本科毕业论文

论文题目：RNA 编辑工具的安全性评估

姓　名：王奕蘅　　学　号：20307110376

院　系：生命科学学院

专　业：生物技术

指导教师：伊成器　　职　称：教授

单　位：北京大学生命科学学院

完成日期：　　2024 年 5 月 18 日

RNA 编辑工具的安全性评估

完成人

王奕蘅

指导小组成员

伊成器 教授

目 录

摘 要

基因编辑技术在医疗和科研领域具有广阔的应用前景,但编辑工具的脱靶效应是一个亟待解决的重大问题。本研究针对 RNA 单碱基 A-to-I 编辑工具 REPAIRv1 和 REPAIRv2，比较评估它们在人体细胞中的脱靶情况，并分析脱靶位点的序列特征、偏好性等。使用 GATK、VarScan2、BCFtools 等主流变异检测工具以及自主开发的 de novo 流程，从全转录组 RNA-seq 测序数据中鉴定出编辑工具可能导致的脱靶位点。通过 IGV 工具核实真假阳性，对不同工具和不同过滤条件下的结果进行比较分析。对脱靶位点的序列特征、基因组注释等进行统计和分析。研究结果发现 de novo 流程在识别真实脱靶位点方面表现优异。同时，REPAIR 编辑工具的脱靶主要是 Cas 非依赖型，且存在一定的序列偏好性，REPAIRv2 相比 REPAIRv1 具有更低的脱靶率。大多数脱靶未对重要基因的表达造成明显影响。本研究系统评估了 RNA A-to-I 单碱基编辑工具 REPAIR 的编辑特异性，为优化编辑工具设计和降低脱靶风险提供了理论依据。同时，研究方法可推广应用于其他编辑工具的脱靶评估。

关键词：基因编辑，RNA 编辑，脱靶效应，变异检测

Abstract

Gene editing technologies have broad application prospects in medicine and scientific research, but the off-target effects of editing tools are a major issue that needs to be addressed. This study compares and evaluates the off-target situations of several RNA single-base A-to-I editing tools such as REPAIRv1 and REPAIRv2 in human cells, and analyzes the sequence characteristics and preferences of the off-target sites. Mainstream variant calling tools such as GATK, VarScan2, BCFtools, and developed de novo pipeline are used to identify potential single nucleotide variants caused by the editing tools from whole transcriptome RNA-seq data. The true and false positives are verified using the IGV tool, and the results under different tools and filtering conditions are compared and analyzed. Statistical analyses are performed on the sequence characteristics, genomic annotations, etc. of the off-target sites. The de novo pipeline is superior at identifying true off-target sites. The off-target effects of the REPAIR editing tools are mainly Cas-independent and exhibit certain sequence preferences. REPAIRv2 has a lower off-target rate compared to REPAIRv1. Most off-target events do not significantly affect the expression of important genes. This study systematically evaluates the off-target situations of the REPAIR editing tools, providing theoretical basis for optimizing tool design and reducing off-target risks. The research methods can be extended to evaluate off-target effects of other RNA editing tools.

Key words: Gene editing, RNA editing, Off-target effects, Variant calling

一、前 言

1.1 基因编辑技术

基因编辑技术的兴起可以追溯到 20 余年前，研究者发现利用人工核酸酶剪切引入 DNA 双链断裂可以大幅提高靶向载体的同源重组效率，此后基因靶向修饰方法逐渐被广泛应用于定向修饰和基因组改造工程中。早期基因编辑技术包括锌指核酸酶（ZFNs）和 TALE 核酸酶（TALENs），二者根据保守的核心结构域通过蛋白质工程化设计来靶向特定 DNA 基序的位置实现修饰和编辑[1]。2013 年，George Church，张锋和 Jennifer A. Doudna 等人开发的 CRISPR/Cas 系统推动基因编辑技术广泛应用，该系统最早由日本学者发现源自细菌的适应性免疫过程。其中，II 型和 V 型 CRISPR 系统利用 RNA 引导的核酸酶系统来切割靠近 PAM 序列并与 gRNA 互补的 DNA 靶点。目前，应用最广泛的 Cas 蛋白源自于化脓性链球菌的 Cas9（SpCas9）。Cas9 酶可以引导导向 RNA（gRNA）靶向到目标位点上，从而通过非同源末端连接（NHEJ）修复或通过模板依赖的同源性（HDR）修复机制进行精确基因编辑来驱动靶向位点的修饰或载体的替换[2]。便捷和高效的特性使得 CRISPR/Cas9 系统迅速成为基因编辑领域研究的重要工具。基因编辑不仅可以影响基因表达，也可以实现精确的基因插入、修复和重组，为单基因遗传病的基因治疗提供了希望。

近年来，许多研究小组和公司基于 CRISPR/Cas 系统研究疾病潜在机制的研究，同时也为临床试验和预后提供了重要数据和参考[3]。例如，利用 CRISPR/Cas9 系统编辑和 DNA 碱基编辑工具纠正导致单基因遗传性疾病（镰状细胞病、地中海贫血、肉芽肿凝血因子缺乏症等）发生的靶点突变矫正基因缺陷实现基因治疗[3]；以及通过基因编辑技术优化 T 细胞的抗肿瘤活性，如敲低 T 细胞受体基因的表达以治疗植物抗宿主疾病；敲低 PD-1 基因表达或敲除整段基因提高巨噬细胞杀伤力，或整合 Chimeric Antigen Receptor（CAR）基因赋予免疫细胞识别癌细胞的能力，从而促进肿瘤免疫细胞疗法的有效治疗[1]。此外，在基础研究中，CRISPR/Cas9 可高效编辑动物胚胎或体细胞基因，用于建立人类疾病模型或在动物体内纠正致病基因[4]。在农业育种领域，基因编辑技术可

赋予农作物抗虫、抗旱、增产等优良性状[4]。

然而，基因组双链断裂可能会导致有害的染色体易位、倒位和副产物增加问题[5,6]，同时 CRISPR/Cas 系统介导的基因编辑在人体中还面临许多问题，包括基于同源重组修复的效率较低、自身免疫反应和脱靶等[7]。此外，目前单碱基突变遗传性疾病在人类致病性突变中占比较高[8]，而新型单碱基编辑技术（single-base editing）不仅能避免 DNA 双链断裂带来的额外风险也更适用于纠正和治疗单碱基突变造成的遗传性疾病。

DNA 单碱基碱基编辑器由催化失活的 Cas9 蛋白和脱氨酶模块[9] 组成，主要包括胞嘧啶单碱基编辑器 CBE 和腺嘌呤单碱基编辑器 ABE，其中 CBE 工具的脱氨酶由 APOBEC 家族构成，ABE 工具的脱氨酶由 TadA 构成。引入调控 DNA 修复的蛋白质（如 UGI），可实现在没有供体模板的情况下生成位点特异性单碱基转换[8]。CBE 和 ABE 系统分别催化 C·G 到 T·A 和 A·T 到 G·C，均可用于 DNA 水平的单碱基编辑。除了 DNA 编辑技术外，RNA 编辑技术同样是重要的基因编辑技术。相比于 DNA 编辑，RNA 编辑编辑靶向 RNA 不会影响靶向基因组基因的序列，具有可逆、快速的特点，不会导致基因组发生永久性变化，使得它在实际应用中较 DNA 编辑更具有优势[10] 。尤其是在治疗转录本表达成熟 RNA 的可变剪接过程发生异常相关疾病时，DNA 编辑由于对转录过程的控制有限，因此 RNA 编辑更适合用于进行 RNA 转录过程相关疾病的治疗。除此之外，大量、多种类型的 RNA 表观修饰广泛存在于不同类型的 RNA 上，因此 RNA 编辑除了可以对靶位点的碱基进行转换外还可以实现表观修饰，调控细胞的生长、互作等表观转录组过程。

2017 年，张峰团队开发了一种 RNA 单碱基 A-to-I 编辑技术[12]，该技术利用一种催化失活的 RNA 导向蛋白 Cas13b 蛋白（dPspCas13b）将腺苷脱氨酶（ADAR）定位到目标位点实现 A-to-I 的编辑。作者在 ADAR2 脱氨酶中引入 E488Q 突变，获得 $ADAR2_{DD}$，将该突变体与 dPspCas13b 融合以生成更加特异靶向目标 RNA 位点的 RNA 编辑工具，以上系统称为可编程的 A-to-I 转换的 RNA 编辑技术 REPAIR（RNA Editing for Programmable A-to-I Replacement）。在使用 REPAIRv1 靶向编辑特定位点结果中，作者发现针对内源转录本目标位点的编辑效率较低，并且同时观察到较高的邻近脱靶现象。由于 ADAR 蛋白在

REPAIR 系统中的外源过表达方法，作者猜测靶向转录本和全基因组的 RNA 脱靶编辑由 ADAR 脱氨酶造成。为了提高 REPAIRv1 的特异性，张峰团队在 ADAR 蛋白中引入了突变，提高 dCas13b-ADAR2$_{DD}$ 融合的特异性，得到了 REPAIRv2。尽管在特异性更高的 REPAIRv2 系统中仍然存在少量的邻近编辑，但相较于 REPAIRv1 系统在维持靶向编辑效率的基础上实现了特异性的显著特高。因此高特异性版本的 REPAIRv2 系统在哺乳动物系统中的 RNA 编辑具有巨大潜力和广泛的应用方向[8]。

1.2 单碱基编辑工具的特异性

随着碱基编辑工具的开发，研究发现单碱基编辑系统可能会在 DNA 和 RNA 上引入额外的碱基转换[11]，即脱靶编辑。此外，当这些编辑器在人体内使用时，携带碱基编辑工具的递送载体有时可能会错误定位到非靶向的器官、组织或特定类型细胞内，从而导致脱靶效应[12]。为了提高基因组编辑工具在临床中应用的安全性，脱靶评估是安全性评估中至关重要的部分。

造成脱靶的原因主要分为两个方面。Cas 核酸酶可以容忍 DNA 底物上的一些错配或插入缺失，从而导致对非目标位点的核酸序列产生编辑[13–15]。这种错配主要是在引导序列负责结合 DNA 的靶向链并赋予靶向的特异性的过程中产生的。这些脱靶位点与 gRNA 具有序列相似性，并且依赖于核糖核蛋白复合物，因此被称为 Cas 依赖型脱靶。此外，脱氨酶的过度表达或结构域特性可能会导致随机的脱氨作用发生，该脱靶编辑的底物具有与编辑器靶向编辑底物相似的二级结构和序列特征[16,17]，造成碱基编辑器在非目标位点的编辑。因此这类由脱氨酶产生的脱靶称为 Cas 非依赖型脱靶。对于 Cas 依赖型脱靶，先后有团队开发 Cas-OFFinder、CRISPOR、GUIDE-seq、CIRCLE-seq 和 CHANGE-seq 等检测方法检测全基因组的脱靶现象[15,18–21]。对于 Cas 非依赖性的脱靶，也先后开发出了 GOTI、正交 R 环测定以及 Detect-seq 等检测技术鉴定全基因组脱靶[22–25]。

除了以上针对 DNA 水平的脱靶的检测技术，针对 RNA 水平的脱靶问题，广泛、快速且成熟的 RNA-seq 测序是一种目前检测 RNA 编辑工具在全转录组水

平脱靶的主要方法。因为全转录组在全基因组的占比较低（~3%），利用 RNA-seq 可以实现较深通量的测序，实现全转录组脱靶检测的灵敏度和可靠性。

1.3 基于 RNA-seq 的脱靶检测方法

中心法则中 RNA 作为 DNA 和蛋白质之间的关键中间过程，对调控上游基因表达和影响下有翻译都是重要的存在，转录过程和基因表达的研究是分子生物学中的重要方向。目前开发的大多数变异检测方法都是基于全基因组或外显子组测序数据，这些方法中大多数都是基于原始序列读数（reads）与参考基因组的比对的高通量测序获得[26]。这种方法有一些缺点，包括测序基因组组装的不完整性[27]、测序错误、单核苷酸多态性对读数映射的干扰、个体基因组的结构变异[28]。因此，在应用于结构多样、读长较短的 RNA 测序时具有一定的局限性，此前已有多项研究比较了针对短读长测序数据的变异检测工具[29–36]，其中广泛使用且检测性能表现较好的分析工是 GATK（The Genome Analysis Toolkit）系列[37]。此外，VarScan[38,39]和 BCFtools[40]也被广泛用于检测 SNP（Single Nucleotide Polymorphism）、插入（Insertion）和缺失变异（Delete）检测。

为了更安全的使用 REPAIR 编辑工具，针对 REPAIR 编辑工具的安全性评估是目前亟待考察和解决的问题。在除了上文中提到的 GATK、VarScan2 和 BCFtools，我们还将使用到一套组内研发中的分析流程——de novo。de novo 流程基于 RNA-seq 分析流程，并在此基础上加入了数步可调节的过滤步骤，使得用户可以根据结果调整参数以获得精度较高的真是结果。本研究将使用上述四种工具/流程对 REPAIR 工具的脱靶效应和也异性进行分析和评估。

通过对 REPAIR 系列 RNA 单碱基编辑工具进行深入的脱靶效应评估，本研究期望可以全面认识这一技术的优缺点，为未来将其安全高效地应用于临床诊疗和基础研究奠定基础，从而推动单碱基遗传性疾病的治疗和生命科学的发展。同时，本研究围绕 RNA 编辑工具的特异性分析流程也将为其他基因编辑技术的脱靶风险检测提供参考，促进精准高效的基因编辑技术不断开发和应用，实现临床和预后的有效治疗。

二、材料与方法

2.1 数据来源

构建质粒 REPAIRv1_TAG、REPAIRv1_NT、REPAIRv2_TAG 和 REPAIRv2_NT，转染至 HEK293T 细胞中，并将样本在 Illumina Hiseq 平台上进行测序，得到 RNA-seq 数据。对原始 fastq 文件进行去接头、质控、比对等处理，得到 BAM 文件。样本中是否有 TAG（TAG/NT）对于本课题的结果不产生影响，只作为一个平行实验组。因此，本课题最终用到的数据源为 Evector、REPAIRv1、REPAIRv1_NT、REPAIRv2、REPAIRv2_NT 五个 BAM 文件。

2.2 服务器环境和工具资源

在实验室服务器集群上进行，Linux CentOS 7，主要用到以下工具包：

snakemake（7.32.3），GATK（4.0.5.1），VarScan2（2.3.9），BCFtools（1.14），SAMtools（1.18），Python（2.7.18、3.10.10），R（4.1.3），BEDtools（2.30.0），Intervene（0.6.5），featureCounts（2.0.1），HOMER（4.11）。

GATK 使用 QD < 2.0，FS > 60.0，MQ < 30.0，MQRankSum < -12.5，ReadPosRankSum < -8.0，DP < 20 的标准进行粗过滤；

VarScan2 使用 default 参数；BCFtools 使用 Multiallelic calling（-m）。

2.3 本地环境和资源

Python（3.10.11），R（4.3.1），IGV（2.17.3）

Python 工具包均用 pip install 下载，R 工具包均在清华镜像下用 install.packages()下载，IGV 由官网下载安装。

在本地操作的流程有，表达量分析、找点统计结果分析绘图以及最后的错义突变分析。

2.4 变异检测的准确性评估

为了评估各个工具变异检测的性能，本文定义了概念以示区分：

真阳性（TP）：由变异检测工具识别出来，并且经过人工 IGV 截图检验为

真脱靶位点；

真阴性（TN）：不是脱靶位点且工具未识别出；

假阳性（FP）：由变异检测工具识别出来，但是经过人工 IGV 截图检验不是脱靶位点；

假阴性（FN）：工具未识别出的脱靶位点。

三、研究结果

3.1 不同基因组变异检测工具/方法的比较和变异位点识别

3.1.1 工具/方法的选择和安装

先前开题的调研环节，选择了目前主流的基因组变异检测工具作为本文的主要研究对象，包括：Genome Analysis Toolkit（GATK），VarScan2，BCFtools和组内自行研发的变异检测流程（以下称为 de novo 流程）。基于编辑器造成脱靶的化学原理，在实际安装和应用的过程中，将着重考察以上工具/方法对于单核苷酸多态性（SNV）的识别和检测能力，而不考察其对于基因组上插入或缺失（Indel）的调用能力。

首先需要辨析两种基因突变类型，即生殖细胞突变（germline）和体细胞突变（somatic），大多数变异检测工具对于这两种变异类型的分析流程（pipeline）或者应用的程序包存在差异，因此为了结果的准确性，需要在一开始明确本研究适用的流程或工具。本实验数据来源于在人类胚胎肾细胞（HEK293T）中进行的前期实验，同时由于编辑造成的突变具有可传递性，因此属于 germline 类型突变，这将指导后续对于变异检测工具的具体分析流程的准确选择。

GATK 作为一种已经被广泛应用于高通量测序数据分析的工具集，具有丰富的工具包和相对成熟的分析流程[41]。依据本研究的数据类型（RNA-seq 数据）、实验目的（SNV 检测）和基于此的 GATK 官方建议[42]，选择使用 SplitNCigarReads — BaseRecalibrator — HaplotypeCaller — VariantFiltration — SelectVariants 的分析流程。HaplotypeCaller 是 GATK 工具中一款通过单倍型的局部重组调用 germline 突变类型的单核苷酸突变位点的工具包[43]。SplitNCigarReads 用于将 RNA 测序读长（reads）分割成更易比对和分析的长度[44]，可供 HaplotypeCaller 直接调用。BaseRecalibrator 用于对碱基质量分数的重新校正，在变异检测的过程中，碱基质量分数在权衡支持或反对可能的变异等位基因的证据方面发挥着重要作用，因此纠正数据中观察到的任何系统偏差非常重要[45]。VariantFiltration 是可以通过人工设定的过滤标准对初识别的 SNP 位点进行筛选的硬过滤（hard-filtering）工具包[46]，并且相对于用机器学习的方法

进行变异位点质量得分校正（VQSR）的过滤方式，更适合目前训练数据较少的 RNA 测序结果。而 SelectVariants 工具包则用于从变异检测结果（VCF 格式文件）中选择特定的变异记录[47]。此前，已有研究证明了运用 HaplotypeCaller 和 VariantFiltration 组合的 pipeline，进行 RNA-seq 数据 SNP 位点提取和过滤的可行性[10]。运行 conda install -c bioconda gatk4 命令，在服务器上安装 GATK 工具集。

VarScan 工具在 2009 发表之后又在 2013 年推出了 VarScan2，并且底层的启发式算法也使得 VarScan2 对于非常规情况的位点识别表现优于其他基于贝叶斯算法的工具[38,39]。本研究将依据官方文件中提供的关于 germline 细胞系突变检测的推荐流程。运行 wget 命令从 GitHub 下载 JAVA 程序，运行 java -jar VarScan.v2.3.9.jar 命令，调用 VarScan2。

SAMtools 最初发布于 2009 年[48]。该工具集不仅包括用于转换和操作 SAM 和 BAM 文件的实用工具包，还包括一个突变检测的工具包，即 BCFtools（2010 年，版本 0.1.9）[49]。BCFtools 的处理对象通常是二进制 VCF（Variant Call Format）文件。它提供了一系列命令行工具，用于对 VCF 文件进行各种操作，包括过滤、统计、转换格式等。

de novo 流程是由课题组自行研发设计的脱靶位点检测流程，由前期的数据处理、格式转换、剔除对照组存在的已知突变和最终的突变位点检测组成。其中，突变位点检测的原理是通过多次调整突变 reads 数（mut count）、覆盖的 reads 数（cover count）、突变率（mut ratio）、最大覆盖 reads 数（max cover）和 q 值（q value）等组合条件的数值进行过滤，最终得到趋于稳定的脱靶结果，即为编辑工具在全转录组造成的可能脱靶位点。以上用于过滤的条件，统称为 cutoff。

3.1.2 SNP 位点调用和粗过滤

依托于实验室服务器已有的 snakemake 工具，此部分大多数脚本以 snakemake 文件的形式进行编写。首先，依据分析工具分类，分为 GATK，VarScan，BCFtools，de novo 五组，分别输入经过预处理的 Evector、REPAIRv1、REPAIRv1_NT、REPAIRv2、REPAIRv2_NT 五个样本的 BAM 文

件，提取其中的变异位点。为了更直观全面的比较各种工具的脱靶检测能力，将在 3.1.3 中使用统一的标准（cutoff）对假阳性位点进行过滤，所以在此不设置严格的过滤标准，沿用工具的默认设置或者官网推荐的标准。

3.1.3 统一过滤标准

为了更好的比较工具之间的表现及工具间平衡比较，本研究对于不同变异检测工具得到的粗过滤 SNP 位点结果使用相同的标准过滤。选取 3.1.1 中提及的 mut count、cover count 和 q 值作为过滤控制条件，其中 q 值固定为 0.001。mut count 和 cover count 作为产生主要影响的过滤控制条件，分别从宽松到严格设置了三个不同的梯度，经组合得到 M1C3、M3C5、M6C30 这三个 cutoff。将提取 mut count 和 cover count 同时满足大于等于 cutoff 的突变位点。

同时，在预实验的过程中，发现以 Evector 样本作为对照，并在过滤之前先去除其他四个实验组中与对照组相同的突变位点，会对整体结果造成较大影响。这种影响表现为不去除对照组位点，在过滤后得到的点数量较多且大部分为假阳性点，而去除对照组位点后再过滤后得到的点数几乎为 0。这两者之间的巨大差异促使正式实验中将实验组分为了去除对照组前过滤（without fix）和去除对照组后过滤（with fix）这两大类进行比较，以验证去除对照组这一步骤对于降低假阳性的必要性。由样本（Evector、REPAIRv1、REPAIRv1_NT、REPAIRv2、REPAIRv2_NT）、cutoff（M1C3、M3C5、M6C30）、去除对照组前后（without fix、with fix）和分析工具（GATK、BCFtools、VarScan2、de novo）这四个变量组合，最终得到 120 个变异检测组合结果（点集）。

3.1.4 结果比较

3.1.4.1 去除对照组前后结果对比

由于 REPAIRv1 和 REPAIRv2 都是针对腺嘌呤的 RNA 单碱基编辑工具，因此 3.1.4 部分对于脱靶位点的统计分析也集中于 A > I 位点，其他突变类型的脱靶位点不作分析，以下不再赘述。

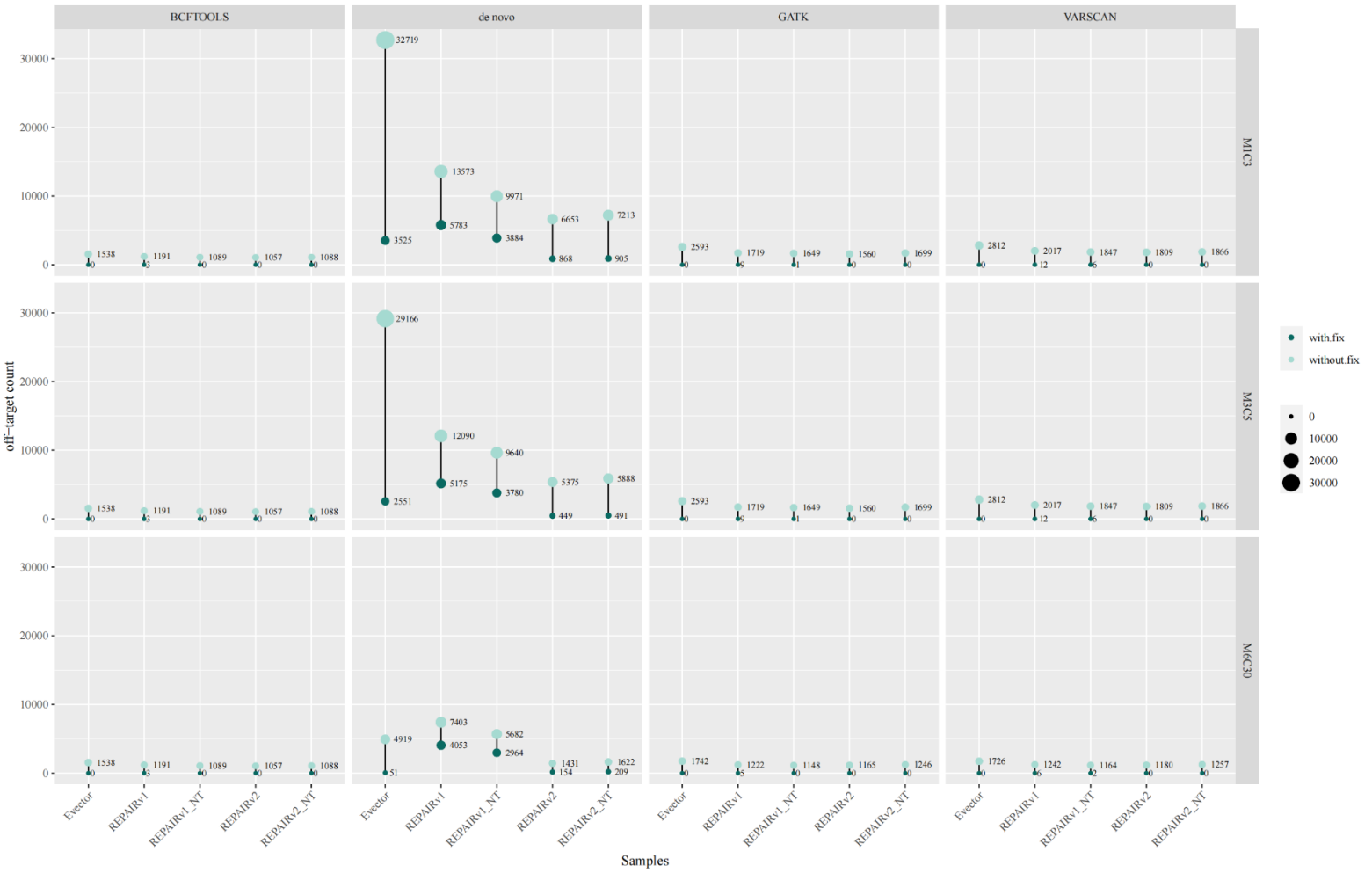


图 1. 去除对照组样本前后脱靶位点数量统计

浅蓝色代表去除对照组（Evector）前经过过滤留下的点数，深蓝色表示去除对照组后经过滤留下的点数，圆点大小与数量正相关，每个圆点右侧的数值为改点的准确计数；横向以分析工具为分组，纵向以 cutoff 为分组；纵坐标为脱靶位点计数，横坐标为样本来源

以上结果可以看出，无论是否加入去除对照组这一步，在相同的过滤条件下 de novo 流程找出的 SNP 位点都显著多于其他三种已发表工具；并且对于所有工具，在全部 3 个 cutoff 的分类下，去除对照组这一步对于最终得到的点数都产生了较大的影响，可以导致数量上 50%-100%的缺失。同时值得注意的是，GATK、VarScan2 和 BCFtools 分别在 M1C3 和 M3C5 两个 cutoff 下得到的结果是一致的。然而，数量上的一致并不能完全说明在这两个 cutoff 下找到的点集就是相同的；同时数量上的巨大差异也不能完全说明去除对照组这一步的作用，目前为止并不能排除剩下的点集中依旧存在大量假点、去除的点集中存

在大量真点的情况。因此为了增加严谨性，需要进一步验证点集中的 SNP 位点的真假。

以上段落中提及的“真点”在这里定义为由编辑工具造成的脱靶位点并且被突变调用工具成功找到，“假点”定义为除编辑工具造成的脱靶之外的其他 SNP 并且被突变调用工具成功找到。由定义可知，如果大量“真点”被排除在最终点集之外，则会增加假阴性；如果大量“假点”被包含在最终点集之内，则会增加假阳性。为了验证 SNP 位点的“真假”，首先编写 Python 脚本从点集中随机选取 100 个 SNP 位点（不满 100 的则选取全部）构成截图脚本，随后导入 IGV（Integrative Genomics Viewer）进行自动截图，通过判断 IGV 截图得到该 SNP 的“真假”性（图 2），并统计结果。需要注意的是，从图 1 的结果可以发现在点的数量上 M1C3 和 M3C5 的差距较小，而 M3C5 和 M6C30 的差距较大，因而在此我们认为在 M6C30 的 cutoff 下所得到的结果，真点的含量会更高，故对 M6C30 的结果进行如上描述的 IGV 验证。

如图 3 所展示的结果，本研究得出以下两个结论：

① de novo 流程在查找识别由脱靶造成的 SNP 位点上，相较于其他工具有显著优势，具有较低的假阳性；

② 去除对照组这一步骤对于降低结果的假阳性有较为明显的作用，对于整个分析流程具有必要性。

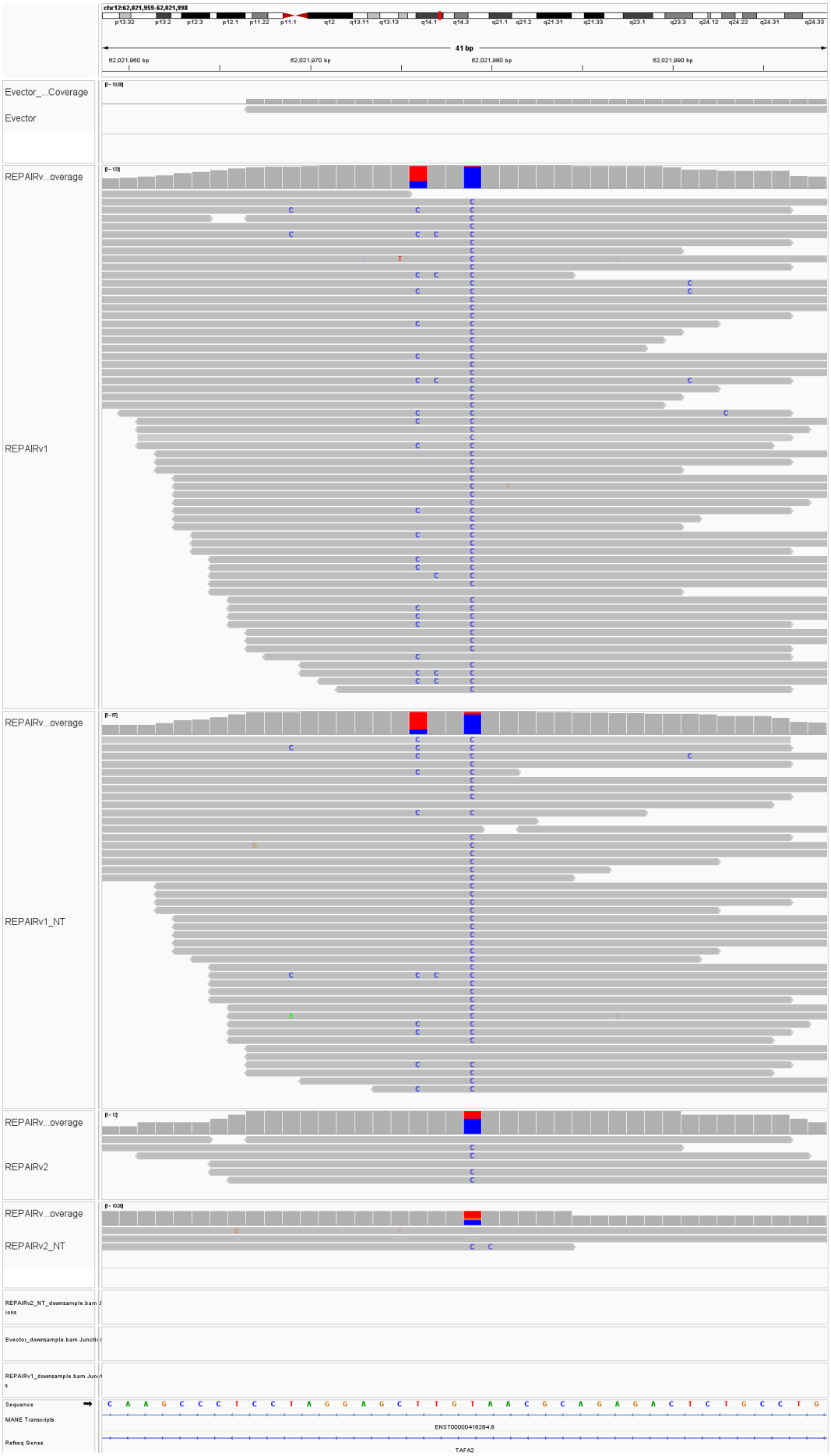

chr12:62,021,959-62,021,998
41 bp
62,021,960 bp
62,021,970 bp
62,021,980 bp
62,021,990 bp
Evector_...Coverage
Evector
REPAIRv...overage
REPAIRv1
REPAIRv...overage
REPAIRv1_NT
REPAIRv...overage
REPAIRv2
REPAIRv...overage
REPAIRv2_NT
REPAIRv2_NT_downsample.bam J
ions
Evector_downsample.bam Junctio
REPAIRv1_downsample.bam Junct
s
Sequence
MANE Transcripts
ENST00000416284.8
Refseq Genes
TAFA2

图 2. IGV 截图判定 SNP 真假性标准示例

如图所示为人 12 号染色体上的第 62021979 号碱基（图片正中大部分都是 C 碱基的位点），因为满足①突变率大于 50%②所有测到的样本都有大量突变，从而被认定为是细胞本身的 SNV；左侧 62021976 号碱基则在满足突变率小于 50%的同时，是 REPAIRv1 系列样本的独特位点，因此判定为真点。编写的脚本针对的是每张图正中间的碱基位置，所以虽然这张图上有一个真点一个假点，但是由突变识别工具找到的是这个假点而非真点，因此在统计结果时这张图将被记为“假点”

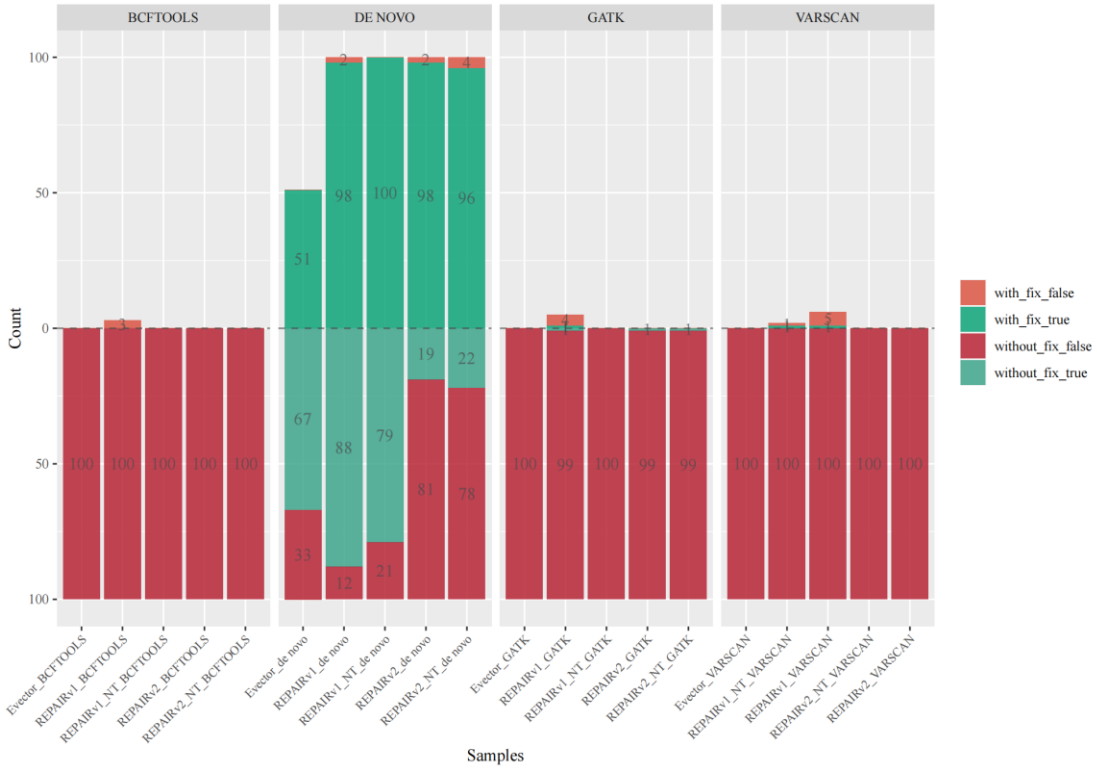


图 3. M6C30 条件下去除对照组前后真假位点统计

如图所示，红色系均代表经 IGV 截图验证后的假点，绿色系均代表经 IGV 截图验证后的真点；y = 0 的上方代表去除对照组后的，下方代表去除对照组前的；柱子上的数字代表这一类的位点计数；可以发现除 de novo 流程之外的工具，在去除对照组之后只留下很少的位点；所有工具在去除对照组之后，假阳性率都有所下降；de novo 流程在去除对照组前后的假阳性率均低于其他三个工具

3.1.4.2 突变位点散点图

张锋实验室开发的 REPAIRv1 和 REPAIRv2 的编辑效率分别约为 20%-30%且 REPAIRv2 的特异性相较于 REPAIRv1 提升了 919 倍[50]，基于此推断，在 REPAIR 系列编辑工具较为安全的情况下，它们的脱靶效率应当低于编辑效率即 20%，并且 REPAIRv2 的脱靶率应当低于 REPAIRv1。

进一步延伸，本文认定在找到的位点中，突变率在 20%左右和以下的 SNP 位点是真点的概率更大，而突变率越大越有可能是实验细胞本身自带的 SNV。

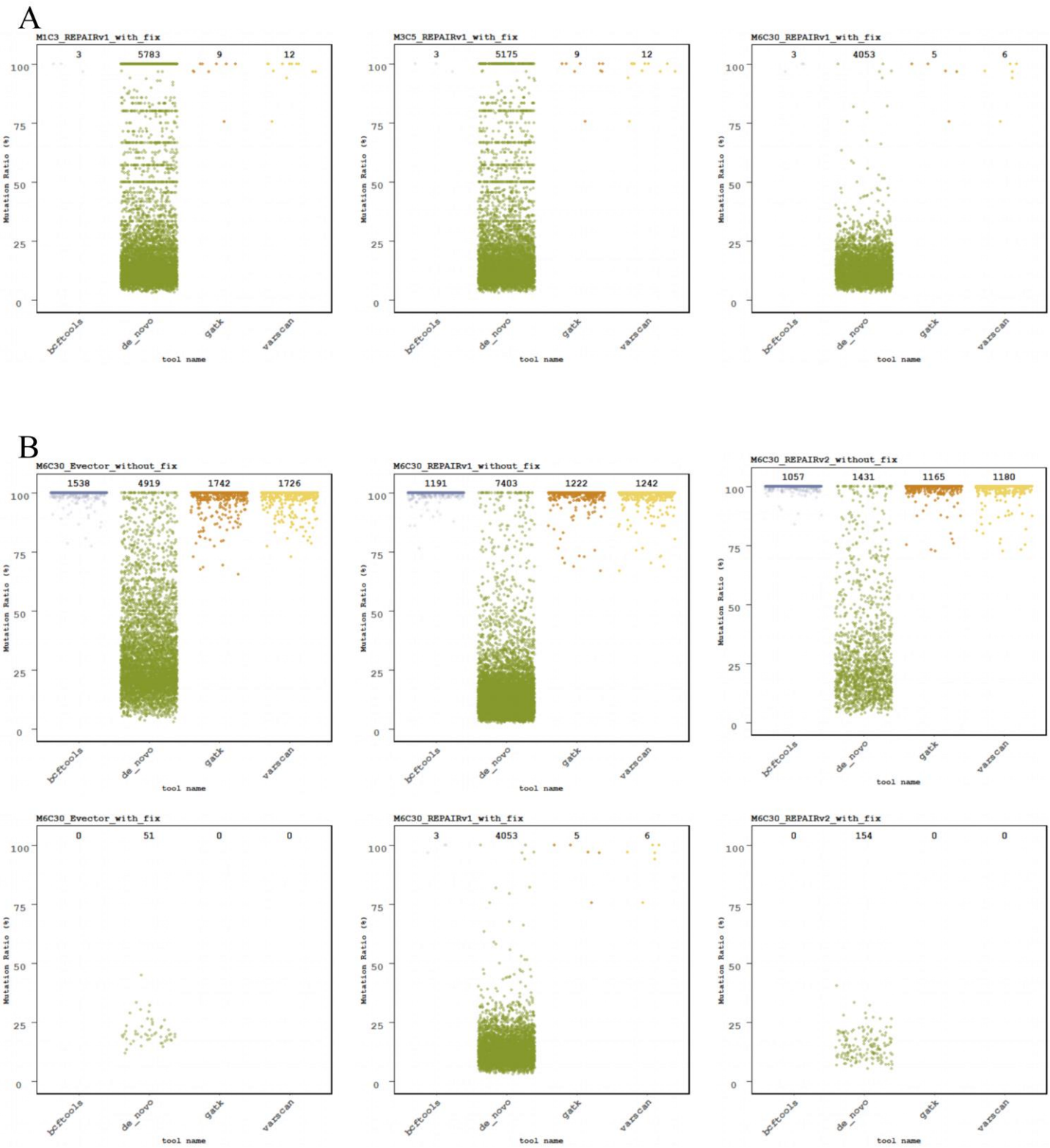


图 4. 相对于工具的位点突变率抖动散点图

A. REPAIRv1 样本去除对照组后在三个 cutoff 下的突变率抖动图，横坐标为工具名称从左至右依次为 bcftools，de novo，gatk 和 varscan，纵坐标为突变率的百分数，图中每个点都代表一个被检测出的位点，可以发现 M6C30 的突变率相较于其他两组更小，并且 de novo 找到的点中，大部分点的突变率都低于 25%；B. 对照组和 REPAIR 系列在三个 cutoff 下去除对照组前后的突变率抖动图，三个 cutoff 间展现的趋势与 A 相同，同时可以发现，去除对照组这一步去除的点大部分突变率在 75%-100%之间，而这些点大概率是细胞本身的 SNV，也验证了去除对照组这一步对于降低假阳性的重要性

如图 4 所展示的结果，M6C30 的 cutoff 和去除对照组这一步骤都较好的去除掉原始点集中疑似假阳性的部分，从另一个角度验证了 3.1.4.1 的结论。此外，无论是在什么样的外加条件下，de novo 流程获得的结果的突变率处于 25%以下的比例都显著高于其他三种工具。因此，本文得出以下结论：

① 去除对照组的；

② 在 M6C30 的 cutoff 下的；

③ 由 de novo 流程分析识别得到的，

脱靶位点结果，是以上所有组合方案中假阳性最低的一种。

而在此条件下得到的 REPAIRv2 脱靶位点个数也远低于 REPAIRv1 的脱靶个数，比较符合张峰实验室所提供的背景数据。可以判定，以上工作流程是合理且可靠的。

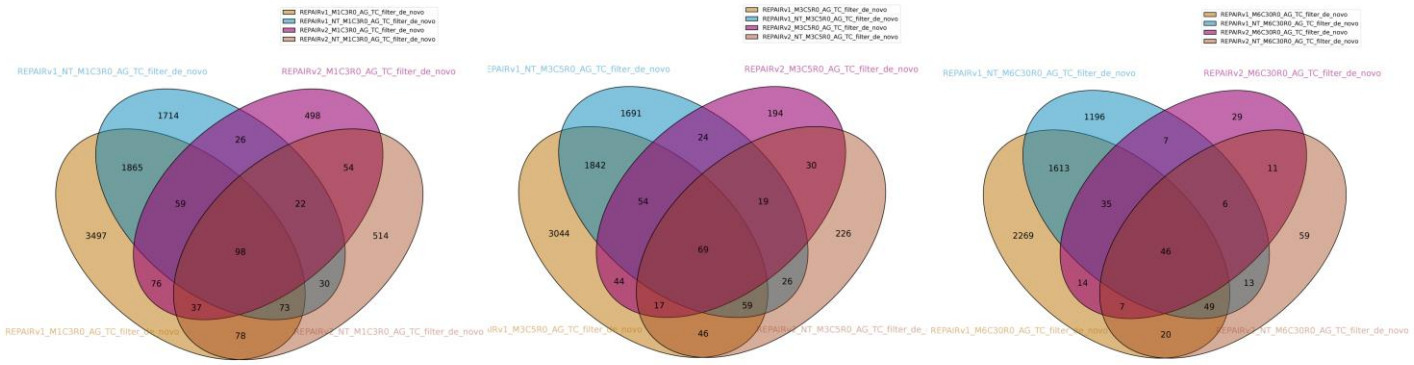
3.1.4.3 Venn 图比较

在 3.1.4.2 的基础上，本实验继续对去除对照组后并经过 M6C30 过滤的 SNP 位点集之间的特异性和共性进行分析。如图 5A 所示，在 M6C30 情况下，至少有 50%以上的脱靶位点是与其他样本共享的，在总脱靶位点较少的 REPAIRv2 系列样本中，这个数字更将达到 70%以上；至少被 3 个样本所共享的点数达到 143 个。这提示 REPAIR 系列工具的脱靶可能存在某种特定的内在模式或者偏好性。

而在 3.1.4.2 已知 de novo 流程的优越性的基础上，图 5B 所显示的工具间的几乎为零的重叠数也某种程度上暗示其他三种工具所存在的高假阴性问题。

由此，因为具有相对低的假阳性和假阴性，de novo 流程在工具比较环节脱颖而出。本文之后部分的分析将用到的数据均来自于去掉对照组后、由 de novo 流程分析并通过 M6C30 过滤所得的 A-to-I 位点信息。

A



B

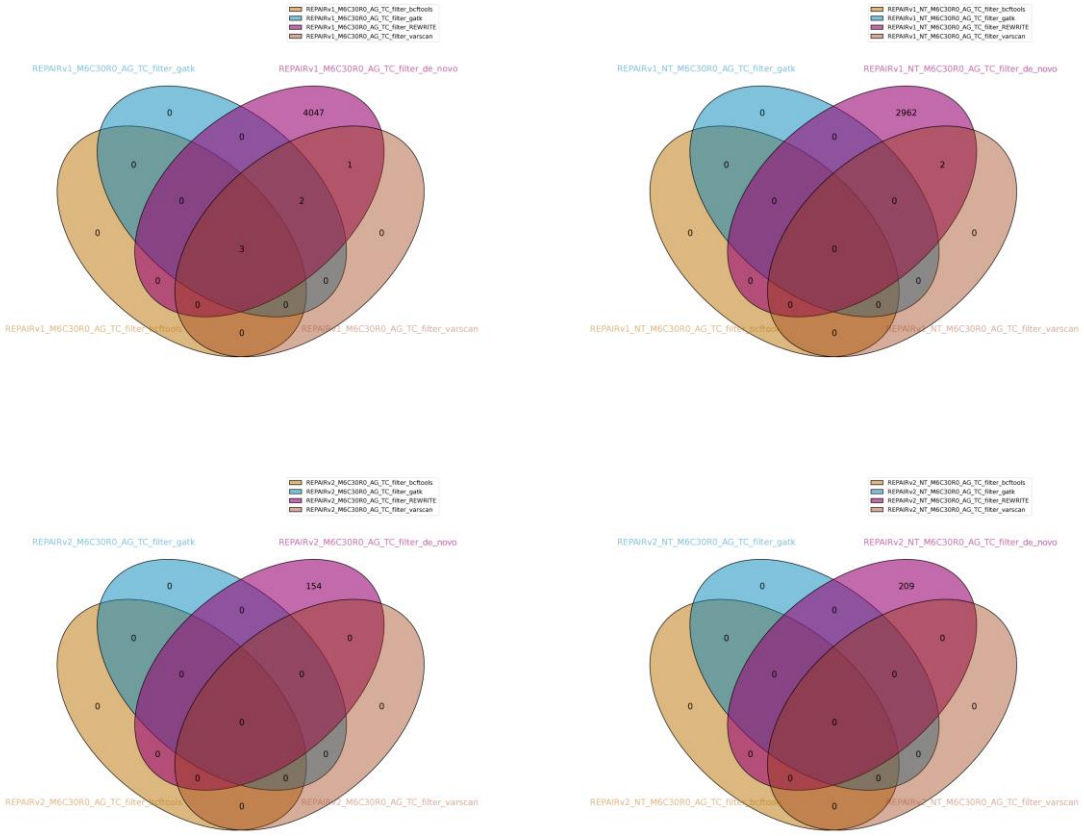
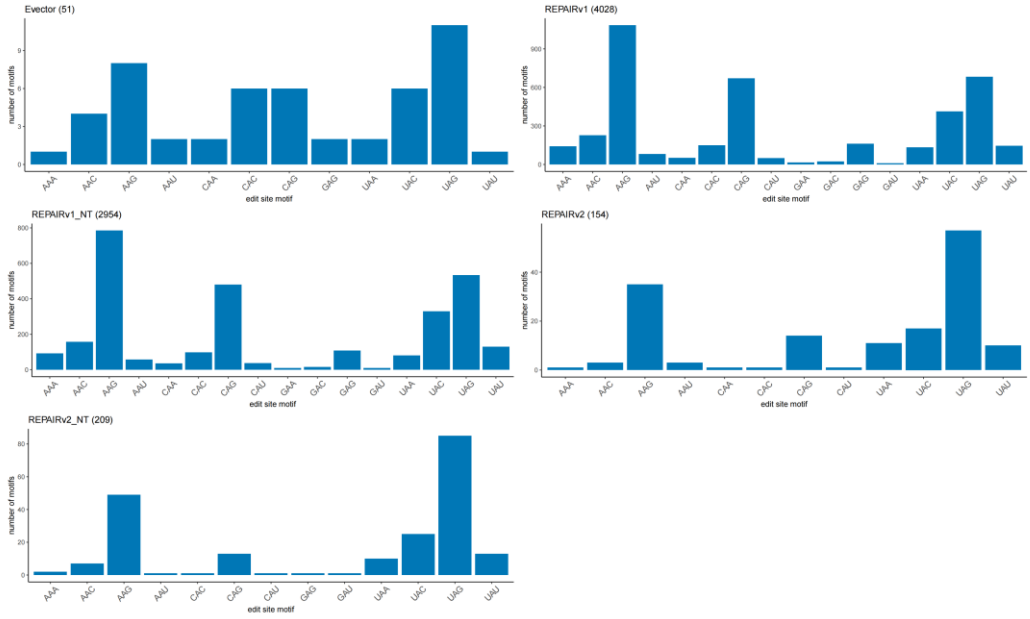


图 5. 工具比较和样本比较的 Venn 图

A. 三个 cutoff 下 de novo 流程所得的不同样本间的 venn 图，从左上角顺时针分别为 REPAIRv1_NT，REPAIRv2，REPAIRv2_NT，REPAIRv1；B. M6C30 cutoff 下 REPAIR 系列样本在不同流程/工具间的 venn 图，从左上角以 Z 字形顺序分别为 REPAIRv1，REPAIRv1_NT，REPAIRv2，REPAIRv2_NT，每幅图内从左上角顺时针分别为 gatk，de novo，varscan，bcftools

A



B

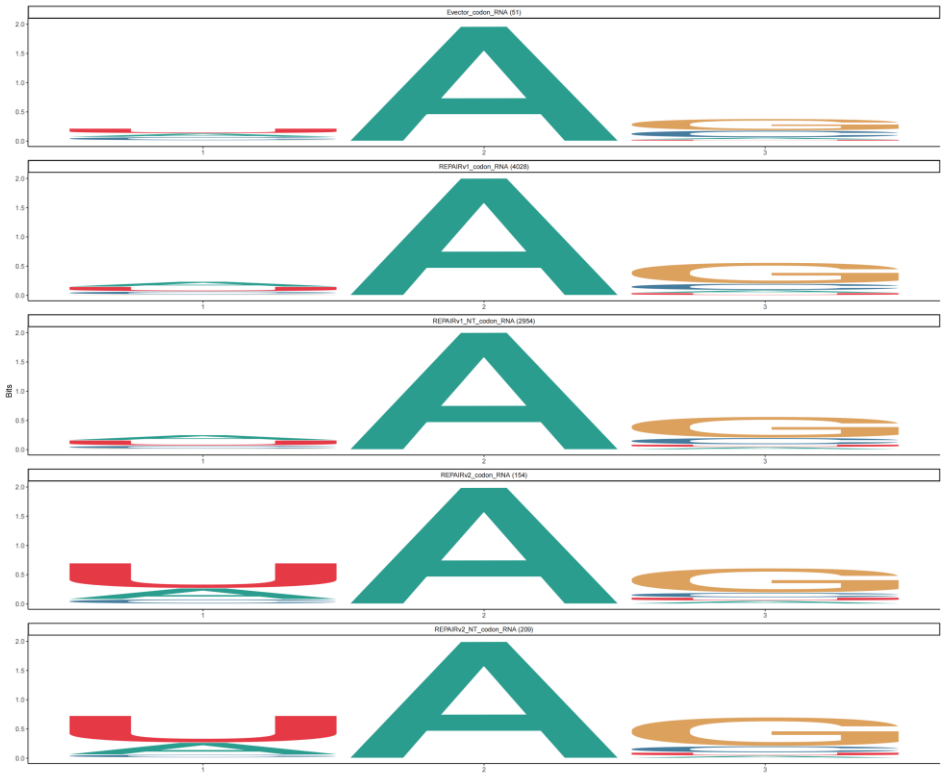


图 6. 序列偏好性分析结果

A. 每个样本中 3bp motif 序列的统计柱状图，从左上角 Z 字形顺序依次为 Evector，REPAIRv1，REPAIRv1_NT，REPAIRv2，REPAIRv2_NT，每张图内横坐标为 motif 种类，纵坐标为计数；B. 每个样本中 3bp 短序列的序列标识图（sequence logo），从上到下依次为 Evector，REPAIRv1，REPAIRv1_NT，REPAIRv2，REPAIRv2_NT

3.2 REPAIR 系列工具脱靶位点的偏好性统计分析

3.2.1 脱靶位点的序列偏好性

为了进一步了解 ADAR 酶的序列偏好性，本实验将脱靶位点及其前后各 1bp 碱基组成的 3bp 碱基单位命名为“motif”，motif 将作为 3.2.1 部分的主要分析对象。

首先，运用脚本提取脱靶位点所在的 motif 序列，并按样本对 motif 进行分类统计。如图 6A 所示，统计得到每个样本中占主导的 motif 类型：Evector（UAG、AAG），REPAIRv1（AAG、UAG），REPAIRv1_NT（AAG、UAG），REPAIRv2（UAG、AAG），REPAIRv2_NT（UAG、AAG）。从图 6B 的 sequence logo 图中也可以发现，motif 第三位是 G 的概率大于是其他三种碱基的概率；第一位是 A 和 U 的概率都较大，是 C 的概率次之。REPAIRv1 系列更偏向于 AAG 类型而 REPAIRv2 系列更偏向于 UAG。

然而值得注意的是，作为对照组的 Evector 样本理论上不应该存在编辑工具造成的脱靶位点，但是因为找点过程中人为设定的标准并不绝对精确，遗留下来的少数位点为细胞本身自带的 SNP。鉴于此，在 Evector 样本中最多的 UAG 类型 motif 可能并不是 ADAR 酶的偏好性 motif。此外，已知 REPAIRv1 的脱靶率高于 REPAIRv2，在不能消除对于造成 UAG 大量富集的内源性疑虑之前，REPAIRv1 所展示的对于 AAG 序列的脱靶偏好性可信度要高于 REPAIRv2 所展示的对于 UAG 序列的偏好性。

以上的结果为后续研究在更安全的环境下运用 REPAIR 系列工具提供了建议，即在实验设计时选择规避在目标位点的编辑窗内存在 AAG 和 UAG 序列的情况，从而减低可能的脱靶的概率；同时对于 NAG 的偏好性也可以被利用到编辑位点的设计当中，从而提高编辑效率。

3.2.2 脱靶位点的 Cas 蛋白依赖性探索

通常情况下，脱靶分为 Cas 蛋白依赖性脱靶和 Cas 蛋白非依赖性脱靶。依赖性脱靶是因为包含脱靶位点的一段序列与 gRNA 具有序列相似性，Cas 蛋白会因此将脱氨酶引导到这段相似但错误的序列上从而造成脱靶。而 Cas 非依赖性脱靶则可能是因为脱氨酶的过度活跃，需要降低其活性以达到降低脱靶率、提高安全性的目的。

为了探究包含脱靶位点的序列是否具有相似性，本研究采取在脱靶位点的左右各延长 20bp，对这些 41bp 的序列进行 sequence logo 绘制。如图 7 所示，所有的样本中都未呈现某一特定序列模式的倾向，因此可以判定 REPAIR 系列工具的脱靶为 Cas 非依赖性脱靶。

对于这种由过度活跃的脱氨酶造成的脱靶，张峰实验室在 v1 到 v2 的升级中就采取了在脱氨酶结构域引入特高特异性的点突变（T375G）来降低 v2 的脱靶。但同时这也牺牲了靶向编辑效率。因此，需要额外的方法来生成既精准又高效的编辑器。

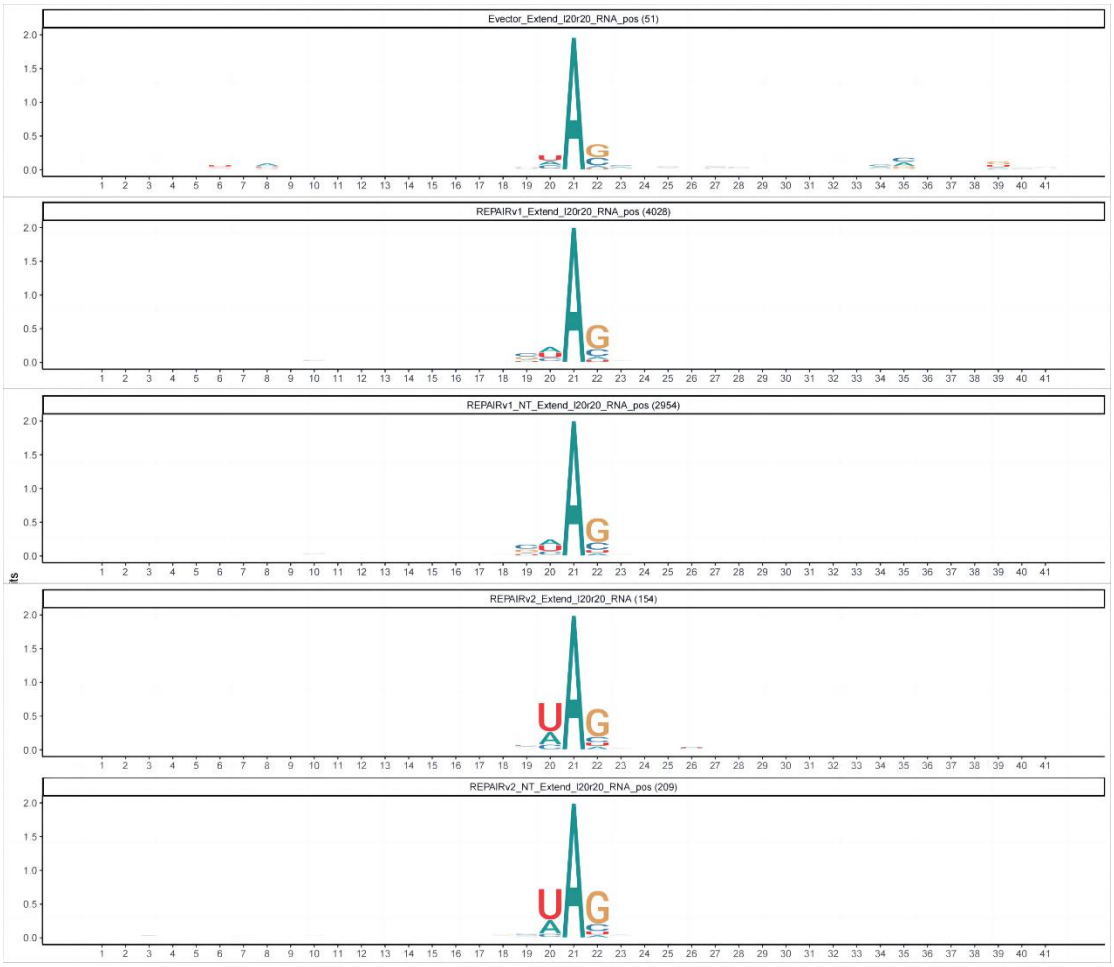


图 7. 脱靶位点及其前后 20bp 序列的 sequence logo

3.3 通过基因表达量分析初步判断脱靶对细胞表达的影响

在具体分析每个脱靶位点造成的变化之前，需要先确认所有的脱靶是否已经对细胞的生存情况或者重要功能造成较大的影响。

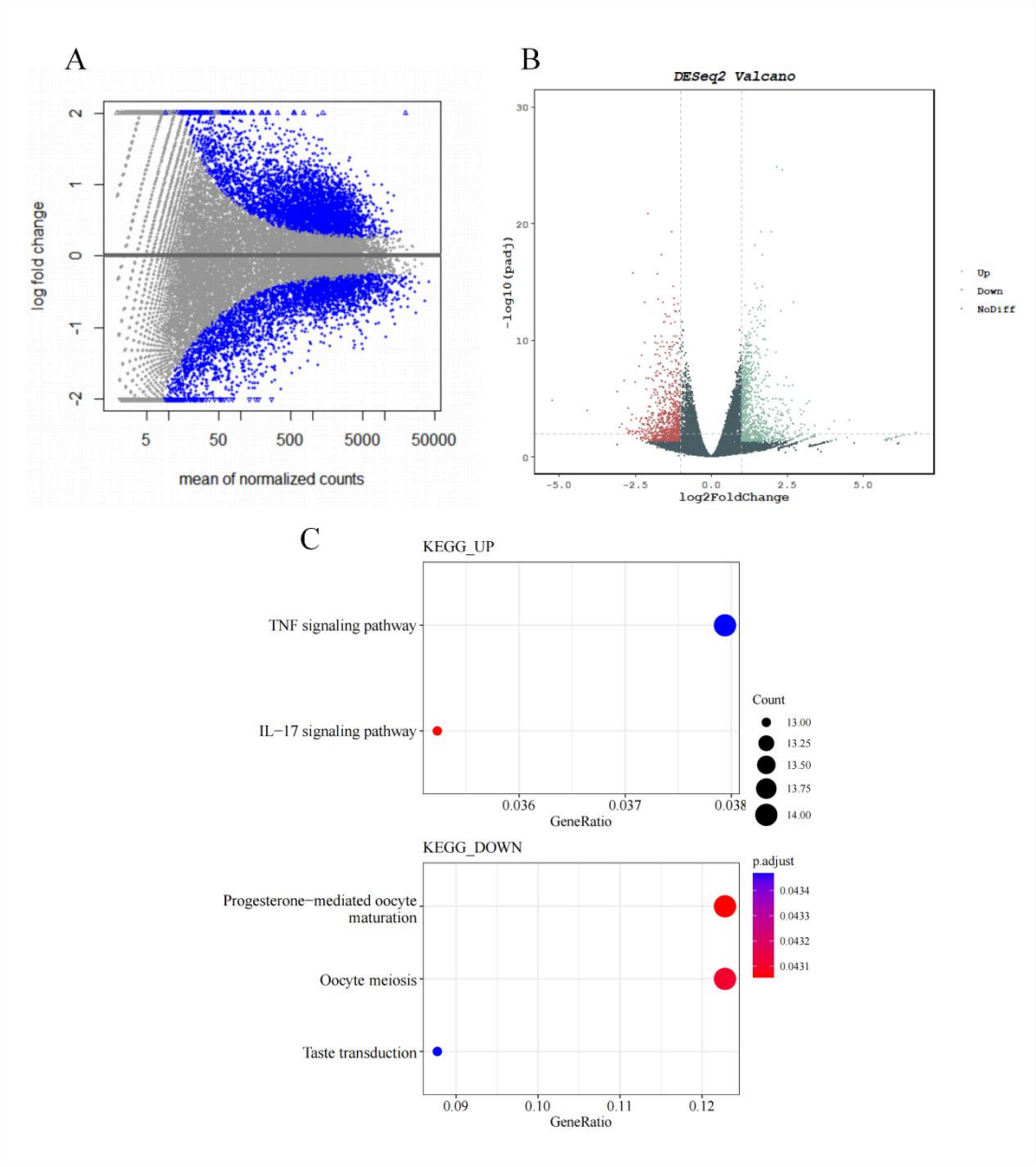


图 8. 差异表达分析和富集分析结果

A. Mean-Dispersion（MA）Plot，x 轴表示基因的平均表达水平，y 轴表示基因表达水平的变异性或差异性（对数倍变化），每个点代表一个基因，灰色的部分没有差异性，蓝色的部分有差异性，参考水平为 Evector；B. Valcano Plot，x 轴表示基因的 $\log_2$ 倍数变化（$\log_2$FC），y 轴表示-$\log_{10}$ 调整后的 p 值（padj），用于表示基因的显著性，其中“Up”表示在实验组相对于对照组（Evector）上调（$\log_2$FC>1）

的基因，“Down”表示在实验组相对于对照组下调（$log_2FC<-1$）的基因，“NoDiff”表示在两组条件之间没有显著差异的基因；C. Kyoto Encyclopedia of Genes and Genomes（KEGG）通路富集气泡图，对 DESeq2 步骤得到的差异表达基因进行筛选，保留 p 值小于 0.05 同时 log_2FC 绝对值大于 1 的基因，同时依据 $log_2FC\geq1$ 和 <1 区分出上调和下降的基因，分别通过 KEGG 将其富集到通路上，并绘制出气泡图

首先对原始的 BAM 文件使用 featureCounts 工具包计数，然后使用 DESeq2，将 Evector 设置为参考水平，分析 REPAIRv1、REPAIRv1_NT、REPAIRv2、REPAIRv2_NT 相对其的差异表达统计，并将得到的结果绘制成 MA 图。如图 8A 所示，MA 图分布相较于 0 轴较为对称，说明数据表达不存在明显异常。接着用 DESeq2 的结果绘制火山图（图 8B），发现大多数基因 log_2FC 的绝对值都在 5 以内，不存在大于 10 的异常值。同时，对 DESeq2 的结果进行 KEGG 富集分析并绘制出的气泡图也没有出现 p 值非常显著（小于 0.01）、基因计数较大的通路，也可以反映对于细胞整体的影响较小。

因此，经过初步分析，本实验找到的脱靶位点并未对细胞整体的功能或者细胞生存造成不可逆的恶劣影响，REPAIR 系列工具的安全性值得进一步评估。

3.4 由脱靶造成的氨基酸变化预测

3.4.1 无义突变和错义突变

基因通过调控蛋白质的合成来影响细胞活动和生成，因此在 3.3 已知 REPAIR 系列工具所造成的的脱靶位点对细胞整体正常运转不造成巨大威胁的基础之上，本研究将进一步探索每个脱靶位点具体会导致的氨基酸的变化。由于关注点在于蛋白编码，所以我们只选取在 CDS 区域上的脱靶位点进行分析。

首先编写一个由序列、CDS（coding sequence）起止位置和脱靶位点的位置得到碱基变化前后所在密码子翻译得到的氨基酸的 Python 函数，在此过程中需要注意脱靶位点在密码子上的相对位置。之后，对每个位点注释，得到位点所属的基因名称。通过基因名称去参考文件中查找并提取出对应的转录本序列以及 CDS 序列的起始位置。再适当延长脱靶位点所在序列（左右各延长 20bp），将这条延长序列比对到参考转录本序列上，从而得到脱靶位点在参考

序列上的位置。整合以上信息，输入到先前编写的 Python 函数中，从而得到位点突变导致的氨基酸变化。

在注释突变靶点之前，除去低质量和覆盖数较低的位点，得到的新的脱靶位点数见表 1 中的“输入点数”；经过注释、能在参考转录本文件中找到对应基因、最后得到碱基改变后氨基酸类型的位点个数见表 1 中的“输出点数”。可以发现，大部分的脱靶位点都在非蛋白编码序列上，被去除。同时值得注意的是，在“输出点数”中计数的位点即最终 Python 函数有输出结果的脱靶位点，也可能存在突变前后对应的氨基酸不发生变化的情况，即无义突变。为了更细致的区分无义突变和相对生物体影响更大的错义突变，在结果中也加上了对于“是否为错义突变”的判断。

表 1 输入点数和输出点数统计表

样本	Evector	REPAIRv1	REPAIRv1_NT	REPAIRv2	REPAIRv2_NT
输入点数	51	4008	2942	153	207
输出点数	2	355	304	29	30

表 2 错义突变前的原始氨基酸统计

AA	Lys	Thr	Gln	Tyr	Ser	Asn	Glu	Arg	Ile	Asp	His	Met	Stop
计数	125	62	49	41	40	39	38	33	31	16	11	11	3

表 3 错义突变后的新氨基酸统计

AA	Arg	Gly	Ala	Cys	Val	Ser	Glu	Met	Asp	Trp
计数	164	127	62	41	30	27	21	12	12	3

由图 9 可以发现，REPAIR 系列的升级，不仅大幅度降低了在蛋白编码序列上的脱靶，并且还大幅降低了有害的错义突变的比例，从原先的 70%降低到了 40%-50%，这使得 REPAIR 工具的安全性得到了巨大的提升。

接着，合并所有样本的数据，并提取出错义突变的部分进一步分析。如表 2 所示，在全部的 499 个错义突变位点中，有 125 个位点的原始氨基酸为赖氨酸，占比将近 25%。赖氨酸所对应的密码子为 AAG 或 AAA，这与 3.2.1 得到的 motif 偏好性 AAG 一致。虽然 motif 不等同于密码子，但是某种程度上还是暗示了大量赖氨酸发生突变可能和 REPAIR 工具的脱靶序列偏好性有关。

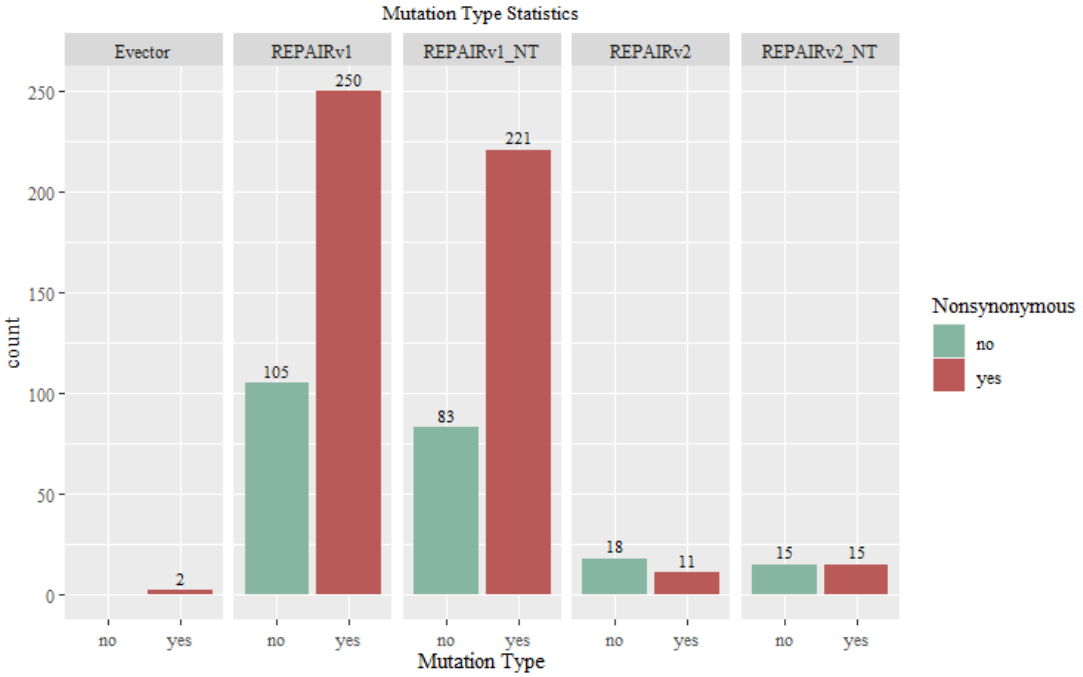


图 9. 无义突变和错义突变个数统计

红色柱子代表错义突变，绿色柱子代表无义突变

除此之外，本实验还对突变后产生的新氨基酸进行了统计。如表 3 所示，精氨酸是突变后得到的最多的氨基酸种类，有 164 个位点。这可能是因为精氨酸是对应密码子最多的氨基酸之一。但通过对 164 个精氨酸位点的原始氨基酸统计分析，发现其中有 104 个都突变自赖氨酸，即 AAG > AGG 或 AAA > AGA。而剩余的 21 个赖氨酸都突变为了谷氨酸，即 AAG > GAG 或 AAA > GAA。在存在序列偏好性的前提下，倾向于认为赖氨酸到精氨酸是源自 AAG > AGG，赖氨酸到谷氨酸是源自 AAA > GAA，还需要进一步一对一去到原序列中回溯进行验证。

3.4.2 错义突变的空间分布和功能分布

3.4.2.1 在染色体上的分布

由图 10A 可以发现，错义突变主要集中分布在 1 号、19 号、11 号、3 号、2 号和 17 号染色体上，这些染色体上的错义突变总和已经将近总数的 50%。这指示后续研究在使用 REPAIR 系列工具并设计编辑位点时，要考虑规避这些较高风险的序列区域。

3.4.2.2 在通路上的富集结果

对错义突变所在的基因进行 KEGG 富集，选取所有显著的通路（图 10B），并依据通路分类绘制每个基因的数量统计图（图 10C）。发现显著性排名前三的通路为 Polycomb 抑制复合体（Polycomb repressive complex, PRC），肌萎缩性侧索硬化症（Amyotrophic lateral sclerosis, ALS），细胞周期（Cell cycle），基因数量排名前三的通路是肌萎缩性侧索硬化症（Amyotrophic lateral sclerosis, ALS），癌症中的蛋白多糖（Proteoglycans in cancer），细菌侵入上皮细胞（Bacterial invasion of epithelial cells）。其中，肌萎缩性侧索硬化症、亨廷顿病、脊髓小脑性共济失调均为神经退行性疾病。在缺少进一步实验验证的情况下，本文暂时无法确认由脱靶造成的错义突变一定会导致某种病理性的结果，但是目前的实验结果已经提供了 REPAIR 系列工具的脱靶效应会集中于神经退行性疾病、癌症以及细菌感染的通路上的直接证据，并将启发未来对于脱靶在这些通路上产生的影响的更深入的研究。

3.4.3 根据实验结果给出的更安全使用 REPAIR 系列工具的建议

1. 在 REPAIR 系列更完善的工具推出之前，更推荐使用 REPAIRv2 进行目标位点编辑，以其造成的错义突变更少，安全性相对更高；

2. 尽量在编辑窗区域内避免精氨酸密码子的存在，以降低因为序列偏好性而造成的精氨酸上的脱靶突变；

3. 在 1 号、19 号、11 号、3 号、2 号和 17 号染色体上进行的编辑设计，需考虑编辑窗区域或者基因组上下游更广的范围内是否存在行使重要编码功能的基因，如有可以选择规避或者通过表观修饰等手段对敏感区域进行保护；

4. 该系列工具不建议使用在家族遗传谱系中存在癌症和神经退行性疾病高发病率的人群或者已经发病的人群中，有诱导发病或者加重病程的可能性。

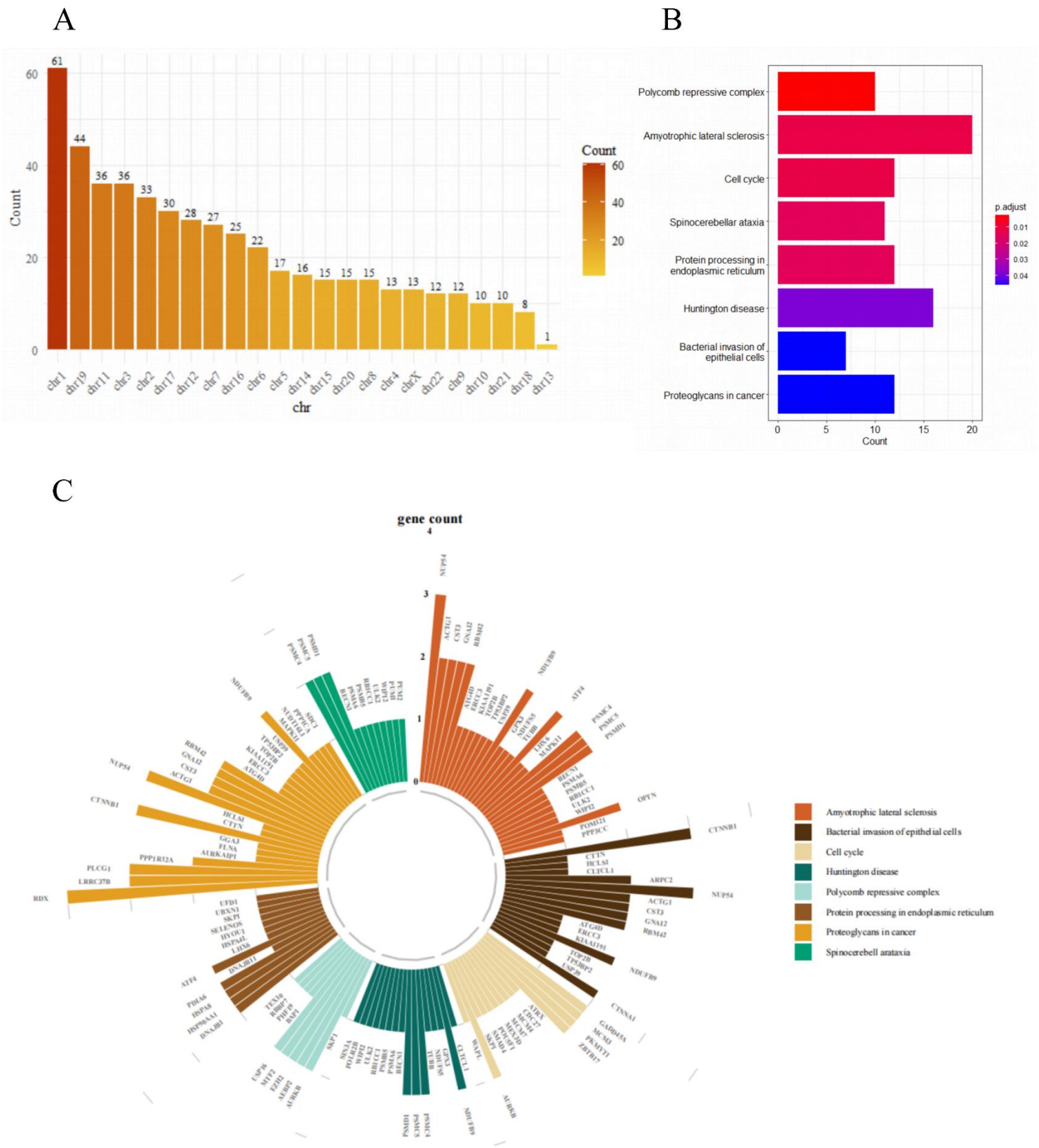


图 10. 错义突变在染色体上的分布和通路上的分布

A. 错义突变在染色体上的分布，由高到低在一号染色体上分布的错义突变最多，有 61 个；B. 对错义突变所在的基因进行 KEGG 通路富集，如图显示具有显著性（p-value < 0.05）的 8 条通路，显著性从高到低分别是：Polycomb 抑制复合体（Polycomb repressive complex, PRC），肌萎缩性侧索硬化症（Amyotrophic lateral sclerosis, ALS），细胞周期（Cell cycle），脊髓小脑性共济失调（Spinocerebellar

ataxia, SCA），蛋白质在内质网中的加工（Protein processing in endoplasmic reticulum），亨廷顿病（Huntington disease, HD），细菌侵入上皮细胞（Bacterial invasion of epithelial cells），癌症中的蛋白多糖（Proteoglycans in cancer）；C. 对错义突变所在基因的分类环形柱状统计图，相同颜色表示所属同一通路，条形的长短表示所有样本中该基因的总个数，条形顶端是对应的基因名

四、讨论

就如前言中所提及的，GATK-Haplotypecaller 和 VarScan2 都是经过验证并且实用性和适用性都很强的 variant caller，但是在本课题中的表现却不尽如人意。这可能是源于本课题所用的数据的特殊性和工具本身的算法不相适合。数据上本实验采用 RNA-seq 数据，缺乏类似于 WGS 等的大量数据可用于模型训练从而取出假阳性；对于脱靶的定义上，本课题需要过滤得到突变率低于编辑效率的位点，这可能使得很多工具在低覆盖度时由于内置算法无法很好的辨别出变异。这里需要理解 GATK-Haplotypecaller、BCFtools 和 VarScan2 所用到的算法。

其中，GATK HaplotypeCaller 主要基于贝叶斯方法[51]，在通过打分识别变异区域后，再使用图算法重构可能的单倍型，并计算每个候选单倍型的似然，最后使用贝叶斯推理框架，结合先验信息和计算出的单倍型似然，来判断每个位点上的变异。HaplotypeCaller 会评估不同单倍型组合的概率，选择具有最高后验概率的组合，并据此进行变异检测。本实验使用的 BCFtools 算法（Multiallelic calling，-m）也是依据贝叶斯算法建立[52,53]：将读取质量得分计算每个样本的基因型似然 $Gi(xy)$作为观测数据给定基因型的概率 $P(\text{data}|\text{genotype})$；将遗传学理论计算等位基因集合的先验概率 $P(S)$设定为先验概率；最后结合观测数据的似然和先验概率，得到后验概率 Ls，用于评估给定等位基因集合的可信度。

对于贝叶斯算法的过于依赖就导致在本课题中，由于脱靶数量少且规律性较低，可能在每个位点只有少量的测序读数覆盖，使得变异检测的统计功效降低，而偶然的测序错误更可能被误认为是变异。此外，由于贝叶斯模型依赖于先验假设，例如等位基因频率的先验分布。如果先验假设与实际数据分布不匹配，可能会导致过度调用变异。例如，如果模型假设变异频率较高或者模型假设过于简单不符合更复杂更无规律的脱靶数据分布，则可能会倾向于识别出本身自带的 SNV 而无法识别出突变率较低的脱靶位点，从而增加假阳性。

VarScan2 主要使用的则是启发式算法[54]，摆脱了对于先验概率的依赖，虽然能通过阈值设置识别出低覆盖度区域内的少数变异位点，但同样可能因此而

囊括因测序低质量产生的错误位点而假阳性增高。同时对于阈值的设置在脱靶这种无同一性的研究对象上也很难精准把握，从而很难兼顾“找的全”和假阳性低。

在使用以上方法进行脱靶检测时，需要对测序数据的质量、参考数据的选取以及结果的检验校正进行严格的控制，虽然可以提高检测效率并降低假阳性，但是操作的繁琐也使得它们的可推广度大大降低。

当然，脱靶检测的初衷是为了为更安全的使用基因编辑工具提供参考建议，在未来我们期望有更高效且安全的编辑工具出现，与更精准的检测工具的合作，使得基因编辑广泛用于疾病治疗的愿景能够早日实现。

参考文献

[1] BAK R O, GOMEZ-OSPINA N, PORTEUS M H. Gene Editing on Center Stage[J]. Trends in Genetics, 2018, 34(8): 600-611.

[2] KIM H, KIM J S. A guide to genome engineering with programmable nucleases[J]. Nature Reviews. Genetics, 2014, 15(5): 321-334.

[3] COX D B T, PLATT R J, ZHANG F. Therapeutic genome editing: prospects and challenges[J]. Nature Medicine, 2015, 21(2): 121-131.

[4] KHALIL A M. The genome editing revolution: review[J]. Journal of Genetic Engineering and Biotechnology, 2020, 18(1): 68.

[5] HAAPANIEMI E, BOTLA S, PERSSON J, et al. CRISPR–Cas9 genome editing induces a p53-mediated DNA damage response[J]. Nature Medicine, 2018, 24(7): 927-930.

[6] EGLI D, ZUCCARO M V, KOSICKI M, et al. Inter-homologue repair in fertilized human eggs?[J]. Nature, 2018, 560(7717): E5-E7.

[7] DOUDNA J A. The promise and challenge of therapeutic genome editing[J]. Nature, 2020, 578(7794): 229-236.

[8] REES H A, LIU D R. Base editing: precision chemistry on the genome and transcriptome of living cells[J]. Nature Reviews Genetics, 2018, 19(12): 770-788.

[9] GAUDELLI N M, KOMOR A C, REES H A, et al. Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage[J]. Nature, 2017, 551(7681): 464-471.

[10] YI Z, QU L, TANG H, et al. Engineered circular ADAR-recruiting RNAs increase the efficiency and fidelity of RNA editing in vitro and in vivo[J]. Nature Biotechnology, 2022, 40(6): 946-955.

[11] GRÜNEWALD J, ZHOU R, GARCIA S P, et al. Transcriptome-wide off-target RNA editing induced by CRISPR-guided DNA base editors[J]. Nature, 2019, 569(7756): 433-437.

[12] RAGURAM A, BANSKOTA S, LIU D R. Therapeutic in vivo delivery of gene editing agents[J]. Cell, 2022, 185(15): 2806-2827.

[13] MURUGAN K, SEETHARAM A S, SEVERIN A J, et al. CRISPR-Cas12a has widespread off-target and dsDNA-nicking effects[J]. Journal of Biological Chemistry, 2020, 295(17): 5538-5553.

[14] KLEINSTIVER B P, TSAI S Q, PREW M S, et al. Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells[J]. Nature Biotechnology, 2016, 34(8): 869-874.

[15] LAZZAROTTO C R, MALININ N L, LI Y, et al. CHANGE-seq reveals genetic and epigenetic effects on CRISPR–Cas9 genome-wide activity[J]. Nature Biotechnology, 2020, 38(11): 1317-1327.

[16] PECORI R, DI GIORGIO S, PAULO LORENZO J, et al. Functions and consequences of AID/APOBEC-mediated DNA and RNA deamination[J]. Nature Reviews Genetics, 2022, 23(8): 505-518.

[17] RALLAPALLI K L, KOMOR A C. The Design and Application of DNA-Editing Enzymes as Base Editors[J]. Annual Review of Biochemistry, 2023, 92(1): 43-79.

[18] TSAI S Q, ZHENG Z, NGUYEN N T, et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases[J]. Nature Biotechnology, 2015, 33(2): 187-197.

[19] TSAI S Q, NGUYEN N T, MALAGON-LOPEZ J, et al. CIRCLE-seq: a highly sensitive in

vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets[J]. Nature Methods, 2017, 14(6): 607-614.
[20] BAE S, PARK J, KIM J S. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases[J]. Bioinformatics (Oxford, England), 2014, 30(10): 1473-1475.
[21] CONCORDET J P, HAEUSSLER M. CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens[J]. Nucleic Acids Research, 2018, 46(W1): W242-W245.
[22] JIN S, ZONG Y, GAO Q, et al. Cytosine, but not adenine, base editors induce genome-wide off-target mutations in rice[J]. Science (New York, N.Y.), 2019, 364(6437): 292-295.
[23] DOMAN J L, RAGURAM A, NEWBY G A, et al. Evaluation and minimization of Cas9-independent off-target DNA editing by cytosine base editors[J]. Nature Biotechnology, 2020, 38(5): 620-628.
[24] RICHTER M F, ZHAO K T, ETON E, et al. Phage-assisted evolution of an adenine base editor with improved Cas domain compatibility and activity[J]. Nature Biotechnology, 2020, 38(7): 883-891.
[25] LEI Z, MENG H, RAO X, et al. Detect-seq, a chemical labeling and biotin pull-down approach for the unbiased and genome-wide off-target evaluation of programmable cytosine base editors[J]. Nature Protocols, 2023, 18(7): 2221-2255.
[26] LI H. Toward better understanding of artifacts in variant calling from high-coverage samples[J]. Bioinformatics, 2014, 30(20): 2843-2851.
[27] MEYER L R, ZWEIG A S, HINRICHS A S, et al. The UCSC Genome Browser database: extensions and updates 2013[J]. Nucleic Acids Research, 2013, 41(D1): D64-D69.
[28] TUNG N V, LIEN N T K, HOANG N H. A comparison of three variant calling pipelines using simulated data[J]. Academia Journal of Biology, 2021, 43(2): 47-53.
[29] YU X, SUN S. Comparing a few SNP calling algorithms using low-coverage sequencing data[J]. BMC Bioinformatics, 2013, 14(1): 274.
[30] YI M, ZHAO Y, JIA L, et al. Performance comparison of SNP detection tools with illumina exome sequencing data—an assessment using both family pedigree information and sample-matched SNP array data[J]. Nucleic Acids Research, 2014, 42(12): e101.
[31] PIROOZNIA M, KRAMER M, PARLA J, et al. Validation and assessment of variant calling pipelines for next-generation sequencing[J]. Human Genomics, 2014, 8(1): 14.
[32] PABINGER S, DANDER A, FISCHER M, et al. A survey of tools for variant analysis of next-generation genome sequencing data[J]. Briefings in Bioinformatics, 2014, 15(2): 256-278.
[33] O'RAWE J, JIANG T, SUN G, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing[J]. Genome Medicine, 2013, 5(3): 28.
[34] LIU X, HAN S, WANG Z, et al. Variant Callers for Next-Generation Sequencing Data: A Comparison Study[J]. PLOS ONE, 2013, 8(9): e75619.
[35] CHENG A Y, TEO Y Y, ONG R T H. Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals[J]. Bioinformatics, 2014, 30(12): 1707-1713.
[36] BAUER D. Variant calling comparison CASAVA1.8 and GATK[J]. Nature Precedings, 2011: 1-1.
[37] MCKENNA A, HANNA M, BANKS E, et al. The Genome Analysis Toolkit: A MapReduce

framework for analyzing next-generation DNA sequencing data[J]. Genome Research, 2010, 20(9): 1297-1303.
[38] KOBOLDT D, ZHANG Q, LARSON D, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing[J]. Genome Research, 2012.
[39] DC K, K C, T W, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples.[J]. Bioinformatics (Oxford, England), 2009, 25(17): 2283-2285.
[40] NARASIMHAN V, DANECEK P, SCALLY A, et al. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data[J]. Bioinformatics, 2016, 32(11): 1749-1751.
[41] GATK[EB/OL]. [2024-05-13]. https://gatk.broadinstitute.org/hc/en-us.
[42] RNAseq short variant discovery (SNPs + Indels)[EB/OL]. (2024-05-07)[2024-05-13]. https://gatk.broadinstitute.org/hc/en-us/articles/360035531192-RNAseq-short-variant-discovery-SNPs-Indels.
[43] HaplotypeCaller[EB/OL]. (2023-03-20)[2024-05-13]. https://gatk.broadinstitute.org/hc/en-us/articles/13832687299739-HaplotypeCaller.
[44] SplitNCigarReads[EB/OL]. (2023-03-20)[2024-05-13]. https://gatk.broadinstitute.org/hc/en-us/articles/13832774383643-SplitNCigarReads.
[45] BaseRecalibrator[EB/OL]. (2023-03-20)[2024-05-13]. https://gatk.broadinstitute.org/hc/en-us/articles/13832708374939-BaseRecalibrator.
[46] VariantFiltration[EB/OL]. (2023-03-20)[2024-05-13]. https://gatk.broadinstitute.org/hc/en-us/articles/13832750065947-VariantFiltration.
[47] SelectVariants[EB/OL]. (2023-03-20)[2024-05-13]. https://gatk.broadinstitute.org/hc/en-us/articles/13832694334235-SelectVariants.
[48] LI H, HANDSAKER B, WYSOKER A, et al. The Sequence Alignment/Map format and SAMtools[J]. Bioinformatics, 2009, 25(16): 2078-2079.
[49] DANECEK P, BONFIELD J K, LIDDLE J, et al. Twelve years of SAMtools and BCFtools[J]. GigaScience, 2021, 10(2): giab008.
[50] COX D B T, GOOTENBERG J S, ABUDAYYEH O O, et al. RNA editing with CRISPR-Cas13[J]. Science, 2017, 358(6366): 1019-1027.
[51] GATK-HaplotypeCaller 变异检测详解_gatk haplotypecaller 参数-CSDN 博客[EB/OL]. [2024-05-18]. https://blog.csdn.net/qq_28723681/article/details/123564608.
[52] Issues · samtools/bcftools[EB/OL]. [2024-05-18]. https://github.com/samtools/bcftools.
[53] DANECEK P, SCHIFFELS S, DURBIN R. Multiallelic calling model in bcftools (-m)[J].
[54] KOBOLDT D C, LARSON D E, WILSON R K. Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection[J]. Current Protocols in Bioinformatics, 2013, 44(1): 15.4.1-15.4.17.

致 谢

这个课题起源于2023年的暑期，中间因为留学申请停止了一段时间，随后又于今年春节后重启。非常感激伊老师能在那个时候给予了我这个“门外汉”机会，让我得以能够接触到全新的科研环境和优秀的师兄师姐们。然后一切好像就顺水推舟，我逐渐学习逐渐适应，自己曾经的忧虑都得到了化解，不再质疑自己的能力和选择。非常感谢我的师兄乌浩和师姐庄元，他们给予了我无限的科研上的指导、情绪上的支持和前途规划上的建议，没有他们这个课题和这段时光将会艰难许多。

毕设进行的这段日子我都是住在北京，而这段时间正好也是留学申请出结果的时候。前三年在上海在复旦的所有积淀，都在这短短的几个月内显像化，而我又是在距离我所有的奋斗和耕耘的千里之外的北京收获所有的好消息和坏消息，这种时空上的错位感也是非常奇妙。但正是这种地理意义上的分隔，让我的心绪很好的从申请的后劲中剥离出来，全身心的投入到毕设中。所以北京的这段日子在我的记忆力中只留下了脚本跑通的喜悦、阳光和晚风。

最后，还想感谢我的家人和挚友郭女士、李女士、高女士和吕女士，他们也是我重要的情感支撑。

其他无需多言，不用记录也不会忘却。