Delineating inter- and intra-antibody repertoire evolution with AntibodyForests

Daphne van Ginneken[1], Valentijn Tromp[1], Lucas Stalder[2], Tudor-Stefan Cotet[2], Sophie Bakker[1], Anamay Samant[2], Sai T. Reddy[2,3], Alexander Yermanos[1,2,3,*]

[1]Center for Translational Immunology, University Medical Center Utrecht, Lundlaan 6, Utrecht 3584EA, The Netherlands.
[2]Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland.
[3]Botnar Institute of Immune Engineering, Basel, Switzerland
*Corresponding author. Email: a.d.yermanos@umcutrecht.nl

Abstract

Motivation: The rapid advancements in immune repertoire sequencing, powered by single-cell technologies and artificial intelligence, have created unprecedented opportunities to study B cell evolution at a novel scale and resolution. However, fully leveraging these data requires specialized software capable of performing inter- and intra-repertoire analyses to unravel the complex dynamics of B cell repertoire evolution during immune responses.
Results: Here, we present AntibodyForests, software to infer B cell lineages, quantify inter- and intra-antibody repertoire evolution, and analyze somatic hypermutation using protein language models and protein structure.
Availability and implementation: This R package is available on CRAN and Github at https://github.com/alexyermanos/AntibodyForests, a vignette is available at https://cran.case.edu/web/packages/AntibodyForests/vignettes/AntibodyForests_vignette.html

Introduction

Rapid progress in immune repertoire sequencing and artificial intelligence are advancing the field by providing high-quality datasets at single-cell resolution and pre-trained large protein language models (PLMs). These datasets of paired heavy- and light-chain sequences can additionally include informative labels on cellular phenotype, antigen-binding and specificity, as well as protein structure. Moreover, the integration of bulk RNA sequencing data can significantly enhance the resolution of immune repertoire analyses and reduce undersampling issues common to single-cell experiments. Pre-trained PLMs have demonstrated the ability to understand structural and functional properties from protein sequences and have been used to predict features of antibody function (1–6). The modalities of this data play a critical role in unraveling the evolutionary processes of B cells during immune responses. Numerous analyses and tools focus on inferring and quantifying individual antibody lineages (7–12), but there is a lack of software dedicated to studying antibody sequence and structural evolution at the repertoire level. In response to this gap, we introduce AntibodyForests, a comprehensive software designed to thoroughly investigate and quantify the inter- and intra-antibody repertoire evolution (**Figure 1**). AntibodyForests encompasses pipelines for analyzing sequences and their metadata, tree topology, protein structure, and PLM features.

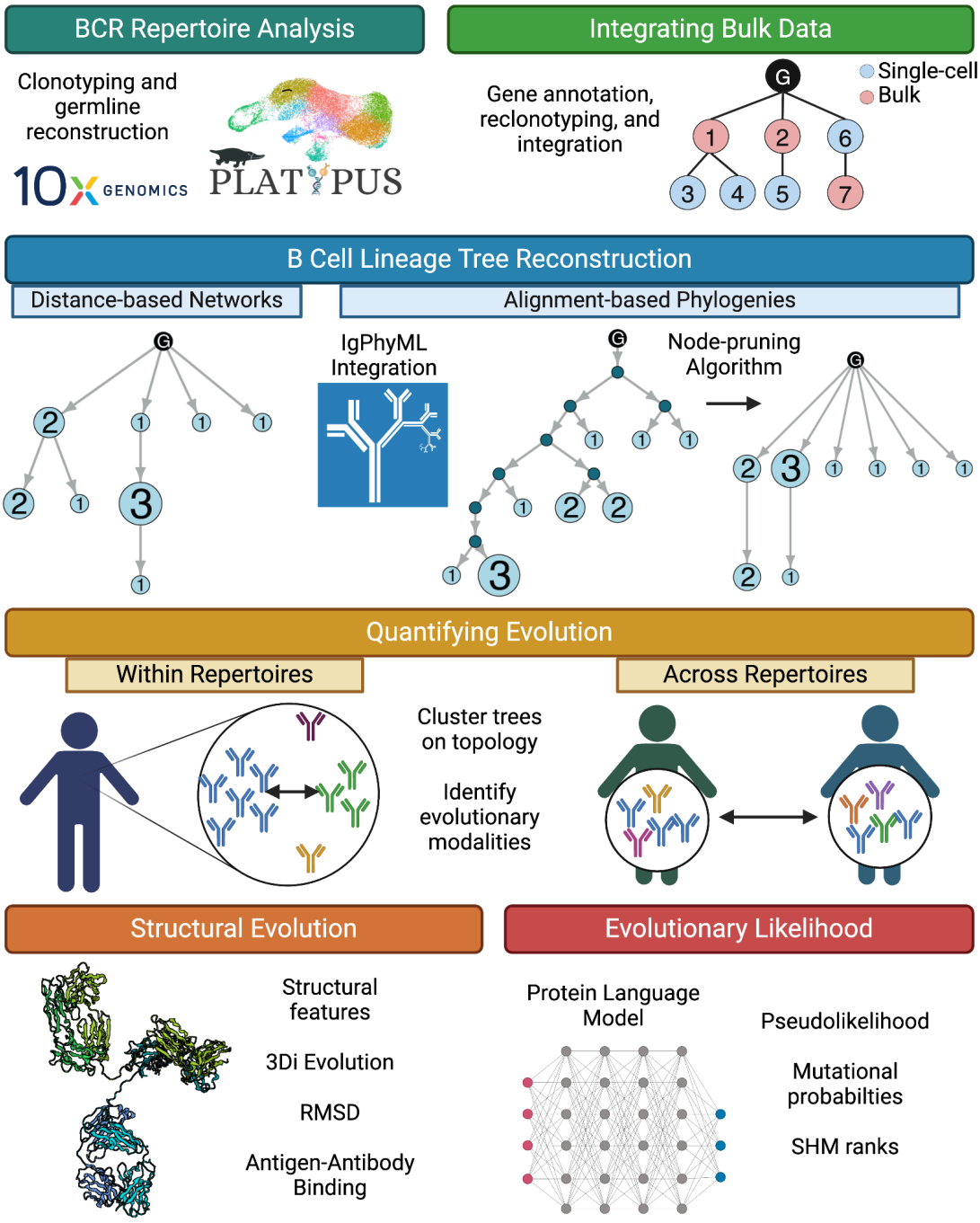

Figure 1. Overview of the AntibodyForests package. This R package is currently composed of a pipeline to reconstruct lineage trees from V(D)J sequencing data preprocessed with the Platypus package and compare trees within and across repertoires. Furthermore, it has modalities to integrate bulk RNA sequencing data, features of protein structure, and evolutionary likelihoods generated with protein language models.

Usage and application

Tree reconstruction

AntibodyForests infers B cell mutational networks for reconstructed clonal lineages (clonotypes) **(Figure S1)**. These clonotypes are defined as B cells arising from the same V(D)J recombination event that have undergone somatic hypermutation (SHM) relative to an unmutated reference germline**.** In AntibodyForests, each clonal lineage is represented as a graph in which the nodes refer to unique antibody sequences, and the edges separating the nodes dictate the clonal relationship between variants. These nodes can be full-length V(D)J sequences, or only certain gene segments or complementarity-determining regions (CDR) from the heavy and/or light chain. Our framework diverges from the traditional B cell phylogenetic tree structure by enabling recovered sequences to serve as either internal or terminal nodes, in addition to allowing for multifurcation events. Some phylogenetic methods generate internal nodes serving as the most recent common ancestors (MRCA). In a BCR repertoire, these MRCAs can be recovered in the sequencing data (7,13,14), resulting in nodes with zero branch length to the internal parent node (**Figure S2**). For this reason, AntibodyForests includes various algorithms to remove these internal nodes allowing for multifurcation events.

AntibodyForests contains several tree construction algorithms based on either distance matrices or multiple sequence alignments and incorporates minimum spanning tree algorithms, phylogenetic methods, and external lineage tools such as IgPhyML (15). Users can supply previously computed phylogenetic trees, thereby enabling the integration of different mutational models, such as those developed specifically for antibodies (8,16,17). AntibodyForests can further interrogate and visualize network robustness by comparing tree topology across different tree construction methods using metrics such as the generalized branch length distance (GBLD) (18) **(Figure S4)**. In the case of a tie during phylogenetic network construction, AntibodyForests has several options that allow the user to select whether breadth or depth (i.e., whether new sequences are added to the node with the most or least descendants), mutational load (whether they are added to the node closest or furthest from the germline), clonal expansion (whether they are added to the node with the highest or lowest cell count), or random addition dictates network topology. Although the package is designed for single-cell immune repertoire data, the pipeline can also be used with bulk repertoire sequences and even has a function to integrate the two data modalities **(Figure S3).**

Upon completion of the network, the color of each node can correspond to the fraction of cells with a given isotype, transcriptional cluster (if such information is available), or a custom BCR feature **(Figure S2)**. To highlight receptor sequences that are identical across multiple B cells, the node size can be scaled to match the relative clonal expansion, while the label can be adjusted to depict the variant frequency. To understand the evolutionary distance separating two nodes, it is possible to include edge labels that correspond to the distance separating nodes**.**

Quantifying intra- and inter-antibody repertoire evolution

A quantitative interpretation of antibody lineage trees is crucial to identify different patterns of clonal selection and expansion that drive B cell evolution within and across repertoires. A repertoire-wide analysis of these tree topologies can identify features of SHM and could identify potentially interesting therapeutically relevant antibodies or vaccine targets. For example, Wu et al. discovered a broadly neutralizing antibody against HIV-1 by identifying progenitor sequences preceding viral neutralization capabilities using phylogenetics (19).

Phylogenetic methods have been extensively used to investigate single B cell clonal lineages (11,19–23). However, the evolution of the repertoire as a system has not been fully explored using the collection of lineage trees. AntibodyForests includes two modules explicitly dedicated to comparing evolution within and across repertoires, in addition to comparing topologies consisting of identical sequences **(Figure S4).** Tree topology metrics can be either calculated and stored within the AntibodyForests object or into a separate matrix that can be supplied for downstream clustering and visualization. We have included various metrics to describe tree topologies including general graph theory metrics, and more specialized metrics such as the Sackin index for tree imbalance and selective pressure, and Laplacian spectral density for characterizing evolutionary patterns (24). Together, AntibodyForests' metrics can be used to project topology features using dimensionality reduction techniques and clustering, thereby identifying topologically similar networks. Clusters of trees can be related to features specific to single-cell immune repertoire sequencing, such as transcriptional phenotype, isotype, or expansion, as well as other custom metadata features provided by the user..

We analyzed single-cell BCR data collected from the blood and lymph nodes of individuals post SARS-CoV-2 vaccination (21) and performed an across- and within-repertoire analysis using AntibodyForests. The across-repertoire comparison revealed the blood repertoire to have more diversity, while the deeper trees in the lymph node repertoire suggest more sequential evolution **(Figure S5A)**. Next, we clonotyped B cells from all tissues and time points for each individual and clustered the trees based on the Laplacian spectrum **(Figure S5B)**. This within-repertoire comparison demonstrated a subset of trees displaying deep branching events, suggesting that cells with a small degree of SHM were recovered. This same subset was characterized by multiple topology modalities, indicating various diversification events **(Figure S5C)**. We further used AntibodyForests' functions to quantify whether V-gene usage, isotype, or sampling time point were correlated with lineage topology clustering **(Figure S5D)**. Additional analysis with AntibodyForests revealed that SHM increased with the time after vaccination and that blood-derived BCRs were located closer to the germline than those from lymph nodes (**Figure S5E**). This use case displays the suitability of AntibodyForests to comprehensively reconstruct and analyze whole BCR repertoires and study patterns of SHM upon immune activation.

Protein language model likelihood

PLMs have demonstrated success in understanding features of sequence, structure, and function of natural proteins (general PLMs) (25–28). Following this success, additional PLMs trained specifically on antibody sequences were able to learn B cell specific features such as

germline usage and SHM (antibody-specific PLMs) (1–5). The learning objective of PLMs allows for the computation of a likelihood score for each residue in the input sequence. These scores can be used to calculate a so-called pseudolikelihood for the entire sequence. PLM-based pseudolikelihoods have been leveraged to capture global evolutionary trajectories (29). Furthermore, general PLMs have been demonstrated to improve antibody affinity through in silico mutagenesis by mutating for increasing evolutionary likelihood (6). Given the relevance of PLMs in studying and engineering antibodies, we included PLM-guided metrics within AntibodyForests to describe B cell evolution at the level of an entire repertoire and a single lineage. Both per-residue likelihoods and per-sequence pseudolikelihoods can be incorporated into AntibodyForests to evaluate correlations between tree topology and PLM-based likelihoods **(Figure S6A)**.

To demonstrate the ability of AntibodyForests to employ PLMs for identifying patterns of B cell evolution, we further analyzed B cell repertoire data after vaccination against SARS-CoV-2 (21). Pseudolikelihoods and per-residue likelihoods were computed with the antibody-specific PLM Ablang (1) on the heavy chain sequences. A strong negative correlation was observed between the distance to the germline and the pseudolikelihood (r^2 = -0.74) **(Figure S6BC)**. The mutations along the edges of the trees were usually towards one of the most likely residues, but never the most likely, thereby suggesting that PLM likelihoods capture information about the rules governing SHM **(Figure S6D)**. The original residue at the mutating site usually had the highest likelihood **(Figure S6E)**, which could be expected given antibody-specific PLMs are inherently biased towards germline-like sequences due to their training data (4,30).

Structural evolution

Recent computational efforts have massively accelerated the ability to predict protein structures from sequence alone (27,28,31,32). Although SHM acts on the sequence directly, the structure of the antibody determines binding affinity with antigens. We therefore included a structure-specific module in AntibodyForests to integrate antibody structural evolution (**Figure S7A**). This allows users to quantify structural features as a function of sequence distance from the germline. Examples of these features are: 1) structure similarity (root-mean-square deviation, RMSD), 2) biophysical properties (hydrophobicity, charge, 3Di alphabet (33), free energy, and pKa (34)), and 3) global confidence of the prediction (pLDDT). Furthermore, we have added the option to include antibody-antigen complexes and calculate how properties at the binding interface change with SHM. In addition to RMSD to the germline, AntibodyForests can calculate a pairwise RMSD between the structures connected with edges in the lineage trees and calculate their correlation with the number of mutations. Although structural modeling methods still struggle with the complex and variable nature of antibody CDR loops and individual mutations on structural predictions (32,35,36), the framework included in AntibodyForests will remain relevant as deep learning algorithms continue to improve.

We predicted the structure of antibody-antigen complexes using AlphaFold3 (37) from a single B cell clone that is known to bind Ovalbumin (OVA) (20) and used AntibodyForests to analyze the correlation between structure and sequence evolution. Lower confidence scores from

AlphaFold3 were observed for a specific branch in the lineage tree, suggesting more intrinsically disordered regions (**Figure S7B**). We calculated biophysical properties using AntibodyForests and analyzed their correlation with distance to the germline in the lineage tree. Charge, hydrophobicity, and 3Di distance showed variable correlations with sequence distance (r^2 ranging between 0.13 and -0.48), depending on whether they were calculated for the full V(D)J sequence or just the binding residues of the antibody (**Figure S7C**). Additionally, we observed that the free energy of the full antibody-antigen complex increased with sequence distance (r^2 = 0.46) (**Figure S7C**). The RMSD to the germline showed no significant correlation with sequence distance to the germline (**Figure S7D**). However, when analyzing the structural change over each edge in the tree, we observed a strong correlation with the number of substitutions (R^2 = 0.67) (**Figure S7E**).

Concluding remarks

Taken together, AntibodyForests can both infer and visualize individual clonal lineages from single-cell V(D)J sequencing data and incorporate data such as bulk V(D)J sequences and transcriptional phenotypes. AntibodyForests can incorporate trees constructed with existing tools, and transform its bifurcating phylogenetic output to a more accurate multifurcating representation of the inferred antibody lineages. The package hosts various algorithms to compare evolution within and across repertoires using graph theory, protein language models, and protein structure. With the ongoing increase in immune repertoire sequencing data and the continuous improvement of deep learning strategies, we believe AntibodyForests is an invaluable toolbox to analyze B cell evolution and selection at both the clonal and repertoire levels.

Supplementary data

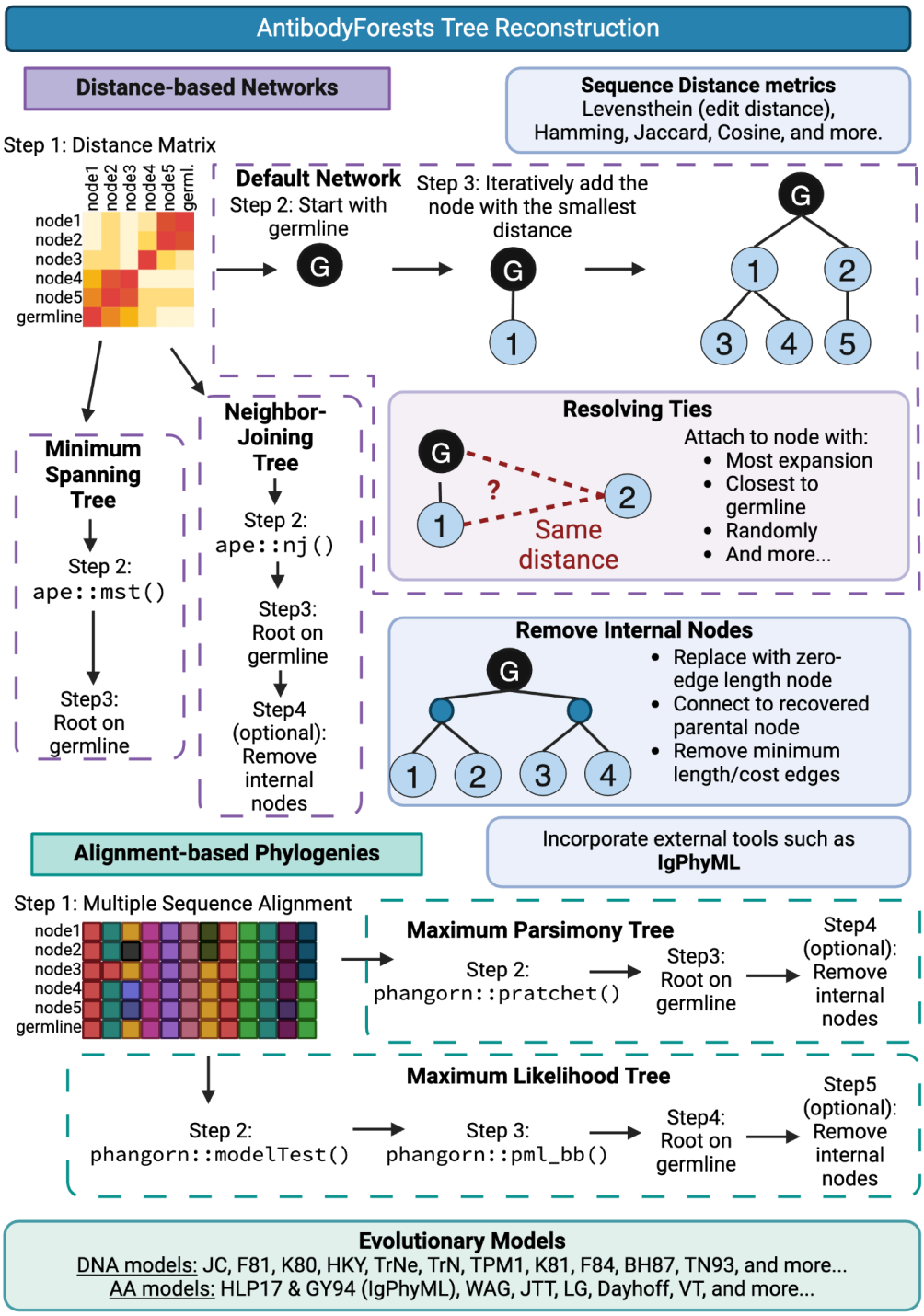

Figure S1. Overview of tree reconstruction algorithms in AntibodyForests. AntibodyForests contains five methods for lineage tree reconstruction and can integrate output

from external tools such as IgPhyML (15). AntibodyForests works with various metrics for sequence distance and a wide range of evolutionary substitution models. Additional algorithms are available to resolve ties and remove internal nodes during network construction. Created with biorender.com.

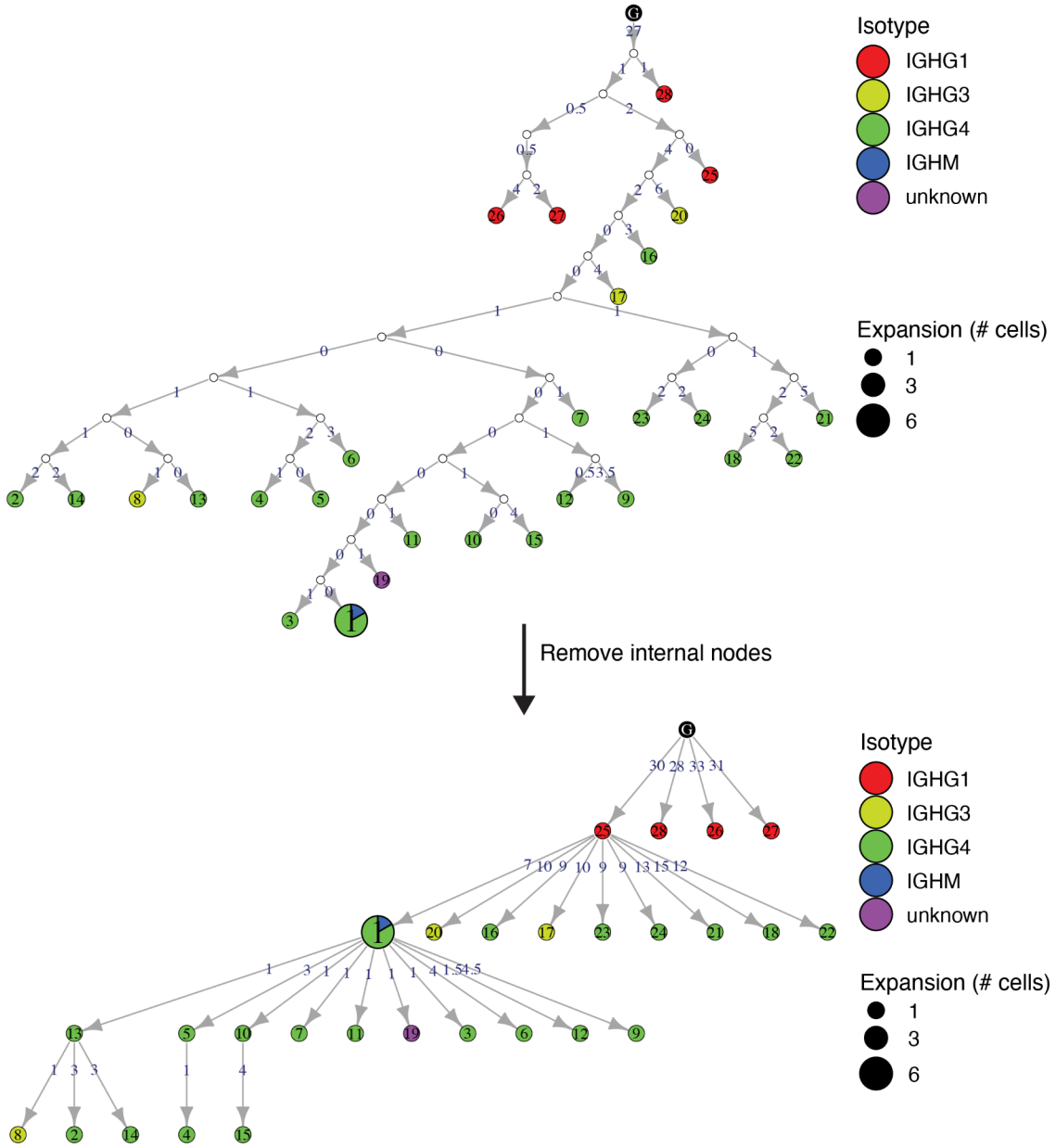

Figure S2. Examples of AntibodyForests trees. An example AntibodyForests object was generated using the maximum parsimony algorithm. The top tree is before internal node pruning, the bottom tree is after node pruning. The black node on top of the network refers to the germline as indicated by 10x Genomics, the white nodes (top tree) are internal nodes, and the colored nodes are recovered sequences. Node colors correspond to the isotypes. If multiple isotypes are presented within a node, the node is represented as a pie chart. Node size

corresponds to the number of cells with that unique heavy+light chain combination. Node labels can be matched to additional information within the AntibodyForests object. Edge labels correspond to the edit distance between the sequence of the attached nodes. Single-cell BCR data for this analysis was derived from Kim et al. (21)

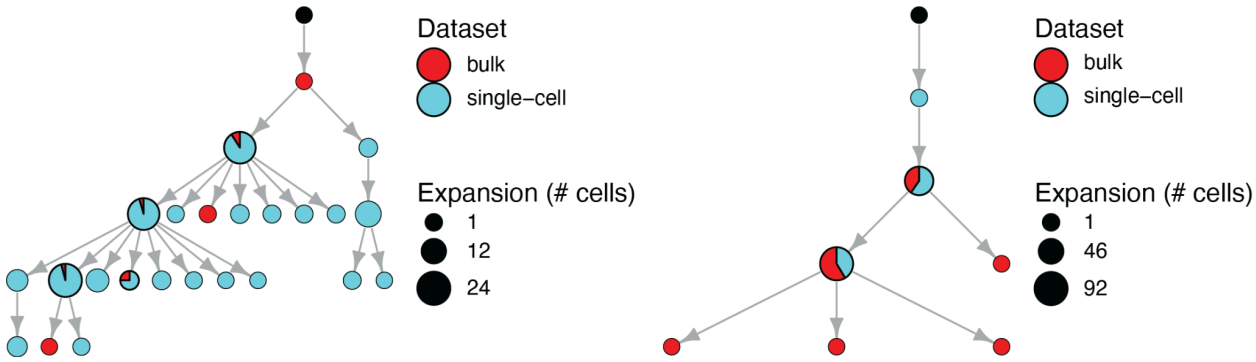

Figure S3. Example bulk RNA integration. An example AntibodyForests object was generated using both single-cell and bulk heavy chain BCR sequencing using the default parameter settings. The black nodes correspond to the reconstructed germlines. These two trees serve as examples to show the identical sequences between the bulk and single-cell datasets (pie chart nodes) and the integration of unique bulk sequences into the single-cell based clonotypes (red nodes). Single-cell and bulk BCR data for this analysis were derived from Neumeier et al. (20).

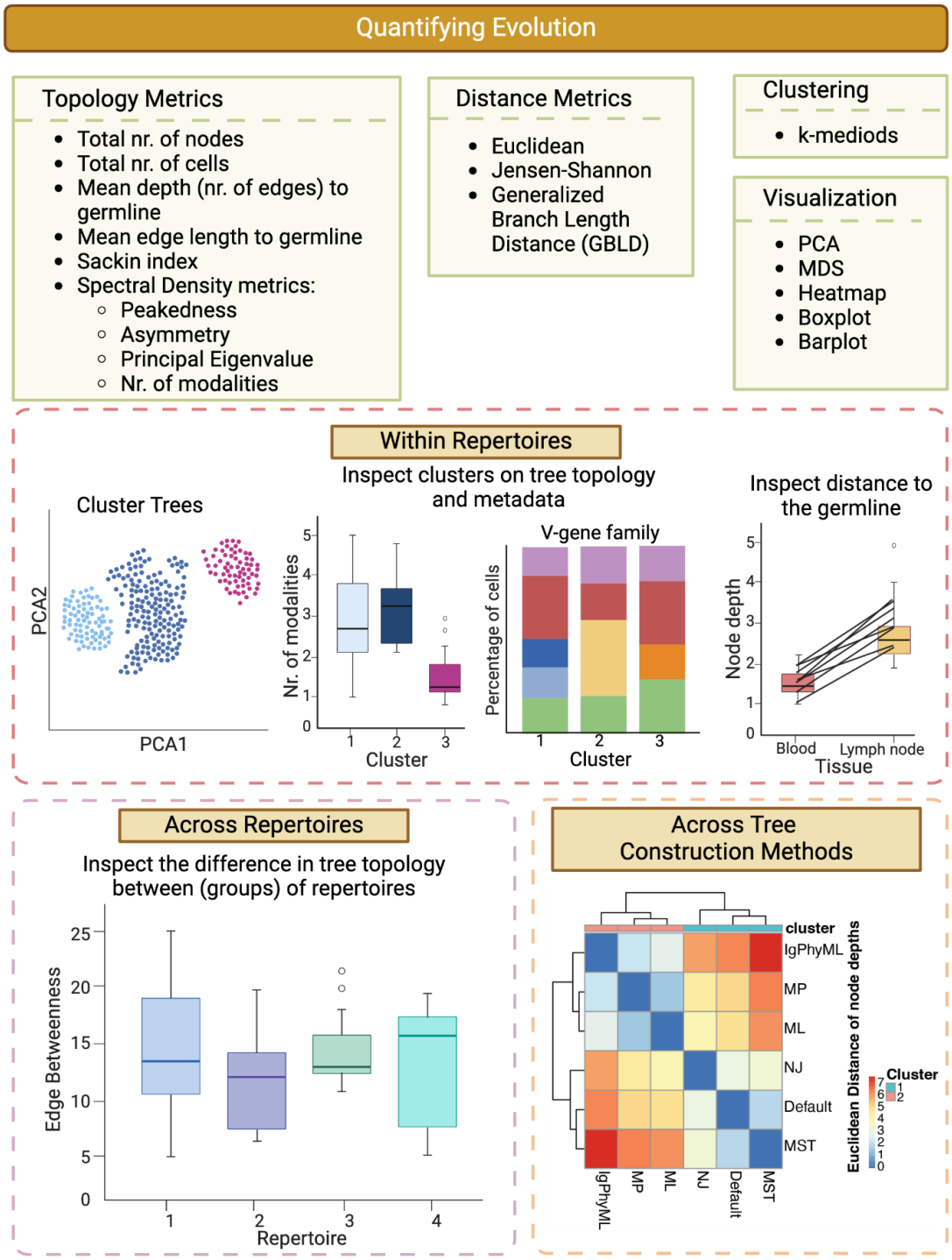

Figure S4. Overview of methods to quantify evolution. The package contains metrics to compare within and across repertoires or the same set of trees with different methods. Trees can be clustered based on various distance metrics and visualized in for example dimensionality reduction plots or heatmaps. Created with biorender.com.

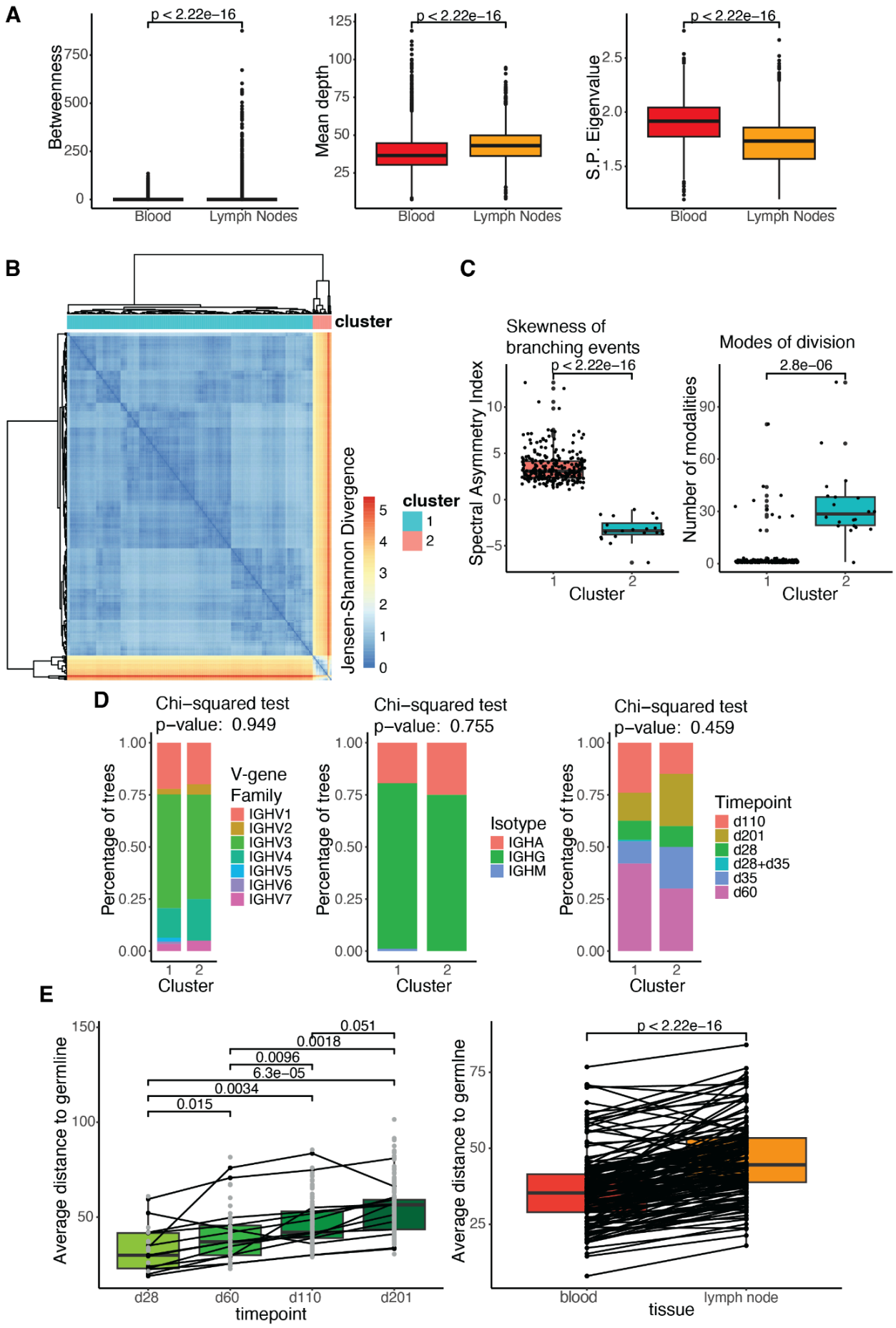
A
p < 2.22e−16
Betweenness
Blood
Lymph Nodes
Mean depth
S.P. Eigenvalue
B
cluster
Jensen-Shannon Divergence
C
Skewness of branching events
Spectral Asymmetry Index
Modes of division
2.8e−06
Number of modalities
Cluster
D
Chi−squared test
p−value: 0.949
Percentage of trees
V-gene Family
IGHV1
IGHV2
IGHV3
IGHV4
IGHV5
IGHV6
IGHV7
p−value: 0.755
Isotype
IGHA
IGHG
IGHM
p−value: 0.459
Timepoint
d110
d201
d28
d28+d35
d35
d60
E
Average distance to germline
0.051
0.0018
0.0096
6.3e−05
0.0034
0.015
d28
d60
d110
d201
timepoint
blood
lymph node
tissue

Figure S5. Quantifying antibody repertoire evolution example. B cell lineage trees were constructed with the default algorithm using the heavy+light chains. A) Comparison of tree topology metrics (node betweenness, average edge length to the germline, and spectral principal eigenvalue) between BCR repertoires in the blood and lymph nodes. B) Heatmap of the Jensen-Shannon divergence in Laplacian spectral density between trees. C) Spectral density metrics of trees in both clusters. D) Proportion of trees in each cluster and their predominant labels (V-gene family, isotype, and timepoint of sampling). E) Average distance (sum of edge lengths) from the nodes of each group to the germline. Grey dots represent trees not containing all groups. Single-cell BCR data for this analysis was derived from Kim et al. (21)

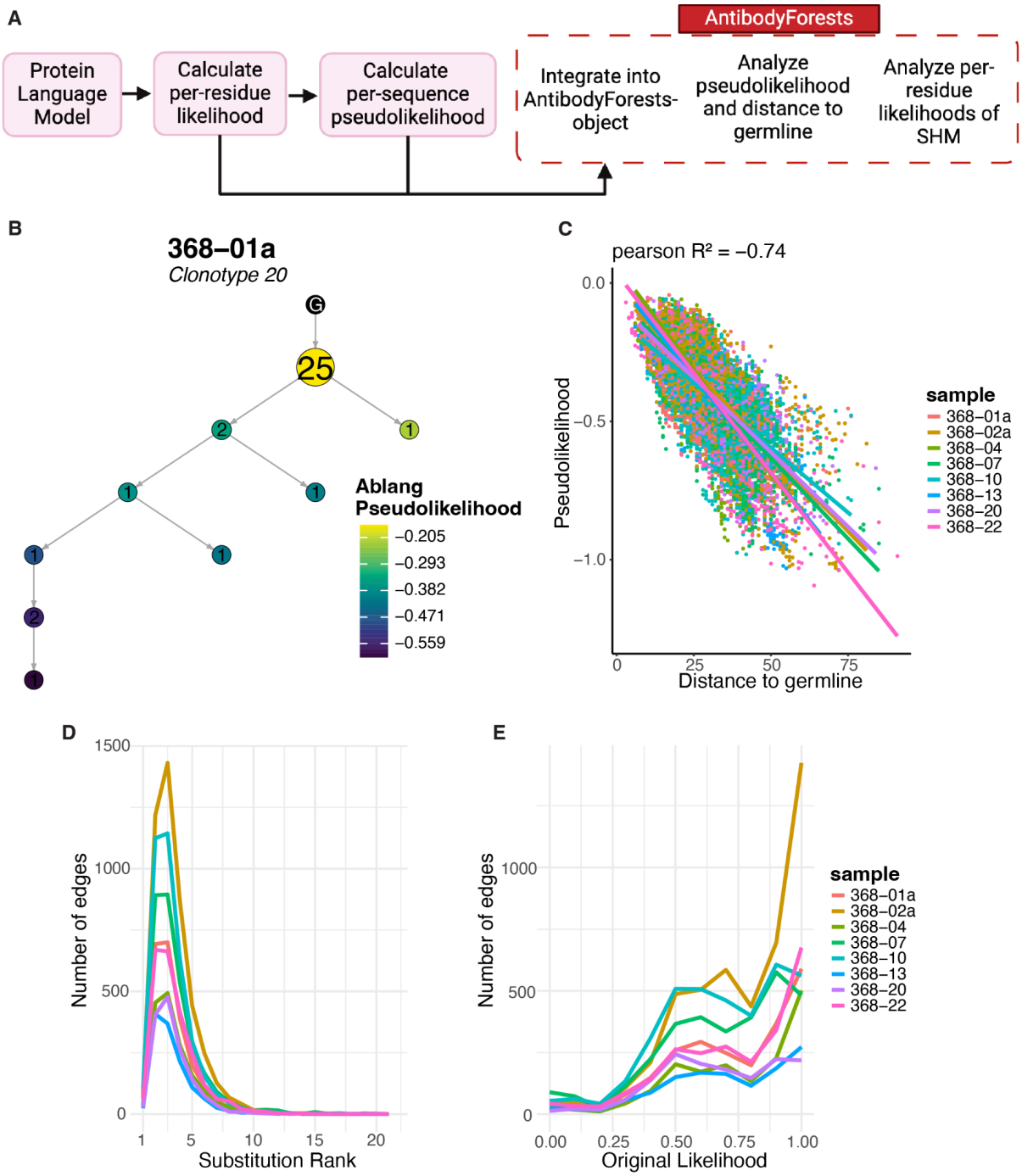

Figure S6. Protein language model (PLM) example. A) Overview of the PLM workflow integrated with AntibodyForests. After calculating the PLM-based likelihoods of the BCR sequences, AntibodyForests can be employed for integration and downstream analysis. B) An example tree colored on the heavy chain pseudolikelihood using the PLM Ablang (1). C) The correlation between the pseudolikelihood and the Levenshtein distance to the germline. D) The per-residue likelihood rank of the substitution along the edges of the trees. E) The per-residue likelihood of the original amino acid at the mutating positions along the edges of the trees. Single-cell BCR data for this analysis was derived from Kim et al. (21)

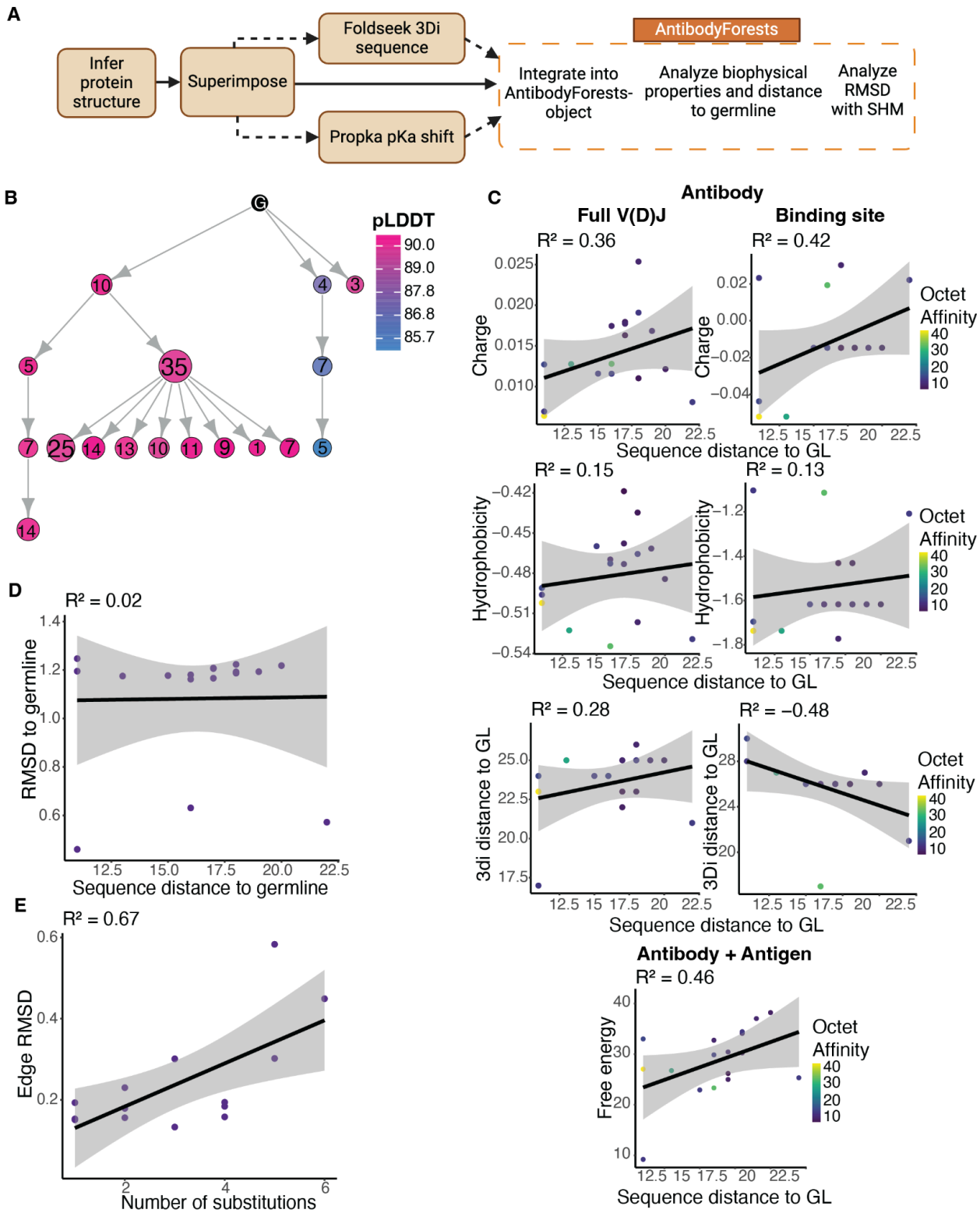

Figure S7. Structural evolution example. A) Overview of the antibody structure workflow within AntibodyForests. After inferring the 3D structure of the BCR sequences (and antigens), AntibodyForests can be employed for integration and downstream analysis. B) An example tree

colored on the average pLDDT score provided by AlphaFold 3 (37). C) The correlation between biophysical properties and the Levenshtein distance to the germline for the full length V(D)J sequence (left) and for the antibody-antigen binding residues (right), and the full antibody-antigen complex (bottom). D) The correlation between the root-mean-square deviation (RMSD) from the germline and the Levenshtein distance to the germline for the full length antibody. E) The correlation between the RSMD and the number of substitutions over each edge in the tree. Single-cell BCR and binding affinity data for this analysis were derived from Neumeier et al. (20)**.**

Online methods

Data analysis: For **Figures S2**, **S5,** and **S6**, the raw BCR reads from all samples generated by Kim et al. (21) were processed with 10X Genomics' Cell Ranger and the VDJ_build function of the Platypus package (version 3.6.0) in R (v4.4.0) (38). We applied the following parameters to the VDJ_build() function: trim.germlines = T, remove.divergent.cells = T, complete.cells.only = T. These parameters trim the germline to the V(D)J region, remove cells with more than 1 V(D)J transcript, and remove cells missing either a heavy or light chain transcript. Next, all samples per donor (retrieved at different time points after vaccination and from both the blood and lymph nodes), were reclonotyped using the Platypus function VDJ_clonotype_v3_w_enclone. We applied the following parameters to the VDJ_clonotype_v3_w_enclone() function: clone.strategy = "enclone" and global.clonotype = T. Under these parameters, the function uses 10X Genomics' Enclone tool to group cells from all samples by V- and J-gene and CDR3 length, and shared SHM with a CDR3 nucleotide identity of at least 85%. For **Figure S5A** we only reclonotyped the time points and kept the tissues as separate clonotypes. This reclonotyped VDJ-dataframe obtained with Platypus is then supplied to AntibodyForests.

For **Figure S2**, we subset the VDJ-dataframe to only include donor 368-10. We run the function Af_build to construct lineage trees for each clonotype with the following parameters: sequence.columns = c("VDJ_sequence_aa_trimmed","VJ_sequence_aa_trimmed"), germline.columns = c("VDJ_germline_aa_trimmed","VJ_germline_aa_trimmed"), and construction.method = "phylo.tree.mp". With these parameters, nodes in the trees represent both the heavy and light chain sequence, and the maximum parsimony method is used to construct the trees, by default the Levensthein distance is used as the distance metric. To create the figures, we use the function Af_plot_tree with the following parameters: sample = "368-10", clonotype = "clonotype40", label.by = "name", edge.label = "original", show.inner.nodes = F (for subfigure A) and show.inner.nodes = T (for subfigure B). With these parameters, the lineage tree of clonotype 40 is plotted, the nodes are labeled by their assigned node names from the AntibodyForests-object, the edges are labeled by Levensthein distance between the respective nodes, and the nodes are by default colored on isotype and sized by expansion.

For **Figure S5**, we run the function Af_build on all samples with the following parameters: sequence.columns = c("VDJ_sequence_aa_trimmed","VJ_sequence_aa_trimmed"), germline.columns = c("VDJ_germline_aa_trimmed","VJ_germline_aa_trimmed"), and construction.method = "phylo.network.default". For the across-repertoire comparison we run the function Af_compare_across_repertoire with the parameters: metrics = c("spectral.density",

"mean.edge.length", "betweenness"), plot = “boxplot”, significance = T. For clustering of the trees, we supply the AntibodyForests-object to the function Af_compare_within_repertoires with the following parameters: min.nodes = 20, distance.method = "jensen-shannon", clustering.method = "mediods", visualization.methods = "heatmap". With these parameters, we only consider trees with at least 20 nodes for topology analysis. We perform a k-mediods clustering on Jensen-Shannon divergence and create a heatmap. To inspect the difference in topology metrics between the clusters, we use the function Af_cluster_metrics with the parameters: clusters = output$clustering (output from Af_compare_within_repertoires), metrics = "spectral.density", min.nodes = 20, and significance = T. Next, additional node features from the VDJ-dataframe are added to the AntibodyForests-object with the function Af_add_node_feature. The influence of these parameters on the clustering is then analyzed with the function Af_cluster_node_features with the parameters: features = c("vgene_family", "timepoint", "isotype"), fill = "max", clusters = output$clustering (output from Af_compare_within_repertoires), and significance = T. These parameters take the most abundant node feature per tree in each cluster and calculates if there is a significant difference in cluster composition. Lastly, to analyze the potential difference in average distance to the germline of certain node features, we use the function Af_distance_boxplot with the parameters: distance = "edge.length", min.nodes = 10, node.feature = "timepoint" or “tissue”, groups = c("d28", "d60", "d110", "d201") or NA (when node.feature is “tissue”), significance = T, and unconnected = T. These create boxplot for the average sum of edge lengths per group per tree of at least 10 nodes and calculate the T-test p-value.

For **Figure S6**, we run the function Af_build() on all samples with the following parameters: sequence.columns = "VDJ_sequence_aa_trimmed", germline.columns = "VDJ_germline_aa_trimmed", and construction.method = "phylo.network.default". Next, we run the PLM Ablang (1) to construct probability matrices and calculate pseudolikelihoods using the code in this repository: https://github.com/dvginneken/PLM-pipeline. We use the function Af_add_node_feature to add the pseudolikelihoods to the AntibodyForests-object. The correlation between pseudolikelihood and distance to the germline was analyzed with the function Af_distance_scatterplot with the parameters: min.nodes = 5, color.by = “sample”, color.by.numberic = F, correlation = “pearson”. This creates a scatterplot of the pseudolikelihood of sequences in the trees with a least 5 nodes, the dots are categorically colored on the donor ID and a pearson correlation coefficient is calculated. Finally, mutations along the edges of the tree were analyzed by supplying the PLM probability matrices to the function Af_PLM_dataframe and plotted with Af_plot_PLM with the parameters: group_by = “sample_id” and values = “substitution_rank” or “original_probability”.

For Figure **S3**, the preprocessed BCR reads one of the mice from Neumeier et al. (20) were processed with the VDJ_build function of the Platypus package (version 3.6.0) in R (v4.4.0) (38). We applied the following parameters to the VDJ_build() function: trim.germlines = T, remove.divergent.cells = T, complete.cells.only = T. A data frame with the bulk transcript was integrated into this VDJ-dataframe with the AntibodyForests function VDJ_integrate_bulk using the parameters: organism = “mouse”, trim.FR1 = T, and tie.resolvement = “all”. This function annotates the single-cell and bulk transcripts using IgBLAST (39), trims the FR1 regions from all the sequences and reconstructed germline sequences to account for variation in primer design, and merges the bulk transcripts into the existing single-cell clonotypes based on identical CDR3

length, V and J gene usage, and CDR3 sequence similarity. We run Af_build on the integrated VDJ-dataframe with the parameters: sequence.columns = "VDJ_sequence_aa_trimmed", germline.columns = "VDJ_germline_aa_trimmed", construction.method = "phylo.network.default", parallel = F, node.features = c("dataset"). Next, we plot clonotype 3 and 5 using the function Af_plot_tree.
For Figure **S7**, we used the data from the largest binding IgG clone of Mouse1 from Neumeier et al. (20). A lineage tree was created in R (v4.4.0) with AntibodyForests Af_build using the parameters: sequence.column = c("VDJ_sequence_aa", "VJ_sequence_aa"), germline.columns = c("VDJ_germline_aa", "VJ_germline_aa"), node.features = c("octet.affinity"). Next, we predicted the antibody-antigen complex 3D structure of the heavy and light chain of each node together with the antigen Ovalbumin (SERPINB14) using AlphaFold3 (37). The resulting structures were superimposed on the C-alpha carbon positions. 3Di sequences were determined using mini3di (v0.2.1) (33) and pKa values and free energy were computed using propka (v3.5.1) (34), both in Python (v3.11.10). Structural properties were then calculated incorporated into AntibodyForests with the function VDJ_3d_properties using the parameters: properties = c("charge", "3di_germline", "hydrophobicity", "RMSD_germline", "pKa_shift", "free_energy", "pLDDT"), chain = "HC+LC" (for S7C left and S7E,F) or chain = "whole.complex" (for S7B,C right), sequence.region = "full.sequence" (for S7C left, S7D,E,F) or "binding.residues" (for S7C right.). These structural properties were added to the AntibodyForests object using the function Af_add_node_feature and distance to the germline was plotted with Af_distance_scatterplot with parameters: correlation = "pearson", color.by = "octet.affinity", color.by.numeric = T, geom_smooth.method = "lm". Figure S7F was created with the function Af_edge_RMSD.

Data visualization: Figures **1, S1, S4, S6A** and **S7A** were created with Biorender.com. Lineage trees in Figures **S2, S3, S6B**, and **S7B** were generated with the function Af_plot_tree. The heatmap of the across tree construction method in Figure **S4** was generated with Af_compare_methods. The heatmap in Figure **S5B** was generated with Af_compare_within_repertoires. The boxplots in Figure **S5C** were generated with Af_cluster_metrics. The barplots in Figure **S6C** were generated with Af_cluster_node_features. The boxplots in Figure **S5E** were generated with Af_distance_boxplot. The scatterplot in Figure **S6C, S7C,D,E** were generated with Af_distance_scatterplot. The distribution plots in Figures **S6D,E** were generated with Af_plot_PLM. The scatterplot in Figure S7F was generated with Af_edge_RMSD.

Data availability: Data for Figures **S2, S5** and **S6** was downloaded from the NCBI Sequence Read Archive using SRA Toolkit with ID PRJNA731610 (21). Data for Figures **S3** and **S7** was provided by Neumeier et al. (20)**.**

Funding: This work was supported by an SNSF Ambizione Grant (PZ00P3_208734 to A.Y.),
Competing interests: S.T.R. is a co-founder and holds shares of Engimmune Therapeutics AG and Encelta and Fy Cappa Biologics. S.T.R. may hold shares of Alloy Therapeutics. S.T.R. is on the scientific advisory board of Engimmune Therapeutics, Alloy Therapeutics, Encelta and Fy

Cappa Biologics. S.T.R. is a member of the board of directors for Engimmune Therapeutics and GlycoEra

References

1. Olsen TH, Moal IH, Deane CM. AbLang: an antibody language model for completing antibody sequences. Bioinform Adv. 2022 Jun 17;2(1):vbac046.

2. Ruffolo JA, Gray JJ, Sulam J. Deciphering antibody affinity maturation with language models and weakly supervised learning [Internet]. 2021 [cited 2025 Jan 16]. Available from: http://arxiv.org/abs/2112.07782

3. Singh R, Im C, Qiu Y, Mackness B, Gupta A, Joren T, et al. Learning the language of antibody hypervariability. Proceedings of the National Academy of Sciences. 2025 Jan 7;122(1):e2418918121.

4. van Ginneken D, Samant A, Daga-Krumins K, Agrafiotis A, Kladis E, Reddy ST, et al. Protein language model pseudolikelihoods capture features of in vivo B cell selection and evolution [Internet]. bioRxiv. 2024. Available from: http://biorxiv.org/lookup/doi/10.1101/2024.12.09.627494

5. Leem J, Mitchell LS, Farmery JHR, Barton J, Galson JD. Deciphering the language of antibodies using self-supervised learning. Patterns (N Y). 2022 Jul 8;3(7):100513.

6. Hie BL, Shanker VR, Xu D, Bruun TUJ, Weidenbacher PA, Tang S, et al. Efficient evolution of human antibodies from general protein language models. Nat Biotechnol. 2024 Feb;42(2):275–83.

7. Barak M, Zuckerman NS, Edelman H, Unger R, Mehr R. IgTree: creating Immunoglobulin variable region gene lineage trees. J Immunol Methods. 2008 Sep 30;338(1-2):67–74.

8. Hoehn KB, Lunter G, Pybus OG. A Phylogenetic Codon Substitution Model for Antibody Lineages. Genetics. 2017 May;206(1):417–27.

9. Gupta NT, Vander Heiden JA, Uduman M, Gadala-Maria D, Yaari G, Kleinstein SH. Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. Bioinformatics. 2015 Oct 15;31(20):3356–8.

10. DeWitt WS 3rd, Mesin L, Victora GD, Minin VN, Matsen FA 4th. Using Genotype Abundance to Improve Phylogenetic Inference. Mol Biol Evol. 2018 May 1;35(5):1253–65.

11. Jensen CG, Sumner JA, Kleinstein SH, Hoehn KB. Inferring B Cell Phylogenies from Paired H and L Chain BCR Sequences with Dowser. J Immunol. 2024 May 15;212(10):1579–88.

12. Hoehn KB, Pybus OG, Kleinstein SH. Phylogenetic analysis of migration, differentiation, and class switching in B cells. PLoS Comput Biol. 2022 Apr;18(4):e1009885.

13. Davidsen K, Matsen FA IV. Benchmarking tree and ancestral sequence inference for B cell receptor sequences [Internet]. bioRxiv; 2018. Available from: http://dx.doi.org/10.1101/307736

14. Hoehn KB, Kleinstein SH. B cell phylogenetics in the single cell era. Trends Immunol. 2024 Jan;45(1):62–74.

15. Hoehn KB, Vander Heiden JA, Zhou JQ, Lunter G, Pybus OG, Kleinstein SH. Repertoire-wide phylogenetic models of B cell molecular evolution reveal evolutionary signatures of aging and vaccination. Proc Natl Acad Sci U S A. 2019 Nov 5;116(45):22664–72.

16. Yaari G, Vander Heiden JA, Uduman M, Gadala-Maria D, Gupta N, Stern JNH, et al. Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. Front Immunol. 2013 Nov 15;4:358.

17. Mirsky A, Kazandjian L, Anisimova M. Antibody-specific model of amino acid substitution for immunological inferences from alignments of antibody sequences. Mol Biol Evol. 2015 Mar;32(3):806–19.

18. Farnia M, Tahiri N. New generalized metric based on branch length distance to compare B cell lineage trees. Algorithms for Molecular Biology. 2024 Oct 5;19(1):1–14.

19. Wu X, Zhou T, Zhu J, Zhang B, Georgiev I, Wang C, et al. Focused Evolution of HIV-1 Neutralizing Antibodies Revealed by Structures and Deep Sequencing. Science [Internet]. 2011 Sep 16 [cited 2024 Dec 17]; Available from: https://www.science.org/doi/10.1126/science.1207532

20. Neumeier D, Yermanos A, Agrafiotis A, Csepregi L, Chowdhury T, Ehling RA, et al. Phenotypic determinism and stochasticity in antibody repertoires of clonally expanded plasma cells. Proc Natl Acad Sci U S A. 2022 May 3;119(18):e2113766119.

21. Kim W, Zhou JQ, Horvath SC, Schmitz AJ, Sturtz AJ, Lei T, et al. Germinal centre-driven maturation of B cell response to mRNA vaccination. Nature. 2022 Feb 15;604(7904):141–5.

22. Liao HX, Lynch R, Zhou T, Gao F, Alam SM, Boyd SD, et al. Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. Nature. 2013 Apr 3;496(7446):469–76.

23. Cizmeci D, Lofano G, Rossignol E, Dugast AS, Kim D, Cavet G, et al. Distinct clonal evolution of B-cells in HIV controllers with neutralizing antibody breadth. Elife [Internet]. 2021 Apr 12;10. Available from: http://dx.doi.org/10.7554/eLife.62648

24. Lewitus E, Morlon H. Characterizing and Comparing Phylogenies from their Laplacian Spectrum. Syst Biol. 2016 May;65(3):495–507.

25. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci U S A [Internet]. 2021 Apr 13;118(15). Available from: http://dx.doi.org/10.1073/pnas.2016239118

26. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. IEEE Trans Pattern Anal Mach Intell. 2022 Oct;44(10):7112–27.

27. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of

atomic-level protein structure with a language model. Science. 2023 Mar 17;379(6637):1123–30.

28. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug;596(7873):583–9.

29. Hie BL, Yang KK, Kim PS. Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. Cell Syst. 2022 Apr 20;13(4):274–85.e6.

30. Olsen TH, Moal IH, Deane CM. Addressing the antibody germline bias and its effect on language models for improved antibody design. Bioinformatics [Internet]. 2024 Nov 1;40(11). Available from: http://dx.doi.org/10.1093/bioinformatics/btae618

31. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science [Internet]. 2021 Aug 20 [cited 2025 Jan 20]; Available from: https://www.science.org/doi/10.1126/science.abj8754

32. Ruffolo JA, Chu LS, Mahajan SP, Gray JJ. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. Nature Communications. 2023 Apr 25;14(1):1–13.

33. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, et al. Fast and accurate protein structure search with Foldseek. Nature Biotechnology. 2023 May 8;42(2):243–6.

34. Olsson MHM, Søndergaard CR, Rostkowski M, Jensen JH. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. J Chem Theory Comput. 2011 Feb 8;7(2):525–37.

35. Abanades B, Georges G, Bujotzek A, Deane CM. ABlooper: fast accurate antibody CDR loop structure prediction with accuracy estimation. Bioinformatics (Oxford, England) [Internet]. 2022 Mar 28 [cited 2025 Jan 20];38(7). Available from: https://pubmed.ncbi.nlm.nih.gov/35099535/

36. Ruffolo JA, Sulam J, Gray JJ. Antibody structure prediction using interpretable deep learning. Patterns (New York, NY) [Internet]. 2021 Dec 9 [cited 2025 Jan 20];3(2). Available from: https://pubmed.ncbi.nlm.nih.gov/35199061/

37. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024 May 8;630(8016):493–500.

38. Cotet TS, Agrafiotis A, Kreiner V, Kuhn R, Shlesinger D, Manero-Carranza M, et al. ePlatypus: an ecosystem for computational analysis of immunogenomics data. Bioinformatics [Internet]. 2023 Sep 2;39(9). Available from: http://dx.doi.org/10.1093/bioinformatics/btad553

39. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W34–40.