Analysis of SARS-CoV-2 Spike Protein Mutations with Logistic Regression

Simón Rodríguez Santana, Roi Naveiro, Daniel García Rasines, Paula Ruiz-Rodriguez, Miguel Álvarez-Herrera, David Ríos Insua, Nuria E. Campillo, Eugenia Ulzurrun, Mireia Coscollá

i2sysbio
CNB CENTRO NACIONAL DE BIOTECNOLOGIA
cib Margarita Salas
ICMAT INSTITUTO DE CIENCIAS MATEMÁTICAS

Objective

Which mutations (individually or by pairs) of the COVID-19 genome are associated to important aspects of the infection?

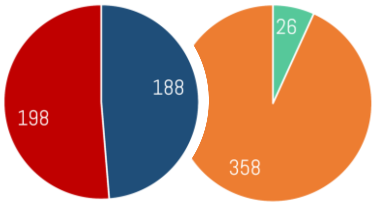
○ Severity - **Hospitalization** (possibly death)

○ Vaccine failure - **Breakthrough** (full vacc. + hosp.)

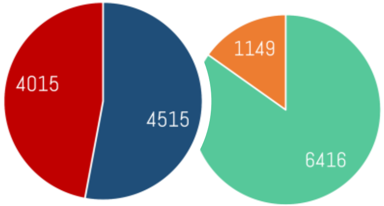
Data

- Data sources: FISABIO (8.534) + GM hospital (386)
- Covariates: *sex*, *age*, *sample month* and *genomic sequences* (AA)
  - Hospitalization study: vaccination status as covariate

GM: 04/2021 - 04/2022

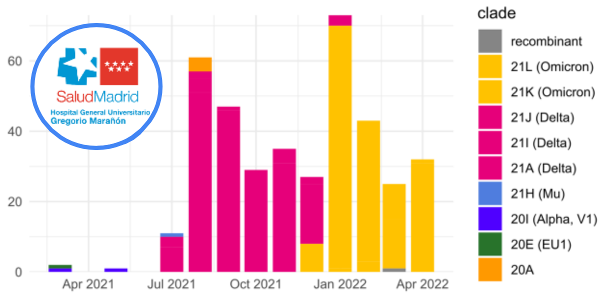


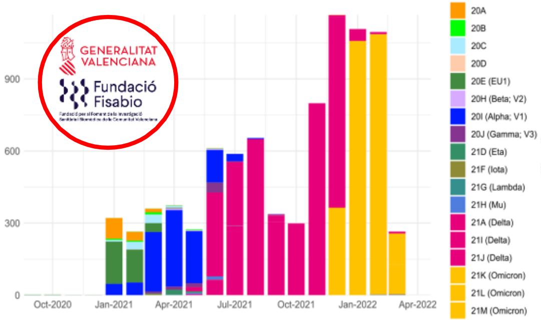
FISABIO: 09/2020 - 05/2022

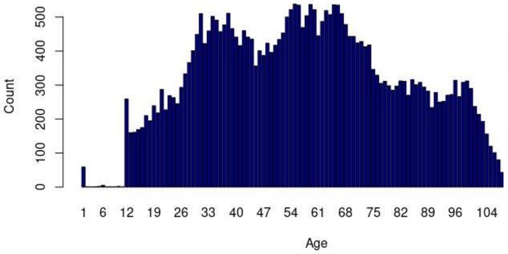


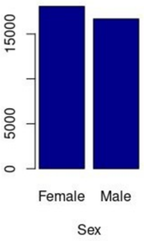
🟦 False
🟥 True
Vaccination status

🟩 False
🟧 True
Hospitalization status









Preprocessing

- Clean the dataset:
  - Remove rows with >10% of missing values
  - Patients with partial information samples
  - Samples before 01/01/2021
  - Genome positions without mutations (at least >1 type of AA)

- Full preprocessing only for Spike protein

→ 331 Spike genome positions (out of 1.272)

→ 5.928 cases (out of 8.920)

Preprocessing

- Clean the dataset:
  - Remove rows with >10% of missing values
  - Patients with partial information samples
  - Samples before 01/01/2021
  - Genome positions without mutations (at least >1 type of AA)

- Full preprocessing only for Spike protein

  → 331 Spike genome positions (out of 1.272)

  → 5.928 cases (out of 8.920)

- Preserve more data with imputation? Possible changes in the preprocessing?

Model

- Logistic regression with Hierarchical Group Lasso regularization

$$\text{logit}[P(Y = 1|\mathbf{X})] = \beta_0 + \sum_{i=1}^{p} X_i \beta_i + \sum_{i<j} X_{i:j} \beta_{i:j}$$

Model

- Logistic regression with Hierarchical Group Lasso regularization

$$\text{logit}[P(Y = 1|\mathbf{X})] = \beta_0 + \sum_{i=1}^{p} X_i\beta_i + \sum_{i<j} X_{i:j}\beta_{i:j}$$

$$\text{argmin}_\beta\, \mathcal{L}(\mathbf{Y}, \mathbf{X}, \beta) + \lambda \sum_{i=1}^{p} \gamma_i||\beta_i||_2$$

  - Negative log-likelihood loss function
  - *L1* reg. + *k*-fold CV regularization strength

Model

- Logistic regression with Hierarchical Group Lasso regularization

$$\text{logit}[P(Y = 1|\mathbf{X})] = \beta_0 + \sum_{i=1}^{p} X_i\beta_i + \sum_{i<j} X_{i:j}\beta_{i:j}$$

$$\text{argmin}_\beta\, \mathcal{L}(\mathbf{Y}, \mathbf{X}, \beta) + \lambda \sum_{i=1}^{p} \gamma_i||\beta_i||_2$$

  - Negative log-likelihood loss function
  - *L1* reg. + *k*-fold CV regularization strength

Model

- Logistic regression with Hierarchical Group Lasso regularization

$$\text{logit}[P(Y = 1|\mathbf{X})] = \beta_0 + \sum_{i=1}^{p} X_i\beta_i + \sum_{i<j} X_{i:j}\beta_{i:j}$$

$$\text{argmin}_{\beta}\, \mathcal{L}(\mathbf{Y}, \mathbf{X}, \beta) + \lambda \sum_{i=1}^{p} \gamma_i ||\beta_i||_2$$

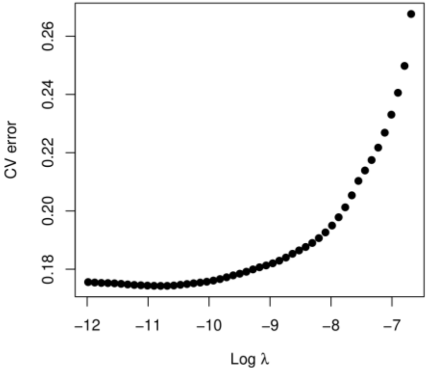
  - Negative log-likelihood loss function
  - *L1* reg. + *k*-fold CV regularization strength

Model

- Logistic regression with Hierarchical Group Lasso regularization

$$\text{logit}[P(Y=1|\mathbf{X})] = \beta_0 + \sum_{i=1}^{p} X_i\beta_i + \sum_{i<j} X_{i:j}\beta_{i:j}$$

$$\text{argmin}_{\beta}\, \mathcal{L}(\mathbf{Y}, \mathbf{X}, \beta) + \lambda \sum_{i=1}^{p} \gamma_i ||\beta_i||_2$$



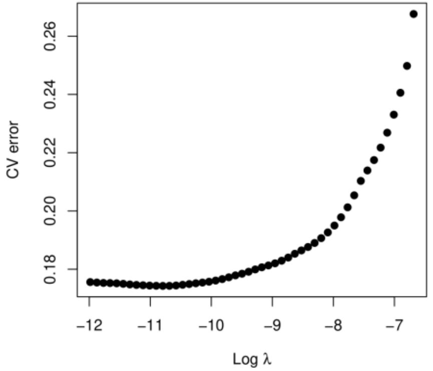
  - Negative log-likelihood loss function
  - *L1* reg. + *k*-fold CV regularization strength

Model

- Logistic regression with Hierarchical Group Lasso regularization

$$\text{logit}[P(Y=1|\mathbf{X})] = \beta_0 + \sum_{i=1}^{p} X_i\beta_i + \sum_{i<j} X_{i:j}\beta_{i:j}$$

$$\text{argmin}_\beta\, \mathcal{L}(\mathbf{Y}, \mathbf{X}, \beta) + \lambda \sum_{i=1}^{p} \gamma_i ||\beta_i||_2$$



  - Negative log-likelihood loss function
  - *L1* reg. + *k*-fold CV regularization strength

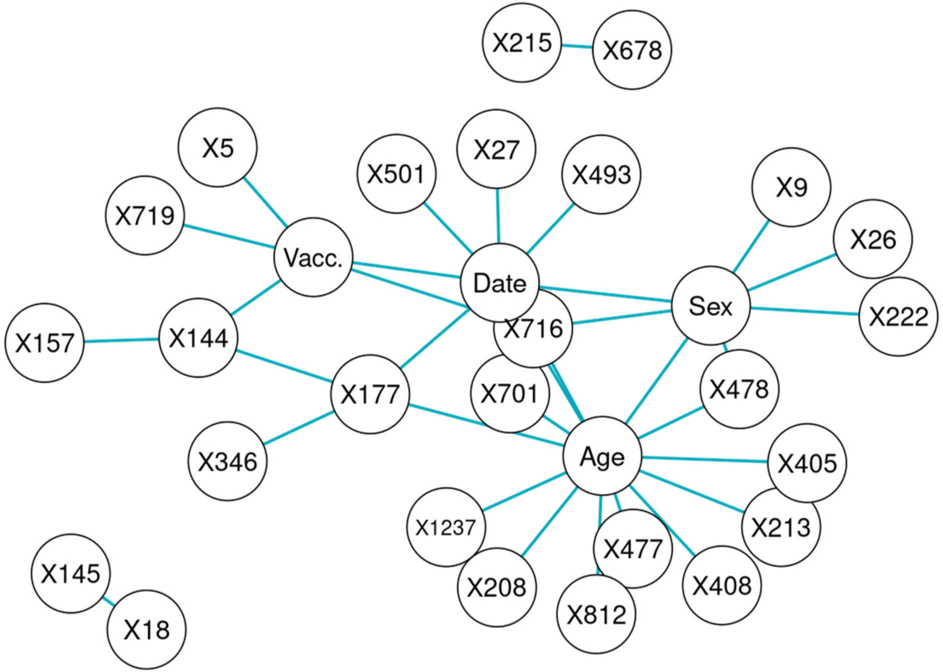
Strong hierarchy:

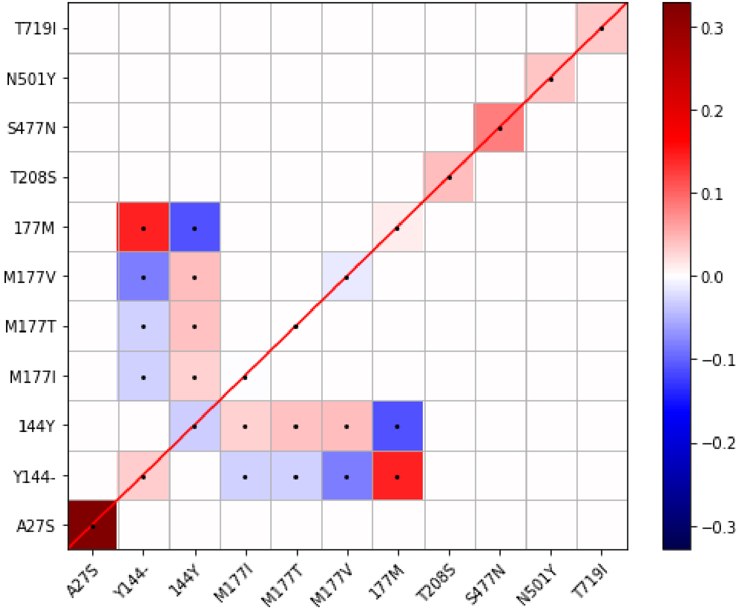
$$\beta_{i:j} \neq 0 \Rightarrow \beta_i \neq 0\,, \beta_j \neq 0$$

Overparametrization:

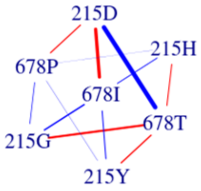
For each position, the sum of its main effects is 0, as well as for its interaction coefficients

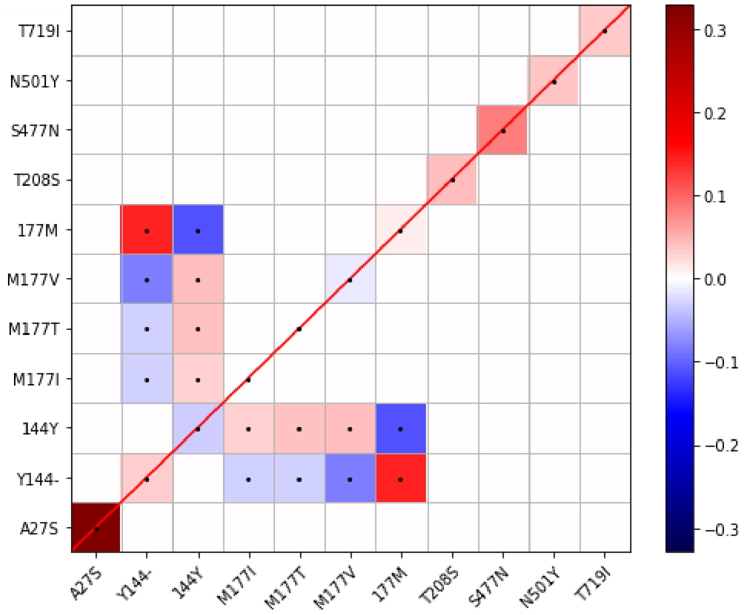
Hospitalization results

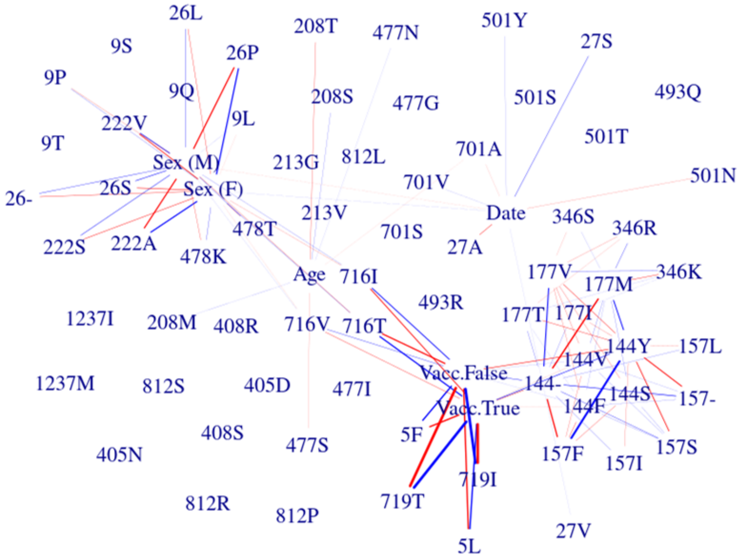


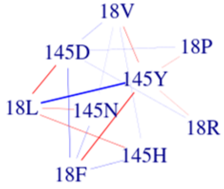


Hospitalization results

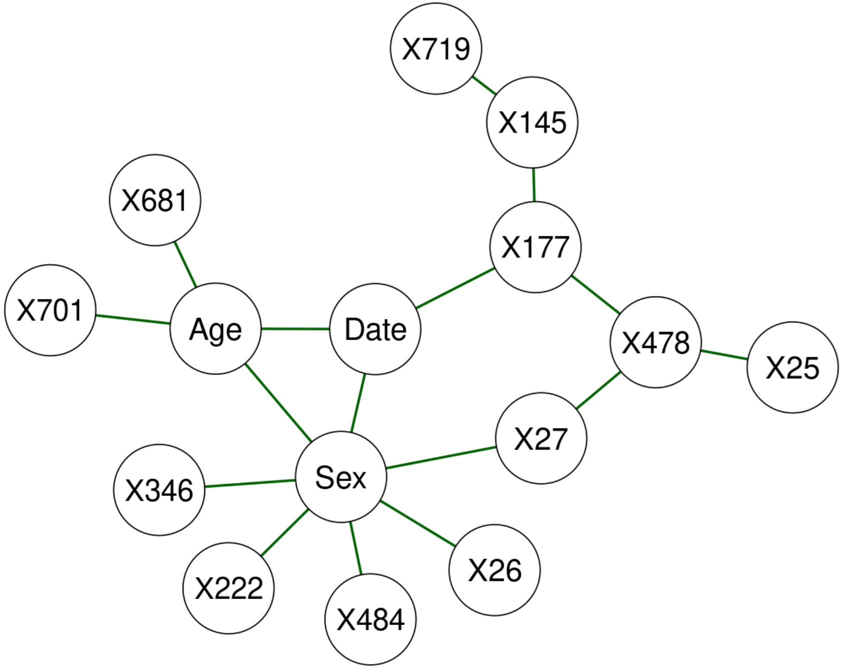


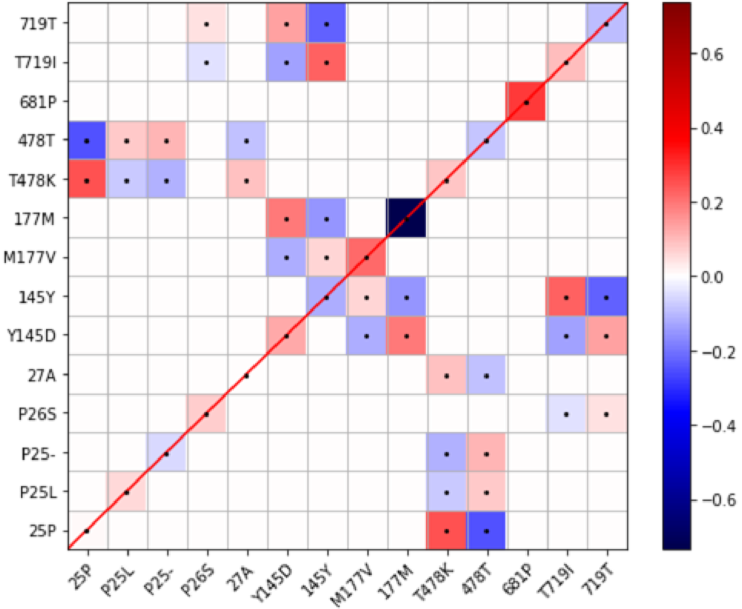




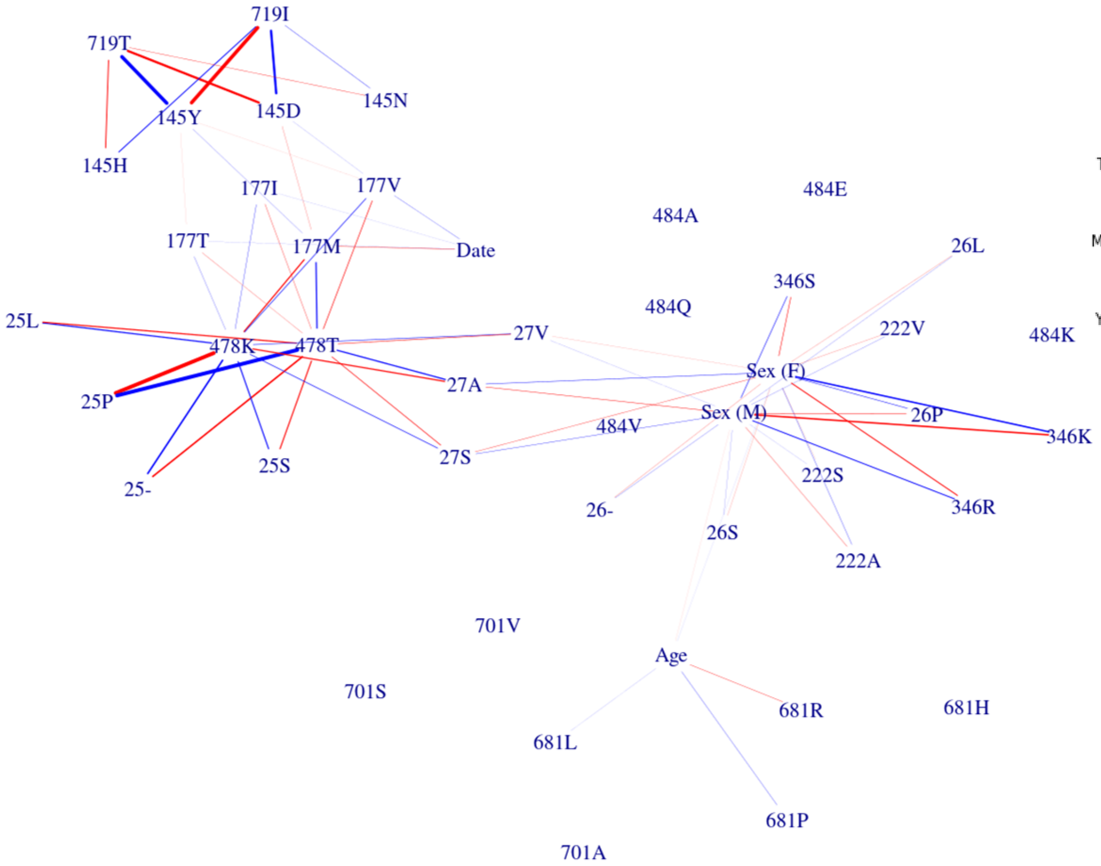


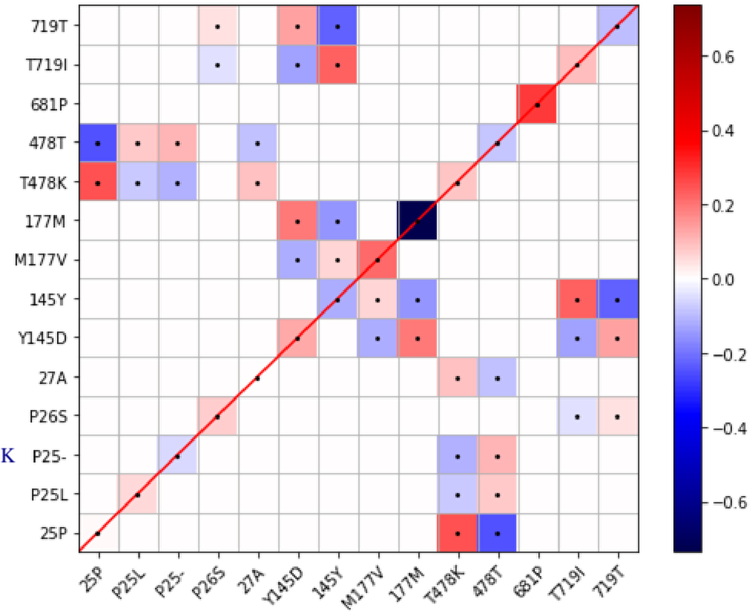
Breakthrough results





Breakthrough results





Conclusions

- Several novel interaction found, some of interest
- Effects of well-known mutations are enhanced or diminished by mutations in other positions
    - Example: **T478K** vs. **478T** in combination with **25P** (hosp.)
- Further analysis:
    - Remaining parts of the genome
    - Characterization of the effects of the preprocessing pipeline
    - Augment with other data sources (available)

Conclusions

- Several novel interaction found, some of interest
- Effects of well-known mutations are enhanced or diminished by mutations in other positions
  - Example: **T478K** vs. **478T** in combination with **25P** (hosp.)
- Further analysis:
  - Remaining parts of the genome
  - Characterization of the effects of the preprocessing pipeline
  - Augment with other data sources (available)

Thanks!

simon.rodriguez@icmat.es