

# **CorGAT and CorGAT-tracker**

**Functional annotation of SARS-CoV-2 genomes and tracking mutations  
and variants of concern**

# CorGAT

- Coronavirus Genome Analysis Tool or CorGAT (DOI:10.1093/bioinformatics/btaa1047).
- Collection of Perl utilities and annotation files.
- Performs the functional annotation of SARS-CoV-2 genetic variants.

**Galaxy / CorGAT** Analyze Data Workflow Visualize Shared Data Help Login or Register Using 79.7 KB

**Tools** search tools

**Get Data**

- [Fasta Extract Sequence](#) Extract a single sequence from a fasta file.
- [Upload File](#) from your computer
- [UCSC Main](#) table browser
- [UCSC Archaea](#) table browser
- [EBI SRA](#) ENA SRA
- [modENCODE fly](#) server
- [InterMine](#) server
- [Flymine](#) server
- [modENCODE modMine](#) server
- [MouseMine](#) server
- [Ratmine](#) server
- [YeastMine](#) server
- [modENCODE worm](#) server
- [WormBase](#) server
- [ZebrafishMine](#) server

**Welcome, to CorGAT!**  
See the [CorGAT manual](#) for an explanation of what you can do with CorGAT

[Configuring Galaxy >>](#) [Installing Tools >>](#)

Take an interactive tour: [Galaxy UI](#) [History](#) [Scratchbook](#)

Galaxy is an open platform for supporting data intensive research. Galaxy is developed by The Galaxy Team with the support of many contributors.

The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Johns Hopkins University.

**History** search datasets

**Unnamed history**  
11 deleted  
79.68 KB

**i** This history is empty. You can load your own data or get data from an external source

- Brief explanation of functioning:

1. Alignment of complete assemblies of SARS-CoV-2 genomes to the reference sequence.
2. Obtain a list of polymorphic regions.
3. Functional annotation of the identified variants.

**Galaxy / CorGAT** | Analyze Data | Workflow | Visualize | Shared Data | Help | Login or Register | Using 150.3 KB

**Tools**  
 search tools

**Graph/Display Data**  
 Phenotype Association  
 genome\_alignment  
 Coronavirus Genome Annotation Tool

**multiFC** Process multi-fasta files to derive a phenetic matrix of genetic variants. (Galaxy Version 1)

Options

Input: 15: Test.fa

What it does:  
 This tool is used to align SARS-CoV-2 genes, in multifasta format. Genomes will be aligned to the reference SARS-CoV-2 genome using nucmer. The output will consist in a single tabular file with as many columns as the number of genomes provided in input. And as many rows as the number of variants observed in the genomes. For every genome assembly and variant a simple binary code 1= present, 0=absent will be used to indicate whether that genome carries a specific variant. This table should be provided to the FunAnn tool to obtain the functional annotation of the variants.

**History**  
 search datasets

Unnamed history  
 2 shown, 14 deleted  
 150.26 KB

16: GCA\_009858895.3\_ASM985889v3\_genomic.fna  
 15: Test.fa

**Galaxy / CorGAT** | Analyze Data | Workflow | Visualize | Shared Data | Help | Login or Register | Using 151.7 KB

**Tools**  
 search tools

**Graph/Display Data**  
 Phenotype Association  
 genome\_alignment  
 Coronavirus Genome Annotation Tool

**FunAnn** Performs functional annotation of genetic variants (Galaxy Version 1)

Options

Input: 18: Test\_GenVar.tsv

What it does:  
 This program reads a tabular formatted file, in pseudo vcf format, as obtained from the join\_nucmer utility and performs functional annotation of SARS-CoV-2 variants. Please notice that the program performs minimum error checks, and that it is designed to work exclusively with the reference annotation of the SARS-CoV-2 genome as available from Genbank. A copy of the genome in fasta format can be found also in this Galaxy, under Shared Data -> Data Libraries -> SARS-CoV-2-REF.

**History**  
 search datasets

Unnamed history  
 4 shown, 14 deleted  
 151.71 KB

18: Test\_GenVar.tsv  
 17: multiFC on data 15: log file  
 16: GCA\_009858895.3\_ASM985889v3\_genomic.fna  
 15: Test.fa

**Galaxy / CorGAT** | Analyze Data | Workflow | Visualize | Shared Data | Help | Login or Register | Using 159.2 KB

**Tools**  
 search tools

**Graph/Display Data**  
 Phenotype Association  
 genome\_alignment  
 Coronavirus Genome Annotation Tool

**FunAnn** Performs functional annotation of genetic variants (Galaxy Version 1)

Options

Input: 18: Test\_GenVar.tsv

What it does:  
 The output file is again, a tabular file delineated by tabs, and providing different types of annotations for the variants included in the input file. A more detailed description of the output format can be found at: [https://github.com/matteo14c/SARS-CoV-2\\_annot](https://github.com/matteo14c/SARS-CoV-2_annot)

POS	REF	ALT	annot
241	C	T	5'UTR:nc.C241T,NA,NA;
733	T	C	nsp1:c.468T>C;p.D156D,synonymous;orf1abc:468T>C;p.D156D,synonymous;
1926	C	T	orf1abc:1661C>T;p.T554I,misense;nsp2:c.1121C>T;p.T374I,misense;
2035	G	T	orf1abc:1770G>T;p.L590F,misense;nsp2:c.1230G>T;p.L410F,misense;
2749	C	T	orf1abc:2484C>T;p.D828D,synonymous;nsp3:c.30C>T;p.D10D,synonymous;
3037	C	T	orf1abc:2772C>T;p.F924F,synonymous;nsp3:c.318C>T;p.F106F,synonymous;
3798	T	.	orf1abc:3533T>.p.E1192I,frameshiftDel;nsp3:c.1079T>.p.E374I,frameshiftDel;
3828	C	T	orf1abc:3563C>T;p.S1188L,misense;nsp3:c.1109C>T;p.S370L,misense;
4178	A	T	orf1abc:3913A>T;p.K130S,stopGain;nsp3:c.1459A>T;p.K487I,stopGain;
5096	AA	.	orf1abc:4831AA>.p.S1612I,frameshiftDel;nsp3:c.2377AA>.p.S794I,frameshiftDel;
5111	AA	.	orf1abc:4846AA>.p.L1621I,frameshiftDel;nsp3:c.2392AA>.p.L803I,frameshiftDel;
5224	T	C	orf1abc:4959T>C;p.T1653T,synonymous;nsp3:c.2505T>C;p.T835T,synonymous;
5367	G	T	orf1abc:5102G>T;p.R1701I,misense;nsp3:c.2648G>T;p.R883I,misense;
6319	A	G	orf1abc:6054A>G;p.P2018P,synonymous;nsp3:c.3600A>G;p.P1200P,synonymous;
6613	A	G	orf1abc:6348A>G;p.V2116V,synonymous;nsp3:c.3894A>G;p.V1298V,synonymous;
8017	G	T	orf1abc:7752G>T;p.A2584A,synonymous;nsp3:c.5298G>T;p.A1766A,synonymous;
11098	TTTACCTTT	.....	orf1abc:10833TTTACCTTT>.....p.FLFP3611F,inframeDel;nsp6:c.126TTTACCTTT>.....p.FLFP42F,inf
11291	G	A	orf1abc:11026G>A;p.G3676S,misense;nsp6:c.319G>A;p.G107S,misense;
11296	T	G	orf1abc:11031T>G;p.F3677L,misense;nsp6:c.324T>G;p.F108L,misense;
11483	G	.	orf1abc:11218G>.p.G3746I,frameshiftDel;nsp6:c.511G>.p.G177I,frameshiftDel;
11653	C	A	orf1abc:11388C>A;p.L3796L,synonymous;nsp6:c.681C>A;p.L227L,synonymous;
12778	C	T	orf1abc:12513C>T;p.V4171V,synonymous;nsp9:c.93C>T;p.Y31V,synonymous;

**History**  
 search datasets

Unnamed history  
 6 shown, 14 deleted  
 159.24 KB

20: FunAnn on data 18: log file  
 19: Functional annotation of SARS-CoV-2 genomes in tabular format  
 18: Test\_GenVar.tsv  
 17: multiFC on data 15: log file  
 16: GCA\_009858895.3\_ASM985889v3\_genomic.fna  
 15: Test.fa

- The output is a simple table containing:

1. Genomic position.
2. Reference allele.
3. Alternative allele.
4. Functional annotation.
5. Allele frequency
6. Epitopes annotation.
7. Annotation of sites under selective pressure.
8. MFE annotation.

POS	REF	ALT	annot
POS	REF	ALT	annot
241	C	T	5'UTR:nc.C241T,NA,NA;
733	T	C	nsp1:c.468T>C,p.D156D,synonymous;orf1ab:c.468T>C,p.D156D,synonymous;
1926	C	T	orf1ab:c.1661C>T,p.T554I,missense;nsp2:c.1121C>T,p.T374I,missense;
2035	G	T	orf1ab:c.1770G>T,p.L590F,missense;nsp2:c.1230G>T,p.L410F,missense;
2749	C	T	orf1ab:c.2484C>T,p.D828D,synonymous;nsp3:c.30C>T,p.D10D,synonymous;
3037	C	T	orf1ab:c.2772C>T,p.F924F,synonymous;nsp3:c.318C>T,p.F106F,synonymous;
3798	T	.	orf1ab:c.3533T>.,p.E1192*,frameshiftDel;nsp3:c.1079T>.,p.E374*,frameshiftDel;

AF	Epitopes
AF	
94.382501	
0.213728	
0.031701	
0.110145	410:LATNNLVVM,7,HLA-B*35:01;HLA-B*46:01;HLA-C*01:0;
0.214636	7:FGDDTVIEV,1,HLA-C*08:01;5
96.462537	98:LASHMYCSF,5,HLA-B*15
0	

Hyphy	MFE
Hyphy	MFE
NA	NA
NA	NA
fel:true;meme:false;kind:negative;	NA
fel:true;meme:true;kind:positive;	NA
NA	NA
fel:true;meme:false;kind:negative;	NA
NA	NA

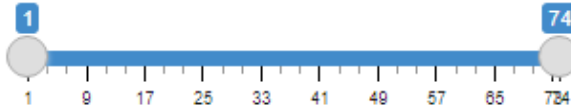
# CorGAT-tracker

- Shiny based dashboard for the visualization of the prevalence of SARS-CoV-2 lineages and mutations of concern.
- Based on CorGAT derived annotations.
- A Galaxy release is under development.
- Data are represented in an interactive way.
- Users can personalize data visualization through a series of widgets that allow to modify as many parameters, among which:
  1. The country of origin of the data.
  2. The interval of time to be displayed.
  3. The minimum number of sequenced genomes.
  4. A mutation of interest to be visualized.
  5. A lineage of interest to be represented.

### Country

Visualize data for the selected country

### Weeks range



Time lapse of interest (number of weeks from a fixed date)

### Min number of genomes (Lineages)

- 1
- 25
- 50
- 100
- 500
- 1000

Minimum number of sequenced genomes required to display a Lineage

### Min number of genomes (Lineages+)

- 1
- 5
- 10
- 15
- 25
- 50

Minimum number of sequenced genomes required to display a Lineage+

### Lineage

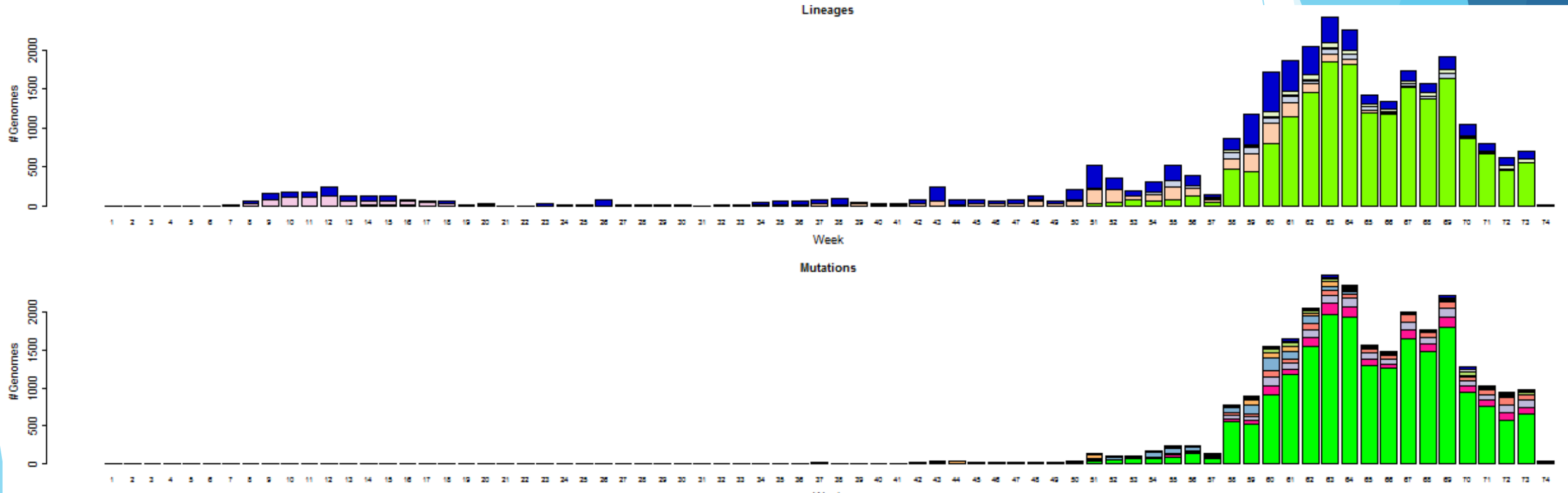
Produce a scatterplot for the selected Lineage

### Mutation

Produce a scatterplot for the selected Mutation

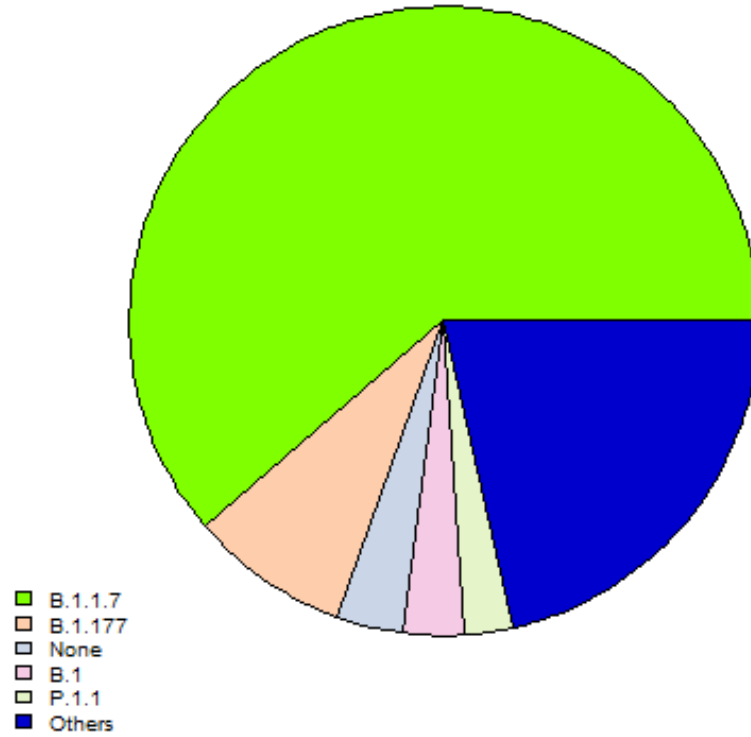
- CorGAT-tracker produces 3 different kinds of plot:

## 1. Barplots

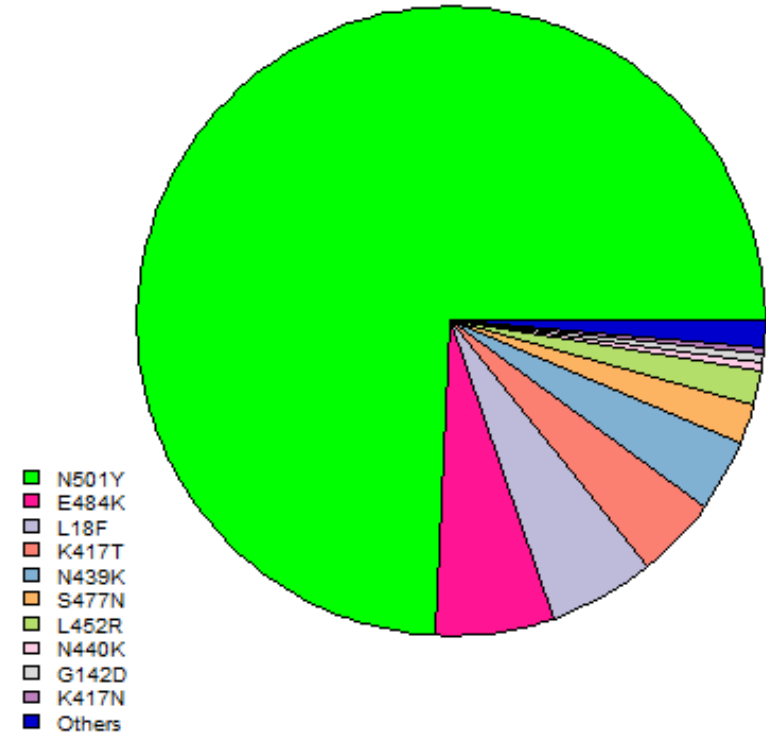


## 2. Pie-charts

Lineages



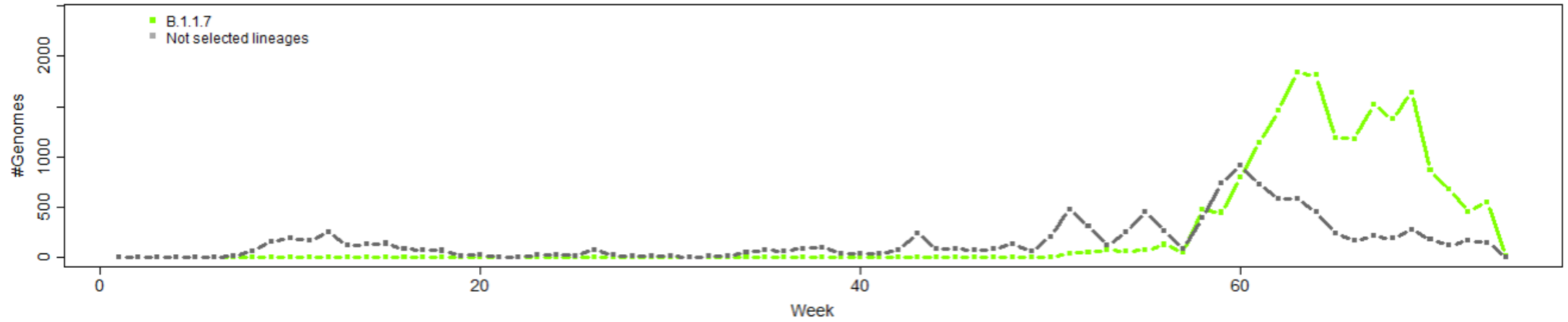
Mutations



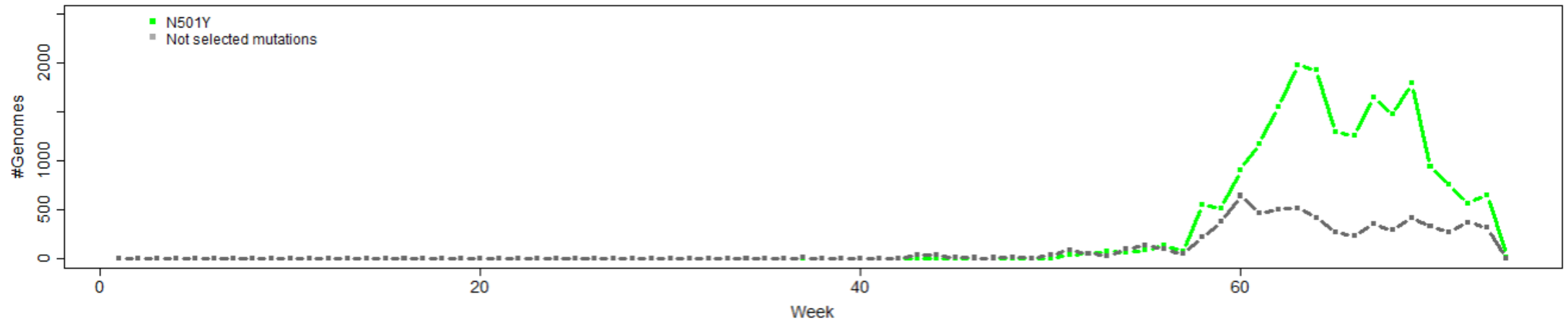


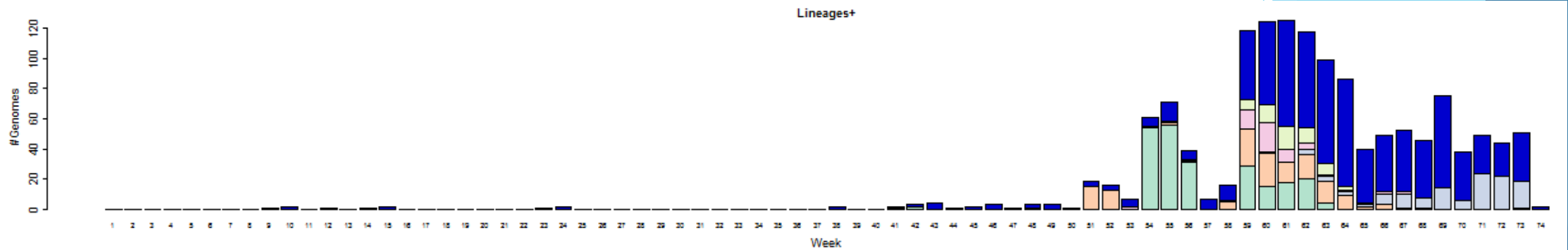
### 3. Scatterplots

**B.1.1.7**

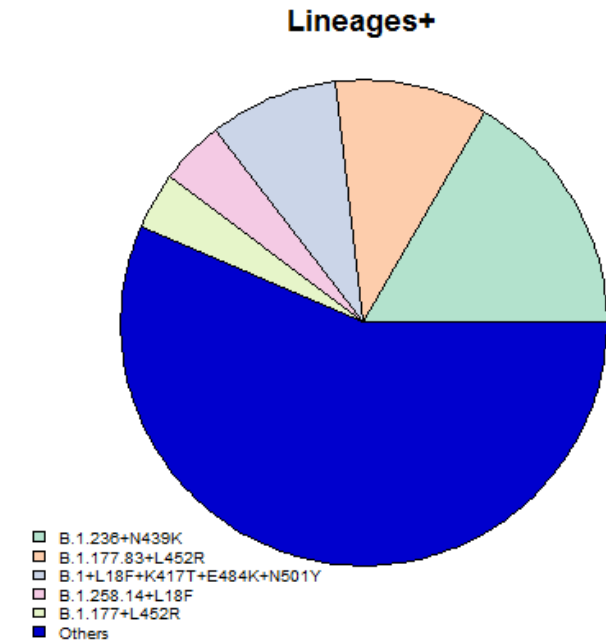


**N501Y**





- In CorGAT-tracker lineage annotations can be “augmented” by reporting the list of MOC that are observed in a genome, but are not specific to its assigned lineage.
- Augmented annotations are called Lineages+ in-app.
- Lineages+ prevalence in time is represented using a barplot and a pie-chart

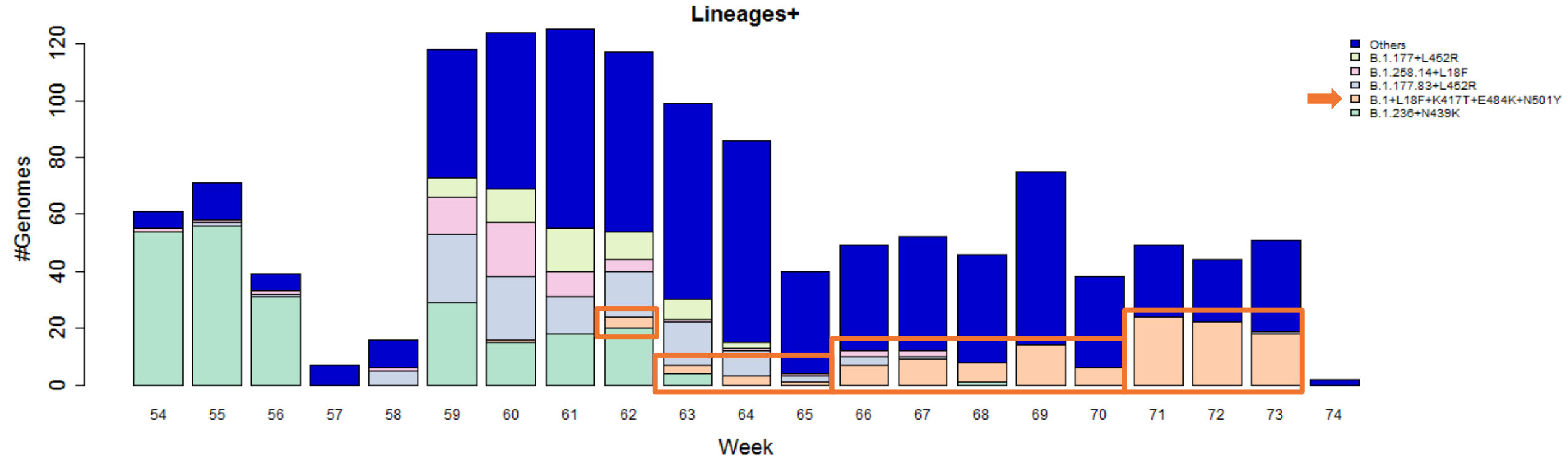


Min number of genomes (Lineages)

# Final considerations

- CorGAT and CorGAT-tracker will provide a useful addition to the currently available "arsenal" of bioinformatics methods for the genomic surveillance of SARS-CoV-2.
- CorGAT has a sensitivity comparable to other similar tools, but also provides additional layers of annotation.
- Example: Identification of misclassified SARS-CoV-2 genomes in Italy.

- Genomes classified as B.1 but presented additional mutations on the spike protein.



- In depth studies highlighted that:
  1. The majority of the additional spike mutations in the misclassified B.1 were in common with P.1.
  2. A “group specific” mutation, P681H, can be identified in the spike protein of the misclassified genomes.
- It is possible to speculate that these genomes represent a newly emerged lineage, however further investigations are required.
- P.1+P681H was recently added to the ECDC list of Variants Under Monitoring

# Availability

- CorGAT is already available through Galaxy at the following link:
  - ❖ <http://corgat.cloud.ba.infn.it/galaxy>
- A Galaxy release for CorGAT-tracker is under development.
- Further information about the tools can be found in the respective GitHub repositories:
  - ❖ <https://github.com/matteo14c/CorGAT> (CorGAT)
  - ❖ <https://github.com/F3rika/CorGAT-tracker> (CorGAT-tracker)

# Thank you!

**Special thanks to:**

- Matteo Chiara
- Federico Zambelli
- Marco Antonio Tangaro
- Pietro Mandreoli
- David S. Horner
- Graziano Pesole