Universal CRISPR-based RNA virus detection and inhibition via foundation model

Guohui Chuai[1,2,3, *], Jieyu Cui [4, *], Chen Zhang[1,2,3, *],Qinchang Chen[1,2,3, *], Jiaying Wen[1,2,3], JiaXin Han[1,2,3], Kejing Dong[1,2,3], Lina Wu[7], Xingxu Huang[5,6], Xinjie Wang[4,5, #], Qi Liu[1, 2, 3, #]

#Corresponding authors

*These authors contribute equally to this work

1 Bioinformatics Department, School of Life Sciences and Technology, Tongji University, Shanghai, 200092, China

2 National Key Laboratory of Autonomous Intelligent Unmanned Systems, Frontiers Science Center for Intelligent Autonomous Systems, Ministry of Education, Shanghai Key Laboratory of Intelligent Autonomous Systems 201804, China

3 Shanghai Qi Zhi Institute, 200232, China

4 Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China

5 Guangzhou National Laboratory, Bio-island, Guangzhou, Guangdong, 510005, China

6 Laboratory of Pancreatic Disease, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310058, China

7 School of Food Science and Pharmaceutical Engineering, Nanjing Normal University, Nanjing, 210023, PR China; Food Laboratory of Zhongyuan, Luohe, 462300, Henan, PR China

Abstract

CRISPR-based technologies offer promising avenues for RNA virus detection and intracellular inhibition. However, the computational design of optimal guide RNAs (gRNAs) is bottlenecked by data-hungry algorithms that struggle to rapidly adapt to newly discovered Cas effectors and emerging viral variants. Furthermore, existing tools largely rely on static reference genomes, ignoring individual genetic polymorphisms (SNPs) and tissue-specific transcriptomes, which risks severe target-activated collateral toxicity. Here, we present CRISPR-viva, a sequence-to-function foundation model for universal CRISPR-based RNA virus detection and inhibition. Pre-trained on a massive unlabelled RNA corpus, CRISPR-viva captures the universal syntax of RNA targeting. This enables few-shot adaptation, rapidly generating highly efficacious gRNAs for novel CRISPR systems and emerging viral variants using minimal experimental training data. Crucially, CRISPR-viva features a dynamic host-context integration pipeline that rigorously filters off-target candidates based on patient-specific SNPs and tissue expression profiles, effectively preventing host-cell toxicity and diagnostic false-positives. We validated the framework across 8 CRISPR systems, demonstrating its precision through in vitro LbuCas13a-based detection and cell-based Cas13d viral inhibition assays. To showcase its scalable inference capability, we deployed CRISPR-viva across 200 segmented viral genomes and 23 individual-derived human cell types, profiling over 300 million gRNA candidates. Overall, CRISPR-viva shifts the paradigm of gRNA design from isolated, task-specific algorithms to a highly adaptable foundation model, providing an agile and precise computational infrastructure for combatting emerging viral threats.

Introduction

Emerging RNA viruses pose a significant threat to global health, causing highly contagious diseases such as influenza, dengue fever, Ebola haemorrhagic fever, and severe respiratory syndromes including SARS[1], MERS[2], and COVID-19[3]. The sensitive, specific, and field-deployable detection of causal RNA viruses, alongside effective intracellular viral inhibition, is critical for disease surveillance and therapeutic intervention. With the rapid evolution of clustered regularly interspaced short palindromic repeats (CRISPR) technologies, programmable RNA-targeting Cas effectors (e.g., Cas13[4], Cas12[5]) have emerged as highly attractive tools for both viral diagnostics and antiviral treatments. Various CRISPR-based methods have been developed to tackle these threats[5-7], including SHERLOCK[8], DETECTR[9], and SENSR[10] for highly sensitive nucleic acid detection, and systems like PAC-MAN[11] for Cas13d-mediated intracellular viral neutralization.

Despite these remarkable experimental advances, the computational design of optimal guide RNAs (gRNAs) remains a fundamental bottleneck for their rapid clinical deployment, facing three critical challenges:

(1) The lackness of universal RNA-targeting models: Existing deep learning algorithms for CRISPR predictive modelling (such as ADAPT[12], TIGER[13], and DeepCas13[14]) are typically task-specialized and built from scratch for a single Cas effector. These isolated models are highly data-hungry, relying entirely on massive amounts of explicitly labelled screening data, and fundamentally fail to capture the universal syntax and interaction rules of RNA targeting.

(2) The inability of rapidly adaption to novel CRISPR systems: The continuous discovery of novel CRISPR effectors (e.g., Cas7-11[15], Cas14[16]) and the rapid mutation rate of RNA viruses demand computational tools capable of rapid adaptation. Traditional models struggle to generalize in time-sensitive emergency scenarios where only highly limited experimental data are available for a newly developed CRISPR system or a newly emerged viral variant.

(3) The neglection of the dynamic and personalized host context: Current computational tools largely evaluate gRNA efficacy based on static reference genomes, completely ignoring individual genetic polymorphisms (SNPs) and tissue-specific transcriptomic expression profiles. This is particularly perilous for Cas13-based viral inhibition; as Cas13 exhibits target-activated non-specific collateral cleavage, erroneous binding to a highly expressed host endogenous transcript, which caused by an overlooked individual SNP or a tissue-specific RNA isoform that can trigger catastrophic collateral degradation and severe cellular toxicity.

To address these interwoven challenges of data scarcity and off-target toxicity, we conceptualize gRNA design not merely as a local sequence-matching problem, but as a representation learning challenge of RNA sequence syntax. Leveraging recent breakthroughs in large language models, we present CRISPR-viva (the **CRISPR**-based RNA **vi**rus detection and inhibition **v**ia found**a**tion model), a sequence-to-function foundation model designed for universal CRISPR-based RNA virus detection and inhibition. Instead of training isolated models for individual Cas effectors, CRISPR-viva was pre-trained on a massive corpus of unlabelled viral and human RNA sequences using self-supervised masked language modelling. This pre-training enables the foundation model to capture the deep latent representations and universal interactions of RNA sequence grammar.

Serving as the core computational engine, the foundation model-driven CRISPR-viva framework achieves three major methodological advances that directly resolve the aforementioned bottlenecks:
(1) Universal sequence-to-function representation: By learning the intrinsic grammar of RNA sequences, the foundation model establishes a unified predictive architecture that significantly outperforms existing isolated methods in predicting RNA-targeting efficacy across multiple CRISPR tasks, without relying on massive scratch-built datasets.
(2) Few-shot rapid response adaptation: Exploiting the latent knowledge acquired during pre-training, CRISPR-viva can be rapidly generalized to newly discovered CRISPR systems and viral variants using only a limited amount of functional data (few-shot learning), successfully breaking the dependency on exhaustive high-throughput screening and provide a rapid-response system for emergent virus outbreaks.
(3) Dynamic host-context integration: Unlike static reference-based tools, CRISPR-viva introduces an automated, context-aware filtering pipeline that meticulously accounts for individual genomic variations and tissue-specific transcriptomes. This ensures the generation of strict off-target-free gRNAs, effectively preventing host-cell toxicity and diagnostic false-positives.

To validate the generalizability and robustness of our foundation model, we systematically evaluated CRISPR-viva across 8 different CRISPR systems for both detection (e.g., LbuCas13a) and inhibition (e.g., Cas13d) tasks. Our computational benchmarks and wet experiments underscore the critical necessity of incorporating host transcriptomic contexts to mitigate off-target toxicity. Finally, to demonstrate the scalable inference capability of our foundation model for field-deployable applications, we deployed CRISPR-viva to process over 200 segmented genomes of RNA viruses across 23 human cell types, yielding a comprehensive profile of over 300 million evaluated gRNA candidates. Overall, CRISPR-viva shifts the paradigm of CRISPR guide design for virus detection and inhibition from data-hungry, task-specific algorithms to a highly adaptable foundation model, providing an agile, universal, and highly precise computational infrastructure for combatting emerging RNA viral threats.

Results

1. The framework of CRISPR-viva

CRISPR-viva is designed as a universal, foundation model-driven framework to generate optimized guide RNA (gRNA) candidates for highly sensitive viral detection and precise viral inhibition. Crucially, the framework operates in a host cell context-aware manner, meaning it dynamically incorporates individual genetic polymorphisms and tissue-specific transcriptomes to minimize off-target toxicity. The CRISPR-viva framework consists of two core components (Figure 1a): the off-target-free module and the foundation model module. The pipeline operates sequentially: first, the off-target-free module generates gRNA candidates from the viral genome and rigorously filters out any sequences with potential off-target homology to the personalized host transcriptome; subsequently, the foundation model module evaluates the surviving context-safe candidates to predict their on-target RNA-cleavage efficacy. Through this streamlined pipeline, an optimized package of highly effective and safe gRNAs can be obtained for diverse RNA viruses across different CRISPR systems.

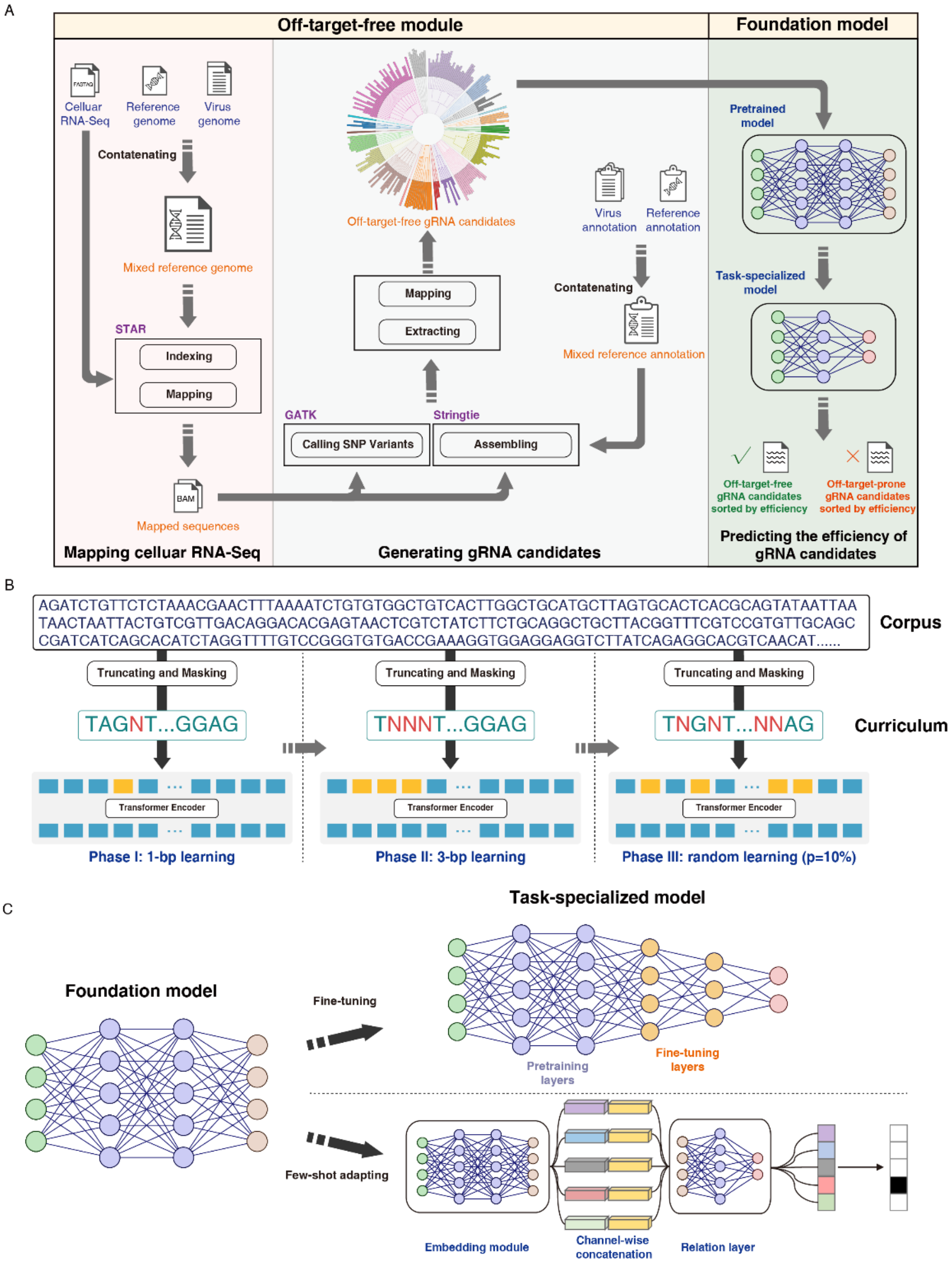

Figure 1. The framework of CRISPR-viva. a. The overall structure of CRISPR-viva, including the off-target-free module and the foundation model; b. Encoding and attention mechanism of the CRISPR-viva foundation model; c. Foundation model and downstream models utilizing two specialized strategies. The foundation model is shown in the left, the fine-tuning-based downstream model is shown in the upper right, and the few-shot adaptation-based downstream model is shown in the lower right.

1.1. The off-target-free module

The primary function of the off-target-free module is to ensure the clinical safety and diagnostic specificity of the designed gRNAs by filtering out candidates prone to cross-react with the host endogenous transcriptome (Figure 1a).

For the virus detection task, the module searches for gRNAs that perfectly match the target viral RNA but exhibit mismatches with the host endogenous RNAs beyond an adjustable mismatch threshold. The default threshold is set to 6 mismatches, a conservatively strict parameter informed by previous studies on Cas13 mismatch tolerance, ensuring the rigorous prevention of false-positive diagnostic signals or target-activated collateral cleavage. The detection pipeline proceeds through a logical workflow: (1) viral RNA sequence input and naive gRNA candidate generation; (2) dynamic host transcriptome acquisition (integrating specific host cell RNA-seq data to account for tissue-specific isoforms and individual SNPs); and (3) candidate stamping, which computationally eliminates any gRNA with potentially dangerous host homology (see Methods for detailed sequence mapping parameters).

For the virus inhibition task, the operational assumption shifts since the target virus has already infected the host, meaning the treated cell expresses both host endogenous RNA and the genomic RNA of the target virus. Thus, the target viral sequences are obtained from public repositories (e.g., the NCBI Virus Variation Resource or GISAID) and merged with the host genome. The inhibition pipeline comprises four steps: (1) reference genome concatenation (host + virus); (2) personalized transcriptome assembly; (3) host-aware gRNA candidate generation; and (4) strict candidate stamping to eliminate off-target-prone gRNAs that could otherwise trigger lethal collateral degradation of the host transcriptome (see Methods).

1.2. The foundation model module

The foundation model module serves as the core predictive engine of CRISPR-viva. Instead of being trained from scratch on limited CRISPR screening data, CRISPR-viva utilizes a Large Language Model (LLM) architecture based on self-supervised masked language modelling. The objective of this pre-training is to capture the intrinsic sequence syntax and local nucleotide interactions (e.g., secondary structure motifs and sequence dependencies) inherent in natural RNA molecules. To pre-train the foundation model (Figure 1b), we constructed a massive nucleotide sequence corpus—named CRISPRviva-3B—comprising three billion nucleotides generated by extracting 50-bp sequence windows from the genomes and transcriptomes of 23 human cell lines and 264 segmented genomes of RNA viruses (see Methods). We applied a curriculum learning strategy (progressing from 1-bp masking to 10% random masking tasks) to stabilize the representation learning. By learning these universal representations from raw RNA sequences rather than specific CRISPR editing events, the foundation model establishes a highly informative prior.

This pre-trained prior significantly accelerates the learning of downstream, CRISPR-specific functional tasks. We applied the pre-trained foundation model to downstream efficacy prediction tasks for 9 different CRISPR systems (Cas12a[5], Cas12b[17], Cas12c[18], LwCas13a[7], LbuCas13a[9], Cas13b[19], Cas13d[8], Cas14[16] for detection; LwCas13a, Cas13d, and Cas7-11 for inhibition). Depending on the availability of experimental screening data for each Cas effector, we adopted two specialized training strategies (Figure 1c). For medium-data scenarios (e.g., Cas12a and LwCas13a for detection, Cas13d for inhibition), we utilized a fine-tuning strategy to update the model weights. Conversely, for newly developed CRISPR systems or applications suffering from severe data scarcity (e.g., LbuCas13a, Cas14, or Cas7-11, where available labelled data $<10^3$), we deployed a few-shot adaptation strategy. This allows the foundation model to achieve robust generalizability and accurate efficacy predictions for novel CRISPR effectors using only a minimal number of experimental data points.

1.3. Scalable computational infrastructure for high-throughput inference

To accommodate the computational demands of integrating individual transcriptomes and evaluating massive candidate pools across numerous viral variants, CRISPR-viva was deployed on a high-performance, highly extendable computational infrastructure. This system is orchestrated via Kubernetes—an open-source container orchestration platform—allowing agile management across multiple servers for massive parallel processing. Furthermore, we integrated AIStore, a lightweight object storage system capable of linearly scaling out peta-scale nodes, ensuring that tens of terabytes of RNA sequencing data and foundation model parameters can be managed and retrieved efficiently during large-scale downstream inference tasks.

2. Integrating individual genetic polymorphisms and tissue-specific transcriptomes reveals off-target vulnerabilities

The issue of off-target effects during CRISPR-based RNA virus detection and inhibition can be decomposed into two subproblems: accurately depicting the dynamic host endogenous transcriptome, and selecting gRNA candidates with a sufficiently large edit distance from this transcriptome to prevent unintended binding. In clinical and field applications, relying on a single static human reference genome is fundamentally flawed. Off-target vulnerability is profoundly affected by two compounding biological factors: inter-individual genetic polymorphisms (e.g., population-level SNPs) and tissue-specific gene expression which dictates whether a mutated transcript is actually present to be erroneously targeted.

To illustrate the critical necessity of this personalized host context during gRNA design, we analysed RNA-seq data for 23 human cell lines from ENCODE[20] to perform genome-wide and transcriptome-wide SNP profiling. Crucially, as these cell lines are derived from diverse human donors, the observed sequence variations accurately represent the vast inter-individual genetic diversity present in the human population. We grouped these diverse donor cell lines into six distinct human organs or functional systems: the dermal, respiratory, circulatory, immune, liver, and kidney systems (Figure 2a, Supplementary Table 1). This grouping allowed us to evaluate patient-specific SNP profiles within the context of their actively transcribed, tissue-specific expression environments (see Methods).

By mapping the variants against the reference genome, we annotated the genome-wide SNP profiles for each unique donor cell line[21,22]. We particularly focused on the essential gene set, comprising 1,878 critical genes required for fundamental cellular survival[23]. The distribution, density, and mutation types (e.g., synonymous vs. non-synonymous) of these SNPs varied drastically across the different genetic backgrounds (Figure 2b). At the system level, while cell lines derived from closely related donor populations (such as specific immune or circulatory lineages) exhibited some similarities in SNP composition, the overall genetic landscape across the 23 cell lines was highly heterogeneous.

To quantify this off-target risk, we determined the absolute SNP counts in the whole genome and specifically within the exonic (transcribed) regions of essential genes (Figure 2c). At the whole-genome level, total SNP counts varied wildly among individuals, ranging from approximately 150,000 (e.g., in LAEpC and PTEpC) to massive mutational burdens exceeding 1,300,000 (e.g., in DFb, NEpC, and TAEnC). More alarmingly for CRISPR interventions, within the exonic regions of the 1,878 essential genes, the SNP counts ranged from 5,000 to 10,000 across different individuals.

This indicates that at least 63.9% of essential genes contained multiple patient-specific variant loci.

Collectively, these findings reveal a severe hidden risk in CRISPR guide design: (1) extensive inter-individual genetic variation (SNPs) exists in the population; (2) a large proportion of these SNPs reside in highly expressed essential genes; and (3) these expression patterns are highly tissue-specific. Consequently, if a gRNA is naively designed based solely on the standard reference genome, it might unknowingly form a perfect match with a patient's essential transcript bearing an unmapped SNP. In applications using Cas13, this off-target binding would activate the enzyme's non-specific collateral cleavage activity, leading to massive destruction of the host transcriptome and lethal cellular toxicity. Therefore, a one-size-fits-all reference genome approach is demonstrably inadequate, proving that the dynamic integration of individual genetic polymorphisms and tissue-specific transcriptomes is an absolute prerequisite for safely designing CRISPR gRNAs.

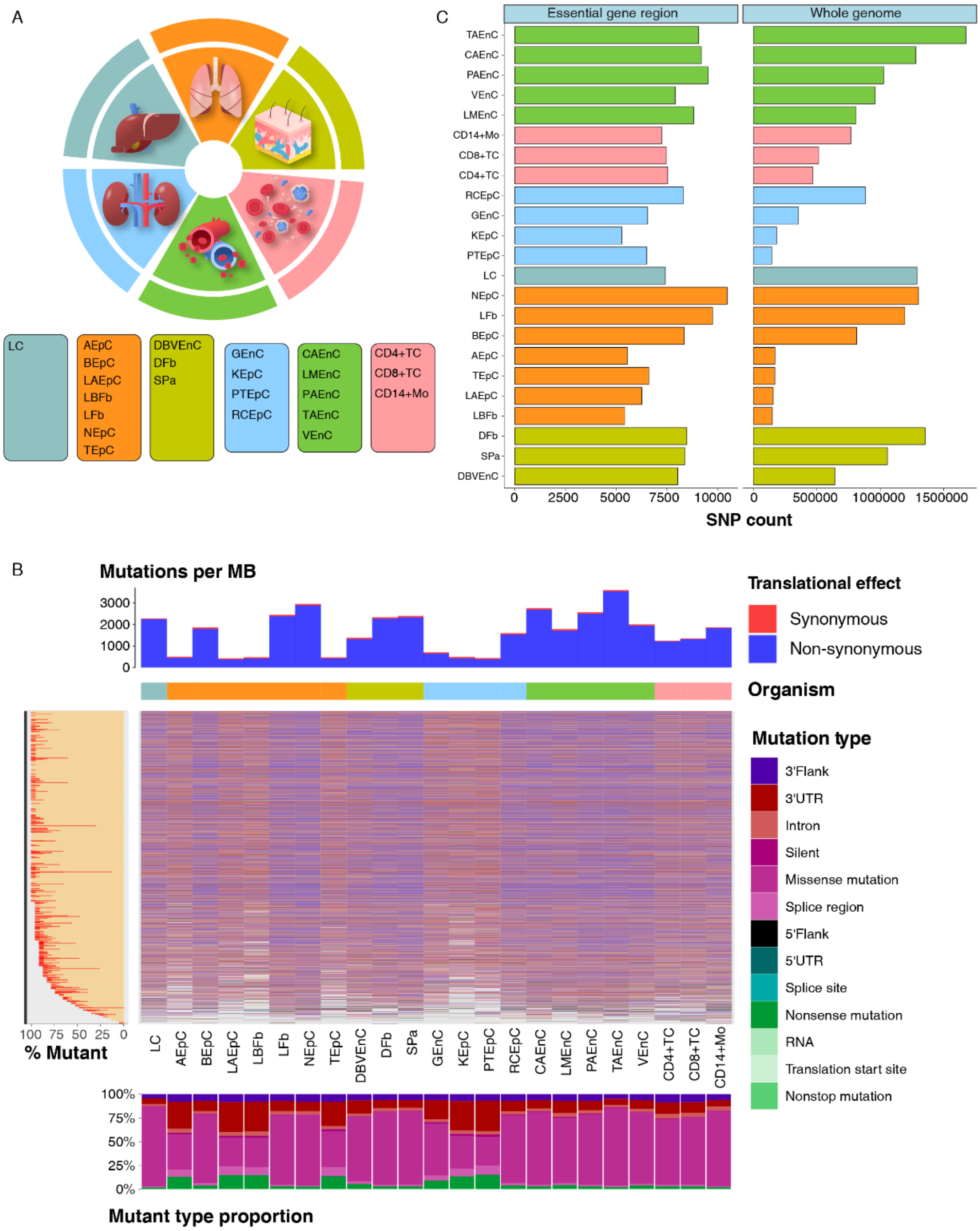

Figure 2. Profiling of inter-individual genetic polymorphisms (SNPs) across 23 human donor-derived cell lines from diverse tissue origins. a. Classification of the 23 distinct individual-derived cell lines, grouped by their six respective human organs or functional systems of origin (respiratory, dermal, immune, circulatory, liver, and kidney). This grouping captures both diverse genetic backgrounds and tissue-specific expression environments. b. Landscape of genome-wide SNP distributions and mutation patterns across the 23 cell lines. The varying mutation rates per megabase (MB) and mutant type proportions (e.g., synonymous vs. non-synonymous) highlight the extensive inter-individual genetic diversity, rather than intra-organismal somatic variation. c. Absolute SNP counts mapped within the exonic (transcribed) regions of 1,878 essential genes (left) and the whole genome (right) for each donor cell line.

3. CRISPR-viva outperformed existing RNA targeting efficacy prediction models with interpretability

Given the foundation model, we can establish downstream task-specialized models. We compared the RNA-targeting efficacy prediction ability of CRISPR-viva with that of well-established existing tools, including ADAPT[12], DeepCas13[13] and TIGER[14], in virus detection and inhibition tasks. In particular, the fine-tuned model was trained based on labelled data, whose related CRISPR system was LwCas13a for the detection task and Cas13d for the inhibition task. For the detection task, a set of 19,209 guide RNAs was compiled, of which 80% were used as the training dataset, while the remainder were used as the test dataset (see Methods)[12,14]. For the inhibition task, the training dataset contained 10,279 guide RNAs, and the test dataset contained 10,829 guide RNAs compiled from independent Cas13d screening experiments[13] (see Methods). The metrics for the performance comparison were the area under the receiver operating characteristic (ROC) curve (ROC-AUC) and the area under the precision-recall (PR) curve (PR-AUC). For the detection task, CRISPR-viva outperformed ADAPT, with a ROC-AUC of 0.90 and a PR-AUC of 0.98 (Figure 3a). For the inhibition task, CRISPR-viva outperformed DeepCas13 and a retrained version of TIGER, with a ROC-AUC of 0.91 and a PR-AUC of 0.76 (Figure 3b).

Furthermore, we analysed the interpretability of the foundation model. This evaluation considered the following three aspects: (1) the optimized sequence motif preferred by fine-tune-based task-specialized model of Cas12a and LwCas13a systems for detection task as well as Cas13d system for inhibition task (Figure 3c); (2) the locational sequence importance (Figure 3d); and (3) the attention map of the model, which characterizes the interaction of the input sequence with itself and its complementary sequence (Figure 3e). Each interpretation conveys a different level of the insight of the model (see Methods). The optimized sequence motif represents the sequence pattern most strongly preferred by the task-specialized model, meaning that the designed or selected guide RNA candidates that are similar to the optimized sequence motif are likely to have high efficacy for the specialized task. In contrast to the optimized sequence motif, which reflects the general preference of the task-specialized model, the locational sequence importance focuses on the precise input guide RNA candidates, representing the different influences of base pairs in each location. Finally, the attention map of the task-specialized layers of the task-specialized model represents the mechanism of the model, reflecting the base pair recognition process within the model.

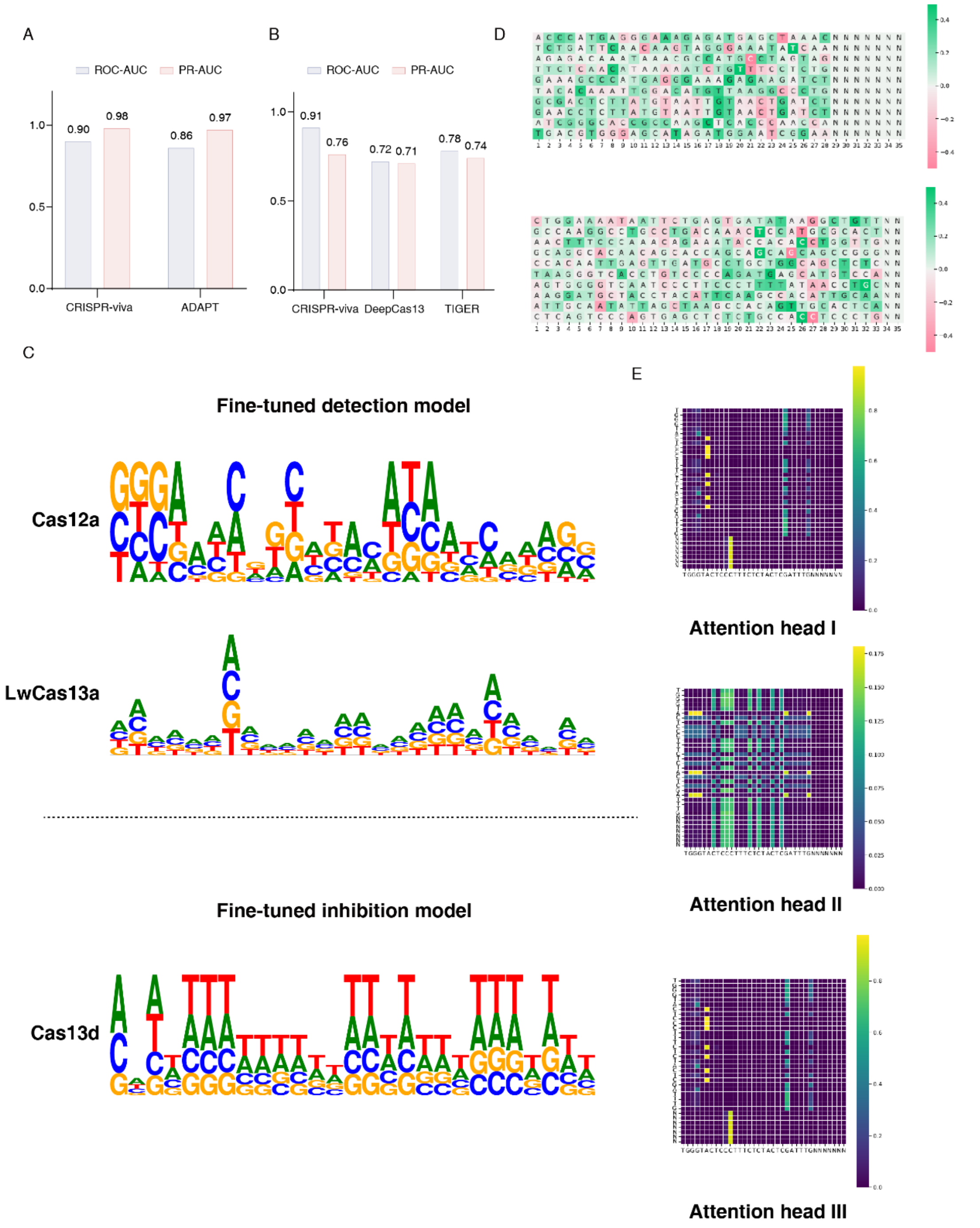

Figure 3. The benchmark of CRISPR-viva and its interpretability in terms of RNA target efficacy prediction. a. Comparison of LwCas13a-based detection efficacy prediction in the test dataset with the ROC-AUC and PR-AUC; b. Comparison of Cas13d-based inhibition efficacy prediction in the test dataset with the ROC-AUC and PR-AUC; c. Optimized sequence motif preferred by each fine-tuned-based task-specialized model for each CRISPR system; d. Locational sequence importance for the LwCas13a-based detection task model and the Cas13d-based inhibition task model.

4. CRISPR-viva enables precise detection of RNA viruses

4.1 Necessity of considering the host endogenous transcriptome in virus detection

There is significant sequence homology between pathogen and host transcriptomes, which increases the risk of off-target effects in guide RNA design. Incorrectly designed guide RNAs could bind to host RNA, leading to false-positive signals that compromise diagnostic accuracy (Figure 4a). To test this hypothesis, we designed 16 guide RNAs targeting Japanese encephalitis virus (JEV) for the CRISPR/LbuCas13a system using CRISPR-viva's off-target-free module (Figure 4b). We prepared genomic RNA from JEV-infected HEK293T cells, uninfected HEK293T cells, and PK-15 cells. Guide RNAs cr1–cr10 were predicted to be prone to off-targeting in HEK293T cells, with cr1–cr4 and cr8–cr10 exhibiting single-nucleotide mismatches, while cr5-cr7 perfectly matched endogenous transcripts. The remaining guide RNAs (cr11–cr16) had more than five mismatches, suggesting a low risk of cross-reactivity (Supplementary Table 4). None of the 16 guide RNAs were predicted to off-target in PK-15 cells.

We first assessed the cleavage activity of each guide RNA using in vitro transcribed (IVT) target RNA (10^{12} copies/µL). CRISPR-viva’s efficacy predictions showed a high accuracy of 0.8 for RNA-targeting efficacy (Figure 4g). Fluorescence intensities generated by each guide RNA varied significantly (Figure 4c, 4d). Notably, due to the limited data available for LbuCas13a, we applied a few-shot adaptation strategy to train the model (see Methods).

Next, we investigated whether off-target-prone guide RNAs could erroneously target human transcripts. We created two mixtures of guide RNAs (cr1–cr10 and cr11–cr16) and performed in vitro cleavage assays with genomic RNA from both infected and uninfected HEK293T cells. As expected, the cr1–cr10 mixture exhibited fluorescence in both infected and uninfected HEK293T samples, indicating off-target activity. In contrast, the cr11–cr16 mixture selectively detected JEV-infected HEK293T cells without fluorescence in uninfected HEK293T or PK-15 cells, confirming accurate targeting (Figure 4e, 4f, and 4h). To identify the specific guide RNAs responsible for off-target effects, we performed individual assays using HEK293T genomic RNA. Guide RNAs cr1–cr3, cr5, and cr7 showed significant fluorescence, confirming their off-target potential, while guide RNA cr6 produced only weak signals (Figure 4i).

Collectively, these findings highlight the critical importance of careful guide RNA design, as off-target effects can lead to false-positive results. Our results also demonstrate CRISPR-viva’s ability to generate off-target-free guide RNAs in low-data-volume scenarios, enabling accurate and reliable viral detection.

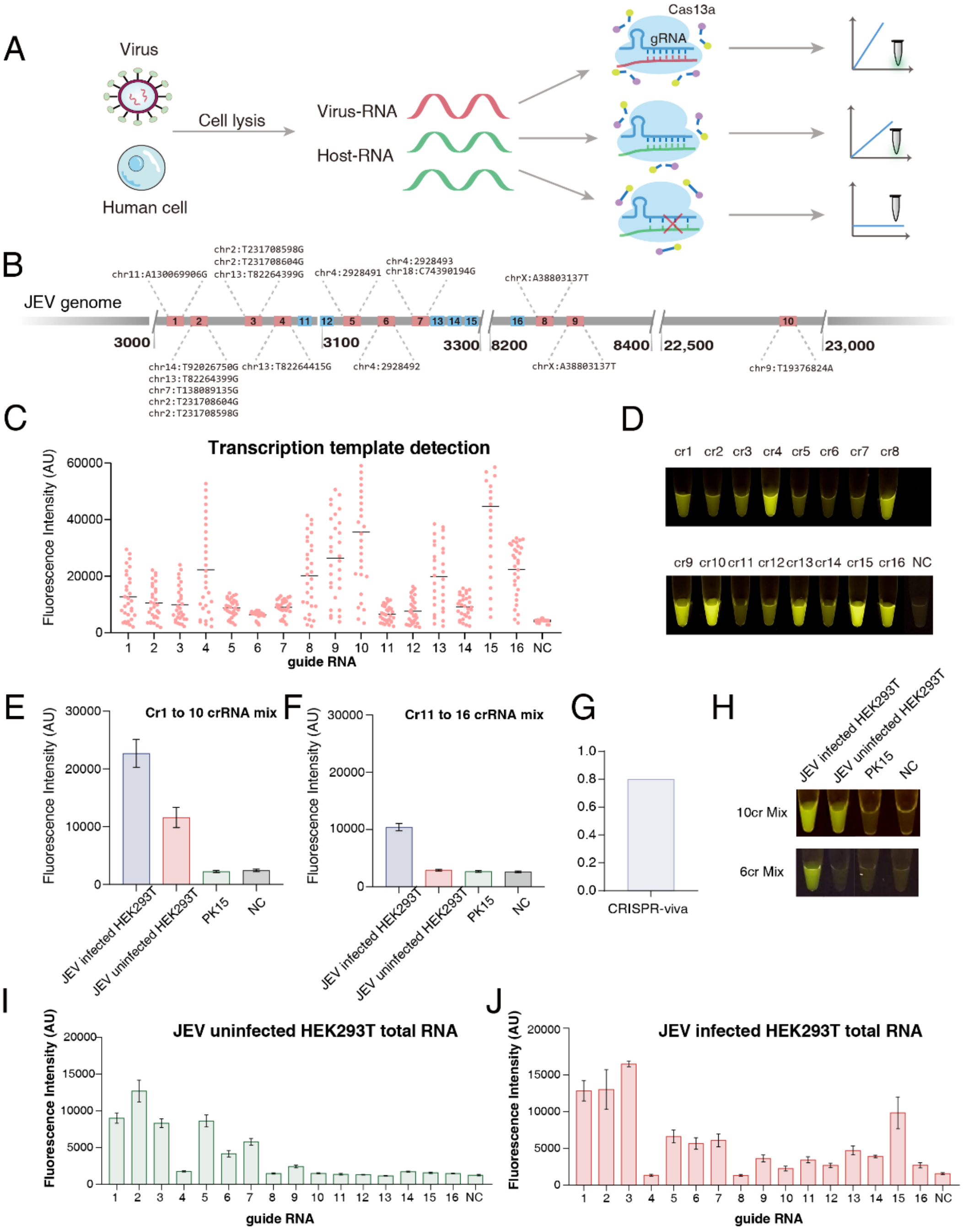

Figure 4. Nucleic acid detection of JEV using CRISPR/LbuCas13a and assessment of off-target effects in human cell genomes. a. Schematic representation of off-target activity in CRISPR/LbuCas13a-based detection of JEV within human cell genomes. b. Locations of the 16 guide RNAs on the JEV genome and their corresponding off-target sites in the human genome. c. Fluorescence signal values measured within first 30 minutes for each guide RNA when assayed with its corresponding IVT target RNA (10^{12} copies/μL). d. Fluorescence images captured at the 30th minute from the reaction in (c). e-f. Fluorescence signal values at the 30th minute for mixtures of cr1–cr10 (e) and cr11–cr16 (f) when assayed with different genomic RNA samples. g. Accuracy of the CRISPR-viva model in predicting the efficacy of the 16 guide RNAs. h. Fluorescence images captured at the 30th minute from reactions in (e) and (f). i-j. Fluorescence signal values at the 30th minute for each guide RNA when assayed with JEV-uninfected HEK293T cell genomic RNA (i)

and infected HEK293T cell genomic RNA (j). Data are presented as the means ± standard deviations (SDs) from three technical replicates.

4.2 CRISPR-viva is deployable for field-deployable virus detection by different CRISPR systems

We employed CRISPR-viva to provide the most comprehensive quickly deployable system for field-deployable virus detection, covering over 200 segmented human-hosted RNA virus curated by the National Center for Biotechnology Information (NCBI) Virus Variation Resource covering 20 families, 42 genus, 98 species and 200 segmented genome of RNA virus, relating to 23 cell types mentioned above utilizing 8 different available CRISPR systems, including three commonly adopted CRISPR systems and five newly emerged systems, namely, Cas12a, LwCas13a, and Cas13d, which are utilized by DETECTR[5], SHERLOCK[7] and SENSR[24], respectively, and the newly developed Cas12b, Cas12c, LbuCas13a, Cas13b and Cas14 systems, were selected for evaluation. We generated more than 300,000,000 evaluated guide RNAs in total and the profile is shown in Figure 5a. The statistical pattern observed for these viruses was highly similar: specifically, at the gene, CRISPR system, cell type, and viral variant levels, 40% to 60% of the guide RNA candidates were off-target-free (Figure 5a). These analysis results were integrated to formulate a comprehensive system freely available for field-deployable virus detection using available CRISPR system (https://www.crisprviva.top).

4.3 CRISPR-viva facilitates the precise detection of SARS-CoV-2 by different CRISPR systems

To further demonstrate the ability of CRISPR-viva to identify off-target-free guide RNA candidates for CRISPR-based virus detection assays, we utilized CRISPR-viva on 20 variants of currently circulating SARS-CoV-2, including the reference virus (GenBank ID: NC_045512.2) and its 19 major variants under monitoring and variants of concern designated by the Centers for Disease Control and Prevention (CDC; Supplementary Table 6). For each variant, we randomly selected 3 samples from the NCBI virus database to construct a dataset containing 58 viruses.

We generated guide RNA candidates for the 3 widely-used CRISPR systems for detection of the abovementioned 20 SARS-CoV-2 types in all 23 cell types listed in Figure 2a. Since the lengths of guide RNA candidates usually vary across CRISPR systems, a length of 20-24 bp for LbuCas13a, 22-30 bp for Cas13d and 22-30 bp for LwCas13a based on previous studies of these CRISPR systems[9,25,26]. Regarding the targeted viral genes, because the ORF1ab and N genes are utilized as indicators for nucleic acid amplification tests (NAATs) for SARS-CoV-2 worldwide because of their conserved nature, these genes were naturally selected to validate their conservation in the CRISPR-based virus detection assays described above. In addition, we included other essential genes, namely, the S, M and E genes, in the list of targeted genes to comprehensively analyse the guide RNA candidates for the viruses. To satisfy the targeting specificity for the practical implementation of the CRISPR-based virus detection assays, we set 7 mismatches as the safe edit distance, meaning that if a guide RNA candidate has no more than 6 mismatches with the transcriptome of the targeted cell, it is considered an off-target-prone candidate for the virus detection assay; otherwise, it is applicable for this assay.

For each of the 23 cell types, an average of 244170.45, 81618.68 and 195861.05 guide RNA candidates were acquired for the Cas13d, LbuCas13a and LwCas13a CRISPR systems, in terms of each SARS-CoV-2 variant respectively. The three SARS-CoV-2 genes among all the SARS-CoV-2 variants were ORF1ab, N, S, M and E. Among the acquired guide RNA candidates, 127076.32 (52%), 3217.952 (3.9%) and 107714.12 (55%) were off-target-free for Cas13d, LbuCas13a and LwCas13a, respectively (Figure 5b). With a maximum percentage of off-target-free guide RNA candidates across all SARS-CoV-2 variants of 55.04% and a minimum percentage of 51.1% across the 3 CRISPR systems, this statistic was also relatively robust across the 23 cell types (Supplementary Figure 7). Collectively, these results suggest that a great proportion of the guide RNA candidates were off-target-prone (in LbuCas13a case, the number is over 95%), considering the different CRISPR systems, SARS-CoV-2 genes, SARS-CoV-2 variants and human cell types, whereas the quantitative difference in guide RNA candidates is generally due to genomic sequence diversity.

For the off-target-prone guide RNA candidates, a longer viral gene or shorter guide RNA would result in more off-target-prone candidates (Figure 5c), since a longer viral gene (the longest gene, ORF1ab, contains approximately 21,280 bp, in contrast to the shortest gene E, with 228 bp) always has a larger guide RNA candidate pool, while a shorter guide RNA not only results in more candidates but also is more likely to be within the edit distance, causing specificity issues.

Additionally, for the off-target-free guide RNA candidates, we further investigated their profile attributes and application modes for advanced utilization. Similar to the off-target-prone candidates, longer viral genes have more guide RNA candidates because of the larger candidate pool. However, regarding the guide RNA length, a longer guide RNA length results in more off-target-free guide RNA candidates, as fewer candidates are filtered despite the larger candidate pool (Figure 5d).

Finally, considering gene conservation, we present two guide RNA selection strategies, i.e., *broad-spectrum detection* and *narrow-spectrum detection*, to utilize off-target-free guide RNA candidate datasets produced by the detection function of CRISPR-viva for SARS-CoV-2 and its variants.

For broad-spectrum detection, the aim is to identify guide RNA candidates targeting conserved viral genes, i.e., ORF1ab and N for SARS-CoV-2, which are shared by various variants of SARS-CoV-2 and used as general virus marker guide RNAs, to be utilized in SARS-CoV-2 assays for public service. This strategy exploits the advantages, i.e., the convenience and sensitivity, of CRISPR-based virus detection assays and, to a large extent, decreases the false-positive rate resulting from the specificity issue of CRISPR systems. Therefore, we aimed to develop an off-target-free guide RNA candidate dataset shared by all 23 human cell types and 20 SARS-CoV-2 variants for the ORF1ab and N genes (for the full dataset, see Data availability). Considering the total calculated number of shared off-target-free guide RNA candidates, the proportion and number of candidates for the ORF1ab gene were greater than those for the N gene because of its conservation and length. Regarding CRISPR systems, systems with shorter guide RNA candidates clearly had fewer guide RNA candidates than systems with longer ones, but the proportion of guide RNA candidates was not vastly different between systems, since different systems usually affect the guide RNA candidate pool rather than the conservativeness conservation of their targeting regions (Figure 5e). However, the distribution of the guide RNA length of the shared off-target-free guide RNA candidates was also determined (Figure 5f). In general, a greater proportion of guide RNAs are relatively long, since massive numbers of shorter guide RNAs are filtered because of specificity.

In contrast to broad-spectrum detection, narrow-spectrum detection has a different aim, i.e., authenticating the subtypes or variants of a given virus. For example, for SARS-CoV-2, we generated off-target-free guide RNA candidates for the given variants (for the full dataset, see Data availability). We aimed to acquire SARS-CoV-2 variant marker guide RNAs for the 3 abovementioned CRISPR systems (Figure 5g). For LwCas13a and Cas13d systems, 17 variants were covered, while 13 variants were covered for LbuCas13a. Investigation of the gene distribution revealed the abundance of guide RNAs targeting ORF1ab, since ORF1ab is much longer than the other four target genes, which allows a much larger guide RNA candidate pool. However, contrasting results were observed, for example, for Omicron BA.5, for which the marker guide RNAs were derived only from the M gene. Regarding the distribution of the guide RNA length, longer guide RNAs accounted for a larger proportion, as observed for broad-spectrum detection.

Collectively, we present a comprehensive investigation of the detection of SARS-CoV-2 by different CRISPR systems, providing two different virus detection strategies and further demonstrating the utility of CRISPR-viva for host cell context-aware virus detection.

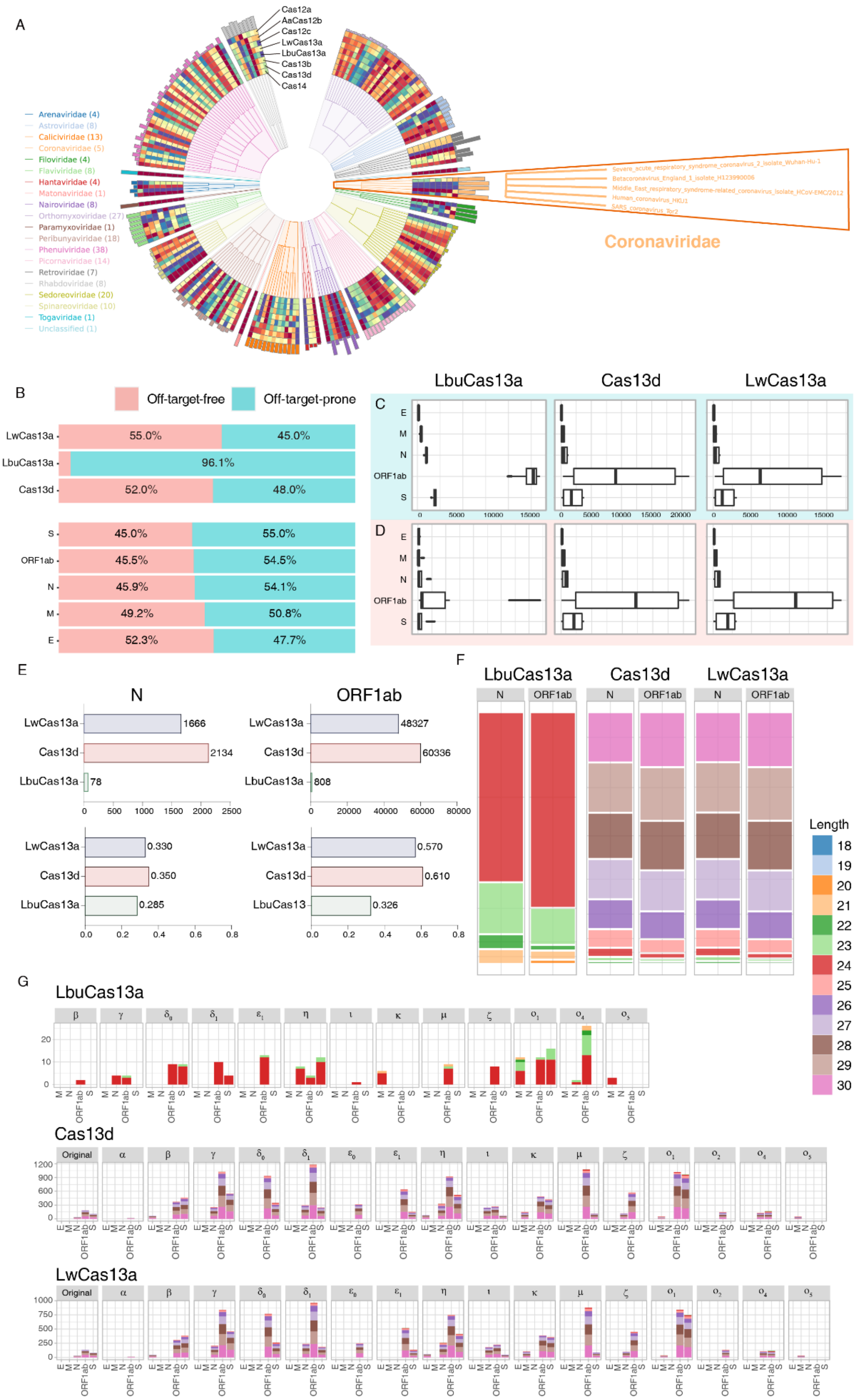

Figure 5. Application of CRISPR-viva for precise detection of SARS-CoV-2 by different CRISPR systems. a. Overall off-target-free guide RNA candidates' profile of 200 segmented human-hosted RNA virus with

number of segmented genomes of each viral family in the brackets; b. Overall ratio between off-target-free guide RNA candidates and off-target-prone candidates in terms of various CRISPR systems and SARS-CoV-2 viral genes; c. Off-target-prone guide RNA candidates' distribution among multiple SARS-CoV-2 variants with respect to each viral gene; d. Off-target-free guide RNA candidates' distribution among multiple SARS-CoV-2 variants with respect to each viral gene; e. Total number and ratio of shared off-target-free guide RNA candidates among multiple SARS-CoV-2 variants; f. Length distribution of shared off-target-free guide RNA candidates among multiple SARS-CoV-2 variants; g. Total number of unique off-target-free guide RNA candidates for each SARS-CoV-2 variants.

5. CRISPR-viva enables precise RNA virus inhibition

5.1 Necessity of considering the host endogenous transcriptome in virus inhibition

The use of guide RNA for viral inhibition may result in off-target effects that may disrupt the structural integrity of the endogenous transcriptome. To verify our hypothesis, we conducted a virus inhibition experiment with Cas13d system upon JEV (Figure 6a, b). The CRISPR-viva off-target-free module was employed to generate both off-target-free and off-target-prone guide RNA candidates for JEV and then the fine-tuned CRISPR-viva Cas13d inhibition model was used to predict the efficacy of each guide RNA candidate. We selected 5 off-target-prone guide RNAs (Supplementary Table 8) targeting three genes (SF23, SMIM13 and DST) to form 6 groups (Figure 6c): SF23 (targeting SF23), SM25, SM26, and SM27 (all three targeting SMIM13), DS24 (targeting DST), and NC (negative control with JEV and Cas13d but no guide RNA). After that, we assessed the Cas13d-mediated knock-down effect around the off-target loci in the endogenous RNA as well as on-target loci in the viral RNA through sequencing assay. We first measured the viral titer to confirm the existence of JEV in each group (Figure 6d) and attained the JEV expression level with the sequencing data (Figure 6e). Clearly, JEV exhibited varying levels of inhibition in the experimental groups and we further calculated the spearman correlation coefficient between this result with the efficacy predicted by CRISPR-viva, achieving a high value of 0.8 (Figure 6c). This result further validates the predictive ability of the CRISPR-viva model.

Moreover, significant RNA knock-down presented around the off-target loci in the endogenous RNA in all experimental groups(Figure 6f). This finding demonstrates that improperly designed guide RNAs can result in significant cleavage of the endogenous transcriptome, causing false knock-down, which severely impacts the safety usage of the CRISPR system.

Collectively, it is necessary to consider the host endogenous transcriptome in virus inhibition. CRISPR-viva can generate off-target-free guide RNA candidates and provide packages of high-efficacy guide RNAs for virus inhibition considering the host cell context.

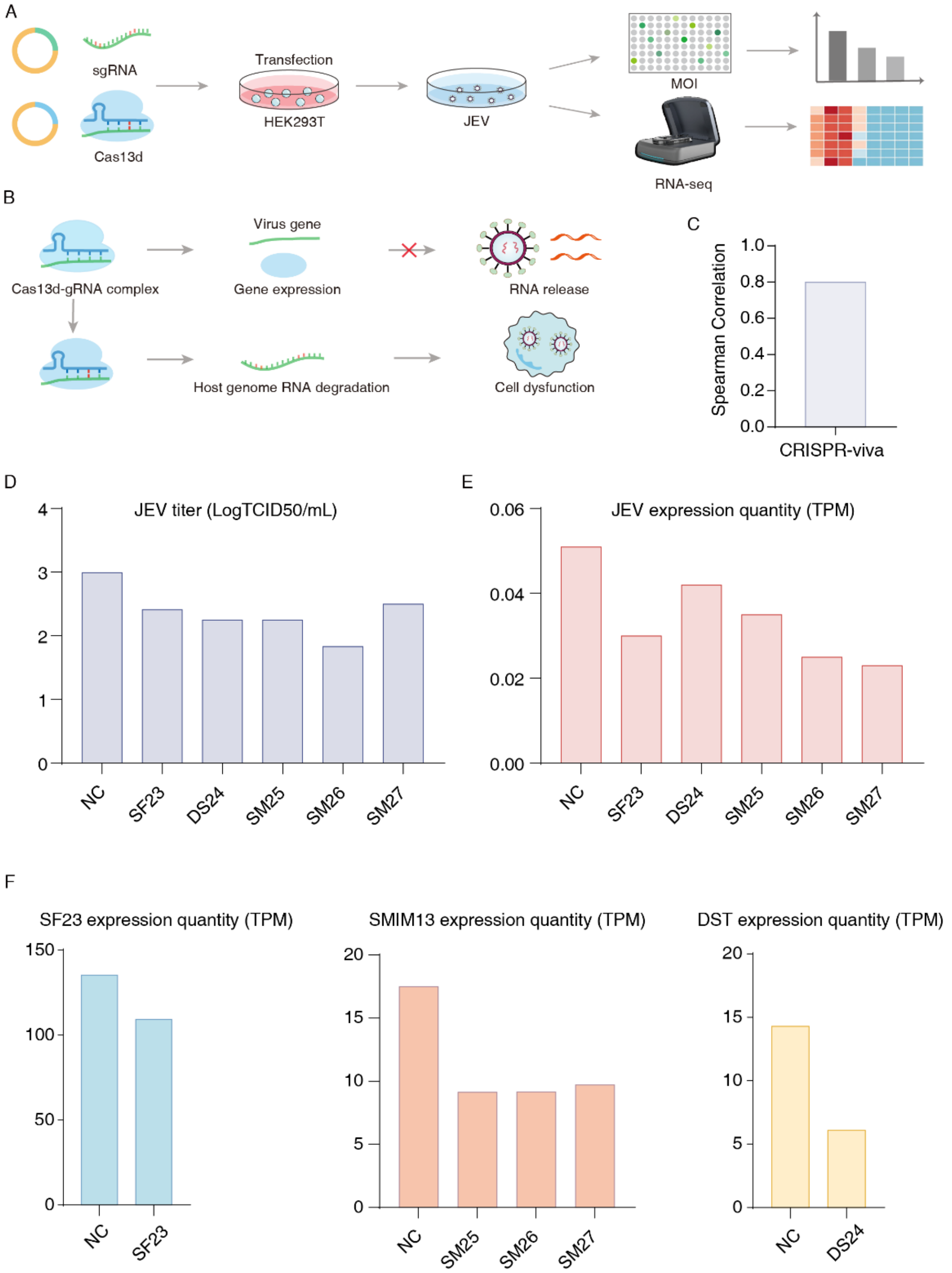

Figure 6. False cleavage events in the host endogenous transcriptome during the virus inhibition task with the Cas13d system. a. Schematic of Cas13d-based JEV nucleic acid inhibition; b. Schematic of Cas13d-based endogenous transcriptome cleavage activity; c. Spearman rank correlation of guide RNA efficacy predicted by CRISPR-viva; d. JEV titre in each group; e. JEV expression quantity (transcript per million, TPM) in each group; f. SF23/SMIM13/DST gene expression quantity (transcript per million, TPM) in control and experimental groups respectively.

5.2 CRISPR-viva facilitates precise inhibition of SARS-CoV-2 via different CRISPR systems

The virus inhibition task strongly requires off-target-free guide RNA candidates since the off-target effects are likely to disrupt physiological cellular functions. To demonstrate the inhibition function of CRISPR-viva, we again selected SARS-CoV-2 to demonstrate this application in three CRISPR systems, including Cas7-11, LwCas13a and Cas13d. For LwCas13a, we set guide RNA length ranging from 22-30nt with H as PFS. For Cas7-11 and Cas13d, we set guide RNA length ranging from 22-30nt with no PFS. We also performed similar analysis for MERS-Cov virus that can be assessed in supplementary materials (Figure S10).

We examined six datasets containing data for two types of cell samples, namely, Calu3 and NHBE cell samples, infected with SARS-CoV-2. These six datasets were generated by the tenOever laboratory of the Icahn School of Medicine at Mount Sina[27] (Supplementary Table 9). For each dataset, we utilized CRISPR-viva to obtain off-target-free guide RNA candidates. In general, regarding the availability of guide RNA candidates, the percentage of off-target-free guide RNA candidates ranged from 56% to 64% at the cell sample, CRISPR system and viral target gene levels (Figure 7a). The average numbers of off-target-free guide RNA candidates targeting the above six cell samples were 142562, 133959 and 113298 for the Cas7-11, Cas13d and LwCas13a systems, respectively, with corresponding coefficients of variation of 7.7%, 7.68% and 7.72% (Figure 7b). These values showed relatively minor differences among the six samples. Similar to the scenario for virus detection, this difference in the number of guide RNAs originates from the difference in PFS complexity among the CRISPR systems. Furthermore, we investigated the targeting distribution of the off-target-free guide RNA candidates at the gene level (Figure 7c, d). According to our results, the guide RNA candidates were able to target all five essential genes, including the SARS-CoV-like RNA-dependent RNA polymerase (RdRp) encoded within the ORF1ab region, in 5 of the samples. The exception was sample C1, in which the S and E RNAs are not in the transcriptome of these virus-infected cells; hence, there were no off-target-free guide RNA candidates targeting the S and E genes.

We further performed calculations using the reference genome of the host and the virus without considering the actual host transcriptome or viral genome. More than 6.3% of the guide RNA candidates were off-target-free, and 5% were off-target-prone, indicating the probability of designing erroneous guide RNA candidates that may either be off-target-prone to prevent cleavage activity or cleave the endogenous transcriptome, causing cellular dysfunction.

5.3 Comparison of CRISPR-viva with the PAC-MAN system for virus inhibition

We compared CRISPR-viva with the PAC-MAN system developed by Lei S. Qi et al.[4] to determine the necessity and feasibility of an optimized host-aware, off-target-free guide RNA design approach for virus inhibition.

In the PAC-MAN system, off-target-free guide RNA candidates are selected in three steps: (1) extraction of guide RNA candidates from conserved regions of the virus or virus family; (2) removal of candidates that have fewer than 3 mismatches with the reference human transcriptome; and (3) removal of candidates with poly-T (> 2 T) bases within the sequence, which prevents guide RNA expression. We utilized the dataset with 3,203 guide RNAs of PAC-MAN on the abovementioned SARS-CoV-2 instances. In the PAC-MAN guide RNA dataset, we filtered 2,758 guide RNAs targeting the ORF1ab, S, E, M and N genes of SARS-CoV-2 virus. We then calculated the differences in the sets of off-target-free and off-target-prone guide RNA candidates (Figure 7e, f). It is clear that PAC-MAN cannot precisely obtain off-target-free guide RNA candidates since it uses the reference human transcriptome and the conserved regions of the virus or virus family with public database, neglecting the sequence variation between the host cell and the reference genome as well as that between the virus genome in the public database and the genome of the infecting virus inside the host cell. The variation between these sequences is over 100 times (Figure 7c,f), which may result in erroneous cleavage of the endogenous transcriptome or shortage of optimized designed guide RNA candidate.

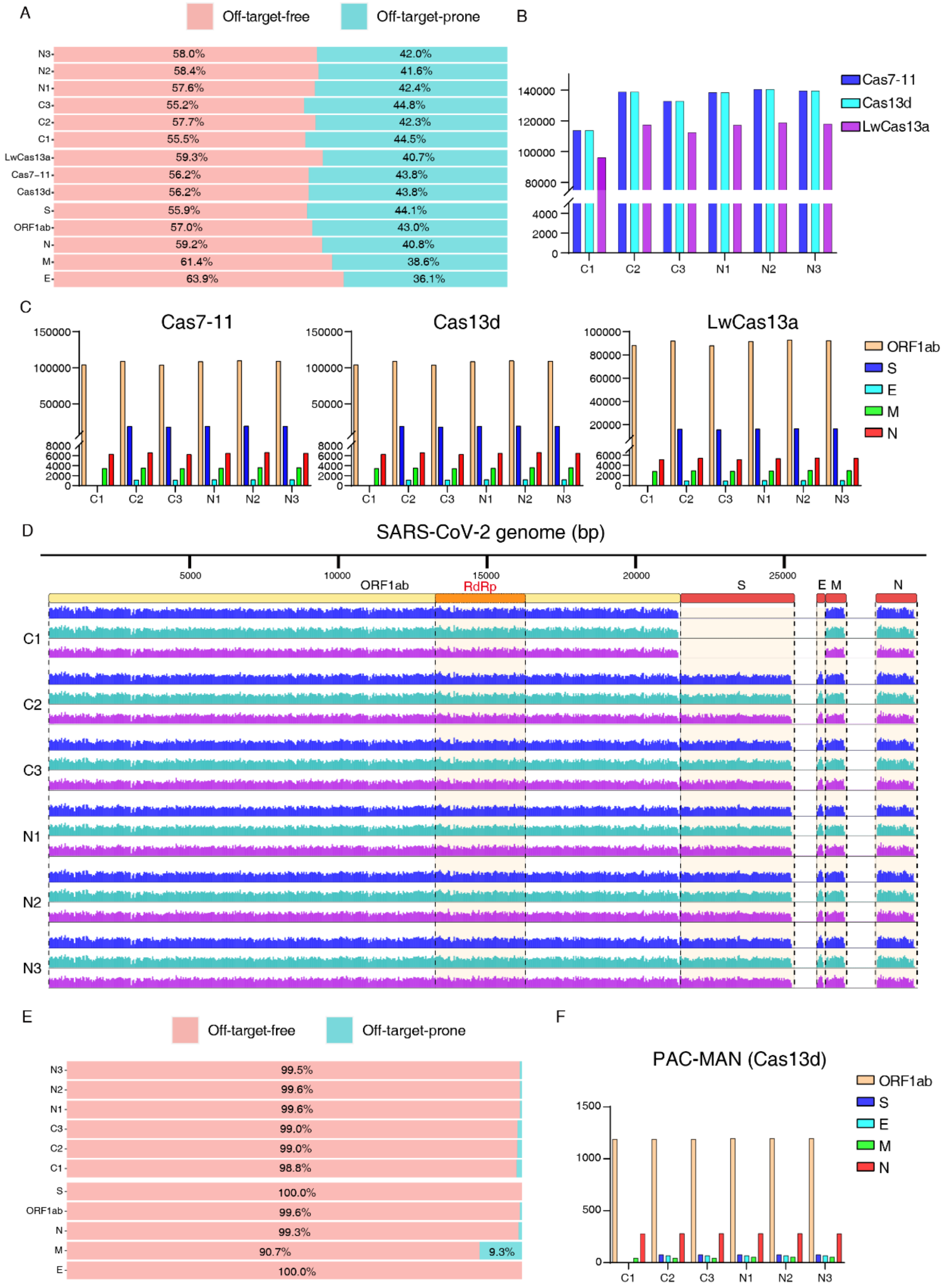

Figure 7. Application of CRISPR-viva for precise inhibition of SARS-CoV-2 by different CRISPR systems. a. Overall ratio between off-target-free guide RNA candidates and off-target-prone candidates in terms of various infected cell samples, CRISPR systems and SARS-CoV-2 viral genes; b. Total number of off-target-free guide RNA candidates in multiple infected cell samples in terms of various CRISPR systems; c. Total number of off-target-free guide RNA candidates of each SARS-CoV-2 viral genes in multiple infected cell samples in terms of various CRISPR systems; d. Genomic distribution of off-target -free guide RNA candidates in multiple infected cell samples in terms of various CRISPR systems; e. Overall ratio between off-target-free guide RNA candidates and off-target-prone candidates in terms of various infected cell samples

given PAC-MAN guide RNA set; f. Total number of off-target-free guide RNA candidates of each SARS-CoV-2 viral genes in multiple infected cell samples with PAC-MAN guide RNA set.

Discussion

In this study, we developed CRISPR-viva, a universal and dynamically host context-aware guide RNA design framework powered by a sequence-to-function foundation model. Unlike existing computational tools that are built from scratch for isolated Cas effectors and heavily rely on massive labelled screening datasets, CRISPR-viva fundamentally shifts the paradigm by conceptualizing gRNA design in virus detection and inhibition as a representation learning challenge. By learning the universal grammar and latent structural interactions of RNA sequences through self-supervised pre-training on a massive, unlabelled corpus (CRISPRviva-3B), our foundation model overcomes the data-scarcity bottleneck. It demonstrates remarkable few-shot adaptation capabilities, enabling the rapid and accurate prediction of RNA-targeting efficacy for newly discovered CRISPR systems (e.g., Cas7-11, Cas14) and newly emerged viral threats with minimal experimental data.

Beyond on-target efficacy, the safe clinical translation of CRISPR-based viral detection and inhibition relies strictly on the absolute avoidance of host off-target interactions. As our computational analyses and experimental validations underscore, relying on a static human reference genome is highly inadequate due to extensive inter-individual genetic polymorphisms (SNPs) and distinct tissue-specific transcriptomic expression profiles. This is especially critical for Cas13-based viral inhibition therapies. For effectors like Cas13, off-target binding to highly expressed endogenous transcripts can trigger target-activated, non-specific collateral cleavage, leading to catastrophic RNA degradation and severe cellular toxicity. Therefore, integrating individual genetic variations and tissue-specific expression levels into an automated filtering pipeline is not merely an optional enhancement, but an absolute prerequisite for ensuring therapeutic safety and diagnostic specificity.

In practical applications, we propose a hierarchical deployment strategy balancing rapid emergency response with precision medicine. In the event of a sudden viral outbreak where personalized host data is unavailable, researchers can leverage the pre-computed, universally "context-safe" gRNA profiles generated by CRISPR-viva to formulate rapid, field-deployable diagnostic kits. Conversely, for targeted antiviral therapeutics where cellular toxicity must be absolutely minimized, the framework allows for de novo gRNA design utilizing personalized patient RNA-seq data, fully customizing the CRISPR intervention to the patient's unique genetic background and the targeted tissue's specific expression profile.

Despite these advancements, we acknowledge several limitations in the current study that outline essential future directions. First, while our current off-target module employs a strict, conservative mismatch threshold (e.g., ≤ 6 mismatches) to robustly prevent target-activated Cas13 collateral activity, this hard-filtering approach could be further refined. Future iterations of CRISPR-viva will incorporate nuanced thermodynamic modelling and position-specific mismatch weighting (e.g., heavily penalizing seed-region mismatches) to provide a more granular, quantitative off-target scoring system. Second, real-world clinical samples (such as respiratory swabs or gut biopsies) often contain complex commensal microbiota, and human genetic diversity extends far beyond the individual donor cell lines evaluated here. Future updates to the CRISPR-viva infrastructure will integrate metagenomic sequence databases and large-scale population-level genetic cohorts (such as the 1000 Genomes Project) to comprehensively computationally shield against all potential sources of cross-reactivity. Finally, we emphasize that computational prediction, no matter how advanced, serves as a high-throughput triage step. The predicted optimal, off-target-free gRNA candidates must continue to undergo rigorous, standardized downstream in vitro and in vivo wet-lab validation to ensure their absolute efficacy and safety before clinical deployment.

Overall, CRISPR-viva provides a highly adaptable, universal, and computationally precise infrastructure, paving the way for the rapid development of personalized, safe, and highly effective CRISPR-based interventions against emerging RNA viral threats.

Author Contributions Statement

Qi Liu, Xinjie Wang and Guohui Chuai designed the framework of this work. Guohui Chuai, Chen Zhang, Qinchang Chen, Jiaying Wen, Jiaxing Han, Kejing Dong performed the analyses. Xinjie Wang, Xingxu Huang and Lina Wu designed the experimental validations and Jieyu Cui performed the related experiments. Guohui Chuai, Xinjie Wang and Qi Liu wrote the manuscript with the help of other authors. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (Grant No. 2025YFC3409300), National Natural Science Foundation of China (Grant No. T2425019, 32341008, 32550745, 82521002, 82541010, U25D9023), New Generation Artificial Intelligence-National Science and Technology Major Project (Grant No. 2025ZD0122802), Shanghai Pilot Program for Basic Research, Shanghai Science and Technology Innovation Action Plan-Key Specialization in Computational Biology, Shanghai Shuguang Scholars Project, Shanghai Excellent Academic Leader Project, Shanghai Municipal Science and Technology Major Project (Grant No. 2021SHZDZX0100), Key Program of Guangzhou National Laboratory (EKPG21-18), Fundamental Research Funds for the Central Universities, and Tongji University “Medicine + X” Cross Research Program [2025080107].

Data availability

CRISPRviva-3B, the corpus to train the CRISPR-viva foundation model, is available on the Huggingface Dataset Hub at https://huggingface.co/datasets/bm2-lab/CRISPRviva-3B. To facilitate the field-deployable amplification-free virus detection, we have released a database with generated guide RNA candidates at https://www.crisprviva.top. The sequence data of Cas13d-JEV inhibition experiments in this study have been deposited in Sequence Read Archive (SRA) of NCBI under the accession number PRJNA1174202 at https://www.ncbi.nlm.nih.gov/sra/PRJNA1174202.

Code availability

The pretrained CRISPR-viva foundation model and fine-tuned downstream models are available on the Huggingface Model Hub at https://huggingface.co/bm2-lab/CRISPR-viva. The code of off-target-free module in CRISRP-viva is available on the Github at https://github.com/bm2-lab/CRISPR-viva.

Reference

1. Chan-Yeung, M., and Xu, R.H. (2003). SARS: epidemiology. Respirology *8*, S9-S14.
2. Zumla, A., Hui, D.S., and Perlman, S. (2015). Middle East respiratory syndrome. The Lancet *386*, 995-1007.
3. Gorbalenya, A.E., Baker, S.C., Baric, R.S., de Groot, R.J., Drosten, C., Gulyaeva, A.A., Haagmans, B.L., Lauber, C., Leontovich, A.M., and Neuman, B.W. (2020). The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nature microbiology *5*, 536-544.
4. Abbott, T.R., Dhamdhere, G., Liu, Y., Lin, X., Goudy, L., Zeng, L., Chemparathy, A., Chmura, S., Heaton, N.S., and Debs, R. (2020). Development of CRISPR as an antiviral strategy to combat SARS-CoV-2 and influenza. Cell *181*, 865-876. e812.
5. Chen, J.S., Ma, E., Harrington, L.B., Da Costa, M., Tian, X., Palefsky, J.M., and Doudna, J.A. (2018). CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. Science *360*, 436-439.
6. Freije, C.A., and Sabeti, P.C. (2021). Detect and destroy: CRISPR-based technologies for the response against viruses. Cell host & microbe *29*, 689-703.
7. Kellner, M.J., Koob, J.G., Gootenberg, J.S., Abudayyeh, O.O., and Zhang, F. (2019). SHERLOCK: nucleic acid detection with CRISPR nucleases. Nature protocols *14*, 2986-3012. 10.1038/s41596-019-0210-2.
8. Rauch, J.N., Valois, E., Solley, S.C., Braig, F., Lach, R.S., Audouard, M., Ponce-Rojas, J.C., Costello, M.S., Baxter, N.J., and Kosik, K.S. (2021). A scalable, easy-to-deploy protocol for Cas13-based detection of SARS-CoV-2 genetic material. Journal of clinical microbiology *59*, e02402-02420.
9. Fozouni, P., Son, S., de León Derby, M.D., Knott, G.J., Gray, C.N., D’Ambrosio, M.V., Zhao, C., Switz, N.A., Kumar, G.R., and Stephens, S.I. (2021). Amplification-free detection of SARS-CoV-2 with CRISPR-Cas13a and mobile phone microscopy. Cell *184*, 323-333. e329.
10. Safari, F., Afarid, M., Rastegari, B., Borhani-Haghighi, A., Barekati-Mowahed, M., and Behzad-Behbahani, A. (2021). CRISPR systems: Novel approaches for detection and combating COVID-19. Virus Research *294*, 198282.
11. Tong, H., Huang, J., Xiao, Q., He, B., Dong, X., Liu, Y., Yang, X., Han, D., Wang, Z., and Wang, X. (2023). High-fidelity Cas13 variants for targeted RNA degradation with minimal collateral effects. Nature biotechnology *41*, 108-119.
12. Metsky, H.C., Welch, N.L., Pillai, P.P., Haradhvala, N.J., Rumker, L., Mantena, S., Zhang, Y.B., Yang, D.K., Ackerman, C.M., and Weller, J. (2022). Designing sensitive viral diagnostics with machine learning. Nature biotechnology *40*, 1123-1131.
13. Cheng, X., Li, Z., Shan, R., Li, Z., Wang, S., Zhao, W., Zhang, H., Chao, L., Peng, J., and Fei, T. (2023). Modeling CRISPR-Cas13d on-target and off-target effects using machine learning approaches. Nature Communications *14*, 752.
14. Wessels, H.-H., Stirn, A., Méndez-Mancilla, A., Kim, E.J., Hart, S.K., Knowles, D.A., and Sanjana, N.E. (2024). Prediction of on-target and off-target activity of CRISPR–Cas13d guide RNAs using deep learning. Nature Biotechnology *42*, 628-637.
15. Özcan, A., Krajeski, R., Ioannidi, E., Lee, B., Gardner, A., Makarova, K.S., Koonin, E.V.,

Abudayyeh, O.O., and Gootenberg, J.S. (2021). Programmable RNA targeting with the single-protein CRISPR effector Cas7-11. Nature *597*, 720-725.

16. Hu, J., Zhou, J., Liu, R., and Lv, Y. (2021). Element probe based CRISPR/Cas14 bioassay for non-nucleic-acid targets. Chemical Communications *57*, 10423-10426.

17. Tong, X., Zhang, K., Han, Y., Li, T., Duan, M., Ji, R., Wang, X., Zhou, X., Zhang, Y., and Yin, H. (2024). Fast and sensitive CRISPR detection by minimized interference of target amplification. Nature Chemical Biology, 1-9.

18. Wang, Z., and Zhong, C. (2021). Cas12c-DETECTOR: A specific and sensitive Cas12c-based DNA detection platform. International Journal of Biological Macromolecules *193*, 441-449.

19. Gootenberg, J.S., Abudayyeh, O.O., Kellner, M.J., Joung, J., Collins, J.J., and Zhang, F. (2018). Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a, and Csm6. Science *360*, 439-444.

20. Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57.

21. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., and Daly, M. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research *20*, 1297-1303.

22. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The ensembl variant effect predictor. Genome biology *17*, 1-14.

23. Wang, T., Birsoy, K., Hughes, N.W., Krupczak, K.M., Post, Y., Wei, J.J., Lander, E.S., and Sabatini, D.M. (2015). Identification and characterization of essential genes in the human genome. Science *350*, 1096-1101.

24. Brogan, D.J., Chaverra-Rodriguez, D., Lin, C.P., Smidler, A.L., Yang, T., Alcantara, L.M., Antoshechkin, I., Liu, J., Raban, R.R., Belda-Ferre, P., et al. (2021). Development of a Rapid and Sensitive CasRx-Based Diagnostic Assay for SARS-CoV-2. ACS Sensors *6*, 3957-3966. 10.1021/acssensors.1c01088.

25. Zhang, C., Konermann, S., Brideau, N.J., Lotfy, P., Wu, X., Novick, S.J., Strutzenberg, T.S., Griffin, P.R., Hsu, P., and Lyumkis, D. (2018). Structural Basis for the RNA-Guided Ribonuclease Activity of CRISPR-Cas13d. Cell *175*, 212-223.

26. Abudayyeh, O.O., Gootenberg, J.S., Konermann, S., Joung, J., Slaymaker, I.M., Cox, D.B., Shmakov, S., Makarova, K.S., Semenova, E., and Minakhin, L. (2016). C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. Science *353*, aaf5573.

27. Blanco-Melo, D., Nilsson-Payant, B.E., Liu, W.-C., Uhl, S., Hoagland, D., Møller, R., Jordan, T.X., Oishi, K., Panis, M., and Sachs, D. (2020). Imbalanced host response to SARS-CoV-2 drives development of COVID-19. Cell *181*, 1036-1045. e1039.

Methods

1. Cellular RNA-seq data acquisition and preparation

Our cellular RNA-seq data for 23 cell lines were acquired from ENCODE[1]. As shown in Figure 2a, the cell lines were derived from 6 types of tissues or organs, namely, the respiratory system, skin, immune system, circulatory system, liver and kidney. Supplementary Table 1 lists the details of these cell lines. We downloaded the raw sequencing data for these cell lines in fastq or fastq.gz format for further analysis.

2. Virus genome data acquisition and preparation

The virus genome data in this research were acquired mainly from the (NCBI virus database, which is a comprehensive portal for acquisition of viral sequence data from RefSeq, GenBank and other NCBI repositories. We firstly acquired 20 families, 42 genus, 98 species and 200 segmented genome of RNA virus (Supplementary Figure 1, Supplementary Table 1, 2, 3). We further investigated six types of viruses in our research, namely, SARS-CoV-2, dengue, Zika, Ebola, MERS-CoV and SARS-CoV-1. For SARS-CoV-2, which is taken as the most important demonstrative example in our study, we collected 20 variants, including one reference strain (GenBank ID: NC_045512.2) and its 11 variants labelled by the World Health Organization. Moreover, the main subtypes of each of the 11 labelled variants were obtained, and we obtained the viral genomes of three biological samples for each subtype, where the subtype is defined in terms of the Phylogenetic Assignment of Named Global Outbreak (PANGO) lineage. Therefore, the total number of SARS-CoV-2 viral genomes was 58, which is listed in Supplementary Table 6.

For the other 5 viruses, namely, dengue, Zika, Ebola, MERS-CoV and SARS-CoV-1, we also selected the main subtype or variant of the virus. Detailed information on the viruses is listed in Supplementary Table 7.

For all the abovementioned virus genome data, we downloaded two types of files: a fasta file for viral genomic sequences and a gff3 file for viral genomic annotations. The gff3 annotation file was then input into gffread utility in the GTF utilities software package to be converted to a gtf format annotation file for further analysis[2].

3. The framework of CRISPR-viva

CRISPR-viva consists of two modules: (1) the off-target-free module, and (2) the foundation model module. With the guide RNA candidates generated via the screening process of both modules, a package of high-efficacy guide RNAs for virus detection and inhibition by any CRISPR system can be obtained.

3.1 The off-target-free modules

We formulated two functional modules: one for virus detection and one for virus inhibition. The aim of the off-target-free functional modules is to filter out potential off-target guide RNA candidates targeting the endogenous transcriptome.

The function of the off-target-free modules is to search for guide RNAs that are perfect matches for the target virus RNA but are not matches for the endogenous RNAs in the host cell within a six-mismatch range to prevent cleavage through off-target effects. There are four steps in the process: (1) mapping cellular RNA-seq data; (2) preparing the reference transcriptome; (3) generating the guide RNA candidates; and (4) Stamping the guide RNA candidates. The technical details are described below.

Mapping cellular RNA-seq data

In this step, the inputs are cellular RNA-seq fastq files and a cellular reference genome (for the detection scenario) or a mixed reference genome (for the inhibition scenario) composed of a cellular reference genome and a virus genome. The original or mixed reference genome is used to construct an index with STAR[3] for the subsequent mapping process, in which the default parameters are set. Using the constructed index, the mapping function of STAR is executed with the compressed cellular RNA-seq fastq files as input; the essential parameters are listed below. The results of cellular RNA-seq data mapping are bam files.

STAR parameter	**Value**
--twopassMode	Basic
--outSAMtype	BAM SortedByCoordinate
--readFilesCommand	zcat

Preparing the reference transcriptome

This step has two branches: (1) Calling of SNP variants to individualize the reference genome formulated via cellular RNA-seq data mapping. In this branch, the bam files are input into the GATK[4] pipeline, which is composed of modules such as AddOrReplaceReadGroups, MarkDuplicates, SplitNCigarReads, HaplotypeCaller and MergeVcfs. In this process, the AddOrReplaceReadGroups command is used to assign all the reads in the file to a single read group, and in our study, the parameter --RGSM or -SM is added to the name of the input cellular RNA-seq sample. The output of this command is a bam file. The output of the AddOrReplaceReadGroups command is then input into the MarkDuplicates command to locate and tag duplicate reads, which are generated during sample preparation, especially in the library construction phase, via PCR. The output of this command is a bam file. The SplitNCigarReads command uses the output of the MarkDuplicates command, which is in the form of a bam file, to split reads that contain Ns in their cigar strings, which is an essential step in the postprocessing of RNA reads aligned to the reference genome. Similar to the above procedure, the output of this command is a bam file. With the output of the SplitNCigarReads command, we use the HaplotypeCaller command to identify the SNPs in the input cellular RNA-seq data compared to the reference genome, and the output of this step is in the form of a vcf file. Finally, since each cell line is associated with multiple fastq RNA-seq data files and every fastq RNA-seq data file is eventually converted into a vcf file that annotates the individualized SNP status of each biological sample of the cell line, we use the MergeVcfs command to combine all the vcf files generated from each biological sample of each cell line to represent the integrated SNP status of the cell line. With the output as integrated vcf files for each cell line, the reference genome is updated to the individualized version. (2) Assembly of the bam files generated in the cellular RNA-seq mapping step into transcripts. In this branch, we use StringTie[5] to assemble the bam files generated in the cellular RNA-seq mapping step into transcripts. The default parameters are used, and the output is in GTF format.

Generating guide RNA candidates

In this step, the output of the previous step is utilized to generate the sequences of guide RNA candidates and the transcriptome; specifically, branch 1 provides individualized genome sequences, and branch 2 provides the coordinates of the transcripts. For the detection scenario, the guide RNA candidates are extracted directly from a given viral genome, while for the inhibition scenario, they are selected from the generated transcripts mapped onto the viral regions of the mixed reference genome.

Stamping guide RNA candidates

In this step, Bowtie2[6] is utilized to stamp guide RNA candidates. First, the generated transcriptome is used as a reference to construct the Bowtie index with the default parameters. Second, the guide RNA candidates are mapped to the constructed index with the essential parameters listed below:

Bowtie2 parameters	**Value**
--score-min	C,-(1+6×number_of_mismatch),0
-L	10
--gbar	31

Finally, the guide RNA candidates are categorized into off-target-free and off-target-prone candidates based on the Bowtie mapping results, in which guide RNA candidates with no mapping results are stamped as off-target-free; otherwise, they are labelled as off-target-prone.

3.2 The foundation model module

The foundation model module of CRISPR-viva is an LLM that harnesses the recent development of the masked language modeling[7] to capture the interactions within the target location during CRISPR editing events, where each guide RNA target location is subject to influence from adjacent sequences, both upstream and downstream. The foundation model module consists of two parts: (1) a large pretrained language model and (2) downstream applications for virus detection and inhibition with respect to 9 CRISPR systems, where Cas12a, Cas12b, Cas12c, LwCas13a, LbuCas13a, Cas13b, Cas13d, Cas14 for detection task and LwCas13a, Cas13d and Cas7-11 for inhibition task.

The large pretrained language model

Construction of the CRISPRviva-3B corpus

To train the large pretrained language model, we assembled a corpus of 3.7 billion nucleotide sequences named CRISPRviva-3B. First, we extracted the whole genome and transcriptome of 23 cell lines and 264 segmented genomes of RNA viruses. Second, we used max sequence length of 50 bp to cleave the genome and transcriptome into guide RNA candidates with front and end dangling nucleotide fragments. Since the overall size was greater than 3 billion nucleotides, we named this corpus CRISPRviva-3B.

Sequence encoding

We utilized and modified the attention block, which is widely used in various research and industrial applications, to construct our sequence attention module[8], which functions as the building block of our CRISPR-viva system. The input of the module is a sequence of nucleotide characters composed of "A", "C", "G", "T" and "N", where "N" acts as both mask token for pretraining task and padding token for downstreaming task. For example, for the input guide RNA candidate "GGAAANCAGCAGATGGCNGGACATGGGCTGGAG", there are two locations being masked and the pretraining task is to recover the original sequence, which is "GGAAAGCAGCAGATGGCAGGACATGGGCTGGAG".

The structure of the large pretrained language model

The pretrained model was built with 12 attention layers. The structural parameters of the whole network are listed below:

Structural parameter	**Value**
Number of pretraining layers	12
Number of attention heads	12
Intermediate size	1024
Hidden size	384
Max position embeddings	64
Activation of hidden units	GELU
Dropout rate of attention units	0.1
Dropout rate of hidden units	0.1
Total number of parameters	**16,746,247**

The pretraining strategy for the large pretrained language model

In the pretraining process, we designed a curriculum learning-based training strategy that follows an easy-to-hard pattern, including 1-bp masking, 3-bp masking and 10% random masking tasks. For the 1-bp masking task, we randomly masked 1 bp of the input sequence and trained the model to predict the nucleotide at the masked position. Similarly, we randomly masked 3 bp and trained the model to predict all 3 masked nucleotides. For the 10% random masking task, we randomly masked each nucleotide with a probability of 10% and trained the model to predict all masked nucleotides. We used four NVIDIA GeForce GTX 4090 GPUs to train our model, and the training parameters are listed below:

Training parameter	**Value**
Optimizer	AdamW
Warm-up learning rate	0.001
Warm-up epochs	1
Epochs	100
Learning rate	0.005
Betas	(0.9, 0.999)
eps	1e-8

The structure of the downstream task-specialized model

Two strategies, i.e., the fine-tuning strategy and few-shot adaptation strategy, were applied here for different tasks with different numbers of training samples.

Fine-tuning strategy

The fine-tuning strategy is applied for medium data volume scenarios, where Cas12a and LwCas13a-based detection task as well as Cas13d-based inhibition task. Given the vast difference in terms of data volume between pretraining dataset and fine-tuning datasets, we only fine-tune the attention modules and output header layer. We utilized the low-rank adaptation (LoRA) method[9] with a very low learning rate to perform the fine-tune tasks in order to maintain training stability. Using LoRA, the update of each fine-tuned weight matrix can be decomposed into two low-rank matrices by the following equation:

$$h = W_0 x + \Delta W x = W_0 x + BAx \qquad (1)$$

where W_0 is the pretrained weight matrix; ΔW is the gradient calculated from the loss function; x is the input fine-tuning guide RNA sequences; A and B are two low-rank matrices, which are in fact being trained and whose multiplication can approximate towards ΔW but with much fewer weights.

We used one NVIDIA GeForce GTX 4090 GPU to train each task-specialized model, and the training parameters are listed below:

Training parameter	**Value**
Optimizer	AdamW
Epochs	100
Learning rate	0.001
Betas	(0.9, 0.999)
eps	1e-8
LoRA rank	8
LoRA alpha	1
LoRA dropout	0.1
Total number of parameters	**16,894,468**
Total number of trainable parameters	**148,226 (0.88%)**

Few-shot adaptation strategy

The few-shot adaptation strategy is applied for newly developed CRISPR systems or applications whose data volume is small ($<10^3$), including the Cas12b, Cas12c, LbuCas13a, Cas13b and Cas14 systems for detection task as well as LwCas13a and Cas7-11 systems for inhibition task. In this context, the goal is to learn a function that can capture the relationships or interactions between different elements or components of a given problem or task, meaning that this strategy takes both labelled data (supported data) and query data as input and predicts the label of the query data. We implemented a relation network[10] of two components: the embedding module and the relation module. The embedding module encodes the input data (both the query and supported data) into a meaningful representation, and we fix the weights of the first 10 layers of the pretrained model and utilize them to generate embeddings for the input sequence pairs. Then, the relation module takes the embedded representations as the input and calculates the relationships or interactions between them. The comparison pattern can be represented by the following equation.

$$r_{ij} = g_\phi(\mathcal{C}(f_\varphi(s_i); f_\varphi(x_j))) \qquad (5)$$

where r_{ij} is the relation score, g_ϕ is the relation layer, $\mathcal{C}$ is the depthwise concatenation operation, f_φ is the embedding module, s_i is the supported data, and x_j is the query data. The relation layer often involves pairwise comparisons or computations between the embedded representations, such as calculation of the similarities, distances, or correlations between pairs of elements. This process enables the network to capture the dependencies and relationships within the input data. By incorporating the relation network in our pretrained model, the downstream models can effectively leverage the learned relationships for generalization to new tasks or adaptation to different contexts. The network's ability to capture and model the relationships between pairs of elements can increase its ability to transfer knowledge and make accurate predictions from unseen data.

We use one NVIDIA GeForce GTX 4090 GPU to train each task-specialized model, and the training parameters are listed below:

Training parameter	**Value**
Optimizer	AdamW
Epochs	100
Learning rate	0.001
Betas	(0.9, 0.999)
eps	1e-8

4. Experimental Materials and Methods

4.1 Transcription template preparation

DNA primers were synthesized by Sangon Biotech (Shanghai, China). The DNA fragments were amplified via PCR using forward primers, reverse primers and 2×Fast Pfu Master Mix (Novoprotein, China), with water added to a total volume of 50 μL. The PCR products were purified with a SanPrep Column PCR Product Purification Kit (Sangon Biotech, China) according to the manufacturer's instructions. Then, the DNA fragments were reverse transcribed into RNA fragments using a T7 High Efficiency Transcription Kit (TransGen Biotech, China), and the reverse transcription products were purified with an EasyPure® RNA Purification Kit (TransGen Biotech, China) according to the manufacturer's instructions. Then, the purified reverse transcription products were diluted to 10^{12} copies/μL and stored at -80°C for further experiments.

4.2 Total cellular RNA extraction

HEK 293T cells, JEV-infected HEK 293T cells and PK15 cells were collected using TRIzol, and total RNA was extracted from the cells using an RNA Easy Fast Tissue/Cell Kit according to the manufacturer's instructions. The RNA was stored at -80°C for further experiments.

4.3 Fluorescent LbuCas13a nuclease assays

The LbuCas13a protein was purchased from GenScript (Nanjing, China), and the guide RNAs were synthesized by GenScript (Nanjing, China). In a 50 μL volume, the LbuCas13a protein (300 ng), guide RNA (20 nM), and Recombinant RNase Inhibitor (0.1 μL; Novoprotein, China) were mixed with 50 mM potassium acetate, 20 mM Tris-acetate and 10 mM magnesium acetate, the reporter RNA (1 μM; 5’-FAM-rUrUrUrUrUrU-BHQ1-3’) and varying amounts of the target RNA. The reactions were loaded into a 96-well plate (BBI, China), and the plate was incubated in a fluorescence plate reader (Agilent, BioTek Synergy H1) for up to 120 min at 37°C, with fluorescence measurement performed at 1 min intervals. Additionally, the reactions were incubated in a water bath at 37°C, and fluorescence images were acquired at different time points under a blue light glue cutter. The key resources are listed below:

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Guide RNA	GenScript	
DNA primers	Sangon Biotech	
ssRNA reporter	Tsingke Biotechnology	
LbuCas13a protein	GenScript	Z03742
T7 High Efficiency Transcription Kit	TransGen Biotech	JT101-02
SanPrep Column PCR Product Purification Kit	Sangon Biotech	B518141-0100
EasyPure® RNA Purification Kit	TransGen Biotech	ER701-01
RNA Easy Fast Tissue/Cell Kit	TIANGEN	DP451
Recombinant RNase Inhibitor	Novoprotein	E125-01A
Tris-acetate	BBI	A610101-0100
Potassium acetate	Sangon Biotech	A610438-0500
Magnesium acetate	Sangon Biotech	A501341-0500
2×Fast Pfu Master Mix	Novoprotein	E035-01B

Reference

1. Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57.
2. Pertea, G., and Pertea, M. (2020). GFF utilities: GffRead and GffCompare. F1000Research *9*.
3. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.
4. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., and Daly, M. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research *20*, 1297-1303.
5. Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature biotechnology *33*, 290-295.
6. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods *9*, 357-359.
7. Kenton, J.D.M.-W.C., and Toutanova, L.K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. (Minneapolis, Minnesota), pp. 2.
8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I. (2017). Attention is All you Need. pp. 5998-6008.
9. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.
10. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., and Hospedales, T.M. (2018). Learning to compare: Relation network for few-shot learning. pp. 1199-1208.