(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114999567 A
(43)申请公布日 2022.09.02

(21)申请号 202210543544.7
(22)申请日 2016.07.25
(30)优先权数据
62/196,250 2015.07.23 US
62/294,948 2016.02.12 US
(66)本国优先权数据
PCT/CN2016/073753 2016.02.14 CN
(62)分案原申请数据
201680051235.0 2016.07.25
(71)申请人 香港中文大学
地址 中国香港新界
(72)发明人 卢煜明 赵慧君 陈君赐 江培勇
(74)专利代理机构 北京英赛嘉华知识产权代理有限责任公司 11204
专利代理师 王达佐 洪欣
(51)Int.Cl.
G16B 20/00(2019.01)
G16B 20/10(2019.01)
G16B 20/20(2019.01)
G16B 25/00(2019.01)
G16B 25/10(2019.01)
G16B 30/00(2019.01)
G16B 30/10(2019.01)
C12Q 1/6827(2018.01)
C12Q 1/6869(2018.01)
C12Q 1/6883(2018.01)
C12Q 1/6886(2018.01)

权利要求书3页 说明书48页 附图195页

(54)发明名称

游离DNA的片段化模式的分析

(57)摘要

描述影响游离DNA(例如血浆DNA)的片段化模式的因素和游离DNA片段化模式分析的应用，包括在分子诊断学中的应用。各种应用可以使用片段化模式的特性来确定特定组织类型的比例贡献，以确定特定组织类型(例如母体样品中的胎儿组织或来自癌症患者的样品中的肿瘤组织)的基因型，和/或识别特定组织类型的优选的结束位置，其然后可以用于确定特定组织类型的比例贡献。

CN 114999567 A

1.分析生物样品以确定来自包括第一组织类型的多个组织类型的游离DNA分子的混合物中的所述第一组织类型的比例贡献的分类的方法,所述生物样品包括所述混合物,所述方法包括:

识别至少一个具有所述第一组织类型特异性片段化模式的基因组区域;

分析来自所述生物样品的多个游离DNA分子,其中分析游离DNA分子包括:

确定对应于所述游离DNA分子的至少一个末端的参考基因组中的基因组位置;

识别第一基因组位置的第一集合,每个第一基因组位置具有对应于所述第一基因组位置的游离DNA分子的末端的局部最小值;

识别第二基因组位置的第二集合,每个第二基因组位置具有对应于所述第二基因组位置的游离DNA分子的末端的局部最大值;

确定结束于所述至少一个基因组区域的任一个中的所述第一基因组位置的任一个上的游离DNA分子的第一数目;

确定结束于所述至少一个基因组区域的任一个中的所述第二基因组位置的任一个上的游离DNA分子的第二数目;

使用所述第一数目和所述第二数目计算分离值;以及

通过比较所述分离值与一个或多个校准值确定所述第一组织类型的所述比例贡献的分类,所述校准值由一个或多个已知所述第一组织类型的比例贡献的校准样品确定。

2.分析生物样品以确定第一组织类型的基因型的方法,所述生物样品包括来自包括所述第一组织类型的多个组织类型的游离DNA分子的混合物,所述第一组织类型可能具有与所述多个组织类型的其它组织类型不同的基因型,所述方法包括:

识别所述第一组织类型的游离DNA分子的末端以高于阈值的比率出现的第一基因组位置;

通过计算机系统分析来自个体的所述生物样品的第一多个游离DNA分子,其中分析游离DNA分子包括:

确定对应于所述游离DNA分子的至少一个末端的参考基因组中的基因组位置;

基于所述第一多个游离DNA分子的所述分析,识别结束于所述第一基因组位置处的游离DNA分子集合;

对于所述游离DNA分子集合中的每一个:

确定所述第一基因组位置处出现的对应碱基,借此确定所述第一基因组位置处的多个对应碱基;和

使用所述游离DNA分子集合中的所述第一基因组位置处出现的所述多个对应碱基确定所述第一基因组位置处的所述第一组织类型的基因型。

3.分析生物样品的方法,所述生物样品包括来自包括第一组织类型的多个组织类型的游离DNA分子的混合物,所述方法包括:

通过计算机系统分析来自个体的所述生物样品的多个游离DNA分子,所述多个游离DNA分子中的每一个具有左末端和右末端,其中分析游离DNA分子包括:

确定对应于所述游离DNA分子的所述左末端的参考基因组中的左侧结束位置;

确定对应于所述游离DNA分子的所述右末端的所述参考基因组中的右侧结束位置;

识别左侧基因组位置的左侧集合,其各自具有所述多个游离DNA分子的左末端的局部

最大值,所述局部最大值对应于所述左侧基因组位置的左侧集合中的一个;

识别右侧基因组位置的右侧集合,其各自具有所述多个游离DNA分子的右末端的局部最大值,所述局部最大值对应于所述右侧基因组位置的右侧集合中的一个;和

如下识别对所述第一组织类型具有特异性的基因组位置的第一集合:

比较所述左侧集合的左侧基因组位置与所述右侧集合的右侧基因组位置以识别所述基因组位置的第一集合,其中左侧基因组位置与最接近的右侧基因组位置的距离大于第一阈值距离,所述第一阈值距离为所述参考基因组中的至少5个基因组位置。

4.分析个体的生物样品以确定所述个体中癌症水平的方法,所述方法包括:

识别肿瘤组织类型的游离DNA分子的末端以高于阈值的比率出现的基因组位置的第一集合;

通过计算机系统分析来自所述个体的所述生物样品的多个组织类型的第一多个游离DNA分子,其中分析游离DNA分子包括:

确定对应于所述游离DNA分子的至少一个末端的参考基因组中的基因组位置;

确定结束于所述基因组位置的第一集合之一处的所述第一多个游离DNA分子的第一数目;

通过使用游离DNA分子的第二数目对所述第一多个游离DNA分子的所述第一数目进行标准化来确定相对丰度,所述第二数目的游离DNA分子结束于与所述基因组位置的第一集合不同的基因组位置的第二集合处;和

通过比较所述相对丰度与一个或多个截止值确定所述个体的癌症水平。

5.分析生物样品的方法,所述方法包括:

分析来自个体的所述生物样品的多个游离DNA分子,所述多个游离DNA分子包括来自包括第一组织类型的多个组织类型的游离DNA分子的混合物,所述多个游离DNA分子中的每一个具有左末端和右末端,其中分析游离DNA分子包括:

确定对应于所述游离DNA分子的所述左末端的参考基因组中的左侧结束位置;和

确定对应于所述游离DNA分子的所述右末端的所述参考基因组中的右侧结束位置;

识别对所述第一组织类型具有特异性的基因组位置的第一集合,其中所述左侧结束位置和所述右侧结束位置的相邻局部最大值已知相隔大于第一阈值距离;

识别基因组位置的第二集合,其中所述左侧结束位置和所述右侧结束位置的相邻局部最大值已知相隔小于第二阈值距离;

使用结束于所述基因组位置的第一集合中的一个处的所述多个游离DNA分子的第一数目和结束于所述基因组位置的第二集合中的一个处的所述多个游离DNA分子的第二数目确定分离值;和

通过比较所述分离值与一个或多个校准值确定所述第一组织类型的比例贡献。

6.分析生物样品的方法,所述方法包括:

分析来自至少一个第一生物样品的多个游离DNA分子以识别所述多个游离DNA分子的结束位置,其中所述生物样品包括第一组织类型,所述多个游离DNA分子包括来自包括所述第一组织类型的多个组织类型的游离DNA分子的混合物;

对于多个基因组窗口中的每个基因组窗口:

确定结束于所述基因组窗口上的所述多个游离DNA分子的对应数目;且

比较所述对应数目与参考值以确定结束于所述基因组窗口内的一个或多个基因组位置上的游离DNA分子的比率是否高于阈值,和

将来自所述多个基因组窗口中具有高于所述阈值的比率的一个或多个基因组位置的基因组位置的第一集合识别为所述第一组织类型特异性的。

7.分析生物样品的方法,所述方法包括:

提供包括来自包括所述第一组织类型的多个组织类型的游离DNA分子的混合物的生物样品;

识别所述第一组织类型的游离DNA分子的末端以高于阈值的比率出现的基因组位置的第一集合;

分析来自个体的所述生物样品的多个组织类型的第一多个游离DNA分子,其中分析游离DNA分子包括:

确定对应于所述游离DNA分子的至少一个末端的参考基因组中的基因组位置;

基于所述第一多个游离DNA分子的所述分析,识别结束于所述基因组位置的第一集合之一处的游离DNA分子的第一集合;和

使用所述游离DNA分子的第一集合确定所述第一组织类型的DNA中相对于所述个体的健康细胞的一种或多种组织相关改变。

8.包含计算机可读介质的计算机产品,所述计算机可读介质存储用于控制计算机系统执行上述任何方法的操作的多个指令。

9.系统,其包含:

根据权利要求8所述的计算机产品;和

一个或多个处理器,所述一个或多个处理器用于执行存储在所述计算机可读介质上的指令。

10.系统,所述系统包含用于执行上述任何方法的装置。

游离DNA的片段化模式的分析

[0001] 相关申请的交叉引用

[0002] 本申请是申请号为201680051235.0的中国专利申请的分案申请，并且要求2015年7月23日提交的美国临时申请第62/196,250号和2016年2月12日提交的第62/294,948号以及2016年2月14日提交的PCT申请第PCT/CN2016/073753号的优先权，其全部内容出于所有目的以引用的方式并入本文中。

背景技术

[0003] 在先前研究中，已经表明血浆DNA主要由小于200bp的短片段组成(Lo等人《科学·转化医学(Sci Transl Med)》2010;2(61):61ra91)。在血浆DNA的尺寸分布中，峰值可以观测于166bp处。另外，观测到当对母体血浆DNA测序时，测序的标签密度将在转录起始位点(TSS)附近以约180bp的周期性变化(Fan等人《美国国家科学院院刊(PNAS)》2008;105:16266-71)。这些结果是血浆DNA的片段化可能不是随机过程的一组证据。但是，血浆中的DNA片段化的精确模式以及管控所述模式的因素尚不明确。另外，尚未完全实现使用DNA片段化的实际应用。

发明内容

[0004] 各种实施例涉及游离DNA，例如血浆DNA和血清DNA的片段化模式的分析的应用(例如诊断应用)。一种应用的实施例可以确定来自不同组织类型的游离DNA的混合物中的特定组织类型的比例贡献的分类。举例来说，特定百分比、百分比范围或比例贡献是否高于指定百分比，可以确定为分类。在一个实例中，可以识别特定组织类型的优选的结束位置，且结束于优选的结束位置上的游离DNA分子的相对丰度可以用于提供比例贡献的分类。在另一实例中，可以使用对特定组织类型具有特异性的区域中的片段化模式的幅度(例如结束于基因组位置处的游离DNA分子的数目)。

[0005] 另一应用的实施例可以确定来自不同组织类型的游离DNA的混合物中的特定组织类型的基因型。在一个实例中，可以识别特定组织类型的优选的结束位置，且可以使用结束于优选的结束位置上的游离DNA分子确定基因型。

[0006] 另一应用的实施例可以通过比较游离DNA分子的左末端的局部最大值与游离DNA分子的右末端的局部最大值来识别优选的结束位置。可以在对应局部最大值充分分离时识别优选的结束位置。另外，结束于左末端/右末端的局部最大值上的游离DNA分子的量可以与低分离的局部最大值的游离DNA分子的量进行比较以确定组织类型的比例贡献。

[0007] 其它实施例涉及与本文所述的方法相关的系统、便携式消费装置和计算机可读介质。

[0008] 可以参考以下具体实施方式和附图来更好地了解本发明实施例的性质和优点。

附图说明

[0009] 图1显示根据本发明的实施例的完整概率(P_I)的定义的说明性实例。

[0010] 图2A和2B显示根据本发明的实施例，使用25作为z值的跨越染色体6上的片段的P_I的变化。

[0011] 图3显示母体血浆中的母源性和胎儿源性DNA的P_I的同步变化的图示。

[0012] 图4显示母体血浆中的母源性和胎儿源性DNA的P_I的异步变化的图示。

[0013] 图5为显示母体和胎儿DNA分子的P_I变化是否同步的分析的流程图。

[0014] 图6显示关于母体血浆中的母源性(红色/灰色)和胎儿源性(蓝色/黑色)DNA片段的P_I的变化的两个母体血浆样品(S24和S26)的分析。

[0015] 图7显示P_I的变化幅度的图示。

[0016] 图8A显示是脱氧核糖核酸酶超敏性位点但不是TSS的区域处的P_I变化的模式。图8B显示是TSS但不是脱氧核糖核酸酶超敏性位点的区域处的P_I变化的模式。

[0017] 图9显示从不同组织释放的DNA的比例的测量原理的图示。

[0018] 图10显示FR_A与混合物中的组织A相对于DNA的比例贡献之间的关系，所述比例贡献通过分析具有已知比例浓度的来自组织A的DNA的两种或更多种校准样品而确定。

[0019] 图11显示$FR_{胎盘}$与母体血浆中的胎儿DNA百分比之间的相关性。

[0020] 图12显示$FR_{血液}$与母体血浆中的胎儿DNA浓度之间的相关性。

[0021] 图13为分析生物样品以确定根据本发明的实施例的第一组织类型的比例贡献的分类的方法1300的流程图。

[0022] 图14显示肿瘤或胎儿源性DNA的循环DNA片段的差异原理的图示。

[0023] 图15为分析包括来自多个组织类型(包括第一组织类型)的游离DNA分子的混合物的生物样品的方法的流程图。

[0024] 图16为显示对HCC案例具有特异性、对孕妇具有特异性和由两种案例共用的频繁结束位点的数目的文氏图。

[0025] 图17显示校准曲线，其显示结束于癌症特异性结束位置上的经测序DNA片段的比例与具有已知的血浆中的肿瘤DNA分数的癌症患者的血浆中的肿瘤DNA分数之间的关系。

[0026] 图18显示携有胎儿特异性等位基因和由母体和胎儿共用的等位基因的血浆DNA的非随机片段化模式的说明性实例。

[0027] 图19显示基因组坐标作为跨越具有信息性单核苷酸多态性(SNP)的区域的母体血浆DNA片段的结束位置的概率曲线。

[0028] 图20显示跨越在母体中为纯合且在胎儿中为杂合的SNP的血浆DNA片段的结束位置的分析。

[0029] 图21显示跨越在胎儿中为纯合且在母体中为杂合的SNP的血浆DNA片段的结束位置的分析。

[0030] 图22显示具有反复性胎儿(集合A)和母体(集合X)末端的血浆DNA分子的相对丰度(比率(F/M))与胎儿DNA分数之间的相关性。

[0031] 图23A-23E显示关于结束于胎儿优选的结束位置上的片段和结束于母体优选的结束位置上的片段的血浆DNA尺寸分布的数据。

[0032] 图24A-24E显示关于结束于胎儿优选的结束位置上的片段和结束于母体优选的结束位置上的片段的来自26个早期妊娠孕妇的合并的血浆DNA样品中的血浆DNA尺寸分布的数据。

[0033] 图25显示HCC患者的血浆DNA的非随机片段化模式的说明性实例。

[0034] 图26是基因组坐标作为跨越具有突变位点的区域的血浆DNA片段的结束位置的概率曲线。

[0035] 图27A显示跨越突变存在于肿瘤组织中的基因组位置的血浆DNA片段的结束位置的分析。

[0036] 图27B显示71个HCC患者的比率$_{M/WT}$与血浆中的肿瘤DNA分数之间的相关性。

[0037] 图28A显示孕妇和HCC患者的血浆DNA的优选的结束位置的数目。集合P含有2千9百万个在孕妇中为优选的结束位置。

[0038] 图28B显示在71个HCC患者的比率$_{HCC/Preg}$与血浆中的肿瘤DNA分数之间观测到正相关。

[0039] 图29A显示优选的末端终止比(PETR)的概念的图示。每条线表示一个血浆DNA片段。

[0040] 图29B显示11个HCC患者的血浆中的肿瘤DNA分数与集合H位置处的PETR之间的相关性。

[0041] 图30显示在结束于HCC优选的末端、HBV优选的末端或共用末端的血浆DNA分子中检测到的短DNA(＜150bp)的比例。

[0042] 图31A显示w-PETR的原理的说明。w-PETR的值计算为结束于窗口A与窗口B内的DNA片段的数目之间的比率。

[0043] 图31B显示11个HCC患者中的肿瘤DNA分数与w-PETR的值之间的相关性。

[0044] 图32显示当相比于脐带血血浆样品(210×单倍体基因组覆盖度)时,所研究样品中的每一个的血浆样品中检测的通常共用的优选的结束位置的比例。

[0045] 图33显示文氏图,所述文氏图显示通常观测于两个或更多个样品中的优选的结束位置以及仅观测于任何一个样品中的优选的结束位置的数目。

[0046] 图34A显示血浆中的胎儿DNA分数与通过“产前”与“产后”血浆DNA样品之间的比较识别的位置集合上的平均PETR之间的相关性。图34B显示血浆中的胎儿DNA分数与通过“产前”与“产后”血浆DNA样品之间的比较识别的位置集合上的平均PETR之间的相关性。

[0047] 图35A显示在18周(怀孕个体1)和38周怀孕期(怀孕个体2)的两个孕妇中的顶部1百万个最频繁观测的血浆DNA优选的结束位置。

[0048] 图35B显示两个孕妇的血浆中的顶部1百万个最频繁观测的优选的结束位置的PETR值的比较。

[0049] 图36是分析生物样品以确定根据本发明的实施例的混合物中的第一组织类型的比例贡献的分类的方法的流程图。

[0050] 图37显示当与胎儿优选的结束位置附近的参考基因组比对时携有不同等位基因的母体血浆DNA分子。

[0051] 图38是根据本发明的实施例分析生物样品以确定第一组织类型的基因型的方法3800的流程图。

[0052] 图39显示可可用于根据本发明实施例的系统和方法的示例计算机系统10的框图。

具体实施方式

[0053] 术语

[0054] “组织”对应于集合在一起作为功能单元的细胞组。可以在单一组织中发现超过一种类型的细胞。不同类型的组织可以由不同类型的细胞(例如肝细胞、肺泡细胞或血细胞)组成,但也可以对应于来自不同生物体(母亲相对于胎儿)的组织或对应于健康细胞相对于肿瘤细胞。

[0055] “生物样品”是指取自个体(例如人类,如孕妇、患有癌症的个体、或疑似患有癌症的个体、器官移植受者或疑似具有涉及器官(例如心肌梗塞的心脏,或中风的大脑。或贫血的造血系统)的疾病过程的个体的任何样品且含有一个或多个所关注的核酸分子。生物样品可以是体液,如血液、血浆、血清、尿液、阴道流体、来自水囊肿(例如睾丸水囊肿)的流体、阴道冲洗流体、胸膜液、腹水流体、脑脊髓液、唾液、汗液、泪液、痰液、支气管肺泡灌洗液、来自乳头的排出流体、来自身体的不同部分(例如甲状腺、乳房)的抽吸流体等。也可以使用粪便样品。在各种实施例中,已富集游离DNA的生物样品(例如经由离心方案获得的血浆样品)中的大部分DNA可以是游离的,例如大于50%、60%、70%、80%、90%、95%或99%的DNA可以是游离的。离心方案可以包括例如3,000g×10分钟,获得流体部分,且再次在例如30,000g下再离心10分钟以去除残余细胞。

[0056] “癌症相关的变化”或“癌症特异性变化”包括(但不限于)癌源突变(包括单核苷酸突变、核苷酸的缺失或插入、基因或染色体片段的缺失、易位、倒位),基因、基因片段或染色体片段的扩增,病毒相关序列(例如病毒游离基因和病毒插入)、异常甲基化概况或肿瘤特异性甲基化签名、异常游离DNA尺寸概况、异常组蛋白修饰标记和其它表观遗传修饰和癌症相关或癌症特异性的游离DNA片段的结束位置。

[0057] “信息性癌症DNA片段”对应于带有或携有癌症相关或癌症特异性变化或突变中的任何一个或多个的DNA片段。“信息性胎儿DNA片段”对应于携有未发现于亲本的任一个基因组中的突变的胎儿DNA片段。“信息性DNA片段”可以指以上类型的DNA片段中的任一个。

[0058] “序列读取”是指从核酸分子的任何部分或全部测序的核苷酸串。举例来说,序列读取可以是从核酸片段测序的短核苷酸串(例如20-150个)、在核酸片段的一个或两个末端处的短核苷酸串或测序生物样品中存在的整个核酸片段。序列读取可以通过多种方式获得,例如使用测序技术或使用探针,例如杂交阵列或捕获探针,或扩增技术,如聚合酶链反应(PCR)或使用单引物的线性扩增或等温扩增。

[0059] “结束位置”或“末端位置”(或仅仅“末端”)可以指游离DNA分子,例如血浆DNA分子的最外碱基(即在末端处)的基因组坐标或基因组标识或核苷酸标识。末端位置可以对应于DNA分子的任一个末端。以此方式,如果一端是指DNA分子的起点和末端,那么两个都将对应于结束位置。在实践中,一个末端位置是通过分析方法检测或确定的游离DNA分子的一个末端上的最外碱基的基因组坐标或核苷酸标识,所述分析方法为如(但不限于)大规模平行测序或下一代测序、单分子测序、双链或单链DNA测序文库制备方案、聚合酶链反应(PCR)或微阵列。此类体外技术可以改变游离DNA分子的真实体内实体末端。因此,每个可检测末端可以表示生物学上的真实末端或所述末端是一个或多个朝内的核苷酸或一个或多个从分子的原始末端延伸的核苷酸,例如非平末端双链DNA分子的突出端通过克列诺片段的5′钝化和3′填充。结束位置的基因组标识或基因组坐标可以从序列读取与人类参考基因组,例如

hg19的比对的结果得出。其可以从表示人类基因组的原始坐标的指数或代码的目录得出。其可以指通过(但不限于)标靶特异性探针、微测序、DNA扩增读取的游离DNA分子上的位置或核苷酸标识。

[0060] “优选的末端”(或“反复性结束位置”)是指在具有生理(例如怀孕)或病理(疾病)状态(例如癌症)的生物样品中,比不具有此类状态的生物样品或比相同病理或生理状态的不同时间点或阶段,例如在治疗之前或之后更高度表示或普遍的末端(例如如通过比率所测量)。优选的末端因此具有相对于其它状态增加的在相关生理或病理状态下被检测到的似然性或概率。增加的概率可以在病理状态与非病理状态之间,例如在患有癌症与无癌症的患者之间比较且定量为似然比或相对概率。似然比可以基于检测所测试样品中的至少阈值数目的优选的末端的概率或基于检测患有此类病况的患者相比于无此类病况的患者中的优选的末端的概率来确定。似然比的阈值的实例包括(但不限于)1.1、1.2、1.3、1.4、1.5、1.6、1.8、2.0、2.5、3.0、3.5、4.0、4.5、5、6、8、10、20、40、60、80和100。此类似然比可以通过比较具有和不具有相关状态的样品的相对丰度值来测量。由于检测相关生理或疾病状态下的优选的末端的概率较高,此类优选的结束位置将见于超过一个具有相同生理或疾病状态的个体中。随着概率增加,超过一个游离DNA分子可以检测为结束于相同优选的结束位置上,甚至在分析的游离DNA分子的数目远小于基因组的大小时也如此。因此,优选的或反复性结束位置也称为“频繁的结束位置”。在一些实施例中,定量阈值可以用于要求将在相同样品或相同样品等分试样内检测至少多次(例如3、4、5、6、7、8、9、10、15、20或50次)的末端视为优选的末端。相关生理状态可以包括个体健康、无疾病或无所关注的疾病时的状态。类似地,“优选的结束窗口”对应于优选的结束位置的连续集合。

[0061] 结束于位置上的DNA分子的“速率”涉及DNA分子结束于所述位置上的频繁程度。速率可以可以基于相对于所分析的DNA分子的数目标准化的结束于所述位置上的DNA分子的数目。因此,速率对应于多少DNA分子结束于位置上的频率,且不涉及具有结束于所述位置上的DNA分子的数目中的局部最大值的位置的周期性。

[0062] “校准样品”可以对应于组织特异性DNA分数是已知或经由校准方法,例如使用对所述组织具有特异性的等位基因确定的生物样品。作为另一实例,校准样品可以对应于可以从其确定优选的结束位置的样品。校准样品可以用于两种目的。

[0063] “校准数据点”包括“校准值”和所关注的DNA(即,特定组织类型的DNA)的经测量或已知比例分布。校准值可以是对于已知组织类型比例分布的校准样品确定的相对丰度。校准数据点可以通过多种方式界定,例如界定为离散点或校准函数(也被称为校准曲线或校准表面)。校准函数可以从校准数据点的额外数学转换导出。

[0064] 术语“测序深度”是指基因座被与基因座比对的序列读取覆盖的次数。基因座可以小至核苷酸,或与染色体臂一样大,或与全基因组一样大。测序深度可以表达为50×、100×等,其中“×”是指基因座被序列读取覆盖的次数。测序深度也可以应用于多个基因座,或全基因组,在此情况下,×可以分别指基因座或单倍体基因组或全基因组经测序的平均次数。超深测序可以指测序深度为至少100×。

[0065] “分离值”对应于涉及两个值的差值或比率。分离值可以是简单的差值或比率。举例来说,x/y的正比以及x/(x+y)是分离值。分离值可以包括其它因数,例如相乘因数。作为其它实例,可以使用所述值的函数的差值或比率,例如两个值的自然对数(ln)的差值或比

率。分离值可以包括差值和比率。

[0066] “相对丰度”是涉及结束于基因组位置的一个窗口内的游离DNA分子的量(一个值)与结束于基因组位置的另一窗口内的游离DNA分子的量(另一个值)的一种类型的分离值。两个窗口可以重叠,但将具有不同尺寸。在其它实施方案中,两个窗口将不重叠。另外,窗口可以具有一个核苷酸的宽度,并且因此等效于一个基因组位置。

[0067] 如本文所用,术语“分类”是指与样品的特定特性有关的任何数目或其它字符。举例来说,“+”符号(或词语“正性”)可以表示样品归类为具有缺失或扩增。分类可以是二元(例如正性或负性)或具有更多分类等级(例如1到10或0到1的标度)。术语“截止值”和“阈值”是指操作中所使用的预定数目。举例来说,截止尺寸可以是指一种尺寸,高于所述尺寸则排除片段。阈值可是一种值,高于或低于所述值则适用特定分类。这些术语中的任一个可以在这些背景中的任一背景下使用。

[0068] 术语“癌症水平”可以指癌症是否存在(即,存在或不存在)、癌症阶段、肿瘤尺寸、是否存在癌转移、身体的总肿瘤负荷和/或癌症严重度的其它量度(例如癌症复发)。癌症水平可以是数字或其它标志,诸如符号、字母和颜色。水平可以是零。癌症水平还包括与突变或多种突变相关的癌变前或癌前期病状(状态)。癌症水平可以按各种方式使用。举例来说,筛选可以检查已知先前不患有癌症的某人是否存在癌症。评估可以调查已经诊断患有癌症的某人以监测癌症随时间的进展、研究疗法的有效性或确定预后。在一个实施例中,预后可以表示为患者死于癌症的机率,或在具体持续期或时间之后癌症进展的机率,或癌症转移的机率。检测可以意指‘筛选’或可以意指检查具有癌症的暗示特征(例如症状或其它阳性测试)的某人是否患有癌症。

[0069] “局部最大值”可以指当相比于相邻位置时,获得所关注的参数的最大值的基因组位置(例如核苷酸)或是指此类基因组位置处的所关注的参数的值。举例来说,相邻位置可以在50bp到2000bp范围内。所关注的参数的实例包括(但不限于)结束于基因组位置上的片段的数目、与所述位置重叠的片段的数目或大于阈值尺寸的覆盖基因组位置的片段的比例。许多局部最大值可以在所关注的参数具有周期性结构时出现。全域最大值是局部最大值中的特定最大值。类似地,“局部最小值”可以指当相比于相邻位置时,获得所关注的参数的最小值的基因组位置或是指此类基因组位置处的所关注的参数的值。

[0070] 描述影响游离DNA(例如血浆DNA)的片段化模式的因素和游离DNA片段化模式分析的应用,包括在分子诊断学中的应用。各种应用可以使用片段化模式的特性来确定特定组织类型的比例贡献,以确定特定组织类型(例如母体样品中的胎儿组织或来自癌症患者的样品中的肿瘤组织)的基因型,和/或识别特定组织类型的优选的结束位置,其可以接着用于确定特定组织类型的比例贡献。在一些实施例中,特定组织的优选的结束位置也可以用于测量样品中的特定组织类型的绝对贡献,例如以每单位体积(例如每毫升)的基因组数目。

[0071] 比例贡献的分类的实例包括特定百分比、百分比范围或高于指定百分比的比例贡献是否可以确定为分类。为了确定比例贡献的分类,一些实施例可以识别对应于特定组织类型(例如胎儿组织或肿瘤组织)的优选的结束位置。此类优选的结束位置可以不同方式确定,例如通过分析游离DNA分子结束于基因组位置上的的速率、比较此类速率与其它样品(例如不具有相关病况)和对于病况不同的不同组织和/或不同样品比较具有游离DNA分子

的末端的高出现率的基因组位置集合。结束于优选的结束位置处的游离DNA分子相对于结束于其它基因组位置处的游离DNA分子的相对丰度可以相比于一个或多个校准值，所述一个或多个校准值由已知特定组织类型的比例贡献的一个或多个校准生物样品确定。本文提供的数据显示相对丰度的各种量度与样品中的各种组织的比例贡献之间的正关系。

[0072] 为了确定比例贡献的分类，一些实施例可以使用片段化模式中的幅度(例如结束于基因组位置处的游离DNA分子的数目)。举例来说，可以通过分析结束于多个基因组位置处的游离DNA分子的数目来识别一个或多个局部最小值和一个或多个局部最大值。一个或多个局部最大值处的游离DNA分子的第一数目与一个或多个局部最小值处的游离DNA分子的第二数目的分离值(例如比率)显示为与特定组织类型的比例贡献正相关。

[0073] 在一些实施例中，可以关于游离DNA样品的体积或重量测量所关注的组织的浓度。举例来说，定量PCR可以用于测量单位体积或单位重量的提取的游离DNA样品中的结束于一个或多个优选的末端处的游离DNA分子的数目。可以对于校准样品进行类似测量，且因此比例贡献可以确定为比例贡献，因为贡献为每单位体积或单位重量的浓度。

[0074] 为了确定来自不同组织类型的游离DNA的混合物中的特定组织类型(例如胎儿组织或肿瘤组织)的基因型，一些实施例可以识别特定组织类型的优选的结束位置。对于结束于优选的结束位置上的游离DNA分子集合中的每个游离DNA分子，可以确定优选的结束位置处出现的对应碱基。对应碱基可以用于确定优选的结束位置处的基因型，例如基于可见的不同碱基的百分比。在各种实施方案中，高百分比的仅一个碱基(例如高于90％)可以指示基因型对于碱基是纯合的，而两个具有类似百分比(例如在30-70％之间)的碱基可能使得确定基因型是杂合的。

[0075] 为了识别优选的结束位置，一些实施例可以比较游离DNA分子的左末端的局部最大值与游离DNA分子的右末端的局部最大值。可以在对应局部最大值经充分地分离时识别优选的结束位置。另外，左末端/右末端的结束于局部最大值上的游离DNA分子的量可以相比于具有低分离的局部最大值的游离DNA分子的量以确定组织类型的比例贡献。

[0076] 在下文的描述中，首先描述片段化和技术的概述，接着描述片段化模式和其定量实例的特性，且其它描述与确定比例贡献、识别优选的结束位置和确定基因型相关。

[0077] I.片段化和技术的概述

[0078] 在本发明中，我们显示存在游离DNA的非随机片段化方法。非随机片段化方法在一定程度上发生于不同类型的生物样品中，所述生物样品含有游离DNA，例如血浆、血清、尿液、唾液、脑脊髓液、胸膜液、羊膜液、腹膜流体和腹水流体。游离DNA天然地以短片段形式出现。游离DNA片段化是指借此高分子量DNA(如细胞的细胞核中的DNA)在产生或释放游离DNA分子时裂解、破碎或消化为短片段的方法。

[0079] 并非全部游离DNA分子具有相同长度。一些分子短于其它分子。已经证明游离DNA，如血浆DNA在开放染色质域内，包括在转录起始位点周围，和在核小体核心之间的位置处，如在连接子位置处一般较短且较不完整，即具有不佳完整概率或较差完整性(Strayer等人《产前诊断(Prenat Diagn)》2016,36:614-621)。各个不同组织具有其特征型基因表达概况，所述特征型基因表达概况转而通过包括染色质结构和核小体定位的方式调节。因此，某些基因组位置处的完整概率或完整性的游离DNA模式，如血浆DNA的所述游离DNA模式是那些DNA分子的组织来源的签名或标志。类似地，当疾病过程，例如癌症改变细胞的基因组的

基因表达概况和功能时，源自患病细胞的游离DNA完整概率概况将反映那些细胞。游离DNA概况因此将为疾病存在提供证据或是疾病存在的标志。

[0080] 一些实施例另外增强研究游离DNA片段化概况的分辨力。代替仅仅经核苷酸的伸长部对读数求和以识别具有较高或较低完整概率或完整性的区域，我们研究个别游离DNA分子，尤其是血浆DNA分子的实际结束位置或末端。值得注意地，我们的数据展示切割游离DNA分子的特定位置是非随机的。在体外剪切或声波处理的高分子量基因组组织DNA显示具有跨越基因组随机分散的结束位置的DNA分子。但是，存在某些高度表示于如血浆的样品内的游离DNA分子的结束位置。此类结束位置的出现或表示数目在统计学上显著高于单独的偶然预期值。这些数据使得我们对游离DNA片段化的理解超出完整性的区域性变化一步(Snyder等人《细胞(Cell)》2016,164:57-68)。此处，我们显示游离DNA片段化的方法甚至往下精心安排到切割或裂解的特定核苷酸位置。我们将游离DNA结束位置的这些非随机位置称为优选的结束位置或优选的末端。

[0081] 在本发明中，我们显示存在通常跨越不同生理状态或疾病状态的个体出现的游离DNA结束位置。举例来说，存在由怀孕和非怀孕个体共用、由怀孕和癌症患者共用、由患有癌症和不患有癌症的个体共用的共同优选末端。另一方面，存在主要仅出现于孕妇中、仅出现于癌症患者中或仅出现于不患有癌症的非怀孕个体中的优选末端。有趣的是，这些怀孕特异性或癌症特异性或疾病特异性末端也高度表示于具有可比生理或疾病状态的其它个体中。举例来说，一个孕妇的血浆中识别的优选末端可检测于其它孕妇的血浆中。此外，此类优选末端的比例的量与其它孕妇的血浆中的胎儿DNA分数相关。此类优选末端实际上与怀孕或胎儿相关，因为其量在产后母体血浆样品中大量减少。类似地，在癌症中，一个癌症患者的血浆中识别的优选末端可检测于另一癌症患者的血浆中。此外，此类优选末端的比例的量与其它癌症患者的血浆中的肿瘤DNA分数相关。此类优选末端与癌症相关，因为其量在癌症治疗，例如手术切除后减少。

[0082] 存在多种用于分析游离DNA优选末端的应用或公用程序。其可以提供关于孕期的胎儿DNA分数和因此的胎儿健康的信息。举例来说，已报道多种怀孕相关病症，如先兆子痫、早产、宫内生长受限(IUGR)、胎儿染色体非整倍性等与相比于孕龄匹配的对照怀孕的胎儿DNA的分数浓度，即胎儿DNA分数，或胎儿分数的扰动相关。与癌症相关的游离血浆DNA优选末端在血浆样品中展现肿瘤DNA分数或分数浓度。知道肿瘤DNA分数提供关于癌症阶段、预后和帮助监测治疗功效或癌症复发的信息。游离DNA优选末端的概况也将展示向含有游离DNA的生物样品，例如血浆中贡献DNA的组织的组成。因此可能能够识别癌症或其它病变的组织来源，例如脑血管意外(即中风)，全身性红斑狼疮的器官表现。

[0083] 可以通过比较具有不同生理或病理状态的个体中的优选末端的游离DNA概况，例如非怀孕相比于怀孕样品、癌症相比于非癌症样品或未患癌症的孕妇的概况相比于非怀孕癌症患者的概况来识别与特定生理状态或病理状态相关的优选末端的目录。另一种方法为比较生理(例如怀孕)或病理(例如癌症)过程的不同时间处的优选末端的游离DNA概况。此类时间点的实例包括怀孕之前和之后、胎儿分娩之前和之后、跨越怀孕期间的不同孕龄收集的样品、癌症治疗(例如靶向疗法、免疫疗法、化疗、手术)之前和之后、癌症诊断之后的不同时间点、癌症进展之前和之后、产生癌转移之前和之后、疾病严重度增加之前和之后或产生并发症之前和之后。

[0084] 另外,可以使用与特定组织相关的遗传标记来识别优选末端。举例来说,含有胎儿特异性SNP等位基因的游离DNA分子将适用于识别如母体血浆的样品中的胎儿特异性优选末端。反过来,含有母体特异性SNP等位基因的血浆DNA分子将适用于识别母体血浆中的母体特异性优选末端。含有肿瘤特异性突变的血浆DNA分子可以用于识别与癌症相关的优选末端。在器官移植的情况下含有供体或受体特异性SNP等位基因的血浆DNA分子适用于识别移植或非移植器官的优选末端。举例来说,对供体具有特异性的SNP等位基因将适用于识别代表移植器官的优选末端。

[0085] 当优选末端具有在一种生理或病理状态下被检测到的高可能性或概率时,可以将所述优选末端视为与所述生理或疾病状态相关。在其它实施例中,优选末端具有相比于其它状态,更可能在相关生理或病理状态下被检测到的一定概率。由于检测相关生理或疾病状态下的优选末端的概率较高,此类优选或反复性末端(或末端位置)将见于超过一个具有所述相同生理或疾病状态的个体中。高概率也将使得此类优选或反复性末端在相同个体的相同游离DNA样品或等分试样中可检测多次。在一些实施例中,可以设定定量阈值以限制被视为优选末端的在相同样品或相同样品等分试样内检测至少指定次数(例如5、10、15、20等)的末端的纳入。

[0086] 在对于任何生理或病理状态建立游离DNA优选末端的目录之后,靶向或非靶向方法可以用于检测其在例如血浆的游离DNA样品或其它个体中的存在以确定具有类似健康、生理或疾病状态的其它测试个体的分类。游离DNA优选末端可以通过随机非靶向测序检测。将需要考虑测序深度以使得可以实现识别相关优选末端的全部或一部分的合理概率。或者,可以对游离DNA样品进行具有高密度的优选末端的基因座的杂交捕获以在不限于通过测序、微阵列或PCR的检测之后富集具有此类优选末端的游离DNA分子的样品。但是,替代地,基于扩增的方法可以用于特异性地扩增和富集具有优选末端的游离DNA分子,例如反向PCR、滚环扩增。扩增产物可以通过测序、微阵列、荧光探针、凝胶电泳和所属领域的技术人员已知的其它标准方法识别。

[0087] 在实践中,一个末端位置可以是通过分析方法检测或测定的游离DNA分子的一个末端上的最外碱基的基因组坐标或核苷酸标识,所述分析方法为如(但不限于)大规模平行测序或下一代测序、单分子测序、双链或单链DNA测序文库制备方案、PCR、用于DNA扩增(例如等温扩增)的其它酶方法或微阵列。此类体外技术可以改变游离DNA分子的真实体内实体末端。因此,每个可检测末端可以表示生物学上的真实末端或末端为一个或多个向内的核苷酸或一个或多个从分子的原始末端延伸的核苷酸。举例来说,克列诺片段(Klenow fragment)用于通过5′突出端的钝化和3′突出端的填充在DNA测序文库构筑期间产生平末端双链DNA分子。尽管此类程序可以展示不与生物末端相同的游离DNA末端位置,但仍可以建立临床相关性。这是因为与特定生理或病理状态相关或有关的优选项的识别可以基于将在校准样品和测试样品两者中对游离DNA末端产生一致和可再现改变的相同实验室方案或方法原理。多种DNA测序方案使用单链DNA文库(Snyder等人《细胞》2016,164:57-68)。单链文库的序列读取的末端可以比双链DNA文库的末端更向内或进一步延伸。

[0088] 末端位置的基因组标识或基因组坐标可以从序列读取与人类参考基因组,例如hg19的比对的结果得出。其可以从表示人类基因组的原始坐标的指数或代码的目录得出。尽管末端是游离DNA分子的一个或两个末端处的核苷酸,末端的检测可以通过识别血浆DNA

分子上的其它核苷酸或核苷酸的其它伸长部来进行。举例来说，具有经由结合到扩增子的中间碱基的荧光探针检测的优选末端的血浆DNA分子的阳性扩增。举例来说，末端可以通过结合到血浆DNA分子的中间区段上的一些碱基的荧光探针的阳性杂交识别，其中已知片段尺寸。通过这种方式，可以通过算出多少碱基在具有已知序列和基因组标识的荧光探针外部而确定末端的基因组标识或基因组坐标。换句话说，末端可以通过检测相同血浆DNA分子上的其它碱基而识别或检测。末端可以是通过(但不限于)标靶特异性探针、微测序和DNA扩增读取的游离DNA分子上的位置或核苷酸标识。

[0089] II.血浆DNA的片段化模式

[0090] 为了分析母体血浆DNA的片段化模式，我们对来自从妇产科(Department of Obstetrics and Gynaecology)招收的12周孕龄的孕妇的血浆DNA测序(Lo等人《科学·转化医学》2010;2(61):61ra91)。获自母体的血浆DNA使用Illumina基因组分析仪(Genome Analyzer)平台进行大规模平行测序。可以使用其它大规模平行或单分子测序仪。进行血浆DNA分子的双末端测序。每个分子在每个末端测序50bp，因此每分子总计100bp。每个序列的两个末端使用SOAP2程序与参考人类基因组(Hg18NCBI.36)比对(Li R等人《生物信息学(Bioinformatics)》2009,25:1966-7)。也从父体和母体的血沉棕黄层样品和CVS样品提取DNA。这些DNA样品使用Affymetrix全基因组人类SNP阵列6.0系统进行基因分型。

[0091] A.片段化的示例性定量

[0092] 为了反映片段化模式，可以基于母体血浆DNA的测序结果关于基因组对每个核苷酸确定完整概率(P_I)。

[0093] $$P_I = \frac{N_z}{N_T}$$

[0094] 其中N_z是覆盖靶核苷酸的两侧(5′和3′)上的至少z个核苷酸(nt)的全长测序读取的数目;且N_T是覆盖靶核苷酸的测序读取的总数。

[0095] P_I的值可以反映具有在长度为z值加1的两倍(2z+1)的特定位置处居中的完整DNA分子的概率。完整概率(P_I)的值越高，血浆DNA越不可能在特定核苷酸位置处片段化。为了进一步对此进行说明，在图1中说明完整概率的定义。

[0096] 图1显示完整概率(P_I)的定义的说明性实例。T是计算P_I的靶核苷酸位置。A和B分别是T上游的z个核苷酸(nt)(5′)和下游的z个nt(3′)处的两个位置。从a到j标记的黑线表示来自母体血浆的经测序血浆DNA片段。片段a到d覆盖所有三个位置A、B和T。因此，覆盖靶核苷酸的两侧(5′和3′)上的至少z个nt的片段数目(N_Z)为4。另外，片段e、f和g也覆盖位置T，但不覆盖位置A和B，因此，存在总共7个覆盖位置T的片段(N_T=7)。片段h和j覆盖A或B但不覆盖T。这些片段未在N_z或N_T中计数。因此，此特定实例中的P_I为4/7(57%)。

[0097] 在一个实施例中，P_I可以使用25作为z值来计算。因此，完整血浆DNA片段将定义为覆盖目标位置上的至少25个nt到目标位置下游的25个nt的片段。在其它实施例中，可以使用其它z值，例如(但不限于)10、15、20、30、35、40、45、50、55、60、65、70、75和80。

[0098] P_I为基因组位置的窗口内的游离DNA分子末端的相对丰度的实例。可以使用其它度量，例如P_I的倒数，其将与具有完整DNA分子的概率具有相反的关系。P_I的倒数的较高值将指示作为末端位置或末端窗口的较高概率。其它实例为末端DNA片段的测量数目相对于末端DNA片段的预期数目的p值、所有比对的DNA片段中的DNA片段末端的比例或优选的末端终

止比(PETR)的比例,其全部更详细地描述于下文。相对丰度的所有此类度量测量窗口内的游离DNA片段结束的速率,例如以2z+1的宽度,其中z可以是零,借此使得窗口等效于基因组位置。

[0099] B.片段化模式的周期性

[0100] 基因组的某些区域倾向于特定组织中的染色体区的较高断裂速率(频率),且因此具有所述区中的窗口内的较高游离DNA片段结束速率。相对丰度的曲线显示片段化模式,其可以具有周期性结构。周期性结构显示最大结束位置(高裂解)的位置和最小结束位置(低裂解)的位置。当使用P_I时,最大值对应于低裂解的窗口,因为P_I测量与裂解概率(末端位置概率)相反的完整概率,其彼此具有反比关系。

[0101] 图2A和2B显示根据本发明的实施例,使用25作为z值的跨越染色体6上的片段的P_I的变化。在图2A中,P_I的变化以不同灰色强度呈现,如左侧的键所示。在图2B中,P_I的变化观测于较短片段中。x轴为核苷酸(nt)中的基因组坐标且y轴为P_I。P_I的变化具有约180bp的表观周期性。

[0102] C.母体血浆中的母体和胎儿DNA的PI的同步变化

[0103] 尽管P_I以大致180bp的周期性跨越基因组变化,我们进一步研究P_I的变化是否将对于胎儿源性和母源性血浆DNA分子同步。同步变化意指PI的峰值(最大值)和谷值(最小值)在整个基因组或足够高比例的基因组中出现于相同的相对核苷酸位置处。界定足够高比例的阈值可以关于特定应用调节,例如(但不限于)＞20％、＞25％、＞30％、＞35％、＞40％、＞45％、＞50％、＞55％、＞60％、＞65％、＞70％、＞75％、＞80％、＞85％、＞90％和＞95％。以下两图(图3和图4)显示母体血浆中的母源性和胎儿源性DNA的P_I的变化之间的两种可能的关系。

[0104] 图3显示母体血浆中的母源性和胎儿源性DNA的P_I的同步变化的图示。P_I的峰值和谷值跨越基因组或在大部分基因组中出现于母体和胎儿DNA的相同的相对位置。如果区域中存在同步变化,那么胎儿源性DNA和母源性DNA将具有相同片段化模式,借此阻碍使用区域中的片段化模式的周期性作为组织类型中的一种的签名。

[0105] 图4显示母体血浆中的母源性和胎儿源性DNA的P_I的异步变化的图示。母体和胎儿DNA的P_I的峰值和谷值不具有跨越基因组的恒定相对关系。在区域I,母体DNA的P_I的峰值符合胎儿DNA的峰值。在区域II,母体DNA的P_I的峰值符合胎儿DNA的谷值。在区域III和IV,母体DNA的P_I的峰值在胎儿DNA的峰值与谷值中间。如果变化并非同步,那么胎儿和母体片段化模式中的此类差异可以用作签名以识别可能来自胎儿或母体的DNA。另外,此类差异可以用于确定胎儿或母体组织的比例贡献,如下文更详细地描述。举例来说,区域II中的一个峰值处的DNA片段末端更可能是胎儿DNA,且相比于其它基因组位置的此类峰值处的DNA片段末端的相对丰度将随着胎儿DNA分数增加而增加。

[0106] 图5为显示母体和胎儿DNA分子的P_I的变化是否同步的分析500的流程图。分析500研究P_I的变化是否在母体血浆中的母源性与胎儿源性DNA之间同步。分析500可以使用计算机系统。尽管如上文所述,使用测序进行分析500,但可以使用其它技术,例如如本文所述。

[0107] 在步骤510处,分析500识别孕妇为纯合(AA)且胎儿为杂合(AB)的SNP。这些SNP被称为信息性SNP。B等位基因为胎儿特异性等位基因。此类信息性SNP可以通过分析仅为或主要为母体来源的母体样品而识别。举例来说,可以使用血液样品的血沉棕黄层,因为白血细

胞将主要来自母体。仅出现一种核苷酸(或高百分比的一种核苷酸,例如高于80%,其可以取决于胎儿DNA分数)的基因组位置可以识别为在母体中纯合。可以分析血浆以识别母体中纯合的位置,其中识别足够百分比的具有另一经识别的等位基因的DNA片段。

[0108] 在步骤520处,识别具有胎儿特异性等位基因B的血浆DNA分子。由于等位基因B经识别,这些DNA分子可以识别为对应于胎儿组织。

[0109] 在步骤530处,对于母体血浆中的游离DNA确定P_I的值。P_I的这些值包括胎儿和母体DNA。给定基因组位置的P_I的值是通过分析与参考基因组的所述基因组位置比对的序列读取而获得。

[0110] 在步骤540处,通过分析步骤530的输出而确定P_I的峰值。可以不同方式识别峰值,且每个峰值可以仅限于一个基因组位置或允许对应于超过一个基因组位置。我们观测到对于母体血浆中的大多数目源性DNA,P_I跨越全基因组以具有大致180bp的周期性的正弦曲线样模式变化。

[0111] 在步骤550处,对于总母体血浆确定信息性SNP与最接近P_I(步骤540)之间的距离。我们对于主要源自孕妇自身的总血浆DNA识别相对于P_I变化的最接近峰值的SNP的位置。

[0112] 在步骤560处,聚集所有胎儿源性DNA片段。聚集所有携有胎儿特异性等位基因的检测的血浆DNA片段以计算胎儿源性DNA的P_I。接着参考总母体血浆DNA的最接近P_I峰值的位置计算聚集的胎儿源性DNA片段的P_I。以与计算总母体血浆DNA的P_I类似的方式进行胎儿源性DNA的P_I的计算。

[0113] 在步骤570处,确定胎儿源性DNA片段的P_I相对于总母体血浆DNA的P_I的峰值的变化。变化显示于图6中。

[0114] 图6显示关于母体血浆样品中的胎儿源性(红色/灰色)和总(蓝色/黑色)DNA片段的P_I的变化的两个母体血浆样品(S24和S26)的分析。纵轴显示百分比形式的P_I。横轴显示信息性SNP与P_I的最接近峰值之间的碱基对(bp)的距离。

[0115] 总值包括来自胎儿和母体DNA的贡献。跨越所有峰值P_I聚集总值。如可见,SNP距峰值P_I越近,P_I的值越高。实际上,对于胎儿源性DNA片段,峰值P_I位于约位置0处。因此,对于母源性和胎儿源性DNA片段,P_I峰值在大约相同位置。从这些数据,我们得出母源性和胎儿源性DNA的P_I的变化同步的结论。

[0116] 尽管片段化模式看起来同步,以下描述显示除周期性以外的其它属性可以用于区分片段化模式,借此允许确定特定组织类型的签名。举例来说,已发现某些基因组区域的峰值和谷值的幅度差异,借此允许那些区域内的某些位置用于确定组织特异性片段化模式。

[0117] D.影响血浆DNA的片段化模式的变化的因素

[0118] 在前述研究中,已经表明血浆DNA的片段化在TSS附近并非随机(Fan等人《美国国家科学院院刊》2008;105:16266-71)。特定核苷酸上的任何血浆DNA末端的概率将随着与TSS的距离,以大致核小体尺寸的周期性变化。一般认为此片段化模式是DNA的凋亡降解的结果。因此,血浆DNA的尺寸总体上类似于与组蛋白复合物相关的DNA的尺寸。

[0119] 在前述研究中,也显示血浆DNA的尺寸总体上类似于与核小体相关的DNA的尺寸(Lo等人《科学·转化医学》2010;2(61):61ra91)。相信血浆DNA是通过细胞DNA(细胞核DNA和线粒体DNA)的凋亡降解产生。此观点另外由循环线粒体DNA中不具有此核小体模式支持,因为线粒体DNA不与细胞中的组蛋白相关。尽管已经就核苷酸位置表明血浆DNA片段末端在

转录起始位点附近并非随机(Fan等人《美国国家科学院院刊》2008;105:16266-71),但调节血浆DNA的片段化模式的精确机制仍不明确。

[0120] 最近,已另外显示血浆DNA的尺寸将在具有不同序列环境的区域中不同(Chandrananda等人《BMC医学基因组学(BMC Med Genomics)》2015;8:29)。后面的数据也支持游离DNA片段更可能在核小体连接子区域上,而不是在核小体核心处起始和结束的前述假设。这些发现与如前述部分中所论述的我们关于完整概率的核苷酸之间的变化的发现一致。此处,我们另外假设完整概率的变化幅度将跨越不同基因组区域变化。片段化变化性的此区域间变化尚未在任何前述研究中充分探索或定量。以下图式说明P_I的局部和区域性变化的概念。

[0121] 图7显示P_I的变化幅度的图示。在前述部分中,我们展示了在DNA的短伸长部上存在P_I变化的正弦曲线样模式。此处,我们另外分析跨越较大基因组区域的变化幅度。变化幅度是指具有指定尺寸的特定区域处的PI的最高峰值与谷值变化之间的P_I差异。在一个实施例中,特定区域的尺寸可以是1000bp。在其它实施例中,可以使用其它尺寸,例如(但不限于)600bp、800bp、1500bp、2000bp、3000bp、5000bp和10000bp。

[0122] 如图7所示,区1的幅度高于区2中的幅度。此特性见于下文数据中。如果此类高幅度出现率出现于不同组织的不同基因组区域,那么幅度的测量值可以用于在分析幅度在组织类型之间不同的区域时确定组织类型的比例贡献。举例来说,如果不同组织类型的幅度不同,那么比例贡献将随着来自特定组织类型(例如胎儿组织或肿瘤组织)的DNA的量增加而按比例变化。因此,幅度的测量值将对应于特定比例贡献。实施例可以使用来自样品的校准数据,其中比例贡献是经由另一技术(例如通过等位基因分析、甲基化签名、扩增/缺失程度)测量,如以全文引用的方式并入的美国专利公开案第2009/0087847号、第2011/0276277号、第2011/0105353号、第2013/0237431号和第2014/0100121号中所述。

[0123] 在我们的测序数据中,我们观测到P_I的变化幅度跨越不同基因组区域而变化。我们假设P_I的变化幅度与染色质对细胞凋亡期间的降解的可接近性有关。因此,我们研究变化幅度与基因组中的脱氧核糖核酸酶超敏性位点之间的可能的关系。在先前研究中,观测到血浆DNA的片段化模式受其与TSS的相对定位影响。在我们的分析中,我们研究TSS和脱氧核糖核酸酶超敏性位点对血浆DNA的片段化模式的效应的相对重要性。可以使用幅度对应于测试组织的其它位点。所述类型的位点的一个实例是使用通过高通量测序分析转座酶可接近的染色质(ATAC-Seq)识别的位点(Buenrostro等人《自然方法(Nat Methods)》2013;10:1213-1218)。所述类型的位点的另一实例是使用微球菌核酸酶(MNase)识别的位点。

[0124] 我们比较两种类型的基因组区域中的P_I变化幅度:

[0125] ii.是TSS但不是脱氧核糖核酸酶超敏性位点的区域;和

[0126] iii.是脱氧核糖核酸酶超敏性位点但不是TSS的区域。

[0127] TSS和脱氧核糖核酸酶超敏性位点的坐标检索自ENCODE数据库(genome.ucsc.edu/ENCODE/downloads.html)。

[0128] 使用以下方法剖析TSS和脱氧核糖核酸酶I位点周围的P_I模式。

[0129] 1)目标参考位点周围的上游和下游2kb区域经检索。

[0130] 2)接着根据与参考位点的距离将绝对基因组坐标重缩放。举例来说,如果尺寸为60bp的特定窗口沿上游方向距参考位点50bp,那么其将标记为-50。另外,如果尺寸为60bp

的特定窗口沿下游方向距参考位点50bp，那么其将标记为+50。

[0131] 3)具有相同重缩放的新坐标的特定窗口中的P_I值将使用完整片段和所有与所述窗口重叠的片段的计数重新计算。

[0132] 图8A显示是脱氧核糖核酸酶超敏性位点但不是TSS的区域处的P_I变化的模式。图8B显示是TSS但不是脱氧核糖核酸酶超敏性位点的区域处的P_I变化的模式。如所示出，是脱氧核糖核酸酶超敏性位点但不是TSS的区域中的变化幅度比是TSS但不是脱氧核糖核酸酶超敏性位点的区域中的变化幅度高得多。这些观测结果表明影响血浆DNA的片段化模式的一个因素是进行片段化的区域与脱氧核糖核酸酶超敏性位点的相对位置。

[0133] III.使用峰值和谷值确定组织比例

[0134] 已展示与脱氧核糖核酸酶超敏性位点的相对位置是调节血浆DNA的片段化模式的重要因素，我们研究此观测结果是否可以转化为临床应用。已观测到脱氧核糖核酸酶超敏性位点的概况在不同类型的组织中不同。概况对应于位点的基因组位置；不同组织的脱氧核糖核酸酶超敏性位点的位置不同。因此，我们推论从不同类型的组织释放的血浆DNA将展现组织特异性片段化模式。以类似方式，可以使用区域的幅度在组织间变化的其它区域。

[0135] A.脱氧核糖核酸酶超敏性位点的实例

[0136] 图9显示从不同组织释放的DNA的比例的测量原理的图示。源自组织A的血浆DNA在具有高P_I(峰值，通过P指示)的核苷酸位置处的片段化概率较低。因此，源自组织A的血浆DNA的末端位于这些核苷酸位置的概率较低。相比之下，源自组织A的血浆DNA的末端位于具有低P_I(谷值，通过T指示)的核苷酸位置的概率较高。另一方面，由于此位点不是组织B的脱氧核糖核酸酶超敏性位点，对于源自组织B的血浆DNA来说，P_I变化幅度较低。因此，位置P和位置T上的来自组织B末端的血浆DNA的概率将类似，至少相对于关于组织A可见的变化的量。

[0137] 我们如下定义是组织A的脱氧核糖核酸酶超敏性位点的区域处的片段末端比(FR_A)：

[0138] FR

[0139] A＝NTNP

[0140] 其中N_T为结束于P_I的谷值的核苷酸位置上的血浆DNA片段的数目且N_P为结束于P_I的峰值的核苷酸位置上的血浆DNA片段的数目。FR_A为分离值的实例，且更确切地说结束于谷值上相对于结束于峰值上的DNA片段的相对丰度的实例。在其它实施例中，可以确定相邻谷值(局部最小值)和峰值(局部最大值)的分离比，且可以确定分离比的平均值。

[0141] 对于组织A，FR_A将大于1，因为N_T将大于N_P。对于组织B，FR_A将为大致1，因为N_T和N_P将类似。因此，在含有源自组织A和B两者的血浆DNA的混合物中，FR_A的值将与组织A的比例贡献具有正相关。在实践中，组织B的FR_A不需要为1。只要组织B的FR_A不同于组织A的FR_A，可以从FR_A确定两种类型的组织的比例贡献。

[0142] 在此类区域中，DNA片段结束于谷值的似然性的高变化将导致相比于结束于峰值处，结束于此类位置处的DNA片段的数目较高(应注意，对于不同定义的相对丰度值，可能对于峰值出现较高似然性)。当更多DNA片段来自组织类型A时，结束于谷值和峰值处的DNA片段的数目的差异将更大。因此，随着组织A的比例贡献增加，结束于谷值的DNA片段的数目与结束于峰值的DNA片段的数目之间的分离将越来越大。此分离值对应于关于组织A在图9中

示出的似然函数中的高幅度。

[0143] B.相对丰度与比例贡献之间的关系

[0144] 图10显示FR_A与混合物中的组织A相对于DNA的比例贡献之间的关系，所述比例贡献通过分析具有已知比例浓度的来自组织A的DNA的两种或更多种校准样品而确定。在示出的实例中，分析具有x_1和x_2的组织A的比例贡献的两个样品。两个样品的FR_A值分别确定为y_1和y_2。可以基于x_1、x_2、y_1和y_2的值确定FR_A与比例贡献之间的关系。

[0145] 值y1和y2是校准值的实例。数据点(x1,y1)和(x2,y2)是校准数据点的实例。校准数据点可以拟合到函数以获得校准曲线1010，其可以是线性的。当对于新样品测量新FR_A(或其它相对丰度值)时，新FR_A可以相比于校准值中的至少一个以确定新样品的比例贡献的分类。可以不同方式进行与校准值的比较。举例来说，校准曲线可以用于发现对应于新FR_A的比例贡献x。作为另一实例，新FR_A可以相比于第一校准数据点的校准值y1以确定作为比例贡献的新样品大于或小于x1。

[0146] 在其它实施例中，可以对于含有超过两种类型的组织的混合物类似地分析组织A的比例贡献，只要其它组织的FR_A相对恒定。此类方法实际上适用于分析不同临床情境，例如(但不限于)癌症检测、移植监测、外伤监测、感染和产前诊断。

[0147] 在一个实施例中，可以确定癌症患者的血浆中的受影响组织的分数浓度。举例来说，在患有肝癌的患者中，可以通过分析肝特异性开放染色质区，例如脱氧核糖核酸酶超敏性位点来确定肝DNA的分数贡献。在一个实施例中，这可以使用DNase-Seq进行(Boyle等人《细胞》2008;132:311-322;Madrigal等人《基因学前沿(Front Genet)》2012;16:123-131)。在另一实施例中，这可以通过甲醛辅助的调节元件分离(FAIRE)-Seq进行(Giresi等人《基因组研究(Genome Res)》2007;17:877-885)。在另一实施例中，这可以通过ATAC-Seq进行(Buenrostro等人《自然方法》2013;10:1213-1218)。$FR_{肝}$可以在这些位点测定且相比于正常的健康个体。在肝特异性脱氧核糖核酸酶超敏性位点处，峰值与谷值区域之间的P_I变化将主要由肝贡献。通过类似于图10地与校准曲线比较，可以确定肝的贡献。测试案例的$FR_{肝}$的值可以相比于健康个体中的肝的贡献的范围。可以使用在结束于混合物的各种组织中的基因组位置处的DNA片段的似然函数中具有高幅度变化的其它区域。此类其它区域的实例更详细地描述于后续部分中。

[0148] 类似地，可以通过此方法确定已接受器官移植的患者中的移植器官的贡献。在前述研究中，已经表明具有排斥反应的患者将导致从移植器官释放的DNA增加，使得血浆中来自移植器官的DNA的浓度升高。移植器官的FR的分析将是检测和监测器官排斥反应的适用方式。用于此类分析的区域可以取决于所移植的器官而变化。

[0149] 在另一实施例中，此方法可以用于确定母体血浆中的胎儿DNA浓度。在母体血浆中，携有胎儿基因型的DNA分子实际源自胎盘。因此，如果我们集中于对胎盘具有特异性但不存在于血细胞中的脱氧核糖核酸酶超敏性位点，那么我们将能够通过分析$FR_{胎盘}$确定胎盘对血浆DNA的比例贡献。

[0150] 图11显示根据本发明的实施例的$FR_{胎盘}$与母体血浆中的胎儿DNA百分比之间的相关性。纵轴对应于如使用位于一个或多个脱氧核糖核酸酶超敏性位点的一个或多个局部最大值和局部最小值确定的$FR_{胎盘}$。横轴为使用独立的测量技术测量的胎儿DNA分数。如可见，$FR_{胎盘}$的值与胎儿DNA分数相关。在此实例中，基于母体为纯合且胎儿为杂合的SNP处的胎儿

特异性等位基因的比例确定胎儿DNA分数。因此，可以使用基于母体血浆DNA的测序结果的$FR_{胎盘}$估计胎儿DNA百分比。

[0151] 或者，由于母体血浆中的两种关键组分是胎盘源性DNA和源自血细胞的DNA(不同组织类型)，我们推论$FR_{血液}$将与血浆中的胎儿DNA的分数浓度负相关。因此，识别对血细胞具有特异性的脱氧核糖核酸酶超敏性位点且确定$FR_{血液}$。

[0152] 图12显示$FR_{血液}$与母体血浆中的胎儿DNA浓度之间的相关性。纵轴对应于如使用位于一个或多个脱氧核糖核酸酶超敏性位点的一个或多个局部最大值和局部最小值确定的$FR_{血液}$。横轴为基于母体血浆中的胎儿特异性等位基因的比例测量的胎儿DNA分数。可以在$FR_{血液}$与胎儿DNA百分比之间观测到负相关。因此，可以使用基于母体血浆DNA的测序结果的$FR_{血液}$估计胎儿DNA百分比。因此，基因组区域可以具有对多种组织类型具有特异性的片段化模式，例如与一些组织正相关且与其它组织负相关。

[0153] C.使用最大值和最小值的方法

[0154] 图13为分析生物样品以确定根据本发明的实施例的第一组织类型的比例贡献的分类的方法1300的流程图。生物样品包括来自多个组织类型(包括第一组织类型)的游离DNA分子的混合物。如同本文所述的其它方法，方法1300可以使用计算机系统。可以基于特定个体选择第一组织类型(例如肝脏组织或胎儿组织)。举例来说，如果个体先前患有肝癌，那么可以进行筛选以检查肝癌是否复发，肝癌复发将导致来自肝脏组织的比例贡献增加。此类选择标准适用于本文所述的其它方法。

[0155] 在步骤1310处，识别至少一个具有对第一组织类型具有特异性的片段化模式的基因组区域。举例来说，至少一个基因组区域可以包括一个或多个脱氧核糖核酸酶超敏性位点。至少一个具有对第一组织类型具有特异性的片段化模式的基因组区域中的每一个可以包括至少一个额外样品中的一个或多个第一组织特异性等位基因，例如如将在第VI部分中描述。作为另一实例，至少一个基因组区域可以包括一个或多个ATAC-seq或微球菌核酸酶位点。第一组织类型可以对应于特定器官或甚至对应于器官的特定癌症。

[0156] 在步骤1320处，分析来自生物样品的多个游离DNA分子。分析游离DNA分子包括确定对应于游离DNA分子的至少一个末端的参考基因组中的基因组位置(结束位置)。因此，可以确定游离DNA分子的两个结束位置，或仅一个结束位置。

[0157] 可以不同方式确定结束位置，如本文所述。举例来说，可以对游离DNA分子进行测序以获得序列读数，且可以将序列读数与参考基因组对应(比对)。如果生物体是人类，那么参考基因组是潜在地来自特定亚群的参考人类基因组。作为另一实例，可以用不同探针分析游离DNA分子(例如在PCR或其它扩增之后)，其中每个探针对应于基因组位置，所述基因组位置可以覆盖至少一个基因组区域。

[0158] 可以分析游离DNA分子的统计显著数目以精确确定来自第一组织类型的比例贡献。在一些实施例中，分析至少1,000个游离DNA分子。在其它实施例中，可以分析至少10,000或50,000或100,000或500,000或1,000,000或5,000,000个或更多个游离DNA分子。

[0159] 在步骤1330处，识别第一基因组位置的第一集合。每个第一基因组位置具有对应于第一基因组位置的游离DNA分子的末端的局部最小值。多个相邻基因组位置可以定义为局部极值(最大值或最小值)，且因此局部最大值不限于仅仅一个位置。

[0160] 在一些实施例中，可以对于多个基因组位置中的每一个确定比率。可以确定结束

于基因组位置且向基因组位置的两侧延伸至少指定数目的核苷酸的第一量的游离DNA分子，例如如关于图1所描述。位于所述基因组位置的第二量的游离DNA分子可以与第一量一起使用以确定比率。可以在比率中识别多个局部最小值和多个局部最大值，例如通过步进贯穿比率值以识别一个或多个存在于极值(最大值或最小值)中的每一个处的连续基因组位置。

[0161] 在步骤1340处，识别第二基因组位置的第二集合。每个第二基因组位置具有对应于第二基因组位置的游离DNA分子的末端的局部最大值。可以与第一集合类似的方式识别第二集合。

[0162] 在步骤1350处，确定结束于至少一个基因组区域中的任一个中的第一基因组位置中的任一个上的游离DNA分子的第一数目。可以不同方式确定第一数目，例如以跨越所有第一基因组位置的总和形式。作为另一实例，可以确定每个基因组位置处的单独的量。因此，确定游离DNA分子的第一数目可以包括确定结束于每个第一基因组位置上的游离DNA分子的第一量，借此确定多个第一量。

[0163] 在步骤1360处，确定结束于至少一个基因组区域中的任一个中的第二基因组位置中的任一个上的游离DNA分子的第二数目。可以与第一数目类似的方式确定第二数目。因此，确定游离DNA分子的第二数目可以包括确定结束于每个第二基因组位置上的游离DNA分子的第二量，借此确定多个第二量。

[0164] 在步骤1370处，使用第一数目和第二数目计算分离值。分离值可以不同方式计算，例如通过第一数目与第二数目的比率，如第1II.A部分中所述。在使用多个最大值和最小值的另一实施方案中，可以确定每个此类基因组位置处的量。计算分离值可以包括确定多个分离比，多个第一量中的一个与多个第二量中的一个的每个分离比。分离值可以使用多个分离比，例如分离比的平均值或中值确定。

[0165] 在步骤1380处，通过比较分离值与一个或多个校准值确定第一组织类型的比例贡献的分类，所述校准值由一个或多个已知第一组织类型的比例贡献的校准样品确定。

[0166] D.无扩增分析

[0167] 步骤1310中的游离DNA分子的分析可以是无扩增。当使用PCR时，测序深度(即覆盖特定核苷酸或结束于参考基因组中的特定核苷酸上的序列读取的数目)不直接反映分析了多少覆盖特定核苷酸的血浆DNA分子。这是因为一个血浆DNA分子可以在PCR过程期间产生多个复本，且多个序列读取可以来源于单一血浆DNA分子。此复制问题将在以下情况下变得更重大:i)用于扩增测序文库的PCR循环的较高数目;ii)增加的测序深度，和iii)原始血浆样品中的DNA分子的较小数目(例如较小血浆体积)。

[0168] 另外，PCR步骤引入另外的误差(Kinde等人《美国国家科学院院刊(Proc Natl Acad Sci USA)》2011;108:9530-9535)，因为DNA聚合酶的保真度不是100％，且偶尔，错误核苷酸将并入到PCR子链中。如果此PCR误差出现于早期PCR循环期间，那么将产生显示相同误差的子分子克隆体。错误碱基的分数浓度可以在来自误差将被曲解为例如胎儿源性或肿瘤源性突变的相同基因座的其它DNA分子中达到此类高比例。无PCR方案的实例包括:Berry Genomics(investor.illumina.com/mobile.view?c＝121127&v＝203&d＝1&id＝1949110);Illumina(www.illumina.com/products/truseq-dna-pcr-free-sample-prep-kits.html)和各种单分子测序技术。无扩增分析的其它细节可见于PCT申请第PCT/CN2016/

073753号中。

[0169] 因此,一些实施例可以包括从待分析的生物样品获得模板DNA分子;使用模板DNA分子制备可分析的DNA分子的测序文库,制备可分析的DNA分子的测序文库不包括对模板DNA分子进行DNA扩增的步骤;对可分析的DNA分子的测序文库测序以获得多个对应于第一多个游离DNA分子的序列读取。分析第一多个游离DNA分子可以包括在计算机系统处接收多个序列读取和通过计算机系统比对多个序列读取与参考基因组以确定多个序列读取的基因组位置。

[0170] IV.左侧和右侧核苷酸的相对丰度

[0171] 图14显示肿瘤或胎儿源性DNA的循环DNA片段的差异原理的图示。在前述研究中,已经证明循环DNA的尺寸紧密类似于核小体DNA的尺寸。血浆DNA的尺寸分布中的166bp的主峰值表示DNA与组蛋白复合物的核心以及连接两个连续组蛋白复合物的连接子DNA相关。

[0172] 也已观测到胎儿和肿瘤源性DNA分子的尺寸分布短于癌症患者和孕妇的血浆中的非肿瘤和非胎儿源性DNA的尺寸分布(Lo等人《科学·转化医学》2010;2(61):61ra91和Jiang等人《美国国家科学院院刊》2015;112:E1317-25.)。对于血浆中的肿瘤和胎儿源性DNA的尺寸分布,166bp的峰值减弱且144bp处的峰值更显著。144bp峰值可能是由于连接两个连续组蛋白复合物的约20bp连接子DNA的降解。

[0173] 为了说明此方法的原理,我们使用癌症患者的情境作为实例。相同原理可接着应用于其它情境,包括分析孕期的母体血浆中的循环胎儿DNA,和分析已接受移植的患者的血浆。实施例可以分析血浆DNA分子的末端,在图14中表示为左末端和右末端。

[0174] 当来自非恶性组织的DNA经片段化且释放至血浆中时,两个分子的连接末端均将位于核苷酸位置A。换句话说,对于右侧上的分子,左侧的最外核苷酸正好紧邻核苷酸位置A。对于左侧上的分子,右侧的最外核苷酸也正好紧邻核苷酸位置A。当相对于核苷酸坐标标绘结束于特定核苷酸处的分子的相对丰度时,末端的峰值丰度将处于位置A以使左侧和右侧的最外核苷酸映射到此区域。对于源自肿瘤细胞的DNA分子,将在片段化过程之后从分子去除20bp片段。

[0175] 因此,将在右侧上的分子的左侧与左侧上的分子的右侧之间存在20bp的间隙。当相对于核苷酸坐标标绘结束于特定核苷酸处的分子的相对丰度时,右侧的最外核苷酸(位于B处)的峰值和左侧的最外核苷酸(位于C处)的峰值将间隔开20bp。因此,结束于核苷酸位置B和C上的分子的丰度与结束于位置A上的分子的丰度之间的比率将表示血浆样品中的肿瘤源性DNA的分数浓度。

[0176] 相同原理可以应用于具有差分尺寸分布的DNA物种的定量,例如(但不限于)测量孕妇的血浆中的胎儿DNA和测量来自移植器官的DNA。

[0177] 图15为分析包括来自多个组织类型(包括第一组织类型)的游离DNA分子的混合物的生物样品的方法1500的流程图。方法1500的一部分可以用于实施步骤1310和识别优选的结束位置的其它步骤。

[0178] 在步骤1510处,分析游离DNA分子以确定参考基因组中的左侧和右侧结束位置。步骤1510可以与步骤1320类似的方式执行。在步骤1510中,可以分析来自个体的生物样品的第一多个游离DNA分子,其中第一多个游离DNA分子中的每一个具有左末端和右末端。可以确定对应于游离DNA分子的左末端的参考基因组中的左侧结束位置,例如通过比对(映射)

DNA片段的序列读取与参考基因组或经由在参考基因组中已知位置的探针。取决于选择用于界定参考基因组的坐标系，左末端可以指任一末端。类似地，可以确定对应于游离DNA分子的右末端的参考基因组中的右侧结束位置。可以在两个独立比对步骤中确定两个结束位置，例如在两个末端具有独立序列读取的情况下。

[0179] 在步骤1520处，识别左侧基因组位置的左侧集合。左侧集合的每个基因组位置具有第一多个游离DNA分子的左末端的局部最大值，所述局部最大值对应于基因组位置的左侧集合中的一个。左侧集合可以与关于方法1300的最大值所描述类似的方式确定。

[0180] 在步骤1530处，识别右侧基因组位置的右侧集合。右侧集合的每个基因组位置具有第一多个游离DNA分子的右末端的局部最大值，所述局部最大值对应于基因组位置的右侧集合中的一个。右侧集合可以与关于方法1300的最大值所描述类似的方式确定。

[0181] 在步骤1540处，基因组位置的第一集合识别为对第一组织类型具有特异性。左侧集合的左侧基因组位置的全部或一部分可以相比于设定成识别基因组位置的第一集合的右侧的右侧基因组位置的全部或一部分，其中左侧基因组位置到最接近的右侧基因组位置的距离大于参考基因组中的基因组位置(例如核苷酸)的第一阈值距离。第一阈值距离的实例是5、6、7、8、9、10、15和20个核苷酸。

[0182] 在步骤1550处，识别基因组位置的第二集合。左侧集合的左侧基因组位置的全部或一部分可以相比于设定成识别基因组位置的第二集合的右侧的右侧基因组位置的全部或一部分，其中左侧基因组位置到最接近的右侧基因组位置的距离小于参考基因组中的基因组位置的第二阈值距离。第二阈值距离的实例是2、3、4和5个基因组位置(例如核苷酸)。

[0183] 在步骤1560处，使用结束于基因组位置的左侧集合中的一个处的第一多个游离DNA分子第一数目和结束于基因组位置的右侧集合中的一个处的第一多个游离DNA分子的第二数目确定分离值。分离值(例如相对丰度值)可以确定于第一数目与第二数目之间。

[0184] 在一个实施例中，识别基因组位置的第一集合与基因组位置的第二集合的配对。配对可以是彼此最接近的位置。对于一个或多个配对中的每一个，可以确定结束于第一基因组位置处的游离DNA分子的第一量，且可以确定结束于第一基因组位置处的游离DNA分子的第二量。游离DNA分子的第一量对应于多个游离DNA分子的第一数目且游离DNA分子的第二量对应于多个游离DNA分子的第二数目。举例来说，第一量可以总计为第一数目且第二量可以总计为第二数目，且可以从第一数目和第二数目直接确定分离值。作为另一实例，可以从对于一个配对各自包括第一量和第二量的多个比率确定分离值。在各种实施方案中，比率的平均值或中值可以用作分离值。配对的对应第一和第二量可以其它方式用于确定个别分离值，所述分离值用于确定总分离值。

[0185] 在步骤1570处，通过比较分离值与一个或多个校准值确定第一组织类型的比例贡献的分类，所述校准值由一个或多个已知第一组织类型的比例贡献的校准样品确定。步骤1570可以与比例贡献的其它确定类似的方式进行。

[0186] 在各种实施例中，左侧和右侧集合均可以用作基因组位置的第一集合；仅可以使用左侧集合；仅可以使用右侧集合；或可以使用来自左侧集合的一些和来自右侧集合的一些。对于整个左侧位置集合，存在左侧位置的子集，其具有以核苷酸的阈值数目与左侧位置的子集分离的对应右侧位置集合。因此，有可能使用左侧位置的子集或右侧位置的对应子集进行计算。

[0187] V.使用组织特异性结束位置

[0188] 我们假设衍生自癌细胞、胎盘细胞的循环DNA的片段化模式和细胞类型将不同。基于此假设，循环DNA片段的一个或两个末端处的末端核苷酸的坐标可以用于预测携有假定突变的DNA片段是否实际上源自肿瘤。可以在血浆DNA片段中识别癌症特异性和怀孕特异性结束位置。

[0189] A.使用肝细胞癌(HCC)的癌症实例

[0190] 为了说明此方法的可行性，分析患有肝细胞癌(HCC)的患者和孕妇的血浆DNA的测序数据。出于说明的目的，分析集中于染色体8上。相同方法可以应用于全基因组或任何其它染色体。

[0191] 确定每个经测序血浆DNA片段的两端处的末端核苷酸的坐标。接着，对结束于染色体8上的每个核苷酸上的片段的数目计数。对于HCC案例和孕妇确定具有最高数目的结束于其上的DNA片段的顶部1百万个核苷酸。顶部一百万可以视为高于阈值。

[0192] 图16为显示对HCC案例具有特异性、对孕妇具有特异性和由两种案例共用的频繁结束位点的数目的文氏图。作为对HCC案例具有特异性的最频繁结束位置的536,772个核苷酸的坐标显示于附录A中。作为对孕妇具有特异性的最频繁结束位置的536,772个核苷酸的坐标在附录B中列出。两种情况共用的最频繁结束位置的463,228个核苷酸的坐标被省去。

[0193] 我们推论具有恰好结束于536,772个HCC特异性结束位置处的末端核苷酸的血浆DNA片段将更可能源自肿瘤。基于此假设，结束于HCC特异性结束位置上的经测序血浆DNA片段的数目可以用于指示存在或不存在HCC或其它具有相同血浆DNA片段化模式的癌症。在另一实施例中，此参数也可以用于反映癌症水平，例如(但不限于)肿瘤尺寸、癌症阶段、肿瘤负荷和癌转移的存在。

[0194] 在另一实施例中，对于具有已知的血浆中的肿瘤DNA分数的样品，结束于HCC特异性结束位置上的片段的数目可以与血浆中的癌源性DNA的分数浓度相关。血浆中的肿瘤DNA分数可以通过例如(但不限于)定量血浆中的癌症突变或血浆DNA中的拷贝数偏差的量值而确定(Chan等人《临床化学(Clin Chem)》2013;59:211-24)。此相关性可以用作校准曲线(图1)。对于具有未知的血浆中的肿瘤DNA分数的患者，可以确定结束于HCC特异性结束位置上的DNA片段的量。接着，可以基于校准曲线和结束于HCC特异性结束位置上的DNA片段的量确定血浆中的肿瘤DNA分数。在一个实施方案中，结束于HCC特异性结束位置上的DNA片段的量可以标准化为测序的DNA片段的总数、可比对读取的总数或与某些染色体区域比对的DNA片段的数目。因此，结束于癌症特异性位置上的经测序DNA片段的比例可以用作参数。

[0195] 图17显示校准曲线，其显示结束于癌症特异性结束位置上的经测序DNA片段的比例与具有已知的血浆中的肿瘤DNA分数的癌症患者的血浆中的肿瘤DNA分数之间的关系。此概念图显示肿瘤DNA分数与结束于癌症特异性结束位置上的序列DNA片段的比例之间的校准曲线的相关性。校准曲线可以通过拟合确定自校准样品的数据点而确定，所述校准样品的肿瘤DNA分数是经由其它技术确定。

[0196] 在本发明的另一实施例中，可以确定罹患不同类型的癌症的患者的血浆DNA片段化模式。这些癌症患者的重叠末端可以视为癌症特异性末端，而个别癌症类型的结束位置可以视为对特定癌症类型具有特异性。对于任何怀疑患有癌症的个体，经测序血浆DNA片段可以首先相比于癌症特异性结束位置以确定个体患有癌症的可能性。如果个体可能患有癌

症,那么可以分析经测序片段的癌症类型特异性结束位置以确定个体最可能罹患的癌症。

[0197] 在本发明的另一实施例中,可以确定源自不同器官的DNA的结束位置且可以用于确定来自不同器官的DNA对血浆的相对贡献。

[0198] B.胎儿实例

[0199] 在另一实施例中,此方法可以用于确定母体血浆样品中的胎儿DNA分数。校准曲线可以通过结束于怀孕特异性结束位置上的经测序血浆DNA片段的比例(首先确定)与具有已知胎儿DNA分数的多种母体血浆样品的胎儿DNA分数的相关性建立。胎儿DNA分数可以通过多种方法确定,例如(但不限于)确定样品中的胎儿特异性等位基因、对于男性怀孕定量染色体Y上的标靶和分析胎儿特异性甲基化标记物。对于具有未知胎儿DNA分数的怀孕血浆样品,可以确定结束于怀孕特异性结束位置上的经测序血浆DNA片段的比例。利用此信息,可以基于校准曲线确定经测试血浆DNA样品中的胎儿DNA分数。

[0200] C.使用优选的结束位置的试剂盒

[0201] 在一些实施例中,提供试剂盒以分析含有多种组织类型的游离DNA分子的混合物的生物样品中的DNA。试剂盒可以包括一种或多种寡核苷酸以特异性地杂交到附录A和B中列出的基因组区域的至少一部分。在一个实施例中,试剂盒包括一种或多种寡核苷酸以特异性地杂交到附录A中列出的基因组区域的至少一部分以用于对个体测试HCC。在另一实施例中,试剂盒包括一种或多种寡核苷酸以特异性地杂交到附录B中列出的基因组区域的至少一部分以用于测试怀孕女性,例如确定来自怀孕女性的母体生物样品中的胎儿DNA分数。

[0202] VI.使用多态性的结束位置分析

[0203] 在一些实施例中,可以使用组织特异性等位基因识别具有组织特异性片段化模式的区域。举例来说,可以通过分析母体血浆样品和比较检测的等位基因与仅母体样品中检测的等位基因而识别胎儿特异性等位基因,如本文所述。相对于组织展现共用等位基因(即,胎儿和母体共用)的比率具有胎儿DNA分子结束于其上的高比率的基因组位置可以识别为具有胎儿组织特异性片段化模式。这些胎儿优选结束位置可以是或可以不是脱氧核糖核酸酶超敏性位点,借此显示各种基因组区域可能具有关于片段化模式的组织特异性幅度,且实施例不限于脱氧核糖核酸酶超敏性位点。可以对于来自经肿瘤筛选的个体的样品进行类似分析。

[0204] A.胎儿实例

[0205] 优选的结束位置可以通过分析来自孕妇的血浆DNA获得。胎儿源性和母源性血浆DNA片段可以通过基于多形性的方法区分。携有胎儿和母体特异性等位基因的片段可以用于确定胎儿源性和母源性DNA的优选的结束位置。

[0206] 在知情同意的情况下,在38周妊娠时从香港威尔士王子医院(Prince of Wales Hospital,Hong Kong)妇产科为此研究募集男性单胎妊娠的孕妇。血液样品在4℃下在1,600g下离心10min。收获血浆部分且在4℃下在16,000g下再离心10min以去除血细胞。血细胞部分在2,500g下再离心,且去除任何残余血浆。来自血细胞和来自母体血浆的DNA分别用QIAamp DNA血液微型试剂盒(Blood Mini Kit)和QIAamp DSP DNA血液微型试剂盒(Qiagen)的血液和体液方案萃取。来自胎盘的DNA根据制造商的组织方案用QIAamp DNA微型试剂盒(Qiagen)萃取。使用Illumina TruSeq无PCR文库制备方案对测序文库进行测序。在双末端模式下使用短寡核苷酸比对程序2(SOAP2)分析双末端测序数据(Li等人《生物信

息学(Bioinformatics)》2009;25:1966-1967)。双末端读取与非重复遮蔽参考人类基因组(Hg19)比对。对于每个末端的比对,允许至多2个核苷酸错配。然后分析2个末端的这些潜在比对的基因组坐标以确定任何组合是否将允许2个末端以正确定向与相同染色体比对,横跨≤600bp的插入物大小,并且映射到参考人类基因组中的单一位置。母体血浆样品测序到单倍体人类基因组的270×覆盖度的深度。母体血细胞、父体血细胞和脐带血细胞使用相同测序方案分别测序到40×、45×和50×单倍体人类基因组覆盖度。

[0207] 为此目的,分析母体血浆DNA中的反复性末端序列。

[0208] 1.识别胎儿特异性结束位置

[0209] 伴随使用非PCR扩增文库的母体血浆DNA样品的极高测序深度的表现,我们研究母体和胎儿基因组中是否可能存在将在血浆DNA产生中优先裂解的位点。为了展示此效应,识别母体为纯合(基因型表示为AA)且胎儿为杂合(基因型表示为AB)的信息性SNP基因座。在此说明性实例中,B等位基因将为胎儿特异性的且A等位基因将被母体和胎儿共用。代表性实例显示于图18中。作为对照,显示获自血细胞且使用声处理人工片段化的DNA样品的测序结果。

[0210] 在血浆DNA中观测到非随机片段化模式。关于作为DNA片段的末端的概率曲线,对于携有胎儿特异性等位基因和由母体共用的等位基因的两组片段中的每一个观测到三个峰。这些峰分别表示母体血浆中的胎儿和母体源性DNA的末端位置的热点。峰的位置在这两组之间很大程度上重叠。相比之下,声处理的DNA的片段化模式似乎是随机的且片段末端概率在整个区域中类似。

[0211] 图18显示携有胎儿特异性等位基因和由母体和胎儿共用的等位基因的血浆DNA的非随机片段化模式的说明性实例。在图的上部部分上,每条水平直线表示一个经测序的DNA片段。DNA片段的末端表示经测序读取的结束位置。片段是根据左侧最外的核苷酸的坐标(最小基因组坐标)分选。在图的下部部分上,显示结束于特定位置上的片段的百分比。X轴表示基因组坐标且SNP位于由虚线指定的中心。

[0212] 我们另外检索具有增加的作为血浆DNA片段的结束位置的概率的坐标。我们基于覆盖信息性SNP的片段聚焦我们的搜索以使得可以分别评估携有胎儿特异性等位基因和由母体和胎儿共用的等位基因的片段。我们使用泊松概率函数确定人类基因组内的某些位置是否具有显著增加的作为血浆DNA片段的结束位置的概率。对于母体为纯合(基因型AA)且胎儿为杂合(基因型AB)的SNP的分析,A等位基因将为“共用等位基因”且B等位基因将为胎儿特异性等位基因。将计数携有共用等位基因和胎儿特异性等位基因的测序读取的数目。在血浆DNA的尺寸分布中,将对于胎儿源性和母体源性DNA观测到166bp处的峰值。如果血浆DNA的片段化是随机的,那么两个末端将跨越信息性SNP上游166bp和下游166的区域均匀分布。

[0213] 可以计算p值以基于泊松概率函数确定特定位置是否具有显著增加的作为携有共用等位基因或胎儿特异性等位基因的读取的末端的概率。

[0214] p值=泊松($N_{实际}$,$N_{预测}$)

[0215] 其中泊松()为泊松概率函数;$N_{实际}$为结束于特定核苷酸处的读取的实际数目;且$N_{预测}$为读取总数除以166。<0.01的p值用作截止值以界定携有胎儿特异性等位基因或共用等位基因的读取的优选的结束位置。对于携有共用等位基因和胎儿特异性等位基因的DNA

片段独立地确定统计显著的结束位置(图19)。可以使用其它概率分布,例如二项分布、负二项分布和正态分布。

[0216] 图19显示基因组坐标作为跨越具有信息性SNP的区域的母体血浆DNA片段的结束位置的概率曲线。具有显著增加的作为携有共用等位基因和胎儿特异性等位基因的血浆DNA片段的末端的概率的核苷酸位置的结果分别以红色和蓝色显示。X轴表示基因组坐标且突变位于通过虚线指定的中心。如所示出,存在仅对于胎儿特异性等位基因、仅对于共用等位基因具有结束位置的高出现率的坐标,且一些坐标为两者共用的。

[0217] 我们识别总共4,131个(集合A)和10,021个(集合B)分别具有显著增加的作为携有胎儿特异性等位基因和共用等位基因的血浆DNA片段的末端的机率的核苷酸位置。集合C是重叠集合且含有4,258个核苷酸位置(图3)。这些结束位置获自总共跨越1.42Mb且覆盖4,303个SNP的区域。因此,胎儿特异性片段的优选的结束位置占所分析区域的0.29%。分别存在24,500、22,942和31,925个携有结束于集合A、集合B和集合C位置上的胎儿特异性等位基因的血浆DNA片段。分别存在27,295、158,632和87,804个携有结束于集合A、集合B和集合C位置上的共用等位基因的血浆DNA片段。预期优选的结束位置的数目或发生率高得多且出现于其它基因组坐标处。

[0218] 如此处所描述的基于多态性的方法仅识别与此胎儿-母体对的信息性SNP相关的优选的结束位置。因此,识别的优选的末端将表示基因组中的此类末端的子集。我们已开发并非基于多态性的方法来识别优选的末端。实际上,识别了许多使用基于非多态性的方法的更优选的结束方法。请参考下文所述的其它实验。

[0219] 图20显示跨越在母体中为纯合且在胎儿中为杂合的SNP的血浆DNA片段的结束位置的分析。集合A包括携有胎儿特异性等位基因的片段的优选的结束位置。集合B包括携有共用等位基因的片段的优选的结束位置。集合C包括两种类型的血浆DNA片段的优选的结束位置。

[0220] 使用相同原理,我们另外分析跨越在母体(基因型AB)中为杂合且在胎儿(基因型AA)中为纯合的SNP的母源性DNA片段的结束位置。我们识别总共7,527个(集合X)和18,829个(集合Y)分别具有显著增加的作为携有胎儿特异性等位基因和共用等位基因的血浆DNA片段的结束位置的机率的核苷酸位置。集合Z为重叠集合且含有10,534个位置(图4)。这些结束位置获自总共跨越3.1Mb且覆盖9,489个SNP的区域。因此,母体特异性片段的优选的结束位置占此母体和胎儿对的所分析区域的0.24%。分别存在69,136、82,413和121,607个携有结束于集合X、集合Y和集合Z位置上的母体特异性等位基因的血浆DNA片段。分别存在46,554、245,037和181,709个携有结束于集合X、集合Y和集合Z位置上的共用等位基因的血浆DNA片段。再次,此分析聚焦于覆盖至少信息性SNP的血浆DNA分子,识别的优选的末端仅表示整个基因组中的此类非随机末端的子集。

[0221] 图21显示跨越在胎儿中为纯合且在母体中为杂合的SNP的血浆DNA片段的结束位置的分析。集合X包括携有母体特异性等位基因的片段的优选的结束位置。集合Y包括携有共用等位基因的片段的优选的结束位置。集合Z包括两种类型的血浆DNA片段的优选的结束位置。

[0222] 2.使用反复性结束位置来推导胎儿DNA分数

[0223] 在识别源自母体和胎儿的血浆DNA片段的反复性结束位置之后,我们推论结束于

这些核苷酸位置集合上的血浆DNA的相对丰度将反映胎儿DNA分数。为了确认这个推论，我们对各自怀有男性胎儿的26个早期妊娠怀孕(10到13周)妇女的血浆DNA测序。中值映射读取计数是1600万(范围：1200万到2200万)。与染色体Y比对的测序读取的比例用于计算每个血浆样品中的实际胎儿DNA分数。可以在具有反复性胎儿(集合A)和母体(集合X)末端的血浆DNA的相对丰度(表示为F/M比率)与胎儿DNA分数之间观测到正相关(R＝0.63，P＝0.0004，皮尔逊相关，图22)。有趣的是尽管优选的结束位置是基于一对胎儿和母体的信息性SNP识别且仅表示基因组中的此类末端的子集，识别的末端也与其它怀孕相关且与胎儿分数的相关性甚至在优选的末端的仅仅这个子集的情况下实现。

[0224] 图22显示具有反复性胎儿(集合A)和母体(集合X)末端的血浆DNA分子的相对丰度(比率(F/M))与胎儿DNA分数之间的相关性。数据点中的每一个可以对应于对应的校准样品，且因此被视为校准数据点。拟合校准数据点的线是校准函数的实例。

[0225] 除集合A和集合X以外，可以使用其它集合。举例来说，可以获取集合A相对于集合C以及集合A相对于集合B的比率(或其它相对丰度或比率的函数)。作为另一实例，可以获取集合X与集合Z的比率或集合X与集合Y之间的比率，其将提供母体DNA分数，可以假设所述母体DNA分数是胎儿DNA分数的倒数。在此实例中，母体组织可以是比例贡献经确定的第一组织类型(即使暗含地)。

[0226] 3.使用尺寸

[0227] 结束于胎儿特异性结束位置上的血浆DNA片段的尺寸分布提供位置具有胎儿特异性的另外的证据。为了进一步支持集合A和集合X位置分别是胎儿源性和母源性DNA片段的优选的结束位点，我们比较结束于这两个位置集合上的血浆DNA的尺寸分布。对于得到这些位置的样品，结束于集合A位置上的片段的尺寸分布较短，短于结束于集合X位置上的那些(图23A)。

[0228] 图23A显示结束于胎儿优选的结束位置上的片段(集合A)(以蓝色)和结束于母体的优选的结束位置上的片段(集合X)(以红色)的血浆DNA尺寸分布。相比于结束于集合X位置上的片段，对于结束于集合A位置上的片段观测到较短尺寸分布。图23B显示两个片段集合的尺寸分布的累积曲线。图23C显示相对于片段尺寸的两个片段集合的累积频率的差异(ΔS)。图23D显示在将集合A和集合X末端位置平移0到5bp到具有较大基因组坐标的位置的情况下，相对于尺寸的ΔS。图23E显示在沿反向方向将集合A和集合X结束位置平移0到5bp(具有较小基因组坐标的位置)的情况下，相对于尺寸的ΔS。

[0229] 为了进一步定量尺寸分布的差异，标绘两个曲线的累积频率(图23B)。由ΔS表示的两个曲线的差异标绘于图23C中。我们观测到最大差异观测于166bp处。这与可以在166bp处观测到胎儿源性与母源性DNA之间的最大差异的前述报导一致(Yu等人《美国国家科学院院刊》2014;111:8583-8)。本发明发现表明相比于结束于母体优选的结束位置的片段(集合X)，结束于胎儿优选的结束位置上的片段(集合A)存在胎儿源性DNA的富集。

[0230] 我们另外通过将集合A和集合X结束位置在基因组上游或下游平移1到5bp来研究这些结束位置的特异性。在沿两个方向平移集合A和集合X结束位置的情况下相对于尺寸标绘ΔS值(图23D和23E)。平移的正数表示平移到具有较大基因组坐标的位置(图23D)且平移的负数表示平移到具有较小基因组坐标的位置(图23E)。将胎儿和母体优选的位置平移甚至1bp也将显著减小结束于这两个位置集合上的DNA片段之间的尺寸差(ΔS)。平移5bp几乎

完全消除尺寸差。这些结果表明相比于结束于通过我们的算法识别的那些优选的结束位置处的读取，结束于那些替代位置处的读取不具有相同的胎儿或母体特异性。这些数据进一步支持我们的如下解释：血浆或游离DNA分子片段或在那些优选的末端位置处极精确地裂解。换句话说，非随机游离DNA片段化方法精确到特定核苷酸的水平。

[0231] 接着，我们分析来自用于胎儿DNA分数分析的26个早期妊娠血浆样品的合并的测序读取。相比于结束于集合X位置上的片段，对于结束于集合A位置上的片段观测到较短尺寸分布(图24A)。

[0232] 图24A显示关于结束于胎儿优选的结束位置上的片段(集合A)(以蓝色)和结束于母体优选的结束位置上的片段(集合X)(以红色)的来自26个早期妊娠孕妇的合并的血浆DNA样品中的血浆DNA尺寸分布。相比于结束于集合X位置上的片段，对于结束于集合A位置上的片段观测到较短尺寸分布。图24B显示两个片段集合的尺寸分布的累积曲线。图24C显示相对于片段尺寸的两个片段集合的累积频率的差异(ΔS)。图24D显示在将集合A和集合X位置平移0到5bp(较大基因组坐标)的情况下的ΔS相对于尺寸。图24E显示在沿反向方向将集合A和集合X位置平移0到5bp(较小基因组坐标)的情况下的ΔS相对于尺寸。结束于两个位置集合上的血浆DNA片段之间的尺寸差(ΔS)将随着这些位置的平移而减小，指示这些位置将精确到单核苷酸水平。

[0233] B.癌症实例

[0234] 相同策略也可以应用于分析癌源性片段的优选的结束位置。在此实例中，我们对罹患肝细胞癌(HCC)的患者的血浆(220×覆盖度)、血沉棕黄层(48×)和肿瘤组织(45×)测序。通过比较肿瘤组织和血沉棕黄层的基因型获得患者的突变概况。为了确定癌源性血浆DNA片段的优选的结束位置，我们分析携有癌症突变的血浆DNA片段。如图24A-24E中所示，HCC患者中的血浆DNA的片段化模式不是随机的。某些核苷酸位置具有增加的作为血浆DNA片段的末端的概率。

[0235] 1.识别癌症特异性结束位置

[0236] 图25显示HCC患者的血浆DNA的非随机片段化模式的说明性实例。在图的上部部分上，每条水平直线表示一个经测序的DNA片段。红线和蓝线分别表示携有野生型和突变等位基因的DNA片段。DNA片段的末端表示经测序读取的结束位置。片段是根据左侧最外的核苷酸的坐标(最小基因组坐标)分选。在图的下部部分上，显示结束于特定位置上的片段的百分比。X轴表示基因组坐标且突变位于通过虚线指定的中心。

[0237] 我们使用如先前描述的泊松概率分布函数识别具有增加的作为携有突变等位基因和野生型等位基因的血浆DNA片段的末端的概率的基因组位置。0.01的p值用作阈值。相反的也是正确的，如PCT申请第PCT/CN2016/073753号中所述，即当识别具有特定末端的血浆DNA分子时，取决于用于血浆DNA数据解释的末端集合，分子上的SNP等位基因或突变更可能是癌源性的、疾病相关的或怀孕相关的。

[0238] 图26是基因组坐标作为跨越具有突变位点的区域的血浆DNA片段的结束位置的概率曲线。具有显著增加的作为携有野生型等位基因和突变等位基因的血浆DNA片段的末端的概率的核苷酸位置的结果分别以红色和蓝色显示。X轴表示基因组坐标且突变位于通过虚线指定的中心。如所示出，存在仅对于突变体特异性等位基因、仅对于野生型等位基因具有结束位置的高出现率的坐标，且一些坐标为两者共用的。

[0239] 图27A显示跨越突变存在于肿瘤组织中的基因组位置的血浆DNA片段的结束位置的分析。集合E包括携有突变等位基因的片段的优选的结束位置。集合F包括携有野生型等位基因的片段的优选的结束位置。集合G包括两种类型的血浆DNA片段的优选的结束位置。

[0240] 2.使用反复性结束位置来推导肿瘤DNA分数

[0241] 由于集合E位置是癌源性DNA的优选的结束位点且集合F位置是主要源自非肿瘤组织的背景DNA的优选的结束位点,我们假设结束于这两个位置集合上的片段之间的比率将与源自肿瘤的DNA相关。因此,我们分析血浆含有至少1%的肿瘤源性DNA的71个HCC患者的血浆。先前分析这些患者的血浆DNA的拷贝数偏差且通过拷贝数偏差的量值来估计肿瘤DNA分数。(Jiang等人《美国国家科学院院刊》2015;112:E1317-25)。结束于这两个位置集合上的片段之间的比率(比率$_{M/WT}$)被定义为:

[0242] 比率

[0243] /WT=结束于集台E位置上的血浆DNA片段的数目结束于集台F位置上的血浆DNA片段的数目

[0244] 图27B显示71个HCC患者的比率$_{M/WT}$与血浆中的肿瘤DNA分数之间的相关性。观测到比率$_{M/WT}$与血浆中的肿瘤DNA分数之间的正相关(r=0.53,p<0.001,皮尔逊相关)。这些结果表明结束于这些癌症优选的结束位置上的片段的数目将适用于预测癌症患者的血浆中的肿瘤源性DNA的量。

[0245] 一些实施例可以通过多种癌症特异性或癌症相关的变化,例如单核苷酸突变,与癌症特异性或癌症相关的DNA甲基化签名(例如5-甲基胞嘧啶和羟甲基化的位置)、癌症特异性或癌症相关的短血浆DNA分子、癌症特异性或癌症相关的组蛋白修饰标记和癌症特异性或癌症相关的血浆DNA末端位置组合的组合检测来增加可接近的信息性癌症DNA片段的数目。某些癌症特异性或癌症相关的变化可以用作识别突变中的过滤标准。

[0246] VII.多态性非依赖性结束位置分析

[0247] 在其它实施例中,可以如下获得优选的结束位置:(A)比较来自不同个体的血浆DNA片段的结束位置或(B)比较在不同时间点从一个个体获取的样品的血浆DNA片段的结束位置。

[0248] A.罹患不同病理性和生理性病况的个体中的优选的结束位置之间的比较

[0249] 1.使用高于阈值的排它性集合

[0250] 基于泊松分布概率函数,我们已识别具有增加的作为前述部分中描述的孕妇和HCC患者的血浆片段的结束位置的概率的基因组位置。在此分析中,虚无假设(null hypothesis)为所有血浆DNA片段将随机片段化以使得每个基因组位置将具有相同的作为血浆DNA片段的末端的概率。假设血浆DNA片段的平均尺寸为166bp。p值计算为

[0251] p值=泊松($N_{实际}$,$N_{预测}$)

[0252] 其中泊松()为泊松概率函数;$N_{实际}$为结束于特定核苷酸处的读取的实际数目;且$N_{预测}=\text{总读取数目}_{3\times\frac{10}{9}\times 166}$,分母中的$3\times10^9$表示基因组中的核苷酸的数目。

[0253] 使用本亚明和霍赫贝格校正(Benjamini and Hochberg correction)(Bejamini等人《皇家统计学会杂志(Journal of the Royal Statistical Society)》,1995;57:289-300)调节p值以实现<1%的预期错误发现率(FDR)。

[0254] 图28A显示孕妇和HCC患者的血浆DNA的优选的结束位置的数目。集合P含有2千9百

万个在孕妇中为优选的结束位置。集合Q含有6百万个在HCC患者中为优选的结束位置。集合S为重叠集合且含有1千5百万个结束位置。

[0255] 我们假设当相比于结束于怀孕优选的结束位置上的那些片段(集合P)时,结束于HCC优选的结束位置上的片段(集合Q)将富集癌源性DNA。

[0256] 因此,我们将比率$_{HCC/Preg}$计算为

[0257] 比率

[0258] CC/Preg=结束于集合Q位置上的血浆DNA片段的数目结束于集台P位置上的血浆DNA片段的数目

[0259] 且将此比率与上文所提及的71个HCC患者中的肿瘤DNA分数相关联。

[0260] 图28B显示在71个HCC患者的比率$_{HCC/Preg}$与血浆中的肿瘤DNA分数之间观测到正相关。这些结果表明结束于特定病况的优选的结束位点上的片段的数目或比例能够适用于检测所述病况或定量从患病器官释放的DNA的量。

[0261] 2.使用具有较高结束速率的基因组位置的集合

[0262] 在另一实施例中,可以通过确定结束于此类位置上的片段的数目与覆盖所述位置但不结束于所述位置上的片段的数目之间的比率来识别优选的结束位点。图29A说明优选的末端终止比(PETR)的计算。

[0263] PE

[0264] 但不结束于其上的DNA片段的数目

[0265] 图29A显示PETR的概念的图示。每条线表示一个血浆DNA片段。这些片段被标记为a到g。片段a、b、c和d封端于所关注的核苷酸上。片段e、f和g覆盖所关注的核苷酸但不结束于此类位置上。在此说明性实例中,PETR等于4/3,即1.33。在其它实施例中,分母可以是覆盖核苷酸的DNA片段的数目,无论DNA片段是否结束于所述位置上。

[0266] PETR的的计算可以用于识别在罹患不同疾病病况的个体中作为优选的末端的核苷酸位置。以下实例表明PETR的效用。比较先前提及的HCC患者和具有慢性B型肝炎病毒(HBV)感染但无癌症的个体(HBV携带者)的血浆样品。HBV携带者的血浆DNA样品测序到215×单倍体基因组覆盖度。对于每个个体的每个基因组位置计算PETR。7,350,067个基因组位置(集合H)识别为相比于HBV携带者,在HCC患者中具有高至少4倍的PETR。相比于HBV携带者,这些位置在HCC患者中具有至少4倍增加的作为血浆DNA片段的末端的机率。可以使用其它倍数差,例如1.5倍、2倍和3倍。

[0267] 来自11个独立HCC患者的血浆样品另外测序到低得多的测序深度。2千8百万测序读取的平均值获自这11个血浆样品。对于这11个HCC患者中的每一个计算7,350,067个集合H位置处的平均PETR且与血浆中的肿瘤DNA分数相关联。基于如先前所述的血浆中的拷贝数偏差的量值计算血浆中的肿瘤DNA分数(Chan等人《美国国家科学院院刊》2015;112:E1317-25)。

[0268] 图29B显示11个HCC患者的血浆中的肿瘤DNA分数与集合H位置处的PETR之间的相关性。可以在两个参数之间观测到正相关,表明HCC优选的(集合H)位置处的平均PETR将适用于指示血浆中的肿瘤DNA的量。

[0269] 3.确认结束位置为肝相关的

[0270] 为了显示存在于HCC血浆DNA样品或HBV血浆DNA样品中的优选的结束位置为肝相

关的，我们在手术去除HCC之前和之后从患者收集的血浆样品中搜索其存在。数据显示于表1中。手术前和手术后样品分别测序到17×和20×单倍体基因组覆盖度。

[0271]

	HCC 优选的结束位点	HBV 优选的结束位点
HCC 1 中的手术前优选的结束位点	92	16
HCC 1 中的手术后优选的结束位点	5	4

[0272] 表1显示在去除患有HCC的患者中的肝肿瘤的手术之前和之后收集的血浆样品中的HCC优选的结束位置和HBV优选的结束位置。

[0273] 如表1中可见，HCC和HBV优选的结束位置的数目均减少。HBV数据表明大部分优选的结束位置是肝源性的且其减少是由手术之后的肝细胞块的减少所致。因此释放至血浆中的肝源性游离DNA分子减少。有趣的是应注意，手术后消失的手术前样品中存在多超过5倍的HCC优选的结束位置。显示手术后消失的优选的末端中的一些为肝源性的。鉴于在相同手术前样品中检测到相比于HBV优选的末端多得多的HCC优选的末端的观测结果，表明那些末端中的大部分是HCC特异性的且并非仅仅属类上肝相关的。

[0274] 存在多种可以从这些数据导出的应用。数据指示游离DNA或血浆DNA优选末端的检测可以用于癌症治疗监测。举例来说，优选末端的手术后减少指示手术去除HCC的成功。如果肿瘤未完全或成功地去除，那么血浆DNA优选末端的量或数量将不展示手术之后的大量减少。这是因为其余的肿瘤或转移性病灶将是HCC优选的结束位置继续释放游离DNA或血浆DNA的来源。数据显示可以在相对较浅测序深度下实现基于游离DNA优选末端的分析的治疗监测。

[0275] 数据也显示组织相关或癌症相关的血浆DNA优选的结束位置可以用于识别病变组织，包括具有癌症的组织。举例来说，可以使用源自不同器官的游离DNA优选末端的多个集合。接着将能够确定源自各种组织的游离DNA的相对量。因此，这可以充当游离DNA组织解卷积的方法。通过此方法显示与从对照样品确立的参考值具有最大偏差(显著增加或显著减少)的组织将为具有病变(例如发炎或病毒感染，正如在慢性B型肝炎病毒携带者中)或癌症的器官组织。

[0276] 用于支持血浆DNA HCC优选的末端具有癌症或HCC特异性的另一项证据，我们研究显示HCC或HBV优选的末端的血浆DNA分子的尺寸概况(图30)。

[0277] 图30显示在结束于HCC优选的末端、HBV优选的末端或共用末端的血浆DNA分子中检测到的短DNA(＜150bp)的比例。图30显示展现HCC优选的末端的血浆DNA分子一般比显示HBV优选的末端的那些短得多(高比例的短DNA)。Jiang等人(Jiang等人《美国国家科学院院刊(Proc Natl Acad Sci U S A.)》2015;112:E1317-25)先前使用另一种方法来显示肿瘤源性血浆DNA分子短于背景非肿瘤DNA。由于具有HCC优选的末端的血浆DNA分子短得多，其高度可能是肿瘤源性的。因此，可以提高在甚至更低测序深度下检测具有HCC优选的末端的血浆DNA分子的机率，可以使样品富含短DNA。

[0278] 4.基于窗口的结束速率

[0279] 在另一实施例中，可以延伸HCC优选的位置以包括相邻核苷酸。图31A说明此方法。将确定结束于窗口A内的片段的数目与结束于窗口B内的那些之间的基于窗口的PETR(w-PETR)比率。可以调节窗口A和窗口B的大小以实现所需性能。可以实验方式获得不同窗口大小的性能。可以设定窗口A的大小，例如(但不限于)5bp、6bp、7bp、8bp、9bp、10bp、15bp、

20bp、25bp和30bp。窗口B的大小将大于窗口A的大小且可以经设定，例如(但不限于)20bp、25bp、30bp、40bp、50bp、60bp、70bp、80bp、100bp、120bp、140bp、160bp、180bp和200bp。在以下说明性实例中，窗口A和窗口B的大小分别设定为20bp和150bp。

[0280] 图31A显示w-PETR的原理的说明。w-PETR的值计算为结束于窗口A与窗口B内的DNA片段的数目之间的比率。窗口A较大且可以在实施标准PETR时宽度为一。窗口B经显示较大。两个窗口均显示为在优选的结束位置处居中，但可以使用窗口的其它定位。在一些实施例中，窗口A可以对应于优选的结束窗口。

[0281] 图31B显示11个HCC患者中的肿瘤DNA分数与w-PETR的值之间的相关性。这些结果表明w-PETR将适用于确定癌症患者的血浆中的肿瘤源性DNA的量。

[0282] 5.使用每个样品的最高结束位置

[0283] 我们比较来自孕妇、一个慢性B型肝炎病毒携带者(HBV)、一个肺癌患者和两个HCC患者的数据之间的顶部1百万个最频繁表示的游离DNA结束位置。对于HCC患者，使用无PCR方案制备一个案例(HCC)的测序文库且使用基于PCR的方案制备另一样品(HCC(PCR)。使用无PCR方案制备所有其它样品。图32显示当相比于脐带血血浆样品(210×单倍体基因组覆盖度)时，所研究样品中的每一个的血浆样品中检测的通常共用的优选的结束位置的比例。

[0284] 图32显示当相比于脐带血血浆样品(210×单倍体基因组覆盖度)时，所研究样品中的每一个的血浆样品中检测的通常共用的优选的结束位置的比例。对于使用PCR检测的怀孕、HCC、HBV、肺癌和HCC中的每一个的常染色体显示百分比。

[0285] 高通用性水准再次支持血浆DNA片段化并非随机过程的概念。HCC和HCC(PCR)数据显示可以使用具有或不具有PCR的任何文库制备方案进行优选的结束位置分析。有趣的是应注意，仍存在一定比例的未显示共同末端的血浆DNA分子。非共同末端是代表生理状态，例如怀孕、胎儿或用于样品的胎盘;或疾病状态，例如癌症的优选末端。血浆DNA优选末端的较详细比较显示于图33中。

[0286] 图33显示文氏图，所述文氏图显示通常观测于两个或更多个样品中的优选的结束位置以及仅观测于任何一个样品中的优选的结束位置的数目。肺癌患者的血浆DNA在175×单倍体基因组覆盖度下测序。

[0287] 从图33值得注意的是115,305个优选的末端在所有三个样品中是共同的。这些可能源自背景血浆DNA的主要来源，例如血细胞。分析也显示存在61,035个在HCC患者和肺癌患者的血浆样品中观测到的优选的结束位置。这些优选的末端可能为多种癌症所共用。因此，其为癌源性的。然而，存在仅检测于HCC患者(479,766个末端)或肺癌患者(749,237个末端)但未检测于两者的血浆DNA分子中的末端。这些优选的末端因此显示较高特异性水平。其对特定癌症组织类型具有特异性。基于相同基本原理，可能能够使用类似挖掘策略来识别对特定器官和特定组织学类型的癌症具有特异性的末端。展现不同类别的末端的血浆DNA分子可以用于各种应用。举例来说，可以旨在检测HCC或肺癌特异性末端以直接检测或筛选特定癌症类型。可以使用HCC和肺癌样品共用的末端来检测或筛选一般的癌症。可以使用最一般的共同末端作为将检测的疾病相关的优选末端的量标准化的分母。也可以出于筛选任何疾病征象(如一般健康状况筛选)的目的检测的一般共同末端。此类测试的阳性结果可以充当向开业医生问诊以进行较详细研究的警示。

[0288] B.在不同时间点从样品个体收集的样品之间的优选的结束位置之间的比较

[0289] 特定病况的优选的结束位置也可以通过比较在不同时间点收集的样品的片段末端而获得。举例来说，在癌症患者中，可以在诊断时收集一个血浆样品且可以在治疗之后(例如在手术切除肿瘤之后)收集另一样品。结束位置的差异可以潜在地反映后者中的癌源性DNA的贡献或对癌症的身体反应的不存在。在另一实例中，可以在在胎儿分娩之前与之后从孕妇收集的血浆样品之间进行比较。

[0290] 在以下实例中，分析从8个孕妇收集的血浆样品。对于每个孕妇，在分娩之前收集血浆样品。在8个妇女中的6个中，在分娩时收集额外血浆样品。在分娩之后6小时向前从八个孕妇收集多个样品且收集总共28个产后血浆样品。血浆DNA样品测序到6.49×单倍体基因组覆盖度的平均深度。将在分娩之前和分娩时收集的样品的测序读数合并在一起用于PETR分析且这些读数将被称作“产前读数”。将分娩之后6小时或随后收集的样品的测序读数合并用于PETR分析且这些读数将被称作“产后”读数。为了识别作为关于怀孕的优选末端的核苷酸位置，检索相比于“产后”读数，PETR在“产前”读数中高至少4倍的位置。识别总共45,281个位点。

[0291] 募集各自怀有男性胎儿的8个早期妊娠孕妇的独立群体且对其血浆DNA测序。2千万测序读取的中值获自这些血浆DNA样品。对于8个孕妇中的每一个确定45,281个位点的平均PETR值且这些值与血浆中的胎儿DNA分数相关，所述胎儿DNA分数估计自与Y染色体比对的读取的比例(Chiu等人《英国医学杂志(BMJ)》2011;342:c7401)。

[0292] 图34A显示血浆中的胎儿DNA分数与通过“产前”与“产后”血浆DNA样品之间的比较识别的位置集合上的平均PETR之间的相关性。这些结果表明识别的位置集合将对于胎儿源性DNA优选且PETR分析将适用于定量母体血浆中的胎儿DNA。

[0293] 与先前描述的方法类似，我们已将w-PETR分析应用于此怀孕优选的位置集合。窗口A和窗口B的大小分别设定为20bp和150bp。在其它实施例中，可以使用其它窗口大小。

[0294] 图34B显示血浆中的胎儿DNA分数与通过“产前”与“产后”血浆DNA样品之间的比较识别的位置集合上的平均PETR之间的相关性。这些结果表明对这些怀孕优选的位置的w-PETR分析将适用于定量母体血浆中的胎儿DNA。

[0295] C.相同病况中的共同端点

[0296] 我们比较两个孕妇的血浆中的顶部1百万个最频繁观测的优选的结束位置(图35A)。

[0297] 图35A显示在18周(怀孕个体1)和38周怀孕期(怀孕个体2)的两个孕妇中的顶部1百万个最频繁观测的血浆DNA优选的结束位置。数据显示这些妇女共用217,947个优选的末端。鉴于两个妇女均怀孕，这些末端源自胎儿、胎盘或在怀孕期间具有增加的细胞死亡(血浆DNA产生)的器官。这些标记物因此最适用于监测怀孕或胎儿的健康。

[0298] 我们计算此样品集合的PETR值。有趣的是，在两个母体血浆样品中的血浆DNA分子的PETR值之间观测到相关性(皮尔逊相关＝0.52，p值＜0.0001)(图35B)。

[0299] 图35B显示两个孕妇的血浆中的顶部1百万个最频繁观测的优选的结束位置的PETR值的比较。较高程度的相关性再次指示血浆DNA片段化高度精心安排的。一些结束位点比其它结束位点更“优选”。有趣的是，甚至在顶部1百万个“最优选”位点中存在PETR的相对宽动态范围。如果要选择若干个优选末端或优选末端的子集用于目标检测，例如对疾病进行测试，那么应选择在所关注的疾病组中通常共用的那些，理想地未在无疾病的对照组中

观测到或较不普遍且尤其是具有极高PETR的结束位置。

[0300] VIII.使用组织特异性结束位置的方法

[0301] 图36是分析生物样品以确定根据本发明的实施例的混合物中的第一组织类型的比例贡献的分类的方法3600的流程图。生物样品包括来自多个组织类型(包括第一组织类型)的游离DNA分子的混合物。

[0302] 在步骤3610处,识别所述第一组织类型的游离DNA分子的末端以高于阈值的比率出现的基因组位置的第一集合。关于步骤3610,以及关于进行优选的结束位置的识别的其它步骤的其它细节在第X.B部分中。其它方法的其它步骤的细节也可发现于第X部分中。

[0303] 在步骤3620处,分析来自个体的生物样品的第一多个游离DNA分子。分析游离DNA分子包括确定对应于游离DNA分子的至少一个末端的参考基因组中的基因组位置。步骤3620可以用于分析游离DNA分子的其它步骤,例如步骤1320类似的方式进行。

[0304] 在步骤3630处,确定第一数目的第一多个游离DNA分子结束于多个窗口中的一个内。所述确定是基于第一多个游离DNA分子的分析而进行。每个窗口包括基因组位置的第一集合中的至少一个。

[0305] 在步骤3640处,计算结束于多个窗口中的一个内的第一多个游离DNA分子的相对丰度。相对丰度可以通过使用游离DNA分子的第二数目对第一多个游离DNA分子的第一数目标准化而确定。第二数目的游离DNA分子包括结束于基因组位置的第二集合处的游离DNA分子,所述基因组位置的第二集合在包括基因组位置的第一集合的多个窗口之外。

[0306] 如关于图27A所描述,基因组位置的第二集合可以使得第二组织类型的游离DNA分子的末端以高于阈值的比率出现于至少一个额外样品中,其中第二组织类型在至少一个额外样品中具有多个第二组织特异性等位基因。基因组位置的第二集合可以使用包括多个第二组织特异性等位基因中的至少一个的至少一个额外样品的游离DNA分子确定。由于集合G可以从用于确定图27B的两个集合排除,在第一组织类型与第二组织类型之间具有共用等位基因的游离DNA分子的末端以高于阈值的第二比率出现的基因组位置可以从基因组位置的第一集合排除且从基因组位置的第二集合排除。

[0307] 在步骤3650处,通过比较相对丰度与一个或多个校准值确定第一组织类型的比例贡献的分类,所述校准值由一个或多个已知第一组织类型的比例贡献的校准样品确定。

[0308] 如果比例贡献高,那么可以进行另外的行动,如个体的治疗性干预或成像(例如在第一组织类型对应于肿瘤的情况下)。举例来说,研究可以使用成像模式,例如计算机断层(CT)扫描或可以进行个体(整个个体或身体的特定部分(例如胸部或腹部),或确切地说,候选器官)的磁共振成像(MRI)以确认或排除个体中肿瘤的存在。如果确认存在肿瘤,那么可以进行治疗,例如手术(通过手术刀或通过辐射)或化疗。

[0309] 治疗可以根据确定的癌症水平、识别的突变和/或来源组织提供。举例来说,识别的突变(例如用于多态实施)可以用特定药物或化疗来靶向。来源组织可以用于指导手术或任何其它形式的治疗。并且,癌症水平可以用于确定任何类型的治疗的侵袭性程度,其也可以基于癌症水平确定。

[0310] IX.确定基因型

[0311] 鉴于可以对于特定组织类型确定优选的结束位置,结束于此类优选的结束位置处的游离DNA分子具有来自所述组织的高可能性。在一些情况下,游离DNA混合物中的特定组

织类型可以在相对于其它组织类型的特定基因组位置处具有不同基因型。举例来说，胎儿组织或肿瘤组织可以具有不同基因型。由于游离DNA分子具有来自所关注的组织类型的高可能性，可以分析结束于此类位置处的游离DNA分子以确定所述位置处的组织类型的基因型。以此方式，优选的结束位置可以用作过滤器以识别来自所述组织类型的DNA。

[0312] A.胎儿基因型

[0313] 关于经测序血浆DNA片段的结束位置的信息可以用于确定哪个母体等位基因已被来自孕妇的胎儿遗传。此处，我们使用假设的实例来说明此方法的原理。我们假设母体、父体和胎儿的基因型分别为AT、TT和TT。为了确定胎儿基因型，我们需要确定胎儿从母体遗传A或T等位基因。我们先前描述了称作相对突变剂量(RMD)分析的方法(Lun等人《美国国家科学院院刊》2008;105:19920-5)。在此方法中，将比较母体血浆中的两个母体等位基因的剂量。如果胎儿遗传母体T等位基因，那么胎儿将对于T等位基因纯合。在此情境下，相比于A等位基因，T等位基因将过度表示于母体血浆中。另一方面，如果胎儿从母体遗传A等位基因，那么胎儿的基因型将为AT。在此情境下，A和T等位基因将以大致相同剂量存在于母体血浆中，因为母体和胎儿都将对于AT杂合。因此，在RMD分析中，将比较母体血浆中的两个母体等位基因的相对剂量。可以分析经测序读取的结束位置以提高RMD方法的精确性。

[0314] 图37显示当与胎儿特异性结束位置附近的参考基因组比对时携有不同等位基因的母体血浆DNA分子。呈实线的分子源自母体且呈虚线的分子源自胎儿。胎儿DNA分子更可能结束于怀孕特异性结束位置上。在一个实施例中，结束于怀孕特异性结束位置上的分子可以在RMD分析中被给予更高权重。在另一实施例中，仅结束于怀孕特异性位置上的血浆DNA片段用于下游分析。此选择可以潜在地富集胎儿源性血浆DNA片段以用于下游分析。

[0315] 图37显示基因型为AT的孕妇中的血浆DNA分子。源自母体组织的DNA片段呈实线且源自胎儿的DNA片段呈虚线。胎儿DNA分子更可能结束于怀孕特异性结束位置上。

[0316] 在此说明性实例中，结束于怀孕特异性结束位置上的两个分子均携有T等位基因。在一个实施例中，仅结束于怀孕特异性结束位置上的两个分子用于下游分析且胎儿基因型将被推论为TT。在另一实施例中，携有T等位基因的两个胎儿源性分子将在RMD分析中被给予更高权重，因为这两个分子结束于怀孕特异性结束位置上。可以给予结束于怀孕特异性结束位置上的分子不同权重，例如(但不限于)1.1、1.2、1.3、1.4、1.5、2、2.5、3和3.5。

[0317] 作为一个实例，确定基因座是否是杂合的标准可以是两个等位基因的阈值，所述等位基因各自以至少预定百分比(例如30％或40％)的与基因座比对的读取呈现。如果一种核苷酸以足够百分比(例如70％或更大)呈现，那么可以将基因座确定为CG中纯合的。

[0318] B.癌症基因型

[0319] 可以对于癌症特异性结束位置执行类似技术。举例来说，可以如上文所述地识别癌症优选的结束位置。可以识别和分析结束于癌症优选的结束位置上的游离DNA分子。可以对于此集合的每个游离DNA分子确定与此位置对应(例如比对)的碱基，且可以对于每个碱基计算总碱基的百分比。举例来说，可以确定结束于所述位置处的游离DNA分子上可见的Cs的百分比。如果C在个体的健康组织中不可见，那么C可以在识别足够数目，例如高于阈值数目(其可以取决于样品中的测量的肿瘤DNA分数)的Cs的情况下识别为突变。

[0320] C.过滤技术

[0321] 除使用结束位置以外的其它标准可以用于过滤来自肿瘤组织的游离DNA分子。其

它标准也可以用于胎儿情境。

[0322] 可以通过向一个或多个序列读取具有已比对的突变的基因座应用过滤标准而改进识别癌症基因型(例如包括癌症特异性突变)和使用此类基因型的任何测试(例如使用突变负荷来确定癌症水平)的特异性。作为癌症的实例,可以通过仅在基因或基因组签名与癌症相关存在高置信度时将其评定为阳性而实现高特异性。这可以通过使可能误识别为突变的测序和比对误差的数目最小化,例如通过与健康对照组的基因组概况比较实现,和/或可以通过与个人自身的组成DNA比较实现和/或可以通过与更早时间的个人的基因组概况比较实现。

[0323] 各种标准可以应用作过滤标准以评估游离DNA片段源自肿瘤且因此有资格作为信息性癌症DNA片段的可能性。每个过滤标准可以个别地、独立地、与相同权重或不同权重集体地,或以指定顺序连续地,或取决于先前过滤步骤的结果条件性地使用。对于条件性使用,可以使用基于贝叶斯(Bayesian)的方法,以及基于分类或决策树的方法。个别地使用标准可以意指仅使用一个标准。独立使用可以涉及超过一个过滤标准,但相比于以特定顺序的连续应用,每个过滤标准不取决于另一过滤标准的应用(例如可以进行平行应用)。作为使用权重的集体使用,可以使用机器学习技术。举例来说,监督学习可以使用具有已知分类的样品的经测量突变负荷来训练任何模型。来自大量个体(例如数百个、数千个或数百万个)的测序数据可以用于训练模型。在更简单形式中,此类已知样品可以用于对于一个或多个确定自过滤标准的评分确定阈值以确定突变是否有效。

[0324] 如果DNA片段显示超过一个癌症特异性变化,那么可以给予其较高权重的信息量或癌症特异性。举例来说,许多癌症为全域低甲基化的,尤其在非启动子区。已显示癌症DNA短于血浆中的非癌症DNA。肿瘤源性血浆DNA片段倾向于在一些特定位置处分段。因此,尺寸较短(例如<150bp)(Jiang等人《美国国家科学院院刊》2015;112:E1317-1325)、具有一个或两个落入癌症相关的末端位置上的末端、显示单核苷酸突变且定位到非启动子区,且具有低甲基化的CpG位点的血浆DNA片段将被认为更可能与癌症相关。低甲基化的DNA的检测可以通过使用亚硫酸氢盐DNA转化或可以区分甲基胞嘧啶与非甲基胞嘧啶的直接单分子测序实现。在本申请中,我们描述增加信息性癌症DNA片段识别中的特异性的方法、方案和步骤。举例来说,一个或多个过滤标准可以用于增加特异性。举例来说,一个或多个过滤标准可以用于增加特异性,如达到大约至少80%、90%、95%或99%的特异性。

[0325] 1.使用血浆DNA末端位置

[0326] 如上文所述,可以进行基于末端核苷酸的坐标(结束位置)过滤潜在的癌症特异性或癌症相关的或胎儿突变。如上文所述,我们已识别并非随机且基于来源组织而变化的DNA片段的末端位置。因此,末端位置可以用于确定具有假定突变的序列读取实际上来自胎儿组织或肿瘤组织的可能性。

[0327] 最近,已经证明血浆DNA的片段化模式为非随机的(Snyder等人《细胞》2016;164:57-68和PCT WO 2016/015058A2)。血浆DNA片段化模式受贡献血浆DNA分子的细胞的基因组中的核小体定位、转录因子结合位点、脱氧核糖核酸酶切割或高敏位点、表达谱(Snyder等人《细胞》2016;164:57-68和PCT WO2016/015058;Ivanov等人《BMC基因组学》2015;16增刊13:S1)和DNA甲基化概况(Lun等人《临床化学》2013;59:1583-1594)影响。因此,不同组织来源的细胞的片段化模式不同。尽管存在显示更频繁片段的基因组区域,但所述区域内的实

际血浆DNA切割位点仍可能是随机的。

[0328] 我们假设不同组织与具有不同切割位点或末端位置的血浆DNA片段的释放相关。换句话说，甚至特异性切割位点都是非随机的。实际上，我们显示癌症患者中的血浆DNA分子的子集显示与无癌症患者不同的末端位置。一些实施例可以使用具有此类癌症相关的末端位置的血浆DNA分子作为信息性癌症DNA片段，或使用此类末端位置信息作为过滤标准，例如连同一个或多个其它过滤标准。因此，在识别此类癌症相关的血浆DNA结束位置的情况下，可以对血浆DNA片段评分来作为信息性癌症DNA片段或基于此类片段的末端位置的性质对权重差异作出归属。此类标准可以用于评估源自癌症、某些器官或某些器官的癌症的片段的可能性。此类权重可以用于将特定DNA片段的特定碱基的贡献修改为所述位置处可见的特定碱基的总百分比。

[0329] 因此，如果血浆DNA片段显示假定突变和/或癌症相关的甲基化改变，以及癌症相关的末端位置，那么血浆DNA片段为信息性癌症DNA片段的机率将高得多。各种实施例也可以将此类片段的状态和其长度，或此类参数与其它参数的任何组合考虑在内。对于具有两个末端(或潜在地至多四个末端，如以下部分中所述)的血浆DNA片段，可以通过考虑其末端中的一个或两个是否与癌症相关或来自与癌症相关的组织类型而进一步修改权重以将其识别为癌源性片段。在一个实施例中，基于末端位置的类似方法也可以用于检测与其它病变或生物过程相关的突变(例如由老化过程所致的突变或由环境诱变因素所致的突变)。

[0330] 类似方法也可以用于通过对怀有胎儿的孕妇的血浆中的DNA测序而识别胎儿的重新突变。因此，如果母体血浆中的此类DNA片段也携有胎盘特异性或胎盘富集结束位置，那么在识别对胎盘具有特异性或相对特异性的末端位置后，可以将较高权重归于假定的胎儿重新突变是真实突变。由于血浆DNA片段具有两个末端，可以通过考虑其末端中的一个或两个是否与胎盘相关而进一步修改权重以将其识别为胎儿源性片段。

[0331] 如图16中所展示，具有恰好结束于536,772个HCC特异性结束位置处的末端核苷酸的血浆DNA片段将更可能源自肿瘤。相比之下，具有恰好结束于怀孕特异性结束位置或两种情况共用的位置处的末端核苷酸的血浆DNA片段将不大可能源自肿瘤，其中怀孕特异性结束位置潜在地不大可能且在任何使用权重的实施例中被给予较低权重。

[0332] 因此，对HCC案例具有特异性的顶部结束位置的清单可以用于选择癌症相关的突变，且对怀孕案例具有特异性或由两种情况共用的顶部结束位置的清单可以用于滤出假阳性突变。类似程序可以用于对于非侵入性产前测试识别胎儿突变和滤出假阳性突变。

[0333] 一般来说，为了识别此类生物学上相关的血浆DNA末端位置，来自具有不同疾病或流行病背景或生理概况的个体组的血浆DNA样品可以相比于无此类疾病或背景或概况的另一个体组的样品。在一个实施例中，这些样品中的每一个可以深入地测序以使得可以在每个样品内识别血浆DNA片段的共同末端位置。在另一实施例中，来自具有互补概况的个人组的序列数据可以合并在一起以识别代表疾病或生理概况的共同末端位置。

[0334] 样品中的每个血浆DNA片段可以个别地询问且基于末端位置分配似然评分。某一末端位置的似然评分可以依赖于结束于目标个体(例如癌症)的末端位置处的序列读取的量(例如通过跨越样品的测序深度标准化的序列读取的百分比或其它值)相对于结束于对照组的末端位置处的序列读取的量的分离。较大分离将导致较高特异性，且因此可以应用较高似然评分。因此，可以进行将具有特异性末端位置的血浆DNA片段分类为可能疾病相关

的或非疾病相关的、胎儿的或母体的等。

[0335] 或者，源自相同区域的血浆DNA片段可以共同地解释，即结束于特定核苷酸处的比率可以通过标准化为测序深度计算。以这种方式，某些核苷酸可以相对于基因组中的其它位置识别为共同末端位置，例如仅基于特定类型的一个样品的分析，尽管可以使用更多样品。因此，可以进行将具有特异性末端位置的血浆DNA片段分类为可能疾病相关的或非疾病相关的、胎儿的或母体的等。对于显示高频率的具有此类生物学上相关的血浆DNA末端位置的血浆DNA片段的位置，可以确定此类基因座富含生物学上相关的DNA且因此包括为高度可能地是癌症相关的或胎儿特异性的或与其它疾病或生物过程相关的一组血浆DNA片段。可能性水平可以基于给定核苷酸相对于其它核苷酸的比率高低程度，以与跨越不同组进行比较类似的方式，如上文所述。

[0336] 2.结果

[0337] 为了说明此方法的功效，直接从HCC患者的血浆DNA测序数据识别潜在的癌症相关突变。存在于至少两个血浆DNA片段的序列读取中的单核苷酸变化被视为潜在的癌症相关突变。也对肿瘤组织测序且存在于肿瘤组织中的突变被视为真实的癌症相关突变。

[0338] 在染色体8上，在不使用动态截止分析的情况下从HCC患者的血浆DNA测序数据识别总共20,065个潜在突变。如果序列变异体存在于至少两个经测序DNA片段中，那么序列变异体将被看作是潜在突变。从肿瘤组织的测序结果识别884个真实体细胞突变。20,065个假定突变包括884个真实突变中的802个(91％)。因此，仅4％的假定突变是得到4％的PPV的肿瘤组织中的真实体细胞突变。

[0339] 为了增强检测体细胞突变的精确性，借此产生癌症基因型，我们使用基于携有假定突变的序列读取的末端核苷酸位置的以下过滤算法。(1).对于任何假定突变，如果存在至少一个携有突变且结束于HCC特异性结束位置上的序列读取，那么突变将对于下游突变分析来说合格。(2).将去除携有假定突变但结束于任何怀孕特异性结束位置或两种情况共用的位置上的序列读取。仅当存在两个或更多个在基于此算法去除读取之后显示相同突变的序列读取时，突变才将对于下游突变分析来说合格。

[0340] 应用上述1和2过滤算法，获得表2中的结果。应用基于携有假定突变的DNA片段的末端核苷酸的位置或末端位置的不同过滤算法的效应。

[0341]

	无过滤	包括具有 HCC 特异性末端的突变（过滤 1）	去除具有共用或怀孕特异性末端的读取（过滤 2）	应用两种过滤算法
识别的假定突变的数目	20,065	1,526	2,823	484
检测的真实突变的百分比	91%	29%	88%	40%
PPV	4%	17%	28%	71%

[0342] 表2

[0343] 通过采用需要末端位置具有HCC特异性的三种算法中的任一种或滤出怀孕特异性或共用位置的算法，PPV显著提高。通过应用两种算法，PPV提高到71％。

[0344] 可以对于每个染色体，或实际上对于另一基因组区域，或实际上对于全基因组识别其它数目的HCC和怀孕相关的末端位置，例如(但不限于)五十万、2百万、3百万、4百万、5

百万、6百万、7百万、8百万、9百万或一千万。在各种实施例中，可以在一个或多个癌症患者群体中确定血浆DNA分子中最频繁可见的末端位置，每个群体具有一种癌症类型。另外，可以对于无癌症个体确定血浆DNA分子中最频繁的末端位置。在一个实施例中，患有癌症的此类患者和无癌症个体可以另外细分为具有不同临床参数，例如性别、吸烟状况、既往健康(例如肝炎状态、糖尿病、体重)等的组。

[0345] 作为使用此类过滤标准的一部分，统计分析可以用于识别具有作为末端核苷酸或末端位置的较高概率的位置以对于不同生理和病理病况循环DNA。统计分析的实例包括(但不限于)司徒登氏t测试(Student t-test)、卡方测试(Chi-square test)和基于二项分布或泊松分布(Poisson distribution)的测试。对于这些统计分析，可以使用不同p值截止值，例如(但不限于)0.05、0.01、0.005、0.001和0.0001。也可以对于多重比较调节p值截止值。

[0346] D.确定基因型的方法

[0347] 图38是根据本发明的实施例分析生物样品以确定第一组织类型的基因型的方法3800的流程图。生物样品包括来自多个组织类型(包括第一组织类型)的游离DNA分子的混合物。第一组织类型潜在地具有与多个组织类型的其它组织类型不同的基因型。可以确定多个基因组位置处的基因型。

[0348] 在步骤3810处，识别所述第一组织类型的游离DNA分子的末端以高于阈值的比率出现的第一基因组位置。步骤3810可以与步骤3610类似的方式进行。第X.B部分提供进行步骤3810的其它实例。

[0349] 在步骤3820处，分析来自个体的生物样品的第一多个游离DNA分子。分析游离DNA分子包括确定对应于游离DNA分子的至少一个末端的参考基因组中的基因组位置。步骤3620可以与分析游离DNA分子的其它步骤类似的方式进行。

[0350] 在步骤3830处，基于第一多个游离DNA分子的分析识别结束于第一基因组位置处的游离DNA分子的集合。举例来说，集合可以使用具有已知结束位置的检测探针的序列读取的比对识别。其它实例提供于本文中。

[0351] 在一些实施例中，可以进行另外的过滤，例如如上文所述。举例来说，游离DNA分子的尺寸可能需要小于规定量，例如由于胎儿组织和肿瘤组织一般短于来自健康细胞的DNA片段。在一个实施方案中，可以过滤游离DNA分子的集合以排除结束于第一基因组位置处的游离DNA分子中的至少一个或修改结束于第一基因组位置处的游离DNA分子中的至少一个的权重。可以使用游离DNA分子的经过滤集合确定基因型。

[0352] 在各种实施例中，过滤可以使用以下中的至少一个：游离DNA分子的尺寸、游离DNA分子在一个或多个位置处的甲基化状态(例如CpG位点经甲基化或未经甲基化)和游离DNA分子是否覆盖第一组织类型的游离DNA分子的末端以高于阈值的比率出现的一个或多个其它基因组位置。甲基化状态可以提供第一组织类型的签名，如上文所述。

[0353] 在步骤3840处，对于游离DNA分子集合中的每个游离DNA分子，确定出现于第一基因组位置处的对应碱基(核苷酸)。可以确定每个碱基的分子总数且可以对于每个碱基计算百分比。

[0354] 在步骤3850处，使用出现于游离DNA分子集合中的第一基因组位置处的对应碱基确定第一基因组位置处的第一组织类型的基因型。在各种实施方案中，高百分比的仅一个

碱基(例如高于80％、85％或90％)可以指示基因型对于碱基是纯合的,而两个具有类似百分比(例如在30-70％之间)的碱基可能使得确定基因型是杂合的。因此,每个碱基的百分比可以相比于基因型的截止值。在一些实施例中,可以基于第一组织类型对样品的比例贡献确定截止值。

[0355] 因此,在一些实施例中,确定第一基因组位置处的第一组织类型的基因型可以包括确定多个碱基中的每一个的百分比贡献且比较百分比贡献中的每一个与一个或多个截止值。在一个实例中,当第一碱基的百分比贡献高于第一截止值时,第一截止值可以对应于第一碱基的纯合基因型。在另一实例中,当第一碱基和第二碱基的百分比贡献高于第一截止值且低于第二截止值时,第一截止值和第二截止值可以对应于第一碱基和第二碱基的杂合基因型。

[0356] 在一些实施例中,可以对于步骤3830中识别的集合中的每个游离DNA分子进行加权。举例来说,如果游离DNA分子来自第一组织类型的可能性是80％,那么0.8可以是权重。可以对特定碱基的所有权重的总贡献求和以确定每个碱基的对应量。对应量可以用于确定每个碱基的百分比贡献,其中百分比可以用于确定基因型。

[0357] 因此,过滤可以对应于游离DNA分子来自第一组织类型的可能性将加权分配到游离DNA分子。可以对于多个碱基(例如仅检测的那些,其可以是2、3或4个)中的每一个确定加权和。如果仅检测一个碱基,那么可以对于所述的一个碱基确定纯合基因型。可以使用加权和确定多个碱基中的每一个的百分比贡献,其中基因型是使用百分比贡献确定。

[0358] X.其它细节

[0359] 上文所述的各种实施例对于特定组织识别优选的结束位置,其中优选的结束位置中的一些可以连续,借此形成优选的结束窗口。不同度量可以用于识别游离基因组窗口处的DNA分子的出现率(例如最小窗口的基因组位置)。关于此类操作的其它细节,以及关于确定参考基因组中的游离DNA分子的结束位置的细节提供于下文。此类特定技术可以用于上文所述的实施例。

[0360] A.确定结束位置

[0361] 当对游离DNA分子测序时,DNA片段的结束模式存在各种可能性。血浆DNA的末端一般存在四种配置:(A)具有两个平整末端的双链DNA分子;(B)具有一个平整末端和一个非平整末端的双链DNA分子(显示两种情境中的每一个,因为两个链中的任一个可以突出);(C)具有两个非平整末端的双链DNA分子,其具有突出末端的不同组合;和(D)单链DNA分子。

[0362] 对于具有非平整末端的配置,取决于DNA分子的5′或3′端突出而存在不同模式。对于(B),双链DNA分子具有一个平整末端和一个非平整末端。在实例B1中,5′端突出且在实例B2中,3′端突出。对于(C),当两个末端为非平整时,存在三种可能的模式。在(C1)中,5′端在两侧上突出。在(C2)中,3′端在两侧上突出。在(C3)中,5′端在一侧上突出且3′端在另一侧上突出。

[0363] 对于测序,双末端测序方案通常对每一个链的一个末端测序。因此将其视为双链DNA测序方案。当两个末端为非平整时,方案可以切断核苷酸或将核苷酸添加到末端以使其平整。克列诺片段为可以进行此类操作的酶。领域中的其它方案使用单链DNA测序方案。

[0364] 不管使用的特定技术(包括使用探针),只要结束位置可重复且显示相关性,如此处所展示,是否在测序中获得DNA片段的真实末端不影响结果,因为任何偏移是可重复的,

且因此抵消。另外，某些技术可以用于识别结束位置，按术语部分中所描述。

[0365] B.识别组织特异性结束位置

[0366] 如上文所述，在特定组织类型中，相比于其它区域，某些基因组区域关于游离DNA分子将结束于特定位置上的似然性具有较大变化。举例来说，肝组织可能具有作为脱氧核糖核酸酶超敏性位点的区域，但其它组织不具有作为脱氧核糖核酸酶超敏性位点的所述区域。因此，相对于其它位置，结束于此类区域内的某些位置上的游离DNA分子的数目将较高。举例来说，此类位置可以识别为关于已知对于特定组织具有高裂解量(因此，似然函数中的高幅度)的区域的游离DNA分子的比率的最大值，例如如第1II部分中所述。在其它实例中，可以识别基因组位置，其中左峰和右峰充分地分离，例如如第1V部分中所述。

[0367] 在其它实例中，具有和不具有病况(例如怀孕或癌症，可能为特定类型)的样品的高比率结束位置(例如高于阈值的比率)集合中的差异可以用于对与所述病况相关的特定组织类型识别优选的结束位点，例如如在第V、VI和VII部分中关于使用文氏图所描述。作为其它实例，具有病况的一种样品中相比于不具有病况的另一样品的显著较高比率可以提供特定组织类型的优选的结束位点。在各种实施例中，此类实例技术中的一些或全部可以一起使用。比率可以通过相对丰度的任何度量测量。

[0368] 在以上方法的一些实施例中，第一组织类型的游离DNA分子的末端以高于阈值的比率出现的基因组位置的第一集合可以如下方式识别。校准样品可以与测试样品类似的方式分析，其中已知相同类型的两个样品(例如血浆、血清、尿液等)和校准样品包括第一组织类型(例如来自怀孕女性的样品的胎儿组织或HCC患者的肝的肿瘤组织)。结束于基因组窗口中的多个游离DNA分子(例如关于宽度一或更大)可以相比于参考值以确定结束位置的比率是否高于所述位置的阈值。在一些实施例中，如果比率超过参考值，那么当对应数目超过参考值时，第一基因组窗口内的基因组位置中的每一个可以识别为具有高于阈值的比率。此类方法可以识别优选的结束窗口，其包括优选的结束位置。

[0369] 参考值可以使得仅前N个基因组窗口具有高于阈值的比率。举例来说，基因组位置的第一集合可以具有关于对应数目的最高N值。举例来说，N可以是至少10,000;50,000;100,000;500,000;1,000,000;或5,000,000。

[0370] 作为另一实例，参考值可以是根据样品中的游离DNA分子的概率分布和平均长度的结束于基因组窗口内的游离DNA分子的预期数目，例如如第VI.A.1部分中所述。p值可以使用对应数目和预期数目确定，其中阈值对应于截止p值(例如0.01)。p值小于截止p值指示比率高于阈值。作为另一另一实例，参考值可以包括结束于基因组窗口内的游离DNA分子的测量数目，所述基因组窗口来自识别为具有减少量的第一组织类型的样品，例如如关于图29A和29B所描述。

[0371] 满足比率阈值的基因组位置不一定添加至基因组位置的第一集合。可以添加其它过滤标准。此类过滤标准的实例在第VI.A.3和IX.C部分中指定。对于尺寸的过滤标准，游离DNA分子的尺寸(例如长度或质量)可以例如如美国专利公开案2011/0276277、2013/0040824和2013/0237431中所述地测量，所述专利公开案全部以全文引用的方式并入。第一统计值可以确定结束于经确定具有高于阈值的比率的第一基因组窗口内的游离DNA分子的尺寸分布(例如当窗口的宽度为一时，在基因组位置上)。当第一统计值不超过尺寸阈值，例如相比于所有游离DNA分子或较大范围内的那些，平均尺寸不足够小或不存在足够数目的

小DNA片段(例如低于指定尺寸)时,可以从基因组位置的第一集合排除第一基因组窗口的基因组位置。

[0372] 第一统计值可以相比于经确定不具有高于阈值的比率的游离DNA分子的尺寸分布的第二统计值。如果两个值类似(例如其将不对于胎儿或肿瘤组织预期),那么可以从优选的结束位置的集合排除第一基因组窗口。比较对应数目与参考值可以包括计算对应数目与覆盖一个样品的基因组窗口的任何部分,且任选地不结束于基因组窗口中的游离DNA分子的数目的第一比率(例如PETR),如第VII.A.2部分中所述。参考值可以包括结束于基因组窗口内的读取的测量数目与覆盖基因组窗口且不结束于另一样品的基因组窗口内的游离DNA分子的数目的参考比率。第一比率可能需要大于相乘因数(例如4)乘以参考比率。

[0373] 另一过滤标准可以是基因组位置的第一集合的每个基因组位置可能需要具有至少指定数目的结束于基因组位置上的游离DNA分子。使用这些技术中的任一种,基因组位置的第一集合可以包括600与10,000个之间的基因组位置。

[0374] 在取用集合之间的差异(例如使用文氏图)的实施例中,比率(例如如从基因组窗口确定)高于阈值的基因组位置包括第一超集,例如如在图28A中显示为集合P和集合S。第三多个游离DNA分子可以从具有减少量的第一组织类型(例如较少或无胎儿组织或HCC组织,如图28A中所描绘)的至少一个第二额外样品分析以识别第二超集,例如集合Q和集合S。取决于分析的组织类型,基因组位置的第一集合可以包括在第一超集中且不在第二超集,例如集合P或集合S中的基因组位置。

[0375] 如第VI部分中所述,第一组织类型可以具有第一组织特异性等位基因。计数可以由结束于基因组位置上且包括多个第一组织特异性等位基因中的至少一个的游离DNA分子组成。游离DNA分子的此计数(数目)可以相比于参考值。

[0376] C.相对丰度

[0377] 相对丰度值的各种实例提供于本文中,例如完整概率(P_I)、第VI.A.1部分中描述的p值和使用基因组窗口或基因组位置(当窗口具有宽度一时)确定的PETR值。对于基因组位置(宽度一的窗口)的PETR,可以对于基因组位置的第一集合的每个基因组位置计算结束于基因组位置上的第一多个游离DNA分子的对应数目。这可以作为确定结束于基因组位置的第一集合中的任一个上的第一多个游离DNA分子的第一数目(例如分子)的一部分进行。覆盖基因组位置且不结束于基因组位置上的游离DNA分子的第三数目(例如分母)可以作为确定游离DNA分子的第二数目的一部分而计算。可以确定对应数目与第三数目的第一比率,且第一比率的平均值用作相对丰度。

[0378] 对于w-PETR,可以对于基因组位置的第一集合的每个基因组位置计算结束于包括基因组位置的第一窗口(例如图31A中的窗口A)内的游离DNA分子的对应数目。可以计算结束于包括基因组位置的第二窗口(例如图31A中的窗口B)内的游离DNA分子的第三数目。对应数目与第三数目的第一比率的平均值可以用作相对丰度。

[0379] 相对丰度值的另一实例为结束于基因组窗口上的游离DNA分子的比例,例如测量为结束于优选的结束位置上的经测序DNA片段的比例。因此,基因组位置的第二集合可以包括对应于第一多个游离DNA分子中的至少一个的末端的所有基因组位置。

[0380] D.校准值

[0381] 在各种实施例中,校准值可以对应于确定自校准样品的校准数据点的校准值或自

其确定的任何校准值，例如接近校准数据点的校准函数。一个或多个校准样品可以包括或可以不包括用于确定优选的结束位点的任何额外样品。

[0382] 对于一个或多个校准样品中的每一个，可以例如使用组织特异性等位基因测量第一组织类型的对应比例贡献。对应相对丰度可以使用结束于对应于基因组位置的第一集合的多个窗口内的游离DNA分子的对应数目确定。测量的比例贡献和相对丰度可以提供校准数据点。一个或多个校准数据点可以是形成接近多个校准数据点的校准函数的多个校准数据点。使用校准值的其它细节可以见于美国专利公开案2013/0237431中。

[0383] E.比例贡献的分类

[0384] 在一些实施例中，特定组织的优选的结束位置也可以用于测量样品中的特定组织类型的绝对贡献，例如以每单位体积(例如每毫升)的基因组数目。举例来说，可以关于游离DNA样品的体积或重量测量所关注的组织的浓度。在一个实施方案中，定量PCR可以用于测量单位体积或单位重量的提取的游离DNA样品中的结束于一个或多个优选末端处的游离DNA分子的数目。可以对于校准样品进行类似测量，且因此比例贡献可以确定为比例贡献，因为贡献为每单位体积或单位重量的浓度。

[0385] 在各种实施例中，当第一组织类型对应于肿瘤组织时，分类可以选自由以下组成的群组：个体中的肿瘤组织的量、个体中的肿瘤的尺寸、个体中的肿瘤的阶段、个体中的肿瘤负荷和个体中的肿瘤转移的存在

[0386] XI.其它实施例

[0387] 实施例1包括分析生物样品以确定混合物中的第一组织类型的比例贡献的分类的方法，所述生物样品包括来自包括第一组织类型的多个组织类型的游离DNA分子的混合物，所述方法包含：识别第一组织类型的游离DNA分子的末端以高于阈值的比率出现的基因组位置的第一集合；通过计算机系统分析来自个体的生物样品的第一多个游离DNA分子，其中分析游离DNA分子包括：确定对应于游离DNA分子的至少一个末端的参考基因组中的基因组位置；基于第一多个游离DNA分子的分析，确定第一数目的第一多个游离DNA分子结束于多个窗口中的一个内，每个窗口包括基因组位置的第一集合中的至少一个；通过使用游离DNA分子的第二数目对第一多个游离DNA分子的第一数目进行标准化来计算结束于多个窗口中的一个内的第一多个游离DNA分子的相对丰度，其中第二数目的游离DNA分子包括结束于基因组位置的第二集合处的游离DNA分子，所述基因组位置的第二集合在包括基因组位置的第一集合的多个窗口之外；和通过比较相对丰度与确定自已知第一组织类型的比例贡献的一个或多个校准样品的一个或多个校准值来确定第一组织类型的比例贡献的分类。

[0388] 实施例2包括根据实施例1所述的方法，其中识别基因组位置的第一集合包括：通过计算机系统分析来自至少一个第一额外样品的第二多个游离DNA分子以识别第二多个游离DNA分子的结束位置，其中已知至少一个第一额外样品包括第一组织类型且与生物样品具有相同样品类型；对于多个基因组窗口中的每个基因组窗口：计算结束于基因组窗口上的第二多个游离DNA分子的对应数目；和比较对应数目与参考值以确定结束于基因组窗口内的一个或多个基因组位置上的游离DNA分子的比率是否高于阈值。

[0389] 实施例3包括实施例2的方法，其中多个基因组窗口中的第一基因组窗口具有大于一个基因组位置的宽度，且其中第一基因组窗口内的基因组位置中的每一个在对应数目超过参考值时识别为具有高于阈值的结束于基因组位置上的游离DNA分子的比率。实施例4包

括实施例2或3的方法，其中基因组位置的第一集合具有对应数目的最高N值，其中N为至少10,000。

[0390] 实施例5包括实施例2、3或4的方法，进一步包含：确定第二多个游离DNA分子中的每一个的尺寸，其中识别基因组位置的第一集合进一步包括：确定第二多个游离DNA分子中结束于第一基因组窗口内的游离DNA分子的尺寸分布的第一统计值，所述第一基因组窗口经确定具有高于阈值的比率；比较第一统计值与大小阈值；和当第一统计值不超过大小阈值时，从基因组位置的第一集合排除第一基因组窗口。实施例6包括实施例2-5中的任一个的方法，其中一个或多个校准样品包括至少一个第一额外样品。实施例7包括实施例1到6中的任一个的方法，进一步包含：对于一个或多个校准样品中的每一个：测量第一组织类型的对应比例贡献；和使用结束于与基因组位置的第一集合相对应的多个窗口内的第二多个游离DNA分子的对应数目确定对应相对丰度，借此获得校准数据点，其中每个校准数据点指定第一组织类型对额外生物样品的经测量比例贡献和对应相对丰度。实施例8包括实施例7的方法，其中一个或多个校准数据点为形成接近多个校准数据点的校准函数的多个校准数据点。

[0391] 实施例9包括实施例2到8中的任一个的方法，其中基因组位置的第一集合的每个基因组位置具有至少指定数目的结束于基因组位置上的第二多个游离DNA分子中的游离DNA分子。实施例10包括实施例2到9中的任一个的方法，其中参考值为根据至少一个第一额外样品中的游离DNA分子的概率分布和平均长度，结束于基因组窗口内的游离DNA分子的预期数目。实施例11包括实施例10的方法，其中概率分布为泊松分布(Poisson distribution)，且其中确定结束于基因组窗口内的一个或多个基因组位置上的游离DNA分子的比率是否高于阈值包括：使用对应数目和预期数目确定对应p值，其中阈值对应于截止p值，对应p值小于截止p值指示结束于基因组窗口内的游离DNA分子的比率高于阈值。

[0392] 实施例12包括实施例2到11中的任一个的方法，其中结束于基因组位置上的第二多个游离DNA分子的比率高于阈值的基因组位置包括第一超集，且其中识别基因组位置的第一集合进一步包括：通过计算机系统分析来自至少一个第二额外样品的第三多个游离DNA分子，以识别结束于基因组位置上的第三多个游离DNA分子中高于所述阈值的第二超集，所述第二额外样品被识别为具有减少量的所述第一组织类型；和将基因组位置的第一集合识别为包括在第一超集中且不在第二超集中的基因组位置。

[0393] 实施例13包括实施例2到12中的任一个的方法，其中参考值包括结束于基因组窗口内的游离DNA分子的测量数目，所述测量数目确定自至少一个识别为不具有第一组织类型的第二额外样品的第三多个游离DNA分子。实施例14包括实施例13的方法，进一步包含：确定第二多个游离DNA分子中的每一个的尺寸，其中识别基因组位置的第一集合进一步包括：确定结束于第一基因组位置上的第二多个游离DNA分子的游离DNA分子的第一尺寸分布的第一统计值，所述第一基因组位置经确定具有高于阈值的比率；确定结束于一个或多个第二基因组位置上的第三多个游离DNA分子的游离DNA分子的第二尺寸分布的第二统计值，所述第二基因组位置经确定具有高于阈值的比率；比较第一统计值与第二统计值；当第一统计值不超过第二统计值至少指定量时，从基因组位置的第一集合排除第一基因组位置，所述指定量指示第一尺寸分布小于第二尺寸分布。实施例15包括实施例13或14的方法，其中比较对应数目与参考值包括：计算对应数目与覆盖基因组窗口的第二多个游离DNA分子

的第三数目的第一比率;和比较第一比率与参考值,所述参考值包括结束于基因组窗口内的读取的测量数目与覆盖基因组窗口且不结束于基因组窗口内的第三多个游离DNA分子的第四数目的参考比率。实施例16包括实施例15的方法,其中第三数目的第二多个游离DNA分子不结束于基因组窗口内。实施例17包括实施例15或16的方法,其中确定结束于基因组窗口内的游离DNA分子的比率是否高于阈值包括:确定第一比率是否大于相乘因数乘以参考比率。

[0394] 实施例18包括实施例2到17中的任一个的方法,其中生物样品和至少一个第一额外样品的样品类型选自由以下组成的群组:血浆、血清、脑脊髓液和尿液。实施例19包括实施例2到18中的任一个的方法,其中基因组窗口为基因组位置,且其中第一组织类型具有多个第一组织特异性等位基因,且其中计算结束于基因组位置上的第二多个游离DNA分子的对应数目包括:识别结束于基因组位置上的游离DNA分子是否包括多个第一组织特异性等位基因中的至少一个;当游离DNA分子包括第一组织特异性等位基因时,在对应数目中包括游离DNA分子;且当游离DNA分子不包括第一组织特异性等位基因时,在对应数目中不包括游离DNA分子。

[0395] 实施例20包括实施例1到19中的任一个的方法,其中第一组织类型在至少一个额外样品中具有多个第一组织特异性等位基因,且其中基因组位置的第一集合是使用包括多个第一组织特异性等位基因中的至少一个的至少一个额外样品的游离DNA分子确定。实施例21包括实施例20的方法,其中基因组位置的第二集合使得第二组织类型的游离DNA分子的末端以高于阈值的比率出现于至少一个额外样品中,其中第二组织类型在至少一个额外样品中具有多个第二组织特异性等位基因,且其中基因组位置的第二集合是使用包括多个第二组织特异性等位基因中的至少一个的至少一个额外样品的游离DNA分子确定。实施例22包括实施例21的方法,其中至少一个额外样品来自怀孕女性,且其中第一组织类型为胎儿组织且第二组织类型为母体组织。实施例23包括实施例21或22的方法,其中将在第一组织类型与第二组织类型之间具有共用等位基因的游离DNA分子的末端以高于阈值的第二比率出现的基因组位置从基因组位置的第一集合排除且从基因组位置的第二集合排除。

[0396] 实施例24包括实施例1到23中的任一个的方法,其中相对丰度包括第一数目与第二数目的比率。实施例25包括实施例1到24中的任一个的方法,其中多个窗口具有一个基因组位置的宽度,且其中如下计算相对丰度:对于基因组位置的第一集合的每个基因组位置:计算结束于基因组位置上的第一多个游离DNA分子的对应数目,作为确定第一数目的第一多个游离DNA分子结束于基因组位置的第一集合中的任一个上的一部分;计算覆盖基因组位置且不结束于基因组位置上的第一多个游离DNA分子的第三数目,作为确定游离DNA分子的第二数目的一部分;计算对应数目与第三数目的第一比率;计算第一比率的平均值作为相对丰度。实施例26包括实施例1到24中的任一个的方法,其中如下计算相对丰度:对于基因组位置的第一集合的每个基因组位置,计算结束于包括基因组位置的第一窗口内的第一多个游离DNA分子的对应数目,作为确定第一数目的第一多个游离DNA分子结束于多个窗口中的一个内的一部分;计算结束于包括基因组位置的第二窗口内的第一多个游离DNA分子的第三数目,第二窗口大于第一窗口;计算对应数目与第三数目的第一比率;计算第一比率的平均值作为相对丰度。

[0397] 实施例27包括实施例1到26中的任一个的方法,其中基因组位置的第二集合和基

因组位置的第一集合不重叠。实施例28包括实施例1到27中的任一个的方法，其中基因组位置的第二集合包括对应于第一多个游离DNA分子中的至少一个的末端的所有基因组位置。实施例29包括实施例1到28中的任一个的方法，其中分析游离DNA分子中的一个或多个包括确定两个基因组位置对应于游离DNA分子的两个末端。实施例30包括实施例1到29中的任一个的方法，其中比例贡献的分类对应于高于指定百分比的范围。实施例31包括实施例1到30中的任一个的方法，其中第一组织类型为肿瘤。实施例32包括实施例31的方法，其中分类选自由以下组成的群组：个体中的肿瘤组织的量、个体中的肿瘤的尺寸、个体中的肿瘤的阶段、个体中的肿瘤负荷和个体中的肿瘤转移的存在。

[0398] 实施例33包括实施例1到32中的任一个的方法，其中一个或多个额外生物样品来自个体且与生物样品在不同时间获得。实施例34包括实施例1到33中的任一个的方法，进一步包含：从待分析的生物样品获得模板DNA分子；使用模板DNA分子制备可分析DNA分子的测序文库，制备可分析DNA分子的测序文库不包括DNA扩增模板DNA分子的步骤；对可分析DNA分子的测序文库测序以获得对应于第一多个游离DNA分子的多个序列读取，其中分析第一多个游离DNA分子包括：在计算机系统处接收多个序列读取；通过计算机系统比对多个序列读取与参考基因组以确定多个序列读取的基因组位置。实施例35包括实施例1到34中的任一个的方法，进一步包含基于分类提供治疗性干预或基于分类对个体进行成像。实施例36包括实施例1到35中的任一个的方法，其中基因组位置的第一集合包括600与10,000个之间的基因组位置。

[0399] 实施例37包括分析生物样品以确定混合物中的第一组织类型的比例贡献的分类的方法，所述生物样品包括来自包括第一组织类型的多个组织类型的游离DNA分子的混合物，所述方法包含：识别至少一个具有对第一组织类型具有特异性的片段化模式的基因组区域；分析来自生物样品的多个游离DNA分子，其中分析游离DNA分子包括：确定对应于游离DNA分子的至少一个末端的参考基因组中的基因组位置；识别第一基因组位置的第一集合，每个第一基因组位置具有对应于第一基因组位置的游离DNA分子的末端的局部最小值；识别第二基因组位置的第二集合，每个第二基因组位置具有对应于第二基因组位置的游离DNA分子的末端的局部最大值；确定结束于至少一个基因组区域中的任一个中的第一基因组位置中的任一个上的游离DNA分子的第一数目；确定结束于至少一个基因组区域中的任一个中的第二基因组位置中的任一个上的游离DNA分子的第二数目；使用第一数目和第二数目计算分离值；和通过比较分离值与确定自已知第一组织类型的比例贡献的一个或多个校准样品的一个或多个校准值，确定第一组织类型的比例贡献的分类。

[0400] 实施例38包括实施例37的方法，其中第一基因组位置的第一集合包括多个基因组位置，其中第二基因组位置的第二集合包括多个基因组位置，其中确定游离DNA分子的第一数目包括确定结束于每个第一基因组位置上的游离DNA分子的第一量，借此确定多个第一量，其中确定游离DNA分子的第二数目包括确定结束于每个第二基因组位置上的游离DNA分子的第二量，借此确定多个第二量，且其中计算分离值包括：确定多个分离比，多个第一量中的一个与多个第二量中的一个的每个分离比和使用多个分离比确定分离值。实施例39包括实施例37或38的方法，其中至少一个基因组区域包括一个或多个脱氧核糖核酸酶超敏性位点。实施例40包括实施例37到38的方法，其中具有对第一组织类型具有特异性的片段化模式的至少一个基因组区域中的每一个包括至少一个额外样品中的一个或多个第一组织

特异性等位基因。实施例41包括实施例37或38的方法，其中至少一个基因组区域包括一个或多个ATAC-seq或微球菌核酸酶位点。实施例42包括实施例37到41中的任一个的方法，其中与基因组位置的第一集合的一个基因组位置比对的游离DNA分子向一个基因组位置的两侧延伸指定数目的核苷酸。实施例43包括实施例42的方法，其中指定数目为10到80个核苷酸。实施例44包括实施例37到43中的任一个的方法，其中识别第一基因组位置的第一集合包括：对于多个基因组位置中的每一个：确定位于基因组位置且向基因组位置的两侧延伸指定数目的核苷酸的游离DNA分子的第一量；确定位于基因组位置的游离DNA分子的第二量；和确定第一量与第二量的比率；以及识别比率中的多个局部最小值和多个局部最大值。实施例45包括实施例37到44中的任一个的方法，其中混合物为血浆或血清。实施例46包括实施例37到45中的任一个的方法，其中多个游离DNA分子为至少1,000个游离DNA分子。实施例47包括实施例37到46中的任一个的方法，其中对于多个基因组位置的给定基因组位置，第二量对应于与给定基因组位置比对的游离DNA分子的总数。

[0401] 实施例48包括分析生物样品以确定第一组织类型的基因型的方法，所述生物样品包括来自包括第一组织类型的多个组织类型的游离DNA分子的混合物，第一组织类型潜在地具有与多个组织类型的其它组织类型不同的基因型，所述方法包含：识别第一组织类型的游离DNA分子的末端以高于阈值的比率出现的第一基因组位置；通过计算机系统分析来自个体的生物样品的第一多个游离DNA分子，其中分析游离DNA分子包括：确定对应于游离DNA分子的至少一个末端的参考基因组中的基因组位置；基于第一多个游离DNA分子的分析，识别结束于第一基因组位置处的游离DNA分子集合；对于所述游离DNA分子集合中的每一个：确定第一基因组位置处出现的对应碱基，借此确定第一基因组位置处的多个对应碱基；使用游离DNA分子的集合中的第一基因组位置处出现的对应碱基确定第一基因组位置处的第一组织类型的基因型。实施例49包括实施例48的方法，进一步包含：过滤游离DNA分子的集合以排除结束于第一基因组位置的游离DNA分子中的至少一个或修改结束于第一基因组位置处的游离DNA分子中的至少一个的权重，其中基因型是使用游离DNA分子的经过滤集合确定。实施例50包括实施例49的方法，其中过滤使用以下中的至少一个：游离DNA分子的尺寸、游离DNA分子在一个或多个位置处的甲基化状态和游离DNA分子是否覆盖第一组织类型的游离DNA分子的末端以高于阈值的比率出现的一个或多个其它基因组位置。实施例51包括实施例49或50的方法，其中过滤对应于游离DNA分子来自第一组织类型的似然性而将加权分配到游离DNA分子，所述方法进一步包含：确定多个碱基中的每一个的加权和；以及使用加权和确定多个碱基中的每一个的百分比贡献，其中基因型是使用百分比贡献确定。实施例52包括实施例48到51中的任一个的方法，其中确定第一基因组位置处的第一组织类型的基因型包括：确定多个碱基中的每一个的百分比贡献；和将百分比贡献中的每一个与一个或多个截止值比较。实施例53包括实施例52的方法，其中当第一碱基的百分比贡献高于一个或多个截止值中的第一截止值时，第一截止值对应于第一碱基的纯合基因型。实施例54包括实施例52的方法，其中当第一碱基和第二碱基的百分比贡献高于一个或多个截止值中的第一截止值且低于一个或多个截止值中的第二截止值时，第一截止值和第二截止值对应于第一碱基和第二碱基的杂合基因型。实施例55包括实施例48到54中的任一个的方法，其中第一组织类型对应于肿瘤。实施例56包括实施例48到55中的任一个的方法，其中第一组织类型对应于胎儿，且其中个体怀有胎儿。

[0402] 实施例57包括分析生物样品的方法，所述生物样品包括来自包括第一组织类型的多个组织类型的游离DNA分子的混合物，所述方法包含：通过计算机系统分析来自个体的生物样品的多个游离DNA分子，多个游离DNA分子中的每一个具有左末端和右末端，其中分析游离DNA分子包括：确定对应于游离DNA分子的左末端的参考基因组中的左侧结束位置；确定对应于游离DNA分子的右末端的参考基因组中的右侧结束位置；识别左侧基因组位置的左侧集合，其各自具有多个游离DNA分子的左末端的局部最大值，所述局部最大值对应于基因组位置的左侧集合中的一个；识别右侧基因组位置的右侧集合，其各自具有多个游离DNA分子的右末端的局部最大值，所述局部最大值对应于基因组位置的右侧集合中的一个；如下识别对第一组织类型具有特异性的基因组位置的第一集合：比较左侧集合的左侧基因组位置与右侧集合的右侧基因组位置以识别基因组位置的第一集合，其中左侧基因组位置与最接近的右侧基因组位置的距离大于第一阈值距离，第一阈值距离为参考基因组中的至少5个基因组位置。实施例58包括实施例57的方法，进一步包含：如下识别基因组位置的第二集合：比较左侧集合的左侧基因组位置与右侧集合的右侧基因组位置以识别基因组位置的第二集合，其中左侧基因组位置与最接近的右侧基因组位置的距离小于第二阈值距离；使用结束于左侧基因组位置的左侧集合中的一个处的多个游离DNA分子的第一数目和结束于右侧基因组位置的右侧集合中的一个处的多个游离DNA分子的第二数目确定分离值；和通过比较分离值与一个或多个校准值确定第一组织类型的比例贡献的分类，所述校准值确定自已知第一组织类型的比例贡献的一个或多个校准样品。实施例59包括实施例58的方法，其中确定分离值包括：识别基因组位置的第一集合和基因组位置的第二集合的配对；对于配对中的每一个：确定结束于配对的第一基因组位置处的游离DNA分子的第一量；且确定结束于配对的第二基因组位置处的游离DNA分子的第二量，其中游离DNA分子的第一量对应于多个游离DNA分子的第一数目且游离DNA分子的第二量对应于多个游离DNA分子的第二数目。实施例60包括实施例59的方法，其中确定分离值包括：对于配对中的每一个：确定包括第一量和第二量的比率；和从所述比率确定分离值。实施例61包括实施例59或60的方法，其中基因组位置的第一集合和基因组位置的第二集合的配对彼此最接近。实施例62包括实施例57到61中的任一个的方法，其中第二阈值距离小于参考基因组中的5个基因组位置。实施例63包括实施例57到62中的任一个的方法，其中基因组位置的第一集合包括左侧基因组位置和右侧基因组位置两者。

[0403] 实施例64包括用于确定DNA混合物中的第一组织的比例贡献的方法，所述方法包含：识别对第一组织具有特异性的脱氧核糖核酸酶超敏性位点；分析来自生物样品的多个游离DNA分子，其中分析游离DNA分子包括：识别参考人类基因组中的游离DNA分子的位置，所述位置包括游离DNA分子的两个末端；识别第一基因组位置的第一集合，其各自具有与基因组位置比对的游离DNA分子的局部最小值且向基因组位置的两侧延伸指定数目的核苷酸；识别第二基因组位置的第二集合，其各自具有与基因组位置比对的游离DNA分子的局部最大值且向基因组位置的两侧延伸指定数目的核苷酸；计算结束于脱氧核糖核酸酶超敏性位点中的一个中的第一基因组位置中的一个上的游离DNA分子的第一数目；计算结束于脱氧核糖核酸酶超敏性位点中的一个中的第一基因组位置中的一个上的游离DNA分子的第二数目；确定第一数目和第二数目的比例；和基于所述比例确定第一组织的比例贡献。实施例65包括实施例64的方法，其中识别第一基因组位置的第一集合包括：对于多个基因组位置

中的每一个:确定位于基因座且向基因座的两侧延伸指定数目的核苷酸的游离DNA分子的第一量;确定位于基因座的游离DNA分子的第二量;和确定第一量和第二量的第一比率;以及识别所述比率中的多个局部最小值。实施例66包括实施例64或65中的任一个的方法,其中DNA混合物为血浆或血清。实施例66包括实施例64到66中的任一个的方法,其中多个游离DNA分子为至少1,000个游离DNA分子。

[0404] 实施例67包括用于确定DNA混合物中的第一组织的比例贡献的方法,所述方法包含:对于第一组织的DNA片段的一个末端,识别DNA片段具有高于阈值的频率的基因组位置;和分析来自生物样品的多个游离DNA分子,其中分析游离DNA分子包括:识别参考人类基因组中的游离DNA分子的位置,所述位置包括游离DNA分子的两个末端;计算结束于脱氧核糖核酸酶超敏性位点中的一个中的经识别基因组位置中的一个上的游离DNA分子的第一数目;从第一数目和经测序DNA的量计算比例;和基于所述比例确定第一组织的比例贡献。实施例68包括根据权利要求67所述的方法,其中第一组织为肿瘤。实施例69包括根据权利要求67所述的方法,其中第一组织为胎儿组织。

[0405] 实施例70包括预测携有假定突变的DNA片段是否实际上源自肿瘤的方法,所述方法包含:对于DNA片段的一个末端,识别DNA片段具有高于阈值的频率的基因组位置;和基于结束于经识别基因组位置中的一个处的DNA片段确定概率。

[0406] 实施例71包括包含计算机可读介质的计算机产品,所述计算机可读介质存储多个用于控制计算机系统以执行实施例1到70中的任一个的操作的指令。实施例72包括一种系统,其包含:实施例71的计算机产品;和一个或多个用于执行存储于计算机可读介质上的指令的处理器。实施例73包括一种系统,其包含用于执行实施例1到70中的任一个的装置。实施例74包括被配置成执行实施例1到70中的任一个的系统。实施例75包括一种系统,其包含分别执行实施例1到70中的任一个的步骤的模块。

[0407] XII.计算机系统

[0408] 本文中提及的任何计算机系统都可以利用任何适合数目的子系统。此类子系统的实例在图39中以计算机设备10示出。在一些实施例中,计算机系统包括单个计算机设备,其中子系统可以是计算机设备的组件。在其它实施例中,计算机系统可以包括多个具有内部组件、各自是子系统的计算机设备。计算机系统可以包括台式电脑和笔记本电脑、平板电脑、移动电话和其它移动装置。

[0409] 图39中示出的子系统经由系统总线75互连。显示额外子系统,如打印机74、键盘78、存储装置79、耦接到显示适配器82的监视器76等。与I/O控制器71耦接的外围设备和输入/输出(I/O)装置可以通过所属领域中已知的任何数目的接头(如输入/输出(I/O)端口77(例如USB、FireWire®))连接到计算机系统。举例来说,I/O端口77或外部接口81(例如以太网、Wi-Fi等)可以用于将计算机系统10连接到如因特网的广域网、鼠标输入装置或扫描仪。经由系统总线75的互连允许中央处理器73与每个子系统连通且控制来自系统存储器72或存储装置79(例如固定磁盘,如硬盘驱动器或光盘)的多个指令的执行,以及子系统之间的信息交换。系统存储器72和/或一个或多个存储装置79可以体现为计算机可读介质。另一子系统为数据采集装置85,如相机、麦克风、加速计等。在此提及的任何数据可以从一个组件输出到另一个组件且可以输出到用户。

[0410] 计算机系统可以包括例如通过外部接口81或通过内部接口连接在一起的多个相

同组件或子系统。在一些实施例中，计算机系统、子系统或设备可以经由网络通信。在所述情况下，一个计算机可以视为客户端并且另一个计算机视为服务器，其中每一个可以是同一计算机系统的一部分。客户端和服务器可以各自包括多个系统、子系统或组件。

[0411] 实施例的方面可以如下形式实施：使用硬件(例如专用集成电路或现场可编程门阵列)和/或使用具有以模块化或一体化方式的一般可编程处理器的计算机软件的逻辑控制。如本文所用，处理器包括单核处理器、在同一集成芯片上的多核处理器，或在单个电路板上或网络化的多个处理单元。基于在此提供的揭露内容和教示内容，所属领域的普通技术人员将知道并且了解使用硬件以及硬件与软件的组合来实施本发明的实施例的其它方式和/或方法。

[0412] 本申请中所述的任一种软件组件或函数可以作为软件代码实施，所述软件代码可通过使用任何适合计算机语言(如Java、C、C++、C#、面向对象的C语言、Swift，或脚本语言，如使用例如常规或面向对象技术的Perl或Python)的处理器执行。软件代码可以存储为用于存储和/或传输的计算机可读介质上的一系列指令或命令。适合的非暂时性计算机可读介质可以包括随机存取存储器(RAM)、只读存储器(ROM)、如硬盘驱动器或软盘的磁性媒体或如光盘(CD)或DVD(数字通用光盘)的光学媒体、闪存等。计算机可读介质可以是此类存储或传输装置的任何组合。

[0413] 所述程序还可以使用适合于经由符合多种方案的有线、光学和/或无线网络(包括因特网)传输的载波信号来编码和传输。因此，计算机可读介质可以使用以此类程序编码的数据信号产生。以程序代码编码的计算机可读介质可以与兼容装置一起封装或与其它装置分开提供(例如经由因特网下载)。任何此类计算机可读介质可以存在于单个计算机产品(例如硬盘驱动器、CD或整个计算机系统)上或内部，并且可以存在于系统或网络内的不同计算机产品上或内部。计算机系统可以包括监视器、打印机，或其它适合显示器以便将在此提及的任何结果提供给用户。

[0414] 在此所述的任何方法可以完全或部分地用计算机系统执行，所述计算机系统包括一个或多个可经配置以执行所述步骤的处理器。因此，实施例可以涉及经配置以执行在此所述的任何方法步骤的计算机系统，其潜在地用不同组件执行相应步骤或相应的步骤群。尽管本文中方法的步骤以经编号步骤的形式呈现，但其可以同时或以不同顺序执行。另外，这些步骤的部分可以与其它方法的其它步骤的部分一起使用。此外，步骤的全部或部分可以是任选的。另外，任何方法的任何步骤都可以用执行这些步骤的模块、单元、电路或其它装置来执行。

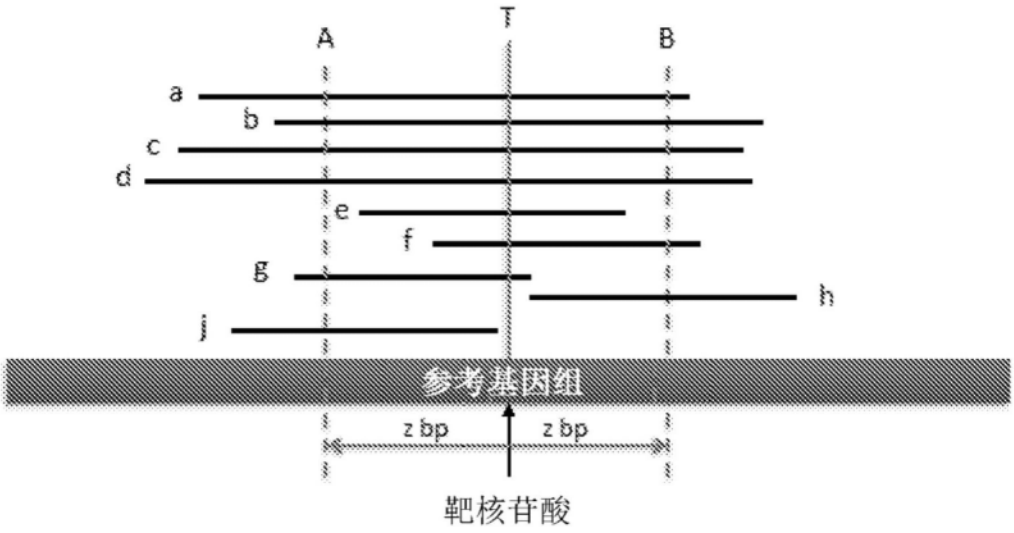
[0415] 特定实施例的具体细节可以按任何适合的方式组合而不脱离本发明实施例的精神和范围。然而，本发明的其它实施例可以涉及与每个个别方面或这些个别方面的特定组合相关的特定实施例。

[0416] 本发明的实例实施例的以上描述已经为了说明和描述的目的而呈现。其并不打算是穷尽性的或将本发明限制于所描述的精确形式，并且鉴于以上传授许多修改和变化是可能的。

[0417] 除非具体地相反指示，否则“一(a/an)”或“所述(the)”的叙述打算意指“一个或多个”。除非具体相反地指示，否则“或”的使用旨在意指“兼或”，而非“异或”。提及“第一”组件未必要求提供第二组件。此外，除非明确陈述，否则提及“第一”或“第二”组件不将提及的组

件限制于特定位置。

[0418] 本文提及的所有专利、专利申请、公开案和描述都出于所有目的以全文引用的方式并入。不承认任一个是现有技术。

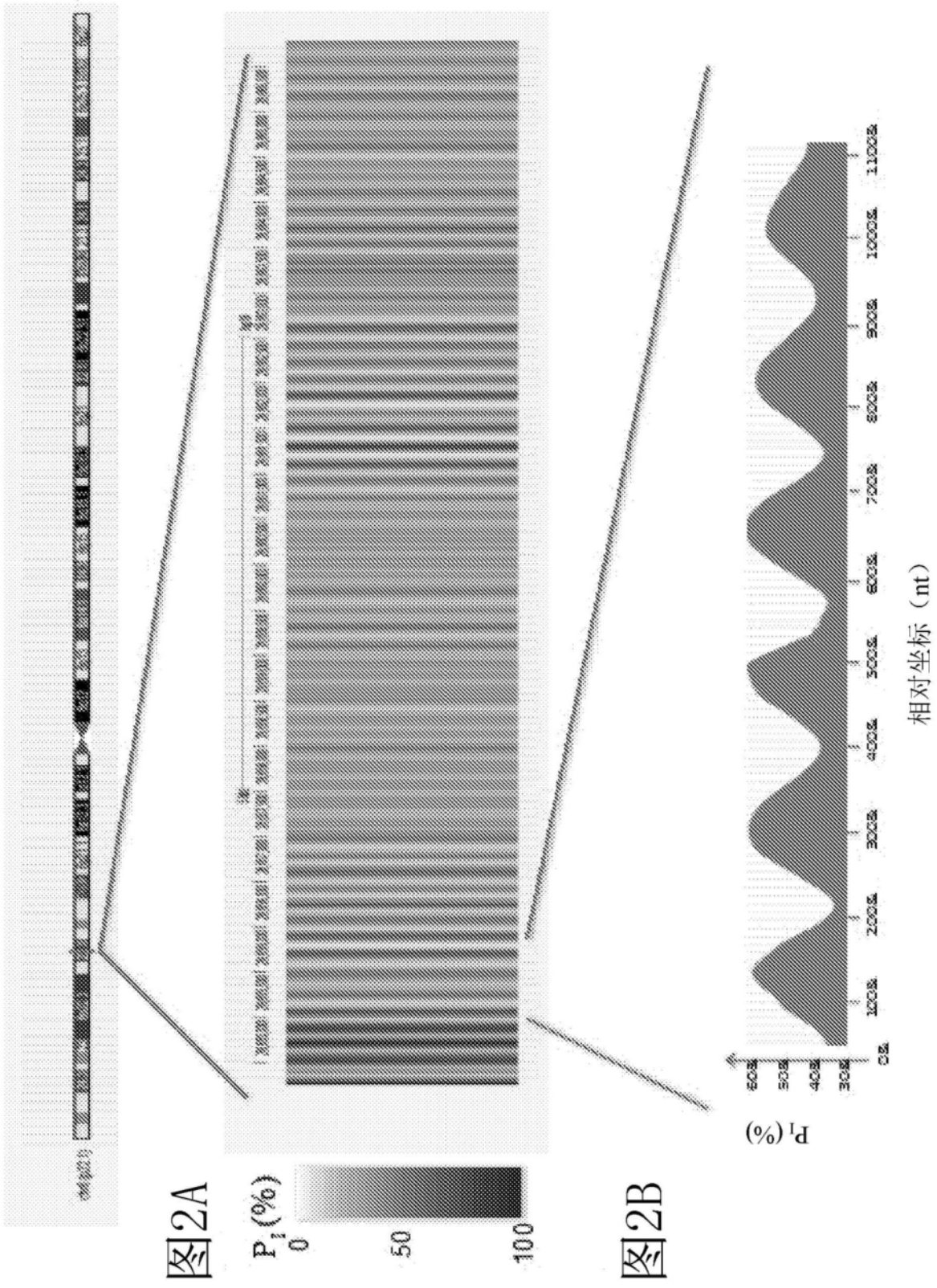


覆盖靶核苷酸上游的z个nt到下游的z个nt的片段：a、b、c和d

覆盖靶核苷酸的片段：a、b、c、d、e、f和g

$$p_i = \frac{4}{7}$$

图1


P_I (%)
0
50
100
图2A
P_I (%)
50
30
40
50
60
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
相对坐标（nt）
图2B

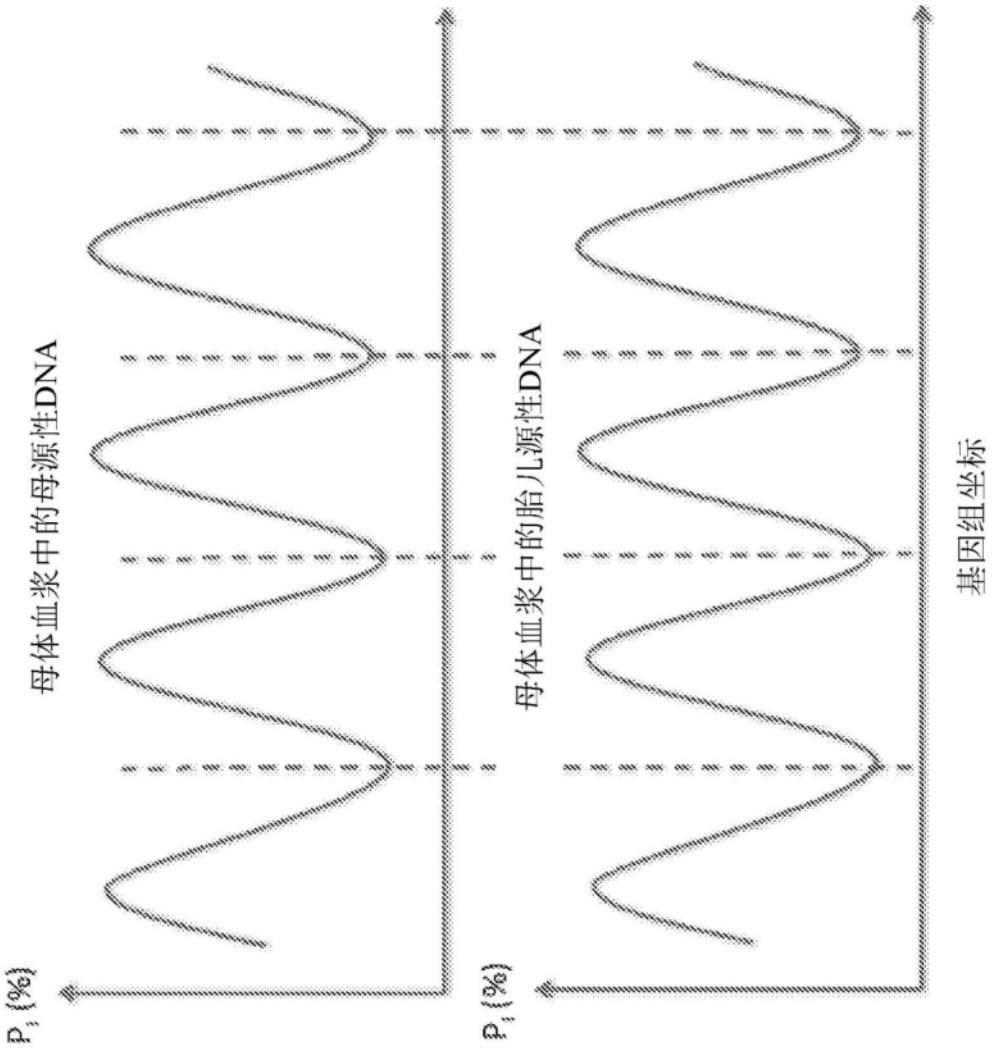


图3

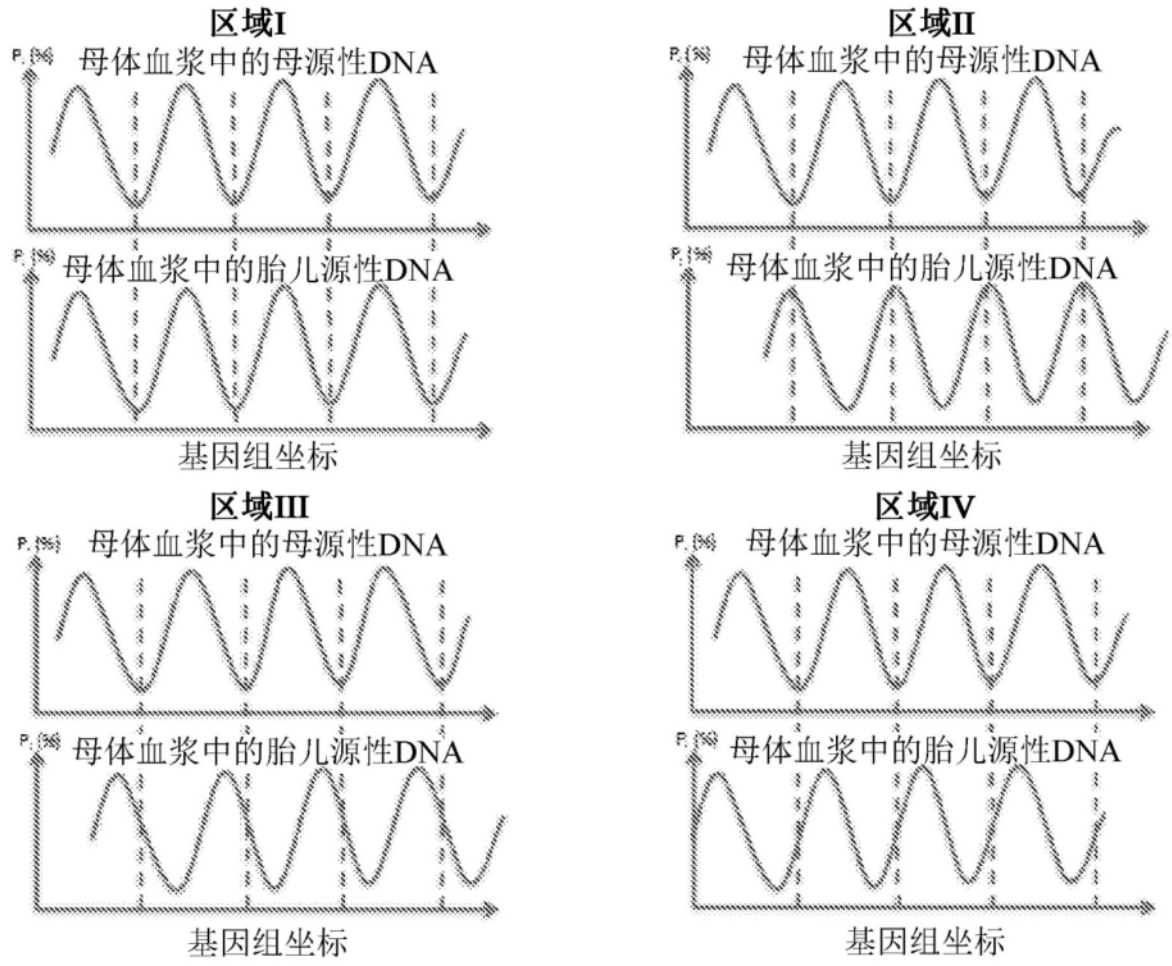

区域I
母体血浆中的母源性DNA
母体血浆中的胎儿源性DNA
基因组坐标
区域II
母体血浆中的母源性DNA
母体血浆中的胎儿源性DNA
基因组坐标
区域III
母体血浆中的母源性DNA
母体血浆中的胎儿源性DNA
基因组坐标
区域IV
母体血浆中的母源性DNA
母体血浆中的胎儿源性DNA
基因组坐标

图4

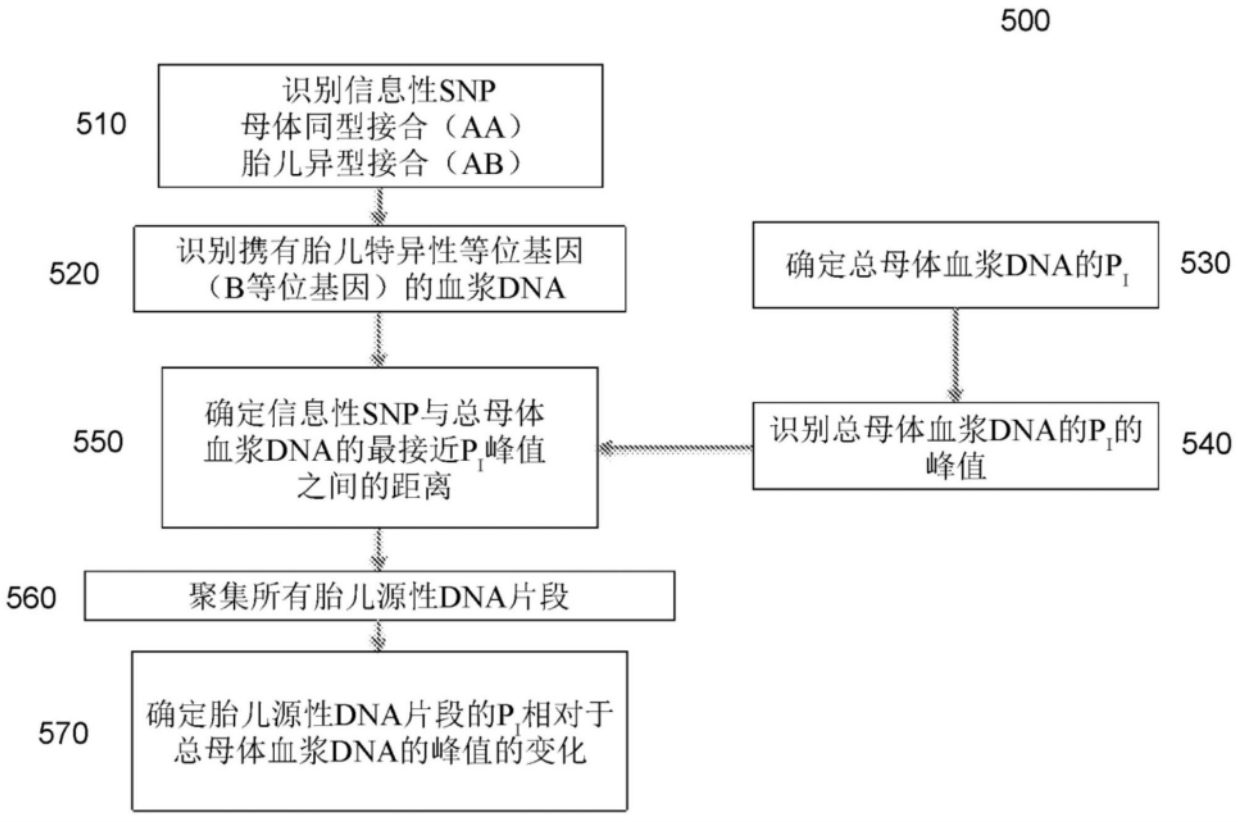

500
510
识别信息性SNP
母体同型接合（AA）
胎儿异型接合（AB）
520
识别携有胎儿特异性等位基因
（B等位基因）的血浆DNA
530
确定总母体血浆DNA的P_I
540
识别总母体血浆DNA的P_I的
峰值
550
确定信息性SNP与总母体
血浆DNA的最接近P_I峰值
之间的距离
560
聚集所有胎儿源性DNA片段
570
确定胎儿源性DNA片段的P_I相对于
总母体血浆DNA的峰值的变化

图5

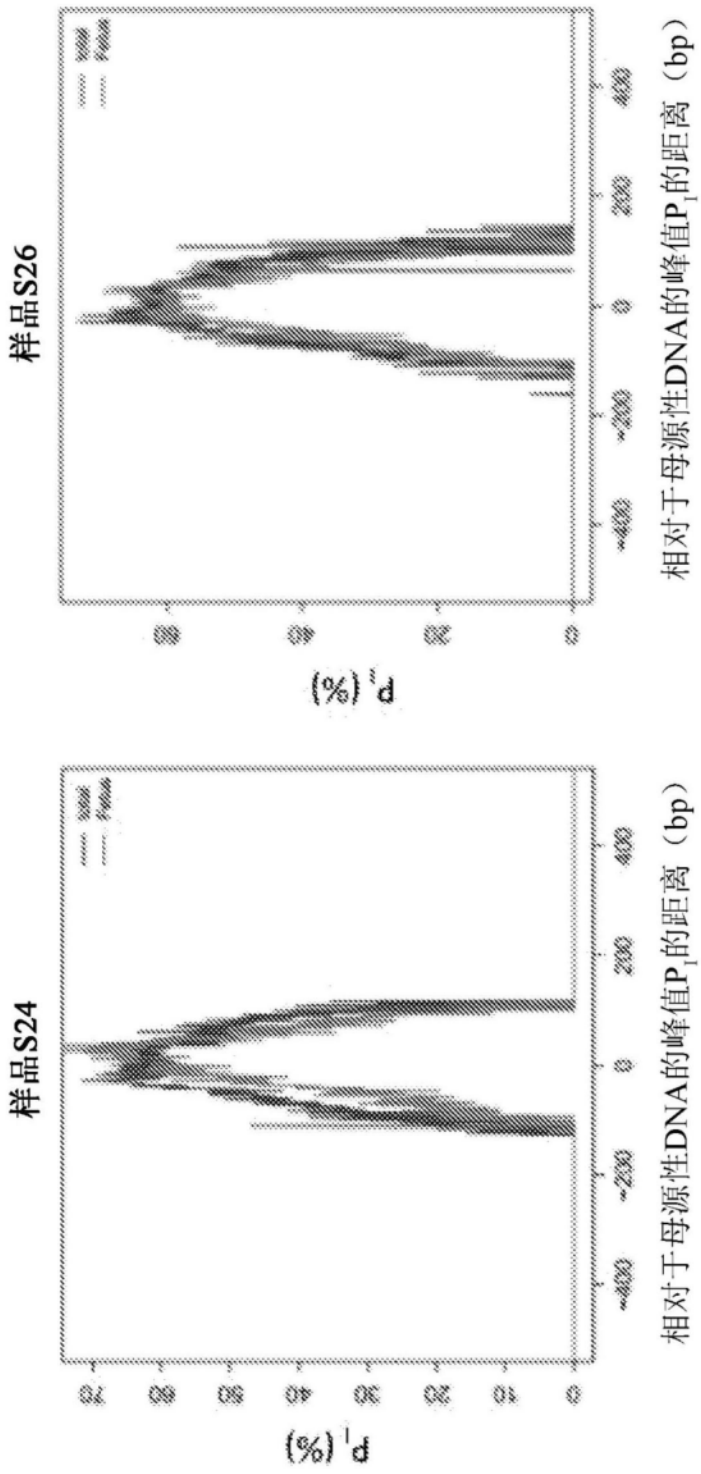


图6

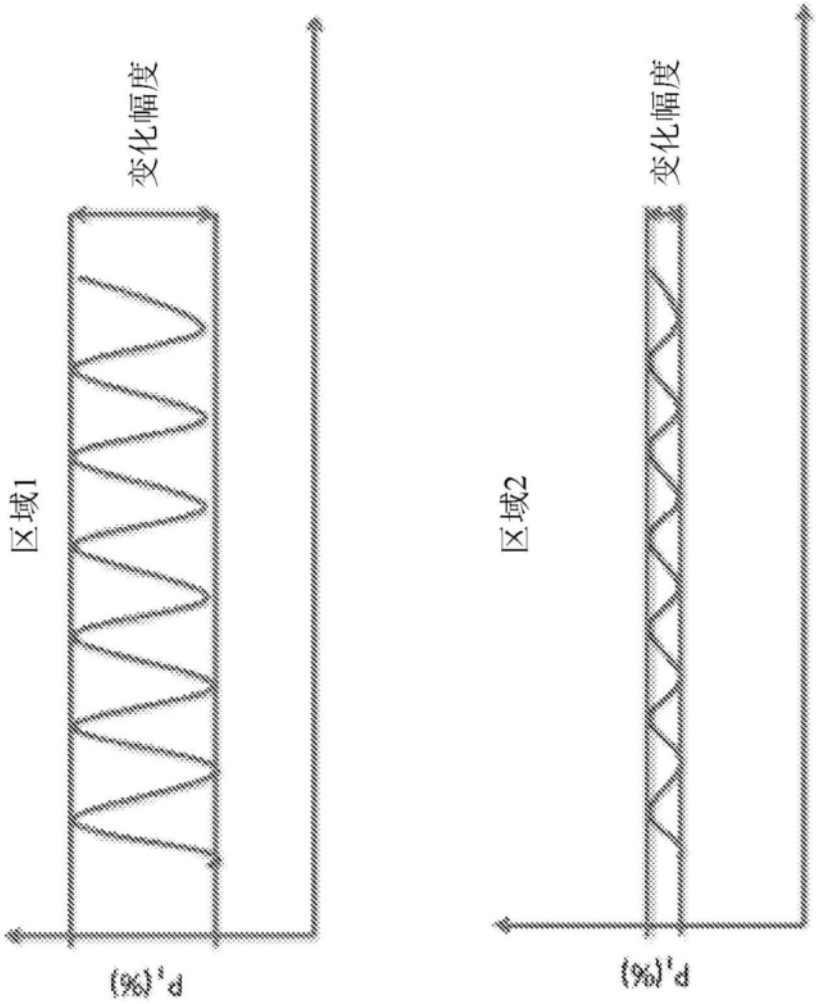


图7

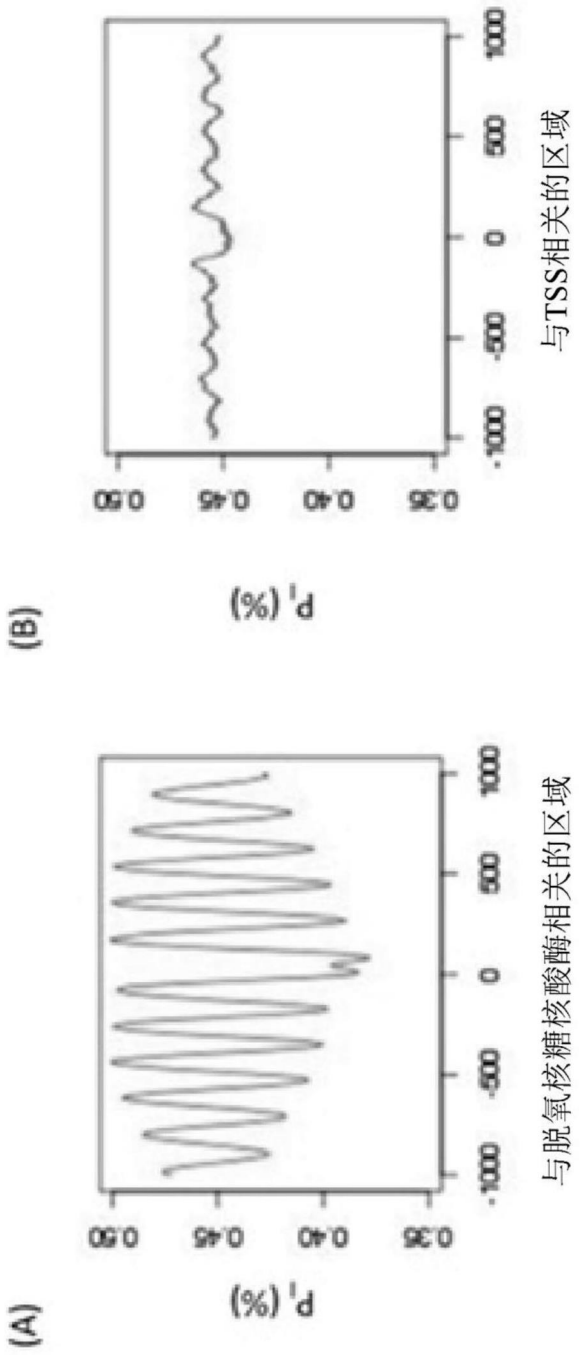

(B)
(A)
P_I (%)
0.35
0.40
0.45
0.50
-1000
-500
0
500
1000
与TSS相关的区域
与脱氧核糖核酸酶相关的区域

图8

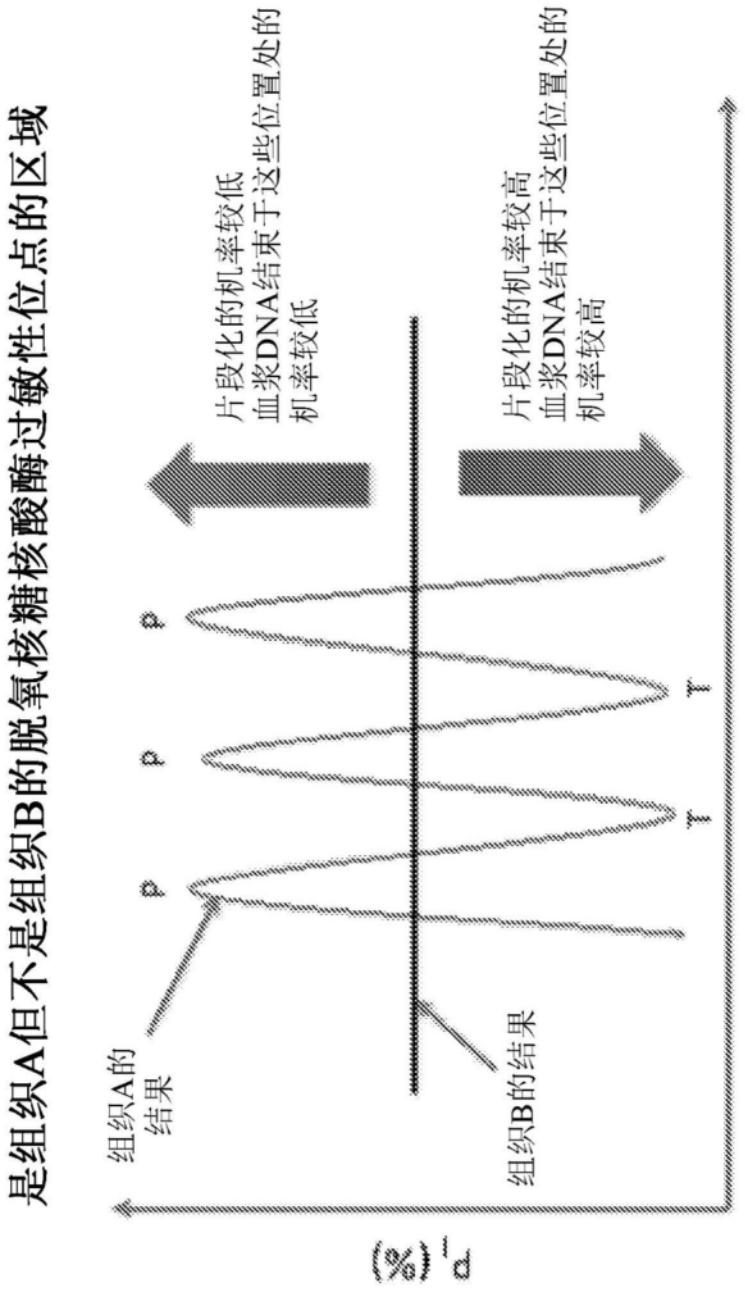

是组织A但不是组织B的脱氧核糖核酸酶过敏性位点的区域
片段化的机率较低
血浆DNA结束于这些位置处的
机率较低
片段化的机率较高
血浆DNA结束于这些位置处的
机率较高
组织A的
结果
组织B的结果
P
P
P
T
T
(%)

图9

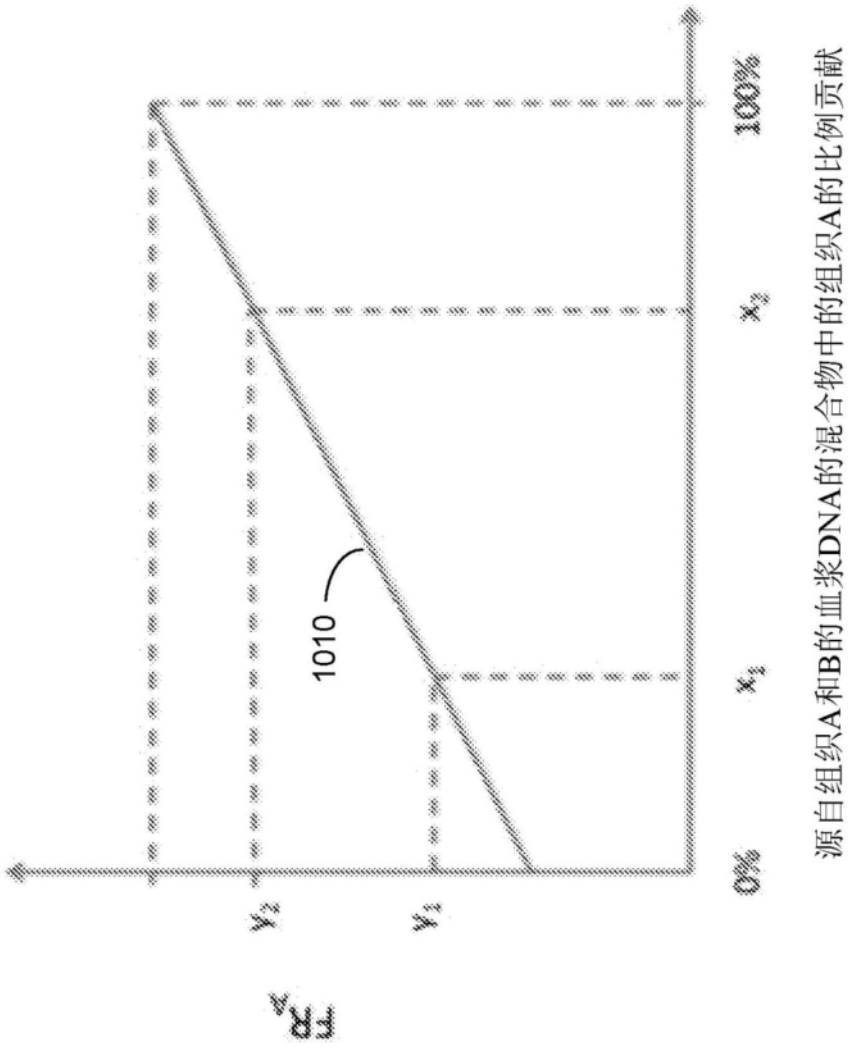


图10

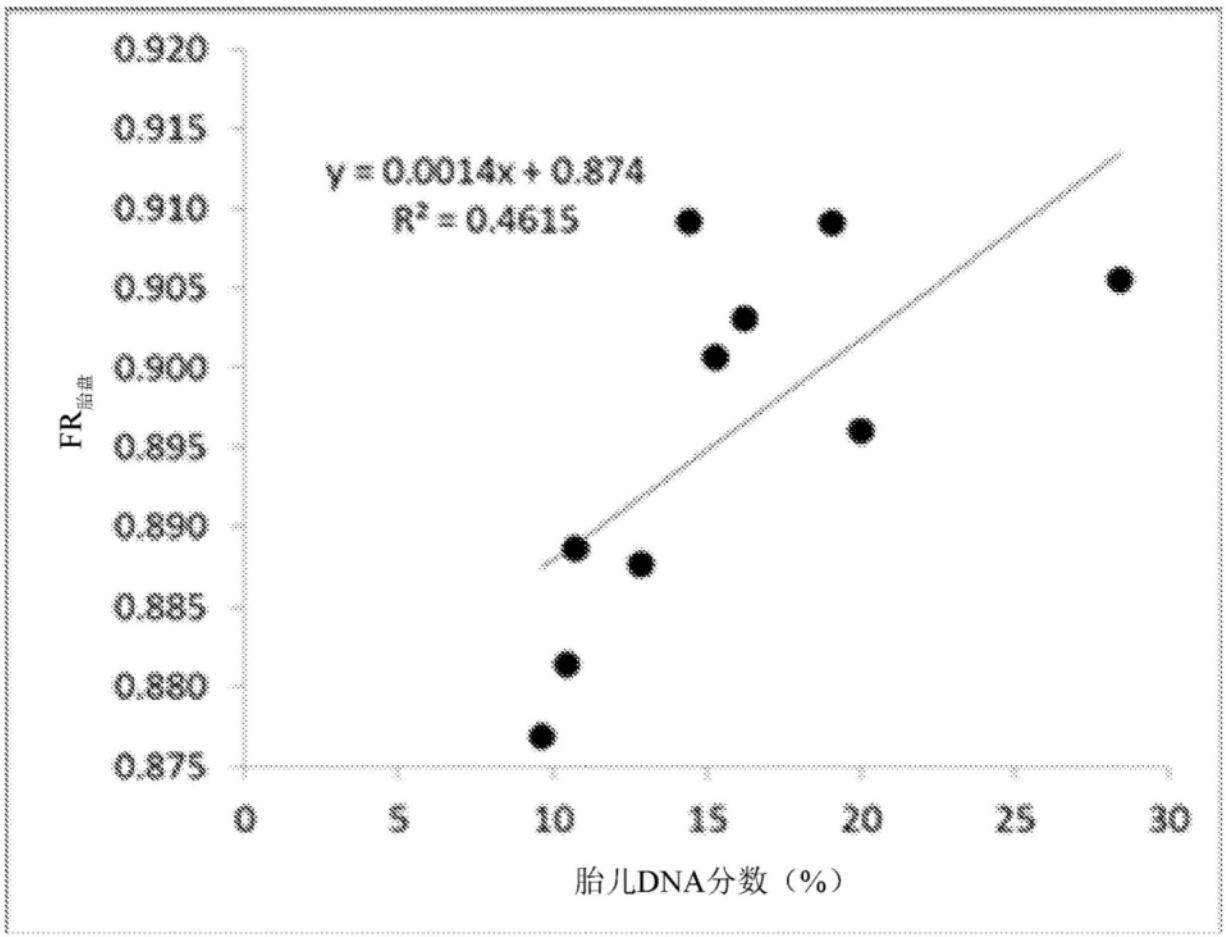

y = 0.0014x + 0.874
R² = 0.4615
FR胎盘
胎儿DNA分数（%）
0.920
0.915
0.910
0.905
0.900
0.895
0.890
0.885
0.880
0.875
0
5
10
15
20
25
30

图11

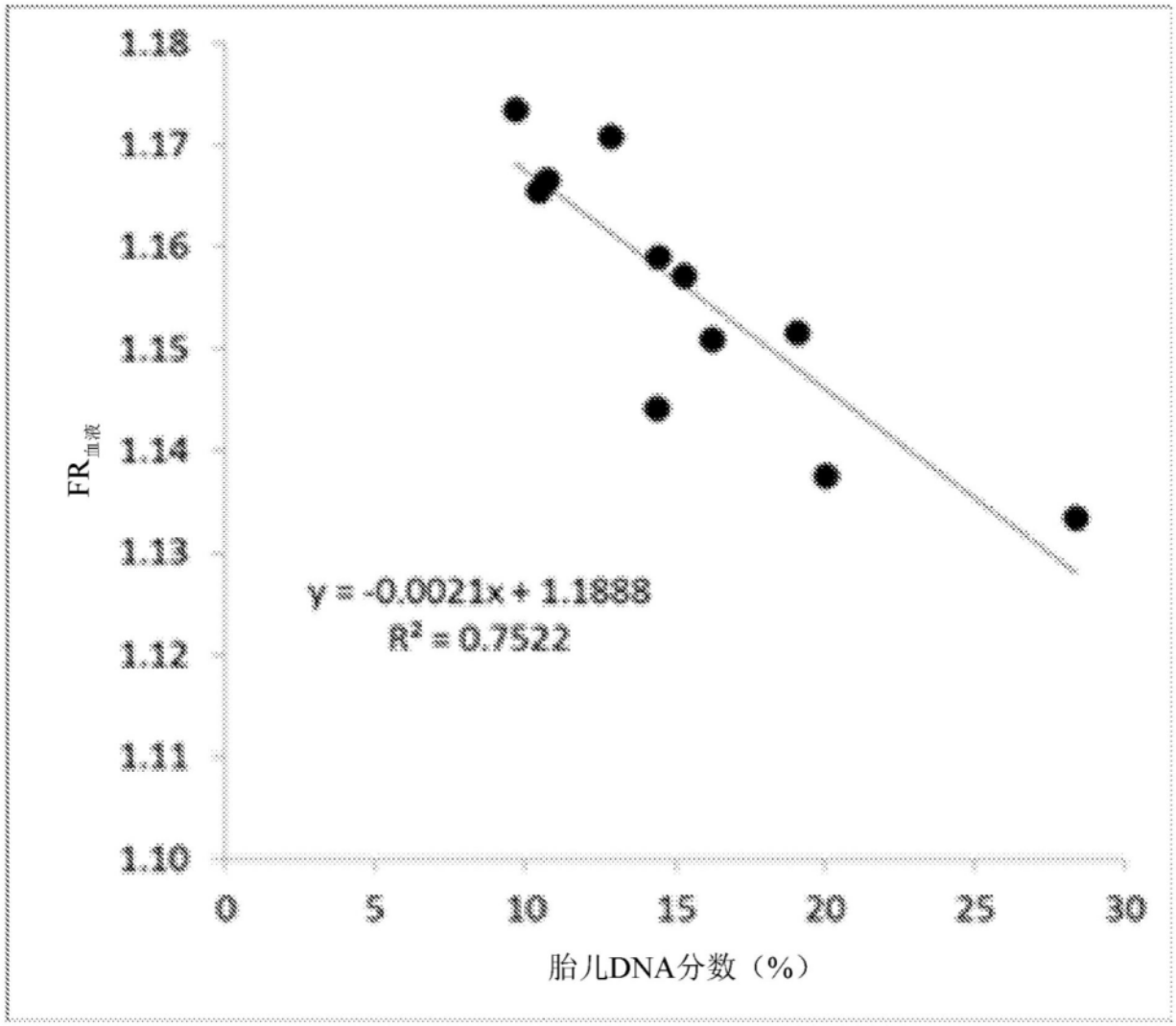

1.18
1.17
1.16
1.15
1.14
1.13
1.12
1.11
1.10
FR血液
y = -0.0021x + 1.1888
R² = 0.7522
0
5
10
15
20
25
30
胎儿DNA分数（%）

图12

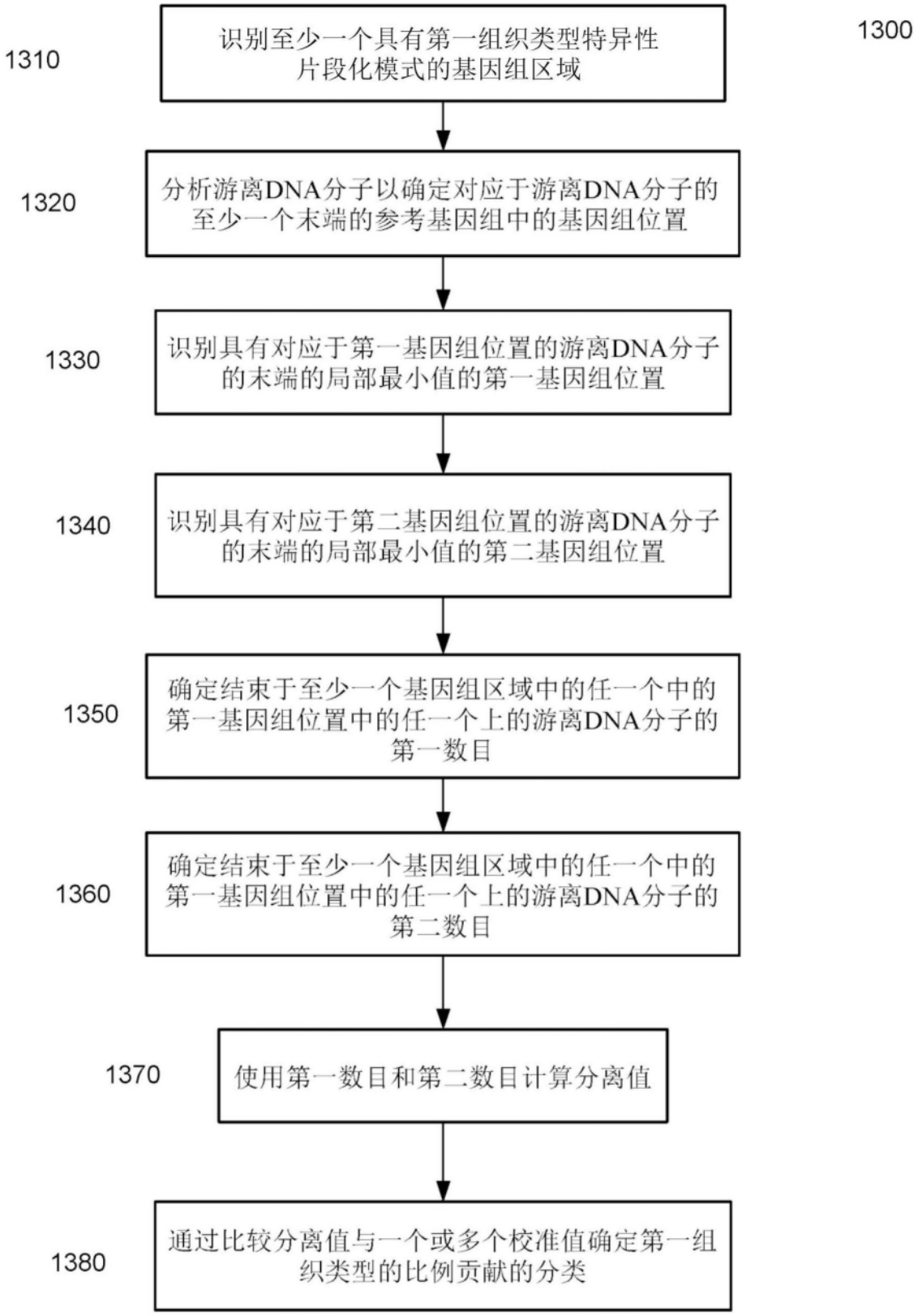

1300
1310
识别至少一个具有第一组织类型特异性片段化模式的基因组区域
1320
分析游离DNA分子以确定对应于游离DNA分子的至少一个末端的参考基因组中的基因组位置
1330
识别具有对应于第一基因组位置的游离DNA分子的末端的局部最小值的第一基因组位置
1340
识别具有对应于第二基因组位置的游离DNA分子的末端的局部最小值的第二基因组位置
1350
确定结束于至少一个基因组区域中的任一个中的第一基因组位置中的任一个上的游离DNA分子的第一数目
1360
确定结束于至少一个基因组区域中的任一个中的第一基因组位置中的任一个上的游离DNA分子的第二数目
1370
使用第一数目和第二数目计算分离值
1380
通过比较分离值与一个或多个校准值确定第一组织类型的比例贡献的分类

图13

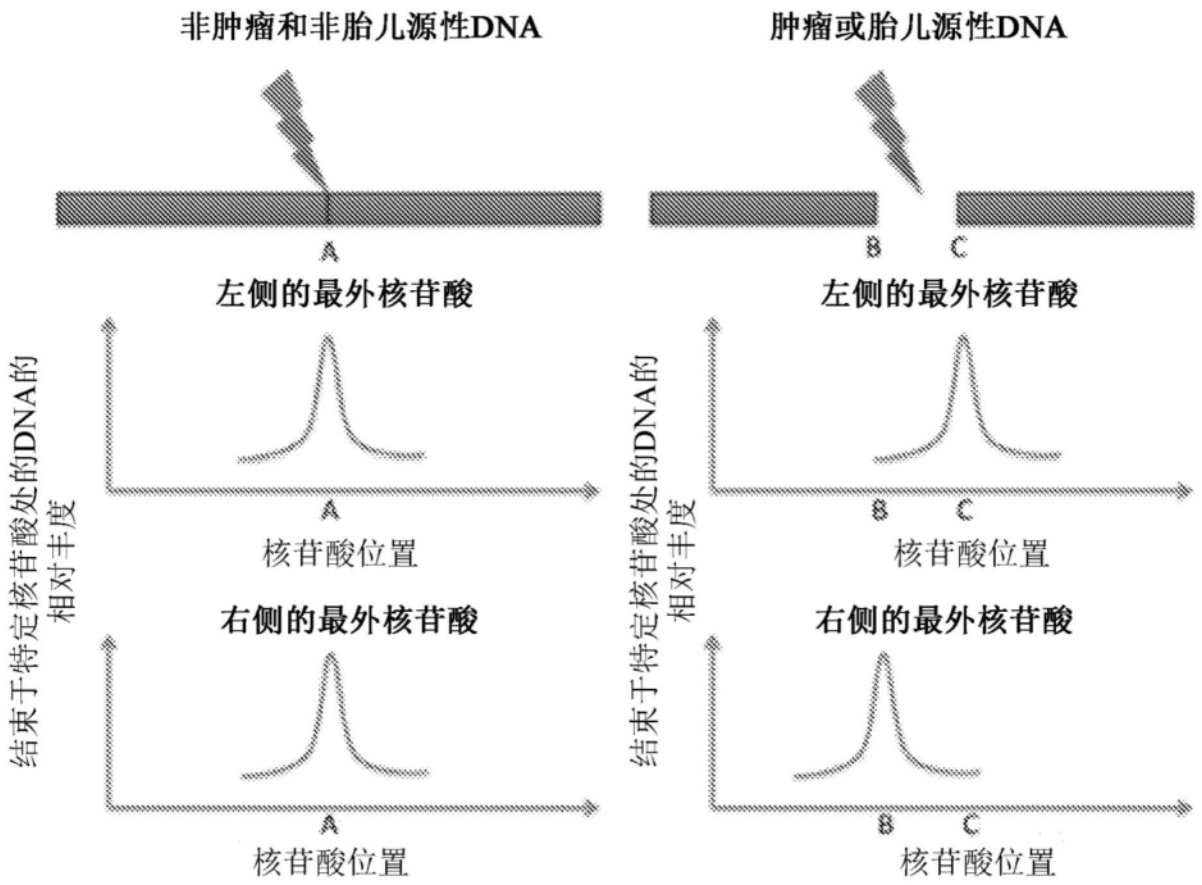

非肿瘤和非胎儿源性DNA
肿瘤或胎儿源性DNA
A
B
C
左侧的最外核苷酸
右侧的最外核苷酸
结束于特定核苷酸处的DNA的相对丰度
核苷酸位置

图14

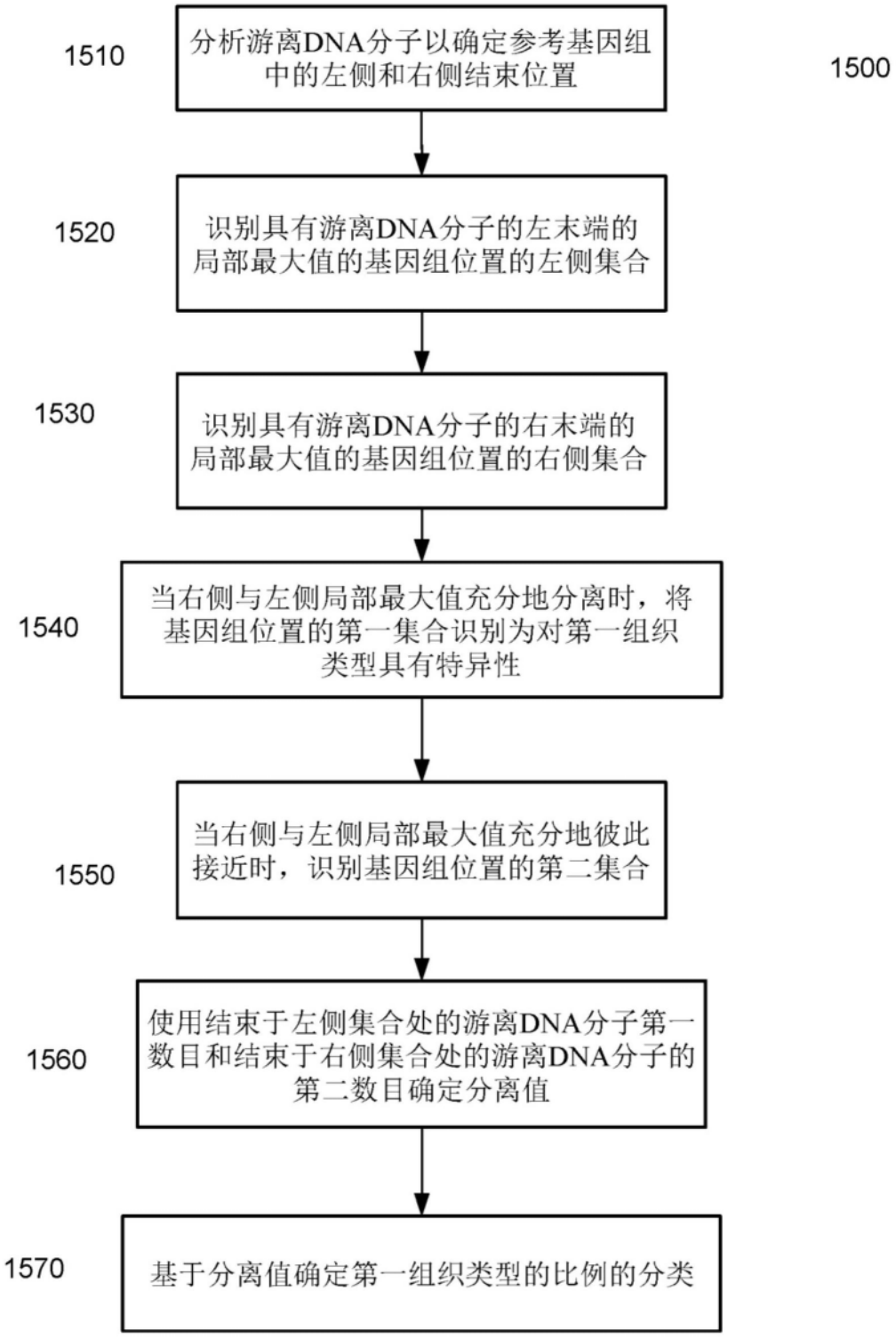

1500
1510
分析游离DNA分子以确定参考基因组中的左侧和右侧结束位置
1520
识别具有游离DNA分子的左末端的局部最大值的基因组位置的左侧集合
1530
识别具有游离DNA分子的右末端的局部最大值的基因组位置的右侧集合
1540
当右侧与左侧局部最大值充分地分离时，将基因组位置的第一集合识别为对第一组织类型具有特异性
1550
当右侧与左侧局部最大值充分地彼此接近时，识别基因组位置的第二集合
1560
使用结束于左侧集合处的游离DNA分子第一数目和结束于右侧集合处的游离DNA分子的第二数目确定分离值
1570
基于分离值确定第一组织类型的比例的分类

图15

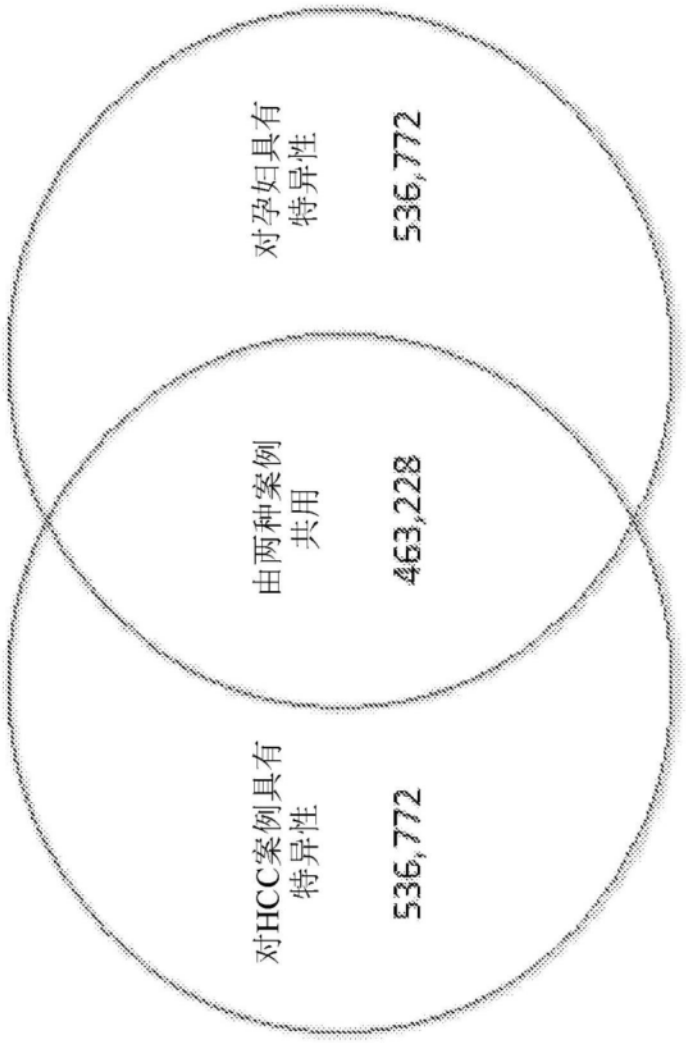


图16

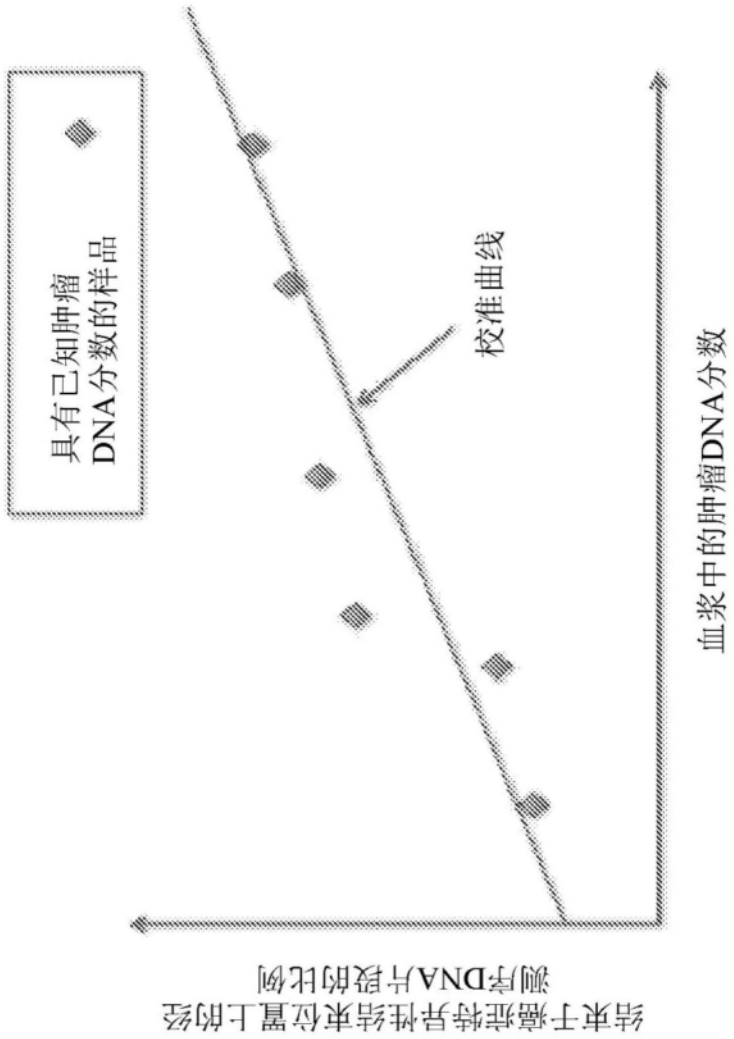

具有已知肿瘤
DNA分数的样品
校准曲线
血浆中的肿瘤DNA分数
结束于癌症特异性结束位置上的经
测序DNA片段的比例

图17

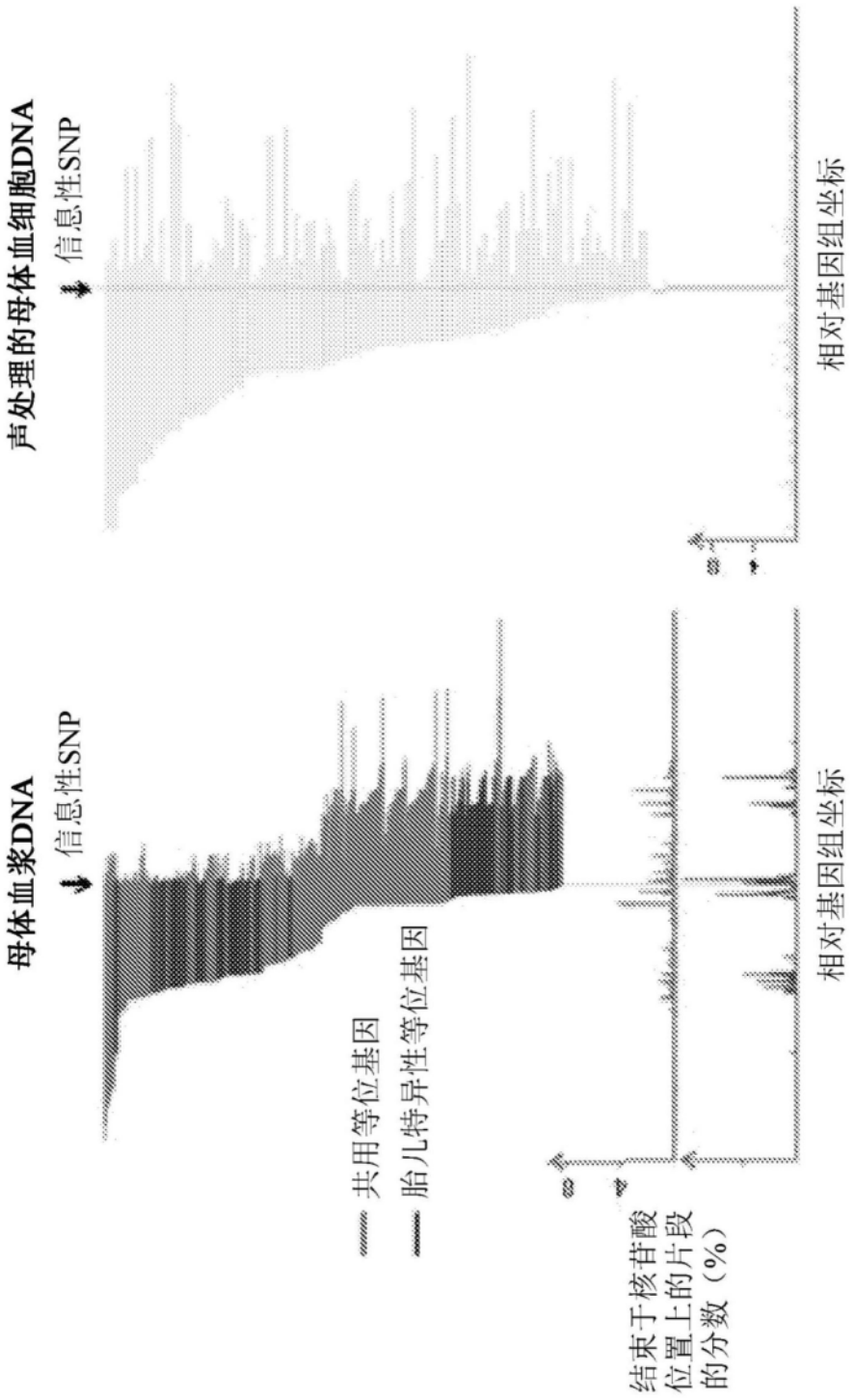

母体血浆DNA
信息性SNP
共用等位基因
胎儿特异性等位基因
结束于核苷酸位置上的片段的分数（%）
相对基因组坐标
声处理的母体血细胞DNA
信息性SNP
相对基因组坐标

图18

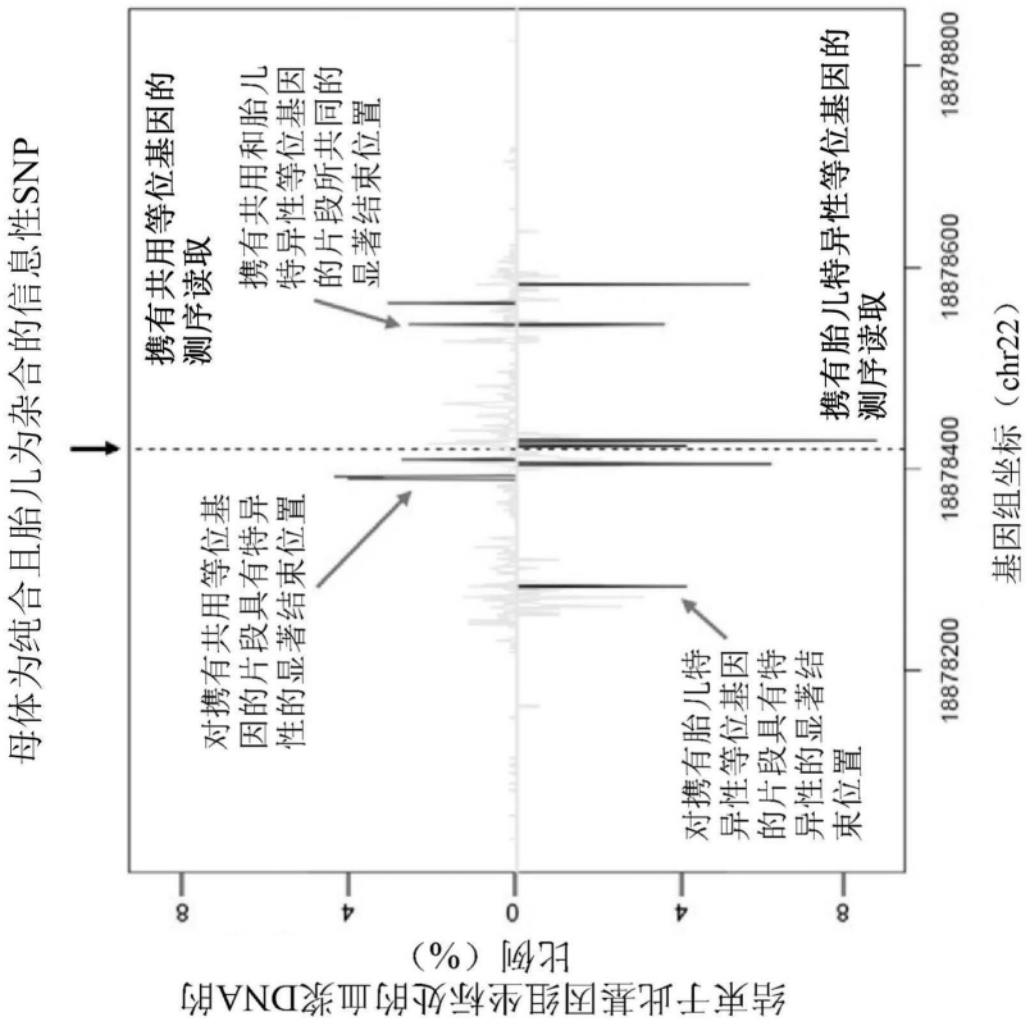

母体为纯合且胎儿为杂合的信息性SNP
携有共用等位基因的测序读取
携有共用和胎儿特异性等位基因的片段所共同的显著结束位置
对携有共用等位基因的片段具有特异性的显著结束位置
携有胎儿特异性等位基因的测序读取
对携有胎儿特异性等位基因的片段具有特异性的显著结束位置
8
4
0
4
8
比例（%）
18878200
18878400
18878600
18878800
基因组坐标（chr22）
关于生基因组坐标处的血浆DNA的

图19

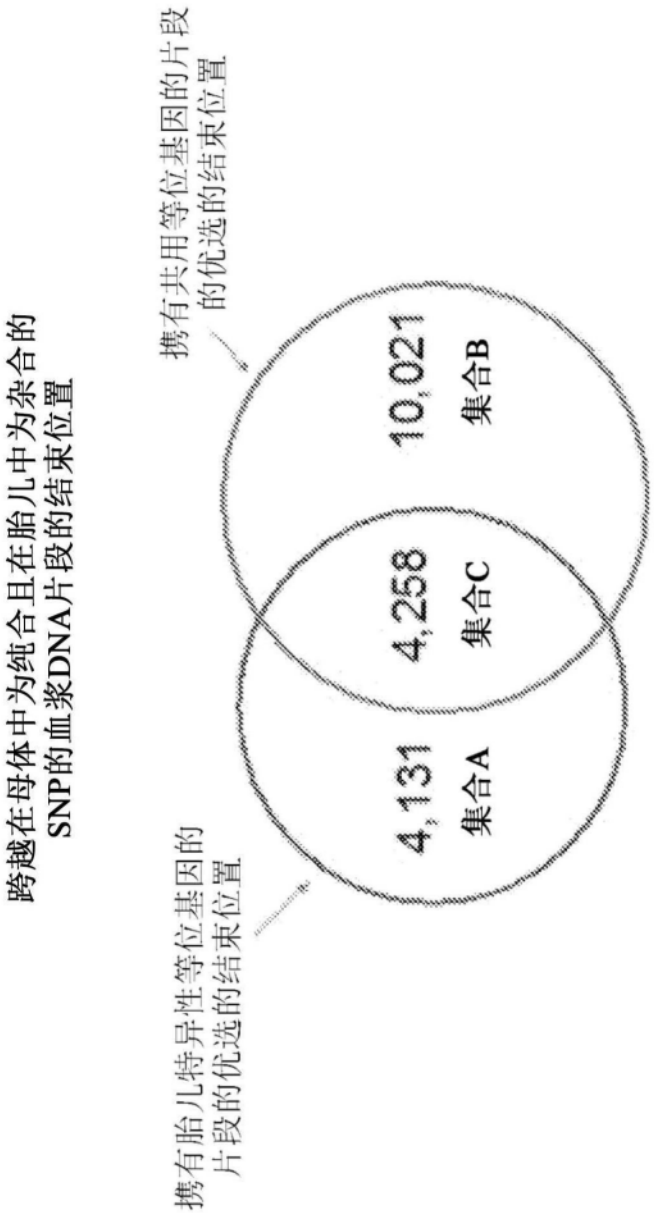

跨越在母体中为纯合且在胎儿中为杂合的
SNP的血浆DNA片段的结束位置
携有共用等位基因的片段
的优选的结束位置
携有胎儿特异性等位基因的
片段的优选的结束位置
4,131
集合A
4,258
集合C
10,021
集合B

图20

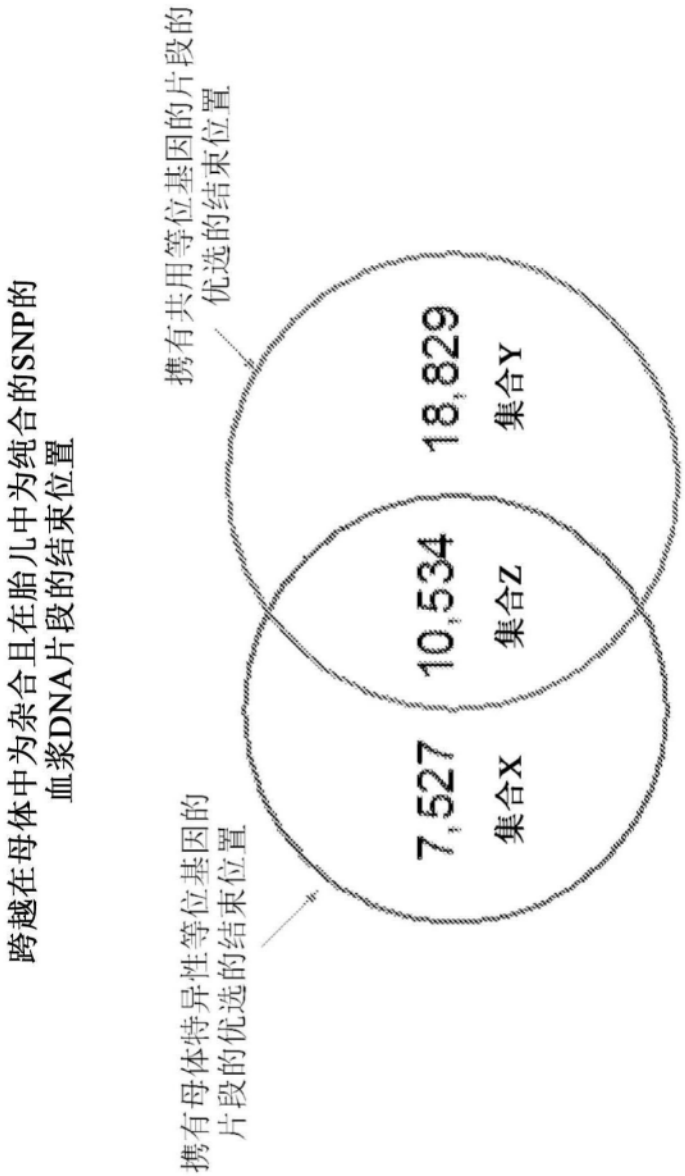

跨越在母体中为杂合且在胎儿中为纯合的SNP的
血浆DNA片段的结束位置
携有共用等位基因的片段的
优选的结束位置
携有母体特异性等位基因的
片段的优选的结束位置
7,527
集合X
10,534
集合Z
18,829
集合Y

图21

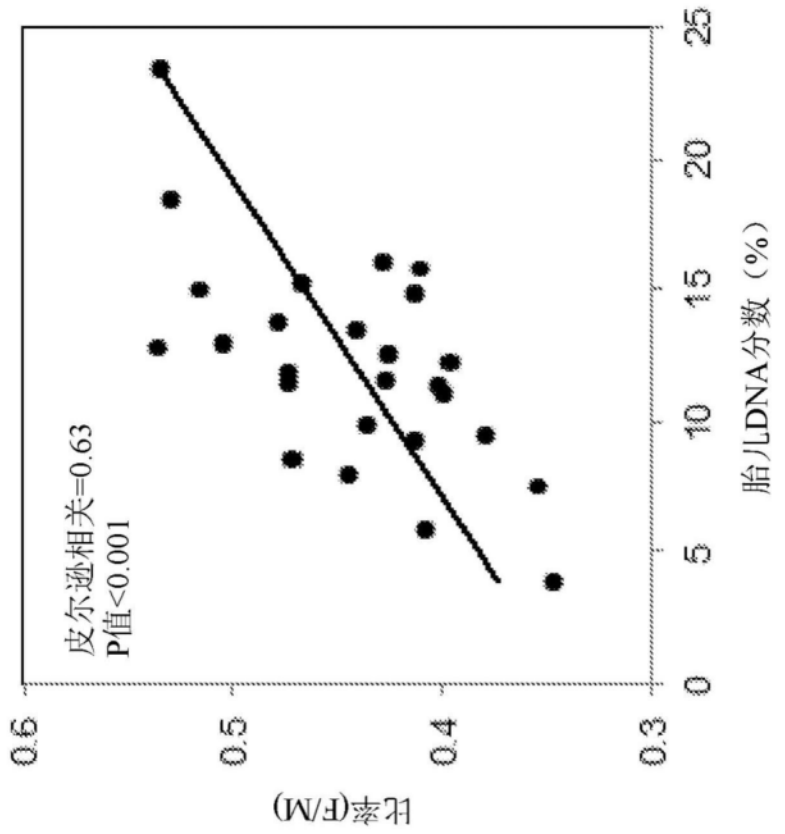


图22

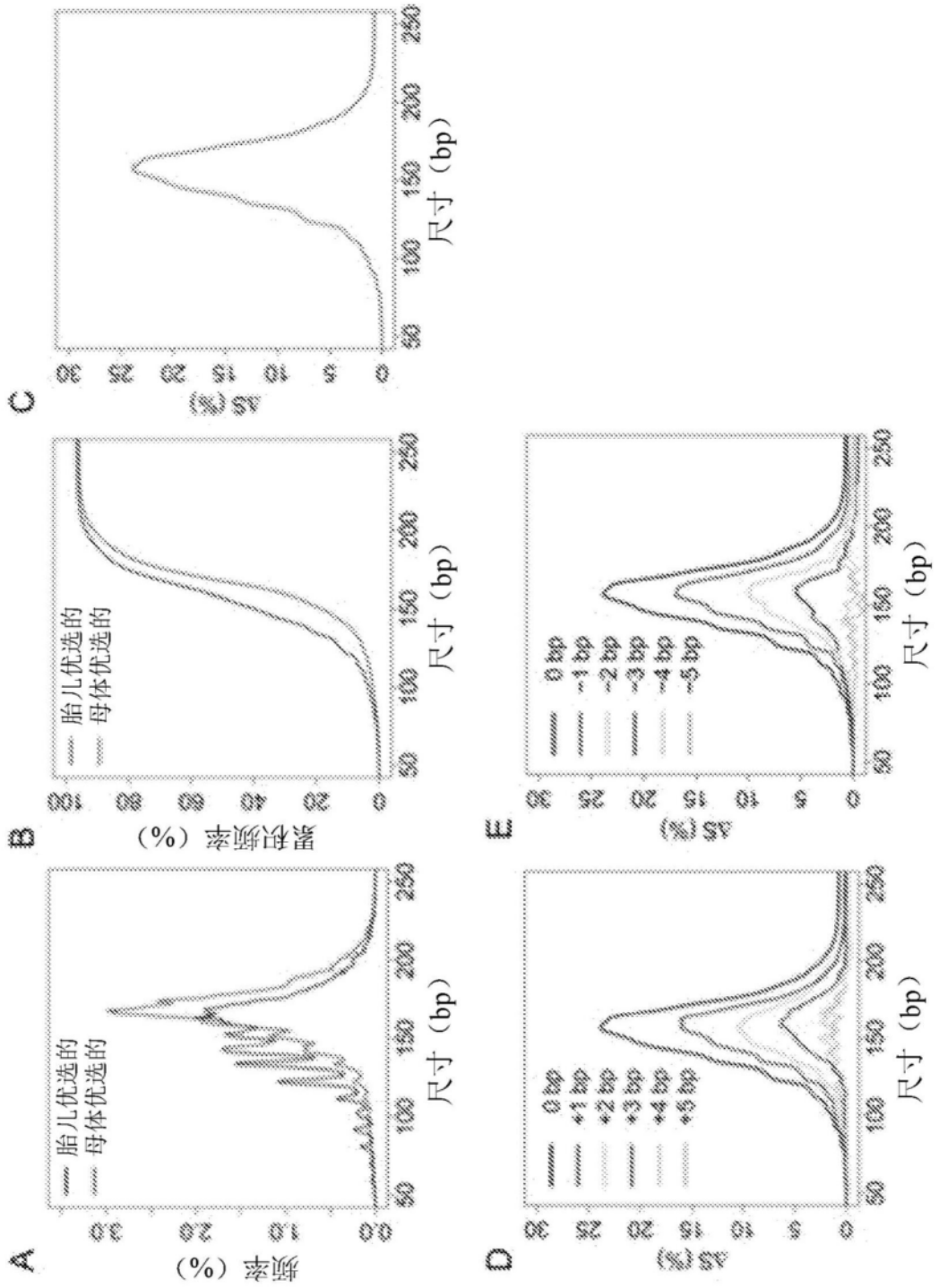

A
频率（%）
胎儿优选的
母体优选的
尺寸（bp）
B
累积频率（%）
胎儿优选的
母体优选的
尺寸（bp）
C
AS（%）
尺寸（bp）
D
AS（%）
0 bp
+1 bp
+2 bp
+3 bp
+4 bp
+5 bp
尺寸（bp）
E
AS（%）
0 bp
-1 bp
-2 bp
-3 bp
-4 bp
-5 bp
尺寸（bp）

图23

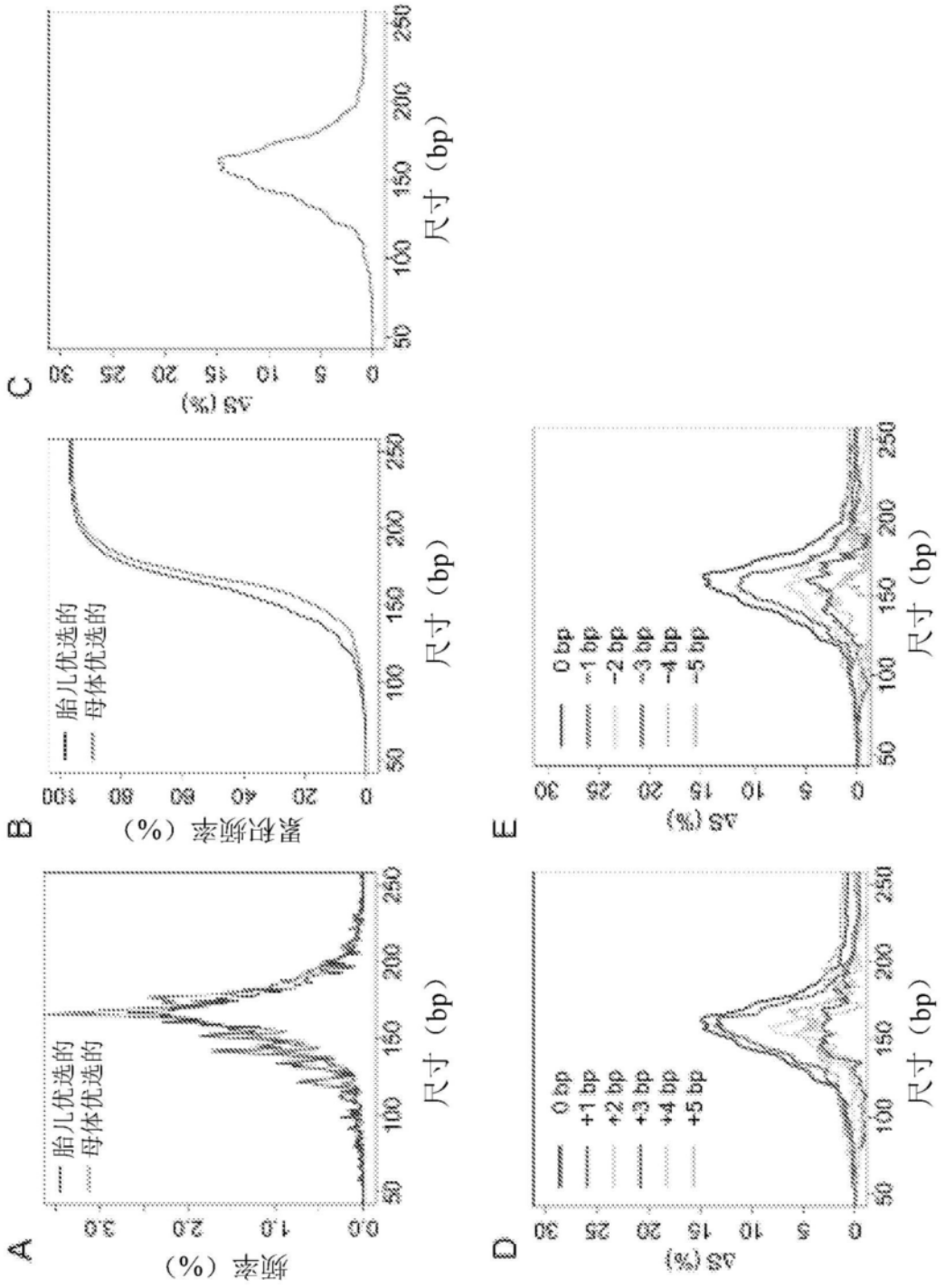

A
频率（%）
胎儿优选的
母体优选的
尺寸（bp）
B
累积频率（%）
胎儿优选的
母体优选的
尺寸（bp）
C
ΔS（%）
尺寸（bp）
D
ΔS（%）
0 bp
+1 bp
+2 bp
+3 bp
+4 bp
+5 bp
尺寸（bp）
E
ΔS（%）
0 bp
-1 bp
-2 bp
-3 bp
-4 bp
-5 bp
尺寸（bp）

图24

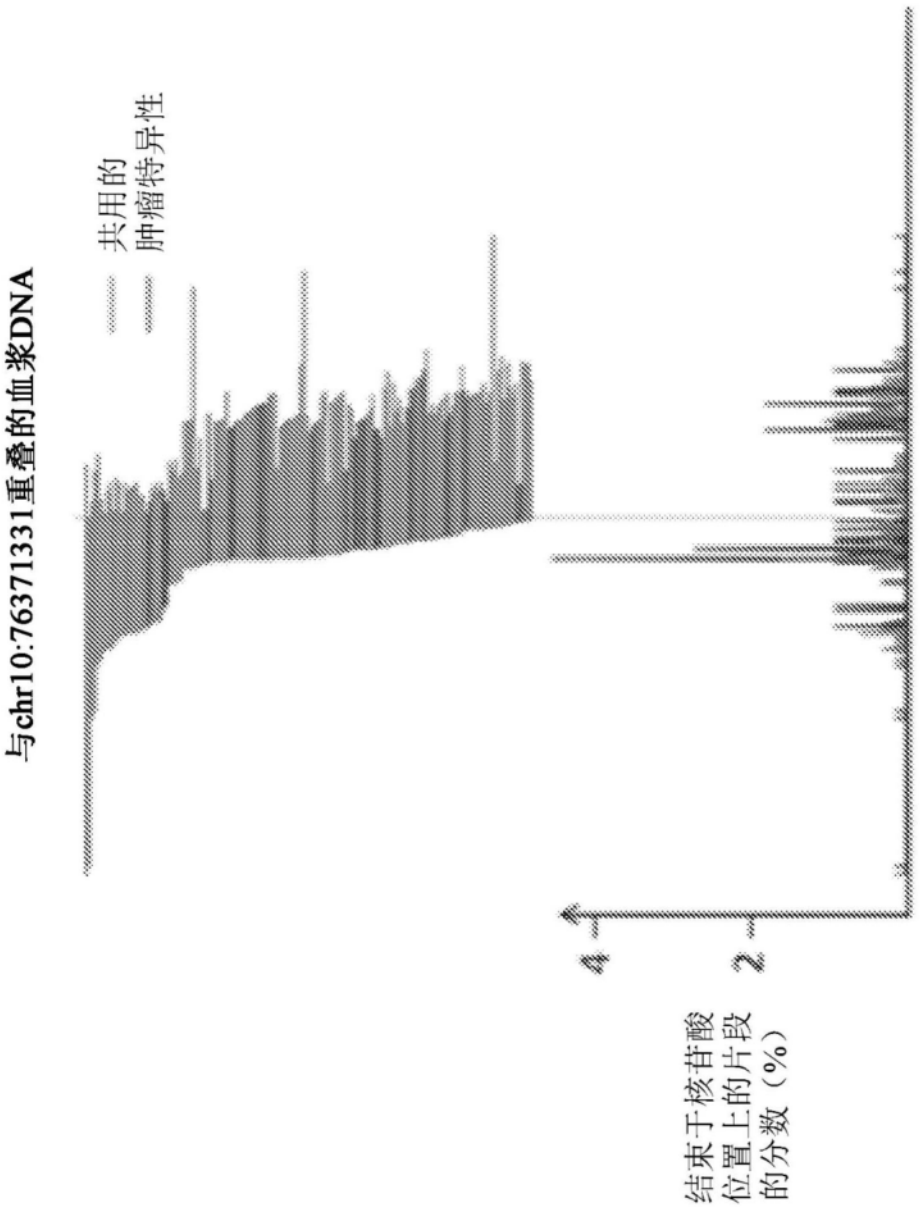


图25

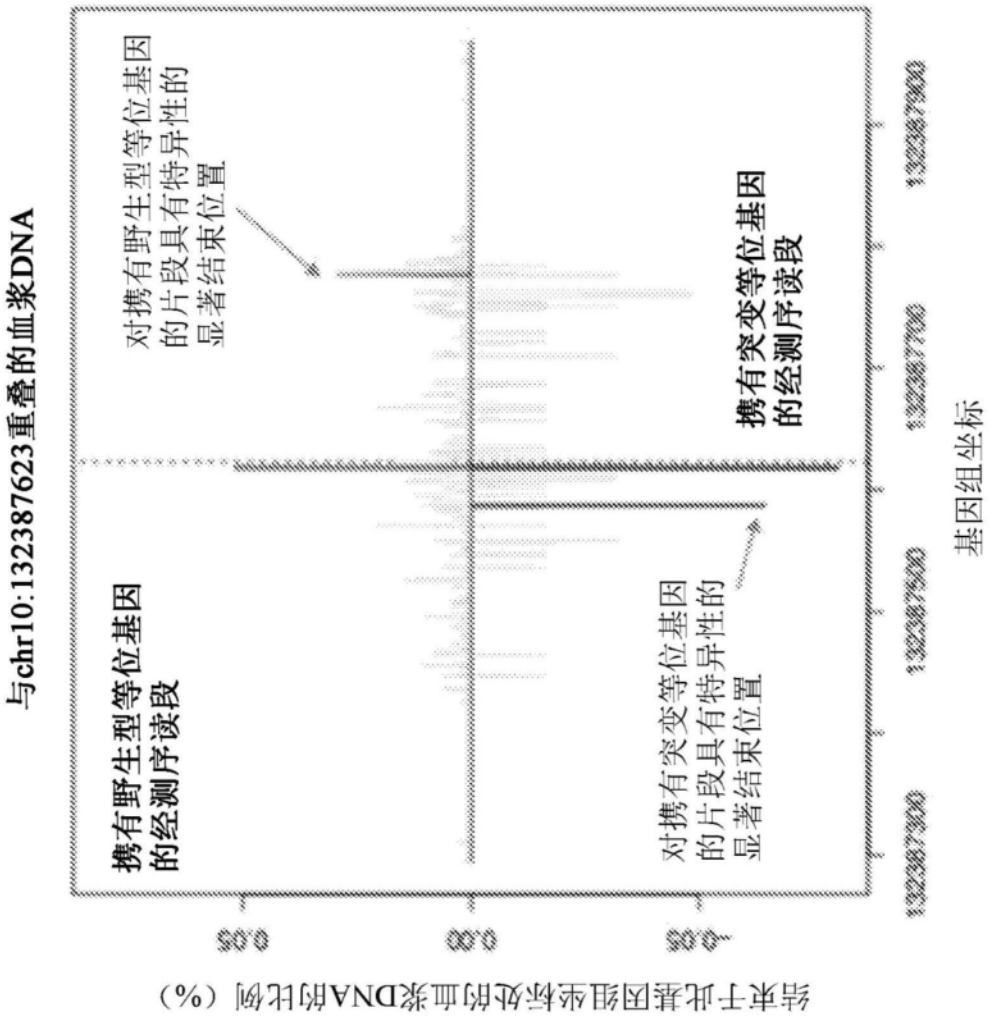

与chr10:132387623重叠的血浆DNA
携有野生型等位基因的经测序读段
对携有野生型等位基因的片段具有特异性的显著结束位置
携有突变等位基因的经测序读段
对携有突变等位基因的片段具有特异性的显著结束位置
结束于此基因组坐标处的血浆DNA的比例（%）
基因组坐标
132387300
132387500
132387700
132387900

图26

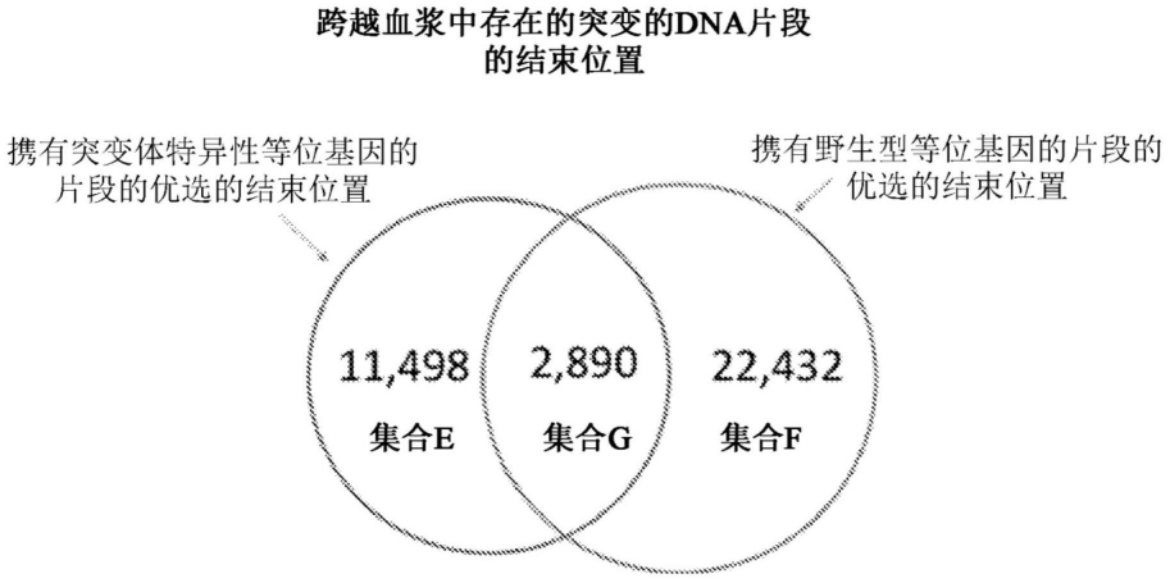

跨越血浆中存在的突变的DNA片段
的结束位置
携有突变体特异性等位基因的
片段的优选的结束位置
携有野生型等位基因的片段的
优选的结束位置
11,498
集合E
2,890
集合G
22,432
集合F

图27A

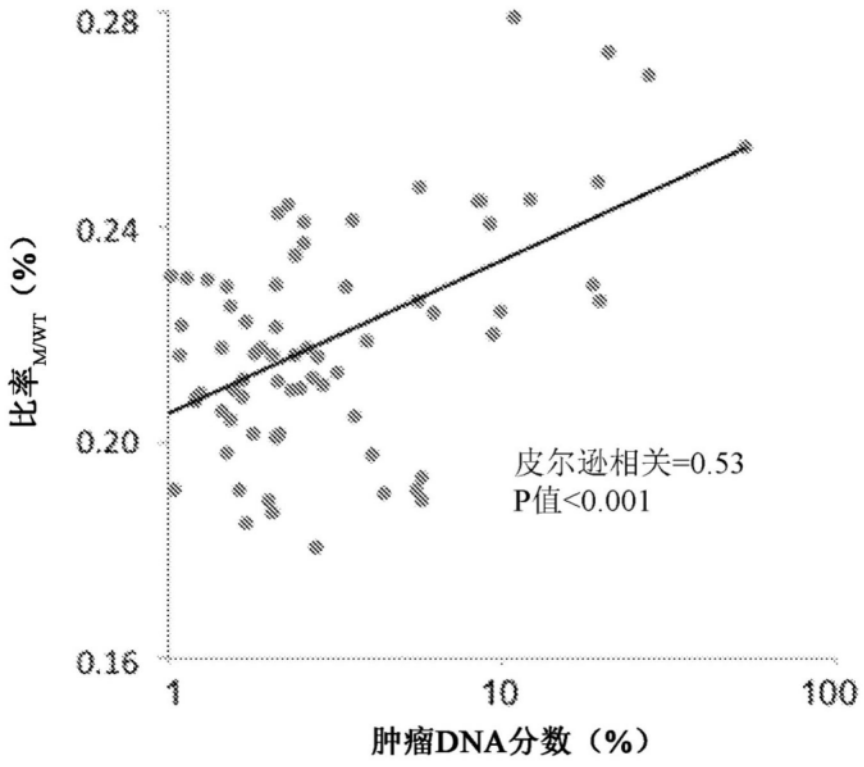

0.28
0.24
0.20
0.16
比率M/WT（%）
皮尔逊相关=0.53
P值<0.001
1
10
100
肿瘤DNA分数（%）

图27B

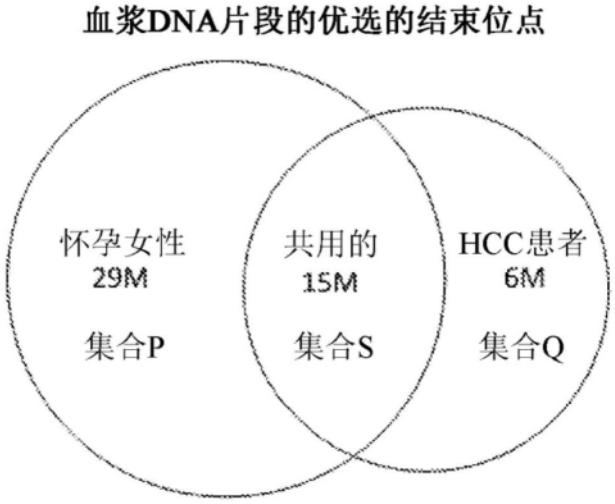

血浆DNA片段的优选的结束位点
怀孕女性
29M
集合P
共用的
15M
集合S
HCC患者
6M
集合Q

图28A

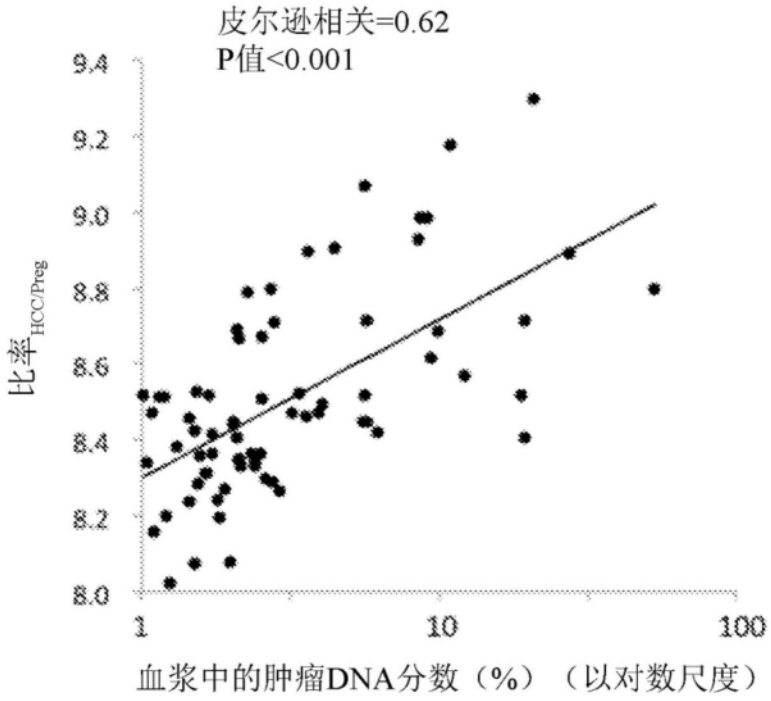

皮尔逊相关=0.62
P值<0.001
9.4
9.2
9.0
8.8
8.6
8.4
8.2
8.0
比率HCC/Preg
1
10
100
血浆中的肿瘤DNA分数（%）（以对数尺度）

图28B

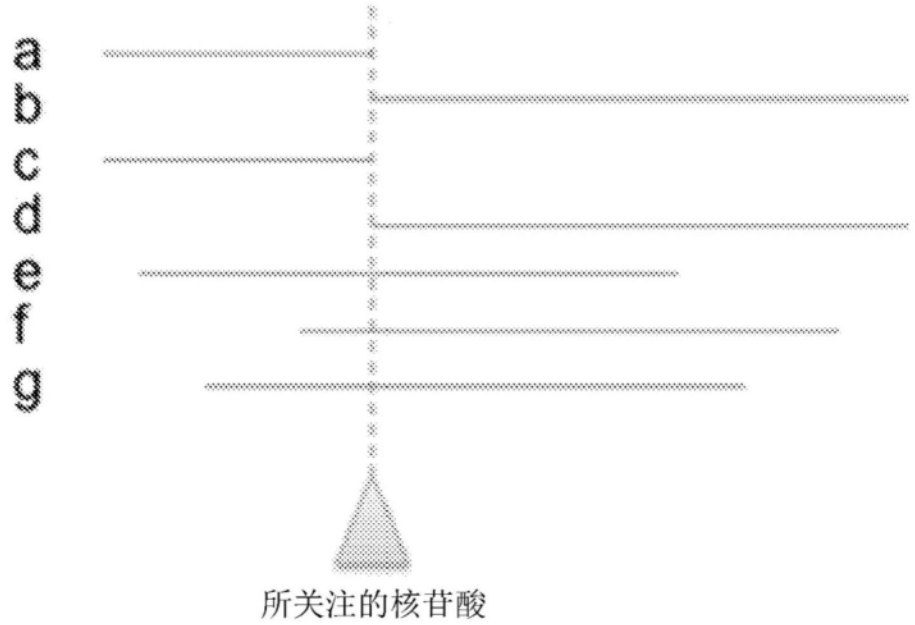

a
b
c
d
e
f
g
所关注的核苷酸

图29A

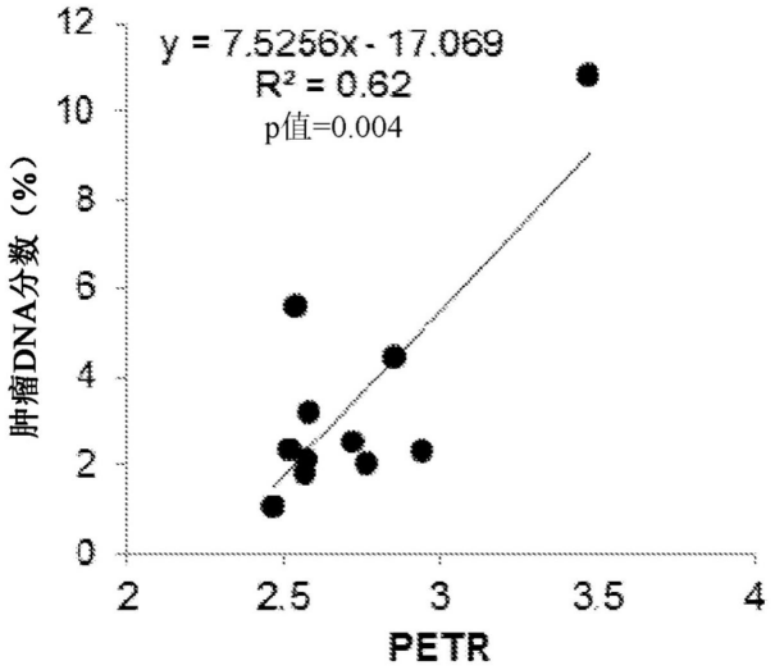

y = 7.5256x - 17.069
R² = 0.62
p值=0.004
肿瘤DNA分数（%）
12
10
8
6
4
2
0
2
2.5
3
3.5
4
PETR

图29B

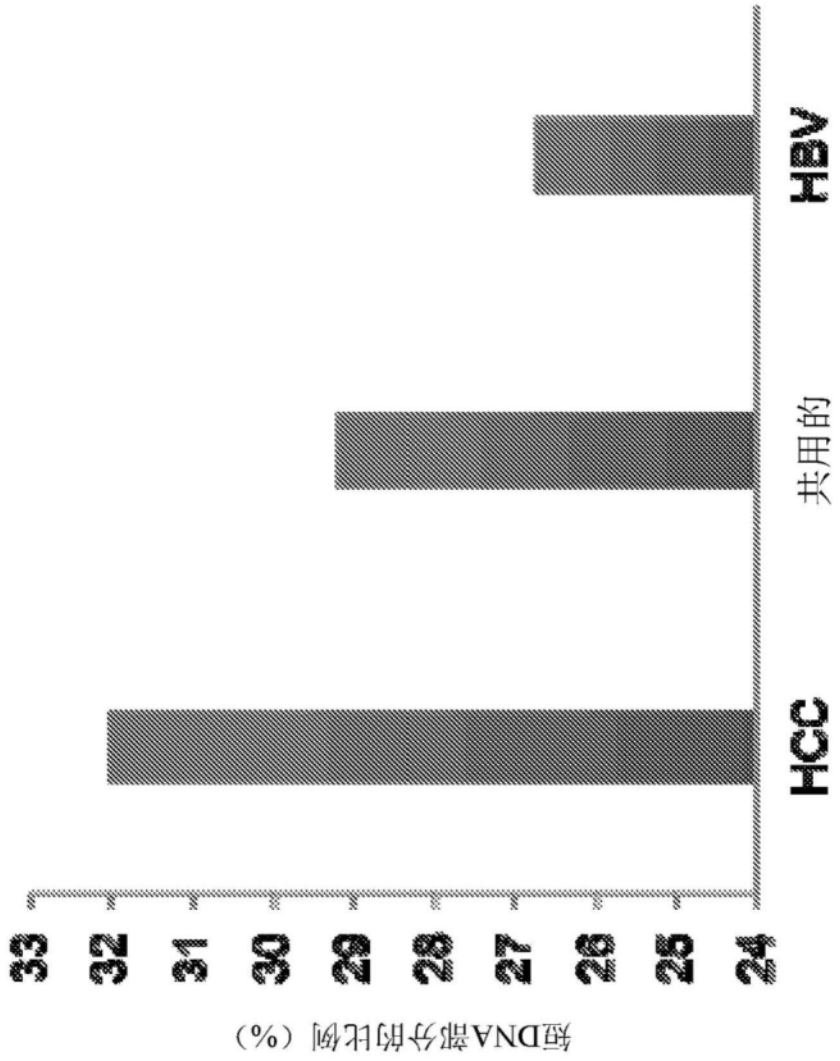

HBV
共用的
HCC
33
32
31
30
29
28
27
26
25
24
短DNA部分的比例（%）

图30


优选的结束位置
基因组坐标
窗口A
窗口B

图31A

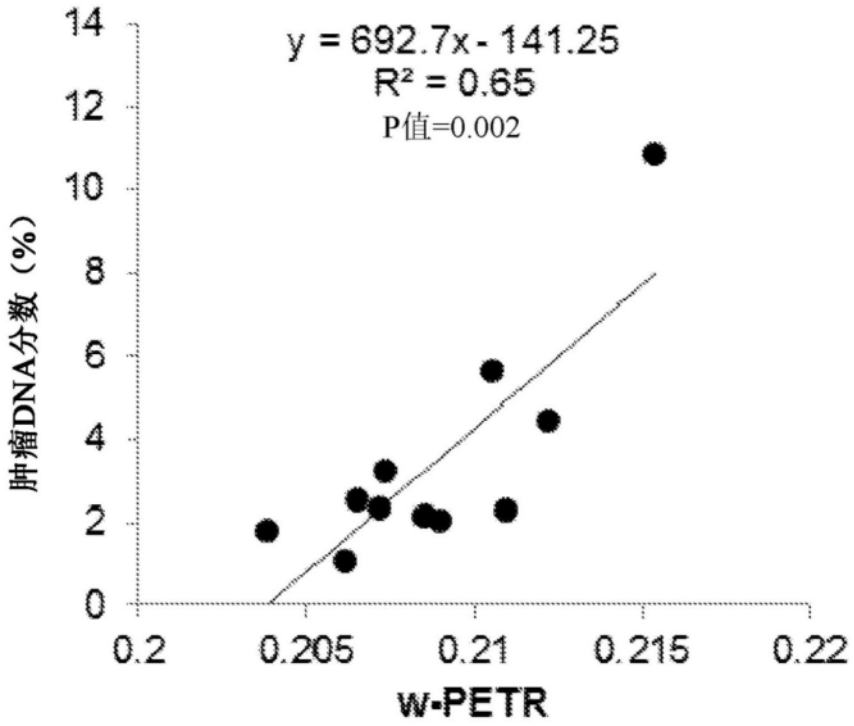

y = 692.7x - 141.25
R² = 0.65
P值=0.002
肿瘤DNA分数（%）
14
12
10
8
6
4
2
0
0.2
0.205
0.21
0.215
0.22
w-PETR

图31B

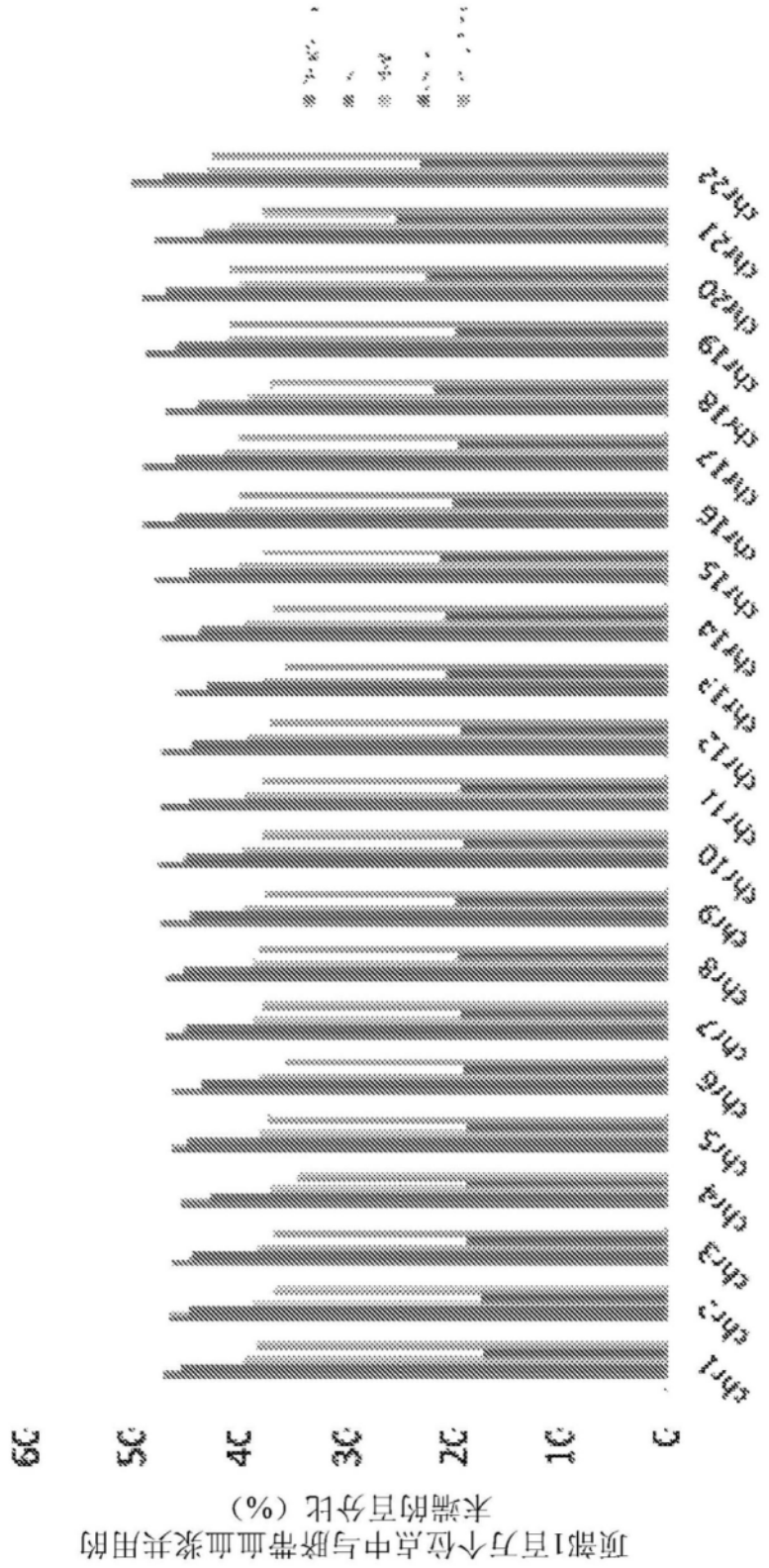

chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
0
10
20
30
40
50
60
顶部1百万个位点中与豚带血血浆共用的
末端的百分比（%）

图32

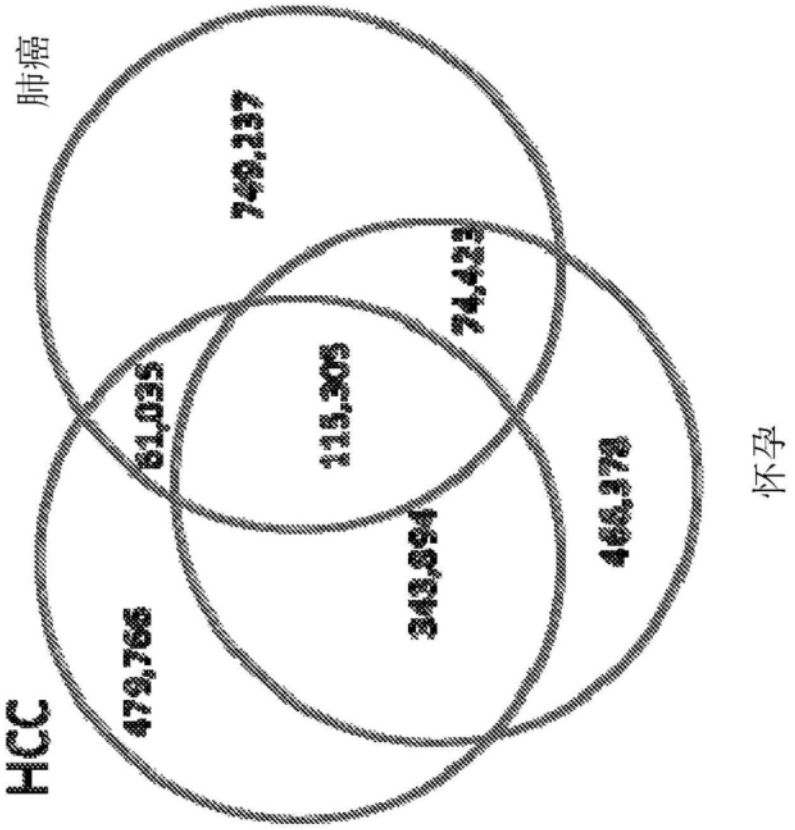

肺癌
749,237
74,423
81,035
115,205
343,894
464,378
479,766
HCC
怀孕

图33

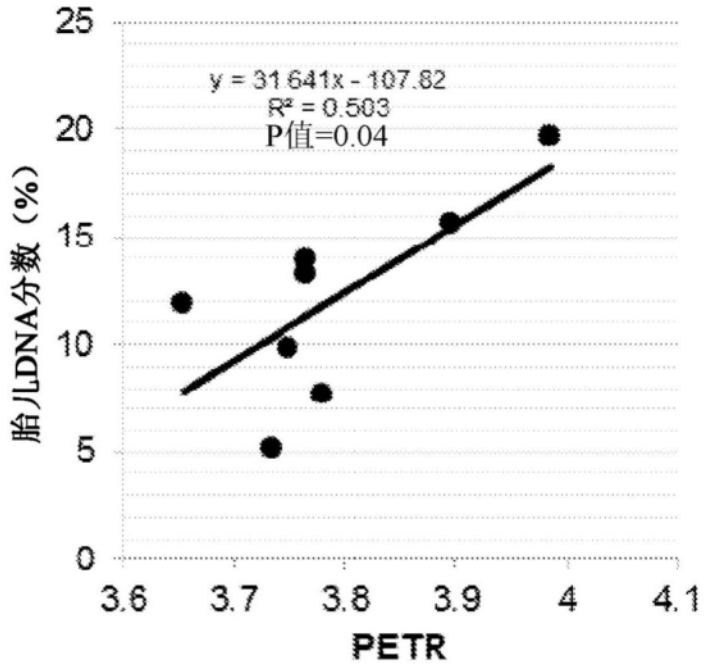

y = 31.641x - 107.82
R² = 0.503
P值=0.04
胎儿DNA分数（%）
25
20
15
10
5
0
3.6
3.7
3.8
3.9
4
4.1
PETR

图34A

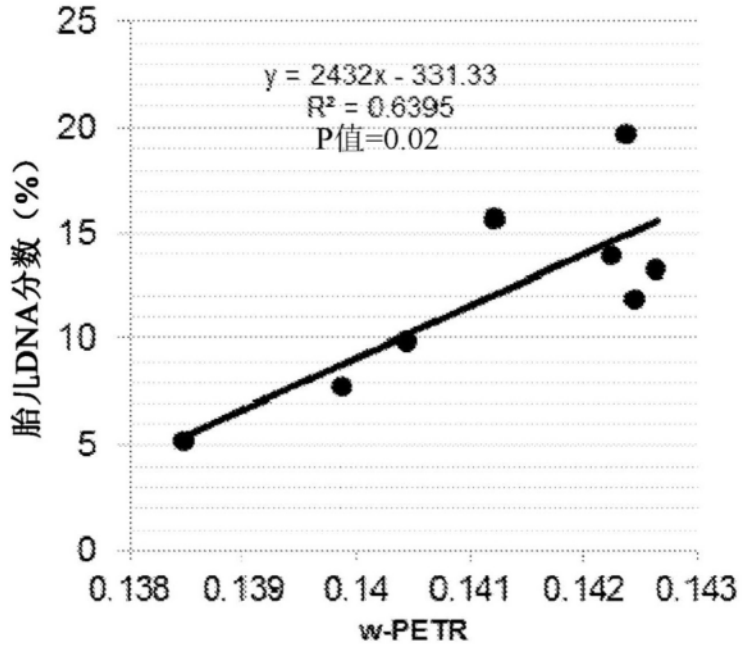

y = 2432x - 331.33
R² = 0.6395
P值=0.02
胎儿DNA分数（%）
25
20
15
10
5
0
0.138
0.139
0.14
0.141
0.142
0.143
w-PETR

图34B

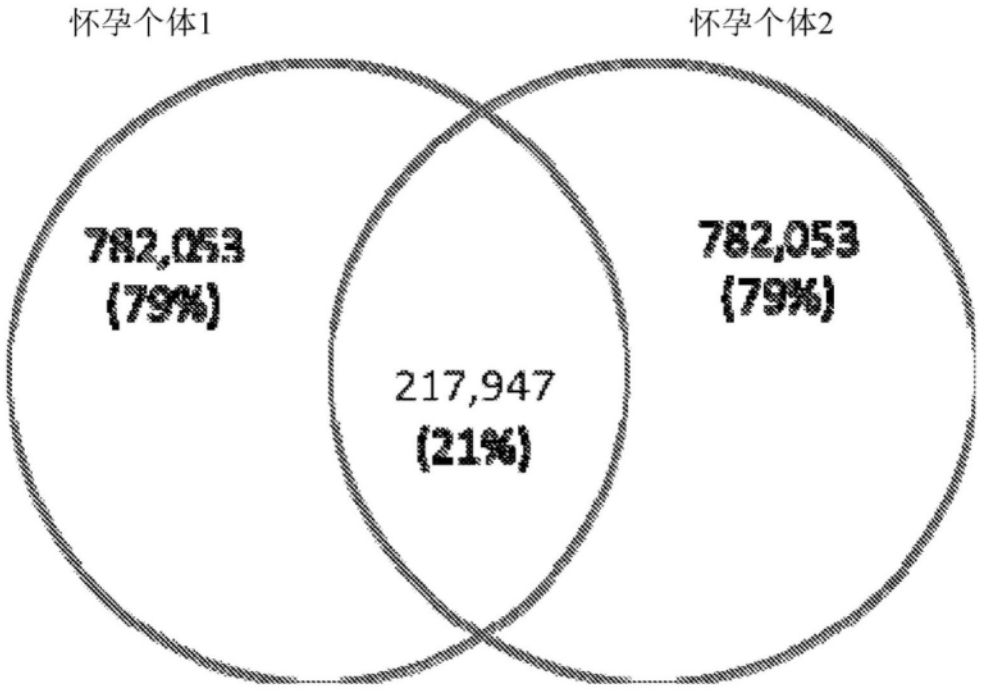

怀孕个体1
怀孕个体2
782,053
(79%)
217,947
(21%)
782,053
(79%)

图35A

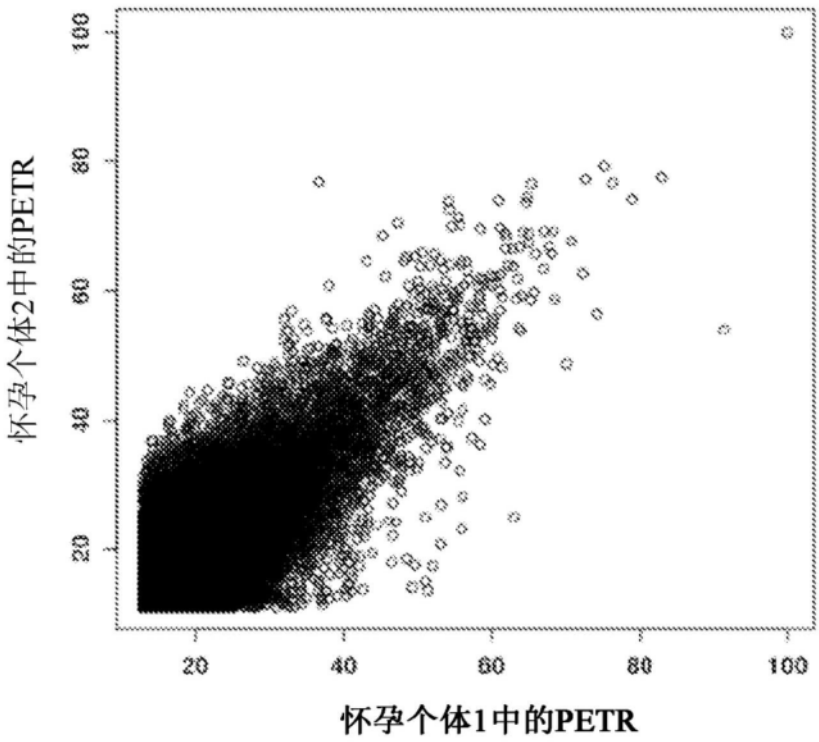


图35B

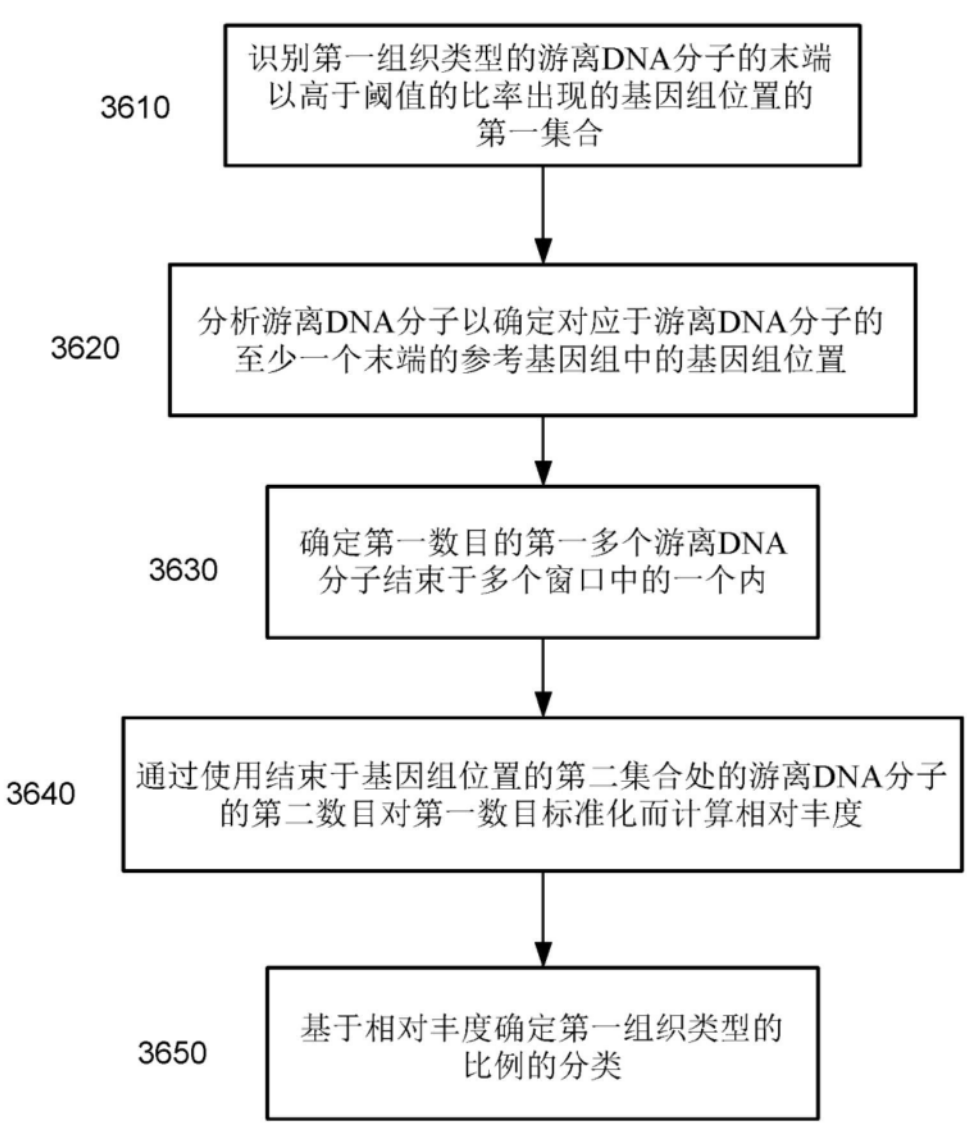

3600
3610
识别第一组织类型的游离DNA分子的末端以高于阈值的比率出现的基因组位置的第一集合
3620
分析游离DNA分子以确定对应于游离DNA分子的至少一个末端的参考基因组中的基因组位置
3630
确定第一数目的第一多个游离DNA分子结束于多个窗口中的一个内
3640
通过使用结束于基因组位置的第二集合处的游离DNA分子的第二数目对第一数目标准化而计算相对丰度
3650
基于相对丰度确定第一组织类型的比例的分类

图36

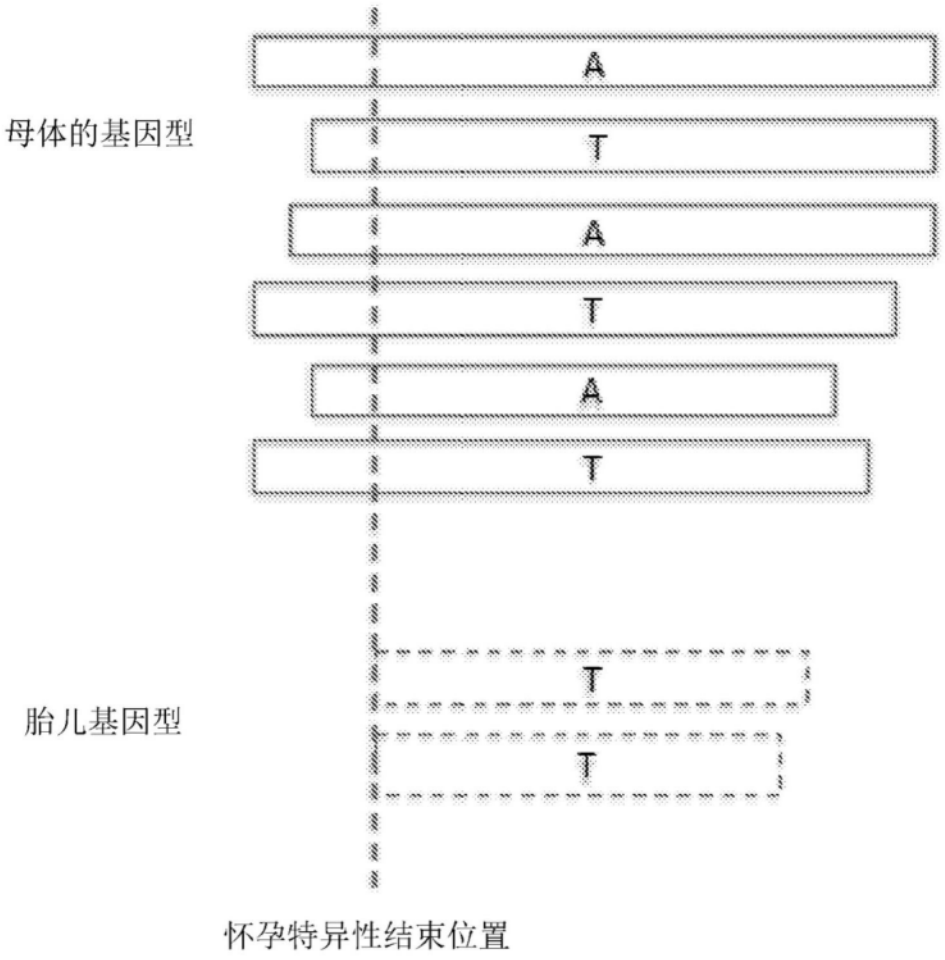


图37

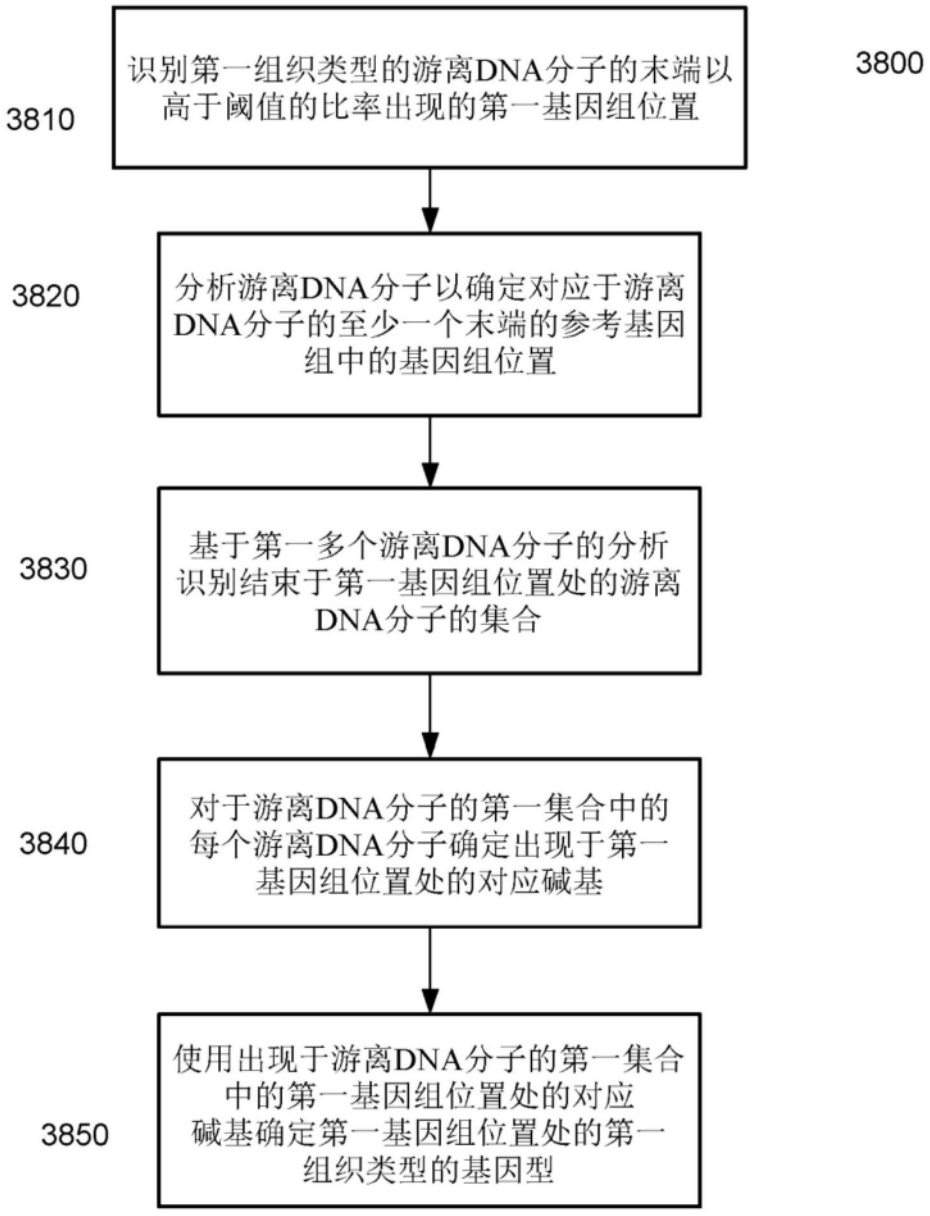

3800
3810
识别第一组织类型的游离DNA分子的末端以高于阈值的比率出现的第一基因组位置
3820
分析游离DNA分子以确定对应于游离DNA分子的至少一个末端的参考基因组中的基因组位置
3830
基于第一多个游离DNA分子的分析识别结束于第一基因组位置处的游离DNA分子的集合
3840
对于游离DNA分子的第一集合中的每个游离DNA分子确定出现于第一基因组位置处的对应碱基
3850
使用出现于游离DNA分子的第一集合中的第一基因组位置处的对应碱基确定第一基因组位置处的第一组织类型的基因型

图38

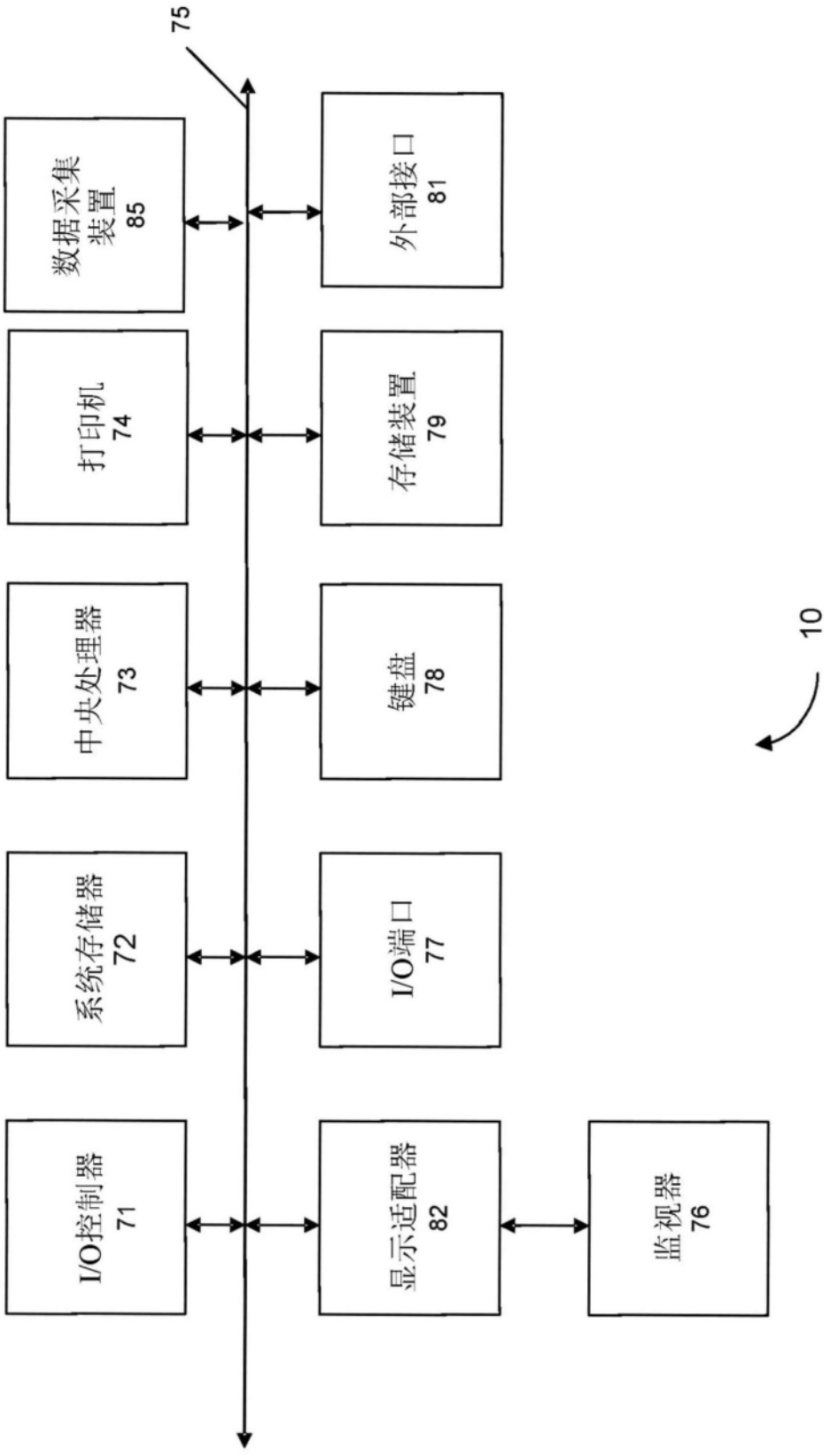
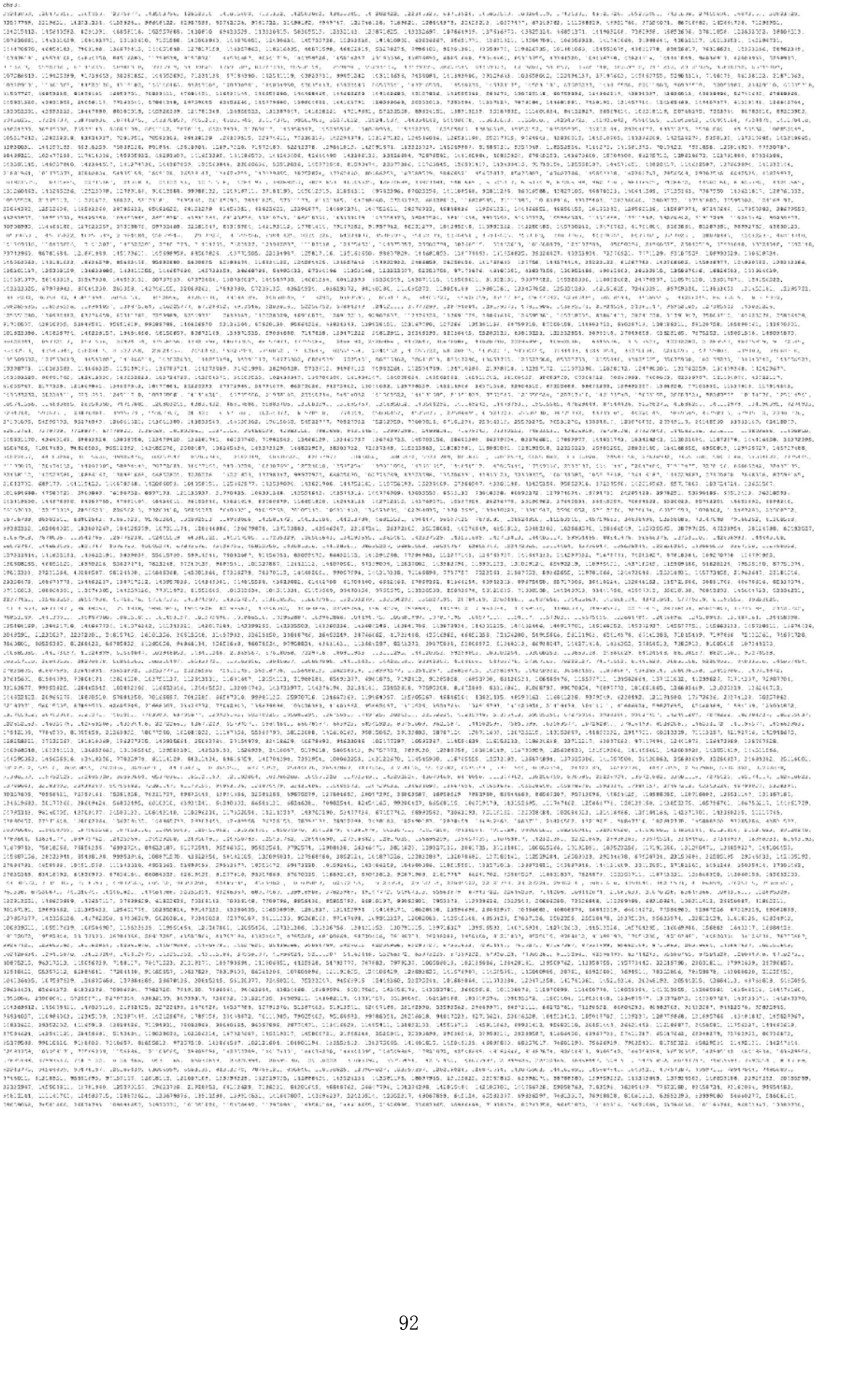
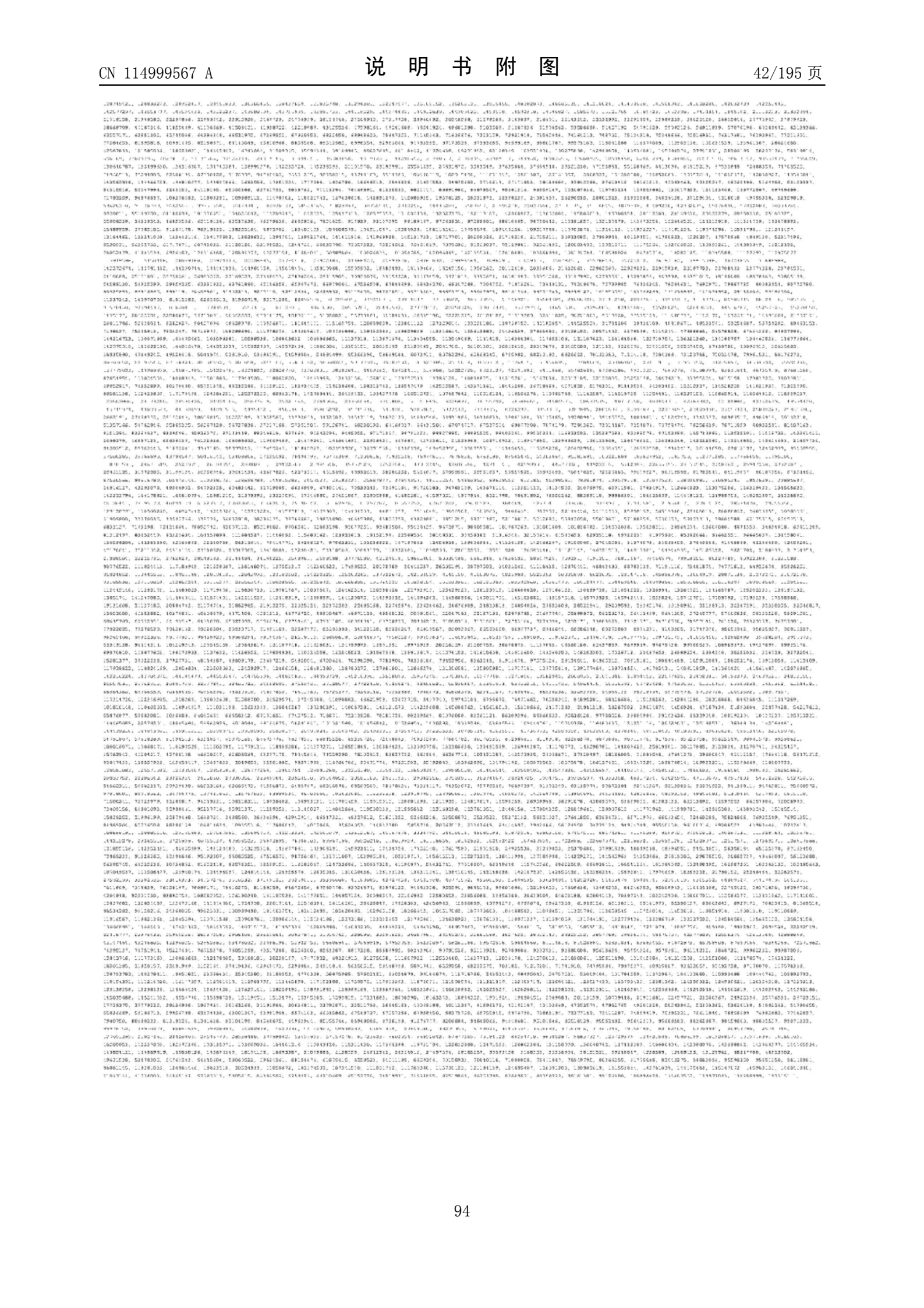

75
数据采集装置 85
外部接口 81
打印机 74
存储装置 79
中央处理器 73
键盘 78
系统存储器 72
I/O端口 77
I/O控制器 71
显示适配器 82
监视器 76
10

图39

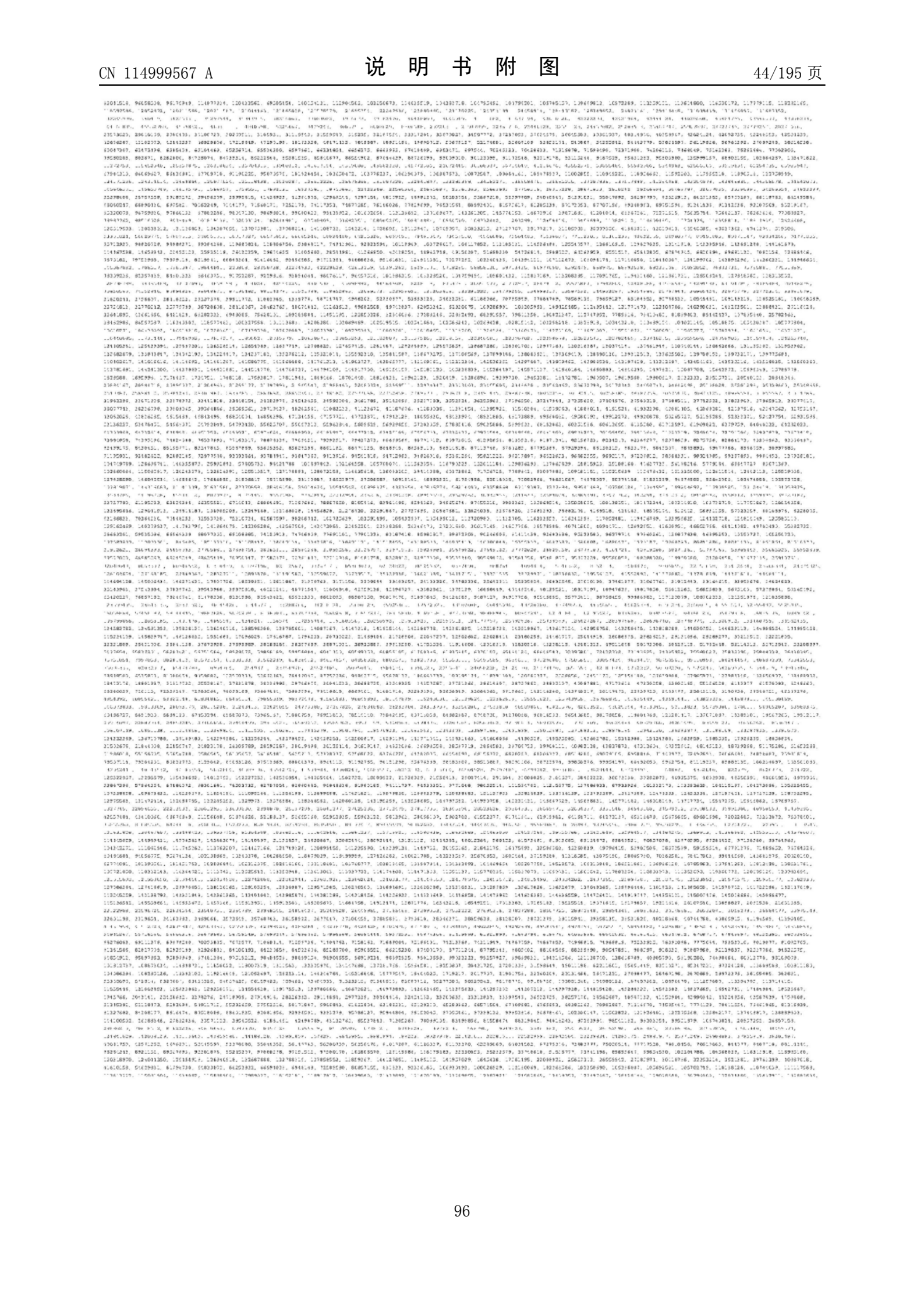
附录 A

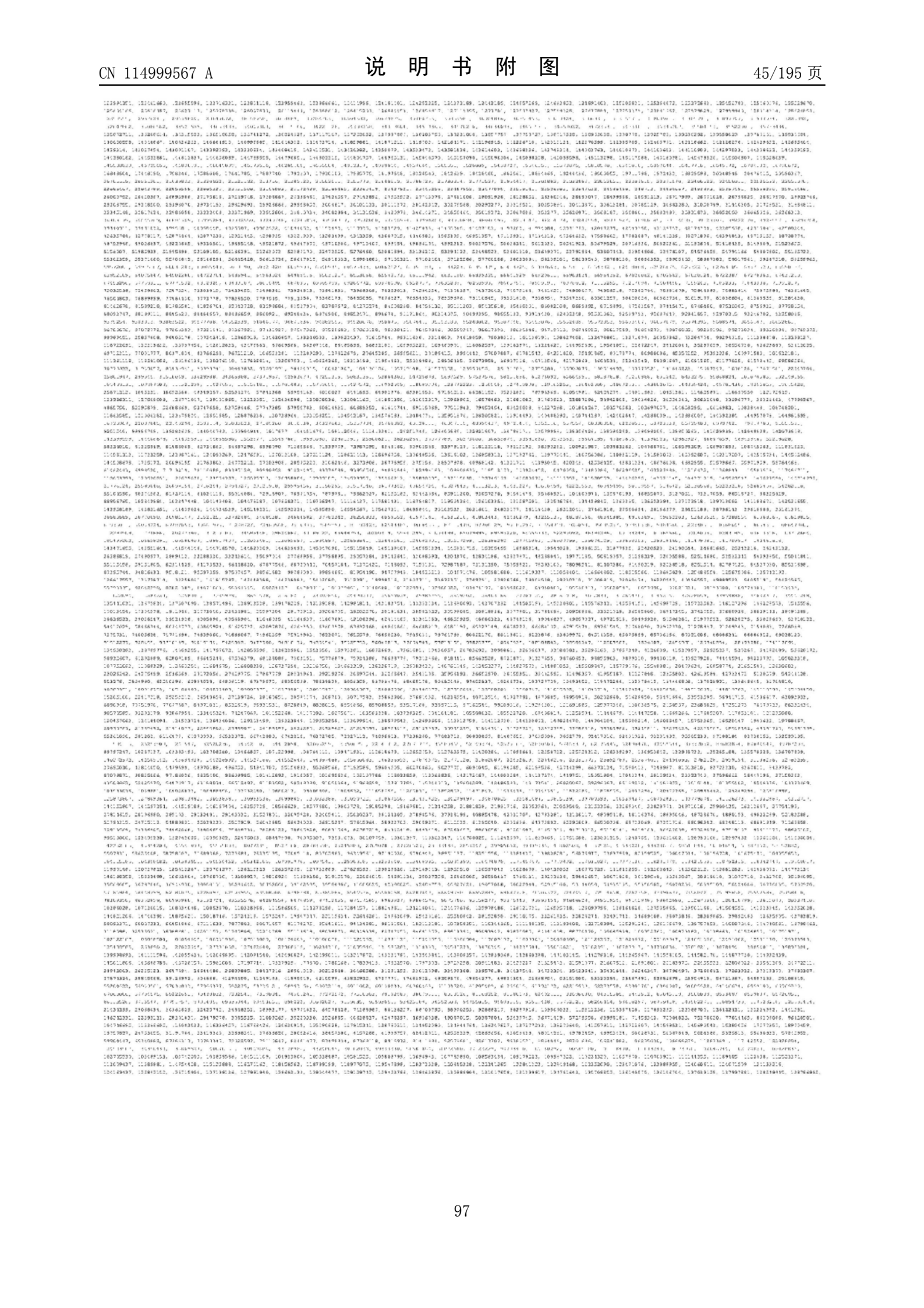


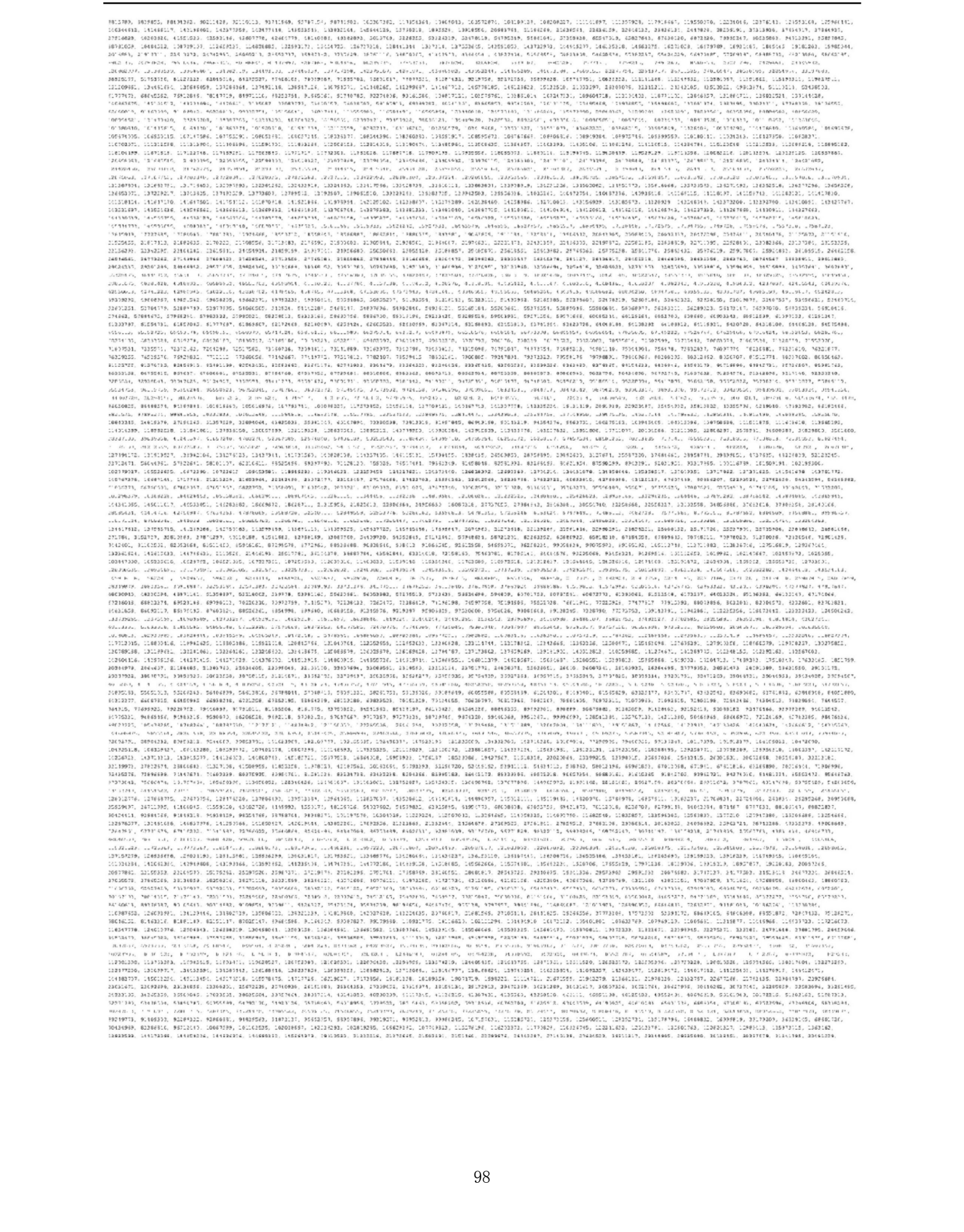


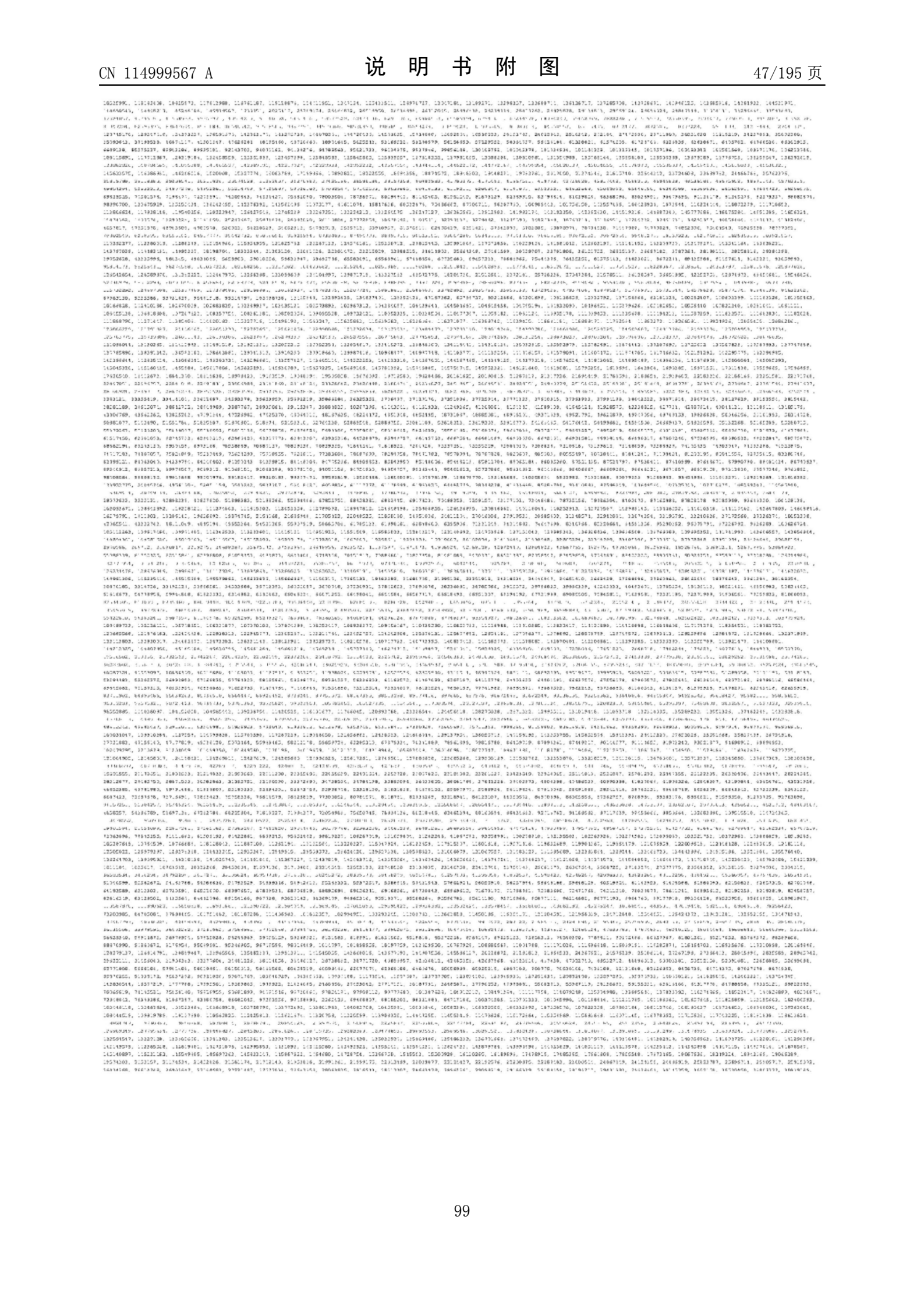


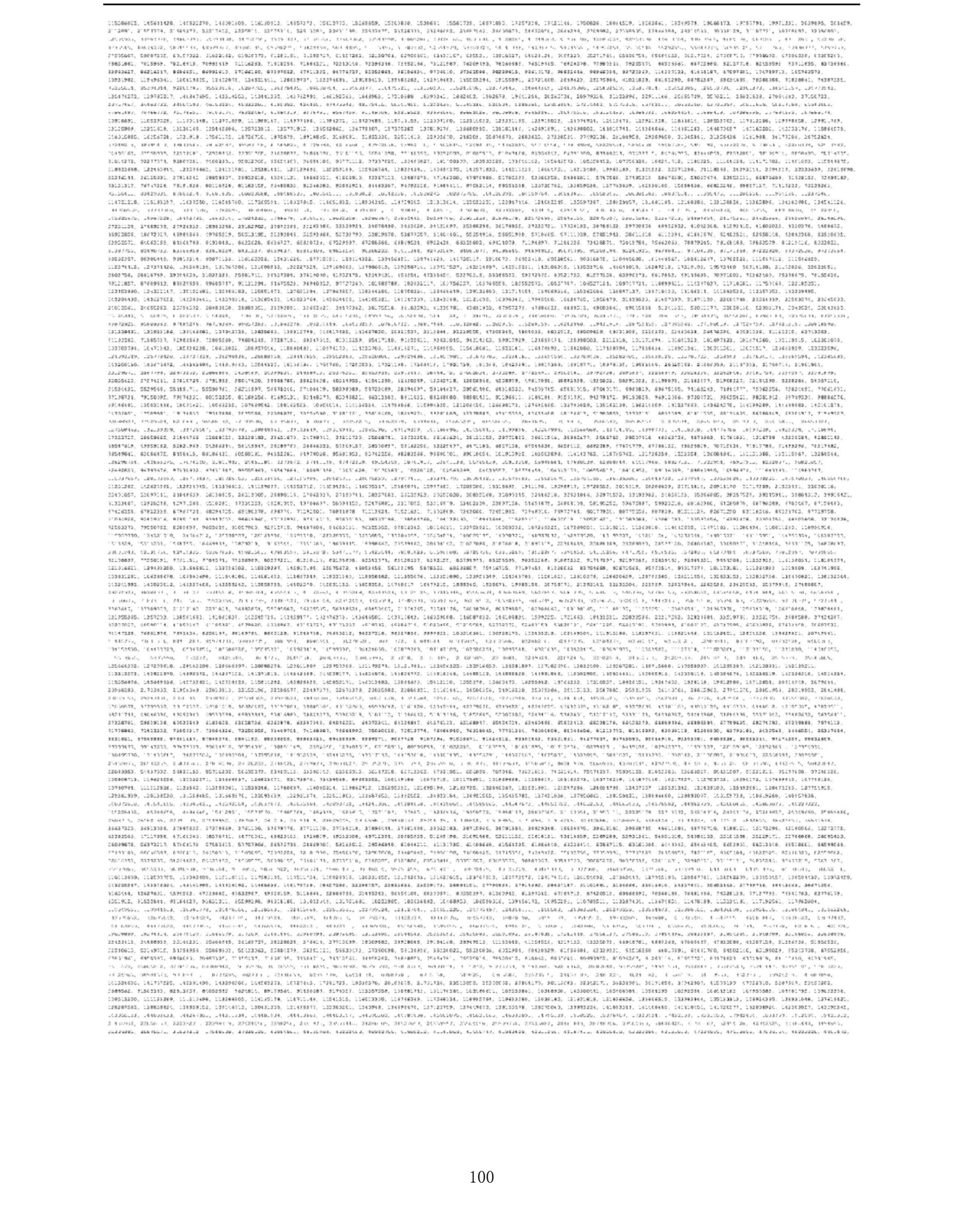




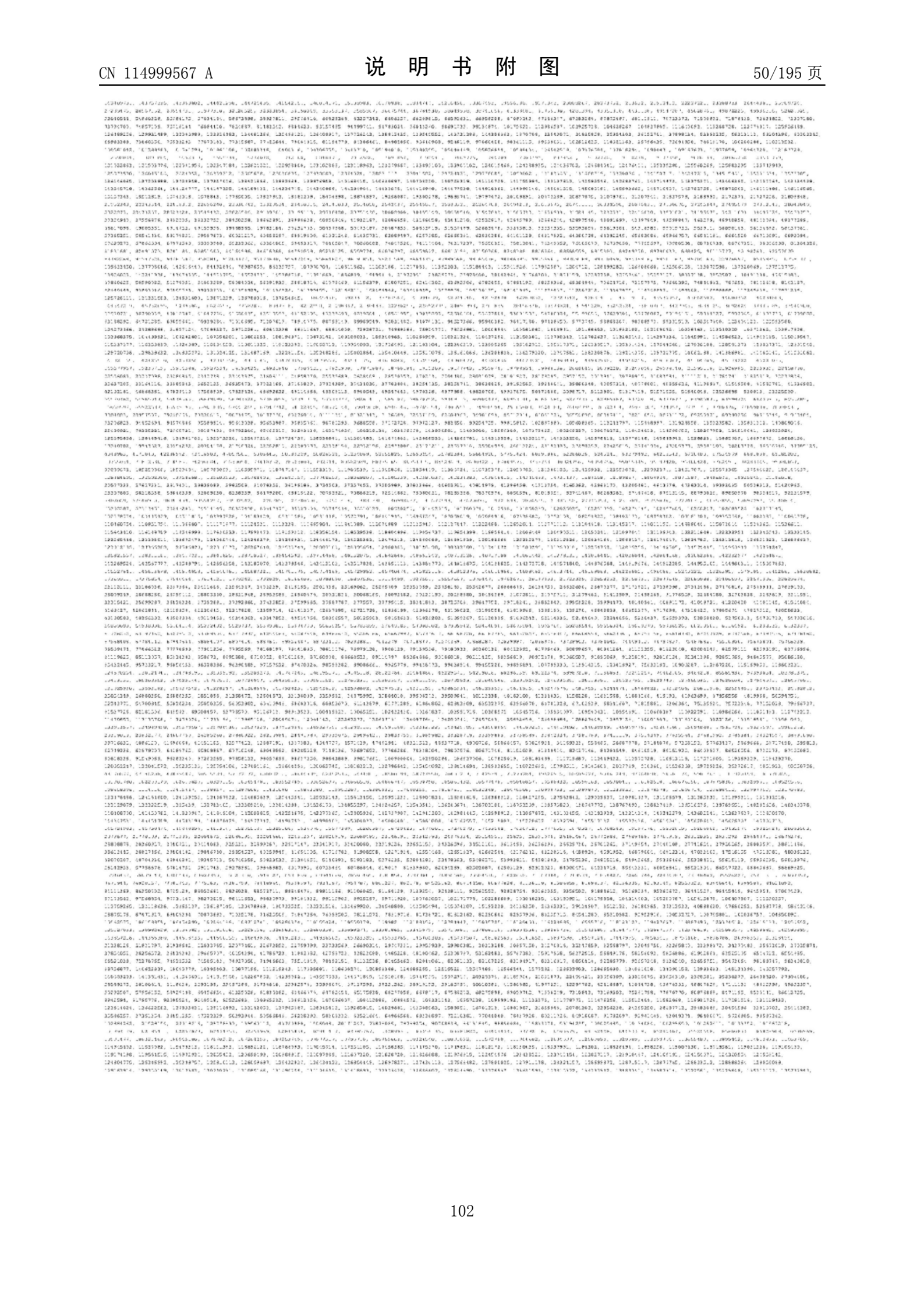


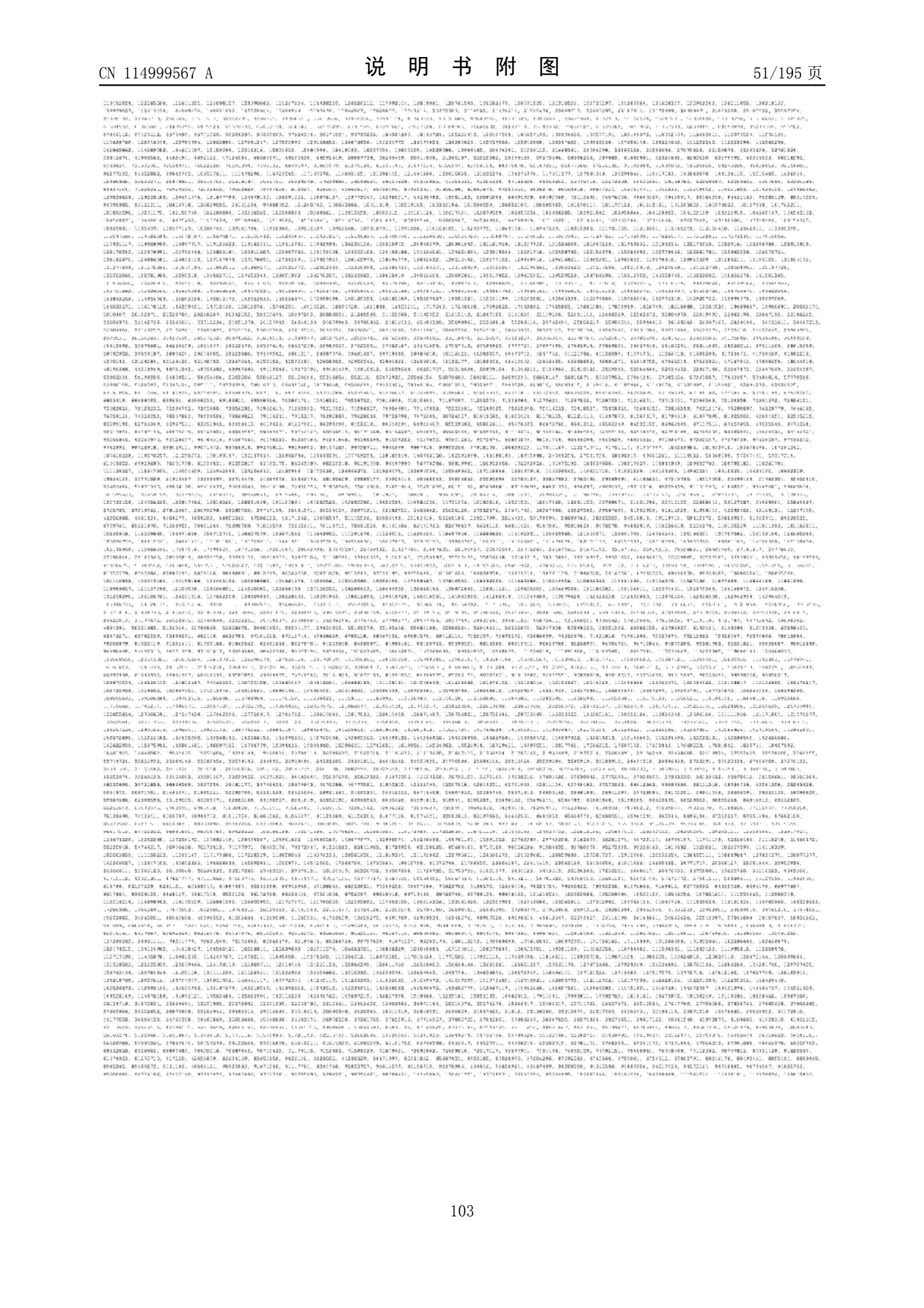


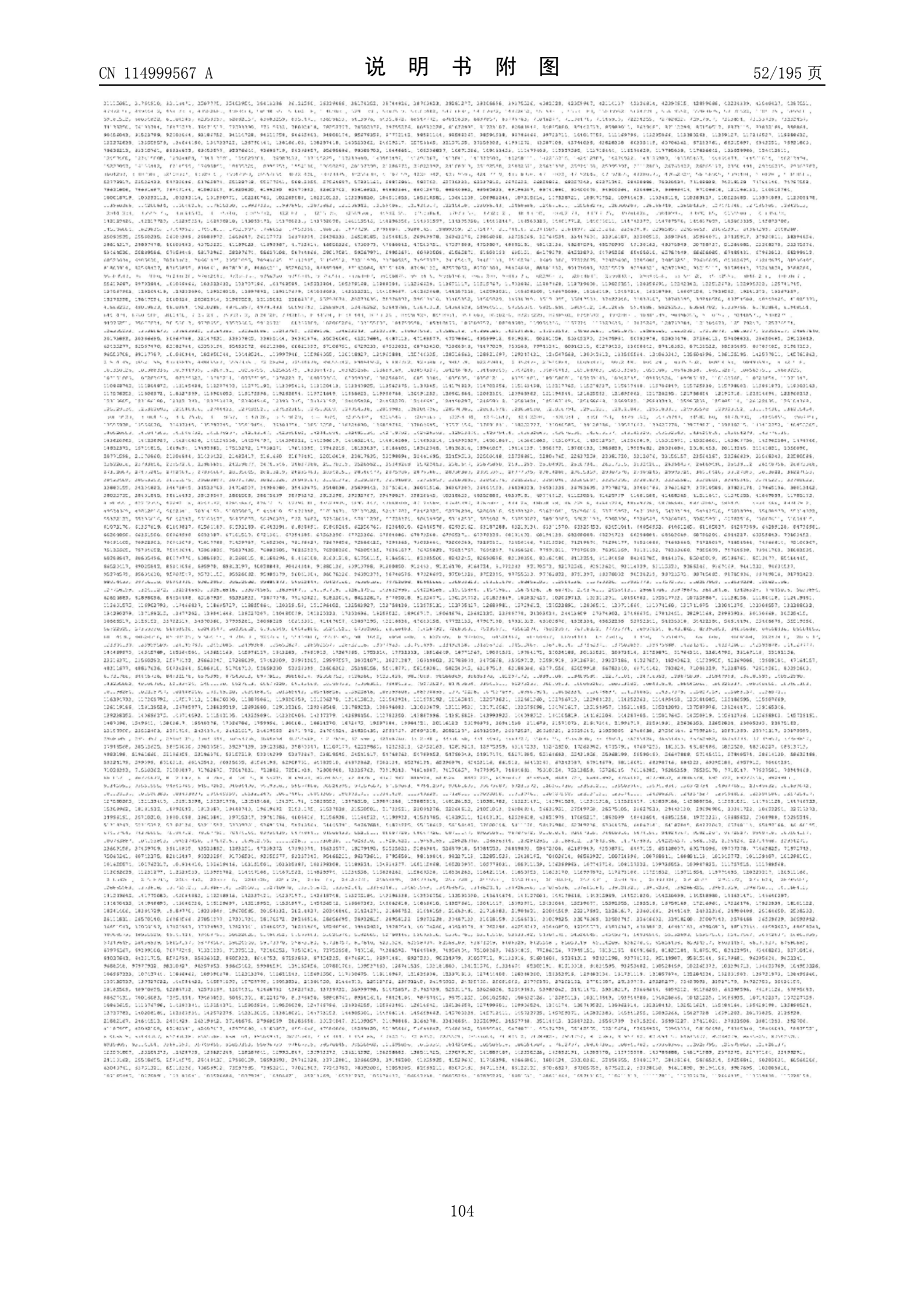


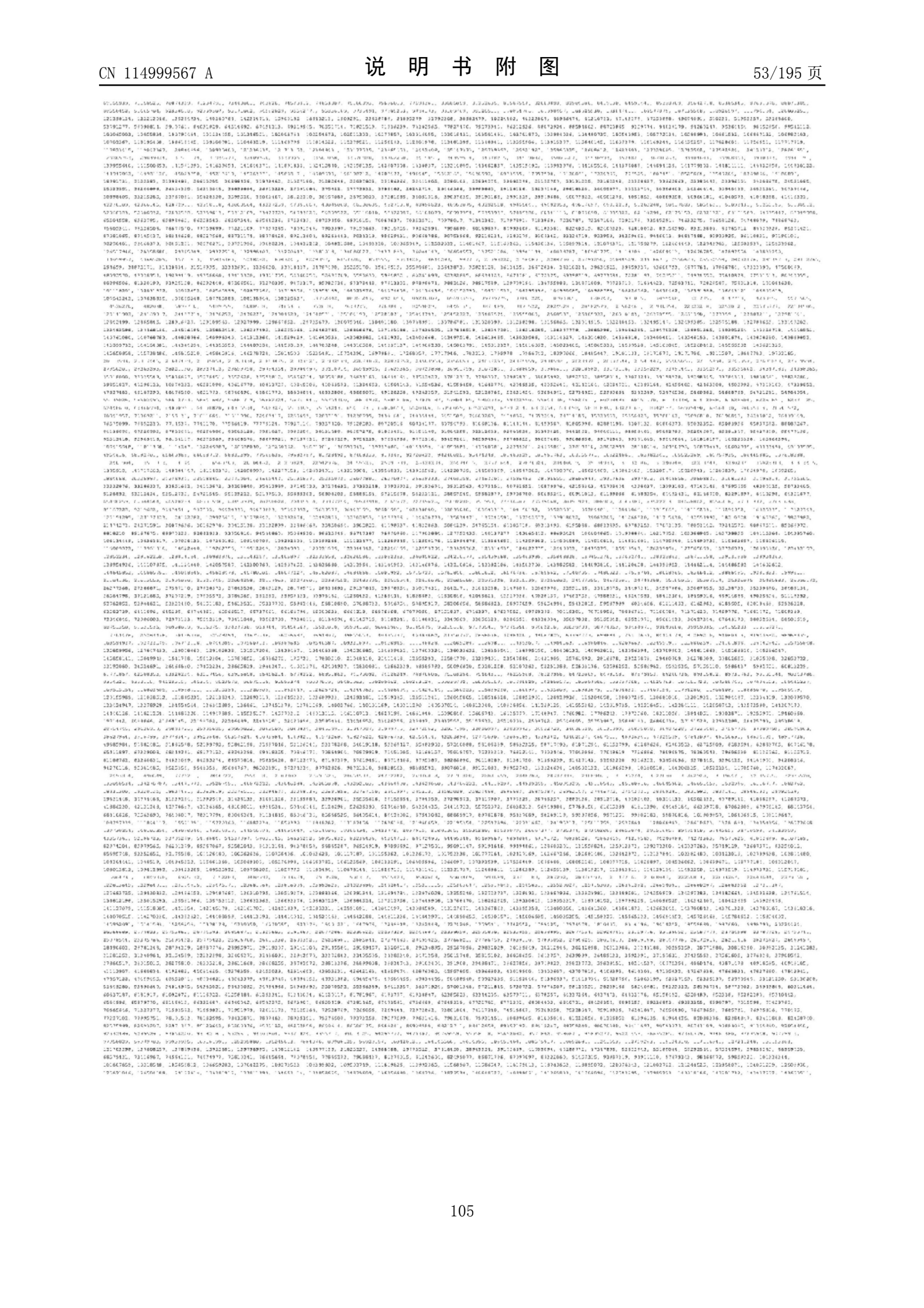


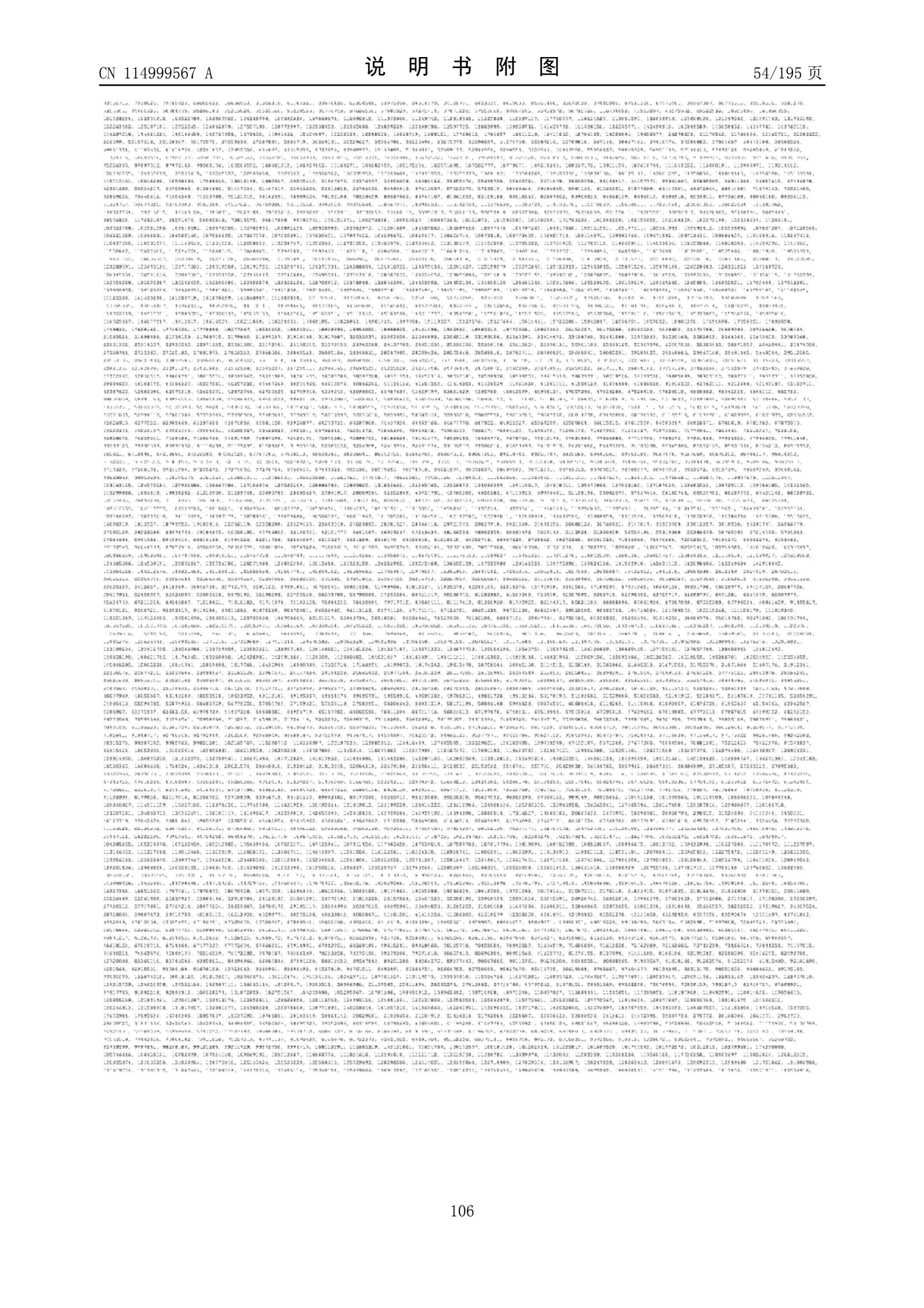




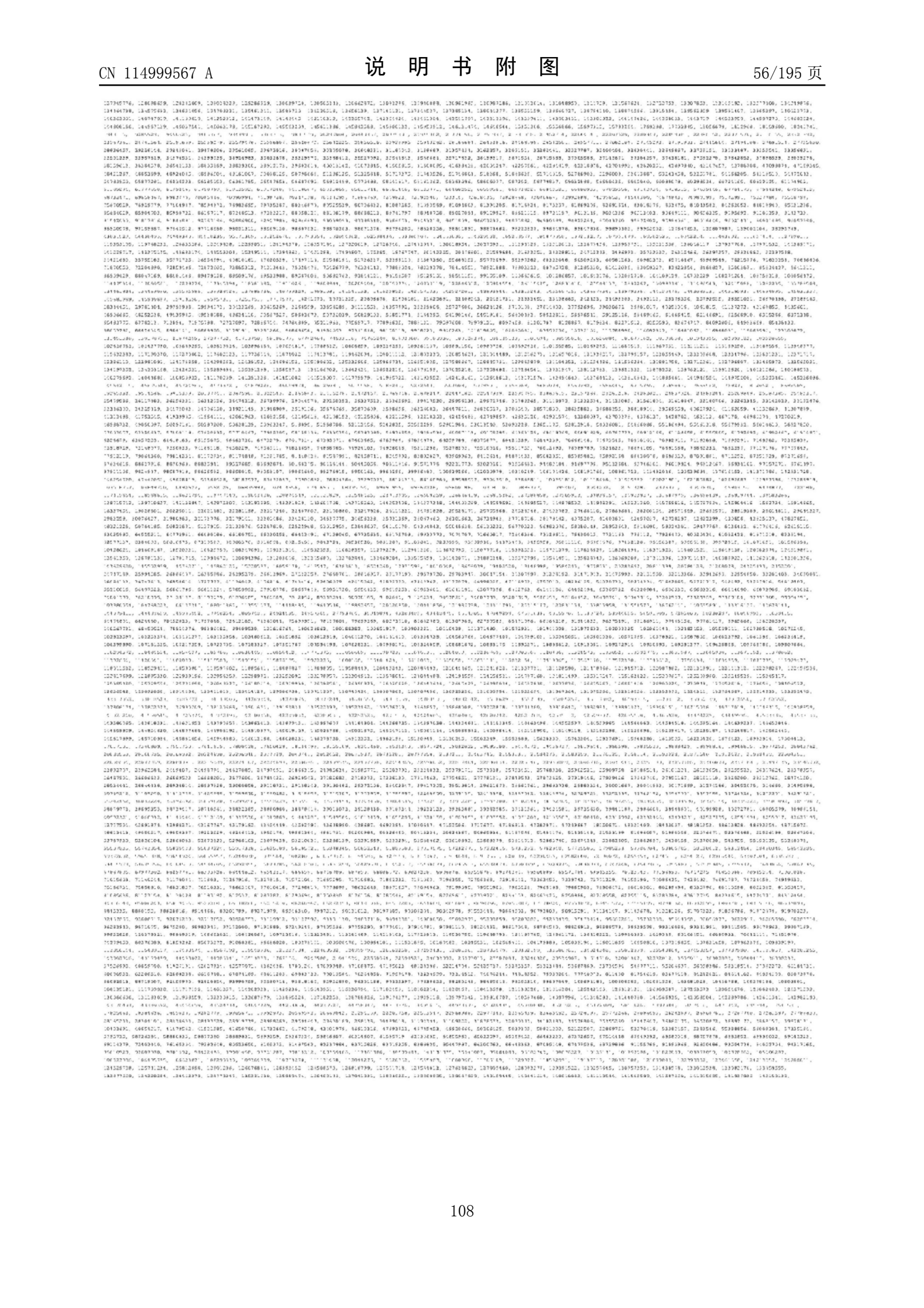


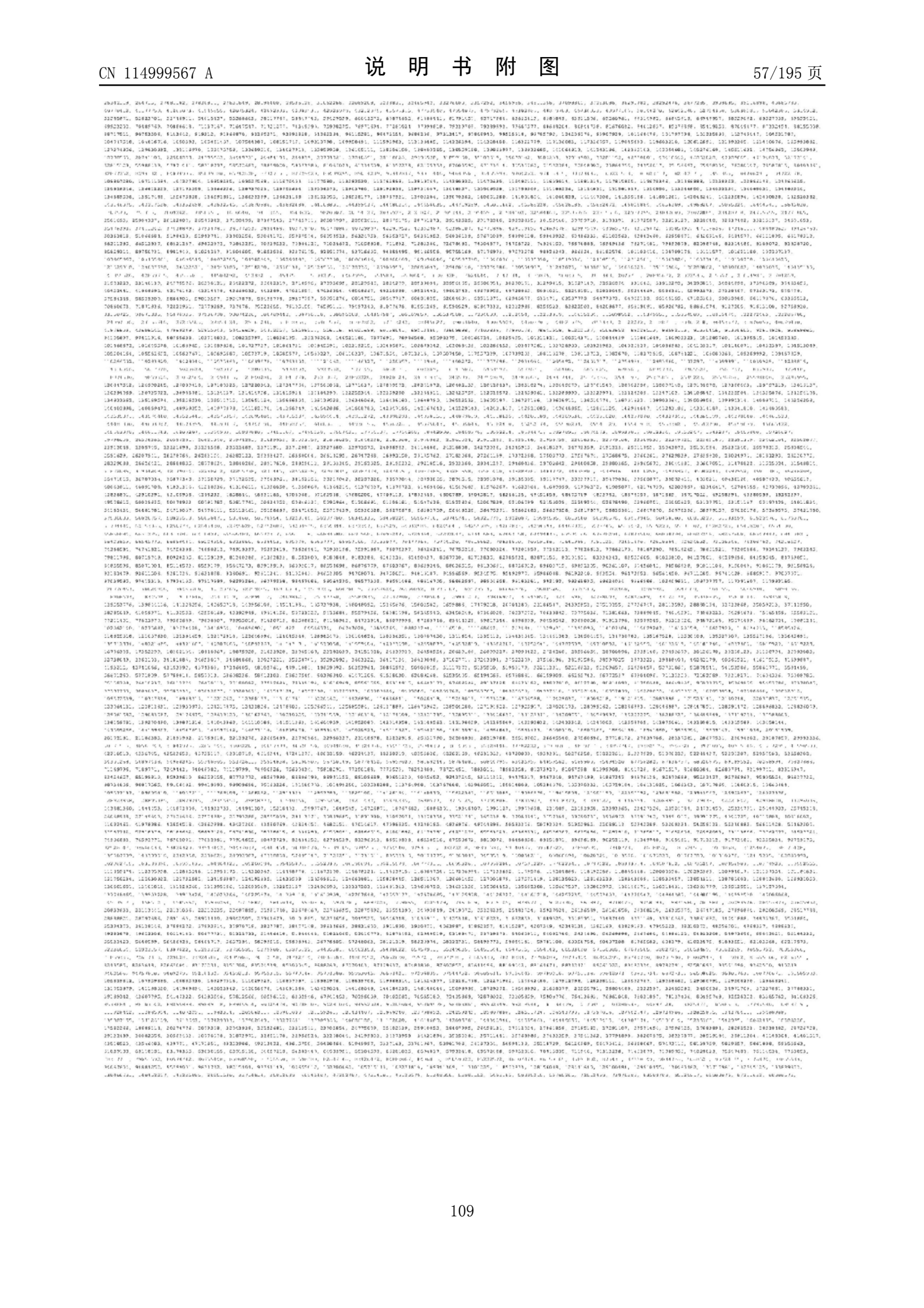


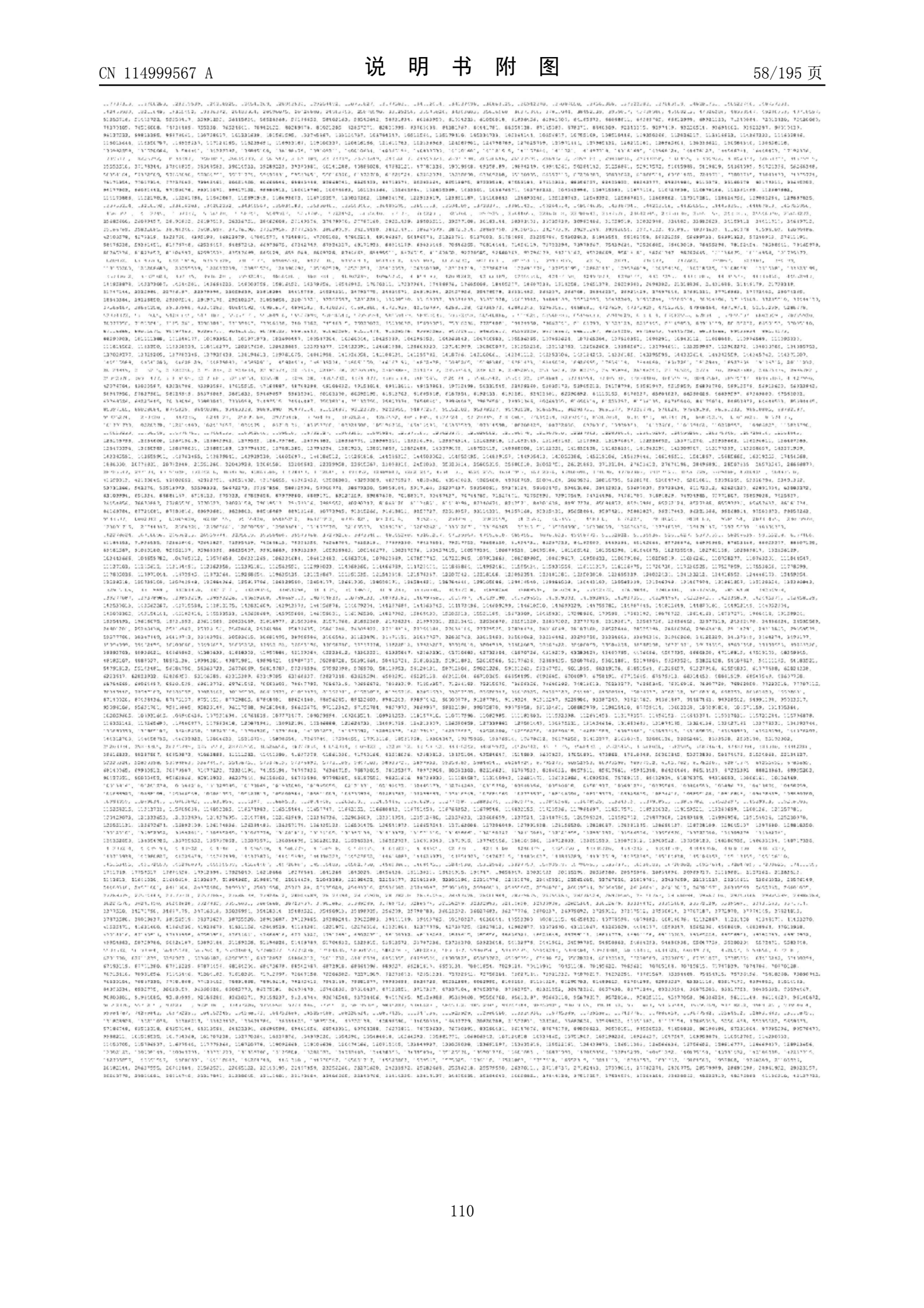


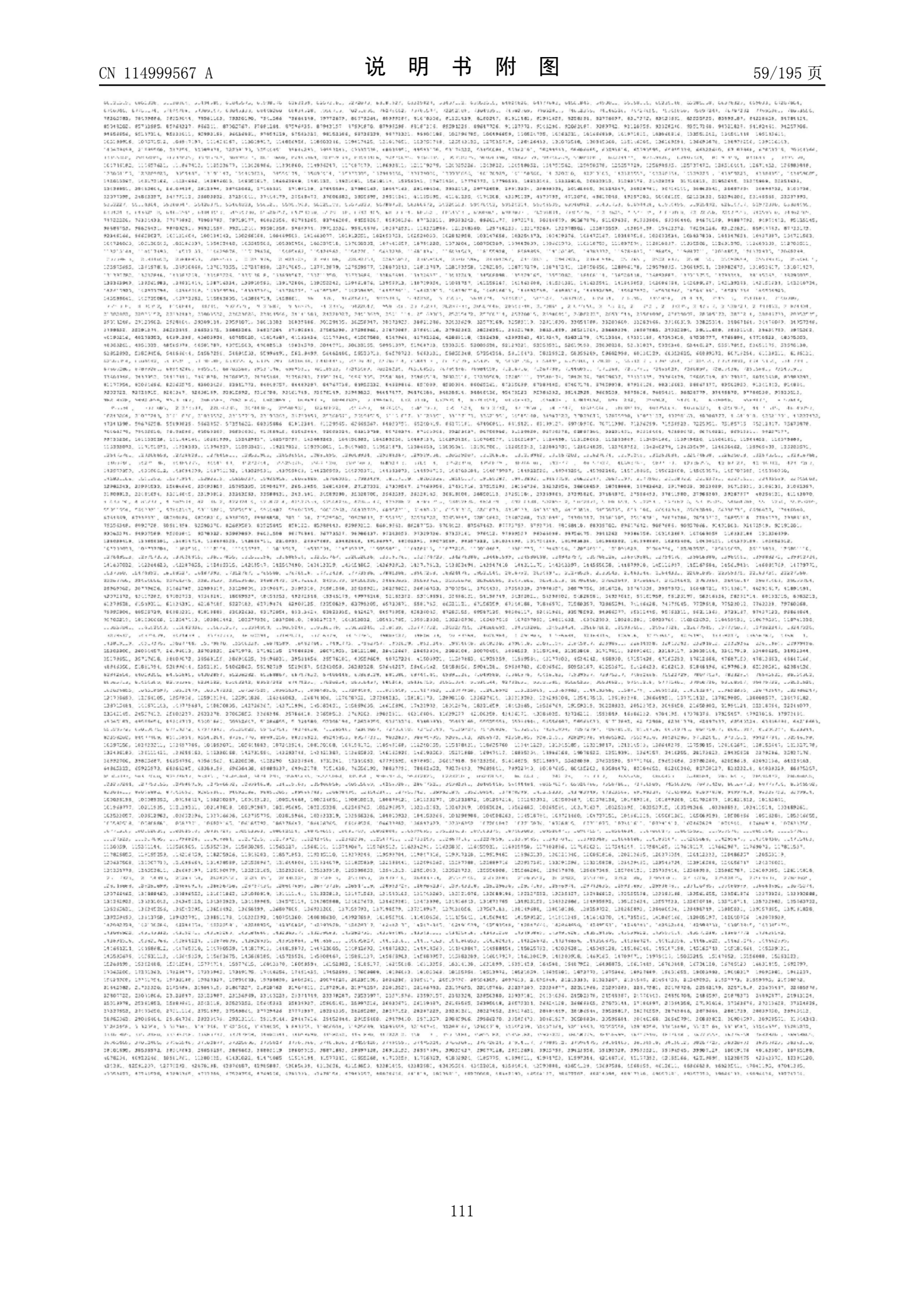


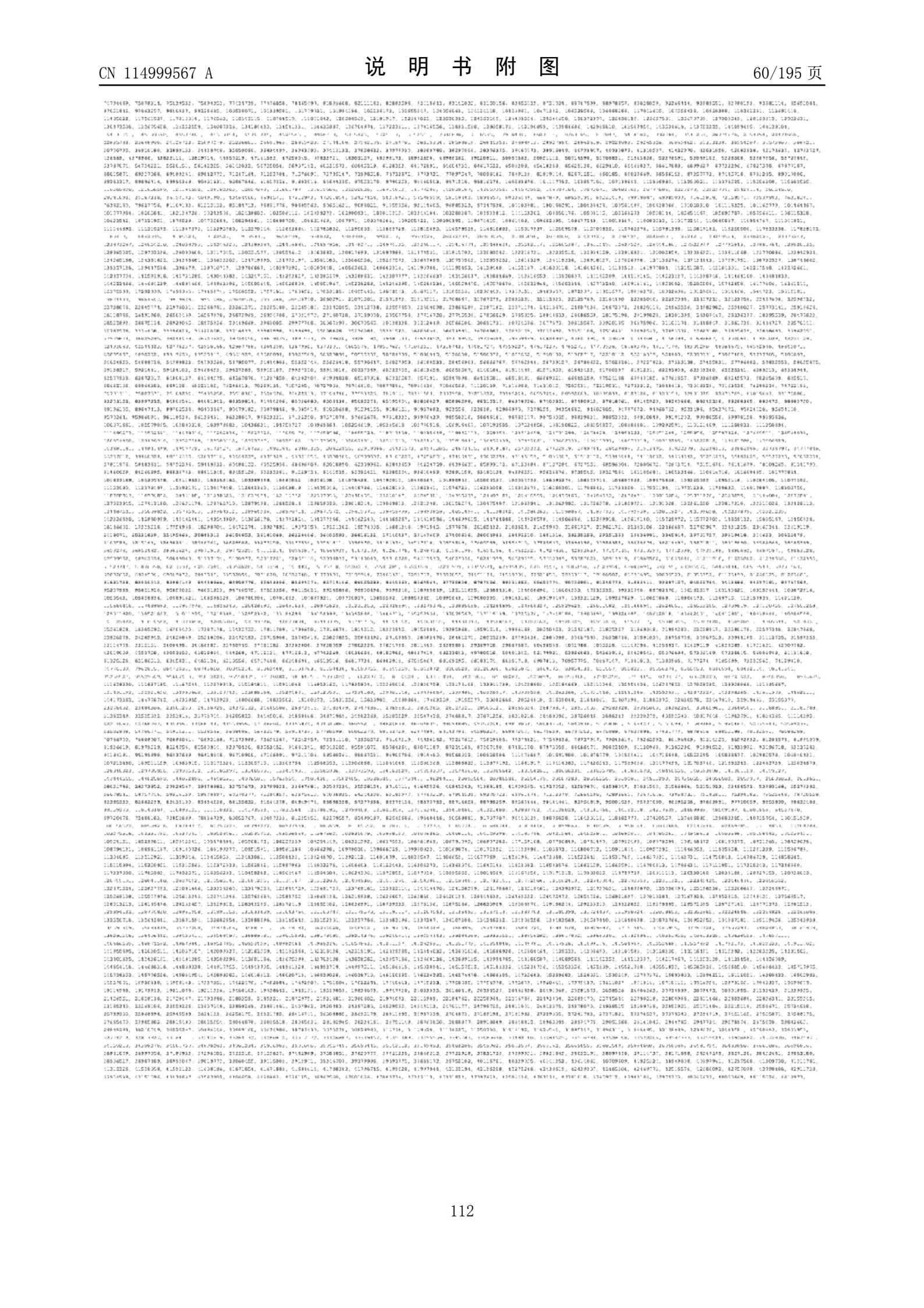




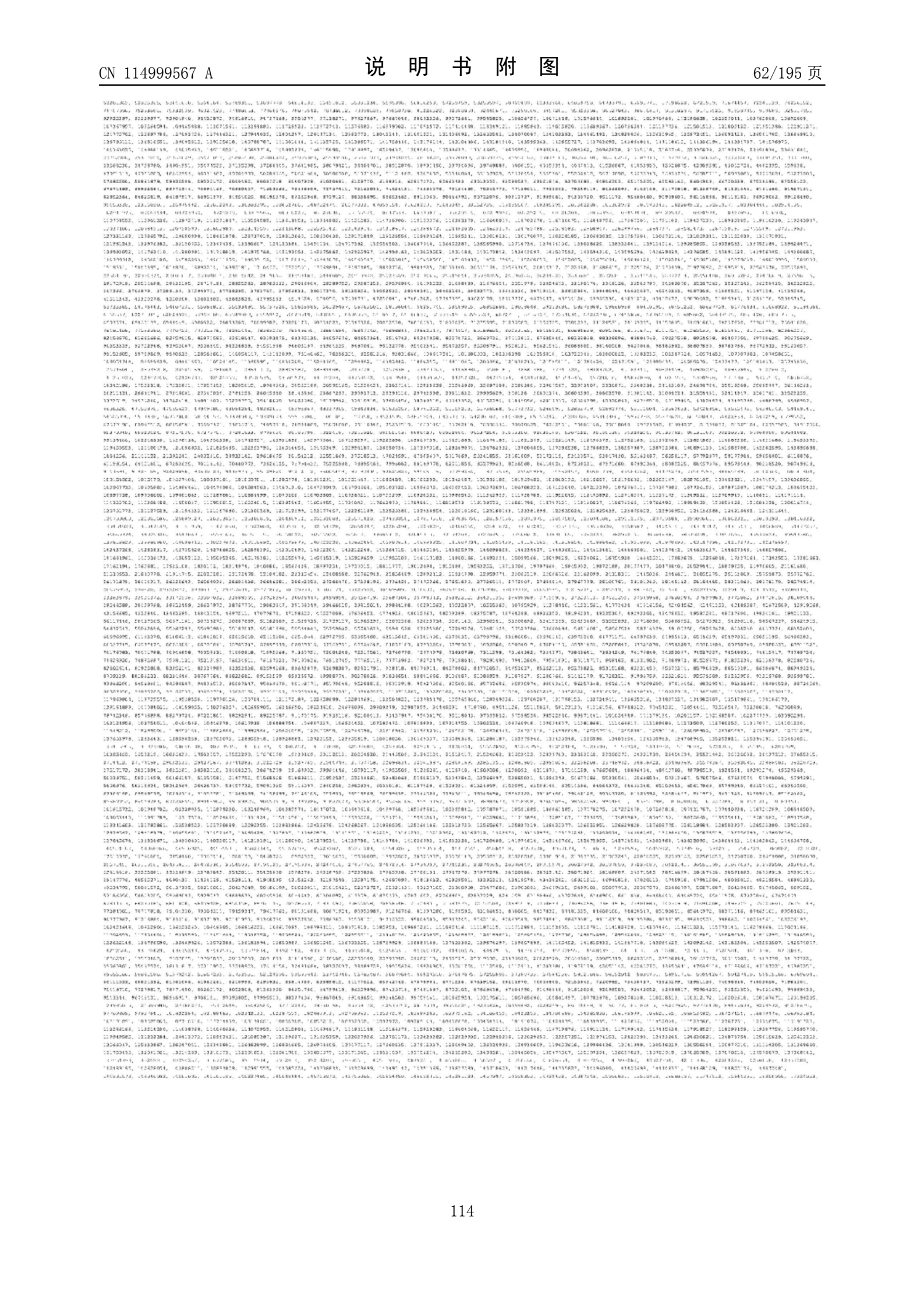


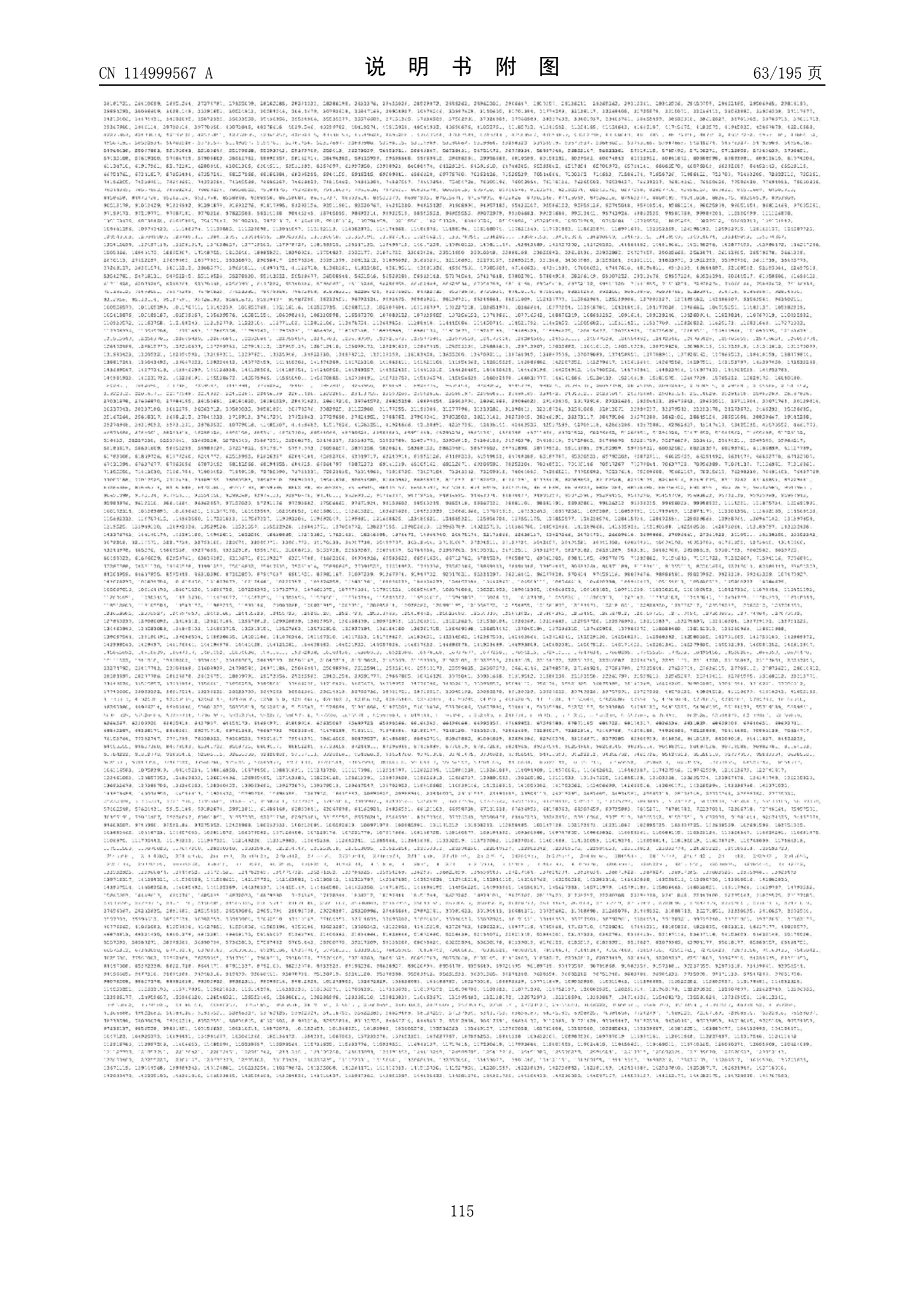


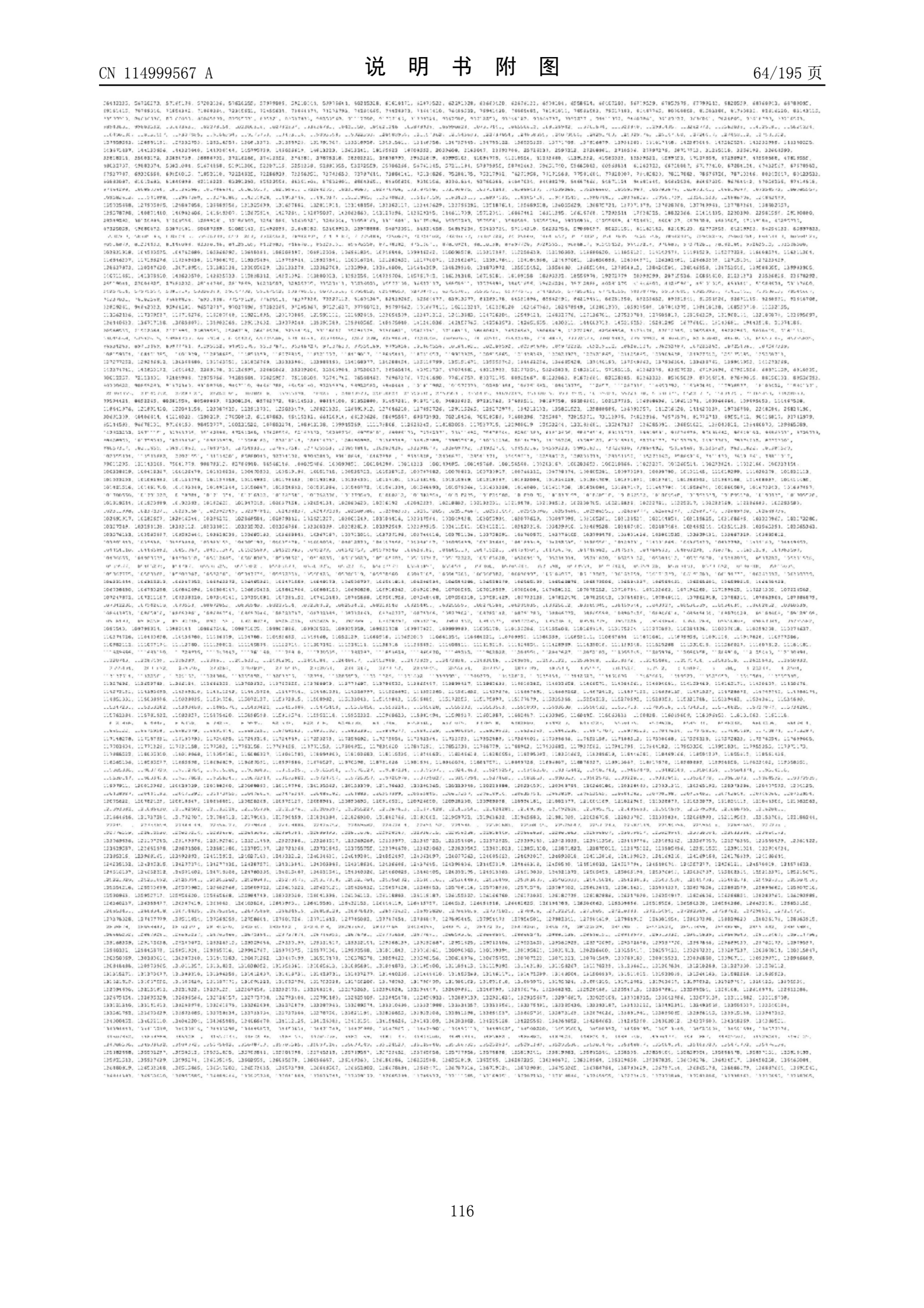




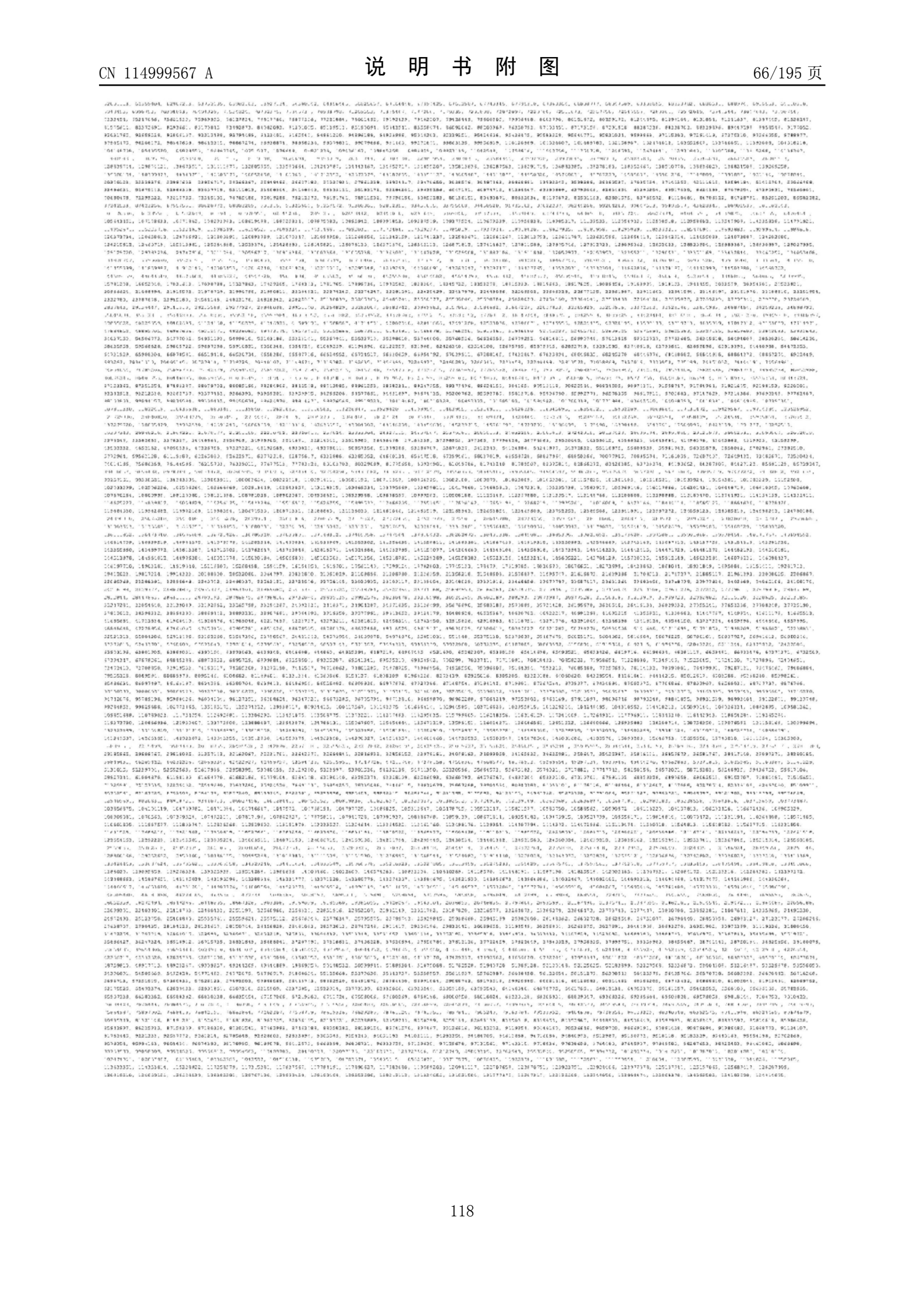


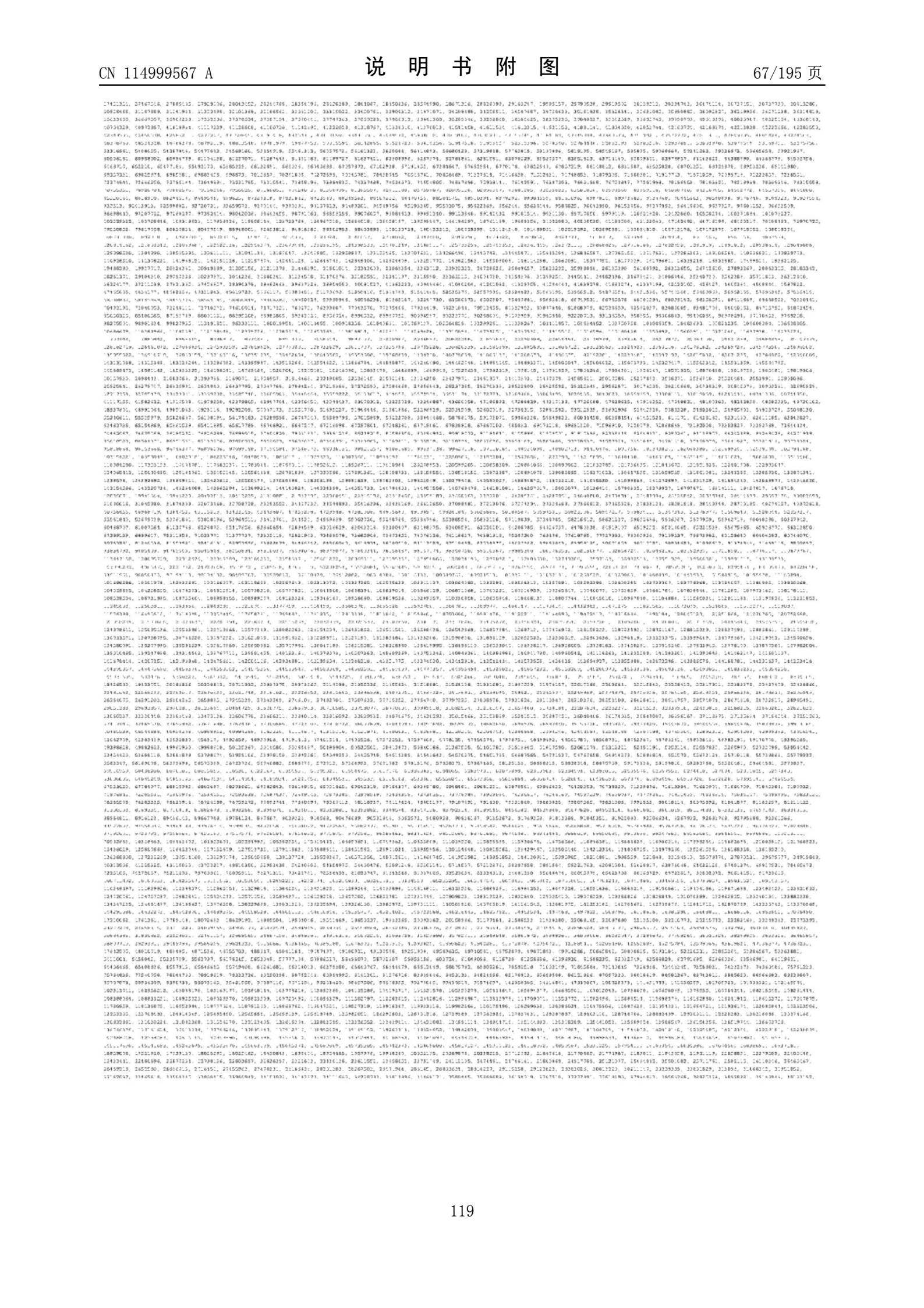


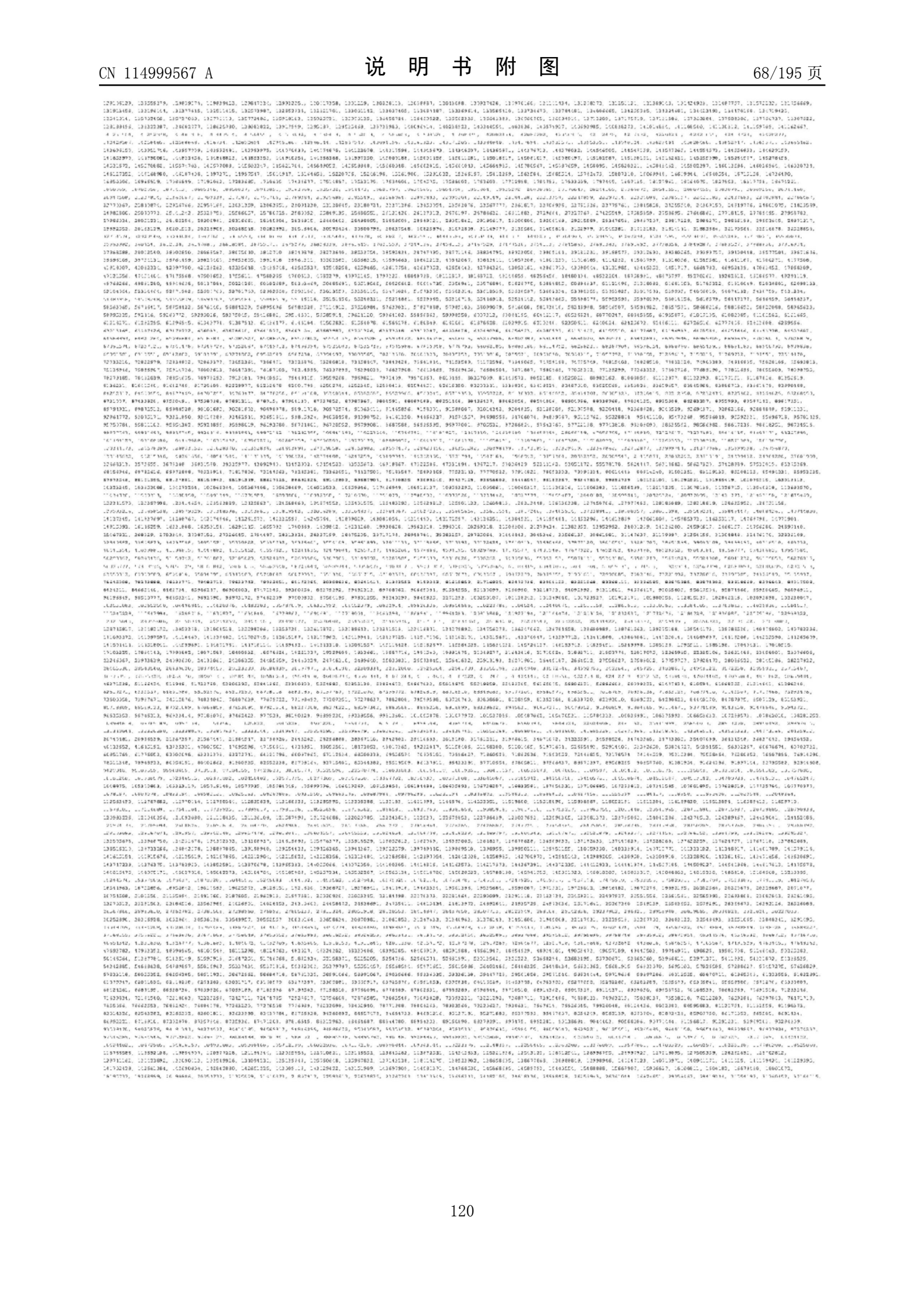




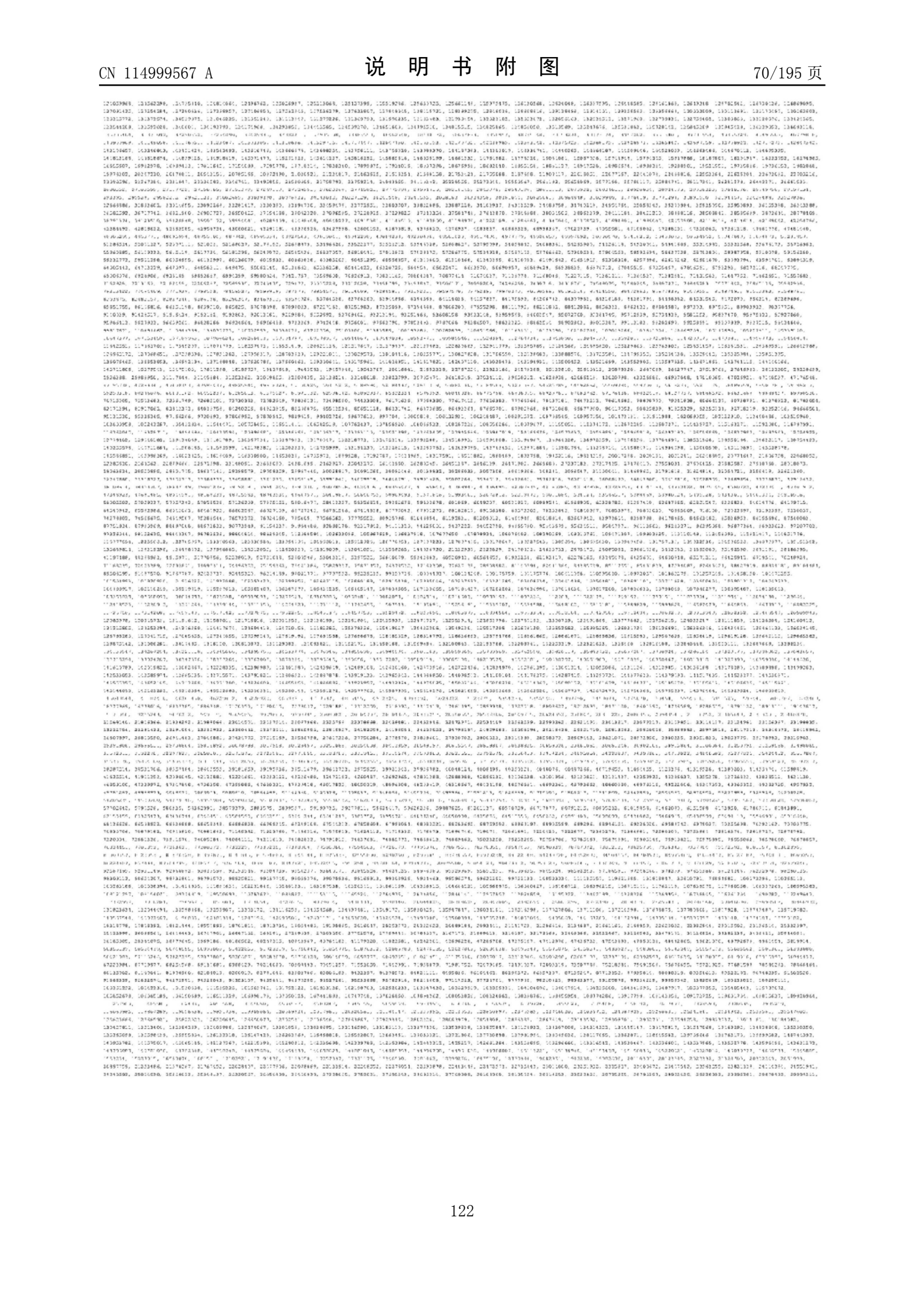


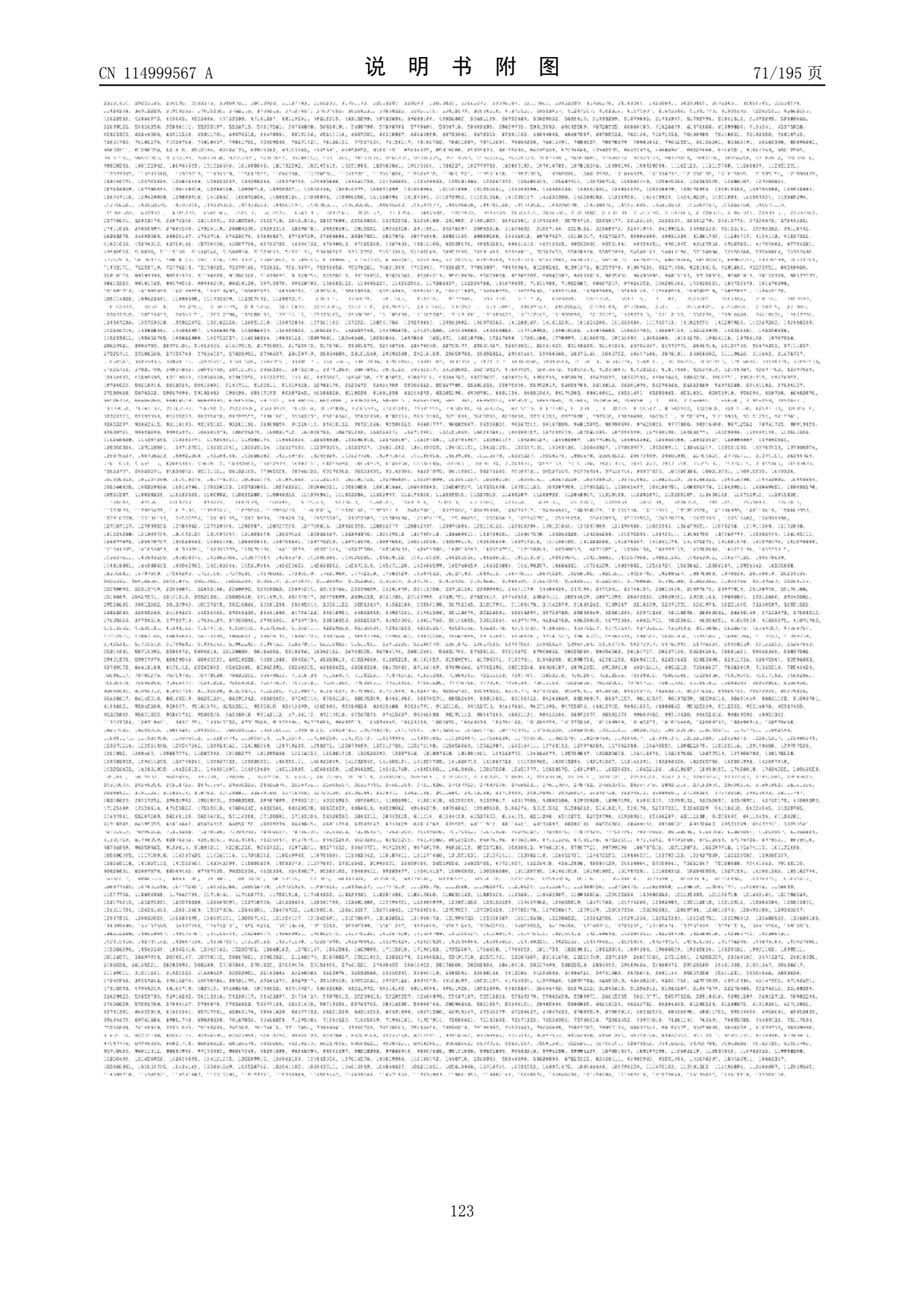


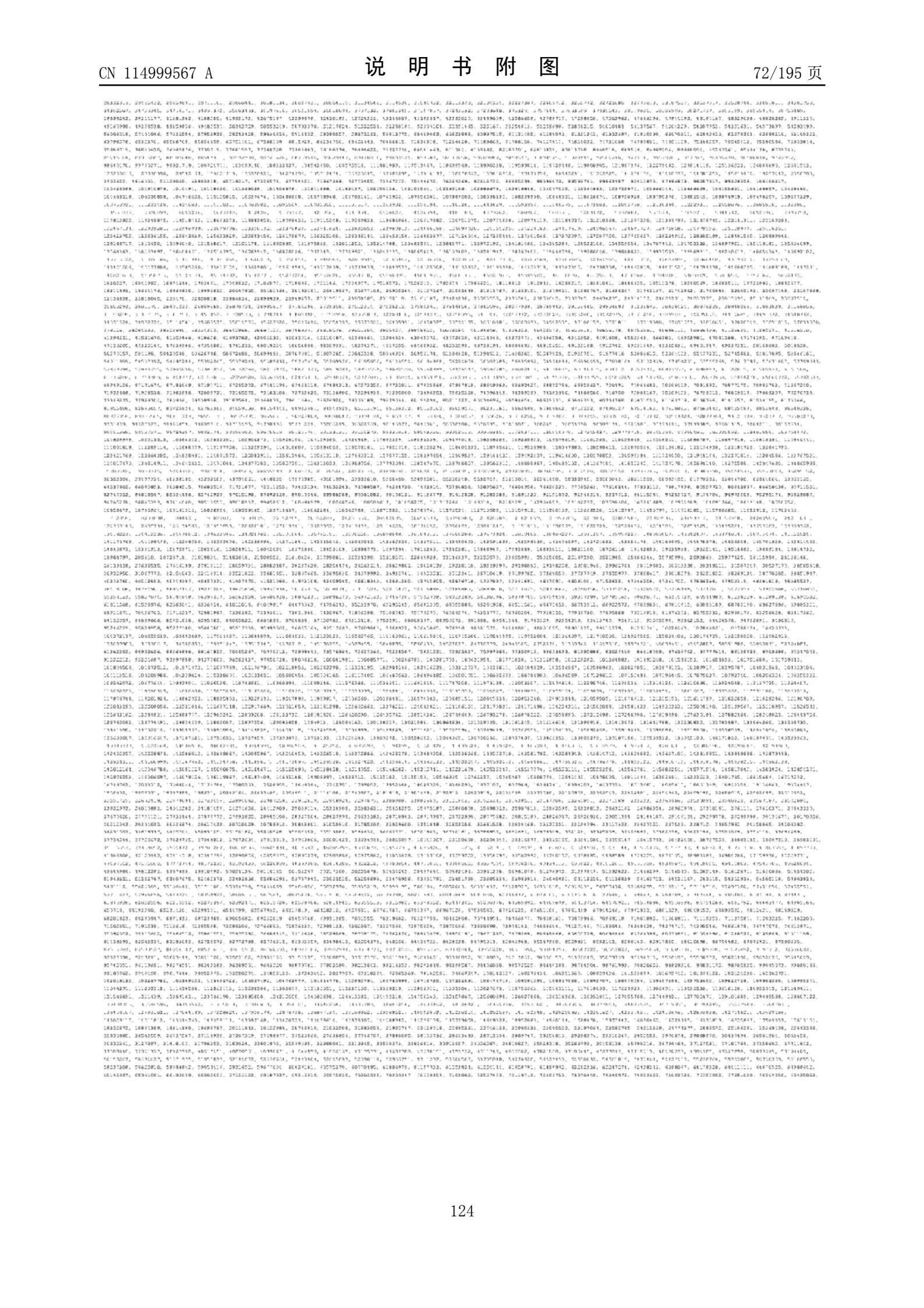


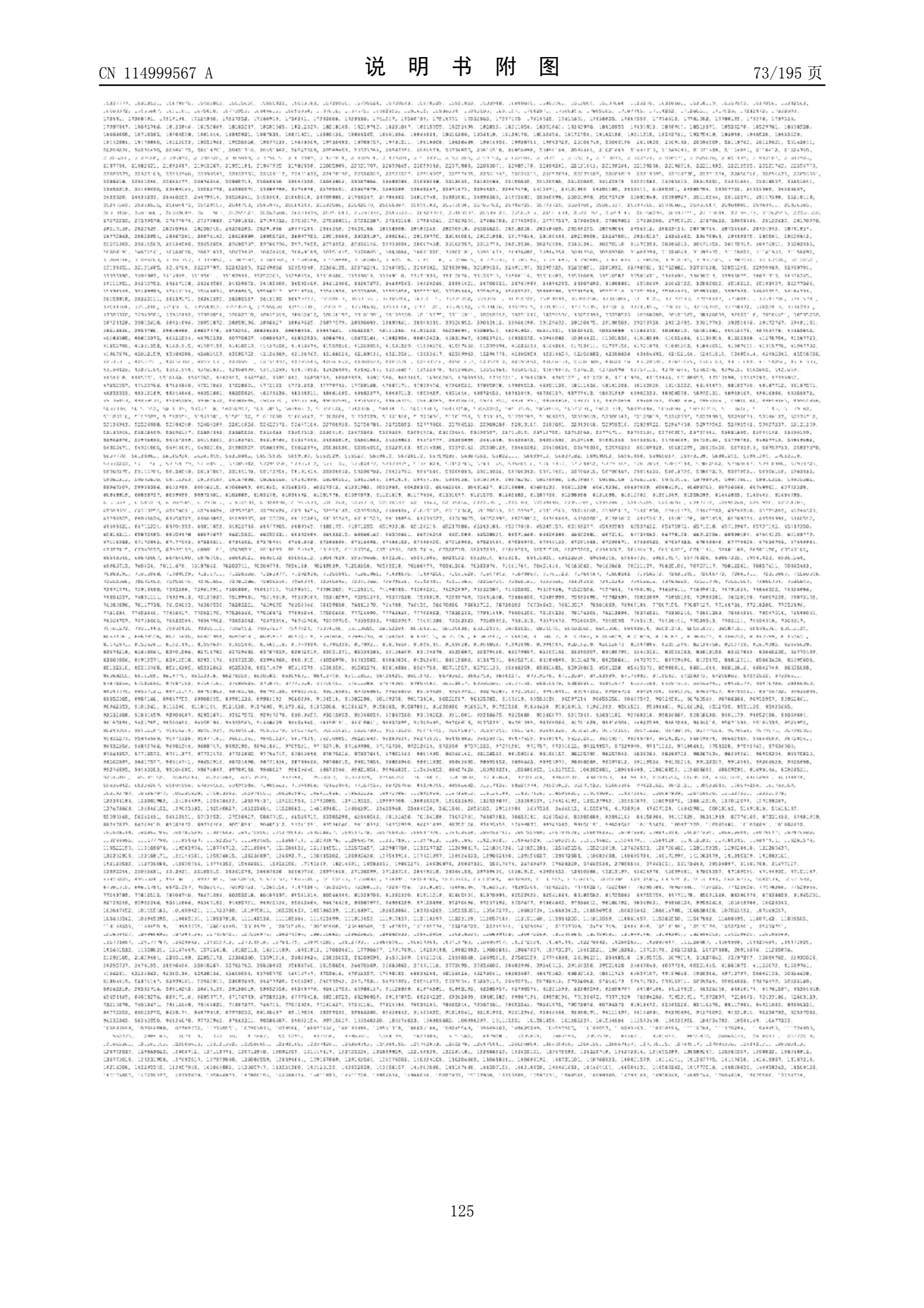




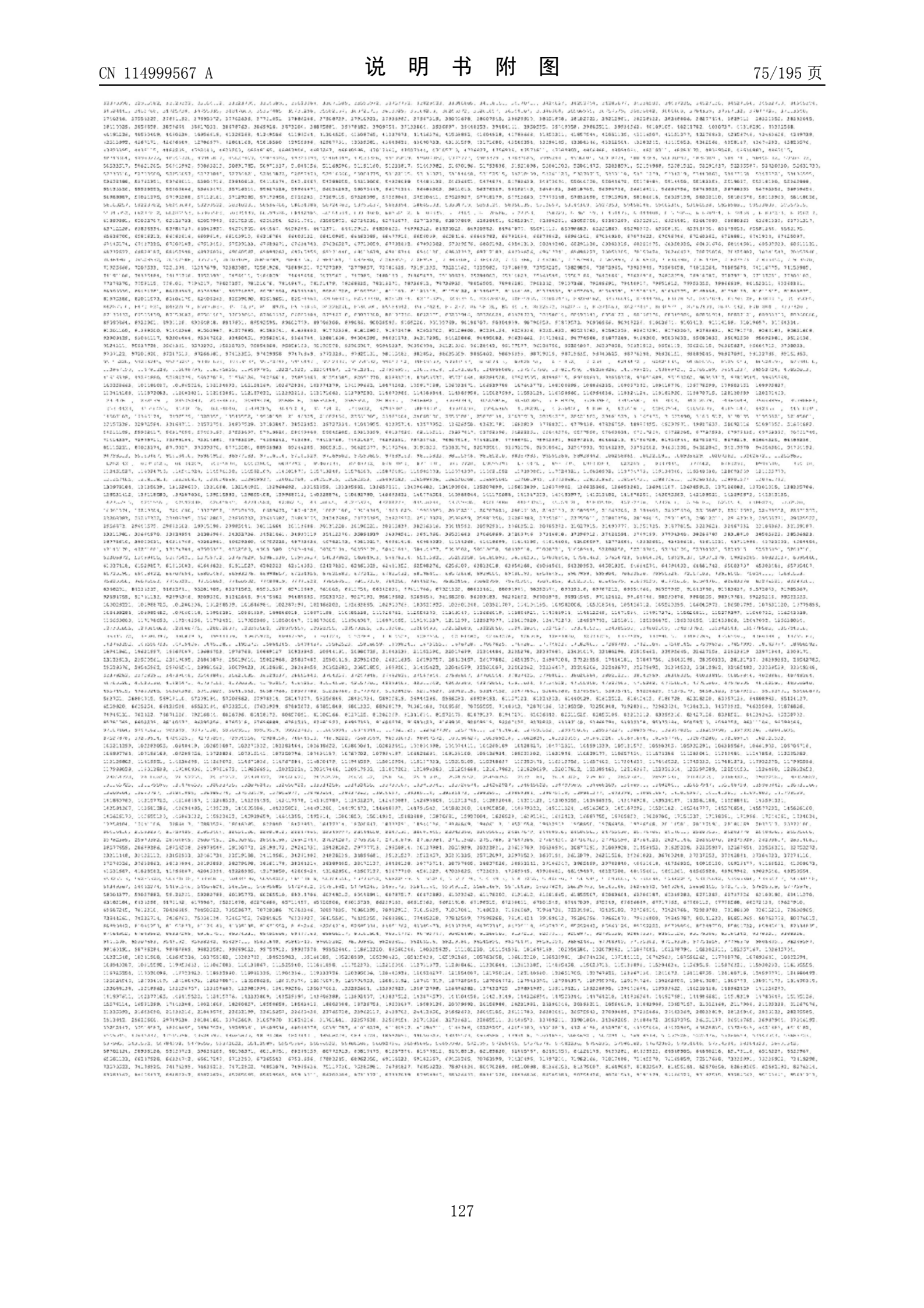


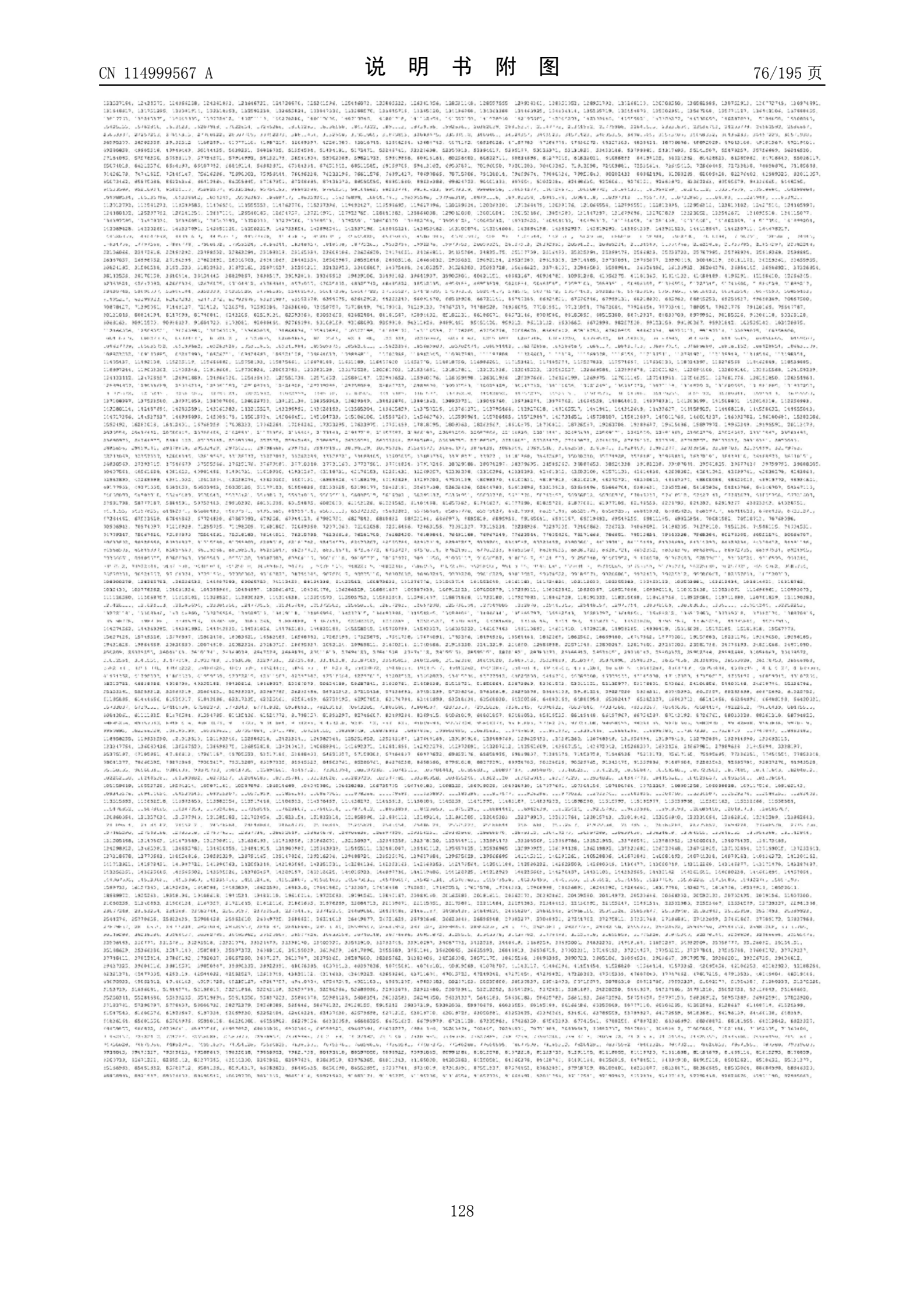


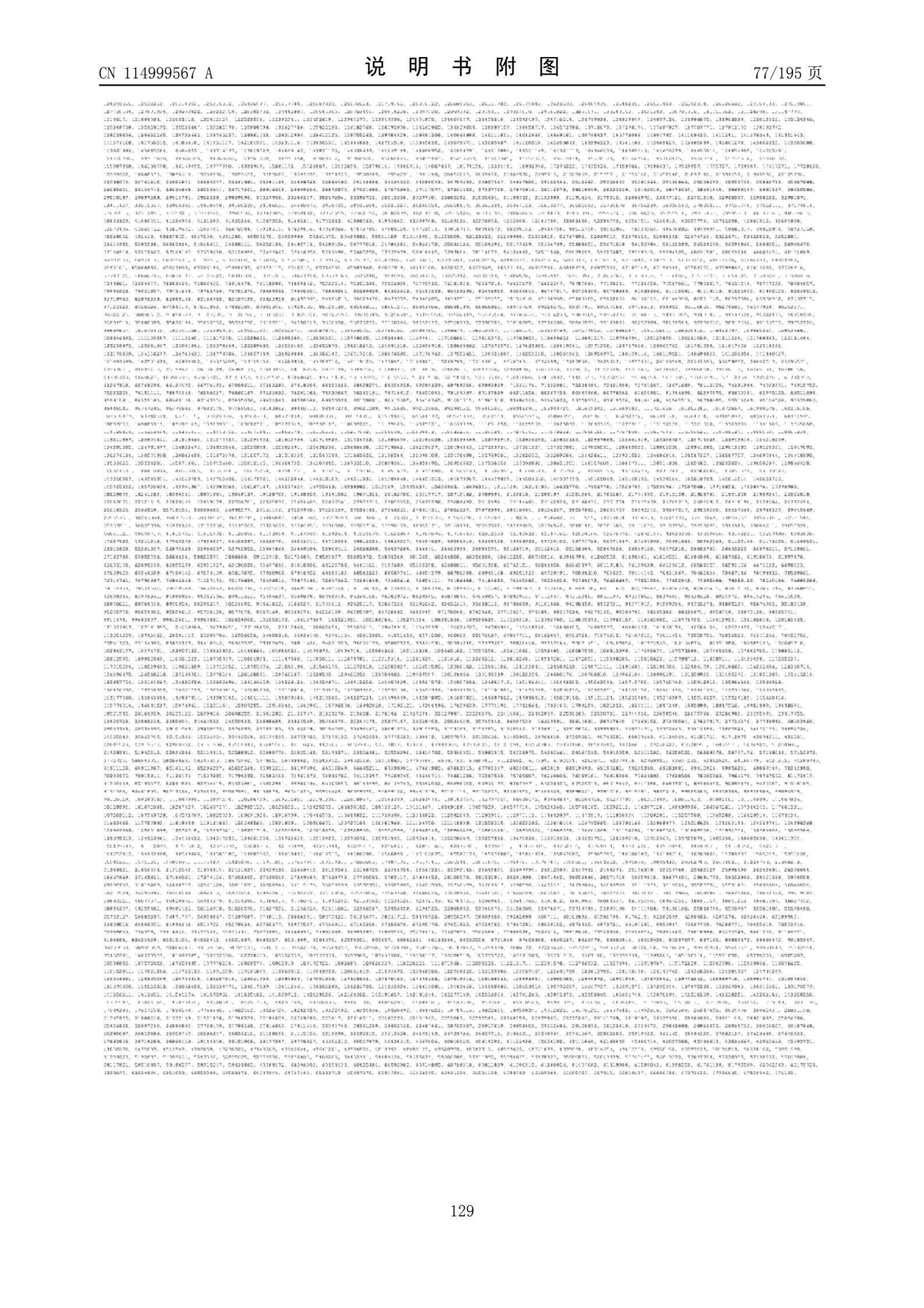


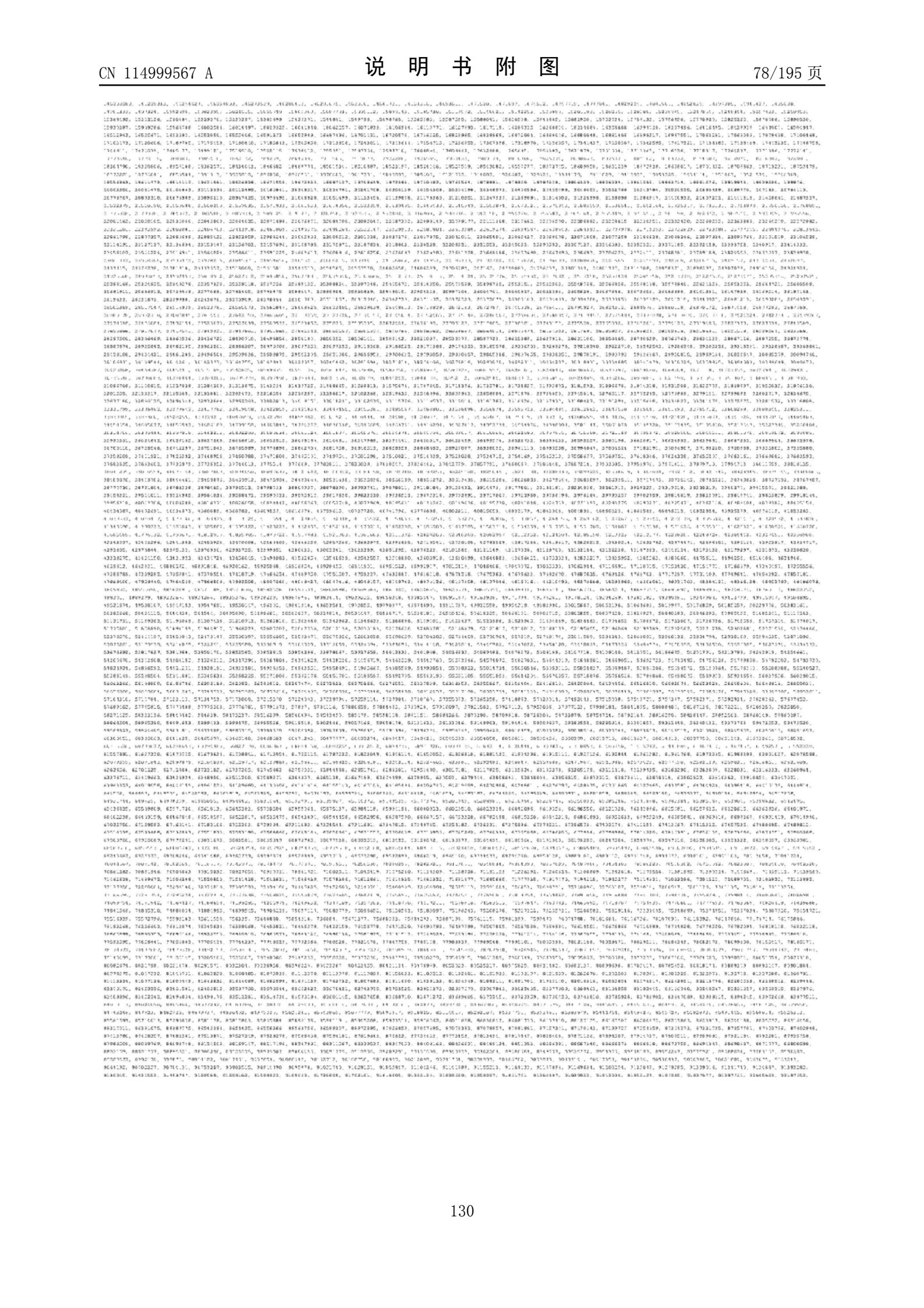


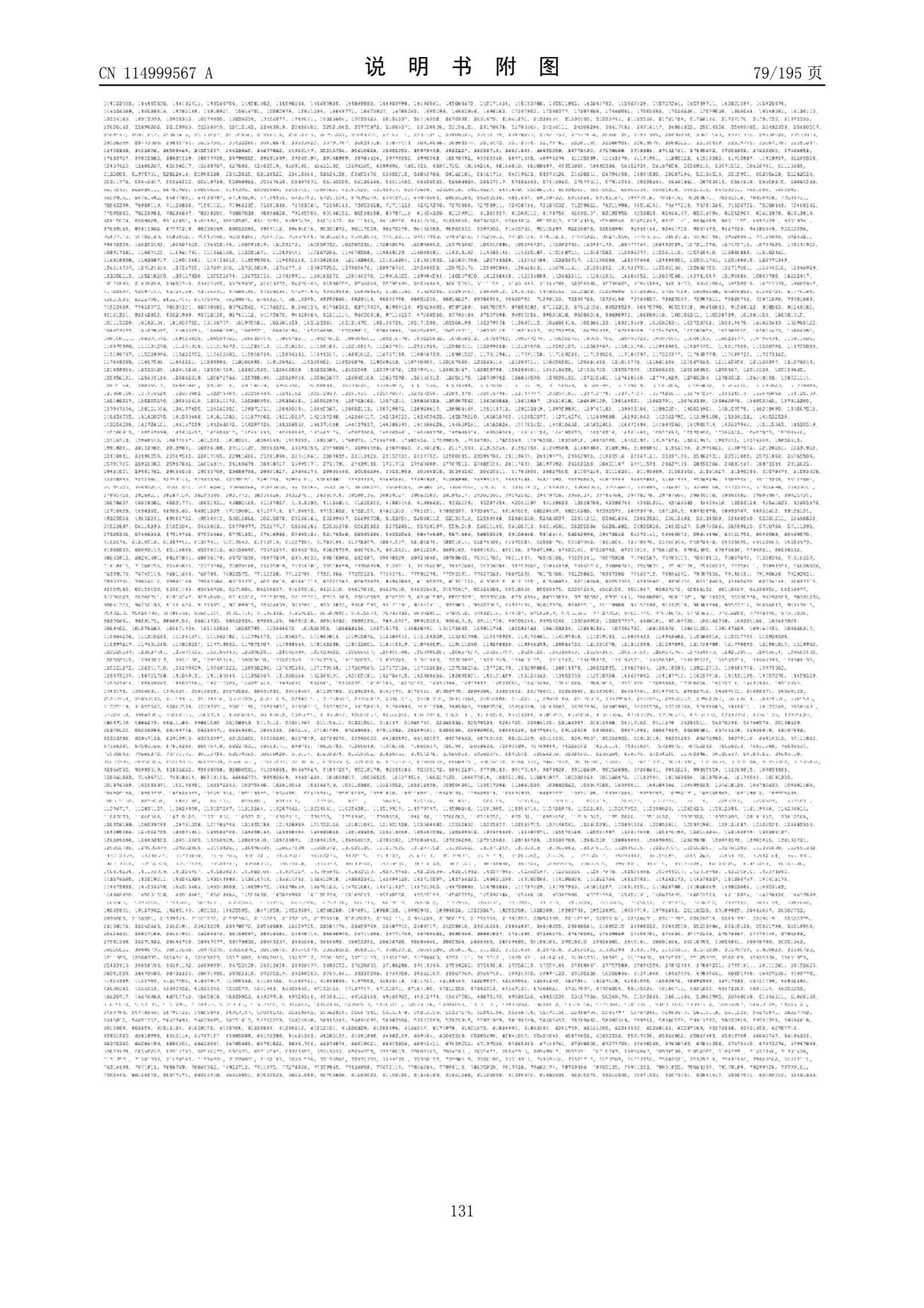


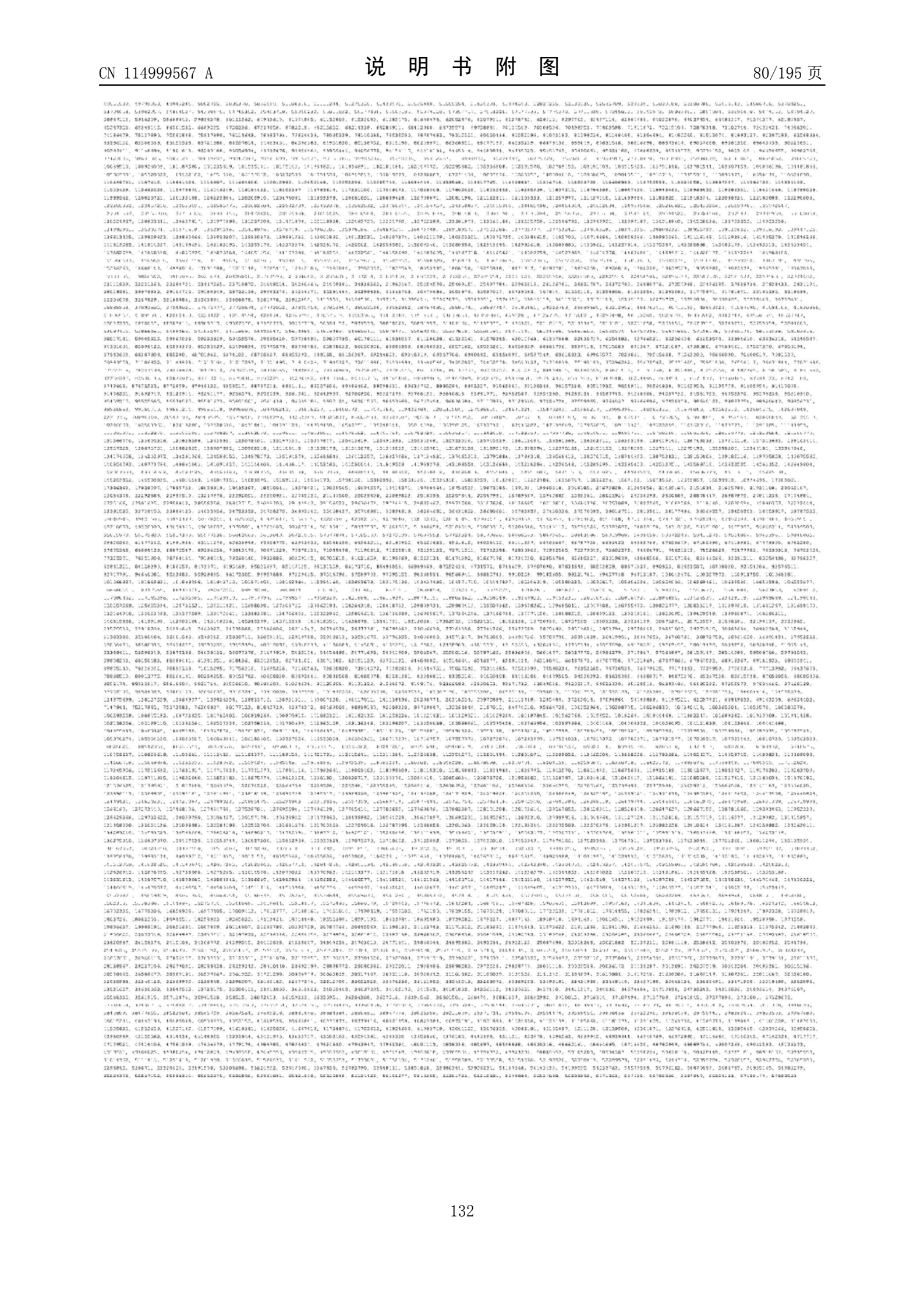


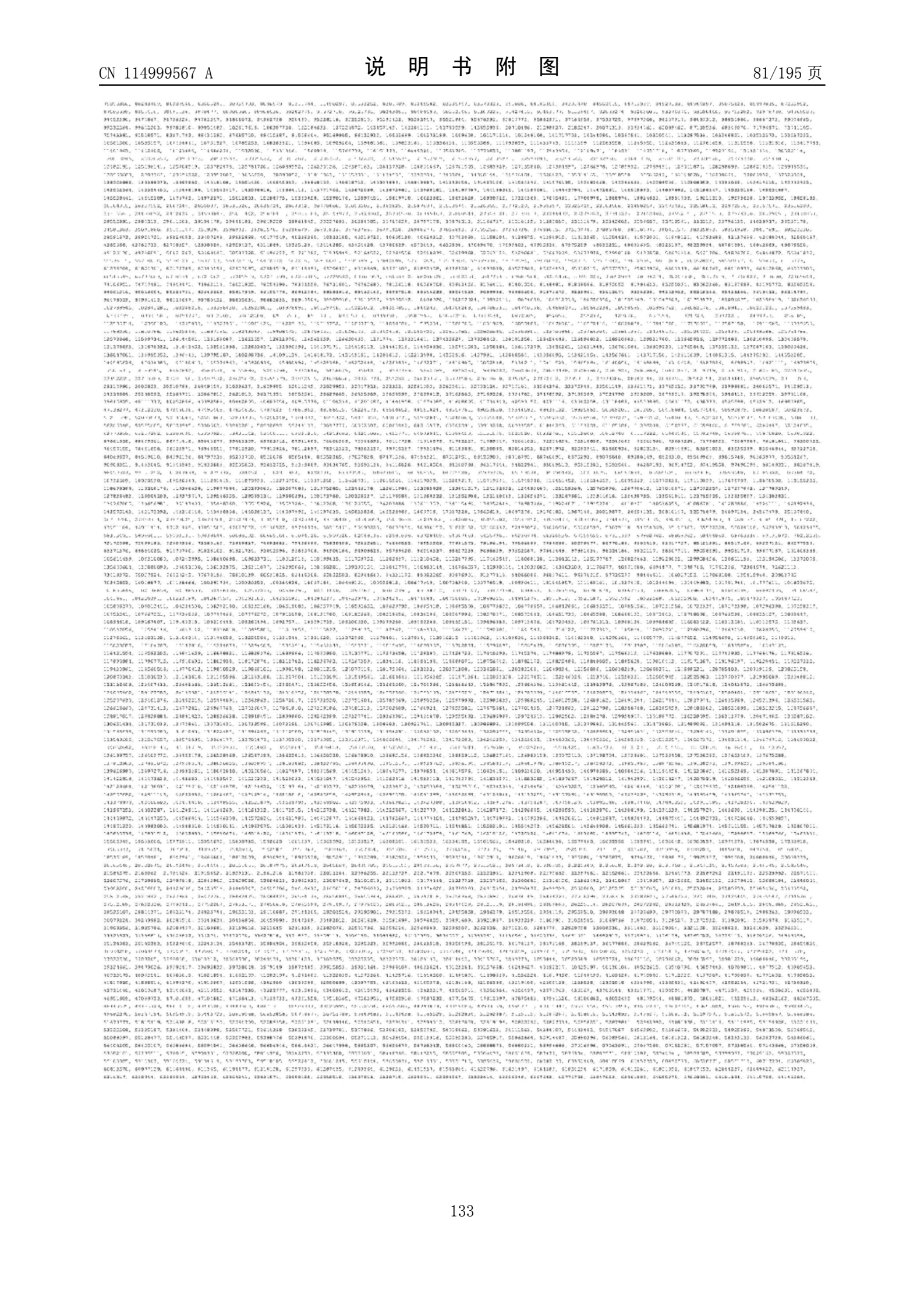


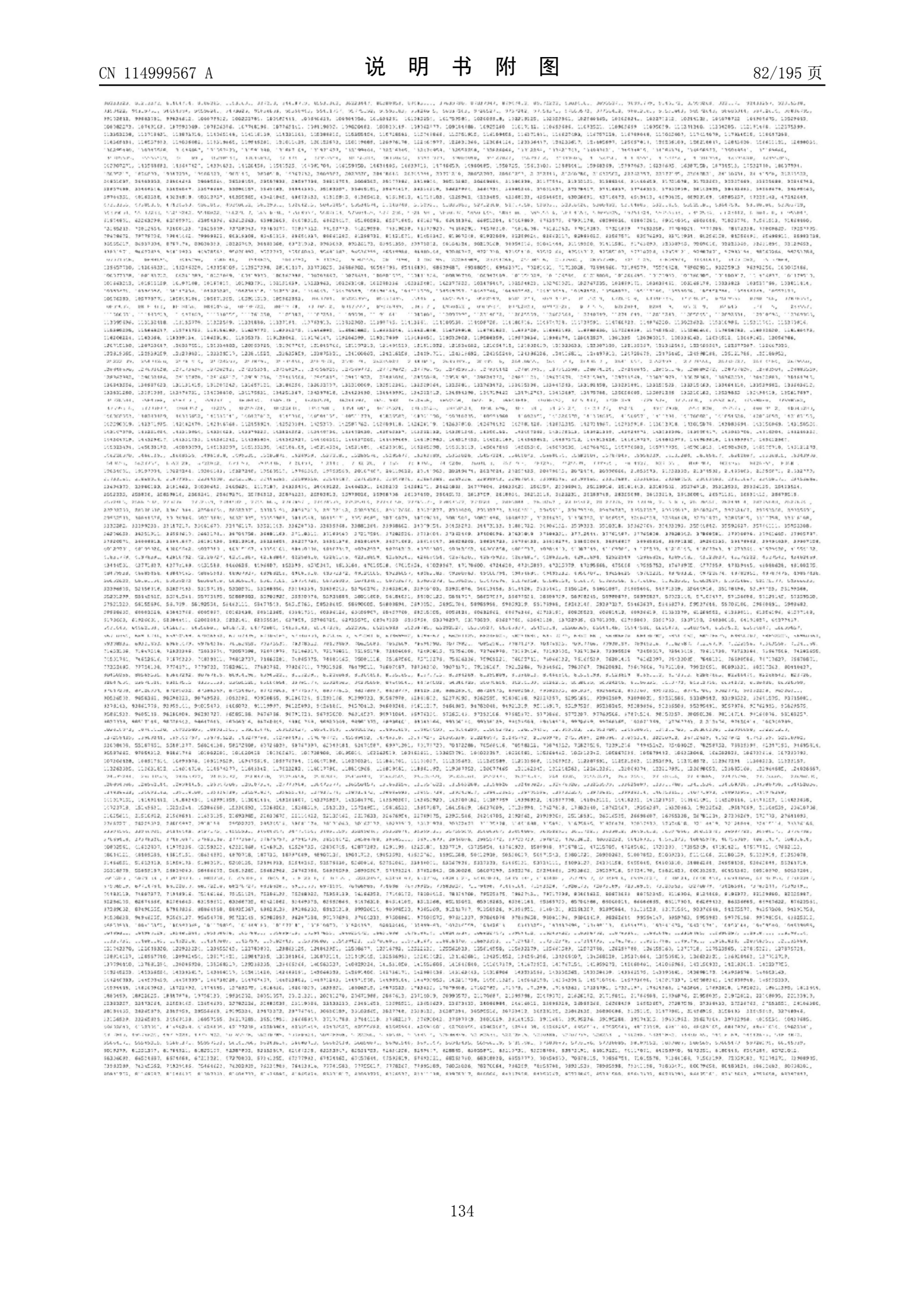




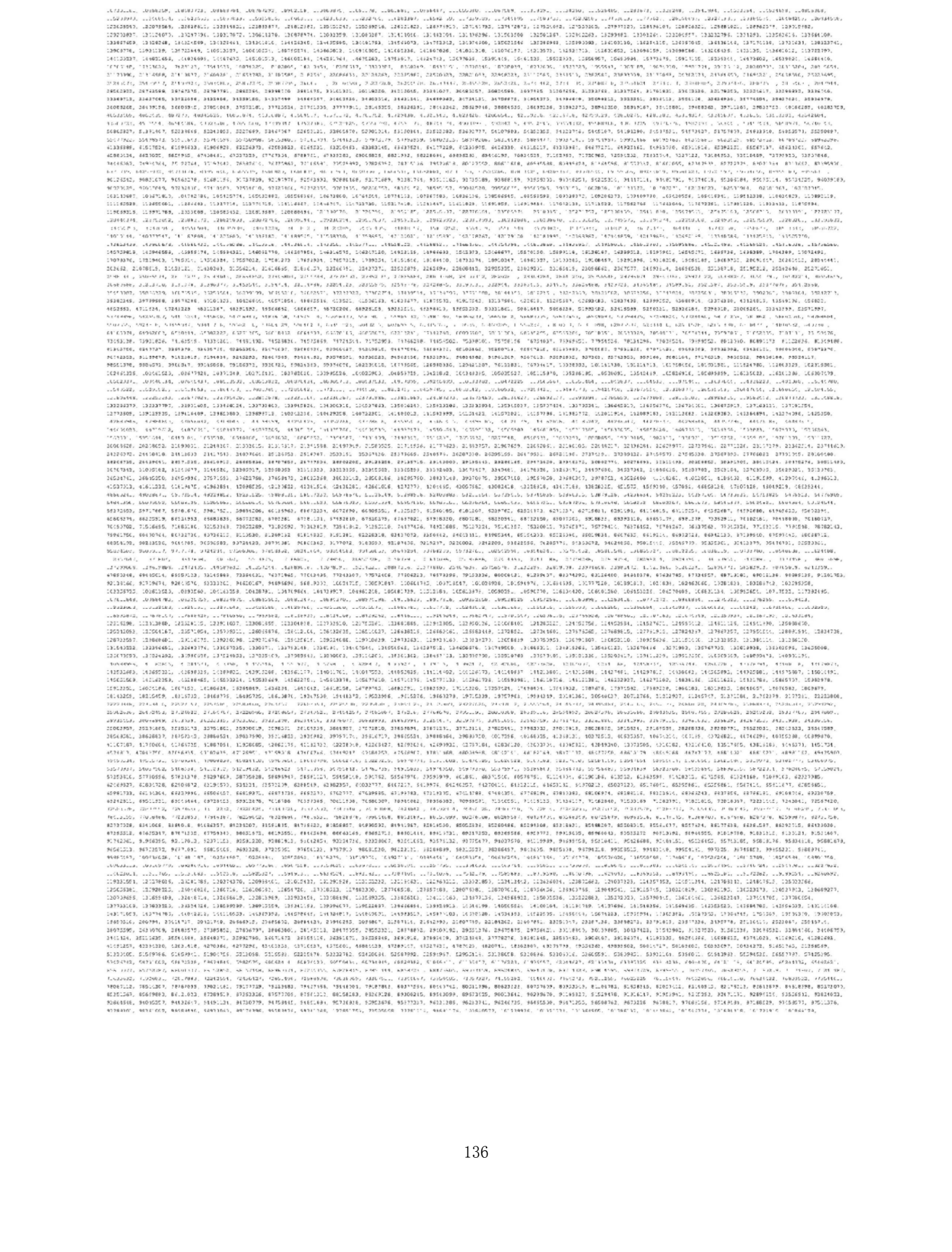


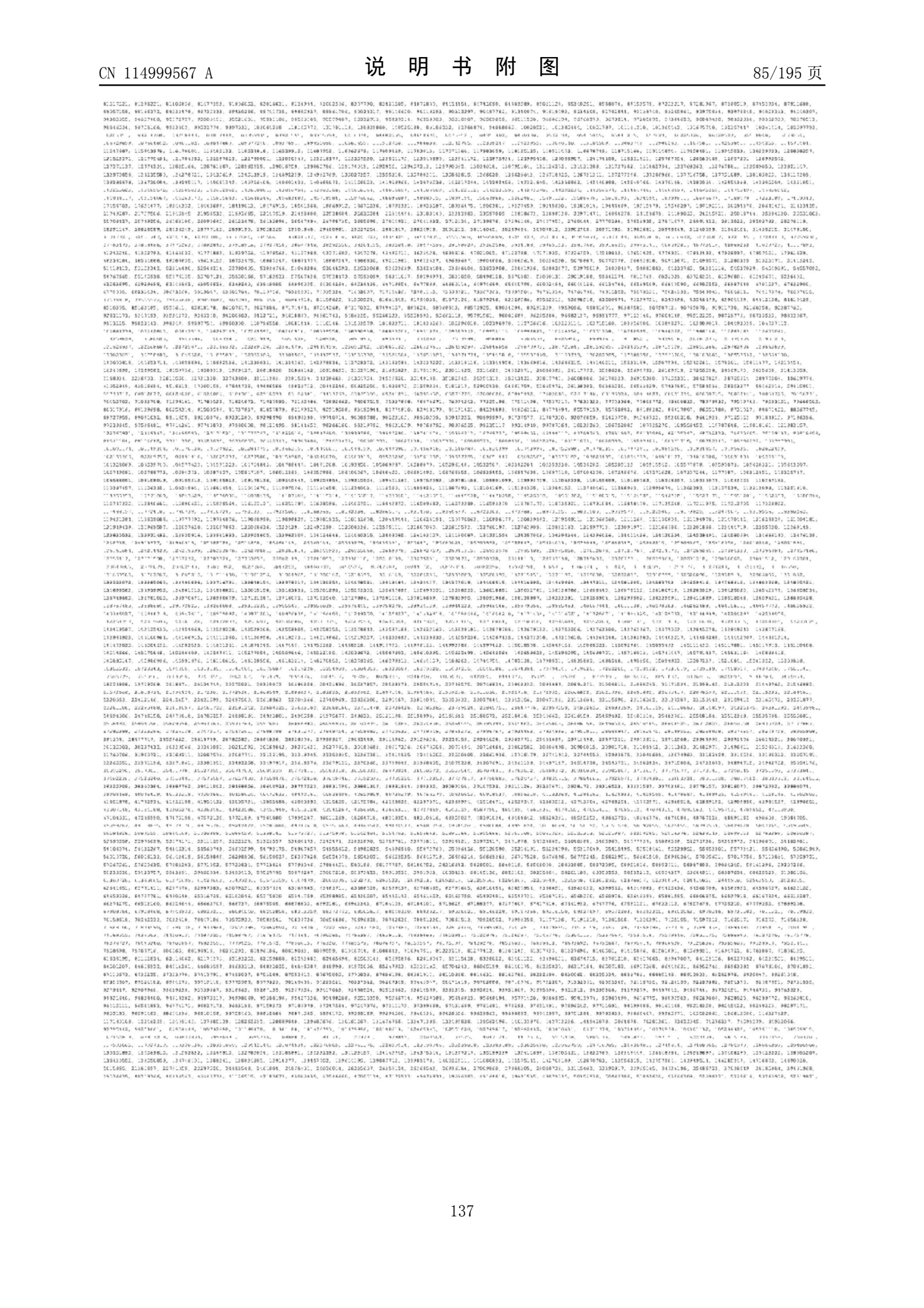


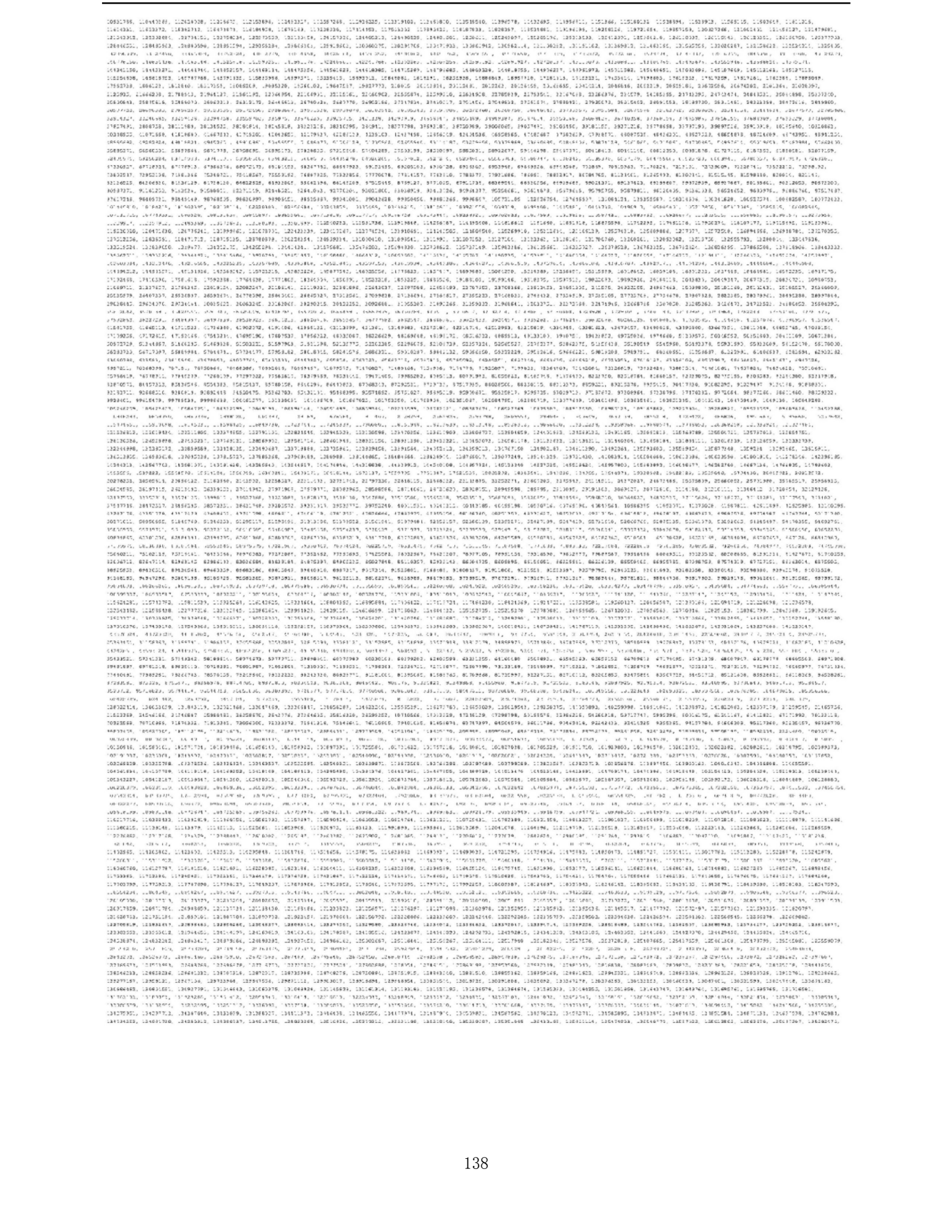


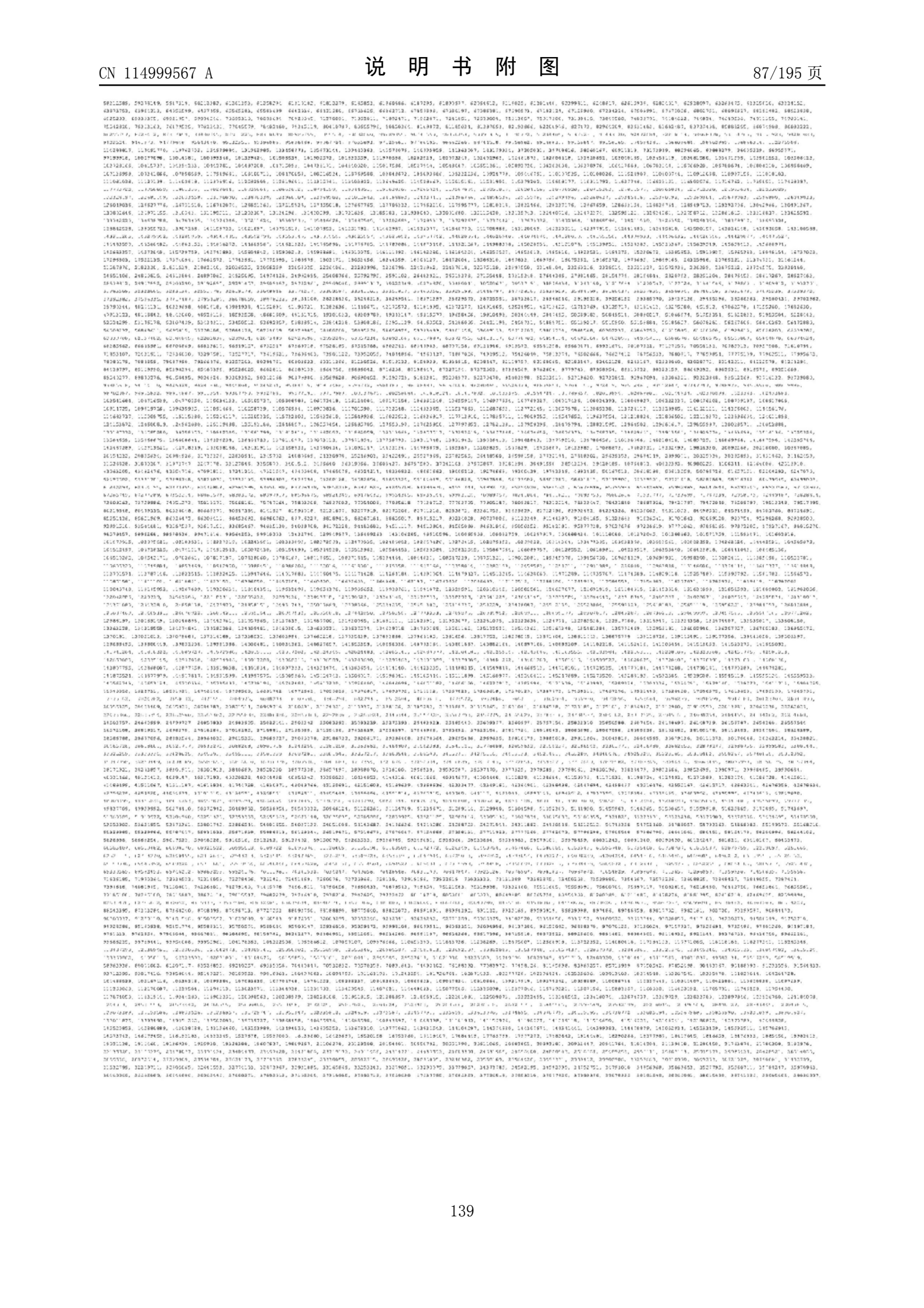


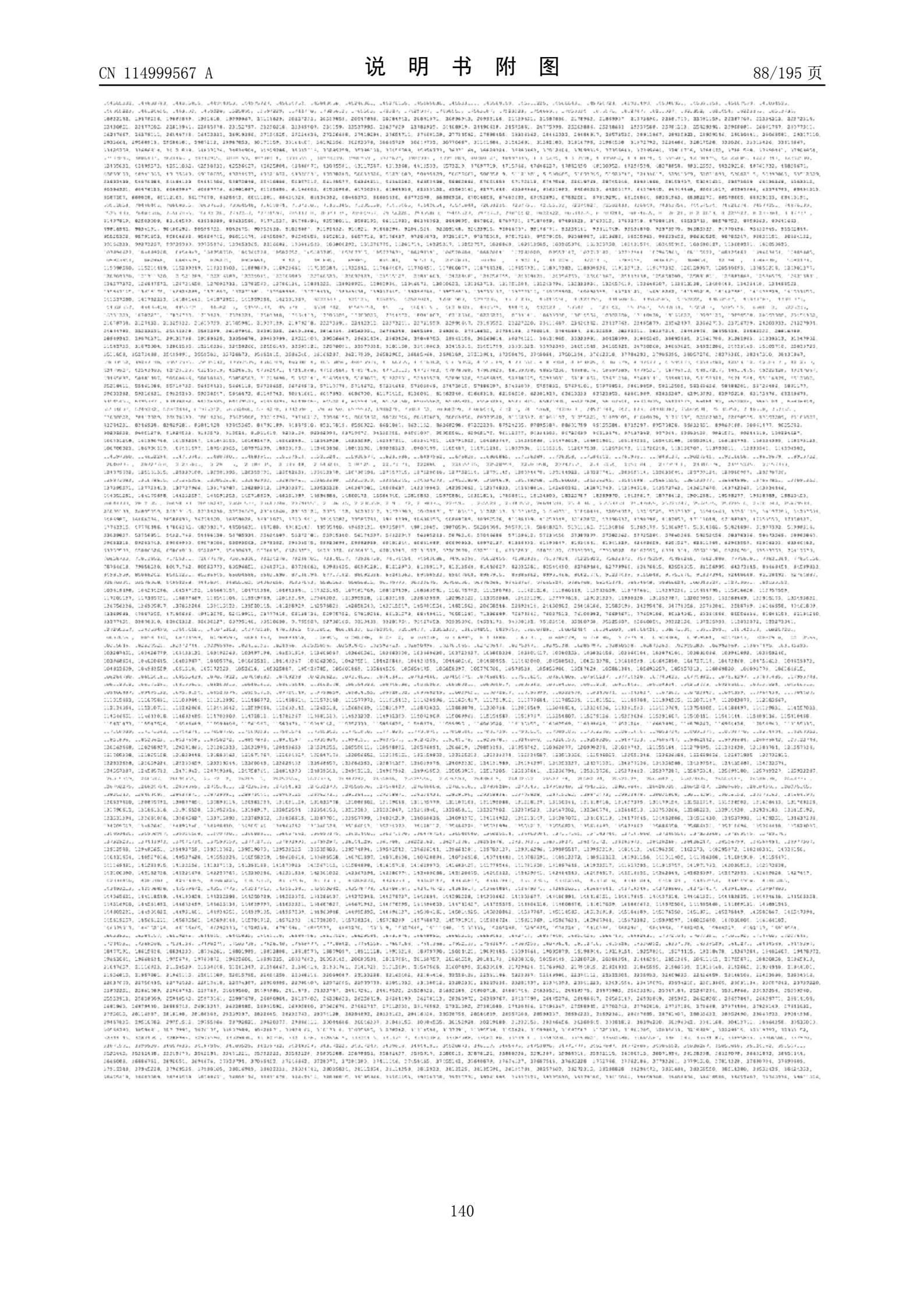


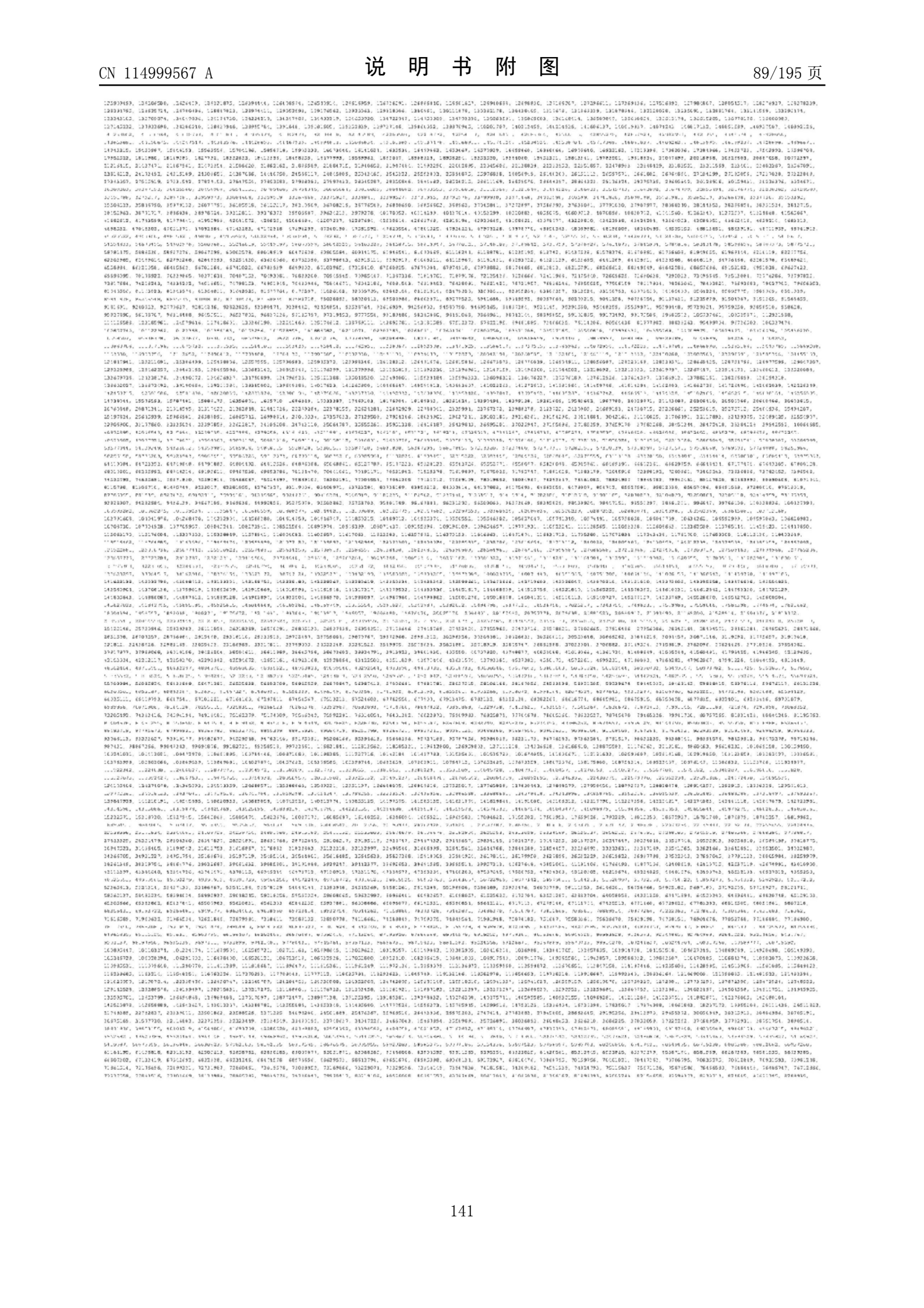


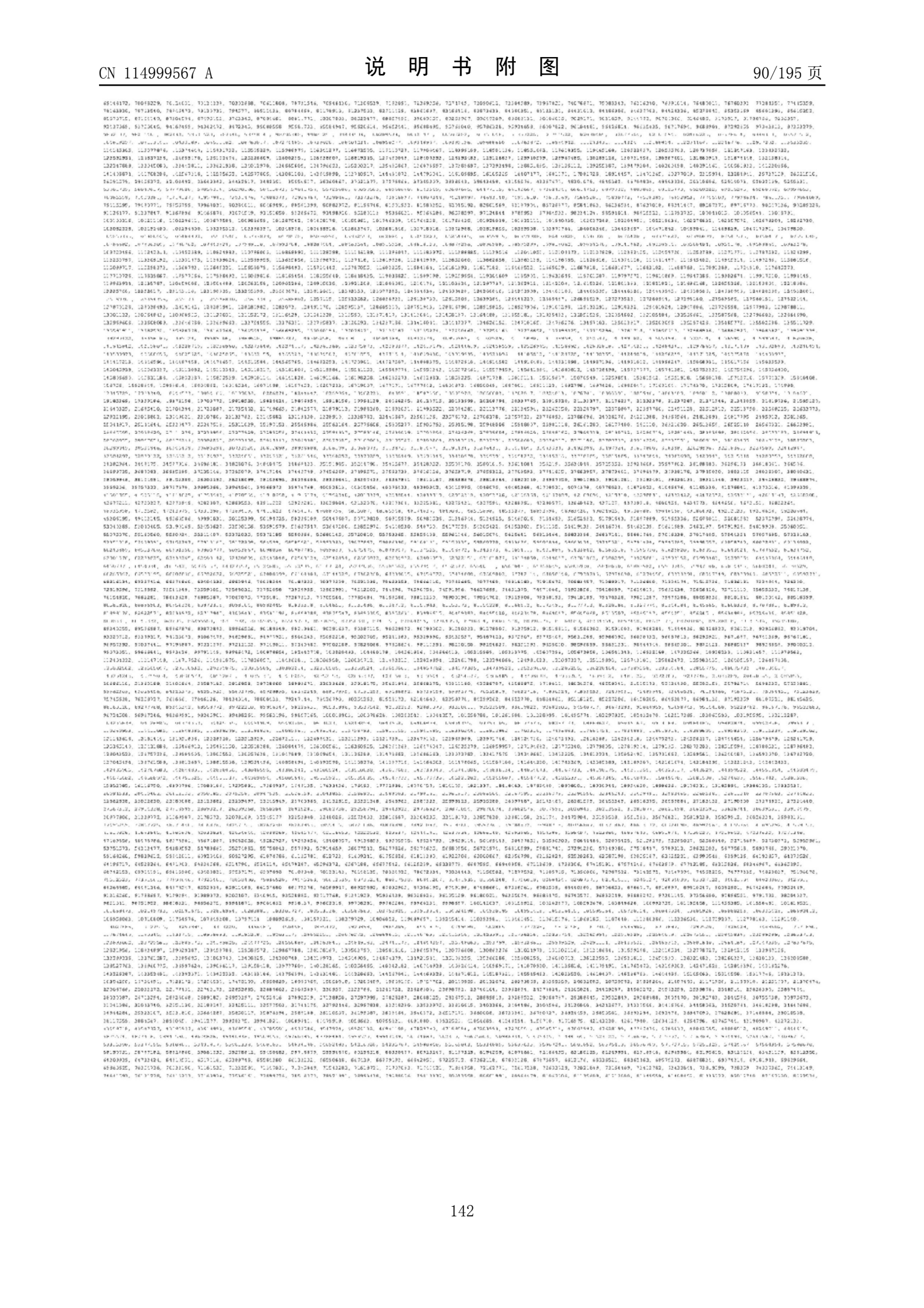


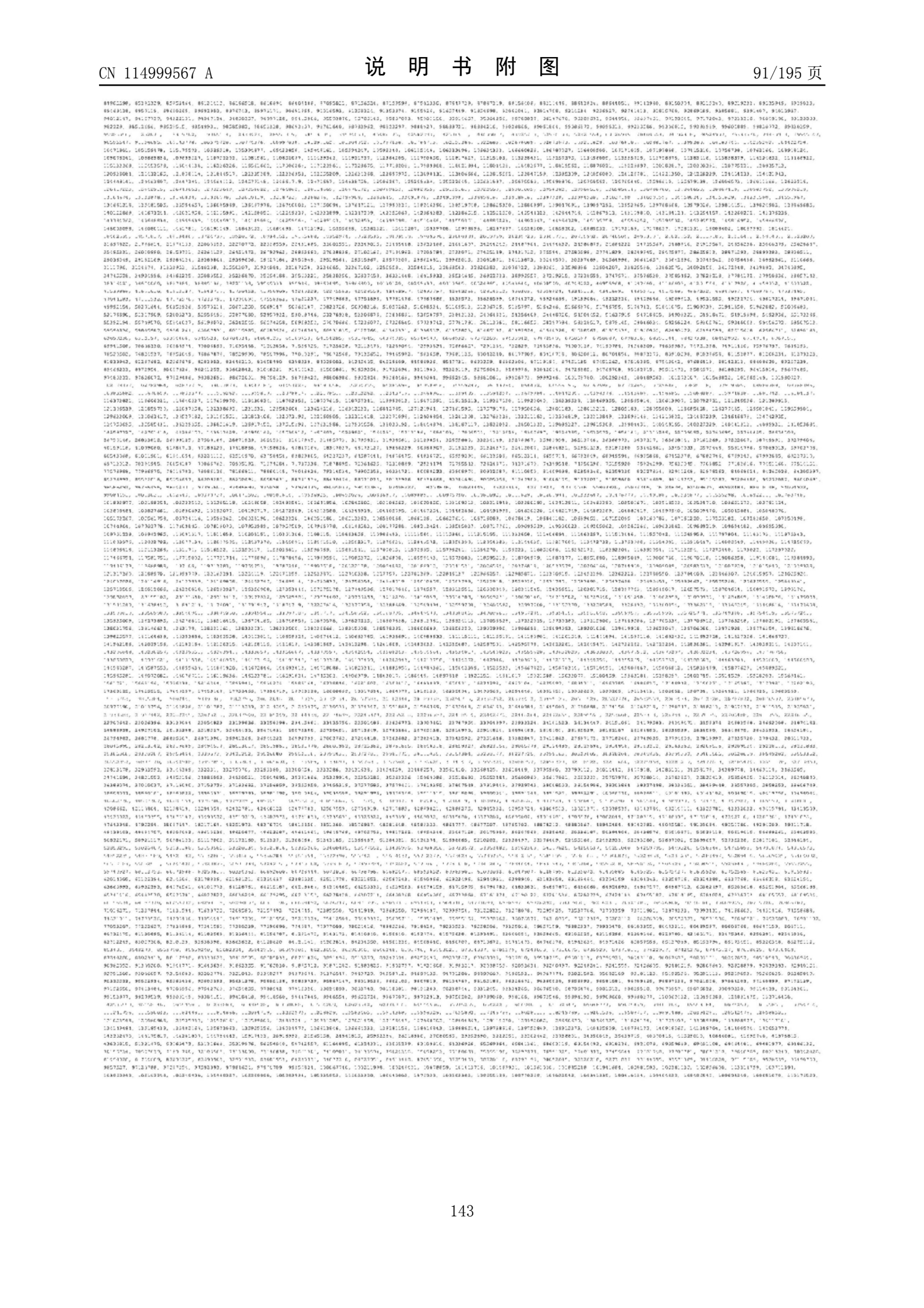


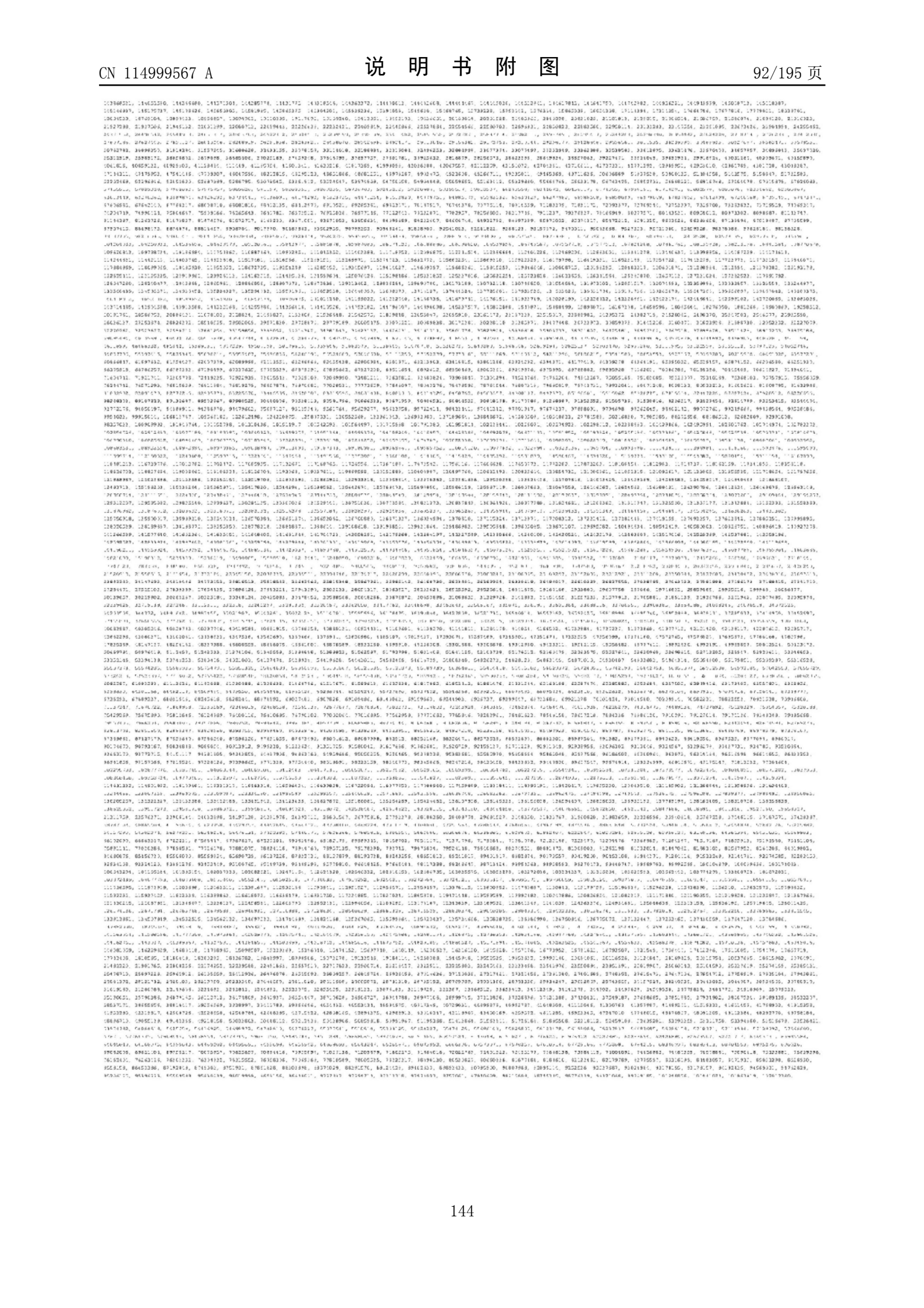


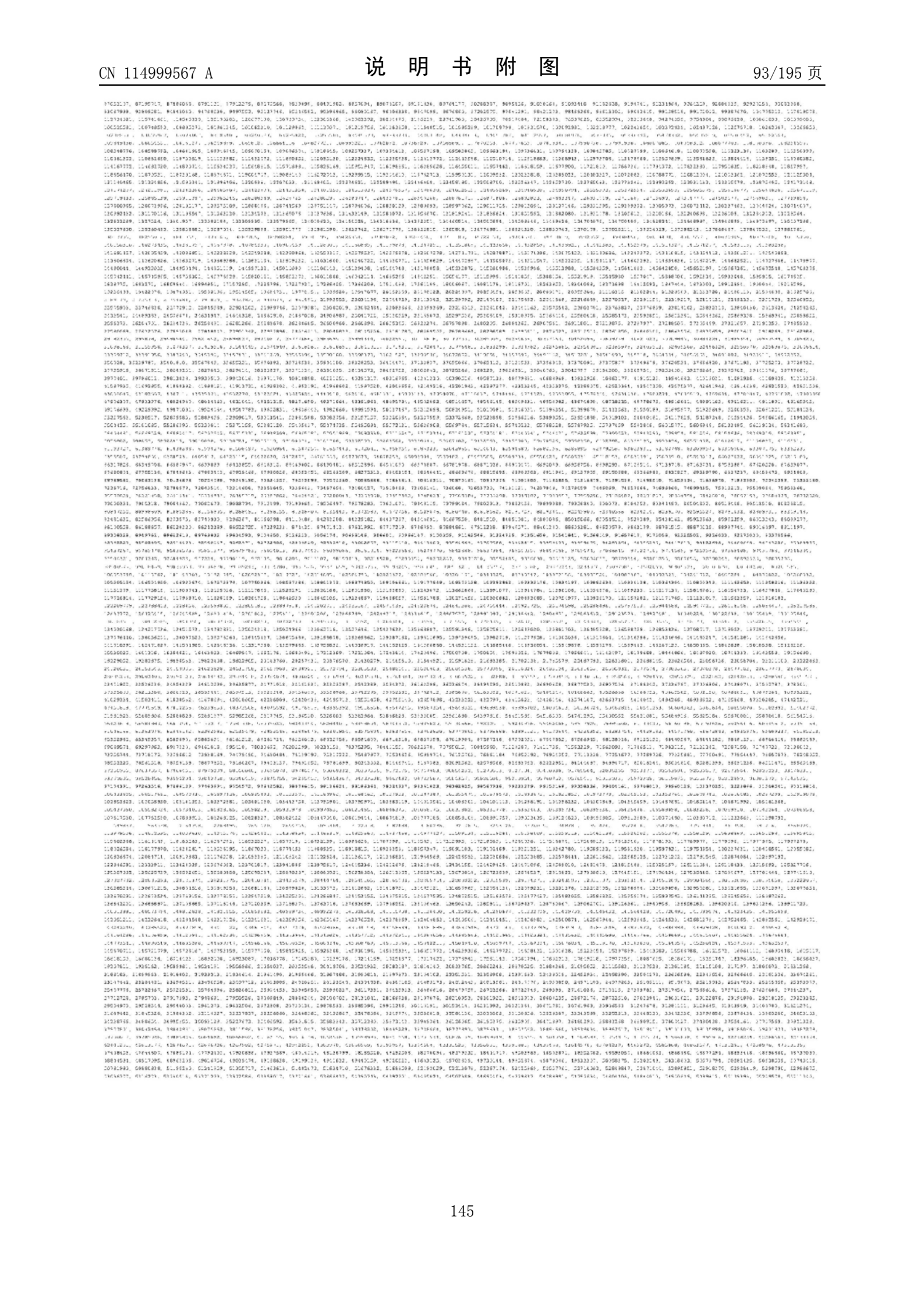


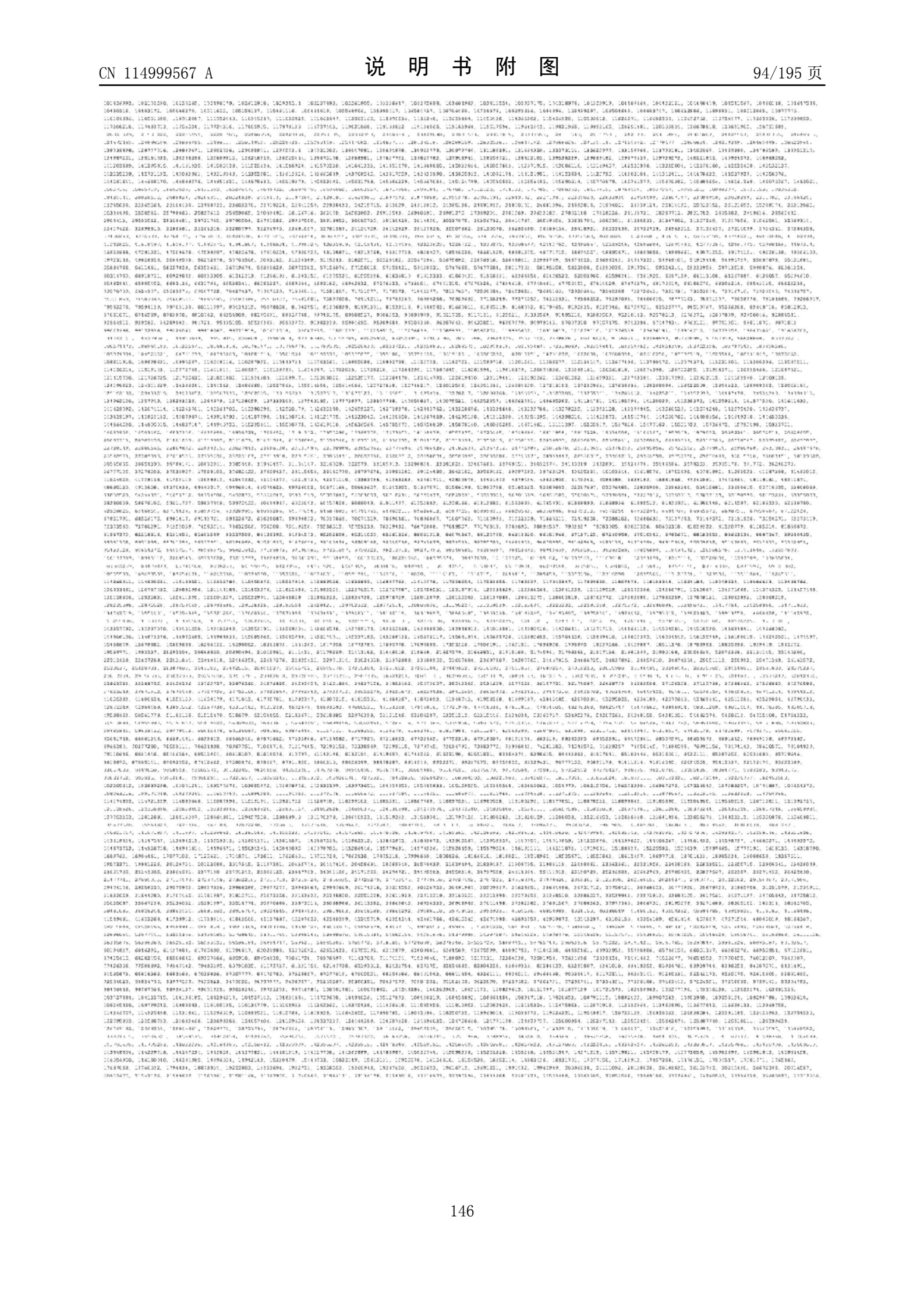


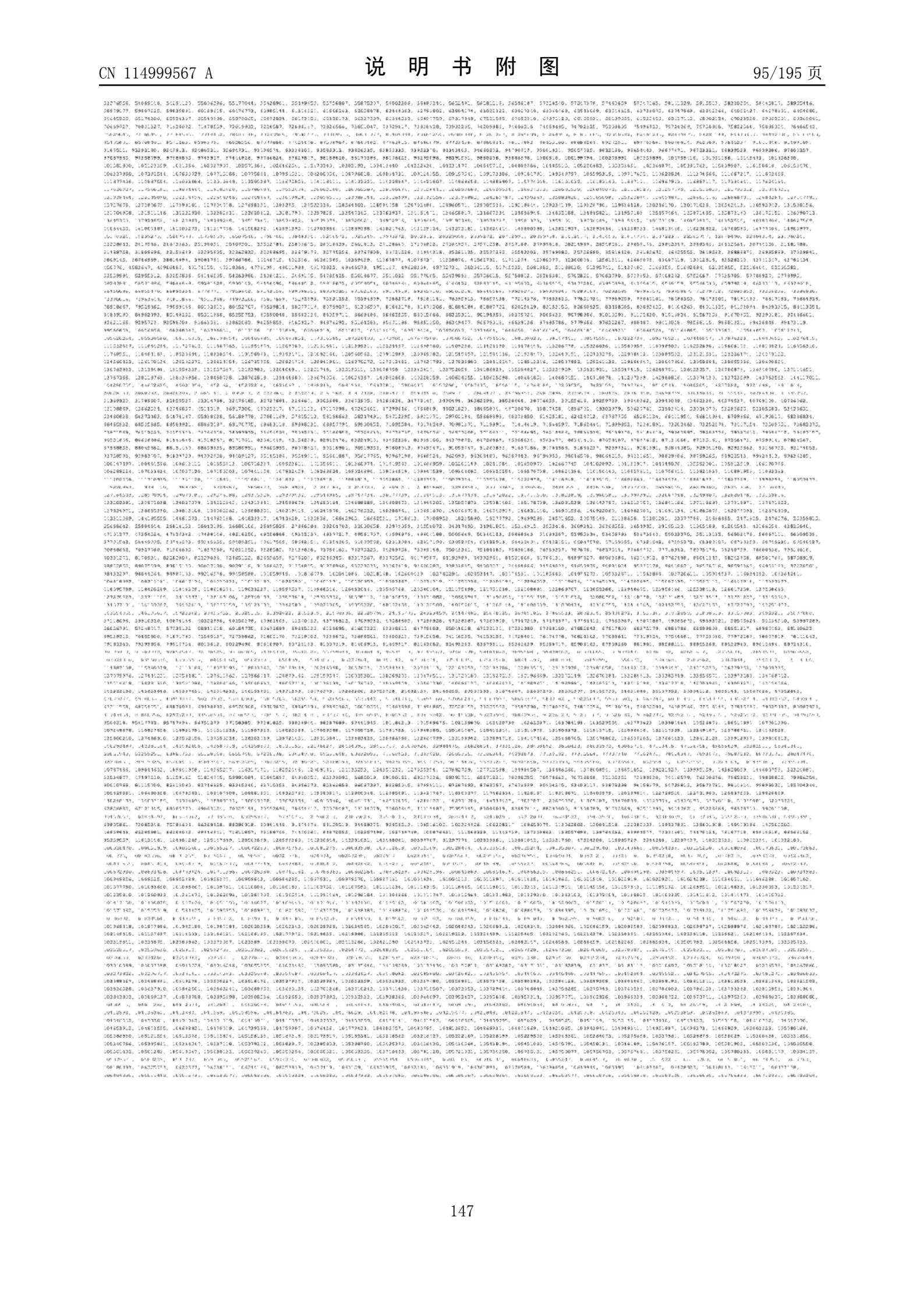


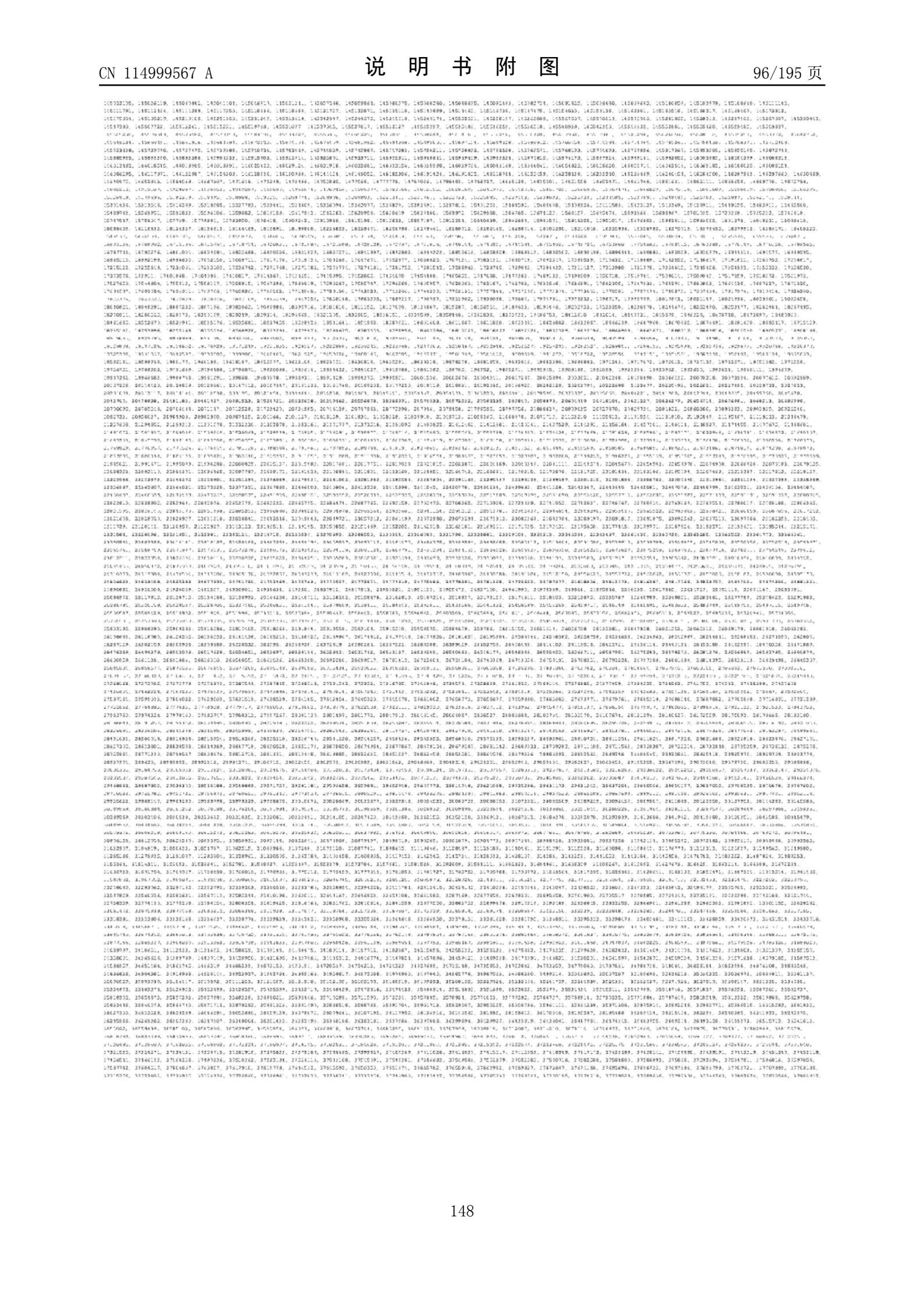


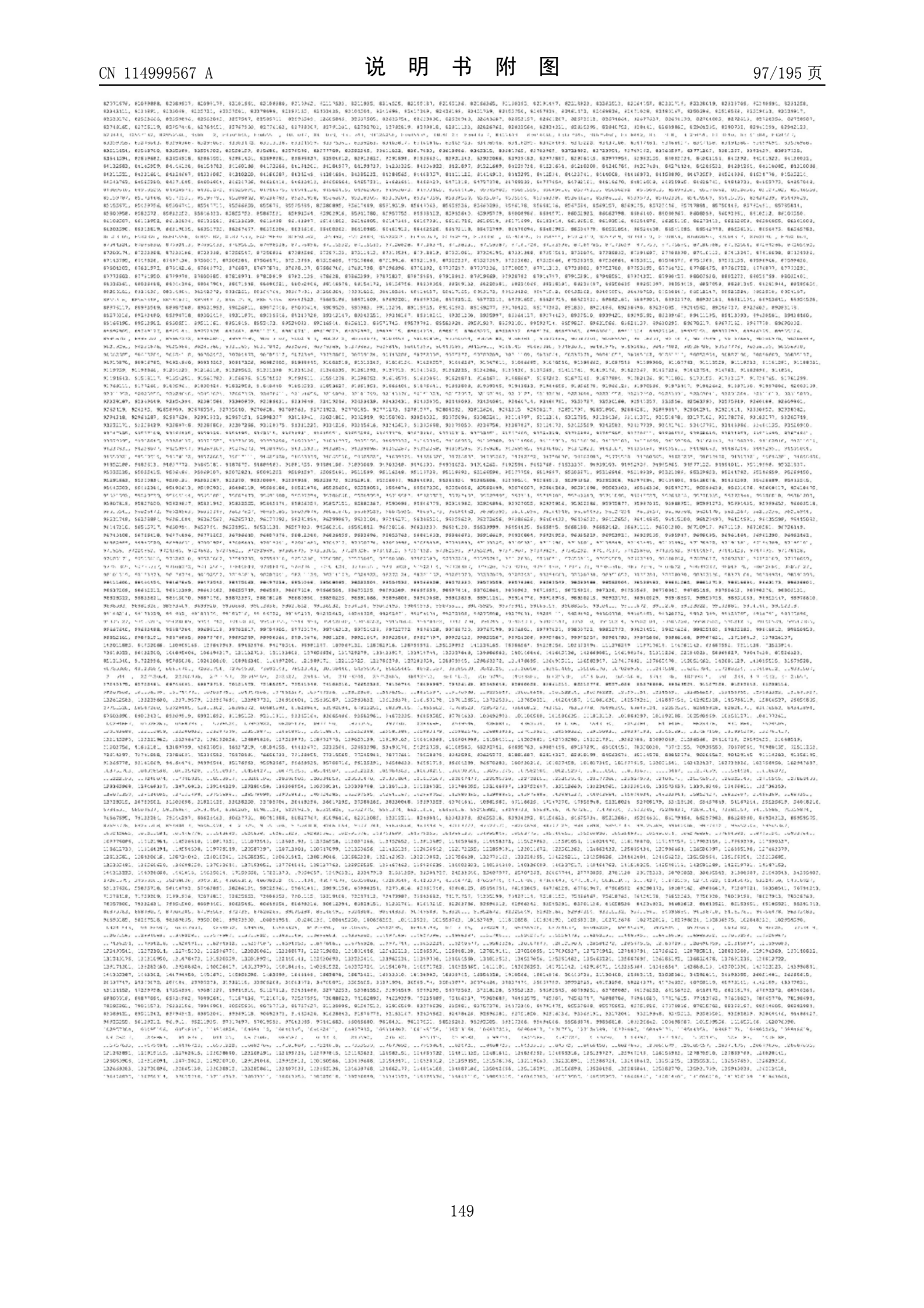


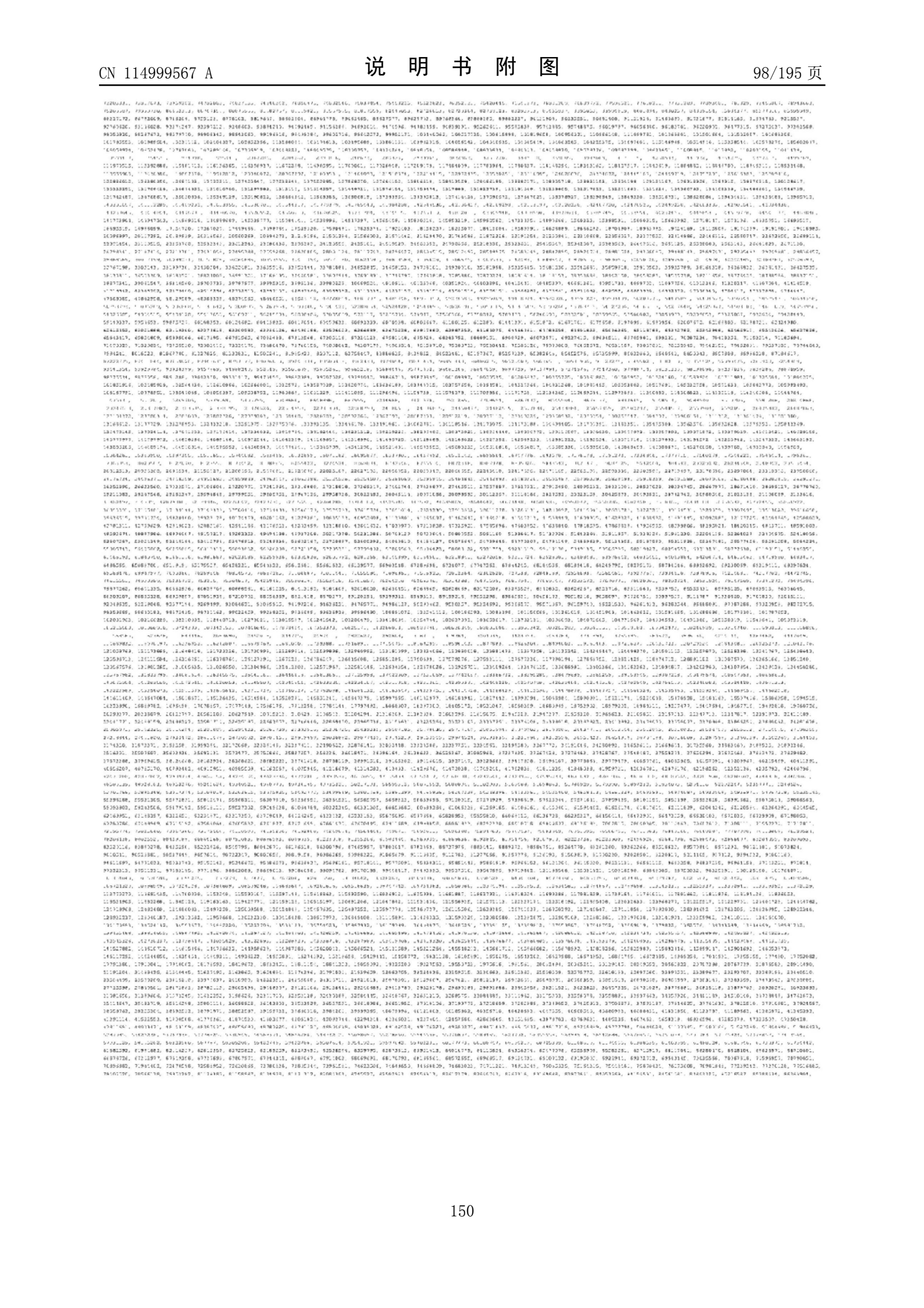


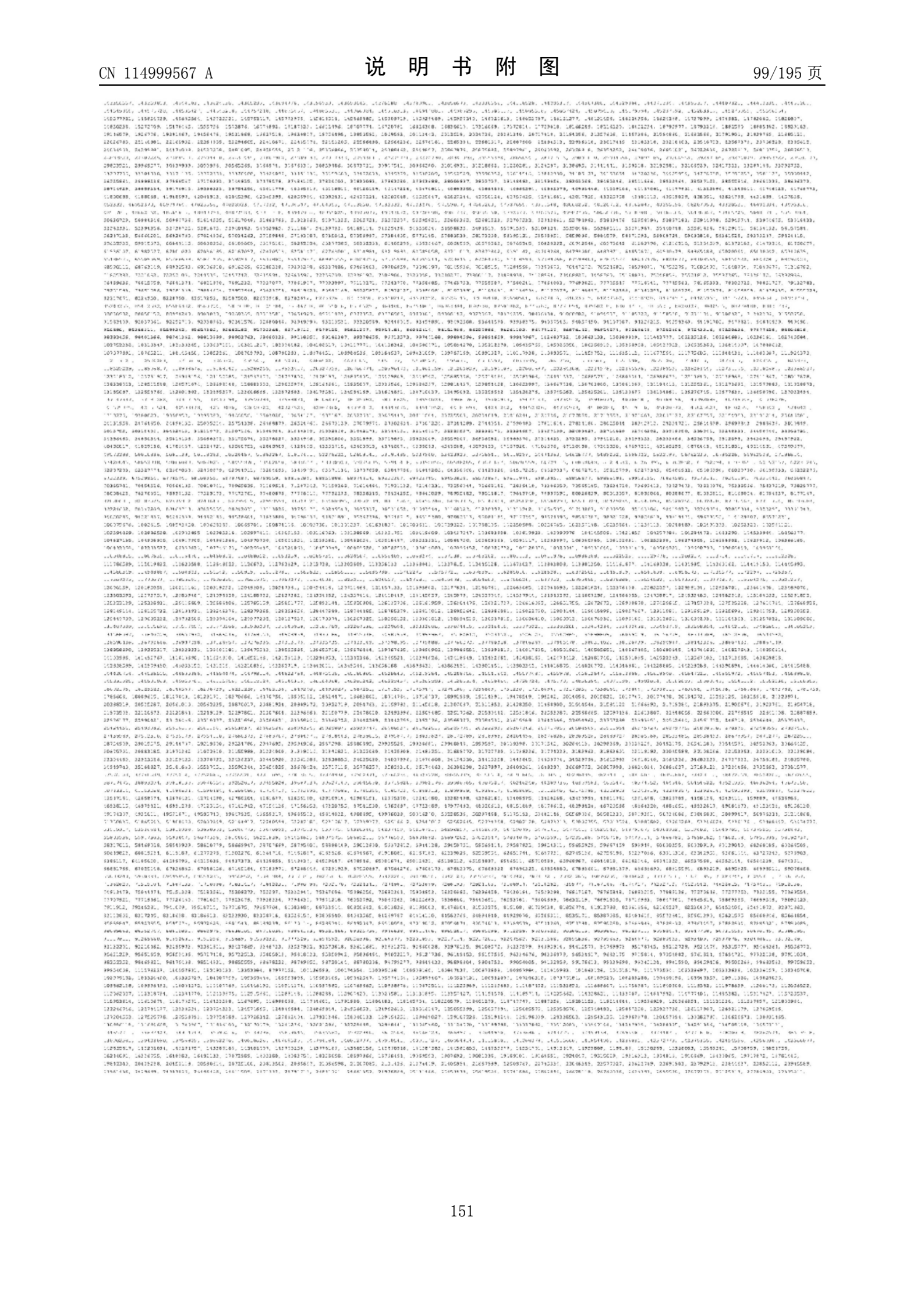


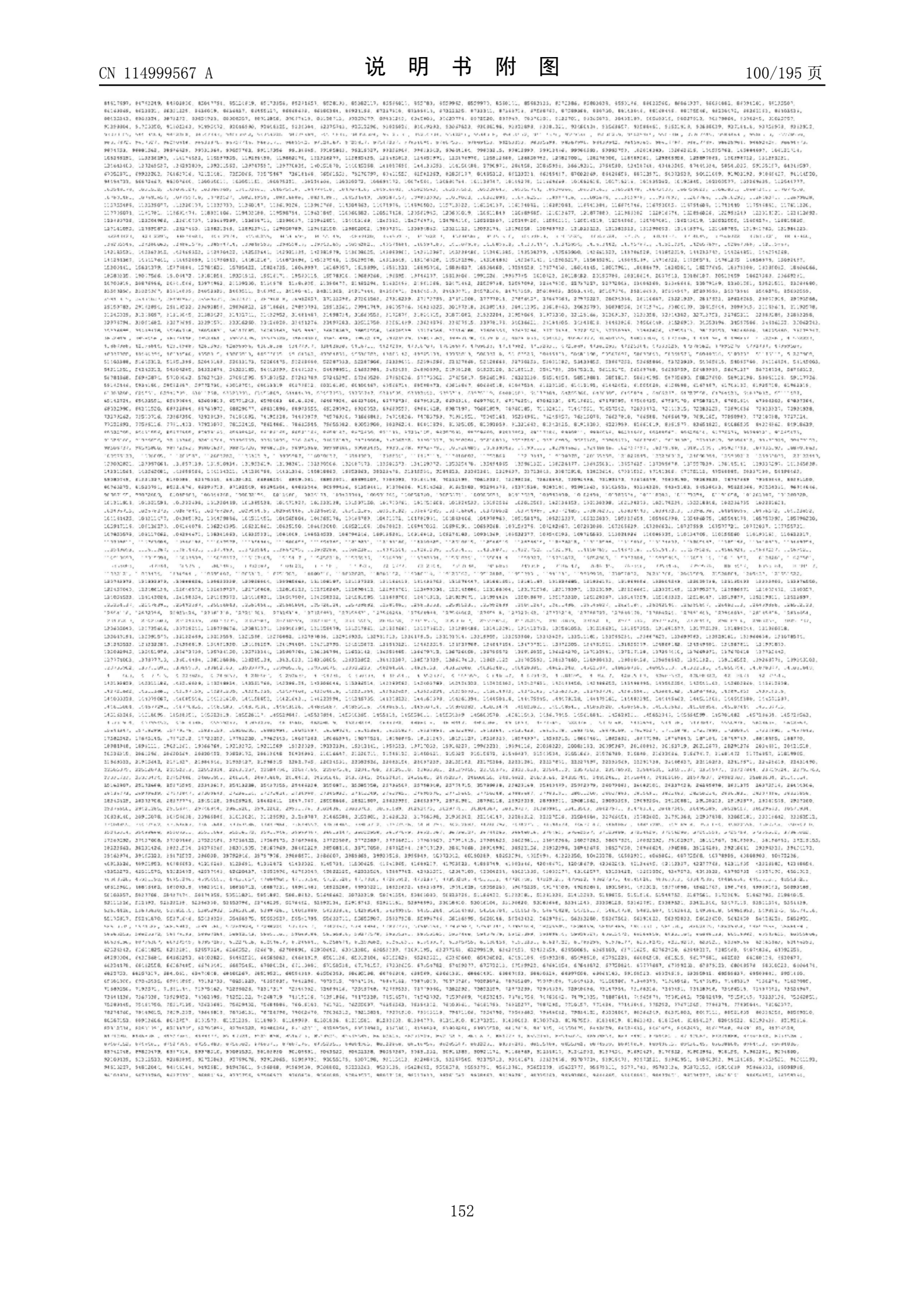


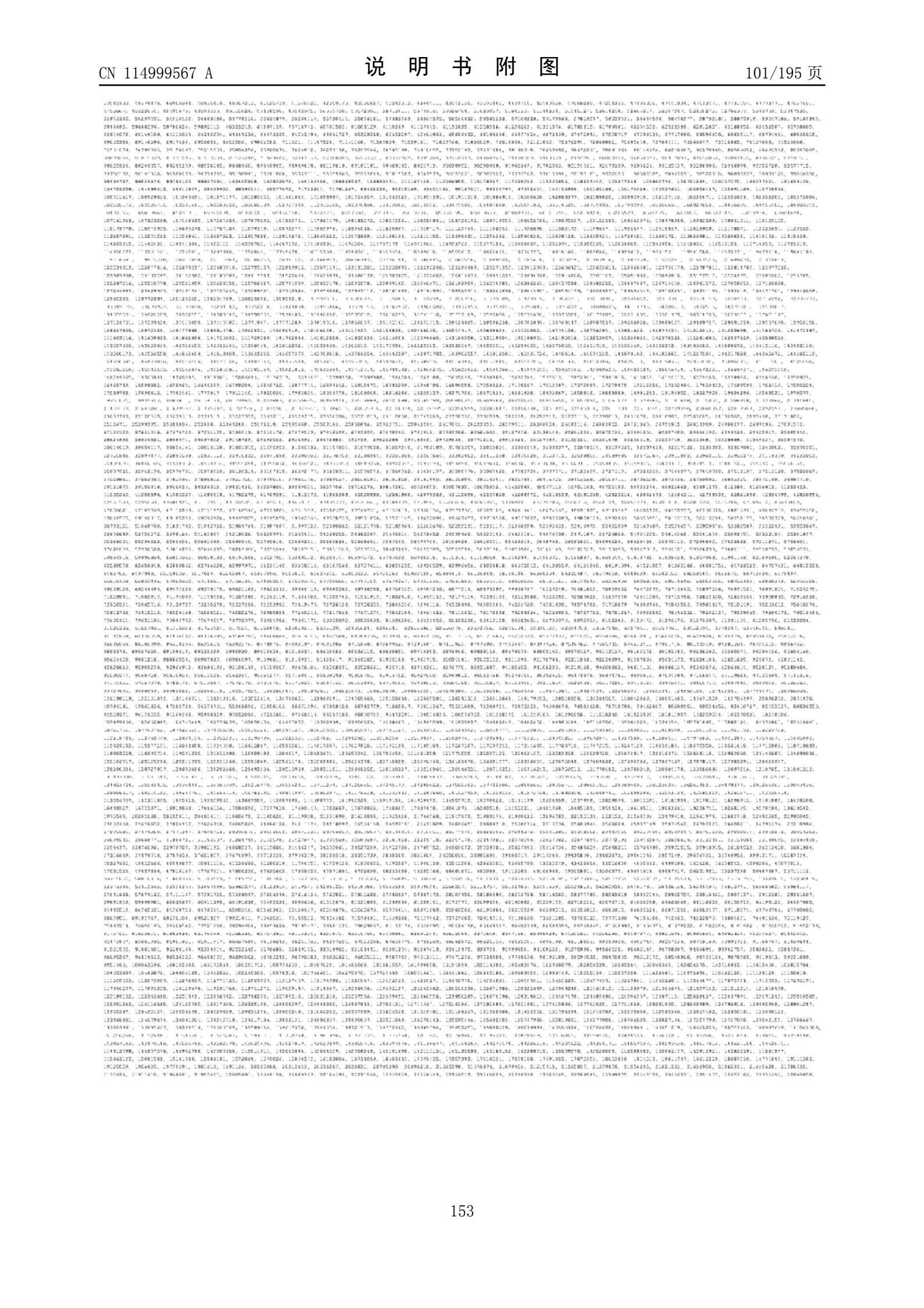


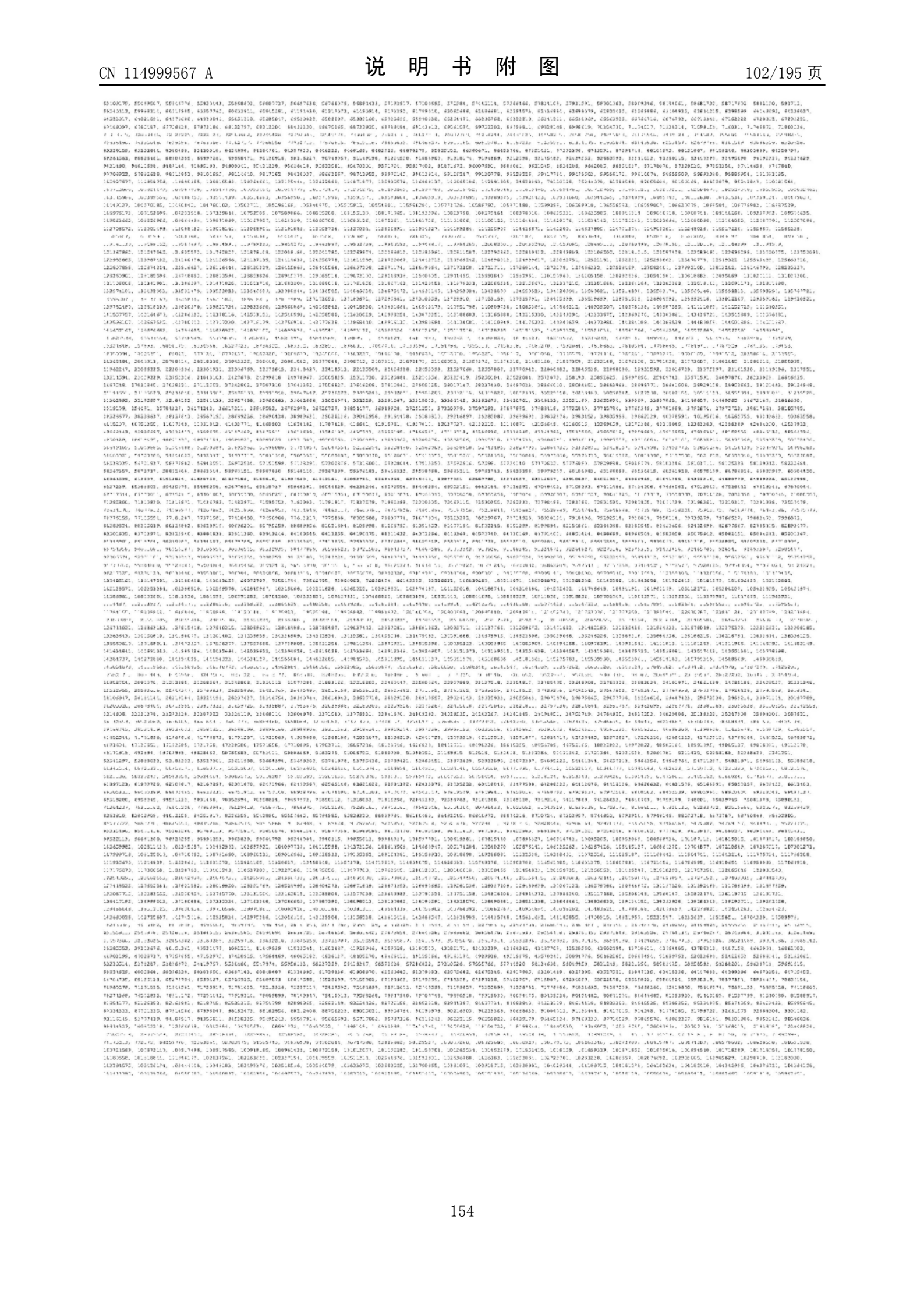


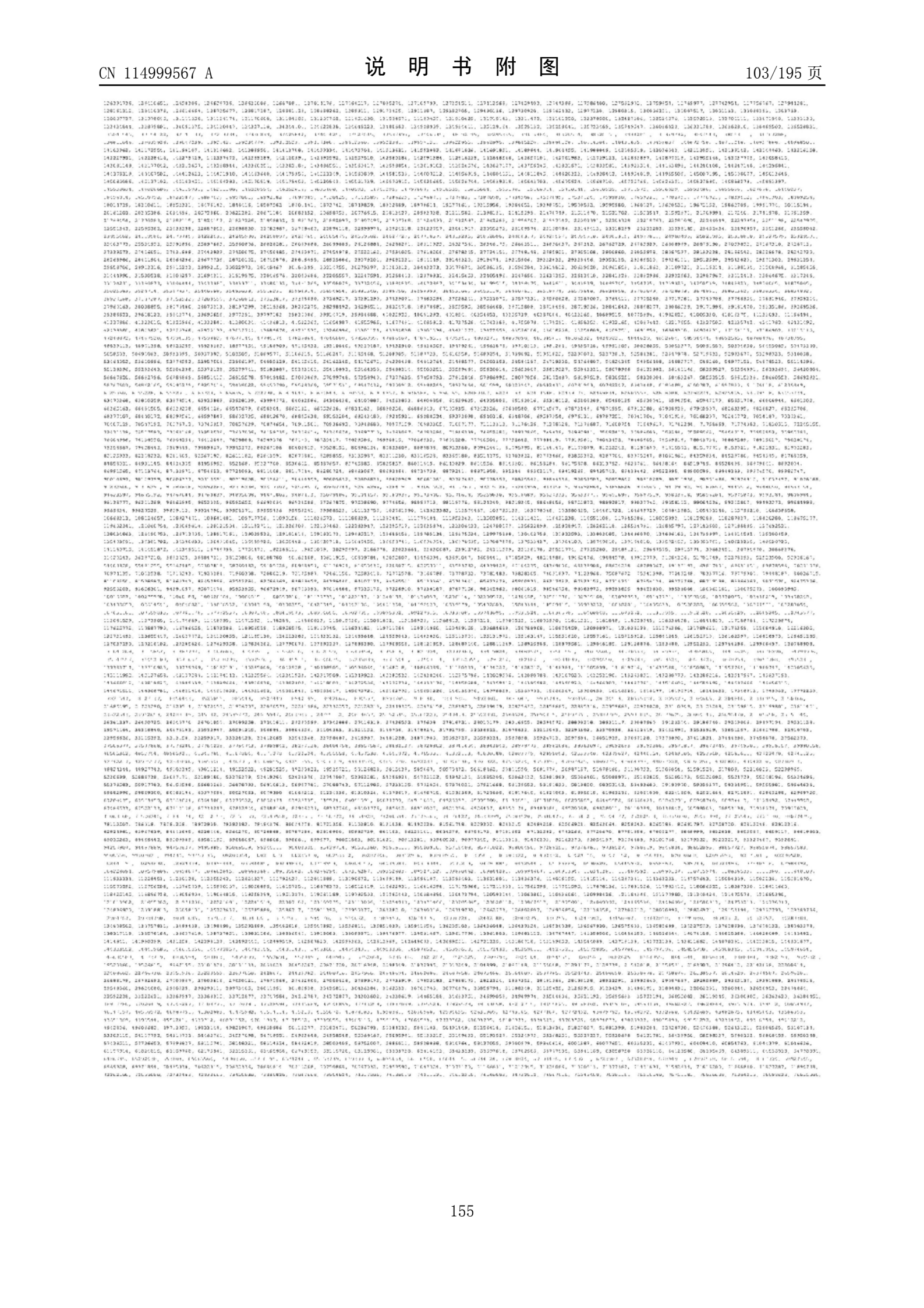


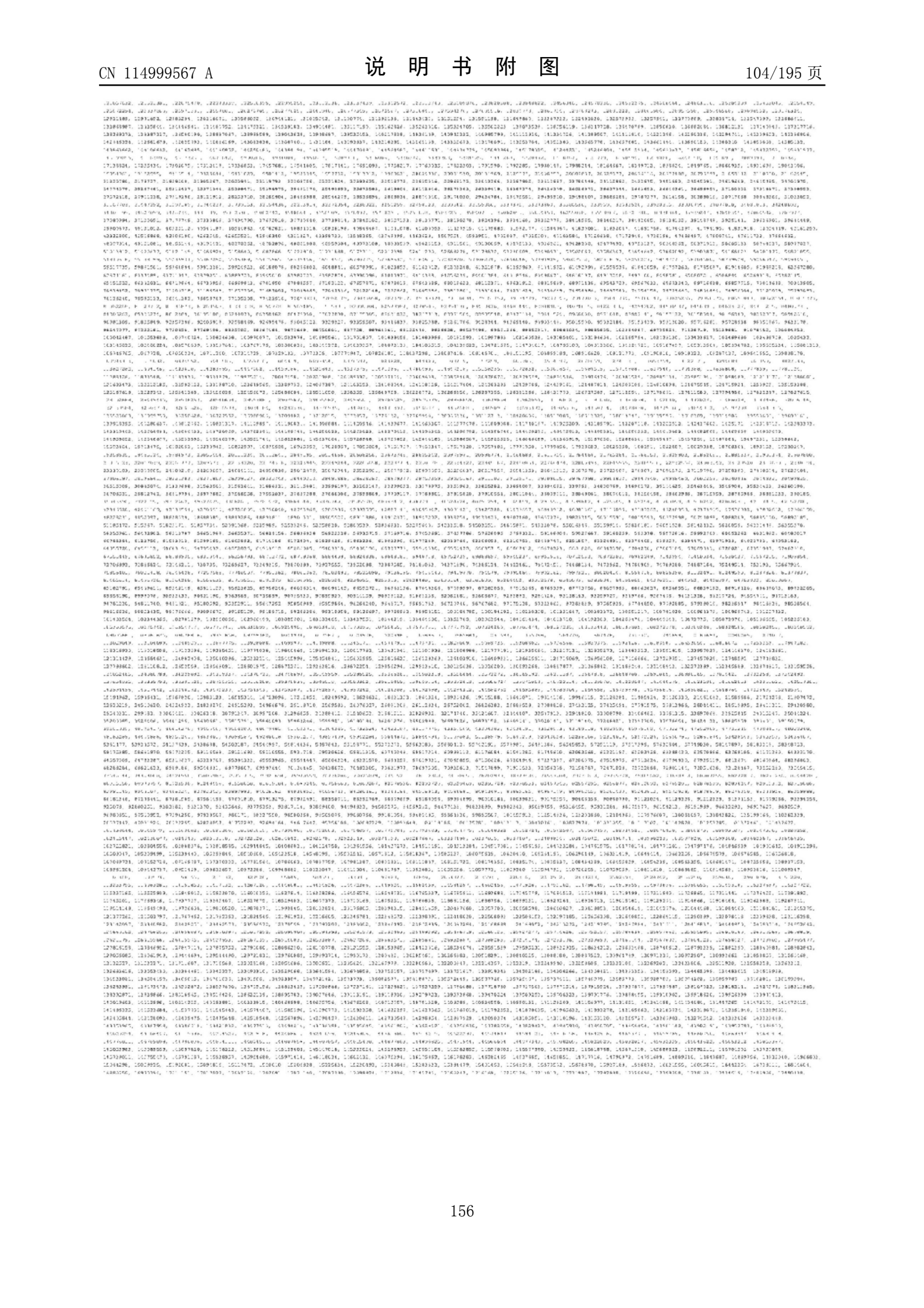


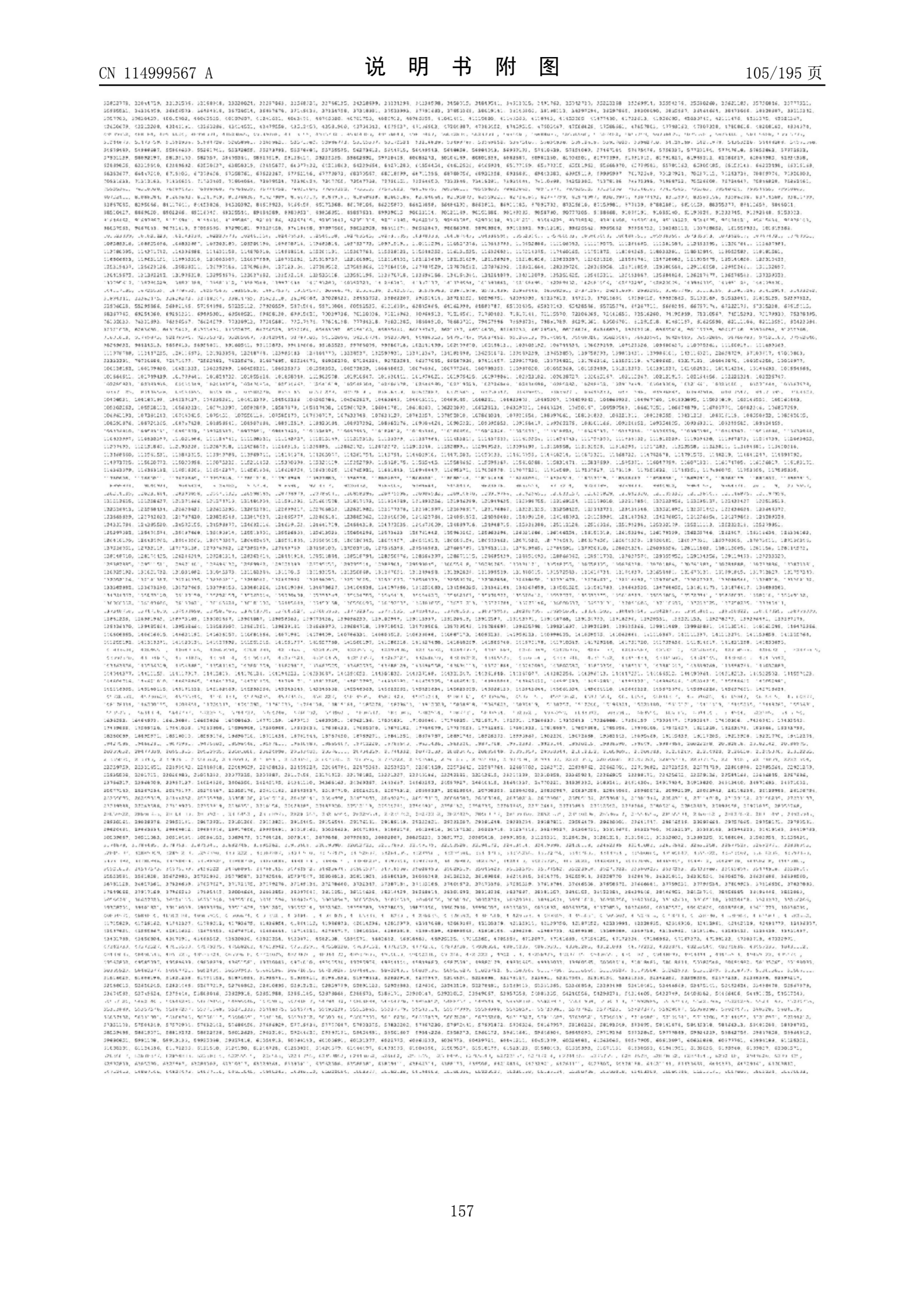


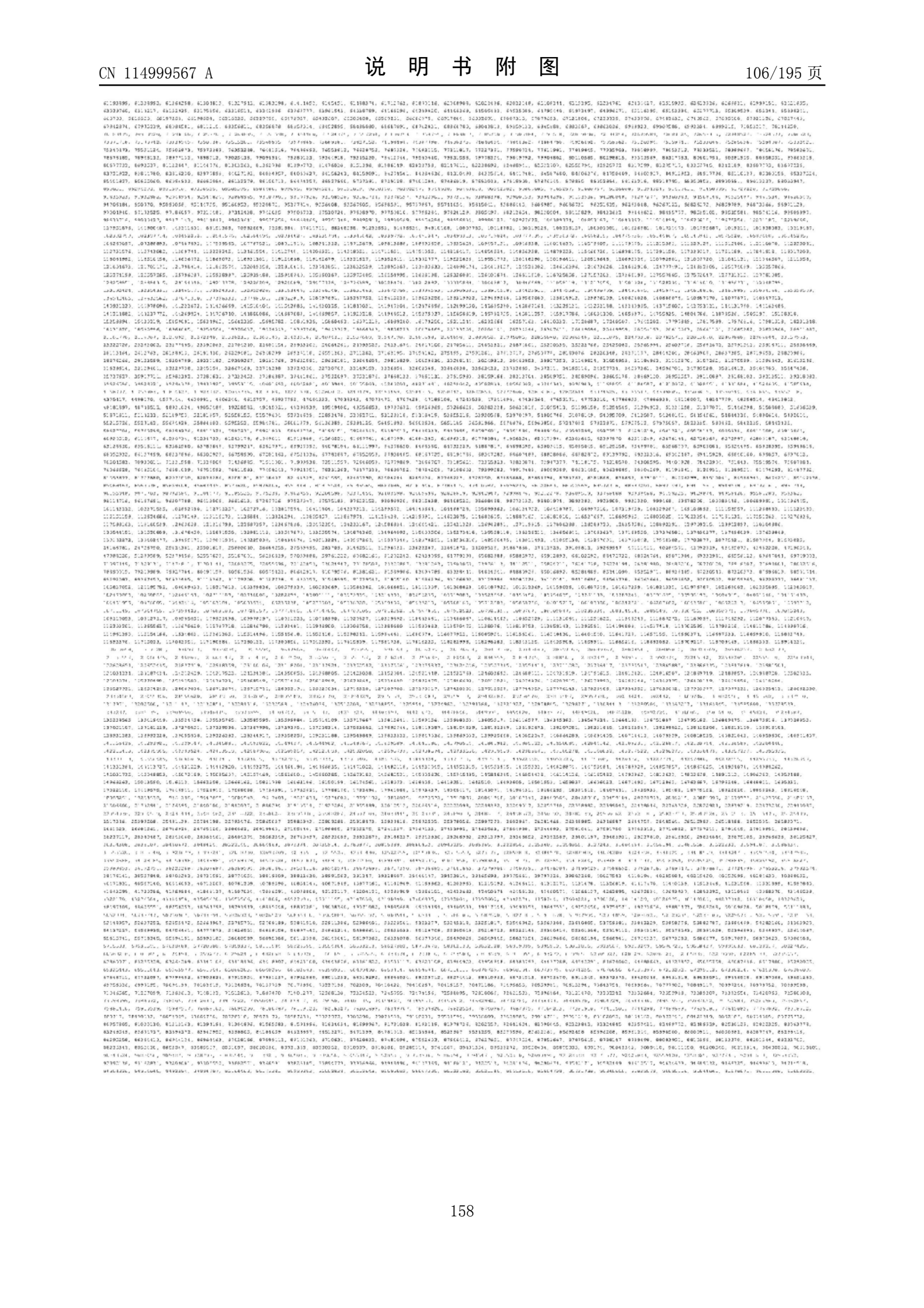


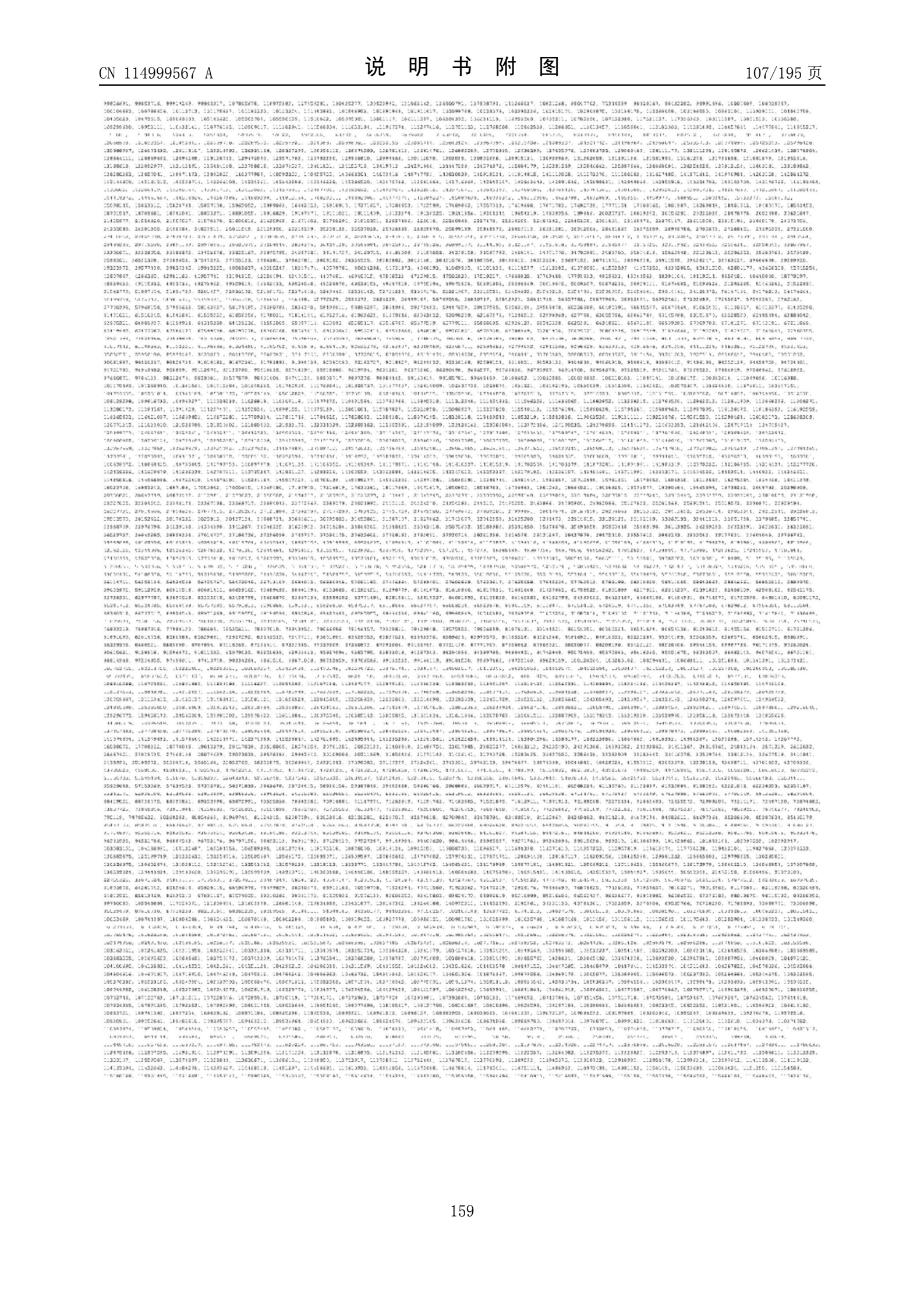


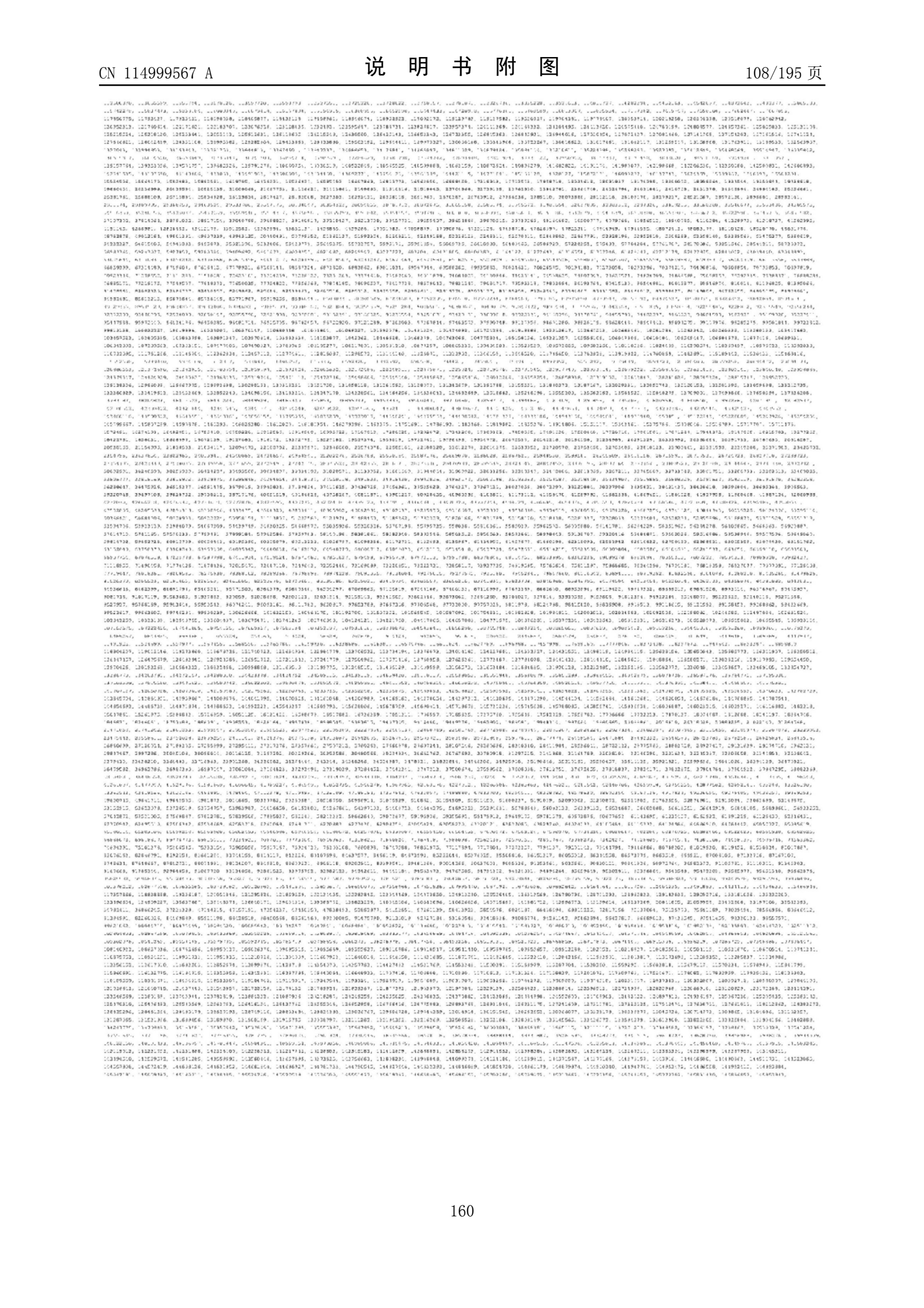


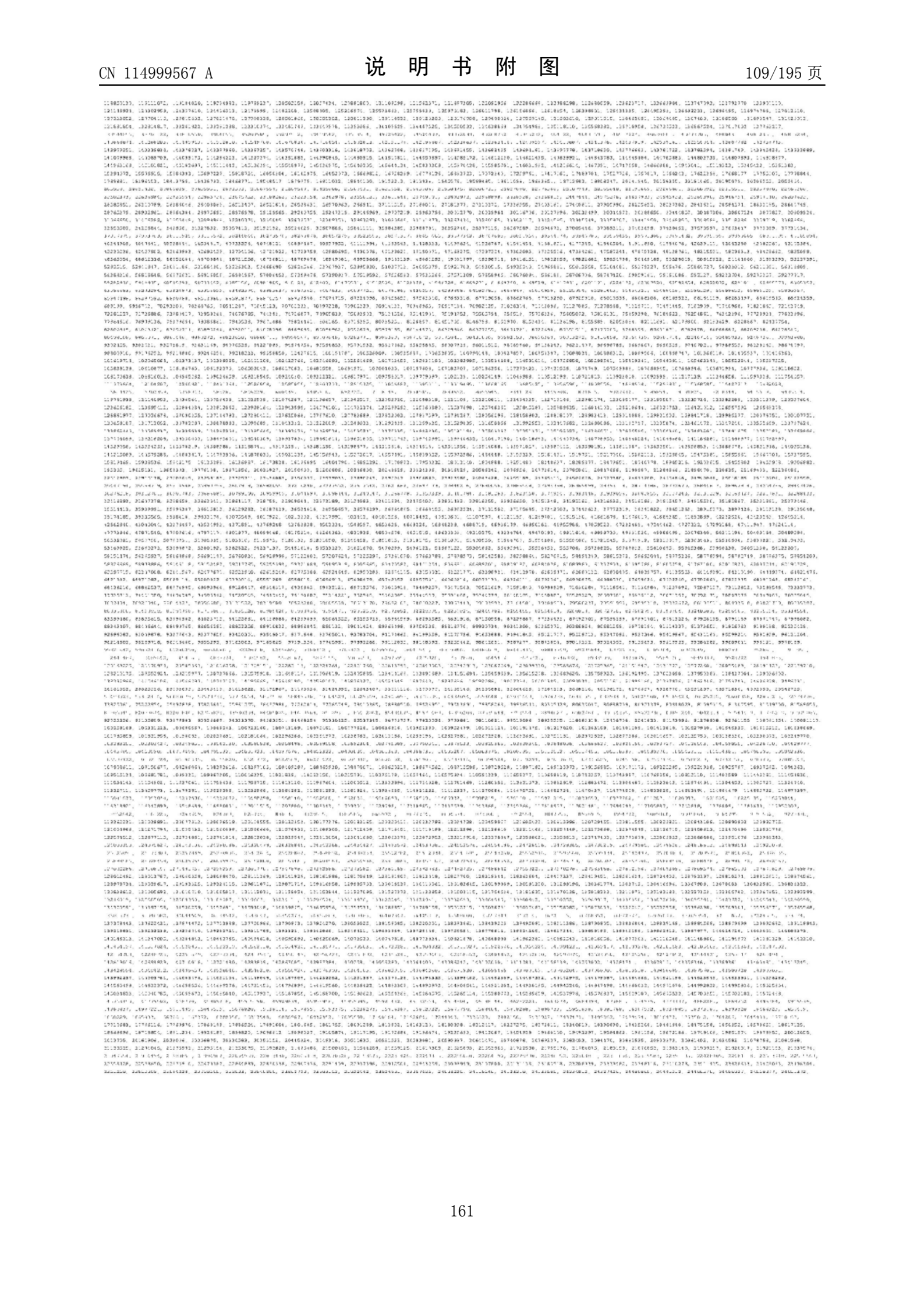


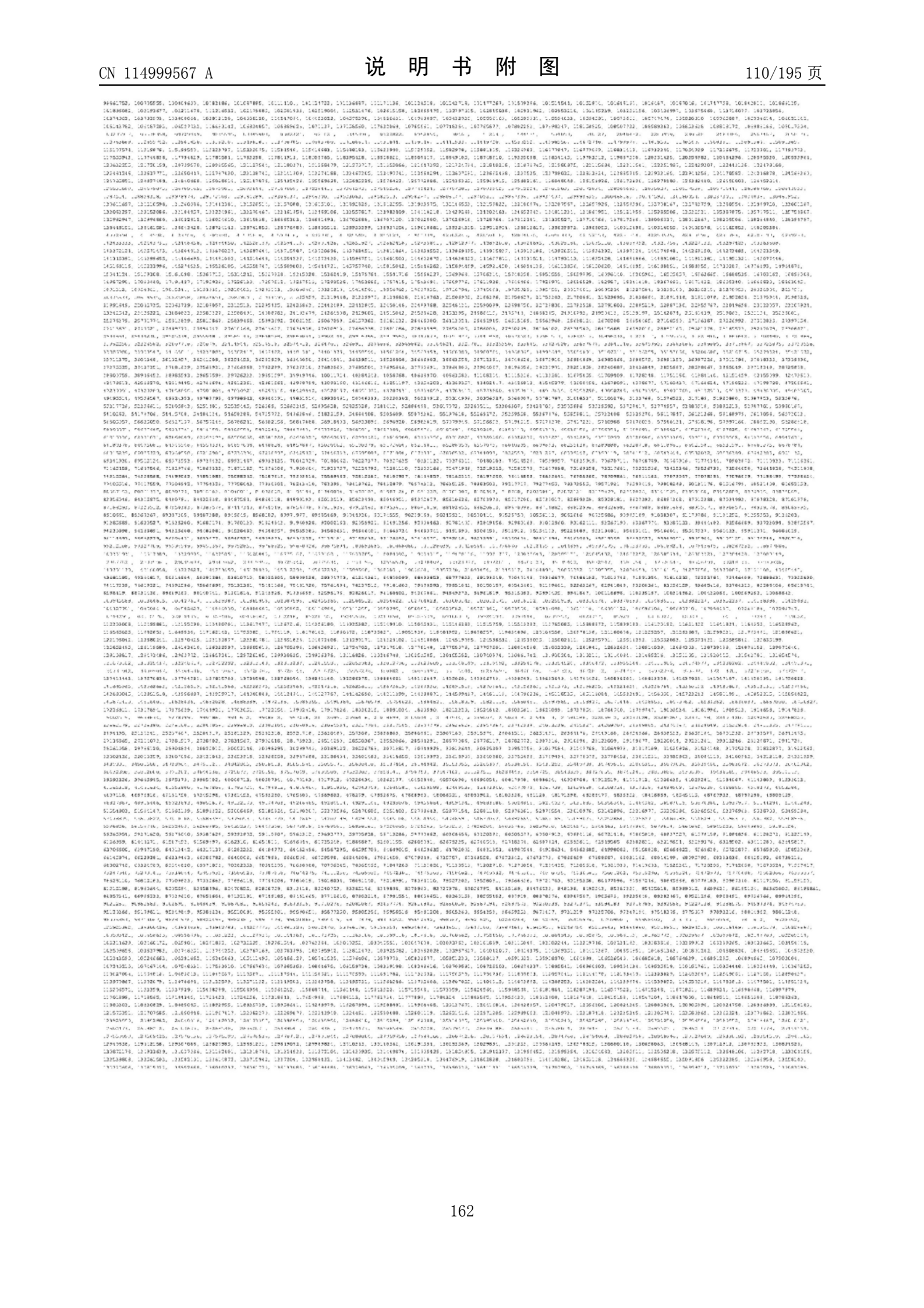


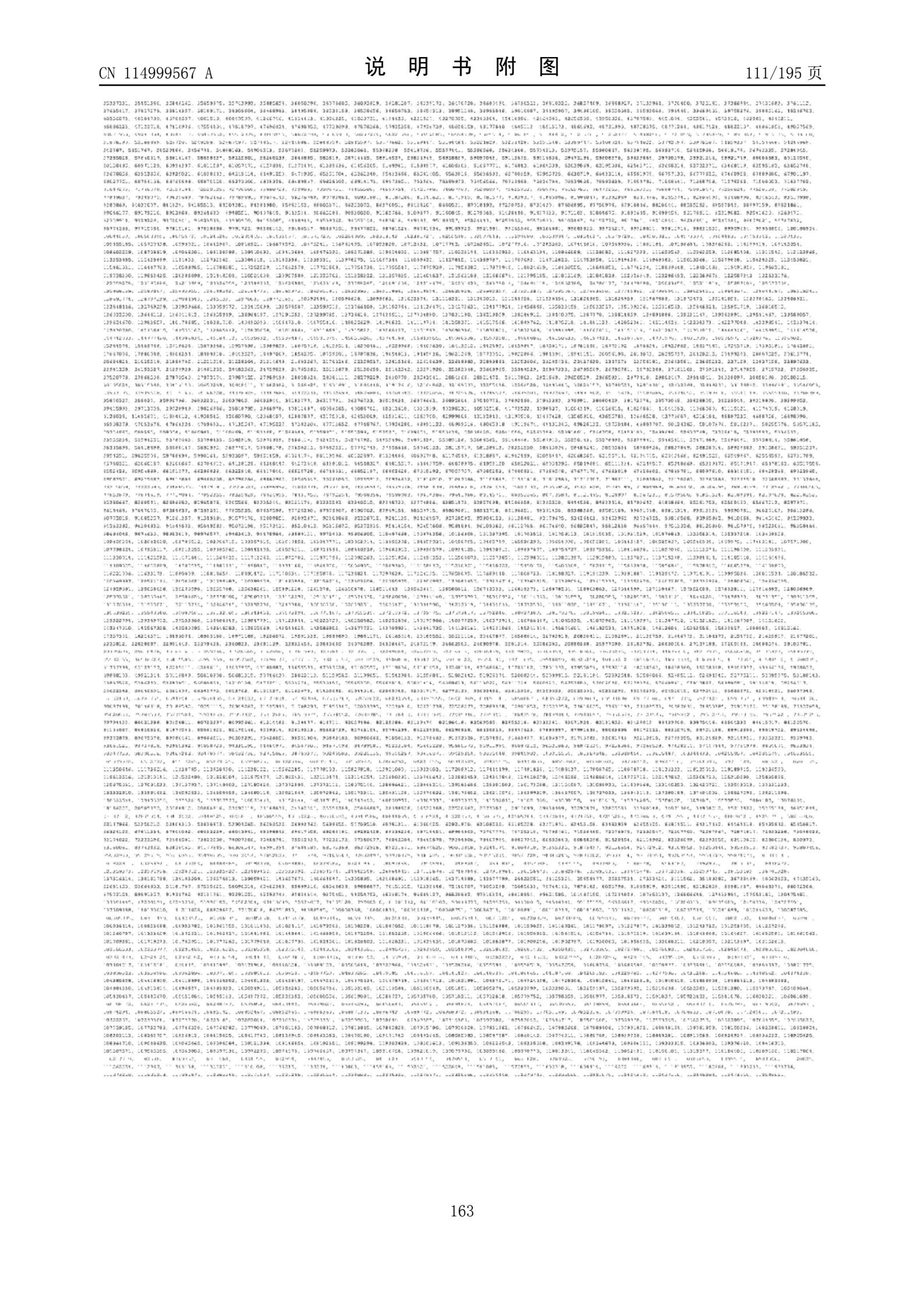


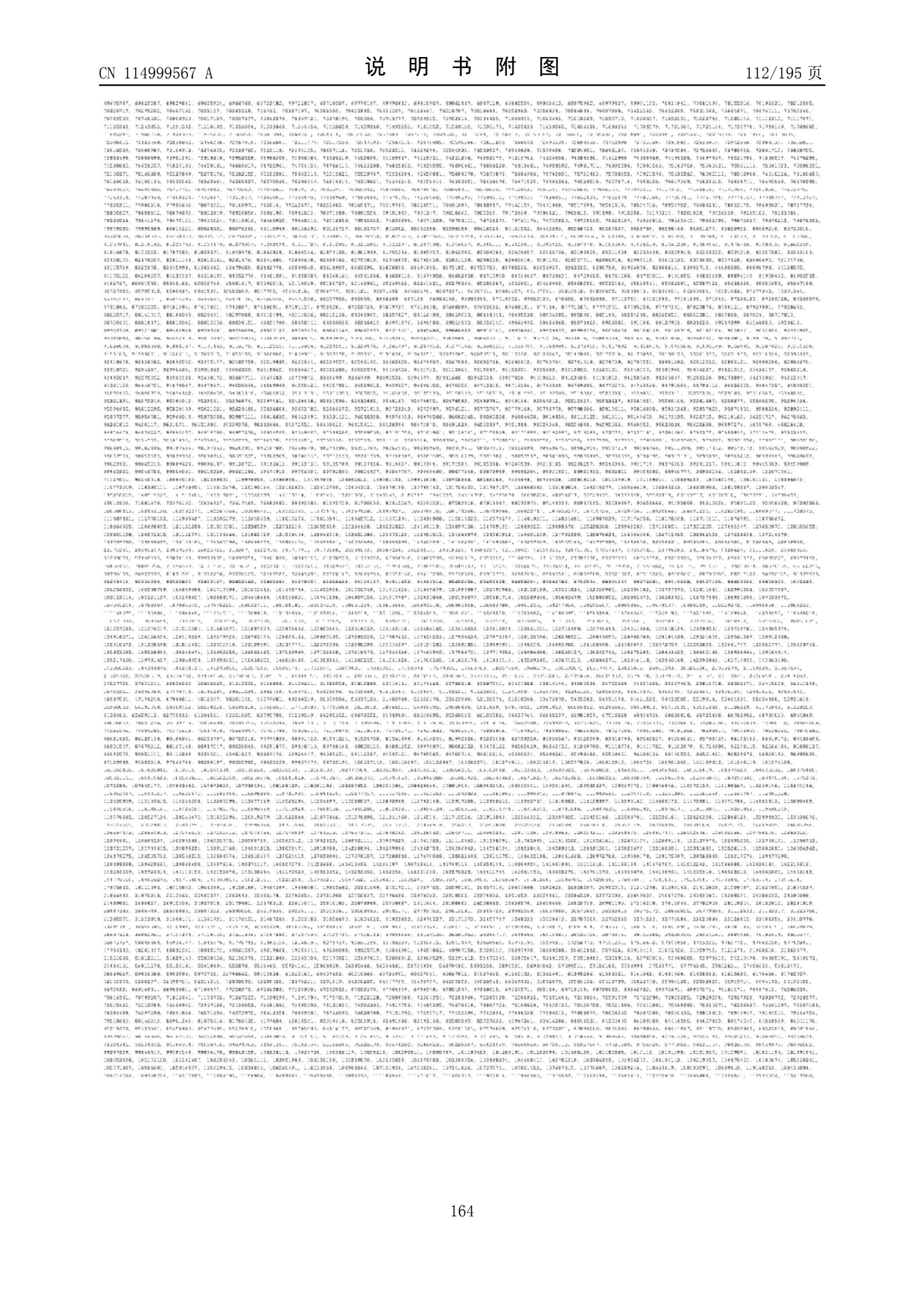


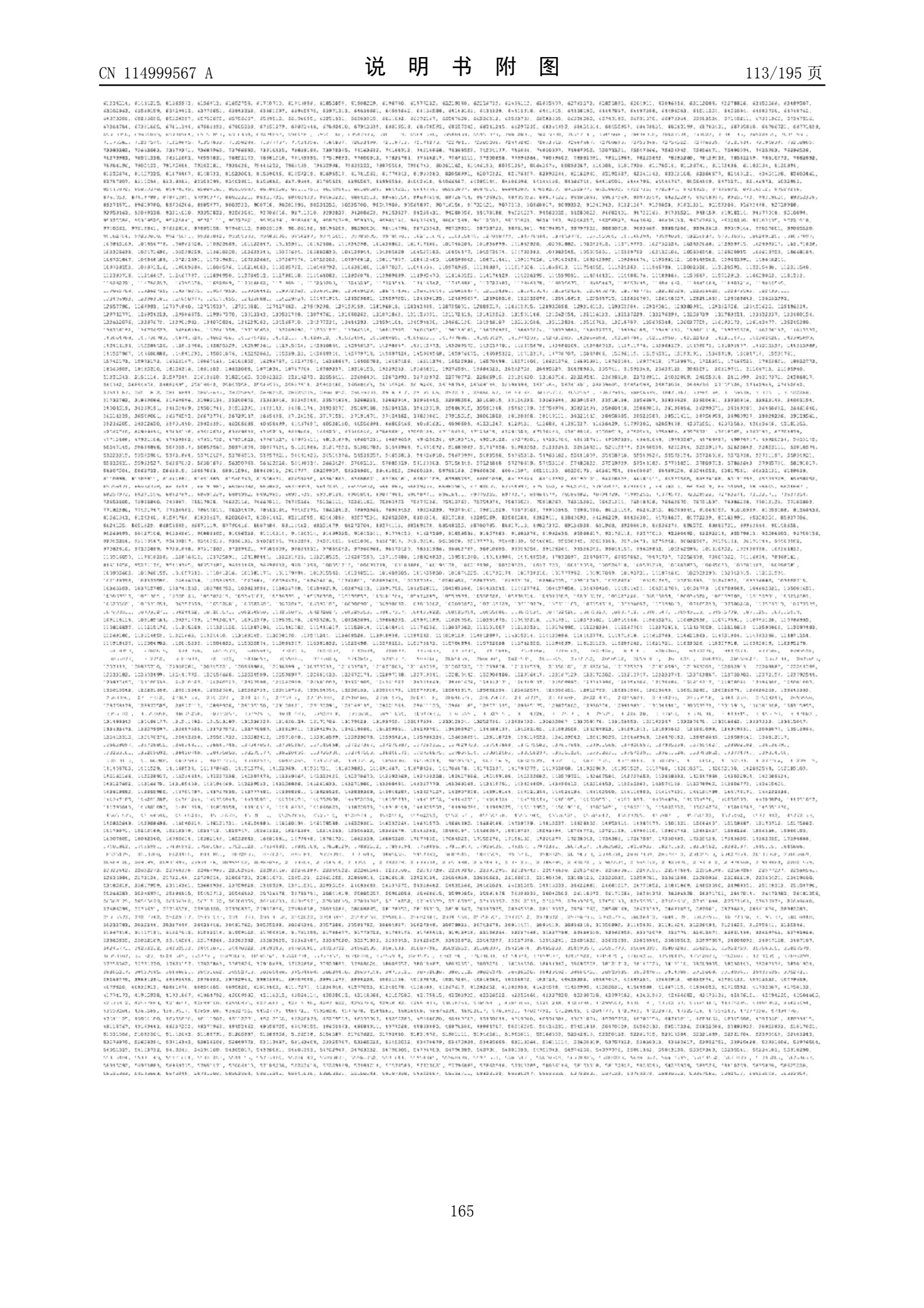


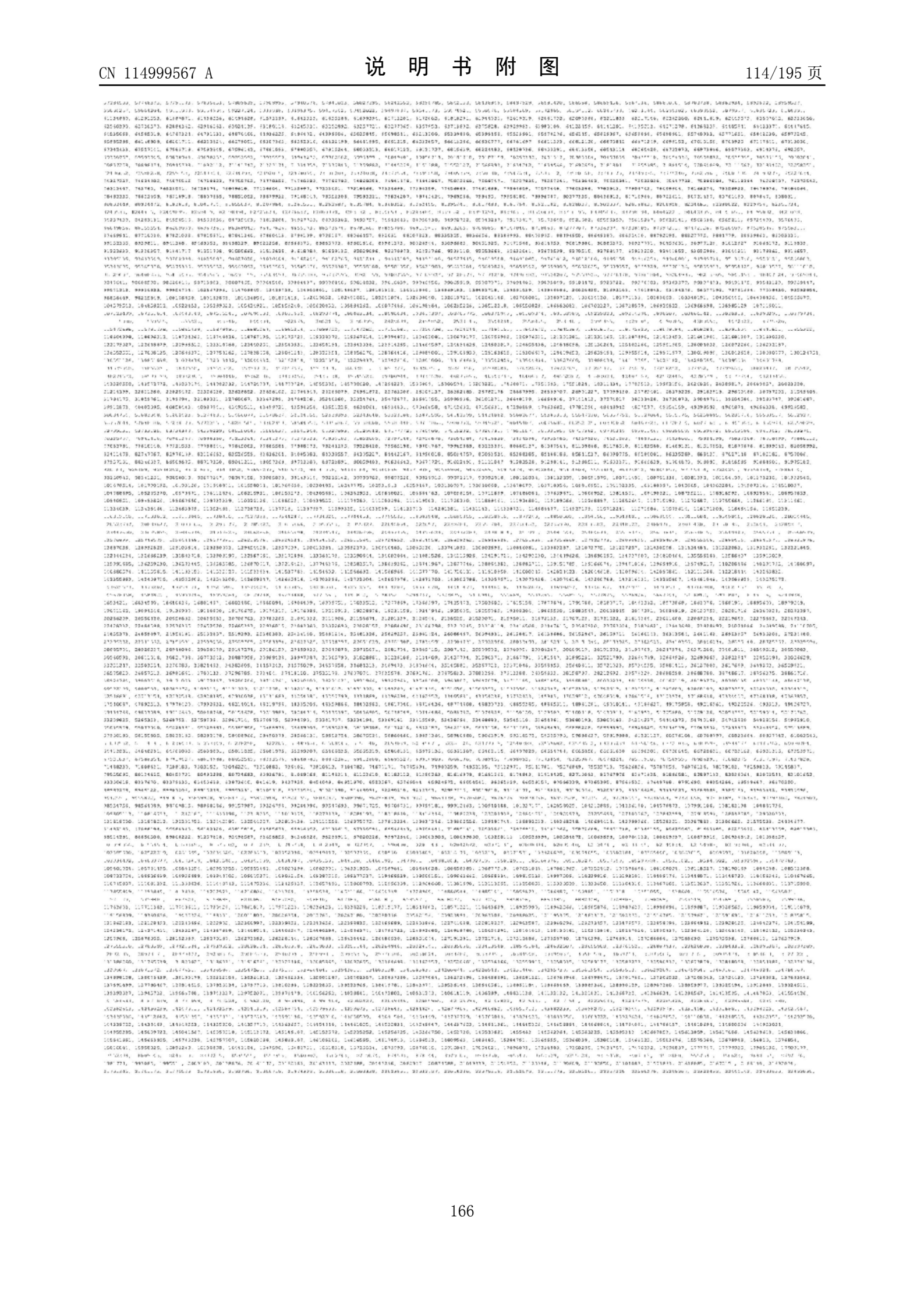


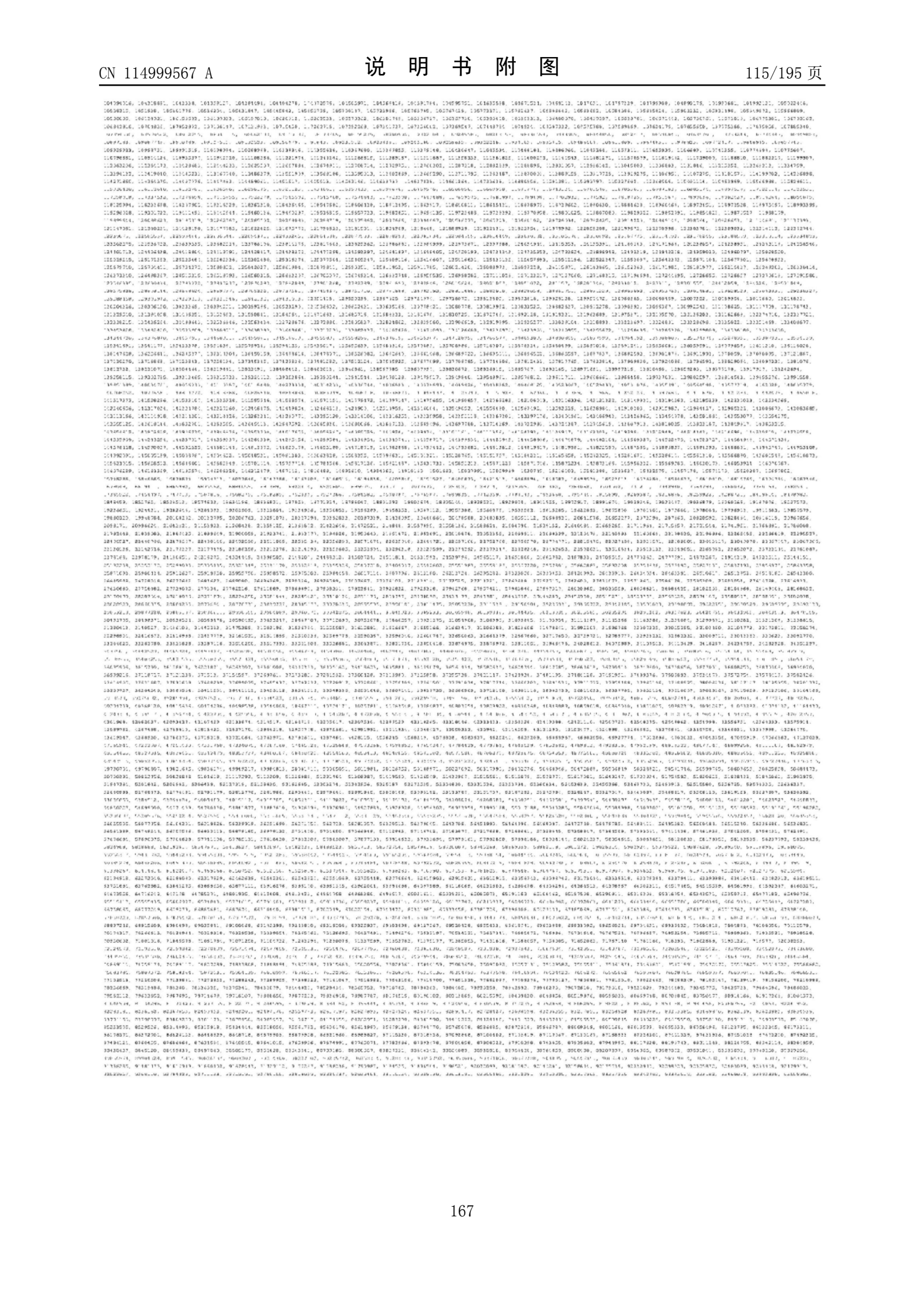


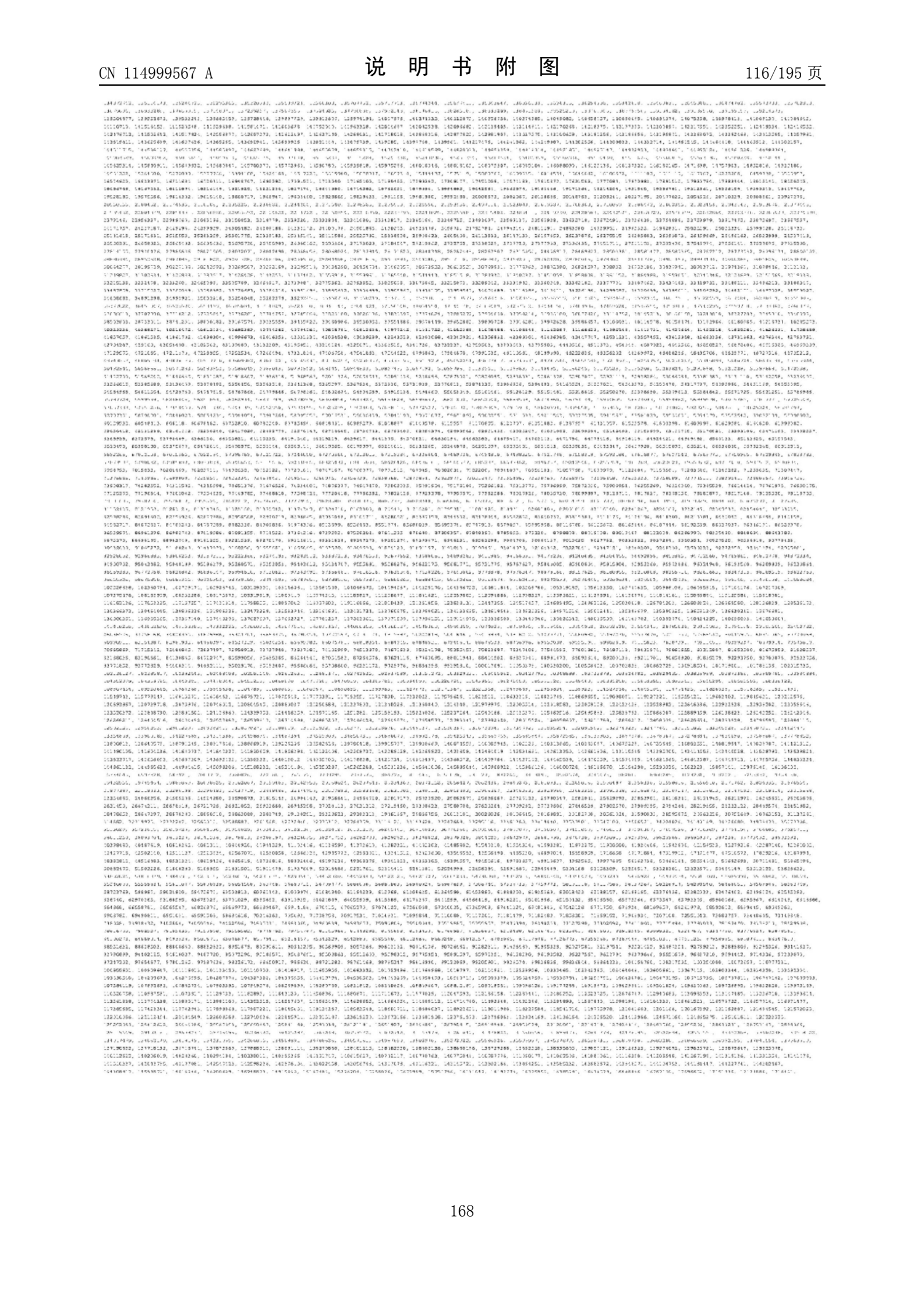


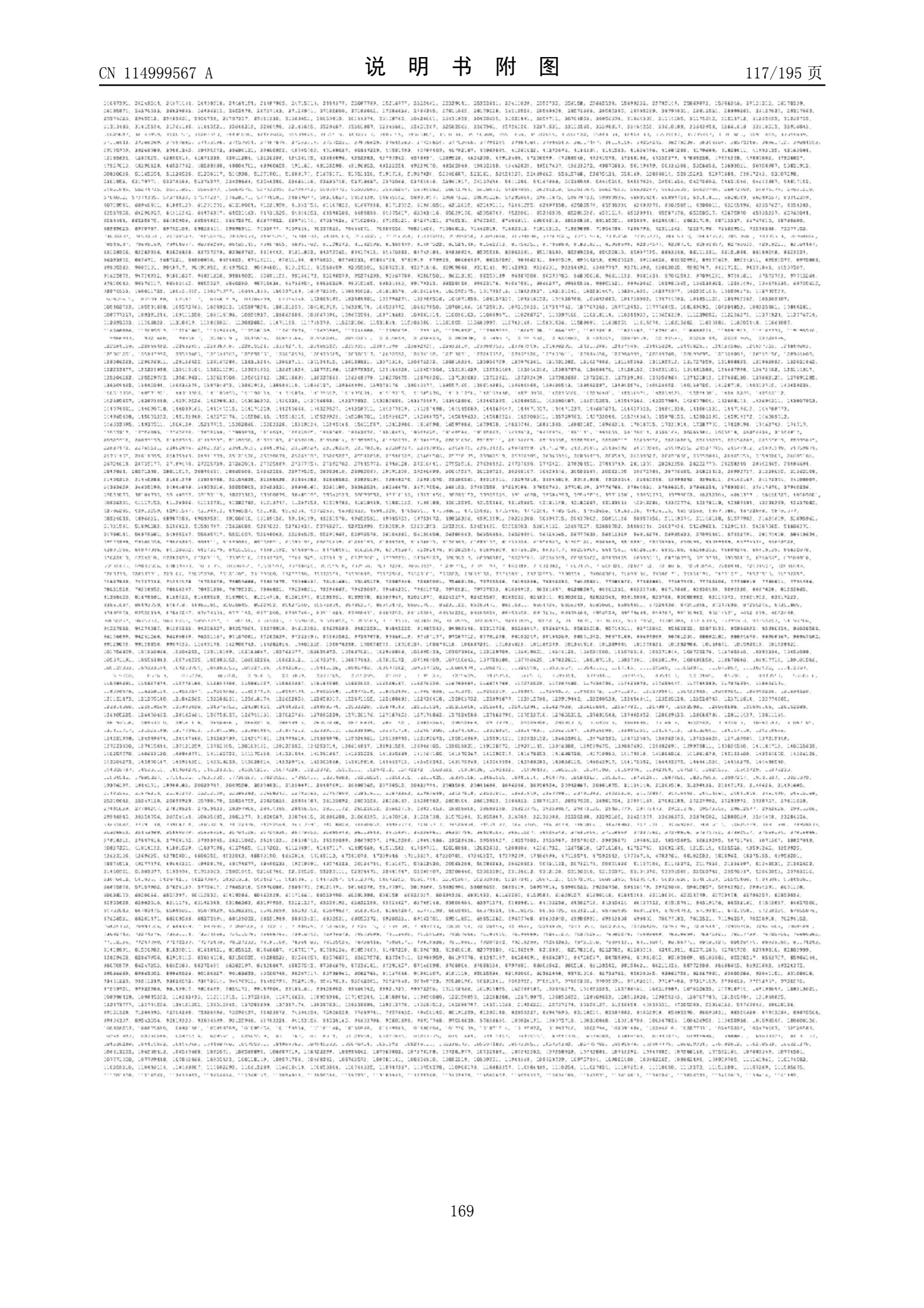












附录B

