Functionally Convergent B Cell Receptor Sequences in Transgenic Rats Expressing a Human B Cell Repertoire in Response to Tetanus Toxoid and Measles Antigens

Jean-Philippe Bürckert*[1†], Axel R.S.X. Dubois[1†], William J. Faison[1], Sophie Farinelle[1], Emilie Charpentier[1], Regina Sinner[1], Anke Wienecke-Baldacchino[1], and Claude P. Muller[1]*

[†] both authors contributed equally to this work

[1] Department of Infection and Immunity, Luxembourg Institute of Health, Esch-sur-Alzette, Luxembourg

Words: 7289

Running title: Ig convergence in transgenic humanized rats

Correspondence:

Jean-Philippe Bürckert - jean.buerckert@gmail.com
Claude P. Muller - claude.muller@lih.lu

Keywords:

Transgenic rats; B cell repertoire; Next-generation sequencing; repertoire convergence; public CDR3s; AIRR-seq; DESeq2

Abstract (212 Words)

The identification and tracking of antigen-specific immunoglobulin (Ig) sequences within total Ig repertoires is central to high-throughput sequencing (HTS) studies of infections or vaccinations. In this context, public Ig sequences shared by different individuals exposed to the same antigen could be valuable markers for tracing back infections, measuring vaccine immunogenicity, and perhaps ultimately allow the reconstruction of the immunological history of an individual. Here, we immunized groups of transgenic rats expressing human Ig against tetanus toxoid (TT), Modified Vaccinia virus Ankara (MVA), measles virus hemagglutinin and fusion proteins expressed on MVA and the environmental carcinogen Benzo[a]Pyrene, coupled to TT. We showed that these antigens impose a selective pressure causing the Ig Heavy chain (IgH) repertoires of the rats to converge towards the expression of antibodies with highly similar IgH CDR3 amino acid sequences. We present a computational approach, similar to differential gene expression analysis, that selects for clusters of CDR3s with 80% similarity, significantly overrepresented within the different groups of immunized rats. These IgH clusters represent antigen-induced IgH signatures exhibiting stereotypic amino acid patterns including previously described TT and measles specific IgH sequences. Our data suggest, that with the presented methodology, transgenic Ig rats can be utilized as a model to identify antigen-induced, human IgH signatures to a variety of different antigens.

Introduction

Immunoglobulin (Ig) molecules are the primary effectors of the humoral immune response. In theory, Ig can bind to every possible antigen through the large variety of immunoglobulin V (variable), D (diversity) and J (joining) gene rearrangements in the bone marrow and target-oriented affinity maturation in germinal centers (1). All B cells of a germinal center are clonally related to a common ancestor and target the same antigen with varying affinities, iteratively selecting for improved affinity and avidity (2). The Ig molecules of the emerging B cells bind to the target epitope in a lock-and-key principle which is mediated mainly by the heavy chain complementary-determining region 3 (CDR3) loop on top of the Ig (1, 3). The CDR3 is the most variable part of the Ig sequence and the main antigen-binding determinant. The repertoire of CDR3s sufficiently describes the entire functional immunoglobulin heavy chain (IgH) repertoire of an individual (3, 4).

High-throughput sequencing (HTS) has been widely applied to study the IgH repertoire in response to vaccination and infection (5). With this technique, it has become possible to investigate the evolutionary affinity maturation processes after antigenic challenge and to compare their outcome across individuals (6). The IgH repertoire is essentially private (7), but it appears that individuals also produce a public response to a common antigenic stimulus characterized by a certain degree of similarity at the CDR3 sequence level (8–11). Studies investigating this concept of public Ig CDR3s mainly used human blood-derived PBMCs. These represent only a miniscule part of the complete Ig repertoire (12) and it is critical to capture the affinity-matured B cells during their brief transit from the germinal centers through peripheral blood to the bone marrow. Public CDR3s were identified in humans in response to dengue infection, H1N1 seasonal influenza vaccination and repetitive polysaccharide antigens (5, 11, 13). Such CDR3s provided signatures of past immunological exposures allowing for sequence-based monitoring of vaccination or infectious diseases, and perhaps ultimately to reconstruct an individual's antigenic history.

Here we applied HTS on bone marrow B cells rich in serum antibody producing plasma cells from rats carrying human germline IgH and light chain (IgL) loci, the OmniRat™ (14–17). These transgenic rats were immunized with viral (Modified Vaccinia virus Ankara, MVA), protein (measles virus hemagglutinin and fusion proteins, HF and Tetanus toxoid, TT) and chemically defined hapten-conjugate antigens (Benzo[a]Pyrene-TT, BaP-TT) to study the evolution of convergent CDR3 amino acid sequences. We showed that OmniRat™ mount convergent Ig responses characterized by CDR3s with high amino acid sequence similarity. The level of similarity was consistent for all investigated antigens. We applied an approach similar to differential expression analysis to identify overrepresented clusters of highly similar, antigen-driven CDR3s. These could be grouped into antigen-associated signatures matching previously described measles virus hemagglutinin-specific OmniRat™ hybridomas (18) and human TT-specific antibodies (13, 19–21). Our results suggested that humanized Ig transgenic rats can be used as a model to study human-like Ig repertoire dynamics and to determine antigen-associated CDR3 signatures to characterize the history of antigen exposure in human individuals.

Materials and Methods

Animals and immunizations

Humanized Ig transgenic rats (OmniRat™, Open Monoclonal Technology Inc., Palo Alto, USA) were developed and bred as previously described (14–17). Animals were separated into 6 groups of 4 to 6 individuals. They received 3 intraperitoneal injections at 2-weeks intervals and were sacrificed 7 days after the last injection. Injections either contained 100 µg of tetanus toxoid (TT group; Serum Institute of India, Pune, IN) or of a benzo[a]pyrene-TT conjugate construct (BaP-TT group) (22), both formulated with 330 µg of aluminum hydroxide. Other rats were injected with 10^7 PFU of a recombinant Modified Vaccinia virus Ankara (MVA) expressing the hemagglutinin (H) and fusion (F) glycoproteins of the measles virus (MVA-HF group) or the MVA viral vector only (MVA group) without adjuvant. The control animals received either 330 µg of aluminum hydroxide alone (ALUM

group) or were left untouched (NEG group). Antigen-specific IgG responses were monitored by ELISA 10 days after immunizations and at sacrifice. All animal procedures were in compliance with the rules described in the Guide for the Care and Use of Laboratory Animals (23) and accepted by the 'Comité National d'Éthique de Recherche' (CNER, Luxembourg).

Antigens for immunization and ELISA

BaP was coupled to ovalbumin (OVA, Sigma-Aldrich) for ELISA and to purified tetanus toxoid as previously described (22). The recombinant Modified Vaccinia virus Ankara (MVA) and the recombinant MVA carrying measles virus H and F proteins of the Edmonston strain (MV vaccine strain, clade A) viruses were propagated on BHK-21 cells (ATTC™ CCL-10™) as previously described (24–26). Antigen-specific IgG antibody levels in sera were determined in 384-well microtiter plates (Greiner bio-one, Wemmel, BE), coated overnight at 4°C with either 250 ng of MV antigen (Measles grade 2 antigens, Microbix Biosystems, Mississauga, USA), 2.5×10^5 PFU of sonicated MVA (~314 ng), 187.5 ng of TT or 0.25 µM of BaP-OVA in carbonate buffer (100 mM, pH9.6). Free binding sites were saturated with 1% bovine serum albumin (BSA) in Tris-buffered saline at room temperature for 2h. Serial dilutions of the sera were added for 90 min at 37°C, and developed with alkaline phosphatase-conjugated goat anti-rat IgG (1/750 dilution, ImTec Diagnostics, Antwerp, BE) and the appropriate substrate. Absorbance was measured at 405 nm. Endpoint titers (EPT) were determined as the serum dilutions corresponding to 5 times the background.

Sample preparation, amplification and IonTorrent PGM Sequencing

Lymphocytes were isolated from bone marrow samples by density-gradient centrifugation (ficoll® Paque Plus, Sigma-Aldrich). Total RNA was extracted from 10^8 cells with an RNeasy midi kit following the manufacturer's protocol (Qiagen) and enriched for mRNA using paramagnetic separation (µMACS mRNA Isolation kit, Miltenyi Biotech, Leiden, NL). cDNA was prepared from 300ng of mRNA using dT_{18} primers and Superscript III reverse transcriptase (Thermo Fisher Scientific) at 50°C for 80 min. Recombined IgH fragments were subsequently amplified by multiplex PCR using primers for human IgHV region and rat Cγ region with Q5 Hot Start High Fidelity polymerase (NEB, Ipswich, USA) as described previously (18). Amplicons were size selected on a 2% agarose gel and quantified. Quality was checked with a Bioanalyzer (High Sensitivity DNA, Agilent Technologies, Diegem, BE). Four randomly-selected libraries were pooled in equimolar concentrations and sequenced on a 318™ Chip v2 (Thermo Fisher Scientific) using multiple identifiers (MIDs) with the Ion OneTouch™ Template OT2 400 Kit and the Ion PGM Sequencing 400 Kit (Thermo Fisher Scientific) on the Ion Torrent Ion Personal Genome Machine (PGM™) System (Thermo Fischer Scientific).

Quality control and sequence annotation

BAM files were extracted from the Torrent Suite ™ software (version 4.0.2, standard settings) and demultiplexed by multiplex identifiers (MID). Only reads with an unambiguously assigned MID (0 mismatches), identified primers at both ends (2 mismatches allowed) and more than 85% of the bases with a quality score above 25 were considered for further analysis. After clipping MIDs and primers, sequences were collapsed and submitted to the ImMunoGeneTics database (IMGT) HighV-QUEST webserver (www.imgt.org, (27)) for IgHV gene annotation and CDR3 delineation (28). IgHV and IgHJ genes for the in-frame, productive sequences were subsequently assigned using a local installation of IgBlast (29), including only the genes present in the genome of the OmniRat™ as references. Only sequences with an unambiguously assigned IgHV and IgHJ gene were considered for further analysis.

CDR3 similarity threshold for public immune responses

The number of matches for the 200 most frequent CDR3s (top 200) of a rat A in a rat B was obtained for a series of similarity thresholds and returned as ratio from 0 to 1 (i.e. all top 200 CDR3s of rat A have a match in rat B). Ratios were determined from 50% to 100% sequence similarity in one percent increments. The averages for the top 200 matching ratios at each increment were then calculated for all rats within a vaccination group and all rats vaccinated with unrelated antigens. Rats with related antigens were excluded in the pairwise comparison (e.g. MVA as intra-group for MVA-HF). The average top 200 matching ratios were plotted against sequence similarity along with the first derivatives in GraphPad Prism 5 (www.graphpad.com).

Identification of antigen-driven CDR3 clusters

Only CDR3s longer than 4 amino acids were included in the analysis. CDR3s with a minimum of 80% amino acid similarity, were considered as relatives. One amino acid length difference was allowed, to account for insertion and deletions introduced by SHM and occasional differential CDR3-IgHJ region alignments by IMGT (30, 31). These length difference was penalized the same way as a substitution. For each CDR3, the cumulative count of all its 80% relatives per rat (CDR3-count) was calculated and stored in a fuzzy match count table. Data was imported and analyzed with DESeq2 according to the standard workflow for RNA-seq, treating CDR3-counts as expression values (32). Briefly, data was imported as a count-data matrix and converted into a DESeq2-object with conditions according to the antigens used for vaccination. Correct sample grouping was confirmed using variance stabilizing transformed count data (VST-counts). Euclidian distance computation was performed on VST-counts as described in the DESeq2 vignette (33). Principle component analysis plots were generated using the 'PlotPCA' function on VST-counts of the DESeq2 package. P-values were adjusted for multiple testing and to determine the false discovery rate (FDR) using Benjamini-Hochberg correction (34). Based on an FDR of 1%, over-represented CDR3 sequences were extracted if their adjusted p-values were lower than 0.01. Log2-fold change cut-offs were determined manually per antigen group. The extracted CDR3 sequences were grouped using single-seed iterative clustering based on maximum difference of 80% sequence similarity. All analytical scripts were written in Python 2.7 and R 3.2.3 (35).

3D modeling

Selected Ig nucleotide sequences were uploaded to IMGT for annotation. Sequences were elongated to full length by adding the missing nucleotides from the closest germline gene as predicted by the IMGT algorithm. Full length sequences were submitted to the "Rosetta Online Server that Includes Everyone" (ROSIE, http://rosie.rosettacommons.org/, (36–38)) with enabled H3 loop modeling option. ROSIE-output PDB files of the grafted and relaxed models were visualized using PyMol (version 1.7.4, http://pymol.org, (39)).

Results

High-throughput sequencing of OmniRat™ IgH mRNA transcripts

To study convergent IgH repertoires in response to vaccination, 32 transgenic IgHumanized rats (OmniRat™) were immunized with different antigens (**Table 1**; TT, BaP-TT, MVA, MVA-HF). Aluminum hydroxide (ALUM) was used as an adjuvant for TT and BaP-TT. Two control groups received either the adjuvant alone or were left untouched (NEG). All animals exhibited a specific antibody response against the immunizations and mock immunized (ALUM group) and non-immunized animals (NEG group) showed no detectable antigen-specific antibodies (Supplemental **Fig. S1**). MVA-HF and BaP-TT vaccinated animals exhibited a specific immune response against the MVA vector or the TT carrier protein respectively, albeit at lower levels than the animals immunized with these antigens only (**Supplemental Fig. S1**). Rearranged heavy chain IgG genes were amplified from mRNA extracted from bone marrow (BM) lymphocytes and sequenced on a high-throughput sequencing (HTS) Ion Torrent PGM™ platform. A total of 37,473,982 raw reads with MID were obtained (range: 850,298 – 1,879,372 per animal, **Supplemental Table S1**). After quality control and annotation, on average 86,619 unique nt sequences per animal were retained for analysis. The rats expressed a diverse IgH repertoire, including varying frequencies of all human IgHV and IgHJ genes. All possible IgHVJ combinations were found in all vaccination groups with no bias in IgHV, IgHJ genes or IgHVJ recombination usage.

Unique and highly similar CDR3 sequences in response to the same antigen

We first investigated to what extent rats that received the same antigen expressed the same CDR3 amino acid sequences. Pairs of rats from different vaccination groups (369 pairs) shared less CDR3s with each other than pairs of rats within the same group (71 pairs, p-value $< 2\times10^{-16}$, Kruskal-Wallis with Nemenyi post-hoc test) or immunized with related antigens (56 pairs, p-value $= 3.4\times10^{-14}$), indicating that mutual CDR3s are essentially induced by the immunizations (**Fig. 1**). Among a total of 11,643 identical CDR3s (i.e. 100% similarity) that were shared by any set of two or more rats irrespective of the antigen, 5,346 CDR3s (45.9%) were shared exclusively by animals of the same

group and 1,912 (16.4%) were shared between animals immunized with a related antigen (TT and BaP-TT, MVA and MVA-HF). Most of the CDR3s shared within groups were common to only two animals of the same group (6,467; 89.1% of CDR3s shared within groups only). CDR3s present in all animals of a group were rare (**Table 2**). However, multiple CDR3s which differed only by one or two amino acids were shared by all animals within a vaccination group but not by animals from other groups (**Table 3**). Interestingly, these differences occurred preferentially at certain positions of a CDR3 amino acid sequence. This suggested that the vaccinations seemed to have induced identical CDR3s as well as clusters of highly similar CDR3s.

Shared antigen-related CDR3s at 80% sequence similarity.

We compared CDR3s within and across the different vaccination groups to estimate the degree of similarity between these antigen-related clusters. We determined which of the top 200 CDR3s of any rat A had a related CDR3 in a rat B either within the same group (intra-group comparison) or between groups (inter-group comparison) allowing for a single amino acid substitution. The same analysis was repeated for two, three and up to eight amino acid substitutions. The number of top 200 CDR3s found to be present inter- and intra-group were plotted against the amino acid substitutions expressed as percent of CDR3 length (**Fig. 2A, Supplemental Fig. S2**). The resulting sigmoidal curves showed a similar shape for all vaccination groups. In the exponential phase between 100% and 90-95%, intra-group overlap was higher than inter-group overlap. In the linear phase between 90-95% and 75%, overlap increased faster for the inter-group comparison. In the asymptotic phase below 75% similarity, both inter- and intra-group overlap leveled off towards 1, indicating that all top 200 CDR3s of a rat had relatives in any other rat, irrespective of the antigen administered. The first derivatives of the curves showed that in all cases the inflection point was at around 80% (**Fig. 2B, Supplemental Fig. S2**). Thus, at this similarity threshold a maximum number of related CDR3s can be found within the same group while keeping the number of related CDR3s between groups at a minimum. In conclusion, all antigens induced in these rats a public IgH response that can best be characterized by clusters of CDR3s with at least 80% similarity.

Hierarchical clustering of CDR3 repertoires at 80% sequence similarity

Based on the above observation, we identified antigen-driven CDR3s using a workflow developed for differential gene expression analysis of RNA-seq data (32). For each CDR3 within a rat, counts of CDR3 sequences with 80% similarity (CDR3-counts) were used analogous to RNA-seq read counts. Rats of the same vaccination group were considered as replicates. The CDR3-counts followed a negative binomial distribution (**Fig. 3A**). Compared to RNA-seq data, CDR3s usually lack a baseline expression and are essentially private, resulting mostly in zero-counts for individuals across the study, while some shared CDR3s have very high counts in a single animal (**Fig. 3B**). To account for this distribution, we applied variance stabilizing transformation (VST) to the CDR3-counts reducing the variance of the standard deviations over ranked means **(Fig. 3C)**. Hierarchical clustering of VST-counts revealed three clusters (**Fig. 3D**). Cluster I included all animals immunized with MVA (with or without MV HF protein expression). Interestingly, within this cluster, animals of the MVA-HF group and of the MVA group emerged from two separate branches indicating, that additional presentation of HF antigens leaves a distinct imprint in the CDR3 repertoire. Cluster II contained the three groups of animals that received alum as an adjuvant (TT, BaP-TT and ALUM). Again, each of the three groups clustered on separated sub-branches. Cluster III contained only untreated animals (NEG group) and was distinct from all immunized animals. The low variance and the specific grouping of the samples through both principle components showed that the VST-counts cluster the data by vaccination group. This indicated that the different antigens had distinct impact on a subset of the Ig repertoire of the rats (**Fig. 3E**). When the data were reanalyzed applying an 85% or 75% threshold, the clear clustering of rats by vaccination group was lost (**Supplemental Fig. S3**), thus confirming that the 80% similarity threshold was optimal to identify antigen-associated responses on the CDR3 repertoire of the rats. Additionally, it showed that VST CDR3-count data can be analyzed analogously to RNA-seq count data.

Large numbers of antigen-associated CDR3s group into stereotypic signatures

Similar to RNA-seq expression experiments, we aimed to identify CDR3s that are differentially represented between groups of rats. Based on a false discovery rate (FDR) of 1%, 16,727 of the

249,657 (6.9%) unique CDR3s across all groups were found to be overrepresented. One hundred-fold differences in numbers of overrepresented CDR3s were identified in each of the six antigen-groups (**Table 4**). The highest number of overrepresented CDR3s was found in the two combined groups MVA and MVA-HF (n=11,080, 10.4% of the unique CDR3s for this combined group), and TT and BaP-TT (n=2,451, 4.4%), which reflected the high immunogenicity of the antigens TT and MVA common within these groups. Less overrepresented CDR3s were found in the MVA (n=1,689, 2.6%), the MVA-HF (n=804, 1.8%) and TT group (n=540, 1.8%). The lowest number of overrepresented CDR3s was found in the BaP-TT group (n=163, 0.6%). These overrepresented CDR3s could be considered group-specific, and thus immunization induced.
Overrepresented CDR3s were grouped into clusters of 80% sequence similarity (**Fig. 4**). The larger the antigen, the more clusters were found. For instance, 20 clusters were found for the BaP-hapten while 109 clusters were found for the TT protein (46 for TT alone, and 63 for TT and BaP-TT combined). The largest number of clusters was found for the MVA virus antigen (518, with 99 for MVA alone and 419 for MVA and MVA-HF combined). These complex antigen-driven clusters of CDR3s, typical for each group, represented up to 46.5% of the bone marrow IgH repertoire of the rats (**Fig. 5**). The fraction of the repertoire corresponding to these CDR3 clusters varied between the groups but was relatively consistent among animals of the same antigen-group. All together we showed that OmniRat™ exhibited large fractions of highly similar, stereotypic CDR3s in response to the applied vaccinations, even across groups with shared antigens.

Stereotypic signatures match known MV-specific and TT-specific CDR3s

MVA-HF signatures were compared to the previously described CDR3s of MV-specific hybridoma clones derived from an independent set of OmniRat™ immunized with whole MV antigen (18). The largest of the identified HF associated clusters (244 members) matched three CDR3s of MV-specific hybridoma cells, suggesting that this cluster is an MV-H or F protein induced CDR3 signature (**Fig. 6**). Similarly, our TT-associated clusters were compared to known human TT-specific IgH sequences (13, 19–21). The CDR3s from the TT-associated *cluster 4* matched 12 published human CDR3s (**Fig. 7A)**. This OmniRat™ CDR3 signature as well as the human CDR3s consisted of 15-mer CDR3s following the same amino acid pattern. Both humans and rats elicited a conserved paratope defined by a static motif '+QWLV' (+= R/K) at the center of the CDR3, flanked by variable positions that are connected to the torso of the CDR3 (**Fig. 7B**). This indicates that similar key positions are used even across species. The sequence similarity between the human and rat CDR3s ranged from 67% to 87% resulting from different torso amino acid compositions at the positions flanking the conserved binding motif (**Fig. 7C+D**). To compare the structures of these CDR3s from human and rat origin, we performed 3D-homology modeling on their Fab-fragments. Four human antibodies with available heavy and light chain sequences (20) and four selected OmniRat™ heavy chain sequences paired with the human light chains were modeled with Rosetta Antibody. Within the OmniRat™-human chimeric Fab-fragments, the CDR3s formed torso structures ranging from unconstrained amino acid formations over short beta-sheets to rigid beta-sheet hairpin constructs (**Fig. 7C**). Like the rats, human CDR3s exposed the key binding residues at the very tip of the CDR3 loop by a rigid beta-sheet hairpin formation of the torso that protruded out of the IgH core structure (**Fig. 7D**). Together our results corroborate the evolution of functionally convergent CDR3s in different individuals and by different vaccines delivering the same antigen. Also, this strongly indicates that OmniRat™ and humans, albeit the lower sequence similarity between their TT-associated CDR3s, produce antibodies with highly homologous CDR3s in response to the same antigen.

Discussion

We analyzed more than 2,700,000 functional IgH sequences derived from the bone marrow of transgenic rats expressing human B cell receptor genes immunized with different antigens. Our study showed that these rats produced identical as well as highly similar CDR3 amino acid sequences in response to common antigenic challenges. When shared CDR3 repertoire fractions were investigated at different levels of sequence similarity, overlaps between rats from the same vaccination group were optimal around 80% CDR3 amino acid similarity. Applying a differential gene expression workflow to

the counts of 80% similar CDR3s, we presented a novel way to identify convergent, stereotypic CDR3 sequences in response to an antigenic stimulus. These included known CDR3s induced by different measles antigens, indicating that the identified CDR3s are specific for the measles virus H or F proteins which were shared in both immunizations. In addition, our approach also identified CDR3s in response to tetanus toxoid that were remarkably similar to known tetanus-specific CDR3s from human samples. Our findings highlighted the presence of convergent IgH transcripts at high levels in the bone marrow of the transgenic rats and that these sequences are highly similar to those of humans.

Pairs of rats within the same group shared more identical CDR3s than pairs from different groups, but very few CDR3s were shared among all rats of a group. Given the tremendous size and diversity of the Ig repertoire, finding identical sequences in several individuals is indeed unlikely (40). Because of private processes during B cell development including stochastic affinity maturation of the Ig molecules, a certain variability in CDR3s converging towards reactivity with the same antigen is to be expected (6, 41, 42). Galson *et al.* found that for the identification of public repertoires in humans an 87.5-91.6% cut-off (1 in 12 to 1 in 8 amino acids) was optimal to identify TT- and influenza-related CDR3 clusters (43). In the present study, we explored the relation between CDR3 sequence similarity and the overlap between CDR3 repertoires, by inter- and intra-group cross-comparisons at different levels of sequence similarity. Our data showed that 80% amino acid similarity optimized the intra-group overlap between CDR3 repertoires while keeping the inter-group overlap at a minimum. The identified antigen-associated clusters were absent in rats outside immunization groups, which provides a strong support for their underlying biological relevance.

We showed that between 6% and 46% of the bone marrow IgH repertoire correspond to convergent CDR3 sequences. Similar proportions (15-50%) of antigen-specific CDR3 sequences were reported in peripheral blood B cells of patients with acute dengue infections (5). In contrast, in the reconvalescent dengue patients as well as in influenza patients, convergent sequences represented only to less than 1% of peripheral B cell sequences. Such human studies are normally restricted to peripheral blood where only a small fraction of the repertoire can be found and assessed (12). In mice, elevated levels (37%) of public clones were observed in response to HBsAg and less to NP-HEL (22%) and OVA (14%) when examining bone marrow derived long-lived plasma cells ($CD138^+$ $CD22^-$ $MHCII^-$ $CD19^-$ IgM^- PI^-) (44). Here, we analyzed Ig mRNA from bulk rat bone marrow cell isolates, an organ rich in serum antibody-producing long-lived plasma cells (45). Because long-lived plasma cells with high levels of Ig mRNA and only class switched BCR (IgG) were targeted, those are overrepresented in our datasets (42, 43). This explains the high levels of converging CDR3s in the bone marrow. We thus primarily target antigen-associated effector B cells, facilitating the tracking of antigen-specific sequences induced by similar antigens.

Potential influence on the repertoire composition could also result from PCR amplification biases introduced during library preparation as well as sequencing errors (46). We did not account for potential errors and sequencing bias by using molecular barcodes or similar methods (47–50). However, the data analysis of the present study was based on collapsed, unique nucleotide Ig sequences, minimizing the influence of potential PCR amplification bias. Analysis of the nucleotide sequences before and after collapsing to unique nucleotide sequences revealed no major difference in our findings, indicating that PCR amplification bias did not falsifying our results. The IonTorrent PGM sequencing platform is prone to insertion and deletion errors, especially within homopolymer repeats (51). Such errors cause frameshifts within the Ig sequence which are detected by IMGT with 98% efficiency in a benchmarking setup, missing only indels at the beginning and end of the sequence or if placed in close proximity to each other masking the resulting frameshift (Bürckert *et al.*, manuscript in preparation). Sequences with detected indels are marked by IMGT as productive with detected errors and were not included in the described analysis. Furthermore, our analysis is based on the CDR3 amino acid sequence. An insertion or deletion within the CDR3 encoding nucleotides results in the sequence being labelled as unproductive, with no correction attempts undertaken by IMGT (Bürckert *et al.*, manuscript in preparation). Less than 1% of indel combinations remain undetected by IMGT and could be present within the CDR3 encoding nucleotides (Bürckert *et al.*, manuscript in preparation). These rare combination of sequencing errors would then result in

artefactual CDR3s either covered by the applied 80% sequence similarity clustering threshold or missed because of higher sequence variation. Therefore, such CDR3 artefacts can be expected to induce only a small underrepresentation of CDR3s by lowering CDR3-counts. In conclusion, the presented workflow is well protected from potential sequencing errors or PCR bias, that could impact our conclusions.

We found that certain CDR3s have high counts of 80% relatives within a group but very few to none in the unrelated groups. This is in principle comparable to differential gene expression in RNA-seq data. The CDR3-counts followed a negative binomial distribution but, unlike in RNA-seq experiments, our data contained large amounts of CDR3s with zero-counts over different samples. These correspond to private CDR3s that are absent in other rats of the same or other groups. On the other hand, some CDR3s exhibited very high counts of 80% relatives within an animal. While such a data distribution is uncommon in RNA-seq, they were nevertheless compatible with our computational approach (DESeq2, (32)) as demonstrated by negative binomial data distribution and perfect sample grouping after variance stabilizing transformation of the CDR3-counts. Interestingly, the Euclidian distance grouping of MVA-HF and MVA rats remained unchanged for 85% and even for 75% CDR3-counts in contrast to TT-associated rats. Similarly, Trück *et al.* found highly similar ($\leq$ 2 mismatches) Hib- and TT-related sequences enriched seven days post-vaccination, but could not identify H1N1- and MenC-related sequences at the same threshold (13). Correspondingly, statistical evidence of convergent CDR3s in pairs of donors against influenza with a mean genetic distance of ~75% were reported (52). Together with our data this indicated that the identification of convergent Ig repertoire responses using amino acid similarity thresholds was applicable. Future research will tell to what extent the 80% threshold can be applied to other antigens.

Identified convergent CDR3 matched to sequences of previously described human monoclonal antibodies against TT protein (13, 19–21). Despite the relatively low sequence similarity (67% to 87%) between OmniRat^{TM} and human TT-specific CDR3s, they shared a common sequence and structural motif at the center of the CDR3. The center part of the CDR3 is exposed at the tip of the loop structure which directly interacts with the antigen while the adjacent amino acids act as a supporting scaffold. Similarly, Victor Greiff and coworkers observed stereotypical motifs at the center of the CDR3 amino acid sequences in specific antibodies following NP vaccination in mice (53). While structural similarity cannot readily be used to determine antibody specificity, algorithms to identify convergent CDR3s could be further improved by including structural parameters drawn from the expanding amount of available crystal structures.

In conclusion, we demonstrated a strong public IgH response with converging and overlapping CDR3 repertoires in animals exposed to the same antigens. These converging repertoires consisted of similar CDR3 sequences that can be best described using an 80% amino acid similarity threshold. Additionally, we presented an approach to identify such CDR3s by adopting a group-wise expression analysis, similar to RNA-seq approaches. This provides also a valuable tool for large-scale HTS data-mining to identify potential candidates for high-affinity targeted antibody design.

Conflict of Interest

The authors declare no conflict of interest.

Author's contributions

JB and AD contributed equally to the work. JB designed and developed the bioinformatics approach, interpreted data, performed data processing and wrote the manuscript. AD designed and carried out research, prepared samples, interpreted data and wrote the manuscript. WF supported bioinformatics approaches and data processing, corrected the manuscript. AB set up the raw data processing script

and performed data processing, data interpretation and corrected the manuscript. SF, EC provided technical assistance with immunizations, ELISA and virus culture. RS performed IonTorrent PGM sequencing. CM designed research, interpreted data, corrected the manuscript and supervised work. All authors have read and approved the final version of the manuscript.

Funding

JB and AD were supported by the AFR (Aides à la Formation Recherche) fellowships #7039209 and #1196376, respectively, from the FNR (Fonds National de la Recherche) Luxembourg.

Acknowledgements

We are grateful to R. Buelow from Open Monoclonal Technology Inc. (Palo Alto, CA, USA) for providing the OmniRat™. We thank Dr. B. Moss, NIAID, National Institutes of Health (Bethesda, USA) for providing the MVA and recombinant MVA viruses. We thank Josiane Kirpach for her valuable discussions and Fleur AD Leenen for critically revising the manuscript.

References

1. Tonegawa, S. 1983. Somatic generation of antibody diversity. *Nature* 302: 575–581.

2. MacLennan, I. C. M. 1994. Germinal Centers. *Ann Rev Immunol* 12: 117–39.

3. Xu, J. L., and M. M. Davis. 2000. Diversity in the CDR3 Region of V H Is Sufficient for Most Antibody Specificities. *Immunity* 13: 37–45.

4. Ippolito, G. C., R. L. Schelonka, M. Zemlin, I. I. Ivanov, R. Kobayashi, C. Zemlin, G. L. Gartland, L. Nitschke, J. Pelkonen, K. Fujihashi, K. Rajewsky, and H. W. Schroeder. 2006. Forced usage of positively charged amino acids in immunoglobulin CDR-H3 impairs B cell development and antibody production. *J. Exp. Med.* 203: 1567–78.

5. Parameswaran, P., Y. Liu, K. M. Roskin, K. K. L. Jackson, V. P. Dixit, J. Y. Lee, K. L. Artiles, S. Zompi, M. J. Vargas, B. B. Simen, B. Hanczaruk, K. R. McGowan, M. A. Tariq, N. Pourmand, D. Koller, A. Balmaseda, S. D. Boyd, E. Harris, and A. Z. Fire. 2013. Convergent antibody signatures in human dengue. *Cell Host Microbe* 13: 691–700.

6. Jiang, N., J. He, J. a Weinstein, L. Penland, S. Sasaki, X.-S. He, C. L. Dekker, N.-Y. Zheng, M. Huang, M. Sullivan, P. C. Wilson, H. B. Greenberg, M. M. Davis, D. S. Fisher, and S. R. Quake. 2013. Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci. Transl. Med.* 5: 171ra19.

7. Glanville, J., T. C. Kuo, H.-C. von Büdingen, L. Guey, J. Berka, P. D. Sundar, G. Huerta, G. R. Mehta, J. R. Oksenberg, S. L. Hauser, D. R. Cox, A. Rajpal, and J. Pons. 2011. Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proc. Natl. Acad. Sci.* 108: 20066–71.

8. Henry Dunand, C. J., and P. C. Wilson. 2015. Restricted, canonical, stereotyped and convergent immunoglobulin responses. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 370: 20140238-.

9. Beltramello, M., K. L. Williams, C. P. Simmons, A. MacAgno, L. Simonelli, N. T. H. Quyen, S. Sukupolvi-Petty, E. Navarro-Sanchez, P. R. Young, A. M. De Silva, F. A. Rey, L. Varani, S. S. Whitehead, M. S. Diamond, E. Harris, A. Lanzavecchia, and F. Sallusto. 2010. The human immune response to dengue virus is dominated by highly cross-reactive antibodies endowed with neutralizing and enhancing activity. *Cell Host Microbe* 8: 271–283.

441 10. Galson, J. D., E. A. Clutterbuck, J. Trück, M. N. Ramasamy, M. Münz, A. Fowler, V. Cerundolo,
442 A. J. Pollard, G. Lunter, and D. F. Kelly. 2015. BCR repertoire sequencing: different patterns of B-cell
443 activation after two Meningococcal vaccines. *Immunol. Cell Biol.* 93: 885–95.

444 11. Jackson, K. J. L., Y. Liu, K. M. Roskin, J. Glanville, R. A. Hoh, K. Seo, E. L. Marshall, T. C.
445 Gurley, M. A. Moody, B. F. Haynes, E. B. Walter, H. X. Liao, R. A. Albrecht, A. García-Sastre, J.
446 Chaparro-Riggers, A. Rajpal, J. Pons, B. B. Simen, B. Hanczaruk, C. L. Dekker, J. Laserson, D.
447 Koller, M. M. Davis, A. Z. Fire, and S. D. Boyd. 2014. Human responses to influenza vaccination
448 show seroconversion signatures and convergent antibody rearrangements. *Cell Host Microbe* 16: 105–
449 114.

450 12. Apostoaei, I. A., and J. R. Trabalka. 2010. Review, Synthesis, and Application of Information on
451 the Human Lymphatic System to Radiation Dosimetry for Chronic Lymphocytic Leukemia. 1–51.

452 13. Trück, J., M. N. Ramasamy, J. D. Galson, R. Rance, J. Parkhill, G. Lunter, A. J. Pollard, and D. F.
453 Kelly. 2015. Identification of Antigen-Specific B Cell Receptor Sequences Using Public Repertoire
454 Analysis. *J. Immunol.* 194: 252–261.

455 14. Geurts, A. M., G. J. Cost, Y. Freyvert, B. Zeitler, C. Jeffrey, V. M. Choi, S. S. Jenkins, A. Wood,
456 X. Cui, X. Meng, A. Vincent, S. Lam, M. Michalkiewicz, R. Schilling, S. Kalloway, H. Weiler, S.
457 Ménoret, I. Anegon, G. D. Davis, L. Zhang, E. J. Rebar, P. D. Gregory, F. D. Urnov, H. J. Jacob, and
458 R. Buelow. 2010. Knockout Rats Produced Using Designed Zinc Finger Nucleases. *Science (80-.).*
459 325: 2009–2011.

460 15. Ménoret, S., A.-L. Iscache, L. Tesson, S. Rémy, C. Usal, M. J. Osborn, G. J. Cost, M.
461 Brüggemann, R. Buelow, and I. Anegon. 2010. Characterization of immunoglobulin heavy chain
462 knockout rats. *Eur. J. Immunol.* 40: 2932–41.

463 16. Osborn, M. J., B. Ma, S. Avis, J. Dilley, X. Yang, K. Lindquist, A.-L. Iscache, L.-H. Ouisse, I.
464 Anegon, M. S. Neuberger, M. Brüggemann, M. J. Osborn, B. Ma, S. Avis, A. Binnie, J. Dilley, X.
465 Yang, K. Lindquist, S. Ménoret, A.-L. Iscache, L.-H. Ouisse, A. Rajpal, I. Anegon, M. S. Neuberger,
466 R. Buelow, and M. Brüggemann. 2013. High-affinity IgG antibodies develop naturally in Ig-knockout
467 rats carrying germline human IgH/Igκ/Igλ loci bearing the rat CH region. *J. Immunol.* 190: 1481–90.

468 17. Muellenbeck, M. F., B. Ueberheide, B. Amulic, A. Epp, D. Fenyo, C. E. Busse, M. Esen, M.
469 Theisen, B. Mordmüller, and H. Wardemann. 2013. Atypical and classical memory B cells produce
470 Plasmodium falciparum neutralizing antibodies. *J. Exp. Med.* 210: 389–99.

471 18. Dubois, A. R. S. X., J. P. Buerckert, R. Sinner, W. J. Faison, A. M. Molitor, and C. P. Muller.
472 2016. High-resolution analysis of the B cell repertoire before and after polyethylene glycol fusion
473 reveals preferential fusion of rare antigen-specific B cells. *Hum. Antibodies* 24: 1–15.

474 19. de Kruif, J., A. Kramer, T. Visser, C. Clements, R. Nijhuis, F. Cox, V. van der Zande, R. Smit, D.
475 Pinto, M. Throsby, and T. Logtenberg. 2009. Human Immunoglobulin Repertoires against Tetanus
476 Toxoid Contain a Large and Diverse Fraction of High-Affinity Promiscuous VH Genes. *J. Mol. Biol.*
477 387: 548–558.

478 20. Meijer, P. J., P. S. Andersen, M. Haahr Hansen, L. Steinaa, A. Jensen, J. Lantto, M. B.
479 Oleksiewicz, K. Tengbjerg, T. R. Poulsen, V. W. Coljee, S. Bregenholt, J. S. Haurum, and L. S.
480 Nielsen. 2006. Isolation of Human Antibody Repertoires with Preservation of the Natural Heavy and
481 Light Chain Pairing. *J. Mol. Biol.* 358: 764–772.

482 21. Poulsen, T. R., P.-J. Meijer, A. Jensen, L. S. Nielsen, and P. S. Andersen. 2007. Kinetic, affinity,
483 and diversity limits of human polyclonal antibody responses against tetanus toxoid. *J. Immunol.* 179:
484 3841–3850.

485 22. Grova, N., E. J. F. Prodhomme, M. T. Schellenberger, S. Farinelle, and C. P. Muller. 2009.
486 Modulation of carcinogen bioavailability by immunisation with benzo[a]pyrene-conjugate vaccines.
487 *Vaccine* 27: 4142–4151.

488 23. Edition, E. 2011. *Guide*, 8th ed. National Academies Press (US), Washington (DC).

489 24. Carroll, M. W., and B. Moss. 1997. Host range and cytopathogenicity of the highly attenuated
490 MVA strain of vaccinia virus: propagation and generation of recombinant viruses in a nonhuman
491 mammalian cell line. *Virology* 238: 198–211.

492 25. Drexler, I., K. Heller, B. Wahren, V. Erfle, and G. Sutter. 1998. Highly attenuated modified
493 vaccinia virus Ankara replicates in baby hamster kidney cells, a potential host for virus propagation,
494 but not in various human transformed and primary cells. *J. Gen. Virol.* 79: 347–352.

495 26. Staib, C., and G. Sutter. 2003. Live Viral Vectors: Vaccinia Virus. *Vaccine Protoc.* 87: 51–68.

496 27. Lefranc, M.-P. 2004. IMGT, The International ImMunoGeneTics Information System,
497 http://imgt.cines.fr. *Methods Mol. Biol.* 248: 27–49.

498 28. Alamyar, E., P. Duroux, M. P. Lefranc, and V. Giudicelli. 2012. IMGT® tools for the nucleotide
499 analysis of immunoglobulin (IG) and t cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG
500 mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. In *Methods in Molecular Biology*
501 vol. 882. 569–604.

502 29. Ye, J., N. Ma, T. L. Madden, and J. M. Ostell. 2013. IgBLAST: an immunoglobulin variable
503 domain sequence analysis tool. *Nucleic Acids Res.* 41: W34--40.

504 30. Kepler, T. B., H.-X. Liao, S. M. Alam, R. Bhaskarabhatla, R. Zhang, C. Yandava, S. Stewart, K.
505 Anasti, G. Kelsoe, R. Parks, K. E. Lloyd, C. Stolarchuk, J. Pritchett, E. Solomon, E. Friberg, L.
506 Morris, S. S. A. Karim, M. S. Cohen, E. Walter, M. A. Moody, X. Wu, H. R. Altae-Tran, I. S.
507 Georgiev, P. D. Kwong, S. D. Boyd, A. Z. Fire, J. R. Mascola, and B. F. Haynes. 2014.
508 Immunoglobulin gene insertions and deletions in the affinity maturation of HIV-1 broadly reactive
509 neutralizing antibodies. *Cell Host Microbe* 16: 304–13.

510 31. Lavinder, J. J., Y. Wine, C. Giesecke, G. C. Ippolito, A. P. Horton, O. I. Lungu, K. H. Hoi, B. J.
511 DeKosky, E. M. Murrin, M. M. Wirth, A. D. Ellington, T. Dörner, E. M. Marcotte, D. R. Boutz, and
512 G. Georgiou. 2014. Identification and characterization of the constituent human serum antibodies
513 elicited by vaccination. *Proc. Natl. Acad. Sci. U. S. A.* 111: 2259–64.

514 32. Love, M. I., W. Huber, and S. Anders. 2014. Moderated estimation of fold change and dispersion
515 for RNA-seq data with DESeq2. *Genome Biol.* 15: 550.

516 33. Love, M. I., S. Anders, and W. Huber. 2014. *Differential analysis of count data - the DESeq2*
517 *package*,.

518 34. Author, T., Y. Benjamini, Y. Hochberg, and Y. Benjaminit. 1995. Controlling the False Discovery
519 Rate: A Practical and Powerful Approach to Multiple Controlling the False Discovery Rate: a Practical
520 and Powerful Approach to Multiple Testing. *J. R. Stat. Soc.* 57: 289–300.

521 35. Team, R. D. C. 2004. R: A language and environment for statistical computing. *Vienna, Austria R*
522 *Found. Stat. Comput.* .

523 36. Sivasubramanian, A., A. Sircar, S. Chaudhury, and J. J. Gray. 2009. Toward high-resolution
524 homology modeling of antibody F v regions and application to antibody-antigen docking. *Proteins*
525 *Struct. Funct. Bioinforma.* 74: 497–514.

526 37. Lyskov, S., F. C. Chou, S. Ó. Conchúir, B. S. Der, K. Drew, D. Kuroda, J. Xu, B. D. Weitzner, P.
527 D. Renfrew, P. Sripakdeevong, B. Borgo, J. J. Havranek, B. Kuhlman, T. Kortemme, R. Bonneau, J. J.
528 Gray, and R. Das. 2013. Serverification of Molecular Modeling Applications: The Rosetta Online
529 Server That Includes Everyone (ROSIE). *PLoS One* 8: 5–7.

530 38. Sircar, A., E. T. Kim, and J. J. Gray. 2009. RosettaAntibody: Antibody variable region homology
531 modeling server. *Nucleic Acids Res.* 37: W474–W479.

532 39. Schrödinger, L. L. C. 2015. The PyMOL molecular graphics system, version 1.8. *There is no*

533 *Corresp. Rec. this Ref.* .

534 40. Glanville, J., T. C. Kuo, H.-C. von Büdingen, L. Guey, J. Berka, P. D. Sundar, G. Huerta, G. R.
535 Mehta, J. R. Oksenberg, S. L. Hauser, D. R. Cox, A. Rajpal, and J. Pons. 2011. Naive antibody gene-
536 segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proc. Natl. Acad. Sci.*
537 *U. S. A.* 108: 20066–71.

538 41. Ippolito, G. C., K. H. Hoi, S. T. Reddy, S. M. Carroll, X. Ge, T. Rogosch, M. Zemlin, L. D. Shultz,
539 A. D. Ellington, C. L. VanDenBerg, and G. Georgiou. 2012. Antibody repertoires in humanized NOD-
540 scid-IL2R??null mice and human B cells reveals human-like diversification and tolerance checkpoints
541 in the mouse. *PLoS One* 7: e35497.

542 42. Georgiou, G. 2014. The promise and challenge of high-throughput sequencing of the antibody
543 repertoire. *Nat. Biotechnol.* 32: 1–11.

544 43. Galson, J. D., J. Trück, A. Fowler, M. Münz, V. Cerundolo, A. J. Pollard, G. Lunter, and D. F.
545 Kelly. 2015. In-depth assessment of within-individual and inter-individual variation in the B cell
546 receptor repertoire. *Front. Immunol.* 6: 531.

547 44. Greiff, V., U. Menzel, E. Miho, C. Weber, R. Riedel, S. Cook, A. Valai, T. Lopes, A. Radbruch, T.
548 H. Winkler, and S. T. Reddy. 2017. Systems Analysis Reveals High Genetic and Antigen-Driven
549 Predetermination of Antibody Repertoires throughout B Cell Development. *Cell Rep.* 19: 1467–1478.

550 45. Halliley, J. L., C. M. Tipton, J. Liesveld, A. F. Rosenberg, J. Darce, I. V Gregoretti, L. Popova, D.
551 Kaminiski, C. F. Fucile, I. Albizua, S. Kyu, K.-Y. Chiang, K. T. Bradley, R. Burack, M. Slifka, E.
552 Hammarlund, H. Wu, L. Zhao, E. E. Walsh, A. R. Falsey, T. D. Randall, W. C. Cheung, I. Sanz, and
553 F. E.-H. Lee. 2015. Long-Lived Plasma Cells Are Contained within the CD19(-)CD38(hi)CD138(+)
554 Subset in Human Bone Marrow. *Immunity* 43: 132–45.

555 46. Baum, P. D., V. Venturi, and D. A. Price. 2012. Wrestling with the repertoire: The promise and
556 perils of next generation sequencing for antigen receptors. *Eur. J. Immunol.* 42: 2834–2839.

557 47. Shugay, M., O. V Britanova, E. M. Merzlyak, M. a Turchaninova, I. Z. Mamedov, T. R.
558 Tuganbaev, D. a Bolotin, D. B. Staroverov, E. V Putintseva, K. Plevova, C. Linnemann, D. Shagin, S.
559 Pospisilova, S. Lukyanov, T. N. Schumacher, and D. M. Chudakov. 2014. Towards error-free
560 profiling of immune repertoires. *Nat. Methods* 11: 653–5.

561 48. Turchaninova, M. A., A. Davydov, O. V Britanova, M. Shugay, V. Bikos, E. S. Egorov, V. I.
562 Kirgizova, E. M. Merzlyak, D. B. Staroverov, D. A. Bolotin, I. Z. Mamedov, M. Izraelson, M. D.
563 Logacheva, O. Kladova, K. Plevova, S. Pospisilova, and D. M. Chudakov. 2016. High-quality full-
564 length immunoglobulin profiling with unique molecular barcoding. *Nat Protoc.* 11: 1599–1616.

565 49. Vollmers, C., R. V Sit, J. a Weinstein, C. L. Dekker, and S. R. Quake. 2013. Genetic measurement
566 of memory B-cell recall using antibody repertoire sequencing. In *Proceedings of the National*
567 *Academy of Sciences of the United States of America* vol. 110. 13463–8.

568 50. Bashford-Rogers, R. J., A. L. Palser, S. F. Idris, L. Carter, M. Epstein, R. E. Callard, D. C. Douek,
569 G. S. Vassiliou, G. a Follows, M. Hubank, and P. Kellam. 2014. Capturing needles in haystacks: a
570 comparison of B-cell receptor sequencing methods. *BMC Immunol.* 15: 29.

571 51. Bragg, L. M., G. Stone, M. K. Butler, P. Hugenholtz, and G. W. Tyson. 2013. Shining a Light on
572 Dark Sequencing: Characterising Errors in Ion Torrent PGM Data. *PLoS Comput. Biol.* 9.

573 52. Strauli, N. B., and R. D. Hernandez. 2016. Statistical inference of a convergent antibody repertoire
574 response to influenza vaccine. *Genome Med.* 8: 60.

575 53. Greiff, V., U. Menzel, E. Miho, C. Weber, R. Riedel, S. Cook, A. Valai, T. Lopes, A. Radbruch, T.
576 H. Winkler, and S. T. Reddy. 2017. Systems Analysis Reveals High Genetic and Antigen-Driven
577 Predetermination of Antibody Repertoires throughout B Cell Development. *Cell Rep.* 19: 1467–1478.

578 54. Pommié, C., S. Levadoux, R. Sabatier, G. Lefranc, and M. P. Lefranc. 2004. IMGT standardized

579 criteria for statistical analysis of immunoglobulin V-Region amino acid properties. *J. Mol. Recognit.*
580 17: 17–32.

581

Abbreviations

Abbreviations used in this article: ALUM, aluminum hydroxide; BaP, Benzo[a]Pyrene; BaP-TT, Benzo[a]Pyrene tetanus toxoid conjugate; BM, bone marrow; EPT, endpoint titers; F, fusion glycoprotein of the measles virus; FDR, false discovery rate; H, hemagglutinin glycoprotein of the measles virus; HF, hemagglutinin and fusion protein of the measles virus; HTS, high-throughput sequencing; IgL, immunoglobulin light chain; IMGT, ImMunoGeneTics database; MV, measles virus; MVA, Modified Vaccinia virus Ankara; ROSIE, rosetta online server that includes everyone; TT, tetanus toxoid; VST, variance stabilizing transformation

Tables

Table 1 Study design: antigen and vaccination groups. TT used in the BaP-TT vaccination group was chemically modified during the coupling process and therefore does not present the same antigenic surface as native TT in the TT group.

Antigen	TT group		MVA group		Controls	
	TT	BaP-TT	MVA	MVA-HF	ALUM	NEG
BaP	-	x	-	-	-	-
TT	x	x*	-	-	-	-
MV H + F	-	-	-	x	-	-
MVA	-	-	x	x	-	-
Alum	x	x	-	-	x	-

Table 2: CDR3s shared between rats in the same vaccination group

GROUP	Not shared	CDR3s shared by numbers of animals per group										
		2	3	4	5	6	7	8	9	10	11	12
BaP-TT	27,167	510	74	12	0							
TT	27,317	988	75	15								
MVA	61,139	1,123	168	40	1	0						
MVA-HF	43,778	686	51	4	0	1						
NEG	33,675	455	37	3	0							
BaP-TT + TT	/	258	101	24	11	0	1	1	0			
MVA + MVA-HF	/	924	316	102	72	35	29	15	8	8	4	3

Table 3: Selected CDR3 sequences shared by rats in the MVA-HF group only

MVA-HF associated CDR3s												No. of animals
A	R	I	V	G	A	T	T	E	F	D	Y	6
·	·	·	·	·	·	·	·	D	·	·	·	4
·	·	V	·	·	·	·	·	·	·	·	·	4
·	·	A	·	·	·	·	·	·	·	·	·	3
·	·	·	·	·	·	S	·	D	·	·	·	2
·	·	V	·	·	·	·	·	D	·	·	·	2
·	·	·	·	·	·	S	·	·	·	·	S	2
·	·	A	·	·	·	S	·	·	·	·	·	2
·	·	·	·	·	·	S	·	·	·	·	C	2
·	·	·	·	·	·	S	N	·	·	·	·	2
·	·	·	·	·	D	S	·	·	·	·	·	2
·	·	G	·	·	·	·	·	·	·	·	·	2

A	R	H	R	T	Y	Y	Y	G	S	G	S	P	L	F	D	Y	
·	·	·	·	·	F	·	·	·	·	·	·	·	·	·	·	·	4
·	·	·	·	·	·	·	·	·	·	·	·	·	P	·	·	·	3
·	·	·	·	·	·	·	·	·	·	·	·	·	H	·	·	·	2
·	·	·	·	·	H	·	·	·	·	·	·	·	·	·	·	·	2
·	·	·	·	·	·	·	F	·	·	·	·	·	·	·	·	·	2
·	·	·	Q	·	·	·	·	·	·	·	·	·	R	·	·	P	2
·	·	·	·	·	F	·	F	·	·	·	·	·	·	·	·	·	2
·	·	·	·	·	H	·	·	·	·	·	·	·	I	·	·	·	2
·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	P	2
·	·	·	·	·	F	·	F	·	·	·	·	·	R	·	·	P	2
·	·	·	K	·	F	·	·	·	·	·	·	·	R	·	·	·	2
·	·	·	·	·	·	·	·	·	·	·	·	·	I	·	·	·	2
·	·	·	·	·	·	·	·	·	·	·	·	·	R	·	·	·	2

598

599 **Table 4:** Antigen-driven sequences and 80% similarity clusters

Vaccination group	Antigen	Overexpressed CDR3s	No. of Clusters (n>10)	Total clusters
BaP-TT	BaP + (TT)	163	3	20
TT	TT	540	14	46
MVA-HF	MV H+F	804	13	79
MVA	MVA	1689	28	99
BaP-TT & TT	TT backbone	2451	25	63
MVA-HF & MVA	MVA backbone	11080	233	419

600

601 ## Figure Legends

602 **Figure 1.** Shared CDR3s in OmniRat™ pairs. Box-whisker plots represent the number of identical
603 CDR3s shared between pairs of rats from different (369 pairs, orange), related (56 pairs, light blue), or
604 the same vaccination group (71 pairs, green). More CDR3s were shared between rats from the same
605 (p-value 3.4×10^{-14}) or related (p-value $< 2 \times 10^{-16}$) antigen group than between rats of different
606 antigen groups (Kruskal-Wallis test followed by Nemenyi post hoc test).

607 **Figure 2.** Influence of CDR3 sequence similarity on CDR3 repertoire overlap between rats. (A)
608 Average fractions of top 200 CDR3s of the MVA-HF vaccination group shared with all CDR3s of
609 other samples. Samples were divided into two groups having either the same antigen (HF group
610 samples, blue curve) or different antigens (ALUM, BaP-TT, TT and NEG samples, red curve). Both
611 curves follow a similar sigmoidal behavior. (B) First derivative of both curves. Inflection points align
612 at 80% CDR3 amino acid similarity.

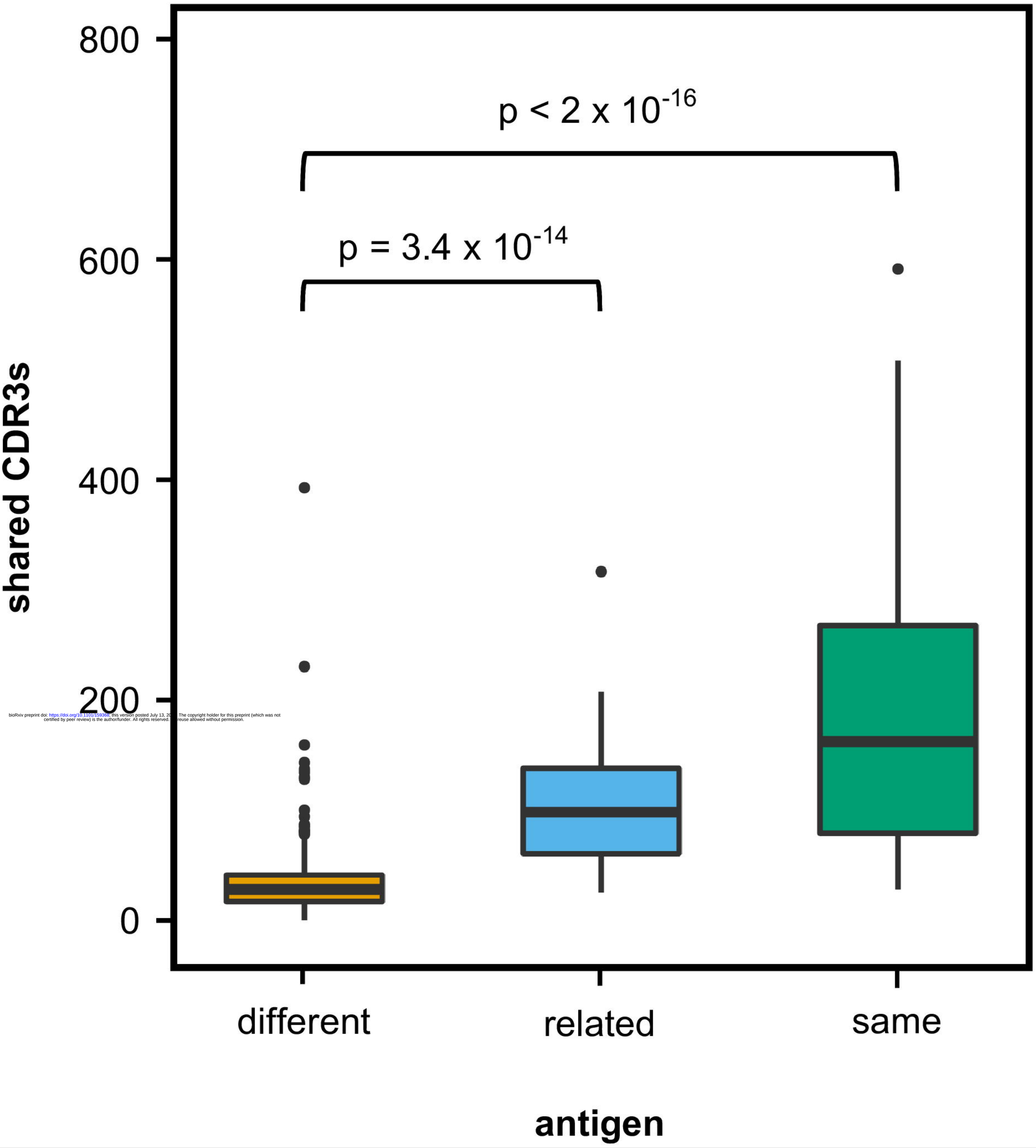
613 **Figure 3.** DESeq2 statistics and sample grouping. (A) Density-histogram representing the distribution
614 of CDR3-counts (log(x+1) transformed). The CDR3-counts follow a negative binomial distribution
615 (B) Sparsity-plot displaying the count distribution per CDR3. The sum of counts for every CDR3 is
616 plotted in $\log_{10}$-scale against the highest count for the CDR3 divided by the sum of all counts for the
617 CDR3. Density of data is indicated by hue. (C) CDR3-wise standard deviation of ranked means of
618 counts after variance stabilizing transformation (VST-counts). The black line shows the standard
619 deviation for all ranked means of VST-counts across all samples, the blue area indicates the data
620 distribution and density by hue. (D) Dendrogram of the Euclidian sample distances calculated for
621 VST-counts. Three main clusters are indicated by coloration (Cluster I: red, Cluster II: green, Cluster
622 III: blue). (E) Scatterplot for the first two principal components of VST-counts. Samples are colored
623 by vaccination-group (MVA: light blue, MVA-HF: green, TT: pink, BaP-TT: gold, ALUM: red, NEG:
624 blue).

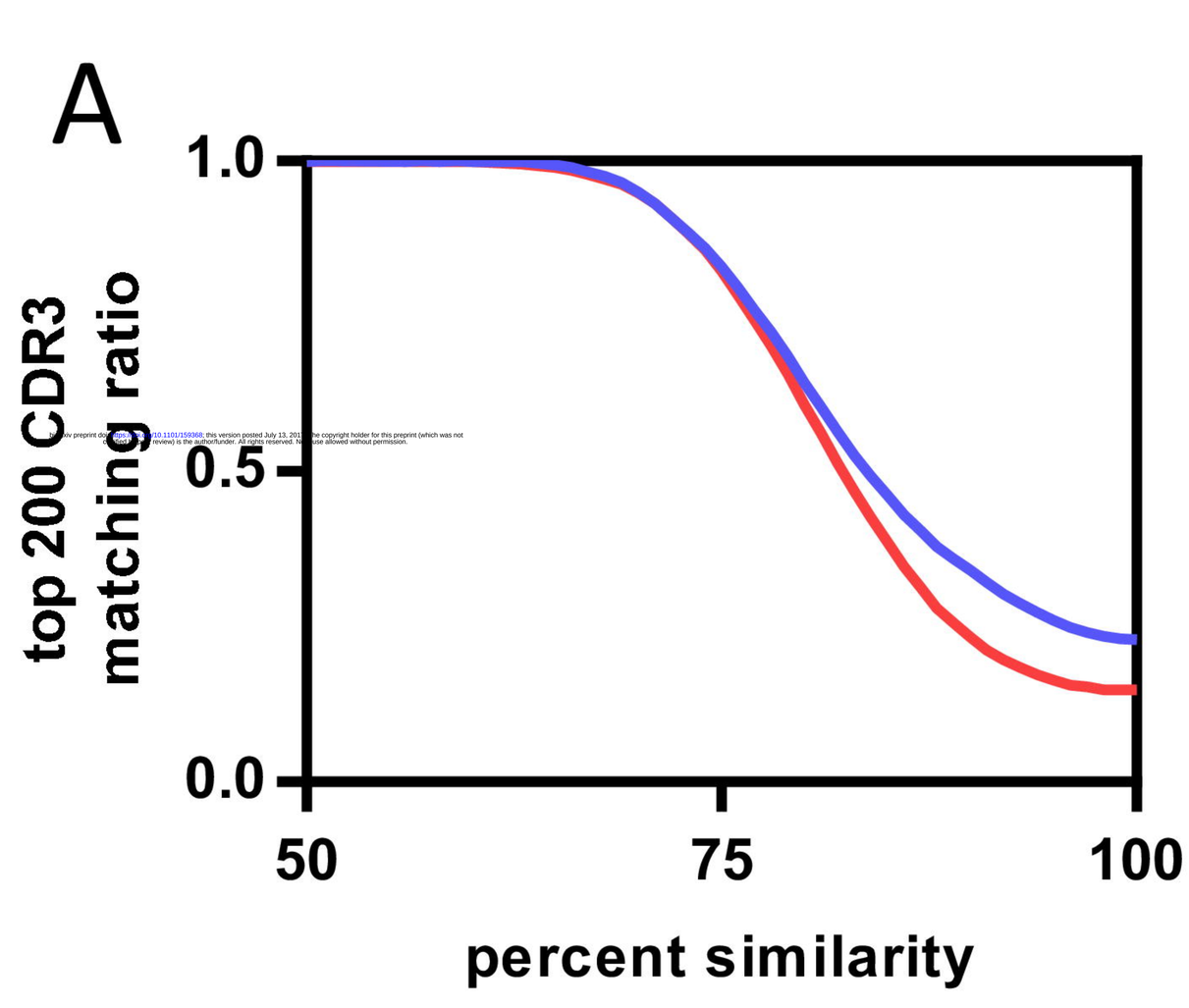
Figure 4. Antigen-associated CDR3-similarity clusters. The top 5 clusters of 80% similar CDR3s overexpressed in response to the antigens are shown as Weblogos. Coloration follows IMGT amino acid coloration scheme (54). Numbers represent the unique CDR3s in each cluster. (A) Clusters associated with the antigens BaP-TT, TT and the combined antigen groups BaP-TT and TT. The red box indicates TT-associated OmniRat™ CDR3s bearing an amino acid pattern also found in human anti-TT PBMC CDR3 sequences from for independent studies. (B) Clusters associated with the antigens MVA-HF, MVA and the combined antigens MVA and MVA-HF. The red box indicates MV-HF associated CDR3-signature also identified in OmniRat™ hybridomas generated in an independent experiment in response to MV antigens.

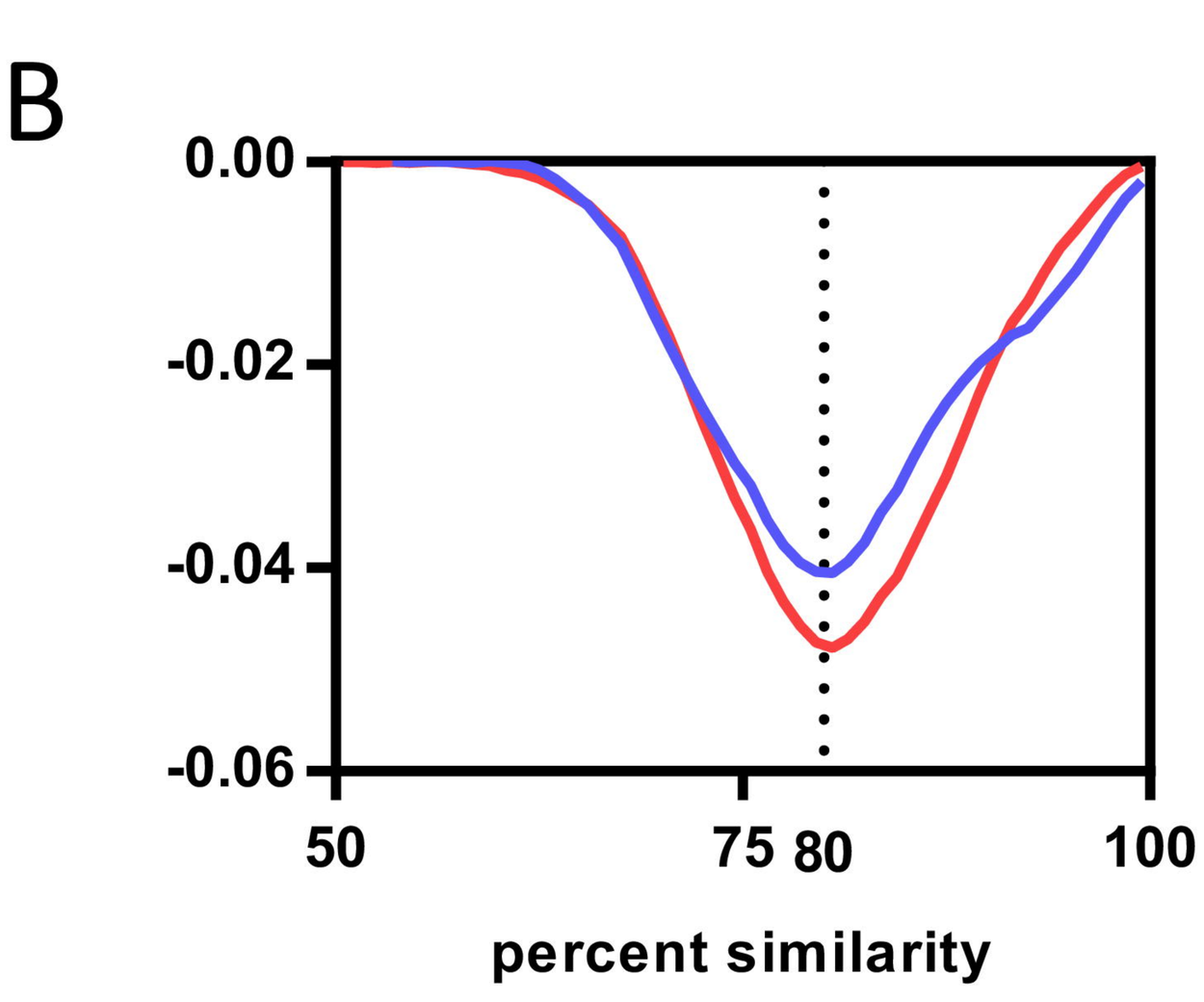
Figure 5. Fractions of the nucleotide Ig repertoire encoding for CDR3 signatures. The Ig repertoire per sample is displayed using numbers of full length nucleotide sequences. Nucleotide sequences encoding for CDR3s that are part of a signature are colored by associated antigen.

Figure 6. OmniRat™ MV-specific CDR3 signature. The clusters of CDR3s overrepresented in response to MVA-HF (see also **Fig. 4A**) and the CDR3s from 3 monoclonal hybridomas specific for MV proteins are shown as Weblogos. The differences between the sequences were calculated as Levenshtein distances in percent of CDR3 length.

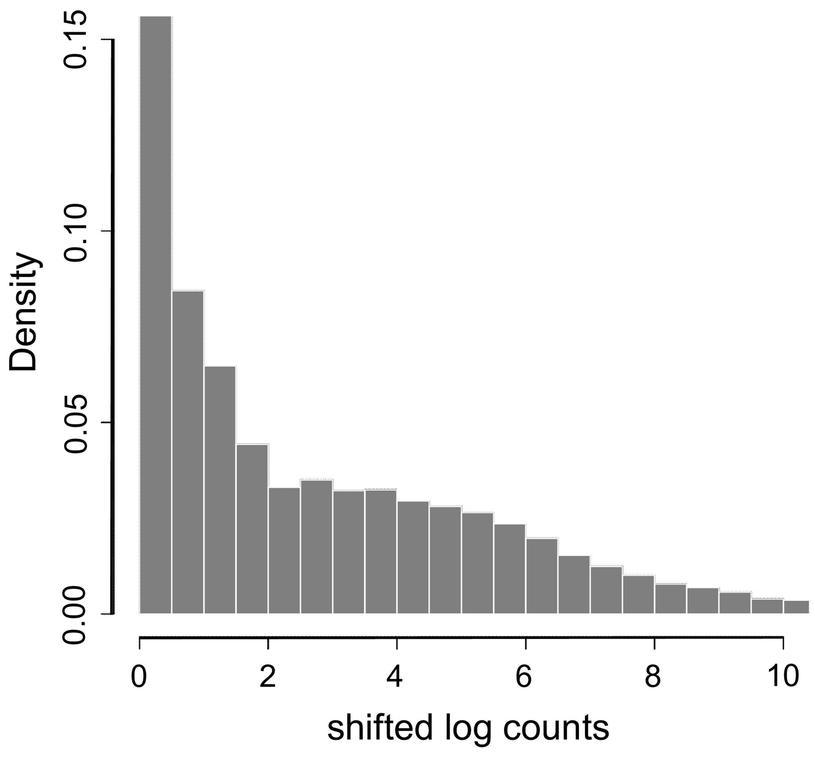
Figure 7. Omnirat™ and human antibodies against tetanus toxoid with similar properties and structures. (A) Sequence similarity range between OmniRat™ TT-associated *cluster 4* (**Fig. 4B**) CDR3s and human TT-specific CDR3s (Levenshtein distance as percent of sequence length). (B) Amino acid pattern for the combined TT-specific human and TT-associated rat CDR3s. The Weblogo shows the conserved binding motif '+QWLV' (+ indicates K/R), and torso amino acids with variable positions are highlighted. (C) 3D-homology models of four OmniRat™-HC-human-LC chimeric antibody Fab-fragments. Heavy chains are colored in orange and light chains in blue, both visualized with 50% transparent surface. CDR3 torsos are shown as cartoon and colored in red. Binding motifs are displayed as sticks and colored in green with only polar hydrogens shown. Views were enlarged to focus on the CDR3 structure. (D) 3D-homology models of four human Fab-fragment visualized as described for (C). Motif, variable and torso structures are highlighted with boxes and arrows as described in (B).



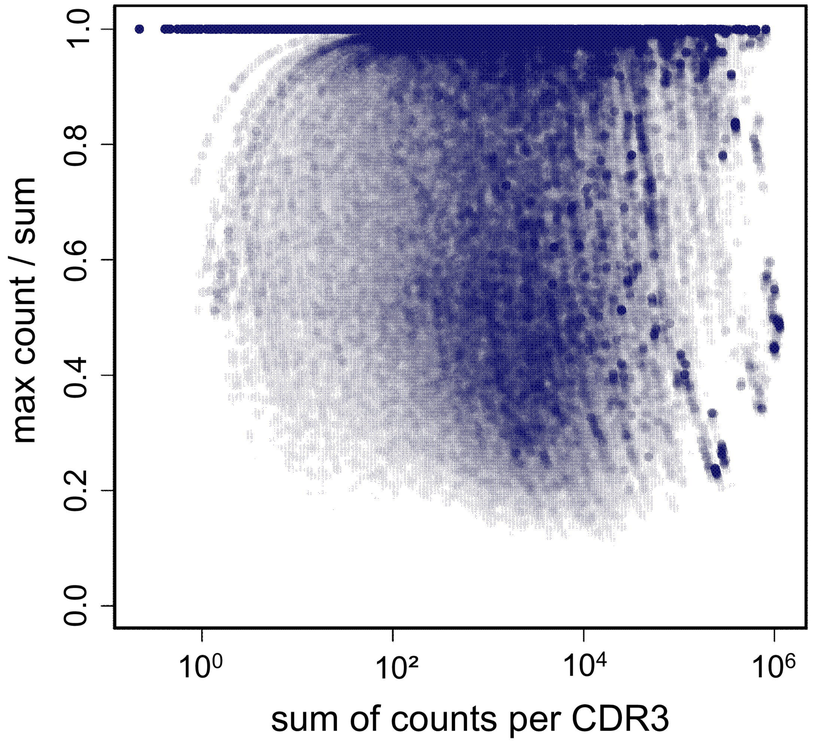




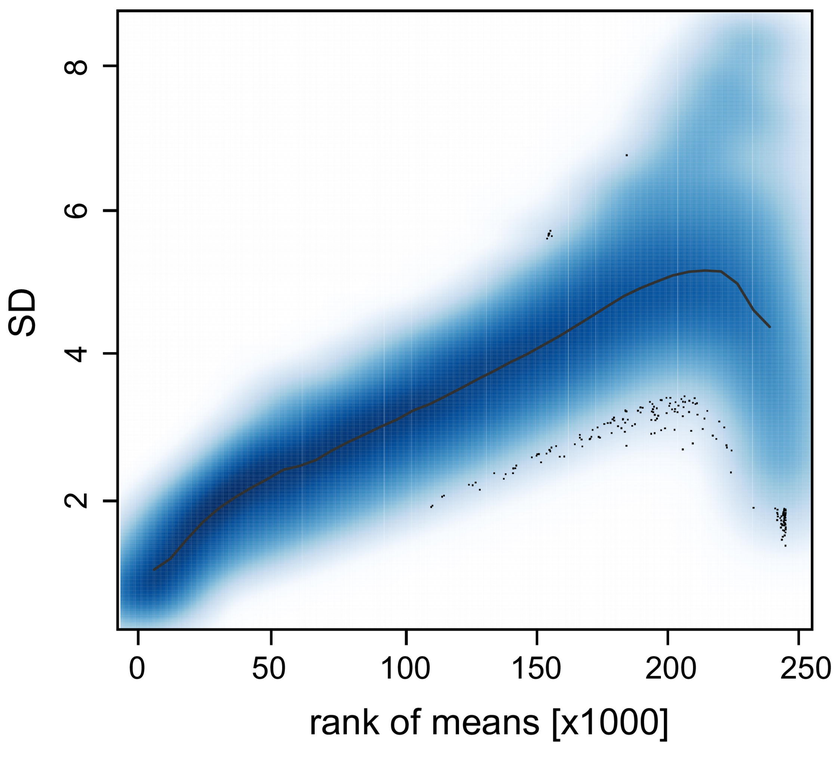
A


Density
0.00
0.05
0.10
0.15
0
2
4
6
8
10
shifted log counts

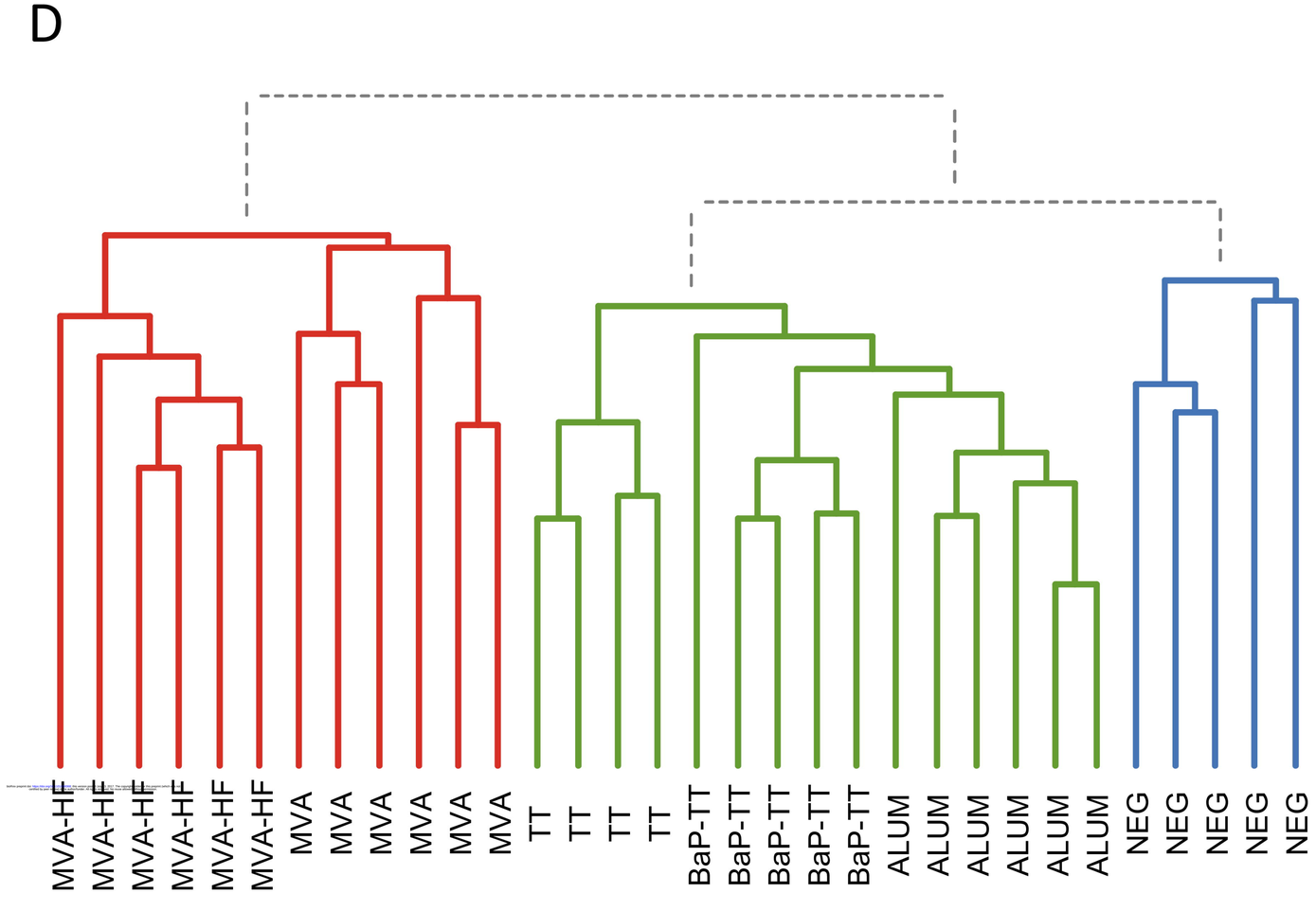
B


max count / sum
0.0
0.2
0.4
0.6
0.8
1.0
10^0
10^2
10^4
10^6
sum of counts per CDR3

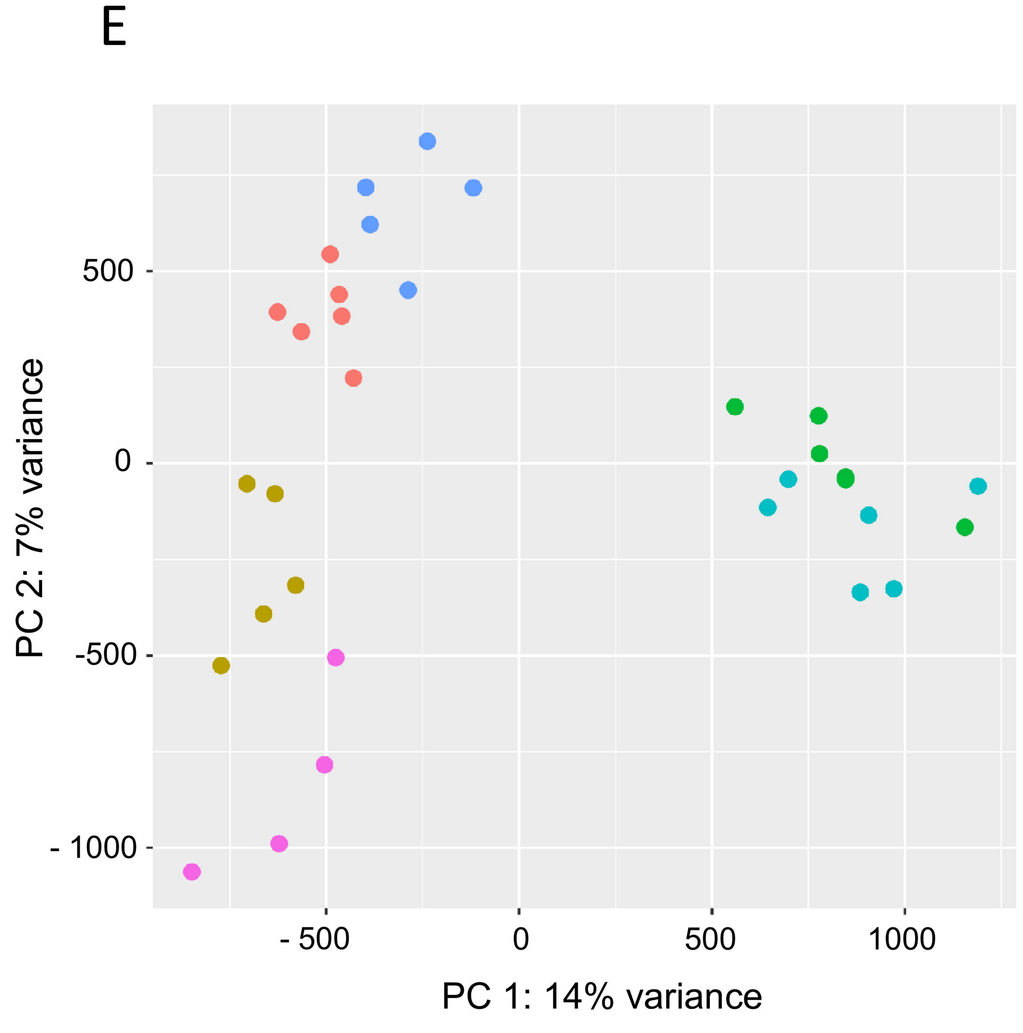
C


SD
2
4
6
8
0
50
100
150
200
250
rank of means [x1000]

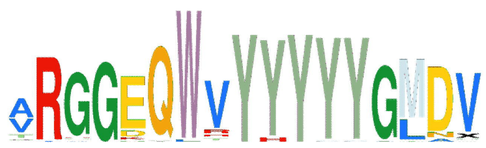
D


MVA-HF
MVA-HF
MVA-HF
MVA-HF
MVA-HF
MVA-HF
MVA
MVA
MVA
MVA
MVA
MVA
TT
TT
TT
TT
BaP-TT
BaP-TT
BaP-TT
BaP-TT
BaP-TT
ALUM
ALUM
ALUM
ALUM
ALUM
ALUM
NEG
NEG
NEG
NEG
NEG

E

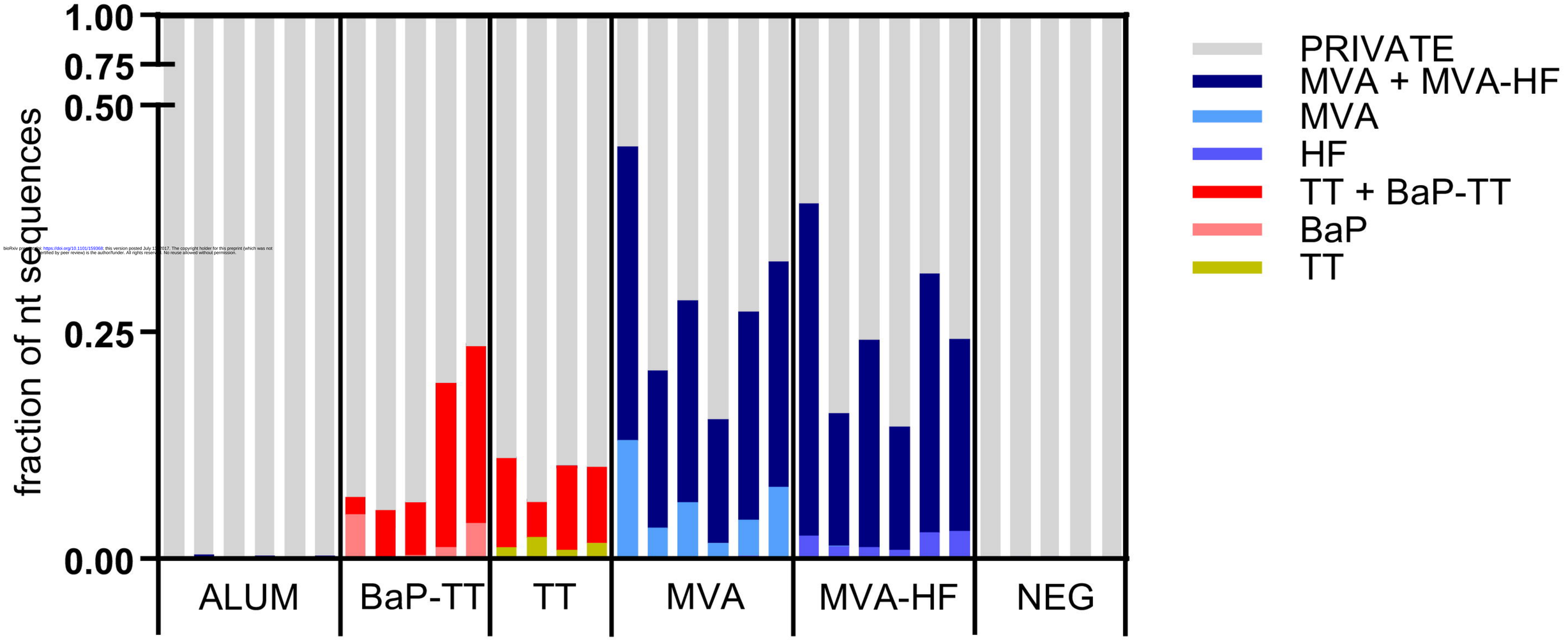

500
0
-500
- 1000
PC 2: 7% variance
- 500
0
500
1000
PC 1: 14% variance

A

Cluster	BaP-TT		TT		BaP-TT + TT	
1		34		63		1283
2		15		39		198
3		14		35		107
4		9		28		80
5		9		25		78

B

Cluster	MVA-HF		MVA		MVA-HF + MVA	
1		244		401		870
2		37		119		857
3		30		64		688
4		28		51		467
5		23		44		408



OmniRat™ MVA-HF Cluster 1 (n=244)

OmniRat™ MV-specific hybridomas (n=3)

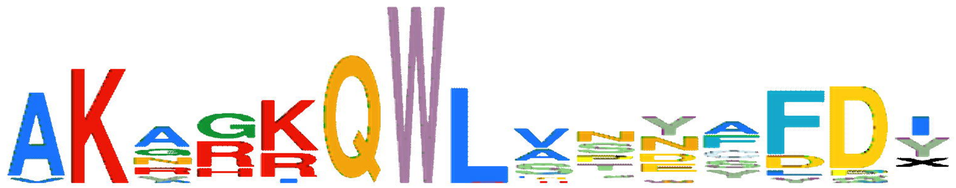
80-100% similarity

A

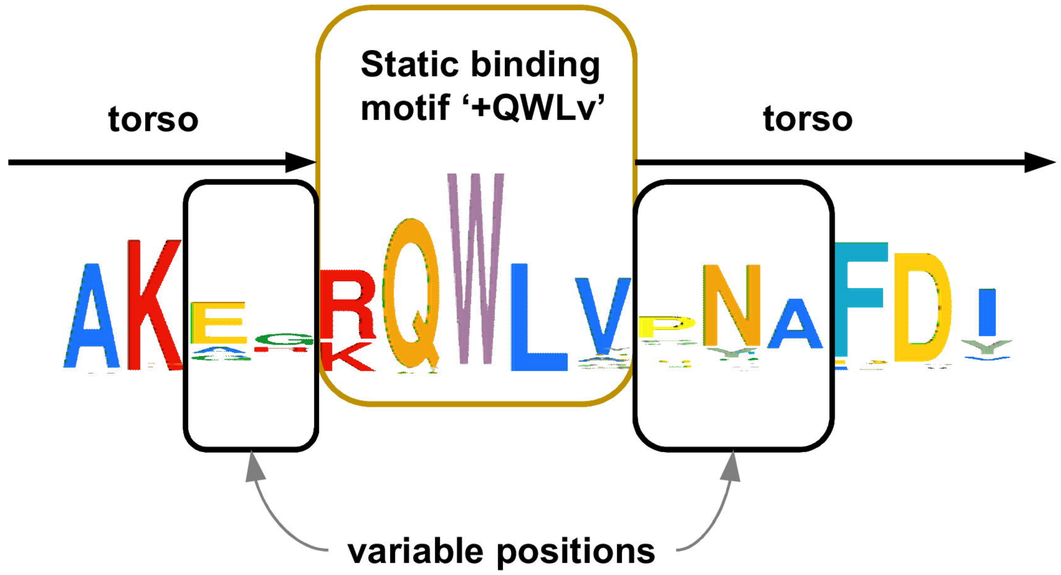
OmniRat™ TT Cluster 4 (n=27)

67-87% similarity

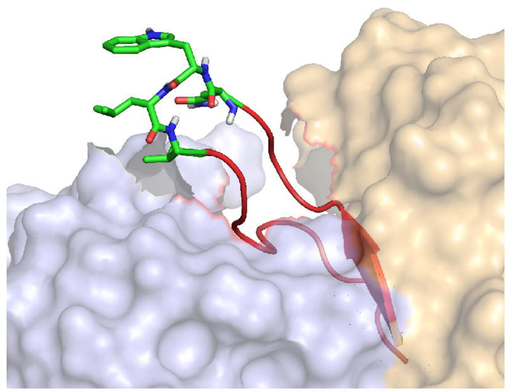
Human TT-specific CDR3s (n=16)



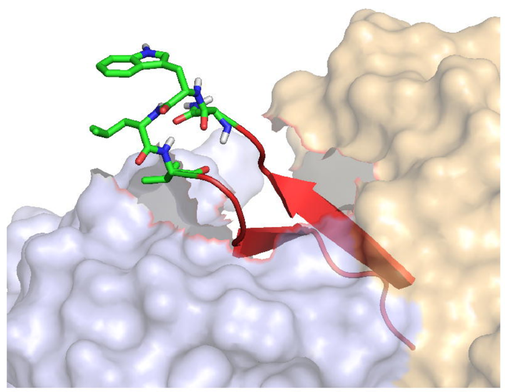
B



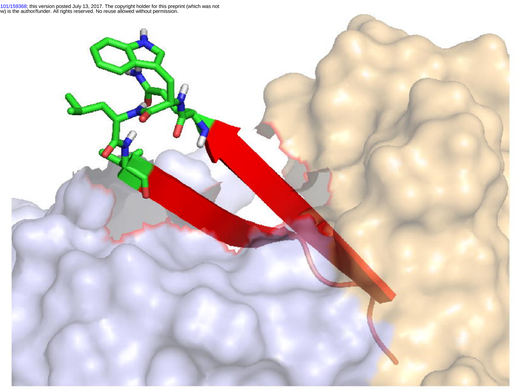
C



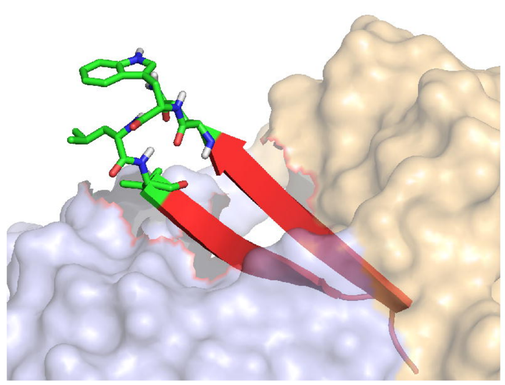
CDR3: **AKEGKQWLVPNAFDI**



CDR3: **AKELRQWLVPNAFDI**

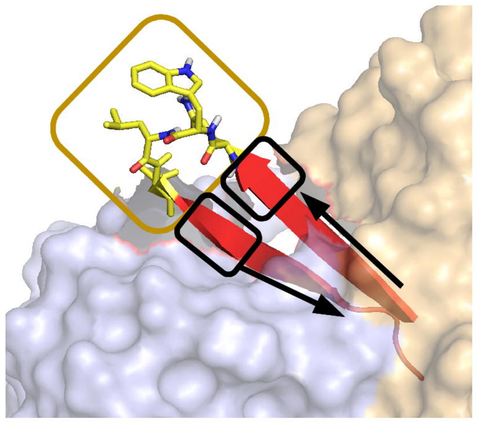


CDR3: **AKENRQWLVHNAFDI**

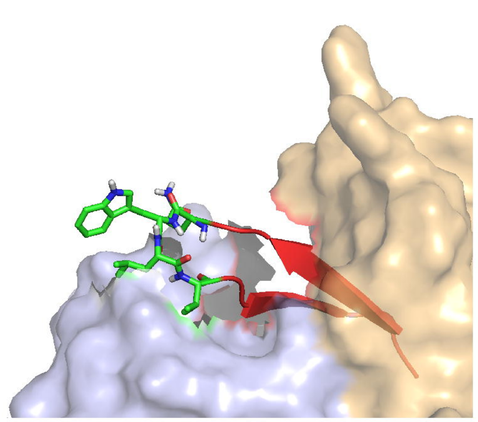


CDR3: **AKSLKQWLVPNAFDI**

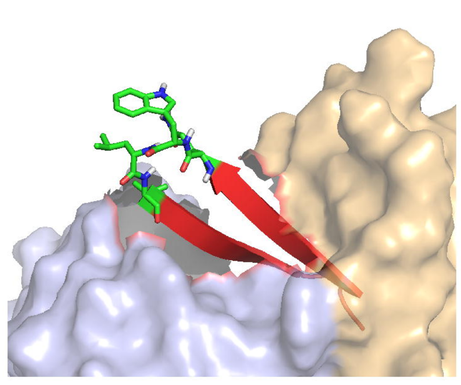
D



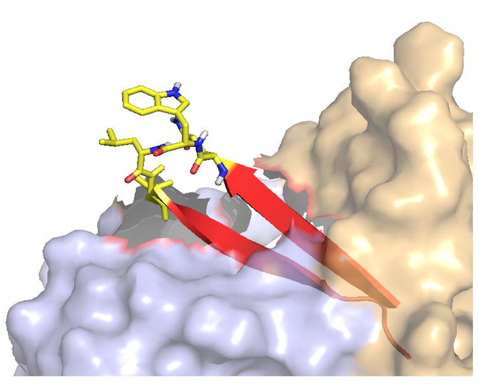
CDR3: **VKRRRQWLVNSSFDF**



CDR3: **AKGRKQWLVPDFDS**



CDR3: **VKRRRQWLVNSSFDF**



CDR3: **VKRRRQWLVNSSFDF**