GUIdEStaR (G-quadruplex, uORF, IRES, Epigenetics, Small RNA, Repeats), the integrated matadatabase: transcript-indexed binary information creation for chaining with neural network analysis

Jee Eun Kang[1 §]

[1]BK21 4th phase, Life Sciences Education and Research, Korea University, Seoul, Republic of Korea, Division of Life Sciences, Korea University, Seoul, Republic of Korea

[§] Corresponding author

Email addresses:

JEK: jekang_39@yahoo.com

Abstract

GUIdEStaR integrates existing databases of important compositional and structural elements of sequences- various types of G-quadruplex, upstream open reading frame (uORF), Internal Ribosome Entry Site (IRES), epigenetic modification (histone protein and RNA), and repeats. It contains binary information (presence/absence of the elements) that are organized into 5 regions (5'UTR, 3'UTR, exon, intron, and biological region) per transcript and per gene. These elements are highly interdependent in controlling functional interaction of a gene. The database contains information of approx. 40,000 genes and 320,000 transcripts, where each transcript has 845 presence/absence information. Recently, artificial intelligence (AI) based analysis of sequencing data has been gaining popularity in the area of bioinformatics. To create a dataset that can be used as an input to AI methods, GUIdEStaR comes with example Java codes. Here, we demonstrates the database usage with three neural network classification examples- 1) small RNA example for identifying the attributes that are unique to transcription factor (TF) genes mediated by small RNAs originated from SARS-CoV-2 vs. from human, 2) cell membrane receptor study for classifying virus interacting vs. non-interacting

receptors, and 3) receptors targeted by nonsense mediated mRNA decay (NMD) vs. of non-target. GUIdEStaR is freely available at www.guidestar.kr and https://sourceforge.net/projects/guidestar.

Background

Composition and structure of nucleotide sequences of a gene have a variety of elements that are interwoven for orchestrated gene function. It is important to identify which elements are present in a gene to understand its function. Here, we introduce the integrated database, GUIdEStaR that integrates information of **G**-quadruplex, **u**ORF, **I**RES, **E**pigenetics, **S**mall RNA, and **R**epeats into a single resource [1-7]. It was created based on existing databases and indexed the information by transcript ID and gene symbol for simple data retrieval.

GQ is topolgical change required of many of conserved regulatory elements [8]. It has become increasingly important in recent years as its diverse roles in biological functions have been unveiled in epigenetic modification: it is revealed that GQ is mediator in epigenetic modification: histone, DNA, and RNA modification [9]. uORFs are common feature in corona-virus and in mammals; it has been discovered that half of human transcripts have uORFs [10,11]. Furthermore, it is discovered that miRNA regulates a specific type of genes carrying uORFs [12]. Another element, IRES along with an uORF is known to select a specific RNA variant for translation and cellular localization [13]. Small RNA has been extensively studied for their major roles in gene regulation and epigenetic modulator. Its binding affinity with target is regulated by epigenetic modification [14]. Genes containing small RNA loci implicate various functions in cellular signaling and response through controlling epigenetic modification [15,16]. Repeats, the component with a largest number of attributes in GUIdEStaR, have been studied as gene regulator in transcription regulation repertoire [17,18,19]. They also function as recognition signature for cellular localization of a transcript variant [20]. Finally, the last component of the database is epigenetic modification: presence/ absence information of the target sites of the RNA m6A writers, readers, and erasers- Fto, KIAA1429, METTL3, METTL14, WTAP, YTHDC1, and YTHDF1 and of histone modification. GUIdEStaR provides example Java codes that can be used as guides to access 845 attributes

(169 presence/absence information from each of 5'UTR, 3'UTR, exon, intron, and biological region) of a gene. It consists of approx. 40,000 genes and 320,000 transcripts. It creates a dataset that can be used as an input to AI methods.

Covid 19 pandemic has changed our lives globally. It is important to understand what functions small RNAs that are originated from corona-virus play to reprogram host transcription [21, 22]. SARS-CoV-2 encoded vs. human encoded small RNAs may have different requirements or affinities for target interaction. Here, we demonstrate three example studies with neural network methods- classification and attribute selection. In addition to GUIdEStaR, differentially expressed gene (DEG) information of eight different cell types (ciliated cells, basal cells, club cells, goblet cells, neuroendocrine cells, ionocytes, and tuft cells) from single-cell (sc) RNA-seq were included in the datasets [23,24]. Presence/absence information of G-quadruple (PU1), SINE (Short interspersed nuclear element) family, and m^6A epigenetic modification sites of WTAP, METTL3, and Fto, and log2 (fold change) (log2 FC) of DEG information were among top ranked attributes that contributed to classifying intron regions into either one of human TF genes mediated by SARS-COV-2 small ncRNA or by human miRNA. The binary neural network achieved the accuracy of 91% (Table 1, Table S1, supplementary data).

Recent study documented that protease functioned as co-receptor to facilitate virus entry, and interacted with methyltransferase to modulate activation of receptor [25,26]. It is known that cell membrane receptor bound by a growth factor increases transcript abundances of specific TFs and proteases by the intracellular signaling cascades [27,28]. In addition to data from GUIdEStaR, we included binary information of log2 FC of DEG data of diverse TFs, growth factors, protease, and methyltransferases in comparison to those of a receptor sample. Interestingly, top attributes were mostly comparison data of log2 FC of DEG for binary classification of intron regions of virus interacting cell membrane receptors vs. of non-interacting receptors, which resulted in accuracy of 78%. In contrast, top 10 attributes selected in cell membrane receptor vs. nuclear receptor study and the receptors of NMD target vs. of non-NMD target study were elements of sequence composition and epigenetic modification (repeats, small RNA, and RNA m6A modification sites) in addition to binary

information of log2 FC of DEG data, which resulted in the accuracies were 88% and 77%, respectively. GUIdEStaR is available at www.guidestar.kr and https://sourceforge.net/projects/guidestar.

Implementation

Main DB creation:

Ensembl gff file (hg38 version: 101) was used to determine transcript location. A transcript ID (Ensembl ENST) file consisted of 845 attributes where each of 5UTR, 3UTR, exon, intron, and biological region contained169 presence/ absence attributes, if the transcript was mRNA. ENST files were stored in the folders of its parent gene ID (Ensembl ENSG) and gene symbol. Intergenic regions were included at both ends of a gene. If it was overlapped with another gene, approx. 4000 nucleotides in upstream and downstream were included [29]. GQ, uORF, IRES, Epigenetics, small RNA, and repeats information were mapped to each transcript. Specifically, GQ locations provided by Chariker et al. in bed format, uORF locations from uORF-Tools in gff format, gene information from IRESbase, epigenetic modification to the Histone H3 from Choi et al., presence/absence of target sites of m6A writers, readers, and erasers from Trew database of Met-DB V2.0, locations of small RNAs from DASHR 2.0, and different types of repeats from RepeatMasker were included in the GUIdEStaR. G-quadruplex, uORF, IRES, small RNA, and repeats have different binary values per region (ex: intron) per transcript, while epigenetic information (Trew database) have different values per region and per gene. Histone epigenetic modification have the same values for all five regions for a given gene. The database contains example Java codes to retrieve the presence/absence information according to input gene symbols and to create a datasheet that can be readily inputted to neural network methods.

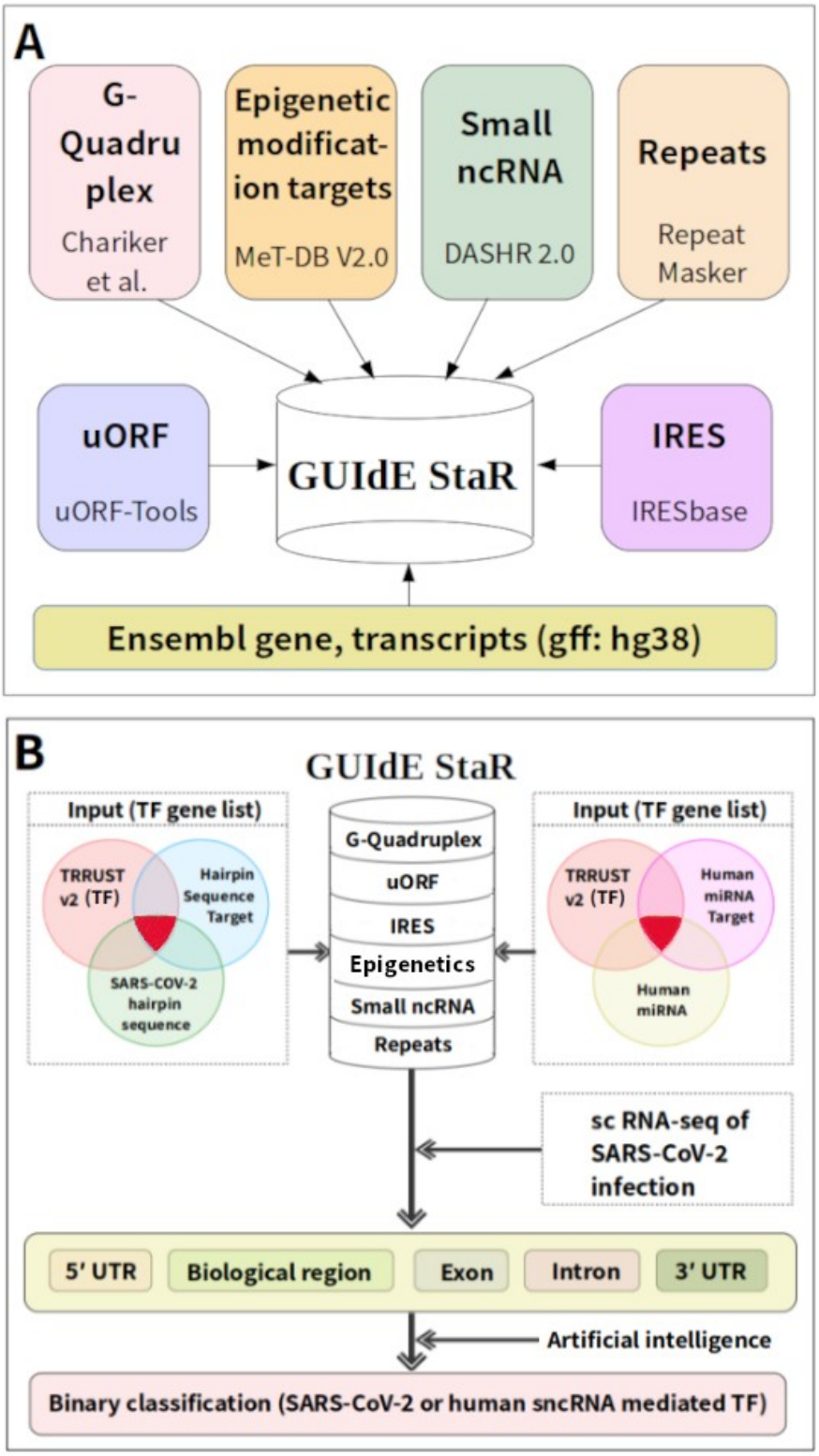

Fig1. (A) Data contained in GUIdEStaR. (B) Example demonstration workflow of the small RNA study

Database usage demonstration:

The presence/ absence data of 845 attributes of each TF gene were retrieved from the GUIdEStaR, and reorganized into 5 datasets in the order of five prime UTR, three prime UTR, exon, intron, biological_region. For example, the first 169 presence/absence information of the GUIdEStaR were from five prime UTR region of a transcript. In small RNA study dataset, we included log2 FC of DEG of infected cell, bystander cell, and control cell measured at day 1,2, and 3 in eight different cell types in scRNA-seq data of SARS-CoV-2 infection [24]. We built binary classifiers using Weka's meta classifier, MultiClassClassifierUpdateable with default options (method: against-all, random number seed: 1, base classifier: Stochastic Gradient Descent (SGD), loss function: hinge loss (SVM), learning rate: 0.01, lambda: 1.0E-4, epochs: 500, epsilon: 0.001, normalization) for classifiers [30]. Resampling preprocessing step was applied in prior of classification due to the small sample number of datasets. CorrelationAttributeEval that ranked attributes in terms of its correlation with the class was used for attribute selection [30]. For receptor studies, 1048 binary values were generated by the same sign function that assessed whether x and y had the same sign where x and y were the log2 FCs of a receptor sample gene (row) and an attribute gene (column) in a dataset, respectively. An attribute gene belonged to one of the following families: growth factor families (bone morphogenetic protein, ciliary neurotrophic factor, LIF, interleukin, epidermal growth factor, erythropoietin, fibroblast growth factor, GDNF family, IGF like family, macrophage, nerve growth factor, neurotrophin, platelet derived growth factor, T cell, megakaryocyte, transforming, and tumor necrosis factor), TF families (AP-2, ARID/BRIGHT, AT hook, BED ZF, Brinker, C2H2 ZF, C2H2 ZF AT hook, C2H2 ZF BED ZF, C2H2 ZF Homeodomain, C2H2 ZF Myb/SANT, CBF/NF-Y, CCCH ZF, CENPB, CG-1, CSD, CSL, CUT Homeodomain, CxxC, CxxC AT hook, DM, E2F, EBF1, Ets, Ets AT hook, FLYWCH, FYVE-type ZF, Forkhead, GATA, GTF2I-like, Grainyhead, HMG/Sox, HMG/Sox bHLH, HSF, Homeodomain, Homeodomain POU, Homeodomain Paired box, IRF, LYAR-type C2H2 ZF, MADF, MADS box, MBD, MBD AT hook, MBD CxxC ZF, MYM-type ZF, MYND-type ZF, Myb/SANT, Myb/SANT GATA, NFX, NOA36-type ZF, Ndt80/PhoG, Nuclear receptor, Paired box, Pipsqueak, Prospero, RFX, Rel, Runt, SAND, SART-1, SBP, SMAD, STAT, T-box, TBP, TCR/CxC, TEA, THAP finger, Unknown, ZZ-type ZF, bHLH,

bZIP, mTERF, and p53), protease (peptidase, transamidase, retropepsin, separase, elastase, glutamyltransferase, transglutaminase, hydrolase, serine kinase, CTP synthetase, thioesterase, eupitrilysin, cathepsin, CTSLL, interphotoreceptor retinoid, LIPF, insulysin, urokinase, caspase, dihydropyrimidinase, thrombin, ovochymase, pepsin, calpain, otubain, methylesterase, spinesin, matriptase, hin1L, imidazolonepropionase, lysosomal, chymase, granzyme, proteasome subunit, abhydrolase, furin, signalase, testisin, tryptase, prosemin, polyserase, polycystin, kallikrein, paracaspase, marapsin, nicalin, hepsin, napsin, latrophilin, nardilysin, presenilin, renin, acrosin, TESP, trypsin, PRSS, TRPM), histone methyltransferases, DNA methyltransferases, RNA methyltransferases, or enzymes modifying amino acid residues.

Small RNA study:

Small RNAs in SARS-CoV-2 and in human were selected based on the study of Demirci et al. [31]. They developed a machine learning method and attempted to identify human target genes mediated by small RNAs encoded by SARS-CoV-2 and virus target genes mediated by human miRNA. The small RNAs of viral and of human origin and TF target genes mediated by SARS-CoV-2 encoded small RNAs were selected based on the study, while those mediated by the human miRNA were based on miRDB and TRRUST v2 [32,33].

Virus interacting cell membrane receptor study:

For gene samples, cell membrane receptors that were known to bind to growth factors were included in the data. Among the receptors, those interacting with different types of viruses were selected based on the database, Viral Receptor [34].

Cell membrane and nuclear receptors of NMD target vs. non-target:

Nuclear receptors were selected based on wikipedia (https://en.wikipedia.org/wiki/Nuclear_receptor). Genes containing NMD-target transcripts were selected based on gene annotation of gff file from Ensembl Homo sapiens 101.

Database availability

GUIdEStaR is freely available to academic and non-academic communities. GUIdEStaR and example codes written in Java are provided at www.guidestar.kr and www.sourceforge.net (https://sourceforge.net/projects/guidestar/).

Example results and discussion

Accuracies of each binary classifier estimated with the cross-validation (folds=10) are shown in Table 1 (Table S1). Top 10 attributes with high discriminating power in the intron dataset in the small RNA study were GQ (PU1), epigenetic modification (WTAP overlap sncRNA, WTAP t2 overlap five prime UTR, Fto t2Fto overlap five prime UTR, WTAP overlap lncRNA), repeats (SINE-tRNA-Deu, SINE-5S-Deu-L2), and log2 FC of DEG data (basal cells, tuft cells, and ciliated cells of infected/bystander cell data) (Supplementary data).

Table 1. Accuracies of binary classifiers

Accuracy	5'UTR	Biol. Reg.	Exon	Intron	3'UTR
Small RNA targets (virus origin vs.- human origin)	88 %	88 %	93 %	91 %	89 %
Receptors with virus interaction vs. receptors without virus interaction	73 %	68 %	72 %	78 %	71 %
Cell membrane receptors vs. nuclear receptors	83 %	80%	85 %	88 %	82 %
Receptors of NMD target vs. receptors of non-NMD target	75 %	70%	69 %	77 %	70 %

It was interesting to find that top 10 attributes in virus vs. non-virus receptors were all from ciliated cells of scRNA-seq data: binary information of the DEG data of receptor samples and the following attribute families- protease family (nicalin), transcription factor family (NFX), growth factor families (erythropoietin, epidermal growth factor, ciliary neurotrophic factor, fibroblast growth factor, GDNF family, neurotrophin, T cell, platelet derived growth factor). It was consistent with the scRNA-seq study that ciliated cells were mainly the targets of the SARS-COV-2 infection [24]. Top 10 attributes in the cell membrane and nuclear receptor study were repeats (DNA, Unknown), epigenetic modification (YTHDF1 overlap three prime UTR), and binary information of DEG data of the protease, TF, growth factor, and

methyltransferase families- granzyme, CxxC AT hook, CSL, platelet derived growth factor, CBF/NF-Y, histone methyltransferase gene (NSD1), methylesterase) in basal cells. It was noteworthy that top 10 attributes in the cell membrane and nuclear receptor study included the binary data of DEG only from basal cells, which was in stark contrast with those in virus vs. non-virus receptor study. Top 10 attributes in the receptors of NMD target and non-target study were repeats (LTR, DNA-TcMar, LTR-ERV1?, LINE-L2, DNA-hAT-Blackjack, LTR-ERVL-MaLR), RNA, and epigenetic modification (METTL14 overlap mRNA, METTL14 overlap three prime UTR, YTHDF1 overlap mRNA).

Although omics data availability rapidly increases, the bottleneck of the data usage lies in the lack of bioinformatics software and databases that provide users with a convenient means to extract meaningful information that conveys valuable knowledge. GUIdEStaR provides an easy access to the key composition and structure elements of a transcript/gene with a single inquiry. Neural network methods such as association rule may reveal useful insights for studying requirements interdependencies of the elements in a given transcript. Due to the highly dynamic nature of epigenetic modification, association rule mining among compositional and structural sequence features and epigenetic modification may reveal a required set of particular features depending on a specific condition of cellular microenvironment. In future, we may develop a bioinformatics tool for identifying such features at single nucleotide resolution.

References

1. Chariker JH, Miller DM, Rouchka EC. Computational Analysis of G-Quadruplex Forming Sequences across Chromosomes Reveals High Density Patterns Near the Terminal Ends. PLoS One. 2016 Oct 24;11(10):e0165101.

2. Scholz A, Eggenhofer F, Gelhausen R, Grüning B, Zarnack K, Brüne B, Backofen R, Schmid T. uORF-Tools-Workflow for the determination of translation-regulatory upstream open reading frames. PLoS One. 2019 Sep 12;14(9):e0222459.

3. Zhao J, Li Y, Wang C, Zhang H, Zhang H, Jiang B, Guo X, Song X. IRESbase: A Comprehensive Database of Experimentally Validated Internal Ribosome Entry Sites. Genom. Proteom. Bioinform. 18(2): 129-139 (2020).
4. Choi S. H. , Heo K. , Byun H. M. , An W. , Lu W. , Yang A. S. (2011). Identification of preferential target sites for human DNA methyltransferases. Nucleic Acids Research, 39, 104–118. 10.1093/nar/gkq774
5. Liu H, Wang H, Wei Z, Zhang S, Hua G, Zhang SW, et al. MeT-DB V2.0: elucidating context-specific functions of N6-methyl-adenosine methyltranscriptome. Nucleic Acids Res. 2018;46(D1):D281–D2D7.
6. Kuksa PP, Amlie-Wolf A, Katanic´ Zˇ , Valladares O, Wang LS, Leung YY. DASHR 2.0: integrated database of human small non-coding RNA genes and mature products. Bioinformatics. 2019 Mar 15;35(6):1033-1039.
7. Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013-2015 <http://www.repeatmasker.org>.
8. Kaplan OI, Berber B, Hekim N, Doluca O. G-quadruplex prediction in E. coli genome reveals a conserved putative G-quadruplex-Hairpin-Duplex switch. Nucleic Acids Res. 2016 Nov 2; 44(19): 9083–9095.
9. Mukherjee AK, Sharma S, Chowdhury S. Non-duplex G-Quadruplex Structures Emerge as Mediators of Epigenetic Modifications. Trends Genet. 2019 Feb;35(2):129-144.
10. Somers J, Pöyry T, Willis AE. A perspective on mammalian upstream open reading frame function. Int J Biochem Cell Biol. 2013 Aug;45(8):1690-700.
11. Wu HY, Guan BJ, Su YP, Fan YH, Brian DA. Reselection of a genomic upstream open reading frame in mouse hepatitis coronavirus 5'-untranslated-region mutants. J Virol. 2014 Jan;88(2):846-58.
12. Baldrich P, Campo S, Wu MT, Liu TT, Hsing YI, San Segundo B. MicroRNA-mediated regulation of gene expression in the response of rice plants to fungal elicitors. RNA Biol. 2015;12(8):847-63.

13. Alexandrova J, Paulus C, Rudinger-Thirion J, Jossinet F, Frugier M. Elaborate uORF/IRES features control expression and localization of human glycyl-tRNA synthetase. RNA Biol. 2015 Dec; 12(12): 1301–1313.
14. Morales L, Oliveros JC, Fernandez-Delgado R, tenOever BR, Enjuanes L, Sola I. SARS-CoV-Encoded Small RNAs Contribute to Infection-Associated Lung Pathology. Cell Host Microbe. 2017 Mar 8;21(3):344-355.
15. Aravin AA, Sachidanandam R, Bourc'his D, Schaefer C, Pezic D, Toth KF, Bestor T, Hannon GJ. A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. Mol Cell. 2008 Sep 26;31(6):785-99.
16. Pasero P, Bensimon A, Schwob E. Single-molecule analysis reveals clustering and epigenetic regulation of replication origins at the yeast rDNA locus. Genes Dev. 2002 Oct 1;16(19):2479-84.
17. Lu JY, Shao W, Chang L, Yin Y, Li T, Zhang H, Hong Y, Percharde M, Guo L, Wu Z, Liu L, Liu W, Yan P, Ramalho-Santos M, Sun Y, Shen X. Genomic Repeats Categorize Genes with Distinct Functions for Orchestrated Regulation. Cell Rep. 2020 Mar 10;30(10):3296-3311.e5.
18. Chen HY, Ma SL, Huang W, Ji L, Leung VH, Jiang H, Yao X, Tang NL. The mechanism of transactivation regulation due to polymorphic short tandem repeats (STRs) using IGF1 promoter as a model. Sci Rep. 2016 Dec 2;6:38225.
19. Thompson PJ, Macfarlan TS, Lorincz MC. Long Terminal Repeats: From Parasitic Elements to Building Blocks of the Transcriptional Regulatory Repertoire. Mol Cell. 2016 Jun 2;62(5):766-76.
20. Buckley PT et al., Cytoplasmic Intron Sequence-Retaining Transcripts Can Be Dendritically Targeted via ID Element Retrotransposons. Neuron. 69, 877–884 (2011).
21. Merino GA, Raad J, Bugnon LA, Yones C, Kamenetzky L, Claus J, Ariel F, Milone DH, Stegmayer G. Novel SARS-CoV-2 encoded small RNAs in the passage to humans. Bioinformatics. 2020 Nov 27;btaa1002.
22. Morales L, Oliveros JC, Fernandez-Delgado R, tenOever BR, Enjuanes L, Sola I. SARS-CoV-Encoded Small RNAs Contribute to Infection-Associated Lung Pathology. Cell Host Microbe. 2017 Mar 8;21(3):344-355.

23. Tabansky I, Stern JN, Pfaff DW. Implications of Epigenetic Variability within a Cell Population for "Cell Type" Classification. Front Behav Neurosci. 2015 Dec 16;9:342.
24. Ravindra NG, Alfajaro MM, Gasque V, Habet V, Wei J, Filler RB, Huston NC, Wan H, Szigeti-Buck K, Wang B, Wang GQ, Montgomery RR, Eisenbarth SC, Williams AJ, Pyle AM, Iwasaki A, Horvath TL, Foxman EF, Pierce RW, van Dijk D, Wilen CB. 2020. Single-cell longitudinal analysis of SARS-CoV-2 infection in human airway epithelium. bioRxiv doi: 10.1101/2020.05.06.081695.
25. Qi F., Qian S., Zhang S., Zhang Z. Single cell RNA sequencing of 13 human tissues identify cell types and receptors of human coronaviruses. Biochem. Biophys. Res. Commun.
26. Alumkal JJ, Herman JG. Distinct epigenetic mechanisms distinguish TMPRSS2-ERG fusion-positive and -negative prostate cancers. Cancer Dis. 2012;2:979–981.
27. Kar A., Gutierrez-Hartmann A. Molecular mechanisms of ETS transcription factor mediated tumorigenesis. Crit. Rev. Biochem. Mol. Biol. 2013;48:522–543. doi: 10.3109/10409238.2013.838202.
28. Henkart PA. ICE family proteases: mediators of all apoptotic cell death? Immunity. 1996;4(3):195–201.
29. 3 Characterization of intergenic regions and gene definition. *Nature* (2019). https://doi.org/10.1038/nature28172
30. Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.
31. Saçar Demirci MD, Adan A. Computational analysis of microRNA-mediated interactions in SARS-CoV-2 infection. PeerJ 2020;2020:1−17.
32. Chen Y, Wang X. miRDB: an online database for prediction of functional microRNA targets. Nucleic Acids Res. 2020 Jan 8;48(D1):D127-D131.
33. Han H, Cho JW, Lee S, Yun A, Kim H, Bae D, Yang S, Kim CY, Lee M, Kim E, Lee S, Kang B, Jeong D, Kim Y, Jeon HN, Jung H, Nam S, Chung M, Kim JH, Lee I. TRRUST v2:an expanded

reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res 46(D1):D380–D386 (2018).
34. Zhang Z., Zhu Z., Chen W., Cai Z., Xu B., Tan Z., Wu A., Ge X., Guo X., Tan Z., et al. Cell membrane proteins with high N-glycosylation, high expression and multiple interaction partners are preferred by mammalian viruses as receptors. Bioinformatics. 2019;35:723–728.