MBE, 2022, 1–10
doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

Accurate Identification of Transcription Regulatory Sequences and Genes in Coronaviruses

Chuanyi Zhang,[1,*] Palash Sashittal,[2,3,*] Michael Xiang,[2] Yichi Zhang,[2] Ayesha Kazi[2] and Mohammed El-Kebir[2,†]

[1]Department of Electrical & Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, [2]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL and [3]Present address: Department of Computer Science, Princeton University, Princeton, NJ

[*]Joint first authorship; [†]Corresponding author: melkebir@illinois.edu

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Transcription regulatory sequences (TRSs), which occur upstream of structural and accessory genes as well as the 5' end of a coronavirus genome, play a critical role in discontinuous transcription in coronaviruses. We introduce two problems collectively aimed at identifying these regulatory sequences as well as their associated genes. First, we formulate the TRS Identification problem of identifying TRS sites in a coronavirus genome sequence with prescribed gene locations. We introduce CORSID-A, an algorithm that solves this problem to optimality in polynomial time. We demonstrate that CORSID-A outperforms existing motif-based methods in identifying TRS sites in coronaviruses. Second, we demonstrate for the first time how TRS sites can be leveraged to identify gene locations in the coronavirus genome. To that end, we formulate the TRS and Gene Identification problem of simultaneously identifying TRS sites and gene locations in unannotated coronavirus genomes. We introduce CORSID to solve this problem, which includes a web-based visualization tool to explore the space of near-optimal solutions. We show that CORSID outperforms state-of-the-art gene finding methods in coronavirus genomes. Furthermore, we demonstrate that CORSID enables *de novo* identification of TRS sites and genes in previously unannotated coronavirus genomes. 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.

Key words: Core sequences, Gene identification, Coronavirus, Motif finding, Local alignment, Maximum weight independent set, Interval graph

Introduction

Coronaviruses are enveloped viruses comprised of a positive-sense, single-stranded RNA genome that is ready to be translated by the host ribosome. 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 several structural, non-structural and accessory genes (fig. 1a). These genes are necessary for the viral life cycle and are expressed and translated using three distinct mechanisms (Sola et al. 2015).

First, upon cell entry, the viral genome is translated to produce polypeptides corresponding to one or two overlapping open reading frames (ORFs). The resulting polypeptides undergo auto-cleavage, producing many non-structural proteins, including the RNA-dependent-RNA-polymerase (RdRP). Second, the viral RdRP mediates the expression of the remaining viral genes via *discontinuous transcription* (Sola et al. 2015). That is, the RdRP is prone to perform template switching, predominantly 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 viral genes — called TRS-B where B stands for body (fig. 1b). Note that while previous studies have found evidence of TRS-independent template switching leading to non-canonical transcripts, the function of these transcripts is still unknown (Kim et al. 2020; Finkel et al. 2021; Sashittal et al. 2021). Third, occasionally

certain genes are expressed via *leaky scanning*, where a weak initiation codon leads to the translation of the next downstream ORF (Jungreis et al. 2021). 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 (supplementary table S1). More specifically, since each TRS contains a $6-7$ nt long conserved sequence, called a *core sequence* (Sola et al. 2015; Finkel et al. 2021), general-purpose motif finding methods (Pavesi et al. 2004; Bailey et al. 2009; Down and Hubbard 2005; Yao et al. 2006) can be employed to identify TRS-L and TRS-Bs in coronaviruses. For instance, MEME (Bailey et al. 2009) 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 (Yang et al. 2021), 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 (Salzberg et al. 1998; Delcher et al. 2007), Prodigal (Hyatt et al. 2010, 2012) and VADR (Schäffer et al. 2020). 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. While Glimmer3 and Prodigal are designed for prokaryotic genomes, VADR is specifically designed for identifying genes in viral genomes. To that end, VADR first classifies the input sequence and finds the most similar sequence in a pre-specified database, maps the curated annotations to the input based on a covariance model, and then uses BLAST (Altschul et al. 1990) to validate the annotated genes. Importantly, current gene finding tools do not 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 (Smith et al. 1981) for local sequence alignment. Additionally, we introduce CORSID to solve the TRS-Gene-ID problem via a maximum-weight independent set problem (Hsiao et al. 1992) on an interval graph defined by the candidate ORFs in the genome with weights obtained from the previous DP.

CORSID enables *de novo* identification of viral genes in coronaviruses using only the nucleotide sequence of the viral genome. CORSID-A, on the other hand, is designed to identify all the TRSs in a coronavirus genome annotated with gene locations. 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 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.

New Approaches

Viewing TRS and gene identification as a multiple sequence alignment is a novel approach that we will outline in this section. We begin by introducing notation and key definitions, followed by stating the TRS Identification problem and then the TRS and Gene Identification problem. Finally, we overview the key methodological contributions of our algorithms.

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 also call a contiguous subsequence $\mathbf{x}$ of $\mathbf{v}$ 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 $\mathbf{b}_0, \dots, \mathbf{b}_n \in \Sigma^*$ 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 $\mathbf{b}_i$.

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

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

Intuitively, the first sequence $\mathbf{a}_0$ in the alignment A represents TRS-L, whereas $\mathbf{a}_1, \dots, \mathbf{a}_n$ 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 (Sola et al. 2005). For the same

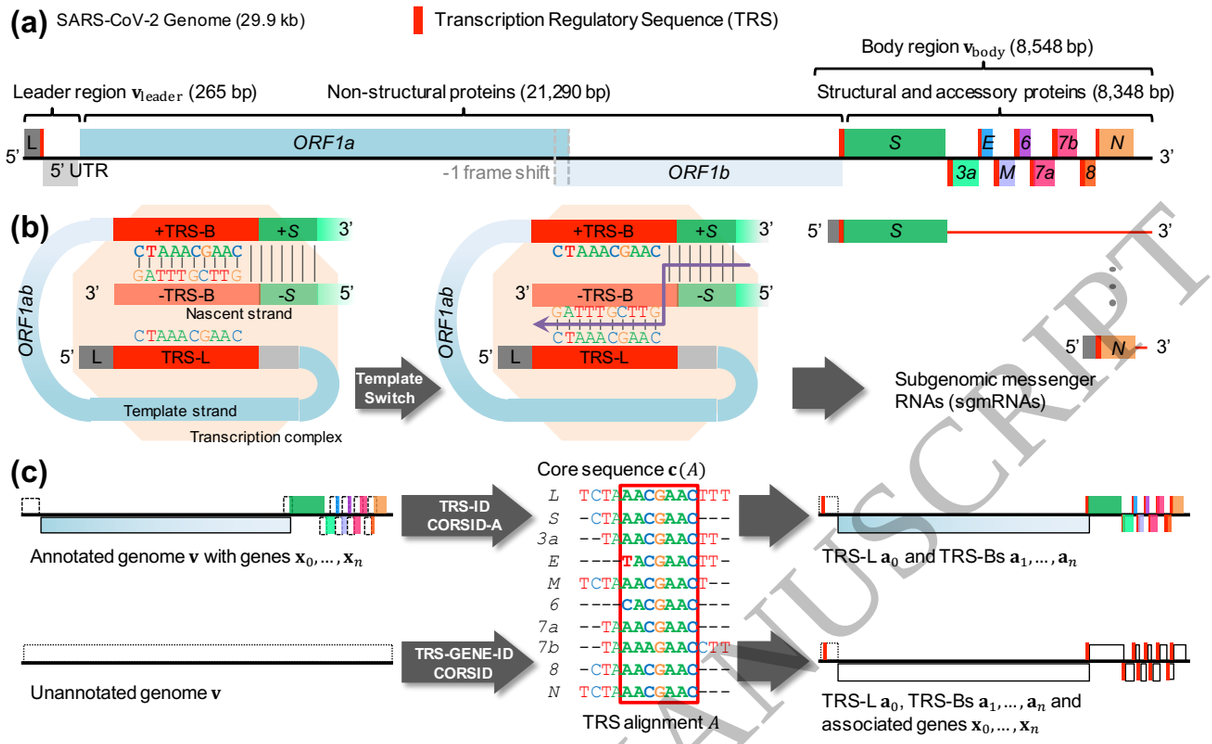

Fig. 1: **Overview.** (a) A coronavirus genome $\mathbf{v}$ consists of a leader region $\mathbf{v}_{\text{leader}}$ and a body region $\mathbf{v}_{\text{body}}$. (b) Structural and accessory genes are expressed via discontinuous transcription with template switching occurring at transcription regulatory 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 $\mathbf{x}_0, \ldots, \mathbf{x}_n$. 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).

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 $A = [\mathbf{a}_0, \ldots, \mathbf{a}_n]^\top$ 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 (Carrillo and Lipman 1988) 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 $A = [\mathbf{a}_0, \ldots, \mathbf{a}_n]^\top$ 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 (Sola et al. 2005).

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 $\mathbf{x}_1, \ldots, \mathbf{x}_n$. We detail in Materials and Methods 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 $A = [\mathbf{a}_0, \ldots, \mathbf{a}_n]^\top$ 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 $s(A)$.

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 $\mathbf{v}_{\text{body}}$. We describe in Materials and Methods 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.* $x_1 \ldots x_3 = \text{ATG}$, (iii) ends at a stop codon, *i.e.* $x_{|\mathbf{x}|-2} \ldots x_{|\mathbf{x}|} \in \{\text{TAA}, \text{TAG}, \text{TGA}\}$, and (iv) does not contain an internal in-frame stop codon, *i.e.* for all $j \in \{1, \ldots, |\mathbf{x}|/3 - 1\}$ it holds that $x_{3(j-1)+1} \ldots x_{3(j-1)+3} \notin \{\text{TAA}, \text{TAG}, \text{TGA}\}$.

Naively, to identify the ORF associated with TRS-B $\mathbf{a}_i$, 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 leaky scanning into account, where the ribosome does not initiate translation at the first encountered 'ATG'. We provide a more robust definition of a downstream ORF that takes leaky scanning into account in Materials and Methods. To summarize, we have that a TRS alignment $A = [\mathbf{a}_0, \ldots, \mathbf{a}_n]^\top$ uniquely determines a set $\Gamma(A)$ of candidate genes.

Definition 6. A set $\Gamma(A)$ of ORFs are *induced genes* of a TRS alignment $A = [\mathbf{a}_0, \ldots, \mathbf{a}_n]^\top$ 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 there may *not* be an ORF downstream of each TRS-B $\mathbf{a}_i$. As such, we have that $|\Gamma(A)| \leq n$. While in theory multiple TRS-Bs of a TRS alignment $A = [\mathbf{a}_0, \ldots, \mathbf{a}_n]^\top$ may induce the same gene in $\mathbf{v}_{\text{body}}$, in practice each coronavirus gene typically has a unique TRS-B. Moreover, these viral genes are typically non-overlapping in the genome. To that end, we have the following definition.

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

In practice, since some coronavirus genes may overlap, we later relax this definition to allow some overlap between ORFs (details in supplementary methods). Finally, coronavirus genomes tend to be compact with most positions coding for genes. To capture these biological constraints, we introduce the following definition.

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 subsequently maximum score $s(A)$.

In other words, there are two objectives, genome coverage $g(A)$ and alignment score $s(A)$, that are ordered lexicographically. That is, if there exist multiple TRS alignments that have maximum genome coverage, we break ties using the alignment score.

Overview of CORSID and CORSID-A

Our algorithm CORSID-A solves the TRS-ID problem to optimality. The key algorithmic insight is that the problem decomposes into n independent pairwise alignment problems when fixing a window in the leader candidate region $\mathbf{w}_0$ that contains the core sequence. Our second problem, the TRS-Gene-ID problem, is solved by CORSID. We use the same insight of sliding a window through the leader region $\mathbf{v}_{\text{leader}}$ and show that the constrained problem corresponds to a maximum weight independent set on an interval graph, which can be solved in polynomial time. We implemented both methods in Python. Moreover, we implemented a web-based visual analytics tool for exploring the space of near-optimal solutions. The source code is available at `https://github.com/elkebir-group/CORSID`. The results of CORSID and CORSID-A are available at `https://github.com/elkebir-group/CORSID-data` and we also built a web application to visualize these results for easier exploration of the solution space and manual annotation (`https://elkebir-group.github.io/CORSID-viz/`). We refer to Materials and Methods for further details, including a description of scope, recommendations, practical considerations and heuristics for obtaining the required input to each problem.

Results

To evaluate the performance of CORSID-A and CORSID, we downloaded the same set of 505 assembled coronavirus genomes previously analyzed by SuPER (Yang et al. 2021) 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

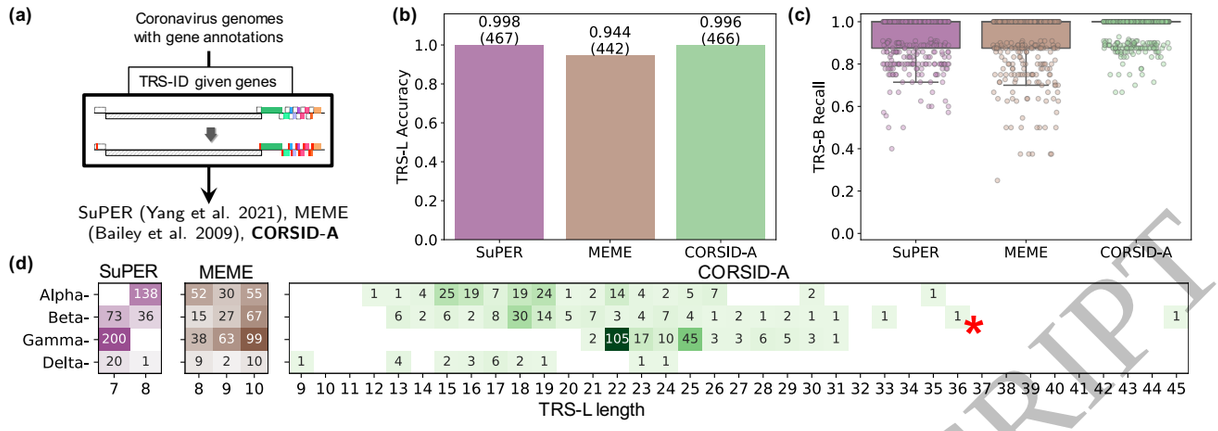

Fig. 2: **CORSID-A accurately identifies TRS-Ls and TRS-Bs.** (a) We used SuPER (Yang et al. 2021), MEME (Bailey et al. 2009) and CORSID-A to identify TRS sites in 468 coronavirus genome with known gene locations. (b) The fraction of genomes for which the three methods identified the TRS-L correctly. (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. The TRS alignment identified by CORSID-A for the genome indicated by '*' is shown in supplementary fig. S7.

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. To account for missing genes in annotation GFF files, we used BLASTx to identify an extended set of ground-truth genes (Altschul et al. 1990) (supplementary fig. S1). We refer to supplementary methods for additional details on how we established the set of genes and locations of TRS sites in the coronavirus genomes (supplement section 2.1 and supplementary fig. S2). We excluded 35 genomes with 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 (supplementary table S2).

CORSID-A Finds TRS-L and TRS-Bs with High Accuracy

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 (Bailey et al. 2009), whereas SuPER is specifically designed for identifying core sequences within coronavirus genomes annotated with genes (Yang et al. 2021). To run CORSID-A, we extracted candidate regions $\mathbf{w}_1, \ldots, \mathbf{w}_n$ upstream of annotated genes $\mathbf{x}_0, \ldots, \mathbf{x}_n$ (see supplementary methods for a precise definition of candidate region). The minimum length ω of core sequence is set to 7 following existing literature (Alonso et al. 2002; Sola et al. 2015), and we use a minimum alignment score of $\tau = 2$. We provided MEME with the same candidate regions $\mathbf{w}_0, \ldots, \mathbf{w}_n$, and ran it in "zero or one occurrence per sequence" mode. For SuPER, we analyzed the previously reported results on the same 468 sequences. We refer to the supplementary results for detailed commands and parameters.

As shown in fig. 2b, 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 (supplementary results, supplementary fig. S3 and supplementary fig. S4). We discuss the one genome (MN996532) where SuPER failed to identify TRS-L correctly in supplementary fig. S5, 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 (SRR11085797). Split reads from RNA-sequencing data map to non-contiguous regions of the viral genome and provide direct evidence of template switching at TRS sites during viral replication in infected cells (Sashittal et al. 2021).

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. 2c). 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 for *ORF7b* predicted by CORSID-A is supported by 246 reads in sample SRR1942956 and 233 reads in sample SRR1942957. On the other hand, SuPER found a different TRS-B region for this gene, which it marked as not recommended, and is supported by only 1 read in each sample (supplementary fig. S6a). We suspect our method successfully identified the TRS-B region by using matching flanking positions of the core sequence rather than restricting the search to a short $6-7$ nt motif in a fixed length region as done by SuPER.

As CORSID-A does not restrict the length of regulatory sequences, our method is able to find evidence for homologous recombination and/or putative TRS-L derived insertions. Specifically, even though the length of the core sequence is fixed at 7, the length of the TRSs identified by our method can be longer due to matching sequences in regions flanking the core sequence. While the core sequences identified by SuPER and MEME (fig. 2d and supplementary fig. S9) are at most 10 nt long, the length of TRSs identified by CORSID-A ranges from 9 to 45 (median: 22, supplementary fig. S7). Across all the genomes for which CORSID-A identifies a TRS-L longer than 25 nt (42 genomes), the median length of the core sequences is 7 and the median number of mismatches of the between core sequences within the longest TRS-B and the core sequences within TRS-L is only 1 (supplementary fig. S8). In particular, in betacoronavirus genome NC_006577, CORSID-A identifies a TRS-B upstream of *ORF4* with a length of 36 nucleotides that perfectly matches the TRS-L as well as another TRS-B preceding gene *HE* with a length of 27 nucleotides with only 1 mismatch, showing strong evidence of recombination and/or TRS-L derived insertion (supplementary fig. S7). Thus, we corroborate previous findings showing numerous genomic insertions of 5'-UTR in betacoronaviruses (Pataca and Haseltine 2022) and that recombination hotspots in coronaviruses are colocated with TRS sites (Yang et al. 2021). Furthermore, we note that there is experimental evidence that, besides the core sequences, flanking nucleotides also play an important role in discontinuous transcription (Sola et al. 2005). In summary, by considering matches in the regions flanking the core sequences using the TRS alignment, CORSID-A finds evidence for putative recombination and/or TRS-L derived insertion events and more accurately identifies regulatory sequences compared to existing motif finding methods such as SuPER and MEME.

CORSID Identifies Genes with High Accuracy

We now focus on the TRS-Gene-ID problem, where we compared CORSID to three gene finding methods: Glimmer3 (Salzberg et al. 1998; Delcher et al. 2007), Prodigal (Hyatt et al. 2010, 2012) and VADR (Schäffer et al. 2020). Each method was given as input the complete, unannotated genome sequence of each of the 468 coronaviruses.i 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. We ran VADR using their recommended parameters as well as their released database for coronaviruses. For CORSID, we used window length $\omega = 7$ and progressively reduced the score threshold τ from 7 to 2. These parameter values were determined from a 5-fold cross validation study (details in supplementary section 2.3, supplementary fig. S10, and supplementary fig. S11). We refer the reader to supplementary results for the precise commands used to run previous tools and details on how the predicted set of genes are compared to ground truth.

Fig. 3a shows that CORSID outperformed Glimmer3 and Prodigal in terms of both precision and recall, and achieved higher recall than VADR. The median precision and recall of CORSID is 0.889 and 1.00, respectively, whereas the median precision and recall is 0.625 and 0.600, respectively, for Glimmer3, 0.714 and 0.636, respectively, for Prodigal. Although VADR has a higher precision (1.00) than CORSID, its median recall is lower (0.900), and its median F_1 score is 0.909, less than CORSID's F_1 score of 0.923.

The same trends are observed when pooling all gene predictions as shown in fig. 3b. CORSID achieved the highest pooled recall (0.926) and F_1 score (0.895), while the precision (0.865) is only slightly lower than VADR's (0.876). The higher precision achieved by VADR can be explained by the fact that its reference database contains 55 coronavirus sequences, 48 of which are included in the 468 complete genomes we test on. If these 48 genomes are removed from the test set, CORSID achieves better overall performance than VADR in the remaining 420 genomes, (supplementary fig. S12).

While Prodigal, Glimmer3, and VADR 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 (supplementary fig. S13). 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 (supplementary fig. S6b). Although VADR found all genes, including three genes missed by CORSID, SARS-CoV-1 is contained in its reference database as mentioned earlier.

In summary, CORSID accurately identifies TRS sites and genes given just the unannotated genome, outperforming existing gene finding methods.

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, Prodigal and VADR. We note that this genome is absent from VADR's reference database. 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. 4). By contrast, VADR found eight genes, missing gene *4b*, covering 88.03% of the genome. Glimmer3 identified a total of six genes spanning 80.52% of the genome, five of which match genes in the BLASTx database. Finally, Prodigal

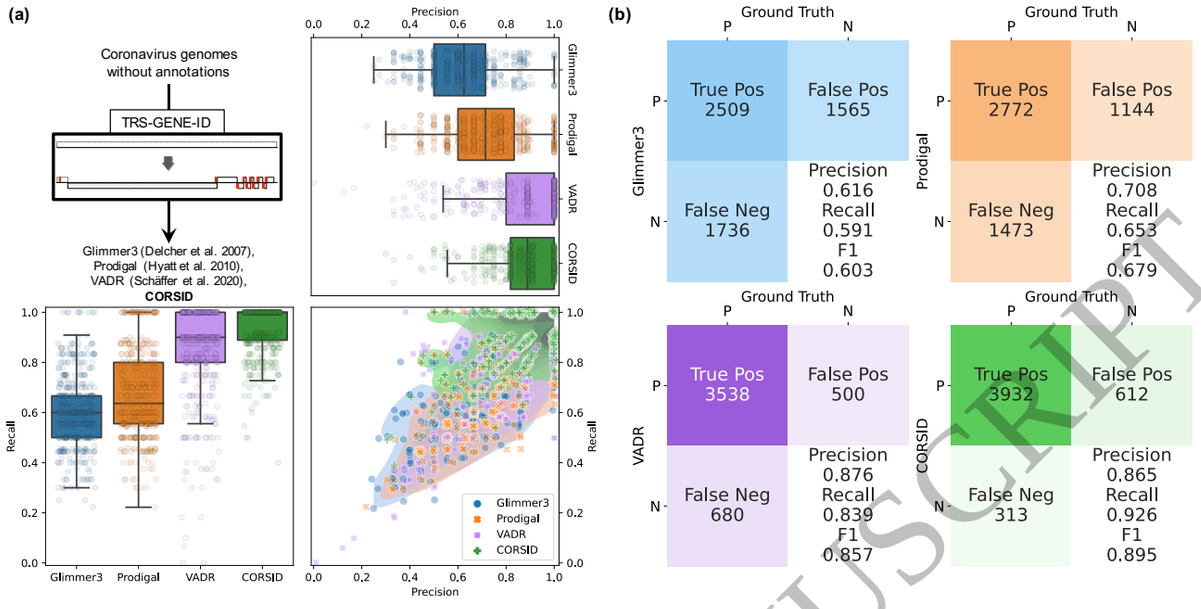

Fig. 3: **CORSID accurately identifies TRS-Ls, TRS-Bs, and genes.** (a) Precision and recall of Glimmer3 (Delcher et al. 2007), Prodigal (Hyatt et al. 2010), VADR (Schäffer et al. 2020) and CORSID for gene prediction in 468 genomes. For clarity, we added a small jitter (drawn from $N(0, 2.5 \times 10^{-5})$) to the 2D distribution plot. (b) Confusion matrices of the ground truth genes and the predicted genes by the each method. Supplementary fig. S8 shows that CORSID incurs a modest reduction in TRS-L accuracy compared to CORSID-A (0.955 vs. 0.996).

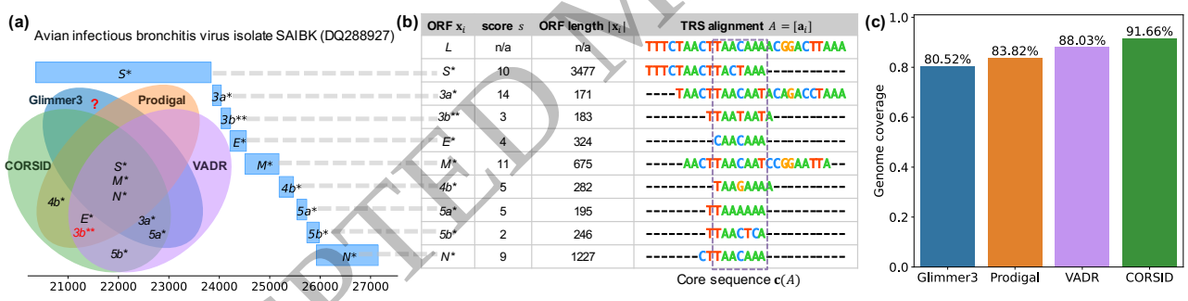

Fig. 4: **CORSID accurately finds genes in an unannotated *Igacovirus* genome (DQ288927).** (a) The position of the genes identified by CORSID. The Venn diagram shows every gene found by CORSID, Glimmer3, Prodigal and VADR. "*" indicates $\geq$ 95% query/hit coverage by BLASTx, "**" indicates a BLASTx hit with query/hit coverage less than 95%, and '?' represents a predicted gene with no BLASTx hit. (b) TRS alignment for genes identified by CORSID. (c) The fraction of positions in $\mathbf{v}_{\text{body}}$ covered by genes identified by the four methods.

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.

Discussion

In this paper, we demonstrated that 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 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 (Smith et al.

1981). 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 (Bailey et al. 2009) and SuPER (Yang et al. 2021). Additionally, we showed that CORSID outperformed two general-purpose gene finding algorithms, Glimmer3 (Salzberg et al. 1998; Delcher et al. 2007) and Prodigal (Hyatt et al. 2010). 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 limitations and avenues for future research. First, the accuracy of identifying genes can be improved by accounting for alternative start codons (supplementary table S3) to improve recall and incorporating Kozak sequence information to improve precision. Second, CORSID is designed for *de novo* gene annotation of novel coronaviruses given only the nucleotide sequence of the genome. However, RNA sequencing data when aligned to the reference genome contain split reads, *i.e.* reads that span non-contiguous regions of the genome, which can be leveraged for identifying candidate regions that contain TRSs. We plan to extend our method by supporting the use of RNA sequencing data to improve gene annotation. Third, CORSID currently requires the complete genome as input to identify the TRS sites and the genes. CORSID can be extended to allow gene identification in the several coronaviruses available in GenBank with only partial reference genomes, such as NC_014470, by leveraging knowledge from other coronaviruses with complete genomes with similar TRS sites. Fourth, 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 (Zirkel et al. 2013; Vasilakis et al. 2014). 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 perform *de novo* assembly jointly with core sequence and TRS site identification, facilitating comprehensive analysis from raw sequencing data of novel coronaviruses.

Materials and Methods

We begin by discussing CORSID-A, which solves the TRS-ID problem. Next, we introduce CORSID, which solves the TRS-Gene-ID problem. Finally, we discuss a web-based visual analytics tool to the space of near-optimal solutions.

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 sgRNA starting at p will lead to the translation of ORF $\mathbf{x}_i$ by the ribosome. SuPER (Yang et al. 2021), 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 $v_{x^- - 170} \cdots v_{x^- - 1}$, *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 leaky scanning into account by skipping over previous ORFs with length smaller than 100 nt (details in supplementary methods section 1.1, supplementary fig. S14 and supplementary fig. S15).

Recall that in a TRS alignment $A = [\mathbf{a}_0, \ldots, \mathbf{a}_n]^\top$ 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 supplementary methods, fig. 5a, and supplementary fig. S16). 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.

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 propose a heuristic to partition a genome $\mathbf{v}$ into $\mathbf{v}_{\text{leader}}$ and $\mathbf{v}_{\text{body}}$, which takes $O(m^2)$ time where m is the number of ORFs in $\mathbf{v}$ that incorporates a classifier to identify truncated genomes missing TRS-L in the 5' UTR (supplementary methods and supplementary fig. S17).

We will now define the relationship between a TRS alignment $A = [\mathbf{a}_0, \ldots, \mathbf{a}_n]^\top$ 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 $\mathbf{c}(A)$. 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 (supplementary fig. S18). As such, we have the following definition.

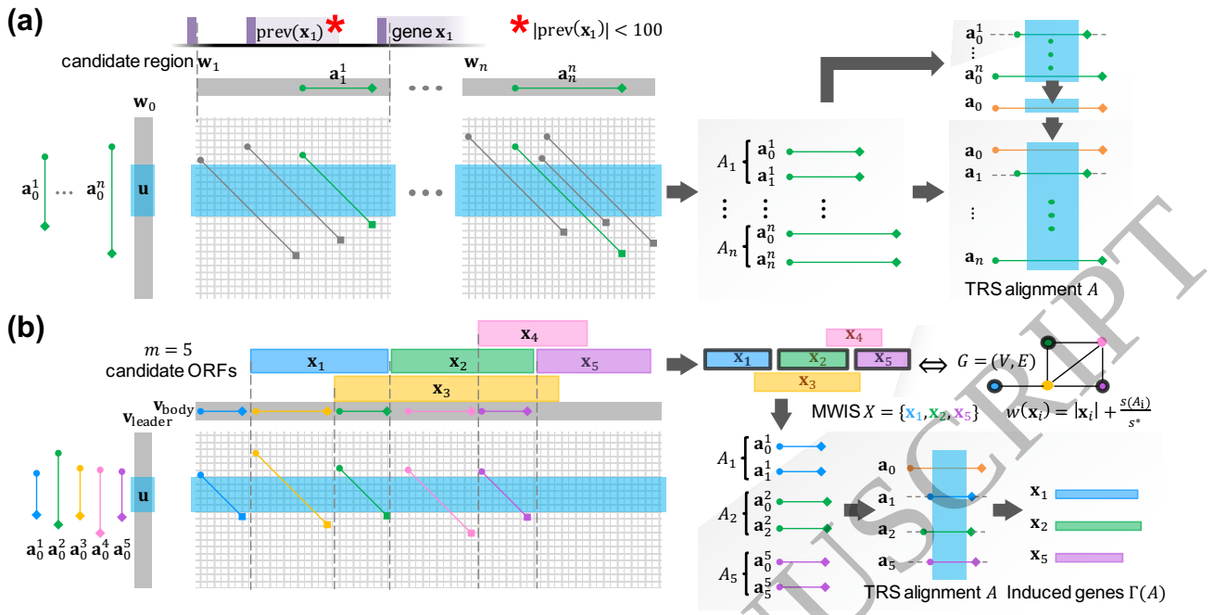

Fig. 5: **Algorithm details.** (a) Given genes $\mathbf{x}_0, \ldots, \mathbf{x}_n$, we obtain candidate regions $\mathbf{w}_0, \ldots, \mathbf{w}_n$ by identifying upstream ORFs, skipping over ORFs if they are of length less than 100 nt (indicated by '*'). CORSID-A solves the TRS-ID problem by sliding 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 (Hsiao et al. 1992) on an interval graph defined by the candidate ORFs to simultaneously identify an optimal pair $(A, \Gamma(A))$ for window $\mathbf{u}$.

Definition 9. Let $A = [\mathbf{a}_0, \ldots, \mathbf{a}_n]^\top$ 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^+ 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 $A = [\mathbf{a}_0, \ldots, \mathbf{a}_n]^\top$, 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 $\mathbf{v}_{\text{body}}$ (supplementary methods and fig. 5b). 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 supplementary methods for details and supplementary fig. S19).

Web Application to Explore Solution Space

In order to present a comprehensive overview of identified TRS sites and genes across solutions, we created a web application that visualizes all solutions and allows for manual annotation. After obtaining solutions from CORSID and CORSID-A, users can launch the application with the output JSON file, then inspect all possible solutions. Specifically, we show a summary table of all solutions, followed by the optimal solution for which we show a sequence logo of the identified TRS-L and TRS-Bs, a genome coverage map, and a detailed table of each identified gene. Users can click the summary table and show other alternative solutions below the fixed optimal solution for comparison. A demo of the visualization can be found at `https://elkebir-group.github.io/CORSID-viz`. We also made an integrated dockerized workflow including CORSID, CORSID-A, BLASTx, and the visualization web

application. After obtaining the docker images, users can easily analyze a new genome by running the workflow without any manual configuration. Additional details about the scope and recommendations for using CORSID and CORSID-A, including the combination of running CORSID-A after VADR, are provided in Section 2.7 of the supplement (supplementary fig. S20).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgements.

This material is based upon work supported by the National Science Foundation under award numbers CCF-1850502, CCF-2027669 and CCF-2046488. This work used resources, services, and support provided via the Greg Gulick Honorary Research Award Opportunity supported by a gift from Amazon Web Services.

References

Sara Alonso, Ander Izeta, Isabel Sola, and Luis Enjuanes. Transcription regulatory sequences and mRNA expression levels in the coronavirus transmissible gastroenteritis virus. *Journal of Virology*, 76(3):1293–1308, 2002.

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.

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.

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

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.

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.

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.

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.

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.

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.

Irwin Jungreis, Rachel Sealfon, and Manolis Kellis. SARS-CoV-2 gene content and COVID-19 mutation impact by comparing 44 Sarbecovirus genomes. *Nature Communications*, 12(1): 1–20, 2021.

Dongwan Kim, Joo-Yeon Lee, Jeong-Sun Yang, Jun Won Kim, V Narry Kim, and Hyeshik Chang. The architecture of SARS-CoV-2 transcriptome. *Cell*, 181(4):914–921, 2020.

Roberto Patarca and William A Haseltine. Intragenomic rearrangements of sars-cov-2 and other β-coronaviruses. *bioRxiv*, 2022.

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.

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.

Palash Sashittal, Chuanyi Zhang, Jian Peng, and Mohammed El-Kebir. Jumper enables discontinuous transcript assembly in coronaviruses. *Nature Communications*, 12(6728), 2021.

Alejandro A Schäffer, Eneida L Hatcher, Linda Yankie, Lara Shonkwiler, J Rodney Brister, Ilene Karsch-Mizrachi, and Eric P Nawrocki. VADR: validation and annotation of virus sequence submissions to GenBank. *BMC Bioinformatics*, 21: 1–23, 2020.

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

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. ISSN 0022-538X, 1098-5514. doi: 10.1128/JVI.79.4.2506-2516.2005. URL `https://JVI.asm.org/content/79/4/2506`.

Isabel Sola, Fernando Almazan, Sonia Zúñiga, and Luis Enjuanes. Continuous and discontinuous RNA synthesis in coronaviruses. *Annual Review of Virology*, 2:265–288, 2015.

Nikos Vasilakis, Hilda Guzman, Cadhla Firth, Naomi L Forrester, Steven G Widen, Thomas G Wood, Shannan L Rossi, Elodie Ghedin, Vsevolov Popov, Kim R Blasdell, et al. Mesoniviruses are mosquito-specific viruses with extensive geographic distribution and host range. *Virology Journal*, 11(1):1–12, 2014.

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.

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

Florian Zirkel, Hanna Roth, Andreas Kurth, Christian Drosten, John Ziebuhr, and Sandra Junglen. Identification and characterization of genetically divergent members of the newly established family Mesoniviridae. *Journal of Virology*, 87(11):6346–6358, 2013.