# Generation of small interfering RNA (siRNA) database from SARS-CoV-2 genome sequences

**Inácio Gomes Medeiros**
Bioinformatics Graduate Program, Metrópole Digital Institute, Federal University of Rio Grande do Norte, Natal, Rio Grande do Norte, 59078-400, Brazil https://orcid.org/0000-0002-3407-218X

**André Salim Khayat**
Instituto de Ciências Biológicas, Universidade Federal do Pará, Belém, Pará, 66075-110, Brazil

**Beatriz Stransky**
Bioinformatics Multidisciplinary Environment (BioME), Metrópole Digital Institute, Federal University of Rio Grande do Norte, Natal, Rio Grande do Norte, 59078-400, Brazil

**Sidney Emanuel Batista dos Santos**
Instituto de Ciências Biológicas, Universidade Federal do Pará, Belém, Pará, 66075-110, Brazil

**Paulo Pimentel de Assumpção**
Núcleo de Pesquisas em Oncologia, Universidade Federal do Pará, Belém, Pará, 66073-110, Brazil

**Jorge Estefano Santana de Souza** (jorge@imd.ufrn.br)
Bioinformatics Multidisciplinary Environment (BioME), Metrópole Digital Institute, Federal University of Rio Grande do Norte, Natal, Rio Grande do Norte, 59078-400, Brazil

**Method Article**

**Keywords:** Database, siRNA, SARS-CoV-2

**DOI:** https://doi.org/10.21203/rs.3.pex-1207/v1

# Abstract

This protocol aims to describe the building of a database of SARS-CoV-2 targets for siRNA approaches. Starting from the virus reference genome, we will derive sequences from 18 to 21nt-long and verify their similarity against the human genome and coding and non-coding transcriptome, as well as genomes from related viruses. We will also calculate a set of thermodynamic features for those sequences and will infer their efficiencies using three different predictors. The protocol has two main phases: at first, we align sequences against reference genomes. In the second one, we extract the features. The first phase varies in terms of duration, depending on computational power from the running machine and the number of reference genomes. Despite that, the second phase lasts about thirty minutes of execution, also depending on the number of cores of running machine. The constructed database aims to speed the design process by providing a broad set of possible SARS-CoV-2 sequences targets and siRNA sequences.

# Introduction

The rapid transformation of coronavirus disease 2019 (COVID-19) into a global pandemic raised a demand for the development of antivirals solutions capable of targeting the SARS-CoV-2 RNA genome, for which RNA interference (RNAi) approaches[1] using small-interference RNAs (siRNAs) were pointed out as possible solutions[2,3]. Past experiences with SARS proved the applicability of siRNAs in this context[4,5], leading to the publication of diverse patents (for a resumed list of them, refer to Liu *et al.*[3]). Despite that, the fast ending of SARS and MERS epidemics in the past decade slowed down the efforts on continuing the development of methodologies regarding siRNAs discovery and design, which was rewarmed again with the COVID-19 pandemic.

The development of *in silico* siRNAs design protocols is essential not only during pandemic epochs but also as a way of preparing for new possible events in the future, such that a framework is ready for application on producing new antiviral solutions. Embedded in this context, we propose a protocol for building a database of siRNAs targets and sequences devoted to the development of antiviral solutions against SARS-CoV-2, with the proper sequences together with a set of features with which is possible to assess their quality and efficiency.

We are going to employ Human, SARS, and MERS reference genomes, so as the knowledge acquired with SARS-CoV-2 genome sequencing projects in Brazil, China, England, Germany, Italy, Russia, Spain, and the USA, seeking to be capable of assessing issues related to possible cross-reaction of siRNAs with the human genome, so as the capability os siRNAs being able to target with higher specificity SARS-CoV-2 strains from any of the mentioned countries.

The protocol has two main phases: a first in which sequences are aligned against Human, SARS, and MERS reference genomes, besides Human coding and non-coding transcriptomes; and a second one, in which the features are extracted, regarding base context, thermodynamic information, and efficiency prediction. Codes employed for the extraction of these features are freely available at https://github.com/inaciomdrs/sirna_db_building_protocol. They cover scripts made by authors to build the database, translations of JavaScript codes from OligoCalc[6] to Python programming language, parts of SSD[7] software, and copies of binaries from ThermoComposition21[8] and si_shRNA_selector[9], to assure that users have the option to use the same software versions that authors use in this protocol.

Once constructed, the database will have more than 170 features, including thermodynamic information, base context, target genes, and alignment information, all of them calculated from the reference genome of interest virus and the application of alignments against human genome and coding and non-coding transcriptome, and to related viruses genomes.

# Reagents

# Equipment

For the development of this protocol, we used a workstation with 40 cores, 300GB of RAM, and 1TB of disk space, with Linux as Operating System (authors used Red Hat Enterprise Linux distribution). Equipment with lower computational power can also be used. This, however, will strongly impact on protocol execution time.

# Procedure

**Step 1: Data acquisition**

Download the following genomic sequence data:

- SARS-CoV-2 GISAID Brazil, China (Wuhan region only and whole country less Wuhan), England, Germany, Italy, Russia, Spain and USA strains genomes, available at https://www.gisaid.org/. We recommend keeping strains genomes from different countries in different files.

- The human genome, available at http://igenomes.illumina.com.s3-website-us-east-1.amazonaws.com/Homo_sapiens/Ensembl/GRCh37/Homo_sapiens_Ensembl_GRCh37.tar.gz;

- Human coding transcriptome, available at ftp://ftp.ensembl.org/pub/release-100/fasta/homo_sapiens/cds/Homo_sapiens.GRCh38.cds.all.fa.gz;

- Human non-coding transcriptome, available at ftp://ftp.ensembl.org/pub/release-100/fasta/homo_sapiens/ncrna/Homo_sapiens.GRCh38.ncrna.fa.gz

- SARS-CoV-2, SARS, MERS, and H1N1 genomes, available at https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/

- SARS-CoV-2 Wuhan strain from NCBI Assembly (code ASM985889v3), available at https://www.ncbi.nlm.nih.gov/assembly/GCF_009858895.2/

### Step 2: Scripts acquisition

Download the scripts that will be used to extract features from siRNAs sequences:

- Auxiliary scripts and files implemented by authors at https://github.com/inaciomdrs/sirna_db_building_protocol

- Softwares ThermoComposition21[8] and si_shRNA_selector[9], whose copies are available at https://github.com/inaciomdrs/sirna_db_building_protocol/bin

*Disclaimers*

- Scripts files OligoCalc.py, OligoCalcCompare.py, SelfAnnealingSites.py, complementarity3.py, and HairpinCalc.py are Python translations from original JavaScript scripts used in OligoCalc webserver[6], whose codes are open-source;

- Script file deltacalculator.py is a modified part of SSD7 software destined for siRNA efficiency prediction and features calculation.

### Step 3: Strains cleaning

To assess siRNAs' efficiency against SARS-CoV-2 strains from different countries, the genomes from these strains must be at least 90% complete. Thus, remove from SARS-CoV-2 GISAID strains genomes the ones whose percentage of N-type nucleotides are higher or equal to 10 (not-closed regions). We recommend using ref_clean.pl script for performing this task. An example of how to use this script follows below:

```
$ ./ref_clean.pl uk_sars_cov_2_strains.fa > uk_sars_cov_2_clean.fa
```

This step must also be performed on SARS, MERS, and H1N1 genomes.

### Step 4: Genomes indexing

Index all the genomes downloaded in Step 1 with Bowtie[10] version 1.1.0. An example of how to make this indexing follows below:

```
$ bowtie-build uk_sars_cov_2_clean.fa uk_sars_cov_2_clean
```

This step must be performed for every genome downloaded in Step 1.

### Step 5: Path settings and siRNAs sequences generation

Edit step_0_seq.pl script file informing, where indicated, the path to the fasta file of SARS-CoV-2 Wuhan strain from NCBI Assembly (code ASM985889v3), downloaded in Step 1. After that, edit aln_commands.pl script file informing, where indicated, the paths of prefixes of indexed genomes in Step 4. Then, run the following commands:

```
$ ./step_0_seq.pl 21  > input

$ ./aln_commands.pl input 21 > aln_commands_21.sh
```

Where step_0_seq.pl generates siRNAs of 21 nucleotides length, saving them on input; and aln_commands.pl generates a shell-script file responsible for executing the alignments of generated siRNAs sequences against indexed genomes, producing STS files that report the minimum number of needed mismatches for those siRNAs to have a match with those genomes.

After that, the user has the option of generating sequences between 18 and 21-nt long by changing the 21 in the above commands by the size of interest. Generated aln_commands_21.sh script file can be either run sequentially (one command at a time) or by batches of commands run in parallel. Finally, create a directory called STS and move all generated STS files to it.

**Step 6: Database building**

Run script file run.pl for generating siRNA targets database, using the following command:

```
$ ./run.pl 21
```

Where 21 is the length of siRNAs sequences. The user must assure that this number is the same as the used in Step 5, at step_0_seq.pl script. Important note: the user also must ensure that run.pl script, bin directory, STS folder, NC_045512.bed file, and db_olig_seq2.pl script file are in the same folder. All of these files are available at https://github.com/inaciomdrs/sirna_db_building_protocol. run.pl calculates features regarding base context, thermodynamic information, and efficiency prediction, using for these latter two ThermoComposition21[8] and si_shRNA_selector[9] software and the other downloaded scripts in Step 2 for the whole process of database building. It also organizes information in STS files across the produced table. It's important to note that run.pl triggers processes that are run in background. Use top program to track the execution of such processes and know when they are finished. When this finally happens, execute the following two commands:

```
$ cat *.res > db.txt

$ rm input.*
```

Where db.txt is the generated siRNAs database for the chosen size.

# Troubleshooting

Step 5 is the longest one and requires more computational effort. Based on our experience designing this protocol, it is common that this step fills out either a huge RAM memory volume either a huge disk space. In face of that, aln_commands_21.sh script file is optimized for saving disk space. In the case that the user wants to save RAM memory, he or she needs to replace every line with the following structure:

```
bowtie -S -a –pairtries 4 -p 30 -n 3 -e 10 -l 7 genome -f NAT.21 | ./summarize.pl - > NAT.genome.21.sts
```

By these two:

```
bowtie -S -a –pairtries 4 -p 30 -n 3 -e 10 -l 7 genome -f NAT.21 > NAT.21.genome.sam
```

```
./summarize.pl NAT.21.genome.sam > NAT.genome.21.sts
```

## Time Taken

The alignment step is the longest one and dependent on the computational power of the user’s machine. Aside from that, database building lasts about thirty minutes, considering the described machine at the Equipment section.

## References

1. Qureshi, A., Tantray, V. G., Kirmani, A. R. & Ahangar, A. G. A review on current status of antiviral siRNA. Reviews in Medical Virology vol. 28 e1976 (2018).

2. Ghosh, S., Firdous, S. M. & Nath, A. siRNA could be a potential therapy for COVID-19. EXCLI J. 19, 528–531 (2020).

3. Liu, C. et al. Research and Development on Therapeutic Agents and Vaccines for COVID-19 and Related Human Coronavirus Diseases. ACS Cent Sci 6, 315–331 (2020).

4. Shi, Y. et al. Inhibition of genes expression of SARS coronavirus by synthetic small interfering RNAs. Cell Res. 15, 193–200 (2005).

5. Wang, Z. et al. Inhibition of severe acute respiratory syndrome virus replication by small interfering RNAs in mammalian cells. J. Virol. 78, 7523–7527 (2004).

6. Kibbe, W. A. OligoCalc: an online oligonucleotide properties calculator. Nucleic Acids Res. 35, W43–6 (2007).

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

8. Shabalina, S. A., Spiridonov, A. N. & Ogurtsov, A. Y. Computational models with thermodynamic and composition features improve siRNA design. BMC Bioinformatics 7, 65 (2006).

9. Matveeva, O. V. et al. Optimization of duplex stability and terminal asymmetry for shRNA design. PLoS One 5, e10180 (2010).

10. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25 (2009).