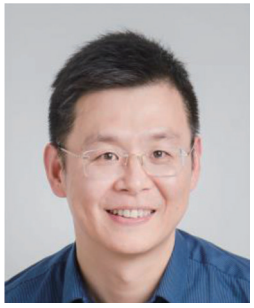

杨力，博士、研究员、博士生导师，长期从事生物大数据组学分析及相关技术创新体系研究，主要利用计算生物学和实验生物学相结合的交叉研究思路和方法手段，针对多维生物组学大数据开展分析研究，并取得了一系列突破性进展，为环形RNA这一全新长非编码RNA前沿研究领域开拓了新的思路，为转录组多水平的复杂性调控研究奠定了基础，为基因组编辑技术在单碱基水平的应用提供了新的体系。至今，共发表研究论文及综述/专评等70余篇，包括在*Cell*、*Nat Biotechnol*、*Cell Stem Cell*、*Mol Cell*和*Genome Res*等期刊作为通讯或共同通讯作者发表研究和综述论文30余篇，被引用4 000余次。

计算生物学分析在基因组编辑研究中的应用

王滢[1#]，熊义春[1#]，陈佳[2]，杨力[1,2*]

([1]中国科学院计算生物学重点实验室，中国科学院-德国马普学会计算生物学伙伴研究所，
中国科学院上海生命科学研究院营养与健康研究院，中国科学院大学，上海 200031；
[2]上海科技大学生命科学与技术学院，上海 200031)

摘要：基因组编辑是对基因组遗传信息进行定向改造的技术，其中CRISPR/Cas系统是目前应用最广泛的基因组编辑新技术。将先进的高通量测序以及相关计算生物分析应用于基因编辑研究，可进一步优化基因编辑效率和精度等检测流程，实现对全基因组功能基因筛选的监测。同时，利用基于生物信息及机器学习和深度学习等新方法，可对向导RNA(gRNA)的高效设计和实现对编辑效果的预测。本文将对计算生物学分析在CRISPR/Cas基因编辑系统的应用及研究进展等进行概述。

关键词：基因组编辑；CRISPR/Cas；gRNA；高通量测序；计算生物学；生物信息；机器学习；深度学习

Computational analysis in CRISPR/Cas genome editing

WANG Ying[1#], XIONG Yichun[1#], CHEN Jia[2], YANG Li[1,2*]

([1]Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology,
Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences,University
of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China;
[2]School of Life Science and Technology, ShanghaiTech University, Shanghai 200031, China)

Abstract: Genome editing is a type of genetic engineering technique to alter genomic information, especially to insert, delete or change genomic DNA sequences. Currently, CRISPR/Cas system has been emerging as a precise and powerful tool for genome editing, which is widely used in basic research and holds great potential in therapeutics. The application of high-throughput sequencing and related computational pipelines

收稿日期：2018-12-30
基金项目：国家自然科学基金项目(31730111)
#共同第一作者：王滢，E-mail: wangying@picb.ac.cn；熊义春，E-mail: xiongyichun@picb.ac.cn
***通信作者：**E-mail: liyang@picb.ac.cn

in CRISPR/Cas-mediated genome editing achieves gene functional screening in a genome wide scale, and also improves the analysis of genome editing efficiency and accuracy. Importantly, bioinformatic analyses and machine learning, including deep learning, can also be used to evaluate/optimize gRNA design and to predict the genome editing outcome. Here, we briefly describe the application of computational analyses in CRISPR/Cas-mediated genome editing.

Key Words: genome editing; CRISPR/Cas; gRNA; high-throughput sequencing; computational biology; bioinformatics; machine learning; deep learning

成簇规则间隔回文重复序列(clustered regular interspaced palindromic repeats, CRISPR)是一类在细菌中发现的短重复DNA序列簇，其被细菌基因组中的间隔序列(spacer sequence)隔开，而与CRISPR序列相关的蛋白基因被称为*Cas*基因(CRISPR-associated genes)[1-2]。CRISPR和Cas与细菌的抗噬菌体免疫功能密切相关[3-8]，Cas蛋白广泛地参与了这种细菌免疫过程中的免疫记忆获得、CRISPR RNA(crRNA)生成和对噬菌体DNA的降解。研究表明，当噬菌体入侵细菌时，来自于外源的噬菌体DNA序列可以被细菌的CRISPR/Cas系统识别、处理形成protospacer，并最终插入到细菌CRISPR中成为新的中间间隔序列，从而建立免疫记忆；当细菌再次遭受同一种外源噬菌体DNA入侵时，特定的Cas蛋白可以转录CRISPR并加工产生crRNA，其含有与外源DNA序列互配的中间间隔序列，因此crRNA可以作为向导RNA序列(guide RNA, gRNA)，识别并靶向介导Cas核酸内切酶对入侵的外源DNA进行降解，从而实现细菌对外源DNA的免疫反应[9-10]。利用crRNA可以靶向引导Cas核酸内切酶切割进而降解特定DNA序列的这一特性，Jinek等[3]和Cong等[11]首次将CRISPR/Cas系统应用于对基因组DNA序列的编辑改造，实现了在基因组水平对靶向基因的遗传操作(genetic engineering)。目前，来源于酿脓链球菌(*Streptococcus pyogenes*)的CRISPR/Cas9系统在基因组编辑领域的应用最广泛[12]。Cas9是一类RNA介导的DNA内切酶，通过sgRNA(single guide RNA)序列引导及其对PAM (protospacer adjacent motif)序列的识别，实现靶向DNA序列的切割[12]。其他类型的Cas蛋白，如特异识别A/T富集PAM序列的Cas12(Cpf1)[13-14]、特异识别RNA序列的Cas13家族蛋白[15]也被陆续发现并应用于基因组编辑。值得注意的是，不同Cas蛋白需要的向导RNA序列不同，Cas9需要整合tracrRNA (trans-activating crRNA)和crRNA产生的sgRNA序列作为引导，而Cas12和Cas13只需要crRNA作为gRNA序列引导基因组编辑。在本文中，除了在具体介绍基于Cas9的基因组编辑时，其他情况我们都使用gRNA进行描述。

基于CRISPR/Cas系统的应用得到了井喷式的发展，在诸如基因表达调控、表观基因组编辑、染色体成像、基因组碱基编辑等领域取得了一系列前所未有的突破[16-19]。更为重要的是，将先进的高通量深度测序(high-throughput deep sequencing)及相关计算生物分析应用于基因编辑研究，实现了在全基因组水平的基因筛选和功能研究，也极大地拓展了我们对生命活动研究的深度和维度。新一代高通量深度测序又称第二代测序技术(next-generation sequencing)，不同于传统的Sanger测序技术，其可同时对数以百万到亿级的DNA分子进行序列测定，使得对不同物种、不同组织、单个细胞的基因组以及转录组的全貌分析成为可能。由于高通量测序带来的海量数据，许多特定的算法和工具被建立起来，对包括序列比对、序列拼接、差异性表达分析、共调控网络等方面开展研究，这些都可以被应用于基于CRISPR/Cas的基因组编辑。同时，许多的生物信息学方法也被广泛地应用于gRNA的设计、打靶(on-target)效应以及脱靶(off-target)效应的预测和评估等；全基因组高通量测序及后续分析可以全面有效地评估基因组编辑的脱靶效应；结合测序及相关分析也可以同时对多样本和/或多靶点的基因组编辑效率进行效果评估，节省了大量的时间和经费。最后，利用基因组编辑所产生的高通量测序大数据开展机器学习，包括深度学习的方法，可以实现对未知位点基因组编辑效果的人工智能预测，这也将大

大提高基因组编辑应用的效率。

本文从计算生物学分析在gRNA设计、CRISPR/Cas介导的基因组编辑效率和特异性检测、CRISPR/Cas介导的全基因组筛选等方面展开论述，并总结了机器学习等人工智能分析在CRISPR/Cas介导的基因组编辑中的应用，以期为读者提供较为全面的相关研究总结。

1 gRNA设计中的计算生物学分析

CRISPR/Cas系统的靶向效应主要是Cas核酸内切酶通过gRNA序列引导及其对靶向位点附近的PAM序列识别来实现的。PAM序列通常很短并直接与间隔序列相邻，其对Cas蛋白的功能发挥是必需的[20]。对于每一种Cas蛋白来说，PAM序列通常固定不变，如spCas9的PAM序列为5'-NGG-3'。Cas9参与的基因组编辑需要sgRNA的介导[3]。其中，构成sgRNA中的tracrRNA序列固定不变，其主要是通过形成茎环结构介导sgRNA与Cas结合；而构成sgRNA中的crRNA序列则被设计成可以与靶向DNA互补进从而引导sgRNA-Cas蛋白复合体识别靶向DNA位点。因此，如果设计的sgRNA可与靶向DNA位点上的序列互补配对，且靶向DNA位点附近存在合适的PAM序列，则sgRNA-Cas9复合物就可以结合到位点上进行DNA切割。但是，在实际应用过程中不同sgRNA的打靶效率有显著的差别，这与sgRNA上crRNA序列的特异性以及靶向位点附近染色体开放程度等因素有关，因此使用综合分析多种因素的预测模型可以用于辅助设计具有高效靶位点编辑效率的sgRNA[21-23]。

基因组编辑过程中存在的另一个主要问题是脱靶效应，由于gRNA在与基因组DNA序列结合时会存在一定几率的错配[24]，造成在非靶向位点的脱靶突变[25-30]。同时，虽然spCas9最优结合的PAM序列是5'-NGG-3'，然而spCas9也可以低频率地结合到5'-NAG-3'或者5'-NGA-3'上造成脱靶突变[25,27,31-32]。

综上，虽然gRNA的设计主要是通过扫描基因组序列查看靶向位点附近是否有合适的PAM序列来完成的，但是考虑到编辑的效率与脱靶效应，设计出好的gRNA也存在着一定的挑战。目前，许多算法模型和计算工具被开发出来用于gRNA序列的设计(表1)，以服务于基因组编辑的研究和应用。值得注意的是，机器学习和深度学习的方法最近也被用于gRNA的设计中，极大地促进了gRNA设计的效率[33-34]。例如，sgRNA Designer通过SVM (support vector machine)模型进行特征提取，使用逻辑回归分类器(logistic regression classifier)预测sgRNA的打靶效率[35-36]。

2 利用高通量测序对CRISPR/Cas介导的基因组编辑效果进行评估

对编辑效果，包括打靶效率(efficiency)和脱靶效应(即特异性，specificity)的检测是评估基因组编辑实验成败的关键。常用的编辑效果检验方法包括错配切割实验(mismatch cleavage assays)、Sanger测序(Sanger sequencing)检验和高通量测序检验等。相对于错配切割实验和Sanger测序检验的通量低和灵敏度低等缺点[48-49]，高通量测序方法可一次性对来自于多个编辑实验的样本进行检测，在后续的计算分析中利用不同的条形码等区分不同的编辑实验组，并分别计算获得所有编辑实验的效率(图1)。

对编辑位点的靶向高通量深度测序(targeted deep-sequencing)，可以对目的基因区域成千上万次的测序覆盖，实现对编辑效率的精准衡量[50]。与计算生物学分析中的变体识别(variant calling)过程类似，对编辑效率的数据分析包括点突变率(mutation rate)、插入(insertion)和缺失(deletion)率等的计算[51]。变体识别的高通量测序分析流程[52]通常为：(1)处理和过滤低质量的测序reads；(2)将reads比对到参考基因组；(3)利用算法识别每个位点的突变、插入和缺失情况；(4)根据计算分析模型筛选并确认变体。基因编辑效率检测也是借鉴此流程来计算突变率、插入与缺失率，目前针对基因编辑效率的计算软件和流程有CRISPR-GA、GRISPRResso、BATCH-GE、Cas-analyzer以及CRISPR-DAV和CRISPRMatch等(表2)。

靶向深度测序虽然也可用于评估预测的脱靶位点，但是由于存在预测的偏差和遗漏等，靶向深度测序很难全面检测基因组编辑的脱靶效应。基因组编辑过程中，Cas蛋白切割基因组DNA产生双链断裂(double strand break, DSB)，通过识别基因组双链断裂区域可鉴定Cas蛋白结合位置，进而

表1 gRNA设计工具

工具名	核酸酶	用户输入	打靶预测方法	脱靶预测方法	参考文献
CRISPR.mit	SpCas9	DNA序列	无	Hsu, 2013[27]	Hsu, 2013[27]
sgRNA Designer	SpCas9；SaCas9	DNA序列；转录本ID；基因ID；基因名	Rule Set 2[35]	CFD[35]	Doench, 2016[35]；Doench, 2014[36]
E-CRISP	Cas9；自定义PAM	DNA序列；基因ID；基因名	E-score[37]；Xu, 2015[38]；Doench, 2014[36]	S-score[37]	Heigwer, 2014[37]
CRISPRscan	SpCas9；LbCpf1；AsCpf1	DNA序列；基因ID；基因名	Moreno-Mateos, 2015[39]	CFD[35]；Cong, 2013[11]；Hsu, 2013[27]	Moreno-Mateos, 2015[39]
WU-CRISPR	SpCas9	DNA序列；基因ID；基因名	Wong, 2015[40]	Wong, 2015[40]	Wong, 2015[40]
SSC	SpCas9	DNA序列	Xu, 2015[38]	无	Xu, 2015[38]
CRISPOR	9种PAM	DNA序列；基因组区域	Rule Set 2[35]；Prox GC[41-42]；Xu, 2015[38]；Chari, 2015[43]；Housden, 2015[44]；Doench, 2014[36]	CFD[35]；Hsu, 2013[27]	Haeussler, 2016[45]
sgRNA Scorer	6种PAM；自定义PAM	DNA序列	Chari, 2015[43]	无	Chari, 2015[43]；Chari, 2017[32]
GuideScan	SpCas9；AsCf1；LbCpf1	基因组区域；基因名	Rule Set 2[35]	CFD[35]	Perez, 2017[46]
CASPER	自定义PAM	DNA序列	CASPER[47]	CASPER[47]	Mendoza, 2018[47]

可检测脱靶效应。多种方法包括BLESS、GUIDE-seq、HIGTS以及Digenome-seq等都是基于此原理检测基因编辑的全基因组脱靶效应。BLESS是将双链断裂的DNA末端用生物素连接，再用链酶亲和素(streptavidin)将其富集并测序和进行分析[59-61]；GUIDE-seq是在DNA双链断裂区域插入dsODN (double-stranded oligodeoxynucleotide)来标记DSB并测序和进行分析[62]；HTGTS通过将两个双链断裂区域进行连接，以易位(translocation)来标记并测序和进行分析[63]；Digenome-seq则在体外将切割位点进行保护标记并测序和进行分析[64]。这几种方法都不依赖于预测，而是直接利用高通量测序来检测Cas蛋白切割产生的DNA双链断裂来发现脱靶位点，其灵敏度高，能有效地发现低频的脱靶位点。当然，这些方法也存在各自的局限性，如BLESS只能检测固定细胞瞬时的断裂位点、GUIDE-seq需要较高的dsODNs转化效率、HIGTS依赖于两个以上的DSB位点的重组，而Digenome-seq是在体外检测DSB的存在等[65]。此外，全基因组测序(whole genome sequencing)通过扫描整个基因组来发现脱靶位点，可检测出更多的未知脱靶位点及大片段插入和缺失等[48,66]，但是其受到

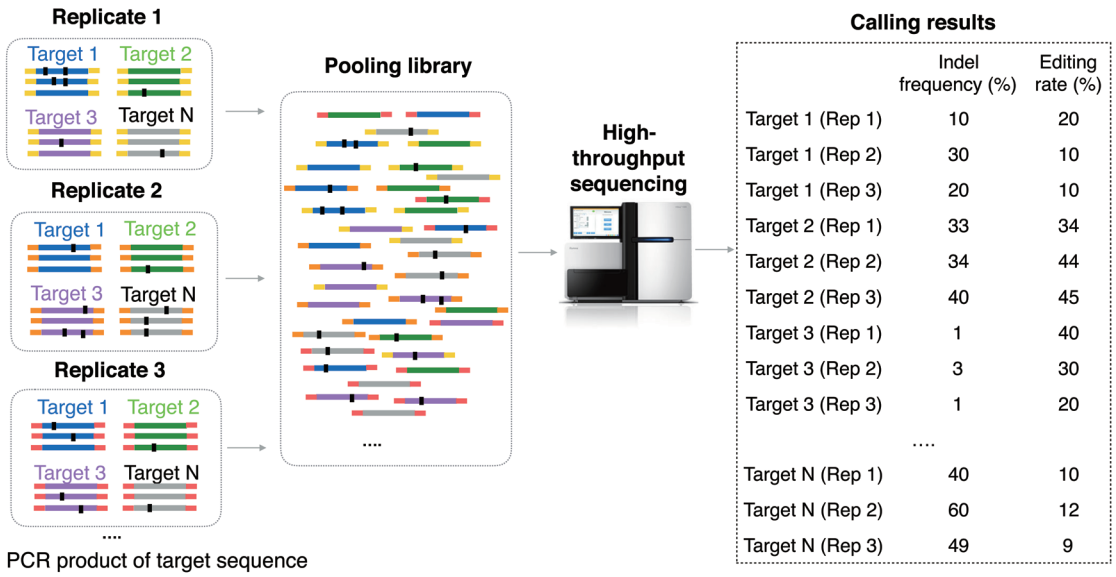

图1 高通量测序检测基因编辑效果示意图

表2 CRISPR/Cas基因编辑高通量测序数据分析工具

工具名	功能	算法	参考文献
CRISPR-GA	计算同源重组率以及插入和缺失率	BLAT alignment	Guell, 2014[54]
CRISPR-Resso	批量计算插入和缺失率、点突变率	改进GRISPR-GA算法，有效考虑sequencing error和mapping error的去除	Pinello, 2016[55]
BATCH-GE	分析靶向深度测序数据(一个样本多靶点)，计算以基因为单位的重组率	Burrows-Wheeler Aligner (BWA)	Boel, 2016[56]
Cas-analyzer	计算插入缺失率，提供比对结果信息	A javascript based algorithm	Park, 2017[57]
CRISPR-DAV	分析和可视化CRISPR/Cas基因编辑高通量数据，检测小片段或大片段的插入和缺失	BWA; Assembly Based ReAlignment	Wang, 2017[53]
CRISPRMatch	计算插入和缺失率，可视化比对结果	BWA for alignment, SAMtools, Picard for mutation calling	You, 2018[58]

测序深度和覆盖度等的影响[67]。最后，脱靶效应的数据分析也与变体识别的计算分析类似。

3 整合高通量测序与相关计算分析，利用CRISPR/Cas介导的基因组编辑实现全基因组水平基因的功能筛选

利用构建全基因组水平的gRNA文库，CRISPR/Cas介导的基因组编辑可以在全基因组水平对不同的功能基因/序列进行编辑改造，通过整合高通量测序和相关计算分析实现功能基因筛选鉴定。与传统的RNA干扰技术(RNA interference, RNAi)不同，CRISPR/Cas介导的筛选实现了在基因组DNA水平对基因表达的调控，除了基因转录序列区域，其也可作用于基因间区、启动子、增强子等非转录序列区域进行功能筛选研究，且其脱靶水平明显低于RNAi技术[68]。根据原理的不同，基于CRISPR/Cas(主要是CRISPR/Cas9)技术的全基因组筛选主要分为3类：CRISPR敲除(knock out)[69]、

CRISPR干扰(CRISPR interference, CRISPRi)[68]和CRISPR激活(CRISPR activation, CRISPRa)[70]。CRISPR敲除直接利用Cas蛋白的核酸酶活性对基因组DNA双链进行切割，然后主要通过内源的非同源末端连接(nonhomologous end joining, NHEJ)方式对切割位点进行DNA修复，由于非同源末端连接可以造成DNA切割位置的序列插入或缺失(insertion and deletion, INDEL)，进而使得基因编码区域发生移码突变丧失表达特定活性蛋白的功能[71]；CRISPR干扰技术是利用切割能力失活的Cas9(dead Cas9, dCas9)蛋白，靶向结合但是不切割DNA序列，通过占位效应阻碍转录因子的结合/移动，又或引入转录抑制因子的结合来阻碍基因的转录达到抑制基因表达的目的，避免了基因组DNA双链断裂引起的细胞毒性[68]；CRISPR激活则是通过dCas9蛋白将转录激活因子引入到转录起始位置进而激活基因的表达[72]。由于利用了dCas蛋白的占位效应，而非其核酸酶切割活性，CRISPRi与CRISPRa对基因的表达呈现可逆的调控[73]。

基于CRISPR的基因组筛选效果主要是通过整合高通量测序和相关计算分析完成。实验过程中，每一个细胞只被一个慢病毒包装的gRNA文库序列所感染，以保证每一个细胞只有一个靶基因序列被更改沉默。细胞在相应条件(如加入药物)筛选处理后，通过高通量测序和相关计算生物学分析检测gRNA丰度在筛选处理前后的变化，推断相应gRNA靶向基因功能与生物表型之间的相关性[69,74-75]。在利用计算生物学的方法分析CRISPR高通量筛选的测序结果时，排除假阳性和假阴性对寻找相关功能基因至关重要。分析技术流程主要是计算gRNA丰度在筛选处理前后的变化，gRNA在设计之初就带有了独特的序列标签(barcode)，通过将筛选处理前后测序序列比对到gRNA和特殊序列标签组成的参考序列上，可以将测序结果通过标准化来实现更为精准的计算分析[74]。现已有一系列的计算分析工具包可以对CRISPR高通量筛选进行数据分析(表3)。其中，MAGeCK是第一款针对于CRISPR高通量筛选结果进行数据处理的软件[76]，后续也开发了包括HiTSelect、CaRpools、BAGEL、HiTSeekR和ENCoRE等计算分析工具包用于开展基于CRISPR的高通量筛选。另外，用于处理RNAi高通量筛选结果的软件如RIGER，也可以用于CRISPR的高通量筛选结果处理[69,77]。

4 机器学习在基因组编辑研究中的应用

机器学习主要通过对数据的归纳和总结自动改进算法性能，获得已有数据集的规律，并最终对未知数据实现人工智能分析[84]。利用机器学习

表3 CRISPR/Cas高通量功能基因筛选数据分析工具

工具名	功能	算法	参考文献
RIGER	RNAi及CRISPRi高通量功能基因排序及筛选	Rank sgRNA score	Luo, 2008[77]
MaGeCK	全基因组敲除实验筛选功能基因及信号通路	A negative binomial model,α-RRA algorithm	Li, 2014[76]
MaGeCK-VISPR	可视化CRISPR筛选的各个分析步骤(包括质控及基因筛选)	Maximum-likelihood algorithm	Li, 2015[78]
HiTSelect	全基因组敲除实验筛选功能位点及信号通路	A random effect model	Diaz, 2015[79]
CaRpools	全基因组敲除实验筛选功能位点	Wilcox, DESeq2 or a rank based model	Winter, 2016[80]
BAGEL	全基因组敲除筛选功能基因	Bayes probability	Hart, 2016[81]
HiTSeekR	全基因组敲除筛选功能基因及功能性富集分析	T test, SSMD, Bayes probability	List, 2016[82]
ENCoRE	全基因组敲除实验筛选功能位点及信号通路	A normal z-test	Trumbach, 2017[83]

的方法预测基因编辑效率，也就是利用计算机实现对已有的CRISPR/Cas系统基因编辑效率结果的大数据集的整理和归纳，通过算法自动学习获得特征(如靶向位点的序列特征、二级结构等)与编辑效果的规律，实现输入已知特征来准确预测编辑效果的目的。自2014年以来，机器学习的方法在基因组编辑研究领域，包括gRNA设计[85-86]、脱靶效应预测[87-89]、Cas蛋白切割活性预测[89-90]、CRISPR/Cas高通量筛选功能基因[81,91]等方面得到了广泛的应用[92]。

例如，CRISPRpred方法利用FC-RES数据集(包括靶向17个基因，共5 310条gRNA的打靶效应的数据)，通过提取位置碱基的特征，基于支持向量机的方法对gRNA的打靶效应进行预测[93]。CRISTA则利用检测脱靶效应的高通量测序大数据(包括GUIDE-seq、BLESS、HTGTS)，基于随机森林以及回归模型，除了考虑核苷酸组成、gRNA与靶向序列的互补配对和热力学特征外，将DNA凸起和RNA凸起影响也考虑在内，构建了预测切割位点及脱靶效应的机器学习方法[89]。BAGEL利用gRNA表达倍数变化在必需基因和非必需基因的分布情况的数据集作为测试集，基于贝叶斯分布来训练模型鉴定必需基因[81]。

选择合适的特征进行训练对机器学习的成功构建及预测效果至关重要，对于CRISPR/Cas系统而言，通常考虑提取的特征有：(1)序列特征[88]，包括GC含量、ATCG的频率特征、连续碱基的频率特征、gRNA与靶向DNA的错配特征、切割位点碱基序列的偏好性等[86,94-95]；(2)形成二级结构的倾向性及热力学特征[95]；(3)染色质状态特征，包括DNA甲基化、核小体占位、是否处于超敏感位点等[96-97]；(4)其他相关特征，如靶向位点与PAM序列的相对位置、 细胞系特征等[98-99]。选择不同的特征进行机器学习可能会产生相差较大的结果，故应慎重选择用于训练的特征。

基于卷积神经网络(convolutional neural network, CNN)的深度学习是机器学习方法的一种，直接将多维数据输入，通过输入层、卷积层、池化层、全连结层等多个数据层的处理，得到分类结果来训练各个层的参数[100]。因为层与层之间、各个因子之间的复杂连接与神经元之间的复杂连接类似，故名为卷积神经网络[101]。深度学习实现对象识别与分类的方式是直接输入多维数据，把特征提取完全交给机器，避免了因为人为特征的选择而造成的结果偏差[102]。将卷积神经网络深度学习的方法应用于基因编辑效果的预测，则是将gRNA序列和靶序列通过字母编码法则将序列信息直接转为多维度的数字信息，直接用于模型的训练与预测，避免了人为选择序列特征对结果预测造成偏差。例如，预测Cpf1切割效率的DeepCpf1方法[96]，是对CRISPR/Cpf1全基因组gRNA编辑实验结果通过深度学习建立的。

5 总结与展望

基于CRISPR/Cas系统的基因组编辑技术，已经广泛地应用于基因敲除、基因敲入、转录调控以及碱基编辑等基础研究和临床前期应用研究，而通过高通量测序和计算生物分析方法可以有效地在诸如gRNA的设计、预测打靶和脱靶效应、提升编辑效果检测效率等方面提高基因组编辑的应用。伴随着基因组编辑技术的迅猛发展，许多研究方向都需要高通量测序和计算生物学方法的进一步参与。例如，对于新型的基因组碱基编辑系统来说，目前还没有特殊的计算分析工具用于高效gRNA的设计，或是对不同位点编辑效率以及脱靶效应的预测分析；除了针对DNA的编辑改造，利用发现的Cas13家族蛋白可以对RNA开展编辑修饰研究，相关的RNA水平编辑数据集和计算方法都存在很大的空白；如何将基于不同Cas蛋白的基因组编辑数据进行汇总和预测分析也面临着巨大的计算分析挑战；最后，虽然机器学习(包括CNN等)应用于CRISPR/Cas编辑技术的研究尚处于起步阶段，但未来在更多相关大数据集产生的基础上其预测性能还有很大的进步空间。

参 考 文 献

[1] Ishino Y, Shinagawa H, Makino K, et al. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. J Bacteriol, 1987, 169(12): 5429-5433

[2] Jansen R, Embden JD, Gaastra W, et al. Identification of genes that are associated with DNA repeats in prokary-

otes. Mol Microbiol, 2002, 43(6): 1565-1575
[3] Jinek M, Chylinski K, Fonfara I, et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science, 2012, 337(6096): 816-821
[4] Barrangou R, Fremaux C, Deveau H, et al. CRISPR provides acquired resistance against viruses in prokaryotes. Science, 2007, 315(5819): 1709-1712
[5] Deltcheva E, Chylinski K, Sharma CM, et al. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature, 2011, 471(7340): 602-607
[6] Makarova KS, Grishin NV, Shabalina SA, et al. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. Biol Direct, 2006, 1: 7
[7] Bolotin A, Quinquis B, Sorokin A, et al. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology, 2005, 151(Pt 8): 2551-2561
[8] Garneau JE, Dupuis ME, Villion M, et al. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature, 2010, 468(7320): 67-71
[9] Shmakov S, Smargon A, Scott D, et al. Diversity and evolution of class 2 CRISPR-Cas systems. Nat Rev Microbiol, 2017, 15(3): 169-182
[10] Makarova KS, Wolf YI, Alkhnbashi OS, et al. An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol, 2015, 13(11): 722-736
[11] Cong L, Ran FA, Cox D, et al. Multiplex genome engineering using CRISPR/Cas systems. Science, 2013, 339(6121): 819-823
[12] Adli M. The CRISPR tool kit for genome editing and beyond. Nat Commun, 2018, 9(1): 1911
[13] Schunder E, Rydzewski K, Grunow R, et al. First indication for a functional CRISPR/Cas system in Francisella tularensis. Int J Med Microbiol, 2013, 303(2): 51-60
[14] Zetsche B, Gootenberg JS, Abudayyeh OO, et al. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell, 2015, 163(3): 759-771
[15] Shmakov S, Abudayyeh OO, Makarova KS, et al. Discovery and functional characterization of diverse class 2 CRISPR-Cas systems. Mol Cell, 2015, 60(3): 385-397
[16] Chen B, Gilbert LA, Cimini BA, et al. Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. Cell, 2013, 155(7): 1479-1491
[17] Liu XS, Wu H, Ji X, et al. Editing DNA methylation in the mammalian genome. Cell, 2016, 167(1): 233-247 e17
[18] Komor AC, Kim YB, Packer MS, et al. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature, 2016, 533(7603): 420-424
[19] Larson MH, Gilbert LA, Wang X, et al. CRISPR interference (CRISPRi) for sequence-specific control of gene expression. Nat Protoc, 2013, 8(11): 2180-2196
[20] Hu JH, Miller SM, Geurts MH, et al. Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. Nature, 2018, 556(7699): 57-63
[21] Koike-Yusa H, Li Y, Tan EP, et al. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. Nat Biotechnol, 2014, 32(3): 267-273
[22] Shalem O, Sanjana NE, Hartenian E, et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. Science, 2014, 343(6166): 84-87
[23] Wang T, Wei JJ, Sabatini DM, et al. Genetic screens in human cells using the CRISPR-Cas9 system. Science, 2014, 343(6166): 80-84
[24] Lin Y, Cradick TJ, Brown MT, et al. CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. Nucleic Acids Res, 2014, 42(11): 7473-7485
[25] Mali P, Aach J, Stranges PB, et al. CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. Nat Biotechnol, 2013, 31(9): 833-838
[26] Cradick TJ, Fine EJ, Antico CJ, et al. CRISPR/Cas9 systems targeting beta-globin and CCR5 genes have substantial off-target activity. Nucleic Acids Res, 2013, 41(20): 9584-9592
[27] Hsu PD, Scott DA, Weinstein JA, et al. DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol, 2013, 31(9): 827-832
[28] Pattanayak V, Lin S, Guilinger JP, et al. High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. Nat Biotechnol, 2013, 31(9): 839-843
[29] Cho SW, Kim S, Kim Y, et al. Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. Genome Res, 2014, 24(1): 132-141
[30] Tsai SQ, Zheng Z, Nguyen NT, et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. Nat Biotechnol, 2015, 33(2): 187-197
[31] Jiang W, Bikard D, Cox D, et al. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nat Biotechnol, 2013, 31(3): 233-239
[32] Chari R, Yeo NC, Chavez A, et al. sgRNA scorer 2.0: A species-independent model to predict CRISPR/Cas9 activity. ACS Synth Biol, 2017, 6(5): 902-904
[33] Chuai GH, Wang QL, and Liu Q. In silico meets *in vivo*:

towards computational CRISPR-Based sgRNA design. Trends Biotechnol, 2017, 35(1): 12-21

[34] Cui Y, Xu J, Cheng M, et al. Review of CRISPR/Cas9 sgRNA design tools. Interdiscip Sci, 2018, 10(2): 455-465

[35] Doench JG, Fusi N, Sullender M, et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol, 2016, 34(2): 184-191

[36] Doench JG, Hartenian E, Graham DB, et al. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nat Biotechnol, 2014, 32(12): 1262-1267

[37] Heigwer F, Kerr G, Boutros M. E-CRISP: fast CRISPR target site identification. Nat Methods, 2014, 11(2): 122-123

[38] Xu H, Xiao T, Chen CH, et al. Sequence determinants of improved CRISPR sgRNA design. Genome Res, 2015, 25(8): 1147-1157

[39] Moreno-Mateos MA, Vejnar CE, Beaudoin JD, et al. CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting *in vivo*. Nat Methods, 2015, 12(10): 982-988

[40] Wong N, Liu W, and Wang X. WU-CRISPR: characteristics of functional guide RNAs for the CRISPR/Cas9 system. Genome Biol, 2015, 16: 218

[41] Ren X, Yang Z, Xu J, et al. Enhanced specificity and efficiency of the CRISPR/Cas9 system with optimized sgRNA parameters in *Drosophila*. Cell Rep, 2014, 9(3): 1151-1162

[42] Farboud B and Meyer BJ. Dramatic enhancement of genome editing by CRISPR/Cas9 through improved guide RNA design. Genetics, 2015, 199(4): 959-971

[43] Chari R, Mali P, Moosburner M, et al. Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. Nat Methods, 2015, 12(9): 823-826

[44] Housden BE, Valvezan AJ, Kelley C, et al. Identification of potential drug targets for tuberous sclerosis complex by synthetic screens combining CRISPR-based knockouts with RNAi. Sci Signal, 2015, 8(393): rs9

[45] Haeussler M, Schonig K, Eckert H, et al. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. Genome Biol, 2016, 17(1): 148

[46] Perez AR, Pritykin Y, Vidigal JA, et al. GuideScan software for improved single and paired CRISPR guide RNA design. Nat Biotechnol, 2017, 35(4): 347-349

[47] Mendoza BJ, Trinh CT. Enhanced guide-RNA design and targeting analysis for precise CRISPR genome editing of single and consortia of industrially relevant and non-model organisms. Bioinformatics, 2018, 34(1): 16-23

[48] Fu Y, Foden JA, Khayter C, et al. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nat Biotechnol, 2013, 31(9): 822-826

[49] Vouillot L, Thélie A, Pollet N. Comparison of T7E1 and surveyor mismatch cleavage assays to detect mutations triggered by engineered nucleases. G3: Genes, Genomes, Genetics, 2015: g3. 114.015834

[50] Bell CC, Magor GW, Gillinder KR, et al. A high-throughput screening strategy for detecting CRISPR-Cas9 induced mutations using next-generation sequencing. BMC Genomics, 2014, 15(1): 1002

[51] Pirooznia M, Kramer M, Parla J, et al. Validation and assessment of variant calling pipelines for next-generation sequencing. Human Genomics, 2014, 8(1): 14

[52] Nielsen R, Paul JS, Albrechtsen A, et al. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet, 2011, 12(6): 443-451

[53] Wang X, Tilford C, Neuhaus I, et al. CRISPR-DAV: CRISPR NGS data analysis and visualization pipeline. Bioinformatics, 2017, 33(23): 3811-3812

[54] Güell M, Yang L, Church GM. Genome editing assessment using CRISPR Genome Analyzer (CRISPR-GA). Bioinformatics, 2014, 30(20): 2968-2970

[55] Pinello L, Canver MC, Hoban MD, et al. Analyzing CRISPR genome-editing experiments with CRISPResso. Nat Biotechnol, 2016, 34(7): 695-697

[56] Boel A, Steyaert W, De Rocker N, et al. BATCH-GE: Batch analysis of Next-Generation Sequencing data for genome editing assessment. Scientific reports, 2016, 6: 30330

[57] Park J, Lim K, Kim J-S, et al. Cas-analyzer: an online tool for assessing genome editing results using NGS data. Bioinformatics, 2017, 33(2): 286-288

[58] You Q, Zhong Z, Ren Q, et al. CRISPRMatch: an automatic calculation and visualization tool for high-throughput CRISPR genome-editing data analysis. Int J Biol Sci, 2018, 14(8): 858-862

[59] Crosetto N, Mitra A, Silva MJ, et al. Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. Nat Methods, 2013, 10(4): 361-365

[60] Ran FA, Cong L, Yan WX, et al. *In vivo* genome editing using *Staphylococcus aureus* Cas9. Nature, 2015, 520(7546): 186-191

[61] Slaymaker IM, Gao L, Zetsche B, et al. Rationally engineered Cas9 nucleases with improved specificity. Science, 2016, 351(6268): 84-88

[62] Tsai SQ, Zheng Z, Nguyen NT, et al. GUIDE-seq enables

genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. Nat Biotechnol, 2015, 33(2): 187-189

[63] Frock RL, Hu J, Meyers RM, et al. Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. Nat Biotechnol, 2015, 33(2): 179-186

[64] Kim D, Bae S, Park J, et al. Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. Nat Methods, 2015, 12(3): 237-243

[65] Zischewski J, Fischer R, and Bortesi L. Detection of on-target and off-target mutations generated by CRISPR/Cas9 and other sequence-specific nucleases. Biotechnol Adv, 2017, 35(1): 95-104

[66] Cho SW, Kim S, Kim Y, et al. Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. Genome Res, 2014, 24(1): 132-141

[67] Smith C, Gore A, Yan W, et al. Whole-genome sequencing analysis reveals high specificity of CRISPR/Cas9 and TALEN-based genome editing in human iPSCs. Cell Stem Cell, 2014, 15(1): 12-13

[68] Gilbert LA, Larson MH, Morsut L, et al. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell, 2013, 154(2): 442-451

[69] Shalem O, Sanjana NE, Hartenian E, et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. Science, 2014, 343(6166): 84-87

[70] Platt RJ, Chen S, Zhou Y, et al. CRISPR-Cas9 knockin mice for genome editing and cancer modeling. Cell, 2014, 159(2): 440-455

[71] Sander JD and Joung JK. CRISPR-Cas systems for editing, regulating and targeting genomes. Nat Biotechnol, 2014, 32(4): 347-355

[72] Perez-Pinera P, Kocak DD, Vockley CM, et al. RNA-guided gene activation by CRISPR-Cas9-based transcription factors. Nat Methods, 2013, 10(10): 973-976

[73] Gilbert LA, Horlbeck MA, Adamson B, et al. Genome-scale CRISPR-mediated control of gene repression and activation. Cell, 2014, 159(3): 647-661

[74] Koike-Yusa H, Li Y, Tan E-P, et al. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. Nat Biotechnol, 2014, 32(3): 267-273

[75] Sanjana NE, Shalem O, and Zhang F. Improved vectors and genome-wide libraries for CRISPR screening. Nat Methods, 2014, 11(8): 783-784

[76] Li W, Xu H, Xiao T, et al. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. Genome Biol, 2014, 15(12): 554

[77] Luo B, Cheung HW, Subramanian A, et al. Highly parallel identification of essential genes in cancer cells. Proc Natl Acad Sci USA, 2008, 105(51): 20380-20385

[78] Li W, Köster J, Xu H, et al. Quality control, modeling, and visualization of CRISPR screens with MAGeCK-VISPR. Genome Biol, 2015, 16(1): 281

[79] Ramalho-Santos M, Diaz A, Song J, et al. HiTSelect: A comprehensive tool for high-complexity-pooled screen analysis. Nucleic Acids Res. 2015, 43(3): e16

[80] Winter J, Breinig M, Heigwer F, et al. caRpools: an R package for exploratory data analysis and documentation of pooled CRISPR/Cas9 screens. Bioinformatics, 2015, 32(4): 632-634

[81] Hart T and Moffat J. BAGEL: a computational framework for identifying essential genes from pooled library screens. BMC Bioinformatics, 2016, 17(1): 164

[82] List M, Schmidt S, Christiansen H, et al. Comprehensive analysis of high-throughput screens with HiTSeekR. Nucleic Acids Res, 2016, 44(14): 6639-6648

[83] Trümbach D, Pfeiffer S, Poppe M, et al. ENCoRE: an efficient software for CRISPR screens identifies new players in extrinsic apoptosis. BMC Genomics, 2017, 18(1): 905

[84] Robert C. Machine learning, a probabilistic perspective. CHANCE, 2014, 27(2): 62-63

[85] Kuan PF, Powers S, He S, et al. A systematic evaluation of nucleotide properties for CRISPR sgRNA design. BMC Bioinformatics, 2017, 18(1): 297

[86] Najm FJ, Strand C, Donovan KF, et al. Orthologous CRISPR-Cas9 enzymes for combinatorial genetic screens. Nat Biotechnol, 2018, 36(2): 179-189

[87] Listgarten J, Weinstein M, Kleinstiver BP, et al. Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. Nat Biomed Eng, 2018, 2(1): 38-47

[88] Wong N, Liu W, and Wang X. WU-CRISPR: characteristics of functional guide RNAs for the CRISPR/Cas9 system. Genome Biol, 2015, 16(1): 218

[89] Abadi S, Yan WX, Amar D, et al. A machine learning approach for predicting CRISPR-Cas9 cleavage efficiencies and patterns underlying its mechanism of action. PLoS Comput Biol, 2017, 13(10): e1005807

[90] Erard N, Knott SR, and Hannon GJ. A CRISPR resource for individual, combinatorial, or multiplexed gene knockout. Mol Cell, 2017, 67(6): 1080

[91] Horlbeck MA, Gilbert LA, Villalta JE, et al. Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. Elife, 2016, 5: e19760

[92] 张桂珊, 杨勇, 张灵敏等. 机器学习方法在CRISPR/Cas9系统中的应用. 遗传, 2018, 40(9): 704-723

[93] Rahman MK and Rahman MS. CRISPRpred: A flexible and efficient tool for sgRNAs on-target activity predic-

tion in CRISPR/Cas9 systems. PLoS One, 2017, 12(8): e0181943

[94] Kuscu C, Arslan S, Singh R, et al. Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. Nat Biotechnol, 2014, 32(7): 677-683

[95] Doench JG, Fusi N, Sullender M, et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol, 2016, 34(2): 184-191

[96] Kim HK, Min S, Song M, et al. Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity. Nat Biotechnol, 2018, 36(3): 239-241

[97] Wu X, Scott DA, Kriz AJ, et al. Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. Nat Biotechnol, 2014, 32(7): 670-676

[98] Chari R, Yeo NC, Chavez A, et al. sgRNA Scorer 2.0: a species-independent model to predict CRISPR/Cas9 activity. ACS Synth Biol, 2017, 6(5): 902-904

[99] Prykhozhij SV, Rajan V, Gaston D, et al. CRISPR multitargeter: a web tool to find common and unique CRISPR single guide RNA targets in a set of similar sequences. PLoS One, 2015, 10(3): e0119372

[100] Bouvrie J. Notes on convolutional neural networks[R]. Massachusetts: Center for Biological and Computational Learning, 2006: 38-44

[101] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436-444

[102] Schmidhuber J. Deep learning in neural networks: An overview. Neural Netw, 2015, 61: 85-117