Development of

varSEAK Online

virSEAK
JSI's SARS-CoV-2 tool

developed by

JSI medical systems GmbH
Tullastr. 18
77955 Ettenheim
GERMANY
phone: +49-7822/440150-21
fax: +49-7822/440150-20
email: support@varSEAK.bio
web: www.jsi-medisys.com

for research use only

2020-06-05

Table of Contents

1 Introduction

The COVID-19 pandemic is currently a high priority issue that is being addressed intensively by a large number of scientists and laboratories worldwide. The enormous scientific interest is currently leading to a rapidly growing amount of sequence data. The databases of NCBI Virus[6] and GISAID[5] contained about 3000 or 18000 fully sequenced virus genomes for download by the middle of May 2020. Like many viruses, SARS-CoV-2 shows an evolutionary development of the viral genome in humans. These genetic variations lead to a differentiation of the original viral genome into different types and allow a grouping of the different SARS-CoV-2 isolates. The grouping of viral genomes could be helpful to understand the evolution and spatial-temporal spread of SARS-CoV-2 as well as to trace infection chains.

One of the early published studies uses a phylogenetic network approach to group virus genomes. Forster *et al.*[3] analyzed 160 SARS-CoV-2 genomes resulting in the suggestion to use the 3 types A, B and C, whereby type A is further divided into two different types. Group A is defined as the root of the network including the closest known bat coronavirus as an outgroup. Group B is divided from A by the two variants 8782C>T and 28144T>C. Group C additionally carries the variant 26144G>T.

I. J. M. Júnior *et al.*[4] have taken a different approach. They clustered the available genomes from NCBI Virus[6] and GISAID[5] into 593 different haplotypes and identified 12 genomic positions with widely shared polymorphisms. They identified 16 SARS-CoV-2 types by clustering the genomes according to the genotypes at these 12 positions. The 6 most common of these types represent about 97,5% of all genomes. Based on this data they proposed to subdivide the global SARS-CoV-2 population into 6 common and 10 less common types.

In addition to the large database for SARS-CoV-2 genomes, GISAID[5] offers a set of downloadable slides on the GISAID EpiCoV™[5] website. On the slide “Full genome tree derived from all outbreak sequences 2020-05-12” the current updates to the phylogenetic tree of the SARS-CoV-2 virus are shown. The virus population is divided into 3 clades. Clade G is based on the marker variant S-D614G (23403A>G), clade S on ORF8-L84S (28144T>C) and clade V on NS3-G251V (26144G>T).

The rapid increase in available sequence data highlights the need to implement a fast and reliable method of grouping the SARS-CoV-2 virus genomes. We developed a system for grouping SARS-CoV-2 genotypes based on sequence comparisons of all SARS-CoV-2 genomes that were available in full length on NCBI Virus[6] (3498 by the middle of May 2020) and in the GISAID[5] EpiCoV™ database (18.298 by the middle of May 2020). The goal of this system is to support a clear and reliable assignment of sequences to representative SARS-CoV-2 genotypes and their global distribution. The spatial-temporal approach in combination with it’s extensibility will also allow to assign possible new types that may develop in the future.

2 Development and results

By aligning the available genomes from NCBI Virus[6] and GISAID[5] against the reference SARS-CoV-2 genome published by Wu *et al.*[1] (NC_045512), we identified polymorphic sites that are unevenly distributed over the virus genome (Figure 1). By analyzing multiple sequence alignments of the available SARS-CoV-2 genomes at 3 different time points, we decided to focus on 12 frequently polymorphic sites to develop a system for the genotyping and clustering of the SARS-CoV-2 population. The use of more positions lead to a more branched tree with more different genotypes. At the same time, fewer isolates could be assigned to these higher number of genotypes. The 12 positions are a good compromise, balancing the complexity of the resulting tree and minimizing the number of isolates that are not assignable to any of these genotypes. The underlying mutations in the SARS-CoV-2 genome are not spread uniformly, they are concentrated on these polymorphic sites. The possible impacts and roles of these SNPs on the pathogenicity and transmission ability of SARS-CoV-2 are not known until now, but 11 of the 12 mutations are located in the coding region and 8 of them are non-synonymous. 1 mutation is located in the genomic leader sequence (5' UTR) of the SARS-CoV-2 genome. This region is an unique characteristic in coronavirus replication and plays a critical role in the gene expression of the virus during its replication. Therefore the highly frequent variants at the 12 positions might have an impact on both the pathogenicity and the transmission ability of SARS-CoV-2. A detailed discussion over the possible effects of the variants for the respective proteins is given in the paper of Changchuan Yin[2].

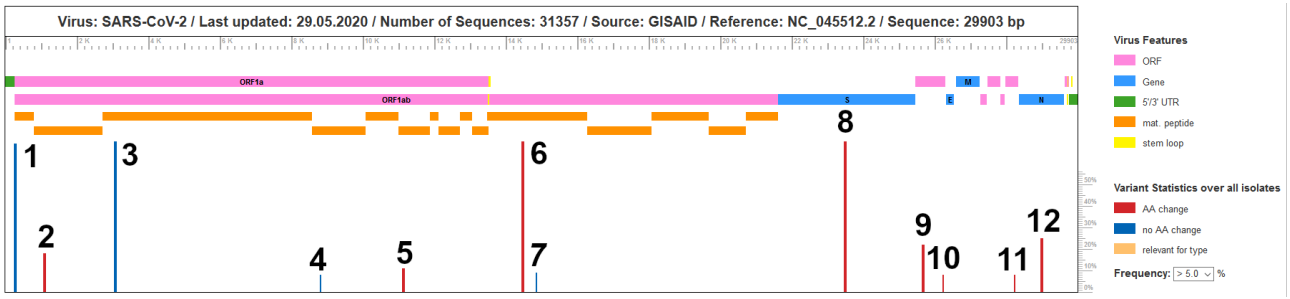

Figure 1: ***Screenshot of virSEAK (www.varSEAK.bio).*** *In the upper part the SARS-CoV-2 genome features are shown. The bars in the histogram represent the frequency of the variants in the analyzed GISAID[5] genomes. The 12 variants that are used to define the SARS-CoV-2 types are numbered like in Table 1.*

Due to an over-representation of genomes sequenced in the USA in the NCBI Virus database[4], we decided to consider the GISAID EpiCoV™ database[5] to define the 12 most frequent polymorphic sites. The isolates included in the GISAID EpiCoV™ database are globally much more evenly distributed. Each of the underlying 12 variants is present in more than 5 % of all analyzed genomes in the GISAID EpiCoV™ database[5] (Table 1). Note: The GISAID[5] data is freely available upon registration, but not licensed to be published elsewhere. All data given in (Table 1) is based on NCBI Virus[4] data. Based on this database, the variant 26144G>T is present at 4.6 %.

#	Position	Nuc. Change	AA Change	# Of Findings NCBI	% Of Genomes NCBI	Gene/Region
6	14408	C>T	Pro4715Leu	2012	60.06	ORF1ab
8	23403*	A>G	Asp614Gly	2010	60.00	S
3	3037	C>T		1988	59.34	ORF1a
1	241	C>T		1942	57.97	5'UTR
9	25563	G>T	Gln57His	1403	41.88	ORF3a
2	1059	C>T	Thr265Ile	1228	36.66	ORF1a
11	28144*	T>C	Leu84Ser	797	23.79	ORF8
4	8782	C>T		781	23.31	ORF1a
5	11083	SNP G>T	Leu3606Phe	232	6.93	ORF1a
12	28881	INDEL GGG>AAC	Arg202_Gly203 delinsLysArg	222	6.63	N
7	14805	C>T		195	5,82	ORF1ab
10	26144*	G>T	Gly251Val	154	4,6	ORF3a

Table 1: ***Polymorphic sites in SARS-CoV-2 genomes from the NCBI Virus[4] database sorted by # of findings.*** *The multiple sequence alignment of 2935 SARS-CoV-2 genomes from NCBI Virus[4] revealed 12 polymorphic sites in SARS-CoV-2 genomes with variants that are present at a high percentage. These polymorphic sites are combined to build different genotype patterns. The GISAID[5] marker variants are indicated by an*. Note: All data given is based on NCBI Virus[4] data, where the variant 26144 G>T (in contrast to the GISAID data) is represented in only 4.6 % of the genomes.*

148 of the 3498 genomes available on the NCBI Virus database[4] were excluded from further analyses because the sequence was not complete. Genomes were considered complete if they were at least 29000 bp long and contained not more than 500 positions with unknown nucleotides (N). Another 415 genomes were discarded due to missing sequence or ambiguity at one or more of the 12 positions. Especially the outermost position 241 is frequently represented by ambiguity codes.
In order to create a clear system that covers as many genomes (isolates) as possible and is open for future developments, we propose to use these 12 positions to build a system for the clustering of the genomes in the SARS-CoV-2 population. The 12 most frequent polymorphic sites (frequency above 5%) are combined to genotype patterns in order of their genomic positions (see Table 2).

Name	Genotype Pattern	# Of Findings	# Of Countries	Parent	Mismatch to Parent	Mismatch to pattern A
A	C|C|C|C|G|C|C|A|G|G|T|G	1903	54		0	0
B	C|C|C|C|T|C|C|A|G|G|T|G	547	43	A	1	1
C	C|C|C|T|G|C|C|A|G|G|C|G	1843	42	A	2	2
D	T|C|T|C|G|T|C|G|G|G|T|G	5104	63	A	4	4
E	C|C|C|T|G|C|T|A|G|G|C|G	207	23	C	1	3
F	C|C|C|C|T|C|T|A|G|T|T|G	1828	39	B	2	3
G	T|C|T|C|G|T|C|G|G|G|T|A	5838	65	D	1	5
H	T|C|T|C|G|T|C|G|T|G|T|G	1094	46	D	1	5
I	T|T|T|C|G|T|C|G|T|G|T|G	4656	49	H	1	6

Table 2: ***The 9 Types of the global SARS-CoV-2 population.*** *All types that represent more than 0.5% of the isolates (based on the GISAID data[5]) are considered. The names of the patterns remain fixed, indepen-*

dent of any new types that may be created. The similarity based parent of each type is given with the number of changes to the parent and to the pattern of type A (reference).

We suggest a designation based on the similarity of each genotype pattern to the genotype pattern of the reference SARS-CoV-2 genome. The differences between the genotype patterns (Table 3) are used to build a pedigree (Figure 2). The *reference SARS-CoV-2 genome is represented by the genotype pattern C|C|C|C|G|C|C|A|G|G|T|G which is named type "A" and defined as root of the pedigree.*

	241	1059	3037	8782	11083	14408	14805	23403	25563	26144	28144	28881
A	C	C	C	C	G	C	C	A	G	G	T	G
B					T							
C				T							C	
D	T		T			T		G				
E				T			T				C	
F					T		T			T		
G	T		T			T		G	T			
H	T		T			T		G				A
I	T	T	T			T		G	T			

*Table 3: **Differences between the 9 SARS-CoV-2 types.** The position of the respective variant is shown in the first row and the suggested naming for the type is given in the first column.*

A threshold that includes all genotype patterns that represent more than 0,5% of the isolates results in the 9 SARS-CoV-2 genotype patterns shown in Table 3. For good traceability each pattern will always keep its name (so for example, C|C|C|C|G|C|C|A|G|G|T|G will always be "A"), even if new types become more frequent and are therefore included in the system. The pedigree based on GISAID[5] data (end of May) is given in figure 2. The naming of the genotype patterns allows a traceable allocation with the required flexibility and maintains clarity if the system has to be expanded.

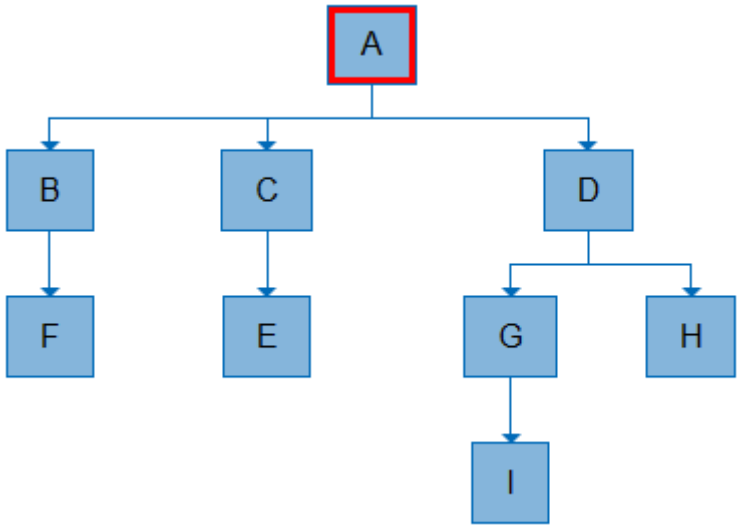

*Figure 2: **Pedigree of the SARS-CoV-2 types.** The pedigree is based on the similarity of the type's genotype patterns compared to the reference SARS-CoV-2 genome published by Wu et al.[1].*

3 References

1 A new coronavirus associated with human respiratory disease in China.
Wu, F., Zhao, S., Yu, B. *et al. Nature* 579, 265–269 (2020).
https://doi.org/10.1038/s41586-020-2008-3

2 Genotyping coronavirus SARS-CoV-2: methods and implications
Changchuan Yin, Genomics, Received 25 March 2020; Received in revised form 13 April 2020
https://doi.org/10.1016/j.ygeno.2020.04.016

3 Phylogenetic network analysis of SARS-CoV-2 genomes
Peter Forster, Lucy Forster, Colin Renfrew, Michael Forster
PNAS April 28, 2020 117 (17) 9241-9243; first published April 8, 2020
https://doi.org/10.1073/pnas.2004999117

4 The global population of SARS-CoV-2 is composed of six major types
Ivair José Morais Júnior, Richard Costa Polveiro, Gabriel Medeiros Souza, Daniel Inserra Bortolin, Flávio Tetsuo Sassaki, Alison Talis Martins Lima
bioRxiv 2020.04.14.040782;
https://doi.org/10.1101/2020.04.14.040782

5 GISAID (accessed 05/03/2020)
https://www.gisaid.org/
Shu, Y., McCauley, J. (2017) GISAID: from vision to reality *EuroSurveillance* 22(13)
doi:10.2807/1560-7917.ES.2017.22.13.30494 PMCID: PMC5388101

6 NCBI Virus (accessed 05/03/2020)
https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=SARS-CoV-2,%20taxid:2697049

4 Acknowledgment

We gratefully acknowledge the following Authors from the Originating laboratories responsible for obtaining the specimens and the Submitting laboratories where genetic sequence data were generated and shared via the **GISAID** Initiative, on which this research is partly based. All submitters of data may be contacted directly via www.gisaid.org.

Albert et al, Li et al, Baumeister et al, Jodar et al, Eden et al, Rockett et al, Draper et al, Sim et al, Bachmann et al, Lam et al, Gray et al, Timms et al, Gall et al, Arnott et al, Sadsad et al, Carter et al, Rahman et al, Holmes et al, OSullivan et al, Sintchenko et al, Chen et al, Maddocks et al, Kok et al, Dwyer et al, Huang et al, Meumann et al, Pyke et al, Chamings et al, Caly et al, Seemann et al, Sikazwe et al, Popa et al, Kashofer et al, Saha et al, Tan et al, Fan et al, Corman et al, Joan Marti-Carreras et al, Vanmechelen et al, Joan Marti-Carerras et al, Selhorst et al, Wawina et al, Lambrechts et al, Vereecke et al, Durkin et al, Keith et al, Resende et al, Nascimento et al, Adelino et al, Melo et al, Romero et al, Jaqueline Goes de Jesus et al, Claudio Tavares Sacchi et al, Mak Tze Minn et al, Karlsson et al, Smith et al, Harrigan et al, Majer et al, Eshaghi et al, Tyson et al, Castillo et al, Medina et al, Shi et al, Nie et al, Linlin et al, Sheng et al, Marcela Mercado-Reyes et al, Duarte et al, Rokic et al, Kramna et al, Nagy et al, Broz et al, Kirkegaard et al, Placide Mbala-Kingebeni et al, Rasmussen et al, Marquez et al, Juan Jose Guadalupe et al, Belen Prado-Vivar et al, Mohamed Ahmed Ali et al, Galiano et al, Platt et al, Meredith et al, Clark et al, Thushan de Silva et al, Smura et al, Kang et al, Bal et al, Djillali et al, V. Caro et al, Terrier et al, Croville et al, Wei et al, Qi et al, Ge et al, Sesay et al et al, Kotaria et al, Murtskhvaladze et al, Chanturia et al, Machablishvili et al, Kotorashvili et al, Muenchhoff et al, Walter et al, Adams et al, Ngoi et al, Bampali et al, Pogka et al, Kassela et al, Lu et al, et al, Zhao et al et al, Hua et al, Yu hua et al, Yao et al, Wang et al, Lau et al, Kenneth Siu-Sing LEUNG et al, Chong-Yee YAU et al, Chun Hang et al, Chan et al, To et al, Alan K.L. Tsang et al, Dominic N.C. Tsang et al, Urban et al, Gudbjartsson et al, Potdar et al, Yadav et al, Pragya Yadav. Savita Patil et al, Pandit et al, Kanani et al, Verma et al, Savaliya et al, Kumar et al, Saiyed et al, Kinariwala et al, Patel et al, Aring et al, Khandelwal et al, Vaghela et al, Trivedi et al, Barve et al, Modi et al, Joshi et al, Shrimali et al, Sood et al, Shah et al, Hinsu et al, Sabara et al, Puvar et al, Raval et al, Gandhi et al, Pandya et al, Nagamani et al, Putty et al, Radhakrishna et al, Raja Rao Mesipogu et al, Thrilok Chander et al, Pramod Kumar# et al, Maitra et al, Pattabiraman et al, Rahardjo et al, Shimizu et al, Johar et al, Zeinali et al, Khosravi et al, Carr et al, Dempsey et al, Zuckerman et al, Inbar Cohen-Gihon et al, Lorusso et al, Stefanelli et al, Licastro et al, Capobianchi et al, Lalle et al, Rueca et al, Cesare M. Gruber et al, Messina et al, Bartolini et al, Antonino Di Caro et al, Castilletti et al, Carletti et al, Zehender et al, R.A Diotti et al, Bagnarelli et al, Milani et al, Sekizuka et al, Kosaki et al, Hishiki et al, Imai et al, Kumagai et al, Braun et al, Cui et al, Zhao et al, Fang et al, Issa Abu-Dayyeh et al, Shevtsov et al, Queen et al, Kim et al, Fahd Al-Mulla et al, Silamikelis et al, Xiaoguang et al, Anke Wienecke-Baldacchino et al, Nieuwenhuijse et al, Yoong Min CHONG et al, Suppiah.J et al, Mohd Noor Mat Isa et al, Guillermo Ruiz-Palacios et al, Joel Armando Vazquez Perez et al, Ernesto et al, Adnan Araiza Rodriguez et al, Jose Ernesto Ramirez Gonzalez et al, Munoz-Medina et al, Irma Lopez Martinez et al, Fabiola Garces Ayala et al, Gisela Barrera Badillo et al, Pilailuk et al, Li jian Xiong et al, Sah et al, Bas Oude Munnink et al, Storey et al, Wellington SCL et al, M.E. Quinones-Mateu et al, Oluniyi et al, Curran et al, Kathrine Stene-Johansen et al, Javed et al, Zainab et al, Franco et al, Carlos Padilla Rojas et al, Lapid et al, Medado et al, Robakowska et al, Milewska et al, Branicki et al, Rabalski et al, Herud et al, Kujawa et al, Guiomar et al, Costa et al, Guiomar et al et al, Santiago et al, Abdul-

latif Al-Khal et al, Pyankov et al, Bodnev et al, Kozlovskaya et al, A. Pavlenko et al, Shchetinin et al, Komissarov et al, Alghoribi et al, Hala et al, Mfarrej et al, Albarrag et al, Thomson et al, Smollett et al, Ana da Silva Filipe et al, McHugh et al, Dia et al, Dejan Vidanovic. Bojana Tesovic et al, Jiang et al, Zhang et al, Yang et al, Mak et al, Octavia et al, Anderson et al, Slavikova et al, Zakotnik et al, Tomaz Mark Zorec et al, Mahnic et al, Giandhari et al, Allam et al, Park et al, Iglesias-Caballero et al, Armengol et al, Dahdouh et al, Gonzalez et al, Viedma et al, Recio et al, Andres et al, Navarro et al, Maria Alma Bracho et al, Maria Dolores Ocete et al, Griselda De Marco et al, Beamud et al, Lidia Ruiz Roldan et al, Marta Pla Diaz et al, Neris Garcia-Gonzalez et al, Loreto Ferrus Abad et al, Giuseppe D'Auria et al, Juan Alberola Enguidanos et al, Inma Galan Vendrell et al, Carbo et al, Paula Ruiz-Hueso et al, Mariana Reyes-Prieto et al, Vicente Soriano Chirona et al, Ansari et al, Gimeno et al, Llucia Martinez-Priego et al, Mendoza et al, Jeewandara et al, Bengner et al, Yun et al, Nilsson et al, Hammar et al, Janers et al, Lindback et al, Henriksson et al, Kerstin Persson Moberg et al, Ersson et al, Espmark et al, Holmqvist et al, Jarbur et al, Mia Settergren Hammer et al, Dagner et al, Carlson et al, Olsson et al, Sengpiel et al, Embring et al, Eiback et al, Klanger et al, Reimer et al, Hjerten et al, Kotz et al, Kollberg et al, Svartstrom et al, Oskar Karlsson Lindsjo et al, Sundqvist et al, FOI Bioinformatics team et al, FOI bioinformatics team et al, Schmutz et al, Hirsch et al, LAUBSCHER Florian et al. et al, Laubscher et al, Tsao et al, Yeh et al, Perng et al, """No. 325 et al", Batty et al, Rodpan et al, Bayrakdar et al, Shaikh Terkis Islam Pavel et al, Karacan et al, Fatma Nilay Tutak et al, Maqsood et al, Bowers et al, Larsen et al, Tao et al, Deng et al, CZB Cliahub Consortium et al, Arevalo et al, Zeller et al, SEARCH Alliance San Diego with Christina Clarke et al, SEARCH Alliance San Diego et al, SEARCH Alliance San Diego with Michael Quigley et al, Uehara et al, Maria Aguero-Rosenfeld et al, Roychoudhury et al, Fauver et al, Paden et al, Schmedes et al, Elbadry et al, Holzer et al, Feehan et al, Kamil et al, Scott et al, Vanchiere et al, Smither et al, Thielen et al, Blankenship et al, Plumb et al, Nemudryi et al, Bailey et al, UNMC COVID-19 Response Team et al, Ana Gonzalez-Reiche et al, Black et al, Butler et al, Kirsten St. George et al, Tu et al, Brendan O'Connell et al, Snyder et al, S. Wesley Long et al, Young et al, DCLS et al, Division of Consolidated Laboratory Services et al, Division of Consolidated Laboratories et al, Virginia DCLS et al, Virginia Division of Consolidated Laboratories et al, Chu et al et al, Chu etl al et al, Chu et al, Katarina Braun and Gage Moreno et al, Moreno et al, Gage Moreno and Katarina Braun et al, Craig Richmond & Paraic Kenny et al, Richmond et al, Domman et al, Ahmad Abou Tayoun et al, Salazar et al, Cao et al, Nguyen et al, Le Quynh Mai et al, Nguyen Thi Tam et al, Ung Thi Hong Trang et al, Moore et al, Ren et al, Zhou et al, Sun et al, Zhu et al, Thomas S.H. Chung et al, Sophie Le Poder et al, Rimoldi et al, Shankar et al, Liu et al, eta Zuckerman et al, Oreshkova et al, Shen et al, Mitchell et al.

The complete list, compliant to the “GISAID EPIFLU™ DATABASE ACCESS AGREEMENT”, is available here.