Insights into immune system development and function from mouse T-cell repertoires

Zachary Sethna[a,1], Yuval Elhanati[a,b,1], Chrissy S. Dudgeon[c], Curtis G. Callan Jr.[a,2], Arnold J. Levine[d], Thierry Mora[e], and Aleksandra M. Walczak[b]

[a]Joseph Henry Laboratories, Princeton University, Princeton, NJ 08544; [b]Laboratoire de Physique Théorique, UMR8549, CNRS, École Normale Supérieure, 75005 Paris, France; [c]Cancer Institute of New Jersey, Rutgers University, New Brunswick, NJ 08903; [d]Institute for Advanced Study, Princeton NJ 08540; and [e]Laboratoire de Physique Statistique, UMR8550, CNRS, École Normale Supérieure, 75005 Paris, France

Contributed by Curtis G. Callan Jr., January 5, 2017 (sent for review November 9, 2016; reviewed by Benjamin Chain and Arup K. Chakraborty)

The ability of the adaptive immune system to respond to arbitrary pathogens stems from the broad diversity of immune cell surface receptors. This diversity originates in a stochastic DNA editing process (VDJ recombination) that acts on the surface receptor gene each time a new immune cell is created from a stem cell. By analyzing T-cell receptor (TCR) sequence repertoires taken from the blood and thymus of mice of different ages, we quantify the changes in the VDJ recombination process that occur from embryo to young adult. We find a rapid increase with age in the number of random insertions and a dramatic increase in diversity. Because the blood accumulates thymic output over time, blood repertoires are mixtures of different statistical recombination processes, and we unravel the mixture statistics to obtain a picture of the time evolution of the early immune system. Sequence repertoire analysis also allows us to detect the statistical impact of selection on the output of the VDJ recombination process. The effects we find are nearly identical between thymus and blood, suggesting that our analysis mainly detects selection for proper folding of the TCR receptor protein. We further find that selection is weaker in laboratory mice than in humans and it does not affect the diversity of the repertoire.

immunology | VDJ recombination | T cells | sequencing | mouse

The adaptive immune system relies on cell surface receptors to recognize an unpredictable array of foreign pathogens. T cells perform their surveillance function through a highly diverse repertoire of T-cell receptors (TCRs): Any individual TCR recognizes only a small subset of the foreign peptides that it may encounter, and the system defends against a broad range of pathogens by having TCRs of many different specificities. This receptor diversity is created by a stochastic DNA editing process (VDJ recombination) that acts on the TCR gene each time a new immune cell is created from a stem cell and acts independently on each of the two chains (alpha and beta) that compose the receptor.

The resulting repertoire diversity can now be studied in great detail, using high-throughput sequencing of lymphocyte receptor repertoires (1–8). In previous work, we used human sequence repertoires to develop methods for inferring the details of the stochastic process of VDJ recombination (9, 38) and for characterizing the statistical effects of thymic selection (10). In this paper, we apply these methods to T-cell β-chain sequence data collected from mice at several stages of development, from embryos to young adults, to study the T-cell repertoire maturation process. We exploit the flexibility of the mouse model to study aspects of immune system development and function that are not readily accessible in humans.

It is known that B- and T-cell receptors formed in embryonic or neonatal individuals are less diverse than in adults: They have fewer nontemplated insertions (11) due to the absence of terminal deoxynucleotidyl transferase (TdT) expression, the enzyme responsible for these insertions (12–14). Whereas this observation has been confirmed by deep sequencing of human TCR repertoires (15, 16), the precise form and time-resolved dynamics of VDJ recombination have not been assessed.

We analyze mouse T cells taken simultaneously from thymus and blood in the same individual to gain insights into the dynamics of repertoire development. First, because the periphery accumulates T cells produced at different times, whereas new T cells pass quickly through the thymus and reflect conditions at a single time point, the statistical structure of the blood T-cell repertoire can be very different from that of the thymus. This difference poses a challenge for statistical inference, and in this paper we develop a method for describing such repertoires derived from a mixture of different conditions. Second, the ability to compare thymus with blood sequence repertoires allows for a more refined view of the stages of receptor selection for functionality, from initial receptor generation to eventual passage into the periphery, and leads to a qualitatively different picture for mice than previously reported for humans.

Results

Inferring the Statistics of VDJ Recombination. A new TCR gene is created from germline DNA by a series of stochastic events: choosing gene segments, deleting bases from the ends of the chosen gene segments, and inserting nucleotides between the modified gene segments. Because the same sequence can be generated

Significance

The immune system defends against pathogens in part via a diverse population of T cells that display different surface receptor proteins [T-cell receptors (TCRs)] designed to recognize MHC-presented foreign peptides. Receptor diversity is produced by an initial random gene recombination process, followed by selection for proteins that fold correctly and bind weakly to self-peptides. Using data from mice of different ages, from embryo to young adult, we quantify the changes with time in the way receptors are generated and selected for function. We find a strong increase in repertoire diversity, occurring shortly after birth, due to a sharp increase in the number of random nucleotide insertions in the primitive TCR gene recombination process. Differences between thymic and blood TCR sequence distributions allow us to infer subtle details of this "turning on" of the mouse immune system.

Author contributions: Z.S., Y.E., C.G.C., A.J.L., T.M., and A.M.W. designed research; Z.S., Y.E., C.S.D., C.G.C., T.M., and A.M.W. performed research; Z.S., Y.E., C.G.C., T.M., and A.M.W. analyzed data; and Z.S., Y.E., C.G.C., A.J.L., T.M., and A.M.W. wrote the paper.

Reviewers: B.C., University College London; and A.K.C., Massachusetts Institute of Technology.

The authors declare no conflict of interest.

[1] Z.S. and Y.E. contributed equally to this work.

[2] To whom correspondence should be addressed. Email: ccallan@princeton.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1700241114/-/DCSupplemental.

by distinct recombination events, standard tools that assign a unique recombination scenario to each sequence (17–20) give biased estimates that would limit our ability to detect the developmental changes that interest us here. In previous work (9, 10) we showed how to overcome this problem, using an approach that assigns probabilities to different ways of generating a sequence (see *Materials and Methods* for details). This approach allows us to accurately quantify and track diversity as a function of developmental age, in both the thymus and the periphery.

Here, we apply these methods to thymic and peripheral sequence repertoires of TCR beta-chain (TRB) genes of mice of varying ages: 17 d after conception and 4 d, 21 d, and 42 d after birth (see *Materials and Methods* and Table S1 for a complete summary of data). A distinct set of generation parameters was inferred for each of these repertoires. Only out-of-frame sequences, which are nonproductive and thus thought to be free of selection effects, were included in the inference.

Thymic Repertoires Reveal Mechanism of Diversity Maturation. The best-documented element of VDJ recombination known to change between fetal and adult life is the number of nontemplated (N) insertions at the junctions. In Fig. 1 we plot the marginal distributions of the number of N insertions at the VD and DJ junctions inferred from out-of-frame thymic sequences of mice at a sequence of ages. During the passage from fetal to mature animal, this distribution changes dramatically and rapidly. In the embryonic mouse, 90% of the sequences have no insertions, whereas this fraction drops to 10% in adult mice. This trend is consistent with previous observations in neonates (11) and is explained by the low level of expression of TdT before birth (12, 14). TdT is turned on after birth, and Fig. 1 indicates that the asymptotic level of TdT in adults must be reached before 21 d, as there are no noticeable differences between 21 d and 42 d. The distribution at 4 d shows an intermediate situation, roughly halfway between embryonic and adult. Whereas Fig. 1 shows that the inferred distribution of insertions is identical at the VD and DJ junctions in the thymus of fetal or adult mice, it is different at 4 d. This difference could be due to the temporal ordering of VDJ recombination: DJ recombination occurs before VD recombination, with a short time delay between the two, and the rise of TdT expression during this delay could explain the increased mean number of VD insertions relative to that of DJ insertions.

We asked whether features of the recombination process other than the number of insertions changed between embryonic and mature mice. We found that D and J gene choice did not change significantly ($P > 0.05$, t test corrected for multiple testing), whereas a few V genes had their use significantly ($P < 0.05$) increase (V4, V12-1, V26) or decrease (V14, V16, V17, V20, V22) with age (Figs. S1 and S2). The profiles of deletion showed no significant changes with age (Figs. S3 and S4) and neither did the frequencies of inserted N nucleotides (Fig. S5). These results confirm quantitatively that an increase in the number of untemplated insertions is the primary driver of diversity expansion in early life.

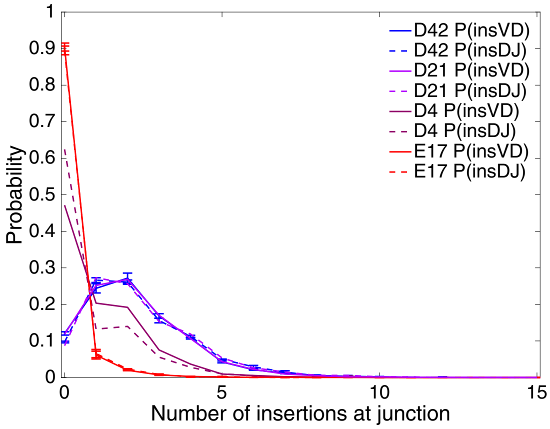


Fig. 1. Age-dependent insertion length distributions. Shown are distributions of the number of N insertions at the VD junction, *P*(insVD), and at the DJ junction, *P*(insDJ), inferred from individual mouse thymus datasets at different ages: embryo day 17 (E17) and 4 d, 21 d, and 42 d postbirth (D4, D21, and D42). The error bars indicate the variation across individuals (Table S1).

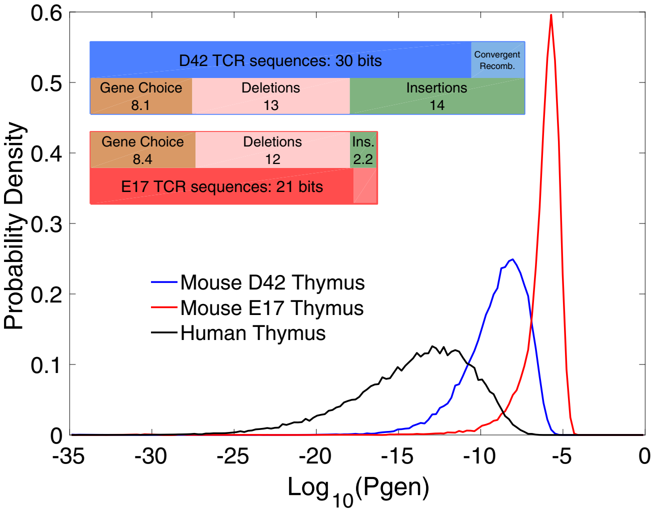


Fig. 2. Sequence entropy for thymic repertoires. Shown is distribution of the log generation probability for mouse thymic repertoires derived from 17 d embryo (E17) and 42 d postbirth animals (D42). The human generative probability distribution (9) is plotted for comparison. Shannon entropy is minus the mean over these distributions. *Inset* shows Shannon entropy at ages 17 d embryo and 42 d postbirth, decomposed into the components of the recombination scenario. Sequence entropy is recombination scenario entropy minus a correction for convergent recombination. Note that only the insertion component of scenario entropy changes significantly between embryonic and mature.

To quantify the overall change in diversity between TRB sequence repertoires at different ages, we calculated the Shannon entropy of their distributions. Entropy can be decomposed as a sum of contributions from gene choices, deletions, and insertions, from which a correction for convergent recombination must be subtracted (9). We find that the diversity of generated nucleotide sequences increased from 21 bits in fetal mice to 30 bits in adult mice. The change in repertoire diversity during this transition is almost entirely due to the change in the insertion profile, as can be seen in Fig. 2, *Inset*, where the different contributions to the sequence entropies of the fetal and mature sequence repertoires are compared.

The entropy is mathematically equal to the negative of the mean over the sequence repertoire of the logarithm of the generation probability. A plot of the distribution of generation probabilities, P_{gen}, over a typical repertoire (Fig. 2) shows that the generation probability of individual sequences ranges from a few parts per million to less than 1 part in 10^{18}.

Peripheral Repertoires Reflect Past History of Thymic Diversity. Our analysis of thymic repertoires has shown that the generative probability distribution for VDJ recombination changes dramatically in the days and weeks after birth. To understand how the evolution of VDJ recombination impacts repertoires, we need to account for the fact that peripheral compartments accumulate cells generated across earlier times, as sketched in Fig. 3. As a result, sequence repertoires must be described by a mixture of generative models with varying parameters that reflect past states of the generation process.

To use our inference procedure to quantify the state of mixing of repertoires, we must make some simplifying assumptions. First, given our observation that other features vary rather little with age (Figs. S1–S4), we assume that only the statistics of the untemplated insertions change with time. Second, we assume that the instantaneous insertion distribution function, $P_\alpha(n)$, interpolates linearly between the embryonic and adult distributions by setting $P_\alpha(n) = \alpha P_{\mathrm{emb}}(n) + (1-\alpha)P_{\mathrm{mat}}(n)$, where n is the number of insertions at a junction, P_{emb} is the distribution for the 17-d embryo, and P_{mat} is the adult distribution at 42 d, and $0 \le \alpha \le 1$ is an effective level of TdT measured by its impact on the number of insertions. This interpolation describes the data at day 4 (Fig. 2) accurately: The Kullback–Leibler divergence between P_α (for the optimal choice of α) and the directly inferred distribution is 0.1 bits, much less than the 2.6-bit entropy of the distribution. A more thorough validation would require data that more densely cover the early life period when the recombination machinery is changing.

The distribution P_α describes the TRB generation process at a fixed TdT level α. As explained above, repertoires in general represent the accumulated output of recombination events at earlier times and must be described by a mixture of processes at various α values. The generic mixture model for insertions n_1, n_2 at the VD and DJ junctions can thus be written as

$$P(n_1, n_2) = \int_0^1 d\alpha\, g(\alpha) P_\alpha(n_1) P_\alpha(n_2)$$
$$= P_{\bar\alpha}(n_1) P_{\bar\alpha}(n_2) + \mathrm{var}(\alpha)\Delta P(n_1)\Delta P(n_2), \quad \textbf{[1]}$$

where $g(\alpha)$ is the distribution of α in the repertoire reflecting the distribution of the past developmental ages at which its receptors were produced, $\bar\alpha$ and $\mathrm{var}(\alpha)$ are its mean and variance, and $\Delta P = P_{\mathrm{mat}} - P_{\mathrm{emb}}$. Conveniently, per the second line of Eq. **1**, the mixture distribution depends only on the mean and variance of α. The variance is constrained by $0 \le \mathrm{var}(\alpha) \le \bar\alpha(1-\bar\alpha)$ and gives a measure of the level of mixing in the repertoire. Zero variance means no mixing; i.e., all cells were created at a single effective TdT level $\alpha = \bar\alpha$. Maximal variance and mixing are attained when a fraction $\bar\alpha$ of cells fully expresses TdT ($\alpha = 1$), whereas the remaining fraction $1-\bar\alpha$ does not express TdT at all ($\alpha = 0$).

We estimate $\bar\alpha$ and $\mathrm{var}(\alpha)$ for our datasets by first inferring the joint distribution $P(\mathrm{insVD,insDJ})$ of Eq. **2** in *Materials and Methods*, using our inference technique, and then adjusting $\bar\alpha$ and $\mathrm{var}(\alpha)$ to obtain the best fit to Eq. **1**. The data points in Fig. 4*A* show the mean effective TdT level $\bar\alpha$ as a function of age for thymus or blood datasets, whereas the data points in Fig. 4*B* report the associated values of $\mathrm{var}(\alpha)$. Fig. 4*A* shows that the blood repertoire transitions from embryonic ($\alpha = 0$) to mature ($\alpha = 1$) with a time delay relative to the thymic repertoire. This result is expected because blood T cells are first produced in the thymus. The rise in TdT level results in an increase of diversity, measured by the entropy of recombination events (Fig. 4*A*, *Inset*). Fig. 4*B* shows that, although embryonic and adult repertoires have no mixing, $\mathrm{var}(\alpha) \approx 0$, all intermediate repertoires are mixed, with $\mathrm{var}(\alpha)$ significantly larger than 0. Although the thymus does not accumulate cells, T cells do spend a finite time in the thymus and, when TdT levels are changing fast, thymic repertoires are described by mixtures. Still, blood repertoires are substantially

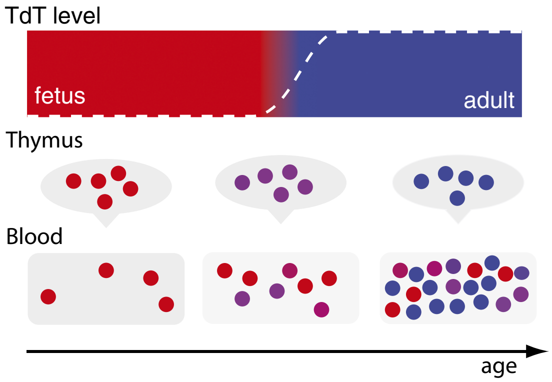


Fig. 3. Age-dependent recombination affects repertoire statistics. T cells recombine according to an insertion profile that depends on a time-dependent effective TdT level α (*Top*, dashed white line). Thymic repertoires have a unique statistical profile at any time and cells' output to the periphery at different times are described by different values of α (indicated by colors). The accumulating peripheral T-cell repertoire is described by a mixture model that accounts for the different numbers of T cells emitted at different times.

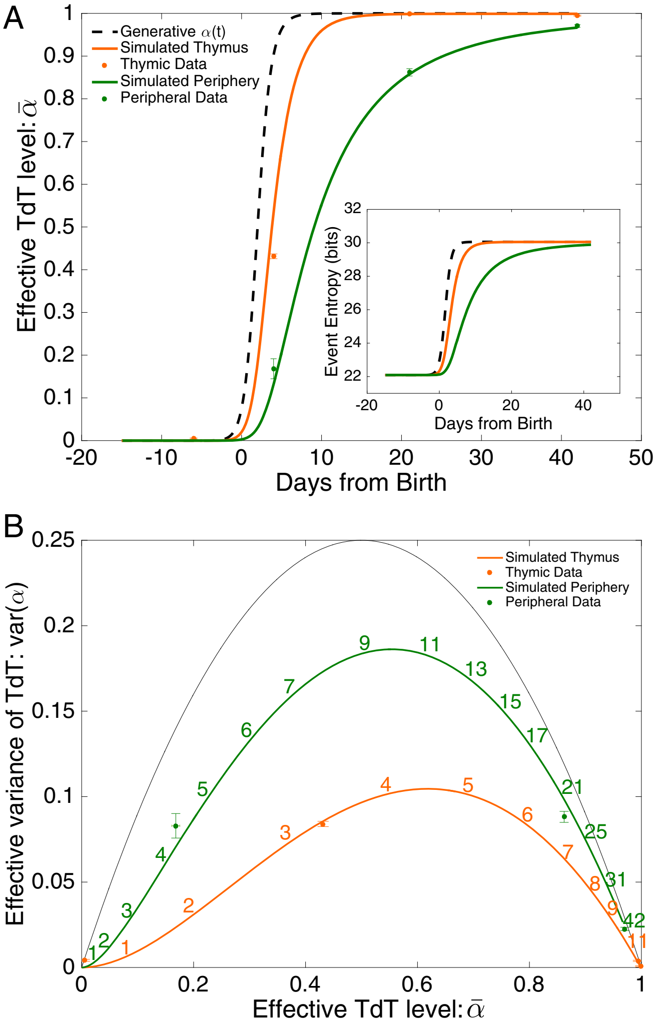


Fig. 4. Age dependence of the recombination process. The effective TdT level $0 \le \alpha \le 1$ is estimated as an interpolation parameter between the recombination statistics of embryonic and mature animals. (*A*) Mean effective TdT $\bar\alpha$ at various ages (17 d embryo and 4 d, 21 d, and 42 d postbirth), from different tissues: thymus and periphery. Periphery is taken from blood, except for day 4 for which it is taken from spleen. The data points are compared with a minimal model of thymic entry, residence, and output, with a sharply increasing effective TdT level $\alpha(t)$ represented by the dashed line (*Materials and Methods*). *A*, *Inset* shows recombination entropy as a function of age, as predicted by the model. (*B*) The variance of α, which measures the level of mixing in the repertoires, is shown as a function of its mean for both data (symbols) and the prediction of the minimal model (lines). The black line shows the maximal possible variance $\bar\alpha(1-\bar\alpha)$. Numbers represent age from birth in days.

more mixed than thymic repertoires, as expected because the thymus contains cells that have recombined over a narrow range of TdT levels α, whereas blood contains cells with a greater range of ages and of values of α.

The behavior of the data displayed in Fig. 4 can be better understood by comparison with a simple model (*Materials and Methods* and *Mixture Model*). In this model the effective TdT level $\alpha(t)$ in the thymus is given by a sharply rising Hill function (Fig. 4*A*, dashed curve), recombined cells are created at a rate that increases rapidly with time, and cells reside in the thymus for 3 d on average, after which they are released into the periphery. Although model parameters were chosen to reproduce the observed behavior quantitatively, we did not attempt a formal fit to the data, because of the paucity of data points. Results for $\bar{\alpha}(t)$ and $\mathrm{var}(\alpha)(t)$ are displayed in Fig. 4 (orange and green curves). The model recapitulates the delay in maturation between thymus and blood (Fig. 4*A*) and also accounts for the observed level of mixing as a function of time in blood and thymus (Fig. 4*B*). The model curves in Fig. 4*B* are parametric in time (time stamps added for clarity) and it is significant that the data points lie close to points on the model curves at the right age.

Selection Shapes the In-Frame Repertoire. Our discussion so far has focused on the evolution of the generative model for VDJ recombination, a model inferred from nonproductive, out-of-frame sequences. We now discuss what can be learned from in-frame sequences. Because they can code for functional surface receptor proteins, their statistics will be modified, relative to the generative model statistics, by selection effects. To quantify selection, we focus on the complementarity-determining region 3 (CDR3) of the beta chain, the region thought to encode most of the functional diversity of the T-cell repertoire. We define the CDR3 as the amino acid sequence running from a conserved cysteine in the V segment to a conserved phenylalanine in the J segment. We associate to each possible CDR3 amino acid sequence σ a selection factor $Q(\sigma)$, defined as the ratio of its probability of being observed in the data to its probability of having been generated. For computability, Q is taken to be a product of factors $q_{i,L}(a)$ reflecting the selection effect of each amino acid a at each position i in a CDR3 of length L. The collection of these subfactors defines a selection motif. The algorithms for inferring the subfactors from the data were developed in previous work on human TRB sequences (21) (see *Materials and Methods* for details).

We inferred selection motifs for a variety of thymic and blood repertoires (Fig. S6). These motifs are very consistent between thymus and blood of mice of the same age, with weaker consistency between mice of different ages (Fig. S7). Similarity between blood and thymus may seem surprising, as we could have expected a significant fraction of TRBs from thymic cells to have been sequenced before any selection effect, making them statistically closer to out-of-frame sequences. These observations suggest that our selection factors primarily capture selection for the ability of the coded protein to fold into a displayable receptor and may not capture more subtle effects such as negative selection against self-recognition. In Fig. S6, we also display patterns of correlation between mature mouse selection factors and quantitative amino acid biochemical properties; significant, but hard-to-interpret, patterns are apparent.

Our method attaches two hidden variables to each in-frame sequence: its probability P_{gen} of being generated in a VDJ recombination event and the selection factor Q governing its probability of then appearing in a thymic or peripheral sequence repertoire. Distributions of sequence repertoires over these variables give interesting insights into selection. We recall that, for humans, we found that the distribution of in-frame sequences was strongly skewed to higher P_{gen}: If a sequence was more likely to be created, it was more likely to be selected (21). Fig. 5*A* shows that this correlation does not hold for mice: The P_{gen} distribution for in-frame repertoires is virtually the same as that created by VDJ recombination. The difference between humans and mice is even more apparent in the distribution of sequence repertoires over the selection factor Q (Fig. 5*B*). For mice, selection is a weak effect: The distribution over Q is narrow, nearly centered about $Q = 1$ (no selection), and moves to only slightly higher values of Q in going from generated to selected repertoires. For humans, the primitively generated repertoire has a large fraction of sequences with a low probability of being selected (Fig. 5*B*). Consequently, selection purges a large fraction of sequences and substantially modifies the repertoire statistics.

Discussion

VDJ recombination is a stochastic process that produces the initial diversity on which the adaptive immune system relies to develop a functional and diverse repertoire of receptor specificities. Previous studies have shown that this diversity is limited in neonates compared with adults, either by biasing the choice of gene segments (22–25) or by having a small number of N insertions (11, 26). Combining high-throughput sequencing with statistical analysis of murine T-cell receptor beta chains, we analyzed the dynamics of maturation of VDJ recombination. This analysis allowed us to precisely quantify, in bits, how diversity increases with age, from embryo to adult. We found that the most significant change in the recombination statistics was the number of untemplated N insertions, which sharply increases around the age of 4 d, from almost no insertions to the amount found in adults. Low numbers of insertions in neonates and during

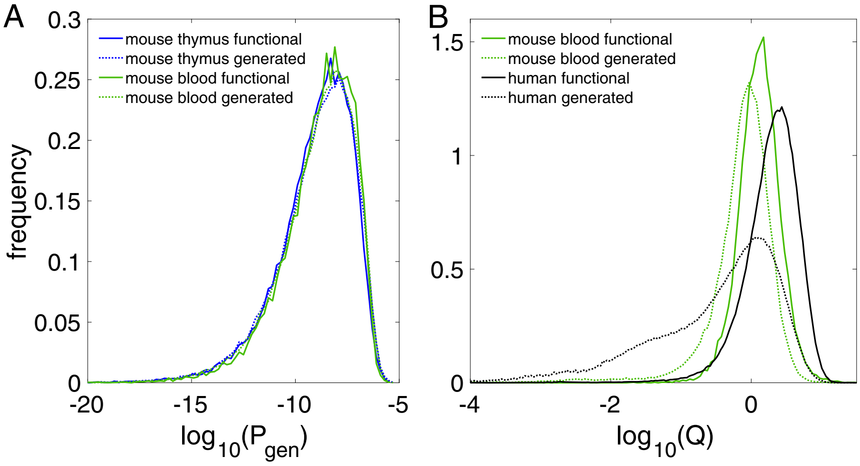


Fig. 5. Distributions with respect to generation probability and selection factor. (*A*) The distribution of mature mouse in-frame sequences with respect to primitive generation probability. The distributions for thymus and blood are essentially the same; the predicted distribution using Q factors on primitively generated sequences agrees perfectly with data. (*B*) Distribution over selection factor Q of primitively generated sequences vs. in-frame data sequences. For mice, both distributions are concentrated around $Q = 1$; for humans, one-third of primitively generated sequences have $Q < 0.1$ (i.e., are strongly selected against).

embryonic development are common to both B- and T-cell receptors (27) and are attributed to low TdT expression (12–14). Diversity can be further reduced in embryos by concentrating gene use on only a few combinations, as was shown for Ig in mice (22, 23), humans (24), and more recently zebrafish, using high-throughput sequencing (25). Similar observations were made on human TCR beta chains (28, 29). By contrast, we found only minor differences in TRB gene use between embryonic and adult mice (Fig. S1), meaning that the reduced number of N insertions is the only factor limiting diversity in the embryo relative to the adult.

One can only speculate about the biological function of the lack of N insertions in embryos and very young individuals. Rearrangements with no insertions may encode particular specificities that are effectively innate. The invariant TCRs of mucosal-associated invariant T cells (MAIT) and natural killer T cells (NKT) are specific examples of such genetically encoded receptors (30). These TCRs, which lack N insertions, are formed with high probability by VDJ recombination (31) and are further selected to be very conserved. Receptors lacking N insertions may provide neonates with a minimal set of innate-like specificities, ensuring basic immunity (32), which is later completed by the full diversity of receptors endowed with N insertions.

Our analysis highlights the importance of focusing on the underlying statistical ensembles from which repertoires are drawn, rather than looking for significance in the sequences themselves. Although sequence repertoires are contingent and noisy, with little to no overlap between individuals, their statistical properties are consistent between individuals, as was already noted for humans (9, 21). Crucially, a statistical treatment is essential for tracking the precise dynamics of N insertions with age, as deterministic assignments give systematically biased estimates of these numbers (Fig. S8). In our study of the development of repertoire diversity, we avoided the confounding factor of possibly time-dependent selection by analyzing out-of-frame rearrangements.

Using an analysis tailored to the productive repertoire, we were able to assign to any CDR3 sequence a statistical measure of selection, quantifying the probability that an amino acid sequence, once generated, goes on to become a functional receptor. The inferred selection motifs were very similar in thymic and blood repertoires. This lack of difference suggests that our measure of selection is mainly sensitive to basic selection against misfolding of the encoded receptor protein (which would have the same effect on thymic and blood repertoires) and is relatively insensitive to thymic selection against self-recognizing receptors (which would be present in blood but not in thymus). Negative selection in the thymus of course exists, but previous work on statistically characterizing its nature has shown that strong effects are localized in the few residues that actually contact presented peptides (33–35). Because we are statistically characterizing selection across the more extensive CDR3 region (11 aa long on average), it is perhaps not surprising that localized negative selection effects are washed out. It would clearly be productive to incorporate such insights into our approach.

The study of mice raises interesting and puzzling questions about sequence diversity. We found that the mature mouse repertoire is 9 bits (or $2^9 \approx 500$-fold) more diverse than the embryonic repertoire. This wide diversity of sequences is accompanied by a wide diversity of generation probabilities: In the mature repertoire, typical generation probabilities vary from more than 10^{-6} to less than 10^{-18} (Fig. 3). TCRs with very low generation probabilities should be private, i.e., not likely to be generated independently in two mice, whereas TCRs with the highest generation probabilities can be public, i.e., frequently found in different mice (36). The estimate of generation probabilities afforded by our model could therefore be useful for studying the origin of public TCR repertoires in mice (37).

The mature mouse repertoire is 14 bits (or $2^{14} \approx 16{,}000$-fold) less diverse than the human T-cell repertoire, owing to a lower number of N insertions (typically $\sim$3 per junction in mice vs. $\sim$5 in humans). Humans and mice have to deal with presumably equally complex pathogen environments, and it would be natural to expect their immune systems to have similar levels of sequence diversity. It is intriguing that this $\sim$10,000-fold difference in potential diversity closely reflects the difference in the number of T cells in the two species ($\sim 10^7$ in mice vs. $\sim 10^{11}$ in humans). Another difference with humans is the timing of the transition. The number of TCR N insertions increases as early as the first semester of gestation in humans (28) and from the second semester for B-cell receptors (BCRs) (27). By contrast, our results for mice show a sharp transition soon after birth. Finally, we found that the inferred selection factors are weaker in mice than in humans. They are also not correlated with generation probability. As a result, selection does not affect the entropy of the mouse repertoire, as it does to the human repertoire (21). The reason for this stark difference is not clear, and it would be interesting to see whether wild mice, as opposed to the inbred laboratory mice we have studied, show the same effect.

Materials and Methods

Datasets. The data used in our analyses are 87-bp (and 60-bp) nucleotide sequences covering the variable region of the rearranged mouse TRB gene. The sequences were obtained by Adaptive Biosciences, using their TRB DNA sequencing protocol (including error correction on the basis of multiple reads of each unique DNA sequence) applied to biological samples provided by two of the authors (A.J.L. and C.S.D.). The samples comprised blood, spleen, and thymus samples from mice sacrificed at four different ages: 17-d embryo and 4 d, 21 d, and 42 d postbirth (the library preparation and sequencing for day 42 thymic samples were replicated). The mice were Black 6 laboratory mice (Jackson Laboratories) raised in standard laboratory conditions. The animal care committee of the New Jersey Medical School provided approvals for the experiments carried out with mice in this publication. The numbers of unique sequences in the various datasets were a few tens of thousands on average (with a few datasets providing more than 10^5 unique sequences). The sequencing of the mature (D42) thymus samples was replicated once. Detailed statistics on the datasets are provided in Table S1. The full sequence datasets are available, along with an explanatory README file, at princeton.edu/~ccallan/MousePaper/data/.

Stochastic Model for VDJ Recombination. Out-of-frame data sequences were used to infer the statistical ensemble of sequences produced directly by the VDJ recombination process. We assume that the probability distribution for the generative events involved in VDJ recombination of TRB has the form (9, 38)

$$\begin{aligned} P(S) = {} & P(\mathrm{V})P(\mathrm{D},\mathrm{J})P(\mathrm{insVD},\mathrm{insDJ}) \\ & P(\mathrm{delV}|\mathrm{V})P(\mathrm{delDl},\mathrm{delDr}|\mathrm{D})P(\mathrm{delJ}|\mathrm{J}) \\ & P(s_1)P(s_2|s_1)\cdots P(s_{\mathrm{insVD}}|s_{\mathrm{insVD}-1}) \\ & P(t_1)P(t_2|t_1)\cdots P(t_{\mathrm{insDJ}}|t_{\mathrm{insDJ}-1}), \end{aligned} \qquad [2]$$

where S is a recombination scenario (defined by gene choice, numbers of deletions, and number and identity of insertions) and where each factor in the equation is a distribution over the possible elements of the scenario, $P(\mathrm{V})P(\mathrm{D},\mathrm{J})$ is the distribution of choices of the three kinds of gene segments (note that a correlation between the two D genes and the two clusters of J genes is imposed by genome topology), and $P(\mathrm{delV}|\mathrm{V})$ is the distribution of numbers of deletions from the end of a particular gene V (and likewise for D and J). A scenario includes specific N nucleotide insertions $s_1 \ldots s_{\mathrm{insVD}}$ and $t_1 \ldots t_{\mathrm{insVD}}$ at the VD and DJ junctions, and $P(\mathrm{insVD},\mathrm{insDJ})$ is the distribution of the total numbers of such insertions, whereas $P(s_i|s_{i-1})$, etc. describes the probability of inserting particular N nucleotides. Note that Eq. **2** gives the probability of recombination scenarios, not sequences. To obtain the probability of generating a specific sequence, one must sum the expression in Eq. **2** over all of the recombination scenarios that result in that sequence. We determine the component probability distributions in Eq. **2**, $P(\mathrm{V})$, $P(\mathrm{D},\mathrm{J})$, $P(\mathrm{insVD},\mathrm{insDJ})$, and so forth, directly from the data using the principle of maximum likelihood. The likelihood of a whole dataset is given by the product, over all of the unique out-of-frame sequences in the dataset, of the generation probabilities of those sequences according to the model. In practice, likelihood maximization is performed using an expectation–maximization algorithm, as explained in ref. 9.

The main assumption underlying Eq. **2** is its simple product structure, reflecting the independence of the enzymes that carry out different steps of the process. Another assumption is that the probability of inserting a given N nucleotide depends only on the identity of the nucleotide that precedes it (Markov assumption). We self-consistently checked the validity of these assumptions by verifying a posteriori that almost no unaccounted correlations between the recombination events were left in the data that were not explicitly assumed (*Validation of the Structure of the Sequence Generation Model* and Fig. S9) and by showing that the statistics of triplets of N insertions were well predicted by the Markov model (Fig. S5). We also compared our probabilistically inferred distributions of recombination scenario variables with distributions assembled from assignments made by a standard VDJ alignment software package (17). We found that these nonprobabilistic alignment methods greatly overestimate the fraction of sequences with no N nucleotide insertions and significantly violate the D-J pairing rule imposed by genome topology, whereas the probabilistic method does not (Fig. S8). This discrepancy is what motivates our use of a probabilistic approach. The inferred model features were very reproducible across individuals of the same age (Figs. S2 and S4).

Selection Model. Following previous work on human TRB sequences (21), we associate to each possible CDR3 amino acid sequence σ of the TRB repertoire a selection factor $Q(\sigma)$ defined as the ratio between the probability of generation of σ in VDJ recombination and its probability of occurrence in unique in-frame data sequences. The selection factor Q is assumed to be a product of subfactors related to the V and J gene choice (q_{VJ}), CDR3 length L (q_L), and amino acid identity a at each position i of the CDR3 ($q_{i,L}(a)$):

$$P_{\mathrm{obs}}(\sigma) = Q(\sigma)\cdot P_{\mathrm{gen}}(\sigma) = \frac{1}{Z} q_L q_{V\,J} \prod_{i=1}^{L} q_{i,\,L}(a)\cdot P_{\mathrm{gen}}(\sigma). \qquad [3]$$

The $q_{i,\,L}(a)$ are normalized such that their sum over amino acids at each i and L is unity and Z enforces an overall normalization. The set of all these subfactors defines a motif of selection across all possible TRB sequences, and a likelihood maximization procedure allows us to infer the best selection factors from the data.

To test the consistency of the selection model, Eq. **3**, we also learned a more general model where, instead of taking the $q_{i,L}(a)$ to depend on amino acid identity, they were functions of the 62 codons. Using the same inference procedure, we found that selection factors for degenerate codons for the same amino acid were consistent (Fig. S10). This agreement justifies the assumption that selection depends only on amino acid composition.

Code Availability. The Matlab software for implementing the inference procedures is available at princeton.edu/∼ccallan/MousePaper/software/. The results of the inference, along with instructions on how to use these files to recreate the figures in this paper, are available at princeton.edu/∼ccallan/MousePaper/results/.

Model of Repertoire Maturation. TCRs are produced in the thymus with a time-dependent effective TdT level $\alpha(t) = [1 + ((T_{\mathrm{half}} - T_{\mathrm{start}})/(t - T_{\mathrm{start}}))^{20}]^{-1}$, with a production rate $\theta(t) \propto (1 + (t - T_{\mathrm{start}} + 1)^{2.1})$ (arbitrary units). Time is in days, with birth at $t = 0$, $T_{\mathrm{start}} = -15$, and $T_{\mathrm{half}} = 2$. Cells reside in the thymus for an average of 3 d (exponentially distributed time), after which they are released into the periphery. The simulation is followed from $t = T_{\mathrm{start}}$ (early embryo) to $t = 42$ (age of oldest dataset).

ACKNOWLEDGMENTS. The work of Y.E., T.M., and A.M.W. was supported in part by European Research Council Starting Grant 306312. The work of Y.E. was supported in part by The V Foundation for Cancer Research Grant D2015-032. The work of C.G.C. and Z.S. was supported in part by National Science Foundation Grant PHY-1305525.

1. Weinstein JA, Jiang N, White RA, Fisher DS, Quake SR (2009) High-throughput sequencing of the zebrafish antibody repertoire. *Science* 324:807–810.
2. Boyd SD, et al. (2009) Measurement and clinical monitoring of human lymphocyte clonality by massively parallel {VDJ} pyrosequencing. *Sci Transl Med* 1:12ra23.
3. Robins HS, et al. (2009) Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* 114:4099–4107.
4. Benichou J, Ben-Hamo R, Louzoun Y, Efroni S (2012) Rep-Seq: Uncovering the immunological repertoire through next-generation sequencing. *Immunology* 135(3):183–191.
5. Baum PD, Venturi V, Price DA (2012) Wrestling with the repertoire: The promise and perils of next generation sequencing for antigen receptors. *Eur J Immunol* 42(11):2834–2839.
6. Six A, et al. (2013) The past, present and future of immune repertoire biology - the rise of next-generation repertoire analysis. *Front Immunol* 4:413.
7. Georgiou G, et al. (2014) The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat Biotechnol* 32:158–168.
8. Calis JJ, Rosenberg BR (2014) Characterizing immune repertoires by high throughput sequencing: Strategies and applications. *Trends Immunol* 35(12):581–590.
9. Murugan A, Mora T, Walczak AM, Callan CG (2012) Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc Natl Acad Sci USA* 109:16161–16166.
10. Elhanati Y, Marcou Q, Mora T, Walczak AM (2016) repgenHMM: A dynamic programming tool to infer the rules of immune receptor generation from sequence data. *Bioinformatics* 32(13):1943–1951.
11. Feeney AJ (1990) Lack of N regions in fetal and neonatal mouse immunoglobulin V-D-J junctional sequences. *J Exp Med* 172:1377–1390.
12. Bogue M, Gilfillan S, Benoist C, Mathis D (1992) Regulation of N-region diversity in antigen receptors through thymocyte differentiation and thymus ontogeny. *Proc Natl Acad Sci USA* 89:11011–11015.
13. Komori T, Okada A, Stewart V, Alt FW (1993) Lack of N regions in antigen receptor variable region genes of TdT-deficient lymphocytes. *Science* 261:1171–1175.
14. Gilfillan S, Dierich A, Lemeur M, Benoist C, Mathis D (1993) Mice lacking TdT: Mature animals with an immature lymphocyte repertoire. *Science* 261:1175–1178.
15. Britanova OV, et al. (2016) Dynamics of individual T cell repertoires: From cord blood to centenarians. *J Immunol* 196(12):5005–5013.
16. Pogorelyy MV, et al. (2016) Persisting fetal clonotypes influence the structure and overlap of adult human T cell receptor repertoires. arXiv qbio:1–21.
17. Brochet X, Lefranc MP, Giudicelli V (2008) IMGT/V-QUEST: The highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res* 36:W503–W508.
18. Thomas N, Heather J, Ndifon W, Shawe-Taylor J, Chain B (2013) Decombinator: A tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. *Bioinformatics* 29:542–550.
19. Bolotin DA, et al. (2015) MiXCR: Software for comprehensive adaptive immunity profiling. *Nat Methods* 12:380–381.
20. Yu Y, Ceredig R, Seoighe C (2015) LymAnalyzer: A tool for comprehensive analysis of next generation sequencing data of T cell receptors and immunoglobulins. *Nucleic Acids Res* 44(4):e31.
21. Elhanati Y, Murugan A, Callan CG, Mora T, Walczak AM (2014) Quantifying selection in immune receptor repertoires. *Proc Natl Acad Sci USA* 111:9875–9880.
22. Yancopoulos GD, et al. (1984) Preferential utilization of the most JH-proximal VH gene segments in pre-B-cell lines. *Nature* 311:727–733.
23. Perlmutter RM, Kearney JF, Chang SP, Hood LE (1985) Developmentally controlled expression of immunoglobulin VH genes. *Science* 227:1597–1601.
24. Schroeder HW, et al. (1988) Physical linkage of a human immunoglobulin heavy chain variable region gene segment to diversity and joining region elements. *Proc Natl Acad Sci USA* 85:8196–8200.
25. Jiang N, et al. (2011) Determinism and stochasticity during maturation of the zebrafish antibody repertoire. *Proc Natl Acad Sci USA* 108:5348–5353.
26. Feeney AJ (1991) Junctional sequences of fetal T cell receptor beta chains have few N regions. *J Exp Med* 174:115–124.
27. Schroeder HW, Zhang L, Philips JB (2001) Slow, programmed maturation of the immunoglobulin HCDR3 repertoire during the third trimester of fetal life. *Blood* 98:2745–2751.
28. George JF, Schroeder HW (1992) Developmental regulation of D beta reading frame and junctional diversity in T cell receptor-beta transcripts from human thymus. *J Immunol* 148:1230–1239.
29. Raaphorst FM, Kaijzel EL, van Tol MJ, Vossen JM, van den Elsen PJ (1994) Non-random employment of V beta 6 and J beta gene elements and conserved amino acid usage profiles in CDR3 regions of human fetal and adult TCR beta chain rearrangements. *Int Immunol* 6:1–9.
30. Le Bourhis L, et al. (2011) Mucosal-associated invariant T cells: Unconventional development and function. *Trends Immunol* 32:212–218.
31. Greenaway HY, et al. (2013) NKT and MAIT invariant TCRα sequences can be produced efficiently by VJ gene recombination. *Immunobiology* 218:213–224.
32. Gilfillan S, et al. (1995) Efficient immune responses in mice lacking N-region diversity. *Eur J Immunol* 25:3115–3122.
33. Kosmrlj A, Jha AK, Huseby ES, Kardar M, Chakraborty AK (2008) How the thymus designs antigen-specific and self-tolerant T cell receptor sequences. *Proc Natl Acad Sci USA* 105:16671–16676.
34. Kosmrlj A, et al. (2010) Effects of thymic selection of the T-cell repertoire on HLA class I-associated control of HIV infection. *Nature* 465:350–354.
35. Stadinski BD, et al. (2016) Hydrophobic CDR3 residues promote the development of self-reactive T cells. *Nat Immunol* 17(8):946–955.
36. Venturi V, Price DA, Douek DC, Davenport MP (2008) The molecular basis for public T-cell responses? *Nat Rev Immunol* 8:231–238.
37. Madi A, et al. (2014) T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. *Genome Res* 24:1603–1612.
38. Elhanati Y, et al. (2015) Inferring processes underlying B-cell repertoire diversity. *Philos Trans R Soc Lond B Biol Sci* 370:20140243.
39. Treves A, Panzeri S (1995) The upward bias in measures of information derived from limited data samples. *Neural Comput* 7(2):399–407.
40. Monod MY, Giudicelli V, Chaume D, Lefranc MP (2004) IMGT/JunctionAnalysis: The first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. *Bioinformatics* 20(Suppl 1):i379–i385.

Supporting Information

Sethna et al. 10.1073/pnas.1700241114

Validation of the Structure of the Sequence Generation Model

Our inference procedure rests on a presumption of independence of the various factors in the generative model and a verification of that independence is an important aspect of our analysis. To address this issue, we compute the mutual information—a nonparametric measure of dependence between random variables—between the various recombination scenario variables and compare these numbers between the data and the generative model inferred from the same data.

The mutual information between two random variables x and y jointly distributed according to $p(x, y)$ is defined as

$$I(x, y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}. \qquad \textbf{[S1]}$$

The mutual information between the variables defining recombination scenarios, as computed from the inferred model itself, can be calculated exactly and is shown in the below-diagonal halves of the matrices of Fig. S9. By construction, the generative model has zero mutual information between certain variable pairs, e.g., the number of VD insertions and the choice of J gene, and nonzero mutual information between variables that correlate with each other either directly or indirectly, for example, between D and J gene choice, or between D choice and number of D deletions.

On the other hand, the inference procedure assigns multiple scenarios, each with its own probability, to each data sequence. For any pair of scenario variables (e.g., insVD and delJ) we can use these assignments over all of the data sequences to populate a list of pairs of values, weighted by scenario probabilities. From this list we can then compute the mutual information between the two elements of the pair, using the Treves–Panzeri correction (39) to account for small sample sizes. The mutual information computed in this fashion is displayed in the above-diagonal halves of the matrices of Fig. S9. The model form is considered accurate if the obtained mutual information agrees with that predicted by the model, i.e., if the matrices of Fig. S9 are symmetric.

Mixture Model

In this model, T cells are introduced into the thymus with rate $\theta(t)$, with an effective TdT level $\alpha(t)$. Cells leave the thymus into the periphery (blood and spleen) with constant rate τ_{thy}^{-1}. Under these assumptions, the total number of cells in the thymus, $N_{\text{thy}}(t)$, and the periphery, $N_{\text{peri}}(t)$, read

$$N_{\text{thy}}(t) = \int_{-\infty}^{t} \theta(t') e^{-t'/\tau_{\text{thy}}} \, dt', \qquad \textbf{[S2]}$$

$$N_{\text{peri}}(t) = \int_{-\infty}^{t} \theta(t') \left(1 - e^{-t'/\tau_{\text{thy}}}\right) dt'. \qquad \textbf{[S3]}$$

The distributions of α in these compartments are given by

$$g_{\text{thy}}(\alpha; t) = \frac{1}{N_{\text{thy}}(t)} \int^{t} \theta(t') e^{-\frac{t'}{\tau_{\text{thy}}}} \delta(\alpha(t') - \alpha) \, dt', \qquad \textbf{[S4]}$$

$$g_{\text{peri}}(\alpha; t) = \frac{1}{N_{\text{peri}}(t)} \int^{t} \theta(t')(1 - e^{-\frac{t'}{\tau_{\text{thy}}}}) \delta(\alpha(t') - \alpha) \, dt'.$$

The mean and variance of α, plotted in Fig. 5 of the main text, are calculated from these expressions. They are sufficient to calculate the joint distribution of insertions at the two junctions, per Eq. **2** of the main text.

Selection Model

The selection factor is defined as the fold change between the probability of generation of a TRB sequence s, $P_{\text{gen}}(s)$, and its probability among productive in-frame sequences, $P_{\text{post}}(s)$:

$$P_{\text{post}}(s) = Q(s) P_{\text{gen}}(s). \qquad \textbf{[S5]}$$

We assume that $Q(s)$ depends on s only through the amino acid translation of its CDR3 and that it takes the factorized form

$$Q(a) = q_L \prod_{i=1}^{L} q_{i;L}(a_i), \qquad \textbf{[S6]}$$

where $a = (a_1, \ldots, a_L)$ is the amino acid sequence of the CDR3, and L is its length. The q_L factors are length-specific factors, whereas $q_{i;L}(a)$ are composition-specific factors. All factors are simultaneously inferred by maximizing the likelihood of the in-frame sequences using gradient ascent, as explained in detail in ref. 21. Fig. S6 shows the values of $q_{i;L}(a)$ inferred from D42 mouse thymic sequences, and Fig. S7 compares the values of $q_{i;L}(a)$ inferred from different datasets. The selection factors were normalized for these figures such that $q_{i;L}(a) > 1$ indicates a positive contribution to the overall selection, whereas a value below 1 indicates a negative contribution.

Table S1. Sequence datasets used in analyses reported in main text

Age	Tissue	Read length, bp	Productive	Nonproductive
17 -d embryo	Thymus 2	87	11,199	15,368
17 -d embryo	Thymus 3	87	9,589	11,485
4 d	Thymus 1	87	9,165	4,687
4 d	Thymus 2	87	27,794	14,417
4 d	Spleen 1	87	447	325
4 d	Spleen 2	87	213	112
4 d	Spleen 3	87	1,356	668
21 d	Thymus	60	95,164	43,418
21 d	Blood	60	14,469	7,510
42 d	Thymus 1	87	33,028	16,159
42 d	Thymus 2	87	24,292	10,864
42 d	Thymus 3	87	24,846	12,006
42 d	Thymus 1 library replicate	87	137,233	67,232
42 d	Thymus 2 library replicate	87	61,990	30,390
42 d	Thymus 3 library replicate	87	83,591	40,425
42 d	Blood 1	87	16,642	7,550
42 d	Blood 2	87	3,256	1,235
42 d	Blood 3	87	16,858	7,306

Fig. S1. V and J gene use as a function of age. Shown are gene use probabilities derived from aggregated thymic datasets at different ages. Aggregated datasets, used to reduce small sample noise, are constructed by combining sequence repertoires from several individual mice at the same age. Whereas the overall use pattern is fairly stable with age, some genes undergo quite substantial changes in use.

Fig. S1

Fig. S2. Variation of V and J gene use across biological replicates. For each V or J gene we plot the thymic use fraction for all available individuals (one, two, or three, depending on the case) for the embryonic day 17 and for the 42-d mature mouse. Different individuals are indicated by repetitions of the same symbol, and their dispersion gives a rough measure of the biological sample variance of gene use.

Fig. S2

Fig. S3. Deletion profiles for different ages. The plots show the deletion profiles inferred from aggregated thymic datasets at three different ages. Deletion profiles depend on the identity of the gene being deleted and negative deletions are used to account for P nucleotides. Variations across sample ages are significantly larger than individual-to-individual variations at any given age. Plot headings record the gene identities and their use probabilities at age 42 d (only genes with use probability greater than 0.02 are displayed).

Fig. S3

Fig. S4. Variation of deletion profiles across biological replicates. Shown is the same convention as Fig. S2. Plot headings record the gene identities and their use probabilities (only genes with use probability greater than 0.02 are displayed). There is virtually no biological sample variance between individuals in a given sequencing run and very little between runs.

Fig. S4

Fig. S5. Frequencies of N insertions. (*Upper*) Dinucleotide insertion bias values $P(s_i|s_{i-1})$ inferred from individual thymus data for D4 and D42. Data are not presented for E17 because there are virtually no insertions in embryonic recombination. The 5′; nt refers to s_{i-1}, whereas the x axis corresponds to s_i. Sample variance at a given age is small, and frequencies are consistent from day 4 to day 42. (*Lower*) Predicted vs. observed trinucleotide insertion frequencies $P(s_i, s_{i+1}, s_{i+2})$ for D42 thymic data. The small scatter about the equality line indicates that the untemplated insertions are well described by a dinucleotide Markov process.

Fig. S5

Fig. S6. Selection factors. (*Upper*) Selection factors $q_{i,L}(a)$ inferred from D42 mouse thymic datasets. (*Lower*) (*A–K*) Pearson correlations between selection factors and amino acid biochemical properties. For each position i, CDR3 length L, and biochemical property, we display the Pearson correlation between $q_{i,L}(a)$ and the values of the biochemical property for the 20 amino acids a (the latter listed in the tables in *A–K*, *Left*).

Fig. S6

Fig. S7. Selection factor correlations. Scatter plot shows selection factors inferred from pairs of datasets. The pairings allow us to compare different tissues (blood or thymus) and ages. $R^2 = 1 - \mathrm{Var}(q' - q)/(\mathrm{Var}(q') + \mathrm{Var}(q))$ is a measure of the difference between two datasets; for identical datasets, one would find $R^2 = 1$.

Fig. S7

Fig. S8. Comparison of deterministic vs. model inference results. We compare the results of our model inference procedure used in this paper to a deterministic alignment as described in ref. 40. (*Upper*) The insertion profiles for D42 thymus (error bars are over individual mice). We see that the deterministic result gives a much higher probability for zero insertions and the VD and DJ junctions differ slightly. (*Lower*) Comparison of the deterministic and the model-inferred joint DJ uses for D42 and E17 thymus data. Note that the deterministic alignment gives a nonzero probability of DJ pairings that are topologically impossible (i.e., TRBJ1 genes cannot be paired with TRBD2). By contrast the model-inferred use is completely consistent with the topological constraint.

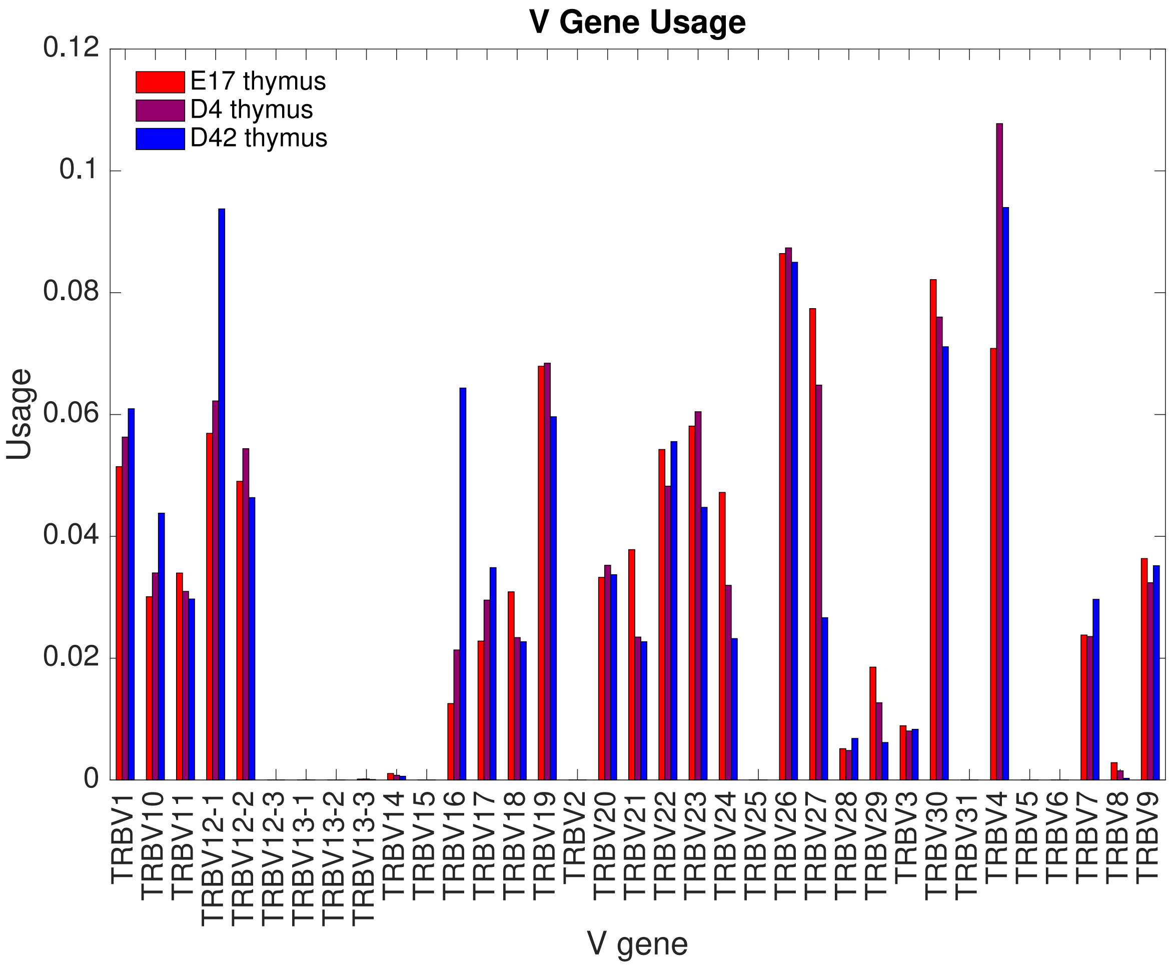
Fig. S8

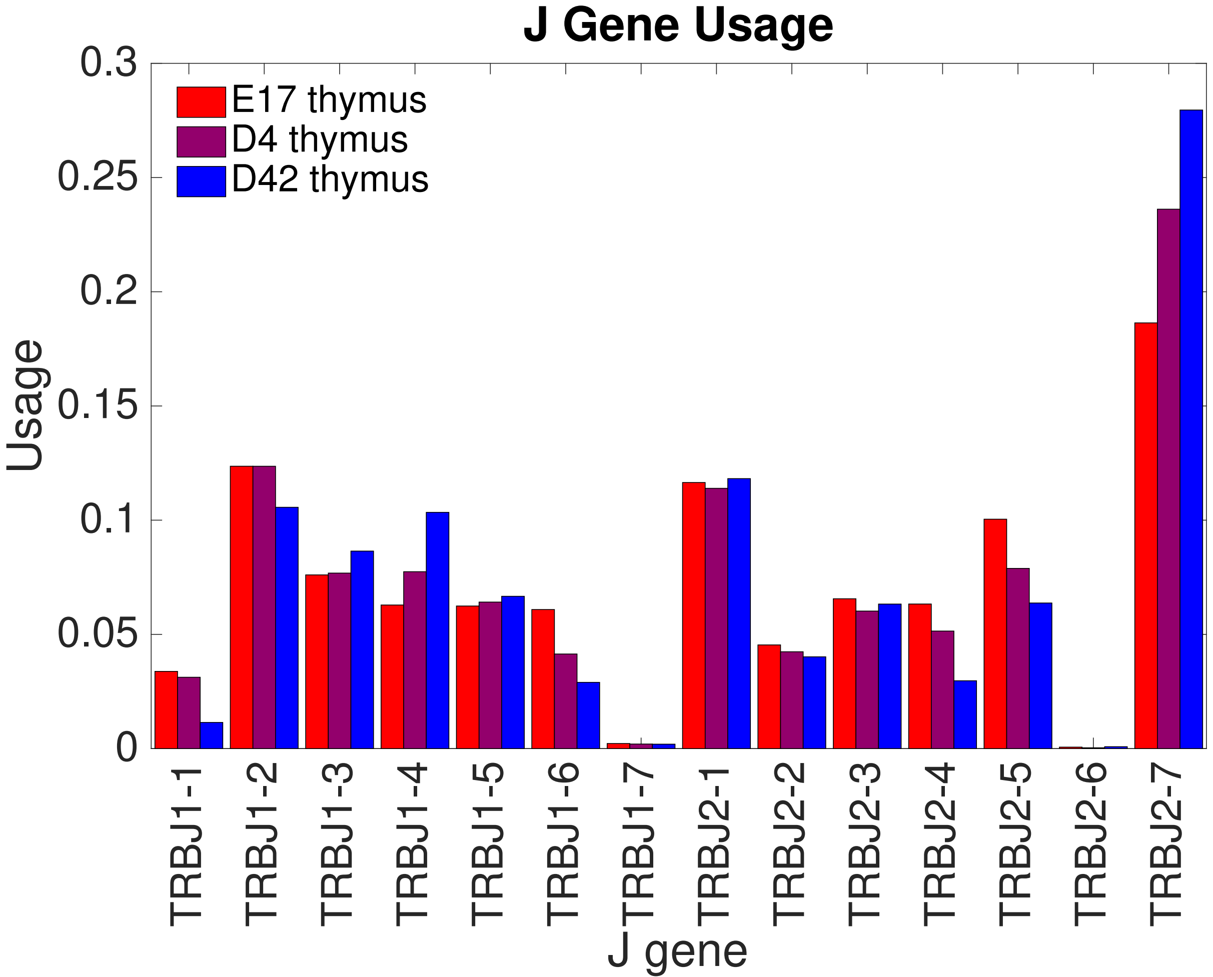
Fig. S9. Mutual information (MI) between hidden scenario variables. Each plot is based on the dataset indicated in the plot heading and the "correlated insertions" model inferred from that dataset. The correlated insertions models allow a general joint distribution $P(\mathrm{insVD}, \mathrm{insDJ})$ of VD and DJ insertions. Squares below the diagonal (*Lower Right* of each plot) display MI values for the model itself, whereas squares above the diagonal (*Upper Left* of each plot) display MI values derived from the data. The sum of all of the MI values calculated from the data (entries above the diagonal) that were not predicted by the model (zeros below the diagonal) does not exceed 0.14 bits in all considered cases. That number is very small compared with the individual entropies of each the scenario variables (each of the order of bits), indicating that these correlations are negligible.

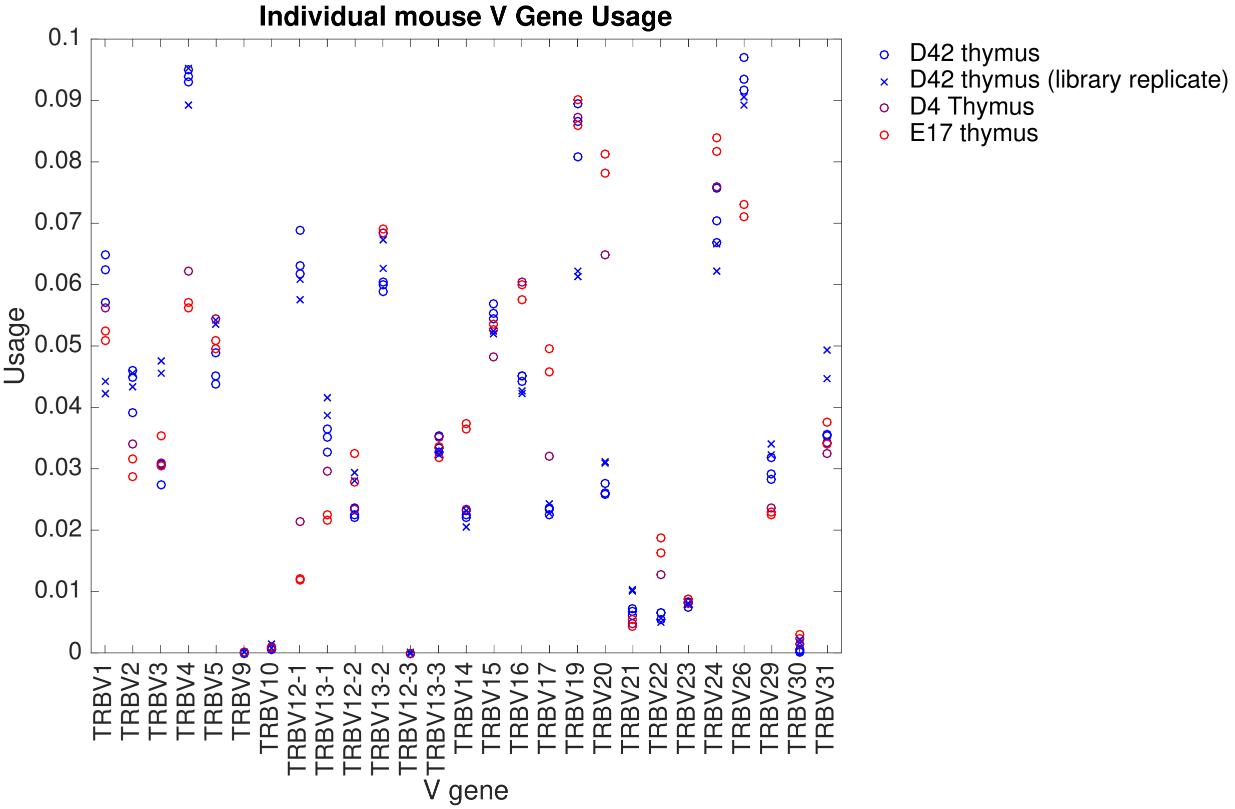
Fig. S9

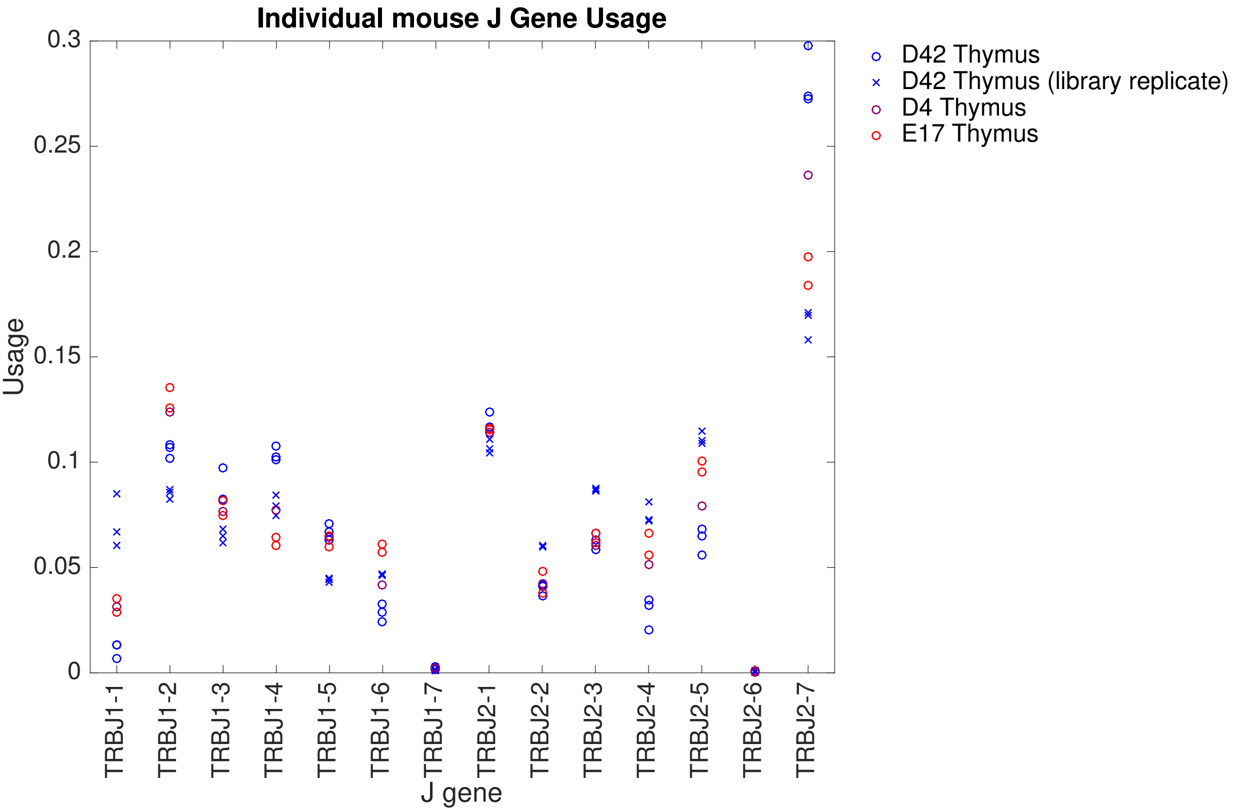
Fig. S10. Selection factors inferred for triplet codons and for amino acids are consistent. Selection factors, $qi;L(a)$, learned both for codons (red crosses) and for amino acids (blue circles), are plotted for each position in the CDR3 region for CDR3 sequences of length 12, as a function of the amino acids at each position. Generally, the codon selection factors follow the amino acid ones, signifying that selection indeed acts on the amino acid level. Codons for which there were not enough data to infer the selection factors are omitted.

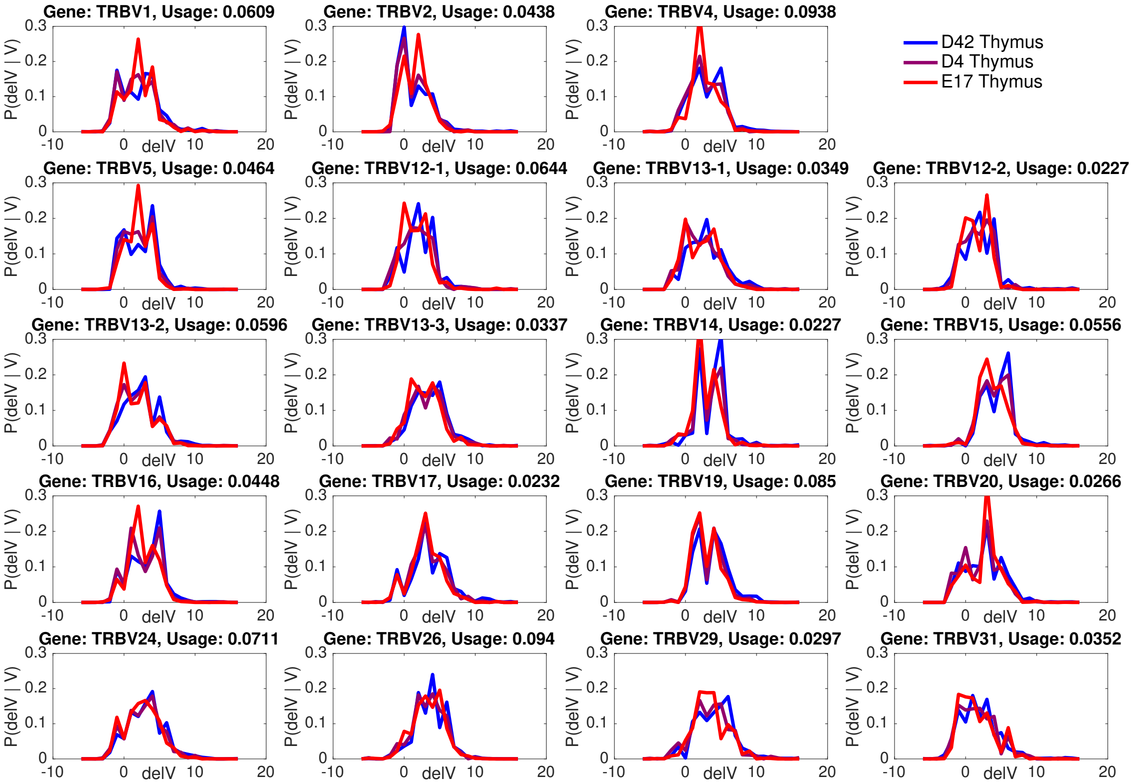
Fig. S10

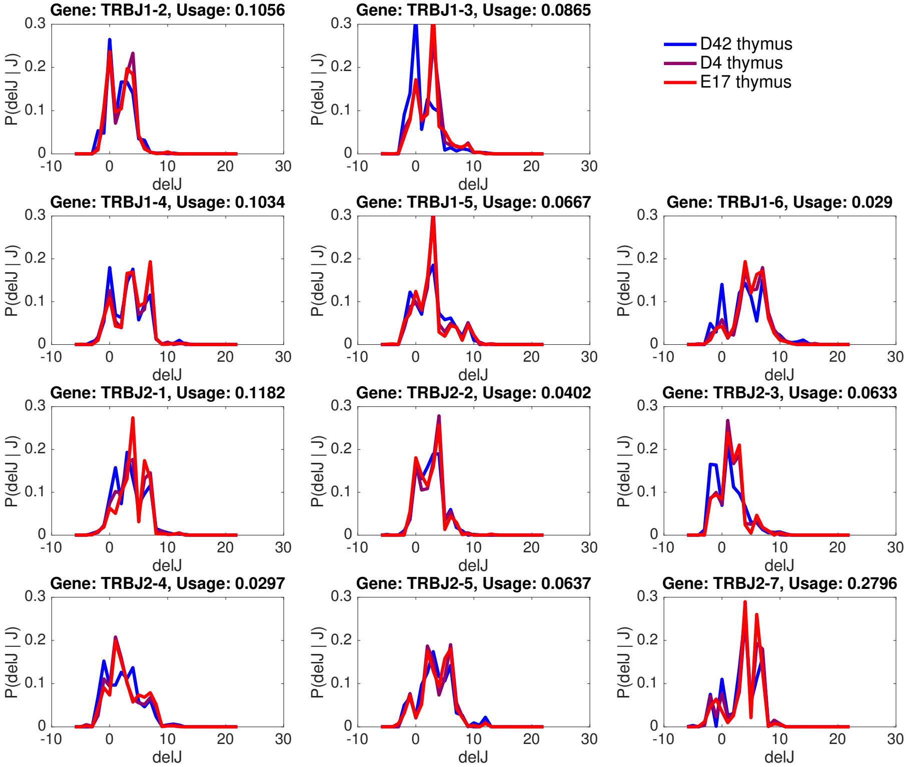

V Gene Usage
E17 thymus
D4 thymus
D42 thymus
Usage
0.12
0.1
0.08
0.06
0.04
0.02
0
TRBV1
TRBV10
TRBV11
TRBV12-1
TRBV12-2
TRBV12-3
TRBV13-1
TRBV13-2
TRBV13-3
TRBV14
TRBV15
TRBV16
TRBV17
TRBV18
TRBV19
TRBV2
TRBV20
TRBV21
TRBV22
TRBV23
TRBV24
TRBV25
TRBV26
TRBV27
TRBV28
TRBV29
TRBV3
TRBV30
TRBV31
TRBV4
TRBV5
TRBV6
TRBV7
TRBV8
TRBV9
V gene

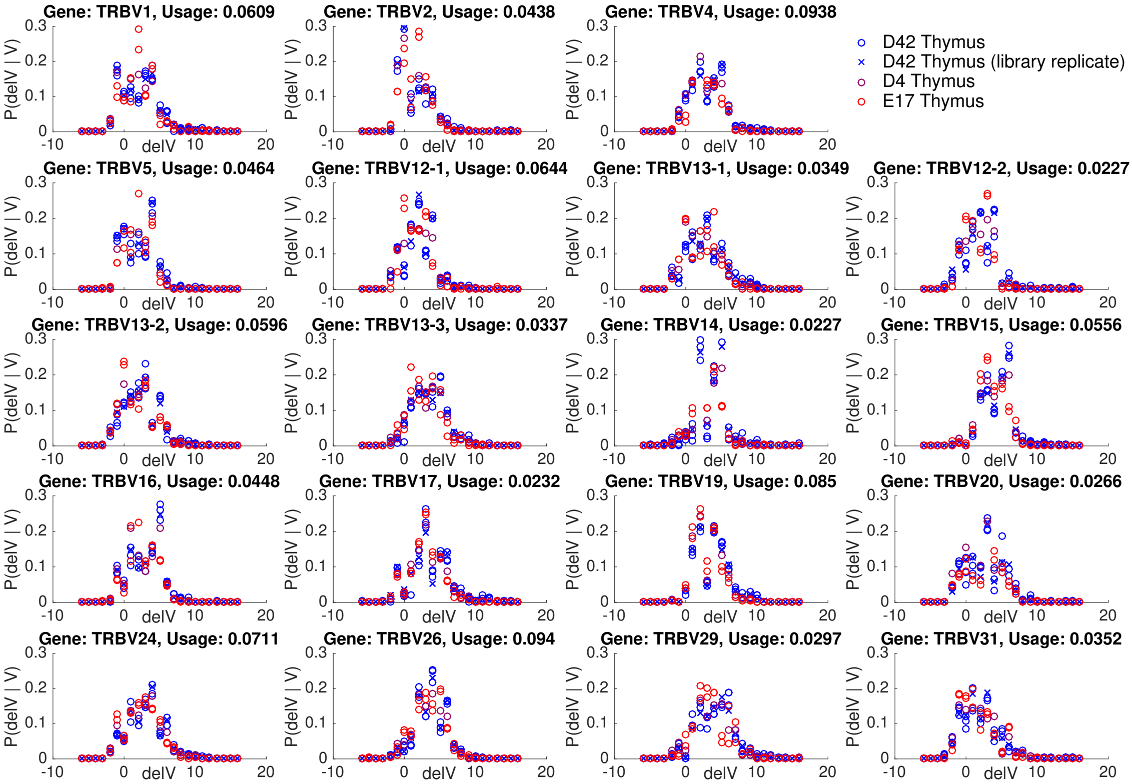

J Gene Usage
E17 thymus
D4 thymus
D42 thymus
Usage
0.3
0.25
0.2
0.15
0.1
0.05
0
TRBJ1-1
TRBJ1-2
TRBJ1-3
TRBJ1-4
TRBJ1-5
TRBJ1-6
TRBJ1-7
TRBJ2-1
TRBJ2-2
TRBJ2-3
TRBJ2-4
TRBJ2-5
TRBJ2-6
TRBJ2-7
J gene

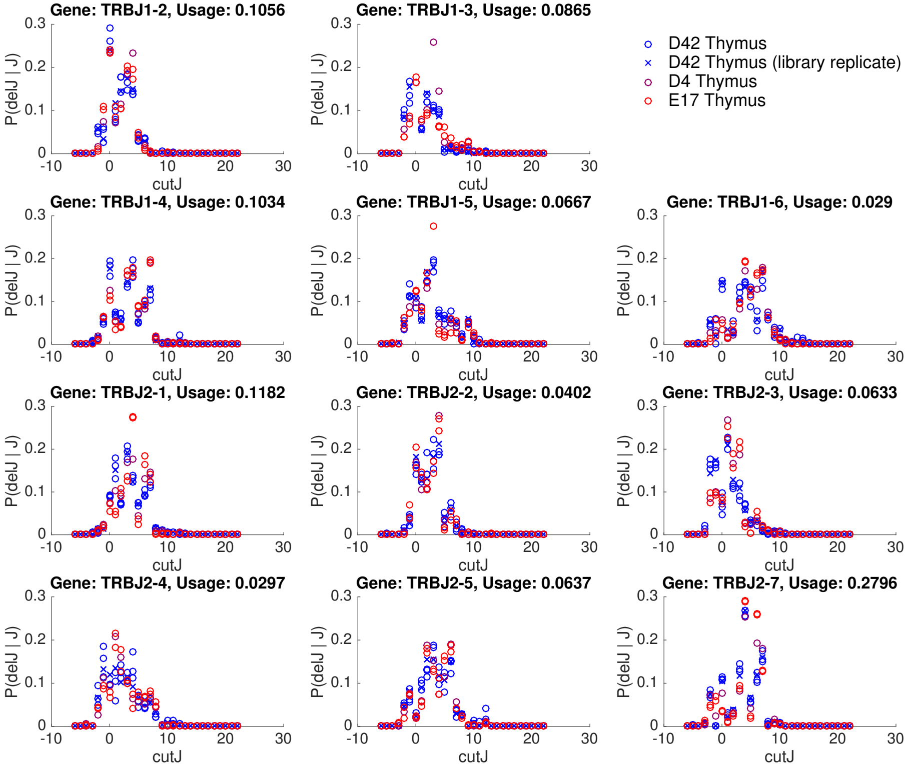

Individual mouse V Gene Usage
○ D42 thymus
× D42 thymus (library replicate)
○ D4 Thymus
○ E17 thymus
Usage
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.1
TRBV1
TRBV2
TRBV3
TRBV4
TRBV5
TRBV9
TRBV10
TRBV12-1
TRBV13-1
TRBV12-2
TRBV13-2
TRBV12-3
TRBV13-3
TRBV14
TRBV15
TRBV16
TRBV17
TRBV19
TRBV20
TRBV21
TRBV22
TRBV23
TRBV24
TRBV26
TRBV29
TRBV30
TRBV31
V gene

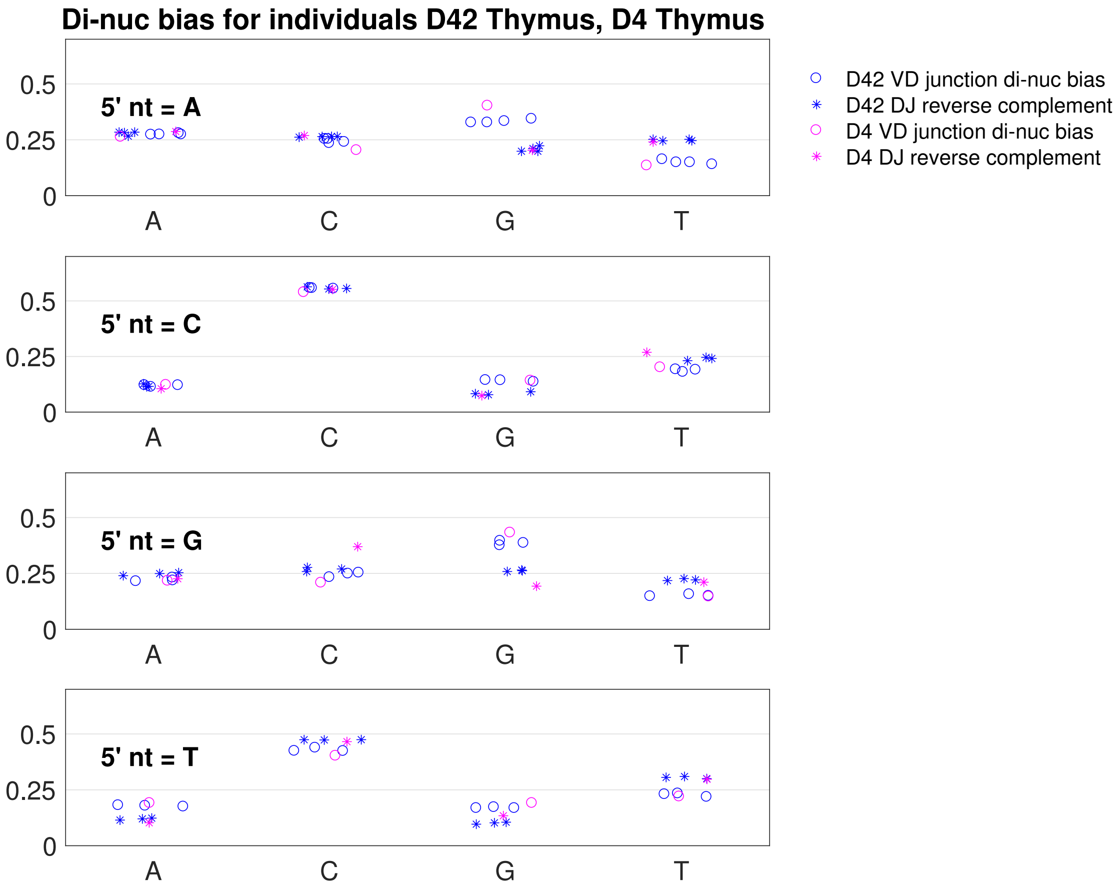

Individual mouse J Gene Usage
○ D42 Thymus
× D42 Thymus (library replicate)
○ D4 Thymus
○ E17 Thymus
Usage
0
0.05
0.1
0.15
0.2
0.25
0.3
TRBJ1-1
TRBJ1-2
TRBJ1-3
TRBJ1-4
TRBJ1-5
TRBJ1-6
TRBJ1-7
TRBJ2-1
TRBJ2-2
TRBJ2-3
TRBJ2-4
TRBJ2-5
TRBJ2-6
TRBJ2-7
J gene

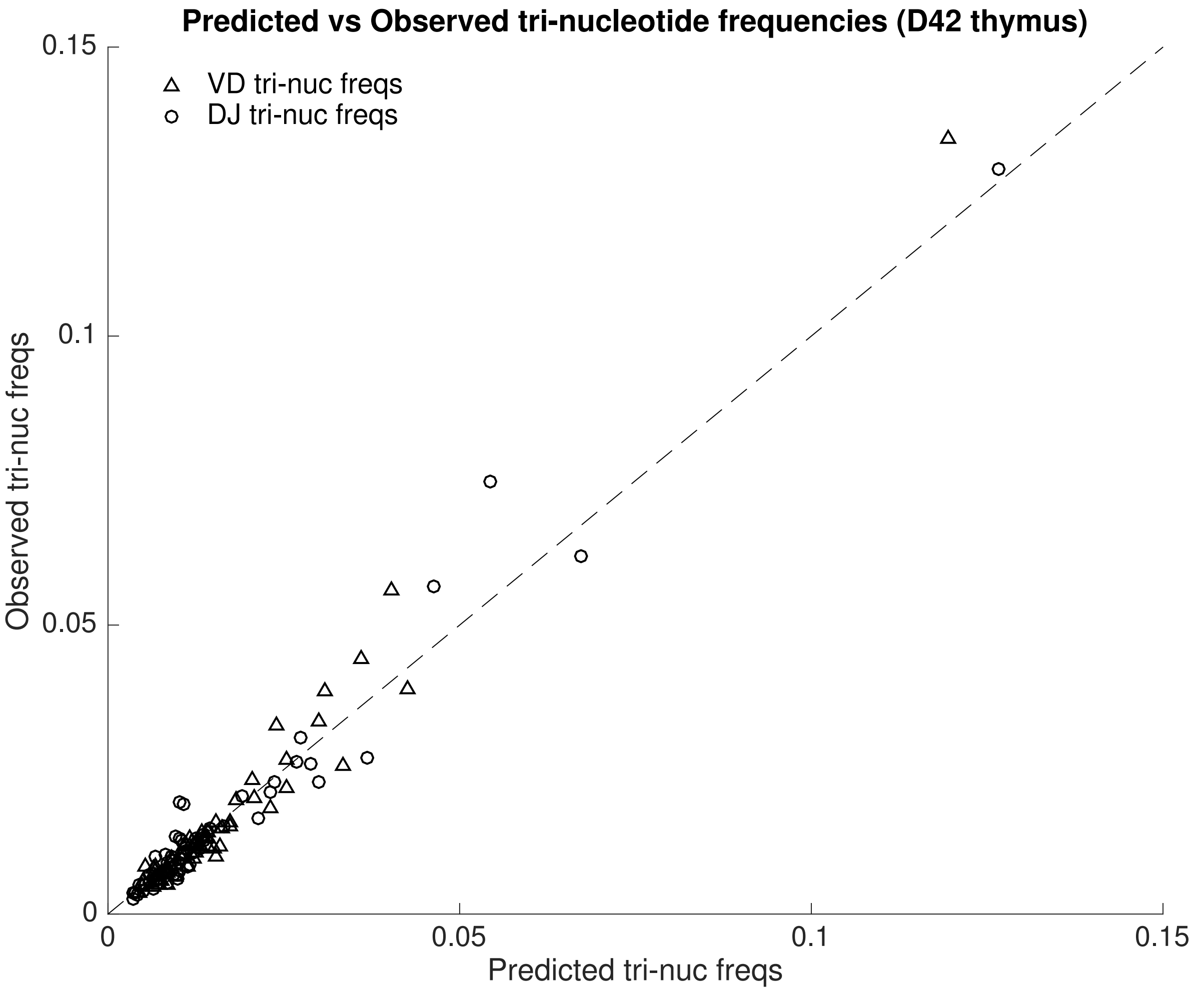

Gene: TRBV1, Usage: 0.0609
Gene: TRBV2, Usage: 0.0438
Gene: TRBV4, Usage: 0.0938
D42 Thymus
D4 Thymus
E17 Thymus
Gene: TRBV5, Usage: 0.0464
Gene: TRBV12-1, Usage: 0.0644
Gene: TRBV13-1, Usage: 0.0349
Gene: TRBV12-2, Usage: 0.0227
Gene: TRBV13-2, Usage: 0.0596
Gene: TRBV13-3, Usage: 0.0337
Gene: TRBV14, Usage: 0.0227
Gene: TRBV15, Usage: 0.0556
Gene: TRBV16, Usage: 0.0448
Gene: TRBV17, Usage: 0.0232
Gene: TRBV19, Usage: 0.085
Gene: TRBV20, Usage: 0.0266
Gene: TRBV24, Usage: 0.0711
Gene: TRBV26, Usage: 0.094
Gene: TRBV29, Usage: 0.0297
Gene: TRBV31, Usage: 0.0352
P(delV | V)
delV
0
0.1
0.2
0.3
-10
0
10
20

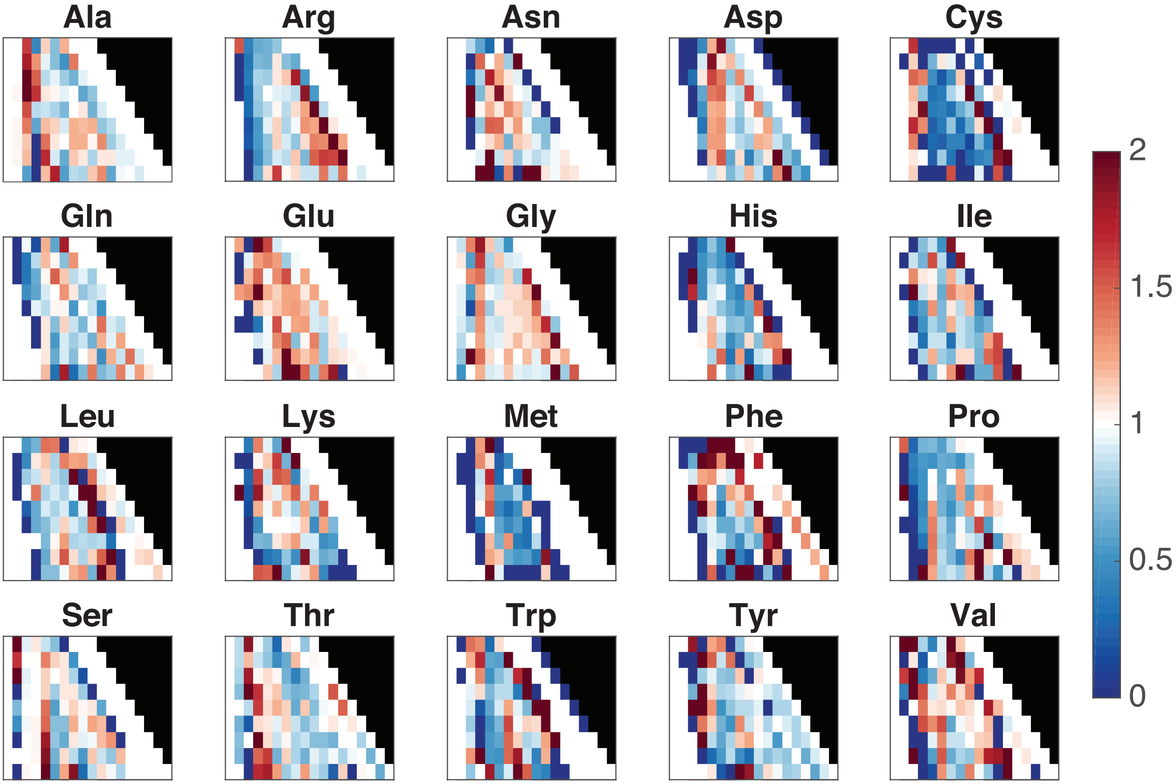

Gene: TRBJ1-2, Usage: 0.1056
Gene: TRBJ1-3, Usage: 0.0865
D42 thymus
D4 thymus
E17 thymus
Gene: TRBJ1-4, Usage: 0.1034
Gene: TRBJ1-5, Usage: 0.0667
Gene: TRBJ1-6, Usage: 0.029
Gene: TRBJ2-1, Usage: 0.1182
Gene: TRBJ2-2, Usage: 0.0402
Gene: TRBJ2-3, Usage: 0.0633
Gene: TRBJ2-4, Usage: 0.0297
Gene: TRBJ2-5, Usage: 0.0637
Gene: TRBJ2-7, Usage: 0.2796
P(delJ | J)
delJ
0
0.1
0.2
0.3
-10
0
10
20
30

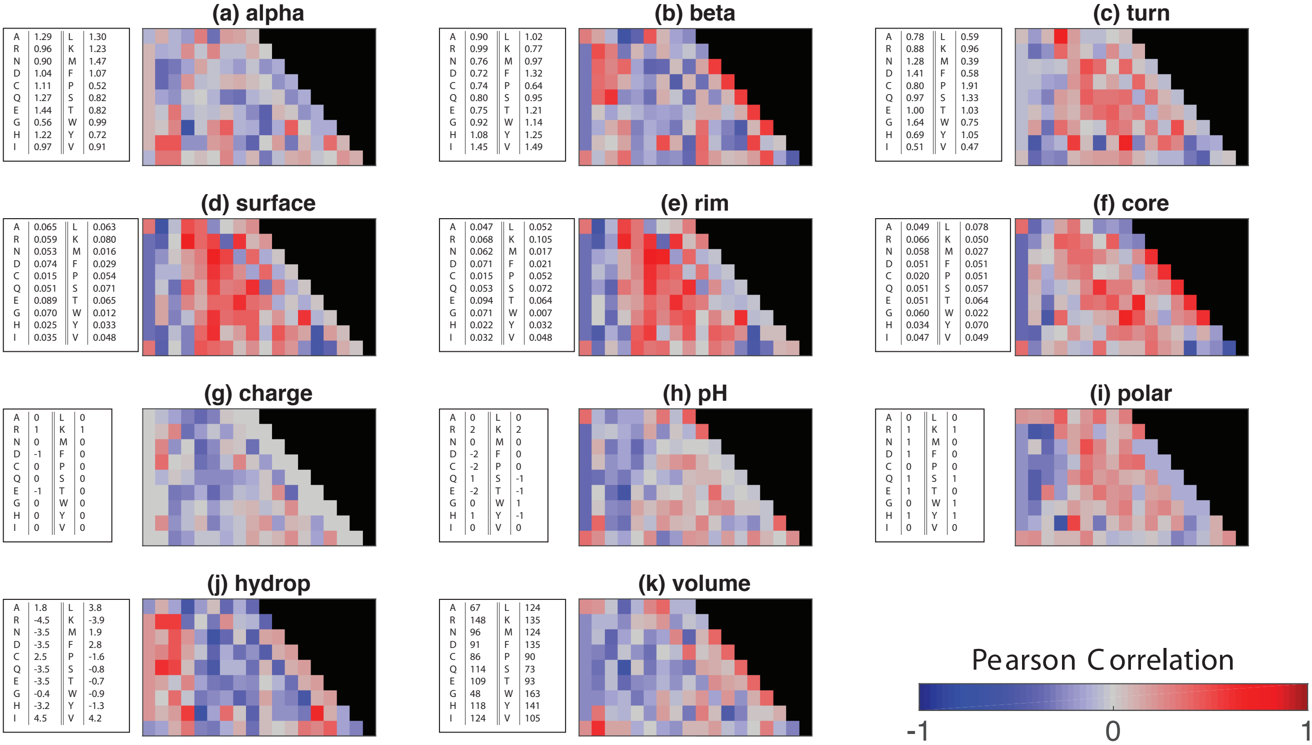

Gene: TRBV1, Usage: 0.0609
Gene: TRBV2, Usage: 0.0438
Gene: TRBV4, Usage: 0.0938
Gene: TRBV5, Usage: 0.0464
Gene: TRBV12-1, Usage: 0.0644
Gene: TRBV13-1, Usage: 0.0349
Gene: TRBV12-2, Usage: 0.0227
Gene: TRBV13-2, Usage: 0.0596
Gene: TRBV13-3, Usage: 0.0337
Gene: TRBV14, Usage: 0.0227
Gene: TRBV15, Usage: 0.0556
Gene: TRBV16, Usage: 0.0448
Gene: TRBV17, Usage: 0.0232
Gene: TRBV19, Usage: 0.085
Gene: TRBV20, Usage: 0.0266
Gene: TRBV24, Usage: 0.0711
Gene: TRBV26, Usage: 0.094
Gene: TRBV29, Usage: 0.0297
Gene: TRBV31, Usage: 0.0352
P(delV | V)
0
0.1
0.2
0.3
delV
-10
0
10
20
D42 Thymus
D42 Thymus (library replicate)
D4 Thymus
E17 Thymus


Gene: TRBJ1-2, Usage: 0.1056
Gene: TRBJ1-3, Usage: 0.0865
Gene: TRBJ1-4, Usage: 0.1034
Gene: TRBJ1-5, Usage: 0.0667
Gene: TRBJ1-6, Usage: 0.029
Gene: TRBJ2-1, Usage: 0.1182
Gene: TRBJ2-2, Usage: 0.0402
Gene: TRBJ2-3, Usage: 0.0633
Gene: TRBJ2-4, Usage: 0.0297
Gene: TRBJ2-5, Usage: 0.0637
Gene: TRBJ2-7, Usage: 0.2796
P(delJ | J)
0
0.1
0.2
0.3
cutJ
-10
0
10
20
30
D42 Thymus
D42 Thymus (library replicate)
D4 Thymus
E17 Thymus


Di-nuc bias for individuals D42 Thymus, D4 Thymus
5' nt = A
5' nt = C
5' nt = G
5' nt = T
0
0.25
0.5
A
C
G
T
○ D42 VD junction di-nuc bias
* D42 DJ reverse complement
○ D4 VD junction di-nuc bias
* D4 DJ reverse complement


Predicted vs Observed tri-nucleotide frequencies (D42 thymus)
△ VD tri-nuc freqs
○ DJ tri-nuc freqs
Observed tri-nuc freqs
Predicted tri-nuc freqs
0
0.05
0.1
0.15
0
0.05
0.1
0.15





(a) alpha

A	1.29	L	1.30
R	0.96	K	1.23
N	0.90	M	1.47
D	1.04	F	1.07
C	1.11	P	0.52
Q	1.27	S	0.82
E	1.44	T	0.82
G	0.56	W	0.99
H	1.22	Y	0.72
I	0.97	V	0.91

(b) beta

A	0.90	L	1.02
R	0.99	K	0.77
N	0.76	M	0.97
D	0.72	F	1.32
C	0.74	P	0.64
Q	0.80	S	0.95
E	0.75	T	1.21
G	0.92	W	1.14
H	1.08	Y	1.25
I	1.45	V	1.49

(c) turn

A	0.78	L	0.59
R	0.88	K	0.96
N	1.28	M	0.39
D	1.41	F	0.58
C	0.80	P	1.91
Q	0.97	S	1.33
E	1.00	T	1.03
G	1.64	W	0.75
H	0.69	Y	1.05
I	0.51	V	0.47

(d) surface

A	0.065	L	0.063
R	0.059	K	0.080
N	0.053	M	0.016
D	0.074	F	0.029
C	0.015	P	0.054
Q	0.051	S	0.071
E	0.089	T	0.065
G	0.070	W	0.012
H	0.025	Y	0.033
I	0.035	V	0.048

(e) rim

A	0.047	L	0.052
R	0.068	K	0.105
N	0.062	M	0.017
D	0.071	F	0.021
C	0.015	P	0.052
Q	0.053	S	0.072
E	0.094	T	0.064
G	0.071	W	0.007
H	0.022	Y	0.032
I	0.032	V	0.048

(f) core

A	0.049	L	0.078
R	0.066	K	0.050
N	0.058	M	0.027
D	0.051	F	0.051
C	0.020	P	0.051
Q	0.051	S	0.057
E	0.051	T	0.064
G	0.060	W	0.022
H	0.034	Y	0.070
I	0.047	V	0.049

(g) charge

A	0	L	0
R	1	K	1
N	0	M	0
D	-1	F	0
C	0	P	0
Q	0	S	0
E	-1	T	0
G	0	W	0
H	0	Y	0
I	0	V	0

(h) pH

A	0	L	0
R	2	K	2
N	0	M	0
D	-2	F	0
C	-2	P	0
Q	1	S	-1
E	-2	T	-1
G	0	W	1
H	1	Y	-1
I	0	V	0

(i) polar

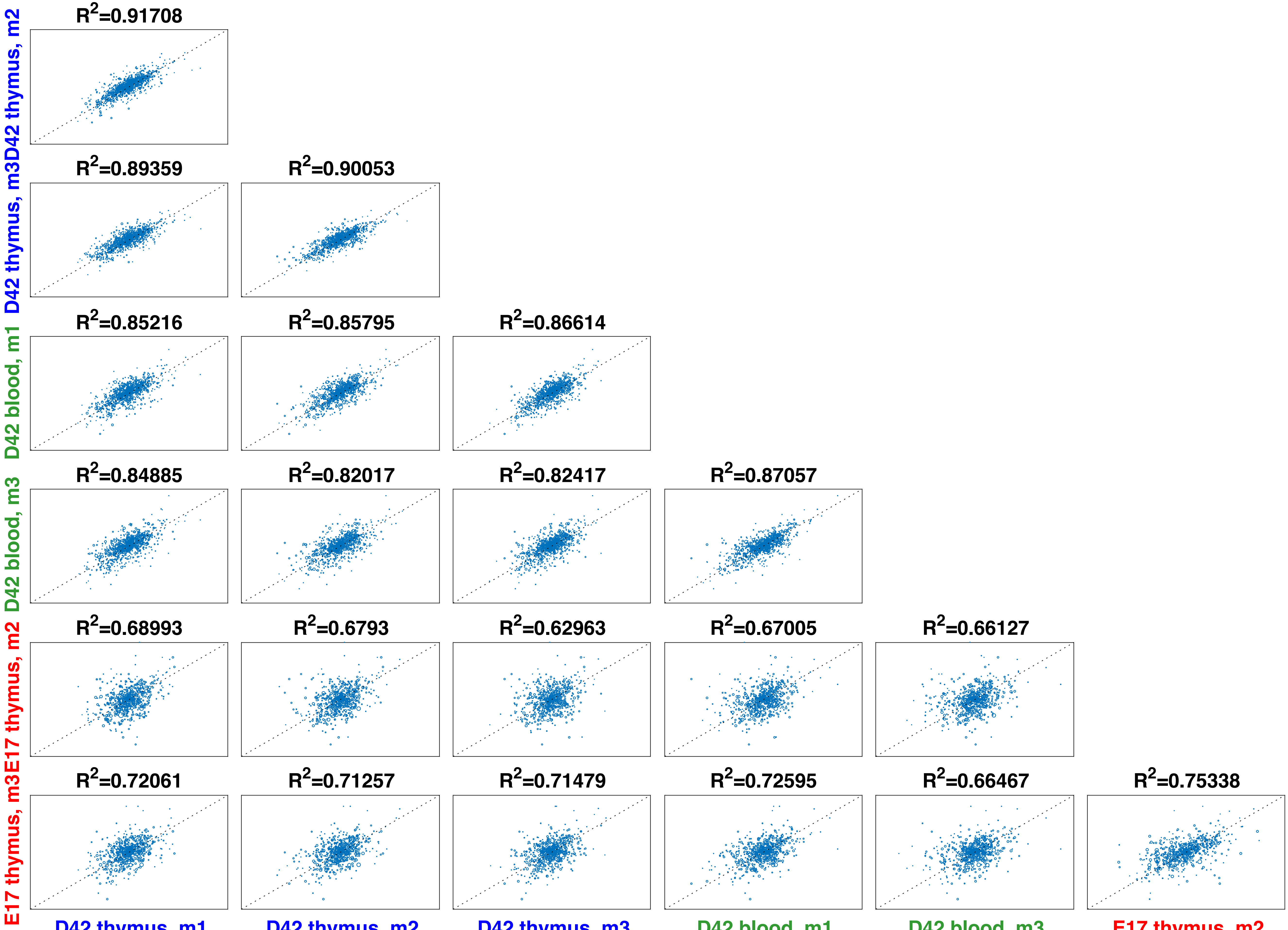
A	0	L	0
R	1	K	1
N	1	M	0
D	1	F	0
C	0	P	0
Q	1	S	1
E	1	T	0
G	0	W	1
H	1	Y	1
I	0	V	0

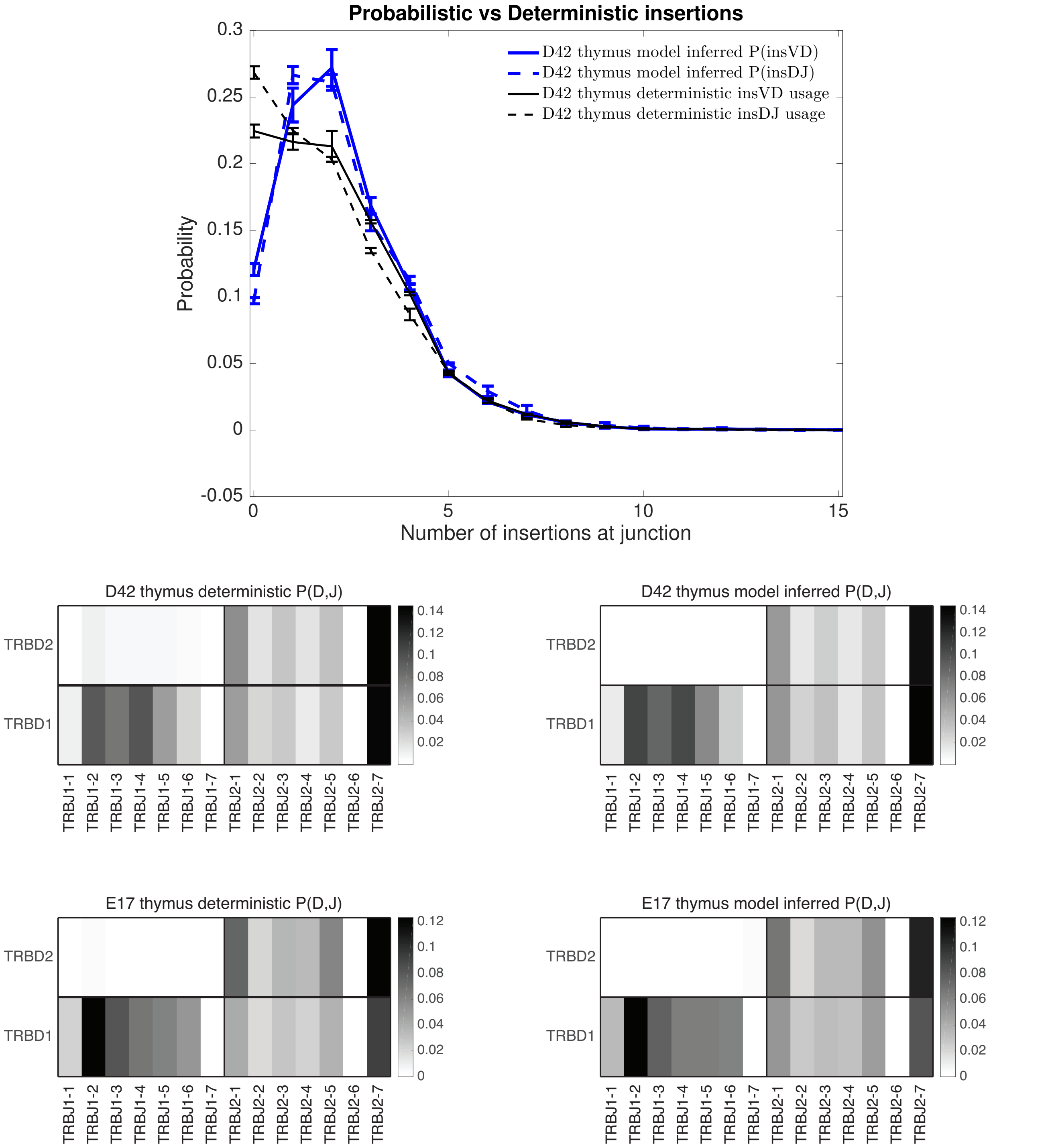
(j) hydrop

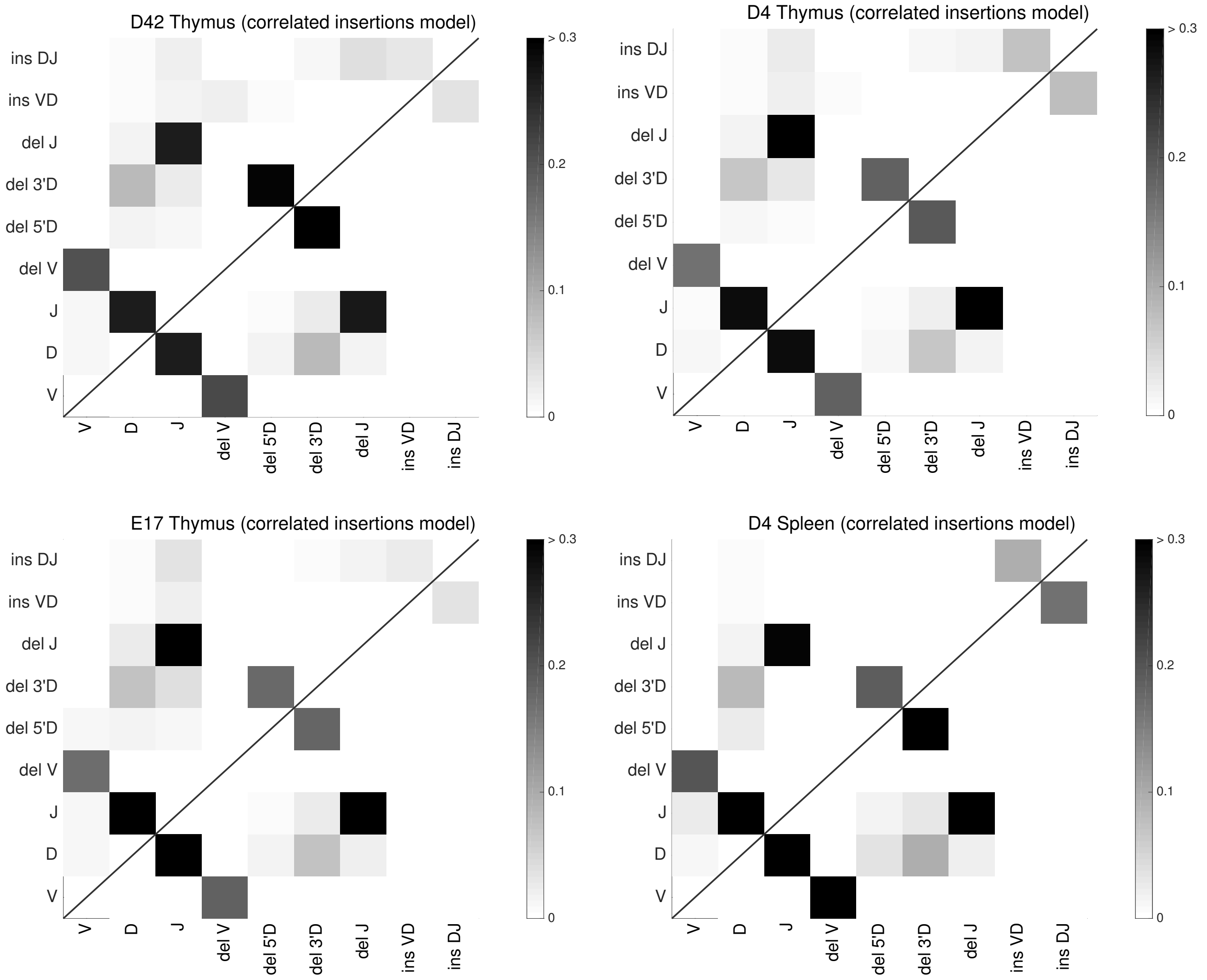
A	1.8	L	3.8
R	-4.5	K	-3.9
N	-3.5	M	1.9
D	-3.5	F	2.8
C	2.5	P	-1.6
Q	-3.5	S	-0.8
E	-3.5	T	-0.7
G	-0.4	W	-0.9
H	-3.2	Y	-1.3
I	4.5	V	4.2

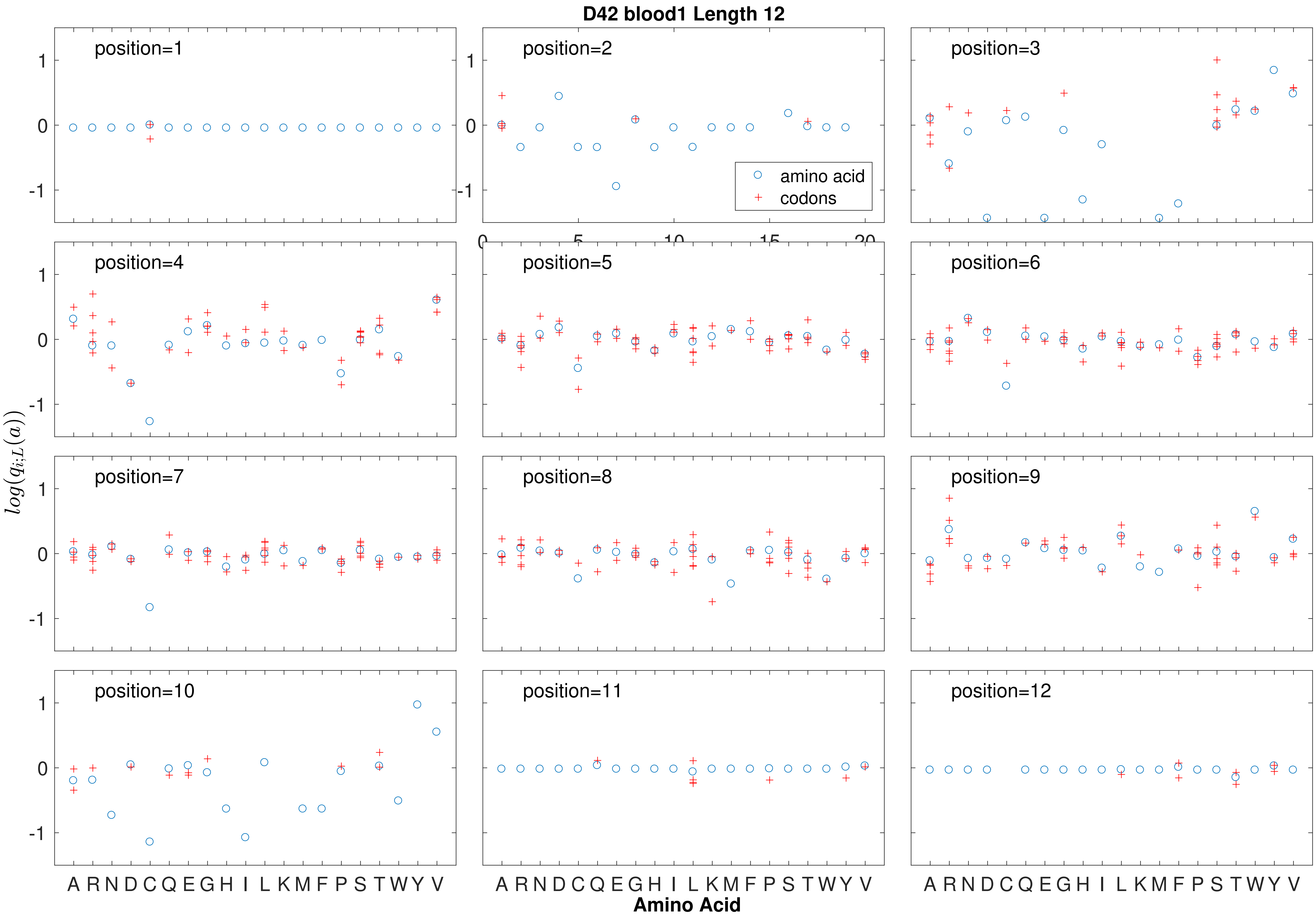
(k) volume

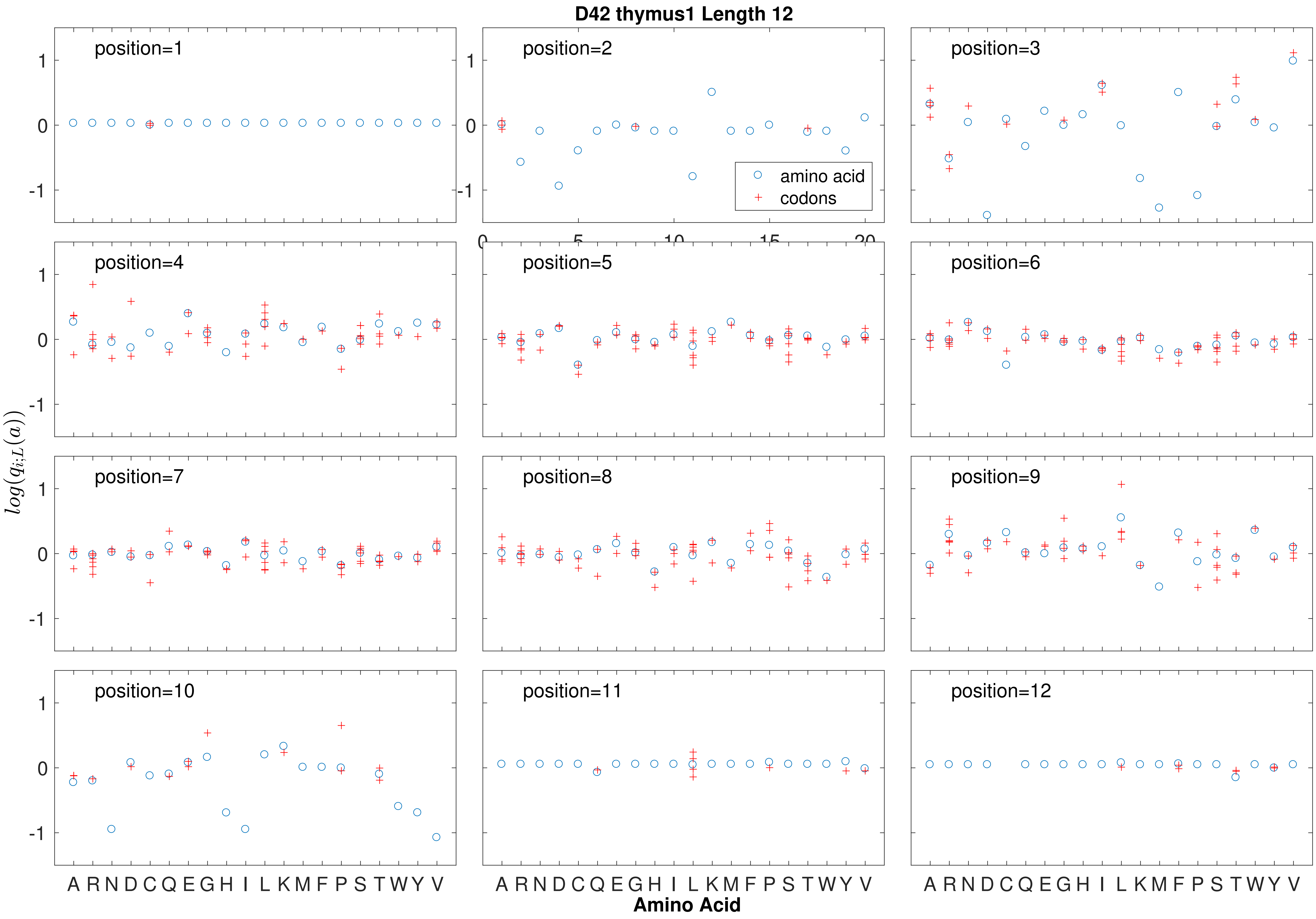
A	67	L	124
R	148	K	135
N	96	M	124
D	91	F	135
C	86	P	90
Q	114	S	73
E	109	T	93
G	48	W	163
H	118	Y	141
I	124	V	105

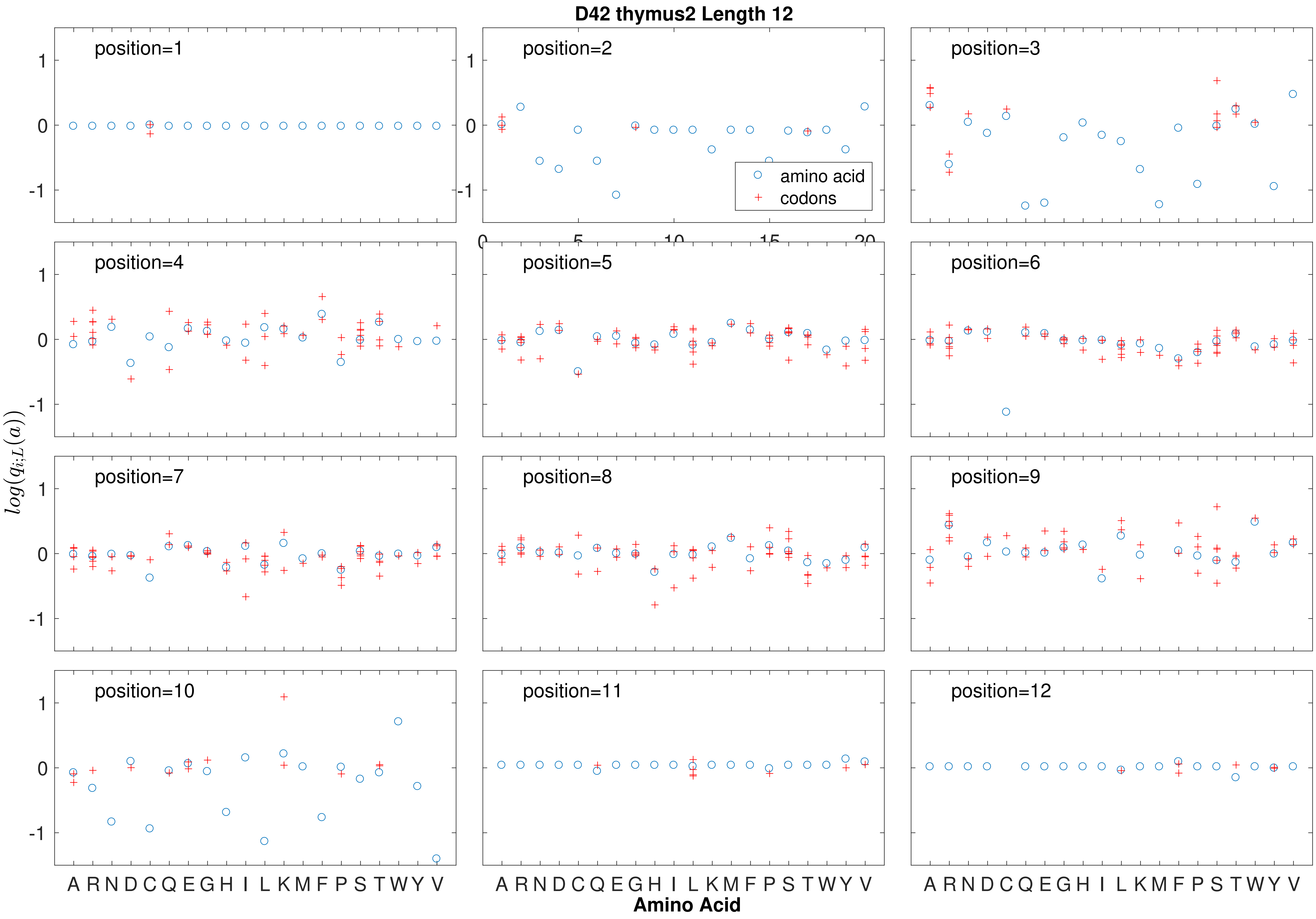

D42 thymus, m2
D42 thymus, m3
D42 blood, m1
D42 blood, m3
E17 thymus, m2
E17 thymus, m3
R^2=0.91708
R^2=0.89359
R^2=0.90053
R^2=0.85216
R^2=0.85795
R^2=0.86614
R^2=0.84885
R^2=0.82017
R^2=0.82417
R^2=0.87057
R^2=0.68993
R^2=0.6793
R^2=0.62963
R^2=0.67005
R^2=0.66127
R^2=0.72061
R^2=0.71257
R^2=0.71479
R^2=0.72595
R^2=0.66467
R^2=0.75338
D42 thymus, m1
D42 thymus, m2
D42 thymus, m3
D42 blood, m1
D42 blood, m3
E17 thymus, m2

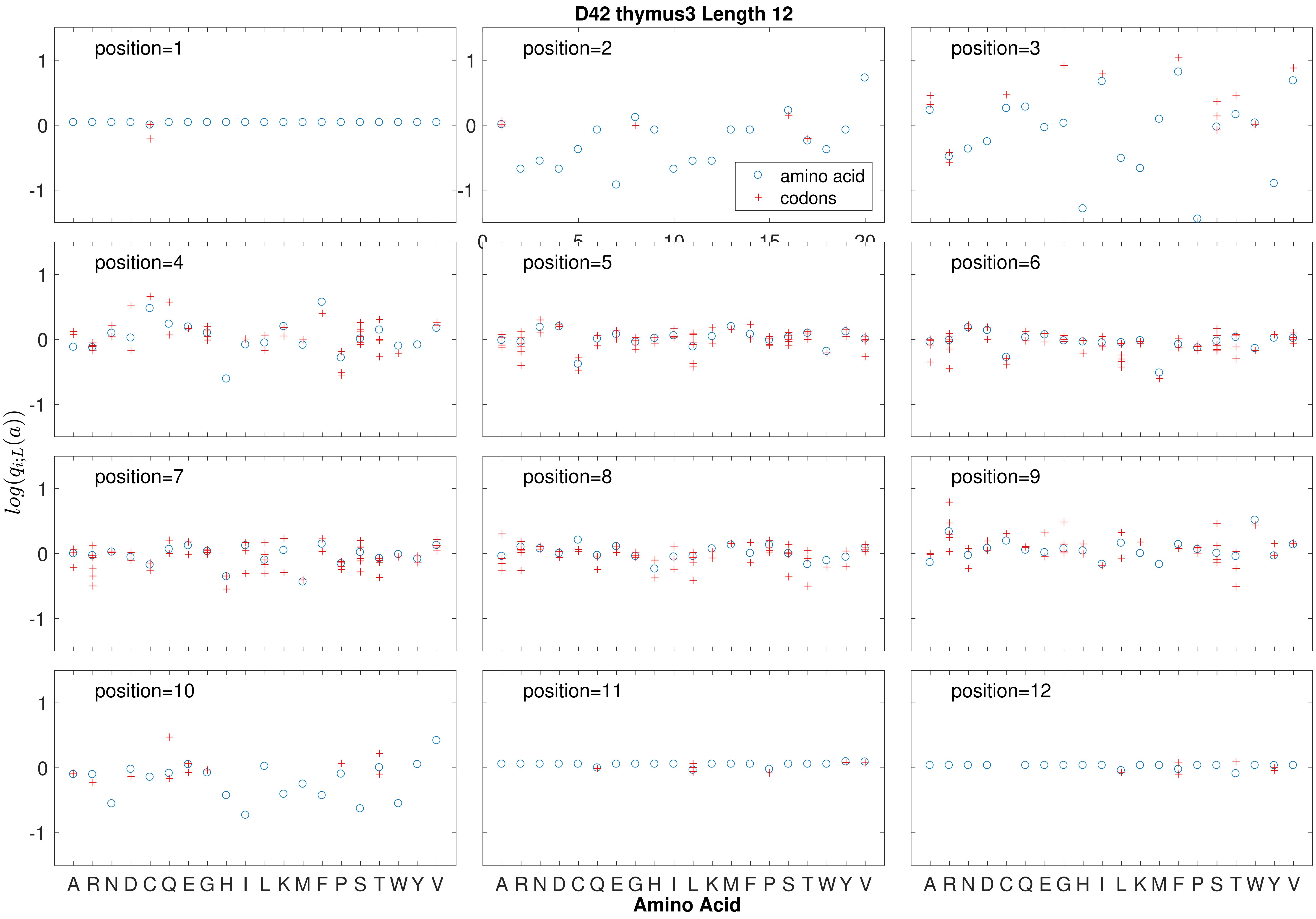

Probabilistic vs Deterministic insertions
D42 thymus model inferred P(insVD)
D42 thymus model inferred P(insDJ)
D42 thymus deterministic insVD usage
D42 thymus deterministic insDJ usage
Probability
0.3
0.25
0.2
0.15
0.1
0.05
0
-0.05
0
5
10
15
Number of insertions at junction
D42 thymus deterministic P(D,J)
D42 thymus model inferred P(D,J)
E17 thymus deterministic P(D,J)
E17 thymus model inferred P(D,J)
TRBD2
TRBD1
TRBJ1-1
TRBJ1-2
TRBJ1-3
TRBJ1-4
TRBJ1-5
TRBJ1-6
TRBJ1-7
TRBJ2-1
TRBJ2-2
TRBJ2-3
TRBJ2-4
TRBJ2-5
TRBJ2-6
TRBJ2-7
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0

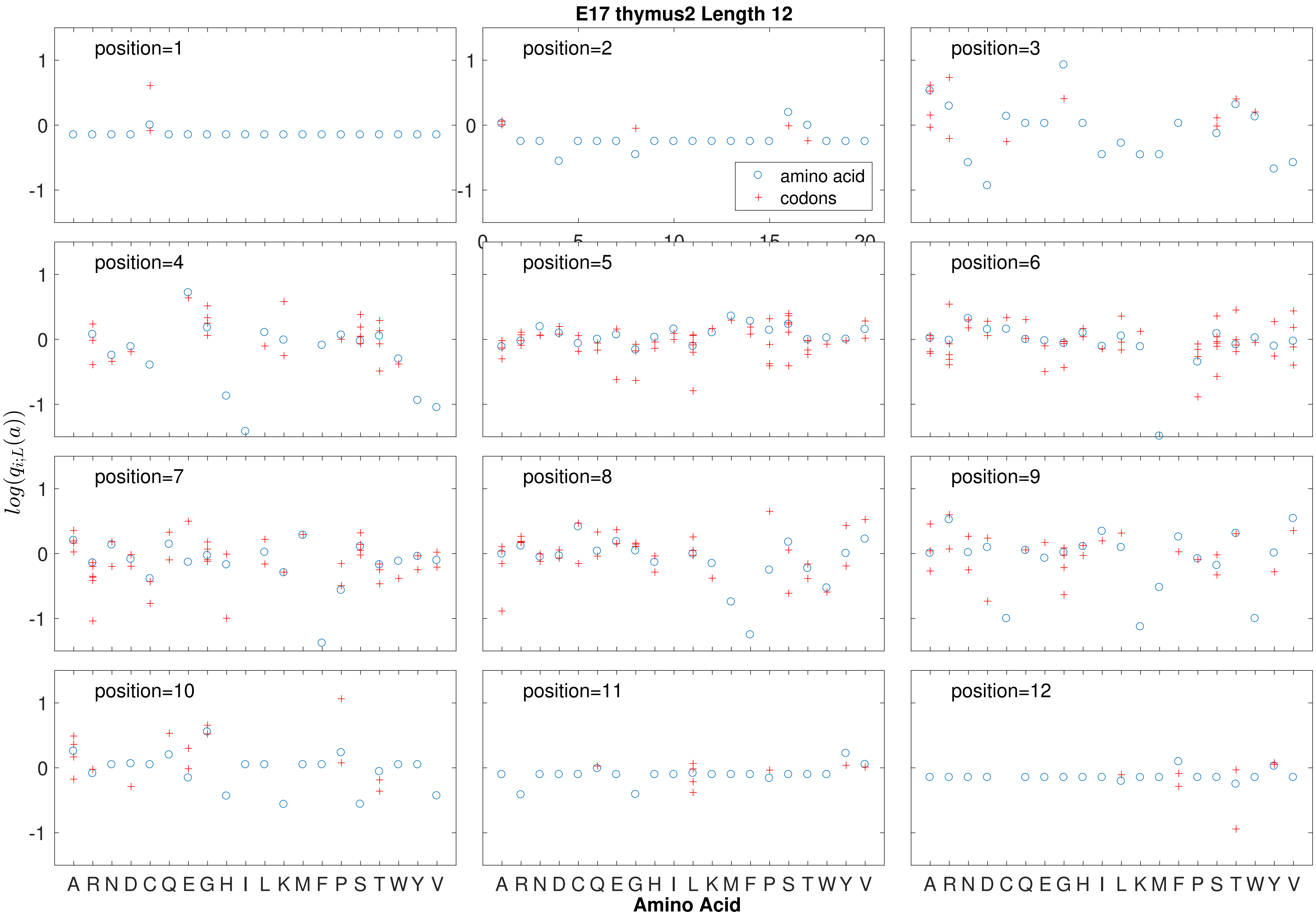

D42 Thymus (correlated insertions model)
D4 Thymus (correlated insertions model)
E17 Thymus (correlated insertions model)
D4 Spleen (correlated insertions model)
ins DJ
ins VD
del J
del 3'D
del 5'D
del V
J
D
V
del V
del 5'D
del 3'D
del J
ins VD
ins DJ
> 0.3
0.2
0.1
0

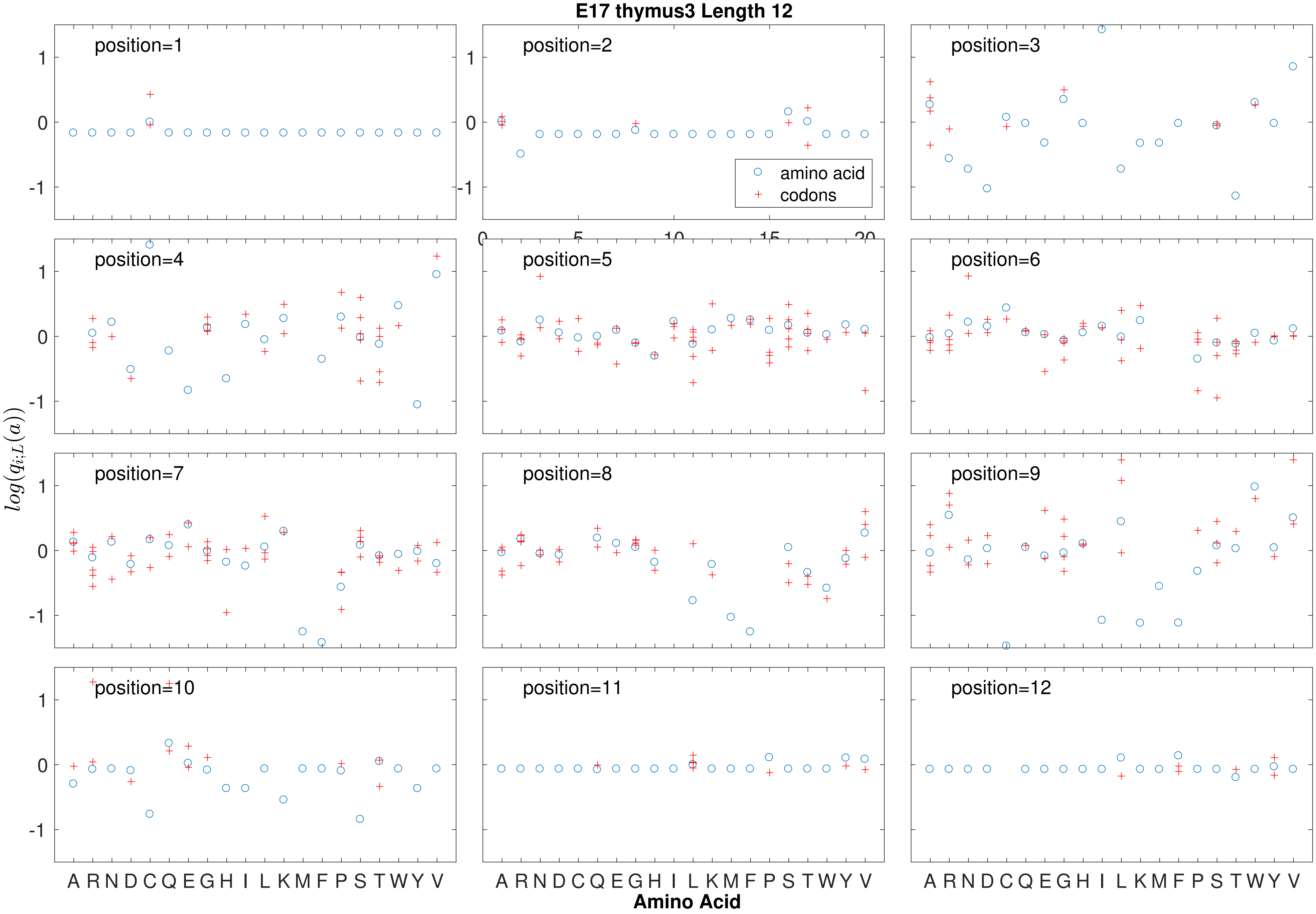

D42 blood1 Length 12
position=1
position=2
position=3
position=4
position=5
position=6
position=7
position=8
position=9
position=10
position=11
position=12
amino acid
codons
1
0
-1
0
5
10
15
20
$log(q_{i;L}(a))$
A R N D C Q E G H I L K M F P S T W Y V
Amino Acid


D42 thymus1 Length 12
position=1
position=2
position=3
position=4
position=5
position=6
position=7
position=8
position=9
position=10
position=11
position=12
amino acid
codons
1
0
-1
0
5
10
15
20
$log(q_{i;L}(a))$
A R N D C Q E G H I L K M F P S T W Y V
Amino Acid


D42 thymus2 Length 12
position=1
position=2
position=3
position=4
position=5
position=6
position=7
position=8
position=9
position=10
position=11
position=12
1
0
-1
0
5
10
15
20
amino acid
codons
$log(q_{i;L}(a))$
A R N D C Q E G H I L K M F P S T W Y V
Amino Acid


D42 thymus3 Length 12
position=1
position=2
position=3
position=4
position=5
position=6
position=7
position=8
position=9
position=10
position=11
position=12
amino acid
codons
1
0
-1
0
5
10
15
20
$log(q_{i;L}(a))$
A R N D C Q E G H I L K M F P S T W Y V
Amino Acid


E17 thymus2 Length 12
position=1
position=2
position=3
position=4
position=5
position=6
position=7
position=8
position=9
position=10
position=11
position=12
amino acid
codons
1
0
-1
0
5
10
15
20
$log(q_{i;L}(a))$
A R N D C Q E G H I L K M F P S T W Y V
Amino Acid


E17 thymus3 Length 12
position=1
position=2
position=3
position=4
position=5
position=6
position=7
position=8
position=9
position=10
position=11
position=12
1
0
-1
○ amino acid
+ codons
0
5
10
15
20
$log(q_{i;L}(a))$
A R N D C Q E G H I L K M F P S T W Y V
Amino Acid