



(12) 发明专利

(10) 授权公告号 CN 110870020 B

(45) 授权公告日 2025. 01. 10

(21) 申请号 201880043867.1

(22) 申请日 2018.10.15

(65) 同一申请的已公布的文献号
申请公布号 CN 110870020 A

(43) 申请公布日 2020.03.06

(30) 优先权数据
62/573125 2017.10.16 US
62/573131 2017.10.16 US
62/573135 2017.10.16 US
62/726158 2018.08.31 US

(85) PCT国际申请进入国家阶段日
2019.12.28

(86) PCT国际申请的申请数据
PCT/US2018/055923 2018.10.15

(87) PCT国际申请的公布数据
W02019/079202 EN 2019.04.25

(73) 专利权人 因美纳有限公司
地址 美国加利福尼亚州

(72) 发明人 K·贾加纳坦 K-H·法尔

S·卡拉佐普罗潘吉欧托普卢
J·F·麦克雷

(74) 专利代理机构 北京泛华伟业知识产权代理
有限公司 11280

专利代理师 王勇

(51) Int.Cl.
G16B 40/30 (2006.01)
G16B 20/00 (2006.01)

(56) 对比文件
Yi Zhang et al..DeepSplice: Deep
classification of novel splice junctions
revealed by RNA-seq.《IEEE Xplore》.2017,全
文.

Xueqiu Jian et al..In Silico
Prediction of Deleteriousness for
Nonsynonymous and Splice-Altering Single
Nucleotide Variants in the Human Genome.
《Nucleic Acids Research》.2016,42(22),全
文.

审查员 赵亚丽

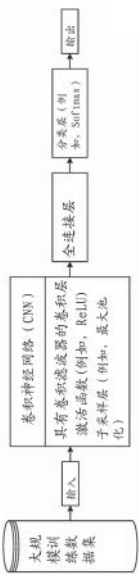
权利要求书4页 说明书87页 附图86页

(54) 发明名称

利用卷积神经网络 (CNNS) 进行异常剪接检测

(57) 摘要

公开的技术涉及构造用于变体分类的基于卷积神经网络的分类器。具体地,本发明涉及使用基于反向传播的梯度更新技术在训练数据上训练基于卷积神经网络的分类器,所述技术将基于卷积神经网络的分类器的输出与相应的真实值标签逐渐匹配。基于卷积神经网络的分类器包括残差块组,每个残差块组由残差块中的卷积滤波器的数量,残差块的卷积窗口大小和残差块的空洞卷积率来参数化,卷积窗口的大小在残差块组之间变化,空洞卷积率在残差块组之间变化。训练数据包括从良性变体和致病性变体产生的良性训练实例和致病性训练实例的转译序列对。



1. 一种用于预测前体mRNA基因组序列中剪接位点的可能性的计算机实现的方法,所述方法包括:

在前体mRNA核苷酸序列的训练实例上训练卷积神经网络CNN,所述训练实例包括至少50000个供体剪接位点的训练实例、至少50000个受体剪接位点的训练实例以及至少100000个非剪接位点的训练实例,其中训练包括:

输入单热编码的前体mRNA核苷酸序列的训练实例,其中每个核苷酸序列包含由至少100个上游上下文核苷酸和至少100个下游上下文核苷酸侧接的至少一个靶核苷酸;以及

通过后向传播来调整所述CNN中的滤波器的参数,以预测所述前体mRNA核苷酸序列中每个靶核苷酸是供体剪接位点、受体剪接位点或非剪接位点的可能性的得分;

从而训练过的CNN被配置为接受前体mRNA核苷酸序列作为输入,所述前体mRNA核苷酸序列是单热编码的并包括每一侧上侧接有至少100个核苷酸的至少一个靶核苷酸,并对每个靶核苷酸是供体剪接位点、受体剪接位点或非剪接位点的可能性进行评分。

2. 根据权利要求1所述的计算机实现的方法,其中,每个靶核苷酸由200个上游上下文核苷酸和200个下游上下文核苷酸侧接。

3. 根据权利要求1所述的计算机实现的方法,进一步包括在至少150000个供体剪接位点的训练实例、150000个受体剪接位点的训练实例以及1000000个非剪接位点的训练实例上训练所述CNN。

4. 根据权利要求1所述的计算机实现的方法,其中,所述CNN通过卷积层的数量、卷积滤波器的数量和子采样层的数量来参数化。

5. 根据权利要求1所述的计算机实现的方法,其中,所述CNN包括一个或多个全连接层和终端分类层。

6. 根据权利要求1所述的计算机实现的方法,其中,所述CNN包括维度改变层,所述维度改变层重塑先前输入的空间和特征维度。

7. 根据权利要求1所述的计算机实现的方法,其中,所述CNN通过残差块的数量、跳跃连接的数量和残差连接的数量来参数化。

8. 根据权利要求7所述的计算机实现的方法,其中,每个残差块包括至少一个批归一化层、至少一个修正线性单元ReLU层、至少一个维度改变层以及至少一个残差连接。

9. 根据权利要求7所述的计算机实现的方法,其中,每个残差块包括两个批归一化层、两个修正线性单元ReLU非线性层、两个维度改变层以及一个残差连接。

10. 根据权利要求1所述的计算机实现的方法,其中,所述CNN在一个时期期间分批评估所述训练实例。

11. 根据权利要求1所述的计算机实现的方法,其中,将所述训练实例随机采样成批,其中,每个批具有预定的批大小。

12. 根据权利要求10所述的计算机实现的方法,其中,所述CNN在至少10个时期内迭代所述训练实例的评估。

13. 根据权利要求1所述的计算机实现的方法,其中,所述输入包括具有两个相邻的靶核苷酸的靶核苷酸序列。

14. 根据权利要求13所述的计算机实现的方法,其中,所述两个相邻的靶核苷酸是缩写为A的腺嘌呤和缩写为G的鸟嘌呤。

15. 根据权利要求13所述的计算机实现的方法, 其中, 所述两个相邻的靶核苷酸是缩写为G的鸟嘌呤和缩写为D的尿嘧啶。

16. 一种用于预测前体mRNA基因组序列中剪接位点的可能性的系统, 包括耦合到存储器的一个或多个处理器, 所述存储器装载有计算机指令, 所述计算机指令在所述一个或多个处理器上执行时, 实施以下动作:

在前体mRNA核苷酸序列的训练实例上训练卷积神经网络CNN, 所述训练实例包括至少50000个供体剪接位点的训练实例、至少50000个受体剪接位点的训练实例以及至少100000个非剪接位点的训练实例, 其中训练包括:

输入单热编码的前体mRNA核苷酸序列的训练实例, 其中每个核苷酸序列包含由至少100个上游上下文核苷酸和至少100个下游上下文核苷酸侧接的至少一个靶核苷酸; 以及

通过后向传播来调整所述CNN中的滤波器的参数, 以预测所述前体mRNA核苷酸序列中的每个靶核苷酸是供体剪接位点、受体剪接位点或非剪接位点的可能性的得分;

从而训练过的CNN被配置为接受前体mRNA核苷酸序列作为输入, 所述前体mRNA核苷酸序列是单热编码的并包括每一侧上侧接有至少100个核苷酸的至少一个靶核苷酸, 并对每个靶核苷酸是供体剪接位点、受体剪接位点或非剪接位点的可能性进行评分。

17. 根据权利要求16所述的系统, 其中, 在150000个供体剪接位点的训练实例、150000个受体剪接位点的训练实例以及1000000个非剪接位点的训练实例上对所述CNN进行训练。

18. 一种用于异常剪接确定的系统, 包括与存储器耦合的并行运行的多个处理器, 所述存储器装载有用于确定异常剪接的计算机指令, 所述系统包括:

在所述多个处理器上运行的用于处理前体mRNA序列的训练过的卷积神经网络CNN, 包括:

输入模块, 其

接收至少101个靶核苷酸的变体序列, 所述靶核苷酸在每一侧上由至少400个核苷酸侧接; 以及

访问包含至少101个靶核苷酸的参考序列, 所述靶核苷酸在每一侧上由至少400个核苷酸侧接;

层, 在所述变体序列和所述参考序列中的靶核苷酸上进行卷积并给所述靶核苷酸分配剪接位点得分, 以给出每个靶核苷酸是供体剪接位点、受体剪接位点或非剪接位点的可能性; 以及

输出模块, 其根据所述变体序列和所述参考序列中靶核苷酸的剪接位点得分的位置差异, 确定所述变体序列中的变体是否引起所述变体序列中任何靶核苷酸的异常剪接并且因此是致病性的。

19. 根据权利要求18所述的系统, 其中, 所述输出模块被配置为当对于至少一个靶核苷酸位置, 所述剪接位点得分的差异高于预定阈值时, 确定为所述变体序列中的变体引起异常剪接。

20. 根据权利要求18所述的系统, 其中, 在所述靶核苷酸之间的任何位置处的剪接位点得分的最大差异低于预定阈值, 进一步被配置为将所述变体序列分类为不引起异常剪接并且因此是良性的。

21. 根据权利要求19-20中任一项所述的系统, 其中, 所述预定阈值由以下确定:

对于多个候选阈值，
处理由良性常见变体产生的第一组参考和变体序列对，以产生第一组异常剪接检测；
以及

处理由致病性罕见变体产生的第二组参考和变体序列对，以产生第二组异常剪接检测；以及

选择至少一个阈值以供所述输出模块使用，所述阈值使所述第二组参考和变体序列对中的异常剪接检测的计数最大化，并使所述第一组参考和变体序列对中的异常剪接检测的计数最小化。

22. 根据权利要求18所述的系统，进一步被配置为识别引起自闭症谱系障碍ASD的变体。

23. 根据权利要求18所述的系统，其进一步被配置为识别引起发育迟缓障碍DDD的变体。

24. 根据权利要求18所述的系统，其中，至少101个连续的靶核苷酸的每个序列在每一侧上由至少1000个核苷酸的扩展上下文侧接。

25. 根据权利要求18所述的系统，其中，所述参考序列中的靶核苷酸的剪接位点得分被编码在所述训练过的CNN的第一输出中，并且所述变体序列中的靶核苷酸的剪接位点得分被编码在所述训练过的CNN的第二输出中。

26. 根据权利要求25所述的系统，其中，所述第一输出被编码为第一 101×3 矩阵，并且所述第二输出被编码为第二 101×3 矩阵。

27. 根据权利要求26所述的系统，其中，所述第一 101×3 矩阵中的每一行唯一地表示剪接位点得分，以给出所述参考序列中的靶核苷酸是供体剪接位点、受体剪接位点或非剪接位点的可能性。

28. 根据权利要求27所述的系统，其中，所述第二 101×3 矩阵中的每一行唯一地表示剪接位点得分，以给出所述变体序列中的靶核苷酸是供体剪接位点、受体剪接位点或非剪接位点的可能性。

29. 根据权利要求26-28中任一项所述的系统，其中，将所述第一 101×3 矩阵和所述第二 101×3 矩阵的每一行中的剪接位点得分指数式地进行归一化并求和为一。

30. 根据权利要求26-28中任一项所述的系统，其中，所述输出模块执行所述第一 101×3 矩阵和所述第二 101×3 矩阵的行到行比较，并且逐行地确定剪接位点得分分布的变化。

31. 根据权利要求30所述的系统，其中，对于所述行到行比较的至少一个实例，对应的分布的变化高于预定阈值，进一步被配置为将所述变体分类为引起异常剪接并且因此是致病性的。

32. 根据权利要求18所述的系统，其中，所述参考序列和所述变体序列是单热编码的。

33. 一种检测异常剪接的方法，所述方法包括：

接受包含至少801个核苷酸的前体mRNA核苷酸序列作为训练过的卷积神经网络CNN的输入，所述前体mRNA核苷酸序列包括在每一侧上由至少400个核苷酸侧接的至少101个靶核苷酸；

通过所述训练过的CNN处理参考序列和变体序列以产生剪接位点得分，以给出所述参考序列和所述变体序列中的每个靶核苷酸是供体剪接位点、受体剪接位点或非剪接位点的

可能性,其中所述参考序列和所述变体序列各自具有在上游和下游、在每一侧上由至少400个核苷酸侧接的至少101个靶核苷酸;以及

根据所述参考序列和所述变体序列中靶核苷酸的剪接位点得分的差异,确定所述变体序列中的变体是否引起异常剪接并且因此是致病性的。

34.一种非暂时性计算机可读存储介质,其上印有用于检测异常剪接的计算机程序指令,所述指令在处理器上执行时,实施包括以下的方法:

接受包含至少801个核苷酸的输入序列中的前体mRNA核苷酸作为训练过的卷积神经网络CNN的输入,所述输入序列包括在每一侧上由至少400个核苷酸侧接的至少101个靶核苷酸;

通过所述训练过的CNN处理参考序列和变体序列以产生剪接位点得分,以给出所述参考序列和所述变体序列中的每个靶核苷酸是供体剪接位点、受体剪接位点或非剪接位点的可能性,其中所述参考序列和所述变体序列各自具有至少101个靶核苷酸,并且每个靶核苷酸在每一侧上由至少400个核苷酸侧接;以及

根据所述参考序列和所述变体序列中靶核苷酸的剪接位点得分的差异,确定所述变体序列中的变体是否引起异常剪接并且因此是致病性的。

35.一种非暂时性计算机可读存储介质,其上存储有计算机程序指令,所述指令在处理器上执行时,实施根据权利要求1-15中任一项所述的计算机实现的方法。

利用卷积神经网络(CNNS)进行异常剪接检测

[0001] 附录

[0002] 附录包括发明人撰写的论文中列出的潜在相关参考文献的书目。本文的主题涵盖在本申请要求其优先权/权益的美国临时专利申请中。这些参考资料可由法律顾问根据要求提供,也可以经由全球档案查阅。

[0003] 优先权申请

[0004] 本申请要求以下的优先权或权益:由Kishore Jaganathan、Kai-How Farh、Sofia Kyriazopoulou Panagiotopoulou和Jeremy Francis McRae于2017年10月16日提交的标题为“基于深度学习的剪接位点分类”的美国临时专利申请No.62/573,125(代理人案卷号:ILLM1001-1/IP-1610-PRV);由Kishore Jaganathan、Kai-How Farh、Sofia Kyriazopoulou Panagiotopoulou和Jeremy Francis McRae于2017年10月16日提交的标题为“基于深度学习的异常剪接检测”的美国临时专利申请No.62/573,131(代理人案卷No.ILLM 1001-2/IP-1614-PRV);由Kishore Jaganathan、Kai-How Farh、Sofia Kyriazopoulou Panagiotopoulou和Jeremy Francis McRae于2017年10月16日提交的标题为“使用卷积神经网络进行异常剪接检测”的美国临时专利申请No.62/573,135(代理人案卷No.ILLM1001-3/IP-1615-PRV);以及由Kishore Jaganathan、Kai-How Farh、Sofia Kyriazopoulou Panagiotopoulou和Jeremy Francis McRae于2018年8月31日提交的标题为“利用深度学习从原始序列预测剪接”的美国临时专利申请No.62/726,158(代理人案卷No.ILLM1001-10/IP-1749-PRV)。出于所有目的,所述临时申请特此以引用的方式并入本文中。

[0005] 并入

[0006] 出于所有目的,以下内容通过引用并入本文,如同在本文中进行了完整阐述一样:

[0007] 由Kishore Jaganathan、Kai-How Farh、Sofia Kyriazopoulou Panagiotopoulou和Jeremy Francis McRae于2018年10月15日同时提交的标题为“基于深度学习的剪接位点分类”的PCT专利申请号PCT/US18/55915(代理案卷号ILLM 1001-7/IP-1610-PCT),随后被公布为PCT公开号WO_____。

[0008] 由Kishore Jaganathan、Kai-How Farh、Sofia Kyriazopoulou Panagiotopoulou和Jeremy Francis McRae于2018年10月15日同时提交的标题为“基于深度学习的异常剪接检测”的PCT专利申请号PCT/US18/55919(代理案卷号ILLM 1001-8/IP-1614-PCT),随后被公布为PCT公开号WO_____。

[0009] 由Kishore Jaganathan、Kai-How Farh、Sofia Kyriazopoulou Panagiotopoulou和Jeremy Francis McRae同时提交的标题为“基于深度学习的剪接位点分类”的美国非临时专利申请(代理人案卷No.ILLM1001-4/IP-1610-US)。

[0010] 由Kishore Jaganathan、Kai-How Farh、Sofia Kyriazopoulou Panagiotopoulou和Jeremy Francis McRae同时提交的标题为“基于深度学习的异常剪接检测”的美国非临时专利申请(代理人案卷No.ILLM 1001-5/IP-1614-US)。

[0011] 由Kishore Jaganathan、Kai-How Farh、Sofia Kyriazopoulou Panagiotopoulou和Jeremy Francis McRae同时提交的标题为“使用卷积神经网络(CNN)的异常剪接检测”的

美国非临时专利申请(代理人案卷No.ILLM 1001-6/IP-1615-US)。

[0012] 文献1-S.Dieleman,H.Zen,K.Simonyan,O.Vinyals,A.Graves,N.Kalchbrenner,A.Senior,and K.Kavukcuoglu,“WAVENET:A GENERATIVE MODEL FOR RAW AUDIO,”arXiv:1609.03499,2016;

[0013] 文献2-S.Ö.Arik,M.Chrzanowski,A.Coates,G.Diamos,A.Gibiansky,Y.Kang,X.Li,J.Miller,A.Ng,J.Raiman,S.Sengupta and M.Shoebybi,“DEEP VOICE:REAL-TIME NEURAL TEXT-TO-SPEECH,”arXiv:1702.07825,2017;

[0014] 文献3-F.Yu and V.Koltun,“MULTI-SCALE CONTEXT AGGREGATION BY DILATED CONVOLUTIONS,”arXiv:1511.07122,2016;

[0015] 文献4-K.He,X.Zhang,S.Ren,and J.Sun,“DEEP RESIDUAL LEARNING FOR IMAGE RECOGNITION,”arXiv:1512.03385,2015;

[0016] 文献5-R.K.Srivastava,K.Greff,and J.Schmidhuber,“HIGHWAY NETWORKS,”arXiv:1505.00387,2015;

[0017] 文献6-G.Huang,Z.Liu,L.van der Maaten and K.Q.Weinberger,“DENSELY CONNECTED CONVOLUTIONAL NETWORKS,”arXiv:1608.06993,2017;

[0018] 文献7-C.Szegedy,W.Liu,Y.Jia,P.Sermanet,S.Reed,D.Anguelov,D.Erhan,V.Vanhoudke,and A.Rabinovich,“GOING DEEPER WITH CONVOLUTIONS,”arXiv:1409.4842,2014;

[0019] 文献8-S.Ioffe and C.Szegedy,“BATCH NORMALIZATION:ACCELERATING DEEP NETWORK TRAINING BY REDUCING INTERNAL COVARIATE SHIFT,”arXiv:1502.03167,2015;

[0020] 文献9-J.M.Wolterink,T.Leiner,M.A.Viergever,and I. Išgum,“DILATED CONVOLUTIONAL NEURAL NETWORKS FOR CARDIOVASCULAR MR SEGMENTATION IN CONGENITAL HEART DISEASE,”arXiv:1704.03669,2017;

[0021] 文献10-L.C.Piqueras,“AUTOREGRESSIVE MODEL BASED ON ADEEP CONVOLUTIONAL NEURAL NETWORK FOR AUDIO GENERATION,”Tampere University of Technology,2016;

[0022] 文献11-J.Wu,“Introduction to Convolutional Neural Networks,”Nanjing University,2017;

[0023] 文献12-I.J.Goodfellow,D.Warde-Farley,M.Mirza,A.Courville,and Y.Bengio,“CONVOLUTIONAL NETWORKS”,Deep Learning,MIT Press,2016;以及

[0024] 文献13-J.Gu,Z.Wang,J.Kuen,L.Ma,A.Shahrudy,B.Shuai,T.Liu,X.Wang,和 G.Wang,“RECENT ADVANCES IN CONVOLUTIONAL NEURAL NETWORKS,”arXiv:1512.07108,2017。

[0025] 文献1描述了深度卷积神经网络架构,其使用具有相同卷积窗口尺寸的卷积滤波器的残差块的组、批归一化层、修正线性单元(缩写ReLU)层、维度改变层、具有指数增长的空洞卷积率的空洞卷积层、跳跃连接和softmax分类层来接受输入序列并产生对输入序列中的条目进行评分的输出序列。由所公开的技术使用文献1中描述的神经网络部件和参数。在一个实现方式中,所公开的技术修改了文献1中描述的神经网络部件的参数。例如,不同于文献1,所公开的技术中的空洞卷积率从较低残差块组到较高残差块组非指数地演进。在

另一个示例中,不同于文献1,所公开的技术中的卷积窗口尺寸在残差块的组之间变化。

[0026] 文献2描述了在文献1中描述的深度卷积神经网络架构的细节。

[0027] 文献3描述了由所公开的技术使用的空洞卷积。如本文中所使用的,空洞卷积也称为“膨胀卷积”。空洞/膨胀卷积允许具有很少的可训练参数的大感受野。空洞/膨胀卷积是指通过以一定步长跳过输入值(也称为空洞卷积率或膨胀因子)将内核应用于大于其长度的区域的卷积。空洞/膨胀卷积增加了卷积滤波器/内核的元素之间的间距,使得当执行卷积操作时,以较大的间距考虑相邻的输入条目(例如,核苷酸、氨基酸)。这使得能够在输入中并入长距离上下文相关性。空洞卷积保留部分卷积计算,以便在处理相邻核苷酸时重复使用。

[0028] 文献4描述了由所公开的技术使用的残差块和残差连接。

[0029] 文献5描述了由所公开的技术使用的跳跃连接。如本文所使用的,跳跃连接也称为“公路网络”。

[0030] 文献6描述了由所公开的技术使用的密集连接的卷积网络架构。

[0031] 文献7描述了由所公开的技术使用的维度改变卷积层和基于块的处理流水线。维度改变卷积的一个示例是 1×1 卷积。

[0032] 文献8描述了由所公开的技术使用的批归一化层。

[0033] 文献9还描述了由所公开的技术使用的空洞/膨胀卷积。

[0034] 文献10描述了可由所公开的技术使用的深度神经网络的各种架构,其包括卷积神经网络、深度卷积神经网络和具有空洞/膨胀卷积的深度卷积神经网络。

[0035] 文献11描述了可由所公开的技术使用的卷积神经网络的细节,其包括用于训练具有子采样层(例如,池化)和全连接层的卷积神经网络的算法。

[0036] 文献12描述了可由所公开的技术使用的各种卷积操作的细节。

[0037] 文献13描述了可由所公开的技术使用的卷积神经网络的各种架构。

[0038] 通过引用以电子方式提交申请的表进行并入

[0039] 以ASCII文本格式的三个表文件与本申请一起提交并以引用的方式并入。文件的名称、创建日期和大小为:

[0040] 表_S4_突变_率.txt 2018年8月31日2,452KB

[0041] 表_S5_基因_富集.txt 2018年8月31日362KB

[0042] 表_S6_验证.txt 2018年8月31日362KB

技术领域

[0043] 所公开的技术涉及人工智能型计算机和数字数据处理系统以及相对应的用于智能仿真的数据处理方法和产品(即,基于知识的系统、推理系统和知识获取系统);并且包括用于不确定性推理的系统(例如模糊逻辑系统)、自适应系统、机器学习系统和人工神经网络。特别地,所公开的技术涉及使用基于深度学习的技术来训练深度卷积神经网络。

背景技术

[0044] 本部分中讨论的主题不应仅由于在本部分中提及而被认为是现有技术。类似地,在本部分中提及的或与作为背景提供的主题相关联的问题不应被假定为先前已经在现有

技术中被认识到。本部分中的主题仅表示不同的方法,这些方法本身也可以对应于所要求保护的技术的实现方式。

[0045] 机器学习

[0046] 在机器学习中,输入变量用于预测输出变量。输入变量通常称为特征并且由 $X = (X_1, X_2, \dots, X_k)$ 表示,其中每个 $X_i, i \in 1, \dots, k$ 是特征。输出变量通常称为响应变量或因变量,并且由变量 Y_i 表示。 Y 和相对应的 X 之间的关系可以用一般形式写成:

$$[0047] \quad Y = f(X) + \epsilon$$

[0048] 在上面的等式中, f 是特征 $(X_1, X_2, \dots, X_k)$ 的函数并且 ϵ 是随机误差项。误差项独立于 X 且具有零的平均值。

[0049] 在实践中,特征 X 是可用的而无需 Y 或无需知晓 X 和 Y 之间的确切关系。因为误差项的平均值为零,所以目标是估计值 f 。

$$[0050] \quad \hat{Y} = \hat{f}(X)$$

[0051] 在上面的等式中, $\hat{f}$ 是 ϵ 的估计值,通常被认为是一个黑盒,这意味着仅 $\hat{f}$ 的输入和输出之间的关系是已知的,但是它为什么工作的问题仍然没有答案。

[0052] 使用学习来找到该函数 $\hat{f}$ 。监督学习和无监督学习是机器学习中用于此任务的两种方式。在监督学习中,标签的数据用于训练。通过示出输入和相对应的输出(=标签),函数 $\hat{f}$ 被优化,使得它接近输出。在无监督学习中,目标是从未标签的数据中找到隐藏的结构。该算法对输入数据的准确度没有任何度量,这使其与监督学习有所差异。

[0053] 神经网络

[0054] 单层感知器(SLP)是神经网络的最简单模型。如图1所示,它包括一个输入层和一个激活函数。输入通过加权图传递。函数 f 使用输入的和作为自变量,并将其与阈值 θ 进行比较。

[0055] 图2示出了具有多层的全连接的神经网络的一个实现方式。神经网络是在彼此之间交换消息的互连的人工神经元(例如, a_1, a_2, a_3)的系统。所示的神经网络具有三个输入,隐藏层中的两个神经元和输出层中的两个神经元。隐藏层具有激活函数 $f(\cdot)$ 并且输出层具有激活函数 $g(\cdot)$ 。连接具有在训练过程期间被调谐的数字权重(例如, $w_{11}, w_{21}, w_{12}, w_{31}, w_{22}, w_{32}, v_{11}, v_{22}$),使得正确训练的网络在馈送图像时正确地响应以识别。输入层处理原始输入,隐藏层基于输入层与隐藏层之间的连接的权重来处理来自输入层的输出。输出层从隐藏层获取输出且基于隐藏层与输出层之间的连接的权重来处理所述输出。该网络包括多层特征检测神经元。每层具有响应于来自先前层的输入的不同组合的许多神经元。这些层被构造成使得第一层检测输入图像数据中的一组原始图案,第二层检测图案的图案,并且第三层检测这些图案的图案。

[0056] 深度学习在基因组学中的应用的调查可以在以下出版物中找到:

[0057] T.Ching et al., Opportunities And Obstacles For Deep Learning In Biology And Medicine, www.biorxiv.org/142760, 2017;

[0058] Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep Learning For Computational Biology. *Mol Syst Biol*. 2016;12:878;

- [0059] Park Y, Kellis M. 2015 Deep Learning For Regulatory Genomics. Nat. Biotechnol. 33, 825-826. (doi:10.1038/nbt.3313);
- [0060] Min, S., Lee, B. & Yoon, S. Deep Learning In Bioinformatics. Brief. Bioinform. bbw068 (2016);
- [0061] Leung MK, DeLong A, Alipanahi B et al. Machine Learning In Genomic Medicine: A Review of Computational Problems and Data Sets 2016; and Libbrecht MW, Noble WS. Machine Learning Applications In Genetics and Genomics. Nature Reviews Genetics 2015; 16(6): 321-32.

附图说明

[0062] 在附图中, 贯穿不同的视图, 相同的附图标记通常指代相同的部分。此外, 附图不一定按比例绘制, 而是总体上将重点放在说明所公开的技术的原理上。在以下描述中, 参考以下附图来描述所公开的技术的各种实现方式, 在附图中:

- [0063] 图1示出了单层感知器 (SLP)。
- [0064] 图2示出了具有多层的前馈神经网络的一个实现方式。
- [0065] 图3描绘了卷积神经网络的工作的一个实现方式。
- [0066] 图4描绘了根据所公开的技术的一个实现方式的训练卷积神经网络的框图。
- [0067] 图5示出了根据所公开的技术的一个实现方式的ReLU非线性层的一个实现方式。
- [0068] 图6示出了膨胀卷积。
- [0069] 图7是根据所公开的技术的一个实现方式的子采样层 (平均/最大池化) 的一个实现方式。
- [0070] 图8描绘了卷积层的两层卷积的一个实现方式。
- [0071] 图9描绘了经由特征映射图添加在下游重新注入初始信息的残差连接。
- [0072] 图10描绘了残差块和跳跃连接的一个实现方式。
- [0073] 图11示出了堆栈的膨胀卷积的一个实现方式。
- [0074] 图12示出了批归一化前向传递。
- [0075] 图13示出了在测试时间的批归一化变换。
- [0076] 图14示出了批归一化后向传递。
- [0077] 图15描绘了具有卷积或密集连接层的批归一化层的使用。
- [0078] 图16示出了1D卷积的一个实现方式。
- [0079] 图17示出了全局平均池化 (GAP) 是如何工作的。
- [0080] 图18示出了具有可以用于实现所公开的技术的训练服务器和生产服务器的计算环境的一个实现方式。
- [0081] 图19描绘了空洞卷积神经网络 (缩写为ACNN) 的架构的一个实现方式, 在本文中称为“SpliceNet”。
- [0082] 图20示出了可由ACNN和卷积神经网络 (缩写为CNN) 使用的残差块的一个实现方式。
- [0083] 图21描绘了ACNN的架构的另一个实现方式, 在本文中称为“SpliceNet80”。
- [0084] 图22描绘了ACNN的架构的又一实现方式, 在本文中称为“SpliceNet400”。

- [0085] 图23描绘了ACNN的架构的又一实现方式,在本文中称为“SpliceNet2000”。
- [0086] 图24描绘了ACNN的架构的又一实现方式,在本文中称为“SpliceNet10000”。
- [0087] 图25、图26和图27示出了由ACNN和CNN处理的各种类型的输入。
- [0088] 图28示出了ACNN可以在至少800万个非剪接位点上被训练,并且CNN可以在至少1百万个非剪接位点上被训练。
- [0089] 图29示出了单热(one-hot)编码器。
- [0090] 图30描绘了ACNN的训练。
- [0091] 图31示出了CNN。
- [0092] 图32示出了ACNN和CNN的训练,验证和测试。
- [0093] 图33描绘了参考序列和替代序列。
- [0094] 图34示出了异常剪接检测。
- [0095] 图35示出了用于剪接位点分类的SpliceNet10000的处理金字塔。
- [0096] 图36描绘了用于异常剪接检测的SpliceNet10000的处理金字塔。
- [0097] 图37A、图37B、图37C、图37D、图37E、图37F、图37G和图37H示出了利用深度学习从原始序列预测剪接的一个实现方式。
- [0098] 图38A、图38B、图38C、图38D、图38E、图38F和图38G描绘了RNA测序数据中的罕见隐秘剪接突变的验证的一个实现方式。
- [0099] 图39A、图39B和图39C示出了经常创建组织特异性替代剪接的隐秘剪接变体的一个实现方式。
- [0100] 图40A、图40B、图40C、图40D和图40E描绘了预测的隐秘剪接变体在人类群体中强烈有害的一个实现方式。
- [0101] 图41A、图41B、图41C、图41D、图41E和图41F示出了在患有罕见遗传疾病的患者中新生隐秘剪接突变的一个实现方式。
- [0102] 图42A和图42B描绘了对LincRNA上的各种剪接预测算法的评估。
- [0103] 图43A和图43B示出TACTAAC分支点和GAAGAA外显子-剪接增强子基序的位置相关效果。
- [0104] 图44A和图44B描绘了核小体定位对剪接的影响。
- [0105] 图45示出了计算具有复杂效果的剪接破坏变体的效果大小的示例。
- [0106] 图46A、图46B和图46C示出了在单例和常见变体上对SpliceNet-10k模型的评估。
- [0107] 图47A和图47B描绘了由变体的位置拆分的剪接位点创建变体的验证率和效果大小。
- [0108] 图48A、图48B、图49C和图49D描绘了对训练和测试染色体上的SpliceNet-10k模型的评估。
- [0109] 图49A、图49B和图49C仅从同义、内含子或非平移区位点示出了患有罕见遗传病的患者的新生隐秘剪接突变。
- [0110] 图50A和图50B示出了ASD中的隐秘剪接新生突变以及作为致病性DNM的比例。
- [0111] 图51A、图51B、图51C、图51D、图51E、图51F、图51G、图51H、图51I和图51J描绘了ASD患者中的预测的隐秘剪接新生突变的RNA测序验证。
- [0112] 图52A和图52B示出了仅在规范转录本上训练的模型的RNA测序的验证率和灵敏

度。

[0113] 图53A、图53B和图53C示出了总体建模改善了SpliceNet-10k性能。

[0114] 图54A和图54B示出了在不同外显子密度的区域中的SpliceNet-10k的评估。

[0115] 图55是表S1,其描绘了用于展示效果大小计算和组织特异性剪接的GTEx采样的一个实现方式。

[0116] 图56是表S2,其描绘了用于评估不同算法的验证率和灵敏度的临界值的一个实现方式。

[0117] 图57示出了每基因富集分析的一个实现方式。

[0118] 图58示出了全基因组富集分析的一个实现方式。

[0119] 图59是可用于实现所公开的技术的计算机系统的简化框图。

具体实施方式

[0120] 给出以下讨论以使本领域任何技术人员能够制造和使用所公开的技术,并且在特定应用及其要求的上下文中提供以下讨论。对所公开的实现方式的各种修改对于本领域技术人员来说将是显而易见的,并且在不脱离所公开的技术的精神和范围的情况下,本文定义的一般原理可以应用于其他实现方式和应用。由此,所公开的技术并非意图被限定于所示出的实现方式,而是应被赋予与本文所公开的原理和特征一致的最广范围。

[0121] 介绍

[0122] 卷积神经网络

[0123] 卷积神经网络是一种特殊类型的神经网络。密集连接层和卷积层之间的基本差异是:密集层在其输入特征空间中学习全局图案,而卷积层学习局部图案:在图像的情况下,在输入的小2D窗口中找到的图案。这一关键特性给卷积神经网络带来了两个有趣的特性:(1)他们学习的图案是平移不变的,(2)他们可以学习图案的空间层次。

[0124] 关于第一个,在学习了图片右下角的某个图案之后,卷积层可以在任何地方识别它:例如,在左上角。如果密集连接的网络出现在新的位置,它将不得不重新学习这种图案。这使得卷积神经网络数据高效,因为它们需要较少的训练采样来学习它们具有泛化能力的表示。

[0125] 关于第二个,第一卷积层可以学习小的局部图案,诸如边缘,第二卷积层将学习由第一层的特征构成的较大图案,等等。这允许卷积神经网络有效地学习日益复杂和抽象的视觉概念。

[0126] 卷积神经网络通过将布置在许多不同层中的人工神经元的层与使这些层相关的激活函数互连来学习高度非线性的映射。它包括散布有一个或多个子采样层和非线性层的一个或多个卷积层,这些层典型地跟随有一个或多个全连接层。卷积神经网络的每个元素从先前层中的一组特征接收输入。卷积神经网络同时学习,因为同一特征映射图中的神经元具有相同的权重。这些局部共享权重降低了网络的复杂性,使得当多维输入数据进入网络时,卷积神经网络避免了特征提取和回归或分类过程中的数据重构的复杂性。

[0127] 卷积在具有两个空间轴(高度和宽度)以及深度轴(也称为通道轴)的3D张量(称为特征映射图)上操作。对于RGB图像,深度轴的维度是3,因为图像具有三个颜色通道;红色、绿色和蓝色。对于黑白图像,深度为1(灰度级)。卷积操作从其输入特征映射图提取子块,并

将相同的变换应用于所有这些子块,从而产生输出特征映射图。此输出特征映射图仍然是3D张量:它具有宽度和高度。其深度可以是任意的,因为输出深度是层的参数,并且该深度轴上的不同通道不再代表如在RGB输入中的具体颜色;相反,它们代表滤波器。滤波器对输入数据的具体方面进行编码:例如,在高度级别上,单个滤波器可以对“输入中存在人脸”的概念进行编码。

[0128] 例如,第一卷积层获取大小为(28,28,1)的特征映射图,并输出大小为(26,26,32)的特征映射图:它在其输入上计算32个滤波器。这些32个输出通道中的每一个都包含26x26个值的网格,其是滤波器在输入上的响应映射,指示该滤波器图案在输入中的不同位置处的响应。这就是术语“特征映射图”的含义:深度轴中的每个维度都是特征(或滤波器),并且2D张量输出[:, :, n]是此滤波器在输入上的响应的2D空间映射。

[0129] 卷积由两个关键参数定义:(1)从输入中提取的子块的大小-这些通常是1x1、3x3或5x5;以及(2)输出特征映射图的深度,是由卷积计算的滤波器的数量。通常,这些以32的深度开始,继续到64的深度,并且以128或256的深度终止。

[0130] 卷积通过在3D输入特征映射图上滑动大小为3x3或5x5的这些窗口、在每个位置处停止、以及提取周围特征的3D子块(形状(窗口_高度、窗口_宽度、输入_深度))来工作。每个此类3D子块然后被(经由具有相同学习权重的矩阵的张量乘积,称为卷积内核)变换为形状的1D向量(输出_深度)。然后将所有这些向量在空间上重新组合成形状(高度、宽度、输出_深度)的3D输出映射。输出特征映射图中的每个空间位置对应于输入特征映射图中的相同位置(例如,输出的右下角包含关于输入的右下角的信息)。例如,对于3x3窗口,向量输出[i, j, :]来自3D子块输入[i-1:i+1, j-1:j+1, :]。在图3中详细描述了整个过程。

[0131] 卷积神经网络包括卷积层,所述卷积层在输入值和卷积滤波器(权重的矩阵)之间执行卷积操作,所述卷积滤波器在训练期间通过多次梯度更新迭代学习。让(m,n)为滤波器大小,并且w为权重的矩阵,然后卷积层通过计算点积 $W \cdot x + b$ 来执行W与输入X的卷积,其中x是X的实例,并且b是偏差。卷积滤波器在输入上滑动的步长称为步幅,并且滤波器区域(m x n)称为感受野。相同的卷积滤波器应用于输入的不同位置,这减少了所学习的权重的数量。它还允许位置不变学习,即,如果输入中存在重要图案,则卷积滤波器无论在序列中的何处都可以学习它。

[0132] 训练卷积神经网络

[0133] 图4描绘了根据所公开的技术的一个实现方式的训练卷积神经网络的框图。调整或训练卷积神经网络,以使输入数据导致具体的输出估计值。使用基于输出估计值与真实值(ground truth)的比較的后向传播来调整卷积神经网络,直到输出估计值渐进地匹配或接近真实值为止。

[0134] 通过基于真实值与实际输出之间的差异来调整神经元之间的权重来训练卷积神经网络。这在数学上被描述为:

$$[0135] \quad \Delta w_i = x_i \delta$$

[0136] 其中 $\delta = (\text{真实值}) - (\text{实际输出})$

[0137] 在一个实现方式中,训练规则被定义为:

$$[0138] \quad w_{nm} \leftarrow w_{nm} + \alpha(t_m - \varphi_m) a_n$$

[0139] 在上面的等式中:所述箭头指示所述值的更新; t_m 是神经元m的目标值; φ_m 是所计

算的神经元m的当前输出; a_n 是输入n;以及 α 是学习率。

[0140] 训练中的中间步骤包括使用卷积层从输入数据生成特征向量。计算关于每层中的权重(在输出处开始)的梯度。这称为后向传递或后退。使用负梯度和先前权重的组合来更新网络中的权重。

[0141] 在一个实现方式中,卷积神经网络使用随机梯度更新算法(诸如ADAM),该算法借助于梯度下降来执行误差的后向传播。下面描述了基于S形函数的后向传播算法的一个示例:

$$[0142] \quad \varphi = f(h) = \frac{1}{1 + e^{-h}}$$

[0143] 在上面的S形函数中,是由神经元计算的加权和。该S形函数具有以下导数:

$$[0144] \quad \frac{\partial \varphi}{\partial h} = \varphi(1 - \varphi)$$

[0145] 该算法包括计算网络中的所有神经元的激活,产生用于前向传递的输出。隐藏层中的神经元m的激活被描述为:

$$[0146] \quad \varphi_m = \frac{1}{1 + e^{-h_m}}$$

$$[0147] \quad h_m = \sum_{n=1}^N a_n w_{nm}$$

[0148] 对所有隐藏层执行此操作以获取激活,描述如下:

$$[0149] \quad \varphi_k = \frac{1}{1 + e^{-h_k}}$$

$$[0150] \quad h_k = \sum_{m=1}^M \varphi_m v_{mk}$$

[0151] 然后,计算每层的误差和正确权重。输出的误差计算为:

$$[0152] \quad \delta_{ok} = (t_k - \varphi_k) \varphi_k (1 - \varphi_k)$$

[0153] 隐藏层中的误差被计算为:

$$[0154] \quad \delta_{hm} = \varphi_m (1 - \varphi_m) \sum_{k=1}^K v_{mk} \delta_{ok}$$

[0155] 输出层的权重更新为:

$$[0156] \quad v_{mk} \leftarrow v_{mk} + \alpha \delta_{ok} \varphi_m$$

[0157] 使用以下学习率 α 来更新隐藏层的权重:

$$[0158] \quad v_{nm} \leftarrow v_{nm} + \alpha \delta_{hm} a_n$$

[0159] 在一个实现方式中,卷积神经网络使用梯度下降优化来计算跨所有层的误差。在此类优化中,对于输入特征向量x和预测输出 $\hat{y}$,对于当目标为y时预测 $\hat{y}$ 的成本,损失函数被定义为1,即 $l(\hat{y}, y)$ 。使用函数f从输入特征向量x变换预测输出 $\hat{y}$ 。函数f由卷积神

经网络的权重参数化,即 $\hat{y} = f_w(x)$ 。损失函数描述为

$l(\hat{y}, y) = l(f_w(x), y)$, 或 $Q(z, w) = l(f_w(x), y)$, 其中 z 是输入和输出数据对 (x, y) 。通过根据以下条件更新权重来执行梯度下降优化:

$$[0160] \quad v_{t+1} = \mu v_t - \alpha \frac{1}{n} \sum_{i=1}^N \nabla_{w_t} Q(z_t, w_t)$$

$$[0161] \quad w_{t+1} = w_t + v_{t+1}$$

[0162] 在上面的等式中, α 是学习率。此外, 损失被计算为一组 n 数据对上的平均值。在线性收敛时, 当学习率 α 足够小时, 计算终止。在其他实现方式中, 仅使用被馈送到内斯特罗夫加速梯度和自适应梯度的选定数据对来计算梯度, 以注入计算效率。

[0163] 在一个实现方式中, 卷积神经网络使用随机梯度下降 (SGD) 来计算成本函数。SGD 通过仅从一个, 随机, 数据对, z_t , 计算出损失函数中的权重来近似梯度, 描述如下:

$$[0164] \quad v_{t+1} = \mu v_t - \alpha \nabla_{w_t} Q(z_t, w_t)$$

$$[0165] \quad w_{t+1} = w_t + v_{t+1}$$

[0166] 在上面的等式中: α 是学习率; μ 是动量; 并且 t 是更新之前的当前权重状态。当学习率 α 足够快和足够慢两者时, SGD 的收敛速度接近 $O(1/t)$ 。在其他实现中, 卷积神经网络使用不同的损失函数, 诸如欧几里得损失和 softmax 损失。在另一个实现方式中, 由卷积神经网络使用 Adam 随机优化器。

[0167] 卷积层

[0168] 卷积神经网络的卷积层用作特征提取器。卷积层充当能够学习并将输入数据分解为分层特征的自适应特征提取器。在一个实现方式中, 卷积层将两个图像作为输入, 并且产生第三图像作为输出。在此类实现方式中, 卷积在二维 (2D) 中的两个图像上操作, 其中一个图像是输入图像并且另一个图像 (称为“内核”) 被应用为输入图像上的滤波器, 从而产生输出图像。因此, 对于长度为 n 的输入向量 f 和长度为 m 的内核 g , f 和 g 的卷积 $f * g$

[0169] 被定义为:

$$[0170] \quad (f * g)(i) = \sum_{j=1}^m g(j) \cdot f(i - j + m/2)$$

[0171] 卷积操作包括在输入图像上滑动内核。对于内核的每个位置, 将内核和输入图像的重叠值相乘, 并添加结果。乘积之和是输入图像中内核居中的点处输出图像的值。许多内核产生的不同输出称为特征映射图。

[0172] 一旦对卷积层进行了训练就可以将它们应用于对新的推理数据执行识别任务。由于卷积层从训练数据中学习, 因此它们避免了显式特征提取并且隐式地从训练数据中学习。卷积层使用卷积滤波器内核权重, 所述卷积滤波器内核权重被确定并更新为训练过程的一部分。卷积层提取输入的不同特征, 所述不同特征在较高层进行组合。卷积神经网络使用不同数量的卷积层, 每个卷积层具有不同的卷积参数, 诸如内核大小、步幅、填充、特征映射图和权重的数量。

[0173] 非线性层

[0174] 图5示出了根据所公开的技术的一个实现方式的非线性层的一个实现方式。非线性层使用不同的非线性触发函数来对每个隐藏层上可能的特征发出不同的识别信号。非线性

性层使用各种具体函数来实现非线性触发,其包括经修正线性单元(ReLU)、双曲正切、双曲正切的绝对值、S形和连续触发(非线性)函数。在一个实现方式中,ReLU激活实现函数 $y = \max(x, 0)$ 并且保持层的输入和输出大小相同。使用ReLU的优点是卷积神经网络的训练速度快了很多倍。ReLU是非连续,非饱和的激活函数,如果输入值大于零,则相对于输入是线性的,否则为零。在数学上,ReLU激活函数被描述为:

$$[0175] \quad \varphi(h) = \max(h, 0)$$

$$[0176] \quad \varphi(h) = \begin{cases} h & \text{如果 } h > 0 \\ 0 & \text{如果 } h \leq 0 \end{cases}$$

[0177] 在其他实现中,卷积神经网络使用功率单元激活函数,该函数是连续的,非饱和函数,描述如下:

$$[0178] \quad \varphi(h) = (a + bh)^c$$

[0179] 在上面的等式中,a,b和c相应地是控制移位、缩放和功率的参数。如果c为奇数,则功率激活函数能够产生x和y-反对称激活,如果c为偶数,则能够产生y-对称激活。在一些实现方式中,所述单元产生非修正线性激活。

[0180] 在其他实现方式中,卷积神经网络使用S形单位激活函数,该函数是一个连续的饱和函数,由以下逻辑函数描述:

$$[0181] \quad \varphi(h) = \frac{1}{1 + e^{-\beta h}}$$

[0182] 在上面的等式中, $\beta = 1$ 。S形单元激活函数不产生负激活,并且仅相对于y-轴反对称。

[0183] 膨胀卷积

[0184] 图6示出了膨胀卷积。膨胀卷积有时称为空洞卷积,其在字面上意味着具有孔。该法语名称源于计算快速二进小波变换的算法a trous。在这些类型的卷积层中,对应于滤波器的感受野的输入不是相邻点。这在图6中示出。输入之间的距离取决于膨胀因子。

[0185] 子采样层

[0186] 图7是根据所公开的技术的一个实现方式的子采样层的一个实现方式。子采样层降低卷积层提取的特征的分辨率,以使提取的特征或特征映射图对噪声和失真具有稳健性。在一个实现方式中,子采样层采用两种类型的池化操作:平均池化和最大池化。池化操作将输入划分成不重叠的二维空间。对于平均池化,计算该区域中的四个值的平均值。对于最大池化,选择四个值中的最大值。

[0187] 在一个实现方式中,子采样层包括通过将其输出映射到最大池化中的输入中的仅一个并且通过将其输出映射到平均池化中的输入的平均值来对先前层中的一组神经元进行池化操作。在最大池化中,池化神经元的输出是驻留在输入内的最大值,如下所描述的:

$$[0188] \quad \varphi_o = \max(\varphi_1, \varphi_2, \dots, \varphi_N)$$

[0189] 在上面的等式中,N是神经元集中的元素的总数。

[0190] 在平均池化中,池化神经元的输出是驻留在输入神经元集内的输入值的平均值,如下所描述的:

$$[0191] \quad \varphi_o = \frac{1}{N} \sum_{n=1}^N \varphi_n$$

[0192] 在上面的等式中,N是输入神经元集中的元素的总数。

[0193] 在图7中,输入的大小为 4×4 。对于 2×2 子采样,将 4×4 图像划分成大小为 2×2 的四个非重叠矩阵。对于平均池化,四个值的平均值是全整数输出。对于最大池化, 2×2 矩阵中的四个值的最大值是全整数输出。

[0194] 卷积示例

[0195] 图8描绘了卷积层的两层卷积的一个实现方式。在图8中,对大小为2048个维度的输入进行卷积。在卷积1处,通过卷积层对输入进行卷积,所述卷积层包括大小为 3×3 的十六个内核的两个通道。然后,借助于在ReLU1处的ReLU激活函数对所得到的16个特征映射图进行修正,然后借助于使用具有大小为 3×3 的内核的16个通道池化层的平均池化,在池化1中池化得到的16个特征映射图。在卷积2处,池化1的输出随后被另一个卷积层卷积,该另一个卷积层包括大小为 3×3 的三十个内核的十六个通道。接下来是内核大小为 2×2 的池化2中的又一个ReLU2和平均池化。卷积层使用不同的步幅和填充,例如零、一、二和三。根据一个实现方式,所得到的特征向量是五百一十二(512)个维度。

[0196] 在其他实现方式中,卷积神经网络使用不同数量的卷积层、子采样层、非线性层和全连接层。在一个实现方式中,卷积神经网络是每层具有较少的层和更多神经元的浅网络,例如每层具有一百(100)至两百(200)个神经元的一个、两个或三个全连接层。在另一个实现方式中,卷积神经网络是每层具有更多层和较少神经元的深度网络,例如每层五(5)、六(6)或八(8)个全连接层,每层具有三十(30)至五十(50)个神经元。

[0197] 前向传递

[0198] 第1个卷积层中第x行,第y列的神经元和特征映射图中第k个特征映射图的输出由以下等式确定:

$$[0199] \quad O_{x,y}^{(l,k)} = \tanh\left(\sum_{t=0}^{f-1} \sum_{r=0}^{k_h} \sum_{c=0}^{k_w} W_{(r,c)}^{(k,t)} O_{(x+r,x+c)}^{(l-1,t)} + Bias^{(l,k)}\right)$$

[0200] 第1个子采样层中的第x行,第y列和第k个特征映射图的神经元的输出由以下等式确定:

$$[0201] \quad O_{x,y}^{(l,k)} = \tanh\left(W^{(k)} \sum_{r=0}^{S_h} \sum_{c=0}^{S_w} O_{(x \times S_h + r, y \times S_w + c)}^{(l-1,k)} + Bias^{(l,k)}\right)$$

[0202] 第1个输出层的第i个神经元的输出由以下等式确定:

$$[0203] \quad O_{(l,i)} = \tanh\left(\sum_{j=0}^H O_{(l-1,j)} W_{(i,j)}^l + Bias^{(l,i)}\right)$$

[0204] 后向传播

[0205] 输出层中的第k个神经元的输出偏差由以下等式确定:

$$[0206] \quad d(O_k^o) = y_k - t_k$$

[0207] 输出层中的第k个神经元的输入偏差由以下等式确定:

$$[0208] \quad d(I_k^o) = (y_k - t_k) \varphi'(v_k) = \varphi'(v_k) d(O_k^o)$$

[0209] 输出层中的第k个神经元的权重和偏差变化由以下等式确定：

$$[0210] \quad \Delta W_{k,x}^o = d(I_k^o) y_{k,x}$$

$$[0211] \quad \Delta Bias_k^o = d(I_k^o)$$

[0212] 隐藏层中的第k个神经元的输出偏差由以下等式确定：

$$[0213] \quad d(O_k^H) = \sum_{i=0}^{i<84} d(I_i^o) W_{i,k}$$

[0214] 隐藏层中的第k个神经元的输入偏差由以下等式确定：

$$[0215] \quad d(I_k^H) = \varphi'(v_k) d(O_k^H)$$

[0216] 从隐藏层中的k个神经元接收输入的初始层的第m个特征映射图中的行x、列y中的权重和偏差变化由以下等式确定：

$$[0217] \quad \Delta W_{m,x,y}^{H,k} = d(I_k^H) y_{x,y}^m$$

$$[0218] \quad \Delta Bias_k^H = d(I_k^H)$$

[0219] 子采样层S的第m个特征映射图中的行x、列y的输出偏差由以下等式确定：

$$[0220] \quad d(O_{x,y}^{S,m}) = \sum_k^{170} d(I_{m,x,y}^H) W_{m,x,y}^{H,k}$$

[0221] 子采样层S的第m个特征映射图中的行x、列y的输入偏差由以下等式确定：

$$[0222] \quad d(I_{x,y}^{S,m}) = \varphi'(v_k) d(O_{x,y}^{S,m})$$

[0223] 子采样层S和卷积层C的第m个特征映射图中的行x、列y中的权重和偏差变化由以下等式确定：

$$[0224] \quad \Delta W^{S,m} = \sum_{x=0}^{fh} \sum_{y=0}^{fw} d(I_{[x/2],[y/2]}^{S,m}) O_{x,y}^{C,m}$$

$$[0225] \quad \Delta Bias^{S,m} = \sum_{x=0}^{fh} \sum_{y=0}^{fw} d(O_{x,y}^{S,m})$$

[0226] 卷积层C的第k个特征映射图中的行x、列y的输出偏差由以下等式确定：

$$[0227] \quad d(O_{x,y}^{C,k}) = d(I_{[x/2],[y/2]}^{S,k}) W^k$$

[0228] 卷积层C的第k个特征映射图中的行x、列y的输入偏差由以下等式确定：

$$[0229] \quad d(I_{x,y}^{C,k}) = \varphi'(v_k) d(O_{x,y}^{C,k})$$

[0230] 第1个卷积层C的第k个特征映射图的第m个卷积内核中的行r、列c中的权重和偏差变化：

$$[0231] \quad \Delta W_{r,c}^{k,m} = \sum_{x=0}^{fh} \sum_{y=0}^{fw} d(I_{x,y}^{C,k}) O_{x+r,y+c}^{l-1,m}$$

$$[0232] \quad \Delta Bias^{C,k} = \sum_{x=0}^{fh} \sum_{y=0}^{fw} d(I_{x,y}^{C,k})$$

[0233] 残差连接

[0234] 图9描绘了经由特征映射图添加在下游重新注入初始信息的残差连接。残差连接包括通过将过去的输出张量添加到稍后的输出张量来将先前的表示重新注入到下游数据流中,这有助于防止沿着数据处理流的信息丢失。残差连接解决了困扰任何大规模深度学习模型的两个常见问题:消失梯度和代表性瓶颈。通常,向具有10个以上层的任何模型添加残差连接有可能是有益的。如以上所讨论的,残差连接包括使较早层的输出可用作较后层的输入,从而有效地在顺序网络中创建快捷方式。不是被级联到稍后的激活,而是将较早的输出与稍后的激活相加,这假设两个激活都是相同的大小。如果它们具有不同的大小,则可以使用将较早激活重新成形为目标形状的线性变换。

[0235] 残差学习和跳跃连接

[0236] 图10描绘了残差块和跳跃连接的一个实现方式。残差学习的主要思想是残差映射比原始映射更容易学习。残差网络堆栈多个残差单元,以减轻训练准确度的下降。残差块利用特殊的加性跳跃连接来对抗深度神经网络中的消失梯度。在残差块的开始处,数据流被分成两个流:第一个流携带块的未变化输入,而第二个流应用权重和非线性。在块的末尾,使用逐元素的和来合并这两个流。此类构造的主要优点是允许梯度更容易地流过网络。

[0237] 受益于残差网络,可以容易地训练深度卷积神经网络(CNN),并且已经实现了用于图像分类和对象检测的改善的准确度。卷积前馈网络将第1层的输出作为输入连接到第(1+1)层,这引起以下层过渡: $x_1 = H_1(x_{1-1})$ 。残差块添加了一个跳跃连接,该跳跃连接使用恒等函数: $x_1 = H_1(x_{1-1}) + x_{1-1}$ 绕过非线性变换。残差块的优点是梯度可以直接通过恒等函数从后面的层流向前面的层。然而,通过求和来组合恒等函数和 H_1 的输出,这可能妨碍网络中的信息流。

[0238] WaveNet

[0239] Wavenet是用于生成原始音频波形的深度神经网络。WaveNet与其他卷积网络不同,因为它能够以低成本获得相对较大的“视野”。此外,它能够在局部地和全局地添加信号的调节,这允许Wavenet用作具有多个语音的文本到语音(TTS)引擎,其中TTS给出局部调节和特定语音到全局调节。

[0240] WaveNet的主要构建块是因果膨胀卷积。作为因果膨胀卷积上的膨胀,WaveNet还允许这些卷积的堆栈,如图11所示。为了在该图中获得具有膨胀卷积的相同感受野,需要另一个膨胀层。堆栈是膨胀卷积的重复,将膨胀卷积层的输出连接到单个输出。这使得WaveNet能够以相对较低的计算成本得到一个输出节点的大“视野”。为了比较,为了获得512个输入的视野,全卷积网络(FCN)将需要511层。在膨胀卷积网络的情况下,我们将需要八层。堆栈的膨胀卷积只需要具有两个堆栈的七层或具有四个堆栈的六层。为了了解覆盖同一视野所需的计算能力差异,下表示出了假设每层有一个滤波器,滤波器宽度为两个时网络所需的权重的数量。此外,假设网络正在使用8位的二进制编码。

	网络类型	堆 栈 数	每个通道的权重 数	权重的总数
[0241]	FCN	1	$2.6 \cdot 10^5$	$2.6 \cdot 10^6$
	WN	1	1022	8176
	WN	2	1022	8176
	WN	4	508	4064

[0242] WaveNet在建立残差连接之前添加跳跃连接,所述跳跃连接将绕过以下所有残差块。在传递一系列激活函数和卷积之前,对这些跳跃连接中的每一个进行求和。直观地说,这是在每层提取的信息的和。

[0243] 批归一化

[0244] 批归一化是用于通过使数据归一化成为网络架构的组成部分来加速深度网络训练的方法。即使在训练期间随着时间的平均和方差变化,批归一化也可以自适应地归一化数据。它通过在内部维持在训练期间看到的数据的分批均值和方差的指数移动平均值来工作。批归一化的主要效果是它有助于梯度传播(非常类似于残差连接),并且因此允许深度网络。一些非常深的网络仅在包含多个批归一化层时才能被训练。

[0245] 批归一化可以看作是可以插入模型架构的又一层,就像全连接或卷积层一样。通常在卷积或密集连接层之后使用批归一化层。它也可以在卷积或密集连接层之前使用。这两种实现方式都可由所公开的技术来使用,并且在图15中示出。批归一化层获取轴自变量,该自变量指定应归一化的特征轴。此自变量默认为-1,即输入张量中的最后一个轴。当使用数据格式设置为“通道_最后一个”的Dense层、Conv1D层、RNN层和Conv2D层时,这是正确的值。但是在数据格式设置为“通道_优先”的Conv2D层的利基用例中,特征轴是轴1;批归一化中的轴自变量可以设置为1。

[0246] 批归一化提供了用于对输入进行前馈的定义,并且经由后向传递相对于参数及其自己的输入来计算梯度。在实践中,在卷积或全连接层之后,但是在输出被馈送到激活函数之前,插入批归一化层。对于卷积层,相同特征映射图的不同元素(即,不同位置处的激活)以相同的方式被归一化,以便服从卷积属性。因此,小批中的所有激活在所有位置而不是每次激活被归一化。

[0247] 内部协变量移位是众所周知深度架构训练缓慢的主要原因。这源于以下事实:深度网络不仅需要在每层学习新的表示,而且还需要考虑它们分布的变化。

[0248] 协变量移位通常是深度学习领域中的已知问题,并且经常发生在现实世界的问题中。常见的协变量移位问题是训练集和测试集的分布差异,这可能导致次优的泛化性能。这个问题通常通过归一化或白化预处理步骤来处理。然而,尤其是白化操作在计算上是昂贵的,并且因此在在线设置中是不切实际的,尤其是如果协变量移位发生在贯穿不同的层中。

[0249] 内部协变量移位是指网络激活的分布因训练期间网络参数的变化而跨层变化的现象。理想地,每层应被变换成空间,在该空间中它们具有相同的分布,但是函数关系保持不变。为了避免在每层和每一步对协方差矩阵进行昂贵的计算以去相关和白化数据,我们将每层中的每一个输入特征跨每一个小批的分布归一化为具有零均值和一的标准偏差。

[0250] 前向传递

[0251] 在前向传递期间,计算小批平均值和方差。利用这些小批统计,经由减去平均值并除以标准偏差来归一化数据。最后,用学习到的缩放和移位参数对数据进行缩放和移位。在图12中描绘了批归一化前向传递 f_{BN} 。

[0252] 在图12中,相应地 μ_p 是批均值并且 σ_p^2 是批方差。学习的缩放和移位参数相应地由 γ 和 β 表示。为了清楚起见,在本文中针对每次激活描述了批归一化过程,并省略了相对应的索引。

[0253] 由于归一化是可微分的变换,所以误差被传播到这些学习的参数中,并且因此能够通过学习恒等变换来恢复网络的代表性功率。相反地,通过学习与相对应的批统计相同的缩放和移位参数,如果是要执行的最佳操作,则批归一化变换将对网络没有影响。在测试时间,批均值和方差由各自的总体统计代替,因为输入不依赖于来自小批的其他采样。另一个方法是在训练期间保持批统计的运行平均值,并且在测试时间使用这些平均值来计算网络输出。在测试时间,可以如图13所示表示批归一化变换。在图13中, μ_D 和 σ_D^2 相应地表示总体均值和方差,而不是批统计。

[0254] 后向传递

[0255] 由于归一化是可微分操作,所以可以如图14中所描绘的那样计算后向传递。

[0256] 1D卷积

[0257] 1D卷积从序列中提取局部1D子块或子序列,如图16所示,1D卷积从输入序列中的时间子块获得每个输出时间步长。1D卷积层识别序列中的局部图案。因为在每个子块上执行相同的输入变换,所以在输入序列中的某个位置学习的图案可以稍后在不同的位置被识别,使得1D卷积层的平移对于时间平移不变。例如,使用大小为5的卷积窗口处理碱基序列的1D卷积层应能够学习长度为5或更短的碱基或碱基序列,并且应能够识别输入序列中任何上下文中的碱基基序。因此,基本级别的1D卷积能够学习基本形态。

[0258] 全局平均池化

[0259] 图17示出了全局平均池化(GAP)是如何工作的。通过获取最后一层元素的空间平均值进行评分,可以使用全局平均池化来替换用于分类的全连接(FC)层。这减少了训练负荷并绕过了过度拟合的问题。全局平均池化在模型之前应用结构,并且其等价于具有预定义权重的线性变换。全局平均池化减少了参数的数量并且消除了全连接层。全连接层通常是最大参数和连接密集的层,并且全局平均池化提供了较低成本的方法来实现类似的结果。全局平均池化的主要思想是从每个最后一层特征映射图中生成平均值,作为评分的置信度因子,直接馈送到softmax层。

[0260] 全局平均池化具有三个益处:(1)在全局平均池化层中不存在额外参数,因此在全局平均池化层处避免过度拟合;(2)由于全局平均池化的输出是整个特征映射图的平均值,全局平均池化对于空间平移将是更稳健的;以及(3)由于全连接层中的大量参数,在整个网络的所有参数中通常需要超过50%,通过全局平均池化层来替换它们可以显著地减小模型的大小,并且这使得全局平均池化在模型压缩中非常有用。

[0261] 全局平均池化是有意义的,因为在最后一层中较强的特征预计将具有较高的平均值。在一些实现方式中,全局平均池化可以用作用于分类得分的代理。全局平均池化下的特征映射图可以解译为置信度映射,并强制特征映射图和类别之间的对应。如果最后一层的

特征具有足够的抽象度用于直接分类,则全局平均池化可以是特别有效的;然而,如果多级特征应被组合成如零件模型的组,则单独的全局平均池化不够,这通过在全局平均池化之后添加简单的全连接层或其他分类器来最佳地执行。

[0262] 术语

[0263] 本申请中引用的所有文献和类似材料,包括但不限于专利、专利申请、文章、书籍、论文和网页,无论这些文献和类似材料的格式如何,均通过引用其全部内容明确地并入。如果一个或多个所并入的文献和类似材料与本申请不同或矛盾,包括但不限于所定义的术语、术语用法、所描述的技术等,则以本申请为准。

[0264] 如本文所用,以下术语具有所指示的含义。

[0265] 碱基是指核苷酸碱基或核苷酸、A(腺嘌呤)、C(胞嘧啶)、T(胸腺嘧啶)或G(鸟嘌呤)。

[0266] 本申请可互换地使用术语“蛋白质”和“转译序列”。

[0267] 本申请可互换地使用术语“密码子”和“碱基三联体”。

[0268] 本申请可互换地使用术语“氨基酸”和“转译单元”。

[0269] 本申请可互换地使用短语“变体致病性分类器”、“用于变体分类的基于卷积神经网络的分类器”和“用于变体分类的基于深卷积神经网络的分类器”。

[0270] 术语“染色体”是指活细胞的带有遗传的基因载体,其衍生自包含DNA和蛋白质成分(尤其是组蛋白)的染色质链。本文采用了常规的国际公认的人类个体基因组染色体编码系统。

[0271] 术语“位点”是指参考基因组上的独特位置(例如,染色体ID、染色体位置和取向)。在某些实现方式中,位点可以是残基、序列标签或序列上的片段的位置。术语“基因座”可用于指参考染色体上的核酸序列或多态性的特定位置。

[0272] 本文中的术语“样品”是指通常衍生自包含核酸或包含至少一种待测序和/或分阶段进行的核酸序列的核酸混合物的生物流体、细胞、组织、器官或生物体的样品。此类样品包括但不限于痰液/口液、羊水、血液、血液分数、细针穿刺活检样品(例如,外科活检、细针穿刺活检等)、尿液、腹膜液、胸膜液、组织外植体、器官培养物 and 任何其他组织或细胞制剂,或其馏分或衍生物或从中分离的。尽管样品通常取自人类受试者(例如患者),但样品可以取自具有染色体的任何生物体,包括但不限于狗、猫、马、山羊、绵羊、牛、猪等。样品可以从生物来源获得时那样直接使用或在修饰样品特性的预处理之后使用。例如,这样的预处理可包括从血液、稀释粘性流体等制备血浆。预处理方法可能还包括但不限于过滤、沉淀、稀释、蒸馏、混合、离心、冷冻、冻干、浓缩、扩增、核酸片段化、干扰组分失活、试剂添加、裂解等。

[0273] 术语“序列”包括或表示彼此偶联的核苷酸链。核苷酸可基于DNA或RNA。应当理解,一个序列可包括多个子序列。例如,(例如,PCR扩增子的)单个序列可具有350个核苷酸。样品解读可在这350个核苷酸内包括多个子序列。例如,样品解读可包括具有例如20-50个核苷酸的第一和第二侧翼子序列。第一和第二侧翼子序列可位于具有相应子序列(例如40-100个核苷酸)的重复片段的任一侧。每个侧翼子序列可包括(或包括部分的)引物子序列(例如,10-30个核苷酸)。为了易于阅读,术语“子序列”将被称为“序列”,但是应理解,两个序列在公共链上不一定彼此分开。为了区分本文所述的各种序列,可以给序列赋予不同的

标记(例如,靶序列、引物序列、侧翼序列、参考序列等)。诸如“等位基因”的其他术语可被赋予不同的标签以区分相似的对象。

[0274] 术语“配对末端测序”是指对靶片段的两端进行测序的测序方法。配对末端测序可以促进基因组重排和重复片段的检测,以及基因融合和新型转录本的检测。在PCT公开W007010252、PCT申请序列号PCTGB2007/003798和美国专利申请公开US 2009/0088327中描述了用于配对末端测序的方法,其各自通过引入并入本文。在一个实例中,一系列操作可以如下进行;(a)产生核酸簇;(b)使核酸线性化;(c)杂交第一测序引物,并如上所述进行延伸、扫描和去封闭的重复循环;(d)通过合成互补拷贝“反转”流细胞表面的靶核酸;(e)使重新合成的链线性化;(f)杂交第二测序引物,并如上所述进行延伸、扫描和去封闭的重复循环。可以通过传递如上所述的用于桥式扩增的单个循环的试剂来进行转化操作。

[0275] 术语“参考基因组”或“参考序列”是指可用于参考来自受试者的鉴定序列的任何生物体的任何特定的已知基因组序列,无论是部分的还是完整的。例如,在国家生物技术信息中心(ncbi.nlm.nih.gov)上可以找到用于人类受试者以及许多其他生物体的参考基因组。“基因组”是指以核酸序列表达的生物体或病毒的完整遗传信息。基因组既包括基因又包括DNA的非编码序列。参考序列可能比与之比对的解读大。例如,它可以大至少约100倍,或大至少约1000倍,或大至少约10,000倍,或大至少约105倍,或大至少约106倍,或大至少约107倍。在一个实例中,参考基因组序列是全长人类基因组的序列。在另一个实例中,参考基因组序列限于特定的人类染色体,例如13号染色体。在一些实施方案中,参考染色体是来自人类基因组版本hg19的染色体序列。这样的序列可被称为染色体参考序列,尽管术语参考基因组旨在覆盖这样的序列。参考序列的其他实例包括其他物种的基因组,以及任何物种的染色体、子染色体区域(例如链)等。在各种实施方案中,参考基因组是共有序列或衍生自多个个体的其他组合。然而,在某些应用中,参考序列可以取自特定个体。

[0276] 术语“解读”是指描述核苷酸样品或参照物的片段的序列数据的集合。术语“解读”可以指样品解读和/或参考解读。通常,尽管不是必须的,解读代表样品或参考中的连续碱基对的短序列。解读可由样品或参考片段的碱基对序列(在ATCG中)象征性地表示。可以将其存储在存储设备中,并进行适当的处理,以确定解读是否与参考序列匹配或满足其他标准。可以直接从测序设备获得解读,或从与样品有关的存储序列信息间接获得解读。在某些情况下,解读是具有足够长度(例如,至少约25bp)的DNA序列,可用于识别更大的序列或区域,例如可比对并特异性分配给染色体或基因组区域或基因的更大的序列或区域。

[0277] 下一代测序方法包括,例如,通过合成技术(Illumina)测序、焦磷酸测序(454)、离子半导体技术(离子激流测序)、单分子实时测序(Pacific Biosciences)和通过连接测序(SOLiD测序)。根据测序方法的不同,每个解读的长度可能在大约30bp至大于10,000bp。例如,使用SOLiD测序仪的Illumina测序方法产生约50bp的核酸解读。又例如,离子激流测序产生高达400bp的核酸解读,而454焦磷酸测序产生约700bp的核酸解读。又例如,单分子实时测序方法可产生10,000bp至15,000bp的解读。因此,在某些实施方案中,核酸序列解读的长度为30-100bp、50-200bp或50-400bp。

[0278] 术语“样品解读”、“样品序列”或“样品片段”是指来自样品的感兴趣的基因组序列的序列数据。例如,样品解读包括来自具有正向和反向引物序列的PCR扩增子的序列数据。序列数据可从任何选择的序列方法中获得。样品解读可以是,例如,来自边合成边测序

(SBS) 反应、边连接边测序或需要确定重复序列的长度和/或同一性的任何其他合适的测序方法。样品解读可以是衍生自多个样品解读的共有(例如,平均或加权)序列。在某些实施方案中,提供参考序列包括基于PCR扩增子的引物序列鉴定感兴趣的基因座。

[0279] 术语“原始片段”是指感兴趣的基因组序列的一部分的序列数据,其至少部分地重叠在样品解读或样品片段内的指定的感兴趣位置或次要位置。原始片段的非限制性实例包括双工缝合片段,单工缝合片段、双工非缝合片段和单工非缝合片段。术语“原始”用于指示原始片段包括与样品解读中的序列数据有某种关系的序列数据,而不管原始片段是否显示出对应于并鉴定或确认样品解读中的潜在变异的支持变体。术语“原始片段”并不表示该片段必然包含支持变体,该变体可验证样品解读中的变体识别。例如,当变体识别应用确定样品解读显示出第一变体时,变体识别应用可以确定一个或多个原始片段缺少相应类型的“支持”变体,否则可预期在样品解读中的给定变体的条件下可能发生。

[0280] 术语“映射(mapping)”、“比对(aligned)”、“比对(alignment)”或“比对(aligning)”是指将阅读或标签与参考序列进行比较,从而确定参考序列是否包含阅读序列的过程。如果参考序列包含解读,则该解读可以被映射到参考序列,或者在某些实现方式中,可以被映射到参考序列中的特定位置。在某些情况下,比对只是告诉解读是否为特定参考序列的成员(即,解读在参考序列中存在还是不存在)。例如,解读与人染色体13的参考序列的比对将表明该解读是否存在于13号染色体的参考序列中。提供此信息的工具可以称为设置成员资格测试仪。在某些情况下,比对还指示解读或标签所映射的参考序列中的位置。例如,如果参考序列是整个人类基因组序列,则比对可以指示在13号染色体上存在解读,并且可以进一步指示该解读在13号染色体的特定链和/或位点上。

[0281] 术语“插入/缺失”是指生物体DNA中碱基的插入和/或缺失。微型插入缺失表示导致1至50个核苷酸的净变化的插入缺失。在基因组的编码区域中,除非插入缺失的长度是3的倍数,否则它将产生移码突变。插入缺失可以与点突变形成对比。插入缺失插入物从序列中插入和删除核苷酸,而点突变是一种替换形式,可以替换一个核苷酸而不改变DNA的总数。插入缺失也可以与串联碱基突变(TBM)形成对比,串联碱基突变可以定义为在相邻核苷酸处的取代(主要是在两个相邻核苷酸处的取代,但已经观察到在三个相邻核苷酸处的取代)。

[0282] 术语“变体”是指与核酸参照不同的核酸序列。典型的核酸序列变体包括但不限于单核苷酸多态性(SNP)、短缺失和插入多态性(Indel)、拷贝数变异(CNV)、微卫星标记或短串联重复和结构变异。体细胞变异识别是鉴定DNA样品中低频存在的变异的工作。在癌症治疗的背景下,体细胞变异识别是令人感兴趣的。癌症是由DNA突变的积累引起的。来自肿瘤的DNA样品通常是异质的,包括一些正常细胞,一些处于癌症进展早期的细胞(突变较少)和一些晚期细胞(突变较多)。由于这种异质性,当对肿瘤进行测序(例如,从FFPE样品中)时,体细胞突变通常会以较低的频率出现。例如,SNV可能仅在覆盖给定碱基的10%解读中可见。被变体分类器分类为体细胞或种系的变体在本文中也称为“被测变体”。

[0283] 术语“噪声”是指由于测序过程和/或变异识别应用程序中的一个或多个错误而导致的错误变体识别。

[0284] 术语“变体频率”表示群体中特定基因座的等位基因(基因变异)的相对频率,以分数或百分比表示。例如,分数或百分比可以是携带该等位基因的群体中的所有的染色体的

分数。举例来说,样品变体频率表示在“群体”上沿着感兴趣的基因组序列在特定基因座/位置的等位基因/变体的相对频率,该“群体”对应于针对来自个体的感兴趣的基因组序列获得的解读和/或样品的数量。作为另一实例,基线变体频率表示沿着一个或多个基线基因组序列在特定基因座/位置的等位基因/变体的相对频率,其中“群体”对应于来自正常个体群体的一个或多个基线基因组序列获得的解读和/或样品的数量。

[0285] 术语“变异等位基因频率 (VAF)”是指观察到的与变体匹配的测序解读的百分比除以靶位置的总覆盖率。VAF是携带变体的测序解读的比例的测度。

[0286] 术语“位置”、“指定位置”和“基因座”是指一个或多个核苷酸在核苷酸序列内的位置或坐标。术语“位置”、“指定位置”和“基因座”也指核苷酸序列中的一个或多个碱基对的位置或坐标。

[0287] 术语“单倍型”是指在染色体上相邻位点的等位基因的组合,其是一起遗传的。单倍型可以是一个基因座、多个基因座或整个染色体,取决于给定基因座组之间(如果有的话)发生的重组事件的数量。

[0288] 术语“阈值”在本文中是指用作临界值以表征样品、核酸或其部分(例如,解读)的数值或非数值。阈值可基于经验分析而变化。可以将阈值与测量值或计算值进行比较,以确定是否应以特定方式对产生该值的来源进行分类。阈值可以凭经验或分析来识别。阈值的选择取决于用户希望进行分类的置信度。可以出于特定目的(例如,以平衡灵敏度和选择性)来选择阈值。如本文所用,术语“阈值”表示可以改变分析过程的点和/或可以触发动作的点。阈值不需要是预定数量。相反,阈值可以是例如,基于多个因素的函数。该阈值可以适应环境。此外,阈值可以指示上限、下限或极限之间的范围。

[0289] 在一些实现方式中,可以将基于测序数据的度量或得分与阈值进行比较。如本文所用,术语“度量”或“得分”可以包括从测序数据确定的值或结果,或者可以包括基于从测序数据确定的值或结果的函数。像阈值一样,度量或得分可以适应环境。例如,度量或得分可以是归一化值。作为得分或度量的实例,一个或多个实现方式可以在分析数据时使用计数得分。计数得分可以基于样品解读的数量。样品解读可能已经历一个或多个过滤阶段,以使样品解读具有至少一种共同的特征或质量。例如,用于确定计数得分的每个样品解读可能已经与参考序列比对,或者可能被指定为潜在等位基因。可以对具有共同特征的样品解读次数进行计数以确定解读计数。计数得分可以基于解读计数。在一些实施方案中,计数得分可以是等于解读计数的值。在其他实施方案中,计数得分可以基于解读计数和其他信息。例如,计数得分可以基于遗传基因座的特定等位基因的解读计数和遗传基因座的解读总数。在一些实现方式中,计数得分可以基于解读计数和遗传基因座先前获得的数据。在一些实现方式中,计数得分可以是预定值之间的归一化得分。计数得分还可以是来自样品其他基因座的解读计数的函数,或者是与感兴趣样品同时运行的其他样品的解读计数的函数。例如,计数得分可以是特定等位基因的解读计数和样品中其他基因座的解读计数和/或来自其他样品的解读计数的函数。作为一个实例,来自其他基因座的解读计数和/或来自其他样品的解读计数可用于归一化特定等位基因的计数得分。

[0290] 术语“覆盖率”或“片段覆盖率”是指针对序列的相同片段的多个样品解读的计数或其他度量。解读计数可以代表覆盖相应片段的解读数目的计数。可备选地,可以通过将解读计数乘以基于历史知识、样品知识、基因座知识等的指定因子来确定覆盖率。

[0291] 术语“解读深度”(通常为数字,后跟“×”)是指在靶位置处具有重叠比对的测序解读的数目。这通常表示为在一组时间间隔(例如外显子、基因或面板)上超过临界值的平均值或百分比。例如,一份临床报告可能会说面板平均覆盖率是1,105倍,其中98%的靶碱基覆盖率超过100倍。

[0292] 术语“碱基识别质量得分”或“Q得分”是指范围为0-20的PHRED缩放的概率,范围与单个测序碱基正确的概率成反比。例如,Q为20的T碱基识别可能被认为正确,具有0.01的置信度P值。Q<20的任何碱基识别均应被认为是低质量的,而鉴定出支持该变体的相当一部分测序解读的低质量的任何变体都应被视为潜在的假阳性。

[0293] 术语“变体解读”或“变体解读数目”是指支持变体存在的测序解读的数目。

[0294] 测序过程

[0295] 本文阐述的实施方案可适用于分析核酸序列以鉴定序列变异。可以使用实现方式来分析遗传位置/基因座的潜在变体/等位基因,并确定遗传基因座的基因型,或者换句话说,为该基因座提供基因型识别。举例来说,可以根据美国专利申请公开号2016/0085910和美国专利申请公开号2013/0296175中描述的方法和系统分析核酸序列,其全部主题明确地通过引用全部内容并入本文。

[0296] 在一个实现方式中,测序过程包括接收样品,该样品包含或怀疑包含核酸,例如DNA。样品可以来自已知或未知来源,例如动物(例如人)、植物、细菌或真菌。样品可直接从来源获取。例如,血液或唾液可直接取自个体。可备选地,可能无法直接从来源获得样品。然后,一个或多个处理器指导系统准备用于测序的样品。制备可包括去除外来物质和/或分离某些物质(例如DNA)。可以制备生物学样品以包括用于特定测定的特征。例如,可以制备生物样品以进行合成测序(SBS)。在某些实施方案中,制备可包括扩增基因组的某些区域。例如,制备可以包括扩增已知的包括STR和/或SNP的预定遗传基因座。可以使用预定的引物序列扩增遗传基因座。

[0297] 接下来,一个或多个处理器指导系统对样品进行测序。可以通过多种已知的测序方案进行测序。在特定的实现方式中,测序包括SBS。在SBS中,多个荧光标记的核苷酸用于对存在于光学基质表面(例如,至少部分地限定流动池中的通道的表面)上的多个扩增的DNA簇(可能数百万个簇)进行测序。流动池可包含用于测序的核酸样品,其中将流动池放置在适当的流动池支架内。

[0298] 可以制备核酸使得它们包含与未知靶序列相邻的已知引物序列。为了启动第一个SBS测序循环,一个或多个不同标记的核苷酸和DNA聚合酶等可以通过流体流动子系统流入/流过流动池。可以一次添加一种类型的核苷酸,或者可以对测序过程中使用的核苷酸进行特殊设计,使其具有可逆的终止特性,从而使测序反应的每个循环在几种类型的标记核苷酸(例如,A、C、T、G)存在下同时发生。核苷酸可以包括可检测的标记部分,例如荧光团。在四个核苷酸混合在一起的情况下,聚合酶能够选择正确的碱基进行整合,并且每个序列都可以延伸一个碱基。可通过使洗涤液流过流动池来洗去未结合的核苷酸。一个或多个激光可以激发核酸并诱导荧光。从核酸发出的荧光是基于掺入碱基的荧光团,并且不同的荧光团可以发射不同波长的发射光。可以将去封闭剂添加到流动池中,以从延伸和检测到的DNA链中去除可逆终止子基团。然后,可以通过使洗涤溶液流过流动池来洗去去封闭剂。然后,流动池准备好进行进一步的测序循环,从引入上述标记的核苷酸开始。流体和检测操作可

以重复几次以完成测序运行。举例的测序方法描述于例如,Bentley等人,Nature 456:53-59(2008)、国际公开号W0 04/018497;和美国专利号7,057,026;国际公开号W0 91/06678;国际公开号W0 07/123744;美国专利号7,329,492;美国专利号7,211,414;美国专利号7,315,019;美国专利号7,405,281和美国专利申请公开号2008/0108082,其每一个均通过引用并入本文。

[0299] 在一些实施方案中,可在测序之前或期间将核酸附着至表面并扩增。例如,可以使用桥式扩增来进行扩增以在表面上形成核酸簇。有用的桥式扩增方法描述于例如美国专利号5,641,658;美国专利申请公开号2002/0055100;美国专利号7,115,400;美国专利申请公开号2004/0096853;美国专利申请公开号2004/0002090;美国专利申请公开号2007/0128624;和美国专利申请公开号2008/0009420,其每一个均通过引用其全部内容并入本文。扩增表面上的核酸的另一种有用的方法是滚环扩增(RCA),例如,如Lizardi等人,Nat.Genet.19:225-232(1998)所述,和美国专利申请公开号2007/0099208A1,其每一个均通过引用并入本文。

[0300] 一个示例性的SBS协议利用具有可去除的3'嵌段的修饰的核苷酸,例如,如国际公开号W0 04/018497、美国专利申请公开号2007/0166705A1和美国专利号7,057,026中所描述的,其每一个均通过引用并入本文。例如,作为桥式扩增方案的结果,例如,可以将重复循环的SBS试剂传递至具有与其连接的靶核酸的流动池。可以使用线性化溶液将核酸簇转化成单链形式。线性化溶液可以包含例如,能够切割每个簇的一条链的限制性核酸内切酶。裂解的其他方法可以用作限制性内切酶或切口酶的替代方法,尤其包括化学裂解(例如,与高碘酸盐的二醇键的裂解)、通过用核酸内切酶裂解的无碱基位点的裂解(例如“USER”,如由NEB,Ipswich,Mass.,USA,part number M5505S供应的“USER”),通过暴露于热或碱下,将掺入扩增产物中的核糖核苷酸裂解,否则其由脱氧核糖核苷酸组成、光化学裂解或肽接头裂解。在线性化操作之后,可以在用于使测序引物与待测序的靶核酸杂交的条件下将测序引物递送至流动池。

[0301] 然后,可将流动池与SBS延伸试剂接触,该SBS延伸试剂具有修饰的核苷酸(带有可移动的3'嵌段)和荧光标记,条件是通过添加单个核苷酸即可扩展与每个靶核酸杂交的引物。仅将一个核苷酸添加到每个引物中,因为一旦将修饰的核苷酸掺入了与正在测序的模板区域互补的正在生长的多核苷酸链中,就没有游离的3'-OH基团可用于指导进一步的序列延伸,因此聚合酶不能添加其他核苷酸。可以去除SBS扩展试剂,并用扫描试剂代替,该试剂包含在辐射激发下保护样品的组分。扫描试剂的示例性组分在美国专利申请公开号2008/0280773A1和美国专利申请号13/018,255中进行了描述,其各自通过引用并入本文。然后可以在扫描试剂的存在下荧光检测延伸的核酸。一旦检测到荧光,就可以使用适合所用封闭基团的去封试剂去除3'嵌段。在W0004018497、US 2007/0166705A1和美国专利号7,057,026中描述了可用于各个封闭基团的示例性去封闭试剂,其各自通过引用并入本文。可以洗去去封闭剂,剩下的靶核酸与具有3'-OH基团的延伸引物杂交,现在可以添加其他核苷酸。因此,可以重复添加延伸试剂、扫描试剂和去封闭试剂的循环,并在一个或多个操作之间进行可选的洗涤,直到获得所需的序列。当每个修饰的核苷酸具有与其连接的不同标记(已知对应于特定碱基)时,可以在每个循环中使用单个延伸试剂递送操作来进行上述循环。不同的标记促进在每次掺入操作期间添加的核苷酸之间的区分。可备选地,每个循环可

包括延伸试剂递送的单独操作,然后是扫描试剂递送和检测的单独操作,在这种情况下,两个或多个核苷酸可以具有相同的标记并且可以基于已知的递送顺序来区分。

[0302] 尽管上面已针对特定的SBS方案讨论了测序操作,但应理解,可以根据需要执行用于对各种其他分子分析中的任何一种进行测序的其他方案。

[0303] 然后,系统的一个或多个处理器接收测序数据以进行后续分析。测序数据可以各种方式格式化,例如以.BAM文件格式。测序数据可以包括例如,许多样品解读。测序数据可包括具有核苷酸的相应样品序列的多个样品解读。尽管仅讨论了一个样品解读,但是应当理解,测序数据可以包括例如,数百、数千、数十万或数百万的样品解读。不同的样品解读可能具有不同数量的核苷酸。例如,样品解读的范围可以在10个核苷酸至约500个核苷酸或更多之间。样品解读可能跨越来源(多个)的整个基因组。作为一个实例,样品解读直接针对预定的遗传基因座,例如具有可疑STR或可疑SNP的那些遗传基因座。

[0304] 每个样品解读可以包括核苷酸序列,其可以称为样品序列、样品片段或靶序列。样品序列可以包括例如引物序列、侧翼序列和靶序列。样品序列内的核苷酸数目可包括30、40、50、60、70、80、90、100或更多。在一些实施方案中,一个或多个样品解读(或样品序列)包括至少150个核苷酸、200个核苷酸、300个核苷酸、400个核苷酸、500个核苷酸或更多。在一些实施方案中,样品解读可包括多于1000个核苷酸、2000个核苷酸或更多。样品解读(或样品序列)可在一端或两端包括引物序列。

[0305] 接下来,一个或多个处理器分析测序数据以获得潜在的变体识别(多个)和样品变体识别(多个)的样品变异频率。该操作也可以称为变体识别应用或变异识别器。因此,变体识别器识别或检测变体,且变体分类器将检测到的变体分类为体细胞或种系。可以根据本文的实现方式利用备选的变体识别器,其中可以基于所执行的测序操作的类型,基于感兴趣的样品特征等使用不同的变异识别器。变体识别应用的一个非限制性示例,例如Illumina Inc. (San Diego, CA)的PiscesTM应用程序,托管在<https://github.com/Illumina/Pisces>上,并在文章Dunn, Tamsen&Berry, Gwenn&Emig-Agius, Dorothea&Jiang, Yu&Iyer, Anita&Udar, Nitin & Strömberg, Michael. (2017)中进行了描述。Pisces:一个准确而多功能的单一样本体细胞和种系变体识别器595-595.10.1145/3107411.3108203,通过引用将其全部内容明确地合并于此。

[0306] 这样的变体识别应用可以包括四个顺序执行的模块:

[0307] (1)Pisces解读缝合器:通过将BAM中的配对解读(同一分子的解读一和解读二)缝合成共有解读来减少噪音。输出是缝合的BAM。

[0308] (2)Pisces变体识别器:识别小的SNV、插入和删除。Pisces包括变体折叠算法以合并被解读边界分解的变体、基本过滤算法以及简单的基于Poisson的变体置信度评分算法。输出为VCF。

[0309] (3)Pisces变体质量重新校准器(VQR):如果变体识别不可抵抗地遵循与热损伤或FFPE脱氨相关的模式,则VQR步骤将降低可疑变体识别的变体Q得分。输出为调整后的VCF。

[0310] (4)Pisces变体定相器(Scylla)使用读后的贪婪聚类方法将小的变体从克隆的子群组装成复杂的等位基因。这可以通过下游工具更准确地确定功能后果。输出为调整后的VCF。

[0311] 附加地或可替代地,该操作可以利用由托管在<https://github.com/Illumina/>

strelka的Illumina Inc.的和在文章T Saunders,Christopher&Wong,Wendy&Swamy,Sajani&Becq,Jennifer&J Murray,Lisa&Cheetham,Keira. (2012) 中描述的变体识别应用Strelka™应用。Strelka:从已测序的肿瘤正常样品对中准确的体细胞小变异识别。Bioinformatics(Oxford,England).28.1811-7.10.1093/生物信息学/bts271,其全部主题明确地通过引用全部内容并入本文。此外,附加地或可替代地,该操作可以利用托管在<https://github.com/Illumina/strelka>的Illumina Inc.和在文章Kim,S.,Scheffler,K.,Halpern,A.L.,Bekritsky,M.A.,Noh,E.,**Källberg**,M.,Chen,X.,Beyter,D.,Krusche,P.,and Saunders,C.T. (2017) 中描述的变体识别应用程序Strelka2™。Strelka2:用于临床测序应用的快速和准确的变体识别,其全部主题明确地通过引用全部内容并入本文。此外,附加地或可替代地,该操作可以利用变体注释/识别工具,例如托管在<https://github.com/Illumina/Nirvana/wiki>的Illumina Inc.的和在文章Stromberg,Michael& Roy,Rajat&Lajugie,Julien&Jiang,Yu&Li,Haochen&Margulies,Elliott. (2017) 中描述的Nirvana™应用程序。Nirvana:临床等级变体注释器。596-596.10.1145/3107411.3108204,通过引用将其全部内容明确地合并于此。

[0312] 这样的变体注释/识别工具可以应用不同的算法技术,例如Nirvana中公开的那些算法:

[0313] a.使用区间数组识别所有重叠的转录本:对于功能注释,我们可以识别所有与变体重叠的转录本,并且可以使用间隔树。然而,由于一组间隔可以是静态的,因此我们能够将其进一步优化为间隔数组。间隔树以 $O(\min(n, k \lg n))$ 时间返回所有重叠的转录本,其中 n 是树中的间隔数,而 k 是重叠的间隔数。实际上,由于与大多数变体的 n 相比, k 确实很小,因此间隔树上的有效运行时间为 $O(k \lg n)$ 。通过创建一个将所有间隔都存储在排序数组中的间隔数组,我们将其改进为 $O(\lg n + k)$,因此我们只需要找到第一个重叠间隔,然后枚举剩余的 $(k-1)$ 。

[0314] b.CNVs/SVs (Yu):可以提供“拷贝数变异”和“结构变体”的注释。与小变体的注释类似,可在在线数据库中注释与SV重叠的转录本以及以前报告的结构变体。与小的变体不同,并不是所有重叠的转录本都需要注释,因为太多的转录本会与大型SV重叠。相反,可以注释属于部分重叠基因的所有重叠转录本。具体地,对于这些转录本,可以报道受影响的内含子、外显子以及由结构变体引起的后果。提供允许输出所有重叠转录本的选项,但是可以报告这些转录本的基本信息,例如基因符号,标记是与转录本规范重叠还是部分重叠。对于每个SV/CNV,了解是否已经研究了这些变体以及它们在不同人群中的频率也很有趣。因此,我们报告了外部数据库(例如1000个基因组、DGV和ClinGen)中的重叠SV。为了避免使用任意的临界值来确定哪个SV重叠,相反可以使用所有重叠的转录本,并且可以计算出相互的重叠,即重叠长度除以这两个SV长度的最小值。

[0315] c.报告补充注释:补充注释有两种类型:小型和结构变体(SV)。SV可以建模为间隔,并使用上面讨论的间隔数组来标识重叠的SV。小变体被建模为点,并通过位置和(可选)等位基因进行匹配。这样,它们使用类似于二进制搜索的算法进行搜索。由于补充注释数据库可能很大,因此创建了一个更小的索引,以将染色体位置映射到补充注释所驻留的文件位置。索引是对象的排序数组(由染色体位置和文件位置组成),可以使用位置进行二进制搜索。为了使索引大小较小,将多个位置(最多达到某个最大计数)压缩到一个对象,该对象

存储第一个位置的值和仅存储后续位置的增量。由于我们使用二进制搜索,因此运行时间为 $O(\lg n)$,其中 n 是数据库中的项目数。

[0316] d. VEP缓存文件

[0317] e. 转录本数据库:转录本缓存(缓存)和补充数据库(SAdb)文件是数据对象(例如转录本和补充注释)的序列化转储。我们使用Ensembl VEP缓存作为缓存的数据源。为了创建缓存,将所有转录本插入间隔数组中,并将数组的最终状态存储在缓存文件中。因此,在注释期间,我们只需要加载一个预先计算的时间间隔数组并对它执行搜索。由于缓存已加载到内存中并且搜索非常快(如上所述),因此在Nirvana中查找重叠的笔录非常快(配置为少于总运行时间的1%)。

[0318] f. 补充数据库:SAdb的数据源在补充材料下列出。小型变体的SAdb是通过对所有数据源进行k路合并而产生的,以便数据库中的每个对象(由参考名称和位置标识)都具有所有相关的补充注释。Nirvana主页中详细记录了解析数据源文件时遇到的问题。为了限制内存使用,仅将SA索引加载到内存中。该索引允许快速查找文件位置以获取补充注释。然而,由于必须从磁盘中获取数据,因此添加补充注释已被确定为Nirvana的最大瓶颈(配置为总运行时间的30%)。

[0319] g. 后果和序列本体:Nirvana的功能注释(当提供时)遵循序列本体论(SO)(<http://www.sequenceontology.org/>)准则。有时,我们有机会识别当前SO中的问题,并与SO团队合作以改善注释状态。

[0320] 这样的变体注释工具可以包括预处理。例如,Nirvana包括来自外部数据源的大量注释,例如ExAC、EVS、1000基因组项目、dbSNP、ClinVar、Cosmic、DGV和ClinGen。为了充分利用这些数据库,我们必须清理来自它们的信息。我们实施了不同的策略来应对来自不同数据源的不同冲突。例如,对于同一位置有多个dbSNP条目和交替的等位基因的情况,我们将所有ID合并为一个逗号分隔的ID列表;如果同一等位基因有多个具有不同CAF值的条目,则使用第一CAF值。对于有冲突的ExAC和EVS条目,我们考虑样品计数数量,并使用具有更高样品计数的条目。在1000个基因组计划中,我们删除了冲突的等位基因的等位基因频率。另一个问题是不正确信息。我们主要从1000个基因组计划中提取了等位基因频率信息,然而,我们注意到,对于GRCh38,info字段中报告的等位基因频率并未排除基因型不可用的样品,从而导致并非所有样品都可用的变体频率降低。为了保证注释的准确性,我们使用所有个体水平基因型来计算真实的等位基因频率。众所周知,基于不同的比对,相同的变体可以具有不同的表示形式。为了确保我们可以准确报告已识别变体的信息,我们必须对来自不同资源的变体进行预处理,以使其具有一致的表示形式。对于所有外部数据源,我们修剪了等位基因以去除参考等位基因和替代等位基因中的重复核苷酸。对于ClinVar,我们直接解析xml文件,并对所有变体执行了五个素数对齐,这在vcf文件中经常使用。不同的数据库可以包含相同的信息集。为了避免不必要的重复,我们删除了一些重复的信息。例如,由于我们已经报告了1000个基因组中的这些变体,并且提供了更详细的信息,因此我们删除了DGV中具有1000个基因组项目数据源的变体。

[0321] 根据至少一些实现方式,变体识别应用程序提供对低频变体的识别、种系识别等。作为非限制性实例,变异识别应用可以在仅肿瘤样品和/或肿瘤正常配对样品上运行。变体识别应用可以搜索单核苷酸变异(SNV)、多核苷酸变异(MNV)、插入缺失等。变体识别应用识

别变异,同时过滤由于测序或样品制备错误导致的不匹配。对于每个变体,变体识别器识别参考序列、变体的位置以及潜在的一个或多个变体序列(例如,A至C SNV或AG至A缺失)。变体识别应用识别样品序列(或样品片段)、参考序列/片段和变异识别以指示存在变体。变体识别应用可以标识原始片段,并输出原始片段的名称,验证潜在变体识别的原始片段数量的计数,原始片段在其中发生支持变体的位置以及其他相关信息。原始片段的非限制性实例包括双工缝合片段,单工缝合片段、双工非缝合片段和单工非缝合片段。

[0322] 变体识别应用可以各种格式输出识别,例如.VCF或.GVCF文件。仅作为实例,变体识别应用可以被包括在MiSeqReporter管线中(例如,当在MiSeq®测序仪上实现时)。任选地,可以利用各种工作流程来实现该应用。该分析可以包括以指定方式分析样品解读以获得所需信息的单个方案或方案的组合。

[0323] 然后,一个或多个处理器执行与潜在变体识别有关的验证操作。验证操作可以基于质量得分和/或分层测试的层次结构,如下文所述。当验证操作验证或验证潜在的变体识别时,验证操作将变体识别信息(来自变体识别应用)传递给样品报告生成器。可替代地,当验证操作使潜在的变体识别无效或取消资格时,验证操作将相应的指示(例如,否定指示符、无识别指示符、无效识别指示符)传递给样品报告生成器。验证操作还可以传递与变体识别正确或无效识别指定正确的置信度有关的置信度得分。

[0324] 接下来,一个或多个处理器生成并存储样品报告。样品报告可以包括例如,关于样品的关于多个遗传基因座的信息。例如,对于预定的一组遗传基因座的每个遗传基因座,样品报告可以至少提供以下一项:提供基因型识别;指示无法进行基因型识别;提供基因型识别的确定性的置信度得分;或指出有关一个或多个遗传基因座的测定法可能存在的问题。样品报告还可以指示提供样品的个人的性别和/或指示样品包括多个来源。如本文所用,“样品报告”可以包括遗传基因座或遗传基因座的预定集合的数字数据(例如,数据文件)和/或遗传基因座或遗传基因座的集合的打印报告。因此,生成或提供可以包括创建数据文件和/或打印样品报告,或显示样品报告。

[0325] 样品报告可能指示已确定变体识别,但未通过验证。当确定变体识别无效时,样品报告可以指示有关确定不验证变体识别的基础的其他信息。例如,报告中的附加信息可以包括原始片段的描述以及原始片段支持或与变体识别相抵触的程度(例如,计数)。附加地或可替代地,报告中的附加信息可以包括根据本文所述的实现方式获得的质量得分。

[0326] 变体识别应用

[0327] 本文公开的实现方式包括分析测序数据以识别潜在的变异识别。可以对存储的数据执行变体识别,以进行先前执行的排序操作。附加地或可替代地,它可以在执行测序操作的同时实时地进行。每个样品解读被分配给相应的遗传基因座。可以基于样品解读的核苷酸的序列,或者换句话说,样品解读内的核苷酸的顺序(例如,A、C、G、T),将样品解读分配给相应的遗传基因座。基于该分析,可以将样品解读指定为包括特定遗传基因座的可能变体/等位基因。样品解读可以与已经指定为包括遗传基因座的可能变体/等位基因的其他样品解读一起收集(或聚集或分类)。分配操作也可以称为识别操作,其中样品解读被识别为可能与特定的遗传位置/基因座相关联。可以分析样品解读以定位一个或多个区分样品解读与其他样品解读的核苷酸的鉴定序列(例如,引物序列)。更具体地,一个或多个识别序列可以将来自其他样品解读的样品解读识别为与特定遗传基因座相关。

[0328] 分配操作可以包括分析识别序列的n个核苷酸序列以确定该识别序列的n个核苷酸序列是否与一个或多个选择序列有效匹配。在特定实现方式中,分配操作可以包括分析样品序列的前n个核苷酸以确定样品序列的前n个核苷酸是否与一个或多个选择序列有效匹配。数字n可以具有各种值,可以将其编程到协议中或由用户输入。例如,数量n可以定义为数据库内最短选择序列的核苷酸数量。数量n可以是预定数量。预定数量可以是例如10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29或30个核苷酸。然而,在其他实现方式中可以使用更少或更多的核苷酸。数字n也可以由个人(例如系统的用户)选择。数量n可以基于一个或多个条件。例如,数量n可以定义为数据库内最短引物序列的核苷酸数量或指定数量,以较小的数量为准。在一些实现方式中,可使用n的最小值,例如15,使得小于15个核苷酸的任何引物序列可被指定为例外。

[0329] 在某些情况下,识别序列的n个核苷酸序列可能与选择序列的核苷酸不完全匹配。然而,如果识别序列与选择序列几乎相同,则识别序列可以有效地与选择序列匹配。例如,如果识别序列的n个核苷酸序列(例如,前n个核苷酸)的序列与选择序列具有不超过指定数目的错配(例如3个)和/或指定的移位数量(例如2个)的选择序列匹配,则可为遗传座位识别样品解读。可以建立规则,使得每个错配或移位都可以算作样品解读和引物序列之间的差异。如果差异的数目小于指定数目,则可以为对应的遗传基因座(即,分配给相应的遗传基因座)识别样品解读。在一些实现方式中,可以基于样品解读的识别序列和与基因座相关的选择序列之间的差异数来确定匹配得分。如果匹配得分超过指定的匹配阈值,则可以将与选择序列相对应的遗传基因座指定为样品解读的潜在基因座。在一些实现方式中,可以进行后续分析以确定样品解读是否被称为遗传基因座。

[0330] 如果样品解读有效匹配数据库中的选择序列之一(即,如上所述完全匹配或几乎匹配),则将样品解读分配或指定给与选择序列相关的遗传基因座。这可以称为基因座识别或临时基因座识别,其中针对与选择序列相关的遗传基因座识别样品解读。然而,如上所述,可能需要多于一个的基因座进行样品解读。在这样的实现方式中,可以执行进一步的分析以识别或分配仅针对潜在遗传基因座之一的样品解读。在一些实现方式中,与参考序列数据库比较的样品解读是来自配对末端测序的第一解读。当执行配对末端测序时,将获得与样品解读相关的第二个解读(代表原始片段)。分配后,使用分配的解读进行的后续分析可以基于已为分配的解读识别的遗传基因座的类型。

[0331] 接下来,分析样品解读以识别潜在的变体识别。其中,分析结果确定了潜在变体识别、样品变体频率、参考序列以及感兴趣的基因组序列中变异发生的位置。例如,如果已知遗传基因座包含SNP,则可以对该遗传基因座识别的指定解读进行分析以识别指定解读的SNP。如果已知遗传基因座包含多态性重复DNA元素,则可以分析指定的解读,以鉴定或表征样品解读中的多态性重复DNA元素。在一些实现方式中,如果分配的解读与STR基因座和SNP基因座有效匹配,则可以将警告或标志分配给样品解读。样品解读可以被指定为STR基因座和SNP基因座。该分析可以包括根据比对协议来比对所分配的解读,以确定所分配的解读的序列和/或长度。比对方案可以包括2013年3月15日提交的国际专利申请号PCT/US2013/030867(公开号W0 2014/142831)中描述的方法,其通过引用其全部内容并入本文。

[0332] 然后,一个或多个处理器分析原始片段,以确定在原始片段内的对应位置是否存在支持变体。可以识别各种类型的原始片段。例如,变体识别器可以识别出原始片段的类

型,该原始片段表现出验证原始变体识别的变体。例如,原始片段的类型可以表示双工缝合片段、单工缝合片段、双工未缝合片段或单工未缝合片段。替代上述实例或除上述实例之外,还可以任选地识别其他原始片段。与识别每种类型的原始片段有关,变体识别器还识别原始片段在其中发生支持变体的位置,以及表现出支持变体的原始片段的数量。例如,变体识别器可以输出指示,其识别出10个原始片段的解读以表示在特定位置X具有支持变体的双工缝合片段。变体识别器还可以输出指示,其五个原始片段的解读被标识为代表在特定位置Y具有支持变体的单工未缝合片段。变体识别器还可以输出许多与参考序列相对应的原始片段,因此不包括支持变体,否则其提供证据来验证在感兴趣的基因组序列处的潜在变体呼叫。

[0333] 接下来,维护包括支持变体在内的原始片段的计数,以及支持变体发生的位置。附加地或可替代地,可以保持原始片段的计数,该原始片段在感兴趣的位置(相对于样品解读或样品片段中潜在的变异识别的位置)不包括支持变体。附加地或可替代地,可以维持对应于参考序列并且不认证或确认潜在变体识别的原始片段的计数。确定的信息将输出到变体识别验证应用程序,包括支持潜在变体识别的原始片段的计数和类型,原始片段中支持变异的位置,不支持潜在变体识别等的原始片段的计数。

[0334] 当识别出潜在变体识别时,该过程输出潜在变体识别、变体序列、变体位置和与其相关联的参考序列的指示。变体识别被指定为代表“潜在”变体,因为错误可能导致识别过程识别错误的变量。根据本文的实现方式,分析了潜在的变体识别以减少和消除错误变体或假阳性。附加地或可替代地,该过程分析与样品解读相关联的一个或多个原始片段,并输出与原始片段相关联的对应变体识别。

[0335] 基因组学中的深度学习

[0336] 遗传变异可以帮助解释许多疾病。每个人都有独特的遗传密码,并且一群人中有许多遗传变体。大多数有害的遗传变体都已通过自然选择从基因组中清除。重要的是要确定哪些遗传变异可能是致病的或有害的。这将有助于研究人员关注可能的致病基因变体,并加快许多疾病的诊断和治愈速度。

[0337] 对变体的特性和功能效应(例如致病性)进行建模是基因组学领域一项重要但具有挑战性的任务。尽管功能基因组测序技术发展迅速,但由于细胞类型特异性转录调控系统的复杂性,对变体的功能后果进行解释仍然是一个巨大的挑战。

[0338] 关于致病性分类器,深度神经网络是一种利用多个非线性和复杂的转换层来连续对高级特征建模的人工神经网络。深度神经网络通过反向传播提供反馈,该反馈携带观察到的和预测的输出之间的差异以调整参数。深度神经网络随着大型训练数据集的可用性、并行和分布式计算的强大功能以及复杂的训练算法而不断发展。深度神经网络促进了计算机视觉、语音识别和自然语言处理等众多领域的重大进步。

[0339] 卷积神经网络(CNN)和递归神经网络(RNN)是深度神经网络的组成部分。卷积神经网络在图像识别方面尤其成功,其架构包括卷积层、非线性层和采样层。递归神经网络旨在利用输入数据的顺序信息,并在感知器、长短期记忆单元和门控递归单元等构建块之间建立循环连接。此外,还针对有限的上下文提出了许多其他新出现的深度神经网络,如深时空神经网络、多维递归神经网络和卷积自动编码器。

[0340] 训练深度神经网络的目标是优化每一层的权重参数,从而逐渐将较简单的特征合

并为复杂的特征,以便从数据中学习最合适的层次表示。优化过程的单个循环组织如下。首先,给定一个训练数据集,正推法按顺序计算每一层中的输出,并通过网络向前传播函数信号。在最后的输出层中,目标损失函数测量推断输出和给定标签之间的误差。为了尽量减少训练误差,逆推法使用链式法则来反向传播错误信号,并计算相对于整个神经网络中所有权重的梯度。最后,使用基于随机梯度下降的优化算法更新权重参数。批量梯度下降法对每个完整数据集执行参数更新,而随机梯度下降法通过对每个小型数据示例集执行更新来提供随机近似。随机梯度下降法产生了几种优化算法。例如,Adagrad和Adam训练算法执行随机梯度下降,同时根据每个参数的更新频率和梯度矩自适应地修改学习速率。

[0341] 深度神经网络训练中的另一个核心要素是正则化,它是指旨在避免过度拟合从而实现良好泛化性能的策略。例如,权重衰减为目标损失函数增加一个惩罚项,使重量参数收敛到更小的绝对值。在训练过程中,丢弃(dropout)从神经网络中随机删除隐藏的单元,可以将其视为可能的子网络的集合。为了增强丢弃的能力,提出了一种新的激活函数maxout和一种称为rnnDrop的递归神经网络的丢弃变体。此外,批归一化通过对小批量中每次激活的标量特征进行归一化并学习每个均值和方差作为参数,提供了一种新的正则化方法。

[0342] 考虑到序列数据是多维和高维的,深度神经网络由于其广泛的适用性和增强的预测能力,在生物信息学研究方面具有广阔的前景。卷积神经网络已被应用于解决基因组学中基于序列的问题,如基序发现、致病性变体识别和基因表达推断等。卷积神经网络使用一种权重共享策略,这种策略对于研究DNA特别有用,因为它可以捕获序列基序,这些基序是DNA中很短并且重复出现的局部模式,被认为具有重要的生物学功能。卷积神经网络的一个标志是卷积滤波器的使用。与基于精心设计和手动制作特征的传统分类方法不同,卷积滤波器执行特征的自适应学习,类似于将原始输入数据映射到知识的信息表示这一过程。从这个意义上说,卷积滤波器用作一系列的基序扫描器,因为一组这样的滤波器能够识别输入中的相关模式,并在训练过程中自我更新。递归神经网络可以在不同长度的序列数据(如蛋白质或DNA序列)中捕获长距离依赖关系。

[0343] 因此,一个预测变体致病性的强大计算模型可以为基础科学和转化研究带来巨大的好处。

[0344] 目前,只有25-30%的罕见疾病患者通过检测蛋白编码序列获得分子诊断,这表明剩余的诊断率可能存在于99%的非编码基因组中。在这里,我们描述一个新型的深度学习网络,该网络从任意的前体mRNA转录本序列中准确预测剪接点,从而能够精确预测非编码变体的剪接改变效应。具有预测剪接改变结果的同义和基因内突变在RNA测序(RNA-seq)上具有很高的验证率,并且在人群中具有极大的危害性。与健康对照组相比,在自闭症和智力残疾患者中,具有预测剪接改变结果的新生突变在自闭症和智力障碍患者中显著丰富,并且在28名患者中有21人根据RNA测序数据进行了验证。我们估计,在罕见遗传性疾病患者中,有9-11%的致病性突变是由这类以前未被充分认识的疾病变异引起的。

[0345] 外显子组测序改变了罕见遗传性疾病患者和家庭的临床诊断,当作为一线检测方法使用时,大大减少了诊断过程的时间和成本(Monroe等人,2016年;Stark等人,2016年;Tan等人,2017年)。然而,在罕见遗传疾病群组中,外显子组测序的诊断率约为最高25-30%,即使结合外显子组和微阵列检测,大多数患者仍无法诊断(Lee等人,2014年;Trujillano等人,2017年;Yang等人,2014年)。非编码区域在基因调控中起着重要作用,占

人类复杂疾病无偏见全基因组关联研究中发现的因果疾病基因位点的90% (Ernst等人, 2011年; Farh等人, 2015年; Maurano等人, 2012年), 表明外显非编码变体也可能是罕见遗传疾病中重要的因果突变负担。事实上, 尽管位于基本GT和AG剪接二核苷酸之外, 但破坏mRNA剪接正常模式的外显非编码变体 (通常被称为隐秘剪接变体) 一直被认为在罕见遗传疾病中发挥着重要作用 (Cooper等人, 2009年; Padgett, 2012年; Scotti和Swanson, 2016年; Wang和Cooper, 2007年)。然而在临床实践中, 由于我们对剪接代码不完全理解, 以及由此难以准确识别基本GT和AG二核苷酸以外的剪接改变变体, 因此隐秘剪接突变往往被忽视 (Wang和Burge, 2008年)。

[0346] 最近, RNA测序已经成为检测孟德尔疾病中剪接异常的一种有前途的测定方法 (Cummings等人, 2017年; Kremer等人, 2017年), 但到目前为止, 其在临床环境中的效用仍然仅限于少数已知相关细胞类型并可进行活检的病例。潜在剪接改变变体的高通量筛选 (Soemedi等人, 2017年) 扩大了剪接变异的表征, 但由于可能出现剪接改变突变的基因组空间非常大, 用于评估遗传疾病中随机新生突变的实用性较差。从任意前体mRNA序列剪接的一般预测将有可能精确预测非编码变体的剪接改变后果, 从而显著改善遗传疾病患者的诊断。迄今为止, 尽管在特定应用中取得了进展, 例如对核心剪接基序的序列特征进行建模 (Yeo和Burge, 2004年), 表征外显子剪接增强子和沉默子 (Fairbrother等人, 2002年; Wang等人, 2004年), 以及预测盒式外显子的包涵性 (Barash等人, 2010年; Jha等人, 2017年; Xiong等人, 2015年), 但是从接近剪接体特异性的原始序列剪接的一般预测模型仍然难以捉摸。

[0347] 将长前体mRNA剪接到成熟转录本中以其精确性和剪接改变突变的临床严重性而引人注目, 但其细胞机制决定其特异性的机制仍不完全清楚。在这里, 我们训练了一个接近电脑模拟的剪接体准确性的深度学习网络, 从前体mRNA序列识别外显子-内含子边界的准确度为95%, 在RNA测序预测功能性隐秘剪接突变的验证率超过80%。预测会改变剪接的非编码变体在人类中具有极大的危害性, 新产生的隐秘剪接突变中有80%经历了负选择, 类似于其他类型的蛋白质截断变异的影响。自闭症和智力障碍患者中新生隐秘剪接突变与通过蛋白质截断突变而反复突变的基因是一致的, 从而可以发现更多的候选疾病基因。我们估计, 在罕见遗传病患者中, 高达24%的外显因果突变是由于这类以前未被充分认识的疾病变异所致, 这强调了需要改进99%临床测序应用的非编码基因组解释。

[0348] 临床外显子组测序彻底改变了罕见遗传疾病患者和家庭的诊断方法, 当作为一线检测方法使用时, 大大减少了诊断过程的时间和成本。然而据报道, 在多个大型的罕见疾病患者及其父母群体中, 外显子组测序的诊断率为25%-30%, 即使结合外显子组和微阵列检测, 大多数患者仍无法诊断。非编码基因组在基因调控中非常活跃, 非编码变体占常见疾病GWAS采样数的最多90%, 这表明非编码基因组中的罕见变异可能也占罕见遗传病、肿瘤病等外显疾病中因果突变的相当比例。然而, 难以解释非编码基因组中的变体意味着, 在大型结构变体之外, 相对于对临床管理影响最大的罕见外显变体, 非编码基因组目前几乎没有提供额外的诊断价值。

[0349] 在规范GT和AG剪接二核苷酸之外, 剪接改变突变的作用早已受到人们的高度重视。事实上, 这些隐秘剪接变体是一些罕见遗传疾病, 如糖原贮积病XI (庞贝病) 和红细胞原卟啉症的最常见突变。内含子5'和3'末端的延伸剪接基序是高度退化的, 基因组中频繁出

现同等好的基序,这使得现有方法无法准确预测哪些非编码变体可能导致隐秘剪接。

[0350] 为了更好地了解剪接体如何实现其特异性,我们训练了一个深度学习神经网络,仅使用转录本序列作为输入来预测前体mRNA转录本中的每个核苷酸,无论是剪接受体、剪接供体,还是两者都不是(图37A)。使用偶数染色体上的规范转录本作为训练集,而奇数染色体上的转录本进行测试(不包括旁系同源基因),深度学习网络以95%的准确率调用外显子-内含子边界,甚至超过100KB的转录本,如CFTR通常重建重建到核苷酸精度(图37B)。

[0351] 接下来,我们试图了解网络用如此卓越的精度识别外显子-内含子边界所使用的特异性决定子。与以前对统计或人工设计特征进行操作的分类器不同,深度学习以分层方式直接从序列中学习特征,从而实现从长程序列上下文中赋予额外的特异性。事实上,我们发现网络的准确度高度依赖于作为输入提供给网络的侧接于待预测核苷酸的序列上下文的长度(表1),并且当我们训练的深度学习模型时仅使用40-nt的序列,性能仅略高于现有统计方法。这表明,深度学习在识别单个9-23nt剪接基序的现有统计方法基础上几乎没有增加,但更广泛的序列上下文是区分功能性剪接位点与具有相同强度的基序的非功能性位点的关键。要求网络预测出现序列扰动的外显子表明破坏供体基序通常也会导致受体信号消失(图37C),正如在体内外显子跳读事件经常观察到的一样,表明仅仅通过要求在可接受的距离上进行强受体和供体基序的配对即可获得相当程度的特异性。

[0352] 虽然大量证据表明,外显子长度的实验性扰动对外显子增加与外显子跳读有强烈的影响,但这并不能解释为什么深度学习网络的准确度继续增加到超过1000-nt上下文。为了更好地区分本地剪接基序驱动的特异性和长距离特异性决定子,我们训练了一个仅接收100-nt上下文输入的本地网络。使用本地网络对已知的进行评分,我们发现外显子和内含子都有最佳长度(外显子为不超过115nt,内含子为不超过1000nt),在该长度处基序强度最小(图37D)。这种关系在10000-nt深度学习网络中并不存在(图37E),表明在宽上下文的深度学习网络中,内含子和外显子长度的变化已经被充分考虑在内。值得注意的是,从未将内含子和外显子的边界赋予宽上下文的深度学习模型,表明它能够仅从序列中推断内含子和外显子的位置来推导出这些距离。

[0353] 对六聚体空间的系统搜索还表明,深度学习网络利用了外显子-内含子定义中的基序,尤其是从-34至-14位置的分支点基序TACTAAC,外显子末端附近良好表征的外显子剪接增强子GAAGAA,以及通常是多聚嘧啶序列一部分但也表现为外显子剪接沉默子的聚尿苷酸基序(图21、22、23和24)。

[0354] 我们通过预测参考转录本序列以及包含变体的替代转录本序列上的外显子-内含子边界,并寻找外显子-内含子边界的任何变化,将深度学习网络扩展到评估剪接改变功能的遗传变体。来自60,706人的最新聚合外显子组数据使我们能够通过检查等位基因频谱中的分布来评估阴性选择对预测会改变剪接功能的变体的影响。我们发现,预测的隐秘剪接变体强烈地表现在负选择之下(图38A),这可以通过它们与预期计数相比,在高等位基因频率下的相对损耗来证明,而且它们的损耗幅度与AG或GT剪接破坏变体和停止增益变体相当。在考虑可能导致移码的隐秘剪接变体超过引起帧内改变的隐秘剪接变体时,负选择的影响更大(图38B)。根据移码隐秘剪接变体与其他类型蛋白质截断变异相比的损耗,我们估计88%的可信预测的隐秘剪接突变是功能性的。

[0355] 虽然没有像外显子组数据那样多的聚合全基因组数据可用,从而限制了在深度内

含子区域检测自然选择影响的能力,但我们还能够计算隐秘剪接突变在远离外显子区域的观测计数与预期计数。总体而言,我们观察到在距离外显子-内含子边界>50nt处的隐秘剪接突变中有60%的损耗(图38C)。与外显子组相比,衰减后的信号可能是样本量较小、具有完整基因组数据的组合,预测深度内含子变体的难度更大。

[0356] 我们还可以使用观察到的与预期的隐秘剪接变体数量对比来估计待选择的隐秘剪接变体的数量,以及如何将其与其他类型的蛋白质截断变异进行比较。由于隐秘剪接变体可能仅部分废除了剪接功能,我们还在更宽松的阈值下评估了观察到的隐秘剪接变体与预期的隐秘剪接变体的数量,并且估计与ExAC数据集中的罕见AG或GT剪接破坏变体相比,有害的罕见隐秘剪接变体的数量大约多3倍(图38D)。每个个体携带约最多20个罕见的隐秘剪接突变,约等于蛋白质截断突变的数目(图38E),尽管并非所有这些变体都完全废除剪接功能。

[0357] 最新发布的GTEx数据包括148个同时具有多个组织部位的全基因组测序和RNA测序的个体,使我们能够直接在RNA测序数据中寻找罕见隐秘剪接变体的影响。为了估计罕见疾病测序中遇到的情况,我们仅考虑了罕见变体(GTEx群组中的单例,在1000个基因组中的等位基因频率<1%),并将它们与具有该变体的个体独有的剪接事件配对。尽管基因表达和组织表达的差异以及剪接异常的复杂性导致难以评估深度学习预测的灵敏度和特异性,但我们发现在严格的特异性阈值下,超过90%的罕见隐秘剪接突变在RNA测序上得到验证(图39A)。RNA测序中出现的大量异常剪接事件似乎与深度学习分类器预测的效果有限的变体有关,表明它们仅部分影响剪接功能。在这些更敏感的阈值下,预计大约75%的新型剪接点会导致剪接功能的畸变(图38B)。

[0358] 深度学习网络成功预测了对种群测序数据有严重危害并在RNA测序上具有很高验证率的隐秘剪接变体,这表明该方法可用于识别罕见疾病测序研究中的其他诊断。为了验证这一假设,我们检查了用自闭症和神经发育障碍的外显子组测序研究中的新生变体,并证明与其健康的兄弟姐妹相比,受影响的个体的隐秘剪接突变显著丰富(图40A)。此外,隐秘剪接突变的富集程度比蛋白质截断突变的富集程度略低,这表明我们预测的隐秘剪接突变中大约90%是功能性的。根据这些值,大约不超过20%的致病蛋白质截断变体可归因于外显子和紧邻外显子的核苷酸的隐秘剪接突变(图40B)。将这个数字外推到能够查询整个内含子序列的全基因组研究,我们估计在罕见遗传性疾病中,24%的因果突变是由于隐秘剪接突变引起的。

[0359] 我们估计了每个单独基因出现新生隐秘剪接突变的概率,使我们能够估计候选疾病基因中与偶然性相比隐秘剪接突变的富集程度。新生隐秘剪接突变在以前被蛋白质截断变异破坏的基因中高度富集(图40C),表明它们主要通过单倍型不足而不是其他作用方式引起疾病。与仅使用蛋白质截断变体相比,将预测的隐秘剪接突变添加到蛋白质截断变体列表中,我们可以识别3个额外的自闭症疾病基因和11个额外的智力障碍疾病基因(图40D)。

[0360] 为了评估在无法获得可能的疾病组织(在这种情况下为大脑)的患者中验证隐秘剪接突变的可行性,我们对来自Simon的单纯形集合的具有预测新生隐秘剪接突变的37位个体进行了深入的RNA测序,并寻找该个体中存在以及该实验中所有其他个体以及GTEx群组149个个体中都没有的异常剪接事件。我们发现37名患者中的NN显示出对RNA测序的独特

的异常剪接(图40E),这可以用预测的隐秘剪接变体解释。

[0361] 综上所述,我们验证了一个深度学习模型,该模型能够准确地预测隐秘剪接变体,并且具有足够的精度,可以用于识别罕见遗传病中的因果致病突变。我们估计,仅考虑蛋白质编码区,忽略了很大一部分由隐秘剪接引起的罕见疾病诊断,并强调需要开发一种方法来解释非编码基因组中外显罕见变异的影响。

[0362] 结果

[0363] 使用深度学习从原始序列准确预测剪接

[0364] 我们仅使用前体mRNA转录本的基因组序列作为输入,构建了一个深度残差神经网络(He等人,2016年a),用来预测前体mRNA转录本中的每个位置是剪接供体,剪接受体还是两者都不是(图37A和图21、22、23和24)。因为剪接供体和剪接受体可能被成千上万个核苷酸分开,我们采用了一种新型网络架构,由32个膨胀的卷积层组成(Yu和Koltun,2016年),可以识别跨越非常大的基因组距离的序列决定子。与以前只考虑附近外显子-内含子边界的短核苷酸窗口(Yeo和Burge,2004年),或依赖人为设计特征(Xing等人,2015年)或如表达或剪接因子结合等实验数据(Jha等人,2017年)的方法,我们的神经网络评估侧翼上下文序列的10,000个核苷酸来预测前体mRNA转录本中每个位置的剪接功能,从而直接从原始序列中学习剪接决定子。

[0365] 我们在人类染色体子集上使用了带有GENCODE注释的前体mRNA转录本序列(Harrow等人,2012来)来训练神经网络的参数,并在不包括旁系同源基因的剩余染色体上使用转录本来测试该网络的预测。对于测试数据集中的前体mRNA转录本,网络预测剪接点具有95%的top-k准确度,这是预测位点数等于测试数据集中实际剪接位点数的阈值处,正确预测的剪接位点的比例(Boyd等人,2012年;Yeo和Burge,2004年)。即使是超过100kb的基因,如CFTR,也常常能够完美地重建到核苷酸精度(图37B)。为了确认网络不仅仅依赖于外显子序列偏倚,我们还用长的非编码RNA对网络进行了测试。尽管非编码转录本注释不完整可能会降低我们的准确度,但该网络预测lincRNA中已知的剪接点的top-k准确度为84%(图42A和42B),表明它可以近似模拟没有蛋白质编码选择压力的任意序列的剪接体行为。

[0366] 对于测试数据集中每个带有GENCODE注释的外显子(不包括每个基因的第一个和最后一个外显子),我们还根据来自基因和组织表达图谱(GTEx)的RNA测序数据,检查了网络的预测得分是否与支持外显子增加和外显子跳读的解读比例相关(GTEx联盟等,2015年)(图37C)。在GTEx组织中进行组成性剪接的外显子的预测得分分别接近1或0,而经历了较大程度选择性剪接(在样品中平均10%-90%的外显子增加)的外显子则趋向于中间值(皮尔逊相关系数=0.78, $P \approx 0$)。

[0367] 接下来,我们试图了解网络利用的序列决定子来实现其非凡的准确度。我们对带注释的外显子附近的每个核苷酸进行了系统的电脑模拟替代,测量了在相邻剪接位点对网络预测得分的影响(图37E)。我们发现,打乱剪接供体基序的序列经常导致网络预测上游剪接受体的位点也将丢失,如体内外显子跳读事件所观察到的,表明配对的上游受体基序和设置在最佳距离的下游供体基序之间的外显子定义赋予了相当程度的特异性(Berget, 1995年)。有助于剪接信号的其他基序包括SR-蛋白家族和分支点的良好表征的结合基序(图43A和43B)(Fairbrother等人,2002年;Reed和Maniatis,1988年)。这些基序的作用高度依赖于它们在外显子中的位置,表明它们的作用包括通过区分竞争受体和供体位点来指定

内含子-外显子边界的精确定位。

[0368] 使用不同的输入序列上下文训练网络会显著影响剪接预测的准确度(图37E),表明距离剪接位点最远10,000nt的长距离序列决定子对于从大量具有接近最优基序的非功能位点中识别功能性剪接点至关重要。为了检查远距离和短距离特异性决定子,我们比较了按80nt序列上下文训练的模型(SpliceNet-80nt)与按10000nt上下文训练的完整模型(SpliceNet-10k)分配给带注释剪接点的得分。在80nt序列上下文训练的网络将更低的得分分配给与典型长度的外显子或内含子相连的剪接点(外显子为150nt,内含子为最多1000nt)(图37F),这与先前的观察到的与外显子和内含子异常长或短的剪接位点相比,此类位点的剪接基序往往更弱相一致(Amit等人,2012年;Gelfman等人,2012年;Li等人,2015年)。相比之下,在10,000nt序列上下文训练的网络显示出对平均长度的内含子和外显子的偏爱,尽管它们的剪接基序更弱,因为它可以解释外显子或内含子长度所赋予的远距离特异性。长不间断内含子中更弱基序的跳读与在没有外显子停顿的情况下通过实验观察到的更快的RNA聚合酶II延伸相一致,这可能使剪接体识别次优基序的时间更少(Close等人,2012年;Jonkers等人,2014年;Veloso等人,2014年)。我们的发现表明,平均的剪接点具有良好的长距离序列决定子,这些决定子具有相当大的特异性,这解释了在大多数剪接基序较高的序列退化耐受程度。

[0369] 由于剪接是共同转录发生的(Cramer等人,1997年;Tilgner等人,2012年),染色质状态与共转录剪接之间的相互作用也可能指导外显子的定义(Luco等人,2011年),并有可能在染色质状态可从原始序列预测的情况下被网络利用。特别是,对核小体定位的全基因组研究表明,外显子中核小体的占有率更高(Andersson等人,2009年;Schwartz等人,2009年;Spies等人,2009年;Tilgner等人,2009年)。为了测试网络是否使用核小体定位的序列决定子进行剪接位点预测,我们让一对间隔150nt(大约是平均外显子的大小)的最佳受体和供体基序穿过基因组,并要求网络预测基序对是否将导致在该基因座处的外显子增加(图37G)。我们发现,即使在基因间区域(斯皮尔曼相关系数=0.36, $P \approx 0$),预测有利于外显子增加的位置也与高核小体占有率相关,并且这种效应在控制GC含量后仍然存在(图44A)。这些结果表明,该网络已经隐式学习了从原始序列预测核小体的位置,并将其用作外显子定义的特异性决定子。与平均长度的外显子和内含子相似,位于核小体上的外显子具有更弱的局部剪接基序(图44B),与存在补偿因子情况下对退化基序的更大耐受性相一致(Spies等人,2009年)。

[0370] 尽管已有多项研究报告了外显子与核小体占有率之间的相关性,但尚未明确确定外显子定义中核小体定位的因果作用。利用来自基因型组织表达(GTE_x联盟等,2015年)群组的149个具有RNA测序和全基因组测序的个体的数据,我们发现了单个个体专有的新外显子,并且与私有剪接位点创建基因突变相对应。这些私有外显子创造事件与K562和GM12878细胞中现有的核小体定位显著相关(通过排列检验, $P=0.006$,图37H),即使这些细胞系最有可能缺少相应的私有基因突变。我们的结果表明,如果产生的新外显子覆盖现有核小体占据的区域,则遗传变异更可能触发新外显子的产生,这支持了核小体定位在促进外显子定义中的因果作用。

[0371] 验证RNA测序数据中预测的隐秘剪接突变

[0372] 我们通过预测参考前体mRNA转录本序列以及包含变体的替代转录本序列的外显

子-内含子边界,并取得分之间的差值(Δ 得分),将深度学习网络扩展到评估剪接改变功能的遗传变体(图38A)。重要的是,该网络只根据参考转录本序列和剪接点注释进行了训练,在训练过程中从未见过变异数据,这使得变异效果的预测成为对网络准确建模剪接序列决定子能力的一项具有挑战性的测试。

[0373] 我们在GTEx群组(GTEx联盟等,2015年)中研究了RNA测序数据中隐秘剪接变体的影响,该研究包括同时具有全基因组测序和来自多个组织的RNA测序的149个个体。为了模拟罕见疾病测序中遇到的情况,我们首先关注罕见的私有突变(仅在GTEx群组中的一个个体出现)。我们发现,被神经网络预测具有功能性后果的私有突变在新颖的私有剪接接点和私有外显子跳读事件的跳过外显子边界处得到了强烈的富集(图38B),表明这些预测中有很大部分是功能性的。

[0374] 为了量化剪接位点产生变体对正常和异常剪接同种型的相对产生的影响,我们测量了支持新型剪接事件的解读数,作为覆盖该位点的解读总数的一部分(图38C)(Cummings等人,2017年)。对于剪接位点的破坏变体,我们观察到许多外显子的外显子跳读基线率较低,并且变异的作用是增加外显子跳读解读的分数。因此,我们采用了两种效果中的较大者,计算了破坏的剪接点处剪接的解读分数的减少和跳读外显子的解读分数的增加(图45和STAR方法)。

[0375] 自信地预测的隐秘剪接变体(Δ 得分 ≥ 0.5)以基本GT和AG剪接破坏速率四分之三的的速率对RNA测序进行了验证(图38D)。隐秘剪接变体的验证率和效果大小均密切跟踪其 Δ 得分(图38D和38E),表明该模型的预测得分可以很好地反映变体的剪接改变潜力。经验证的变体,尤其是得分较低的变体(Δ 得分 < 0.5),通常不完全外显,并导致在RNA测序数据中产生异常和正常转录本混合物的选择性剪接(图38E)。由于未考虑到剪接同种型的变化和无意义介导的衰变,我们对验证率和影响大小的估计是保守的,可能会低估真实值,这会优先降解异常剪接的转录本,因为它们经常引入过早的终止密码子(图38C和图45)。这可以通过破坏基本GT和AG剪接二核苷酸的变体的平均作用大小来证明,其小于完全外显杂合变体预期值的50%。

[0376] 对于在至少十分之三的观察到的mRNA转录本拷贝中产生异常剪接同种型的隐秘剪接变体,当变体是附近外显子时,网络的灵敏度为71%,当变体处于深度内含子序列时,网络的灵敏度为41%(Δ 得分 ≥ 0.5 ,图38F)。这些发现表明,深度内含子变体的预测更具挑战性,这可能是由于深度内含子区域包含较少的被选择存在于外显子附近的特异性决定子。

[0377] 为了对照现有方法对我们的网络性能进行基准测试,我们选择了三种在文献中被引用用于罕见遗传疾病诊断的热门分类器:GeneSplicer(Pertea等人,2001年),MaxEntScan(Yeo和Burge,2004年)和NNSplice(Reese等人,1997年),并绘制了在不同阈值下的RNA测序验证率和灵敏度(图38G)。正如其他人在这个领域的经验一样(Cummings等人,2017年),考虑到全基因组范围内可能会影响剪接的非编码变体数量非常多,我们发现现有分类器的特异性不足,大概是因为它们注重局部基序并在很大程度上不能解释长距离特异性决定子。

[0378] 考虑到与现有方法在性能上的巨大差距,我们进行了额外的控制,以排除RNA测序数据中结果可能因过度拟合而混淆的可能性。首先,我们分别针对私人变体和GTEx群组中

一个以上个体中存在的变体重复了验证和灵敏度分析(图46A, 46B和46C)。由于无论是剪接机制还是深度学习模型都无法获得等位基因频率信息,因此验证网络在等位基因频谱上具有相似的性能是一项重要的控制措施。我们发现在相同的 Δ 得分阈值下,私有和常见的隐秘剪接变体在RNA测序的验证率没有显著差异($P>0.05$, 费雪精确检验),表明该网络的预测对等位基因频率是可靠的。

[0379] 其次,为了验证模型在可创建新型剪接点的不同类型隐秘剪接变体中的预测能力,我们分别评估了产生新型GT或AG二核苷酸的变体,影响扩展受体或供体基序的那些变体以及发生在更远端区域的变体。我们发现隐秘剪接变体在三组之间大致均等地分布,并且在相同的 Δ 得分阈值下,各组之间的验证率或效果大小没有显著差异($P>0.3\chi^2$ 的一致性检验和 $P>0.3$ 的曼-惠特尼检验,图47A和47B)。

[0380] 第三,我们分别对用于训练的染色体上的变体和其余染色体上的变体进行了RNA测序验证和灵敏度分析(图48A和48B)。尽管仅根据参考基因组序列和剪接注释对网络进行了训练,并且在训练过程中没有接触变体数据,但我们想排除由于网络已经在训练染色体中看到了参考序列这一事实而导致变体预测存在偏差的可能性。我们发现该网络在来自训练和测试染色体的变体上表现同样出色,验证率或灵敏度没有显著差异($P>0.05$, 费雪精确检验),这表明该网络的变体预测不太可能通过过度拟合训练序列来解释。

[0381] 从我们的模型以及其他剪接预测算法的结果(比较图37E和图38G)可以看出,预测隐秘剪接变体比预测带注释的剪接点要困难得多。一个重要的原因是这两种类型的分析之间外显子增加速率的基本分布存在差异。绝大多数带有GENCODE注释的外显子具有很强的特异性决定子,导致组成性剪接以及预测得分接近1(图37C)。相反,大多数隐秘剪接变体仅部分外显(图38D和38E),预测得分较低或中等,并经常导致产生正常和异常转录本混合物的选择性剪接。这使得预测隐秘剪接变体影响这一问题在本质上比识别带注释的剪接位点更加困难。其他因素,如无意义介导的衰变,未解释的同种型变化以及RNA测序分析的局限性,进一步降低了RNA测序的验证率(图38C和图45)。

[0382] 组织特异的可变剪接通常是由弱隐秘剪接变体引起的可变剪接是基因调节的主要模式,可增加不同组织和发育阶段的转录本多样性,其失调与疾病进程有关(Blencowe, 2006年;Irimia等人,2014年;Keren等人,2010年;Licatalosi and Darnell, 2006年;Wang等人,2008年)。出乎意料的是,我们发现由隐秘剪接突变产生的新型剪接点的相对用法不同组织中可能存在很大差异(图39A)。此外,在多个个体之间可重现导致剪接中组织特异性差异的变体(图39B),表明组织特异性生物学可能是这些差异的基础,而不是随机效应。我们发现,35%具有较弱和中等预测得分(Δ 得分0.35-0.8)的隐秘剪接变体在跨组织产生的正常和异常转录本分数中表现出显著差异(对于 χ^2 测试,邦弗朗尼校正的 $P<0.01$,图39C)。这与具有较高预测得分(Δ 得分 >0.8)的变体形成对比,后者的产生组织特异性效果的可能性明显更低($P=0.015$)。我们的发现与之前的观察结果相吻合,即可变剪接外显子往往具有中等预测得分,相比之下,组成性地剪接的外显子分别具有接近1或0的得分(图37C)。

[0383] 这些结果支持了一个模型,在该模型中,组织特异性因素,例如染色质上下文和RNA结合蛋白的结合,可能会在两个亲和度接近的剪接点之间产生竞争(Gelfman等人,2013年;Luco等人,2010年;Shukla等人,2011年;Ule等人,2003年)。无论表观遗传背景如何,强大的隐秘剪接变体都可能将剪接从正常同种型完全转移到异常同种型,而较弱的变体使剪

接点选择更接近决策边界,从而导致在不同组织类型和细胞环境中使用替代性剪接点。这突显了隐秘剪接突变在产生新型可变剪接多样性中所发挥的出乎意料的作用,因为自然选择将有机会保留产生有用的组织特异性可变剪接的突变。

[0384] 预测的隐秘剪接变体在人类群体中具有很强的有害性

[0385] 虽然预测的隐秘剪接变体在RNA测序中具有很高的验证率,但在许多情况下,效果并不会完全外显,并且会产生正常和异常剪接同种型的混合物,提高一小部分隐秘剪接改变变体在功能上并不重要的可能性。为了探索自然选择在预测的隐秘剪接变体上的特征,我们对来自外显子组聚合联盟(ExAC)数据库(Lek等人,2016年)的60,706个人类外显子组中存在的每个变体都进行了评分,并确定了预计会改变外显子-内含子边界的变体。

[0386] 为了衡量负选择对预测的剪接改变变体的作用程度,我们计算了在常见等位基因频率(在人群中为 $\geq 0.1\%$)下发现的预测的剪接改变变体的数量,并将其与ExAC中处于单等位基频率(即60,706个个体中的1个)的预测的剪接改变变体数量进行了比较。在以单例等位基因改变变体。由于最近人口数量呈指数级增长,单例变体代表了最近创建的突变,这些突变已通过净化选择进行了最小限度的过滤(Tennessen等人,2012年)。相反,常见变体代表已通过净化选择筛选的中性突变的子集。因此,相对于单例变体,常见等位基因频谱中的预期剪接改变变体的损耗提供了对有害且因此是功能性的预测剪接改变变体的分数的估计。为避免混淆对蛋白质编码序列的影响,我们将分析限制在基本GT或AG二核苷酸之外的同义变体和内含子变体,不包括预计也具有剪接改变效果的错义突变。

[0387] 在常见的等位基因频率,自信预测的隐秘剪接变体(Δ 得分 ≥ 0.8)处于强烈的负选择下,如相对于预期的相对损耗就证明了这一点(图40A)。在此阈值下,预计大多数变体将接近在RNA测序数据中完全外显(图38D),预测的同义和内含子隐秘剪接突变在常见的等位基因频率减少了78%,这与移码、停止增益和基本GT或AG剪接破坏变体的82%损耗相当(图40B)。当考虑将导致移码的隐秘剪接变体而不是引起帧内改变的隐秘剪接变体时,负选择的影响更大(图40C)。具有移码结果的隐秘剪接变体的损耗几乎与其他类别的蛋白质截断变体相同,表明在近内含子区域(距离已知的外显子-内含子边界 $\leq 50\text{nt}$)的绝大多数自信预测的隐秘剪接突变是功能性的,对人类具有严重的有害影响。

[0388] 为了将此分析扩展到距已知外显子-内含子边界 $>50\text{nt}$ 的深内含子区域,我们使用了来自基因组聚合数据库(gnomAD)群组(Lek等人,2016年)的15,496人的汇总全基因组测序数据来计算在常见等位基因频率下观察到的和预期的隐秘剪接突变计数。总而言之,我们在距离外显子-内含子边界 $>50\text{nt}$ 处观察到了常见隐秘剪接突变(Δ 得分 ≥ 0.8)56%的损耗(图40D),这与预测深度内含子变体的影响时的较大难度一致,就像我们在RNA测序数据中观察到的那样。

[0389] 接下来,我们通过测量gnomAD群组每个个体中罕见隐秘剪接突变数,来估计相对于其他类型的蛋白质编码变异而言,隐秘剪接突变导致外显遗传疾病的可能性。根据负选择下预测的隐秘剪接突变的分数(图40A),一般人携带最多5个罕见的功能性隐秘剪接突变(等位基因频率 $< 0.1\%$),而携带最多11个罕见的蛋白截断突变(图40E)。隐秘剪接变异数量约为基本GT或AG剪接破坏变体的2倍。我们警告说,这些隐秘剪接变体的很大一部分可能不会完全丧失基因功能,这可能是因为它们产生帧内改变,或者是因为它们没有完全将剪接转变为异常同种型。

[0390] 新生隐秘剪接突变是罕见遗传病的主要原因

[0391] 对自闭症谱系障碍和严重智力障碍患者的大规模测序研究表明,新生蛋白质编码突变(错义、无义、移码和基本剪接二核苷酸)可破坏神经发育途径中的基因(Fitzgerald等人,2015年;Iossifov等人,2014年;McRae等人,2017年;Neale等人,2012年;De Rubeis等人,2014年;Sanders等人,2012年)。为了评估通过改变的剪接发挥作用的非编码突变的临床影响,我们应用了神经网络预测了来自解读发育障碍群组(DDD)的4,293名智力障碍人士(McRae等人,2017年)、来自Simons Simplex Collection(De Rubeis等人,2014年;Sanders等人,2012年;Turner等人,2016年)和自闭症测序联盟的3,953名患有自闭症谱系障碍(ASD)的人士以及来自Simons Simplex Collection的2,073名未受影响的同胞对照中新生突变的影响。为了控制各个研究中新生变体确定的差异,我们对新生变体的预期数目进行了归一化,这样每个个体的同义突变数量在群体间是相同的。

[0392] 与健康对照组相比(Δ 得分 ≥ 0.1 ,图41A、图43A和图43B),预计破坏剪接的新生突变在智力障碍($P=0.000416$)中富集1.51倍,在自闭症谱系障碍($P=0.0203$)中富集1.31倍。当仅考虑同义和内含子突变时(图49A、图49B和图49C),与对照组相比,剪接破坏突变也显著富集,排除了仅通过具有双重蛋白质编码和剪接效果的突变来解释富集的可能性。基于患病个体和未患病个体中新生突变的过量,调整每项每项研究中缺乏测序覆盖或变体确定的区域的突变预期比例,估计隐秘剪接突变在自闭症谱系障碍的致病突变中约占11%,在智力障碍中占9%(图41B)。在受影响的个体中,大多数预测的新生隐秘剪接突变的 Δ 得分 < 0.5 (图41C、图50A和图50B),并且根据GTEx RNA测序数据集中得分相似的变体,预计可以生成正常和异常转录本的混合物。

[0393] 为了估计候选疾病基因中与偶然性相比隐秘剪接突变的富集程度,我们使用三核苷酸上下文调整突变率来计算每个个体基因出现新生隐秘剪接突变的概率(Samocha等人,2014年)(表S4)。在新的基因发现中将隐秘剪接突变和蛋白质编码突变结合在一起,产生了5个与智力障碍相关的额外候选基因和2个与自闭症谱系障碍相关的额外基因(图41D和图45),当只考虑蛋白质编码突变时,这些基因会低于发现阈值($FDR<0.01$)(Kosmicki等人,2017年;Sanders等人,2015年)。

[0394] 自闭症患者新生隐秘剪接突变的实验验证

[0395] 我们从Simons Simplex Collection的36个个体中获得了外周血来源的淋巴母细胞样细胞系(LCL),这些个体在至少具有最小LCL表达水平的基因中包含了预测的新生隐秘剪接突变(De Rubeis等人,2014年;Sanders等人,2012年);每个个体都是直系亲属中唯一的自闭症病例。与大多数罕见的遗传疾病一样,相关的组织和细胞类型(可能是发育中的大脑)无法获得。因此,我们进行了高深度mRNA测序(每个样本最多3.5亿x 150bp的单次解读,大约是GTEx的10倍),从而补偿LCL中许多这些转录本的弱表达。为确保我们正在验证一组有代表性的预测隐秘剪接变体,而不是仅对最高预测进行验证,我们应用了相对宽松的阈值(剪接损耗变体的 Δ 得分 >0.1 ,剪接增益变体的 Δ 得分 >0.5 ;STAR方法)并对符合这些条件的所有新生变体进行了实验验证。

[0396] 在排除了8个在感兴趣基因上RNA测序覆盖率不足的个体后,我们在28位患者中的21位发现了与预测的新生隐秘剪接突变相关的独特异常剪接事件(图41E和图51A、51B、51C、51D、51E、51F、51G、51H、51I和51J)。获得深层LCL RNA测序的其他35个个体以及GTEx群

组中的149个个体没有这些异常剪接事件。在21个已确认的新生隐秘剪接突变中,我们观察到9例新的剪接点产生,8例外显子跳读和4例内含子保留,以及更复杂的剪接像差(图41F、图46A、图46B和图46C)。尽管有足够的转录本表达,但有7例未在LCL中显示异常剪接。尽管它们的子集可能代表假阳性预测,但在这些实验条件下,一些隐秘剪接突变可能会导致在LCL中无法观察到的组织特异性可变剪接。

[0397] 尽管RNA测序测定法存在局限性,但自闭症谱系障碍患者中预测的隐秘剪接突变的高验证率(75%)表明大多数预测是有效的。然而,与对照相比(DDD为1.5倍,ASD为1.3倍,图41A),新生隐秘剪接变体的富集仅为新生蛋白质截断变体观察到的效果大小的38%(DDD为2.5倍,ASD为1.7倍)(Iossifov等人,2014年;McRae等人,2017年;De Rubeis等人,2014年)。这使我们能够量化,功能性隐秘剪接突变的临床外显率约为经典形式的蛋白质截断突变(停止增益,移码和基本剪接二核苷酸)的50%,因为其中许多突变仅部分干扰了正常转录本的产生。事实上,孟德尔病中一些最良好表征的隐秘剪接突变,例如FECH中的c.315-48T>C(Gouya等人,2002年)和GAA中c.-32-13T>G(Boerkoel等人,1995年)是与较温和表型或较晚发病年龄有关的亚效等位基因。对于满足相对允许阈值(Δ 得分 ≥ 0.1)的所有新生变体,计算出临床外显率的估计值,并且预计具有较高预测得分的变体将具有相应较高的外显率。

[0398] 基于ASD和DDD群组中病例相对于对照的新生突变过量,与新生蛋白质截断变体可以解释的909例相比,可以通过新生隐秘剪接突变来解释250例(图41B)。一旦隐秘剪接突变的外显率降低被考虑在内,这与我们之前对罕见隐秘剪接突变(不超过5)与一般人群中罕见的蛋白质截断突变(不超过11)的平均数量的估计一致(图38A)。隐秘剪接突变在基因组中的广泛分布表明,神经发育障碍中由隐秘剪接突变的病例比例(9-11%,图41B)可能普遍适用于原生疾病机制是功能蛋白丧失的其他罕见遗传性疾病。为便于解释剪接改变突变,我们预先计算了全基因组所有可能的单核苷酸取代的 Δ 得分预测值,并将它们作为一种资源提供给科学界。我们相信,这种资源将促进人们对这种先前未被充分认识的遗传变异来源的理解。

[0399] 特定实施方式

[0400] 本文描述了应用训练的空洞卷积神经网络检测基因组序列(如核苷酸序列或氨基酸序列)中剪接位点的制造系统,方法和文章。可将实施方式的一个或多个特征与基本实施方式相结合。将互不排斥的实施方式教导为可结合。可将实施方式的一个或多个特征与其他实施方式相结合。本公开定期提醒用户这些选项。重复这些选项的某些实施方式叙述的遗漏不应视为对前面部分结合的限制,并通过引用将这些叙述向前合并到以下各个实施方式中。

[0401] 本部分交替使用术语:模块和阶段。

[0402] 本技术公开的系统实施方式包括耦合到存储器的一个或多个处理器。存储器加载有计算机指令以训练识别基因组序列(如,核苷酸序列)中剪接位点的剪接位点检测器。

[0403] 如图所示30,系统通过至少50000个供体剪接位点训练样品,至少50000个受体剪接位点训练样品和至少100000个非剪接位点训练样品对空洞卷积神经网络(简称ACNN)进行训练。每个训练样品为具有至少一个靶核苷酸的靶核苷酸序列,该靶核苷酸各侧翼具有至少20个核苷酸。

[0404] ACCN为使用允许具有少数可训练参数的大接收域的空洞/膨胀卷积。空洞/膨胀卷积的内核通过以一定的步骤跳跃输入值应用到大于其长度的区域,也称为空洞卷积率或膨胀因子。空洞/膨胀卷积增加卷积滤波器/内核元件的间距,使得进行卷积运算时,间隔更大的相邻输入条目(如核苷酸、氨基酸)得以考虑。这就使得远程语境依赖并入到输入中。空洞卷积保留部分卷积计算,以便在处理相邻核苷酸时重复使用。

[0405] 如图30所示,系统提供两侧具有至少40个上游上下文核苷酸和至少40个下游上下文核苷酸的靶核苷酸序列作为ACNN输入,用以使用ACNN评估训练样品。

[0406] 如图30所示,基于该评估,ACNN产生三重得分作为输出,使得靶核苷酸中各个核苷酸为供体剪接位点,受体剪接位点或非剪接位点成为可能。

[0407] 公开的系统实施方式和其他系统任选地包括以下一个或多个特征。系统也可包括与公开的方法相关的特征。简洁起见,系统特征的选择性组合不再一一列举。对于基本特征的各个法定类集,适用于制造系统,方法和文章的特征不再赘述。读者应当了解本部分识别的特征是如何容易地与其他法定中的基本特征相结合的。

[0408] 如图25、26和27所示,输入可包括具有靶核苷酸的靶核苷酸序列,该靶核苷酸两翼均具有2500个核苷酸。在这样一个实施方式中,该靶核苷酸序列两翼具有5000个上游上下文核苷酸和5000下游上下文核苷酸。

[0409] 输入可包括具有靶核苷酸的靶核苷酸序列,该靶核苷酸两翼均具有100核苷酸。在这样一个实施方式中,该靶核苷酸序列两翼具有200个上游上下文核苷酸和200下游上下文核苷酸。

[0410] 输入可包括具有靶核苷酸的靶核苷酸序列,所述靶核苷酸两侧翼均具有500核苷酸。在这样一个实施方式中,所述靶核苷酸序列两侧翼具有1000个上游上下文核苷酸和1000下游上下文核苷酸。

[0411] 如图28所示,系统通过150000个供体剪接位点训练样品,至少150000个受体剪接位点训练样品和至少800000000个非剪接位点训练样品对ACNN进行训练。

[0412] 如图19所示,ACNN可包括多组从最低到最高以序列布置的残差块。通过残差块中多个卷积滤波器,残差块卷积窗口尺寸和残差块空洞卷积率对每组残差块进行参数化。

[0413] 如图21,22,23和24所示,在ACNN中,空洞卷积率由低残差块组到高残差块组进行非指数型演变。

[0414] 如图21,22,23和24所示,在ACNN中,各组残差块之间卷积窗口尺寸不同。

[0415] ACNN可配置为评估包括靶核苷酸序列的输入,所述靶核苷酸两侧翼具有40个上游上下文核苷酸和40个下游上下文核苷酸。在这样一个实施方式中,ACNN包括一个四残差块组和至少一个跳跃连接。每个残差块具有32个卷积滤波器,卷积窗口尺寸11和空洞卷积率1。如图21所示,本文中ACNN的实施方式是指SpliceNet80。

[0416] ACNN可配置为评估包括靶核苷酸序列的输入,所述靶核苷酸两侧翼具有200个上游上下文核苷酸和200个下游上下文核苷酸。在这样一个实施方式中,ACNN包括至少两个四残差块组和至少两个跳跃连接。第一组中每个残差块具有32个卷积滤波器,卷积窗口尺寸11和空洞卷积率1。第二组中每个残差块具有32个卷积滤波器,卷积窗口尺寸11和空洞卷积率4。如图22所示,本文中ACNN的实施方式是指SpliceNet400。

[0417] ACNN可配置为评估包括靶核苷酸序列的输入,所述靶核苷酸两侧翼具有1000个上

游上下文核苷酸和1000个下游上下文核苷酸。在这样一个实施方式中,ACNN包括至少三个四残差块组和至少三个跳跃连接。第一组中每个残差块具有32个卷积滤波器,卷积窗口尺寸11和空洞卷积率1。第二组中每个残差块具有32个卷积滤波器,卷积窗口尺寸11和空洞卷积率4。第三组中每个残差块具有32个卷积滤波器,卷积窗口尺寸21和空洞卷积率19。如图23所示,本文中ACNN的实施方式是指SpliceNet2000。

[0418] ACNN可配置为评估包括靶核苷酸序列的输入,所述靶核苷酸两侧翼具有5000个上游上下文核苷酸和5000个下游上下文核苷酸。在这样一个实施方式中,ACNN包括至少四组残差块(每组四个)和至少四个跳跃连接。第一组中每个残差块具有32个卷积滤波器,卷积窗口尺寸11和空洞卷积率1。第二组中每个残差块具有32个卷积滤波器,卷积窗口尺寸11和空洞卷积率4。第三组中每个残差块具有32个卷积滤波器,卷积窗口尺寸21和空洞卷积率19。第四组中每个残差块具有32个卷积滤波器,卷积窗口尺寸41和空洞卷积率25。如图24所示,本文中ACNN的实施方式是指SpliceNet10000。

[0419] 靶核苷酸中每个核苷酸的三重得分可指数式地进行归一化以求和为一。在这样一个实施方式中,基于各自三重得分的最高得分,系统将靶核苷酸中每个核苷酸分类为供体剪接位点,受体剪接位点或非剪接位点。

[0420] 如图35所示,ACNN输入的维度可定义为 $(C^u+L+C^d) \times 4$,其中C为上游上下文核苷酸的数量, C^d 为下游上下文核苷酸的数量,L为靶核苷酸序列中核苷酸的数量。在一个实施方式中,输入维度为 $(5000+5000+5000) \times 4$ 。

[0421] 如图35所示,ACNN输入的维度可定义为 $L \times 3$ 。在一个实施方式中,输入维度为 5000×3 。

[0422] 如图35所示,每组残差块可通过处理前述输入产生一中间输出。中间输出的维度可定义为 $(I - [(W-1) \times D] \times A) \times N$,其中I为前述输入的维度,W为残差块的卷积窗口尺寸,D为残差块的空洞卷积率,A为各组空洞卷积层数,N为残差块中卷积滤波器数。

[0423] 如图32所示,CANN在一个时间段内批量评估所述训练样本。训练样本随机批量取样。每一批次具有预定的批次大小。ACNN在多个时间段(如,1-10个时间段)内对所述训练样品进行评估迭代。

[0424] 输入可包括具有两个相邻靶核苷酸的靶核苷酸序列。两个相邻靶核苷酸可为腺嘌呤(简称A)和鸟嘌呤(简称G)。两个相邻靶核苷酸可为鸟嘌呤(简称G)和尿嘧啶(简称U)。

[0425] 系统包括单热编码器(如图29所示),单热编码器对训练样本进行稀疏编码并提供单热编码作为输入。

[0426] 通过多个残差块、多个跳跃连接和多个残差连接对ACNN进行参数化。

[0427] ACNN可包括改造前述输入的空间和特征维度的维度变换卷积层。

[0428] 如图20所示,每个残差块可包括至少一个批归一化层、至少一个修正线性层(简称ReLU)、至少一个空洞卷积层和至少一个残差连接。在这样一个实施方式中,每个残差块包括两个批归一化层、两个ReLU非线性层、两个空洞卷积层和一个残差连接。

[0429] 其他实施方式可包括存储指令的非暂存性计算机可读存储介质,处理器执行所述指令以完成上述系统动作。其他实施方式也可包括执行上述系统动作的方法。

[0430] 本公开技术的其他系统实施方式包括在大量并行处理器上运行的训练的剪接位点预测器,所述处理器耦合到内存。系统通过至少5000个供体剪接位点训练样本,至少5000

个受体剪接位点训练样本及至少100000个非剪接位点训练样本对运行在大量处理器上的空洞卷积神经网络(简称ACNN)进行训练。训练所用的每个训练样本为核苷酸序列,所述核苷酸序列包括两侧翼均具有至少400个核苷酸的靶核苷酸。

[0431] 系统包括ACNN的输入级,ACNN在至少一个所述大量处理器上运行并提供至少801个核苷酸的输入序列以评估靶核苷酸。每个靶核苷酸两侧翼均具有至少400个核苷酸。在其他实施方式中,系统包括ACNN的输入模块,ACNN在至少一个所述大量处理器上运行并提供至少801个核苷酸的输入序列以评估靶核苷酸。

[0432] 系统包括ACNN的输入级,ACNN在至少一个所述大量处理器上运行并将ACNN的分析转化为分类得分,使得每个靶核苷酸为供体剪接位点,受体剪接位点或非剪接位点成为可能。在其他实施方式中,系统包括ACNN的输入级,ACNN在至少一个所述大量处理器上运行并将ACNN的分析转化为分类得分,使得每个靶核苷酸为供体剪接位点,受体剪接位点或非剪接位点成为可能。

[0433] 用于第一系统实施方式的这一特定实施方式部分讨论的每个特征同样适用于这一系统实施方式。如上所述,本文不再赘述所有系统特征,通过引用,所有系统特征视为重复的。

[0434] ACNN可通过150000个供体剪接位点训练样品,150000受体剪接位点训练样品和800000000个非剪接位点训练样品进行训练。在系统另一实施方式中,ACNN包括多组从最低到最高以序列布置的残差块。在系统另一实施方式中,通过残差块中多个卷积滤波器,残差块卷积窗口尺寸和残差块空洞卷积率对每组残差块进行参数化。

[0435] ACNN可包括多组从最低到最高以序列布置的残差块。通过残差块中多个卷积滤波器,残差块卷积窗口尺寸和残差块空洞卷积率对每组残差块进行参数化。

[0436] 在ACNN中,空洞卷积率由低残差块组到高残差块组进行非指数型演变。同样,在ACNN中,各组残差块之间卷积窗口尺寸不同。

[0437] 如图18所示,ACNN可通过一个或多个训练服务器进行训练。

[0438] 如图18所示,训练的ACNN可部署到一个或多个从请求客户接收输入序列的生产服务器。如图18所示,在这样一个实施方式中,生产服务器通过ACNN的输入级和输出级处理输入序列以产生输出,输出发送到客户端。如图18所示,在其他实施方式中,生产服务器通过ACNN的输入级和输出级处理输入序列以产生输出,输出发送到客户端。

[0439] 其他实施方式可包括存储指令的非暂存性计算机可读存储介质,处理器执行所述指令以完成上述系统动作。其他实施方式也可包括完成上述系统动作的方法。

[0440] 本公开技术的其他方法实施方式包括对基因组序列(如核苷酸序列)中识别剪接位点的剪接位点预测器进行训练。

[0441] 方法包括提供空洞卷积神经网络(简称ACNN)和具有至少801个核苷酸的输入序列以评估靶核苷酸,靶核苷酸两侧翼均具有至少400个核苷酸。

[0442] ACNN可通过5000个供体剪接位点训练样品,5000个受体剪接位点训练样品和100000个非剪接位点训练样品进行训练。训练所用的每个训练样本为核苷酸序列,所述核苷酸序列包括两侧翼均具有至少400个核苷酸的靶核苷酸。

[0443] 方法还包括将ACNN的分析转化为分类得分,使得每个靶核苷酸为供体剪接位点,受体剪接位点或非剪接位点成为可能。

[0444] 用于第一系统实施方式的这一特定实施方式部分讨论的每个特征同样适用于这一系统实施方式。如上所述,本文不再赘述所有系统特征,通过引用,所有系统特征视为重复的。

[0445] 其他实施方式可包括存储指令的非暂存性计算机可读存储介质,处理器执行所述指令以完成上述系统动作。其他实施方式还可包括用以执行所述方法的系统,系统包括存储器和一个或多个存储在存储器中的指令可操作的处理器。

[0446] 本文描述了应用训练的空洞卷积神经网络检测基因组序列(如核苷酸序列)中异常剪接的制造系统,方法和文章。实施方式的一个或多个特征可与基本实施方式相结合。互不排斥的实施方式教导为可结合。实施方式的一个或多个特征可与其他实施方式相结合。本公开定期提醒用户这些选项。重复这些选项的某些实施方式叙述的遗漏不应视为对前面部分结合的限制,通过引用将这些叙述向前合并到一下各个实施方式中。

[0447] 本技术公开的系统实施方式包括耦合到存储器的一个或多个处理器。存储器加载有计算机指令以实现在大量处理器上运行的异常剪接检测器,处理器并行运行并耦合到存储器。

[0448] 如图34所示,系统包括在大量处理器上运行的训练的空洞卷积神经网络(简称ACNN)。ACNN为使用允许具有少数可训练参数的大接收域的空洞/膨胀卷积。空洞/膨胀卷积的内核通过以一定的步长跳过输入值应用到大于其长度的区域,也称为空洞卷积率或膨胀因子。空洞/膨胀卷积增加卷积滤波器/内核元件的间距,使得进行卷积运算时,间隔更大的相邻输入条目(如核苷酸,氨基酸)得以考虑。这就使得远程语境依赖并入到输入中。空洞卷积保留部分卷积计算,以便在处理相邻核苷酸时重复使用。

[0449] 如图34所示,ACNN将输入序列中的靶核苷酸进行分类并将剪接位点得分进行分配,使得靶核苷酸中每个核苷酸为供体剪接位点,受体剪接位点或非剪接位点成为可能。输入序列包括至少801个核苷酸,每个核苷酸两侧翼均具有至少400个核苷酸。

[0450] 如图34所示,系统还包括在至少一个大量处理器上运行的分类器,分类器通过ACNN对参考序列和变体序列进行处理,使得参考序列和变体序列中每个核苷酸为供体剪接位点,受体剪接位点或非剪接位点成为可能。每个参考序列和变体序列具有至少101个靶核苷酸,每个靶核苷酸两侧翼均具有至少400个核苷酸。图33描述了参考序列和可变、变体序列。

[0451] 如图34所示,根据参考序列和变体序列中靶核苷酸的剪接位点得分的差异,系统还确定产生变体序列的变体是否导致异常剪接及是否具有致病性。

[0452] 本公开系统实施方式和其他系统任选地包括以下一个或多个特征。系统还可包括与本公开方法所描述相关的特征。简洁起见,系统特征的选择性组合不再单独列举。对于基本特征的各个法定类集,适用于制造系统,方法和文章的特征不重复。读者应当了解本部分识别的特征是如何容易地与其他法定中的基本特征相结合的。

[0453] 如图34所示,剪接位点得分的差异可通过参考序列和变体序列中靶核苷酸的位置来确定。

[0454] 如图34所示,对于至少一个靶核苷酸位置而言,当剪接位点得分的全局最大差值大于预设阈值时,ACNN将引起异常剪接和具有致病性的变体进行分类。

[0455] 如图17所示,对于至少一个靶核苷酸位置而言,当剪接位点得分的全局最大差值

小于预设阈值时,ACNN将未引起异常剪接的良性变体进行分类。

[0456] 阈值可由多个候选阈值确定。包括处理第一组良性常见变体产生的参考序列和变体序列以生成第一组异常剪接检测,处理第二组致病性罕见变体产生的参考序列和变体序列以生成第二组异常剪接检测以及选取至少一个用于分类器的阈值,所述阈值将第二组的异常剪接检测数量最大化,将第一组的异常剪接检测数量最小化。

[0457] 在一个实施方式中,ACNN识别引起自闭症谱系障碍(简称ASD)的变体。在另一实施方式中,ACNN识别引起发育迟缓障碍(简称DDD)的变体。

[0458] 如图36所示,每个参考序列和变体序列具有至少101个靶核苷酸,每个靶核苷酸两侧翼均具有至少500个核苷酸。

[0459] 如图36所示,参考序列中靶核苷酸的剪接位点得分可在ACNN的第一输出中进行编码,变体序列中靶核苷酸的剪接位点得分可在ACNN的第二输出中进行编码。在另一实施方式中,第一输出编码为第一101x3矩阵,第二输出编码为第二101x3矩阵。

[0460] 如图36所示,在这样一个实施方式中,第一101x3矩阵的每行仅表示剪接位点得分,使得参考序列中靶核苷酸为供体剪接位点,受体剪接位点或非剪接位点成为可能。

[0461] 同样,如图36所示,在这样一个实施方式中,第二101x3矩阵的每行仅表示剪接位点得分,使得变体序列中靶核苷酸为供体剪接位点,受体剪接位点或非剪接位点成为可能。

[0462] 如图36所示,在一些实施方式中,第一101x3矩阵和第二101x3矩阵的每行剪接位点得分可指数式地进行归一化以求和为一。

[0463] 如图36所示,分类器可将第一101x3矩阵和第二101x3矩阵进行行行比较并在所述行的基础上确定剪接位点得分在分配上的变化。对于至少一个行行比较实例而言,当分配变化大于预定阈值时,ACNN将引起异常剪接且具有致病性的变体进行分类。

[0464] 系统包括单热编码器(如图29所示),单热编码器对参考序列和变体序列进行稀疏编码。

[0465] 用于其他系统和方法的实施方式的这一特定实施方式部分讨论的每个特征同样适用于这一系统实施方式。如上所述,本文不再赘述所有系统特征,通过引用,所有系统特征视为重复的。

[0466] 其他实施方式可包括存储指令的非暂存性计算机可读存储介质,处理器执行所述指令以完成上述系统动作。其他实施方式也可包括执行上述系统动作的方法。

[0467] 本公开技术的方法实施方式包括检测引起异常剪接的基因组变体。

[0468] 方法包括通过将靶子序列中每个核苷酸分类为供体剪接位点,受体剪接位点和非剪接位点,通过训练的空洞卷积神经网络(简称ACNN)对参考序列进行处理,以检测输入序列的靶子序列中差异剪接方式。

[0469] 基于处理,方法包括通过将靶子序列中每个核苷酸分类为供体剪接位点,受体剪接位点和非剪接位点对参考靶子序列中的差异剪接方式进行检测。

[0470] 方法包括通过ACNN对变体序列进行处理。变体序列和参考序列之间至少相差一个变体核苷酸,所述变体核苷酸位于变体靶子序列中。

[0471] 基于处理,方法包括通过将变体靶子序列中每个核苷酸分类为供体剪接位点,受体剪接位点和非剪接位点对变体靶子序列中的第二差异剪接方式进行检测。

[0472] 基于核苷酸到核苷酸,方法包括通过比较参考靶子序列和变体靶子序列的剪接位

点分类来确定第一异常剪接方式和第二异常剪接方式的差异。

[0473] 当差值大于预定阈值时,方法包括将引起异常剪接并具有致病性的变体进行分类并存储在存储器中。

[0474] 用于其他系统和方法的实施方式的这一特定实施方式部分讨论的每个特征同样适用于这一系统实施方式。如上所述,本文不再赘述所有系统特征,通过引用,所有系统特征视为重复的。

[0475] 差异剪接方式可识别靶子序列中剪接事件的位置分布。剪接事件实例包括以下至少一种:隐秘剪接、外显子跳跃、互斥外显子、可变给位、可变受位和内含子保留。

[0476] 参考靶子序列和变体靶子序列可与核苷酸位置对齐且相差至少一个变体核苷酸。

[0477] 每个参考靶子序列和变体靶子序列具有至少40个靶核苷酸,每个每个参考靶子序列和变体靶子序列两侧翼均具有至少40个核苷酸。

[0478] 每个参考靶子序列和变体靶子序列具有至少101个靶核苷酸,每个每个参考靶子序列和变体靶子序列两侧翼均具有至少5000个核苷酸。

[0479] 变体靶子序列可包括两个变体。

[0480] 其他实施方式可包括存储指令的非暂存性计算机可读存储介质,处理器执行所述指令以完成上述系统动作。其他实施方式还可包括用以执行所述方法的系统,系统包括存储器和一个或多个存储在存储器中的指令可操作的处理器。

[0481] 本文描述了应用训练的卷积神经网络检测基因组序列(如核苷酸序列)中剪接位点和异常剪接的制造系统,方法和文章。实施方式的一个或多个特征可与基本实施方式相结合。互不排斥的实施方式教导为可结合。实施方式的一个或多个特征可与其他实施方式相结合。本公开定期提醒用户这些选项。重复这些选项的某些实施方式叙述的遗漏不应视为对前面部分结合的限制,通过引用将这些叙述向前合并到一下各个实施方式中。

[0482] 本技术公开的系统实施方式包括耦合到存储器的一个或多个处理器。存储器加载有计算机指令以训练识别基因组序列(如核苷酸序列)中剪接位点的剪接位点检测器。

[0483] 系统通过至少50000个供体剪接位点训练样品,至少50000个受体剪接位点训练样品和至少100000个非剪接位点训练样品对卷积神经网络(简称CNN)进行训练。每个训练样品为具有至少一个靶核苷酸的靶核苷酸序列,该靶核苷酸各侧翼具有至少20个核苷酸。

[0484] 系统提供两侧具有至少40个上游上下文核苷酸和至少40个下游上下文核苷酸的靶核苷酸序列作为CNN输入,用以使用CNN评估训练样品。

[0485] 基于该项评估,CNN随后产生三重得分作为输出,使得靶核苷酸中各个核苷酸为供体剪接位点,受体剪接位点或非剪接位点成为可能。

[0486] 公开的系统实施方式和其他系统任选地包括以下一个或多个特征。系统也可包括与公开的方法相关的特征。简洁起见,系统特征的选择性组合不再一一列举。对于基本特征的各个法定类集,适用于制造系统,方法和文章的特征不重复。读者应当了解本部分识别的特征是如何容易地与其他法定中的基本特征相结合的。

[0487] 输入可包括具有靶核苷酸的靶核苷酸序列,该靶核苷酸两翼均具有100个核苷酸。在这样一个实施方式中,靶核苷酸序列两翼具有200个上游上下文核苷酸和200下游上下文核苷酸。

[0488] 如图28所示,系统通过至少150000个供体剪接位点训练样品,至少150000个受体

剪接位点训练样品和至少100000个非剪接位点训练样品对CNN进行训练。

[0489] 如图31所示,通过多个卷积层,多个卷积滤波器及多个子抽样层(如最大池化和平均池化)对CNN进行参数化。

[0490] 如图31所示,CNN可包括一个或多个全连接层和一个终端分类层。

[0491] CNN可包括改造前述输入的空间和特征维度的维度变换卷积层。

[0492] 靶核苷酸中每个核苷酸的三重得分可指数式地进行归一化以求和为一。在这样一个实施方式中,基于各自三重得分的最高得分,系统将靶核苷酸中每个核苷酸分类为供体剪接位点,受体剪接位点或非剪接位点。

[0493] 如图32所示,CNN在一段时间内批量评估所述训练样本。训练样本随机批量取样。每一批次具有预定的批次大小。CNN在多个时间段(如,1-10个时间段)内对所述训练样品进行评估迭代。

[0494] 输入可包括具有两个相邻靶核苷酸的靶核苷酸序列。两个相邻靶核苷酸可为腺嘌呤(简称A)和鸟嘌呤(简称G)。两个相邻靶核苷酸可为鸟嘌呤(简称G)和尿嘧啶(简称U)。

[0495] 系统包括单热编码器(如图29所示),单热编码器对训练样本进行稀疏编码并提供单热编码作为输入。

[0496] 通过多个残差块、多个跳跃连接和多个残差连接对CNN进行参数化。

[0497] 每个残差块可包括至少一个批归一化层、至少一个修正线性单元(简称ReLU)层、至少一个维度变化层和至少一个残差连接。每个残差块可包括两个批归一化层、两个ReLU非线性层、两个维度变化层和一个残差连接。

[0498] 用于其他系统和方法的实施方式的这一特定实施方式部分讨论的每个特征同样适用于这一系统实施方式。如上所述,本文不再赘述所有系统特征,通过引用,所有系统特征视为重复的。

[0499] 其他实施方式可包括存储指令的非暂存性计算机可读存储介质,处理器执行所述指令以完成上述系统动作。其他实施方式也可包括执行上述系统动作的方法。

[0500] 本公开技术的其他系统实施方式包括在大量并行处理器上运行的训练的剪接位点预测器,所述处理器耦合到内存。系统通过至少5000个供体剪接位点训练样本,至少5000个受体剪接位点训练样本及至少100000个非剪接位点训练样本对运行在大量处理器上的卷积神经网络(简称ACNN)进行训练。训练所用的每个训练样本为核苷酸序列,所述核苷酸序列包括两侧翼均具有至少400个核苷酸的靶核苷酸。

[0501] 系统包括ACNN的输入级,CNN在至少一个所述大量处理器上运行并提供至少801个核苷酸的输入序列以评估靶核苷酸。每个靶核苷酸两侧翼均具有至少400个核苷酸。在其他实施方式中,系统包括CNN的输入模块,ACNN在至少一个所述大量处理器上运行并提供至少801个核苷酸的输入序列以评估靶核苷酸。

[0502] 系统包括CNN的输入级,ACNN在至少一个所述大量处理器上运行并将CNN的分析转化为分类得分,使得每个靶核苷酸为供体剪接位点,受体剪接位点或非剪接位点成为可能。在其他实施方式中,系统包括CNN的输入级,ACNN在至少一个所述大量处理器上运行并将CNN的分析转化为分类得分,使得每个靶核苷酸为供体剪接位点,受体剪接位点或非剪接位点成为可能。

[0503] 用于其他系统和方法的实施方式的这一特定实施方式部分讨论的每个特征同样

适用于这一系统实施方式。如上所述,本文不再赘述所有系统特征,通过引用,所有系统特征视为重复的。

[0504] 通过150000个供体剪接位点训练样品,150000个受体剪接位点训练样品和800000000个非剪接位点训练样品对CNN进行训练。

[0505] CNN可通过一个或多个训练服务器进行训练。

[0506] 训练的CNN可部署到一个或多个从请求客户接收输入序列的生产服务器。在这样一个实施方式中,生产服务器通过CNN的输入级和输出级处理输入序列以产生输出,输出发送到客户端。在其他实施方式中,生产服务器通过ACNN的输入级和输出级处理输入序列以产生输出,输出发送到客户端。

[0507] 其他实施方式可包括存储指令的非暂存性计算机可读存储介质,处理器执行所述指令以完成上述系统动作。其他实施方式也可包括执行上述系统动作的方法。

[0508] 本公开技术的方法实施方式包括对基因组序列(如核苷酸序列)中识别剪接位点的剪接位点预测器进行训练。方法包括提供卷积神经网络(简称CNN)和具有至少801个核苷酸的输入序列以评估靶核苷酸,靶核苷酸两侧翼均具有至少400个核苷酸。

[0509] 通过至少50000个供体剪接位点训练样品,至少50000个受体剪接位点训练样品和至少100000个非剪接位点训练样品对CNN进行训练。训练所用的每个训练样本为核苷酸序列,所述核苷酸序列包括两侧翼均具有至少400个核苷酸的靶核苷酸。

[0510] 方法还包括将CNN的分析转化为分类得分,使得每个靶核苷酸为供体剪接位点,受体剪接位点或非剪接位点成为可能。

[0511] 用于其他系统和方法的实施方式的这一特定实施方式部分讨论的每个特征同样适用于这一系统实施方式。如上所述,本文不再赘述所有系统特征,通过引用,所有系统特征视为重复的。

[0512] 其他实施方式可包括存储指令的非暂存性计算机可读存储介质,处理器执行所述指令以完成上述系统动作。其他实施方式还可包括用以执行所述方法的系统,系统包括存储器和一个或多个存储在存储器中的指令可操作的处理器。

[0513] 本技术公开的另一系统实施方式还包括耦合到存储器的一个或多个处理器。存储器加载有计算机指令以实现在大量处理器上运行的异常剪接检测器,处理器并行运行并耦合到存储器。

[0514] 系统包括在大量处理器上运行的训练的卷积神经网络(简称CNN)。

[0515] 如图34所示,CNN将输入序列中的靶核苷酸进行分类并将剪接位点得分进行分配,使得靶核苷酸中每个核苷酸为供体剪接位点,受体剪接位点或非剪接位点成为可能。输入序列包括至少801个核苷酸,每个核苷酸两侧翼均具有至少400个核苷酸。

[0516] 如图34所示,系统还包括在至少一个大量处理器上运行的分类器,分类器通过CNN对参考序列和变体序列进行处理,使得参考序列和变体序列中每个核苷酸为供体剪接位点,受体剪接位点或非剪接位点成为可能。每个参考序列和变体序列具有至少101个靶核苷酸,每个靶核苷酸两侧翼均具有至少400个核苷酸。

[0517] 如图34所示,根据参考序列和变体序列中靶核苷酸的剪接位点得分的差异,系统还确定产生变体序列的变体是否导致异常剪接及是否具有致病性。

[0518] 用于其他系统和方法的实施方式的这一特定实施方式部分讨论的每个特征同样

适用于这一系统实施方式。如上所述,本文不再重复所有系统特征,通过引用,所有系统特征视为重复的。

[0519] 剪接位点得分的差异可通过参考序列和变体序列中靶核苷酸的位置来确定。

[0520] 对于至少一个靶核苷酸位置而言,当剪接位点得分的全局最大差值大于预设阈值时,CNN将引起异常剪接和具有致病性的变体进行分类。

[0521] 对于至少一个靶核苷酸位置而言,当剪接位点得分的全局最大差值小于预设阈值时,ACNN将未引起异常剪接的良性变体进行分类。

[0522] 阈值可由多个候选阈值确定。包括处理第一组良性常见变体产生的参考序列和变体序列以生成第一组异常剪接检测,处理第二组致病性罕见变体产生的参考序列和变体序列以生成第二组异常剪接检测以及选取至少一个用于分类器的阈值,所述阈值将第二组的异常剪接检测数量最大化,将第一组的异常剪接检测数量最小化。

[0523] 在一个实施方式中,ACNN识别引起自闭症谱系障碍(简称ASD)的变体。在另一实施方式中,ACNN识别引起发育迟缓障碍(简称DDD)的变体。

[0524] 每个参考序列和变体序列具有至少101个靶核苷酸,每个靶核苷酸两侧翼均具有至少1000个核苷酸。

[0525] 参考序列中靶核苷酸的剪接位点得分可在CNN的第一输出中进行编码,变体序列中靶核苷酸的剪接位点得分可在CNN的第二输出中进行编码。在一实施方式中,第一输出编码为第一101x3矩阵,第二输出编码为第二101x3矩阵。

[0526] 在这样一个实施方式中,第一101x3矩阵的每行仅表示剪接位点得分,使得参考序列中靶核苷酸为供体剪接位点,受体剪接位点或非剪接位点成为可能。

[0527] 同样,在这样一个实施方式中,第二101x3矩阵的每行仅表示剪接位点得分,使得变体序列中靶核苷酸为供体剪接位点,受体剪接位点或非剪接位点成为可能。

[0528] 在一些实施方式中,第一101x3矩阵和第二101x3矩阵的每行剪接位点得分可指数式地进行归一化以求和为一。

[0529] 分类器可将第一101x3矩阵和第二101x3矩阵进行行行比较并在所述行的基础上确定剪接位点得分在分配上的变化。对于至少一个行行比较实例而言,当分配变化大于预定阈值时,CNN将引起异常剪接且具有致病性的变体进行分类。

[0530] 系统包括单热编码器(如图29所示),单热编码器对参考序列和变体序列进行稀疏编码。

[0531] 其他实施方式可包括存储指令的非暂存性计算机可读存储介质,处理器执行所述指令以完成上述系统动作。其他实施方式也可包括执行上述系统动作的方法。

[0532] 本公开技术的方法实施方式包括检测引起异常剪接的基因组变体。

[0533] 方法包括通过将靶子序列中每个核苷酸分类为供体剪接位点,受体剪接位点和非剪接位点,通过训练的空洞卷积神经网络(简称CNN)对参考序列进行处理,以检测输入序列的靶子序列中差异剪接方式

[0534] 基于处理,方法包括通过将靶子序列中每个核苷酸分类为供体剪接位点,受体剪接位点和非剪接位点对参考靶子序列中的差异剪接方式进行检测。

[0535] 方法包括通过CNN对变体序列进行处理。变体序列和参考序列之间至少相差一个变体核苷酸,所述变体核苷酸位于变体靶子序列中。

[0536] 基于处理,方法包括通过将变体靶子序列中每个核苷酸分类为供体剪接位点,受体剪接位点和非剪接位点对变体靶子序列中的第二差异剪接方式进行检测。

[0537] 基于核苷酸到核苷酸,方法包括通过比较参考靶子序列和变体靶子序列的剪接位点分类来确定第一异常剪接方式和第二异常剪接方式的差异。

[0538] 当差值大于预定阈值时,方法包括将引起异常剪接并具有致病性的变体进行分类并存储在存储器中。

[0539] 用于其他系统和方法的实施方式的这一特定实施方式部分讨论的每个特征同样适用于这一系统实施方式。如上所述,本文不再赘述所有系统特征,通过引用,所有系统特征视为重复的。

[0540] 差异剪接方式可识别靶子序列中剪接事件的位置分布。剪接事件实例包括以下至少一种:隐秘剪接、外显子跳跃、互斥外显子、可变给位、可变受位和内含子保留。

[0541] 参考靶子序列和变体靶子序列可与核苷酸位置对齐且相差至少一个变体核苷酸。

[0542] 每个参考靶子序列和变体靶子序列具有至少40个靶核苷酸,每个参考靶子序列和变体靶子序列两侧翼均具有至少40个核苷酸。

[0543] 每个参考靶子序列和变体靶子序列具有至少101个靶核苷酸,每个参考靶子序列和变体靶子序列两侧翼均具有至少1000个核苷酸。

[0544] 变体靶子序列可包括两个变体。

[0545] 其他实施方式可包括存储指令的非暂存性计算机可读存储介质,处理器执行所述指令以完成上述系统动作。其他实施方式还可包括用以执行所述方法的系统,系统包括存储器和一个或多个存储在存储器中的指令可操作的处理器。

[0546] 上述说明使得本公开技术得以制造和使用。本公开实施方式的各种修改是显而易见的,在不脱离所公开技术的精神和范围的情况下,本文中限定的基本原理可被应用到其它实施方式和应用中。因此,本公开技术并非局限于所述实施方式,而是符合与本文公开的原理和特征相一致的最宽范围。本公开技术的范围由附加的权利要求定义。

[0547] 每基因富集分析

[0548] 图57为每基因富集分析的一个实施方式。在一个实施方式中,异常剪接检测器进一步配置为执行每基因富集分析,每基因富集分析用于确定已确定会引起异常剪接的变体的致病性。对于从具有基因缺陷的一群个体中取样的特定基因而言,每基因富集分析包括应用训练的ACNN识别引起异常剪接的特定基因中的候选变体,基于候选变体的观测三核苷酸突变率的和及所述和与传输数和群组大小的乘积确定特定基因突变的基线数,应用训练的ACNN识别引起异常剪接的特定基因中的新生变体以及将突变的基线数与新生变体数进行比较。基于比较结果,每基因富集分析确定特定基因与基因缺陷有关,新生变体具有致病性。在某些实施方式中,基因缺陷为自闭症谱系障碍(简称ASD)。在其他实施方式中,基因缺陷为发育迟缓障碍(简称DDD)。

[0549] 如图57实施例所示,异常剪接检测器将引起异常剪接的特定基因中的5个候选变体进行分类。这5个候选变体均观察到三核苷酸突变率,分别为 10^{-8} 、 10^{-2} 、 10^{-1} 、 10^5 和 10^1 。基于观察到的5个候选变体的三核苷酸突变率的和及所述和与传输数/染色体数(2)及群组(1000)大小的乘积,特定基因的突变基线数确定为 10^{-5} 。再与新生变体数(3)进行比较。

[0550] 在某些实施方式中,异常剪接检测器进一步配置为运用生成p-值作为输出的统计

测试进行比较。

[0551] 在其他实施方式中,异常剪接检测器进一步配置为将突变的基线数与新生变体数进行比较,以及基于比较结果确定特定基因与基因曲线无关,新生变体为良性。

[0552] 在一个实施方式中,至少一些候选变体为蛋白截断变体。

[0553] 在另一实施方式中,至少一些候选变体为错义变体、。

[0554] 全基因组富集分析

[0555] 图58为全基因组富集分析的一个实施方式。在另一实施方式中,异常剪接检测器进一步配置为进行全基因组富集分析,全基因组富集分析用于确定已确定会引起异常剪接的变体的致病性。全基因组富集分析包括应用训练的ACNN识别引起异常剪接的、从健康个体群组中取样的多个基因的第一组新生变体,应用训练的ACNN识别引起异常剪接的、从具有基因缺陷的个体群组中取样的多个基因的第二组新生变体,以及将第一组和第二组新生变体的数量进行比较并基于比较结果确定第二组新生变体富集在具有基因缺陷的个体组中且第二组新生变体具有致病性。在某些实施方式中,基因缺陷为自闭症谱系障碍(简称ASD)。在其他实施方式中,基因缺陷为发育迟缓障碍(简称DDD)。

[0556] 在某些实施方式中,异常剪接检测器进一步配置为运用生成p-值作为输出的统计测试进行比较。在一个实施方式中,可利用各自的群组大小进一步将比较参数化。

[0557] 在某些实施方式中,异常剪接检测器进一步配置为将第一组和第二组各自的数量进行比较,并基于比较结果确定第二组新生变体未富集在具有基因缺陷的个体群组中且第二组新生变体为良性。

[0558] 如图58所示,除了每个个体突变量(4)外,还描述了健康群组(0.001)和影响群组(0.004)的突变率。

[0559] 讨论

[0560] 尽管外显子测序在严重基因缺点病人中的诊断率有限,临床测序侧重于罕见编码突变,由于很难解释,很大程度上忽略了非编码基因组变异。本文中我们介绍了一种准确预测原始核苷酸顺序剪接的深度学习网络,从而识别出非编码突变,这些非编码突变破坏了产生的蛋白上具有严重后果的外显子和内含子的正常模式。我们展示了RNA测序技术高倍率验证的预测的隐秘剪接突变对人类具有剧毒,是引起罕见的遗传性疾病的主要原因。

[0561] 通过使用作为剪接体的计算机模型的深度学习模型,我们能够重建特异性决定子,使得剪接体的体内精度非常精确。我们重申了过去四十多年所做的对于剪接机制的研究的众多发现,指出剪接体在其决策中结合大量短期和长期特异性决定子。尤其,我们发现由于短期和长期决定子的存在,如外显子及内含子长度和核小体定位,解释了感知到的大部分剪接图案的退化,这比在图形水平补偿和渲染额外特异性更为必要。我们的发现展示了对提供生物见解的深度学习模型的承诺,而不是仅仅作为黑盒分类器。

[0562] 深度学习是生物学中一种较新的技术,不是没有潜在权衡的。通过学习从序列中自动提取特征,深度学习模型可使用人类专家未做详细描述的新型序列决定子,但是也存在风险,即模型可结合不反映剪辑体真实行为的特征。这些无关特征可增加预测标注的外显子-内含子边界的明显准确度,降低预测遗传变异引起的任意序列变化的剪接变化效果的准确度。由于变体的准确预测为模型能概括为真正的生物学提供了最为有力的证据,我们可使用三种完全正交方法对预测的剪接变化变体进行验证:RNA测序技术,人类种群自然

选择和对比对照群组,新生变体富集。然而并不完全排除将无关特征并入到模型中的情况,产生的模型似乎对于真正的剪接生物学足够忠诚,具有重要的实际应用意义,例如,识别患有基因疾病患者中隐秘剪接突变。

[0563] 与蛋白截断突变的其他种类相比,由于不完全外显剪接改变变体,隐秘剪接突变特别有趣的一方面是可变剪接普遍存在的现象,这倾向于相对于可变剪接位点减弱规范剪接位点,使得RNA测序技术中产生变异转录本和正常转录本的混合体。通过观察,这些变体频繁驱使组织变异性可变剪接,强调了引起新型可变剪接多样性的隐秘剪接突变所发挥的意想不到的作用。通过相关组织的RNA测序技术在剪接点注解上训练深度学习模型将会是一个潜在的发展方向,以便得到可变剪接的组织特异性模型。对直接来源于RNA测序技术数据的注解进行网络训练同样有助于填补GENCODE注解中的空白,改善了模型在变体预测方面的性能(如图52A和52B所示)。

[0564] 我们对于非编码基因组中突变导致人类疾病的理解远未完成。儿童神经发育障碍中外显新生隐秘剪接突变的发现表明对非编码基因组的改进说明使得具有严重基因缺陷的患者直接受益。隐秘剪接突变在癌症(Jung et al.,2015;Sanz et al.,2010;Supek et al.,2014)也具有重要作用,剪接因子中频发体细胞突变已被证明剪接特异性会发生广泛变化。为理解不同组织和细胞上下文中剪接调控,特别是在突变情况下直接影响剪接体中的蛋白,仍然有很多工作要做。根据寡核苷酸疗法的研究进展,该寡核苷酸疗法可潜在性地以序列特异性地(Finkel et al.,2017)靶向剪接缺陷,更好的理解调控这一非凡过程的调控机制,这可为新候选用于干预治疗铺平了道路。

[0565] 图37A,37B,37C,37D,37E,37F,37G和37H描述了从具有深度学习的原始序列来预测剪接的一个实施方式。

[0566] 如图37A,对于前体mRNA转录本的每个位置,SpliceNet-10k使用10000个侧翼序列的核苷酸作为输入并预测所述位置为剪接受体或剪接供体亦或所述位置既不是剪接受体也不是剪接供体。

[0567] 图37B表示了使用MaxEntScan(顶部)和SpliceNet-10k(底部)用于CFTR基因得分的全前体mRNA转录本,还表明了预测的受体(红色箭头)和供体(绿色箭头)位点及外显子的实际位置(黑盒)。对于每一种方法,我们运用阈值,所述阈值使得预测位点数等于实际位点总数。

[0568] 如图37C,我们测量了RNA测序技术中每个外显子的包含率并显示不同包含率处外显子的SpliceNet-10k得分分布。表明了外显子受体和供体得分的最大值。

[0569] 如图37D,通过计算机模拟U2SURP基因中外显子9周围每个核苷酸的影响。当所述核苷酸突变(Δ 得分)时,每个核苷酸的垂直尺寸表明预测受体位点(黑色箭头)强度降低。

[0570] 如图37F,输入序列上下文的大小影响网络的准确度。Top-k准确度为阈值处正确预测的剪辑位点的得分,预测位点数等于现有位点的实际数量。PR-AUC为精度召回曲线下方的区域。我们还展示了用于剪接位点检测的三种其他算法的Top-k准确度和PR-AUC。

[0571] 如图37F,由SpliceNet-80nt(局部基序得分)和SpliceNet-10k来预测相邻剪接位点的外显子/内含子长度和强度的关系。上下文中显示了外显子长度(黄色)和内含子长度(粉色)的全基因组分布。x轴为对数分度。

[0572] 如图37G,相距150nt的一对剪接受体和剪接供体基序沿着HMGR基因向前。如

SpliceNet-10k所预测的,显示了每个位置处K562核小体信号和在此位置成对形成外显子的可能性。

[0573] 如图37H,显示了SpliceNet-10k模型预测的私有突变附近平均K562和GM12878核小体信号,用于在GTEx群组中创建新型外显子。显示了置换测试获得的P值。

[0574] 图38A,38B,38C,38D,38E,38F及38G描述了RNA测序技术数据中罕见隐秘剪接突变验证的一个实施方式。

[0575] 如图38A,对于rs397515893:一个与心肌病相关MYBPC3内含子中的致病性隐秘剪接变体,为评估突变的剪接变化影响,SpliceNet-10k预测突变或不突变前体mRNA序列基因中每个位置的受体和供体得分。突变 Δ 得分为距变体50nt以内剪接预测得分的最大变化。

[0576] 如图38B,我们使用SpliceNet-10k模型对私有基因变体(通过在GTEx群组中149个个体中的一个观察得到)进行评分。显示了预测会改变剪接(Δ 得分大于0.2,蓝色)的私有变体或私有外显子跳跃点(顶部)或私有受体和供体诶点(底部)附近的、不影响剪接(Δ 得分小于0.01,红色)的私有变体的富集。与置换所得预期数量相比,y轴表示同一个体中私有剪接事件和附近私有基因变体同时发生的次数。

[0577] 如图38C,显示了创建具有不完全外显率的新型供体位点的PYGB中一个杂合同义突变变体的实例。显示了变体个体和对照个体的RNA测序技术范围,接点解读数和接点位置(蓝色和灰色箭头)。计算效应量作为带变体个体和不带变体个体之间新接点用法的差值。在下方叠加条状图表中,我们展示了使用注释或新接点(分别为无剪接和新接点)的参考等位基因或交互等位基因的解读数量。参考等位基因解读总数明显不同于交互等位基因解读总数,这就表明RNA测序技术数据中新接点处60%的转录体剪接缺失,这大概是因为无意义介导的衰变(NMD)。

[0578] 如图38D,显示了部分通过GTExRNA测序技术数据进行验证的SpliceNet-10k模型所预测的隐秘剪接突变。根据覆盖率和无意义介导的衰变,基本受体或供体二核苷酸(虚线所示)的破坏验证率小于100%。

[0579] 如图38E,显示了用于验证的隐秘剪接预测的效应量的分布。虚线(50%)与全外显杂合变体的预计效应量相对应。根据无意义介导的衰变或未予说明的剪接形式变化,所测得的基本受体或供体二核苷酸破坏的效应量小于50%。

[0580] 如图38F,显示了不同 Δ 得分临界值GTEx群组中检测剪接变化私有变体的SpliceNet-10k的灵敏度。变体分裂成深度基因内变体(距离外显子大于50nt)和附近外显子变体(与外显子重合或距离外显子-内含子边界 ≤ 50 nt)。

[0581] 如图38G,显示了不同置信临界值剪接位点及三种其他方法预测的SpliceNet10-k的验证率和灵敏度。SpliceNet10-k曲线上的三个点表示 Δ 得分临界值处SpliceNet10-k的性能:0.2、0.5及0.8。对于其他三种算法,曲线上的三个点表明阈值处的性能,所述阈值预测隐秘剪接变体的相同数量为 Δ 得分临界值处SpliceNet10-k:0.2、0.5及0.8。

[0582] 图39A,39B及39C表示频繁创造组织特异性可变剪接的隐秘剪接变体的一个实施方式。

[0583] 如图39A,显示了创造一新型供体位点的CDC25B中一个杂合外显变体的实例。变体仅属于GTEx群组中一个单独的个体并显示出组织特异性可变剪接,与纤维母细胞相比,所述可变剪接有利于肌肉中很大一部分新型剪接形式(根据费希尔精确检验,P等于0.006)。

在肌肉和纤维母细胞中,显示了变体个体和对照个体的RNA测序技术范围,接点解读数和接点位置(蓝色和灰色箭头)。

[0584] 如图39B,显示了FAM229B中一杂合外显受体创造变体的实例,所述FAM229B展示了具有变体的GTEx群组中全部三个个体的一致性组织特异性效应。显示了用于具有变体的三个单体和对照单体的动脉和肺RNA测序技术。

[0585] 如图39C,通过卡方检验评估跨域表达组织、与新型接点有显著联系的GTEx中的部分剪接位点创建变体。具有中低 Δ 得分的验证的隐秘剪接变体更可能引起组织特异性可变剪接(根据费希尔精确检验,P等于0.015)。

[0586] 图40A、40B、40C、40D和40E描述了在人群中具有剧毒性的预测的隐秘剪接变体的一个实施方式。

[0587] 如图40A,相对于仅在60706个个体中观察得到的罕见变体,人群中具有可信赖的预测剪接变化效应(Δ 得分 ≥ 0.8)的同义突变和基因内变体以一般等位基因频率大大耗散。比值率4.58(通过卡方检验,P小于 10^{-127})表明78%的最近出现的预测隐秘剪接变体具有足够的毒害性,由自然选择去除。

[0588] 如图40B,显示了部分有害ExAC数据组中蛋白截断变体和预测的同义突变和基因内隐秘剪接变体,所述部分变体计算在(A)内。

[0589] 如图40C,显示了部分有害(Δ 得分 ≥ 0.8)ExAC数据组中同义突变和基因内隐秘剪接获得变体,所述部分变体基于变体是否引起码移进行拆分。

[0590] 如图40D,显示了部分有害gnomAD数据组中蛋白截断变体和预测的深度基因内(距已知外显子-内含子边界大于50nt)隐秘剪接变体。

[0591] 如图40E,显示了每个个体人类基因组罕见(等位基因频率小于0.1%)蛋白截断变体和罕见功能隐秘剪接变体的平均数。基于有害预测分数,评估了预计功能化的隐秘剪接突变数。预测总数较高。

[0592] 图41A、41B、41C、41D、41E和41F表示患有罕见遗传疾病患者的新生隐秘剪辑突变的一个实施方式。

[0593] 如图41A,显示了每个译码发育障碍群组患者的预测隐秘剪接新生突变,来自孤独症儿童样本数据库自闭症测序联盟的患有孤独症谱系障碍个体以及健康对照。调整超过健康对照的DDD和ASD群组富集用以DDD和ASD群组间的变体确定。误差线表示95%的置信区间。

[0594] 如图41B,对比健康对照,基于每个分类的富集,显示了DDD和ASD群组功能分类的致病新生突变的评估比例。

[0595] 如图41C,对比不同 Δ 得分阈值的健康对照,显示了DDD和ASD群组中隐秘剪接新生突变的富集和过量。

[0596] 如图41D,当富集分析中同时包括预测隐秘剪接突变和蛋白编码突变时,显示了DDD和ASD群组($FDR < 0.01$)中新生突变富集的新型候选疾病基因列表。显示了存在于多个个体中的表型。

[0597] 如图41E,在RNA测序技术上验证的自闭症患者的预测新生隐秘剪接突变的三个实例分别引起内含子保留、外显子跳跃和外显子扩展。每个实例中,顶端显示了受影响单体的RNA测序技术范围和结点数,底端显示了未突变的对照单体。正义链上显示了与基因转录相

关的序列。蓝色箭头和灰色箭头分别区分具有变体的个体和对照个体的接点位置。

[0598] 图41F为用于通过RNA测序技术的实验验证的36个预测隐秘剪接位点的验证状态。

[0599] 实验模型与受试者详细信息

[0600] 早前Iosifov等人在Nature 2014(表S1)上发布了36名自闭症患者的受试者详细信息,这些信息可以采用我们论文中表S4第1列中的匿名标识符进行交叉引用。

[0601] 方法细节

[0602] I.对剪接预测的深度学习

[0603] SpliceNet结构

[0604] 我们训练了若干基于超深卷积神经网络的模型来计算预测前体mRNA核苷酸序列的剪接。我们设计了四种结构,即,SpliceNet-80nt、SpliceNet-400nt、SpliceNet-2k和SpliceNet-10k,它们分别以目的位置每一侧上的40、200、1,000和5,000个核苷酸作为输入,并输出该位置为剪接受体和供体的概率。更精确地,模型的输入为单热编码的核苷酸序列,其中A、C、G和T(或相当于U)分别被编码为[1,0,0,0]、[0,1,0,0]、[0,0,1,0]和[0,0,0,1],模型的输出由三个得分组成(总和为1),对应于目的位置为剪接受体、剪接供体和二者都不是的概率。

[0605] SpliceNet结构的基本单元是残差块(He等人,2016b),它由批归一化层(Ioffe和Szegedy,2015),修正线性单元(ReLU)和以特定方式组织的卷积单元组成(图21、22、23和24)。在设计深度神经网络时常采用残差块。在开发残差块之前,由于梯度爆炸/消失的问题,很难训练由许多一个接一个堆叠的卷积单元组成的深度神经网络(Glorot和Bengio,2010),并且增加这种神经网络的深度常常会导致较大的训练误差(He等人,2016a)。通过一系列的实验,表明由许多一个接一个堆叠的残差块组成的结构克服了这些问题(He等人,2016a)。

[0606] 图21、22、23和24提供了完整的SpliceNet结构。该结构由K个将输入层连接至倒数第二层的堆叠残差块和具有softmax激活作用的将倒数第二层连接至输出层的卷积单元组成。堆叠残差块,从而使第i个残差快的输出连接至第i+1个残差块的输入。进一步地,将每隔三个残差块的输出加入到倒数第二层的输入中。在深度神经网络中常采用这种“跳跃连接”来增加训练期间的收敛速度(Oord等人,2016)。

[0607] 每个残差块具有三个超参数,N,W和D,其中N表示卷积内核的数目,W表示窗口尺寸,D表示每个卷积内核的膨胀率(Yu和Koltum,2016)。由于具有窗口尺寸W和膨胀率D的卷积内核可以提取跨越(W-1)D个相邻位置的特征,具有超参数N,W和D的残差块可以提取跨越 $2(W-1)D$ 个相邻位置的特征。因此,由 $S = \sum_{i=1}^K 2(W_i - 1)D_i$ 得出SpliceNet结构的总相邻跨度,其中 N_i , W_i 和 D_i 为第i个残差块的超参数。对于SpliceNet-80nt、SpliceNet-400nt、SpliceNet-2k和SpliceNet-10k结构而言,选择残差块的数目和每个残差块的超参数,以使S分别等于80、400、2,000和10,000。

[0608] 除了含有卷积单元外,SpliceNet结构仅有归一化和非线性激活单元。因而,模型可以在具有可变序列长度的序列到序列模式下使用(Oord等人,2016)。例如,SpliceNet-10k模型的输入为一个长度为 $S/2+1+S/2$ 的单热编码的核苷酸序列,输出为 1×3 矩阵,对应于输入中1中心位置即排除第一个和最后 $S/2$ 个核苷酸后剩余的位置的三个得分。可利用该特征来节省训练与测试期间的巨大计算量。这是由于大多数对彼此接近的位置的计算都是

通用的,而当共享计算以序列到序列模式使用时,其只需要由模型执行一次。

[0609] 我们的模型采用残差块结构,其因在图像分类方面的成功已被广泛采用。残差模块包括卷积重复单元,其中穿插有使更早层的信息跳过残差块的跳跃连接。在每个残差块中,首先对输入层进行批归一化,然后使用修正线性单元(ReLU)对激活层进行归一化。然后激活穿过1D卷积层。1D卷积层的中间输出再次被批归一化并被ReLU激活,接着是另一个1D卷积层。在第二次1D卷积结束时,我们将其输出与原始输入求和加入到残差块中,通过使原始输入信息绕过残差块来充当跳跃连接。在这种被作者称为深度残差学习网络的结构中,输入保持其原始状态,剩余连接不受模型的非线性激活,从而对更深层的网络进行有效训练。

[0610] 在残差块之后,softmax层计算每个氨基酸的三种状态的概率,其中最大softmax概率确定氨基酸的状态。利用ADAM优化器,采用整个蛋白序列的累积分类交叉熵损失函数对模型进行训练。

[0611] 空洞/膨胀卷积容许有大的感受野,很少有可训练的参数。空洞/膨胀卷积是一种通过以一定的步长跳过输入值(也称为空洞卷积率或膨胀因子)来将内核应用于大于其长度的区域的卷积,也称为空洞卷积率或膨胀因子。空洞卷积/膨胀卷积在卷积滤波器/内核的元件之间增加了间隔,以便在执行卷积操作时可以考虑更大间隔的相邻输入项(例如,核苷酸、氨基酸)。这使得可以将远程语境依赖并入输入中。空洞卷积保留了部分褶积计算,以在处理相邻核苷酸时再次使用。

[0612] 所示出的示例使用1D卷积。在其他实施方案中,模型可以使用不同类型的卷积,如2D卷积、2D卷积、膨胀或空洞卷积、转置卷积、可分离卷积和深度可分离卷积。与S形或双曲线正切等饱和非线性相比,一些层还使用了ReLU激活函数,大大地加速了随机梯度下降的收敛。被所公开技术使用的激活函数的其它示例包括参数ReLU、漏ReLU和指数线性单元(ELU)。

[0613] 一些层还使用批归一化(Ioffe和Szegedy 2015)。关于批归一化,卷积神经网络(CNN)中各层的分布在训练过程中会发生变化,并且各层的分布彼此不同。这降低了优化算法的收敛速度。批归一化是克服这个问题的技术。用x表示批归一化层的输入,用z表示其输出,批归一化对x应用以下转换:

$$[0614] \quad z = \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} \gamma + \beta$$

[0615] 批归一化通过使用 μ 和 σ 以及线性标度对输入x应用平均方差归一化,并通过使用 γ 和 β 使它移位。使用被称为指数移动平均法的方法计算训练集上当前层的归一化参数 μ 和 σ 。换言之,它们不是可训练参数。相反, γ 和 β 为可训练参数。训练过程中计算的 μ 和 σ 值在推理过程中用于前向通过。

[0616] 模型训练与测试

[0617] 我们从UCSC表浏览器下载了GENCODE(Harrow等人,2012)V24lift37基因注释表,并提取了20,287个蛋白质编码基因注释,从而在有多个亚型可用时选择主转录本。我们去除了没有任何剪接点的基因,将剩余的基因拆分为以下的训练和测试集基因:使用属于2、4、6、8、10-22、X和Y号染色体的基因训练模型(13,384个基因,130,796个供体-受体对)。我们随机选择了10%的训练基因,用它们来确定训练过程中早期停止的点,其余的用于训练

模型。为了测试模型,我们使用了来自1、3、5、7和9号染色体的基因,它们没有任何旁系同源基因(1,652个基因,14,289个供体-受体对)。为此,我们参考了来自<http://grch37.ensembl.org/biomart/martview>的人类基因的旁系同源基因列表。

[0618] 我们使用以下程序,以序列到序列模式来训练和测试模型,块大小为 $l=5,000$ 。对于每个基因,从hg19/GRCh37组合物中提取规范转录起始和终止位点之间的mRNA转录序列。输入的mRNA转录序列被以下序列单热编码:分别映射到 $[1,0,0,0]$ 、 $[0,1,0,0]$ 、 $[0,0,1,0]$ 和 $[0,0,0,1]$ 的A、C、G、T/U。对单热编码的核苷酸序列进行零填充,直到长度为5,000的倍数,然后在开始和终止处进一步进行零填充,侧翼序列长度为 $S/2$,其中对于SpliceNet-80nt、SpliceNet-400nt、SpliceNet-2k和SpliceNet-10k模型, S 分别等于80、400、2,000和10,000。然后将填充的核苷酸序列拆分为 $S/2+5,000+S/2$ 长度的块,使得第 i 个块由 $5,000(i-1)-S/2+1$ 到 $5,000i+S/2$ 的核苷酸位置组成。类似地,剪接输出标签序列被单热编码,如下所示:非剪接位点、剪接受体(对应外显子的第一个核苷酸)和剪接供体(对应外显子的最后一个核苷酸)分别映射到 $[1,0,0]$ 、 $[0,1,0]$ 和 $[0,0,1]$ 。单热编码的剪接输出标签序列被零填充,直到长度为5,000的倍数,然后再被拆分成5,000长度的块,使得第 i 个块由 $5,000(i-1)+1$ 到 $5,000i$ 的位置组成。单热编码的核苷酸序列和相应的单热编码的标签序列分别作为模型的输入和模型的目标输出。

[0619] 在两台NVIDIA GeForce GTX 1080Ti-gpu上对模型进行了10次训练,批次大小为12。在训练过程中使用Adam优化器(Kingma和Ba,2015)最小化目标和预测输出之间的类别交叉熵损失。优化器的学习率在前6次被设置为0.001,然后在随后的每次降低二分之一。对于每个结构,我们重复训练过程5次,得到5个训练模型(图53A和53B)。在测试过程中,使用所有5个训练模型对每个输入进行评估,并将其输出的平均值作为预测输出。我们使用用于图中37A及其他相关附图中分析的模式。

[0620] 对于在涉及剪接改变变体的识别的图38A-G、39A-C、40A-E和41A-F中分析,我们增加了GENCODE注释的训练集以包括通常在2、4、6、8、10-22、X、Y号染色体(67,012个剪接供体和62,911个剪接受体)上的GTEx群组中观察到的新剪接点。这使训练集中的剪接连接注释的数目增加大约50%。与仅在GENCODE注释上训练的网络相比,在组合数据集上训练网络提高了在RNA测序数据中检测剪接改变变体,特别是用于预测深内含子剪接改变变体的灵敏度(图52A和52B),我们使用该网络进行涉及变体评估的分析(图38A-G、39A-C、40A-E和41A-F及有关附图)。为了确保GTEx-RNA-seq数据集中不包含训练和评估之间的重叠,我们只在训练数据集中包含存在于5个以上个体中的接点,并且只评估4个或更少个体中存在的变体上的网络的性能。在方法的GTEx分析部分的“剪接连接检测”中描述了新剪接连接识别的细节。

[0621] Top-k准确度

[0622] 由于大多数位置都不是剪接位置,像正确分类的位置百分比这样的准确度度量在很大程度上是无效的。相反,我们使用两个在这种情况下有效的度量来评估模型,即top-k准确度和精度-召回曲线下的面积。特定类的top-k准确度定义为:假设测试集具有属于该类的 k 个位置。我们选择阈值,从而就可以准确地预测 k 个测试集位置是否属于该类。这些 k 个预测位置中真正属于该类的部分被报告为top-k准确度。实际上,这等于选择阈值时的准确度,以便精度和召回具有相同的值。

[0623] 对lincRNA的模型评估

[0624] 我们获得了基于GENCODE V24lift37注释的所有lincRNA转录本的列表。与蛋白质编码基因不同,lincRNA在GENCODE注释中没有被分配一个主转录本。为了最小化验证集中的冗余,我们鉴定了每个lincRNA基因的最长总外显子序列的转录本,并将其称为该基因的规范转录本。由于lincRNA注释的可靠性低于蛋白质编码基因注释,并且这种错误注释会影响我们对top-k准确度的估计,因此我们使用GTEx数据来消除具有潜在注释问题的lincRNA(有关这些数据的详细信息,请参阅下面的“GTEx数据集分析”一节)。对于每个lincRNA,我们计算了所有GTEx样本的lincRNA长度上所映射的所有拆分解读(有关详细信息,请参见下面的“剪接点检测”)。这是对使用注释连接或新连接的lincRNA的总连接-跨越解读的估计。我们还计算了跨越规范转录本接点的解读次数。我们只考虑了所有GTEx样本中至少95%接点-跨越解读与规范转录本对应的lincRNA。我们还要求在GTEx群组中至少观察一次规范转录本的所有接点(不包括跨越长度<10nt的内含子的接点)。为了计算top-k准确度,我们只考虑了通过上述过滤器的lincRNA的规范转录本的连接(781个转录本,1047个接点)。

[0625] 从前体mRNA序列识别剪接点

[0626] 在图37B中,我们比较了MaxEntScan和SpliceNet-10k的关于从序列中识别基因的规范外显子边界的性能。我们将存在于我们的测试集中且有26个典型的剪接受体和供体的CFTR基因作为一个案例进行研究,并使用MaxEntScan和SpliceNet-10k获得了从规范转录起始位点(chr7:117120017)到规范转录结束位点(chr7:117308719)之间的188,703个位置中每个位置的受体和供体得分。如果在评估top-k准确度的同时其相应得分大于所选阈值,则将该位置归类为剪接受体或供体。MaxEntScan预测了49个剪接受体和22个剪接供体,其中9个受体和5个供体分别是真实的剪接受体和供体。为了更好的可视化,我们展示了Max扫描的预对数得分(剪除到最大值2,500)。SplitNet-10k预测了26个剪接受体和26个剪接供体,均正确。对于图42B,我们使用LINC00467基因重复分析。

[0627] GENCODE-注释的剪接点处外显子包含的评估

[0628] 我们根据GTEx RNA测序数据计算了所有GENCODE注释的外显子的包含率(图37C)。对于每个不包括每个基因的第一个和最后一个外显子的外显子,我们将包含率计算为:

$$[0629] \quad \frac{(L + R)/2}{S + (L + R)/2}$$

[0630] 其中,L为所有GTEx样本中从上一个规范外显子到考虑中的外显子之间的接点的总解读计数,R是从考虑中的外显子到下一个规范外显子之间的接点的总解读计数,S是从上一个规范外显子到下一个规范外显子之间的跳跃接点的总解读计数。

[0631] 各个核苷酸对剪接位点识别的意义

[0632] 在图37D中,我们鉴定了被SplitNet-10k认为对将一个位置分类为剪接受体很重要的核苷酸。为此,我们考虑了存在于我们的测试集中的U2SUPR基因中chr3:142740192处的剪接受体。核苷酸相对于剪接受体的“重要性得分”定义如下: s_{ref} 表示所考虑的剪接受体的受体得分。通过用A、C、G和T替换所考虑的核苷酸来重新计算受体得分。让这些得分分别用 S_A 、 S_C 、 S_G 和 S_T 表示。将核苷酸的重要性得分估计为:

$$[0633] \quad s_{ref} - \frac{S_A + S_C + S_G + S_T}{4}$$

[0634] 该过程通常被称为in-silico诱变(Zhou和Troyanskaya,2015)。我们绘制了从chr3:142740137到chr3:142740263的127个核苷酸,以使每一个核苷酸的高度为其在关于chr3:142740192处剪接受体的重要得分。绘图功能改编自DeepLIFT(Shrikumar等人,2017)软件。

[0635] TACTAAC与GAAGAA基序对剪接的作用

[0636] 为了研究分支点序列的位置对受体强度的影响,我们首先利用SplitNet-10k获得14,289个测试集剪接受体的受体得分, y_{ref} 表示包含这些得分的向量。对于从0到100的每个i值,我们做了以下工作:对于每个测试集剪接受体,我们用TACTAAC替换剪接受体前的第i到i-6位置的核苷酸,并使用SplitNet-10k重新计算受体得分。包含这些得分的向量用 $y_{alt,i}$ 表示。

[0637] 我们在图43A中将下列数量绘制为i的函数:

$$[0638] \quad \text{mean}(y_{alt,i} - y_{ref})$$

[0639] 对于图43B,我们利用SR-蛋白基序GAAGAA重复了相同的程序。在这种情形下,我们还研究了存在于剪接受体之后时基序的影响以及对供体强度的影响。基于对k-mer空间的综合研究,GAAGAA和TACTAAC是对受体和供体强度影响最大的基序。

[0640] 外显子与内含子长度在剪接中的作用

[0641] 为了研究外显子长度对剪接的作用,我们筛选出测试集外显子,其为第一个或最后一个外显子。该筛选步骤从14,289个外显子中去掉了1,652个。我们将剩下的12,637个外显子按长度增加的顺序进行排序。对于每一个,我们利用SplicNet-80nt通过对剪接受体位点处的受体得分和剪接供体位点处的供体得分求平均数来计算剪接得分。在图37F中,我们将剪接得分绘制为外显子长度的函数。在绘制之前,我们应用了以下平滑过程: x 表示包含外显子长度的向量, y 表示包含其相应剪接得分的向量。我们使用大小为2,500的平均窗口平滑 x 和 y 。

[0642] 我们通过使用SplicNet-10k计算拼接得分来重复此分析。在背景中,我们显示了用于此分析的12,637个外显子的长度直方图。我们应用了一个相似的分析来研究内含子长度对剪接的影响,主要的区别是不需要排除第一个和最后一个外显子。

[0643] 核小体对剪接的作用

[0644] 我们从UCSC基因组浏览器下载了K562细胞系的核小体数据。我们使用测试集中的HMGR基因作为一个轶事示例来证明核小体定位对SplicNet-10k得分的影响。对于基因中的每个位置p,我们计算其“栽培剪接得分”如下:

[0645] • 从p+74到p+81位置的8个核苷酸被供体基序AGGTAAGG取代。

[0646] • 从p-78到p-75的4个核苷酸被受体基序TAGG取代。

[0647] • 从p-98到p-79位置的20个核苷酸被多嘧啶束CCTCCTTCGCCCTC取代。

[0648] • 从p-105到p-99的7个核苷酸被分支点序列CACTAAC取代。

[0649] • 将p-75处受体得分与由SplitNet-10k预测的p-75处的供体得分的平均值作为植入的剪接得分。

[0650] 图37G显示了从chr5:74652154到chr5:74657153的5,000个位置处的K562核小体信号和植入的剪接得分。

[0651] 为了计算这两个轨迹之间的全基因组Spearman相关性,我们随机选取了100万个

基因间位置,这些位置距离所有规范基因至少100,000nt。对于每一个位置,我们计算其栽培剪接得分以及其平均K562核小体信号(窗口尺寸为50用于求平均数)。这两个值在100万个位置上的相关性如图37G所示。我们根据这些位置的GC含量(使用植入的受体和供体基序之间的核苷酸估计)进一步对其进行分类,二进制大小为0.02。我们在图44A中显示了每个二进制的全基因组Spearman相关性。

[0652] 对于14,289个测试集剪接受体中的每一个,我们提取每侧50个核苷酸内的核小体数据,并将其核小体富集计算为外显子侧的平均信号除以内含子侧的平均信号。我们按照剪接受体核小体富集的增加顺序对其进行分类,并使用SplitNet-80nt计算其受体得分。在图44B中,受体得分是核小体富集的函数。在绘制之前,应用了图37F中使用的平滑程序。我们使用SplitNet-10k重复了该分析并用于14,289测试剪接供体。

[0653] 核小体信号在新外显子处的富集

[0654] 在图37H中,我们想看看预测的新外显子周围的核小体信号。为了确保我们研究的是高度可信的新外显子,我们只选择了单变体(存在于单个GTEx个体中的变体),其中预测获得的接点对于具有变体的个体是完全私有的。此外,为了消除附近外显子的混淆效应,我们只观察了距注释外显子至少750nt的内含子变体。我们从UCSC浏览器下载GM12878和K562细胞株的核小体信号,并从每个预测的新受体或供体位点提取750nt以内的核小体信号。我们对两个细胞系之间的核小体信号求平均数,并翻转了负链上重叠基因的变异的信号载体。我们将信号从受体位点向右移动了70nt,从供体位点向左移动了70nt。移位后,受体和供体的核小体信号都集中在长度为140nt的理想外显子的中间,这是GENCODE v19注释中外显子的中间长度。最后,我们对所有移位的信号求平均数,并通过计算以每个位置为中心的11nt窗口内的平均值来平滑得到的信号。

[0655] 为了检验关联性,我们选择了随机的单SNV,它们至少离注释外显子750nt远,并且被模型预测对剪接没有影响(Δ 得分 <0.01)。我们创建了1000个这样的SNV随机样本,每个样本的SNV数目与用于图37H(128个位点)的剪接位点增益位点集的SNV数目相同。对于每个随机样本,我们计算了一个平滑平均信号,如上所述。由于预测随机SNV不能产生新的外显子,我们将每个SNV的核小体信号集中在SNV本身,并随机地向左移动70nt或向右移动70nt。然后,我们将图37H中间碱基处的核小体信号与从该碱基处1000个模拟所得的信号进行比较。将经验p值计算为具有大于或等于剪接位点增益变量观测值的中间值的模拟集的分。

[0656] 网络对外显子密度差异的鲁棒性

[0657] 为了研究网络预测的通用性,我们评估了不同外显子密度区域的SplitNet-10k。我们首先根据10,000个核苷酸窗口(每侧5,000个核苷酸)中出现的规范外显子数目将测试集位置分为5类(图54)。为了确保外显子计数为每个位置的整数值,我们将窗口中的外显子起始作为代理项。对于每一类,我们计算了top-k准确度和精度-召回曲线下的面积。不同类别的位置数目和k值不同(详见下表)。

	外显子计数	#位置	#剪接受体	#剪接供体
[0658]	1 个外显子	15,870,045	1,712	1,878
	2 个外显子	10,030,710	2,294	2,209
	3 个外显子	6,927,885	2,351	2,273
	4 个外显子	4,621,341	2,095	2,042
	≥5 个外显子	7,247,582	5,679	5,582

[0659] 总体5个模型中每一个的网络的鲁棒性

[0660] 训练多个模型并将其预测的平均值作为输出是机器学习中获得更好预测性能的常用策略,称为总体学习。在图53A中,我们展示了我们所训练的5个SplitNet-10k模型的top-k准确度和精度-召回曲线下的面积来构建总体。结果清楚地表明了训练过程的稳定性。

[0661] 我们还计算了他们的预测之间的Pearson相关性。由于基因组中大多数位置不是剪接位点,大多数模型的预测之间的相关性将接近1,使得分析变得毫无意义。为了克服这个问题,我们只考虑测试集中那些至少被一个模型分配了一个大于或等于0.01的受体或供体得分的位置。53,272个位置(大约相同数目的剪接位点和非剪接位点)都满足该标准。图53B中概括该结果。模型预测之间非常高的Pearson相关性进一步说明了它们的鲁棒性。

[0662] 在图53C中,我们展示了用于构建总体的模型数目对性能的影响。结果表明,随着模型数目的增加,性能会提高,但回报率会降低。

[0663] II. 对GTEx RNA测序数据基的分析

[0664] 单核苷酸变体的 Δ 得分

[0665] 我们将单核苷酸变异引起的剪接变化量化如下:我们首先使用参考核苷酸,计算变体周围101个位置(每侧50个位置)的受体和供体得分。假设这些得分分别由向量 a_{ref} 和 d_{ref} 表示。然后我们使用替代核苷酸并重新计算受体和供体得分。这些得分分别用向量 a_{alt} 和 d_{alt} 表示。

[0666] 我们估计以下四个数量:

[0667] Δ 得分受体增益) = $\max(a_{alt} - a_{ref})$

[0668] Δ 得分受体损失) = $\max(a_{ref} - a_{alt})$

[0669] Δ 得分供体增益) = $\max(d_{alt} - d_{ref})$

[0670] Δ 得分供体损失) = $\max(d_{ref} - d_{alt})$

[0671] 这四个得分的最大值称为变体的 Δ 得分。

[0672] 用于质量控制和变体的滤波的标准

[0673] 我们从dbGaP下载了GTEx VCF和RNA测序数据(研究登录号phs000424.v6.p1; https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v6.p1)。

[0674] 我们评估了在GTEx群组中出现最多的4个个体的常染色体SNV上的SplitNet的性能。特别是,如果一个变体在至少一个个体A中满足以下标准,则应予以考虑:

[0675] 1. 未筛选变体(VCF的FILTER字段为PASS)。

[0676] 2. 变体在个体A的VCF的INFO字段中未标记为MULTI_ALLELIC,且VCF在ALT字段中

含有单个等位基因。

[0677] 3.个体A与该变体杂合。

[0678] 4. $\text{alt_depth}/(\text{alt_depth}+\text{ref_depth})$ 的比值在0.25到0.75之间,其中 alt_depth 和 ref_depth 分别是单个A中支持替代等位基因和参考等位基因的解读。

[0679] 5.在个体A的VCF中,总深度 $\text{alt_depth}+\text{ref_depth}$ 在20到300之间。

[0680] 6.该变体重叠了一个基因体区域。基因体被定义为GENCODE (V24lift37) 的规范转录本的转录起始和终止之间的区域。

[0681] 对于在至少一个个体中满足这些标准的变体,我们认为出现变体的所有个体(即使不满足上述标准)都具有该变体。我们将单个个体中出现的变体称为单体,将2-4个个体中出现的变体称为普通变体。我们没有评估5个以上个体中出现的变体以防止与训练数据集重叠。

[0682] RNA测序解读比对

[0683] 我们用OLego(Wu等人,2013)将GTEx样本的解读与hg19参考映射,使查询解读和参考之间的编辑距离不超过4(参数-M 4)。注意,OLego可以完全从头开始操作,不需要任何基因注释。由于OLego在拼接解读的末端寻找剪接基序的存在,其比对可以偏向或背离分别破坏或创建剪接位点的SNV周围的参考。为了消除这种偏差,我们通过将带有PASS筛选器的个体的所有SNV插入hg19参考中来进一步为每个GTEx个体创建替代的参考序列。我们用具有相同参数的OLego将每个个体的所有样本映射到该个体的备用参考序列。对于每个样本,我们通过选取每个读对的最佳比对来将两个比对集(针对hg19参考和个体的替代参考)组合在一起。为了选择解读对P的最佳比对,我们使用了以下步骤:

[0684] 1.如果两组比对中两个解读P都未被映射,我们随机选择hg19或P的替代比对。

[0685] 2.如果P在一组比对中有更多未映射的末端(例如,P的两端映射到备选参考,但只有一端映射到hg19),我们选择P的两端被映射的比对。

[0686] 3.如果P的两端都在两组比对中都被映射,我们选择总错配最少的比对,或者随机选择错配次数相同的比对。

[0687] 比对的RNA测序数据中剪接点的检测

[0688] 我们用leafcutter_cluster,leafcutter包中的一个实用程序(Li et al.,2018)来检测和计数每个样本中的剪接点。我们需要一个单一的拼接解读来支持一个接点,并假设最大内含子长度为500kb(参数-m 1-1500000)。为了获得用于训练深度学习模型的高置信度连接集,我们编译了所有样本中所有切叶刀接点的联合,然后从满足以下任一标准的考虑接点中除去:

[0689] 1.与ENCODE黑名单区(UCSC基因组浏览器中hg19中的表WgEncodedCapabilityConsensusIncluded)重叠的接点的任意一端或一简单重复序列(来自UCSC基因组浏览器的hg19中的简单重复轨迹)。

[0690] 2.接点的两端都位于非规范外显子上(基于GENCODE版本V24lift37的规范转录本)。

[0691] 3.接点的两端位于不同的基因上,或任意一端位于非基因区。

[0692] 4.两端缺乏必需的GT/AG二核苷酸。

[0693] 在5个以上个体中存在的接点被用于增加GENCODE注释的剪接点的列表,以用于对

分析变体预测(图38A-G、39A-C、40A-E和41A-F)。关键资源表中提供了指向包含用于培训模型的剪接点的列表的文件的链接。

[0694] 虽然我们用leafcutter检测到的接点来增强训练数据集,但我们注意到,尽管使用了松弛参数,leafcutter在RNA测序数据中过滤了许多接点,并提供了良好的支持。这人为地降低了我们的验证率。因此,对于GTEx RNA测序验证分析(图38A-G和39A-C),我们直接从RNA测序解读数据中重新计算了一组接点和接点计数。我们统计了所有不重复的拆分映射解读,其中接点每侧上排列有至少10和5nt的MAPQ。允许一个解读跨越两个以上的外显子,在这种情况下,该解读被计数到每个连接处,两侧至少有5nt的映射序列。

[0695] 私有接点的定义

[0696] 如果一个接点至少满足以下条件之一,则认为该接点在A中是私有的:

[0697] 1. 在至少一个来自A的样品中,该接点具有至少3个解读,并且从未在任何其他个体中观察到。

[0698] 2. 至少有两个组织同时满足以下两个条件:

[0699] a. 来自组织中A个体的样品中接点的平均解读计数至少为10。

[0700] b. 平均而言,A个体的归一化解读至少是该组织中任何其他个体的两倍。在此,将样品中接点的归一化解读计数定义为对应基因的所有接点上的解读总数进行归一化后的接点的解读数。

[0701] 该测试忽略了来自其他个体(非A个体)的少于5个样本的组织。

[0702] 私有接点周围单例SNV的富集

[0703] 如果基于GENCODE注释对私有接点一末端进行了准确的注释,我们将其视为受体或供体增益的候选对象,并搜索距离未注释末端150nt以内的同一个体专有的单例SNV(出现在单个GTEx个体中的SNV)。如果私有接点的两末端都带有注释,则根据GENCODE注释,如果它跳过了同一基因的至少一个但不超过3个同一外显子,则认为它是一个私有外显子跳跃事件的候选对象。然后,我们从每个跳过的的外显子的末端开始,在150nt内搜索单例SNV。GENCODE外显子注释中缺少两端的私有接点将被忽略,这是因为其中很大一部分具有比对错误。

[0704] 为了计算新的私有受体或供体周围单例SNV的富集(图38B,底部),我们汇总了相对于私有接点的每个位置上的单例SNV的计数。如果重叠基因在负链上,则将相对位置翻转。我们将SNV分为两组:具有私有接点的个体私有SNV和具有不同个体的私有SNV。为了对结果信号进行平滑处理,我们在每个位置居中的7nt窗口中计算平均数。然后,我们计算了第一组(同一个体私有)的平滑计数与第二组(另一个体中私有)的平滑计数之比。对于新的私有外显子跳跃项(图38B,顶部),我们遵循类似的过程,在跳过的的外显子的末端周围汇总了单例SNV的计数。

[0705] GTEx RNA测序数据中模型预测的验证

[0706] 对于私有变体(在GTEx群组中出现在一个个体中)或常见变体(在GTEx群组中出现在两个到四个个体中),我们获得了参考和其他等位基因的深度学习模型的预测值并计算 Δ 得分。我们还获得了模型预测异常(新的或破坏的)接点的位置。然后,我们试图确定RNA测序数据中是否有证据支持,在预测位置的具有变体的个体中是否具有异常剪接。在许多情况下,模型可以用于预测同一变体的多种影响,例如破坏了注释的剪接供体的变体也可

能增加次优供体的使用,如图45所示。在这种情况下,该模型可以预测带注释的剪接位点处的供体损失以及次优位点处的供体增加。但是,出于验证目的,我们仅考虑了每个变体具有最高预计 Δ 得分的效果。因此,对于每个变体,我们分别考虑了预测的剪接位点创建和剪接位点破坏的作用。请注意,在模型训练期间排除了出现在不到五个个体中的接点,以避免对模型进行训练时评估新的接点。

[0707] 基于私有剪接点预测的隐秘剪接突变的验证

[0708] 对于每个预计会引起新的接点形成的私有变体,我们使用网络预测新创建的异常剪接点的位置,并查看RNA测序数据以验证这种新的接点是否仅出现在具有SNV并且没有其他GTE_x的个体中。同样,对于预测会引起影响外显子X剪接位点的剪接位点丢失的变体,我们从前一个规范外显子(根据GENCODE注释在X的上游)到下一个规范外显子(X的下游)寻找新的外显子跳跃事件,该外显子仅出现在具有变体并且没有其他GTE_x的个体中。如果模型预测损失的剪接位点未在GENCODE中标注或从未在没有变体的GTE_x个体中观察到,我们则排除预期损失。如果在GENCODE中已经标注了预期获得的剪接位点,则我们也排除了预期增益。为了将这一分析扩展到常见的变体(存在于两个到四个个体中),我们还验证了有变体的个体中至少一半存在,并且在没有变体的所有个体中均不存在的新接点。

[0709] 我们可以通过使用预期的异常剪接事件对于具有变体的个体是私有的这一要求来验证40%预期高分(Δ 得分0.5)受体和供体增益,但仅可以验证3.4%的预测高分损失和5.6%的必需GT或AG的破坏情况(基于置换的错误确认率为<0.2%-请参阅“估计错误确认率”部分)。损益确认率差异的原因有两个。首先,与增益不同,外显子跳跃事件很少对具有变体的个体来说是完全私有的,因为外显子通常在较低的基线水平被跳过,这可以在足够深的RNA测序中观察到。其次,除了增加外显子跳跃以外,剪接位点的损失还可能具有其他影响,例如增加内含子保留或增加其他次优剪接位点的使用。基于这些原因,我们并不能完全依靠使用新的私有接点来验证模型的预测情况,我们还要基于定量证据对变体进行了验证,以增加或减少预计会受到影响具有变体的个体的接点使用量。

[0710] 通过定量标准验证预测的隐秘剪接突变

[0711] 对于样本s中的接点j,我们获得了归一化的接点数 c_{js} :

$$[0712] \quad c_{js} = \operatorname{asinh} \left(\frac{r_{js}}{\sum_g r_{gs}} \right) \quad (1)$$

[0713] 在此, r_{js} 是样本中的接点j的原始接点计数,并且分母中的总和取代带注释的受体和与j相同的基因的供体之间的所有其他接点(使用GENCODE v19的注释)。 asinh 变换定义为 $\operatorname{asinh}(x) = \ln(x + \sqrt{x^2 + 1})$ 。它类似于通常用于转换RNA测序数据的对数转换(Lonsdale等人,2013),但是将其定义为0,从而消除了对伪计数的需求,因为许多接点尤其是新的接点计数低或为零,这会导致伪计数严重失真。对于大值, asinh 变换的行为类似于对数变换,但对于小值,其接近线性。因此,它经常用在具有大量接近零值的数据集(例如RNA-seq或ChIP-seq数据集)中,以防止少数大值主导信号(Azad等,2016;Herring等,2018;Hoffman等,2012;Kasowski等,2013;SEQC/MAQC-III Consortium,2014)。如下所述,在“验证的考虑标准”部分中,将方程(1)中的分母小于200的样品排除在所有验证分析之外,从而避免了数值问题。

[0714] 对于预测由一组个体I中出现的由SNV引起的每个增加或失去的接点j,我们分别

计算每个组织t中的以下Z得分：

$$[0715] \quad z_{jt} = \frac{\text{mean}_{s \in A_t}(c_{js}) - \text{mean}_{s' \in U_t}(c_{js'})}{\text{std}_{s' \in U_t}(c_{js'})} \quad (2)$$

[0716] 其中 A_t 是来自组织t中I中的个体的样本集,而 U_t 是来自组织t中所有其他个体的样本集。请注意,对于相同的个体和组织,GTE_x数据集中可能有多个样本。如前所述, c_{js} 是样本中的接点j的计数。对于预测的损失,我们还为跳过推定的外显子的接点k计算了相似的z得分：

$$[0717] \quad z_{kt} = \frac{\text{mean}_{s' \in U_t}(c_{ks'}) - \text{mean}_{s \in A_t}(c_{ks})}{\text{std}_{s' \in U_t}(c_{ks'})} \quad (3)$$

[0718] 注意,引起跳跃的损失会导致丢失的接点相对减少,而跳跃会相对增加。这证明了将 z_{jt} 和 z_{kt} 分子的差值恢复情况,因此对于实际的剪接位点损失,这两个得分都趋于为负值。

[0719] 最后,我们计算了所有考虑的组织的Z得分的中位数。对于损失,我们分别根据方程式(2)和(3)计算了每个z得分的中位数。如果以下任一情况成立,则认为受体或供体的损失预测是有效的：

[0720] 1. 方程(2)中的z得分的中位数量化了接点的相对损耗,其小于置换数据中相应值的第五百分位值(-1.46),方程(3)中的z得分的中位数量化了跳跃的相对变化,其为非正数(零,负数或不存在,如果在任何个体中均未观察到跳跃接点的话)。换句话说,有强有力的证据表明减少了受影响的接点的使用,但是没有证据表明受影响的个体的跳跃有所减少。

[0721] 2. 来自方程(3)的z得分的中位数小于置换数据中相应值的第五百分位值(-0.74),来自方程(3)的z得分的中位数为非正数。

[0722] 3. 方程(2)中的z得分的中位数小于置换数据中相应值的第一个百分位数(-2.54)。

[0723] 4. 方程(3)中的z得分的中位数小于置换数据中相应值的第一个百分位数(-4.08)。

[0724] 5. 至少有一半的具有变体的个体中观察到接点跳过受影响的外显子,但是其他个体没有观察到(如上文“基于私有剪接点的预测隐秘剪接突变的验证”部分所述)。

[0725] 在“估计错误的验证率”部分中描述了用于获得上述临界值的置换情况。

[0726] 根据经验,我们观察到,与增益相比,我们需要对损失采用更严格的验证标准,因为如“基于私有剪接点预测的隐秘剪接突变的验证”部分中所述,损失比增益会引起更大的混合效应。在私有SNV附近偶然观察到一个新接点的可能性很小,因此即使是接点的微小证据也足以进行验证。相反,大多数预测的损失会减弱现有接点,并且这种减弱比增益引起的变化更难检测,并且更可能归因于RNA测序数据中的噪声值。

[0727] 验证分析的纳入标准

[0728] 为了避免在计数较低或覆盖率较差的情况下计算z得分,我们使用以下标准来过滤变体以进行验证分析：

[0729] 1. 仅当样本表达了基因(等式(1)中($\sum_g r_{gs} > 200$))时,才考虑将其用于上述z得分计算。

[0730] 2. 如果没有变体的个体中丢失或“参考”接点的平均计数分别小于10,则不考虑将

组织用于损失或增益 z 得分计算。“参考”接点是基于GENCODE注释在新的接点前使用的规范接点(有关详细信息,请参见效果大小计算部分)。直觉是,我们不应尝试验证影响控制对照个体中未表达的接点的剪接变体。同样地,如果对照个体不能充分表达跨越受影响位点的转录本,我们也不应尝试验证剪接变体。

[0731] 3. 在预测的剪接位点丢失的情况下,仅当样本中至少有10个丢失的接点计数时,才考虑从没有变体的个体获得的样本。在预期的受体或供体增益的情况下,仅当对照个体的“参考”接点计数至少为10时才考虑对照个体的样本。直觉是,即使在受影响的接点平均表达量大(即通过标准2)的组织中,不同的样品也可能具有截然不同的测序深度,因此仅应包括表达量足够的对照样品。

[0732] 4. 仅当至少有一个样本符合具有变体的个体的上述标准,以及至少5个样本符合至少2个不同的对照个体的标准时,才考虑这组织。

[0733] 没有满足上述标准的组织的变体被认为是不可确定的,并且在计算确认率时被排除在外。对于拼接增益变体,我们过滤了已经存在于GENCODE标注的剪接位点处的那些变体。类似地,对于拼接损耗的变体,我们仅考虑那些会降低现有GENCODE注释的剪接位点得分的变体。总体而言,分别认为55%和44%的高分(Δ 得分 ≥ 0.5)预测的收益和损失是可确定的,并可用于验证分析。

[0734] 估计错误的验证率

[0735] 为了确保上述程序具有合理的真实验证率,我们首先查看出现在1-4个GTEx个体中并破坏必需的GT/AG二核苷酸的SNV。我们认为,此类突变几乎肯定会影响剪接,因此其验证率应接近100%。在上述破坏中,根据上述标准可确定为39%,在可确定的之中,确认率为81%。为了估计错误的验证率,我们对SNV数据的个体标签进行了置换。对于出现在 k 个GTEx个体中的每个SNV,我们选择了 k 个GTEx个体的随机子集,并将SNV分配给它们。我们创建了10个这样的随机数据集,并对它们进行重复验证。置换的数据集中的验证率为:增益为1.7-2.1%,损失为4.3-6.9%,中位数分别为1.8%和5.7%。损失的错误确认率较高,而破坏的确认率较低,正如“基于私有剪接点预测的隐秘剪接突变的验证”一节所强调的那样,这是由于难以验证拼接位点的损失引起的。

[0736] 计算RNA测序数据中隐秘剪接变体的效应大小

[0737] 我们将变体的“效应大小”定义为由于变体而改变剪接模式的受影响基因转录本的得分(例如,转换为新受体或供体的得分)。作为预测的拼接增益变体的参考示例,考虑图38C中的变体。对于预测的获得的供体A,我们首先确定最接近注释的受体C的接点(AC)。然后我们确定“参考”接点(BC),其中 $B \neq A$ 是最接近A的注释的供体。然后,我们在每个样本 s 中计算新接点(AC)与参考接点(BC)的相对使用量:

$$[0738] \quad u_{(AB)s} = \frac{r_{(AC)s}}{r_{(AC)s} + r_{(BC)s}} \quad (4)$$

[0739] 在此, $r_{(AC)s}$ 是样本 s 中接点(AC)的原始解释计数。对于每个组织,我们计算了带有变体的个体与所有其他个体之间的接点(AC)使用的变化:

$$[0740] \quad \text{mean}_{s \in A_t} u_{(AC)s} - \text{mean}_{s' \in U_t} u_{(AC)s'} \quad (5)$$

[0741] 其中 A_t 是来自组织 t 中具有变体的个体的样本集,而 U_t 是来自组织 t 中其他个体的样本集。最终效应大小计算为所有组织中的上述差值的中位数。在获得受体的情况下或在

产生可变体的剪接位点是内含子的情况下,计算是相似的。效应大小计算的简化形式(假设来自具有和不具有变体的个体的单个样本)在图38C中示出。

[0742] 对于预测的损失,我们首先计算了跳过受影响外显子的转录本的分。计算如图45中说明。为了预测供体C的损失,我们确定了下一个下游带注释的外显子的接点(CE),以及从上游外显子到假定受影响的一个外显子的接点(AB)。我们对跳过受影响外显子的转录本的比例进行了量化,如下所示:

$$[0743] \quad k_{(AE)s} = \frac{r_{(AE)s}}{r_{(AE)s} + \text{mean}(r_{(AB)s} + r_{(CE)s})} \quad (6)$$

[0744] 至于增益,我们计算了具有变体的个体的样本与没有变体的个体的样本之间的跳跃分数的变化:

$$[0745] \quad \text{mean}_{s \in A_t} k_{(AE)s} - \text{mean}_{s' \in U_t} k_{(AE)s'} \quad (7)$$

[0746] 如上所计算的跳过的转录本的分不能完全包含受体或供体损失的影响,因为这样的破坏也可能导致内含子保留水平的提高或使用次优的剪接位点。为了解决其中一些影响,我们还计算了相对于具有相同受体E的其他接点的丢失接点(CE)的使用情况:

$$[0747] \quad l_{(CE)s} = \frac{r_{(CE)s}}{\sum r_{(E)s}} \quad (8)$$

[0748] 此处, $\sum r_{(E)s}$ 是任何(带注释的或新的)受体与供体E的所有接点的总和。这包括受影响的接点(CE),跳跃的接点(AE)以及补偿了C的损失其他次优供体的潜在接点,如图45所示实例。然后,我们计算了受影响的接点的相对使用的变化:

$$[0749] \quad \text{mean}_{s' \in U_t} l_{(CE)s'} - \text{mean}_{s \in A_t} l_{(CE)s} \quad (9)$$

[0750] 注意,与(5)和(7)不同,它们测量具有变体的个体中获得的或跳过的接点的使用增加量,在(9)中,我们要测量丢失的接点使用下降量,因此还原了两个部分的差异。对于每个组织,效应大小计算为(7)和(9)中的最大值。至于增益,变体的最终效应大小是整个组织的效应大小的中值。

[0751] 效应大小分析的纳入标准

[0752] 仅当根据上一节中所述的标准认为变体被验证为有效时,才考虑对变体的效应大小进行计算。为了避免计算非常小的异常转录本的分,我们仅考虑异常和参考接点计数均至少为10的样本。因为大多数隐秘剪接变体在内含子中,所以效应大小不能直接通过计算与变体重叠的参考和替代解读的数量来计算。因此,损失的效应大小是根据正常剪接接点的相对使用量的减少间接计算得出的。对于新接点增益的效应大小,异常转录本可能会受到无义介导的降解的影响,从而削弱观察到的效应大小。尽管这些测量存在局限性,但在增益和损失事件中,对于得分较低的隐秘剪接变体,我们观察到一致的趋势,即效应大小较小。

[0753] 完全外显性杂合私有SNV的预期效应大小

[0754] 对于完全外显的剪接位点创建变体,该变体导致带有变体的个体的变体单倍型的所有转录本都切换到新接点,并假定在对照个体中不出现新接点,那么根据方程式(5),预期的效应大小将为0.5。

[0755] 同样,如果杂合SNV引起新的外显子跳跃事件,并且受影响单倍型的所有转录本都切换到跳跃接点,则方程式(7)中的预期效应大小为0.5。如果来自具有变体的个体的所有

转录本都切换到不同的接点(跳跃接点或另一补偿点),则在方程式(8)中,来自具有变体的个体的样本的比率为0.5,来自其他样本的比率为1,因此方程式(9)中的差值为0.5。假定没有变体的个体中没有受体E的跳跃或其他接点。它还假定剪接位点破坏不会触发内含子保留。在实践中,至少低水平的内含子保留经常与剪接位点破坏有关。此外,即使在没有可变剪接变体的情况下,外显子跳跃也非常广泛。这就解释了为何对于即使破坏了必要GT/AG二核苷酸的变体来说,其测得的效应大小也都低于0.5。

[0756] 对于预期完全外显的杂合变体的效应大小为0.5而言,我们还假定了该变体未触发无义介导的降解(NMD)。在存在NMD的情况下,方程式(4),(6)和(8)的分子和分母都将下降,从而减小了观察到的效应大小。

[0757] 通过无义介导降解的转录本的分数的

[0758] 对于图38C,由于变体是外显子,我们可以算出跨过变体并具有参考或替代等位基因(分别为“Ref(无剪接)”和“A1t(无剪接)”)的解读数。我们还计算了新的剪接位点上剪接的解读数,以及可能带有替代等位基因(“A1t(新接点)”)的解读数。在图38C的示例以及其他观察到的情况下,我们观察到来自具有交替等位基因的单倍型的解读的总数(“A1t(无剪接)”和“A1t(新接点)”的总和)小于来自参考等位基因的解读数(“Ref(无剪接)”)。由于我们相信我们已在解读映射时消除了参考偏差,通过映射到参考和替代单倍型,并假设解读的数量与每个等位基因的转录本的数量成正比,我们期望参考等位基因会占变体基因座解读数的一半。我们假设“缺失”的替代等位基因解读对应于来自替代等位基因单倍型的转录本,该转录本在新接点处剪接并通过无义介导的降解(NMD)进行降解。我们称这个组为“A1t(NMD)”。

[0759] 为了确定观察到的参考数和替代解读数之间的差异是否显著,我们采用在二项式分布下成功概率为0.5和A1t(无剪接)+A1t(新接点)+Ref(无剪接)的试验总数来计算A1t(无剪接)+A1t(新接点)(或少量)解读的概率。这是一个保守的p值,因为我们没有通过计算可能降解了的转录本,这低估了“试验”的总数。图38C中NMD转录本的分数的计算为占新接点(A1t(NMD)+A1t(新接点))处剪接的解读总数中的“A1t(NMD)”解读的数量。

[0760] 网络在检测隐秘剪接接点时的灵敏度

[0761] 为了评估SpliceNet模型的灵敏度(图38F),我们使用了距离受影响的剪接位点(即新的或受破坏的受体或供体)20nt以内且未与注释的外显子必需GT/AG二核苷酸重叠的SNV,且估计效应大小至少为0.3(请参阅“效应大小计算”一节)。在所有灵敏度图中,如果SNV与带注释的外显子重叠或在带注释的外显子边界的50nt之内,则被定义为“附近外显子”。所有其他SNV都被认为是“深度内含子”。使用强烈支持的隐秘剪接位点的真实数据集,我们在变化的 Δ 得分阈值下评估了我们的模型,并报告了该临界值模型预测的真实数据集中的隐秘剪接位点的比例。

[0762] 与现有剪接预测模型的比较

[0763] 我们就各种指标对SpliceNet-10k,MaxEntScan(Yeo和Burge,2004),GeneSplicer(Pertea等,2001)和NNSplice(Reese等,1997)进行了正面对比。我们分别从<http://genes.mit.edu/burgelab/maxent/download/>和<http://www.cs.jhu.edu/~genomics/GeneSplicer/>下载了MaxEntScan和GeneSplicer软件。NNSplice不能作为可下载的软件使用,因此我们从http://www.fruitfly.org/data/seq_tools/datasets/Human/GENIE_96/

spliceassets/下载了训练和测试集,并且训练了具有最佳性能架构的模型(Reese et al, 1997)。作为完整性检查,我们复制了(Reese et al, 1997)中报告的测试集指标。

[0764] 为了评估这些算法的Topk准确度和精度-召回率曲线下的面积,我们对每种算法在测试集基因和lincRNA中的所有位置进行了评分(图37D)。

[0765] MaxEntScan和GeneSplicer输出对应于对数优势比,而NNSplice和SpliceNet-10k输出对应于概率。为了确保我们给MaxEntScan和GeneSplicer最佳机会,我们使用它们以及默认输出以及转换后的输出(我们首先转换其输出以使其与概率相对应)的 Δ 得分进行计算。更准确地说,MaxEntScan的默认输出对应于

$$[0766] \quad x = \log_2 \frac{p(\text{剪接位点})}{p(\text{非剪接位点})}$$

[0767] 在进行变换 $\frac{2^x}{2^x+1}$ 之后,它对应于所需的数量。我们对GeneSplicer软件进行了两次编译,一次是将RETURN_TRUE_PROB标志设置为0,一次是将其设置为1。我们选择了对RNA测序数据的验证率最高的输出策略(MaxEntScan:转换输出, GeneSplicer:默认输出)。

[0768] 为了比较各种算法的确认率和灵敏度(图38G),我们发现了所有算法预测全基因组相同数目的增益和损失的临界值。也就是说,对于SpliceNet-10k得分的每个临界值,我们发现每个竞争算法在该临界值处做出与SpliceNet-10k相同数量的增益预测和相同数量的损耗预测。选择的临界值在表S2中示出。

[0769] 单例变体与常见变体预测比较

[0770] 我们分别对出现在2-4个GTEx个体中的单例SNV和SNV(图46A, 46B和46C)进行了验证和灵敏度分析(如“灵敏度分析”和“模型预测的验证”中所述)。为了检验单例和常见变体之间的确认率是否存在显著差异,我们进行了Fisher精确检验,比较了每个 Δ 得分组(0.2-0.35、0.35-0.5、0.5-0.8、0.8-1)的确认率以及每个预测的效果(受体或供体的增益或损失)。经过Bonferroni校正以进行16个测试后,所有P值均大于0.05。我们同样检测了单例或常见变体的灵敏度,并进行了比较。我们使用Fisher精确检验来检验两组变体之间的验证率是否存在显著差异。我们分别考虑了深部内含子变体和外显子附近的变体,并对两个测试进行了Bonferroni校正。使用0.05临界值时,所有P值均不显著。因此,我们将单例和常用GTEx变体组合在一起,并结合图48A, 48B, 48C, 48D, 48E, 48F, 和48G以及图39A, 39B, 和39C一起考虑进行分析。

[0771] 对训练和测试染色体上的变体预测的比较

[0772] 我们比较了训练期间使用的染色体上的变体与其余染色体上的变体之间的RNA测序验证率和SpliceNet-10k的灵敏度(图48A和48B)。Bonferroni校正后,所有P值均大于0.05。我们还分别计算了训练和测试染色体上的变体的有害变体的分数,如“有害变体的分数”部分所述(图48C)。对于每个 Δ 得分组和每种类型的变体,我们使用Fisher精确检验比较了训练和测试染色体之间的常见和稀有变体的数量。经过Bonferroni校正12次测试后,所有P值均大于0.05。最后,我们按照“每个群组的新生突变的富集”部分所述,计算了训练和测试染色体上的隐秘剪接新生变体的数量(图48D)。

[0773] 不同类型隐秘剪接变体的变体预测的比较

[0774] 我们将预测的剪接位点创建变体分为三类:创建新的GT或AG剪接二核苷酸的变体,变体与其余剪接基序重叠的变体(外显子-内含子边界附近的位置最多3nt插入外显子

和8nt插入内含子),以及剪接基序之外的变体(图47A和47B)。对于每个得分组(0.2-0.35、0.35-0.5、0.5-0.8、0.8-1),我们进行了 χ^2 检验,以检验在三种剪接位点创建的变体的验证率是相同的这一假说。甚至在进行多个假设校正之前,所有测试的P值均大于0.3。为了比较三种类型的剪接位点创建变体的效应大小分布,我们使用了Mann-Whitney U检验,并对每个 Δ 得分组比较了所有三对变体类型(总计 $4 \times 3 = 12$ 个测试)。经过Bonferroni校正12次测试后,所有P值均 >0.3 。

[0775] 组织特异性剪接增益变体的检测

[0776] 对于图39C,我们想测试在表达受影响基因的组织中新接点的使用率是否一致。我们关注的是创造了新的私有剪接位点的SNV,即导致获得的剪接点的SNV,该接点仅出现在至少一半具有变体的个体中,而没有其他个体出现。对于每个这样的新接点j,我们在每个组织t中计算组织中具有变体的个体的所有样本中的接点总数: $\sum_{s \in A_t} r_{js} \cdot He$ 。这里 A_t 是来自组织t的具有变体的个体的样本集。同样,对于相同的样本 $\sum_{s \in A_t} \sum_g r_{gs}$,我们计算了该基因的所有带注释的接点的总数,其中g表明了该基因的带注释的接点。组织t中新接点的相对使用量,在进行该基因的本底计数归一化后,可以测量为:

$$[0777] \quad m_t = \frac{\sum_{s \in A_t} r_{js}}{\sum_{s \in A_t} (r_{js} + \sum_g r_{gs})}$$

[0778] 我们还计算了跨组织的接点的平均用量:

$$[0779] \quad m = \frac{\sum_t \sum_{s \in A_t} r_{js}}{\sum_t \sum_{s \in A_t} (r_{js} + \sum_g r_{gs})}$$

[0780] 我们想检验以下假设:接点的相对使用在组织之间是均匀的,等于m。因此,我们进行了 χ^2 检验,将观察到的组织计数 $\sum_{s \in A_t} r_{js}$ 与假设均匀率的预期计数进行比较 $m \sum_{s \in A_t} (r_{js} + \sum_g r_{gs})$ 。如果Bonferroni校正的 χ^2 p值小于 10^{-2} ,则认为形成剪接位点的变体是组织特异性的。测试的自由度为T-1,其中T是要考虑的组织数。测试中仅使用满足验证部分中所述标准的组织。此外,为避免计数较低的情况,如果均匀性测试的功能不足,我们仅对均匀变体进行了测试,该变体具有至少三个考虑过的组织,平均每个组织至少具有一个异常解读(即 $m > 1$),并且至少在所有考虑的组织中总共有15个异常解释(即 $\sum_t \sum_{s \in A_t} r_{js} > 15$)。我们忽略了所有 Δ 得分小于0.35的变体,因为此类变体的效应大小通常较低,而接点数却很少。我们观察到该类别的组织特异性变体的比例非常低,但是我们认为这是由于统计功效问题引起的。

[0781] III. 对ExAC和gnomAD数据集进行分析

[0782] 变体过滤

[0783] 我们从ExAC浏览器(Lek等人,2016)下载了Sites VCF版本0.3文件(60,706个外显子组),并从gnomAD浏览器中下载了Sites VCF版本2.0.1文件(15,496个全基因组)。我们从中筛选出了变体列表,以评估SpliceNet-10k。尤其我们要考虑满足以下条件的变体:

[0784] • FILTER字段为PASS。

[0785] • 变体是一个单核苷酸变体,只有一个替代核苷酸。

[0786] • AN字段(称为基因型的等位基因总数)的值至少为10,000。

[0787] • 变体位于规范的GENCODE转录本的转录起点和终点之间。

[0788] 总共7,615,051和73,099,995个变体分别通过了ExAC和gnomAD数据集中的这些过滤器。

[0789] 有害变体的分数

[0790] 对于该分析,我们仅考虑了ExAC和gnomAD过滤列表中的那些变体,这些变体在同类人群中是单例或常见的(等位基因频率(AF)0.01%)。我们根据GENCODE规范注释,基于它们的基因组位置对这些变体进行了子分类:

[0791] • 外显子的:该组由同义的ExAC变体(676,594个单例和66,524个常见变体)组成。这里没有考虑错义变体,以确保该组中大多数变体的有害性是由于剪接变化引起的。

[0792] • 附近内含子:这组包含内含子ExAC变体,位于规范外显子边界3到50nt之间。更准确地说,在分析受体增益/损失和供体增益/损失变体时,仅考虑分别来自剪接受体和供体的3-50nt的那些变体(575,636个单例和48,362个共同变体的受体增益/损失,567,774个单例和50,614个共同变体的供体增益/损失)。

[0793] • 深度内含子:该组由内含子的gnomAD变体组成,它们距规范外显子边界超过50nt(共34,150,431个单例和8,215,361个共同变体)。

[0794] 对于每个变体,我们使用SpliceNet-10k计算了四种接头类型的 Δ 得分。然后,对于每种剪接类型,我们构造了一个 2×2 卡方列联表,其中两行对应于预测的可变剪接变体(Δ 得分在剪接类型的范围内)与预测的不可变剪接变体(所有剪接类型的 Δ 得分均小于0.1),并且这两列对应于单例与常见变体。对于拼接增益变体,我们过滤了已经存在于GENCODE标注的剪接位点处的那些变体。类似地,对于拼接损耗的变体,我们仅考虑那些会降低现有GENCODE注释的剪接位点得分的那些变体。然后计算比值比,将有害变体的分数估算为

$$[0795] \quad \left(1 - \frac{1}{\text{比值比}}\right) \times 100\%$$

[0796] ExAC和gnomAD过滤列表中的截短蛋白的变体鉴定如下:

[0797] • 无义变体:VEP(McLaren等人,2016)的结果是“stop_gained”(ExAC中有44,046个单例和722个常见变体,gnomAD中有20,660个单例和970个常见变体)。

[0798] • 移码变体:VEP结果是“移码变体”。为了创建该组,放宽了在变体过滤过程中的单核苷酸变体标准(在ExAC中48,265个单例和896个常见变体,在gnomAD中30,342个单例和1,472个常见变体)。

[0799] • 必需的受体/供体损失:变体位于规范内含子的前两个位置或最后两个位置(ExAC中29,240个单例和481个常见变体,gnomAD中12,387个单例和746个常见变体)。

[0800] 为ExAC和gnomAD过滤列表构建了蛋白质截断变体的 2×2 卡方平方列联表,用于估计有害变体的比例。在这里,两行对应于蛋白质截断与同义变体,两列对应于单例与常见变体。

[0801] ExAC(外显子和近内含子)和gnomAD(深部内含子)变体的结果分别展示于图40B和图40D中。

[0802] 移码与帧内剪接增益

[0803] 在此分析中,我们将注意力集中在ExAC变体上,它们是外显子(仅同义突变)或接近内含子,并且在该群组中是单例或常见变体(AF 0.01%)。为了将受体增益变体分类为帧

内或移码变体,我们测量了规范的剪接受体与新创建的剪接受体之间的距离,并检查它是否为3的倍数。我们通过测量规范拼接供体和新创建的拼接供体之间的距离,对供体增益变体进行了类似的分类。

[0804] 从 2×2 卡方列联表中估算有害的帧内拼接增益变体的比例,其中两行对应于预测的帧内拼接增益变体(对于受体或供体增益, Δ 得分0.8)与预测没有可变剪接变体(所有剪接类型的 Δ 得分 <0.1),并且两列对应于单例与常见变体。通过将列联表中的第一行替换为预测的移码剪接增益变体,对移码剪接增益变体重复此过程。

[0805] 为了计算图40C中所示的p值,我们仅使用预测的拼接剪接变体来构建 2×2 卡方列联表。在此,两行对应于帧内与移码剪接增益变体,两列与以前一样对应于单例与常见变体。

[0806] 每个个体隐秘剪接变体的数量

[0807] 为了估计每个个体的稀有功能性隐秘剪接变体的数量(图40E),我们首先将每个gnomAD变体包括在每个等位基因中,从而模拟100个gnomAD个体,其概率等于其等位基因频率。换句话说,每个个体的每个变体独立采样两次,以模拟二倍体。我们计算了每个人的稀有($AF < 0.1\%$)外显子(仅同义突变),近内含子和深部内含子变体的数量,其 Δ 得分分别大于或等于0.2、0.2和0.5。这些是相对宽松的 Δ 得分阈值,可在确保至少40%的预测变体有害的同时优化灵敏度。在这些临界值下,每人平均获得7.92个同义突变/近内含子和3.03个深度内含子稀有隐秘剪接变体。因为并非所有这些变体都具有功能,所以我们将计数乘以在这些临界值处有害的变体的分数。

[0808] IV. 对DDD和ASD数据集进行分析

[0809] 隐秘剪接新生突变

[0810] 我们获得了已发布的新生突变(DNM)。其中包括3953名患有自闭症谱系障碍的先证者(Dong等人,2014;Iossifov等人,2014;De Rubeis等人,2014),来自解码发育异常症群组的4293名先证者(McRae等人,2017)和2073健康对照(Iossifov等,2014)。低质量DNM被排除在分析之外(ASD和健康对照:置信度=低置信度,DDD:PP(DNM) <0.00781 , (McRae et al, 2017))。使用网络对DNM进行了评估,我们根据具体情况,使用 Δ 得分(请参见上面的方法)对隐秘剪接突变进行了分类。我们仅考虑了以VEP后果注释的突变,即同义突变_变体,剪接_区域_变体,内含子_变体,5_prime_UTR_变体,3_prime_UTR_变体或无义_变体。在图41A,41B,41C,41D,41E,41F,50A和50B中我们使用了 Δ 得分 >0.1 的位点,在图49A,49B和49C中我们使用了 Δ 得分 >0.2 的位点。

[0811] 图20、21、22、23和24展示了SpliceNet-80nt,SpliceNet-400nt,SpliceNet-2k和SpliceNet-10k体系结构的详细说明。四种体系结构分别在靶位置的每一侧使用长度分别为40、200、1,000和5,000的侧翼核苷酸序列作为输入,并输出该位置是剪接受体,剪接供体和两者都不是的概率。该体系结构主要由卷积层(N,W,D)组成,其中N,W和D分别是卷积层中卷积内核的数量,每个卷积内核的窗口尺寸和膨胀率。

[0812] 图42A和42B描述了对lincRNA的各种剪接预测算法的评估。图42A显示了在lincRNA上评估时,各种剪接预测算法的top-k准确度和精度-召回率曲线下面积。图42B显示了使用MaxEntScan和SpliceNet-10k评分的LINC00467基因的完整前体mRNA转录本,以及预测的受体(红色箭头)和供体(绿色箭头)位点以及外显子的实际位置。

[0813] 图43A和43B示出了TACTAAC分支点和GAAGAA外显子剪接增强子基序的位置依赖性效应。如图43A所示,对14,289个测试集剪接受体中的每一个,在不同距离处引入最佳分支点序列TACTAAC,并使用SpliceNet-10k计算出受体得分。将预测的受体得分的平均变化绘制为距剪接受体的距离的函数。当距剪接受体的距离在20到45nt之间时,预测得分会增加。在小于20nt的距离处,TACTAAC会破坏多聚嘧啶序列,因此预测的受体得分非常低。

[0814] 在图43B中,将SR蛋白六聚体基序GAAGAA类似地引入到距14,289个测试集剪接受体和供体的每一个的不同的距离处。将预测的SpliceNet-10k受体和供体得分的平均变化分别绘制为距剪接受体和供体的距离的函数。当基序在外显子侧且距剪接位点小于~50 nt时,预测得分会增加。在距外显子的较大距离处,GAAGAA基序趋向于不利于使用所考虑的剪接受体或供体位点,大概是因为它现在优先支持更近端的受体或供体基序。当GAAGAA放置在非常靠近内含子的位置时,受体和供体的得分非常低,这是由于扩展的受体或供体剪接基序受到破坏所致。

[0815] 图44A和44B描绘了核小体定位对剪接的影响。在图44A中,在100万个随机选择的基因间位置处,引入了间隔150nt的强受体和供体基序,并使用SpliceNet-10k计算了外显子增加的概率。为了显示SpliceNet-10k预测与核小体定位之间的相关性与GC成分无关,基于其GC含量对位置进行了分类(使用引入的剪接位点之间的150个核苷酸计算得出),并且对每个分类,对SpliceNet-10k预测和核小体信号间的Spearman相关性进行了作图。

[0816] 在图44B中,使用SpliceNet-80nt(称为局部基序评分)和SpliceNet-10k对测试集中的剪接受体和供体位点进行了评分,并将这些得分绘制为核小体富集的函数。核小体富集的计算方法是:在剪接位点外显子侧的50nt处的平均核小体信号除以在剪接位点内含子侧的50nt处的平均核小体信号。SpliceNet-80nt得分是基序强度的替代指标,与核小体富集呈负相关,而SpliceNet-10k得分与核小体富集呈正相关。这表明核小体定位是可以补偿微弱局部剪接基序的一个远程特异性决定因素。

[0817] 图45示出了计算具有复杂效果的剪接破坏变体的效果大小的示例。内含子变体chr9:386429A>G破坏了正常供体位点(C),并激活了先前被抑制的内含子下游供体(D)。显示的是具有变体的个体和对照个体全血中RNA测序的覆盖率和接点解读计数。具有变体的个体和对照个体的供体位点分别用蓝色和灰色箭头标记。红色粗体字母对应于接点端点。为了更明显,与内含子长度相比,外显子长度被夸大了四倍。为了估计效应大小,相对于具有相同供体E的所有其他接点,我们计算了外显子跳跃接点(AE)使用增加量和破坏接点(CE)使用的减少量。最终效应大小是两个值的最大值(0.39)。突变样品中还存在内含子保留量的增加。这些可变效应在外显子跳跃事件中很常见,并增加了验证稀有变体的复杂性,这些变体预计会导致受体或供体位点丢失。

[0818] 图46A,46B和46C示出了在单例和常见变体中对SpliceNet-10k模型的评估。图46A是SpliceNet-10k预测的隐秘剪接突变的一部分,该隐秘剪接突变针对GTEx RNA测序数据进行了验证。对GTEx群组中出现在至多四个个体中的所有变体进行了模型评估。对具有预测的可变剪接效应的变体,已针对RNA测序数据进行了验证。对于出现在单个GTEx个体中的变体(左)和出现在两到四个GTEx个体中的变体(右),分别显示了验证率。预测按其 Δ 得分分组。我们比较了每个 Δ 得分组中四种类别的变体中的单例与常见变体之间的确认率(受体或供体的增益或损失)。差异不显著($P>0.05$,采用Bonferroni校正的Fisher精确检验,进

行了16次检验)。

[0819] 在图46B中,SpliceNet-10k在不同的 Δ 得分临界值下检测GTEx群组中的可变剪接变体的灵敏度。图中分别显示了单例(左)和常见变体(右)的模型灵敏度。对于外显子附近的变体或深部内含子变体,单例和常见变体在 Δ 得分的临界值为0.2时的灵敏度差异均不显著($P>0.05$,经Bonferroni校正后的Fisher Exact检验,用于两项检验)。

[0820] 图46C是经验证的单例和常见变体的 Δ 得分的分布。P值用于Mann-Whitney U检验,比较单例和常见变体的得分。常见变体的 Δ 得分明显较低,这是由于自然选择过滤了剪接破坏突变的巨大影响。

[0821] 图47A和47B描绘了剪接位点创建变体的验证率和效应大小,并按变体的位置进行划分。根据变体是否产生新的必需GT或AG剪接二核苷酸,是否与其余剪接基序重叠(外显子-内含子边界周围的3nt范围的所有位点插入到外显子,8nt范围的所有位点插入到内含子,不包括必需的二核苷酸),或者是否位于剪接基序之外,对预测的剪接位点创建变体进行分组。

[0822] 如图47A所示,显示了剪接位点创建变体的三个类别的每一个的确认率。条形上方显示了每个类别中的变体总数。在每个 Δ 得分组中,三组变体之间的验证率差异不显著($P>0.3$, χ^2 均匀性检验)。

[0823] 图47B中显示了剪接位点创建变体的三个类别的每一个的分布情况。在每个 Δ 得分组中,三组变体之间的效应大小差异不显著($P>0.3$,采用Bonferroni校正的Mann-Whitney U检验)。

[0824] 图48A,48B,49C和49D描绘了在训练和测试染色体上的SpliceNet-10k模型的评估。图48A是SpliceNet-10k模型预测的隐秘剪接突变的一部分,该模型已针对GTEx RNA测序数据进行了验证。图中分别显示了训练期间使用的染色体上的变体的验证率(除chr1, chr3, chr5, chr7和chr9以外的所有染色体;左图)和其余的染色体上的变体的验证率(右图)。预测按其 Δ 得分分组。我们比较了每个 Δ 得分组中四种类别的变体(受体或供体的增益或损失)在训练和测试染色体之间的验证率。这说明了在训练和测试染色体之间的预测 Δ 得分分布中的潜在差异。验证率的差异不显著($P>0.05$,采用Bonferroni校正的Fisher精确检验,进行了16次检验)。

[0825] 图48B显示了SpliceNet-10k在不同的 Δ 得分临界值下检测GTEx群组中的可变剪接变体的灵敏度。分别显示了用于训练的染色体(左)和其余染色体(右)上的变体的模型的灵敏度。我们使用Fisher精确检验比较了训练和测试染色体之间 Δ 得分临界值为0.2时模型的灵敏度。对于外显子附近的变体或深部内含子变体而言,差异均不显著(Bonferroni校正后两次检验的P值 >0.05)。

[0826] 图48C显示了ExAC数据集中的预测的同义突变和内含子隐秘剪接变体的部分是有害的,分别针对用于训练的染色体上的变体(左)和其余染色体上的变体(右)进行计算。图4A显示所计算分数和P值。我们比较了每个 Δ 得分组中四类变体(受体或供体的增益或损失)中训练染色体和测试染色体之间常见和稀有变体的数量。他们之间差异不显著($P>0.05$,采用Bonferroni校正的Fisher精确检验,进行了12次检验)。

[0827] 图48D显示用于DDD,ASD和对照组的人均预测的隐秘剪接接新生突变(DNM),分别针对所用染色体上的变体。

[0828] 用于训练的染色体(左)和其余的染色体(右)。误差线显示95%置信区间(CI)。对于测试集,每人的隐秘剪接新生变体的数量要少一些,因为它大约是训练集大小的一半。由于样本量小,数字比较嘈杂。

[0829] 图49A,49B和49C仅从同义的,内含子的或非翻译的区域位点显示了具有罕见遗传病的患者中的隐秘剪接新生突变。图49A显示了解密发育障碍群组(DDD),西门子单个孤独症儿童样本数据库和自闭症测序联盟的患有自闭症谱系障碍(ASD)的患者的隐秘剪接 Δ 得分 >0.2 /人,以及健康对照人群。显示了在健康对照上方的DDD和ASD群组中的富集,针对群组之间的变体确定进行了调整。误差线显示95%的置信区间。

[0830] 图49B显示了基于DDD和ASD群组的功能类别的致病性DNM的估计比例,以及基于与健康对照相比的每个类别的富集。调整了隐秘剪接的比例,以确保没有错义和深部内含子位点。

[0831] 图49C显示了与在不同 Δ 得分阈值下的健康对照相比,DDD和ASD群组中隐秘剪接DNM的富集和过量。隐秘剪接过量的调整是针对缺少错义和深部内含子位点而言的。

[0832] 图50A和50B显示了ASD中隐秘剪接新生突变,并作为致病性DNM的一部分。图50A显示了ASD先证者中的隐秘剪接DNM在不同的 Δ 得分阈值下的富集和过量,用于预测隐秘剪接位点。

[0833] 在图50B中,显示了通过使用不同的 Δ 得分阈值预测隐秘剪接位点,在所有类别的致病性DMN(包括蛋白质编码突变)中归因于隐秘剪接位点的致病性DMN的比例。更大范围的 Δ 得分阈值会增加比本底预期更高的隐秘剪接位点的数量,但要以较低的优势比为代价。

[0834] 图51A,51B,51C,51D,51E,51F,51G,51H,51I和51J描述了ASD患者中预测的隐秘剪接新生突变的RNA测序验证情况。从36个预测的隐秘剪接位点中选择RNA测序进行实验验证,验证RNA表达的覆盖率和剪接点数。对于每个样品,受影响的个体的RNA测序覆盖率和接点数显示在顶部,而没有突变的对照个体显示在底部。这些图按验证状态和剪接突变类型分组。

[0835] 图52A和52B示出了仅在规范转录本上训练的模型的验证率和对RNA测序的灵敏度。在图52A中,我们仅使用规范GENCODE转录本中的接点训练了SpliceNet-10k模型,并将该模型的性能与GTEx群组中至少五个个体中出现的规范接点和剪接接点的训练模型进行了比较。我们在每个 Δ 得分组中比较了二种模型对四种类别的变体(受体或供体的增益或损失)的验证率。两个模型之间的验证率差异不显著($P>0.05$,采用Bonferroni校正的Fisher精确检验,进行了16次检验)。

[0836] 图52B显示了在不同的 Δ 得分临界值下,检测GTEx群组中的可变剪接变体的规范接点的训练模型。该模型在深部内含子区域的灵敏度低于图2的模型($P<0.001$,采用Bonferroni校正的Fisher精确检验)。外显子附近的灵敏度没有显著差异。

[0837] 图53A,53B和53C显示了集成建模改善了SpliceNet-10k性能。图53A示出了5个个体SpliceNet-10k模型的top-k准确度和精度-召回率曲线下的面积。这些模型具有相同的架构,并使用相同的数据集进行了训练。但是,由于训练过程涉及各种随机方面,例如参数初始化,数据改组等,因此它们彼此不同。

[0838] 图53B中显示5个个体SpliceNet-10k模型的预测高度相关。在本研究中,我们仅考虑了测试集中分配给受体或供体得分大于或等于0.01的那些位置。通过将模型#i的预测与

模型#j的预测作图来构造子图(i,j)(相应的Pearson相关性显示在子图上方)。

[0839] 在图53C中,随着用于构建SpliceNet-10k集成的模型数量从1增加到5,模型性能提高。

[0840] 图54A和54B示出了在外显子密度变化的区域中SpliceNet-10k的评估。如图54A所示,根据10,000个核苷酸窗口中存在的规范外显子的数量,将测试集的位置分类为5个部分。对于每个部分,我们计算了SpliceNet-10k的top-k准确度和精度-召回率曲线下面积。

[0841] 图54B中,我们以MaxEntScan作为比较重复进行了分析。请注意,通过top-k准确度和精度-召回率曲线下面积来衡量,两个模型的性能在较高的外显子密度下都会提高,这是因为阳性测试用例的数量相对于阴性测试用例的数量增加了。

[0842] 每个群组中新生突变的富集

[0843] 在三个群组的每个群组中,都对候选对象隐秘剪接DNM进行了计数。DDD群组没有报告离外显子>8nt的内含子DNM,因此出于富集分析的目的,从所有群组中排除了离外显子>8nt的区域,以实现DDD和ASD群组之间的等效比较(图41A)。我们还进行了单独的分析,该分析排除了具有双重隐秘剪接和蛋白质编码功能影响的突变,以证明该富集不是由于受影响群组中具有蛋白编码作用的突变的富集而产生的(图49A,49B和49C)。通过使用健康对照群组作为基线,经过归一化群组之间每位个体的同义DNM的比率后,为了区分群组之间对DNM的确定度,进行计数。我们使用E检验比较了每个群组的隐秘剪接DNM的比率,以比较两个泊松比率(Krishnamoorthy和Thomson,2004年)。

[0844] 通过使用三核苷酸序列背景模型,按比例将所有离外显子9-50nt的隐秘剪接DNM的比例向上扩展,针对离外显子的>8nt的DNM缺乏的情况,调整了绘制的富集率和预期图(图41C)(参见下文,每个基因的新生突变的富集)。通过放大误义突变位点和同义突变位点下要发生的隐秘剪接位点的隐秘计数的比例,调整了缺乏误义突变位点处无声诊断比例和过量的隐秘位点(图49B和49C)。通过计算一系列临界值内DDD群组中隐秘剪接DNM的富集,可以评估 Δ 得分阈值对富集的影响。对于这些中的每一个,都将计算观察到的预期比值比,以及隐秘剪接DNMs的过量值。

[0845] 致病性DNM的比例

[0846] 与基线突变率相比,DNM过量可被视为群组中的致病性结果。我们估计了相对于健康对照群组的本底,在ASD和DDD群组中按功能类型的DNM的过量(图41B)。如上所述,将DNM计数归一化为每个个体的同义DNM比率。如上所述,针对缺少距内含子9-50nt的DNM来调整DDD隐秘剪接的计数。对于ASD和DDD群组,我们还使用来自阴性选择分析的近内含子(<50nt)与深部内含子(>50nt)隐秘剪接变体比率,来调整离外显子>50nt的深部内含子变体的缺失确定度(图38G)。

[0847] 每个基因的新生突变的富集

[0848] 我们使用三核苷酸序列背景模型确定了基因组中每个变体的无效突变率(Samocha等人,2014)。我们使用网络预测外显子内和内含子中最多8nt处所有可能的单核苷酸取代的 Δ 得分。基于无效突变率模型,我们获得了每个基因的隐秘剪接新生突变的预期数量(使用 Δ 得分>0.2作为临界值)。

[0849] 根据DDD研究(McRae et al,2017),在两种模型下评估了基因相对于偶然性的DNM富集程度,一种模型仅考虑蛋白截短(PTV)DNM,另一种模型考虑所有蛋白改变的DNM(PTV,

错义和帧内插入缺失)。对于每个基因,我们选择了最重要的模型,并针对多个假设检验调整了P值。在不考虑隐秘剪接DNM或隐秘剪接率(默认测试,在原始DDD研究中使用默认值)的情况下运行了这些检验一次,在我们考虑了隐秘剪接DNM或突变率的情况下,我们运行了这些检验一次。我们报告了其他的候选基因,这些基因当包含隐秘剪接DNM时,经FDR调整后的P值<0.01,但是当不包含隐秘剪接DNM时经FDR调整后的P值>0.01(默认测试)。对ASD群组进行了类似的富集检验。

[0850] 预测的隐秘剪接位点的验证

[0851] 我们从西门子单个孤独症儿童样本数据库的受影响先证者中选择了高信度的新生变体,它们在淋巴母细胞细胞系中至少RPKM>1RNA测序表达。我们基于剪接损失变体的 Δ 得分阈值>0.1和剪接增益变体的 Δ 得分阈值>0.5的标准,选择了新生隐秘剪接变体进行验证。因为需要提前采购细胞系,所以与我们在本文其他地方采用的阈值相比,这些阈值反映了我们的方法的早期迭代(图38G和图41A,41B,41C和41D),并且该网络不包含用于模型训练的GTEx新剪接接点。

[0852] 从SSC获得了这些先证者的淋巴母细胞系。在培养基(RPMI1640、2mM L-谷氨酰胺,15%胎牛血清)中培养细胞,使其最大细胞密度为 1×10^6 细胞/ml。当细胞达到最大密度时,通过上下吸移4或5次使细胞解离,并接种至密度为200,000-500,000活细胞/ml,从而使它们传代。细胞在37°C,5%CO₂条件下生长10天。然后分离约 5×10^5 个细胞,并在4°C以300×g离心5分钟。按照制造商的规程,使用RNeasy®Plus微型试剂盒(QIAGEN)提取RNA。使用安捷伦RNA 6000Nano试剂盒(安捷伦科技公司)评估RNA质量,并在Bioanalyzer 2100(安捷伦科技公司)上运行。RNA测序文库通过带有Ribo-Zero Gold Set A(Illumina)的TruSeq®链式总RNA文库制备试剂盒生成。使用150nt单一解读测序技术,在先进技术中心(UCSF)的HiSeq 4000仪器上对文库进行了测序,覆盖范围为270-388百万个解读数(中位数为358百万个解读数)。

[0853] 通过使用相应的替代等位基因替代患者的新生变体(Iossifov等人,2014),将每个患者的测序解释与OLego(Wu等人,2013)相对于hg19创建的参考进行了比对。测序覆盖率,剪接点的使用和转录本的位置用MISO的sashimi plot工具绘制(Katz等人,2010)。我们参照上述在模型预测的验证部分中评估了预测的隐秘剪接位点。确认了13个新的剪接位点(9个新接点,4个跳跃外显子),因为它们仅在包含隐秘剪接位点的样品中观察到,而在149个GTEx样品或其他35个测序样品中均未观察到。对于另外4个外显子跳跃事件,在GTEx中经常观察到低水平的外显子跳跃。在这些情况下,我们计算了使用跳跃接点的解读分数,并验证了与其他样本相比,该分数在包含隐秘剪接位点的样本中最高。根据其他样本中不存在或存在显著低的内含子保留率,我们对另外4个实例进行了验证。对照样本中适度的内含子保留使我们无法解决DDX11和WDR4中的事件。因为在测序解读中不存在变体,(在CSAD和GSAP中)的两个事件被归类为验证失败。

[0854] 可使用的数据和软件

[0855] 训练和测试数据、参考基因组中所有单核苷酸取代的预测得分、RNA测序验证结果、RNA测序接点和源代码公开发布在:

[0856] <https://basespace.illumina.com/s/5u6Th0blecrh>

[0857] 将36种淋巴母细胞系的RNA测序数据以登录号为E-MTAB-xxxxx保存在

ArrayExpress数据库中的EMBL-EBI (www.ebi.ac.uk/arrayexpress) 处。

[0858] 预测得分和源代码是在开放源代码修改的Apache License v2.0下公开发布的,可免费用于学术和非商业软件应用。为了减少该领域关注循环性问题,作者明确要求不要将该方法的预测得分整合到其他分类方法中,而是要求相关方直接使用提供的源代码和数据来训练和改进自己的深度学习模型。

[0859] 关键资源表

反应物或资源 存放资料	来源	标识符
RNA 测序数据和 GTEx 群组的变体识别	https://www.ncbi.nlm.nih.gov/projects/gap	dbGAP 登 录 : phs000424.v6.p1
自闭症患者和健康对照者的新变异	(Iossifov et al., 2014)	不适用
解密发育障碍群组中的新生突变	(McRae et al., 2017)	不适用
用于训练规范的 SpliceNet 模型的 GENCODE 主体转录本中的剪接点	本研究	https://basespace.illumina.com/s/5u6ThOblecrh
用于增强训练数据集的 GTEx 剪接点	本研究	https://basespace.illumina.com/s/5u6ThOblecrh
用于测试模型的 GENCODE 主体转录本中的剪接点,但不包括旁系同源基因	本研究	https://basespace.illumina.com/s/5u6ThOblecrh
用于测试模型的 lincRNA 剪接点	本研究	https://basespace.illumina.com/s/5u6ThOblecrh
规范模型的预测	本研究	https://basespace.illumina.com/s/5u6ThOblecrh
GTEx 补充模型的预测	本研究	https://basespace.illumina.com/s/5u6ThOblecrh
所有 GTEx v6.p1 示例中的所有 GTEx 接点	本研究	https://basespace.illumina.com/s/5u6ThOblecrh
Δ 得分 > 0.1 的经过验证的 GTEx 私有变体列表	本研究	https://basespace.illumina.com/s/5u6ThOblecrh
36 名自闭症患者 RNA 测序的 BAM 比对文件	本研究	ArrayExpress accession: E-MTAB-xxxx
软件和算法		
SpliceNet 源代码	本研究	https://basespace.illumina.com/s/5u6ThOblecrh

[0861] 附表标题

[0862] 表S1示出了用于演示效应大小计算和组织特有的剪接效应的GTEx样本。参见图38A, 38B, 38C, 38D, 38E, 38F, 38G, 39A, 9B和45。

[0863] 表S2显示了SpliceNet-10k, GeneSplicer, MaxEntScan和NNSplice的匹配置信度临界值, 在这些临界值下所有算法都预测了全基因组中相同数目的增益和损失。参见图

38G。

[0864] 表S3显示了每个群组中预测的隐秘剪接DNM的数量。参见图41A, 41B, 41C, 41D, 41E和41F和以下表：

[0865]	群组	先证者 (n)	每个先证 者的同义 新生突变	外显子+内含子高达 8 nt		内含子> 8 nt	
				未经调整 的	归一化 为同义 突变的	未经调整 的	归一化 为同义 突变的
	DDD	4293	0.28744	347	298.7	14	12.1
	ASD	3953	0.24462	236	238.7	64	64.7
	对照	2073	0.24747	98	98	20	20

[0866] 表S4示出了每个突变类别的每个基因的预期新生突变率。参见图41A, 41B, 41C, 41D, 41E和41F。

[0867] 表S5示出了用于DDD和ASD中的基因富集的p值。参见图41A, 41B, 41C, 41D, 41E和41F。

[0868] 表S6描绘了自闭症患者中36个预测的隐秘剪接DNM的验证结果。参见图41A, 41B, 41C, 41D, 41E和41F。

[0869] 计算机系统

[0870] 计算机系统

[0871] 图59是可用于实现所公开技术的计算机系统的简化框图。计算机系统通常包括至少一个处理器, 该处理器通过总线子系统与多个外围设备通信。这些外围设备可以包括存储子系统, 该存储子系统包括例如存储器设备和文件存储子系统、用户接口输入设备、用户接口输出设备和网络接口子系统。输入和输出设备允许用户与计算机系统交互。网络接口子系统提供至外部网络的接口, 包括至其他计算机系统中相应接口设备的接口。

[0872] 在一种实施方式中, 诸如ACNN和CNN之类的神经网络可通信地链接到存储子系统和用户接口输入设备。

[0873] 用户接口输入设备可以包括键盘; 定点设备, 例如鼠标、轨迹球、触摸板或图形输入板; 扫描仪; 结合到显示器中的触摸屏; 音频输入设备, 例如语音识别系统和麦克风; 以及其他类型的输入设备。通常, 使用术语“输入设备”旨在包括将信息输入计算机系统的所有可能类型的设备和方式。

[0874] 用户接口输出设备可以包括显示子系统、打印机、传真机或诸如音频输出设备的非视觉显示器。显示子系统可以包括阴极射线管(CRT)、诸如液晶显示器(LCD)的平板设备、投影设备或用于创建可视图像的一些其他机构。显示子系统还可以提供非视觉显示, 例如音频输出设备。通常, 使用术语“输出设备”旨在包括从计算机系统向用户或另一机器或计算机系统输出信息的所有可能类型的设备和方式。

[0875] 存储子系统存储编程和数据结构, 这些程序和数据结构提供本文描述的一些或所有模块和方法的功能。这些软件模块通常由处理器单独执行或与其他处理器结合执行。

[0876] 存储子系统中使用的存储器可以包括多个存储器, 包括用于在执行程序期间存储指令和数据的主随机存取存储器(RAM)和存储固定指令的只读存储器(ROM)。文件存储子系统可以为程序和数据文件提供持久存储, 并且可以包括硬盘驱动器、软盘驱动器以及相关

联的可移动介质、CD-ROM驱动器、光驱或可移动介质盒。实现某些实施方式的功能的模块可以由文件存储子系统存储在存储子系统中,或者存储在处理器可访问的其他机器中。

[0877] 总线子系统提供了一种机制,让计算机系统的各个组件和子系统按照预期相互通信。虽然将总线子系统示意性地显示为单个总线,但是总线子系统的替代实施方式可以使用多个总线。

[0878] 计算机系统本身可以是不同类型的,包括个人计算机、便携式计算机、工作站、计算机终端、网络计算机、电视、大型机、服务器群、广泛分布的松散联网计算机组,或任何其他数据处理系统或用户设备。由于计算机和网络的性质不断变化,图59中描述的计算机系统仅旨在作为说明所公开的技术的特定实例。计算机系统的许多其他配置可能具有比图59中描绘的计算机系统更多或更少的组件。

[0879] 深度学习处理器可以是GPU或FPGA,可以由深度学习云平台托管,如谷歌云平台、Xilinx和Cirrascle。深度学习处理器的实例包括谷歌的张量处理单元(TPU),机架式解决方案(如GX4 Rackmount Series、GX8 Rackmount Series)、英伟达DGX-1、微软的Stratix V FPGA、Graphcore的智能处理器单元(IPU)、高通的带Snapdragon处理器的Zeroth平台、英伟达的Volta、英伟达的DRIVE PX、英伟达的JETSON TX1/TX2 MODULE、英特尔的Nirvana、Movidius VPU、富士通DPI、ARM的DynamicIQ、IBM TrueNorth以及其它。

[0880] 前面的描述是为了能够制造和使用所公开的技术。对所公开的实施方式的各种修改将是显而易见的,并且在不脱离所公开的技术的精神和范围的情况下,本文定义的一般原理可以应用于其他实施方式和应用。因此,所公开的技术不限于所示的实施方式,而是符合与本文公开的原理和特征一致的最宽范围。所公开技术的范围由所附权利要求限定。

[0881] 附录

[0882] 以下是发明人撰写的论文中列出的潜在相关参考文献的书目。论文的主题涵盖在本申请要求其优先权/权益的美国临时专利申请中。这些参考资料可由法律顾问根据要求提供,也可以经由全球档案查阅。

[0883] 1.Amit,M.,Donyo,M.,Hollander,D.,Goren,A.,Kim,E.,Gelfman,S.,Lev-Maor,G.,Burstein,D.,Schwartz,S.,Postolsky,B.,et al.(2012).Differential GC Content between Exons and Introns Establishes Distinct Strategies of Splice-Site Recognition.Cell Rep.1,543-556.

[0884] 2.Andersson,R.,Enroth,S.,Rada-Iglesias,A.,Wadelius,C.,and Komorowski,J.(2009).Nucleosomes are well positioned in exons and carry characteristic histone modifications.Genome Res.19,1732-1741.

[0885] 3.Azad,A.,Rajwa,B.,and Pothen,A.(2016).flowVS:channel-specific variance stabilization in flow cytometry.BMC Bioinformatics 17,291.

[0886] 4.Barash,Y.,Calarco,J.A.,Gao,W.,Pan,Q.,Wang,X.,Shai,O.,Blencowe,B.J.,and Frey,B.J.(2010).Deciphering the splicing code.Nature 465,53-59.

[0887] 5.Berget,S.M.(1995).Exon recognition in vertebrate splicing.J.Biol.Chem.270,2411-2414.

[0888] 6.Blencowe,B.J.(2006).Alternative Splicing:New Insights from Global Analyses.Cell 126,37-47.

- [0889] 7.Boerkoel,C.F.,Exelbert,R.,Nicastri,C.,Nichols,R.C.,Miller,F.W.,Plotz,P.H.,and Raben,N. (1995).Leaky splicing mutation in the acid maltase gene is associated with delayed onset of glycogenosis type II.Am.J.Hum.Genet.56,887-897.
- [0890] 8.Boyd,S.,Cortes,C.,Mohri,M.,and Radovanovic,A. (2012).Accuracy at the Top.In Advances in Neural Processing Systems,pp.953-961.
- [0891] 9.Close,P.,East,P.,Dirac-Svejstrup,A.B.,Hartmann,H.,Heron,M.,Maslen,S.,Chariot,A.,**Söding**,J.,Skehel,M.,and Svejstrup,J.Q. (2012).DBIRD complex integrates alternative mRNA splicing with RNA polymerase II transcript elongation.Nature 484,386-389.
- [0892] 10.Cooper,T.A.,Wan,L.,and Dreyfuss,G. (2009).RNA and Disease.Cell 136,777-793.
- [0893] 11.Cramer,P.,Pesce,C.G.,Baralle,F.E.,and Kornblihtt,a R. (1997).Functional association between promoter structure and transcript alternative splicing.Proc.Natl.Acad.Sci.U.S.A.94,11456-11460.
- [0894] 12.Cummings,B.B.,Marshall,J.L.,Tukiainen,T.,Lek,M.,Donkervoort,S.,Foley,A.R.,Bolduc,V.,Waddell,L.B.,Sandaradura,S.A.,O'Grady,G.L.,et al. (2017).Improving genetic diagnosis in Mendelian disease with transcriptome sequencing.Sci.Transl.Med.9,eaal5209.
- [0895] 13.Dong,S.,Walker,M.F.,Carriero,N.J.,DiCola,M.,Willsey,A.J.,Ye,A.Y.,Waqar,Z.,Gonzalez,L.E.,Overton,J.D.,Frahm,S.,et al. (2014).De novo insertions and deletions of predominantly paternal origin are associated with autism spectrum disorder.Cell Rep.9,16-23.
- [0896] 14.Ernst,J.,Kheradpour,P.,Mikkelsen,T.S.,Shores,N.,Ward,L.D.,Epstein,C.B.,Zhang,X.,Wang,L.,Issner,R.,Coyne,M.,et al. (2011).Mapping and analysis of chromatin state dynamics in nine human cell types.Nature 473,43-49.
- [0897] 15.Fairbrother,W.G.,Yeh,R.F.,Sharp,P.A.,and Burge,C.B. (2002).Predictive identification of exonic splicing enhancers in human genes.Science(80-.).297,1007-1013.
- [0898] 16.Farh,K.K.H.,Marson,A.,Zhu,J.,Kleinewietfeld,M.,Housley,W.J.,Beik,S.,Shores,N.,Whitton,H.,Ryan,R.J.H.,Shishkin,A.A.,et al. (2015).Genetic and epigenetic fine mapping of causal autoimmune disease variants.Nature 518,337-343.
- [0899] 17.Finkel,R.S.,Mercuri,E.,Darras,B.T.,Connolly,A.M.,Kuntz,N.L.,Kirschner,J.,Chiriboga,C.A.,Saito,K.,Servais,L.,Tizzano,E.,et al. (2017).Nusinersen versus Sham Control in Infantile-Onset Spinal Muscular Atrophy.N.Engl.J.Med.377,1723-1732.
- [0900] 18.Fitzgerald,T.W.,Gerety,S.S.,Jones,W.D.,van Kogelenberg,M.,King,

D.A.,McRae,J.,Morley,K.I.,Parthiban,V.,Al-Turki,S.,Ambridge,K.,et al.(2015).Large-scale discovery of novel genetic causes of developmental disorders.Nature 519,223-228.

[0901] 19.Gelfman,S.,Burstein,D.,Penn,O.,Savchenko,A.,Amit,M.,Schwartz,S.,Pupko,T.,and Ast,G.(2012).Changes in exon-intron structure during vertebrate evolution affect the splicing pattern of exons.Genome Res.22,35-50.

[0902] 20.Gelfman,S.,Cohen,N.,Yearim,A.,and Ast,G.(2013).DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon-intron structure.Genome Res.23,789-799.

[0903] 21.Glorot,X.,and Bengio,Y.(2010).Understanding the difficulty of training deep feedforward neural networks.Proc.13th Int.Conf.Artif.Intell.Stat.9,249-256.

[0904] 22.Gouya,L.,Puy,H.,Robreau,A.-M.,Bourgeois,M.,Lamoril,J.,Silva,V.Da,Grandchamp,B.,and Deybach,J.-C.(2002).The penetrance of dominant erythropoietic protoporphyria is modulated by expression of wildtype FECH.Nat.Genet.30,27-28.

[0905] 23.Graubert,T.A.,Shen,D.,Ding,L.,Okeyo-Owuor,T.,Lunn,C.L.,Shao,J.,Krysiak,K.,Harris,C.C.,Koboldt,D.C.,Larson,D.E.,et al.(2012).Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes.Nat.Genet.44,53-57.

[0906] 24.Harrow,J.,Frankish,A.,Gonzalez,J.M.,Tapanari,E.,Diekhans,M.,Kokocinski,F.,Aken,B.L.,Barrell,D.,Zadissa,A.,Searle,S.,et al.(2012).GENCODE: the reference human genome annotation for The ENCODE Project.Genome Res.22,1760-1774.

[0907] 25.He,K.,Zhang,X.,Ren,S.,and Sun,J.(2016a).Deep Residual Learning for Image Recognition.In IEEE Conference on Computer Vision and Pattern Recognition(CVPR),pp.770-778.

[0908] 26.He,K.,Zhang,X.,Ren,S.,and Sun,J.(2016b).Identity mappings in deep residual networks.In European Conference on Computer Vision,pp.630-645.

[0909] 27.Herring,C.A.,Banerjee,A.,McKinley,E.T.,Simmons,A.J.,Ping,J.,Roland,J.T.,Franklin,J.L.,Liu,Q.,Gerdes,M.J.,Coffey,R.J.,et al.(2018).Unsupervised Trajectory Analysis of Single-Cell RNA-Seq and Imaging Data Reveals Alternative Tuft Cell Origins in the Gut.Cell Syst.6,37-51.e9.

[0910] 28.Hoffman,M.M.,Buske,O.J.,Wang,J.,Weng,Z.,Bilmes,J.A.,and Noble,W.S.(2012).Unsupervised pattern discovery in human chromatin structure through genomic segmentation.Nat.Methods 9,473-476.

[0911] 29.Ioffe,S.,and Szegedy,C.(2015).Batch Normalization:Accelerating Deep Network Training by Reducing Internal Covariate Shift.Proc.Mach.Learn.Res.37,448-456.

- [0912] 30. Iossifov, I., O'Roak, B. J., Sanders, S. J., Ronemus, M., Krumm, N., Levy, D., Stessman, H. A., Witherspoon, K. T., Vives, L., Patterson, K. E., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. *Nature* 515, 216-221.
- [0913] 31. Irimia, M., Weatheritt, R. J., Ellis, J. D., Parikshak, N. N., Gonatopoulos-Pournatzis, T., Babor, M., Quesnel-Vallières, M., Tapial, J., Raj, B., O'Hanlon, D., et al. (2014). A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* 159, 1511-1523.
- [0914] 32. Jha, A., Gazzara, M. R., and Barash, Y. (2017). Integrative deep models for alternative splicing. 274-282.
- [0915] 33. Jonkers, I., Kwak, H., and Lis, J. T. (2014). Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *Elife* 3, e02407.
- [0916] 34. Jung, H., Lee, D., Lee, J., Park, D., Kim, Y. J., Park, W.-Y., Hong, D., Park, P. J., and Lee, E. (2015). Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nat. Genet.* 47, 1242-1248.
- [0917] 35. Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., Zaugg, J. B., Kundaje, A., Liu, Y., Boyle, A. P., Zhang, Q. C., Zakharia, F., Spacek, D. V., et al. (2013). Extensive variation in chromatin states across humans. *Science* (80-.). 342.
- [0918] 36. Katz, Y., Wang, E. T., Airolidi, E. M., and Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* 7, 1009-1015.
- [0919] 37. Keren, H., Lev-Maor, G., and Ast, G. (2010). Alternative splicing and evolution: Diversification, exon definition and function. *Nat. Rev. Genet.* 11, 345-355.
- [0920] 38. Kingma, D. P., and Ba, J. L. (2015). Adam: A Method for Stochastic Optimization. *Int. Conf. Learn. Represent.*
- [0921] 39. Kosmicki, J. A., Samocha, K. E., Howrigan, D. P., Sanders, S. J., Slowikowski, K., Lek, M., Karczewski, K. J., Cutler, D. J., Devlin, B., Roeder, K., et al. (2017). Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat. Genet.* 49, 504-510.
- [0922] 40. Kremer, L. S., Bader, D. M., Mertes, C., Kopajtich, R., Pichler, G., Iuso, A., Haack, T. B., Graf, E., Schwarzmayr, T., Terrile, C., et al. (2017). Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat. Commun.* 8, 15824.
- [0923] 41. Krishnamoorthy, K., and Thomson, J. (2004). A more powerful test for comparing two Poisson means. *J. Stat. Plan. Inference* 119, 23-35.
- [0924] 42. Lee, H., Deignan, J. L., Dorrani, N., Strom, S. P., Kantarci, S., Quintero-

Rivera, F., Das, K., Toy, T., Harry, B., Yourshaw, M., et al. (2014). Clinical Exome Sequencing for Genetic Identification of Rare Mendelian Disorders. *JAMA* 312, 1880-1887.

[0925] 43. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285-291.

[0926] 44. Li, Y.I., Sanchez-Pulido, L., Haerty, W., and Ponting, C.P. (2015). RBFox and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts. *Genome Res.* 25, 1-13.

[0927] 45. Li, Y.I., Knowles, D.A., Humphrey, J., Barbeira, A.N., Dickinson, S.P., Im, H.K., and Pritchard, J.K. (2018). Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* 50, 151-158.

[0928] 46. Licatalosi, D.D., and Darnell, R.B. (2006). Splicing Regulation in Neurologic Disease. *Neuron* 52, 93-101.

[0929] 47. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* 45, 580-585.

[0930] 48. Luco, R.F., Pan, Q., Tominaga, K., Blencowe, B.J., Pereira-Smith, O.M., and Misteli, T. (2010). Regulation of alternative splicing by histone modifications. *Science* 327, 996-1000.

[0931] 49. Luco, R.F., Allo, M., Schor, I.E., Kornblihtt, A.R., and Misteli, T. (2011). Epigenetics in Alternative Pre-mRNA Splicing. *Cell* 144, 16-26.

[0932] 50. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* (80-.). 337, 1190-1195.

[0933] 51. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biol.* 17, 122.

[0934] 52. McRae, J.F., Clayton, S., Fitzgerald, T.W., Kaplanis, J., Prigmore, E., Rajan, D., Sifrim, A., Aitken, S., Akawi, N., Alvi, M., et al. (2017). Prevalence and architecture of de novo mutations in developmental disorders. *Nature* 542, 433-438.

[0935] 53. Monroe, G.R., Frederix, G.W., Savelberg, S.M.C., de Vries, T.I., Duran, K.J., van der Smagt, J.J., Terhal, P.A., van Hasselt, P.M., Kroes, H.Y., Verhoeven-Duif, N.M., et al. (2016). Effectiveness of whole-exome sequencing and costs of the traditional diagnostic trajectory in children with intellectual disability. *Genet. Med.* 18, 949-956.

- [0936] 54. Neale, B.M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K.E., Sabo, A., Lin, C.-F., Stevens, C., Wang, L.-S., Makarov, V., et al. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* 485, 242-245.
- [0937] 55. Oord, A. van den, Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio. *ArXiv:1609.03499[Cs.SD]*.
- [0938] 56. Padgett, R.A. (2012). New connections between splicing and human disease. *Trends Genet.* 28, 147-154.
- [0938] 57. Pertea, M., Lin, X., and Salzberg, S.L. (2001). GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res.* 29, 1185-1190.
- [0939] 58. Reed, R., and Maniatis, T. (1988). The role of the mammalian branchpoint sequence in pre-mRNA splicing. *Genes Dev.* 2, 1268-1276.
- [0940] 59. Reese, M.G., Eeckman, F.H., Kulp, D., and Haussler, D. (1997). Improved splice site detection in Genie. *J. Comput. Biol.* 4, 311-323.
- [0941] 60. Robinson, E.B., Samocha, K.E., Kosmicki, J.A., McGrath, L., Neale, B.M., Perlis, R.H., and Daly, M.J. (2014). Autism spectrum disorder severity reflects the average contribution of de novo and familial influences. *Proc. Natl. Acad. Sci. U.S.A.* 111, 15161-15165.
- [0942] 61. De Rubeis, S., He, X., Goldberg, A.P., Poultney, C.S., Samocha, K., Ercument Cicek, A., Kou, Y., Liu, L., Fromer, M., Walker, S., et al. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* 515, 209-215.
- [0943] 62. Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., **Rehnström**, K., Mallick, S., Kirby, A., et al. (2014). A framework for the interpretation of de novo variation in human disease. *Nat. Genet.* 46, 944-950.
- [0944] 63. Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L., et al. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* 485, 237-241.
- [0945] 64. Sanders, S.J., He, X., Willsey, A.J., Ercan-Sencicek, A.G., Samocha, K.E., Cicek, A.E., Murtha, M.T., Bal, V.H., Bishop, S.L., Dong, S., et al. (2015). Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* 87, 1215-1233.
- [0946] 65. Sanz, D.J., Acedo, A., Infante, M., Durán, M., Pérez-Cabornero, L., Esteban-Cardenosa, E., Lastra, E., Pagani, F., Miner, C., and Velasco, E.A. (2010). A high proportion of DNA variants of BRCA1 and BRCA2 is associated with aberrant splicing in breast/ovarian cancer patients. *Clin. Cancer Res.* 16, 1957-1967.
- [0947] 66. Schwartz, S., Meshorer, E., and Ast, G. (2009). Chromatin organization marks exon-intron structure. *Nat. Struct. Mol. Biol.* 16, 990-995.

- [0948] 67.Scotti,M.M.,and Swanson,M.S.(2016).RNA mis-splicing in disease.Nat.Rev.Genet.17,19-32.
- [0949] 68.SEQC/MAQC-III Consortium,S.(2014).A comprehensive assessment of RNA-seq accuracy,reproducibility and information content by the Sequencing Quality Control Consortium.Nat.Biotechnol.32,903-914.
- [0950] 69.Shirai,C.L.,Ley,J.N.,White,B.S.,Kim,S.,Tibbitts,J.,Shao,J.,Ndonwi,M.,Wadugu,B.,Duncavage,E.J.,Okeyo-Owuor,T.,et al.(2015).Mutant U2AF1 Expression Alters Hematopoiesis and Pre-mRNA Splicing In Vivo.Cancer Cell 27, 631-643.
- [0951] 70.Shrikumar,A.,Greenside,P.,and Kundaje,A.(2017).Learning Important Features Through Propagating Activation Differences.Proc.Mach.Learn.Res.70, 3145-3153.
- [0952] 71.Shukla,S.,Kavak,E.,Gregory,M.,Imashimizu,M.,Shutinoski,B.,Kashlev,M.,Oberdoerffer,P.,Sandberg,R.,and Oberdoerffer,S.(2011).CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing.Nature 479,74-79.
- [0953] 72.Soemedi,R.,Cygan,K.J.,Rhine,C.L.,Wang,J.,Bulacan,C.,Yang,J., Bayrak-Toydemir,P.,McDonald,J.,and Fairbrother,W.G.(2017).Pathogenic variants that alter protein code often disrupt splicing.Nat.Genet.49,848-855.
- [0954] 73.Spies,N.,Nielsen,C.B.,Padgett,R.A.,and Burge,C.B.(2009).Biased chromatin signatures around polyadenylation sites and exons.Mol.Cell 36,245-254.
- [0955] 74.Stark,Z.,Tan,T.Y.,Chong,B.,Brett,G.R.,Yap,P.,Walsh,M.,Yeung,A., Peters,H.,Mordaunt,D.,Cowie,S.,et al.(2016).A prospective evaluation of whole-exome sequencing as a first-tier molecular test in infants with suspected monogenic disorders.Genet.Med.18,1090-1096.
- [0956] 75.Supek,F.,**Miñana**,B.,Valcárcel,J.,Gabaldón,T.,and Lehner,B.(2014).Synonymous mutations frequently act as driver mutations in human cancers.Cell 156,1324-1335.
- [0957] 76.Tan,T.Y.,Dillon,O.J.,Stark,Z.,Schofield,D.,Alam,K.,Shrestha,R., Chong,B.,Phelan,D.,Brett,G.R.,Creed,E.,et al.(2017).Diagnostic impact and cost-effectiveness of whole-exome sequencing for ambulant children with suspected monogenic conditions.JAMA Pediatr.171,855-862.
- [0958] 77.Tennessen,J.A.,Bigham,A.W.,O'Connor,T.D.,Fu,W.,Kenny,E.E.,Gravel,S.,McGee,S.,Do,R.,Liu,X.,Jun,G.,et al.(2012).Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes.Science(80-.).337,64-69.
- [0959] 78.The GTEx Consortium,Welter,D.,MacArthur,J.,Morales,J.,Burdett,T., Hall,P.,Junkins,H.,Klemm,A.,Flicek,P.,Manolio,T.,et al.(2015).The Genotype-Tissue Expression(GTEx)pilot analysis:multitissue gene regulation in

humans. *Science* (80-.). 348,648-660.

[0960] 79. Tilgner, H., Nikolaou, C., Althammer, S., Sammeth, M., Beato, M., Valcárcel, J., and Guigó, R. (2009). Nucleosome positioning as a determinant of exon recognition. *Nat. Struct. Mol. Biol.* 16, 996-1001.

[0961] 80. Tilgner, H., Knowles, D.G., Johnson, R., Davis, C.A., Chakraborty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T.R., and Guigó, R. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.* 22, 1616-1625.

[0962] 81. Trujillano, D., Bertoli-Avella, A.M., Kumar Kandaswamy, K., Weiss, M.E., Köster, J., Marais, A., Paknia, O., Schröder, R., Garcia-Aznar, J.M., Werber, M., et al. (2017). Clinical exome sequencing: Results from 2819 samples reflecting 1000 families. *Eur. J. Hum. Genet.* 25, 176-182.

[0963] 82. Turner, T.N., Hormozdiari, F., Duyzend, M.H., McClymont, S.A., Hook, P.W., Iossifov, I., Raja, A., Baker, C., Hoekzema, K., Stessman, H.A., et al. (2016). Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA. *Am. J. Hum. Genet.* 98, 58-74.

[0964] 83. Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R.B. (2003). CLIP Identifies Nova-Regulated RNA Networks in the Brain. *Science* (80-.). 302, 1212-1215.

[0965] 84. Veloso, A., Kirkconnell, K.S., Magnuson, B., Biewen, B., Paulsen, M.T., Wilson, T.E., and Ljungman, M. (2014). Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. *Genome Res.* 24, 896-905.

[0966] 85. Wang, G.-S., and Cooper, T.A. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet.* 8, 749-761. 86. Wang, Z., and Burge, C.B. (2008). Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* 14, 802-813.

[0967] 87. Wang, E.T., Sandberg, R., Luo, S., Khrebtkova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* 456, 470-476.

[0968] 88. Wang, Z., Rolish, M.E., Yeo, G., Tung, V., Mawson, M., and Burge, C.B.

[0969] (2004). Systematic identification and analysis of exonic splicing silencers. *Cell* 119, 831-845.

[0970] 89. Wu, J., Anczukow, O., Krainer, A.R., Zhang, M.Q., and Zhang, C. (2013). Olego: Fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. *Nucleic Acids Res.* 41, 5149-5163.

[0971] 90. Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K.C., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., et al. (2015). The

human splicing code reveals new insights into the genetic determinants of disease. *Science* (80-.). 347,1254806.

[0972] 91. Yang, Y., Muzny, D.M., Xia, F., Niu, Z., Person, R., Ding, Y., Ward, P., Braxton, A., Wang, M., Buhay, C., et al. (2014). Molecular Findings Among Patients Referred for Clinical Whole-Exome Sequencing. *JAMA* 312,1870.

[0973] 92. Yeo, G., and Burge, C.B. (2004). Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. *J. Comput. Biol.* 11, 377-394.

[0974] 93. Yoshida, K., Sanada, M., Shiraishi, Y., Nowak, D., Nagata, Y., Yamamoto, R., Sato, Y., Sato-Otsubo, A., Kon, A., Nagasaki, M., et al. (2011). Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* 478,64-69.

[0975] 94. Yu, F., and Koltun, V. (2016). Multi-Scale Context Aggregation by Dilated Convolutions. *ArXiv:1511.07122*[Cs.CV].

[0976] 95. Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* 12,931-934.

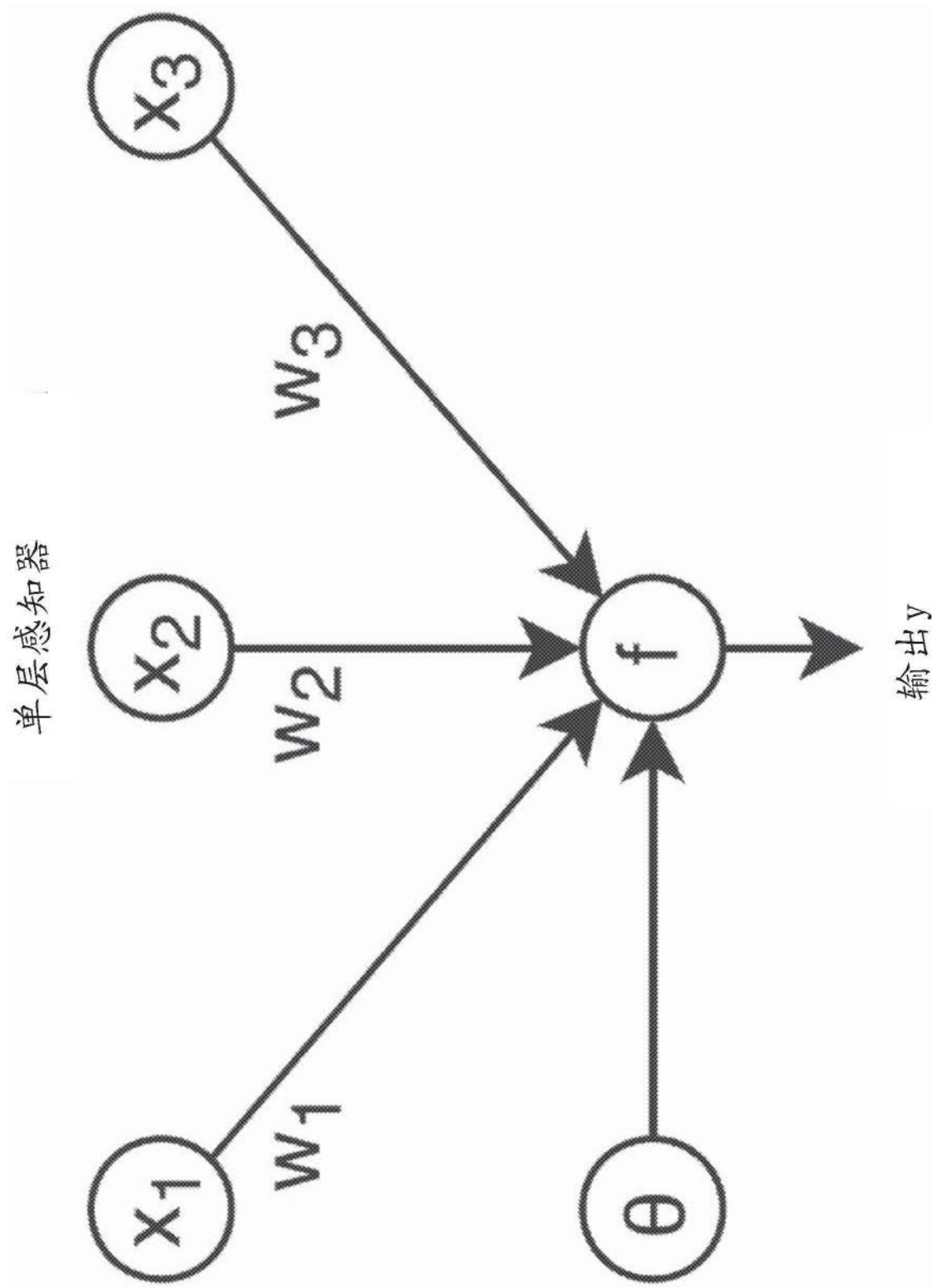


图1

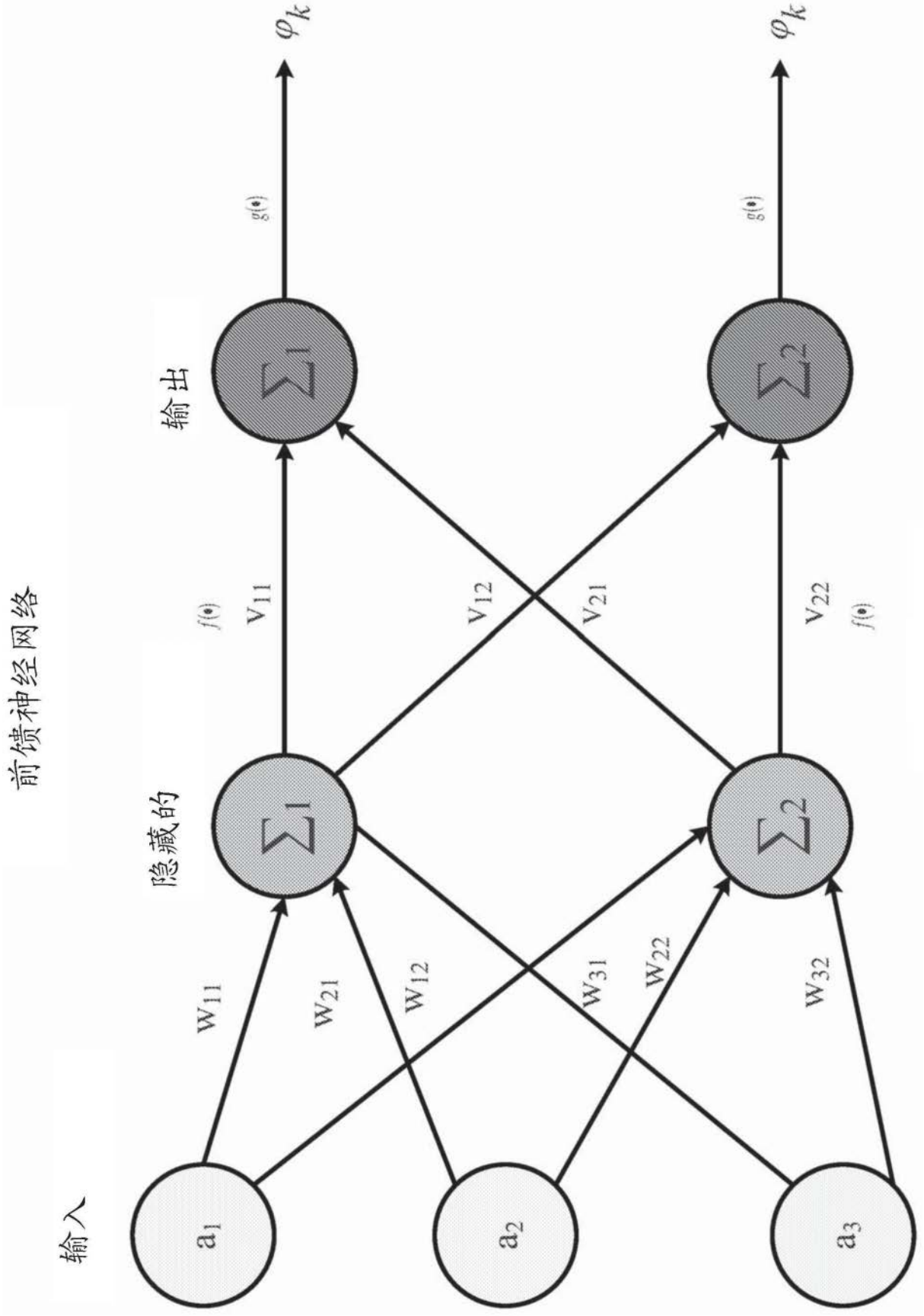


图2

卷积神经网络

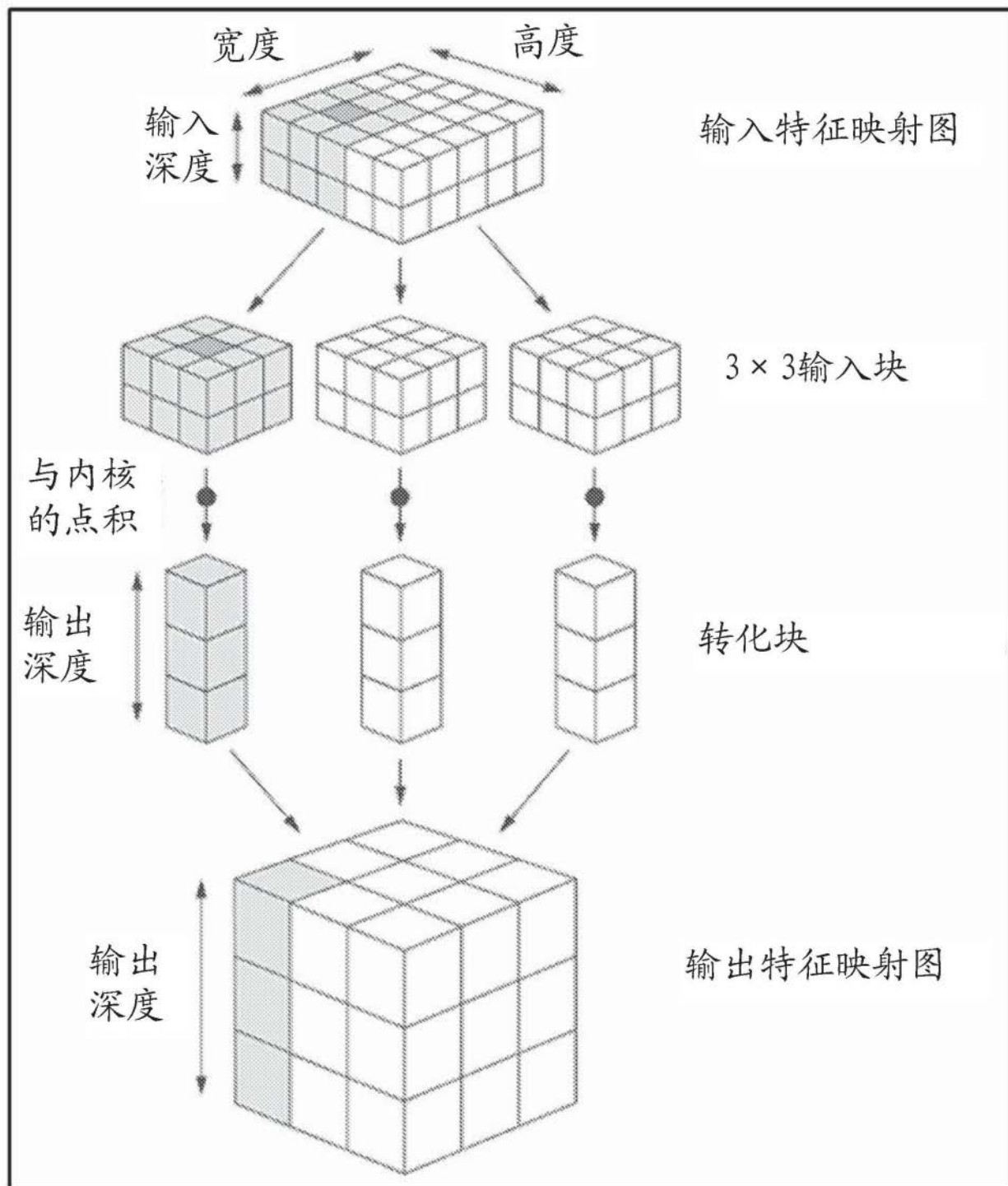


图3

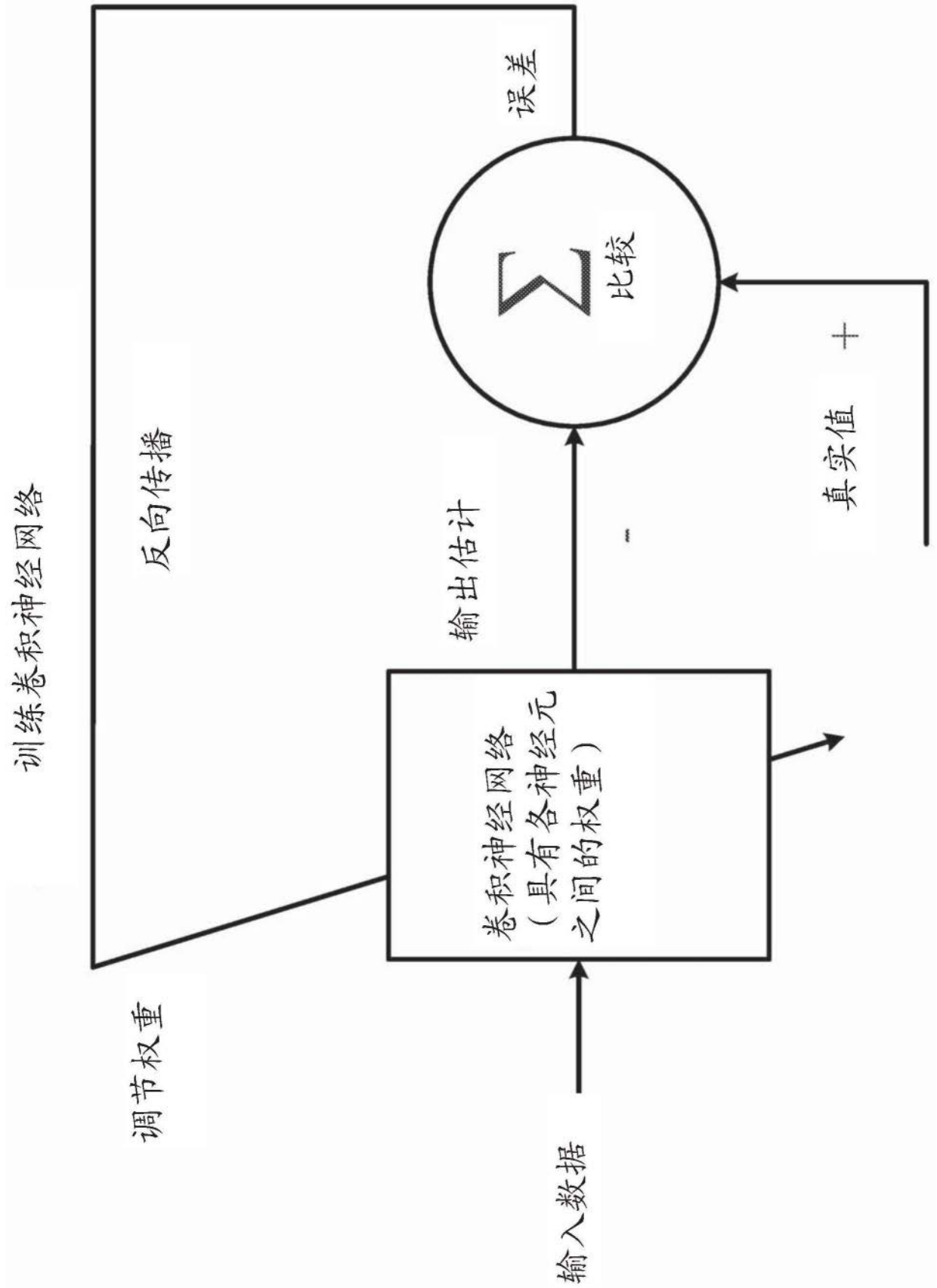


图4

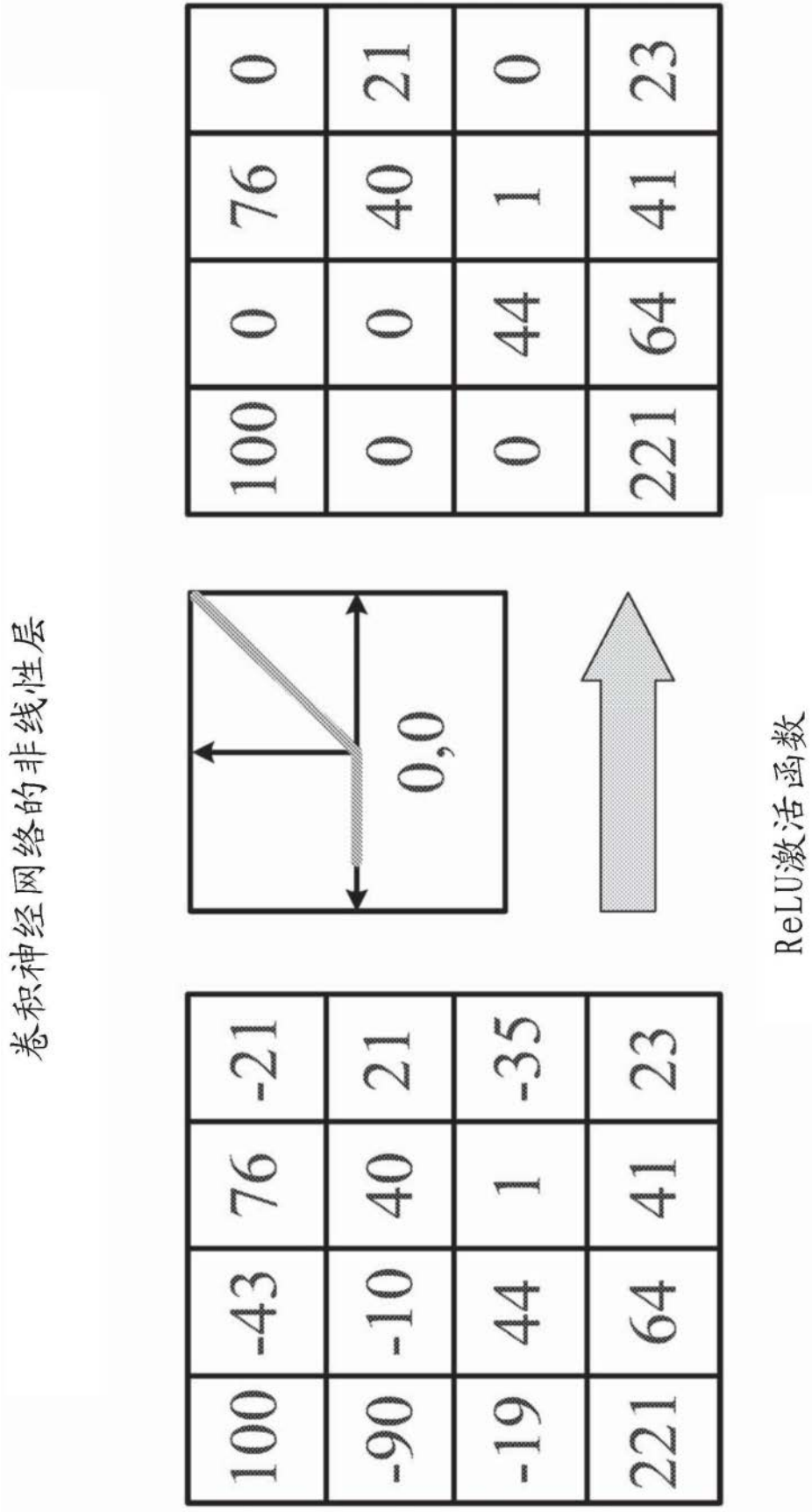


图5

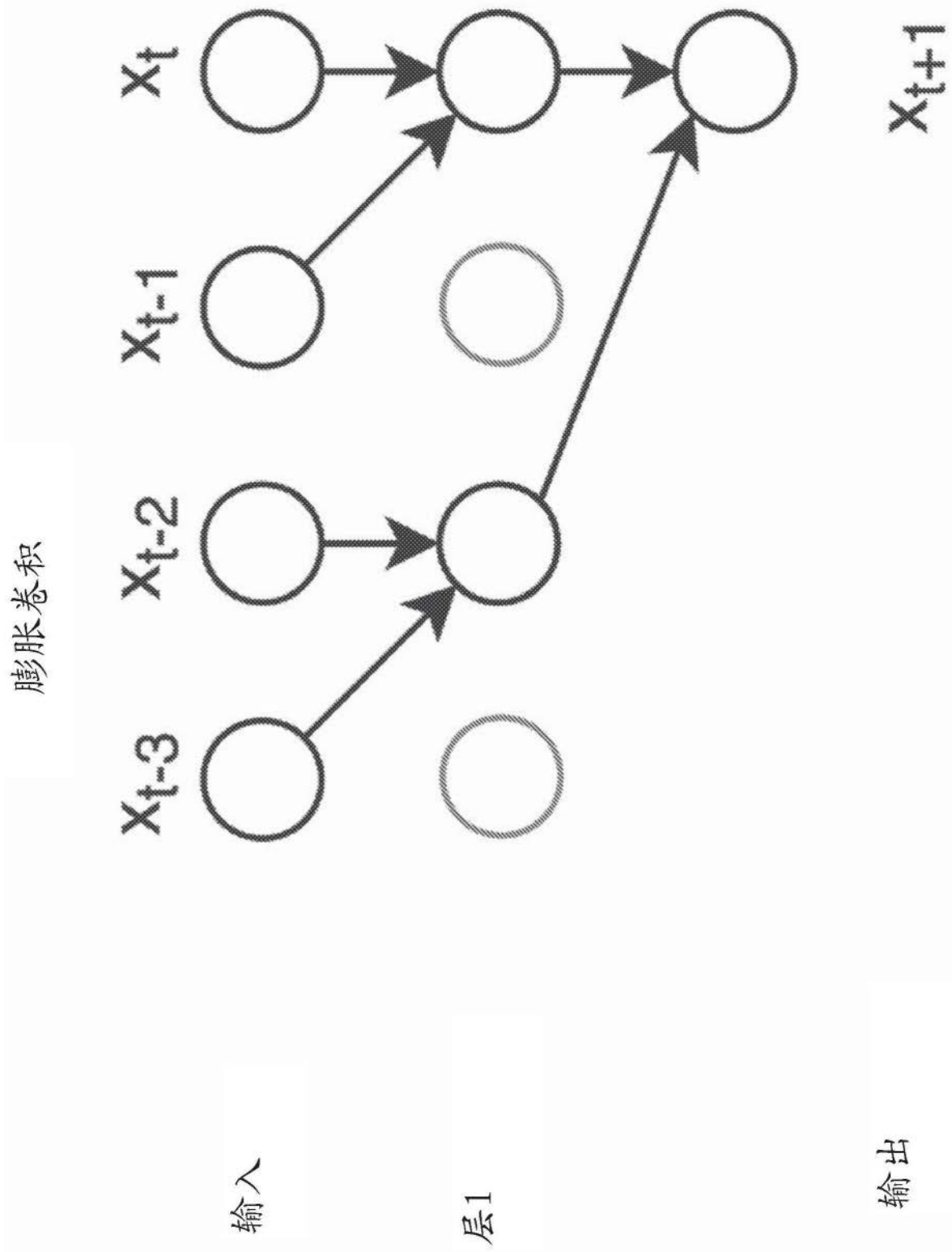


图6

卷积神经网络的子采样层

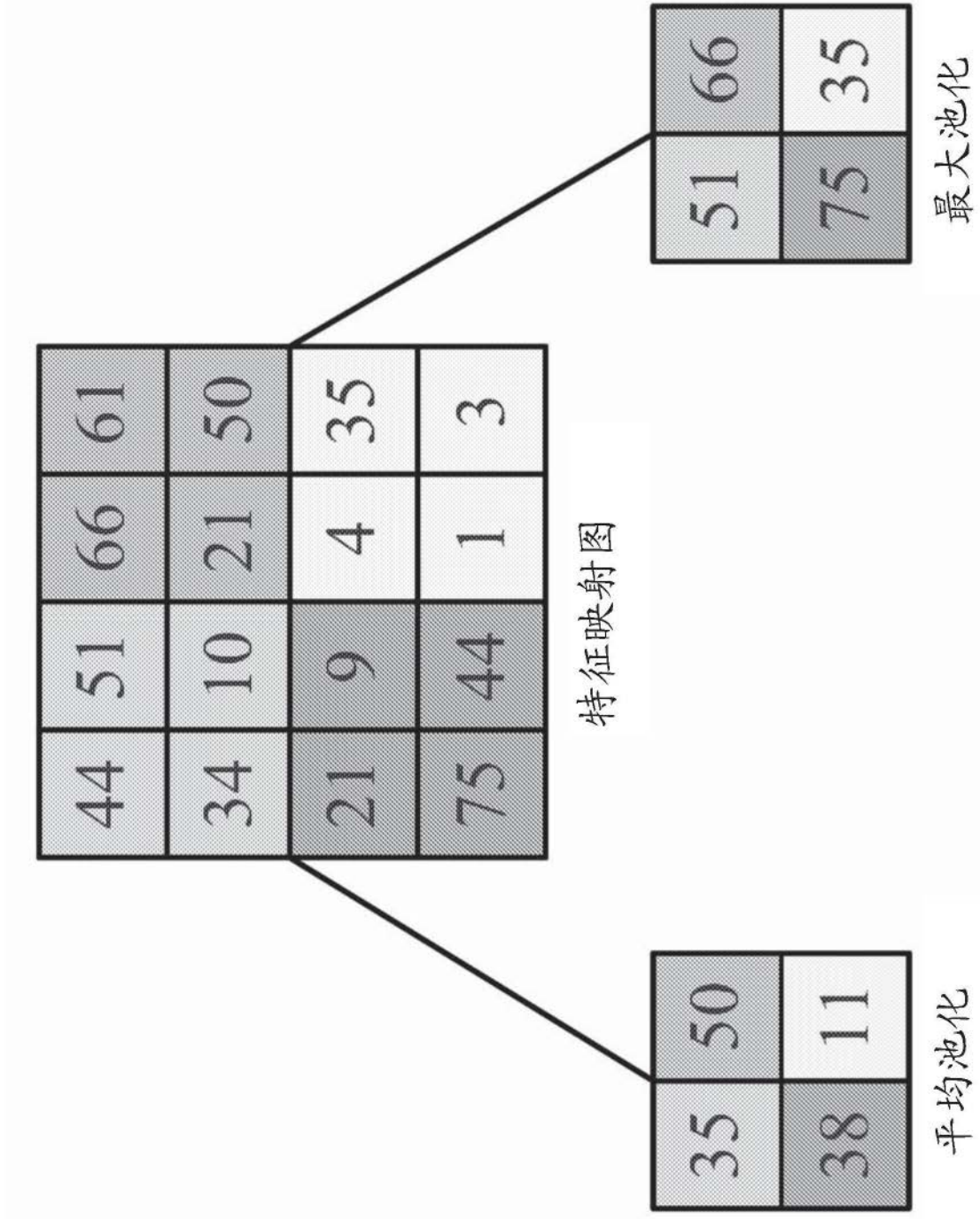


图7

卷积神经网络的卷积层

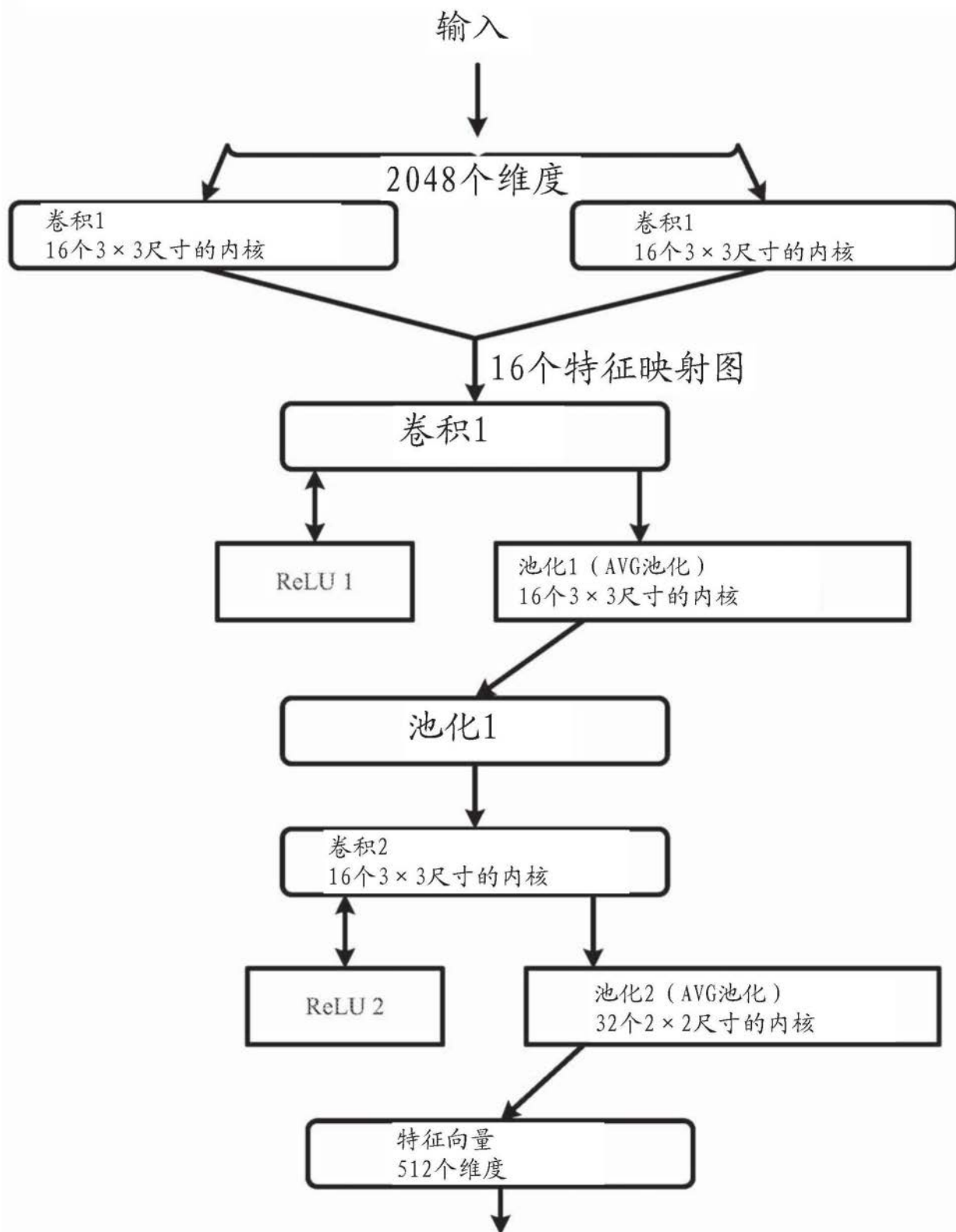
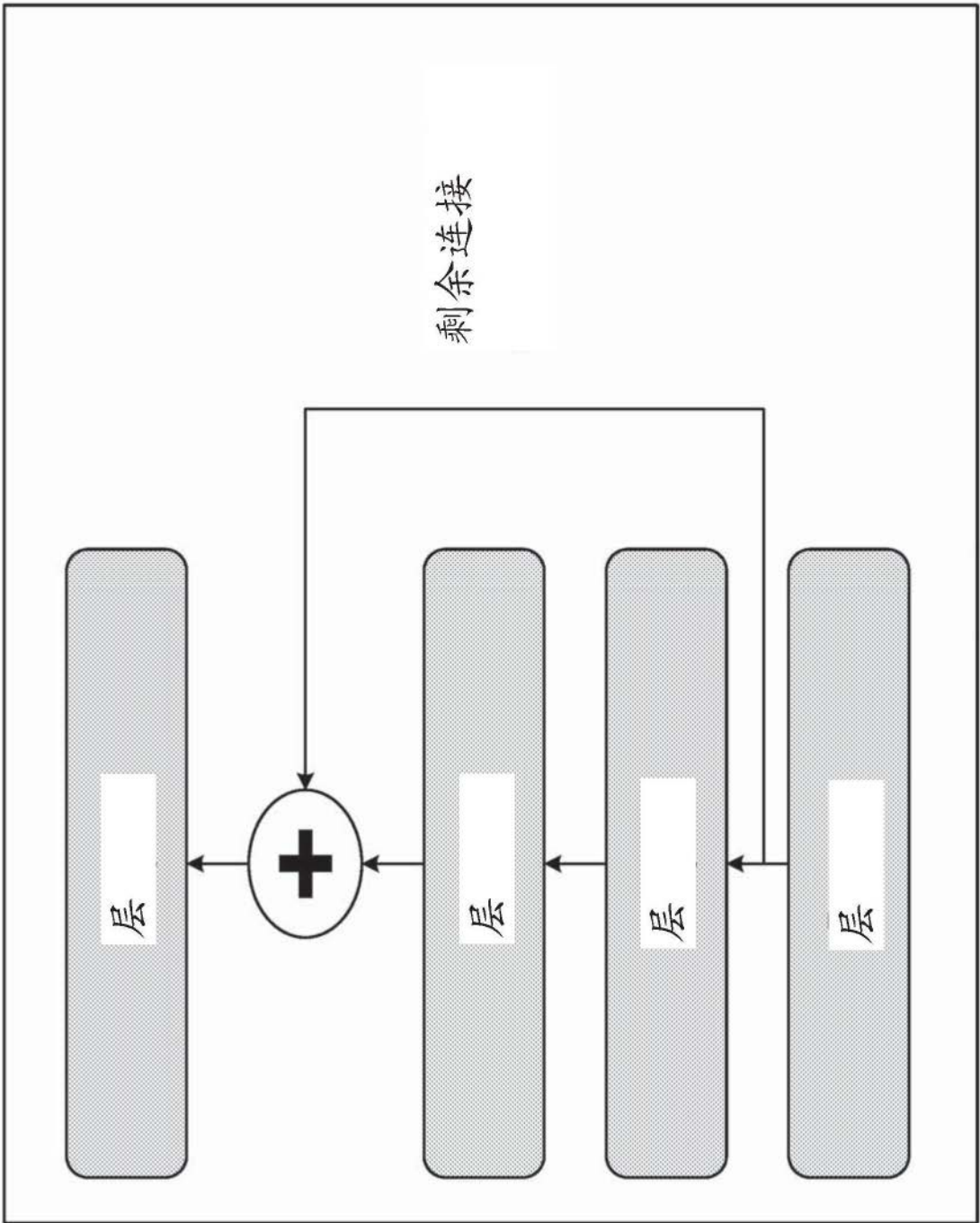


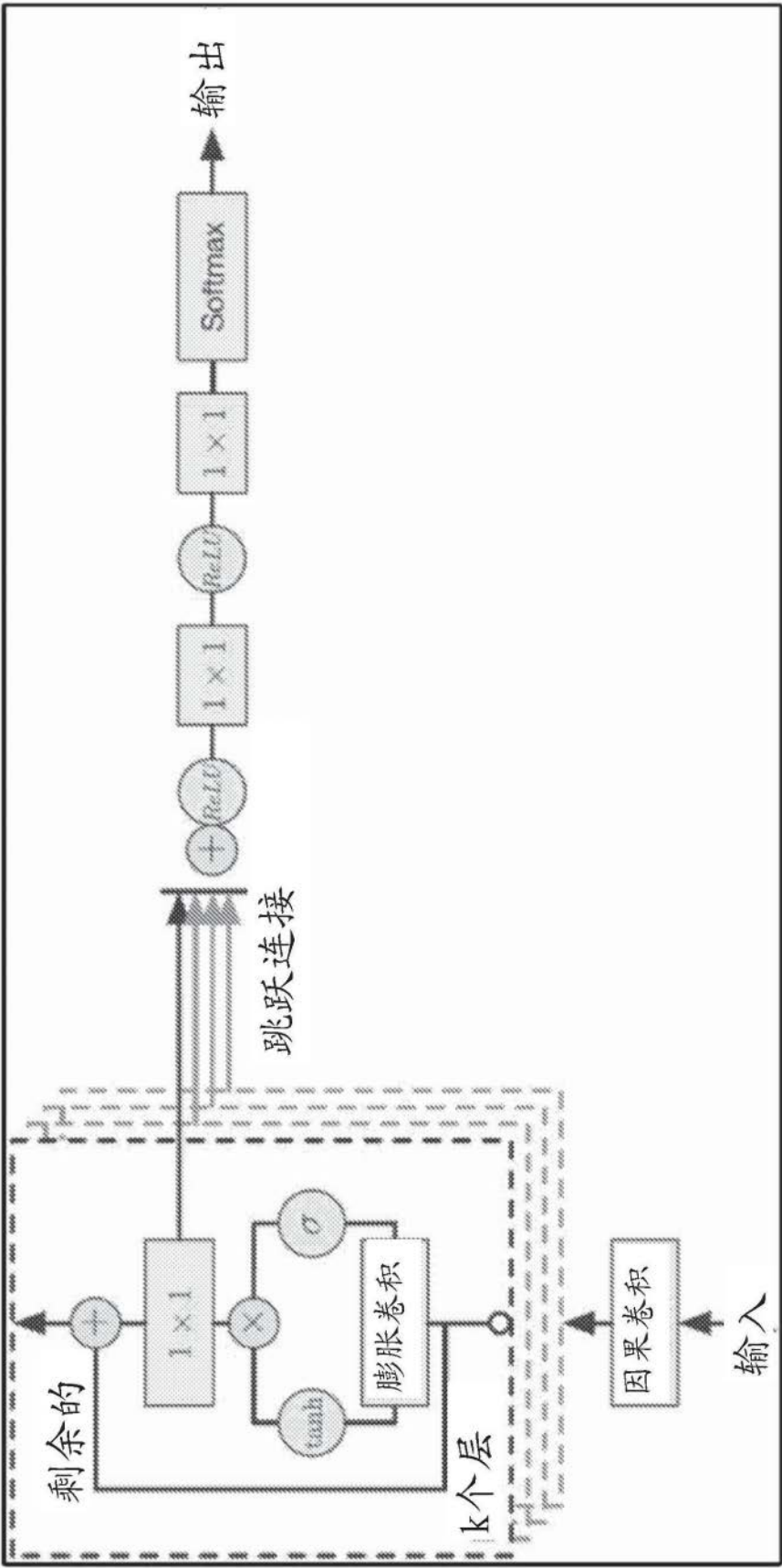
图8

卷积神经网络中使用的剩余连接



在

图9



在卷积神经网络中使用的残差块和跳跃连接

图10

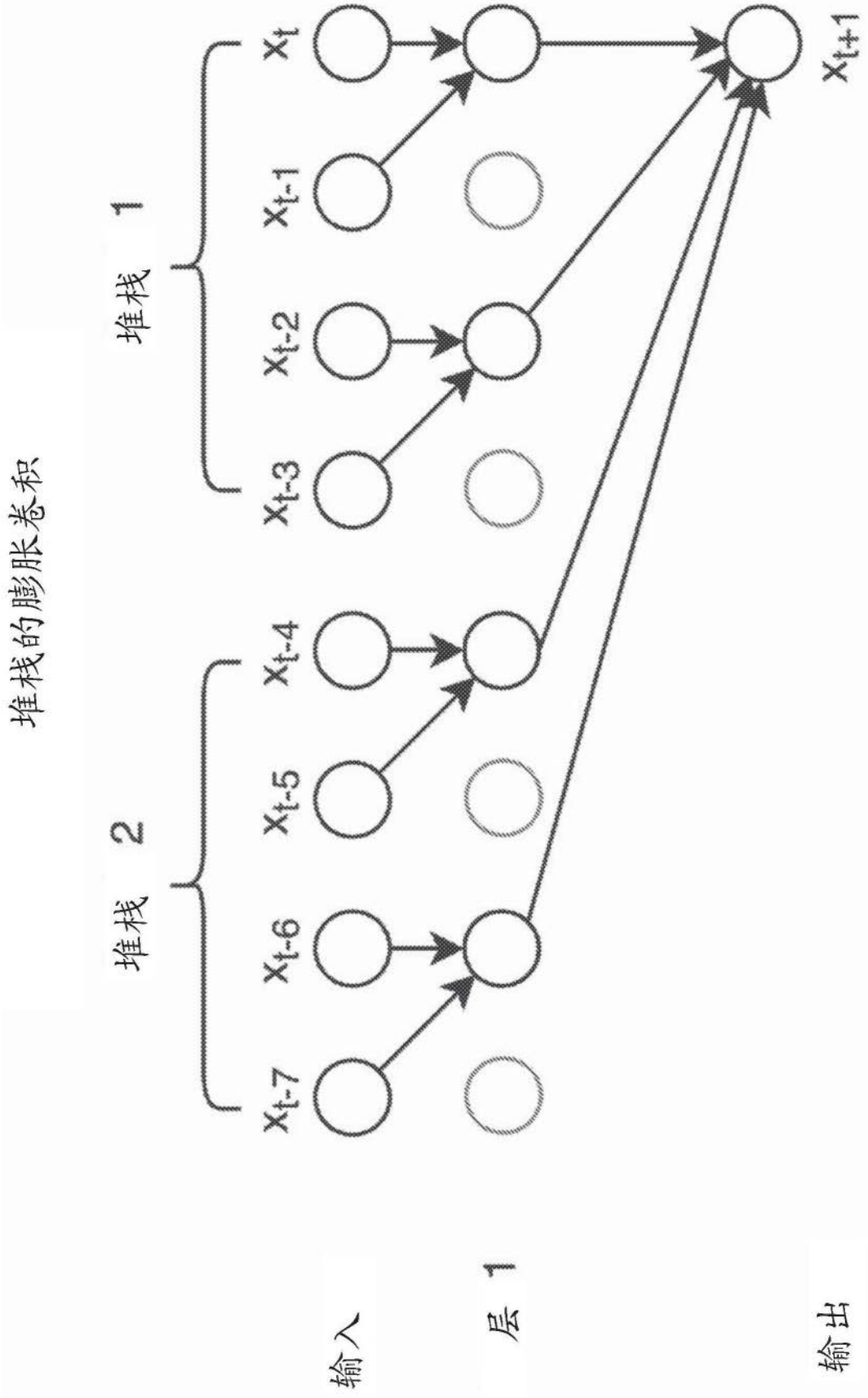


图11

利用卷积神经网络的批归一化前向传递

$$\begin{aligned}
 \mu_{\mathcal{B}} &= \frac{1}{n} \sum_{i=1}^n x_i^{(\ell-1)} \\
 \sigma_{\mathcal{B}}^2 &= \frac{1}{n} \sum_{i=1}^n (x_i^{(\ell-1)} - \mu_{\mathcal{B}})^2 \\
 \hat{\mathbf{x}}^{(\ell-1)} &= \frac{\mathbf{x}^{(\ell-1)} - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \\
 \mathbf{x}^{(\ell)} &= \gamma^{(\ell)} \hat{\mathbf{x}}^{(\ell-1)} + \beta^{(\ell)}
 \end{aligned}$$

图12

利用卷积神经网络的批归一化推理

$$\hat{x}^{(\ell-1)} = \frac{x^{(\ell-1)} - \mu_{\mathcal{D}}}{\sqrt{\sigma_{\mathcal{D}}^2 + \epsilon}}$$
$$x_i^{(\ell)} = \gamma^{(\ell)} \hat{x}_i^{(\ell-1)} + \beta^{(\ell)}$$

图13

利用卷积神经网络的批归一化后向传递

$$\begin{aligned}\nabla_{\gamma^{(l)}} \mathcal{L} &= \sum_{i=1}^n (\nabla_{\mathbf{x}^{(l+1)}} \mathcal{L})_i \cdot \hat{\mathbf{x}}_i^{(l)} \\ \nabla_{\beta^{(l)}} \mathcal{L} &= \sum_{i=1}^n (\nabla_{\mathbf{x}^{(l+1)}} \mathcal{L})_i\end{aligned}$$

图14

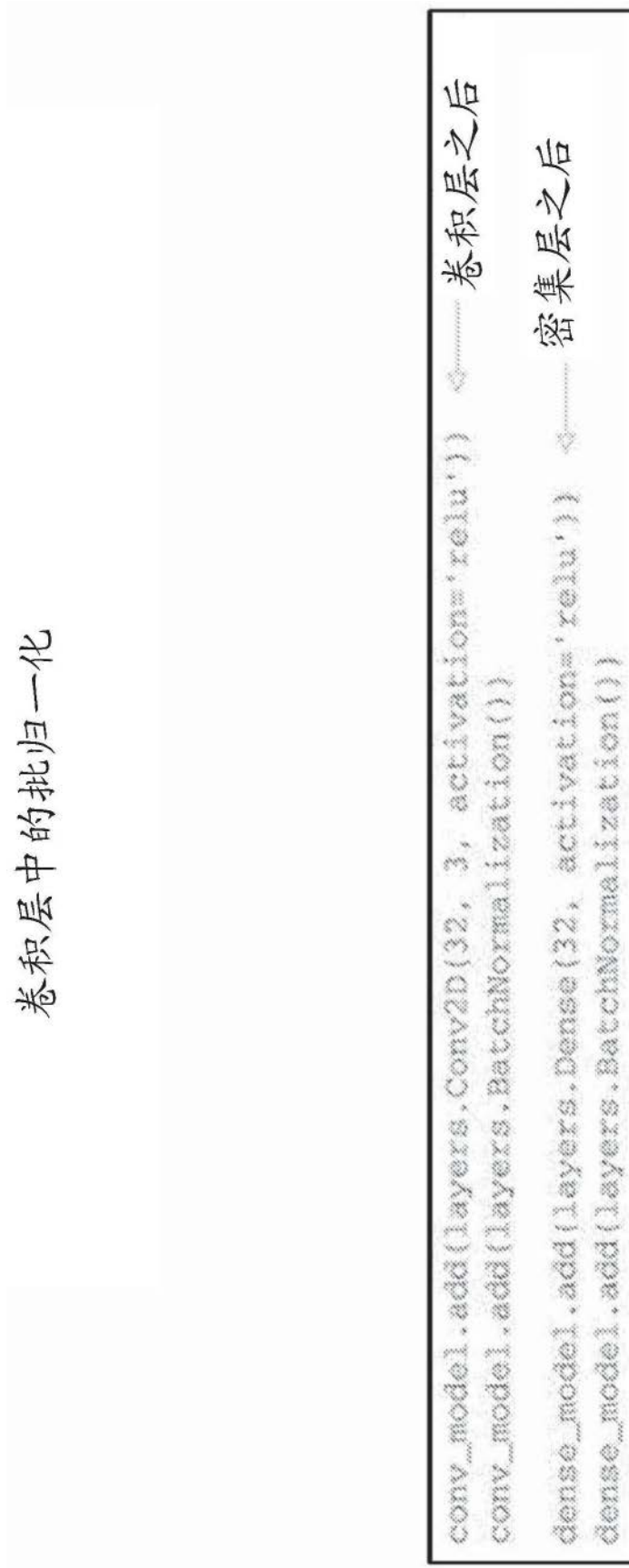


图15

卷积神经网络中使用的ID卷积

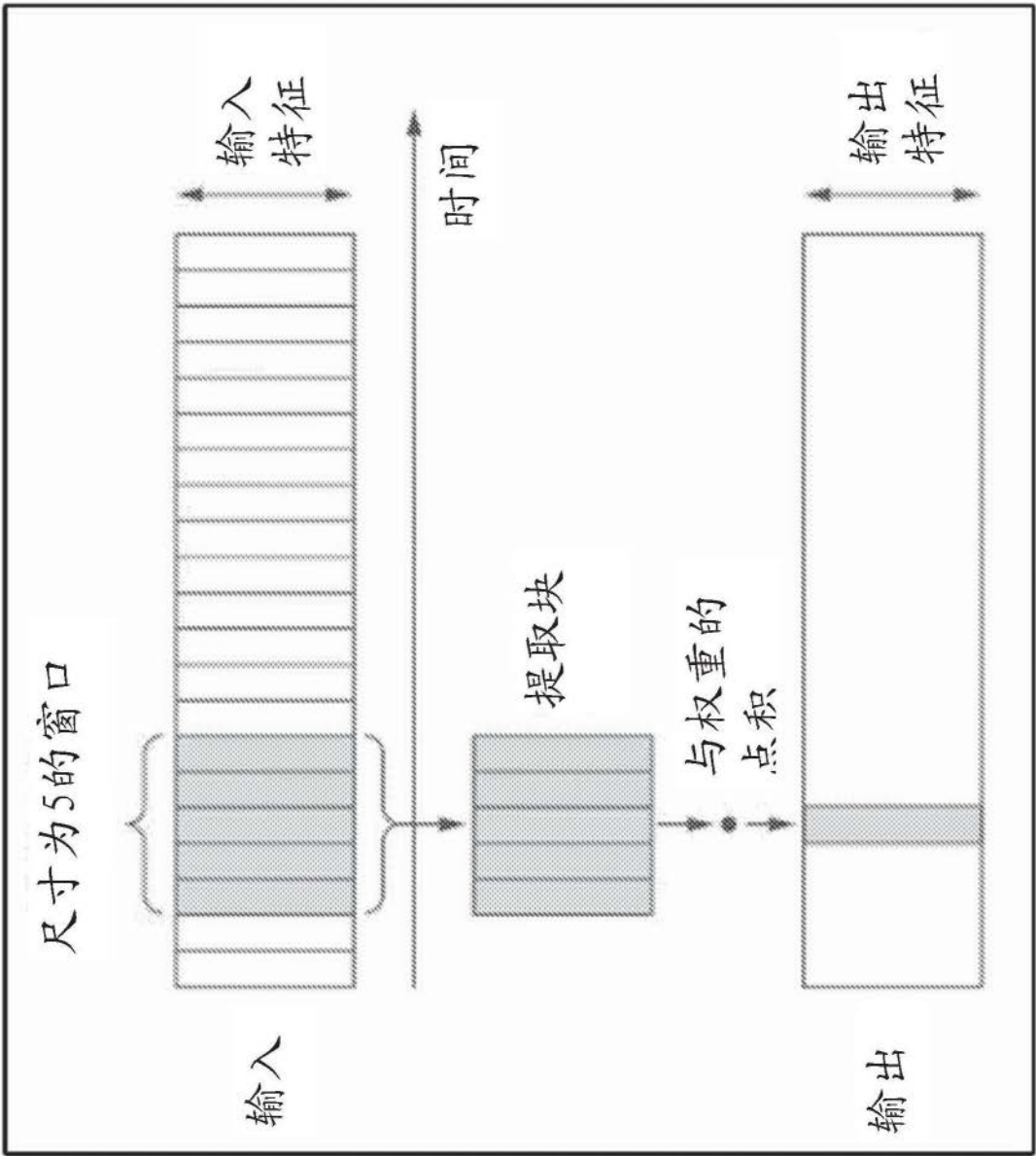


图16

卷积神经网络中的全局平均池化 (GAP)

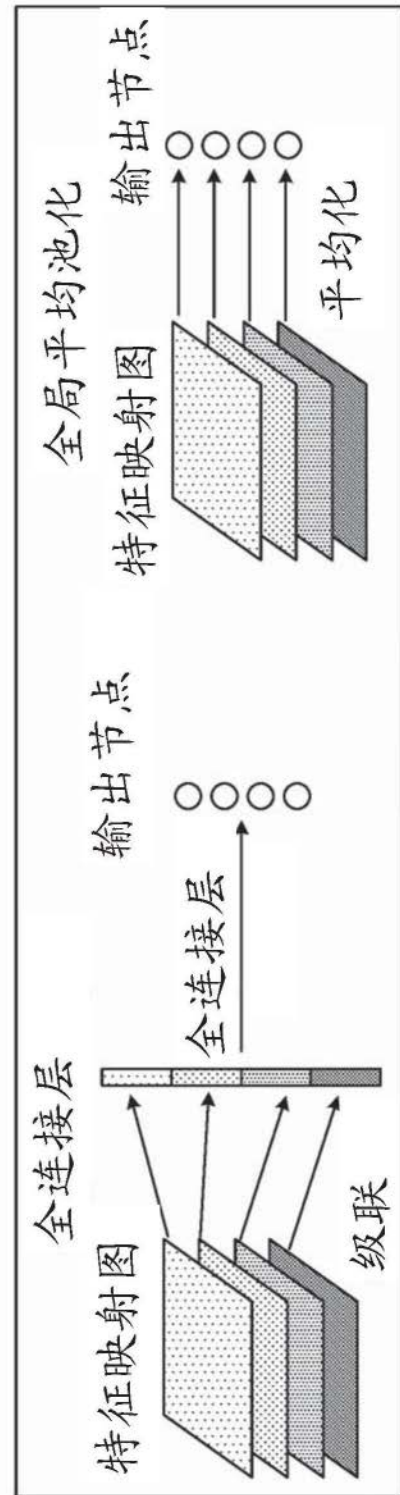
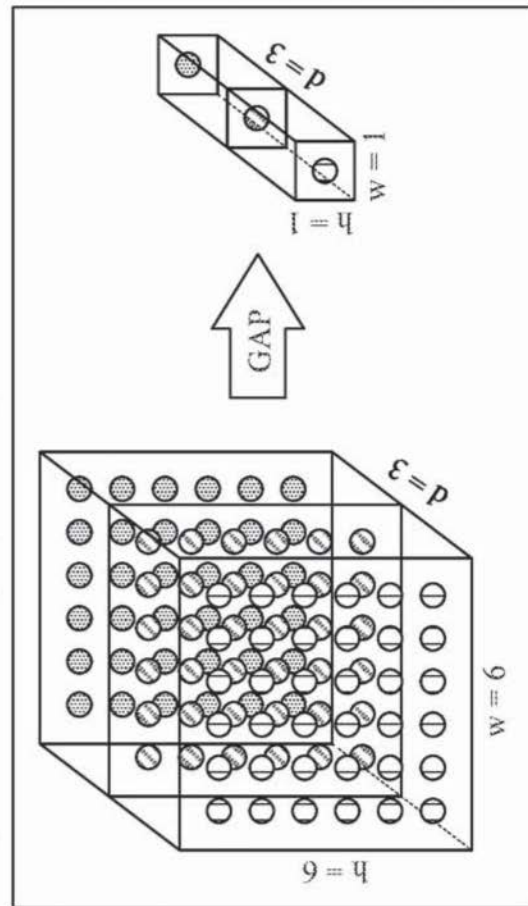


图17

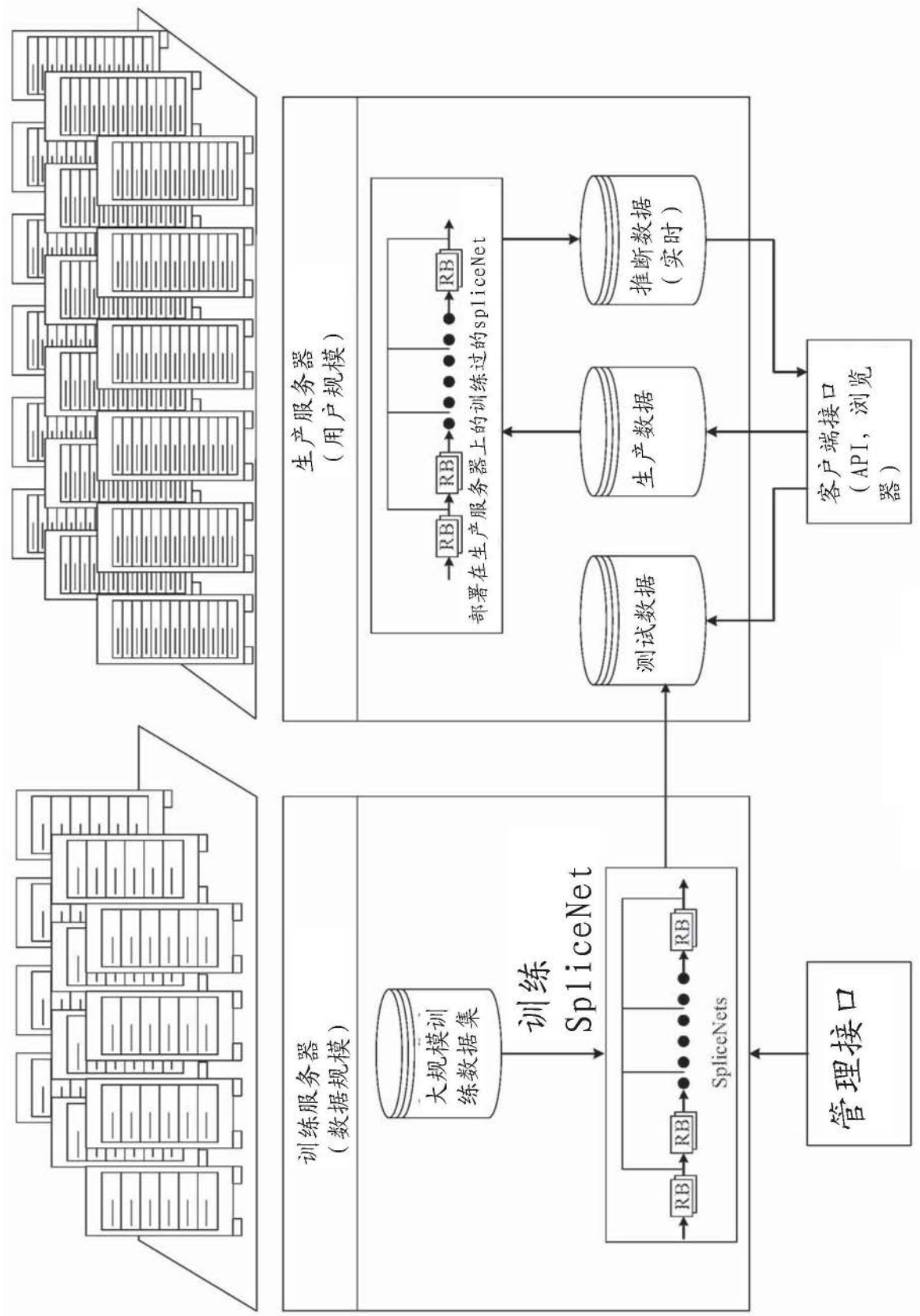


图18

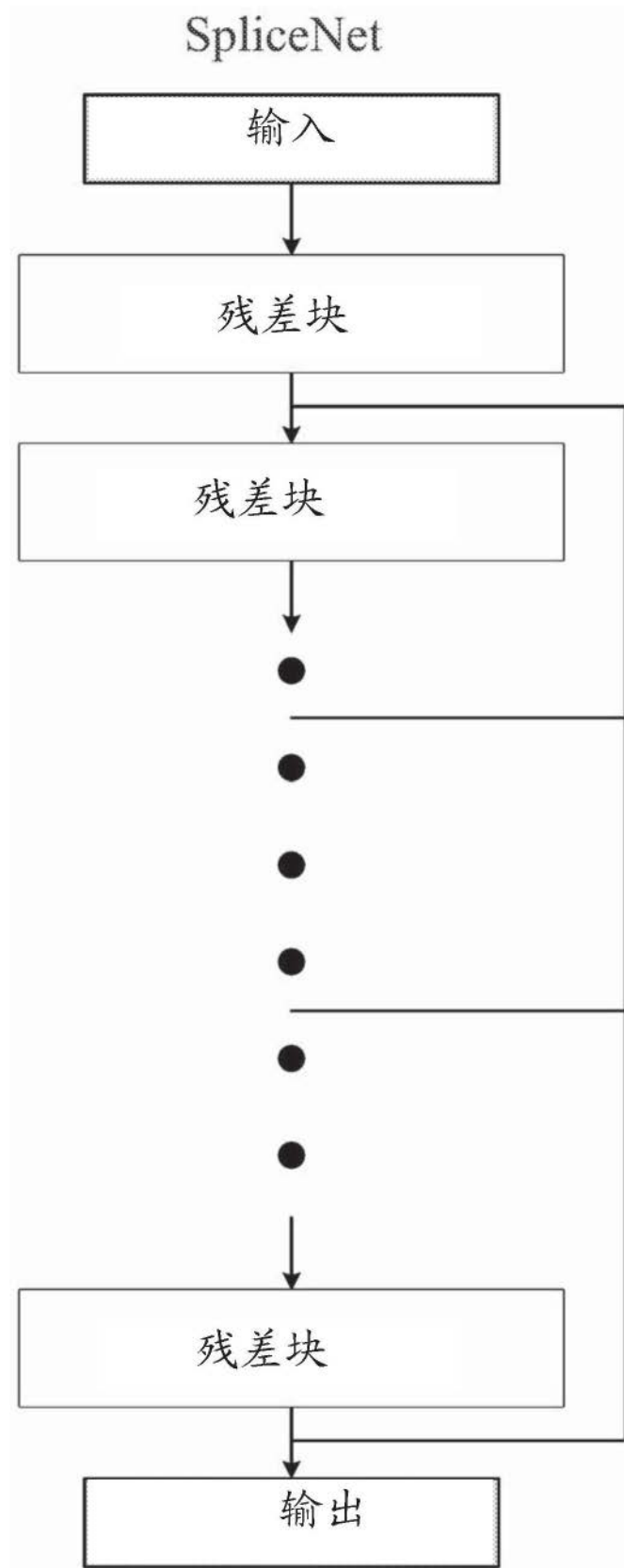


图19

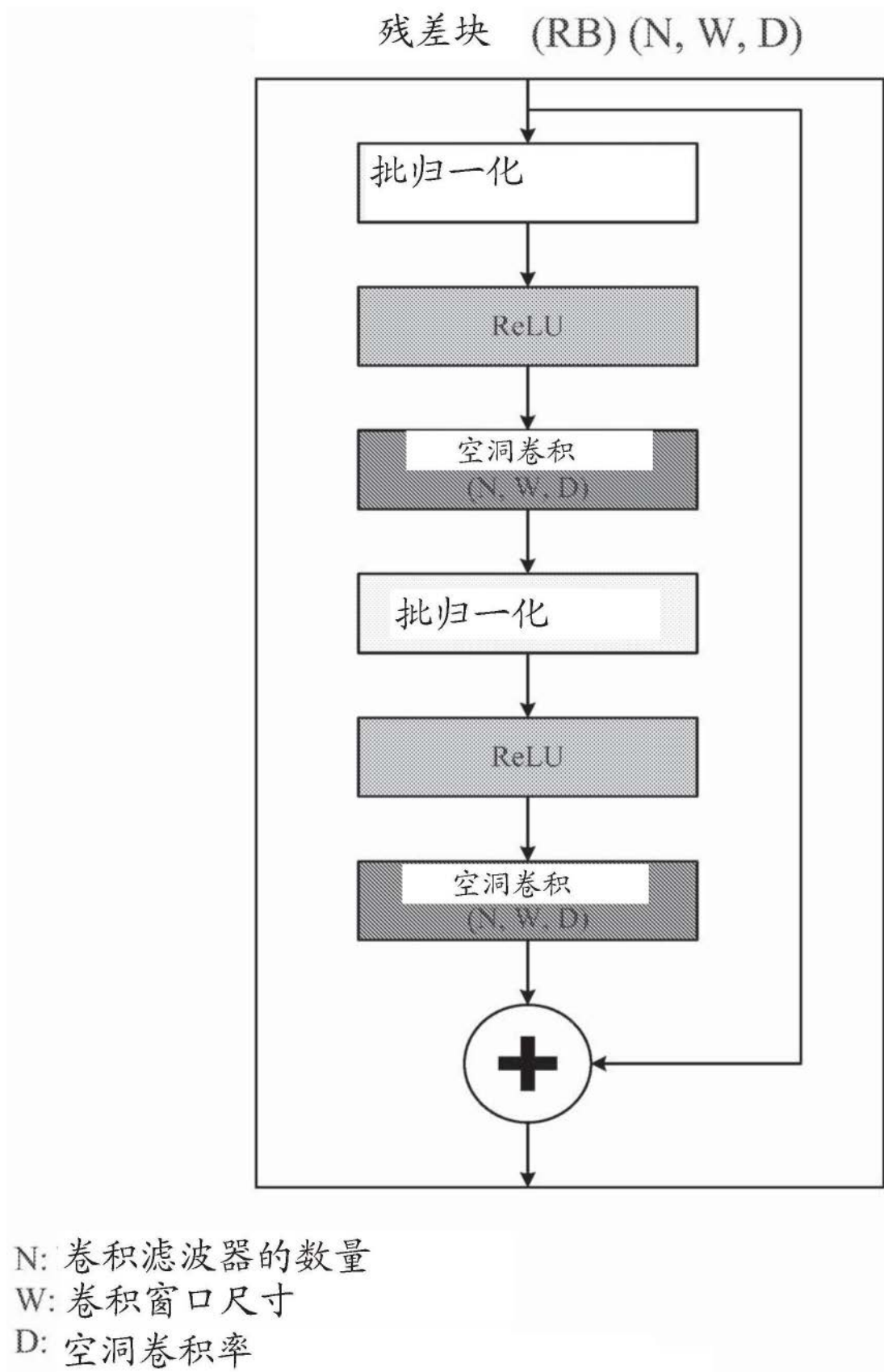


图20

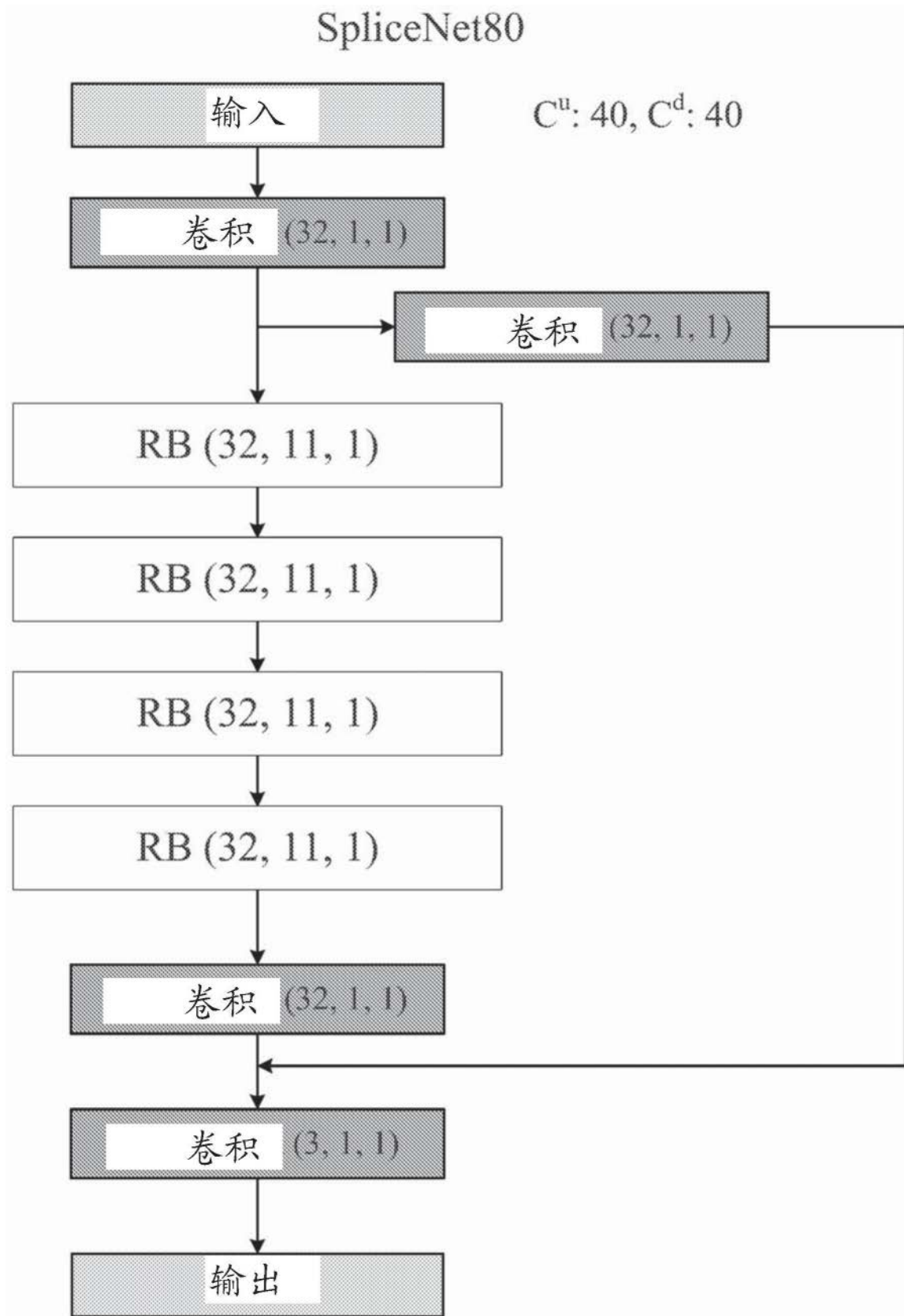


图21

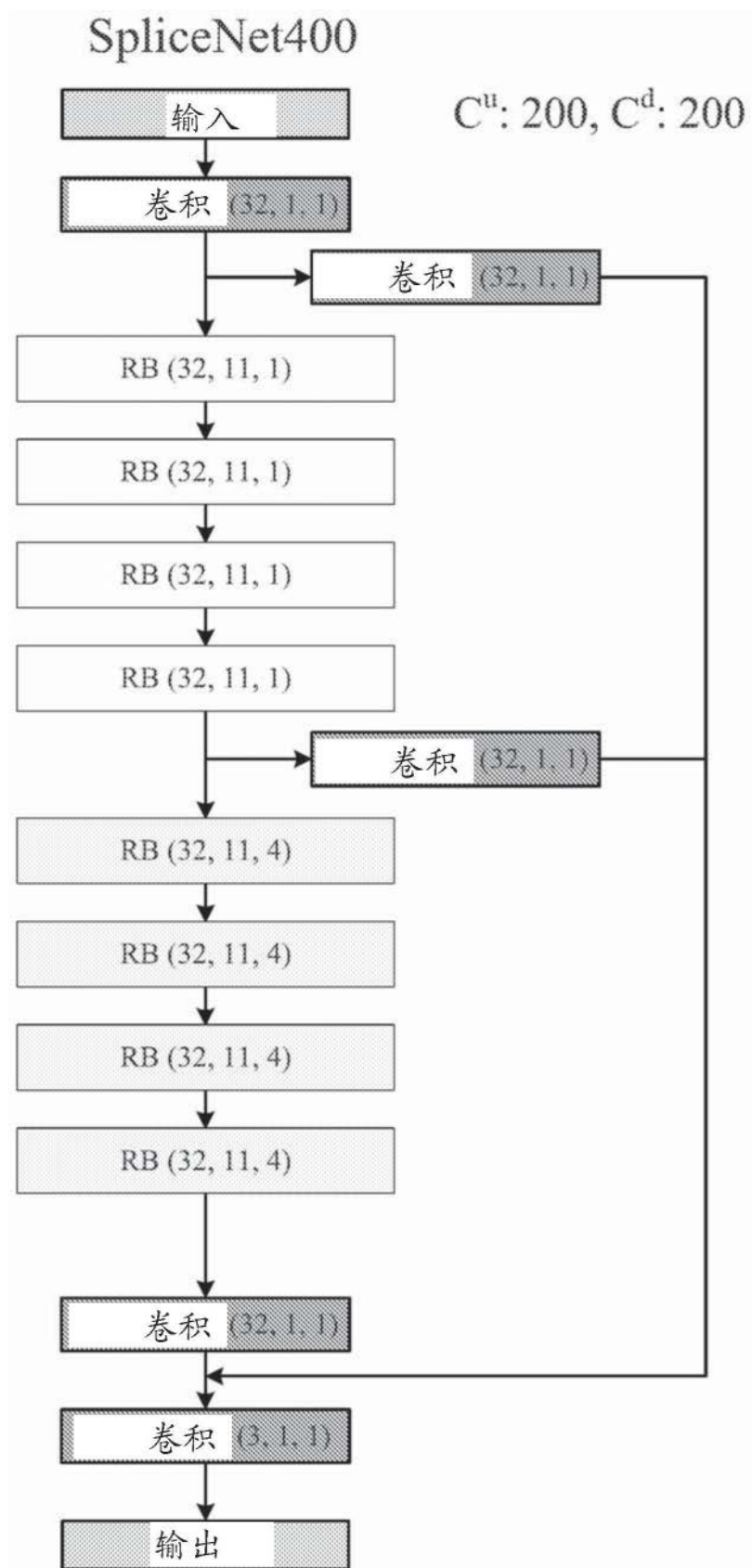


图22

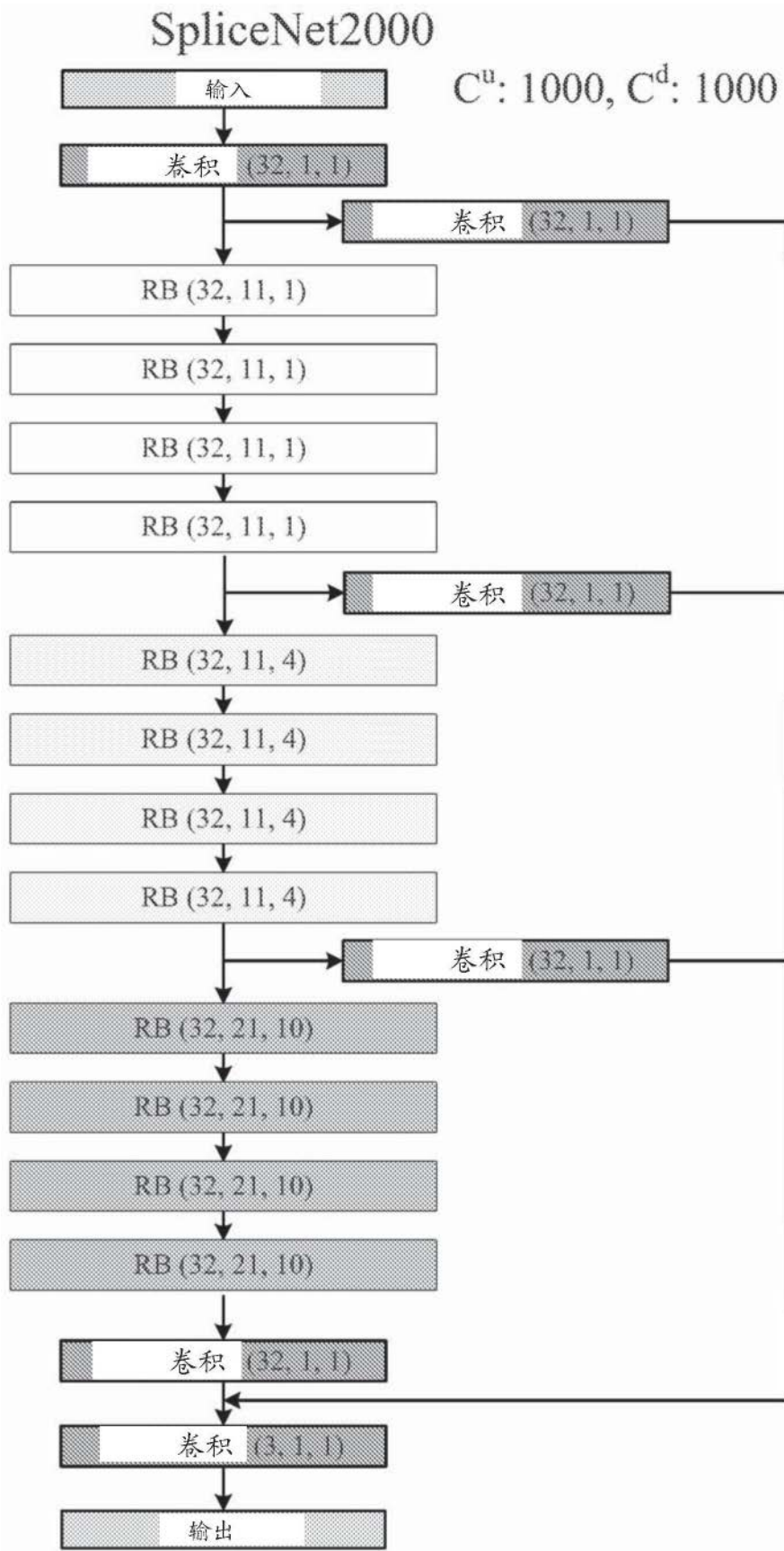


图23

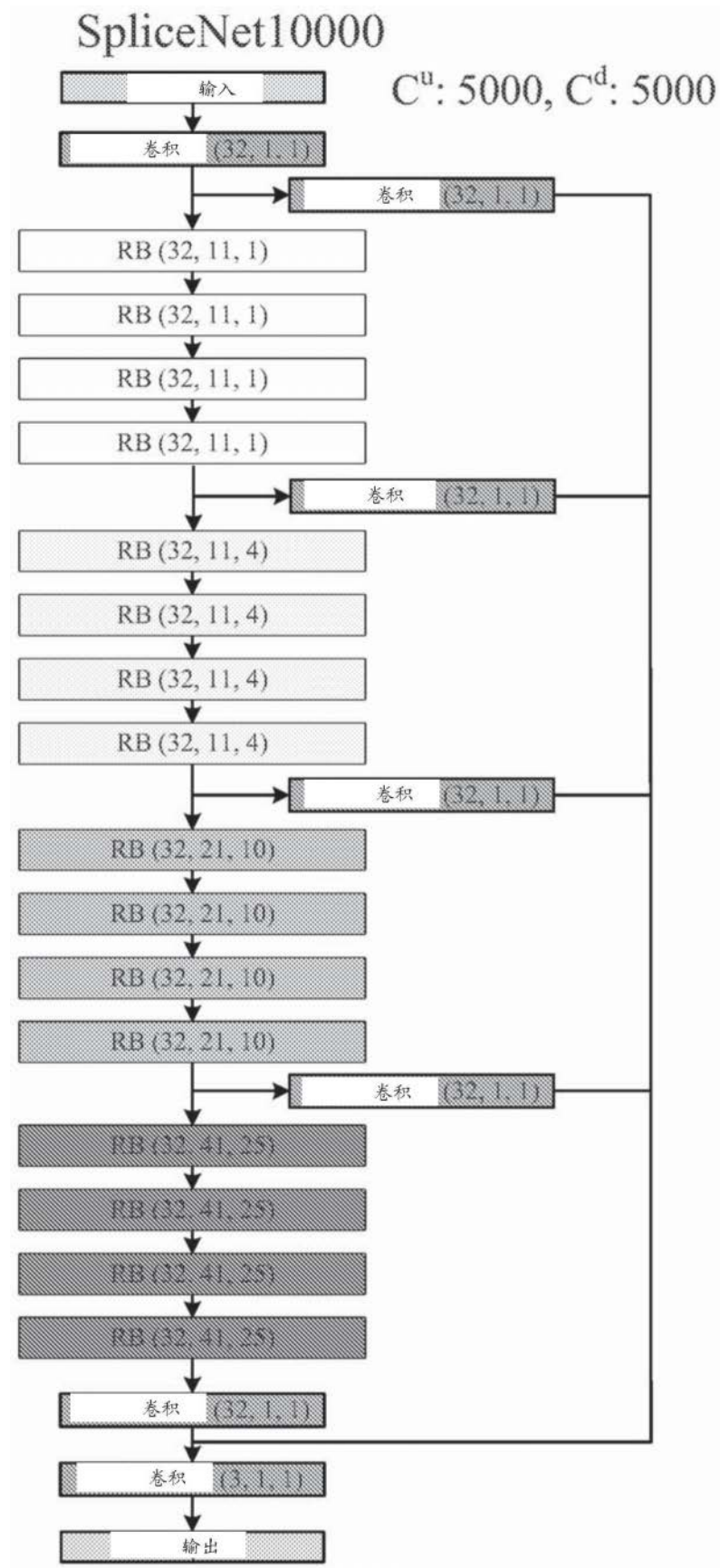


图24

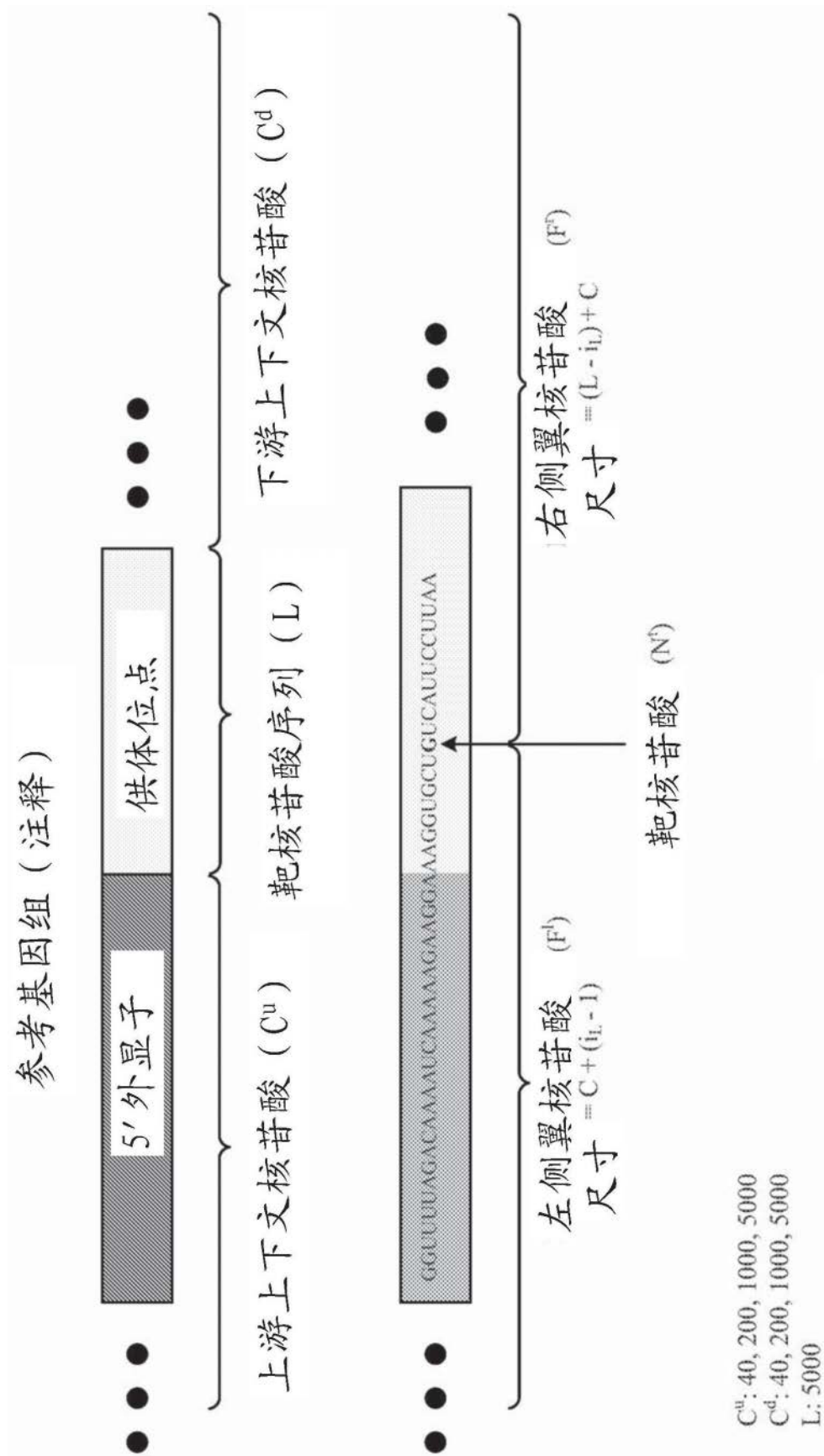


图25

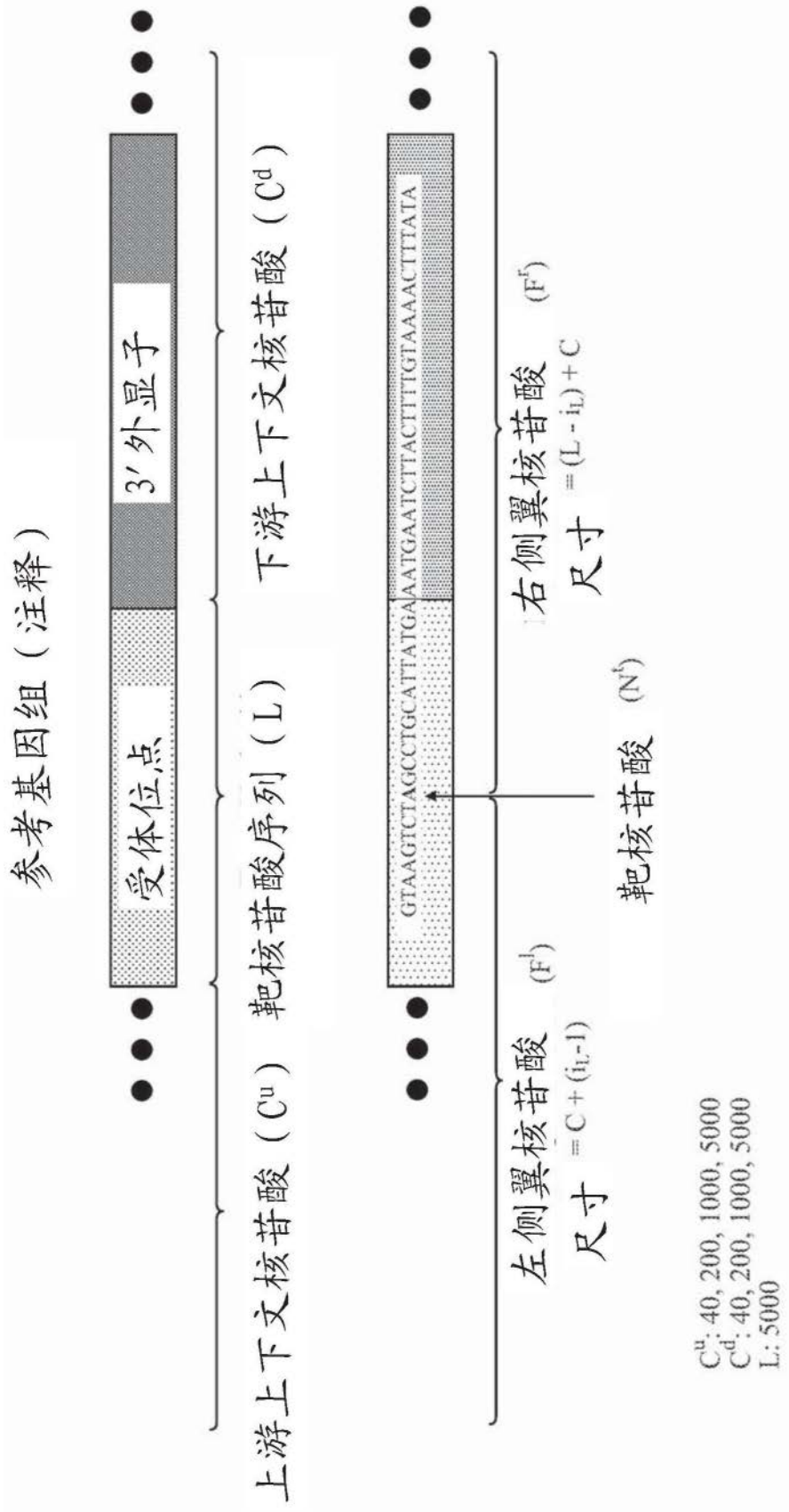


图26

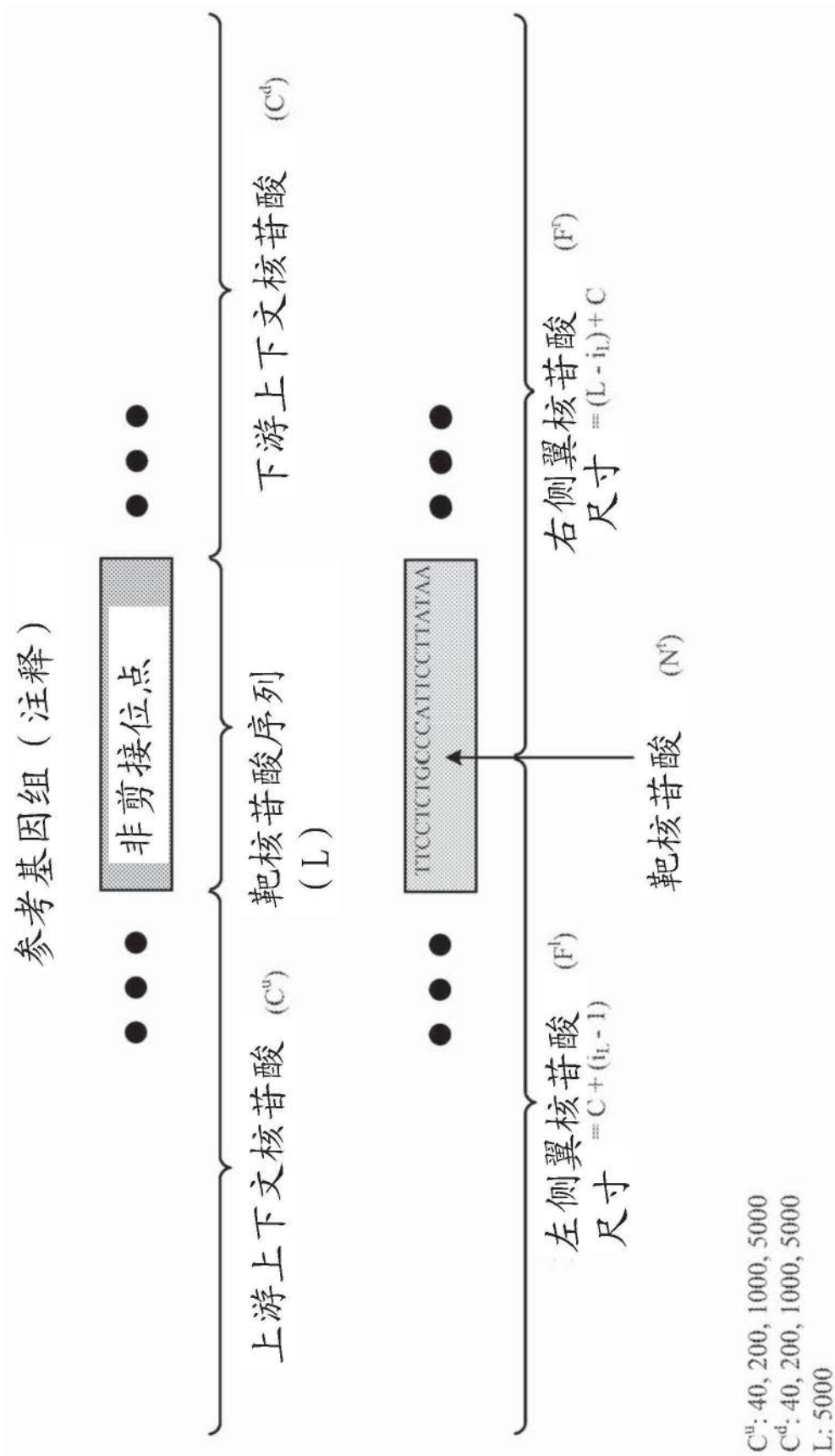


图27

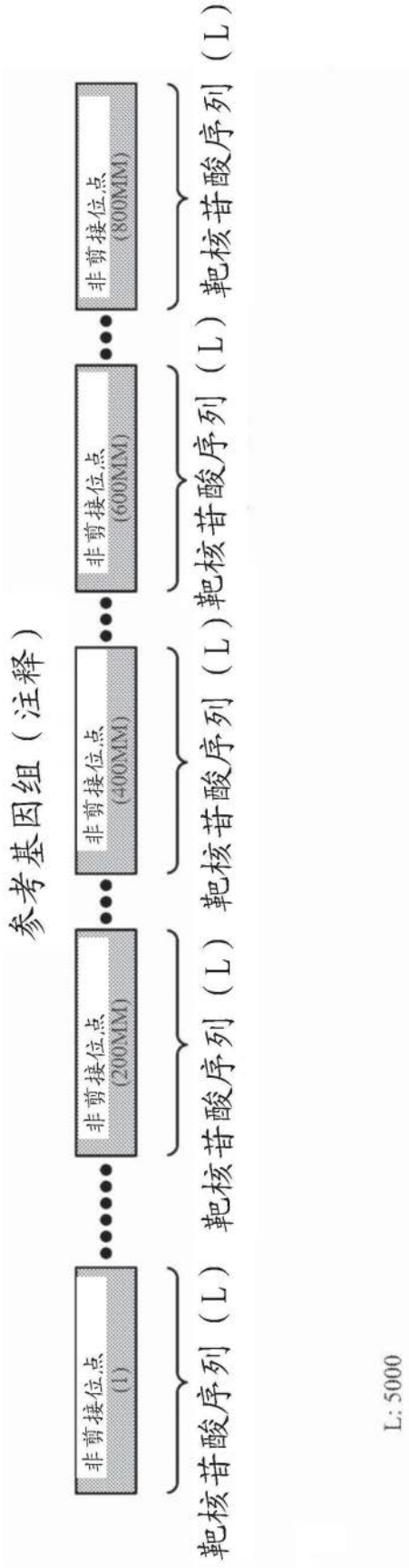


图28

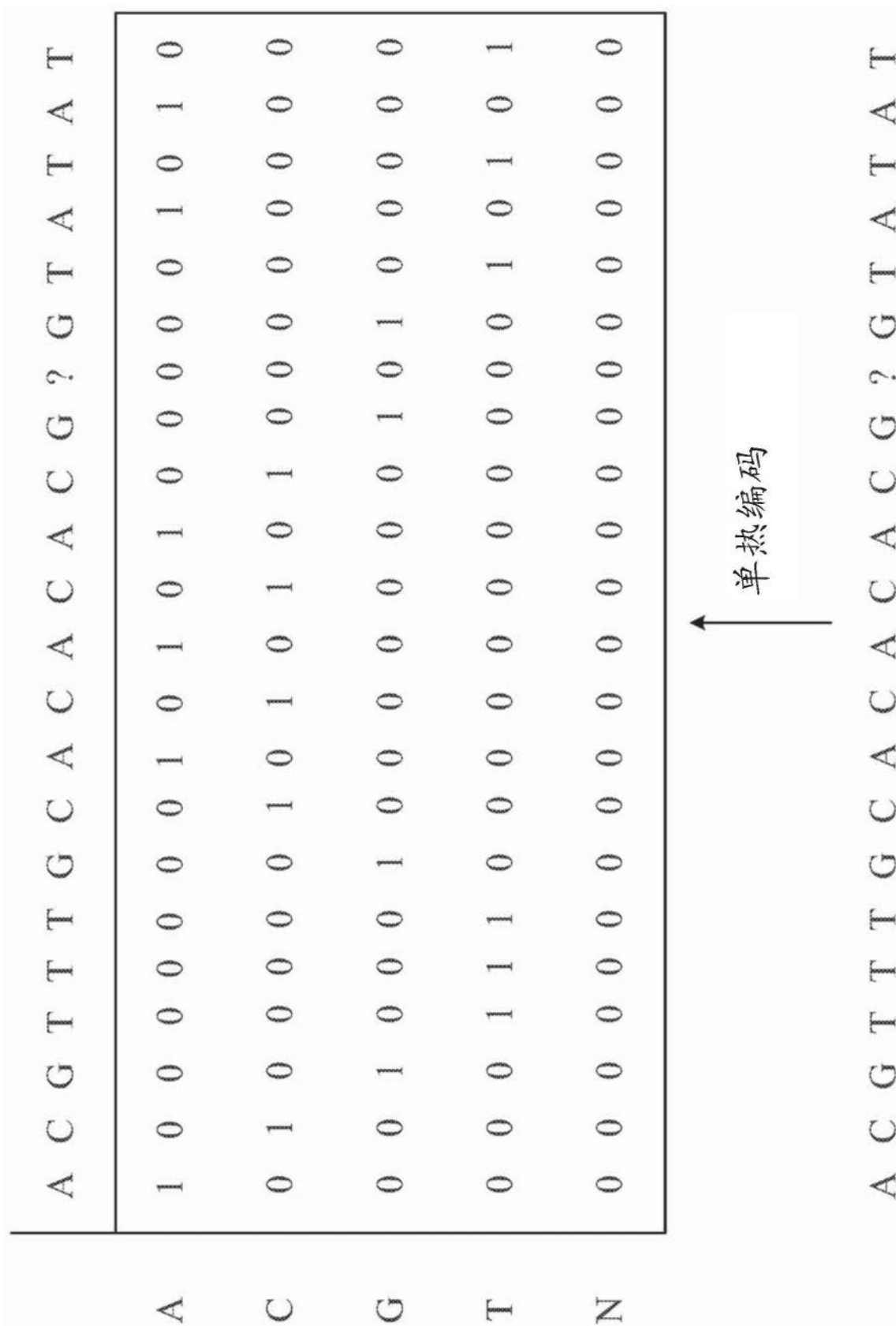


图29

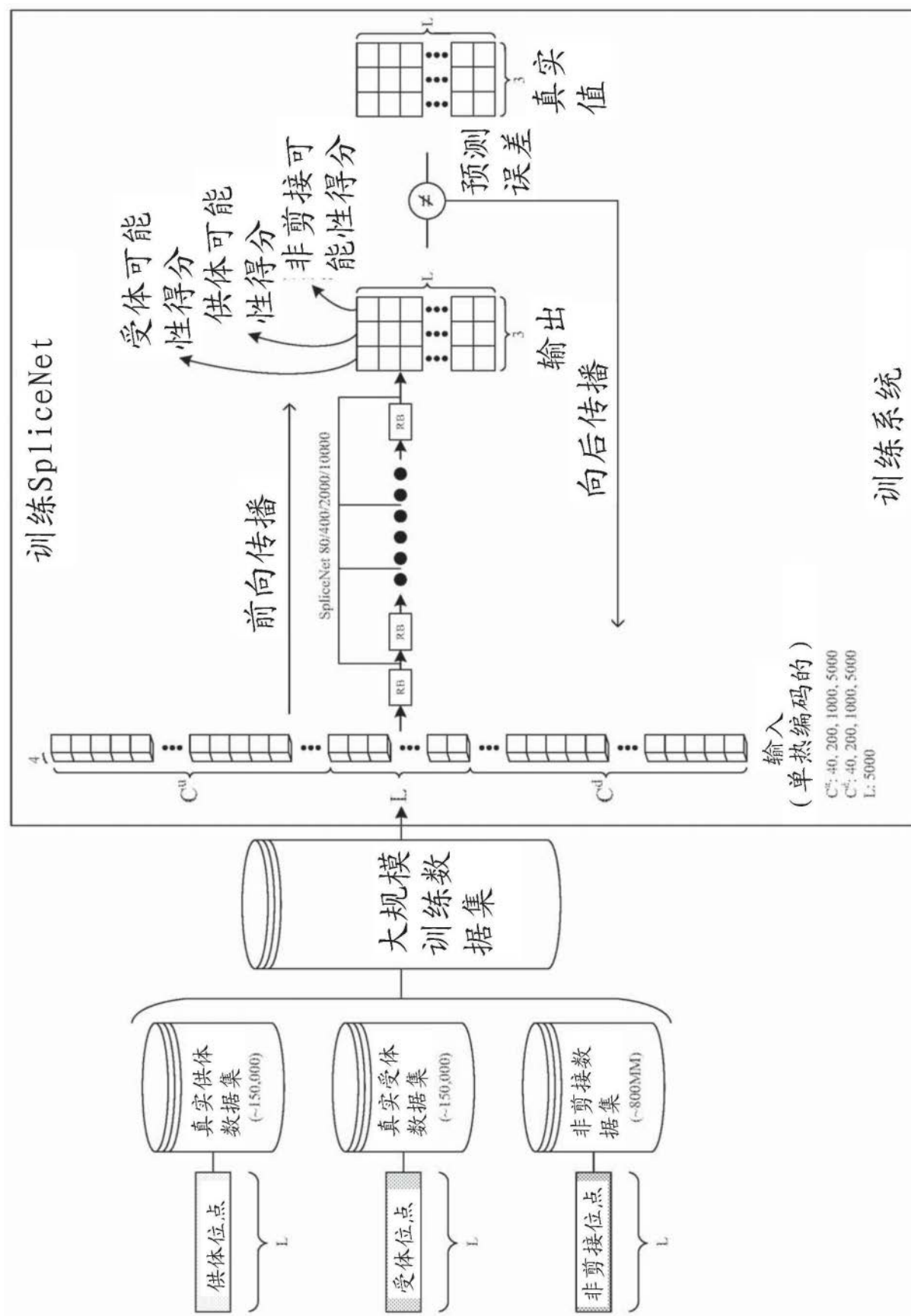


图30

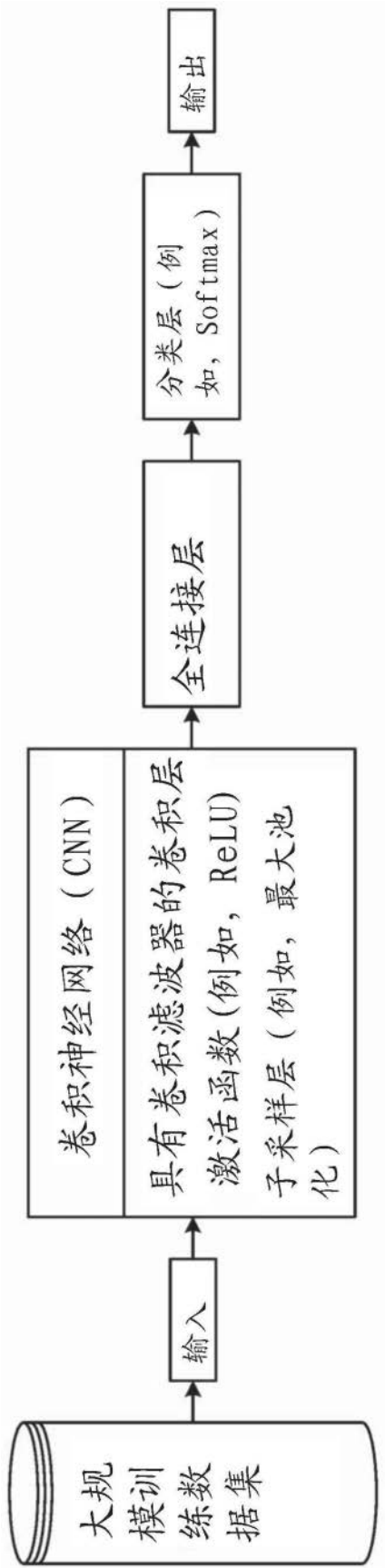


图31

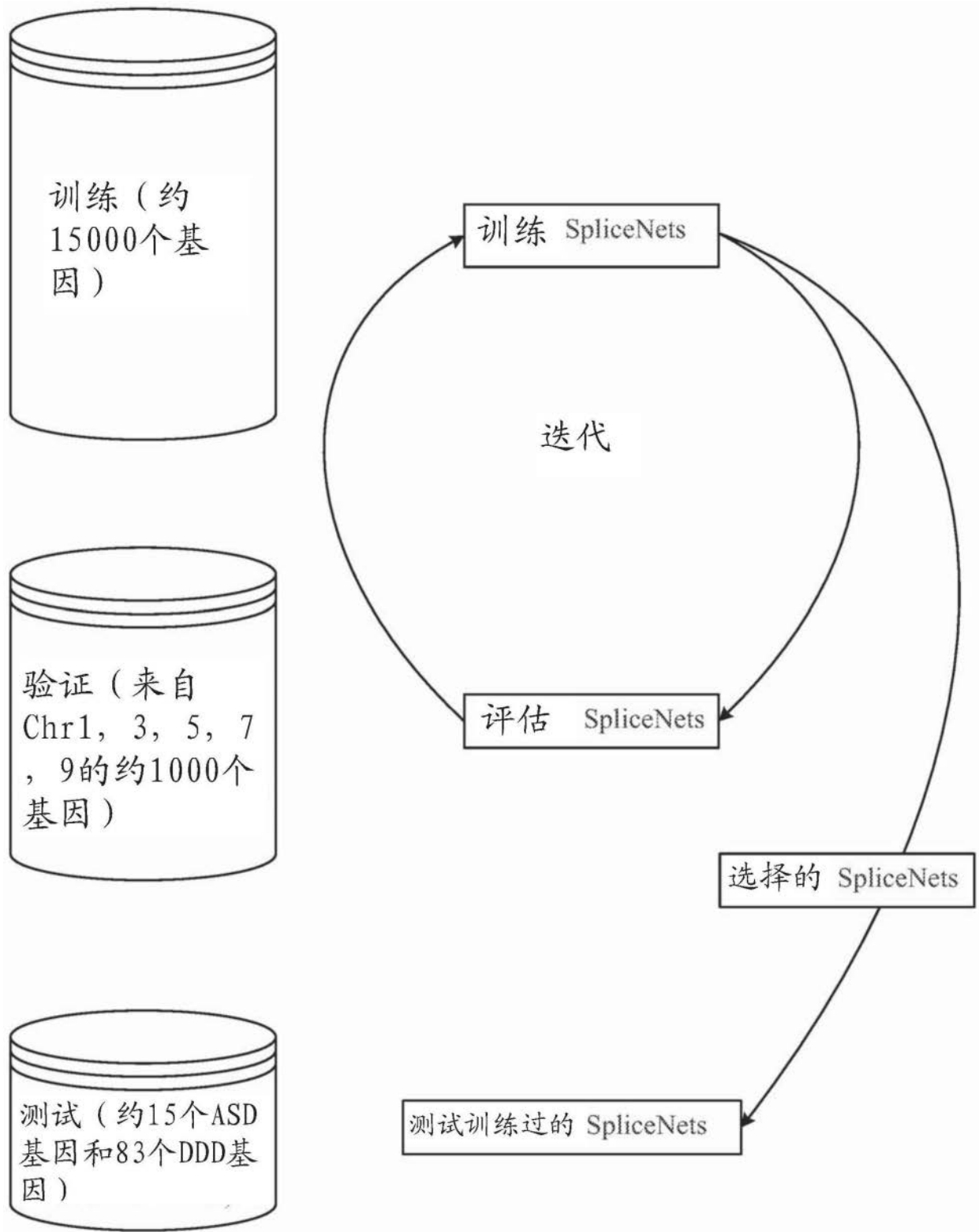


图32

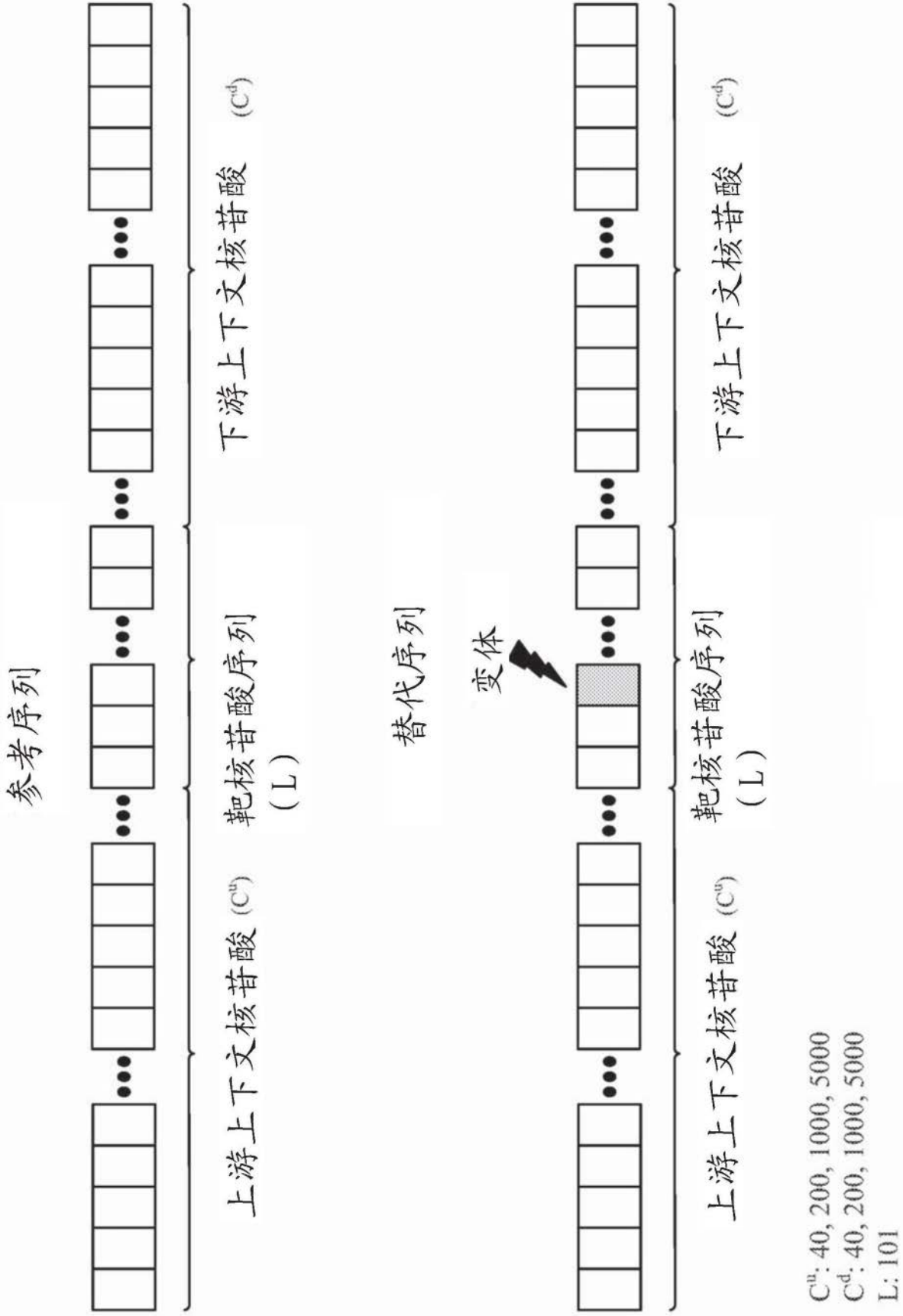


图33

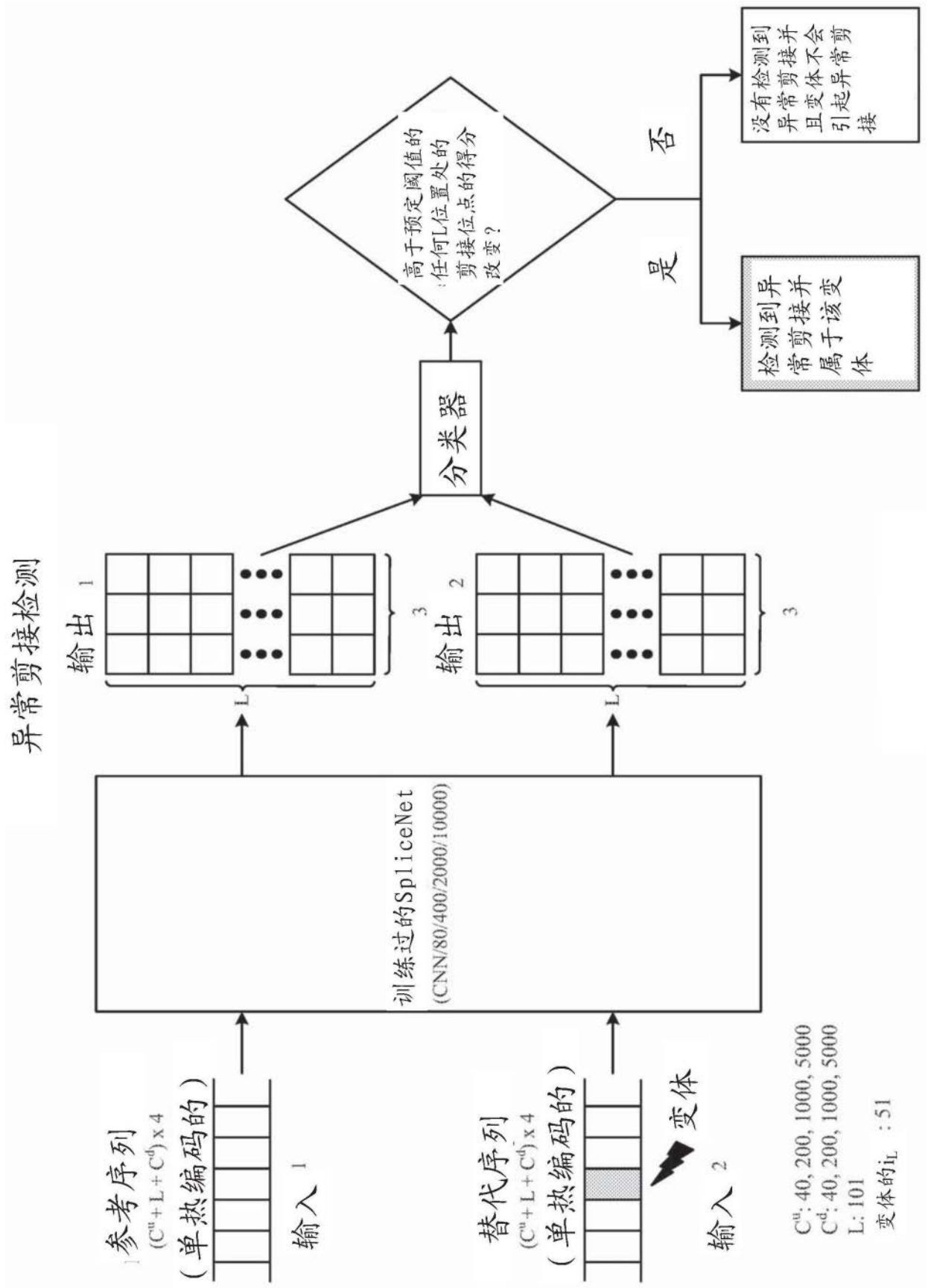


图34

用于剪接位点分类的SpliceNet10000处理金字塔

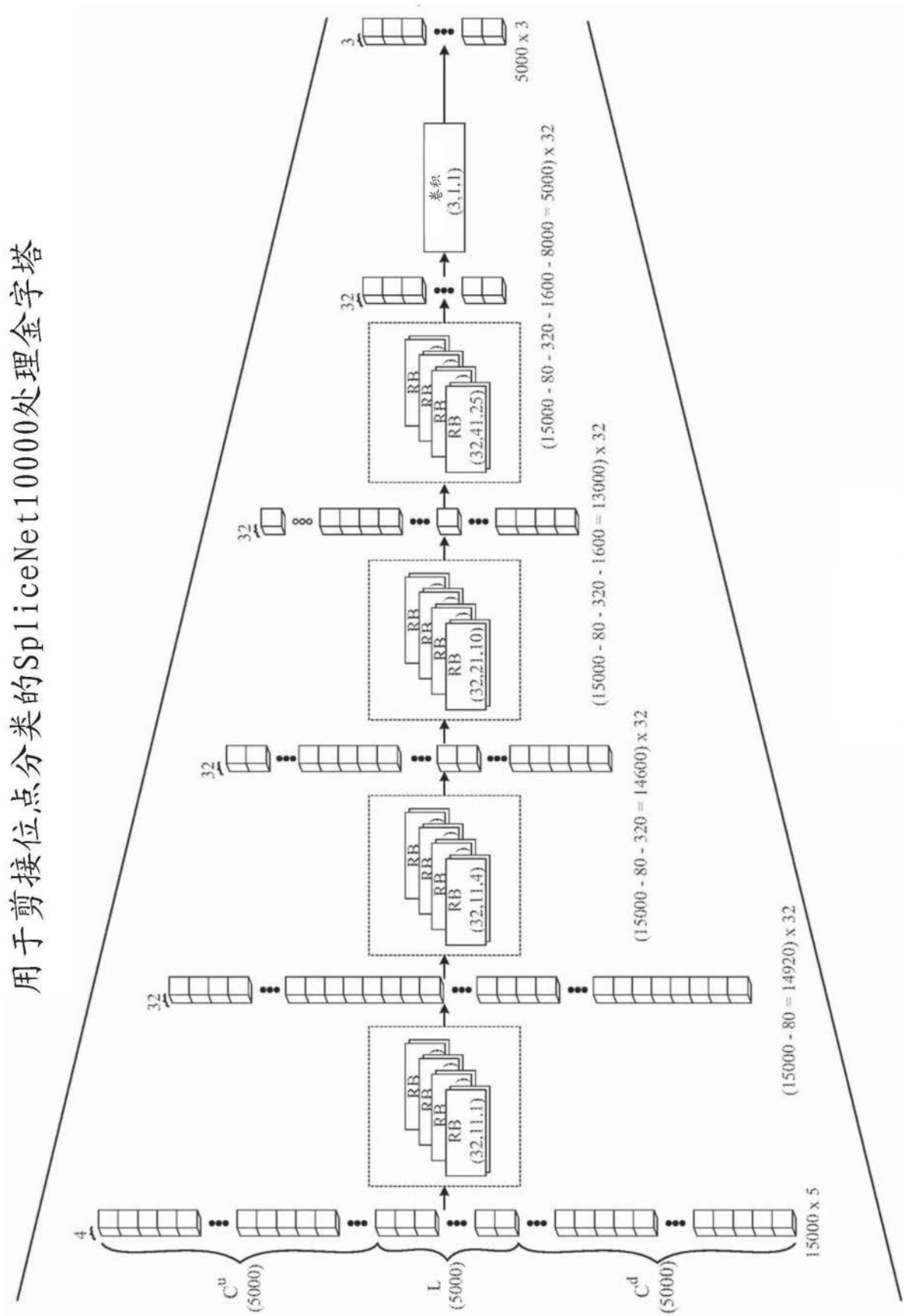


图35

用于异常剪接检测的SpliceNet10000处理金字塔

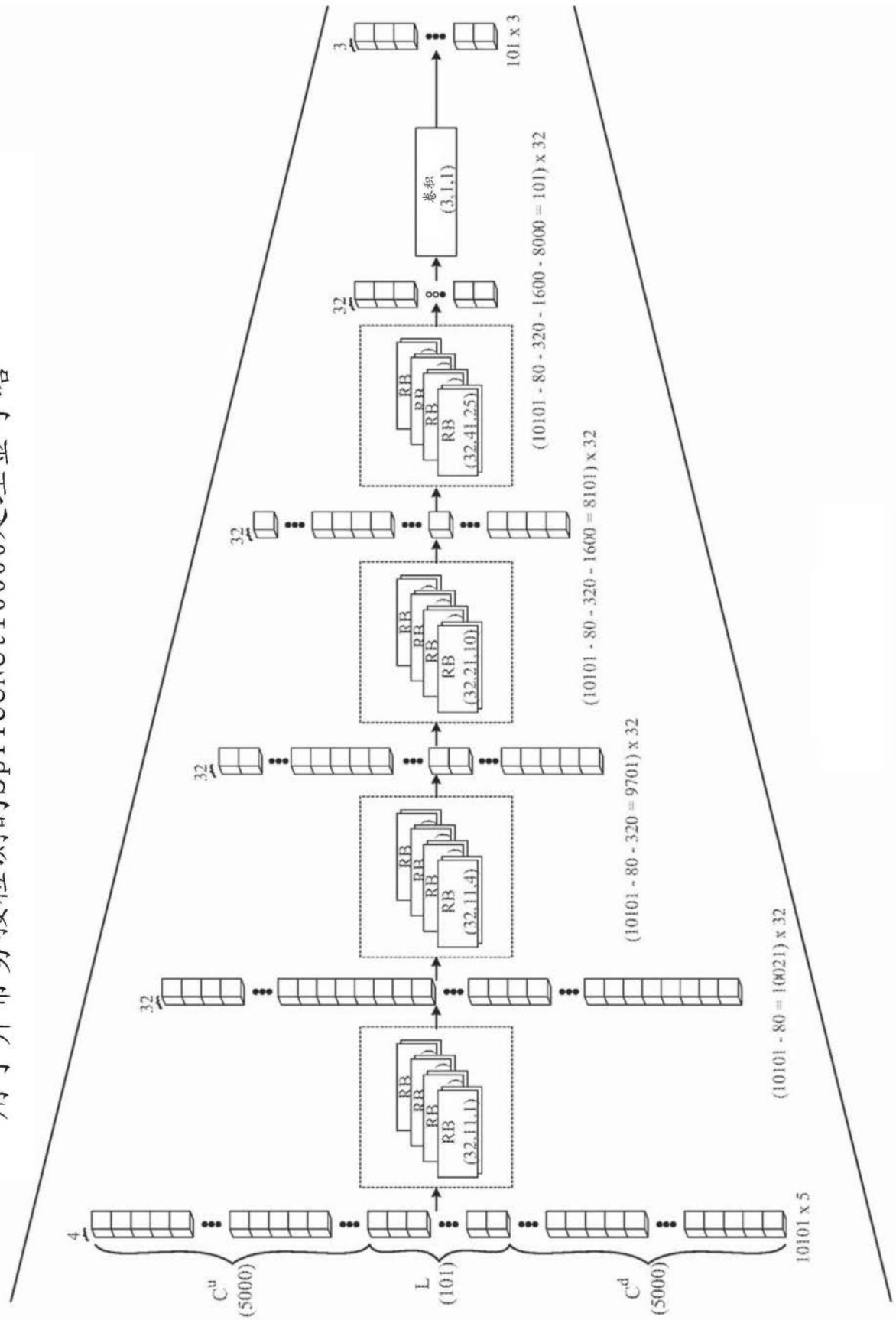


图36

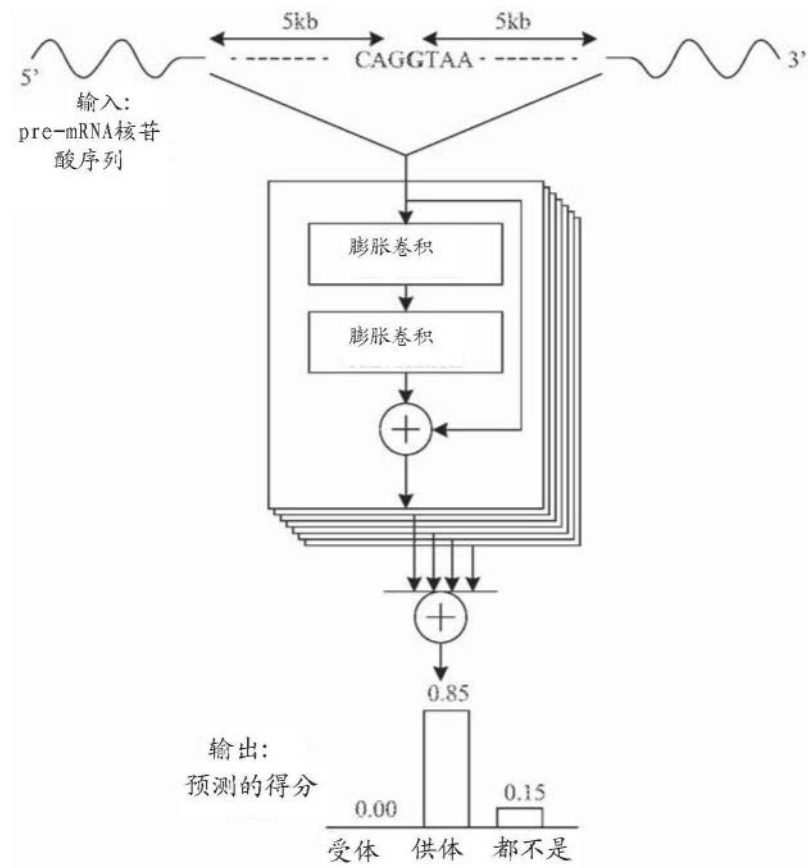


图37A

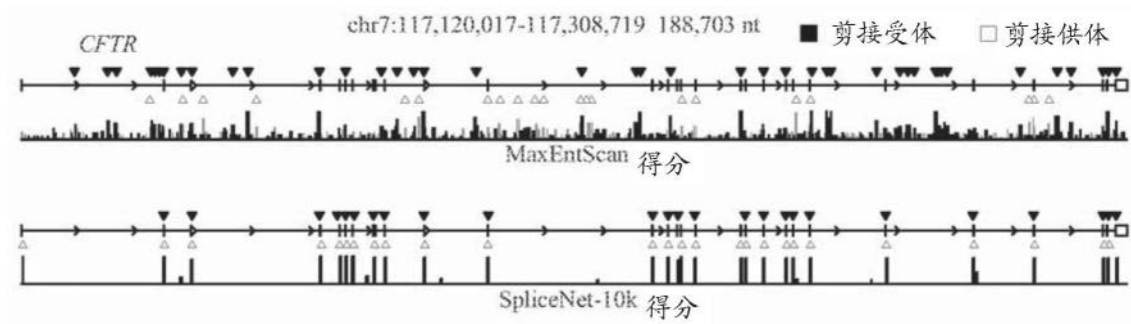


图37B

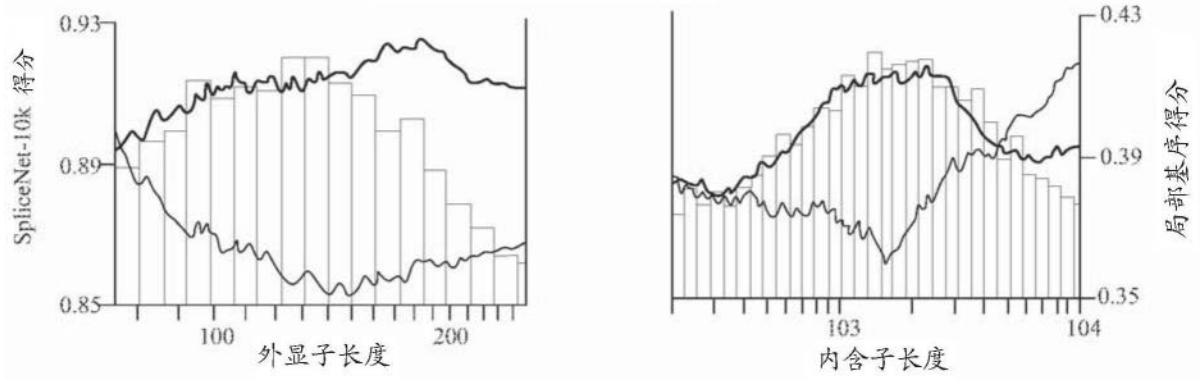


图37F

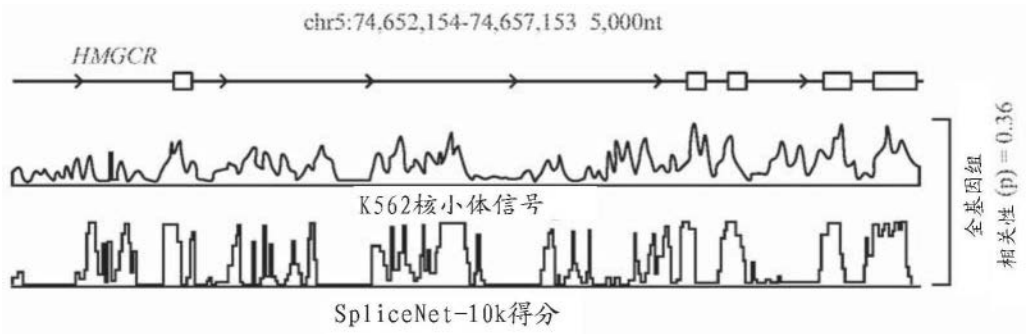


图37G

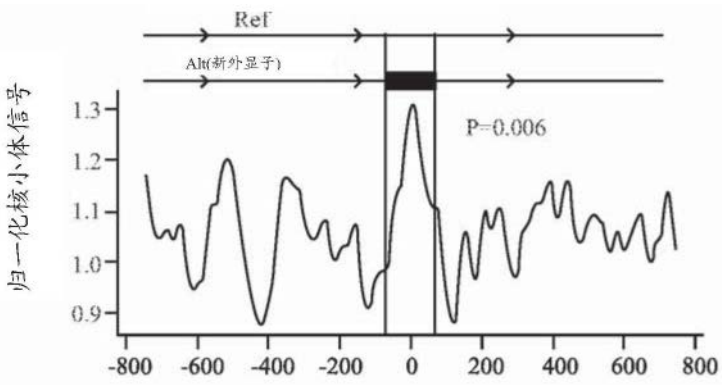


图37H

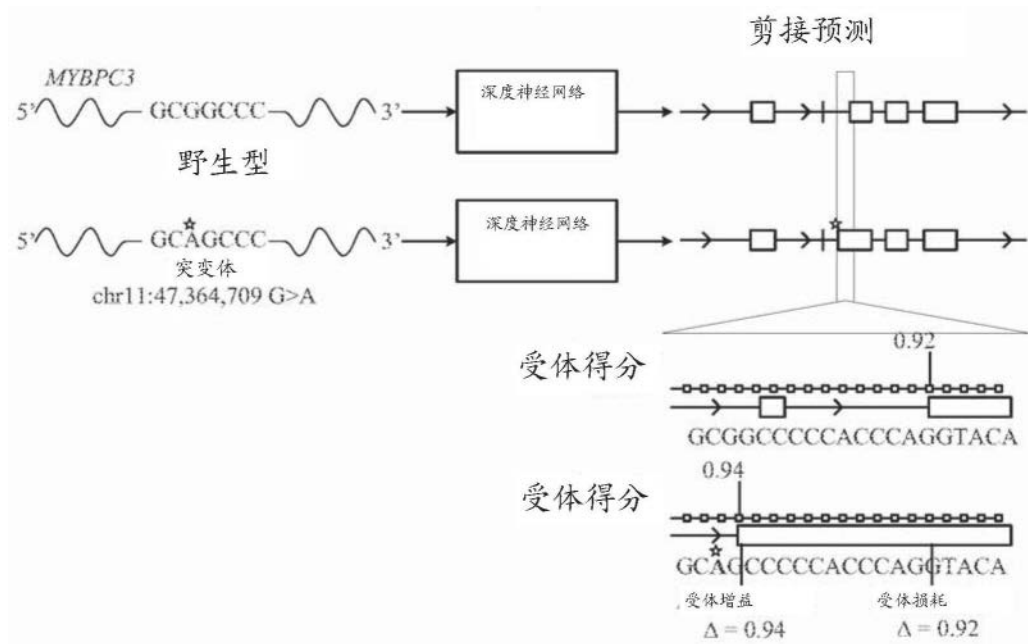


图38A

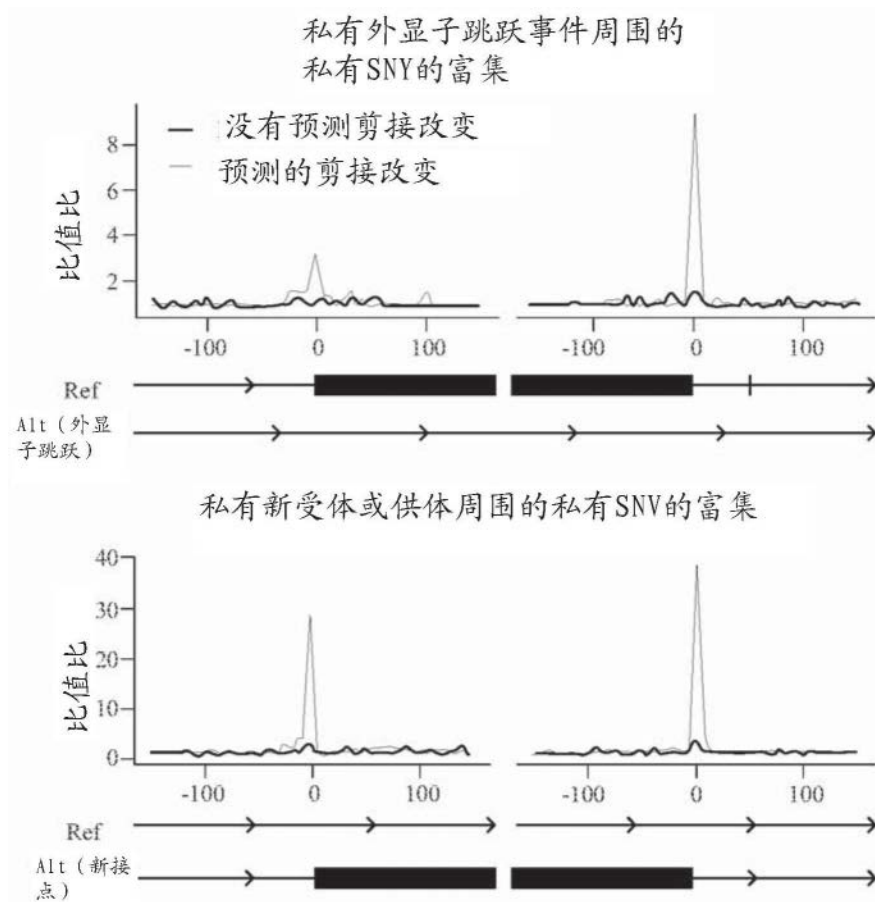


图38B

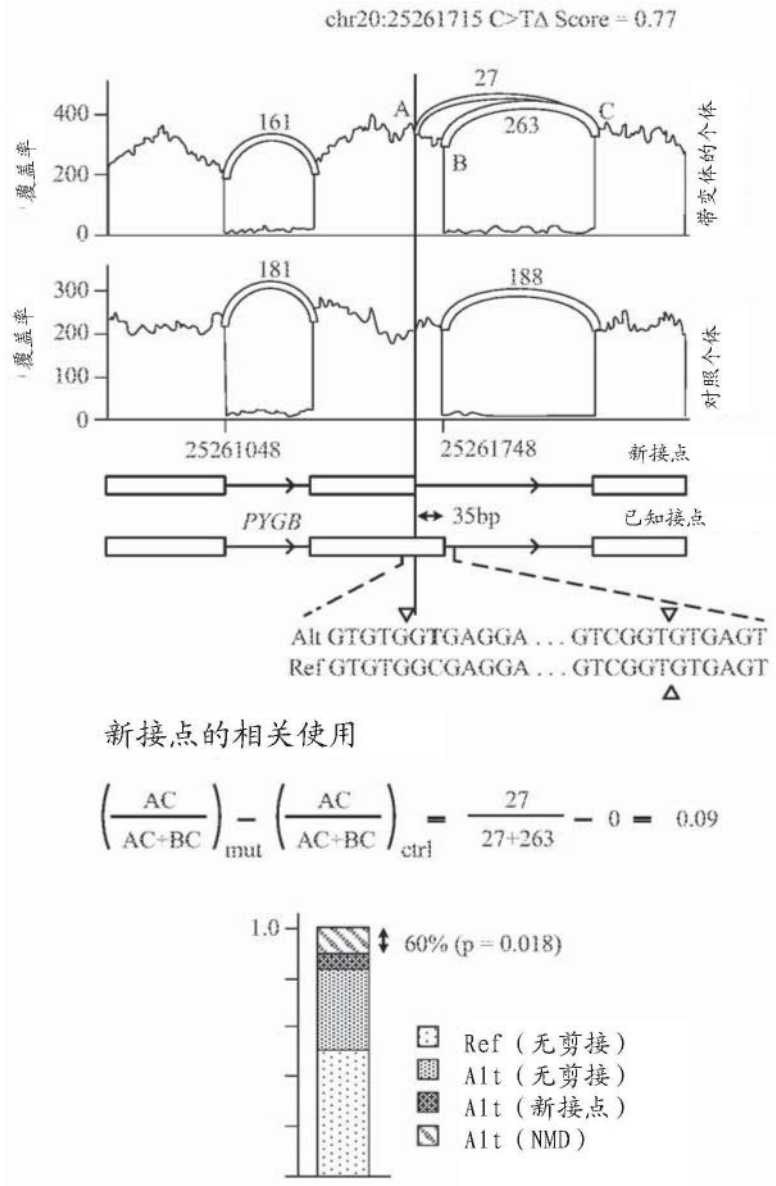


图38C

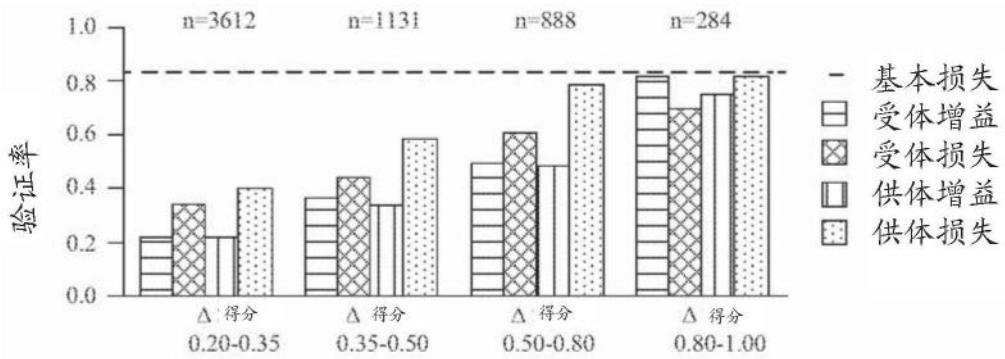


图38D

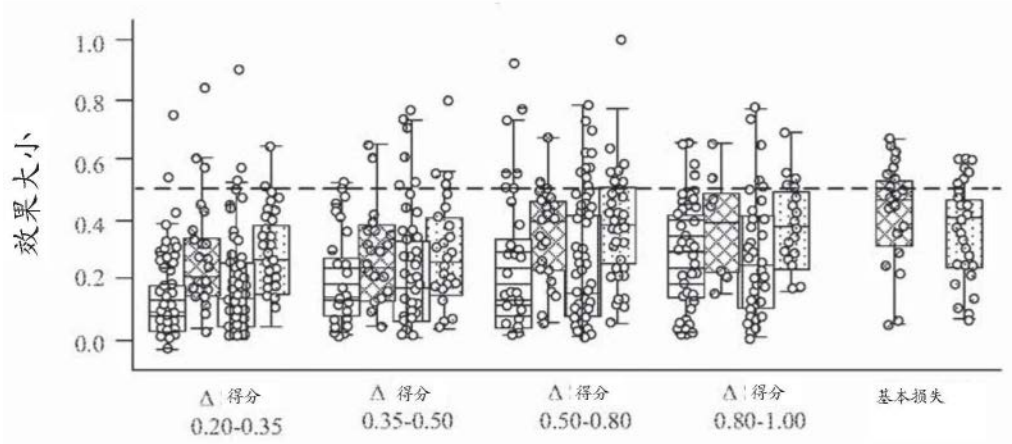


图38E

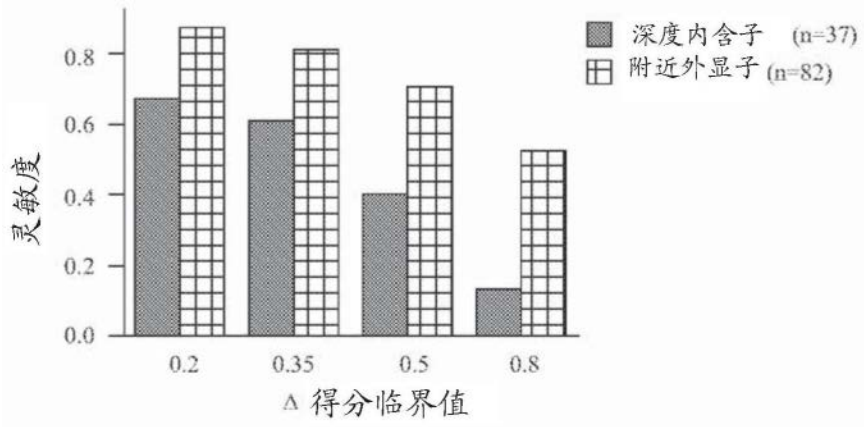


图38F

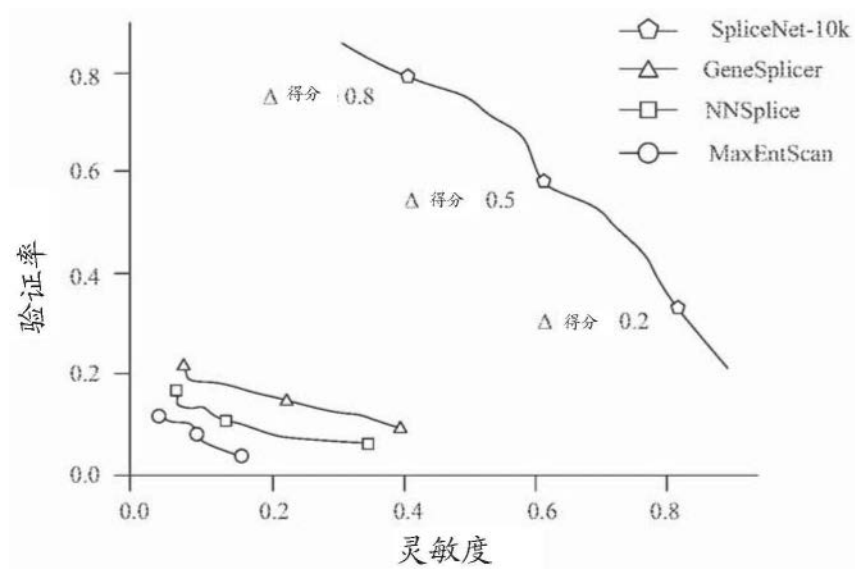


图38G

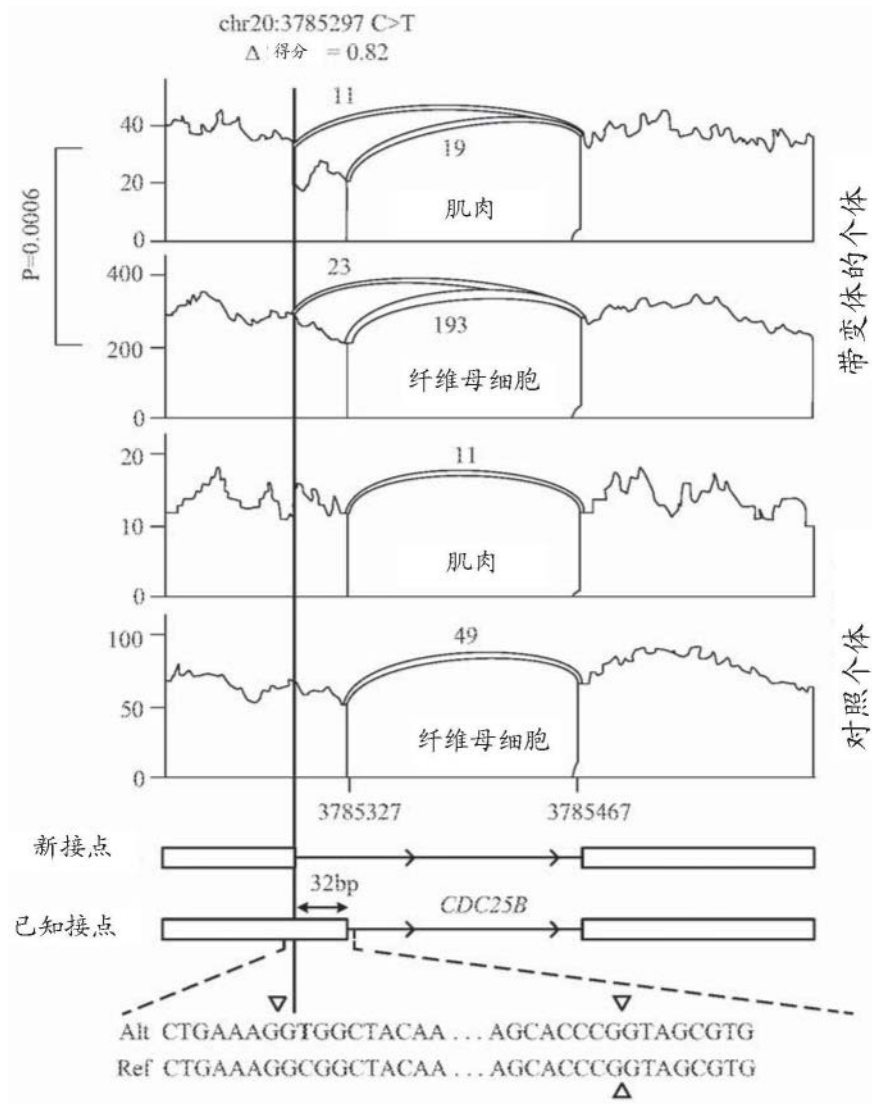


图39A

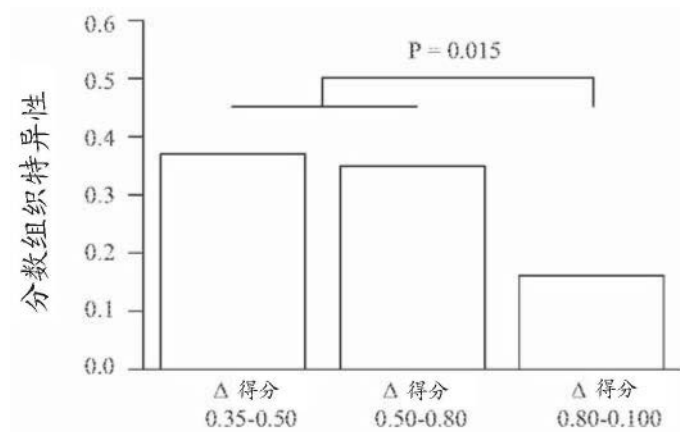


图39C

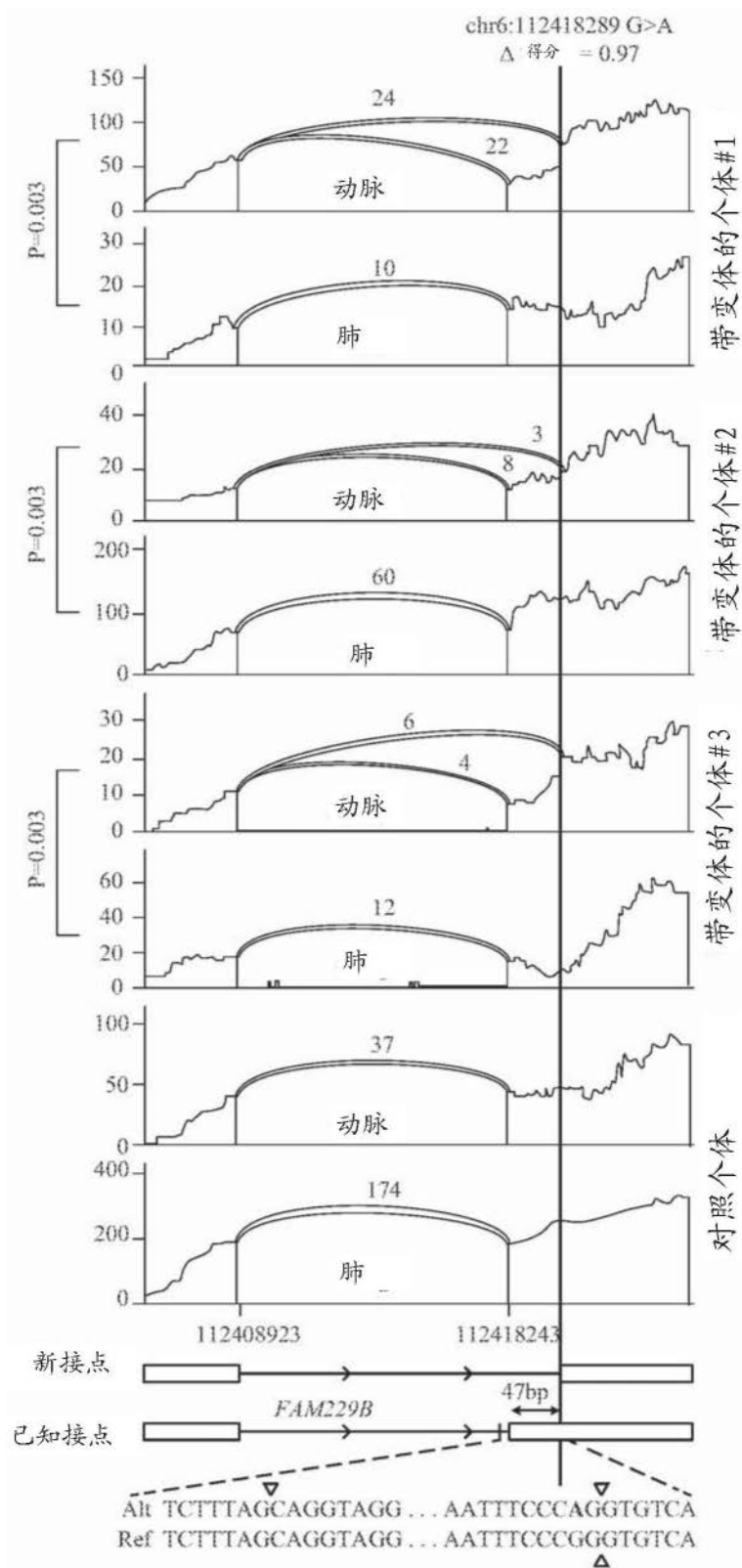


图39B

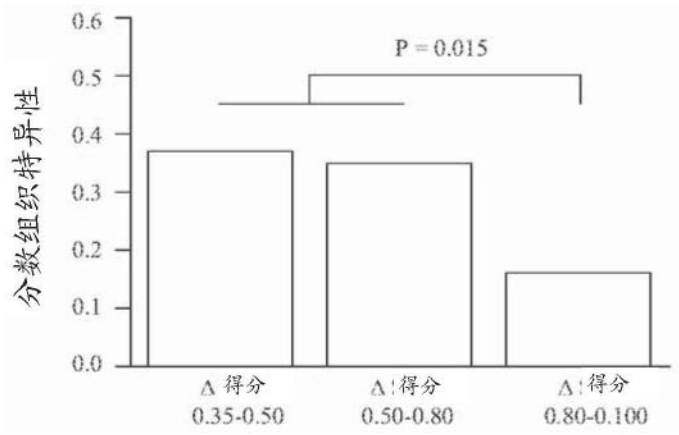


图39C

外显子变体附近的ExAC	
单例	常见 (AF ≥ 0.1%)
具有Δ得分 ≥ 0.8 的SNV	10,369 212
具有Δ得分 < 0.1 的SNV	1,687,004 158,177
比值比 (OR) = 4.58 (P < 10 ⁻¹²⁷) 有害变体的分数 = (1-1/OR)	

图40A

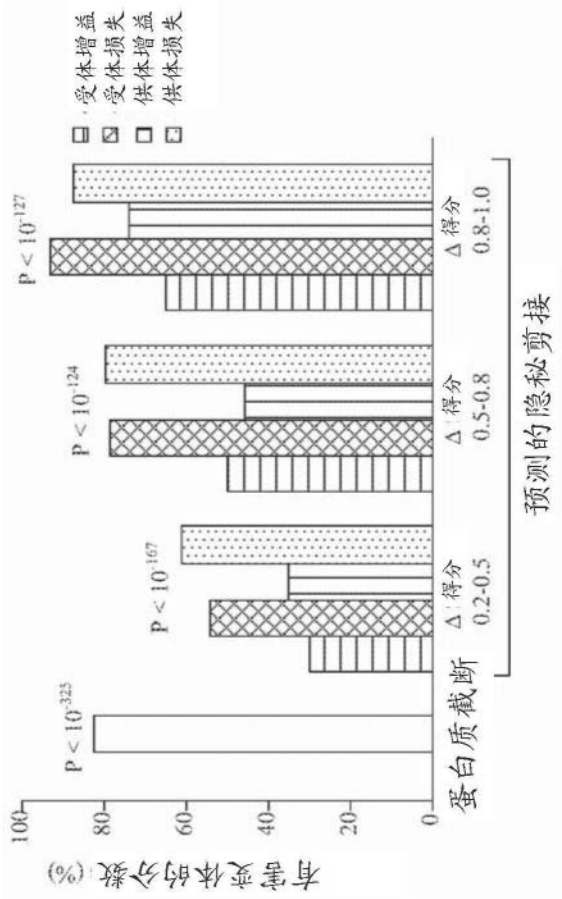


图40B

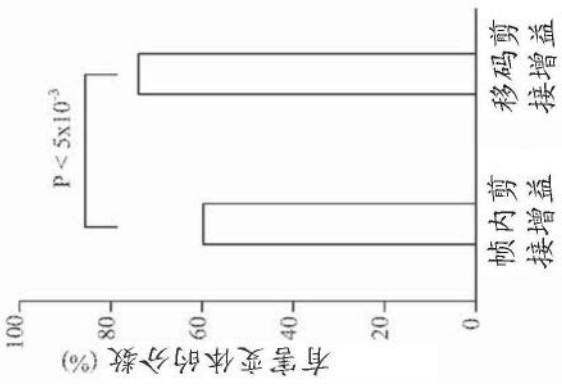


图40C

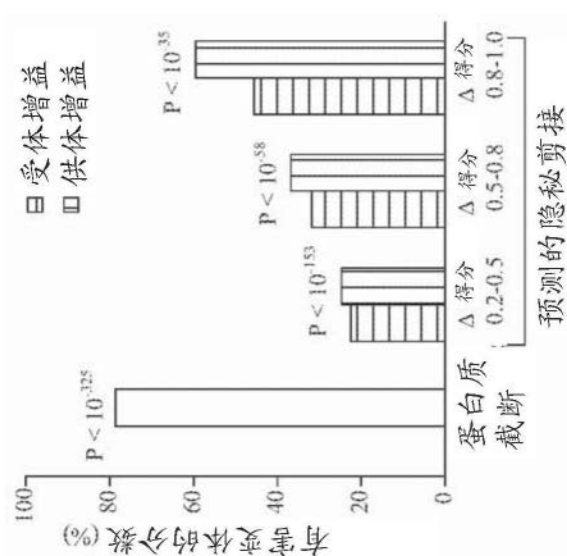


图40D

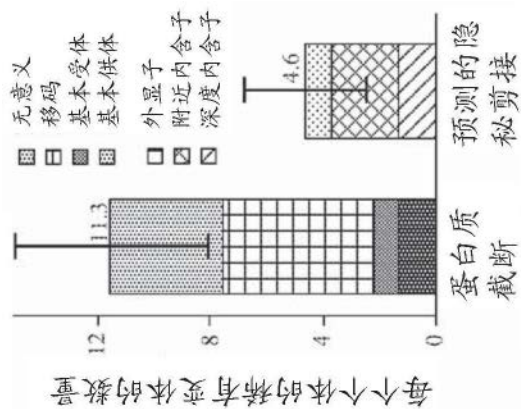


图40E

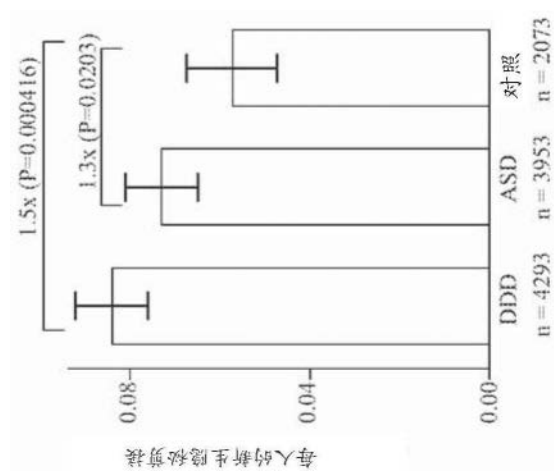


图41A

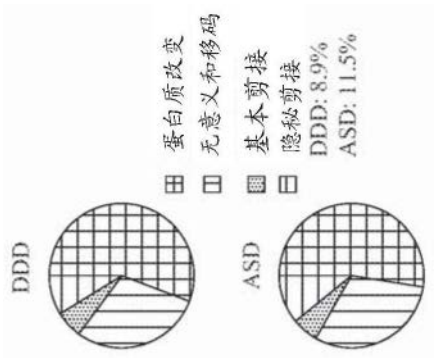


图41B

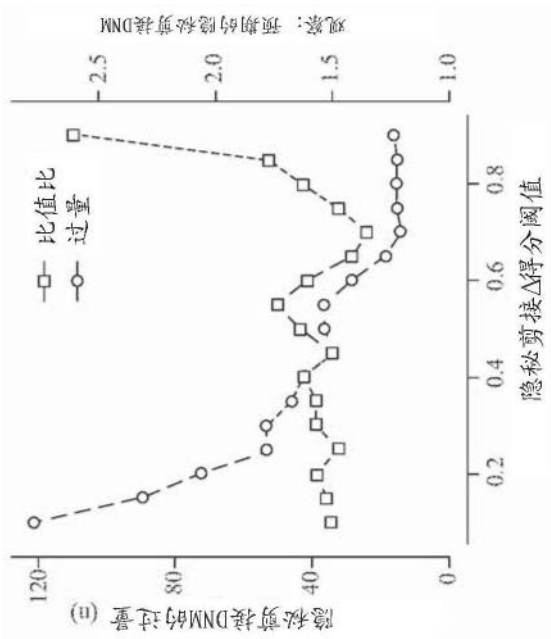


图41C

符号	PTV	错误的	隐秘的	P	表型
CTNND1	2	0	1	4.7×10^{-5}	发育迟缓
HNRNPK	2	1	1	9.7×10^{-6}	发育迟缓, 脊柱侧凸, 唇腭裂, 膀胱输尿管反流
PHF5A	1	0	1	3.1×10^{-5}	发育迟缓
SNAP25	1	0	2	6.8×10^{-5}	失神发作, 内眦赘皮
TRIP12	3	2	1	8.5×10^{-6}	迟语
KDM6B	3	1	1	4.6×10^{-6}	自闭症
SLC6A8	2	0	1	5.8×10^{-6}	自闭症

图41D

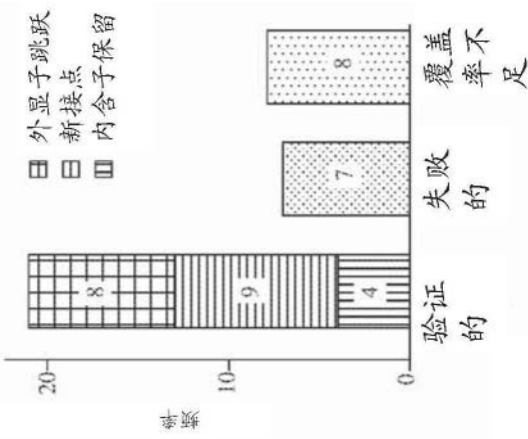


图41F

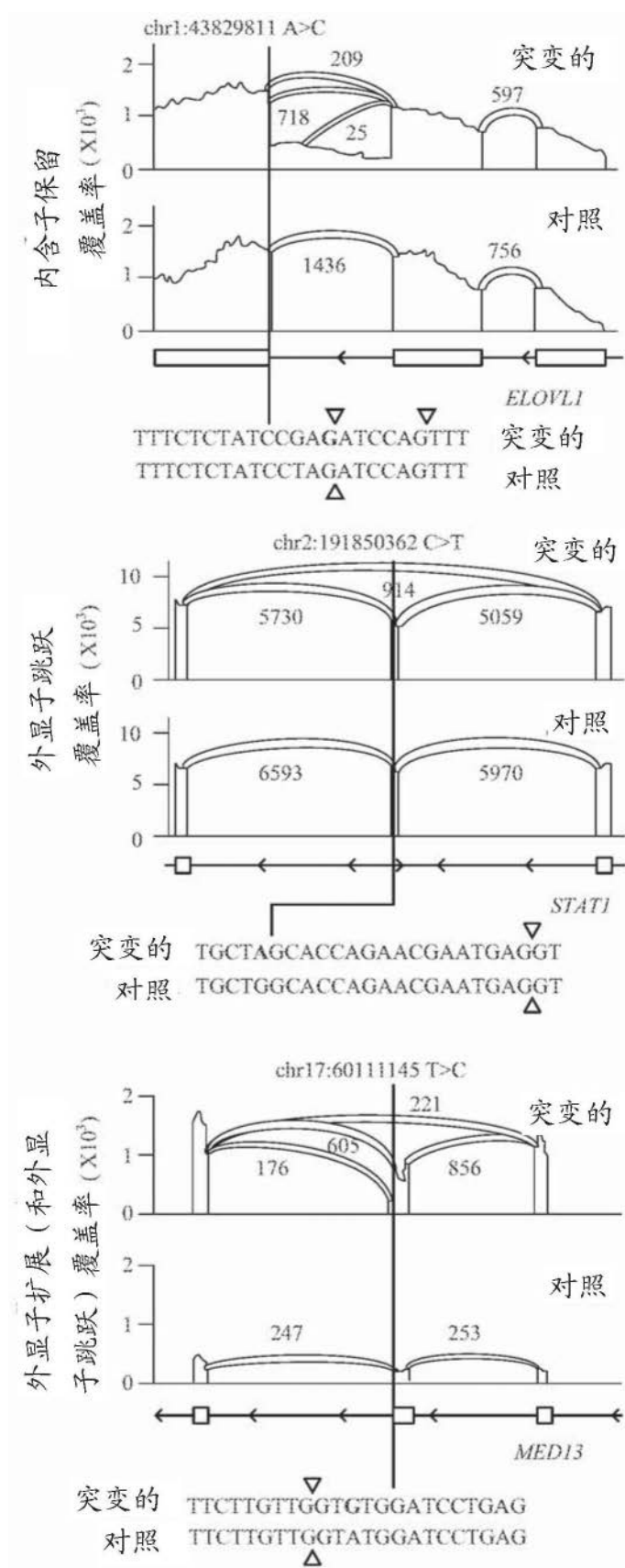


图41E

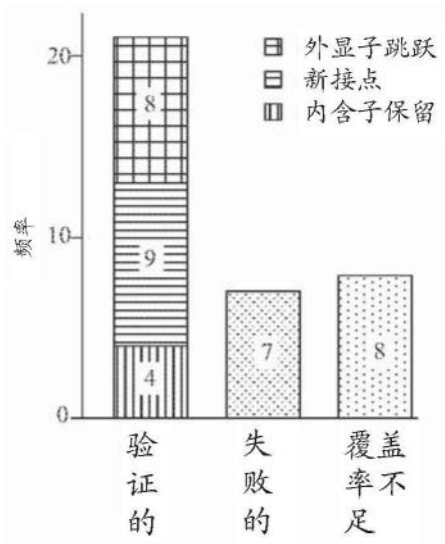


图41F

	top-k精确度		精度召回AUC	
	受体	供体	受体	供体
SpliceNet-80nt	0.5129	0.5110	0.5195	0.5273
SpliceNet-400nt	0.6848	0.7221	0.7320	0.7982
SpliceNet-2k	0.8090	0.8405	0.8742	0.9047
SpliceNet-10k	0.8223	0.8586	0.8896	0.9107
GeneSplicer	0.3238	0.3381	0.2234	0.2982
MaxEntScan	0.2932	0.3687	0.2233	0.2937
NNSplice	0.2655	0.3687	0.1702	0.3003

图42A

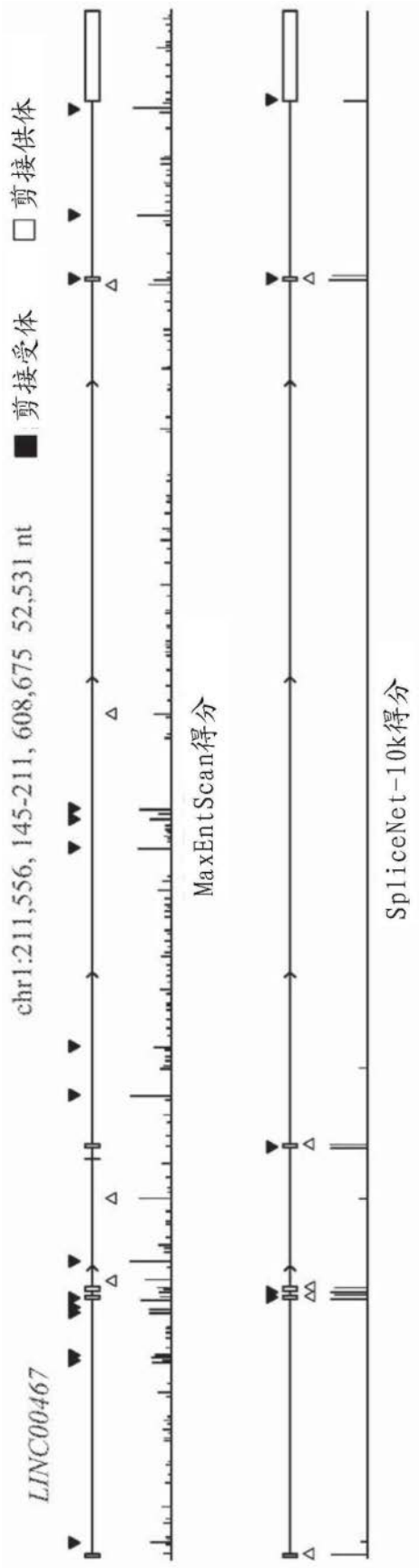


图42B

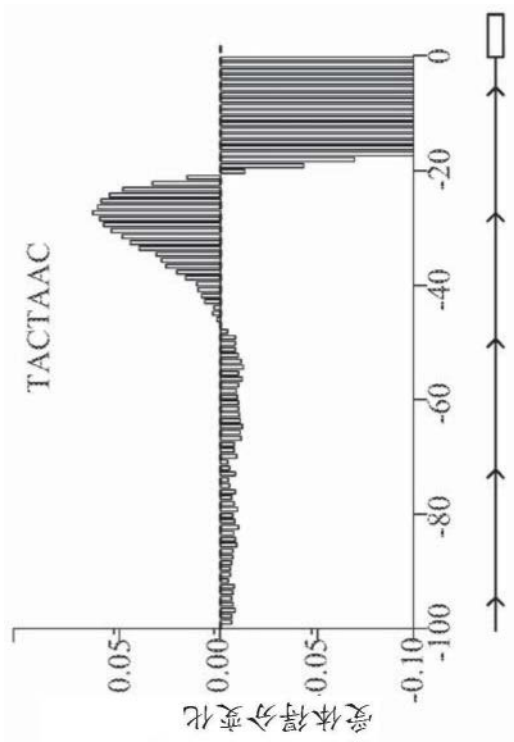


图43A

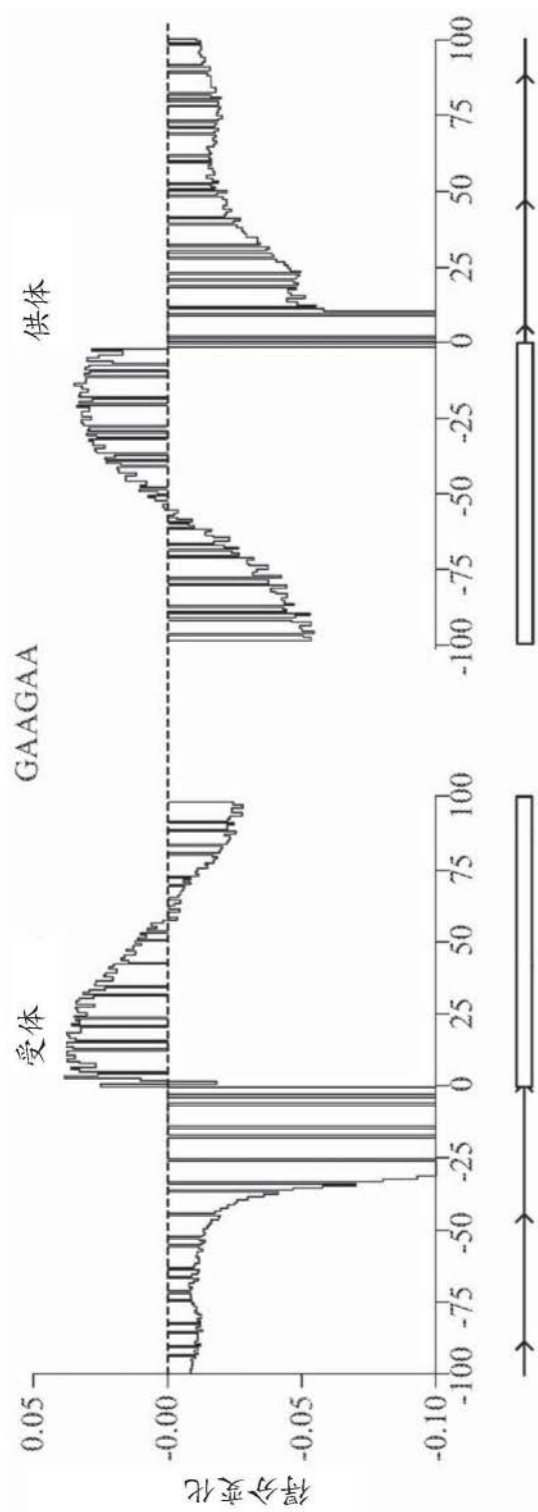


图43B

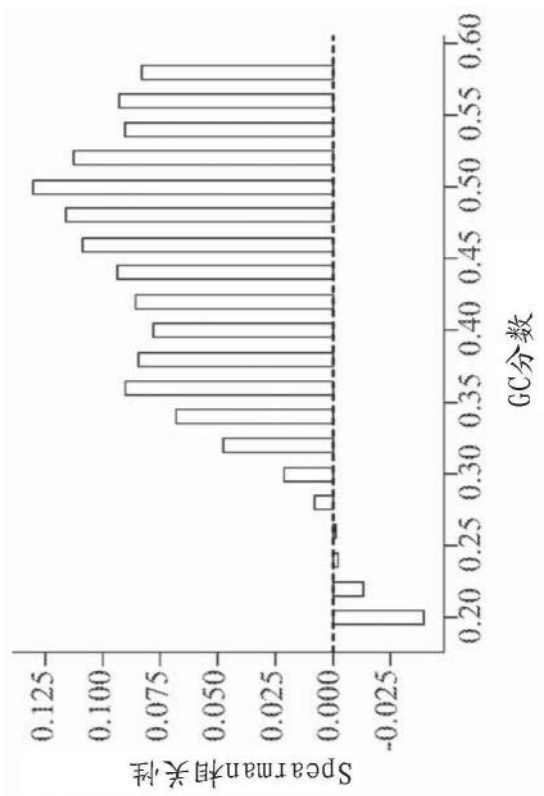


图44A

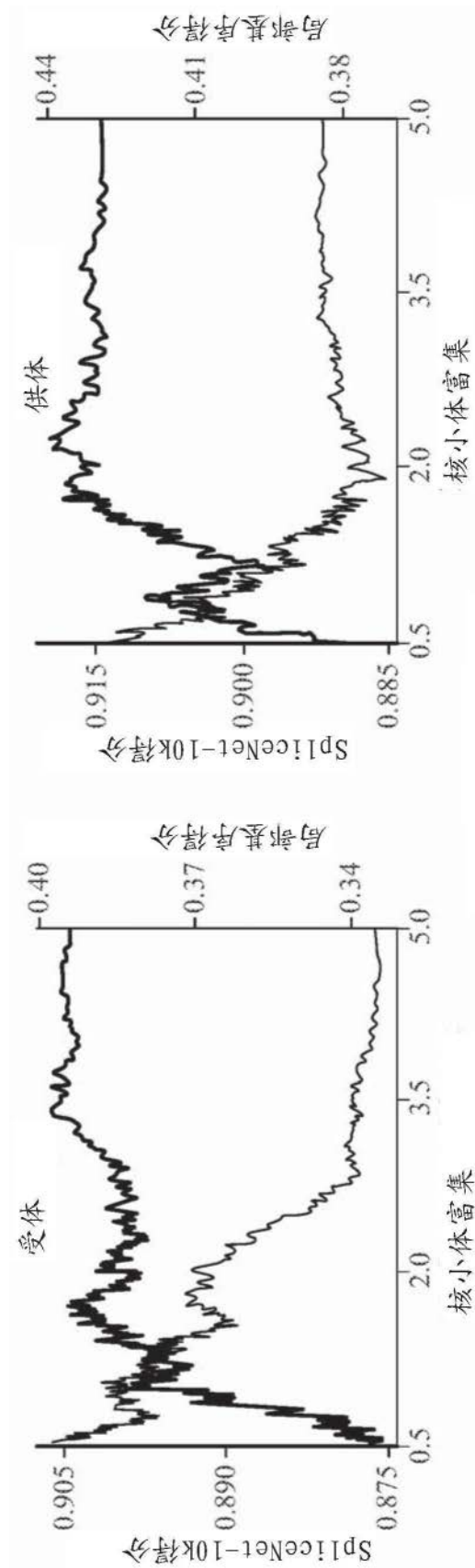


图44B

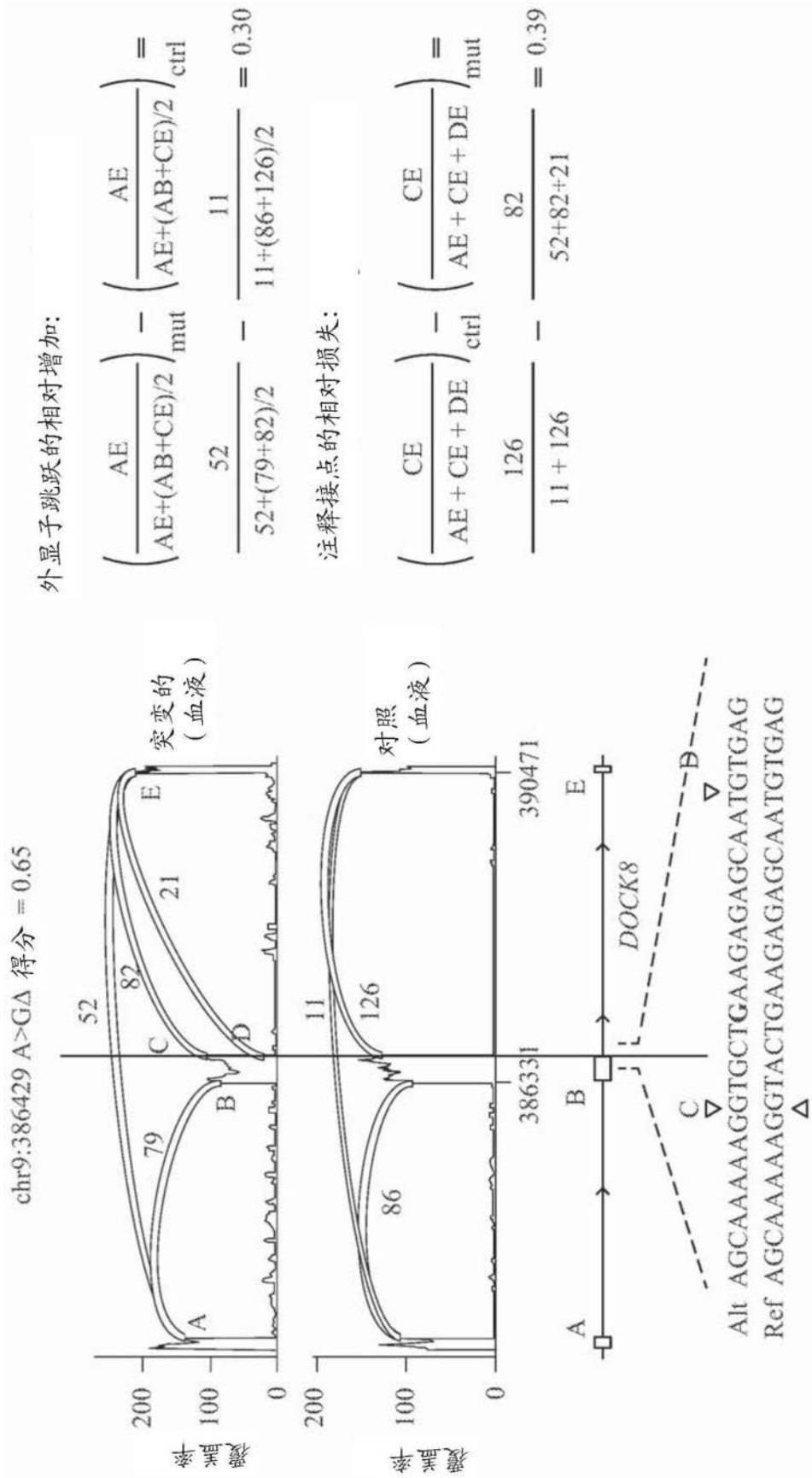


图45

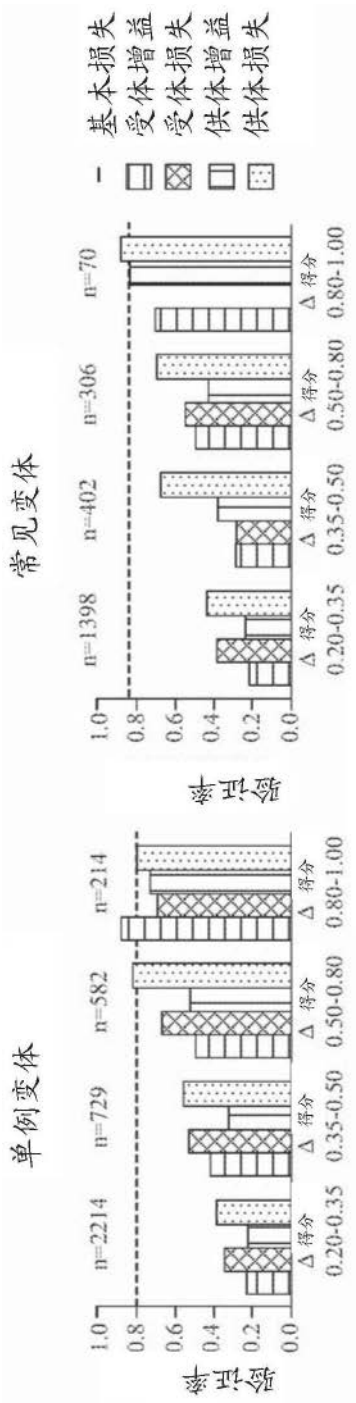


图46A

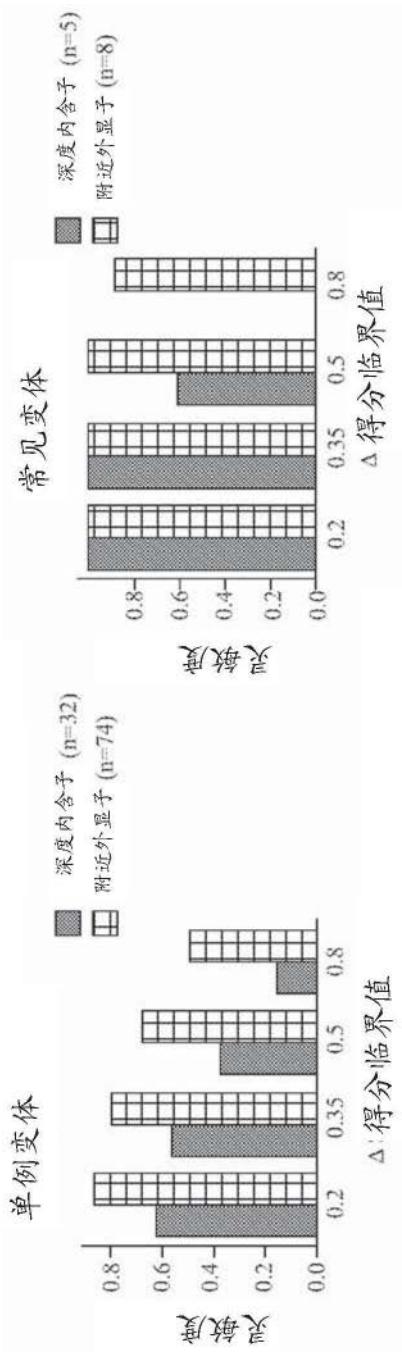


图46B

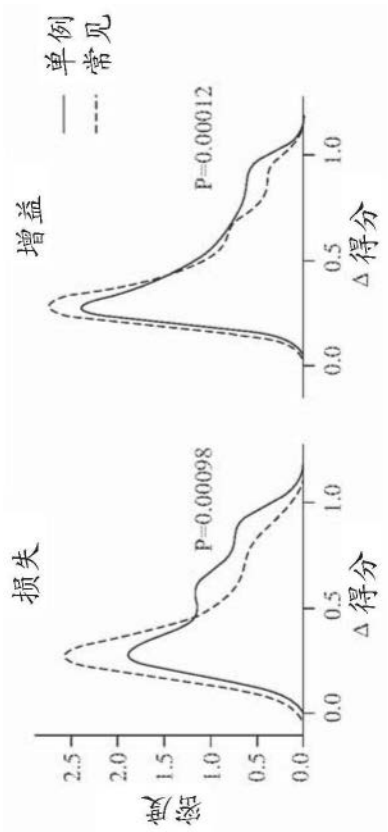


图46C

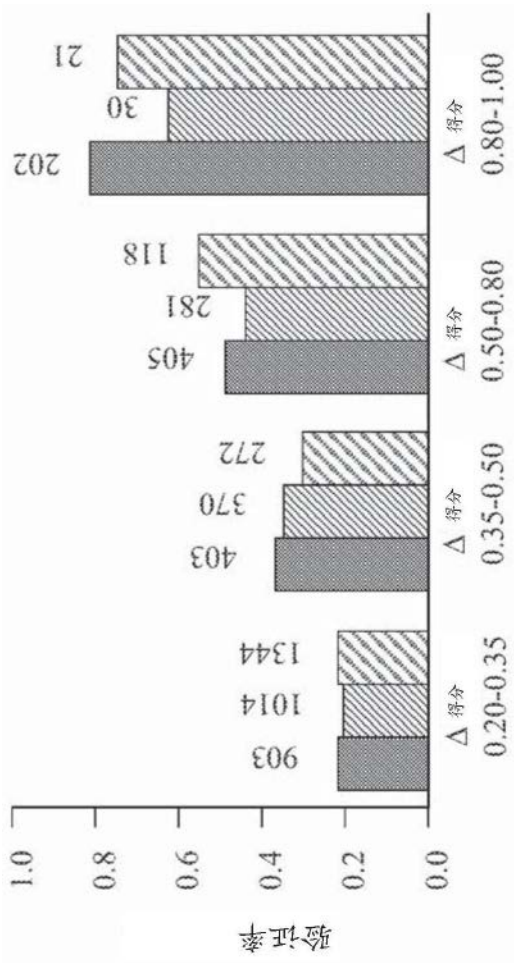


图47A

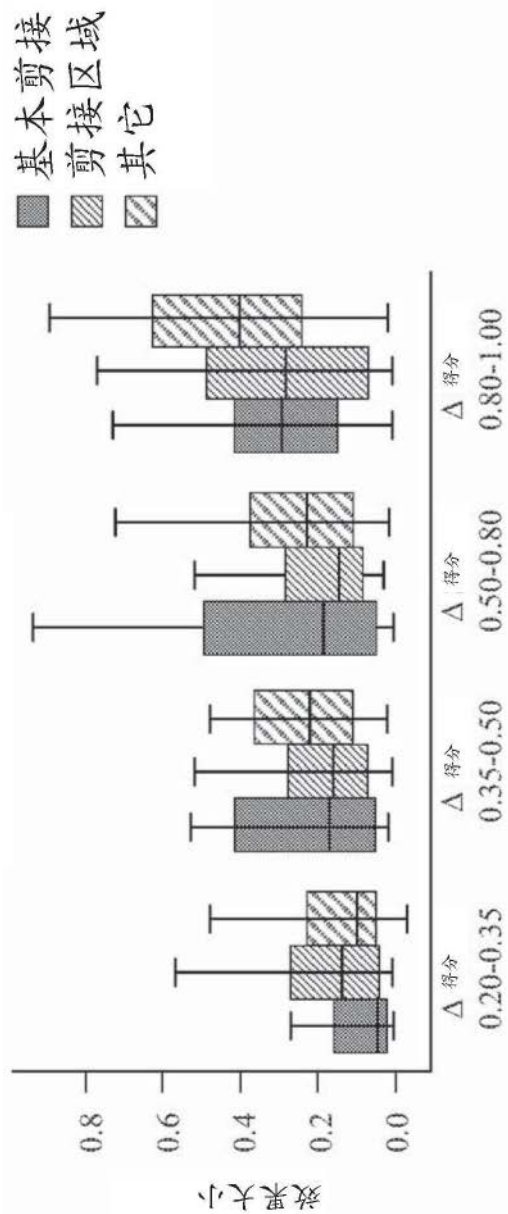


图47B

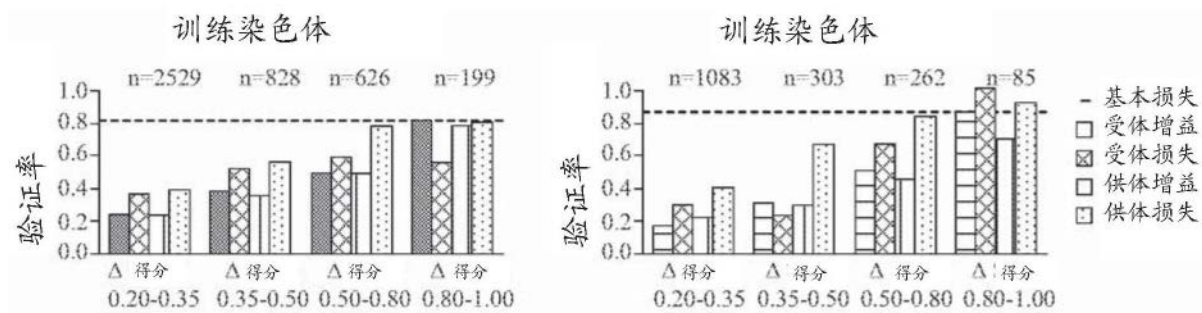


图48A

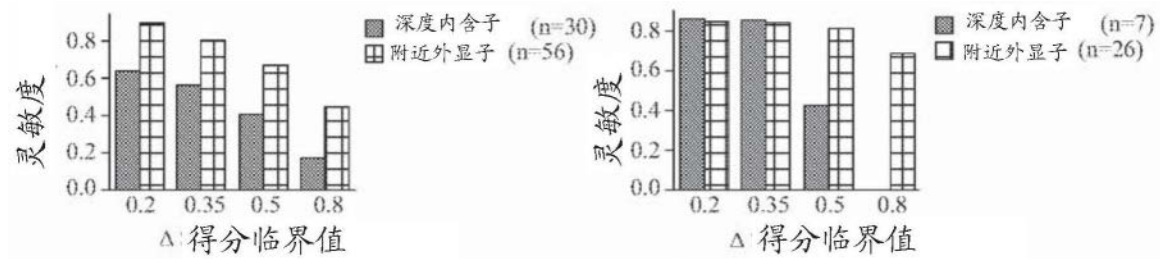


图48B

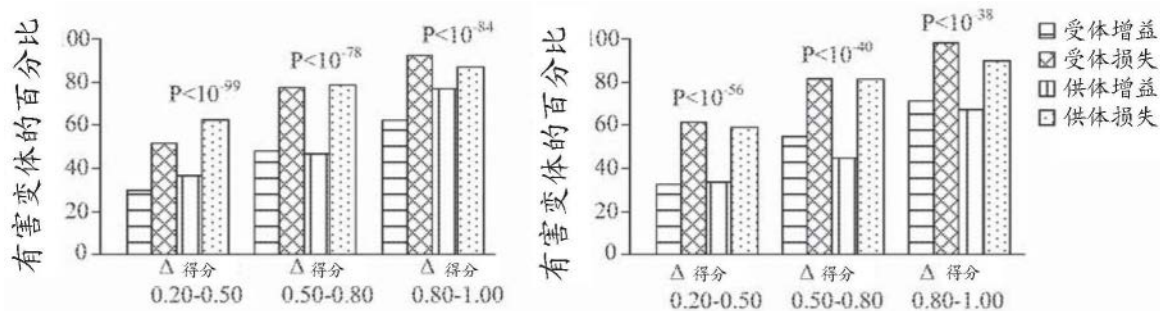


图48C

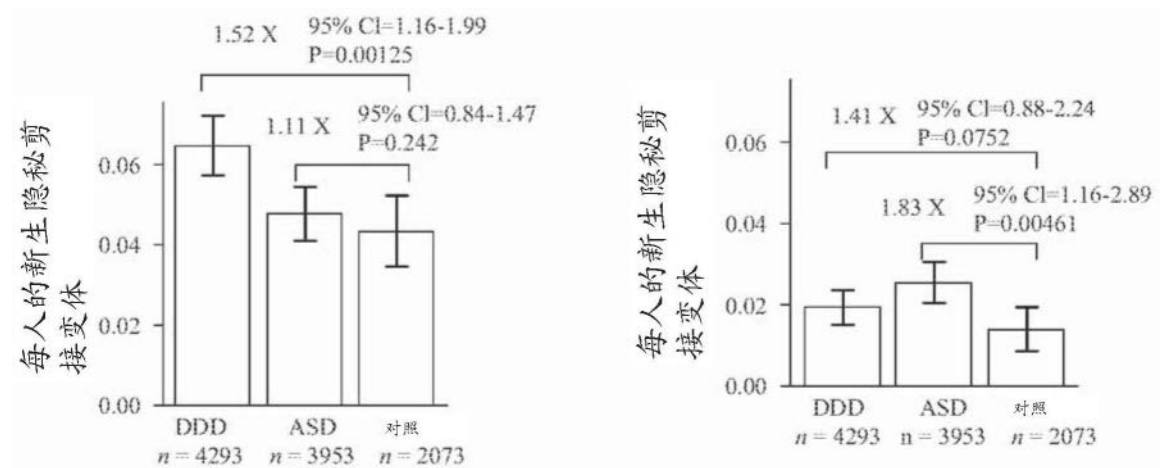


图48D

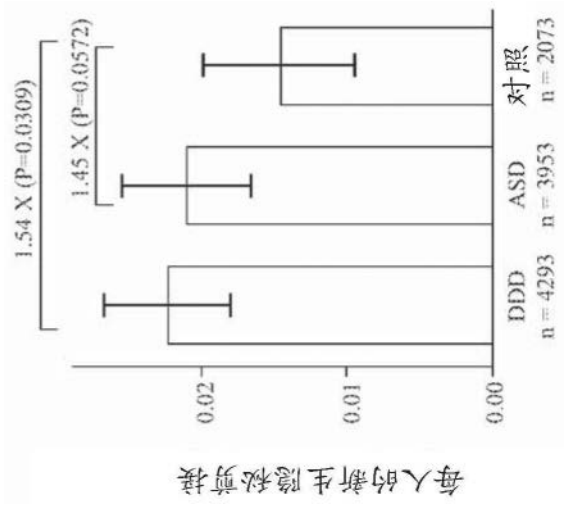


图49A

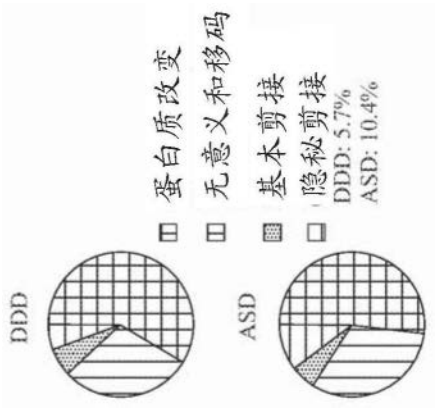


图49B

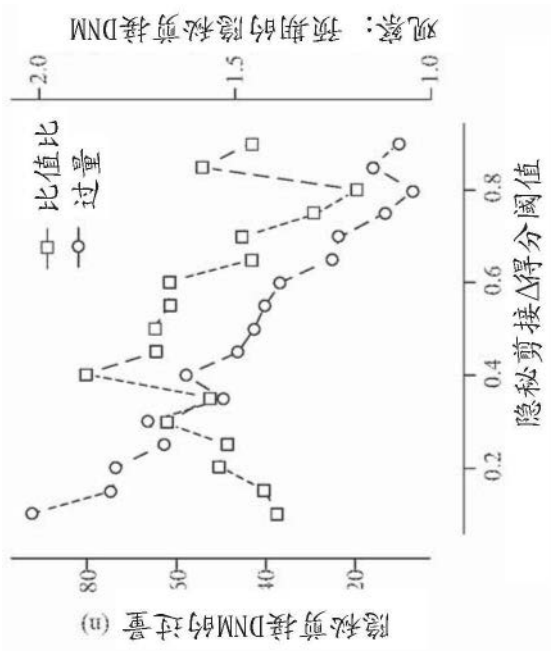


图49C

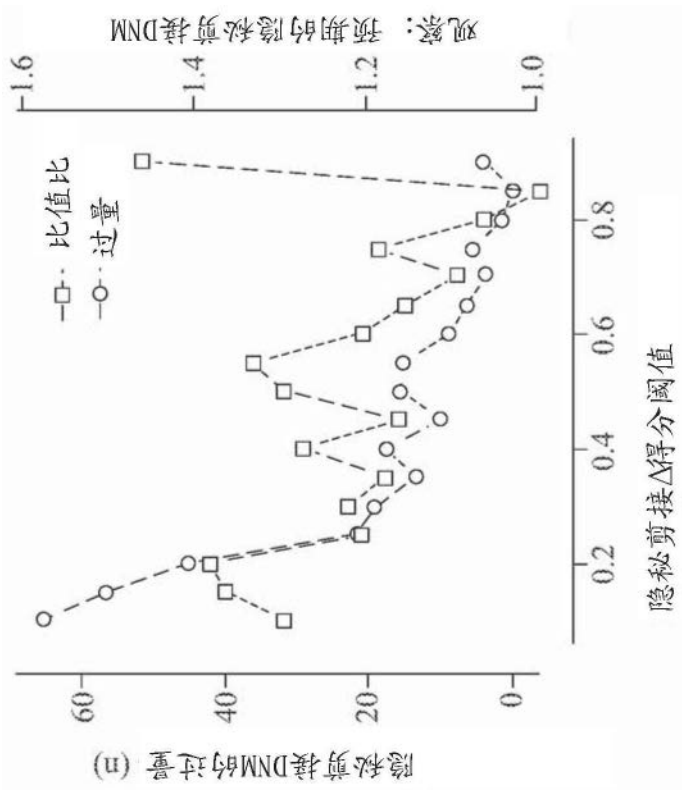


图50A

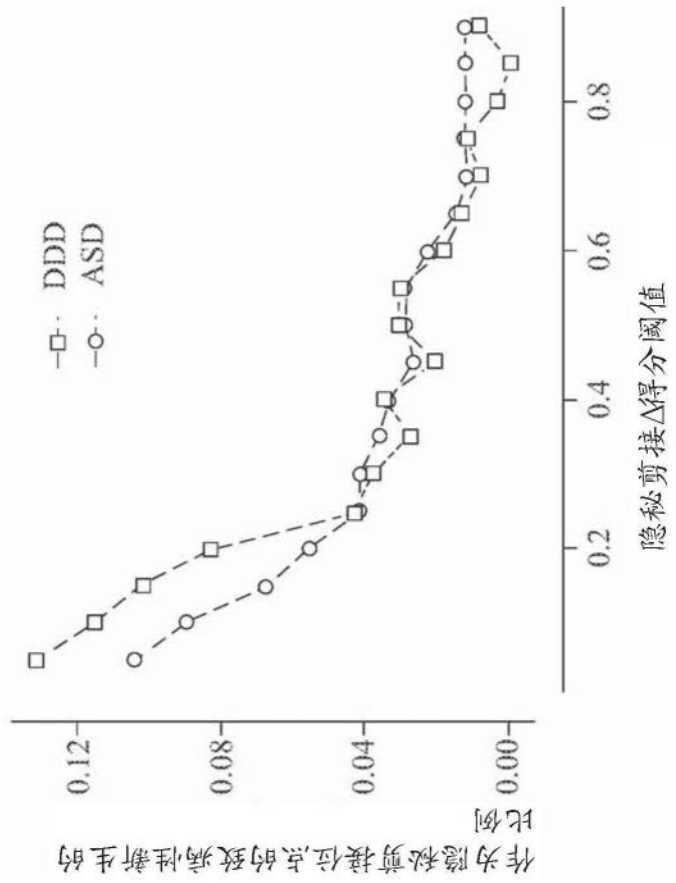


图50B

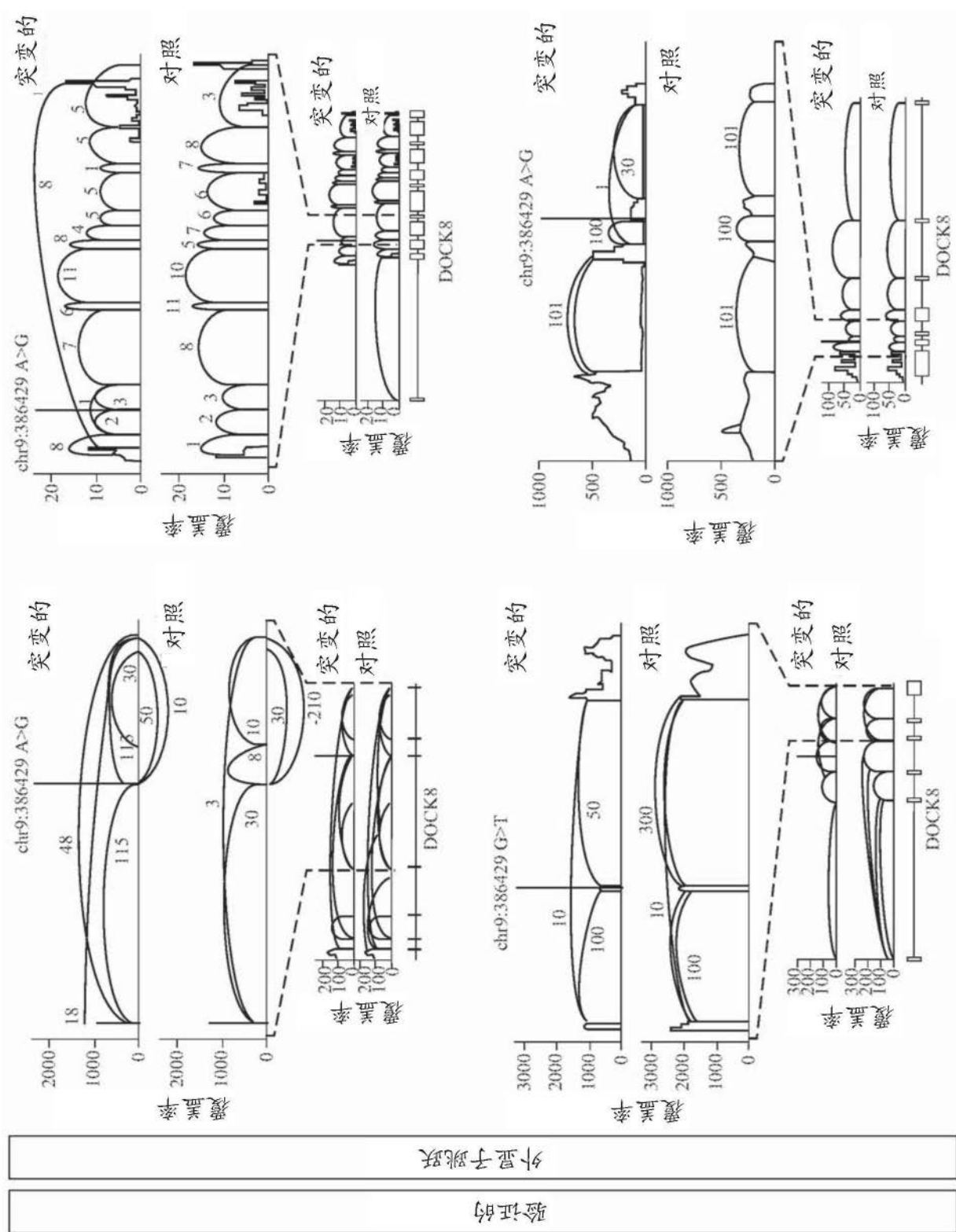


图51A

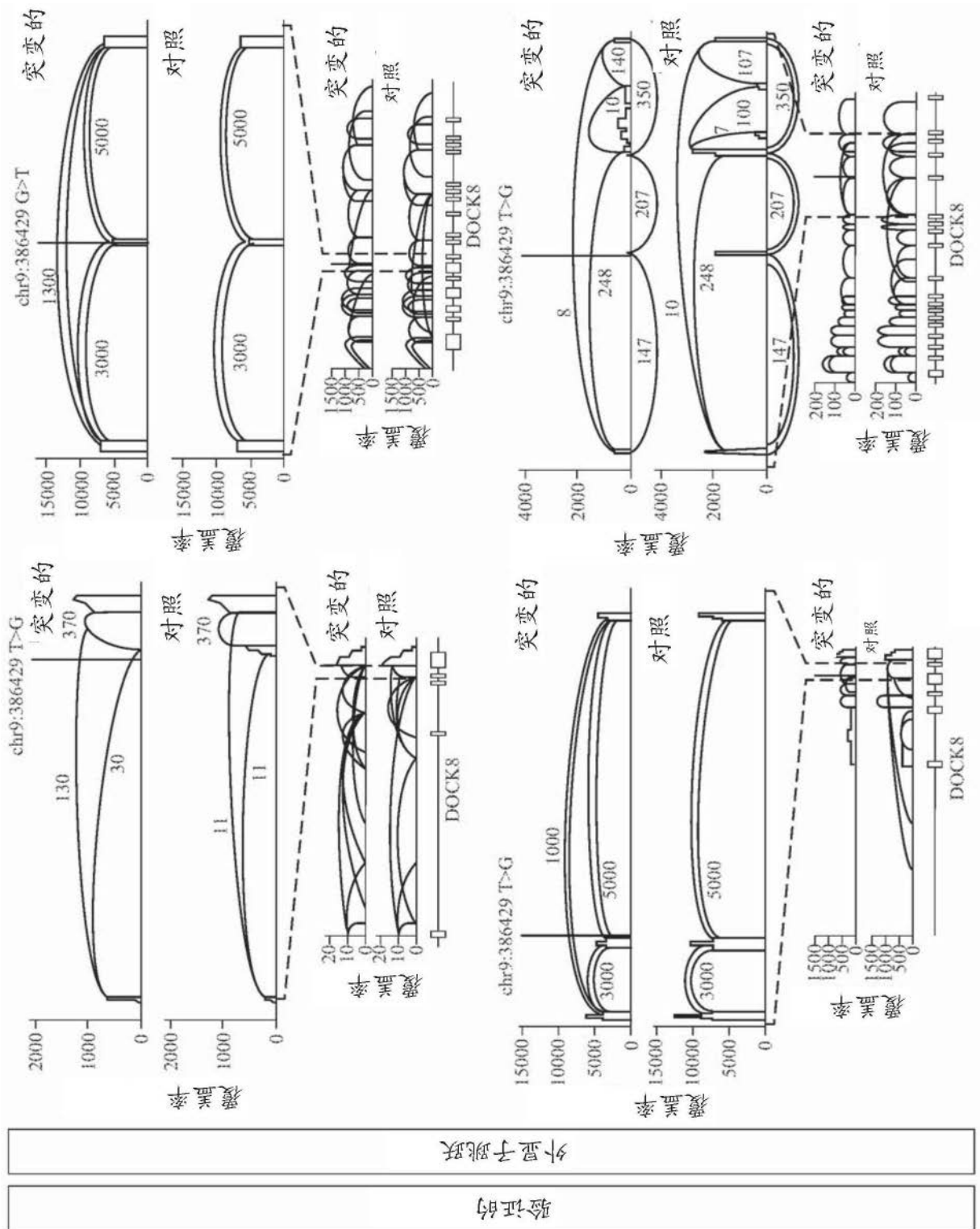


图51B

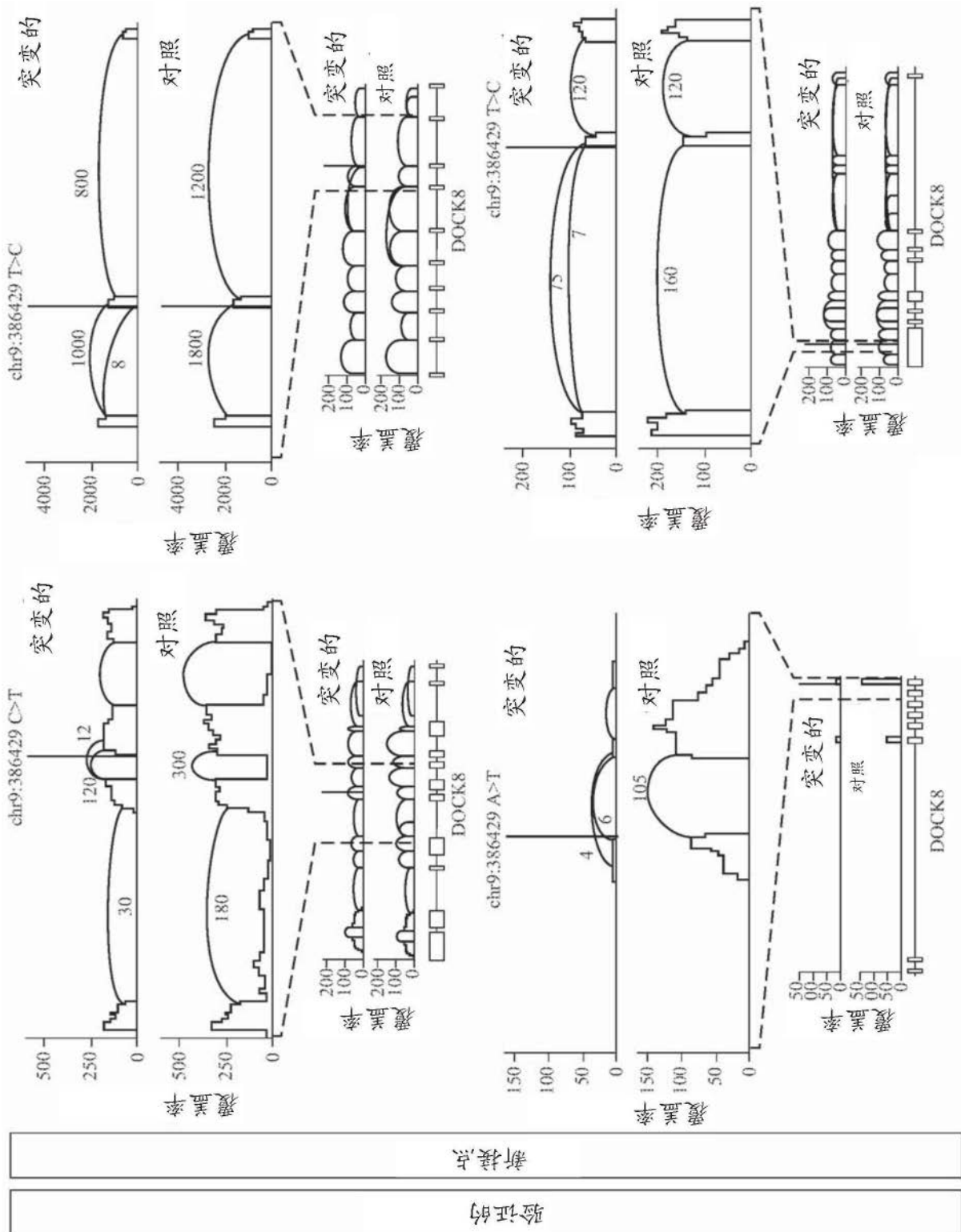


图51C

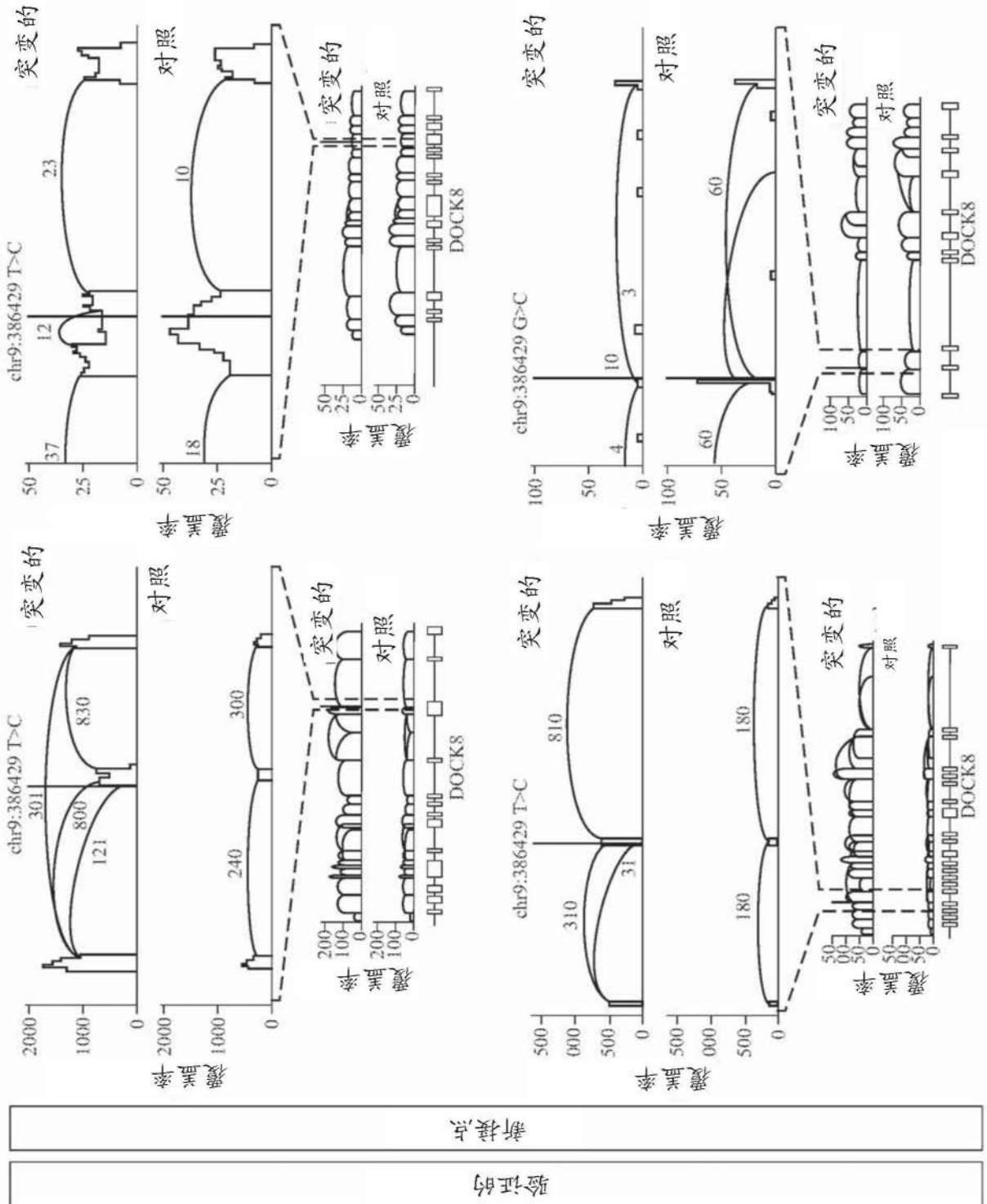


图51D

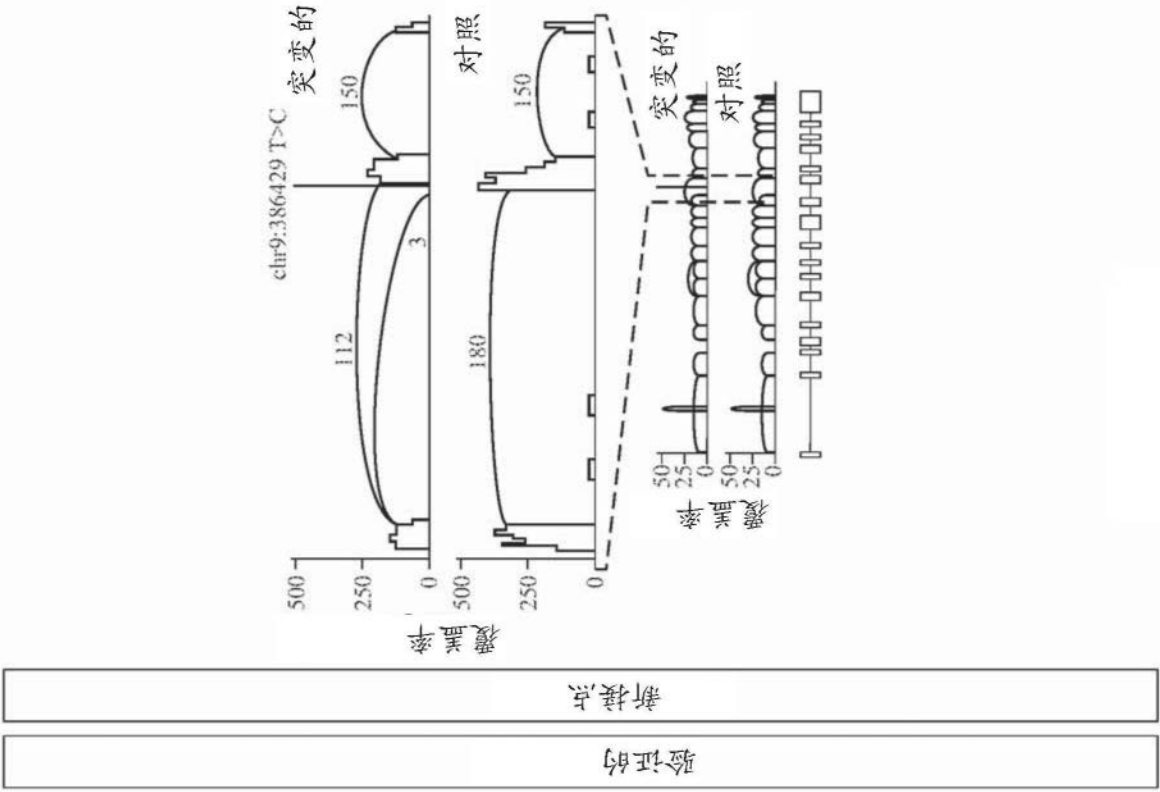


图51E

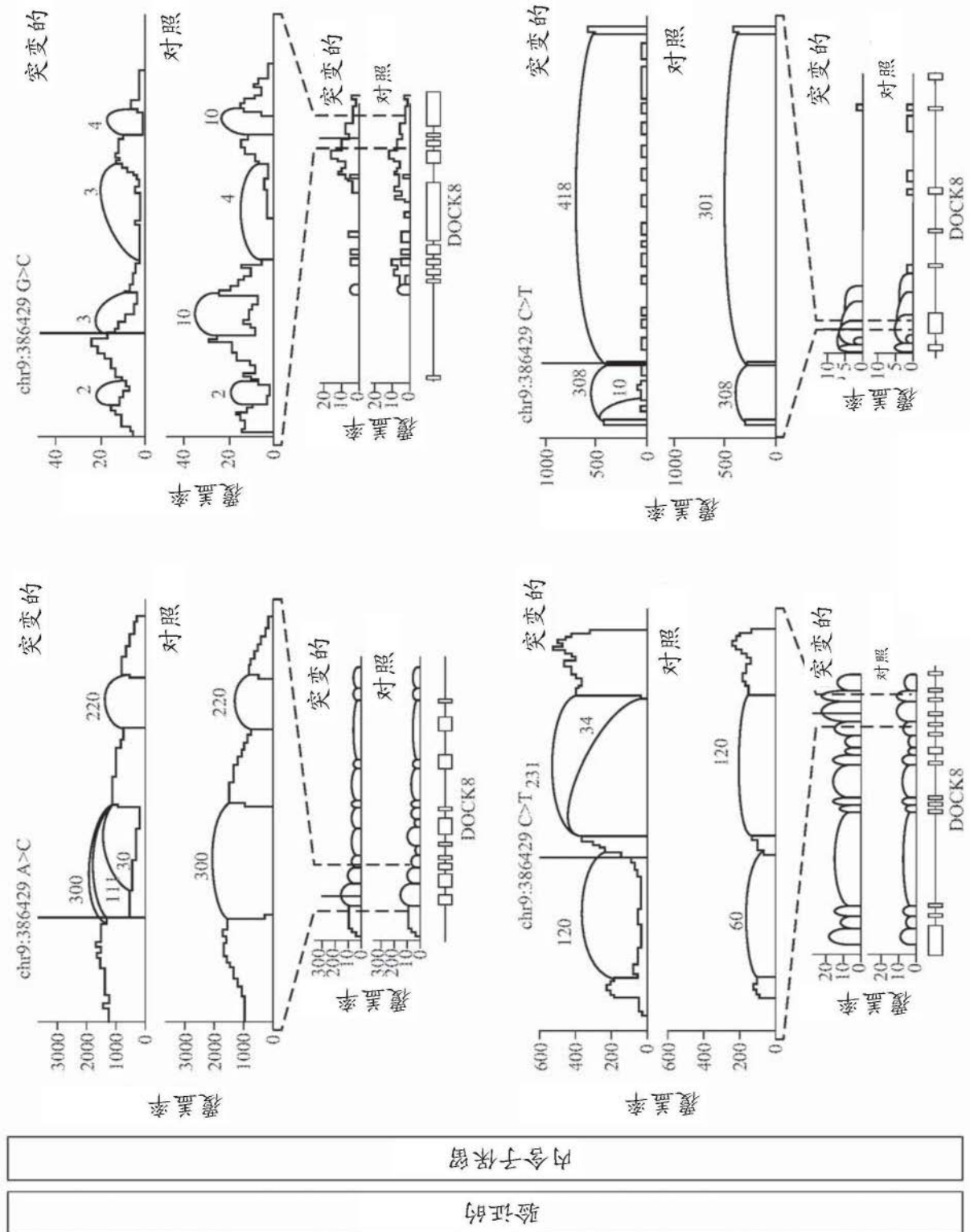


图51F

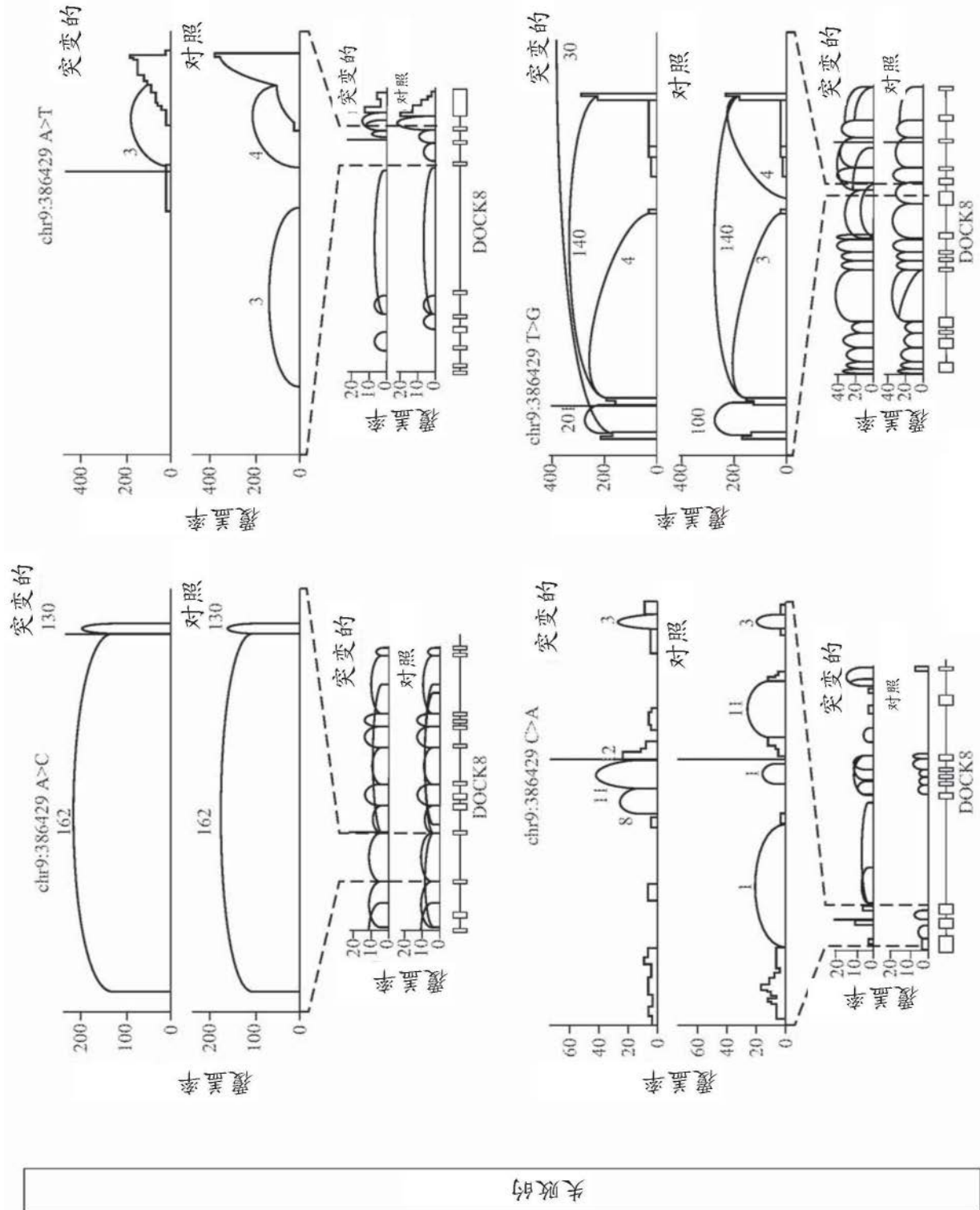


图51G

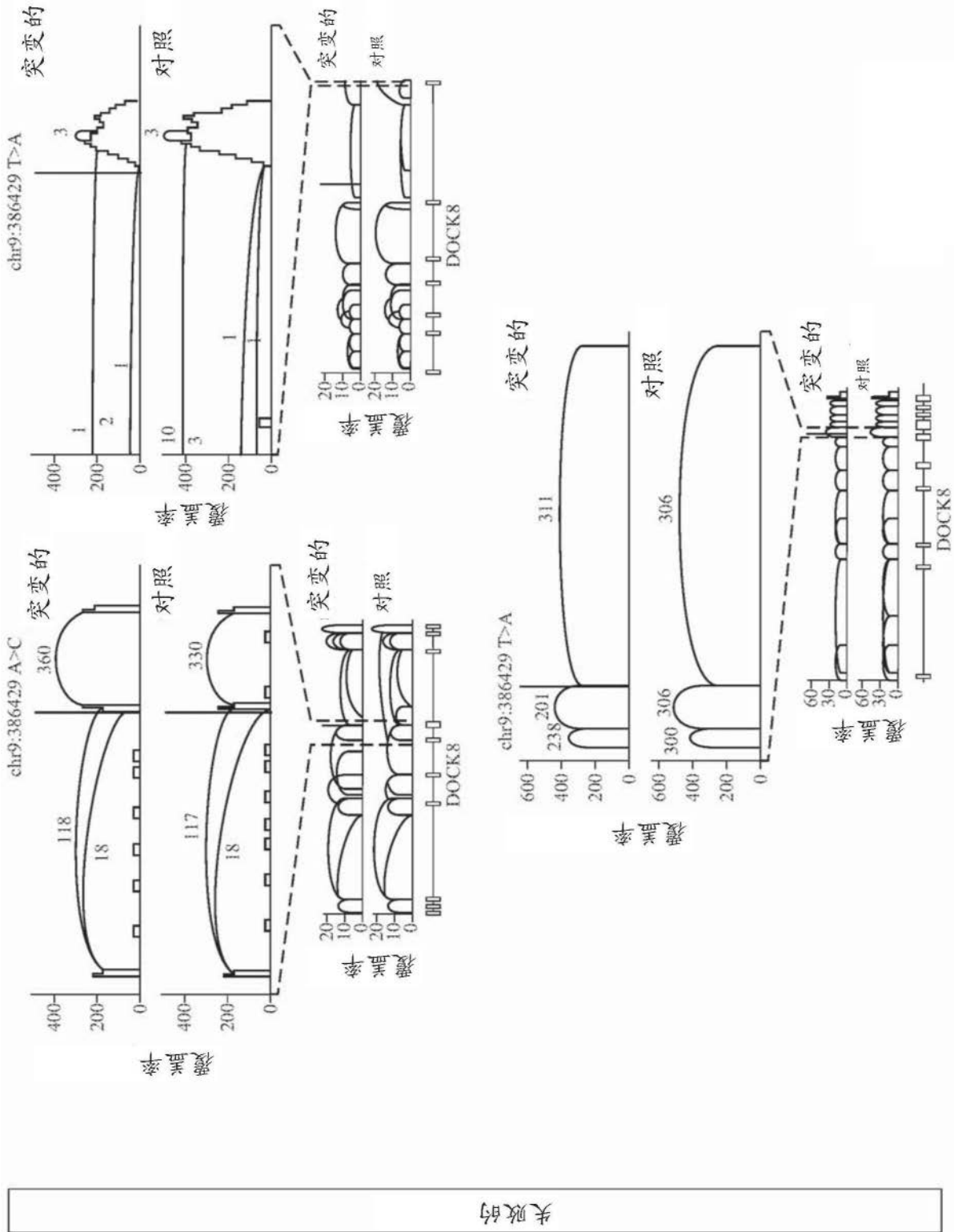


图51H

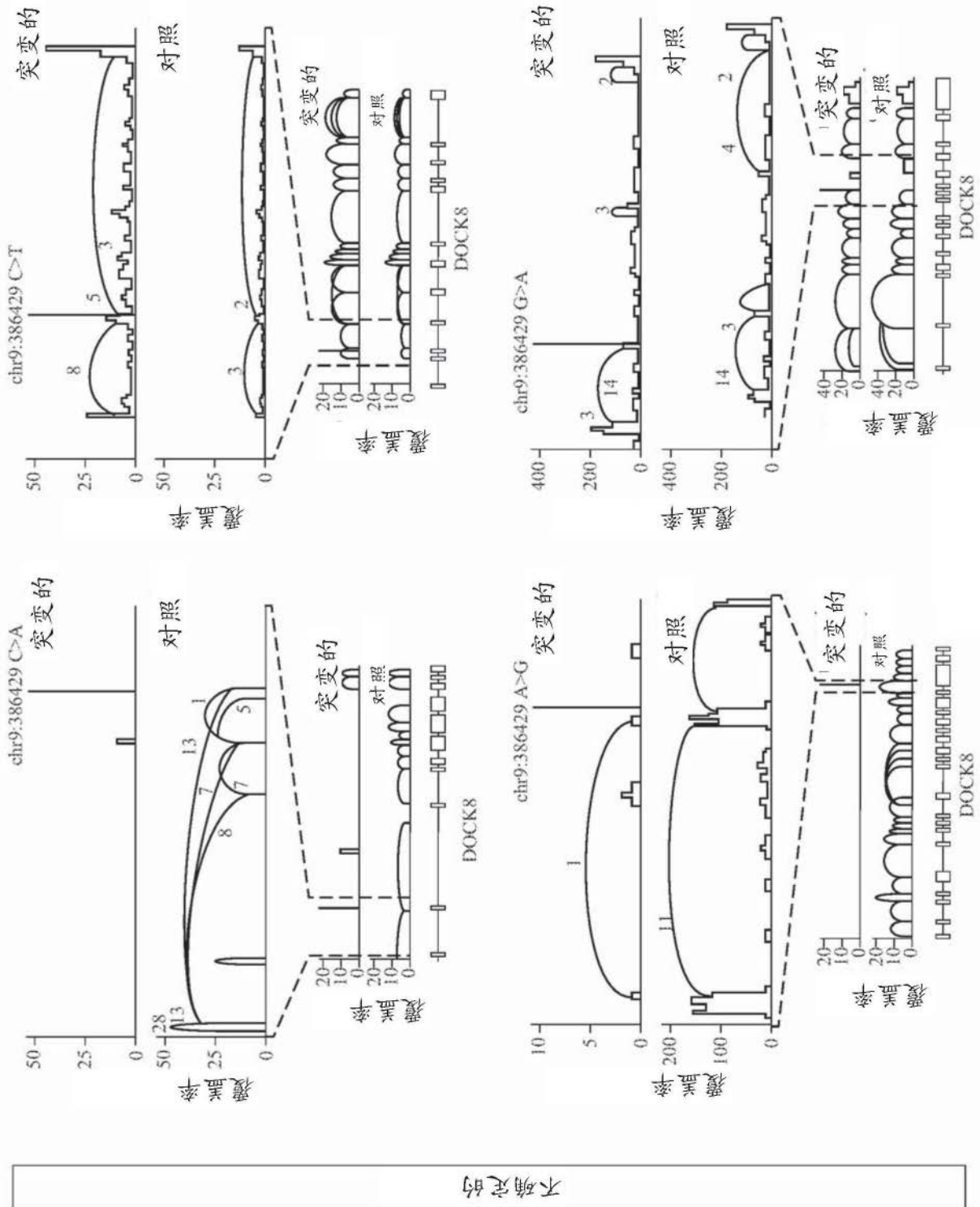


图51I

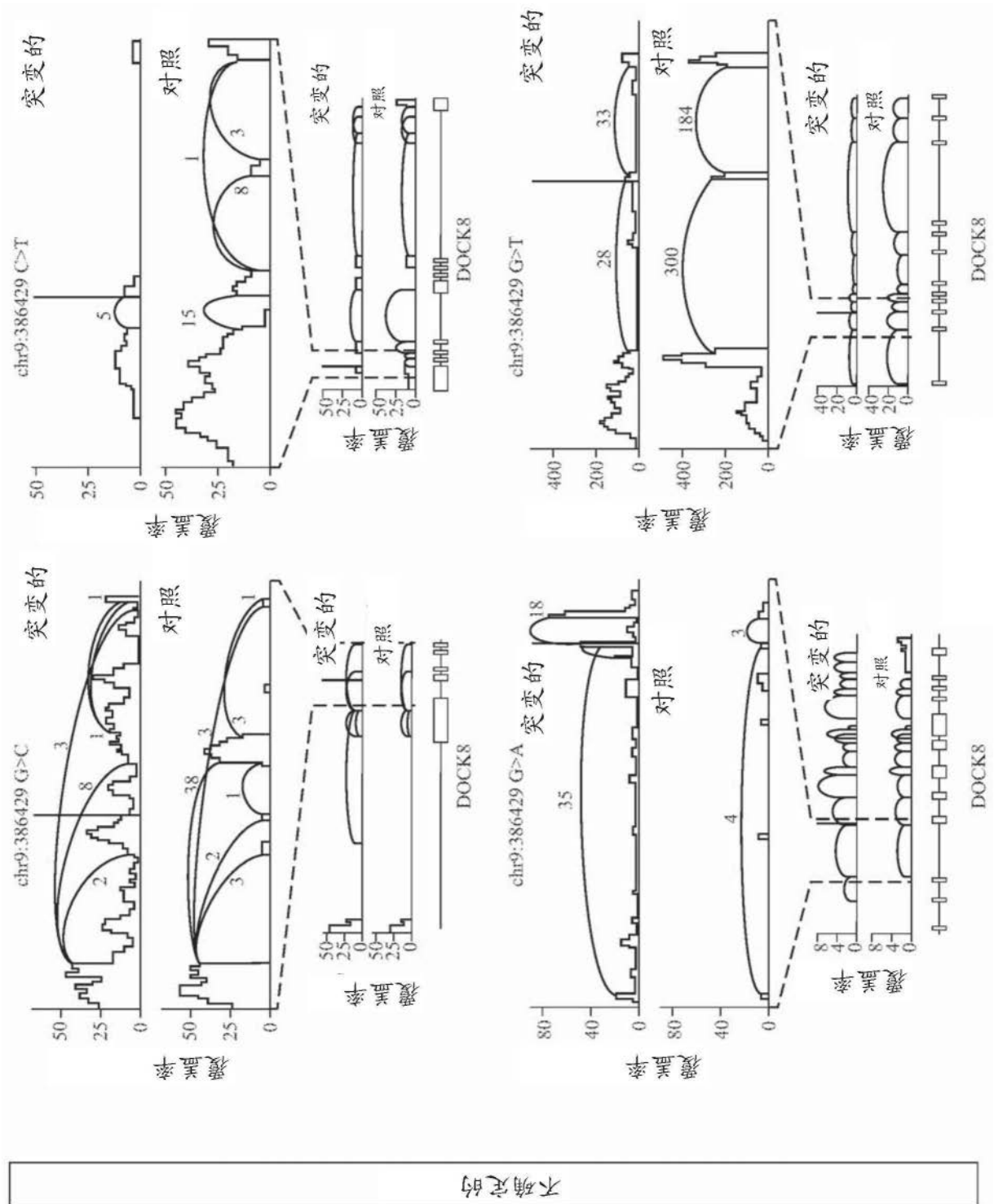


图51J

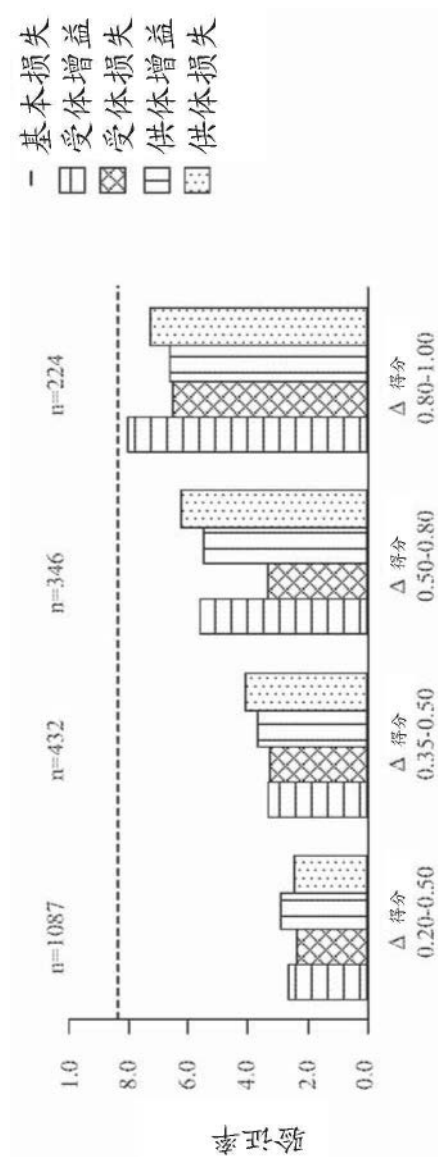


图52A

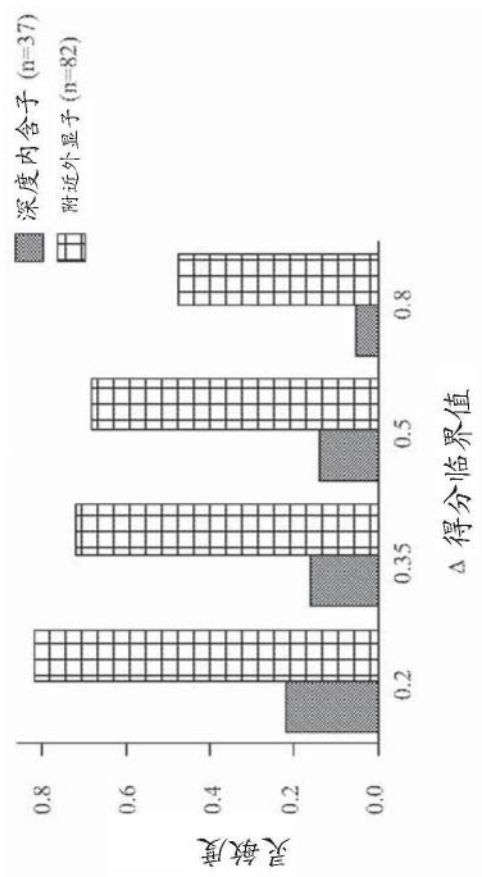


图52B

		top-k精确度		精度召回AUC	
		受体	供体	受体	供体
模型	#1	0.9304	0.9353	0.9630	0.9682
模型	#2	0.9297	0.9354	0.9649	0.9695
模型	#3	0.9288	0.9373	0.9650	0.9702
模型	#4	0.9222	0.9307	0.9616	0.9659
模型	#5	0.9267	0.9297	0.9640	0.9673

图53A

		top-k精确度		精度召回AUC	
		受体	供体	受体	供体
1个模型		0.9304	0.9353	0.9630	0.9682
2个模型		0.9376	0.9428	0.9694	0.9742
3个模型		0.9409	0.9472	0.9715	0.9758
4个模型		0.9424	0.9474	0.9727	0.9765
5个模型		0.9429	0.9479	0.9737	0.9771

图53B

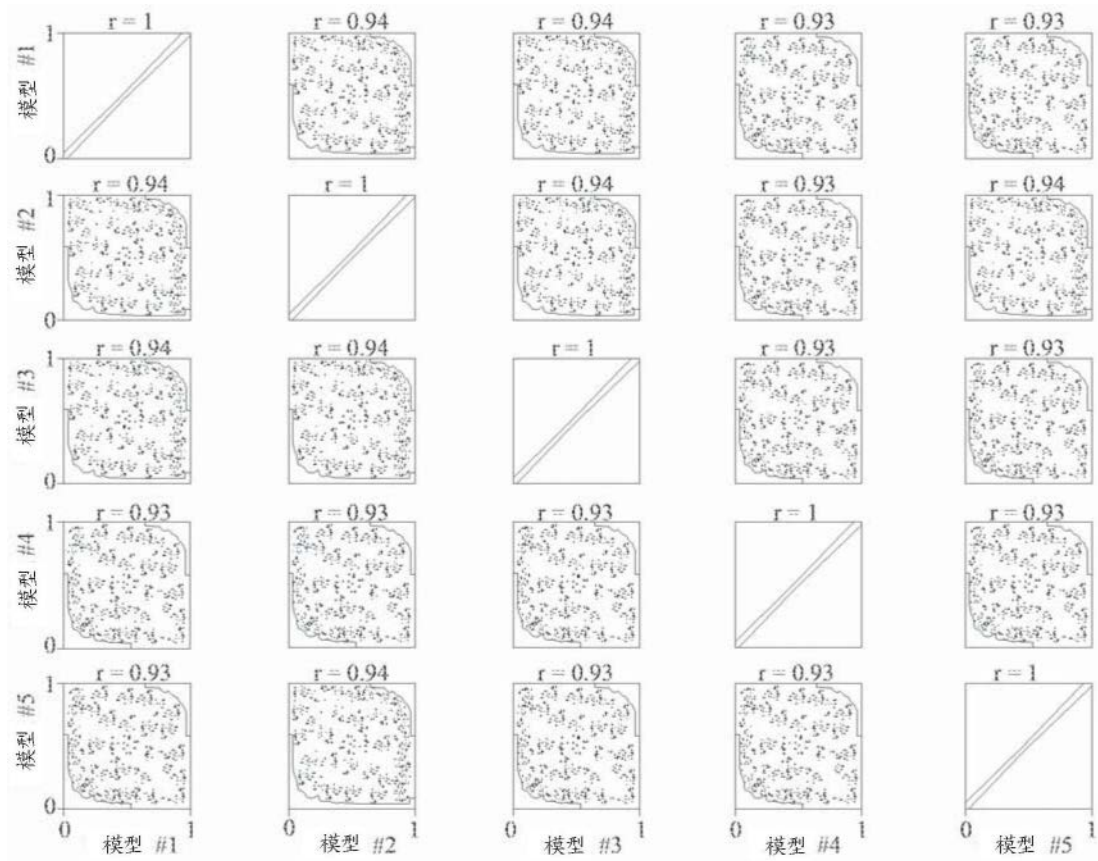


图53C

A SpliceNet-10k					
外显子密度	top-k精确度		精度召回AUC		
	受体	供体	受体	供体	
1个外显子	0.9165	0.9292	0.9592	0.9659	
2个外显子	0.9307	0.9357	0.9645	0.9674	
3个外显子	0.9519	0.9512	0.9816	0.9822	
4个外显子	0.9556	0.9540	0.9833	0.9845	
≥ 5个外显子	0.9579	0.9665	0.9826	0.9872	
B MaxEntScan					
外显子密度	top-k精确度		精度召回AUC		
	受体	供体	受体	供体	
1个外显子	0.1834	0.2487	0.1140	0.1660	
2个外显子	0.2180	0.3124	0.1512	0.2375	
3个外显子	0.2616	0.3348	0.1964	0.2773	
4个外显子	0.2950	0.3820	0.2307	0.3347	
≥ 5个外显子	0.3583	0.4652	0.3103	0.4314	

图54A-B

组织	变体个体	对照个体
皮肤暴露于日光 (小腿)	T6MO	OXRK
肌肉-骨骼的, 转化的纤维母细胞	WHSB	WK11
动脉-胚, 肺	XOT4, S341, POMQ	WFON
全血	TMIMY	T8EM

图55

增益					损失				
SpliceNet-10k	GeneSplicer	NNSplice	MaxEntScan		SpliceNet-10k	GeneSplicer	NNSplice	MaxEntScan	
0.1	103.5145	0.9024	0.901		0.1	1.5183	0.1063		0
0.2	105.6838	0.9288	0.9171		0.2	2.6295	0.2315		0.0068
0.3	107.1061	0.9542	0.9419		0.3	3.4707	0.3554		0.0312
0.4	108.0655	0.967	0.9582		0.4	4.0856	0.4625		0.0646
0.5	108.9011	0.9739	0.9694		0.5	4.6085	0.5349		0.1003
0.6	109.6138	0.9781	0.9759		0.6	5.2064	0.6345		0.1624
0.7	110.388	0.9817	0.9826		0.7	6.4456	0.7573		0.2838
0.8	111.2775	0.9839	0.9875		0.8	8.1053	0.8202		0.4271
0.9	112.6826	0.9858	0.9915		0.9	10.2586	0.8529		0.641

图56

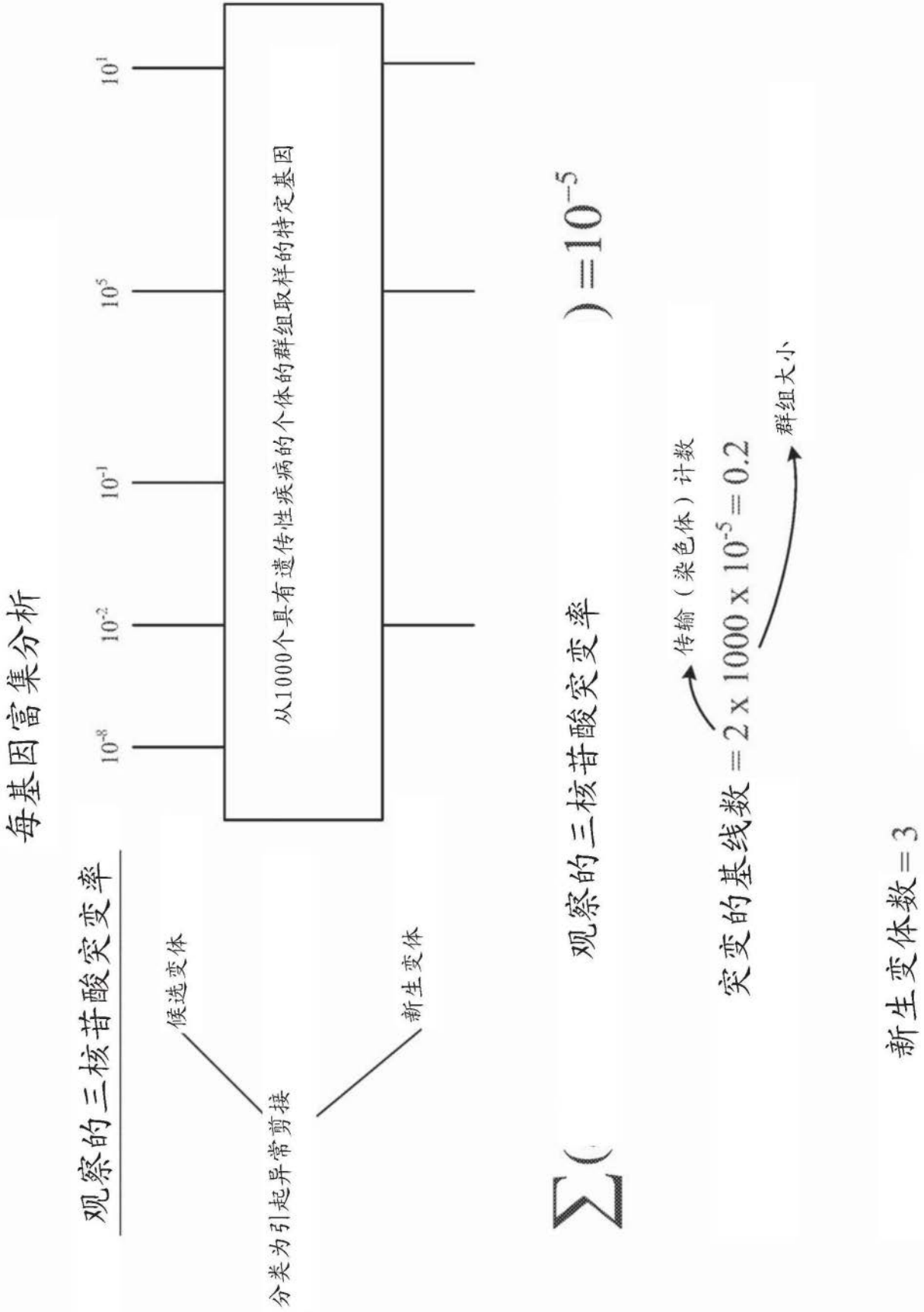


图57

全基因组富集分析

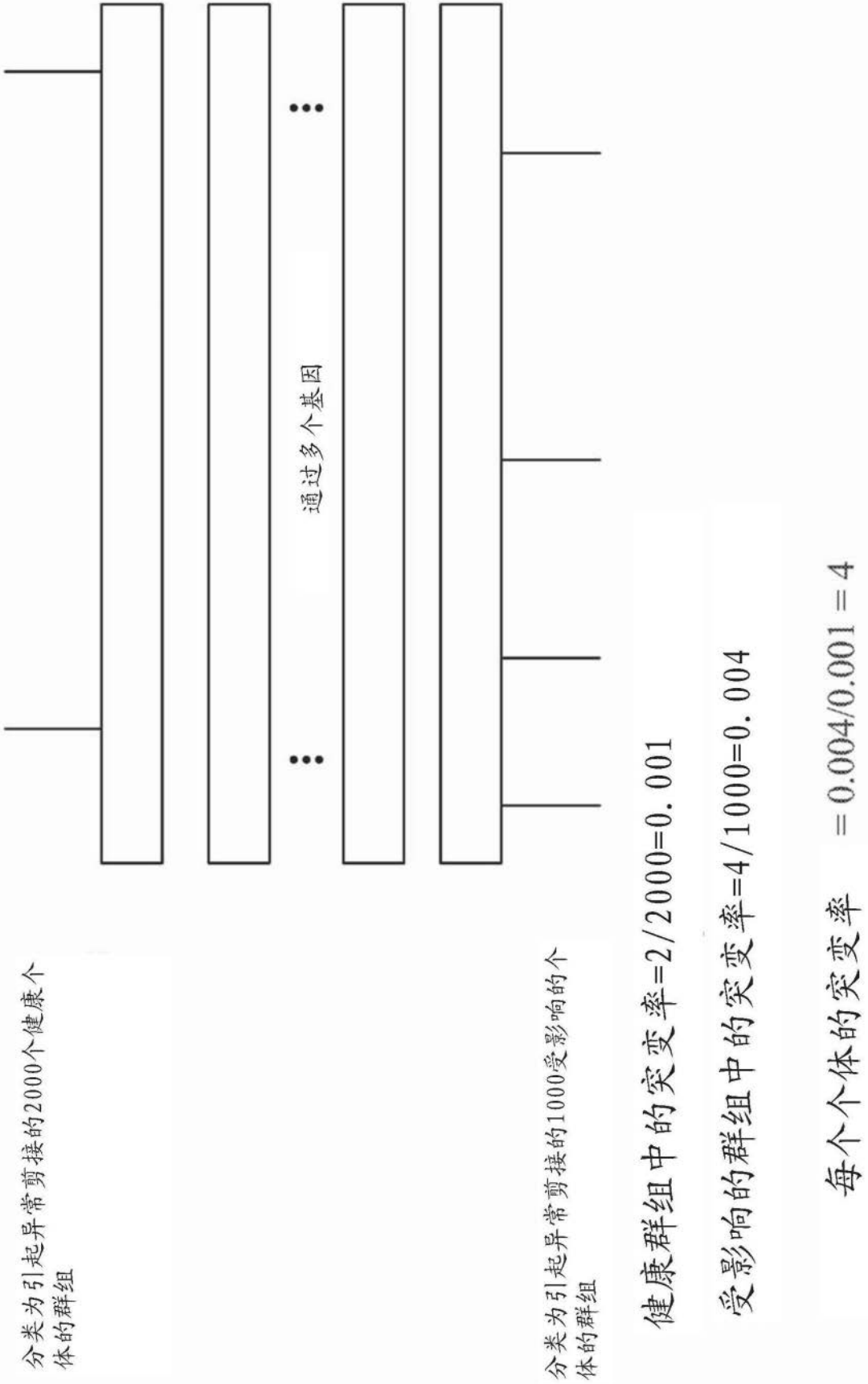


图58

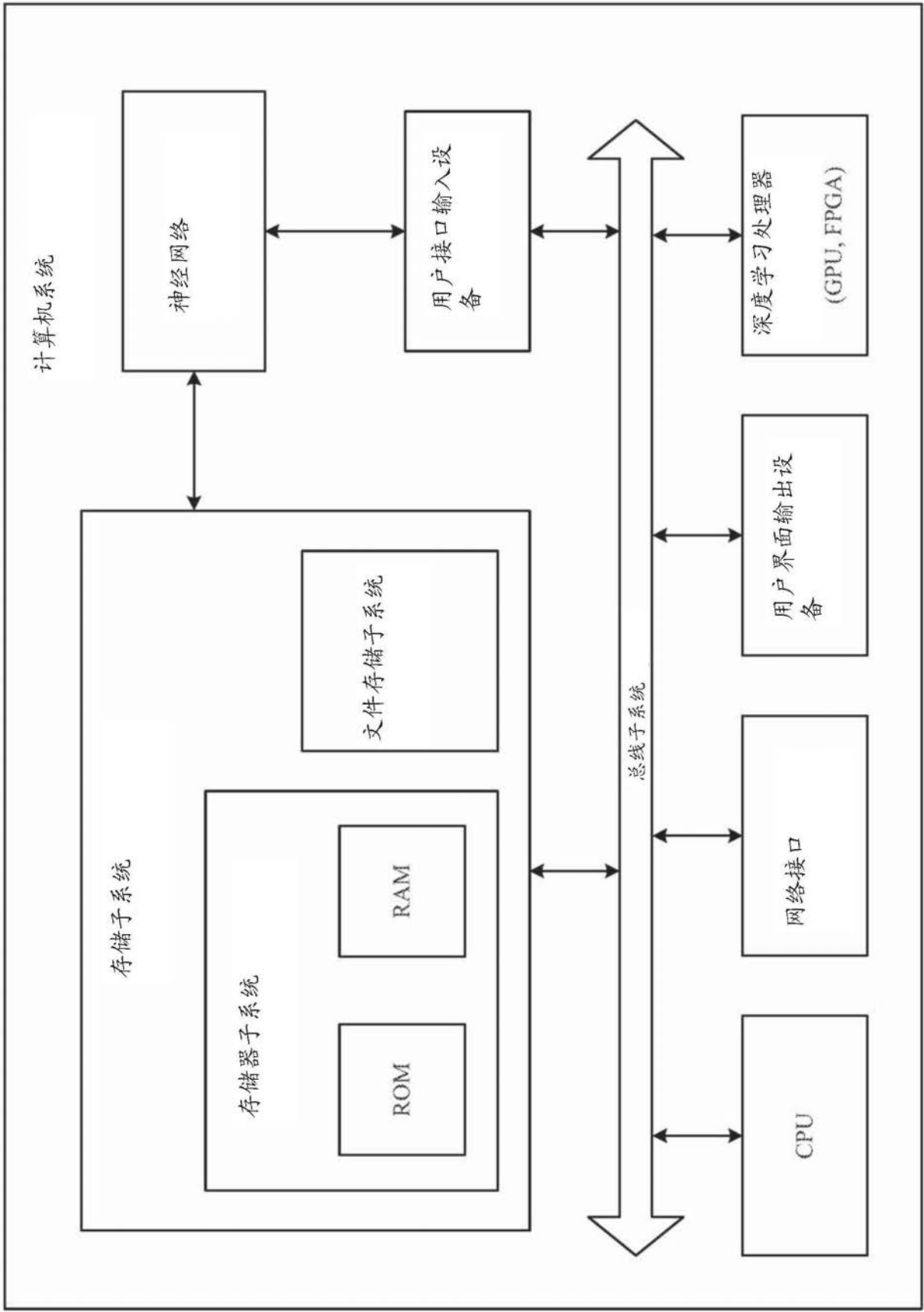


图59