EFFECT OF GRANULARITY ON SARS-COV-2 VARIANT ABUNDANCE ESTIMATES USING DOMESTIC WASTEWATER.

AUTHOR: YASH KALIA
Y.Kalia@student.tudelft.nl

SUPERVISOR: JASMIJN BAAIJENS
J.A.Baaijens@tudelft.nl

1. BACKGROUND

- RNA material present in domestic wastewater can be used to **estimate abundances of COVID-19 variants**[1] using the Baaijens pipeline.
- Predictions of RNA material can be made at **high granularity(HG)** i.e. **lineage level**, or **low granularity(LG)** i.e. **variant level**.
- Granularity level affects prediction accuracy.

2. RESEARCH QUESTION

Do predictions become more accurate at lower granularity?

- How does prediction accuracy differ at different granularities?
- What do the results at different granularities theoretically illustrate about the prediction pipeline.

3. EXPERIMENTS

- Variants simulated - **Alpha, Delta and Mu.**
- Single Lineage experiment – with Connecticut(CT)(small scale) and then US reference set(large scale)
- Combined Lineage experiment – with US reference set with all lineages in the same sample.
- Experiment Steps/Process: Kallisto constructs a graph with k-mers from reference set then pseudo aligns wastewater sequences to find closest lineage in reference set to sequence.
- Key performance indicator - **Relative prediction error**, to measure accuracy:

$$RPE = \frac{|T-E|}{T} \times 100$$

T: True abundance
E: Estimated abundance

[1] Jasmijn A Baaijens et al. "Variant abundance estimation for SARS-CoV-2 in wastewater using RNA-Seq quantification". In: *medRxiv* (2021).

4. RESULTS

Reference Set: **Connecticut,** One lineage per sample.

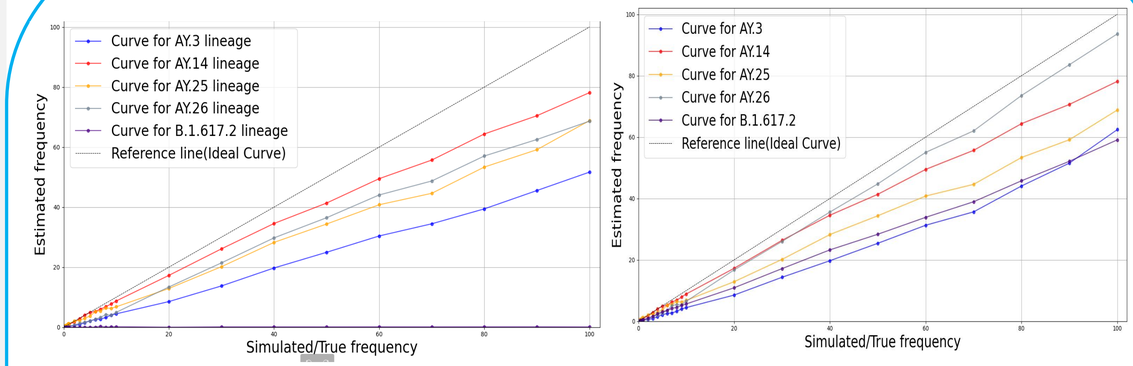

Figure 1: High(left) and Low(right) Granularity Results for Delta variant.

Reference Set: **US,** One lineage per sample.

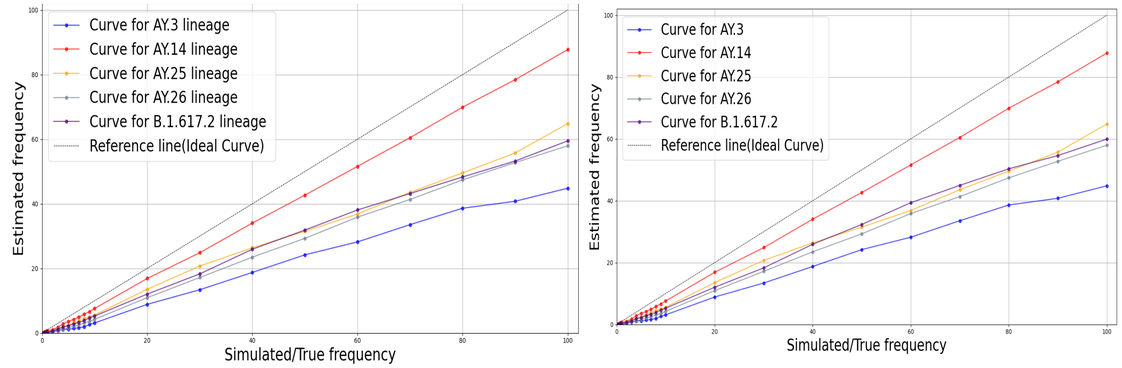

Figure 2: High(left) and Low(right) Granularity Results for Delta variant.

Reference Set: **US,** All lineages combined into one sample.

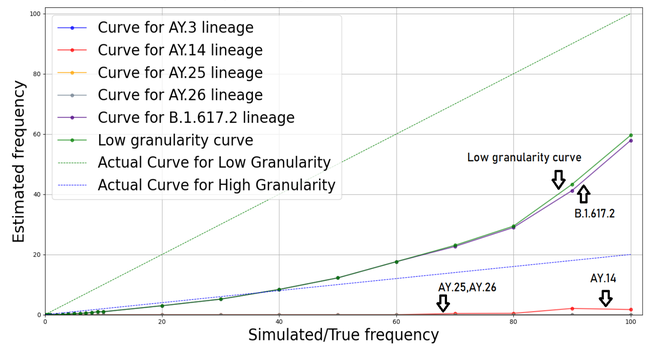

Figure 3: Granularity results for Delta variant(US), Combined Lineage Experiment.

5. ANALYSIS

- US reference set results were more accurate compared to CT due to a more detailed/larger Kallisto index.
- US reference set fixed the misclassification for B.1.617.2.
- Sub-lineage results were more accurate compared to root.
- Results for Delta and Alpha were inaccurate because of the number of lineages & genomes simulated together, except at very high abundances. Results for Mu were best in comparison.

6. CONCLUSION

- Prediction accuracy for **LG was consistently at least as high** as HG for single lineage experiments(CT and US).
- Single lineage results are more **regular** and **predictable** compared to combined lineage results**.**
- Pipeline is **not strong** at making predictions for more than 2 lineages combined together, especially for Alpha and Delta.

TU Delft
Delft University of Technology