



(12) 发明专利

(10) 授权公告号 CN 110832596 B

(45) 授权公告日 2021.03.26

(21) 申请号 201880043817.3

(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 110832596 A

(43) 申请公布日 2020.02.21

(30) 优先权数据

62/573144 2017.10.16 US

62/573149 2017.10.16 US

62/573153 2017.10.16 US

62/582898 2017.11.07 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/US2018/055840 2018.10.15(87) PCT国际申请的公布数据
W02019/079166 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 H·高 K-H·法尔 L·孙达拉姆
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

代理人 王勇

(51) Int.Cl.

G16B 20/20 (2006.01)

G16B 40/20 (2006.01)

(56) 对比文件

CA 2894317 A1, 2016.12.15

CN 103679185 A, 2014.03.26

WO 2016201049 A2, 2016.12.15

WO 2017172958 A1, 2017.10.05

CN 105378104 A, 2016.03.02

Wei, Qiong等.The Role of Balanced
Training and Testing Data Sets for Binary
Classifiers in Bioinformatics.《PLOS ONE》
.2013,第8卷(第7期),全文.

审查员 刘欢

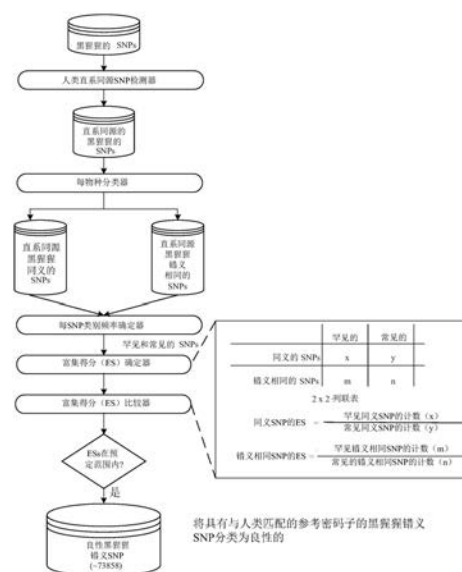
权利要求书4页 说明书100页 附图103页

(54) 发明名称

基于深度学习的深度卷积神经网络训练方
法

(57) 摘要

公开的技术涉及构建用于变体分类的基于卷积神经网络的分类器。更具体地,涉及一种在训练数据上使用基于后向传播梯度更新方法训练基于卷积神经网络的分类器,该方法逐步地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匹配。该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数量、残差块的卷积窗口尺寸和残差块的空洞卷积率参数化,所述卷积窗口尺寸在残差块组之间变化,所述空洞卷积率在残差块组之间变化。所述训练数据包括良性训练实例和从良性变体和致病性变体产生的转化的序列对的致病性训练实例。



1. 一种构建变体致病性分类器的方法,所述方法包括:

使用参考蛋白质序列和替代蛋白质序列的良性训练实例和致病性训练实例作为输入来训练基于卷积神经网络的变体致病性分类器,所述基于卷积神经网络的变体致病性分类器在耦接至存储器的多个处理器上运行,其中所述替代蛋白质序列是从良性变体和致病性变体产生的;以及

其中所述良性变体包括常见人类错义变体和非人类灵长类动物错义变体,所述非人类灵长类动物错义变体出现在替代的非人类灵长类动物密码子序列上,所述替代的非人类灵长类动物密码子序列共用与人类匹配的参考密码子序列,其中所述常见人类错义变体在人类种群变体数据集上具有大于0.1%的次要等位基因频率MAF。

2. 根据权利要求1所述的方法,其中所述常见人类错义变体在从至少100000个人类取样的人类种群变体数据集上具有大于0.1%的次要等位基因频率MAF。

3. 根据权利要求2所述的方法,其中所述取样的人类属于不同的人类亚种群并且在各个人类亚种群变体数据集内所述常见人类错义变体具有大于0.1%的次要等位基因频率。

4. 根据权利要求3所述的方法,其中所述人类亚种群包括非洲人/非洲裔美国人AFR、美国人AMR、德系犹太人ASJ、东亚人EAS、芬兰人FIN、非芬兰欧洲人NFE、南亚人SAS、和其他OTH。

5. 根据权利要求1所述的方法,其中所述非人类灵长类动物错义变体包括来自多个非人类灵长类动物物种的错义变体,所述物种包括黑猩猩、倭黑猩猩、大猩猩、B红毛猩猩、S红毛猩猩、恒河猴和狨猴。

6. 根据权利要求1所述的方法,进一步包括,基于富集分析,接受特定的非人类灵长类动物物种,以将所述特定的非人类灵长类动物物种的错义变体包含在所述良性变体之中,其中所述富集分析包括对于所述特定的非人类灵长类动物物种将所述特定的非人类灵长类动物物种的同义变体的第一富集得分与所述特定的非人类灵长类动物物种的错义相同变体的第二富集得分进行比较;

其中,所述错义相同变体是共享与人类匹配的参考和替代密码子序列的错义变体;

其中,所述第一富集得分是通过确定次要等位基因频率小于0.1%的罕见同义变体与次要等位基因频率大于0.1%的常见同义变体的比率而产生的;以及

其中,所述第二富集得分是通过确定次要等位基因频率小于0.1%的罕见错义相同变体与次要等位基因频率大于0.1%的常见错义相同变体的比率而产生的。

7. 根据权利要求6所述的方法,其中罕见同义变体包括单例变体。

8. 根据权利要求6所述的方法,其中,所述富集分析进一步包括所述第一富集得分与所述第二富集得分之间的差在预定的范围内,接受所述特定的非人类灵长类动物物种,以将所述特定的非人类灵长类动物物种的错义变体包含在所述良性变体之中。

9. 根据权利要求8所述的方法,其中在所述预定范围内的所述差指示所述错义相同变体与所述同义变体处于相同程度的自然选择之下,因此与所述同义变体一样是良性的。

10. 根据权利要求6所述的方法,还包括重复地应用所述富集分析,以接受多个非人类灵长类动物物种,以将所述非人类灵长类动物物种的错义变体包含在所述良性变体中。

11. 根据权利要求6所述的方法,还包括使用同质性的卡方检验,以比较每个所述非人类灵长类动物物种的同义变体的第一富集得分和错义相同变体的第二富集得分。

12. 根据权利要求1所述的方法, 其中所述非人类灵长类动物错义变体的计数至少为100000。

13. 根据权利要求12所述的方法, 其中所述非人类灵长类动物错义变体的计数为385236。

14. 根据权利要求1所述的方法, 其中所述常见人类错义变体的计数至少是50000。

15. 根据权利要求14所述的方法, 其中所述常见人类错义变体的计数是83546。

16. 一种存储有构建变体致病性分类器的计算机程序指令的非暂时性计算机可读存储介质, 当所述计算机程序指令在处理器上执行时实施一种方法, 所述方法包括:

使用参考蛋白质序列和替代蛋白质序列的良性训练实例和致病性训练实例作为输入来训练基于卷积神经网络的变体致病性分类器, 所述基于卷积神经网络的变体致病性分类器在耦接至存储器的多个处理器上运行, 其中所述替代蛋白质序列是从良性变体和致病性变体产生的; 以及

其中所述良性变体包括常见人类错义变体和非人类灵长类动物错义变体, 所述非人类灵长类动物错义变体出现在替代的非人类灵长类动物密码子序列上, 所述替代的非人类灵长类动物密码子序列共用与人类匹配的参考密码子序列, 其中所述常见人类错义变体在人类种群变体数据集上具有大于0.1%的次要等位基因频率MAF。

17. 根据权利要求16所述的非暂时性计算机可读存储介质, 实施所述方法进一步包括, 基于富集分析, 接受特定的非人类灵长类动物物种, 以将所述特定的非人类灵长类动物物种的错义变体包含在所述良性变体之中, 其中所述富集分析包括对于特定的非人类灵长类动物物种将所述特定的非人类灵长类动物物种的同义变体的第一富集得分与所述特定的非人类灵长类动物物种的错义相同变体的第二富集得分进行比较;

其中, 所述错义相同变体是共享与人类匹配的参考和替代密码子序列的错义变体;

其中, 所述第一富集得分是通过确定次要等位基因频率小于0.1%的罕见同义变体与次要等位基因频率大于0.1%的常见同义变体的比率而产生的; 以及

其中, 所述第二富集得分是通过确定次要等位基因频率小于0.1%的罕见错义相同变体与次要等位基因频率大于0.1%的常见错义相同变体的比率而产生的。

18. 根据权利要求17所述的非暂时性计算机可读存储介质, 实施所述方法进一步包括: 使用同质性的卡方检验, 以比较每个所述非人类灵长类动物物种的同义变体的第一富集得分和错义相同变体的第二富集得分。

19. 一种用于构建变体致病性分类器的系统, 包括一个或多个耦接至存储器的处理器, 所述存储器装载有构建变体致病性分类器的计算机指令, 当其在所述处理器上执行时, 所述指令实现下列动作, 所述动作包括:

使用参考蛋白质序列和替代蛋白质序列的良性训练实例和致病性训练实例作为输入来训练基于卷积神经网络的变体致病性分类器, 所述基于卷积神经网络的变体致病性分类器在耦接至存储器的多个处理器上运行, 其中所述替代蛋白质序列是从良性变体和致病性变体产生的; 以及

其中所述良性变体包括常见人类错义变体和非人类灵长类动物错义变体, 所述非人类灵长类动物错义变体出现在替代的非人类灵长类动物密码子序列上, 所述替代的非人类灵长类动物密码子序列共用与人类匹配的参考密码子序列, 其中所述常见人类错义变体在人

类种群变体数据集上具有大于0.1%的次要等位基因频率MAF。

20. 根据权利要求19所述的系统,进一步执行以下动作:基于富集分析,接受特定的非人类灵长类动物物种,以将所述特定的非人类灵长类动物物种的错义变体包含在所述良性变体之中,其中所述富集分析包括对于所述特定的非人类灵长类动物物种将所述特定的非人类灵长类动物物种的同义变体的第一富集得分与所述特定的非人类灵长类动物物种的错义相同变体的第二富集得分进行比较;

其中,所述错义相同变体是共享与人类匹配的参考和替代密码子序列的错义变体;

其中,所述第一富集得分是通过确定次要等位基因频率小于0.1%的罕见同义变体与次要等位基因频率大于0.1%的常见同义变体的比率而产生的;以及

其中,所述第二富集得分是通过确定次要等位基因频率小于0.1%的罕见错义相同变体与次要等位基因频率大于0.1%的常见错义相同变体的比率而产生的。

21. 一种构建用于变体分类的基于卷积神经网络的分类器的计算机实现的方法,所述方法包括:

使用基于后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,所述基于后向传播的梯度更新技术逐渐地使所述基于卷积神经网络的分类器的输出与相应的给定标签匹配,所述基于卷积神经网络的分类器在耦接至存储器的多个处理器上运行;

其中所述基于卷积神经网络的分类器包括残差块组;

其中卷积窗口尺寸在所述残差块组之间变化;

其中空洞卷积率在所述残差块组之间变化;

其中所述训练数据包括转译序列对,所述转译序列对是从良性变体和致病性变体产生的并被用作良性训练实例和致病性训练实例;以及

其中所述良性变体包括常见人类错义变体和非人类灵长类动物错义变体,所述非人类灵长类动物错义变体出现在替代的非人类灵长类动物碱基三联体序列上,所述替代的非人类灵长类动物碱基三联体序列共用与人类匹配的参考碱基三联体序列,其中所述常见人类错义变体在人类种群变体数据集上具有大于0.1%的次要等位基因频率MAF。

22. 一种存储有构建用于变体分类的基于卷积神经网络的分类器的计算机程序指令的非暂时性计算机可读存储介质,当在处理器上执行所述指令时实施一种方法,所述方法包括:

使用基于后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,所述基于后向传播的梯度更新技术逐渐地使所述基于卷积神经网络的分类器的输出与相应的给定标签匹配,所述基于卷积神经网络的分类器在耦接至存储器的多个处理器上运行;

其中所述基于卷积神经网络的分类器包括残差块组;

其中卷积窗口尺寸在所述残差块组之间变化;

其中空洞卷积率在所述残差块组之间变化;

其中所述训练数据包括转译序列对,所述转译序列对是从良性变体和致病性变体产生的并被用作良性训练实例和致病性训练实例;以及

其中所述良性变体包括常见人类错义变体和非人类灵长类动物错义变体,所述非人类灵长类动物错义变体出现在替代的非人类灵长类动物碱基三联体序列上,所述替代的非人类灵长类动物碱基三联体序列共用与人类匹配的参考碱基三联体序列,其中所述常见人类

错义变体在人类种群变体数据集上具有大于0.1%的次要等位基因频率MAF。

23.一种构建用于变体分类的基于卷积神经网络的分类器的系统,包括耦接至存储器的一个或多个处理器,所述存储器装载有构建用于变体分类的基于卷积神经网络的分类器的计算机指令,当所述指令在所述处理器上执行时实施动作,所述动作包括:

使用基于后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,所述基于后向传播的梯度更新技术逐渐地使所述基于卷积神经网络的分类器的输出与相应的给定标签匹配,所述基于卷积神经网络的分类器在耦接至存储器的多个处理器上运行;

其中所述基于卷积神经网络的分类器包括残差块组;

其中卷积窗口尺寸在所述残差块组之间变化;

其中空洞卷积率在所述残差块组之间变化;

其中所述训练数据包括转译序列对,所述转译序列对是从良性变体和致病性变体产生的并被用作良性训练实例和致病性训练实例;以及

其中所述良性变体包括常见人类错义变体和非人类灵长类动物错义变体,所述非人类灵长类动物错义变体出现在替代的非人类灵长类动物碱基三联体序列上,所述替代的非人类灵长类动物碱基三联体序列共用与人类匹配的参考碱基三联体序列,其中所述常见人类错义变体在人类种群变体数据集上具有大于0.1%的次要等位基因频率MAF。

基于深度学习的深度卷积神经网络训练方法

[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Hong Gao、Kai-How Farh、Laksshman Sundaram和Jeremy Francis McRae于2017年10月16日提交的标题为“使用大规模良性训练数据训练深度致病性分类器”的美国临时专利申请No.62/573,144(代理人案卷号:ILLM 1000-1/IP-1611-PRV);由Kai-How Farh、Laksshman Sundaram、Samskruthi Reddy Padigepati和Jeremy Francis McRae于2017年10月16日提交的标题为“基于深度卷积神经网络(CNNS)的致病性分类器”的美国临时专利申请No.62/573,149(代理人案卷号:ILLM 1000-2/IP-1612-PRV);由Hong Gao、Kai-How Farh、Laksshman Sundaram和Jeremy Francis McRae于2017年10月16日提交的标题为“生成大规模致病训练数据的深度半监督学习”的美国临时专利申请No.62/573,153(代理人案卷号:ILLM 1000-3/IP-1613-PRV);以及由Kishore Jaganathan、Kai-How Farh、Sofia Kyriazopoulou Panagiotopoulou和Jeremy Francis McRae于2017年11月7日提交的标题为“使用深度卷积神经网络(CNN)的基因组数据致病性分类”的美国临时专利申请No.62/582,898(代理人案卷号:ILLM 1000-4/IP-1618-PRV)。出于所有目的,这些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Laksshman Sundaram、Kai-How Farh、Hong Gao、Samskruthi Reddy Padigepati和Jeremy Francis McRae于2018年10月15日同时提交的标题为“用于变体分类的深度卷积神经网络”的PCT专利申请号PCT/US2018/_____(代理人案卷号:ILLM 1000-9/IP-1612-PCT),随后被公布为PCT公开号WO_____。

[0008] 由Laksshman Sundaram、Kai-How Farh、Hong Gao和Jeremy Francis McRae于2018年10月15日同时提交的标题为“用于训练一套深度卷积神经网络的半监督学习”的PCT专利申请号PCT/US2018/_____(代理人案卷号:ILLM 1000-10/IP-1613-PCT),随后被公布为PCT公开号WO_____。

[0009] 由Hong Gao、Kai-How Farh、Laksshman Sundaram和Jeremy Francis McRae同时提交的标题为“基于深度学习的深度卷积神经网络训练方法”的美国非临时专利申请(代理人案卷号:ILLM 1000-5/IP-1611-US)。

[0010] 由Laksshman Sundaram、Kai-How Farh、Hong Gao、Samskruthi Reddy Padigepati和Jeremy Francis McRae同时提交的标题为“用于变体分类的深度卷积神经网络”的美国非临时专利申请(代理人案卷号:ILLM 1000-6/IP-1612-US)。

[0011] 由Laksshman Sundaram、Kai-How Farh、Hong Gao和Jeremy Francis McRae同时提交的标题为“用于训练一套深度卷积神经网络的半监督学习”的美国非专利申请(代理人

案卷号:ILLM 1000-7/IP-1613-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.Vanhoecke,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.Işgum,“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 A DEEP 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] ASCII文本格式的以下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:

	补充表 1.txt	2018 年 10 月 2 日	13 KB
	补充表 2.txt	2018 年 10 月 2 日	13 KB
	补充表 3.txt	2018 年 10 月 2 日	11 KB
	补充表 4.txt	2018 年 10 月 2 日	13 KB
	补充表 6.txt	2018 年 10 月 2 日	12 KB
	补充表 7.txt	2018 年 10 月 2 日	44 KB
[0040]	补充表 13.txt	2018 年 10 月 2 日	119 KB
	补充表 18.txt	2018 年 10 月 2 日	35 KB
	补充表 20.txt	2018 年 10 月 2 日	1027 KB
	补充表 20 汇总.txt	2018 年 10 月 2 日	9 KB
	补充表 21.txt	2018 年 10 月 2 日	24 KB
	补充表 21.txt	2018 年 10 月 2 日	24 KB
	补充表 18.txt	2018 年 10 月 4 日	35 KB
	数据文件 S1.txt	2018 年 10 月 4 日	138 MB
	数据文件 S2.txt	2018 年 10 月 4 日	980 MB
[0041]	数据文件 S3.txt	2018 年 10 月 4 日	1.01 MB
	数据文件 S4.txt	2018 年 10 月 4 日	834 KB
	致病性_预测_模型.txt	2018 年 10 月 4 日	8.24 KB

[0042] 补充表1:源自分析中使用的每个物种的变体的详情。该表包括了这些数据源各自渠道的中间结果。注:该表列在补充表1.txt中。

[0043] 补充表2:以常见人类等位基因频率存在于其他物种中的错义变体的损耗。该损耗是使用人和其他物种之间状态相同的变体,基于与罕见变体(<0.1%)相比常见变体(>0.1%)中的错义:同义比率计算的。注:该表列在补充表2.txt中。

[0044] 补充表3:以常见人类等位基因频率存在于其他物种中的错义变体的损耗,仅限于人类和其他哺乳动物之间平均核苷酸保守性(conservation)>50%的基因。该损耗是使用人和其他物种之间状态相同的变体,基于与罕见变体(<0.1%)相比常见变体(>0.1%)中的错义:同义比率计算的。注:该表列在补充表3.txt中。

[0045] 补充表4:以常见人类等位基因频率在相关物种对中作为固定替代存在的错义变体的损耗。该损耗是使用人和相关物种对之间状态相同的变体,基于与罕见变体(<0.1%)

相比常见变体(>0.1%)中的错义:同义比率计算的。注:该表列在补充表3.txt中。

[0046] 补充表6:SCN2A基因的结构域特异性注释。Wilcoxon秩和p值表示与整个蛋白质相比特异性结构域的PrimateAI得分发散。以粗体突出显示的结构域涵盖该蛋白质的约7%,但具有大多数的ClinVar致病注释。这与各结构域的平均PrimateAI得分关联良好,并且是基于PrimateAI模型的前3个致病结构域。注:该表列在补充表6.txt中。

[0047] 补充表7:计算等位基因频率对预期错义:同义比率影响时使用的原始计数。同义变体和错义变体的预期计数是基于内含子区域中的变体,使用三核苷酸背景计算以控制突变率和基因转变。注:该表列在补充表.xlsx中。

[0048] 补充表13:用于训练3-态二级结构和3-态溶剂可及性预测的深度学习模型的蛋白质名称清单标签栏表明各蛋白质是否用于模型训练的训练/验证/测试阶段。注:该表列在补充表13.txt中。

[0049] 补充表18:仅按蛋白质截短变异计算时在DDD研究中对疾病相关性名义上显著($p < 0.05$)的605个基因的清单注:该表列在补充表18.txt中。

[0050] 补充表20:具有至少一个观察到的新生突变(DNM)的所有基因中每个基因的DNM富集测试结果。提供了包含所有DNM时和删除PrimateAI得分<0.803的错义DNM后的P值。按相似方式提供了FDR校正的P值。包括仅由DDD群组和由完整荟萃分析群组观察到的蛋白截短(PTV)和错义DNM的计数。在以下情况时还包括观察到的和预期的错义DNM的相似计数:首先是包含所有错义DNM时,其次是删除PrimateAI得分<0.803的所有错义DNM时。注:该表列在补充表20.txt和补充表20汇总.txt中。

[0051] 补充表21:FDR<0.1的基因中新生突变的富集测试结果。包括观察到的蛋白截短(PTV)新生突变的计数和其他蛋白质改变新生突变的计数,一次具有所有错义新生突变,一次仅具有受损的错义突变。提供了包含所有错义位点时的P值Vs排除低得分错义位点后的P值。注:该表列在补充表21.txt中。

[0052] 数据文件S1:其他物种中存在的所有变体的列表。ClinVar显著性栏列出了现有的无冲突ClinVar注释。注:该表列在数据文件S1.txt中。

[0053] 数据文件S2:源自相关物种对的所有固定替换的列表。注:该表列在数据文件S2.txt中。

[0054] 数据文件S3:与灵长类动物的IBS的隐藏良性测试变体的列表。良性测试变体是与1个以上灵长类物种IBS的非常见人类变体。注:该表列在数据文件S3.txt中。

[0055] 数据文件S4:与隐藏良性测试变体匹配的无标记变体IBS及灵长类动物的列表。将无标记变体与良性测试变体匹配,用于突变率、覆盖率偏差和与灵长类物种比对。注:该表列在数据文件S4.txt中。

[0056] 致病性_预测_模型:用Python编程语言编写的代码,实现按照一个实现方式所公开的技术。注:该表列在致病性_预测_模型.txt中。

技术领域

[0057] 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

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

[0058] 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。

[0059] 机器学习

[0060] 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:

$$[0061] \quad Y = f(X) + \epsilon$$

[0062] 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。

[0063] 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 。

$$[0064] \quad \hat{Y} = \hat{f}(X)$$

[0065] 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。

[0066] 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。

[0067] 神经网络

[0068] 图1A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。

[0069] 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:

[0070] T.Ching et al., Opportunities And Obstacles For Deep Learning In Biology And Medicine, www.biorxiv.org:142760, 2017;

- [0071] Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep Learning For Computational Biology. Mol Syst Biol. 2016;12:878;
- [0072] Park Y, Kellis M. 2015 Deep Learning For Regulatory Genomics. Nat. Biotechnol. 33, 825-826. (doi:10.1038/nbt.3313);
- [0073] Min, S., Lee, B. & Yoon, S. Deep Learning In Bioinformatics. Brief. Bioinform. bbw068 (2016);
- [0074] Leung MK, DeLong A, Alipanahi B et al. Machine Learning In Genomic Medicine: A Review of Computational Problems and Data Sets 2016; and
- [0075] Libbrecht MW, Noble WS. Machine Learning Applications In Genetics and Genomics. Nature Reviews Genetics 2015;16(6):321-32.

附图说明

[0076] 在附图中, 相同的附图标记在不同视图中通常指代相同的部分。而且, 附图不一定按比例绘制, 而是通常将重点放在说明所公开的技术的原理上。在以下描述中, 参考以下附图描述所公开的技术的各种实现方式, 在附图中:

- [0077] 图1A示出了具有多层的前馈神经网络的一个实现方式。
- [0078] 图1B描述了卷积神经网络工作的一个实现方式。
- [0079] 图1C描述了按照所公开技术的一个实现方式训练卷积神经网络的框图。
- [0080] 图1D是按照所公开技术的一个实现方式的子采样层(平均化/最大池化)的一个实现方式。
- [0081] 图1E示出了按照所公开技术的一个实现方式的ReLU非线性层的一个实现方式。
- [0082] 图1F描述了卷积层的双层卷积的一个实现方式。
- [0083] 图1G描述了通过特征映射图添加在下游重新注入先前信息的残差连接。
- [0084] 图1H描述了残差块和跳跃连接的一个实现方式。
- [0085] 图1I示出了批归一化正向传递。
- [0086] 图1J说明了测试时的批归一化转换。
- [0087] 图1K示出了批归一化反向传递。
- [0088] 图1L描述了卷积或全连接层前后批归一化层的使用。
- [0089] 图1M示出了1D卷积的一个实现方式。
- [0090] 图1N说明了全局平均池化(GAP)的工作原理。
- [0091] 图1O说明了空洞卷积。
- [0092] 图1P示出了堆栈的空洞卷积的一个实现方式。
- [0093] 图1Q示出了能够运行所公开技术的示例计算环境。
- [0094] 图2示出了致病性预测的深度残差网络(本文中称为“PrimateAI”)的示例结构。
- [0095] 图3描述了PrimateAI即致病性分类的深度学习网络结构的示意图。
- [0096] 图4A、4B和4C是补充表16, 示出了致病性预测深度学习模型PrimateAI的示例模型结构详情。
- [0097] 图5和6说明了用于预测蛋白质二级结构和溶剂可及性的深度学习网络。
- [0098] 图7A和7B是补充表11, 示出了3-态二级结构预测深度学习(DL)模型的示例模型结

构详情。

[0099] 图8A和8B是补充表12,示出了3-态溶剂可及性预测深度学习模型的示例模型结构详情。

[0100] 图9描述了从良性变体和致病性变体生成参考序列和替代蛋白序列的一个实现方式。

[0101] 图10示出了比对参考序列和替代蛋白序列的一个实现方式。

[0102] 图11是生成位置频率矩阵(缩写PFM)的一个实现方式,其中PFM也称为位置加权矩阵(缩写PWM)或位置特异性评分矩阵(缩写PSSM)。

[0103] 图12、13、14和15示出了二级结构和溶剂可及性子网络的处理。

[0104] 图16变体致病性分类器的运行。如本文中所用,术语变体也称为单核苷酸多态性(缩写SNP),通常也称为单核苷酸变体(缩写SNV)。

[0105] 图17说明了残差框图。

[0106] 图18描述了二级结构和溶剂可及性子网络的神经网络结构。

[0107] 图19示出了变体致病性分类器的神经网络结构。

[0108] 图20示出了SCN2A基因中每个氨基酸位置处的预测致病性得分,对主要功能结构域进行注释。

[0109] 图21D示出了对于一个包含10000个从训练保留的常见灵长类动物变体的测试组预测良性结果的分类器的比较。

[0110] 图21E说明了与未患病的兄弟姐妹相比,解读发育障碍(DDD)患者中出现的新生错义变体的PrimateAI预测得分的分布,以及相应的Wilcoxon秩和检验P值。

[0111] 图21F描述了与对照例(control)相比分离DDD病例中新生错义变体时分类器的比较。示出了每个分类器的Wilcoxon秩和检验P值。

[0112] 图22A示出了605个对新生蛋白截短变异显著($P < 0.05$)的相关基因中,DDD群组的患病个体中超预期新生错义突变的富集。

[0113] 图22B描述了605个相关基因中与未患病兄弟姐妹相比DDD患者出现的新生错义变体的PrimateAI预测得分的分布,以及相应的Wilcoxon秩和检验P值。

[0114] 图22C示出了605个基因中与对照例相比分离各病例中新生错义变体时各分类器的比较。

[0115] 图22D描述了通过每个分类器所示的曲线下面积(AUC)进行的各分类器比较(如受试者工作特征曲线上所示)。

[0116] 图22E描述了每个分类器的分离准确度和曲线下面积(AUC)。

[0117] 图23A、23B、23C和23D示出了用于训练的数据对分类准确度的影响。

[0118] 图24示出了序列覆盖率对常见灵长类动物变体确认影响的校正。

[0119] 图25A、25B、25C和26描述了所公开神经网络对蛋白质基序的识别。图26包括线图,示出了扰乱变体中和变体附近每个位置对该变体的预测深度学习得分的影响。

[0120] 图27说明了模拟BLOSUM62和Grantham得分矩阵的权重相关性模式。

[0121] 图28A、28B和28C示出了深度学习网络PrimateAI和其他分类器的性能评估。

[0122] 图29A和29B说明了四个分类器的预测得分的分布。

[0123] 图30A、30B和30C比较了PrimateAI网络和其他分类器分离605个疾病相关基因中

致病性变体和良性变体的准确度。

[0124] 图31A和31B说明了分类器在人类专家精选的ClinVar变体中的性能和在经验一个实现方式中性能间的关联。

[0125] 图32是补充表14,示出了3-态二级结构和3-态溶剂可及性预测模型在蛋白质数据库的注释样品的性能。

[0126] 图33是补充表15,示出了使用DSSP数据库的人类蛋白质的注释二级结构标记进行的深度学习网络的性能比较。

[0127] 图34是补充表17,示出了我们评估的20个分类器中每个分类器在DDD病例Vs对照例的10000个隐藏灵长类动物变体准确度值和新生变体P值。

[0128] 图35是补充表19,示出了在DDD病例Vs对照例数据库中不同分类器对新生变体的性能比较(限于605个疾病相关基因)。

[0129] 图36示出了所公开半监督学习器的计算环境。

[0130] 图37、38、39、40和41示出了所公开半监督学习的各个循环。

[0131] 图42是迭代平衡采样流程的说明。

[0132] 图43说明了用于生成良性一个实现方式的计算环境的一个实现方式。

[0133] 图44描述了生成良性人错义SNP的一个实现方式。

[0134] 图45示出了人同源错义SNP的一个实现方式。具有与人类匹配的参考和替代密码子的非人类物种中的错义SNP。

[0135] 图46描述了将具有与人类匹配的参考密码子的非人类灵长类物种(例如黑猩猩)的SNP分类为良性的一个实现方式。

[0136] 图47描述了计算富集得分并比较它们的一个实现方式。

[0137] 图48描述了良性SNP一个实现方式的一个实现方式。

[0138] 图49A、49B、49C、49D和49E描述了人类等位基因频率谱中的错义/同义比率。

[0139] 图50A、50B、50C和50D示出了与其他物种状态相同的错义变体的纯化选择。

[0140] 图51示出了不存在纯化选择时人类等位基因频率谱中的预期错义:同义比率。

[0141] 图52A、52B、52C和52D描述了CpG和非CpG变体的错义:同义比率。

[0142] 图53、54和55说明了与六种灵长类动物状态相同的人类变体的错义:同义比率。

[0143] 图56是一种模拟,示出了通过提高所调查人类群组规模发现的新常见错义变体饱和度和度。

[0144] 图57示出了PrimateAI在基因组的不同保守谱中的准确度。

[0145] 图58是补充表5,示出了常见人类变体和非人类灵长类动物中存在的变体对标记良性训练一个实现方式的贡献。

[0146] 图59是补充表8,示出了等位基因频率对预期错义:同义比率的影响。

[0147] 图60是补充表9,示出了ClinVar分析。

[0148] 图61是补充表10,示出了按照一个实现方式在ClinVar中发现的其他物种的错义变体数量。

[0149] 图62是表1,示出了发现智力残疾的14个额外候选基因的一个实现方式。

[0150] 图63是表2,示出了ClinVar中致病性和良性变体间Grantham得分平均差异的一个实现方式。

- [0151] 图64示出了每个基因富集分析的一个实现方式。
- [0152] 图65示出了基因组富集分析的一个实现方式。
- [0153] 图66是能够用于执行所公开技术的计算机系统的简化框图。

具体实施方式

[0154] 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。

[0155] 介绍

[0156] 卷积神经网络

[0157] 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。

[0158] 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。

[0159] 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。

[0160] 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。

[0161] 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。

[0162] 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。

[0163] 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 1×1 、 3×3 或 5×5 ;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。

[0164] 卷积通过在3D输入特征映射图上滑动大小为 3×3 或 5×5 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 3×3 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1B中详细描述了整个过程。

[0165] 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 $(m \times n)$ 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。

[0166] 训练卷积神经网络

[0167] 图1C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)的较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。

[0168] 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:

$$[0169] \quad \Delta w_i = x_i \delta$$

[0170] 其中 $\delta = (\text{真实值}) - (\text{实际输出})$

[0171] 在一个实现方式中,训练规则被定义为:

$$[0172] \quad w_{nm} \leftarrow w_{nm} + \alpha(t_m - \phi_m)a_n$$

[0173] 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。

[0174] 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。

[0175] 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:

$$[0176] \quad \varphi = f(h) = \frac{1}{1 + e^{-h}}$$

[0177] 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:

$$[0178] \quad \frac{\partial \varphi}{\partial h} = \varphi(1 - \varphi)$$

[0179] 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:

$$[0180] \quad \varphi_m = \frac{1}{1 + e^{-h_m}}$$

$$[0181] \quad h_m = \sum_{n=1}^N a_n w_{nm}$$

[0182] 对所有隐藏层执行此操作以获取激活,描述如下:

$$[0183] \quad \varphi_k = \frac{1}{1 + e^{-h_k}}$$

$$[0184] \quad h_k = \sum_{m=1}^M \varphi_m v_{mk}$$

[0185] 然后,计算每层的误差和正确权重。输出的误差计算为:

$$[0186] \quad \delta_{ok} = (t_k - \varphi_k) \varphi_k (1 - \varphi_k)$$

[0187] 隐藏层中的误差被计算为:

$$[0188] \quad \delta_{hm} = \varphi_m (1 - \varphi_m) \sum_{k=1}^K v_{mk} \delta_{ok}$$

[0189] 输出层的权重更新为:

$$[0190] \quad v_{mk} \leftarrow v_{mk} + \alpha \delta_{ok} \varphi_m$$

[0191] 使用以下学习率 α 来更新隐藏层的权重:

$$[0192] \quad v_{nm} \leftarrow w_{nm} + \alpha \delta_{hm} a_n$$

[0193] 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 预测 $\hat{y}$ 的成本,损失函数被定义为 $l(\hat{y}, y)$,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:

$$[0194] \quad v_{t+1} = \mu v_t - \alpha \frac{1}{n} \sum_{i=1}^N \nabla_{w_t} Q(z_t, w_t)$$

$$[0195] \quad w_{t+1} = w_t + v_{t+1}$$

[0196] 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。

[0197] 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:

$$[0198] \quad \mathbf{V}_{t+1} = \mu \mathbf{V} - \alpha \nabla_w Q(\mathbf{z}_t, \mathbf{w}_t)$$

$$[0199] \quad \mathbf{w}_{t+1} = \mathbf{w}_t + \mathbf{V}_{t+1}$$

[0200] 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。

[0201] 卷积层

[0202] 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 $f * g$

[0203] 被定义为:

$$[0204] \quad (f * g)(i) = \sum_{j=1}^m g(j) \cdot f(i - j + m/2)$$

[0205] 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。

[0206] 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。

[0207] 子采样层

[0208] 图1D是按照所公开技术的一个实现方式进行子采样层的一个实现方式。子采样层减少了通过卷积层提取的特征分辨率,使提取特征或特征映射在噪声和变形方面更稳健。在一个实现方式中,子采样层采用两类池化操作:平均池化和最大池化。池化操作将输入分成不重叠的二维空间。就平均池化而言,计算该区域中四个数值的平均值。就最大池化而言,选择四个数值中的最大值。

[0209] 在一个实现方式中,子采样层包括上一层中一组神经元的池化操作:在最大池化中将其输出映射到仅一个输入,在平均池化中将其输出映射到输入的平均值。在最大池化中,池化神经元的输出是输入中驻留的最大值,如以下所述:

$$[0210] \quad \varphi_o = \max(\varphi_1, \varphi_2, \dots, \varphi_N)$$

[0211] 在以上方程中,N是神经元组中元素的总数。

[0212] 在平均池化中,池化神经元的输出是随输入神经元组驻留的输入值的平均值,如下所述:

$$[0213] \quad \varphi_o = \frac{1}{N} \sum_{n=1}^N \varphi_n$$

[0214] 在以上方程中,N是输入神经元组中元素的总数。

[0215] 在图1D中,输入的规模为 4×4 。就 2×2 子采样而言,将 4×4 图像分成四个不重叠的 2×2 规模矩阵。就平均池化而言,四个数值的平均值是整数输出。就最大池化而言, 2×2 矩阵中四个数值的最大值是整数输出。

[0216] 非线性层

[0217] 图1E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(sigmoid)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:

$$[0218] \quad \varphi(h) = \max(h, 0)$$

$$[0219] \quad \varphi(h) = \begin{cases} h & \text{如果 } h > 0 \\ 0 & \text{如果 } h \leq 0 \end{cases}$$

[0220] 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:

$$[0221] \quad \varphi(h) = (a + bh)^c$$

[0222] 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。

[0223] 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:

$$[0224] \quad \varphi(h) = \frac{1}{1 + e^{-\beta h}}$$

[0225] 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。

[0226] 卷积示例

[0227] 图1F描绘了卷积层的两层卷积的一个实现方式。在图1F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。

[0228] 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。

[0229] 前向传递

[0230] 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:

$$[0231] \quad O_{x,y}^{(l,k)} = \tanh\left(\sum_{t=0}^{f-1} \sum_{r=0}^{k_h} \sum_{c=0}^{k_w} W_{(r,c)}^{(k,t)} O_{(x+r,x+c)}^{(l-1,t)} + Bias^{(l,k)}\right)$$

[0232] 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:

$$[0233] \quad O_{x,y}^{(l,k)} = \tanh\left(W^{(k)} \sum_{r=0}^{S_h} \sum_{c=0}^{S_w} O_{(x \times S_h + r, y \times S_w + c)}^{(l-1,k)} + Bias^{(l,k)}\right)$$

[0234] 第1个输出层的第i个神经元的输出由以下等式确定:

$$[0235] \quad O_{(l,i)} = \tanh\left(\sum_{j=0}^H O_{(l-1,j)} W_{(i,j)}^l + Bias^{(l,i)}\right)$$

[0236] 后向传播

[0237] 输出层中的第k个神经元的输出偏差由以下等式确定:

$$[0238] \quad d(O_k^o) = y_k - t_k$$

[0239] 输出层中的第k个神经元的输入偏差由以下等式确定:

$$[0240] \quad d(I_k^o) = (y_k - t_k) \varphi'(v_k) = \varphi'(v_k) d(O_k^o)$$

[0241] 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:

$$[0242] \quad \Delta W_{k,x}^o = d(I_k^o) y_{k,x}$$

$$[0243] \quad \Delta Bias_k^o = d(I_k^o)$$

[0244] 隐藏层中的第k个神经元的输出偏差由以下等式确定:

$$[0245] \quad d(O_k^H) = \sum_{i=0}^{i < 84} d(I_i^o) W_{i,k}$$

[0246] 隐藏层中的第k个神经元的输入偏差由以下等式确定:

$$[0247] \quad d(I_k^H) = \varphi'(v_k) d(O_k^H)$$

[0248] 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：

$$[0249] \quad \Delta W_{m,x,y}^{H,k} = d(I_k^H) y_{x,y}^m$$

$$[0250] \quad \Delta Bias_k^H = d(I_k^H)$$

[0251] 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：

$$[0252] \quad d(O_{x,y}^{S,m}) = \sum_k d(I_{m,x,y}^H) W_{m,x,y}^{H,k}$$

[0253] 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：

$$[0254] \quad d(I_{x,y}^{S,m}) = \varphi'(v_k) d(O_{x,y}^{S,m})$$

[0255] 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：

$$[0256] \quad \Delta W^{S,m} = \sum_{x=0}^{fh} \sum_{y=0}^{fw} d(I_{[x/2],[y/2]}^{S,m}) O_{x,y}^{C,m}$$

$$[0257] \quad \Delta Bias^{S,m} = \sum_{x=0}^{fh} \sum_{y=0}^{fw} d(O_{x,y}^{S,m})$$

[0258] 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：

$$[0259] \quad d(O_{x,y}^{C,k}) = d(I_{[x/2],[y/2]}^{S,k}) W^k$$

[0260] 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：

$$[0261] \quad d(I_{x,y}^{C,k}) = \varphi'(v_k) d(O_{x,y}^{C,k})$$

[0262] 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：

$$[0263] \quad \Delta W_{r,c}^{k,m} = \sum_{x=0}^{fh} \sum_{y=0}^{fw} d(I_{x,y}^{C,k}) O_{x+r,y+c}^{l-1,m}$$

$$[0264] \quad \Delta Bias^{C,k} = \sum_{x=0}^{fh} \sum_{y=0}^{fw} d(I_{x,y}^{C,k})$$

[0265] 残差连接

[0266] 图1G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。关于残差连接的其他信息可见K.He、

X.Zhang、S.Ren和J.Sun的“图像识别的深度残差学习”arXiv:1512.03385,2015,该文章出于所有目的通过引用并入本文,如同在本文中进行了完整阐述一样。

[0267] 残差学习和跳跃连接

[0268] 图1H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。关于残差框图和跳跃连接的其他信息可见A.V.D.Oord、S.Dieleman、H.Zen、K.Simonyan、O.Vinyals、A.Graves、N.Kalchbrenner、A.Senior和K.Kavukcuoglu的“WAVENET:原始比率的生成模型”arXiv:1609.03499,2016。

[0269] 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。

[0270] 空洞卷积

[0271] 图10说明了空洞卷积。空洞卷积,有时称为带孔卷积,它在字面上表示带孔。法文名称在算法中有其来源trous,它计算快速二进位小波变换。在这些类型的卷积层中,对应于滤波器感受野的输入不是相邻点。这在图10中说明。输入间距取决于伸缩系数。

[0272] WaveNet

[0273] Wavenet是用于生成原始音频波形的深度神经网络。

[0274] 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。

[0275] 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1P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
[0276]	FCN	1	$2.6 \cdot 10^5$	$2.6 \cdot 10^6$
	WN	1	1022	8176
	WN	2	1022	8176
	WN	4	508	4064

[0277] 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。

[0278] 批归一化

[0279] 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。

[0280] 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1L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。

[0281] 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。

[0282] 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。

[0283] 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。

[0284] 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。

[0285] 前向传递

[0286] 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1I中描绘了批归一化前向传递 f_{BN} 。

[0287] 在图1I中,相应地 μ_B 是批均值并且 σ_B^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。

[0288] 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1J所示表示批归一化变换。在图1J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。

[0289] 后向传递

[0290] 由于归一化是可微分操作,所以可以如图1K中所描绘的那样计算后向传递。

[0291] 1D卷积

[0292] 1D卷积从序列中提取局部1D子块或子序列,如图1M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。

[0293] 全局平均池化

[0294] 图1N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。

[0295] 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。

[0296] 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。

[0297] 基因组学中的深度学习

[0298] 基因变异可以帮助解释很多疾病。每个人都有特定的遗传密码,一组个体中存在很多基因变体。大部分不良基因变体已通过自然选择从基因组中删除。重要的是确定哪个遗传变异可能是致病性的或有害的。这将帮助研究人员关注可能致病的基因变异,并加快许多疾病的诊断和治愈步伐。

[0299] 对变体的性质和功能作用(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突飞猛进,但由于细胞类型特异性转录调控系统的复杂性,因此对变体功能后果的解读仍是一项巨大的挑战。

[0300] 过去数十年来生物化学技术的进步已经产生下一代测序(NGS)平台,它们能以比以往低得多的成本快速产生基因组数据。如此极大量的测序DNA仍然难以注释。当有大量的标记数据可用时,受监督的机器学习算法一般运行良好。在生物信息学和很多其他数据丰富的学科中,标记实例的过程是昂贵的;但是,未标记的实例是廉价和容易获得的。就标记数据量相对较小并且未标记数据量明显较大的情况而言,半监督学习是具有成本效益的人工标记替代方案。

[0301] 有机会使用半监督算法构建基于深度学习的致病性分类器,它可准确地预测变体的致病性。可能得到不含人类确认偏差的致病性变体数据库。

[0302] 关于致病性分类器,深度神经网络是一类人工神经网络,它们使用多种非线性的复杂转换层来成功地对高级别特征进行建模。深度神经网络通过反向传播提供反馈,所述反向传播携带观察到的输出和预测输出间的差异以调节参数。深度神经网络已经随着大规模训练数据集的可用、并行和分布式计算的能力以及复杂的训练算法发生进化。深度神经网络已经促进众多领域如计算机视觉、语音识别和自然语言处理等的重大进展。

[0303] 卷积神经网络(CNN)和循环神经网络(RNN)是深度神经网络的组成部分。卷积神经网络在图像识别中特别成功,其结构包括卷积层、非线性层和池化层。循环神经网络旨在通过各构建块如感知器、长短期记忆单元和门控循环单元间的环形连接,利用输入数据的序列信息。此外,已对有限背景下提出很多其他新兴的深度神经网络,例如深度时空神经网络、多维循环神经网络和卷积自动编码器。

[0304] 训练深度神经网络的目标是优化每一层的权重参数,逐渐将更简单的特征组合成复杂的特征,从而能够从数据中学习最适合的层次表示。优化流程的单个循环组织如下。首先,给定一个训练数据集后,正向传递按顺序计算每一层的输出并通过网络正向传播函数信号。在最终的输出层中,客观损失函数测量推定输出和给定标签之间的误差。为了使训练误差最小化,反向传递使用链式规则反向传播误差信号,并计算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。而批次梯度下降执行每个完整数据集的参数更新,随机梯度下降通过执行每一小组数据实例的更新提供随机近似。几种优化算法源于随机梯度下降。例如,Adagrad和Adam训练算法执行随机梯度下降,同时分别

根据每个参数的更新频率和梯度矩适应地改变学习速率。

[0305] 深度神经网络训练的另一个核心元素是正则化,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,权重衰减增加了客观损失函数的罚分项,使权重参数收敛于较小的绝对值。丢弃(Dropout)在训练过程中从神经网络中随机移除隐藏单元,并且可以被认为是可能子网络的集合。为提高丢弃能力,已经提出新的激活函数Maxout和用于称为rnnDrop的循环神经网络的丢弃变体。此外,批归一化方法通过归一化小批内的每次激活的标量特征并以每个均值和方差作为参数进行学习,提供了一种新的正则化方法。

[0306] 鉴于测序数据为多维和高维,深度神经网络因其广泛的适用性和高预测能力,在生物信息学研究中具有很大的应用前景。卷积神经网络已被用于解决基因组学中基于序列的问题,如基序发现、致病变体识别和基因表达推断。卷积神经网络使用了一种对研究DNA特别有用的权重分享策略,因为它可以捕获序列基序,这些基序是DNA中短的、反复出现的局部模式,推测具有显著的生物学功能。卷积神经网络的标志是使用卷积滤波器。与基于精心设计且人工制作的特征的传统分类方法不同,卷积滤波器执行适应性的特征学习,类似于将原始输入数据映射成知识的信息表示的过程。从这个意义上说,卷积滤波器起到一系列基序扫描器的作用,因为一组此类滤波器能够识别输入中的相关模式,并在训练程序中自我更新。循环神经网络可以捕获不同长度的序列数据(例如蛋白质或DNA序列)中的长程依赖性。

[0307] 因此,预测变体致病性的强力计算模型对基础科学和转化研究均有巨大的益处。

[0308] 常见的多态现象代表了自然试验,其适应性经过多代的自然选择检验。通过比较人类错义替换和同义替换的等位基因频率分布,我们发现非人类灵长类动物物种中错义变体以高等位基因频率的存在可靠地预测,该变体在人类种群中也处于中性选择。相反,随着进化距离增加,更远的物种中的常见变异会出现负向选择。

[0309] 我们采用来自6个非人类灵长类动物物种的常见变异训练半监督深度学习网络,该网络仅使用序列就能准确分类临床新生错义突变。根据500多个已知物种,灵长类谱系包含足够的常见变异,以便对意义不明的大多数人类变体的影响进行系统地建模。

[0310] 人类参考基因组包含7000多万个潜在蛋白质改变错义替换,其中绝大多数是尚未表征对人类健康影响的罕见突变。这些意义不明的变体对临床应用中的基因组解读构成挑战,并成为长期用于种群筛查测序和个体化医疗的障碍。

[0311] 将不同人类种群的常见变异归类是鉴定临床良性变异的有效策略,但现代人类中现有的常见变异受到我们物种远古瓶颈事件的限制。人类和黑猩猩有99%的序列同一性,表明对黑猩猩变体进行的自然选择,具备对人类中状态相同的变体影响进行建模的可能。人类种群中等位基因作用多形性的平均聚结时间是物种分化时间的一部分,因此除通过平衡选择维持单倍型的罕见情况外,自然发生的黑猩猩变异很大程度上探索了与人类变异不重叠的突变空间。

[0312] 最近从60706个人获得的汇总外显子组数据,使我们能够通过比较错义突变和同义突变的等位基因频率谱来检验这一假设。

[0313] ExAc中的单胎变体与使用三核苷酸背景调整突变率后通过新生突变预测的预期2.2:1错义:同义比率密切相符,但在更高的等位基因频率下,观察到的错义变体数量因通过自然选择从有害变体中滤出而减少。整个等位基因频谱的错义:同义比率模式表明,种群

频率 $<0.1\%$ 的大部分错义变体是轻度有害的,也就是说,致病性既不足以保证立即从种群中清除,中性也不足以允许以高等位基因频率存在,与更有限的种群数据的既往观察结果一致。由于对外显性遗传疾病可能为良性,因此除平衡选择和建立者效应引起各种有据可查的预期外,这些发现支持诊断实验室滤除等位基因频率大于 $0.1\%-1\%$ 的广泛经验实践。

[0314] 用与常见黑猩猩变体(在黑猩猩种群测序中观察到一次以上)状态相同的人类变体子集重复该分析,我们发现错义:同义比率在整个等位基因频率谱中大体上是恒定的。黑猩猩种群中这些变体的高等位基因频率表明,它们已经在黑猩猩中经受自然选择的筛选,它们对人类种群适应性的中性影响提供了令人信服的证据,表明在两个物种中错义变异的选择压力高度一致。在黑猩猩中观察到的较低错义:同义比率,与祖先黑猩猩种群中较大的有效种群规模一致,能够更有效地滤除轻度有害的变体。

[0315] 相比之下,罕见的黑猩猩变体(仅在黑猩猩种群测序中观察到一次)显示较高等位基因频率下错义:同义比率有中度下降。从人类变异数据模拟相同规模的群组,我们估计与该群组中多次观察到的变体的 99.8% 相比,该规模的群组中观察到一次的变体中仅 64% 在普通种群中具有大于 0.1% 的等位基因频率,表明并非所有罕见的黑猩猩变体都经受选择的筛选。总之,我们估计已确认的黑猩猩错义变体中有 16% 在普通种群中的等位基因频率小于 0.1% ,并且在更高的等位基因频率下发生负向选择。

[0316] 我们接下来表征与其他非人类灵长类物种(矮黑猩猩、大猩猩、猩猩、恒河猴和狨猴)中所见变异状态相同的人类变体。与黑猩猩相似,我们观察到错义:同义比率在整个等位基因频率范围内大致相等,高等位基因频率下错义变异的轻微损耗除外,由于少量罕见变体(约 $5-15\%$)的纳入这是可预期的。这些结果意味着,对错义变异的选择力在灵长类谱系内大体上一致,至少在新世界猕猴中如此,估计它们在约3500万年前已与人类祖先谱系分化。

[0317] 与其他灵长类动物中的变体状态相同的人类错义变体,在ClinVar中因良性结果而强力富集。排除注释未知或冲突的变体后,我们发现与一般情况下错义变异的 45% 相比,具有灵长类同源物的人类变体在ClinVar中约 95% 可能被注释为良性或可能良性。来自非人类灵长类动物且分类为致病性的ClinVar变体中一小部分,与通过确认来自相似规模的健康人类群组的罕见变体所观察到的致病性ClinVar变体部分相似。这些变体中一大部分被注释为致病性或可能致病性,表明在大型等位基因频率数据库出现之前收到了它们的分类,并且目前可能被进行了不同的精选。

[0318] 人类遗传学领域长期以来都依赖模型生物来推断人类突变的临床影响,但是与大多数遗传可追溯模型的漫长进化距离使人们担心这些研究结果在多大程度上可以推广到人类。为了检查人类和更远物种中错义变体的自然选择一致性,我们将分析扩展到灵长类谱系以外,以便大体上纳入源自四个额外哺乳动物物种(小鼠、猪、山羊、牛)和两个更远的脊椎动物物种(鸡、斑马鱼)的常见变异。与之前的灵长类分析相反,我们发现与罕见等位基因频率相比,错义变异在常见等位基因频率下明显损耗,尤其是在较大的进化距离,表明在更远的物种中大部分常见错义变异在人类种群将出现负向选择。尽管如此,在更远的脊椎动物中观察到的错义变体仍然增加了良性结果的可能性,因为自然选择导致的常见错义变体部分远低于基线时人类错义变体约 50% 的损耗。与这些结果一致,我们发现在小鼠、犬、猪和母牛中观察到的人类错义变体在ClinVar中约 85% 可能被注释为良性或可能良性,而

灵长类动物变异为95%，ClinVar数据库整体为45%。

[0319] 在不同进化距离上密切相关物种对的存在，也提供了评估人类种群中固定错义替换的功能后果的机会。在哺乳动物家族树上的近亲物种对(分支长度 <0.1)中，我们发现与罕见等位基因频率相比，固定错义变异在常见等位基因频率下损耗，表明在人类甚至在灵长类谱系内，大部分的物种间固定替换将是非中性的。错义损耗程度的比较表明，物种间固定替换的中性显著低于物种内多态性。有趣的是，与物种内常见的多态性相比，近缘哺乳动物之间的物种间变异在ClinVar中致病性实质上并未更强(83%可能被注释为良性或可能是良性)，表明这些变化并未使蛋白质功能丧失，而是反映了蛋白质功能的调节，赋予物种特异性适应性优势。

[0320] 大量意义不明的可能变体以及准确变体分类对临床应用的至关重要性，促使人们多次尝试用机器学习来处理问题，但是这些努力很大程度上受限于人类常见变体的数量不足和精选数据库中注释的质量可疑。来自6个非人类灵长类动物的变异贡献了300,000多个与常见人类变异不重叠且大体上为良性后果的独特错义变体，极大地扩大了可以用于机器学习方法的训练数据集的规模。

[0321] 与采用大量的人机工程特征和元分类器的早期模型不同，我们应用了简单的深度学习残差网络，它只以靶变体侧翼的氨基酸序列和其他物种的直系同源序列比对作为输入。为了给该网络提供蛋白质结构相关的信息，我们训练了两个独立的网络，以便仅从序列中学习二级结构和溶剂可及性，并将这些作为更大深度学习网络的子网络，以预测对蛋白质结构的影响。以序列为起点避免了蛋白质结构和功能结构域注释的潜在偏差，这些偏差可能未完全确认或未一致地应用。

[0322] 我们使用半监督学习来克服训练集只包含带有良性标签的变体的问题，通过初步训练网络的集合将可能的良性灵长类动物变体与突变率和测序覆盖范围匹配的随机未知变体区分开。该网络集合用于对整组的未知变体进行评分，并通过偏向具有更大致病预测后果的未知变体，影响接种分类器下一次迭代的未知变体的选择，在每次迭代时采取渐进步骤，以防止模型过早收敛到次优结果。

[0323] 常见的灵长类变异也为评估完全独立于以前使用的训练数据(因元分类器的增殖而难以客观评估)的现有方法，提供了一个干净的验证数据集。我们使用10000个固定灵长类常见变体，评估我们模型以及四个其他流行分类算法(Sift、Polyphen2、CADD、M-CAP)的性能。由于约50%的人类错义突变可以按常见等位基因频率通过自然选择去除，因此我们在一组随机挑选的错义变体(按照突变率与10000个固定灵长类常见变体匹配)上计算每个分类器的第50百分位得分，并用该阈值评估灵长类常见的变异。通过使用仅在人类常见变体上训练的深度学习网络，或者使用人类常见变体和灵长类动物变体，我们的深度学习模型的准确度在该独立验证数据集上明显优于其他分类器。

[0324] 最新的trio测序研究已经将神经发育障碍患者及其健康兄弟姐妹的数千个新生突变进行归类，从而能够在分离病例Vs对照例的新生错义突变时，评估各种分类算法的强度。就四种分类算法中的每一种算法而言，我们与对照例相比对病例中的新生错义变体进行评分，并报告两种分布间差异的Wilcoxon秩和检验p值，表明在灵长类动物变体($p \sim 10^{-33}$)上训练的深度学习网络，在该临床情况下的表现远远优于其他分类器($p \sim 10^{-13}$ 至 10^{-19})。根据该群组项下先前报道的超预期新生错义变体约1.3倍富集，以及约20%的错义突

变产生功能丧失效应的先前估计,我们期望有完美的分类器分离p值为10-40的两个类别,表明我们的分类器仍有改进空间。

[0325] 深度学习分类器的准确度与训练数据集的规模成比例,六个灵长类物种中每个物种的变异数据对提升分类器的准确度均有独立贡献。现存非人类灵长类动物物种的大数量和多样性,以及证据表明蛋白改变变体的选择压力在灵长类谱系内大体一致,建议将系统性灵长类种群测序作为一个有效策略,以便对当前限制临床基因组判读的数百万意义不明的人类变体进行分类。在504种已知非人类灵长类物种中,大约60%因狩猎和栖息地丧失而面临灭绝,促使人们迫切地需要在全世界范围内开展保护工作,使这些独特的、不可替代的物种和我们自己都受益。

[0326] 虽然没有那么多的聚集体全基因组数据是作为外显子数据获得的,限制了检测深度内含子区域中自然选择的影响,但我们也能够计算远离外显子区域的隐秘剪接突变的实测Vs预期计数。总之,我们在距外显子-内含子边界>50nt的距离处,观察到隐秘剪接突变中60%损耗。衰减信号可能是与外显子组相比具有全基因组数据的样品量较小和预测深度内含子变体影响的难度更大的组合。

[0327] 术语

[0328] 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。

[0329] 如本文所用,以下术语具有所指示的含义。

[0330] 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)。

[0331] 本申请可互换地使用术语“蛋白质”和“转译序列”。

[0332] 本申请可互换地使用术语“密码子”和“碱基三联体”。

[0333] 本申请可互换地使用术语“氨基酸”和“转译单元”。

[0334] 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”。

[0335] 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。

[0336] 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。

[0337] 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。

[0338] 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。

[0339] 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。

[0340] 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。

[0341] 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。

[0342] 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。

[0343] 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。

[0344] 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。

[0345] 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。

[0346] 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)。

[0347] 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”。

[0348] 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。

[0349] 术语“变体频率”表示种群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种群中的所有的染色体的分数。举例来说,样品变体频率表示在“种群”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“种群”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“种群”对应于来自正常个体种群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。

[0350] 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。

[0351] 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。

[0352] 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。

[0353] 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。

[0354] 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。

[0355] 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。

[0356] 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。

[0357] 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。

[0358] 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。

[0359] 测序过程

[0360] 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。

[0361] 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。

[0362] 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。

[0363] 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。

[0364] 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。

[0365] 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。

[0366] 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2008/0280773 A1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。

[0367] 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。

[0368] 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。

[0369] 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。

[0370] 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Pisces™应用程序,托管在<https://github.com/Illumina/Pisces>上,并在文文章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。

[0371] 这样的变体识别应用可以包括四个顺序执行的模块:

[0372] (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) 缝合共有解读来减少噪音。输出是缝合的 BAM。

[0373] (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。

[0374] (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。

[0375] (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亚种群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。

[0376] 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. 的和在文章 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。

[0377] 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:

[0378] 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 $(k-1)$ 。

[0379] 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种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。

[0380] 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。

[0381] d. VEP缓存文件

[0382] 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%)。

[0383] 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%)。

[0384] 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。

[0385] 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度,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。

[0386] 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。

[0387] 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。

[0388] 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。

[0389] 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。

[0390] 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。

[0391] 变体识别应用

[0392] 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。

[0393] 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。

[0394] 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。

[0395] 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。

[0396] 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。

[0397] 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。

[0398] 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。

[0399] 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。

[0400] 良性训练集生成

[0401] 已对数百万的人类基因组和外显子组进行了测序,但是由于从良性遗传变异中区分出导致疾病的突变有难度,这些基因组和外显子组的临床应用仍有局限。在本文中,我们证明了很大程度上,其他灵长类物种中的常见错义变体在人类临床上是良性的,使得致病突变通过消除法进行系统识别。应用来自六个非人类灵长类物种种群测序的成千上万的常见变体,我们对识别罕见疾病患者中致病突变的深度神经网络进行训练,准确度达88%,并在基因范围意义上发现智力残疾中的14个新的候选基因。把来自其他灵长类物种的常见变异编入目录可提高对不确定重要性的数百万变体的诠释,进一步加强人类基因组测序的临

床效应。

[0402] 由于诠释人类种群中罕见遗传变体及推断它们对疾病风险的影响存在难度,诊断测序的临床可实施性是有局限的。由于罕见遗传变体对健康有害作用,临床上具有显著意义的遗传变体在人类种群中及其罕见,对于绝大多数罕见遗传变体而言,它们对人类健康的影响并不确定。这些大量的、罕见的具有不确定临床意义的变体严重妨碍了运用测序进行个体化医疗和全种群健康筛查。

[0403] 大多数外显性孟德尔疾病在种群中流行性很低,因此观察种群中高频次的变体是利于良性结果的有力证据。评估横跨多样性人类种群的常见变异是将良性变体编入目录的有效策略,但是由于我们的物种近代史上出现瓶颈,在此期间,大部分原始多样性缺失,所以当今人类的常见变异总数是有限的。当今人类种群研究由过去15000-65000年间少于10000个个体的有效种群大小(N_e)显著增加,一小组常见多态性追溯到这种大小种群变异的有限容量。在参考基因组超过7亿个潜在蛋白改变错义取代中,只有大概千分之一大于整个种群等位基因频率的0.1%。

[0404] 人类种群以外,黑猩猩包括接下来最为接近的现存物种,其氨基酸序列一致性达99.4%。人类和黑猩猩的蛋白编码序列的近一致性表明对黑猩猩蛋白编码变体进行纯化选择也可模拟出状态一致性人类突变的健康结果。

[0405] 由于中性多态性在人类祖先系谱($-4N_e$)的平均时间为物种分化时间(6百万年前)的一部分,除非平衡选择维持的罕见单倍体实例外,排除偶然情况,自然发生的黑猩猩变异探索的突变空间基本不重叠。如果状态一致性多态性同样影响这两种物种的健康,黑猩猩种群中高等位基因频率变体的存在可能在人类中表示良性结果,这就扩大了纯净选择建立的具有良性结果的已知变体的目录。

[0406] 结果-其他灵长类中常见变体在人类中很大程度上是良性的

[0407] 最近可用的汇总的外显子组数据包括外显子组聚合联盟(ExAC)和基因组聚合数据库(gnomAD)采集的123136名人类,让我们能测量自然选择对横跨等位基因频率谱的错义突变和同义突变的影响。对三核苷酸上下文的影响进行突变率(如图49A、51、52A、52B、52C及52D)调整后,组群中一次观察得到的罕见单变体与新生突变预测的预期的2.2/1错义/同义突变率紧密吻合,但是由于通过自然选择对有害突变进行纯化,在更高等位基因频率观察到的错义变体数减少。尽管在健康个体中也观察到这些变体,等位基因频率升高的错义/同义突变率逐渐降低符合相当大一部分种群频率小于0.1%的错义变体,这些变体具有轻度毒害后果。除了由于平衡选择和始祖效应引起的少数良好记录的异常外,这些发现支持通过过滤掉等位基因频率大于0.1%到1%的变体诊断实验室进行的普遍经验性实践可能对于外显性遗传疾病是良性的。

[0408] 我们对从24个无关个体的群组中两次或多次取样得到的常见黑猩猩变体进行鉴定。我们估计99.8%的变体在普通黑猩猩种群中常见(等位基因频率(AF)大于0.1%,这就表明这些变体已经通过了纯化选择的筛选。对于状态一致性人类变体(图49B),我们测试了人类等位基因频率,排除了扩展的主要组织相容性复杂区域作为平衡选择的已知区域,同时变体在多序列比对中缺乏一一映射。对于与普通黑猩猩变体状态一致的人类变体,在整个人类等位基因频率谱(通过卡方检验(χ^2)测试, P 大于0.5)中,错义/同义突变率很大程度上是恒定的,这与人类种群中对普通黑猩猩变体阴性选择的损耗及两种物种中错义变体的

一致选择系数是一致的。与普通黑猩猩变体状态一致的人类变体中观察得到的低错义/同义突变率与黑猩猩 (Ne-73000) 中较大有效种群大小一致,使得对轻度有害变异的过滤更为有效。

[0409] 相反,对于单个黑猩猩变体(仅从群组中一次取样)而言,我们观察到在常见等位基因频率 (P 小于 5.8×10^{-6}) 上,错义/同义突变率显著下降,如图49C所示。这就表明24%的单个黑猩猩变体通过在等位基因频率大于0.1%处的人类种群的纯化选择过滤。这种损耗意味着相当一部分黑猩猩单变体为罕见有害突变,这些突变对健康的损害作用妨碍其达到两个物种中常见等位基因频率。我们估计仅69%的单个变体在普通黑猩猩种群中是常见的 (AF 大于 0.1%)。

[0410] 我们接下来识别了六个非人类灵长类物种中至少一个观察到的变异状态一致性人类变体。六个物种的每个的变异通过类人猿基因组项目(黑猩猩、倭黑猩猩、大猩猩和红毛猩猩)进行确定,或提交到来自灵长类基因组项目(恒河猴、狨猴)的单核苷酸多态性数据库 (bdSNP),并基于有限数量的测序单例和每个物种观察的低错义:同义突变率(见补充表1),很大程度上代表了常见变体。与黑猩猩类似,我们发现来自六个非人类灵长类物种的变体的错义/同义突变率在整个人类等位基因频率谱中大概相同,其不同于常见等位基因频率上错义变异的轻度损耗(如图49D、53、54及55,补充数据文件1),这是由于包含少数罕见变体(黑猩猩中16%的变体低于等位基因频率0.1%,且由于少数测序个体,其他物种中更低)。这些结果表明状态一致性错义变体的选择系数与至少与新世界猴相关的灵长类血统一致,估计这些错义变体分化自3500万年前的人类祖先血统。

[0411] 我们发现与观察到的灵长类动物变体状态一致的人类错义变体在ClinVar数据库中大量富集,这具有良性结果。在排除具有不确定意义的变体和注释冲突的变体后,在平均90%的时间内,出现在至少一个非人类灵长类物种中的ClinVar变体注释为良性或可能为良性,与之相比,ClinVar错义变体在35%时间内通常为良性 (P 小于 10^{-40} ,如图49E)。除等位基因频率大于1%的人类变体以用于降低精选偏离外,用于灵长类动物变体的ClinVar注释的致病性稍大于观察到的取样自类似大小群组的健康人类的致病性 (95%为良性结果或可能为良性结果, $p=0.07$)。

[0412] 人类遗传学领域长期依赖于模式生物以推断出人类突变的临床影响,对大多数基因可驯动物模型的长期进化距离引起人们对进化程度的关心,模式生物的发现可归纳到人类。我们延伸了超出灵长类血统的分析,这包括来自四种其他哺乳动物物种(老鼠、猪、山羊及牛)和两种更为遥远的脊椎动物(鸡和斑马鱼)的十分常见变异。由于错义/同义突变率远小于2.2/1,我们选择了dbSNP中基因范围内变异充分确定的物种,并确定这些变体为十分常见的变体。与灵长类分析相比,在常见等位基因频率上(如图50A),在更为遥远的物种中,与变异状态相一致的人类错义突变明显损耗,在更长的进化距离中,损耗度增大(如图50B及补充表2和3所示)。

[0413] 在更遥远的物种中,对人类有毒害性地、可耐受高等位基因频率的错义突变表明状态一致的错义突变的选择系数在人类和更多遥远物种间进行充分分化。尽管如此,由于在常见等位基因频率上自然选择损耗的一部分错义变体通常小于观察到的人类错义变体50%的损耗(如图49A),更为遥远的哺乳动物中错义变体的存在仍会增加良性结果的可能。与这些结果一致,我们发现在老鼠、猪、山羊和牛中观察到的ClinVar错义变体73%的可能

性被注释为良性结果或可能为良性结果,而在灵长类变异中为90% (P 小于 2×10^{-8}) (如图50C),在整个ClinVar数据库中为35%。

[0414] 确定进化距离而非驯化遗物是分化选择系数的主要推动力,跨越大范围的进化距离(如图50D,补充表4和补充数据文档2所示),我们使用了取代种内多态性的近缘物种之间的固定替代物重新进行分析。我们发现,与种内固定替代物状态相一致的人类错义变体的损耗随着进化分支长度而增加,与那些驯化的物种相比,野生物种没有明显区别。与苍蝇和酵母早期工作一致,偶然发现与分化血统相比,状态一致性固定错义替代物的数量低于预期。

[0415] 变体致病性分类的深度学习网络

[0416] 本公开技术提供了变体致病性分类的一种深度学习网络。基于变体分类的临床应用意义,进行了大量的尝试以使用监督的机器学习来解决这一问题,然而,由于缺乏包括用于训练的明确标记为良性和致病性的变体的大小合适的真实数据集,这些努力受到了阻碍。

[0417] 现有人类专家精选的变体的数据库并不代表完整基因组,50%的ClinVar数据库中的变体仅来自于200个基因(人类蛋白编码基因占1%)。进一步,系统学习确定多数人类专家注释的支持证据有疑问,低估了诠释仅在一个患者中观察到的罕见变体的难度。尽管人类专家诠释越来越严格,围绕共识实践,制定了大量分类指导方针,这些分类指导方针对加强现有趋势是有风险的。为降低人类诠释偏差,最近的分类器对常见人类多态性或固定的人类-黑猩猩替代物进行训练,这些分类器同时也作为它们的输入:早期分类器的预测得分,这些早期分类器通过人类精选数据库进行训练。在没有独立的、无偏差的真实数据集的情况下,这些不同方法的性能的客观基准是难懂的。

[0418] 来自六个非人类灵长类(黑猩猩、倭黑猩猩、大猩猩、红毛猩猩、恒河猴和狨猴)的变异提供了超过300000个独特的、与常见人类变异不重叠的错义变体,很大程度上代表了通过纯化选择筛选的良性结果的常见变体,大大扩大了可用于机器学习方法的训练数据集。通常,在排除具有不确定意思的变体和注释冲突的变体后,每个灵长类物种比整个ClinVar数据库(42000,截止2017年十一月)提供更多的变体。另外,这部分内容没有人类诠释偏差。

[0419] 通过使用包括常见人类变体(AF大于0.1%)和灵长类变异(补充表5(如图58所示)的数据集,我们训练了一种新的深度残差网络:PrimateAI,用于输入其他物种(如图2和3所示)中侧翼具有感兴趣的变体的氨基酸序列和直系同源序列比对。与现有使用人类工程特征的分类器不同,我们的深度学习网络学习直接来自原始序列的准确特征。为并入关于蛋白结构的信息,我们训练了独立网络用于单独预测二级结构和序列的溶剂可及性,然后将这些作为子网络列入全模型中(如图5和6)。考虑到少量已成功结晶的人类蛋白,推断原始序列的结构具有避免偏差的优势,这是由于蛋白结构不完整和功能化域注释。包括蛋白结构在内,网络的总深度为36个卷积层,包括大概400000个可训练参数。

[0420] 为了只使用具有良性标记的变体对分类器进行训练,由于观察到的给定突变是否为种群中常见突变,我们拟定了预测问题。多个因素影响在高等位基因频率处观察变体的可能性,其中我们仅关心其毒害性。其他因素包括突变率、技术误差,如测序范围及影响中性遗传漂变的因素,如基因转换。

[0421] 我们将良性训练集中的每个变体与来自ExAC数据库的123136个外显子组的缺失的错义突变进行匹配,对各个混淆的因素进行控制,并对深度学习网络进行训练以区分良性变体和匹配对照例(如图24所示)。由于未标记变体数量远超标记的良性训练数据集的大小,我们同时训练了八个网络,每个网络使用与良性训练数据集相匹配的未标记变体的不同集,以获得共识预测。

[0422] 仅使用主要氨基酸序列作为其输入,深度学习网络将高致病性得分准确分配给有用蛋白功能域残差量,如图20所示用于电压门控钠通道SCN2A:癫痫、自闭症和智力障碍的重要疾病基因。SCN2A的结构包括四种同源重复,每种同源重复包括六个跨膜螺旋(S1-6)。对于膜去极化,带正电荷的S4跨膜螺旋向膜细胞外侧移动,使得S5/S6成孔域通过S4-S5连接器打开。临床上与早期发作癫痫性脑病相关的S4和S4-5连接器和S5域中的突变通过网络进行预测,在基因中具有最高致病性得分,也将健康种群中这些突变进行去除。我们还发现网络识别域内重要的氨基酸位置,并将最高致病性得分分配到这些位置处的突变,如转录因子的包含DNA的残基和酶的催化残基(如图25A、25B、25C和26所示)。

[0423] 为了更好地理解深度学习网络如何深入了解主要序列的蛋白结构和功能,我们将来自网络第一个三层的可训练参数进行可视化。在这三层中,我们观察到网络学习不同氨基酸重量的相关性,不同氨基酸重量接近现有氨基酸距离的测量值如Grantham得分(如图27所示)。这些初始层的输出成为后面层的输入,使得深度神经网络构建渐进的高阶数据显示形式。

[0424] 使用训练保留的10000个常见灵长类动物变体,我们通过现有分类算法对网络性能进行比较。由于通过在常见等位基因频率进行纯化选择对50%的所有最新出现的错义单例进行过滤,(如图49A所示),我们通过一组10000个随机选择的变体确定每个分类器50%得分,这些变体与10000个常见灵长类动物变体通过突变率和测序范围进行匹配,并在此阈值处评估每个分类器的准确度(如图21D,28A和补充数据文档4所示)。在分配良性结果到10000个保留常见灵长类动物变体上,我们的学习网络(准确率达91%)性能强于其他分类器(下面最佳模型的准确率为80%)。

[0425] 仅与人类变异数据(如图21D所示)训练网络的准确度相比,几乎一半对现有方法的改进源自对深度学习网络的使用,一半源自对灵长类变异训练数据集的扩增。为了在临床场景中测试具有不确定意义的变体分类,对比健康对照例,我们评估了深度学习网络在区分神经发育障碍患者的新生突变的能力。在流行性方面,神经发育障碍是罕见遗传疾病最大类别之一,最近三重测序研究涉及新生错义及蛋白截断突变的核心作用。

[0426] 我们对来自破译发育障碍(DDD)群组的受影响的4293个个体的每一个确定识别的新生错义变体进行分类,与来自孤独症儿童样本数据库(SSC)群组中2517个未受影响的氏族成员的新生错义变体进行对比,并评估了进行Wilcoxon秩和检验的两种分配在预测得分中的差别(如图21E、29A及29B所示)。在这个任务中,深度训练网络明显胜过其他分类器(P小于 10^{-28} 如图21F和28B所示)。不同分类器在保留灵长类动物变体数据集的性能与DDD案例对比对照例案例数据集相关(相关系数 ρ 等于0.57,P小于0.01),表明尽管使用完全不同来源和方法(如图30A所示),评估致病性的两个数据集一致性良好。

[0427] 接下来,我们旨在评估深度学习网络在同一基因中对良性和致病性突变进行分类的准确度。假定DDD群组主要包括无受影响一级亲属的受影响小孩的索引病例,利于具有新

生主导遗传模式的基因的致病性,分类器的准确度并未夸大,这是很重要的。我们限制了对DDD学习中疾病相关性名义上很重要的605个基因的分析,仅从蛋白截断变异(P小于0.05)进行计算。在这些基因中,相较于预期,新生错义突变富集3/1,(如图22A所示),表明67%的突变是致病的。

[0428] 在同一基因集(如图22B所示,(P小于10⁻¹⁵),深度学习网络能区分致病新生变体和良性新生变体,这很大程度上强于其他方法(如图22C和28C所示)。在大于0.803的二进制临界值处(如图22D和30B所示),案例中65%的新生错义突变被深度学习网络归类为致病的,对照例中这一数据为14%,这与分类准确度88%一致(如图22E和30C所示)。考虑到神经发育障碍的不完全外显率频率和表现度不一致性,由于对照例中包括部分外显致病变体,这一数字可能低估了我们的分类器的准确度。

[0429] 新型候选基因发现

[0430] 应用大于等于0.803的阈值将致病错义突变进行分层增加了对1.5-2.2倍DDD患者中新生错义突变的富集,这与蛋白截断突变(2.5倍)相近,放弃了富集超出预期的低于总数三分之一的变体。这从实质上提高了统计功效,使得智力障碍中14个额外候选基因得以发现。此前,统计功效在原始DDD学习(表1)中并未达到基因范围意义阈值。

[0431] 与人类专家精选进行比较

[0432] 我们测试了各种分类器对ClinVar数据库最近人类专家精选的变体的性能,发现分类器对ClinVar数据库的性能与保留灵长类动物变体数据集和DDD病例对比对照例数据集(P分别为0.12和0.34)无明显相关性(如图31A和31B所示)。我们假定现有分类器对人类专家精选有偏差,而这些人类试探法往往方向是正确的,但不是最佳的。在一个实例中,ClinVar中致病变体和良性变体的Grantham平均差是605个疾病相关基因中DDD病例对比对照例中差值的两倍(如图2所示)。相比之下,人类专家精选似乎未充分利用蛋白结构,尤其是未充分利用暴露于表面的残基的重要性,在表面上,残基能与其他分子发生相互作用。我们观察到ClinVar致病突变和DDD新生突变均与预测的溶剂暴露残基相关,但是良性和致病ClinVar变体之间溶剂可及性的差值仅为DDD病例对比对照例的一半。这些发现表明确认偏差对人类专家更为直接地解释如Grantham和保守性的因素是有益的。人类精选数据库训练的机器学习分类器有望加强这些趋势。

[0433] 结果表明系统灵长类种群测序是对当前限制临床基因组解释、具有不确定意义的数百万人类变体进行分类的有效策略。深度学习网络对保留的常见灵长类动物变体和临床变体的准确度随着用于训练网络的良性变体的数量增加而增加(如图23A所示)。此外,单独对六个非人类灵长类物种中每个的变体进行训练有助于提高网络的性能,而训练更为遥远的哺乳动物的变体对网络性能造成负面影响(如图23B和23C所示)。这些结果支持这种论断,即对于外显性孟德尔疾病,常见灵长类动物变体在人类中多为良性,而在更为遥远的物种变异中,情况并非如此。

[0434] 尽管相对于已测序的人类基因组和外显子组数量,研究检测的非人类灵长类基因组的数量较少,值得注意的是,这些额外的灵长类提供数量不均衡的常见良性变异相关信息。ExAC模拟表明仅几百个个体(如图56所示)后,常见人类变体(小于0.1%等位基因频率)快速停滞,对数百万健康种群进行测序主要提供额外罕见变体。与常见变体不同(已知这些常见变体基于等位基因频率很大程度上在临床上良性的),健康种群中罕见变体可引起

具有不完全外显率的隐性遗传疾病或显性遗传疾病。由于每个灵长类物种携带一组不同的常见变体,对每个物种的数十个灵长类进行测序是系统编目灵长类血统中良性错义变异的有效策略。实际上,本研究中检测的来自六个非人类灵长类物种的134个个体提供的常见错义变体是ExAC研究中123136个个体的近4倍(补充表5(如图58所示)。即使野生动物保护区和动物园中无关个体数量相对较少,涉及数百个体的灵长类物种测序研究同样适用。将对野生种群的干扰降到最低,从对非人类灵长类的保护和善待立场出发,是很重要的。

[0435] 现有类人群携带的遗传多样性远低于非人类灵长类物种,是大约每个黑猩猩、大猩猩和长臂猿携带的单核苷酸变体数量的一半,红毛猩猩携带的变体的三分之一。虽然不知道大多数非人类灵长类物种的遗传多样性水平,根据现有的大量非人类灵长类物种,我们推断大多数可能良性人类错义位置可被至少一种灵长类物种的常见变体所覆盖,使得致病变体通过排除法系统识别(如图23D所示)。即便仅对这些物种的小集进行测序,加大训练数据大小可更为准确地通过机器学习预测错义结果。最后,我们的发现侧重于错义变异,这种策略也可应用到推断非编码变异的结果,尤其是保留的调控区,在此对人类基因组和灵长类基因组进行充分比对以便明确地确定变体是否状态一致。

[0436] 504个已知非人类灵长类物种中,由于非法猎取和广泛栖息地的丧失,大约60%的物种濒临灭绝。种群规模的减小和潜在的物种灭绝对遗传多样性带来无法弥补的损失,迫切鼓励全世界范围内对物种保护作出努力,这对稀罕、不可替代物种及人类本身是有益的。

[0437] 数据生成及比对

[0438] 本公开坐标涉及人类基因组构建UCSC hg19/GRCh37,包括映射到使用多个序列比对的hg19的其他物种变体的坐标。蛋白编码DNA序列的规范转录本及99个脊椎动物基因组和枝长的多个序列对比从UCSC基因组浏览器下载。

[0439] 我们从外显子集合联盟(ExAC)/基因组集合数据库(gnomAD外显子组)v2.0获得人类外显子组多态性数据。我们从类人猿基因组测序项目中获得灵长类变异数据,其包括24只黑猩猩、13只倭黑猩猩、27只大猩猩和10只红毛猩猩的全部基因组测序数据及基因型。我们同样将来自黑猩猩和倭黑猩猩单独研究的35只黑猩猩的变异包括在内,但是由于变体呼叫方法不同,我们排除了来自种群分析的变异并仅使用这些变异对深度学习模型进行训练。另外,使用16只恒河猴个体和9个狨猴个体对这些物种原始基因组项目变异进行试验,个体水平信息不可用。我们获得来自dbSNP的恒河猴、狨猴、猪、牛、山羊、老鼠、鸡和斑马鱼的变异数据。由于种群分析用个体基因型信息不可用,dbSNP的还包括额外红毛猩猩变体,这些额外红毛猩猩变体仅用于训练深度学习模型。为避免平衡选择效应,我们也排除了用于种群分析的扩展的主要组织相容性复杂区域(chr6:28,477,797-33,448,354)的变体。

[0440] 我们使用99个脊椎动物多个物种对比确定与人类蛋白编码区域的直系同源一一映射并防止对假基因的映射。如发生在参考/可变方向,变体状态一致。为了确保变体在人类和其他物种中均具有相同预测蛋白编码结果,对于错义变体和同义突变变体而言,需要物种间密码子中其他两个核苷酸一致。补充数据文档中列出了分析包括的每个物种的多态性,补充表1表示详细计量。

[0441] 对于四个等位基因频率类别的每个(如图49A所示),我们使用基因内区域的变异评估96个可能三核苷酸上下文的每个中同义突变变体和错义变体的期望数并更正突变率

(如图51和补充表7和8(如图9)所示)。我们也分别分析了状态一致CpG二核苷酸和非CpG二核苷酸变体,证实两个类别等位基因频谱的错义/同义突变率是平直的,这就表明我们的分析适用于CpG和非CpG变体,尽管它们的突变率大不相同(如图52A、52B、52C和52D所示)。

[0442] 其他物种中具有多态性的状态一致性人类错义变体的损耗

[0443] 为评价其他物种中变体在人类常见等位基因频率(大于0.1%)上是否耐受,我们识别了其他物种中具有变异性的状态一致性人类变体。对于每个变体,我们基于它们在人类种群(单例、大于单例~0.01%单例,0.01-0.1%和大于0.1%单例)等位基因频率将其分派到四个分类中的其中一个,并评估了罕见(小于0.1%)和常见(大于0.1%)变体之间错义/同义突变率(MSR)的降低。在常见等位基因频率(大于0.1%),状态一致性变体的损耗表明来自具有相当毒害性的其他物种的这部分变体通过人类常见等位基因频率自然选择进行筛选。

$$[0444] \quad \% \text{depletion} = \frac{\text{MSR}_{\text{rare}} - \text{MSR}_{\text{comm}}}{\text{MSR}_{\text{rare}}}$$

[0445] 计算每个物种错义/同义突变率和损耗率,如图50B和补充表2所示。此外,低于常见黑猩猩变体(如图49B)、单个黑猩猩变体(如图49C)及哺乳动物变体(如图50A),我们对2*2列联表进行同源性卡方检验(x2)以测试罕见变体和常见变体之间错义/同义突变率差异是否显著。

[0446] 由于仅对来自类人猿基因组项目的有限量个体进行测序,我们使用来自ExAC的人类等位基因频谱对该部分取样变体进行评估,该部分取样变体在一般黑猩猩种群中为罕见(小于0.1%)变体或常见(大于0.1%)变体。我们基于ExAC等位基因频率对一组24个个体进行取样并识别了该组中一次观察到的或多次观察到的错义变体。99.8%的多次观察到的变体可能为一般种群中常见(大于0.1%)变体,而69%的一次观察到的变体可能为一般种群中常见变体。为了验证在更为遥远的哺乳动物中的错义变体的中观察到的损耗不是由更好保存的基因的混杂作用引起,我们重复了上述分析,与人类相比,在11个灵长类和50个哺乳动物的多个序列比对中,仅对大于50%平均核苷酸相同度的基因进行限定。

[0447] 去除了分析中7%的人类蛋白编码基因,不会对结果造成实质性影响。其次,为了确定结果不受变体呼叫事件或驯化遗物(因为大多数选自dbSNP的物种是驯化物种)的影响,我们使用来自密切相关物种对的固定替代物代替种内多态性进行重复分析(如图50D、补充表4及补充数据文档2)。

[0448] 人类、灵长类、哺乳动物及其他脊椎动物多态性ClinVar分析

[0449] 为了测试与其他物种状态相一致的变体的临床影响,我们下载了ClinVar数据库,不包括致病性注释冲突的变体或仅标识为具有不确定意义的变体。下面,补充表9显示了过滤步骤,致病类别中总计24853个错义变体,良性类别中总计17775个错义变体。

[0450] 我们对与人类、非人类灵长类、哺乳动物和其他脊椎动物变异状态相一致的致病和良性ClinVar变体进行计数。对于人类而言,我们模拟了从ExAC等位基因频率取样的一组30个人。补充表10显示了每个物种的良性和致病性变体的数量。

[0451] 用于模型训练的良性和未标记变体的生成

[0452] 我们从人类和非人类灵长类构建了大量常见良性错义变体的良性训练数据集用于机器学习。数据集包括常见人类变体(大于0.1%等位基因频率,83546个变体)和黑猩猩、

倭黑猩猩、大猩猩、红毛猩猩、恒河猴及狨猴的变体(301690个唯一灵长类动物变体)。补充表5显示了每个来源提供的良性训练变体的数量。

[0453] 我们对深度学习网络进行训练以区分与三核苷酸上下文控制、测序范围及物种和人类间可对比性匹配的一组标记的良性变体和一组为标记的变体。为了获得未标记训练数据集,我们从规范化编码区域中所有可能的错义变体开始。我们排除了来自ExAC的123136个外显子组中观察到的变体和起始密码子和终止密码子中的变体。总共生成68258623个未标记错义变体。当选择用于灵长类动物变体的匹配的未标记变体时,对变体进行过滤以修正测序范围较差的区域以及人类和灵长类基因组非一一比对的区域。

[0454] 通过训练使用同一组标记的良性变体和八组随机取样的未标记的变体的八个模型以及得到其预测的平均数,我们获得了共识预测。我们还选出随机取样的一系列10000个灵长类动物变体用于验证和测试,这些变体通过训练保留(如补充表数据文档3所示)。对于每个数据集而言,我们对三核苷酸上下文匹配的10000个非标记变体进行取样,在比较不同分类算法(如补充数据文档4所示)时,我们使用这些变体将每个分类器的阈值归一化。在其他实施方式中,在集合中可使用较少模型或附加模型,2个到500个不等。

[0455] 我们评估了深度学习网络两个版本的分类准确度,一个版本仅通过常见人类变体进行训练,另一个则通过完整良性标记数据集进行训练,数据集包括常见人类变体和灵长类动物变体。

[0456] 深度学习网络架构

[0457] 对于每个变体而言,致病性预测网络将以感兴趣变体为中心的51-长度氨基酸序列作为输入,并在中心位置代入具有错义变体的二级结构和溶剂可及性网络(如图2和3所示)的输出。三个51-长度位置频率矩阵由99个脊椎动物的多重序列比对生成,一个用于11个灵长类,一个用于50个不包括灵长类的哺乳动物及一个用于不包括灵长类和哺乳动物的38个脊椎动物。

[0458] 二级结构深度学习网络预测每个氨基酸位置的三态二级结构: α 螺旋(H)、 β 折叠(B)及卷曲(C)(如补充表11所示)。溶剂可及性网络描述了每个氨基酸位置的三态溶剂可及性:埋藏的(B)、中间的(I)及暴露的(E)(如补充表12所示)。网络均将侧翼氨基酸序列作为其输入,并使用来自蛋白数据库的已知无冗余晶体结构标签对网络进行训练(如补充表13所示)。对于预训练的三态二级结构和三态溶剂可及性网络的输入而言,我们使用了所有99个脊椎动物的多重序列比对生成的一个单长度位置频率矩阵,长度为51,深度为20。在对网络的蛋白数据库中已知晶体结构进行预训练后,去除二级结构和溶剂模型的最后两层,并将网络输出直接与致病性模型输入相连。实现三状态二级结构预测模型的最佳测试准确度为79.86%(补充表14)。与仅使用预测结构标记相比,使用DSSP-注释的用于近4000个具有晶体结构的人类蛋白(蛋白定义二级结构)结构标记来比较神经网络预测,无实质性差别(如补充表15所示)。

[0459] 用于致病性预测(primatAI)的深度学习网络和用于预测二级结果和溶剂可及性的深度学习网络均采用残差块架构。图3和补充表16(图4A、4B和4C)描述了primatAI的详细架构。图6和补充表11、12(图7A、7B)(图8A、8B)描述了预测二级结构和溶剂可及性网络的详细架构。

[0460] 分类器性能在保留测试组10000个灵长类动物变体的基准

[0461] 我们使用测试数据组中10000个保留灵长类动物变体来测试深度学习网络的基准,同样也对先前公布的其他20个分类器进行基准测试,我们从数据库dbNSFP中获得用于该变体的预测得分。图28A同样表示每个分类器在10000个保留灵长类动物变体测试集中的性能。由于不同分类器具有大不相同的得分分布,我们使用与通过三核苷酸上下文得到的测试集相匹配的10000个随机选择的未标记变体来识别每个分类器的50%的阈值。我们在10000个保留灵长类动物变体测试集中部分变体对每个分类器进行基准化,对于该分类器,这些变体在50%阈值处归类为良性变体,确保各方法之间公平比较。

[0462] 对于每个分类器而言,图28A和补充表17(如图34所示)表示使用50%的阈值的预测为良性的部分保留灵长类测试变体。还表明,变体位置处比对物种数目相关的primateAI性能强大,只要有充分的哺乳动物保守信息,primateAI通常性能良好,这对于大多数蛋白编码序列是正确的(如图57)。

[0463] DDD研究新生变体的分析

[0464] 我们从DDD研究中获得发布的新生变体并获得从SSC自闭症研究中健康成员对照例中的新生变体。DDD研究为新生变体提供置信水平,由于变体呼叫误差,我们排除了阈值小于0.1的DDD数据集变体作为潜在假阳性。在一个实施方式中,总计有来自DDD受影响个体的3512个错义新生变体和来自健康对照例的1208个错义新生变体。用于99-脊椎动物多重序列比对的、UCSC使用的规范转录本注释略不同于DDD使用的转录本注释,使得错义变体的总数略有不同。我们评估了这些分类方法在区分DDD受影响个体中新生错义变体对比来自自闭症研究的未受影响成员对照例中新生错义变体的能力。对于每个分类器而言,我们报告了来自两种分配的预测得分之间差值的Wilcoxon秩和测试的P值(如补充表17(如图34)所示)。

[0465] 为了测量各种分类器在区分相同疾病基因中良性变异和致病性变异的准确度,我们重复分析了一个子集605个基因,对这些基因进行富集用于DDD组中新生蛋白截断变异($P < 0.05$,泊松精密试验)(补充表18所示)。在这605个基因中,基于新生错义突变的3/1富集超出预期,我们估计DDD数据集中三分之二的新生变体为致病性,三分之一为良性。我们假设了最小不完全外显率,并假设健康对照例中新生错义突变为良性。对于每个分类器而言,我们识别产生相同数量良性或致病性预测的阈值作为这些数据集中观察到的经验性比例,并使用该阈值作为二进制临界值来评估每个分类器在区分病例与对照例对比的新生突变的准确度。为了构建受试者工作特征曲线,我们把新生DDD变体的致病分类作为真实阳性呼叫,健康对照例中致病性新生变体分类作为假阳性呼叫。由于DDD数据集包括三分之一的良性新生变体,用于理论完美分类器的曲线(AUC)下方的区域小于1。因此,将良性和致病性变体进行完美分类的分类器可将DDD患者中67%的新生变体分为真阳性,DDD患者中33%的新生变体分为假阴性,对照例中100%的新生变体为真阴性,得到最大可能AUC为0.837(如图29A和29B及补充表19如图35所示)。

[0466] 新型候选基因发现

[0467] 通过对比观察到的新生突变的数量和无效突变模型下预期数量,我们测试了基因中新生变异的富集。我们重复了DDD研究中的富集分析,当仅计算primateAI得分大于0.803的新生错义突变时,我们报告了在全基因组中具有新意义的基因。我们通过部分错义变体调整了新生破坏性错义变异的全基因组的预期,这些错义变体满足primateAI阈值大于

0.803 (大约为全基因组内所有可能错义突变的五分之一)。根据DDD研究,每个基因需要四次测试,一个测试蛋白截断富集,一个测试蛋白改变新生突变富集,这两种测试仅用于DDD组和神经发育三重测序组的较大规模元分析。通过费雪方法,将蛋白改变新生突变富集与编码序列中错义新生突变的聚类测试相结合。(补充表20,21)。每个基因的P值取自四次测试的最小值,由于 $P < 6.757 \times 10^{-7}$,对全基因组内意义进行确定。 $(\alpha = 0.05)$,四次测试使用18500个基因)。

[0468] ClinVar分类准确度

[0469] 由于大多数现有分类器通过ClinVar内容直接或间接训练,如使用过ClinVar训练的分类器的预测得分,我们对ClinVar数据集分析进行限制,仅使用2017年以后增加的ClinVar变体。由于最近ClinVar变体和其他数据库大量重叠,我们过滤去除了ExAC中位于常见等位基因频率(大于0.1%)的变体,或HGMD(人类基因突变数据库)、LOVD(莱顿开放变异数据库)或Uniprot(通用蛋白质资源)的变体。在排除了注释为不确定意义的变体和注释冲突的变体后,留下了177例注释为良性的错义变体和969例注释为致病性的错义变体。我们使用深度学习网络和其他分类方法对ClinVar变体进行评分。对于每个分类器而言,我们识别产生相同数量良性或致病性预测的阈值作为这些数据集中观察到的经验性比例,并使用该阈值作为二进制临界值来评估每个分类器的准确度(如图31A和31B所示)。

[0470] 增加训练数据量和使用训练数据不同来源的影响

[0471] 为了评估训练数据量对深度学习网络性能的影响,我们随机对来自标记的良性训练集的385236个灵长类动物变体和常见人类变体的变体子集进行取样,并保持基本深度学习网络架构相同。为了表明每个个体灵长类物种的变体对分类准确度有益,而每个个体哺乳动物类物种的变体降低了分类准确度,根据一个实施方式,我们使用训练数据集对深度学习网络进行训练,训练数据集包括83546个人类变体外加一定数量的每个物种随机选择的变体,保持底层网络架构相同。增加到训练集的变体常数(23380)为具有最少错义变体的物种,如倭黑猩猩,中可用变体的总数。我们重复训练过程5次以获得每个分类器的中值性能。

[0472] 具有数量增加的测序灵长类种群的所有可能人类错义突变的饱和度

[0473] 通过基于ExAC中观察到的人类常见错义变体(等位基因频率大于0.1%)对变体进行模拟,我们通过504个现有灵长类物种的常见变体对所有7千万个可能人类错义突变的预测饱和度进行了调查。对于每个灵长类物种而言,我们对人类观察到的常见错义变体进行了四次模拟(83500个错义变体的等位基因频率大于0.1%),这是因为人类每个个体变体的数量约为其他灵长类物种的一半,并在等位基因频率大于0.1%处,通过纯化选择对50%的人类错义变体进行滤除(如图49A)。

[0474] 为了模拟考察的越来越多的人类组中发现的人类常见错义变体(等位基因频率大于0.1%)的部分(如图56所示),我们根据ExAC等位基因频率对基因表型进行取样,并对这些模拟组中至少观察到一次的部分常见变体进行报告。

[0475] 在一个实施方式中,对于primateAI得分的实际应用,对比对照例(如图21D),基于对新生变体案例的富集,在具有显性遗传模式的变体中,优选阈值大于0.8可能为致病性类别,小于0.6可能为良性类别,0.6-0.8之间为中间类别,而在具有隐性遗传模式的基因中,优选阈值大于0.7可能为致病性类别,小于0.5可能为良性类别。

[0476] 图2表示用于致病性预测的深度残差网络的一个实例架构,在本文中是指primateAI。图2中,1D是指一维卷积层。预测的致病性的范围从0(良性)到1(致病性)。网络将以变体为中心的人类氨基酸(AA)参考序列和可变序列(51AAs)、99个脊椎动物物种计算得到的位置权重矩阵(PWM)保守谱及二级结构和溶剂可及性预测深度学习网络作为输入,用以预测三态蛋白二级结构(H代表螺旋、B代表beta折叠、C代表卷曲)及三态溶剂可及性(B代表埋藏的、I代表中间的、E代表暴露的)。

[0477] 图3为primateAI的示意图,该primateAI为致病性分类用深度学习网络架构。模型输入包括参考序列和变体取代的序列的侧翼序列的51个氨基酸(AA),来自灵长类、哺乳动物及脊椎动物比对的三个51-AA-长度位置权重矩阵表示的保守性,模型输入还包括预训练二级结构网络及溶剂可及性网络的输入(长度也为51AA)。

[0478] 图4A、4B及4C为补充表16,表示致病性预测深度学习模型primateAI的实例模型架构详情。形状指定模型每层的输出张量的形状,活化为给予改层神经元的活化。模型输入是指变体周围侧翼氨基酸序列的位置特异性频率矩阵(51AA长,20深)、单热编码人类参考序列和可变序列(51AA长,20深)及二级结构和溶剂可及性模型(51AA长,40深)的输入。

[0479] 所述实例使用一维卷积。在其他实施方式中,模型可使用不同类型卷积,如2D卷积、3D卷积、扩张或空洞卷积、转置卷积、分离卷积及深度方向卷积。相对于饱和非线性如S-形曲线或双曲正切,某些层也使用大大加快随机梯度下降收敛性的ReLU活化功能。本公开技术所用的激活功能的其他实例包括参数ReLU、泄露ReLU及指数线性单元(ELU)。

[0480] 某些层也使用批归一化(Ioffe及Szegedy2015)。就批归一化而言,训练期间,改变卷积神经网络(CNN)中每层的分布,从一层到另一层,分布不同。降低了优化算法的收敛速度。批归一化为解决这一问题的技术。使用x对批归一化层的输入进行注释并使用z对其输出进行注释,批归一化应用以下x变形:

$$[0481] \quad z = \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} \gamma + \beta$$

[0482] 批归一化使用 μ 及 σ 对输入进行均值方差规整并使用 γ 和 β 独一输入进行线性缩放和位移。使用指数移动平均线方法对用于训练集当前层的归一化参数 μ 和 σ 进行计算。换言之,这些参数为不可训练参数。相反,gamma和beta为可训练参数。训练期间, μ 值和 σ 值用于推论期间的正推法。

[0483] 图5和6阐明了用于预测蛋白的二级结构和溶剂可及性地深度训练网络。模型输入是使用由RaptorX软件(用于训练蛋白数据库序列)产生的保守性的位置加权矩阵或99个脊椎动物比对(用于训练和干扰人类蛋白序列)。第二层到最后一层的输出,长度为51AAs,为用于致病性分类的深度学习网络的输入。

[0484] 图7A和7B为补充表11,表示用于3-态二级结构预测深度训练(DL)模型的实例模型架构详情。形状指定模型每层的输出张量的形状,活化为给予改层神经元的活化。模型输入为用于变体周围侧翼氨基酸序列的位置特异性频率矩阵(55AA长,20深)。

[0485] 图8A和8B为补充表12,表示用于3-态溶剂可及性预测深度学习模型的实例模型架构详情。形状指定模型每层的输出张量的形状,活化为给予改层神经元的活化。模型输入为用于变体周围侧翼氨基酸序列的位置特异性频率矩阵(51AA长,20深)。

[0486] 图20表示SCN2A基因中每个氨基酸位置的预测致病性得分,对关键功能域进行注

释。沿着基因绘制了每个氨基酸位置错义替代物的瓶颈primateAI得分。

[0487] 图21D表示对分类器预测训练保留10000个常见灵长类动物变体的测试集的结果进行比较。在将每个分类器的阈值规范化为在匹配突变率的一组100000个随机变体上50%的得分后,y轴表示正确分类为良性的灵长类动物变体的百分比。

[0488] 图21E阐明了对比未受影响成员、具有相应Wilcoxon秩和值P的破译发育障碍(DDD)患者中新生错义变体的primateAI预测得分的分布情况。

[0489] 图21F描述了与对照例相比分离DDD病例中新生错义变体时分类器的比较。示出了每个分类器的Wilcoxon秩和检验P值。

[0490] 图22A、22B、22C、22D及22E阐明了在605个DDD基因中分类器的准确度,P小于0.05。图22A表示来自具有605个相关基因的DDD组的受影响个体中超出预期的新生错义突变的富集,605个相关基因对于新生蛋白截断变异相当重要(P小于0.05)。图22B描述了具有605个相关基因的DDD患者对比未受影响成员中新生错义变体的primateAI预测得分的分布情况,具有相应的Wilcoxon秩和P值。

[0491] 图22C表示对比较分类器在605个基因中DDD个体对比对照例分离新生错义变体的描述。y轴表示每个分类器的Wilcoxon秩和测试值P。

[0492] 图22D描述了对受试者工作特征曲线所示各种分类器进行比较,每个分类器由AUC表示。

[0493] 图22E阐明了每个分类器的分类准确度性及AUC。使用阈值,基于图22A所示富集,分类器可对相同数量致病性和良性变体进行预期性预测,所示分类准确度为真阳性率和真阴性率的平均值。考虑到33%的DDD新生错义变体代表背景,用虚线表示完美分类器用最大可获得AUC。

[0494] 图23A、23B、23C及23D表示用于训练分类器准确度的数据的影响。数量增加的灵长类和人类常见变体训练的深度学习网络取决于完整数据集(总计385236个变体)。如图23A,每个网络的分类性能是基于DDD病例对比对照例中10000个保留灵长类动物变体和新生变体的准确度。

[0495] 根据一个实施方式,图23B和23C表示使用数据集训练的网络的性能,该数据集包括83,546个人类常见变体及来自单个灵长类或哺乳动物物种的22380个变体。基于10000个保留灵长类动物变体(如图23B所示)及DDD病例对比对照例新生错义变体(如图23C所示),对用常见变异的不同来源训练的每个网络进行结果显示。

[0496] 图23D描述了504个现存灵长类物种中状态一致性常见变体(大于0.1%)的所有可能人类良性错义位置的预期饱和度。y轴表示至少一个物种中观察到的部分人类错义变体,绿色表示CpG错义变体,蓝色表示所有错义变体。为了模拟每个灵长类物种的常见变体,我们通过取代对所有可能单核苷酸替代物集进行取样,这与ExAC中常见人类变体(大于0.1%等位基因频率)观察的三核苷酸上下文分配相匹配。

[0497] 图24阐明了对测序范围在确定常见灵长类动物变体的效果进行校正。观察非人类灵长类物种给定变体的可能性与ExAC/gnomAD外显子组数据集中该位置的测序深度呈负相关。相反,较小gnomAD读数深度并不影响在该位置(大于0.1%等位基因频率)观察常见人类变体的可能性,这是因为大量测序的人类外显子组保证大多数常见变异得以确定。当选取匹配变体用于训练网络的每个灵长类动物变体时,对测序深度的影响调整选取变体的可能

性,对三核苷酸上下文进行匹配以控制突变率和基因转换。

[0498] 图25A、25B、25C及26描述了本公开神经网络识别的蛋白基序。图25A、25B及25C阐明了对蛋白域的神经网络进行识别,对在三个不同蛋白域中每个氨基酸位置的变体,显示评价primateAI得分。图25A对重复GXX基序中具有甘氨酸的COL1A2胶原蛋白股进行了突出显示。胶原蛋白基因中临床识别的突变主要是由于GXX重复中甘氨酸的错义突变,这是因为这些突变干扰胶原蛋白的正常汇集并具有很强的显性抑制作用。图25B中,IDS硫酸酯酶的活性位点突出显示,包括转译后修饰到甲酰甘氨酸的活性位点处的半胱氨酸。在图25C中,显示了MYC转录本因子的bHLHzip域。基域通过带正电的精氨酸和赖氨酸残基(突出显示)与DNA接触,这些残基与带负电的糖-磷酸骨架相互作用。亮氨酸拉链域包括相隔7个氨基酸(突出显示)的亮氨酸残基,这些残基对于二聚作用至关重要。

[0499] 图26为一线图,表示在变体预测深度学习得分上对变体中及变体周围每个位置的干扰作用。我们系统化的将变体周围临近氨基酸(位置-25到+25)输入归零,并测算变体在神经网络预测的致病性中的变化。该图表示5000个随机选择变体在每个临近氨基酸位置处干扰的预测致病性得分的平均变化。

[0500] 图27阐述了权重模拟BLOSUM62及Grantham得分模型的关联模型。二级结构深度学习网络的头三层权重相关模型表明与BLOSUM62和Grantham得分矩阵相似的氨基酸之间的相关性。左热图表示第一卷积层参数权重的相关性,第一卷积层紧接着位于单热表示编码的氨基酸之间的二级结构深度学习网络的两个初始上采样层。中间热图表示氨基酸对之间的BLOSUM62得分。右边热图表示氨基酸之间的Grantham距离。深度学习权重和BLOSUM62得分之间的皮尔逊相关性为0.63 ($P=3.55 \times 10^{-9}$)。深度学习权重和Grantham得分之间的相关性为-0.59 ($P=4.36 \times 10^{-8}$)。BLOSUM62和Grantham得分之间的相关性为-0.72 ($P=8.09 \times 10^{-13}$)。

[0501] 图28A、28B及28C表示对深度学习网络primateAI及其他分类器的性能评价。图28A描述了深度学习网络primateAI在预测训练保留的、相对于其他分类器的10000个灵长类动物变体的测试集的良好结果的准确度,其他分类器包括SIFT、PolyPhen-2、CADD、REVEL、M-CAP、LRT、MutationTaster、MutationAssessor、FATHMM、PROVEAN、VEST3、MetaSVM、MetaLR、MutPred、DANN、FATHMM-MKL_coding、Eigen、GenoCanyon、integrated_fitCons及GERP。通过使用与灵长类动物变体匹配的一组10000个随机选择变体对三核苷酸上下文进行突变率和基因转换控制,基于将每个分类器的阈值规范化为其50%得分,y轴表示分类为良性的灵长类动物变体的百分比。

[0502] 图28B描述了对primateAI网络在分离DDD病例对比对照例新生错义变体中的性能进行对比,同时在上方列出了20种现有方法。y轴表示每个分类器的Wilcoxon秩和测试的P值。

[0503] 图28C对primateAI网络在分离605个疾病相关基因中DDD病例对比未受影响对照例新生错义变体中的性能进行对比,同时列出了20种现有方法。y轴表示每个分类器的Wilcoxon秩和测试的P值。

[0504] 图29A及29B阐明了四个分类器的预测得分分布。显示了DDD病例对比未受影响对照例中新生错义变体的四个分类器:包括SIFT、PolyPhen-2、CADD及REVEL预测得分的直方图,具有相应的Wilcoxon秩和P值。

[0505] 图30A、30B及30C对primateAI网络和其他分类器在分离605个疾病相关基因中致病性变体和良性变体的准确度进行对比。图30A中散点图表示每个分类器在DDD病例对比对照例(y轴)上的性能和在保留灵长类数据集(x轴)上的良性预测准确度。图30B比较了分离605个基因中病例对比对照例新生错义变体的不同分类器,由受试者工作特征(ROC)曲线表示,曲线下区域(AUC)表示每个分类器。图30C表示primateAI网络的分类准确度和AUC,图28A\28B\28C列举了20个分类器。使用阈值,基于图22A所示富集,分类器可对相同数量致病性和良性变体进行预期性预测,所示分类准确度为真阳性率和真阴性率的平均值。假定DDD病例中新上错义变体67%为致病性变体,33%为菱形变体而对照例中新上错义变体100%为良性,虚线表示完美分类器的最大可得AUC。

[0506] 图31A及31B阐述了分类器对人类专家精选的ClinVar变体的性能和分类器对经验数据集性能的相关性。散点图31A显示了20个其他分类器中每个分类器在10000个保留灵长类动物变体(x轴)ClinVar变体上的分类准确度及仅由人类或人类加灵长类数据训练的primateAI网络。表示了斯皮尔曼相关系数rho及相关值P。为了对未用于训练分类器的数据的评估进行限制,我们仅使用2017年一月和十月之间增加的ClinVar变体并排除ExAC/gnomAD(大于0.1%等位基因频率)的常见人类变体。使用阈值,分类器可对相同数量的ClinVar观察到的致病性和良性变体进行预期性预测,ClinVar分类准确度为真阳性率和真阴性率的平均值。

[0507] 散点图31B表示ClinVar变体的分类器准备下(y轴)和20个分类器中每个分类器及使用仅人类或人类加灵长类数据训练的primateAI网络的DDD病例对比对照例完整数据集(x轴)。

[0508] 图32为补充表14,表示3-态二级结构和3-态溶剂可及性预测模型在来自蛋白数据库的注释样本上的性能,使用3637个无关蛋白序列进行训练,400个用于验证,500个用于测试。仅选择蛋白数据库中序列相似性小于25%的蛋白。我们将深度学习网络的准确度作为性能度量进行报告,这是因为这三种类别在二级结构或溶剂可及性方面严重失衡。

[0509] 图33为补充表15,表示对使用来自DSSP数据库注释的二级结构标记的深度学习网络的性能比较,可用于使用预测二级结构标记的深度学习网络。

[0510] 图34为补充表17,表示评估的10000个保留灵长类动物变体的准确度值和20个分类器中每个分类器的DDD病例对比对照例新生变体的P值。仅具有人类数据的primateAI模型为使用包括唯一常见人类变体(种群中大医院0.1%的83.5K个变体)的标记良性训练数据集的深度学习网络,而具有人类和灵长类数据的primateAI模型为对全集385K个标记良性变体进行训练得到的深度学习网络,变体包括常见人类变体和灵长类动物变体。

[0511] 图35为补充表19,表示对DDD病例对比数据集中新生变体不同分类器的性能进行比较,仅限于605个疾病相关基因。为了使不同方法规范化,对于每个分类器,我们对阈值进行识别,基于DDD及对照例中的富集,分离器预期地预测致病性和良性变体的相同数量。所示分类准确度为该阈值处真阳性和真阴性误差率的平均值。

[0512] 图49A、49B、49C、49D及49E描述了整个人类等位基因频谱的错义/同义突变率。图49A表示ExAC/gnomAD数据库中123136人类个体中观察到的错义变体和同义突变变体通等位基因频率分为四类。阴影灰色条纹表示每个目录中的同义突变变体数,深绿色条纹表示错义变体。每个条纹的高度缩减到每个等位基因频率目录中同义突变变体数,调整突变率

后,显示错义/同义突变数和显示错义/同义突变率。图49B及49C描述了与黑猩猩常见变体(如图49B)和黑猩猩单变体(如图49C)状态一致(IBS)的人类错义及同义突变变体的等位基因频谱。红框表示相对于罕见人类等位基因频率(小于0.1%)在常见等位基因频率(大于0.1%)处黑猩猩错义变体的损耗,同时也表示卡方检验(χ^2)测试值P。

[0513] 图49D表示在至少一个非人类灵长类物种中观察到的人类变体。相对于取样自ExAC/gnomAD等位基因频率(中行)的一组30个人类样本中的ClinVar变体及灵长类中观察到的变体(底行),图49E表示整个ClinVar数据库(顶行)中良性和致病性错义变体数。排除了确定的冲突良性和致病性变体和仅注释为未确定意思的变体。

[0514] 图50A、50B、50C及50D表示对与其他物种状态一致的错义变体进行纯化选择。图50A描述了人类错义变体和同义突变变体的等位基因频谱,这些变体与四个非灵长类哺乳动物物种(老鼠、猪、山羊及牛)中变体状态相一致。红框表明在常见人类等位基因频率(大于0.1%)上错义变体的损耗,同时也表示卡方检验(χ^2)测试值P。

[0515] 图50B为一散布图,表示对比来自人类物种进化距离在常见人类等位基因频率(大于0.1%)上其他物种中观察到的错义替代物的损耗,以平均分支长度为单位表示(每个核苷酸位置替代物的平均数)。紧挨着物种名称,表示每个物种和人类之间的总分支长度。显示了具有变体频率的物种的单个和常见变体的损耗值,大猩猩除外,其包括相关个体。

[0516] 相对于灵长类中观察到的变体(中行)及老鼠、猪、山羊和牛中观察到的变体(底行),图50C阐明了取样自一组ExAC/gnomAD等位基因频率(顶行)的良性和致病性错义变体的数量。排除了确定的冲突良性和致病性变体和仅注释为未确定意思的变体。

[0517] 图50D为一散布图,表示对比来自人类物种进化距离在常见人类等位基因频率(大于0.1%)上近缘种中对观察到的固定错义替代物的损耗(以平均分支长度为单位表示)。

[0518] 图51表示在无纯化选择的情况下整个人类等位基因频谱预期的错义:同义突变率。阴影灰色条纹代表同义突变变体数量,暗绿色条纹代表错义变体数量。虚线表示同义突变变体形成的基线。表明了每个等位基因频率类别的错义:同义突变率。根据一个实施方式,基于用于控制基因变异中突变率和GC偏差的变体的三核苷酸上下文,通过从包括123136个外显子组的ExAC/gnomAD数据集选取基因内变体和使用这些变体来估计预计属于四个等位基因类别中每个类别的部分变体,来对每个等位基因类别中的预计错义及同义突变数进行计算。

[0519] 图52A、52B、52C及52D描述了CpG和非CpG变体的错义:同义突变率。图52A及52B表示整个人类等位基因频谱中CpG变体(如图52A)及非CpG变体(如图52A)的错义:同义突变率,所有变体均来自ExAC/gnomAD外显子组。图52C及52D表示整个人类等位基因频谱中CpG变体(如图52C)及非CpG变体(如图52D)的错义:同义突变率,该错义:同义突变率仅受限于与黑猩猩常见多态性状态一致的人类变体。

[0520] 图53、54及55阐明了与六个灵长类状态一致的人类变体的错义:同义突变率。对于与变异状态相一致的ExAC/gnomAD变体,人类等位基因频谱的错义:同义突变率的模式在黑猩猩、倭黑猩猩、大猩猩、红毛猩猩、恒河猴及狨猴中出现。

[0521] 图56模拟了新常见错义变体的饱和度,这些新常见错义变体通过增大调查的人类组大小来发现。模拟中,根据gnomAD等位基因频率对每个样本的基因型进行取样。部分发现的gnomAD常见变体在10到10000个变体的每个样本量中的100个模拟中较为平均。

[0522] 图57表示基因组中不同保守谱中primateAI的准确度。x轴代表具有99脊椎动物比对的序列周围51AA的百分比比对性。基于10000个保留灵长类动物变体的测试数据集,y轴代表每个保守区域内变体的primateAI准确度的分类性能。

[0523] 图58为补充表5,表示对常见人类变体的标记的良性训练数据集的影响及存在于非人类灵长类中的变体。

[0524] 图59为补充表8,表示等位基因频率对预期的错义:同义突变率的影响。使用三核苷酸上下文对突变率和基因转换偏差,基于至少20-30n远离基因内区域内变体的等位基因频谱,对同义突变和错义变体的预期数量进行计算。

[0525] 图60为补充表9,表示ClinVar分析。根据一实施方式,从下载自2017年10月构建的ClinVar数据库的变体中滤除注释冲突的错义变体并排除意义不确定的变体,剩下17775个良性变体和24853个致病性变体。

[0526] 根据一实施方式,图61为补充表10,表示ClinVar中发现的来自其他物种的错义变体的数量。要求变体与相应人类变体状态相一致,并在读码框架内其他两个位置处就有一致的核苷酸,以保证编码结构相同。

[0527] 图62为表1,表示智力障碍中发现的14个额外候选基因的一个实施方式,此前,该额外候选基因在原始DDD研究中未达到全基因组意义阈值。

[0528] 图63为表2,表示ClinVar中致病性变体和良性变体间Grantham得分的平均差的一个实施方式,该平均差是605个疾病相关基因中DDD病例对比对照例新生变体差值的两倍。

[0529] 数据生成

[0530] 本文中使用的坐标均指人类基因组结构UCSC hg19/GRCh37,包括其他物种中变体的坐标,并使用本节中所述的方法使用多序列比对将其映射到hg19。从UCSC基因组浏览器中下载了人类的蛋白质编码的DNA序列以及和99个脊椎动物基因组的多序列比对,用于hg19构建(<http://hgdownload.soe.ucsc.edu/goldenPath/hg19/multiz100way/alignments/knownCanonical.exonNuc.fa.gz>)。对于具有多个规范基因注释的基因,本文选择了最长的编码转录本。

[0531] 我们从外显子组整合数据库(ExAC)/基因组聚合数据库(gnomAD) v2.0中下载了人类外显子组多态性数据,该数据收集了来自全球8个亚种群(123://gnomad.broadinstitute.org/)的123,136个体的全基因组测序数据(WES)。我们排除了在ExAC VCF文件中注释的,未通过默认质量控制过滤器的变体或落在规范编码区域之外的变体。为避免由于平衡选择带来的影响,我们还从灵长类动物分析的扩展MHC区域(chr6:28,477,797-33,448,354)中排除了变体。类人猿基因组测序项目提供了24只黑猩猩,13只倭黑猩猩,27只大猩猩和10只红毛猩猩的全基因组测序数据和基因型(包括来自苏门答腊亚种的5只和婆罗洲亚种的5只,我们对其折叠以进行下游分析)。对黑猩猩和倭黑猩猩的研究,我们提供了另外35只黑猩猩的基因组序列。但是,由于没有使用与类人猿基因组测序项目相同的方法来识别这些额外的黑猩猩的变体,因此我们将其从等位基因频谱分析中排除,仅将它们用于训练深度学习模型。这些灵长类动物多样性研究的变异已经映射到人类参考序列中(hg19)。此外,对于狨猴和恒河猴,使用了16只恒河猴个体和9只狨猴个体来测定这些物种的基因组原始序列的变异,但是尚未获得个体水平的信息。

[0532] 类人猿基因组测序项目4提供了24只黑猩猩,13只倭黑猩猩,27只大猩猩和10只红

毛猩猩的全基因组测序数据和基因型(包括来自苏门答腊亚种的5个和婆罗洲亚种的5只,我们对其折叠以进行下游分析)。对黑猩猩和倭黑猩猩的研究,我们提供了另外35只黑猩猩的基因组序列。但是,由于没有使用与类人猿基因组类别项目相同的方法来识别这些额外的黑猩猩的变体,因此我们将其从等位基因频谱分析中排除,仅将它们用于训练深度学习模型。这些灵长类动物多样性研究的变异已经映射到人类参考序列中(hg19)。此外,对于狢猴和恒河猴,使用了16只恒河猴个体和9只狢猴个体来测定这些物种的基因组原始序列的变异,但是尚未获得个体水平的信息。

[0533] 为了与其他灵长类动物和哺乳动物进行比较,我们还从dbSNP下载了其他物种的SNP,包括恒河猴,狢猴,猪,牛,山羊,小鼠,鸡和斑马鱼的SNP。dbSNP还包括其他红毛猩猩变体,我们仅将其用于训练深度学习模型,这是因为单个基因型信息无法用于等位基因频谱分析。我们放弃了其他物种,例如狗,猫或绵羊,这是因为dbSNP为这些物种只提供了有限数量的变体。

[0534] 为了将变体映射到人类,我们使用了99种脊椎动物的多物种比对来确保与人类蛋白质编码区直系同源1:1映射。使用直系同源的多物种比对映射变体对于消除由假基因或逆转座序列引起的伪影至关重要,伪影或逆转座序列是在使用多对一映射的工具(例如liftOver)时,直接在物种之间映射SNP时发生的。如果dbSNP中物种的基因组结构与99种脊椎动物多序列比对中的物种的基因组结构不匹配,我们使用liftOver将变体更新为多序列比对中使用的基因组结构。如果变体以参考/替代的方式出现,则我们接受变体为状态一致性,例如,如果人类参考序列中为G,备选等位基因中为A,则认为该变体在其他物种中处于同一状态,其参考序列中为A,备选等位基因中为G。为了确保变体在人类和其他物种中具有相同的预测蛋白质编码结果,我们要求密码子中的其他两个核苷酸在物种间错义和同义变体中均相同。在分析中包括的每个物种的多态性在补充数据文件1中列出,详细的度量标准在补充表1中显示。

[0535] 为了确保每个dbSNP提交批次的变体具有高质量并且和人类的变体进行比对,我们计算了每个批次的错义突变/同义突变率,确认这低于预期的2.2:1的比率;大多数物种的比率均低于1:1,尤其是斑马鱼和老鼠,它们的有效种群数量非常大。我们从进一步分析中排除了错义突变/同义突变比率异常高的母牛的两批SNP(比率为1.391的snpBatch_1000_BULL_GENOMES_1059190.gz和比率为2.568的snpBatch_COFACTOR_GENOMICS_1059634.gz)。其余奶牛批次的平均错义突变/同义突变比率为0.8:1。

[0536] 校正等位基因频率对错义突变/同义突变比率,突变率,遗传漂移和基于GC的基因转化的影响

[0537] 除了纯化选择的作用外,在高等位基因频率下观察到的人类错义变体的损耗也可能受到与自然选择无关的因素的影响。在种群中以特定等位基因频率出现中性突变的可能性是突变率,基因转换和遗传漂移的函数,并且即使没有选择力的情况下,这些因素可能潜在地在整个等位基因频谱的错义突变/同义突变率中引入偏差。

[0538] 为了在没有蛋白编码选择的情况下在每个等位基因频率类别上计算预期的错义突变/同义突变比率,我们在每个外显子上游31-50bp和下游21-50bp的内含子区域内选择了变体。选择这些区域的距离要足够远,以避免扩展的剪接基序的影响。因为这些区域位于ExAC/gnomAD外显子组的外显子组捕获序列的边缘附近,所以为确保清楚地确定变体,我们

删除了所有chrX区,并排除了平均解读深度<30的区域。每个变体及其紧邻的上游和下游核苷酸属于64种三核苷酸背景之一。如果我们将中间核苷酸突变为其他三个碱基,则总共可能有 $64 \times 3 = 192$ 个三核苷酸构型。由于三核苷酸构型及其反向互补是等效的,因此96种三核苷酸背景有效。我们观察到三核苷酸背景对突变率有非常强的影响,而对GC偏向的基因转化的影响较小,这使得三核苷酸背景对于构建这些变量有效。

[0539] 在这些内含子区域中,我们从126,136个ExAC/gnomAD外显子组中选取了每个变体,并根据等位基因频率的四个类别将它们分为 4×192 个类别(单例,超过单例 $\sim 0.01\%$, $0.01\% \sim 0.1\%$, $>0.1\%$)和192种三核苷酸背景。我们通过将可能的变体总数除以该三核苷酸背景(通过以三种不同方式将内含子序列中的每个核苷酸替换而获得),将在 4×192 个类别(等位基因频率 $\times$ 三核苷酸背景)中观察到的变体的数量归一化。因此,对于192种三核苷酸背景中的每一个,在没有蛋白质编码选择的情况下,我们已经获得了落入4个等位基因频率类别中的变体的预期比例。这隐含地模拟了由于三核苷酸背景的差异而导致的突变率、GC偏向基因转换以及遗传漂移的影响(补充表7)。

[0540] 为了获得每个等位基因频率类别中预期的错义突变/同义突变率,我们计算了人类基因组中可通过单核苷酸取代获得的同义和错义突变的总数,并将它们分别分配给192种三核苷酸背景中的一种。对于每种情况,我们使用 4×192 表来计算期望落入4个等位基因频率类别中的每个变体的数量。最后,我们总结了192种三核苷酸背景中同义和错义变体的数量,以获得四个等位基因频率类别中每个类别的同义和错义变体的预期总数(图51和补充表8(图59))。

[0541] 预期的错义突变/同义突变率在等位基因频谱上几乎是恒不变的,并且接近在没有自然选择的情况下新生变体所期望的2.23:1的比率,但是单例变体除外,其预期的错义突变/同义突变比率为2.46:1。这表明由于与蛋白质编码选择压力无关的因素(突变率,基因转换,遗传漂移)的作用,ExAC/gnomAD中具有单例等位基因频率类别的变体的预期错义突变/同义突变比率比默认情况下的新生突变高10%。为了解决这个问题,我们在等位基因频率分析中将单例的错义突变/同义突变率降低了10%(图49A,49B,49C,49D和49E,图50A,50B,50C和50D)。这个很小的调整降低了灵长类动物和其他哺乳动物中存在的常见人类变体的预计错义突变损耗约 $\sim 3.8\%$ (如图49A,49B,49C,49D和49E和图50A,50B,50C和50D所示)。单例变体的错义突变/同义突变率较高是由于过渡突变(更可能产生同义变化)具有较高的等位基因频率,这归因于突变率高于颠换突变(很可能产生错义变化)。

[0542] 此外,这解释了在ExAC/gnomAD中观察到的单例变体的错义突变/同义突变比率为2.33:1,其超过了对新生突变的预期比率,即2.23:1。在考虑了等位基因频谱对错义突变/同义突变率的影响后,这实际上反映了单例变体与预期相比减少了5.3%,这可能是由于选择了对抗具有新生显性遗传模式的致病性错义突变。实际上,当我们仅考虑单倍体不足基因且具有较高的功能丧失的情况时($pLI > 0.9$),ExAC/gnomAD单例变体的错义突变/同义突变率是2.04:1,表明单倍体不足基因大约损耗 $\sim 17\%$ 。此结果与先前的估计相符,即,假设某种程度的不完全外显的情况下,20%的错义突变相当于功能丧失突变。

[0543] 由于它们的突变率差异很大,我们还专门研究了在人类等位基因频谱上CpG和非CpG变体的错义突变/同义突变率(图52A,52B,52C和52D)。我们证实,对于CpG和非CpG突变,与黑猩猩常见多态性状态一致性的人类变体在等位基因频谱上具有几乎不变的错义突变/

同义突变率。

[0544] 在其他物种中具有多态性的状态一致性的人类错义变体的损耗

[0545] 为了评估在人类的共同等位基因频率(>0.1%)下是否可以容忍来自其他物种的变体,我们鉴定了在其他物种中具有变异的状态一致的人类变体。对于每个变体,我们根据其在人类种群中的等位基因频率将其分配为四个类别之一(单例,超过单例~0.01%, 0.01%~0.1%, >0.1%),并估计罕见(<0.1%)和常见(>0.1%)变体之间的错义突变/同义突变比率(MSR)的下降情况。在常见人类等位基因频率(>0.1%)下状态一致性错义变体的损耗显示了可以通过自然选择滤除的其他物种的有害的变体的比例。

$$[0546] \quad \% \text{depletion} = \frac{\text{MSR}_{\text{rare}} - \text{MSR}_{\text{comm}}}{\text{MSR}_{\text{rare}}}$$

[0547] 计算每种物种的错义突变/同义突变率和损耗百分比,并显示在图50B和补充表2中。此外,对于黑猩猩常见变体(图49A),黑猩猩单例变体(图49C)和哺乳动物变体(图50A),我们在2×2列联表上进行了卡方(x2)均一性检验,以测试罕见和常见变体之间的错义突变/同义突变率的差异是否具有显著性。

[0548] 因为测序只针对类人猿多样性项目中少量的个体,我们使用EXAC/gnomAD中的人类等位基因频谱来估计黑猩猩种群中罕见(<0.1%)或常见(>0.1%)采样变体的比例。我们根据ExAC/gnomAD等位基因频率对24名个体群组进行了抽样,并确定了在该群组中一次或多次观察到的错义变体。多次观察到的变体在常见种群中(>0.1%)有99.8%的可能性,而仅观察到一次的变体仅有69%的可能性。在图49B和49C中,我们表明,由于某些黑猩猩单例变体是罕见的有害突变,我们在人等位基因频率较高时观察到了单例黑猩猩变体的损耗,但不是黑猩猩中的常见的变体。在含有24个个体的群组中,大约一半的黑猩猩变体仅被观察到一次,大约一半被观察到多次。

[0549] 为了确认在较远古的哺乳动物中观察到的错义变体的损耗不是由于更保守的基因的混杂效应所致,从而更准确地进行了比对,我们重复了上述分析,但只限于和人类相比的11个灵长类和50个哺乳动物的多序列比对中,平均核苷酸同一性>50%的那些基因(参见补充表3)。这分析中排除了约7%的人类蛋白质编码基因后,基本上没有影响结果。

[0550] 在灵长类,哺乳动物和远缘脊椎动物之间的固定替换

[0551] 为确保我们使用dbSNP变异的结果不受变体数据问题或驯化伪影的影响(因为从dbSNP中选择的大多数物种都已驯化),我们还使用了来自密切相关的物种对的固定替代物重复了分析了种内多态性。我们从UCSC基因组浏览器下载了100种脊椎动物的系统发育树(<http://hgdownload.soe.ucsc.edu/goldenPath/hg19/multiz100way/hg19.100way.commonNames.nh>),他们的系统发育距离以分支长度(平均每个位置的核苷酸取代数)表示。我们选择了密切相关的物种对(分支长度<0.25)进行进一步分析。为了确定密切相关的物种对之间的固定取代,我们从UCSC基因组浏览器下载了99种脊椎动物基因组与人类的多序列比对的编码区,以及19种哺乳动物(16个灵长类)基因组与人类的比对的编码区。因为99种脊椎动物比对中不存在某些灵长类动物,例如倭黑猩猩(<http://hgdownload.soe.ucsc.edu/goldenPath/hg38/multiz20way/alignments/knownCanonical.exonNuc.fa.gz>),所以19种哺乳动物多物种比对是必要的。如图50D和补充表4所示,我们总共获得了15对密切相关的物种,其中包括5种灵长类物种对。

[0552] 我们对19种哺乳动物或99种脊椎动物基因组与规范编码区内的人类基因组进行了多序列比对,并在补充数据文件2中列出了每对选定的脊椎动物之间获取的替换的核苷酸。这些替换的核苷酸被映射到人类基因组,要求在人类和其他物种之间,密码子中的其他两个核苷酸没有变化,并且以参考或其他方式接受变体。使用具有相关物种对的固定替换的状态一致性的人类变体,我们计算了罕见($<0.1\%$)和常见($>0.1\%$)等位基因频率类别中变体的错义突变/同义突变率比率,从而获得了负选择下情况下替换的核苷酸的比例,如补充表4所示。

[0553] 针对人类,灵长类,哺乳动物和其他脊椎动物的多态性数据的ClinVar分析

[0554] 为了检查与其他物种状态一致的变体的临床影响,我们下载了ClinVar数据库的变体总结发行版(ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/clinvar_20171029.vcf.gz于2017年11月2日发布)12。该数据库在hg19基因组结构上包含324,698个变体,其中122,884变体是个错义单核苷酸变体,映射到我们的蛋白质编码基因列表(补充表9)上。ClinVar数据库中的大多数变体没有错义的结果,因此被排除在外。接下来,我们用与致病性相矛盾的解释过滤变体,并仅保留具有良性,可能良性,致病性和可能致病性注释的变体。我们将“变体”与“良性”或“可能的良性”注释合并为一个类别,并将“变体”与“致病性”或“可能是致病性注释”合并。按照补充表9所示的过滤步骤,在病原体类别中共有24,853个变体,在良性类别中总共有17,775个变体;其余的被排除,因为它们是未知意义的或相互矛盾的注释的变体。

[0555] 为了获得人类种群中ClinVar错义变体的基线,我们对来自ExAC/gnomAD等位基因频率的群组中抽样的30名个体研究了ClinVar错义变体。该群组的大小以大致反映灵长类动物多样性项目研究中测序的个体数量为准。我们从100个这样的模拟变体中报告了30个人类群组(图49E)中致病性和良性变体的平均数量。因为管理者已在ClinVar中系统地注释了具有良性后果的常见人类变体,所以我们排除了等位基因频率大于1%的变体,以避免这种偏差。

[0556] 我们分析了在灵长类,哺乳动物和其他脊椎动物中具有状态一致性的ClinVar变体。每个物种的良性和致病性变体的数量如补充表10所示。人类,灵长类和较远古的哺乳动物中存在的ClinVar变体的数量如图49E和50B所示。卡方(χ^2)检验的同质性结果显示了良性与致病性变体之比的差异。

[0557] 用于模型训练的良性变体的生成

[0558] 在人类种群中常见的变体在很大程度上不受影响,除了创始效应或平衡选择的罕见情况外,这使它们适合作为不受人类解释偏差影响的机器学习的良性训练数据集。我们使用了来自ExAC/gnomAD数据库(版本v2.0)中123,136个外显子组的等位基因频率数据,不包括未通过过滤器的变体,这给我们带来了83,546个错义变体,其总体种群等位基因频率 $\geq 0.1\%$ 规范的蛋白编码转录本。

[0559] 根据我们先前的结果,灵长类动物中的变体在人类中基本上是良性的,我们创建了一个良性训练数据集,用于机器学习,其中包括常见人类的变体($>0.1\%$ 等位基因频率),来自类人猿多样性项目以及其他灵长类动物测序项目中的黑猩猩,倭黑猩猩,大猩猩和红毛猩猩的变体,以及dbSNP中的恒河猴,红毛猩猩和狨猴变体。根据一种实施情况,总共将301,690个独特的灵长类动物变体添加到了良性训练集中。不同来源的良性训练变体的数

量在补充表5中显示。

[0560] 需要注意的是,尽管大多数灵长类动物的变体在它们各自的种群中是常见的,但其中少数是罕见的变体。因为非人类灵长类的测序个体数量有限,所以我们希望确定的变体集合通常代表常见的变异。实际上,我们发现每个灵长类物种的变体的错义突变/同义突变率小于新生突变预期的2.23:1比率的一半,这表明这些是大多数已经通过筛选的常见变体。此外,对于黑猩猩群组,我们估计,约84%的确定变体以共同的等位基因频率($>0.1\%$)存在于它们各自的种群中。由于大约50%的新出现的错义突变是通过在常见人类等位基因频率($>0.1\%$) (图49A) 上进行纯化选择而过滤的,因此该数字与~16%的罕见变体相符,占观察到的人类错义变体的8.8%,这些错义变体与观察到的灵长类动物变异具有状态一致性(图49D)。

[0561] 应用约20%的人类错义突变为等价于功能丧失的估计,灵长类动物的变体预计将包含3.2%的完全致病性突变,91.2%的良性突变(容许等位基因频率 $>0.1\%$)和5.6%的中间不能完全消除基因功能的突变,但足以在常见等位基因频率($>0.1\%$)处被滤除。尽管此训练数据集存在已知的缺陷,但与仅有常见的人类变体相比,在包含常见人类变体和灵长类动物变体的良性训练数据集上进行训练时,深度学习网络的分类准确度要好得多。因此,在当前的分类准确度下,可供使用的训练数据量是更严格的限制。由于每个灵长类物种中都有大量个体被测序,因此有可能制备包含更高比例的灵长类动物常见变体的训练数据集,从而减少训练数据集中致病性变体的掺杂并进一步提高分类性能。

[0562] 生成未标记变体以补充良性训练数据集

[0563] 所有可能的错义变体是通过在规范编码区的每个碱基位置上用其他三个核苷酸取代而产生的。我们排除了来自ExAC/gnomAD的123,136个外显子组中观察到的变体,以及起始或终止密码子中的变体。总共产生了68,258,623个未标记变体。我们将每个未标记变体分配到96种不同的三核苷酸背景类别中的一种中去。我们使用半监督方法训练了深度学习网络,该方法是通过从该未标记数据集中(按照三核苷酸背景)采样与良性数据集中变体匹配的未标记数据集中的变体,并训练分类器来区分良性和未标记训练实例。

[0564] 未标记变体的过滤

[0565] 通过展示良性和未标记变体的示例以及侧翼氨基酸序列,深度学习网络可以帮助了解高度不耐受的突变的蛋白区域。但是,蛋白质序列区域内不存在常见的变体,这可能是由于高度的纯化选择,或者可能是由于技术伪像的原因,其阻止了该区域中的变体被识别。为了纠正后者,我们从ExAC/gnomAD数据集平均覆盖范围 <1 的区域的良性和未标记数据集中删除了变体。同样地,当在训练过程中良性数据集中的未标记变体与灵长类动物的变体匹配时,我们从灵长类动物在多序列比对中与人没有直系同源可比序列的区域中排除了未标记的变体。

[0566] 保留灵长类动物变体进行验证和测试,以及受影响和未受影响的个体的新生变体

[0567] 为了验证和测试深度学习网络,我们随机抽取了两套10,000个灵长类动物变体进行验证和测试,但是我们没有进行训练。其余的灵长类动物变体与常见人类变体($>0.1\%$ 等位基因频率)一起用作训练深度学习网络的良性数据集。此外,我们还抽样了两套10,000个未标记变体,它们与保留的灵长类动物变体相匹配,用于验证集合和测试集。

[0568] 我们使用验证集中的10,000个被保留的灵长类动物变体和相匹配的10,000个未

标记变体,通过测量两个集合区分两个变体的网络的能力,我们可以监控深度学习网络的性能。一旦网络性能达到饱和,我们就能够确定训练的停止点并避免过度拟合。

[0569] 我们使用了在测试数据集中保留的10,000个灵长类动物变体来对深度学习网络以及其他20个分类器进行基准测试。由于不同的分类器的得分分布差异很大,因此我们使用这些未标记变体来识别每个分类器的第50个百分点阈值。我们对10,000个保留的灵长类动物变体测试集中的变体比例(在该分类器的第50个百分点阈值处被归类为良性)进行了基准测试,以确保方法之间的合理比较。

[0570] 在临床环境中,为了评估在神经发育障碍患者和健康对照中使用新生变体进行深度学习网络的性能评估,我们从解密发展障碍(DDD)研究和Simons Simplex Collection (SSC)自闭症研究的健康同级对照中下载了新生变体。DDD研究为新生变体提供了置信度,我们从DDD数据集中排除了阈值 <0.1 的变体,将其作为由于变体识别错误而导致的潜在误报。我们共有来自DDD感染者的3,512个错义新生变体和来自健康对照的1,208个错义新生变体。

[0571] 为了更好地模拟现实世界中区分候选疾病基因组中意义不明的良性和致病性变体的临床情况,我们将分析仅限于与DDD研究中疾病相关的605个基因中的新生变体($p < 0.05$),仅根据蛋白质截短变异(补充表18)计算得出。我们通过在给定基因特异性突变率和考虑的染色体数的预期新生突变数的无假设条件下,计算统计的显著性,然后评估蛋白质截短新生突变的基因特异性富集。我们选择了标称P值 <0.05 的605个基因。我们计算了605个基因(图22A)中同义和错义新生突变的过量数,将其作为观察到的新生突变与预期新生突变的计数之比,以及观察到的新生突变减去预期新生突变的差。在这605个基因中,我们观察到了DDD感染者的380个新生错义突变(图22A)。对于每个分类器,包括我们自己的分类器,一小部分变体没有预测,通常是因为它们没有映射到分类器所使用的相同转录本模型。因此,对于我们的深度学习网络,我们使用来自DDD感染者的362新生错义突变和来自健康对照的65新生错义突变进行了下游分析,如图22A,22B,22C,22D和22E中所示。

[0572] 随着灵长类种群测序数量的增加,所有可能的人类错义突变的饱和度

[0573] 我们调查了504种灵长类物种中常见的变体对所有约70M的人类错义突变的饱和度。对于每种灵长类动物,我们模拟的是人类观察到的常见错义变体数量的四倍(约83,500个等位基因频率 $>0.1\%$ 的错义变体),因为人类每个个体的变体数量大约是其其他灵长类动物的一半,并且大约50%的人类错义变体已通过 $>0.1\%$ 等位基因频率处纯化选择而被滤出(图49A)。我们根据观察到的人类常见错义变体在96种三核苷酸背景中的分布,分配了模拟变体。例如,如果2%的人类常见错义变体来自CCG $\rightarrow$ CTG三核苷酸背景,那么我们将要求2%的模拟变体是随机采样的CCG $\rightarrow$ CTG突变。使用三核苷酸背景可以控制突变率,遗传漂移和基因转化偏差的影响。

[0574] 假设我们确定了每种灵长类物种中所有常见的变体($>0.1\%$ 等位基因频率),图23D中的曲线显示了504个灵长类物种中常见变体所造成的约70M可能的人类错义突变的累积饱和度。从图49A中可以看出,大约50%的人类错义突变在人类和其他灵长类动物中均具有足够的有害性,可防止它们升高到常见的等位基因频率($>0.1\%$),因此,图23D中的曲线代表了非有害的人类错义突变的比例,这些错义突变随着灵长类物种数量的增长而被常见的灵长类变异所饱和。我们表明,在504种灵长类动物中,大多数非有害的人类错义突变将

被饱和,由于其较高的突变率,非有害的CpG突变将被较少种类的物种所饱和。

[0575] 为了模拟随调查的人类群组的增加而发现的人类常见的错义变体(>0.1%等位基因频率)的比例(图36),我们根据gnomAD等位基因频率对基因型进行了采样。对于大小为100到100K的样本,对100个模拟样本中发现的gnomAD常见错义变体的比例进行了平均计算。

[0576] 二级结构和溶剂可及性的预测

[0577] 致病性预测的深度学习网络包含36个总卷积层,其中19个用于二级结构和溶剂可及性预测网络的卷积层,以及17个用于主要致病性预测网络的卷积层,其将二级结构和溶剂可及性网络的结果作为输入。因为大多数人类蛋白质的晶体结构是未知的,所以我们训练了两个模型以使网络能够从一级序列中学习蛋白质结构。如图6所示,两种型号都使用相同的网络架构和输入。二级结构和溶剂可及性网络的输入是一个51长度×20个氨基酸的位置频率矩阵,该矩阵编码来自人类与99种其他脊椎动物的多序列比对的保守信息。

[0578] 二级结构网络经过训练可以预测三态二级结构:α螺旋(H),β折叠(B)和卷曲(C)。溶剂可及性网络经过训练可以预测三态溶剂可及性:埋藏(B),中间(I)和暴露(E)。这两个网络仅以一级序列作为输入,并使用Protein DataBank中已知晶体结构的标记进行了训练。该模型预测了每个氨基酸残基的一种状态。

[0579] 二级结构和溶剂可及性的预测数据准备

[0580] 我们使用了蛋白质数据库中无关的晶体结构来训练模型。去除了具有超过25%序列相似性的氨基酸序列。总共使用了6,367个蛋白质序列进行训练,使用400个蛋白质序列用于验证,500个蛋白质序列进行测试(补充表13)。可从RaptorX网站(<http://raptorx.uchicago.edu/download/>)获得用于训练的数据,这些数据包括氨基酸序列,二级结构和溶剂可及性标签。

[0581] 大部分解析的晶体结构都是非人类蛋白质,因此,为了预先训练二级结构和溶剂模型,我们使用了RaptorX套件(基于PSI-BLAST)来获取相关序列,这是因为基于人类的多序列比对通常无法获取。我们使用RaptorX的CNFsearch1.66_release工具为蛋白质生成了多序列比对,并从99个最接近的比对中计算每个位置的氨基酸,以形成位置频率矩阵。例如,使用RaptorX检索1u71A.fasta蛋白的多序列比对的特定命令如下:

[0582] `%./buildFeature-i 1u71A.fasta-c 10-o./TGT/1u71A.tgt`

[0583] `%./CNFsearch-a 30-q 1u71A`

[0584] 对于数据集中的每个氨基酸位置,我们从对应于侧翼51个氨基酸的位置频率矩阵中提取了一个窗口,并以此预测51个长度的氨基酸序列中心的氨基酸的二级结构或溶剂可及性的标签。使用DSSP软件直接从蛋白质的已知3D晶体结构获得二级结构和相对溶剂可及性的标签,并且不需要从一级序列进行预测。为了将二级结构和溶剂可及性网络作为致病性预测网络的一部分,我们从基于人类的99个脊椎动物多序列比对中计算了位置频率矩阵。尽管从这两种方法生成的保守矩阵通常相似,但我们在训练致病性预测的过程中通过二级结构模型和溶剂可及性模型对其进行了反向传播,从而可以对参数权重进行微调。

[0585] 模型架构和训练

[0586] 我们训练了两个单独的深度卷积神经网络模型来预测蛋白质的二级结构和相对溶剂可及性。两种模型的架构和输入数据相同,但预测状态不同。我们进行了详细的超参数

搜索,以优化模型以获得最佳性能。我们用于致病性预测的深度学习网络以及用于预测二级结构和溶剂可及性的深度学习网络都采用了在图像分类方面被广泛采用的残差块架构。剩余块包括卷积的重复单元,并散布有跳跃连接,这些跳跃连接允许来自较早层的信息跳过残差块。在每个残差块中,首先对输入层进行批归一化,然后使用线性修正单元(ReLU)激活层。激活后,通过1D卷积层。来自1D卷积层的中间层输出再次进行批量归一化并激活ReLU,然后再执行另一个1D卷积层。在第二个1D卷积的结尾,我们将其输出与原始输入加到残差块中,该残差块通过允许原始输入信息绕过残差块而充当跳跃连接。在这样的架构中,其作者将其称为深度残差学习网络,其输入保持其原始状态,并且残差连接不受模型的非线性激活的影响,从而可以对较深的网络进行有效的训练。图6和补充表11(图7A和7B)和12(图8A和8B)中提供了详细的架构。

[0587] 在残差块之后,softmax层为每个氨基酸计算三种状态的概率,其中最大的softmax概率决定了氨基酸的状态。使用ADAM优化程序对模型的整个蛋白质序列进行累积分类交叉熵损失函数训练。在对网络进行了二级结构和溶剂可及性的预训练之后,我们不再将网络的输出直接作为致病性预测网络的输入,而是将其置于softmax层之前,以便更多的信息能够通过致病性预测网络。

[0588] 三态二级结构预测模型的最佳测试准确度为79.86%(补充表14),与DeepCNF模型30预测的最新准确度相似。三态溶剂可及性预测模型的最佳测试准确度为60.31%(补充表14),与RaptorX在类似训练数据集上预测的当前最佳准确度相似。我们还比较了使用约有4000种具有晶体结构的人蛋白质的DSSP注释结构标签与仅使用预测结构标签的标准PrimateAI模型时的神经网络的预测情况。当使用带DSSP注释的标签时,我们并没有提高致病性预测的准确度(补充表15)。

[0589] 用于致病性预测的深度学习模型的输入特性

[0590] 过滤后,用于致病性预测网络的训练数据集包含385,236个标记为良性变体和68,258,623个未标记变体。对于每个变体,我们生成了以下输入特征。每个变体的第一个输入特征是其51个长度的侧翼氨基酸序列,即从hg19的参考序列获得的变体的每侧各25个氨基酸,以提供变体的深度学习模型的序列背景。该侧翼参考序列的长度总计为51个氨基酸。通过经验观察,我们发现使用氨基酸代表蛋白质序列比使用核苷酸代表蛋白质编码序列更有效。

[0591] 第二个特征是长度为51的人类侧翼氨基酸序列,其替代氨基酸在中心位置被变体取代。除了序列的中间位置包含替代氨基酸而不是参考氨基酸之外,替代侧翼序列与第一特征中的参考侧翼序列相同。参考和替代人类氨基酸序列均被转换为长度为 51×20 的一位有效编码载体,其中每个氨基酸由19个氨基酸(值为0)和单个氨基酸(值1)的载体表示。

[0592] 从99种脊椎动物的变体的多序列比对中生成三个位置频率矩阵(PFM),其中一个为11种灵长类动物,一个为50种哺乳动物(不包括灵长类),一个为38种脊椎动物(不包括灵长类和哺乳动物)。每个PFM的大小为 $L \times 20$,其中L是变体周围的侧翼序列的长度(在我们的案例中,L代表51个氨基酸)。

[0593] 对于预先训练的三态二级结构和三态溶剂可及性网络的输入,我们使用了由99种脊椎动物的多序列比对生成的单个PFM矩阵,长度也为51,深度为20。在对来自Protein DataBank的已知晶体结构的网络进行预训练之后,删去了二级结构和溶剂模型的最后两层

(整个maxpool层和输出层),上一层的 51×40 形状的输出用作为致病性预测网络的输入。我们允许通过网络的结构层进行反向传播,从而微调参数。

[0594] 半监督学习

[0595] 因为半监督学习算法在训练过程中同时使用了标记和未标记的实例,所以它们可以产生分类器,其性能要比完全监督的学习算法好,因为完全监督的学习算法只有少量的标记数据可用于训练。半监督学习的原理是,可以利用未标记数据中的内在知识来增强仅使用标记实例的监督模型的预测能力,从而为半监督学习提供潜在的优势。从少量标记数据中学到的模型参数的监督分类器可以通过未标记的数据来转向更真实的分布(与测试数据的分布更相似)。

[0596] 生物信息学中普遍存在的另一个挑战是数据不平衡问题。当要预测的类别在数据中的代表性数量不足时,就会出现数据不平衡现象,这是因为属于该类别的实例很少(值得注意的情况)或难以获取。具有讽刺意味的是,那些少数的类型通常是最重要的学习对象,因为它们可能与特殊情况相关。

[0597] 处理不平衡的数据分布的一种算法是基于分类器集合的算法。数量有限的标记数据自然会使得分类器较弱,但是较弱分类器的集合往往会超过任何单个成分的分类器的性能。此外,分类器的集合通常会通过一个因素来提高从单一分类器获得的预测准确度,该因素可验证与学习多个模型相关的工作量和成本。直观而言,汇总几个分类器可引起更好的过拟合控制,因为将各个分类器的高变异性进行平均时,也可平均分类器的过拟合。

[0598] 由于缺乏足够大小的可靠标记致病性变体的数据集,我们采用了半监督学习策略。尽管ClinVar数据库有超过300,000个条目,但除去不确定性显著的变体后,仅剩下约42,000个错义变体,它们致病性解释没有冲突。

[0599] 系统评价还发现,这些条目通常没有足够的临床证据来支持其注释的致病性。而且,人类的精选数据库中的大多数变体往往都位于极少数的基因中,从而使它们与良性训练数据集中的变体不匹配,这些变体使用人类常见的变体或黑猩猩-人类固定代替物在全基因组范围内进行确定。鉴于确定的数据集的差异,以人类精选的变体作为致病性集,以全基因组常用变体作为良性集来训练监督学习模型可能会带来重大的偏差。

[0600] 我们训练了深度学习网络,以区分经过仔细匹配后以消除偏差的一组标记的良性变体和未标记的变体。根据一种实施情况,385,236个标记的良性变体包括ExAC/gnomAD数据库中的人类常见变体($>0.1\%$ 等位基因频率)和六种非人类灵长类动物的变体。

[0601] 我们采样了一组未标记变体,需要与三核苷酸背景上的良性变体匹配(以控制突变率,遗传漂移和基因转换),并调整可比性和序列覆盖对变体确认的影响。由于未标记变体的数量大大超过标记的良性变体,因此我们通过训练八个使用相同标记的良性变体的模型和八个随机采样的未标记变体的模型,并取它们的平均值,从而获得一致性的预测结果。

[0602] 选择半监督学习的动机是,人类精选的变体数据库不可靠且掺杂的,尤其是缺乏可靠的致病性变体。我们从gnomAD和灵长类动物的变体中获得了一组可靠的良性变体。对于致病性变体,我们采用迭代平衡采样方法从一组未知变体(VUS变体,无注释临床意义)中对致病性变体进行采样。

[0603] 为了减少采样偏差,我们训练了八个模型的集合,这些模型使用相同的良性训练变体集和八个不同的致病性变体集。最初,我们随机采样未知变体来代表致病性变体。接下

来,使用模型集以迭代方式,对一组未参与先前训练周期的未知变体进行评分。然后获得得分最高的致病性变体,以替换先前周期中5%的随机未知变体。请注意,我们保留的得分最高的致病性变体比需要保留的多25%,因此我们可以采样八组不同得分的致病性变体,以代替未知变体,这增加了八个模型的随机性。然后形成新的致病性训练集,并执行新的训练周期。重复此过程,直到初始随机采样的未知变体全部被集合模型预测的高置信度致病性变体所代替为止。图42说明了迭代均衡采样过程。

[0604] 平衡良性和未知训练集

[0605] 与良性变体匹配的未知变体的采样方案可减少我们模型训练的偏差。当对未知变体进行随机采样时,深度学习模型通常会提取有偏见的信息并给出简单的解决方案。例如,如果在未知变体中氨基酸替代K→M发生的频率比良性变体高,则深度学习模型往往会将K→M的替代分类为致病性。因此,重要的是平衡两个训练集之间的氨基酸替换的分布。

[0606] 诸如CpG转换之类的较高突变类型在常见良性变体中具有巨大的表示偏差。来自其他灵长类动物的直系同源变体也遵循人类突变率,这意味着在整个良性训练集中高度突变类型的富集。如果未知变体的采样过程未得到很好的控制和平衡,则与诸如易位或非CpG转换等代表性较小的类别相比,深度学习模型更倾向于将CpG转换归类为良性。

[0607] 为了防止深度学习模型收敛到一个微不足道的非生物学解决方案,我们考虑平衡良性变体和未知变体的三核苷酸背景。三核苷酸由变体之前的碱基,变体的参考碱基和变体之后的碱基形成。并且变体的参考碱基可以改变为其他三个核苷酸。总共有64x3个三核苷酸背景。

[0608] 迭代均衡采样

[0609] 第1周期

[0610] 我们对未知变体进行了采样,以匹配每个三核苷酸背景的良性变体的确切数目。换句话说,在第一个周期中,我们从变体的三个核苷酸背景反映了良性和致病性训练集。这种采样方法背后的事实就是,在良性组和未知组之间,突变率相同的变体,其表示方法也相同,这样可以防止模型基于突变率收敛到一个平凡解。

[0611] 第2-20周期

[0612] 对于第2周期,我们应用了第1周期的训练模型对一组在第1周期未涉及的未知变体进行评分,并用预测的致病性最高的变体替换了5%的未知变体。该组变体合完全是由模型生成的,因此我们对该组中的三核苷酸背景未进行任何平衡。对训练所需的其余95%的未知变体进行采样,作为良性变体中每种三核苷酸背景计数的95%。

[0613] 客观情况是,由于第1周期使用完全匹配的训练集,因此生成的预测的致病性最高的变体没有任何突变率偏差。因此,无需考虑该组中的任何偏差。其余95%的数据仍受到三核苷酸背景突变率的控制,以防止模型收敛到一个平凡解。

[0614] 对于每个周期,被替换的未知变体的百分比增加5%。对于第3周期,我们用第3周期模型中预测的致病性最高的变体替换了5%的未知变体。通过累积,致病性变体的比例增加到10%,在三核苷酸背景反映的未知变体减少到90%。其余周期的采样过程相似。

[0615] 第21周期

[0616] 对于周期21(最后一个周期),整个致病性训练集完全由深度学习模型预测的致病性最高的变体组成。由于我们已经明确控制了每个周期的突变率偏差,因此致病性变体可

以用作训练数据,这是非常可靠的,并且不受突变率偏差的影响。因此,训练的最后一个周期产生了最终的深度学习模型,用于致病性预测。

[0617] 匹配标记的良性训练集和未标记的训练集

[0618] 未标记变体的均衡采样对于消除与变体的有害性无关的偏差至关重要。当没有适当控制混淆效果的情况下,深度学习很容易会因疏忽引入偏差,影响类别之间的区分。人类常见的变体往往富含来自高突变类型的变体,例如CpG启动区上的变体。同样,灵长类动物的多态性也遵循人类的突变率,这意味着在整个良性训练集中存在高突变类型的变体的富集。如果未标记变体的采样过程未得到很好的控制和平衡,则深度学习网络倾向于依靠突变率偏差对变体进行分类,因此与代表性较少的类型(例如易位或作非CpG转换)相比,他们更有可能将CpG转换归类为良性。我们在96种三核苷酸背景的每一个中采样了与标记的良性变体数量完全相同的未标记变体(前面讨论过)。

[0619] 当将未标记变体与标记的良性数据集中的灵长类动物变体匹配时,我们不允许从没有进行过灵长类物种多序列比对的人类基因组区域中选择变体,这是因为不能在那个位点识别那种灵长类动物的变体。

[0620] 在96种三核苷酸背景中,我们修正了灵长类动物变体的测序覆盖率。由于测序的人类种群众多,因此经常观察到人类种群中常见的变体,这样,即使在测序覆盖率较低的区域也能很好地确定它们。对于灵长类动物变体而言,情况并非如此,因为只有少数个体进行了测序。根据ExAC/gnomAD外显子组的测序覆盖范围,我们将基因组分为10个区。对于每个区,我们测量了标记的良性数据集中的灵长类动物变体相对未标记的数据集中的比例。我们使用线性回归,仅基于测序的覆盖范围来计算一个灵长类动物变体在标记的良性数据集中的概率(图24)。当选择未标记变体以匹配标记的良性数据集中的灵长类动物变体时,我们使用回归系数根据在该位点的测序覆盖率对变体采样的概率进行了加权。

[0621] 良性变体和未知变体的产生

[0622] 人类种群中常见的变体

[0623] 最近的研究表明,人类种群中常见的变体通常是良性的。根据一种实施方式,gnomAD提供了90,958个在规范编码区域内次要等位基因频率(MAF) $\geq 0.1\%$ 的非同义SNP。那些通过过滤器的变体被保留,不包括插入缺失。删除了起始密码子或终止密码子中出现的变体以及蛋白质截短的变体。根据一种实施方式,通过检查亚种群,每个亚种群中MAF $\geq 0.1\%$ 的错义变体的总数增加到245,360。这些变体构成良性变体训练集的一部分。

[0624] 类人猿的常见多态性

[0625] 由于已知编码区是高度保守的,因此可以直接假设多态性是否以高频率隔离在类人猿种群中,这也可能对人类拟合产生轻微影响。来自类人猿基因组计划和其他研究中的倭黑猩猩,黑猩猩,大猩猩和红毛猩猩的多态性数据可以和来自dbSNP中的恒河猴和狨猴的SNP合并在一起。

[0626] 未知变体的生成

[0627] 通过将每个碱基位置的核苷酸替换为其他三个核苷酸,可以从规范编码区的每个碱基位置生成所有可能的变体。当新的密码子形成,引起该位置氨基酸的潜在变化。同义突变将会被过滤。

[0628] 在gnomAD数据集中观察到的变体将会被删除。删除了起始密码子或终止密码子中

出现的变体以及形成了终止密码子的变体。对于具有多个基因注释的SNP,选择规范基因注释来表示SNP的注释。根据一种实施方式,总共生成了68,258,623个未知变体。

[0629] 变体的过滤

[0630] 人类基因组的某些区域难以比对解读。将这些区域包括在内会对训练和测试数据集造成混淆的影响。例如,在高选择压力下的区域倾向于具有有限数量的多态性。然而,难以测序的区域也具有较少的多态性。为避免对我们的模型造成混淆,我们从那些未通过gnomAD测序的基因中删除了变体。

[0631] 通常,良性变体是在良好测序的区域发现的,这些区域在多个物种中趋于保守。未知变体是在整个基因组中随机抽样的,其中包括一些覆盖较差的区域。这就导致了良性和未知集之间的确定性的偏差。为了减少偏差,我们在gnomAD中过滤了解释深度小于10的变体。我们还过滤了所有哺乳动物物种的侧翼序列比对中缺失数据超过10%的所有变体。

[0632] 验证和测试数据

[0633] 为了验证和测试致病性模型,根据一种实施方式,我们从一大批良性变体中随机抽取了两组分别为10,000个的良性变体进行了验证和测试。良性变体的其余部分用于训练深度学习模型。这些变体是从直系同源的灵长类动物变体中专门取样的,以确保方法之间的公平比较,就如针对人类常见变体进行训练时的那些方法一样。根据一种实施方式,我们还分别随机抽取了两组分别为10,000个未知变体进行了验证和测试。我们确保分别将192个三核苷酸背景中的未知变体的数量与用于验证和测试的良性变体的数量相匹配。

[0634] 我们使用自闭症或解密发展障碍(DDD)患儿及其未患病兄弟姐妹的新生变体在临床环境中评价了多种方法的性能。根据一种实施方式,总共有来自DDD病例的3821个错义新生变体和来自自闭症病例的2736个错义新生变体。根据一种实施方式,有1231个未患病的兄弟姐妹有错义新生变体。

[0635] 深度学习网络架构

[0636] 致病性预测网络通过二级结构和溶剂可及性网络接收五个直接输入和两个间接输入。这五个直接输入是长度为51的氨基酸序列×深度为20的序列(编码20个不同的氨基酸),包括不带变体(1a)的人类参考氨基酸序列,带变体的(1b)中替代的人类替代氨基酸序列,灵长类物种的多序列比对的PFM(1c),哺乳动物物种的多序列比对的PFM(1d)和更远古的脊椎动物的多序列比对的PFM种类(1e)。二级结构和溶剂可及性网络分别从多序列比对(1f)和(1g)接收PFM作为输入,并将其输出,作为输入到主要致病性预测网络,作为其输入。二级结构和溶剂可及性网络已在Protein DataBank的已知蛋白晶体结构上进行了预训练,并在致病性模型训练期间允许反向传播。

[0637] 五个直接输入通道通过40个带有线性激活的内核的上采样卷积层。人类参考氨基酸序列(1a)与来自灵长类,哺乳动物和脊椎动物多序列比对的PFM合并(合并1a)。同样地,人类替代氨基酸序列(1b)与灵长类,哺乳动物和脊椎动物多序列比对的PFM合并(合并1b)。这样会创建两条平行的轨道,一条用于参考序列,另一条带有替换序列,其中的变体已被替换。

[0638] 参考通道和备用通道(合并1a和合并1b)的合并特征映射图通过一系列的六个残差块(层2a至7a,合并2a和层2b至7b,合并2b)传递。残差块的输出(合并2a和合并2b)连接在一起,形成大小为(51,80)(合并3a,合并3b)的特征映射图,该特征映射图将参考通道和备

用通道中的数据完全混合。接下来,数据具有两条并行通过网络的路径,或者通过一系列的六个残差块,其中每个残差块包含两个卷积层,如第2.1节中所定义(合并3到9,第9到46层,不包括第21,34层),或通过跳跃连接,它们通过1D卷积后分别连接每两个残差块的输出(第21层,第37层,第47层)。最后,合并后激活(合并10)被馈送到另一个残差块(第48至53层,合并11)。来自合并11的激活被分配给具有过滤大小为1和S形激活的1D卷积(第54层),然后通过全局最大池化层,该池化层将选择一个代表网络对变体致病性预测的值。该模型的示意图如图3和补充表16(图4A,4B和4C)所示。

[0639] 模型概述

[0640] 我们开发了半监督的深卷积神经网络(CNN)模型来预测变体的致病性。该模型的输入特征包括变体侧翼的蛋白质序列和保守图谱,以及特定基因区域中错义变体的损耗。我们还通过深度学习模型预测了变体导致的二级结构和溶剂可及性的变化,并将其整合到我们的致病性预测模型中。为了训练该模型,我们从人类亚种群的常见变体中产生了良性变体,并且从灵长类动物中产生了直系同源变体。但是,我们仍然缺乏致病性变体的可靠来源。我们最初使用良性和未知变体训练该模型,然后使用半监督迭代平衡采样(IBS)算法逐步将未知变体替换为一组具有高置信度的致病性变体。最后,我们证明了我们的模型在区分导致人类发育障碍和良性疾病的新生变体方面优于现有方法。

[0641] 残差块的采用

[0642] 图17示出了残差块。我们的致病性预测的深度学习模型和用于预测二级结构和溶剂可及性的深度学习模型均采用了残差块的定义,该定义最早在本文件中进行了说明。残差块的结构如下图所示。首先对输入层进行批归一化,然后对非线性激活“ReLU”进行归一化。激活后通过1D卷积层。来自1D卷积层的中间层输出再次进行批归一化并激活ReLU,然后再执行另一个1D卷积层。在第二个1D卷积结束时,我们将其输出与原始输入合并。在这样的架构中,输入保持其原始状态,而剩余的连接则不受模型的非线性激活的影响。

[0643] 带孔/空洞卷积允许具有很少训练参数的大感受野。带孔/空洞卷积是一种卷积,其中通过以一定步长跳过输入值(也称为带孔卷积率或膨胀系数),从而将内核应用于大于其长度的区域。带孔/空洞卷积增加了卷积过滤器/内核元素之间的间隔,以便在执行卷积操作时考虑以较大的间隔相邻输入条目(例如核苷酸,氨基酸)。这样可以在输入中包含远程的背景相关性。带孔卷积保留部分卷积计算,以便在处理相邻核苷酸时重新使用。

[0644] 我们模型的新颖性

[0645] 我们的方法在三个方面不同于现有的预测变体致病性的方法。首先,我们的方法采用了一种新型的半监督深度卷积神经网络架构。其次,可靠的良性变体是从gnomAD和灵长类动物变体的人类常见变体中获得的,而高置信度致病性训练集是通过迭代均衡采样和训练生成的,以避免使用相同的人类精选的变体数据库来对模型进行循环训练和测试。第三,将二级结构和溶剂可及性的深度学习模型整合到我们的致病性模型的架构中。从结构和溶剂模型获得的信息不限于特定氨基酸残基的标记预测。而且,从结构和溶剂模型中删除了读出层,并将预训练的模型与致病性模型合并。在训练致病性模型时,结构和溶剂预训练层也会反向传播,以将误差减少到最小。这样可以使预训练的结构和溶剂模型能专用于致病性预测问题。

[0646] 训练二级结构与溶剂可及性模型

[0647] 数据准备

[0648] 我们训练了深卷积神经网络来预测蛋白质的三态二级结构与三态溶剂可及性。PDB的蛋白质注释用于训练模型。根据一个实施方案,去除与序列谱有25%以上相似性的序列。总计,根据一个实施方案,6,293个蛋白质序列用于训练,392个蛋白质序列用于验证,499个蛋白质序列测试。

[0649] 通过运行E值阈值为0.001和迭代次数为3次的PSI-BLAST来搜索UniRef90,生成蛋白质的位置特异性评分矩阵(PSSM)保守谱。将任意未知的氨基酸及其二级结构设为空白。我们还对所有人类基因进行了具有相似参数设置的PSI-BLAST以收集它们的PSSM保守谱。这些矩阵用于将结构模型整合到致病性预测中。然后将蛋白质序列的氨基酸转化为一位有效编码向量。将蛋白质序列和PSSM矩阵更改为Lx20矩阵,其中L为蛋白质的长度。二级结构的三个预测的标签包括螺旋(H)、 β 折叠(B)和卷曲(C)。溶剂可及性的三个标签包括埋藏的(B)、中间的(I)和暴露的(E)。一个标签对应一个氨基酸残基。标签被编码为维度为3的一位有效编码向量。

[0650] 模型架构与训练

[0651] 我们训练了两个端对端深卷积神经网络模型来分别预测蛋白质的三态二级结构和三态溶剂可及性。这两个模型具有相似的配置,包括两个输入通道,一个输入通道用于蛋白质序列,另一个输入通道用于蛋白质保守谱。每个输入通道的维度为Lx20,其中L表示蛋白质的长度。

[0652] 每个输入通道都通过具有40个内核和线性激活值的1D卷积层(层1a与1b)。该层用于将输入维度从20上取样至40。注意,整个模型的所有其他层使用40个内核。通过对40个维度的每个维度进行求和(即,合并模式=“求和”)来将两个层(1a与1b)激活值合并在一起。合并节点的输出通过一层1D卷积(第2层),然后进行线性激活。

[0653] 层2的激活值通过一系列如上定义的9个残差块(层3至11)。层3的激活值反馈到层4,层4的激活值反馈到层5,以此类推。还存在可直接将每第3个残差块的输出相加(层5、8和11)的跳越连接。然后将合并的激活值反馈至2个具有ReLU激活值的1D卷积(层12和13)中。将层13的激活值赋予softmax读出层。该softmax计算给定输入的三个类别输出的概率。

[0654] 对于最佳二级结构模型,1D卷积的空洞率为1。对于溶剂可及性模型,最后的3个残差块(第9、10和11层)的空洞率为2,以增加内核的覆盖率。蛋白质的二级结构强烈依赖于紧邻氨基酸之间的相互作用。因此,具有较高内核覆盖范围的模型稍微提高了性能。另一方面,溶剂可及性受到氨基酸之间的大范围相互作用的影响。因而,对于采用空洞卷积的具有高内核覆盖度率的模型,其准确度比短覆盖率模型高2%以上。

[0655] 下表提供了根据一个实施方案的关于三态二级结构预测模型的每一层的激活值和参数的详细信息

	层	类型	内核数, 窗口尺寸	塑造	空洞率	激活
[0656]	输入序列(层 1a)	1D 卷积	40,1	(L,40)	1	线性
	输入 PSSM (层 1b)	1D 卷积	40,1	(L,40)	1	线性
	合并序列 + PSSM	合并 (模式= 连接)	-	(L,80)	-	-
[0657]	层 2	1D 卷积	40,5	(L,40)	1	线性
	层 3	1D 卷积	40,5	(L,40)	1	ReLU
	层 4	1D 卷积	40,5	(L,40)	1	ReLU
	层 5	1D 卷积	40,5	(L,40)	1	ReLU
	层 6	1D 卷积	40,5	(L,40)	1	ReLU
	层 7	1D 卷积	40,5	(L,40)	1	ReLU
	层 8	1D 卷积	40,5	(L,40)	1	ReLU
	层 9	1D 卷积	40,5	(L,40)	1	ReLU
	层 10	1D 卷积	40,5	(L,40)	1	ReLU
	层 11	1D 卷积	40,5	(L,40)	1	ReLU
	合并激活值	合并 -层 5、8 和 11.模式='求和'	-	(L,40)	-	-
	层 12	1D 卷积	40,5	(L,40)	1	ReLU
	层 13	1D 卷积	40,5	(L,40)	1	ReLU
	输出层	1D 卷积	1,3	(L,3)	-	Softmax

[0658] 根据一个实施方案,下表示出了溶剂可及性的详细信息

层	类型	内核数, 窗口尺寸	塑造	空洞率	激活
输入序列(层 1a)	1D 卷积	40,1	(L,40)	1	线性
输入 PSSM (层 1b)	1D 卷积	40,1	(L,40)	1	线性
[0659] 合并序列 + PSSM	合并 (模式= 连接)	-	(L,80)	-	-
层 2	1D 卷积	40,5	(L,40)	1	线性
层 3	1D 卷积	40,5	(L,40)	1	ReLU
层 4	1D 卷积	40,5	(L,40)	1	ReLU
层 5	1D 卷积	40,5	(L,40)	1	ReLU

	层 6	1D 卷积	40,5	(L,40)	1	ReLU
	层 7	1D 卷积	40,5	(L,40)	1	ReLU
	层 8	1D 卷积	40,5	(L,40)	1	ReLU
	层 9	1D 卷积	40,5	(L,40)	2	ReLU
	层 10	1D 卷积	40,5	(L,40)	2	ReLU
[0660]	层 11	1D 卷积	40,5	(L,40)	2	ReLU
	合并激活值	合并 -层 5、8 和 11.模式 =‘求和’	-	(L,40)	-	-
	层 12	1D 卷积	40,5	(L,40)	1	ReLU
	层 13	1D 卷积	40,5	(L,40)	1	ReLU
	输出层	1D 卷积	1,3	(L,3)	-	Softmax

[0661] 特异性氨基酸残基的二级结构类别由最大的预测softmax概率确定。采用ADAM优化器以累积分类交叉熵损失函数来训练整个蛋白质序列的模型，优化反向传播算法。

[0662] 三态二级结构预测模型的最佳测试准确度为80.32%，其与由在类似的训练数据集上的DeepCNF模型预测的最佳准确度相似。

[0663] 三态溶剂可及性预测模型的最佳测试准确度为64.83%，与由在类似训练数据集上的RaptorX预测的当前最佳准确度相似。

[0664] 我们将预训练的三态二级结构与溶剂可及性预测模型整合为我们如下所述的致病性预测模型。

[0665] 训练模型来预测变体的致病性

[0666] 致病性预测模型的输入特性

[0667] 如上所述,对于致病性预测问题,存在一个良性变体训练集和一个未知变体训练集,以用于训练致病性模型。对于每个变体,我们准备了以下输入特性以馈送到我们的模型中。

[0668] 每个变体的第一输入特性为其侧翼氨基酸序列,即由hg19的参考序列获得的变体的每一侧上的25个氨基酸,以提供变体的序列上下文的深度学习模型。总的来说,该侧翼参考序列的长度为51个氨基酸。

[0669] 第二个特性为形成变体的替代氨基酸。我们不直接提供参考替代氨基酸对,而是提供模型的替代侧翼序列。该替代侧翼序列与第一特性中的参考侧翼序列相同,除了序列的中间位置含有替代氨基酸,而不是参考氨基酸。

[0670] 然后将这两个序列都转化为长度为 51×20 的一位有效编码的向量,其中每个氨基酸由20个0或1的向量表示。

[0671] 然后通过99种脊椎动物变体的多序列比对(MSA)生成3个位置权重矩阵(PWM),其中1个PWM用于12种灵长类动物,1个PWM用于47种哺乳动物(不包括灵长类动物),1个PWM用于40种脊椎动物(不包括灵长类动物和哺乳动物)。每个PWM的尺寸为 $L \times 20$,其中L为变体周围侧翼序列的长度(在这种情况下,L表示51个氨基酸)。它包含每个类别种属中看到的氨基酸计数。

[0672] 我们还生成了来自psi blast的51个氨基酸的变体-侧翼序列的PSSM矩阵。这用于将三态二级结构和溶剂可及性预测模型整合以进行致病性预测。

[0673] 我们用参考序列(输入1)、交替序列(输入2)、用于灵长类动物(输入3)、哺乳动物(输入4)、脊椎动物(输入5)和三态二级结构信息的PWM矩阵训练致病性模型和溶剂可及性模型。

[0674] 深度学习模型训练

[0675] 图19为提供深度学习模型工作流程综述的框图。致病性训练模型包括五个直接输入和四个间接输入。五个直接输入特征包括参考序列(1a)、替代序列(1b)、灵长类动物保守序列(1c)、哺乳动物保守序列(1d)和脊椎动物保守序列(1e)。间接输入包括基于参考序列的二级结构(1f)、基于替代序列的二级结构(1g)、基于参考序列的溶剂可及性(1h)和基于替代序列的溶剂可及性(1i)。

[0676] 对于间接输入1f和1g,我们加载了二级结构预测模型的预训练层,不包括softmax层。对于输入1f,预训练层基于变体的人类参考序列以及变体的由PSI-BLAST生成的PSSM。同样地,对于输入1g,二级结构预测模型的预训练层以人类替代序列以及PSSM矩阵作为输入。输入1h和1i分别对应于包含变体的参考和替代序列的溶剂可及性信息的相似预训练通道。

[0677] 五个直接输入通道通过具有线性激活值的40个内核的上取样卷积层。层1a、1c和1h与40个特征维度相加的值合并,产生层2a。换言之,参考序列的特征映射图与三种类型的保守特征映射图合并。相似的,层1b、1c、1d和1e与40个特征维度相加的值合并,生成层2b,即,替代序列的特征与三种类型的保守特征合并。

[0678] 使用ReLU激活值对层2a和2b进行批归一化,并且每层都通过滤波器大小为40(3a和3b)的1D卷积层。层3a和3b的输出与1f、1g、1h和1i合并,其中特征映射图彼此关联。换言之

之,具有保守谱的参考序列和具有保守谱的替代序列的特征映射图与参考和替代序列的二级结构特征映射图和参考和替代序列的溶剂可及性特征映射图(层4)合并。

[0679] 层4的输出通过六个残差块(层5、6、7、8、9、10)。1D卷积的最后三个残差块的空洞率为2,以为内核提供更高的覆盖率。层10的输出通过滤波器尺寸为1且激活S型的1D卷积(层11)。层11的输出通过全局maxpool,该全局maxpool会选择一个变体的单一值。此值表示变体的致病性。致病性预测模型的一个实施方案的详细信息如下表所示。

[0680]	层	类型	内核数, 窗口尺寸	塑造	空洞率	激活
	参考序列 (1a)	1D 卷积	40,1	(51,40)	1	线性
	替代序列 (1b)	1D 卷积	40,1	(51,40)	1	线性
	灵长类动物 保守序列 (1c)	1D 卷积	40,1	(51,40)	1	线性
[0681]	哺乳动物保 守序列 (1d)	1D 卷积	40,1	(51,40)	1	线性
	脊椎动物保 守序列(1e)	1D 卷积	40,1	(51, 40)	1	线性
	基于二级结 构的参考序 列(1f)	输入层	-	(51,40)	-	-
	基于二级结 构的替代序	输入层	-	(51,40)	-	-

列(1g)						
[0682]	基于溶剂可及性的参考序列 (1h)	输入层	-	(51,40)	-	-
	基于溶剂可及性的替代序列 (1i)	输入层	-	(51,40)	-	-
	参考序列合并 (2a)	合并 (模式=求和) (1a,1c,1d,1e)	-	(51,40)	-	-
	替代序列合并 (2b)	合并 (模式=求和) (1b,1c,1d,1e)	-	(51,40)	-	-
	3a	卷积 1D (2a)	40,5	(51,40)	1	ReLU
	3b	卷积 1D (2b)	40,5	(51,40)	1	ReLU
	4	合并 (模式=连接)	-	(51,240)	-	-

	(3a,3b,1f,1g,1h,1i)				
5	1D 卷积	40,5	(51,40)	1	ReLU
6	1D 卷积	40,5	(51,40)	1	ReLU
7	1D 卷积	40,5	(51,40)	1	ReLU
8	1D 卷积	40,5	(51,40)	1	ReLU
[0683]					
9	1D 卷积	40,5	(51,40)	2	ReLU
10	1D 卷积	40,5	(51,40)	2	ReLU
11	1D 卷积	40,1	(51,1)	2	S 形
12	全局 max pooling	-	1	-	-
输出层	激活层	-	1	-	S 形

[0684] 集合

[0685] 在一个实施方案中,对于我们的方法的每个周期,我们运行了八个不同的模型,这些模型在相同的良性数据集和八个不同的未知数据集进行训练,并平均八个模型中的评估数据集的预测。当模型中存在多个未知变体的随机取样集时,可以减小和控制取样偏差。

[0686] 此外,采用集成方法可提高我们评估数据集上的模型的性能。CADD使用10个模型的集合并获得所有10个模型的平均得分来对变体评分。此处我们尝试使用类似的集成方法。我们使用一个集合对结果进行基准测试,然后增加集合的数目来评估性能增益。注意,每个集合有8个模型,这8个模型在相同的良性数据集和8个不同的未知数据集上训练。对于不同的集合,随机数产生器的种子值是不同的,使得随机变体集被彼此不同地绘制。

[0687] 下表示出了根据一个实施方案的详细结果。

集合数		-DDD 数据集	-DDD 数据集上的 log(p 值) 上的 log(p 值) (集合的中值) (集合的平均 值)
[0688]	1	3.4e-27	3.4e-27
	5	2.74e-29	3.8e-29
	10	2.98e-29	2.55e-29
	15	4.06e-29	3.88e-29
	20	3.116e-29	3.05e-29
[0689]	25	3.77e-29	3.31e-29
	30	3.81e-29	3.34e-29

[0690] 与一个集合相比,使用DDD数据集评估时,5个集合和10个集合产生了更显著的p值。但增加集合数并不能进一步提高性能,表明集合的饱和度。集合用大量未知变体减少取样偏差。然而,我们还需要在良性和致病性类别之间匹配192个三核苷酸上下文,这大大限制了我们的取样空间,导致快速饱和。我们的结论是,集成方法显著提高了模型的性能,进一步丰富了我们模型的理解。

[0691] 训练致病性模型的早停法

[0692] 由于缺乏可靠的带注释的致病性变体样本,定义模型训练的停止标准是一项挑战。为了避免在模型评估中使用致病性变体,在一个实施方案中,我们使用了来自直系同源灵长类动物的10,000个良性验证变体和10,000个与未知变体匹配的三核苷酸上下文。在训练模型的每个时期(epoch)之后,我们评估了良性验证变体和未知验证变体。我们使用Wilcoxon秩和检验来评估两个验证变体集的概率分布的差异。

[0693] 随着模型区分良性变体与一组位置未知变体的能力的提高,测试的p值变得更加显著。如果在模型训练的任何五个连续时期,模型区分这两种分布的能力没有得到改善,我们就停止训练。

[0694] 早些时候,我们从训练中分离出两组10,000个保留灵长类动物变体,我们称之为验证集和测试集。我们使用10,000个保留灵长类动物变体的验证集和10,000个未标记的与三核苷酸上下文匹配的变体的测试集以用于评估模型训练期间的早停。在每个训练时期之后,我们评估了深度神经网络在标记的良性验证集和未标记的匹配对照中的变体之间进行区分的能力,从而使用Wilcoxon秩和检验来测量预测得分分布的差异。一旦在连续五个训练时期后没有观察到进一步的改善,我们就停止训练,以防止过度拟合。

[0695] 分类器性能的基准测试

[0696] 我们评估了两个版本的深度学习网络的分类准确度,一个版本仅使用普通人类变体进行训练,一个版本使用包含普通人类变体和灵长类动物变体的全良性标记数据集进行训练,此外,我们还评估了以下分类器:SIFT、PolyPhen-2、CADD、REVEL、M-CAP、LRT, MutationTaster, MutationAssessor, FATHMM, PROVEAN, VEST3, MetaSVM, MetaLR, MutPred, DANN, FATHMM-MKL_coding, Eigen, genoconyon和GERP++13, 32-48。为了获得其他分类器中每一个的得分,我们从dbNSFP 49 (<https://sites.google.com/site/jpopgen/dbNSFP>) 中下载了所有错义变体的得分,并在10,000个保留灵长类动物变体测试集和DDD病例与对照的denovo变体上对方法进行了基准测试。我们选择了主要论文中包含的SIFT、PolyPhen-2和CADD,因为它们是使用最广泛的方法,并且选择了REVEL,因为在不同的评价模式中,它是我们评价的20个现有分类器中最好的一个。图28A中提供了我们评估的所有分类器的性能。

[0697] 为了评估可用训练数据大小对深度学习网络性能的影响,我们在图6中的每个数据点对深度学习网络进行了训练,方法是从385,236个灵长类动物和常见人类变异体的标记良性训练集中随机取样。为了降低分类器性能中的随机噪声,我们执行该训练过程5次,每次使用初始参数权重的随机例示,并在图6中示出了10,000个保留灵长类动物变体和DDD病例与对照数据集的中值性能。碰巧的是,带有385,236个标记良性变体的完整数据集的中值分类器的性能略好于我们在DDD数据集上的其他论文中使用的中值分类器(通过Wilcoxon秩和检验, $P < 10^{-29}$ 而不是 $P < 10^{-28}$)。为了表明每个单个灵长类物种的变体有助于分类准确度,而每个单个哺乳动物物种的变体分类准确度较低,根据一个实施方案,我们使用包含83,546个人类变体和每个物种的恒定随机选择变体数的训练数据集来训练深度学习网络。根据一个实施方案,我们添加到训练集(23,380)的变体恒定数目是在错义变体(即倭黑猩猩)数目最低的物种中可用的变体总数。为了降低噪声,我们再次重复了五次训练过程,并报告了分类器的中值性能。

[0698] 模型评估

[0699] 在一个实施方案中,我们按照迭代平衡取样过程训练了21个周期的深度学习模型。我们进行了两种类型的评估来评估我们的分类器的性能。我们还将我们的模型与Polyphen2、SIFT和CADD在这两个指标上进行了比较,并评估了我们的模型在临床注释中的应用潜力。

[0700] 方法1:良性测试集准确度

[0701] 在一个实施方案中,我们通过使用八个不同训练模型的集合计算其预测概率来评估10,000个良性变体和未知变体。我们还通过上述其他现有方法得到它们预测的概率。

[0702] 然后,我们获得了评估中使用的每种方法的未知测试变体的预测概率中值。通过使用中值得分,我们发现得分高于或低于中值的良性变体的数目取决于每种方法使用的良

性和致病性变体的注释。SIFT, CADD和我们的方法将致病性变体标记为1, 良性变体标记为0。因此, 我们计算了得分低于中值的良性变体的数目。Polyphen使用相反的注释, 我们计算了中值以上的良性变体的数目。中值以上/以下评分的良性变体数除以良性变体总数的比率代表良性变体的预测准确度。

[0703] 良性准确度 = 高于 (低于*) 中值的良性变体总数 ÷ 良性变体总数

[0704] 这种评估方法背后的推理依赖于对gnomAD中变体的选择压力的分析。对于gnomAD中的单子, 错义变体与同义变体之比约为2.26:1。而对于gnomAD中常见的变体 (MAF > 0.1%), 错义与同义的比例约为1.06:1。这表明, 从一组随机未知变体中, 大约50%变体预期被自然选择清除, 剩下的50%趋于温和, 并且可能在种群中变得普遍。

方法 保留良性集的准确度

Polyphen 0.74

SIFT 0.69

[0705]

CADD 0.77

我们的模 0.85

型

[0706] 如上表所示, 我们的方法优于第二最佳方法CADD 8%以上。这表明我们的模型对良性变体的分类能力有了显著的提高。虽然这样的演示证明了我们模型的能力, 但是下面的方法2显示了我们在临床数据集上的模型对临床解释的有用性。

[0707] 方法2: 临床数据集评估

[0708] 在一个实施方案中, 我们在临床数据集上评估了这些致病性预测方法, 包括发育障碍 (DDD) 病例-对照数据集。DDD数据集包括来自受影响儿童的3,821个新生错义变体和来自未受影响兄弟姐妹的1,231个新生错义变体。我们假设, 受影响儿童的新生变体往往比未受影响的兄弟姐妹的新生变体更有害。

[0709] 由于临床测试数据集不能清楚地标记致病性变体, 所以我们使用两组新变体 (从受影响和未受影响) 之间的分离来评估这些方法的性能。我们应用Wilcoxon秩和检验来评估这两组新生变体集是如何很好地分离的。

方法 DDD 数据集上的 $\log_{10}(\text{p-值})$

Polyphen 15.02

[0710]

SIFT 13.52

CADD 13.83

DL 28.35

[0711] 根据上表,我们的半监督深度学习模型在区分受影响的新生变体集和未受影响的集方面表现得明显更好。这表明我们的模型比现有的方法更适合临床解释。这也验证了从基因组序列和保守谱中提取特征的一般方法优于基于人类保存数据集的手动构建特征。

[0712] 对10,000个灵长类动物变体的保留测试集的良好预测准确度

[0713] 我们在测试数据集中使用了10,000个保留灵长类动物变体来对深度学习网络以及其他20个分类器进行基准测试。由于不同的分类器具有广泛不同的得分分布,我们使用了10,000个随机选择的未标记的变体,这些变体与由三核苷酸上下文匹配的测试集来确定每个分类器的第50百分位阈值。为了确保方法之间的公平比较,我们根据10,000个保留灵长类动物变体测试集中的在该分类器的第50个百分位阈值处被分类为良性的变体得分对每个分类器进行基准测试。

[0714] 我们使用第50个百分位数来识别良性变体的理由是基于在ExAC/gnomAD数据集中观察到的错义变体的选择压力。对于发生在单等位基因频率的变体,错义:同义比为大约2.2:1,而对于常见变体(>0.1%等位基因频率),错义:同义比为大约1.06:1。这表明在正常等位基因频率下,大约有50%的错义变体有望被自然选择清除,剩下的50%足够温和以有通过遗传漂变在种群中普遍存在的潜力。

[0715] 对于每个分类器,显示了使用第50百分位阈值预测为良性的保留灵长类动物测试变体的分数(图28A和补充表17(图34))。对DDD研究的新生变体的分析

[0716] 我们根据其区分DDD受影响个体中的新生错义变体与未受影响兄弟姐妹对照中的新生错义变体的能力对分类方法进行了基准测试。对于每个分类器,我们报告了两个分布的预测得分之间的差异的Wilcoxon秩和检验的p值(图28B和28C以及补充表17(图34))。

[0717] 鉴于我们用于分析模型性能的两个指标源自不同的来源和方法,我们测试了分类器在两个不同指标上的性能是否相关。事实上,我们发现这两个指标是相关的,在保留灵长类测试集上的良性分类准确度与DDD病例和对照中新生错义变体的Wilcoxon秩和p值之间spearman $\rho=0.57$ ($P<0.01$)。这表明,保留灵长类动物测试集准确度与用于对分类器进行

基准测试的DDD病例与对照p值之间存在良好的一致性(图30A)。

[0718] 此外,我们还测试了深度学习网络是否有助于发现与疾病相关的基因。我们通过将观察到的新生突变数目在零突变模型下预期的数目来测试基因中新生突变的富集程度。

[0719] 我们检查了深度学习网络的性能,从而将所有错义新生突变的结果与得分大于0.803的错义突变的结果进行了比较。测试所有错义新生开始使用默认错义率,而测试过滤的错义新生开始使用从得分 >0.803 的位点计算的错义突变率。每个基因需要四个测试,一个测试蛋白质截短富集,一个测试蛋白质改变新生突变富集,这两个测试都只针对DDD群组,以及一个更大的神经发育三重测序群组元分析。用Fisher法结合编码序列中错义新生突变聚类试验(补充表20和21)对蛋白质改变的新生突变进行富集。每个基因的p值从四个测试中取最小值,全基因组显著性为 $p < 6.757 \times 10^{-7}$ ($\alpha = 0.05$ 18,500个基因,四个测试)。

[0720] 计算605个DDD相关基因内受体-算子曲线特征及分类准确度

[0721] 为了测试深度学习网络是否真的在同一基因内的致病性和良性变体之间进行区分,而不是有利于具有新生显性遗传模式的基因的致病性,我们在DDD群组中鉴定了一组605个与 p 值 <0.05 的神经发育疾病相关的基因(仅使用新生蛋白质截断变种计算)(补充表18)。我们报告了所有分类器的Wilcoxon秩和 p 值,它们能够在DDD和对照数据集中分离605个基因变体的概率分布(图28C和补充表19(图35))。

[0722] 在这组605个基因中,我们观察到新生错义变体的富集率是单靠突变率预期的三倍。这表明DDD受影响的患者中新生错义变体包括大约67%种致病性变体和33%种背景变体,而健康对照中新生错义变体主要由背景变体组成,除了不完全外显的实例。

[0723] 为了计算完美区分致病性和良性变体的分类器的最大可能AUC,我们考虑到在605个基因内的受影响个体中只有67%的新生错义变体是致病性的,其余的是背景。为了构建受体-算子特征曲线,我们将新生DDD变体的致病性分类视为真阳性识别,将健康对照中新生变体的致病性分类视为假阳性识别。因此,一个完美的分类器可以将DDD患者中67%的新生变体归类为真阳性,将DDD患者中33%的新生变体归类为假阳性,将对照中100%的新生变体归类为真阴性。受体-算子曲线的可视化将仅显示一个点,其真阳性率为67%,假阳性率为0%,通过直线连接到绘图的(0%,0%)和(100%,100%)角,从而产生最大AUC为0.837,其中完美区分了良性和致病突变(图30B和补充表19(图35))。

[0724] 我们通过估计组合的DDD和健康对照数据集中605个基因内致病性变体的预期得分来计算用于以二进制阈值分离致病性和良性变体的深度学习网络的分类准确度。由于DDD数据集包含379个新生变体,超过预期249个新生错义变体,而对照数据集包含65个新生变体,我们预计444个总变体中有249个致病变体(图22A)。我们为每个分类器选择阈值,根据这个期望比例将444个新错义变体分为良性或致病性类别,并将其作为二进制截止值来评估每个分类器的准确度。对于我们的深度学习模型,在 ≥ 0.803 的临界点处获得该阈值,真阳性率为65%,假阳性率为14%。为了计算在DDD个体中存在约33%背景变体时校正的分类准确度,我们假设作为背景的33%的新生DDD变体将以与我们在健康对照中观察到的相同的假阳性率进行分类。这对应于实际上是来自背景变量的假阳性的DDD数据集中真阳性分类事件的 $14\% \times 0.33 = 4.6\%$ 的。我们估计深度学习网络的校正真阳性率为 $(65\% - 4.6\%) / 67\% = 90\%$ 。我们报告了真阳性率和真阴性率的平均值,对于深度学习网络,这是88%(图30C和补充表19(图35))。由于神经发育障碍的不完全外显率很高,这一估计可能低

估了分类器的真实准确度。

[0725] ClinVar分类准确度

[0726] 大多数现有分类器都是在ClinVar上训练的;即使通过使用在ClinVar上进行训练的分类器的预测得分不直接在ClinVar上进行训练的分类器也可能受到受到影响。此外,常见的人类变体对于良性的ClinVar结果是高度丰富的,因为等位基因频率是将良性结果分配给变体的标准的一部分。

[0727] 我们设法将ClinVar数据集中的循环性最小化以使其适用于通过仅使用2017年添加的ClinVar变体来进行分析,因为其他分类方法在前些年已发布。即使在2017ClinVar变体中,我们排除了在ExAC中常见等位基因频率(>0.1%)或HGMD、LSDB或Uniprot中存在的任何变体。在过滤了所有这些变体并排除了不确定意义的变体和注释冲突的变体后,我们在ClinVar中留下了177个具有良性注释的变体和969个具有致病性注释的变体。

[0728] 我们使用深度学习网络和现有方法对所有ClinVar变体评分。我们根据该数据集内良性与致病性变体的比例来选择将ClinVar变体分离为良性或致病性类别的每个分类器的阈值,并将该阈值作为二进制截止来评估每个分类器的准确度。我们报道了每个分类器的真阳性率与假阳性率的平均值(图31A和31B)。分类器在ClinVar数据集上的性能与分类器在10,000个保留的灵长类动物变体上的分类准确度或DDD病例与对照数据集的Wilcoxon秩和p值上的性能没有显著相关性(图31A和32B)。

[0729] 我们假设,现有分类器可准确地模拟人类专家的行为,但是对于区分经验数据中的致病和良性突变,人类经验可能不是完全最优的。一个这样的例子是Grantham评分,它提供了距离度量来描述氨基酸取代的相似性或相异性。我们在完整的ClinVar数据集内计算了致病性和良性变体的平均Grantham评分(~42,000个变体),并将其与605个基因中DDD受影响和未受影响个体的新生变体的平均Grantham评分进行了比较。为了纠正DDD受影响个体中约33%的背景变体,我们将DDD病例与对照例之间Grantham评分的差异增加了50%,这仍然小于ClinVar中致病性和良性变体的差异。一种可能是,人类专家过于看重容易测量的指标,如氨基酸取代距离,而低估了蛋白质结构等因素,这些因素对人类专家来说更难量化。

[0730] 解释深度学习模型

[0731] 理解机器学习算法解决问题的方法通常是困难的。我们将深度学习网络的初始层可视化,以了解其已学会提取的特征,以便预测变体的致病性。我们计算了预处理的三态二级结构预测模型的前三层(两个上取样层后的第一卷积层)中不同氨基酸的相关系数,并表明卷积层的权值学习特征与BLOSUM62矩阵或Grantham距离非常相似。

[0732] 为了计算不同氨基酸之间的相关系数,我们从二级结构模型中第一卷积层的权重开始,然后是三个上取样层(层1a、1b和1c)。我们在三层之间进行矩阵乘法,得到一个维度为(20,5,40)的矩阵,其中20是氨基酸的数目,5是卷积层的窗口尺寸,40是内核的数目。我们通过将最后两个维度展平来更改矩阵以获得维度(20,200),得到一个矩阵,在该矩阵中,20个氨基酸中的每一个氨基酸上的操作权重被表示为一个200长度的向量。我们计算了这20种氨基酸之间的相关矩阵。由于每个维度代表每个氨基酸,通过计算相关系数矩阵,我们计算氨基酸之间的相关性,以及它们在深度学习网络中的相似程度,这基于它从训练数据中获得的信息。相关系数矩阵的可视化如图27所示(按BLOSUM62矩阵顺序排序的氨基酸),

并显示两个显著的簇,包括疏水性氨基酸(蛋氨酸、异亮氨酸、亮氨酸、缬氨酸、苯丙氨酸、酪氨酸、色氨酸)和亲水性氨基酸(天冬酰胺、天冬氨酸,谷氨酸、谷氨酰胺、精氨酸和赖氨酸)。这些初始层的输出成为后一层的输入,使得深度学习网络能够构建数据的日益复杂的层次表示。

[0733] 为了说明神经网络在其预测中所使用的氨基酸序列的窗口,我们对随机选择的5000个变体中和周围的每个位置进行扰动,观察其对该变体的预测PrimateAI学会评分的影响(图25B)。我们系统地调零了变体周围每个氨基酸位置(-25到+25)处的输入,测量了神经网络的预测变体致病性变化,并绘制了5000个变体变化的平均绝对值。变体附近的氨基酸影响最大,大致呈对称分布,随着与变体距离的增加而逐渐减少。重要的是,该模型的预测不仅基于变体所在位置的氨基酸,而且通过使用更宽窗口中的信息来识别蛋白质基序。与蛋白质亚结构域相对紧凑的尺寸一致,我们经验性地观察到,将窗口的尺寸扩展到51个以上的氨基酸并不能进一步提高准确度。

[0734] 为了评估深度学习分类器对比对的敏感性,我们测试了比对深度对变体分类准确度的影响,如下所示。我们根据比对中物种的数目将数据分成五个区段,并评估每个区段中网络的准确度(图57)。我们发现,网络在将一组保留良性突变与随机选择的与三核苷酸上下文匹配的突变(如图21D所示,但对每个区段分别执行)分离的准确度在前三个区段最强,在后两个区段明显较弱。99种脊椎动物的多物种比对包括11种非人类灵长类动物、50种哺乳动物和38种脊椎动物,底部的两个区段表示具有来自其他非灵长类哺乳动物的稀疏比对信息的蛋白质。当比对信息遍布灵长类动物和哺乳动物时,深度学习网络是鲁棒和准确的,而来自更遥远脊椎动物的保守信息就不那么重要了。

[0735] 规范编码区的定义

[0736] 为了定义规范编码区,从UCSC基因组浏览器下载了包含人类的用于编码DNA序列(CDS)区(knownCanonical.exonNuc.fa.gz)的99个脊椎动物基因组的多重比对。对于人类来说,外显子的坐标是在hg19的基础上建立的。合并外显子,形成基因。保留常染色体和chrX上的基因。去除非同源基因,从NCBI <ftp://ftp.NCBI.nih.gov/pub/HomoloGene/current/HomoloGene.data>中下载同源基因列表。对于具有多个基因注释的SNP,选择最长的转录本来表示SNP的注释。

[0737] 人类、类人猿和哺乳动物多态性数据

[0738] 我们从最近一项大规模研究中下载了人类外显子组多态性数据,即基因组聚集数据库(gnomAD),该数据库从全世界8个亚种群中收集了123,136个个体的全外显子组测序数据。然后,我们提取通过过滤器并属于规范编码区域的变体。

[0739] 大猩猩基因组测序项目提供了24只黑猩猩、13只倭黑猩猩、27只大猩猩和10只猩猩(包括5只苏门答腊猩猩和5只博尔纳猩猩)的全基因组测序数据。对黑猩猩和倭黑猩猩的研究提供了另外25只类人猿的WGS。当所有的测序数据被映射到hg19时,我们下载了源自这些研究的VCF文件,并直接提取了规范编码区内的变体。

[0740] 为了与其他类人猿和哺乳动物进行比较,我们还从dbSNP下载了一些其他物种的SNP,包括恒河猴、绒猴、猪、牛、山羊、老鼠和鸡。我们摒弃了其他物种,如狗、猫或羊,因为dbSNP为这些物种提供了数量有限的变体。我们最初将每个物种的SNP提升到hg19。结果发现,大约20%的变体被定位到假基因区域。然后,我们从100种典型编码区的脊椎动物的多

重比对文件中获得了每个物种的外显子坐标,并提取了这些外显子中的变体。然后提取的SNP被提升到hg19。如果变体是在不同的基因组构建上,我们首先将变体提升到基因组构建上

[0741] 由于牛SNP数据来自不同的研究,我们从dbSNP下载了所有大批量的牛变体(VCF文件大于100MB的16个批次),并通过计算每个批次的错义与同义比率来评估不同批次牛SNP的质量。错义与同义比率的中值为0.781,中值绝对偏差为0.160(平均值为0.879,标准差为0.496)。两批异常比值(snpBatch_1000_BULL_基因组_1059190.gz比值为1.391,snpBatch_辅因子_1059634.gz比值为2.568)被排除在进一步分析之外。

[0742] 对类人猿和哺乳动物中多态性性能的评估

[0743] 为了说明大猩猩SNP的可用性,我们设计了测量单子数和普通SNP的比值的富集得分(等位基因频率(AF) $>0.1\%$)。已知同义变体是良性的,并且通常在没有任何选择压力的情况下进行中性进化。有害的错义变体通过自然选择逐渐被清除,因此其等位基因频率分布往往比同义变体多。

[0744] 我们关注的是那些与在灵长类动物、哺乳动物和家禽中观察到的SNP重叠的gnomAD SNP。我们计算了每个物种的同义和错义变体的数目。对于错义变体,我们进一步将其分为两类,一类是在另一物种中具有相同氨基酸变化的,称为“错义相同”,另一类是在另一物种中具有不同氨基酸变化的,称为“错义不同”。然后,根据单核数目与普通变体数目之比,计算每个物种的富集得分。

[0745] 此外,我们在2x2列联表上进行了同质性的卡方测试(χ^2),以比较每个物种的同义和错义相同变体之间的富集得分。所有灵长类动物在同义和错义相同变体之间的丰富集得分没有显著差异,而牛、鼠和鸡之间的丰富集得分具有显著差异。

[0746] 研究结果表明,那些具有大猩猩的相同氨基酸变化的SNP往往具有与同义SNP十分相似的富集得分,这意味着它们往往对人类健康有轻微的影响。而那些具有不同氨基酸变化或在大猩猩中不存在的物种,其富集得分与同义SNP显著不同。非灵长类物种错义多态性与同义变体也具有不同的等位基因频率分布。结论是,在大猩猩中具有相同氨基酸变化的SNP可以被添加到良性变体的训练集中。

[0747] 我们假设,大多数变体都是独立衍生的,而不是根据血统同一性(IBD)生成的。因此,我们对IBD SNP中的稀有变体进行了富集分析,以评估其富集得分的不同行为。IBD SNP定义为出现在人类和两个或多个大猩猩物种中的人SNP包括黑猩猩、倭黑猩猩、大猩猩、B红毛猩猩和S红毛猩猩。然后分别计算错义变体和同义变体的富集得分(定义为单子数除以常见变体数(AF $>0.1\%$)),它们被认为是中性的,并作为比较的基线。

[0748] 哺乳动物物种之间的固定替换

[0749] 固定替换的富集分析

[0750] 我们还研究了物种间替换的稀有变异富集分析。我们从UCSC基因组浏览器(<http://hgdownload.soe.ucsc.edu/goldenPath/hg19/multiz100way/hg19.100way.commonNames.nh>)中下载了100种脊椎动物的系统发生树。然后我们计算成对的系统发生距离,选出亲缘关系较近的物种对(距离 <0.3)。为了获得灵长类动物物种对,我们从UCSC基因组浏览器下载了具有人类CDS区域的19个哺乳动物(16个灵长类)基因组的比对(hg38)。在13对脊椎动物中增加4对灵长类动物。下表示出了根据一个实施方案的多对近缘

物种的遗传距离。

物种 1	物种 2	距离
黑猩猩	倭黑猩猩	0.00799195
恒河猴	食蟹猕猴	0.00576807
[0751]		
狨猴	松鼠猴	0.0680206
长鼻猴	金丝猴	0.01339514
狒狒	绿猴	0.045652

马	白犀牛	0.084397
羊	家养山羊	0.1
鼠	大鼠	0.176098
南美栗鼠	刷尾鼠	0.16
金鼯	马岛猴	0.265936
中国仓鼠	金黄地鼠	0.08
[0752]		
鼠耳蝠	微型蝙蝠	0.08254
海豚	虎鲸	0.074368
猎隼	游隼	0.2
鸡	火鸡	0.126972
绿海龟	锦龟	0.2
中华鳖	多刺鳖	0.2

[0753] 我们对19个哺乳动物或99个脊椎动物基因组与人类在规范编码区内进行了多重比对,得到了每对选择的脊椎动物之间的核苷酸替换。这些替换被映射到来自gnomAD的人类外显子组SNP,需要在物种对和人类变体之间进行相同的密码子变化。我们将变体分为三种类型,同义变体,在另一物种中具有相同的氨基酸变化的错义变体,在另一物种中具有不同的氨基酸变化的错义变体。计算每个物种对的每一类的富集得分。

[0754] 种内和种间多态性比较

[0755] 共选择了六个物种进行种内和种间多态性的比较,包括黑猩猩、恒河猴、狨猴、山羊,小鼠和鸡,因为这些物种均由种内和种间变体可用。种内和种间变体的富集得分的比较类似于两个2x2相依表让步比的比较。通常采用伍尔夫检验来评估相依表之间让步比的同质性。因此,我们利用伍尔夫检验来评估种内和种间多态性之间的富集得分差异。

[0756] 每基因富集分析

[0757] 图64示出了每基因富集分析的一种实施方式。在一种实施方式中,进一步配置了基于深度卷积神经网络的变体致病性分类器,来实施每基因富集分析,从而确认已被确定为致病性的变体的致病性。对于从一群遗传性疾病患者中取样的特定基因,每基因富集分析包括应用基于深度卷积神经网络的变体致病性分类器来识别特定基因中具有致病性的候选变体,对候选变体观察到的三核苷酸突变率求和并与传播计数和群组大小相乘来确定特定基因的突变基线数,应用基于深度卷积神经网络的变体致病性分类器识别特定基因中具有致病性的新生错义变体,以及将突变基线数与新生错义变体计数进行比较。根据比较的结果,每基因富集分析确认特定基因与遗传性疾病有关,并且新生错义变体具有致病性。在一些实施方式中,遗传性疾病是自闭症谱系障碍(简称ASD)。在其他实施方式中,遗传性疾病是发育迟延障碍(简称DDD)。

[0758] 在图64所示的示例中,基于深度卷积神经网络的变体致病性分类器已将特定基因中的五个候选变体分类为具有致病性。观察到这五个候选变体分别具有 10^{-8} 、 10^{-2} 、 10^{-1} 、 10^5 和 10^1 的三核苷酸突变率。根据对5个候选突变变体各自观察到的三核苷酸突变率进行累加,并与传播/染色体计数(2)和群组大小(1000)相乘,确定特定基因的突变基线数为 10^{-5} 。然后将其与新生变体计数(3)进行比较。

[0759] 在一些实施方式中,进一步配置了基于深度卷积神经网络的变体致病性分类器,以便使用产生p值作为输出的统计检验执行比较。

[0760] 在其他实施方式中,进一步配置了基于深度卷积神经网络的变体致病性分类器,以便将突变的基线数与新生错义变体的计数进行比较,并且基于比较的输出,确认该特定基因与遗传性疾病无关,并且新生错义变体是良性的。

[0761] 全基因组富集分析

[0762] 图65示出了全基因组富集分析的一种实施方式。在另一个实施方式中,进一步配置了基于深度卷积神经网络的变体致病性分类器,来实施全基因组富集分析,从而确认已被确定为致病性的变体的致病性。全基因组富集分析包括应用基于深度卷积神经网络的变体致病性分类器来识别从健康个体群组取样的多个基因中第一组致病性的新生错义变体,应用基于深度卷积神经网络的变体致病性分类器来识别从遗传性疾病患者群组取样的多个基因中第二组致病性的新生错义变体,并比较第一组和第二组各自的计数,并且根据比较的输出确认第二组新生错义变体在遗传性疾病患者群组中富集,因此具有致病性。在一些实施方式中,遗传性疾病是自闭症谱系障碍(简称ASD)。在其他实施方式中,遗传性疾病是发育迟延障碍(简称DDD)。

[0763] 在一些实施方式中,进一步配置了基于深度卷积神经网络的变体致病性分类器,以便使用产生p值作为输出的统计检验执行比较。在一个实施方式中,通过各自的群组大小进一步对比较进行参数化。

[0764] 在一些实施方式中,进一步配置了基于深度卷积神经网络的变体致病性分类器,

以便比较第一组和第二组各自的计数,并根据比较的输出确认第二组新生错义变体在遗传性疾病患者群组中没有富集,因此是良性的。

[0765] 在图65所示的示例中,说明了健康群组的突变率(0.001)和受影响群组的突变率(0.004),以及每个个体的突变率(4)。

[0766] 特定实施方式

[0767] 我们描述了用于构建变体致病性分类器的系统、方法和制品。一个实施方式的一个或多个特征可以与基本实施方式组合。不互斥的实施方式被教导为可组合的。一个实施方式的一个或多个特征可以与其他实施方式组合。本公开定期地向用户提醒这些选项。省略了重复这些选项的叙述的某些实施方式不应被视为限制了前面章节中所教导的组合,这些叙述在此通过引用的方式被纳入以下每个实施方式中。

[0768] 所公开技术的系统实施方式包括与存储器耦合的一个或多个处理器。存储器中装有计算机指令,用于训练识别基因组序列(例如核苷酸序列)中剪接位点的剪接位点检测器。

[0769] 如图48和图19所示,该系统训练了一个基于卷积神经网络的变体致病性分类器,该分类器在耦合到存储器的多个处理器上运行。该系统使用从良性变体和致病性变体生成的蛋白质序列对的良性训练实例和致病性训练实例。良性变体包括常见的人类错义变体和非人类灵长类动物错义变体,其中非人类灵长类动物错义变体出现在共享与人类匹配的参考密码子序列的替代性非人类灵长类动物密码子序列上。短语“蛋白质序列对”指参考蛋白质序列和替代蛋白质序列,其中参考蛋白质序列包括由参考三联体核苷酸碱基(参考密码子)形成的参考氨基酸,替代蛋白质序列包括由替代三联体核苷酸碱基(替代密码子)形成的替代氨基酸,因此,由于在形成参考蛋白质序列的参考氨基酸的参考三联体核苷酸碱基(参考密码子)中出现变体,产生了替代蛋白质序列。变体可以是SNP、插入或缺失。

[0770] 所公开的该系统实施方式和其他系统可选地包括一个或多个以下特征。系统还可以包括结合所公开的方法描述的特征。为了简明起见,没有单独列举系统特征的可选组合。适用于系统、方法和制品的特征不会针对每个法定类别的基础特征集进行重复。读者将理解本节中识别的特征如何容易地与其他法定类别中的基本特征相结合。

[0771] 如图44所示,在从至少100000人采样的人类种群变体数据集中,常见的人类错义变体的次要等位基因频率(缩写为MAF)大于0.1%。

[0772] 如图44所示,被采样的人类属于不同的人类亚种群,而常见的人类错义变体在各自的人类亚种变体数据集中的MAF大于0.1%。

[0773] 人类亚种群包括非洲/非洲裔美国人(缩写为AFR)、美国人(缩写为AMR)、德系犹太人(缩写为ASJ)、东亚人(缩写为EAS)、芬兰人(缩写为FIN)、非芬兰欧洲人(缩写为NFE)、南亚人(缩写为SAS)和其他人(缩写为OTH)。

[0774] 如图43和44所示,非人类灵长类动物错义变体包括来自多种非人类灵长类动物物种的错义变体,包括黑猩猩、倭黑猩猩、大猩猩、B红毛猩猩、S红毛猩猩、恒河猴和狨猴。

[0775] 如图45和46所示,根据富集分析,该系统接受特定的非人类灵长类动物物种,以便将该特定非人类灵长类动物物种的错义变体包含在良性变体中。对于特定的非人类灵长类动物物种,富集分析包括将该特定非人类灵长类动物物种的同义变体的第一富集得分与该特定非人类灵长类动物物种的错义相同变体的第二富集得分进行比较。

[0776] 图45示出了人类同源错义SNP的一种实施方式。非人物种中具有与人类匹配的参考密码子和替代密码子的错义SNP。如图45所示,错义相同变体是共享与人类匹配的参考和替代密码子序列的错义变体。

[0777] 如图46和47所示,通过确定MAF小于0.1%的罕见同义变体与MAF大于0.1%的常见同义变体的比值,可以得到第一富集得分。通过确定MAF小于0.1%的罕见错义相同变体与MAF大于0.1%的常见错义相同变体的比值,得出第二富集得分。罕见变体包括单例变体。

[0778] 如图46和47所示,第一富集得分与第二富集得分之间的差异在预定范围内,进一步包括接受特定的非人类灵长类动物物种,以便将该特定非人类灵长类动物物种的错义变体包含在良性变体中。差异在预定范围内表明,错义相同变体与同义变体处于相同程度的自然选择之下,因此和同义变体一样是良性的。

[0779] 如图48所示,系统重复地应用富集分析,以便接受多个非人类灵长类动物物种,以便将这些非人类灵长类动物物种的错义变体包含在良性变体中。该系统还包括同质性的卡方检验,以比较每个非人类灵长类动物物种的同义变体的第一富集得分和错义相同变体的第二富集得分。

[0780] 如图48所示,非人类灵长类动物错义变体的数量至少为100000,非人类灵长类动物错义变体的数量为385236。常见人类错义变体的数量至少为50000。常见人类错义变体的数量为83546。

[0781] 其他实施方式可能包括非暂时性的计算机可读存储介质,其存储可由处理器执行以执行上述系统操作的指令。另一个实施方式可能包括完成上述系统操作的方法。

[0782] 所公开的技术的另一个系统实施方式包括构建单核苷酸多态性(缩写SNP)致病性分类器。该系统使用由良性SNP和致病性SNP表示的氨基酸序列的良性训练实例和致病性训练实例,训练一个在与存储器耦合的多个处理器上运行的基于卷积神经网络的SNP致病性分类器。良性训练实例包括表达为氨基酸序列对的第一和第二组核苷酸序列,每个氨基酸序列都包括两侧为上游和下游氨基酸的中央氨基酸。每个氨基酸序列对都包括一个由参考核苷酸序列表达的氨基酸参考序列和一个由包含SNP的替代核苷酸序列表达的氨基酸替代序列。

[0783] 如图9所示,第一组包括人类核苷酸序列对,其中每对包括一个人类替代核苷酸序列,该序列包含一个SNP并且具有被认为在人类种群中常见的次要等位基因频率(缩写为MAF)。第二组包括与非人类灵长类动物替代核苷酸序列配对的非人类灵长类动物参考核苷酸序列。非人类灵长类动物参考核苷酸序列具有直系同源人类核苷酸参考序列。非人类灵长类动物替代核苷酸序列包含一个SNP。

[0784] 在此特定实施方式部分中针对第一个系统实施方式而讨论的每个特征都同样适用于本系统实施方式。如上文所述,此处不重复所有的系统特征,应通过引用将其视为重复。

[0785] 其他实施方式可能包括非暂时性的计算机可读存储介质,其存储可由处理器执行以执行上述系统操作的指令。另一个实施方式可能包括完成上述系统操作的方法。

[0786] 如图48和图19所示,所公开的技术的第一方法实施方式包括构建变体致病性分类器,该方法包括。该方法进一步包括,使用从良性变体和致病性变体生成的蛋白质序列对的良性训练实例和致病性训练实例,训练一个在与存储器耦合的多个处理器上运行的基于卷

积神经网络的变体致病性分类器。良性变体包括常见的人类错义变体和非人类灵长类动物错义变体,其中非人类灵长类动物错义变体出现在共享与人类匹配的参考密码子序列的替代性非人类灵长类动物密码子序列上。

[0787] 在此特定实施方式部分中针对第一个系统实施方式而讨论的每个特征都同样适用于本方法实施方式。如上文所述,此处不重复所有的系统特征,应通过引用将其视为重复。

[0788] 其他实施方式可能包括非暂时性的计算机可读存储介质,其存储可由处理器执行以执行上述方法的指令。另一个实施方式可能包括一个系统,该系统包括存储器和一个或多个处理器,所述一个或多个处理器可用于执行存储在存储器中的指令以执行上述方法。

[0789] 如图48和19所示,所公开的技术的第二种方法实施方式包括构建单核苷酸多态性(缩写SNP)致病性分类器。该方法进一步包括,使用由良性SNP和致病性SNP表示的氨基酸序列的良性训练实例和致病性训练实例,训练一个在与存储器耦合的多个处理器上运行的基于卷积神经网络的SNP致病性分类器。良性训练实例包括表达为氨基酸序列对的第一和第二组核苷酸序列,每个氨基酸序列都包括两侧为上游和下游氨基酸的中央氨基酸,每个氨基酸序列对都包括一个由参考核苷酸序列表达的氨基酸参考序列和一个由包含SNP的替代核苷酸序列表达的氨基酸替代序列。第一组包括人类核苷酸序列对,其中每对包括一个人类替代核苷酸序列,该序列包含SNP并且具有被认为在人类种群中常见的次要等位基因频率(缩写为MAF)。第二组包括与非人类灵长类动物替代核苷酸序列配对的非人类灵长类动物参考核苷酸序列。非人类灵长类动物参考核苷酸序列具有直系同源人类核苷酸参考序列,非人类灵长类动物替代核苷酸序列包含一个SNP。

[0790] 在此特定实施方式部分中针对第二个系统实施方式而讨论的每个特征都同样适用于本方法实施方式。如上文所述,此处不重复所有的系统特征,应通过引用将其视为重复。

[0791] 其他实施方式可能包括非暂时性的计算机可读存储介质,其存储可由处理器执行以完成上述方法的指令。另一个实施方式可能包括一个系统,该系统包括存储器和一个或多个处理器,所述一个或多个处理器可用于执行存储在存储器中的指令以执行上述方法。

[0792] 我们描述了系统、方法和制品以便使用具有二级结构分类器和溶剂可及性分类器的基于深度卷积神经网络的变体致病性分类器。一种实施方式的一个或多个特征可以与基本实施方式组合。不互斥的实施方式被教导为可组合的。一种实施方式的一个或多个特征可以与其他实施方式组合。本公开定期地提醒用户这些选项。重复这些选项的叙述的一些实施方式的省略不应被视为限制前面章节中所教导的组合——这些叙述因此通过引用结合到下面的每个实施方式中。

[0793] 所公开的技术的系统实施方式包括耦合到存储器的一个或多个处理器。存储器装载有计算机指令,以运行具有二级结构分类器和溶剂可及性分类器的基于深度卷积神经网络的变异致病性分类器。

[0794] 系统包括第一二级结构子网络,在耦合到存储器的多个处理器上运行,训练成预测蛋白质序列中氨基酸位置的三态二级结构。系统还包括第二溶剂可及性子网络,在耦合到存储器的多个处理器上运行,训练成预测蛋白质序列中氨基酸位置的三态溶剂可及性。

[0795] 三态二级结构是指多个DNA二级结构状态 α 螺旋(H)、 β 折叠(B)和卷曲(C)。

[0796] 三态溶剂可及性是指多种蛋白质溶剂可及性状态:埋藏的 (Buried)、中间的 (Intermediate) 和暴露的 (Exposed) 之一。

[0797] 位置频率矩阵 (缩写为PFM) 生成器,在多个处理器中的至少一个上运行,应用于灵长类动物、哺乳动物、和不包括灵长类动物和哺乳动物的脊椎动物的三个序列组,以便生成灵长类动物、哺乳动物和脊椎动物PFM。

[0798] 换句话说,这包括将PFM发生器应用于灵长类动物序列数据以生成灵长类动物PFM,将PFM发生器应用于哺乳动物序列数据以生成哺乳动物PFM,以及将PFM发生器应用于不包括灵长类动物和哺乳动物序列数据的脊椎动物序列数据以生成脊椎动物PFM。

[0799] 输入处理器,其接受具有靶变体氨基酸的变体氨基酸序列,该靶变体氨基酸在每个方向上在上游和下游侧接至少25个氨基酸,其中单个核苷酸变体产生靶变体氨基酸。在多个处理器中的至少一个上运行的补充数据分配器,其分配具有目标参考氨基酸的参考氨基酸序列,该目标参考氨基酸在每个方向上在上游和下游侧接至少25个氨基酸,与变体氨基酸序列对齐。之后,它将由第一子网络和第二子网络产生的参考状态分类分配给参考氨基酸序列。此后,补充数据分配器将由第一子网络和第二子网络产生的变体状态分类分配给变体氨基酸序列。最后,它分配与参考氨基酸序列对齐的灵长类动物、哺乳动物和脊椎动物PFM。

[0800] 在本申请的上下文中,短语“与...对齐”是指针对参考氨基酸序列或替代氨基酸序列中的每个氨基酸位置,按位置确定灵长类动物、哺乳动物和脊椎动物PFM,并按照与参考氨基酸序列或替代氨基酸序列中出现的氨基酸位置相同的顺序编码和存储基于位置或顺序位置的确定结果。

[0801] 系统还包括深度卷积神经网络,在多个处理器上运行,训练成基于处理变体氨基酸序列、分配的参考氨基酸序列、分配的参考和变体状态分类以及分配的PFM来将变体氨基酸序列分类为良性的或致病性的。系统包括输出处理器,该处理器至少报告变体氨基酸序列的致病性得分。

[0802] 该系统实施方式和所公开的其他系统可选地包括一个或多个以下特征。系统还可以包括结合所公开的方法描述的特征。为了简明起见,没有单独列举系统特征的可选组合。适用于系统、方法和制品的特征不会针对每个法定类别的基础特征集进行重复。读者将理解本节中识别的特征如何容易地与其他法定类别中的基本特征相结合。

[0803] 包括基于深度卷积神经网络的变体致病性分类器的系统,进一步配置成基于致病性得分将单核苷酸变体分类为良性的或致病性的。

[0804] 系统包括基于深度卷积神经网络的变体致病性分类器,其中深度卷积神经网络并行接受至少变体氨基酸序列、分配的参考氨基酸序列、分配的变体二级结构状态分类、分配的参考二级结构状态分类、分配的变体溶剂可及性状态分类、分配的参考溶剂可及性状态分类、分配的灵长类PFM、分配的哺乳动物PFM和分配的脊椎动物PFM作为输入。

[0805] 系统配置成使用批归一化层、ReLU非线性层和维度改变层来预处理变体氨基酸序列、分配的参考氨基酸序列、分配的灵长类PFM、分配的哺乳动物PFM和分配的脊椎动物PFM。系统还配置成对预处理后的特性求和,并将和与分配的变体二级结构状态分类、分配的参考二级结构状态分类、分配的变体溶剂可及性状态分类和分配的参考溶剂可及性状态分类相级联,以产生级联输入。系统通过维度改变层来处理级联输入,并接受处理后的级联输入

以启动深度卷积神经网络的残差块。

[0806] 深度卷积神经网络包括按从最低到最高的顺序排列的残差块组。深度卷积神经网络由多个残差块、多个跳跃连接和多个残差连接进行参数化,无需非线性激活。深度卷积神经网络包括维度改变层,这些层对先前输入的空间和特征维数进行更改。

[0807] 系统还配置成进行训练以将单核苷酸变体分类为致病性的,单核苷酸变体从目标参考氨基酸中产生靶变体氨基酸,目标参考氨基酸在灵长类动物、哺乳动物和脊椎动物的对齐的参考氨基酸序列中是保守的。

[0808] 保守性代表目标参考氨基酸的功能意义,并由PFW确定。系统还配置成进行训练以将导致变体氨基酸序列和参考变体氨基酸序列之间不同二级结构的单核苷酸变体分类为致病性的。

[0809] 系统还配置成进行训练以将导致变体氨基酸序列和参考变体氨基酸序列之间不同溶剂可及性的单核苷酸变体分类为致病性的。

[0810] PFM表示通过逐个位置地确定人类蛋白质序列中的氨基酸在其他物种的对齐蛋白质序列中的出现频率,人类蛋白质序列中的氨基酸在其他物种的对齐蛋白质序列中的保守性。

[0811] 二级结构的三种状态是螺旋、折叠和卷曲。第一二级结构子网络训练成接受输入蛋白质序列和与输入蛋白质序列中的氨基酸位置对齐的灵长类动物、哺乳动物和脊椎动物PFM,并预测每个氨基酸位置的三态二级结构。溶剂可及性的三种状态是暴露的、埋藏的和中间的。

[0812] 第二溶剂可及性子网络训练成接受输入蛋白质序列和与输入蛋白质序列中的氨基酸位置对齐的灵长类动物、哺乳动物和脊椎动物PFM,并预测每个氨基酸位置的三态溶剂可及性。输入蛋白质序列是参考蛋白质序列。输入的蛋白质序列是替代蛋白质序列。第一二级结构子网络包括按从最低到最高的顺序排列的残差块。第一二级结构子网络由多个残差块、多个跳跃连接和多个残差连接进行参数化,无需非线性激活。

[0813] 第一二级结构子网络包括维度改变层,这些层对先前输入的空间和特征维数进行更改。第二溶剂可及性子网络包括按从最低到最高的顺序排列的残差块。第二溶剂可及性子网络由多个残差块、多个跳跃连接和多个残差连接进行参数化,无需非线性激活。第二溶剂可及性子网络包括维度改变层,这些层对先前输入的空间和特征维数进行更改。

[0814] 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(缩写为ReLU)层、至少一个维度改变层和至少一个残差连接。每个残差块包括两个批归一化层、两个ReLU非线性层、两个维度改变层和一个残差连接。

[0815] 深度卷积神经网络、第一二级结构子网络和第二溶剂可及性子网络均包括最终分类层。最终分类层是基于乙状结肠的层。最终分类层是基于softmax的层。

[0816] 系统还配置成消融第一二级结构子网和第二溶剂可及性子网的最终分类层,以便与深度卷积神经网络协作。

[0817] 系统还配置成在深度卷积神经网络的训练期间,进一步在致病性分类上训练第一二级结构子网络和第二溶剂可及性子网络,包括向子网络反向传播误差和更新子网络权重。

[0818] 第二溶剂可及性子网络包括至少一个空洞卷积层。系统还配置成将引起发育延迟

障碍(缩写为DDD)的变体分类为致病性的。变体氨基酸序列和参考氨基酸序列共享侧翼氨基酸。系统还配置成使用单热编码来编码对深度卷积神经网络的输入。

[0819] 图1Q示出了可以在其中操作所公开的技术的示例计算环境。在一个或多个训练服务器上训练深度卷积神经网络、第一二级结构子网络和第二溶剂可及性子网络。训练后的深度卷积神经网络、训练后的第一二级结构子网络和训练后的第二溶剂可及性子网络部署在一个或多个从请求客户端接收输入序列的生产服务器上。生产服务器通过深度卷积神经网络、第一二级结构子网络和第二溶剂可及性子网络中的至少一个来处理输入序列,以产生传输到客户端的输出。

[0820] 其他实施方式可以包括用于存储指令的非暂时性计算机可读存储介质,指令可由处理器执行以执行上述系统的动作。又一实施方式可以包括执行上述系统的动作的方法。

[0821] 所公开的技术的另一系统实施方式包括基于深度卷积神经网络的变体致病性分类器,在耦合到存储器的多个处理器上运行。系统包括:位置频率矩阵(缩写为PFM)生成器,在多个处理器中的至少一个上运行,应用于灵长类动物和哺乳动物的两个序列组,以生成灵长类动物PFM和哺乳动物PFM。系统还包括:输入处理器,其接受具有靶变体氨基酸的变体氨基酸序列,该靶变体氨基酸在每个方向上在上游和下游侧接至少25个氨基酸,其中单个核苷酸变体产生靶变体氨基酸。系统还包括:补充数据分配器,在多个处理器中的至少一个上运行,其分配具有目标参考氨基酸的参考氨基酸序列,该目标参考氨基酸在每个方向上在上游和下游侧接至少25个氨基酸,与变体氨基酸序列对齐。它还分配与参考氨基酸序列对齐的灵长类动物和哺乳动物PFM。系统还包括深度卷积神经网络,在多个处理器上运行,训练成基于处理变体氨基酸序列、分配的参考氨基酸序列、以及分配的PFM来将变体氨基酸序列分类为良性的或致病性的。最后,系统包括:输出处理器,该处理器至少报告变体氨基酸序列的致病性得分。

[0822] 该系统实施方式和所公开的其他系统可选地包括一个或多个以下特征。系统还可以包括结合公开的方法描述的特征。为了简明起见,没有单独列举系统特征的可选组合。适用于系统、方法和制品的特征不会针对每个法定类别的基础特征集进行重复。读者将理解本节中识别的特征如何容易地与其他法定类别中的基本特征相结合。

[0823] 系统进一步配置成基于致病性得分将单核苷酸变体分类为良性的或致病性的。深度卷积神经网络并行接受并处理变体氨基酸序列、分配的参考氨基酸序列、分配的灵长类PFM和分配的哺乳动物PFM。系统还配置成进行训练以将单核苷酸变体分类为致病性的,单核苷酸变体从目标参考氨基酸中产生靶变体氨基酸,目标参考氨基酸在灵长类动物和哺乳动物的参考氨基酸序列中是保守的。保守性代表目标参考氨基酸的功能意义,并由PFW确定。

[0824] 在此特定实施方式部分中针对第一系统实施方式而讨论的每个特征都同样适用于该系统实施方式。如上所述,所有的系统特征都不在这进行重复,并且通过参考应视为是重复的。

[0825] 其他实施方式可以包括用于存储指令的非暂时性计算机可读存储介质,指令可由处理器执行以执行上述系统的动作。又一实施方式可以包括执行上述系统的动作的方法。

[0826] 所公开的技术的第一方法实施方式包括在耦合到存储器的多个处理器上运行第一二级结构子网络,其训练成预测蛋白质序列中氨基酸位置的三态二级结构。在耦合到存

储器的多个处理器上运行第二溶剂可及性子网络,其训练成预测蛋白质序列中氨基酸位置的三态溶剂可及性。在多个处理器中的至少一个上运行位置频率矩阵(缩写为PFM)生成器,其应用于灵长类动物、哺乳动物、和不包括灵长类动物和哺乳动物的脊椎动物的三个序列组,以生成灵长类动物PFM、哺乳动物PFM和脊椎动物PFM。通过输入处理器接受具有靶变体氨基酸的变体氨基酸序列,该靶变体氨基酸在每个方向上在上游和下游侧接至少25个氨基酸。单核苷酸变体产生靶变体氨基酸。在多个处理器中的至少一个上运行补充数据分配器,其分配具有目标参考氨基酸的参考氨基酸序列,该目标参考氨基酸在每个方向上在上游和下游侧接至少25个氨基酸,与变体氨基酸序列对齐。它还将由第一子网络和第二子网络产生的参考状态分类分配给参考氨基酸序列。它还将由第一子网络和第二子网络产生的变体状态分类分配给变体氨基酸序列。它分配与参考氨基酸序列对齐的灵长类动物、哺乳动物和脊椎动物PFM。在多个处理器上运行深度卷积神经网络,其训练成基于处理变体氨基酸序列、分配的参考氨基酸序列、分配的参考和变体状态分类以及分配的PFM来将变体氨基酸序列分类为良性的或致病性的。通过输出处理器至少报告变体氨基酸序列的致病性得分。

[0827] 在此特定实施方式部分中针对第一系统实施方式而讨论的每个特征都同样适用于该方法实施方式。如上所述,所有的系统特征都不在这进行重复,并且通过参考应视为是重复的。

[0828] 其他实施方式可以包括用于存储指令的非暂时性计算机可读存储介质,指令可由处理器执行以执行上述方法。又一实施方式可以包括一种系统,该系统包括存储器和一个或多个处理器,所述一个或多个处理器可操作以执行存储在存储器中的指令,以便执行上述方法。

[0829] 所公开的技术的第二方法实施方式包括在耦合到存储器的多个处理器上运行基于深度卷积神经网络的变体致病性分类器。在多个处理器中的至少一个上运行位置频率矩阵(缩写为PFM)生成器,其应用于灵长类动物和哺乳动物的两个序列组,以生成灵长类动物PFM和哺乳动物PFM。在输入处理器中接受具有靶变体氨基酸的变体氨基酸序列,该靶变体氨基酸在每个方向上在上游和下游侧接至少25个氨基酸。单核苷酸变体产生靶变体氨基酸。在多个处理器中的至少一个上运行补充数据分配器,其分配具有目标参考氨基酸的参考氨基酸序列,该目标参考氨基酸在每个方向上在上游和下游侧接至少25个氨基酸,与变体氨基酸序列对齐,并且分配与参考氨基酸序列对其的灵长类动物和哺乳动物PFM。在多个处理器上运行深度卷积神经网络,其训练成基于处理变体氨基酸序列、分配的参考氨基酸序列、以及分配的PFM来将变体氨基酸序列分类为良性的或致病性的。在输出处理器中至少报告变体氨基酸序列的致病性得分。

[0830] 在此特定实施方式部分中针对第二系统实施方式而讨论的每个特征都同样适用于该方法实施方式。如上所述,所有的系统特征都不在这进行重复,并且通过参考应视为是重复的。

[0831] 其他实施方式可以包括用于存储指令的非暂时性计算机可读存储介质,指令可由处理器执行以执行上述方法。又一实施方式可以包括一种系统,该系统包括存储器和一个或多个处理器,所述一个或多个处理器可操作以执行存储在存储器中的指令,以便执行上述方法。

[0832] 所公开的技术的又一系统实施方式包括一种系统,该系统生成用于训练单核苷酸

多态性(缩写为SNP)的致病性分类器的大规模致病训练数据。

[0833] 如图19所示,该系统使用从组合生成的SNP的合成集中筛选出的良性SNP的训练集和精英预测致病SNP的训练集来训练SNP致病性分类器,该分类器在耦合到存储器的多个处理器上运行。在本申请的上下文中,精英预测的致病性SNP是在每个循环结束时根据集合输出的平均或最大致病性得分产生/选择的那些SNP。术语“精英”是从遗传算法词汇中借用的,意在具有遗传算法出版物中通常给出的含义。

[0834] 如图37、图38、图39、图40、图41和42所示,系统在多个循环内迭代地构建精英集,从没有预测的SNP开始并且通过从合成集中筛选出异常(outlier)SNP来累积预测的SNP的完整集。合成集包括伪致病SNP,它们是组合生成的SNP,不存在于良性集中并且由于从合成集中迭代筛选出异常SNP以将其包含在精英集中而集合成员减少。在本申请的上下文中,术语“筛选”是指用新种群过滤、替换、更新或选择先前种群。术语“筛选”是从遗传算法词汇中借用的,意在具有遗传算法出版物中通常给出的含义。

[0835] 如图37、图38、图39、图40、图41和42所示,系统训练并应用一组SNP致病性分类器,以便在多个循环内迭代地从合成集中筛选出异常SNP。这包括使用良性SNP的公用训练集、精英预测致病SNP的公用训练集以及从合成集中采样而无需替换的伪致病SNP的单独训练集来训练集合。这还包括通过应用训练后的集合对合成集合中的至少一些SNP(在当前循环中未用于训练集合)进行评分,并且使用这些得分从评分后的SNP中选择当前循环异常SNP以便累积在公共精英集中来应用训练后的集合从合成集中筛选出异常SNP并将筛选出的SNP累积在公共精英集中。

[0836] 在本申请的上下文中,“伪致病SNP”是那些出于训练目的而标记为致病性的并从合成生成的变体中取样而在训练过程中没有替换的SNP。

[0837] 而且,精英预测致病SNP的训练集在多个循环内迭代构建。

[0838] 如图37、图38、图39、图40、图41和图42所示,然后系统将通过训练得出的分类器参数、在循环内并在公共良性集的预定范围内完成的公共精英集、和用于训练SNP致病性分类器的公共良性集存储在存储器中。

[0839] 如图37、图38、图39、图40、图41和图42所示,精英预测的致病性SNP是由集合所预测的SNP的前5%。在一些实施方式中,它们是固定数量的得分最高的SNP,例如20000。

[0840] SNP致病性分类器和SNP致病性分类器的集合都是深度卷积神经网络(缩写为DCNN)。该集合包括4到16个DCNN。如图37、图38、图39、图40、图41和图42所示,该集合包括8个DCNN。

[0841] 如图37、图38、图39、图40、图41和图42所示,系统在循环过程中的各个时期训练DCCN的集合,从而当关于验证样本的预测形成良性和致病预测的离散概率分布集群时,结束特定循环的训练。

[0842] 如图37、图38、图39、图40、图41和图42所示,系统使用得分通过对来自DCCN的集合的得分求和来选择当前循环的异常SNP。

[0843] 如图37、图38、图39、图40、图41和图42所示,系统使用得分通过对由DCNN的集合评分的每个SNP取最大平均值来选择当前循环的异常SNP。

[0844] 如图37、图38、图39、图40、图41和42所示,在当前循环内不进行替换的采样导致当前循环内伪致病SNP的不相交的独立训练集。

[0845] 系统继续循环,直到达到终止条件。终止条件可以是预定的循环数量。如图37、图38、图39、图40、图41和图42所示,预定的循环数量是21。

[0846] 如图37、图38、图39、图40、图41和图42所示,终止条件是精英预测的致病集大小在良性集大小的预定范围内。

[0847] 分类器参数至少可以是卷积滤波器权重和学习速率。

[0848] 系统可以选择集合中的SNP致病性分类器之一作为该SNP致病性分类器。所选择的SNP致病性分类器可以是在最终循环内评估的验证样本上超过集合中其他SNP致病性分类器的分类器。

[0849] 如图37、图38、图39、图40、图41和图42所示,在整个循环内完成的普通精英集可以具有至少400000个精英预测的致病性SNP。

[0850] 如图37、图38、图39、图40、图41和图42所示,系统在每个循环中都可以匹配良性SNP与采样的伪致病性SNP之间的三核苷酸上下文,以防止精英预测的致病性SNP中的突变率偏好。

[0851] 如图37、图38、图39、图40、图41和图42所示,在每个连续循环中,来自合成组的伪致病性SNP采样可以减少5%。

[0852] 如图37、图38、图39、图40、图41和图42所示,该系统可以过滤合成SNP,这些合成SNP在当前循环内通过当前循环中采样用于训练的伪致病性SNP、精英预测致病SNP和在当前循环中用于训练的良性SNP来进行评分。

[0853] 在此特定实施方式部分中针对第一系统实施方式而讨论的每个特征都同样适用于该系统实施方式。如上所述,所有的系统特征都不在这进行重复,并且通过参考应视为是重复的。

[0854] 其他实施方式可以包括用于存储指令的非暂时性计算机可读存储介质,指令可由处理器执行以执行上述系统的动作。又一实施方式可以包括一种系统,该系统包括存储器和一个或多个处理器,所述一个或多个处理器可操作以执行存储在存储器中的指令,以便执行上述系统的动作。

[0855] 所公开的技术的又一实施方式包括基于卷积神经网络(缩写为CNN)的半监督学习器,如图36所示。

[0856] 如图36所示,半监督学习器可以包括在耦合到存储器的多个处理器上运行的CNN的集合,该集合在良性训练集和致病性训练集上迭代训练。

[0857] 如图36所示,半监督学习器可以包括在至少一个处理器上运行的集合增强器,该集合增强器基于所训练的集合对合成集的评估,逐步增加病原训练集的集合大小;

[0858] 在每次迭代中,评估都产生一个精英预测的致病集,该致病集由集合增强器添加到致病训练集中。

[0859] 半监督学习器可以包括构建器,该构建器使用CNN、增强的致病训练集和良性训练集中的至少一个来构建和训练单核苷酸多态性(缩写为SNP)致病性分类器。

[0860] 在此特定实施部分中针对第一种系统实施方式而讨论的每个特征都同样适用于该系统实施方式。如上所述,所有的系统特征都不在这进行重复,并且通过参考应视为重复的。

[0861] 其他实施方式可以包括非暂时性计算机可读存储介质,其存储可由处理器执行以

执行上述系统的动作的指令。又一实施方式可以包括一种系统,该系统包括存储器和一个或多个处理器,所述一个或多个处理器可操作以执行存储在存储器中的指令,以便执行上述系统的动作

[0862] 呈现前面的描述是为了能够制订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,并且在不偏离所公开的技术的精神和范围的情况下,本文中定义的一般原理可以应用于其他实施方式和申请。因此,所公开的技术并不旨在限于所示的实施方式,而是旨在符合与本文中所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。

[0863] 计算机系统

[0864] 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。

[0865] 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。

[0866] 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。

[0867] 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。

[0868] 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。

[0869] 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。

[0870] 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。

[0871] 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66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66

中描绘的计算机系统更多或更少的组件。

[0872] 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。

[0873] 附录

[0874] 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。

[0875] 1.Laksshman Sundaram,Hong Gao,Samskruthi Reddy Padigepati,Jeremy F.McRae,YanJun Li,Jack A.Kosmicki,Nondas Fritzilas,JörgHakenberg,Anindita Dutta,John Shon,Jinbo Xu,Serafim Batzoglou,Xiaolin Li&Kyle Kai-How Farh.Predicting the clinical impact of human mutation with deep neural networks.Nature Genetics volume 50,pages1161-1170(2018).Accessible at <https://www.nature.com/articles/s41588-018-0167-z>.

[0876] 2.MacArthur,D.G.et al.Guidelines for investigating causality of sequence variants in human disease.Nature 508,469-476,doi:10.1038/nature13127 (2014) .

[0877] 3.Rehm,H.L.,J.S.Berg,L.D.Brooks,C.D.Bustamante,J.P.Evans,M.J.Landrum,D.H.Ledbetter,D.R.Maglott,C.L.Martin,R.L.Nussbaum,S.E.Plon,E.M.Ramos,S.T.Sherry,M.S.Watson.ClinGen--the Clinical Genome Resource.N.Engl.J.Med.372,2235-2242 (2015) .

[0878] 4.Bamshad,M.J.,S.B.Ng,A.W.Bigham,H.K.Tabor,M.J.Emond,D.A.Nickerson,J.Shendure.Exome sequencing as a tool for Mendelian disease gene discovery.Nat.Rev.Genet.12,745-755 (2011) .

[0879] 5.Rehm,H.L.Evolving health care through personal genomics.Nature Reviews Genetics 18,259-267 (2017) .

[0880] 6.Richards,S.et al.Standards and guidelines for the interpretation of sequence variants:a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology.Genet Med 17,405-424,doi:10.1038/gim.2015.30 (2015) .

[0881] 7.Lek,M.et al.Analysis of protein-coding genetic variation in 60,706 humans.Nature 536,285-291,doi:10.1038/nature19057 (2016) .

[0882] 8.Mallick,S.et al.The Simons Genome Diversity Project:300 genomes from 142 diverse populations.Nature 538,201-206,doi:10.1038/nature18964 (2016) .

[0883] 9.Genomes Project Consortium et al.A global reference for human

genetic variation. *Nature* 526,68–74,doi:10.1038/nature15393 (2015) .

[0884] 10.Liu,X.,X.Jian,E.Boerwinkle.dbNSFP:A lightweight database of human nonsynonymous SNPs and their functional predictions. *Human Mutation* 32,894–899 (2011) .

[0885] 11.Chimpanzee Sequencing Analysis Consortium.Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437,69–87,doi:10.1038/nature04072 (2005) .

[0886] 12.Takahata,N.Allelic genealogy and human evolution. *Mol Biol Evol* 10, 2–22 (1993) .

[0887] 13.Asthana,S.,Schmidt,S.&Sunyaev,S.A limited role for balancing selection. *Trends Genet* 21,30–32,doi:10.1016/j.tig.2004.11.001 (2005) .

[0888] 14.Leffler,E.M.,Z.Gao,S.Pfeifer,L.Ségurel,A.Auton,O.Venn,R.Bowden, R.Bontrop,J.D.Wall,G.Sella,P.Donnelly.Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* 339,1578–1582 (2013) .

[0889] 15.Samocha,K.E.et al.A framework for the interpretation of de novo mutation in human disease. *Nat Genet* 46,944–950,doi:10.1038/ng.3050 (2014) .

[0890] 16.Ohta,T.Slightly deleterious mutant substitutions in evolution. *Nature* 246,96–98 (1973) .

[0891] 17.Reich,D.E.&Lander,E.S.On the allelic spectrum of human disease. *Trends Genet* 17,502–510 (2001) .

[0892] 18.Whiffin,N.,E.Minikel,R.Walsh,A.H.O’Donnell-Luria,K.Karczewski, A.Y.Ing,P.J.Barton,B.Funke,S.A.Cook,D.MacArthur,J.S.Ware.Using high-resolution variant frequencies to empower clinical genome interpretation. *Genetics in Medicine* 19,1151–1158 (2017) .

[0893] 19.Prado-Martinez,J.et al.Great ape genome diversity and population history. *Nature* 499,471–475 (2013) .

[0894] 20.Klein,J.,Satta,Y.,O’Huigin,C.&Takahata,N.The molecular descent of the major histocompatibility complex. *Annu Rev Immunol* 11,269–295,doi:10.1146/annurev.iy.11.040193.001413 (1993) .

[0895] 21.Kimura,M.The neutral theory of molecular evolution. (Cambridge University Press,1983) .

[0896] 22.de Manuel,M.et al.Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science* 354,477–481,doi:10.1126/science.aag2602 (2016) .

[0897] 23.Locke,D.P.et al.Comparative and demographic analysis of orang-utan genomes. *Nature* 469,529–533 (2011) .

[0898] 24.Rhesus Macaque Genome Sequencing Analysis Consortium et al.Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 316,222–234,doi:10.1126/science.1139247 (2007) .

[0899] 25.Worley,K.C.,W.C.Warren,J.Rogers,D.Locke,D.M.Muzny,E.R.Mardis,

G.M.Weinstock,S.D.Tardif,K.M.Aagaard,N.Archidiacono,N.A.Rayan.The common marmoset genome provides insight into primate biology and evolution.Nature Genetics 46,850–857 (2014) .

[0900] 26.Sherry,S.T.et al.dbSNP:the NCBI database of genetic variation.Nucleic Acids Res 29,308–311 (2001) .

[0901] 27.Schrager,C.G.&Russo,C.A.Timing the origin of New World monkeys.Mol Biol Evol 20,1620–1625,doi:10.1093/molbev/msg172 (2003) .

[0902] 28.Landrum,M.J.et al.ClinVar:public archive of interpretations of clinically relevant variants.Nucleic Acids Res 44,D862–868,doi:10.1093/nar/gkv1222 (2016) .

[0903] 29.Brannon,E.P.,Idzerda,R.L.&McKnight,G.S.Targeting the mouse genome: a compendium of knockouts (Part II) .Curr Biol 5,758–765 (1995) .

[0904] 30.Lieschke,J.G.,P.D.Currie.Animal models of human disease:zebrafish swim into view.Nature Reviews Genetics 8,353–367 (2007) .

[0905] 31.Sittig,L.J.,P.Carbonetto,K.A.Engel,K.S.Krauss,C.M.Barrios–Camacho, A.A.Palmer.Genetic background limits generalizability of genotype–phenotype relationships.Neuron 91,1253–1259 (2016) .

[0906] 32.Bazykin,G.A.et al.Extensive parallelism in protein evolution.Biol Direct 2,20,doi:10.1186/1745–6150–2–20 (2007) .

[0907] 33.Ng,P.C.&Henikoff,S.Predicting deleterious amino acid substitutions.Genome Res 11,863–874,doi:10.1101/gr.176601 (2001) .

[0908] 34.Adzhubei,I.A.et al.A method and server for predicting damaging missense mutations.Nat Methods 7,248–249,doi:10.1038/nmeth0410–248 (2010) .

[0909] 35.Chun,S.,J.C.Fay.Identification of deleterious mutations within three human genomes.Genome research 19,1553–1561 (2009) .

[0910] 36.Schwarz,J.M.,C. **Rödelsperger**,M.Schuelke,D.Seelow.MutationTaster evaluates disease-causing potential of sequence alterations.Nat.Methods 7, 575–576 (2010) .

[0911] 37.Reva,B.,Antipin,Y.&Sander,C.Predicting the functional impact of protein mutations:application to cancer genomics.Nucleic Acids Res 39,e118, doi:10.1093/nar/gkr407 (2011) .

[0912] 38.Dong,C.et al.Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies.Hum Mol Genet 24,2125–2137,doi:10.1093/hmg/ddu733 (2015) .

[0913] 39.Carter,H.,Duville,C.,Stenson,P.D.,Cooper,D.N.&Karchin, R.Identifying Mendelian disease genes with the variant effect scoring tool.BMC Genomics 14 Suppl 3,S3,doi:10.1186/1471–2164–14–S3–S3 (2013) .

[0914] 40.Choi,Y.,Sims,G.E.,Murphy,S.,Miller,J.R.&Chan,A.P.Predicting the functional effect of amino acid substitutions and indels.PLoS One 7,e46688,

doi:10.1371/journal.pone.0046688 (2012) .

[0915] 41.Gulko,B.,Hubisz,M.J.,Gronau,I.&Siepel,A.A method for calculating probabilities of fitness consequences for point mutations across the human genome.Nat Genet 47,276–283,doi:10.1038/ng.3196 (2015) .

[0916] 42.Shihab,H.A.et al.An integrative approach to predicting the functional effects of non-coding and coding sequence variation.Bioinformatics 31,1536–1543,doi:10.1093/bioinformatics/btv009 (2015) .

[0917] 43.Quang,D.,Chen,Y.&Xie,X.DANN:a deep learning approach for annotating the pathogenicity of genetic variants.Bioinformatics 31,761–763, doi:10.1093/bioinformatics/btu703 (2015) .

[0918] 44.Bell,C.J.,D.L.Dinwiddie,N.A.Miller,S.L.Hateley,E.E.Ganusova, J.Midge,R.J.Langley,L.Zhang,C.L.Lee,R.D.Schilkey,J.E.Woodward,H.E.Peckham, G.P.Schroth,R.W.Kim,S.F.Kingsmore.Comprehensive carrier testing for severe childhood recessive diseases by next generation sequencing.Sci.Transl.Med.3, 65ra64 (2011) .

[0919] 45.Kircher,M.,D.M.Witten,P.Jain,B.J.O’Roak,G.M.Cooper,J.Shendure.A general framework for estimating the relative pathogenicity of human genetic variants.Nat.Genet.46,310–315 (2014) .

[0920] 46.Smedley,D.et al.A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease.Am J Hum Genet 99,595–606,doi:10.1016/j.ajhg.2016.07.005 (2016) .

[0921] 47.Ioannidis,N.M.et al.REVEL:an ensemble method for predicting the pathogenicity of rare missense variants.Am J Hum Genet 99,877–885,doi: 10.1016/j.ajhg.2016.08.016 (2016) .

[0922] 48.Jagadeesh,K.A.,A.M.Wenger,M.J.Berger,H.Guturu,P.D.Stenson, D.N.Cooper,J.A.Bernstein,G.Bejerano.M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity.Nature genetics 48,1581–1586 (2016) .

[0923] 49.Grimm,D.G.The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity.Human mutation 36, 513–523 (2015) .

[0924] 50.He,K.,X.Zhang,S.Ren,J.Sun.in Proceedings of the IEEE conference on computer vision and pattern recognition.770–778.

[0925] 51.Heffernan,R.et al.Improving prediction of secondary structure, local backbone angles,and solvent accessible surface area of proteins by iterative deep learning.Sci Rep 5,11476,doi:10.1038/srep11476 (2015) .

[0926] 52.Wang,S.,J.Peng,J.Ma,J.Xu.Protein secondary structure prediction using deep convolutional neural fields.Scientific reports 6,18962–18962 (2016) .

- [0927] 53.Harpak,A.,A.Bhaskar,J.K.Pritchard.Mutation Rate Variation is a Primary Determinant of the Distribution of Allele Frequencies in Humans.PLoS Genetics 12(2016) .
- [0928] 54.Payandeh,J.,Scheuer,T.,Zheng,N.&Catterall,W.A.The crystal structure of a voltage-gated sodium channel.Nature 475,353-358(2011) .
- [0929] 55.Shen,H.et al.Structure of a eukaryotic voltage-gated sodium channel at near-atomic resolution.Science 355,eaal4326,doi:10.1126/science.aal4326(2017) .
- [0930] 56.Nakamura,K.et al.Clinical spectrum of SCN2A mutations expanding to Ohtahara syndrome.Neurology 81,992-998,doi:10.1212/WNL.0b013e3182a43e57(2013) .
- [0931] 57.Henikoff,S.&Henikoff,J.G.Amino acid substitution matrices from protein blocks.Proc Natl Acad Sci USA 89,10915-10919(1992) .
- [0932] 58.Li,W.H.,C.I.Wu,C.C.Luo.Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications.Journal of Molecular Evolution 21,58-71(1984) .
- [0933] 59.Grantham,R.Amino acid difference formula to help explain protein evolution.Science 185,862-864(1974) .
- [0934] 60.LeCun,Y.,L.Bottou,Y.Bengio,P.Haffner.in Proceedings of the IEEE 2278-2324.
- [0935] 61.Vissers,L.E.,Gilissen,C.&Veltman,J.A.Genetic studies in intellectual disability and related disorders.Nat Rev Genet 17,9-18,doi:10.1038/nrg3999(2016) .
- [0936] 62.Neale,B.M.et al.Patterns and rates of exonic de novo mutations in autism spectrum disorders.Nature 485,242-245,doi:10.1038/nature11011(2012) .
- [0937] 63.Sanders,S.J.et al.De novo mutations revealed by whole-exome sequencing are strongly associated with autism.Nature 485,237-241,doi:10.1038/nature10945(2012) .
- [0938] 64.De Rubeis,S.et al.Synaptic,transcriptional and chromatin genes disrupted in autism.Nature 515,209-215,doi:10.1038/nature13772(2014) .
- [0939] 65.Deciphering Developmental Disorders Study.Large-scale discovery of novel genetic causes of developmental disorders.Nature 519,223-228,doi:10.1038/nature14135(2015) .
- [0940] 66.Deciphering Developmental Disorders Study.Prevalence and architecture of de novo mutations in developmental disorders.Nature 542,433-438,doi:10.1038/nature21062(2017) .
- [0941] 67.Iossifov,I.et al.The contribution of de novo coding mutations to autism spectrum disorder.Nature 515,216-221,doi:10.1038/nature13908(2014) .
- [0942] 68.Zhu,X.,Need,A.C.,Petrovski,S.&Goldstein,D.B.One gene,many

neuropsychiatric disorders:lessons from Mendelian diseases.Nat Neurosci 17, 773–781,doi:10.1038/nn.3713(2014) .

[0943] 69.Leffler,E.M.,K.Bullaughhey,D.R.Matute,W.K.Meyer,L.Ségurel,A.Venkat, P.Andolfatto,M.Przeworski.Revisiting an old riddle:what determines genetic diversity levels within species?PLoS biology 10,e1001388(2012) .

[0944] 70.Estrada,A.et al.Impending extinction crisis of the world's primates:Why primates matter.Science advances 3,e1600946(2017) .

[0945] 71.Kent,W.J.,C.W.Sugnet,T.S.Furey,K.M.Roskin,T.H.Pringle,A.M.Zahler, D.Haussler.The human genome browser at UCSC.Genome Res.12,996–1006(2002) .

[0946] 72.Tyner,C.et al.The UCSC Genome Browser database:2017 update.Nucleic Acids Res 45,D626–D634,doi:10.1093/nar/gkwl134(2017) .

[0947] 73.Kabsch,W.&Sander,C.Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.Biopolymers 22,2577–2637,doi:10.1002/bip.360221211(1983) .

[0948] 74.Joosten,R.P.et al.A series of PDB related databases for everyday needs.Nucleic Acids Res 39,D411–419,doi:10.1093/nar/gkq1105(2011) .

[0949] 75.He,K.,Zhang,X.,Ren,S.&Sun,J.in European Conference on Computer Vision.630–645(Springer) .

[0950] 76.Ionita-Laza,I.,McCallum,K.,Xu,B.&Buxbaum,J.D.A spectral approach integrating functional genomic annotations for coding and noncoding variants.Nat Genet 48,214–220,doi:10.1038/ng.3477(2016) .

[0951] 77.Li,B.et al.Automated inference of molecular mechanisms of disease from amino acid substitutions.Bioinformatics 25,2744–2750,doi:10.1093/bioinformatics/btp528(2009) .

[0952] 78.Lu,Q.et al.A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data.Sci Rep 5,10576,doi:10.1038/srep10576(2015) .

[0953] 79.Shihab,H.A.et al.Predicting the functional,molecular,and phenotypic consequences of amino acid substitutions using hidden Markov models.Hum Mutat 34,57–65,doi:10.1002/humu.22225(2013) .

[0954] 80.Davydov,E.V.et al.Identifying a high fraction of the human genome to be under selective constraint using GERP++.PLoS Comput Biol 6,e1001025, doi:10.1371/journal.pcbi.1001025(2010) .

[0955] 81.Liu,X.,Wu,C.,Li,C.&Boerwinkle,E.dbNSFP v3.0:A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs.Hum Mutat 37,235–241,doi:10.1002/humu.22932(2016) .

[0956] 82.Jain,S.,White,M.&Radivojac,P.in Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence.2066–2072.

[0957] 83.de Ligt,J.et al.Diagnostic exome sequencing in persons with severe

intellectual disability.N Engl J Med 367,1921–1929,doi:10.1056/NEJMoal206524 (2012) .

[0958] 84.Iossifov,I.et al.De novo gene disruptions in children on the autistic spectrum.Neuron 74,285–299,doi:10.1016/j.neuron.2012.04.009 (2012) .

[0959] 85.O’Roak,B.J.et al.Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations.Nature 485,246–250,doi:10.1038/nature10989 (2012) .

[0960] 86.Rauch,A.et al.Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability:an exome sequencing study.Lancet 380,1674–1682,doi:10.1016/S0140-6736(12)61480-9 (2012) .

[0961] 87.Epi,K.C.et al.De novo mutations in epileptic encephalopathies.Nature 501,217–221,doi:10.1038/nature12439 (2013) .

[0962] 88.Euro,E.-R.E.S.C.,Epilepsy Phenome/Genome,P.&Epi,K.C.De novo mutations in synaptic transmission genes including DNMT1 cause epileptic encephalopathies.Am J Hum Genet 95,360–370,doi:10.1016/j.ajhg.2014.08.013 (2014) .

[0963] 89.Gilissen,C.et al.Genome sequencing identifies major causes of severe intellectual disability.Nature 511,344–347,doi:10.1038/nature13394 (2014) .

[0964] 90.Lelieveld,S.H.et al.Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability.Nat Neurosci 19,1194–1196,doi:10.1038/nn.4352 (2016) .

[0965] 91.Famiglietti,M.L.et al.Genetic variations and diseases in UniProtKB/Swiss-Prot:the ins and outs of expert manual curation.Hum Mutat 35,927–935,doi:10.1002/humu.22594 (2014) .

[0966] 92.Horaitis,O.,Talbot,C.C.,Jr.,Phommavanh,M.,Phillips,K.M.&Cotton,R.G.A database of locus-specific databases.Nat Genet 39,425,doi:10.1038/ng0407-425 (2007) .

[0967] 93.Stenson,P.D.et al.The Human Gene Mutation Database:building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine.Hum Genet 133,1–9,doi:10.1007/s00439-013-1358-4 (2014) .

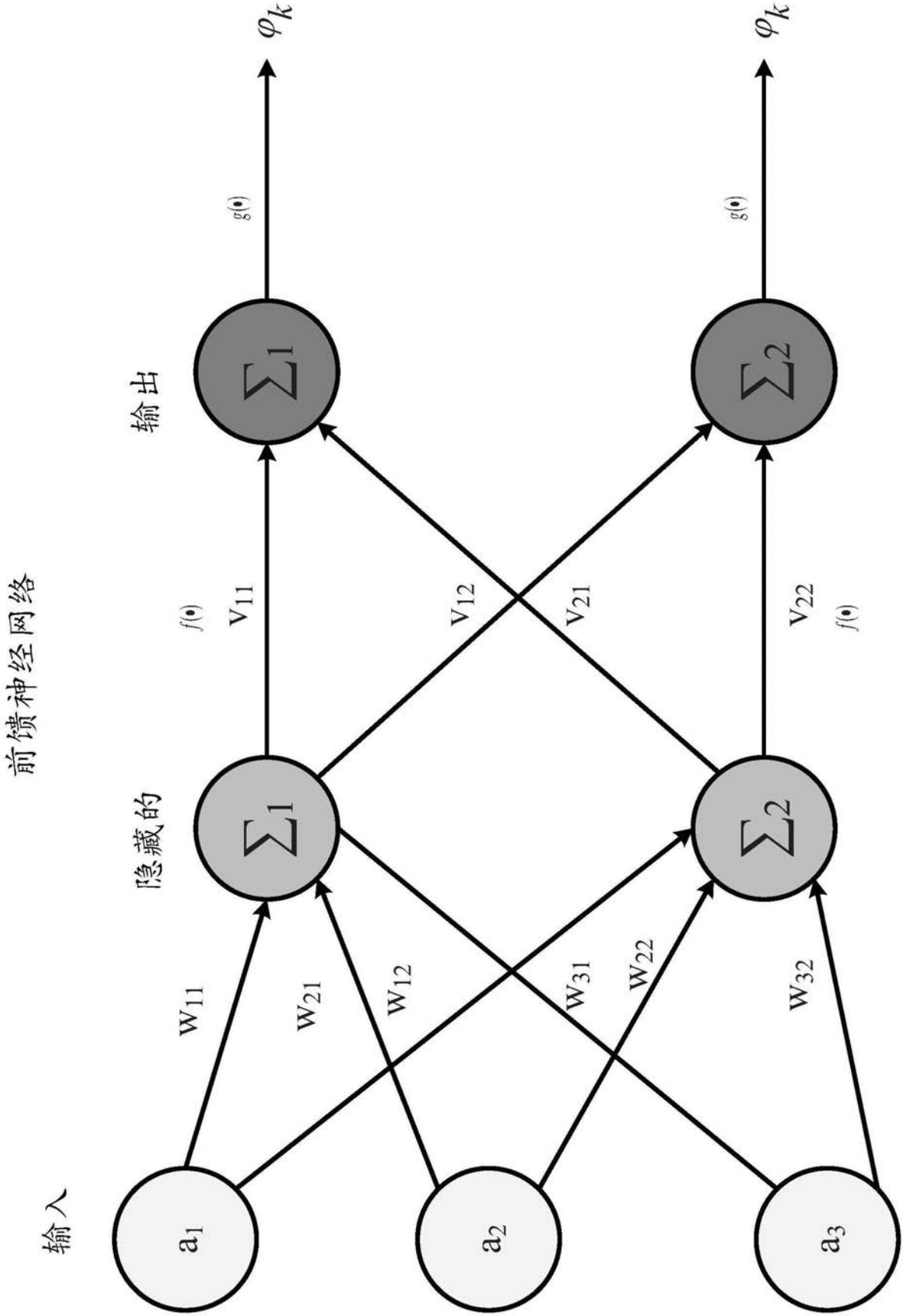


图1A

卷积神经网络

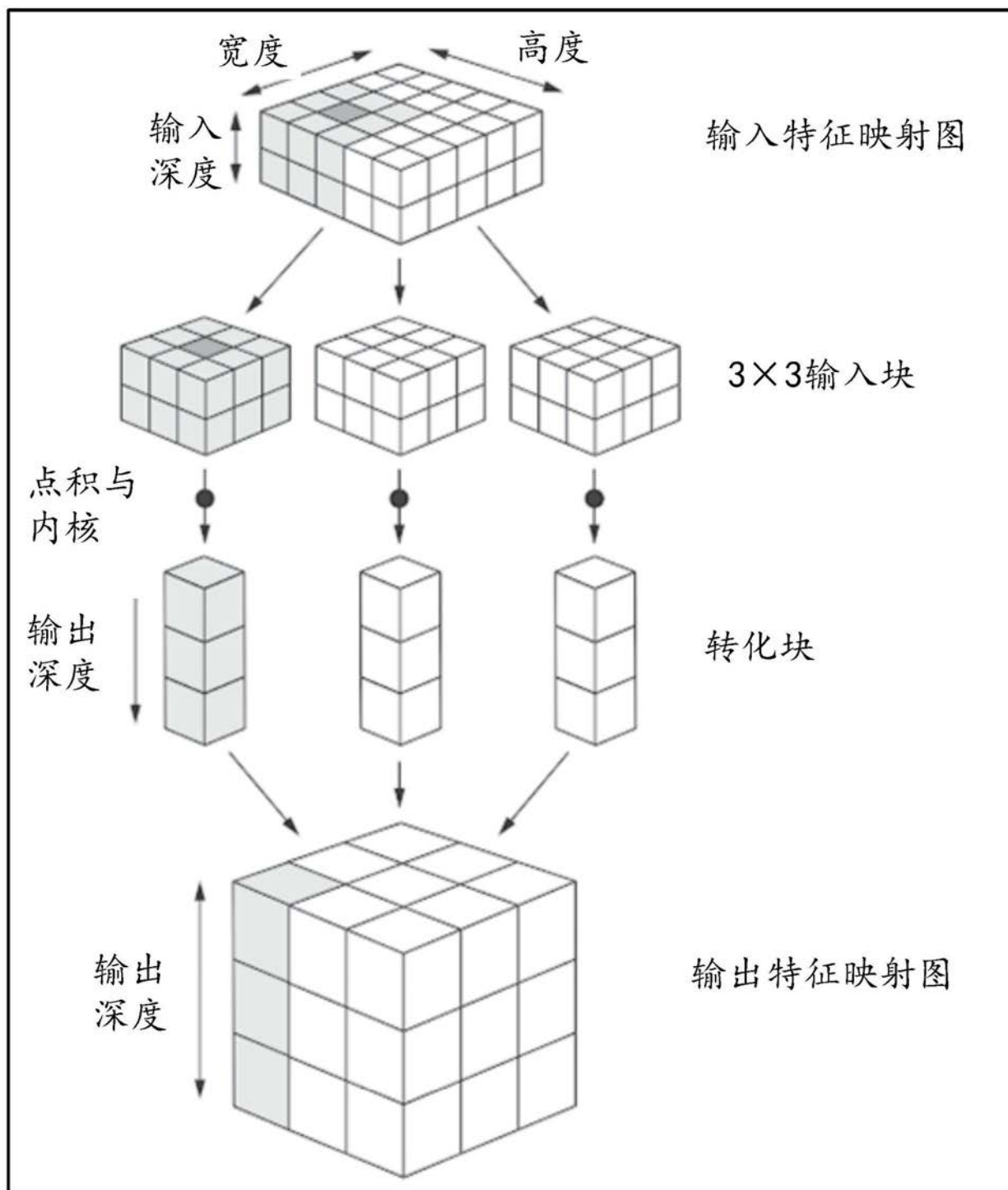


图1B

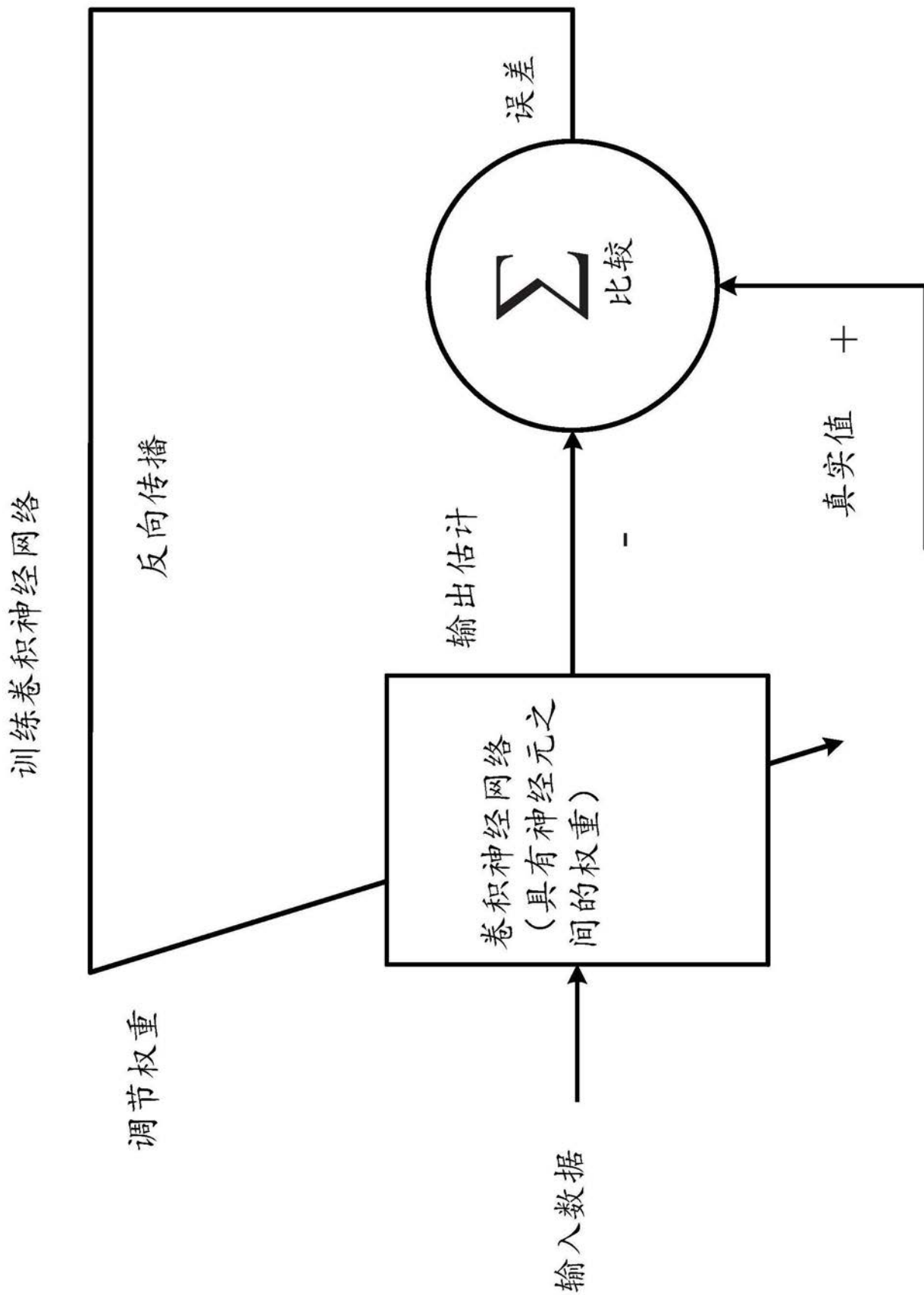


图1C

卷积神经网络的子采样层

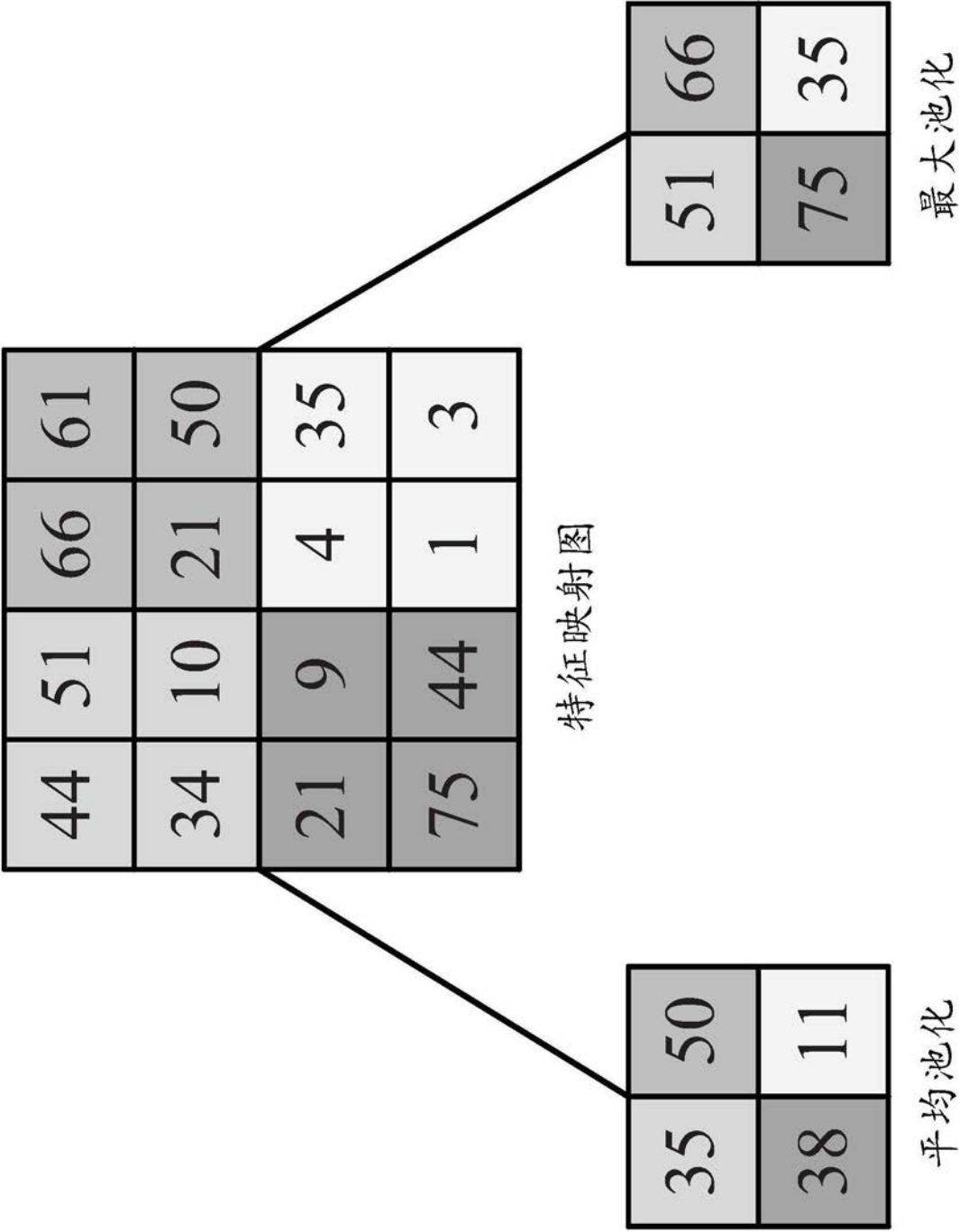
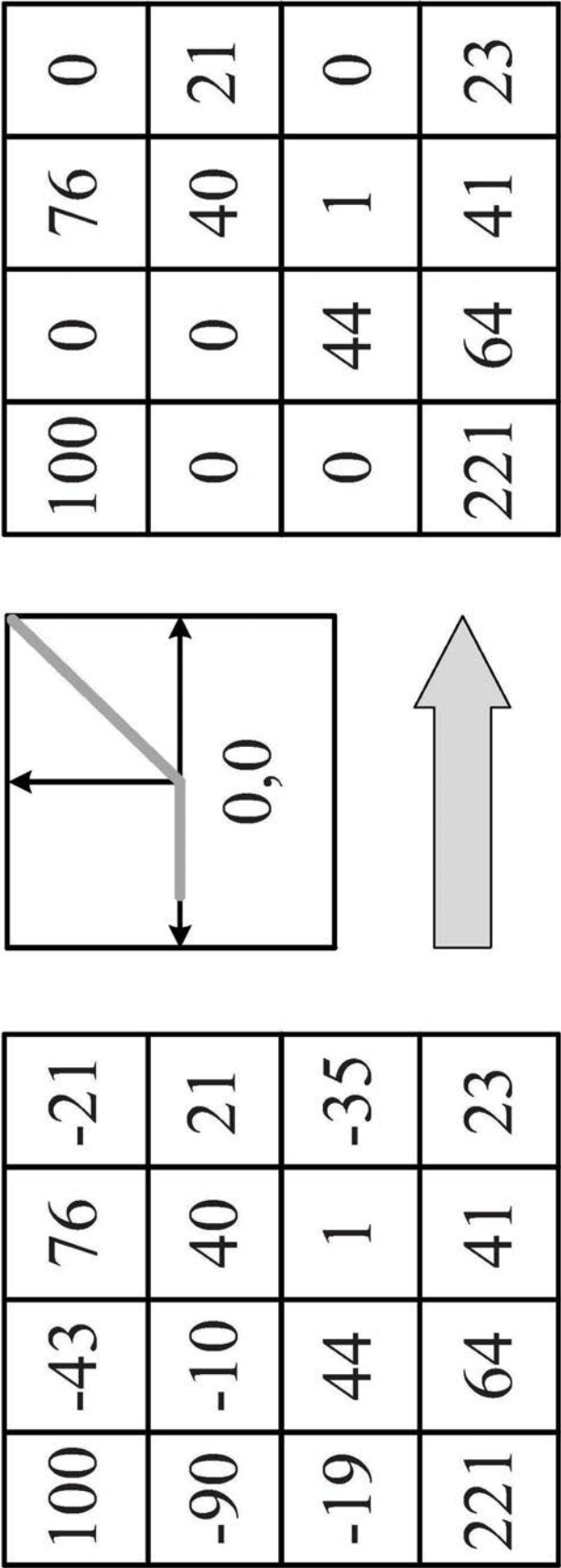


图1D

卷积神经网络的非线性层



ReLU激活函数

图1E

卷积神经网络的卷积层

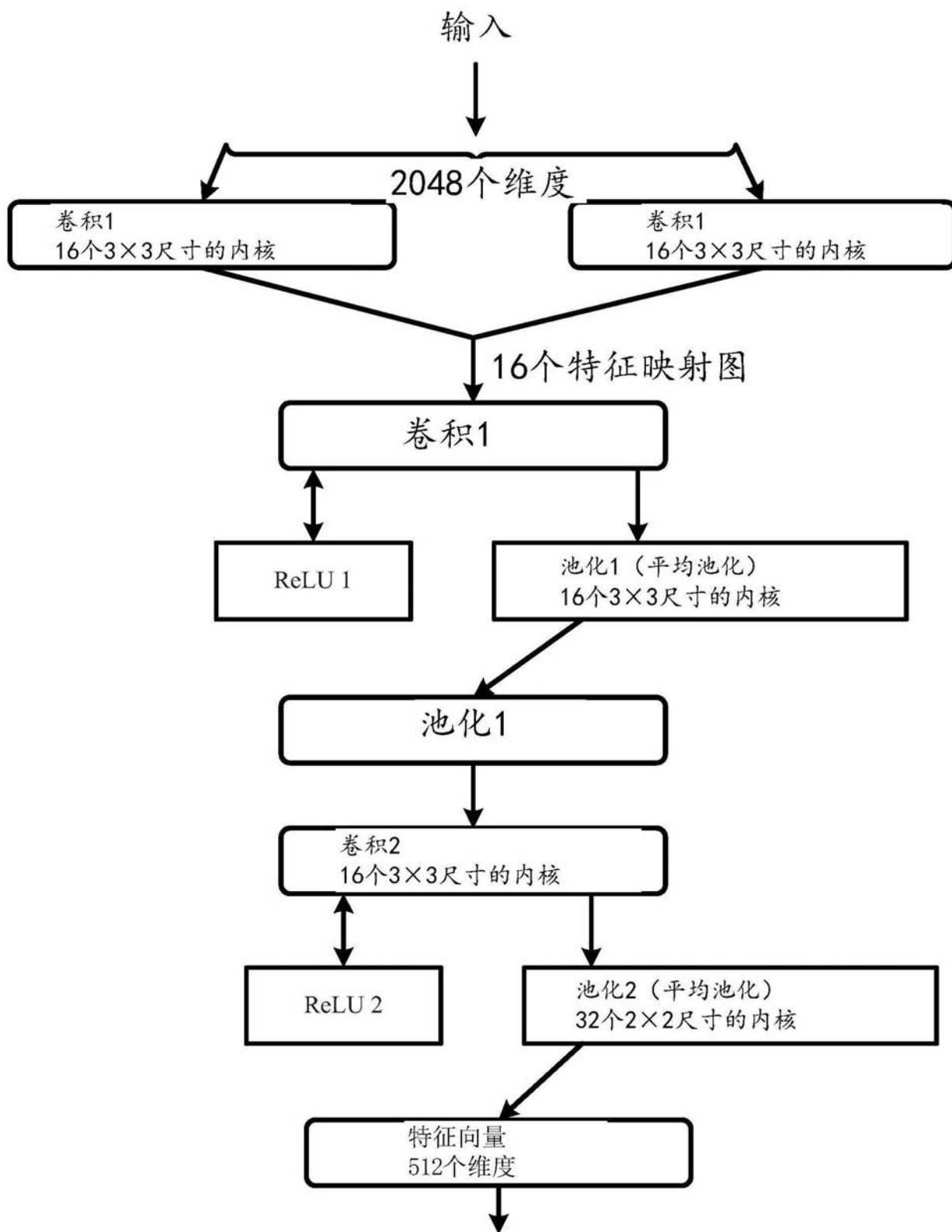


图1F

卷积神经网络中使用的残差连接

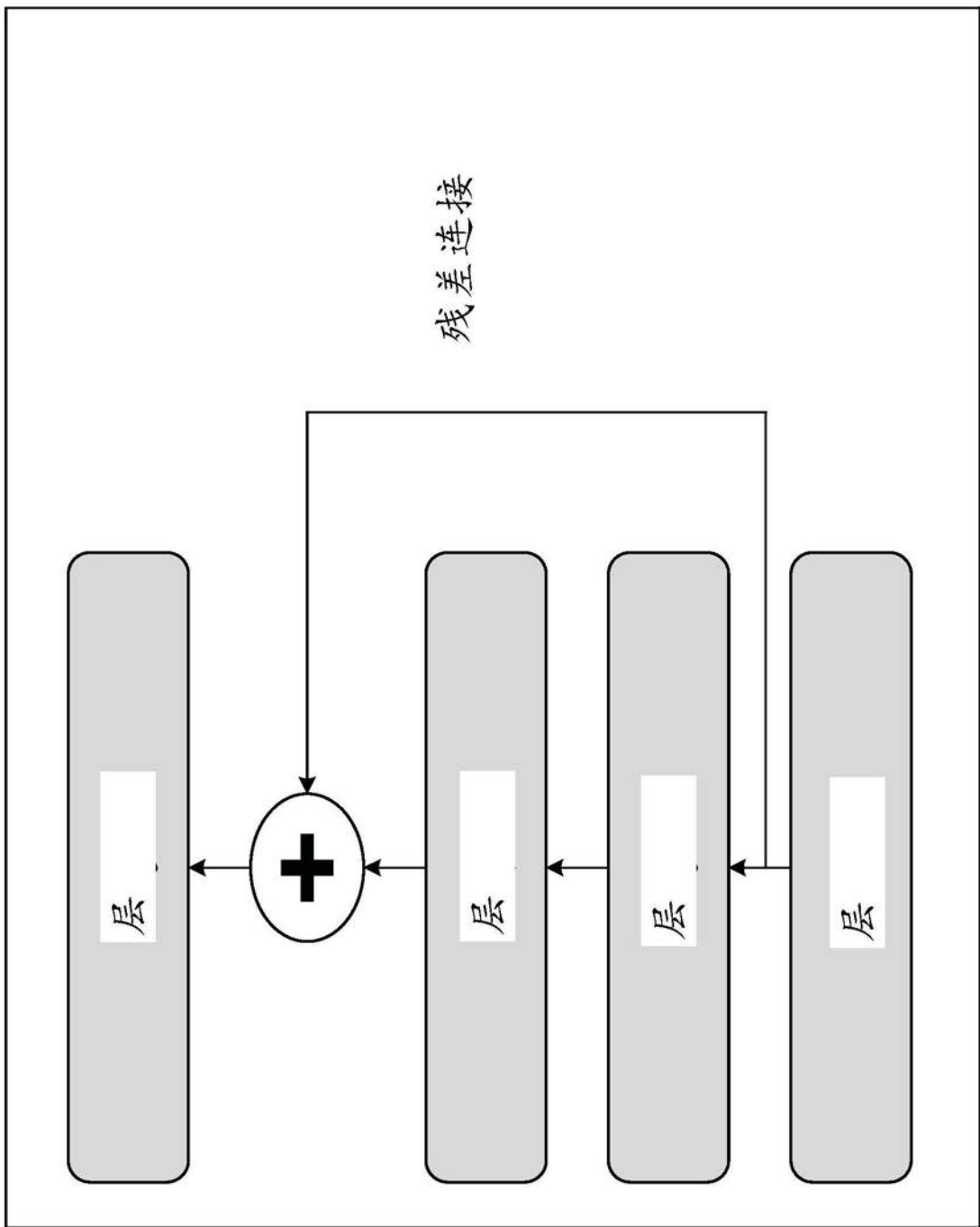
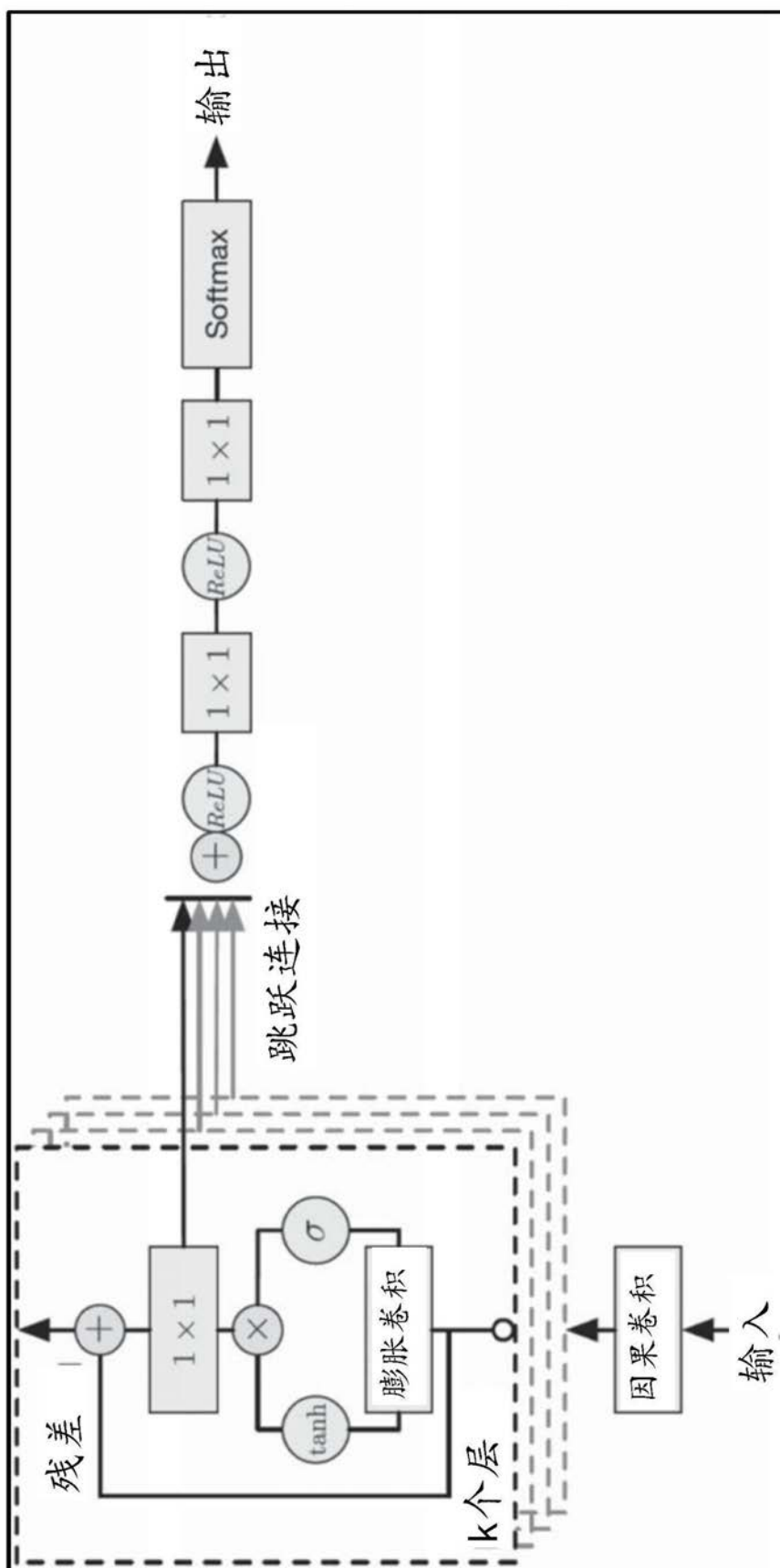


图1G



在卷积神经网络中使用的残差块和跳跃连接

图1H

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1I

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1J

利用卷积神经网络的批归一化后向传递

$$\begin{aligned}\nabla_{\gamma^{(\ell)}} \mathcal{L} &= \sum_{i=1}^n (\nabla_{x^{(\ell+1)}} \mathcal{L})_i \cdot \hat{x}_i^{(\ell)} \\ \nabla_{\beta^{(\ell)}} \mathcal{L} &= \sum_{i=1}^n (\nabla_{x^{(\ell+1)}} \mathcal{L})_i\end{aligned}$$

图1K

卷积层中的批归一化

```
conv_model.add(layers.Conv2D(32, 3, activation='relu')) ← 卷积层之后
conv_model.add(layers.BatchNormalization())

dense_model.add(layers.Dense(32, activation='relu')) ← 密度层之后
dense_model.add(layers.BatchNormalization())
```

图1L

卷积神经网络中使用的1D卷积

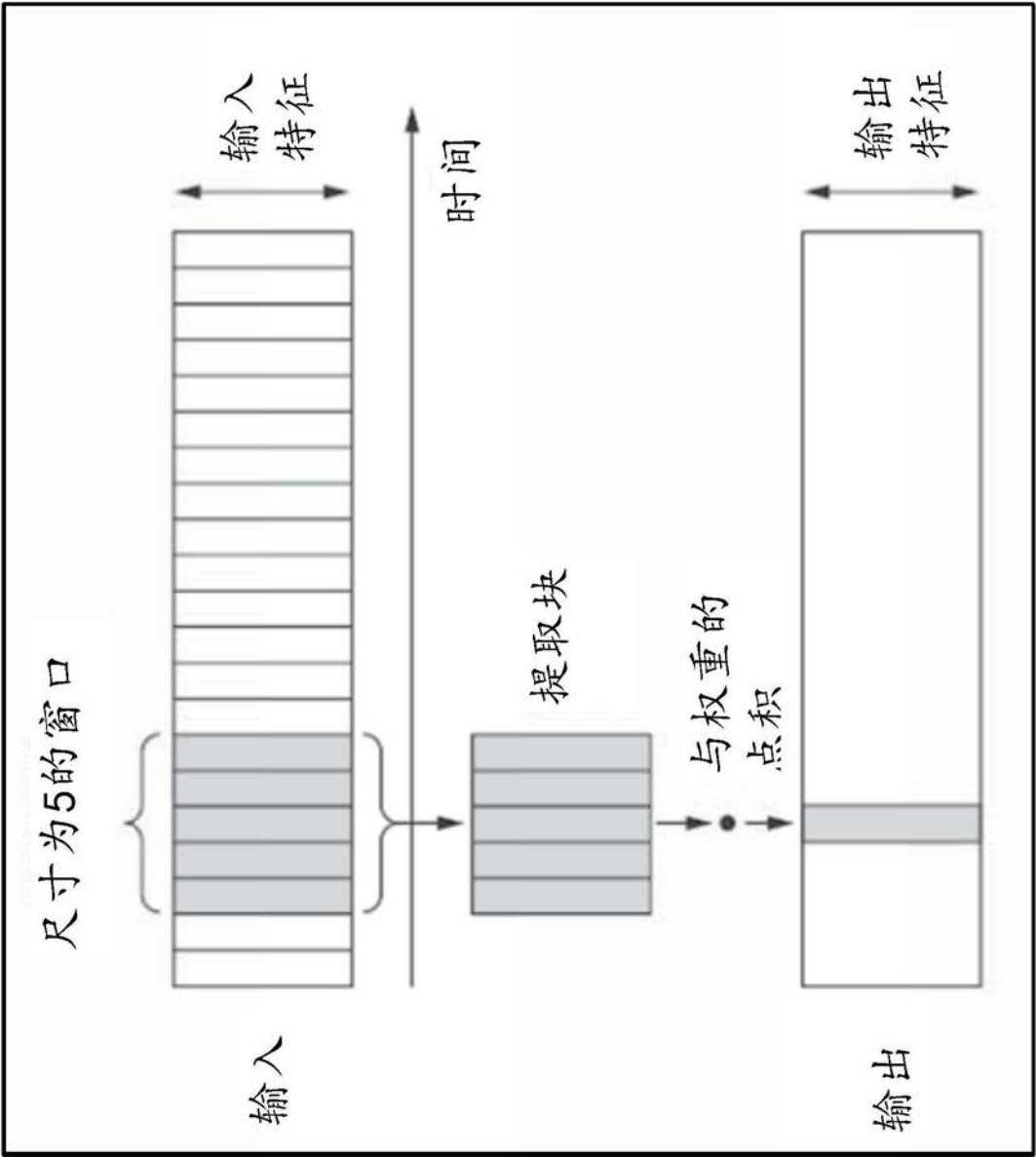


图1M

卷积神经网络中的全局平均池化 (GAP)

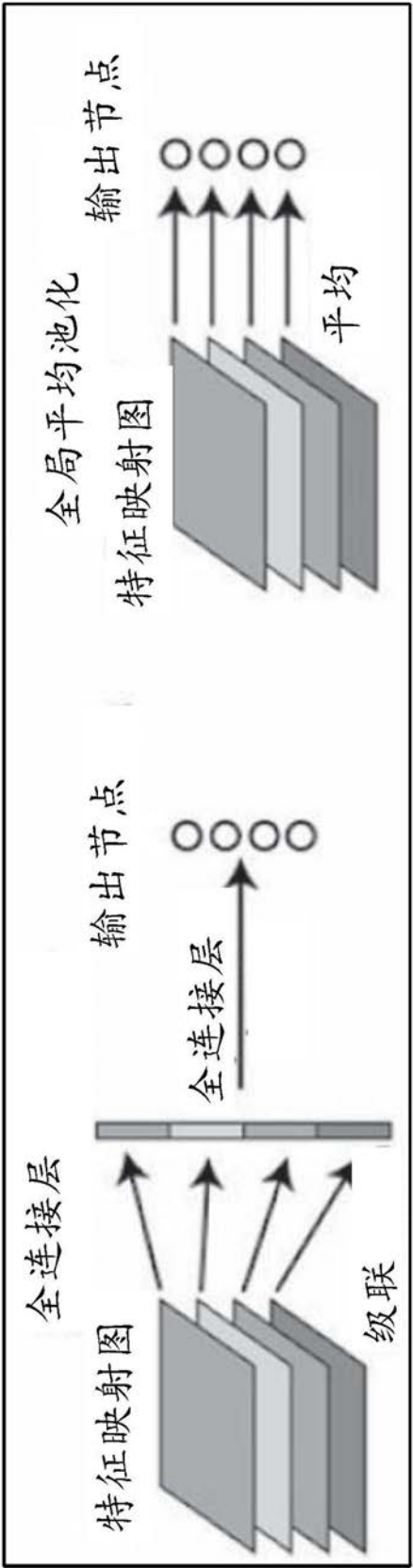
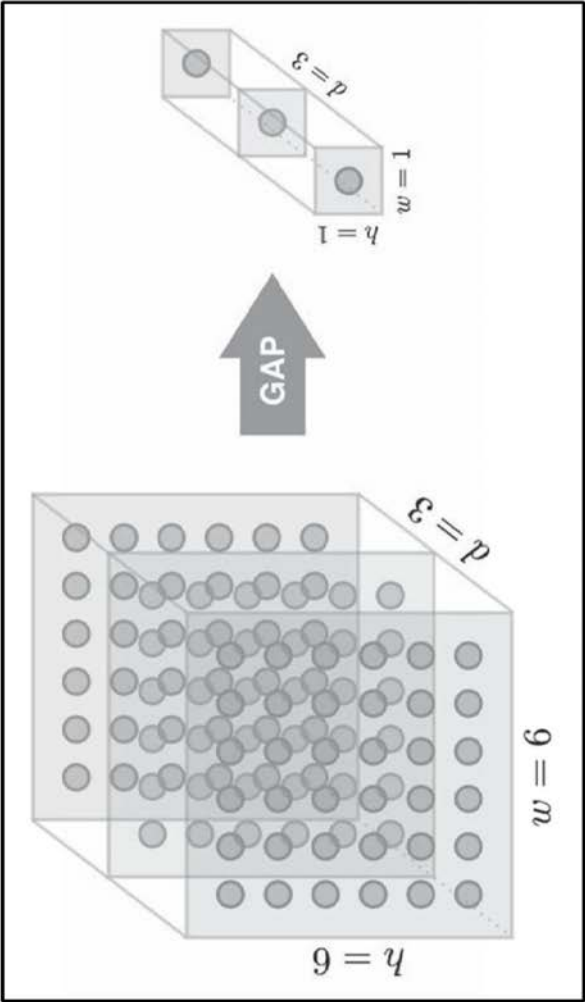


图1N

膨胀卷积

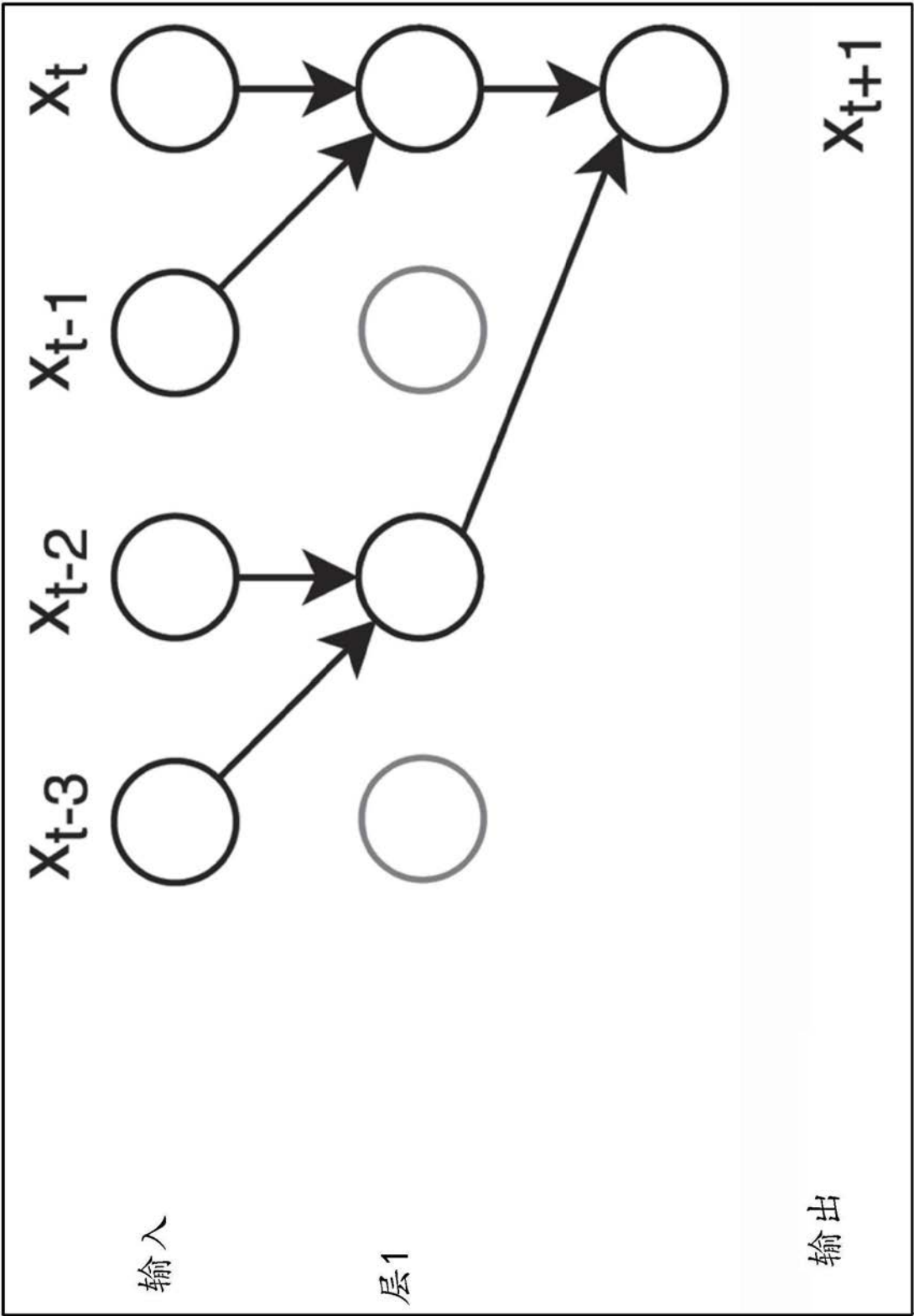


图10

堆叠的膨胀卷积

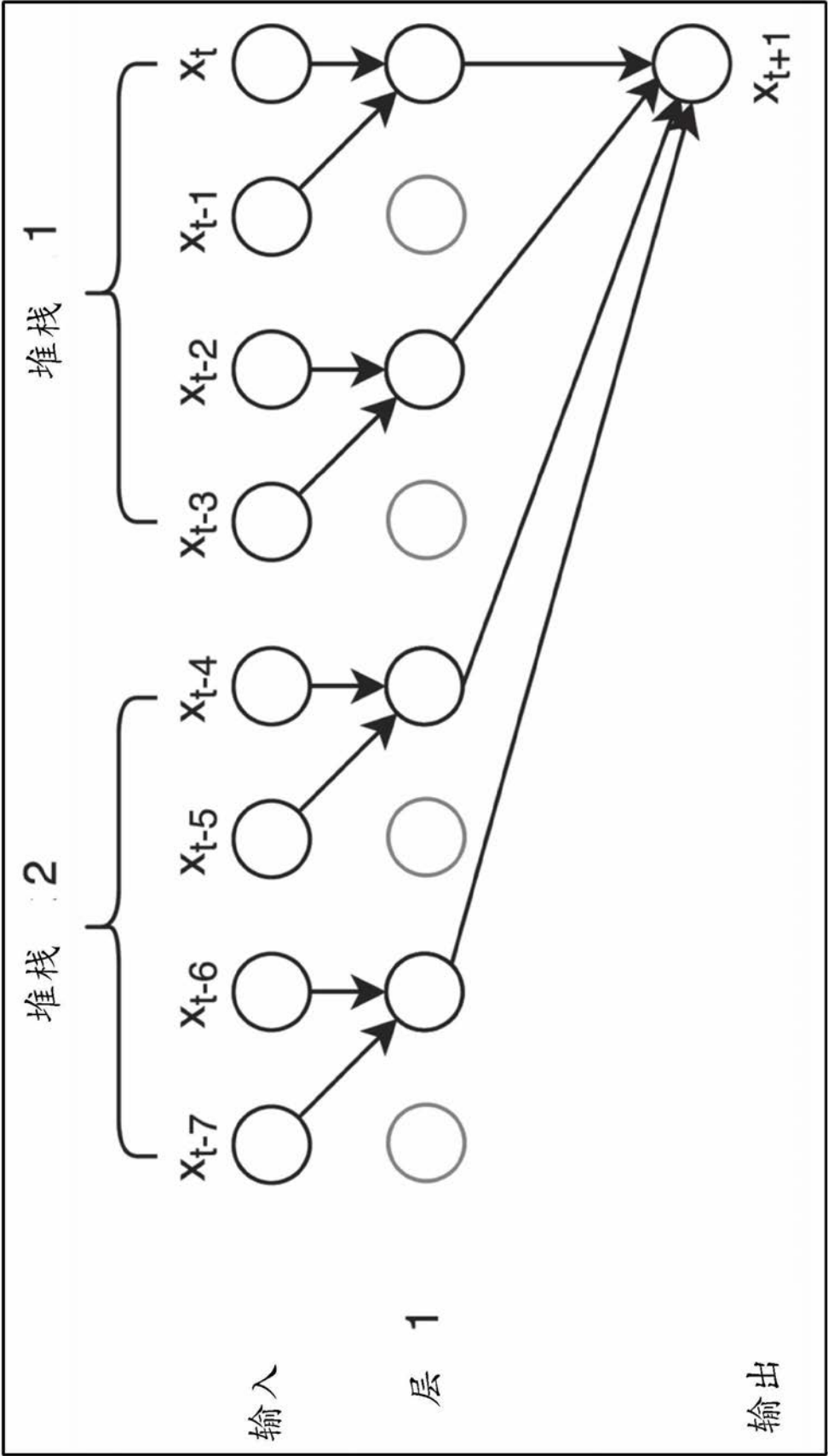


图1P

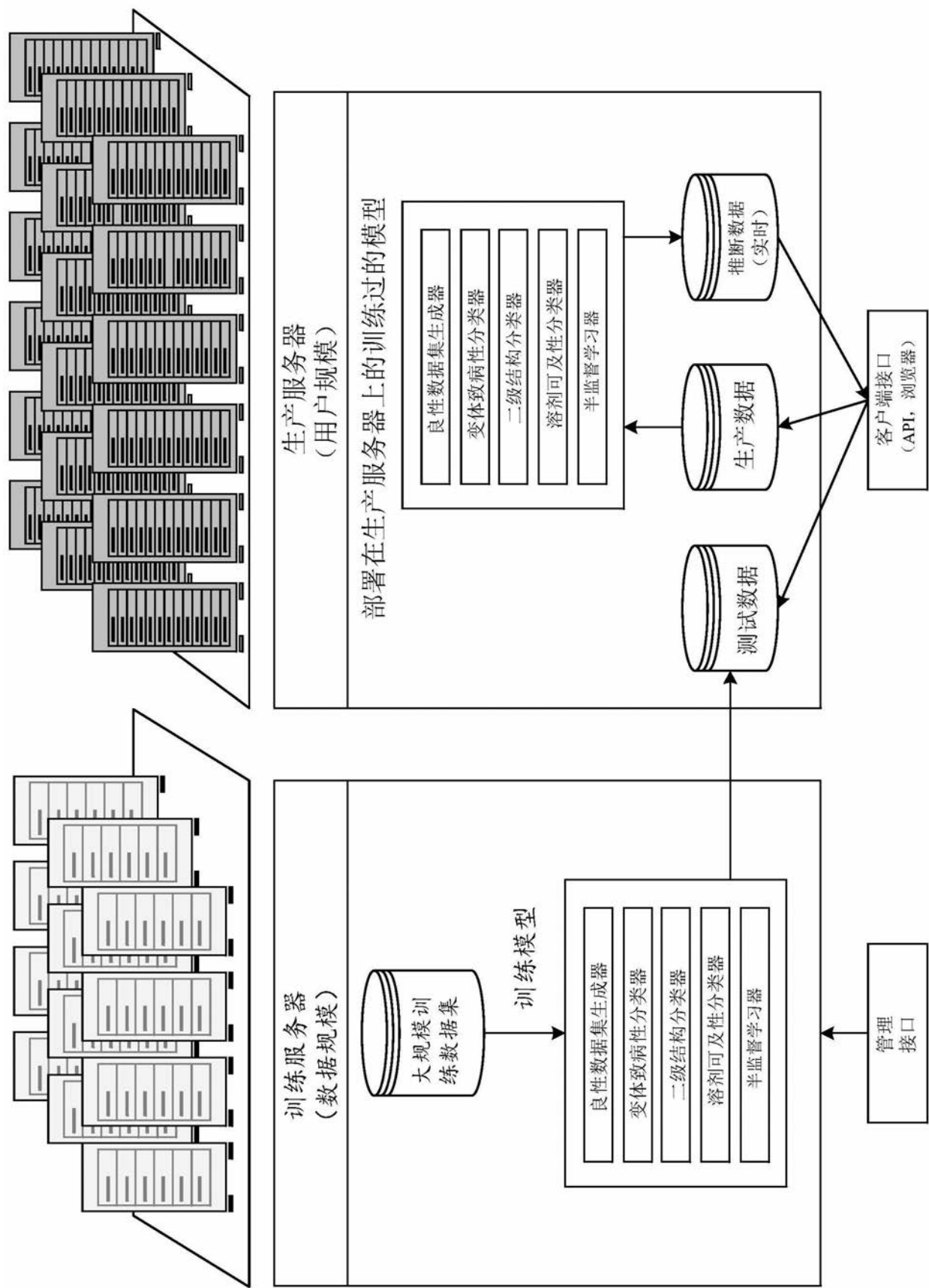


图1Q

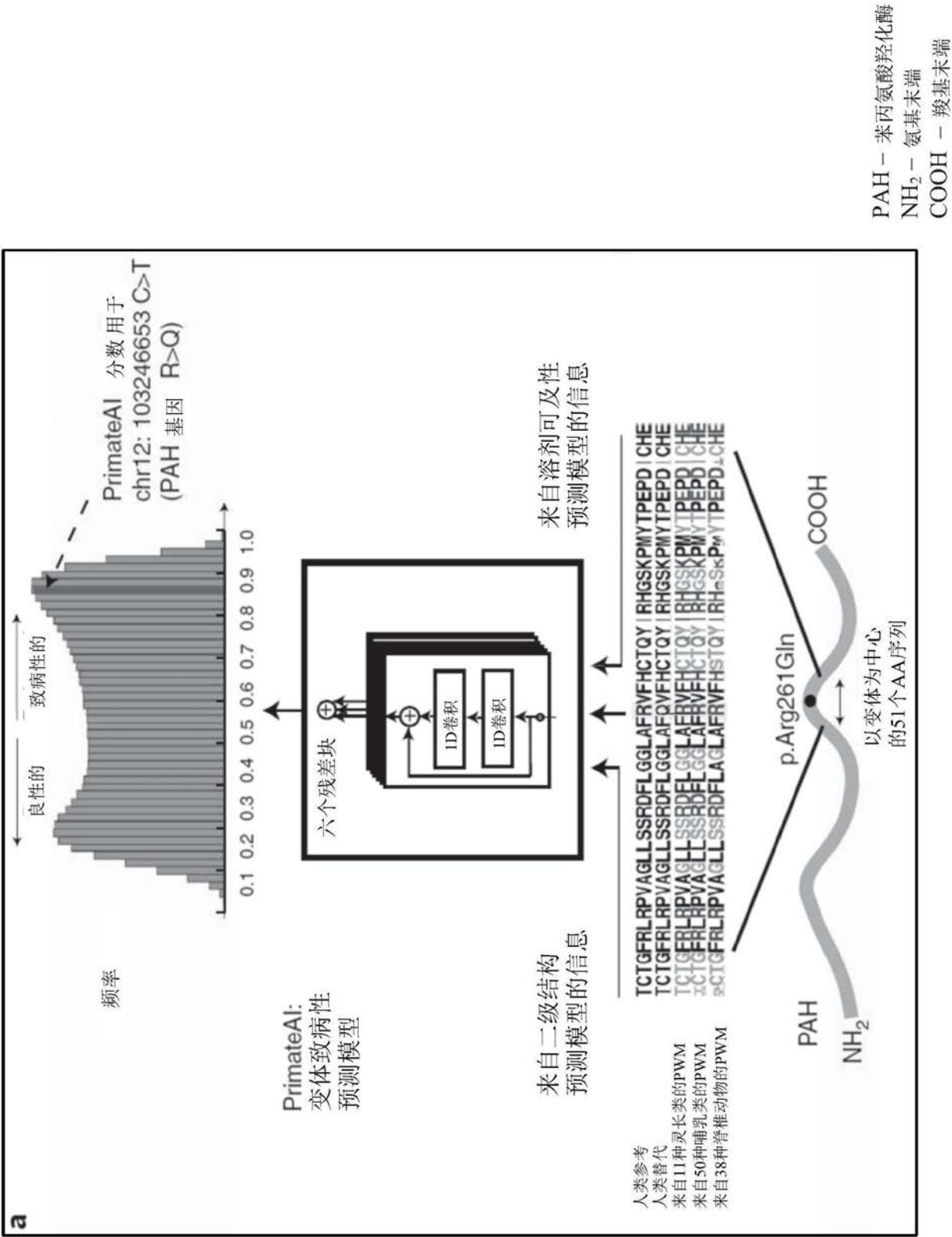


图2

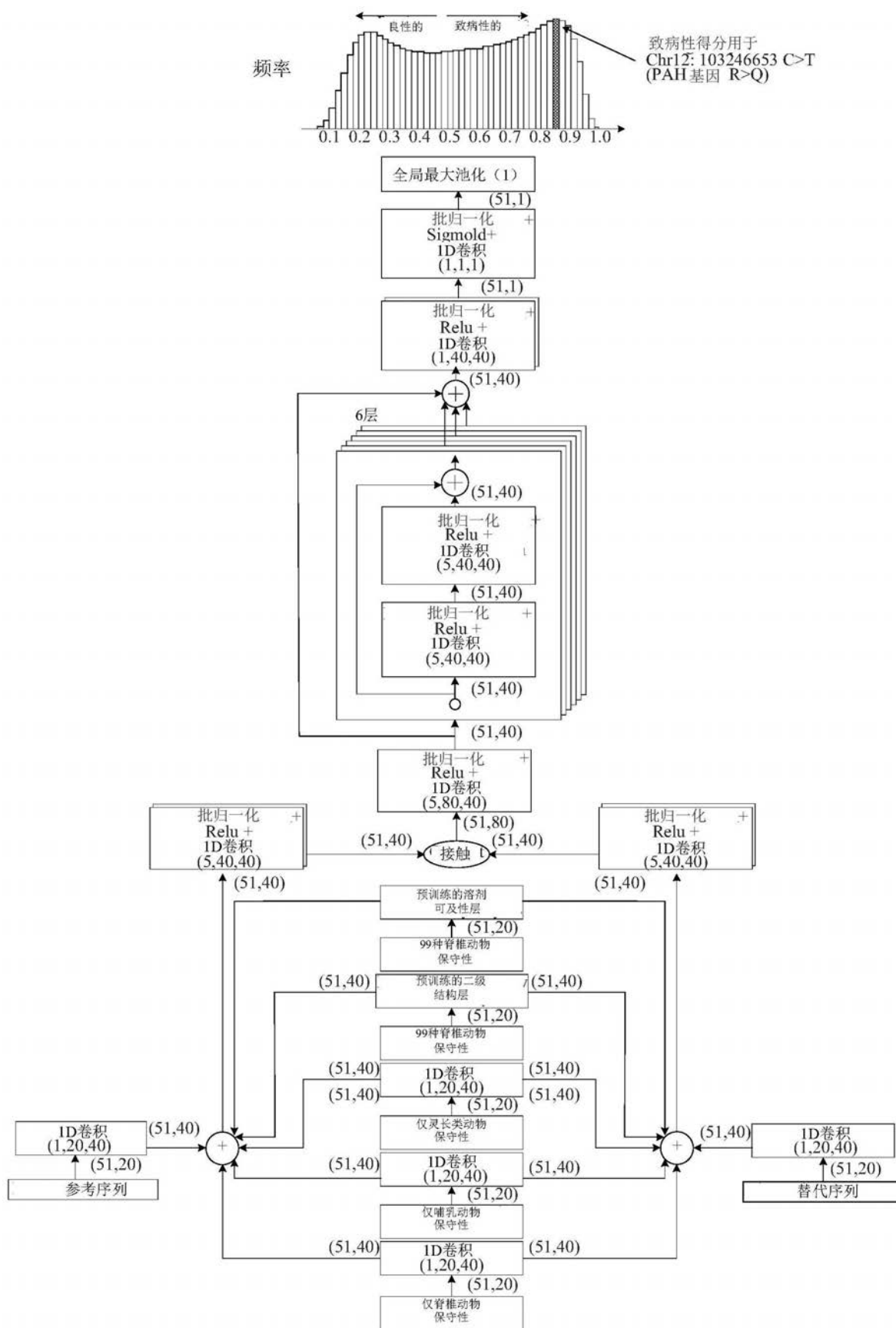


图3

层	层的输入	类型	内核数量 和窗口尺寸	形状	激活
层 1a	参考序列	卷积1D	40,1	(51,40)	线性
层 1b	替代序列	卷积1D	40,1	(51,40)	线性
层 1c	灵长类动物保守性	卷积1D	40,1	(51,40)	线性
层 1d	哺乳动物保守性	卷积1D	40,1	(51,40)	线性
层 1e	脊椎动物保守性	卷积1D	40,1	(51,40)	线性
预先训练的 二级结构 预测模型 (可训练网络)	灵长类动物+ 哺乳动物+ 脊椎动物保守性	-	-	(51,40)	-
预先训练的 溶剂可及性模型 (可训练网络)	灵长类动物+ 哺乳动物+ 脊椎动物保守性	-	-	(51,40)	-
层 1f	预先训练的结构 预测模型 输出	卷积1D	40,1	(51,40)	线性
层 1g	预先训练的溶剂 可及性模型 输出	卷积1D	40,1	(51,40)	线性
合并 1a	层 1a, 层 1c, 层 1d, 层 1e, 层 1f, 层 1g	加	-	(51,40)	-
合并 1b	层 1b, 层 1c, 层 1d, 层 1e, 层 1f, 层 1g	加	-	(51,40)	-
层 2a	合并 1a	批归一化	-	(51,40)	-
层 3a	层 2a	激活	-	(51,40)	Relu
层 4a	层 3a	卷积1D	40,5	(51,40)	线性
层 5a	层 4a	批归一化	-	(51,40)	-
层 6a	层 5a	激活	-	(51,40)	Relu
层 7a	层 6a	卷积1D	40,5	(51,40)	线性
层 2b	合并 1b	批归一化	-	(51,40)	-
层 3b	层 2b	激活	-	(51,40)	Relu
层 4b	层 3b	卷积1D	40,5	(51,40)	线性

图4A

层5b	层4b	批归一化	-	(51,40)	-
层6b	层5b	激活	-	(51,40)	Relu
层7b	层6b	卷积1D	40,5	(51,40)	线性
合并2a	层7a, 层7b	卷积1D	-	(51,80)	-
合并2b	层7a, 层7b	连接	-	(51,80)	-
层8a	合并2a	卷积1D	40,5	(51,40)	线性
层8b	合并2b	卷积1D	40,5	(51,40)	线性
层9	层8a	批归一化	-	(51,40)	-
层10	层9	激活	-	(51,40)	Relu
层11	层10	卷积1D	40,5	(51,40)	线性
层12	层11	批归一化	-	(51,40)	-
层13	层12	激活	-	(51,40)	Relu
层14	层13	卷积1D	40,5	(51,40)	线性
合并3	层8a, 层14	加	-	(51,40)	-
层15	合并3	批归一化	-	(51,40)	-
层16	层15	激活	-	(51,40)	Relu
层17	层16	卷积1D	40,5	(51,40)	线性
层18	层17	批归一化	-	(51,40)	-
层19	层18	激活	-	(51,40)	Relu
层20	层19	卷积1D	40,5	(51,40)	线性
合并4	合并3, 层20	加	-	(51,40)	-
层21	合并4	卷积1D	40,1	(51,40)	线性
层22	合并4	批归一化	-	(51,40)	-
层23	层22	激活	-	(51,40)	Relu
层24	层23	卷积1D	40,5	(51,40)	线性
层25	层24	批归一化	-	(51,40)	-
层26	层25	激活	-	(51,40)	Relu
层27	层26	卷积1D	40,5	(51,40)	线性
合并5	合并4, 层27	加	-	(51,40)	-
层28	合并5	批归一化	-	(51,40)	-
层29	层28	激活	-	(51,40)	Relu
层30	层29	卷积1D	40,5	(51,40)	线性
层31	层30	批归一化	-	(51,40)	-
层32	层31	激活	-	(51,40)	Relu

图4B

层 33	层 32	卷积1D	40,5	(51,40)	线性
合并 6	合并 5, 层 33	加	-	(51,40)	-
层 34	合并 6	卷积1D	40,1	(51,40)	线性
层 35	合并 6	批归一化	-	(51,40)	-
层 36	层 35	激活	-	(51,40)	Relu
层 37	层 36	卷积1D	40,5	(51,40)	线性
层 38	层 37	批归一化	-	(51,40)	-
层 39	层 38	激活	-	(51,40)	Relu
层 40	层 39	卷积1D	40,5	(51,40)	线性
合并 7	合并 6, 层 40	加	-	(51,40)	-
层 41	合并 7	批归一化	-	(51,40)	-
层 42	层 41	激活	-	(51,40)	Relu
层 43	层 42	卷积1D	40,5	(51,40)	线性
层 44	层 43	批归一化	-	(51,40)	-
层 45	层 44	激活	-	(51,40)	Relu
层 46	层 45	卷积1D	40,5	(51,40)	线性
合并 8	合并 7, 层 46	加	-	(51,40)	-
层 47	合并 8	卷积1D	40,1	(51,40)	线性
合并 9	层 8b, 层 21, 层 34, 层 47	加	-	(51,40)	-
层 48	合并 10	批归一化	-	(51,40)	-
层 49	层 48	激活	-	(51,40)	Relu
层 50	层 49	卷积1D	40,1	(51,40)	线性
层 51	层 50	批归一化	-	(51,40)	-
层 52	层 51	激活	-	(51,40)	Relu
层 53	层 52	卷积1D	40,1	(51,40)	线性
合并 10	合并 9, 层 53	加	-	(51,40)	-
层 54	合并 10	卷积1D	1,1	(51,1)	Sigmoid
输出层	层 54	全局最大池化1D	-	1	-

图4C

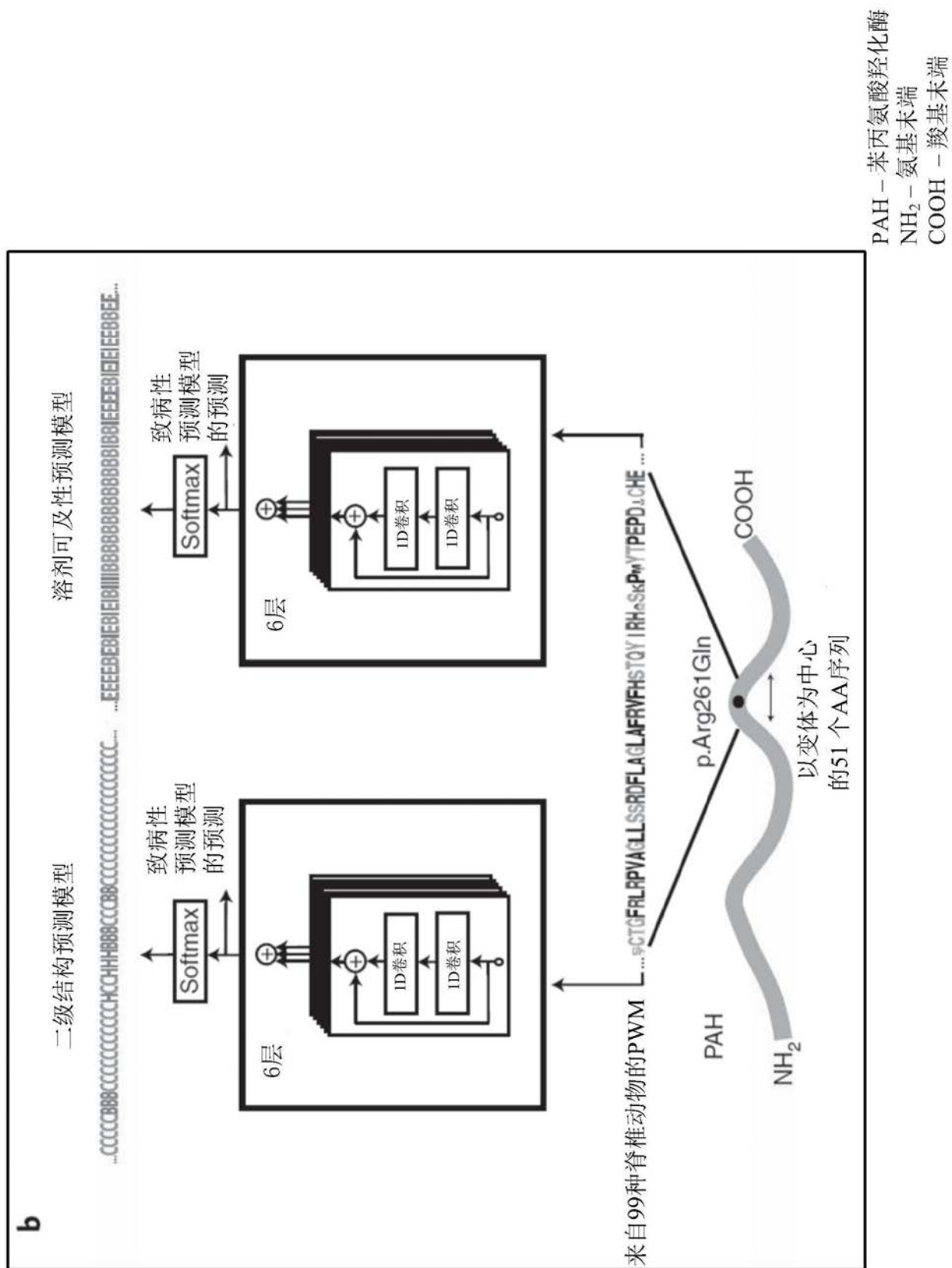


图5

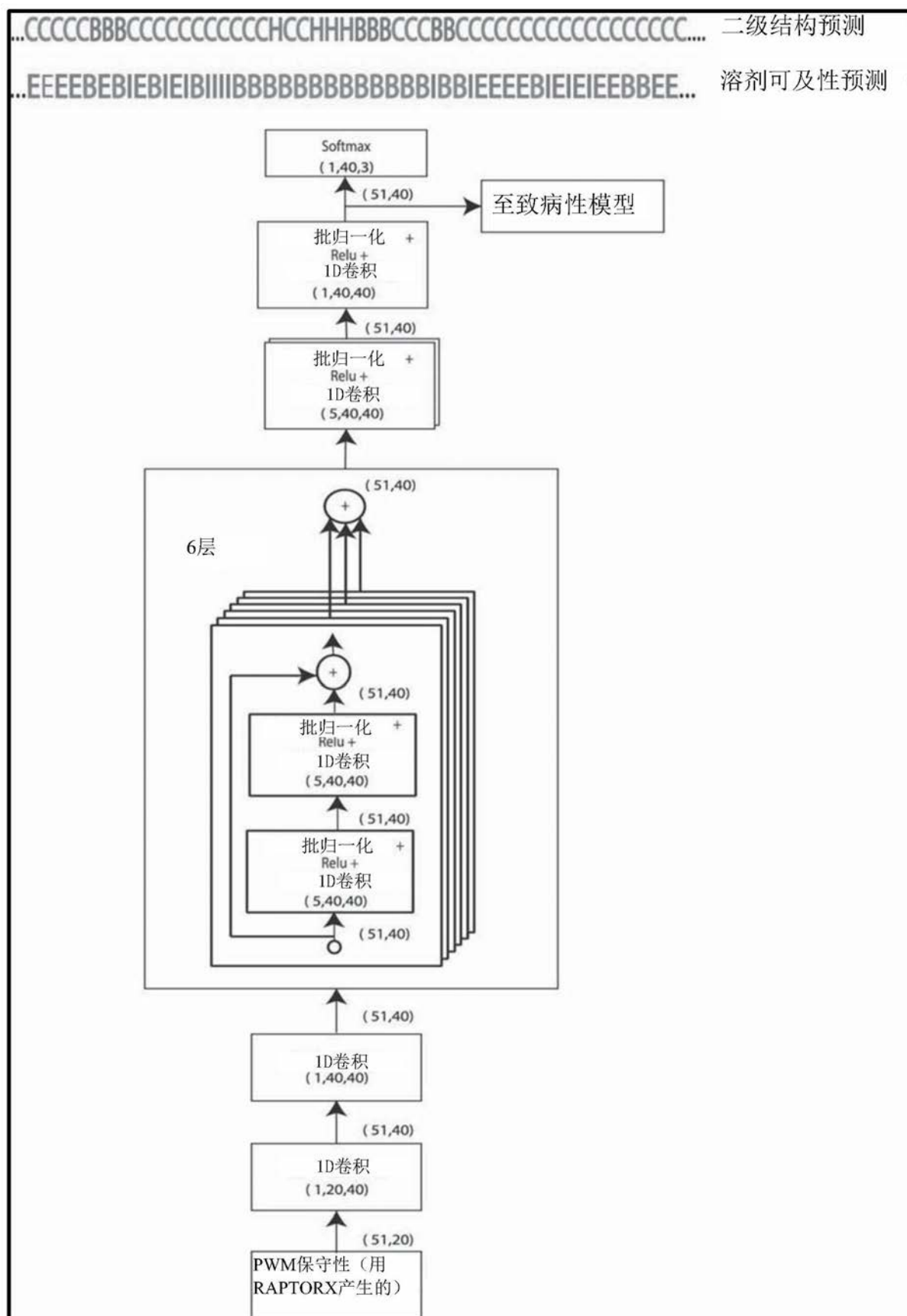


图6

层	层的输入	类型	内核数量 和窗口 尺寸	形状	激活
层 1a	输入 PSFM	卷积1D	40,1	(51,40)	线性
层 2a	层 1a	卷积1D	40,1	(51,40)	线性
层 1b	输入 PSFM	卷积1D	40,1	(51,40)	线性
层 2b	层 1b	卷积1D	40,1	(51,40)	线性
层 3	层 2a	批归一化	-	(51,40)	-
层 4	层 3	激活	-	(51,40)	Relu
层 5	层 4	卷积1D	40,5	(51,40)	线性
层 6	层 5	批归一化	-	(51,40)	-
层 7	层 6	激活	-	(51,40)	Relu
层 8	层 7	卷积1D	40,5	(51,40)	线性
合并 1	层 2a, 层 9	加	-	(51,40)	-
层 10	合并 1	批归一化	-	(51,40)	-
层 11	层 10	激活	-	(51,40)	Relu
层 12	层 11	卷积1D	40,5	(51,40)	线性
层 13	层 12	批归一化	-	(51,40)	-
层 14	层 13	激活	-	(51,40)	Relu
层 15	层 14	卷积1D	40,5	(51,40)	线性
合并 2	合并 1, 层 15	加	-	(51,40)	-
层 16	合并 2	卷积1D	40,1	(51,40)	线性
层 17	合并 2	批归一化	-	(51,40)	-
层 18	层 17	激活	-	(51,40)	Relu
层 19	层 18	卷积1D	40,5	(51,40)	线性
层 20	层 19	批归一化	-	(51,40)	-
层 21	层 20	激活	-	(51,40)	Relu
层 22	层 21	卷积1D	40,5	(51,40)	线性
合并 3	合并 2, 层 22	加	-	(51,40)	-

图7A

层 23	合并 3	批归一化	-	(51,40)	-
层 24	层 23	激活	-	(51,40)	Relu
层 25	层 24	卷积ID	40,5	(51,40)	线性
层 26	层 25	批归一化	-	(51,40)	-
层 27	层 26	激活	-	(51,40)	Relu
层 28	层 27	卷积ID	40,5	(51,40)	线性
合并4	合并 3, 层 28	加	-	(51,40)	-
层 29	合并 4	卷积ID	40,1	(51,40)	线性
层 30	合并 4	批归一化	-	(51,40)	-
层 31	层 30	激活	-	(51,40)	Relu
层 32	层 31	卷积ID	40,5	(51,40)	线性
层 33	层 32	批归一化	-	(51,40)	-
层 34	层 33	激活	-	(51,40)	Relu
层 35	层 34	卷积ID	40,5	(51,40)	线性
合并5	合并 4, 层 35	加	-	(51,40)	-
层 36	合并 5	批归一化	-	(51,40)	-
层 37	层 36	激活	-	(51,40)	Relu
层 38	层 37	卷积ID	40,5	(51,40)	线性
层 39	层 38	批归一化	-	(51,40)	-
层 40	层 39	激活	-	(51,40)	Relu
层 41	层 40	卷积ID	40,5	(51,40)	线性
合并 6	合并 5, 层 41	加	-	(51,40)	-
层 42	合并 6	卷积ID	40,1	(51,40)	线性
合并 7	层 2b, 层 16, 层 29, 层 42	加	-	(51,40)	-
层 43	合并 7	批归一化	-	(51,40)	-
层 44	层 43	激活	-	(51,40)	Relu
层 45	合并 44	卷积ID	40,1	(51,40)	线性
层 46	层 45	批归一化	-	(51,40)	-
层 47	层 46	激活	-	(51,40)	Relu
层 48	层 47	卷积ID	40,1	(51,40)	线性
合并 8	合并 7, 层 48	加	-	(51,40)	-
输出层	合并 8	卷积ID	3,1	(51,3)	Softmax

图7B

层	层的输入	类型	内核数量 和窗口 尺寸	形状	激活
层 1a	输入 PSFM	卷积1D	40,1	(51,40)	线性
层 2a	层 1a	卷积1D	40,1	(51,40)	线性
层 1b	输入 PSFM	卷积1D	40,1	(51,40)	线性
层 2b	层 1b	卷积1D	40,1	(51,40)	线性
层 3	层 2a	批归一化	-	(51,40)	-
层 4	层 3	激活	-	(51,40)	Relu
层 5	层 4	卷积1D	40,5	(51,40)	线性
层 6	层 5	批归一化	-	(51,40)	-
层 7	层 6	激活	-	(51,40)	Relu
层 8	层 7	卷积1D	40,5	(51,40)	线性
合并 1	层 2a, 层 9	加	-	(51,40)	-
层 10	合并 1	批归一化	-	(51,40)	-
层 11	层 10	激活	-	(51,40)	Relu
层 12	层 11	卷积1D	40,5	(51,40)	线性
层 13	层 12	批归一化	-	(51,40)	-
层 14	层 13	激活	-	(51,40)	Relu
层 15	层 14	卷积1D	40,5	(51,40)	线性
合并 2	合并 1, 层 15	加	-	(51,40)	-
层 16	合并 2	卷积1D	40,1	(51,40)	线性
层 17	合并 2	批归一化	-	(51,40)	-
层 18	层 17	激活	-	(51,40)	Relu
层 19	层 18	卷积1D	40,5	(51,40)	线性
层 20	层 19	批归一化	-	(51,40)	-
层 21	层 20	激活	-	(51,40)	Relu
层 22	层 21	卷积1D	40,5	(51,40)	线性
合并 3	合并 2, 层 22	加	-	(51,40)	-

图8A

层 23	合并 3	批归一化	-	(51,40)	-
层 24	层 23	激活	-	(51,40)	Relu
层 25	层 24	卷积1D	40,5	(51,40)	线性
层 26	层 25	批归一化	-	(51,40)	-
层 27	层 26	激活	-	(51,40)	Relu
层 28	层 27	卷积1D	40,5	(51,40)	线性
合并 4	合并 3, 层 28	加	-	(51,40)	-
层 29	合并 4	卷积1D	40,1	(51,40)	线性
层 30	合并 4	批归一化	-	(51,40)	-
层 31	层 30	激活	-	(51,40)	Relu
层 32	层 31	卷积1D	40,5	(51,40)	线性
层 33	层 32	批归一化	-	(51,40)	-
层 34	层 33	激活	-	(51,40)	Relu
层 35	层 34	卷积1D	40,5	(51,40)	线性
合并 5	合并 4, 层 35	加	-	(51,40)	-
层 36	合并 5	批归一化	-	(51,40)	-
层 37	层 36	激活	-	(51,40)	Relu
层 38	层 37	卷积1D	40,5	(51,40)	线性
层 39	层 38	批归一化	-	(51,40)	-
层 40	层 39	激活	-	(51,40)	Relu
层 41	层 40	卷积1D	40,5	(51,40)	线性
合并 6	合并 5, 层 41	加	-	(51,40)	-
层 42	合并 6	卷积1D	40,1	(51,40)	线性
合并 7	层 26, 层 29, 层 16, 层 42	加	-	(51,40)	-
层 43	合并 7	批归一化	-	(51,40)	-
层 44	层 43	激活	-	(51,40)	Relu
层 45	合并 44	卷积1D	40,1	(51,40)	线性
层 46	层 45	批归一化	-	(51,40)	-
层 47	层 46	激活	-	(51,40)	Relu
层 48	层 47	卷积1D	40,1	(51,40)	线性
合并 8	合并 7, 层 48	加	-	(51,40)	-
输出层	合并 8	卷积1D	3,1	(51,3)	Softmax

图8B

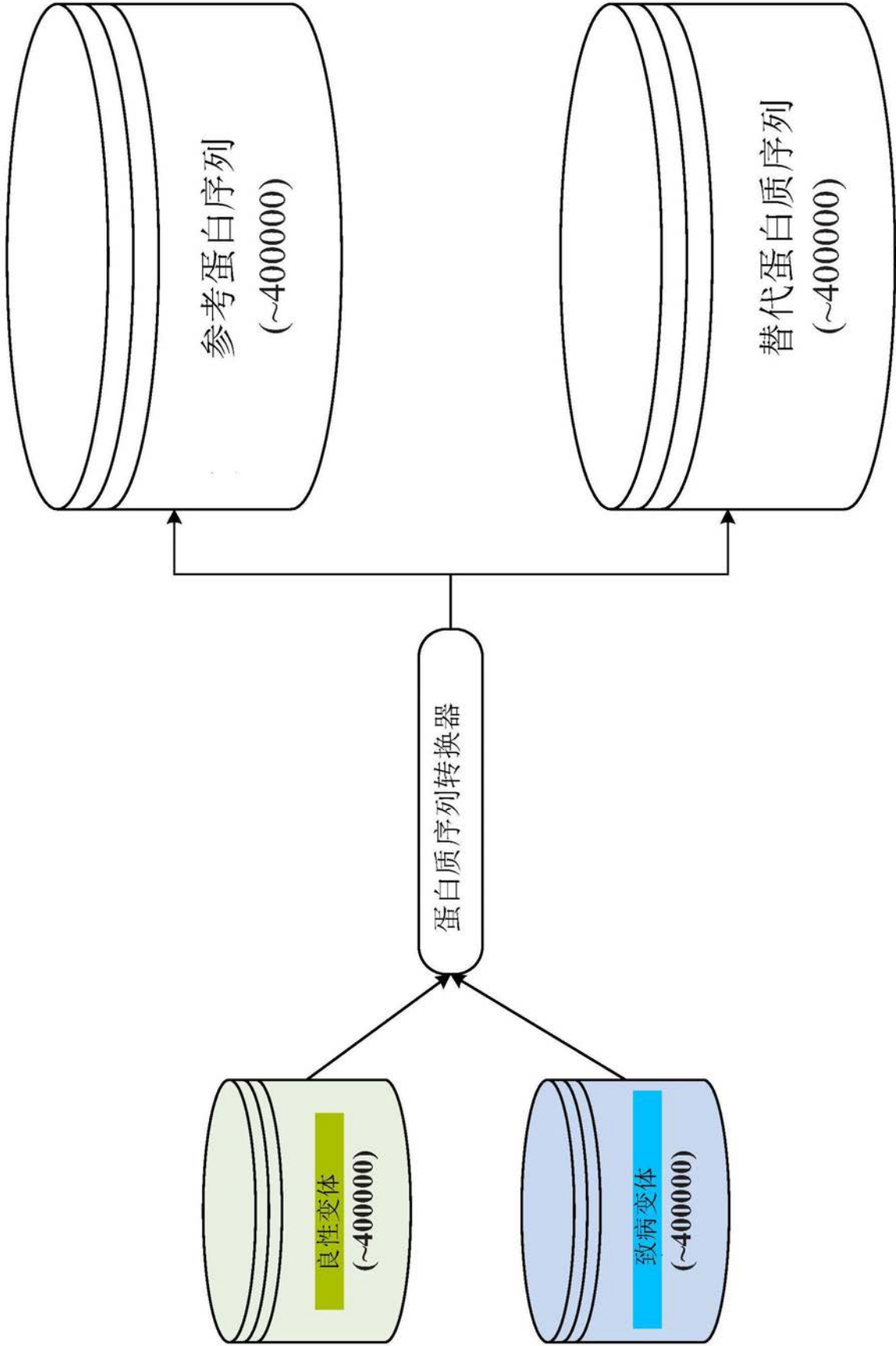


图9

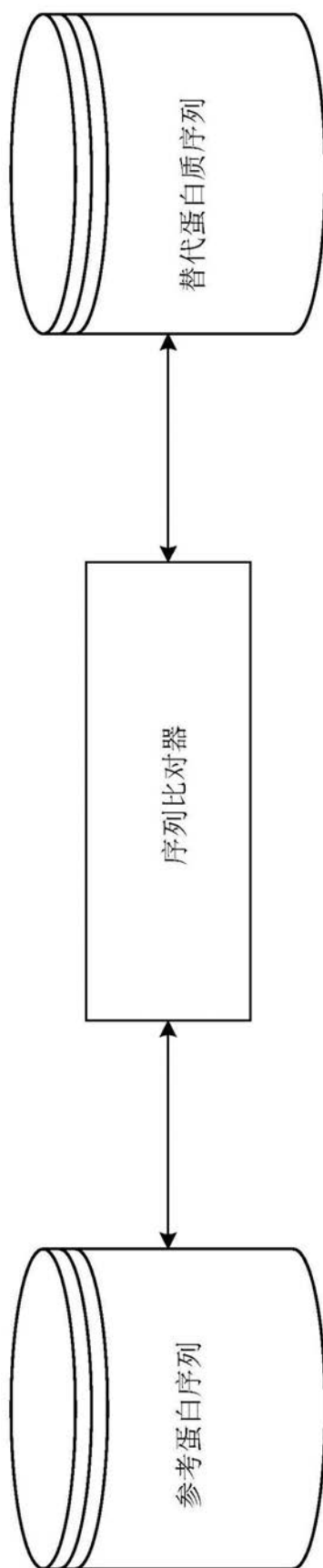


图10

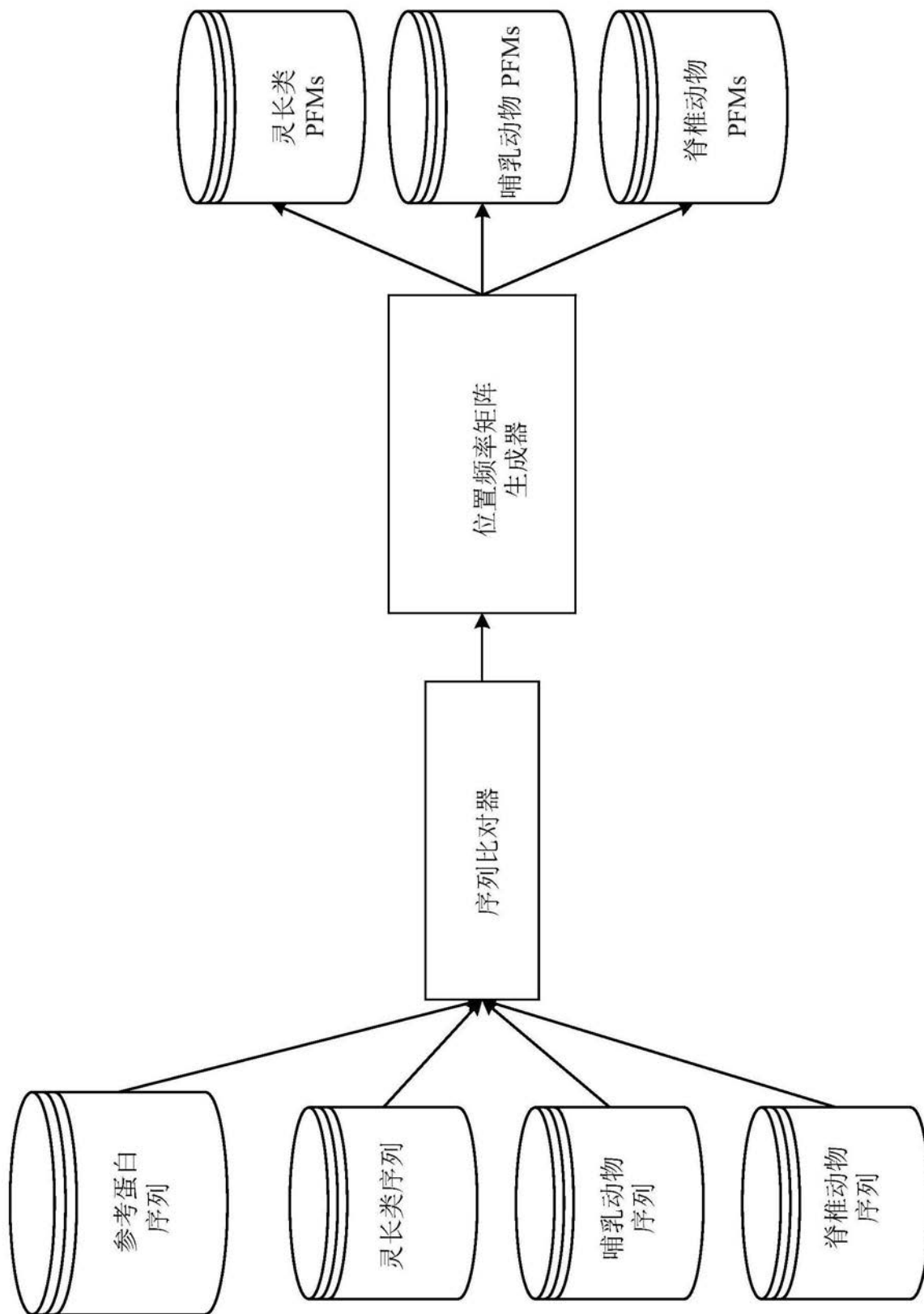


图11

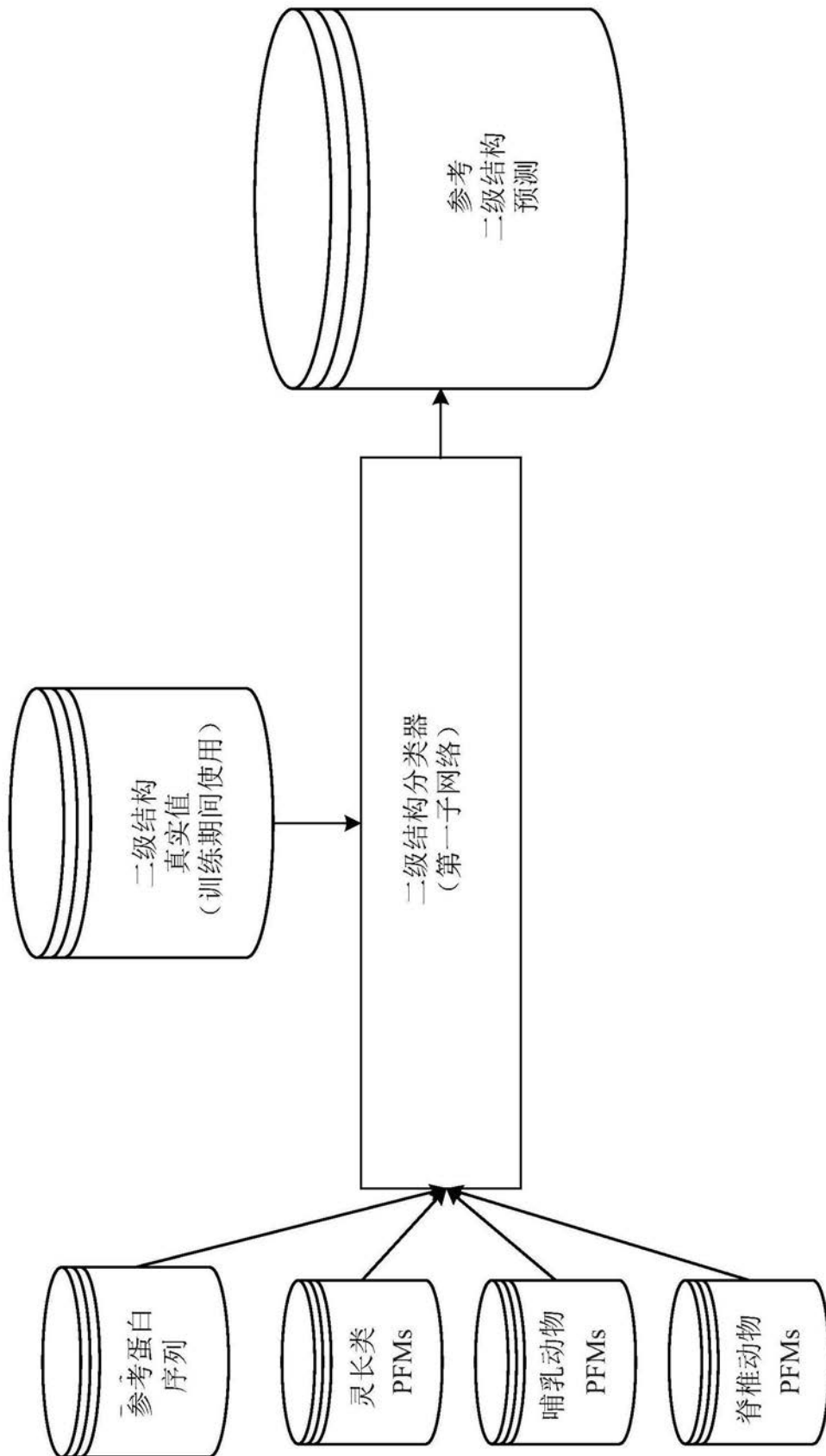


图12

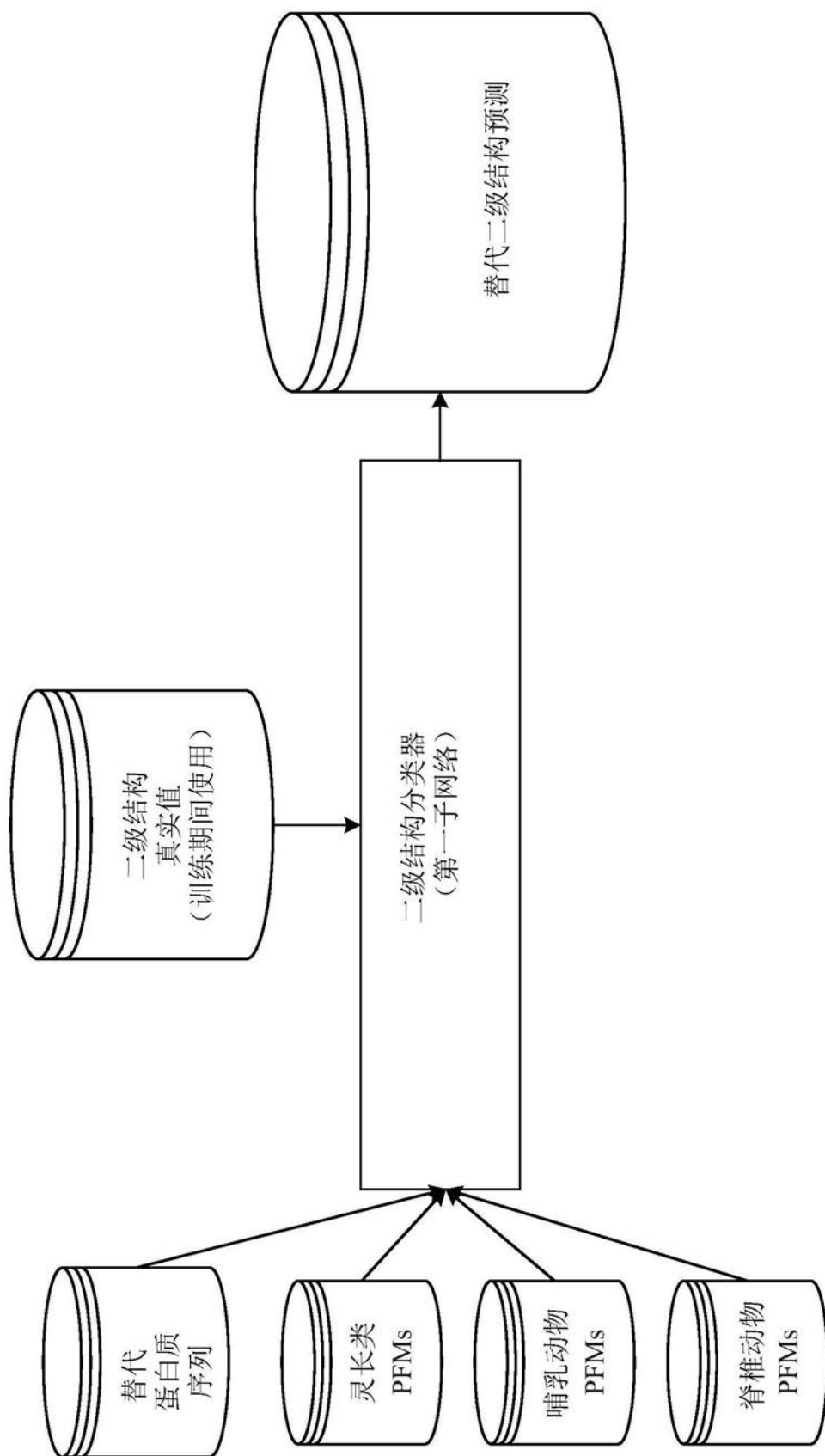


图13

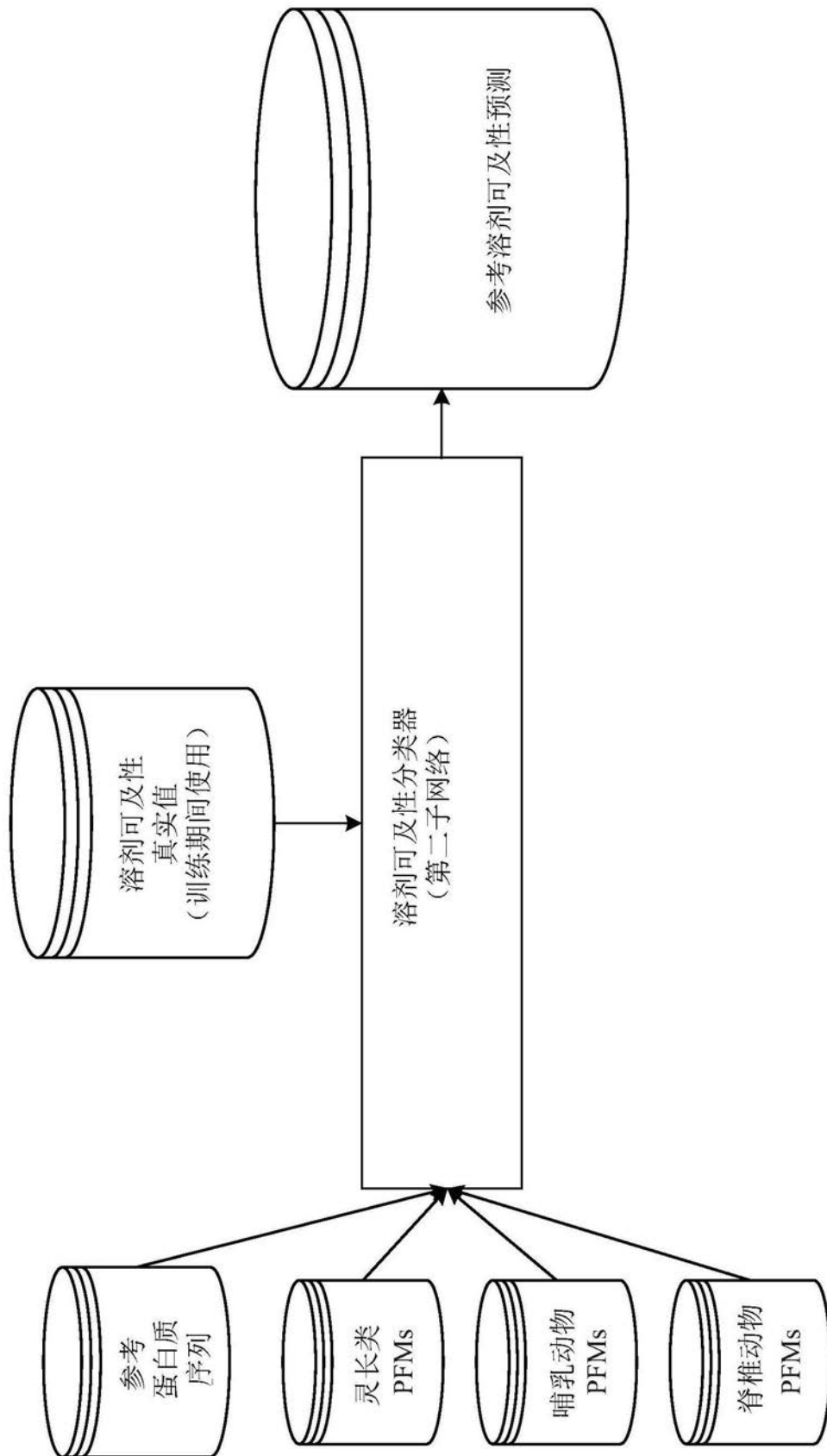


图14

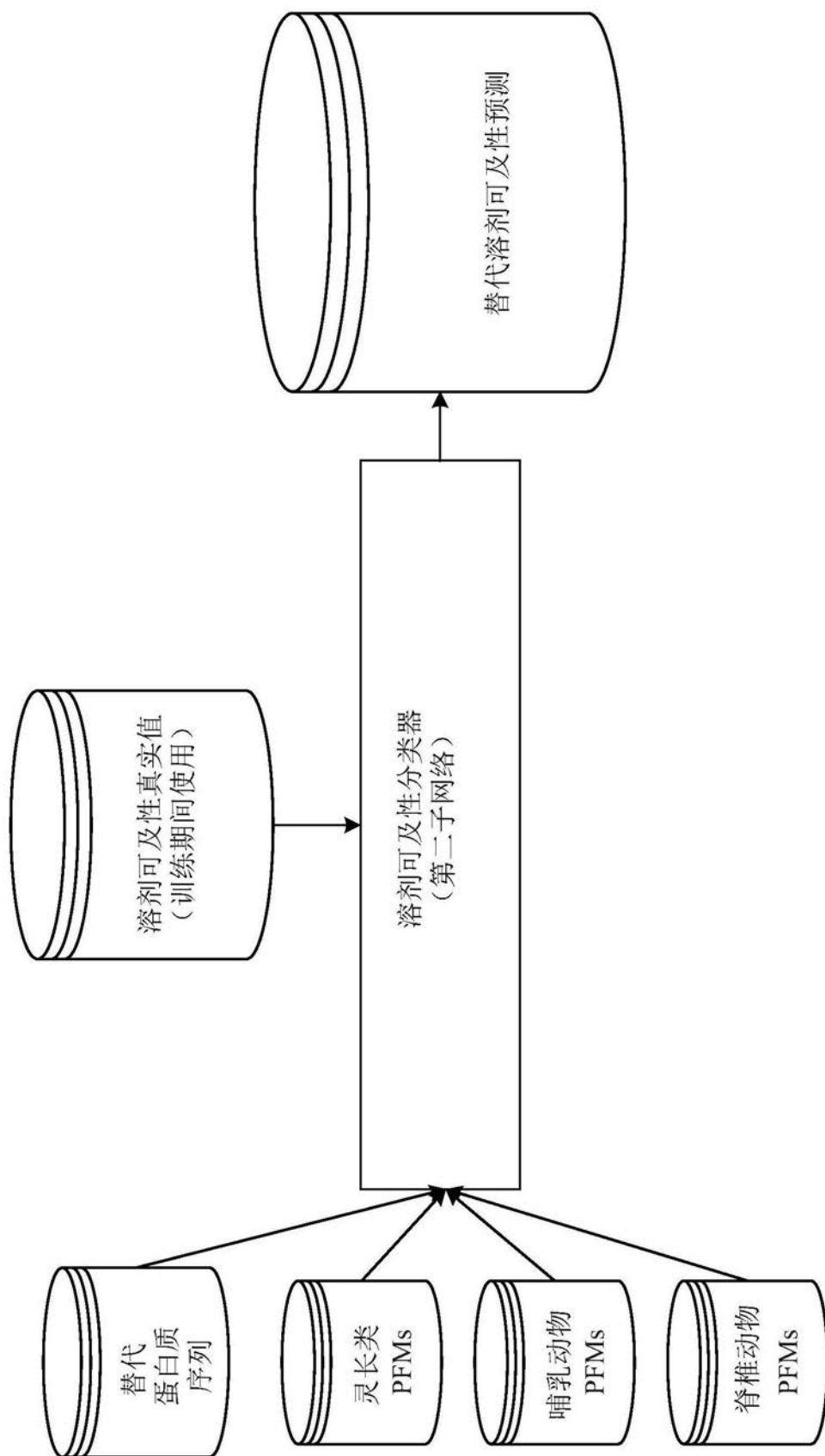


图15

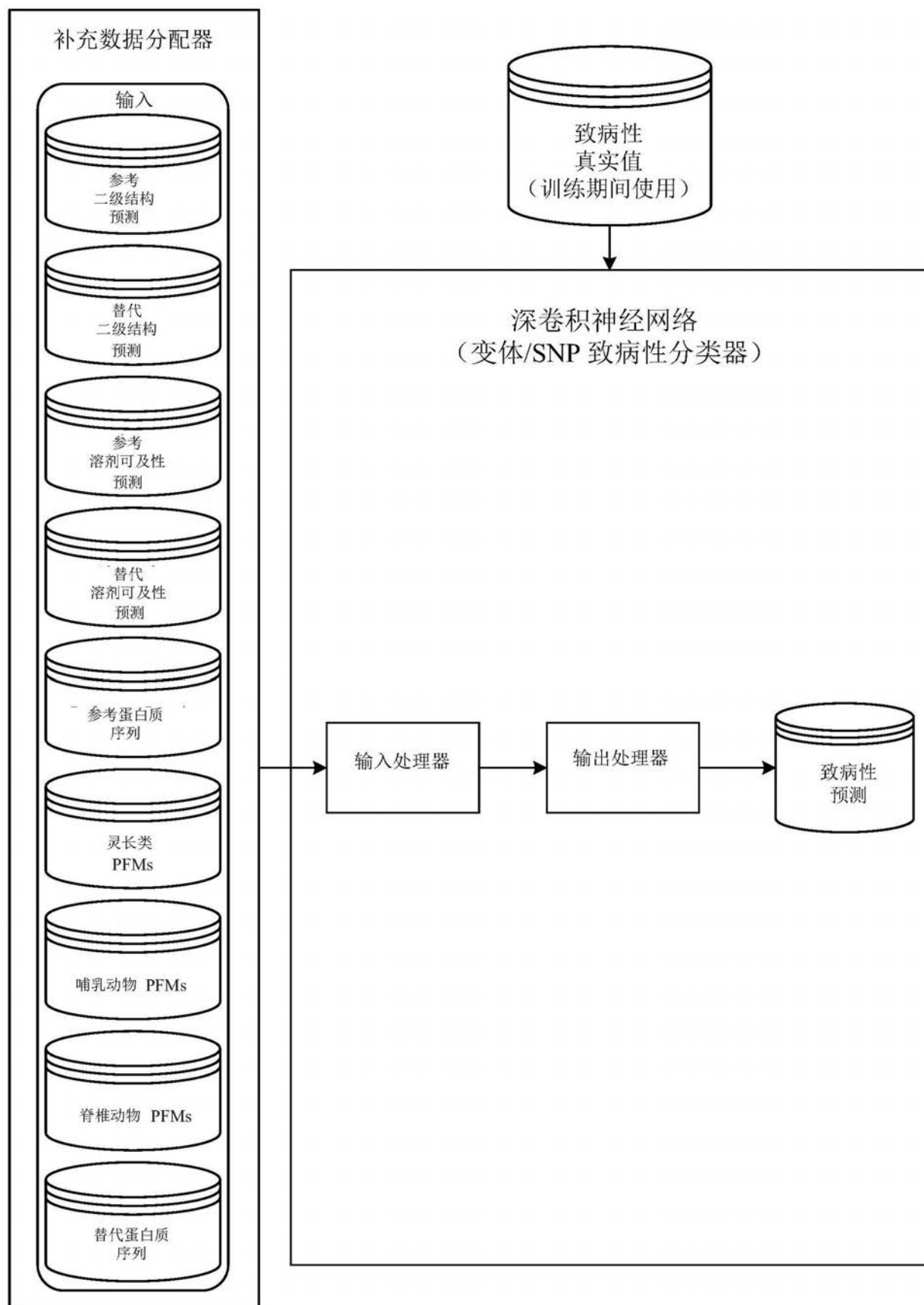


图16

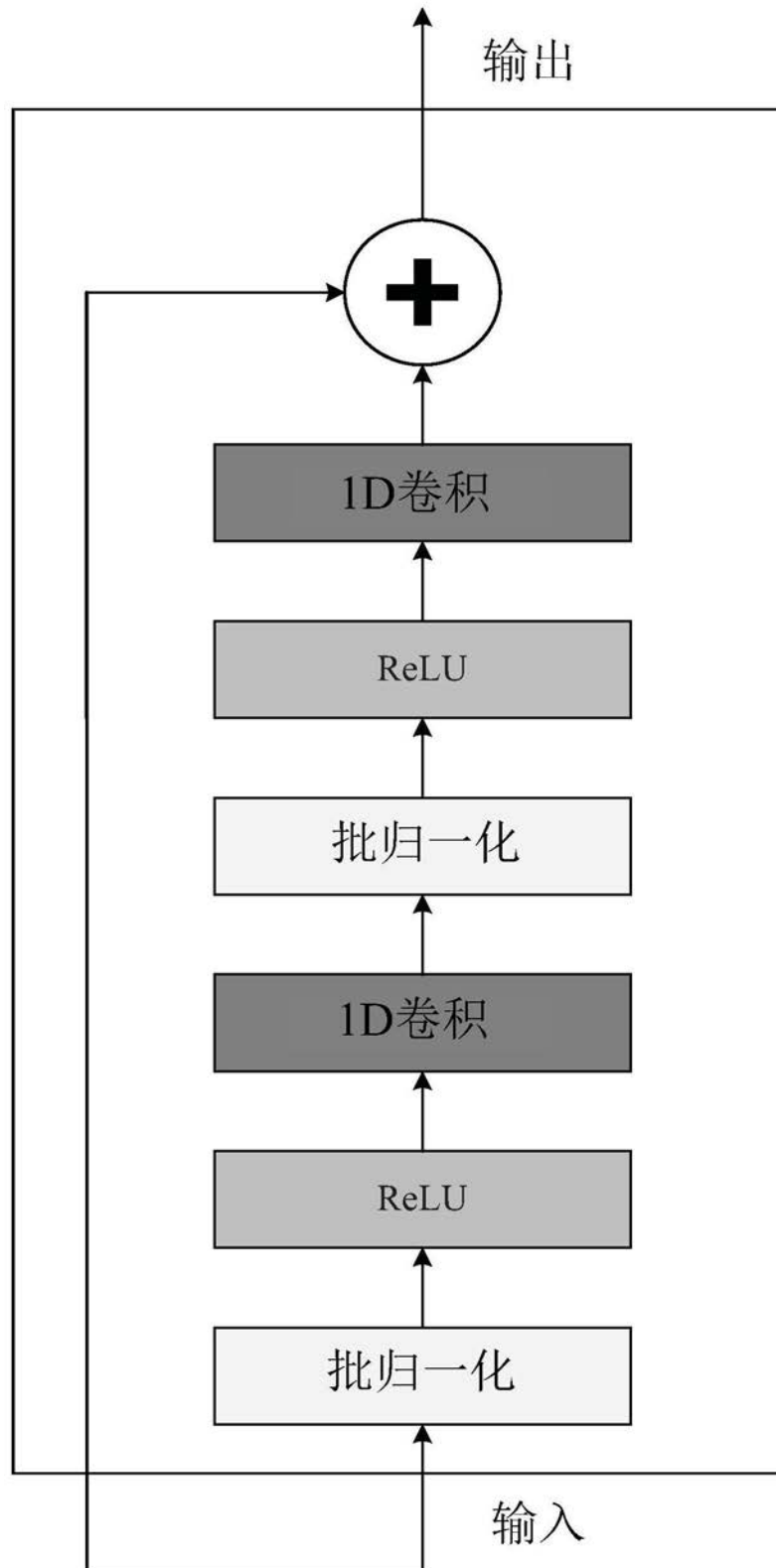


图17

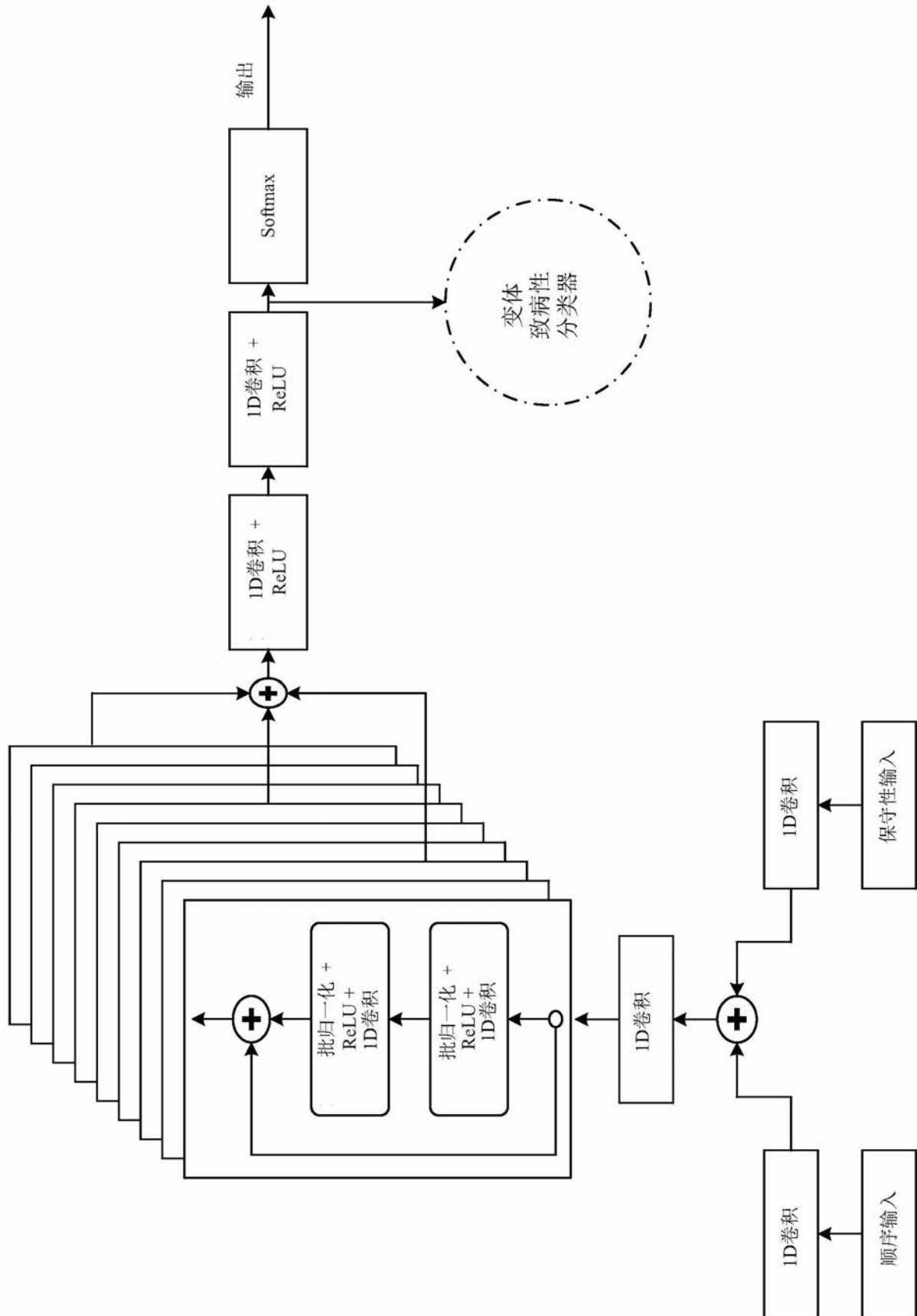


图18

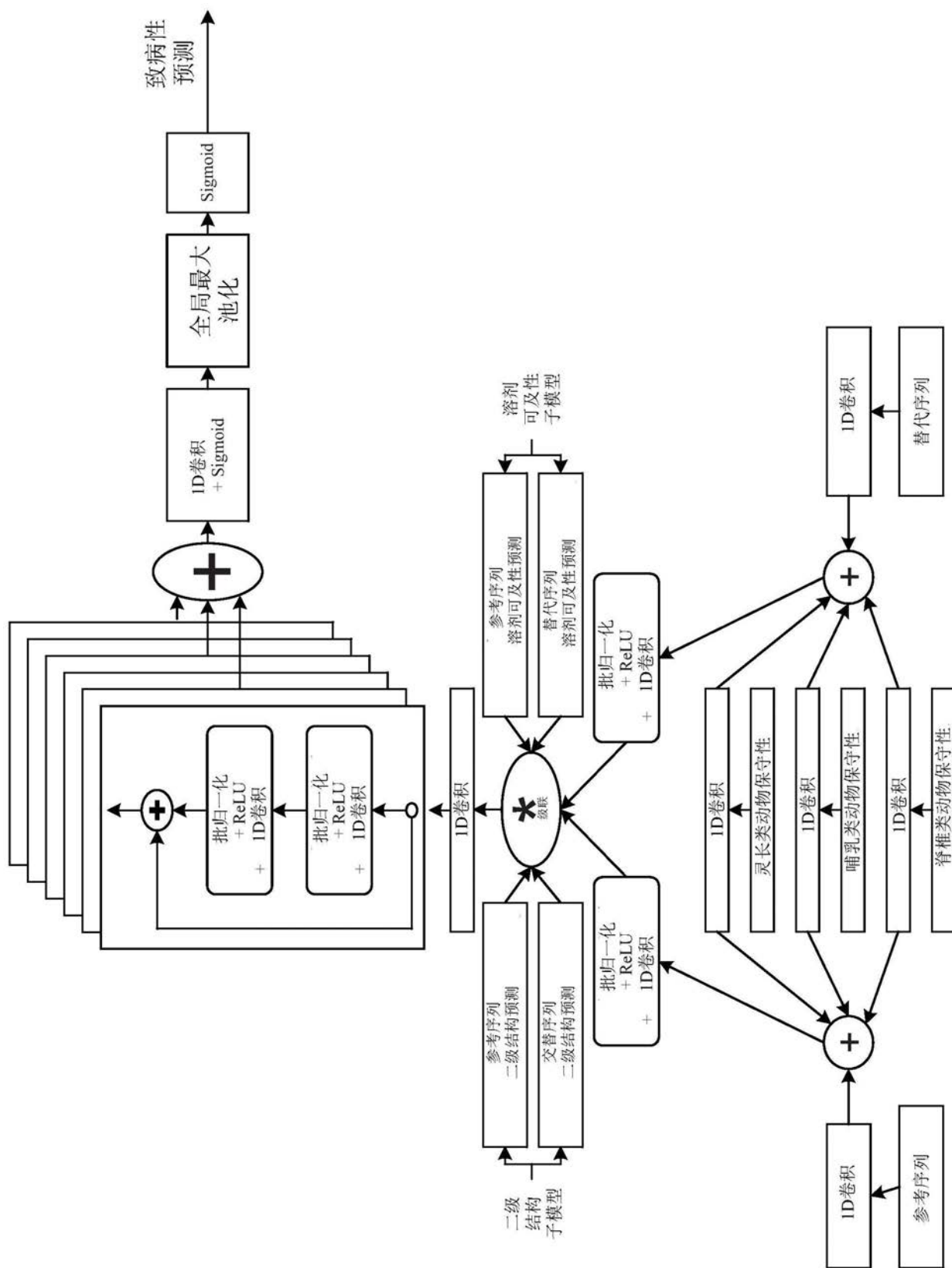


图19

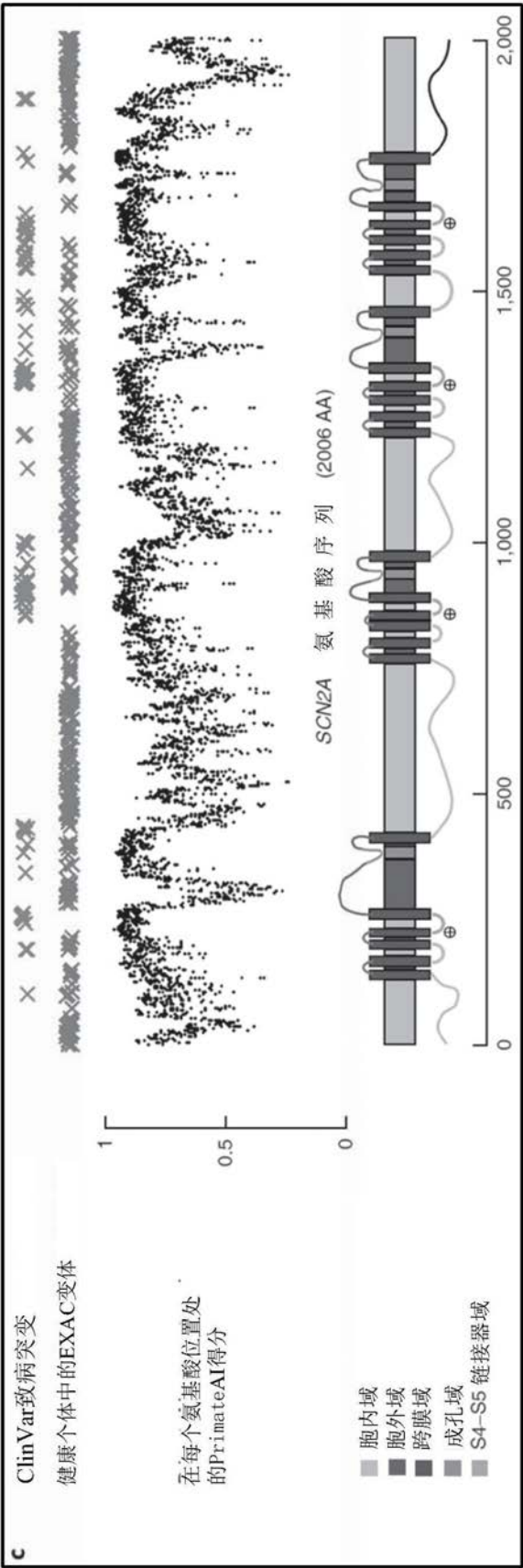


图20

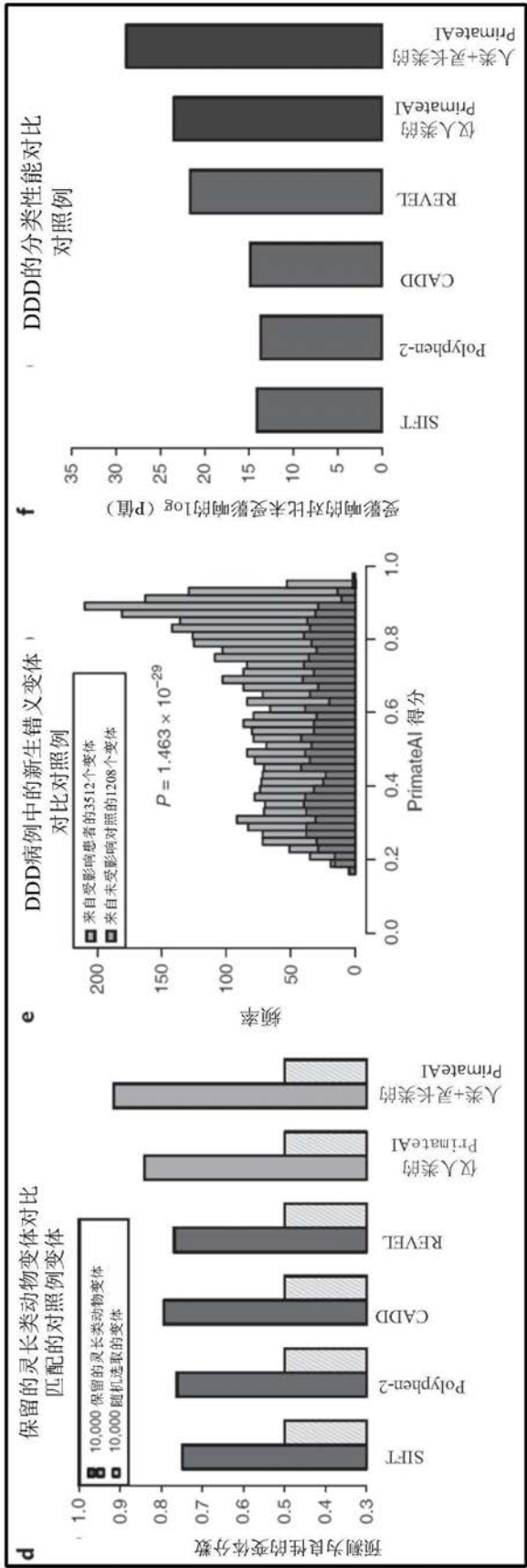


图 21F

图 21E

图 21D

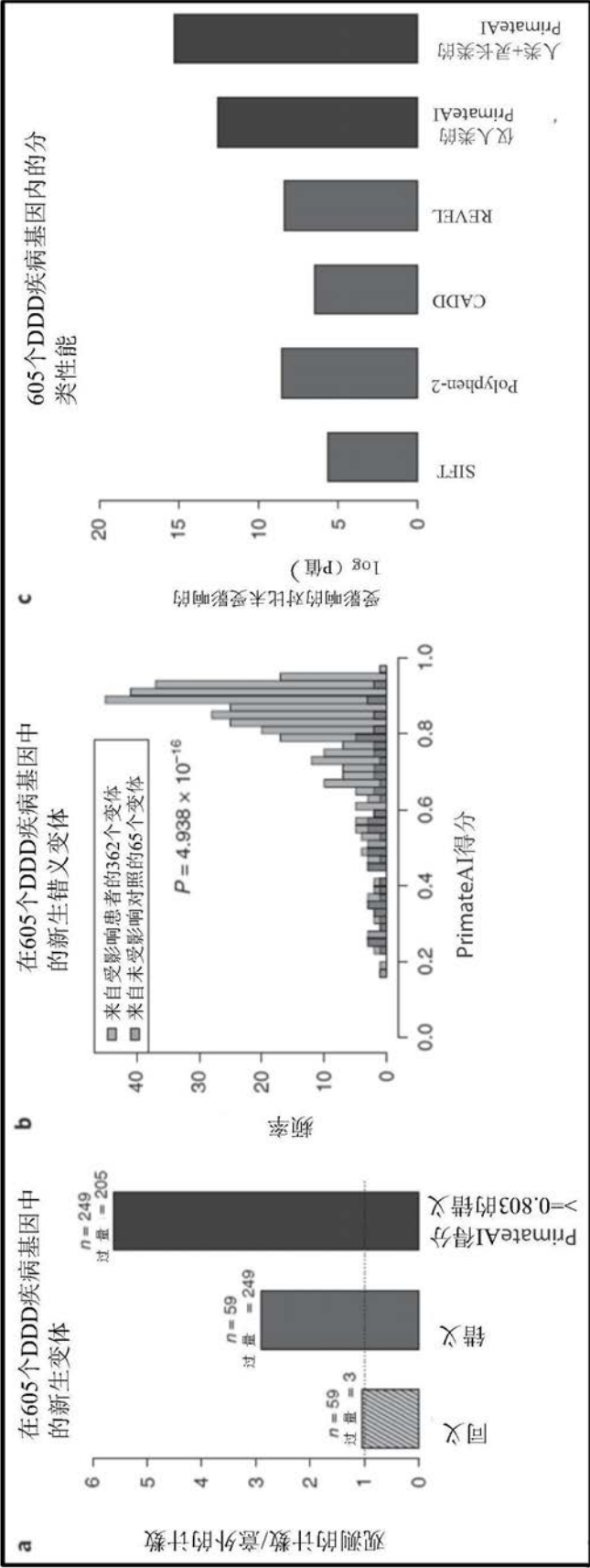


图 22A

图 22B

图 22C

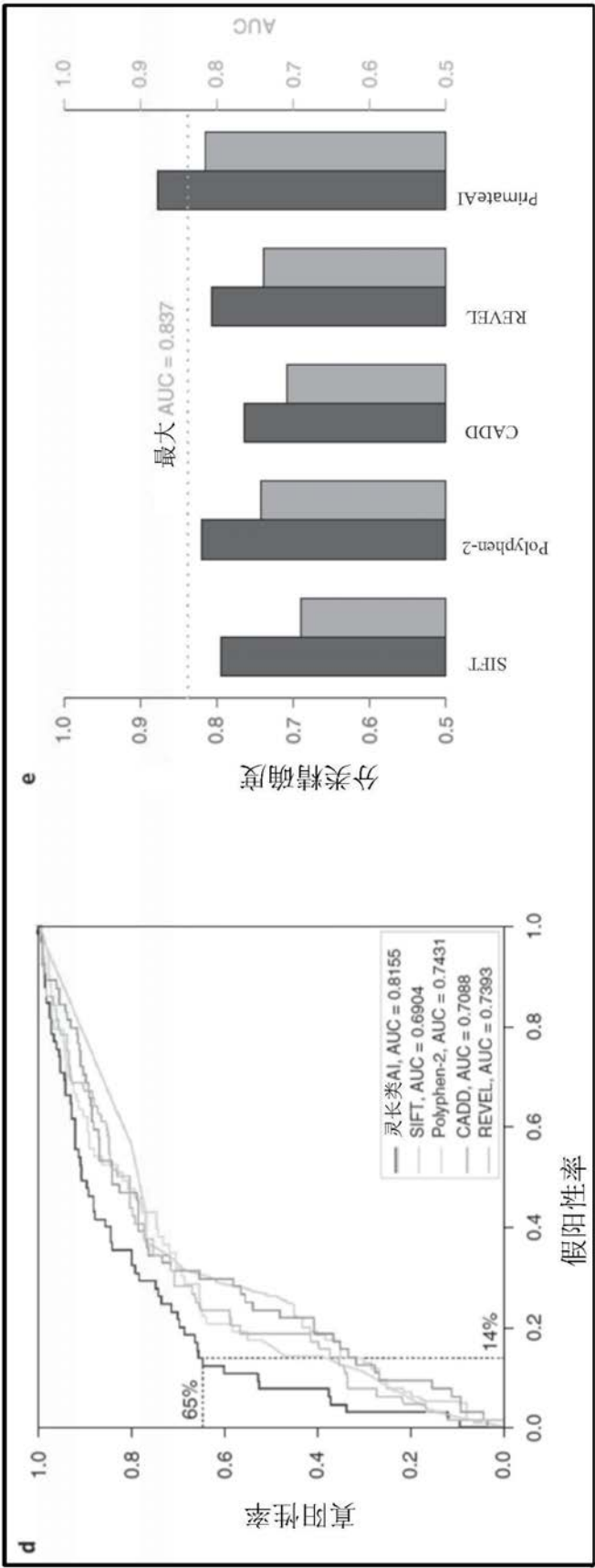


图 22E

图 22D

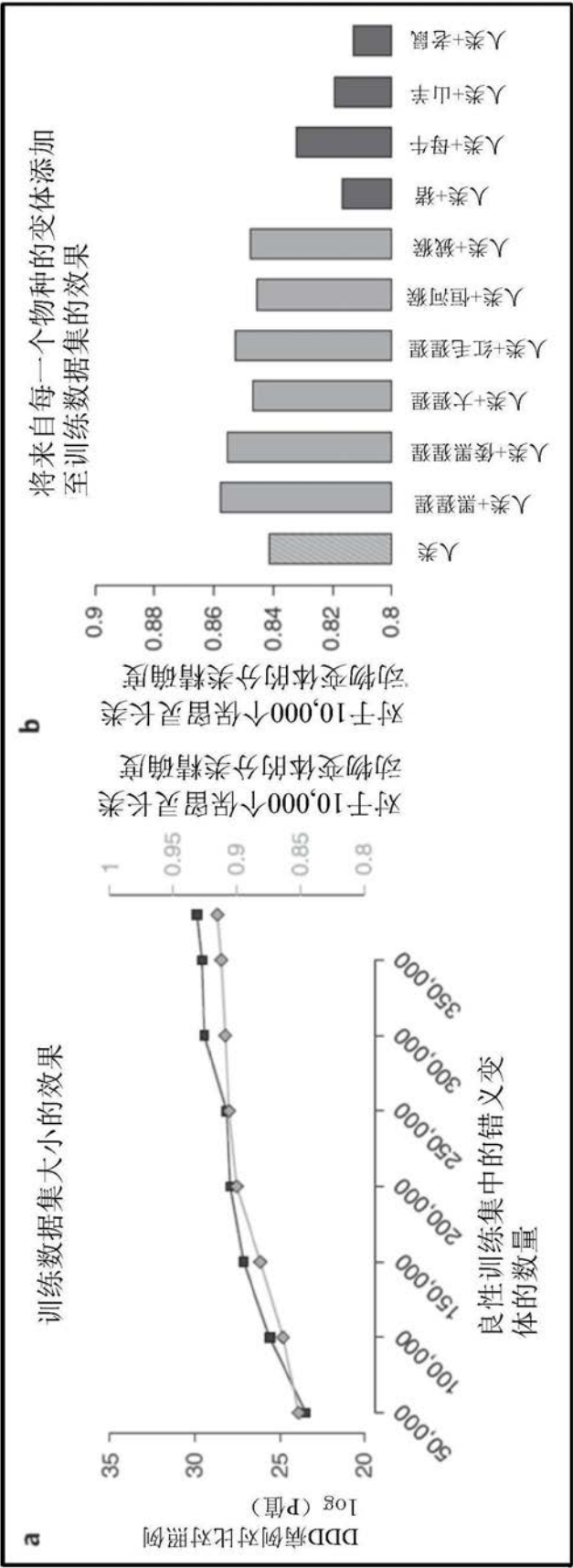


图 23A

图 23B

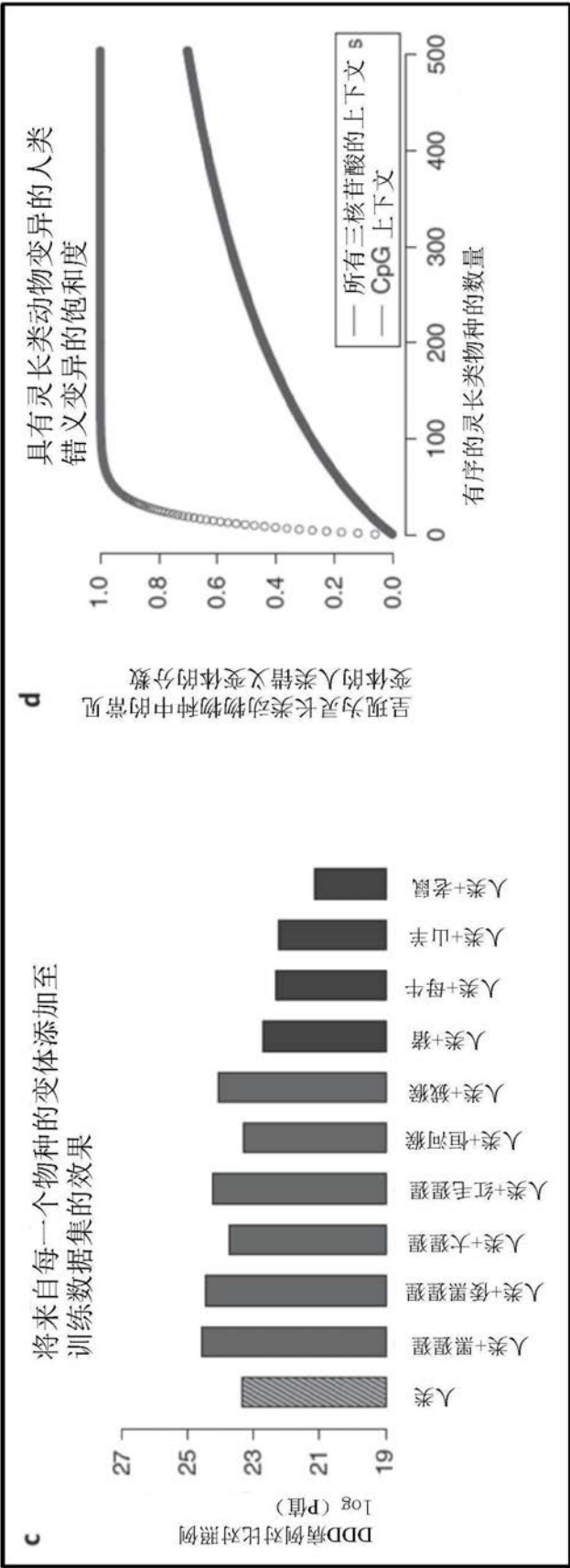


图 23D

图 23C

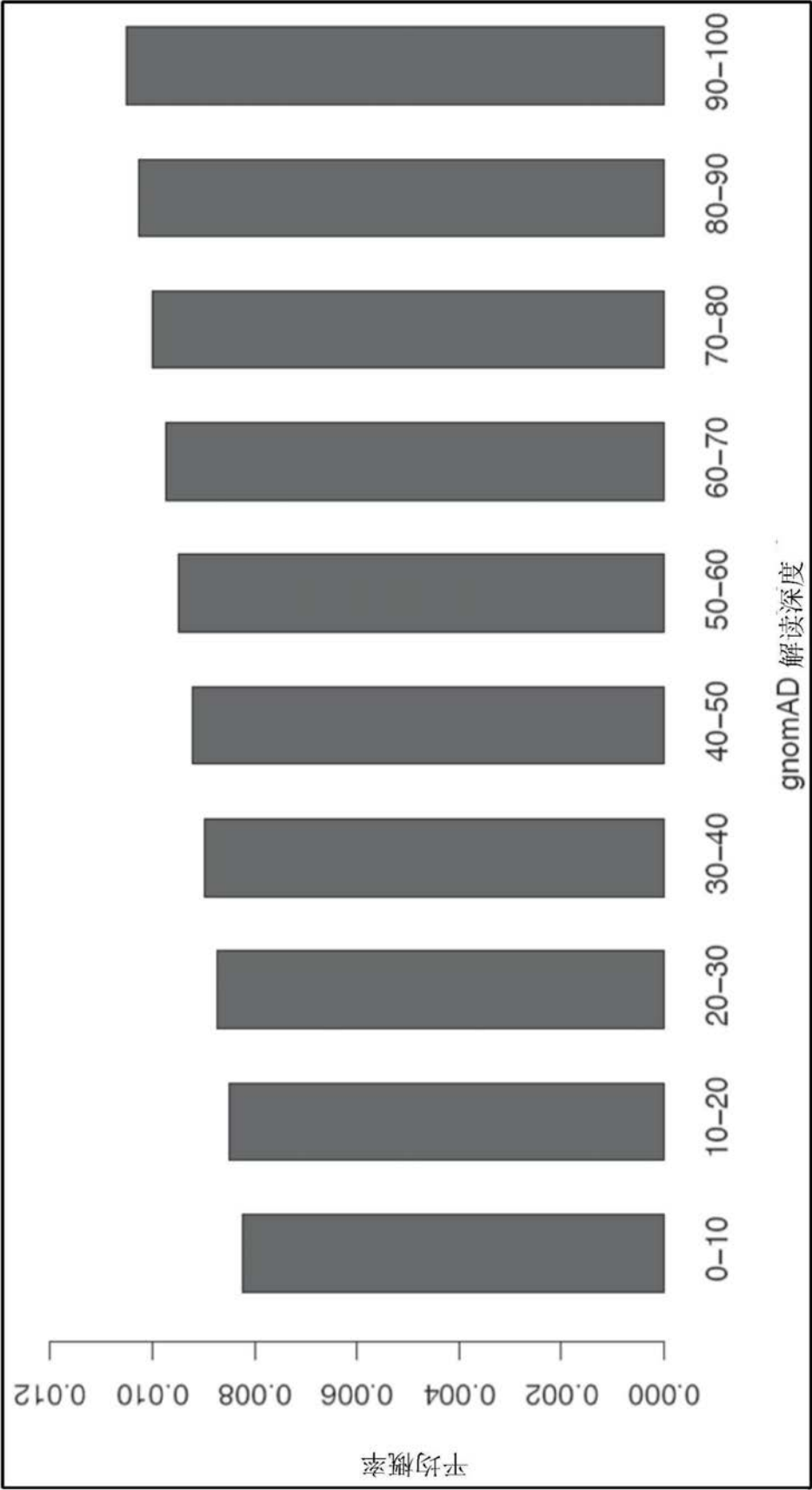


图24

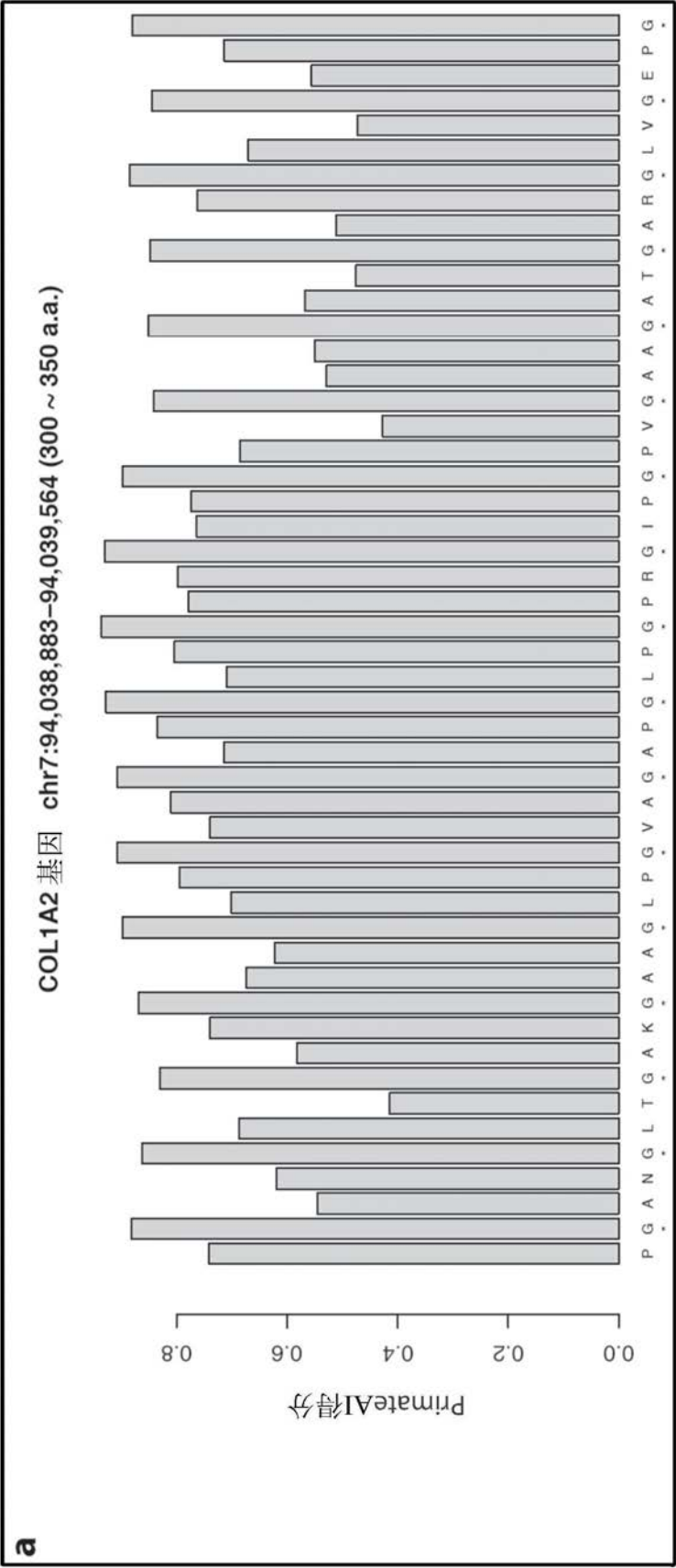


图25A

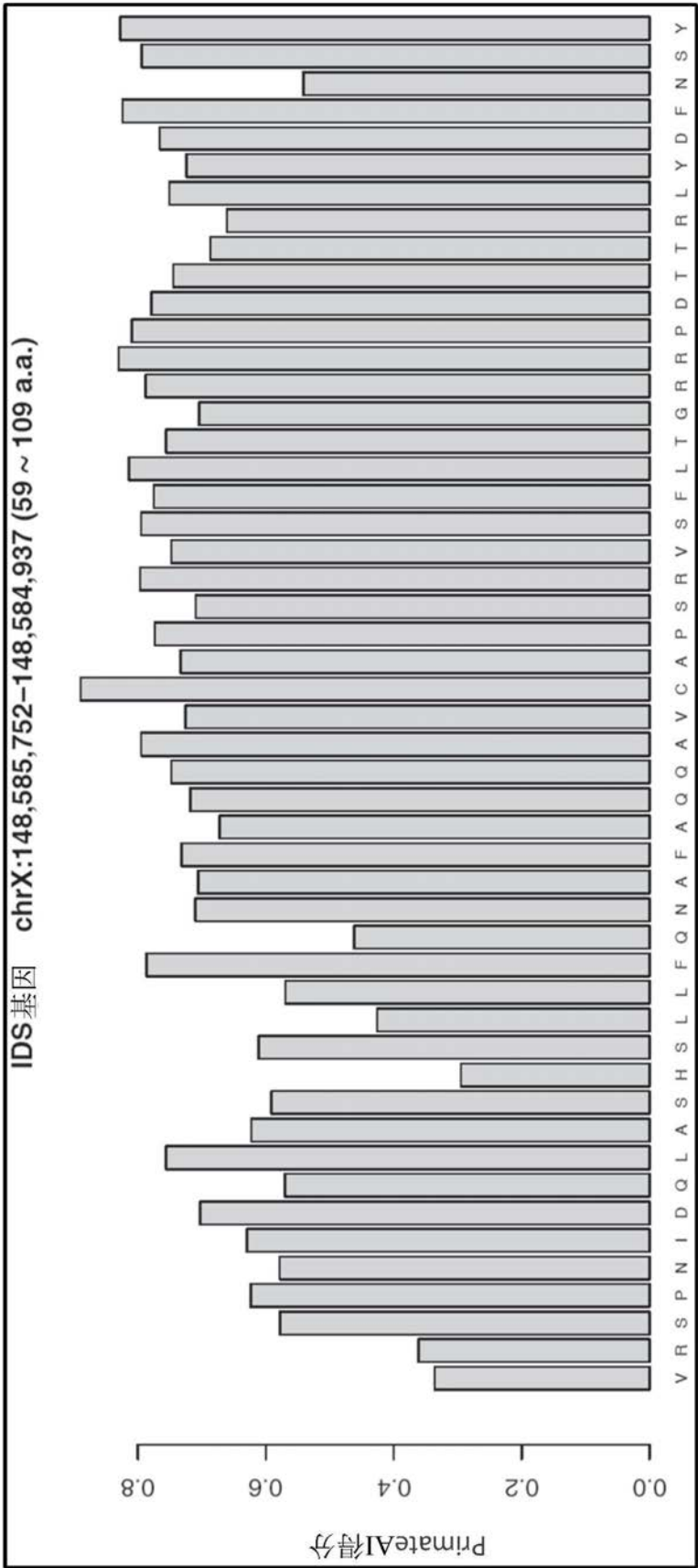
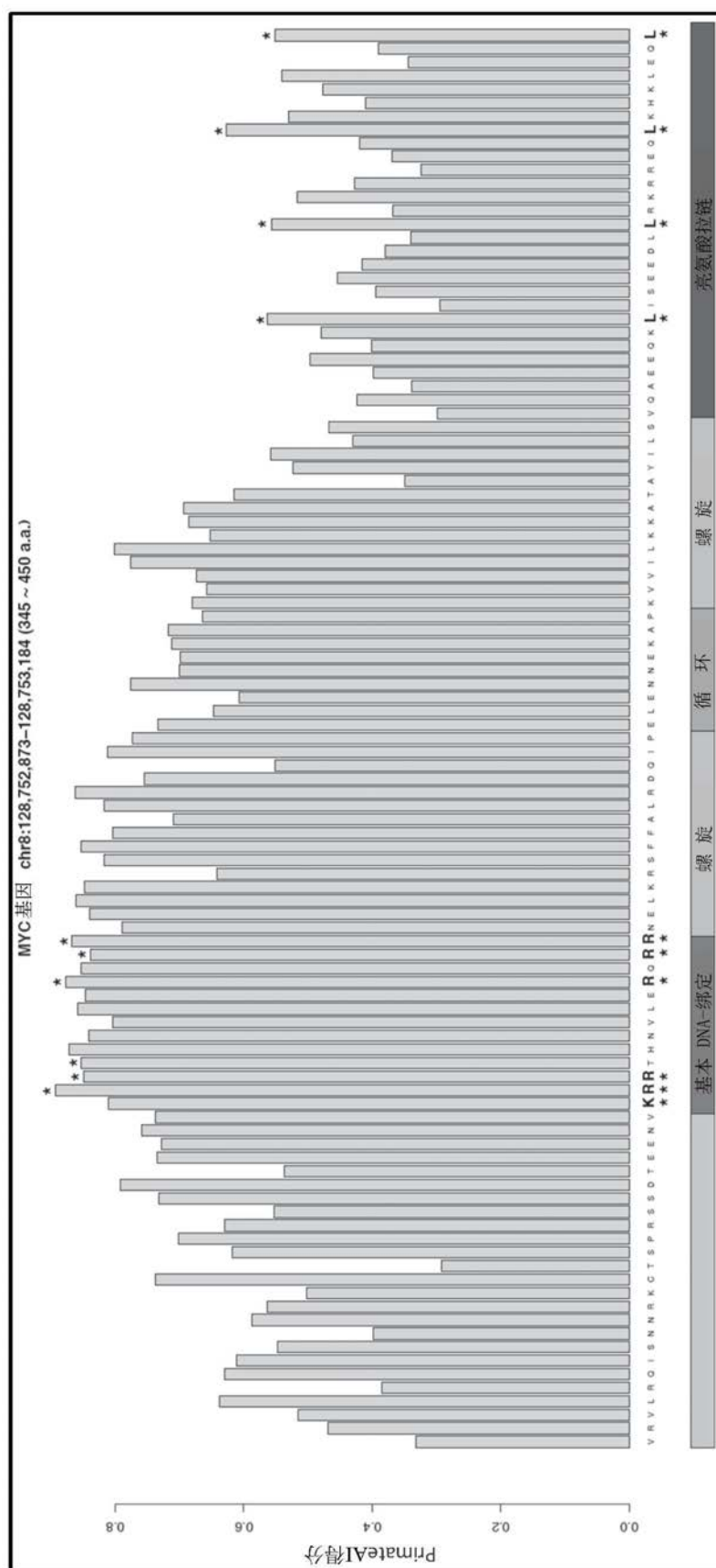


图25B



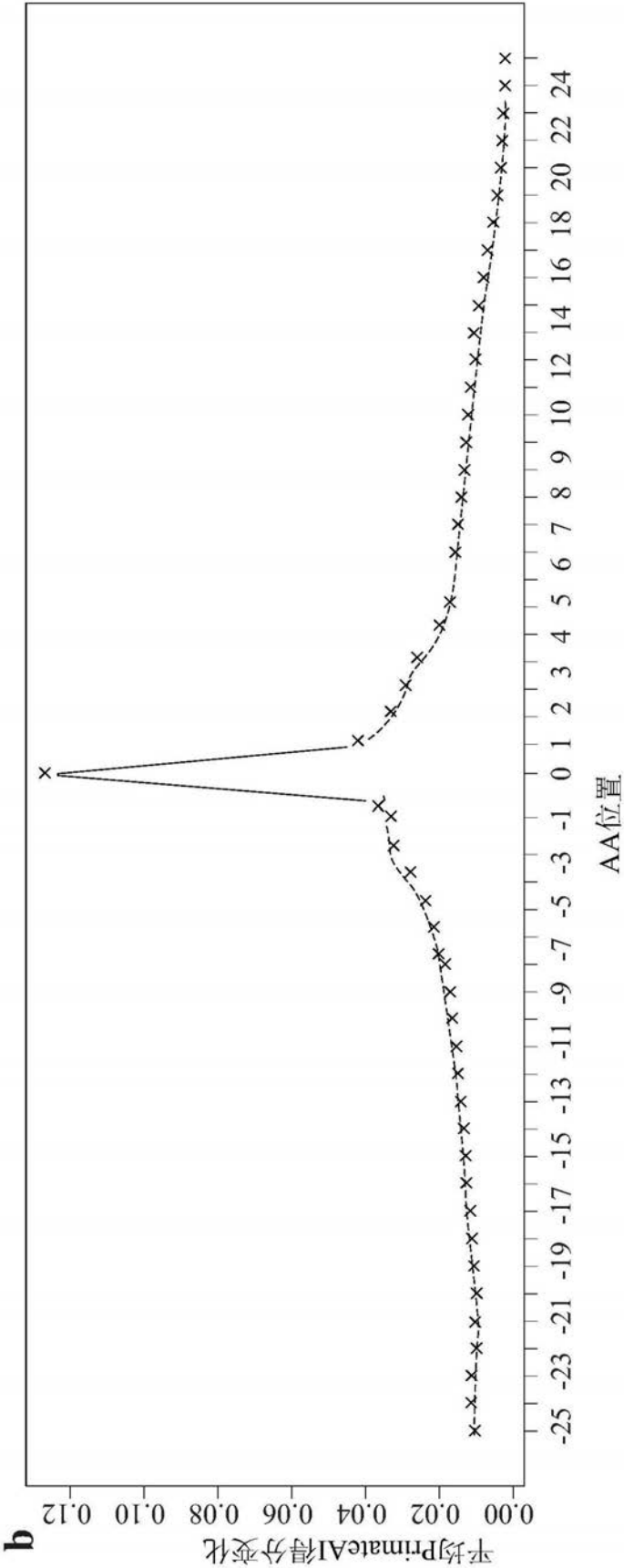


图26

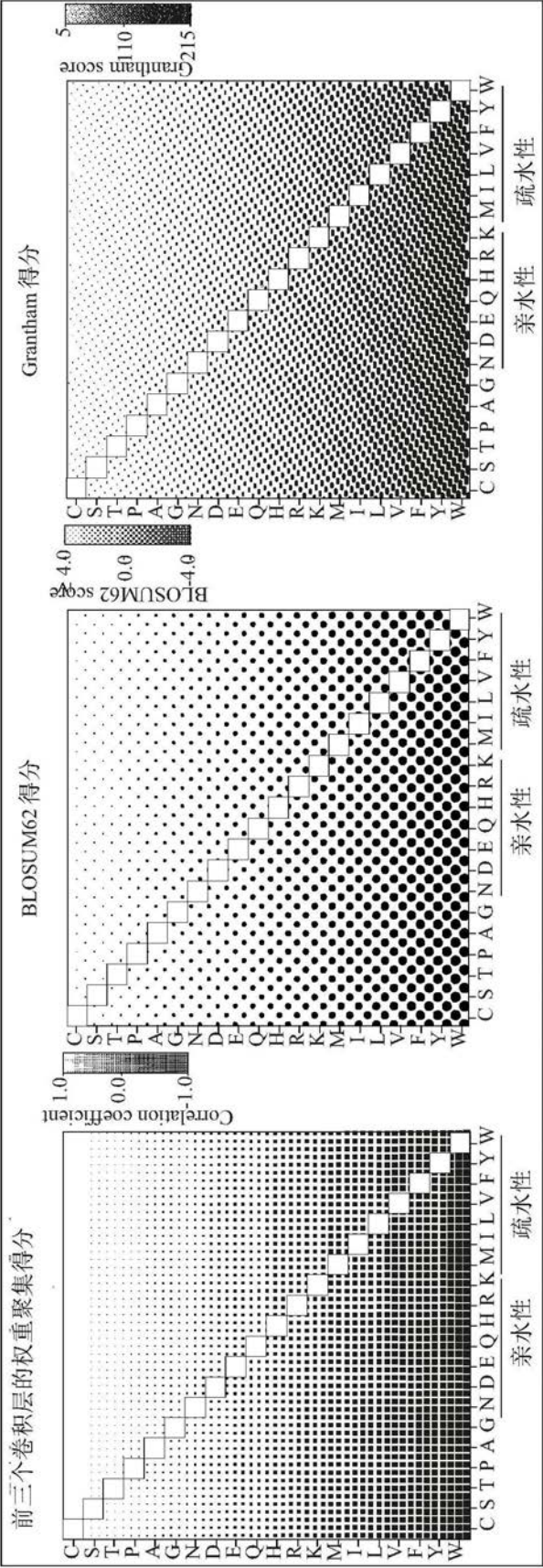


图27

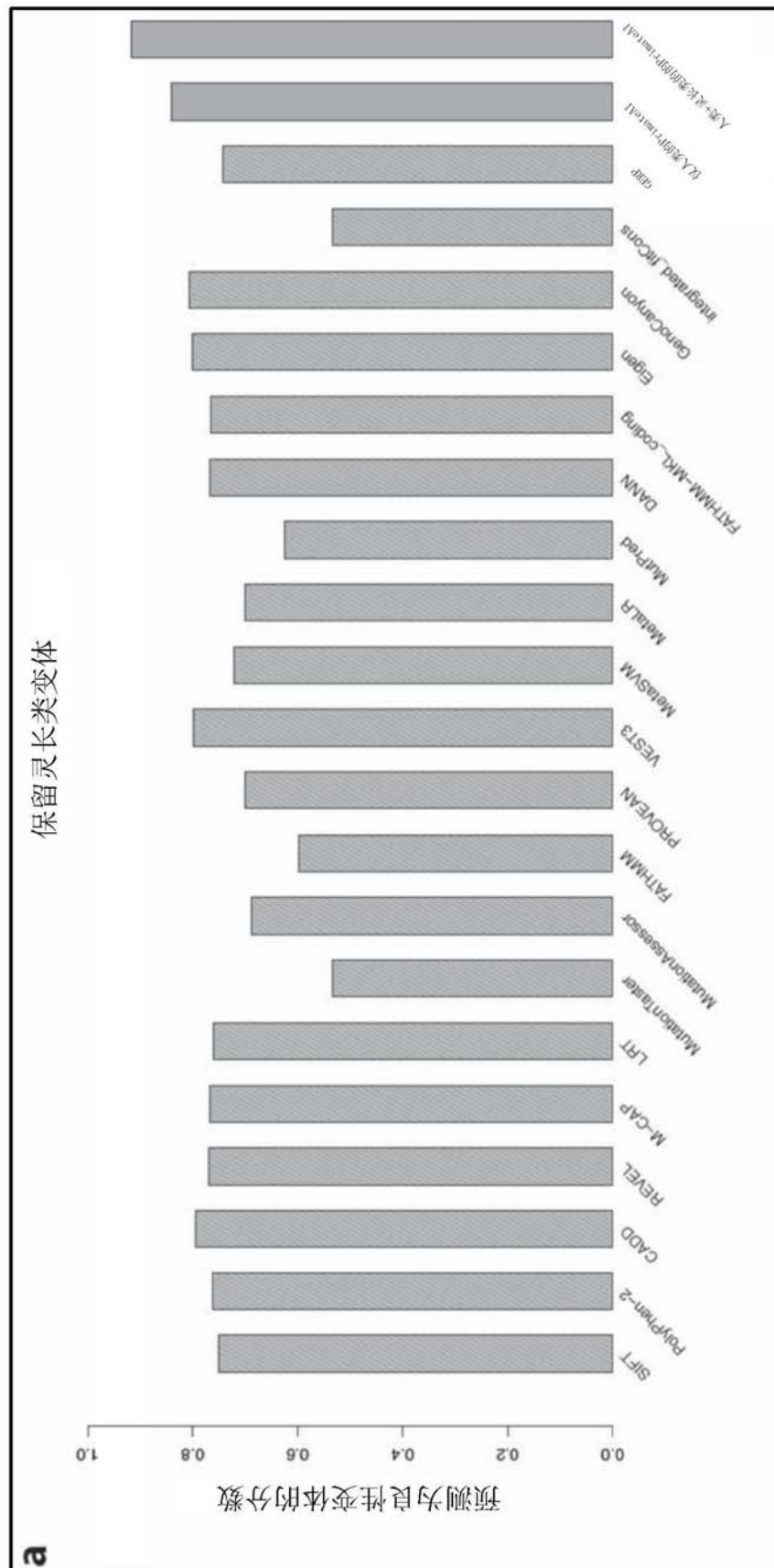


图28A

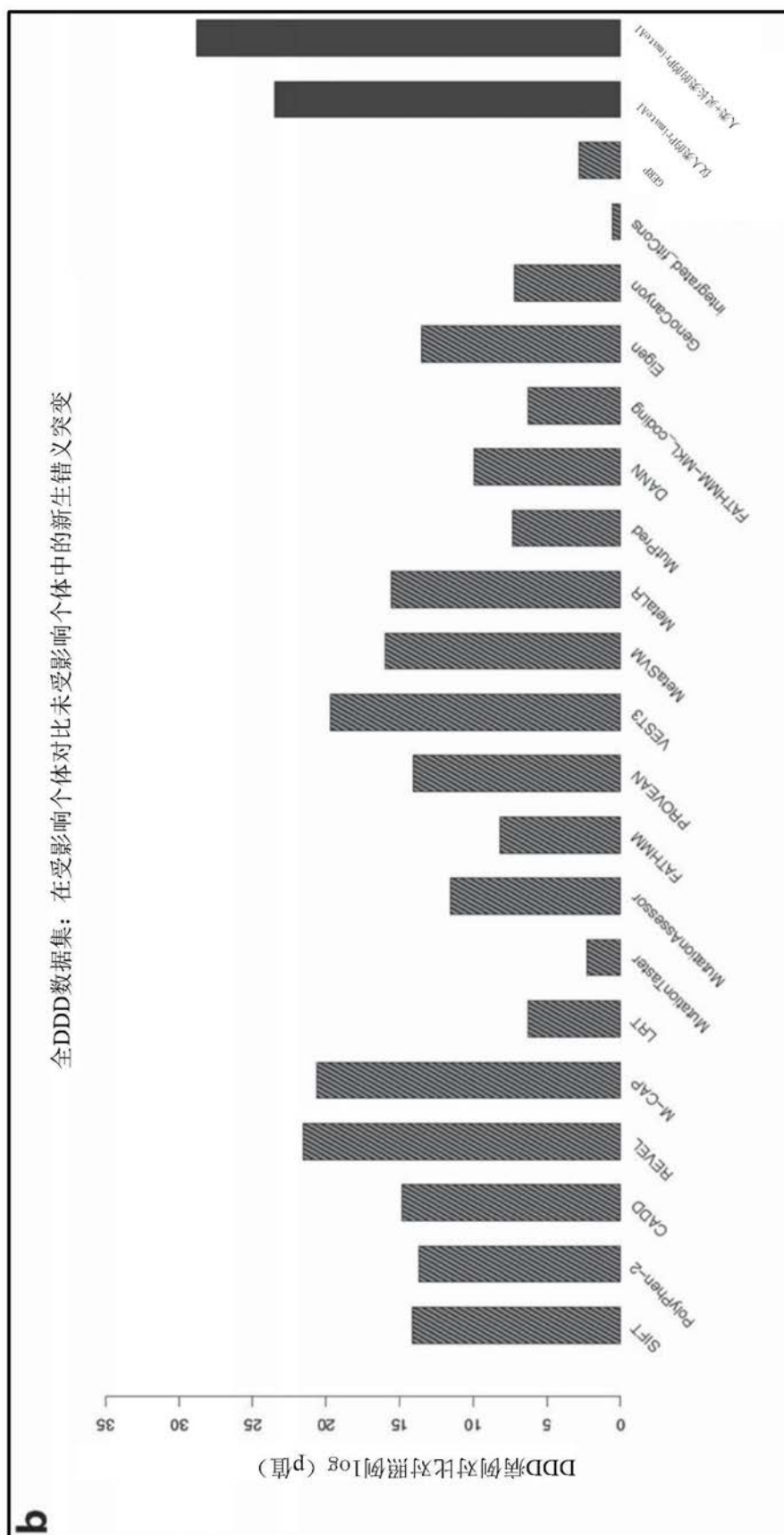


图28B

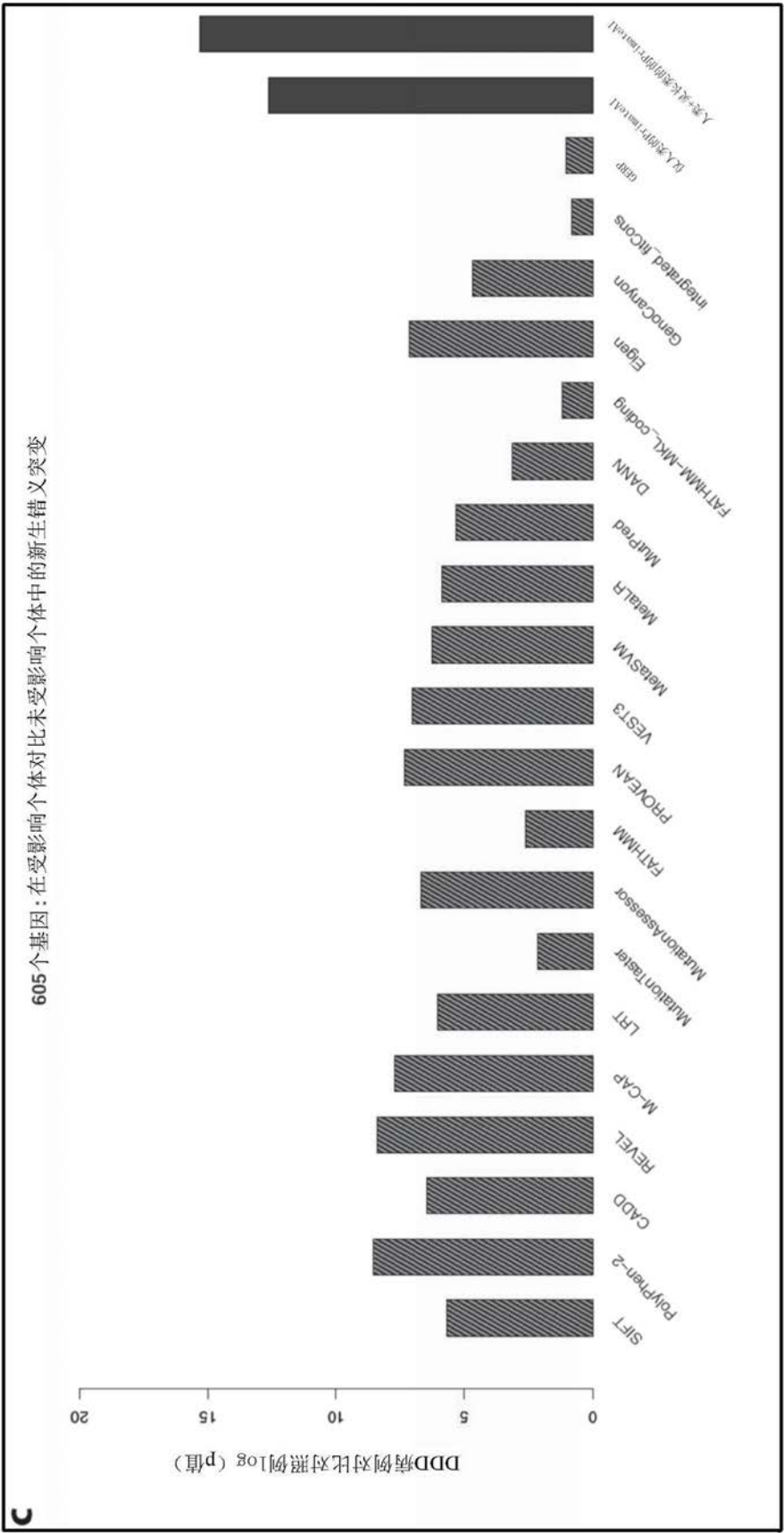


图28C

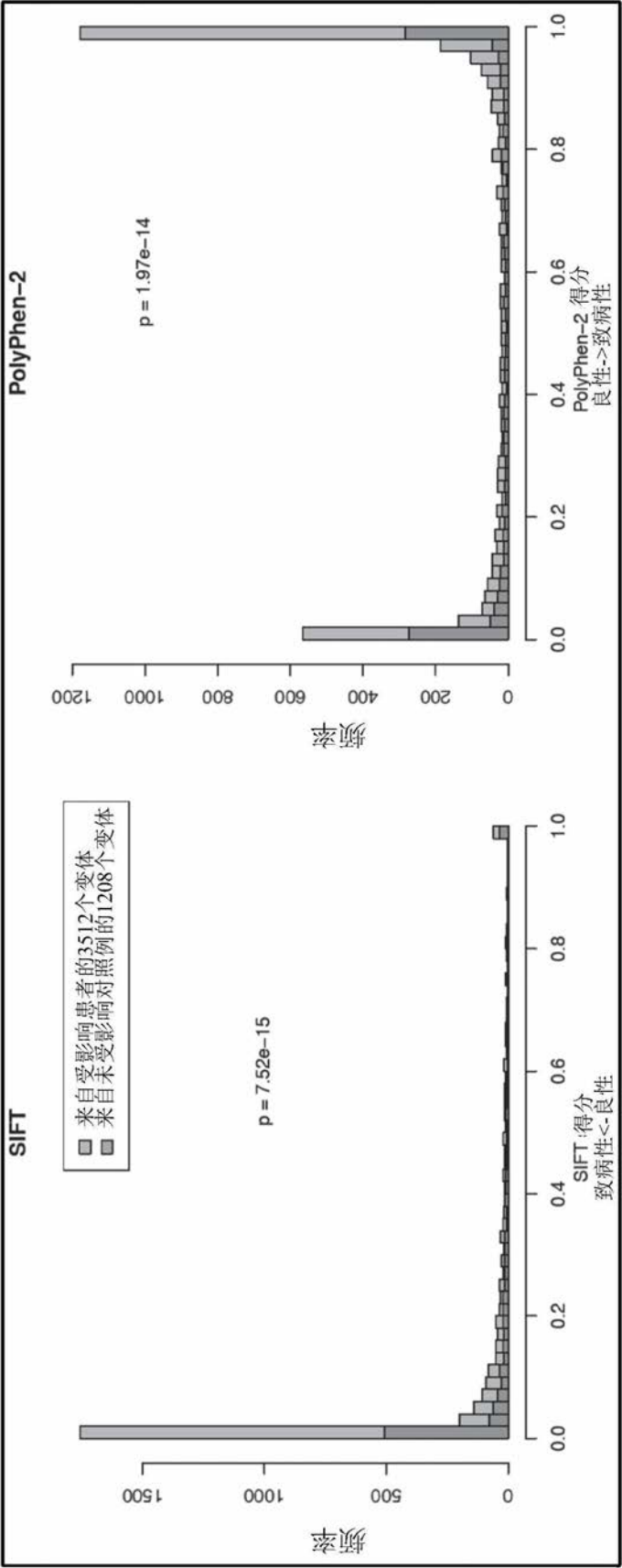


图29A

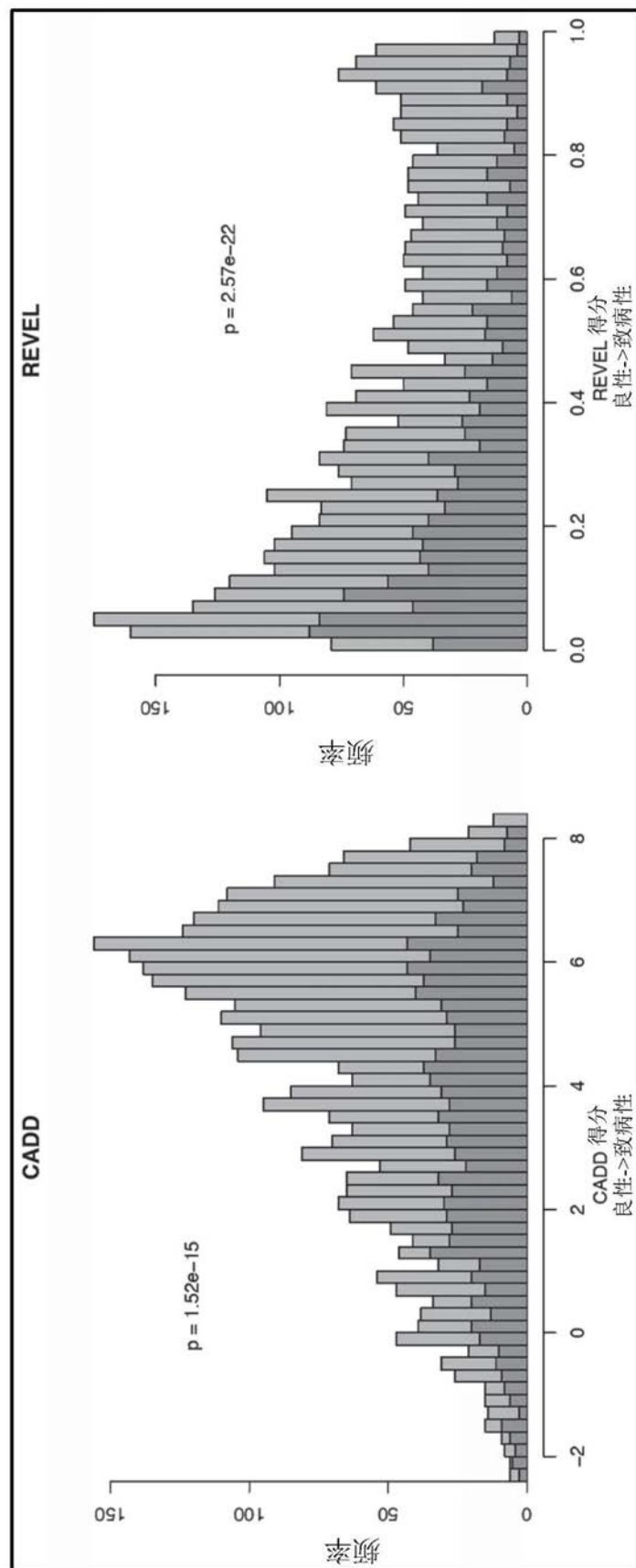


图29B

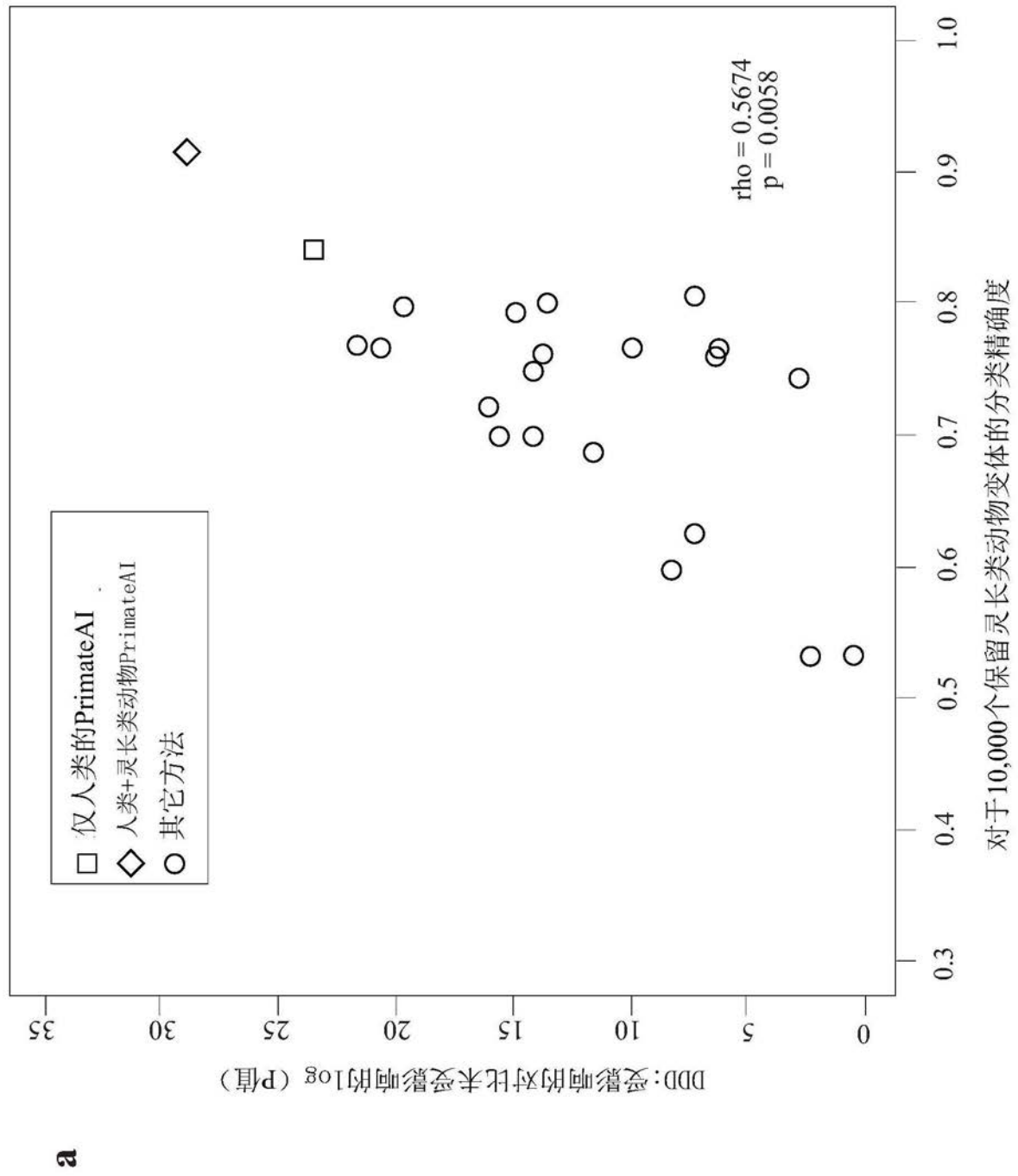


图30A

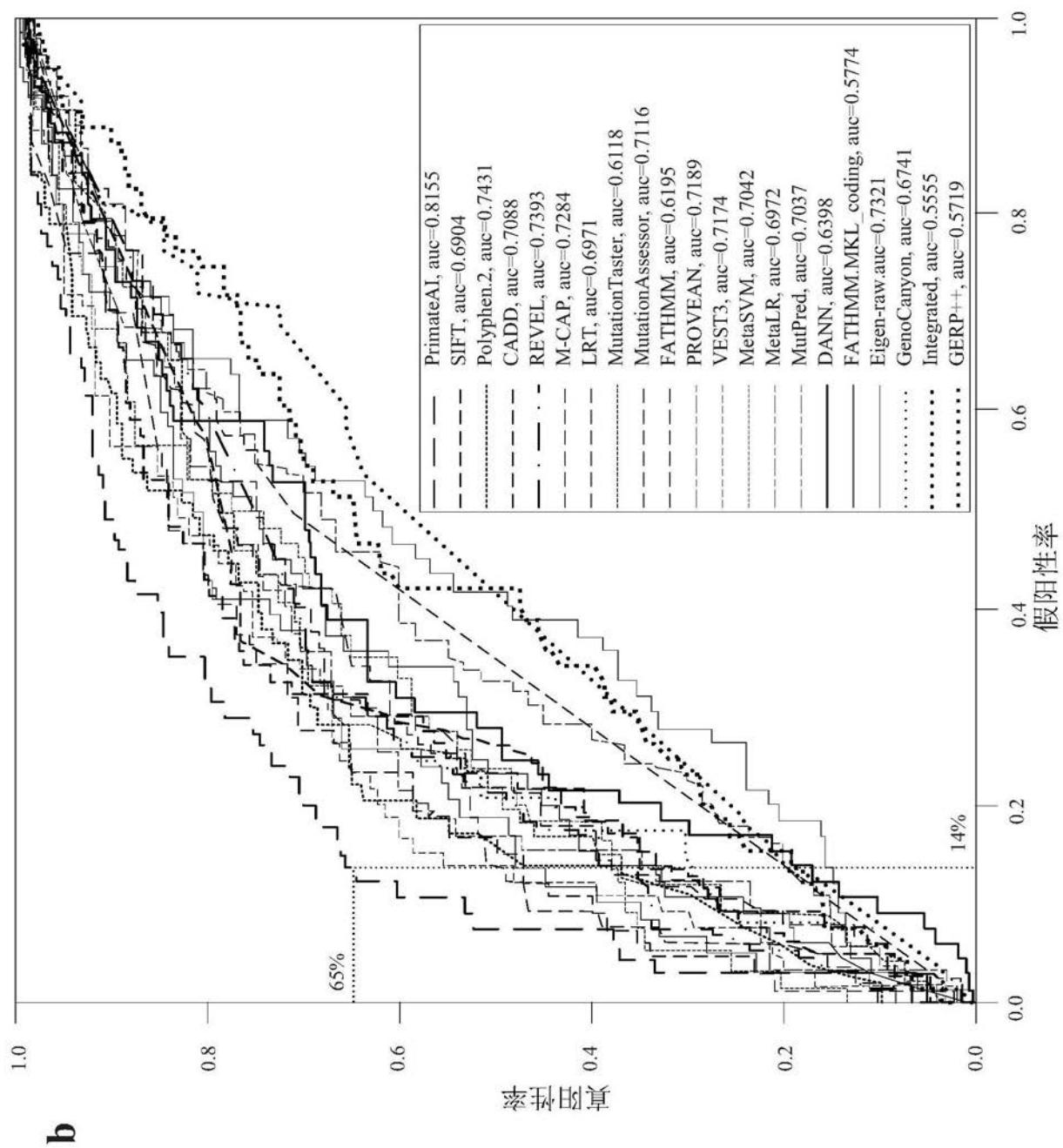


图30B

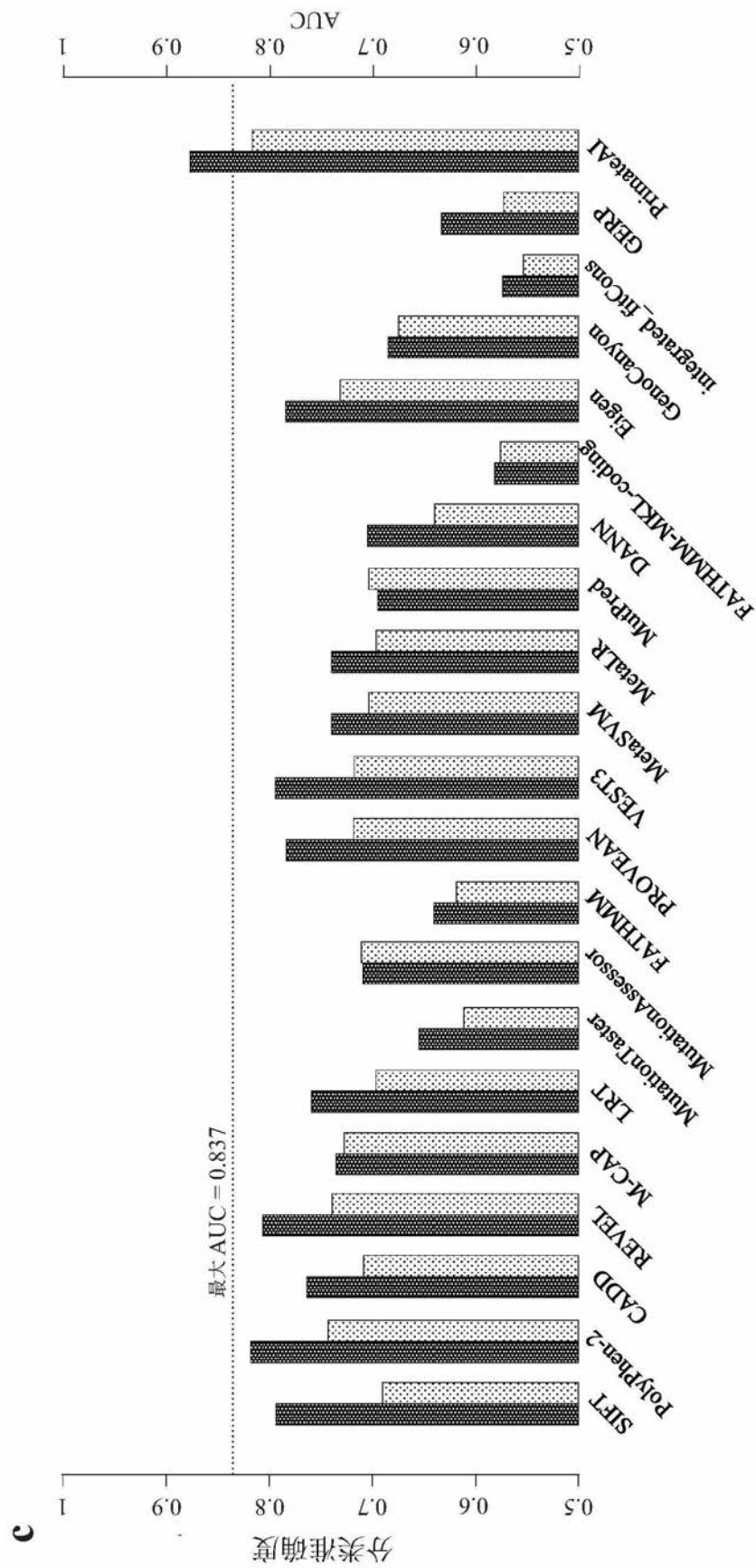


图30C

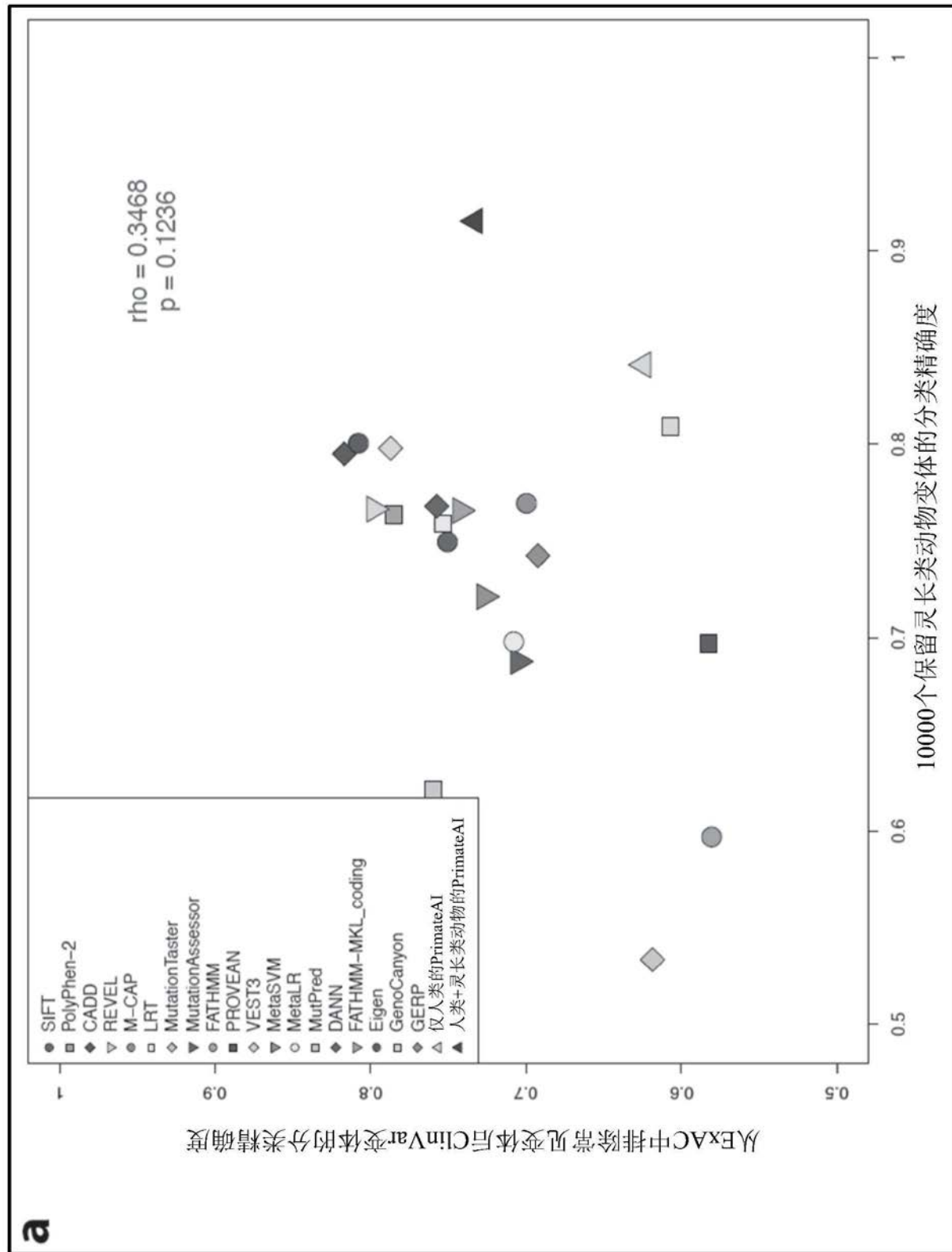


图31A

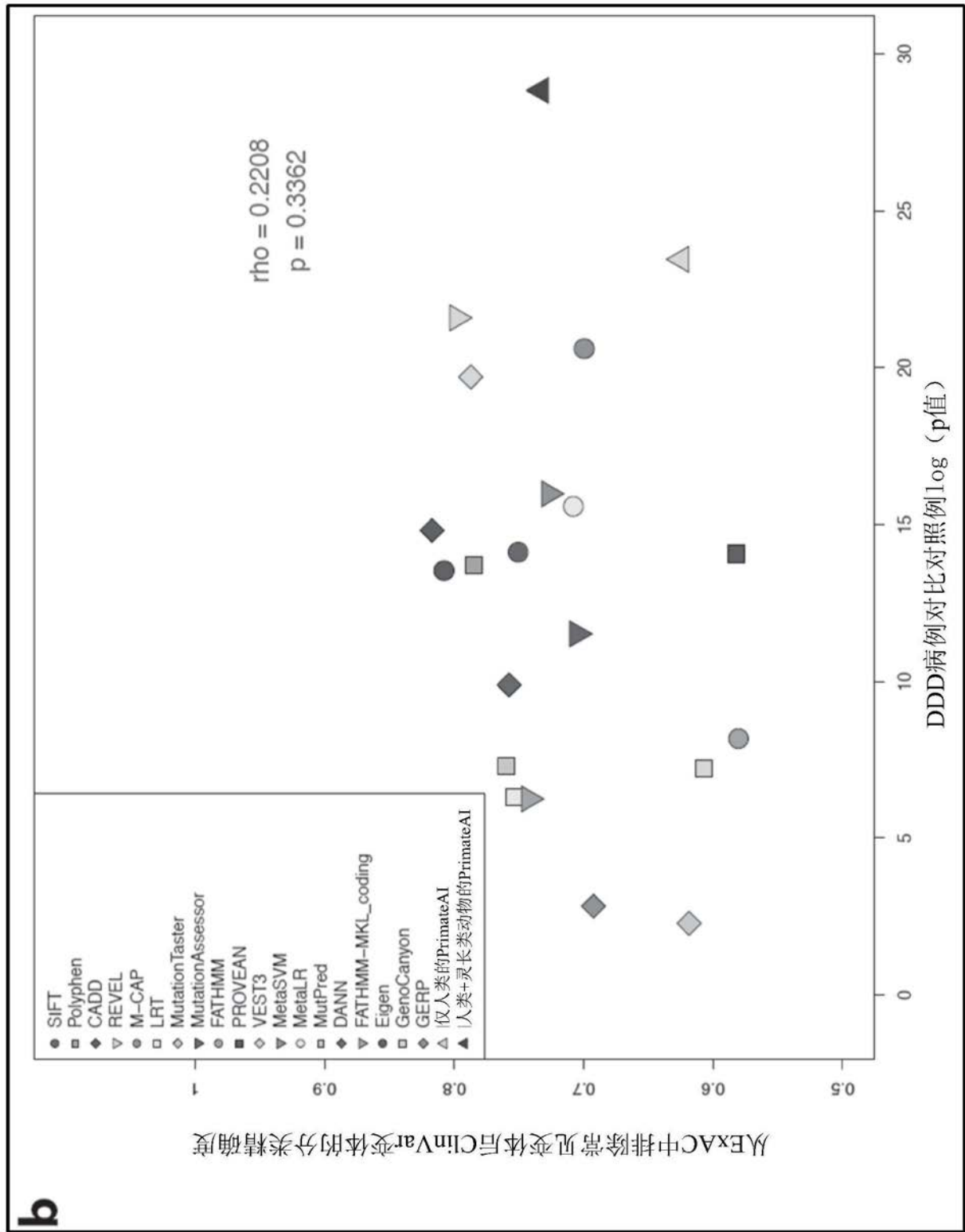


图31B

	训练 精确度	验证 精确度	测试 精确度
三态 二级结构	80.32 %	79.17 %	79.86 %
三态溶剂 可及性	64.83 %	60.53 %	60.31 %

图32

模型	重复	保留灵长类动物 变体精确度	DDD病例对比对照例 log (P值)	DDD 605 个 疾病基因 log (p值)
使用DL预测 的二级结构 标签	1	0.914	28.60	15.61
	2	0.919	32.32	15.94
	3	0.915	29.48	16.38
	4	0.917	30.35	15.62
	5	0.916	29.11	16.22
	中值	0.916	29.48	15.94
当可用时用 人类蛋白质的 DSSP二级结构 标签替换DL 预测的 二级结构标签	1	0.916	29.48	15.94
	2	0.913	31.72	16.24
	3	0.915	31.09	15.69
	4	0.915	30.49	16.09
	5	0.915	30.57	14.79
	中值	0.915	30.79	16.09

图33

	保留灵长类动物 变体精确度	DDD病例对比对照例 log (P值)
SIFT	0.7490	14.1233
PolyPhen-2	0.7618	13.7044
CADD	0.7939	14.8167
M-CAP	0.7667	20.6078
LRT	0.7598	6.2960
MutationTaster	0.5332	2.2811
MutationAssessor	0.6879	11.5202
FATHMM	0.5981	8.1690
PROVEAN	0.6995	14.0747
VEST3	0.7978	19.7068
MetaSVM	0.7215	15.9829
MetaLR	0.6991	15.5842
REVEL	0.7686	21.5892
MutPred	0.6240	7.3044
DANN	0.7666	9.8976
MKL_coding	0.7657	6.2469
Eigen	0.8003	13.5364
GenoCanyon	0.8058	7.2196
integrated_fitCons	0.5331	0.4954
GERP	0.7430	2.8285
人类的PrimateAI	0.8411	23.4536
人类+灵长类动物的 PrimateAI	0.9156	28.8346

图34

	-log(pvalue)	AUC	阈值	分类精确度
SIFT	5.6793	0.6903	0.001	0.7949
PolyPhen-2	8.5522	0.7430	0.956	0.8197
CADD	6.4772	0.7087	5.040	0.7643
M-CAP	7.7143	0.7283	0.081	0.7348
LRT	6.0285	0.6970	1.00E-06	0.7604
MutationTaster	2.1366	0.6117	1.000	0.6564
MutationAssessor	6.6779	0.7115	2.240	0.7096
FATHMM	2.5979	0.6194	-1.260	0.6401
PROVEAN	7.3447	0.7188	-4.310	0.7833
VEST3	7.0357	0.7173	0.679	0.7953
MetaSVM	6.2556	0.7041	-0.3371	0.7406
MetaLR	5.8707	0.6971	0.349	0.7407
REVEL	8.4115	0.7392	0.456	0.8056
MutPred	5.3370	0.7036	0.526	0.6949
DANN	3.1588	0.6397	0.996	0.7046
MKL_coding	1.1649	0.5773	0.965	0.5840
Eigen	7.1659	0.7320	0.577	0.7844
GenoCanyon	4.6763	0.6740	0.999	0.6875
integrated_fitCons	0.8315	0.5554	0.706	0.5736
GERP	1.0330	0.5718	5.060	0.6330
人类和灵长动物的 PrimateAI	15.3064	0.8154	0.802	0.8772

图35

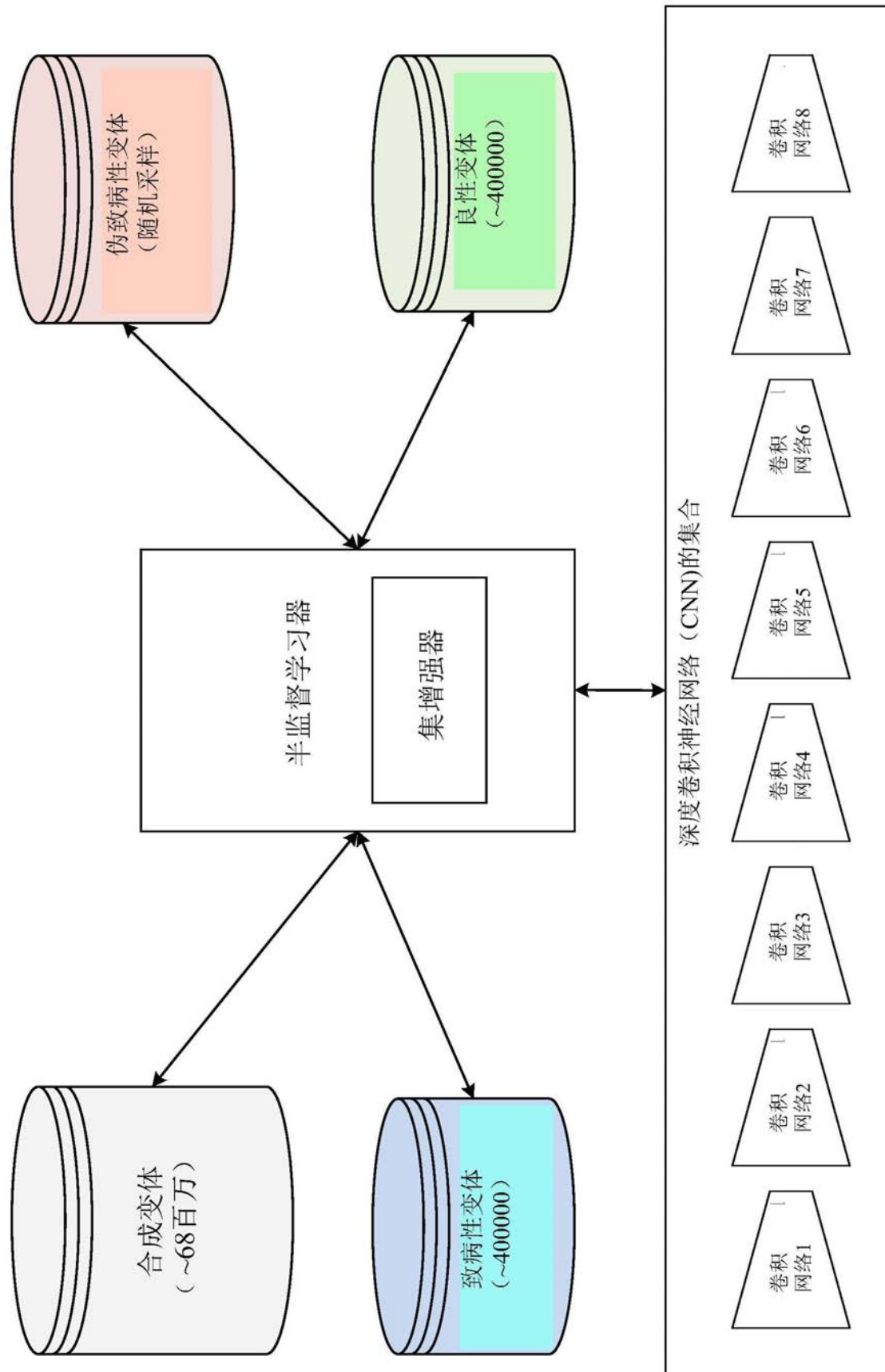


图36

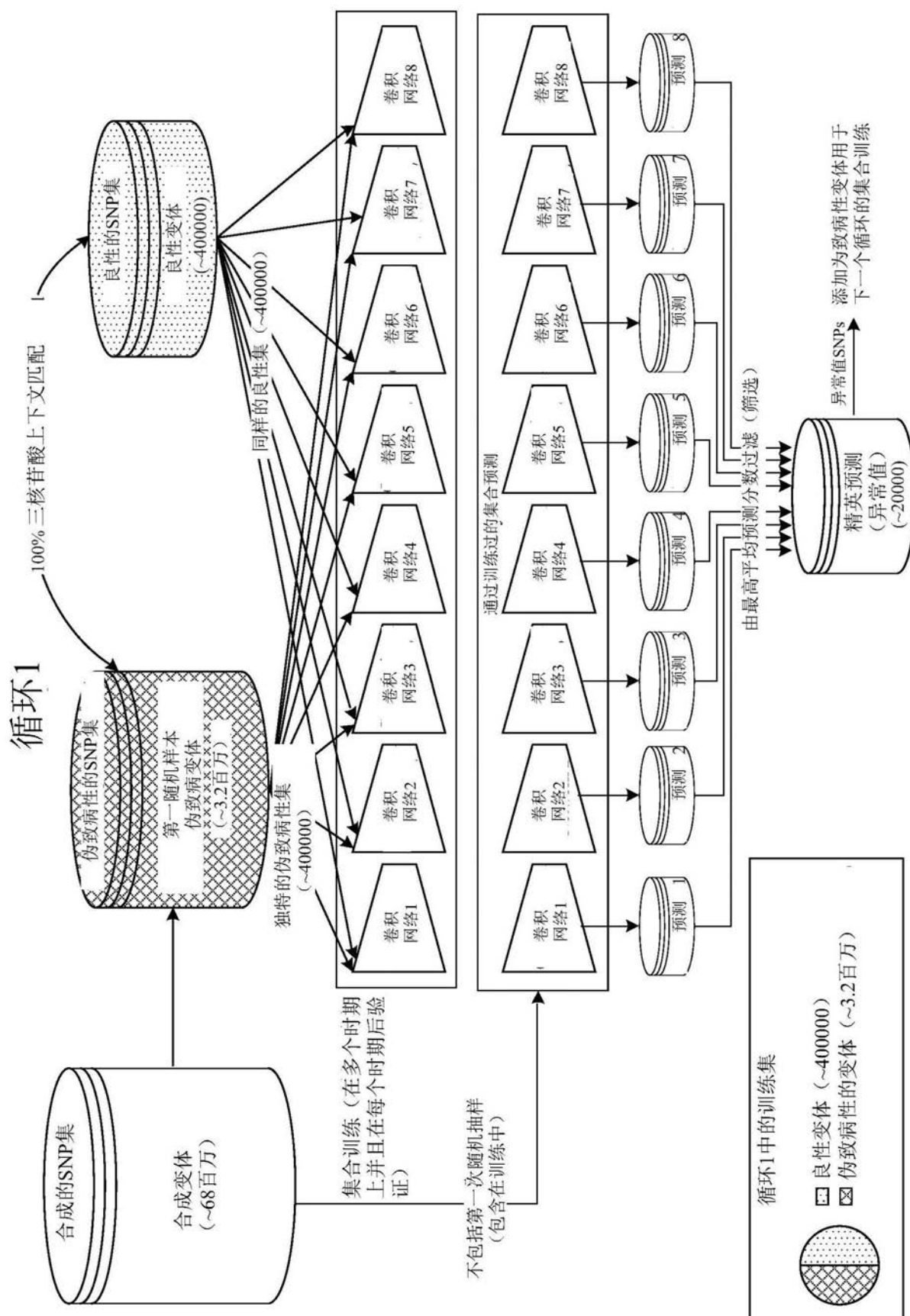


图37

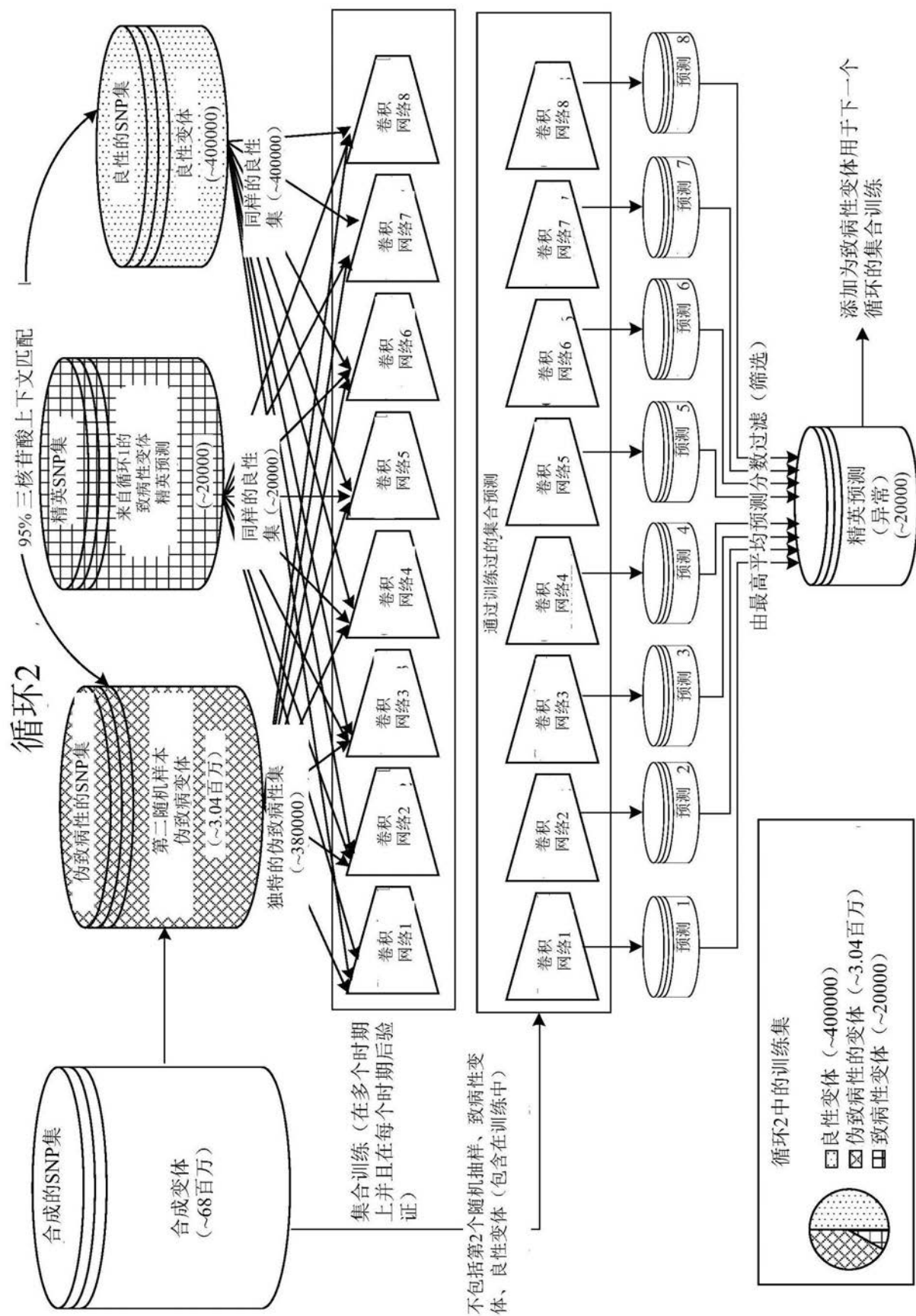


图38

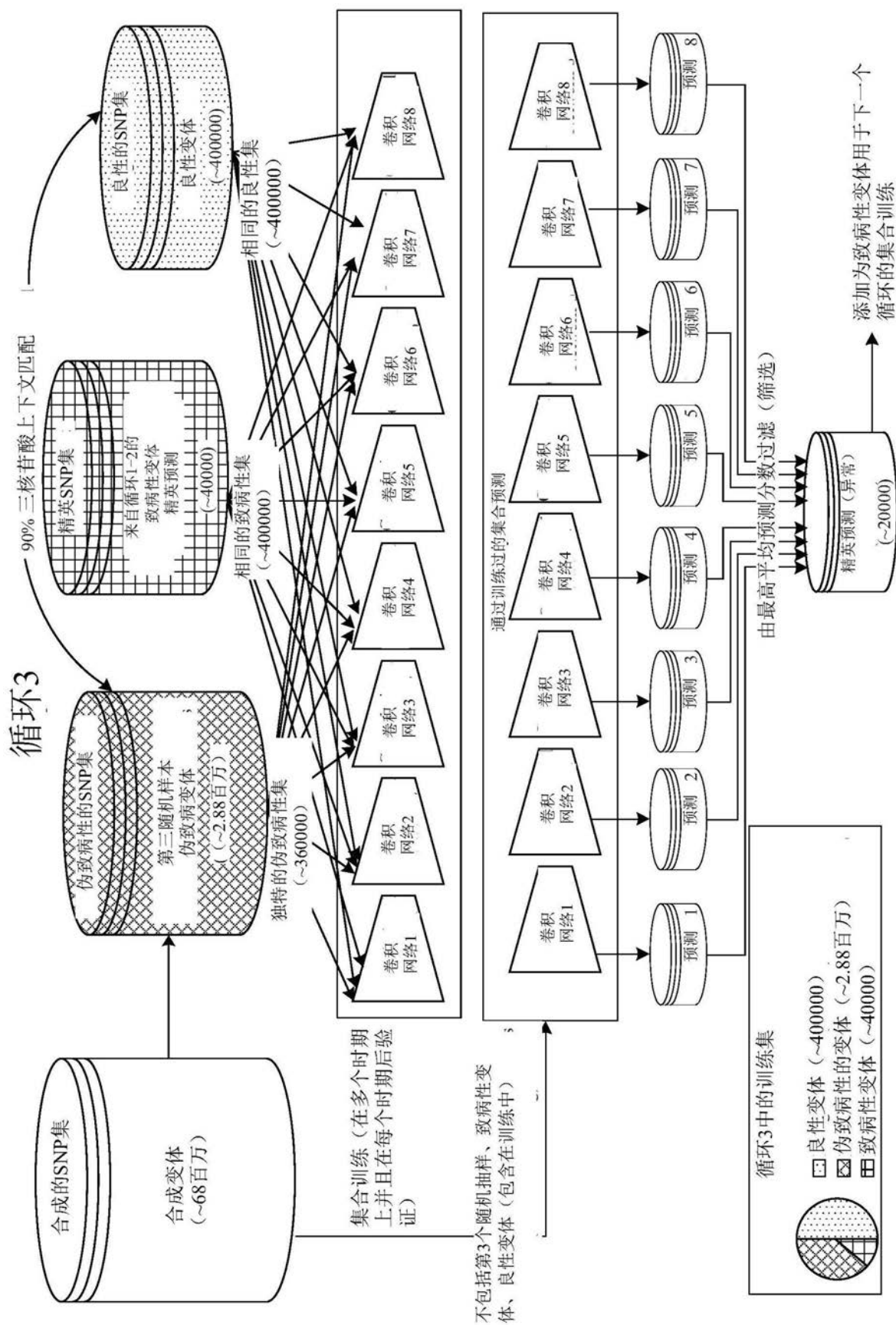


图39

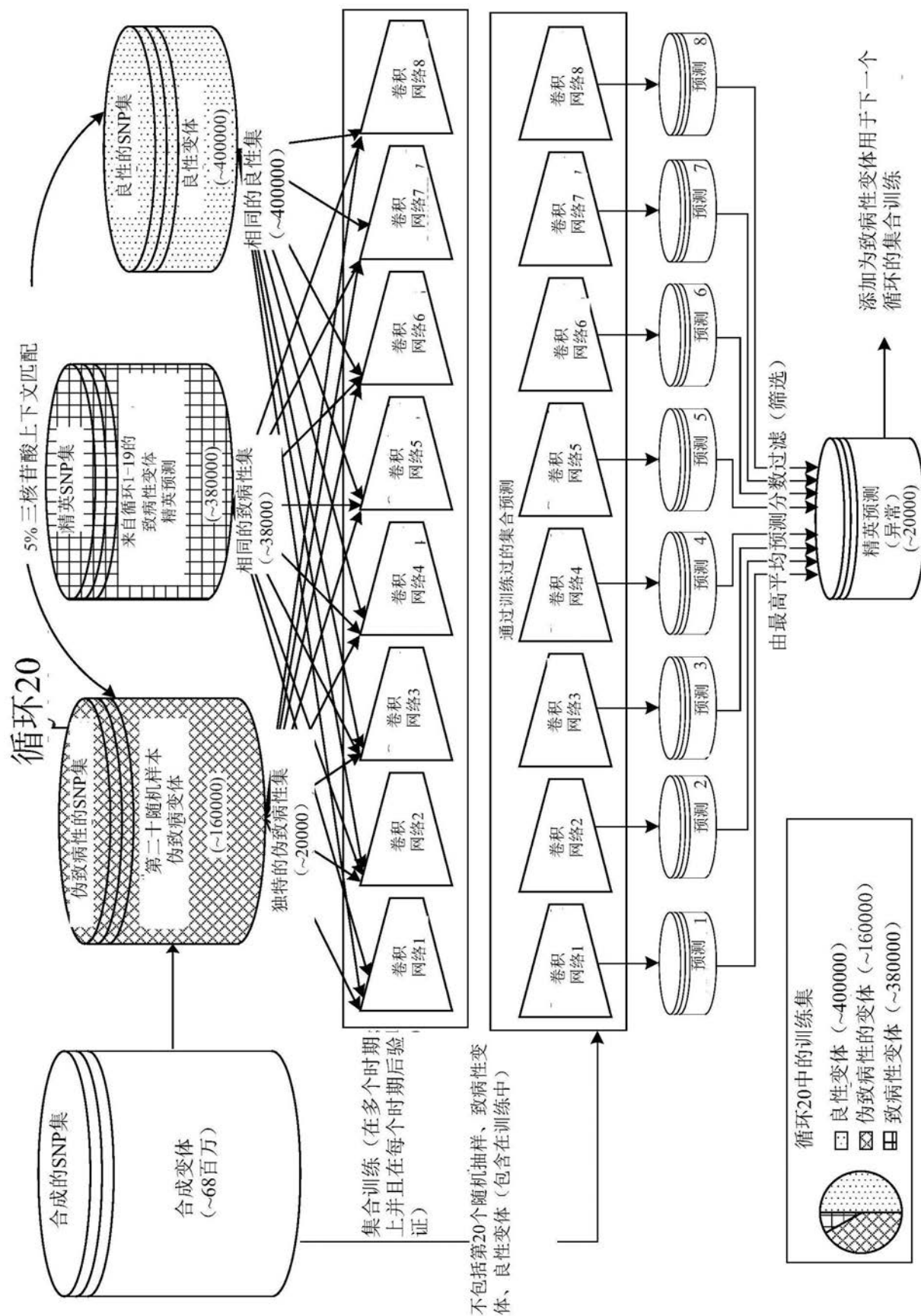


图40

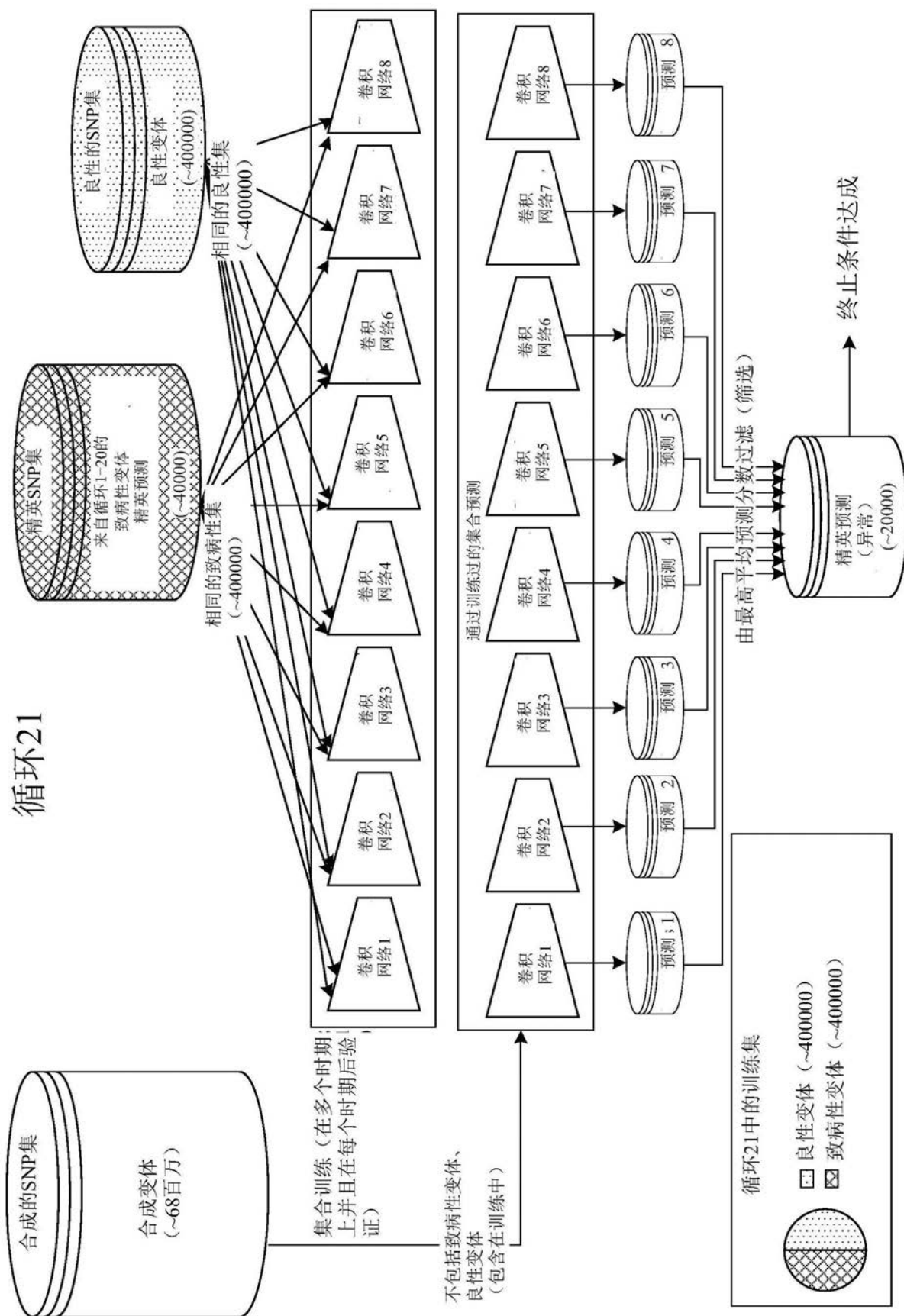


图41

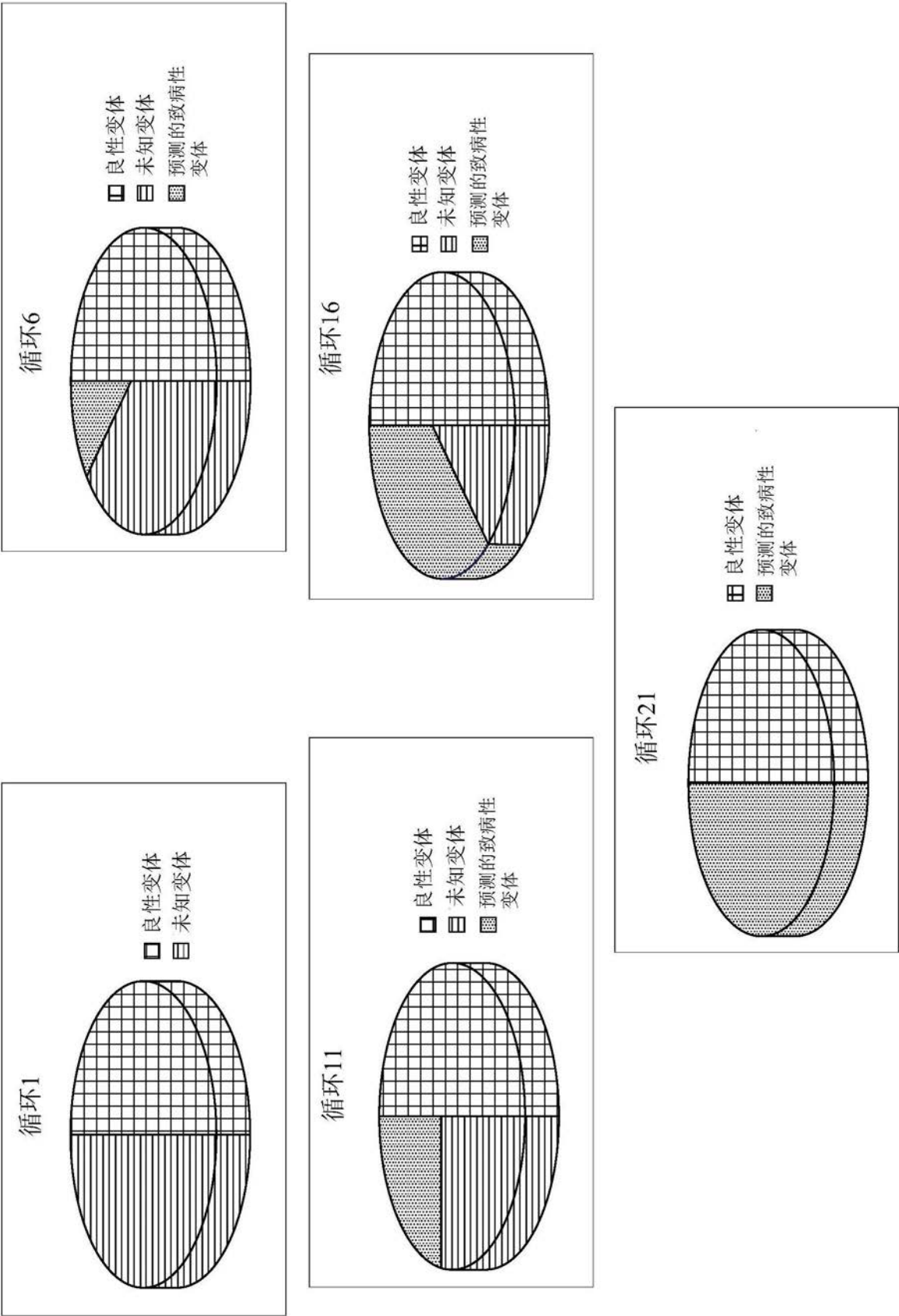


图42

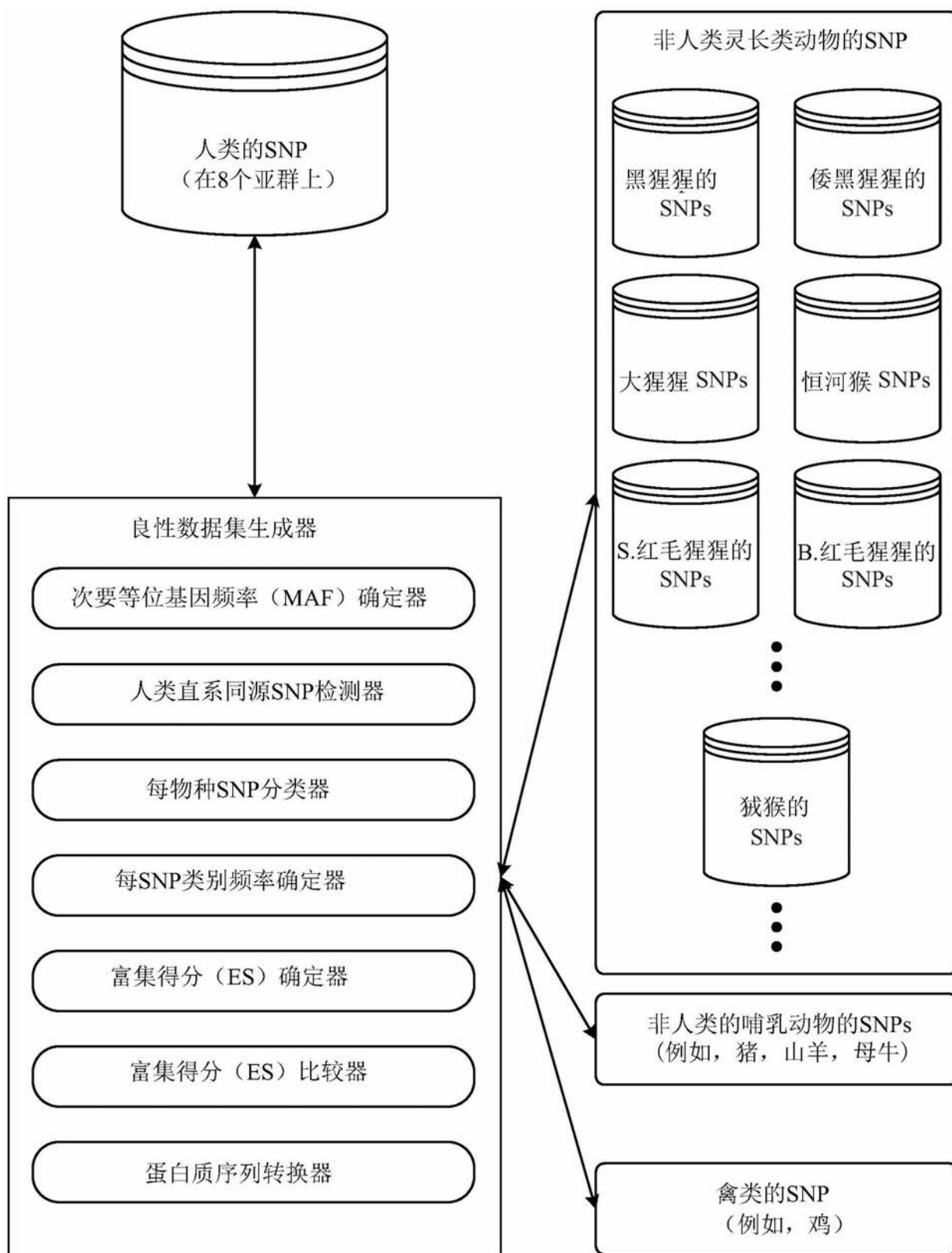


图43

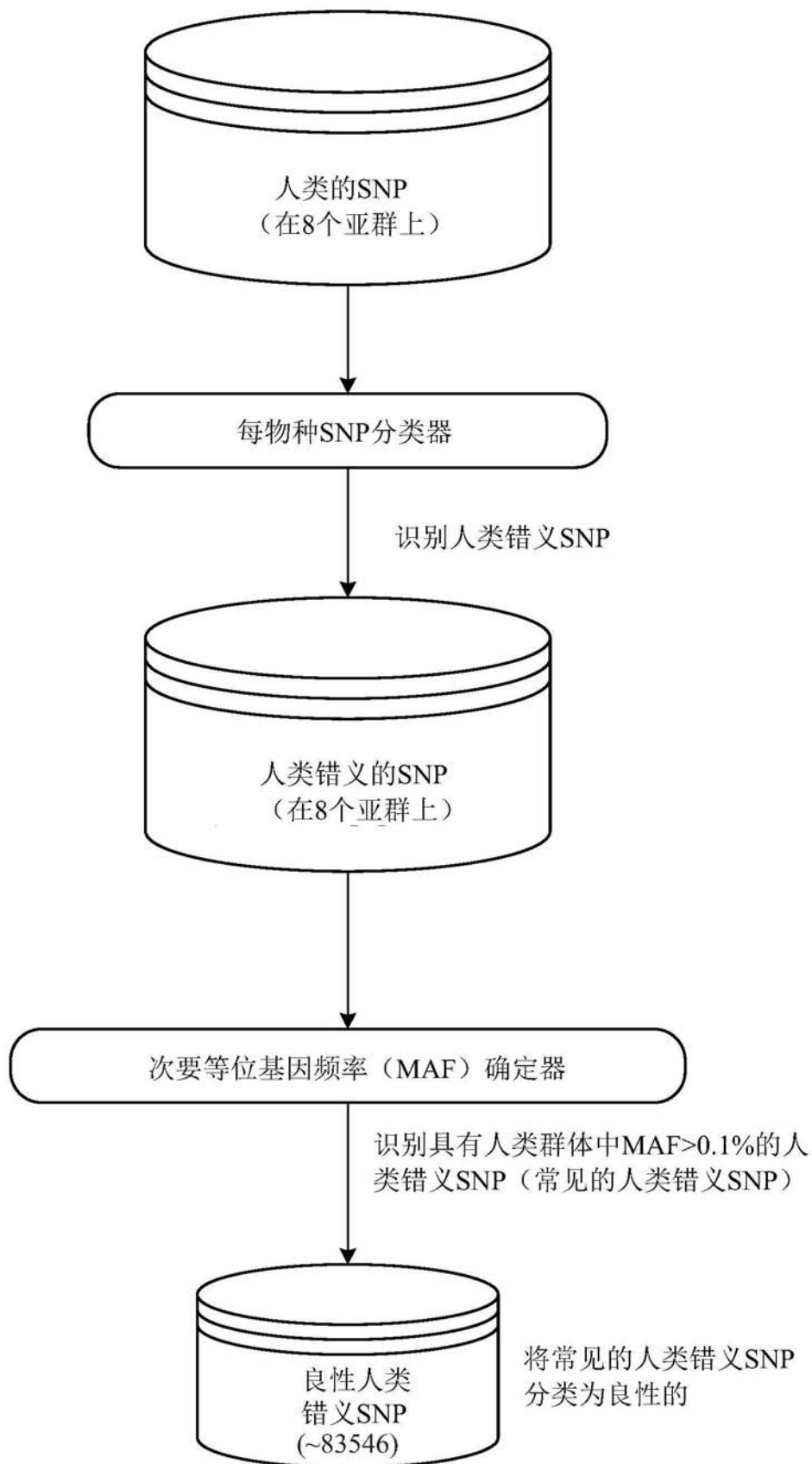
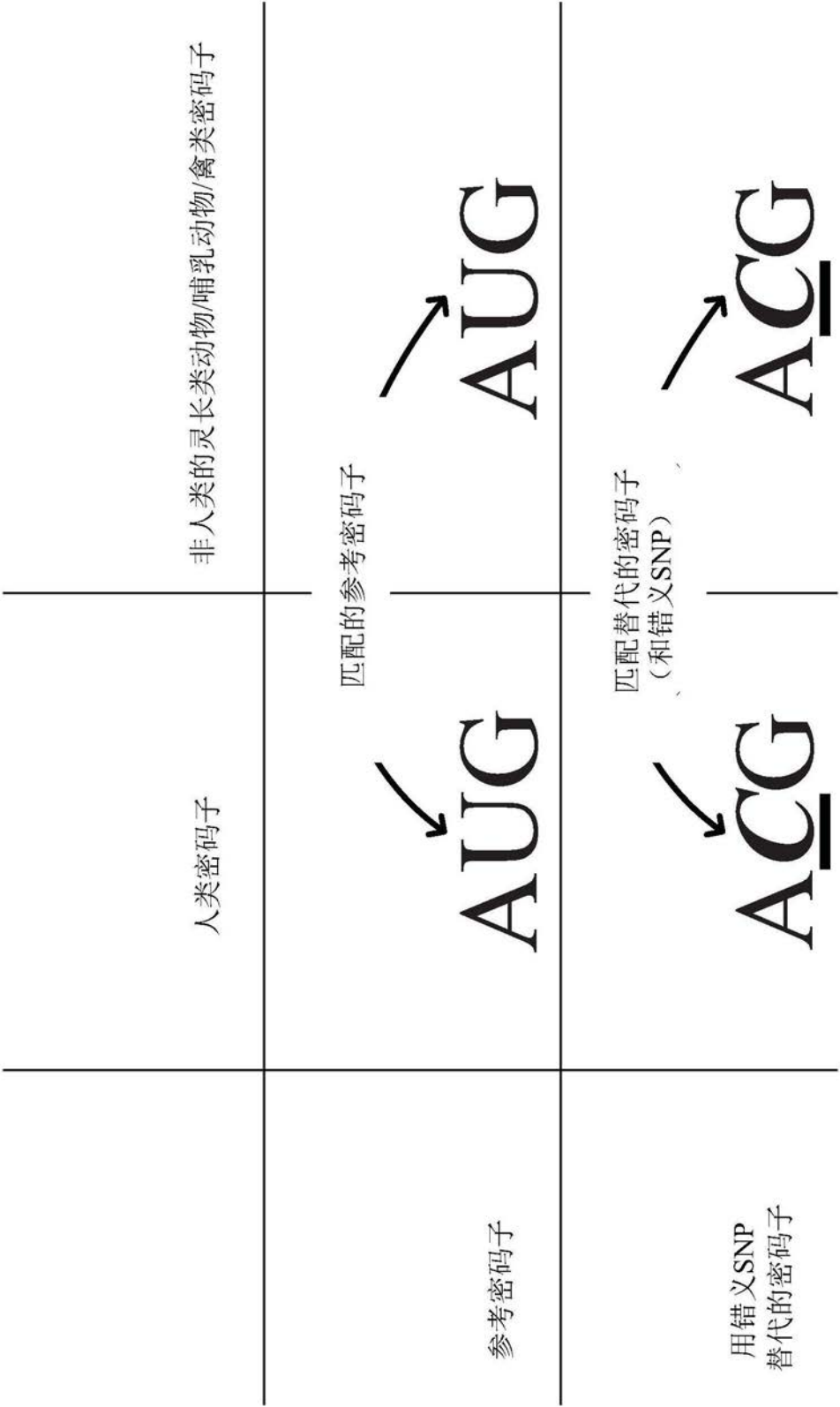


图44



人类直系同源错义SNP : 在具有与人类匹配的参考和替代密码子的非人类物种中的错义SNP

图45

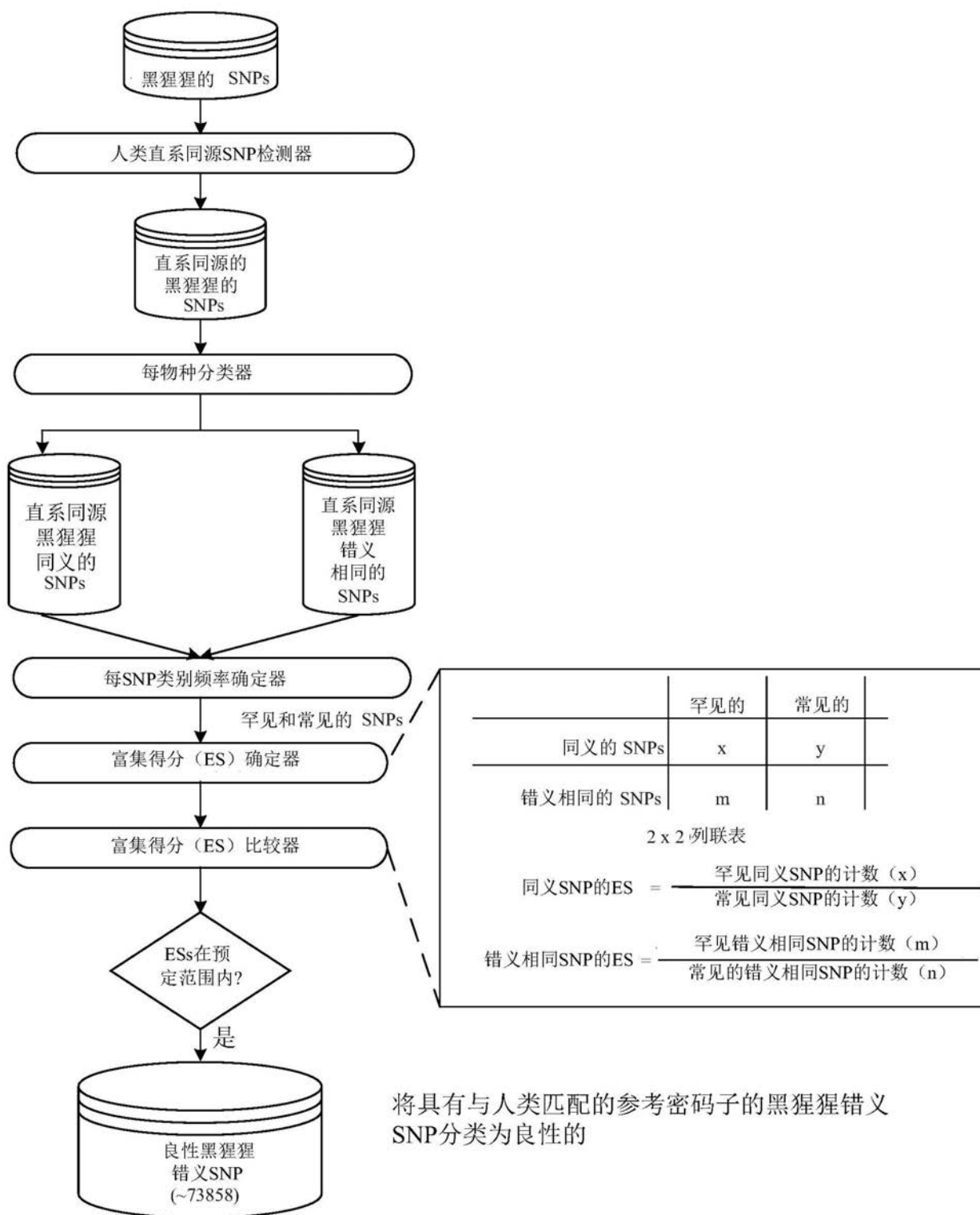


图46

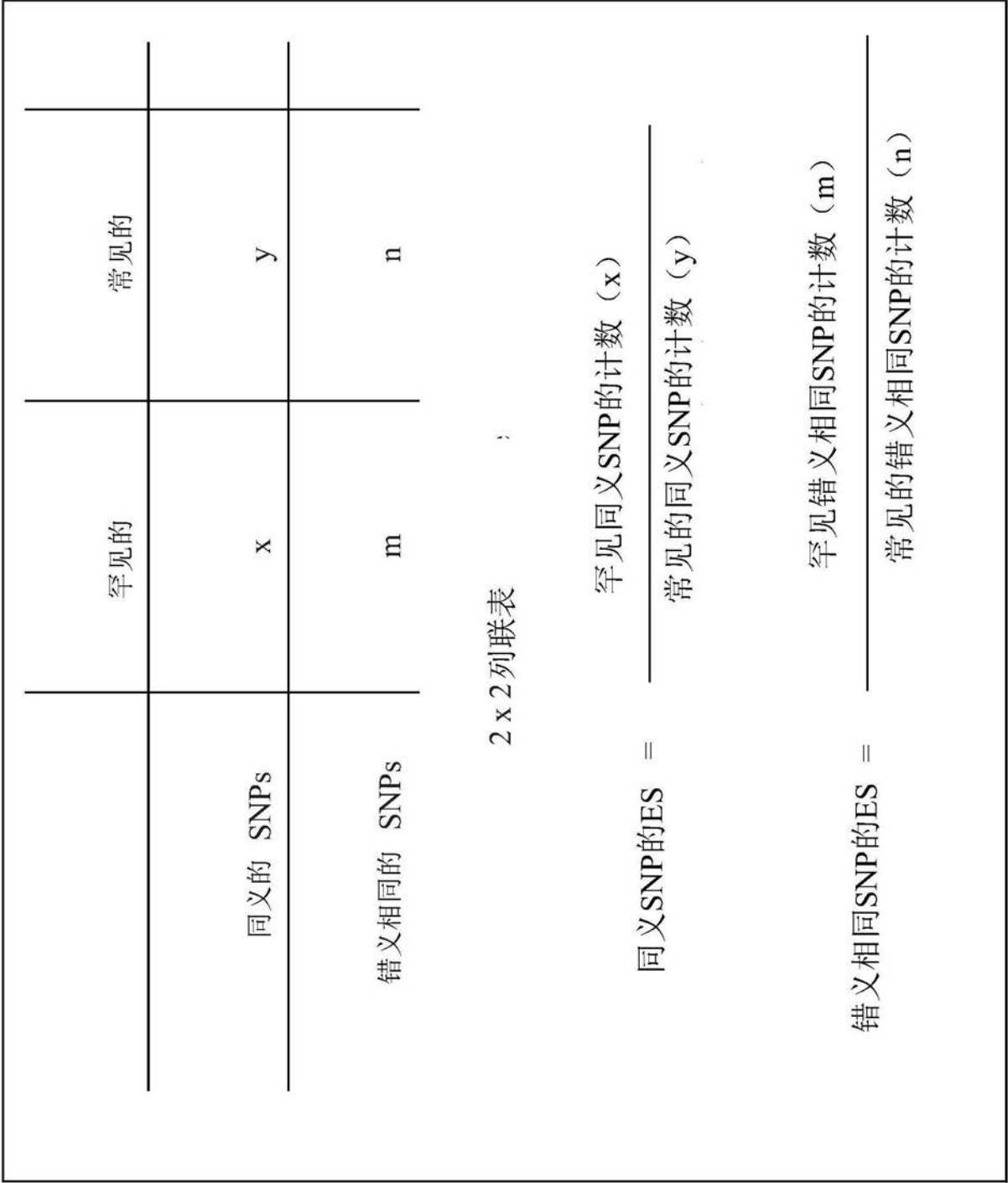


图47

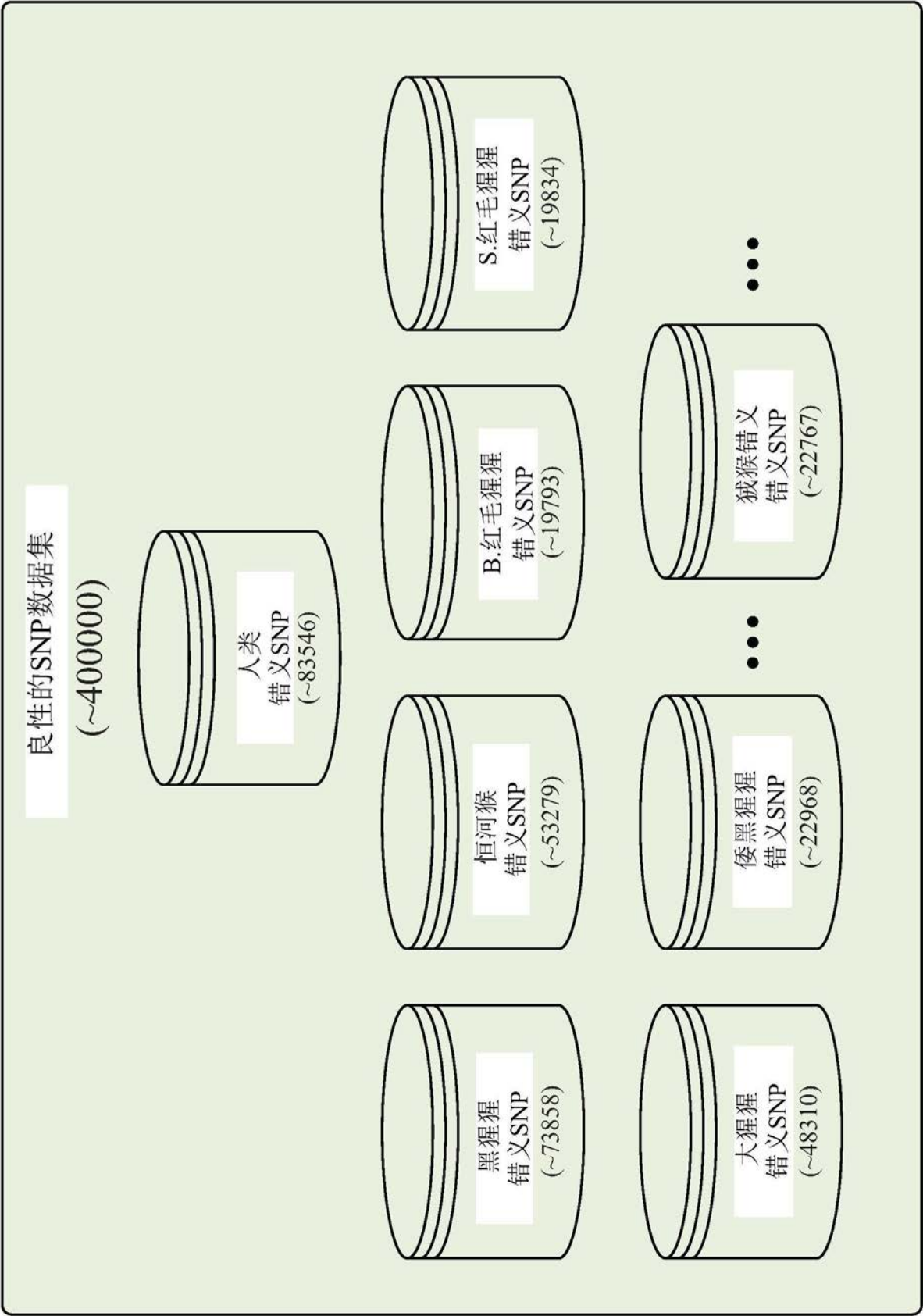


图48

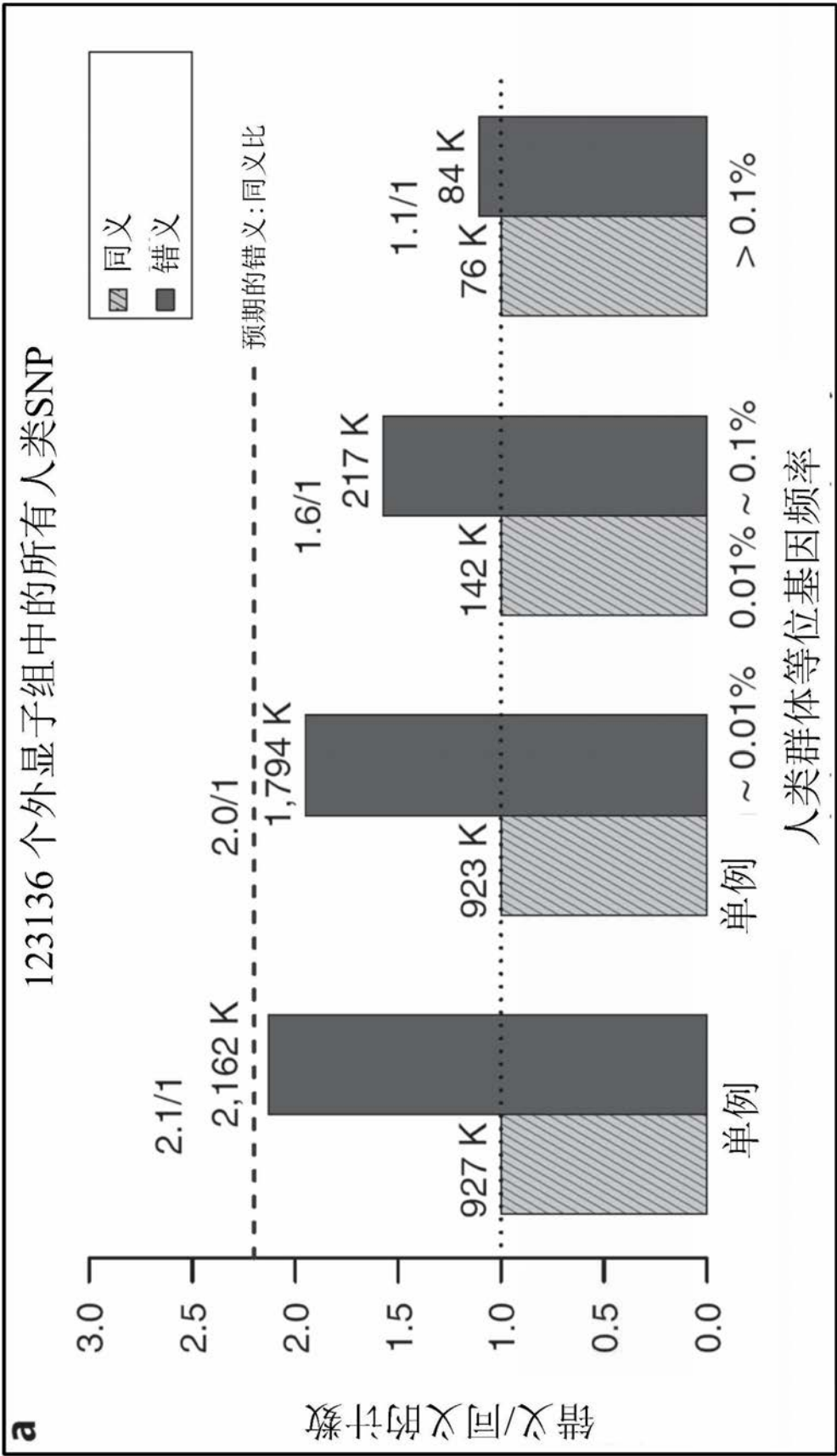


图49A

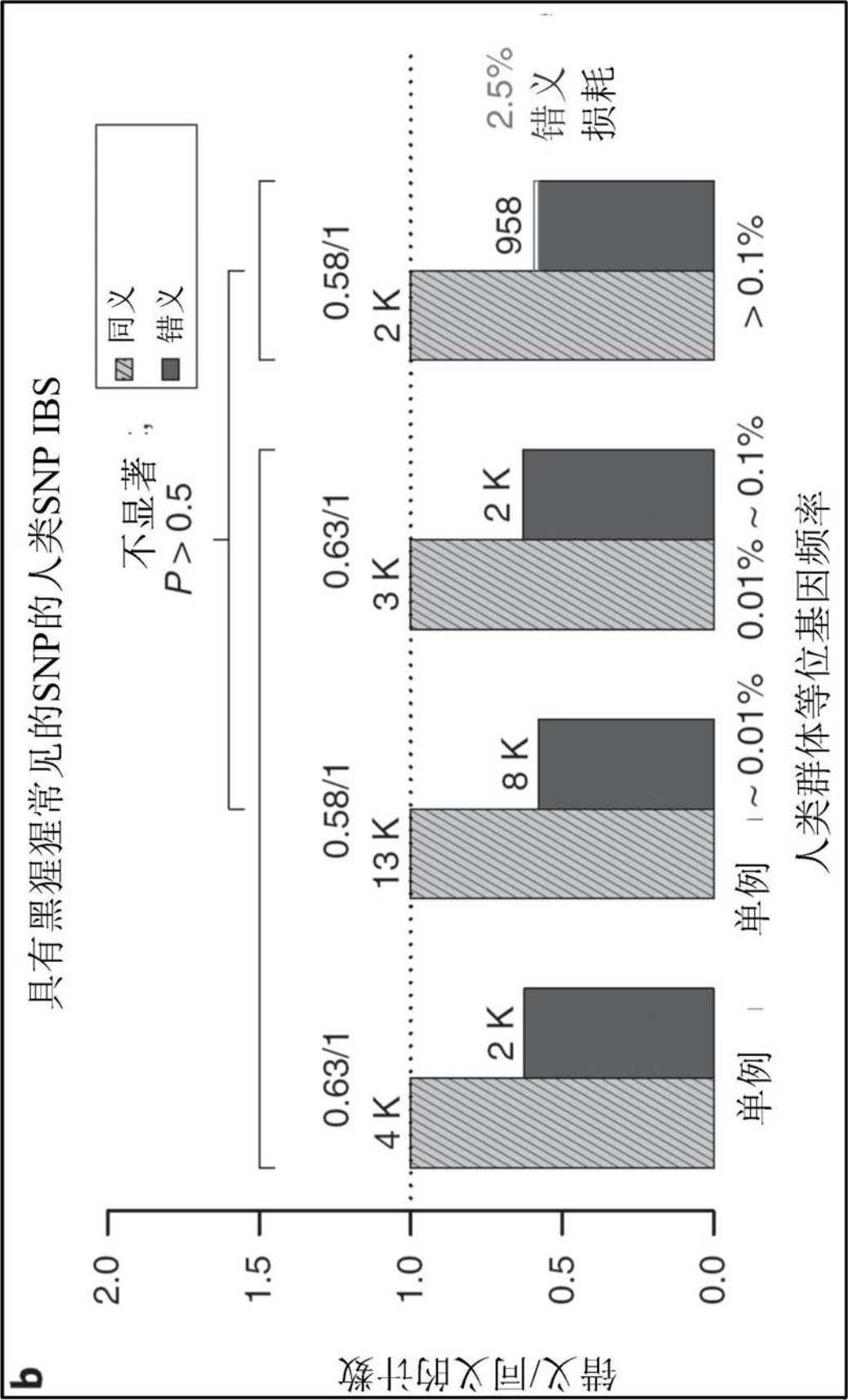


图49B

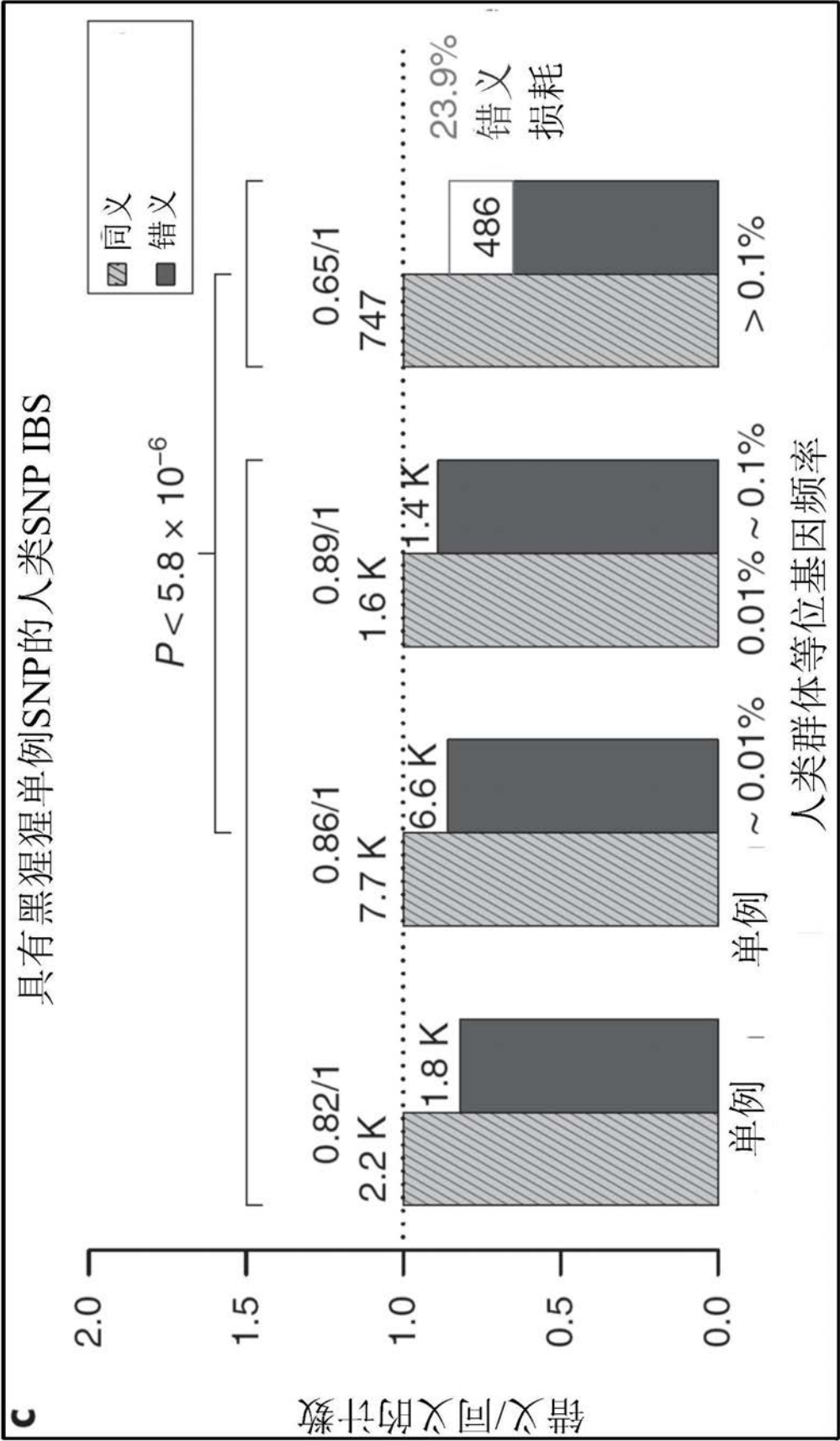


图49C

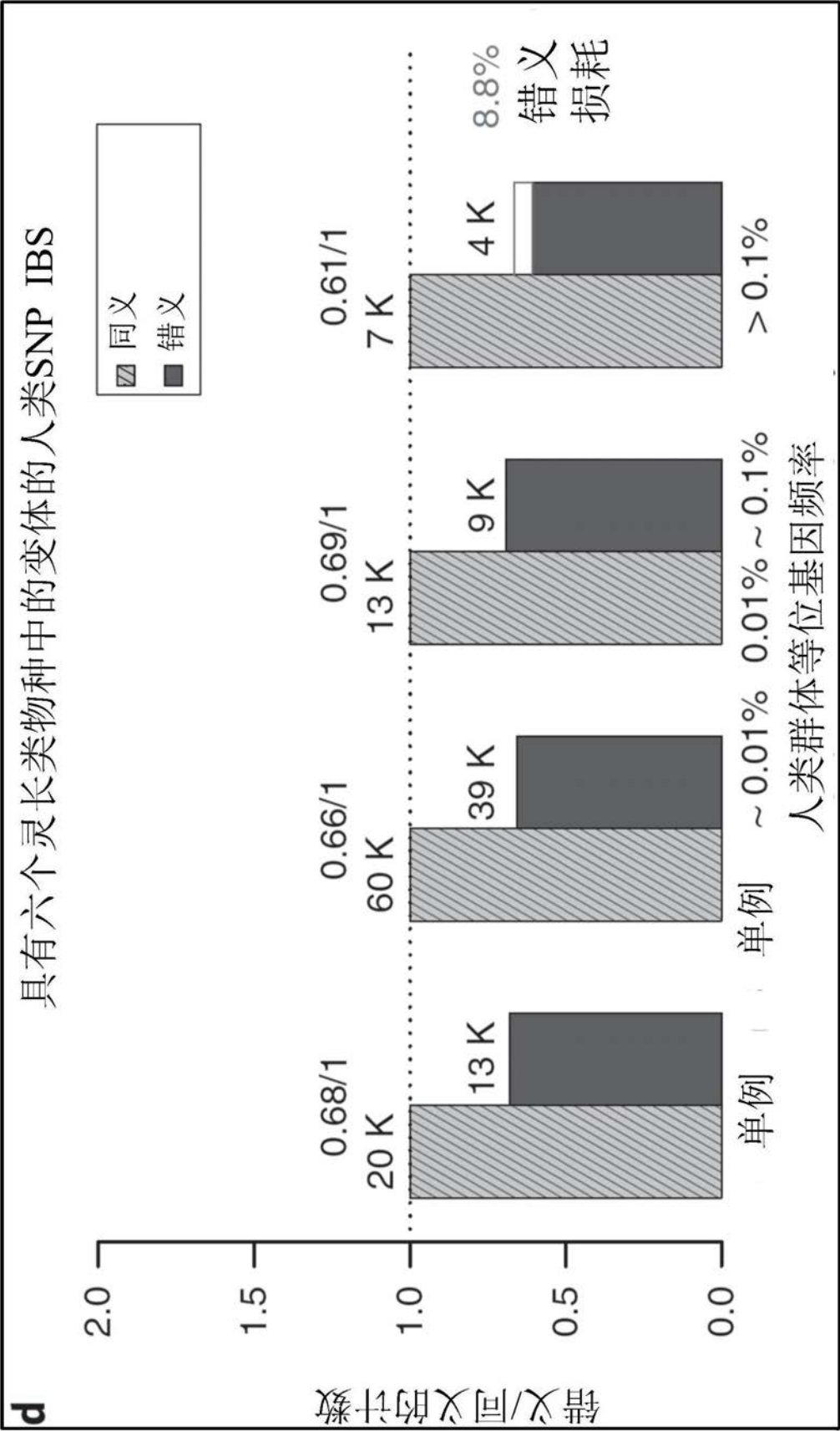


图49D

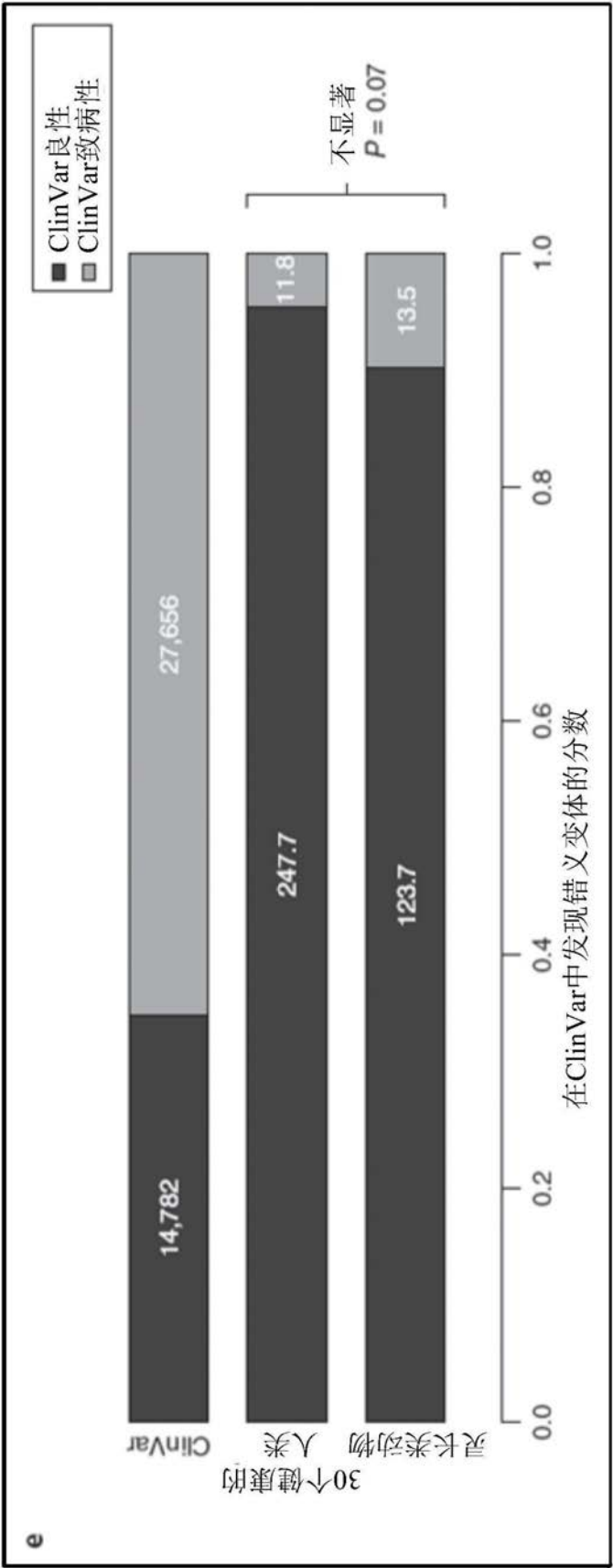


图49E

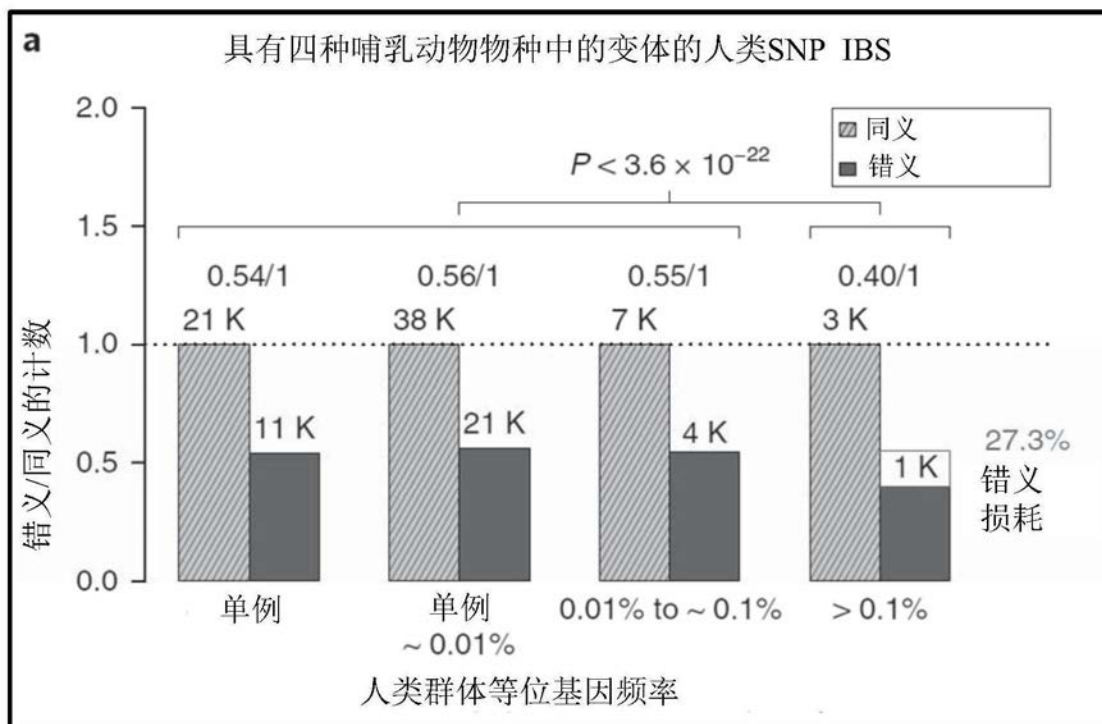


图50A

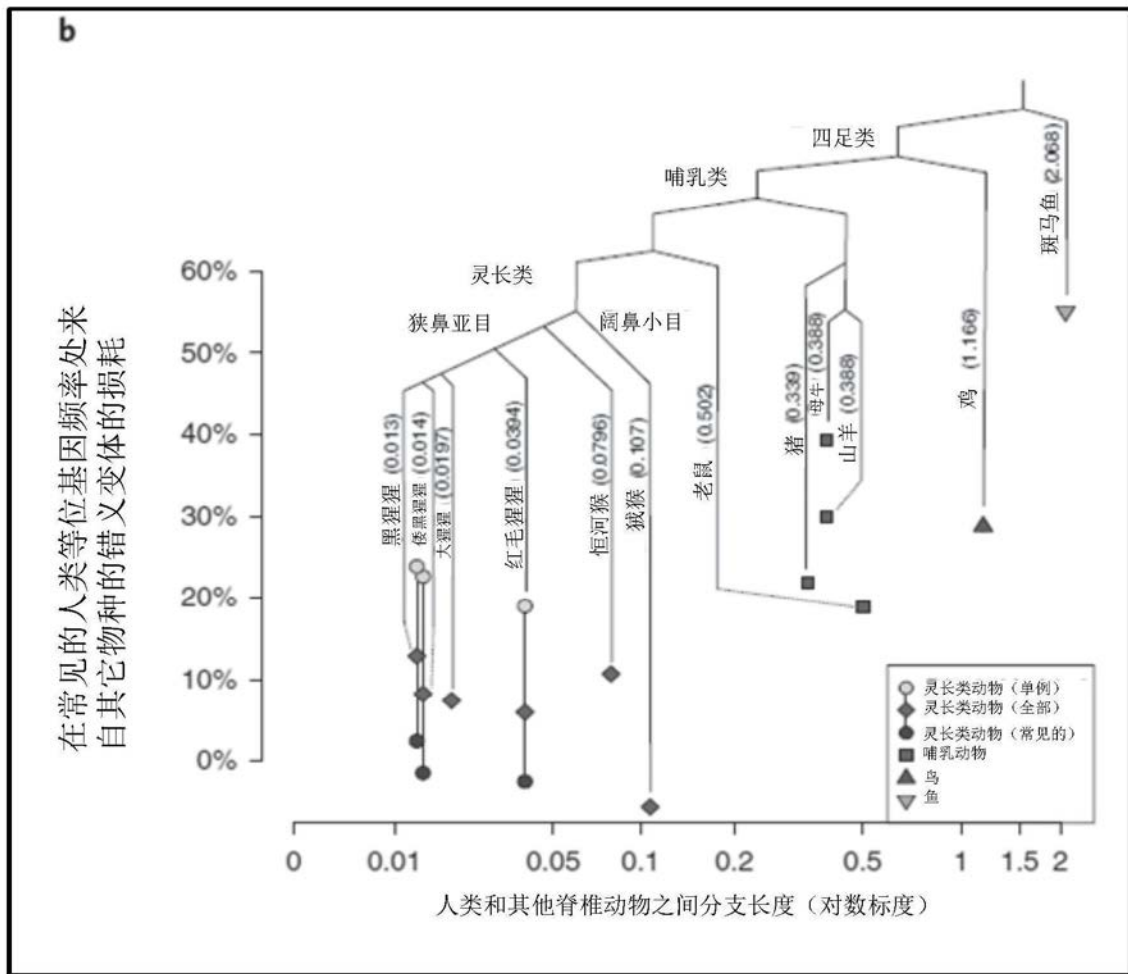


图50B

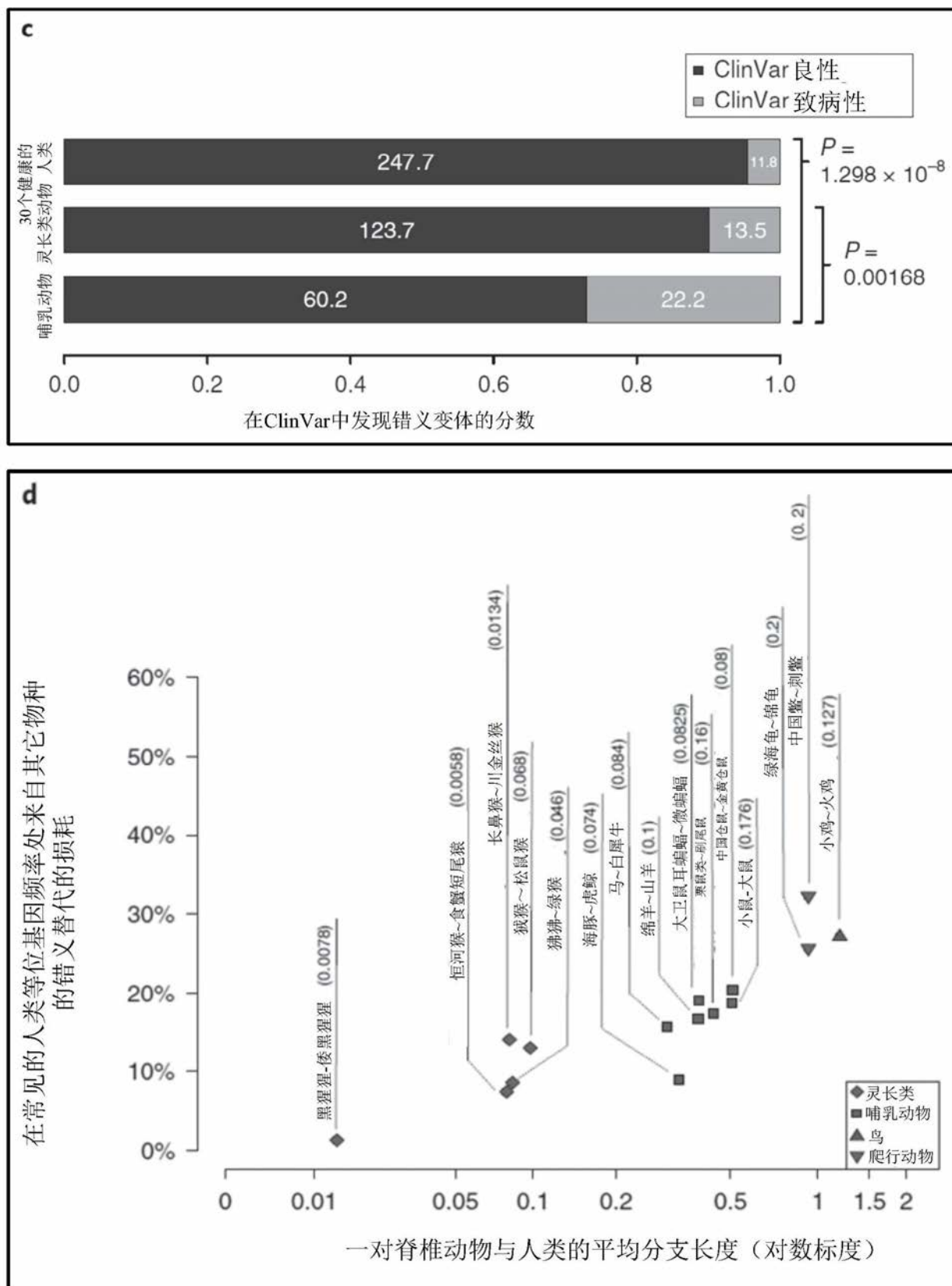


图50C 50D

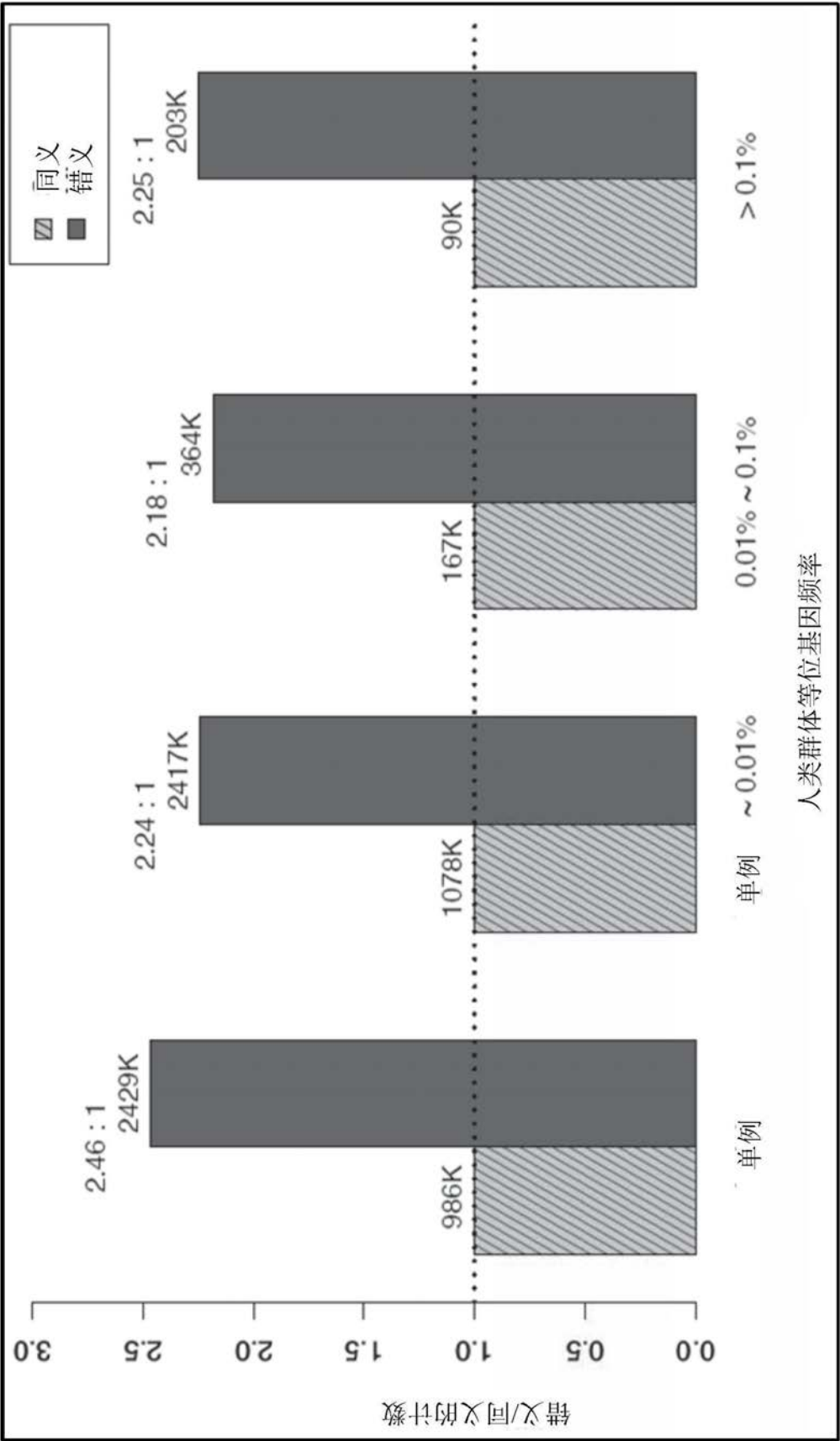


图51

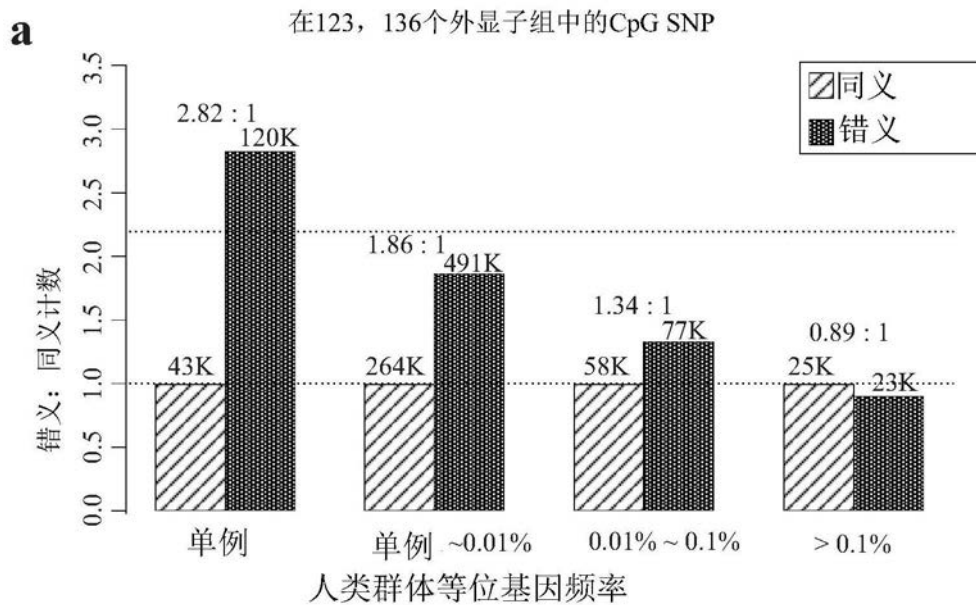


图52A

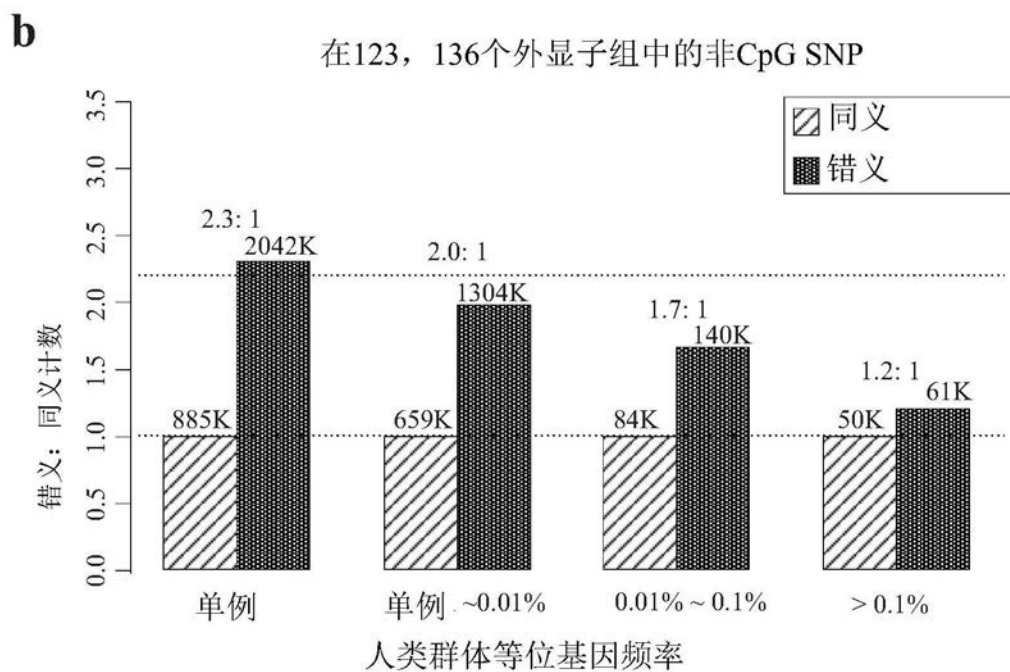


图52B

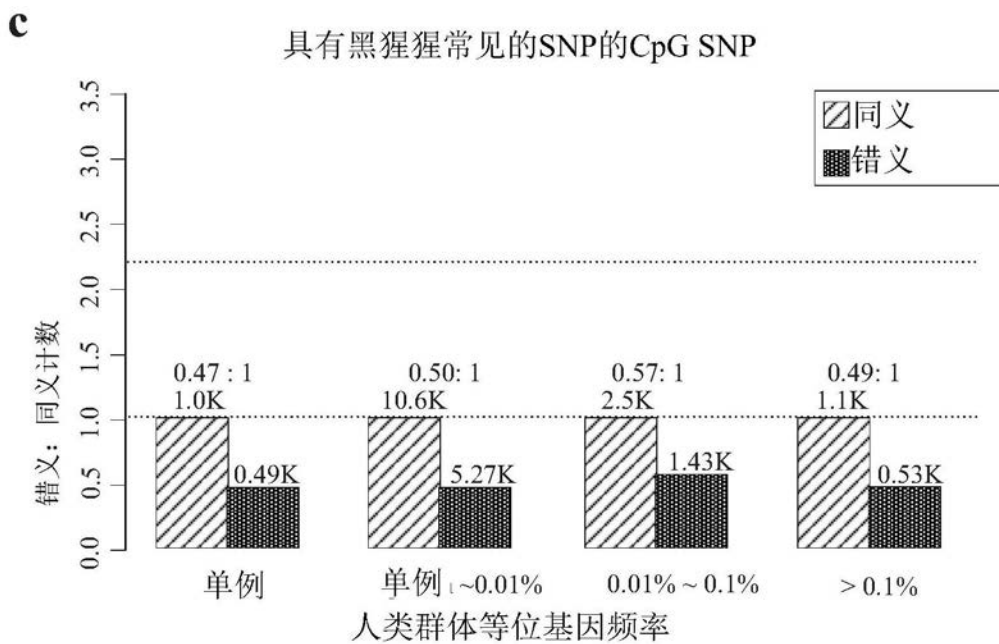


图52C

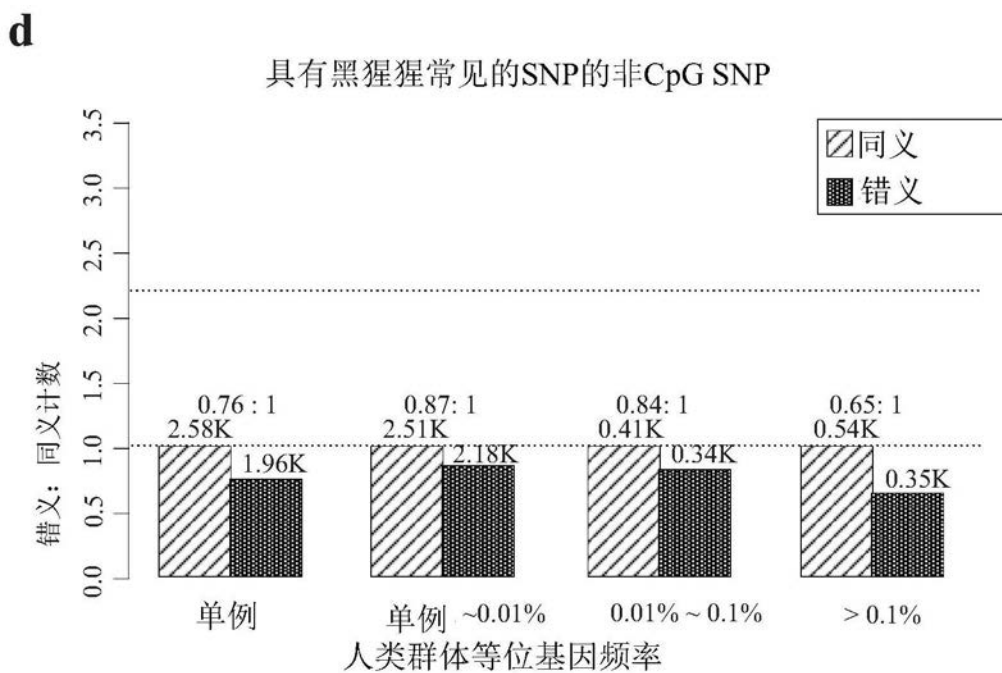


图52D

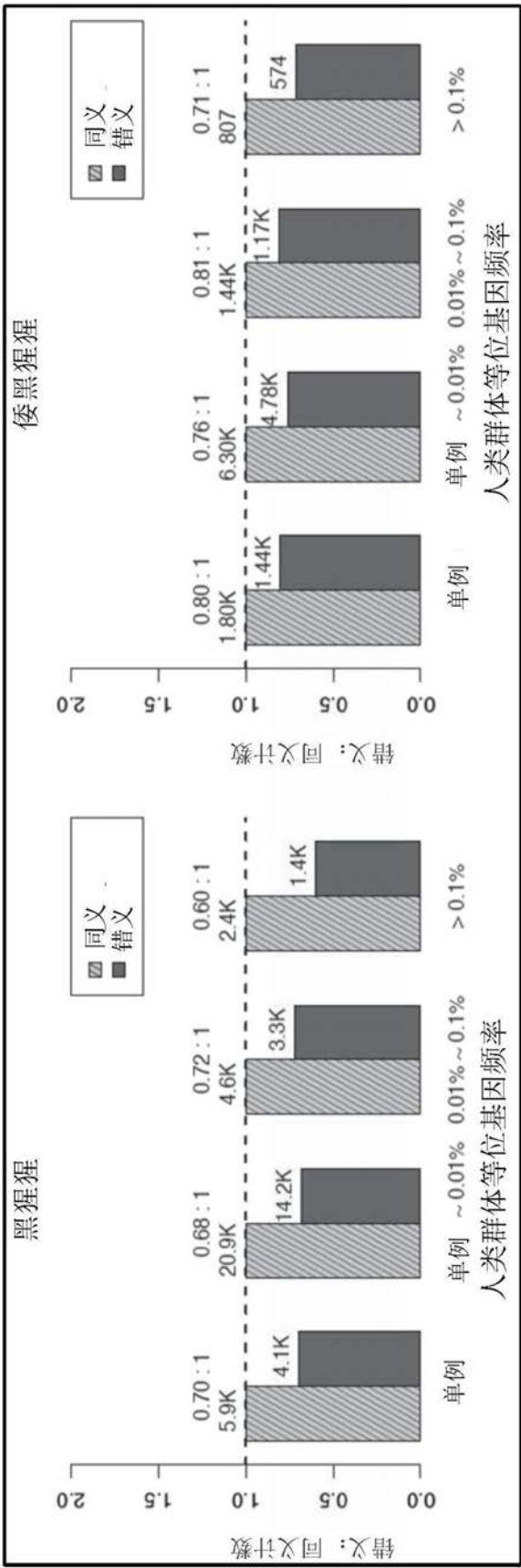


图53

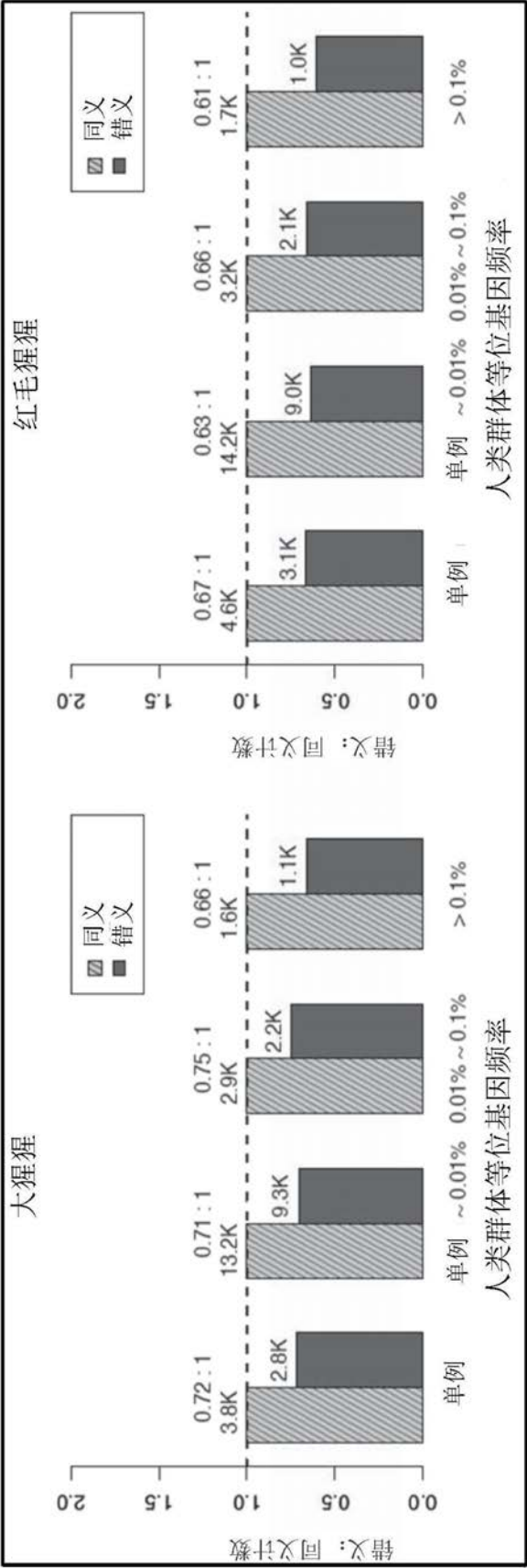


图54

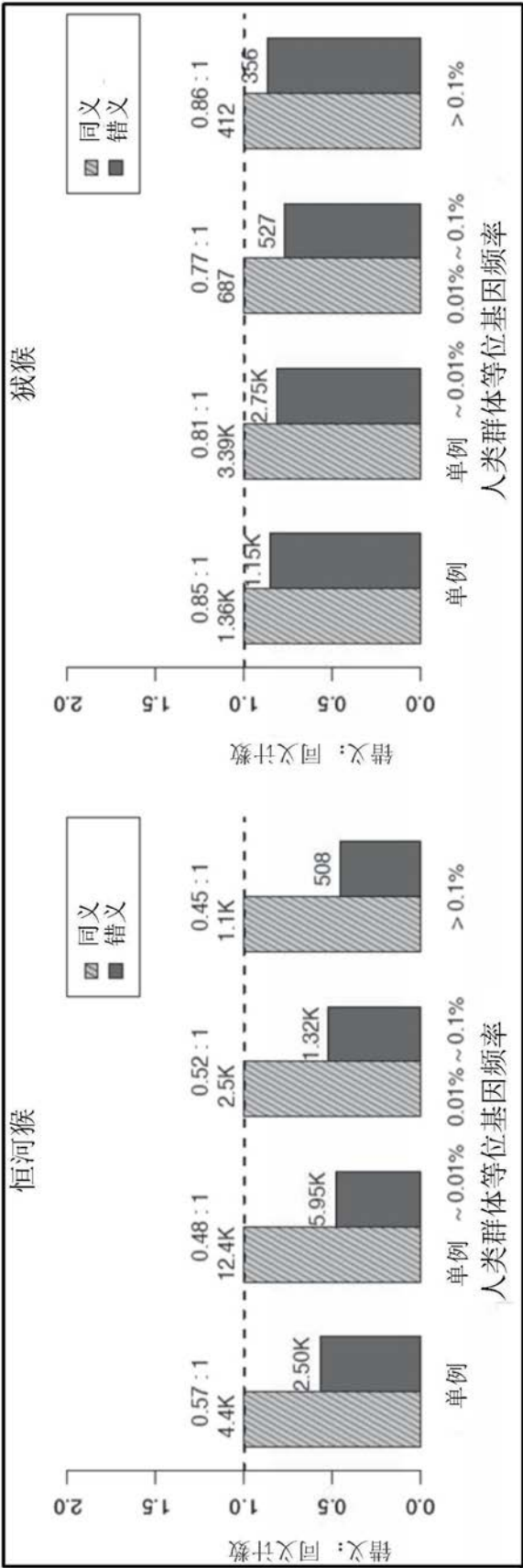


图55

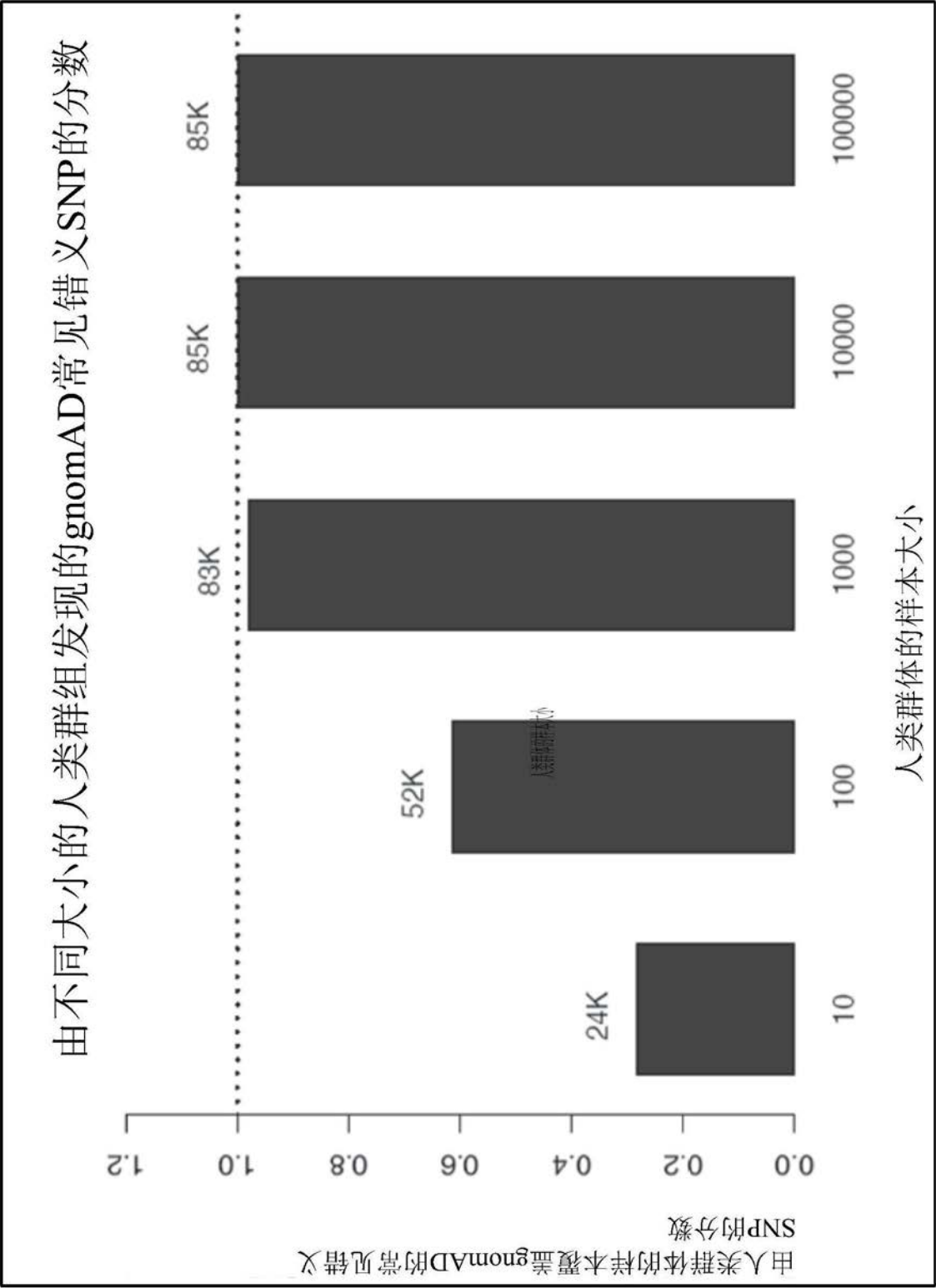


图56

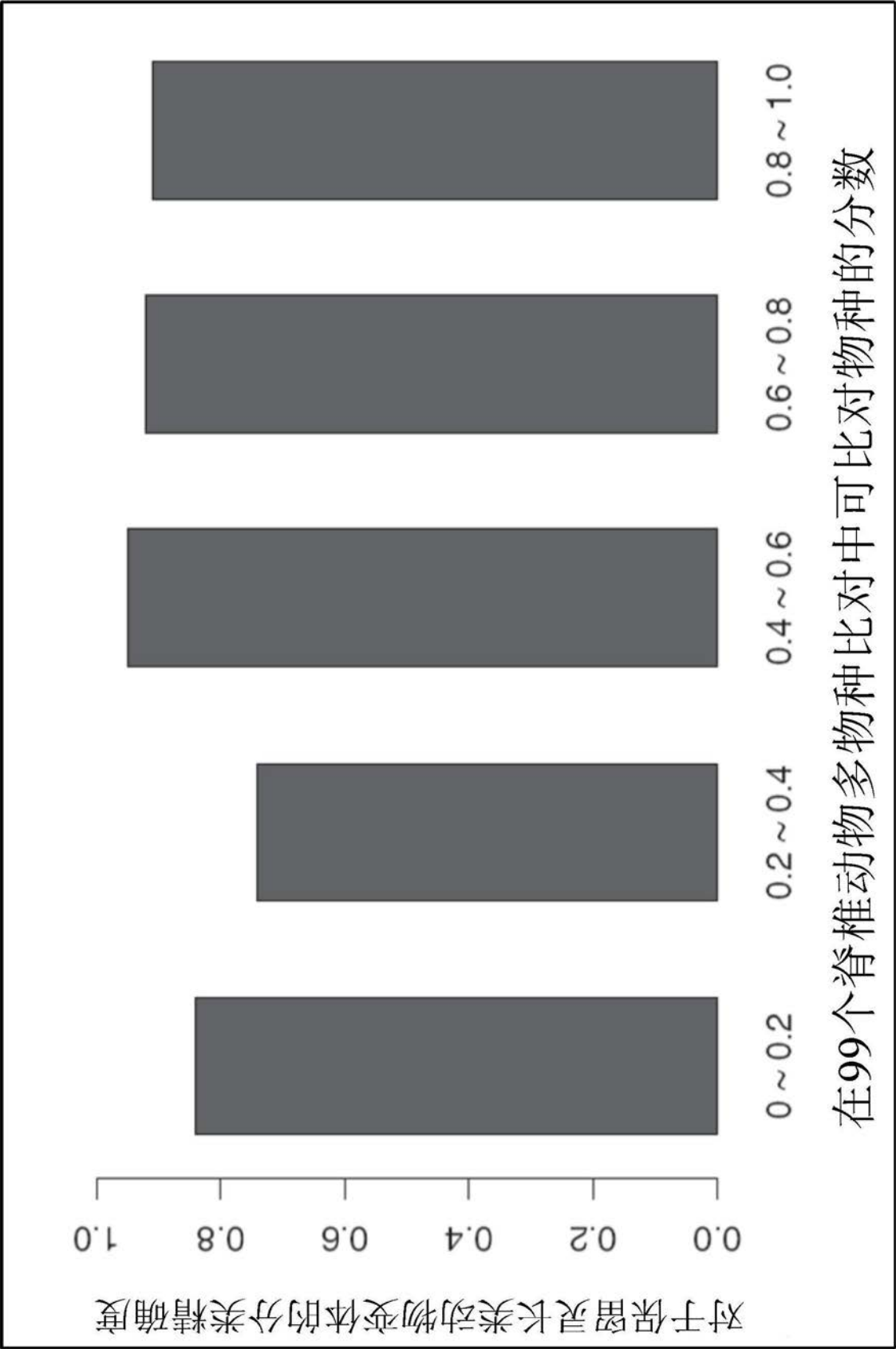


图57

物种 (和个体数)	来源	与人类密码子匹配的 错义变体的数量	贡献的独特错义 变体的数量
人类常见 变体 (MAF > 0.1%, 123136 个体)	gnomAD / ExAC 数据库	83,546	83,546
黑猩猩 (24)	Prado-Martinez 等人 自然 (2013)	76,293	73,858
倭黑猩猩 (13)	Prado-Martinez 等人 自然 (2013)	25,277	22,968
大猩猩 (27)	Prado-Martinez 等人 自然 (2013)	52,052	48,310
红毛猩猩, 波罗洲 亚种 (5)	Prado-Martinez 等人 自然 (2013)	21,621	19,793
红毛猩猩苏门答腊 亚种 (5)	Prado-Martinez 等人 自然 (2013)	27,567	19,834
黑猩猩 (35)	de Manuel 等人 科学 (2016)	75,580	30,327
红毛猩猩 (无个体数据)	dbSNP	18,627	10,554
恒河猴 (16, 无个体数据)	dbSNP	62,393	53,279
狨猴 (9, 无个体数)	dbSNP	25,048	22,767
			385,236

图58

在123136个外显子组 中的人类群体等位 基因频率	预期的 同义变体 的数量	预期的 错义变体的 数量	错义:同义比
单例	986,141.22	2,429,287.85	2.46
单例 ~0.01%	1,077,630.37	2,416,751.61	2.24
0.01%~0.1%	166,849.26	364,292.75	2.18
>0.1%	90,221.42	202,918.38	2.24

图59

感兴趣的ClinVar 变体总数	
666,403	原始的ClinVar文件
324,698	仅保存的hg19记录
264,623	仅提取单个核苷酸变体
186,769	除去非编码的变体
	忽略所有的同义变化和错义变化
122,884	创建或破坏终止密码子
42,438	排除意义不明的变体和具有不一致注释的变体

图60

物种	在ClinVar中具有良性 注释的变体的数量	在ClinVar中具有致病性 注释的变体的数量
黑猩猩	218	21
倭黑猩猩	85	7
大猩猩	167	23
红毛猩猩	160	12
恒河猴	87	11
狨猴	25	7
老鼠	30	4
猪	74	30
母牛	77	39
山羊	60	16
鸡	9	8
斑马鱼	2	6

图61

表1/当只考虑具有PrimateAI得分>=0.803的错义新生突变（DNM）时在智力障碍中实现全基因组显著性的额外基因					
HGNC 符号	蛋白质 截断变体	错义	p值	在多个个体中观察到的 表型异常	
		PrimateAI 得分≥0.803	所有错义	PrimateAI 得分≥0.803	所有错义
ACTL6B	0	3	3	1.5×10 ⁻⁷	2.4×10 ⁻⁶
EBF3	3	3	3	5.2×10 ⁻⁸	5.4×10 ⁻⁶
EFTUD2	2	4	4	1.5×10 ⁻⁷	1.5×10 ⁻⁵
HECW2	1	8	8	2.8×10 ⁻¹⁰	6.7×10 ⁻⁷
KDM6A	2	3	3	2.3×10 ⁻⁷	9.8×10 ⁻⁶
KIF5C	0	3	3	3.0×10 ⁻⁷	2.8×10 ⁻⁶
MAP2K1	0	5	5	3.1×10 ⁻⁸	2.7×10 ⁻⁶
PPP1CB	0	6	6	1.5×10 ⁻⁸	1.6×10 ⁻⁶
PRKD1	0	6	6	8.6×10 ⁻⁸	1.7×10 ⁻⁵
SOX11	1	3	3	3.1×10 ⁻⁷	2.4×10 ⁻⁵
TBR1	4	4	4	1.3×10 ⁻¹⁰	4.2×10 ⁻⁷
TLK2	3	5	5	4.7×10 ⁻⁹	6.3×10 ⁻⁷
TRIP12	6	2	4	1.4×10 ⁻⁷	5.4×10 ⁻⁷
U2AF2	0	4	4	2.6×10 ⁻⁷	1.2×10 ⁻⁵
提供了蛋白质截断和错义DNMS的计数。当只用primateAI得分>=0.803的错义突变运行统计测试并且当其对于所有错义突变都重复时，显示了基因富集的p值。					

图62

表2 | 在ClinVar中的人类专家注释的变体与DDD病例对比对照例的新生变体之间在Grantham得分蛋白质表面暴露和氨基酸序列保守性方面差异的比较

	Grantham 得分	蛋白质 表面 暴露	序列 保守性
ClinVar致病 变体	91.1	0.53	0.87
ClinVar良性 变体	67.4	0.41	0.54
人类专家 注解中的差异	+23.7	+0.12	+0.33
DDD患者中 的新生变体	84.9	0.51	0.90
健康对照中的 新生变体	72.7	0.29	0.73
在受影响个体对比未 受影响个体中的差异	+12.2	+0.22	+0.17

对于ClinVar数据库中具有非矛盾注释的错义突变，并且对于在605个疾病-关联的基因内存在于DDD病例对比对照例中的新生变体，显示了平均得分。蛋白表面暴露反映由溶剂可及性神经网络预测为暴露的残差的氨基酸的分数，而序列保守性示出了具有在100个脊椎动物比对都序列相同的氨基酸的分数。粗体数字突出显示了人类专家偏爱的试探法与经验数据相比的差异。

图63

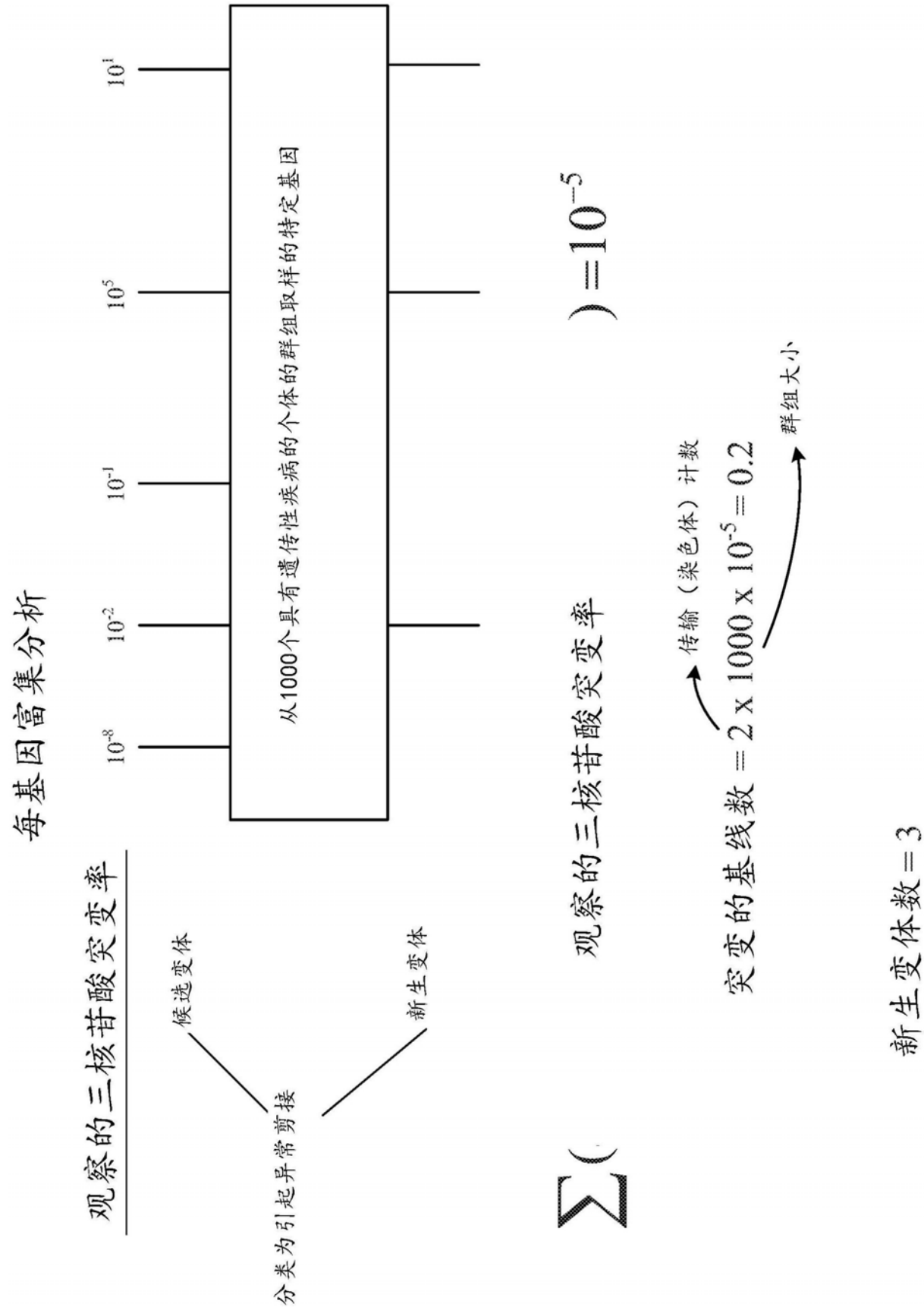


图64

全基因组富集分析

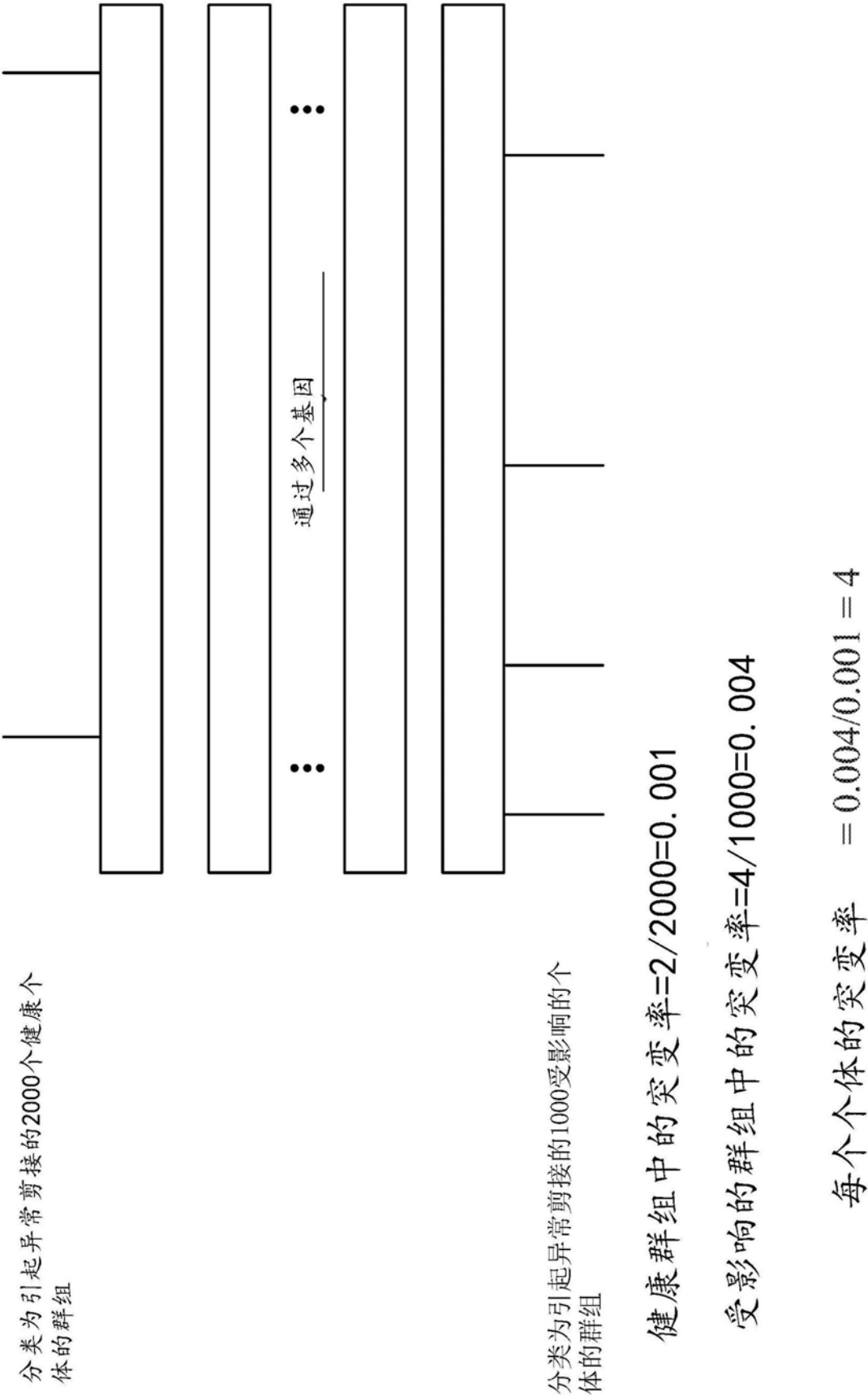


图65

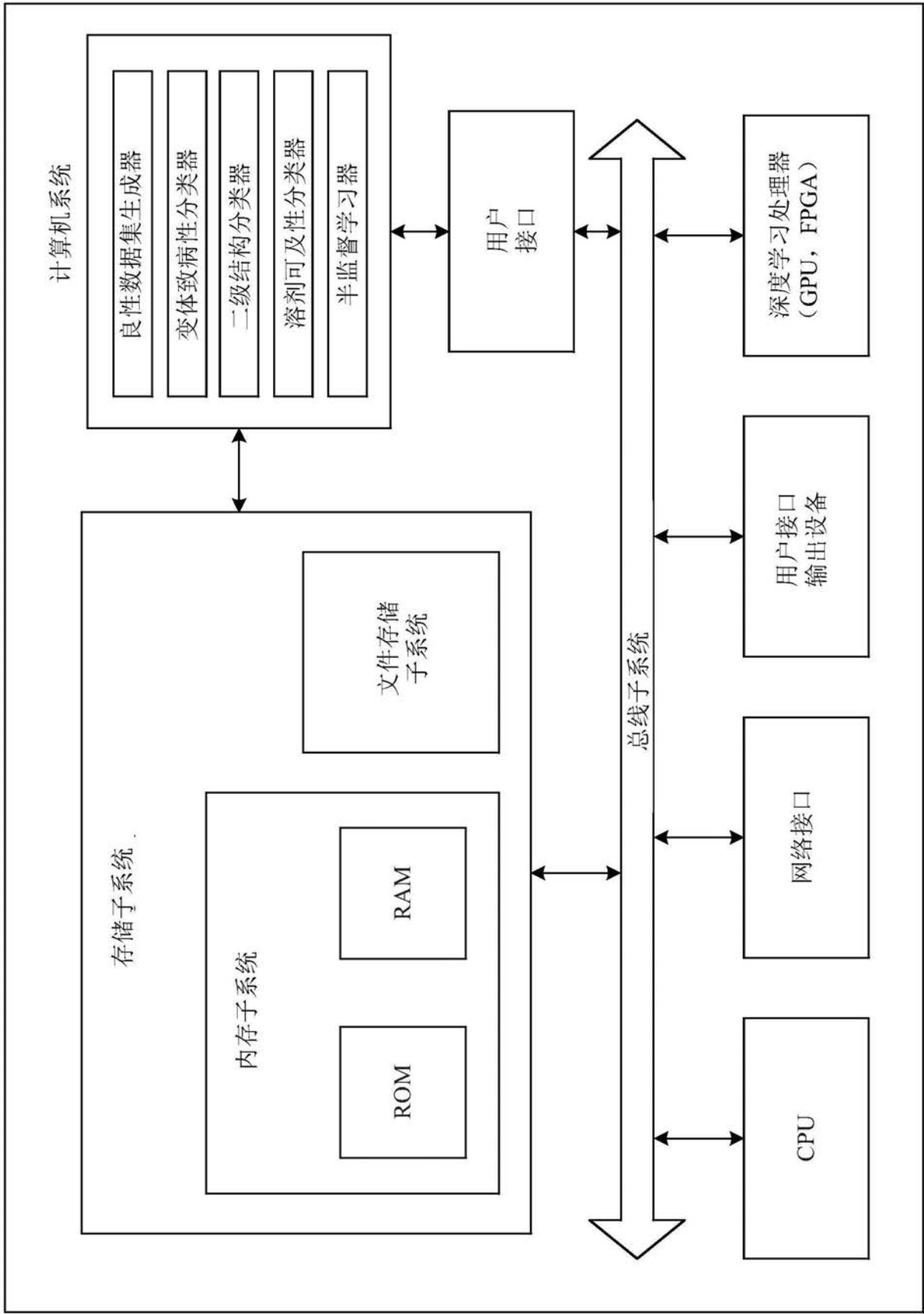


图66