Supplemental material for:

V(DD)J recombination is an important and evolutionary conserved mechanism for generating antibodies with unusually long CDR3s

Yana Safonova and Pavel A. Pevzner*

Computer Science and Engineering Department, University of California San Diego, La Jolla, USA

* Corresponding author: ppevzner@ucsd.edu

This document includes Supplemental Methods, Supplemental References, Supplemental Tables S1–4, and Supplemental Figures S1–7.

Supplemental Methods

Supplemental Method: Analyzing correlation between the usage of D genes and the likelihoods of heptamers/nonamers in their RSSs

The correlation between the usage of human D genes and the RSS probabilities of their canonical heptamers/nonamers is far from being perfect (Supplemental Fig. S3). For example, D genes IGHD1-7 (red dot in Supplemental Fig. S3) and IGHD1-20 (green dot in Supplemental Fig. S3) with high heptamers/nonamers probabilities rarely contribute to CDR3s across multiple immunosequencing datasets (Briney et al., 2012; Elhanati et al., 2015; Safonova and Pevzner, 2019a; Bhardwaj et al., 2020). This may reflect the limitations of the profile model of RSSs that does not adequately reflect the "strengths" of various RSSs (Lee et al., 2003).

We thus do not expect to see a good correlation in a more complex case when we compare the tandem usage and the "strength" of heptamers/nonamers in two RSSs contributing to tandem fusions of genes D and D*. Analyzing this correlation is further complicated by the observation that tandem coefficients are highly variable while probabilities of cryptic nonamers are less variable (form a bimodal distribution) (Supplemental Fig. S4).

Supplemental Method: Challenge of analyzing tandem fusions in mouse Rep-Seq datasets

The only non-human species with publicly available (not antigen-stimulated) Rep-Seq datasets and known IGHD genes is mouse. However, analyzing the correspondence between RSSs and tandem CDR3s for mouse D genes is a more challenging problem than for human D genes since (i) many mouse D genes listed in the IMGT database do not occur in the mouse immunoglobulin locus, pointing to potential assembly problems, (ii) available mouse Rep-Seq dataset may have been generated from mouse strains with a different immunoglobulin locus, (iii) many mouse D genes are duplicated, (iv) since many mouse D genes are very short, IgScout cannot reliably identify tandem fusions formed by these genes, and (v) mouse D genes have extremely non-uniform usage, with a single D gene contributing to over 80% of all CDR3s. Indeed, although the IMGT database includes 38 mouse IGHD genes, only 26 of them occur in the immunoglobulin locus in the reference mouse genome (version GRCm38.p6). 5 out of these 26 D genes have multiple occurrences in the genome (D2-5 (2 occurrences), 4-1 (3), 5-1 (2), 5-2 (5), and 6-1 (3)), making it difficult to figure out which specific RSS was used for generating tandem fusions and thus compute p-values. Only four mouse genes have usage exceeding 0.5% even after duplicated mouse D genes are counted as a single gene: 1-1 (usage 80.8%), 2-3 (10.4%), 3-2 (4.7%), and 4-1 (0.6%).

Supplemental references

Bhardwaj V, Franceschetti M, Rao R, Pevzner PA, Safonova Y. 2020. Automated analysis of immunosequencing datasets reveals novel immunoglobulin D genes across diverse species. *PLOS Comp Bio* **16(4):** e1007837.

Briney BS, Willis JR, Hicar MD, Thomas JW, Crowe JE. 2012. Frequency and genetic characterization of V(DD)J recombinants in the human peripheral blood antibody repertoire. *Immunology* **137:** 56–64.

Elhanati Y, Sethna Z, Marcou Q, Callan CG Jr, Mora T, Walczak AM. 2015. Inferring processes underlying B-cell repertoire diversity. *Philos Trans R Soc Lond B Biol Sci* **370:** 1676.

Safonova Y, Pevzner PA. 2019a. De novo Inference of Diversity Genes and Analysis of Non-canonical V(DD)J Recombination in Immunoglobulins. *Front Immunol* **10:** 987.

Supplemental Table S1. Immunosequencing datasets.

Dataset	NCBI accession numbers	Description	# Rep-Seq datasets
ALLERGY	PRJEB18926	B cells from PBMC and bone marrow of donors with allergy	24
INTESTINAL	PRJNA355402	Memory and plasma B cells from intestinal tissues	32
MOUSE	PRJEB18631	pre-B cells, naive B cells, plasma cells from healthy and vaccinated mice	71

Supplemental Table S2. Twelve cryptic nonamers of human IGHD genes. Conserved positions in nonamers are shown by upper-case letters. Positions coinciding with the consensus sequence of the canonical nonamers are bolded and underlined.

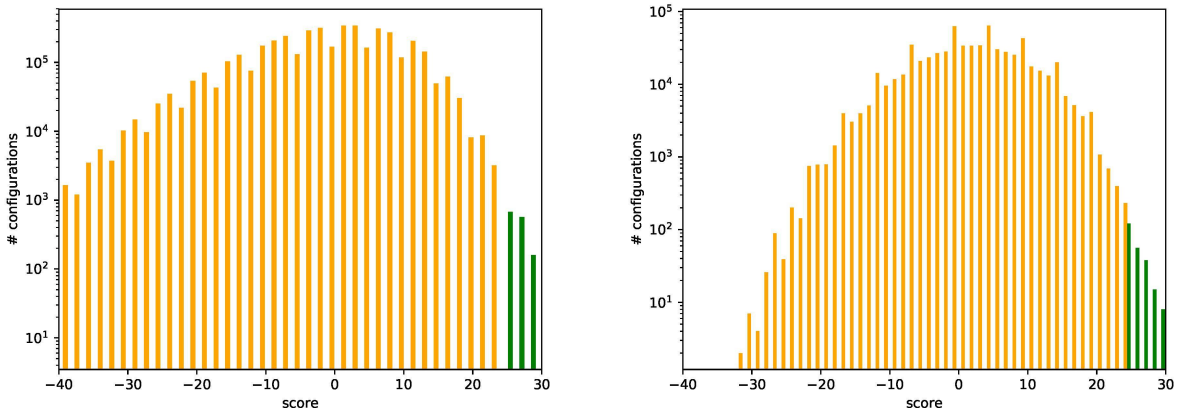
Index	D gene	Orientation	Cryptic nonamer	# turns	Spacer (bp)	Fusions
1	IGHD3-16	left	**GGTTT**ccCc	3	34	(*, D3-16)
2	IGHD3-10	left	**GGTTT**ccCc	3	34	(*, D3-10)
3	IGHD3-9	left	**GGTTT**ccCc	3	34	(*, D3-9)
4	IGHD3-22	left	**GGTTT**ccCc	3	34	(*, D3-22)
5	IGHD3-3	left	**GGTTT**ccCc	3	34	(*, D3-3)
6	IGHD6-19	right	gTCcC**AA**GT	2	23	(D6-19, *)
7	IGHD2-8	right	a**C**TgC**AA**A**C**	2	24	(D2-8, *)
8	IGHD2-21	right	a**C**TgC**AA**A**C**	2	24	(D2-21, *)
9	IGHD2-15	right	a**C**TgC**AA**A**C**	2	24	(D2-15, *)
10	IGHD2-2	right	cTGt**AAAC**G	2	25	(D2-2, *)
11	IGHD6-13	right	aTTcC**AA**GT	2	23	(D6-13, *)
12	IGHD5-12	right	aGGcC**AA**GT	2	21	(D5-12, *)

Supplemental Table S3. Eight optimal configurations explaining the fusion graph in Figure 2. Red (blue) cells highlight nonamers 2-2R, 2-8R, 5-12R, 6-13R, 2-15R, 6-19R (3-3L, 3-9L, 3-16L) if they contribute 1s to the corresponding configuration.

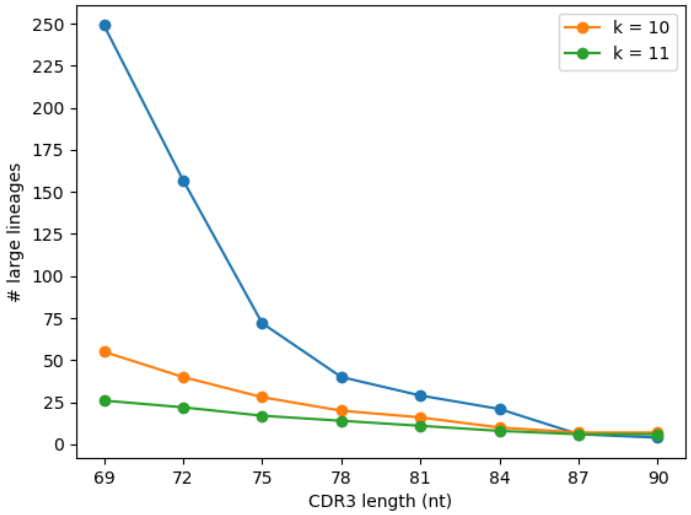
2-2R	3-3R	3-3L	2-8R	2-8L	3-9R	3-9L	3-10R	3-10L	5-12R	5-12L	6-13R	6-13L	2-15R	2-15L	3-16R	3-16L	6-19R	6-19L	2-21R	2-21L	3-22L
1	0	1	1	0	0	1	0	0	1	0	1	0	1	0	0	1	1	0	0	0	0
1	0	1	1	0	0	1	0	0	1	0	1	0	1	0	1	1	1	0	0	0	0
1	0	1	1	0	0	1	0	0	1	1	1	0	1	0	0	1	1	0	0	0	0
1	0	1	1	0	0	1	0	0	1	1	1	0	1	0	1	1	1	0	0	0	0
1	0	0	1	0	0	1	0	0	1	0	1	0	1	0	0	1	1	0	0	0	0
1	0	0	1	0	0	1	0	0	1	0	1	0	1	0	1	1	1	0	0	0	0
1	0	0	1	0	0	1	0	0	1	1	1	0	1	0	0	1	1	0	0	0	0
1	0	0	1	0	0	1	0	0	1	1	1	0	1	0	1	1	1	0	0	0	0

Supplemental Table S4. Information about IGHD-contigs in VGP assemblies of mammalian species. We used the maternal assembly for finding IGHD genes in the genome of common marmoset (mCalJac1).

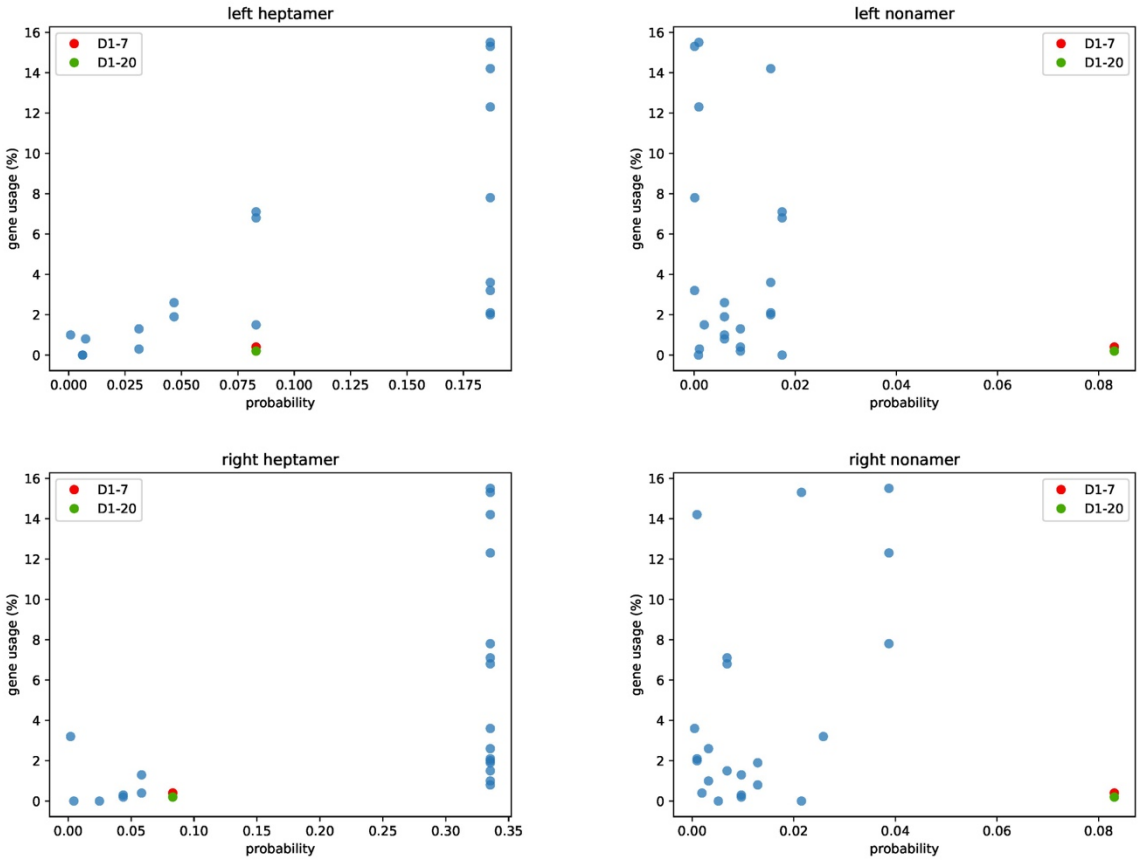
Common name of species	VGP ID of species	Contig covering IGHD locus	Contig length (Mbp)
common marmoset	mCalJac1	Super_scaffold_mat_12	136.9
ring-tailed lemur	mLemCat1	scaffold_1|arrow	285.8
European otter	mLutLut1	HiC_scaffold_5_arrow_ctg1	144.1
Canada lynx	mLynCan4	Super_Scaffold_2	146.1
stoat	mMusErm1	scaffold_5_arrow_ctg1	148,1
vaquita	mPhoSin1	scaffold_2_arrow_ctg1	178.6
pale spear-nosed bat	mPhyDis1	scaffold_175_arrow_ctg2	6.2
greater horseshoe bat	mRhiFer1	scaffold_58_arrow_ctg1	101.1
grey squirrel	mSciCar1	SUPER_2	199.8
Eurasian red squirrel	mSciVul1	HiC_scaffold_2_arrow_ctg1	204.4
California sea lion	mZalCal1	S6	146.9



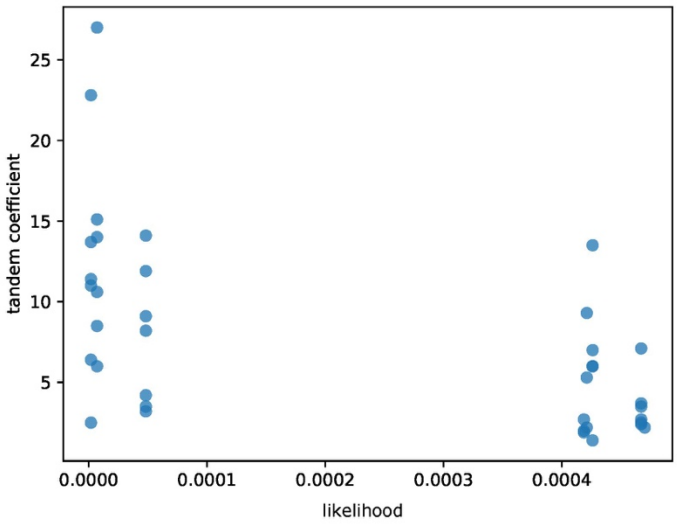
Supplemental Figure S1. Distribution of scores of all 2^{22} configurations (left) and $C_{22,12}$ configurations with twelve 1s (right). Bars corresponding to scores 24 (the score of a configuration formed by the twelve 2- and 3-turning cryptic nonamers) and above are shown in green. The distribution is shown in the logarithmic scale.



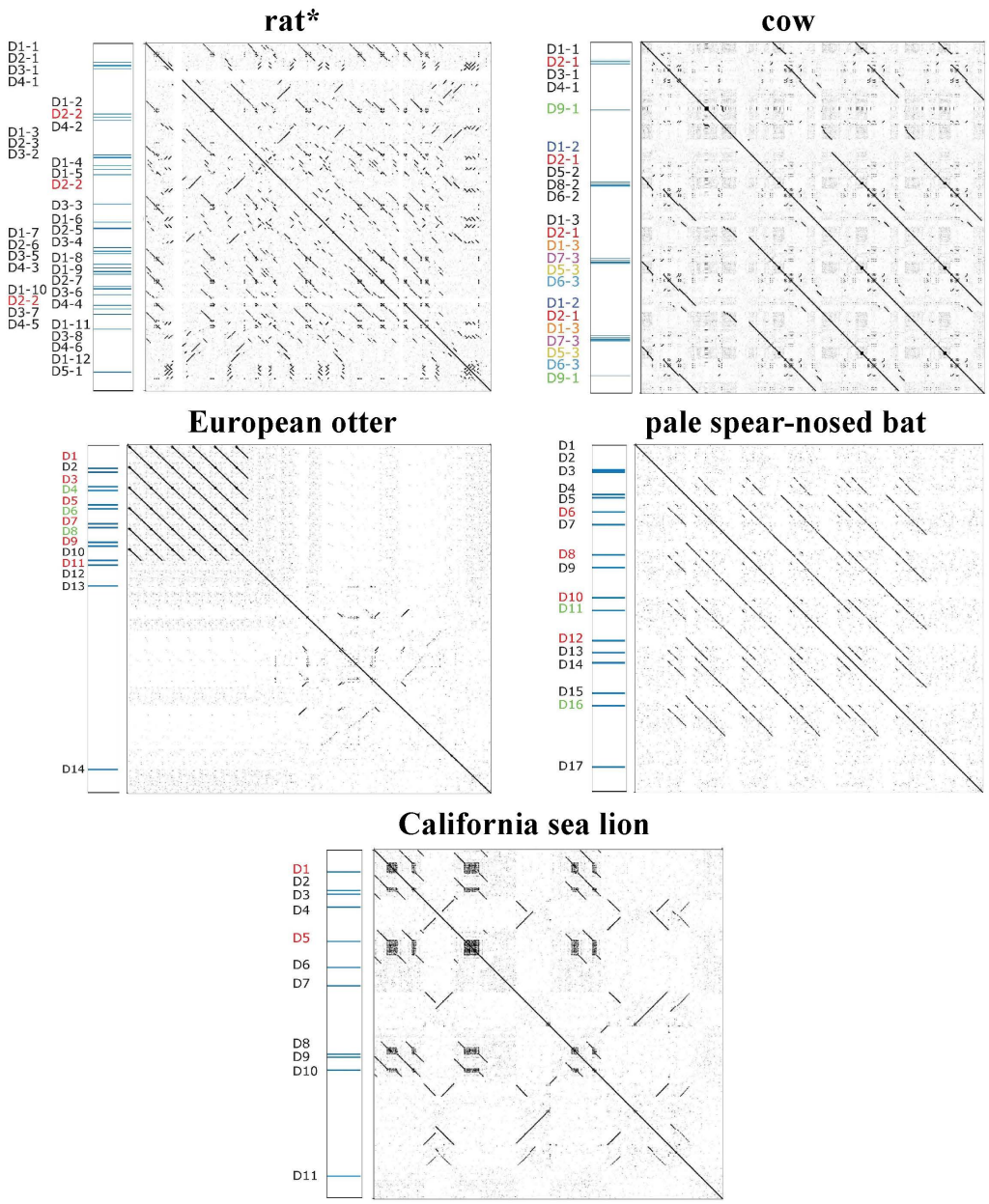
Supplemental Figure S2. Clonal lineages with long CDR3s in the INTESTINAL dataset. Each dot (x, y) of the blue line shows the number of large clonal lineages (y) with CDR3s of length at least x nt. Dots (x, y) of the orange and green lines show the number of large clonal lineages (y) with tandem CDR3s of length at least x detected by IgScout using $k = 10$ and $k = 11$, respectively. There are 157 large clonal lineages with long CDR3s (length at least 72 nt). 22 (40) out of them are formed by tandem fusions identified for $k = 11$ ($k = 10$) with false discovery rates 12.5% (29%). There are 21 large clonal lineages with ultralong CDR3s (length at least 84 nt). 11 (16) out of them are formed by tandem fusions identified for $k = 11$ ($k = 10$).



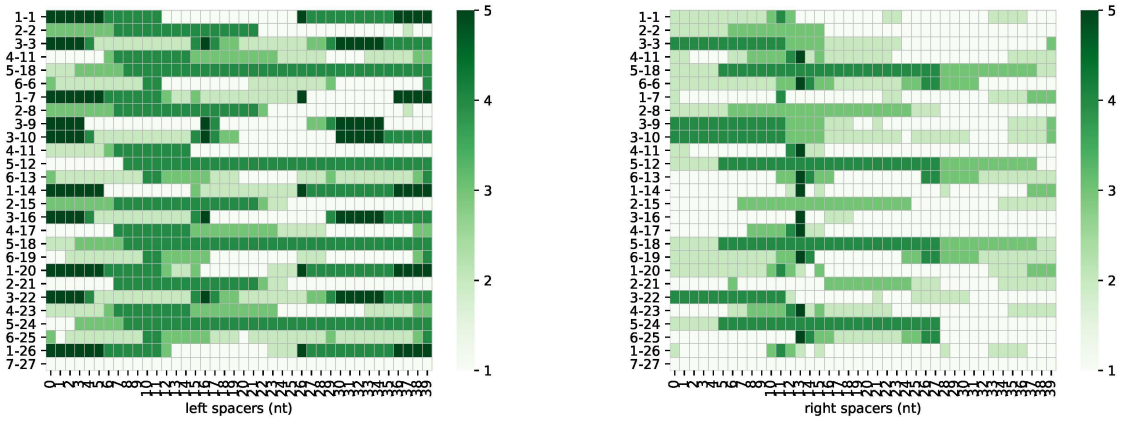
Supplemental Figure S3. Usage of human IGHD genes vs probabilities of canonical heptamers (left) and nonamers (right). Each dot corresponds to a human IGHD gene. Usage of a D gene is defined as the percentage of CDR3s derived from this gene. The correlation coefficients for the left heptamers, right heptamers, left nonamers, and right nonamers are 0.65, 0.49, –0.15, and 0.51, respectively. D genes IGHD1-7 and IGHD1-20 are shown as red and green dots, respectively.



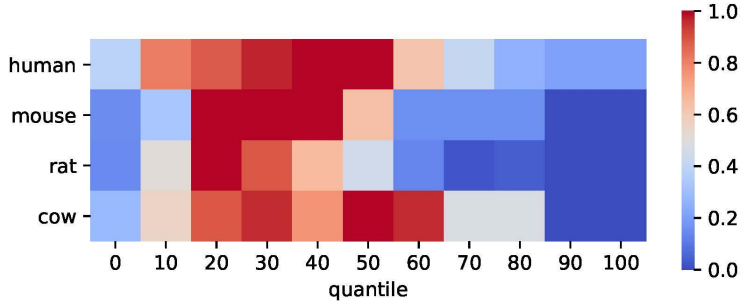
Supplemental Figure S4. Likelihoods of cryptic nonamers vs tandem coefficients. Dots represent tandem fusions shown in Figure 2. For each tandem fusion, we identify cryptic nonamers that explain it and compute their probabilities. If we identified two cryptic nonamers for the same fusion, we compute the likelihood that at least one of them works.



Supplemental Figure S5. IGHD loci of mammalian species have highly repetitive structure. Duplicated D genes are shown by the same (non-black) color. For better resolution, we show a 121,147 bp long fragment of 753,916 bp long rat IGHD loci (rat*) that covers 35 out of 38 IGHD genes and does not cover genes IGHD2-7, IGHD2-7, and IGHD2-1 following gene IGHD5-1 (the last gene in the dot plot). The distance between IGHD5-1 and IGHD2-7 is 404,988 bp.



Supplemental Figure S6. Multiplicities of left and right nonamers corresponding to 0–39 nt long spacers.



Supplemental Figure S7. Benchmarking SEARCH-D. The matrix shows values of the product of sensitivity and precision computed for quantile *q* varying from 0 to 100. Values in each row are normalized by the maximum product value and vary from 0 (blue) to 1 (red).