T cell receptor repertoires of mice and humans are clustered in similarity networks around conserved public CDR3 sequences

Asaf Madi[1,*], Asaf Poran[1,*], Eric Shifrut[1], Shlomit Reich-Zeliger[1], Erez Greenstein[1], Irena Zaretsky[1], Tomer Arnon[1,2], Francois Van Laethem[3], Alfred Singer[3], Jinghua Lu[4], Peter D. Sun[4], Irun R. Cohen[1], and Nir Friedman[1,†]

1) Department of Immunology, Weizmann Institute of Science, Rehovot 76100, Israel.
2) Department of Physics and Astronomy, Alfred University, 1 Saxon Drive, Alfred, NY 14802
3) Experimental Immunology Branch, National Cancer Institute, Bethesda, Maryland 20892.
4) Structural Immunology Section, Laboratory of Immunogenetics, National Institute of Allergy and Infectious Diseases, Rockville, Maryland 20852
*) These authors contributed equally to this work
†) Corresponding author. E-mail: nir.friedman@weizmann.ac.il (N.F.).

Abstract:

Diversity of T cell receptor (TCR) repertoires, generated by somatic DNA rearrangements, is central to immune system function. However, the level of sequence similarity of TCR repertoires within and between species has not been characterized. Using network analysis of high-throughput TCR sequencing data, we found that abundant CDR3-TCRβ sequences were clustered within networks generated by sequence similarity. We discovered a substantial number of public CDR3-TCRβ segments that were identical in mice and humans. These conserved public sequences were central within TCR sequence-similarity networks. Annotated TCR sequences, previously associated with self-specificities such as autoimmunity and cancer, were linked to network clusters. Mechanistically, CDR3 networks were promoted by MHC-mediated selection, and were reduced following immunization, immune checkpoint blockade or aging. Our findings provide a new view of T cell repertoire organization and physiology, and suggest that the immune system distributes its TCR sequences unevenly, attending to specific foci of reactivity.

Introduction:

The T-cell receptor (TCR), which is generated through random rearrangement of genomic V-D-J segments, is the mediator of specific antigen recognition by T lymphocytes. The collective variety of these receptors expressed by an individual, the TCR repertoire, reflects the state of the adaptive immune system and its history, as its composition changes throughout life in response to immune challenges. The individual TCR repertoire is shaped by biases in the process of VDJ recombination (Robins et al. 2010; Miles et al. 2011; Murugan et al. 2012; Ndifon et al. 2012), and by the subsequent expansion and deletion of certain T cell clones upon antigen recognition during T cell development in the thymus, and later in the periphery.

Here, we studied the organization of TCR repertoires using high-throughput TCR sequencing, comparing data from mice and humans. We focused on the CDR3 amino acid (AA) sequence of the TCRβ chain, which is the most diverse segment of the TCR and is positioned to interact with the antigenic peptide epitope presented by an MHC molecule (Davis and Bjorkman 1988). The organization of TCR repertoires of individual mice and humans was evaluated using network analysis, where CDR3 sequences were connected based on their level of sequence similarity.

Results:

Initially, we constructed TCR networks from a dataset of TCRβ AA sequences obtained from splenic CD4+ T cells from 12 healthy C57BL/6 mice (Madi et al. 2014). We obtained on average about 30,000 different CDR3 sequences from each mouse, which were found at varying abundances and had an average length of 13.4±1.4 (mean±SD) AA. **Figure 1A** shows a network obtained using the thousand most frequent CDR3 sequences from a single mouse, which in terms of abundance correspond to 34% of the total sequences obtained for that mouse. CDR3 sequences (nodes) were connected (by edges) if they were separated by one amino acid difference (replacement / addition / deletion of one AA) – a Levenshtein distance of 1(Levenshtein 1966). A cluster was defined as a set of two or more nodes that are connected to each other by any number of edges and intermediate nodes (**Fig. 1A**, inset). A similar analysis had previously revealed the existence of networks of B-cell immunoglobulin heavy-chains, which were attributed to clonally derived sequences generated by somatic hyper-mutations (SHM) (Ben-Hamo and Efroni 2011; Bashford-Rogers et al. 2013). Our analysis demonstrated the existence of networks also for TCRβ sequences. As T cells do not undergo SHM, other factors lead to the formation of TCR similarity networks.

We repeated this analysis for all 12 mice, and found that of the thousand most frequent CDR3 sequences in each mouse (with an accumulated frequency of 34.5±8% of total sequences), 647±104 (mean±SD) were clustered, with 1282±383 edges. In contrast, networks composed of a thousand randomly selected CDR3 sequences from a single mouse (with an accumulated frequency of 5±0.7% of total sequences) were much sparser (**Fig. 1B**), with only 225±64 sequences clustered, and with 152±52 edges (average values for 10 independent randomized sets of sequences). These results were not sensitive to the number of sequences used for the analysis (**Fig. 1S1**).

To contrast the TCR networks with their BCR counterparts, we tested whether these networks are structurally similar. BCR networks have been shown to center around highly abundant clones, representing a snapshot of the individual-specific local evolution driven by SHM. However, we found no correlation (R^2=0.11±0.07) between the abundance of a TCR CDR3 sequence and its degree of connectivity in the network (number of edges connecting it to other sequences). We further found that each cluster typically contained sequences of a single (or in some cases two) specific J segment (**Fig. 1S2**). V usage, in contrast, was not cluster-specific; any cluster contained sequences with many different V segments (**Fig. 1S2**). This reflects the higher number of V segments compared with J segments, as well as their lower overlap with CDR3 and the relative similarity of their 3' ends. Networks of similar connectivity were obtained also for the top1000 CDR3β sequences from CD8 T cells, and for CD4 T cells of a different mouse strain (C3H.HeSnJ), that bears a different MHC haplotype ($H2^k$; **Fig. 1S3**, **Fig. 1S4**).

We found a parallel network organization also in human TCRβ repertoires: we analyzed previously published data containing the TCRβ repertoires of 39 human subjects of different ages (Britanova et al. 2014), and found that the most abundant CDR3 sequences formed connected clusters in human TCR repertoires (**Fig. 1C**, **Supplementary File 1**, and **Fig. 1S1**), though with a lower connectivity than that found in the similarity networks of inbred mice. From the thousand most frequent CDR3 sequences (accumulated frequency of 17.1%±6.6% of total sequences) in each of the 11 young human subjects in that study (ages 6-25 years), 207±79 nodes were clustered, with 367±201 edges. Networks composed of randomly selected sequences from the individual subjects generated only 8±4 clustered nodes with 4±2 edges. We thus conclude that these newly discovered TCR similarity networks are likely to be driven by conserved evolutionary forces, as opposed to BCR networks that are generated by SHM that operates within individuals.

Next, we tested whether these TCR networks reflect our previous finding that TCRβ CDR3 AA sequences express a range of sharing levels between individual mice. As a measure of sharing level, we used a reference dataset of 28 mice (Madi et al. 2014) and assigned to each CDR3 AA sequence in a network a sharing level ranging from 1 (private, found in only one mouse in the reference dataset) to 28 (public, found in all 28 mice in the reference dataset) (Madi et al. 2014). Interestingly, we found a strong association between the sharing level of a CDR3 sequence and its connectivity in the network: highly shared sequences are positioned at the center of network clusters (**Fig. 1A**). This is indicated by a statistically significant correlation between the degree of node connectivity (number of edges connecting it to other nodes in the network) and its sharing level (**Fig. 1D**), ($R=0.69\pm0.03$, P-value $< 2.2e-16$; see also **Supplementary File 1**). An independent method for estimation of node centrality, betweenness centrality, confirmed the correlation between CDR3 sharing and centrality for the 1,000 most abundant CDR3 sequences, but not for a random set of expressed sequences (**Fig. 1S5**, **Supplementary File 1**). As in mice, public CDR3 sequences in humans manifested a higher degree of connectivity than did more private sequences (**Fig. 1C**, **Fig. 1S6**), and sequence abundance was not correlated with its level of connectivity (**Supplementary File 1**). Thus, private and public CDR3 sequences are distributed differently across the mouse and human networks: public sequences are highly connected to other similar sequences and are more central in network clusters; in contrast, more private sequences are found at the edges of clusters, or as un-connected nodes, with rare similarity to other sequences in the network.

These findings of a similar organization of mouse and human TCR networks prompted us to look for the existence of shared CDR3β sequences between the two species. Interestingly, we found that a substantial number of TCRβ CDR3 AA sequences were shared by mice and humans. Out of 5,247,785 unique AA sequences in the human dataset (11 young individuals) and 371,977 in the mouse dataset (28 animals), 27,337 were shared by at least one mouse and one human individual. In general, CDR3 sequences with a higher level of sharing in mice were found to have an increased probability of being found in human repertoires; similarly, sequences more shared in humans were found more frequently in mice (**Fig. 2A**, **Fig. 2S1**). Of note, more than 25% of the public CDR3 sequences (found in all 11 young human subjects, or found in all 28 mice) were found also in at least one individual of the other species (**Fig. 2A**).

We defined a set of cross-species (CS) public CDR3 sequences that were public or relatively public in both mice (found in at least 25 of the 28 mice) and humans (found in all 11 young individuals). All these 86 CS-public sequences contained the human Jβ2.7 or Jβ2.3 segments, and the mouse Jβ2.5 or Jβ2.7 segments. V usage was dominated by Vβ20.1 in humans, but a more diverse V usage was observed in mice. Examples of CS-public sequences are shown in **Fig. 2B.** The CS-public CDR3 sequences manifested a significantly higher degree of connectivity in human and mouse networks than did CDR3 sequences that were public only in humans, only in mice or not public in either (**Fig. 2C, D** and **Fig. 2S2**). Moreover, we found a significant correlation between the mean degrees of CS-public sequences in mouse and human networks (**Fig. 2S3**); CS-public sequences that have more neighbors in mouse networks also tended to have more neighbors in human networks, suggesting an evolutionarily conserved network structure. We note that while CS-public sequences are central in network clusters, their frequency is not higher than that of other public sequences that are found only in humans or in mice. These findings propose that similar driving forces may generate and expand particular public CDR3 TCR sequences that contain conserved sequence motifs in the two species.

To further characterize the mechanisms that contribute to the generation of CS-public sequences, we evaluated their existence in synthetic TCR repertoires that simulate the random generation of TCR sequences (see methods). These simulations do not include any clonal selection, thus allow discriminating between genetic mechanisms that influence the generation of TCRs and selection mechanisms that shape it somatically. We generated 100 datasets of simulated repertoires of 28 mice and 11 humans, the sizes of which matched the sizes of the experimental repertoires. The simulated repertoires contained a somewhat larger number of CS-public CDR3 sequences than observed in the experimental data (average of 221±9 in the simulations, vs. 86 in the data). The simulated CS-public sequences contained the same restricted set of mouse and human J segments, which are highly similar between the two species (J2.7 mouse and human; J2.5 mouse/J2.3 human). Thus, sequence homology of J segments contributes to the formation of CS-public TCRs, but is not sufficient by itself, and is accompanied by other mechanisms that induce bias in the recombination process (e.g. biased V segment usage, statistics of nucleotide deletions and insertions at V-D and D-J junctions). We also asked whether the simulated repertoires contained the same CS-public sequences as those observed experimentally. We found that 54 out of the 86 experimentally observed CS-public sequences were identical to simulated CS-public sequences, while 32 were not CS-public in the simulations (**Fig. 2S4**). The partial overlap between simulations and data may result from inaccuracies in the assumptions of the simulations regarding the random TCR generation process, or indicate that selection mechanisms in the thymus and in the periphery further influence the existence of specific CS-public sequences.

We further evaluated the similarity between public sequences by analyzing the level of connectivity within a network composed of the most highly shared CDR3 sequences. A network formed by the 1,000 most abundant public mouse sequences (found in >25 of the 28 mice) was highly connected, with 965 clustered nodes and 3,387 edges (**Fig. 3A**). In contrast, networks formed by the 1,000 most abundant *private* sequences (found in only one of the 28 mice) were very sparse, manifesting only 38±15 clustered nodes and 20±7 edges (mean±SD, averaged over 28 mice). Similarly, a network formed by the 1,000 most public human CDR3 sequences was also highly connected (with 969 clustered nodes and 4,398 edges, **Fig. 3B**).

The functional TCR is formed by a complex of TCR alpha and beta chains (Davis and Bjorkman 1988), hence one cannot attribute specific antigen recognition to CDR3β segments alone. Moreover, the current level of understanding precludes the development of general predicting tools that can computationally relate a TCR sequence to an antigen that it recognizes. Defining TCR antigen specificity is further complicated by substantial TCR cross-reactivity (Burrows et al. 1997; Wooldridge et al. 2012). Yet, TCRβ sequences that bind the same pMHC antigen do contain shared CDR3β sequence motifs (Klinger et al. 2015; Chen et al. 2017; Sun et al. 2017; Tickotsky et al. 2017). Thus, some insight on antigen specificity can be gained by linking the sequence-similarity networks to previously annotated TCR sequences. We have reported that 124 of the CDR3β sequences in our mouse dataset were associated with various mouse immune reactivities previously described in the literature (Madi et al. 2014). As a step towards relating antigen specificity to the clusters of public CDR3 sequences, we looked for these 124 annotated CDR3β sequences within the clusters of shared CDR3 sequences. The annotated sequences were grouped according to four categories: a) Immunity to foreign pathogens; b) Allograft reactions; c) Tumor-associated T cells; and d) Autoimmune conditions. **Figure 3A** includes these annotations in the network formed by the 1,000 most public CDR3β sequences. Out of the 124 annotated sequences, 63 were either identical to one of the existing nodes (n=11), or linked to an existing node by a Levenshtein distance of 1 (n=52). The clustered annotated nodes were found to be enriched with annotations related to self or self-like autoimmune, cancer or allograft reactions (self-related: 51/63 = 81% of network-clustered sequences vs. 85/124=69% in all 124 annotated sequences, compared to non-self: 12/63 = 19% in clusters vs. 39/124=31%; Fisher exact test p=0.0035).

We find that sequences with a similar annotation tended to be linked in the same cluster. Examples include twelve sequences of tumor infiltrating regulatory T cells (Sainz-Perez et al. 2012) which were found in cluster #2; six COPD related CDR3 sequences (Motz et al. 2008) in cluster #6; and four CDR3 sequences connected with cluster #2 that were associated with type 1 diabetes in NOD mice in two different studies (Nakano et al. 1991; Tikochinski et al. 1999). However, different annotations can also be found in the same cluster (**Fig. 3A**); for example, mouse CDR3 sequences associated with experimental autoimmune encephalomyelitis (EAE; (Menezes et al. 2007)) and collagen-induced arthritis (CIA;(Osman et al. 1993)) were also connected to cluster #2. **Figure 3B** shows that many previously annotated self/self-like sequences of humans and mice were also linked to clusters in the network of public human sequences. Thus, the CDR3 clusters, which serve as repertoire foci, seem to be enriched with TCR sequences that are associated with self (or self-like) reactivates, whereas pathogen-associated TCR sequences are less clustered and so tend to be more evenly spread throughout sequence space.

To analyze mechanisms involved in network cluster formation, we investigated the contribution of antigen selection using two complimentary approaches. First, we analyzed similarity networks formed by CDR3 sequences of CD4$^-$CD8$^-$ double-negative (DN) thymocytes. Rearranged TCRβ chains in DN cells are not subject to MHC-dependent selection, which only occurs at later stages of thymic development. We found that networks formed by DN CDR3 sequences were significantly less connected compared to splenic CD4$^+$ T cells, which have undergone antigen selection (**Fig. 4A** and **Supplementary File 2**). In addition, DN thymocytes and CD4$^+$ spleen T cells manifested different levels of convergent recombination (Venturi et al. 2006; Venturi et al. 2008). Public CDR3 AA sequences in DN thymocytes were encoded on average by a low number of nucleotide (nt) sequences, whereas the same AA sequences were encoded by a much larger number of nt sequences in CD4$^+$ splenic T cells (**Fig. 4B**, **Fig. 4S1**). The finding of relatively increased network clusters in T cells that have undergone antigen selection suggests that the CDR3 AA sequences that are found within clusters are positively selected; this antigen selection would extend any underlying physical bias generated during TCR DNA recombination in the thymus (Murugan et al. 2012; Ndifon et al. 2012).

To further evaluate the impact of selection, we evaluated TCR networks formed in the repertoires of splenic T cells from mice lacking four elements needed for physiological MHC-dependent antigen selection: MHC-I and -II molecules together with CD4 and CD8 co-receptor molecules, so-called Quad-KO mice (Van Laethem et al. 2007; Van Laethem et al. 2013). In contrast to wild-type (WT) mice, the TCR of Quad-KO mice are selected by MHC-independent ligands in the thymus and their T cells express a diverse MHC-independent TCR repertoire in the periphery (Van Laethem et al. 2007; Tikhonova et al. 2012; Van Laethem et al. 2013). We found that similarity networks formed by the top 1,000 CDR3 sequences from Quad-KO mice were significantly less connected than were those of the WT strain (C57BL/6) measured in the same set of experiments (**Fig. 4A** and **Supplementary File 2**). Together, these findings indicate that MHC-dependent thymic selection plays a significant role in promoting the formation of dense clusters of TCR-similarity networks. Lack of MHC-dependent selection in DN thymocytes and in Quad-KO mice is associated with TCR networks of reduced connectivity; in contrast, TCRs that are subject to MHC selection form dense networks with a higher level of convergent recombination. Thus, recombination biases combined with clonal selection generate a TCR repertoire that is not uniform, but rather focused in specific regions of sequence space that are preferentially associated with self-related antigen-reactivities.

Following these observations, we tested if the relative abundance of CS-public clonotypes is increased by MHC-dependent selection. Thus, we compared the frequency of CS-public sequences in repertoires of Quad-KO mice and DN thymocytes to those of control WT mice (**Fig. 4**B). The cumulative frequencies of the CS-public CDR3 sequences between two sets of experiments done with WT mice (the 28 WT mice used in the network analysis, and the WT mice used as controls in the Quad-KO experiment) show no significant difference (P value = 0.293). On the other hand, the Quad-KO repertoires exhibited lower total frequency of the CS-public CDR3s compared with both 28 WT mice (P value = 4.318e-09) and the Quad-WT mice (P value = 0.01781). The cumulative frequency in the DN shows a similar trend, with no statistical significant (P value = 0.1877). Together, these results indicate that, although sequence homology of V and J germline segments and bias in the recombination process influence the probability for a sequence to be shared between the two species, additional selection forces are influencing its abundance.

Since the composition of the TCR repertoire of an individual changes in response to immune challenges throughout life, we tested the effects of both immunization and aging on the network organization of the TCR repertoire. We immunized naïve mice with p277, a self peptide derived from HSP60, or with a foreign peptide, derived from ovalbumin (OVA). Peptide p277 was previously found to be recognized by the C9 public TCR in NOD mice (Tikochinski et al. 1999), and the CDR3β sequence of the C9 clone was also public in C57BL/6 mice (Madi et al. 2014). Additionally, we analyzed the network structures in the TCR repertoires of T cells from the immunized mice that were cultured *in vitro* with antigen presenting cells loaded with the specific peptide. The distribution of sequence abundances and repertoire evenness were evaluated using the Gini inequality coefficient, which ranges from 0 for a repertoire where every sequence is present in equal abundance, to 1 for a repertoire dominated by a single sequence, with other sequences present at zero abundance (Bashford-Rogers et al. 2013; Thomas et al. 2013).

We found that immunization with either peptide resulted in repertoires that contained a set of expanded CDR3 sequences and had an increased abundance inequality. *In vitro* re-stimulation further increased inequality (**Fig. 5A-C** and **Supplementary File 3**). This inequality was associated with the emergence of private clones that dominated the post-immunization repertoire, such that the relative weight of public clones was reduced (**Fig. 5E**). Interestingly, immunization was also associated with network disruption; the number of clustered nodes and the number of edges both fell after immunization *in vivo* and fell further after *in vitro* re-stimulation (**Fig. 5D**, **Fig. 5S1**). Both the increased inequality and the decreased network connectivity reversed spontaneously in the OVA-immunized mice 2 months following immunization (**Fig. 5D, E** (right), **Fig. 5S1**). Similar to immunization, repertoires in aged mice (**Fig. 5F**, **Fig. 5S2**) and in aged humans (**Fig. 5G**, **Fig. 5S3**) were more unequal and less connected than those of young individuals, and private CDR3 sequences became relatively more abundant with age (**Fig. 5S4**). Altogether, we found a strong anti-correlation between the Gini Coefficient of TCR inequality and the number of connected nodes in TCR networks in mice (**Fig. 5F**, Spearman correlation = -0.661) and in humans (**Fig. 5G**, Spearman correlation = -0.865).

Another factor that impacted network structure was immune checkpoint blockade. We used published CDR3β sequence data (Robert et al. 2014) from subjects who had undergone CTLA4 (cytotoxic T–lymphocyte-associated protein 4) blockade with tremelimumab. Previous analysis of these data showed that this treatment diversified the peripheral T-cell pool. Applying TCR similarity network analysis, we now show that the 1000 most abundant CDR3 sequences after check-point blockade are less connected than pre-treatment (P value<0.05 ranked Wilcox paired test, **Fig. 5H** left); moreover, this reduction in connectivity was detected concurrently with a decrease in the number of public CDR3 sequences and an increase in the frequency of private ones (p-value = 0.01947, ranked Wilcox paired test, **Fig. 5H** right, **Fig. 5S5**). Thus, broadening of the peripheral repertoire following CTLA4 blockade reduces the presence of public clones and enhances the expansion of private clones, similar to the changes we observed in aging or after immunization. This finding raises the possibility that check-point associated immune regulation also could be involved in the prominence of network connectivity of public T cells. Finally, we analyzed TCR repertoires of patients with the autoimmune disease Juvenile Idiopathic Arthritis (JIA)(Henderson et al. 2016). We found that there was a strong increase of public (network promoting) TCRs in the peripheral blood of JIA patients compared to healthy donors (P value = 0.0006, **Fig. 5I**). Thus, while immune perturbations such as immunization and aging lead to reduced levels of public clonotypes and network reduction, this specific autoimmune condition is associated with an increased level of public clones which are putatively associated with self-antigens.

Discussion:

Our application of network analysis to TCRβ CDR3 sequencing data reveals a hitherto unrecognized structure of the TCR repertoire in both mice and humans: In young, healthy individuals, the most abundant TCRβ CDR3 sequences are distributed unevenly in sequence-space, with clusters centered around public CDR3s, and in particular around CS-public sequences, which are public both in mice and humans (**Fig. 5J** top-right, even and focused repertoire). The clustering of the most abundant CDR3 sequences in young and healthy individuals results in a repertoire that is much more restricted than would be expected from the random process of TCR somatic recombination. This basic network architecture is modified by immunization and aging due to the dominant expansion of more private CDR3 clonotypes. Thus, public CDR3s that serve as hubs of TCR networks become less prominent, leading to reduced connectivity of TCR networks combined with a more skewed repertoire (**Fig. 5J** bottom-left, skewed and spread repertoire). We find that network organization and repertoire evenness are restored with the resolution of immune responses. It might be the case that incomplete resolution of immune responses throughout life lead to accumulation of changes in the TCR repertoire that eventually result in the skewed and spread (less clustered) repertoires that we observe in aged individuals. Interestingly, TCR repertoires from patients with the autoimmune condition JIA showed increased levels of public TCR sequences. This aligns with our observation that public TCR networks are enriched with self-associated TCRs. Taken together, our analysis supports the idea that the level of network connectivity, frequency of public TCRs and repertoire evenness are linked to each other, and are concurrently modulated by the individual's immune state (disease / immunization / aging).

Mechanistically, we found that MHC-dependent antigen selection contributes to the formation of dense networks, since reduced network connectivity was observed in pre-selection DN thymocytes and also by inhibiting MHC-dependent selection, in the Quad-KO mice. These results can be explained by preferential selection and increased survival, in both the thymus and periphery, of T cells that carry specific CDR3 sequences that recognize self-antigens presented by MHC molecules. Different T cell clones, which carry different CDR3 nt sequences but encode the same AA sequence, would appear to enjoy a common selective advantage and accumulate in the peripheral repertoire. This mechanism can explain our observations of increased convergent recombination in splenic CD4$^+$ T cells compared to DN thymocytes (**Fig. 4A**). Antigen selection can also account for the enhanced network connectivity of TCRs that differ by one AA in their CDR3 sequences; such related CDR3 sequences can be selected by the same peptide-MHC complex, albeit with different affinities (Moss et al. 1991; Serana et al. 2009; Zoete et al. 2013). This working hypothesis needs to be tested experimentally to see if linked CDR3 sequences really cross-react with the same or similar peptide-MHC complexes. MHC-antigen selection of public CDR3 sequences takes place on a background of biases in the biophysical process of DNA recombination (Elhanati et al. 2014). Combined, these processes lead to the formation of dense network clusters of the most abundant public TCR sequences, as we report here. In contrast, the most abundant private TCR sequences generate poorly connected networks. B cell receptor (BCR) sequences (Ben-Hamo and Efroni 2011; Bashford-Rogers et al. 2013), unlike the T-cell repertoire networks we disclose here, have long been known to generate networks in individual subjects by affinity maturation that is mediated by SHM; T cells do not undergo SHM so TCR networks must be generated in the developmental process. Thus, dominant and public T cell clonotypes have a higher sequence similarity than non-dominant and private ones. In contrast, BCR networks have a distinct structure resulting from the SHM process, in which abundance and degree are correlated, which is not the case in TCR networks.

Our finding that TCR CDR3 networks include identical and related sequences that are not confined to individuals but are shared by most individuals of the same species and even cross the species divide between mice and humans, suggests the likelihood of some fundamental evolutionary advantage in such sequences. As noted above, antigen specificity of a TCR cannot be defined based on its CDR3β alone. However, the same or very similar CDR3β sequences are frequently observed within repertoires of T cells specific for a given antigen, in combination with flexible or preferential pairing with TCRα (Klinger et al. 2015; Chen et al. 2017; Tickotsky et al. 2017). Hence, we hypothesize that T cell clones bearing the conserved, CS-public, CDR3 sequences recognize similar antigenic epitopes that are conserved across species. These antigens may be derived from evolutionarily conserved regions of self proteins, forming a core of T cell reactivities to specific self epitopes, with potential implications for self-maintenance, autoimmunity and cancer. Further studies relating TCRα, TCRβ and peptide specificity will enable to experimentally test this hypothesis.

Our results indicate that T lymphocytes "focus their attention" to specific regions in sequence space. These new findings on the organization of TCR repertoires and their dynamics raise intriguing questions, for example, does the existence of network clusters indicate a healthy immune state? Can restoration of network structure reinstate immune function in the elderly or prevent excess inflammation and autoimmune disease? The theory of the immunological homunculus composed of self-recognizing B cells and T cells (Cohen 1992; Cohen 2000) might be relevant here.

Materials and methods:

Mice

Female 5–8 weeks old C57BL/6 mice were obtained from Harlan Laboratories. Analysis of TCR sequences from aged mice is based on data that was previously described in Shifrut et al., 2013 (Shifrut et al. 2013). Analysis of TCR sequences from repertoires which are not subject to MHC-dependent selection, is based on Quad-KO mice, which are lacking four elements needed for physiological MHC-dependent antigen selection: MHC-I and -II molecules together with CD4 and CD8 co-receptor molecules, and matched control WT mice (Van Laethem et al. 2007; Van Laethem et al. 2013) and DN thymocytes, which represent the landscape of generated TCRs before thymic selection.

Human data used in this study

Dataset of 39 healthy Caucasian donors, ages 6–90 years, was obtained from Britanova et. al., 2014 (Britanova et al. 2014; Robert et al. 2014). CTLA4 blockade data was obtained from (Robert et al. 2014). Juvenile Idiopathic Arthritis (JIA) data of patients compared to healthy donors was obtained from (Henderson et al. 2016).

Immunization and *in vitro* stimulation

Mice were injected intra-peritonealy (IP) with 100µg of either Chicken Ovalbumin (OVA) or peptide 277 (p277) emulsified in CFA (1:1 ratio). Spleens were harvested on day 7 post immunization and T cells were extracted for TCR analysis. *in vitro* stimulation: T cells from spleens of immunized mice were harvested on day 7 and were re-stimulated with irradiated splenocytes and the relevant peptide antigen. Five of the OVA-immunized mice received a boost IP injection of 100µg OVA+CFA on day 14, and spleens were harvested on day 60 for TCR analysis (**Supplementary File 3**).

Library preparation for TCR-seq and data pre-processing

Libraries were prepared and pre-processed as published(Ndifon et al. 2012). Briefly, T cells were purified from splenocytes by magnetic bead separation, total RNA was extracted and reverse transcribed using a TCR Cβ-specific primer linked to the 3'-end Illumina sequencing adapter. cDNA was amplified using PCR with a Cβ-3'adpater primer and a set of 20 Vβ-specific 5' primers, followed by ligation of a 5'Illumina adaptor and a 2nd PCR using universal primers for the 5' and 3' Illumina adapters. The libraries were sequenced using Genome Analyzer II or HiSeq 2000 (Illumina). Sequence filtering, VDJ annotation, normalization and translation to AA sequences were performed as published (Ndifon et al. 2012). Libraries for TCR-seq of Quad mice and C57BL/6 controls were sequenced using Illumina sequencers, performed by Adaptive Biotechnologies Corp (Seattle, WA). In brief, αβT cells were isolated by cell sorting, washed in PBS and lysed in Trizol. RNA was extracted using the RNEasy protocol (Qiagen) and 2 µg per sample reverse transcribed to cDNA by oligo (dT) priming with the SuperScript TM III First-Strand Synthesis System (Invitrogen). cDNA was sequenced by Adaptive Biotechnologies Corp.

Statistical analysis and visualization

Statistical analysis was performed using R Software (Core Team R 2013). We used the following packages: "ShortRead" (Morgan et al. 2009) for the pre-processing pipeline; "ineq" (Zeileis 2012) and "reldist" (Handcock 2014) to calculate the Gini coefficient; "Igraph" (Csardi and Nepusz 2006) to create network objects, obtain the degree of a node and its betweeness; "stringdist" (van der Loo 2014) to calculate Levenshtein distances; and "ggplot2" (Wickham 2009) for generating figures. Statistical tests performed are stated in the text. All network figures were made using Cytoscape (http://www.cytoscape.org/) (Cline et al. 2007; Smoot et al. 2011; Saito et al. 2012).

Data access

The sequence data from this study have been made publicly available (https://usegalaxy.org/u/erezgrn/h/network-tcrs).

References

Bashford-Rogers RJ, Palser AL, Huntly BJ, Rance R, Vassiliou GS, Follows GA, Kellam P. 2013. Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome Res* **23**(11): 1874-1884.

Ben-Hamo R, Efroni S. 2011. The whole-organism heavy chain B cell repertoire from Zebrafish self-organizes into distinct network features. *BMC Syst Biol* **5**: 27.

Britanova OV, Putintseva EV, Shugay M, Merzlyak EM, Turchaninova MA, Staroverov DB, Bolotin DA, Lukyanov S, Bogdanova EA, Mamedov IZ et al. 2014. Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. *J Immunol* **192**(6): 2689-2698.

Burrows SR, Silins SL, Khanna R, Burrows JM, Rischmueller M, McCluskey J, Moss DJ. 1997. Cross-reactive memory T cells for Epstein-Barr virus augment the alloresponse to common human leukocyte antigens: degenerate recognition of major histocompatibility complex-bound peptide by T cells and its role in alloreactivity. *Eur J Immunol* **27**(7): 1726-1736.

Chen G, Yang X, Ko A, Sun X, Gao M, Zhang Y, Shi A, Mariuzza RA, Weng NP. 2017. Sequence and Structural Analyses Reveal Distinct and Highly Diverse Human CD8+ TCR Repertoires to Immunodominant Viral Antigens. *Cell Rep* **19**(3): 569-583.

Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B et al. 2007. Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* **2**(10): 2366-2382.

Cohen IR. 1992. The cognitive principle challenges clonal selection. *Immunol Today* **13**(11): 441-444.

Cohen IR. 2000. *Tending Adam's Garden: Evolving the Cognitive Immune Self.* Academic Press, London.

Core Team R. 2013. R: A Language and Environment for Statistical Computing. Vienna.

Csardi G, Nepusz T. 2006. The igraph software package for complex network research. *InterJournal, Complex Systems* **1695**.

Davis MM, Bjorkman PJ. 1988. T-cell antigen receptor genes and T-cell recognition. *Nature* **334**(6181): 395-402.

Elhanati Y, Murugan A, Callan CG, Jr., Mora T, Walczak AM. 2014. Quantifying selection in immune receptor repertoires. *Proc Natl Acad Sci U S A* **111**(27): 9875-9880.

Handcock MS. 2014. Relative Distribution Methods. **Version 1.6-3**.

Henderson LA, Volpi S, Frugoni F, Janssen E, Kim S, Sundel RP, Dedeoglu F, Lo MS, Hazen MM, Beth Son M et al. 2016. Next-Generation Sequencing Reveals Restriction and Clonotypic Expansion of Treg Cells in Juvenile Idiopathic Arthritis. *Arthritis Rheumatol* **68**(7): 1758-1768.

Klinger M, Pepin F, Wilkins J, Asbury T, Wittkop T, Zheng J, Moorhead M, Faham M. 2015. Multiplex Identification of Antigen-Specific T Cell Receptors Using a Combination of Immune Assays and Immune Receptor Sequencing. *PLoS One* **10**(10): e0141561.

Levenshtein VI. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* **10**(8): 707–710.

Madi A, Shifrut E, Reich-Zeliger S, Gal H, Best K, Ndifon W, Chain B, Cohen IR, Friedman N. 2014. T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. *Genome Res* **24**(10): 1603-1612.

Menezes JS, van den Elzen P, Thornes J, Huffman D, Droin NM, Maverakis E, Sercarz EE. 2007. A public T cell clonotype within a heterogeneous autoreactive repertoire is dominant in driving EAE. *J Clin Invest* **117**(8): 2176-2185.

Miles JJ, Douek DC, Price DA. 2011. Bias in the alphabeta T-cell repertoire: implications for disease pathogenesis and vaccination. *Immunol Cell Biol* **89**(3): 375-387.

Morgan M, Anders S, Lawrence M, Aboyoun P, Pages H, Gentleman R. 2009. ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* **25**(19): 2607-2608.

Moss PA, Moots RJ, Rosenberg WM, Rowland-Jones SJ, Bodmer HC, McMichael AJ, Bell JI. 1991. Extensive conservation of alpha and beta chains of the human T-cell antigen receptor recognizing HLA-A2 and influenza A matrix peptide. *Proc Natl Acad Sci U S A* **88**(20): 8987-8990.

Motz GT, Eppert BL, Sun G, Wesselkamper SC, Linke MJ, Deka R, Borchers MT. 2008. Persistence of lung CD8 T cell oligoclonal expansions upon smoking cessation in a mouse model of cigarette smoke-induced emphysema. *J Immunol* **181**(11): 8036-8043.

Murugan A, Mora T, Walczak AM, Callan CG, Jr. 2012. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc Natl Acad Sci U S A* **109**(40): 16161-16166.

Nakano N, Kikutani H, Nishimoto H, Kishimoto T. 1991. T cell receptor V gene usage of islet beta cell-reactive T cells is not restricted in non-obese diabetic mice. *J Exp Med* **173**(5): 1091-1097.

Ndifon W, Gal H, Shifrut E, Aharoni R, Yissachar N, Waysbort N, Reich-Zeliger S, Arnon R, Friedman N. 2012. Chromatin conformation governs T-cell receptor Jbeta gene segment usage. *Proc Natl Acad Sci U S A* **109**(39): 15865-15870.

Osman GE, Toda M, Kanagawa O, Hood LE. 1993. Characterization of the T cell receptor repertoire causing collagen arthritis in mice. *J Exp Med* **177**(2): 387-395.

Robert L, Tsoi J, Wang X, Emerson R, Homet B, Chodon T, Mok S, Huang RR, Cochran AJ, Comin-Anduix B et al. 2014. CTLA4 blockade broadens the peripheral T-cell receptor repertoire. *Clin Cancer Res* **20**(9): 2424-2432.

Robins HS, Srivastava SK, Campregher PV, Turtle CJ, Andriesen J, Riddell SR, Carlson CS, Warren EH. 2010. Overlap and effective size of the human CD8+ T cell receptor repertoire. *Sci Transl Med* **2**(47): 47ra64.

Sainz-Perez A, Lim A, Lemercier B, Leclerc C. 2012. The T-cell receptor repertoire of tumor-infiltrating regulatory T lymphocytes is skewed toward public sequences. *Cancer Res* **72**(14): 3557-3569.

Saito R, Smoot ME, Ono K, Ruscheinski J, Wang PL, Lotia S, Pico AR, Bader GD, Ideker T. 2012. A travel guide to Cytoscape plugins. *Nat Methods* **9**(11): 1069-1076.

Serana F, Sottini A, Caimi L, Palermo B, Natali PG, Nistico P, Imberti L. 2009. Identification of a public CDR3 motif and a biased utilization of T-cell receptor V beta and J beta chains in HLA-A2/Melan-A-specific T-cell clonotypes of melanoma patients. *J Transl Med* **7**: 21.

Shifrut E, Baruch K, Gal H, Ndifon W, Deczkowska A, Schwartz M, Friedman N. 2013. CD4(+) T Cell-Receptor Repertoire Diversity is Compromised in the Spleen but Not in the Bone Marrow of Aged Mice Due to Private and Sporadic Clonal Expansions. *Front Immunol* **4**: 379.

Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. 2011. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**(3): 431-432.

Sun Y, Best K, Cinelli M, Heather JM, Reich-Zeliger S, Shifrut E, Friedman N, Shawe-Taylor J, Chain B. 2017. Specificity, Privacy, and Degeneracy in the CD4 T Cell Receptor Repertoire Following Immunization. *Front Immunol* **8**: 430.

Thomas PG, Handel A, Doherty PC, La Gruta NL. 2013. Ecological analysis of antigen-specific CTL repertoires defines the relationship between naive and immune T-cell populations. *Proc Natl Acad Sci U S A* **110**(5): 1839-1844.

Tickotsky N, Sagiv T, Prilusky J, Shifrut E, Friedman N. 2017. McPAS-TCR: A manually-curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics*.

Tikhonova AN, Van Laethem F, Hanada K, Lu J, Pobezinsky LA, Hong C, Guinter TI, Jeurling SK, Bernhardt G, Park JH et al. 2012. alphabeta T cell receptors that do not undergo major histocompatibility complex-specific thymic selection possess antibody-like recognition specificities. *Immunity* **36**(1): 79-91.

Tikochinski Y, Elias D, Steeg C, Marcus H, Kantorowitz M, Reshef T, Ablamunits V, Cohen IR, Friedmann A. 1999. A shared TCR CDR3 sequence in NOD mouse autoimmune diabetes. *Int Immunol* **11**(6): 951-956.

van der Loo M. 2014. The stringdist package for approximate string matching. *The R Journal* **6**: 111-122.

Van Laethem F, Sarafova SD, Park JH, Tai X, Pobezinsky L, Guinter TI, Adoro S, Adams A, Sharrow SO, Feigenbaum L et al. 2007. Deletion of CD4 and CD8 coreceptors permits generation of alphabetaT cells that recognize antigens independently of the MHC. *Immunity* **27**(5): 735-750.

Van Laethem F, Tikhonova AN, Pobezinsky LA, Tai X, Kimura MY, Le Saout C, Guinter TI, Adams A, Sharrow SO, Bernhardt G et al. 2013. Lck availability during thymic selection determines the recognition specificity of the T cell repertoire. *Cell* **154**(6): 1326-1341.

Venturi V, Kedzierska K, Price DA, Doherty PC, Douek DC, Turner SJ, Davenport MP. 2006. Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. *Proc Natl Acad Sci U S A* **103**(49): 18691-18696.

Venturi V, Price DA, Douek DC, Davenport MP. 2008. The molecular basis for public T-cell responses? *Nat Rev Immunol* **8**(3): 231-238.

Wickham H. 2009. *Ggplot2: Elegant Graphics for Data Analysis.* Springer, New York.

Wooldridge L, Ekeruche-Makinde J, van den Berg HA, Skowera A, Miles JJ, Tan MP, Dolton G, Clement M, Llewellyn-Lacey S, Price DA et al. 2012. A single autoimmune T cell receptor recognizes more than a million different peptides. *J Biol Chem* **287**(2): 1168-1177.

Zeileis A. 2012. *Ineq: Measuring Inequality, Concentration, and Poverty.*

Zoete V, Irving M, Ferber M, Cuendet MA, Michielin O. 2013. Structure-Based, Rational Design of T Cell Receptors. *Front Immunol* **4**: 268.

Acknowledgements:

We thank Benjamin Chain and Shalev Itzkovitz for helpful comments on the manuscript. This research was supported by grants from the Minerva Foundation with funding from the Federal German Ministry for Education and Research and the I-CORE Program of the Planning and Budgeting Committee and the Israel Science Foundation. A.M. was supported by the M.D. Moross Institute for Cancer Research.

Figure Legends

Figure 1. M**ouse and human TCR repertoires manifest dense similarity networks surrounding public CDR3β sequences** (A) Networks formed by the thousand most frequent CDR3 AA sequences expressed in a single mouse. Nodes (CDR3 AA sequences) were connected by edges defined by a Levenshtein distance of 1 (one AA substitution / insertion / deletion). Node size reflects its log frequency. The nodes are colored according to their sharing levels in a reference dataset of 28 mice (Madi et al.

2014), from *Private* CDR3 sequences (found in only one mouse in the reference dataset) to *public* (shared by all 28 mice). Inset shows a blowup of the marked cluster with labeled CDR3β AA sequences (nodes) and edges which represent a Levenshtein distance of 1 between connected nodes. (B) Networks formed by a thousand CDR3 sequences randomly chosen from a single naïve mouse. (C) A Network formed by the thousand most frequent CDR3 sequences of a representative human subject (data from (Britanova et al. 2014)). Nodes are colored by their degree of sharing among the 11 young subjects in that study (ages 6-25 years). (D) Mean degree of node connectivity as a function of sharing level in a network formed by the top 1,000 CDR3 sequences (blue) or by 1,000 randomly chosen sequences (orange). Error bars indicate standard error (SE) across the 12 mice used in this study.

Figure 2. TCR repertoires are focused around public and CS-public CDR3 AA sequences shared by mice and humans. (A) Human (left) or mouse (right) CDR3 sequences are grouped according to their sharing level in the corresponding dataset. For each sharing group, we plotted the percentage of sequences that were shared by at least one subject of the other species. (B) Examples of CS-Public CDR3 sequences, and their V and J segments in mouse and human repertoires. (C) A network formed by the top 1,000 CDR3 sequences of a single human subject. Node color represents its sharing within or between species: Pink - shared by all 11 human subjects; Green - shared by at least 25 of the 28 mice; Black – CS-public nodes shared by all 11 humans and at least 25 mice; Blue - not shared. (D) The mean number of edges per node (degree) in the 11 human and 28 mouse networks, subdivided into the four categories as in B. Error bars mark SE.

Figure 3. Public CDR3 sequences form highly connected similarity networks in mice and humans and are enriched for self-associated immune reactivities. (A) A network formed by the 1,000 most shared mouse CDR3 sequences (found in >25 of 28 mice). Node size corresponds to the mean abundance of the sequence. Nodes are colored according to their cluster association. 124 CDR3 sequences that were previously annotated (see (Madi et al. 2014)) were added to the network and are presented as triangles. 63 annotated sequences were either identical to, or at a Levenshtein distance of 1 from one of the nodes, and are listed next to each cluster (with the corresponding color). Annotations of 61 un-clustered sequences are also listed. (B) A network formed by the 1,000 most frequent public CDR3 sequences in humans (found in all 11 subjects). Previously annotated mouse (n=124) and human (n=30) CDR3 sequences were added to the network as in A (triangles). The clusters were distinctly colored in order to visually match between clusters and their annotated sequences, not to define antigen specificity of a cluster. A list of linked annotated CDR3 sequences

is shown next to each cluster (11 of 30 human and 23 of 124 mouse annotated CDR3 sequences), together with a list of unclustered annotated human sequences.

Figure 4. MHC-dependent public CDR3 sequences form highly connected similarity networks. (A) Mean number of clustered nodes in networks formed by the top 1,000 CDR3 sequences from the following repertoires: DN thymocytes (CD4$^-$CD8$^-$) (n=3), CD4$^+$ spleen T cells (n=3), Quad-K mice(Van Laethem et al. 2007) (lack MHC-I and MHC–II, and CD4 and CD8) (n=4), and their WT controls (C57BL/6) (n=4). Error bars signify standard error. (B) Cumulative frequency of the 86 CS-public CDR3 sequences (observed in the datasets of 28 WT mice and 11 healthy humans) is shown for: DN thymocytes (CD4$^-$CD8$^-$) (n=3), CD4$^+$ spleen T cells (n=3) (left), Quad-KO mice (n=4), and their WT controls (C57BL/6) (n=4). Error bars signify standard error. (C) Cumulative frequency of nucleotide sequences coding for two annotated (C9 and COPD, top) and two unknown (bottom) public AA CDR3 sequences from repertoires of DN thymocytes and CD4$^+$ spleen T cells (sequences from 3 mice are shown). Each color represents a different nucleotide sequence.

Figure 5. Immunization, *in vitro* antigen re-stimulation, anti-CTLA4 antibody treatment and aging perturb repertoire networks coupled with an increase in repertoire skewness. (A-C) Networks of the thousand most frequent CDR3 sequences are shown for (A) a naïve mouse, (B) a mouse Immunized with a self-peptide (p277), and (C) T cells from the spleen of an immunized mouse, which were re-stimulated *in vitro* with the p277 peptide. (D) Mean number of clustered nodes in networks formed by the top 1,000 CDR3 sequences from the following repertoires: Left: naïve mice (n=12); p277 immunized mice, 7d post immunization (n=5); and in-vitro re-stimulated with p277 (n=5). Right: naïve mice (n=12); OVA immunized mice, 7d post immunization (n=5); in-vitro re-stimulated with OVA peptide (n=3); and immunized mice, 2 months post-immunization (n=5). Error bars indicate standard error. (E) Frequency of the top 1,000 most frequent CDR3 sequences by sharing level, for the same repertoires as in (D). Sharing levels were calculated based on sharing in the reference dataset of 28 mice. (F) The Gini Coefficient (a measure for repertoire skewness) plotted vs. the number of clustered nodes, for the top 1,000 CDR3 sequences from the repertoires from (D, E) and from aged mice (n=3). (G) The Gini Coefficient plotted vs. the number of clustered nodes for 39 human samples (Britanova et al. 2014) divided into 4 age groups. (H) The number of clustered nodes (left) and the number of public clonotypes (right, shared by all 11 young human samples in a reference cohort (Britanova et al. 2014)) for the top 1,000 most abundant CDR3 sequences in 21 paired samples of patients at baseline and 30 to 60 days after receiving CTLA4 blockade treatment with tremelimumab (data from (Robert et al. 2014)). (I) Number of public clonotypes (defined as in H) out of the top 1,000 most abundant CDR3 sequences in either healthy

donors (left) or Juvenile Idiopathic Arthritis (right) samples. (J) A conceptual figure of the evolution of repertoire structure. In young and healthy individuals the repertoire is focused and even (top-right), with public and CS-public CDR3 sequences at the center of network clusters. Following an immune response, or with aging, the repertoire becomes more skewed and spread in sequence space (bottom-left), due to preferential expansion of private clones at the expanse of more public clones.

Figure 1- figure supplement 1. Mean number of clustered nodes as a function of the sample size selected for generating the network. (Right panel is a zoomed-in version of the left panel). Results are shown for 4 representative conditions, with different levels of observed network connectivity, as expressed by the number of clustered nodes (degree > 0). These graphs show that regardless of sample size, (A, B) networks from a naïve mouse are the most connected, followed by those of immunized (p277), aged mice, and lastly p277 in vitro stimulation, which is the least connected. (C, D) networks for 39 human samples (Britanova et al. 2014) divided into 4 age groups. Above ~1,000 sequences, the trend is linear; hence the relative fraction of clustered nodes is not sensitive to sample size. Thus, our analysis of network connectivity is not sensitive to the number of sequences used.

Figure 1- figure supplement 2. CDR3β sequences form networks with clusters dominated by J-genes and heterogeneous for V-genes. An example of a network constructed from the 1000 most abundant CDR3β AA sequences from a naïve mouse. Both panels show the same network. In the left panel, nodes are colored by the dominating J-gene, in the right panel, the color indicates the dominating V-gene for each AA sequence. Network clusters mostly consist of a single J-gene, with only a few clusters featuring two or three primary J-genes (left). In contrast, V-gene usage in clusters is heterogeneous, with no obvious dominating gene segment (right). This pattern of clusters with homogenous J-gene and heterogeneous V-gene usage was consistent in all top 1000 CDR3β AA sequence networks we examined.

Figure 1- figure supplement 3: CD8^{+} T cell networks formed by the thousand most frequent CDR3 AA segments expressed in two naïve mice. Nodes (CDR3 AA sequences) were connected by edges defined by a Levenshtein distance of 1.

Figure 1- figure supplement 4: Networks from C3H.HeSnJ mouse strain bearing the H2^{k} MHC haplotype. CD4^{+} T cell networks formed by the thousand most frequent CDR3 AA segments expressed in two naïve mice. Nodes (CDR3 AA sequences) were connected by edges defined by a Levenshtein distance of 1.

Figure 1- figure supplement 5. Evaluating the level of node centrality vs. sharing level. The mean betweenness centrality is presented as a function of the sharing level in the dataset of 28 mice, for networks composed of the 1,000 most frequent CDR3 AA sequences and for networks composed of 1,000 randomly selected sequences from the dataset. Error bars indicate standard error (SE) across the 12 mice used in this study.

Figure 1- figure supplement 6. Node centrality vs. sharing level in human TBRβ repertoires. TCRβ repertoires of 11 healthy young human subjects previously investigated by Britanova *et al.*(Britanova et al. 2014). Shown is the mean degree of nodes as a function of their sharing level in the dataset, for networks composed of the most frequent 1,000 CDR3 aa sequences and for networks composed of 1,000 randomly selected sequences. Note that public human TCRs manifest a higher degree of connectivity than do private TCRs.

Figure 2 - figure supplement 1. Cross-species TCR sharing. (A) All CDR3β sequences in the 28 mouse dataset were categorized according to their sharing level, from private (found in only one mouse, n=1), to public (found in all 28 mice). The graph presents the percent of sequences within each category that were also found in the human dataset (11 young subjects). (B) All CDR3β sequences in the 11 young human subjects were categorized according to their sharing level, from private (found in only one subject, n=1), to public (found in all subjects, n=11). The graph presents the percent of sequences within each group that were also found in the 28 mice. In both cases, the fraction of cross-species sequences increases with the sharing level; sequences that are more public in mice (humans) are more frequently found in the other species.

Figure 2 - figure supplement 2. CS-Public CDR3 sequences are central in mouse TCRβ networks. Shown is a representative network of the 1,000 most frequent sequences from a mouse. Nodes are labeled according to 4 categories: CDR3 sequences that are not public; CDR3 sequences shared by all 11 human samples; CDR3 sequences shared by at least 25 mice; CDR3 sequences shared by at least 25 mice and all 11 humans.

Figure 2 - figure supplement 3. Degree of CS-public sequences is correlated in mouse and human TCR networks. Each dot represents one CS-public sequence that is found among the most abundant 1,000 sequences in at least one mouse and at least one human subject (n=45 sequences). There is a significant correlation between the

degree of CS-public sequences in the two species (R=0.65, spearman); Sequences that are more connected in one species are typically more connected in the other as well.

Figure 2 - figure supplement 4. Sharing properties of the 86 observed CS-public CDR3 sequences in the simulated data. We generated 100 datasets of simulated human and mouse repertoires, with number of individuals (11 humans, 28 mice) and repertoire sizes as in the experimental data. For each of the 86 observed CS-public sequences, we plot its mean sharing level in the simulations, for human repertoires (red – 11 humans) and mouse repertoires (blue – 28 mice). The top panel shows 54 sequences that are CS-public in both experiment and simulations. The lower panel shows 32 sequences that are CS-public in the experimental data but not in the simulations. Note that there were additionally about 200 CS-public sequences in the simulations which were not CS-public in the data.

Figure 3 - figure supplement 1. Public CDR3 sequences form highly connected similarity networks in mice and are enriched for self-associated immune reactivities. Sequance visualization of the red cluster in the mouse CDR3 sequences network shown in Figure 3A. The original full network is formed by the 1,000 most shared mouse CDR3 sequences (found in >25 of 28 mice). 124 CDR3 sequences that were previously annotated (see (Madi et al. 2014)) were added to the network and are presented as triangles. 13 annotated sequences were either identical to, or at a Levenshtein distance of 1 from one of the nodes in this cluster, and their associated pathology / peptide antigen is listed next to the corresponding node.

Figure 4 - figure supplement 1. DN thymocytes manifest lower convergent recombination. Comparison of the number of nt sequences encoding, on average, an AA CDR3 sequence, for public CDR3 AA sequences, found to be shared by more than 25 mice. Public CDR3 sequences coming from DN thymocytes were encoded on average by a lower number of nucleotide (nt) sequences compared to those from CD4+ splenic T cells (p<2.2e-16 for each of these top sharing levels).

Figure 5 - figure supplement 1. Immunization and in vitro antigen stimulation affect network architecture. (A) The number of edges in networks formed by the 1,000 most abundant CDR3 sequences in three TCR datasets: 12 naïve mice; 5 mice immunized with peptide p277 (HSP60 437-460 VLGGGCALLRCIPALDSLTPANED) emulsified in Complete Freund’s Adjuvant (CFA); and 5 mice immunized with p277 whose splenic T cells were stimulated in-vitro with peptide p277. (B) The number of edges in networks

formed by the 1,000 most abundant CDR3 sequences in four TCR datasets: 12 naïve mice; 5 mice immunized with OVA 323-339 ISQAVHAAHAEINEAGR AA sequence peptide in CFA; 3 mice immunized with OVA/CFA whose splenic T cells were stimulated in-vitro with the same OVA peptide; and 5 mice immunized with OVA/CFA whose splenic T cells were analyzed 2 months post-immunization.

Figure 5 - figure supplement 2. Mouse TCR Networks become less connected with aging. A comparison of network clusters in young and aged mice. Network representations of the 1,000 most frequent clones in (A) young and (B) aged mice. The networks composed of the 1,000 most frequent clones in the young mice (n=3) manifested 590.3±61.9 clustered nodes with 992.7±147.4 edges. In contrast, networks composed of the 1,000 most frequent clones in the aged mice (n=3) had 334.7±63.5 clustered nodes with 362.3±153.8 edges. Nodes are colored according to the sharing level of their corresponding CDR3 sequence in the 28 mice dataset.

Figure 5 - figure supplement 3. Human TCR Networks become less connected with aging. A comparison of network connectivity formed by the thousand most frequent CDR3 AA segments expressed in 39 humans at different ages published by (Britanova et al. 2014). The Mean degree was calculated for each human sample and colored according to 4 age groups: 6-25, 34-43, 61-66, and 71-90 years.

Figure 5 - figure supplement 4. With aging, the repertoire becomes more skewed and spread in sequence space due to preferential expansion of private clones at the expanse of more public clones. Frequency of the top 1,000 most frequent CDR3 sequences by sharing level for young (6-8 weeks, n=3) and aged (17-20 months, n=3) mice.

Figure 5 - figure supplement 5. CTLA4 blockade results in a repertoire that is more skewed and spread in sequence space, due to preferential expansion of private clones at the expanse of more public clones. The cumulative frequency (in %) of *relatively private* CDR3 sequences from the top 1000 most frequent sequences in the repertoires of patients pre and post CTLA4 blockade treatment with tremelimumab (Robert et al. 2014). Sharing was defined by comparison with a reference dataset of CDR3 sequences from 11 young healthy individuals (Britanova et al. 2014): *Relatively private* sequences were defined as CDR3 sequences shared by 0-5 individuals out of 11 in the reference dataset, where 0 indicates a sequence not found in any of the 11 individuals. There is a significant increase in the frequency of relatively private sequences (p-value = 0.01947, ranked Wilcox paired test).

Supplementary Files:

Supplementary File 1. Statistics of TCR networks for mouse and human repertoires. Mouse data: 12 mice from (Madi et al. 2014). Human data: 11 young subjects from (Britanova et al. 2014).

Supplementary File 2. Summary of the data for the quad-KO mice, which are lacking four elements needed for physiological MHC-dependent antigen selection: MHC-I and -II molecules together with CD4 and CD8 co-receptor molecules (Van Laethem et al. 2007; Van Laethem et al. 2013), and matched control WT mice. Connected.nodes and edges refers to network statistics generated from the 1,000 most frequent CDR3 sequences in each mouse.

Supplementary File 3. Summary of TCR-seq data used in this study, from 5 experimental conditions: (1) mice that were immunized with either Chicken Ovalbumin (OVA) or (2) peptide 277 (p277), of HSP60. Spleens were harvested on day 7 post immunization and T cells were extracted for TCR analysis. (3) *in vitro* stimulation: T cells from spleens of immunized mice were harvested on day 7 and were re-stimulated with irradiated splenocytes and the relevant peptide antigen. (4) Five of the OVA-immunized mice received a boost IP injection of 100µg OVA+CFA on day 14, and spleens were harvested on day 60 for TCR analysis. (5) DN thymocytes.

Figure 1

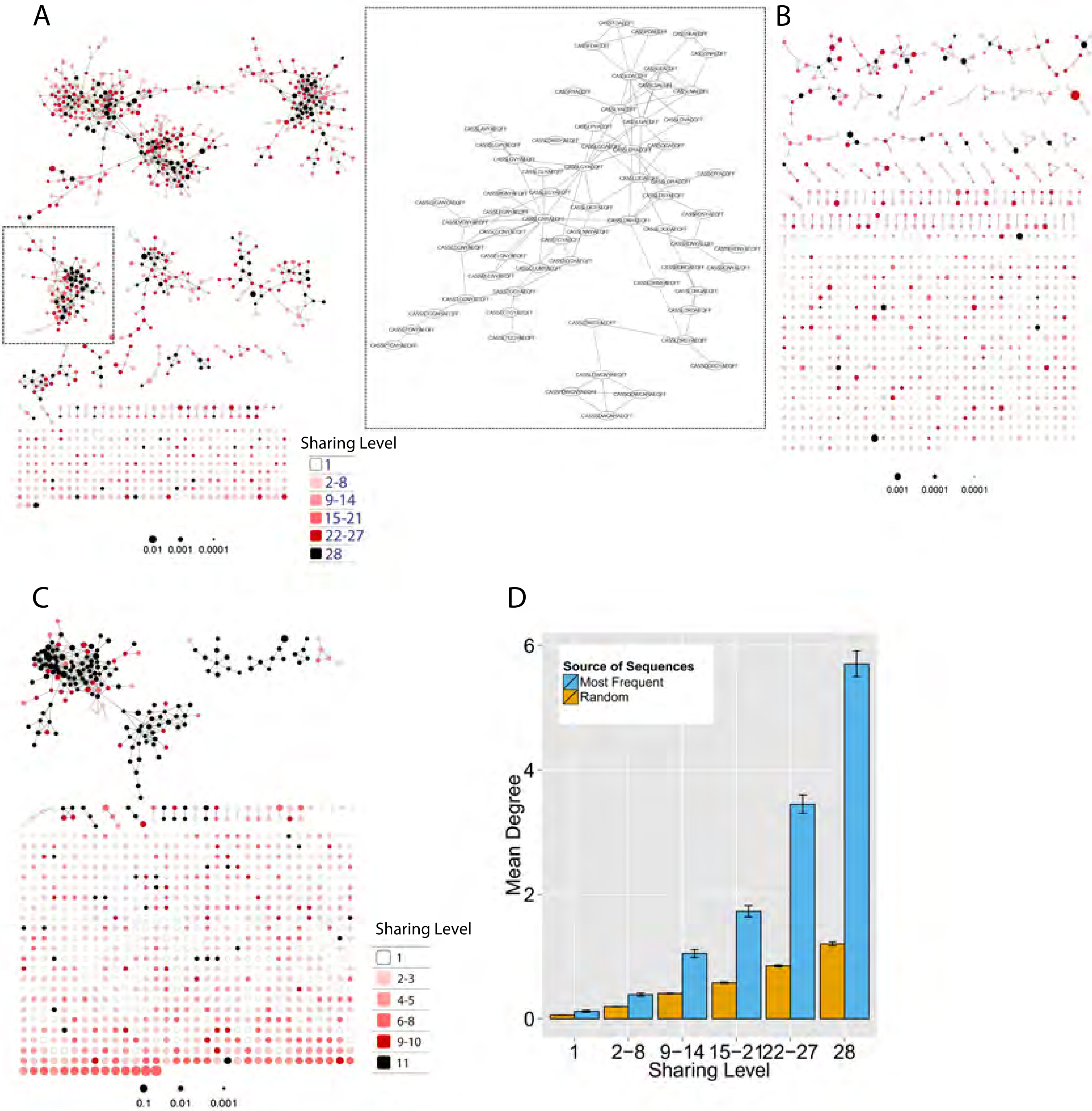
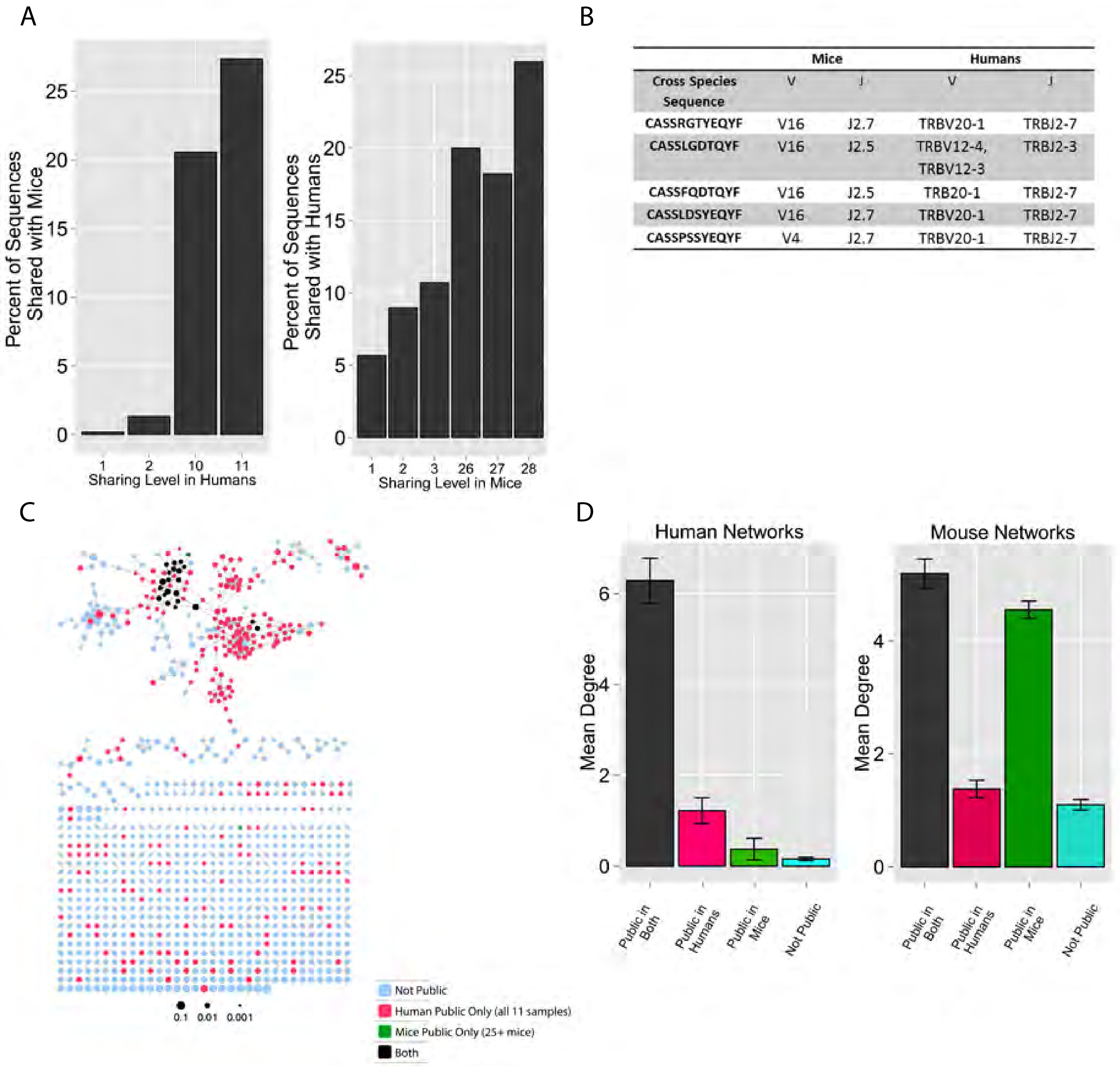

A
B
Sharing Level
□ 1
2-8
9-14
15-21
22-27
28
0.01
0.001
0.0001
0.001
0.0001
0.0001
C
Sharing Level
1
2-3
4-5
6-8
9-10
11
0.1
0.01
0.001
D
Source of Sequences
Most Frequent
Random
Mean Degree
0
2
4
6
1
2−8
9−14
15−21
22−27
28
Sharing Level

Figure 2



	Mice		Humans	
Cross Species Sequence	V	J	V	J
CASSRGTYEQYF	V16	J2.7	TRBV20-1	TRBJ2-7
CASSLGDTQYF	V16	J2.5	TRBV12-4, TRBV12-3	TRBJ2-3
CASSFQDTQYF	V16	J2.5	TRB20-1	TRBJ2-7
CASSLDSYEQYF	V16	J2.7	TRBV20-1	TRBJ2-7
CASSPSSYEQYF	V4	J2.7	TRBV20-1	TRBJ2-7

Figure 3

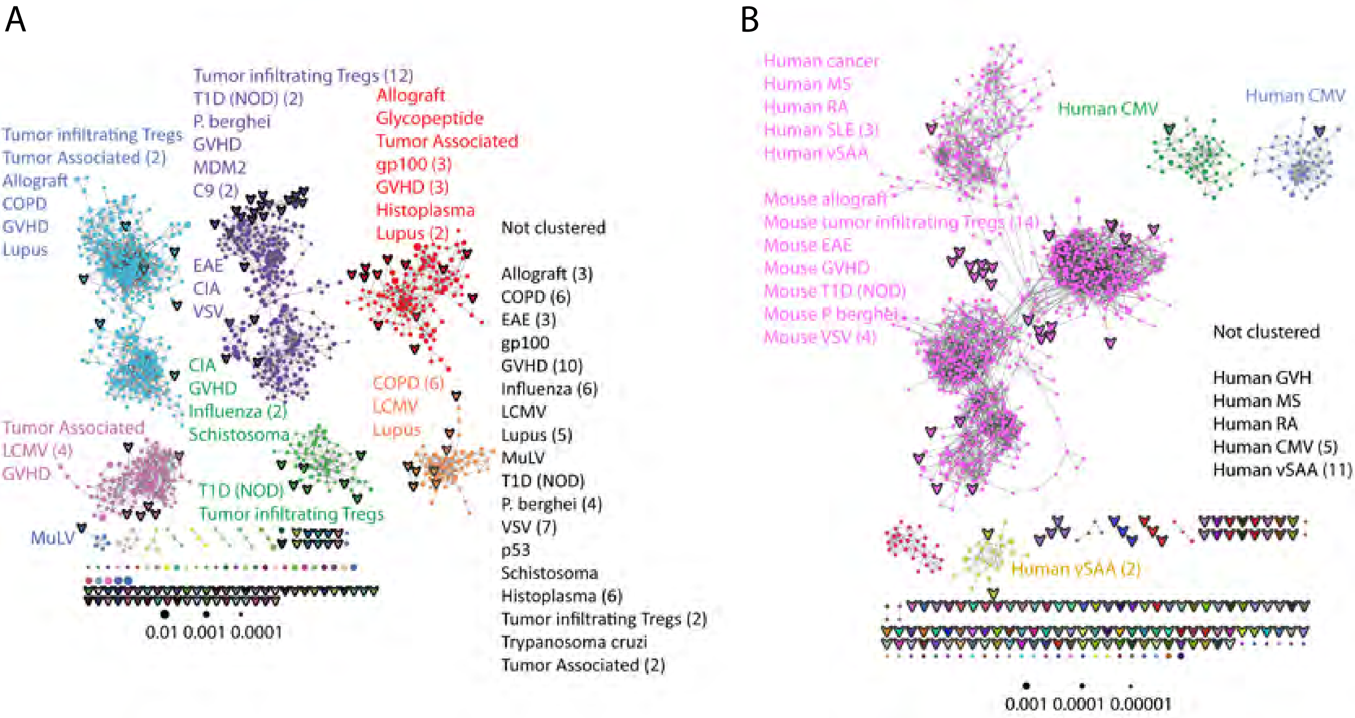

A
Tumor infiltrating Tregs
Tumor Associated (2)
Allograft
COPD
GVHD
Lupus
Tumor infiltrating Tregs (12)
T1D (NOD) (2)
P. berghei
GVHD
MDM2
C9 (2)
EAE
CIA
VSV
Allograft
Glycopeptide
Tumor Associated
gp100 (3)
GVHD (3)
Histoplasma
Lupus (2)
CIA
GVHD
Influenza (2)
Schistosoma
COPD (6)
LCMV
Lupus
Tumor Associated
LCMV (4)
GVHD
T1D (NOD)
Tumor infiltrating Tregs
MuLV
0.01 0.001 0.0001
Not clustered
Allograft (3)
COPD (6)
EAE (3)
gp100
GVHD (10)
Influenza (6)
LCMV
Lupus (5)
MuLV
T1D (NOD)
P. berghei (4)
VSV (7)
p53
Schistosoma
Histoplasma (6)
Tumor infiltrating Tregs (2)
Trypanosoma cruzi
Tumor Associated (2)
B
Human cancer
Human MS
Human RA
Human SLE (3)
Human vSAA
Mouse allograft
Mouse tumor infiltrating Tregs (14)
Mouse EAE
Mouse GVHD
Mouse T1D (NOD)
Mouse P berghei
Mouse VSV (4)
Human CMV
Human CMV
Not clustered
Human GVH
Human MS
Human RA
Human CMV (5)
Human vSAA (11)
Human vSAA (2)
0.001 0.0001 0.00001

Figure 4

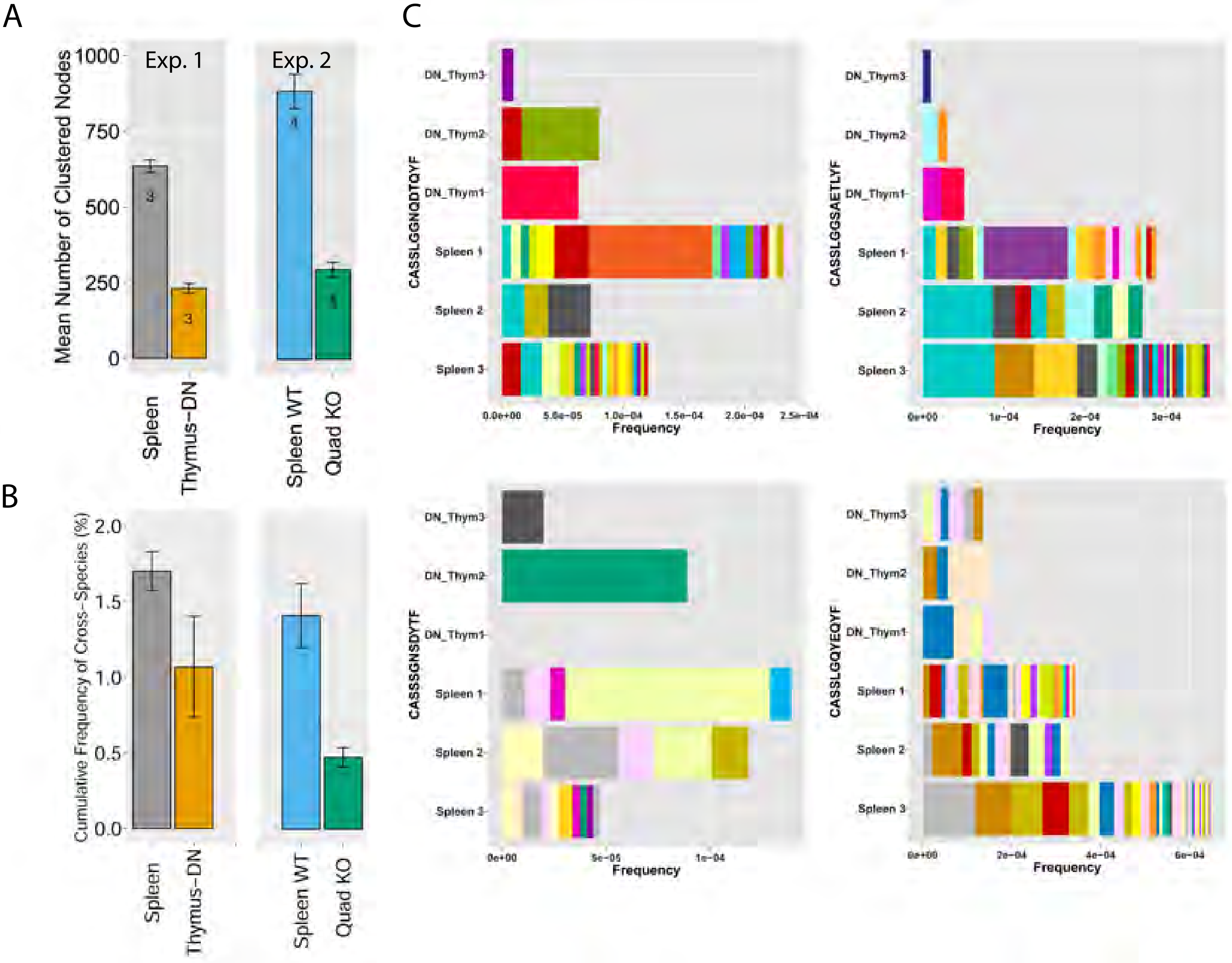

A
Exp. 1
Exp. 2
Mean Number of Clustered Nodes
1000
750
500
250
0
3
3
4
1
Spleen
Thymus–DN
Spleen WT
Quad KO
B
Cumulative Frequency of Cross–Species (%)
2.0
1.5
1.0
0.5
0.0
Spleen
Thymus–DN
Spleen WT
Quad KO
C
CASSLGGNQDTQYF
DN_Thym3
DN_Thym2
DN_Thym1
Spleen 1
Spleen 2
Spleen 3
0.0e+00
5.0e–05
1.0e–04
1.5e–04
2.0e–04
2.5e–04
Frequency
CASSLGGSAETLYF
DN_Thym3
DN_Thym2
DN_Thym1
Spleen 1
Spleen 2
Spleen 3
0e+00
1e–04
2e–04
3e–04
Frequency
CASSSGNSDYTF
DN_Thym3
DN_Thym2
DN_Thym1
Spleen 1
Spleen 2
Spleen 3
0e+00
5e–05
1e–04
Frequency
CASSLGQYEQYF
DN_Thym3
DN_Thym2
DN_Thym1
Spleen 1
Spleen 2
Spleen 3
0e+00
2e–04
4e–04
6e–04
Frequency

Figure 5

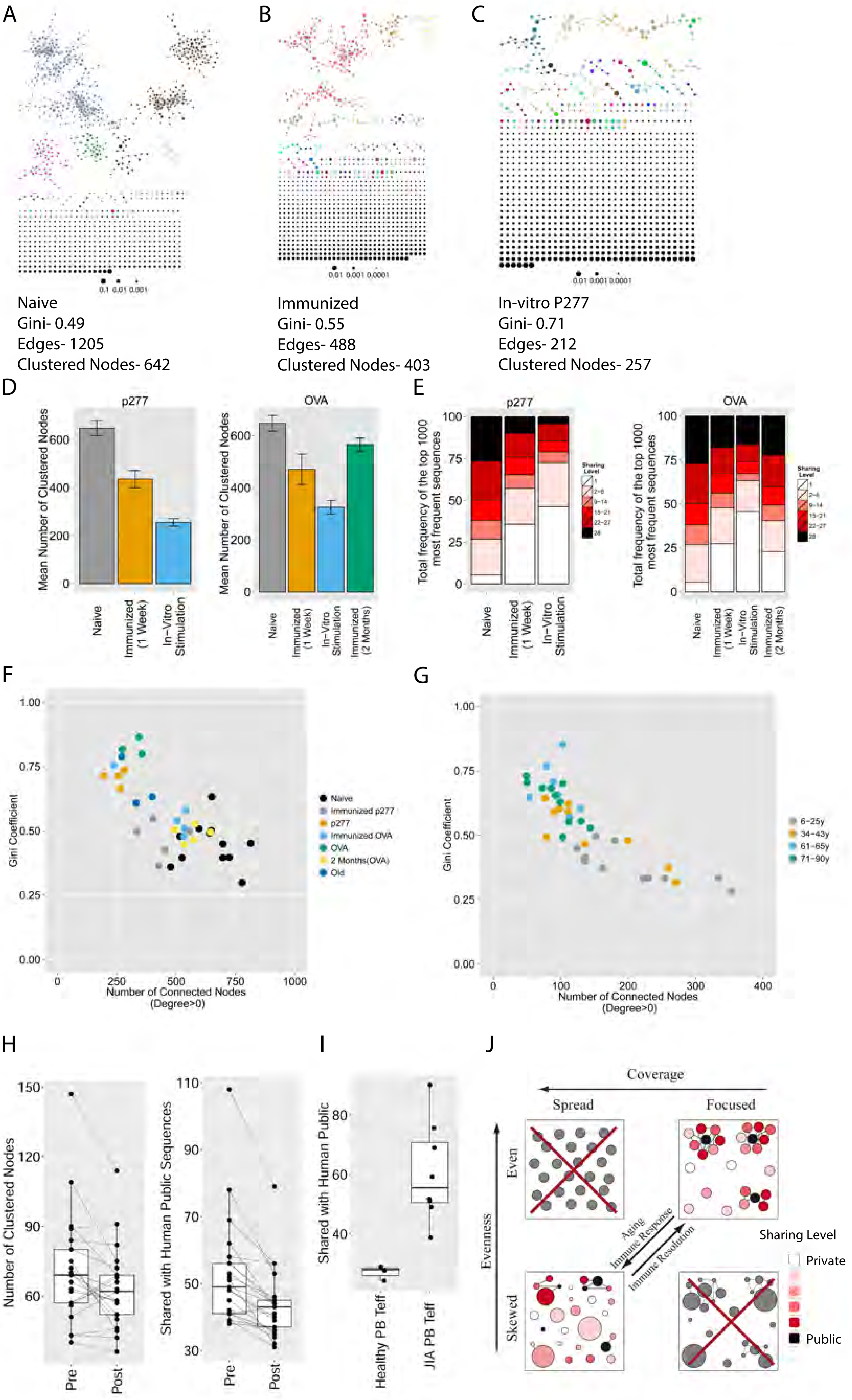

A
0.1
0.01
0.001
Naive
Gini- 0.49
Edges- 1205
Clustered Nodes- 642
B
0.01
0.001
0.0001
Immunized
Gini- 0.55
Edges- 488
Clustered Nodes- 403
C
0.01
0.001
0.0001
In-vitro P277
Gini- 0.71
Edges- 212
Clustered Nodes- 257
D
p277
Mean Number of Clustered Nodes
0
200
400
600
Naive
Immunized (1 Week)
In-Vitro Stimulation
OVA
Mean Number of Clustered Nodes
0
200
400
600
Naive
Immunized (1 Week)
In-Vitro Stimulation
Immunized (2 Months)
E
p277
Total frequency of the top 1000 most frequent sequences
0
25
50
75
100
Sharing Level
1
2-8
9-14
15-21
22-27
28
Naive
Immunized (1 Week)
In-Vitro Stimulation
OVA
Total frequency of the top 1000 most frequent sequences
0
25
50
75
100
Sharing Level
1
2-8
9-14
15-21
22-27
28
Naive
Immunized (1 Week)
In-Vitro Stimulation
Immunized (2 Months)
F
Gini Coefficient
0.00
0.25
0.50
0.75
1.00
0
250
500
750
1000
Number of Connected Nodes (Degree>0)
Naive
Immunized p277
p277
Immunized OVA
OVA
2 Months(OVA)
Old
G
Gini Coefficient
0.00
0.25
0.50
0.75
1.00
0
100
200
300
400
Number of Connected Nodes (Degree>0)
6-25y
34-43y
61-65y
71-90y
H
Number of Clustered Nodes
60
90
120
150
Pre
Post
Shared with Human Public Sequences
30
50
70
90
110
Pre
Post
I
Shared with Human Public
40
60
80
Healthy PB Teff
JIA PB Teff
J
Coverage
Spread
Focused
Evenness
Even
Skewed
Aging
Immune Response
Immune Resolution
Sharing Level
Private
Public

Figure 1- figure supplement 1

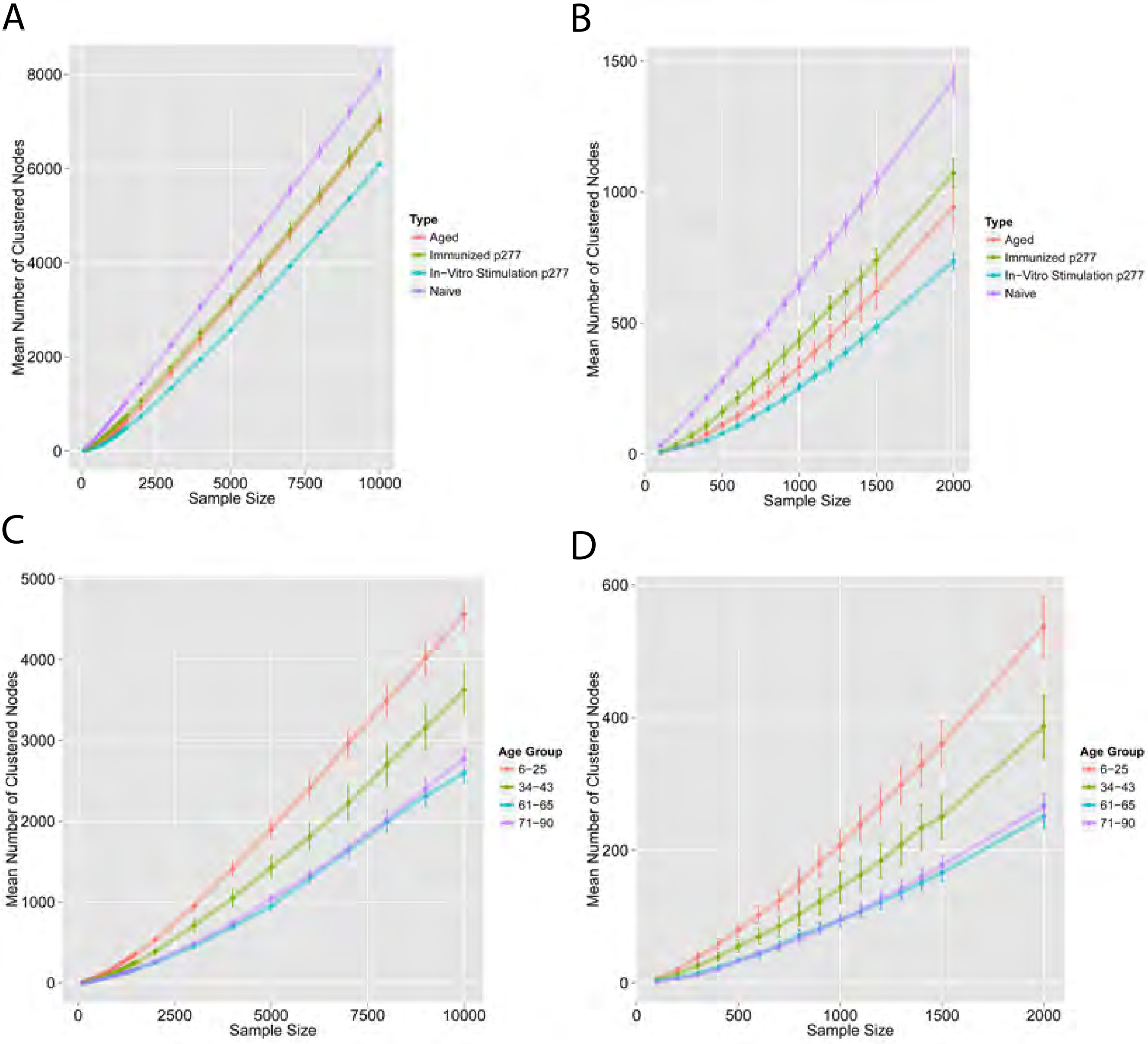


Figure 1- figure supplement 2

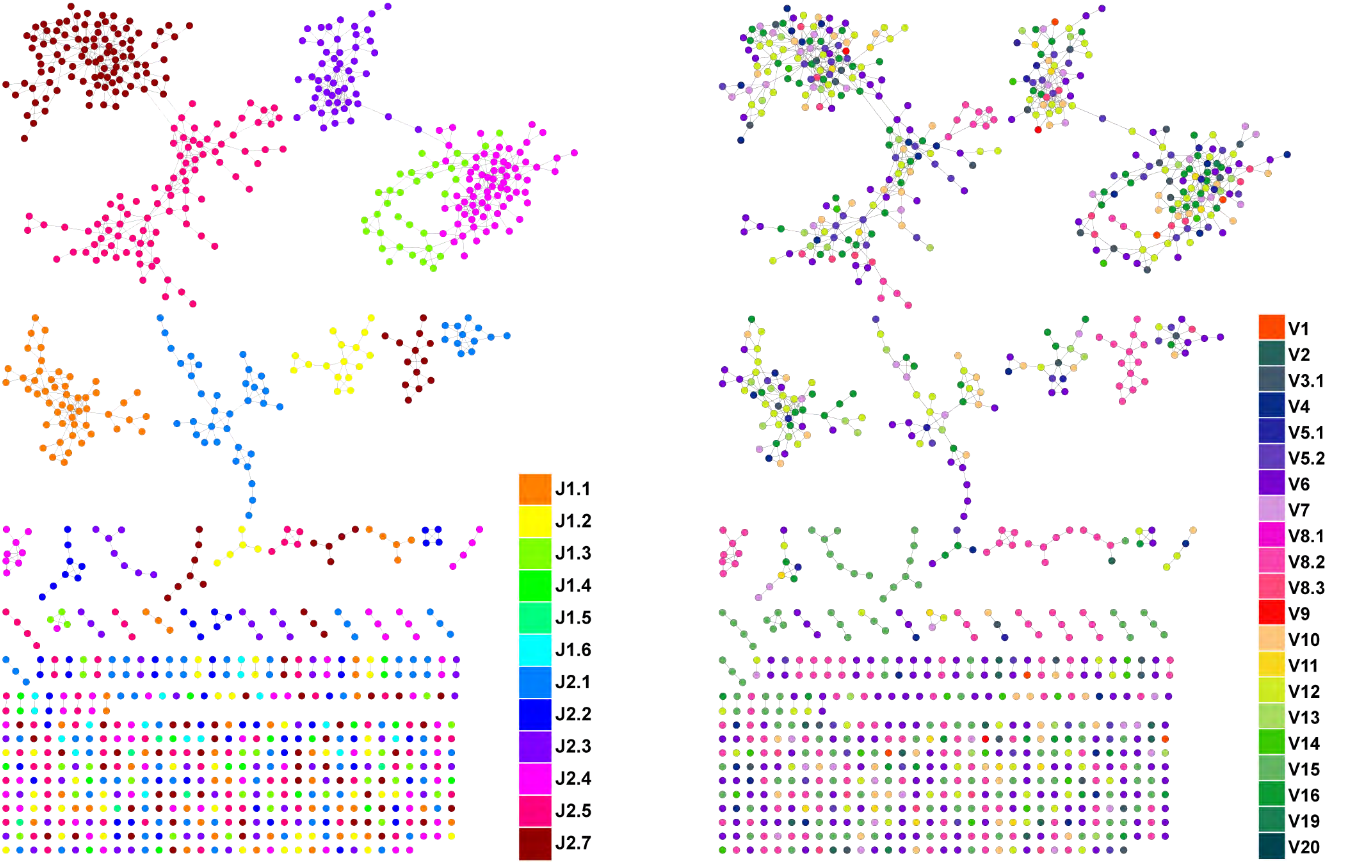


Figure 1- figure supplement 3

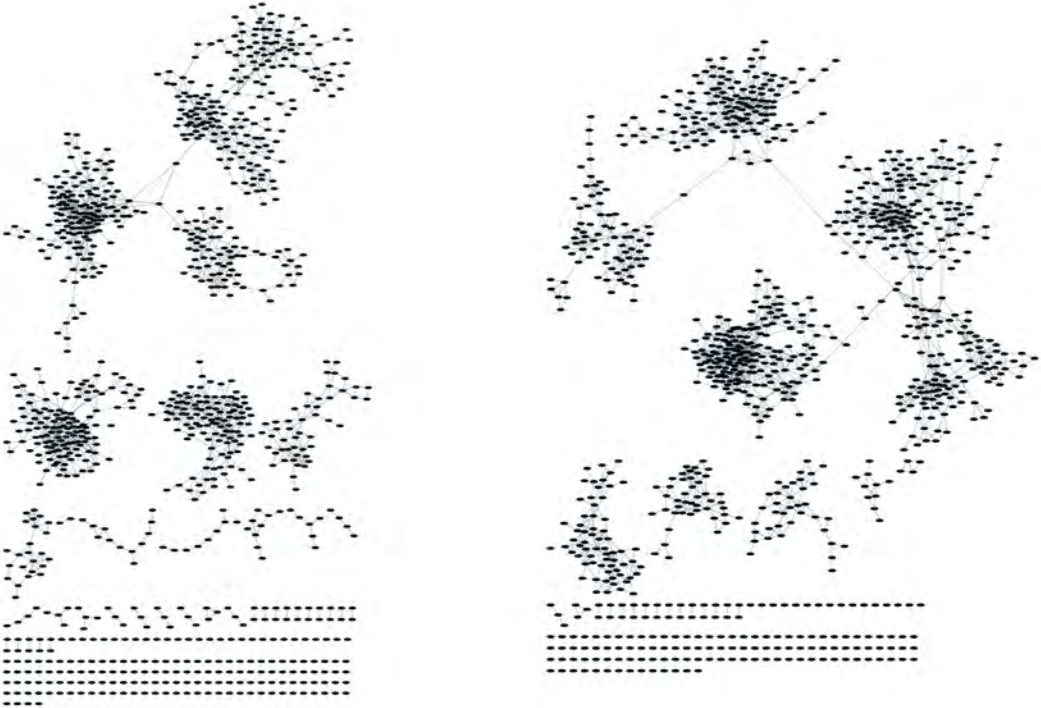


Figure 1- figure supplement 4

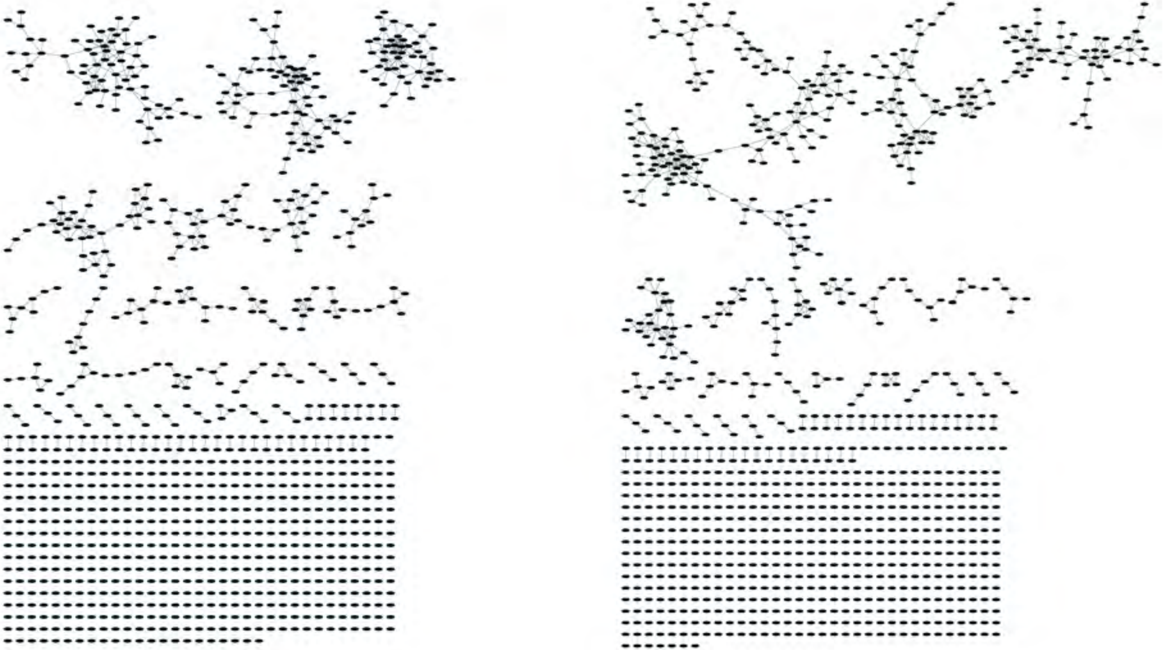


Figure 1- figure supplement 5

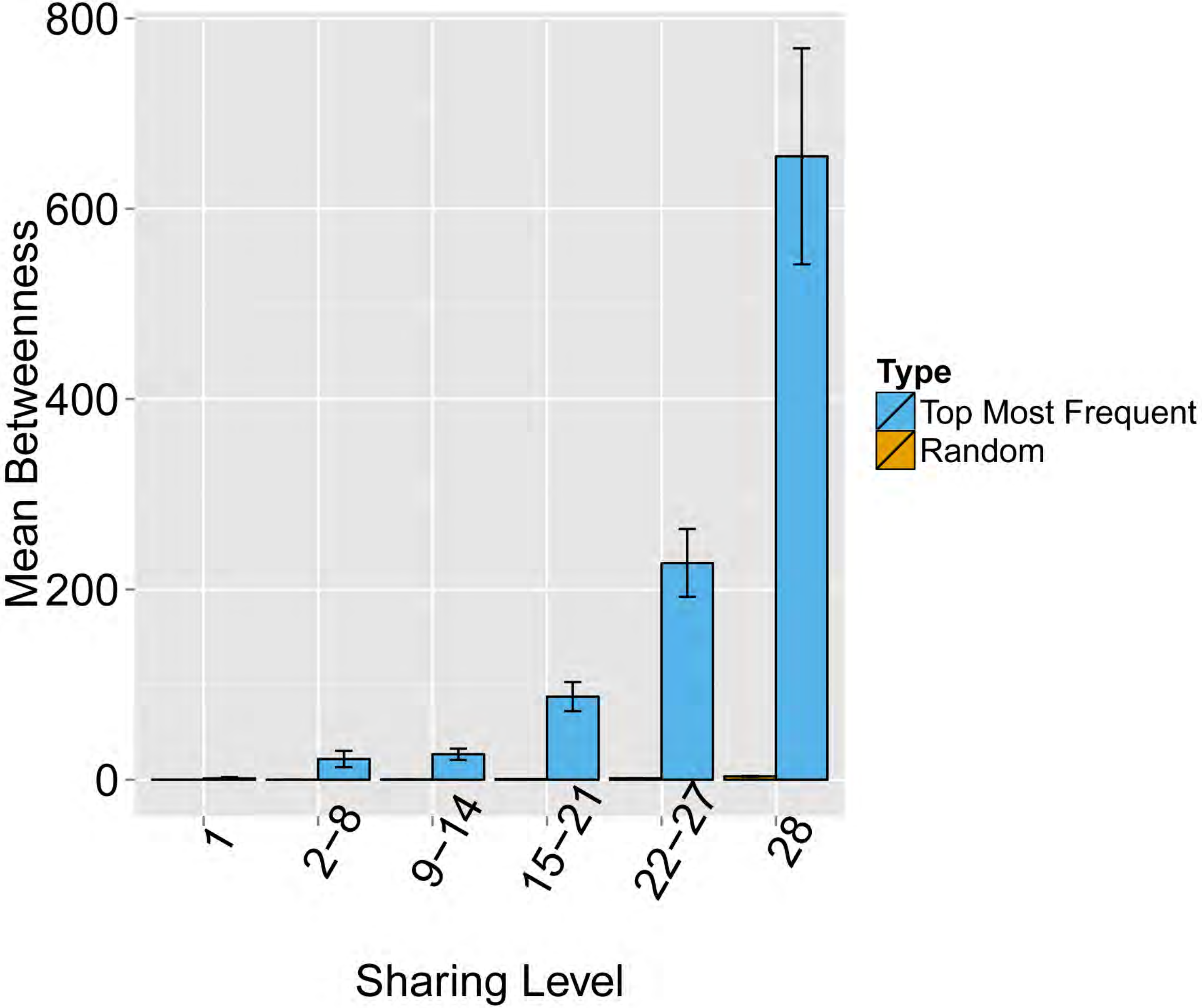


Figure 1- figure supplement 6

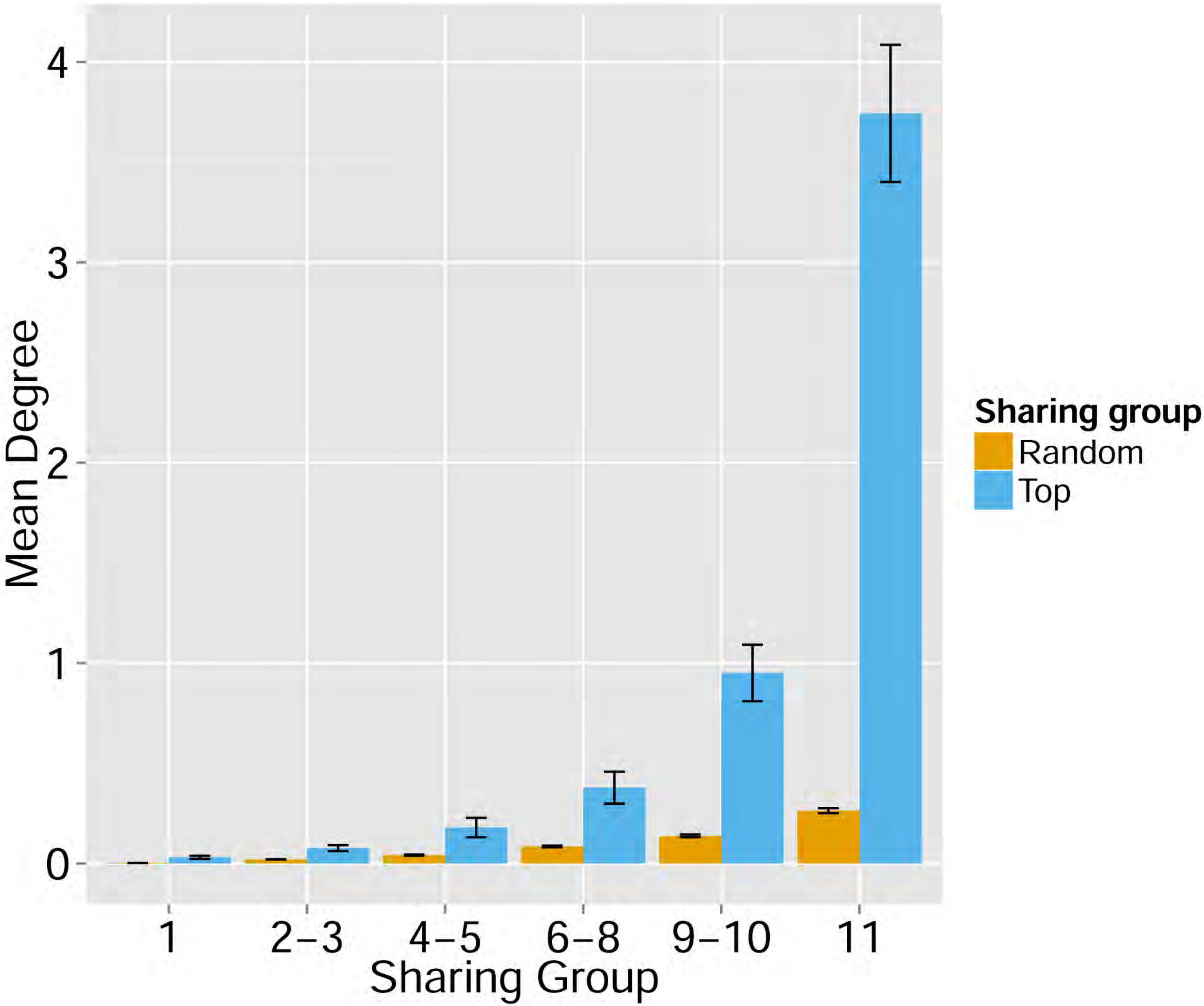


Figure 2 - figure supplement 1

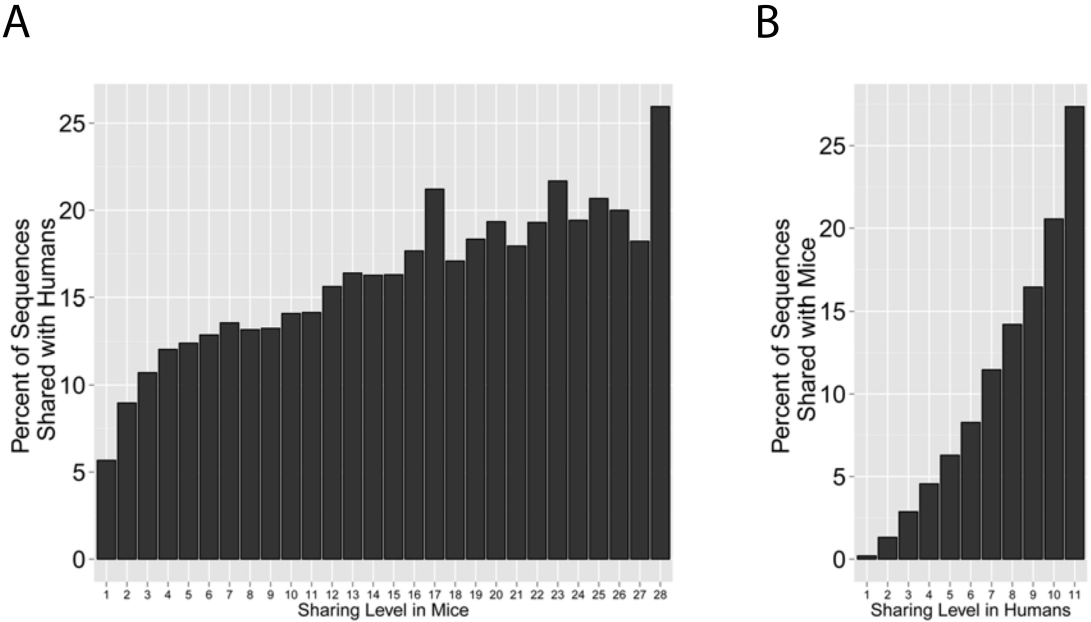


Figure 2 - figure supplement 2

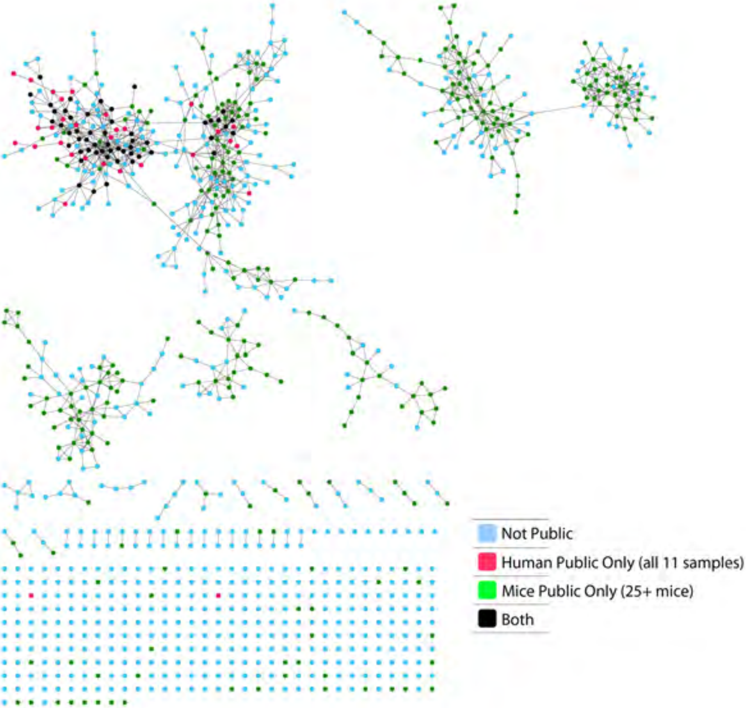

Not Public
Human Public Only (all 11 samples)
Mice Public Only (25+ mice)
Both

Figure 2 - figure supplement 3

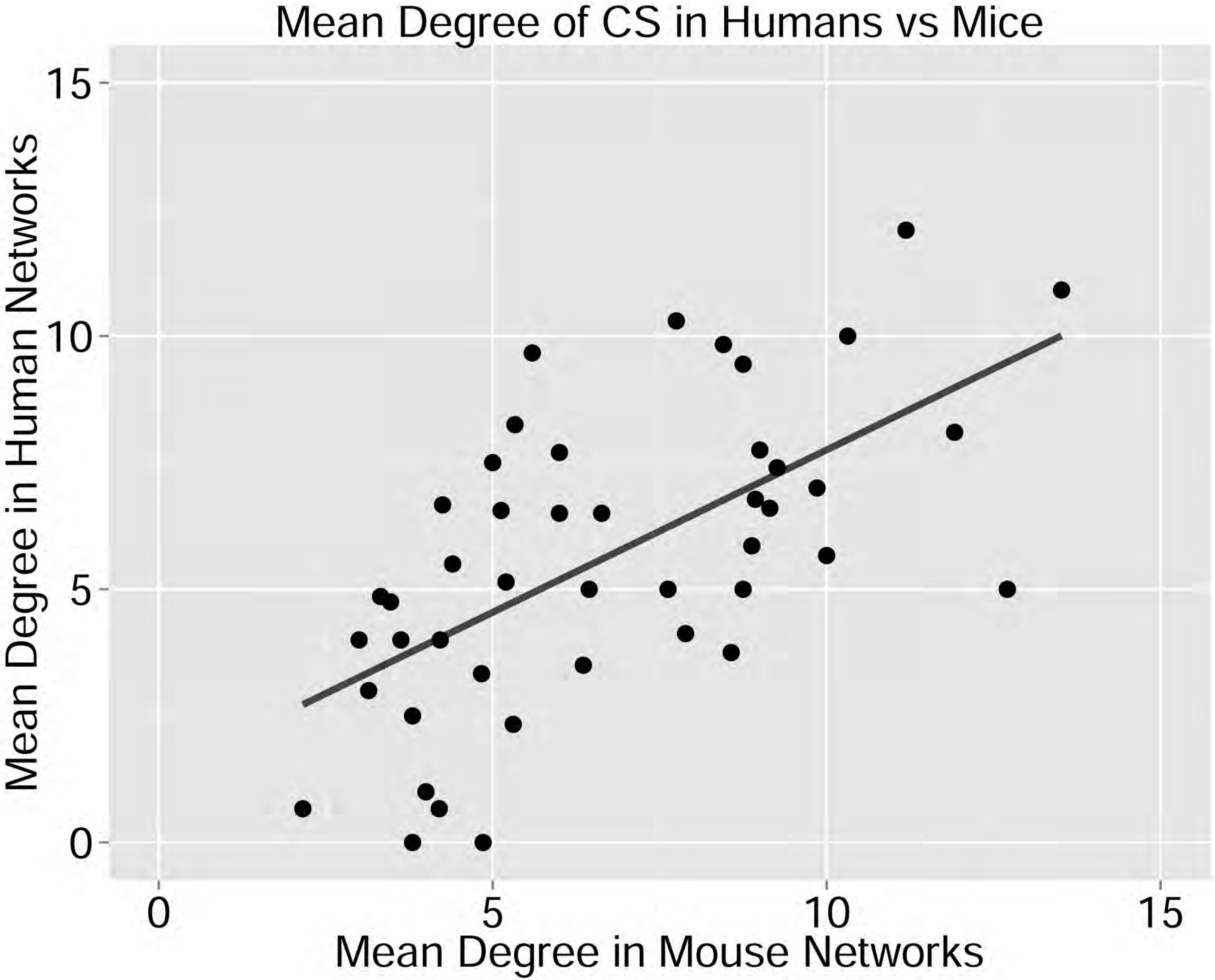


Figure 2 - figure supplement 4

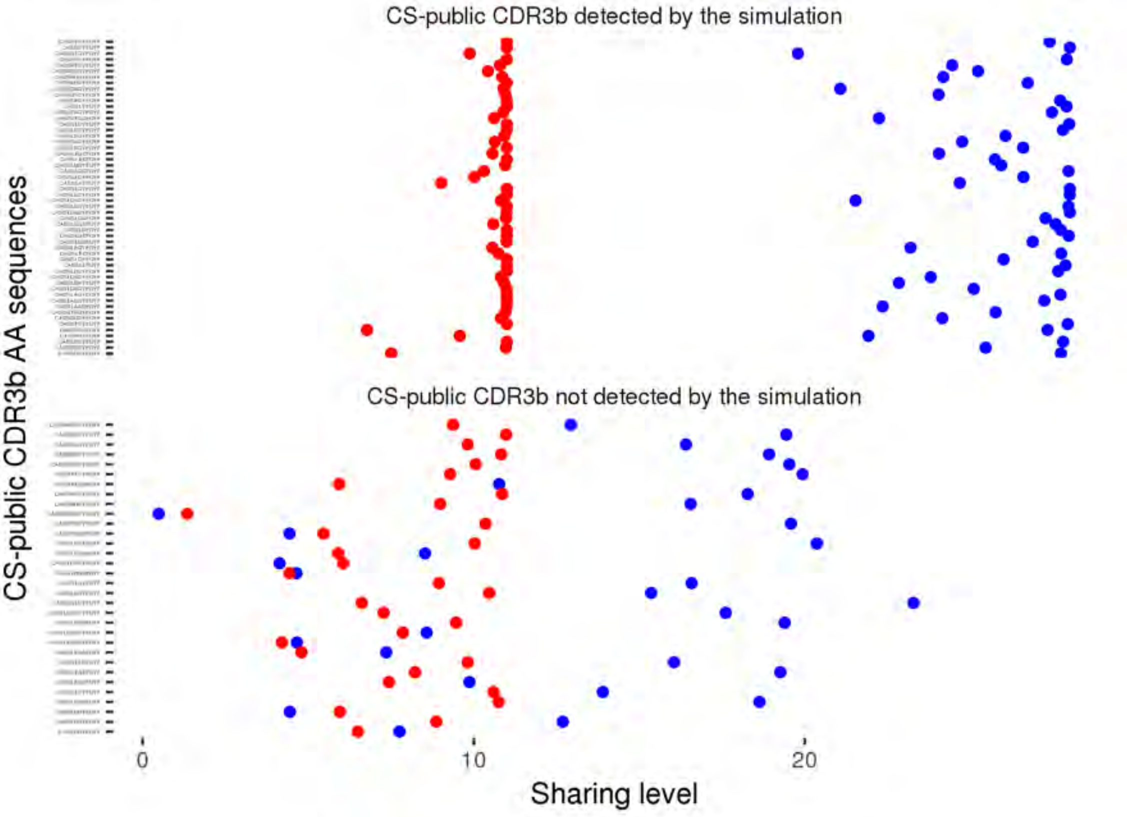


Figure 3 - figure supplement 1

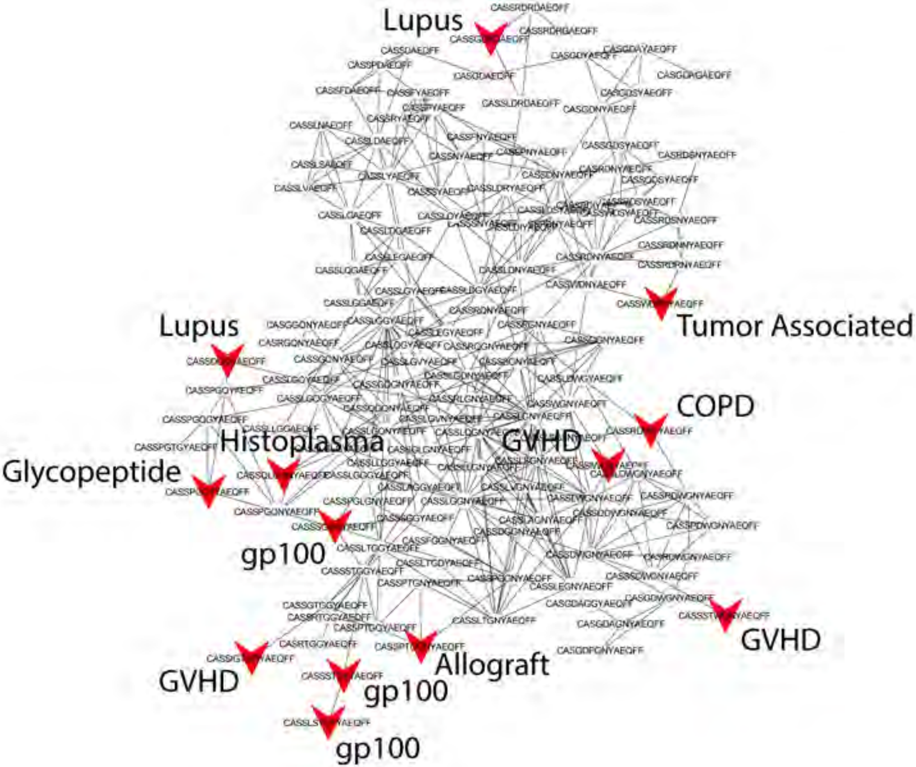


Figure 4 - figure supplement 1

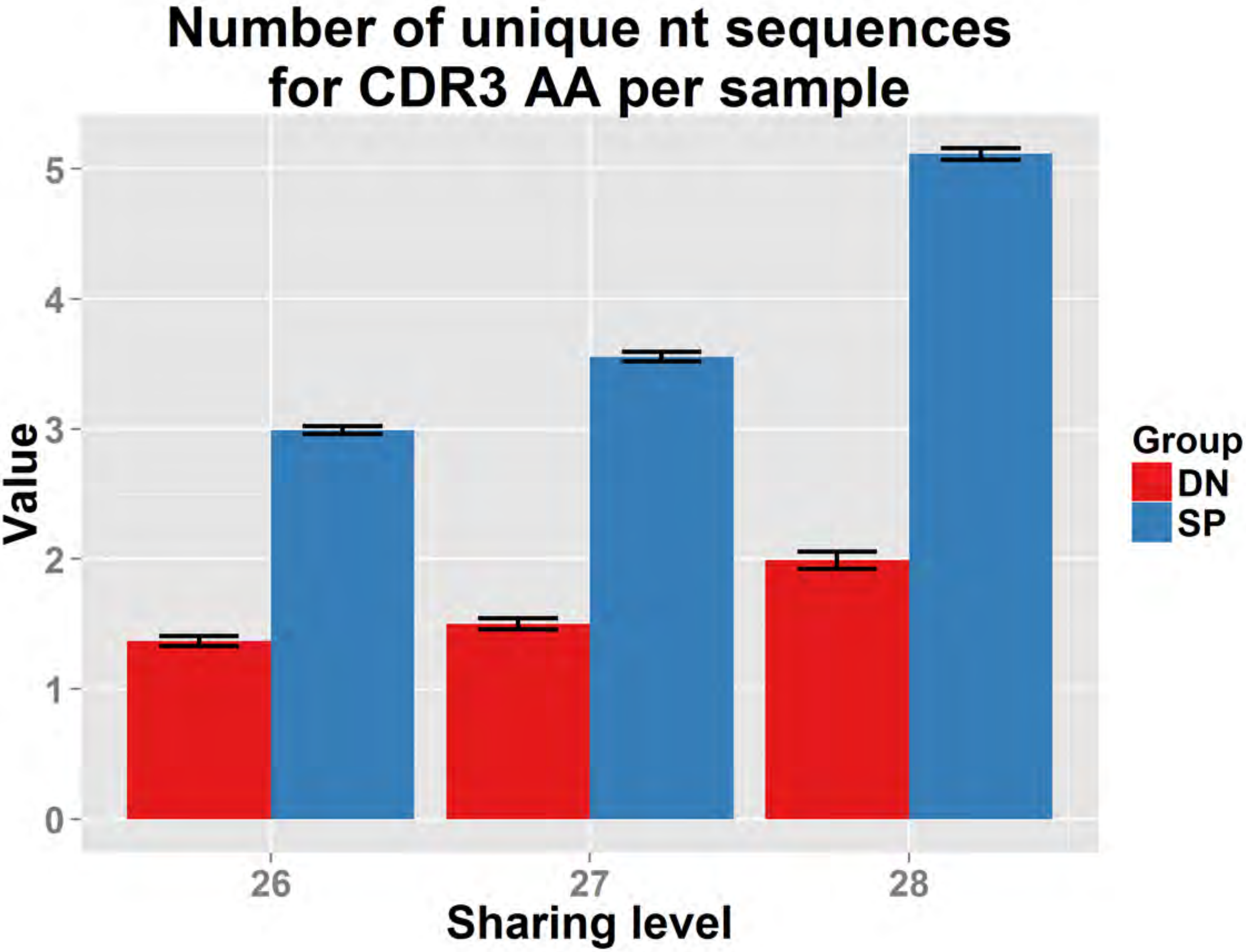


Figure 5 - figure supplement 1

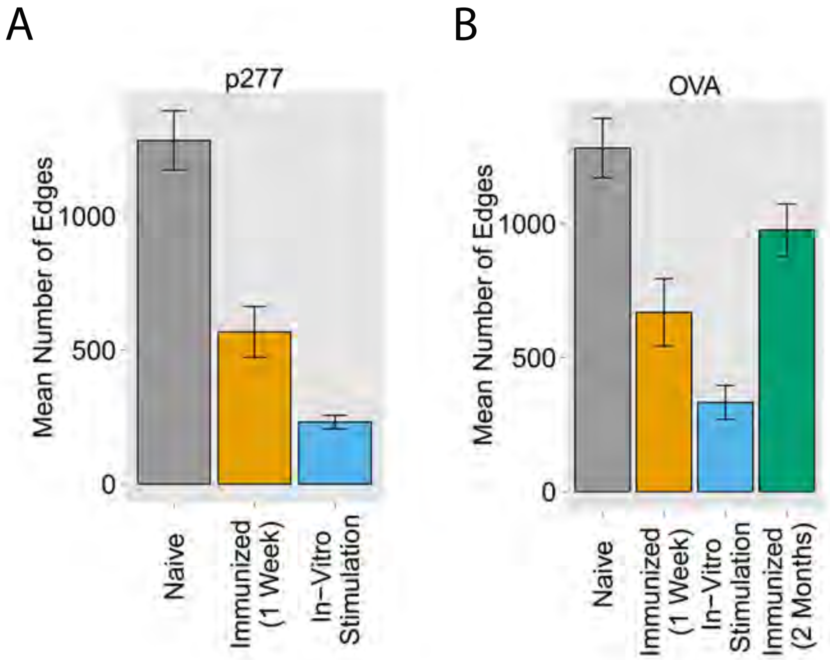


Figure 5 - figure supplement 2

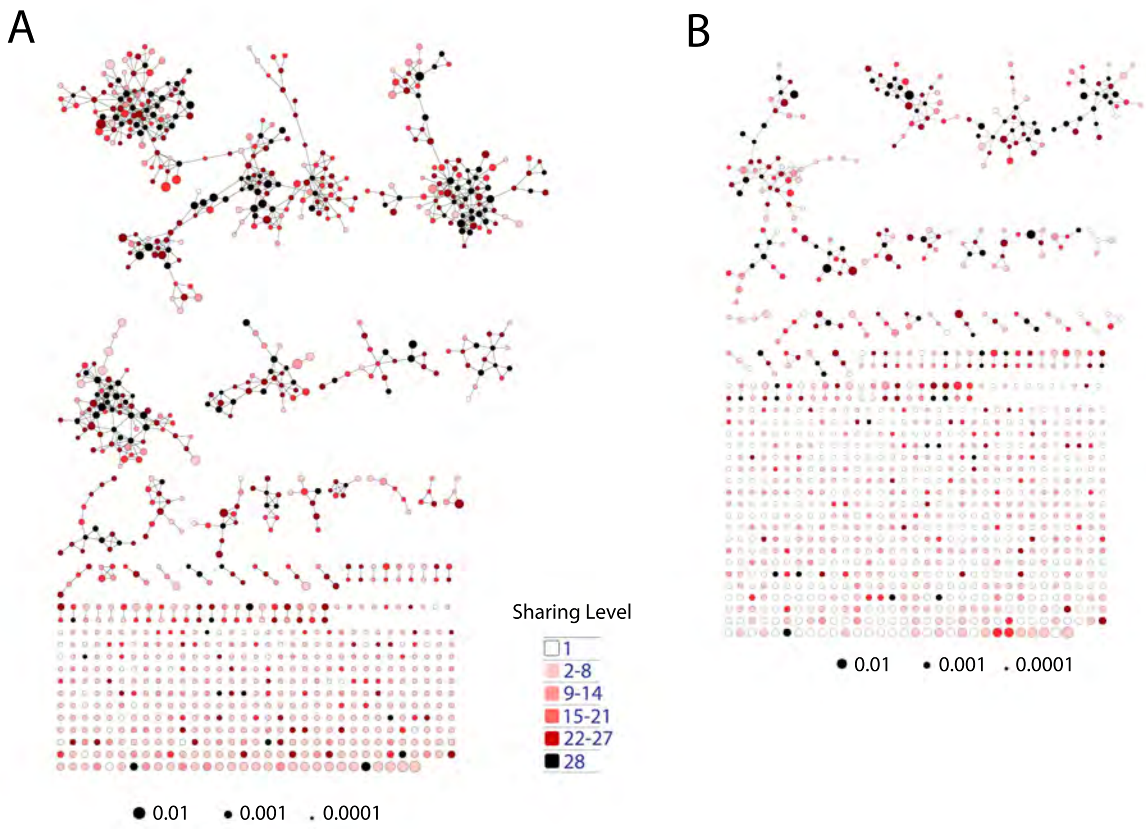


Figure 5 - figure supplement 3

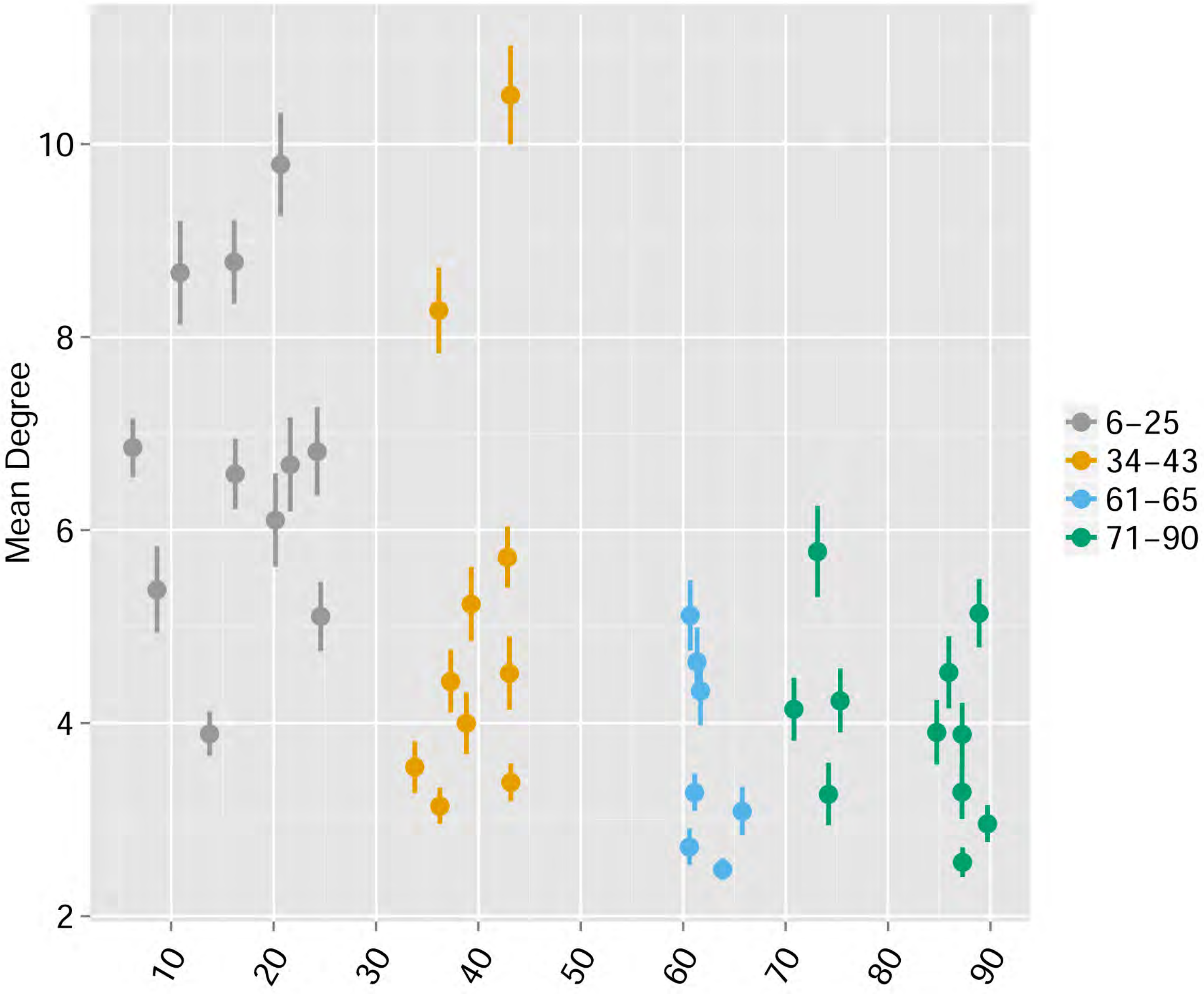


Figure 5 - figure supplement 4

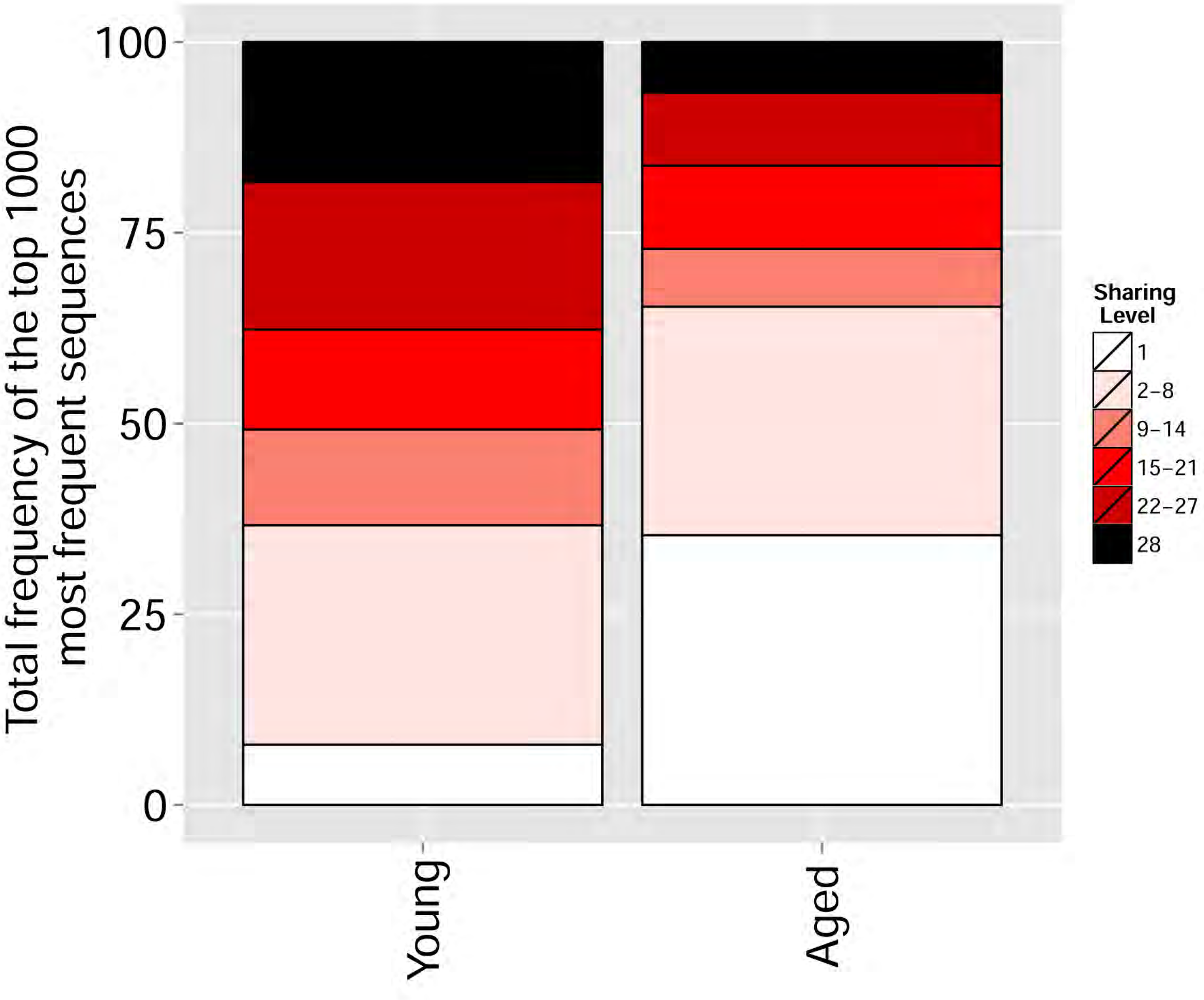
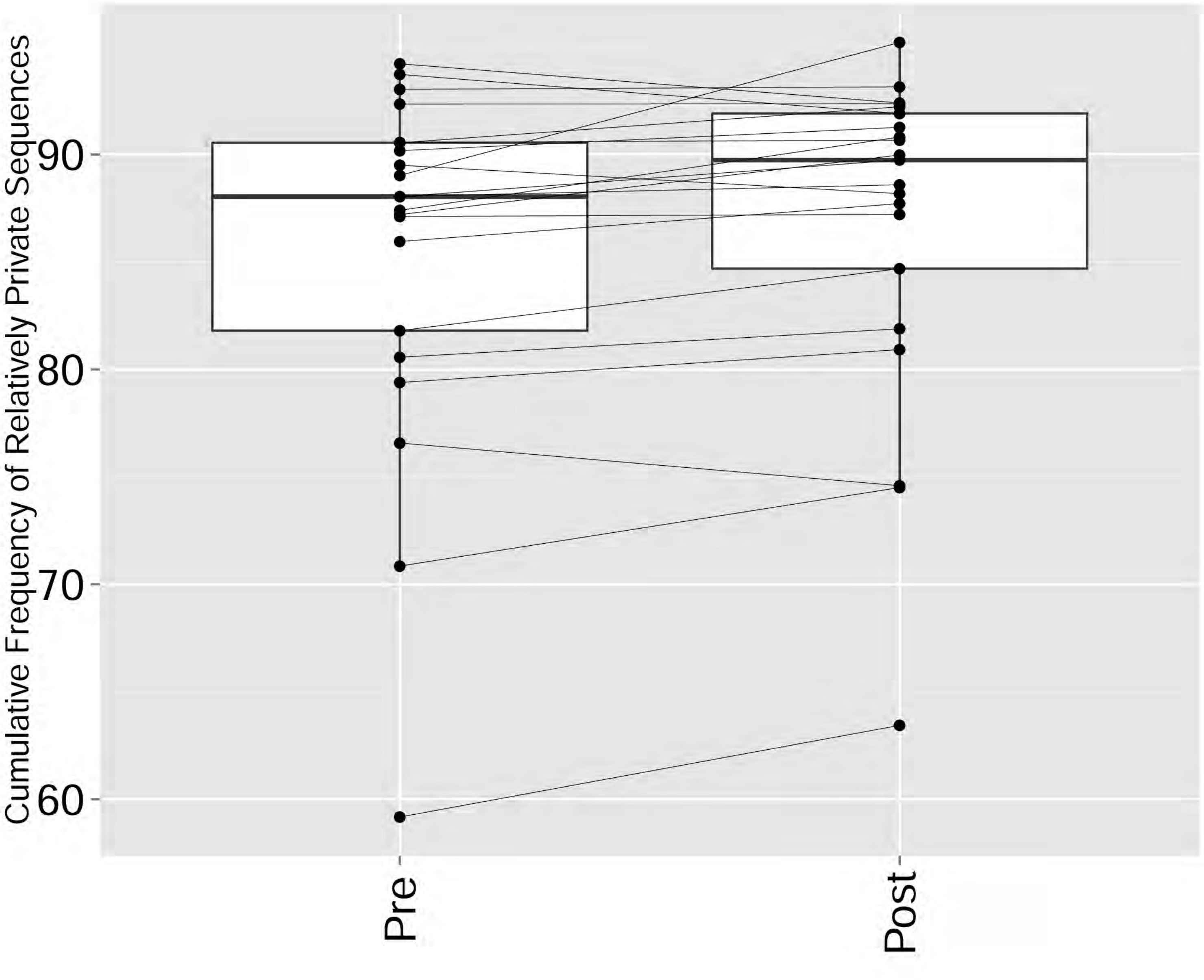


Figure 5 - figure supplement 5


Cumulative Frequency of Relatively Private Sequences
Pre
Post