Compressive Pangenomics using PanMANs

Prof. Yatish Turakhia

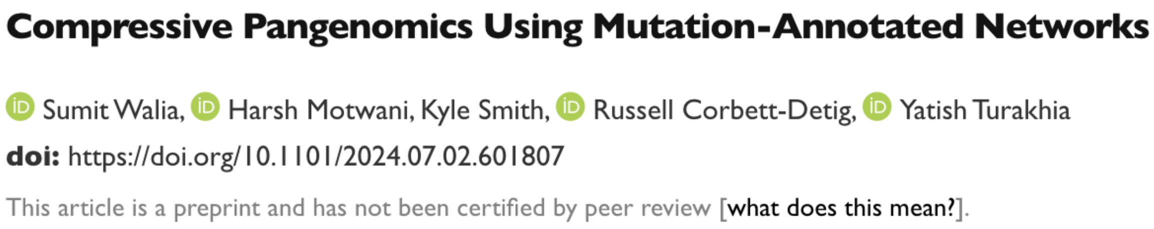
Department of ECE, UCSD

Codebase and Preprint

PANMAN

Pangenome Mutation-Annotated Network

https://github.com/TurakhiaLab/panman



Compressive Pangenomics Using Mutation-Annotated Networks

Sumit Walia, Harsh Motwani, Kyle Smith, Russell Corbett-Detig, Yatish Turakhia

doi: https://doi.org/10.1101/2024.07.02.601807

This article is a preprint and has not been certified by peer review [what does this mean?].

https://www.biorxiv.org/content/10.1101/2024.07.02.601807v1

Sumit Walia

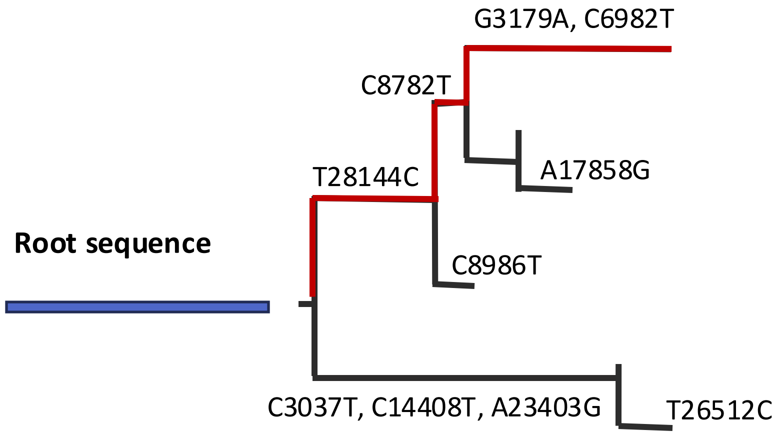
Harsh Motwani

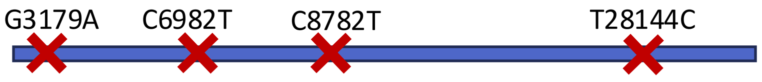
MAT: The data structure powering UShER

- **MAT**: mutation-annotated tree

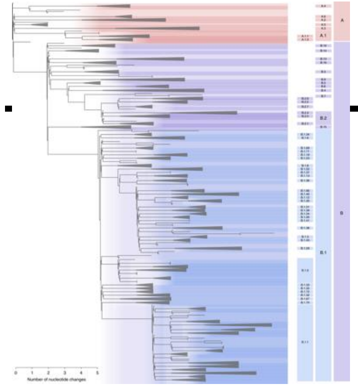
- Stores:
  - **Tree topology** corresponding to the inferred phylogeny
  - A single **root sequence** (could be the reference genome)
  - **Mutations** inferred on each branch

- **Property:** sequence corresponding to every tip or internal node of the tree can be derived from the root sequence and the mutations on its path to the root



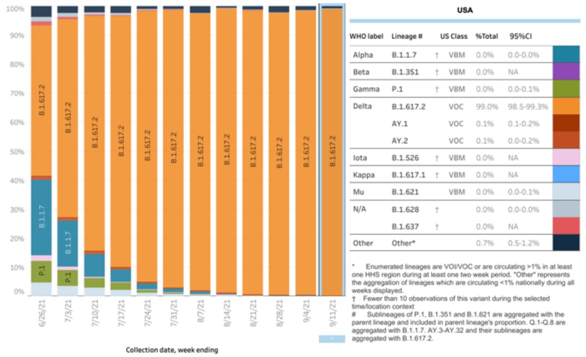


Naming lineages



(Rambaut et al., Nat. Microbiol. 2020)

Monitoring circulating lineages



USA

WHO label	Lineage #		US Class	%Total	95%CI	
Alpha	B.1.1.7	†	VBM	0.0%	0.0-0.0%	
Beta	B.1.351	†	VBM	0.0%	NA	
Gamma	P.1	†	VBM	0.0%	0.0-0.1%	
Delta	B.1.617.2		VOC	99.0%	98.5-99.3%	
	AY.1		VOC	0.1%	0.1-0.2%	
	AY.2		VOC	0.1%	0.0-0.2%	
Iota	B.1.526	†	VBM	0.0%	NA	
Kappa	B.1.617.1	†	VBM	0.0%	NA	
Mu	B.1.621		VBM	0.0%	0.0-0.1%	
N/A	B.1.628	†		0.0%	0.0-0.0%	
	B.1.637	†		0.0%	NA	
Other	Other*			0.7%	0.5-1.2%	

* Enumerated lineages are VOI/VOC or are circulating >1% in at least one HHS region during at least one two week period. "Other" represents the aggregation of lineages which are circulating <1% nationally during all weeks displayed.

† Fewer than 10 observations of this variant during the selected time/location context

# Sublineages of P.1, B.1.351 and B.1.621 are aggregated with the parent lineage and included in parent lineage's proportion. Q.1-Q.8 are aggregated with B.1.1.7. AY.3-AY.32 and their sublineages are aggregated with B.1.617.2.

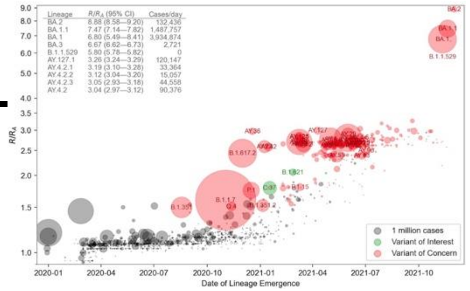
(CDC.gov dashboard 2021)

Identify newly-introduced strains

USA/CA-CDC-FG-069350/2021
USA/CA-CDC-FG-064300/2021
USA/CA-CDC-FG-081289/2021
USA/TX-CDC-FG-094229/2021
USA/TX-CDC-FG-105045/2021
USA/CA-CDC-FG-077556/2021
Ottawa

(McBroome et al., Virus Evol. 2022)

Predicting fitness of a new strain

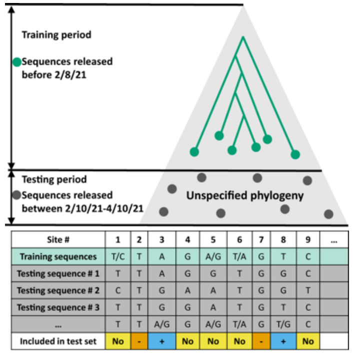


Lineage	R/R_A (95% CI)	Cases/day
BA.2	8.88 (8.58—9.20)	132,436
BA.1.1	7.47 (7.14—7.82)	1,487,757
BA.1	6.80 (5.49—8.41)	3,934,874
BA.3	6.67 (6.62—6.73)	2,721
B.1.1.529	5.80 (5.78—5.82)	0
AY.127.1	3.26 (3.24—3.29)	120,147
AY.4.2.1	3.19 (3.10—3.28)	33,364
AY.4.2.2	3.12 (3.04—3.20)	15,057
AY.4.2.3	3.05 (2.93—3.18)	44,558
AY.4.2	3.04 (2.97—3.12)	90,376

(Obermeyer et al., Science 2022)

UShER

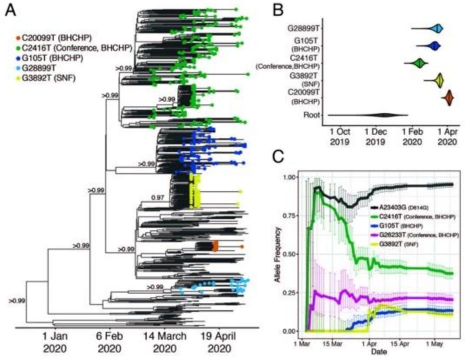
Predicting the next mutation


Training period
Sequences released before 2/8/21
Testing period
Sequences released between 2/10/21-4/10/21
Unspecified phylogeny

Site #	1	2	3	4	5	6	7	8	9	...
Training sequences	T/C	T	A	G	A/G	T/A	G	T	C	
Testing sequence # 1	T	T	A	G	G	T	G	G	C	
Testing sequence # 2	C	T	G	A	A	T	G	G	T	
Testing sequence # 3	T	T	G	G	A	T	G	T	C	
...	T	T	A/G	G	A/G	T/A	G	T/G	C	
Included in test set	No	-	+	No	No	No	-	+	No	

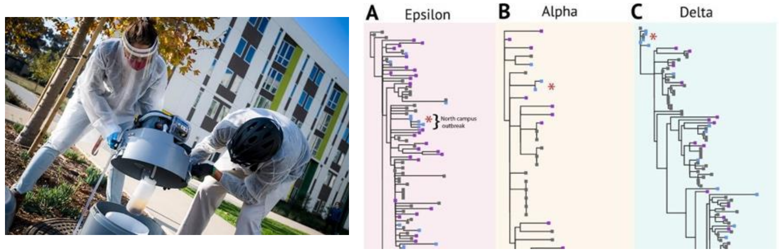
(Hallak et al., Nat. Comm Biol. 2022)

Analyze outbreaks and superspreader events



(Lemieux et al., Science 2021)

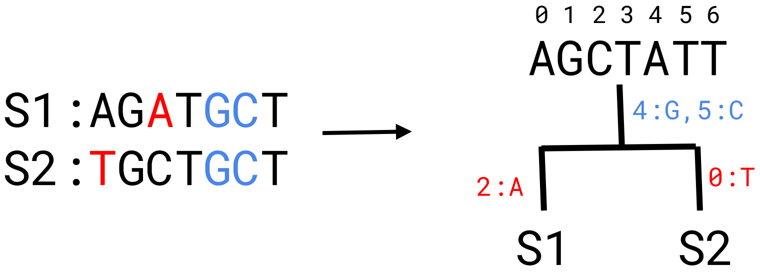
Wastewater surveillance

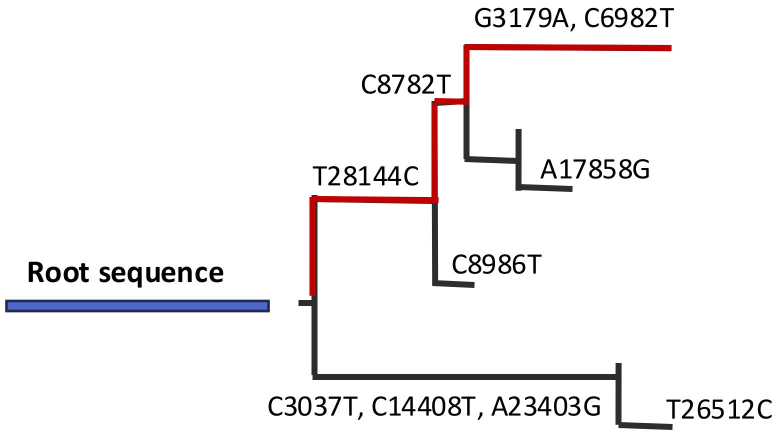


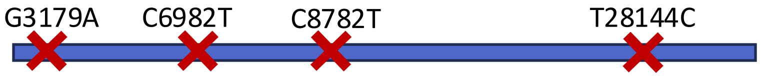
(Karthikeyan et al., Nature 2022)

Limitations of UShER-MAT

- Reference-based



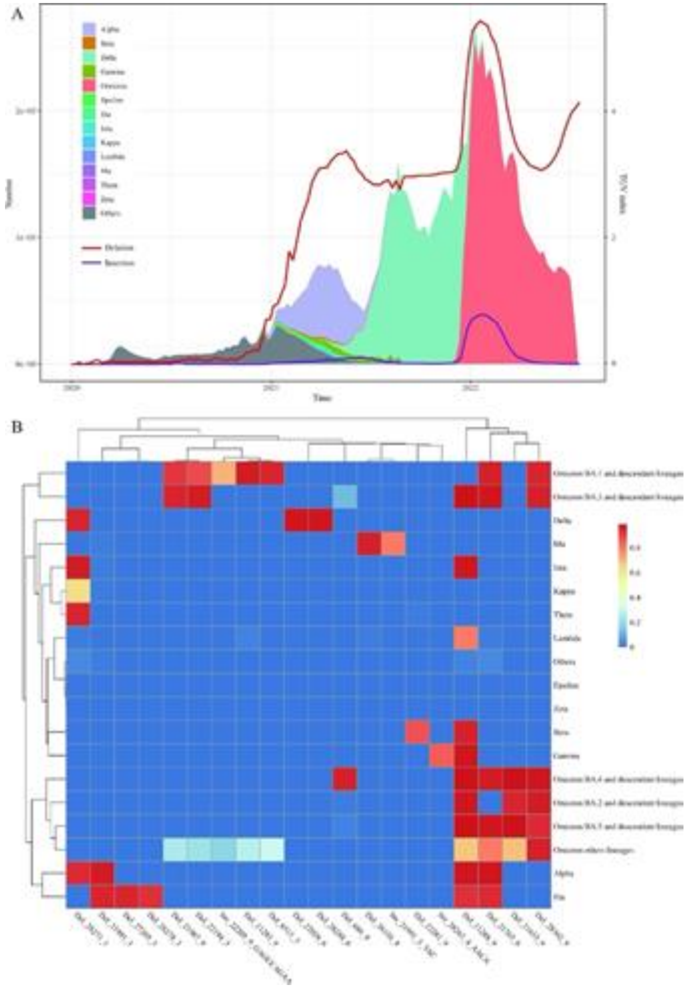




Limitations of UShER-MAT

- **Reference-based**

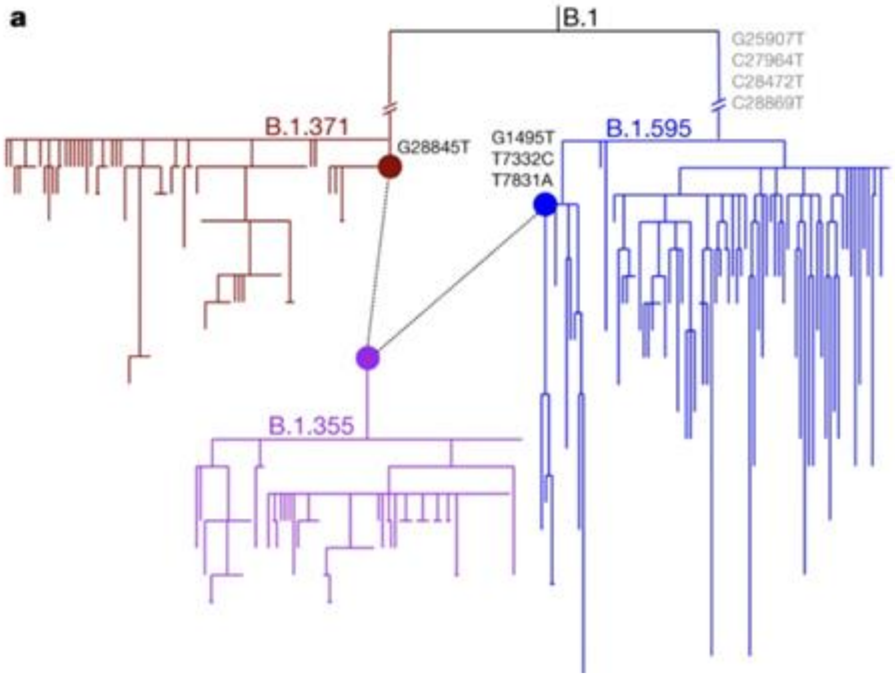
- Only stores substitutions – **ignores indels**
  - Indels sometimes comprise lineage-defining mutations



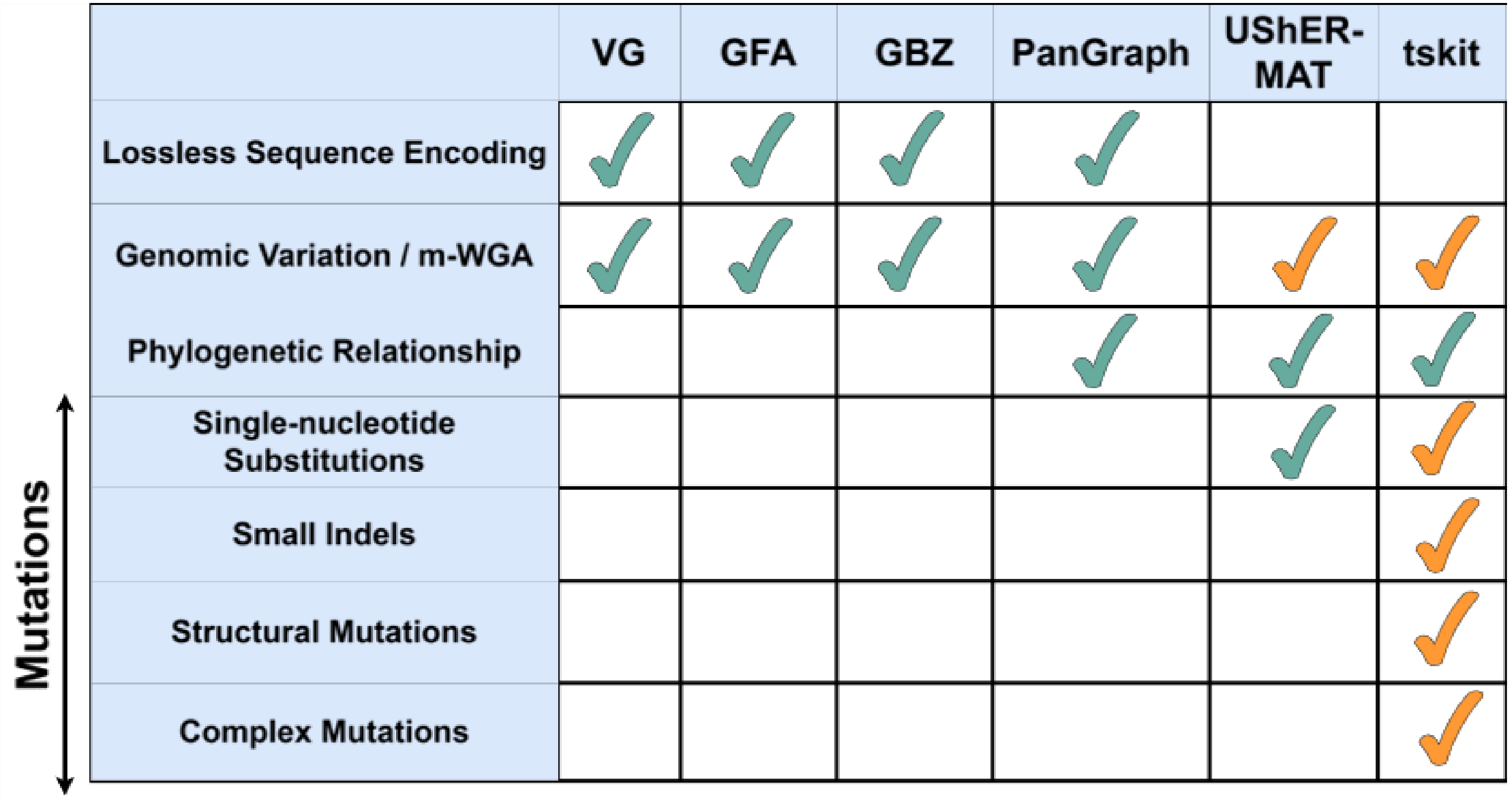
(Li et al., Microbiology Spectrum 2023)

Limitations of UShER-MAT

- **Reference-based**
- Only stores substitutions – **ignores indels**
  - Indels sometimes comprise lineage-defining mutations
- Restricted to a single **tree topology** – cannot represent complex mutations (e.g., recombination or horizontal gene transfer) violating the vertical mode of evolution

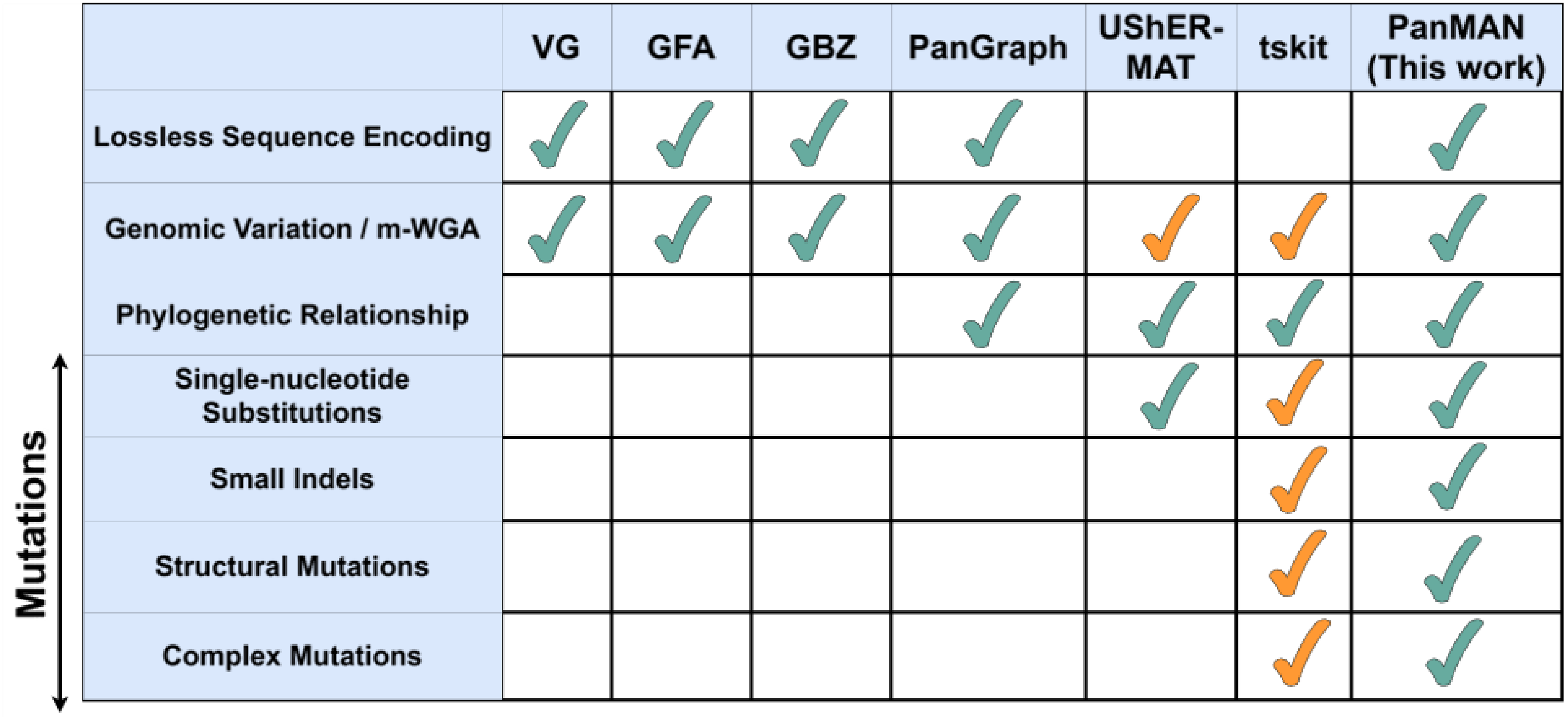


Summary of features in Pangenome formats



		VG	GFA	GBZ	PanGraph	UShER-MAT	tskit
	Lossless Sequence Encoding	✓	✓	✓	✓		
	Genomic Variation / m-WGA	✓	✓	✓	✓	✓	✓
	Phylogenetic Relationship				✓	✓	✓
Mutations	Single-nucleotide Substitutions					✓	✓
	Small Indels						✓
	Structural Mutations						✓
	Complex Mutations						✓

Summary of features in Pangenome formats

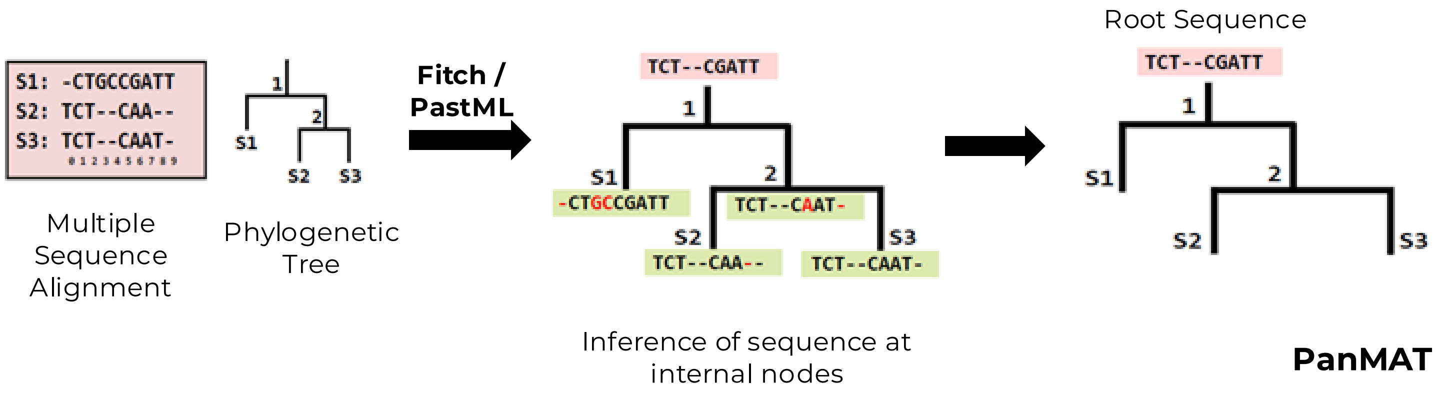


		VG	GFA	GBZ	PanGraph	UShER-MAT	tskit	PanMAN (This work)
	Lossless Sequence Encoding	✓	✓	✓	✓			✓
	Genomic Variation / m-WGA	✓	✓	✓	✓	✓	✓	✓
	Phylogenetic Relationship				✓	✓	✓	✓
Mutations	Single-nucleotide Substitutions					✓	✓	✓
	Small Indels						✓	✓
	Structural Mutations						✓	✓
	Complex Mutations						✓	✓

Inferred MSA, Phylogeny, and **mutations** all in **one format!**
PanMAN is not just **information-rich** but also more **compact** and **scalable**

PanMAT: Pangenome Mutation-Annotated Tree

- Incorporating insertions and deletions (indels) into a MAT
  - MSA defines the coordinate system
  - Gaps treated as special characters



```
S1: -CTGCCGATT
S2: TCT--CAA--
S3: TCT--CAAT-
    0123456789
```

Multiple Sequence Alignment

Phylogenetic Tree

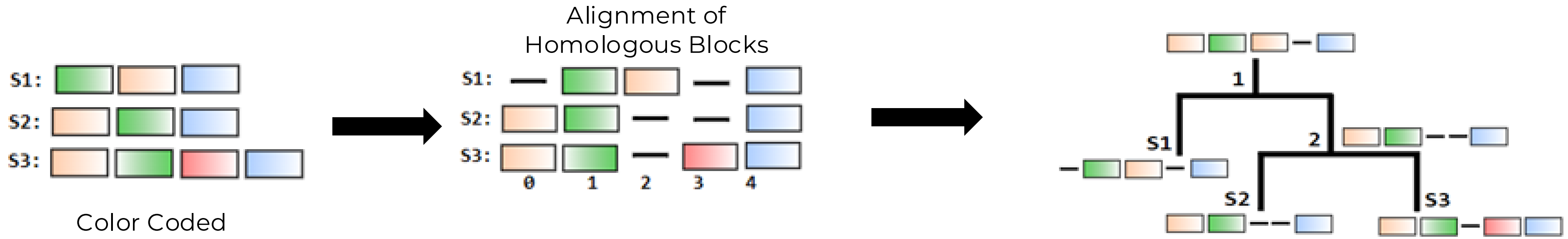
Inference of sequence at internal nodes

PanMAT

Nucleotide Mutation Table		
Node	Pos	Allele
S1	[0,3,4]	[-,G,C]
2	[6,9]	[A,-]
S2	[8]	[-]

PanMAT: Pangenome Mutation-Annotated Tree

- Incorporating **structural changes** and **rearrangements**
  - Identify homologous blocks
  - MSA of homologous blocks
  - Block mutations are like substitutions to or from gaps



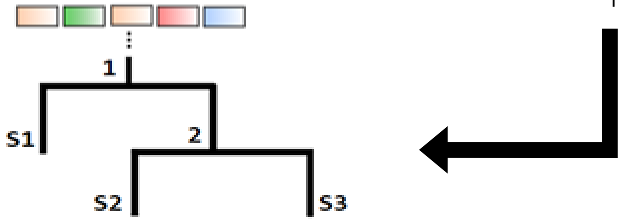
Alignment of Homologous Blocks

Color Coded Homologous Blocks

Inference of blocks at internal nodes’

Block Mutation Table		
Node	Pos	Mutation
1	[3]	[DEL]
S1	[0]	[DEL]
2	[2]	[DEL]
S3	[3]	[INS]

(Pseudo-Root)



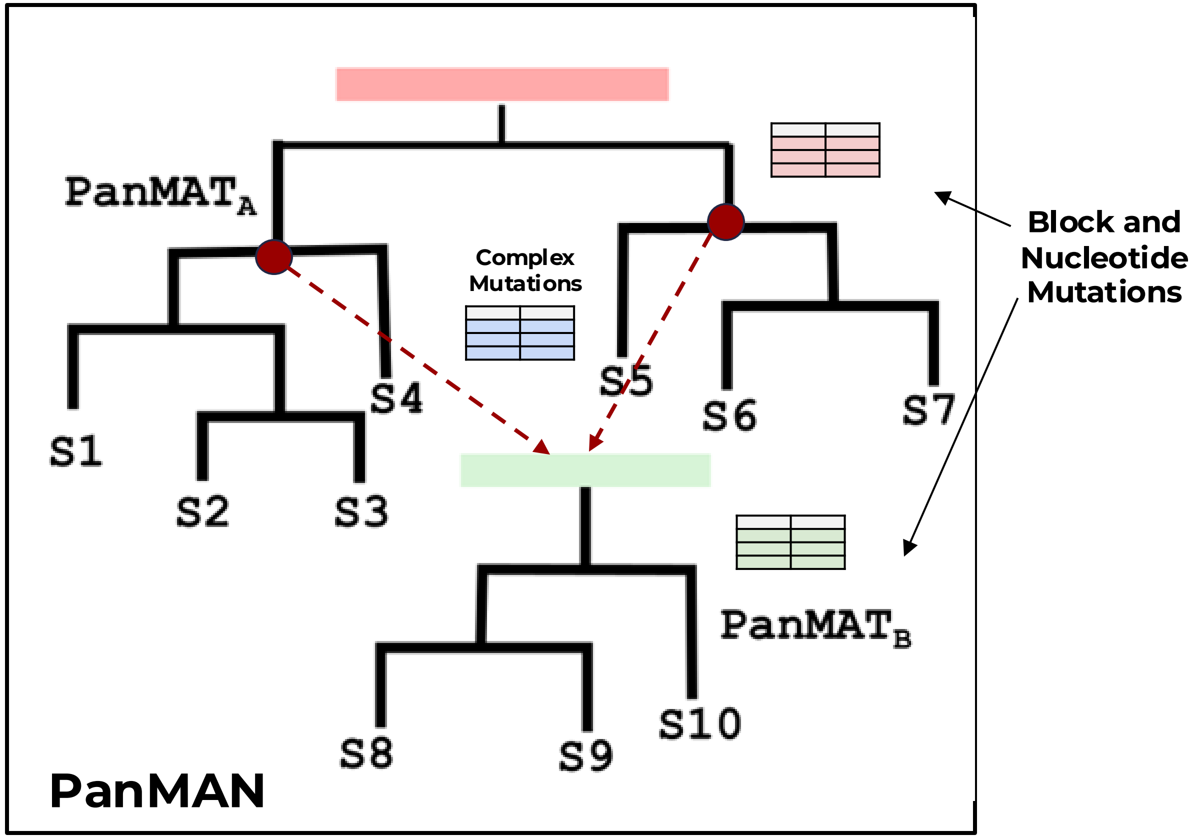
PanMAT

PanMAN: Pangenome Mutation-Annotated Network

- **PanMAN:** Generalization of PanMAT to represent **complex mutations**

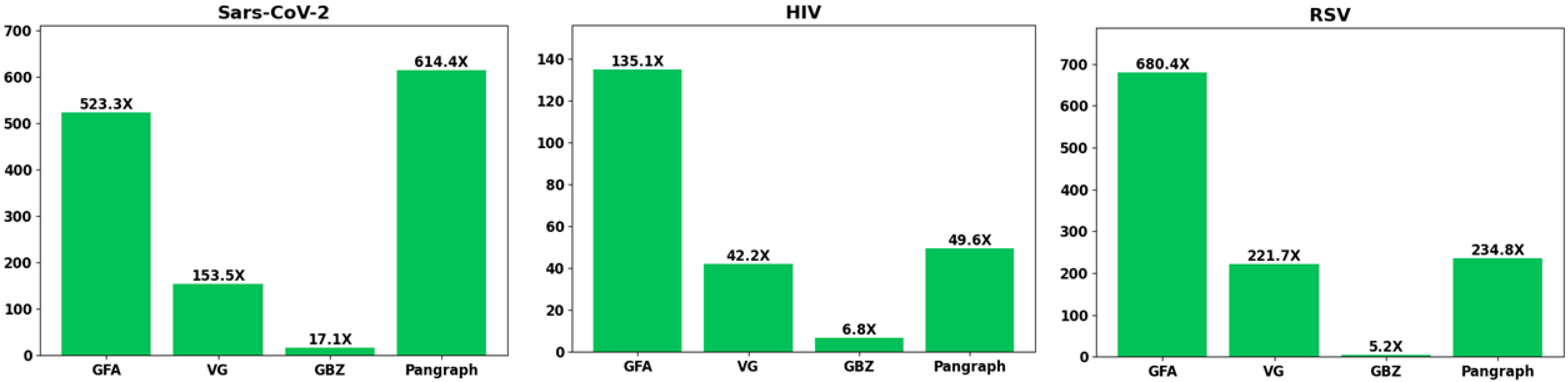
- One or more PanMATs are connected with network edges (red dotted lines)

- Network Edges stores complex mutations (blue table), i.e., Horizontal Gene Transfer (HGT) and Recombination



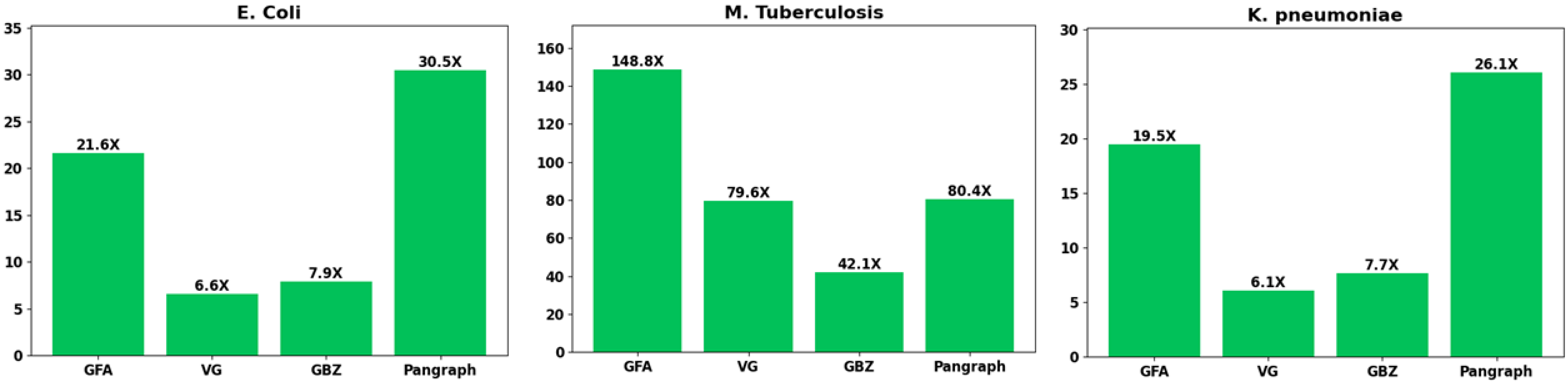
PanMAN is the most compressive pangenomic format

Compression achieved by PanMAN compared to other formats

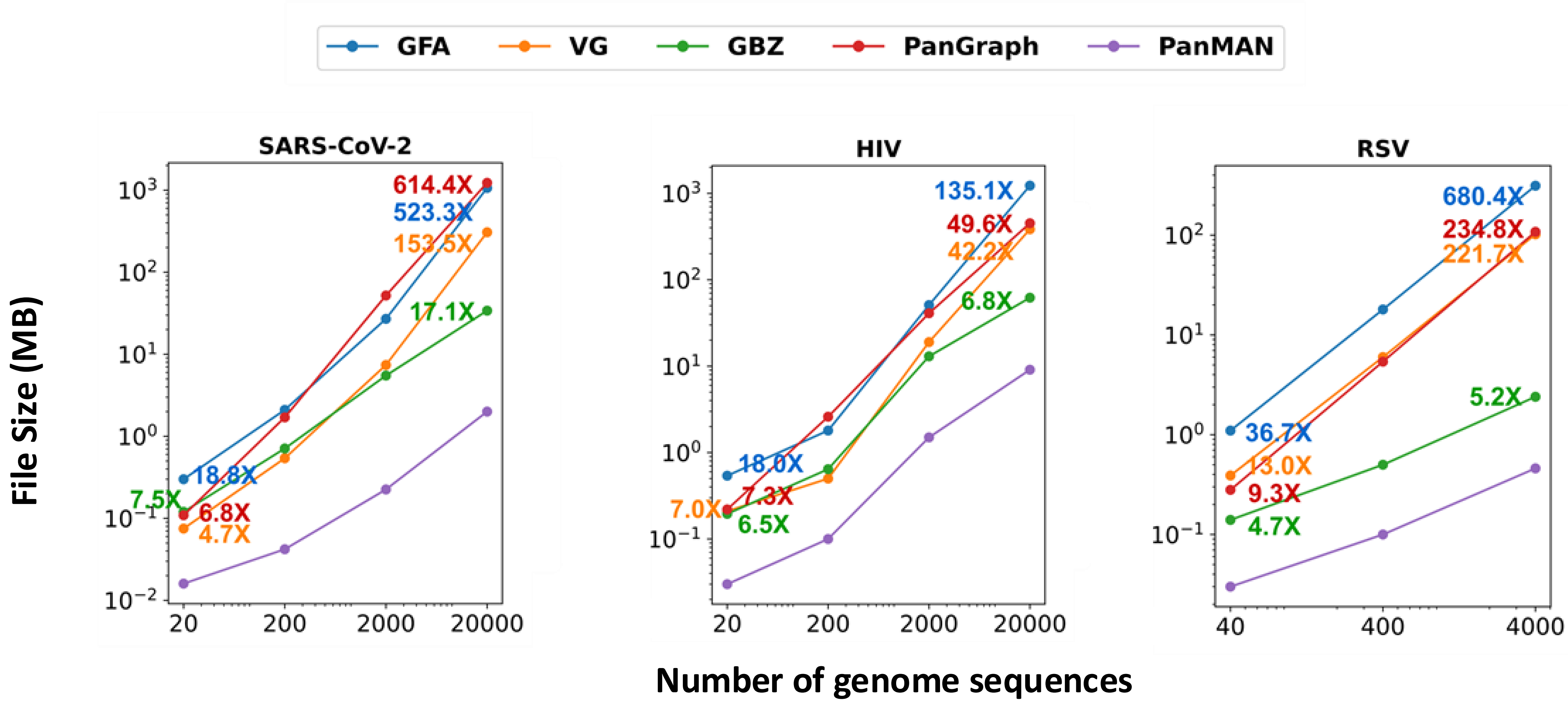


PanMAN is the most compressive pangenomic format

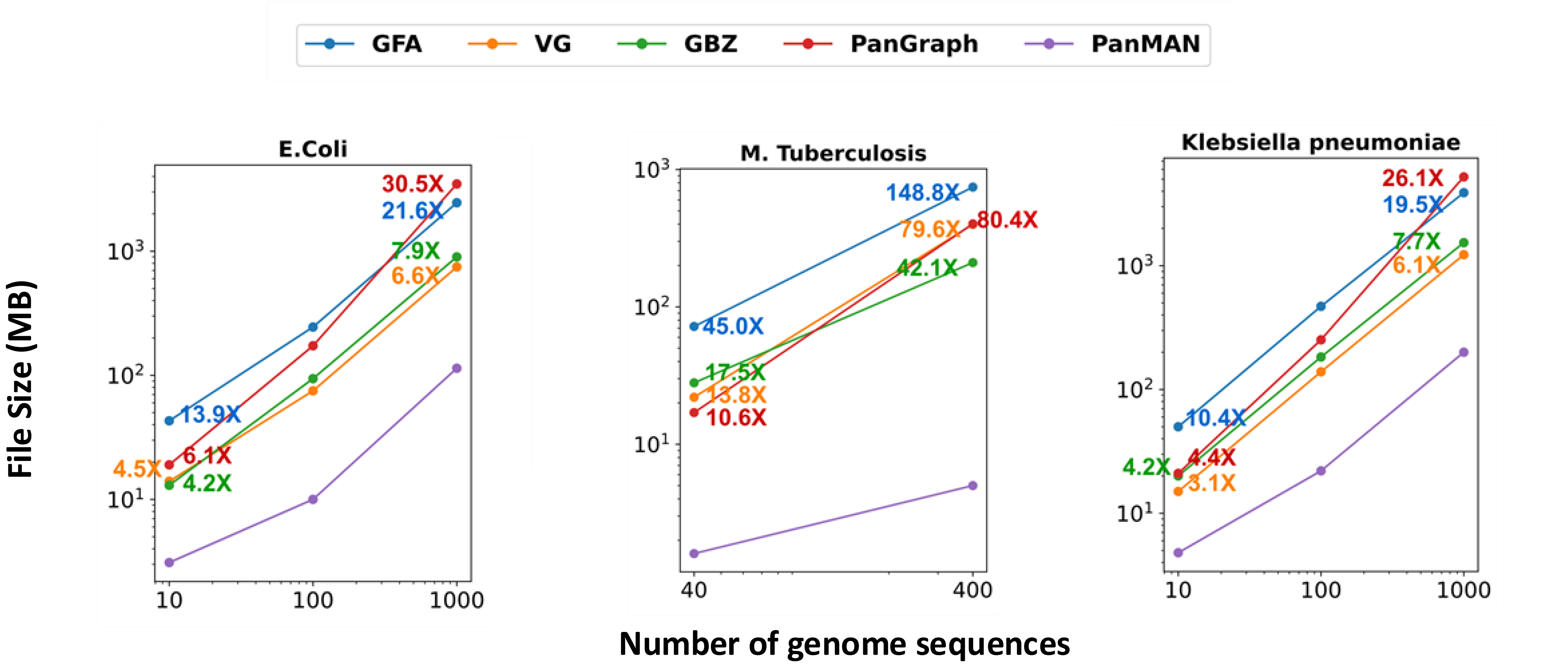
Compression achieved by PanMAN compared to other formats



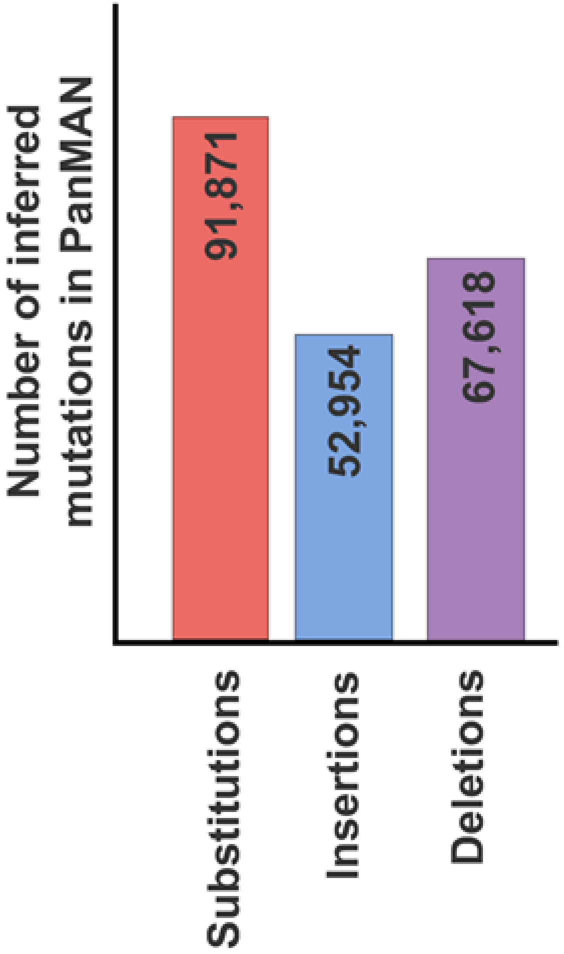
PanMAN scales well relative to other formats

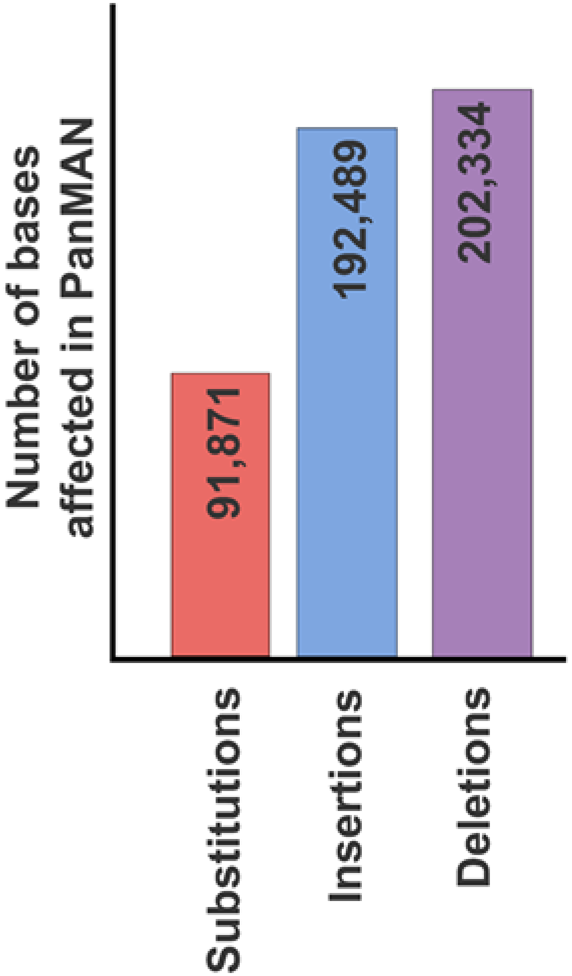


PanMAN scales well relative to other formats

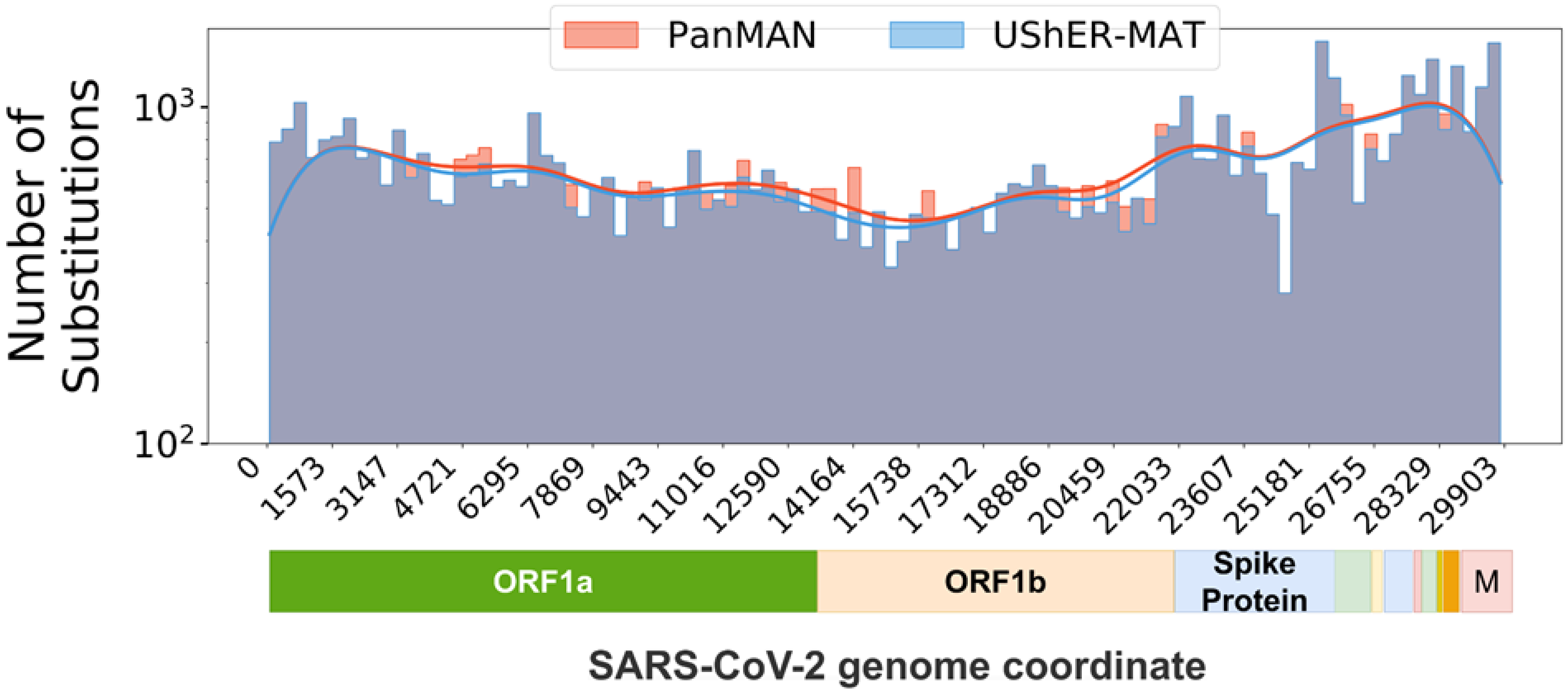


Exploration of SARS-CoV-2 mutational and evolutionary landscape using PanMAN

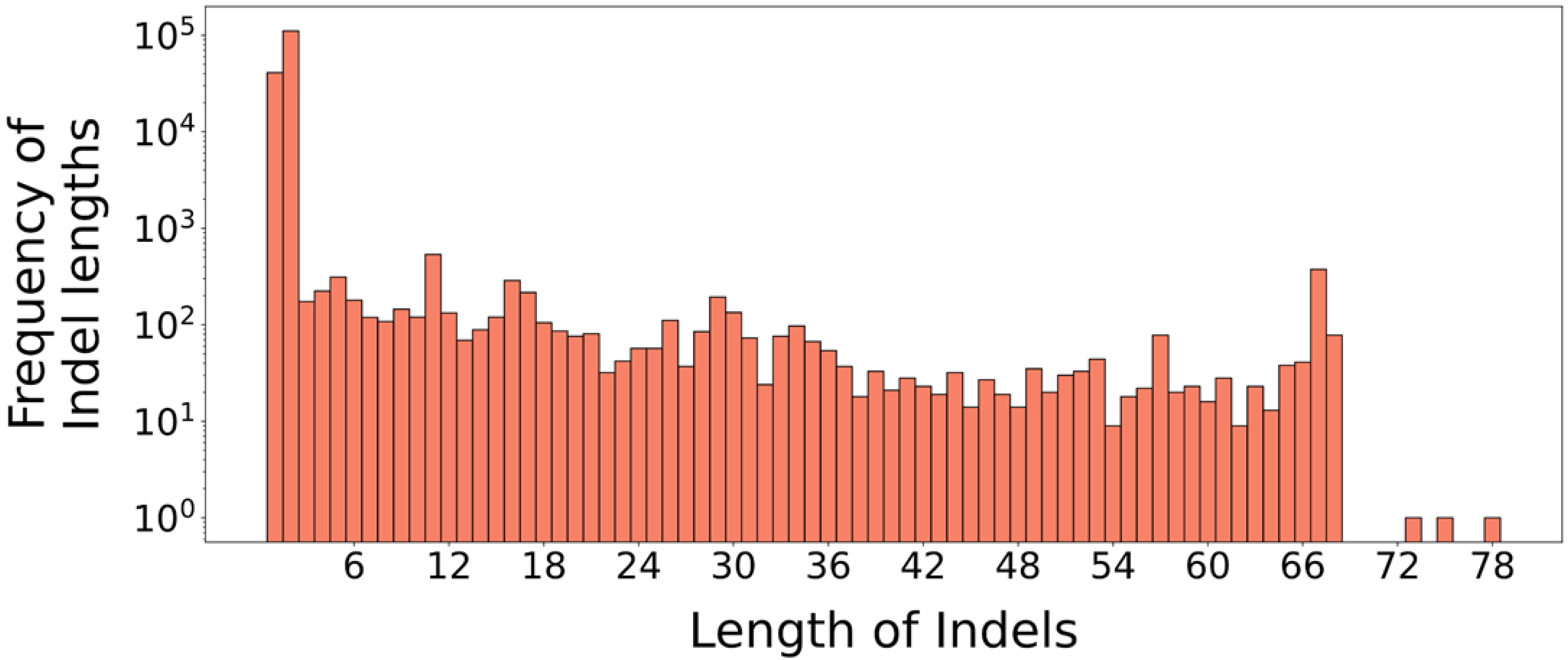




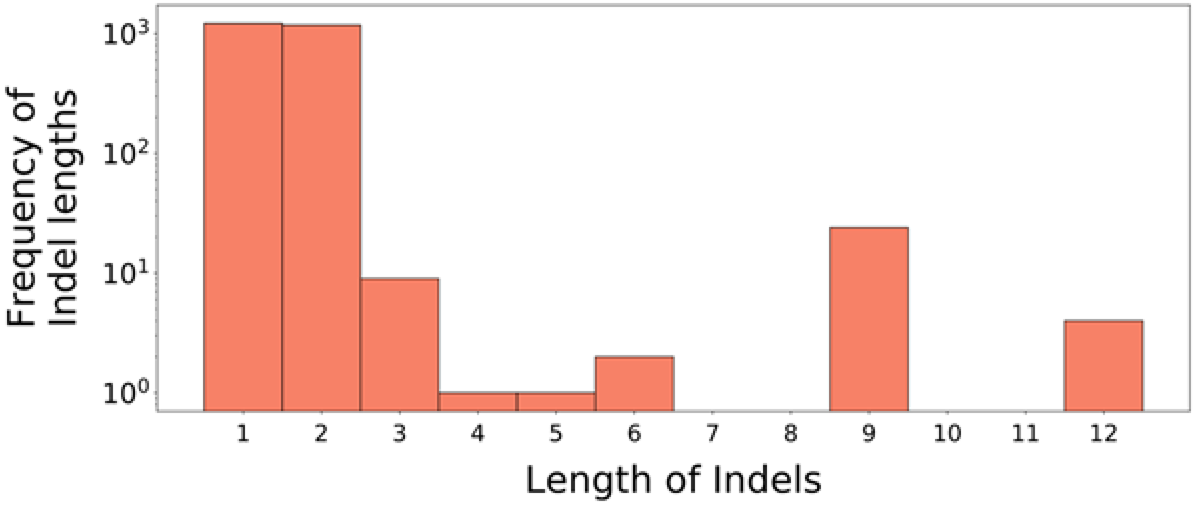
Exploration of SARS-CoV-2 mutational and evolutionary landscape using PanMAN

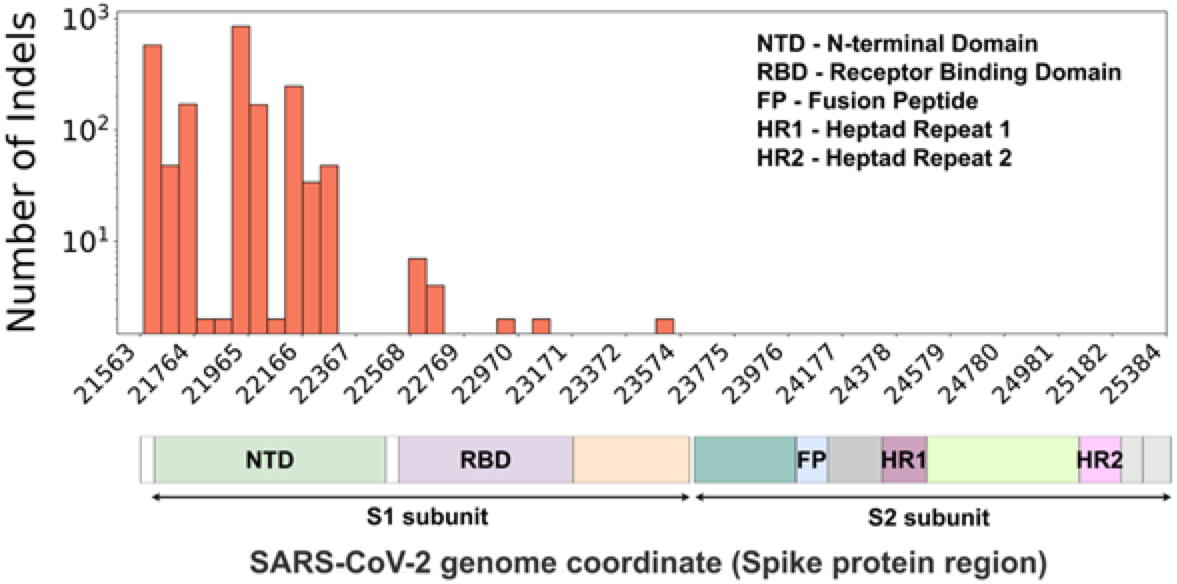


Exploration of SARS-CoV-2 mutational and evolutionary landscape using PanMAN

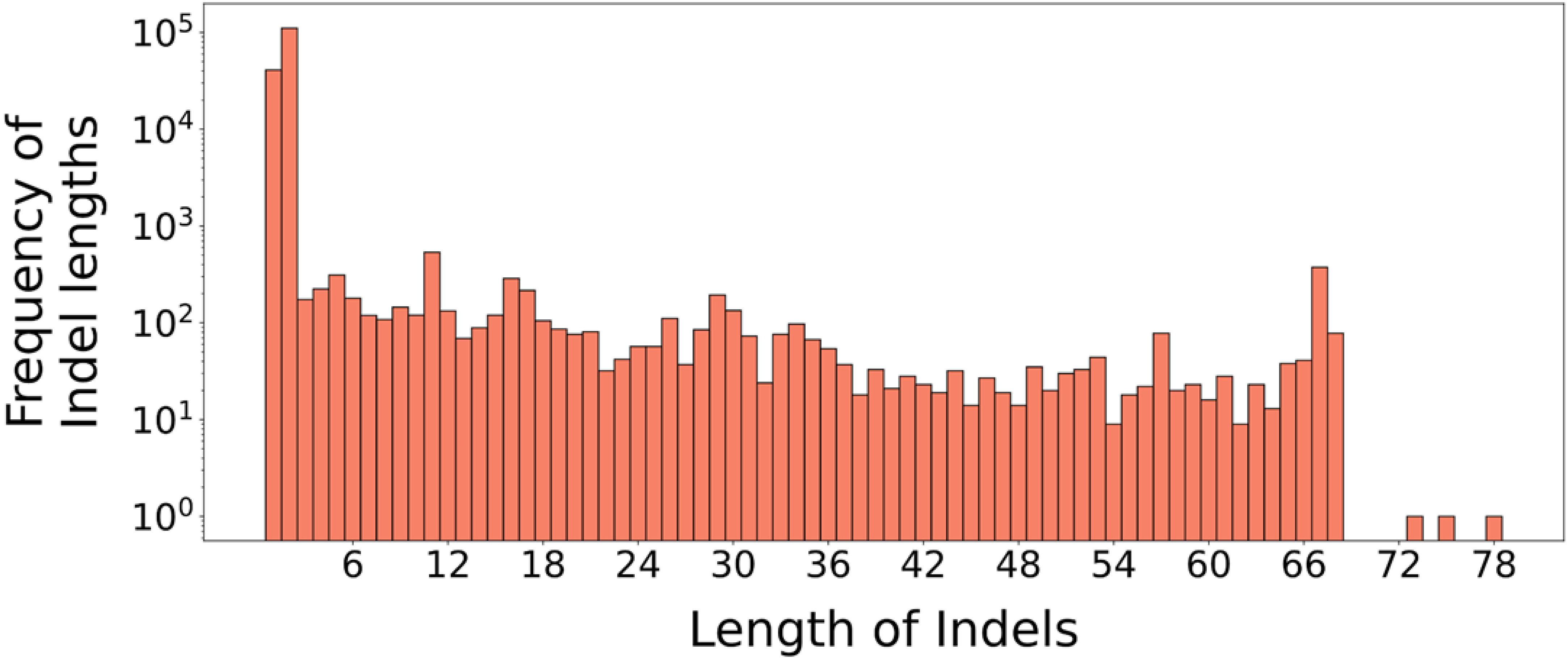


Exploration of SARS-CoV-2 mutational and evolutionary landscape using PanMAN

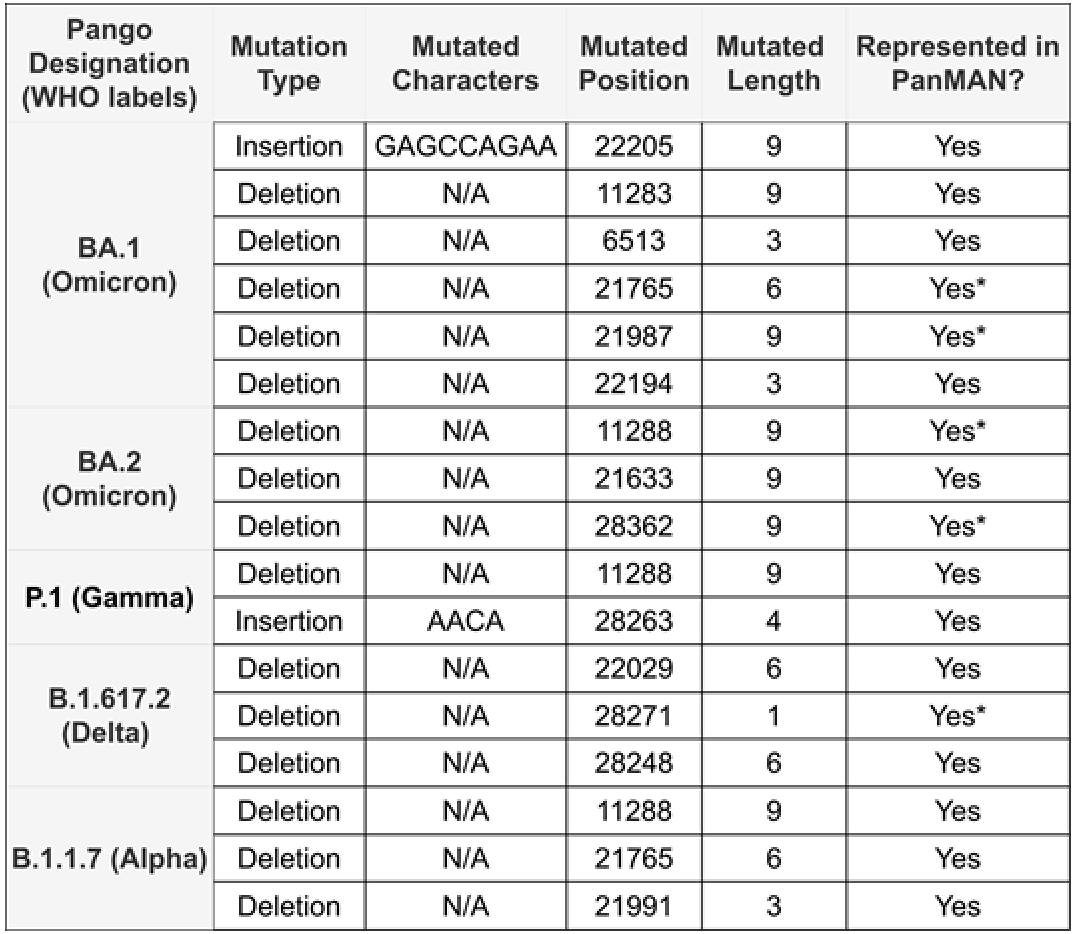




Exploration of SARS-CoV-2 mutational and evolutionary landscape using PanMAN

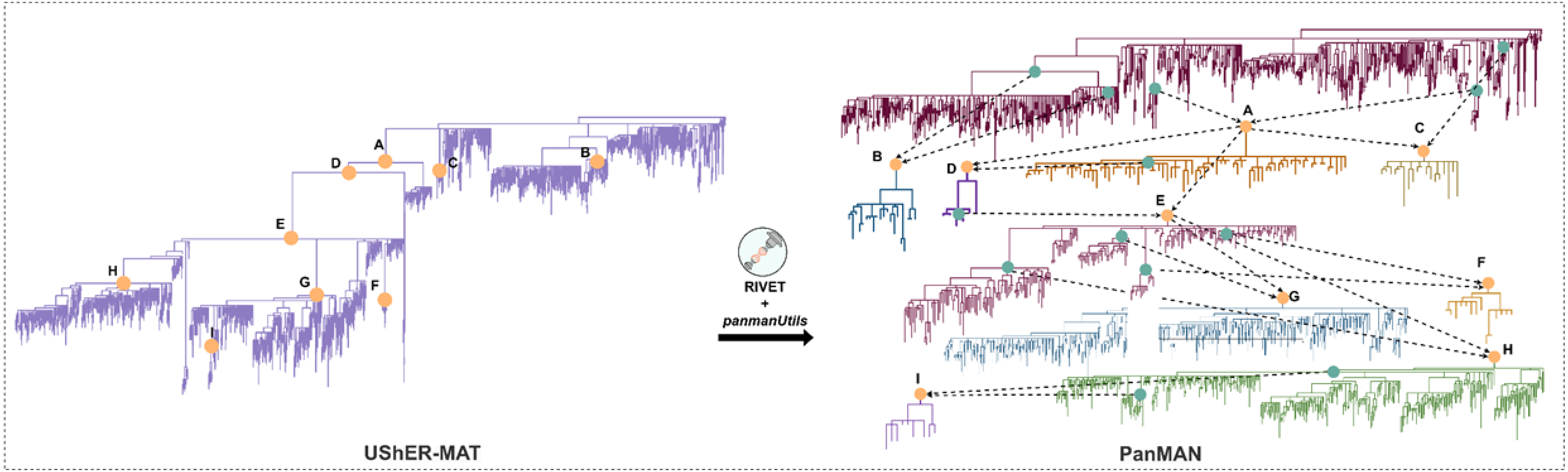


Exploration of SARS-CoV-2 mutational and evolutionary landscape using PanMAN



Pango Designation (WHO labels)	Mutation Type	Mutated Characters	Mutated Position	Mutated Length	Represented in PanMAN?
BA.1 (Omicron)	Insertion	GAGCCAGAA	22205	9	Yes
	Deletion	N/A	11283	9	Yes
	Deletion	N/A	6513	3	Yes
	Deletion	N/A	21765	6	Yes*
	Deletion	N/A	21987	9	Yes*
	Deletion	N/A	22194	3	Yes
BA.2 (Omicron)	Deletion	N/A	11288	9	Yes*
	Deletion	N/A	21633	9	Yes
	Deletion	N/A	28362	9	Yes*
P.1 (Gamma)	Deletion	N/A	11288	9	Yes
	Insertion	AACA	28263	4	Yes
B.1.617.2 (Delta)	Deletion	N/A	22029	6	Yes
	Deletion	N/A	28271	1	Yes*
	Deletion	N/A	28248	6	Yes
B.1.1.7 (Alpha)	Deletion	N/A	11288	9	Yes
	Deletion	N/A	21765	6	Yes
	Deletion	N/A	21991	3	Yes

Exploration of SARS-CoV-2 mutational and evolutionary landscape using PanMAN



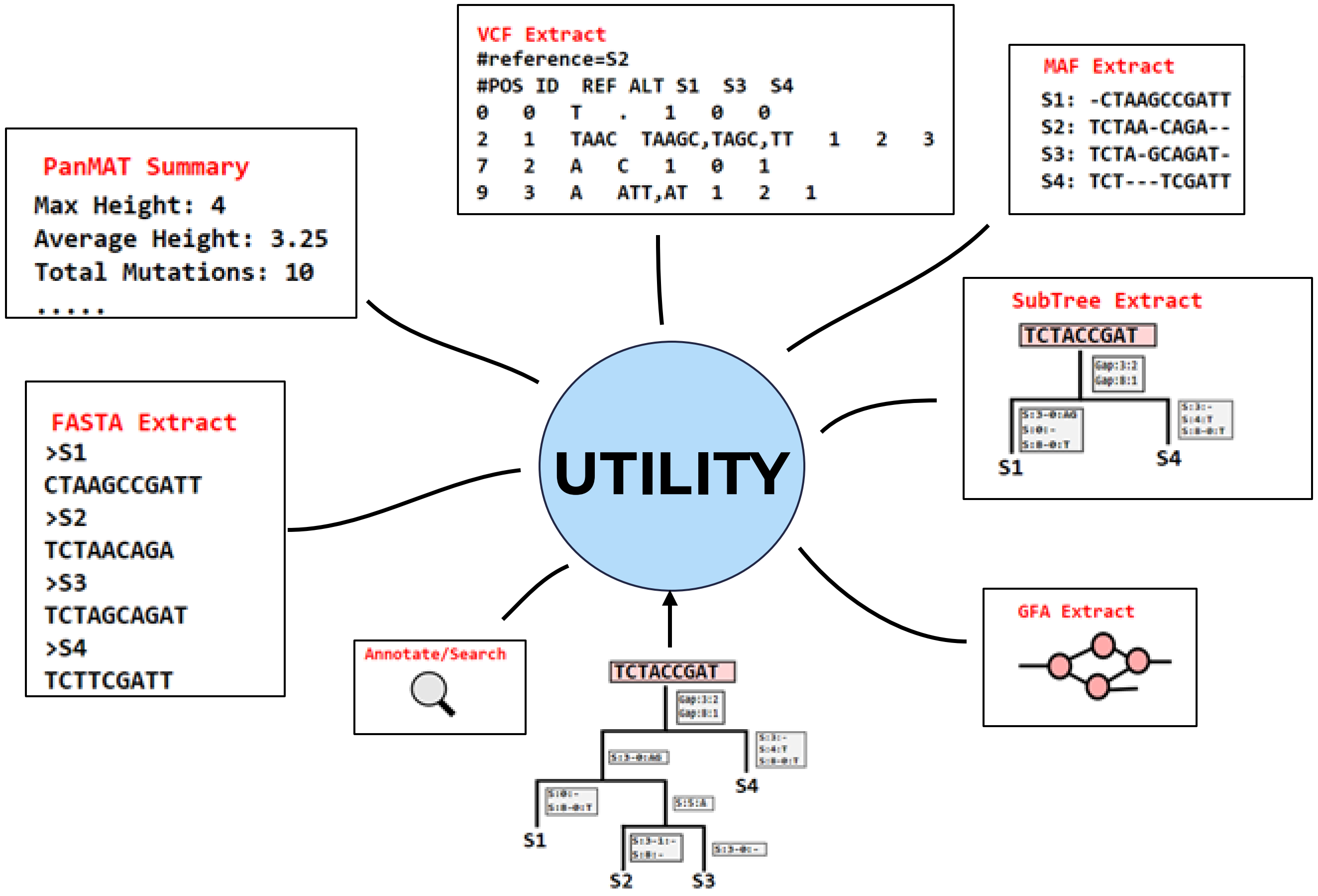
PanMANs using likelihood

- Ancestral sequences in PanMAN can be inferred by a a variety of techniques:
  - Parsimony, e.g. Fitch algorithm
  - Likelihood, e.g. PastML, MPPA

- Appears to have a noticeable impact on the file sizes

Dataset	# sequences	File size (MB)		Ratio (LK/parsimony)
		PanMAT-Parsimony	PanMAT-LK	
Sars-CoV2	20	0.019	0.019	1
	200	0.083	0.1	1.2
	2000	0.68	0.91	1.3
	20000	4.8	6.1	1.3
RSV	50	0.047	0.17	3.6
	500	0.137	0.65	4.7
	5000	1.1	4.3	3.9
TB	40	1.9	9.3	4.9
	400	5.1	39.7	7.8

PanMAN Utility for Common Bioinformatic Analyses



PanMAT Summary

```
Max Height: 4
Average Height: 3.25
Total Mutations: 10
.....
```

VCF Extract

```
#reference=S2
#POS ID  REF ALT S1  S3  S4
0   0   T   .   1   0   0
2   1   TAAC  TAAGC,TAGC,TT   1   2   3
7   2   A   C   1   0   1
9   3   A   ATT,AT  1   2   1
```

MAF Extract

```
S1: -CTAAGCCGATT
S2: TCTAA-CAGA--
S3: TCTA-GCAGAT-
S4: TCT---TCGATT
```

FASTA Extract

```
>S1
CTAAGCCGATT
>S2
TCTAACAGA
>S3
TCTAGCAGAT
>S4
TCTTCGATT
```

UTILITY

SubTree Extract

GFA Extract

Annotate/Search

PanMAN