RASCL: RAPID ASSESSMENT OF SELECTION IN CLADES THROUGH MOLECULAR SEQUENCE ANALYSIS

*Alexander G Lucaci[1], Jordan D Zehr[1], Stephen D Shank[1], Dave Bouvier[2], Alexander Ostrovsky[3], Han Mei[2], Anton Nekrutenko[2], Darren P Martin[4], Sergei L Kosakovsky Pond[1]

Abstract

An important unmet need revealed by the ongoing COVID-19 pandemic is the near-real-time identification of potentially fitness-altering mutations within rapidly growing SARS-CoV-2 lineages. Motivated by the need to analyze new lineage evolution in near-real time using large numbers of genomes, we developed the Rapid Assessment of Selection within CLades (RASCL) pipeline. RASCL applies state of the art phylogenetic comparative methods to evaluate selective processes acting at individual codon sites and across whole genes. By enabling the rapid detection of genome sites evolving under different selective regimes, RASCL is well-suited to near-real-time monitoring of the population-level selective processes that will likely underlie the emergence of future variants of concern in rapidly evolving pathogens with extensive genomic surveillance.

Methods

RASCL is designed for scalable and automatically updated lineage-specific selection analysis reports, even for lineages that include tens or hundreds of thousands of sampled genome sequences. Key to this performance is:

- The dynamic automated generation of high quality down-sampled datasets of gene/ORF sequences drawn from a selected "**query**" viral lineage.
- Contextualization of these query sequences in codon alignments that include high-quality "**background**" sequences representative of global SARS-CoV-2 diversity.
- The extensive parallelization of batteries of computationally intensive selection analysis tests including:
  - ❑ **SLAC**: performs substitution mapping.
  - ❑ **BGM**: identifies groups of sites that might be co-evolving.
  - ❑ **FEL**: locates codon sites with evidence of pervasive positive diversifying or negative selection.
  - ❑ **MEME**: locates codon sites with evidence of episodic positive diversifying selection.
  - ❑ **BUSTED[S]**: tests for gene-wide episodic selection.
  - ❑ **RELAX**: compare gene-wide selection pressure between the query clade and reference sequences.
  - ❑ **Contrast-FEL**: comparison of site-by-site selection pressure between query and reference sequences.
  - ❑ **FADE**: identify amino-acid sites with evidence of directional selection.
  - ❑ **FMM**: identify sites with complex multiple instantaneous substitutions.

Results

Figure 1. RASCL results include a full-feature interactive ObservableHQ notebook with explorable multiple sequence alignments for each gene, and the inferred phylogenetic tree.

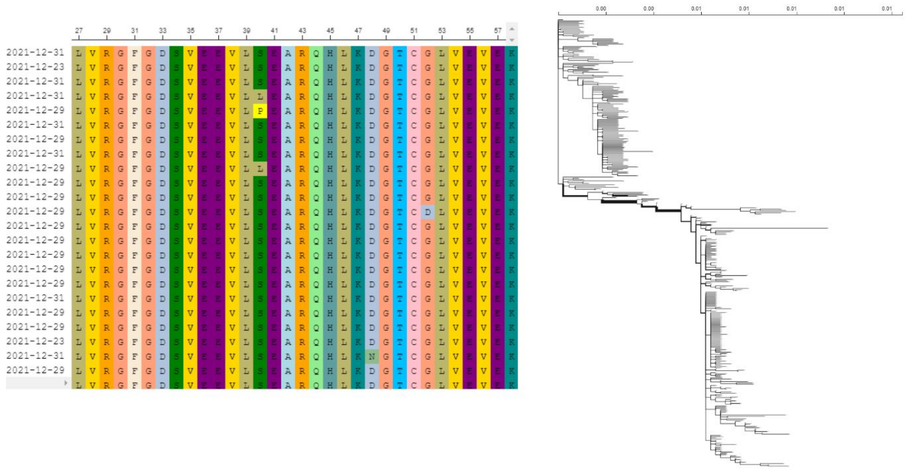


Figure 2. Post-hoc spatial and structural analysis of results are available to be explored using the ecosystem of accessory interactive notebooks we have developed to investigate the relationship between sites of interest and the viral clade under analysis.

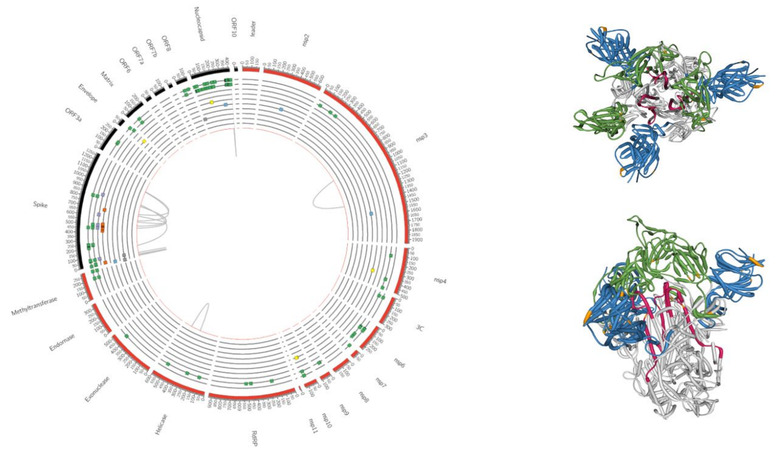
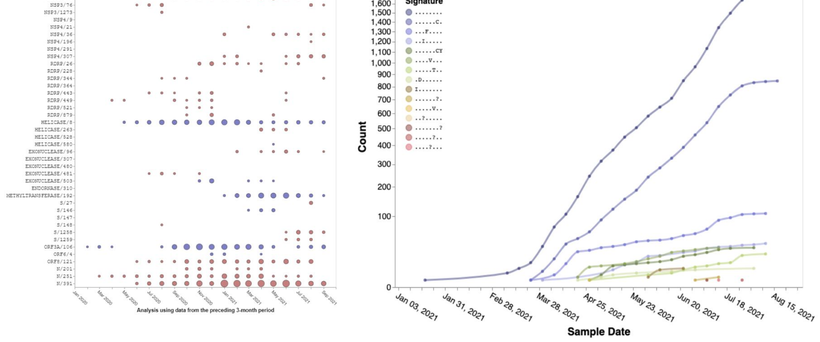


Table 1. We also present summary tables for all selection analyses (see Methods section for details). Here, we highlight the statistically significant sites of interest in the Spike gene. Table columns include the Genomic Coordinate within the SARS-CoV-2 genome, Gene/ORF we are exploring, codon site in the corresponding gene, number of branches under selection (via the MEME method), uncorrected LRT p-value, q-value to correct for FDR, and the physicochemical property (if available) of the substitution (via the PRIME method).

Coordinate (SARS-CoV-2)	Gene/ORF	Codon (in gene/ORF)	# of selected branches	p-value	q-value	Properties
21574	S	5	4	0.0158	0.0188	
21700	S	47	1	0.0120	0.0153	
21748	S	63	1	0.00720	0.0111	
21760	S	67	6	0.00373	0.00671	
21820	S	87	1	0.0149	0.0181	overall, volume (constrained) , composition (changing)
21829	S	90	2	0.00174	0.00377	overall, volume (changing), charge (constrained)
21832	S	91	3	0.000126	0.000514	overall, composition (changing)
21835	S	92	2	0.00163	0.00358	overall, composition (constrained)
21838	S	93	3	0.00148	0.00337	overall, secondary (constrained)
21844	S	95	7	0.000741	0.00199	
21850	S	97	3	0.000342	0.00121	overall, volume (changing)
21853	S	98	0	0.0190	0.0222	composition (changing)
21868	S	103	1	0.0106	0.0140	
21886	S	109	2	0.0118	0.0152	overall, bipolar (changing), secondary (constrained) , composition (constrained)
21895	S	112	2	0.00823	0.0119	overall, charge (changing)

Figure 3. Post-hoc temporal analysis of our results are also available to be explored using data from global SARS-CoV-2 analysis to track codon sites of interest and their amino acid signatures.



Conclusions

RASCL has been used to characterize the role of natural selection in the emergence of the Beta, Gamma, and Omicron VOC lineages, and for identifying patterns of convergent evolution in the Alpha, Beta and Gamma lineages. We are presently using RASCL to monitor the ongoing evolution of a number of current VOI/VOC lineages. Whenever future genomic surveillance efforts reveal new potentially problematic SARS-CoV-2 lineages, we will use RASCL to analyze these too. Finally, RASCL has been designed so that, with minimal modification (reference genomes, genes, and default thresholding settings), it can also be adapted to analyze any other viral pathogens for which sufficient sequencing data is available. The RASCL application and current results are available from dedicated repositories at:

- **Github** as a Snakemake pipeline https://github.com/veg/RASCL.
- As a **Galaxy workflow** https://usegalaxy.eu/u/hyphy/w/rascl.
- Existing clade analysis results are available here: https://observablehq.com/@hyphy_software/rascl.

Contact

Alexander G. Lucaci
Alexander.Lucaci@Temple.edu

Department of Biology
Temple University

Acknowledgements

We thank members of the Datamonkey, HyPhy, and Galaxy teams for their assistance in the development of this application. We are also thankful for the GISAID Contributors from across the world.

Author affiliations

[1]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, Pennsylvania, USA
[2]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA, USA
[3]Krieger School of Arts and Sciences, Johns Hopkins University, Baltimore, MD, USA
[4]Institute of Infectious Diseases and Molecular Medicine, Division Of Computational Biology, Department of Integrative Biomedical Sciences, University of Cape Town, Cape Town 7701, South Africa

For software and results SCAN ME

Check out our preprint! SCAN ME