Supporting Methods

Following Plotkin et al. (1) we defined the changeability of complimentarity determining region (CDR) or framework (FW) [C(g)] as the summed changeability of their codons [C(i)]. We used the following statistical test to determine whether the overall changeability of CDR and FW regions was significantly elevated or depressed compared with the rest of the human genome, while controlling for the length and amino acid composition. We indexed the 45 viable (i.e. not "stop") codon nodes in an arbitrary order i = 1 …45. As explained in Fig. 1 in greater detail, because all codons with *c/t* on the third position are synonymous in the amino acids they encode, we have 45 and not 61 nodes. We used the notation *aa(i)* to denote the amino acid encoded by codon i. We then further defined N_i as the number of occurences of codon i in the entire human genome and n_i as the same in the gene region to be compared (CDR or FW in this case). Similarly M_α and m_α denoted the number of occurences of amino acid α in the entire genome and in the gene region under study, respectively.

Thus, the changeability of a gene region (G) is defined as:

1. $$C(G) = \sum_{i=1}^{45} n_i \times C(i)$$

For each amino acid α, we defined its expected changeability and its variance in changeability, given the codon usage in the entire genome, by the equations:

2. $$E[C(\alpha)] = \sum_{i \text{ such that } aa(i)=\alpha} \left(C(i) \times N_i / M_{aa(i)} \right)$$

3. $$V[C(\alpha)] = \sum_{i \text{ such that } aa(i)=\alpha} \left(C(i)^2 \times N_i / M_{aa(i)} \right) - E[C(\alpha)]^2$$

Based on this, we defined the expected changeability of (G) and its variance by the equations:

4. $$E[C(G)] = \sum_{\alpha=1}^{20} E[C(\alpha)] \times m_{aa(\alpha)}$$

5. $$V[C(G)] = \sum_{\alpha=1}^{20} V[C(\alpha)] \times m_{aa(\alpha)}$$

We calculated the significance of the difference between the expected and observed levels of changeability of CDR and FW. A Kolmagorov-Smirnov test for normality of the product of the distribution of codons and their changeability in the human genome was not significant ($P > 0.1$). We therefore could assume a normal distribution of this trait and used the following equation to determine whether differences between expected and observed changeability were significant ($\alpha = 0.05$):

6. $$p = \frac{1}{2} \times \left(1 \mp erf\left(\frac{C(G) - E[(C(G)]}{\sqrt{2 \times V[C(G)]}}\right)\right)$$

7. $$erf(x) = \frac{2}{\pi} \times \int_0^{\infty} e^{-t^2} \times dt$$

Supporting data sets

There are three datasets in the supporting information. Data set 1 contains the sequences analyzed in the paper (one per subgroup of human λ, κ and heavy chains and also CD8) divided into CDR and FW. Data set 2 contains a sequence for every type of human κ V gene in the IMGT database (http://imgt.cines.fr), whereas data set 3 contains a sequence for every type of human λ V gene in the IMGT database. Both are in FASTA format. The results of the analysis of these larger groups is in tables 5 and 6 in the supplemental information.

1. Plotkin, J. B., Dushoff, J. & Fraser, H. B. (2004) Nature 428, 942-5.