A Model for TCR Gene Segment Use

Aryeh Warmflash and Aaron R. Dinner

*J Immunol* 2006; 177:3857-3864; doi: 10.4049/jimmunol.177.6.3857

http://www.jimmunol.org/content/177/6/3857

**References** This article cites 31 articles, 17 of which you can access for free at:
http://www.jimmunol.org/content/177/6/3857.full#ref-list-1

**Subscriptions** Information about subscribing to *The Journal of Immunology* is online at:
http://jimmunol.org/subscriptions

**Permissions** Submit copyright permission requests at:
http://www.aai.org/ji/copyright.html

**Email Alerts** Receive free email-alerts when new articles cite this article. Sign up at:
http://jimmunol.org/cgi/alerts/etoc
A Model for TCR Gene Segment Use

Aryeh Warmflash* and Aaron R. Dinner2†§¶

The TCR α-chain is assembled by somatic recombination of variable (V) and joining (J) gene segments at the CD4+CD8+ stage of development. In this study, we present the first analytical model for deletional rearrangement and show that it is consistent with almost all available data on VαJα use in mice and humans. A key feature of the model is that both “local” and “express service” models of rearrangement can be obtained by varying a single parameter that describes the number of gene segments accessible at a time. We find that the window is much larger for Vα segments than Jα segments, which reconciles seemingly conflicting data for the former. Implications for the properties of the repertoire as a whole and experiments that seek to probe them are discussed. Special considerations for allelic inclusion are treated in the Appendices. The Journal of Immunology, 2006, 177: 3857–3864.

To provide protection from pathogens, T lymphocytes must react to an enormous variety of foreign molecules. The specificity of a clone for Ag is determined by the TCRs on its surface, which are typically heterodimers of α- and β-chains. The diversity of the TCR repertoire derives in large part from the fact that both α- and β-chains are generated during intrathymic development by somatic recombination of gene segments encoding the constant and variable portions of these molecules. The latter include variable (V), diversity (D), and joining (J) gene segments in the case of β-chains and V and J gene segments in the case of α-chains.

Obtaining a quantitative understanding of how physical factors impact rearrangement of TCR loci is important for assessing the diversity of specificities in the TCR repertoire. Loci with D segments can only rearrange each chromosome once because all but a loose correlation in Vα segments than Jα segments, which reconciles seemingly conflicting data for the former. Implications for the properties of the repertoire as a whole and experiments that seek to probe them are discussed. Special considerations for allelic inclusion are treated in the Appendices.

**Materials and Methods**

We derived an expression for the number of possible ways each VαJα gene configuration can be generated and then used it to estimate probabilities of observing in the periphery: 1) Vα and Jα segments irrespective of with which gene they are paired, 2) VαJα pairs, and 3) Vα/Nα and Iα/Jα pairs in dual TCR cells.

**Counting paths to VαJα gene configurations**

As mentioned in the Introduction, we can treat Vα and Jα rearrangements with the same expressions by encoding the degree to which each set is used sequentially by a parameter (W) that describes the number of accessible gene segments. In other words, we assumed that rearrangements at any time can be made to the W most proximal gene segments remaining; this scheme is discussed in further detail below. A small value of W corresponds to more sequential use, and a large one to more random use.

For clarity, we numbered the Vα segments over 5′ (Iα proximal) and Jα segments (8), but another argued for nonsequential use (“express service”) of Vα segments based on the fact that there is only a loose correlation in Vα/Nα-pairs in cells with productive rearrangements on both chromosomes (6). In the present paper, we develop a mathematical model that reconciles these seemingly conflicting observations. We estimated the Vα and Jα window sizes (Wα and Wα, respectively) from the average separations of segments on different alleles in selected cells and then show that, without further adjustment, these parameters yield good agreement with independent experimental data on Vα and Jα use in the selected TCR repertoire. We find that the window is much larger for Vα segments than Jα segments (37 segments compared with 13 segments, respectively), which accounts for quasiquantitative use of Vα gene segments (8) with only loose correlation between alleles (6). Beyond shedding light on mechanisms of rearrangement, the model is useful for extrapolating statistics on the TCR repertoire from data obtained with limited numbers of Vα-specific mAbs, as demonstrated in the Appendices.

Let us first consider the situation of paired Vα and Jα segments irrespective of which gene they are paired. Then, the number of paths of length k (i.e., those corresponding to k rearrangements) that end in segment n1 prior ones was

\[
\binom{n-1}{k-1} \binom{n-1}{k-1}! = \binom{n-1}{k-1}! (k-1)! (n-k)!\]

(1)

From these unrestricted paths, we subtracted the number of combinations with immediately sequential segments separated by more than W (“forbidden jumps”). For this purpose, we defined the quantity

\[
f(n,k,W,m) = \binom{k}{m} \binom{n-mW-1}{k-1}\]

(2)

where m is the number of forbidden jumps in a path. The first binomial coefficient gives the number of ways of placing the m forbidden jumps.

Copyright © 2006 by The American Association of Immunologists, Inc.

0222-1767/06/002.00
among the \( k \) rearrangements. The second gives the number of ways of choosing the gene segments for the first \( k - 1 \) rearrangements given that certain jumps are required to be of length greater than \( W \). In the advent that \( n - mW < 1 \), we take the second factor on the right side to be 0 here and below.

It is important to note that the quantity \( f(n,k,W,m) \) is not in itself the number of paths with \( m \) forbidden jumps because it also counts those with more than \( m \) forbidden jumps repeatedly. In particular, a path with \( q \geq m \) forbidden jumps contributes \( q/m(q - m)! \) times to \( f(n,k,W,m) \) because there are many ways of choosing the minimum of \( m \) jumps specified to violate the window constraint. To obtain only the number of allowed paths, begin by considering \( f(n,k,W,0) \). By comparing Eqs. 1 and 2, it can be seen that \( f(n,k,W,0) \) is the number of unrestricted paths. It is thus necessary to subtract \( f(n,k,W,1) \) from \( f(n,k,W,0) \). Then paths with one forbidden jump will no longer be counted. However, doing so overcompensates with regard to the paths with more than one forbidden jump because they are counted \( q!/q(q - 1)! = q \) times in \( f(n,k,W,1) \). Adding back \( f(n,k,W,2) \) overcompensates for paths with more than two forbidden jumps in the opposite direction, and so it is necessary to continue subtracting and adding terms with increasing numbers of forbidden jumps until the maximum possible \( (k) \) is reached. Generalizing, in this alternating sum of \( f(n,k,W,m) \) over \( m \), the number of times paths with \( q \geq 1 \) are counted is

\[
\sum_{m=0}^{m=q} (-1)^m \binom{q}{m} = (1 - 1)^q = 0
\]

where the first equality derives from the binomial expansion. In other words, in an alternating sum of \( f(n,k,W,m) \) over \( m \), contributions from paths with forbidden jumps cancel. Thus, the number of paths that satisfy the window constraint is

\[
g(n,k,W) = \sum_{m=0}^{m=k} (-1)^m f(n,k,W,m)
\]

\[
= \sum_{m=0}^{k} (-1)^m \binom{n-mW-1}{k-1}
\]

Eq. 4 is the main result of this paper, and all subsequent expressions are derived from it.

**Probability distributions**

To use Eq. 4 to estimate the probabilities of observing particular \( V_\alpha \) and \( J_\beta \) segments in various contexts, we assume that the likelihoods of making productive rearrangements \( P_s \) and being selected \( P_s \) are constants. Clearly, the latter does in fact vary for specific gene segments, but here we are concerned with the overall statistics of the repertoire: moreover, we effectively average over different \( \beta \)-chains and CDR sequences. Given \( P_s \) and \( P_s \) as well as the window sizes, we combine these variables into case-dependent aggregate probabilities for attempting to generate a particular gene configuration and then being or not being selected \((a \text{ and } b, \text{ respectively})\).

\( V_\alpha \) or \( J_\beta \) probability distributions. Here, we consider the probability of observing a single \( V_\alpha \) or \( J_\beta \) segment (indexed \( n \)), irrespective of with which gene it is paired \([P(n)]\). The likelihood of making a productive rearrangement and then being selected is the product \( P_s \cdot P_s \), and its complement is \( 1 - P_s \cdot P_s \). We normalize these expressions by the window size for \( a \) and \( b \) because there is an equal chance of picking any accessible gene segment:

\[
a = P_s \cdot P_s / W \quad \text{and} \quad b = (1 - P_s \cdot P_s) / W.
\]

To obtain the probability of interest, we perform a sum over the numbers of rearrangements \( k < N_s \), where \( N_s \) is the maximum number of rearrangements, which acts as a surrogate for time in the thymus) weighted according to the number of paths:

\[
P(n) = \frac{a}{Q} \sum_{k=1}^{\infty} \binom{n,k,W}{k} b^{k-1}
\]

where \( Q \) is a normalization factor determined by summing over all possible values of \( n \). The factor \( b^{k-1} \) arises from the fact that to be selected in \( k \) rearrangements, a clone must fail to be selected in \( k - 1 \) previous attempts.

**Evaluation of simplifications**

The model described above makes no reference to underlying molecular details and is consistent with any mechanism that only allows a certain number of sequential gene segments to be accessible at a given time. To make the model mathematically tractable, additional simplifications were made. In particular, we assume that rearrangements can always access the most proximal remaining segments. However, data for the Ig H chain locus suggest that the windows of accessible gene segments are predetermined (12). In this case, the number of segments to which rearrangements can make the model mathematically tractable, additional simplifications were made.

To determine how varying the mechanism for making gene segments accessible impacts the results, we performed stochastic simulations for models with a sliding window of constant size (as in the mathematical
FIGURE 2. Comparison of different mechanisms for how gene segments become accessible. J use for V19, the most J-proximal V<sub>c</sub> gene segment, from the analytical sliding window model (dashed line) and simulation of the fixed block model (solid line). Inset, Mouse V<sub>i</sub> gene use is plotted as a function of index starting at the proximal end with lines as in the main figure. Calculations of the sliding window model used to derive the analytic expressions were performed with W<sub>i</sub> = 37 and W<sub>d</sub> = 13. Simulations of a model in which a new block of segments opens after every other rearrangement were performed with fixed window ends at indices 37, 74, and 104 (assumed to be the last based on Ref. 1).

In this paper, we consider the probability of rearrangements in this model and the effects of the window size on the probability of obtaining V genes. We use the GNU Multiple Precision Arithmetic Library for numeric evaluations of the analytic expressions.

FIGURE 3. Average separation of like segments on different alleles as a function of window size: Σ<sub>n</sub>[n<sub>i</sub> - n<sub>j</sub>]P(n<sub>i</sub>)/W<sub>Wd</sub><sub>Wd</sub>, where P(n<sub>i</sub>)/W is the probability of obtaining V<sub>i</sub>/V<sub>j</sub> or J<sub>i</sub>/J<sub>j</sub> selected cells, computed with Eq. 10 with b as in Eq. 9 and α = α<sub>i</sub> in Eq. B2. Because the combinatorial factors involved are often very large, standard computer arithmetic was not sufficient for the accurate computation of these sums; we used the GNU Multiple Precision Arithmetic Library for this and all other numeric evaluations of the analytic expressions.

In this paper, we consider the probability of rearrangements in this model and the effects of the window size on the probability of obtaining V genes. We use the GNU Multiple Precision Arithmetic Library for numeric evaluations of the analytic expressions.
Choice of parameters

There are five parameters in the model: the $W$ and $P$ window sizes ($W_v$ and $W_J$), the probability that rearrangements are productive ($P_r$), the probability of positive selection following productive rearrangements ($P_s$), and the maximum number of rearrangements per allele ($N_v$). To estimate values for $W_v$ and $W_J$ from a relatively small amount of data, we assume values for $P_r$ and $P_s$. We take $P_r$ to be 0.3; this choice is somewhat less than the one-in-three chance that a rearrangement is in-frame to account for pseudogenes and the possibility of generating stop codons (14). There is little information from experiments to guide the choice of $P_s$. For $P_s$, the probability of selection following productive rearrangement. We expect it to be small because <5% of thymocytes are positively selected (Ref. 15, and references therein). Based on this fact, we take $P_s$ to be 0.03, which yields an overall selection probability of $1 - (1 - 0.3 \times 0.03)^N_v = 4.4\%$ given the choices of $P_r$ and $N_v$ (discussed below). Nearly identical results were obtained with $P_r$ values ranging over an order magnitude.

Data from mice incapable of reintroducing recombination activating gene (RAG) at the double-positive stage suggest that ~35% of the T cell repertoire is formed by primary $\alpha$-chain rearrangements (16) (but see Fig. 5 and associated discussion below). Assuming constant $P_r$ and $P_s$, and that the number of gene segments is not limiting, it is straightforward to show that this percentage is theoretically

$$P_v = \frac{P_r P_s}{\sum_{N_v=1}^{\infty} \left(1 - P_r P_s \right)^{N_v - 1} \times 100\%}$$

(12)

This statistic is not very sensitive to the values of $P_r$ and $P_s$, but it does depend strongly on $N_v$, since it is in the exponent. Reasonable values for $P_v$ in the range 20–34% are obtained with $4 \leq N_v \leq 6$. We take $N_v = 5$, which is consistent with other estimates (15).

Results

In our model for TCR rearrangement, the degree to which a set of gene segments ($V$ or $J$) is used sequentially is encoded in a single parameter that describes the size of the window of accessible gene segments (Fig. 1). Using this idea, we derive in Materials and Methods an analytic expression for the number of possible rearrangement “paths” leading to a selected gene configuration (Eq. 4), as well as probabilities for observing specific $V$, $J$, and $\alpha$ gene segments and their combinations on selected cells (Eqs. 5–10). Here, we estimate the $V$ and $J$ window sizes ($W_v$ and $W_J$, respectively) from the average separations of segments on different alleles in selected cells and then show that, without further adjustment, these parameters yield good agreement with independent experimental data on $V$ and $J$ use in the overall TCR repertoire.

Separation of gene segments on different alleles

To estimate the numbers of $V$ and $J$ segments available for rearrangement at any given time, we calculate the average separation between like types of segments on different alleles in selected cells as a function of window size (Fig. 3). Consistent with intuition, the separation between segments on different chromosomes increases monotonically with $W$. In other words, the correlation between alleles decreases as use becomes less sequential.

For the 61 mouse $J$ genes, the experimentally measured average separation is 7.1 (SD 6.7), and, for the 58 human $V$ genes, it

FIGURE 4. Comparison of calculated (gray) and observed (black) $J$ use for (A) $V_{20S1}$ and (B) $V_{19}$. Probabilities of $V_{20S1}$ use were computed using Eqs. 7 and 8. Only the relative sizes of bars is important; the scale on the vertical axis is arbitrary. Experimental data is from Ref. 8. Inset, $V_{20S1}$ use calculated with the variation of the model in which primary rearrangement can be to either the $V$-proximal end of the locus or $J_{49}$ and the four gene segments downstream of it (13).

FIGURE 5. Comparison of calculated (gray and white) and observed (black) $J$ use in mice incapable of reintroducing RAG at the double-positive stage. We simulated this system by assuming that following each round of rearrangement there is a 70% probability of being incapable of further rearrangements due to insufficient RAG expression; gray contributions are from primary rearrangements, and white contributions are from secondary rearrangements. Experimental data is from Ref. 16.
is 13.8 (SD 9.3) (6). Given these data, we can read the $V_\alpha$ and $J_\alpha$ window sizes off Fig. 3. These averages and SDs correspond to $W_j = 12$ or 15 and $W_\nu = 38$ or 31, respectively (Fig. 3); the fact that the averages and SDs yield relatively close estimates for the window sizes suggests that the large SD values observed are inherent to the rearrangement process rather than due to experimental uncertainty.

From the ranges above, we chose $W_j = 13$ and $W_\nu = 37$ because the mean generally converges more quickly than the second moment of the distribution and the raw data are quite limited. These values confirm the idea that rearrangement of the $J_\alpha$ segments is less random (more sequential) than that of the $V_\alpha$ segments. The value $W_j = 13$ is also consistent with the observation that mice that lack the T early $\alpha$ promoter are unable to use the 10 most proximal $J_\alpha$ segments (5), which suggests a window size of about that number of gene segments.

Due to the limited amount of data, we assume that our values of $W_j$ and $W_\nu$ are common to mice and humans, which appears to be justified given the remarkably good agreement with experiment that we obtain below. Taking the window sizes to be the same in mice and humans also reconciles seemingly conflicting data on $V_\alpha$ segment use (M. Krangel, unpublished observation). Because 37 gene segments represent roughly 65% of the human but only 35% of the murine $V_\alpha$ segments, the former appear to be used randomly while the latter appear to be used sequentially.

$J_\alpha$ distributions for particular $V_\alpha$ genes

We now fix the five parameters in the model at the values estimated above ($P_\nu = 0.3, P_\nu = 0.03, N_\nu = 5, W_j = 13$, and $W_\nu = 37$) and compare calculated $V_\nu J_\alpha$ pair frequencies with measured ones (8, 17–19). We begin by considering the data from Ref. 8 for $J_\alpha$ segments paired with $V_\nu 6$ (located at the J-proximal end of the $V_\alpha$ locus) and $V_\nu 19$ (at the distal end). For these extreme $V_\alpha$ segments, the model and experimentally observed probabilities agree well (Fig. 4). The estimated frequency for $V_\nu 6-J_{\alpha48}$ pairs is very sensitive to whether targeting by the cis-regulatory element at $J_{\alpha49}$ is considered because almost all of these pairs derive from primary rearrangements and $J_{\alpha48}$ falls in the downstream window but not the upstream one. The model predictions compare favorably with the experimental data for $V_\nu 19$ and $V_\nu 20S1$ (located at the proximal end of the locus) studied by Huang and Kanagawa (17) as well (Table I). Again, including the downstream initiation site improves the estimates for pair frequencies involving the proximal $V_\alpha$ segment, $V_\nu 20S1$ (data not shown).

Although the same qualitative trends were observed in Ref. 18, the model cannot reproduce the reported $J_\alpha$ use in detail due to the fact that the distributions are not unimodal. However, it is important to note that these data are putatively for individual members of the $V_\nu 2$ superfamily, which are difficult to conclusively identify, as noted by Huang and Kanagawa (17). Additional data for this superfamily was obtained recently for mice with only a single functional $J_\alpha$ segment and control animals (19). In this case, reasonable agreement with calculated frequencies for the model in which the ends of the windows are restricted to fixed points along the locus (see Fig. 2), but the model exhibits a greater bias toward $J_{\alpha-proximal V_\nu}$ segments than was observed (data not shown).

Regulation of secondary rearrangements

The quasisequential scheme on which our model is based is consistent with the results of knockout experiments directed at elucidating the factors that regulate secondary rearrangements. In mice that are unable to reinduce RAG expression at the double-positive stage because they lack a necessary regulatory element, residual RAG from rearrangement of TCR$\beta$ at the double-negative stage catalyzes only limited rearrangement of the TCR$\alpha$ locus. $J_\alpha$ use is restricted to the 5th (proximal) segments in T cells from these mice (16). To simulate these experiments, we performed simulations of our model in which there was a large probability of losing residual RAG and stopping rearrangement following each round of rearrangement. We found that probabilities between 60 and 80% give good agreement with the experimental data (Fig. 5). Although there are data suggesting that the half-life of RAG is short (~10 min), they were obtained for the destruction of RAG at the G1-S phase of the cell cycle (20); that following the formation of the $\beta$-chain could be slower. In any event, some RAG must be present in these cells to account for the observed $\alpha$-chain rearrangements.

Discussion

In this study, we present a model for TCR gene segment use in which the degree that rearrangement is sequential is determined by the size of a window of accessible gene segments. The model is based only on this notion and the fact that intervening gene segments are deleted when $V_\alpha$ and $J_\alpha$ are brought together. We deduced the window sizes from data on the correlations between the two $V_\nu$ or $J_\alpha$ genes used in selected cells, and showed that without further adjustment, these parameters yield statistics in good agreement between the model and almost all available data on TCR$\alpha$ gene segment use. Although the model cannot reveal the detailed molecular mechanism, it strongly indicates that $V_\alpha$ use is quasisequential, which reconciles seemingly conflicting data for mice and humans.

The model is exactly solvable and provides the first expressions that can be used to extract information directly from data on gene segment use in lymphocytes. Previous theoretical studies were limited to simulations (numeric “experiments”) (21, 22) and focused on the use of L chain segments in B cells, in particular $J_\alpha$. There are only four functional such gene segments, which are used with a slight bias toward the two more proximal segments. By either assigning probabilities to each of the four (21) or varying the ratio of the likelihoods of choosing each gene segment and the one immediately upstream (22), it was found that quasisequential use best explains the available $J_\alpha$ data as well.

One consequence of the mechanism identified is that the most distal $V_\nu$ segments are incapable of pairing with the most proximal $J_\alpha$ segments and vice versa, consistent with Refs. 23 and 24. Although we calculate that 70% of mouse $V_\nu J_\alpha$ and 58% of human $V_\nu J_\alpha$ pairs are expressed at frequencies within an order of magnitude of the uniform distribution (with the differences arising from the fact that the $V_\nu$ locus in humans is roughly twice the size of that in mice), ~4% of $V_\nu J_\alpha$ pairs are incapable of forming in both species. Thus, the use of specific $V_\nu J_\alpha$ pairings can vary dramatically depending on their chromosome locations, and care must be used in extrapolating statistics from experiments specific to particular TCR, as discussed in Appendix A.

In general, as rearrangement becomes increasingly sequential, there is a tradeoff between diversity and cell conservation in the thymus. Consequently, it is natural to ask whether the diversity of the TCR repertoire is limited significantly by the quasisequential nature of rearrangement. If rearrangement were totally uniform, every $V_\nu J_\alpha$ pair would be present in the repertoire with probability $1/N_\nu N_\alpha$, where $N_\nu$ and $N_\alpha$ are the total numbers of $V_\nu$ and $J_\alpha$ genes. To compare different mechanisms, we used our model to calculate the SD of pairing frequencies for all possible values of $W_\nu$ and $W_j$ (Fig. 6). A lower SD corresponds to a more uniform distribution. Interestingly, the window sizes estimated from the experimental data (Fig. 3) fall close to the bottom of the very shallow basin around the minimum (Fig. 6). Large SDs are obtained for either a perfectly sequential model ($W_\nu = 1$ and $W_j =
1) or a random deletional one ($W_\gamma = 104$ and $W_\delta = 61$) because gene segments from the proximal and distal ends, respectively, tend to be used disproportionately. These calculations lead us to speculate that the quasisequential rearrangement mechanism that we identified evolved to maximize the diversity of the repertoire.

**Appendix A: Corrections for phenotypic allelic inclusion frequencies**

Roughly one-quarter of mature $\alpha \beta$ T lymphocytes have productive rearrangements at both their TCR $\alpha$-chain gene loci (genotypic allelic inclusion) (10), but not all of these cells express twoAg receptors on their surfaces (25–29). Dual TCR cells can thus serve as windows to the mechanisms that regulate T cell surface expression in general and, in turn, how events in the lives of T cells can lead to variations in molecular populations underlying autoimmunity (11). To determine the factors that influence phenotypic allelic exclusion, it is important to be able to quantitate its extent accurately. Unfortunately, only a few mAbs for specific $V_\alpha$ protein segments are available, so it is necessary to extrapolate from limited FACS data to estimate the total fraction of mature T cells that express two TCR on their surface. In this Appendix, we use our model for TCR $V_\alpha J_\alpha$ (see below) to improve means for interpreting such experiments.

**Correction for overcounting**

To determine the extent of phenotypic allelic inclusion in a population of T lymphocytes, FACS is used to count the number of cells that bind reagents specific for two different $V_\alpha$ protein segments. Typically, these data are then used to calculate the fraction

$$f_0 = \frac{N_j/N}{(N/N)(N/N)}$$  \hspace{1cm} \text{(A1)}

where $N_j$ is the number of cells that are $V_\alpha i^+ j^+$, $N_i$ is the number of cells that are $V_\alpha i^+ j^+$, and $N$ is the total number of cells. Although often quoted as such (26–29), $f_0$ is not the frequency of phenotypic allelic inclusion in the total population ($F$). Rather, it is an approximation for the fraction of $V_\alpha i^+ j^+$ cells which display two receptors at the cell surface (in other words, $f_0 \approx f_j$ and $f_f$, respectively). To relate $f_0$ to $F$, it is necessary to avoid double-counting cells that express two different $V_\alpha$ protein segments.

The experiments sort cells only according to $V_\alpha$, which effectively averages over $J_\alpha$, CDR, and the $\beta$-chain. Consequently, it is not unreasonable to assume $f_{ij} \approx f_j \approx f_0 \approx f$ is essentially the same for all $V_\alpha$ (discussed below). Then, denoting the number of $V_\alpha i^+ j^+$ cells that express two TCR by $d$, we can write for each $V_\alpha$ protein segment an equation of the form $f N_i = d$. Summing over segments,

$$f \sum_{i} N_i = \sum_{i} d_i$$  \hspace{1cm} \text{(A2)}

The sums above count dual TCR cells with different $V_\alpha$ twice since $N_i f_j$ contributes to $N_j f_i$, $d_i$, and $d_i$. Denoting the total number of dual TCR cells by $d$ and the number of such cells with the same $V_\alpha$ on both alleles by $d_{i\alpha}$,

$$\sum_{i} d_i = 2d - d_{i\alpha} \quad \text{and} \quad \sum_{i} N_i = N + d - d_{i\alpha}$$  \hspace{1cm} \text{(A3)}

Substituting the expressions in Eq. A3 into Eq. A2 and solving for $F = d/N$, we find

$$F = \frac{f}{2 - f} \left[ 1 + d_{i\alpha} \left( \frac{1}{N} - 1 \right) \right] \approx \frac{f}{2 - f}$$ \hspace{2cm} \text{(A4)}

The second (approximate) equality follows from the fact that $d_{i\alpha}$ is expected to be much smaller than $N$ (using the model described in the main text $d_{i\alpha}/N \approx 0.003$), which allows neglect of the term in square brackets. Eq. A4 thus provides a practical means of estimating the fraction of mature T cells that express two TCR on their surfaces from a measurable quantity ($f \approx f_0$, but see below).

**Correction for biases in gene rearrangement**

$V_\alpha$ gene segments paired in dual TCR cells are weakly correlated (6). As discussed in the main text, such biases in the repertoire come from quasisequential use of the two sets of $V_\alpha$ gene segments at the same rate. In particular, application of Eq. 10 shows that the fraction of dual TCR cells observed depends on the separation of the two $V$ segments detected (Fig. A1). In many cases, $f_0$ as computed from Eq. A1 will be a poor approximation for $f$. 

![Figure A1](http://www.jimmunol.org/) 

**FIGURE A1.** Average $f$ as computed by Eq. A1 from simulation data as a function of the distance between the $V_\alpha$ genes studied for $N_i = 5$ (solid line) and $N_i = 10$ (dashed line). The average is performed over all $V_\alpha$ pairs separated by the given distance and is weighted by the numbers of passing single TCR cells which express the $V_\alpha$ genes. In other words, ($f_{i\alpha}(k)$) = \sum_{i\alpha}\sum_{i\alpha}f_{i\alpha}N_i/N. \text{Inset,} \text{ Fraction of cells with } V_\alpha \text{ which are dual TCR cells as a function of } i \text{ calculated from simulation data (} f_0 = d/N, \text{ for } N_i = 5 \text{ (solid line)} \text{ and } N_i = 10 \text{ (dashed line). The same parameters as in the main text were used.}
Here, we show how our model of TCR $V_j\alpha$ rearrangement can be used to mitigate the bias that quasisequential deletional rearrangement introduces to interpretation of measured $f_i$. To this end, we use the model with the parameters given in the main text to calculate the fraction of dual TCR cells as a function of $V_j\alpha$ gene segment position and the calculated value of $f_i$:

$$f_i^{\text{calc}} = \sum_j P(i,j)P(i) + f_i^{\text{calc}} = P(i,j)P(i)P(j)$$

(A5)

where $P(i,j)$ is given by Eq. 10 and $P(i)$ is computed as explained in Appendix B. Note that $P(i)$ includes both single and dual TCR cells. We then scale the measured $f_i$ by the ratio of $f_i^{\text{calc}}$ and its average value in the model ($\langle f_i^{\text{calc}} \rangle$, where the average is weighted by the number of $V_j\alpha^+$ cells) to obtain a corrected fraction for use in Eq. A4:

$$f_i = f_i^{\text{calc}} f_i^{\text{calc}} \langle f_i^{\text{calc}} \rangle.$$  

(A6)

Basing this correction on our model for TCR $V_j\alpha$ rearrangement implicitly assumes that the extent of phenotypic allelic inclusion is directly proportional to the extent of genotypic allelic inclusion. Although this cannot be wholly the case, it is not an unreasonable approximation given, as discussed above, the effective averaging over $J_{\alpha}$, CDR, and the $\beta$-chain.

One feature of Eq. A6 that might be confusing to the reader is that $f_i$ is a property of one gene segment, but $f_i$ is clearly a property of two. However, $f_i$ as defined by Eq. A1 (26–29) is an estimate for $f_i = f_i$. Eq. A6 is used to correct for cases when this is a poor estimate due to the separation between gene segments used in the experiment. It is worth noting that the error in $f_i$ could thus be reduced by instead staining for one $V_j\alpha$ protein segment and CDS as in Ref. 25. Because the ratio of total TCR to CDS is roughly constant, cells with a low ratio of CDS to the $V_j\alpha$ protein segment studied express only that $V_j\alpha$, while those with a high ratio express another $V_j\alpha$ as well.

The discussion immediately above leads to a related point. Given the dependence of $f_i$ on the separation in position of gene segments $V_{ji}$ and $V_{ij}$, it is natural to wonder whether the assumption above that $f_i = f_i$ is essentially the same for all $V_j\alpha$ protein segments is appropriate. In deriving Eq. A4, the $f$ used represents the actual fraction of dual TCR cells as a function of $V_j\alpha$ gene segment position, not the fraction estimated from a specific pair of $V_{ji}$ and $V_{ij}$. Due to averaging over $V_{ji}$, $f_i$ calculated with Eq. A5 is much flatter than $P(i,j)$ calculated with Eq. 10 (Fig. A1, inset). The assumption that $f_i$ is essentially the same for all $V_j\alpha$ is thus reasonable. Moreover, it improves as the number of rearrangements becomes larger, although a kink persists at the transition from primary to secondary rearrangement.

**Total fraction of phenotypic dual TCR cells**

We use the corrections in Eqs. A4 and A6 to re-evaluate previously published data. The reagents used were for the $V_j\alpha$, $V_j\alpha$, and $V_j\alpha$ families. In these cases, $P(i,j)$ is the probability that any gene segments from family $i$ is paired with any member from family $j$ and $P(i)$ is the probability that any member of gene family $i$ is used. Generally speaking, these gene families are distributed throughout the chromosome; the $V_j\alpha$ family is slightly biased toward the 3' end whereas the $V_j\alpha$ family is slightly biased toward the 5' end. The $V_j\alpha$ family is nearly evenly distributed. Thus, it is to be expected that $V_j\alpha$ and $V_j\alpha$ experiments need little correction, while $V_j\alpha$ experiments significantly underestimate the fraction of dual TCR cells. In agreement with this expectation, the calculated correction factors for experiments with $V_j\alpha$, $V_j\alpha$, and $V_j\alpha$ are 1.02, 1.03, and 1.86, respectively. Because the experiments overestimate the fraction of dual TCR cells by nearly a factor of 2 due to overcounting, the data from $V_j\alpha$ experiments remains nearly the same after applying both corrections. However, since no genotypic bias was introduced in $V_j\alpha$ and $V_j\alpha$ experiments, there was significant overestimation of the fraction of dual TCR cells (Table A1). The separation of $V_j\alpha$ and $V_j\alpha$ and the consequent biases in their statistics are likely to account for the fact that the same authors generally found a higher fraction of dual TCR cells in $V_j\alpha$ and $V_j\alpha$ experiments with the exception of two aberrant data points (18–21% in Ref. 28; 31.0% in Ref. 27, the latter of which may have been due to gaiting on CD8). Thus, the corrected statistics suggest that the overall rate of phenotypic allelic inclusion is 2–11%, which supports the idea that posttranslational control mechanisms regulate TCR surface expression (see Refs. 11, 30, and 31 for discussion).

**Appendix B: Normalization for dual TCR cells**

We explicitly considered only one allele in deriving Eq. 6. However, it is important to treat both chromosomes to normalize the probabilities in Eq. A5 consistently. Specifically, in Eq. A5, for the probability of a cell expressing a receptor which uses the $m\alpha$ gene segment regardless of whether it is a single or dual TCR cell, we have

$$P(m) = \frac{\alpha}{Q} \sum_{k=1}^{m} [g(m,k,W)]^{b-1}$$

$$+ \frac{\alpha_2}{Q} \sum_{k=1}^{m} 2g(m,k,W)[W - g(m,k,W)]^{b-1}$$

(B1)

The first term represents paths in which both alleles have rearranged to the $m\alpha$ gene segment and at least one of them is in-frame. The second term represents paths in which only one of the two alleles has rearranged to the $m\alpha$ gene segment, in which case that allele must be in-frame and the other can be either in- or out-of-frame. The aggregate probabilities are thus

$$a_1 = [P_1P_2 + 2(1 - P_2)P_1]/W^2$$

and

$$a_2 = [P_1P_2 + (1 - P_1)P_2]/W^2$$

(B2)

with $b$ given by Eq. 9. Finally, we set the normalization $Q$ such that

$$\sum_i P(i) - \sum_{j<i} P(i,j) = 1.$$  

(B3)

---

<table>
<thead>
<tr>
<th>Previous Result (%)</th>
<th>Corrected Result (%)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_j\alpha$, $V_j\alpha$</td>
<td>10.8</td>
<td>5.8</td>
</tr>
<tr>
<td>$V_j\alpha$, $V_j\alpha$, $V_j\alpha$</td>
<td>14–17, 16</td>
<td>7.6–9.5, 8.9</td>
</tr>
<tr>
<td>$V_j\alpha$, $V_j\alpha$, $V_j\alpha$</td>
<td>3.7, 3.9</td>
<td>1.9, 2.0</td>
</tr>
<tr>
<td>$V_j\alpha$, $V_j\alpha$, $V_j\alpha$</td>
<td>6.6, 8.7, 11.0, 13.6</td>
<td>3.5, 5.5, 5.9, 7.4</td>
</tr>
<tr>
<td>$V_j\alpha$, $V_j\alpha$, $V_j\alpha$</td>
<td>7.1, 11, 18–21</td>
<td>7.0–11.3, 20.1–24.2</td>
</tr>
<tr>
<td>$V_j\alpha$, $V_j\alpha$, $V_j\alpha$</td>
<td>6.8, 8.0, 31.0</td>
<td>6.8, 8.0, 41.0</td>
</tr>
<tr>
<td>$V_j\alpha$, $V_j\alpha$, $V_j\alpha$</td>
<td>16.0</td>
<td>9.0</td>
</tr>
</tbody>
</table>

*Results separated by a comma are independent experiments performed by the same author. Ranges of values result from different methods of background subtraction performed by Heath et al. (28). Data for $V_j\alpha$ from Ref. 26 were not considered because of difficulties involved in determining the total number of $V_j\alpha^+$ cells as noted by the author.*
Acknowledgments
We thank Martin Weigert for helpful discussions and critical reading of the manuscript and Barry Steckman for providing data on V_{\gamma2} use in advance of publication.

Disclosures
The authors have no financial conflict of interest.

References