Unsupervised detection and fitness estimation of emerging SARS-CoV-2 variants, application to wastewater samples (ANRS0160)

Alexandra Lefebvre,
LJLL (Sorbonne Université, Paris) & CIRB (Collège de France, Paris)

with Y. Maday (LJLL, Paris), A. Lambert (IBENS & CIRB, Paris), V. Maréchal (CRSA, Paris), A. Gloagen (CNRGH, Evry)

Statistical Methods for PostGenomic Data (SMPGD)
29 & 30 Jan. 2026 - Grenoble, France.

CONTEXT

The Covid-19 pandemics and repeated waves of emerging variants:

- Revealed the importance of the early detection of emerging variants for rapid adaptation to viral evolution
- Provided an enormous amount of longitudinal genomic data (RNA and protein sequencing, …) → boosted development of time-series analyses applied to viral evolution
- Boosted wastewater analyses as complementary to clinical sample analyses

Wastewater analyses as a complementary approach to clinical sample analyses

- Population based versus clinical based
- Cost effective and non invasive
- Noisy data : Pooled samples containing fragmented, incomplete sequences from multiples strains versus 1 sample = 1 viral sequence

STATE OF THE ART IN UNSUPERVISED DETECTION AND/OR FITNESS ESTIMATION

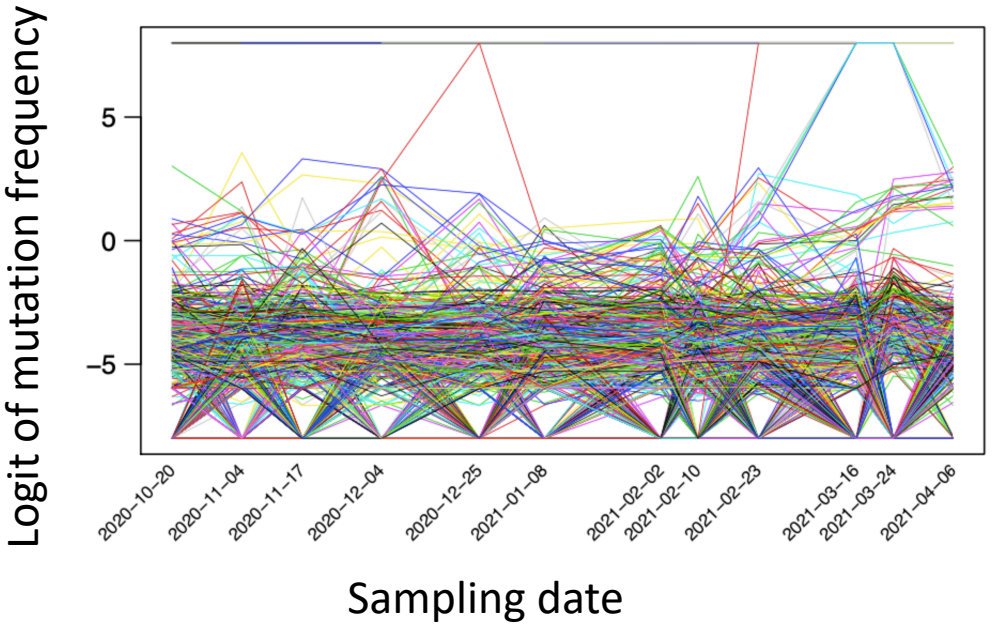
- **Analyze frequency trajectory of sequences through time**
  - ➡ **prior** clustering of sequences
  - ➡ Unsuited to pooled data

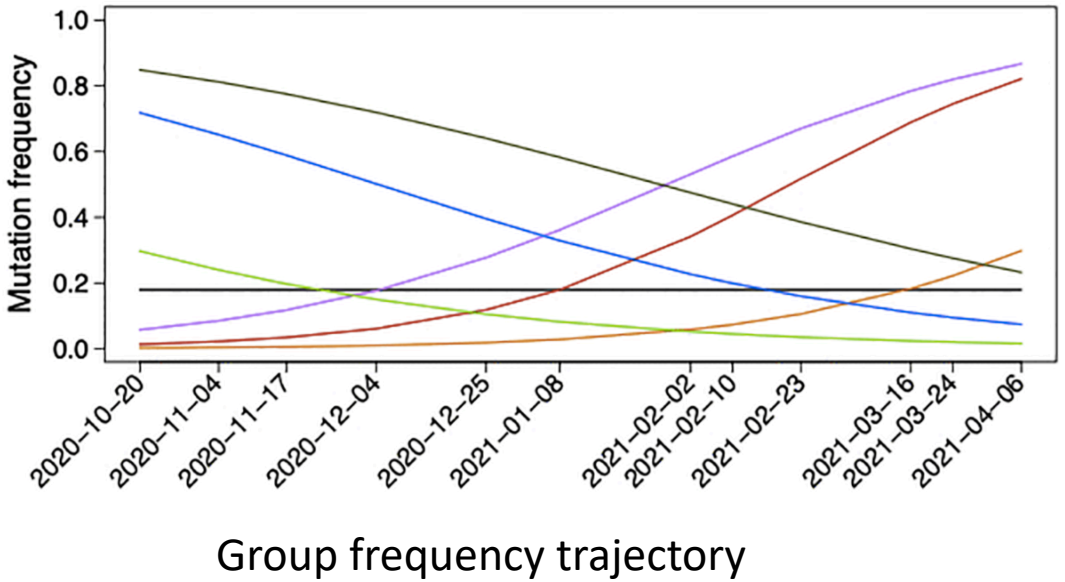
- **Analyze frequency trajectory of mutations through time**
  - ➡ **prior** clustering of mutations (ex. weighted mutation network, latent population genetic structure)
  - ➡ non parametric

- **Idea within the analyses of frequency trajectory of mutations**
  - ➡ **clustering** of mutations according to their **frequency trajectory** itself
  - ➡ suited to pooled data and parametric

GOAL

n trajectories of n mutation frequency ⟶ $K << n$ groups of mutations of similar trajectories





Source: Ifremer, Nantes.

Laure Barbé et al (2022): *SARS-CoV-2 Whole-Genome Sequencing Using Oxford Nanopore Technology for Variant Monitoring in Wastewaters,* (Frontiers in Microbiology, 2022)

THE MODEL

Mixture of

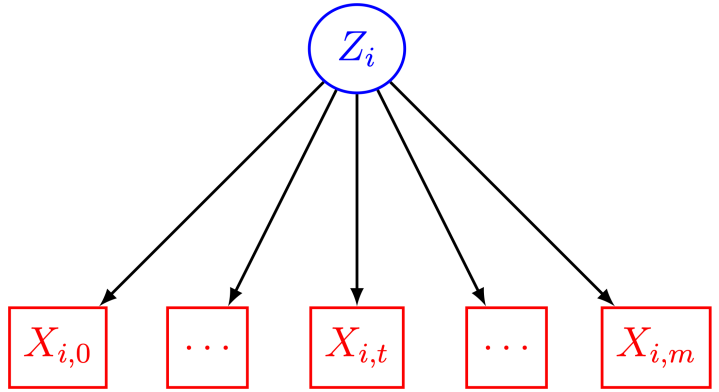
- latent group assignment $Z_i \in \{0,\ldots,K\}$: K groups under selection + 1 neutral group (fixed constant trajectory)

$$\{Z_i\} \sim \text{Multinomial}\,(\pi)$$

- observed mutation counts $X_i = (X_{it})_{t=1,\ldots,m}$

$$\{X_{it} \,|\, Z_i = k, k \neq 0\} \sim \text{Binomial}\left(d_{it}, \frac{1}{1+e^{-(\mu_k + s_k t)}}\right)$$

d_{it} : read depth at time t at position of mutation i



- The neutral group $\{i, Z_i = 0\}$ of constant trajectory (fixed selection coefficient at 0) has a random intercept :

$$\{X_{it} \,|\, Z_i = 0\} \sim \text{Binomial}\left(d_{it}, \frac{1}{1+e^{-u}}\right) \quad \text{whith} \quad u \sim \text{Beta}\,(\alpha, \beta)$$

STATISTICAL TOOLS based on the EM algorithm

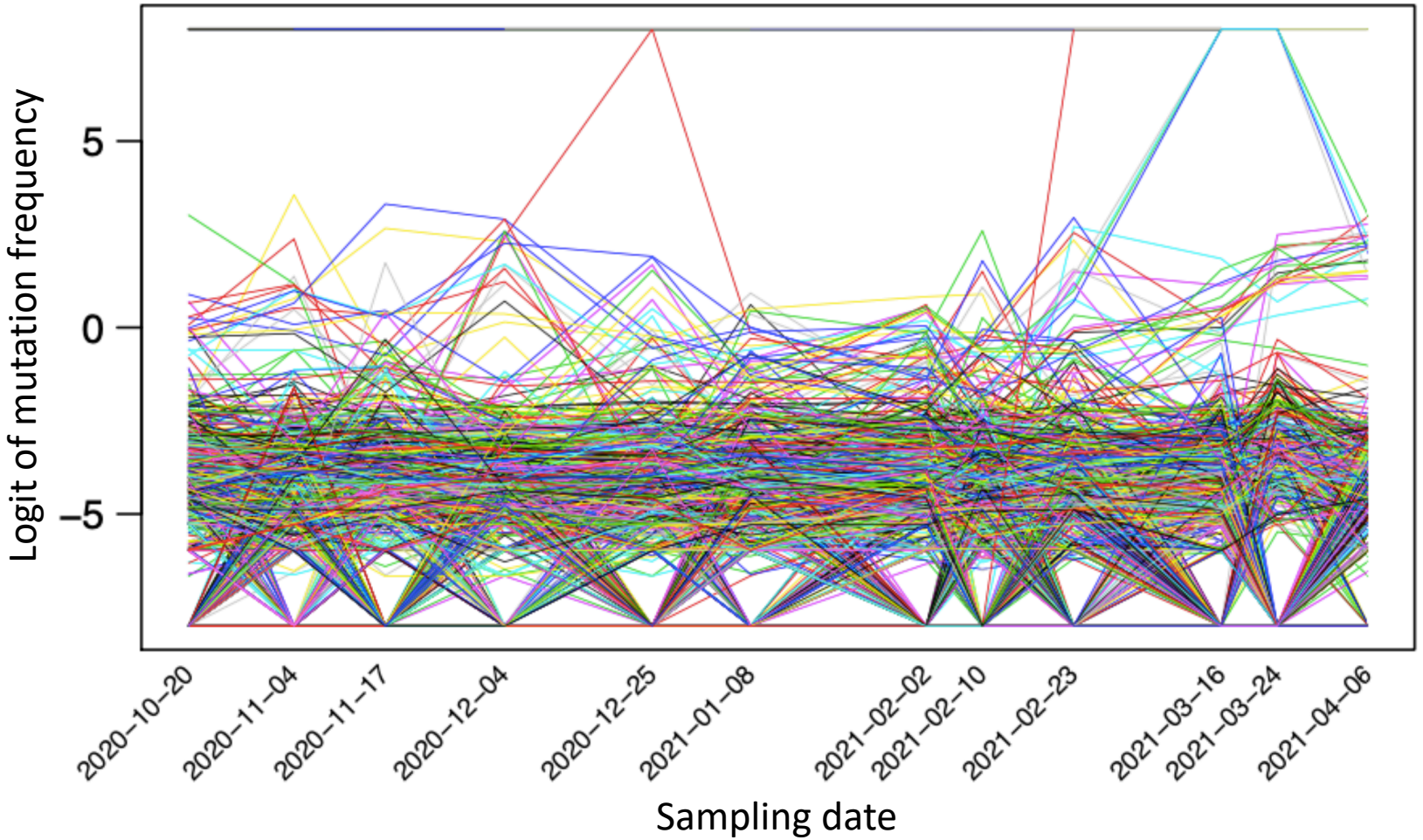
$$\mathscr{P}(X,Z) = \prod_i \mathscr{P}(Z_i) \prod_t \mathscr{P}(X_{it} \mid Z_i)$$

- **Parameter estimation with the Expectation Maximization (EM) algorithm**
  - Initialization with data driven initial posterior probabilities of group assignment computed with EM algorithm over a model with no neutral groups

$$\mathscr{Q}(\theta \mid \theta^{\text{old}}) = \sum_i \sum_k \underbrace{\mathscr{P}(Z_i = k \mid X_i = x_i; \theta)}_{\text{Expectation step}} \underbrace{\log \mathscr{P}(Z_i = k, X_i = x_i \mid \theta^{\text{old}})}_{\text{Maximization step}}$$

- **Selection of the number of groups**:
  - based on lower limit of the size of the smallest group N_k (in terms of number of mutations), $K = \arg\max_k \{N_K \geq 2\}$ (resp. 3 & 5) for analyses covering days (resp. weeks & months).
  - No group contains solely 1 mutation.

RESULTS - DATASETS

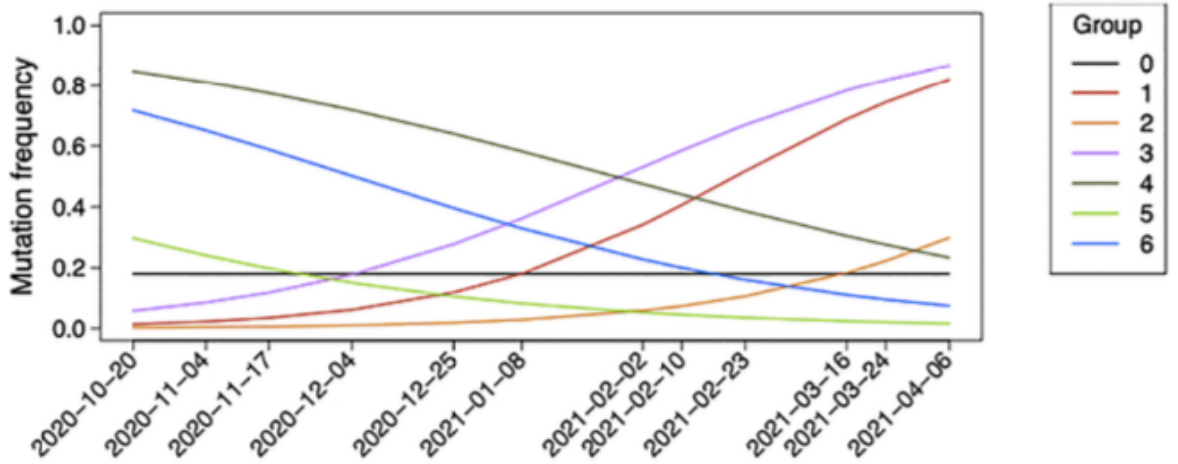


Source: Ifremer, Nantes.

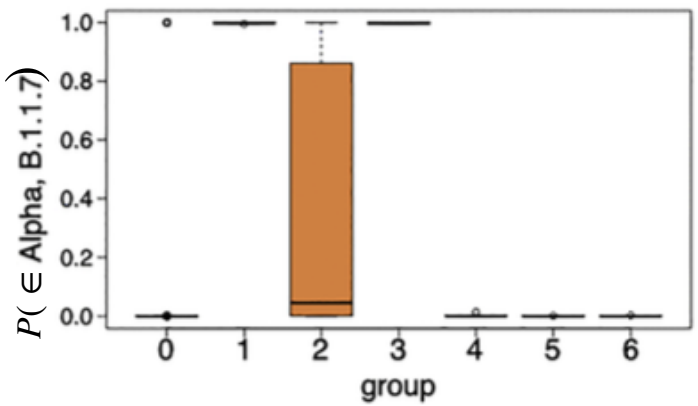
Laure Barbé et al (2022): *SARS-CoV-2 Whole-Genome Sequencing Using Oxford Nanopore Technology for Variant Monitoring in Wastewaters,* (Frontiers in Microbiology, 2022)

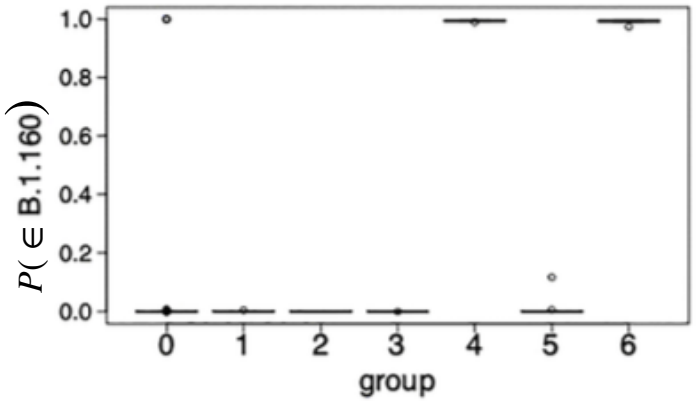
RESULTS - WWTP1 over its entire time period

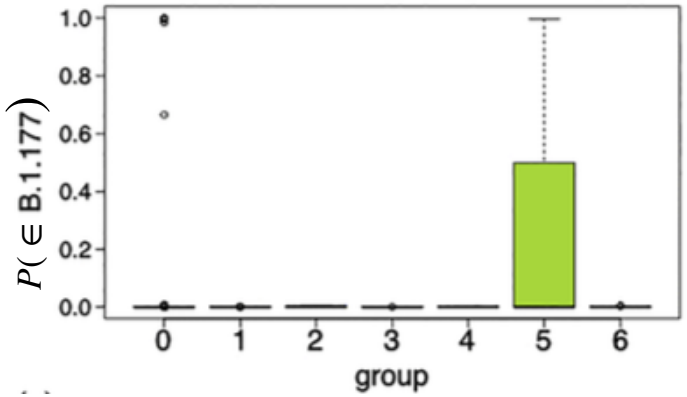
	Group						
	0	1	2	3	4	5	6
π	0.67	0.08	0.04	0.04	0.05	0.07	0.05
μ	-1.52	-4.28	-5.97	-2.79	1.72	-0.86	0.93
[95%CI]	-	[-4.33; -4.24]	[-6.04; -5.89]	[-2.82; -2.75]	[1.68; 1.76]	[-0.91; -0.81]	[0.89; 0.97]
$s \times 100$	0.00	3.46	3.04	2.77	-1.73	-1.94	-2.06
[95%CI]	-	[3.42; 3.49]	[2.99; 3.10]	[2.74; 2.81]	[-1.78; -1.69]	[-2.02; -1.86]	[-2.11; -2.00]



Parameter estimates and estimated frequency trajectories



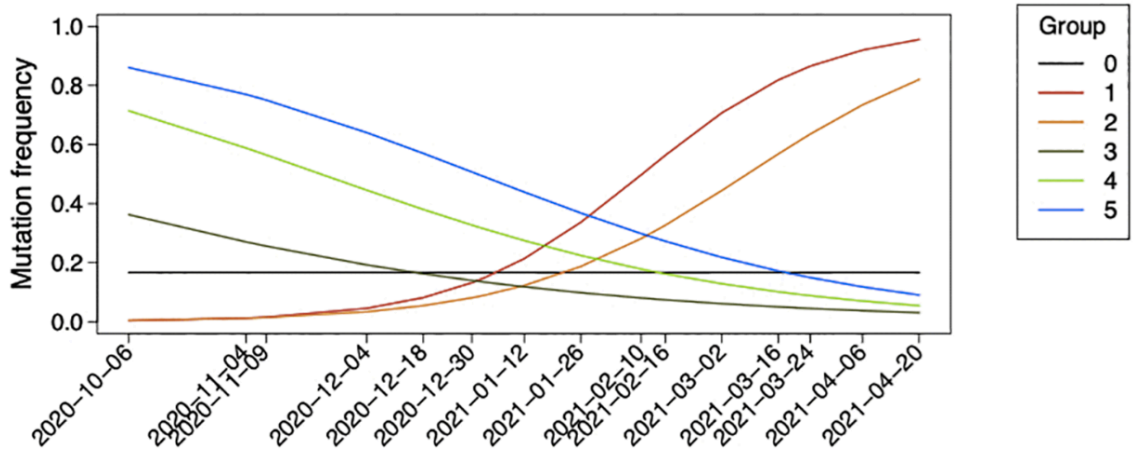




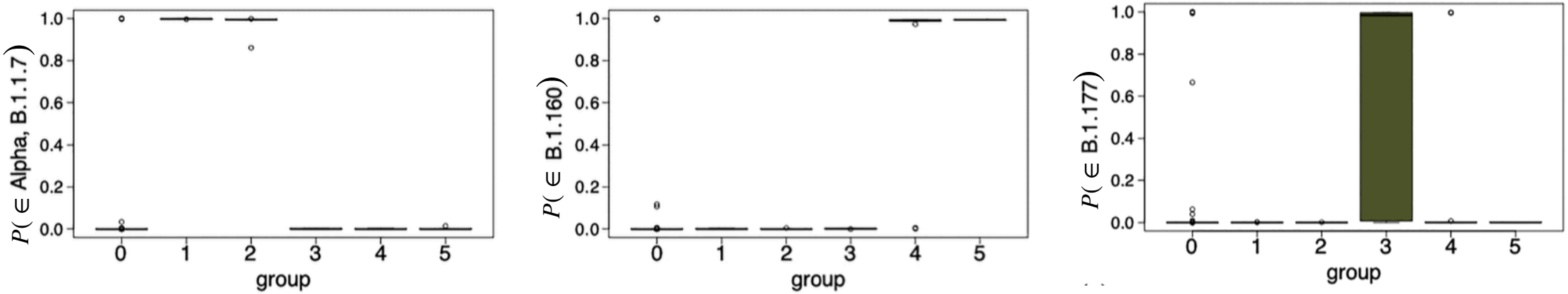
Probability to belong to main circulating variants computed with VirPool package (retrospectively) stratified on posterior group assignment

RESULTS - WWTP2 over its entire time period

	Group					
	0	1	2	3	4	5
π	0.69	0.11	0.04	0.04	0.08	0.03
μ	-1.61	-5.67	-5.45	-0.56	0.92	1.82
[95%CI]	-	[-5.70; -5.64]	[-5.49; -5.41]	[-0.61; -0.51]	[0.89; 0.94]	[1.78; 1.86]
$s \times 100$	0.00	4.46	3.56	-1.47	-1.92	-2.10
[95%CI]	-	[4.43; 4.48]	[3.53; 3.59]	[-1.53; -1.41]	[-1.95; -1.89]	[-2.14; -2.07]



Parameter estimates and estimated frequency trajectories

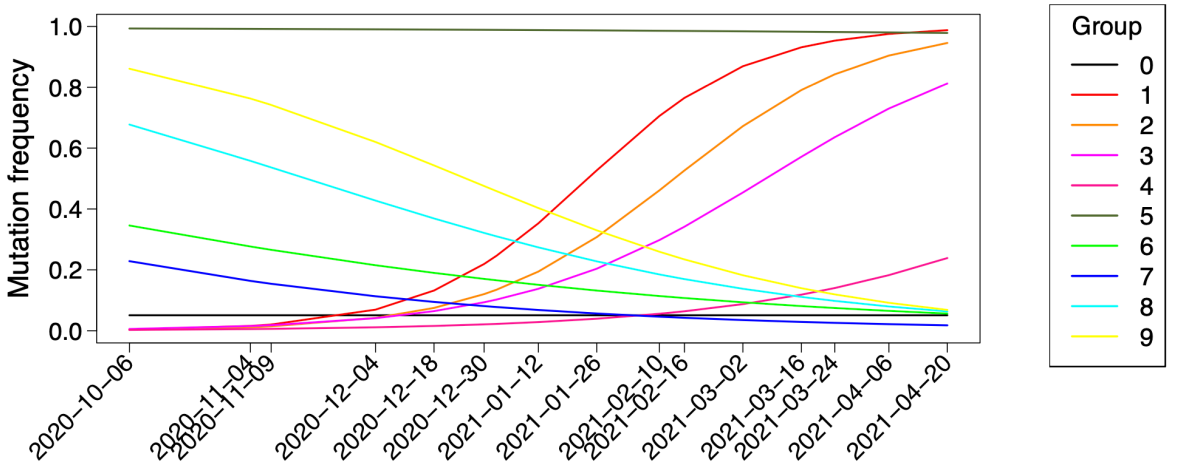


Probability to belong to main circulating variants (computed with VirPool package) stratified on posterior group assignment

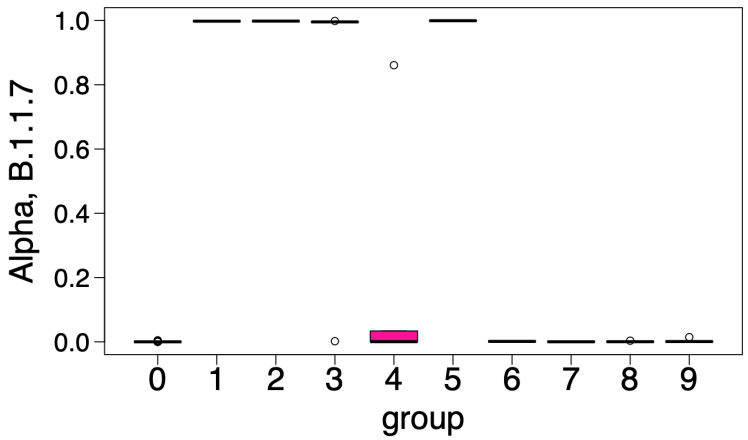
RESULTS - WWTP2 lowering group size limit from 5 to 3 mutations

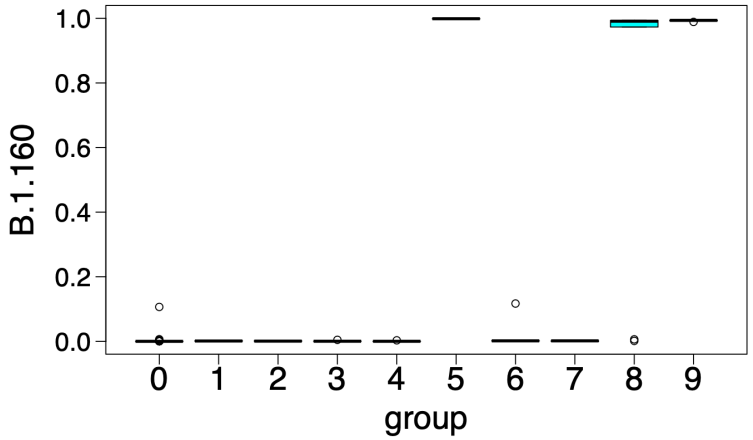
	Group					
	0	1	2	3	4	5
π	0.61	0.03	0.08	0.04	0.03	0.03
μ	-2.93	-5.61	-5.70	-5.13	-5.90	5.00
[95%CI]	-	[-5.70; -5.52]	[-5.74; -5.67]	[-5.18; -5.09]	[-6.02; -5.79]	[4.83; 5.17]
$s \times 100$	0.00	5.10	4.37	3.37	2.42	-0.60
[95%CI]	-	[5.03; 5.18]	[4.34; 4.39]	[3.34; 3.40]	[2.35; 2.50]	[-0.73; -0.47]

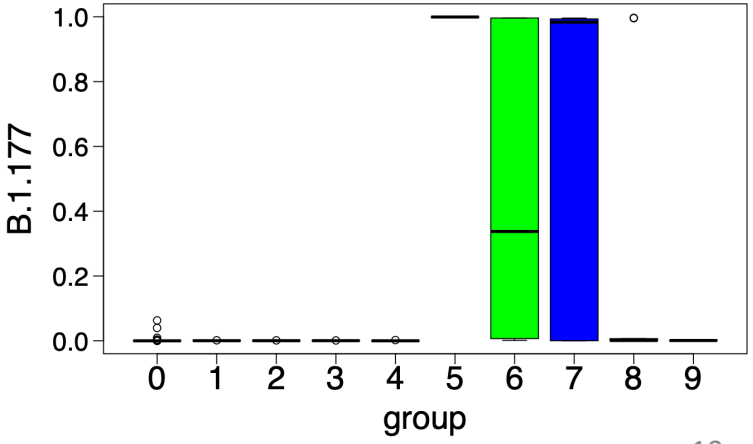
6	7	8	9
0.04	0.03	0.06	0.05
-0.64	-1.22	0.74	1.82
[-0.69; -0.59]	[-1.29; -1.15]	[0.71; 0.77]	[1.79; 1.86]
-1.11	-1.43	-1.75	-2.26
[-1.16; -1.06]	[-1.51; -1.34]	[-1.78; -1.72]	[-2.30; -2.23]



Parameter estimates and estimated frequency trajectories

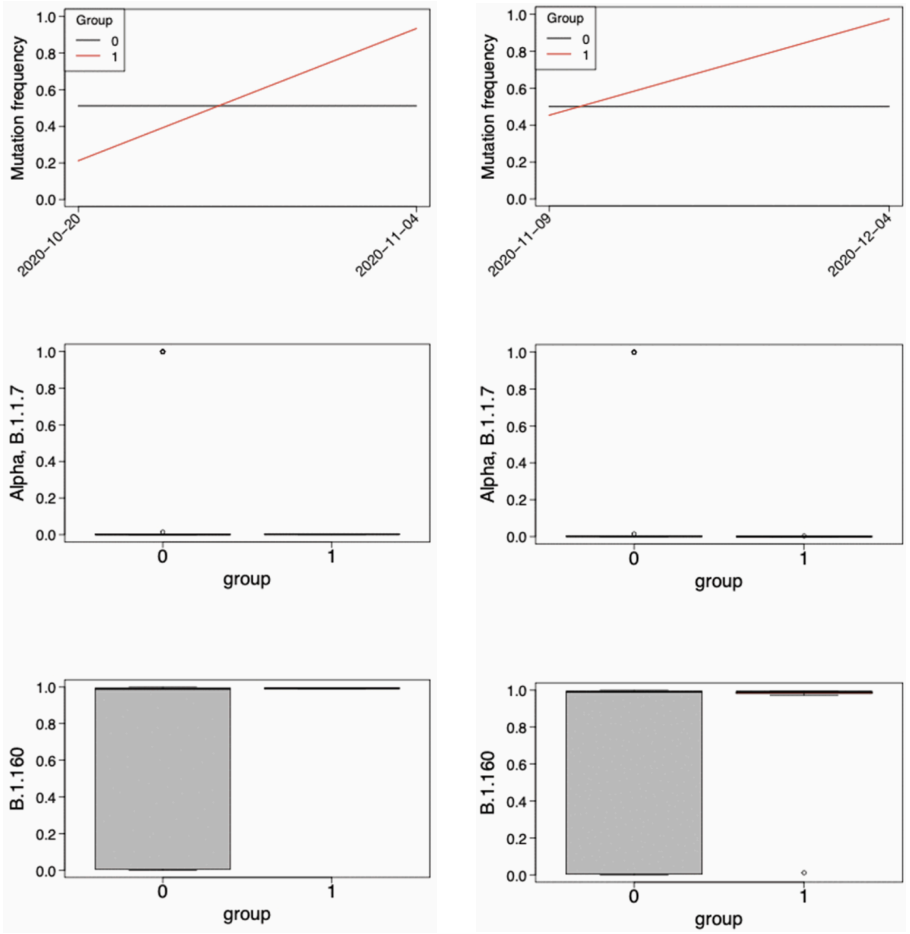






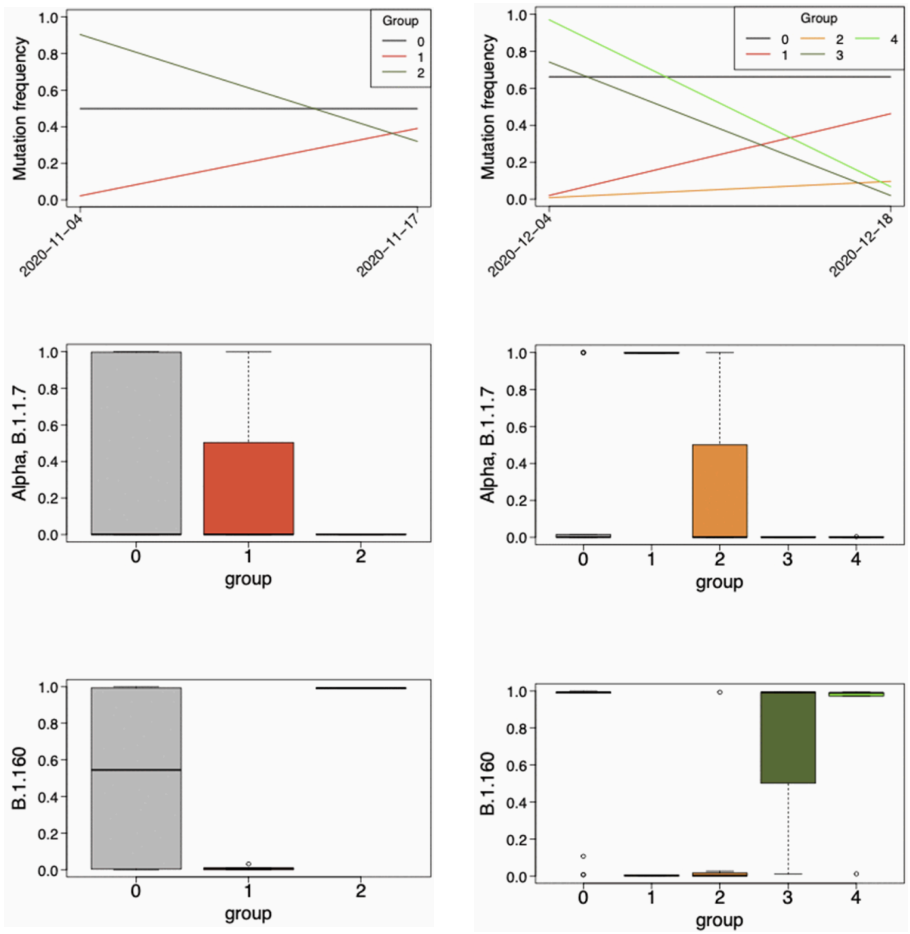
RESULTS - detection (just before and at the beginning of Alpha emergence)

Before Alpha emergence



WWTP1 WWTP2

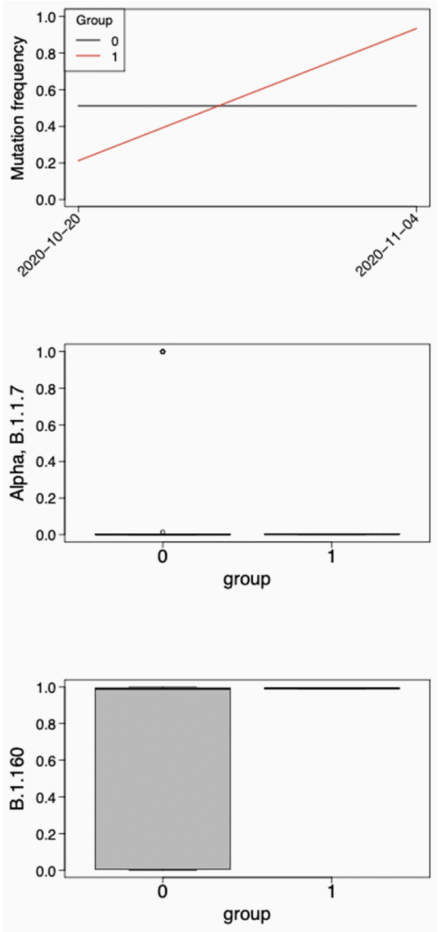
At the beginning of Alpha emergence

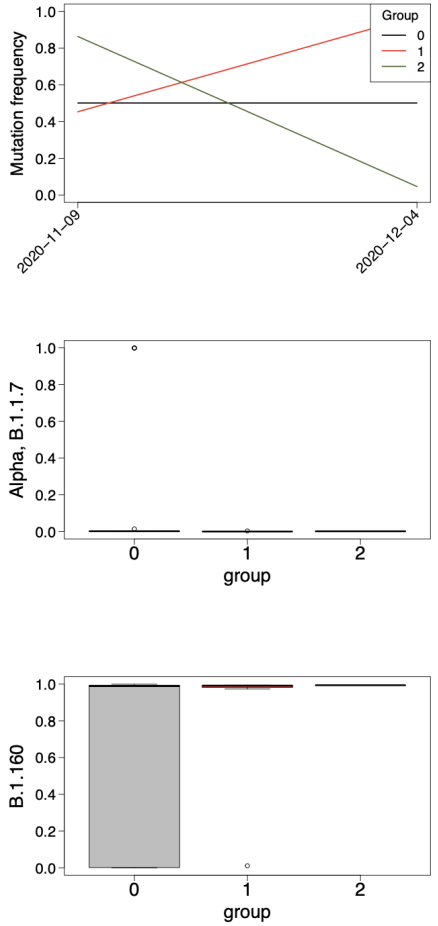


WWTP1 WWTP2

RESULTS - lowering group size limit from 3 to 2 mutations

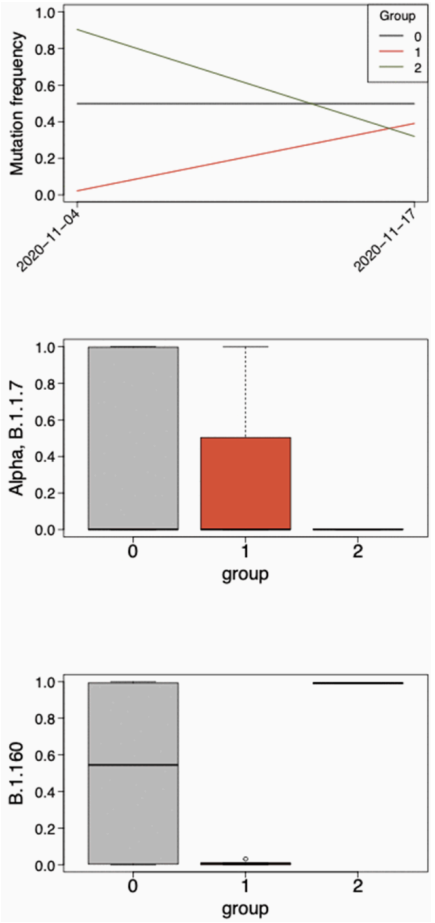
Before Alpha emergence

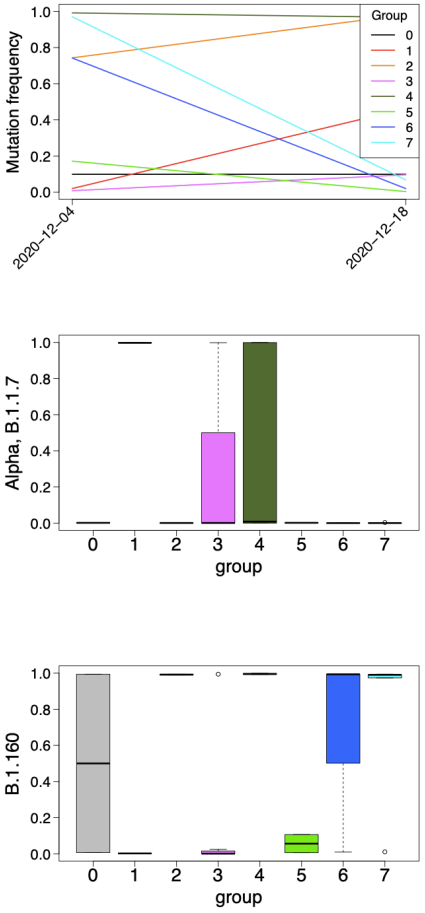




WWTP1 WWTP2

At the beginning of Alpha emergence





WWTP1 WWTP2

PERSPECTIVES

- Leverage general conditions of application (number of groups selection, thresholds, etc.)
- Hidden Gaussian random walk with break point detection

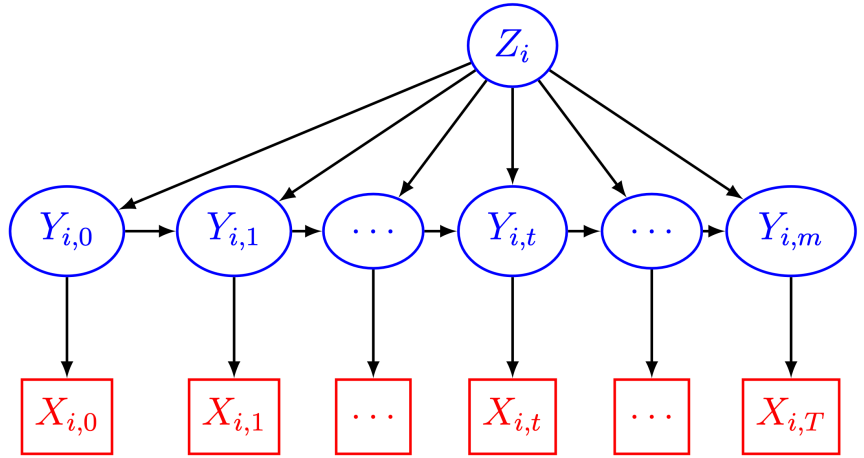
$$\mathscr{P}(X, Y, Z \mid \theta) = \prod_{i=1}^{n} \mathscr{P}(Z_i \mid \pi)\mathscr{P}(Y_{i,0} \mid Z_i; \mu_0, \sigma_0^2) \prod_{t=1}^{m} \mathscr{P}(Y_{i,t} \mid Y_{i,t-1}, Z_i; s, \sigma^2) \prod_{t=0}^{m} \mathscr{P}(X_{i,t} \mid Y_{i,t})$$

- latent group assignment $Z_i \in \{0, \ldots, K\}$
- latent Gaussian random walk $Y_i = (Y_{it})_{t=0,\ldots,m}$
- observed mutation count $X_i = (X_{it})_{t=0,\ldots,m}$

$$\{Z_i\} \sim \text{Multinomial}\,(\pi)$$

$$\{Y_{i,0} \mid Z_i = k\} \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right); \quad \{Y_{i,t} \mid Y_{i,t-1} = y, Z_i = k\} \sim \mathcal{N}\left(y + s_k \Delta_j, \sigma^2 \Delta_j\right)$$

$$\{X_{i,j} \mid Y_{i,j} = y\} \sim \text{Binomial}\left(d_{i,j}, \frac{1}{1 - e^{-y}}\right)$$



This work is published in PLOS Computational Biology under DOI https://doi.org/10.1371/journal.pcbi.1013749

The code is available at the link : https://github.com/AlexandraLefe/FT-mixture

THANK YOU !