##### 1. Tool/Model Name

- \*\*Name:\*\* SONIA

##### 2. Reference(s)

- \*\*Citation:\*\* Sethna, Z., Isacchini, G., Dupic, T., Mora, T., Walczak, A. M., & Elhanati, Y. (2020). Population variability in the generation and selection of T-cell repertoires. \*PLOS Computational Biology\*, 16(12), e1008394.

- \*\*Link:\*\* https://doi.org/10.1371/journal.pcbi.1008394

- \*\*Publication Date:\*\* December 9, 2020

##### 3. Category

- \*\*Category:\*\* Computational Biology/Immunoinformatics, specifically for T-cell receptor repertoire analysis. This falls under the broader umbrella of tools for analyzing adaptive immune receptor repertoires (AIRR), though it is not explicitly labeled as a V(D)J simulator or SHM model, focusing instead on selection analysis.

##### 4. Main Purpose/Function

- \*\*Purpose:\*\* SONIA is designed to infer individual-specific computational models for the generation and selection of TCR beta chains (TRB) from sequenced repertoires. It addresses the problem of quantifying selection patterns by comparing the statistics of observed (post-selection) sequences with background (pre-selection) samples, thereby separating the contributions of V(D)J recombination (generation) and thymic selection to TCR repertoire diversity.

- \*\*Problem Addressed:\*\* The tool tackles the challenge of understanding the stochastic processes that shape TCR diversity, particularly how selection pressures influence the final repertoire, which is critical for studying immune responses in health and disease.

##### 5. Key Features

SONIA’s key features include:

- \*\*Probabilistic:\*\* Yes, it uses probabilistic models to infer selection factors, employing maximum likelihood estimation (MLE) and comparing probability distributions (e.g., P\_gen and P\_post).

- \*\*ML-based:\*\* Partially, it leverages TensorFlow and Keras for iterative model fitting, using Kullback-Leibler divergence as a loss function, but it is more statistical than deep learning-focused.

- \*\*Modular:\*\* Yes, it is flexible and modular, allowing users to define custom feature sets for selection models and apply it to different receptor chains (e.g., TRA, immunoglobulins).

- \*\*AIRR-compliant:\*\* While not explicitly stated, SONIA uses standard immunosequencing data formats (e.g., CDR3 sequences, V/J gene alignments) and integrates with tools like iGOR and OLGA, which are part of the AIRR community, suggesting compatibility.

- \*\*Supports SHM:\*\* No, it focuses on V(D)J recombination and thymic selection for TCRs, not somatic hypermutation (SHM), which is more relevant for B-cell receptors.

##### 6. Input Requirements

- \*\*Data Needed:\*\*

- A list of productive CDR3 amino acid sequences (the hypervariable region of TCRs).

- Optionally, V/J gene alignments for each sequence to provide additional context for generation and selection modeling.

- Background (pre-selection) sequences, which can be provided directly or generated using a VDJ generation model like OLGA.

- \*\*Parameters:\*\* The tool can handle unique sequences or sequences with clonal weights (e.g., accounting for clone sizes in specific contexts like antigen exposure).

##### 7. Output

- \*\*Produced Outputs:\*\*

- Selection factors specific to sequences (Q(α)), which quantify the probability of a sequence being selected post-thymic selection.

- Probability distributions before and after selection (P\_gen for generation, P\_post for post-selection).

- Statistical measures such as feature marginals (e.g., amino acid usage, CDR3 length distributions), variance, and covariance of generation and selection probabilities across individuals.

- Simulated selected sequences, generated via rejection sampling when combined with tools like OLGA for sequence generation.

##### 8. Validation/Evaluation

- \*\*Validation Method:\*\* SONIA was validated using TRB sequences from 651 healthy individuals from a large cohort study (Emerson et al., 2017). It was applied to high-throughput T-cell repertoire sequencing (RepSeq) data to ensure robustness.

- \*\*Evaluation Metrics:\*\*

- Compared predicted post-selection distributions (P\_post) with experimental data using density plots, Pearson correlation, and information theory metrics like Jensen-Shannon divergence.

- Tested two selection models: LengthPosition (with more parameters for higher flexibility) and Left+Right (with fewer parameters, focusing on relative positions of amino acids in CDR3). The Left+Right model was selected for its balance of efficiency and accuracy in fitting experimental frequencies.

- Assessed convergence using L1 distance between data and model feature marginals, ensuring the model could reproduce observed sequence statistics (e.g., CDR3 length distribution, amino acid usage).

- \*\*Usage:\*\* While the article does not explicitly state widespread use, its integration with other established tools (iGOR, OLGA) and open-source availability suggest growing adoption in the immunoinformatics community.

##### 9. Strengths

- \*\*Advantages:\*\*

- High flexibility in defining and inferring selection models, allowing customization of features such as single amino acids, pairs, motifs, gene usage, and CDR3 length, making it adaptable to various research needs.

- Efficient parameter reduction, as seen with the Left+Right model, which balances accuracy and computational efficiency, suitable for large datasets.

- Integration with other tools like iGOR (for generation models) and OLGA (for sequence generation), enhancing its utility in a pipeline for comprehensive repertoire analysis.

- Demonstrated consistency across a large population (651 individuals), with inter-individual variance being only 1.6% of intra-individual variance, indicating robustness for population studies.

- Open-source availability on GitHub, facilitating reproducibility, community contributions, and further development.

##### 10. Limitations

- \*\*Weaknesses and Issues:\*\*

- Focuses on statistical features rather than specific sequence-HLA associations, potentially missing precise selection pressures such as HLA-specific effects or antigen-specific responses, which are crucial for certain immunological studies.

- Limited to T-cell receptors (TRB, with potential extension to TRA and immunoglobulins), and not directly applicable to B-cell receptors with somatic hypermutation (SHM) without modification, limiting its scope for B-cell-related research.

- Dependent on high-quality input data (e.g., sequenced repertoires), and performance may be affected by sampling biases or noise in high-throughput sequencing, which can be a challenge in real-world applications.

- The studied population (651 individuals) may not represent the full diversity of human populations, as acknowledged in the article, potentially affecting generalizability.

##### 11. Applications in Autoimmunity

- \*\*Use in Autoimmune Disease:\*\* The article suggests SONIA has potential for studying autoimmune diseases, though it is not explicitly used in such contexts in the provided study. It could be applied to:

- Identify specific selection pressures in T-cell subsets associated with autoimmune conditions, such as comparing repertoires between healthy individuals and patients with diseases like type 1 diabetes or rheumatoid arthritis.

- Determine if TCR sequences linked to autoimmunity are generated or selected differently, aiding in understanding mechanisms like central tolerance breakdown, which is critical for autoimmune pathology.

- Model selection pressures to compare T-cell repertoires, potentially leading to the identification of biomarkers or therapeutic targets for autoimmune diseases.

- \*\*Suitability:\*\* Given its focus on T-cell selection and flexibility, SONIA seems suitable for autoimmune research, particularly for T-cell-mediated diseases, but further studies are needed to validate its efficacy in clinical settings.

##### 12. Notable Citations/Use Cases

- \*\*Key Papers/Studies:\*\*

- The tool builds on previous work in statistical modeling of immune repertoires, such as Elhanati et al. (2014), "Quantifying selection in immune receptor repertoires," published in \*Proceedings of the National Academy of Sciences\*, and Murugan et al. (2012), "Statistical inference of the generation probability of T-cell receptors from sequence repertoires," published in \*Proceedings of the National Academy of Sciences\*.

- It integrates with iGOR and OLGA, which are referenced for generation models and sequence generation, respectively, enhancing its utility in a broader immunoinformatics pipeline.

- \*\*Use Cases:\*\* Potential applications include studying immune responses to vaccines, infections, or cancer, as well as differentiating T-cell phenotypes (e.g., naive, effector, memory) or sample types (e.g., blood, tissue, tumor), which could extend to autoimmune research contexts.

##### 13. Reviewer’s Comments

- The article does not explicitly include reviewer comments, but the authors acknowledge that SONIA’s approach complements other strategies, such as analyzing TCR-HLA associations, and accept limitations such as the restricted subset of the human population studied (651 individuals).

- SONIA is presented as a significant advancement for flexible modeling of selection pressures, with potential future applications in clinical and biological contexts, such as vaccine design and immunotherapy, suggesting room for further development and validation in diverse settings.

#### Conclusion

SONIA represents a significant tool for analyzing T-cell receptor repertoires, with strong potential for applications in autoimmune disease research by modeling selection pressures and comparing repertoires. However, its limitations, such as the focus on T-cells and dependency on high-quality data, suggest areas for future improvement and broader validation. Researchers are encouraged to explore its integration with other tools and its application in clinical contexts, leveraging its open-source nature for further development. For more detailed insights, refer to the original article at https://doi.org/10.1371/journal.pcbi.1008394.