PAPER ID: PMC6981180

TITLE:
Comparisons of the antibody repertoires of a humanized rodent and humans by high throughput sequencing

ABSTRACT:
The humanization of animal model immune systems by genetic engineering has shown great promise for antibody discovery, tolerance studies and for the evaluation of vaccines. Assessment of the baseline antibody repertoires of unimmunized model animals will be useful as a benchmark for future immunization experiments. We characterized the heavy chain and kappa light chain antibody repertoires of a model animal, the OmniRat, by high throughput antibody sequencing and made use of two novel datasets for comparison to human repertoires. Intra-animal and inter-animal repertoire comparisons reveal a high level of conservation in antibody diversity between the lymph node and spleen and between members of the species. Multiple differences were found in both the heavy and kappa chain repertoires between OmniRats and humans including gene segment usage, CDR3 length distributions, class switch recombination, somatic hypermutation levels and in features of V(D)J recombination. The Inference and Generation of Repertoires (IGoR) software tool was used to model recombination in VH regions which allowed for the quantification of some of these differences. Diversity estimates of the OmniRat heavy chain repertoires almost reached that of humans, around two orders of magnitude less. Despite variation between the species repertoires, a high frequency of OmniRat clonotypes were also found in the human repertoire. These data give insights into the development and selection of humanized animal antibodies and provide actionable information for use in vaccine studies.

Introduction:
A major challenge in human vaccine science is finding appropriate models for studying antibody responses. Animals such as mice, rabbits and monkeys have typically been used in the past and the small animals, in particular, have been favored for ease of immunization, cost reasons and the ability to extensively biopsy post-immunization. One limitation is their use of non-human immunoglobulin (V, D, J) genes in antibodies which can be restricted in their specificity1, and/or lack residues needed for priming by a germline targeting immunogen2. One approach to solving this problem of wild-type animal models is to use humanized immunoglobulin loci-transgenic rodents3,4. The first demonstration of a transgenic rodent with the ability to express human IgM was 30 years ago5. Since then, advances in genetic engineering technologies allowed for the first transgenic mice strains that express fully human antibodies6,7. Today, many new transgenic animal models have been developed including rodents, chickens, rabbits and cows8. These animal models have been used extensively for the discovery of monoclonal antibodies (mAbs)9, tolerance studies10 and more recently for modelling human antibody responses to vaccine candidates3,4. Here we focus on one such animal model: a rat with expression of humanized chimeric antibodies.
The generation of antibody diversity begins with the development of B cells in the bone marrow. Three unlinked loci contain the immunoglobulin gene segments necessary for the assembly of an antibody: one heavy chain locus on chromosome 14 and two light chain loci (lambda and kappa found on chromosomes 2 and 22 respectively). Large pre-B cells derived from common lymphoid progenitors randomly join VH, DH, and JH gene segments to produce a heavy chain. This process requires V(D)J recombinase: a protein complex that contains RAG1, RAG2 and Artemis (among others). P and N nucleotides are added in the VH-DH and DH-JH junctions by Artemis and TdT, dramatically increasing sequence diversity. After successful pairing of this newly formed heavy chain with surrogate light chain (SLC), recombination of a light chain from V and J gene segments of the kappa or lambda loci occurs and the B cell swaps the SLC for this new light chain. Unless the immature B cell is autoreactive or anergic and undergoes receptor editing or clonal deletion, it matures into a naïve B cell and migrates to the periphery whereupon it can become activated by encountering antigen and form germinal centers with help from T-cells. Sequence diversity is again enhanced in the germinal center by somatic hypermutation (SHM) and/or class switch recombination (CSR), two processes that depend on activation induced cytidine deaminase (AID).
The OmniRat was created by genomic integration of human immunoglobulin (Ig) loci on a background of inactivated endogenous rat Ig loci. It expresses chimeric heavy chains (i.e. human V, D, and J genes and rat constant genes) that pair with fully human light chains11,12. We sought

to characterize the circulating antibody repertoire diversity in this animal and make comparisons to humans.
High throughput antibody sequencing has been used to describe the circulating antibody repertoire of organisms, including more recently at unprecedented depth in humans13. Reverse transcription of antibody RNA and combined tagging with unique molecular identifiers (UMIs) have allowed us14 and others15,16 to correct for error and bias in antibody sequencing. Using these methods to gain insight into the antibody repertoire of OmniRats, we ask whether or not it accurately represents that of humans, and by extension allows for usefulness in the approximation of the human antibody response. We postulate that there are major differences in the repertoires due to distinctness in the Ig loci genomic structure and genes that shape antibody diversity between species. Here, we provide the most thorough description of humanized transgenic rodent antibody repertoires to date and leverage a novel extremely deep human dataset to make comparisons with implications of immediate use as a reference for OmniRat immunization studies.

Gene usage comparisons between different tissue sources, between individual OmniRats and between OmniRats and humans ::: Results and Discussion:
We started by making intra-animal comparisons, intra-species comparisons and inter-species comparisons of the immunoglobulin gene segment usage frequencies for each antibody repertoire by performing hierarchical clustering (Fig. 1) and linear regression analysis (Figs. S1 and S2). Repertoires were found to cluster by species and tissue when variable heavy (VH) (Fig. 1a), diversity heavy (DH) (Fig. 1b), joining heavy (JH) (Fig. 1c) and variable kappa (VK) (Fig. S5a), but not joining kappa (JK) (Fig. S5b) gene usage was examined.
Differences between the lymph node and spleens of individual OmniRats were next investigated. VH gene, DH gene and JH gene usage frequencies between these tissues were highly correlated (Fig. S1), although a few VH gene segments were overrepresented in spleen as compared to lymph nodes including VH4-34, VH4-59 and VH5-1 (Figs. 1a and S3a). DH gene and JH gene usage remained highly correlated with minor differences in specific genes (Figs. 1b,c, S1 and S3b,c).
Inter-animal spleen gene usage was highly correlated for all three heavy chain gene segments (Fig. S2). Inter-human comparisons yielded similar, albeit slightly less correlated results (Fig. S2). Intra-species VH and DH usage comparisons made show much weaker correlations with lower R-squared values than any other previous comparison, while surprisingly JH gene usage was highly correlated (Fig. S2). Species specific gene usage biases were more predominant in variable genes (VH and VK) than in joining genes (JH and JK) (Fig. S3a–e). We hypothesize that this may be due to species specific differences in variable gene order, but not joining gene order at the genomic loci11, although no significant correlation between variable gene order and gene usage in the OmniRat was found (data not shown).
OmniRats show a preference for DH gene families of shorter average length such as DH1, DH5, DH6 and DH7 as compared to humans which show a higher representation of longer DH genes from DH2, DH3, and DH4 families with the exception of DH3-9 which appears at similar frequencies between each species (Figs. 1b and S3b). VK and JK gene usage frequencies were very similar for all comparisons made (Figs. S1–3 and S6).

CDR3 comparisons between OmniRats and humans ::: Results and Discussion:
Differences in CDR3 length distributions of each repertoire were next determined. The mean heavy chain CDR3 (CDRH3) length in humans is 14.8 amino acids, while in the OmniRat we observed a mean CDRH3 length that is shorter with a mean length of 12.1 amino acids (Fig. 2a). There are minor differences in the kappa light chain (CDRL3) lengths between species with near identical average lengths of 9.0 and 9.1 for OmniRats and humans respectively (Fig. S6c). The frequency of light chains with a CDR3 of 5 amino acids in length is an important consideration when choosing a model animal for vaccination experiments involving the germline targeting immunogen eOD-GT8 which is in human clinical trials3,14,19. This frequency of 5-amino acid CDRL3s was lower in OmniRats (0.02%) than in humans (0.56%) i.e. a factor of 28 (Fig. S6c).
After observing a tendency for shorter CDRH3 lengths in the OmniRat as compared to humans, we wanted to know if the number of N and P nucleotide additions in the heavy chain V-D and D-J junction sites were different. Figure 2b,c shows average V-D and D-J junction nucleotide addition lengths in the OmniRat are indeed shorter as compared to humans. Nucleotide additions in the V-J junctions of kappa chains are also shorter on average as compared to humans (Fig. S6e).

The longest DH gene segments are found in the DH3 family and the longest JH gene segments come from the JH6 gene family. Gene segments from these families are important contributors to the generation of unusually long CDRH3s in humans and are consistently found in certain broadly neutralizing antibodies (bnAbs) that bind to the human immunodeficiency virus (HIV) envelope (Env) glycoprotein, indicating the importance of these rearrangements in HIV vaccine studies20. On average, the frequency of antibodies with D3-J6 rearrangements in OmniRats is 0.012 with little variation, while in humans the frequency of these antibody species is more variable between subjects with a higher mean of 0.028 (Fig. 2d). The preference of OmniRats for shorter CDRH3 lengths and DH gene segments can be placed in the context of shorter DH gene lengths in the wild-type rat (Rattus norvegicus) as compared to human DH genes (Fig. 2e), indicating a possible biologically intrinsic bias.
We used IGoR21 to infer recombination models for each individual repertoire from 100,000 unmutated sequences allowing for the quantification of differences in features of heavy chain VDJ recombination and generated 1,000,000 synthetic sequences per model. CDRH3 length, VD insertion length and DJ insertion length distributions from the synthetic sequence data (Fig. S4a–c) were found to be very similar to the observed data (Fig. 2a–c). Kullback–Leibler (KL) divergence is a measure of how different two probability distributions are. KL divergence between models (Fig. S4d) and model ‘events’ (Fig. S4e) were computed as previously described13. KL divergence between pairs of OmniRat models was found to be lower than KL divergence between pairs of human models for both complete and all event level calculations. The average pairwise OmniRat model versus human model complete KL divergence calculation was found to be much greater than that of pairwise inter-animal calculations and more than twice that of pairwise inter-human calculations. “D-Gene”, “V-gene trim (3’)”, and “D-gene trim (3’)” were among the events computed to have the mean highest KL divergence from pairwise inter-species event level model comparisons.

Class switch recombination and somatic hypermutation in OmniRats ::: Results and Discussion:
In Supplementary Fig. S5a, the frequency of antibody isotypes is shown. The human repertoire contains average frequencies of 0.84 and 0.16 for IgM and IgG respectively as previously published13, while in the OmniRat antibody repertoire we observe mean frequencies of 0.15 and 0.003 for lymph node and spleen IgG respectively and means of 0.85 and 0.997 for lymph node and spleen IgM respectively. Mean numbers of variable gene mutations in IgM (Fig. S5b), IgG (Fig. S5c) and kappa (Fig. S6d) sequences of the OmniRat were about half of those found in the human repertoire. The observed increase in SHM of class-switched IgG sequences as compared to IgM sequences in the OmniRat demonstrates the ability of the animal to generate memory B cells.

Heavy chain diversity estimates in OmniRats ::: Results and Discussion:
We first examined clonotype (defined as identical VH gene, JH gene and CDRH3 amino acid sequence) diversity of the heavy chain repertoire for each individual animal. All sequences from lymph nodes and spleens were collapsed into unique clonotypes, separately for each tissue and animal. Any clonotype found in both tissue compartments must have originated from different B cells, allowing for the measurement of repeatedly observed clonotypes. Rarefaction curves for each animal were generated (Fig. 3a) and indicate a low frequency of repeatedly observed clonotypes. We estimated diversity using Chao 222,23 and Recon23,24 as previously described13. Diversity estimates were similar between the two estimators, (4.8 × 106–7.4 × 106) for Chao and (9.4 × 106–1.9 × 107) for Recon (Fig. 3b). We next estimated heavy chain sequence diversity for each animal (Fig. 3c) and again found that both estimators broadly agreed, giving similar values of (5.0 × 107–8.1 × 107) for Chao and (5.4 × 107–1.0 × 108) for Recon. Previously published estimates of both clonotype and sequence diversity in individual humans17 only exceed that in the OmniRats by a maximum two orders of magnitude. This is surprising given that the OmniRat is more restricted in CDRH3 length.

Sharing of repertoires between individual OmniRats and between OmniRats and humans ::: Results and Discussion:
For each combination of two or more animals, we computed the frequency of shared unique heavy chain clonotypes (Fig. 4a). There was on average 9.32% of clonotypes shared between each combination of two OmniRats. Surprisingly, we found that 4.90% of clonotypes were shared between all three of the animals.
Next, we pooled unique heavy chain clonotypes from all ten human subjects and measured the percentage of clonotypes in each animal that could be found in the total human pool (Fig. 4b). We

found that (11.9–13.7%) of each OmniRat clonotype repertoire and 13.8% of clonotypes combined from all animals could be found in the total human clonotype pool. Shared clonotypes have shorter CDRH3 lengths than unshared clonotypes on average (Fig. 4c) which is expected given that sequence diversity is expected to increase as the number of amino acids increases giving less of a chance for sharing. VH gene family usage between shared and unshared clonotypes indicates no major differences (Fig. 4d). Sequence logos for 8 amino acid long (Fig. 4e) and 13 amino acid long (Fig. 4e) CDHR3s from both shared and unshared fractions were made and indicate broad similarity between the two fractions.

Conclusion:
We set out to determine commonalities and differences between OmniRat and human antibody repertoires to be used as a reference for vaccine studies. Our results show that there exists substantial variation in gene usage frequencies and elements of recombination, indicating specific limitations of this animal model for predicting the human immune response. We found that by performing hierarchical clustering on gene segment usage, repertoires clustered together by both species and tissue. Differences in gene segment usage between transgenic animal models and humans, as well as between tissues are expected. For example, multiple human Ig loci transgenic rodents are reported to have gene usage profiles that slightly vary from that of humans25,26. Furthermore, antibody repertoires from separate human tissues are known to deviate strongly enough to be clustered by hierarchical clustering27. In our case, the lymph node repertoire from the OmniRat was most likely able to be distinguished from that of the spleen due to the increased presence of antigen-experienced B cells in the latter as shown by somatic hypermutation and class-switched transcript frequencies. This indicates that tissue selection will affect the outcome of an antibody discovery campaign and reinforces evidence for normal B cell development by suggesting the existence of affinity maturation.
Investigation into CDR3 length distributions revealed that the OmniRat prefers shorter CDRH3s as compared to humans. Interestingly, the mechanisms of this preference are due to decreased N additions, and a tendency to incorporate shorter DH gene segments. This result has also been seen in multiple other transgenic and wild-type rodents25,26,28,29. The specific reasons remain unclear, although the observation that wild-type rat germline D gene segment lengths are shorter suggests intrinsic species-specific mechanisms of selection as well as differences in TdT expression during bone marrow B cell development. We further speculate that another possible reason for intra-species variation can be attributed to distinct prior antigen exposure and divergent gut microbiomes30,31.
The diversity of the OmniRat heavy chain repertoire was shown to be slightly lower than that in humans. Our results indicate biased gene usage and decreased junctional diversity are the primary reasons for the resulting repertoire diversity estimate comparisons. We also showed that there is a much higher frequency of ‘public clonotypes’ or clonotypes shared between members of this species than previously reported in humans13,32. Lower sequence diversity combined with identical genetic background and highly similar gene usage are possible reasons for this result.
In summary, we have determined specific differences between the OmniRat and human antibody repertoires which must be taken into careful consideration when evaluating an antibody response in order to make predictions for human subjects. We have also shown that this animal’s antibodies show signs of class switch recombination, somatic hypermutation and large diversity supporting its value for the discovery of monoclonal antibodies to targets that may not be immunogenic in other models. Even though a high degree of variation exists, we still found many clonotypes to be shared between the species pools. Finally, more studies will need to be done in order to characterize OmniRat serum and memory B cell responses to immunogens.

Next-generation sequencing of OmniRat antibody repertoires ::: Methods:
Total RNA from spleens and lymph nodes was extracted (RNeasy Maxi Kit, Qiagen) from each unimmunized heavy chain and kappa chain only transgenic rat (OmniRat, Open Monoclonal Technology Inc., Palo Alto, CA, USA) and antibody sequences were amplified as previously described13 except for different primers used during reverse transcription (Table S2). We chose to interrogate the antibody repertoires found in secondary lymphoid organs as opposed to peripheral blood due to the higher number of B cells. Correct PCR product sizes were verified on an agarose gel (E-Gel EX; Invitrogen) and quantified with fluorometry (Qubit; Life Technologies), pooled at approximately equimolar concentrations and each sample pool was re-quantified before sequencing on an Illumina MiSeq (MiSeq Reagent Kit v3, 600-cycle). All animal experiments were

conducted in accordance with the Institutional Animal Care and Use Committee of Scripps Research and approved by the Institutional Research Boards of Scripps Research.

Processing of next-generation sequencing data ::: Methods:
The Abstar analysis pipeline17 was used as previously described to quality trim, remove adapters and merge paired sequences. Sequences were then annotated with Abstar in combination with UMI based error correction by AbCorrect (https://github.com/briney/abtools/). Resulting annotated consensus sequences were deposited to MongoDB and Spark databases for querying and data analysis in python on Jupyter and Zeppelin Notebooks. For comparisons of frequencies, read counts were scaled for each repertoire as previously described33 due to the large differences in the number of reads between species and number of genes that each species expresses.