GUIdeStaR (G-quadruplex, Uorf, IRES, Small RNA, Repeats), bioinformatics tool for gene characterization- case study: development of attribute selection and classification methods based on GuideStar for studying human transcription factor genes mediated by SARS-COV-2 small ncRNA

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

Recently, artificial intelligence (AI) based analysis of sequencing data is gaining popularity in the area of bioinformatics. This paper introduces a ready-to-plug-in-to-AI database, GuideStar, integrated with five important nucleotide elements and structure, G-quadruplex, Uorf, IRES, Small RNA, Repeats. We demonstrated the database usage with attribute selection and classification methods using Weka software in identifying the features that are unique to transcription factor genes (TF genes) mediated by small RNAs originated from SARS-CoV-2 in comparison to those from human. G-quadruple (18:15:4, 14:11:3, PU1, 6:3:1), SINE (Short interspersed nuclear element) family, LINE (Long interspersed nuclear element) family, hAT family, and Tc1/mariner family were selected as top ranked attributes. However, statistical confidence was not measured due to small sample size (N: 7 vs 49) because a small number of TF genes have been identified as being mediated by small RNAs with viral origin. GuideStar is freely available at www.guidestar.kr.

Background

Structural element such as G-quadruplex (GQ) and nucleotide elements such as repeats and small RNAs are well known gene regulators. GQ is topolgical change required of many of conserved regulatory elements [1]. 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 and DNA modification and RNA modification [2]. Upstream open reading frames (uORFs) are common feature in corona-virus and in mammals; it has been discovered that half of human transcripts have uORFs [3,4]. It is shown that miRNA regulates a specific type of uORF containing gene [5]. Internal Ribosome Entry Site (IRES) and an uORF control selecting specific RNA variant for translation and localization [6]. There are different types of Small RNAs playing diverse functions. Genes containing these loci implicate various functions in cellular signaling and response through controlling epigenetic modification [7,8]. In addition, repeats have been studied as gene regulator, serving as one of members in transcription regulation repertoire [9,10,11]. Features of nucleotide element and structure vary widely across gene types, and they are interwoven into orchestrated gene regulation.

Currently, there is no existing bioinformatics software that integrates these features into a single resource. Here, we introduce a bioinformatics database, GuideStar (G-quadru, Uorf, IRES, Small RNA, Repeats). It contains GQ maps from Chariker et al., uORF-Tools, IRESbase, DASHR 2.0, and repeats from RepeatMasker [12-16]. It provides Easy-to-Use tool by integrating the resources into the presence/ absence of 520 attributes (104 attributes from each of 5'UTR, biological region, exon, intron, and 3'UTR) of a gene, for total of approx. 60,000 genes (Table S1). It is readily applicable to artificial intelligence (AI) methods.

Covid 19 pandemic has changed our lives globally. Since January 2020 (formal news publication on Covid 19 by WHO), variants with a large number of mutations have emerged. It is important to understand what functions small RNAs originated from corona-virus play to reprogram host transcription [17, 18]. We included attribute selection in SARS-CoV-2 infection as a case study to demonstrate the database usage. SARS-CoV-2 Encoded vs. human encoded Small RNAs may have different requirements or affinities for target

interaction in gene regulation, where the features of nucleotide element and structure may play orchestrated roles. We created datasets based on the GuideStar, and compared human transcription factor genes (TF genes) mediated by SARS-COV-2 small ncRNA vs. by human miRNA, using attribute selection and classification methods. Small RNAs are known as important gene regulator. Small ncRNA is also known as epigenetic modulator. Additionally, its binding affinity with target is regulated by epigenetic modification [18].

Epigenetic modification may serve as cell type determinant partially due to the fact that epigenetic modification is strongly related to cell origin and development [19]. We included differential gene expression (DGE) information of eight different cell types from single-cell (sc) RNA-seq provided by Ravindra et al. [20]. They provided longitudinal analysis of the DGE in infected cell and bystander cell. As cell tropism of the viral infection to other cell types also varied over the course of infection, the addition of the scRNA-seq data to our dataset increased the accuracy of the classification [20]. We attempted to identify important features of TF genes mediated by SARS-COV-2 small ncRNA. Attribute selection methods identified various types of GQ and repeats, for example, 18:15:4, 14:11:3, PU1, 6:3:1, SINE (Short interspersed nuclear element) family, LINE (Long interspersed nuclear element) family, hAT family, and Tc1/mariner family. However, statistical significance was not measured due to the small sample number (N: 7 vs. 49).

Implementation

Main DB creation:

Ensembl gff file (hg38 version: 101) was used to determine transcript location. Transcript ID- ENST file contained 520 attributes where 104 presence/ absence attributes were arranged in the order of 5UTR, 3UTR, exon, intron, biological region. ENST was stored in its parent ENSG and gene folders. 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 [21]. GQ, Uorf, IRES, 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, locations of small

RNAs (microRNA (miRNA), Piwi-interacting RNA (piRNA), small nuclear RNA (snRNA), nucleolar RNA (snoRNA), cytoplasmic RNA (scRNA), transfer RNA (tRNA), tRNA fragment (tRF), and ribosomal RNA (rRNA)) from DASHR 2.0, and 64 different types of repeats from RepeatMasker were included in the main DB (Data S1). The database contains an example Java code to create a dataset from the main DB.

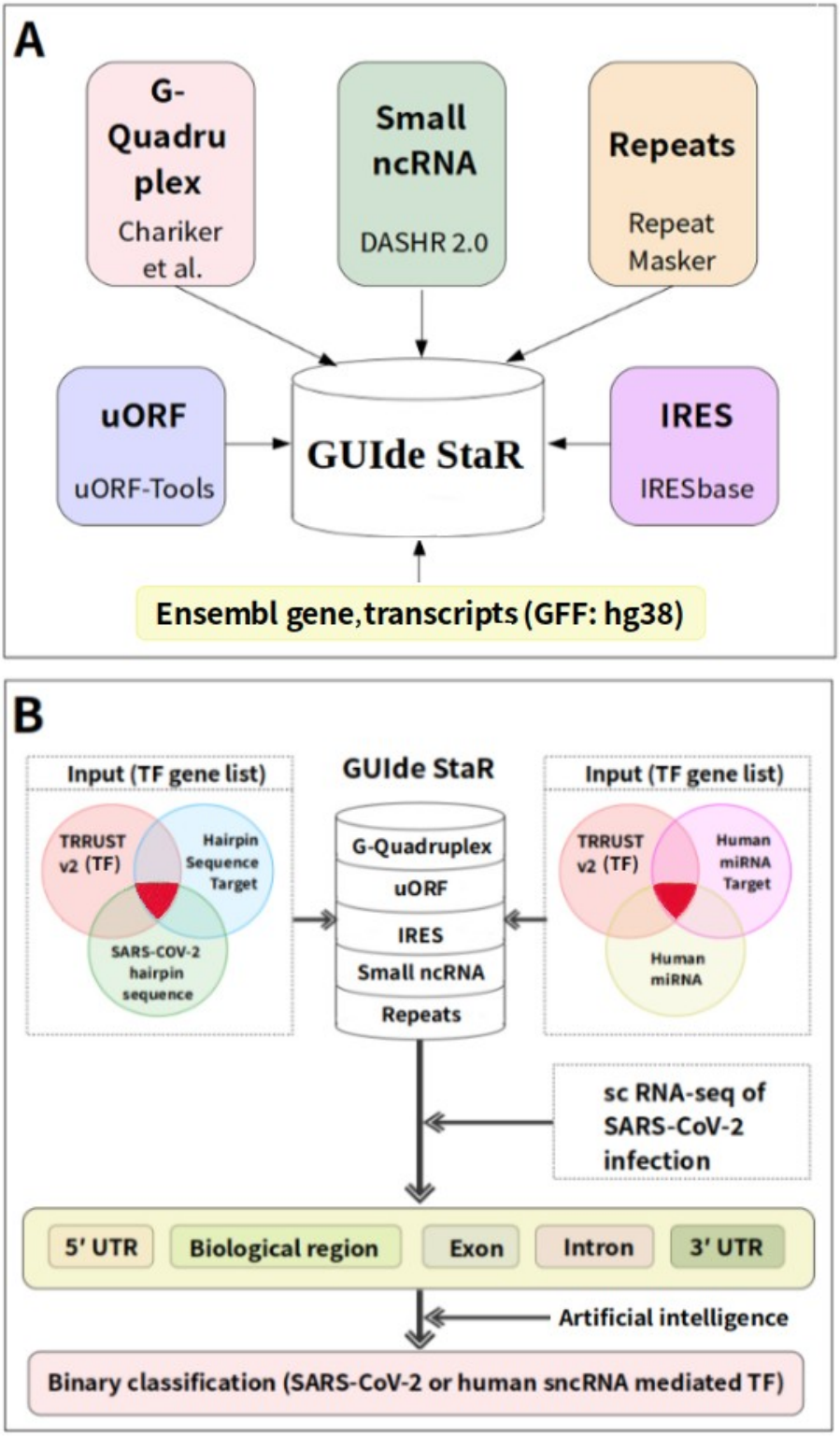

Fig1. (A) Data contained in GuideStar. (B) Case study workflow

Case Study:

Small RNAs in SARS-CoV-2 and in human were selected based on the study of Demirci et al. [22]. They developed a machine learning method and attempted to identify human target genes mediated by SARS-CoV-2 encoded small RNAs and virus target gene by human miRNA. The small RNAs of viral and of human origin and target genes mediated by SARS-CoV-2 encoded small RNAs were selected based on the study, while those by the human miRNA were based on miRDB [23]. TF genes were selected from the target genes based on TRRUST v2 [24]. The presence/ absence data of 520 attributes of each TF gene were retrieved from the main DB. They were reorganized into 5 datasets (5'UTR, 3'UTR, exon, intron, biological_region). In each dataset, we included log2 (fold change) (log2 FC) of DGE 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 [20]. We built a binary classifier that classifies a gene into either of groups medicated by virus originated small RNA or by human originated small RNA. We used 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 binary classifier and for multi-class classifier [25]. CorrelationAttributeEval that ranks attributes in terms of its correlation with the class was used for attribute selection [25]. After we excluded TF genes containing missing values, there were only 7 instances of TF genes mediated by virus originated small RNAs and 49 instances of TF genes by human originated small RNAs. Due to the small number of instances in virus origin, we analyzed attributes with high discriminative power with attribute selection method instead of cross-validation.

(a) Filter, ClassBalancer was applied in preprocess step in order to make each class have the same total weight [25] in prior to application of the binary classifier (a).

(b) Clustering methods, simpleKMeans (k=3) and EM (expectation maximisation) were applied to instances with virus origin and human origin in intron dataset, respectively [25]. We replaced the binary values in the class with cluster memberships (total 6 values for the class) in intron dataset. ClassConditionalProbabilities filter that converted attributes into

class conditional probabilities was applied in preprocess step in prior to the multi-class classifier (b) [25].

Database availability

GuideStar is freely available to academic and non-academic communities. Main DB and an example code (written in Java) to create datasets in case study are provided (www.java.com). It can be downloaded at www.guidestar.kr and www.sourceforge.net.

Case study results and discussion

With ClassBalancer, total weight of each class was 28. The performance estimated with the training with full dataset is shown in Table S1: correct classification divided by the total based on confusion matrix- virus origin and human origin, accuracy (the number of correctly classified instances divided by the total number of instances), Kappa statistic, and Root mean squared error of intron dataset were 24/28, 27.43/28, 91.836 % (90.816 %), 0.836, and 0.285, respectively. The accuracies calculated from the datasets only with the attributes of main DB without scRNA-seq data are indicated inside the parentheses of Table S1. Addition of scRNA-seq data improved the accuracies of the classifier by the average of 6% in four datasets except biological region. It increased by 34%. Accuracy of the classifier (b) that classified instances into 6 different clustered memberships was 100% when estimated with training with full dataset (Data S2). Clustering methods, that subdivided each binary class into 3 classes increased the accuracy of the classification model, which indicated attributes had high intra-class variation in binary class dataset.

Attributes with discriminative power had a non-zero value in classifier (a) (Table S2). It is noteworthy that most of the attributes with non-zero value were identical in the five datasets of 5'UTR, biological region, exon, intron, and 3'UTR where intron dataset had the highest number of non-zero attributes. Ranks of all attributes are included in the Data S3.

Table 1. Frequencies of top 50 attributes from intron dataset with method (a)

	GQ	uORF	IRES	Repeats	Small RNA	scRNA-seq

Frequency (a)	5	0	0	25	6	14

Top 20 selected from intron dataset with method (a) included GQ (18:15:4, 14:11:3, PU1-*), repeats (DNA?, SINE-tRNA-Deu, DNA-TcMar-Mariner, LINE-RTE-X, rRNA, SINE-Alu, DNA?-PiggyBac?, DNA-hAT-Blackjack, DNA-Kolobok, DNA-hAT-Ac, Satellite, DNA-TcMar-Tigger, scRNA, DNA-hAT-Tip100), small RNA (scRNA), and scRNA-seq (infVbystander_Basal, infVbystander_Club). IRES was ranked at 60th, while uORF was ranked almost last. Because the case study was in gene level, rather than transcript level, their ranks were low. In GQ, the format x:y:z was intepreted as “x represents the number of guanine tracts, y represents the number of locations at which a G4 could form, and z represents the number of G4 that could form simultaneously in the sequence” [12]. Instances of virus origin had higher average occurrences of the binary attributes and higher mean values of the numeric attributes in top 20 attributes.

Table 2. Frequencies of top 300 attributes from intron dataset with method (b)

	GQ	uORF	IRES	Repeats	Small RNA	scRNA-seq
Frequency (b)	15	0	0	162	18	105

Top 120 selected from intron dataset with method (b) included GQ (6:3:1), repeats (SINE-tRNA, SINE-MIR, Unknown, LINE-Penelope, LINE-L1-Tx1, DNA-hAT-Tag1, DNA-TcMar?, LTR-ERVL-MaLR, LTR-ERV1?, SINE-Alu, SINE-5S-Deu-L2, srpRNA), small RNA (snoRNA), and scRNA-seq (infVmock_Ciliated, bystanderVmock_Ciliated, bystanderVmock_BC-Club, bystanderVmock_Ionocytes, infVmock_BC-Club, bystanderVmock_Tuft, bystanderVmock_Basal, infVbystander_Ionocytes, infVmock_Goblet, infVmock_Neuroendocrine, bystanderVmock_Club, infVmock_Basal, infVmock_Tuft, infVmock_Ionocytes, infVmock_Club, infVbystander_Goblet, bystanderVmock_Neuroendocrine). Application of Clustering methods and ClassConditionalProbabilities filter decreased the number of attributes from the resources-GQ and small RNA, while they increased that from scRNA-seq data. Clustering methods

transformed binary data where attributes had high within-class variance into more improved data where attributes had high correlation with the class. Ranks of all attributes are included in the Data S4.

One of the biggest barriers to identify the features of TF genes mediated by SARS-COV-2 small ncRNA was the small number of instances, total 7 in this study. With the imbalanced dataset (7 vs. 49 instances), statistical uncertainty was too high either to correctly evaluate the classifiers with cross-validation or to produce statistical confidence in identifying the important features of target genes with SARS-COV-2 small ncRNA interaction. However, with more and more data available in SARS-COV-2 research, identification of key features of human target genes mediated by small RNAs with viral origin may be achieved in near future.

Although omics data availability rapidly increases, the bottleneck of the data usage lies in the lack of bioinformatics software and databases that provide users a convenient means to extract meaningful information from the vast amount of omics data. GuideStar's goal is to provide easy-to-plug-in-to-AI database for coronavirus research, where all the important nucleotide elements and structures are available with a simple query- a transcript ID. With more resources available in future, GuideStar may expand to include additional features such as R-loop and riboswitch that play important roles in viral infection.

References

1. 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.

2. 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.

3. 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.

4. 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.

5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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.

14. 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).

15. 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.

16. Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013-2015 <http://www.repeatmasker.org>.

17. 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.

18. 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.

18. Yao Q, Chen Y, Zhou X. The roles of microRNAs in epigenetic regulation. Curr Opin Chem Biol. 2019 Aug;51:11-17.

19. 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.

20. 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.

21. 3 Characterization of intergenic regions and gene definition. *Nature* (2019). https://doi.org/10.1038/nature28172

22. Duygu M, Demirci S, Adan A. Computational analysis of microRNA-mediated interactions in SARS-CoV-2 infection. PeerJ. 2020; 8: e9369.

23. Chen Y, Wang X. miRDB: an online database for prediction of functional microRNA targets. Nucleic Acids Res. 2020 Jan 8;48(D1):D127-D131.

24. Han H, Cho JW, Lee S, Yun A, Kim H, Bae D, Yang S, Kim CY, Lee M, Kim E, Lee S, KangB, 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).
25. 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.