Supplementary material for "Alignment-free detection and seed-based identification of multi-loci V(D)J recombinations"

Cyprien Borée, Mathieu Giraud, Mikaël Salson

July 3, 2024

49bfdd3cd4 feature-a/one-heuristic-multiple-affects-working

	5'	D-like	3'	size (kB)	seeds		Index load	
					5'	3'	5'	3'
TRA	TRAV		TRAJ+down	74.8	12s	10s	0.198%	1.240%
TRB	TRBV	TRBD	TRBJ+down	73.8	12s	10s	0.188%	0.264%
TRB+	TRBD+up		TRBJ+down	4.8	12s	10s	0.001%	0.264%
TRG	TRGV		TRGJ+down	11.4	10s	10s	0.558%	0.080%
TRD	TRDV	TRDD	TRDJ+down	11.7	12s	10s	0.029%	0.078%
TRA+D	TRDV	TRDD	TRAJ+down	29.1	12s	10s	0.002%	1.240%
	TRDD+up		TRAJ+down	19.0	12s	10s	0.002%	1.240%
TRD+	TRDV		TRDD3+down	10.7	12s	10s	0.001%	0.012%
	TRDD2+up	TRDD	TRDJ+down	1.5	12s	10s	0.001%	0.012%
	TRDD2+up		TRDD3+down	0.4	12s	10s	0.001%	0.012%
IGH	IGHV	IGHD	IGHJ+down	230.7	12s	10s	0.301%	0.179%
IGH+	IGHD+up		IGHJ+down	13.5	12s	10s	0.016%	0.179%
IGK	IGKV		IGKJ+down	63.8	10s	10s	1.721%	0.082%
IGK+	IGKV IGK-INTRON		IGK-KDE	62.1	10s	10s	1.752%	0.053%
IGL	IGLV		IGLJ+down	65.6	10s	10s	2.935%	0.117%

Table 1: For a given seed w, the *index load* is the ratio describing how many of the $4^{weight(w)}$ possible seeds are indexed. Locus or gene types with less sequences can have shorter seeds while keeping low index loads that enables efficient filtering.

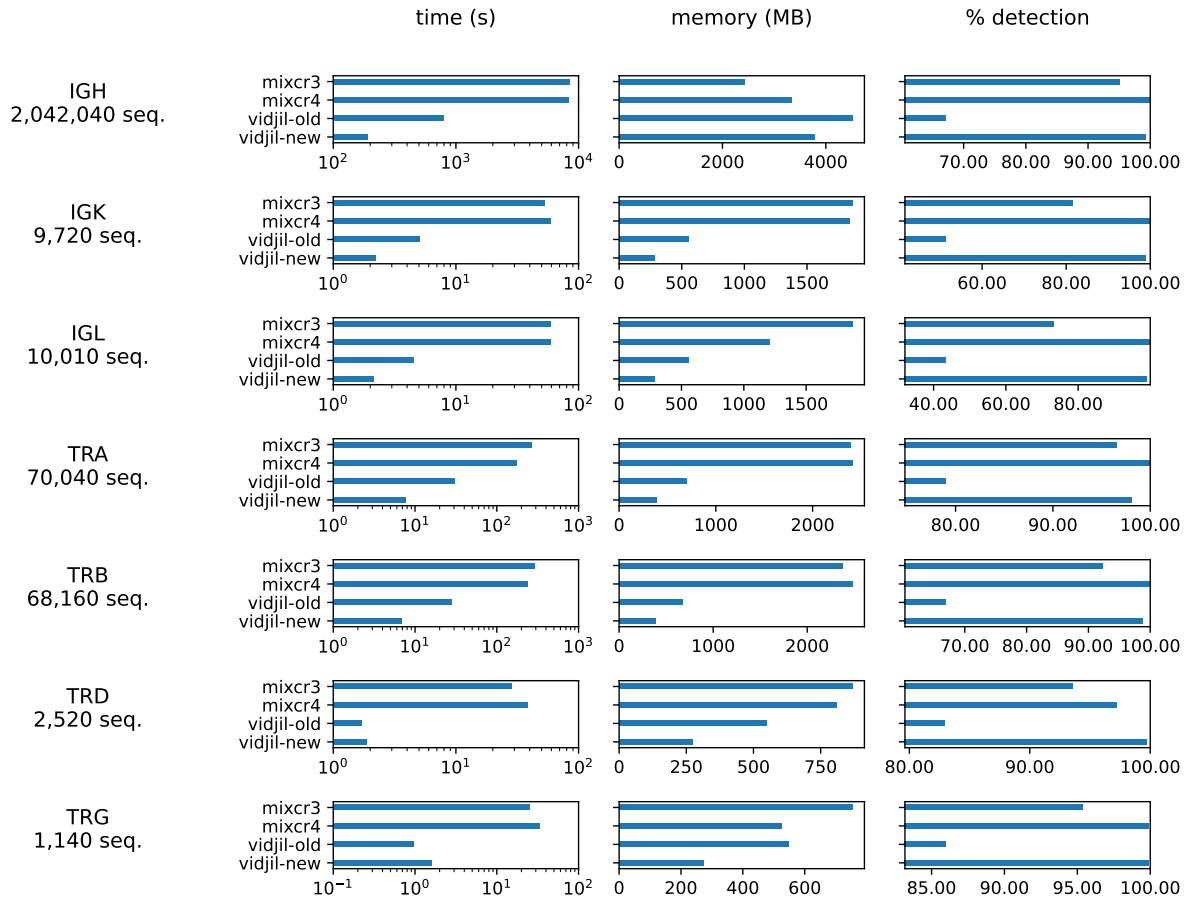

Figure 1: Detection by MiXCR and Vidjil-algo on synthetic V(D)J recombinations on all human loci with 10 % mutations. The X-axis on the time diagrams is logarithmic.

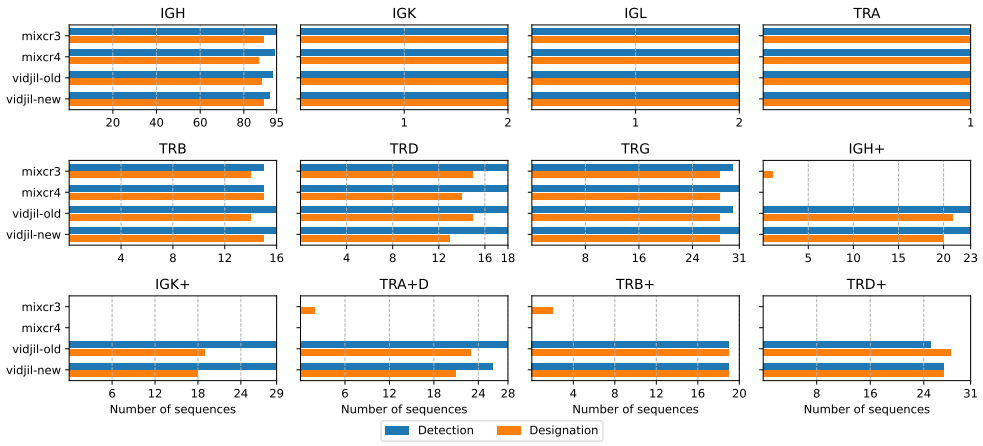

Figure 2: Correct detection and designation of V(D)J recombinations on manually curated V(D)J sequences (dataset E) with MiXCR and Vidjil-algo.

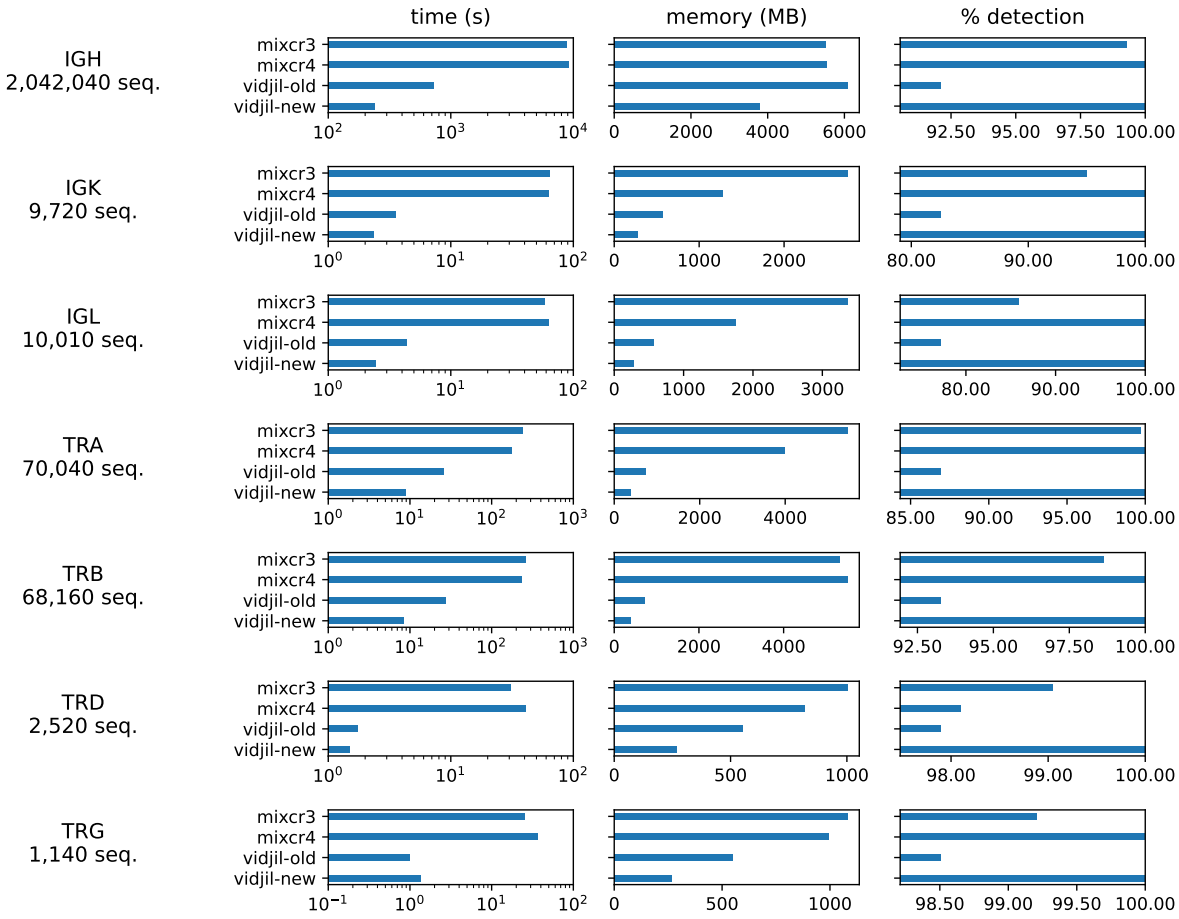

Figure 3: Detection by MiXCR and Vidjil-algo on synthetic V(D)J recombinations on all human loci (dataset B), with 5 % errors (80 % substitutions, 20 % indels). The X-axis on the time diagrams is logarithmic.