Bioinformatics, 38(9), 2022, 2437–2443
https://doi.org/10.1093/bioinformatics/btac144
Advance Access Publication Date: 16 March 2022
Original Paper

Sequence analysis

Evitar: designing anti-viral RNA therapies against future RNA viruses

Dingyao Zhang[1,2], Jingru Tian[1,2], Yadong Wang[1,2,3] and Jun Lu [1,2,3,4,*]

[1]Yale Stem Cell Center, Yale University, New Haven, CT 06520, USA, [2]Department of Genetics, Yale University School of Medicine, New Haven, CT 06510, USA, [3]Yale Center for RNA Science and Medicine, Yale Cancer Center, Yale University, New Haven, CT 06520, USA and [4]Yale Cooperative Center of Excellence in Hematology, Yale University, New Haven, CT 06520, USA

*To whom correspondence should be addressed.
Associate Editor: Valentina Boeva

Received on August 12, 2021; revised on February 7, 2022; editorial decision on February 24, 2022

Abstract

Motivation: The coronavirus disease 2019 (COVID-19) pandemic has highlighted the threat of emerging respiratory viruses and has exposed the lack of availability of off-the-shelf therapeutics against new RNA viruses. Previous research has established the potential that siRNAs and RNA-targeting CRISPR have in combating known RNA viruses. However, the feasibility and tools for designing anti-viral RNA therapeutics against future RNA viruses have not yet been established.

Results: We develop the Emerging-Virus-Targeting RNA (Evitar) pipeline for designing anti-viral siRNAs and CRISPR Cas13a guide RNA (gRNA) sequences. Within Evitar, we develop Greedy Algorithm with Redundancy and Similarity-weighted Greedy Algorithm with Redundancy to enhance the performance. Time simulations using known coronavirus genomes deposited as early as 10 years prior to the COVID-19 outbreak show that at least three SARS-CoV-2-targeting siRNAs are among the top 30 pre-designed siRNAs. In addition, among the top 19 pre-designed gRNAs, there are three SARS-CoV-2-targeting Cas13a gRNAs that could be predicted using information from 2011. Before-the-outbreak design is also possible against the MERS-CoV virus and the 2009-H1N1 swine flu virus. Designed siRNAs are further shown to suppress SARS-CoV-2 viral sequences using *in vitro* reporter assays. Our results support the utility of Evitar to pre-design anti-viral siRNAs/gRNAs against future viruses. Therefore, we propose the development of a collection consisting of roughly 30 pre-designed, safety-tested and off-the-shelf siRNA/CRISPR therapeutics that could accelerate responses to future RNA virus outbreaks.

Availability and implementation: Codes are available at GitHub (https://github.com/dingyaozhang/Evitar).

Contact: jun.lu@yale.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The coronavirus disease 2019 (COVID-19) pandemic, caused by the single-stranded RNA virus SARS-CoV-2 (Lu *et al.*, 2020; Su *et al.*, 2016; Weiss and Leibowitz, 2011; Zhu *et al.*, 2020) has highlighted the challenges posed by newly emerged RNA viruses to human societies. Existing off-the-shelf anti-viral drugs are often not effective in the face of a new virus. We postulated a model in which virus-specific siRNA or CRISPR therapeutics could be designed and developed prior to an outbreak, with a reasonably small sized collection of pre-designed and safety-tested siRNA/CRISPR drugs that could be used off-the-shelf to combat future viral diseases (Fig. 1A). siRNAs are small RNAs that can be chemically synthesized to target longer RNA species inside cells via perfect or near-perfect base-pairing with target RNAs, leading to AGO2-mediated target cleavage and degradation (Ameres *et al.*, 2007; Frank *et al.*, 2010; Schirle and MacRae, 2012). The choice of siRNAs as pre-designed anti-viral reagents is supported by the fact that many respiratory viruses are siRNA-targetable RNA viruses. Prototype siRNA drugs have also been tested in murine and primate models and have shown that they can effectively suppress both respiratory viruses such as SARS-CoV-2 (Gu *et al.*, 2020) and non-respiratory viruses such as EBOLA (Thi *et al.*, 2015) through pulmonary delivery (Bitko and Barik, 2007; Chow *et al.*, 2020; Lam *et al.*, 2012). Similarly, recent studies have shown that the Cas13a-based CRISPR system could be engineered to target RNA virus both *in vitro* and *in vivo*, including SARS-CoV-2 (Abbott *et al.*, 2020; Blanchard *et al.*, 2021), thus supporting the potential use of CRISPR in antiviral therapies.

Multiple tools have been published that design siRNAs based on target sequences, often incorporating design rules that enhance targeting efficacy (de Carli *et al.*, 2020; Good *et al.*, 2016; Huesken *et al.*, 2005; Knott *et al.*, 2014; Lück *et al.*, 2019; Moffat *et al.*, 2006; Naito and Ui-Tei, 2012; Naito *et al.*, 2004; Sciabola *et al.*, 2021; Tafer *et al.*, 2008). Designing tools are also available for anti-viral siRNAs (e.g. siVirus, VIRsiRNApred) (Naito *et al.*, 2006; Qureshi *et al.*, 2013) or Cas13 guide RNAs (gRNAs) against an existing virus (Abbott *et al.*, 2020). However, these tools are aimed at designing against a known virus using its viral sequences as input. A tool that is engineered toward combating future viruses is not yet available, which motivates us to perform this study.

In this study, we developed and implemented the Evitar pipeline (Fig. 1B), the pre-design mode of which takes the reference genomes from a class of existing viruses as input to pre-design a collection of siRNAs/gRNAs, out of which several could target a virus emerging afterwards. We also envisioned the possibility that some pre-designed siRNAs/gRNAs may fail pre-clinical or clinical testing and therefore incorporated a redesign process for 'failed' siRNA/gRNAs. The basis of this approach is the assumption that any future virus would bear some level of sequence similarity to one of the existing input viruses. Notably, a key requirement is that the pre-designed collection of siRNAs cannot be too large, otherwise it would be difficult and cost-prohibitive to go through efficacy and safety evaluations, especially in clinical trials. Using SARS-CoV-2, MERS-CoV and H1N1 swine flu virus as examples, we demonstrate with time simulation that Evitar can be used to design anti-viral siRNA/gRNAs prior to the viral outbreaks. We conclude that Evitar enables the design of siRNAs/gRNAs against future viruses.

2 Materials and methods

2.1 Source of genomic data

Details are included in Supplementary Methods.

2.2 The Evitar pipeline

An overview of the Evitar pipeline is presented in Figure 1B. The Evitar pipeline contains two modes. The mode A (also referred to as the predict mode), which is not extensively discussed in the main text, focuses on designing siRNA candidates against a single known virus by utilizing genomic sequences of multiple strains of the virus as an input. The details of mode A are included in Supplementary Methods.

2.3 The pre-designing pipeline (mode B) of Evitar

The mode B (also referred to as the pre-design mode) of the Evitar pipeline focuses on pre-designing siRNA candidates against future or emerging viruses, with the input being the genomic sequences of a class of viruses. The methods below use coronavirus as an example, but the pipeline could be used for other classes of RNA viruses.

2.3.1 Input viral genome information

We combined the genome sequences for input viral strains (details described in following sections) in a single FASTA file and used this file as the input of the pipeline.

2.3.2 Design candidate siRNAs for each of the input viral genomes

a. Rules for siRNA candidates

Rules used were the same as in the mode A pipeline.

b. Filtering and penalizing candidate siRNAs with off-target effects

Filtering and penalizing based on off-target effects were the same as in mode A.

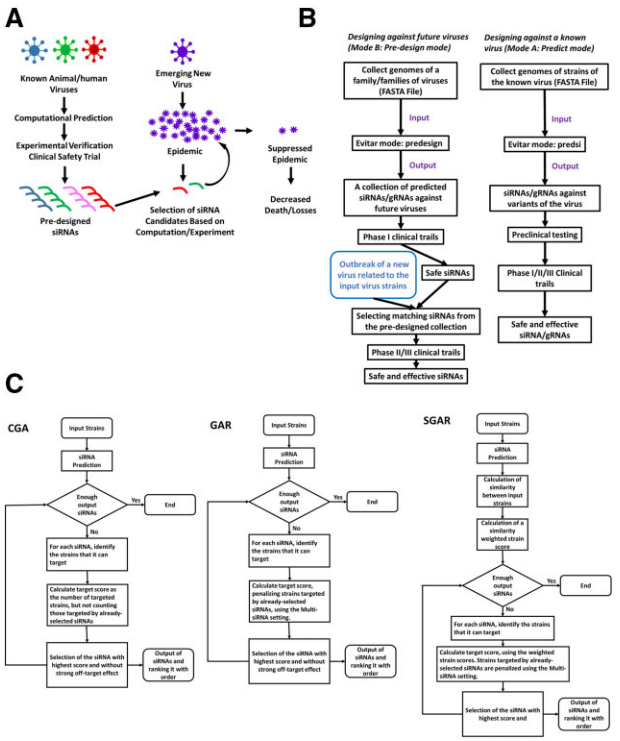

Fig. 1. Pre-designing siRNAs against SARS-CoV-2. (**A**) A schematic of the concept of pre-designing siRNAs to improve the speed of therapeutic response to future viral outbreaks. (**B**) Detailed workflows to use Evitar to pre-design siRNAs/gRNAs against future viruses or to predict siRNAs/gRNAs against an existing virus. (**C**) Flow charts of CGA, GAR and SGAR algorithms, with detailed steps of the algorithms shown.

2.3.3 Use the Greedy Algorithm with Redundancy and Similarity-weighted Greedy Algorithm with Redundancy algorithms to rank siRNAs based on the input viral genomes

a. Similarity-weighted Greedy Algorithm with Redundancy

The Similarity-weighted Greedy Algorithm with Redundancy (SGAR) algorithm has the following steps.

Step 1: Calculate the similarity between input strains based on the level of overlaps of candidate siRNAs. We calculated the similarity between viral strains by quantifying the overlap of candidate siRNAs between any two input strains:

$$S_{ij} = \frac{N_s}{N_a},$$

where S_{ij}, similarity score between viral stains i and j; N_s, the number of candidate siRNAs shared by the two stains i and j; N_a, the union of all unique candidate siRNAs targeting the two strains.

Step 2: Calculate a similarity-adjusted weight for each input strain

$$W_i = \frac{1}{1 + \sum_{j=1}^{N} S_{ij}},$$

where W_i, weight of stain i; S_{ij}, similarity between stains i and j; N, total number of input strains.

Step 3: Start the greedy selection

We obtained the union of all candidate siRNAs targeting any of the input strains. For each siRNA in this union, the target score was calculated as:

$$T_i = \sum_{j=1}^{N} \begin{cases} W_j * \max\left(\frac{R - C_j}{R}, 0\right), & \text{targeting} \\ 0, & \text{not targeting} \end{cases},$$

where T_i, target score for siRNA i; W_j, weight of strain j; R, user-defined Multi-siRNA parameter; C_j, count of selected siRNAs

(selected in previous iterations) that can target strain j; N, total number of input strains.

All candidate siRNAs were sorted based on their target scores in descending order. The first siRNA selected was the one that had the highest target score. If the highest target score was shared by multiple siRNAs, then one of them was randomly picked.

Step 4: Iterative selection of additional siRNAs

For each of the remaining siRNAs to be selected, we re-calculated the target score utilizing the information on which strains could be targeted by the already selected siRNAs. The siRNA with the highest targeting score then became the second selected siRNA. This process reiterated until the number of selected siRNAs reached a pre-defined number of output siRNAs that was set to 100 for this study. The Multi-siRNA parameter was designed to tune the robustness of identifying multiple siRNAs in a collection against future viruses. Users can adjust this parameter via the –repeatnum option in Evitar.

b. Greedy Algorithm with Redundancy (GAR)

GAR is similar to the SGAR algorithm, except that there is no weighting for the stains based on similarity. The target score for a siRNA was defined as:

$$T_i = \sum_{j=1}^{N} \begin{cases} \max(R - C_j, 0), & \text{targeting} \\ 0, & \text{not targeting} \end{cases},$$

where T_i, target score for siRNA i; R, user-defined Multi-siRNA parameter; C_j, count of selected siRNAs that can target strain j; N, total number of input stains.

c. Conventional Greedy Algorithm

Conventional Greedy Algorithm (CGA) was implemented by setting the Multi-siRNA parameter to 1 in the GAR algorithm.

2.4 Pre-designing of Cas13a gRNAs

The pre-design pipeline of Cas13a gRNA is similar to that of siRNAs. The only difference is that for each of the input viral sequence, instead of designing candidate siRNAs, we designed gRNAs using the Cas13design software (Wessels *et al.*, 2020). The gRNAs that Cas13design outputted were then filtered to retain only those with a score higher than or equal to 0.8 for further analysis. The rest of the pipeline is the same as that for siRNAs.

2.5 Evaluating GAR, SGAR and CGA, and using SGAR to pre-design siRNAs

2.5.1 Evaluation of GAR versus CGA

To compare the performance of GAR versus CGA, we performed stimulations. We used data from the Virus Pathogen Resource (ViPR) database as input information for prediction and used 15920 SARS-CoV-2 strains from GISAID database for the evaluation of the prediction results. The input data were *Coronaviridae* genomes collected before or in 2018. We further confirmed the removal of the bat coronavirus sequence RaTG13 (GenBank MN996532) due to its later sequence deposit. Evitar pipeline was run with the GAR algorithms, using the Multi-siRNA parameter ranging from 1 to 4. When the Multi-siRNA parameter was 1, the algorithm and results were identical to CGA.

After the prediction, we counted siRNAs which could target SARS-CoV-2 strains. An siRNA was considered SARS-CoV-2-targeting if it could perfectly match >50% of SARS-CoV-2 strains. In addition, to avoid multiple siRNAs with overlap in their target site locations and therefore not independent from each other, we removed any siRNA that overlapped with a better ranked siRNA.

2.5.2 Evaluation of SGAR versus GAR

Evaluation of SGAR versus GAR was performed similar to the comparison of GAR versus CGA, with the following differences. To build input data, we divided all *Coronaviridae* genome data into two data types and mixed them in various proportions. The first data type included genome sequences of human and bat viruses in *Coronaviridae*. The second data type included genome sequences of all non-human/bat viruses in Coronaviridae with genomes of betaCoVs removed to enhance strain biases. Type 1 input data and Type 2 input data were mixed so that the proportion of Type 2 sequences was 0%, 10%, 20%, 40%, 60%, 80% or 100% of the total. For each mixing ratio, we generated 5 randomly chosen sequence sets of 100 strains per set. Each input sequence set was then subjected to the Evitar pipeline, using SGAR and GAR in separate runs. We used the default Multi-siRNA setting of 3. We also performed analyses by varying the number of output siRNAs (set –limitnum option in Evitar to 50, 100 or 500).

To compare results between GAR and SGAR, we performed Wilcoxon matched-pairs signed rank test using GraphPad Prism 8 software. Since in this analysis, we limited the number of output siRNAs to 100, if none of the top 100 siRNAs could target SARS-CoV-2, we set 100 as the number for the sake of statistical comparison.

2.5.3 Time simulation of pre-designing siRNAs against SARS-CoV-2, MERS-CoV, SARS-CoV-1 and 2009-H1N1

For the time simulation of pre-designing against SARS-CoV-2, MERS-CoV and SARS-CoV-1, we assembled multiple sets of input data using all *Coronaviridae* genomes from the ViPR database with an ending collection time from year 2002 to 2018. We then used the Evitar pipeline with the SGAR algorithm to predict siRNAs using each of the input datasets. We used the default Multi-siRNA setting of 3. Evaluation of performance on SARS-CoV-2-targeting siRNAs was carried out in the same way as described in the section comparing GAR versus CGA. For evaluating results on MERS-CoV, we used 251 MERS-CoV genomes from the ViPR database. For evaluating results on SARS-CoV-1, we used 52 SARS-CoV-1 genomes from the ViPR database.

For the time simulation of pre-designing against 2009-H1N1, we assembled multiple sets of input data using all type-A Influenza genomes from the Influenza Virus Database with an ending collection time ranging from year 2000 to 2008. We further removed any H1N1 strains from these collections. We then used the Evitar pipeline, with the SGAR algorithm, to predict siRNAs using each of the input datasets. We used the default Multi-siRNA setting of 3. Evaluation of performance on 2009-H1N1 was similar to that of SARS-CoV-2 but used a total of 23 2009-H1N1 genomes from the NCBI GenBank.

2.5.4 Pre-designing siRNAs against future coronavirus based on existing coronavirus strains

Coronaviridae genomes from the ViPR database were first divided into those of SARS-CoV-2 and non-SARS-CoV-2 strains. For the first pre-design, we combined 769 human/bat non-SARS-CoV-2-strains with all 5221 SARS-CoV-2 strains in ViPR database. Because the number of SARS-CoV-2 strains was substantially larger than the non-SARS-CoV-2 strains, to test whether the SGAR algorithm could overcome this distorted distribution of the strains, we also performed a parallel analysis using an input set with the 769 human/bat non-SARS-CoV-2 strains and a randomly selected set of 192 SARS-CoV-2 strains with the latter representing 25% of the non-SARS-CoV-2 strains. The Evitar pipeline was run with SGAR with the default Multi-siRNA setting of 3.

For the second pre-design, we used 3200 non-SARS-CoV-2 Coronaviridae genomes available in the ViPR database regardless of the host. These genomes were combined with 5221 SARS-CoV-2 genomes as input. We also did a similar analysis with 3200 non-SARS-CoV-2 strains and 800 SARS-CoV-2 strains for comparison. The Evitar pipeline was run with SGAR with the default Multi-siRNA setting of 3.

2.5.5 Simulating the re-design of siRNAs based on experimental results

We first performed pre-design against SARS-CoV-2 using human/bat Coronaviridae genomes with an ending year in 2018 as described in the time-simulation section above. For the resulting top

25 pre-designed siRNAs, we randomly labeled 13/25 (~50%) to be successful, with the rest being failed. Using Evitar, we redesigned by inputting the lists of successful and failed siRNAs. The re-design process was similar to the pre-design procedures but with the following differences. First, any selected siRNA that overlapped with the successful list was automatically retained. Second, for any selected siRNA that was a member of the failed list, this siRNA was replaced with another siRNA based on the following. (i) The replacement siRNA was not a member of the successful list or failed list. (ii) The replacement siRNA was not an siRNA already selected in the SGAR process. (iii) The replacement siRNA had the highest target similarity score with the siRNA to be replaced. The target similarity score was calculated based on strains that could be targeted by both the replacement siRNA and the siRNA to be replaced (referred to as 'shared target'). The target similarity score was defined as:

$$TS_{ik} = \sum_{j=1}^{N} \begin{cases} W_j * \max\left(\frac{R - C_j}{R}, 0\right), & j \text{ is a shared target} \\ 0, & j \text{ is not a shared target} \end{cases},$$

where TS_{ik}, target similarity score between siRNA i and siRNA k; W_j, weight of strain j (see section on SGAR); R, user-defined Multi-siRNA parameter (see section on SGAR); C_j, count of selected siRNAs that can target strain j (see section on SGAR); N, total number of input stains (see section on SGAR).

2.6 Pre-designing siRNAs based on broadly targeting siRNAs

Details are included in Supplementary Methods.

2.7 Luciferase reporter assays

Details are included in Supplementary Methods.

2.8 Availability of the computational pipeline

The Evitar pipeline is available as a command-line software in GitHub (https://github.com/dingyaozhang/Evitar).

3 Results

To achieve pre-designing a small collection of siRNAs to prepare against future viruses, we initially tested the seemingly straightforward strategy of identifying a collection of siRNAs, each of which could target the most numbers of input viral strains. We used available *Coronaviridae* genomes prior to 2019 as an input (which does not contain SARS-CoV-2 genomes) to design siRNAs and ranked them based on their ability to target the most numbers of input genomes. However, none of the top 100 siRNAs designed by this approach could target SARS-CoV-2 effectively (Supplementary Table S1, and see Section 2).

To implement Evitar (Fig. 1B), we turned to the Greedy algorithm (referred to hereafter as 'conventional Greedy algorithm' or CGA), originally designed to solve the mathematical maximum coverage problem. CGA could be used to identify the smallest number of siRNAs to ensure that each of the input viral strains could be targeted by at least one siRNA. By iterative evaluation of siRNAs to minimize the number of siRNAs to cover most strains (see flow chart in Fig. 1C), CGA could theoretically make it possible to predict one siRNA against a future virus. On the other hand, for the purpose of building a collection of siRNAs against a future viral disease, having a single working siRNA within the collection is not sufficient, because there could be failures of siRNA candidates due to issues of efficacy or safety (Gu *et al.*, 2020). Instead, we need siRNA collections that contain two or more siRNAs against a future virus. We thus developed the Greedy Algorithm with Redundancy (GAR), which retained the CGA concept of maximizing the targeting of input strains with the fewest siRNAs, but was modified to allow siRNA-targeting redundancy, by reducing the penalty of siRNAs on viral strains that were targeted by the already-selected siRNAs in the collection (see flow chart in Fig. 1C). By incorporating a Multi-siRNA parameter in the algorithm (see Section 2 for details), GAR could tune the redundancy, and therefore, the robustness of the pre-designed siRNA collections.

To evaluate both the feasibility of predicting siRNAs against future viruses, and to evaluate the relative performance of GAR versus CGA, we performed time simulation to mimic the emergence of SARS-CoV-2. *Coronaviridae* genomic sequences from the Virus Pathogen Database (Pickett *et al.*, 2012) were filtered to only retain those deposited up to the end of 2018, a time point about a year prior to the COVID-19 outbreak. We further focused on *Coronaviridae* strains that could use human or bat as a host, due to bat being a natural reservoir of many coronaviruses that eventually crossed the species barrier into human (Hu *et al.*, 2015). Using these 717 sequences as input, which did not include any SARS-CoV-2 stains, we predicted and ranked the siRNAs using both CGA and GAR. The prediction of siRNAs considered the biochemical properties of siRNAs, the target RNA accessibility, and potential off-target effects on human lung transcriptome (see Section 2). We evaluated the performance of CGA and GAR by counting the number of top predicted siRNAs that were needed to have at least one (redundancy = 1), two (redundancy = 2) or three (redundancy = 3) non-overlapping siRNAs, each of which could target the vast majority of the 15920 analyzed SARS-CoV-2 strains. Interestingly, both GAR and CGA successfully predicted siRNAs that could target SARS-CoV-2 (Fig. 2A). Comparing results from GAR and CGA (note that CGA is equivalent to running GAR with setting the Multi-siRNA parameter to 1) showed that with higher redundancy requirements (e.g. redundancy of 3), GAR performed substantially better than CGA when larger Multi-siRNA parameters were used, at a minor cost of slightly worse performance for the redundancy requirement of one (Fig. 2A). The analyses above demonstrated the feasibility of pre-designing siRNAs against a future virus using existing viral information. The results further demonstrated that GAR outperformed CGA on the robustness of finding multiple siRNAs against a future virus.

The existing viral genome information tends to be heavily biased toward virus types that have been more intensely studied. On the other hand, it is not guaranteed that a future virus would emerge

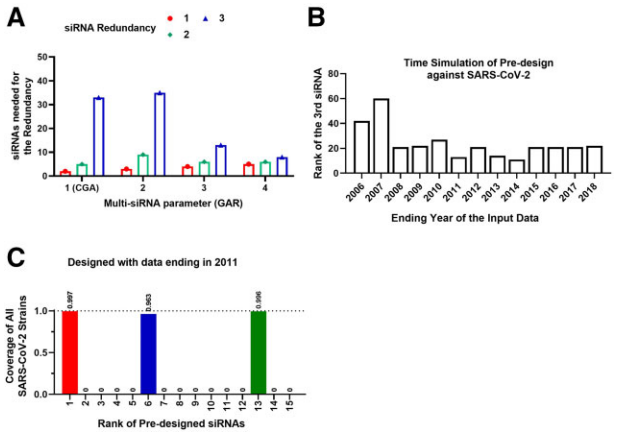

Fig. 2. Pre-designing siRNAs against SARS-CoV-2. (**A**) Human and bat coronavirus genomes with an ending year of 2018 were used as input for the Evitar pipeline, with the GAR algorithm and the Multi-siRNA parameters as indicated. When the Multi-siRNA parameter was 1, the results were identical to CGA. The numbers of top predicted siRNAs that were needed to identify one, two or three (redundancy) siRNAs against SARS-CoV-2 are plotted. (**B**) Time simulations were performed to evaluate the earliest time feasible to pre-design siRNAs against SARS-CoV-2. Human/bat Coronaviridae genomes with the indicated ending years were assembled and used as input for the Evitar pipeline, using the SGAR algorithm. Data indicate the ranks of the third siRNA that target SARS-CoV-2 (reflecting redundancy = 3). (**C**) For the analysis from (B), the detailed results for input data ending in 2011 are plotted, with the top-ranked pre-designed siRNAs shown. Each bar represents the fraction of SARS-CoV-2 genomes that can be targeted by the corresponding siRNA, with the number above the bar indicating the exact fraction

with strong similarity to an intensely studied virus in the past. Both GAR and CGA treated each input stain equally, without considering similarity among input stains, and thus could be more susceptible to the biases in existing data. Therefore, we developed SGAR algorithm, or Similarity-weighted GAR, which was aimed at reducing the effects of such biases. SGAR added weights based on similarity between input viral strains, with strains that differ more from other input sequences weighted more (see flow chart in Fig. 1C). To simulate biased input data, we mixed human/bat coronavirus genomes with non-human/bat coronavirus genomes at various ratios. Pre-designs were then performed, generating a total of 100 pre-designed siRNAs in each case. We then evaluated the pre-designed siRNAs and used the ranks of siRNAs that can target SARS-CoV-2 as a comparison metric. We found that SGAR outperformed GAR when input sequence sets contained fewer human/bat coronavirus genomes (see Section 2, Supplementary Fig. S3A and B). Similar results were obtained when we changed the number of output siRNAs to 500 or 50 (Supplementary Fig. S3C and D). These data support that SGAR is less sensitive to skewed input data than GAR. We thus used SGAR in the subsequent analyses, and enabled users to choose either GAR or SGAR in the Evitar pipeline.

We next asked when was the earliest time possible to pre-design siRNAs against SARS-CoV-2 by simulating the time cutoff of input viral strains. We took the *Coronaviridae* genomes as input for SGAR but limited the input sequences to an ending year that range from 2003 to 2018. We used an evaluation criterion of the number of top-ranked pre-designed siRNAs necessary to identify three siRNAs against SARS-CoV-2. We found that for the ending year from 2008 to 2018, a maximum of 27 top ranked siRNAs was needed, whereas predictions using data ending before 2008 had worse performances (Fig. 2B). As an example, with data ending in year 2011, only 13 siRNAs were needed to find three siRNAs, each of which was predicted to target >96% of analyzed SARS-CoV-2 strains (Fig. 2B and C, Supplementary Table S2). In hindsight, if this strategy were available in 2009 or before, there could have been a 10-year lead period to prepare against the SARS-CoV-2 outbreak. These data suggest that the pre-design concept holds a strong promise.

We next asked whether we could pre-design siRNAs against MERS-CoV and SARS-CoV-1 using Coronaviridae sequences prior to their outbreaks in June 2012 and November 2002, respectively. With 235 input strains of human and bat coronaviruses ending in 2011, the first three pre-designed siRNAs against MERS-CoV were ranked 12, 26, and 52 among all pre-designed siRNAs, each of which could target all MERS-CoV stains (Fig. 3A). Predictions using data ending in year 2010 or before had weaker performances (Supplementary Table S3). Nevertheless, two MERS-CoV-targeting siRNAs could be identified among the top 100 predicted siRNAs using data ending as early as 2008. The reduced performance was associated with reduced numbers of available input strains, with 218 stains for data ending in 2010 and 163 strains for data ending in 2008. We next tested pre-designing against SARS-CoV-1. Only 54 human/bat coronavirus sequences were available prior to 2002, and none of the top 100 SGAR-predicted siRNAs could target SARS-CoV-1. To determine whether the low numbers of input strains was the reason of the failed pre-design against SARS-CoV-1, we used *Coronaviridae* sequences ending in 2018 and removed all MERS-CoV and SARS-CoV-1 sequences from this input set. Among SGAR-pre-designed siRNAs, 8 and 19 top siRNAs were needed to identify three that could target MERS-CoV and SARS-CoV-1, respectively (Fig. 3B and C). These data indicate that the pre-design of siRNAs against MERS-CoV can be achieved using sequence information prior to the outbreak, and suggest that limited numbers of available input sequences weaken the performance of the pre-design against both SARS-CoV-1 and MERS-CoV.

We next asked whether pre-design is feasible against a respiratory RNA virus that is not a coronavirus. We thus explored pre-design against the 2009-H1N1 influenza virus, which underlies the 2009 swine flu pandemic. We took type-A influenza genomes as inputs, with the input-data-ending year ranging from 2000 to 2008. To make the pre-design more challenging, we removed all H1N1

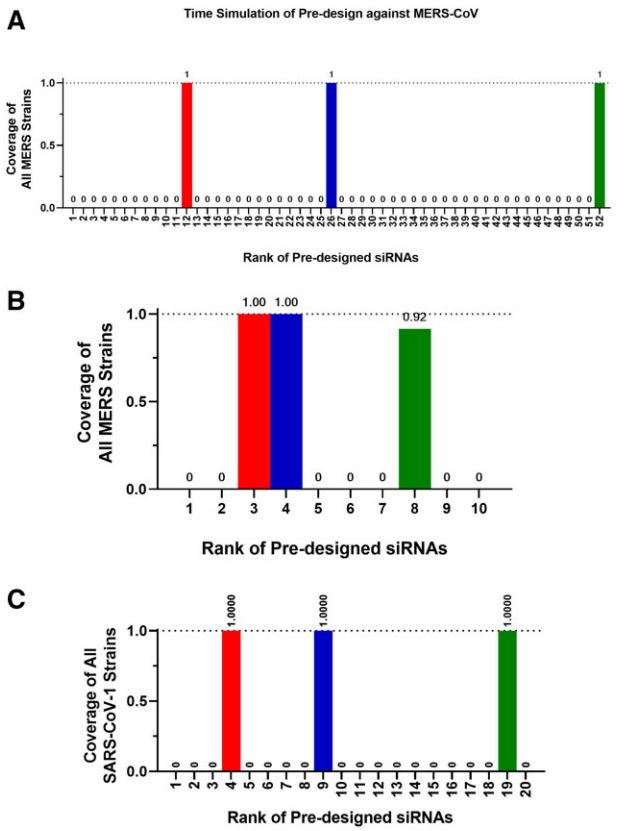

Fig. 3. Predicting siRNAs against MERS-CoV and SARS-CoV-1. (**A**) Time simulations were performed to evaluate the earliest time feasible to pre-design siRNAs against MERS-CoV. Coronaviridae genomes with 2011 as the ending year were assembled and used as input for the Evitar pipeline, using the SGAR algorithm with a Multi-siRNA setting of 3. Top ranked siRNAs were evaluated for their ability to target MERS-CoV. Each bar represents the fraction of MERS-CoV genomes that can be targeted by the corresponding siRNA, with the number above the bar indicating the exact fraction. (**B**) SARS-CoV-1 sequences were removed from Coronaviridae genomes with an ending year of 2018. The Evitar pipeline with the SGAR algorithm was run on these input data, using a Multi-siRNA setting of 3. Top ranked siRNAs were evaluated for their ability to target SARS-CoV-1. Each bar represents the fraction of SARS-CoV-1 genomes that can be targeted by the corresponding siRNA, with the number above the bar indicating the exact fraction. (**C**) Similar data as in (B) are shown for MERS-CoV

strains from the input. Results show that pre-design was possible even with input data ending in year 2000 (1093 total input strains), for which each of the top three pre-designed siRNAs was capable of targeting all 2009-H1N1 strains (Supplementary Table S4).

Some pre-designed siRNAs in a collection could fail efficacy and safety tests, which would lead to fewer siRNAs entering clinical trials and thus reduce the robustness to prepare for future outbreaks. To overcome this problem, we developed a re-design process in which a replacement siRNA could be identified that could target a similar set of input strains as a failed siRNA. We simulated re-design using SARS-CoV-2 as an example. Among the top 25 pre-designed siRNAs against SARS-CoV-2, with input sequence ending in 2018, there were four siRNAs that could each target > 94% of SARS-CoV-2 strains. We randomly assigned these 25 siRNAs into the successful and failed groups, with a ~ 50% success rate, leading to three out of four SARS-CoV-2-targeting siRNAs falling into the failed group. Redesigning yielded the same number (four) of SARS-CoV-2-targeting siRNAs among the top 25 pre-designs (Supplementary Table S5). These data support that additional siRNAs could be easily re-designed to replace failed siRNAs to facilitate preparation against future viruses.

In addition to siRNAs, we incorporated the capability of pre-designing Cas13a gRNAs against future viruses into the Evitar pipeline, with the same SGAR algorithms implemented. Using the pre-

design mode of Evitar on Cas13a gRNAs, as well as input *Coronaviridae* genomes ending in 2011, we found we could identify three anti-SARS-COV-2 gRNAs in the top 19 pre-designed gRNAs, with each targeting >99% of analyzed SARS-CoV-2 strains (Supplementary Fig. S4).

We named the pre-design functions as mode B in the Evitar pipeline and used currently existing coronavirus sequences as input to pre-design siRNAs against possible future coronaviruses. Two lists of pre-designed siRNAs were generated. The first pre-designed list was based on human/bat coronaviruses as an input, with 192 non-SARS-CoV-2 strains and 5221 SARS-CoV-2 strains (Supplementary Table S6). The second list was based on *Coronaviridae* sequences from all host species, with 3200 non-SARS-CoV-2 strains and 5221 SARS-CoV-2 strains. We observed strong overlaps between the top pre-designed siRNAs (Supplementary Tables S6 and S7). Moreover, because the number of SARS-CoV-2 strains were greater than non-SARS-CoV-2 strains in both pre-designs, we tested whether SGAR is insensitive to the over representation of SARS-CoV-2 strains. We reduced the numbers of SARS-CoV-2 strains to 25% of that of non-SARS-CoV-2 strains in the input data, and found that the ranks of SARS-CoV-2-targeting siRNAs in the output collections were similar to the pre-designs with larger numbers of SARS-CoV-2 input strains (Supplementary Fig. S5). These data further support the utility of SGAR on skewed input data. We propose that the two lists of pre-designed siRNAs (Supplementary Tables S6 and S7) could form candidates to prepare against future coronaviruses. In addition, we also established the Evitar mode A, which was aimed at designing siRNAs against a known virus using existing viral strains of that virus. Applying Evitar mode A to SARS-CoV-2, we designed siRNAs that were predicted to target the majority of 15920 input SARS-CoV-2 strains (Supplementary Table S8).

Finally, to experimentally test the ability of the designed siRNAs to target SARS-CoV-2 sequences, we used reporter assays in which parts of SARS-CoV-2 sequences were cloned in the 3'UTR of luciferase. We picked three siRNAs for testing, including two designed against SARS-CoV-2 with Evitar mode A (Supplementary Table S8), and one from the mode B pre-design (Supplementary Table S7) that matched SARS-CoV-2 genomes. Two of these siRNAs were predicted to target the RNA-dependent RNA polymerase and another predicted to target ORF1a in the SARS-CoV-2 genome (Fig. 4A). Of note, all designed/pre-designed siRNAs against SARS-CoV-2 in this study, including the three siRNAs tested here, perfectly matched the alpha, beta, delta and gamma variants. Transfecting each of the three siRNAs, or a combination of two of them, led to significant suppression of the luciferase reporters (Fig. 4B), confirming their activities in mammalian cells.

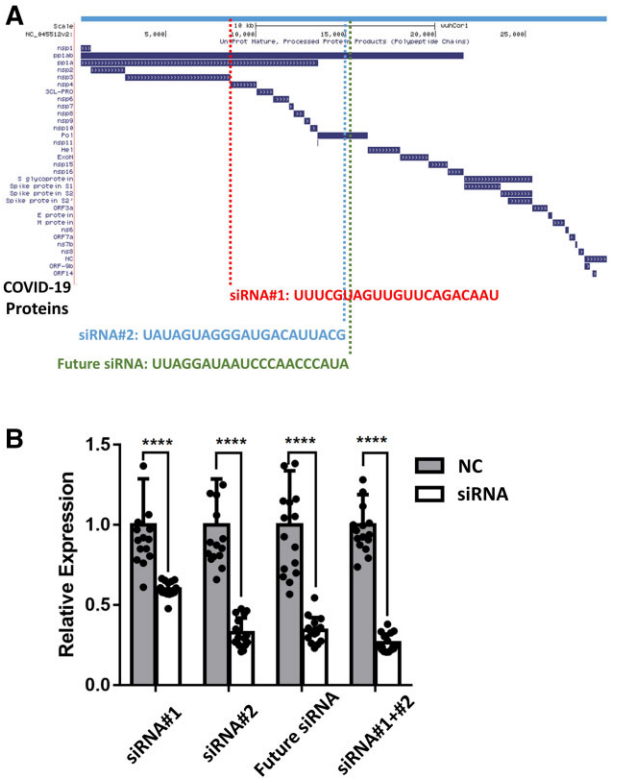

Fig. 4. Candidate siRNAs suppress target sequence expression. (**A**) The locations and sequences of the three experimentally tested siRNAs are shown together with the open reading frames in the SARS-CoV-2 genome. siRNA#1 and #2 were from Mode A design by Evitar, whereas the indicated Future siRNA was from Mode B pre-design against future viruses. (**B**) SARS-CoV-2 sequences containing the siRNA target sites were cloned into luciferase reporters. The indicated siRNAs or negative control (NC) siRNA were co-transfected with reporter plasmids into 293T cells, and luciferase activities were analyzed 2 days after. Data were normalized to those of NC controls. $N = 16$. Data are representative of two independent experiments. Error bars represent standard deviation. *P*-values were calculated using student's *t*-test. $^{****}P < 0.0001$

4 Conclusions and discussion

In this study, we developed the Evitar pipeline and we conclude with time simulation analyses that it is feasible to use Evitar to pre-design siRNAs or Cas13a gRNAs against future respiratory coronaviruses. Evitar can also rationally design siRNA against an existing virus. The prospect of pre-designing anti-viral RNA reagents is exciting and has not been previously explored. Indeed, Evitar is the first tool that is aimed at designing anti-viral siRNA/gRNAs for future viruses. We demonstrated the pre-design of siRNAs using Evitar against SARS-CoV-2, MERS-CoV and 2009-H1N1 using data prior to their outbreaks. In the case of SARS-CoV-2, we showed the feasibility of pre-designing using data with a cutoff as early as year 2008. For SARS-CoV-1, the pre-design using data prior to year 2002 did not generate successful siRNAs. We believe the reason is due to the limited number of human/bat coronavirus genomes prior to 2002 because including later strains (without SARS-CoV-1 sequences) made it possible to design siRNAs against SARS-CoV-1. Similar observations were made for MERS-CoV. Although pre-design of siRNAs against MERS-CoV was possible, performance was substantially improved when later coronavirus strains (without MERS-CoV sequences) were included. These data also strongly argue for the need to sustain and enhance the basic research into respiratory viruses in animal species including bats. Our data suggest that the more information is available, the better the pre-design concept can work.

Evitar pipeline utilizes GAR and SGAR algorithms developed in this study. Both GAR and SGAR have a Multi-siRNA/gRNA setting to improve the robustness of identifying more than one siRNA (or gRNA) against a future viral strain within a small collection of top-ranked candidate siRNAs. Our data show that SGAR have advantages over GAR suggesting that SGAR is the preferred mode. Nevertheless, both GAR and SGAR are included in the Evitar pipeline. The choice of the algorithm, and the choice of the Multi-siRNA/gRNA setting could be defined by users.

In this study, we evaluated pre-designs for a collection of siRNA/gRNAs with the collection size set at or around 30. This choice of collection size is based on our own estimates of the amount of pre-clinical experimental testing that could be reasonably handled by a single laboratory and the amount of resources required for clinical trials in the setting of one or a few pharmacologic corporations. Clearly, the larger the collection size, the higher the chance that several pre-designed siRNAs/gRNAs could target a virus emerging in the future. However, the cost and resources required for follow-up pre-clinical and clinical studies would also increase with a larger collection size. The optimal collection size would thus need to be determined based on existing resources.

The pre-designed siRNAs and gRNAs will need to undergo experimental testing. These include evaluation of the targeting efficiency against viral sequences in tissue culture, engineering their delivery into lung tissues, and testing the efficacy against known viruses in animal models. On the computational side, further

refinement of the pipeline could be performed. For example, it could be beneficial to utilize experimental data to improve the prediction of effective siRNA/gRNAs against viral sequences that tend to be heavily structured. Finally, while we focused our analysis on respiratory RNA viruses in this study, designing siRNAs/gRNAs against non-respiratory viruses, such as the EBOLA virus (Thi *et al.*, 2015) should also be possible using Evitar.

Acknowledgements

The authors thank the contributors to the GISAID database (Supplementary Table S9) and all databases that have been used in this study. They thank colleagues within Yale Cooperative Center of Excellence in Hematology for stimulating discussions and suggestions. They thank Luke Lu for proofreading the manuscript.

Author contributions

D.Z. and J.L. designed the study. D.Z. performed all *in silico* analyses. J.T. and Y.W. performed reporter experiments. J.L. supervised the study. D.Z. and J.L. wrote the manuscript.

Funding

This work was supported in part by National Institutes of Health [R01GM116855, R01GM138856, R33CA225863 and R33CA246711 to J.L.].

Conflict of Interest: none declared.

References

Abbott,T.R. *et al.* (2020) Development of CRISPR as an antiviral strategy to combat SARS-CoV-2 and influenza. *Cell*, **181**, 865–876.e12.

Ameres,S.L. *et al.* (2007) Molecular basis for target RNA recognition and cleavage by human RISC. *Cell*, **130**, 101–112.

Bitko,V. and Barik,S. (2007) Respiratory viral diseases: access to RNA interference therapy. *Drug Discov. Today Therap. Strategies*, **4**, 273–276.

Blanchard,E.L. *et al.* (2021) Treatment of influenza and SARS-CoV-2 infections via mRNA-encoded Cas13a in rodents. *Nat. Biotechnol.*, **39**, 717–726.

de Carli,G.J. *et al.* (2020) SSD – a free software for designing multimeric mono-, bi- and trivalent shRNAs. *Genet. Mol. Biol.*, **43**, e20190300.

Chow,M.Y.T. *et al.* (2020) Inhaled RNA therapy: from promise to reality. *Trends Pharmacol. Sci.*, **41**, 715–715.

Frank,F. *et al.* (2010) Structural basis for 5′-nucleotide base-specific recognition of guide RNA by human AGO2. *Nature*, **465**, 818–822.

Good,R.T. *et al.* (2016) OfftargetFinder: a web tool for species-specific RNAi design. *Bioinformatics*, **32**, 1232–1234.

Gu,S.H. *et al.* (2020) A small interfering RNA lead targeting RNA-dependent RNA-polymerase effectively inhibit the SARS-CoV-2 infection in Golden Syrian hamster and Rhesus macaque. *bioRxiv*, **2020**.07.07.190967.

Hu,B. *et al.* (2015) Bat origin of human coronaviruses. *Virol. J.*, **12**, 221.

Huesken,D. *et al.* (2005) Design of a genome-wide siRNA library using an artificial neural network. *Nat. Biotechnol.*, **23**, 995–1001.

Knott,S.R.V. *et al.* (2014) A computational algorithm to predict shRNA potency. *Mol. Cell*, **56**, 796–807.

Lam,J.K.W. *et al.* (2012) Pulmonary delivery of therapeutic siRNA. *Adv. Drug Deliv. Rev.*, **64**, 1–15.

Lu,R. *et al.* (2020) Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet, **395**, 565–574.

Lück,S. *et al.* (2019) siRNA-Finder (si-Fi) software for RNAi-target design and off-target prediction. *Front. Plant Sci.*, **10**, 1023.

Moffat,J. *et al.* (2006) A Lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell*, **124**, 1283–1298.

Naito,Y. *et al.* (2004) siDirect: highly effective, target-specific siRNA design software for mammalian RNA interference. *Nucleic Acids Res.*, **32**, W124–W129.

Naito,Y. *et al.* (2006) siVirus: web-based antiviral siRNA design software for highly divergent viral sequences. *Nucleic Acids Res.*, **34**, W448–W450.

Naito,Y. and Ui-Tei,K. (2012) SiRNA design software for a target gene-specific RNA interference. *Front. Genet.*, **3**, 102.

Pickett,B.E. *et al.* (2012) ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res.*, **40**, D593–D598.

Qureshi,A. *et al.* (2013) VIRsiRNApred: a web server for predicting inhibition efficacy of siRNAs targeting human viruses. *J. Transl. Med.*, **11**, 305–312.

Schirle,N.T. and MacRae,I.J. (2012) The crystal structure of human argonaute2. *Science*, **336**, 1037–1040.

Sciabola,S. *et al.* (2021) PFRED: a computational platform for siRNA and antisense oligonucleotides design. *PLoS One*, **16**, e0238753.

Su,S. *et al.* (2016) Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. Trends in Microbiology, **24**, 490–502.

Tafer,H. *et al.* (2008) The impact of target site accessibility on the design of effective siRNAs. *Nat. Biotechnol.*, **26**, 578–583.

Thi,E.P. *et al.* (2015) Lipid nanoparticle siRNA treatment of Ebola-virus-Makona-infected nonhuman primates. *Nature*, **521**, 362–365.

Wessels,H.-H. *et al.* (2020) Massively parallel Cas13 screens reveal principles for guide RNA design. *Nat. Biotechnol.*, **38**, 722–727.

Weiss,S.R. and Leibowitz,J.L. (2011) Coronavirus pathogenesis. In: *Advances in Virus Research*. Academic Press Inc., pp. 85–164.

Zhu,N. *et al.* (2020) A Novel Coronavirus from Patients with Pneumonia in China, 2019. The New England Journal of Medicine, **382**, 727–733.