Forecasting and Nowcasting Variant Proportions with Genomic Data at the Regional Level

University of Massachusetts, Amherst

Isaac MacArthur, **Maryclare Griffin**, Evan Ray, Nicholas Reich, Thomas Robacker, and Benjamin Rogers

June 16, 2025

Forecasting and Nowcasting Variant Proportions with Genomic Data at the Regional Level

University of Massachusetts, Amherst

Isaac MacArthur, **Maryclare Griffin**, Evan Ray, Nicholas Reich, *Thomas Robacker*, and Benjamin Rogers

June 16, 2025

Genomic Data and Variant Proportions

For COVID and other viral illnesses, viral samples are

- ▶ Collected from infected individuals
- → Sent to a lab
- → Processed
- → Classified into clades using an estimated phylogenetic tree

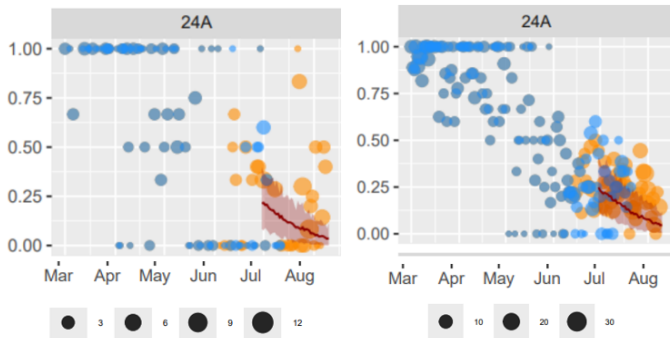Proportion of samples across clades (variant proportions) informs:

- ▶ Mitigation efforts
- ▶ Treatment strategies

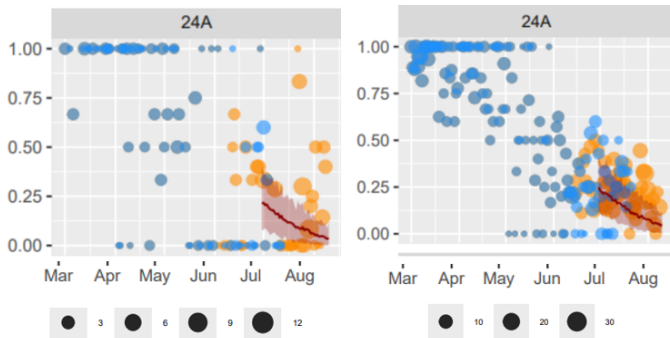There is a need for local (state-level):

- ▶ Forecasts of variant proportions in the near future
- ▶ Nowcasts of variant proportions as samples during processing

---

Bedford Lab provides sample data and estimated phylogenetic trees.

Isaac MacArthur, **Maryclare Griffin**, Evan Ray, Nicholas Reich, *Thomas Robacker*, and Benjamin Rogers

MA and MN Proportion 24A for July 31, 2024

Isaac MacArthur, **Maryclare Griffin**, Evan Ray, Nicholas Reich, *Thomas Robacker*, and Benjamin Rogers
Forecasting and Nowcasting Variant Proportions with Genomic Data at the Regional Level

MA and MN Proportion 24A for July 31, 2024

- ▶ Lags in data collection
- ▶ Low counts per state
- ▶ Poor evaluation data
- ▶ New clades can arise, clade assignments can change*

Isaac MacArthur, **Maryclare Griffin**, Evan Ray, Nicholas Reich, *Thomas Robacker*, and Benjamin Rogers

Previous Approaches to the Problem

Previous approaches support the use of simple models.

Abousamra, Figgins, and Bedford (2024):

▶ Consider country level data

▶ Use local Multinomial Logistic Regression (MLR) models

▶ Measure performance with median absolute error (MAE) relative to retrospective seven day averages

▶ Find MLR models outperforms alternatives
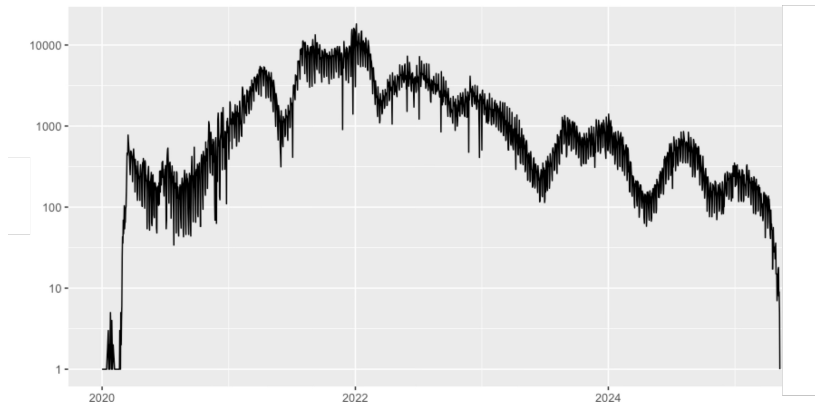
▶ Find a hierarchical MLR models bring benefits

Multinomial Logistic Regression

Let C_{ltv} be the observed count of the v variant at the l location, on the t day, then for $v = 1, \ldots V$ and $l = 1 \ldots L$ and $t = 1 \ldots T$

$$C_{lt1}, \ldots C_{ltV} \mid \alpha_l, \beta_l, n_{lt} \sim$$
$$\text{Multinomial} \left(\frac{\exp\left(\alpha_{lv} + \beta_{lv} t\right)}{\sum_v \exp\left(\alpha_{lv} + \beta_{lv} t\right)}, n_{lt} \right)$$

Isaac MacArthur, **Maryclare Griffin**, Evan Ray, Nicholas Reich, *Thomas Robacker*, and Benjamin Rogers

Need for Sharing Across States

We do not observe many samples!

Some states do not submit viral samples at all.

Isaac MacArthur, **Maryclare Griffin**, Evan Ray, Nicholas Reich, *Thomas Robacker*, and Benjamin Rogers
Forecasting and Nowcasting Variant Proportions with Genomic Data at the Regional Level

Support for Sharing Across States

Variant proportions are similar across states!

Isaac MacArthur, **Maryclare Griffin**, Evan Ray, Nicholas Reich, *Thomas Robacker*, and Benjamin Rogers
Forecasting and Nowcasting Variant Proportions with Genomic Data at the Regional Level
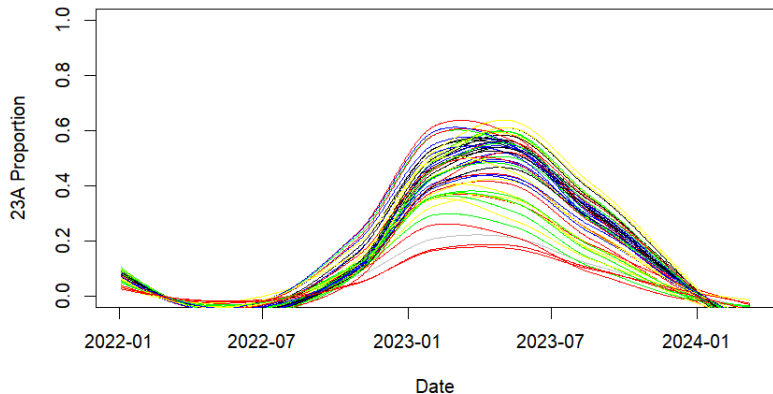
Defining the Hierarchical Structure

As we saw, the MLR model has the form

$$C_{lt1}, \ldots C_{ltV} \mid \alpha_l, \beta_l, n_{lt} \sim$$
$$\text{Multinomial} \left(\frac{\exp\left(\alpha_{lv} + \beta_{lv} t\right)}{\sum_v \exp\left(\alpha_{lv} + \beta_{lv} t\right)}, n_{lt} \right)$$
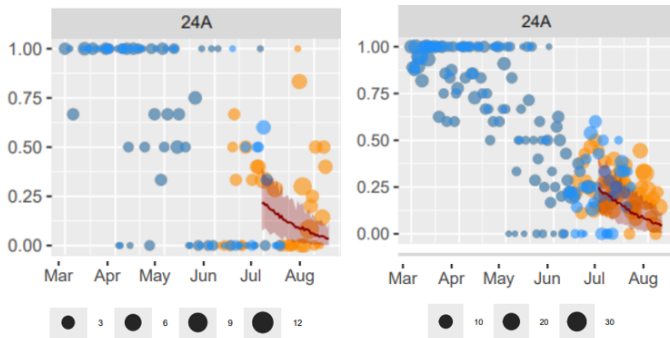
To extend the model we created a hierarchical structure of priors over the α and β. We assumed that for each $v = 1, \ldots, V$

$$\beta_{1v}, \ldots \beta_{Lv} \mid \mu_{\beta v}, \tau_{\beta v}^2 \overset{\text{iid}}{\sim} \text{normal}(\mu_{\beta v}, \tau_{\beta v}^2),$$
$$\alpha_{1v}, \ldots \alpha_{Lv} \mid \mu_{\alpha v}, \tau_{\alpha v}^2 \overset{\text{iid}}{\sim} \text{normal}(\mu_{\alpha v}, \tau_{\alpha v}^2).$$

We then put non-informative priors on the $\mu_\beta, \mu_\alpha, \tau_\beta, \tau_\alpha$.

We also consider a Dirichlet-Multinomial generalization.
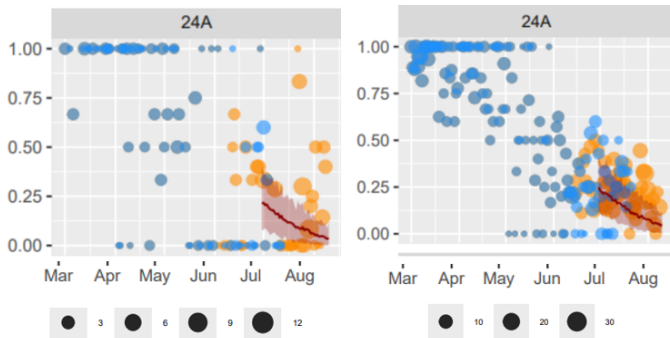
MA and MN Proportion 24A for July 31, 2024

- ▶ Hierarchical model accommodates common trends
- ▶ Dirichlet-Multinomial accommodates overdispersion

Isaac MacArthur, **Maryclare Griffin**, Evan Ray, Nicholas Reich, *Thomas Robacker*, and Benjamin Rogers

Model Fitting and Specifications

▶ Choose a target "nowcast" date

▶ Fix training data to the 150 days prior to "nowcast" date

▶ Use STAN to simulate from the posterior distribution of the parameters

▶ Construct "nowcasts" for "nowcast" date and 30 days prior

▶ Construct forecasts for 10 days after "nowcast" date

Isaac MacArthur, **Maryclare Griffin**, Evan Ray, Nicholas Reich, *Thomas Robacker*, and Benjamin Rogers

MA and MN Proportion 24A for July 31, 2024

Isaac MacArthur, **Maryclare Griffin**, Evan Ray, Nicholas Reich, *Thomas Robacker*, and Benjamin Rogers
Forecasting and Nowcasting Variant Proportions with Genomic Data at the Regional Level

Evaluating the Model

- ▶ Abousamra et. al (2024) used MAE relative to seven day retrospective average
    - ▶ Does not reflect dependence of proportions
    - ▶ Focuses on quality of point prediction
    - ▶ Requires definition of a notion of retrospect "truth"
- ▶ We use approximate energy scores

The Energy Score

▶ The energy score is a proper scoring rule for comparing how similar two distributions are that was introduced in Gneiting et al. (2008)

▶ If we let $\mathbf{y} = (y_1, y_2, \ldots, y_V)$ be the observed values for a given location and time and $\mathbf{x}_1, \ldots \mathbf{x}_m$ be the forecasted values for the forecasted distribution F, then the energy score is then computed as:

$$\text{ES}(F, \mathbf{y}) = \frac{1}{m} \sum_{i=1}^{m} ||\mathbf{y} - \mathbf{x}_i|| - \frac{1}{2m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} ||\mathbf{x}_i - \mathbf{x}_j||$$
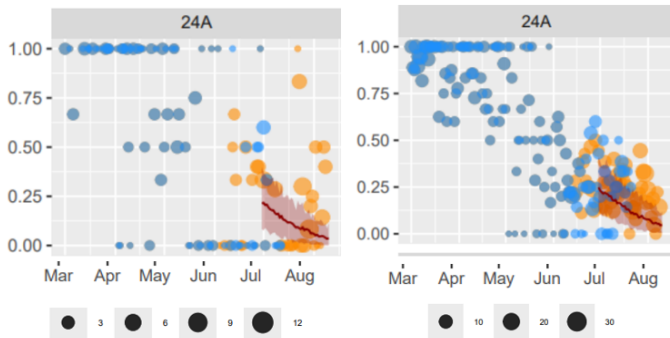
where $||.||$ is the Euclidean norm on R^V, and m is the total number of samples from the predictive distribution.

Isaac MacArthur, **Maryclare Griffin**, Evan Ray, Nicholas Reich, *Thomas Robacker*, and Benjamin Rogers
Forecasting and Nowcasting Variant Proportions with Genomic Data at the Regional Level

Computing the Energy Score

▶ From our model, we obtain posterior distributions of variant proportions on the forecast and nowcast dates

▶ 90 days after our target "nowcast" date, we have observed the number of samples obtained for each forecast and nowcast date

▶ We simulate forecasts and nowcasts of counts by drawing multinomial random variables with the observed totals and posterior predictive variant proportions

$$\mathrm{ES}(F, \mathbf{y}) = \frac{1}{m} \sum_{i=1}^{m} ||\mathbf{y} - \mathbf{x}_i|| - \frac{1}{2m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} ||\mathbf{x}_i - \mathbf{x}_j||$$

Isaac MacArthur, **Maryclare Griffin**, Evan Ray, Nicholas Reich, *Thomas Robacker*, and Benjamin Rogers

MA and MN Proportion 24A for July 31, 2024

Isaac MacArthur, **Maryclare Griffin**, Evan Ray, Nicholas Reich, *Thomas Robacker*, and Benjamin Rogers
Forecasting and Nowcasting Variant Proportions with Genomic Data at the Regional Level

Model Comparison

▶ The energy score measures both the accuracy and variance of the forecasted distribution and a higher score is worse, but the energy score is hard to interpret.

▶ Thus to evaluate the performance of the HMLR model, we want to choose a baseline model that we can use as a comparison.

▶ We have chosen a baseline MLR model that predicts at the country level and makes the same prediction at each state.

Isaac MacArthur, **Maryclare Griffin**, Evan Ray, Nicholas Reich, *Thomas Robacker*, and Benjamin Rogers

Model Testing

▶ To test the HMLR model, we accrued historical versions of Nextstrain datasets for Mondays.

▶ These datasets run from August 2022 to August 2024

▶ Each Monday dataset was used to predict the following Wednesday, for a total for 106 predictions.

▶ Each forecast was scored using the data 91 days after the dataset used.

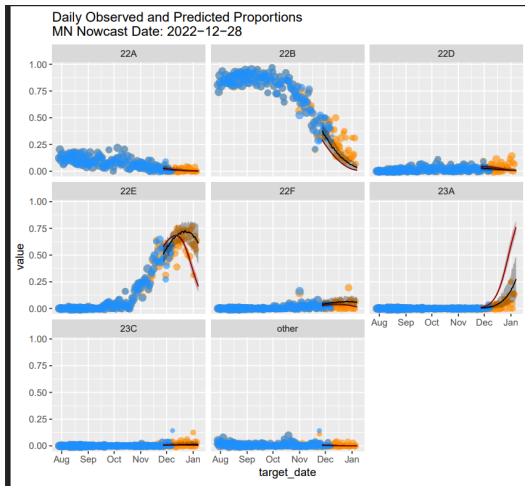*Isaac MacArthur*, **Maryclare Griffin**, Evan Ray, Nicholas Reich, *Thomas Robacker*, and Benjamin Rogers

Example Forecasts

Figure: Mean energy scores: HMLR = 4.37, Baseline = 9.45

Isaac MacArthur, **Maryclare Griffin**, Evan Ray, Nicholas Reich, *Thomas Robacker*, and Benjamin Rogers
Forecasting and Nowcasting Variant Proportions with Genomic Data at the Regional Level
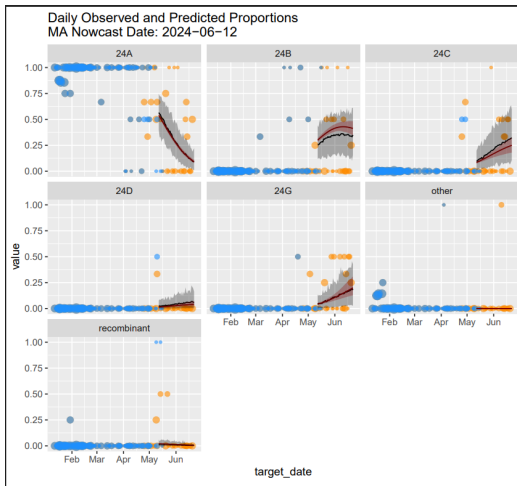
Example Forecasts

Figure: Mean energy scores: HMLR = 0.78, Baseline = 0.81

Isaac MacArthur, **Maryclare Griffin**, Evan Ray, Nicholas Reich, *Thomas Robacker*, and Benjamin Rogers

Forecasting and Nowcasting Variant Proportions with Genomic Data at the Regional Level
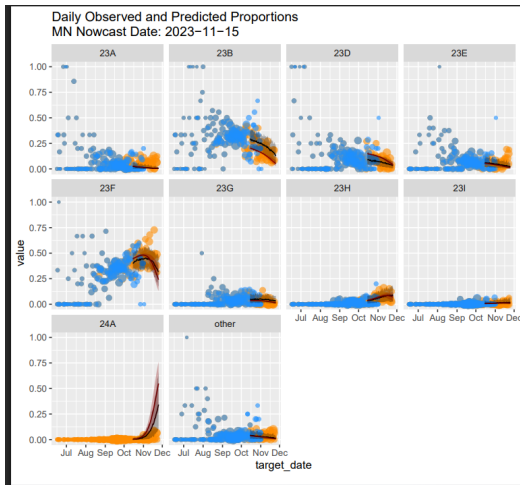
Example Forecasts

Figure: Mean energy scores: HMLR = 4.86, Baseline = 6.48

Isaac MacArthur, **Maryclare Griffin**, Evan Ray, Nicholas Reich, *Thomas Robacker*, and Benjamin Rogers

Forecasting and Nowcasting Variant Proportions with Genomic Data at the Regional Level
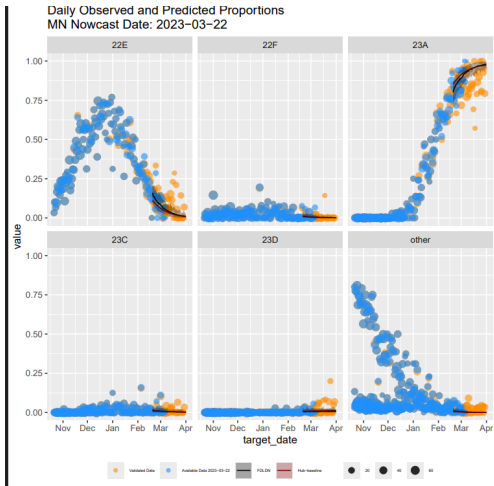
Example Forecasts

Daily Observed and Predicted Proportions
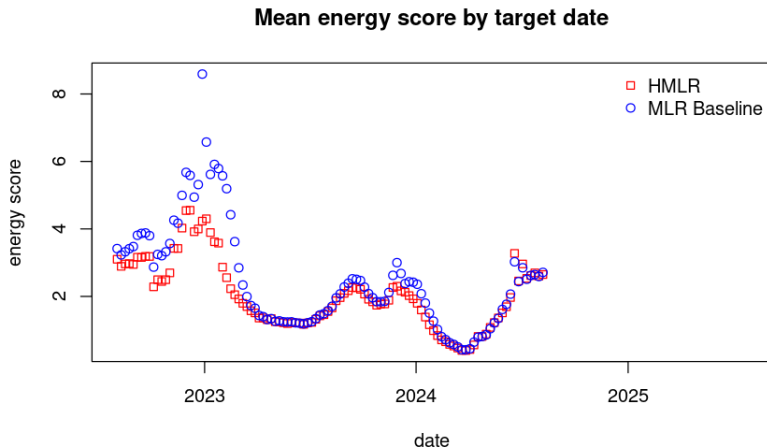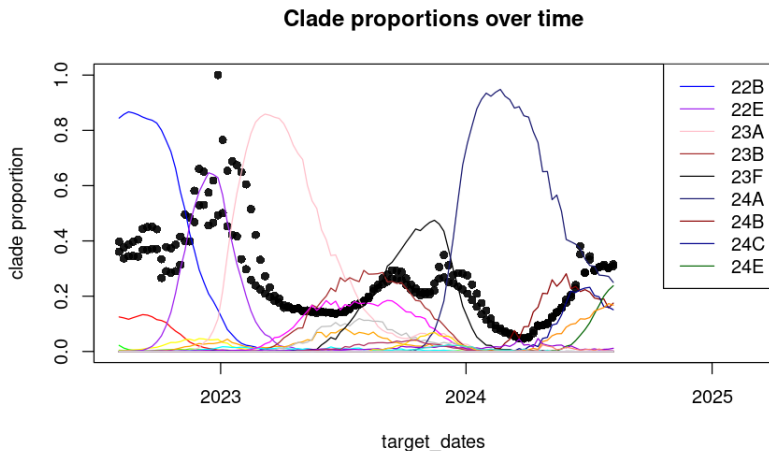MN Nowcast Date: 2023−03−22

Figure: Mean energy scores: HMLR = 2.76, Baseline = 3.02

Isaac MacArthur, **Maryclare Griffin**, Evan Ray, Nicholas Reich, *Thomas Robacker*, and Benjamin Rogers

Forecasting and Nowcasting Variant Proportions with Genomic Data at the Regional Level

Mean Energy Scores

Mean energy score by target date

Isaac MacArthur, **Maryclare Griffin**, Evan Ray, Nicholas Reich, *Thomas Robacker*, and Benjamin Rogers
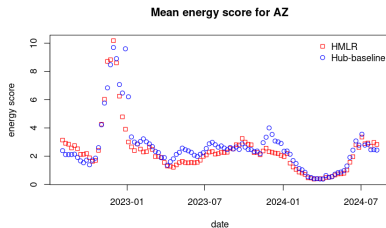Forecasting and Nowcasting Variant Proportions with Genomic Data at the Regional Level

Energy Score Trends
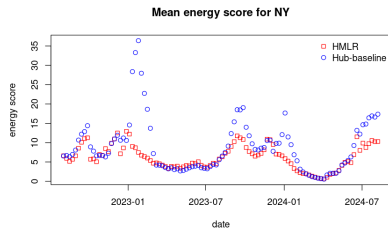

Clade proportions over time

Isaac MacArthur, **Maryclare Griffin**, Evan Ray, Nicholas Reich, *Thomas Robacker*, and Benjamin Rogers
Forecasting and Nowcasting Variant Proportions with Genomic Data at the Regional Level

Model Performance

Model	Mean score	Median score
HMLR	2.05	1.92
MLR Baseline	2.49	2.20

Results: Looking Across States

Isaac MacArthur, **Maryclare Griffin**, Evan Ray, Nicholas Reich, *Thomas Robacker*, and Benjamin Rogers
Forecasting and Nowcasting Variant Proportions with Genomic Data at the Regional Level

Conclusions

- Overall, the HMLR model does a reasonable job of forecasting and nowcasting variant proportions, given the difficulty of the problem and the lack of complete data.

- The HMLR model outperforms a country level MLR model based on Abousamra et. al (2024) in terms of the energy score.

- However, much more work needs to be done on the HMLR model.

Future Work

► Exploring other forms of model growth

► Doing more testing against other models

► Seeing if the model is more broadly applicable