Supplementary Material:

BRepertoire: A user-friendly web server for analysing antibody repertoire data

Christian Margreitter[*], Hui-Chun Lu[†], Catherine Townsend[‡], Alexander Stewart[§], Deborah Dunn-Walters[§] and Franca Fraternali[*]

March 26, 2018

List of Tables

List of Figures

[*]Randall Centre for Cell & Molecular Biophysics, King's College London
[†]Cell and Developmental Biology, University College London
[‡]Department of Immunobiology, King's College London
[§]Faculty of Health and Medical Sciences, University of Surrey

Table S1: List of features in the "IMGT" and "Calculation" branches. Each line represents one tab. In the third column, input and output artefacts are stated. In addition to the resources mentioned in the main manuscript, references 1–7 have been used in the implementation process.

IMGT branch

Upload	Implements the upload of IMGT/V-Quest files (*.tar.gz) and allows selection of desired columns from the 11 files comprising the V-Quest format. Enabled columns are combined into a data table. The table can be downloaded and used in the context of both the "Calculate" and "Analysis" branches. Successful upload enables the "Annotation" tab.	IMGT input Output table
Annotation	Optional function to extend a given table by additional columns uploaded as a data table. The first column serves to match the observations.	Output table

Calculation branch

Upload	Allows the upload of a data table (*.csv), with observations in rows and features in columns, respectively. May be obtained from the "IMGT" branch. Successful upload enables the following tabs.	Input table
Extract	Implements two different methods to extract parts from contents in one column and to store these in a new column, attached at the rear of the table.	Output table
Calculation	Implements the calculation of 23 physico-chemical properties (see main text for details) for amino acid sequences. The necessary columns are attached for every row (observation) at the rear or the table. Columns already present are ignored.	Output table
Clonotype clustering	Allows the clustering of observations into clonotypes based on the distances between nucleotide sequences and hierarchical clustering. In addition, a representative observation per cluster can be obtained, if required.	Output table

Table S2: List of features in the "Analysis" branch. Each line represents one tab. In the third column, input and output artefacts are stated.

Analysis branch		
Upload	Allows the upload of a data table (*.csv), with observations in rows and features in columns, respectively. May be obtained from the "IMGT" or "Calculation" branches. Successful upload enables the following tabs.	Input table
Select	Allows the selection of columns which are maintained for further analysis. Useful to reduce the complexity of a dataset with very many columns. Of course, all columns can be selected.	Output table
Filter	Optional feature, that implements the filtering of columns for certain values. Numerical values can be selected by specifying a range and nominal data by enabling levels.	Output table
Grouping	To allow comparisons within the data provided, at least one column has to be set as grouping column. The data can be split by the levels (numerical or ordinal) in the(se) column(s) in the subsequent analysis steps. The selected columns can be ordered individually.	Grouping table
Boxplot	Shows the selected numerical property for every level in the grouping column in a box-and-whisker representation. The line in the middle represents the median, the box limits the upper and lower quartiles, respectively and the "whiskers" represent the remaining ones (outliers excluded).	Input table p-values Box data Plot
Barplot	Shows the absolute occurrences or relative frequencies of the selected property for every level in the grouping column(s) or over the entire data set. Non-numerical columns (e.g. sequences) can be transformed if necessary. Different targets can be specified for the normalization (by group, over all data, ...).	Input table Output table Plot
Distribution analysis	This function allows to specify two different datasets, which can be compared in terms of five statistical tests (p-values) and three effect size measures. Multiple properties can be selected for simultaneous calculation and the resulting plot will hold two histograms per property, representing the two data sets.	Input table Statistics t. Eff. size t. Protocol Plot

Analysis branch (continuation)

PCA	This function performs a linear transformation on the specified data to maximise the (independent) variance along so-called principal components, which is frequently used for dimensionality reduction of data sets. A minimum of two properties has to be selected. Plotting options are the calculation of means and representing the spread as error bars and / or ellipses.	Input table Output table Rotations Plot
V(D)J gene usage	A critical task in the context of antibody repertoires is the combinatorial frequency with which V, D and J genes occur. This function offers 1D, 2D (circle) and 3D (bubble) plots to show in which proportion each combination occurs.	Output table Plot
Dendrogram	To show the distance between various (sub-)groups in antibody repertoire data, tree-based dendrograms are frequently used. Multiple properties can be specified, which are usually scaled and centred before the calculation starts.	Input table Dist. matrix Plot
t-SNE	A rather new dimensionality reduction algorithm [8] that tries to optimize local and global features in the data simultaneously. The algorithm requires the appropriate adjustment of various hyper-parameters. This calculation is computationally very demanding and will take, depending on the size of the data set, quite some time.	Input table Output table t-SNE results Plot

Table S3: Description of the two (real) datasets used for the demonstration of the server's functions.

Vaccination dataset

The vaccination dataset, obtained by the group of Deborah Dunn-Walters [9] in 2012 ("vaccination") contains B cell repertoire data of six young (aged 19-45) and six elderly (aged 70-89) healthy volunteers. Three samples per donor have been taken: The first prior to vaccination (with Influvac and Pneumovax II) called `Day 0`, the next seven days later (`Day 7`) and the last one 28 days after vaccination (`Day 28`). This allows for a time-resolved monitoring of the immune response for the two different age groups. In total, the data set (as downloaded) contains 45784 observations. This file can be obtained from http://doi.org/10.5281/zenodo.1161143. The largest three clonotypes in this dataset are of a size of 527, 456 and 395, respectively (when clustered using Levenshtein distance and a cut-off of 0.18).

PBMC dataset

The peripheral blood mononuclear cells ("PBMC") dataset has been established by Deborah Dunn-Walters and co-workers and is, as of yet, not publicly available [10]. It contains PBMCs isolated from the repertoires of six young (aged 21-45) and eight elderly (aged 62-87) healthy volunteers. In total, the data set (as used) contains 51909 observations, all of which are representative reads for their respective clones (as described in the main text). The largest three clonotypes in this dataset are of a size of 847, 522 and 309, respectively.

Table S4: Kullback-Leibler divergence [11] results computed for figure 2 (main manuscript) and figure S7. The divergence for the `Young` group (`Day 0` versus `Day 7`) is much higher than the corresponding values for the `Old` one, which is also reflected in the respective figures.

		Young			Old		
		Day 0	Day 7	Day 28	Day 0	Day 7	Day 28
Young	Day 0	0.0000	0.2054	0.0424	0.0587	0.1808	0.0994
	Day 7	0.2162	0.0000	0.2206	0.3035	0.2419	0.2899
	Day 28	0.0394	0.1740	0.0000	0.0245	0.1068	0.0467
Old	Day 0	0.0494	0.2234	0.0249	0.0000	0.1118	0.0469
	Day 7	0.1357	0.2086	0.0954	0.1054	0.0000	0.1371
	Day 28	0.0705	0.1786	0.0363	0.0412	0.1102	0.0000

Table S5: Effect sizes (Cliff's Δ) calculated for the comparison of `IGHV2` and `IGHV3` with the other V gene families respectively (for all ten Kidera factors). The same filtering (excluding `IGHV7` and CDR3H loops longer than 35 amino acids) has been applied prior to calculation as for figures 3 and 4 (main manuscript). For the subsequent analysis of family `IGHV2`, Kidera factors with an effect size $\geq$ 0.1 or -0.1 have been used (supplementary figure S10). The short description of the Kidera factors has been taken from reference 12.

	Cliff's Δ		
Property	IGHV2	IGHV3	Description
Kidera 1	-0.04	-0.02	Helix / bend preference
Kidera 2	0.11	-0.04	Side-chain size
Kidera 3	0.04	-0.07	Extended structure preference
Kidera 4	-0.12	0.05	Hydrophobicity
Kidera 5	0.14	0.12	Double-bend preference
Kidera 6	0.13	-0.11	Partial specific volume
Kidera 7	-0.18	-0.01	Flat extended preference
Kidera 8	-0.02	0.01	Occurrence in α region
Kidera 9	0.22	-0.04	pK-C
Kidera 10	0.00	0.01	Surrounding hydrophobicity

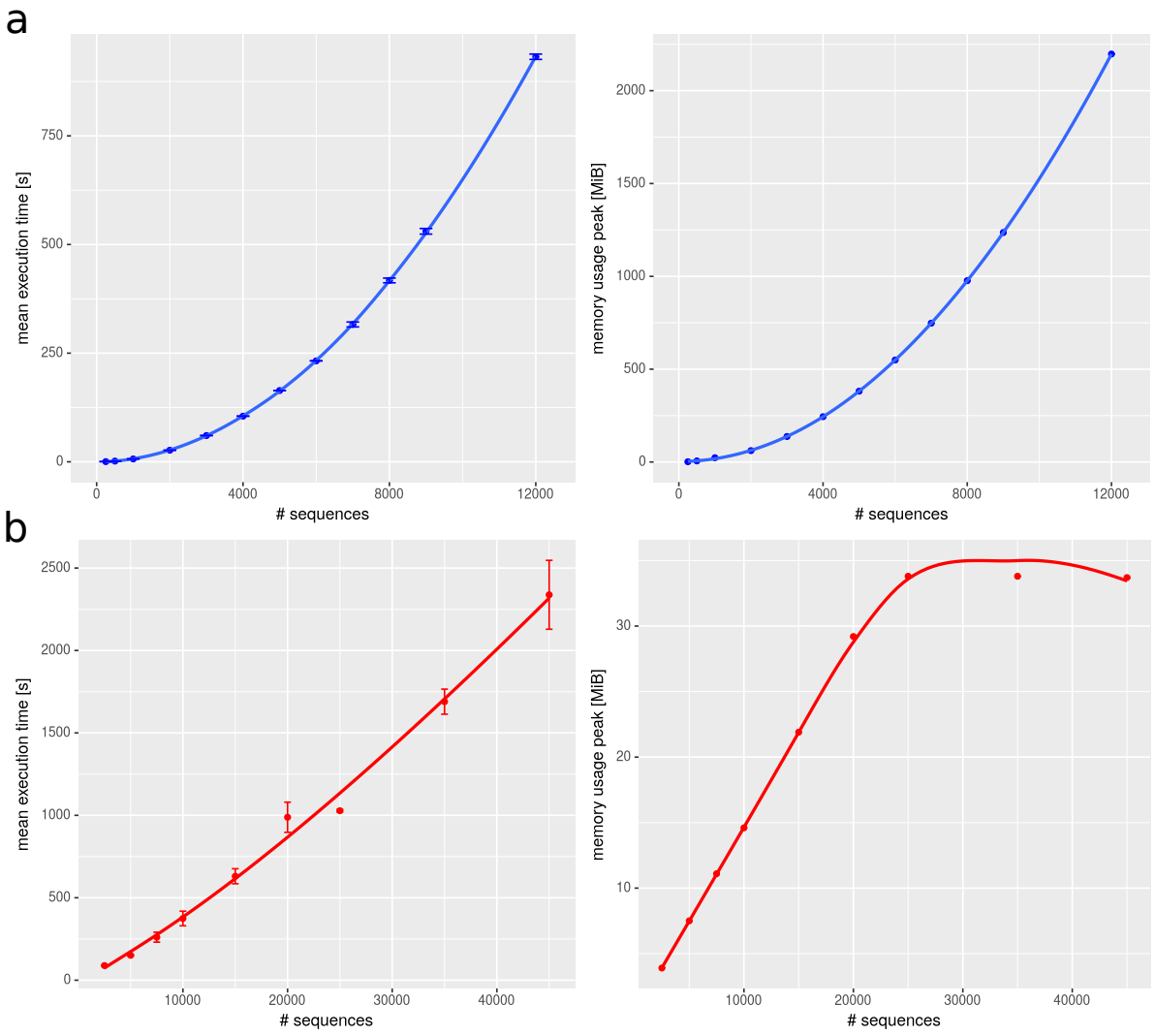

Figure S1: Benchmarks for clonotype clustering (blue) and t-SNE (red), showing runtime requirements (in seconds) and the maximum random-access memory (RAM) allocation (in mebibyte, MiB) during execution (depending on the input size). The runtime is averaged over three trials and shown together with the associated standard deviation (error-bars). In **(a)**, the construction of the distance matrix (the time-limiting step of the clonotype clustering) is shown to increase quadratically with the size of the input, $O(n^2)$, in both execution time and memory requirement. The adjusted R^2 values as calculated by the R function `lm()` are 0.9 for a quadratic model fitted to the points. In conclusion, partitioning the data (see main manuscript and tutorials) may improve the speed of this calculation tremendously. For real data (about 100000 reads, split into comparably large partitions) one could expect the clustering to be completed within one to two hours. In **(b)**, the t-SNE calculation's runtime complexity is proven to be of O(n*log(n)) (the adjusted R^2 value is 0.99), which is achieved by the Barnes-Hut approximation used in the algorithm [4]. In this example, 1000 iterations and ten dimensions (Kidera factors) have been used. The maximum memory requirement at

any given time does not exceed about 35 MiB using our parameter settings. For large data (about 100000 reads, ten dimensions, 3000 iterations) one would expect the t-SNE calculation to complete within 5 hours. Note, that the variation between the individual trials for t-SNE is much higher compared to the clonotype clustering, since the precise execution of the algorithm differs significantly depending on the initial seed set.

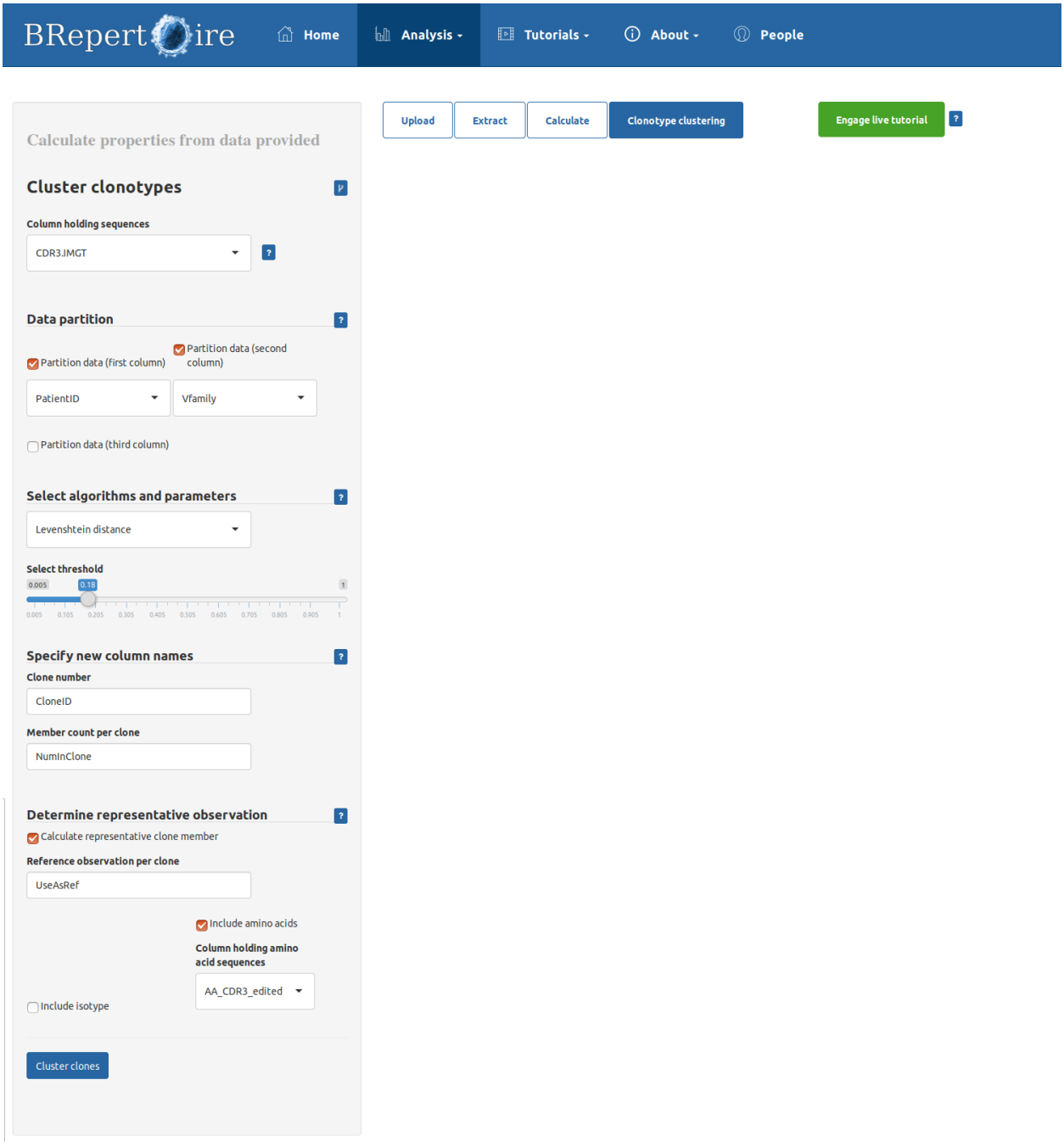

Figure S2: Clonotype clustering interface. Clustering usually is performed using DNA rather than amino acid sequences due to the higher information content. Prior to the calculation of the distance matrix and the following clustering, it might be necessary to split the data to reduce the size of the individual subsets. This helps in speeding up the calculation and meliorates the memory requirement (see also supplementary figure S1). In the Dunn-Walters group, data is usually split by sample or patient ID and the V gene family. The latter is done in order to include also members in a clone that may have a wrong V gene assignment due to hypermutation. However, it is worth mentioning that other groups use the less conservative V gene partitioning instead, which will increase the speed of the calculation dramatically

due to the much smaller partitions. Hierarchical clustering, as applied by this server, requires the specification of a cut-off threshold, by which the tree is cut in order to group the clones. From our experience, we propose 0.18 and 0.05 for heavy and light chain CDR3 sequences for B cell repertoires as meaningful defaults. The server attaches two columns to every observation in the data set, holding the clone ID and the number of members for each clone. Moreover, in order to select a representative, typical observation for every clone, a score is calculated internally by ranking the observed amino acid sequences and classes by their abundance in a given clone. It is also possible, however, to simply select the first member of each cluster. If this "representative" feature is activated, an additional column will be added to the data set, holding either the values `TRUE` (if an observation has been designated to represent the clone) or `FALSE`. We refer also to the (online) tutorials and the tooltips for the interface descriptions (both available at the server's address, http://mabra.biomed.kcl.ac.uk/BRepertoire).

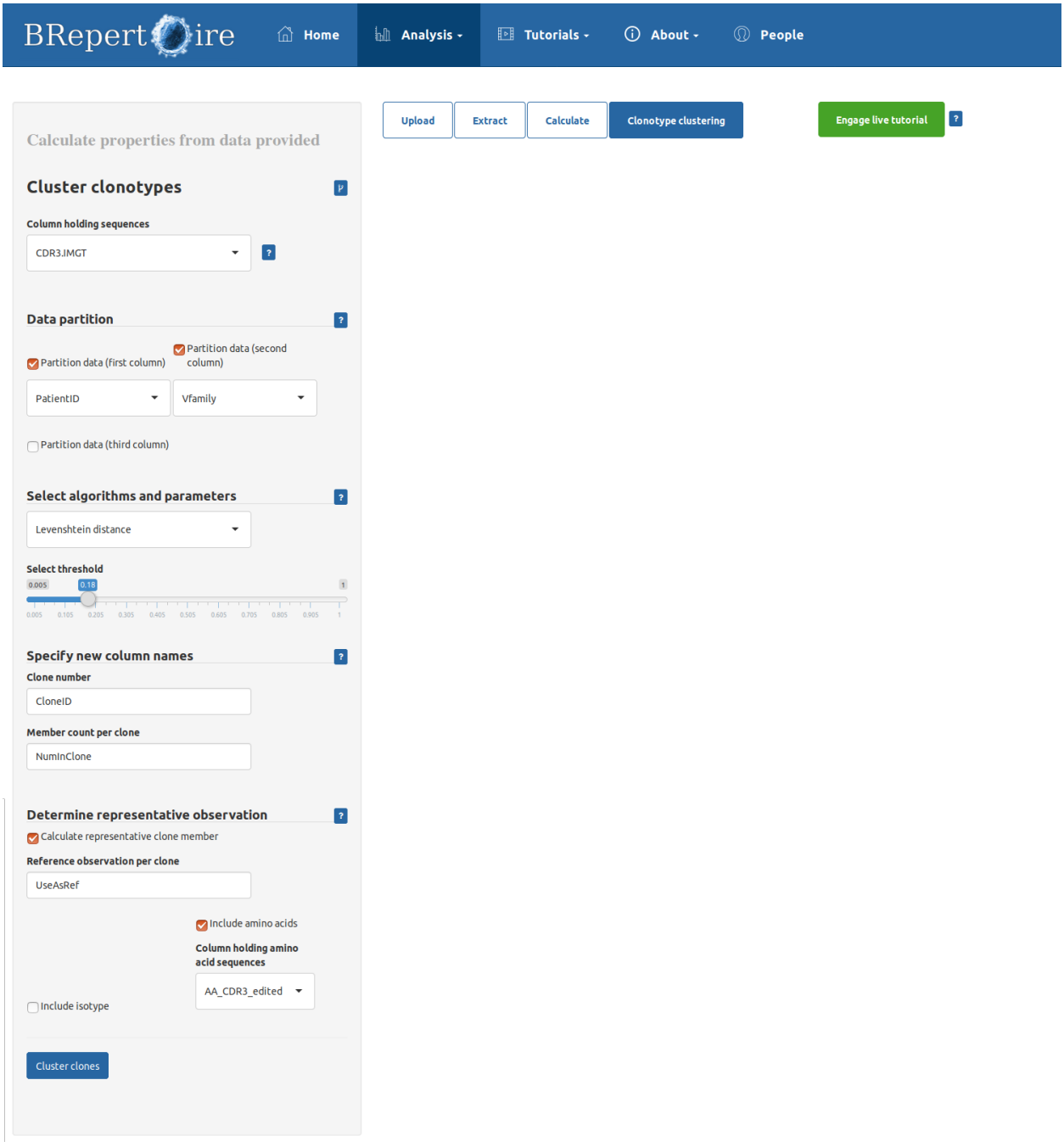

Figure S3: Distribution analysis interface. The statistical tests currently supported are the t-test, the Wilcoxon Rank Sum test (WRST) [13], the Kolmogorov-Smirnov [14] (K-S) and two types of permutational analyses, using the permutational central limit theorem (pclt) and a monte-carlo (mc) implementation [15]. Since a t-test requires the assumption of normally distributed sample means, WRST and K-S have been implemented as alternatives. Moreover, WRST is not sensitive to changes in the shape (only to changes in the median).

If, however, only little knowledge is available on the distribution of the data, the permutational methods might be used. To this date, tests for statistical significance have been often misused [16, 17], predominantly because of the misconception that a p-value below 0.05 proves H_0 false and thereby confirms the initial theory. In order to strengthen reproducibility [18] and to quantify the size of probable effects and confidence intervals, effect size measures can be used. BRepertoire offers three ways to calculate effect sizes: Cohen's d [19], Hedge's g [20] and Cliff's Δ [21]. Note, that the latter uses ranking in contrast to the others and may be used as default.

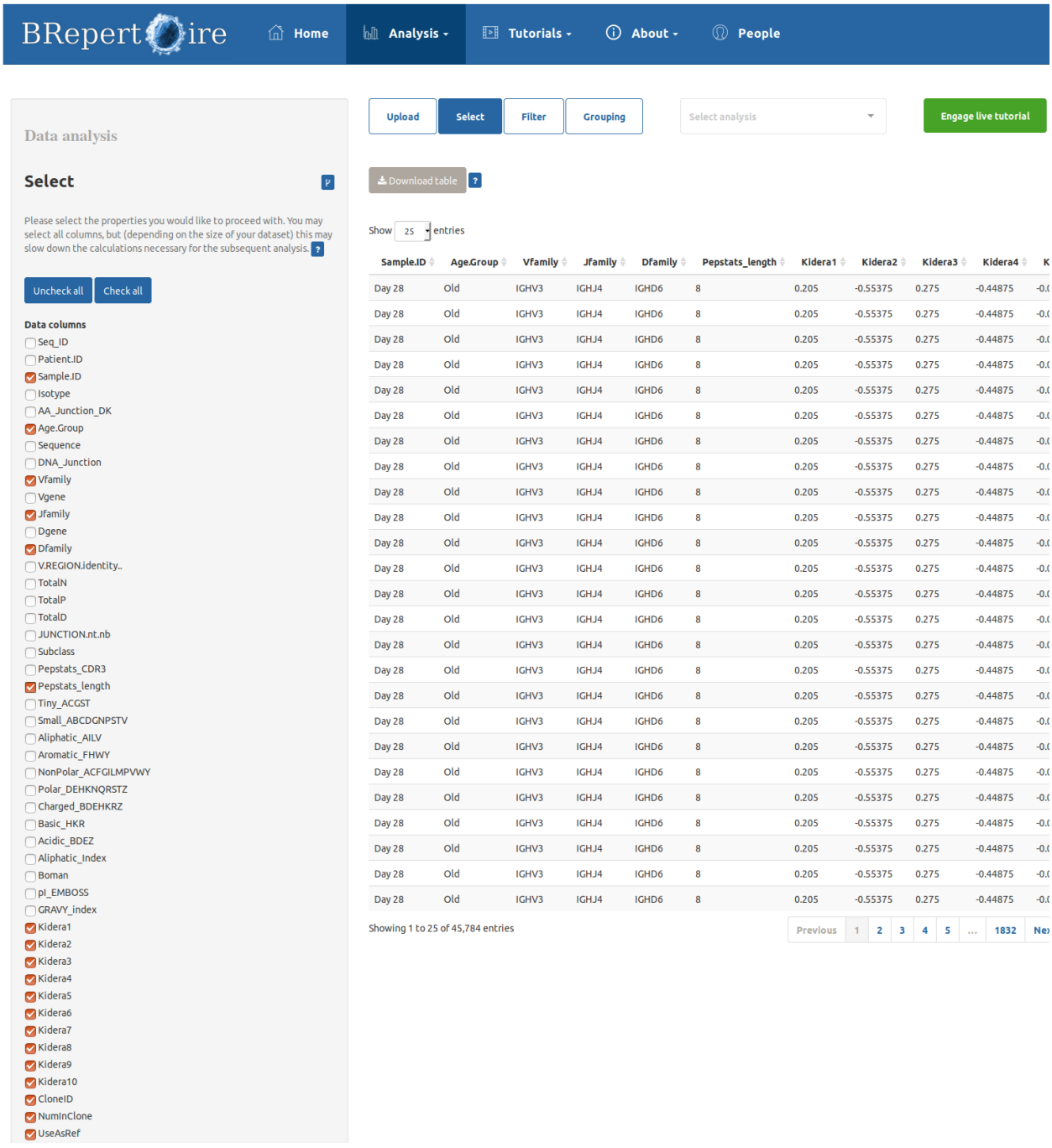

Figure S4: Select columns interface. In order to reduce large datasets to manageable sizes, only selected columns will be maintained. The table on the right hand side is automatically updated.

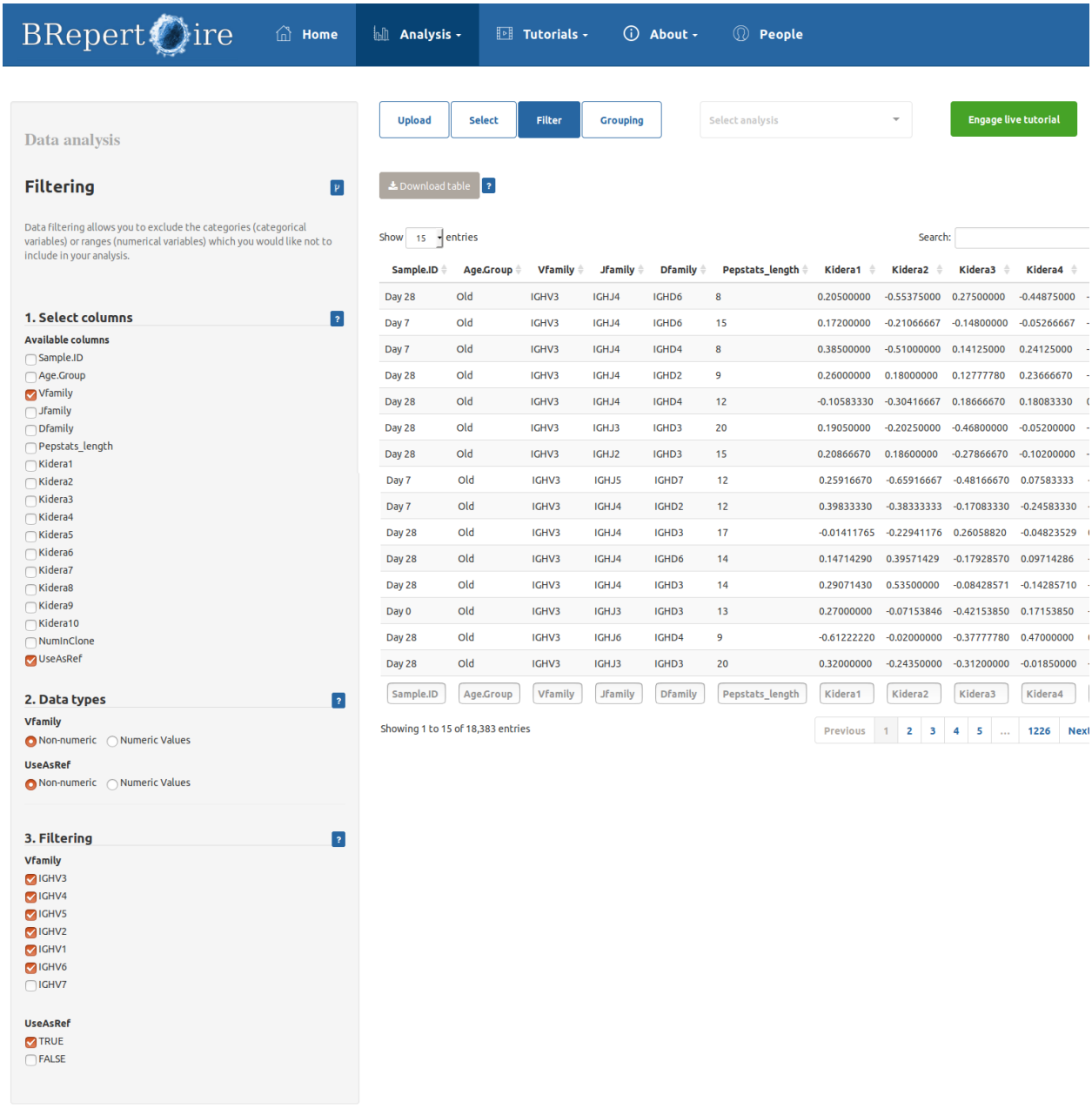

Figure S5: Filter values interface. This tab operates in three steps: First, the columns for which certain values need to be filtered out are selected. Then the data type can be adjusted if the server's guess is wrong. And finally, either checkbox groups (nominal data) or range input sliders (numerical data) can be used to select certain values. The number of remaining observations is displayed right under the (automatically updated) table on the right hand side.

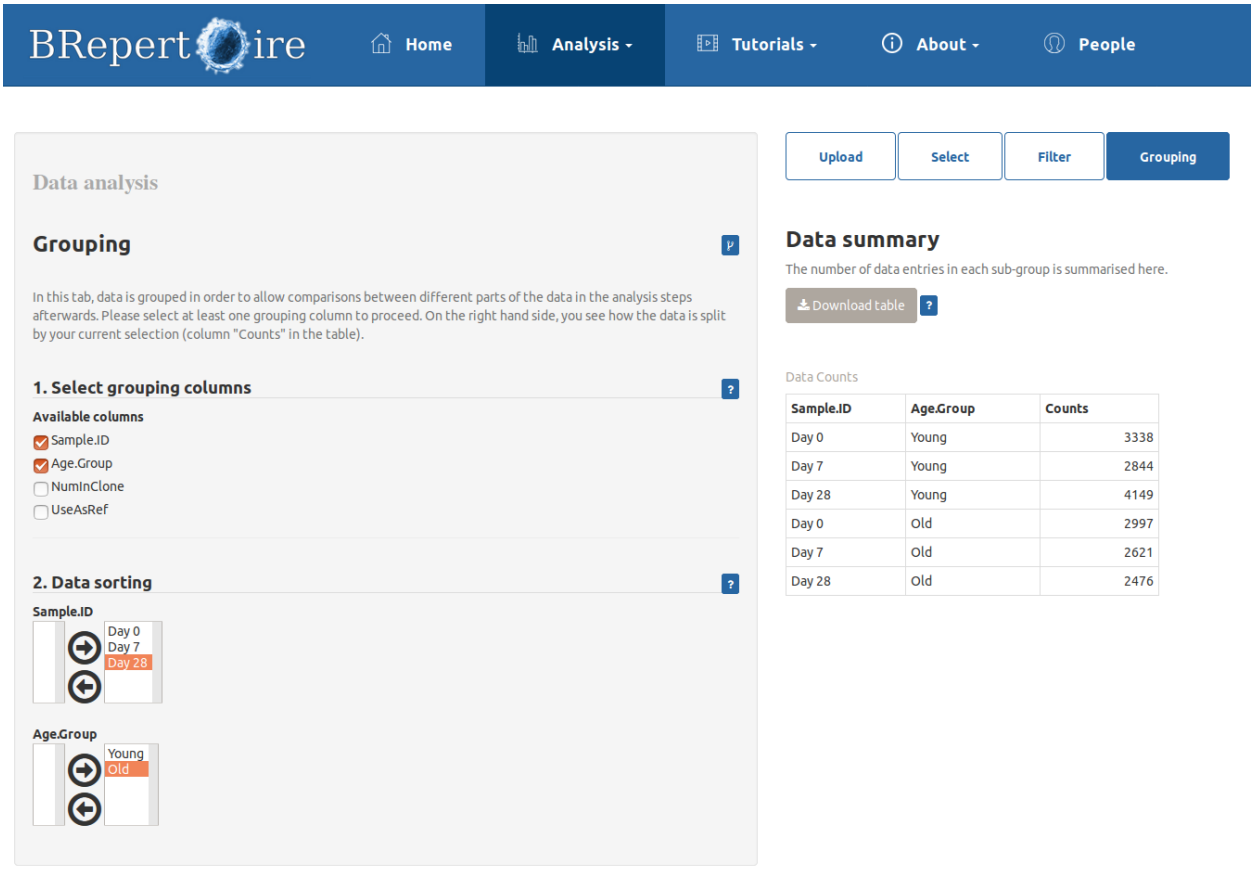

Figure S6: Grouping interface. The columns selected here may later be used to split the data for comparisons. The order of their elements can be adjusted in the "Data sorting" menu. Note, however, that only levels present in the right boxes are available later on. Up to four grouping columns can be specified at once. Columns holding equal or more than 100 different levels (e.g. numerical values), are not available for grouping. The number of observations per group is shown in the table to the right.

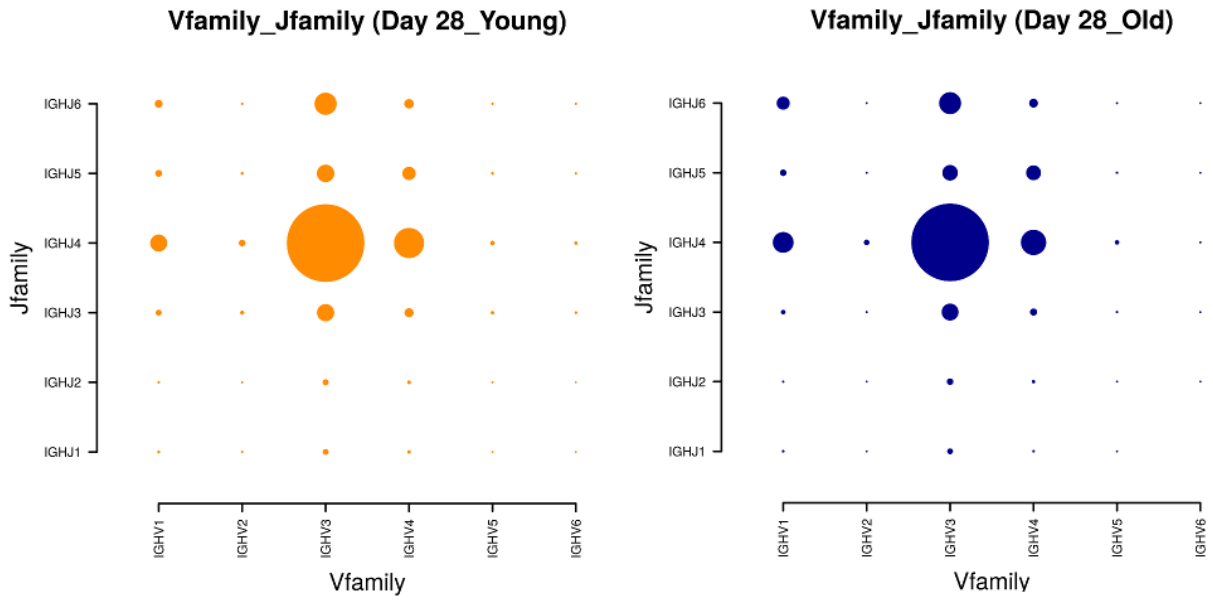

Figure S7: Use case 1: Gene usage plot (2D) reporting the frequencies of gene families present at `Day 28` for the two age groups (vaccination dataset; compare to figure 2 in the main manuscript). The `Young` repertoires seem to have returned to the original state at `Day 0`, which is further illustrated by the Kullback-Leibler divergence values in table S4. In contrast, the `Old` group still shows a slightly different pattern which agrees with the analysis using the CDR3H lengths (figure 3a).

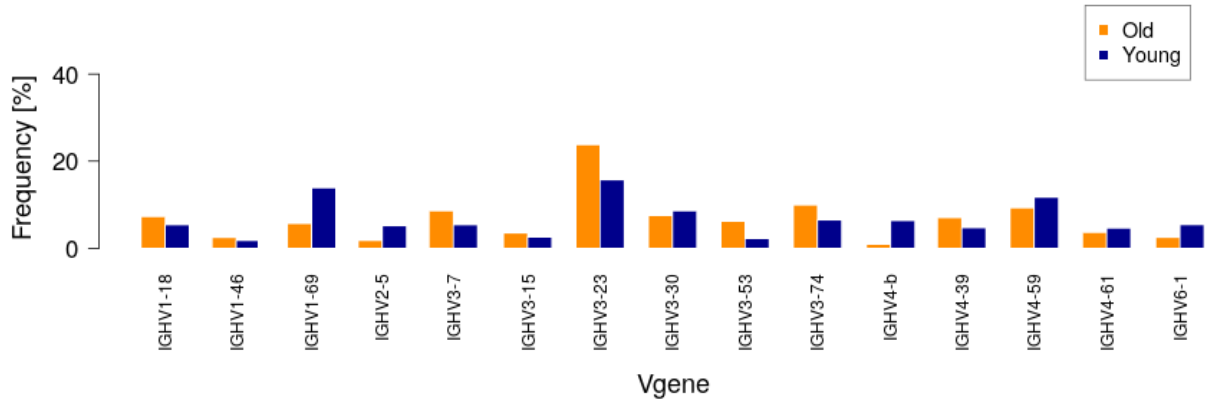

Figure S8: Use case 1: V gene usage plot (1D) for the "Vgene" column (vaccination dataset). Only the significantly populated genes are shown.

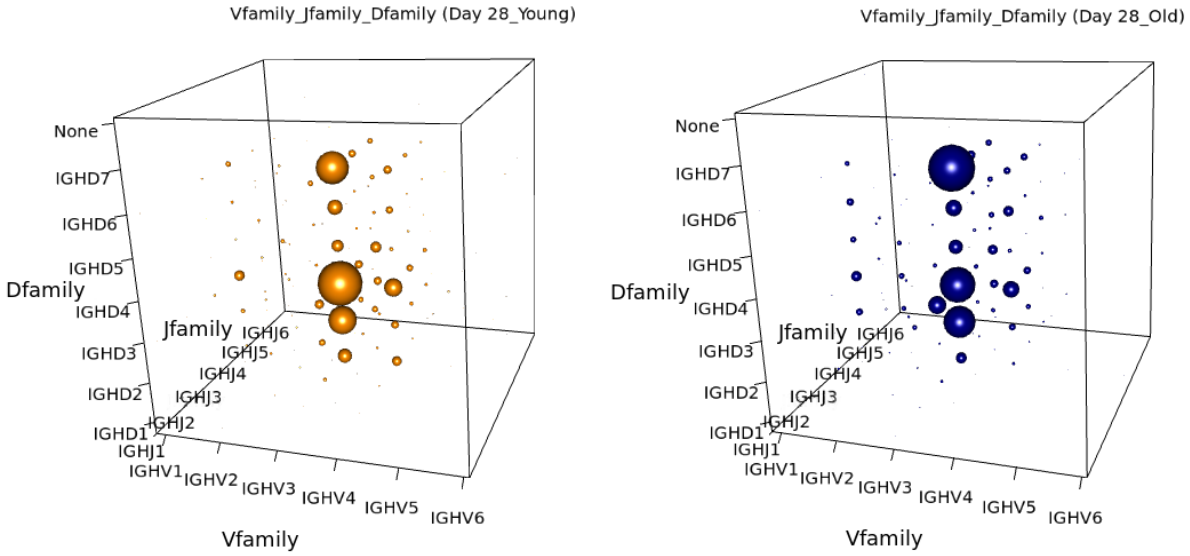

Figure S9: Use case 1: V gene family usage plot (3D) for the "Vfamily", "Jfamily" and "Dfamily" columns at `Day 28` for both the `Young` and `Old` groups of the vaccination dataset. In both cases, `IGHV3` is dominant - but in different combinations. These plots can be rotated and zoomed freely in the web-browser.

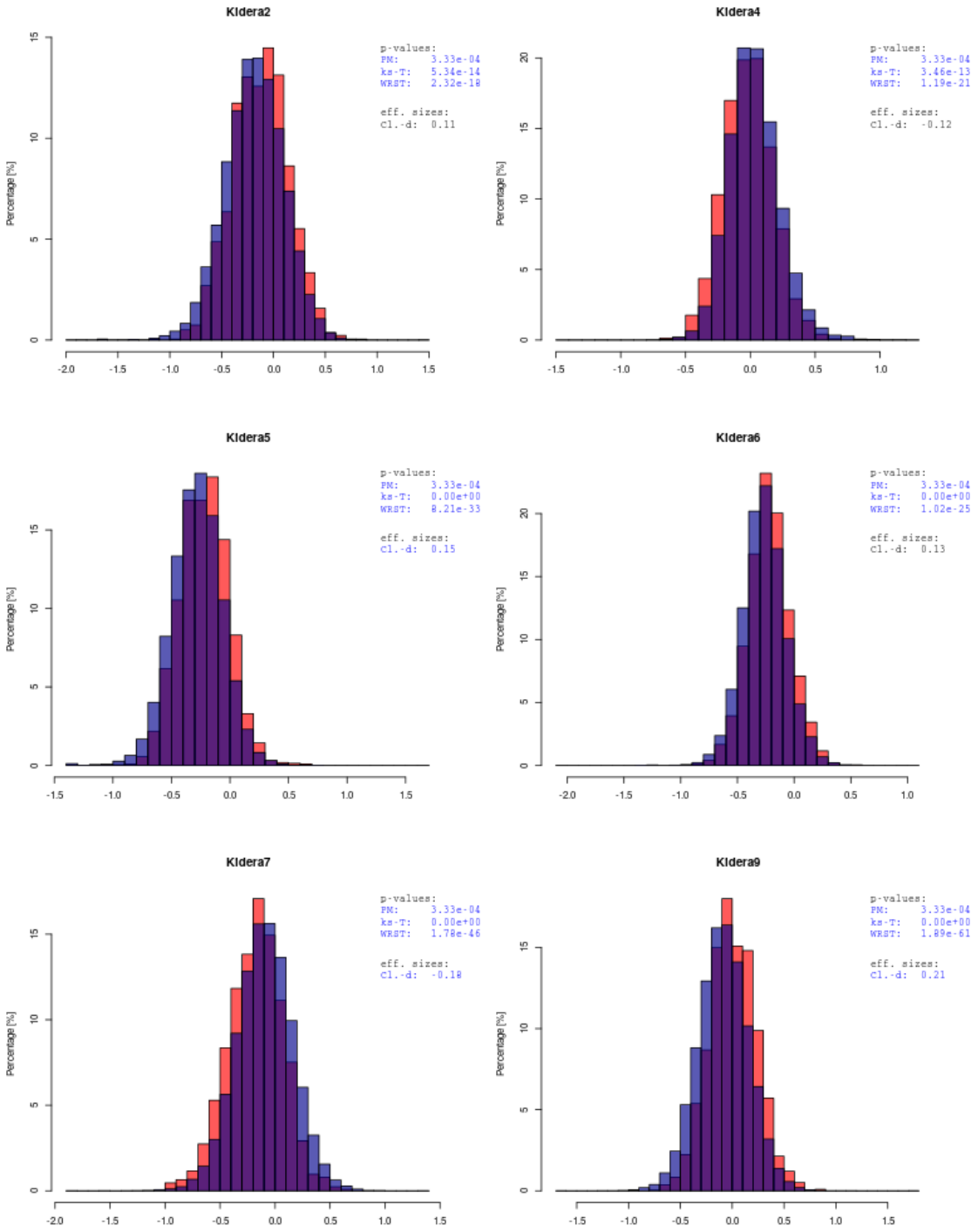

Figure S10: Use case 2: Distribution of CDR3H Kidera factors of variable region sequences encoded by IGHV2 (red) versus all other sequences (excluding those encoded by IGHV7) (blue). In this plot, p-values with a value below 0.05 and effect size measures with non-negligible values (according to reference 22) are shown in blue. The Cliff's Δ values for the remaining Kidera factors are provided in supplementary table S5.

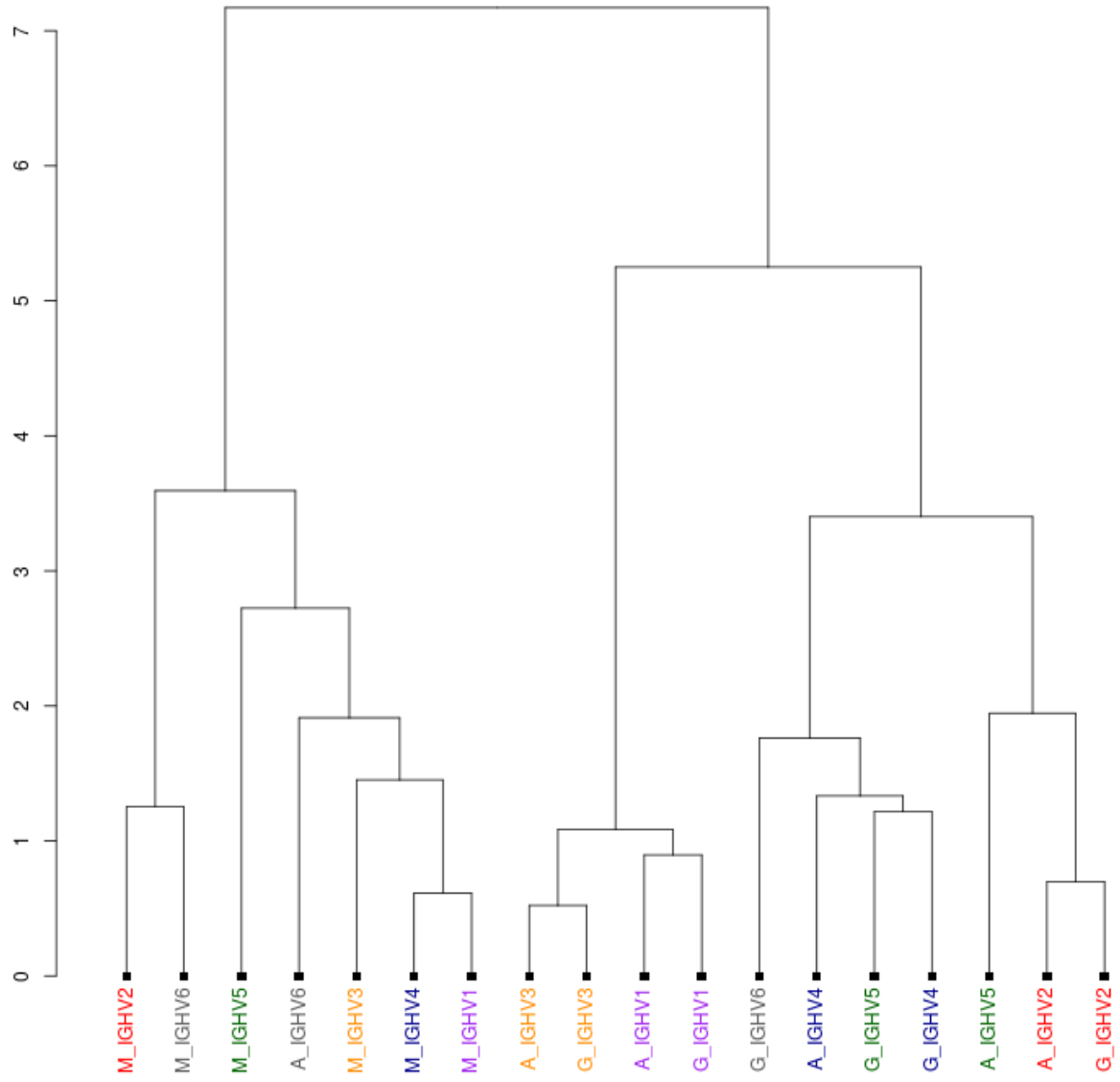

Figure S11: Use case 2: Dendrogram related to figure 4 (main manuscript), showing the result of the hierarchical clustering if only Kidera factors 1, 3, 8 and 10 are used (PBMC). As the main contributors to the separation of `IGHV2` from the other V gene families are excluded, there is no apparent order observable. For this analysis, `IGHV7` and CDR3H loops longer than 35 amino acids have been excluded.

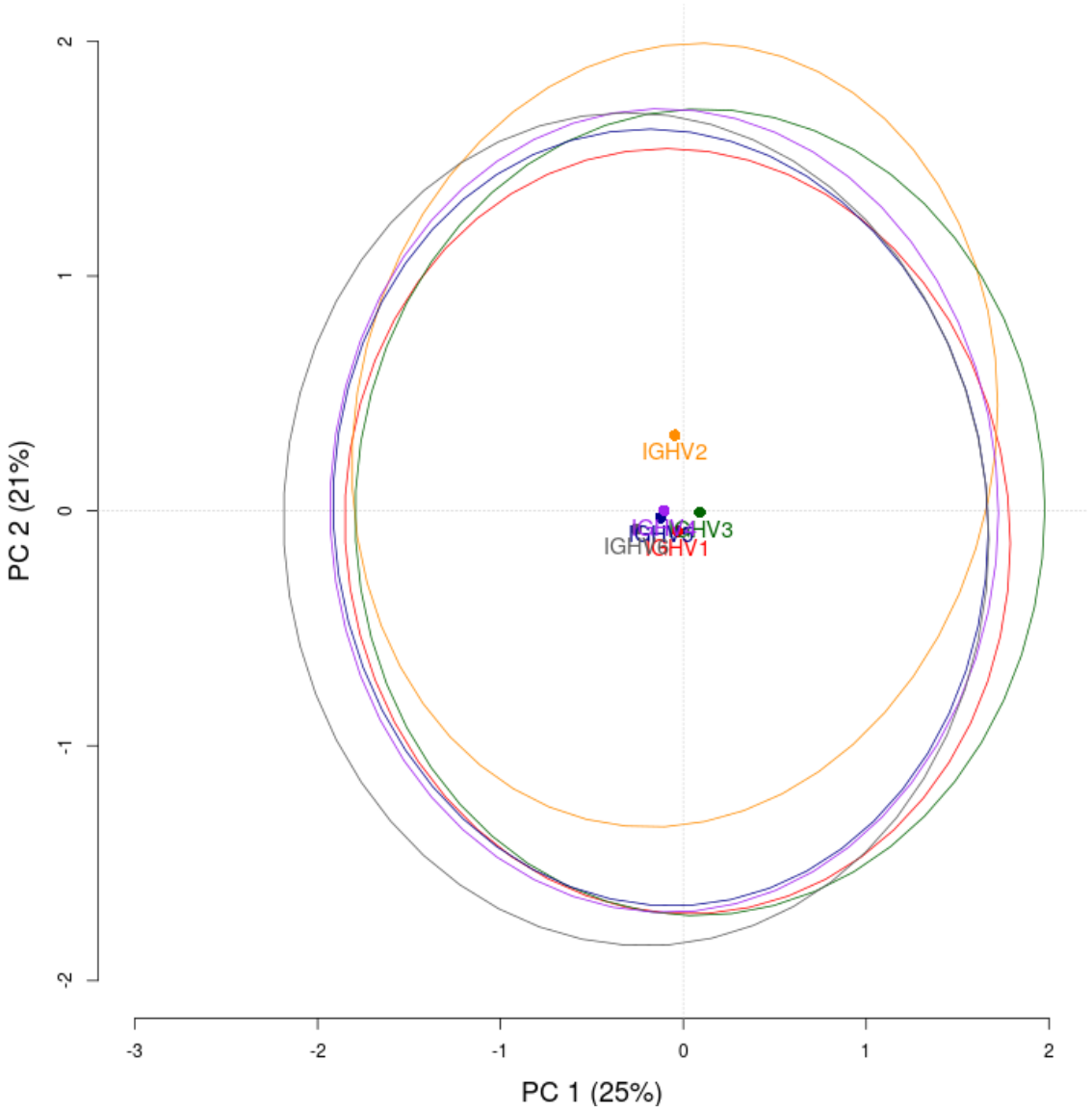

Figure S12: Use case 2: PCA plot showing the separation of `IGHV2` from the other V gene families (PBMC). The same combination of Kidera factors (2, 4, 5, 6, 7 and 9) has been used as in figure 4b (main manuscript). This plot has been generated using the "PCA plot" tab. For this analysis, `IGHV7` and CDR3H loops longer than 35 amino acids have been excluded.

References

[1] Attali, D. shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds, CRAN. (2017).

[2] Bailey, E. shinyBS: Twitter Bootstrap Components for Shiny, CRAN. (2015).

[3] Perrier, V. and Meyer, F. shinyWidgets: Custom Inputs Widgets for Shiny, CRAN. (2017).

[4] Krijthe, J. H. Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation, Rtsne on github. (2015).

[5] Mersmann, O. sendmailR: send email using R, CRAN. (2014).

[6] Sali, A. shinycssloaders: Add CSS Loading Animations to 'shiny' Outputs, CRAN. (2017).

[7] Suzuki, R. and Shimodaira, H. pvclust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling, CRAN. (2015).

[8] van der Maaten, L. J. P. and Hinton, G. E. (2008) Visualizing High-Dimensional Data Using t-SNE. *J. Mach. Learn. Res.,* **9**, 2579–2605.

[9] Wu, Y.-C. B., Kipling, D., and Dunn-Walters, D. K. (2012) Age-Related Changes in Human Peripheral Blood IGH Repertoire Following Vaccination. *Front. Immunol.,* **3**, 193.

[10] Martin, V., Bryan Wu, Y.-C., Kipling, D., and Dunn-Walters, D. (2015) Ageing of the B-cell repertoire. *Philos. Trans. R. Soc. Lond., B., Biol. Sci.,* **370**(1676).

[11] Kullback, S. and Leibler, R. A. (March, 1951) On Information and Sufficiency. *Ann. Math. Stat.,* **22**(1), 79–86.

[12] Osorio, D., Rondón-Villarreal, P., and Torres, R. (2015) Peptides: A Package for Data Mining of Antimicrobial Peptides. *R J.,* **7**(1), 4–14.

[13] Bauer, D. F. (1972) Constructing Confidence Sets Using Rank Statistics. *J. Am. Stat. Assoc.,* **67**(339), 687–690.

[14] Massey, F. J. (1951) The Kolmogorov-Smirnov Test for Goodness of Fit. *J. Am. Stat. Assoc.,* **46**(253), 68–78.

[15] Fay, M. and Shaw, P. (2010) Exact and Asymptotic Weighted Logrank Tests for Interval Censored Data: The interval R Package. *J. Stat. Softw.,* **36**(2), 1–34.

[16] Cohen, J. (1994) The Earth Is Round ($p < .05$). *Am. Psychol.,* **49**, 997–1003.

[17] Wasserstein, R. L. and Lazar, N. A. (2016) The ASA's Statement on p-Values: Context, Process and Purpose. *Am. Stat.,* **70**(2), 129–133.

[18] Nuzzo, R. (February, 2014) Scientific method: Statistical errors. *Nature News,* **506**(7487), 150.

[19] Cohen, J. (1988) Statistical Power Analysis for the Behavioral Sciences, Routledge, 2 edition.

[20] Hedges, L. V. and Olkin, I. (1985) Statistical Methods for Meta-Analysis, Academic Press, 1 edition.

[21] Cliff, N. (1993) Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychol. Bull.,* **114**(3), 494–509.

[22] Torchiano, M. effsize: Efficient Effect Size Computation, CRAN. (2017).