bioRxiv preprint doi: https://doi.org/10.1101/2021.11.10.468129; this version posted November 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

CORSID enables *de novo* identification of transcription regulatory sequences and genes in coronaviruses

Chuanyi Zhang[1,*], Palash Sashittal[2,*], and Mohammed El-Kebir[2,†]

[1]Department of Electrical & Computer Engineering, University of Illinois at Urbana-Champaign, IL 61801
[2]Department of Computer Science, University of Illinois at Urbana-Champaign, IL 61801
[*]Joint first authorship
[†]Correspondence: melkebir@illinois.edu

Abstract

Genes in coronaviruses are preceded by transcription regulatory sequences (TRSs), which play a critical role in gene expression mediated by the viral RNA-dependent RNA-polymerase via the process of discontinuous transcription. In addition to being crucial for our understanding of the regulation and expression of coronavirus genes, we demonstrate for the first time how TRSs can be leveraged to identify gene locations in the coronavirus genome. To that end, we formulate the TRS AND GENE IDENTIFICATION (TRS-GENE-ID) problem of simultaneously identifying TRS sites and gene locations in unannotated coronavirus genomes. We introduce CORSID (CORe Sequence IDentifier), a computational tool to solve this problem. We also present CORSID-A, which solves a constrained version of the TRS-GENE-ID problem, the TRS IDENTIFICATION (TRS-ID) problem, identifying TRS sites in a coronavirus genome with specified gene annotations. We show that CORSID-A outperforms existing motif-based methods in identifying TRS sites in coronaviruses and that CORSID outperforms state-of-the-art gene finding methods in finding genes in coronavirus genomes. We demonstrate that CORSID enables *de novo* identification of TRS sites and genes in previously unannotated coronaviruses. CORSID is the first method to perform accurate and simultaneous identification of TRS sites and genes in coronavirus genomes without the use of any prior information.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.10.468129; this version posted November 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

1 Introduction

Coronaviruses are comprised of a single-stranded RNA genome that is ready to be translated by the host ribosome (Fig. 1a). While the majority of messenger RNA (mRNA) in eukaryotes is *monocistronic*, *i.e.* each mRNA is translated into a single gene product, the coronavirus RNA genome is comprised of many genes, which are expressed and translated using two distinct mechanisms [15]. First, upon entry, the viral genome is translated to produce polypeptides corresponding to one or two overlapping open reading frames (ORFs). Second, the resulting polypeptides undergo auto-cleavage, producing many non-structural proteins, including RNA-dependent-RNA-polymerase (RdRP), which mediates the expression of the remaining viral genes via *discontinuous transcription* [22]. That is, RdRP is prone to perform template switching upon encountering *transcription regulatory sequences* (TRSs) located in the 5' untranslated region (UTR) of the genome — called TRS-L where L stands for leader — and upstream of each viral gene — called TRS-B where B stands for body (Fig. 1b). This mechanism yields multiple subgenomic mRNAs that are translated into the structural and accessory viral proteins, necessary for the viral life cycle. Not only is the identification and characterization of TRS sites crucial to understanding the regulation and expression of the viral proteins, but here we hypothesize that the existence of these regulatory sequences can be leveraged to simultaneously identify TRS sites and associated viral genes in unannotated coronavirus genomes with high accuracy.

While there exist methods for identifying either TRS sites or viral genes, no method exists that does so simultaneously (Table S1). More specifically, since TRSs contain $6-7$ nt long conserved sequences, called *core sequences* [8,25], general-purpose motif finding methods [2,7,20,29] can be employed to identify TRS-L and TRS-Bs in coronaviruses. For instance, MEME [2] is a widely used method that employs expectation maximization to identify multiple appearances of multiple motifs simultaneously. The only method that is specifically developed for identifying TRS sites in coronaviruses is SuPER [28], which takes as input a coronavirus genome sequence with specified gene locations as well as additional taxonomic and secondary structure information. Importantly, SuPER as well as other general-purpose motif finding algorithms are unable to identify viral genes in unannotated coronavirus genome sequences. On the other hand, gene prediction is a well-studied problem with many methods including Glimmer3 [5,21] and Prodigal [12,13]. Glimmer3 uses a Markov model to assign scores to ORFs, and then processes overlapping genes to generate the final list of predicted genes. By contrast, Prodigal employs a more heuristic approach with fine-tuned parameters that are optimized to identify genes in prokaryotes. However, these general-purpose gene finding tools are not designed to leverage the genomic structure of coronaviruses, specifically the TRS sites located upstream of the genes in the genome, nor are they able to directly identify these regulatory sequences.

In this study, we introduce the TRS IDENTIFICATION (TRS-ID) and the TRS AND GENE IDENTIFICATION (TRS-GENE-ID) problems, to identify TRS sites in a coronavirus genome with specified gene annotations, and to simultaneously identifying TRS sites and genes in an unannotated coronavirus genome, respectively (Fig. 1c). Underpinning our approach is the concept of a *TRS alignment*, which is a multiple sequence alignment of TRS sites with additional constraints that result from template switching by RdRP. We introduce CORSID-A, a dynamic programming (DP) algorithm to solve the TRS-ID problem, adapting the recurrence that underlies the Smith-Waterman algorithm [24] for local sequence alignment. Additionally, we introduce CORSID to solve the TRS-GENE-ID problem via a maximum-weight independent set problem [11] on an interval graph defined by the candidate ORFs in the genome with weights obtained from the previous DP. We evaluate the performance of our methods on 468 coronavirus genomes downloaded from GenBank, demonstrating that CORSID-A outperforms MEME and SuPER in identifying TRS sites and, unlike these methods, possesses the ability to identify recombination events. Moreover, we find that CORSID vastly outperforms state-of-the-art gene finding methods. Finally, we illustrate how CORSID enables *de novo* identification of TRS sites and genes in previously unannotated coronaviruses. In summary, CORSID is the first method to perform accurate and simultaneous identification of TRS sites and genes in coronavirus genomes without the use of prior taxonomic or secondary structure information.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.10.468129; this version posted November 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

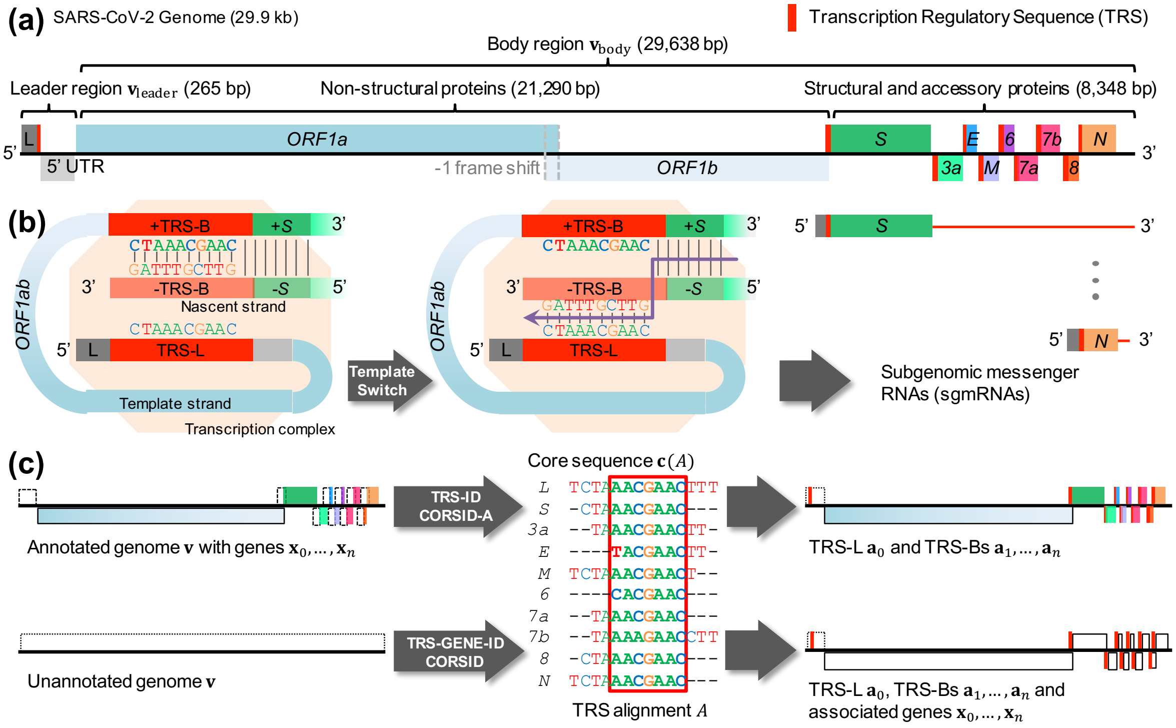

(a)
SARS-CoV-2 Genome (29.9 kb)
Transcription Regulatory Sequence (TRS)
Body region $\mathbf{v}_{\text{body}}$ (29,638 bp)
Leader region $\mathbf{v}_{\text{leader}}$ (265 bp)
Non-structural proteins (21,290 bp)
Structural and accessory proteins (8,348 bp)
5'
L
ORF1a
S
E
6
7b
N
3'
5' UTR
-1 frame shift
ORF1b
3a
M
7a
8
(b)
+TRS-B
+S
3'
CTAAACGAAC
GATTTGCTTG
3'
-TRS-B
-S
5'
Nascent strand
CTAAACGAAC
ORF1ab
5'
L
TRS-L
Template strand
Transcription complex
Template Switch
+TRS-B
+S
3'
CTAAACGAAC
3'
-TRS-B
-S
5'
GATTTGCTTG
CTAAACGAAC
ORF1ab
5'
L
TRS-L
5'
S
3'
5'
N
3'
Subgenomic messenger RNAs (sgmRNAs)
(c)
Annotated genome $\mathbf{v}$ with genes $\mathbf{x}_0, \dots, \mathbf{x}_n$
TRS-ID CORSID-A
Core sequence $\mathbf{c}(A)$
L TCTAAACGAACTTT
S -CTAAACGAAC---
3a --TAAACGAACTT-
E ----TACGAACTT-
M TCTAAACGAACT--
6 ----CACGAAC---
7a --TAAACGAAC---
7b --TAAAAGAACCTT
8 -CTAAACGAAC---
N TCTAAACGAAC---
TRS alignment A
TRS-L $\mathbf{a}_0$ and TRS-Bs $\mathbf{a}_1, \dots, \mathbf{a}_n$
Unannotated genome $\mathbf{v}$
TRS-GENE-ID CORSID
TRS-L $\mathbf{a}_0$, TRS-Bs $\mathbf{a}_1, \dots, \mathbf{a}_n$ and associated genes $\mathbf{x}_0, \dots, \mathbf{x}_n$

Figure 1: Overview. (a) A coronavirus genome $\mathbf{v}$ consists of a leader region $\mathbf{v}_{\text{leader}}$ and a body region $\mathbf{v}_{\text{body}}$. (b) Structural and accessory genes are expressed via discontinuous transcription with template switching occurring at transcription regulatory sequences (TRS, indicated in red), resulting in subgenomic messenger RNAs (sgmRNAs) for each gene. (c) In the TRS IDENTIFICATION (TRS-ID) problem, we wish to identify TRSs given a genome $\mathbf{v}$ with genes $\mathbf{x}_0, \dots, \mathbf{x}_n$. The TRS AND GENE IDENTIFICATION (TRS-GENE-ID) asks to simultaneously identify genes and their associated TRSs given genome $\mathbf{v}$. Throughout this manuscript we use 'T' (thymine) rather than 'U' (uracil).

2 Problem Statement

We begin by introducing notation and key definitions (Section 2.1), followed by stating the TRS IDENTIFICATION problem (Section 2.2) and then the TRS AND GENE IDENTIFICATION problem (Section 2.3).

2.1 Preliminaries

A *genome* $\mathbf{v} = v_1 \dots v_{|\mathbf{v}|}$ is a sequence from the alphabet $\Sigma = \{\mathrm{A}, \mathrm{T}, \mathrm{C}, \mathrm{G}\}$. The first position of the genome is known as the *5' end* whereas the last position of the genome is known as the *3' end*. We denote the contiguous subsequence $v_p \dots v_q$ of $\mathbf{v}$ by $\mathbf{v}[p, q]$. We call a contiguous subsequence $\mathbf{x}$ of $\mathbf{v}$ also a *region*, denoted as $\mathbf{x} = [x^-, x^+]$ such that $\mathbf{x} = \mathbf{v}[x^-, x^+]$. Thus, coordinates x^- and x^+ of a subsequence $\mathbf{x}$ are in terms of the reference genome $\mathbf{v}$, *i.e.* $\mathbf{x} = v_{x^-} \dots v_{x^+}$. Alternatively, we may refer to individual characters in a subsequence $\mathbf{x}$ using relative indices, *i.e.* $\mathbf{x} = x_1 \dots x_{|\mathbf{x}|}$. Our goals are twofold: given a coronavirus genome $\mathbf{v}$, we aim to identify (i) TRS-L and TRS-Bs, and optionally, (ii) the associated genes (Fig. 1c). To begin, recall the following definition of an alignment.

Definition 1. Matrix $A = [a_{ij}]$ with $n + 1$ rows is an *alignment* of sequences $\mathbf{w}_0, \dots, \mathbf{w}_n \in \Sigma^*$ provided (i) entries a_{ij} either correspond to a letter in the alphabet Σ or a gap denoted by '$-$' such that (ii) no column of A is composed of only gaps, and (iii) the removal of gaps of row i of A yields sequence $\mathbf{w}_i$.

Here, we seek an alignment with two additional constraints, called a TRS alignment defined as follows.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.10.468129; this version posted November 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Definition 2. An alignment $A = [\mathbf{a}_0, \ldots, \mathbf{a}_n]^T$ is a *TRS alignment* provided (i) $\mathbf{a}_0$ does not contain any gaps, and (ii) $\mathbf{a}_1, \ldots, \mathbf{a}_n$ do not contain any internal gaps.

Intuitively, the first sequence $\mathbf{a}_0$ in the alignment A represents TRS-L, whereas $\mathbf{a}_1, \ldots, \mathbf{a}_n$ represent TRS-Bs, each upstream of an accessory or structural gene. We do not allow gaps in the TRS-L sequence $\mathbf{a}_0$ as template switching by RdRP occurs due to complementary base pairing between TRS-L and the nascent strand of TRS-B [26]. For the same reason, we do not allow internal gaps in TRS-Bs $\mathbf{a}_i$. However, as each TRS-B may match a different region of the TRS-L, we do allow flanking gaps in these sequences (Fig. 1c). We score a TRS alignment A using a scoring function $\delta : \Sigma \times (\Sigma \cup \{-\}) \to \mathbb{R}$ in the following way.

Definition 3. The *score* $s(A)$ of a TRS alignment $A = [\mathbf{a}_0, \ldots, \mathbf{a}_n]^T$ is given by $\sum_{i=1}^{n} s(\mathbf{a}_0, \mathbf{a}_i) = \sum_{i=1}^{n} \sum_{j=1}^{|\mathbf{a}_0|} \delta(a_{0j}, a_{ij})$, whereas the *minimum score* $s_{\min}(A)$ is defined as $\min_{i \in \{1,\ldots,n\}} s(\mathbf{a}_0, \mathbf{a}_i)$.

In other words, we score each TRS-B $\mathbf{a}_i$ (where $i \geq 1$) by comparing it to the TRS-L sequence $\mathbf{a}_0$ in a way that is consistent with the mechanism of template switching during discontinuous transcription. As such, our scoring function differs from the traditional sum-of-pairs scoring function [3] where every unordered pair $(\mathbf{a}_i, \mathbf{a}_j)$ of sequences contributes to the score of the alignment. Furthermore, each TRS alignment uniquely determines the core sequence as follows.

Definition 4. Sequence $\mathbf{c}(A)$ is the *core sequence* of a TRS alignment $A = [\mathbf{a}_0, \ldots, \mathbf{a}_n]^T$ provided $\mathbf{c}(A)$ is the largest contiguous subsequence of $\mathbf{a}_0$ such that no character of $\mathbf{c}$ is aligned to a gap in any of $\mathbf{a}_1, \ldots, \mathbf{a}_n$.

Note that the core sequence is a subsequence of the TRS sequences. As such, the TRS alignment can include nucleotides immediately flanking the core sequence, which have been shown to play an important role in discontinuous transcription in previous experiments [26].

2.2 The TRS IDENTIFICATION problem

The first problem we consider is that of identifying TRS sites given a viral genome with known genes $\mathbf{x}_0, \ldots, \mathbf{x}_n$. Specifically, we are given a candidate region $\mathbf{w}_0$ that contains the unknown TRS-L $\mathbf{a}_0$ upstream of gene $\mathbf{x}_0$ as well as candidate regions $\mathbf{w}_1, \ldots, \mathbf{w}_n$ that contain the unknown TRS-Bs $\mathbf{a}_1, \ldots, \mathbf{a}_n$ of genes $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Section 3.1 details how to obtain these candidate regions when only given the gene locations. To further guide the optimization problem, we impose an additional constraint on the sought TRS alignment A in the form of a minimum length ω on the core sequence $\mathbf{c}(A)$ as well as a threshold τ on the minimum score $s_{\min}(A)$ of the TRS alignment. We formalize this problem as follows.

Problem 1 (TRS IDENTIFICATION (TRS-ID))**.** Given non-overlapping sequences $\mathbf{w}_0, \ldots, \mathbf{w}_n$, core-sequence length $\omega > 0$ and score threshold $\tau > 0$, find a TRS alignment $A = [\mathbf{a}_0, \ldots, \mathbf{a}_n]^T$ such that (i) $\mathbf{a}_i$ corresponds to a subsequence in $\mathbf{w}_i$ for all $i \in \{0, \ldots, n\}$, (ii) the core sequence $\mathbf{c}(A)$ has length at least ω, (iii) the minimum score $s_{\min}(A)$ is at least τ, and (iv) the alignment has maximum score $s(A)$.

2.3 The TRS AND GENE IDENTIFICATION problem

In the second problem, we are no longer given an annotated genome with gene locations. Rather, we seek to simultaneously identify genes and TRS sites given a viral genome sequence $\mathbf{v}$ split into a leader region $\mathbf{v}_{\text{leader}}$ and body region $\mathbf{v}_{\text{body}}$. Section 3.2 describes a heuristic for identifying these two regions when only given $\mathbf{v}$. The key idea here is that each TRS alignment will uniquely determine a set of genes it encodes. To make this relationship clear, we begin by defining an open reading frame as follows.

Definition 5. A contiguous subsequence $\mathbf{x} = [x^-, x^+]$ of $\mathbf{v}$ is an *open reading frame* provided $\mathbf{x}$ (i) has a length $|\mathbf{x}|$ that is a multiple of 3, (ii) starts with a start codon, *i.e.* $x_1 \ldots x_3 = \text{ATG}$, and (iii) ends at a stop codon, *i.e.* $x_{|\mathbf{x}|-2} \ldots x_{|\mathbf{x}|} \in \{\text{ATG}, \text{TAG}, \text{TGA}\}$.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.10.468129; this version posted November 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Each TRS-B $\mathbf{a}_i$ is associated with at most one ORF that occurs immediately downstream of $\mathbf{a}_i$. Naively, to identify the ORF associated with $\mathbf{a}_i$, one could simply scan downstream of the TRS-B for the first occurrence of a start codon and continue scanning to identify the corresponding in-frame stop codon. However, this would not take ribosomal leaky scanning into account, where the ribosome does not initiate translation at the first encountered 'ATG'. Section 3.2 provides a more robust definition of a downstream ORF that takes ribosomal leaky scanning into account. To summarize, we have that a TRS alignment $A = [\mathbf{a}_0, \ldots, \mathbf{a}_n]^T$ uniquely determines a set $\Gamma(A)$ of candidate genes.

Definition 6. A set $\Gamma(A)$ of ORFs are *induced genes* of a TRS alignment $A = [\mathbf{a}_0, \ldots, \mathbf{a}_n]^T$ provided $\Gamma(A)$ is composed of the ORFs that occur downstream of each TRS-B $\mathbf{a}_1, \ldots, \mathbf{a}_n$ in $\mathbf{v}_{\text{body}}$.

Note that multiple TRS-Bs of a TRS alignment $A = [\mathbf{a}_0, \ldots, \mathbf{a}_n]^T$ can induce the same gene in $\mathbf{v}_{\text{body}}$. Moreover, there may *not* be an ORF downstream of a TRS-B $\mathbf{a}_i$. As such, we have that $|\Gamma(A)| \leq n$. By contrast, in coronaviruses, each viral gene typically has a unique TRS-B. Moreover, these viral genes are typically non-overlapping in the genome. Finally, coronavirus genomes tend to be compact with most positions coding for genes. To capture these biological constraints, we introduce the following definitions.

Definition 7. A TRS alignment $A = [\mathbf{a}_0, \ldots, \mathbf{a}_n]^T$ is *concordant* provided (i) each TRS-B $\mathbf{a}_i$ corresponds to exactly one gene in $\Gamma(A)$, and (ii) there are no two ORFs in $\Gamma(A)$ whose positions in $\mathbf{v}_{\text{body}}$ overlap.

Definition 8. The *genome coverage* $g(A)$ of a TRS alignment A is the number of positions in $\mathbf{v}_{\text{body}}$ that are covered by the set $\Gamma(A)$ of induced genes.

This leads to the following problem.

Problem 2 (TRS AND GENE IDENTIFICATION (TRS-GENE-ID))**.** Given leader region $\mathbf{v}_{\text{leader}}$, body region $\mathbf{v}_{\text{body}}$, core-sequence length $\omega > 0$ and score threshold $\tau > 0$, find a TRS alignment $A = [\mathbf{a}_i]$ such that (i) $\mathbf{a}_0$ corresponds to a subsequence in $\mathbf{v}_{\text{leader}}$, (ii) $\mathbf{a}_i$ corresponds to a subsequence in $\mathbf{v}_{\text{body}}$ for all $i \geq 1$, (iii) the core sequence $\mathbf{c}(A)$ has length at least ω, (iv) the minimum score $s_{\min}(A)$ is at least τ, (v) A is concordant, and (vi) A induces a set $\Gamma(A)$ of genes with maximum genome coverage $g(A)$ and A subsequently has maximum score $s(A)$.

3 Methods

Section 3.1 introduces CORSID-A, which solves the TRS-ID problem. Section 3.2 introduces CORSID, solving the TRS-GENE-ID problem. Both sections discuss practical considerations as well as several heuristics for obtaining the required input to each problem. We implemented both methods in Python. The source code is available at `https://github.com/elkebir-group/CORSID`.

3.1 Solving the TRS IDENTIFICATION problem

Recall that in the TRS-ID problem we seek a TRS alignment A given input candidate regions sequences $\mathbf{w}_0, \ldots, \mathbf{w}_n$ that each occur upstream of genes $\mathbf{x}_0, \ldots, \mathbf{x}_n$. Intuitively, we define the candidate region for a gene $\mathbf{x}_i$ as the region $\mathbf{w}_i = [w_i^-, w_i^+]$ composed of positions $w^- \leq p \leq w^+$ such that any sgmRNA starting at p will lead to the translation of ORF $\mathbf{x}_i$ by the ribosome. SuPER [28], the only other method for identifying TRSs in annotated coronavirus genomes, employs a heuristic by defining the candidate region $\mathbf{w}_i$ of a gene $\mathbf{x}_i$ as $v_{x^- - 170} \ldots v_{x^- - 1}$, *i.e.* the candidate region $\mathbf{w}_i$ is a subsequence of 170 nt immediately upstream of gene $\mathbf{x}_i$ for $1 \leq i \leq n$. Here, we take a more rigorous and flexible approach that takes ribosomal leaky scanning into account by skipping over previous ORFs with length smaller than 100 nt (details in Appendix A.1).

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.10.468129; this version posted November 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

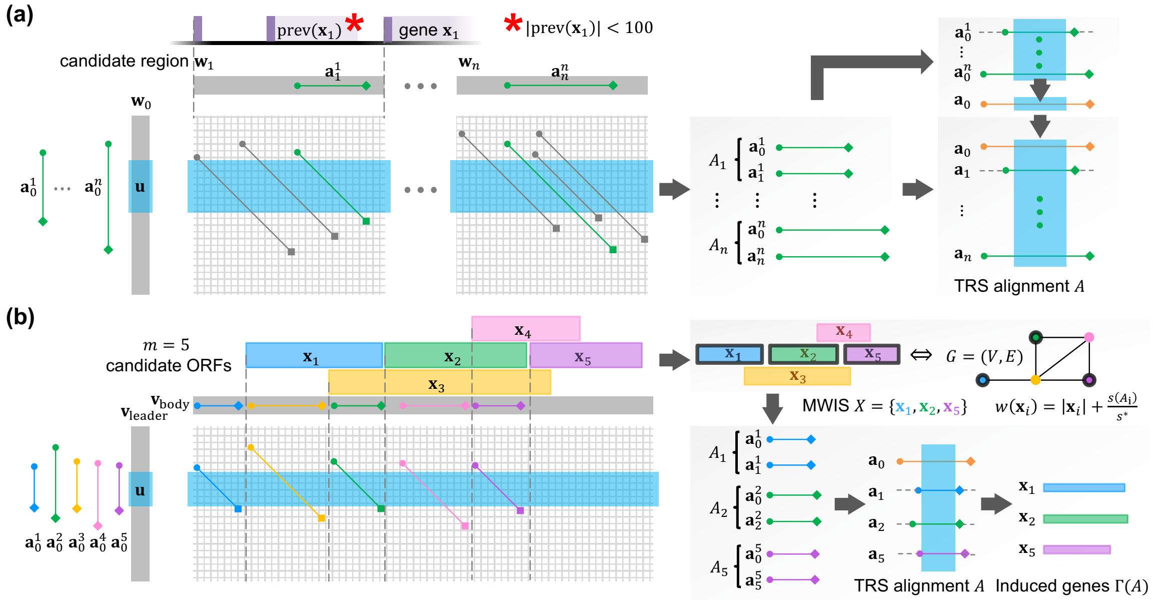


Figure 2: Algorithm details. (a) Given genes $\mathbf{x}_0, \ldots, \mathbf{x}_n$, we obtain candidate regions $\mathbf{w}_0, \ldots, \mathbf{w}_n$ by identifying upstream ORFs, skipping over ORFs if they are of length less than 100 nt (indicated by '*'). CORSID-A solves the TRS-ID problem by sliding a window $\mathbf{u}$ through $\mathbf{w}_0$, solving n independent pair-wise dynamic programming problems, which together yield the optimal TRS alignment A for window $\mathbf{u}$. (b) To solve the TRS-GENE-ID problem, CORSID additionally solves a maximum-weight independent set problem [11] on an interval graph defined by the candidate ORFs to simultaneously identify an optimal pair $(A, \Gamma(A))$ for window $\mathbf{u}$.

Recall that in a TRS alignment $A = [\mathbf{a}_0, \ldots, \mathbf{a}_n]^T$ only the TRS-Bs $\mathbf{a}_1, \ldots, \mathbf{a}_n$ are allowed to have gaps (restricted to the flanks), and that the TRS-L $\mathbf{a}_0$ is gapless. To score a TRS alignment, we use a simple scoring function $\delta : \Sigma \times (\Sigma \cup \{-\}) \to \mathbb{R}$ such that $s(x, y)$ equals $+1$ for matches (*i.e.* $x = y$), -2 for mismatches (*i.e.* $x \neq y$ and $y \neq -$), and 0 for gaps (*i.e.* $y = -$). In other words, while we reward matches and penalize mismatches, we do not penalize flanking gaps.

Recall that the sought TRS alignment A must induce a core sequence $\mathbf{c}(A)$ of length at least ω. Due to this constraint, the input sequences $\mathbf{w}_0, \ldots, \mathbf{w}_n$ depend on one another and cannot be considered in isolation. We break this dependency by considering a subsequence $\mathbf{u}$ within $\mathbf{w}_0$ of length ω, restricting the induced core sequence $\mathbf{c}(A)$ of output TRS alignments A to contain $\mathbf{u}$. We solve this constrained version of the TRS-ID problem using dynamic programming in time $O(|\mathbf{w}_0|L)$ where L is the total length of candidate regions $\mathbf{w}_1, \ldots, \mathbf{w}_n$ (details are in Appendix A.2 and Fig. 2a). We obtain the solution to the original TRS-ID problem by identifying the window $\mathbf{u}$ that induces a TRS alignment A with maximum score. As there are $O(|\mathbf{w}_0|)$ windows in $\mathbf{w}_0$ of fixed length ω, this procedure takes $O(|\mathbf{w}_0|^2 L)$ time.

3.2 Solving the TRS AND GENE IDENTIFICATION problem

In the TRS-GENE-ID problem, we require two sequences: $\mathbf{v}_{\text{leader}}$ which contains TRS-L $\mathbf{a}_0$ and $\mathbf{v}_{\text{body}}$ which contains each TRS-B $\mathbf{a}_1, \ldots, \mathbf{a}_n$. We describe a heuristic to partition a genome $\mathbf{v}$ into $\mathbf{v}_{\text{leader}}$ and $\mathbf{v}_{\text{body}}$ in Appendix A.3. This heuristic is performed in $O(m^2)$ time where m is the number of ORFs in $\mathbf{v}$.

We will now define the relationship between a TRS alignment $A = [\mathbf{a}_0, \ldots, \mathbf{a}_n]^T$ and the set $\Gamma(A)$ of induced genes. Upon removing (flanking) gaps, each aligned sequence $\mathbf{a}_i$ corresponds to a contiguous subsequence $\mathbf{v}_i$ of the viral genome $\mathbf{v}$. Specifically, $\mathbf{v}_0$ occurs in $\mathbf{v}_{\text{leader}}$ and $\mathbf{v}_i$ occurs in $\mathbf{v}_{\text{body}}$ (where $i \geq 1$). By Definition 4, each subsequence $\mathbf{v}_i$ has positions that are aligned with the core sequence $\mathbf{c}(A)$.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.10.468129; this version posted November 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

These aligned positions induce the subsequence $\mathbf{c}_i = [c_i^-, c_i^+]$ of length equal to $|\mathbf{c}(A)|$. Note that while $\mathbf{c}_0 = \mathbf{c}(A)$, it may be that $\mathbf{c}_i \neq \mathbf{c}(A)$ where $i \geq 1$ due to mismatches. Importantly, there are coronaviruses where the last three nucleotides of the core sequence within a TRS-B coincide with the start codon of the associated gene (Fig. S1). As such, we have the following definition.

Definition 9. Let $A = [\mathbf{a}_0, \ldots, \mathbf{a}_n]^T$ be a TRS alignment and let $\mathbf{c}_i = [c_i^-, c_i^+]$ be the subsequence of $\mathbf{a}_i$ that is aligned to the core sequence $\mathbf{c}(A)$. The ORF *associated* with TRS-B $\mathbf{a}_i$ is the unique ORF $\mathbf{x}$ where position $c_i^+ - 2$ occurs within the candidate region of $\mathbf{x}$.

As discussed, there may not exist an ORF associated with a TRS-B $\mathbf{a}_i$, which may happen when the TRS-B is located near the 3' end of the genome. Given a TRS alignment $A = [\mathbf{a}_0, \ldots, \mathbf{a}_n]^T$, the set $\Gamma(A)$ of induced genes equals the set of ORFs that are associated with $\mathbf{a}_1, \ldots, \mathbf{a}_n$.

To solve the TRS-GENE-ID problem, we take a similar sliding window approach that we used to solve the TRS-ID problem. That is, we consider all subsequences $\mathbf{u}$ within $\mathbf{v}_{\text{leader}}$ of length ω and solve a constrained version of the TRS-GENE-ID problem, additionally requiring that the sought TRS alignment A has a core sequence $\mathbf{c}(A)$ that fully contains $\mathbf{u}$, using the following two steps. First, we construct a DP table similar to the previous table used in TRS-ID problem in $O(|\mathbf{v}_{\text{leader}}||\mathbf{v}_{\text{body}}|)$ time, and for each ORF, we select the alignment with the highest score in the corresponding candidate region. Second, given these ORFs and corresponding alignments, we build a vertex-weighted interval graph combining ORF lengths and alignment scores as weights. To identify the optimal TRS alignment A and associated genes $\Gamma(A)$, we solve a maximum-weight independent set (MWIS) on this graph in $O(m)$ time, where m is the number of candidate ORFs in $\mathbf{v}_{\text{body}}$ (Appendix A.4 and Fig. 2b). Each instance of the constrained TRS-GENE-ID problem takes $O(|\mathbf{v}_{\text{leader}}||\mathbf{v}_{\text{body}}| + m)$ time. Since the number of windows of length ω in $\mathbf{v}_{\text{leader}}$ is $O(|\mathbf{v}_{\text{leader}}|)$, the total running time of CORSID to solve the TRS-GENE-ID problem is $O(|\mathbf{v}_{\text{leader}}|^2|\mathbf{v}_{\text{body}}| + |\mathbf{v}_{\text{leader}}|m)$. In practice, the number m of candidate ORFs in $\mathbf{v}_{\text{body}}$ ranges from $21 - 92$, the length $|\mathbf{v}_{\text{leader}}|$ of leader region ranges from $171 - 716$ and the length $|\mathbf{v}_{\text{body}}|$ of the body region ranges from $6280 - 11462$ across all the coronaviruses studied in this paper. Finally, to obtain biologically meaningful solutions, we employ a progressive approach and consider overlapping genes (see Appendix A.5 for details).

4 Results

To evaluate the performance of CORSID-A and CORSID, we downloaded the same set of 505 assembled coronavirus genomes previously analyzed by SuPER [28] from GenBank along with their annotation GFF files, indicating gene locations. To benchmark methods for the TRS-ID problem, we assessed each method's ability to correctly identify TRS-L as well as identify a TRS-B upstream of each gene. For the TRS-GENE-ID problem, we additionally assessed each method's ability to identify ground-truth genes. Appendix B.1 describes how we established the set of genes and locations of TRS sites in the coronavirus genomes. We excluded 35 genomes due to incomplete leader sequences, thus lacking TRS-L. We excluded two more genomes due to empty GFF files, thus lacking gene annotations. The remaining 468 genomes comprised all four genera of the *Coronaviridae* family and spanned a total of 22 subgenera (Table S2).

4.1 CORSID-A finds TRS-L and TRS-Bs with higher accuracy than existing methods

We begin by comparing the performance of CORSID-A with MEME and SuPER for the TRS IDENTIFICATION problem. Recall that MEME is a general-purpose motif detection algorithm [2], whereas SuPER is specifically designed for identifying core sequences within coronavirus genomes annotated with genes [28]. To run CORSID-A, we extracted candidate regions $\mathbf{w}_1, \ldots, \mathbf{w}_n$ upstream of annotated genes $\mathbf{x}_0, \ldots, \mathbf{x}_n$ as described in Definition 11. The minimum length ω of core sequence is set to 7 following existing literature [8], and we use a minimum alignment score of $\tau = 2$. We provided MEME with the same candidate

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.10.468129; this version posted November 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

regions $\mathbf{w}_0, \ldots, \mathbf{w}_n$, and ran it in "zero or one occurrence per sequence" mode. As for SuPER, we analyzed the previously reported results on the same 468 sequences considered here. Detailed commands and parameters can be found in Appendix B.2.

As shown in Fig. 3b, CORSID-A correctly identified TRS-Ls in 466 out of 468 genomes, reaching a higher accuracy (99.6%) than MEME (442 genomes, 94.4%), but was outperformed by SuPER, which was correct in 467 genomes (99.8%). The two genomes where our method failed are outliers in their respective subgenera, indicative of possible sequencing errors (Appendix B.4). Fig. S2 discusses the one genome where SuPER failed to identify TRS-L correctly, showing that the TRS-L sequence identified by our method is supported by both secondary structure information as well as a split read in a corresponding RNA sequencing sample. *Split reads* map to non-contiguous regions of the viral genome and provide direct evidence of template switching at TRS sites.

Of note, SuPER uses additional information to identify TRS-L and TRS-B sites compared to MEME and CORSID-A. That is, SuPER requires the user to specify the genus of origin for each input sequence, which is used to obtain a genus-specific motif of the core sequence from a look-up table. This motif is used to identify matches along the genome. In addition, SuPER takes as input the 5' UTR secondary structure, restricting the region in which the TRS-L occurs until the fourth stem loop (SL4). Importantly, while CORSID-A does not rely on any prespecified motif, taxonomic or secondary structure information, our method identified more TRS-Bs than either SuPER or MEME (Fig. 3c). Specifically, we define the *TRS-B recall* as the fraction of genes for which TRS-Bs were identified. While the median TRS-B recall of all three methods is 1, CORSID-A found putative TRS-Bs of all genes in 387 genomes (82.7%), while SuPER and MEME did so in only 290 (62.0%) and 315 (67.3%) genomes, respectively.

To validate the identified TRS sites, we examined split reads in publicly available RNA-sequencing data of cells infected by coronaviruses. Here we considered two samples, SRR1942956 and SRR1942957, of SARS-CoV-1-infected cells (NC_004718) with a median depth of 2940× and 2765×, respectively. The TRS-B region predicted by CORSID-A is supported by 246 reads in sample SRR1942956 and 233 reads in sample SRR1942957, whereas the TRS-B region predicted by SuPER is supported by only 1 read in each sample (Fig. S3a). Our method was able to identify these positions due to use of flanking positions rather than focusing on identifying a short $6-7$ nt motif as done by SuPER.

Recombination plays a crucial role in the evolution of RNA viruses, and results from homologous or non-homologous template switching. In particular for coronaviruses, template switching occurs at TRS sites during discontinuous transcription [9], making these sites prone to recombination events. CORSID-A uses local alignment to identify TRSs, and unlike SuPER, is not restricted to identifying regulatory sequences of a fixed length. Therefore, as a by-product, our method will be able to find evidence for homologous recombination at these sites. Specifically, even though the length of the core sequence is fixed at 7, the length of the TRS-Ls identified by our method ranges from 9 to 45 (median: 22). This corroborates previous findings showing that recombination hotspots in coronaviruses are colocated with TRS sites [28]. In particular, the longest TRS-B with a length of 45 nucleotides occurs upstream of gene *ns12.9* of NC_006213 with only 6 mismatches showing strong evidence of recombination (Fig. S4). In contrast, the core sequence identified by SuPER and MEME (Fig. 3d and Fig. S5) are at most 10 nt long. Furthermore, we note there is experimental evidence that not only the core sequence but also flanking nucleotides play an important role in discontinuous transcription [26]. This demonstrates the importance of identifying larger regulatory sequences, as done by our method, rather than identifying shorter recurring motifs as done by SuPER and MEME. In summary, CORSID-A outperforms existing methods, such as SuPER and MEME, in identifying TRS sites in coronavirus genomes with given gene locations.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.10.468129; this version posted November 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

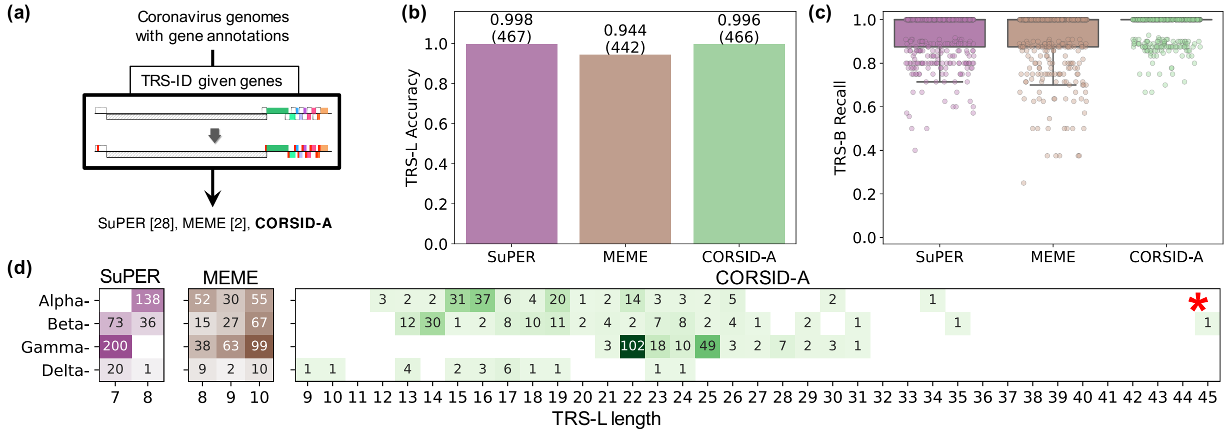


Figure 3: CORSID-A accurately identifies TRS-Ls and TRS-Bs. (a) We used SuPER [28], MEME [2] and CORSID-A to identify TRS sites in 468 coronavirus genome with known gene locations. (b) The fraction of genomes for which the three methods identified the TRS-L correctly. (c) The fraction of genes of the genomes for which the three methods identified the corresponding TRS-B site correctly. (d) Number of coronavirus genomes of the four genera of the *Coronaviridae* family with different lengths of the TRS-L identified by the three methods. Fig. S4 provides the TRS alignment identified by CORSID-A for the genome indicated by '*'.

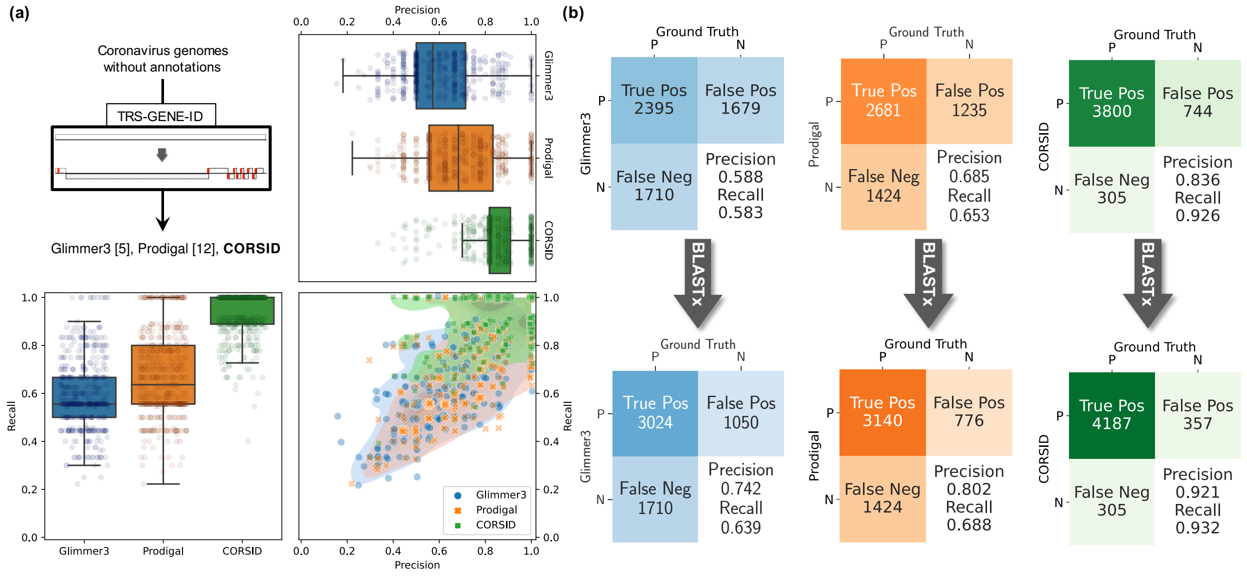
4.2 CORSID identifies genes with higher accuracy than existing methods

We now focus on the TRS-GENE-ID problem, where we compared CORSID to two general-purpose gene finding methods: Glimmer3 [5, 21] and Prodigal [12, 13]. Each method was given as input the complete, unannotated genome sequence of each of the 468 coronaviruses. Following recommended instructions, we ran Glimmer3 by first building the required interpolated context model (ICM) on each genome sequence separately. We ran Prodigal in meta-genomics mode. For CORSID, we used window length $\omega = 7$ and progressively reduced the score threshold τ from 7 to 2. We refer the reader to Appendix B.2 for the precise commands used to run previous tools, and to Appendix B.3 for details on how the predicted set of genes are compared to the ground truth.

Fig. 4a shows that CORSID outperformed Glimmer3 and Prodigal in terms of both precision and recall. The median precision and recall of CORSID is 0.818 and 1.00, respectively, whereas the median precision and recall is 0.556 and 0.556, respectively, for Glimmer3, and 0.636 and 0.667, respectively, for Prodigal. Additionally, the first quartile of both precision and recall for CORSID is greater than the third quartile of these metrics for either Glimmer3 and Prodigal, showing a significant performance advantage for CORSID. While Prodigal and Glimmer3 do not have the capability to identify TRS sites, CORSID identifies these regulatory sites in addition to the genes. Specifically, compared to CORSID-A, which identified TRS-L correctly for 466 (99.6%) genomes, CORSID does so for 443 (94.7%) genomes (Fig. S6). This is a modest reduction in performance, especially when taking into account that CORSID, unlike CORSID-A, is not given any additional information apart from the complete, unannotated genome sequence. Analyzing the previously discussed SARS-CoV-1 genome (NC_004718), we found that CORSID identified the same 10 genes as CORSID-A, while Prodigal missed four genes and Glimmer3 missed two genes (Fig. S3b).

Given the large discrepancy between precision and recall for CORSID, we investigated whether genomes were poorly annotated, leading to incorrect false positives. To that end, we used BLASTx [1] to transfer gene annotation from well-annotated genomes to poorly-annotated genomes. To reduce computation time, we only assessed false positive (FP) genes occurring in $\mathbf{v}_{\text{body}}$. We reclassified a FP gene as a true positive (TP) if the alignment reported by BLASTx spans at least 95% of both the query sequence (the FP gene) and the hit sequence in the database (detailed in Appendix B.2). The resulting confusion

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.10.468129; this version posted November 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



Glimmer3	P	N
P	True Pos 2395	False Pos 1679
N	False Neg 1710	Precision 0.588 Recall 0.583

Prodigal	P	N
P	True Pos 2681	False Pos 1235
N	False Neg 1424	Precision 0.685 Recall 0.653

CORSID	P	N
P	True Pos 3800	False Pos 744
N	False Neg 305	Precision 0.836 Recall 0.926

Glimmer3	P	N
P	True Pos 3024	False Pos 1050
N	False Neg 1710	Precision 0.742 Recall 0.639

Prodigal	P	N
P	True Pos 3140	False Pos 776
N	False Neg 1424	Precision 0.802 Recall 0.688

CORSID	P	N
P	True Pos 4187	False Pos 357
N	False Neg 305	Precision 0.921 Recall 0.932

Figure 4: CORSID accurately identifies TRS-Ls, TRS-Bs, and genes. (a) Precision and recall of Glimmer3 [5], Prodigal [12], and CORSID for gene prediction in 468 genomes. For clarity, we added a small jitter (drawn from $N(0, 2.5 \times 10^{-5})$) to the 2D distribution plot. (b) Confusion matrices of the ground truth genes and the predicted genes by the three methods. In order to account for poorly annotated genomes, we used BLASTx [1] to verify the false positive genes predicted by the three methods. The adjusted confusion matrices are shown in the row below.

matrices of the predicted and the ground-truth genes are shown in Fig. 4b. Although BLASTx found 629 and 459 matches in the database for FPs of Glimmer3 and Prodigal, respectively, which is more than 387 for CORSID, CORSID maintains its lead in absolute numbers for TP and FP. As such, CORSID achieved the highest pooled precision and recall both before and after the BLASTx verification. In summary, CORSID accurately identifies TRSs and genes given just the unannotated genome, outperforming existing gene finding methods.

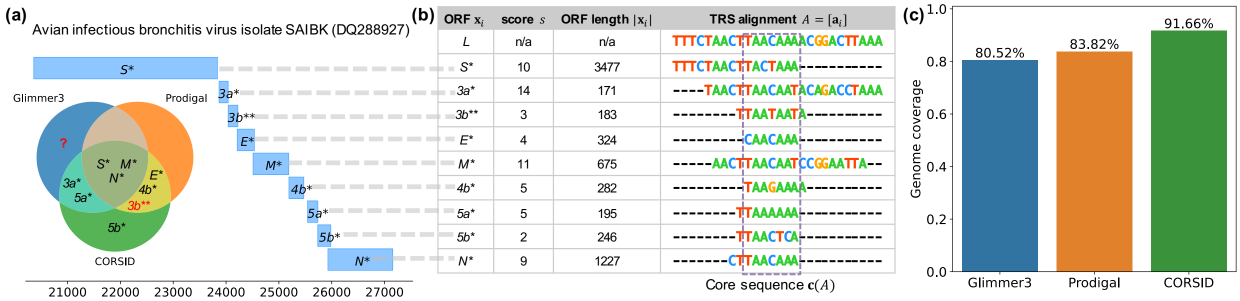
4.3 CORSID enables *de novo* identification of TRS sites and genes

To demonstrate how users can use CORSID to annotate genes and identify TRS-L and TRS-Bs given a newly assembled genome, we analyzed a previously-excluded genome that lacks gene annotation (genome DQ288927). This genome is 27534 nt long, which we provided as input to CORSID, Glimmer3 and Prodigal. CORSID identified nine genes spanning 91.66% of the genome, all of which match annotated genes in other *Igacoviruses* sequences in the BLASTx database (Fig. 5). By contrast, Glimmer3 identified a total of six genes spanning 80.52% of the genome, five of which match genes in the BLASTx database. On the other hand, Prodigal found six genes, all of which were present in the database, spanning 84.22% of the genome. In summary, CORSID identified more genes than existing methods, all of which occurred in homologous previously-annotated genomes in the BLASTx database, demonstrating that CORSID can be used to accurately annotate coronavirus genomes.

5 Discussion

In this paper, we investigated the hypothesis that the presence of transcription regulatory sequences in coronavirus genomes can be leveraged to simultaneously infer these regulatory sequences and their associated genes in a synergistic manner. To that end, we formulated the TRS IDENTIFICATION (TRS-ID) problem of

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.10.468129; this version posted November 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



ORF $\mathbf{x}_i$	score s	ORF length $\lvert\mathbf{x}_i\rvert$	TRS alignment $A = [\mathbf{a}_i]$
L	n/a	n/a	TTTCTAACTTAACAAAACGGACTTAAA
*S**	10	3477	TTTCTAACTTACTAAA-----------
*3a**	14	171	----TAACTTAACAATACAGACCTAAA
*3b***	3	183	--------TTAATAATA----------
*E**	4	324	---------CAACAAA-----------
*M**	11	675	-----AACTTAACAATCCGGAATTA--
*4b**	5	282	---------TAAGAAAA----------
*5a**	5	195	--------TTAAAAAA-----------
*5b**	2	246	--------TTAACTCA-----------
*N**	9	1227	-------CTTAACAAA-----------

Figure 5: CORSID accurately finds genes in an unannotated *Igacovirus* genome (DQ288927). (a) The position of the genes identified by CORSID. The Venn diagram shows the genes found by CORSID, Glimmer3 and Prodigal. "*" indicates $\geq$ 95% query/hit coverage, and "**" indicates a hit coverage of 93.5% and a query coverage of 98.3%. (b) TRS alignment for genes identified by CORSID. (c) The fraction of positions in $\mathbf{v}_{\text{body}}$ covered by genes identified by the three methods.

identifying TRS sites in a coronavirus genome with given gene locations, and the general problem, the TRS AND GENE IDENTIFICATION (TRS-GENE-ID) problem of simultaneous identification of genes and TRS sites given only the coronavirus genome. Underpinning both problems is the notion of a TRS alignment, which extends the previous concept of core sequences to include flanking nucleotides that provide additional signal. Our proposed method for the first problem, CORSID-A, is based upon a dynamic programming formulation which extends the classical Smith-Waterman recurrence [24]. CORSID, which solves the general problem, additionally incorporates a maximum-weight independent set formulation on an interval graph to identify TRS sites and genes.

Using extensive experiments on 468 coronavirus genomes, we showed that CORSID-A outperformed two motif-based approaches, MEME [2] and SuPER [28]. Additionally, we showed that CORSID outperformed two general-purpose gene finding algorithms, Glimmer3 [5, 21] and Prodigal [12]. We performed direct validation of TRS sites predicted for the SARS-CoV-1 genome (NC_004718), showing that the TRS sites identified by our method are more strongly supported by split reads in RNA-seq samples than the TRS sites identified by SuPER. Lastly, we demonstrated that CORSID enables *de novo* identification of TRSs and genes in newly assembled coronavirus genomes by applying it on a previously unannotated coronavirus (DQ288927) belonging to the *Igacovirus* subgenus.

There are several avenues for future research. First, CORSID currently requires the complete genome as input to identify the TRS sites and the genes. We plan to extend our method to allow gene identification in the several coronaviruses available in GenBank with only partial reference genomes by leveraging knowledge from other coronaviruses with complete genomes with similar TRS sites. Second, while in this study we only focused on coronaviruses, discontinuous transcription occurs in all viruses in the taxonomic order of *Nidovirales*. However, CORSID, which assumes a single TRS-L region in the genome, cannot be directly applied to other families of viruses within *Nidovirales* such as the family *Mesoniviridae* that contain multiple TRS-L regions in the genome. Incorporating such features and extending CORSID to all *Nidovirales* viruses is a useful direction of future work. Finally, currently CORSID requires the reference genome of the virus as input. In the future, we plan to extend this method to simultaneously discover the reference genome and the core-sequences of the virus using *de novo* assembly. We envision that this will enable simultaneous genome assembly and gene annotation of coronaviruses.

Acknowledgements. This material is based upon work supported by the National Science Foundation under award numbers CCF-1850502, CCF-2027669 and CCF-2046488. We thank Ayesha Kazi, Michael Xiang, and Yichi Zhang for developing the web-based visualization tool of CORSID solutions.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.10.468129; this version posted November 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

References

[1] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[2] Timothy L Bailey, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, 37(suppl_2):W202–W208, 2009.

[3] Humberto Carrillo and David Lipman. The multiple sequence alignment problem in biology. *SIAM Journal on Applied Mathematics*, 48(5):1073–1082, 1988.

[4] Shih-Cheng Chen and René CL Olsthoorn. Group-specific structural features of the 5'-proximal sequences of coronavirus genomic RNAs. *Virology*, 401(1):29–41, 2010.

[5] Arthur L Delcher, Kirsten A Bratke, Edwin C Powers, and Steven L Salzberg. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, 23(6):673–679, 2007.

[6] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[7] Thomas A Down and Tim JP Hubbard. NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. *Nucleic Acids Research*, 33(5):1445–1453, 2005.

[8] Yaara Finkel, Orel Mizrahi, Aharon Nachshon, Shira Weingarten-Gabbay, David Morgenstern, Yfat Yahalom-Ronen, Hadas Tamir, Hagit Achdout, Dana Stein, Ofir Israeli, et al. The coding capacity of SARS-CoV-2. *Nature*, 589(7840):125–130, 2021.

[9] Rachel L Graham, Damon J Deming, Meagan E Deming, Boyd L Yount, and Ralph S Baric. Evaluation of a recombination-resistant coronavirus as a broadly applicable, rapidly implementable vaccine platform. *Communications Biology*, 1(1):1–10, 2018.

[10] S. Griffiths-Jones. Rfam: an RNA family database. *Nucleic Acids Research*, 31(1):439–441, January 2003.

[11] Ju Yuan Hsiao, Chuan Yi Tang, and Ruay Shiung Chang. An efficient algorithm for finding a maximum weight 2-independent set on interval graphs. *Information Processing Letters*, 43(5):229–235, 1992.

[12] Doug Hyatt, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1):1–11, 2010.

[13] Doug Hyatt, Philip F LoCascio, Loren J Hauser, and Edward C Uberbacher. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics*, 28(17):2223–2230, 2012.

[14] Marilyn Kozak. Initiation of translation in prokaryotes and eukaryotes. *Gene*, 234(2):187–208, 1999.

[15] Michael M. C. Lai and Stephen A Stohlman. Comparative analysis of RNA genomes of mouse hepatitis viruses. *Journal of Virology*, 38(2):661–670, 1981.

[16] Mark A Larkin, Gordon Blackshields, Nigel P Brown, R Chenna, Paul A McGettigan, Hamish McWilliam, Franck Valentin, Iain M Wallace, Andreas Wilm, Rodrigo Lopez, et al. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947–2948, 2007.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.10.468129; this version posted November 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

[17] Ramakanth Madhugiri, Markus Fricke, Manja Marz, and John Ziebuhr. RNA structure analysis of alphacoronavirus terminal genome regions. *Virus Research*, 194:76–89, 2014.

[18] Helena Jane Maier, Erica Bickerton, and Paul Britton. Coronaviruses. *Methods and Protocols*, 2015.

[19] Christian Jean Michel, Claudine Mayer, Olivier Poch, and Julie Dawn Thompson. Characterization of accessory genes in coronavirus genomes. *Virology Journal*, 17(1):1–13, 2020.

[20] Giulio Pavesi, Paolo Mereghetti, Giancarlo Mauri, and Graziano Pesole. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Research*, 32(suppl_2):W199–W203, 2004.

[21] Steven L Salzberg, Arthur L Delcher, Simon Kasif, and Owen White. Microbial gene identification using interpolated Markov models. *Nucleic Acids Research*, 26(2):544–548, 1998.

[22] Palash Sashittal, Chuanyi Zhang, Jian Peng, and Mohammed El-Kebir. Jumper enables discontinuous transcript assembly in coronaviruses. *bioRxiv*, 2021.

[23] Aditi Shukla and Rolf Hilgenfeld. Acquisition of new protein domains by coronaviruses: analysis of overlapping genes coding for proteins N and 9b in SARS coronavirus. *Virus Genes*, 50(1):29–38, 2015.

[24] Temple F Smith, Michael S Waterman, et al. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, 1981.

[25] Isabel Sola, Fernando Almazan, Sonia Zúñiga, and Luis Enjuanes. Continuous and discontinuous RNA synthesis in coronaviruses. *Annual review of virology*, 2:265–288, 2015.

[26] Isabel Sola, José L. Moreno, Sonia Zúñiga, Sara Alonso, and Luis Enjuanes. Role of Nucleotides Immediately Flanking the Transcription-Regulating Sequence Core in Coronavirus Subgenomic mRNA Synthesis. *Journal of Virology*, 79(4):2506–2516, February 2005.

[27] Ke Xu, Bo-Jian Zheng, Rong Zeng, Wei Lu, Yong-Ping Lin, Liang Xue, Li Li, Lei-Lei Yang, Chen Xu, Jie Dai, et al. Severe acute respiratory syndrome coronavirus accessory protein 9b is a virion-associated protein. *Virology*, 388(2):279–285, 2009.

[28] Yiyan Yang, Wei Yan, A Brantley Hall, and Xiaofang Jiang. Characterizing transcriptional regulatory sequences in coronaviruses and their role in recombination. *Molecular Biology and Evolution*, 38(4):1241–1248, 2021.

[29] Zizhen Yao, Zasha Weinberg, and Walter L Ruzzo. CMfinder—a covariance model based RNA motif finding algorithm. *Bioinformatics*, 22(4):445–452, 2006.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.10.468129; this version posted November 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

A Supplementary Methods

A.1 Obtaining candidate regions for the TRS-ID problem

Intuitively, a candidate region for gene $\mathbf{x}_i$ corresponds to the region $\mathbf{w}_i = [w_i^-, w_i^+]$ composed of positions $w^- \leq p \leq w^+$ such that any sgmRNA starting at p will lead to the translation of ORF $\mathbf{x}_i$ by the ribosome. Note that the first ORF $\mathbf{x}_0$ corresponds to *ORF1ab*. Accordingly, we restrict $\mathbf{w}_0 = [w_0^-, w_0^+]$ to match exactly the leader region, spanning the start of the genome at the 5' end until the start codon of $\mathbf{x}_0$, *i.e.* $w_0^- = 1$ and $w_0^+ = x_0^- - 1$. To define the remaining candidate regions $\mathbf{w}_1, \ldots, \mathbf{w}_n$, we must take ribosomal leaky scanning into account, where the ribosome does not initiate translation at the first 'ATG' it encounters [14]. To model this, we make use of the fact that coronovirus genes have a length of at least 100 nucleotides. Specifically, when determining the candidate region of a gene, we skip over a previous ORF in case its length is less than 100. To that end, we introduce the following function.

Definition 10. Function $\mathrm{prev}(p)$ returns the first ORF $\mathbf{x} = [x^-, x^+]$ upstream of position p in the genome, *i.e.* for ORF $\mathbf{x}$ returned by $\mathrm{prev}(p)$ it holds that $x^- < p$ and there exists no ORF $\mathbf{y} = [y^-, y^+]$ such that $x^- < y^- < p$. If no such ORF $\mathbf{x}$ exists then $\mathrm{prev}(p) = [0, 0]$. Moreover, $\mathrm{prev}(0) = [0, 0]$.

Using this function, we define a TRS-B candidate region $\mathbf{w}$ of an ORF $\mathbf{x}$ as follows.

Definition 11. Let $\mathbf{x} = [x^-, x^+]$ be an ORF, and let $\mathbf{y} = [y^-, y^+] = \mathrm{prev}(x^-)$ and $\mathbf{z} = [z^-, z^+] = \mathrm{prev}(y^-)$ be the previous two ORFs. The *candidate region* $\mathbf{w} = [w^-, w^+]$ of ORF $\mathbf{x}$ ends at the start of $\mathbf{x}$, *i.e.* $w^+ = x^- - 1$, and begins at the first position of the genome if $\mathbf{x}$ has no previous ORF or the only preceding ORF $\mathbf{y}$ has a length smaller than 100; $\mathbf{w}$ begins at the first ORF $\mathbf{y}$ if the length of $\mathbf{y}$ is at least 100; otherwise $\mathbf{w}$ begins at the second ORF $\mathbf{z}$ if the first ORF $\mathbf{y}$ has a length smaller than 100 nucleotides. That is,

$$w^- = \begin{cases} 1, & \text{if } \mathbf{y} = [0, 0], \\ 1, & \text{if } \mathbf{y} \neq [0, 0], \mathbf{z} = [0, 0] \text{ and } |\mathbf{y}| < 100, \\ y^- + 3, & \text{if } \mathbf{y} \neq [0, 0], |\mathbf{y}| \geq 100, \\ z^- + 3, & \text{if } \mathbf{y} \neq [0, 0], \mathbf{z} \neq [0, 0] \text{ and } |\mathbf{y}| < 100. \end{cases} \tag{1}$$

Finally, to remove overlap among candidate regions $\mathbf{w}_0, \ldots, \mathbf{w}_n$, we set $w_i^- = w_{i-1}^+ + 1$ if $w_{i-1}^- \geq w_i^-$ for all $i \in \{1, \ldots, n\}$.

A.2 Constrained TRS-ID problem

Here, we introduce the following constrained version of the TRS-ID problem.

Problem 3 (CONSTRAINED TRS IDENTIFICATION (TRS-ID-$\mathbf{u}$)). Given non-overlapping sequences $\mathbf{w}_0, \ldots, \mathbf{w}_n$, and a subsequence $\mathbf{u}$ of $\mathbf{w}_0$, find a TRS alignment $A = [\mathbf{a}_0, \ldots \mathbf{a}_n]^T$ such that (i) $\mathbf{a}_i$ corresponds to a subsequence in $\mathbf{w}_i$ for all $i \in \{0, \ldots, n\}$, (ii) $\mathbf{u}$ is a subsequence of the core sequence $\mathbf{c}(A)$, and (iii) the alignment has maximum score $s(A)$.

Let us focus on solving a single TRS-ID-$\mathbf{u}$ problem instance, where we are given non-overlapping sequences $\mathbf{w}_0, \ldots, \mathbf{w}_n$ and a subsequence $\mathbf{u}$ of $\mathbf{w}_0$. Each such instance decomposes into n TRS-ID-$\mathbf{u}$ instances each with exactly two sequences. That is, for any $i \in \{1, \ldots, n\}$, we seek a TRS alignment $A_i = [\mathbf{a}_0^i, \mathbf{a}_i^i]^T$ of sequences $\mathbf{w}_0 = [w_{0,p}]$ and $\mathbf{w}_i = [w_{i,q}]$ such that the induced core sequence $\mathbf{c}(A_i)$ contains $\mathbf{u} = [u^-, u^+]$. This is a variant of local alignment [24] with three key differences: (i) alignment A_i

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.10.468129; this version posted November 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

may not contain gaps, (ii) A_i must span u^- and (iii) A_i must span u^+. Letting ℓ be the relative position of u^- in $\mathbf{w}_0$, we obtain

$$s[p,q] = \max \begin{cases} 0, & \text{if } p < \ell, \\ s[p-1,q-1] + \delta(u_p, w_q), & \text{if } p \geq 1 \text{ and } q \geq 1, \end{cases} \tag{2}$$

where $s[p,q]$ indicates the optimal score of a constrained TRS alignment between $w_{0,1} \ldots w_{0,p}$ and $w_{i,1} \ldots w_{i,q}$. The desired solution is then as follows.

$$(p^*, q^*) = \operatorname*{arg\,max}_{\ell+|\mathbf{u}|-1 \leq p \leq |\mathbf{w}_0|,\, p \leq q \leq |\mathbf{w}_i|} s[p,q]. \tag{3}$$

Note that the recurrence (2) lacks the two cases of the Smith-Waterman [24] recurrence corresponding to a gap (*i.e.* $s[p-1,q]$ and $s[p,q-1]$), thus satisfying constraint (i). Constraints (ii) and (iii) are satisfied because the first case, corresponding to initiating the alignment, is only enabled when $p < \ell$ thus covering u^-, and by (3), we have that the alignment contains u^+. Similarly to local alignment, we can identify (p^*, q^*) by filling out the table $s[0,0], \ldots, s[|\mathbf{w}_0|, |\mathbf{w}_i|]$ using dynamic programming, and reconstruct the TRS alignment $A_i = [\mathbf{a}_0^i, \mathbf{a}_i^i]^T$ using a backtrace from (p^*, q^*). Letting $L = \sum_{i=1}^n |\mathbf{w}_i|$ be the total length of candidate regions $\mathbf{w}_1, \ldots, \mathbf{w}_n$, solving the n pairwise TRS-ID-$\mathbf{u}$ problems and obtaining the pairwise TRS alignments $A_1, \ldots, A_n$ that cover $\mathbf{u}$ takes $O(|\mathbf{w}_0|L)$ time.

Given these pairwise alignments $A_1, \ldots, A_n$ that each span $\mathbf{u}$, we construct the final TRS alignment $A = [\mathbf{a}_0, \ldots, \mathbf{a}_n]^T$ as follows. First, we exclude alignments A_i that have a score less than the threshold τ. Second, the TRS-L sequence $\mathbf{a}_0$ equals the subsequence of $\mathbf{w}_0$ that spans the positions covered by all pairwise alignments, *i.e.* $\mathbf{a}_0$ spans exactly the positions of $\mathbf{w}_0$ covered by $\mathbf{a}_0^1, \ldots, \mathbf{a}_0^n$. Third, we obtain the remaining gapped sequences $\mathbf{a}_1, \ldots, \mathbf{a}_n$ of A by adding flanking gaps to each (ungapped) sequence $\mathbf{a}_1^1, \ldots, \mathbf{a}_n^n$ so as to match the unaligned letters of $\mathbf{a}_0$ (Fig. 2a). As flanking gaps do not incur a penalty, this operation will not change the total score, *i.e.* $s(A) = \sum_{i=1}^n s(A_i)$. The running time of computing alignments $A_1, \ldots, A_n$ and then subsequently merging them into A is dominated by the first step.

A.3 Partitioning the genome into $\mathbf{v}_{\text{leader}}$ and $\mathbf{v}_{\text{body}}$ for the TRS-GENE-ID problem

To obtain these two sequences, $\mathbf{v}_{\text{leader}}$ and $\mathbf{v}_{\text{body}}$, for a given coronavirus genome, we developed a heuristic for identifying *ORF1ab*, the largest gene in coronavirus genomes. This heuristic begins by enumerating all ORFs $\mathbf{x}_1, \ldots, \mathbf{x}_m$ in the genome (Definition 5). As *ORF1ab* is the result of a frameshift upstream of the stop codon of *ORF1a* [18], we extend each enumerated ORF $\mathbf{x}_i$ by performing either a -1 or -2 frameshift and subsequently scanning for an in-frame stop codon. We select the frameshift that results in the largest extended ORF, obtaining extended ORFs $\mathbf{y}_1, \ldots, \mathbf{y}_m$. We designate the largest ORF among this set as *ORF1ab*. Finally, we set $\mathbf{v}_{\text{leader}}$ as the region from the start of the genome until the 5' coordinate of *ORF1ab*. As the TRS-B of the first gene downstream of *ORF1ab* may reside within *ORF1ab*, we set $\mathbf{v}_{\text{body}}$ as the region starting from 200 nucleotides upstream of the 3' coordinate of *ORF1ab* until the 3' end of the genome.

A.4 Constrained TRS-GENE-ID problem

Here, we introduce the constrained version of the TRS-GENE-ID problem.

Problem 4 (CONSTRAINED TRS AND GENE IDENTIFICATION (TRS-GENE-ID-$\mathbf{u}$)). Given leader region $\mathbf{v}_{\text{leader}}$, body region $\mathbf{v}_{\text{body}}$ and a subsequence $\mathbf{u}$ of $\mathbf{v}_{\text{leader}}$, find a TRS alignment $A = [\mathbf{a}_i]$ such that (i) $\mathbf{a}_0$ corresponds to a subsequence in $\mathbf{v}_{\text{leader}}$, (ii) $\mathbf{a}_i$ corresponds to a subsequence in $\mathbf{v}_{\text{body}}$ for all $i \geq 1$, (iii) $\mathbf{u}$ is a subsequence of the core sequence $\mathbf{c}(A)$, (iv) A is concordant, and (v) A induces the set $\Gamma(A)$ of genes with maximum genome coverage $g(A)$ and subsequently has maximum score $s(A)$.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.10.468129; this version posted November 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

We solve this problem in two steps. First, we use dynamic programming to compute $s[p, q]$ for all values of $0 \leq p \leq |\mathbf{v}_{\text{leader}}|$ and $0 \leq q \leq |\mathbf{v}_{\text{body}}|$, *i.e.* the optimal score $s[p, q]$ of a TRS alignment between $v_{\text{leader},1} \ldots v_{\text{leader},p}$ and $v_{\text{body},1} \ldots v_{\text{body},q}$ constrained to contain $\mathbf{u}$. The quantity $s[p, q]$ is defined using the same recurrence as for the TRS-ID-$\mathbf{u}$ problem – Eq. (2) – and the complete table can be filled out using dynamic programming in $O(|\mathbf{v}_{\text{leader}}||\mathbf{v}_{\text{body}}|)$ time.

Second, let $\mathbf{x}_1, \ldots, \mathbf{x}_m$ be the candidate ORFs in $\mathbf{v}_{\text{body}}$, each with a length of at least 100 nucleotides (Fig. S7). For each ORF $\mathbf{x}_i$ we find the position (p, q) that encodes the maximum scoring alignment $A_i = [\mathbf{a}^i_0, \mathbf{a}^i_i]^T$ where $\mathbf{a}^i_i$ is associated with $\mathbf{x}_i$. We remove ORFs $\mathbf{x}_i$ whose maximum scoring alignment A_i has a score $s(A_i)$ less than the user-specified score threshold τ. Let s^* indicate the maximum score among all TRS alignments $A_1, \ldots, A_m$. Then, we construct a vertex-weighted interval graph $G = (V, E)$ whose vertices V correspond to the intervals $\{[x^-_1, x^-_1], \ldots, [x^-_m, x^+_m]\}$ of the candidate ORFs. There is an edge $(\mathbf{x}_i, \mathbf{x}_j)$ if and only if the two corresponding intervals overlap, *i.e.* $[x^-_j, x^-_i] \cap [x^-_j, x^-_j] \neq \emptyset$. To capture the lexicographical ordering of the objective functions, *i.e.* first the genome coverage and then the score, each vertex/interval $\mathbf{x}_i$ is assigned weight

$$w(\mathbf{x}_i) = |\mathbf{x}_i| + \frac{s(A_i)}{s^*}. \tag{4}$$

In other words, among ORFs with the same length, we prefer those that have an associated TRS alignment with largest score. Finally, we solve a maximum-weight independent set (MWIS) problem, which can be done in $O(|V|)$ time for interval graphs [11]. The maximum-weight independent set $X = \{\mathbf{x}_{\pi(1)}, \ldots, \mathbf{x}_{\pi(|X|)}\}$ directly corresponds to the induced genes $\Gamma(A)$ of the TRS alignment A that can be constructed by merging pairwise TRS alignments $A_{\pi(1)}, \ldots, A_{\pi(|X|)}$ following the same procedure described in Section 3.1. For each window $\mathbf{u}$ of fixed length ω, the first step takes $O(|\mathbf{v}_{\text{leader}}||\mathbf{v}_{\text{body}}|)$ time and the second step takes $O(|\mathbf{v}_{\text{leader}}||\mathbf{v}_{\text{body}}| + m)$ time.

A.5 Practical considerations to solve the TRS-GENE-ID problem

In this section we discuss practical considerations for identifying genes in coronaviruses and the way they are addressed in CORSID.

Overlapping genes. In practice, coronavirus genes may overlap. That is, the start codon of a gene can be located within another gene. To support such cases, we shrink candidate ORFs $\mathbf{x} = [x^-, x^+]$ by 5% prior to solving the MWIS problem, *i.e.* the interval graph $G = (V, E)$ has shortened intervals $\mathbf{x}' = [x^- + \alpha, x^+ - \alpha]$ where $\alpha = 0.05|\mathbf{x}|$.

Progressive approach. To obtain biologically meaningful solutions, we solve the TRS-GENE-ID problem in a progressive manner. More specifically, given user-specified parameters $(\tau_{\text{min}}, \tau_{\text{max}})$, we start with setting $\tau = \tau_{\text{max}}$ and solve the problem for a fixed window $\mathbf{u}$. Then, for every subsequent iteration we decrement τ and require the solution to contain all ORFs that were identified in the previous iteration. The final iteration occurs when $\tau = \tau_{\text{min}}$, yielding the final solution for window $\mathbf{u}$. We consider all sliding windows $\mathbf{u}$ within $\mathbf{v}_{\text{leader}}$ and return the solution with maximum genome coverage and subsequently maximum score. We show that the progressive approach performs better than solving the TRS-GENE-ID problem directly using $\tau = \tau_{\text{min}}$ or $\tau = \tau_{\text{max}}$ in Fig. S8.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.10.468129; this version posted November 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

B Supplementary Results

B.1 Establishing the ground truth of gene and TRS locations in coronavirus genomes

We established a ground-truth set of genes by processing the GFF annotation files and extracting a set of genes for each genome, removing duplicates and incomplete ORFs. In particular, we removed 10 ORFs (8 annotated as *N* and 2 annotated as ‘unknown’) that had duplicate names and were completely covered by another gene with the same name. This resulted in a median number of 8 genes per genome (min 3 and max 14, Fig. S9). We excluded two genomes that had no genes in their annotation (DQ288927 and EU526388).

We used the following approach to establish the ground truth for TRS-B sites. As discussed in Section 3.1, we identified *candidate regions* (Definition 11) for each gene. If a method identified a TRS-B contained within any of the candidate regions, then we counted this as the method recalling the TRS-B for the corresponding gene.

We established ground-truth locations of TRS-Ls using the fact that these regulatory sequences occur between the second (SL2) and fourth stem loop (SL4) in the 5’ untranslated region (UTR) [17]. For *Sarbecovirus* genomes, a subgenus of the *Betacoronavirus* genus, we required TRS-Ls to occur in stem loop (SL3) [4]. Specifically, we analyzed leader regions upstream of *ORF1ab*, its location taken from the GFF annotation or determined by our own heuristic in case this gene was absent, and performed a multiple sequence alignment using ClustalW2 [16] of all sequences within each subgenus. We then superimposed secondary structure information from Rfam [10] onto each alignment to identify the relevant stem loops in each viral sequence, obtaining for each sequence a small range in which TRS-L may occur (Fig. S10).

B.2 Command-line arguments

MEME. We ran MEME v5.3.0 in mode “zero or one occurrence per sequence” (zoops) and maximum width of 10.

```
python corena/candidate_region.py -f ${input_fasta} -g ${input_gff} \
    -o ${fasta} -m 10 --prev2ATG
meme ${fasta} -dna -maxw 10 -oc ${output_dir} -mod zoops -nmotifs 1
```

Glimmer3. We followed the steps written in the script `g3-from-scratch.csh` provided in the Glimmer3 package.

```
long-orfs -n -t 1.15 ${input_fasta} ${long_orfs}
extract -t ${input_fasta} ${long_orfs} > ${train}
build-icm -r ${icm} < ${train}
glimmer3 -g 100 ${input_fasta} ${icm} ${dir}
```

Prodigal. We ran Prodigal v2.6.3 in metagenomic mode.

```
prodigal -i ${input_fasta} -p meta -f gff -o ${output_gff} -s ${output_gene}
```

ClustalW2. We used ClustalW2 v2.1 to align sequences.

```
clustalw2 -infile=${fasta}
```

BLASTx. We ran BLASTx on a FASTA file containing FP genes with the following parameters. We used a subset of protein sequences from the official “nr” database (downloaded on October 7, 2021), containing all species under the taxonomic unit *Coronaviridae* (taxid:11118). From the BLASTx output, we extracted the top hit for each FP gene, and calculated the coverage between the aligned subsequence, and the query and

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.10.468129; this version posted November 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

hit sequences. In particular, the *hit coverage* is fraction of position in the hit sequence that are aligned. On the other hand, the *query coverage* is the fraction the query sequence that are aligned. We set the threshold to 95% for both hit and query coverage, restricting solutions to almost exact matches.

```
blastx -db ${database} -query ${fasta} -word_size 6 -gapopen 11 -gapextend 1 \
    -comp_based_stats 2 -task blastx -max_target_seqs 15 \
    -evalue 0.05 -num_threads 6 -outfmt 15 -out ${json}
```

B.3 Evaluating gene finding methods

To assess performance of gene finding methods, we say that a predicted gene $\mathbf{x} = [x^-, x^+]$ is correct provided there exists a ground-truth gene $\mathbf{y} = [y^-, y^+]$ in the same genome such that $|x^- - y^-| \leq 3$ and $|x^+ - y^+| \leq 3$. In other words, the start and end positions may be off by the length of at most one codon, accounting for variation in annotation (e.g. sometimes the stop codon is omitted from ORFs). Moreover, as *ORF1ab* is not a real ORF, i.e. the corresponding polypeptide 1ab results from a -1 frameshift, we treat this gene differently. That is, we say that a method correctly identified $\mathbf{y} = [y^-, y^+] =$ *ORF1ab* if it found two genes $\mathbf{x}_1 = [x_1^-, x_1^+]$ and $\mathbf{x}_2 = [x_2^-, x_2^+]$ such that the start position of $\mathbf{x}_1$ is at most 3 nucleotides away from y^- and the stop position of $\mathbf{x}_2$ is at most 3 nucleotides away from y^+. Note that CORSID will identify a single gene $\mathbf{x}$ matching *ORF1ab*, in which case $\mathbf{x} = \mathbf{x}_1 = \mathbf{x}_2$. Using this definition we classify each predicted and ground-truth gene as either a *true positive* (TP), i.e. the predicted gene matches a ground-truth gene; *false positive* (FP), i.e. the predicted gene does not match any ground-truth gene); or *false negative* (FN), i.e. the ground-truth gene has no matching predicted gene.

B.4 Two anomalous coronavirus genomes

CORSID-A was unsuccessful in identifying the correct TRS-L site in only two coronaviruses: MK211372 (from subgenus *Pedacovirus* of genus *Alphacoronovirus*) and MK472070 (unclassified subgenus of genus *Alphacoronovirus*).

To understand why CORSID-A failed on MK211372, we performed a multiple sequence alignment of the leader regions of all 45 genomes in the *Pedacovirus* subgenus. Inspecting the alignment, we see that MK211372 is an outlier, with multiple insertion/deletions in the TRS-L region compared to the other sequences (Fig. S11). This explains why CORSID-A was unable to accurately identify the TRS-L and TRS-B sites for this genome.

Since genome MK472070 has a known genus but unknown subgenus, we only aligned it to the covariance model of the alphacoronaviruses. From the alignment result we found a poor alignment in the TRS-L region. Based on the alignment, it resembles some nyctacoviruses, but it is still an outlier, as shown in Fig. S12. Specifically, the TRS-L consensus sequence, 5'-TCAACTAAAC-3', differed significantly from the subsequence 5'-ACAATCTAAT-3' of MK472070, with a Hamming distance of 5. Moreover, MEME also failed to identify TRS-L in this genome, and SuPER assigned a low confidence score to the identified TRS-Bs for important genes such as *S* and *N*.

In summary, we believe further investigation of genomes MK211372 and MK472070 is warranted in order to determine whether they harbor a TRS-L region or whether the deposited genome sequences are incomplete/incorrect.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.10.468129; this version posted November 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

C Supplementary Figures and Tables

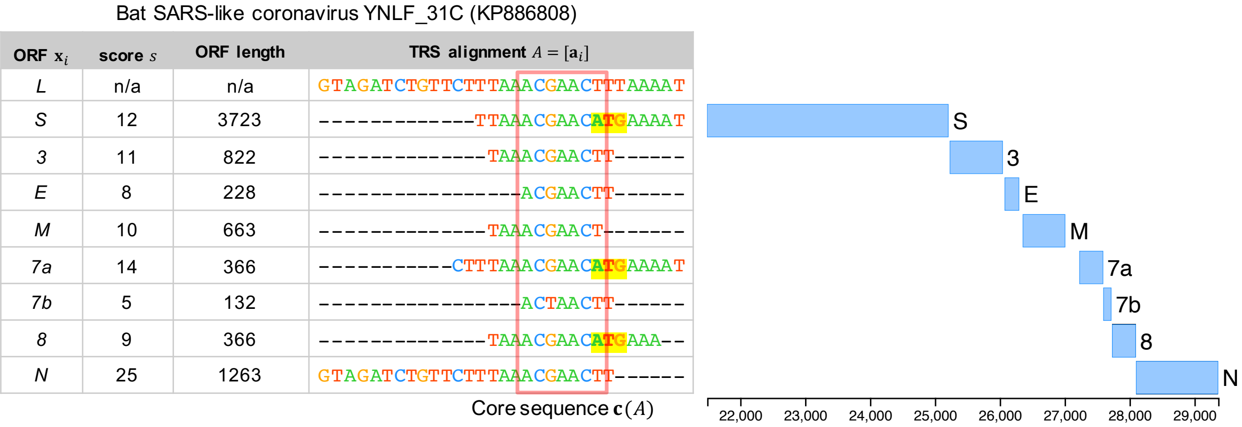
We have the following supplementary figures and tables.

- Fig. S1 shows an example of TRS regions overlapping with the start codon of the corresponding genes in coronaviruses.
- Fig. S2 shows CORSID finds the correct TRS-L in a *Sarbecovirus*, verified using a RNA-seq dataset.
- Fig. S3 shows RNA-seq data supports the TRS-B found by CORSID-A in SARS-CoV-1 (NC_004718), and CORSID identifies more genes than Glimmer3 and Prodigal in the same genome.
- Fig. S4 shows the TRS alignment in genome NC_006213 where CORSID find the longest alignment between TRS-L and TRS-B.
- Fig. S5 shows the number of coronaviruses with varying lengths of TRS-L regions broken down by the genus and the subgenus. Several coronaviruses of each genus have TRS-L regions much longer than the core sequences indicating recombination events.
- Fig. S6 shows that CORSID shows modest reduction in TRS-L accuracy compared to CORSID-A.
- Fig. S7 shows the distribution of length of annotated genes. Only 10 out of 3637 genes are shorther than 100 nt.
- Fig. S8 shows that CORSID achieves better performance when using the progressive approach rather than directly solving $\tau = \tau_{\max} = 7$ or $\tau = \tau_{\min} = 2$.
- Fig. S9 shows the histogram of length of annotated genes from 468 genomes, of which only 10 genes are shorter than 100 nt.
- Fig. S10 shows the distribution of the length of the TRS-L regions, separated by four genera.
- Fig. S11 shows the MSA of leader regions of pedacoviruses with genome MK211372 highlighted, indicating multiple INDELs.
- In Table S1, we compare features of CORSID, CORSID-A, and other methods. We show that CORSID is the first method for simultaneous identification of TRS sites and genes in coronaviruses.
- Table S2 shows the number of included and excluded genomes grouped by genera and subgenera.

	TRS-L identification	TRS-B identification	Gene identification
CORSID	✓	✓	✓
CORSID-A	✓	✓	✗
SuPER [28]	✓	✓	✗
MEME [2]	✓	✓	✗
Glimmer3 [5]	✗	✗	✓
Prodigal [12]	✗	✗	✓

Table S1: CORSID **is the first method for simultaneous identification of TRS sites and genes in coronaviruses**. This table shows the features of CORSID and CORSID-A along with three existing methods: MEME [2], Glimmer3 [5] and Prodigal [12].

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.10.468129; this version posted November 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

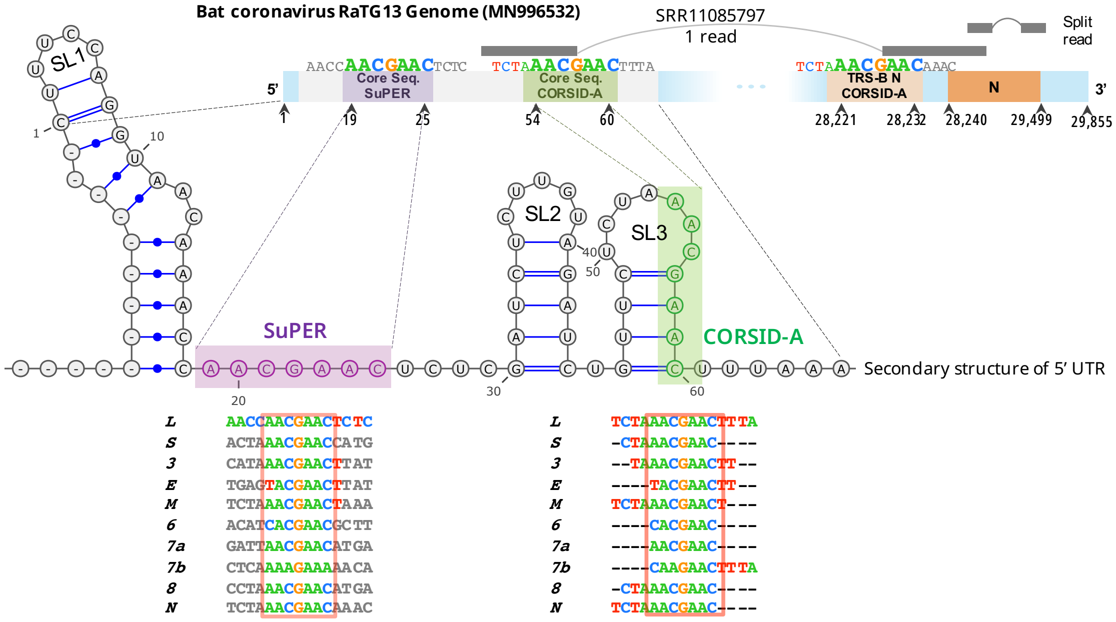


Bat SARS-like coronavirus YNLF_31C (KP886808)

ORF x_i	score s	ORF length	TRS alignment $A = [\mathbf{a}_i]$
L	n/a	n/a	GTAGATCTGTTCTTTAAACGAACTTTAAAAT
S	12	3723	-------------TTAAACGAACATGAAAAT
3	11	822	--------------TAAACGAACTT------
E	8	228	-----------------ACGAACTT------
M	10	663	--------------TAAACGAACT-------
7a	14	366	-----------CTTTAAACGAACATGAAAAT
7b	5	132	-----------------ACTAACTT------
8	9	366	--------------TAAACGAACATGAAA--
N	25	1263	GTAGATCTGTTCTTTAAACGAACTT------

Core sequence $\mathbf{c}(A)$

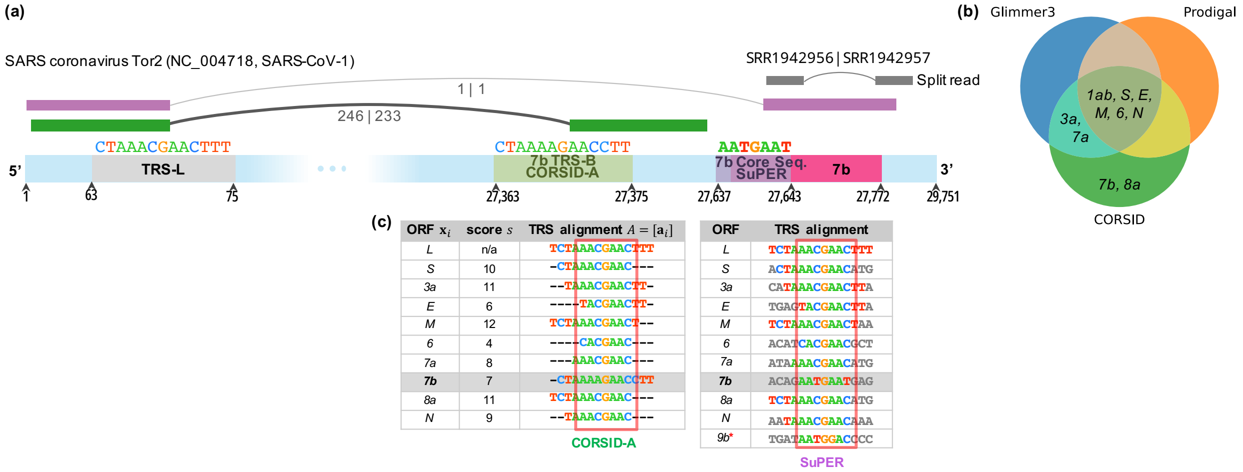
Figure S1: TRS sites may overlap with the start codon of the corresponding genes in coronaviruses. The TRS alignment identified by CORSID when applied to a *Sarbecovirus* genome KP886808, showing th start codons of gene *S, 7a,* and *8* are partially contained in core sequences (highlighted in yellow. We note that MEME identified the same TRS-L as our method. As a side note, the TRS-B of gene *N* matches 25 nucleotides of the TRS-L, indicative of a possible a recombination event.



```
L   AACCAACGAACTCTC     L   TCTAAACGAACTTTA
S   ACTAAACGAACCATG     S   -CTAAACGAAC----
3   CATAAACGAACTTAT     3   --TAAACGAACTT--
E   TGAGTACGAACTTAT     E   ----TACGAACTT--
M   TCTAAACGAACTAAA     M   TCTAAACGAACT---
6   ACATCACGAACGCTT     6   ----CACGAAC----
7a  GATTAACGAACATGA     7a  ----AACGAAC----
7b  CTCAAAAGAAAAACA     7b  ----CAAGAACTTTA
8   CCTAAACGAACATGA     8   -CTAAACGAAC----
N   TCTAAACGAACAAAC     N   TCTAAACGAAC----
```

Figure S2: CORSID-A finds the correct TRS-L site in a *Sarbecovirus* unlike SuPER. Although SuPER uses a hard-coded motif to identify core sequences, it incorrectly identified TRS-L (positions 19 – 25) in genome MN996532 (Bat coronavirus RaTG13). By contrast, CORSID-A found the TRS-L at a different location (54 – 60). We verified CORSID-A's position using the corresponding RNA-seq data (SRR11085797). We aligned the reads to the reference genome using a splice-aware aligner, STAR [6]. The resulting alignment had a single split read, spanning positions 54 and 28221, which matches CORSID-A's TRS-L and the TRS-B for gene *N*. Moreover, TRS-L sites in sarbecoviruses occur in stem loop (SL3) [4], which co-incides with the TRS-L site identified by CORSID-A.

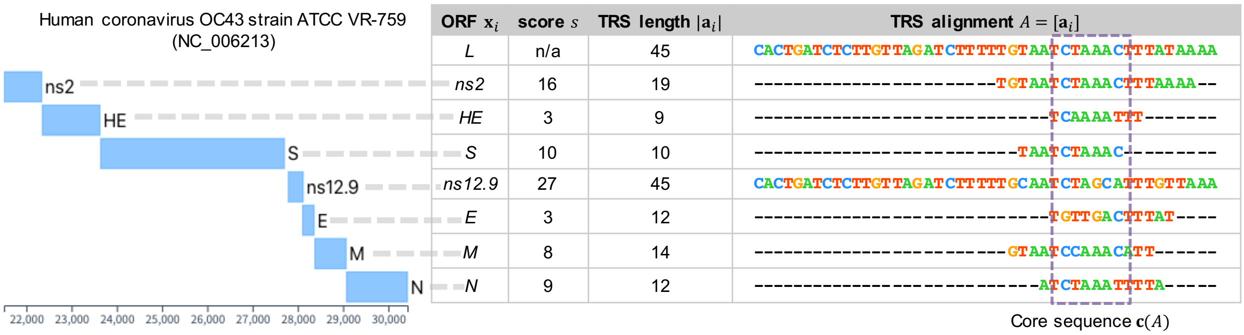
bioRxiv preprint doi: https://doi.org/10.1101/2021.11.10.468129; this version posted November 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



ORF $\mathbf{x}_i$	score s	TRS alignment $A = [\mathbf{a}_i]$
L	n/a	TCTAAACGAACTTT
S	10	-CTAAACGAAC---
3a	11	--TAAACGAACTT-
E	6	----TACGAACTT-
M	12	TCTAAACGAACT--
6	4	----CACGAAC---
7a	8	---AAACGAAC---
7b	7	-CTAAAAGAACCTT
8a	11	TCTAAACGAAC---
N	9	--TAAACGAAC---

ORF	TRS alignment
L	TCTAAACGAACTTT
S	ACTAAACGAACATG
3a	CATAAACGAACTTA
E	TGAGTACGAACTTA
M	TCTAAACGAACTAA
6	ACATCACGAACGCT
7a	ATAAAACGAACATG
7b	ACAGAATGAATGAG
8a	TCTAAACGAACATG
N	AATAAACGAACAAA
*9b**	TGATAATGGACCCC

Figure S3: RNA-seq data supports the TRS-B found by CORSID-A, and CORSID identifies more genes than Glimmer3 and Prodigal in a SARS-CoV-1 genome (NC_004718) (a) For SARS-CoV-1 (NC_004718), CORSID-A identified a TRS-B corresponding to *ORF7b* supported by 246 and 233 split reads in RNA-sequencing samples SRR1942956 and SRR1942957, respectively. For the same genome, SuPER identified a TRS-B site that is supported by only one read in each RNA-sequencing sample. (b) A Venn diagram of the genes correctly identified by the three methods. (c) The TRS alignment of CORSID-A, and aligned core sequence with flanking regions of SuPER. "*": SuPER identified a TRS-B for *9b* but its start codon is located at the second to fourth nucleotide of the core sequence, and the Hamming distance is 2. CORSID-A did not find this TRS-B as it occurs outside the candidate region of *9b*. Moreover, in previous studies *ORF9b* has been hypothesized to be translated via a ribosomal leaky scanning mechanism [19, 23, 27], explaining the absence of an associated TRS-B site.


Human coronavirus OC43 strain ATCC VR-759 (NC_006213)
ns2
HE
S
ns12.9
E
M
N
22,000
23,000
24,000
25,000
26,000
27,000
28,000
29,000
30,000
Core sequence $\mathbf{c}(A)$

ORF $\mathbf{x}_i$	score s	TRS length $\lvert\mathbf{a}_i\rvert$	TRS alignment $A = [\mathbf{a}_i]$
L	n/a	45	CACTGATCTCTTGTTAGATCTTTTTGTAATCTAAACTTTATAAAA
ns2	16	19	------------------------TGTAATCTAAACTTTAAAA--
HE	3	9	-----------------------------TCAAAATTT-------
S	10	10	--------------------------TAATCTAAAC---------
ns12.9	27	45	CACTGATCTCTTGTTAGATCTTTTTGCAATCTAGCATTTGTTAAA
E	3	12	-----------------------------TGTTGACTTTAT----
M	8	14	-------------------------GTAATCCAAACATT------
N	9	12	----------------------------ATCTAAATTTTA-----

Figure S4: Location of genes and the TRS alignment inferred by CORSID for Human coronavirus OC43 strain ATCC VR-759 (genome NC_006213). This genome has the longest TRS-B with a length of 45 nucleotides (with only 6 mismatches) upstream of gene *ns12.9*, indicative of a recombination event.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.10.468129; this version posted November 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

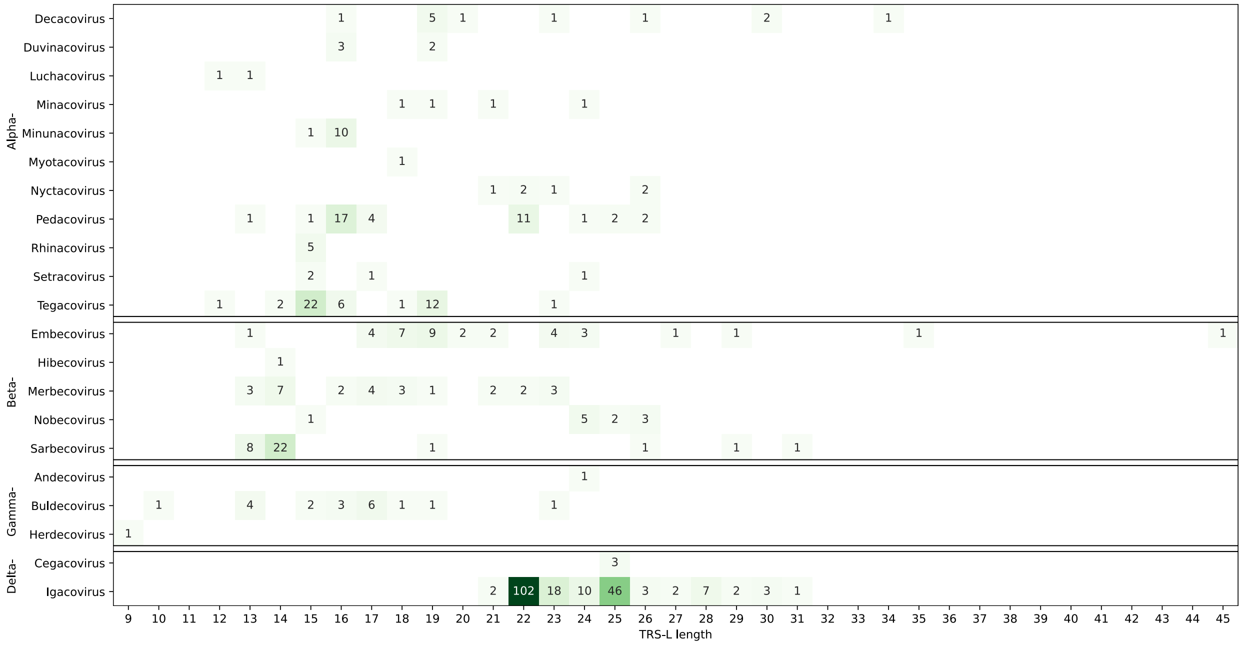


Figure S5: Several coronaviruses of each genus have TRS-L regions much longer (median: 22 nt) than the core sequence length ($6-7$ nt) indicating recombination events.

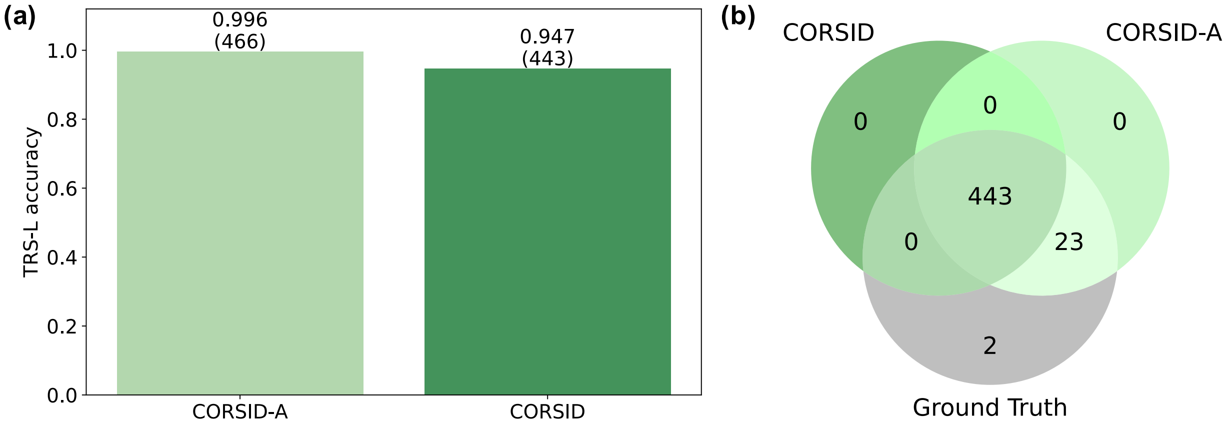


Figure S6: CORSID shows modest reduction in TRS-L accuracy compared to CORSID-A. (a) TRS-L accuracy of CORSID and CORSID-A. (b) Venn diagram of the TRS-L identified by CORSID and CORSID-A compared to the ground truth. Both methods failed on genomes MK211372 and MK472070 (discussed in Appendix B.4).

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.10.468129; this version posted November 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

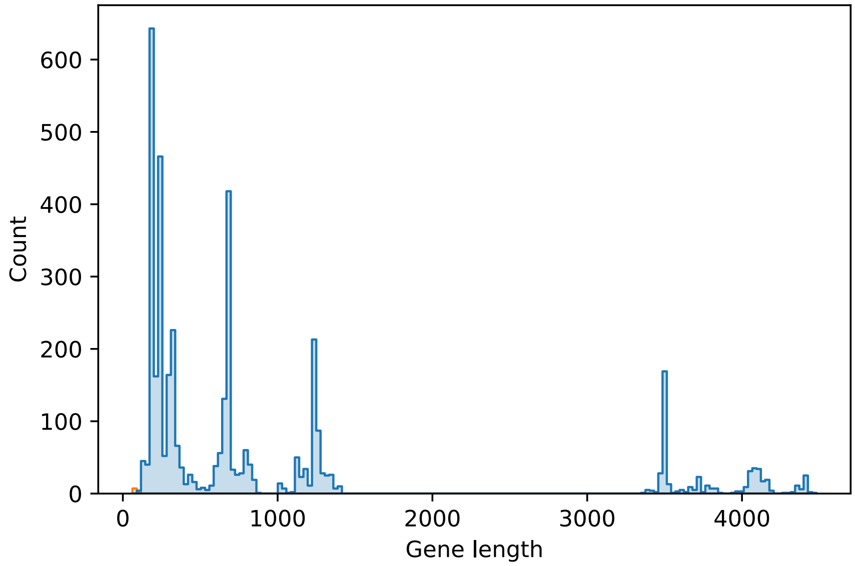


Figure S7: Coronavirus genes are almost always longer than 100 **nt** . Histogram of annotated gene lengths from 468 genomes (*ORF1ab* not included). Only 10 out of the 3637 genes are shorter than 100 nt.

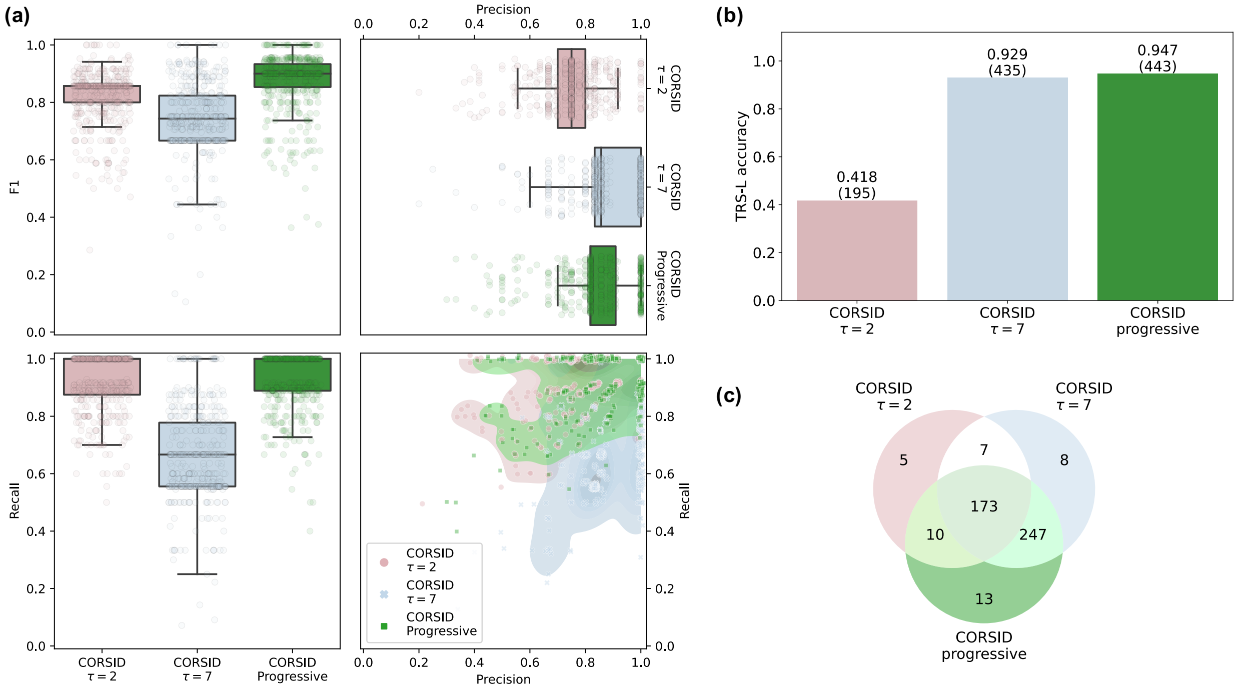


Figure S8: Using the progressive approach rather than directly solving $\tau = \tau_{\max} = 7$ **or** $\tau = \tau_{\min} = 2$ **leads to better performance.** (a) The F_1 score (harmonic mean between precision and recall) is shown in top left panel. We show the precision and recall in the top right and lower left panel, respectively, and the joint distribution in the lower right panel. (b) TRS-L accuracy of setting the minimum matching score threshold to $\tau = 2$ and $\tau = 7$, compared with the progressive approach. (c) Venn diagram of genome sets with correctly identified TRS-L by the three versions of CORSID.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.10.468129; this version posted November 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

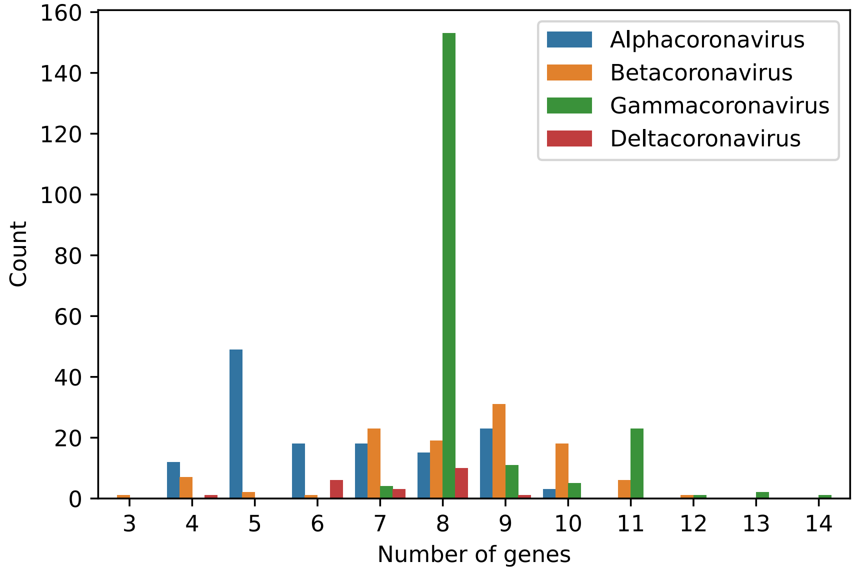


Figure S9: Number of annotated genes varies across the four genera of coronaviruses. (Median: 8, min: 3, max: 14).

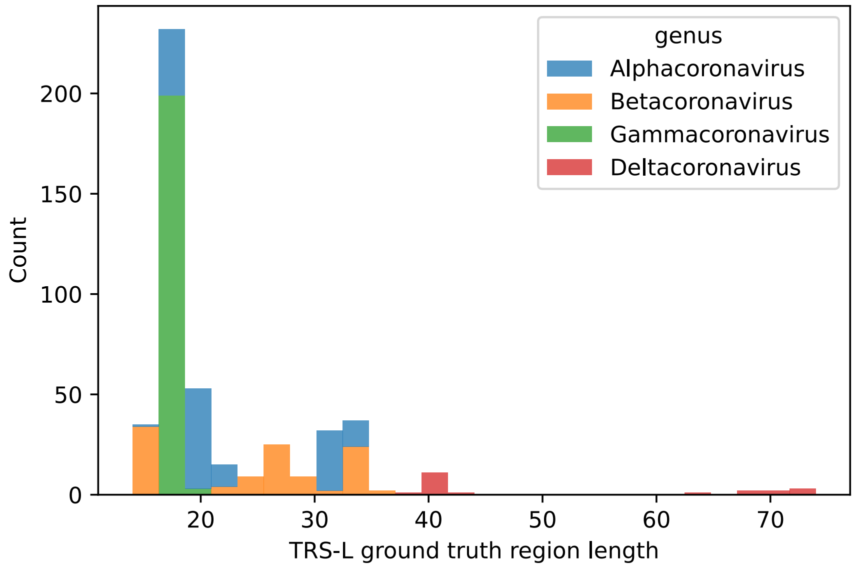


Figure S10: Length of the TRS-L region varies across the four genera of coronaviruses.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.10.468129; this version posted November 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

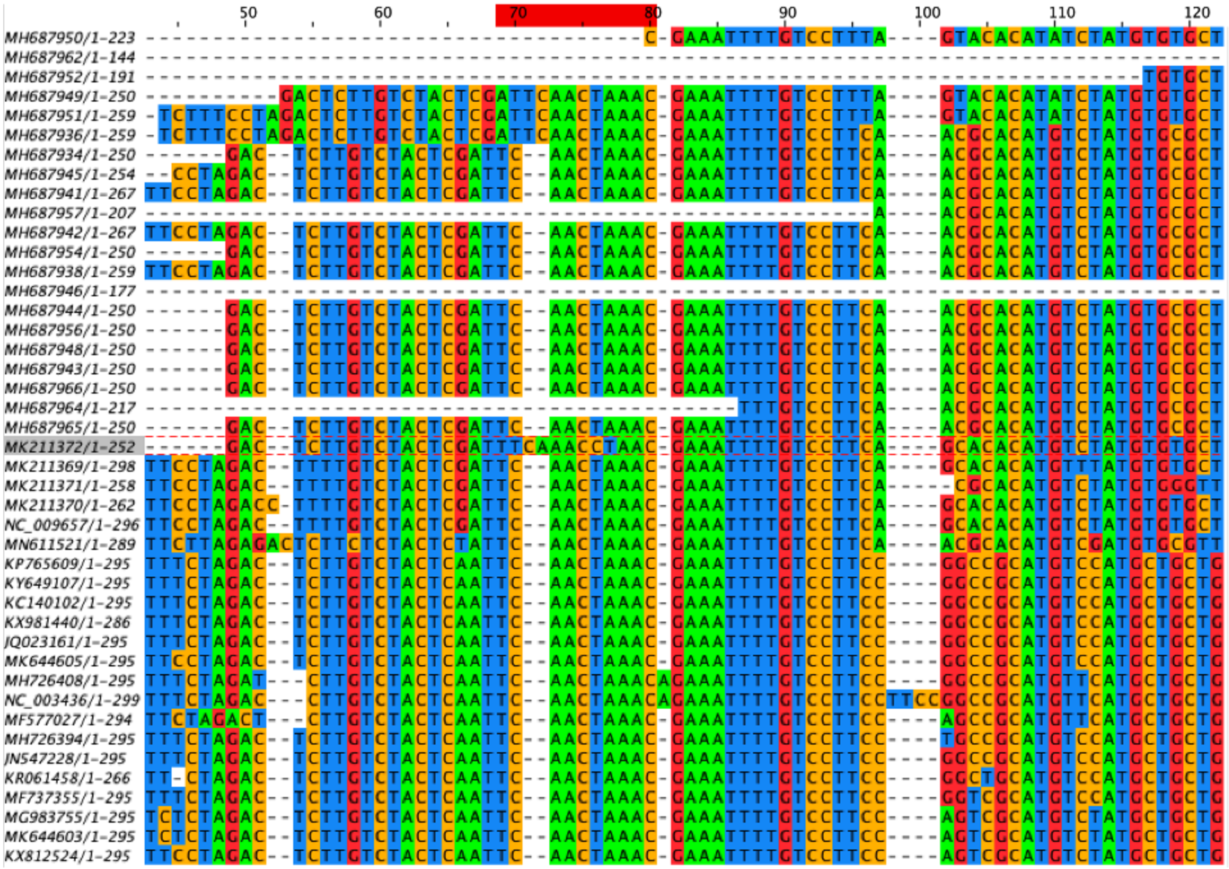


Figure S11: MSA of leader regions of pedacoviruses. We highlight the MK211372 genome and indicate the TRS-L region with a red bar on top. MK211372 differs from other sequences since it contains multiple indels in the TRS-L region. As such, the TRS-L region cannot be found in this genome.

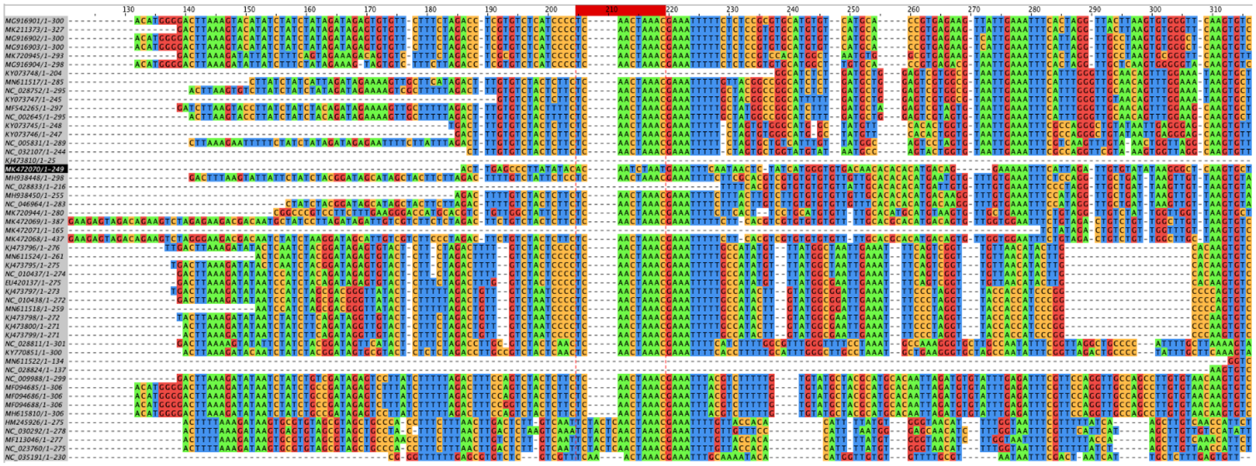


Figure S12: MSA of leader regions of some alphacoronaviruses. We highlight MK472070 genome and indicate the TRS-L region with red bar on top. MK472070 differs from others since it contains multiple indels in the TRS-L region. As such, the TRS-L region cannot be found in this genome.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.10.468129; this version posted November 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Genus	Subgenus	# Included genomes	# Excluded genomes
Alphacoronavirus	Colacovirus	0	2
	Decacovirus	12	2
	Duvinacovirus	5	1
	Luchacovirus	2	0
	Minacovirus	4	2
	Minunacovirus	11	0
	Myotacovirus	1	0
	Nyctacovirus	6	2
	Pedacovirus	39	6
	Rhinacovirus	5	2
	Setracovirus	4	0
	Tegacovirus	45	1
	Unclassified	4	1
Subtotal		138	19
Betacoronavirus	Embecovirus	36	1
	Hibecovirus	1	0
	Merbecovirus	27	1
	Nobecovirus	11	0
	Sarbecovirus	34	6
Subtotal		109	8
Deltacoronavirus	Andecovirus	1	0
	Buldecovirus	19	0
	Herdecovirus	1	0
Subtotal		21	0
Gammacoronavirus	Cegacovirus	3	0
	Igacovirus	196	10
	Unclassified	1	0
Subtotal		200	10
Totals		468	37

Table S2: Number of coronaviruses of each genus and subgenus included and exlcuded in this study.