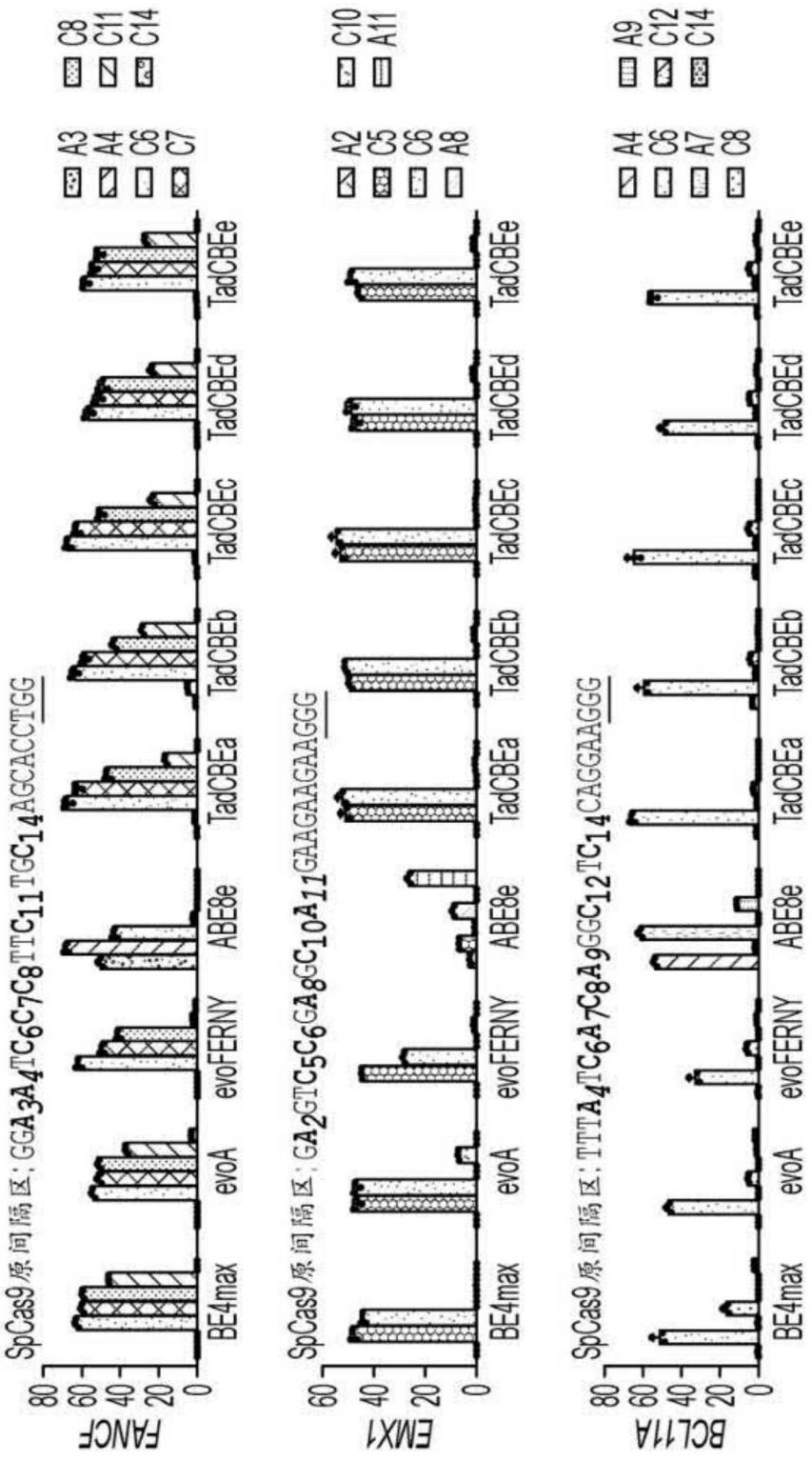

FANCF
EMX1
BCL11A
具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比
SpCas9原间隔区：GGA3A4TC6C7C8TTC11TGC14AGCACCTGG
SpCas9原间隔区：GA2GTC5C6GA8GC10A11GAAGAGAAGGG
SpCas9原间隔区：TTTA4TC6A7C8A9GGC12TC14CAGGAAGGG
BE4max
evoA
evoFERNY
ABE8e
TadCBEa
TadCBEb
TadCBEc
TadCBEd
TadCBEe
A3 A4 C6 C7 C8 C11 C14
A2 C5 C6 A8 C10 A11
A4 C6 A7 C8 A9 C12 C14

图3

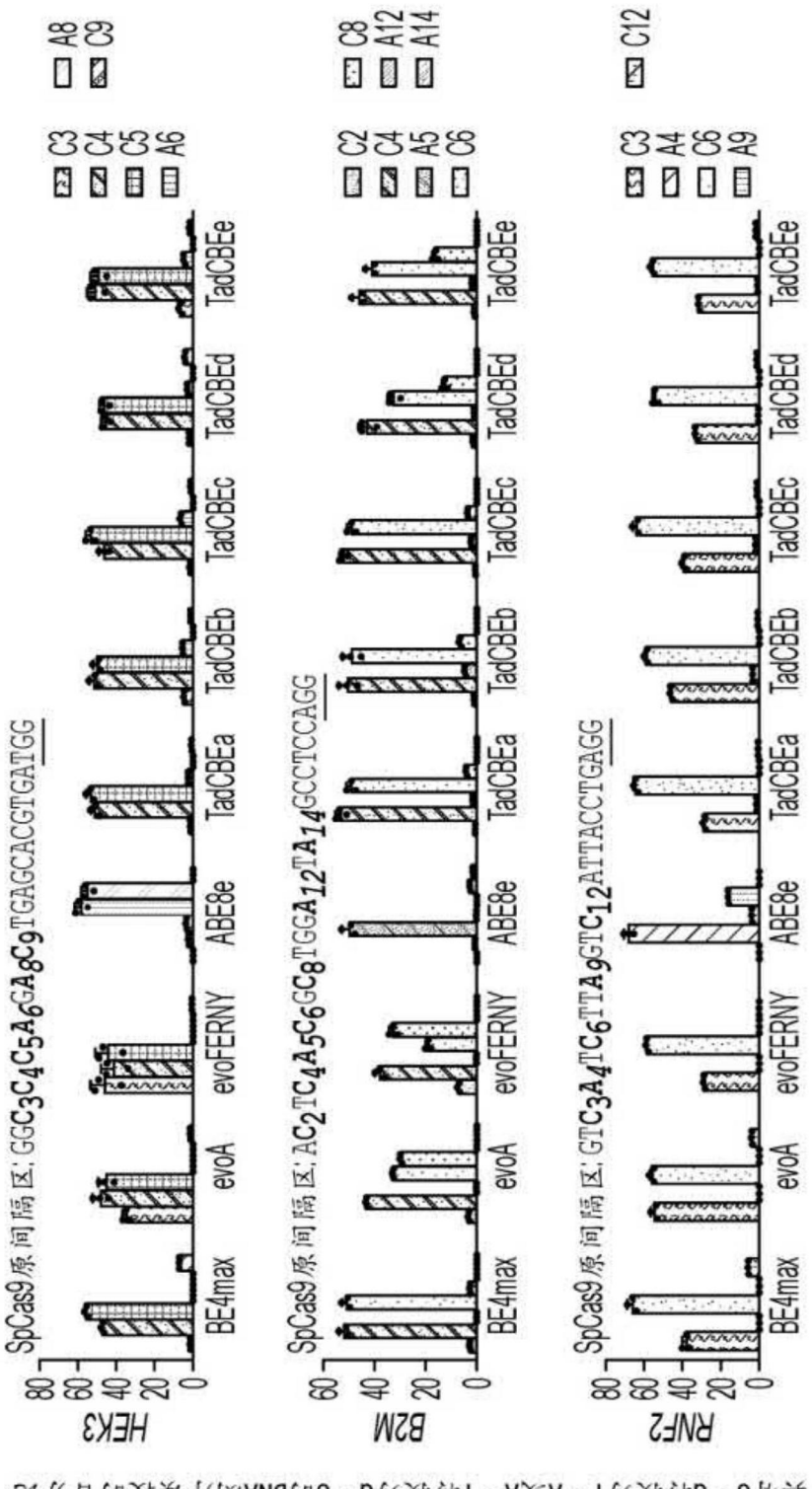

HEK3
B2M
RNF2
SpCas9原间隔区:GGC3C4C5A6GA8C9TGAGCACGTGATGG
SpCas9原间隔区:AC2TC4A5C6GC8TGGA12TA14GCCTCCAGG
SpCas9原间隔区:GTC3A4TC6TTA9GTC12ATTACCTGAGG
BE4max
evoA
evoFERNY
ABE8e
TadCBEa
TadCBEb
TadCBEc
TadCBEd
TadCBEe
C3 C4 C5 A6
A8 C9
C2 C4 A5 C6
C8 A12 A14
C3 A4 C6 A9
C12
具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比

图3(续)

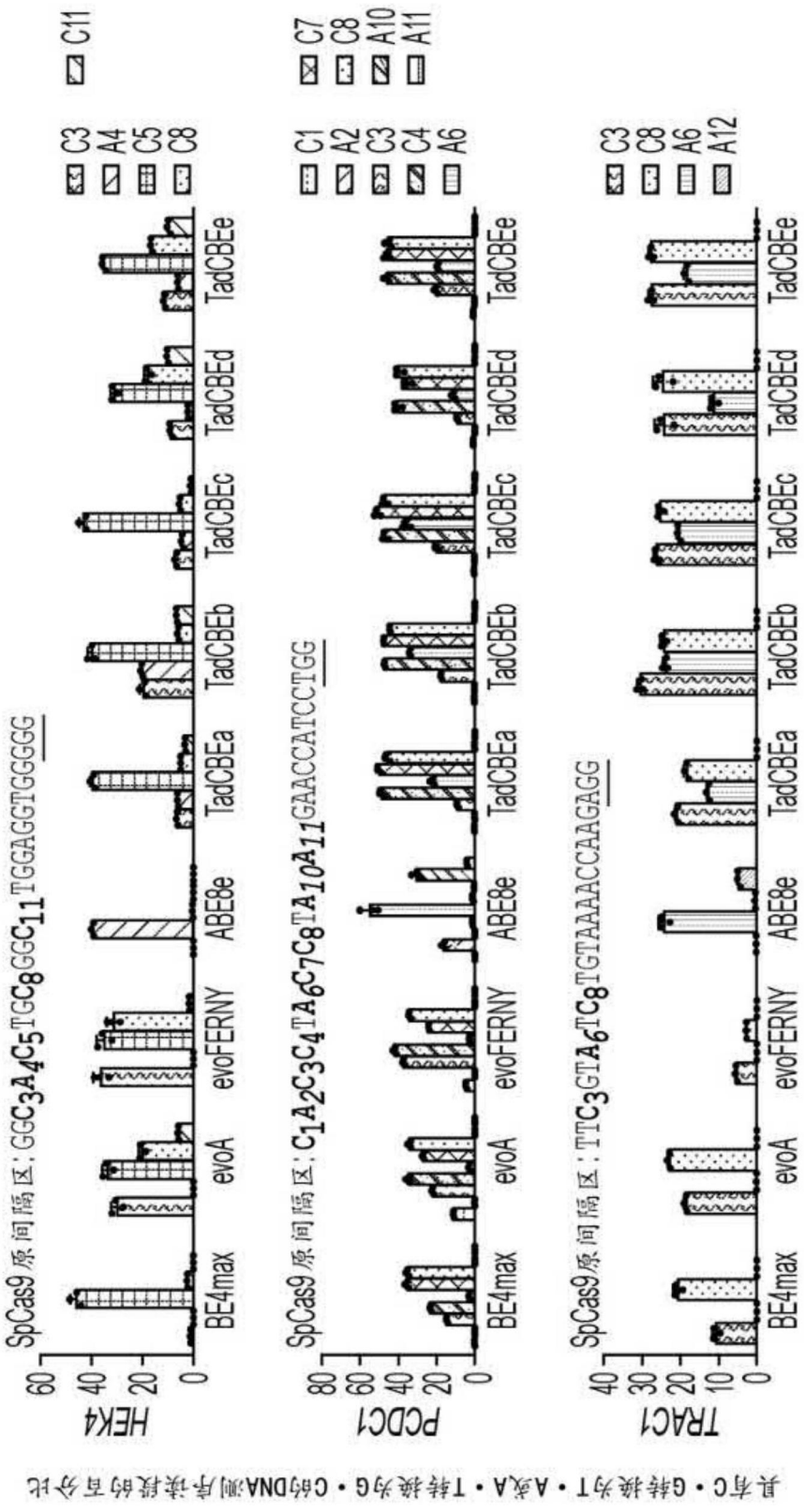

HEK4
PCDC1
TRAC1
具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比
SpCas9原间隔区:GGC3A4C5TGC8GGC11TGGAGGTGGGGG
SpCas9原间隔区:C1A2C3C4TA6C7C8TA10A11GAACCATCCTGG
SpCas9原间隔区:TTC3GTA6TC8TGTAAAACCAAGAGG
C11
C3
A4
C5
C8
C7
C8
A10
A11
C1
A2
C3
C4
A6
C3
C8
A6
A12
BE4max
evoA
evoFERNY
ABE8e
TadCBEa
TadCBEb
TadCBEc
TadCBEd
TadCBEe

图3(续)

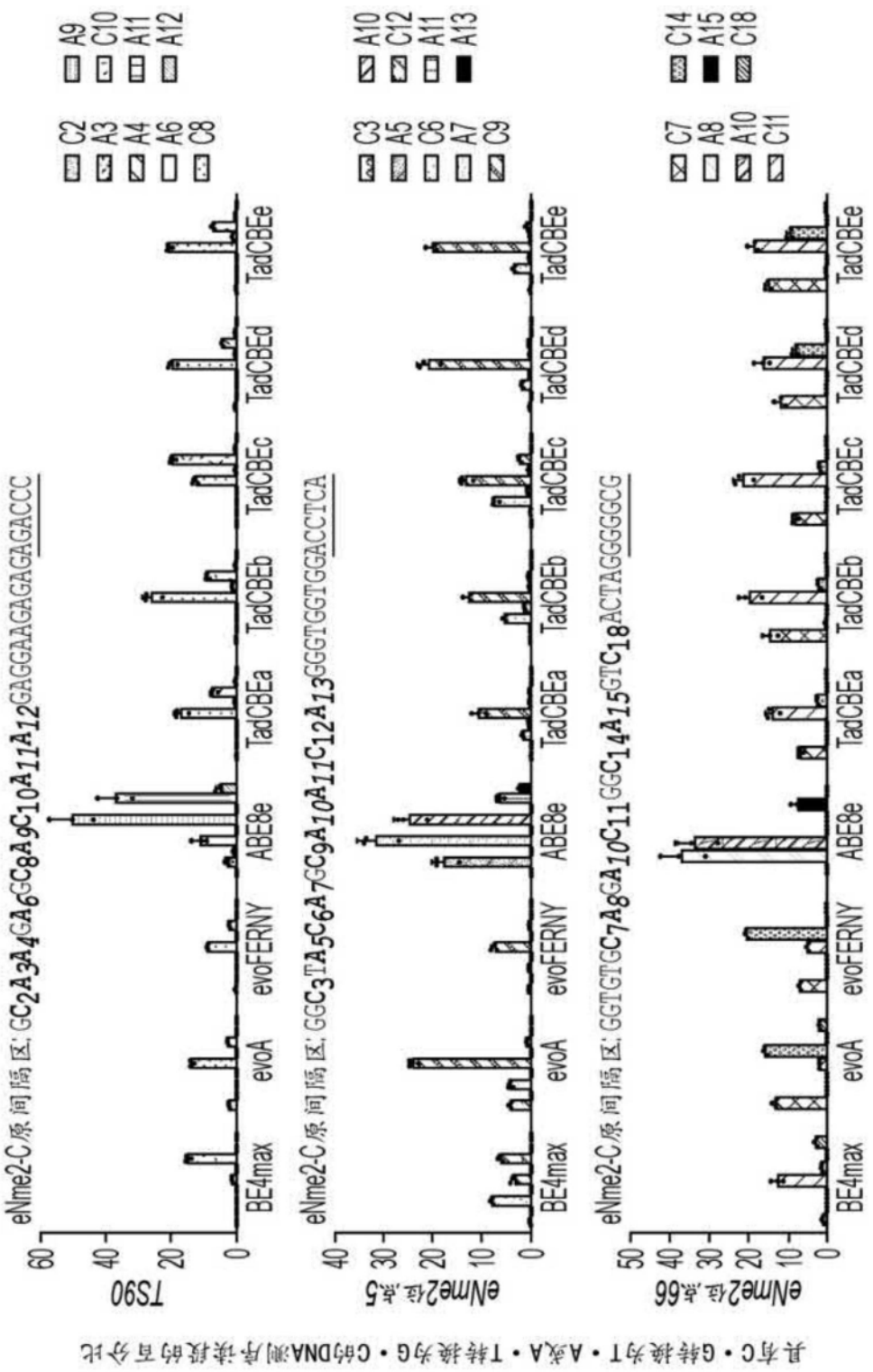

A9
C10
A11
A12
C2
A3
A4
A6
C8
A10
C12
A11
A13
C3
A5
C6
A7
C9
C14
A15
C18
C7
A8
A10
C11
eNme2-C原间隔区:GC2A3A4GA6GC8A9C10A11A12GAGGAAGAGAGAGACCC
eNme2-C原间隔区:GGC3TA5C6A7GC9A10A11C12A13GGGTGGTGGACCTCA
eNme2-C原间隔区:GGTGTGC7A8GA10C11GGC14A15GTC18ACTAGGGGGCG
BE4max
evoA
evoFERNY
ABE8e
TadCBEa
TadCBEb
TadCBEc
TadCBEd
TadCBEe
TS90
eNme2位点5
eNme2位点66
具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比

图4

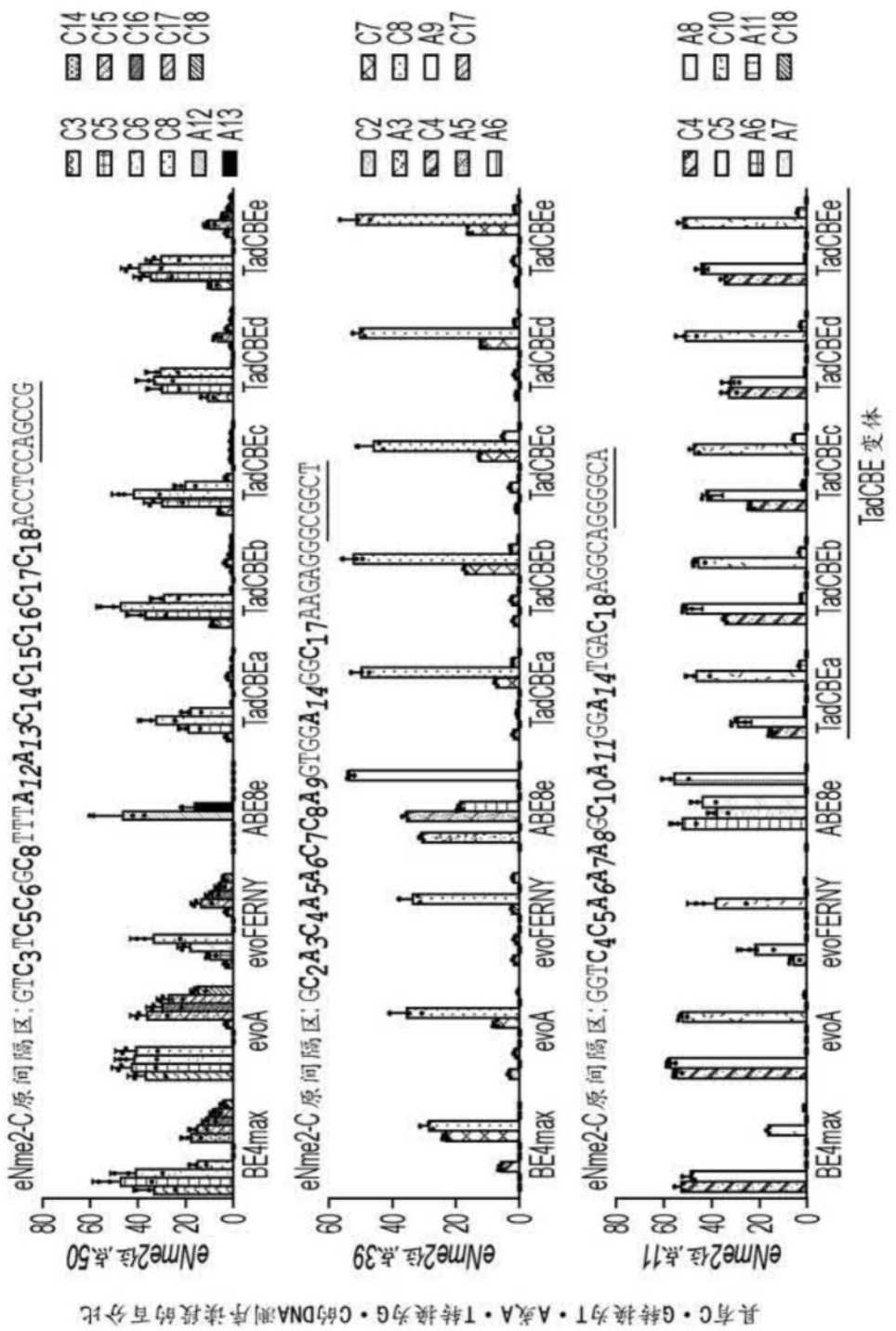

eNme2-C 原间隔区: GTC3TC5C6GC8TTTA12A13C14C15C16C17C18ACCTCCAGCCG
C3 C5 C6 C8 A12 A13
C14 C15 C16 C17 C18
BE4max evoA evoFERNY ABE8e TadCBEa TadCBEb TadCBEc TadCBEd TadCBEe
eNme2位点.50
eNme2-C 原间隔区: GC2A3C4A5A6C7C8A9GTGGA14GGC17AAGAGGGCGGCT
C2 A3 C4 A5 A6
C7 C8 A9 C17
BE4max evoA evoFERNY ABE8e TadCBEa TadCBEb TadCBEc TadCBEd TadCBEe
eNme2位点.39
eNme2-C 原间隔区: GGTC4C5A6A7A8GC10A11GGA14TGAC18AGGCAGGGGCA
C4 C5 A6 A7
A8 C10 A11 C18
BE4max evoA evoFERNY ABE8e TadCBEa TadCBEb TadCBEc TadCBEd TadCBEe
eNme2位点.11
TadCBE 变体
具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比

图4(续)

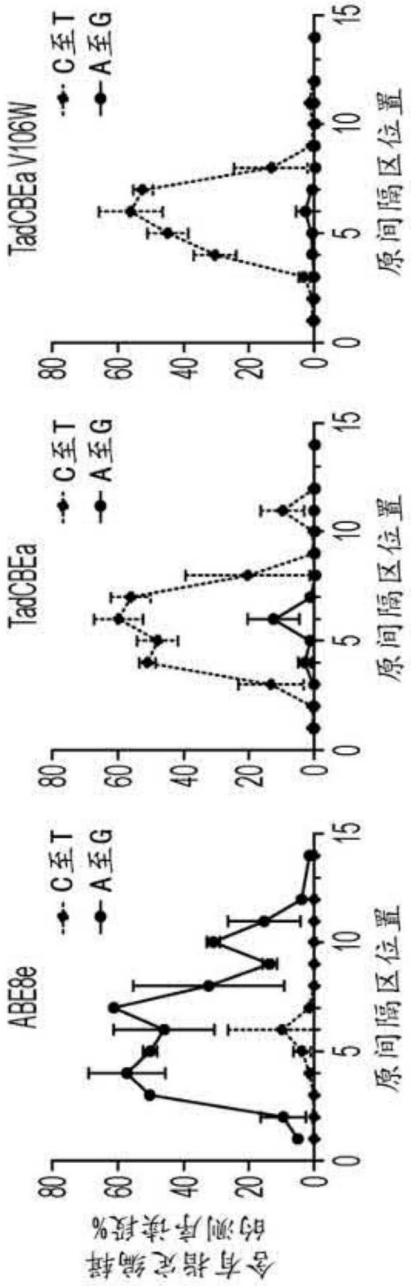

含有指定编辑的测序读段%
ABE8e
TadCBEa
TadCBEa V106W
C至T
A至G
原间隔区位置
0
20
40
60
80
0
5
10
15

图5A

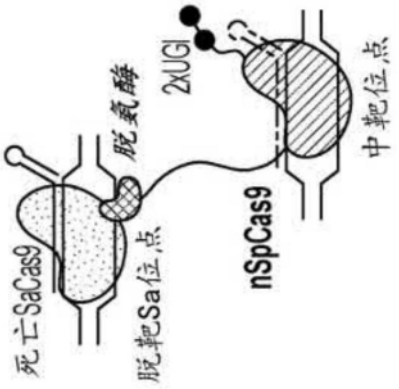


图5B


SaR1
SaR2
SaR3
SaR4
SaR5
SaR6
无BE
Tad CBEd V106W
Tad CBEc V106W
Tad CBEb V106W
Tad CBEa V106W
Tad CBEe
Tad CBEd
Tad CBEc
Tad CBEb
Tad CBEa
YE1
evo FERNY
evoA
BE4max
0
4
8
12
具有C·G转换为T·A的测序读段(%)

图5C

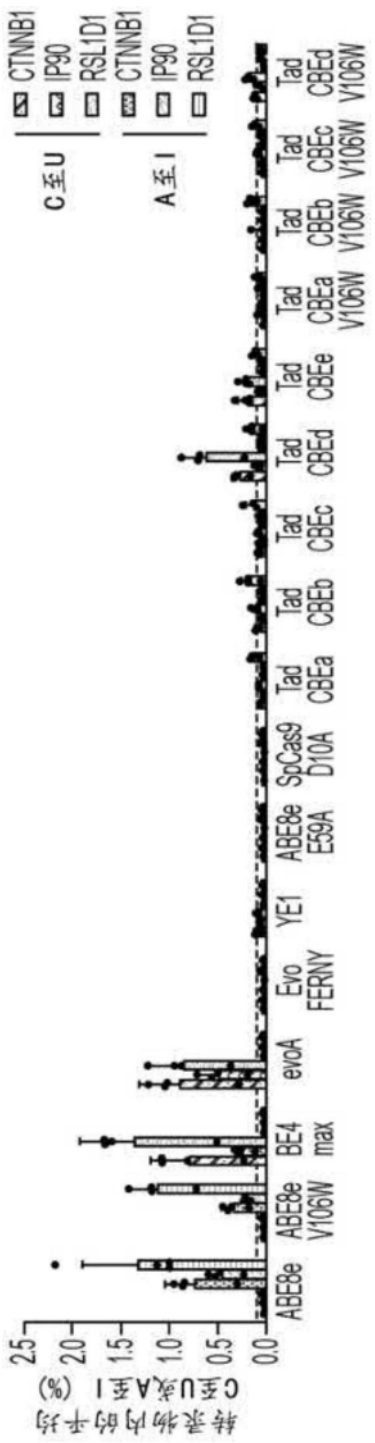

转录物内的平均
C至U或A至I (%)
2.5
2.0
1.5
1.0
0.5
0.0
C至U
A至I
CTNNB1
IP90
RSL1D1
CTNNB1
IP90
RSL1D1
ABE8e
ABE8e V106W
BE4 max
evoA
Evo FERNY
YE1
ABE8e E59A
SpCas9 D10A
Tad CBEa
Tad CBEb
Tad CBEc
Tad CBEd
Tad CBEe
Tad CBEa V106W
Tad CBEb V106W
Tad CBEc V106W
Tad CBEd V106W

图5D

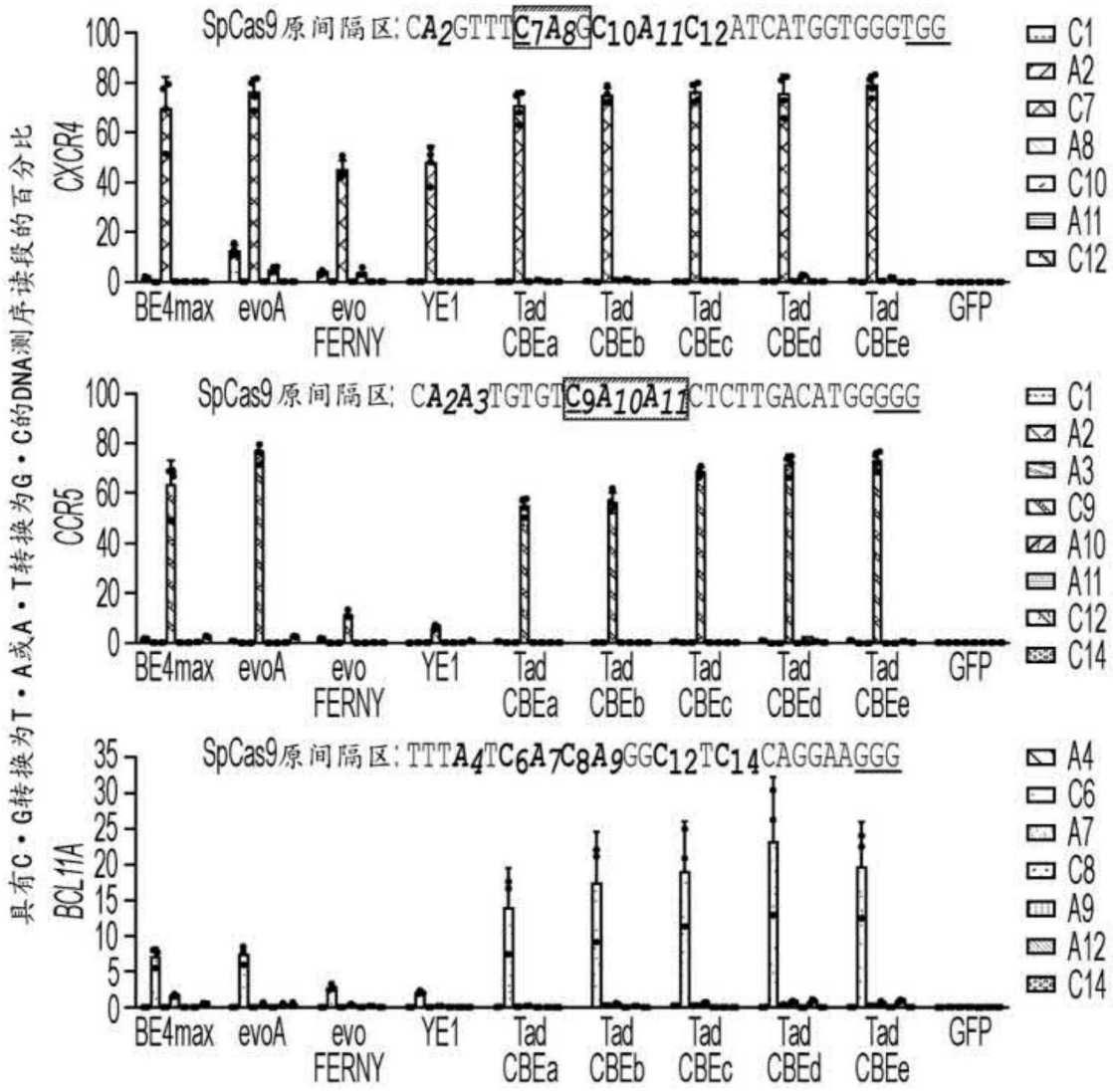

SpCas9原间隔区: CA2GTTTC7A8GC10A11C12ATCATGGTGGGTGG
CXCR4
C1
A2
C7
A8
C10
A11
C12
SpCas9原间隔区: CA2A3TGTGTC9A10A11CTCTTGACATGGGGG
CCR5
C1
A2
A3
C9
A10
A11
C12
C14
SpCas9原间隔区: TTTA4TC6A7C8A9GGC12TC14CAGGAAGGG
BCL11A
A4
C6
A7
C8
A9
A12
C14
具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比
BE4max
evoA
evo FERNY
YE1
Tad CBEa
Tad CBEb
Tad CBEc
Tad CBEd
Tad CBEe
GFP

图6

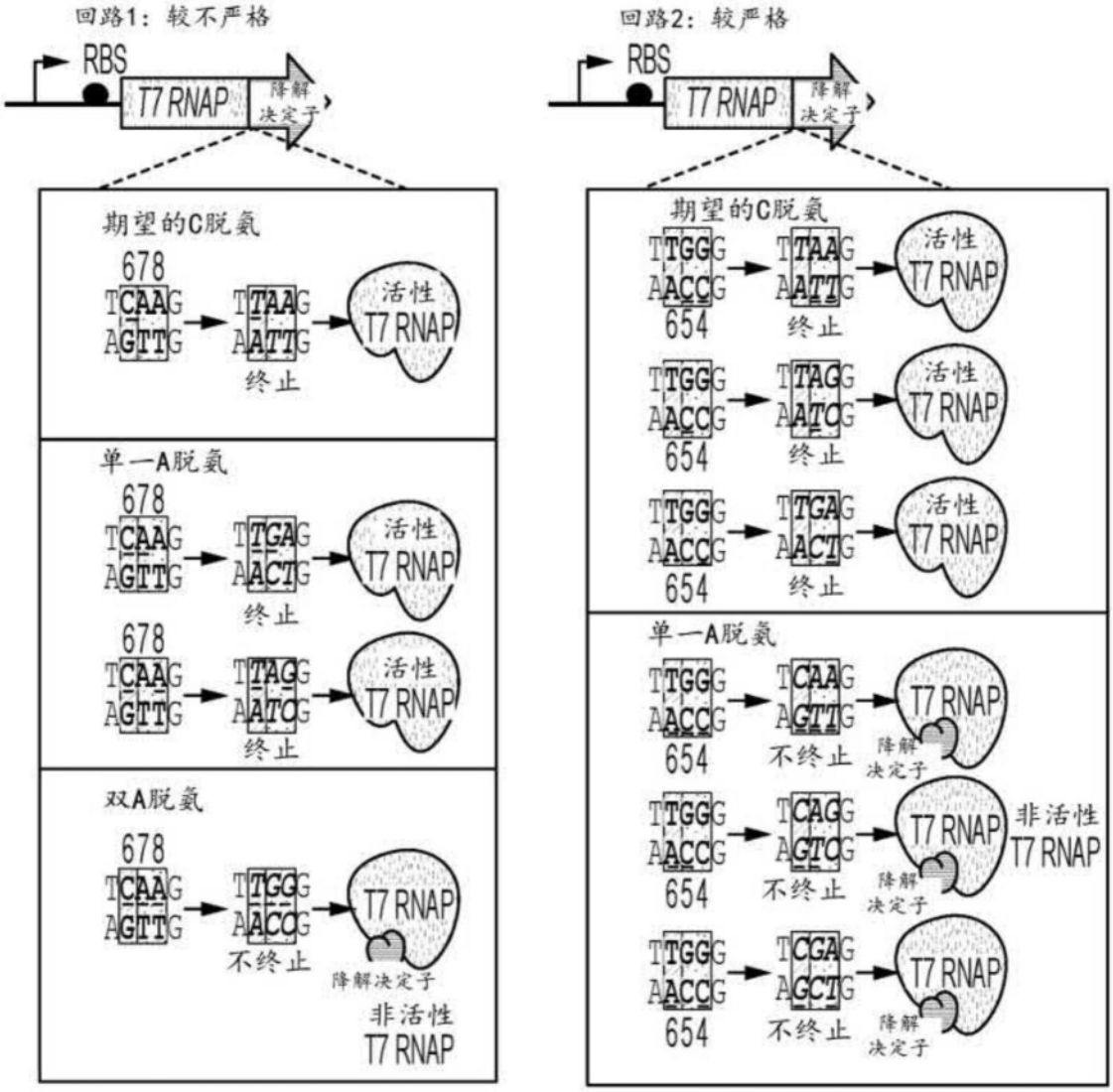

回路1：较不严格
RBS
T7 RNAP
降解决定子
期望的C脱氨
678
TCAAG
AGTTG
TTAAG
AATTG
终止
活性T7 RNAP
单一A脱氨
TTGAG
AACTG
TTAGG
AATCG
双A脱氨
TTGGG
AACCG
T7 RNAP
不终止
非活性T7 RNAP
回路2：较严格
654
TTAAG
AATTG
TTAGG
AATCG
TTGAG
AACTG
TCAAG
AGTTG
TCAGG
AGTCG
TCGAG
AGCTG

图7

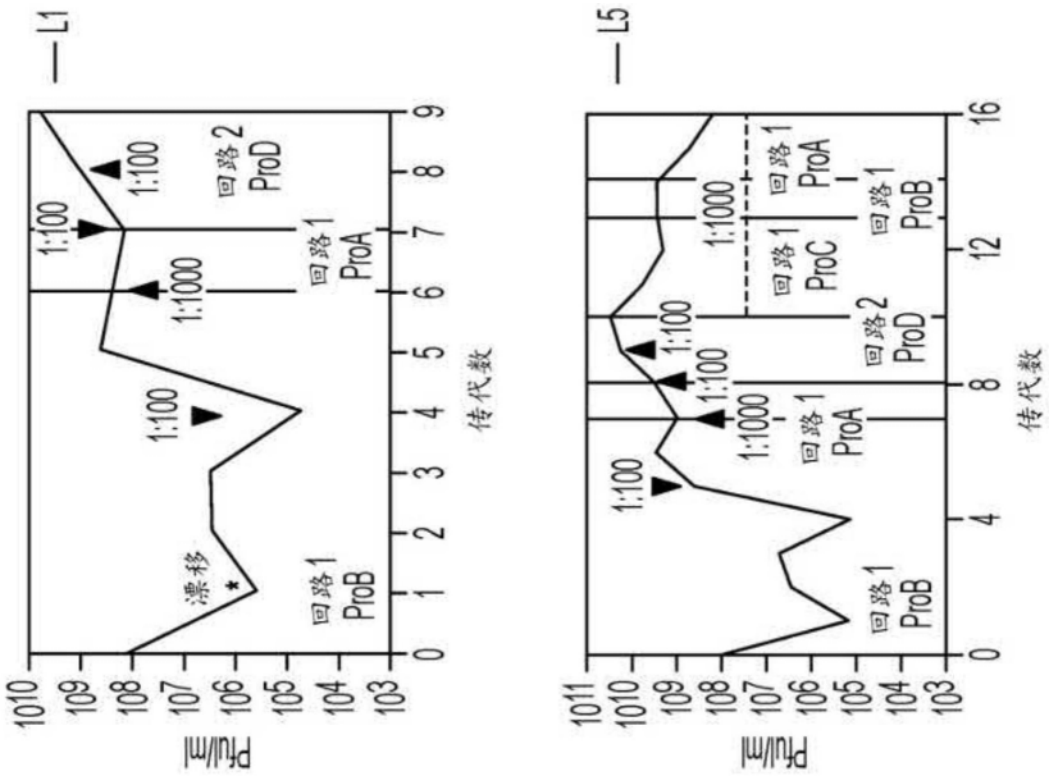


图8A

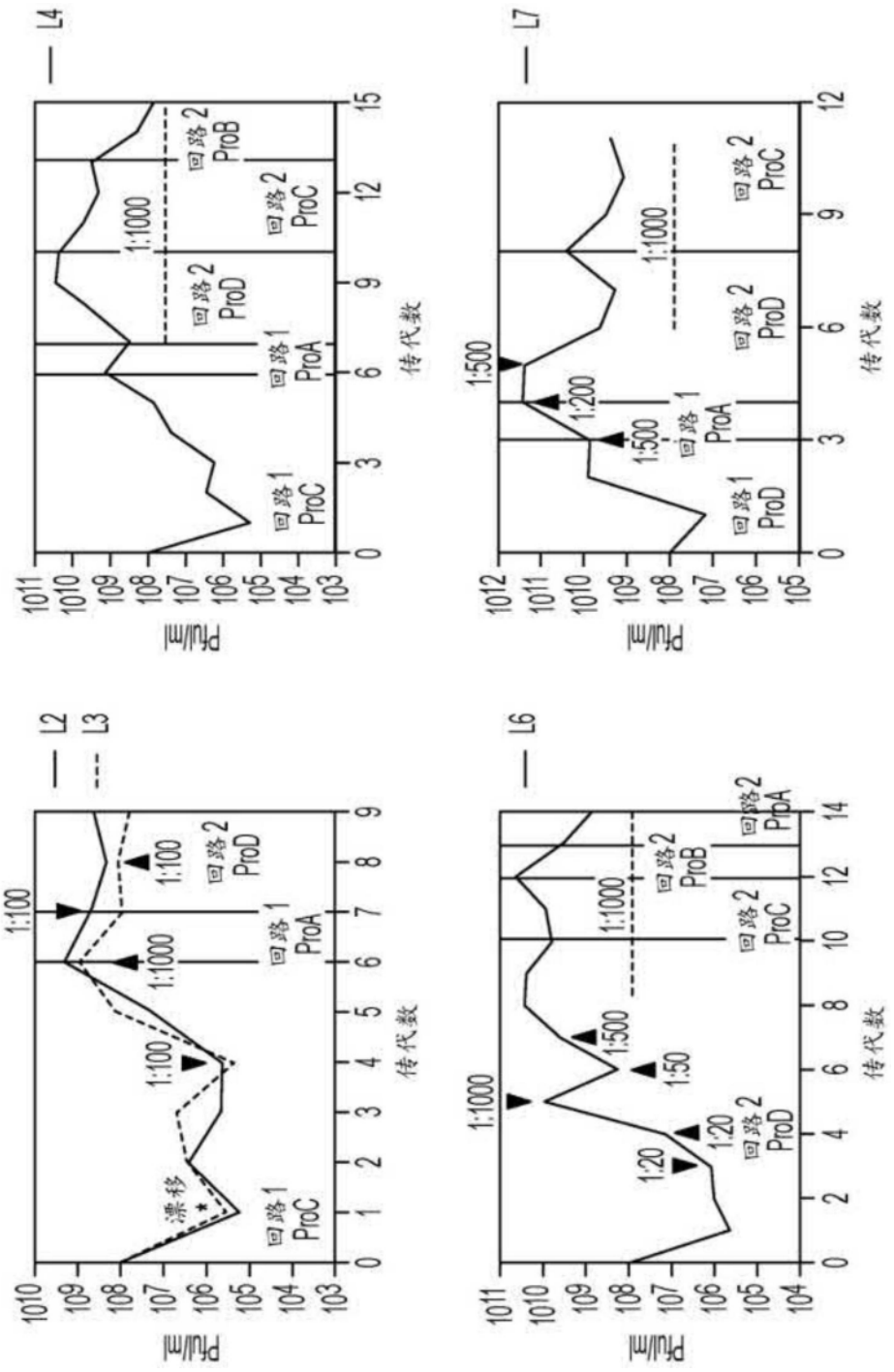

L4
L7
L2
L3
L6
回路1 ProC
回路1 ProA
回路2 ProD
回路2 ProC
回路2 ProB
回路2 ProA
1:100
1:1000
1:500
1:200
1:50
1:20
漂移
传代数

图8A(续)

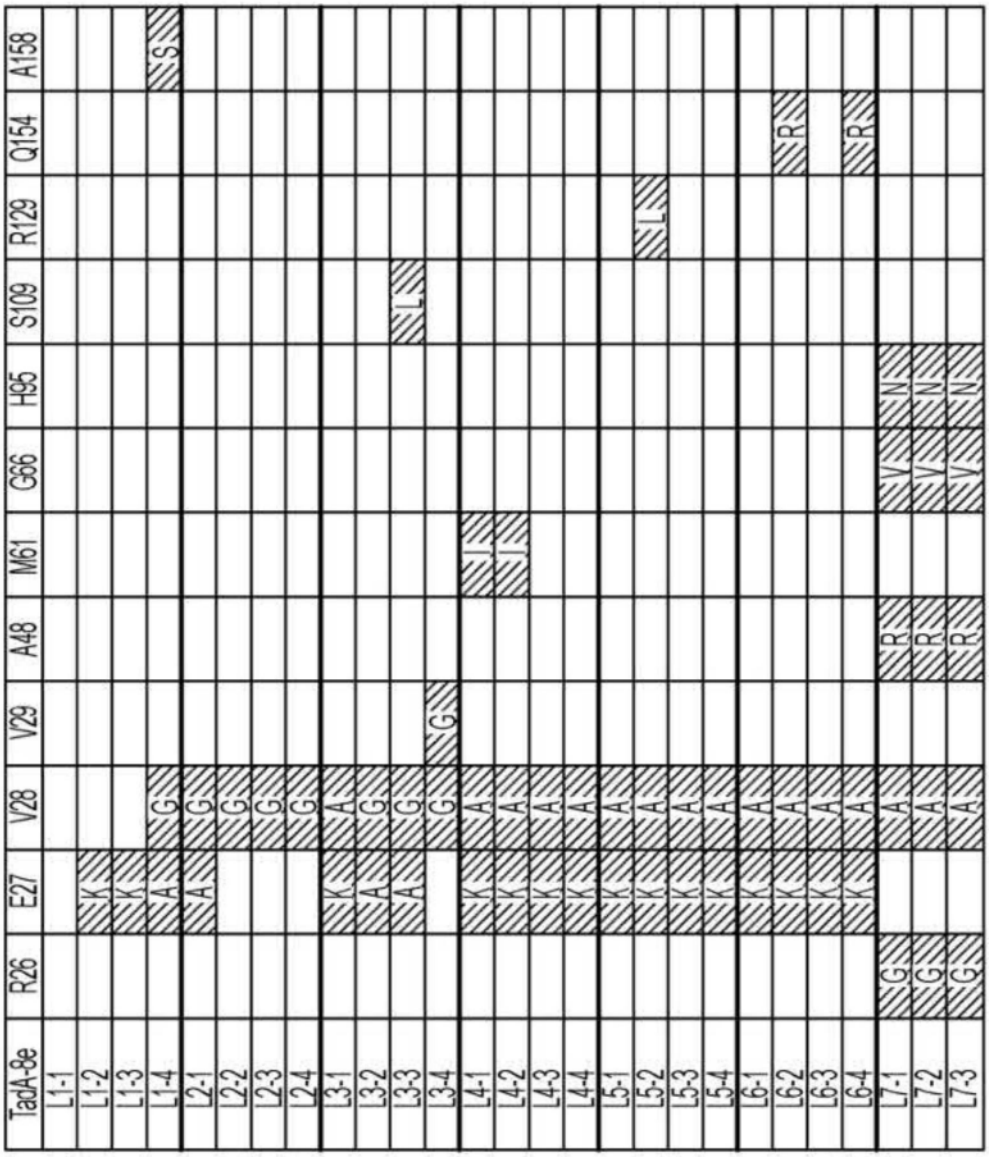

TadA-8e
R26
E27
V28
V29
A48
M61
G66
H95
S109
R129
Q154
A158
L1-1
L1-2
L1-3
L1-4
L2-1
L2-2
L2-3
L2-4
L3-1
L3-2
L3-3
L3-4
L4-1
L4-2
L4-3
L4-4
L5-1
L5-2
L5-3
L5-4
L6-1
L6-2
L6-3
L6-4
L7-1
L7-2
L7-3

图8B

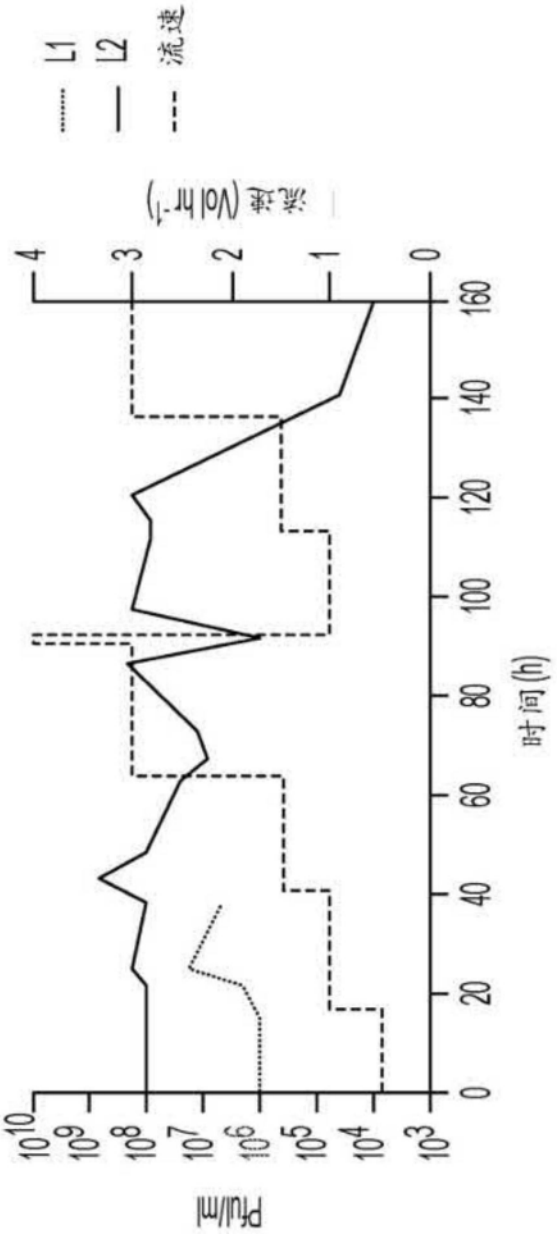

L1
L2
流速
流速 (Vol hr-1)
4
3
2
1
0
160
140
120
100
80
60
40
20
0
时间 (h)
Pfu/ml
10^10
10^9
10^8
10^7
10^6
10^5
10^4
10^3

图9A

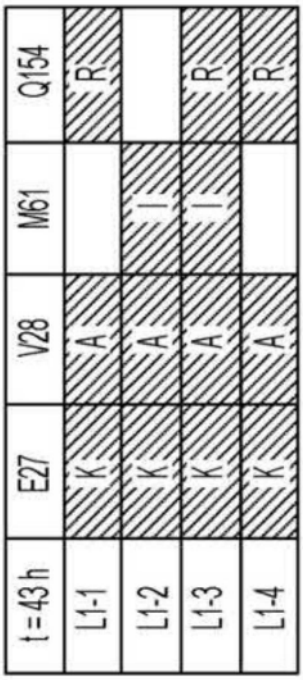


图9B

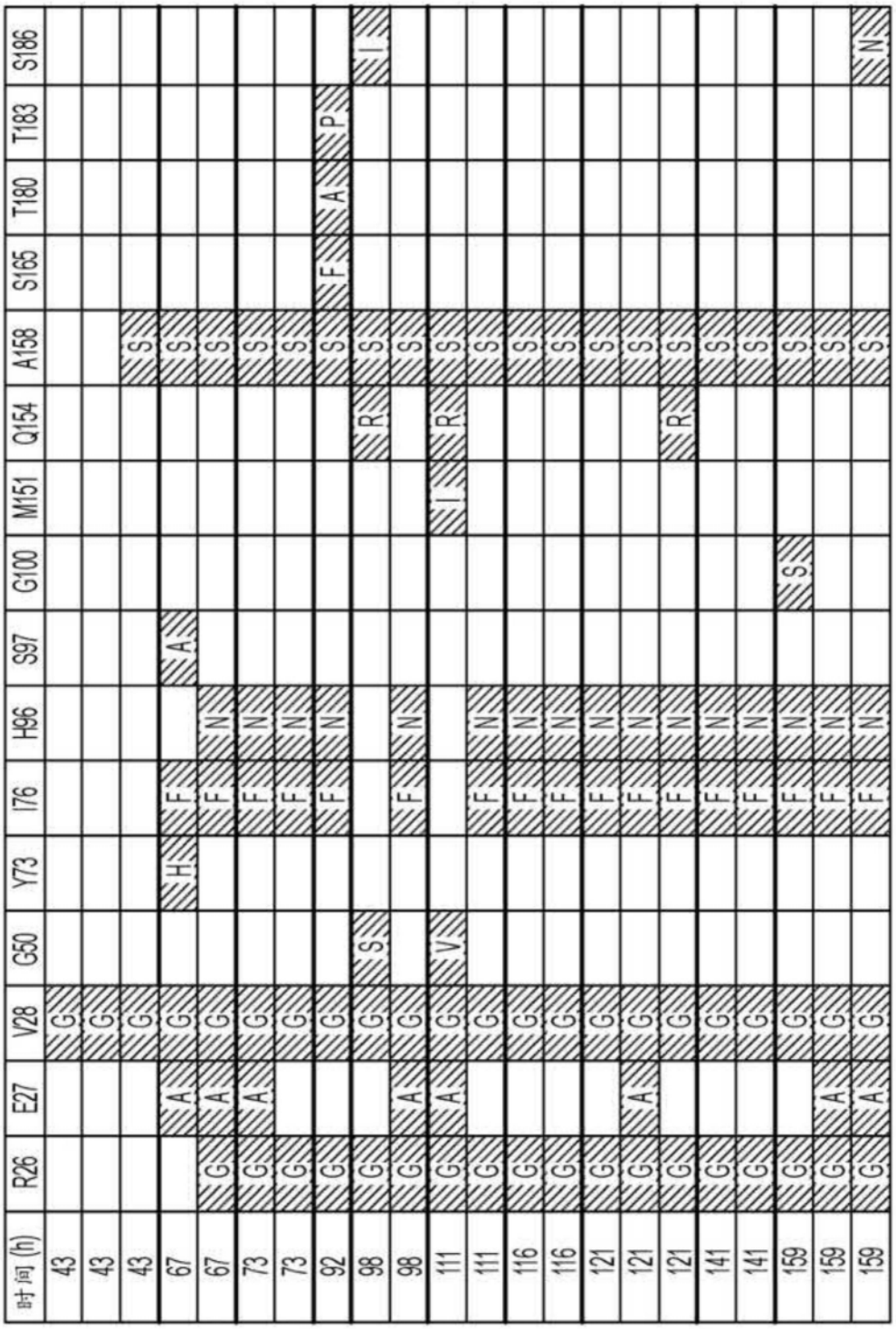
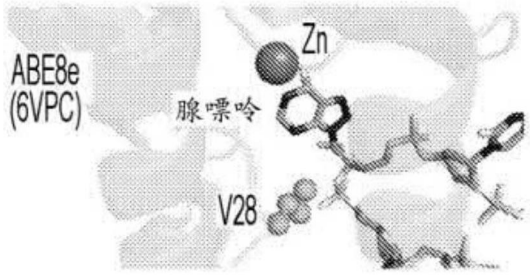
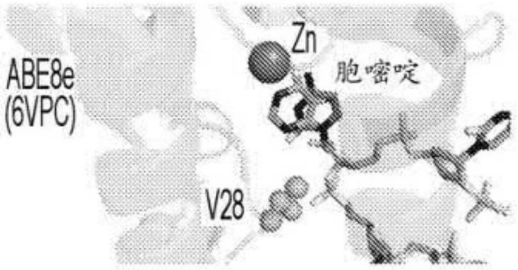

时间 (h)
R26
E27
V28
G50
Y73
I76
H96
S97
G100
M151
Q154
A158
S165
T180
T183
S186
43
67
73
92
98
111
116
121
141
159

图9C



Val 28支持腺嘌呤靠近催化锌的定位

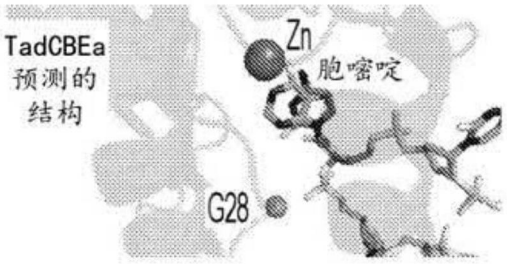
图10A



胞嘧啶的脱氨需要将DNA底物向锌移动

Val 28可能通过空间位阻阻碍这种移动

图10B



具有叠加的来自6VPC的R环的TadCBEa预测的结构

TadCBEa中的Gly 28可能减轻空间约束以允许正确的胞嘧啶定位

图10C

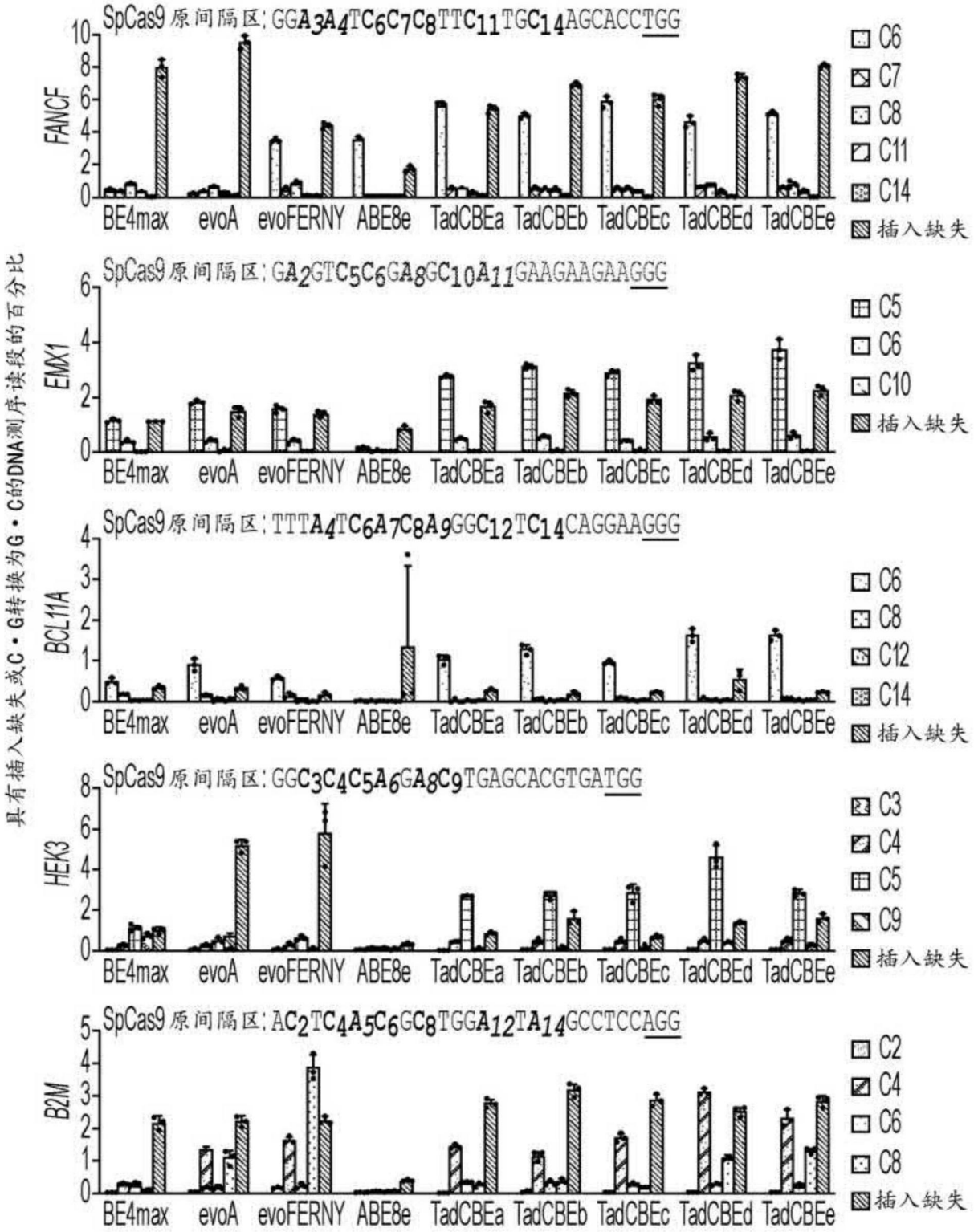

具有插入缺失或C·G转换为G·C的DNA测序读段的百分比
SpCas9 原间隔区: GGA3A4TC6C7C8TTC11TGC14AGCACCTGG
FANCF
C6
C7
C8
C11
C14
插入缺失
SpCas9 原间隔区: GA2GTC5C6GA8GC10A11GAAGAAGAAGGG
EMX1
C5
C6
C10
插入缺失
SpCas9 原间隔区: TTTA4TC6A7C8A9GGC12TC14CAGGAAGGG
BCL11A
C6
C8
C12
C14
插入缺失
SpCas9 原间隔区: GGC3C4C5A6GA8C9TGAGCACGTGATGG
HEK3
C3
C4
C5
C9
插入缺失
SpCas9 原间隔区: AC2TC4A5C6GC8TGGA12TA14GCCTCCAGG
B2M
C2
C4
C6
C8
插入缺失
BE4max
evoA
evoFERNY
ABE8e
TadCBEa
TadCBEb
TadCBEc
TadCBEd
TadCBEe

图11

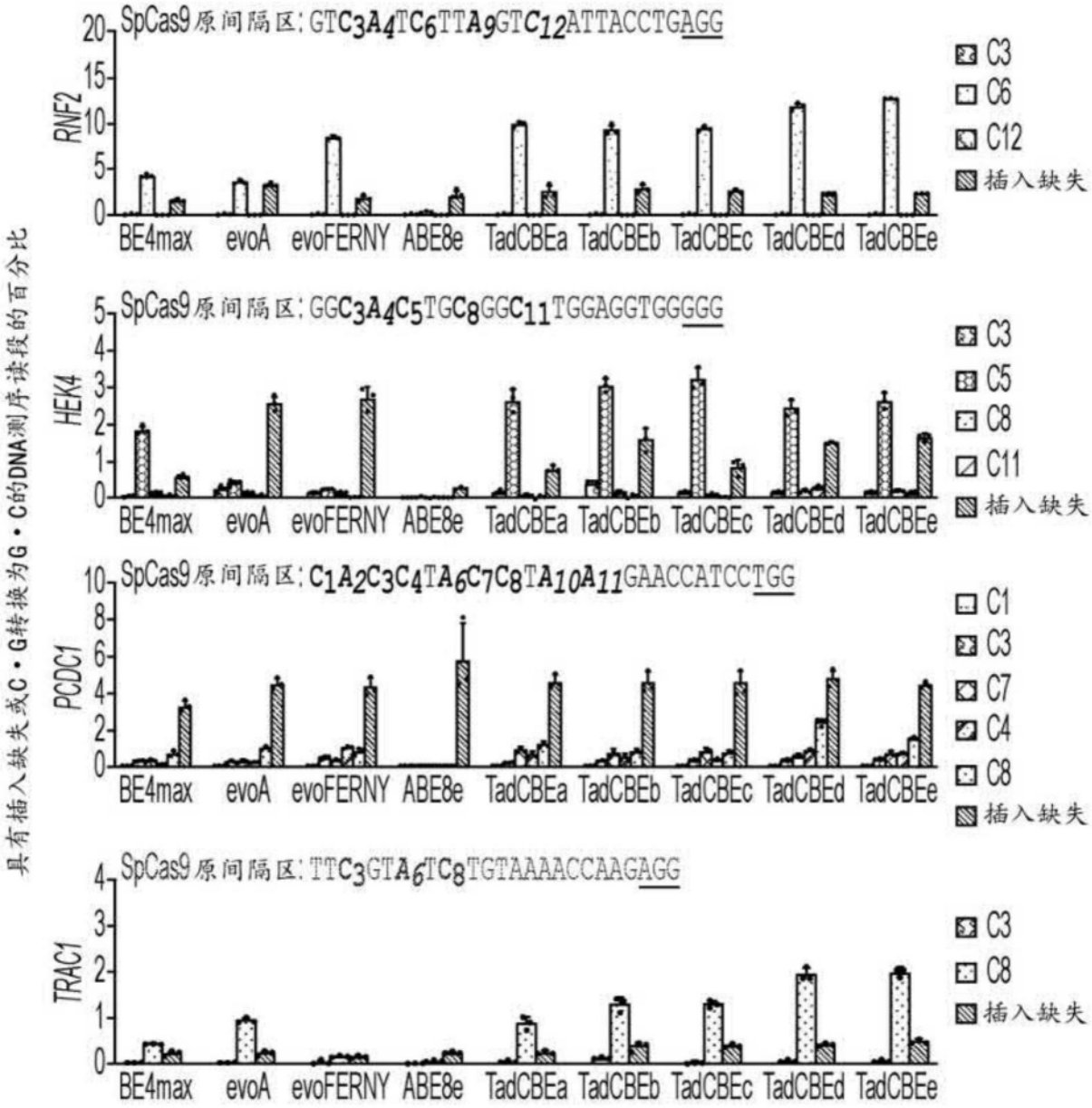

具有插入缺失或C·G转换为G·C的DNA测序读段的百分比
RNF2
SpCas9原间隔区: GTC3A4TC6TTA9GTC12ATTACCTGAGG
C3
C6
C12
插入缺失
HEK4
SpCas9原间隔区: GGC3A4C5TGC8GGC11TGGAGGTGGGGG
C3
C5
C8
C11
插入缺失
PCDC1
SpCas9原间隔区: C1A2C3C4TA6C7C8TA10A11GAACCATCCTGG
C1
C3
C7
C4
C8
插入缺失
TRAC1
SpCas9原间隔区: TTC3GTA6TC8TGTAAAACCAAGAGG
C3
C8
插入缺失
BE4max
evoA
evoFERNY
ABE8e
TadCBEa
TadCBEb
TadCBEc
TadCBEd
TadCBEe

图11(续)

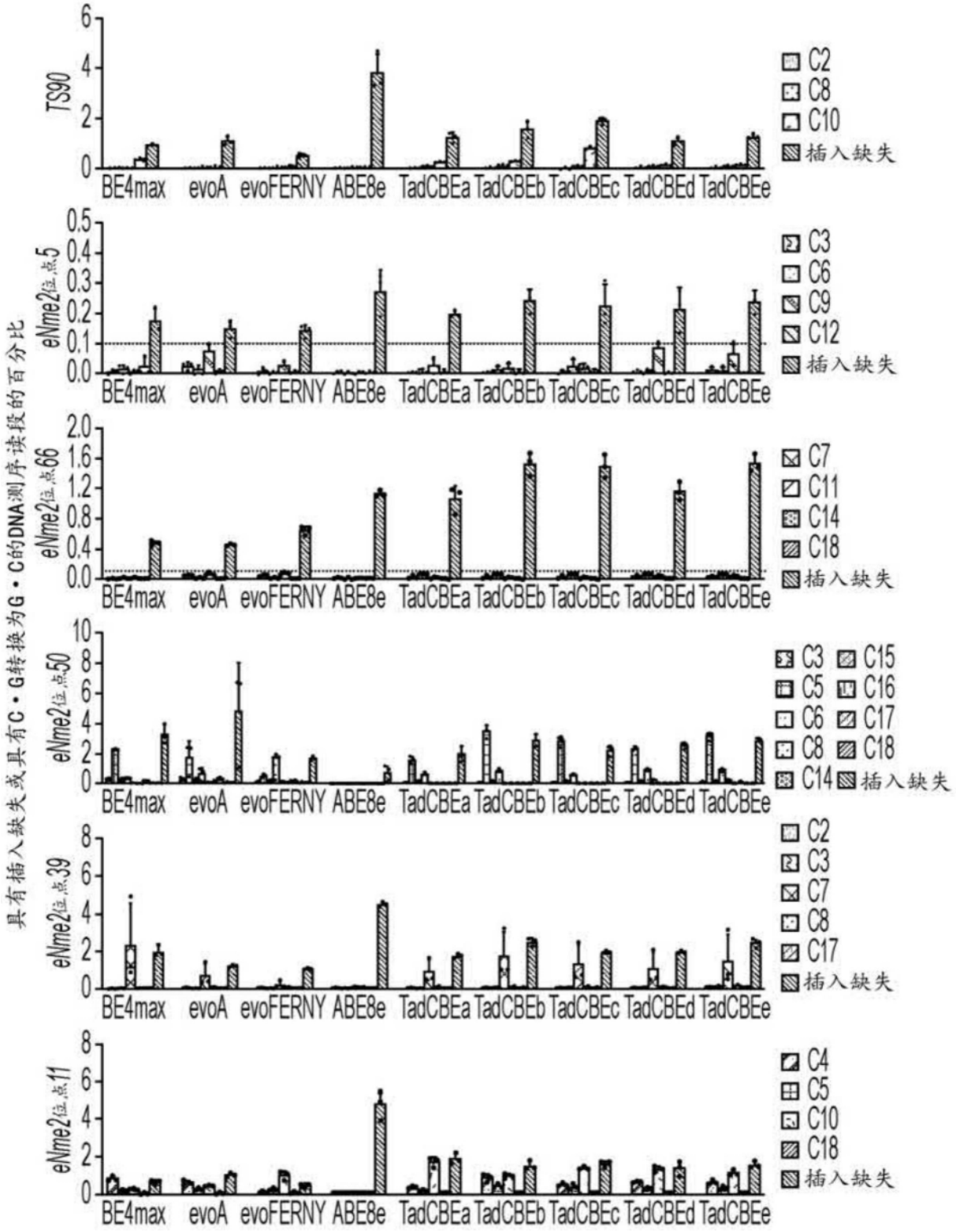

具有插入缺失或具有C·G转换为G·C的DNA测序读段的百分比
TS90
eNme2位点5
eNme2位点66
eNme2位点50
eNme2位点39
eNme2位点11
BE4max
evoA
evoFERNY
ABE8e
TadCBEa
TadCBEb
TadCBEc
TadCBEd
TadCBEe
C2
C8
C10
插入缺失
C3
C6
C9
C12
C7
C11
C14
C18
C5
C15
C16
C17
C4

图12

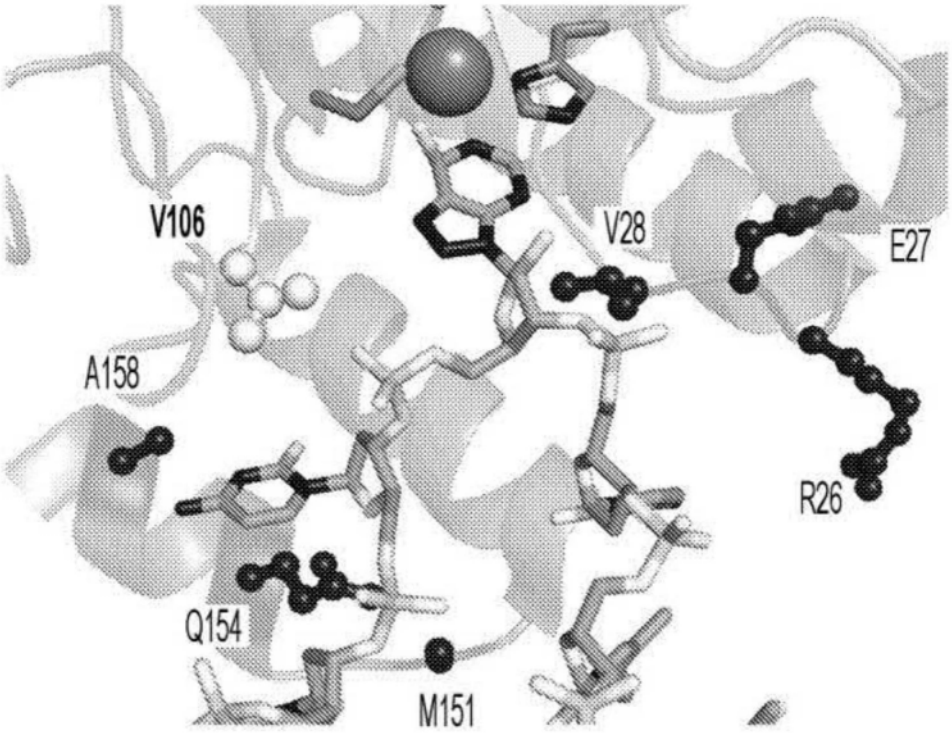


图13

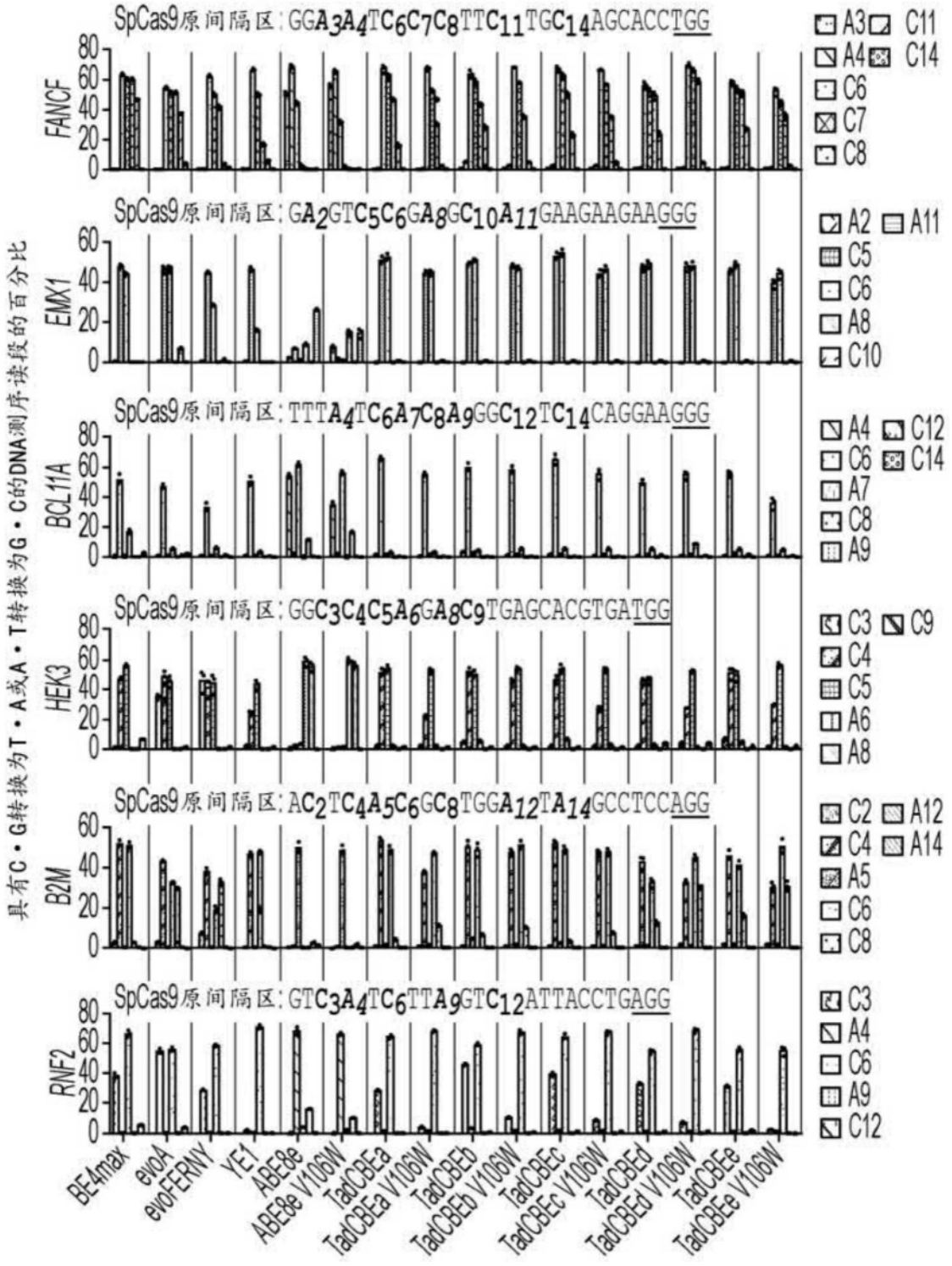

SpCas9原间隔区: GGA3A4TC6C7C8TTC11TGC14AGCACCTGG
FANCF
A3 C11
A4 C14
C6
C7
C8
SpCas9原间隔区: GA2GTC5C6GA8GC10A11GAAGAAGAAGGG
EMX1
A2 A11
C5
C6
A8
C10
SpCas9原间隔区: TTTA4TC6A7C8A9GGC12TC14CAGGAAGGG
BCL11A
A4 C12
C6 C14
A7
C8
A9
SpCas9原间隔区: GGC3C4C5A6GA8C9TGAGCACGTGATGG
HEK3
C3 C9
C4
C5
A6
A8
SpCas9原间隔区: AC2TC4A5C6GC8TGGA12TA14GCCTCCAGG
B2M
C2 A12
C4 A14
A5
C6
C8
SpCas9原间隔区: GTC3A4TC6TTA9GTC12ATTACCTGAGG
RNF2
C3
A4
C6
A9
C12
具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比
BE4max
evoA
evoFERNY
YE1
ABE8e
ABE8e V106W
TadCBEa
TadCBEa V106W
TadCBEb
TadCBEb V106W
TadCBEc
TadCBEc V106W
TadCBEd
TadCBEd V106W
TadCBEe
TadCBEe V106W

图14

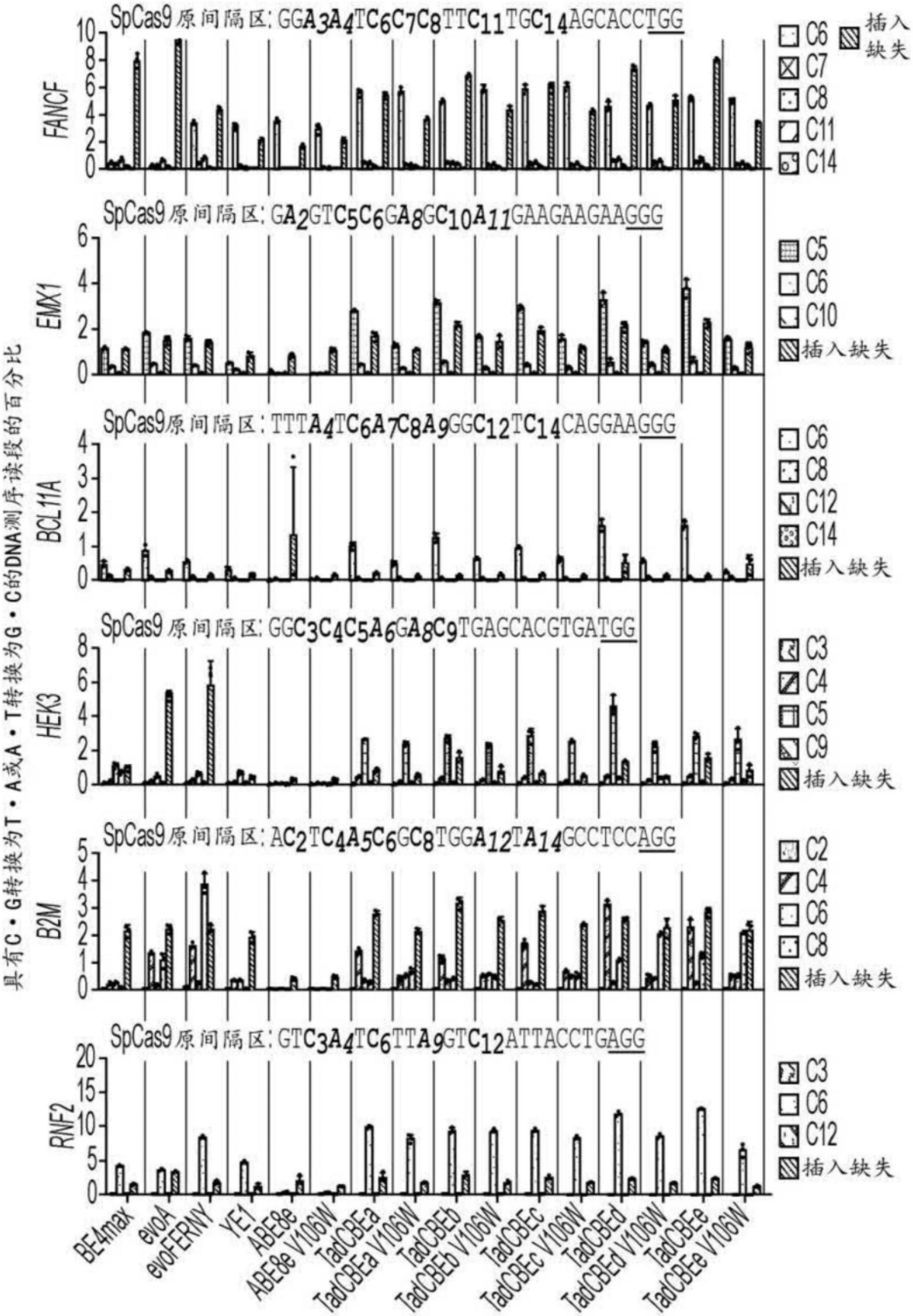

SpCas9原间隔区: GGA3A4TC6C7C8TTC11TGC14AGCACCTGG
FANCF
插入
缺失
C6
C7
C8
C11
C14
SpCas9原间隔区: GA2GTC5C6GA8GC10A11GAAGAAGAAGGG
EMX1
C5
C6
C10
插入缺失
SpCas9原间隔区: TTTA4TC6A7C8A9GGC12TC14CAGGAAGGG
BCL11A
C6
C8
C12
C14
插入缺失
SpCas9原间隔区: GGC3C4C5A6GA8C9TGAGCACGTGATGG
HEK3
C3
C4
C5
C9
插入缺失
SpCas9原间隔区: AC2TC4A5C6GC8TGGA12TA14GCCTCCAGG
B2M
C2
C4
C6
C8
插入缺失
SpCas9原间隔区: GTC3A4TC6TTA9GTC12ATTACCTGAGG
RNF2
C3
C6
C12
插入缺失
具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比
BE4max
evoA
evoFERNY
YE1
ABE8e
ABE8e V106W
TadCBEa
TadCBEa V106W
TadCBEb
TadCBEb V106W
TadCBEc
TadCBEc V106W
TadCBEd
TadCBEd V106W
TadCBEe
TadCBEe V106W

图15

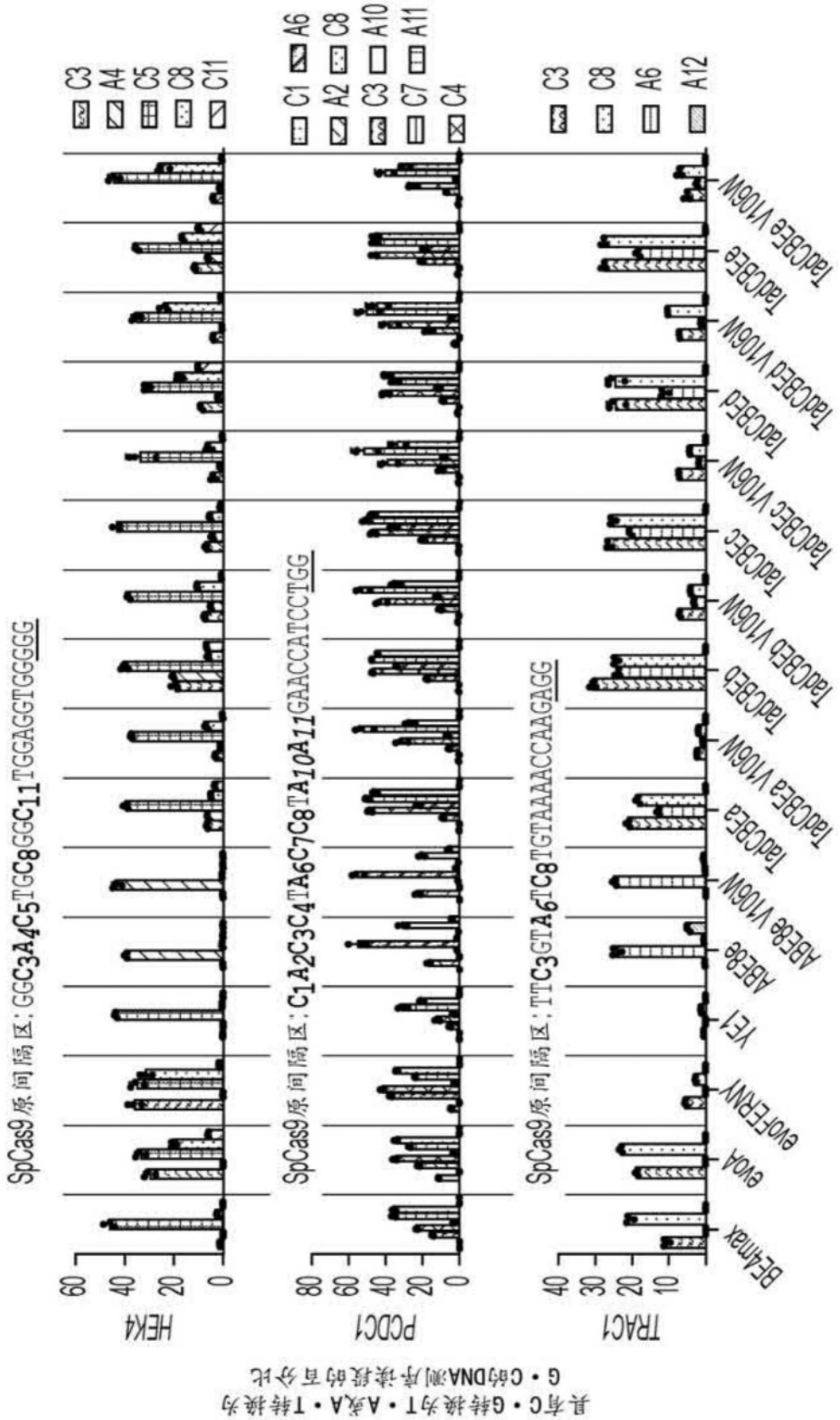

HEK4
PCDC1
TRAC1
具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比
SpCas9原间隔区:GGC3A4C5TGC8GGC11TGGAGGTGGGGG
SpCas9原间隔区:C1A2C3C4TA6C7C8TA10A11GAACCATCCTGG
SpCas9原间隔区:TTC3GTA6TC8TGTAAAACCAAGAGG
C3 A4 C5 C8 C11
C1 A6 A2 C8 C3 A10 C7 A11 C4
C3 C8 A6 A12
BE4max
evoA
evoFERNY
YE1
ABE8e
ABE8e V106W
TadCBEa
TadCBEa V106W
TadCBEb
TadCBEb V106W
TadCBEc
TadCBEc V106W
TadCBEd
TadCBEd V106W
TadCBEe
TadCBEe V106W

图16A

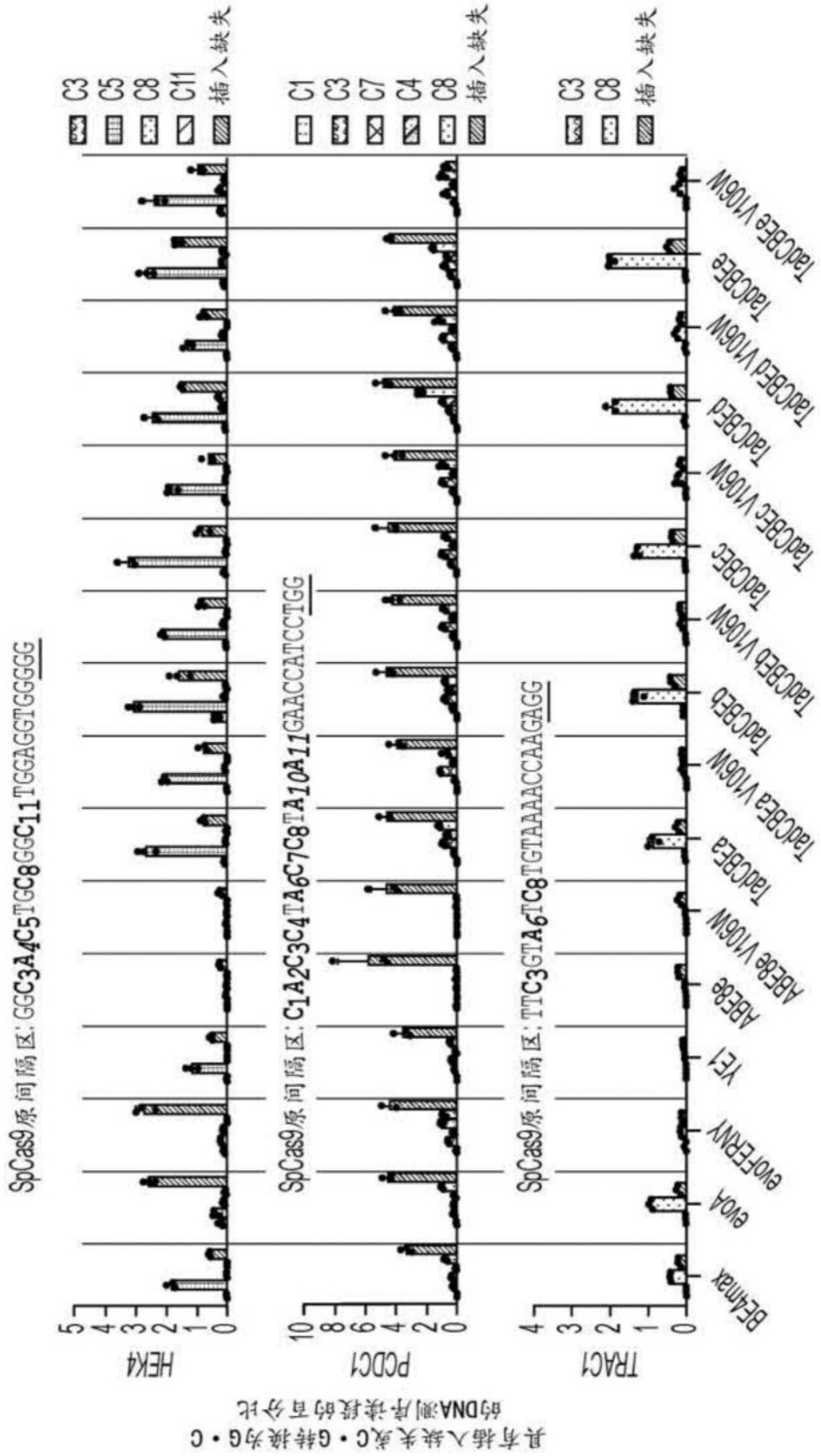

具有插入缺失或C·G转换为G·C的DNA测序读段的百分比
HEK4
PCDC1
TRAC1
SpCas9原间隔区:GGC3A4C5TGC8GGC11TGGAGGTGGGGG
SpCas9原间隔区:C1A2C3C4TA6C7C8TA10A11GAACCATCCTGG
SpCas9原间隔区:TTC3GTA6TC8TGTAAAACCAAGAGG
C3 C5 C8 C11 插入缺失
C1 C3 C7 C4 C8 插入缺失
C3 C8 插入缺失
BE4max
evoA
evoFERNY
VE1
ABE8e
ABE8e V106W
TadCBEa
TadCBEa V106W
TadCBEb
TadCBEb V106W
TadCBEc
TadCBEc V106W
TadCBEd
TadCBEd V106W
TadCBEe
TadCBEe V106W

图16B

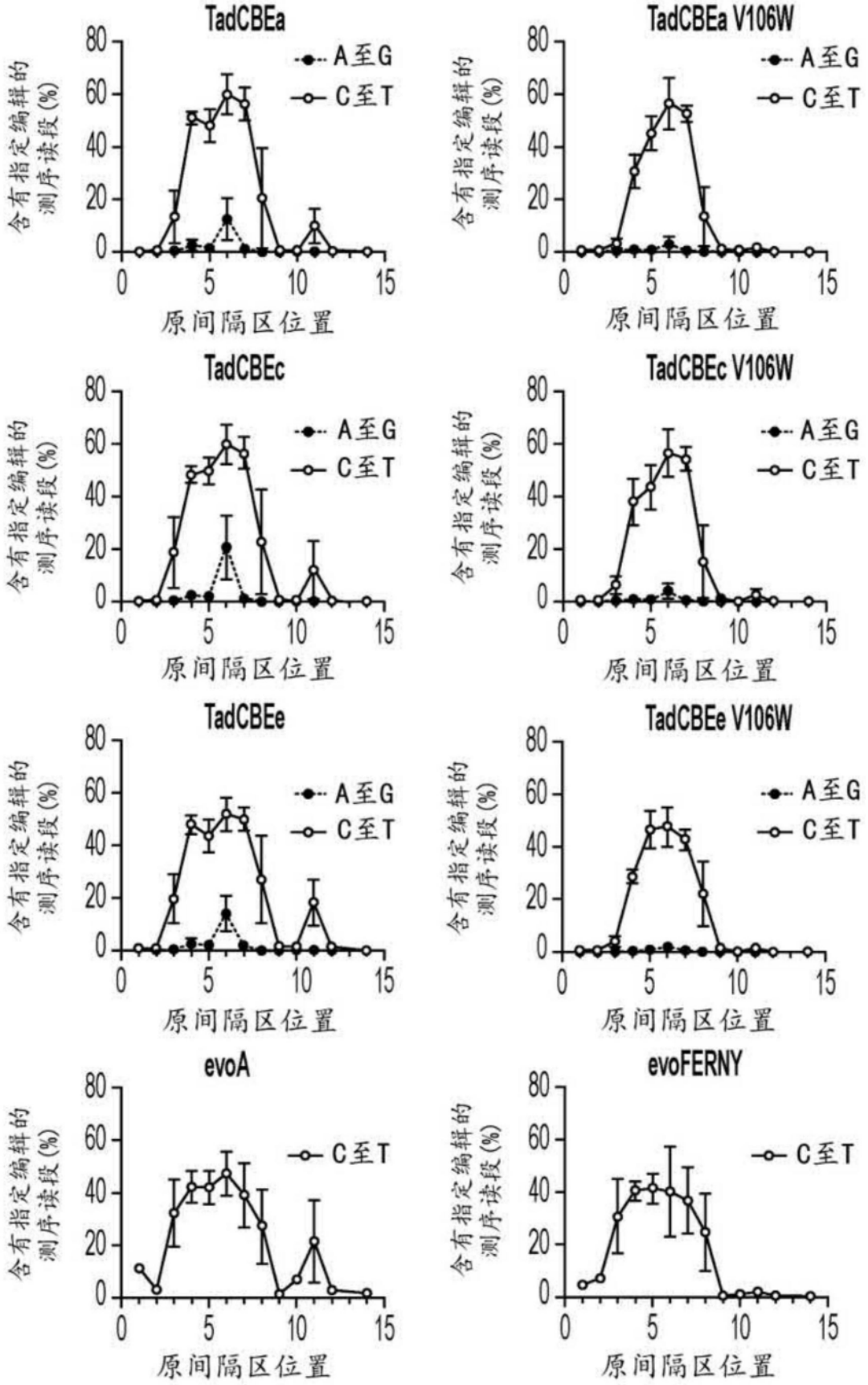

TadCBEa
TadCBEa V106W
TadCBEc
TadCBEc V106W
TadCBEe
TadCBEe V106W
evoA
evoFERNY
A至G
C至T
含有指定编辑的测序读段(%)
原间隔区位置
0
5
10
15
20
40
60
80

图17

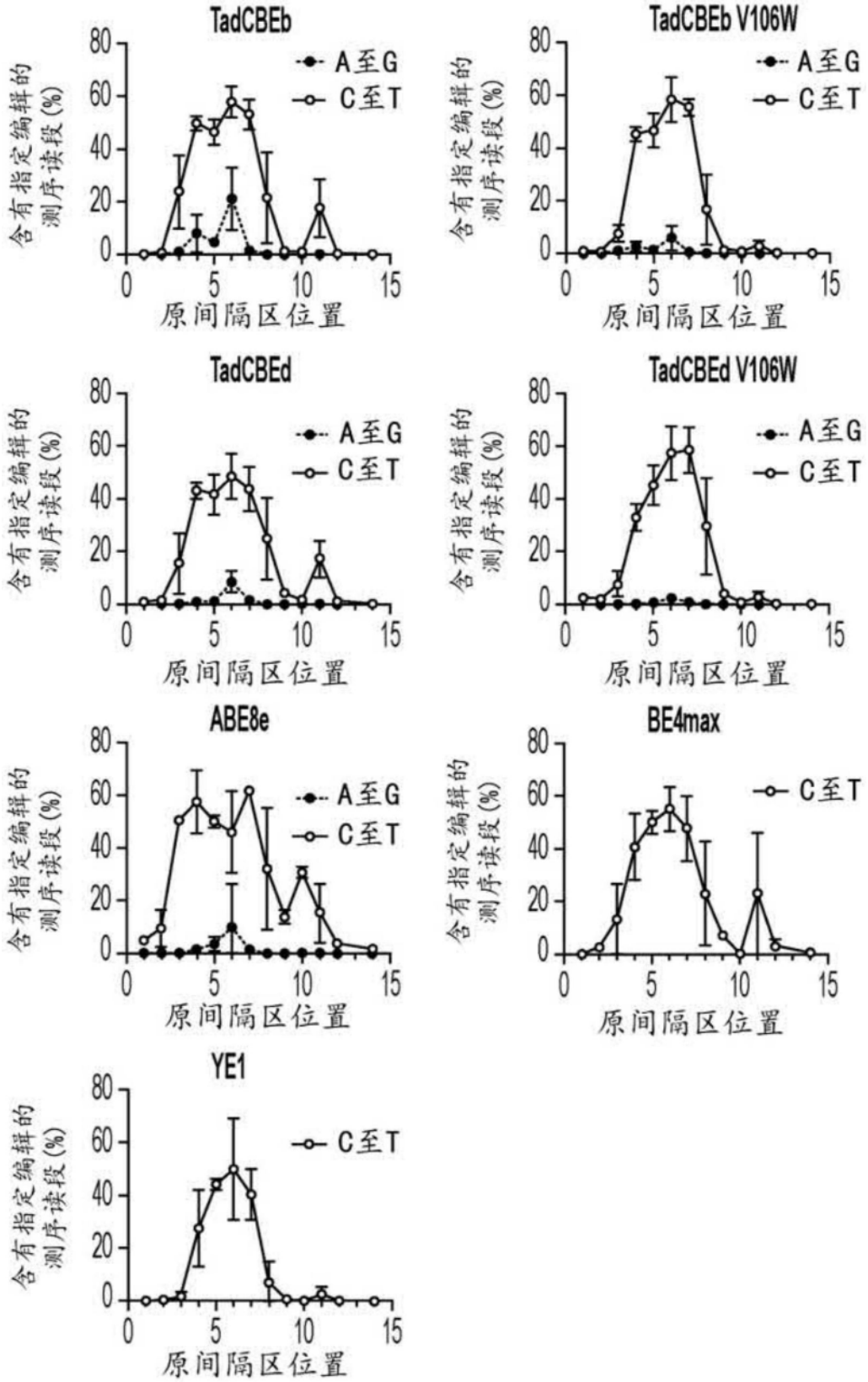

TadCBEb
TadCBEb V106W
TadCBEd
TadCBEd V106W
ABE8e
BE4max
YE1
A至G
C至T
含有指定编辑的测序读段(%)
原间隔区位置
0
20
40
60
80
0
5
10
15

图17(续)

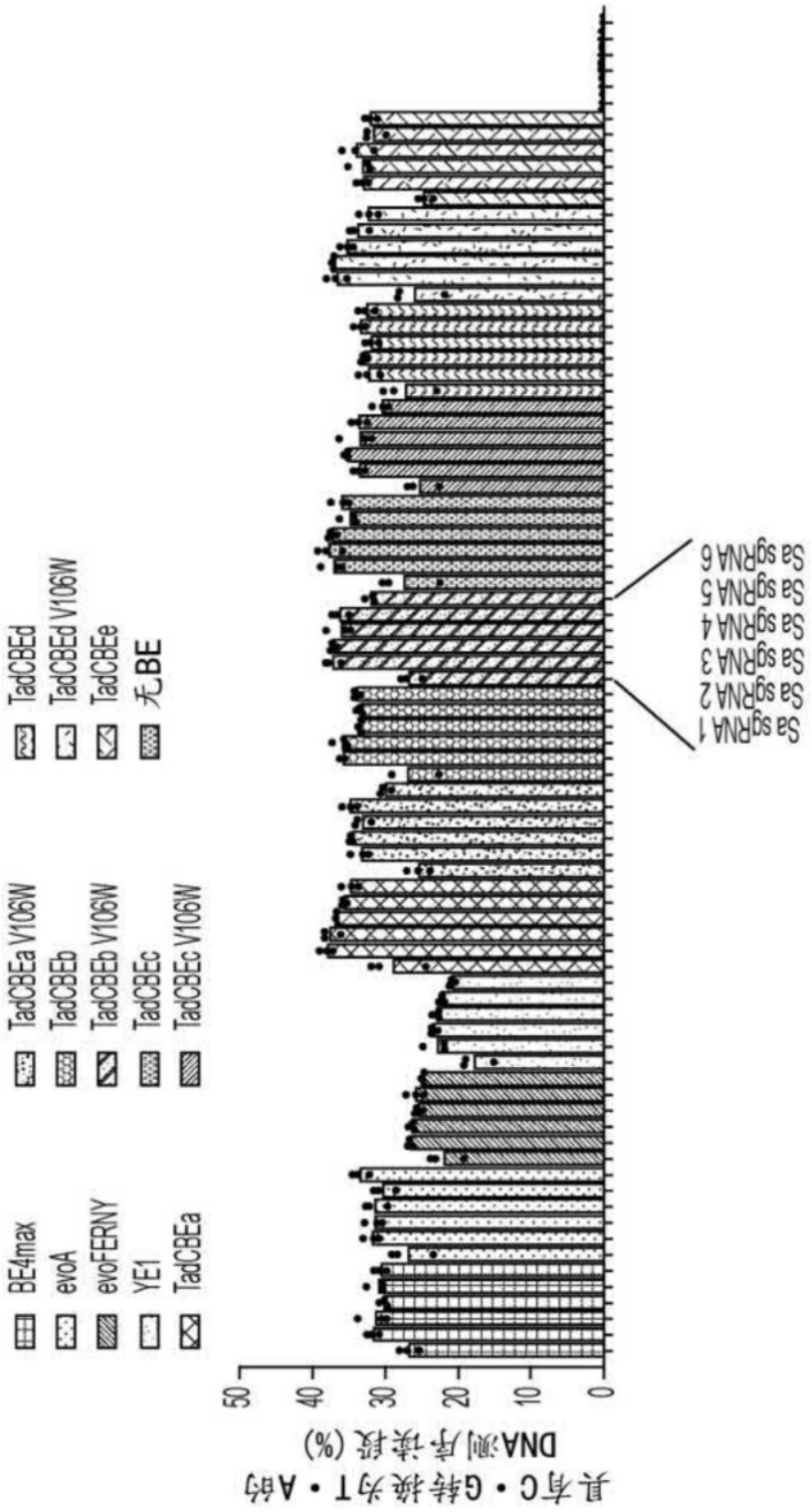

具有C·G转换为T·A的
DNA测序读段(%)
0
10
20
30
40
50
BE4max
evoA
evoFERNY
YE1
TadCBEa
TadCBEa V106W
TadCBEb
TadCBEb V106W
TadCBEc
TadCBEc V106W
TadCBEd
TadCBEd V106W
TadCBEe
无BE
Sa sgRNA 1
Sa sgRNA 2
Sa sgRNA 3
Sa sgRNA 4
Sa sgRNA 5
Sa sgRNA 6

图18

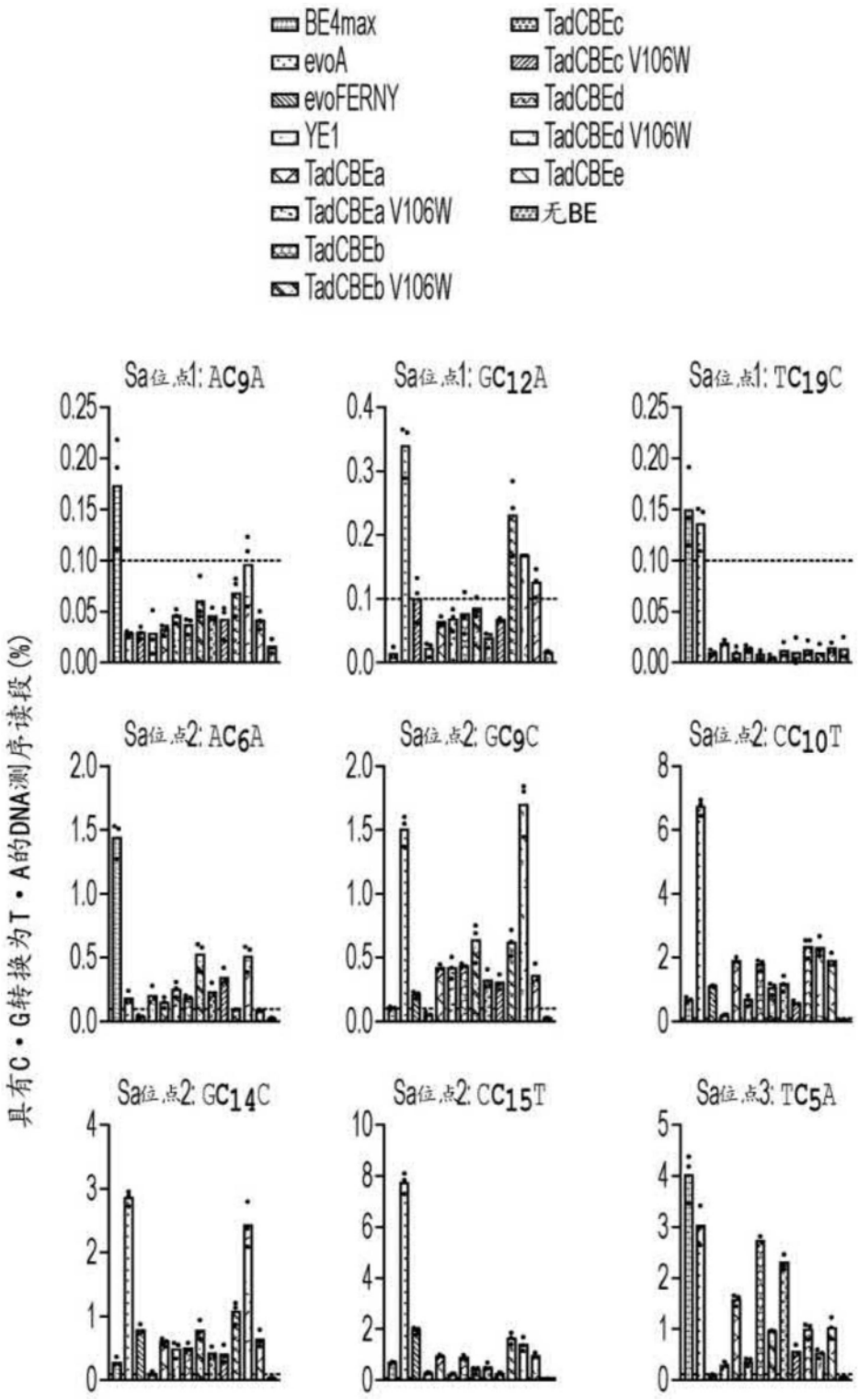

BE4max
evoA
evoFERNY
YE1
TadCBEa
TadCBEa V106W
TadCBEb
TadCBEb V106W
TadCBEc
TadCBEc V106W
TadCBEd
TadCBEd V106W
TadCBEe
无BE
Sa位点1: AC9A
Sa位点1: GC12A
Sa位点1: TC19C
Sa位点2: AC6A
Sa位点2: GC9C
Sa位点2: CC10T
Sa位点2: GC14C
Sa位点2: CC15T
Sa位点3: TC5A
具有C·G转换为T·A的DNA测序读段(%)

图19

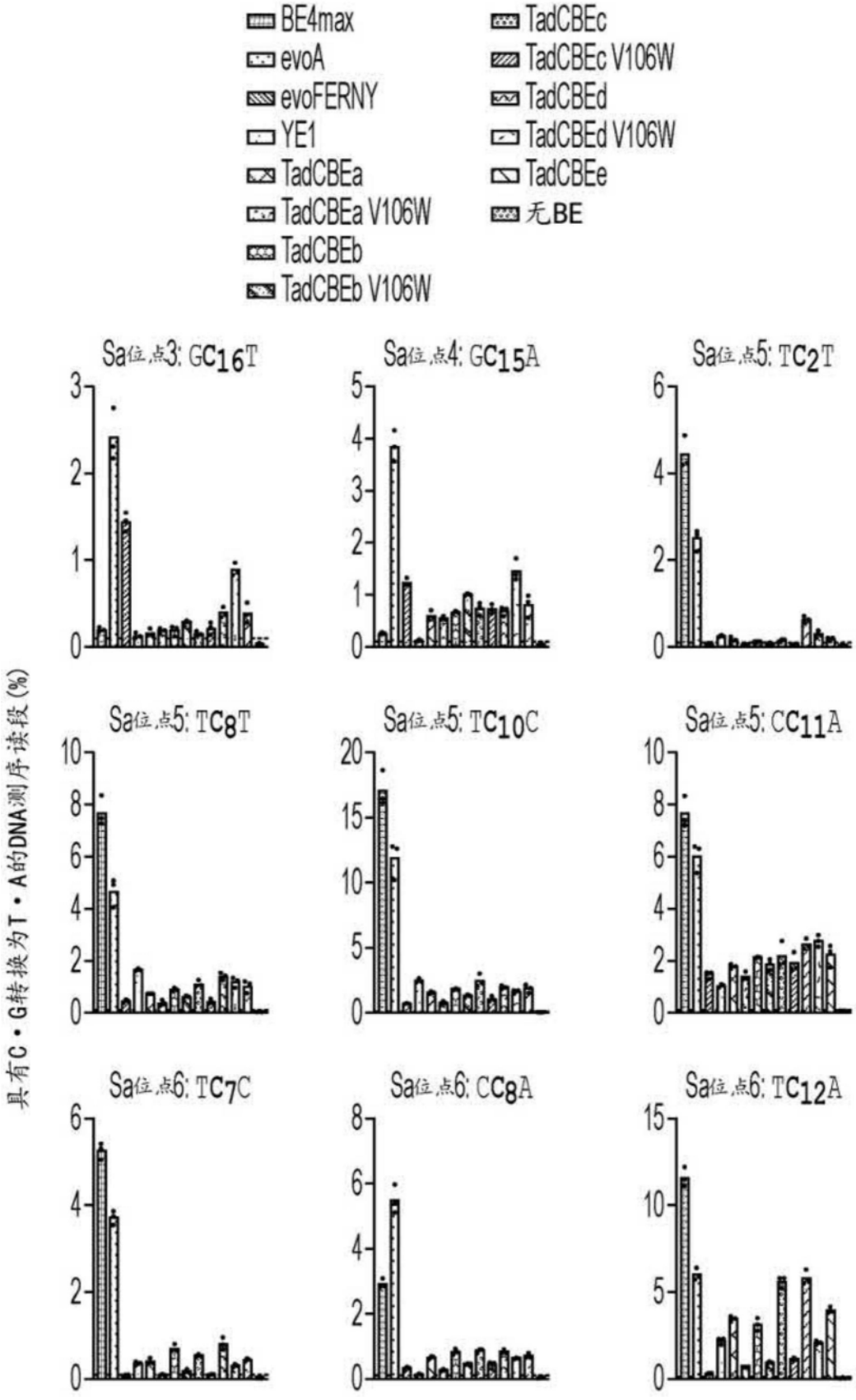

BE4max
evoA
evoFERNY
YE1
TadCBEa
TadCBEa V106W
TadCBEb
TadCBEb V106W
TadCBEc
TadCBEc V106W
TadCBEd
TadCBEd V106W
TadCBEe
无BE
Sa位点3: GC16T
Sa位点4: GC15A
Sa位点5: TC2T
Sa位点5: TC8T
Sa位点5: TC10C
Sa位点5: CC11A
Sa位点6: TC7C
Sa位点6: CC8A
Sa位点6: TC12A
具有C·G转换为T·A的DNA测序读段(%)

图19(续)

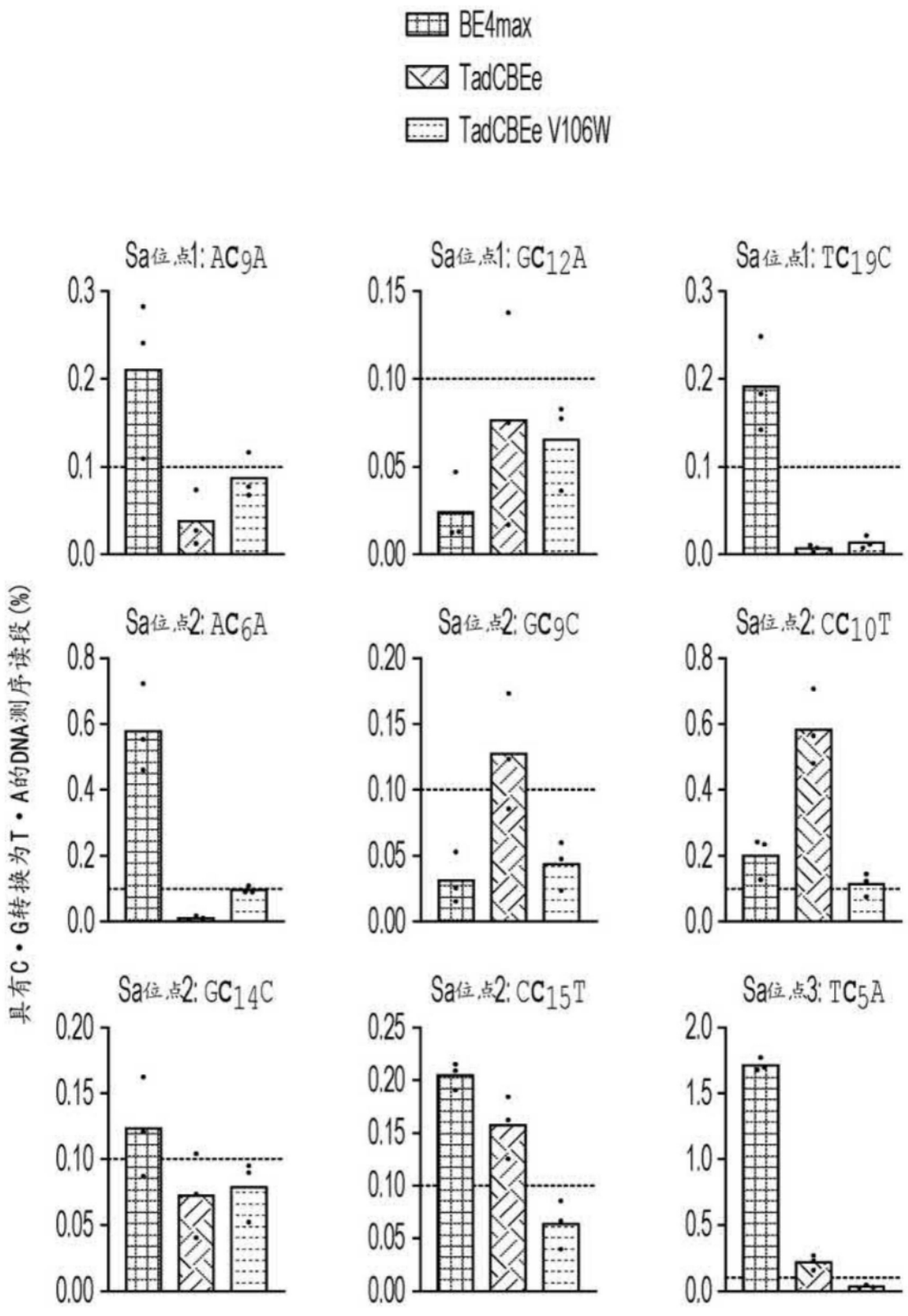

BE4max
TadCBEe
TadCBEe V106W
Sa位点1: AC9A
Sa位点1: GC12A
Sa位点1: TC19C
Sa位点2: AC6A
Sa位点2: GC9C
Sa位点2: CC10T
Sa位点2: GC14C
Sa位点2: CC15T
Sa位点3: TC5A
具有C·G转换为T·A的DNA测序读段(%)

图20

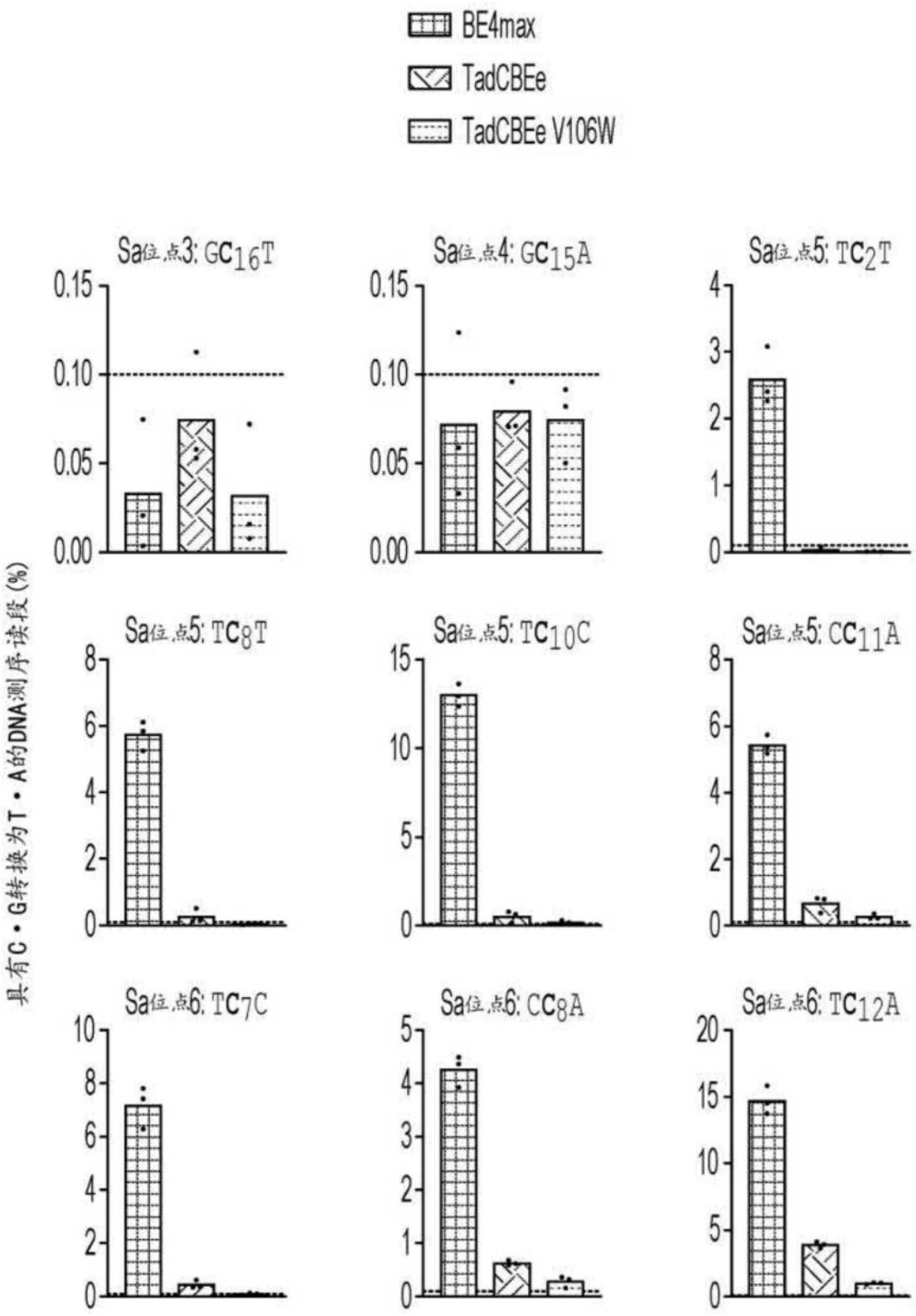

BE4max
TadCBEe
TadCBEe V106W
Sa位点3: GC16T
Sa位点4: GC15A
Sa位点5: TC2T
Sa位点5: TC8T
Sa位点5: TC10C
Sa位点5: CC11A
Sa位点6: TC7C
Sa位点6: CC8A
Sa位点6: TC12A
具有C·G转换为T·A的DNA测序读段(%)

图20(续)

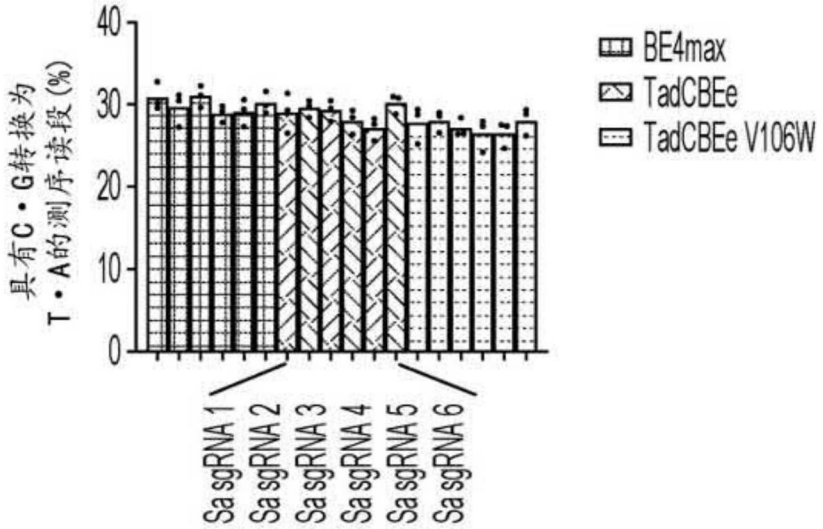


图21A

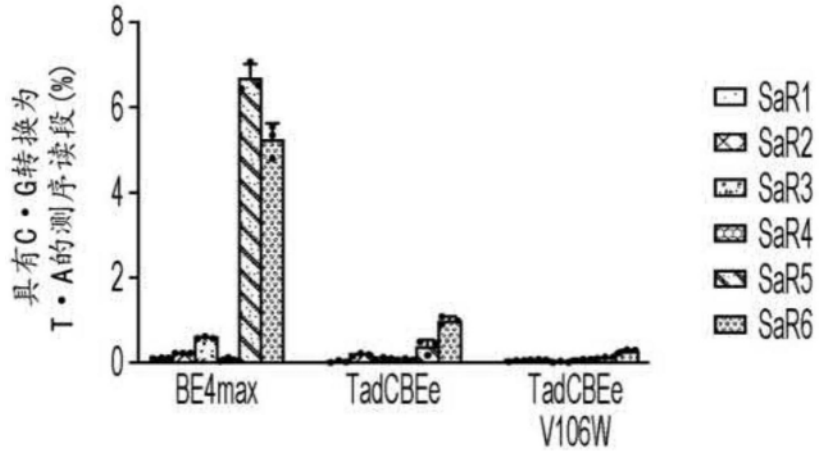


图21B

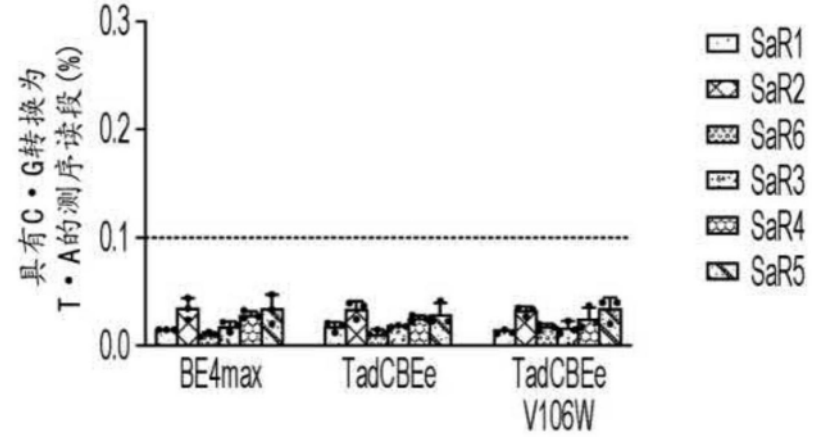


图21C

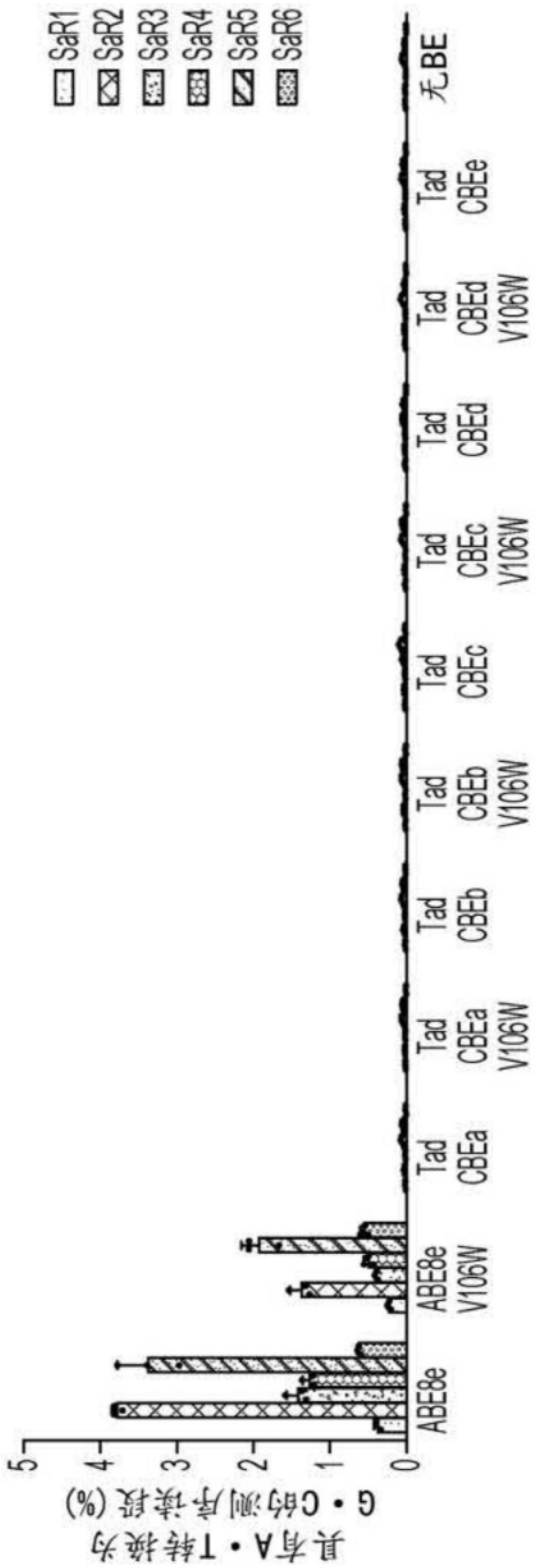

SaR1
SaR2
SaR3
SaR4
SaR5
SaR6
具有A·T转换为G·C的测序读段(%)
0
1
2
3
4
5
ABE8e
ABE8e V106W
Tad CBEa
Tad CBEa V106W
Tad CBEb
Tad CBEb V106W
Tad CBEc
Tad CBEc V106W
Tad CBEd
Tad CBEd V106W
Tad CBEe
无BE

图22

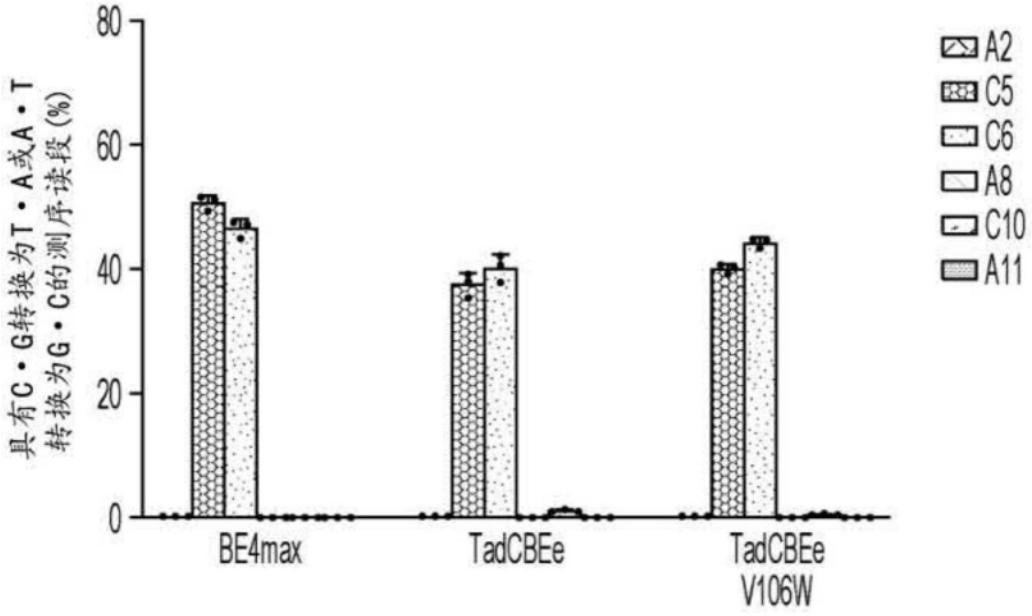

具有C·G转换为T·A或A·T转换为G·C的测序读段(%)
80
60
40
20
0
BE4max
TadCBEe
TadCBEe V106W
A2
C5
C6
A8
C10
A11

图23A

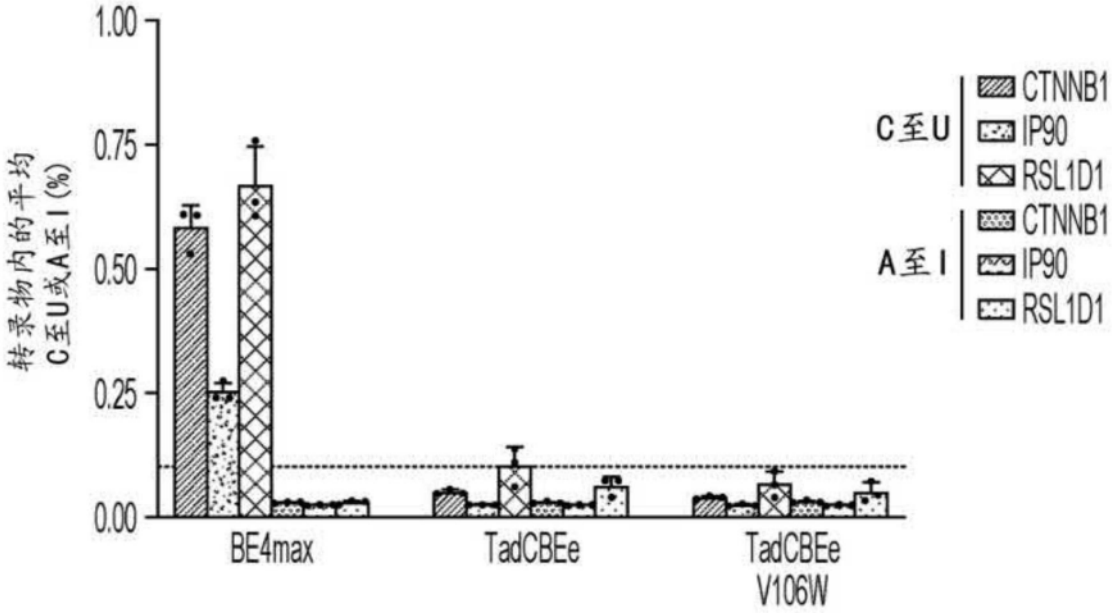

转录物内的平均C至U或A至I(%)
1.00
0.75
0.50
0.25
0.00
BE4max
TadCBEe
TadCBEe V106W
C至U
A至I
CTNNB1
IP90
RSL1D1
CTNNB1
IP90
RSL1D1

图23B

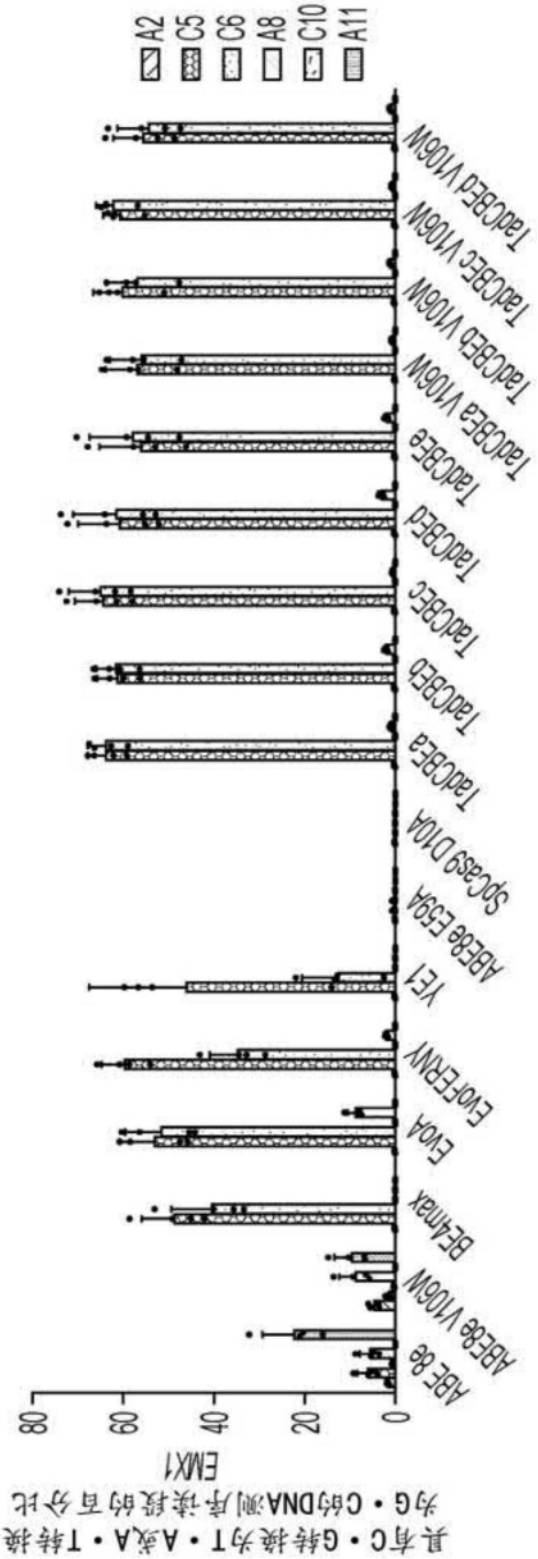

A2
C5
C6
A8
C10
A11
TadCBEd V106W
TadCBEc V106W
TadCBEb V106W
TadCBEa V106W
TadCBEe
TadCBEd
TadCBEc
TadCBEb
TadCBEa
SpCas9 D10A
ABE8e E59A
YE1
EvoFERNY
EvoA
BE4max
ABE8e V106W
ABE8e
80
60
40
20
0
EMX1
具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比

图24

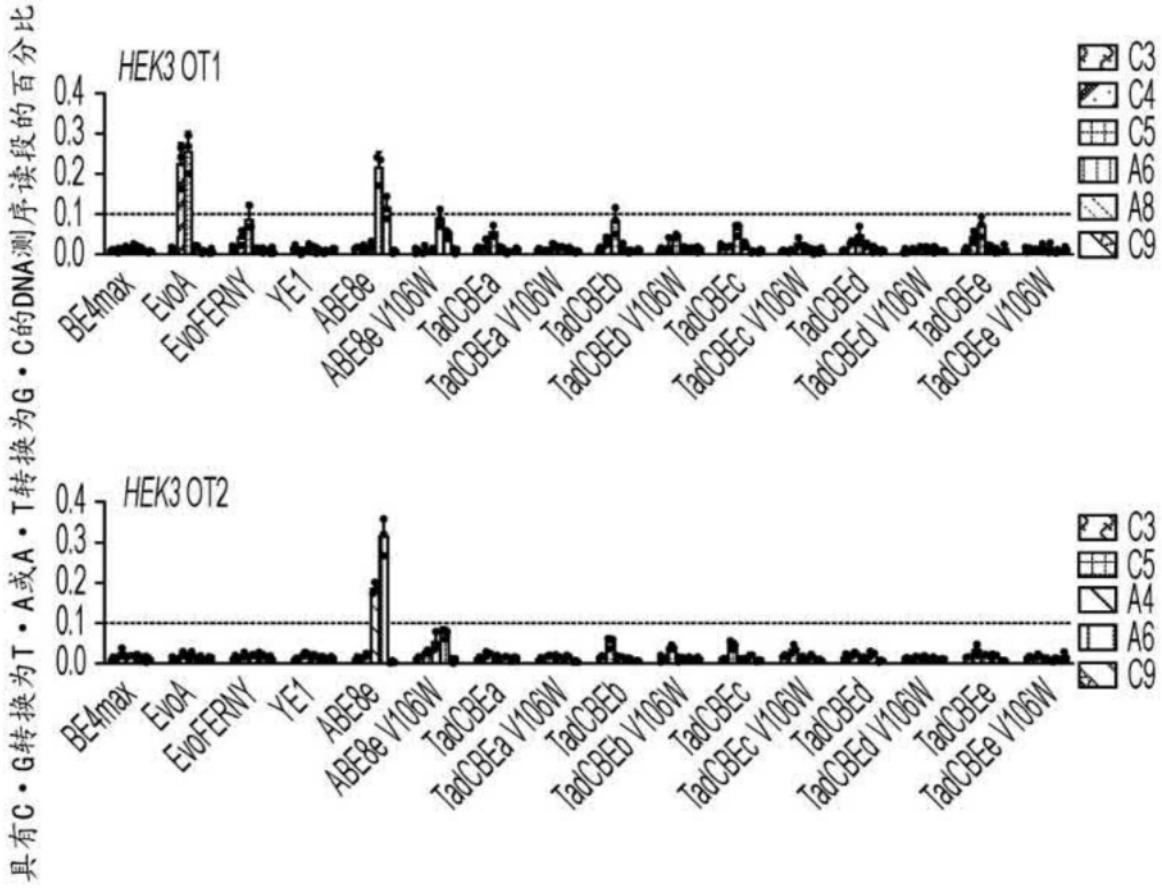

具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比
HEK3 OT1
0.4
0.3
0.2
0.1
0.0
C3
C4
C5
A6
A8
C9
BE4max
EvoA
EvoFERNY
YE1
ABE8e
ABE8e V106W
TadCBEa
TadCBEa V106W
TadCBEb
TadCBEb V106W
TadCBEc
TadCBEc V106W
TadCBEd
TadCBEd V106W
TadCBEe
TadCBEe V106W
HEK3 OT2
0.4
0.3
0.2
0.1
0.0
C3
C5
A4
A6
C9
BE4max
EvoA
EvoFERNY
YE1
ABE8e
ABE8e V106W
TadCBEa
TadCBEa V106W
TadCBEb
TadCBEb V106W
TadCBEc
TadCBEc V106W
TadCBEd
TadCBEd V106W
TadCBEe
TadCBEe V106W

图25

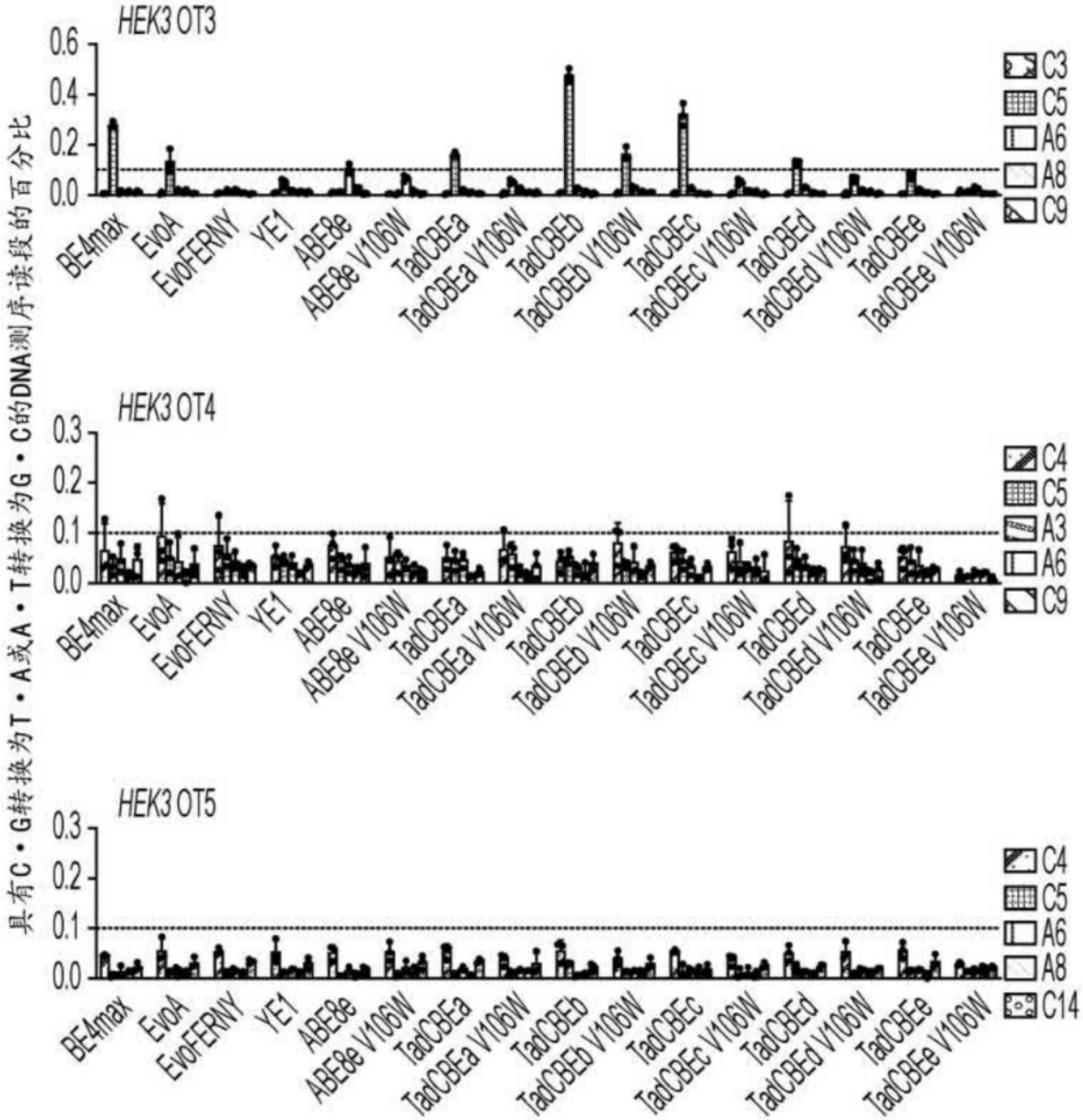

HEK3 OT3
HEK3 OT4
HEK3 OT5
具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比
BE4max
EvoA
EvoFERNY
YE1
ABE8e
ABE8e V106W
TadCBEa
TadCBEa V106W
TadCBEb
TadCBEb V106W
TadCBEc
TadCBEc V106W
TadCBEd
TadCBEd V106W
TadCBEe
TadCBEe V106W
C3
C5
A6
A8
C9
C4
A3
C14

图25(续)

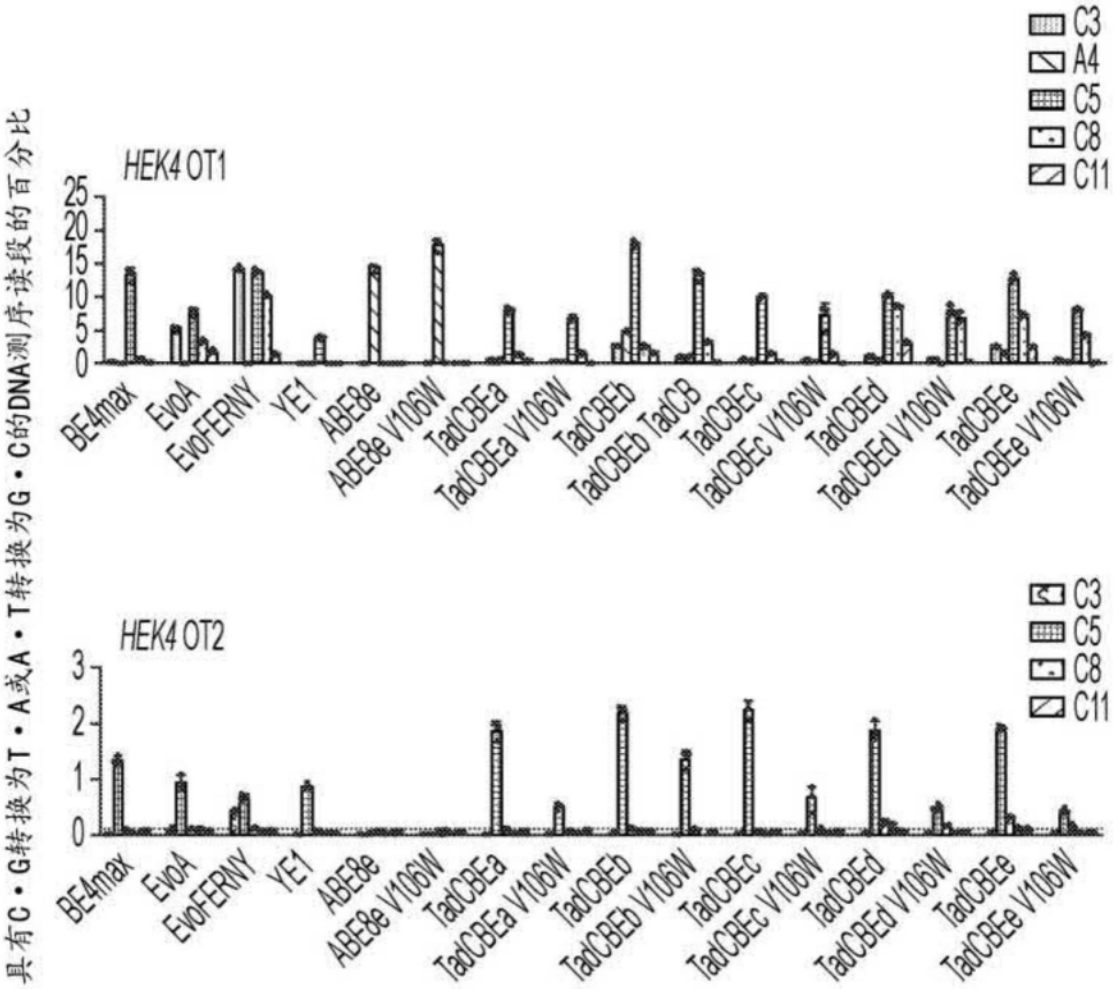

具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比
HEK4 OT1
C3
A4
C5
C8
C11
0
5
10
15
20
25
BE4max
EvoA
EvoFERNY
YE1
ABE8e
ABE8e V106W
TadCBEa
TadCBEa V106W
TadCBEb TadCB
TadCBEc
TadCBEc V106W
TadCBEd
TadCBEd V106W
TadCBEe
TadCBEe V106W
HEK4 OT2
C3
C5
C8
C11
0
1
2
3
BE4max
EvoA
EvoFERNY
YE1
ABE8e
ABE8e V106W
TadCBEa
TadCBEa V106W
TadCBEb
TadCBEb V106W
TadCBEc
TadCBEc V106W
TadCBEd
TadCBEd V106W
TadCBEe
TadCBEe V106W

图26

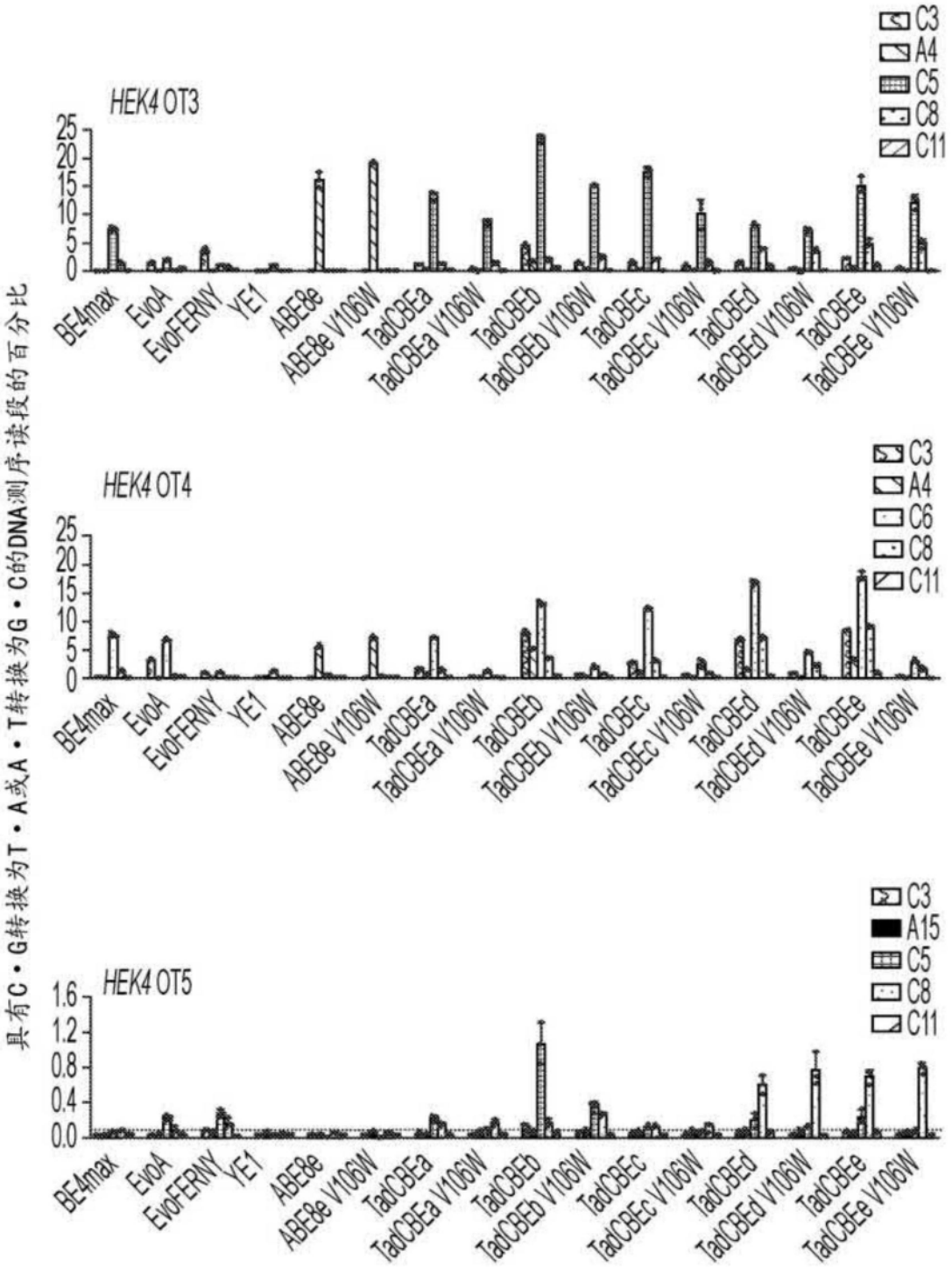

HEK4 OT3
C3
A4
C5
C8
C11
HEK4 OT4
C3
A4
C6
C8
C11
HEK4 OT5
C3
A15
C5
C8
C11
具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比
BE4max
EvoA
EvoFERNY
YE1
ABE8e
ABE8e V106W
TadCBEa
TadCBEa V106W
TadCBEb
TadCBEb V106W
TadCBEc
TadCBEc V106W
TadCBEd
TadCBEd V106W
TadCBEe
TadCBEe V106W

图26(续)

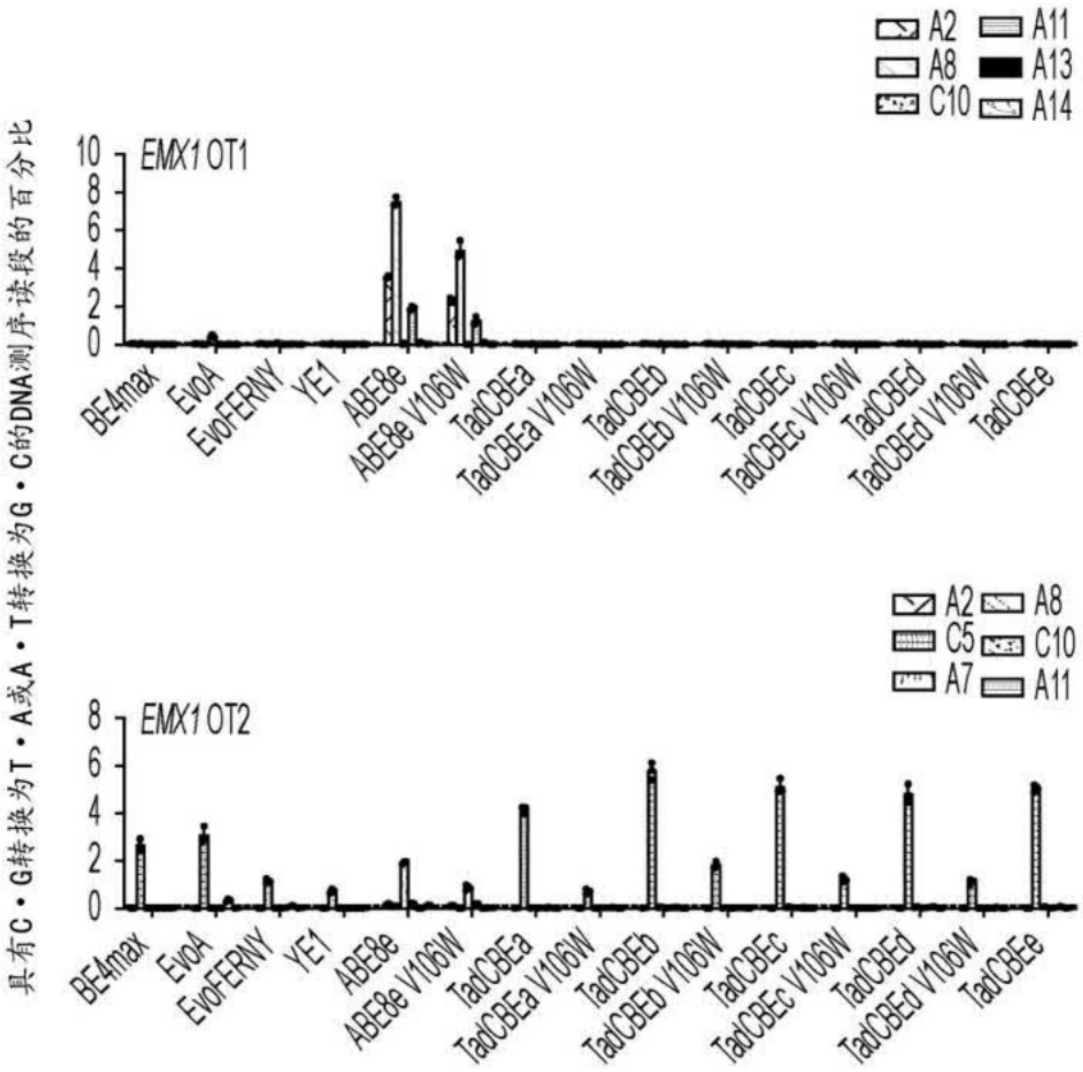

具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比
EMX1 OT1
A2
A11
A8
A13
C10
A14
BE4max
EvoA
EvoFERNY
YE1
ABE8e
ABE8e V106W
TadCBEa
TadCBEa V106W
TadCBEb
TadCBEb V106W
TadCBEc
TadCBEc V106W
TadCBEd
TadCBEd V106W
TadCBEe
EMX1 OT2
A2
A8
C5
C10
A7
A11

图27A

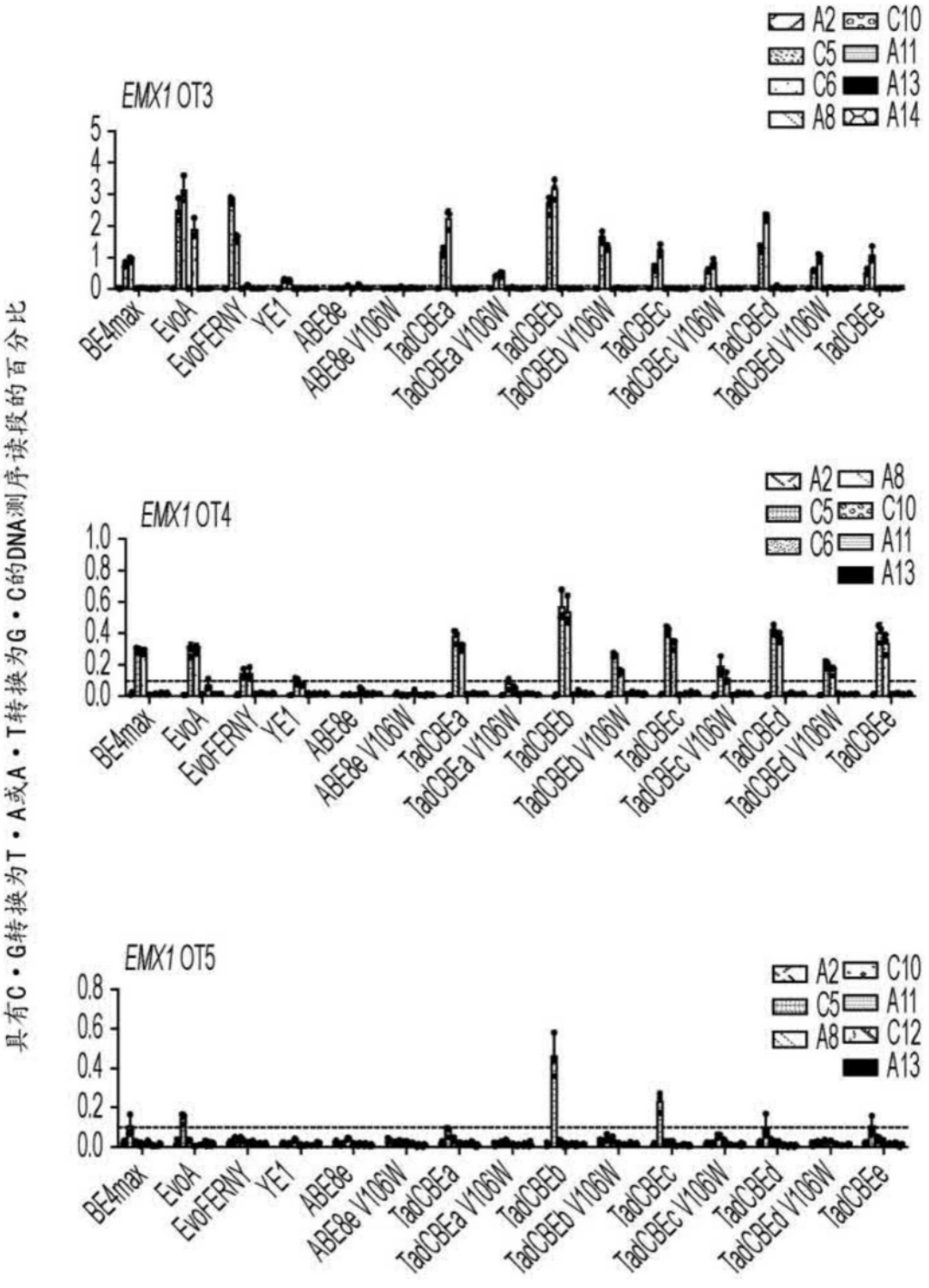

EMX1 OT3
A2
C10
C5
A11
C6
A13
A8
A14
EMX1 OT4
A2
A8
C5
C10
C6
A11
A13
EMX1 OT5
A2
C10
C5
A11
A8
C12
A13
BE4max
EvoA
EvoFERNY
YE1
ABE8e
ABE8e V106W
TadCBEa
TadCBEa V106W
TadCBEb
TadCBEb V106W
TadCBEc
TadCBEc V106W
TadCBEd
TadCBEd V106W
TadCBEe
具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比

图27A(续)

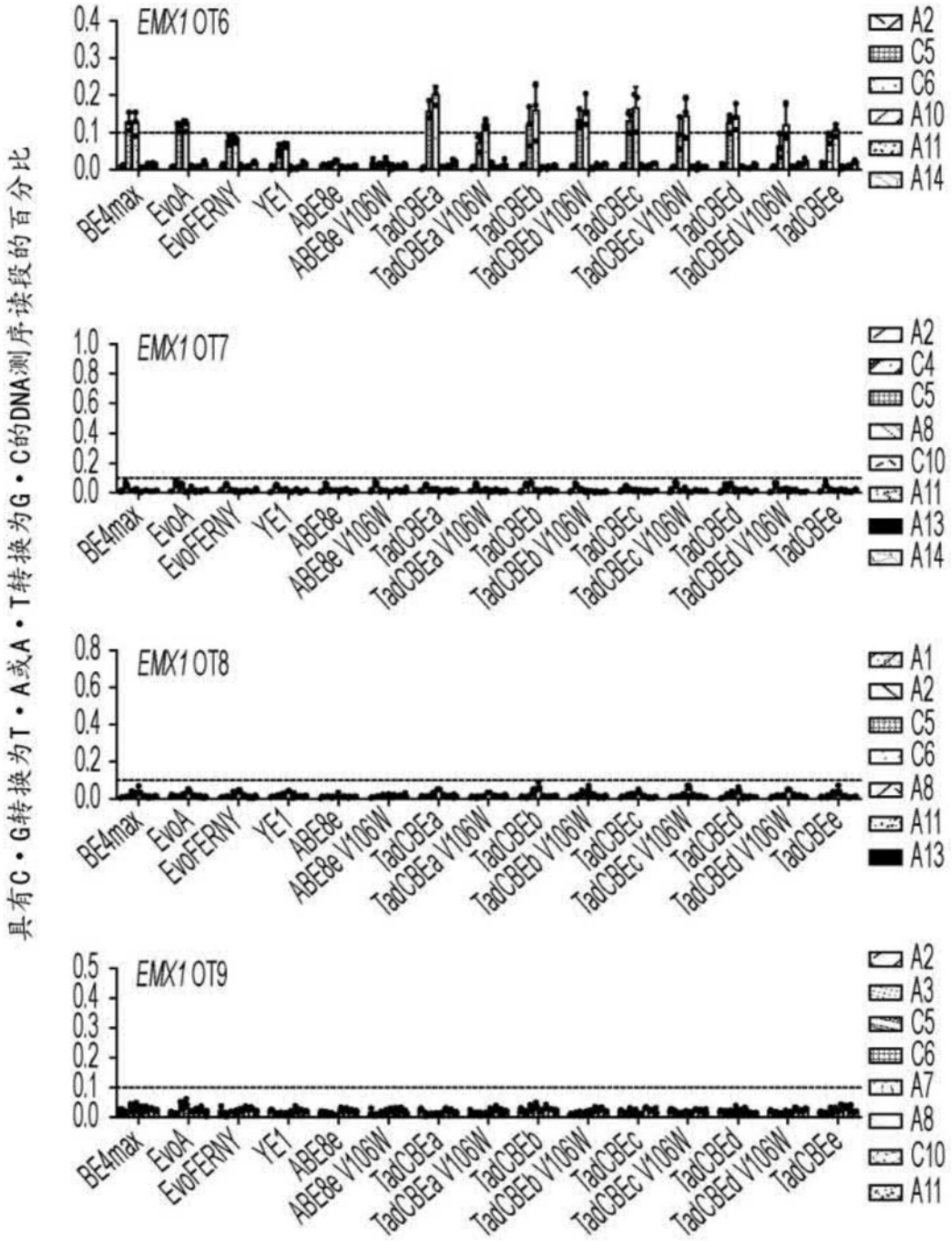

具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比
EMX1 OT6
EMX1 OT7
EMX1 OT8
EMX1 OT9
BE4max
EvoA
EvoFERNY
YE1
ABE8e
ABE8e V106W
TadCBEa
TadCBEa V106W
TadCBEb
TadCBEb V106W
TadCBEc
TadCBEc V106W
TadCBEd
TadCBEd V106W
TadCBEe

图27B

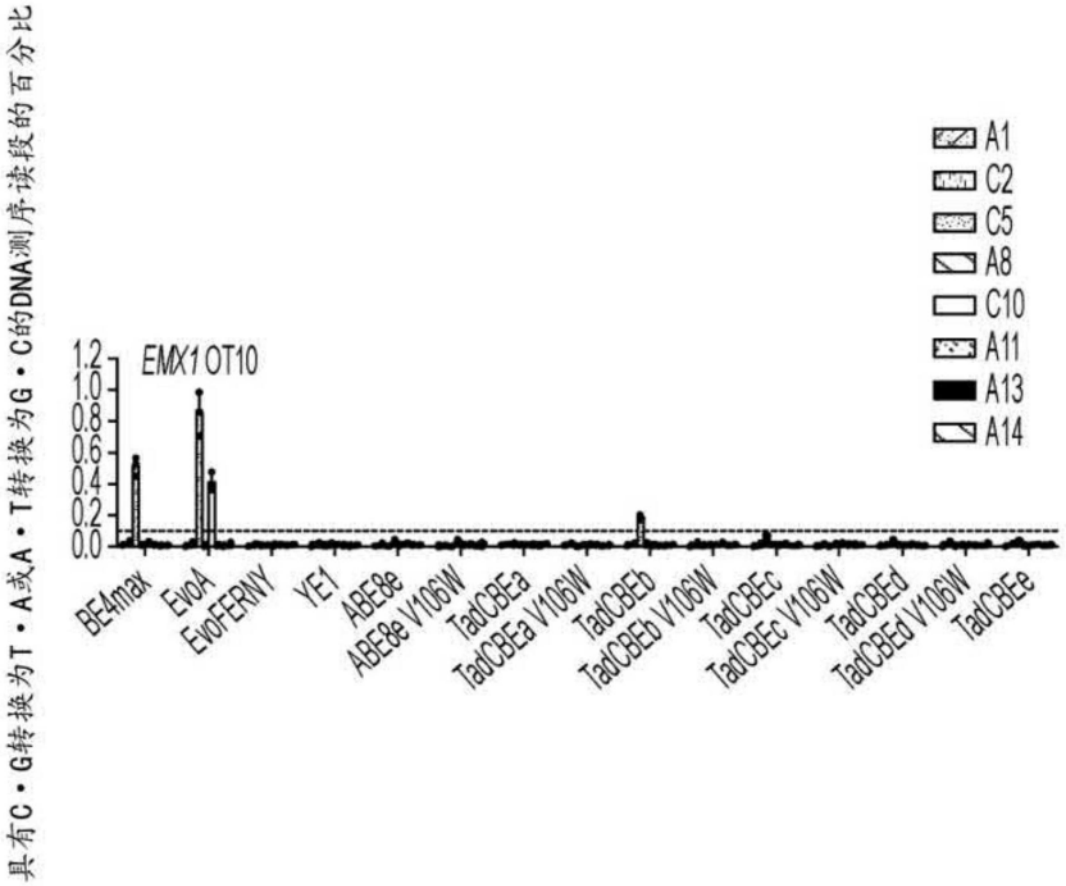

具有C・G转换为T・A或A・T转换为G・C的DNA测序读段的百分比
EMX1 OT10
1.2
1.0
0.8
0.6
0.4
0.2
0.0
A1
C2
C5
A8
C10
A11
A13
A14
BE4max
EvoA
EvoFERNY
YE1
ABE8e
ABE8e V106W
TadCBEa
TadCBEa V106W
TadCBEb
TadCBEb V106W
TadCBEc
TadCBEc V106W
TadCBEd
TadCBEd V106W
TadCBEe

图27B(续)

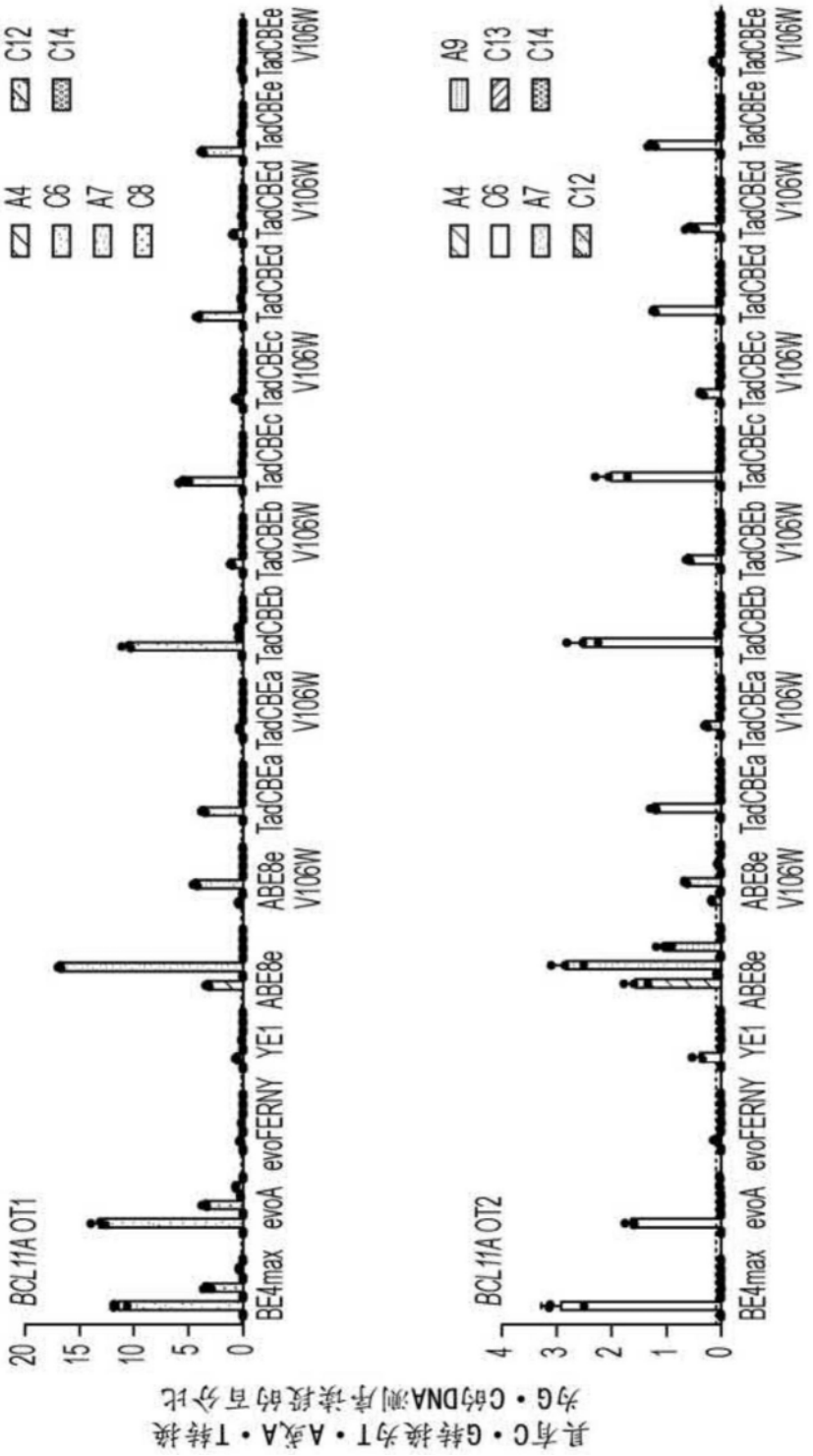

具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比
BCL11A OT1
BCL11A OT2
BE4max
evoA
evoFERNY
YE1
ABE8e
ABE8e V106W
TadCBEa V106W
TadCBEb V106W
TadCBEc V106W
TadCBEd V106W
TadCBEe V106W
A4
C6
A7
C8
C12
C14
A9
C13

图28

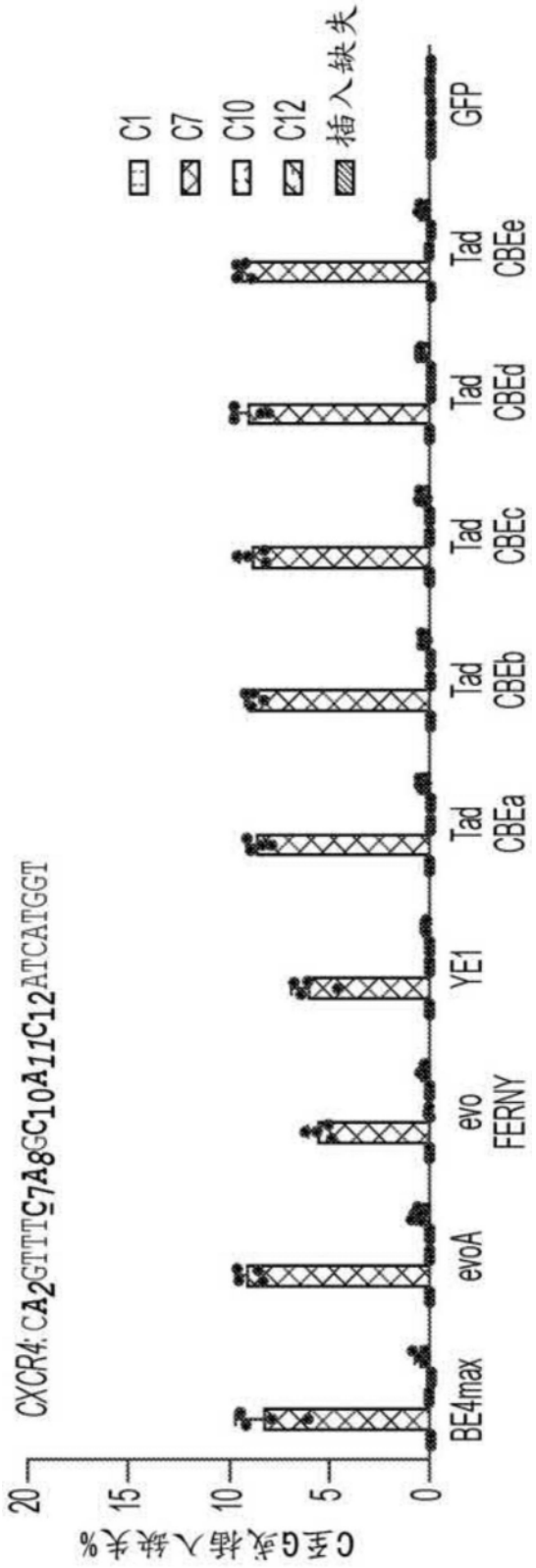

C1
C7
C10
C12
插入缺失
C至G或插入缺失%
20
15
10
5
0
CXCR4: CA2GTTTC7A8GC10A11C12ATCATGGT
BE4max
evoA
evo FERNY
YE1
Tad CBEa
Tad CBEb
Tad CBEc
Tad CBEd
Tad CBEe
GFP

图29A

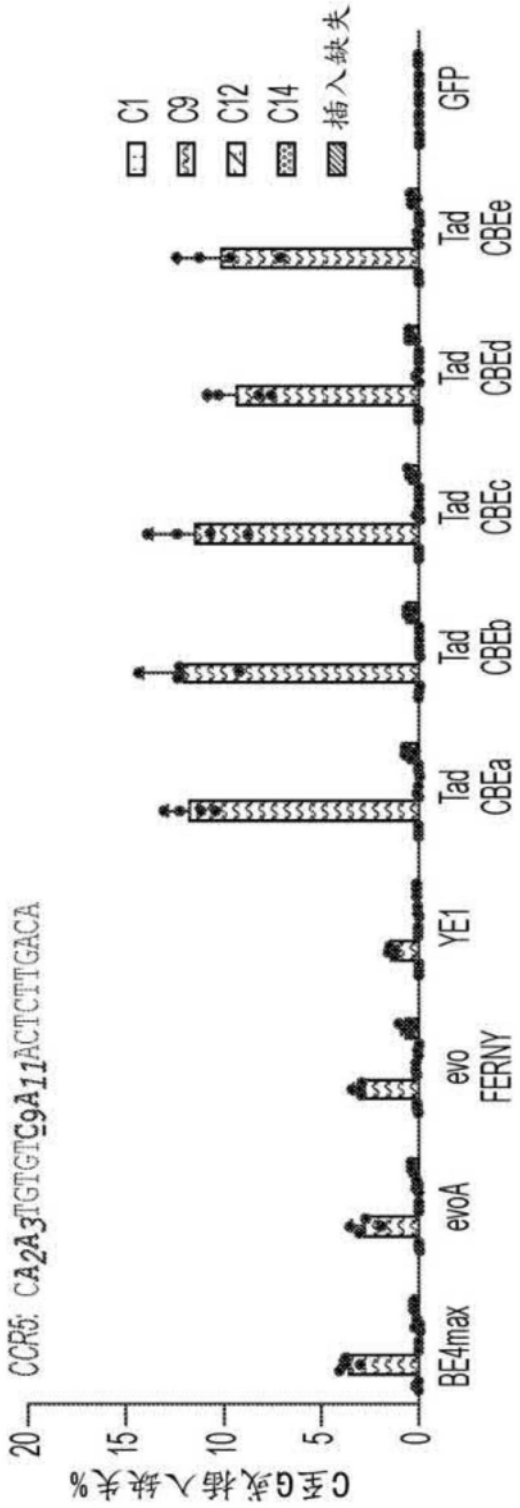

C至G或插入缺失%
20
15
10
5
0
CCR5: CA2A3TGTGTC9A11ACTCTTGACA
BE4max
evoA
evo FERNY
YE1
Tad CBEa
Tad CBEb
Tad CBEc
Tad CBEd
Tad CBEe
GFP
C1
C9
C12
C14
插入缺失

图29B

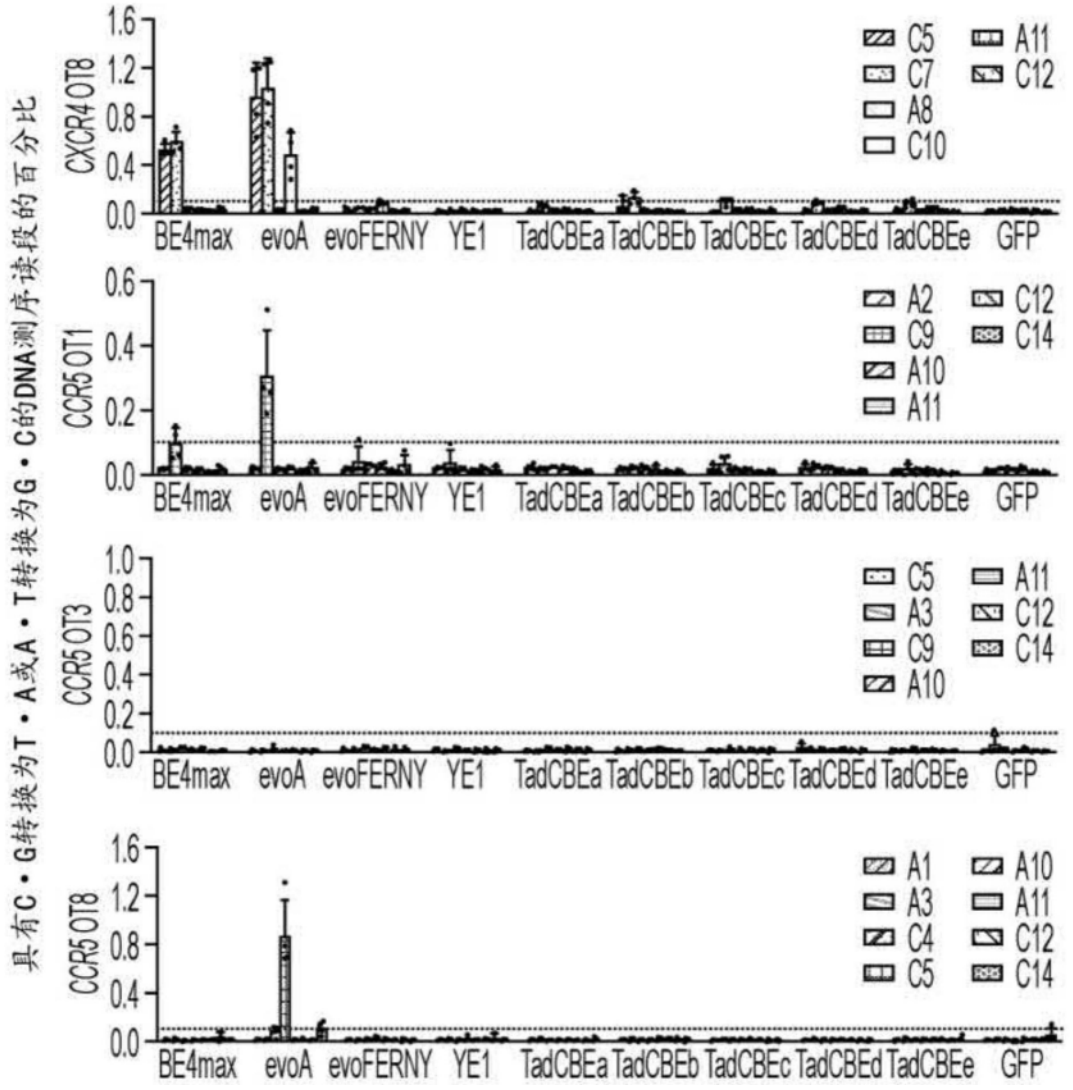

具有C・G转换为T・A或A・T转换为G・C的DNA测序读段的百分比
CXCR4 OT8
C5
A11
C7
C12
A8
C10
CCR5 OT1
A2
C12
C9
C14
A10
A11
CCR5 OT3
C5
A11
A3
C12
C9
C14
A10
CCR5 OT8
A1
A10
A3
A11
C4
C12
C5
C14
BE4max
evoA
evoFERNY
YE1
TadCBEa
TadCBEb
TadCBEc
TadCBEd
TadCBEe
GFP

图30

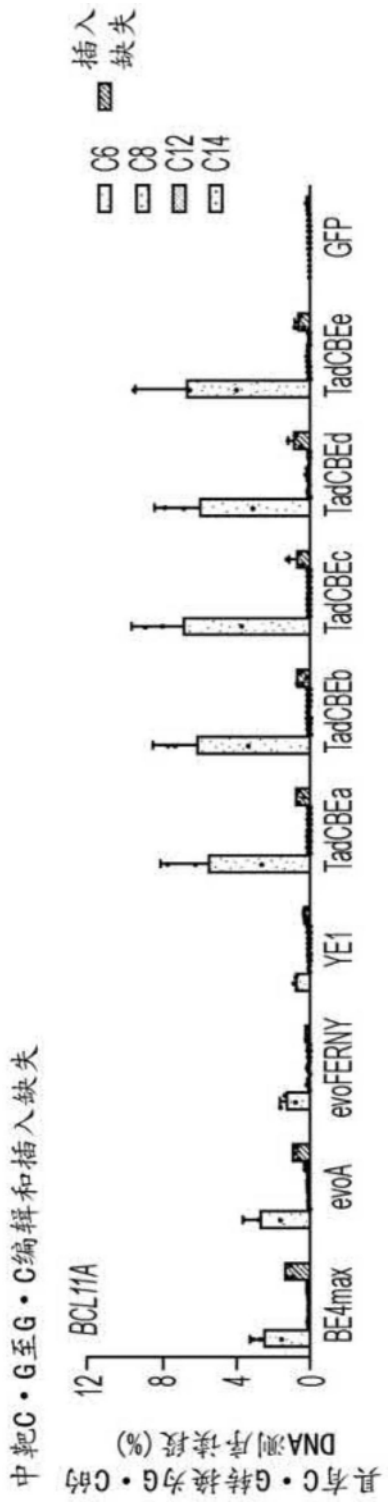

单布C·G转换为G·C的中靶C·G至G·C编辑和插入缺失
DNA测序读段(%)
12
8
4
0
BCL11A
BE4max
evoA
evoFERNY
YE1
TadCBEa
TadCBEb
TadCBEc
TadCBEd
TadCBEe
GFP
C6
C8
C12
C14
插入
缺失

图31A

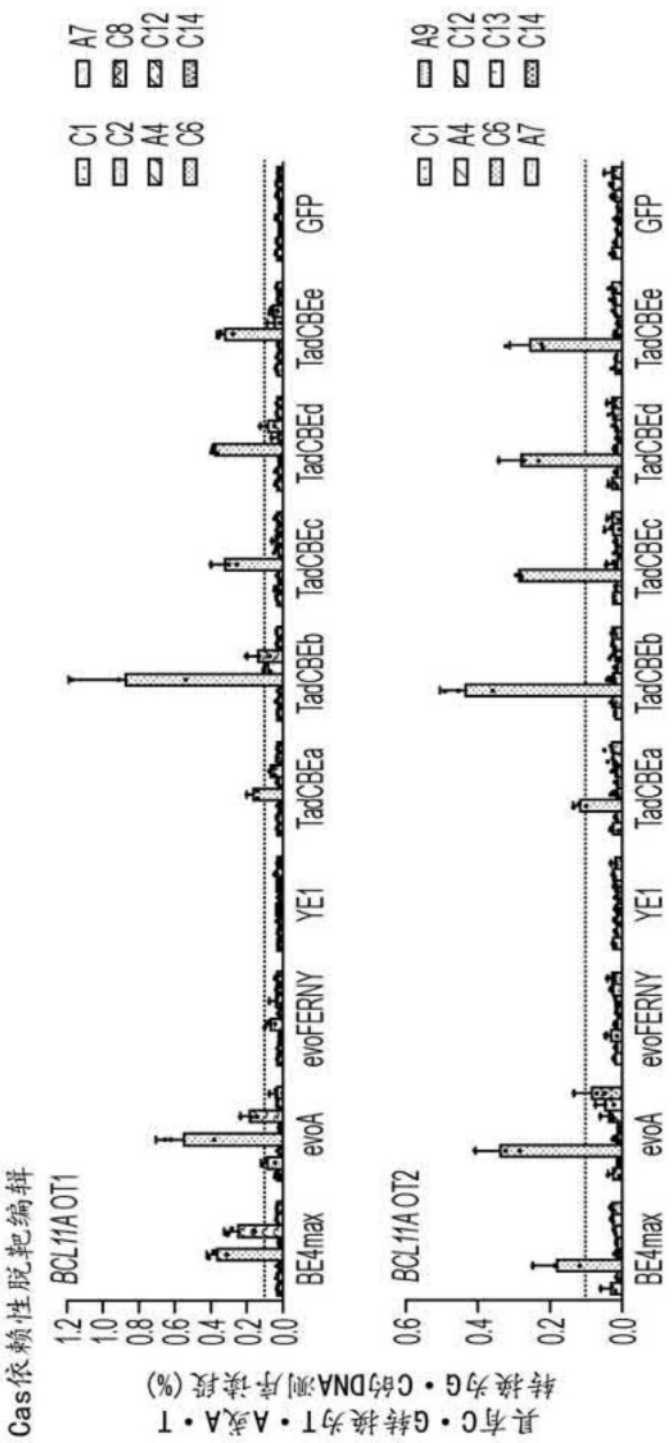

Cas依赖性脱靶编辑
具有C·G转换为T·A或A·T转换为G·C的DNA测序读段(%)
BCL11A OT1
BCL11A OT2
C1
C2
A4
C6
A7
C8
C12
C14
A9
C13
BE4max
evoA
evoFERNY
YE1
TadCBEa
TadCBEb
TadCBEc
TadCBEd
TadCBEe
GFP
0.0
0.2
0.4
0.6
0.8
1.0
1.2

图31B

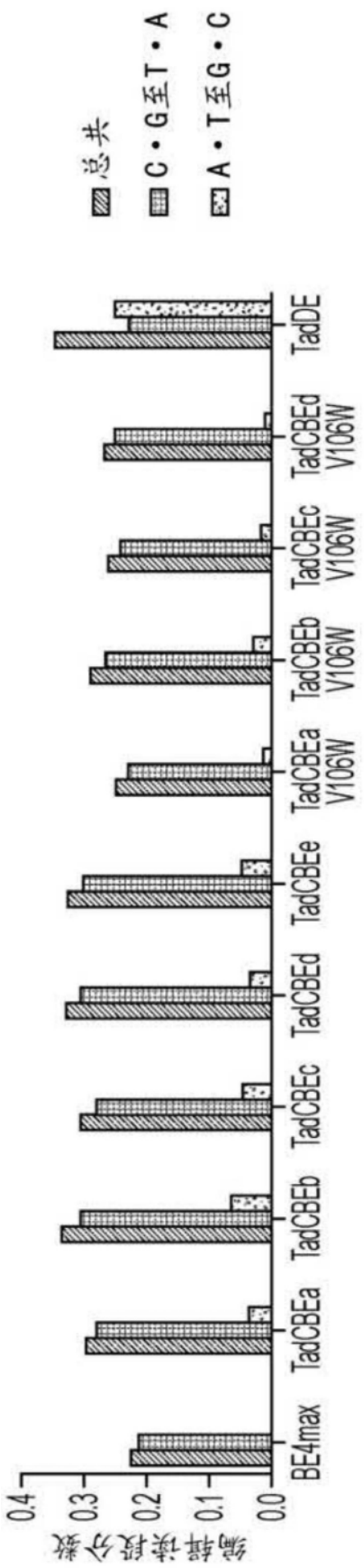

总共
C•G至T•A
A•T至G•C
编辑读段分数
0.0
0.1
0.2
0.3
0.4
BE4max
TadCBEa
TadCBEb
TadCBEc
TadCBEd
TadCBEe
TadCBEa V106W
TadCBEb V106W
TadCBEc V106W
TadCBEd V106W
TadDE

图32A

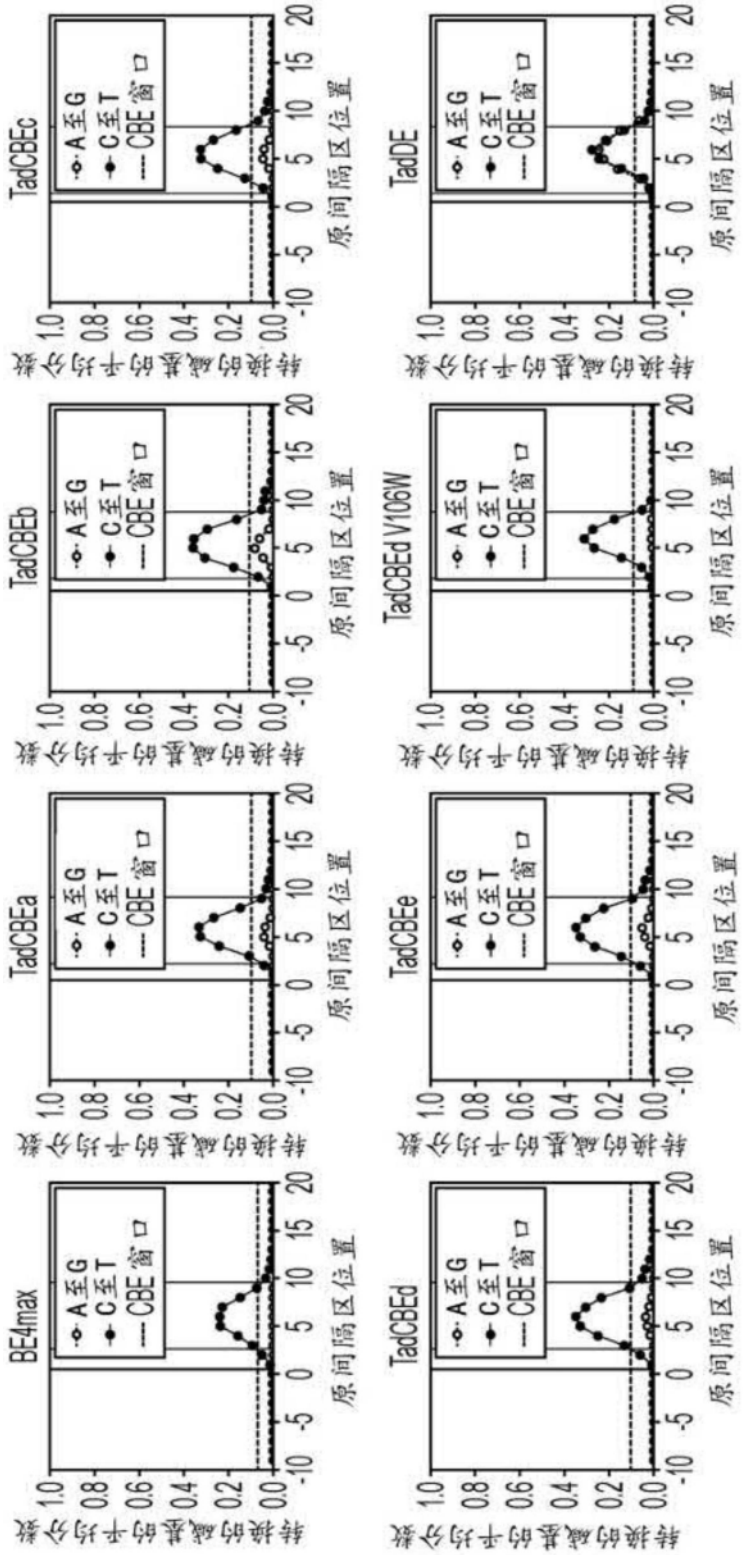

BE4max
TadCBEa
TadCBEb
TadCBEc
TadCBEd
TadCBEe
TadCBEd V106W
TadDE
A至G
C至T
CBE窗口
转换的碱基的平均分数
原间隔区位置

图32B

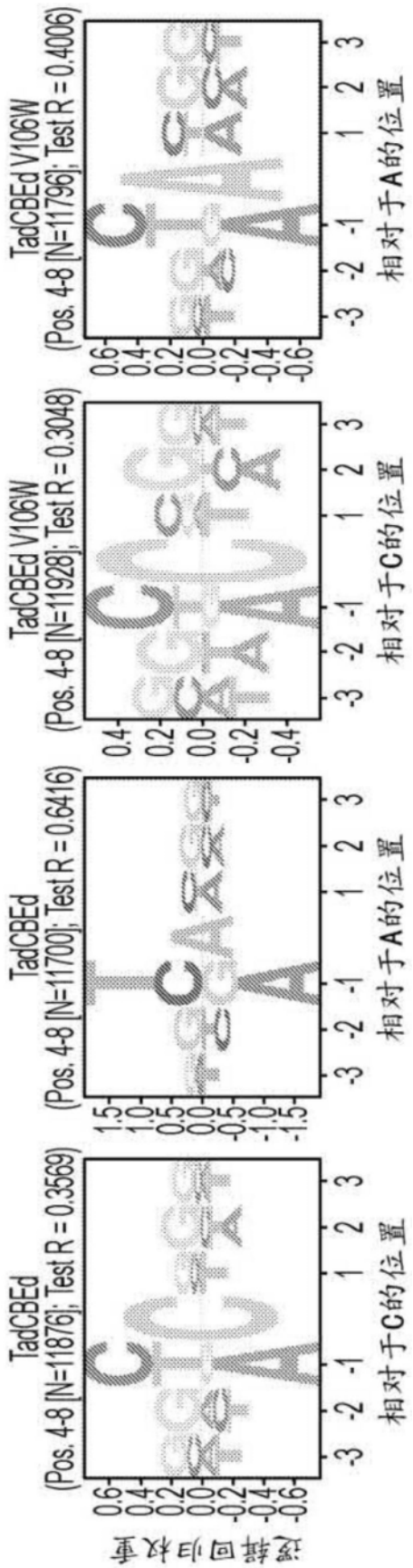

TadCBEd
(Pos. 4-8 [N=11876]; Test R = 0.3569)
TadCBEd
(Pos. 4-8 [N=11700]; Test R = 0.6416)
TadCBEd V106W
(Pos. 4-8 [N=11928]; Test R = 0.3048)
TadCBEd V106W
(Pos. 4-8 [N=11796]; Test R = 0.4006)
逻辑回归权重
相对于C的位置
相对于A的位置
相对于C的位置
相对于A的位置

图32C

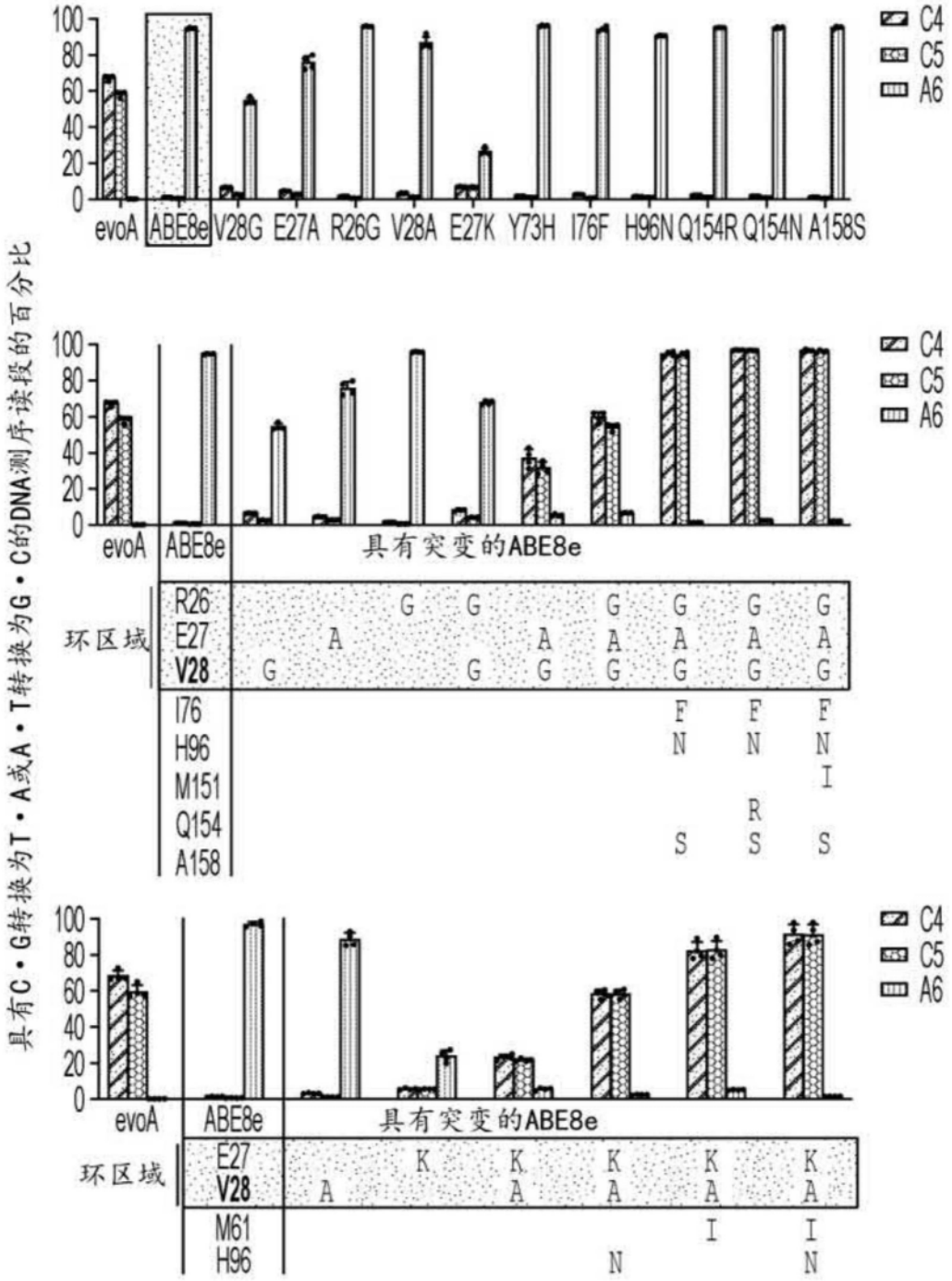

具有C•G转换为T•A或A•T转换为G•C的DNA测序读段的百分比
C4
C5
A6
evoA
ABE8e
V28G
E27A
R26G
V28A
E27K
Y73H
I76F
H96N
Q154R
Q154N
A158S
具有突变的ABE8e
环区域
R26
E27
V28
I76
H96
M151
Q154
A158
M61

图33

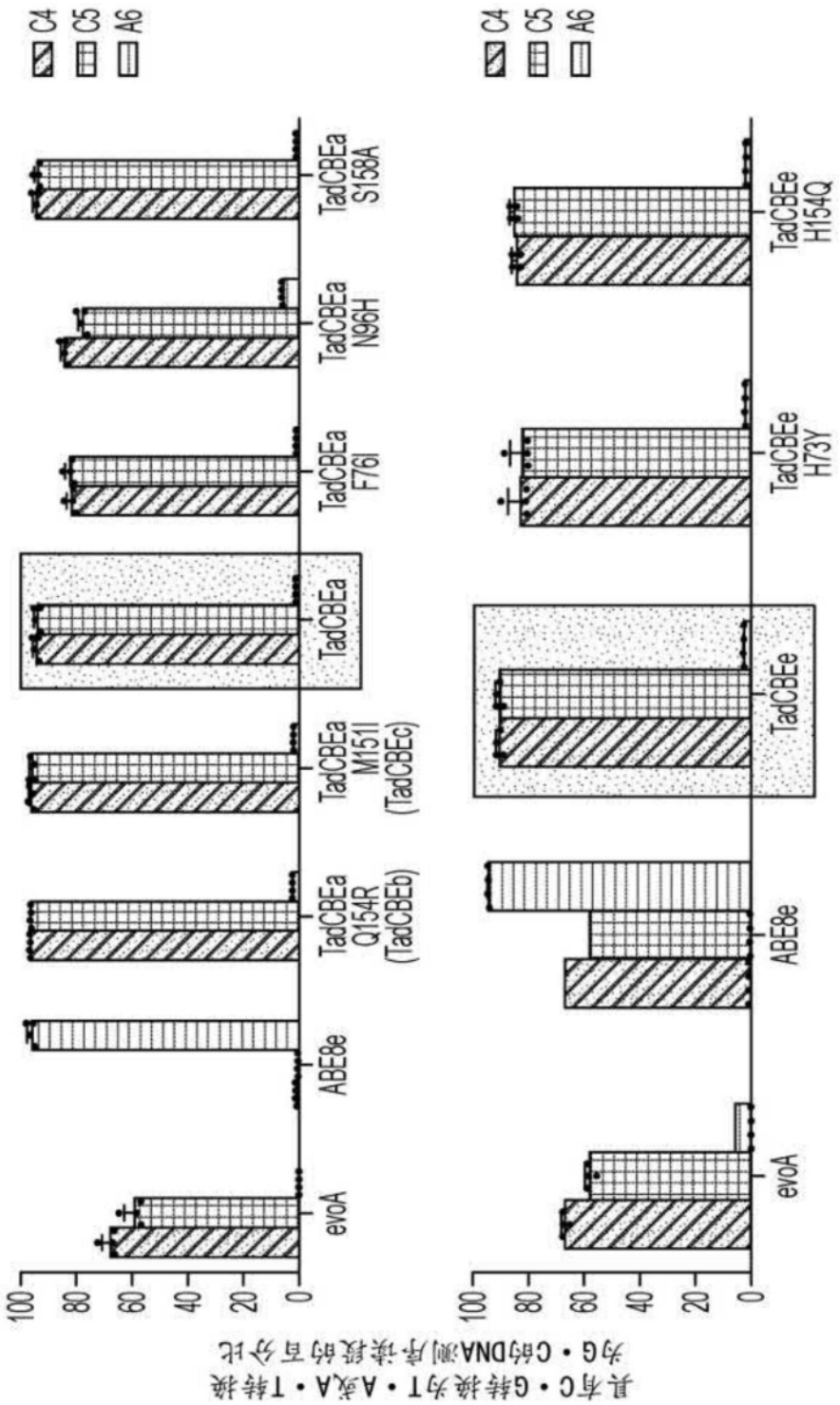

C4
C5
A6
evoA
ABE8e
TadCBEa Q154R (TadCBEb)
TadCBEa M151I (TadCBEc)
TadCBEa
TadCBEa F76I
TadCBEa N96H
TadCBEa S158A
evoA
ABE8e
TadCBEe
TadCBEe H73Y
TadCBEe H154Q
100
80
60
40
20
0
具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比

图34

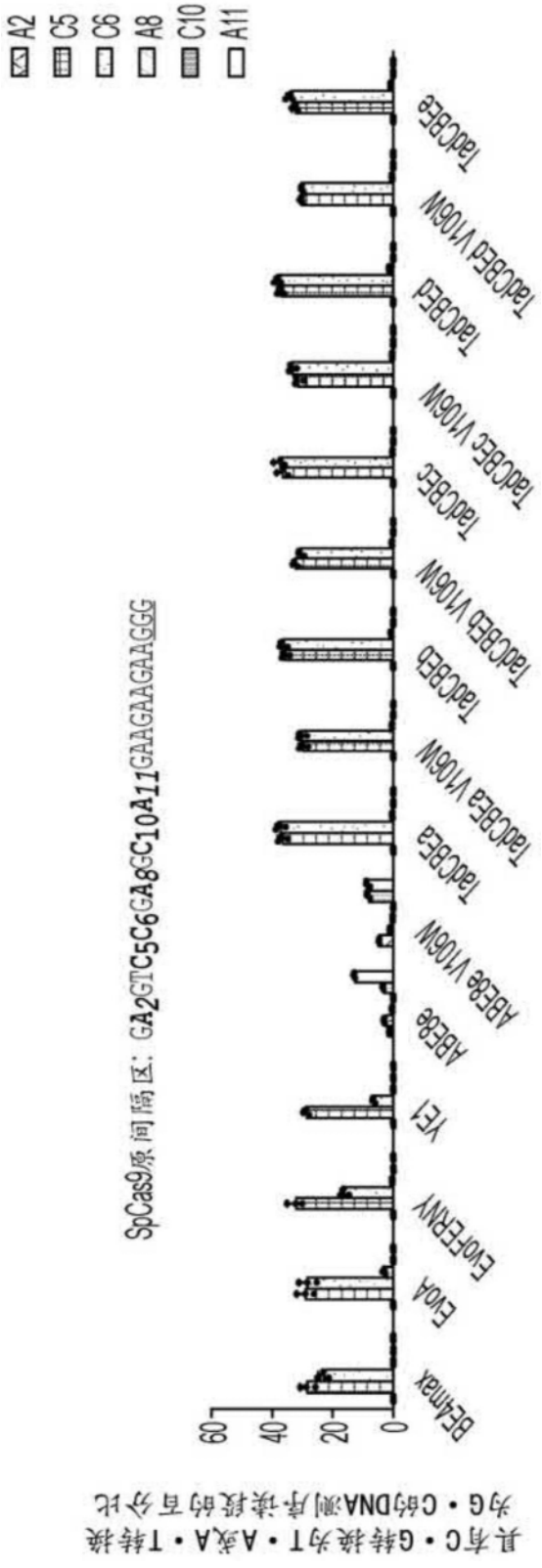

A2
C5
C6
A8
C10
A11
SpCas9原间隔区：GA2GTC5C6GA8GC10A11GAAGAAGAGGG
BE4max
EvoA
EvoFERNY
YE1
ABE8e
ABE8e V106W
TadABEa
TadABEa V106W
TadABEb
TadABEb V106W
TadABEc
TadABEc V106W
TadABEd
TadABEd V106W
TadABEe
0
20
40
60
具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比

图35

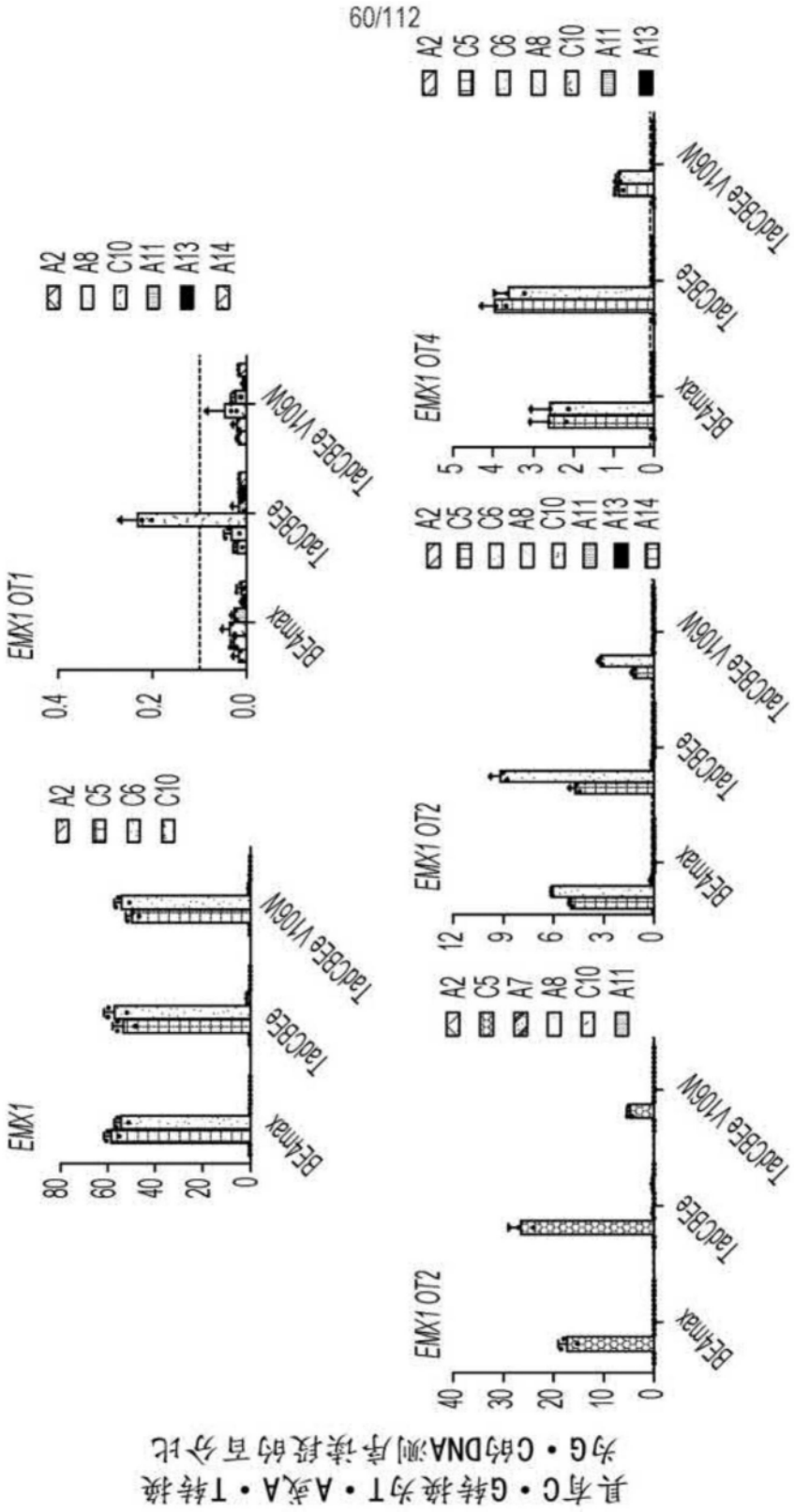

60/112
具有C·G转换为T·A或A·T转换
为G·C的DNA测序读段的百分比
EMX1
EMX1 OT1
EMX1 OT2
EMX1 OT2
EMX1 OT4
A2
A5
A6
A7
A8
A10
A11
A13
A14
BE4max
TadCBEe
TadCBEe V106W

图36

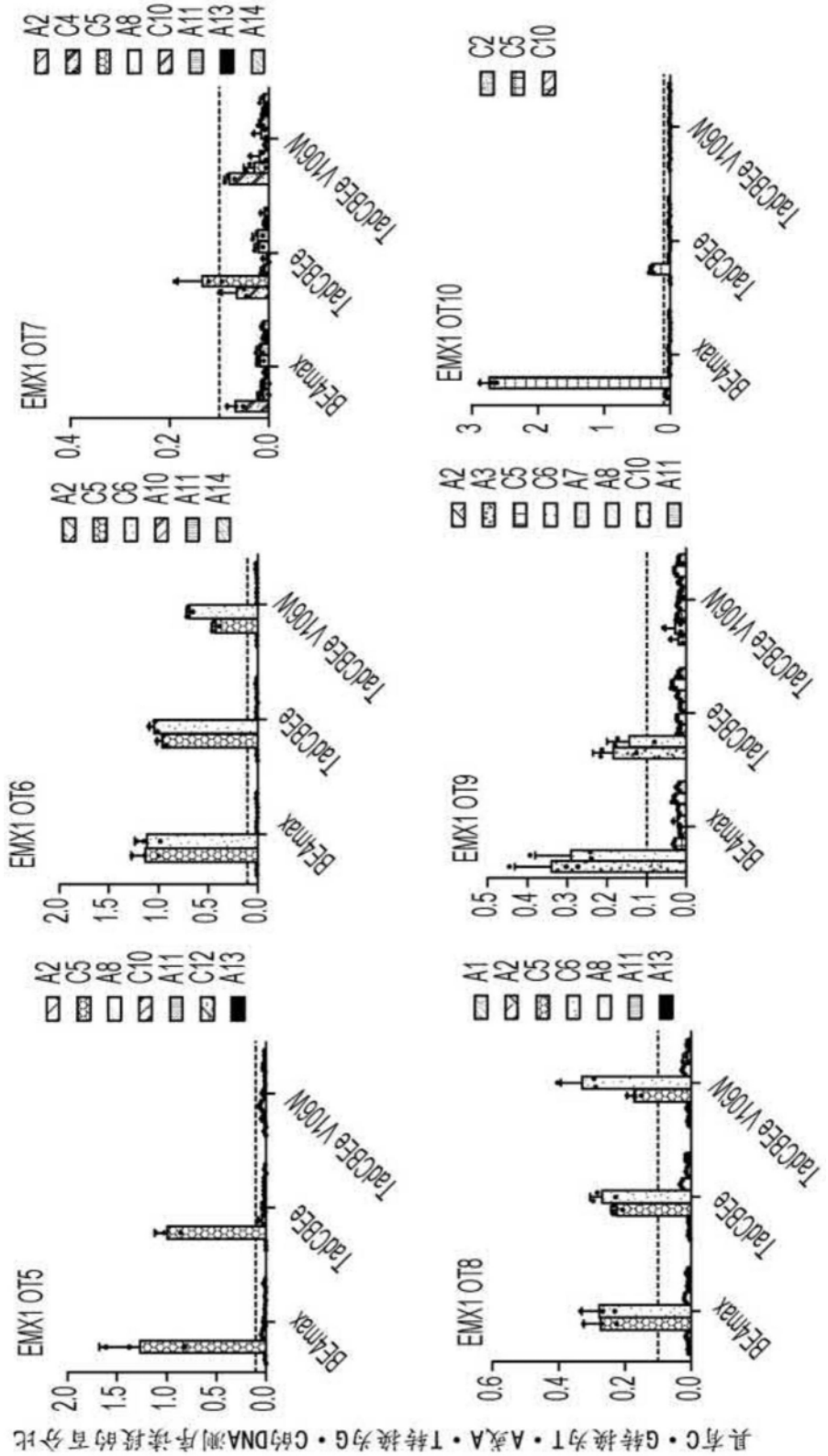

EMX1 OT7
EMX1 OT10
EMX1 OT6
EMX1 OT9
EMX1 OT5
EMX1 OT8
BE4max
TadCBEe
TadCBEe V106W
具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比

图36(续)

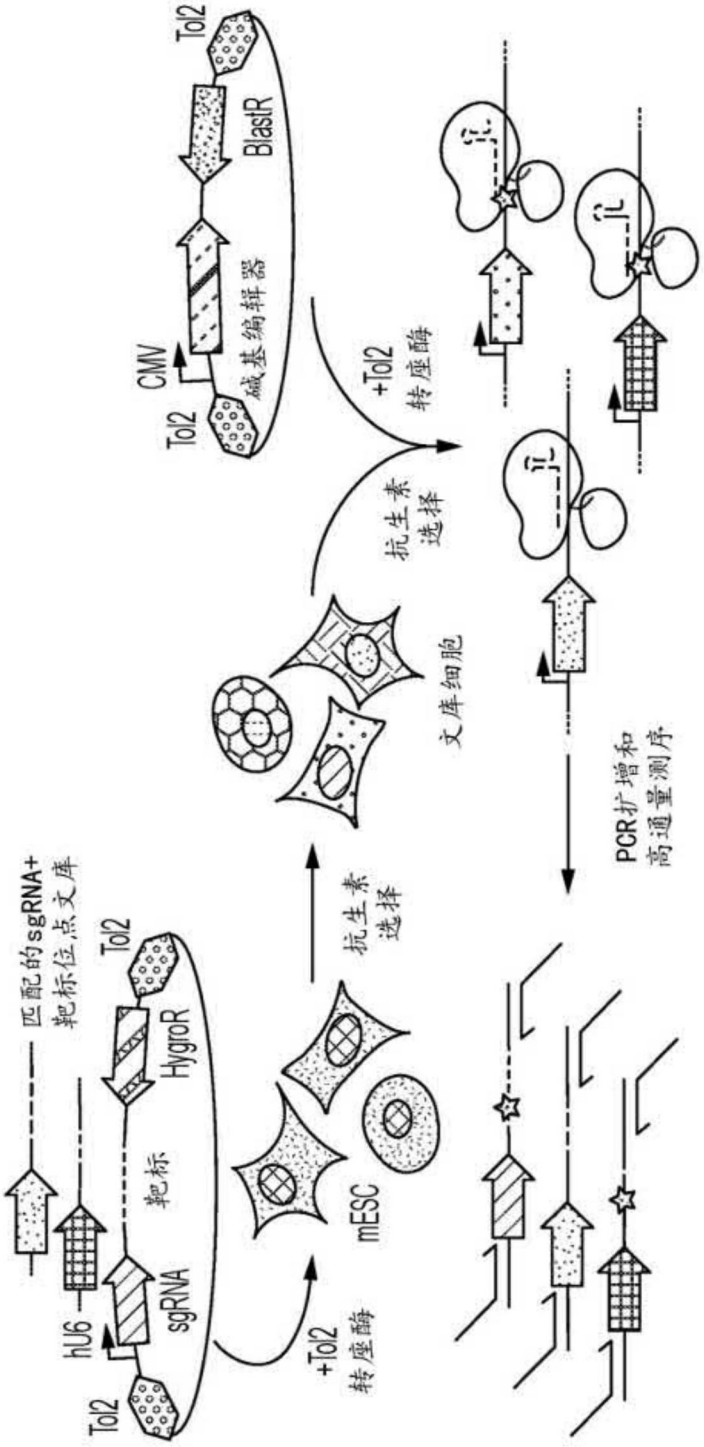

匹配的sgRNA+
靶标位点文库
hU6
sgRNA
靶标
Tol2
HygroR
Tol2
Tol2
+Tol2
转座酶
mESC
抗生素
选择
文库细胞
抗生素
选择
+Tol2
转座酶
CMV
Tol2
碱基编辑器
BlastR
Tol2
PCR扩增和
高通量测序

图37

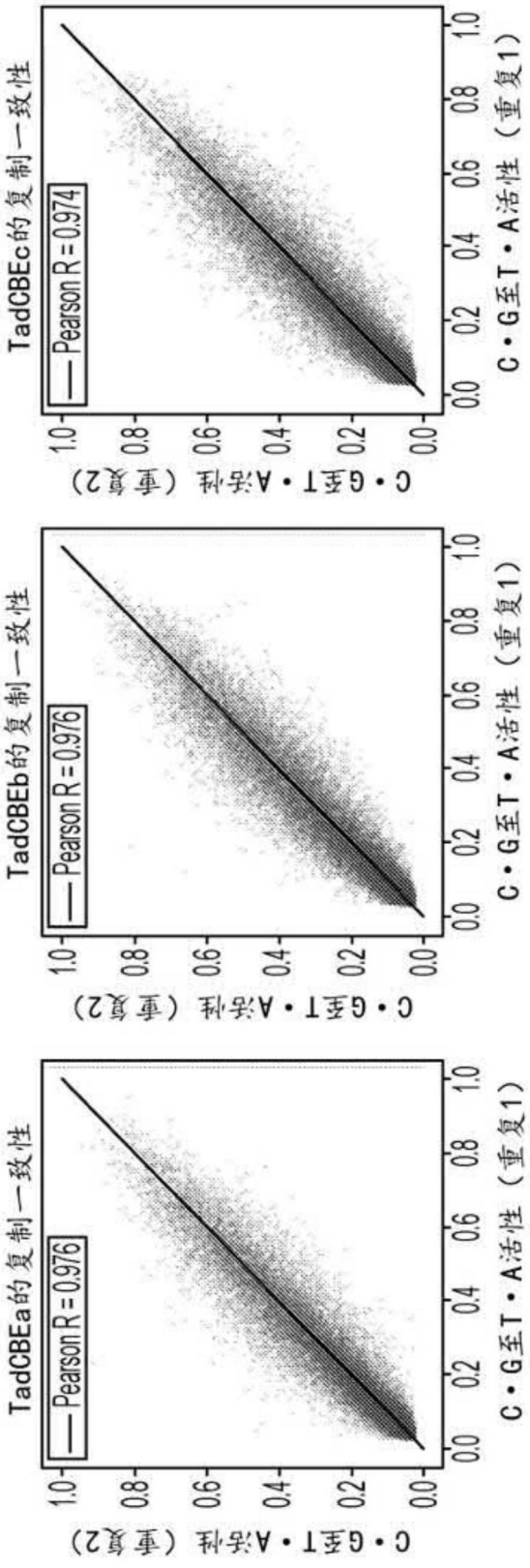

TadCBEa的复制一致性
Pearson R = 0.976
C • G至T • A活性（重复2）
C • G至T • A活性（重复1）
TadCBEb的复制一致性
Pearson R = 0.976
C • G至T • A活性（重复2）
C • G至T • A活性（重复1）
TadCBEc的复制一致性
Pearson R = 0.974
C • G至T • A活性（重复2）
C • G至T • A活性（重复1）
0.0
0.2
0.4
0.6
0.8
1.0

图38

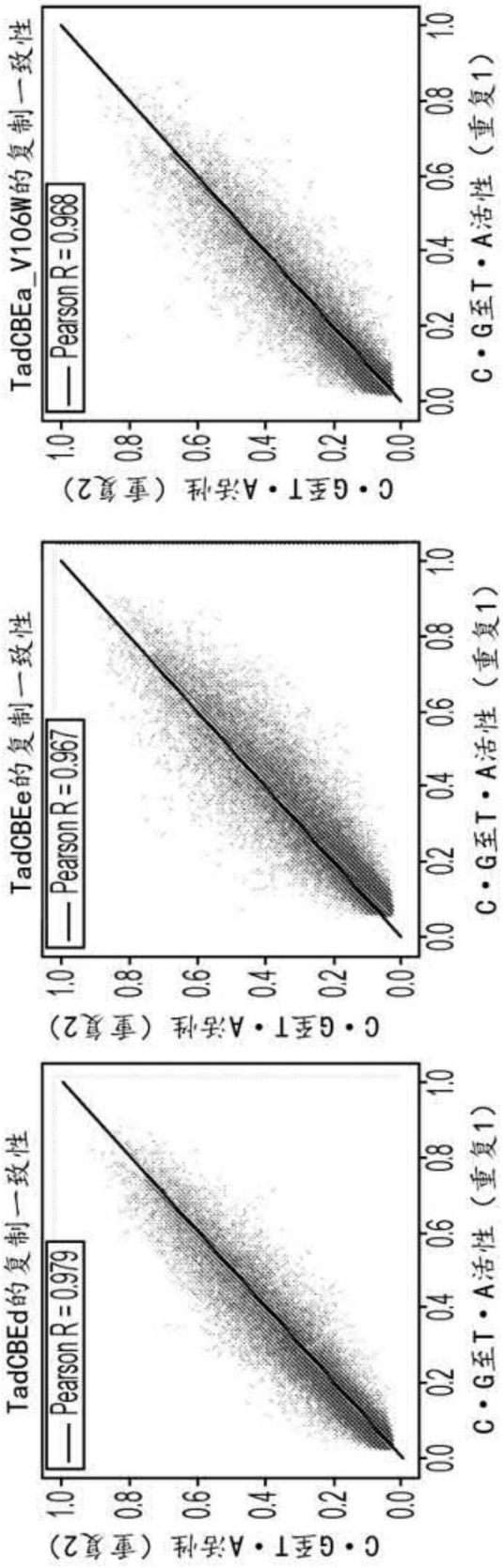

TadCBEd的复制一致性
Pearson R = 0.979
C • G至T • A活性（重复2）
C • G至T • A活性（重复1）
TadCBEe的复制一致性
Pearson R = 0.967
C • G至T • A活性（重复2）
C • G至T • A活性（重复1）
TadCBEa_V106W的复制一致性
Pearson R = 0.968
C • G至T • A活性（重复2）
C • G至T • A活性（重复1）

图38(续)

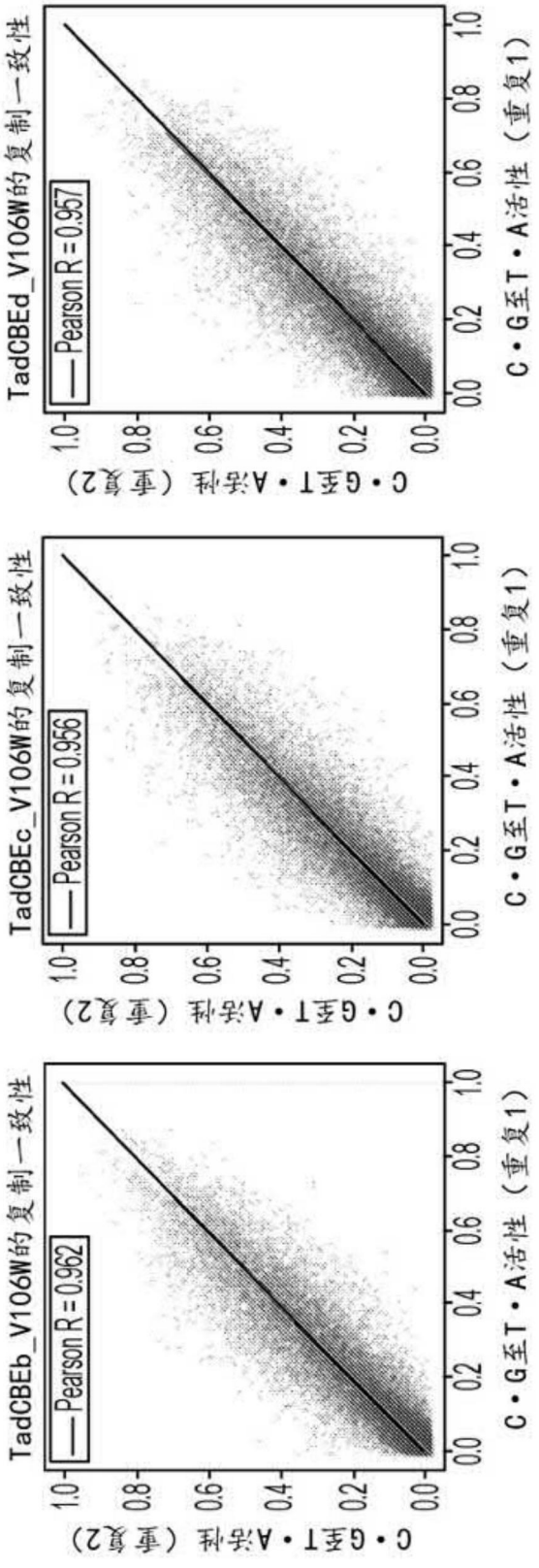

TadCBEb_V106W的复制一致性
Pearson R = 0.962
C • G至T • A活性（重复2）
C • G至T • A活性（重复1）
TadCBEc_V106W的复制一致性
Pearson R = 0.956
C • G至T • A活性（重复2）
C • G至T • A活性（重复1）
TadCBEd_V106W的复制一致性
Pearson R = 0.957
C • G至T • A活性（重复2）
C • G至T • A活性（重复1）
0.0
0.2
0.4
0.6
0.8
1.0

图38(续)

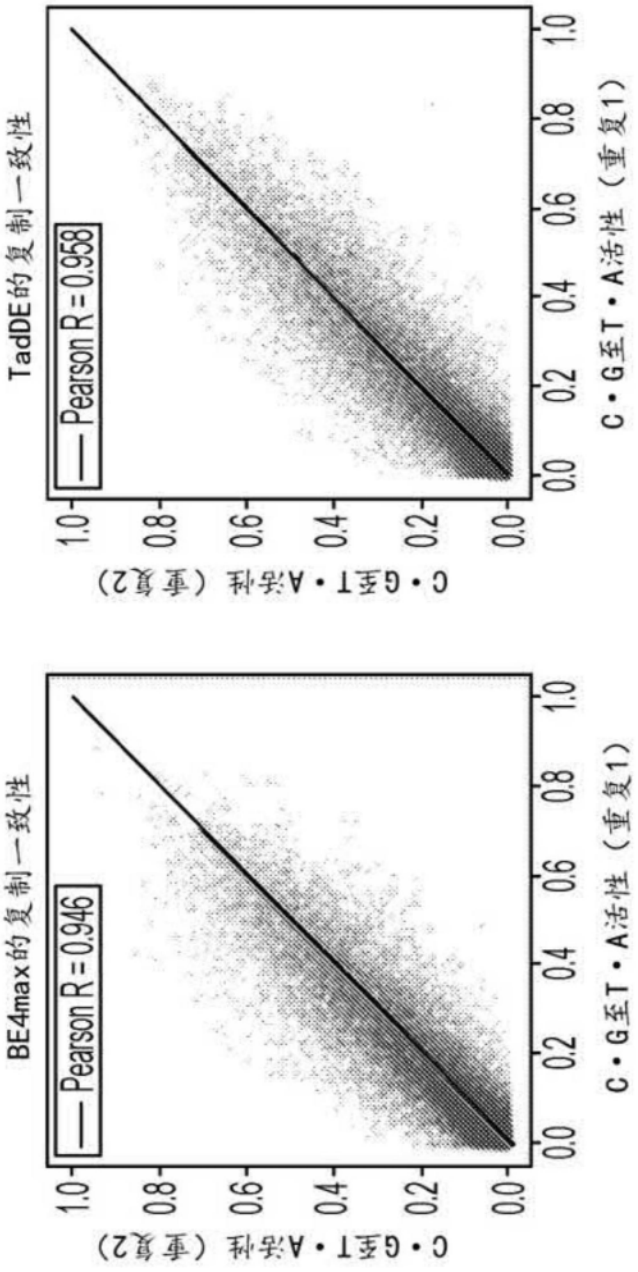

TadDE的复制一致性
Pearson R = 0.958
C • G至T • A活性（重复2）
C • G至T • A活性（重复1）
0.0
0.2
0.4
0.6
0.8
1.0
BE4max的复制一致性
Pearson R = 0.946
C • G至T • A活性（重复2）
C • G至T • A活性（重复1）

图38(续)

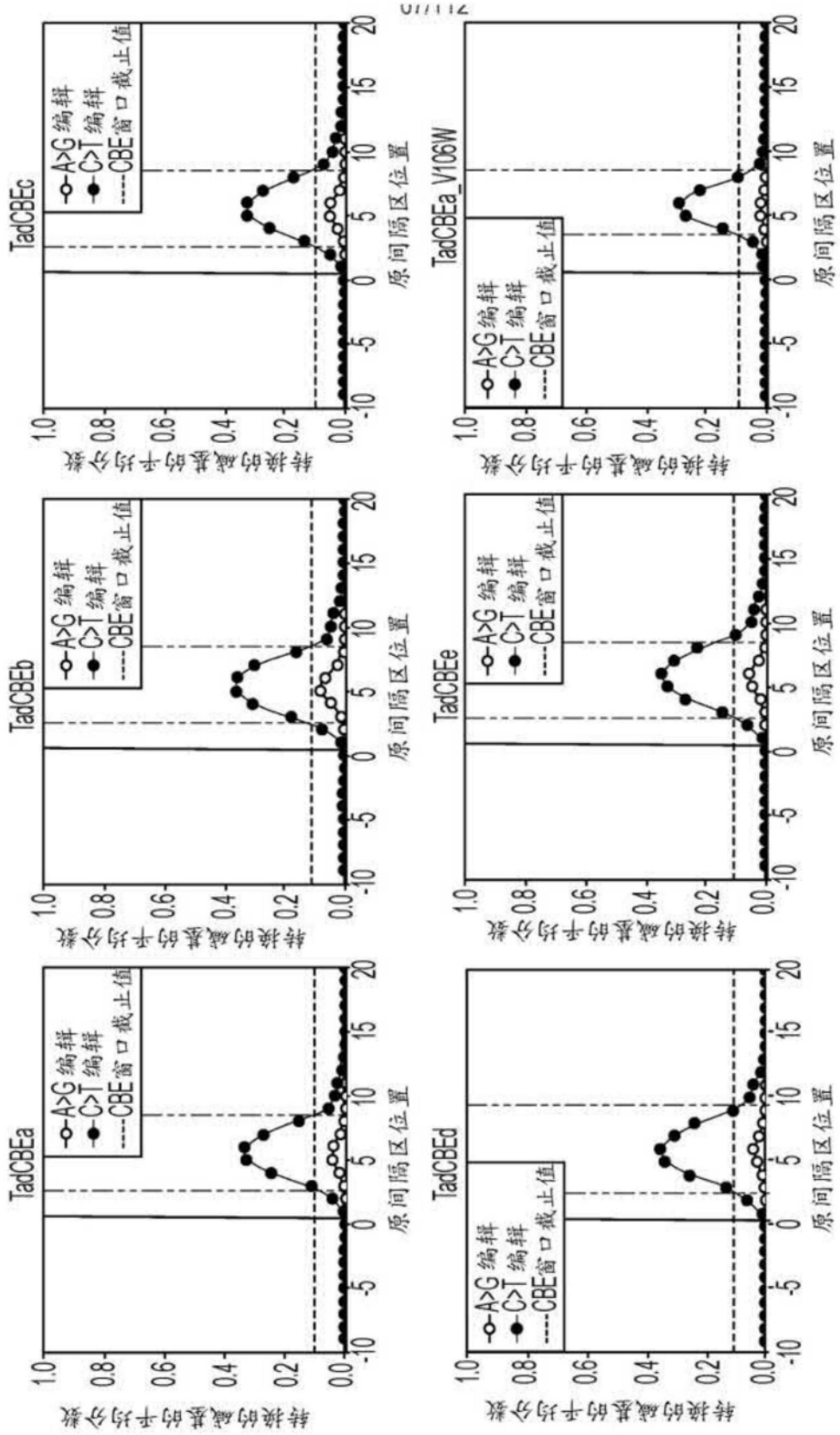

TadCBEa
TadCBEb
TadCBEc
TadCBEd
TadCBEe
TadCBEa_V106W
A>G 编辑
C>T 编辑
CBE窗口截止值
转换的碱基的平均分数
原间隔区位置

图39

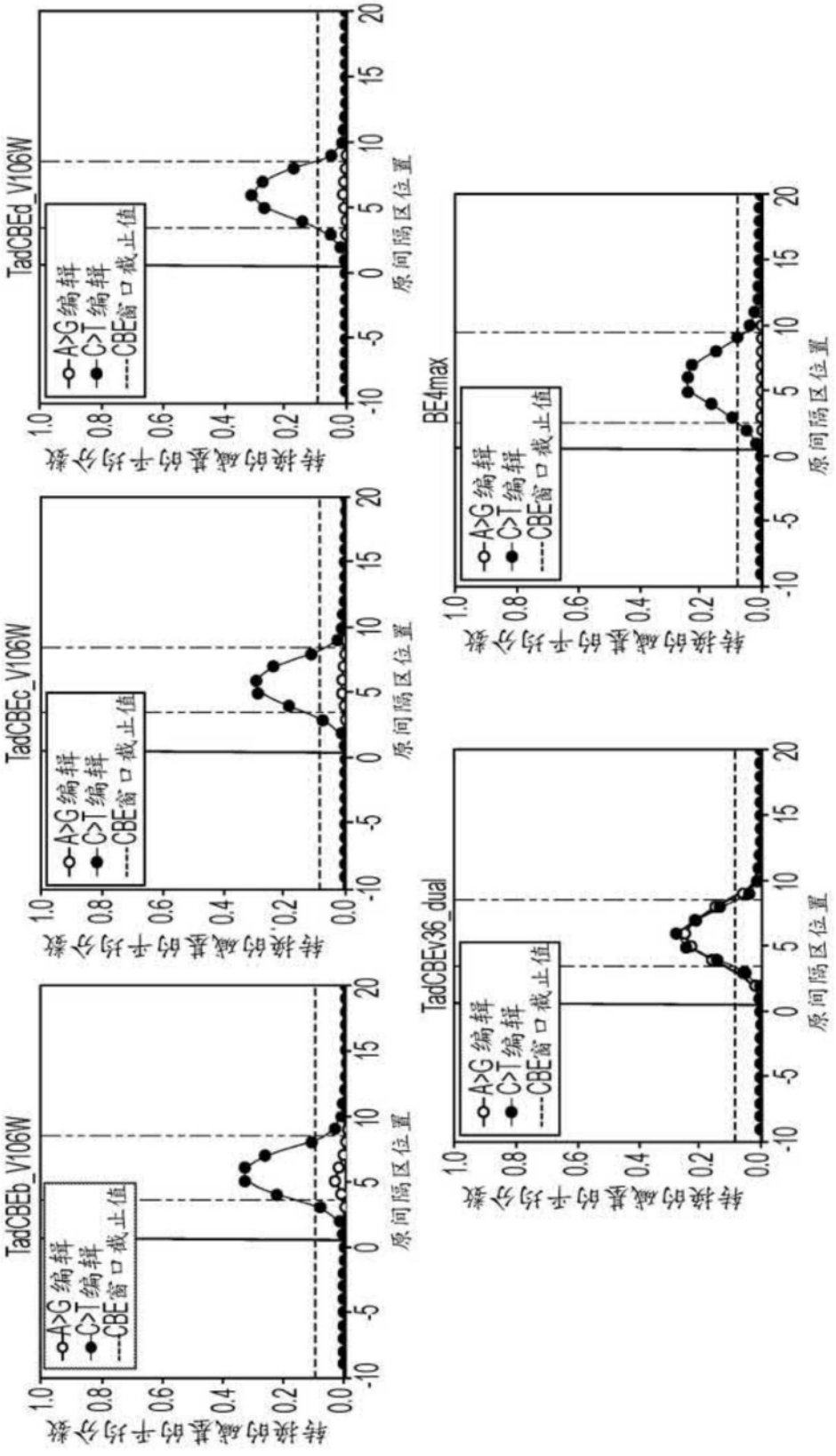

TadCBEb_V106W
TadCBEc_V106W
TadCBEd_V106W
BE4max
TadCBEv36_dual
A>G 编辑
C>T 编辑
CBE窗口截止值
转换的碱基的平均分数
原间隔区位置

图39(续)

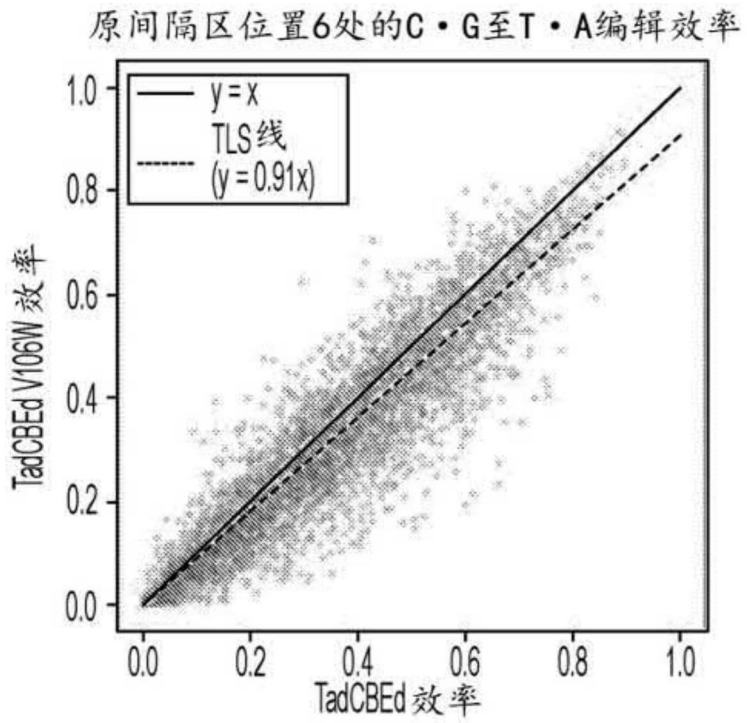

原间隔区位置6处的C•G至T•A编辑效率
y = x
TLS线
(y = 0.91x)
TadCBEd V106W 效率
TadCBEd 效率
0.0
0.2
0.4
0.6
0.8
1.0

图40A

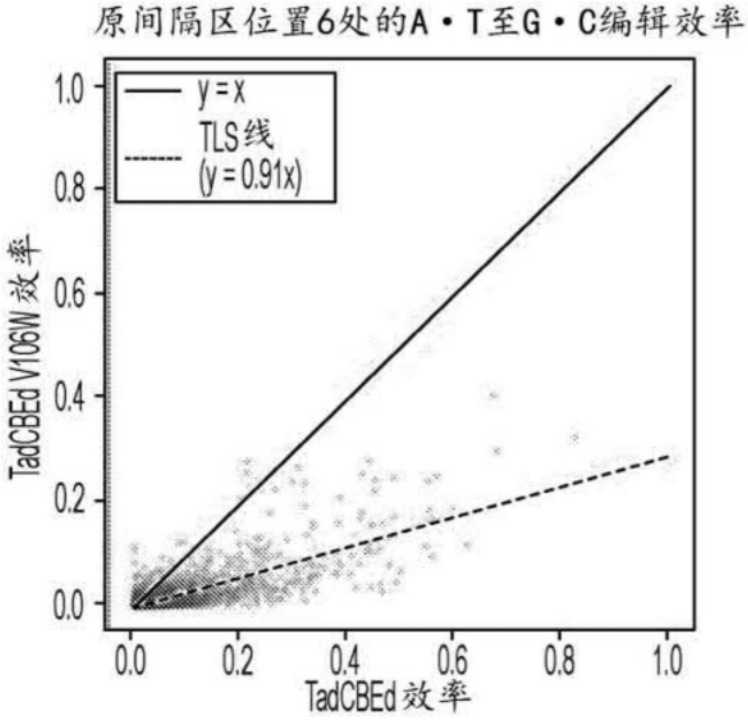

原间隔区位置6处的A•T至G•C编辑效率
y = x
TLS线
(y = 0.91x)
TadCBEd V106W 效率
TadCBEd 效率
0.0
0.2
0.4
0.6
0.8
1.0

图40B

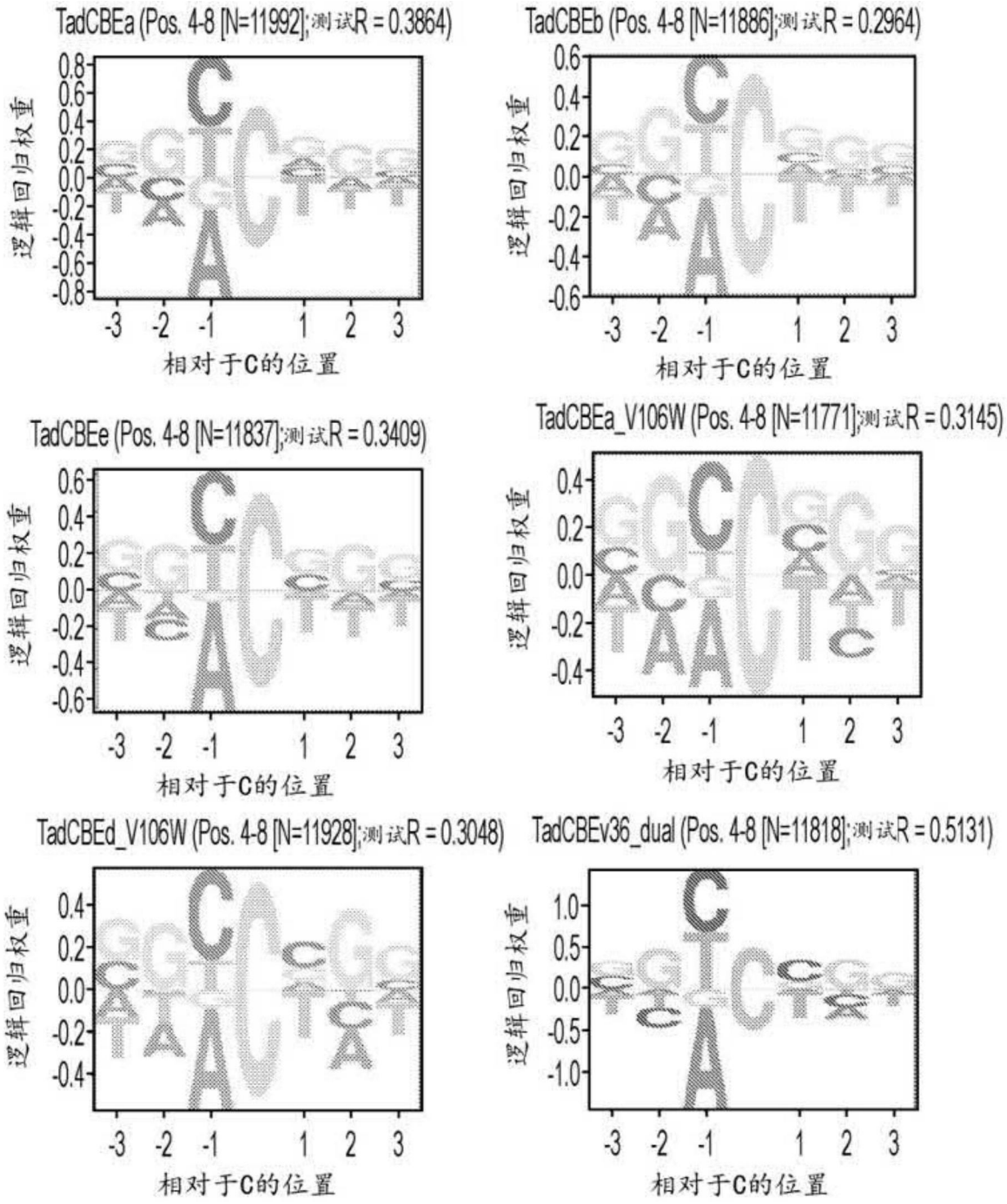

TadCBEa (Pos. 4-8 [N=11992];测试R = 0.3864)
TadCBEb (Pos. 4-8 [N=11886];测试R = 0.2964)
TadCBEe (Pos. 4-8 [N=11837];测试R = 0.3409)
TadCBEa_V106W (Pos. 4-8 [N=11771];测试R = 0.3145)
TadCBEd_V106W (Pos. 4-8 [N=11928];测试R = 0.3048)
TadCBEv36_dual (Pos. 4-8 [N=11818];测试R = 0.5131)
逻辑回归权重
相对于C的位置

图41A

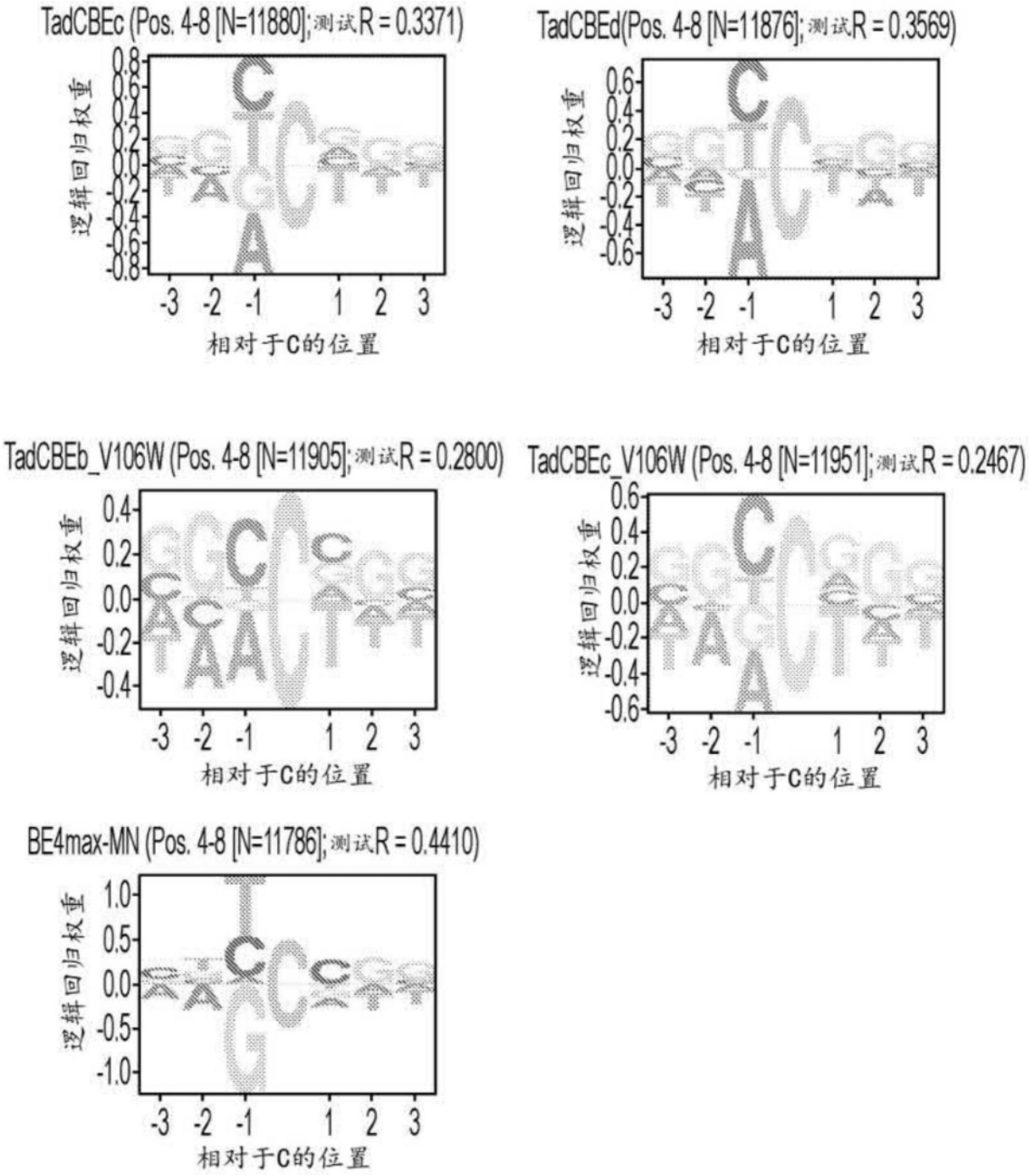

TadCBEc (Pos. 4-8 [N=11880]; 测试R = 0.3371)
TadCBEd(Pos. 4-8 [N=11876]; 测试R = 0.3569)
TadCBEb_V106W (Pos. 4-8 [N=11905]; 测试R = 0.2800)
TadCBEc_V106W (Pos. 4-8 [N=11951]; 测试R = 0.2467)
BE4max-MN (Pos. 4-8 [N=11786]; 测试R = 0.4410)
逻辑回归权重
相对于C的位置

图41A(续)

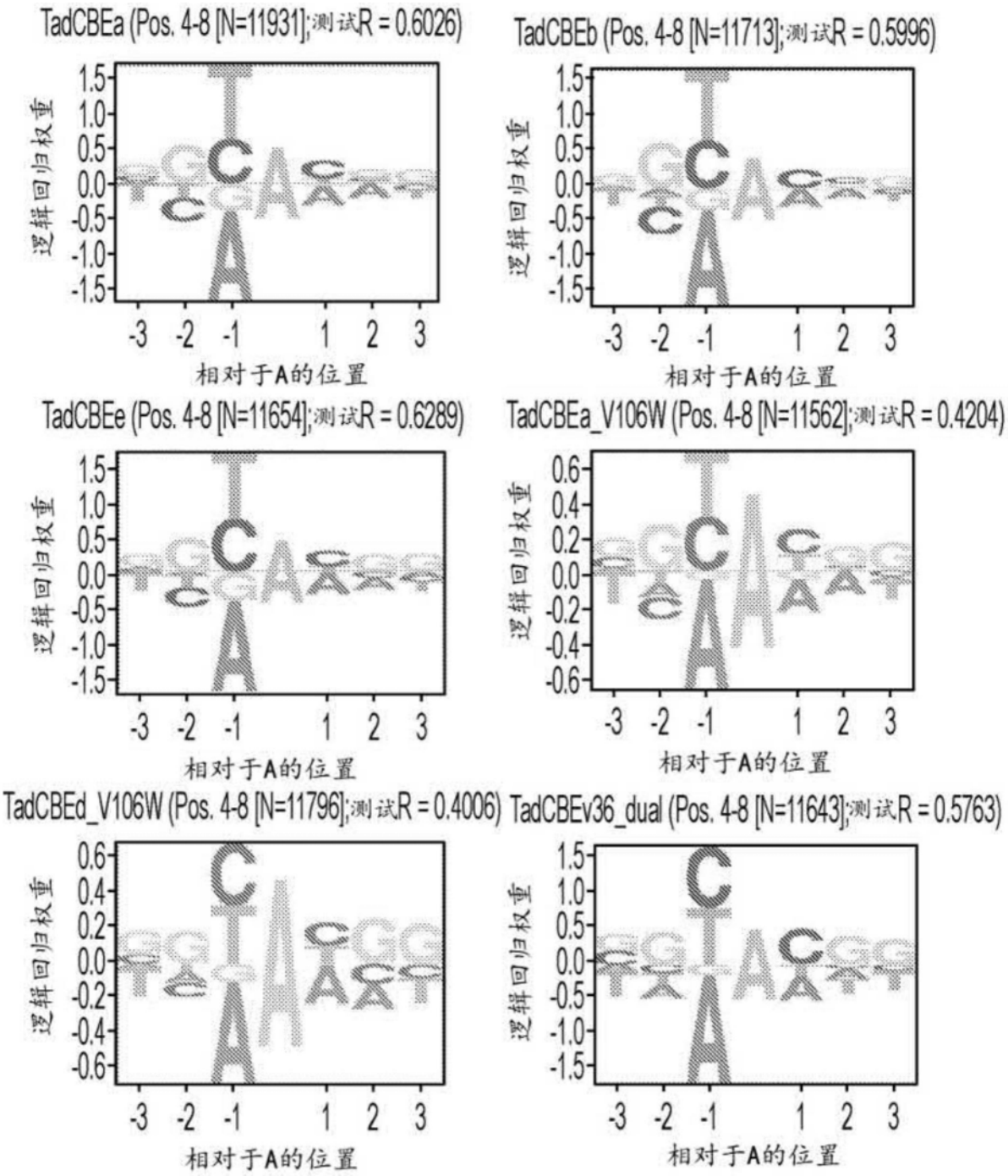

TadCBEa (Pos. 4-8 [N=11931];测试R = 0.6026)
TadCBEb (Pos. 4-8 [N=11713];测试R = 0.5996)
TadCBEe (Pos. 4-8 [N=11654];测试R = 0.6289)
TadCBEa_V106W (Pos. 4-8 [N=11562];测试R = 0.4204)
TadCBEd_V106W (Pos. 4-8 [N=11796];测试R = 0.4006)
TadCBEv36_dual (Pos. 4-8 [N=11643];测试R = 0.5763)
逻辑回归权重
相对于A的位置

图41B

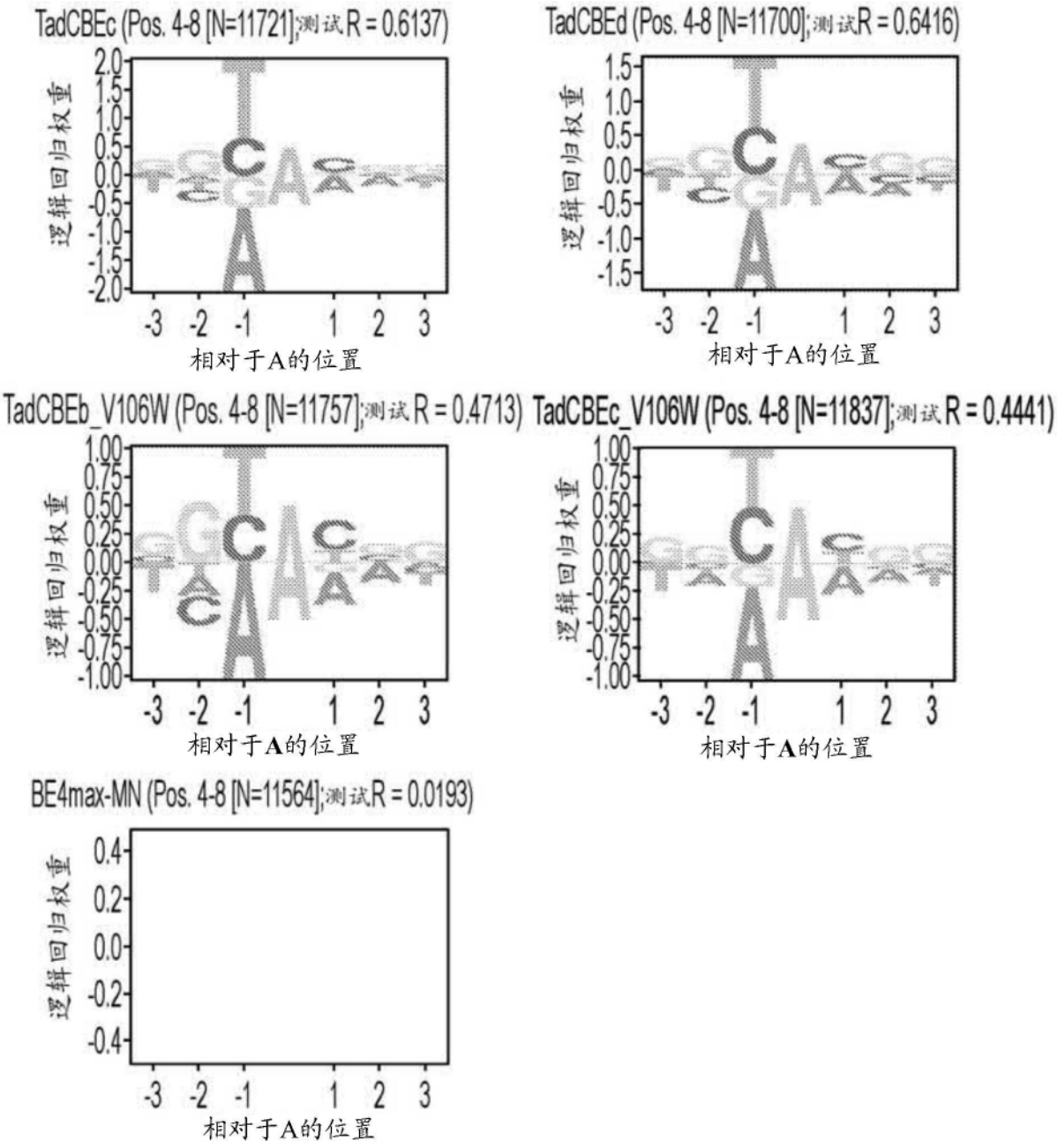

TadCBEc (Pos. 4-8 [N=11721];测试R = 0.6137)
TadCBEd (Pos. 4-8 [N=11700];测试R = 0.6416)
TadCBEb_V106W (Pos. 4-8 [N=11757];测试R = 0.4713)
TadCBEc_V106W (Pos. 4-8 [N=11837];测试R = 0.4441)
BE4max-MN (Pos. 4-8 [N=11564];测试R = 0.0193)
逻辑回归权重
相对于A的位置

图41B(续)

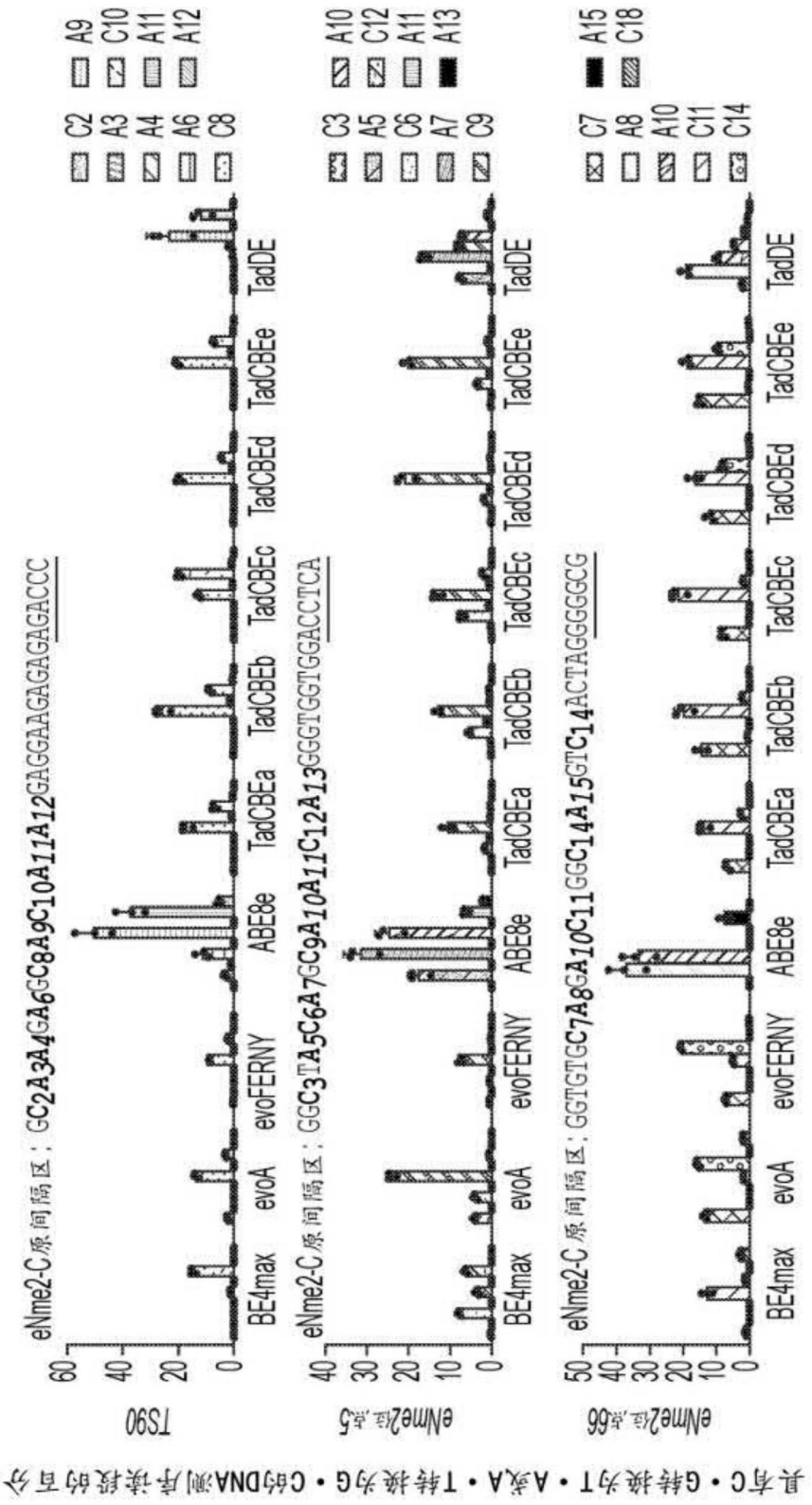

TS90
eNme2-C原间隔区：GC2A3A4GA6GC8A9C10A11A12GAGGAAGAGAGAGACCC
C2 A3 A4 A6 C8 A9 C10 A11 A12
BE4max evoA evoFERNY ABE8e TadCBEa TadCBEb TadCBEc TadCBEd TadCBEe TadDE
eNme2-C原间隔区：GGC3TA5C6A7GC9A10A11C12A13GGGTGGTGGACCTCA
C3 A5 C6 A7 C9 A10 C12 A11 A13
BE4max evoA evoFERNY ABE8e TadCBEa TadCBEb TadCBEc TadCBEd TadCBEe TadDE
eNme2-C原间隔区：GGTGTGC7A8GA10C11GGC14A15GTC14ACTAGGGGGCG
C7 A8 A10 C11 C14 A15 C18
BE4max evoA evoFERNY ABE8e TadCBEa TadCBEb TadCBEc TadCBEd TadCBEe TadDE
eNme2-C原间隔区5
eNme2-C原间隔区66
具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比

图42

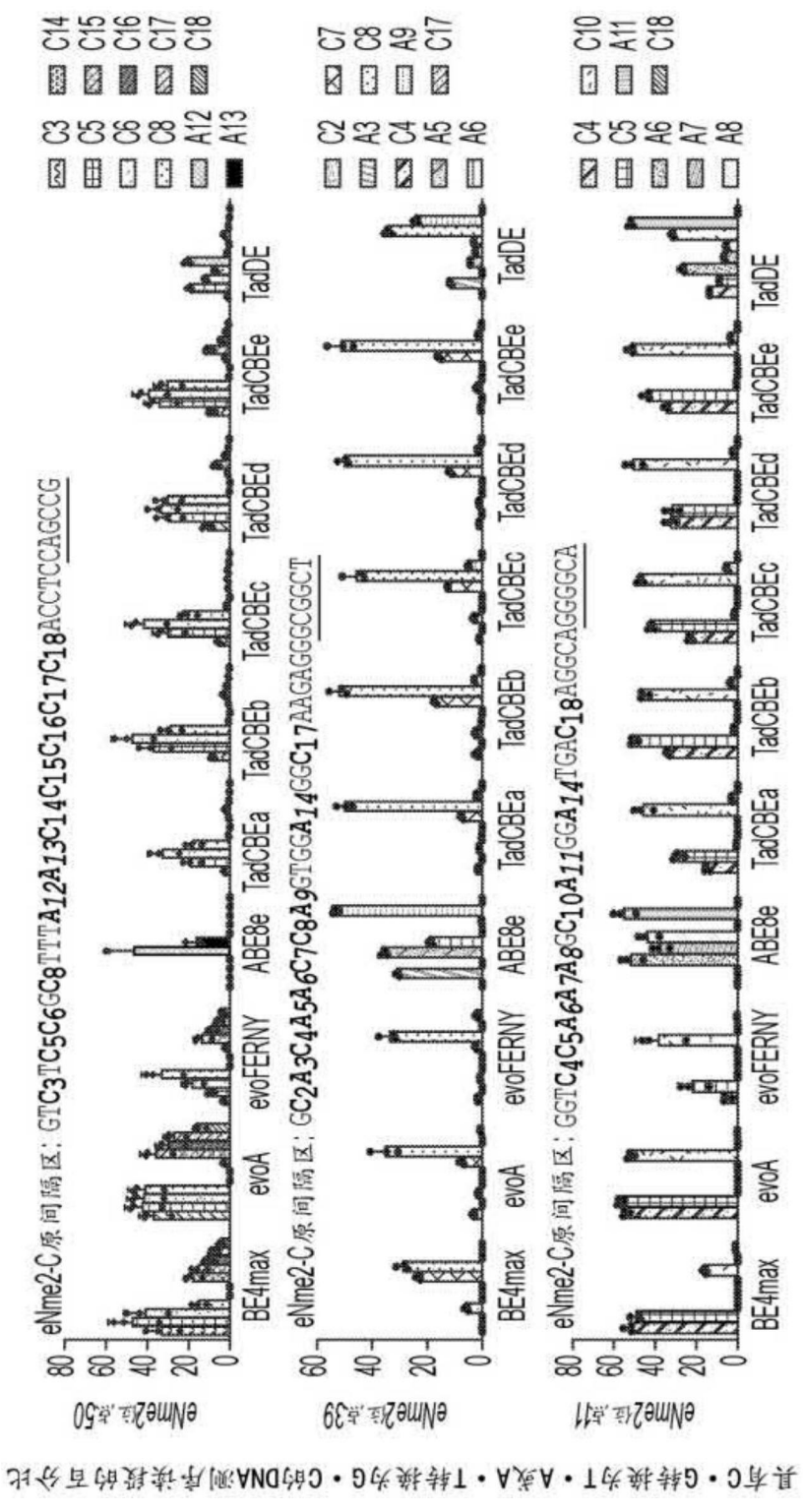

eNme2-C原间隔区: GTC3TC5C6GC8TTTA12A13C14C15C16C17C18ACCTCCAGCCG
eNme2-C原间隔区: GC2A3C4A5A6C7C8A9GTGGA14GGC17AAGAGGGCGGCT
eNme2-C原间隔区: GGTC4C5A6A7A8GC10A11GGA14TGAC18AGGCAGGGGCA
BE4max evoA evoFERNY ABE8e TadCBEa TadCBEb TadCBEc TadCBEd TadCBEe TadDE
C3 C5 C6 C8 A12 A13 C14 C15 C16 C17 C18
C2 A3 C4 A5 A6 C7 C8 A9 C17
C4 C5 A6 A7 A8 C10 A11 C18
eNme2位点50
eNme2位点39
eNme2位点41
具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比

图42(续)

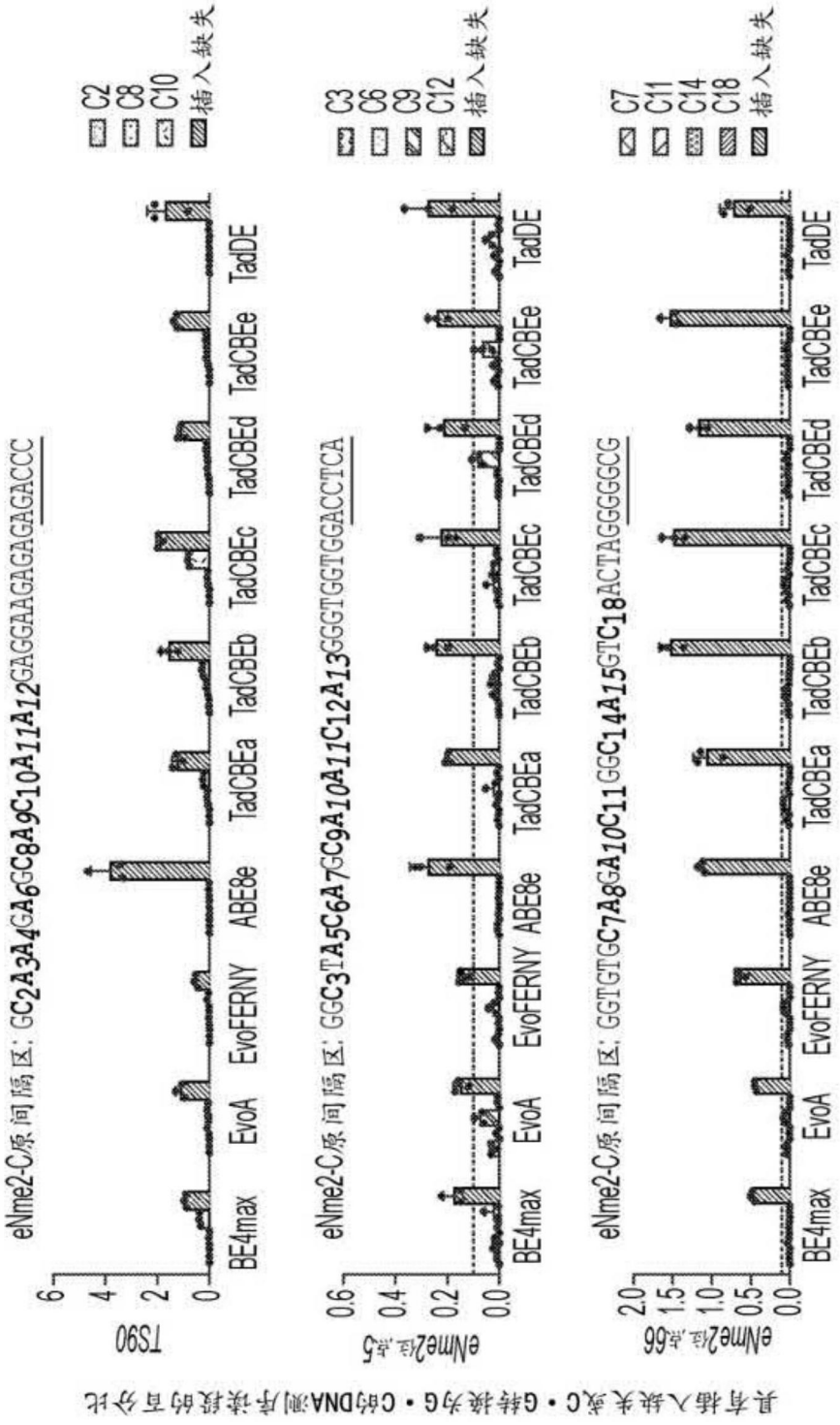

具有插入缺失或C • G转换为G • C的DNA测序读段的百分比
TS90
eNme2位点5
eNme2位点66
eNme2-C原间隔区: GC2A3A4GA6GC8A9C10A11A12GAGGAAGAGAGAGACCC
eNme2-C原间隔区: GGC3TA5C6A7GC9A10A11C12A13GGGTGGTGGACCTCA
eNme2-C原间隔区: GGTGTGC7A8GA10C11GGC14A15GTC18ACTAGGGGGCG
C2
C8
C10
插入缺失
C3
C6
C9
C12
插入缺失
C7
C11
C14
C18
插入缺失
BE4max
EvoA
EvoFERNY
ABE8e
TadCBEa
TadCBEb
TadCBEc
TadCBEd
TadCBEe
TadDE

图43

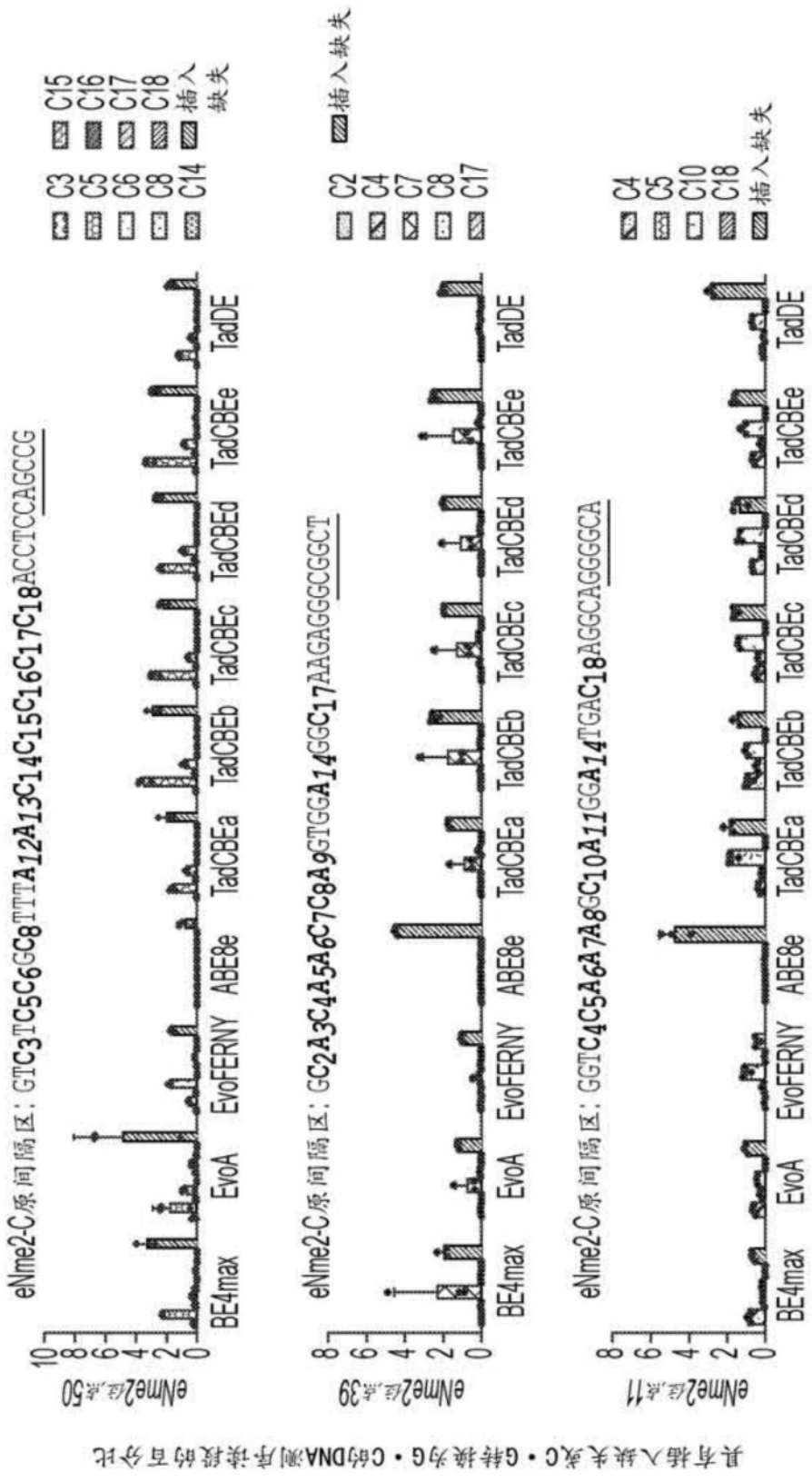

具有插入缺失或C·G转换为G·C的DNA测序读段的百分比
eNme2-C原间隔区：GTC3TC5C6GC8TTTA12A13C14C15C16C17C18ACCTCCAGCCG
eNme2-C原间隔区：GC2A3C4A5A6C7C8A9GTGGA14GGC17AAGAGGGCGGCT
eNme2-C原间隔区：GGTC4C5A6A7A8GC10A11GGA14TGAC18AGGCAGGGGCA
BE4max EvoA EvoFERNY ABE8e TadCBEa TadCBEb TadCBEc TadCBEd TadCBEe TadDE
C3 C5 C6 C8 C14 C15 C16 C17 C18 插入 缺失
C2 C4 C7 C8 C17 插入缺失
C4 C5 C10 C18 插入缺失

图43(续)

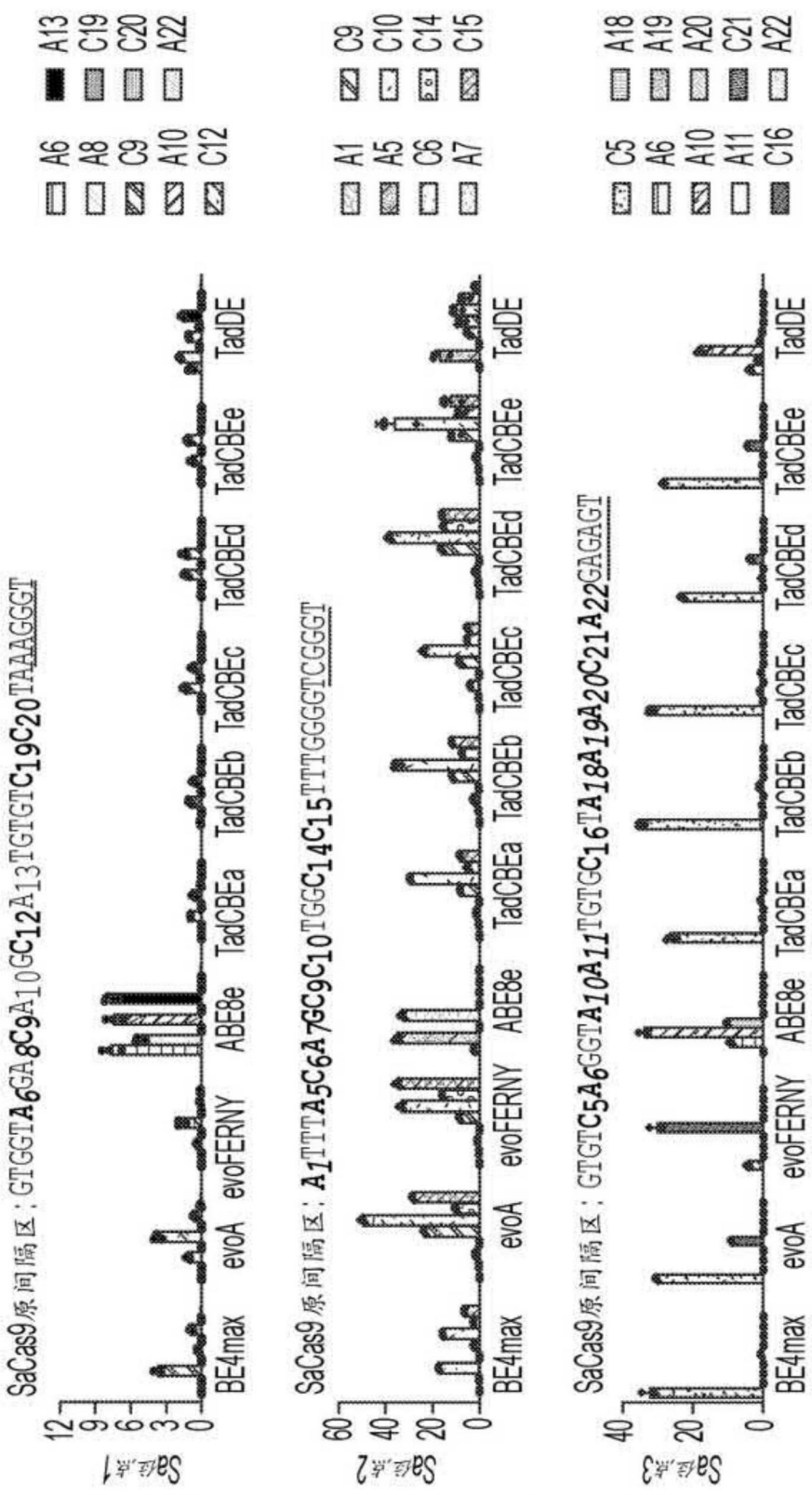

具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比
Sa位点1
Sa位点2
Sa位点3
SaCas9原间隔区：GTGGTA6GA8C9A10GC12A13TGTGTC19C20TAAAGGGT
SaCas9原间隔区：A1TTTA5C6A7GC9C10TGGC14C15TTTGGGGTCGGGT
SaCas9原间隔区：GTGTC5A6GGTA10A11TGTGC16TA18A19A20C21A22GAGAGT
BE4max evoA evoFERNY ABE8e TadCBEa TadCBEb TadCBEc TadCBEd TadCBEe TadDE
A6 A8 C9 A10 C12 A13 C19 C20 A22
A1 A5 C6 A7 C9 C10 C14 C15
C5 A6 A10 A11 C16 A18 A19 A20 C21 A22

图44

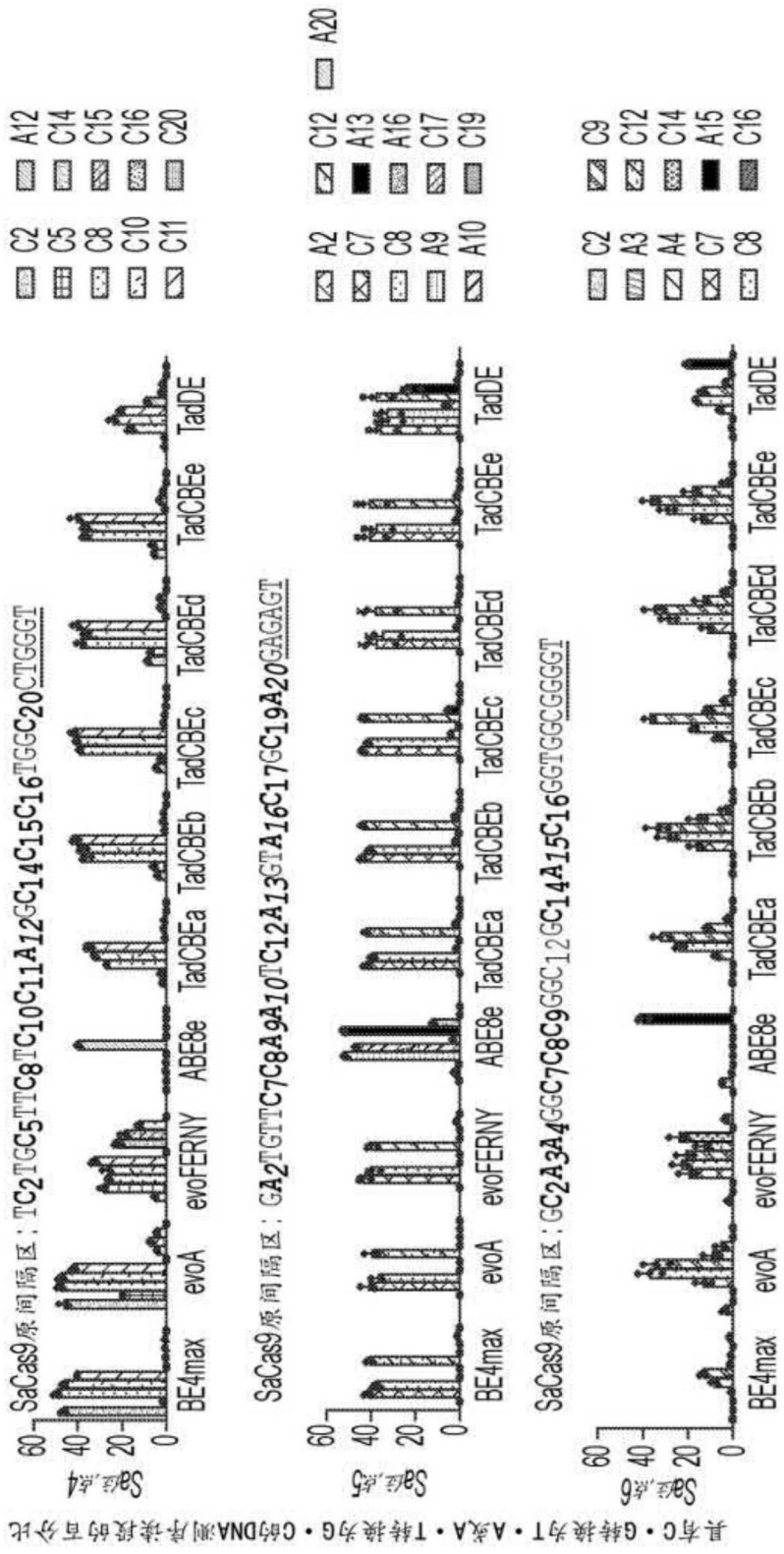

具有C·G转换为T·A或A·T转换为G·G的DNA测序读段的百分比
Sa位点4
Sa位点5
Sa位点6
SaCas9原间隔区：TC2TGC5TTC8TC10C11A12GC14C15C16TGGC20CTGGGT
SaCas9原间隔区：GA2TGTTC7C8A9A10TC12A13GTA16C17GC19A20GAGAGT
SaCas9原间隔区：GC2A3A4GGC7C8C9GGC12GC14A15C16GGTGGCGGGGT
BE4max evoA evoFERNY ABE8e TadCBEa TadCBEb TadCBEc TadCBEd TadCBEe TadDE
A12 C14 C15 C16 C20
C2 C5 C8 C10 C11
A20
C12 A13 A16 C17 C19
A2 C7 C8 A9 A10
C9 C12 C14 A15 C16
C2 A3 A4 C7 C8
0 20 40 60

图44(续)

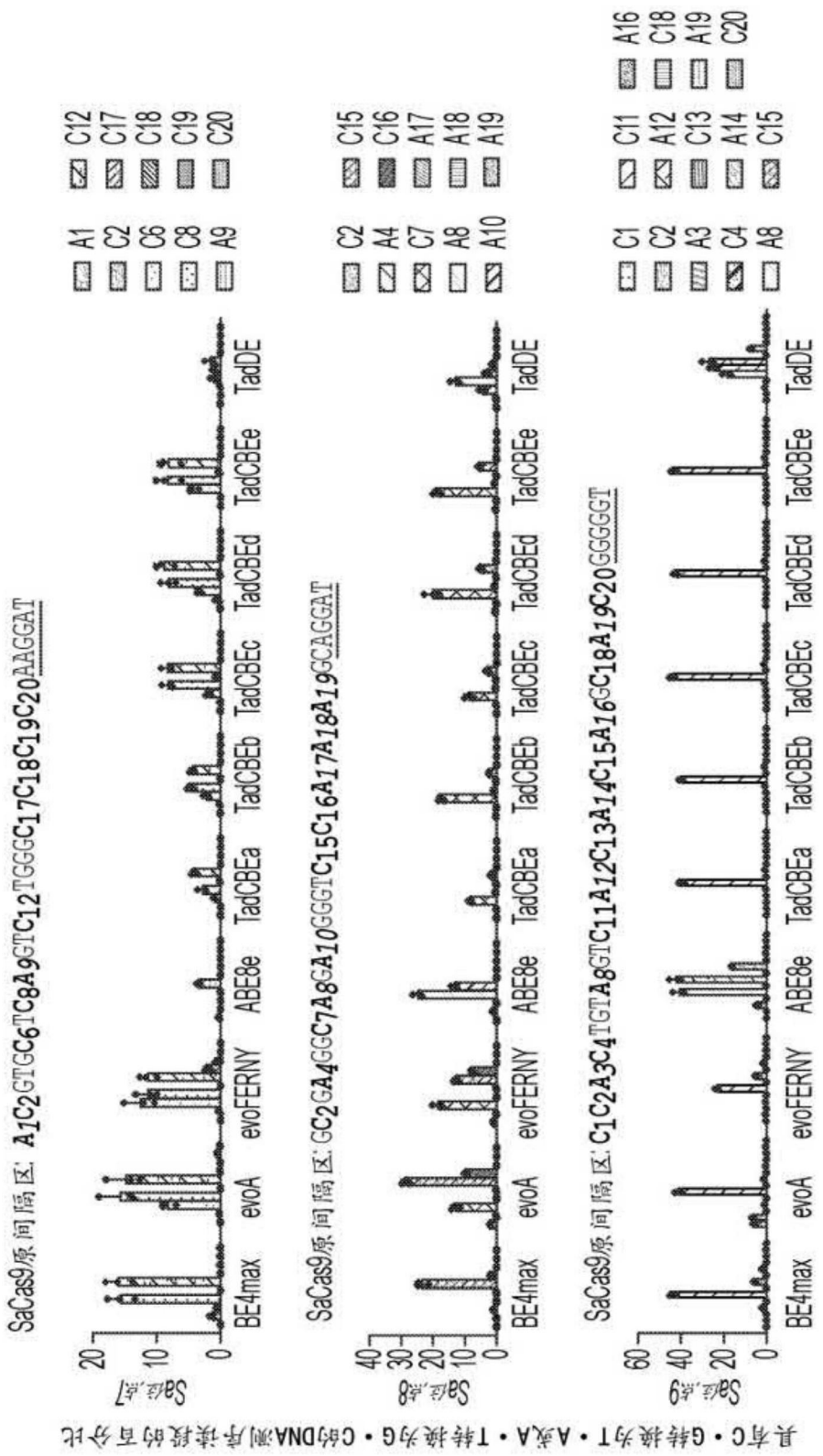

具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比
Sa位点7
Sa位点8
Sa位点9
SaCas9原间隔区: A1C2GTGC6TC8A9GTC12TGGGC17C18C19C20AAGGAT
SaCas9原间隔区: GC2GA4GGC7A8GA10GGGTC15C16A17A18A19GCAGGAT
SaCas9原间隔区: C1C2A3C4TGTA8GTC11A12C13A14C15A16GC18A19C20GGGGGT
BE4max evoA evoFERNY ABE8e TadCBEa TadCBEb TadCBEc TadCBEd TadCBEe TadDE
A1 C2 C6 C8 A9 C12 C17 C18 C19 C20
C2 A4 C7 A8 A10 C15 C16 A17 A18 A19
C1 C2 A3 C4 A8 C11 A12 C13 C15 A16 C18 A19 C20

图44(续)

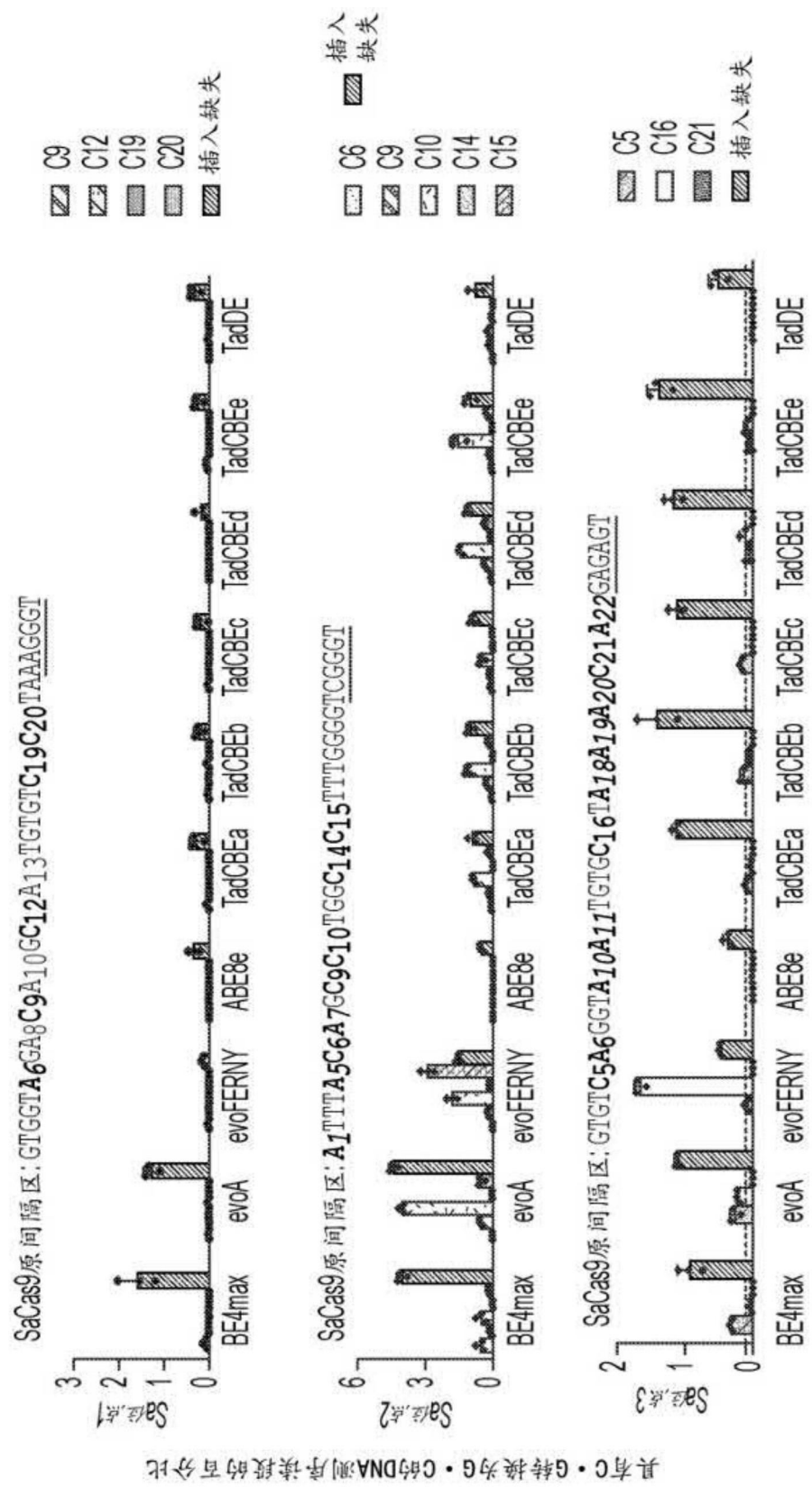

具有C • G转换为G • C的DNA测序读段的百分比
Sa位点1
Sa位点2
Sa位点3
SaCas9原间隔区:GTGGTA6GA8C9A10GC12A13TGTGTC19C20TAAAGGGT
SaCas9原间隔区:A1TTTA5C6A7GC9C10TGGC14C15TTTGGGGTCGGGT
BE4max
evoA
evoFERNY
ABE8e
TadCBEa
TadCBEb
TadCBEc
TadCBEd
TadCBEe
TadDE
C9
C12
C19
C20
插入缺失
C6
C9
C10
C14
C15
插入缺失
C5
C16
C21
插入缺失

图45

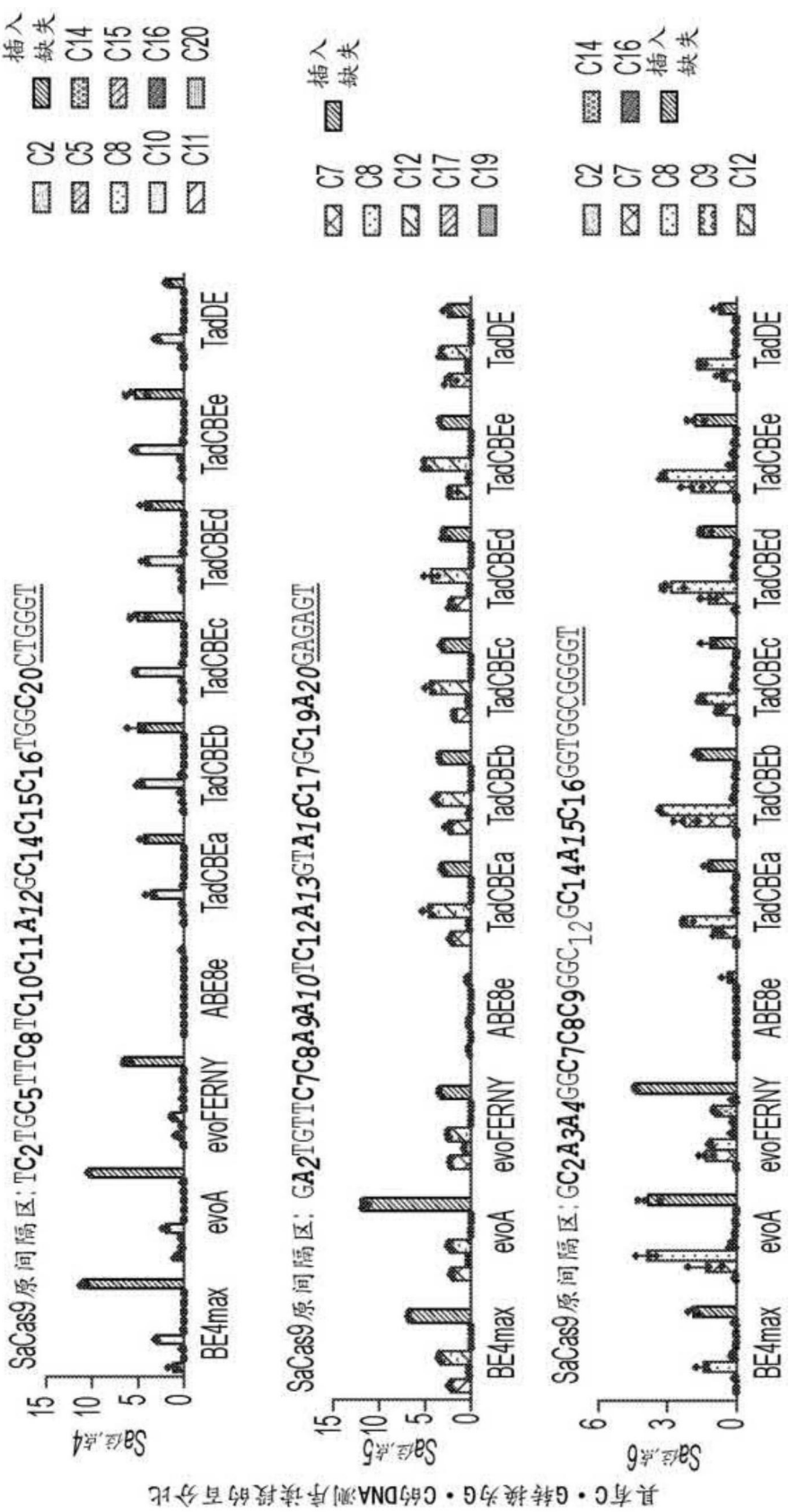

SaCas9原间隔区: TC2TGC5TTC8TC10C11A12GC14C15C16TGGC20CTGGGT
BE4max evoA evoFERNY ABE8e TadCBEa TadCBEb TadCBEc TadCBEd TadCBEe TadDE
插入
缺失
C2 C5 C8 C10 C11 C14 C15 C16 C20
SaCas9原间隔区: GA2TGTTC7C8A9A10TC12A13GTA16C17GC19A20GAGGT
C7 C8 C12 C17 C19
C2 C7 C8 C9 C12 C14 C16
SaCas9原间隔区: GC2A3A4GGC7C8C9GGC12GC14A15C16GGTGGCGGGGT
具有C • G转换为G • C的DNA测序读段的百分比

图45(续)

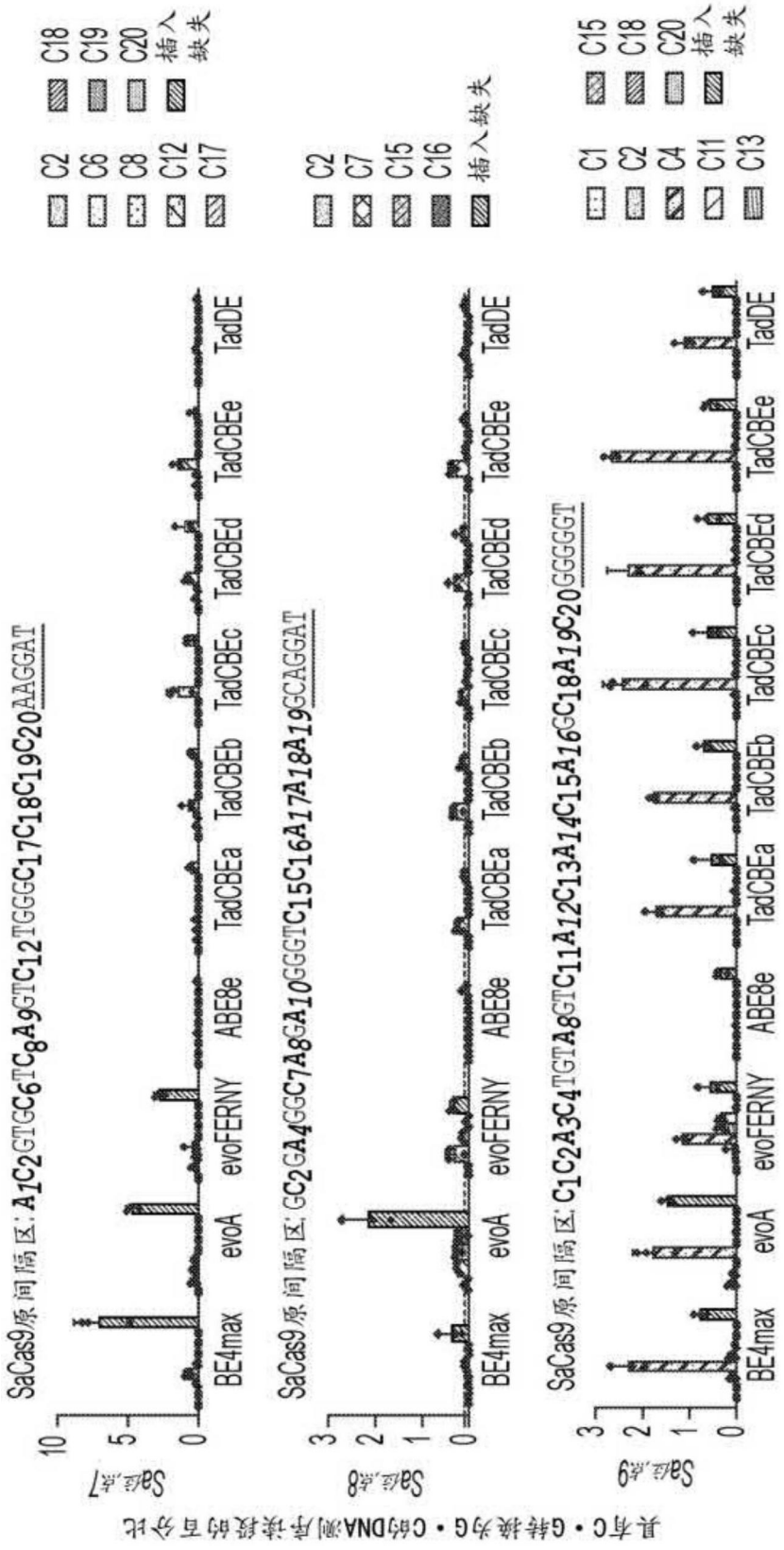

SaCas9原间隔区: A1C2GTGC6TC8A9GTC12TGGGC17C18C19C20AAGGAT
SaCas9原间隔区: GC2GA4GGC7A8GA10GGGTC15C16A17A18A19GCAGGAT
SaCas9原间隔区: C1C2A3C4TGTA8GTC11A12C13A14C15A16GC18A19C20GGGGGT
BE4max evoA evoFERNY ABE8e TadCBEa TadCBEb TadCBEc TadCBEd TadCBEe TadDE
C2 C6 C8 C12 C17 C18 C19 C20 插入 缺失
C2 C7 C15 C16 插入缺失
C1 C2 C4 C11 C13 C15 C18 C20 插入 缺失
SaCas9
SaCas9
SaCas9
具有C • G转换为G • C的DNA测序读段的百分比

图45(续)

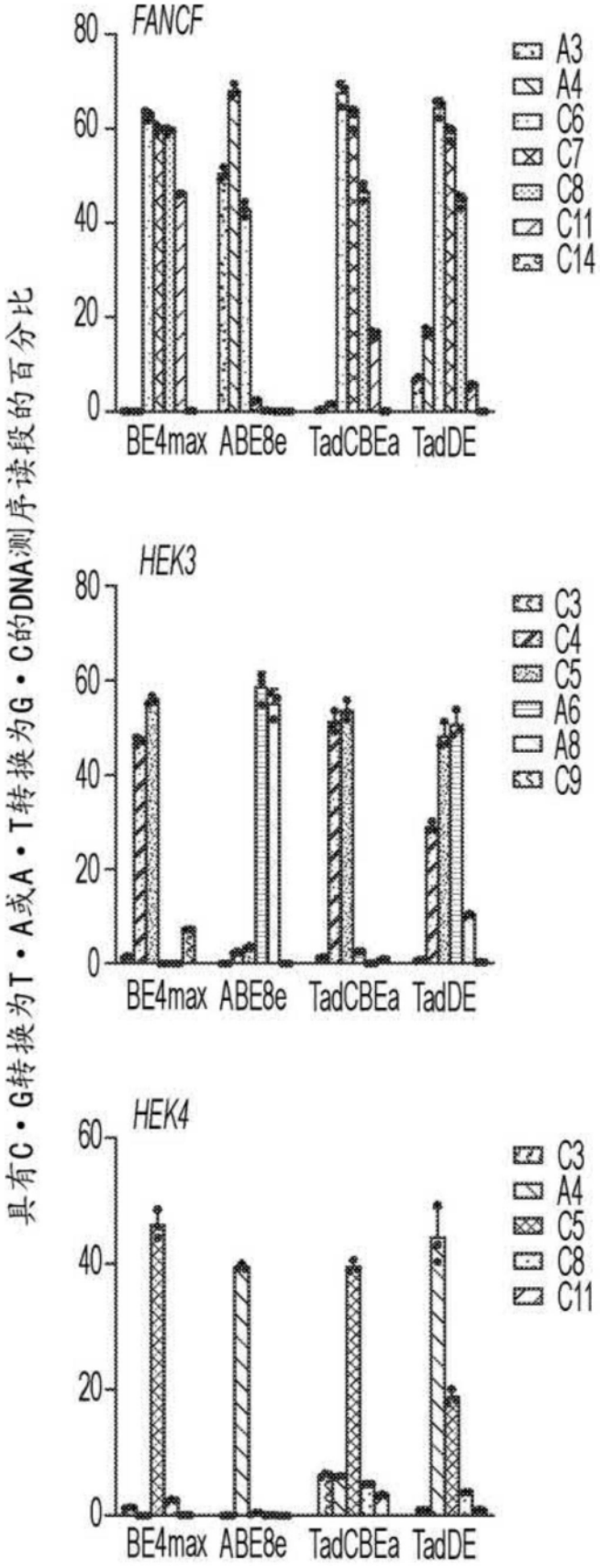

FANCF
80
60
40
20
0
BE4max ABE8e TadCBEa TadDE
A3
A4
C6
C7
C8
C11
C14
HEK3
C3
C4
C5
A6
A8
C9
HEK4
C3
A4
C5
C8
C11
具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比

图46

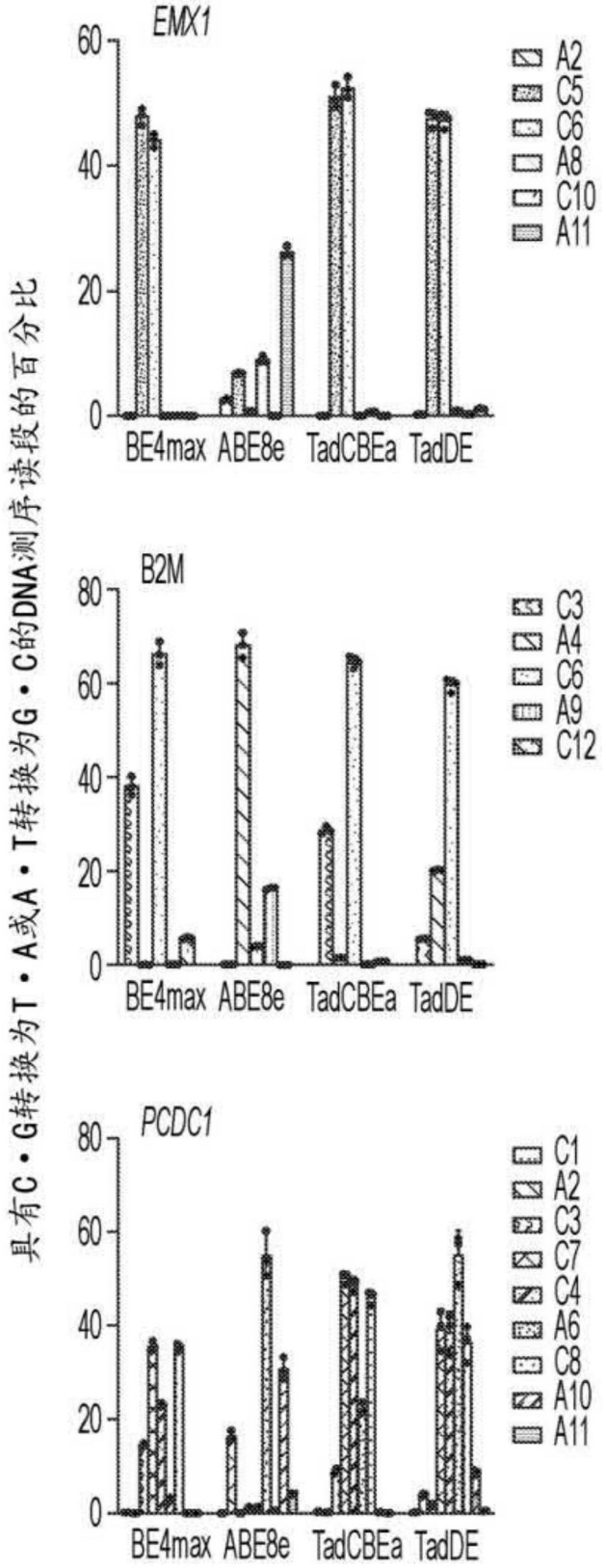

EMX1
60
40
20
0
BE4max ABE8e TadCBEa TadDE
A2
C5
C6
A8
C10
A11
B2M
80
60
40
20
0
BE4max ABE8e TadCBEa TadDE
C3
A4
C6
A9
C12
PCDC1
80
60
40
20
0
BE4max ABE8e TadCBEa TadDE
C1
A2
C3
C7
C4
A6
C8
A10
A11
具有C•G转换为T•A或A•T转换为G•C的DNA测序读段的百分比

图46(续)

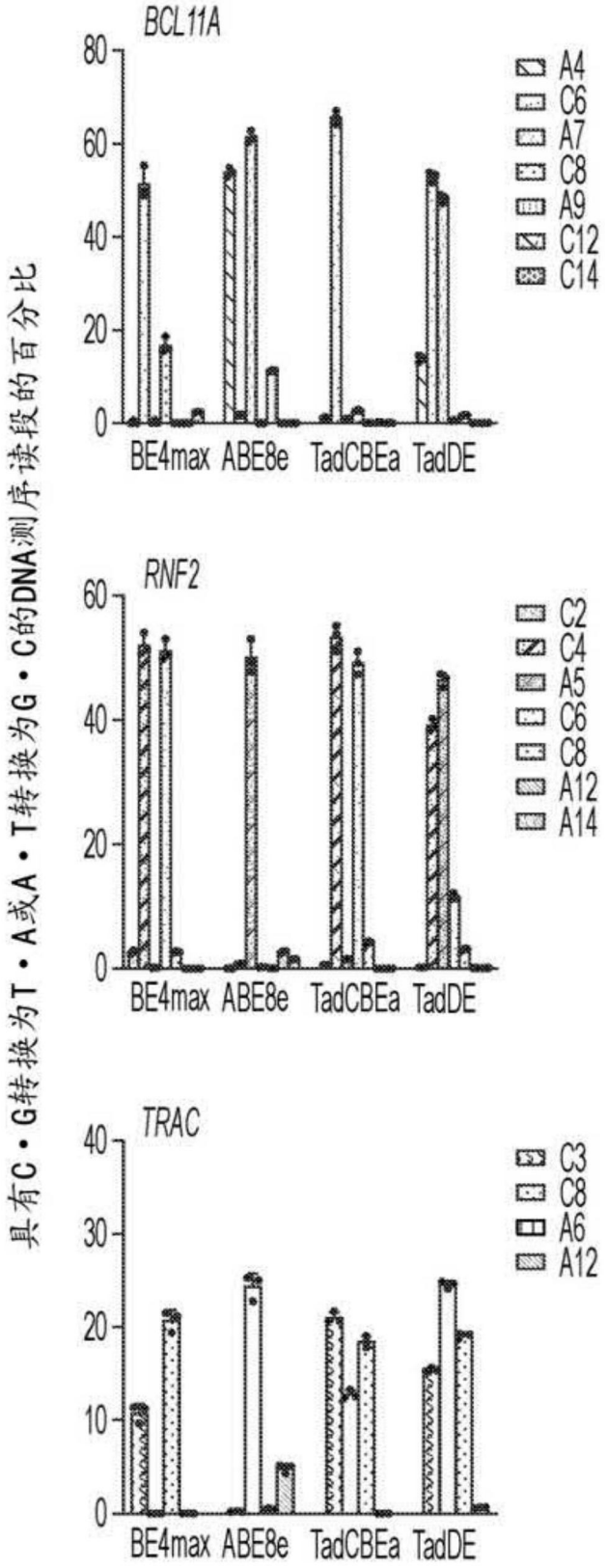

BCL11A
80
60
40
20
0
BE4max ABE8e TadCBEa TadDE
A4
C6
A7
C8
A9
C12
C14
RNF2
60
40
20
0
BE4max ABE8e TadCBEa TadDE
C2
C4
A5
C6
C8
A12
A14
TRAC
40
30
20
10
0
BE4max ABE8e TadCBEa TadDE
C3
C8
A6
A12
具有C • G转换为T • A或A • T转换为G • C的DNA测序读段的百分比

图46(续)

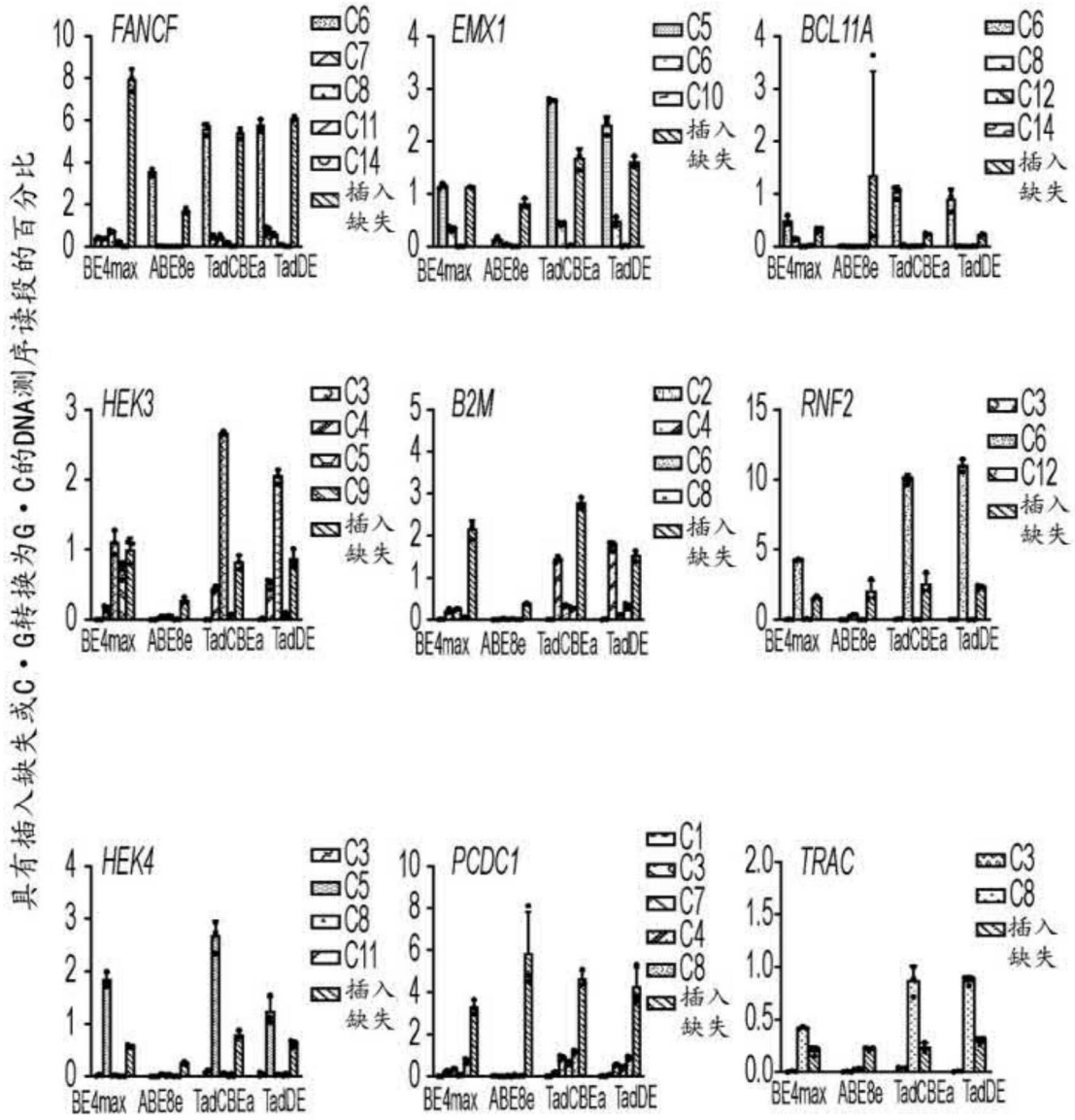

具有插入缺失或C·G转换为G·C的DNA测序读段的百分比
FANCF
C6
C7
C8
C11
C14
插入
缺失
BE4max ABE8e TadCBEa TadDE
EMX1
C5
C6
C10
插入
缺失
BCL11A
C6
C8
C12
C14
插入
缺失
HEK3
C3
C4
C5
C9
插入
缺失
B2M
C2
C4
C6
C8
插入
缺失
RNF2
C3
C6
C12
插入
缺失
HEK4
C3
C5
C8
C11
插入
缺失
PCDC1
C1
C3
C7
C4
C8
插入
缺失
TRAC
C3
C8
插入
缺失

图47

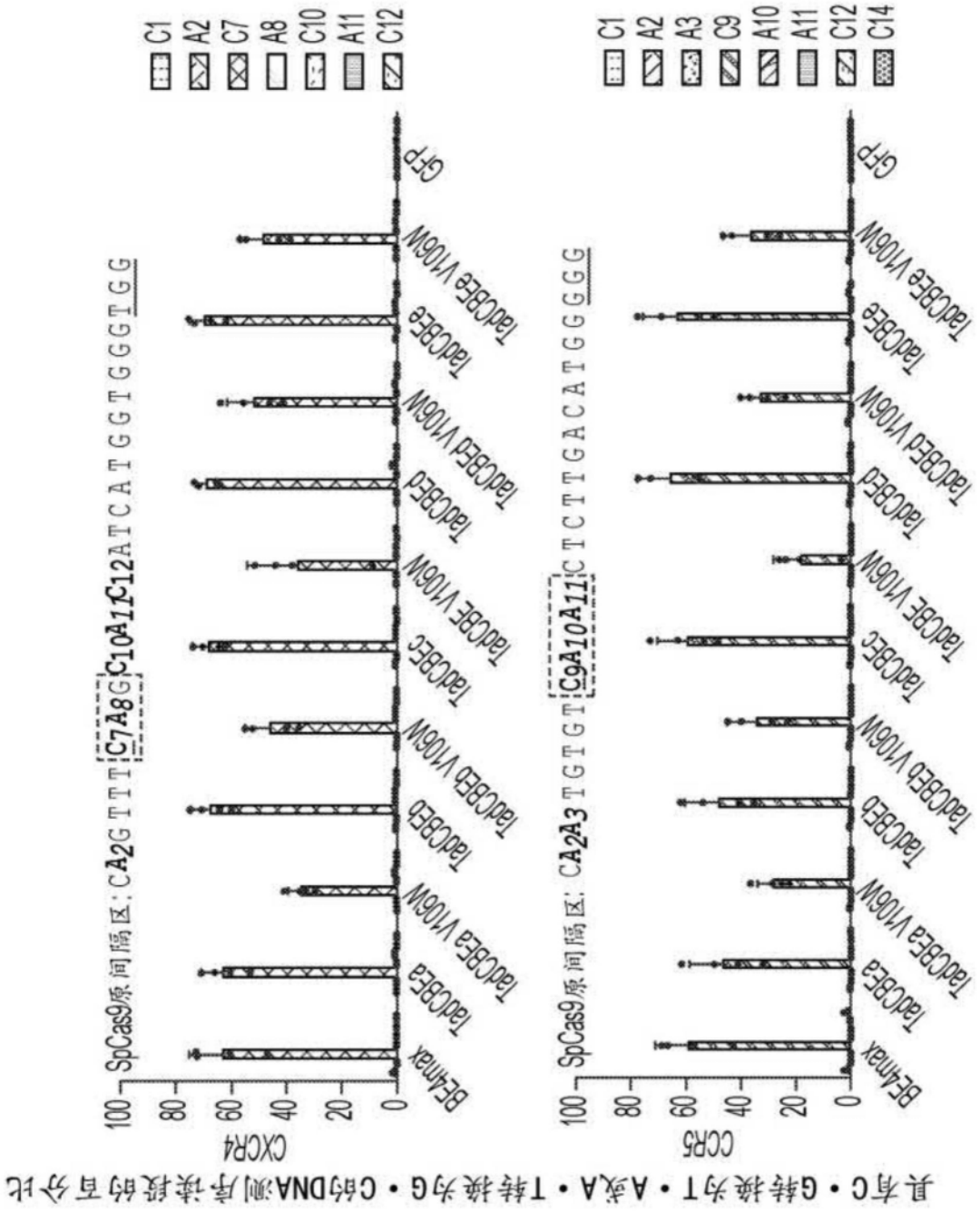

具有C·G转换为T·A或A·T转换为G·G的DNA测序读段的百分比
CXCR4
CCR5
SpCas9原间隔区:CA2GTTTIC7A8GC10A11C12ATCATGGTGGGTGG
SpCas9原间隔区:CA2A3TGTGTIC9A10A11ICTCTTGACATGGGG
C1 A2 C7 A8 C10 A11 C12
C1 A2 A3 C9 A10 A11 C12 C14
BE4max
TadCBEa
TadCBEa V106W
TadCBEd
TadCBEd V106W
TadCBEc
TadCBEc V106W
TadCBEb
TadCBEb V106W
TadCBEe
TadCBEe V106W
GFP
0 20 40 60 80 100

图48

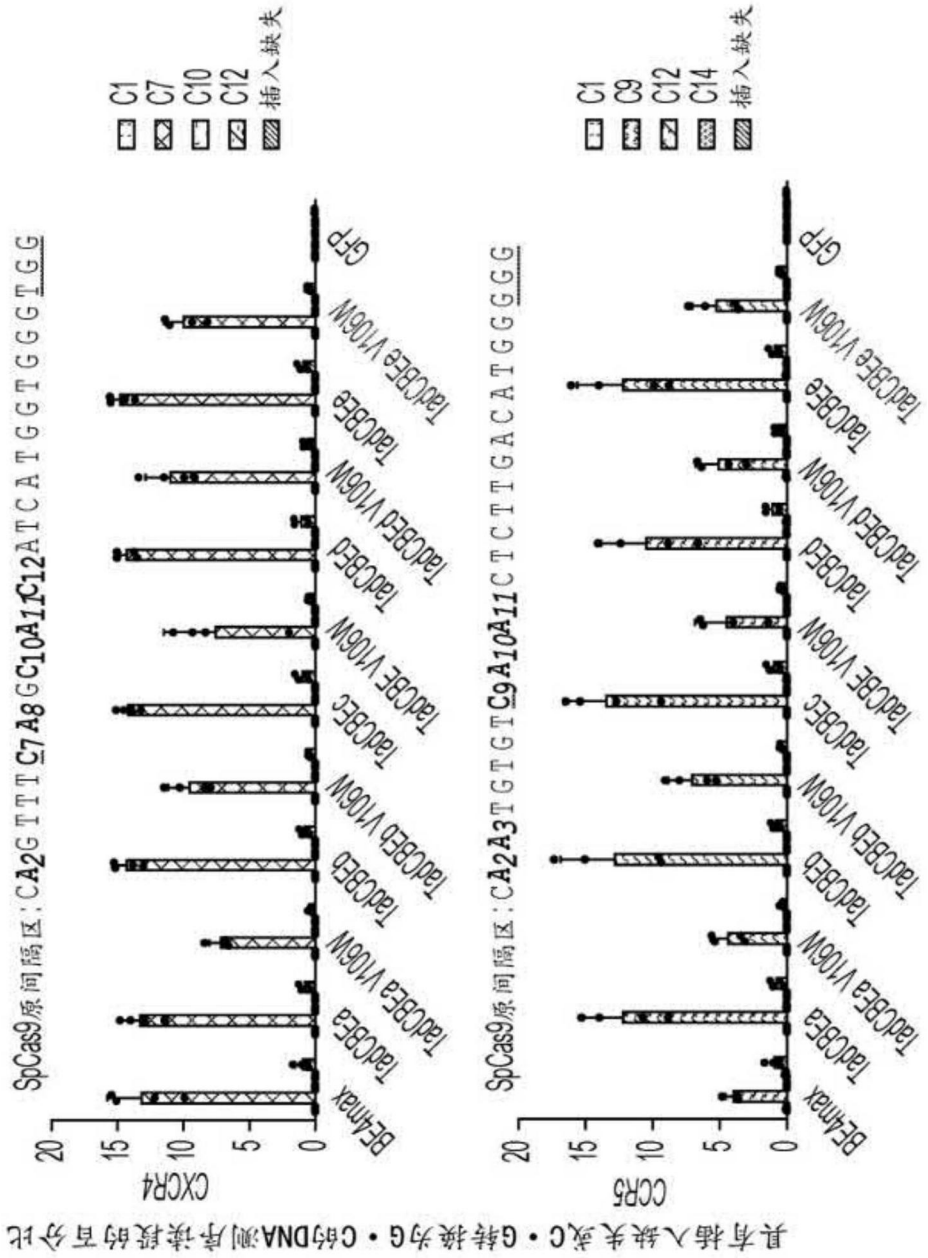

CXCR4
CCR5
SpCas9原间隔区：CA2G T T T C7 A8 G C10A11C12A T C A T G G T G G G T G G G
SpCas9原间隔区：CA2A3T G T G T C9A10A11C T C T T G A C A T G G G G G
C1
C7
C10
C12
插入缺失
C1
C9
C12
C14
插入缺失
BE4max
TadCBEa
TadCBEa V106W
TadCBEc
TadCBEd
TadCBEd V106W
TadCBEe
TadCBEe V106W
GFP
0
5
10
15
20
具有插入缺失或C • G转换为G • C的DNA测序读段的百分比

图49

C5
C7
A8
C10
A11
C12
A2
C9
A10
A11
C12
C14
GFP
TadCBEe V106W
TadCBEe
TadCBEd V106W
TadCBEd
TadCBE V106W
TadCBEc
TadCBEb V106W
TadCBEb
TadCBEa V106W
TadCBEa
BE4max
1.6
1.2
0.8
0.4
0.0
CXCR4 OT8
0.6
0.4
0.2
0.0
CCR5 OT1
具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比

图50

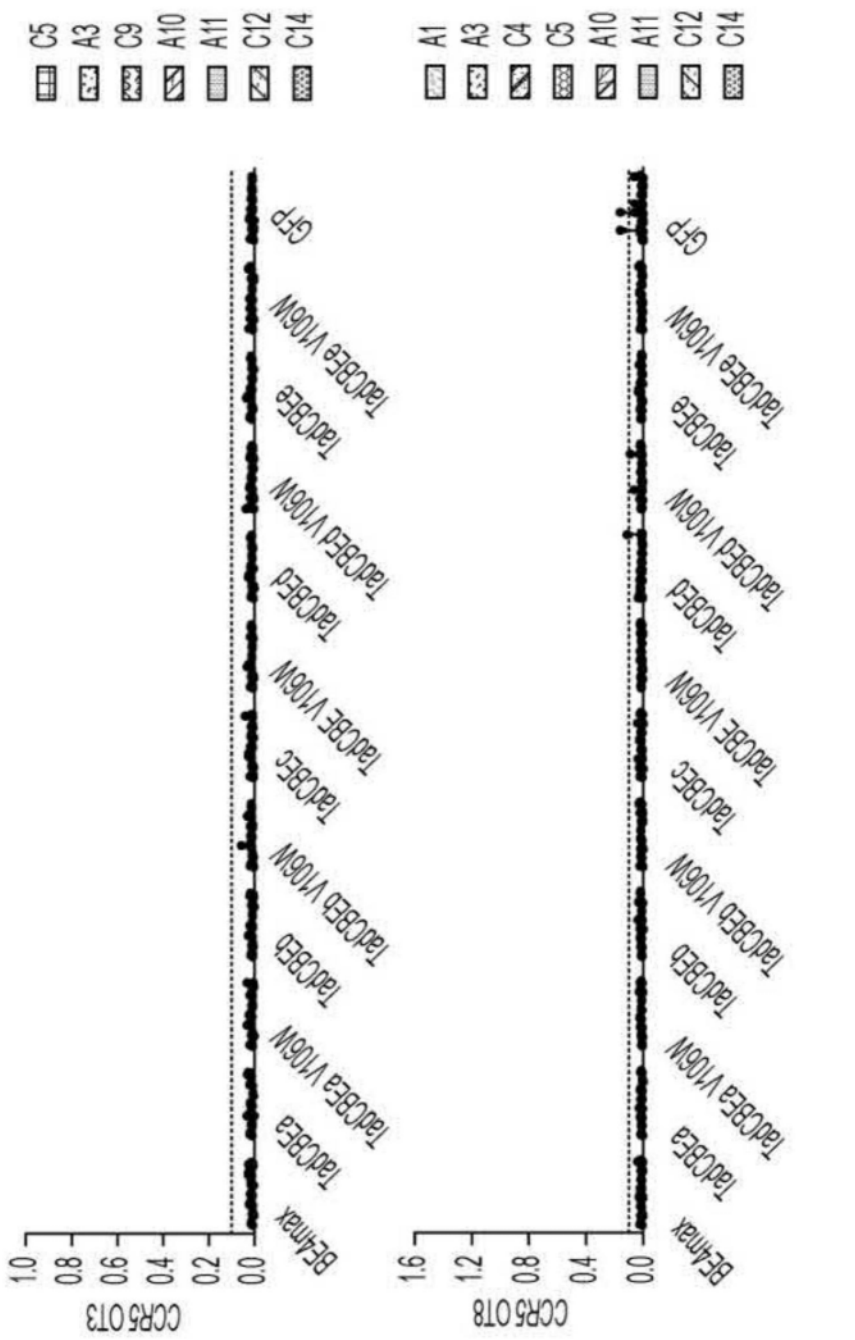

C5
A3
C9
A10
A11
C12
C14
A1
A3
C4
C5
A10
A11
C12
C14
GFP
TadCBEa V106W
TadCBEa
TadCBEd V106W
TadCBEd
TadCBEc V106W
TadCBEc
TadCBEb V106W
TadCBEb
BE4max
1.0
0.8
0.6
0.4
0.2
0.0
CCR5 OT3
1.6
1.2
0.8
0.4
0.0
CCR5 OT8
具有C·G转换为T·A或A·T转换为G·G的DNA测序读段的百分比

图50(续)

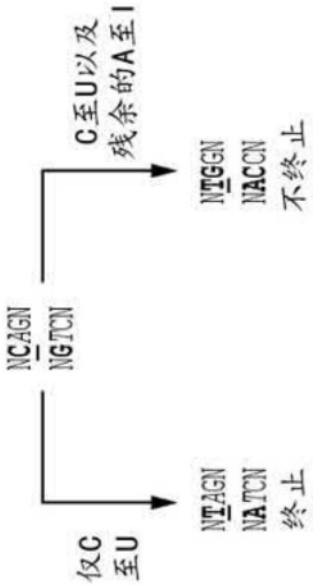

NCAGN
NGTCN
C至U以及
残余的A至I
NTGGN
NACCN
不终止
仅C
至U
NTAGN
NATCN
终止

图51A

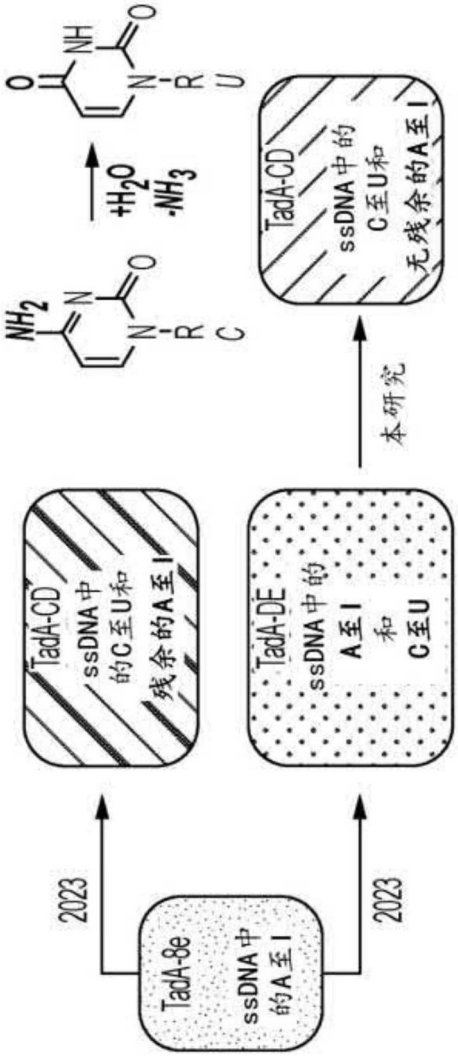

TadA-8e
ssDNA中
的A至I
2023
2023
TadA-CD
ssDNA中
的C至U和
残余的A至I
TadA-DE
ssDNA中的
A至I
和
C至U
本研究
TadA-CD
ssDNA中的
C至U和
无残余的A至I
+H2O
-NH3

图51B

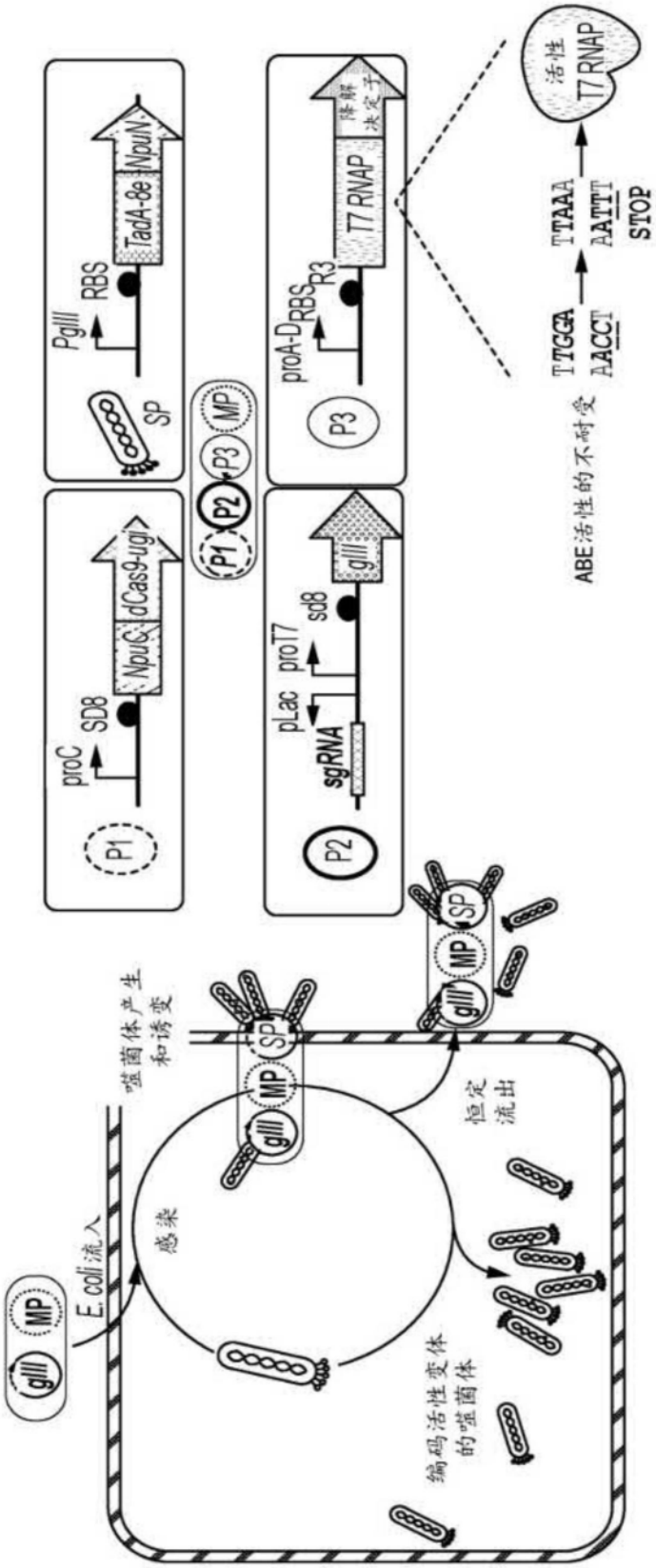

gIII
MP
E. coli 流入
感染
噬菌体产生和诱变
SP
gIII
MP
恒定流出
编码活性变体的噬菌体
SP
gIII
MP
P1
proC
SD8
NpuC
dCas9-ugi
P2
pLac
proT7
sd8
gIII
sgRNA
P3
MP
SP
PgIII
RBS
TadA-8e
NpuN
P3
proA-D
RBS
R3
T7 RNAP
降解决定子
TTGGA
AACCT
TTAAA
AATTT
STOP
ABE活性的不耐受
活性T7 RNAP

图51C

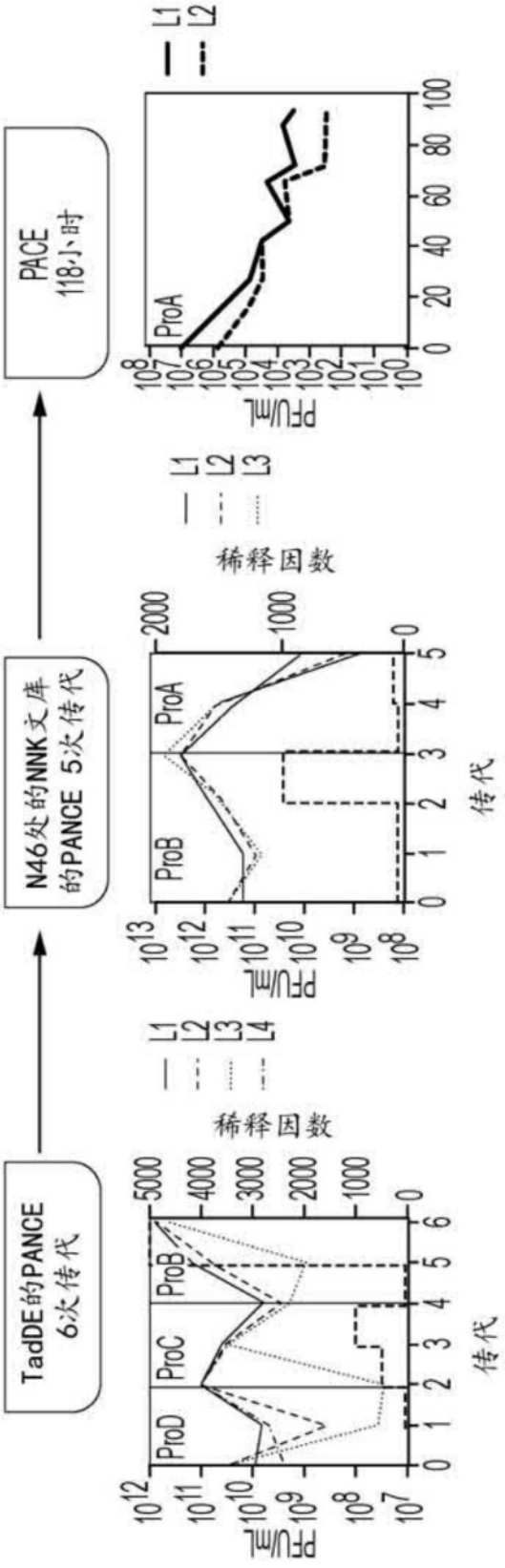

TadDE的PANCE
6次传代
N46处的NNK文库
的PANCE 5次传代
PACE
118小时
ProA
ProB
ProC
ProD
PFU/mL
传代
稀释因数
L1
L2
L3
L4

图51D

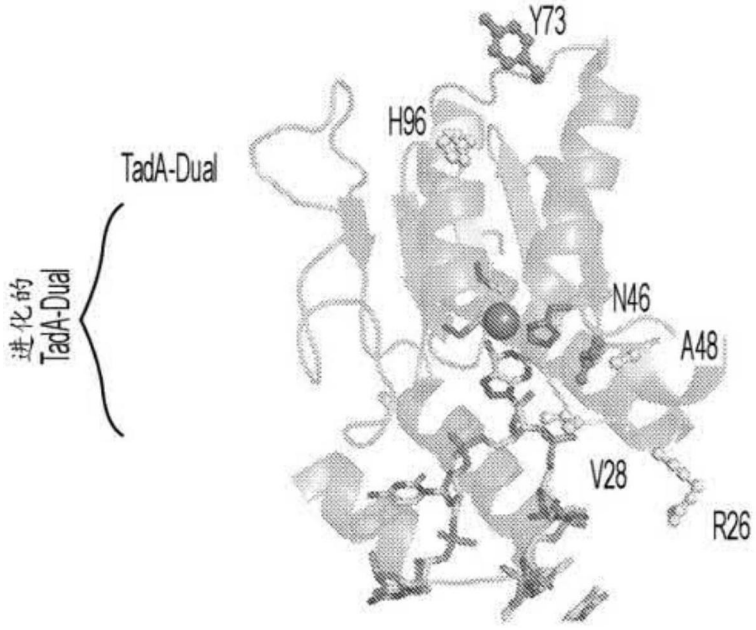


图51E

TadA-Dual	G26	A28	N46	R48	Q71	S73	N96	G105	H154	A162	39
TadA-CD-1	G	A	I	R	Q	P	N	G	Q	A	42
TadA-CD-2	G	A	T	R	Q	S	N	G	H	A	43
TadA-CD-3	G	A	T	R	Q	S	N	G	Q	A	44
TadA-CD-4	G	A	V	R	Q	S	N	G	Q	A	45
TadA-CD-5	G	A	V	R	Q	P	N	S	Q	A	46
TadA-CD-6	G	A	L	R	Q	P	N	G	Q	A	47
TadA-CD-7	R	A	L	P	Q	P	H	G	Q	A	48
TadA-CD-8	G	A	C	R	Q	S	H	G	Q	A	49
TadA-CD-9	G	A	V	R	Q	P	N	G	Q	A	50
TadA-CD-10	G	A	V	R	H	P	N	G	Q	A	51
TadA-CD-11	G	A	L	R	Q	S	N	G	Q	A	52
TadA-CD-12	G	A	C	R	Q	P	N	G	Q	A	53
TadA-CD-13	G	A	C	R	Q	P	N	G	Q	V	54

图52A

TadA-Dual	G26	A28	N46	R48	Q71	S73	N96	G105	H154
TadA-CD-1	G	A	I	R	Q	P	N	G	Q
TadA-CD-1	G	A	I	R	Q	P	N	G	Q
TadA-CD-2	G	A	T	R	Q	S	N	G	H
TadA-dual	G	A	N	R	Q	S	N	G	H
TadA-CD-14	G	A	I	R	Q	S	N	G	Q
TadA-CD-3	G	A	T	R	Q	S	N	G	Q
TadA-CD-15	G	A	N	R	S	S	N	G	Q

图52B

TadA-Dual	G26	A28	N46	R48	S73	N79	N96	G106
16	G	A	L	R	P	T	H	G
17	G	A	L	R	P	T	N	G
18	G	A	N	A	P	T	N	G
19	G	A	V	R	S	T	N	G
20	G	A	V	R	P	T	N	G
21	G	V	L	A	Y	T	H	G
22	G	A	I	R	P	T	N	G
23	G	A	V	R	P	T	N	S
24	G	A	N	P	H	P	N	G
25	G	V	I	A	Y	T	N	G

图52C

TadA-Dual	G26	A28	N46	R48	Q71	S73	N96	A162
26	G	A	V	R	Q	P	N	A
26	G	A	V	R	Q	P	N	A
27	G	A	L	R	Q	S	N	A
28	G	A	C	R	Q	Y	N	V
29	G	A	V	R	H	P	N	A
30	G	A	C	R	Q	P	N	A
30	G	A	C	R	Q	P	N	A
30	G	A	C	R	Q	P	N	A
30	G	A	C	R	Q	P	N	A
30	G	A	C	R	Q	P	N	A
31	G	A	C	R	Q	P	N	V
31	G	A	C	R	Q	P	N	V
31	G	A	C	R	Q	P	N	V
31	G	A	C	R	Q	P	N	V

图52D

TadA-Dual	G26	A28	L34	N46	R48	R64	S73	N96	A162
32	G	A	L	V	R	R	P	N	A
33	G	A	L	V	R	R	S	N	A
34	G	A	L	V	P	R	S	N	A
35	G	A	L	C	R	R	P	N	A
36	G	A	M	L	R	R	P	N	A
37	G	A	L	L	R	R	P	N	A
38	G	A	L	L	P	K	P	N	A

图52E

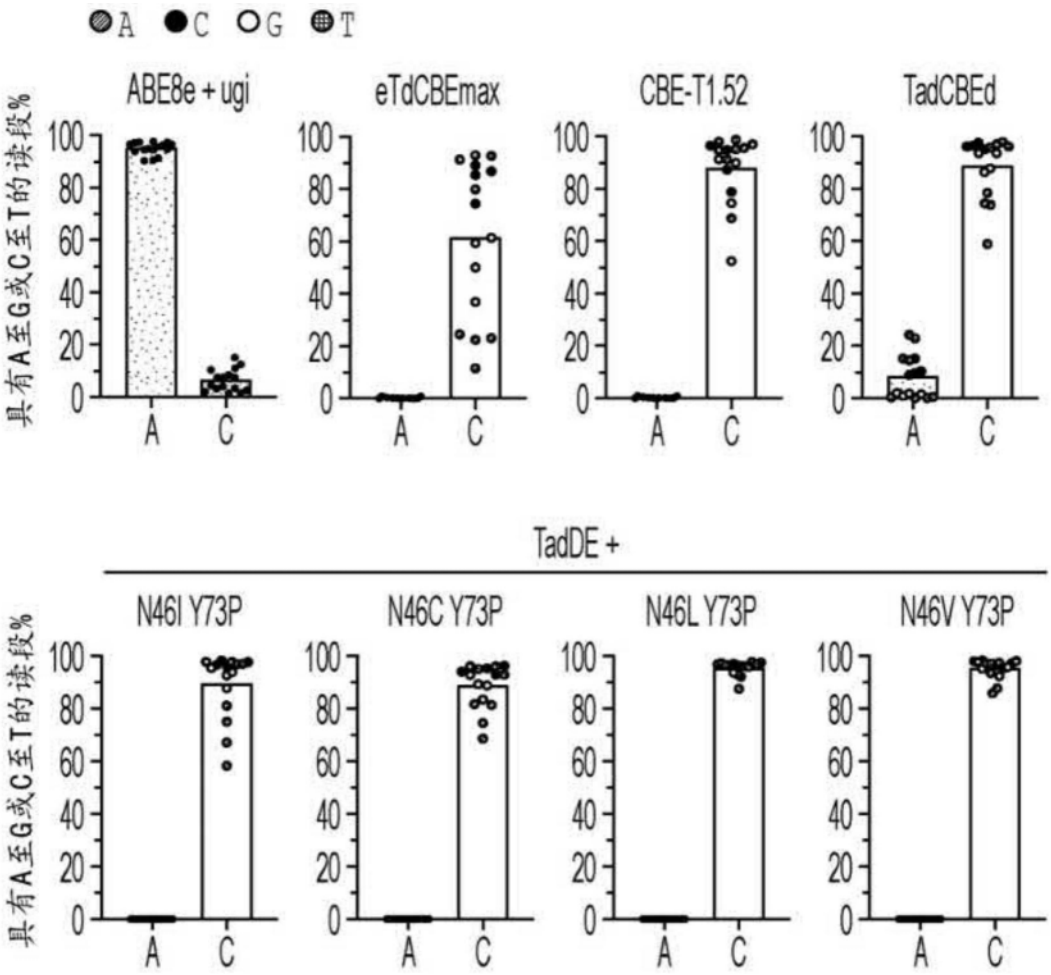

A C G T
ABE8e + ugi
eTdCBEmax
CBE-T1.52
TadCBEd
具有A至G或C至T的读段%
TadDE +
N46I Y73P
N46C Y73P
N46L Y73P
N46V Y73P
具有A至G或C至T的读段%
0 20 40 60 80 100
A C

图53

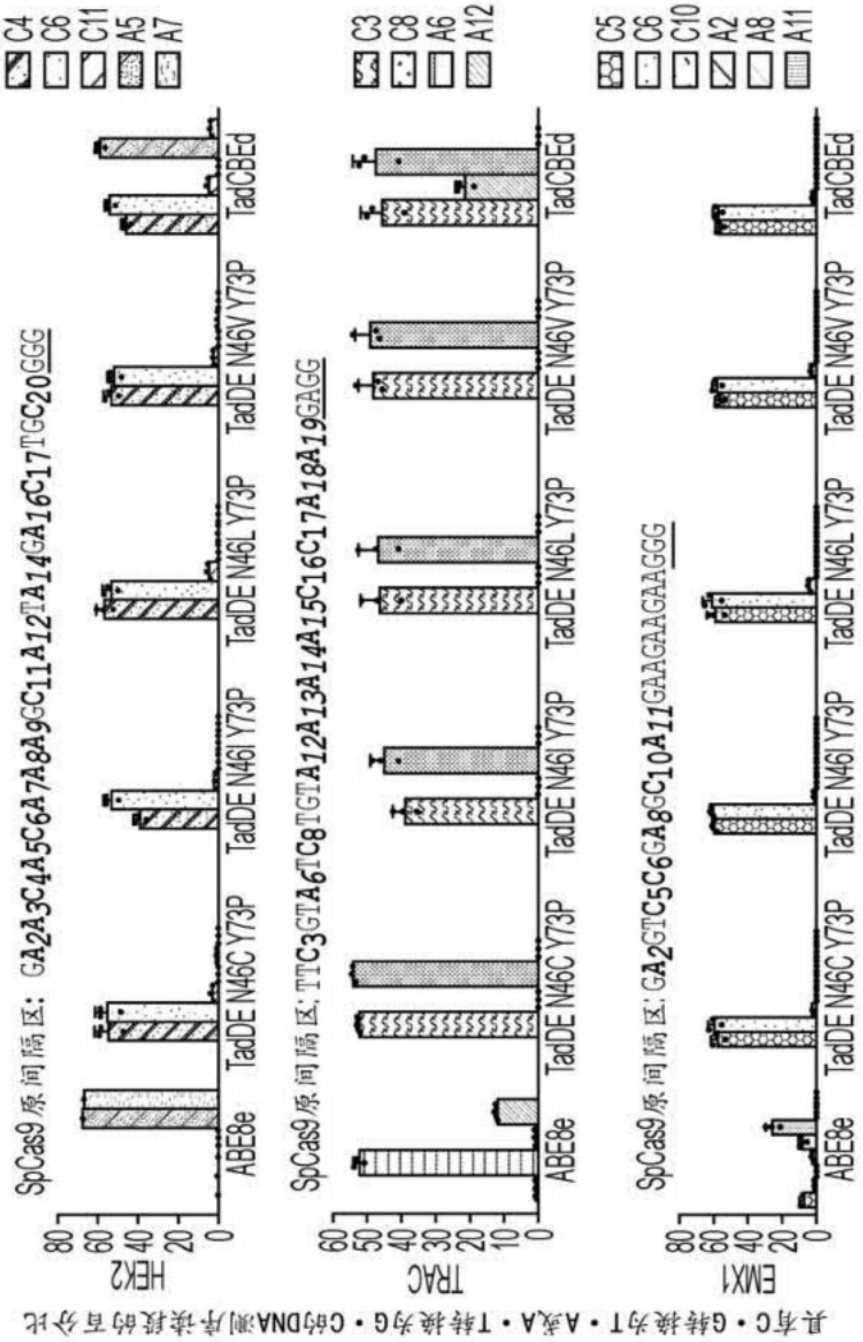

SpCas9 原间隔区: GA2A3C4A5C6A7A8A9GC11A12TA14GA16C17TGC20GGG
SpCas9 原间隔区: TTC3GTA6TC8TGTA12A13A14A15C16C17A18A19GAGG
SpCas9 原间隔区: GA2GTC5C6GA8GC10A11GAAGAAGAAGGG
C4 C6 C11 A5 A7
C3 C8 A6 A12
C5 C6 C10 A2 A8 A11
HEK2
TRAC
EMX1
ABE8e
TadDE N46C Y73P
TadDE N46I Y73P
TadDE N46L Y73P
TadDE N46V Y73P
TadCBEd
具有C・G转换为T・A或A・T转换为G・G的DNA测序读段的百分比

图54A

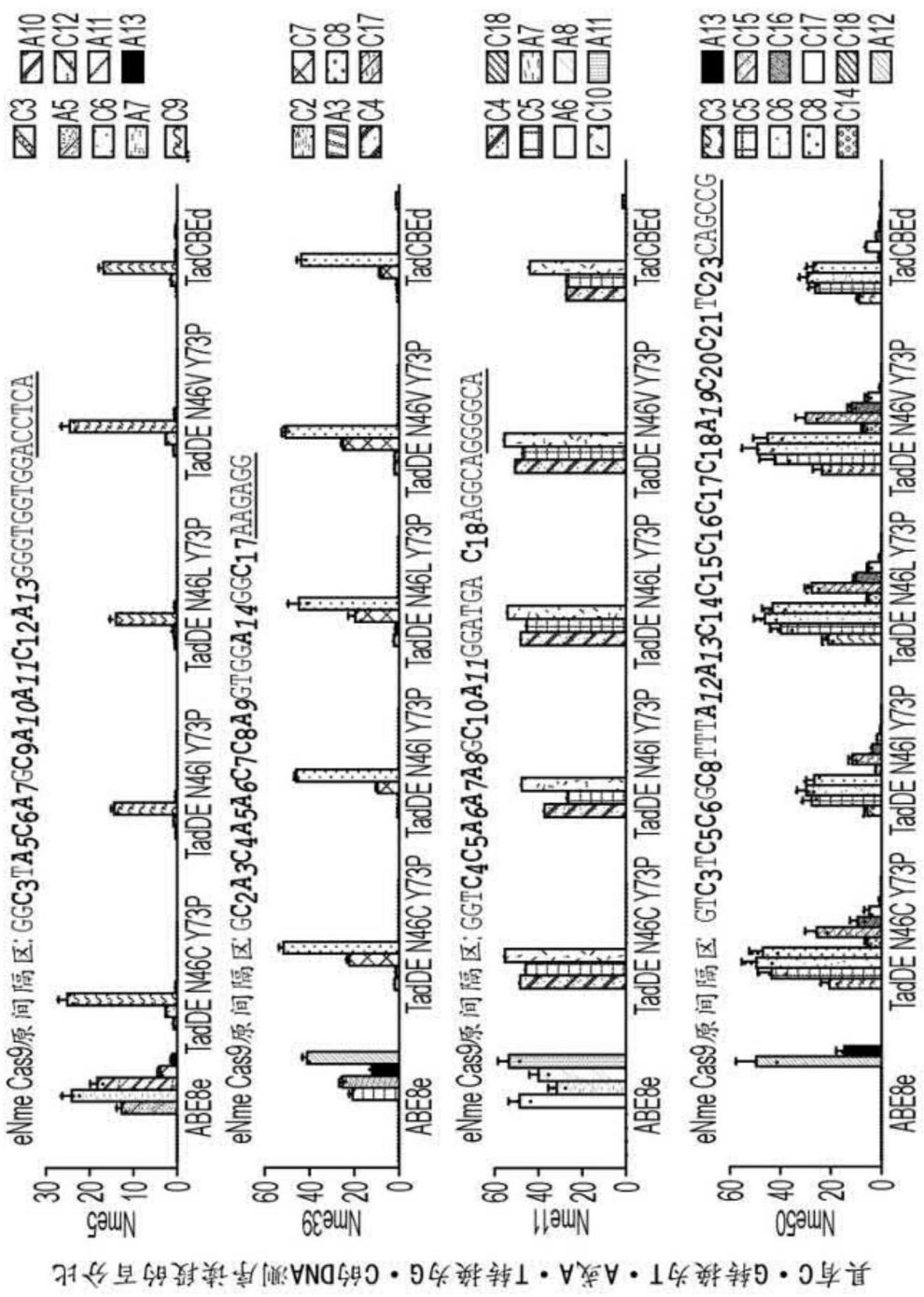

具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比
Nme5
Nme39
Nme11
Nme50
eNme Cas9原间隔区: GGC3TA5C6A7GC9A10A11C12A13GGGTGGTGGACCTCA
eNme Cas9原间隔区: GC2A3C4A5A6C7C8A9GTGGA14GGC17AAGAGG
eNme Cas9原间隔区: GGTC4C5A6A7A8GC10A11GGATGA C18AGGCAGGGGCA
eNme Cas9原间隔区: GTC3TC5C6GC8TTTA12A13C14C15C16C17C18A19C20C21TC23CAGCCG
ABE8e
TadDE N46C Y73P
TadDE N46I Y73P
TadDE N46L Y73P
TadDE N46V Y73P
TadCBEd
C3 A10
A5 C12
C6 A11
A7 A13
C9
C2 C7
A3 C8
C4 C17
C4 C18
C5 A7
A6 A8
C10 A11
C3 A13
C5 C15
C6 C16
C8 C17
C14 C18
A12

图54B

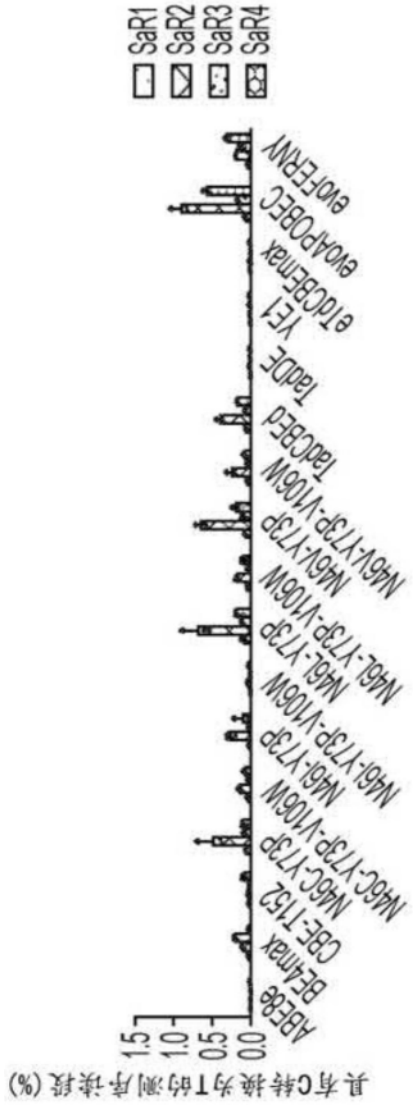

SaR1
SaR2
SaR3
SaR4
ABE8e
BE4max
CBE-T152
N46C-Y73P
N46C-Y73P-V106W
N46K-Y73P
N46K-Y73P-V106W
N46L-Y73P
N46L-Y73P-V106W
N46V-Y73P
N46V-Y73P-V106W
TadCBEcd
TadDE
YE1
etdCBEmax
evoAPOBEC
evoFERNY
0.0
0.5
1.0
1.5
具有C转换为T的测序读段(%)

图55A

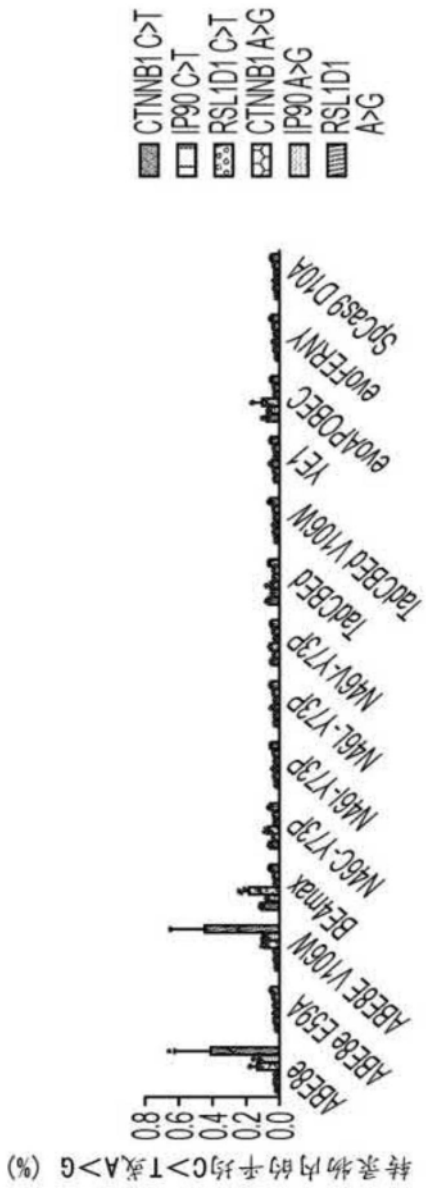

CTNNB1 C>T
IP90 C>T
RSL1D1 C>T
CTNNB1 A>G
IP90 A>G
RSL1D1 A>G
SpCas9 D10A
evoFERNV
evoAPOBEC
YE1
TadCBEd V106W
TadCBEd
N46V-Y73P
N46L-Y73P
N46I-Y73P
N46C-Y73P
BE4max
ABE8e V106W
ABE8e E59A
ABE8e
0.0
0.2
0.4
0.6
0.8
转录物内的平均C>T或A>G (%)

图55B

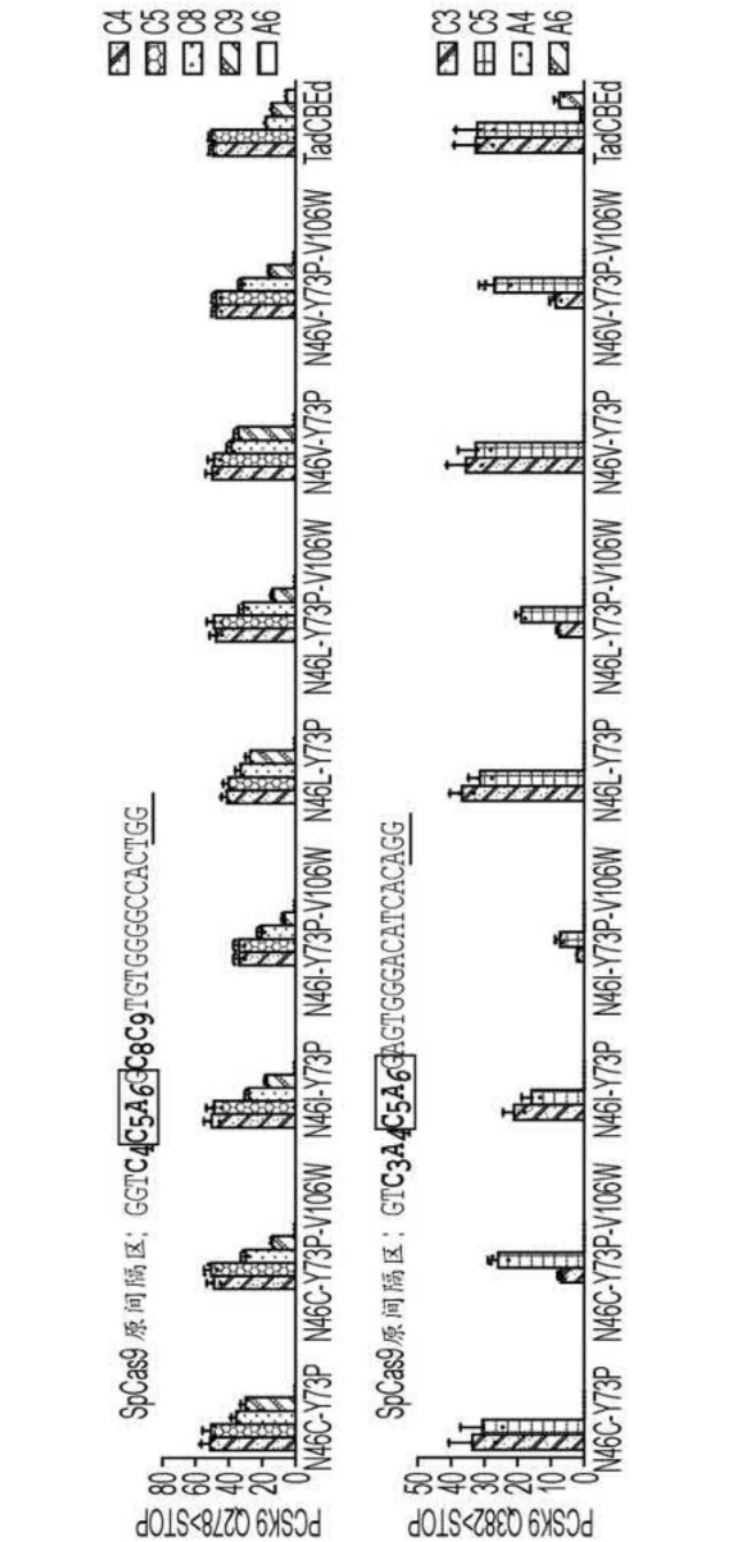

C4
C5
C8
C9
A6
SpCas9原间隔区：GGTC4C5A6C8C9TGTGGGGCCACTGG
PCSK9 Q278>STOP
80
60
40
20
0
N46C-Y73P
N46C-Y73P-V106W
N46I-Y73P
N46I-Y73P-V106W
N46L-Y73P
N46L-Y73P-V106W
N46V-Y73P
N46V-Y73P-V106W
TadCBEd
C3
C5
A4
A6
SpCas9原间隔区：GTC3A4C5A6GAGTGGGACATCACAGG
PCSK9 Q382>STOP
50
40
30
20
10
0
N46C-Y73P
N46C-Y73P-V106W
N46I-Y73P
N46I-Y73P-V106W
N46L-Y73P
N46L-Y73P-V106W
N46V-Y73P
N46V-Y73P-V106W
TadCBEd
具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比

图56

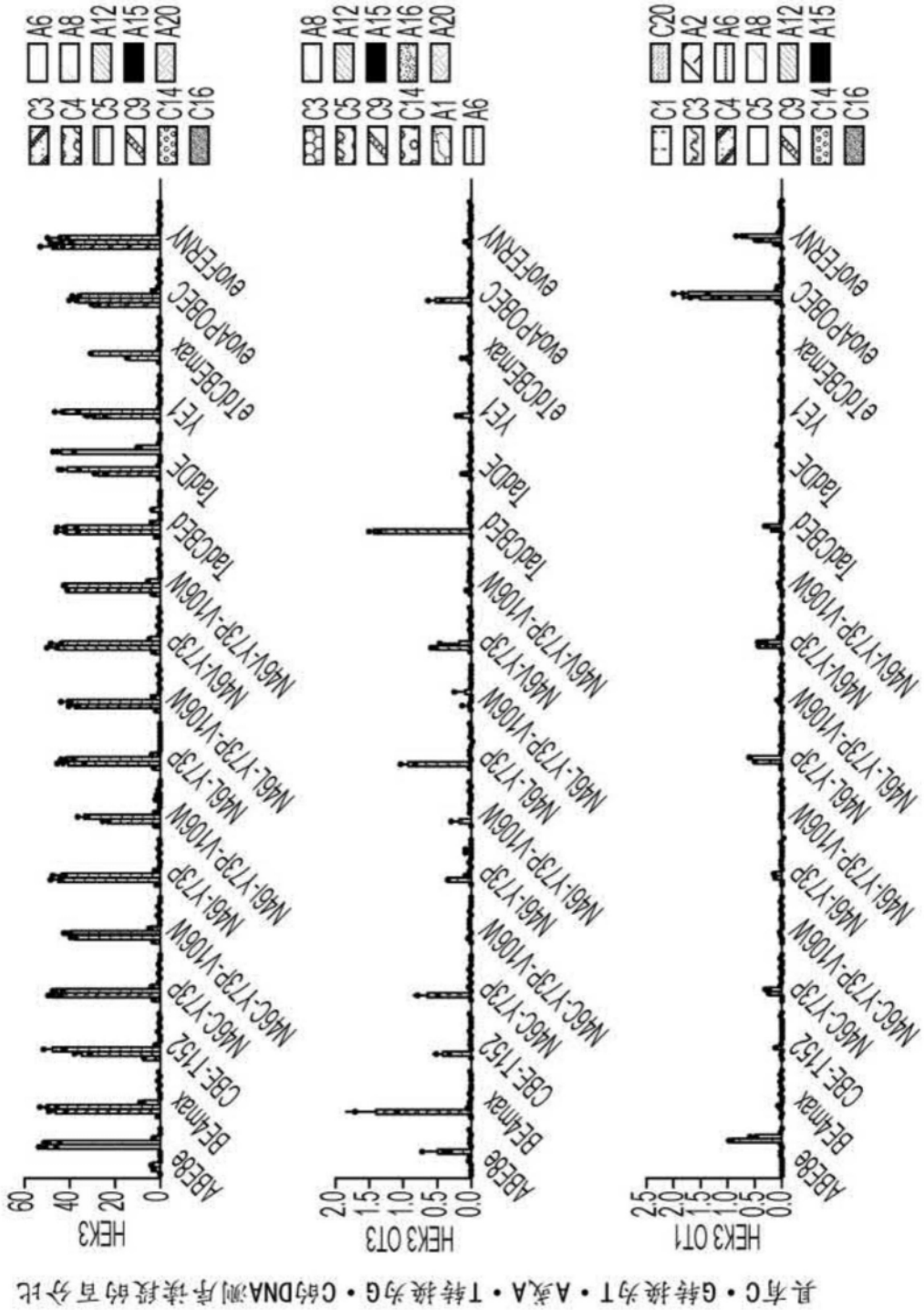

HEK3
HEK3 OT3
HEK3 OT1
具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比
ABE8e
BE4max
CBE-T152
N46C-Y73P
N46C-Y73P-V106W
N46L-Y73P
N46L-Y73P-V106W
N46V-Y73P
N46V-Y73P-V106W
TadCBEd
TadDE
YE1
etCBEmax
evoAPOBEC
evoFERNY

图57

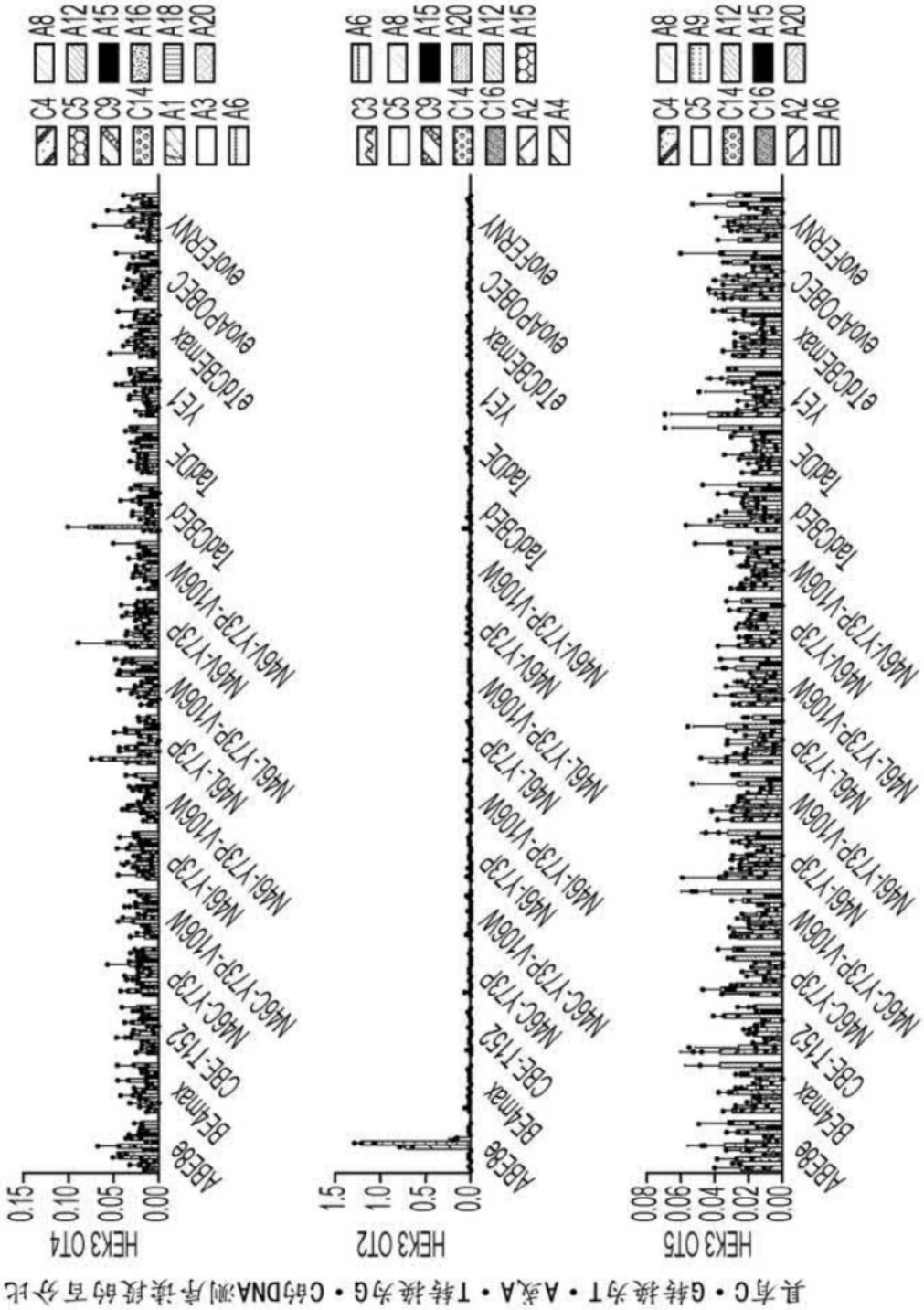

HEK3 OT4
HEK3 OT2
HEK3 OT5
具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比
ABE8e
BE4max
CBE-T152
M46C-Y73P
M46C-Y73P-V106W
M46I-Y73P
M46I-Y73P-V106W
M46L-Y73P
M46L-Y73P-V106W
M46V-Y73P
M46V-Y73P-V106W
TadCBEd
TadDE
YE1
eTdCBEmax
evoAPOBEC
evoFERNY

图57(续)

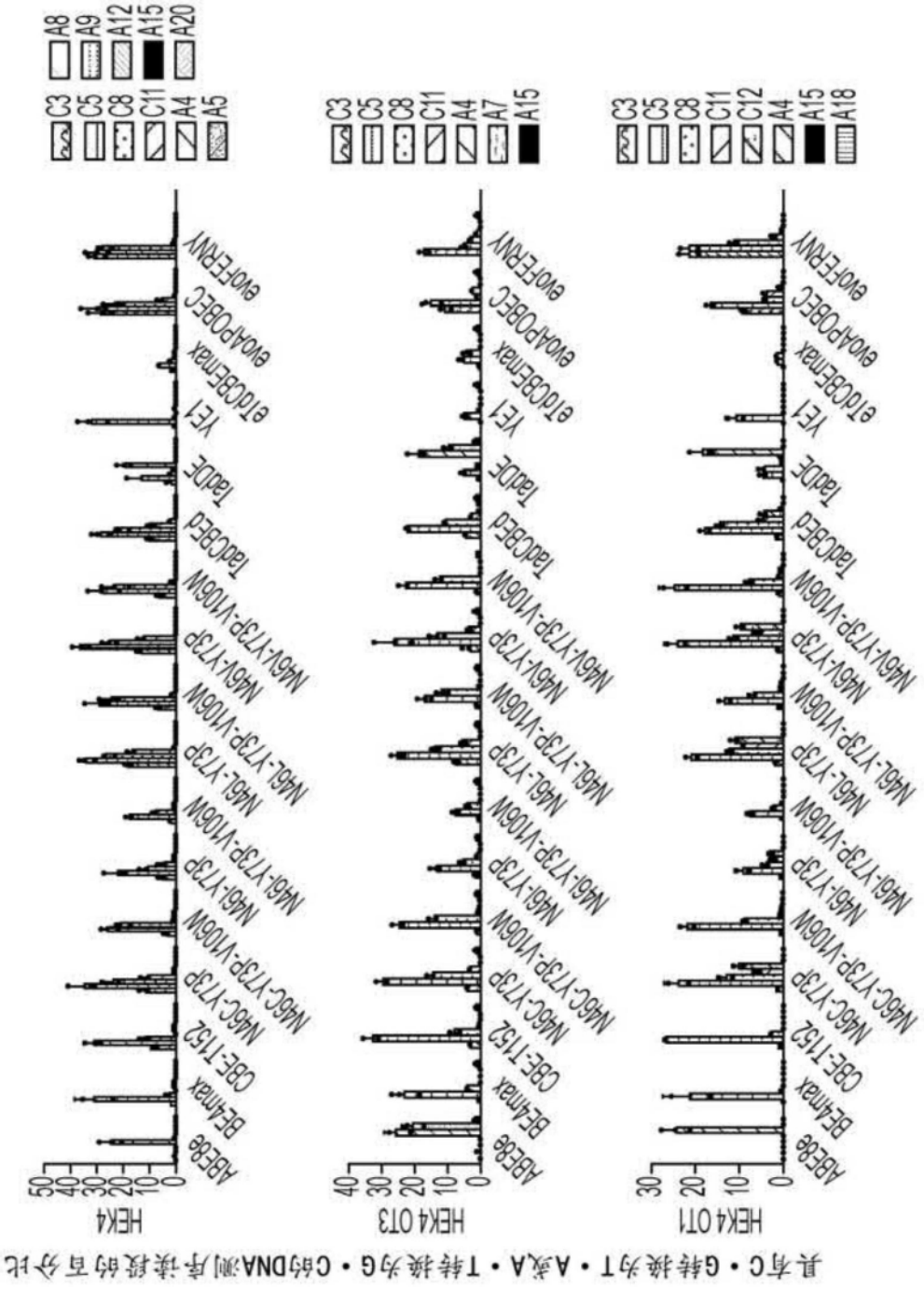

A8
A9
A12
A15
A20
C3
C5
C8
C11
A4
A5
C3
C5
C8
C11
A4
A7
A15
C3
C5
C8
C11
C12
A4
A15
A18
HEK4
HEK4 OT3
HEK4 OT1
具有C·G转换为T·A或A·T转换为G·G的DNA测序读段的百分比

图58

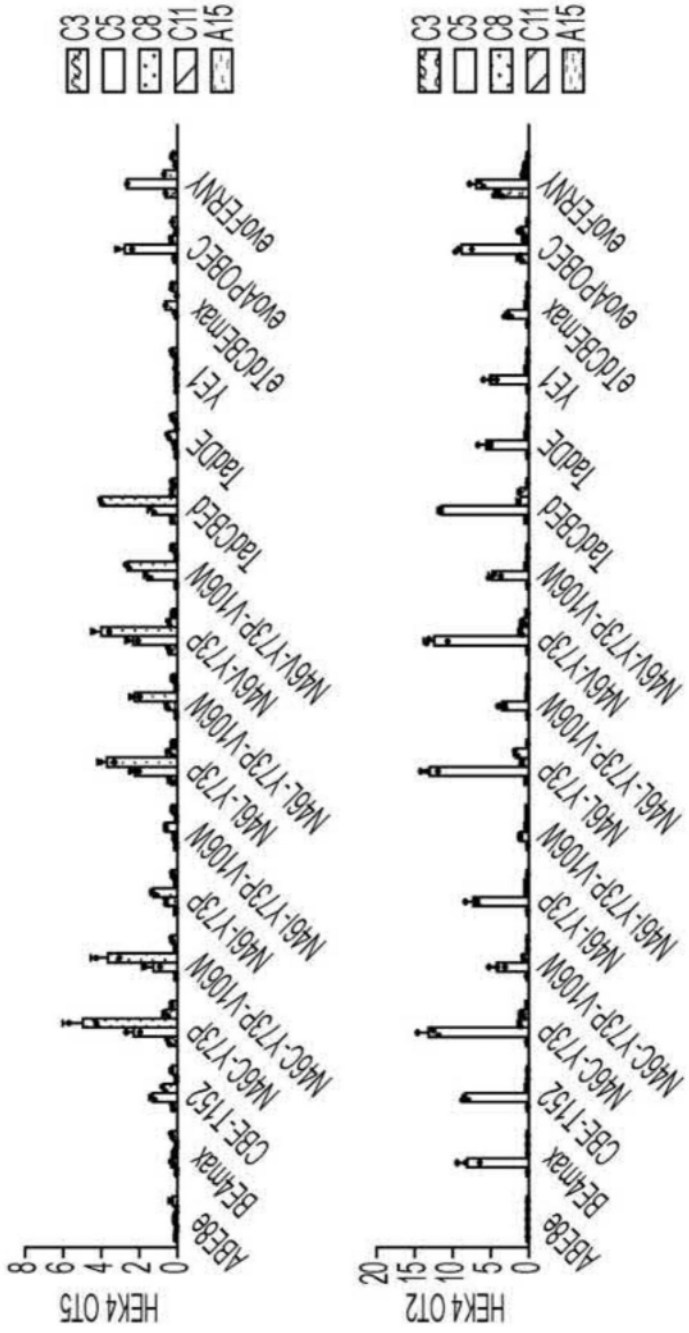

C3
C5
C8
C11
A15
HEK4 OT5
HEK4 OT2
ABC99
BE4max
CBE-T1s2
N4gC-Y73P
N4gC-Y73P-V106W
N4gI-Y73P
N4gI-Y73P-V106W
N4gL-Y73P
N4gL-Y73P-V106W
N4gV-Y73P
N4gV-Y73P-V106W
TadCBEd
TadDE
YE1
etdCBEmax
evoAPOBEC
evoFERNY
具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比

图58(续)

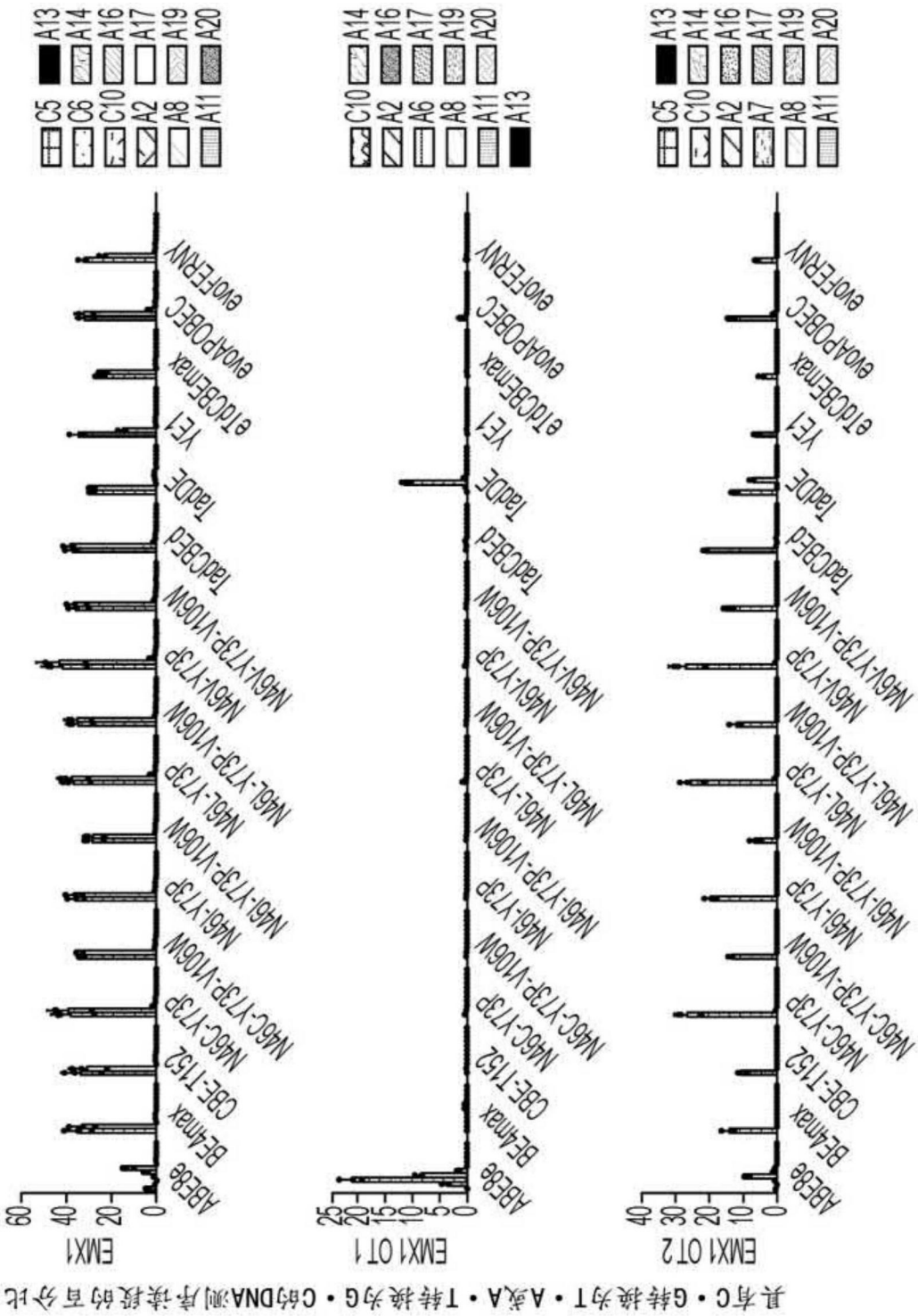

A13
A14
A16
A17
A19
A20
C5
C6
C10
A2
A8
A11
EMX1
EMX1 OT1
EMX1 OT2
ABE8e
BE4max
CBE-T152
N46C-Y73P
N46C-Y73P-V106W
N46I-Y73P
N46I-Y73P-V106W
N46L-Y73P
N46L-Y73P-V106W
N46V-Y73P
N46V-Y73P-V106W
TadCBEd
TadDE
YE1
etdCBEmax
evoAPOBEC
evoFERNY
具有C·G转换为T·A或A·T转换为G·C的DNA测序读段的百分比

图59

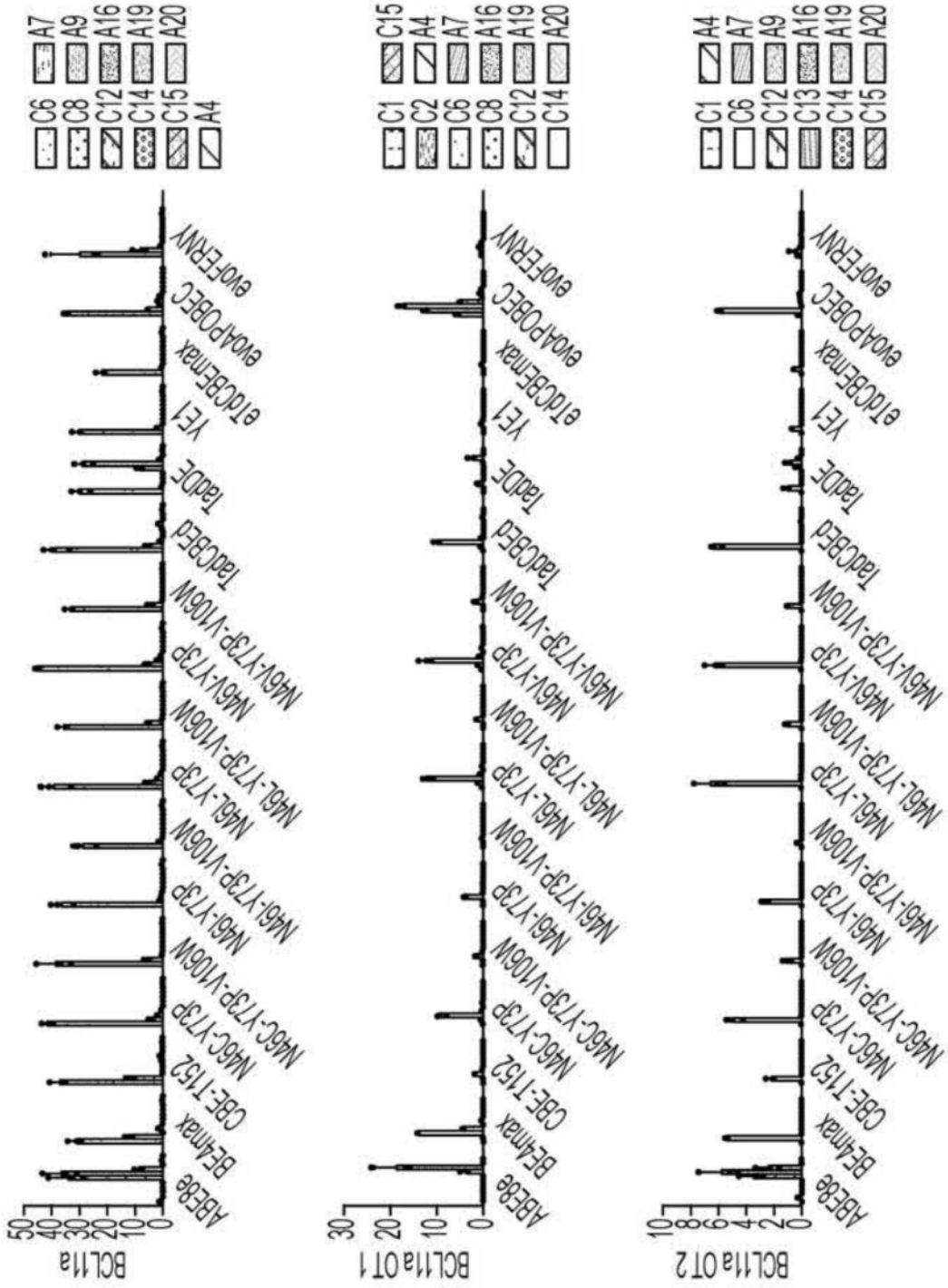

C6 C8 C12 C14 C15 A4
A7 A9 A16 A19 A20
C1 C2 C6 C8 C12 C14
C15 A4 A7 A16 A19 A20
C1 C6 C12 C13 C14 C15
A4 A7 A9 A16 A19 A20
ABC88
BE4max
CBE-T1S2
N46C-Y73P
N46C-Y73P-V106W
N46I-Y73P
N46I-Y73P-V106W
N46L-Y73P
N46L-Y73P-V106W
N46V-Y73P
N46V-Y73P-V106W
TadCBEd
TadDE
YE1
etdCBEmax
evoAPOBEC
evoFERNY
BCL11a
BCL11a OT1
BCL11a OT2
具有C·G转换为T·A或A·T转换为G·G的DNA测序读段的百分比

图60

C5
C6
C10
A2
A8
A11
A13
A14
A16
A17
A19
A20
ABE8e
BE4max
CBE-T1S2
N46C-Y73P
N46C-Y73P-V106W
N46L-Y73P
N46L-Y73P-V106W
N46V-Y73P
N46V-Y73P-V106W
TadCBEd
TadDE
YE1
etdCBEmax
evoAPOBEC
evoFERNY
EMX 1 for SaR1
EMX 1 for SaR3
具有C·G转换为T·A或A·T转换
为G·C的DNA测序读段的百分比

图61

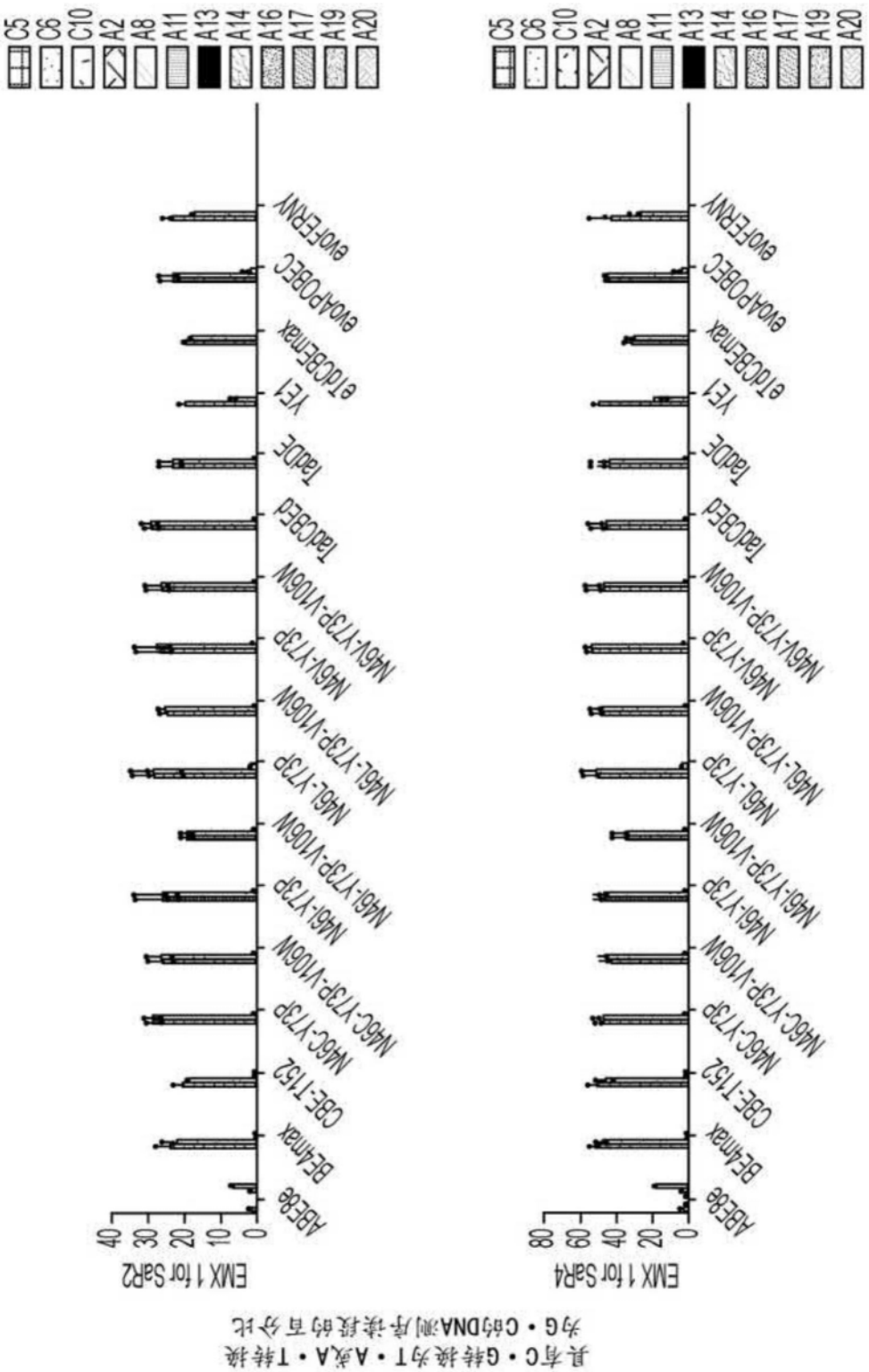

C5
C6
C10
A2
A8
A11
A13
A14
A16
A17
A19
A20
ABE8e
BE4max
CBE-T1S2
M46C-Y73P
M46C-Y73P-V106W
M46I-Y73P
M46I-Y73P-V106W
M46L-Y73P
M46L-Y73P-V106W
M46V-Y73P
M46V-Y73P-V106W
TadCBEd
TadDE
YE1
eTdCBEmax
evoAPOBEC
evoFERNY
EMX 1 for SaR2
EMX 1 for SaR4
具有C·G转换为T·A或A·T转换
为G·C的DNA测序读段的百分比

图61(续)

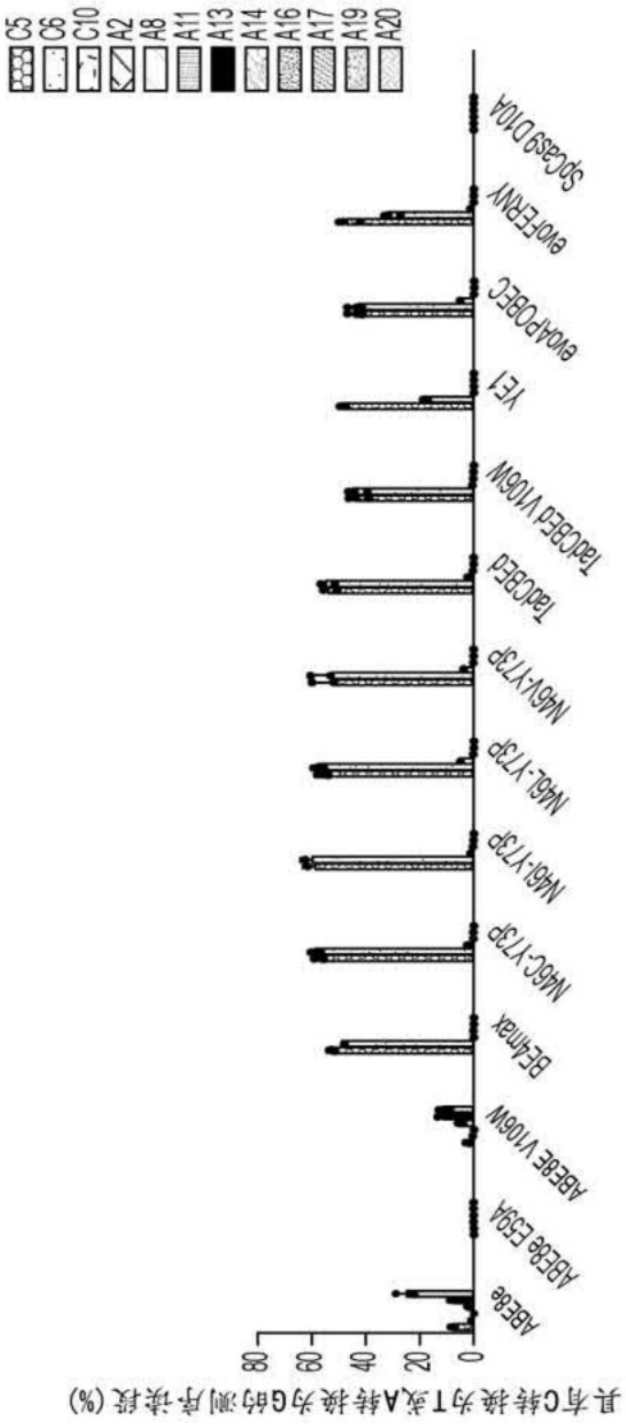

C5
C6
C10
A2
A8
A11
A13
A14
A16
A17
A19
A20
SpCas9 D10A
evoFERNV
evoAPOBEC
VE1
TadCBEd V106W
TadCBEd
Nd6V-Y73P
Nd6L-Y73P
Nd6I-Y73P
Nd6C-Y73P
BE4max
ABE8e V106W
ABE8e E59A
ABE8e
0
20
40
60
80
具有C转换为T或A转换为G的测序读段(%)

图62

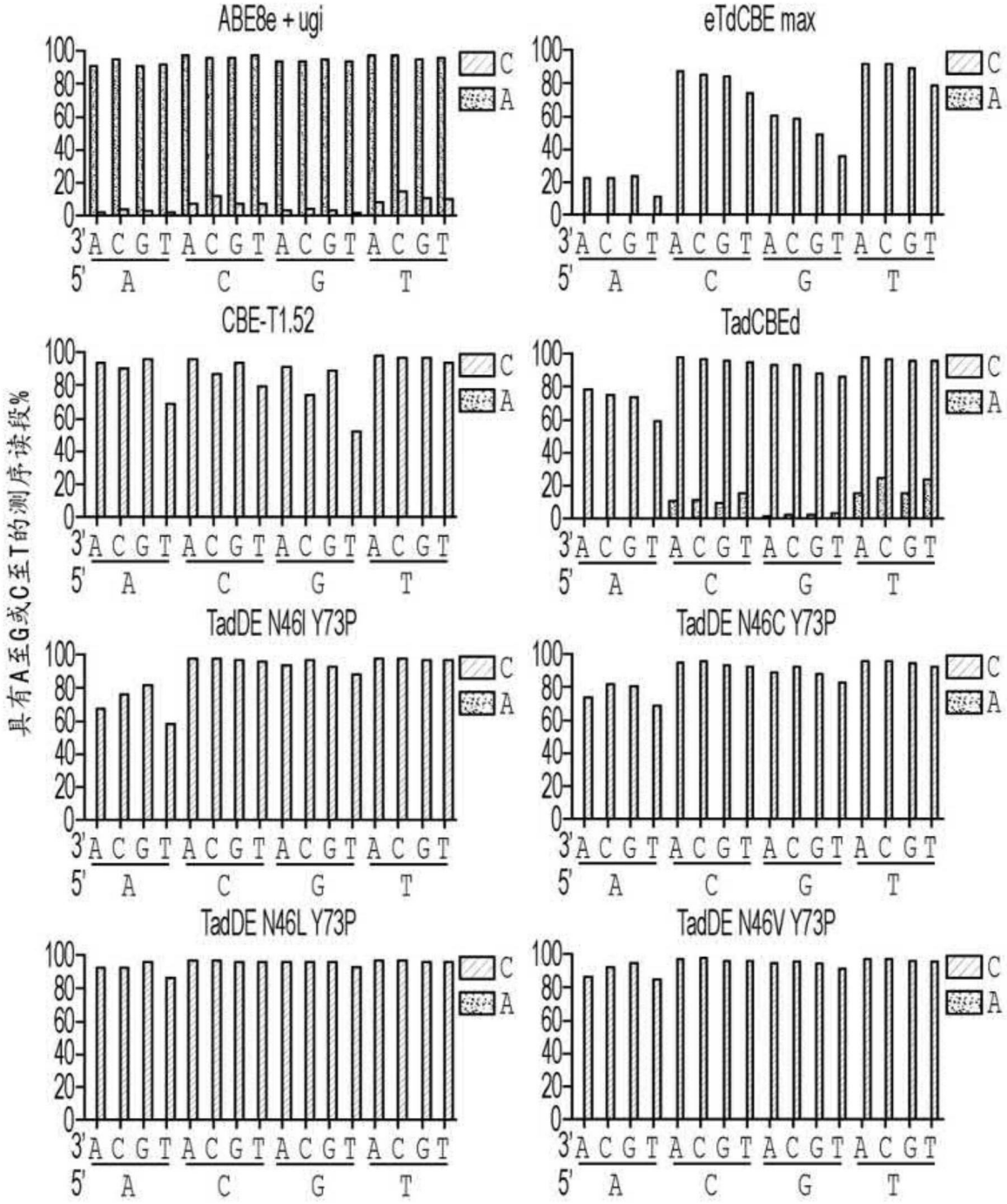

ABE8e + ugi
eTdCBE max
CBE-T1.52
TadCBEd
TadDE N46I Y73P
TadDE N46C Y73P
TadDE N46L Y73P
TadDE N46V Y73P
具有A至G或C至T的测序读段%
C
A
3' A C G T A C G T A C G T A C G T
5' A C G T

图63