Estimation of Covid-19 variants from wastewater without prior lineage classification

Sampling wastewater is a useful approach for monitoring the increase of COVID-19 cases by measuring the viral load.

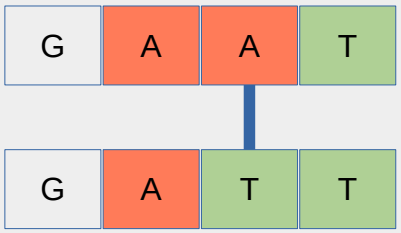
So far, for estimating new variants wastewater-based approaches require the prior definition of lineages which are gained through the sequencing of clinical samples.

→ Goal is to eliminate the need for clinical sampling especially considering the low testing frequencies

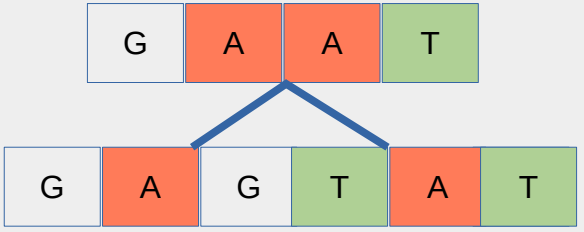
Approach based on tracking changes in **genetic diversity**

Genetic Diversity is defined based on the proportions of mutations between two samples:

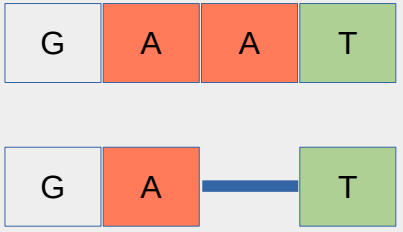
SNPs: Single Nucleotide Polymorphisms



Insertions



Deletions



Idea: Genetic diversity changes when a new variant emerges

Approach based on tracking changes in **genetic diversity: LogK**

H_s: diversity **within the combined populations** from two different time points

H_p: diversity **between the populations** from two different time points

$$LogK = \ln\left(\frac{H_s}{H_p}\right)$$

Approach based on tracking changes in **genetic diversity: LogK**

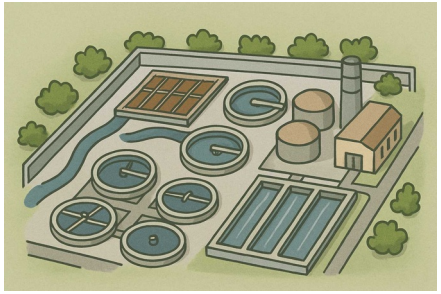
Idea: Genetic diversity changes when a new variant emerges

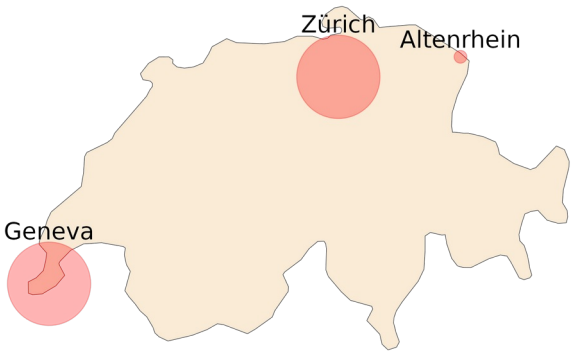
→ calculate LogK for each pairwise sample combination (increase in LnK means increase in genetic diversity)

→ track genetic diversity as a function of time to establish an early warning system for emerging variants

Analysis

Sequencing data from three Swiss wastewater facilities (openly available data, ENA ID: PRJEB44932)[1] available on almost daily basis





catchment	population	catchment area	Analyzed time period
Altenrhein, St.Gall	64000	104 km²	01.02.2021 – 01.02.2022
Geneva	454000	126 km²	01.11.2021 – 01.08.2022
Zurich	471000	102 km²	08.12.2020 – 30.09.2022

Analysis

Bioinformatics Pipeline:

- Alignment to reference sequence
- SNP calling with HaplotypeCaller: calculating SNP proportions at each position → used to calculate the genetic divergence
- filtering data based on genome coverage (>90) and read depth (> 30)

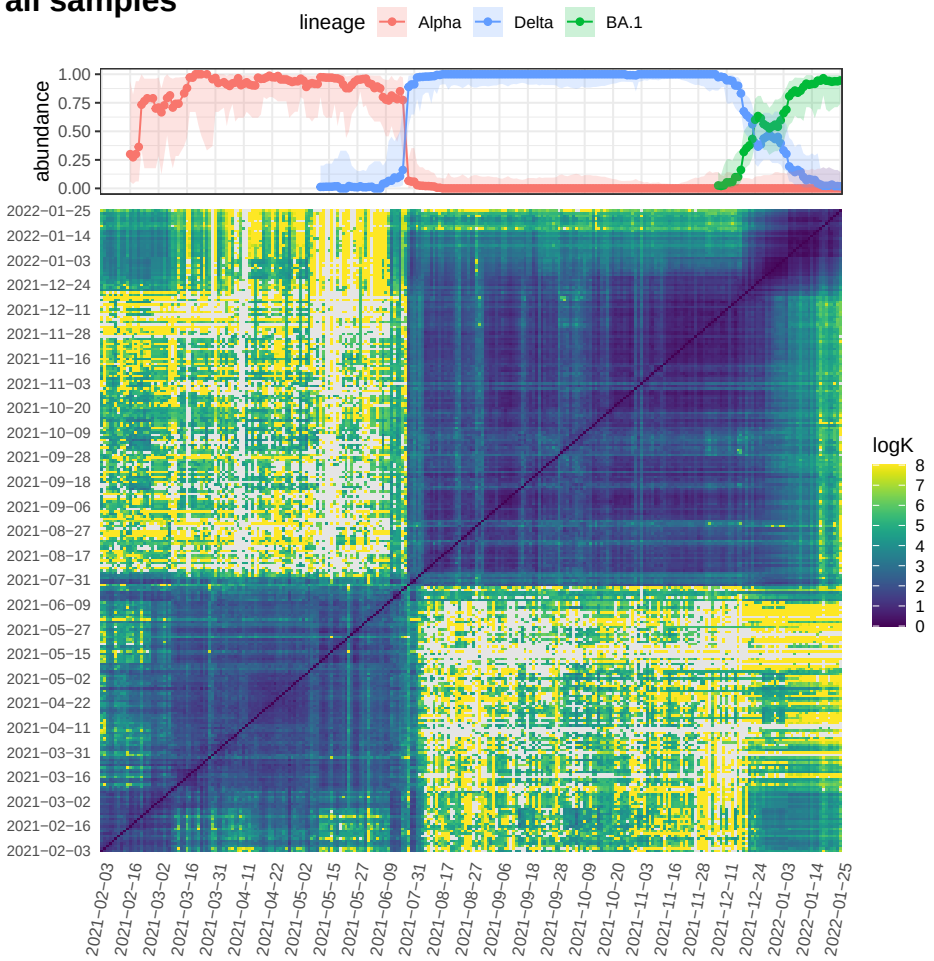
Read Depth: The average number of reads per base

Coverage: The percentage of the genome sequenced at a certain depth

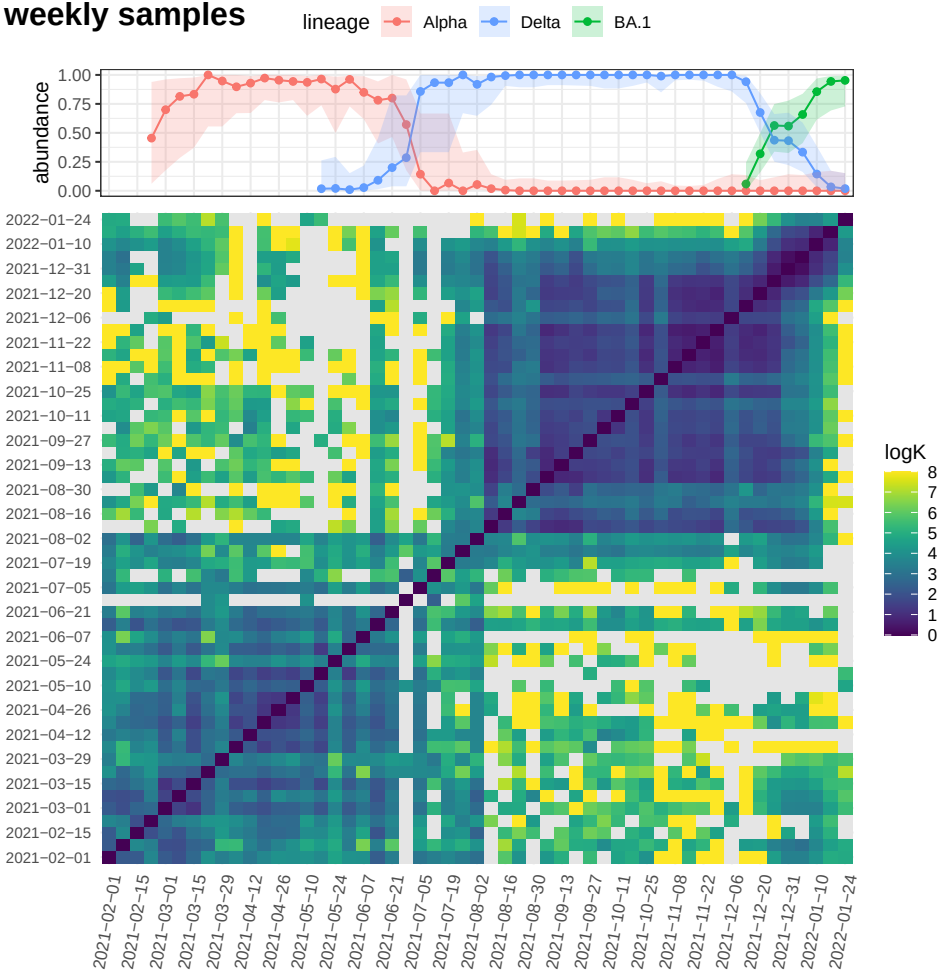
Additionally to daily samples, weekly samples and samples taken on 3 days per week have been analyzed (samples only from Mondays were selected). (Those were not filtered)

Results: LogK for Altenrhein

all samples

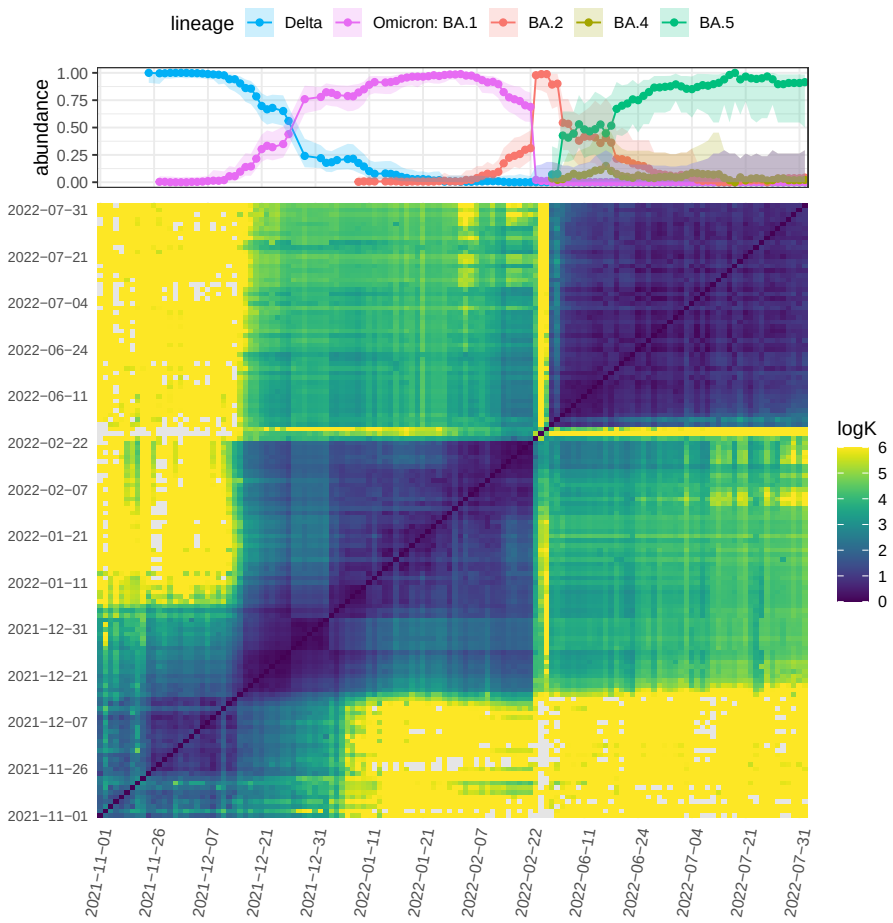


weekly samples

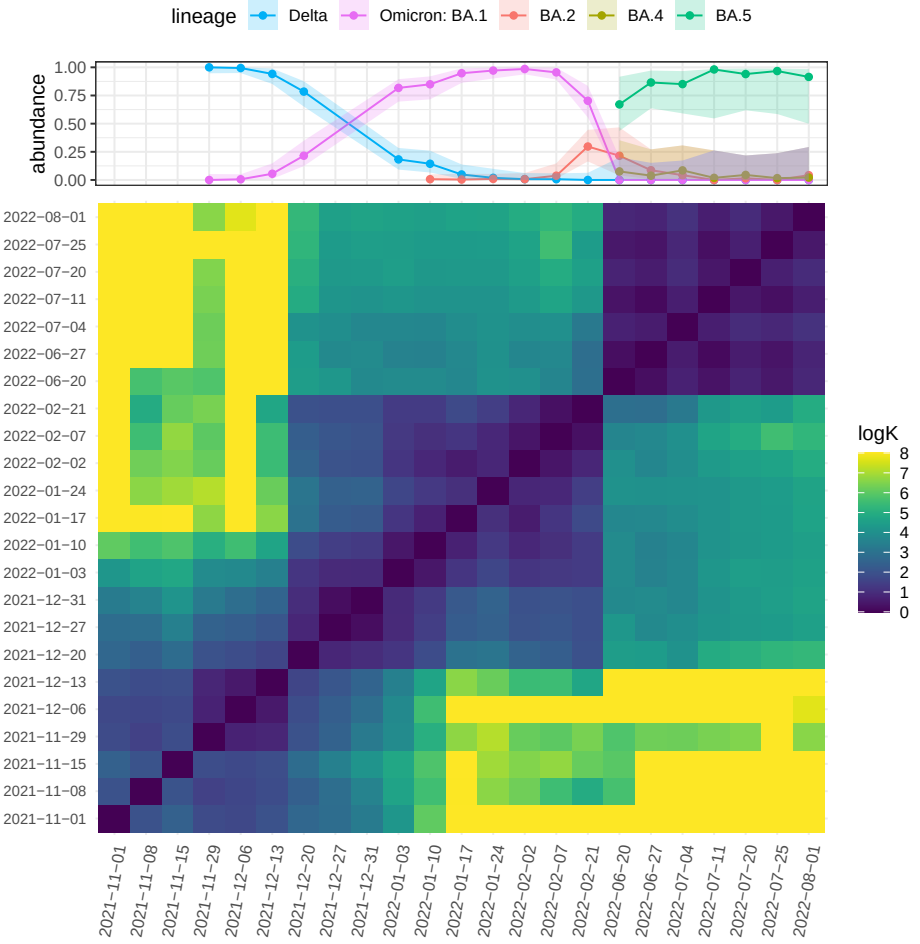


Results: LogK for Geneva

all samples

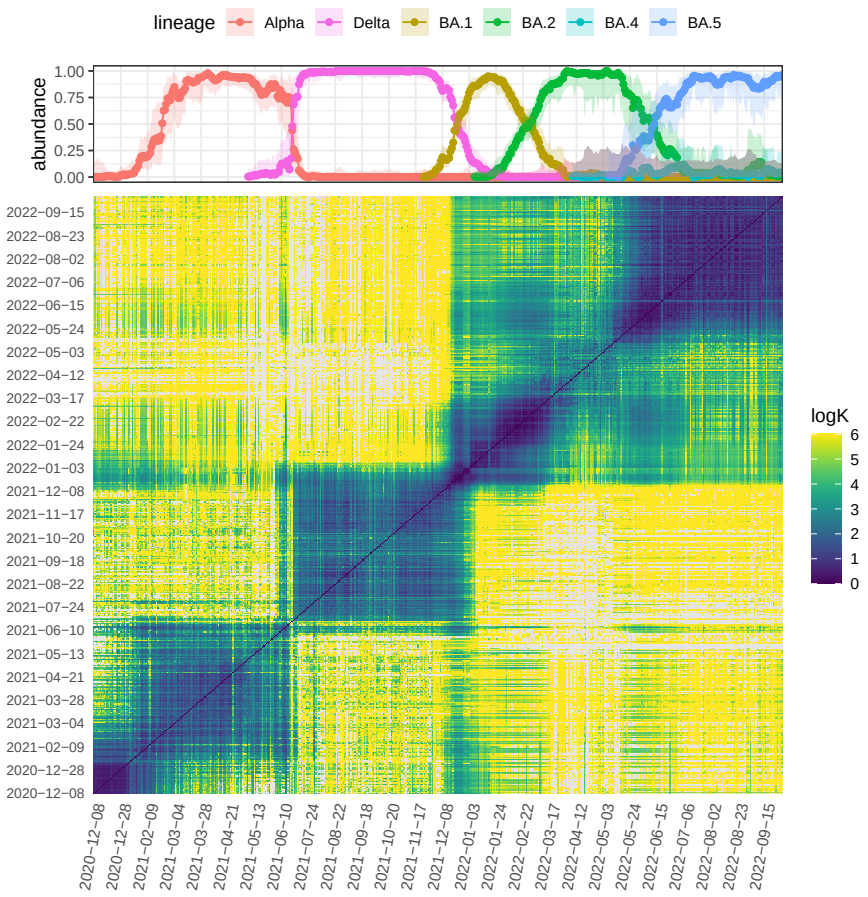


weekly samples

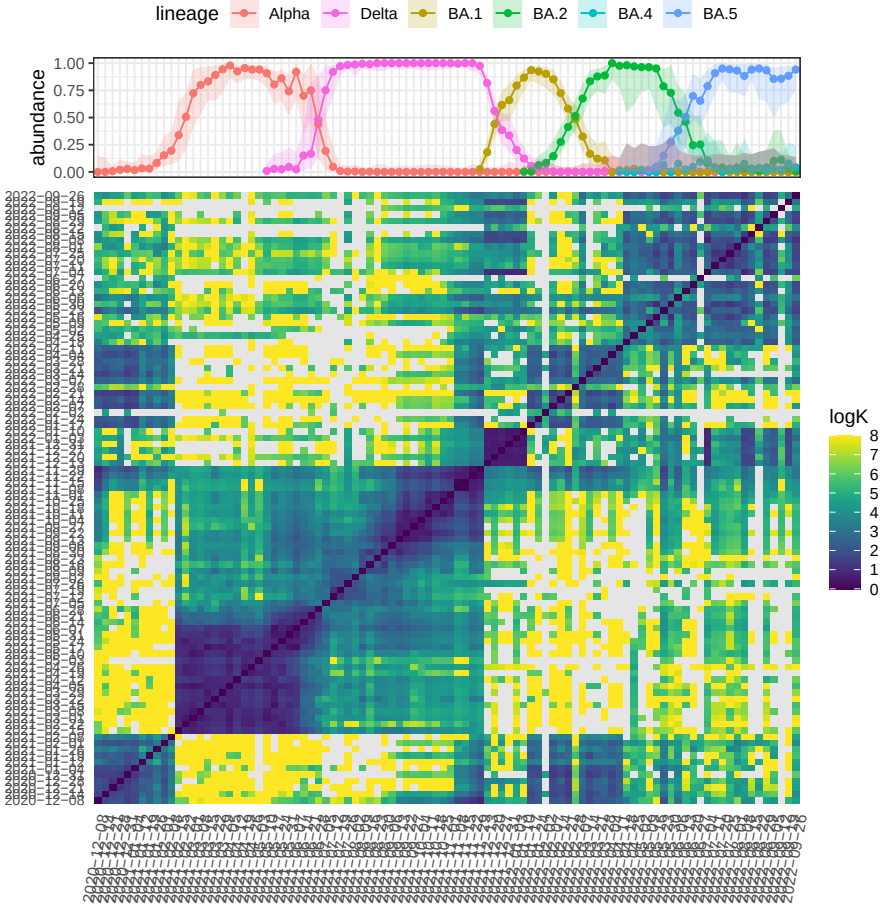


Results: LogK for Zurich

all samples



weekly samples



Approach based on tracking changes in **genetic diversity: LogK**

Idea: Genetic diversity changes when a new variant emerges

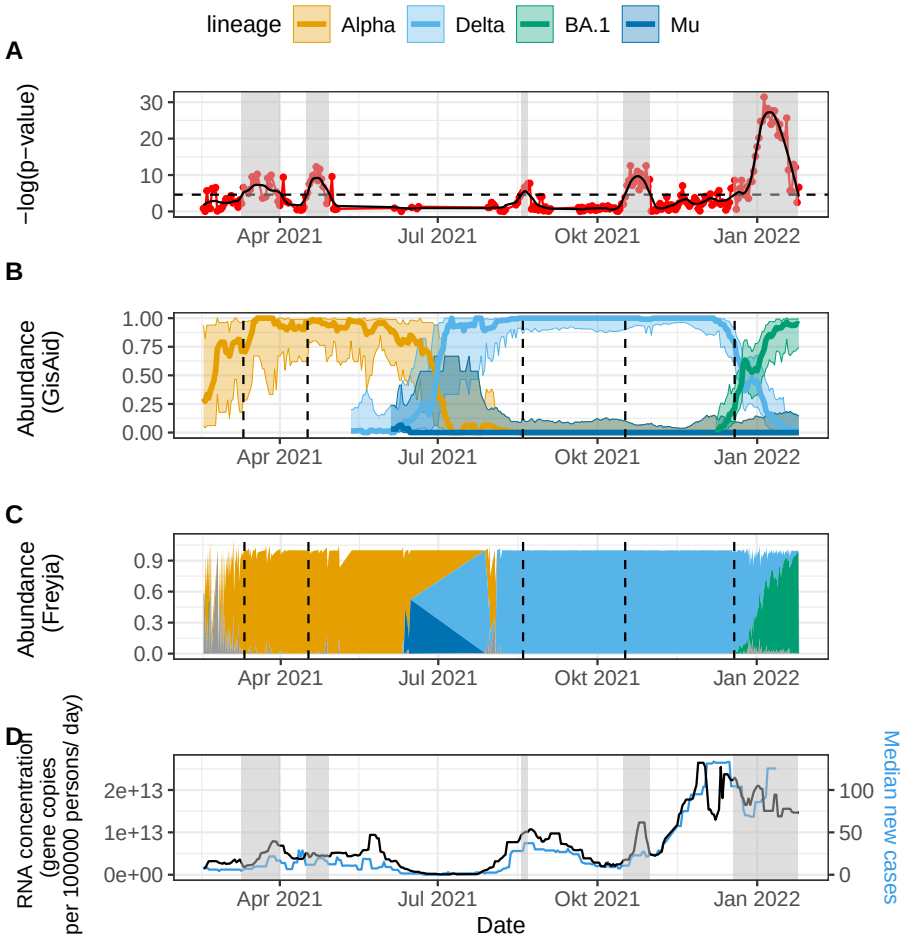
→ track genetic diversity as a function of time to establish an early warning system for emerging variants

→ calculate the median of the differences of measurements of the last 25 days (ΔLogK)

→ correlate ΔLogK with time of the size of previous 21 or 42 days using kendall's Tau ranked correlation

→ plot p-values of the correlation as a measure for emerging variant

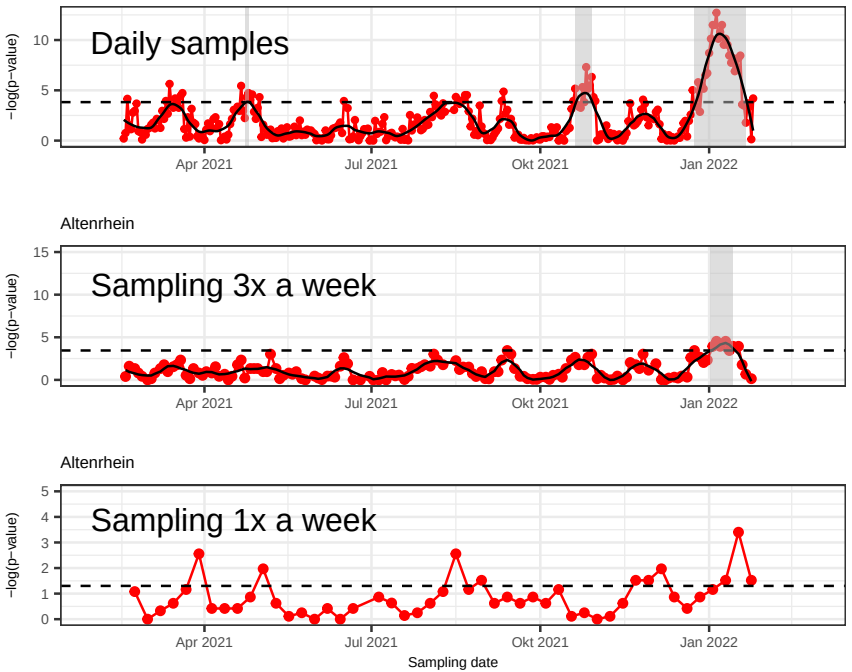
Results: Altenrhein



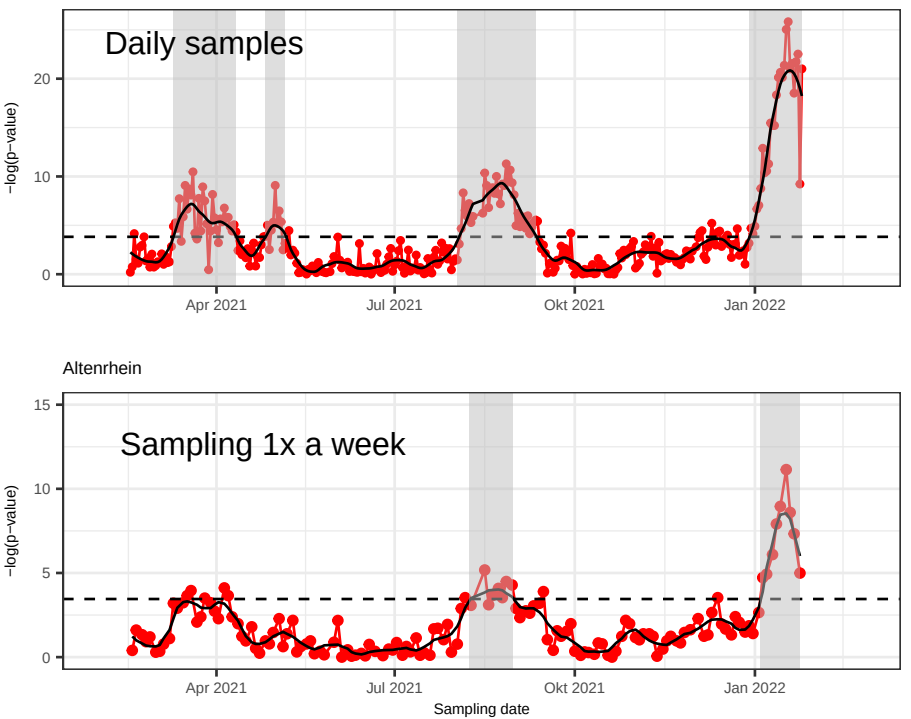
- Threshold: 0.01
- Correlations for 21 days
- GisAid: circulating variants based on clinical samples
- Freyja: gold standard variant estimation from wastewater, relies on preceding variant definition
- Missing data in July 2021 (Delta)

Analysis of sampling frequency and correlation time for unfiltered data for Altenrhein

Correlation time: 21 days

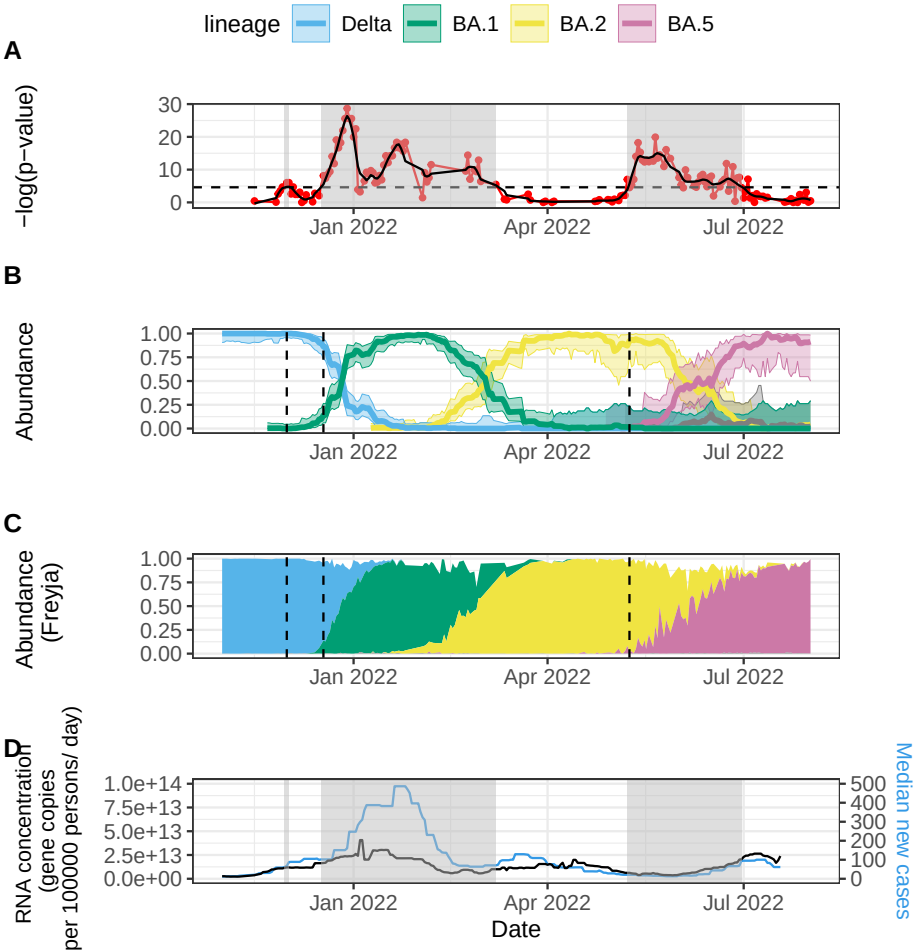


Correlation time: 42 days



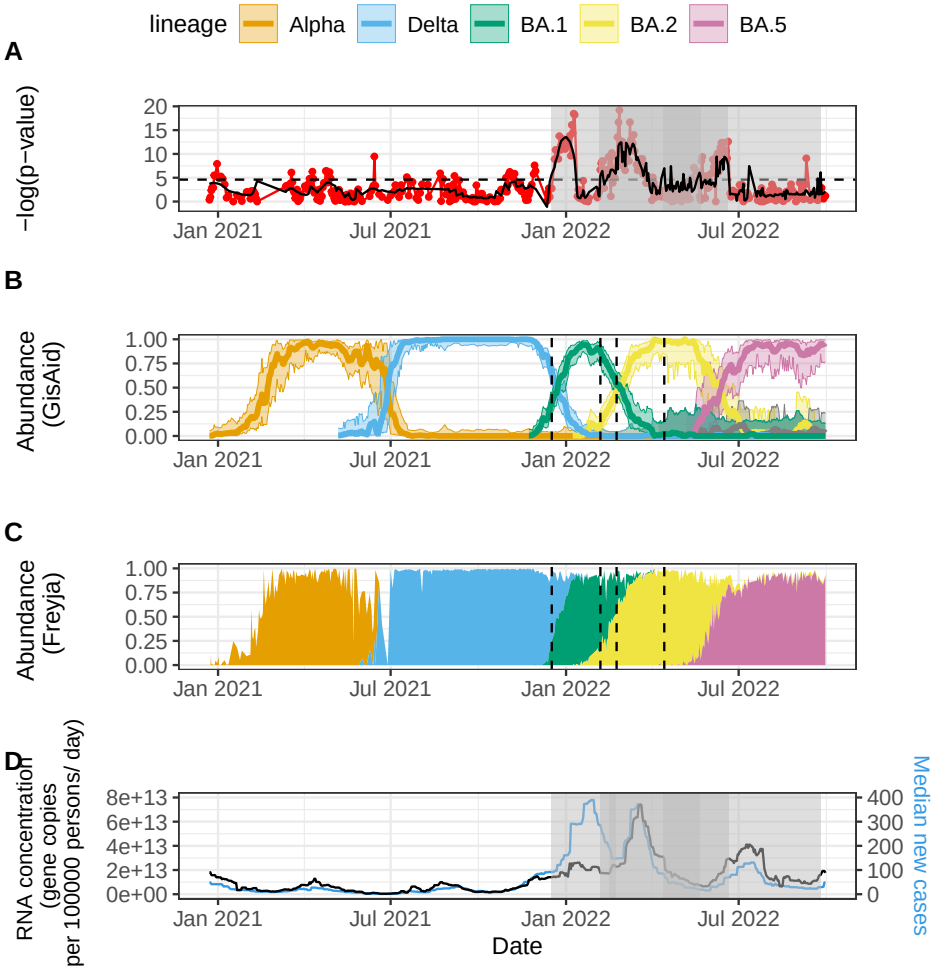
Threshold: 0.05/N; N=number of samples

Results: Geneva



- Threshold: 0.01
- Correlations for 21 days
- GisAid: circulating variants based on clinical samples
- Freyja: gold standard variant estimation from wastewater, relies on preceding variant definition
- Missing data in July 2021 (Delta)

Results: Zurich



- Threshold: 0.01
- Correlations for 21 days
- GisAid: circulating variants based on clinical samples
- Freyja: gold standard variant estimation from wastewater, relies on preceding variant definition
- Missing data in July 2021 (Delta)

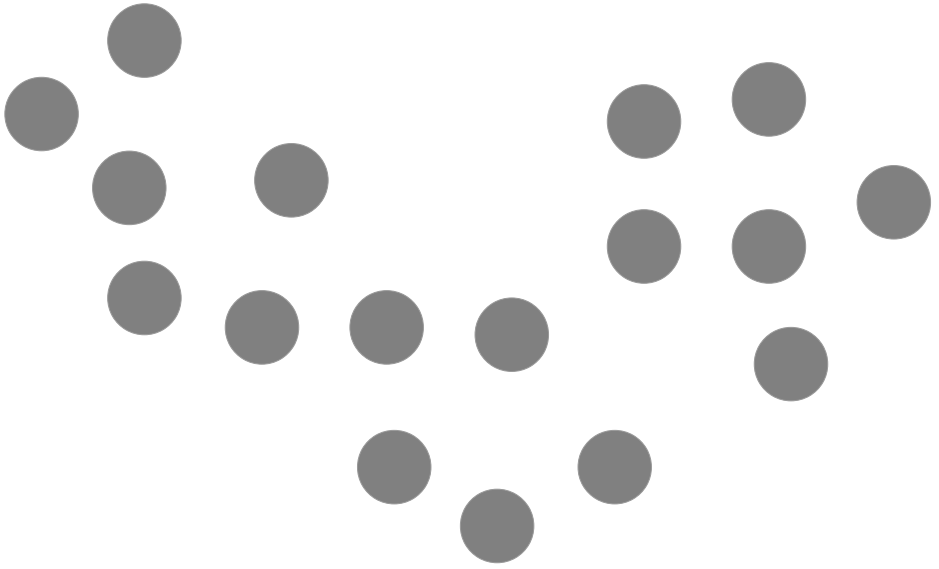
Variant abundance estimation

Idea: Cluster SNPs with the **k-means clustering** method for every data point. The median of the cluster should correlate with the abundance of the variant.

k-mean clustering

- represents a data set with k points in space
- distance function required (euclidean distance)
- minimizes distance between the data point and its closest centroid
- user must set the k-value

k-means clustering



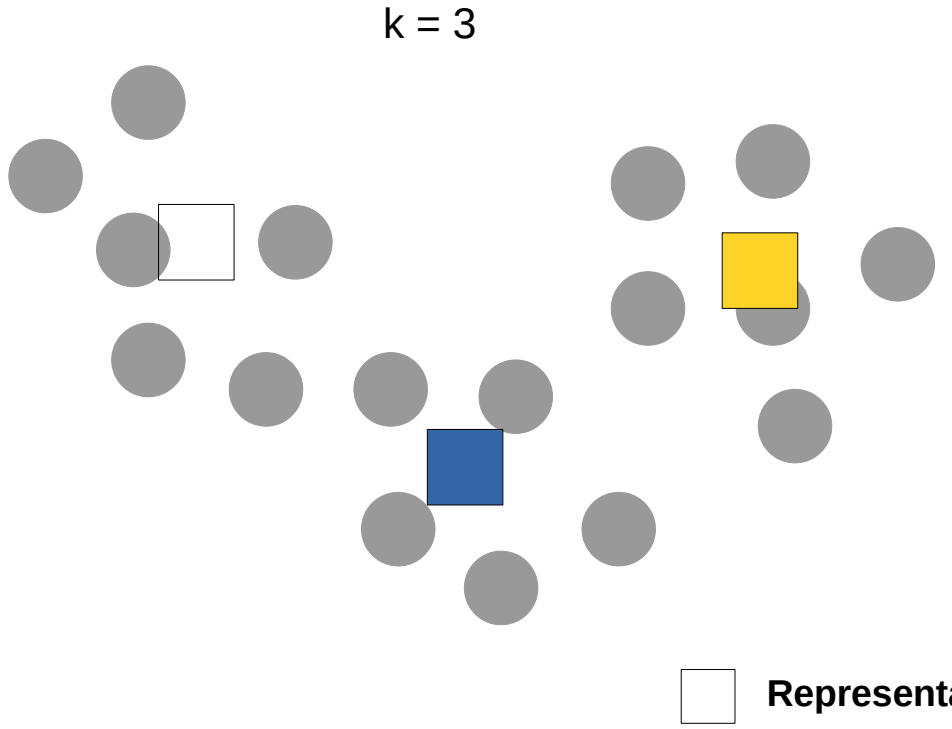
k-means clustering

$$\sum_{k} \sum_{x_i \in C_K} \|x_i - \mu_i\|^2$$

clusters

cluster points

centroid location



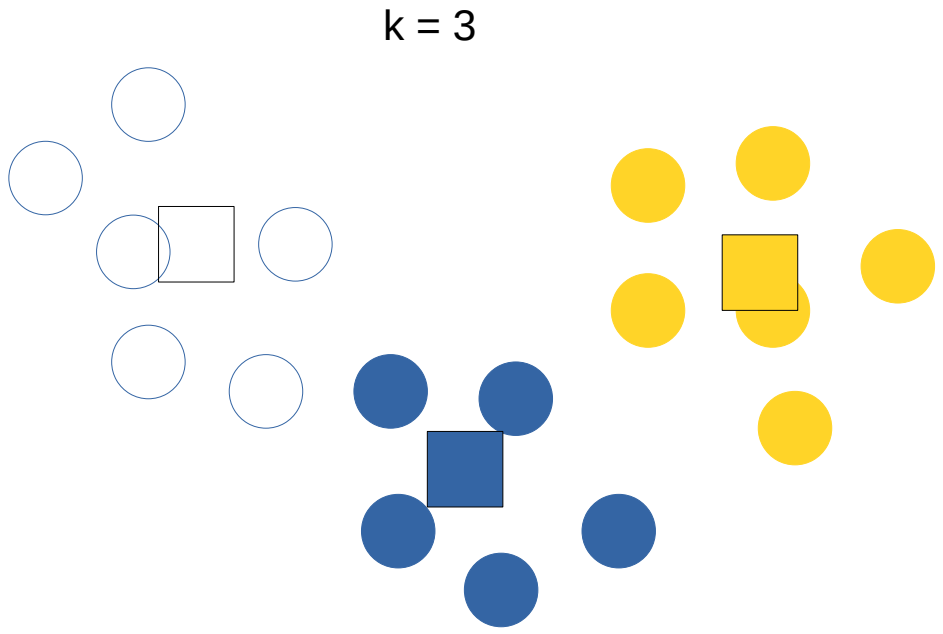
k-means clustering

$$\sum_{k} \sum_{x_i \in C_K} \|x_i - \mu_i\|^2$$

clusters

cluster points

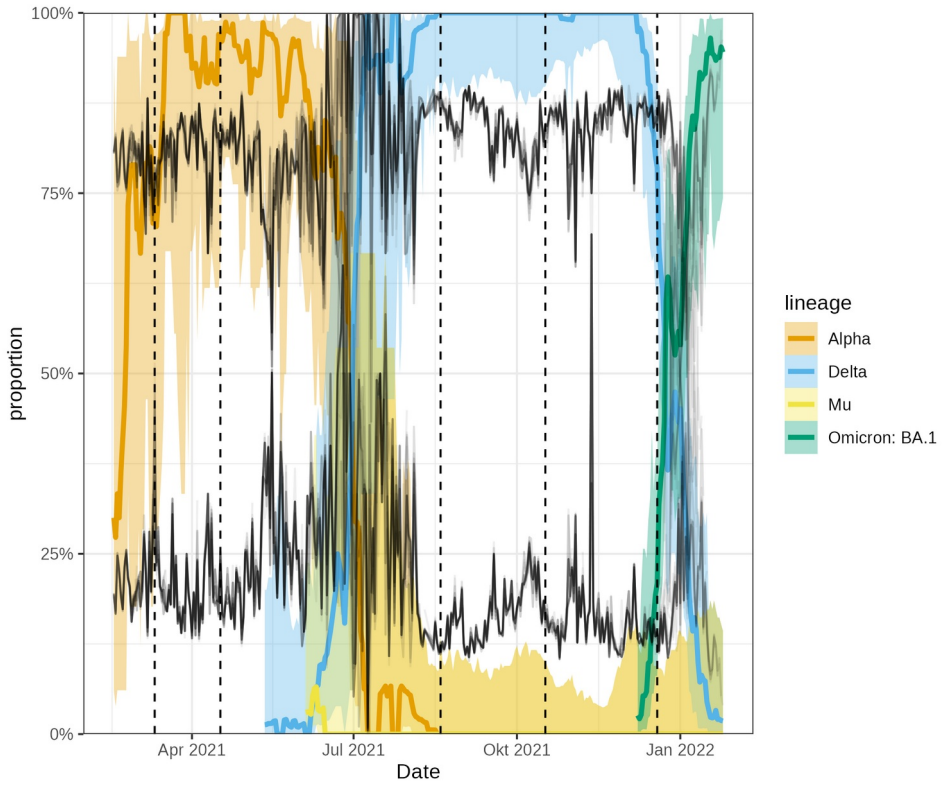
centroid location



Iterative process where
centroid position is updated at each step
by calculating the **mean** of all data points
assigned to that centoid

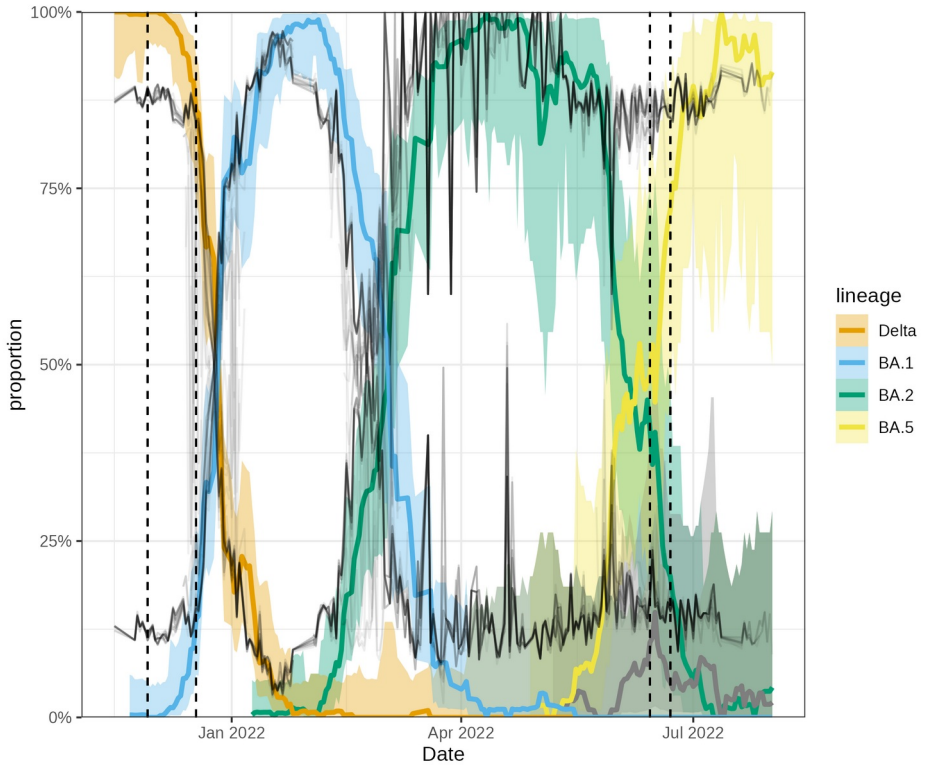
Clustering

Variant abundance estimation with k-means clustering: **Altenrhein**



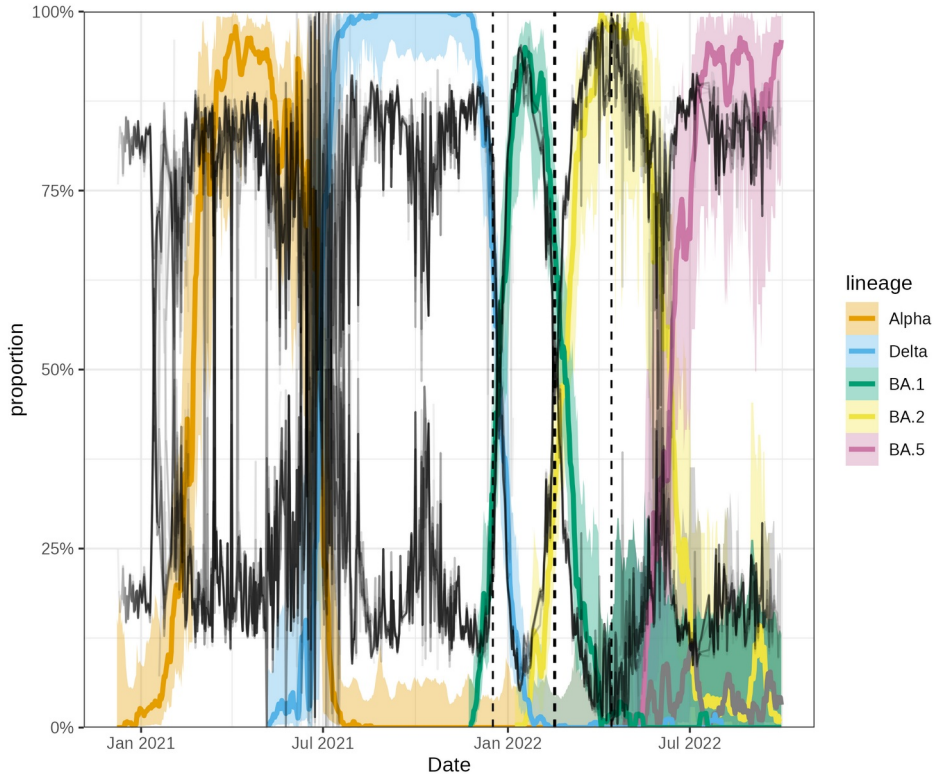
- Cluster mutations in sliding windows of 21 days with k=2
- High noise in July 2021 (Delta)
- Negative Kendalls' Tau **correlation coefficient** between residuals and coverage:
  $R = -0.22$, $p < 0.001$

Variant abundance estimation with k-means clustering: **Geneva**



- Cluster mutations in sliding windows of 21 days with k=2
- Negative Kendalls' Tau **correlation coefficient** between residuals and coverage:
  $R = -0.22$, $p < 0.001$

Variant abundance estimation with k-means clustering: **Zurich**



- cluster mutations in sliding windows of 21 days with k=2
- High noise in July 2021 (Delta)
- Negative Kendalls' Tau **correlation coefficient** between residuals and coverage:
  $R = -0.19$, $p < 0.001$

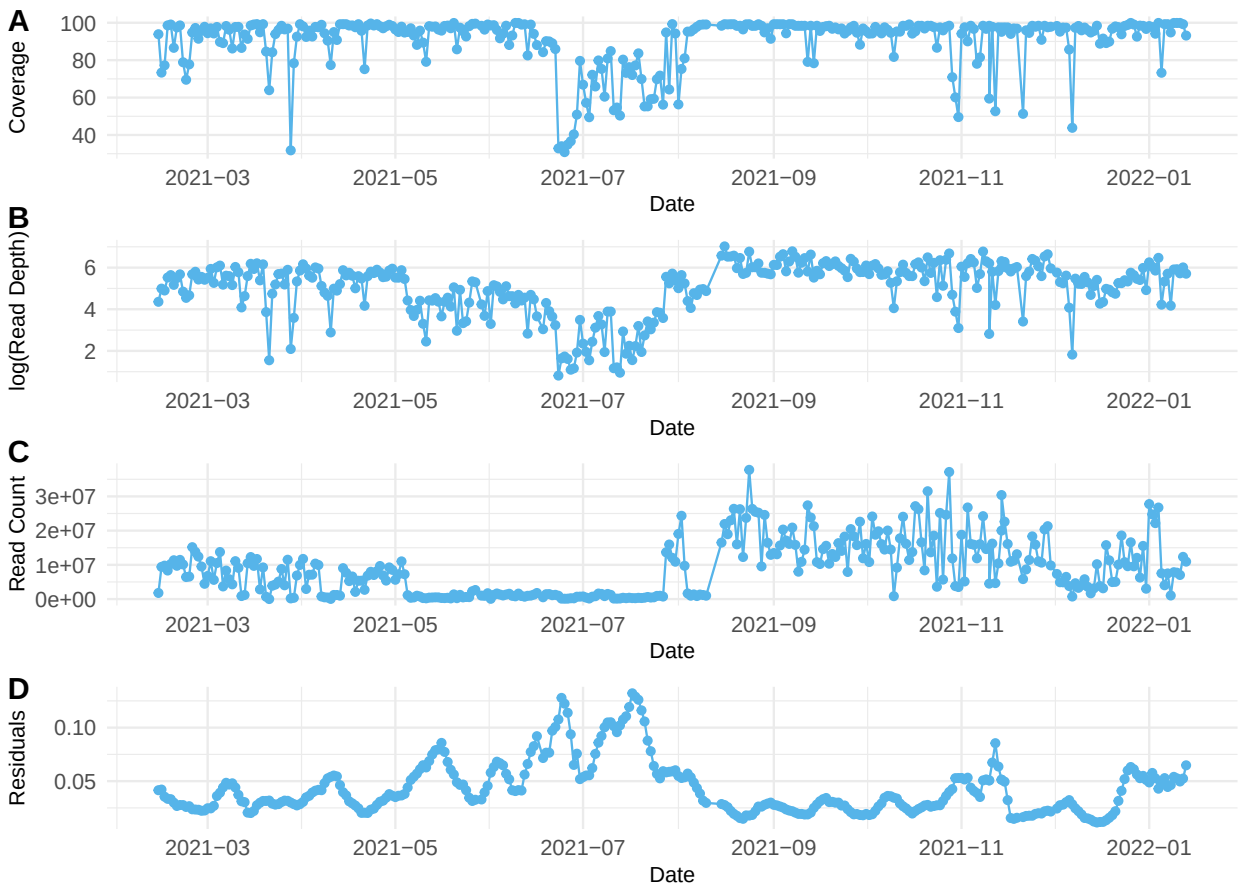
Thank you :)

Any questions?

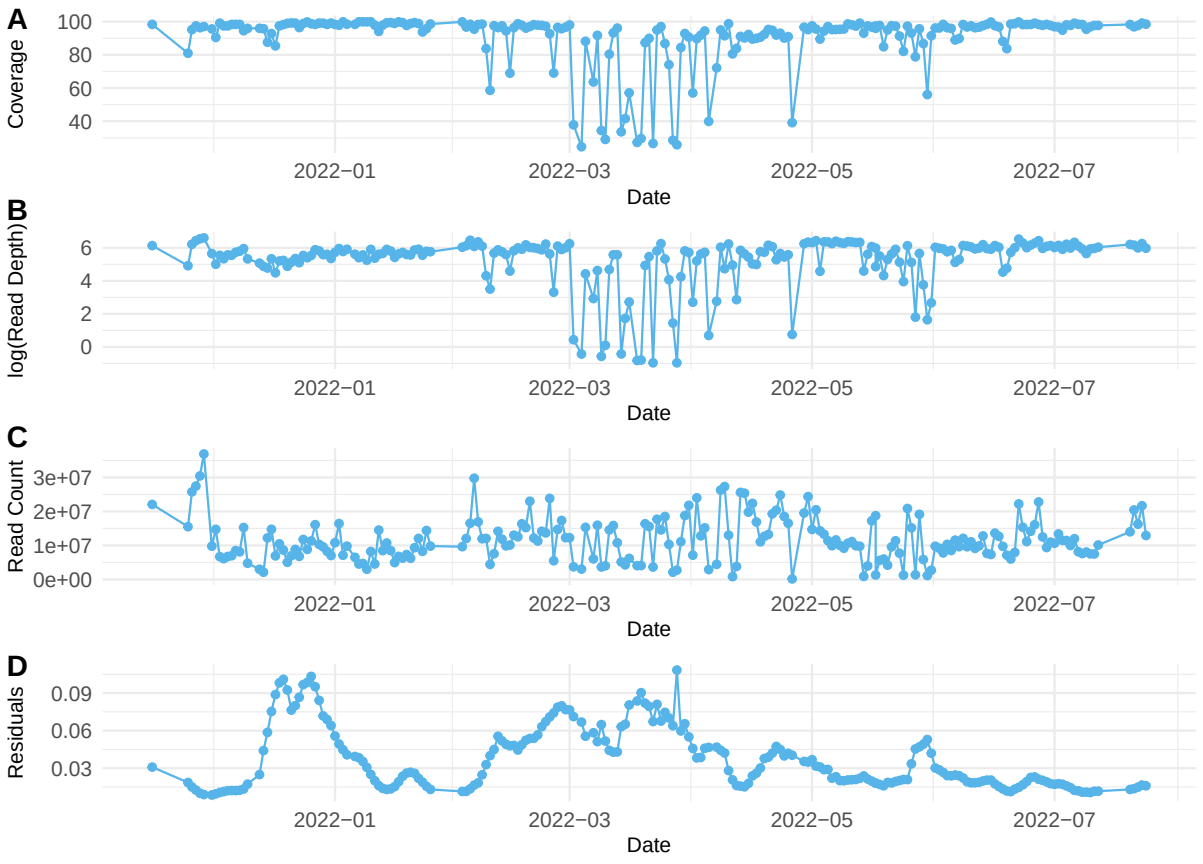
Quality Control and Correlation between Residuals from abundance and sequencing parameters: Altenrhein

Coverage: The percentage of the genome sequenced at a certain depth

Read Depth: The average number of reads per base



Quality Control and Correlation between Residuals from abundance and sequencing parameters: Geneva



Quality Control and Correlation between Residuals from abundance and sequencing parameters: Zurich