

AN UNBIASED COMPUTATIONAL ANALYSIS OF CRISPR/CAS9 OFF-TARGET EFFECTS

Wei Chen, Dyuti Islam, Qingsong Zhao, Marjan Trutschl, Urska Cvek
¹ LSU Shreveport, One University Pl, Shreveport, Louisiana, USA

Corresponding Author: Urska Cvek
Email: ucvek@lsus.edu
doi: 10.34107/UDUK9890258

ABSTRACT

The CRISPR/Cas9 system is a powerful genome-editing tool with therapeutic potential for previously untreatable diseases. However, off-target effects remain under-characterized in both experimental and computational studies. Most current prediction tools rely heavily on sequence homology, which may overlook chromatin context and DNA dynamics. Whole-genome sequencing (WGS) provides an unbiased mean to detect gene editing, including CRISPR-induced cleavage events, where abrupt read drop-offs or pile-ups signal DNA breaks, quantified via start and end coverages. We developed a computational pipeline that leverages clipping coverage signals in WGS data to identify CRISPR-induced DNA cleavage events, enabling visualization of editing outcomes and off-target risks.

Keywords: CRISPR/Cas9, off-target, whole-genome sequencing, CIGAR string, DNA dynamics, homology direct repair

INTRODUCTION

The clustered, regularly interspaced short palindromic repeats (CRISPR) and CRISPR-associated protein 9 (Cas9) is a genome editing system using intracellular double-strand breaks (DSB) followed by the Non-Homologous End Joining (NHEJ) or Homology Direct Repair (HDR) mechanisms [1]. Currently, major assessments of CRISPR/Cas9 on- and off-target activity rely primarily on in silico tools such as CRISPOR, Cas-OFFinder, and CCTop, which base their predictions on sequence homology to the guide RNA and the identification of protospacer adjacent motifs (PAMs). Nevertheless, these methods overlook important biological factors, including PAM flexibility [2, 3], DNA physical properties such as chromatin accessibility and dynamics [4, 5], and guide RNA secondary structure [6]. Recent studies emphasize that Cas9-induced off-targets are not limited to simple mismatches but include structural variants and large genomic alterations detectable only by long-read sequencing technologies [7, 8]. Moreover, deep-learning models and neural-network predictors such as crispAI have begun incorporating uncertainty estimates to improve prediction reliability [9]. Here, we present an unbiased WGS-based framework that refines read classification and introduces a novel CIGAR string operation ('R') to improve detection of HDR events and off-target cleavage, moving beyond sequence homology-based predictions.

MATERIALS & METHODS

Read Classification. Raw sequencing data were obtained from the NCBI Sequence Read Archive (SRA) under accession numbers DRR278262 (or DRR278261 with PAM mutation). Adapter sequences were removed from the raw FASTQ files using Trimmomatic [10]. Filtered reads were aligned to the reference genome using the Burrows-Wheeler Aligner Mem algorithm (BWA-MEM) [11].

| Criterion | Primary Reads | Secondary Reads |
|------------------------|----------------------|-----------------|
| Mapping Quality (MAPQ) | ≥ 20 | > 10 |
| Mismatches | ≤ 5 | > 5 |
| Soft/Hard Clipping | ≤ 10 base pairs (bp) | > 10 bp |
| Indels | < 10 bp | ≥ 10 bp |