aPhyloGeo-Covid: A Web Interface for Reproducible Phylogeographic Analysis SARS-CoV-2 Variation using Neo4j and Snakemake

Wanlin Li and Nadia Tahiri.
Department of Computer Science, Université de Sherbrooke, Canada

ABSTRACT

- This research developed an **interactive analysis platform** that facilitate efficient filtering and organization of input data for phylogeographical studies of SARS-CoV-2.
- The **Neo4j database** was integrated as a comprehensive repository, consolidating COVID-19 pandemic-related sequence information, climate data, and demographic data obtained from public databases.
- Additionally, the platform provides a **scalable and reproducible phylogeographic workflow** for investigating the intricate relationship between geographic features and the patterns of variation in diverse SARS-CoV-2 variants.
- Database currently contains 113,774 nodes and 194,381 relationships.
- The code is freely available to researchers and collaborators on GitHub.

WORKFLOW

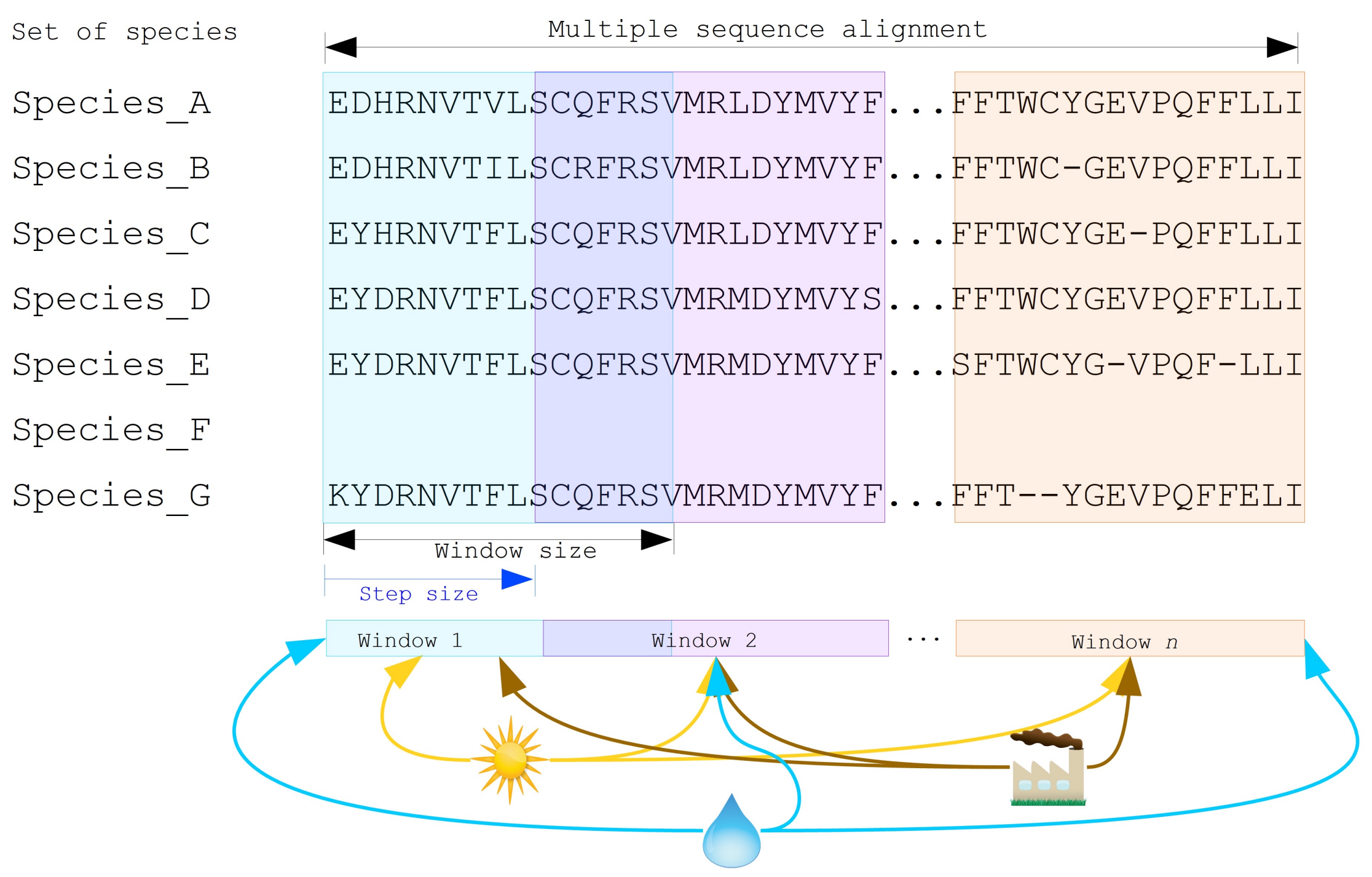

Set of species
Multiple sequence alignment
Species_A EDHRNVTVLSCQFRSVMRLDYMVYF...FFTWCYGEVPQFFLLI
Species_B EDHRNVTILSCRFRSVMRLDYMVYF...FFTWC-GEVPQFFLLI
Species_C EYHRNVTFLSCQFRSVMRLDYMVYF...FFTWCYGE-PQFFLLI
Species_D EYDRNVTFLSCQFRSVMRMDYMVYS...FFTWCYGEVPQFFLLI
Species_E EYDRNVTFLSCQFRSVMRMDYMVYF...SFTWCYG-VPQF-LLI
Species_F
Species_G KYDRNVTFLSCQFRSVMRMDYMVYF...FFT--YGEVPQFFELI
Window size
Step size
Window 1
Window 2
...
Window n

Figure 3: Integrated analysis of genetic data and environmental data

- Multiple sequences are **aligned** and **segmented** into numerous **alignment windows**
- **Robinson and Foulds (RF) metric** was employed to **quantify** the **dissimilarity** between
  - the phylogenetic tree of each window

  and
  - the topological tree of geographic features.
- ➢ **Tuning** of **window size** and **step size** are required to optimize the analysis.
- ➢ **Reproducibility** played a critical role in this process.

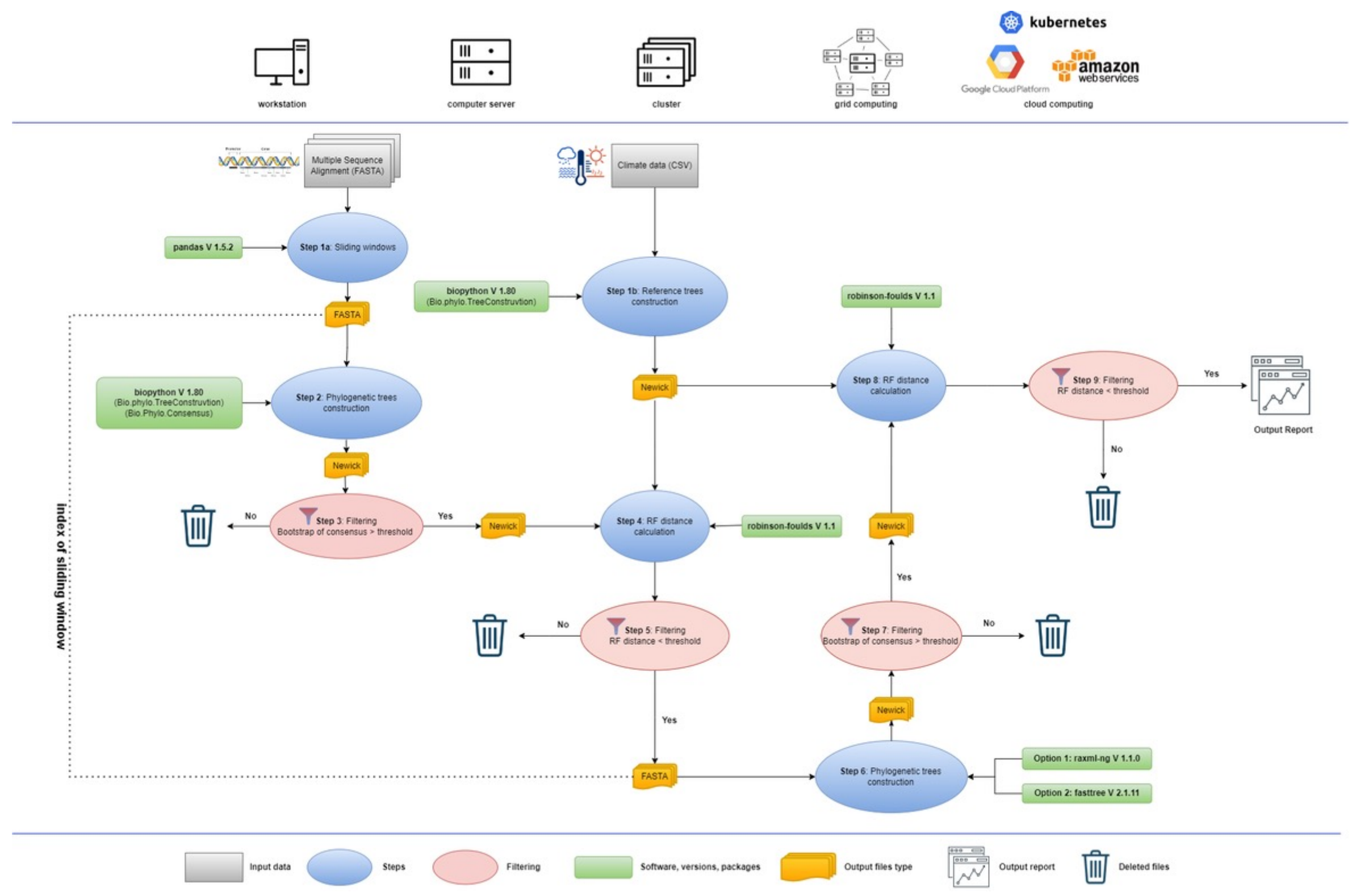

kubernetes
amazon web services
Google Cloud Platform
workstation
computer server
cluster
grid computing
cloud computing
Multiple Sequence Alignment (FASTA)
Climate data (CSV)
pandas V 1.5.2
Step 1a: Sliding windows
FASTA
biopython V 1.80 (Bio.phylo.TreeConstruvtion)
Step 1b: Reference trees construction
robinson-foulds V 1.1
biopython V 1.80 (Bio.phylo.TreeConstruvtion) (Bio.Phylo.Consensus)
Step 2: Phylogenetic trees construction
Newick
Step 8: RF distance calculation
Step 9: Filtering RF distance < threshold
Yes
No
Output Report
Step 3: Filtering Bootstrap of consensus > threshold
Step 4: RF distance calculation
Step 5: Filtering RF distance < threshold
Step 7: Filtering Bootstrap of consensus > threshold
Step 6: Phylogenetic trees construction
Option 1: raxml-ng V 1.1.0
Option 2: fasttree V 2.1.11
index of sliding window
Input data
Steps
Filtering
Software, versions, packages
Output files type
Output report
Deleted files

Figure 4: Snakemake workflow of the algorithm based on the new pipeline aPhyloGeo

DATA INTEGRATION

Neo4j Data Modeling

facilitate
- storage,
- management
- querying

of extensive SARS-CoV-2 variants-related data.

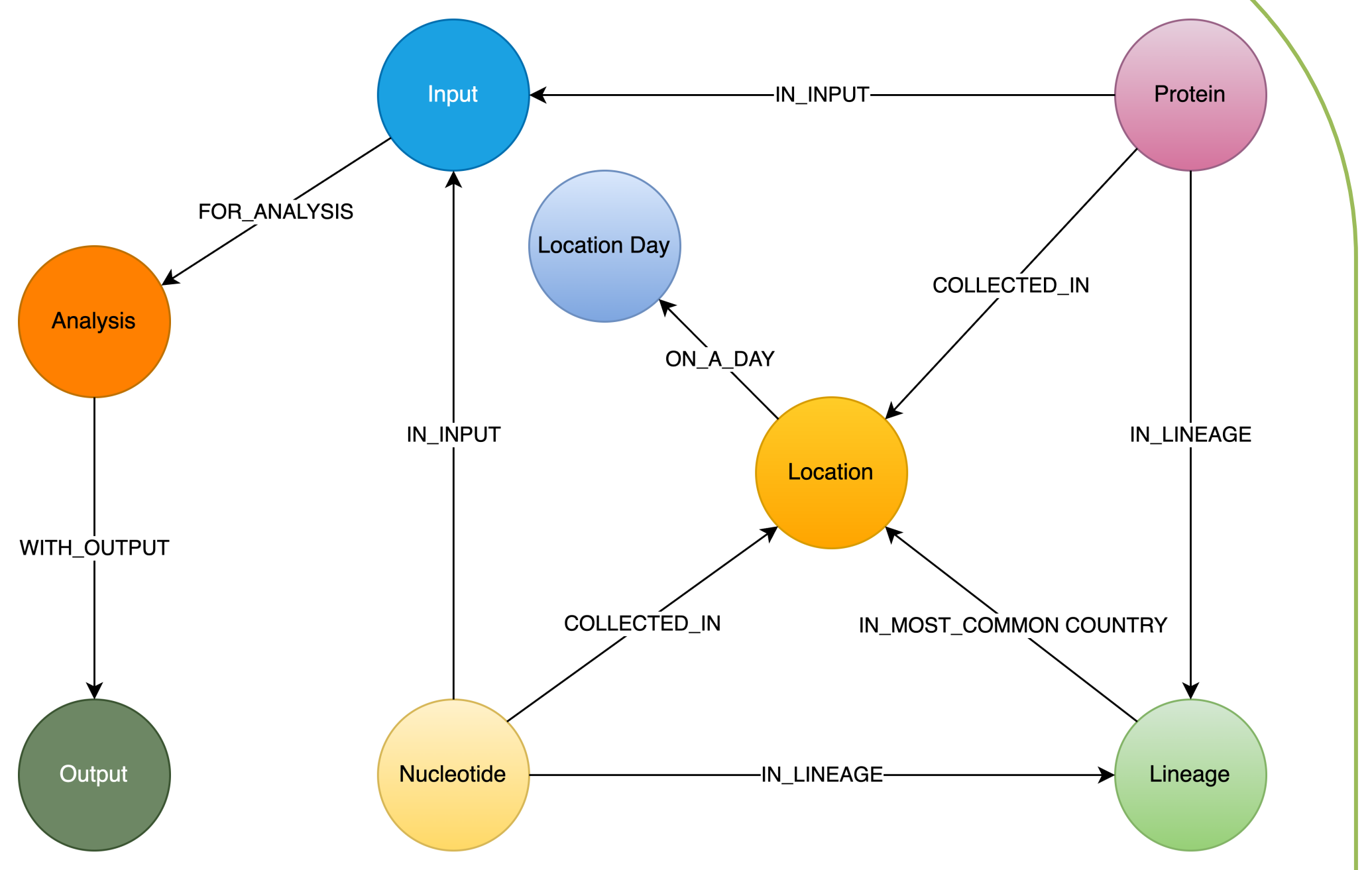


Figure 1: Schema of Neo4j Database for Phylogeographic Analysis of SARS-CoV-2 Variation

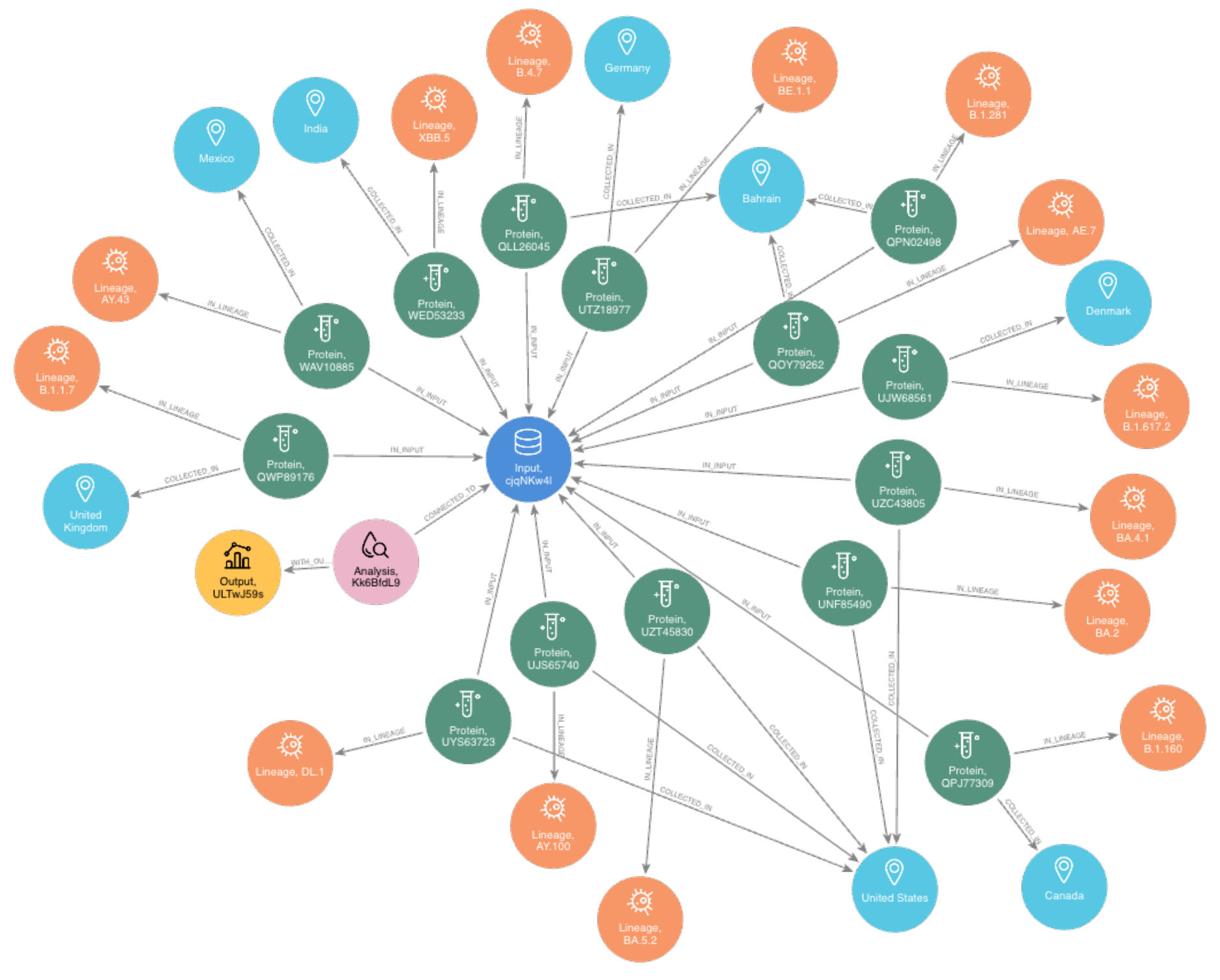


Figure 2: The networks of a single analysis experiment

Neo4j Data Analytics

ensure the **repeatability** and **comparability** of phylogeographic analysis results.

- The network highlights all entities serving as **input data sources** and their **relationships**.
- The **Input node** establishes connections between the data source objects and the specific analysis object.
- The **Analysis node** captures the parameters associated with the analysis.

RESULTS

- ➢ The **sliding window region** with the lower RF distance was exclusively identified in the integrated analysis.
- ➢ Within the regions identified with lower RF distance, a special attention should be given to regions with positions between 792 to 940 residue.

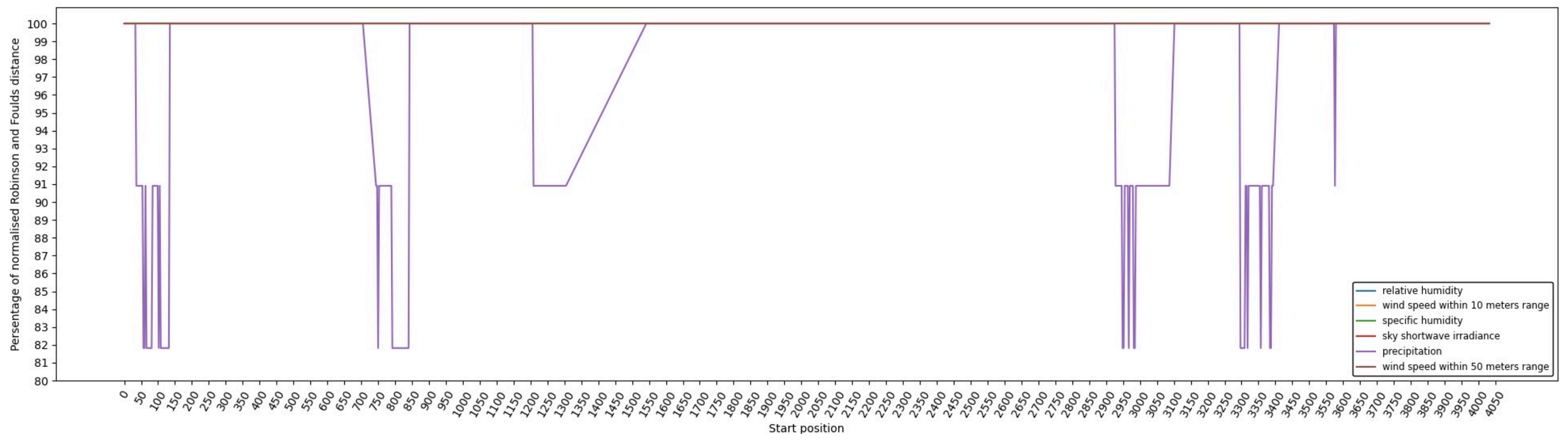
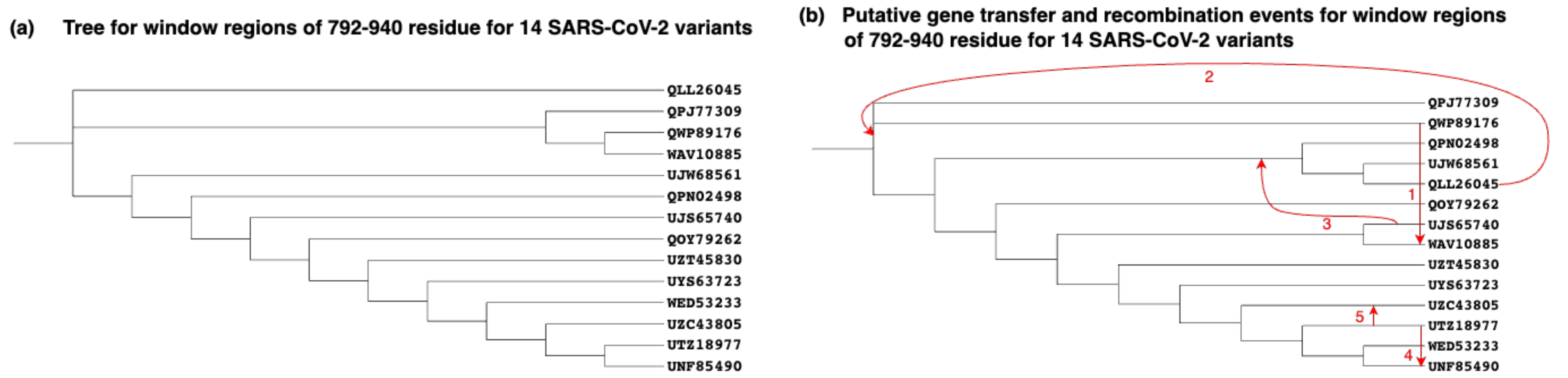


Figure 5: Robinson and Foulds topological distance normalized changes over the alignment windows

- ➢ Putative **horizontal gene transfer (HGT)** events were detected in the window regions spanning residues 792-940 of the amino acid sequences of 14 SARS-CoV-2 variants.


(a) Tree for window regions of 792-940 residue for 14 SARS-CoV-2 variants

(b) Putative gene transfer and recombination events for window regions of 792-940 residue for 14 SARS-CoV-2 variants

Figure 6: Putative horizontal gene transfer events found for the window regions of 792-940 residue (amino acid sequences) of 14 SARS-Cov-2 variants

CONCLUSIONS

- The platform streamlines the extraction, validation, and integration of genetic and environmental data, overcoming the limitations of manual tools.
- aPhyloGeo-Covid workflow utilizes a sliding window approach to identify region-specific subparts of viral genetic sequences affected by environmental conditions.
- The application of Snakemake workflow management ensures that the phylogeografic analysis can be replicated, validated, and used as a reliable basis for further research or analysis.
- The platform facilitates the sharing of research results, encourages collaboration and promotes the exploitation of previous work.

REFERENCES

Brister, J. Rodney, et al. (2015). NCBI viral genomes resource. *Nucleic acids research*, 43.D1, D571-D577.

Mathieu, E. et al. (2021). A global database of COVID-19 vaccinations. *Nature human behaviour*, 5.7: 947-953.

O'Toole, Á. et al. (2021). Tracking the international spread of SARS-CoV-2 lineages B. 1.1. 7 and B. 1.351/501Y-V2 with grinch. *Wellcome Open Research,* 6 (2021).

Tahiri (2012). Un nouvel algorithme pour retrouver les relations phylogénétiques entre la distribution géographique des espèces et leurs compositions génétiques.

Source

GitHub Link
https://github.com/tahiri-lab/iPhyloGeo/tree/iPhylooGeo-neo4j

Tahiri Lab website
https://tahirinadia.github.io/

We are hiring
- Master
- PhD
- Postdoc

Acknowledgements

Centre de recherche en ECOLOGIE DE L'UdeS
NSERC CRSNG
Fonds de recherche Nature et technologies Québec
UNIVERSITÉ DE SHERBROOKE