Supplemental material for

Gene prediction in the immunoglobulin loci

Vikram Sirupurapu[1], Yana Safonova[1,2], and Pavel A. Pevzner[1,*]

[1] Computer Science and Engineering Department, University of California San Diego, San Diego, USA

[2] The Department of Computer Science, Johns Hopkins University, Baltimore, USA

[*] Corresponding author: ppevzner@ucsd.edu

The document includes Supplemental Tables S1–18, Supplemental Figures S1–19, and 21 Supplemental Methods.

Supplemental Tables

Locus	Chromosome	Gene type	# genes	Start position (Mbp)	End position (Mbp)
IGH	14	V	120	105.6	106.9
	14	D	27		
	14	J	6		
IGHorphon15	15	V	9	20	22.2
IGHorphon16	16	V	15	32	33.8
IGHorphon21	21	V	1	10.6	10.6
IGK	2	V	76	88.9	90.2
	2	J	5		
IGKorphon1	1	V	1	144.1	144.1
IGKorphon9.1	9	V	1	40.6	40.6
IGKorphon9.2	9	V	3	64.8	65.8
IGKorphon22	22	V	4	16.9	16.9
IGKorphonY	Y	V	1	11.5	11.5
IGL	22	V	73	22	22.9
	22	J	11		
IGLorphon8	8	V	1	47.2	47.2
TRA	14	V	54	21.6	22.6
	14	J	61		
TRB	7	V	64	142.3	142.8
	7	D	2		
	7	J	14		
TRBorphon9	9	V	9	33.6	33.8
TRD	14	V	8	22.4	22.5
	14	D	3		
	14	J	4		
TRG	7	V	12	38.2	38.4
	7	J	5		

Supplemental Table S1. Information about human IG and TCR loci. Information was collected from the IMGT (for canonical genes) and NCBI (for orphon genes) databases.

V genes			
Species	**Cow**	**Human**	**Mouse**
No. of genes in IMGT Database	44	371	408
No of genes that have valid alignments with reference	42	366	407
No of alignments with percent identity = 100% (perfect matches)	42	345	387
No of alignments with percent identity at least 90% but less than 100%	0	21	20
No of genes after collapsing exact overlapping alignments	36	84	165
No of genes after removing alignments with starting positions located within 100 bp	36	70	154
No of filtered genes with percent identity = 100%	36	68	154
No of filtered genes with percent identity at least 90% but less than 100%	0	2	0
D genes			
Species	**Cow**	**Human**	**Mouse**
No. of genes in IMGT Database	23	34	39
No of genes with >0 alignments with reference	23	31	32
No of alignments with percent identity = 100%	23	31	32
No of alignments with 90%<=percent identity<100%	0	0	0
No of genes after collapsing exact overlapping alignments	14	25	15
No of genes after removing alignments with starting positions located within 100 bp	14	25	15
No of filtered genes with percent identity = 100%	14	12	15
No of filtered genes with 90%<= percent identity < 100%	0	0	0
J genes			
Species	**Cow**	**Human**	**Mouse**
No. of genes in IMGT Database	18	13	9
No of genes with >0 alignments with reference	18	12	9
No of alignments with percent identity = 100%	17	12	9
No of alignments with 90%<=percent identity<100%	1	0	0
No of genes after collapsing exact overlapping alignments	12	6	5
No of genes after removing alignments with starting positions located within 100 bp	12	6	4
No of filtered genes with percent identity = 100%	12	6	4
No of filtered genes with 90%<= percent identity < 100%	0	0	0

Supplemental Table S2. Information about alignments of the IMGT database genes with the IGH reference. The blue highlighted rows show the final number of genes considered for the IterativeIGDetective algorithm.

Latin name	common name	species ID	short common name	total length of 11 (Mbp)	# predicted IGHV RSSs	RSSV density (RSSV/Mbp)	# IGHV pseudogenes	# IGHV genes
Balaenoptera musculus	blue whale	mBalMus1	blue whale	328.2	1006	3.06	1	13
Callithrix jacchus	common marmoset	mCalJac1	marmoset	396.1	2431	6.13	12	50
Choloepus didactylus	southern two-toed sloth	mChoDid1	sloth	492.3	1346	2.73	57	60
Lemur catta	ring-tailed lemur	mLemCat1	lemur	285.8	1067	3.73	4	20
Lutra lutra	European otter	mLutLut1	otter	314.8	984	3.12	28	29
Lynx canadensis	Canada lynx	mLynCan4	lynx	147.2	512	3.47	4	17
Molossus molossus	Pallas's mastiff bat	mMolMol1	mastiff bat	388.3	1233	3.17	4	10
Mustela erminea	stoat	mMusErm1	stoat	153.5	490	3.19	15	15
Ornithorhynchus anatinus	platypus	mOrnAna1	platypus	224.6	1059	4.71	18	37
Phocoena sinus	vaquita	mPhoSin1	vaquita	237.4	729	3.07	4	6
Phyllostomus discolor	pale spear-nosed bat	mPhyDis1	spear-nosed bat	719.7	1279	1.77	52	32
Pipistrellus kuhlii	Kuhl's pipistrelle	mPipKuh1	pipistrelle	11.6	69	5.94	20	9
Rhinolophus ferrumequinu	greater horseshoe bat	mRhiFer1	horseshoe bat	101.4	424	4.17	12	63
Sciurus carolinensis	grey squirrel	mSciCar1	grey squirrel	200.8	749	3.72	40	21
Sciurus vulgaris	Eurasian red squirrel	mSciVul1	red squirrel	204.3	821	4.01	60	36
Tursiops truncatus	bottlenose dolphin	mTurTru1	dolphin	203.7	674	3.30	20	27
Zalophus californianus	California sea lion	mZalCal1	Sea lion	146.9	493	3.35	5	3
Pan troglodytes	Chimpanzee	chimp	chimp	694.3	6570	9.46	23	39
Gorilla gorilla	Western Gorilla	gorilla	gorilla	702.4	7027	10.00	28	47
Pongo pygmaeus	Orangutan	orangutan	orangutan	201.3	2075	10.30	21	47

Supplemental Table S3. Information about IterativeIGDetective predictions of IGHV genes in target species. The column "Species ID" shows VGP IDs for seventeen VGP species and short common names for three great ape species. For the common marmoset (mCalJac1), VGP assembled maternal and paternal haplotypes. For this analysis, we used the maternal assembly only. The RSSV density is defined as the number of predicted RSSVs divided by the total length of all IGH contigs.

Species ID	# IGH-contigs	Species ID	# IGH-contigs
stoat	8	horseshoe bat	6
otter	5	chimp	23
sea lion	5	gorilla	21
lynx	3	orangutan	7
vaquita	5	marmoset	4
dolphin	10	lemur	3
blue whale	6	red squirrel	3
mastiff bat	7	grey squirrel	5
pipistrelle	5	sloth	7
spear-nosed bat	16	platypus	12

Supplemental Table S4. The number of IGH-contigs for twenty target species.

Species	RSSV		$RSSD_{left}$		$RSSD_{right}$		RSSJ	
	7-mer	9-mer	7-mer	9-mer	7-mer	9-mer	7-mer	9-mer
Human	0.0200	0.0900	0.0010	0.0100	0.0070	0.0015	0.1200	0.2100
Cow	0.0050	0.0275	0.0200	0.0025	0.0025	0.0050	0.0100	0.0450
Mouse	0.0500	0.0095	0.0200	0.0075	0.0300	0.0200	0.4200	0.4200
Combined	0.1300	0.0175	0.0150	0.0150	0.0350	0.0050	0.0325	0.0400

Supplemental Table S5. Likelihood ratio thresholds. The shown likelihood ratios (L_{min}) achieve the highest F1 scores.

	Human RSSV (70)				Human RSSD (25)				Human RSSJ (6)			
Reference profile	DS	TP	FP	FN	DS	TP	FP	FN	DS	TP	FP	FN
Human	56	45	11	25	24	21	3	4	4	4	0	2
Cow	61	43	18	27	10	7	3	18	25	5	20	1
Mouse	70	1	69	69	30	10	20	15	1	1	0	5
Combined	64	44	20	26	17	15	2	10	6	5	1	1

Supplemental Table S6. Information about the number of the candidate RSSs captured in the human IGH locus. Candidate RSSs were predicted based on each of four reference profiles (human, mouse, cow, and combined). The "DS" columns show the number of detected signals in IGH locus, the "TP", "FN", and "FP" columns show the number of true positives, false positives, and false negatives, respectively. The total number of canonical signals is shown in parentheses next to signal type.

RSSV												
Reference	Target											
	Human (70)			Cow (36)			Mouse (154)			Combined (260)		
	FDR	TPR	F1	FDR	TPR	F1	FDR	TPR	F1	FDR	TPR	F1
Human	0.196	0.642	0.714	0.310	0.555	0.615	0	0	NA	0.235	0.25	0.376
Cow	0.295	0.614	0.656	0.275	0.805	0.763	0.962	0.006	0.011	0.416	0.280	0.379
Mouse	0.985	0.014	0.014	0	0	NA	0.141	0.707	0.775	0.427	0.423	0.486
Combined	0.312	0.628	0.656	0.130	0.555	0.677	0.145	0.571	0.684	0.194	0.596	0.685
RSSD												
Reference	Target											
	Human (25)			Cow (14)			Mouse (15)			Combined (54)		
	FDR	TPR	F1	FDR	TPR	F1	FDR	TPR	F1	FDR	TPR	F1
Human	0.125	0.84	0.857	0.939	0.428	0.106	0.9	0.466	0.16	0.822	0.629	0.277
Cow	0.3	0.28	0.40	0.333	0.857	0.75	0.2	0.533	0.64	0.289	0.5	0.587
Mouse	0.667	0.4	0.363	0.363	0.5	0.56	0.153	0.733	0.785	0.481	0.518	0.518
Combined	0.117	0.6	0.714	0.307	0.642	0.667	0.111	0.533	0.667	0.179	0.592	0.687
RSSJ												
Reference	Target											
	Human (6)			Cow (12)			Mouse (4)			Combined (22)		
	FDR	TPR	F1	FDR	TPR	F1	FDR	TPR	F1	FDR	TPR	F1
Human	0	0.833	0.909	0	0.25	0.4	0	0	NA	0	0.363	0.533
Cow	0.8	0.833	0.322	0.33	1	0.8	0	0	NA	0.60	0.772	0.523
Mouse	0	0.166	0.285	0	0.083	0.153	0	0.75	0.857	0	0.227	0.370
Combined	0.167	0.833	0.833	0	0.333	0.5	0	0.5	0.667	0.083	0.5	0.647

Supplemental Table S7. RSS detection statistics. Rows refer to the reference species profile, Columns refer to the target species in which we are detecting RSSs using the reference profile. The number inside the parenthesis is the total number of canonical genes for a given species. For each species and the gene type, the highest (the second highest) F1 score given is highlighted in blue (green).

RSS type	heptamer/ nonamer	total number of patterns passing the likelihood threshold	# patterns flanking the candidate genes in target species identified by IterativeIGDetective	Percentage of observed patterns among all patterns passing the likelihood threshold (%)
V	heptamer	1	1	100
	nonamer	690	82	12
D_{left}	heptamer	79	13	16
	nonamer	317	21	7
D_{right}	heptamer	19	10	5
	nonamer	625	34	5
J	heptamer	48	12	3
	nonamer	41	17	41

Supplemental Table S8. The total number of candidate patterns passing the likelihood thresholds for classification as a candidate RSS is identified as described in section "RSS detection algorithm" in Methods).

species	# predicted target-like genes	mean PI of the highest-scoring alignment with a V gene predicted at the previous iteration (%)	mean PI of the highest-scoring alignment with a canonical human V gene (%)	mean length of the longest shared common *k*-mer with a V gene predicted at the previous iteration (%)
sloth	22	78.5	65.9	19.5
lemur	3	73.2	63.9	129
mastiff bat	1	70.2	69.6	17
platypus	9	79.3	64.4	57.4
vaquita	1	62.0	57.0	285
spear-nosed bat	9	71.2	61.7	29.7
pipistrelle	2	71.2	65.5	123
red squirrel	2	65.6	64.5	24
gorilla	1	67.3	69.4	28

Supplemental Table S9. Information about identified target-like V genes across various species. Across the twenty target species, IterativeIGDetective identified 971 human-like V genes at the first iteration and 47 (3) target-like V genes at the second (third) iteration. Target-like V genes were identified in 9 out of the 20 target species.

Species	# candidate RSSs	# isolated vertices	# vertices in giant components	# vertices in V-graph	# edges in V-graph	# of components in V-graph		
						small	large	giant
Human	28394	6942	20768	684	1953	136	43	1
Cow	7795	6104	1136	555	2856	70	36	1
Mouse	13483	8424	3930	1129	33242	97	55	1
Spear-nosed bat	7645	6479	0	1166	12135	85	48	0
Horseshoe bat	6977	6048	0	929	21610	62	30	0

Supplemental Table S10. Information about V-graphs constructed by BlindIGDetective. Number of isolated vertices" is computed as described in Supplementary Note "Alignment speedup for constructing *S*-graphs". BlindIGDetective ignores giant components as they are likely triggered by spurious RSSs within repeat regions.

cluster ID	cluster locus	clump ID	clump size	clump annotation index	#unannotated vertices/ #unannotated and accordant vertices	conservation* (%)	
						minimum	median
H0	IGHV	0.0	27	0.96	1/1	83	83
		0.1	19	0.94	1/0	80	80
		0.2	8	1	0/0	-	-
		0.3	3	0	3/0	61	62
		0.4	2	0	2/1	56	56
H1	IGLV	1.0	21	0.71	6/2	76	83
		1.1	3	1	0/0	-	-
		1.2	3	0.66	1/1	89	89
		1.3	2	1	0/0	-	-
		1.4	2	0	2/0	55	57
H5	TRGV	5.0	6	0.83	1/0	88	88
H6	TRAV	6.0	4	1	0/0	-	-
		6.1	2	1	0/0	-	-
H10	TRBV	10.0	2	0.5	1/1	88	88
		10.1	2	1	0/0	-	

Supplemental Table S11. Information about the annotated human clusters constructed by BlindIGDetective. The cluster IDs correspond to the cluster IDs from Table 2 in results. The conservation* column represents conservation among only the unannotated vertices of a clump. The unannotated and accordant vertices in the IGH cluster H0 have coding lengths 276 (clump 0.0) and 348 nt (in clump 0.4). The unannotated and accordant vertices in the IGL cluster H1 have coding lengths 276 (clump 1.0), 231 (clump 1.0), 279 (clump 1.0), and 348 nt (clump 1.2).

Cow														
cluster ID	cluster size	# clumps / |L.C|	PI	coding length (nt)	cluster density (%)	cluster span (Mb)	center vertex		AI/AI80	conservation wrt species genes		conservation wrt human genes		locus
							chr	coordinate (Mb)		min	med	min	med	
C0	47	8/21	93	336	24	1.113	10	23.28	0.96/0.28	58	100	55	77	TRA
C1	36	3/30	99	408	70	0.406	17	70.85	1/0.06	100	100	73	77	IGL
C2	31	5/10	95	315	20	0.666	10	25.12	1/0.66	100	100	70	77	TRD
C3	28	3/20	98	354	54	0.437	4	105.88	0.93/0	55	100	56	77	TRB
C4	21	5/10	97	219	28	0.318	21	0.33	0.71/0.24	55	100	54	73	IGH
C5	5	1/5	96	231	90	0.099	1	4.96	0/0	56	56	54	56	
C6	4	2/2	100	216	33	0.182	X	36.46	0/0	55	57	55	55	
Mouse														
cluster ID	cluster size	# clumps / |L.C|	PI	coding length (nt)	cluster density (%)	cluster span (Mb)	center vertex		AI/AI80	conservation wrt species genes		conservation wrt human genes		locus
							chr	coordinate (Mb)		min	med	min	med	
M0	106	9/43	93	342	44	2.437	12	114.73	0.93/0.08	56	100	54	76	IGH
M1	53	9/13	99	315	14	1.215	14	53.56	0.94/0.15	59	100	55	75	TRA
M2	7	3/3	100	261	100	2.211	Y	78.80	0/0	58	58	54	54	
M3	6	2/3	97	237	100	1.392	Y	8.80	0/0	56	58	54	55	
M4	5	2/3	100	261	100	1.201	Y	83.80	0/0	58	58	54	54	
M5	4	1/4	87	207	100	0.576	7	47.18	0/0	56	57	55	55	
M6	4	1/4	100	261	100	0.573	Y	85.50	0/0	55	57	54	54	
M7	4	1/4	75	285	100	0.04	3	3.021	0/0	58	59	56	56	

Supplemental Table S12. Information about large clusters derived from the cow and mouse genomes. BlindIGDetective constructed 12 and 44 large clusters (7 and 8 accordant clusters) in the cow and mouse genomes, respectively. "L.C" represents the size of the largest clump in the cluster. Annotation index and annotation80 index are abbreviated to "AI " and "AI80". Annotated clusters are highlighted in blue. Clusters are ordered in the decreasing orders of their sizes. Conservation is shown with respect to predicted V genes from the same species as well as canonical human V genes, with annotation index (annotation80 index) defined with respect to the former (latter). Predicted V genes are the canonical V genes for reference species. The locus column classifies the annotated clusters as one of the families of cow (mouse) IG or TCR genes described in Supplemental Table S1 (highlighted in blue). All annotated clusters, except for cow cluster C4, have coding length greater than 315 nt, consistent with the range of coding lengths in known V genes.

Horseshoe bat														
cluster ID	cluster size	# clumps / |L.C|	PI	coding length (nt)	cluster density (%)	cluster span (Mb)	center vertex		AI/AI80	conservation wrt species genes		conservation wrt human genes		locus
							contig	coordinate (Mb)		min	med	min	med	
RF0	53	5/21	93	366	17	0.663	S6	21.266	0/0.77	54	56	56	84	IGL
RF1	38	2/19	96	357	38	0.262	S56	0.176	1/0.97	100	100	75	89	IGH
RF2	31	2/17	96	360	36	0.206	S58	0.001	1/0.94	100	100	74	90	IGH
RF3	28	8/7	96	312	12	0.463	S58	72.85	0/0.71	52	55	75	82	TRA/ TRD*
RF4	6	2/3	93	363	40	0.036	S2	0.037	1/1	100	100	87	90	IGH
RF5	4	2/2	82	270	100	0.793	S108	41.213	0/0	50	50	54	55	

Supplemental Table S13. Large Clusters reported by BlindIGDetective on the genome of the horseshoe bat. BlindIGDetective reported 48 clusters, including 8 large clusters. Cluster RF3 contains 7 clumps classified as the TRA family and 1 clump classified as the TRD family. Only accordant unannotated clusters are shown. The description of the columns is consistent with the caption to Table 2 in Results. In the "contig" column, "scaffold_<N>_arrow_ctg1" and Super_Scaffold<N> are shortened to S<N> and SS<N> respectively. The "contig" column refers to the VGP assembly version "mRhiFer1.pri.cur.20180907".

$RSSD_{left}$			
9-mer/7-mer	**Cluster 1**	**Cluster 2**	**Cluster 3**
Cluster 1	0	0	9
Cluster 2	3	7	0
Cluster 3	0	6	0

$RSSD_{right}$		
9-mer/7-mer	**Cluster 1**	**Cluster 2**
Cluster 1	6	2
Cluster 2	4	0
Cluster 3	7	6

Supplemental Table S14. Information about clustering of human RSSDs. Given clustering of heptamers into *N* clusters and nonamers into *M* clusters, we assign a label (*i*,*j*) to an RSS formed by a heptamer from cluster *i* and a nonamer from cluster *j*. The table shows the number of RSSDs with a given label.

Signal string	# clusters	# signals in cluster	total usage (%)	mean usage (%)	median usage (%)	P-value
D_{left} heptamer	1	3	1.8	0.6	0.8	0.042
	2	13	80.5	6.1	3.2	
	3	9	17.6	1.9	0.4	
D_{right} heptamer	1	17	94.5	5.5	2.6	0.007
	2	8	5.4	0.6	0.2	
D_{left} nonamer	1	9	17.6	1.9	0.4	0.071
	2	10	28.2	2.8	1.9	
	3	6	54.1	9	10	
D_{right} nonamer	1	8	22.2	2.7	2	0.401
	2	4	21.9	5.4	2.8	
	3	13	55.8	4.2	0.8	
D_{left} 16-mer	1	10	28.2	2.8	1.9	0.071
	2	6	54.1	9.0	10.0	
	3	9	17.6	1.9	0.4	
D_{right} 16-mer	1	8	22.2	2.7	2.0	0.628
	2	8	53.6	6.7	4.8	
	3	9	24.1	2.6	1.3	
l9l7r7r9	1	10	69.5	6.95	6.95	0.02
	2	5	2.2	0.44	0.3	
	3	6	6.3	1.05	0.9	
	4	4	21.9	5.47	2.85	
l7r7	1	6	2.2	0.36	0.25	0.009
	2	15	92.7	6.18	3.6	
	3	4	5	1.25	0.9	
l9r9	1	4	21.9	5.475	2.85	0.028
	2	5	2.2	0.44	0.3	
	3	10	69.5	6.95	6.95	
	4	6	6.3	1.05	0.9	

Supplemental Table S15. Cluster statistics for human RSSDs. For each signal string, P-value is calculated using the Kruskal-Wallis test.

Signal string	# clusters	# signals in cluster	Total usage (%)	Mean usage (%)	Median usage (%)	P-value
V heptamers	1	5	8.44	1.68	0.003	0.366
	2	52	91.56	1.76	1.00	
V nonamers	1	38	74.58	1.96	1.25	0.618
	2	3	3.99	1.33	0.003	
	3	16	21.42	1.33	0.90	
V 16-mer	1	11	17.67	1.60	0.75	0.297
	2	45	82.31	1.82	1.22	
	3	1	0.004	0.004	0.004	

Supplemental Table S16. Cluster statistics for human RSSVs. P-values are calculated using the Kruskal-Wallis test.

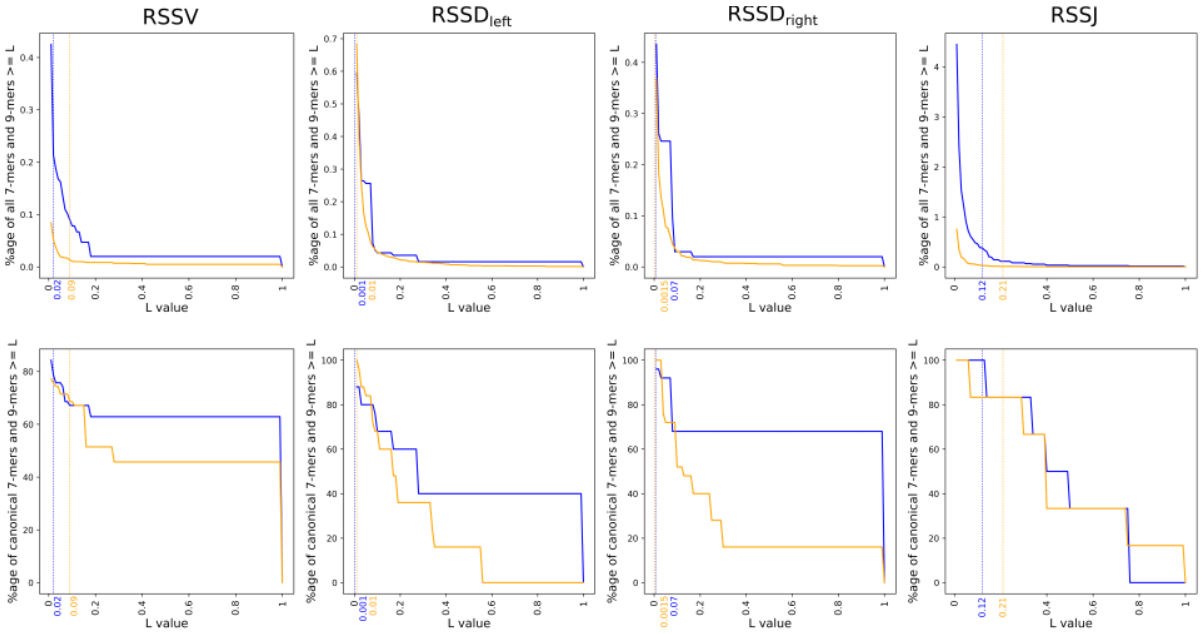
Signal string	Cluster no.	No. of signals in cluster	Total usage (%)	Mean usage (%)	Median usage (%)	P-value
J Heptamers	1	3	82.39	27.46	27.91	0.145
	2	1	3.04	3.04	3.04	
	3	2	12.12	6.06	6.06	
J Nonamers	1	3	33.30	11.10	3.04	0.538
	2	2	50.15	25.07	25.07	
	3	1	14.09	14.09	14.09	
J 16-mer	1	2	12.12	6.06	6.06	0.232
	2	2	68.29	34.14	34.14	
	3	1	14.09	14.09	14.09	
	4	1	3.04	3.04	3.04	

Supplemental Table S17. Cluster statistics for human RSSJs. For each signal string, P-value is calculated using the Kruskal-Wallis test.

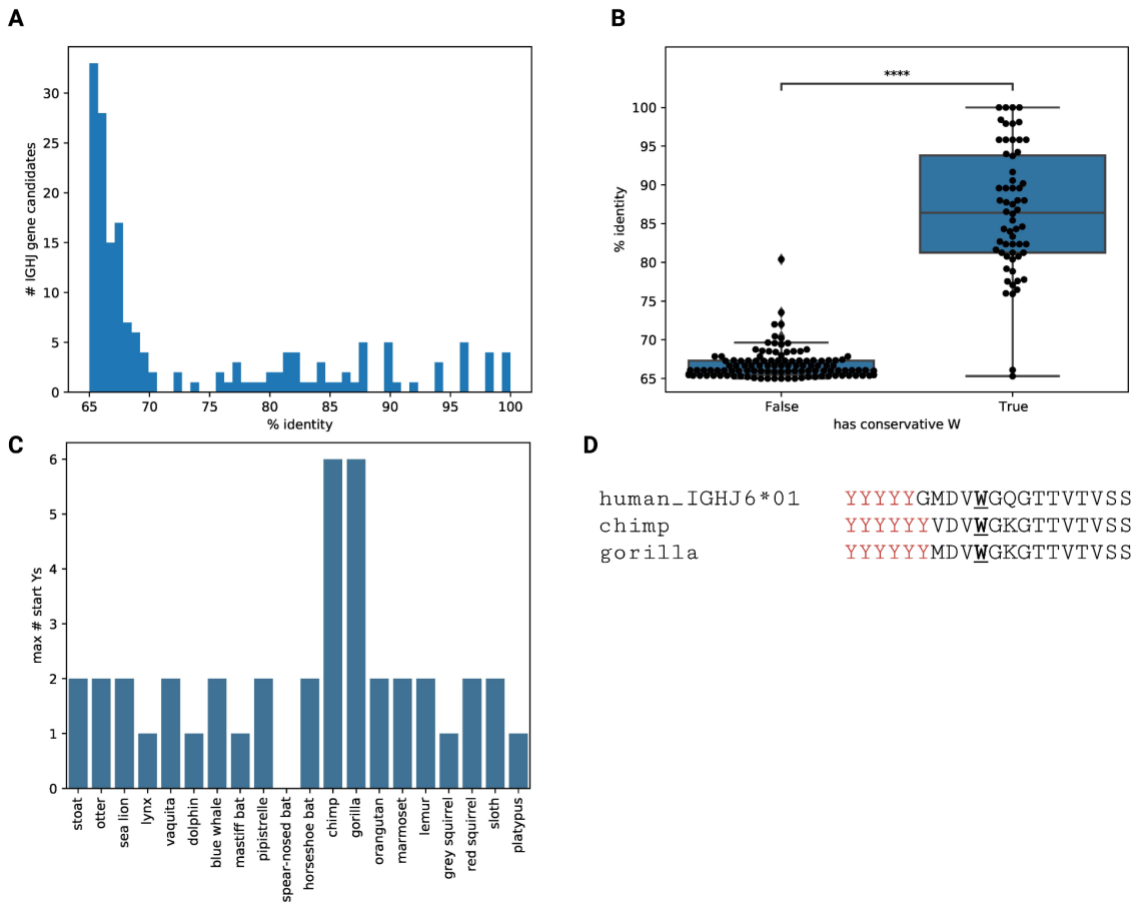
Species	non-iterative mode, pi_{min} = 70%			iterative mode			non-iterative mode, pi_{alt} = 60%			non-iterative mode, pi_{alt} = 55%		
	FDR	TPR	F1	FDR	TPR	F1	FDR	TPR	F1	FDR	TPR	F1
Human	0.140	0.700	0.771	0.140	0.700	0.771	0.155	0.700	0.765	0.169	0.700	0.759
Cow	0.130	0.555	0.677	0.130	0.555	0.677	0.130	0.550	0.677	0.130	0.550	0.677
Mouse	0.055	0.551	0.696	0.070	0.551	0.690	0.132	0.550	0.674	0.132	0.550	0.674

Supplemental Table S18. Percent identity threshold statistics for iterative vs non-iterative modes. The F1 score, True Positive Rate (TPR) and False discovery Rate (FDR) are shown for IterativeIGDetective running in the default non-iterative mode, iterative mode, non-iterative mode with reduced percent identity equal to 60% and 55%.

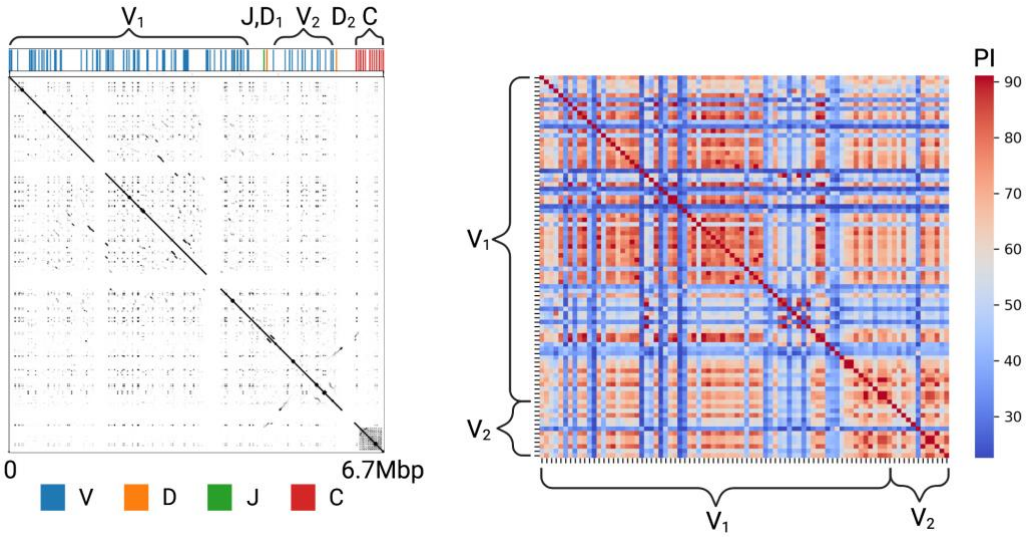
Supplemental Figures



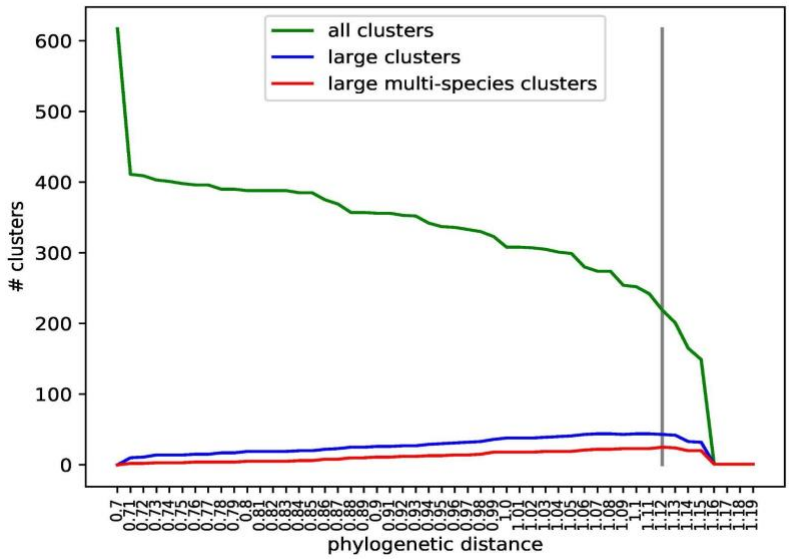
Supplemental Figure S1. Percentage of heptamers and nonamers passing the likelihood threshold. (Top) Percentage of all heptamers and all nonamers seen in the human IGH locus passing the L value threshold for an RSS and (Bottom) Percentage of all canonical human RSSs passing the L value thresholds is shown for all four signal types (RSSV, RSSDleft, RSSDright, and RSSJ from left to right). Heptamers and nonamers are represented by blue and orange color respectively. The dotted vertical lines represent the likelihood threshold (Lmin) chosen for the corresponding RSS (RSSV, RSSDleft, RSSDright, and RSSJ.



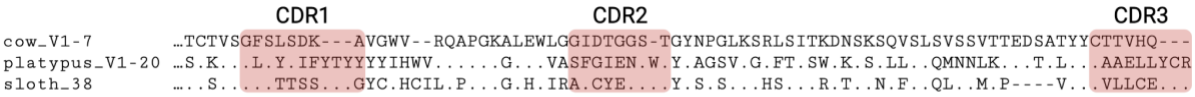
Supplemental Figure S2. Comparative analysis of IGHJ gene candidates. (A) The distribution of percent identities of 174 candidate J genes with respect to the closest human J gene. The distribution reveals the "low PI" and "high-PI" modes with 114 and 60 genes, respectively. **(B)** The distribution of percent identities of candidate J genes depending on whether they have the conservative tryptophan-encoding codon. P-value=6.23×10^{-25}, according to the Kruskal-Wallis test. **(C)** The maximum length of poly-Y prefix across target species. The plot was computed for 60 candidate J gene candidates in the "high PI" mode. **(D)** Two J gene candidates (from chimp and gorilla) with 6 starting tyrosines aligned against the human IGHJ6*01 gene with 5 starting tyrosines.



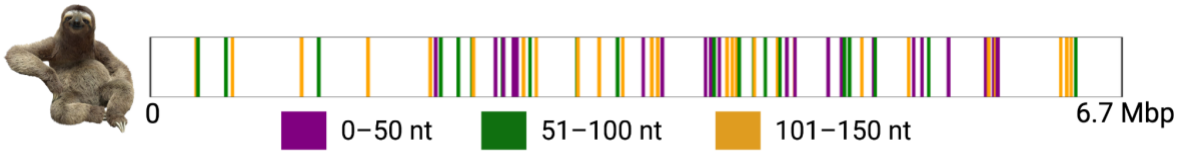
Supplemental Figure S3. The IGH locus in the sloth genome. The dot plot on the left shows the sloth IGH locus aligned against itself. The dot plot is generated using the Gepard tool (Krumsiek et al., 2007). Positions of 86 V (blue), 2 D (orange), 2 J (green), and 17 C (red) genes are shown on the top. The heatmap on the right shows the percent identities of the pairwise alignments of all 86 V genes ordered according to their positions in the IGH locus. Bounds of the V1 and V2 groups are shown by brackets.



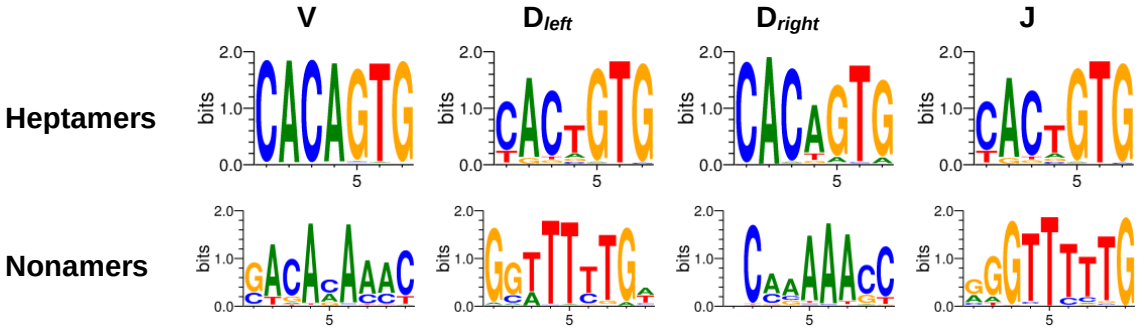
Supplemental Figure S4. The number of V gene clusters vs the maximum phylogenetic distance. The grey vertical line (at *distance*=1.12) corresponds to the maximum number of large multi-species clusters (25).



Supplemental Figure S5. Three V genes with long non-canonical CDR3 suffixes. The shown genes IGHV1-7 (cow), IGHV1-20 (platypus), and sloth_38, have suffixes CTTVHQ, LAAELLYCR, and CVLLCE, respectively. Dots stand for amino acids that are identical with the amino acid at the same position in the cow IGHV1-7 gene.

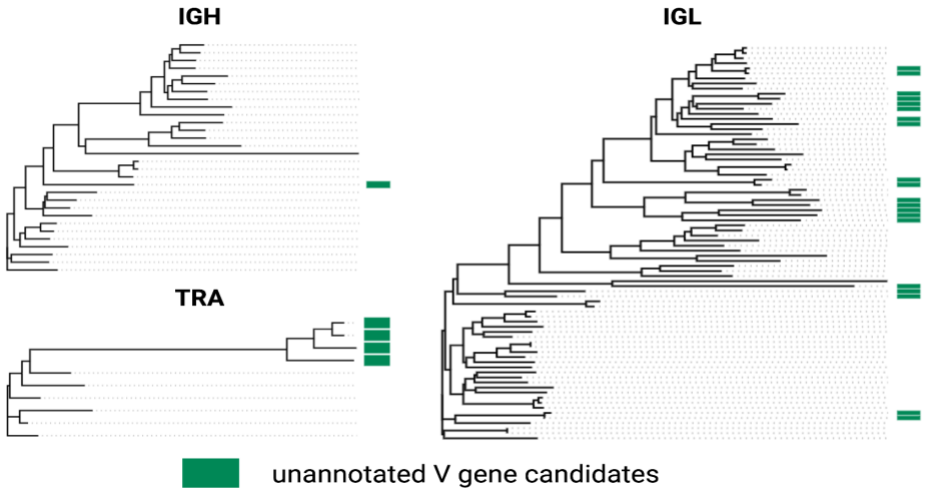


Supplemental Figure S6. The candidate D genes are scattered across the entire IGH locus in sloth. The candidate sloth D genes (identified using SEARCH-D and IterativeIGDetective) are colored according to their lengths: <50 nt (purple), from 50 to 100 nt (green), and ≥100 nt (orange).

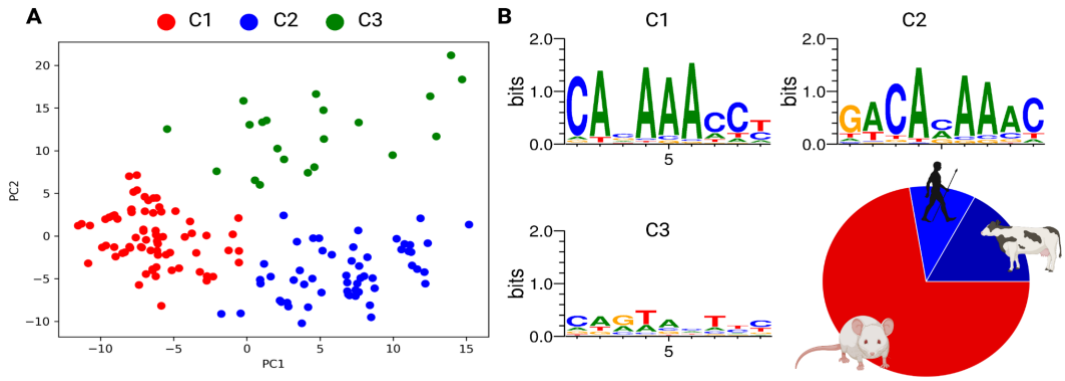


Supplemental Figure S7. Sequence logos of RSSs combined from three reference species and twenty target species.

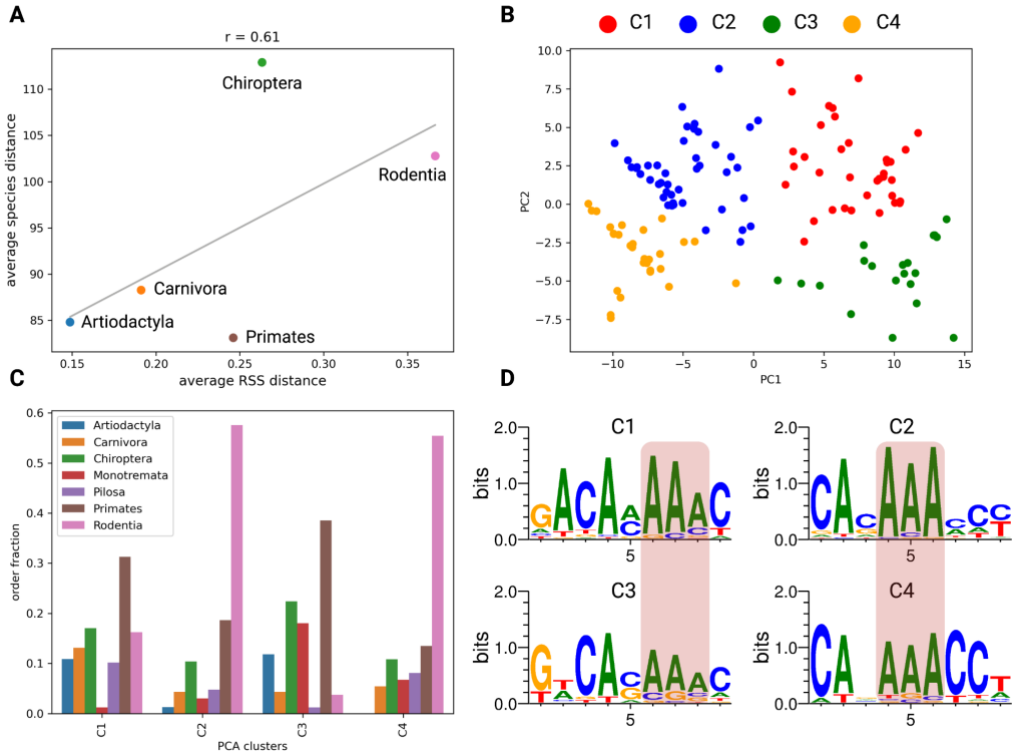
Supplemental Figure S8. Global alignment between two unannotated IGH *v*-fragments (left) and two unannotated IGL *v*-fragments (right) in the human genome. The *v*-fragments are extended in the 3' direction to include the RSSV (highlighted in blue), till an in-frame stop codon is encountered. The alignment is broken up into 7 blocks for better visualization. We also note that the 2 IGH *v*-fragments have a coding length of 597 nt and 117 nt respectively, while the 2 IGL *v*-fragments both have a coding length of 27 nt. Numbers flanking each block represent the index of the block within the *v*-fragments, RSSV and the extended sequence. These 4 sequences represent unannotated *v*-fragments in annotated clusters and have no significant hits in the IgBlast database. Pairs of IGH (IGL) *v*-fragments are aligned with percent identity=88% (95%).



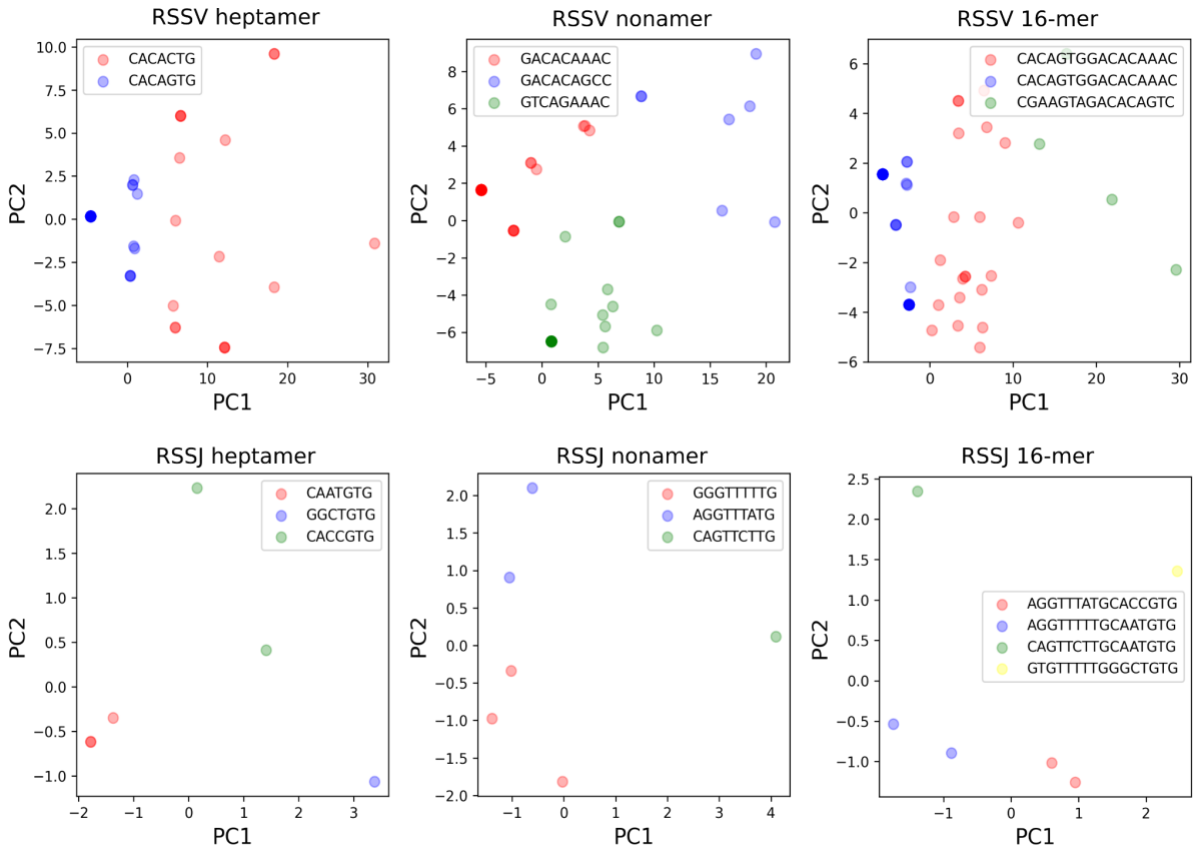
Supplemental Figure S9. The phylogenetic trees of candidate V genes in the spear-nosed bat genome. Phylogenetic trees of productive annotated (white) and unannotated (green) V gene candidates detected by BlindIGDetective for the IGH, IGL, and TRA loci. Trees were computed using the Clustal Omega tool (Sievers et al., 2011).



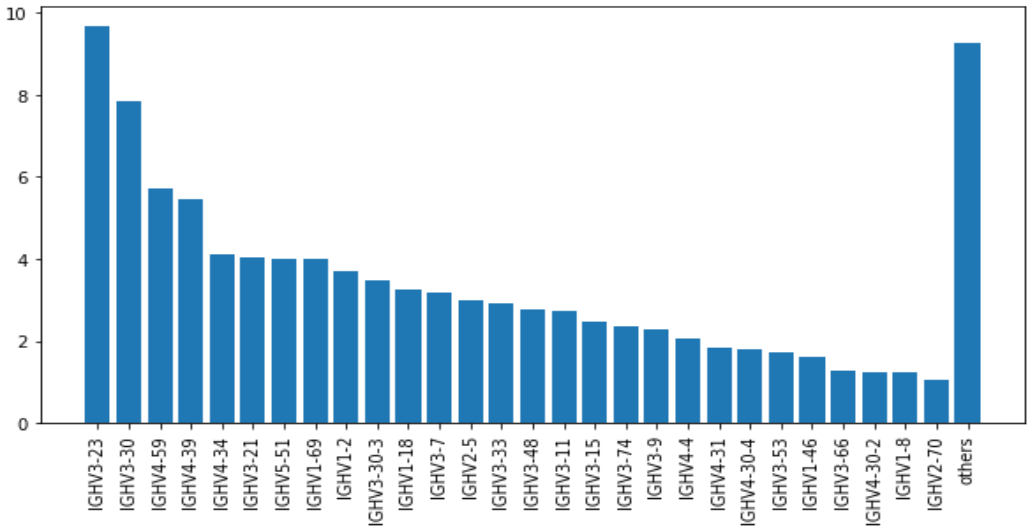
Supplemental Figure S10. Analysis of RSSV nonamers across reference and target species. (A) The principal component analysis on the distance matrix computed for 150 distinct nonamers in RSSVs. Clusters C1–C3, computed using the *k*-means clustering, are shown in red, blue, and green. **(B)** Motif logos of nonamers in RSSVs for clusters C1–C3. The pieplot shows fractions of nonamers corresponding to cow, human, and mouse V genes in the cluster C3 formed by diverse nonamers.



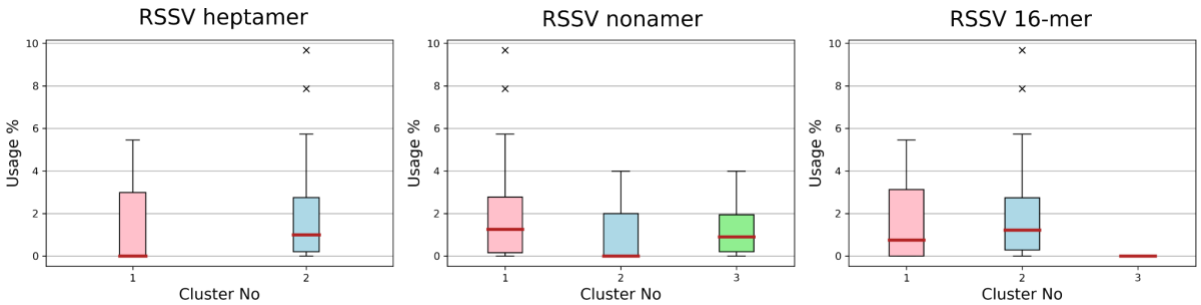
Supplemental Figure S11. The diversity of V gene RSS nonamers. (A) The average RSS distance vs the number species distance for five out of seven orders containing more than one species. (B) The principal component analysis on the distance matrix computed for 129 distinct nonamers of V gene RSSs. Clusters C1–C4 computed using the *k*-means clustering are shown in red, blue, green, and orange. (C) Fractions of each of seven orders in clusters C1–C4. (D) Motif logos of nonamers in RSSVs in clusters C1–C4. Red bars show positions of the AAA-motifs.



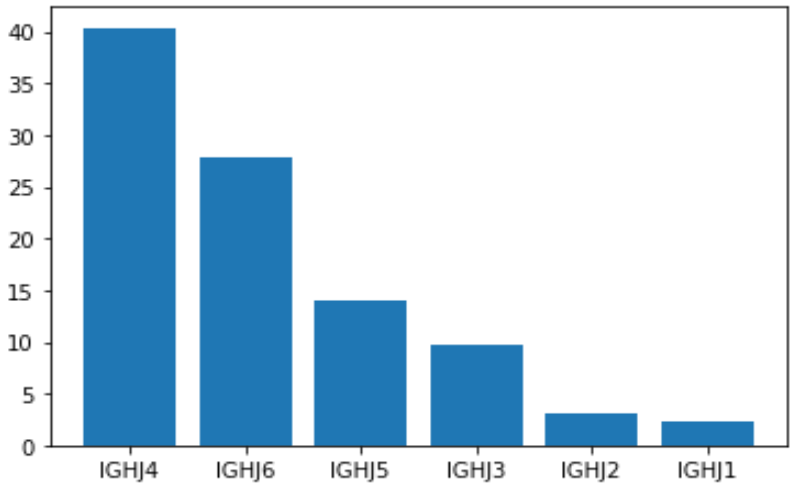
Supplemental Figure S12. Clustering of human IGH signals. The top and bottom rows show RSSVs and RSSJs, respectively. Heptamers, nonamers, and combined 16-mers are shown as left, middle right columns, respectively. We refer to the red, blue, green, and yellow clusters as clusters 1, 2, 3, and 4, respectively.



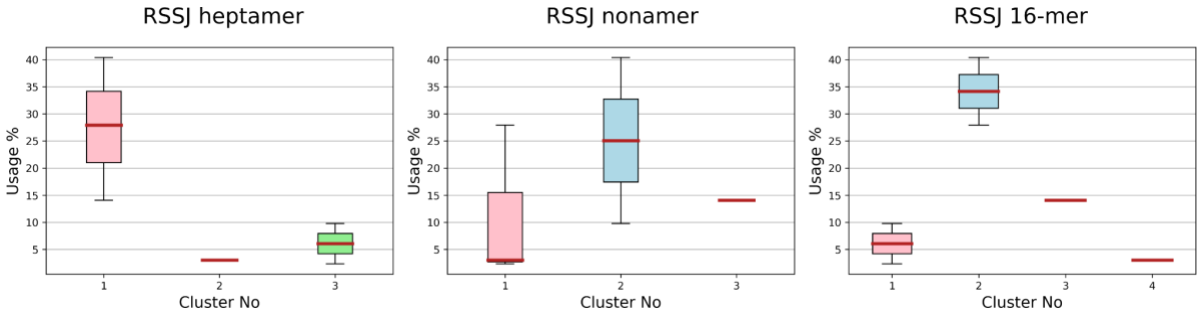
Supplemental Figure S13. Usage of human V genes.



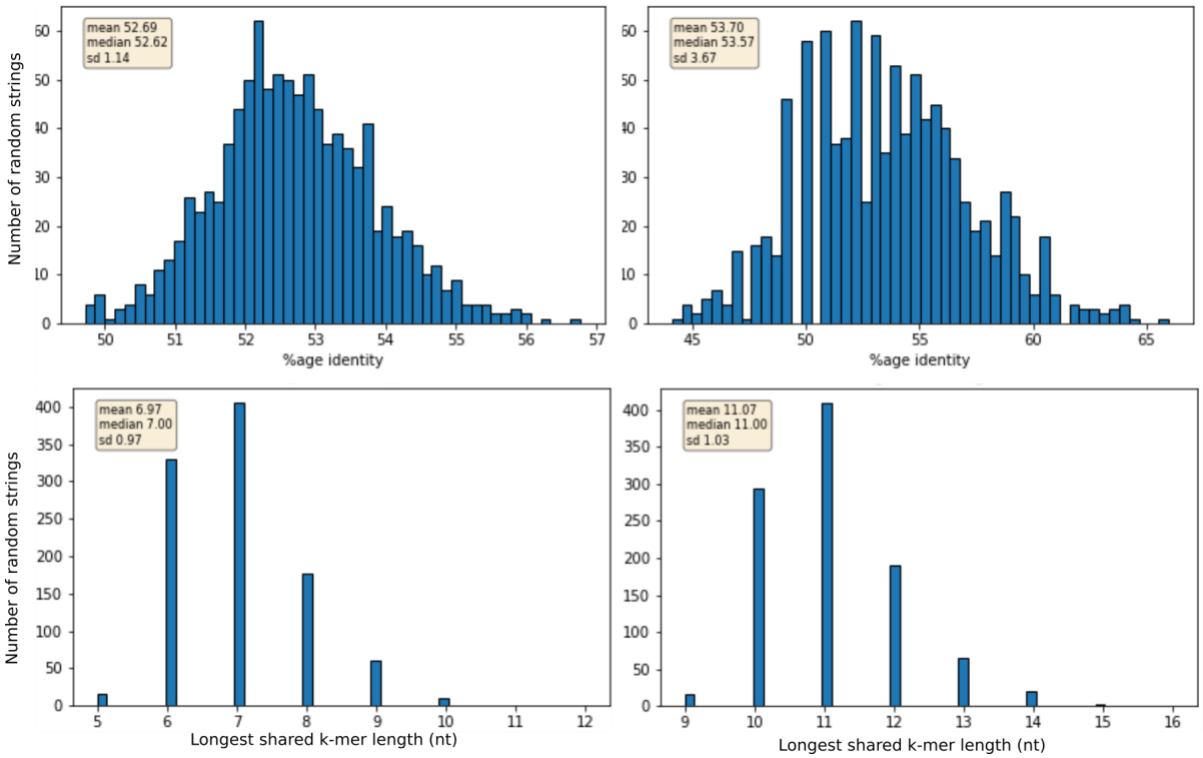
Supplemental Figure S14. The usage distribution per cluster for heptamers (left), nonamers (middle), and combined 16-mers (right) within RSSVs. Cluster ordering and coloring is the same as the ordering and coloring in the Supplemental Figure S12 for the V signals and *k*-mer size (7, 9, 16).



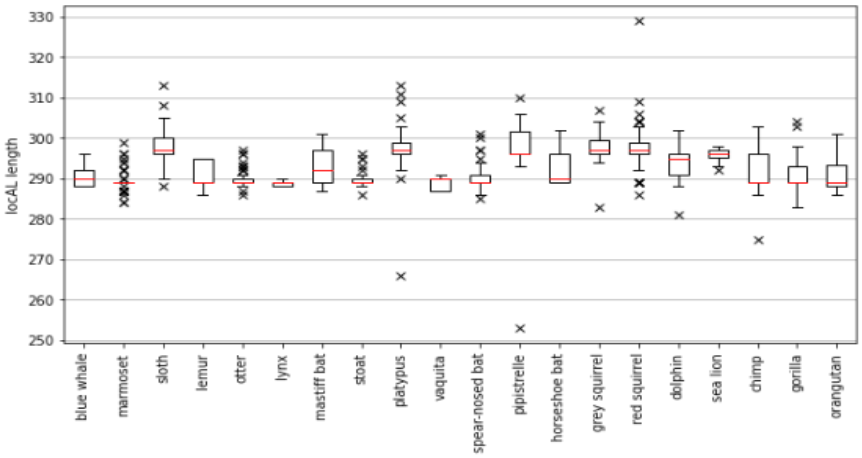
Supplemental Figure S15. Usage of human J genes.



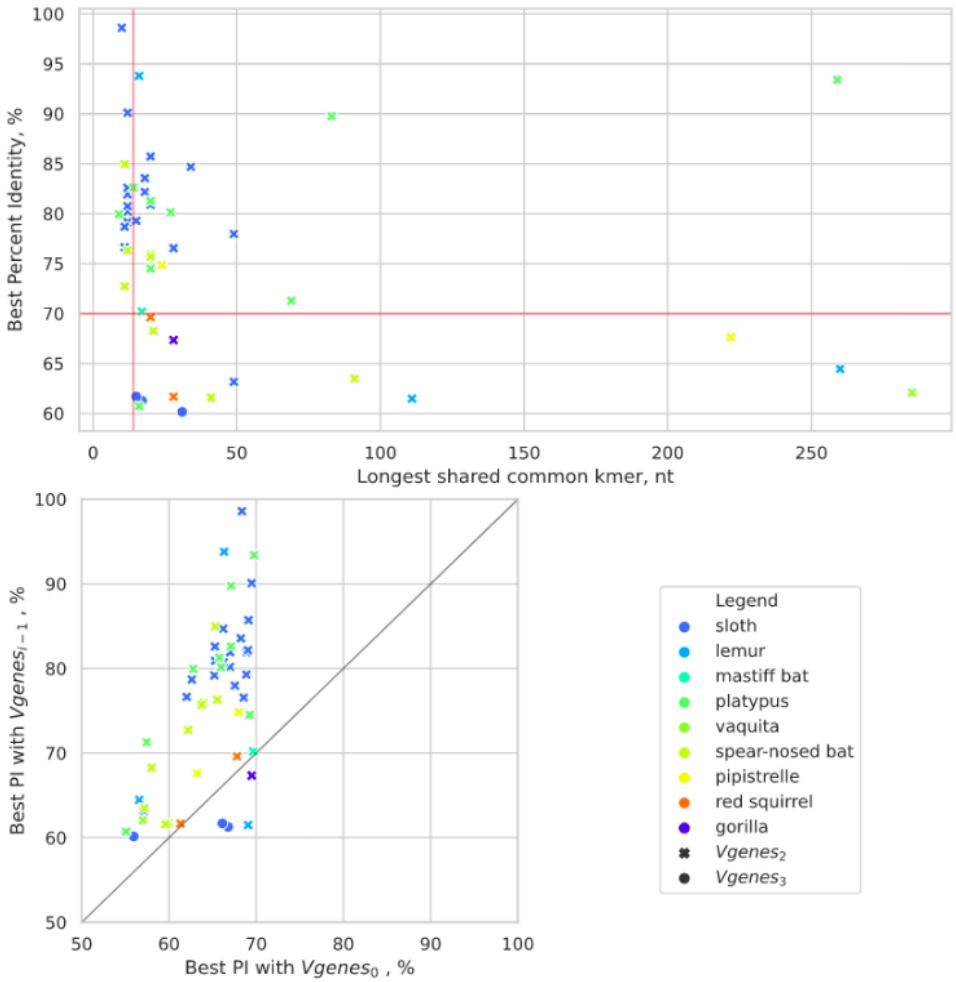
Supplemental Figure S16. The usage distribution per cluster for J signals. Heptamers, nonamers, and combined 16-mers are shown in left, middle, and right, respectively.



Supplemental Figure S17. Distribution of alignment percent identity and longest shared *k*-mer between random strings and their closest V and J genes. The left column shows the distribution of percent identity of 1000 random strings of length 70 against their closest human J gene (Top) and the length of the longest shared *k*-mer between these strings and the best aligned human J gene (Bottom). The right column shows the distribution of percent identity of 1000 random strings of length 350 against their closest human V gene (Top) and the length of the longest shared *k*-mer between these strings and the best aligned human V gene (Bottom). Inside the figure the mean, median and standard deviation (sd) are noted for each distribution.



Supplemental Figure S18. Distribution of alignment lengths for candidate V genes from all target species. Each box plot represents the local alignment lengths of the RSSV-flanking regions containing candidate V genes and its closest canonical human V gene. The alignment lengths are comparable to the mean length of canonical human V genes (306 nt).



Supplemental Figure S19. Alternate similarity thresholds for detection of V genes. (Top) Percent identities (*y*-axis) and longest shared common *k*-mer (*x*-axis) between target-like candidate V genes (predicted in the i^{th} iteration of IterativeIGDetective) and candidate V genes predicted in the $(i\text{-}1)^{th}$ iteration of IterativeIGDetective). Horizontal (vertical) red line shows the percent identity (longest shared *k*-mer) cutoffs used in IterativeIGDetective. (Bottom) For each target-like V gene predicted in iteration *i*, we show the percent identity with the closest V gene predicted in the previous iteration *i-1* (*y*-axis) and the closest canonical human V gene (*x*-axis). Note that many candidate V genes have low percent identity with human V genes but significantly higher percent identity with previously predicted target-like or human-like V genes, and thereby can only be captured through iteratively running IterativeIGDetective.

Supplemental Methods

Supplemental Method: Annotating IG genes in reference genomes

Given a list of IG genes and the IGH locus in a given species, we mapped these genes to the IGH locus using the Bowtie2 tool (Langmead and Salzberg, 2012) with default parameters (match score=0, mismatch penalty=6, affine gap open penalty=5, gap extension penalty=3). We classify an alignment of an IG gene to the IGH locus as *valid* if its score exceeds the default Bowtie2 threshold of $-0.6-0.6L$, where L is the length of the gene. This scoring typically only reports alignments with percent identity (PI) exceeding 90%. Here, percent identity is defined as the percentage of positions in the alignment where the reference and the gene share the same nucleotide.

We apply a filter at this stage by considering the best alignment for each gene and remove those genes that did not have even a single valid alignment. We then identify genes whose alignments are found on the same strand and have starting positions within a distance of 100 nucleotides to other gene(s) and consider only the first (on the 5' end) among them while discarding the other genes. We remove the

gene having fewer allelic variants in the dataset (ties are resolved arbitrarily). **Supplemental Table S2** summarizes the alignment statistics and the number of genes considered after every stage of filtering.

Supplemental Method: Identifying putative IGH-contigs in a genome assembly

To identify all contigs containing fragments of an IGH locus in an assembly, we align all human immunoglobulin IGHV and IGHJ genes against all contigs in this assembly using Bowtie 2 (Langmead and Salzberg, 2012). This step serves as a filter to only identify contigs potentially overlapping with the IGH locus to be considered by IterativeIGDetective for a more sensitive IGH gene prediction. Although Bowtie 2 (designed for finding highly similar sequences) may miss some IGH-contigs that contain only highly-diverged IG genes, these contigs will be revealed by BlindIGDetective that searches the entire genome using a more sensitive local alignment approach. The number of IGH-contigs for 20 target species varies from 3 to 23 (**Supplemental Table S4**) and, for most of the species, the IGH locus was not assembled into a single contig.

Supplemental Method: Extended benchmarking of IterativeIGDetective

Supplemental Table S7 illustrates that, as expected, the F1 score, TPR and FDR for a target species are maximized when IterativeIGDetective uses the reference profile from the same species. The second-best result is usually achieved when IterativeIGDetective uses the reference profile from the combined species. This result is intuitive since, for evolutionary distant species like human, mouse, and cow, the profile of the combined species is typically closer to a target profile than the profile of any other species different from the target. The effect is particularly strong for the mouse target that has poor F1 scores when we use the human and cow references (**Supplemental Table S7**). This observation is especially evident when none of the canonical signals pass the L_{min} threshold for the V(J) signals when we use the human (cow) reference. The converse is true as well, when we use the mouse profile as the reference and the human and cow IG loci as a target. This result stems from the fact that humans and cows have similar RSS profiles that differ from the mouse RSSs profiles.

We note that when IterativeIGDetective identifies V and J genes based on the candidate RSSs, it can potentially remove previously detected true positive and false positive candidate RSSs. In practice, this step only slightly reduces the TPR while significantly reducing the FDR. The F1 score for human RSSs significantly improves, for cow RSSs remains the same, and for mouse RSSs slightly decreases, leading to an overall increase in F1 score over the combined dataset.

Supplemental Method: Analysis of RSSs in reference and target species

Understanding the similarities and differences in the RSSs is paramount to discovering novel IG genes. We observe that 1, 79, 19 and 48 heptamer patterns passed the candidate heptamer thresholds for V, D_{left}, D_{right}, and J genes. Similarly, 690, 317, 625, and 41 nonamer patterns passed the candidate nonamer thresholds. **Supplemental Table S8** shows that for twenty target species, we identified 1, 13, 10, and 12 distinct heptamers for V, D_{left}, D_{right}, and J genes, respectively. We also identified 82, 21, 34, and 17 distinct nonamers for V, D_{left}, D_{right}, and J genes, respectively.

Supplemental Method: Comparative analysis of candidate IGHJ genes

IGDetective reported 174 candidate J gene candidates for twenty target species (the number of predicted IGHJ genes varies from 4 to 21 per target species). Since the number of J genes in the reference species is small (6 in humans, 4 in mice, and 12 in cows), we applied additional filters to discard potential false positives among candidate J genes in target species. **Supplemental Figure S2**A shows that the distribution of percent identities of J gene candidates with respect to the closest human J gene is bimodal where the "high PI" ("low PI") mode is formed by J genes with percent identities at least 75% (below 75%). J genes of three reference species are also highly conservative and are characterized by a conserved tryptophan-encoding codon TGG that indicates the end of CDR3. Multiple alignment of 174

J gene candidates revealed that only 60 (34%) of them have the tryptophan-encoding codon at the same position in the alignment. **Supplemental Figure S2**B shows that percent identities of 59 out of these 60 J gene candidates come from the "high PI" mode. Until Rep-seq data for target species is available, it is unclear if 114 J gene candidates without the conservative tryptophan-encoding codon can generate productive heavy chains.

We also noticed that two candidate J genes (one for chimpanzee and one for gorilla) start with six tyrosines (**Supplemental Figure S2**C). For all other candidate J genes, the number of starting tyrosines does not exceed 2. The only reference J gene with a similar feature is human J gene IGHJ6 that starts with 5 tyrosines (**Supplemental Figure S2**D). Avnir et al., 2014 showed that the poly-Y prefix of IGHJ6 plays an important role in antibodies neutralizing flu viruses. We assume that similar J genes in the chimpanzee and the gorilla may indicate that they also participate in antibody responses to flu viruses.

Supplemental Method: Computing boundaries of the IGH loci using predicted IG genes

To identify bounds of the IGH loci in a target species, we combine positions of V, D, and J genes predicted by IterativeIGDetective with positions of the identified C genes. Two genes are linked if the distance between them does not exceed the distance threshold (the default value 1 Mbp). We identify groups of co-located IG genes using single linkage clustering of linked genes and consider the largest group as the IG genes in the candidate IG locus. For each species, we apply this procedure for all contigs containing IG genes. We also compute the total length of the IGH locus as the sum of the lengths of IG-segments across all contigs containing IG genes.

Supplemental Method: Unusual IGH locus in the sloth genome

In the sloth assembly, we found a D gene and two J genes inside the V region resulting in a surprising $V_1 \rightarrow J \rightarrow D_1 \rightarrow V_2 \rightarrow D_2 \rightarrow C$ arrangement, where V_1 and C genes are located on the direct strand and V_2 and J genes are inverted (**Supplemental Figure S3**). Since D genes have reverse-complementary RSSs and do not have a fixed ORF, we cannot reliably identify the direction of D_1 and D_2 genes. If the assembly of the sloth IGH locus is correct, then it will likely result in a highly skewed usage of V and J genes in expressed antibody repertoires. For example, V_2 genes cannot be recombined with J genes. In this case, we would also assume that IterativeIGDetective missed D genes between V_1 and J genes (that can be recombined with both) and J genes between D_2 and C genes (that can be recombined with all V and D genes). Re-assembly of the sloth IGH locus will be needed to check whether the inversion of V2-D1-J represents an unusual configuration of the IGH locus or an assembly artifact.

Supplemental Method: Analyzing connected components in the similarity graph

Given a set of predicted RSSs in a genome, its RSS-based similarity graph (*S*-graph) is formed by vertices representing *s*-fragments and edges connecting similar *s*-fragments. Two *s*-fragments are similar if the percent identity of their local alignment exceeds a threshold pi_{min} (default value 70%). The edge-weight of an edge (*s,s'*) is defined as the percent identity between the *s*-fragment and the *s'*-fragment. The scoring parameters for alignment are identical to those used in IterativeIGDetective described in **Supplemental Method "Parameter selection for identifying candidate IG genes".** Since construction of the *S*-graph becomes time-consuming when the number of candidate RSSs exceeds 10,000, **Supplemental Method "Speeding-up BlindIGDetective"** describes a speedup mechanism to guide the formation of *S*-graphs.

The *density* of an *n*-vertex connected component in the similarity graph is defined as the number of its edges divided by $n \times (n-1)/2$. We define *percent identity* of a vertex in the similarity graph as the highest weight among all edges incident to this vertex. The *percent identity* (*coding length*) of a connected component in the similarity graph is defined as the median percent identity (coding length) of its

vertices. A connected component is *uniform* if its percent identity is at least *minPI* (default value *minPI*=80%).

Given a set of *predicted genes* in one species (e.g., all genes identified by IterativeIGDetective or all canonical genes), we define the *conservation* of a vertex in the similarity graph of another species as its percent identity with the closest *predicted* gene. The median (minimum) conservation of a connected component is defined as the median (minimum) conservation of its vertices and is denoted as $conservation_{min}$ ($conservation_{med}$).

Given a set of *predicted genes* in a target species (e.g., all genes identified by IterativeIGDetective or canonical genes), we classify a vertex *s* in the similarity graph as *annotated* if it shares a percent identity of at least $PI_{annotation}$ (default value = 90%) with some predicted gene. All other vertices are classified as *unannotated*. The *annotation index* of a connected component is the fraction of annotated vertices in this component. Similarly, we define a more relaxed version of the *annotation index* called *annotation80 index* as the fraction of annotated vertices given $PI_{annotation}$ = 80%. A connected component in the similarity graph is *annotated* if it contains annotated vertices, and *unannotated*, otherwise.

We extend these definitions of density, percent identity, coding length, annotation index and conservation of a connected component to both clumps and clusters.

Supplemental Method: Speeding-up BlindIGDetective

The formation of *S*-graphs involves identifying edges between similar vertices, where similarity is defined as the percent identity of the alignment between vertices (*s*-fragments). Given *N* vertices (each representing an *s*-fragment of length *L*), BlindIGDetective performs $N^2/2$ alignments on strings of length *L* to form an *S*-graph on these N vertices and becomes time-consuming when *N* is large. For example, construction of the V-graph on 28,394 candidate RSSVs in the human genome involves computing ~400M alignments on *v*-fragments of length *L* = 350.

To speed-up BlindIGDetective, we first perform alignments on suffixes of *v*-fragments of length *l*=100 nt and only proceed with computing full-length alignments of entire *v*-fragments of length *L* = 350 if the percent identity between their suffixes exceeds the PI threshold. Since computing alignments takes quadratic time, this procedure results in $(350/100)^2$ times speed-up to construct the V-graph.

Isolated vertices in the V'-graph will likely remain isolated in the V-graph and can be discarded. Similarly, we discard vertices contained in giant connected components of size greater than 500 in the V'-graph as they likely represent spurious RSSs from repeat regions. We then construct a V-graph on the remaining vertices by performing a full alignment on *v*-fragments of length L. Isolated vertices in the resulting V-graph are discarded as well. In the case of the human reference genome, this speedup results in having to perform an alignment on the entire *v*-fragment on only 980 vertices.

Supplemental Method: BlindIGDetective results on cow and mouse genomes

BlindIGDetective constructed 7 (12) accordant (large) clusters in the cow genome (**Supplemental Table S12**). Five out of 7 accordant clusters revealed all known cow IG and TCR loci (except for the IGK locus), with IGH and IGL loci represented by 21 and 36 vertices, respectively. While all vertices in the IGL cluster were annotated, two unannotated vertices in the IGH cluster are accordant. These vertices, as well as vertices in a single unannotated accordant cluster C5 may represent still unknown IG genes in the cow genome.

BlindIGDetective constructed 44 (8) large (accordant) clusters in the mouse genome (**Supplemental Table S12**). Two largest out of these 8 accordant clusters originated from IGH locus (106 candidates) and TRA locus (53 candidates). Nearly all v-fragments in these clusters are annotated except for 7 (3) genes in IGH (TRA) clusters, with the unannotated v-fragments grouped together in 2 (1) clumps within the IGH (TRA) cluster. The significance of the remaining 6 out of 8 accordant clusters in the mouse

genome remains unclear: four of them are located on the highly repetitive Y Chromosome and thus may be formed by spurious RSSs that fall in repetitive regions.

BlindIGDetective benchmarking on the mouse genome identified only two loci (IGH and TRA) and failed to identify others, such as the IGL locus. It turned out that the v-fragments representing mouse V genes from the IGLV locus got absorbed by the giant component in the mouse V-graph. We thus plan to add analysis of the giant component in the next release of BlindIGDetective.

Supplemental Method: Applying BlindIGDetective to the horseshoe bat genome

BlindIGDetective uncovered 8 large clusters in the horseshoe bat genome (**Supplemental Table S13**). It identified 3 large IGH clusters formed by 6 clumps and 75 *v*-fragments, one large IGL cluster formed by 5 clumps and 53 *v*-fragments, and one large TRA cluster formed by 8 clumps and 28 *v*-fragments. One of the clumps in the TRA cluster contains 2 *v*-fragments that were also mapped to the TRD V genes with median conservation of 75%.

Supplemental Method: Clustering nonamers in RSSVs

Analysis of RSSVs from known V genes in reference species and candidate V genes in target species revealed 150 distinct nonamers in RSSVs. Motif logos for all types of combined RSSs are shown in **Supplemental Figure S7**. To analyze these nonamers, we computed pairwise Hamming distances between them and applied *k*-means clustering to the resulting distance matrix for $k = 2,...,10$ followed by the elbow method analysis (Yuan and Yang, 2019) that reported the optimal number of clusters equal to 3. **Supplemental Figure S10**A shows that the computed clusters C1, C2, and C3 consist of 73, 56, and 21 nonamers, respectively. While clusters C1 and C2 are formed by highly conservative nonamers, the cluster C3 is formed by diverse nonamers from three reference species only (**Supplemental Figure S10**B). Thus, to analyze the interspecies diversity of V gene nonamers, we excluded 21 nonamers from cluster C3 and focused on 129 remaining nonamers. Analysis of 129 conserved V gene nonamers is described below.

To explore the diversity of 129 conserved RSSVs, we first analyzed five orders with multiple target species: Artiodactyla, Carnivora, Chiroptera, Primates, and Rodentia (**Figure 2**A). For each such order, we analyzed all RSSV nonamers corresponding to the species from the order and computed the average Hamming distance over all pairs of such nonamers. Also, for each such order, we computed the average distance over all pairs of species using the phylogenetic tree shown in **Figure 2**A. **Supplemental Figure S11**A shows that these two distances positively correlate (r=0.61, P-value=0.28). Even though the correlation is not statistically significant (because of the small number of observations), it implies that the higher diversity of the species, the higher diversity of corresponding RSSs. This also suggests the importance of exploring less conserved RSSs motifs because they can be order- or species-specific.

To analyze similarities between 129 RSSV nonamers, we computed the pairwise Hamming distances for them and applied *k*-means clustering to the resulting distance matrix for $k = 2,...,10$ followed by the elbow method (Yuan and Yang, 2019) that reported the optimal number of clusters equal to 4 (**Supplemental Figure S11**B). The resulting clusters C1, C2, C3, and C4 consist of 25, 44, 31, and 29 nonamers, respectively. For each nonamer and each biological order, we also computed the total number of times this nonamer was detected in IG genes of all species from this order. Then, for each order, we summed up the computed numbers across nonamers from the same cluster, and, for each cluster, computed fractions for all seven orders. While clusters C1 and C3 are dominated by Primate species, clusters C2 and C4 are dominated by Rodentia species (**Supplemental Figure S11**C). Remaining orders have higher (and similar to each other) fractions in clusters C1 and C3 compared to C2 and C4. **Supplemental Figure S11**D shows that four clusters also differ by positions of the AAA-motif important for the activity of the RAG complex initiating VDJ recombination (Nagawa et al., 1998). While clusters C1 and C3 have the AAA-motif at positions 6–8, it is shifted to positions 4–6 in clusters C2 and C4. We hypothesize that differences in clusters C2 and C4 might represent shorter spacers common for rodents and be associated with differences in RAG proteins.

Analysis of RSSDs and RSSJs revealed correlations like the one shown in **Supplemental Figure S11**B but did not reveal clusters with clear order or motif associations.

Supplemental Method: Cluster-based usage of RSSDs

We have thus far identified clustering membership for various strings from RSSDs: heptamers, nonamers, and combined 16-mer as well as the *l9l7r7r9, l7r7* and *l9r9* combination of the paired RSSDs. To propagate the usage of RSSDs to the cluster they belong to, we compute the sum total, mean, and median of the usage for each cluster for all types of signal strings. We also compute a P-value associated with every signal type (calculated using the Kruskal-Wallis statistical test) that provides a confidence score that the distribution of usages between the clusters came from the same distribution. **Supplemental Table S15** lists all statistics calculated for RSSD usages. We plotted box plots for the significant cases where P-value is below 0.05 in the main text in **Figure 6** and the extended usage distribution statistics in **Supplemental Table S15**.

Supplemental Method: Finding association between RSSs and usages of V and J genes

Supplemental Figure S12 depicts the clustering for human RSSVs and RSSJs genes using the same clustering approach that we described for RSSDs. If we define the distance between two RSSs as the Euclidean distance between their vectors determined earlier, we notice that for RSSVs the inter cluster distance is low and the intra cluster distance is high when compared to RSSDs Even though we have only six human RSSJs, we can see some semblance of distinct clusters for both signal types, although it isn't as distinct as the clustering in D genes.

We acquired the usage statistics for 57 of these 70 V genes left after our filtering step. We note that the remaining 13 V genes were mostly allelic variants which had the same cluster membership for heptamers, nonamers and combined 16-mer. Of the 28 out of 57 V genes have usage exceeding 1% and the remaining 29 genes account for only 9.5% of the total usage (**Supplemental Figure S13**). Unlike the D genes, we cannot clearly see any cluster having the majority share of usage for V genes. The distribution of the usage is evenly spread between the clusters for heptamer, nonamer and 16-mer (**Supplemental Figure S14**). Similar to the D signals we carried out an association test correlating the signal type and the gene usage (calculated using the Kruskal-Wallis statistical test). **Supplemental Table S16** lists all statistics calculated for RSSV usages.

The same process of identifying the usage distribution was carried out for J genes. **Supplemental Figure S15** shows the distribution of usage between all 6 J genes. One of the clusters indeed does exhibit higher usage on average than the other clusters for heptamers, nonamers and combined 16-mer (**Supplemental Figure S16).** However, the sample is too small (only six J genes) to establish a correlation between usage and cluster membership. Similar to the RSSDs and RSSVs, we list all clustering statistics in **Supplemental Table S17**.

Supplemental Method: Parameter selection for identifying candidate RSSs

As described in section "Selection of the likelihood ratio threshold" in Methods, the heptamer L_{min} and nonamer L'_{min} thresholds in each reference species are determined through grid search. A tie in F1 score for different grid search heptamer L_{min} and nonamer L'_{min} pairs is broken by selecting the pair with the smaller heptamer L_{min}. If a tie still occurs, the signal pair (L_{min}, L'_{min}) with the smaller nonamer L'_{min} is chosen.

We launched IterativeIGDetective on all heptamers and nonamers of the reference species selected using the grid search L_{min} thresholds. We identify the L_{min} threshold in the grid search giving us the best F1 score and list these in **Supplemental Table S5** for the four reference species (human, mouse, cow, and combined).

Since the heptamers and nonamers are conserved, the canonical signals of the reference species often have a high likelihood score and are generally expected to pass the selected L_{min} threshold. Conversely,

arbitrary non-signal strings usually have a very low likelihood score and will likely fail to clear the L_{min} threshold. However, this is not always the case as there are RSSs which deviate significantly from the consensus (false negatives) and spurious strings with large likelihood scores (false positives).

Supplemental Method: Identification of candidate V and J genes

Below we focus on identification of candidate V genes (J genes are extracted using the same approach and similar parameters). In order to determine the V gene based on its candidate RSSs, IGDetective considers a nucleotide sequence *G* starting from the 5' end of the RSS extending for a length of $MaxLength_V$ in the 5' direction. It takes into account an observation that the local alignment of the nucleotide sequences of two orthologous V genes typically reveals that these genes have a high percent identity above a threshold pi_{min} over the aligned region. We also observed that, in some cases, when the percent identity falls below pi_{min}, some V genes share a long *k*-mer. Thus, IGDetective classifies two V sequences as *similar* if they (i) align with a percent identity greater than or equal to a threshold pi_{min} or (ii) they align with a percent identity greater than or equal to a threshold pi'_{min} ($pi'_{min} < pi_{min}$) and also share a common *k*-mer of length $MinLength_V$. **Supplemental Method "Parameter selection for identifying candidate IG genes"** and **Supplemental Figure S17** show the alignment parameters as well as the determination of thresholds for identifying similar sequences.

Given a sequence *G* of length of $MaxLength_V$ (starting from the 5' end of the RSS and extending in the 5' direction), IterativeIGDetective constructs a local alignment between *G* and all canonical human V genes. If one of these genes is similar to *G*, it classifies *G* as a *candidate* V gene. After identifying a candidate V gene, IterativeIGDetective identifies its start by including nucleotides at the 5' end of *G* to match the start position of the local alignment with the closest human V gene. Finally, if the amino acid translation of *G* contains a stop codon in the reading frame (reading 5' to 3') for a V gene, IGDetective reclassifies it as a *pseudogene*. **Supplemental Method "Aligning candidate IG genes"** and **Supplemental Figure S18** describe the extraction and alignment of the predicted candidate V genes and pseudogenes.

The described approach for identifying V and J genes filters out many spurious RSS-resembling patterns detected in the previous step. However, it misses an IG gene if its RSSs is not classified as a candidate RSS. A potential approach for improving the true positive rate (TPR) would be to optimize for true positive rate rather than F1 score in the candidate RSS detection step. Although this would come at the cost of detecting many false positive RSSs, we will likely reject them in the follow-up gene extraction phase since they are unlikely to be flanked at the 5' end by human-like canonical V gene sequences. However, we did not pursue this approach since we make use of only the candidate RSSs (without using any information about the previously detected genes) to detect new genes which have deviated significantly from any known gene (see subsection "BlindIGDetective pipeline" in Results) It is thus important that we minimize the number of false positive RSSs to prevent erroneous predictions.

Supplemental Method: Identification of candidate D genes

In contrast to identification of V and J genes, since D genes are very diverse (both inter- and intra-species), we classify each sequence between all *paired* candidate $RSSD_{left}$ and $RSSD_{right}$ as a candidate D gene, without any comparison to known D genes. We further discard candidate D genes that contain a stop codon in all three frames of translation in the 5' - 3' direction. It often happens that the reverse complements of a candidate D gene's $RSSD_{left}$ and $RSSD_{right}$ are simultaneously classified as candidate $RSSD_{right}$ and $RSSD_{left}$ respectively. This classification causes the same region between the RSSDs to be reported twice in both the forward and reverse strand of the genome sequence.

Supplemental Method: Iterative extension of the set of candidate IG genes

IterativeIGDetective extracts candidate V genes from candidate RSSVs in a target species based on whether the RSSV-flanking 5' region passes a similarity threshold with a canonical human V gene, thereby resulting in a set of *human-like* candidate genes. However, some *weakly-conserved* V genes may not pass the similarity thresholds because of high divergence from human V genes. Although reducing the similarity threshold may look like a reasonable way to detect weakly-conserved V genes, **Supplemental Method "Alternate similarity thresholds for detection of V genes"**, **Supplemental Table S18,** and **Supplemental Figure S19** demonstrate that it results in a greatly increased false discovery rate. Thus, to detect weakly-conserved V genes in a target species, we extend IterativeIGDetective and classify a candidate RSSV-flanking region as a candidate gene if it passes the similarity threshold with a canonical human V gene *or* a previously identified candidate V gene in this species. Similarly to Olivieri and Gambón-Deza, 2019, we iteratively repeat this step until no new candidate IGHV genes are added and refer to the newly added gene candidates as *target-like* candidate IGHV genes. By definition, no target-like V genes would be similar to any canonical human V gene whereas human-like V genes will be similar to some canonical human V gene.

Supplemental Method: Parameter selection for identifying candidate IG genes

The maximum and average lengths of all human, cow, and mouse V genes combined is 306 nt and 294 nt, respectively, while the maximum and average length of all human, cow, and mouse J genes combined is 63 nt and 52 nt, respectively. We therefore liberally set the $MaxLength_V$=350 nt and $MaxLength_J$=70.

IGDetective uses local alignment with *affine gap penalties* (Compeau and Pevzner, 2018) to identify similar strings. A string G is classified as a candidate V gene if it aligns with any human V gene with percent identity over pi_{min} (with respect to local alignment with scoring parameters +1 for matches, 0 for mismatches, –1 for gap opening, and –0.5 for gap extension). We determine pi_{min} based on the following heuristic. For any RSS flanking regions of $MaxLength_V$ ($MaxLength_J$) we find the highest-scoring alignment with all the human V (J) genes and compute its percent identity. We set a percent identity threshold, pi_{min}=70% for selecting these flanking regions as candidate V (J) genes. We aligned a thousand randomly generated nucleotide strings of length $MaxLength_V$ ($MaxLength_J$) to the canonical human V (J) genes and found that none of them had percent identity exceeding 57% (68%), with a mean best alignment percent identity of 53% (54%). This observation implies that random strings rarely pass the pi_{min}= 70% threshold thereby minimizing the number of potential false positives. The distribution of best alignment percent identity for the random strings for both human V and J genes is visualized in **Supplemental Figure S17**.

We also observed that canonical genes often contained a common conserved k-mer for rather large values of k. This observation was the motivation to introduce a reduced minimum percent identity threshold (pi'_{min}=60% for V genes and pi'_{min}=65% for J genes) as long as the string G shares a k-mer of length $MaxLength_V$ ($MinLength_J$) with the corresponding best aligned canonical human V(J) genes. When we sample a thousand random nucleotide strings of length $MaxLength_V$ ($MaxLength_J$), we note that very few (<1%), if any of these strings share a common k-mer of length 15 (11) with any of the human V(J) genes. However, we notice that a larger (>1%) number of these strings shared a k-mer of length 14 (10) with the human V(J) genes. This result prompted us to select $MinLength_V$ ($MinLength_J$) to be equal to 15 (11). The distribution of the longest shared k-mer between the random strings and human V and J genes is visualized in **Supplemental Figure S17**.

We now analyze the $MinLength_V$ threshold ($MinLength_J$ threshold is analyzed similarly). For simplicity, we assume that all canonical human V genes have the same length equal to the average length of all human V genes.

Consider an arbitrary string G of length $MaxLength_V$. The expected number of exact matches of length k between this string and all human V genes combined is given by $E(k)=num(V)\times(N-k+1)\times(MaxLength_V-k+1)\times 0.25^k$, where *num (V)* is the number of canonical human V genes (70) and N is the

average length of human V genes (294). Under the assumption that matches of length k between a random string G and all human V follow the Poisson distribution, the probability of getting x matches is:

$$P(x) = \frac{E(k)^x e^{-E(k)}}{x!}.$$

The probability of seeing at least one match of length k between G and all human V genes is given by

$$P(x \geq 1) = 1 - P(0) = 1 - e^{-E(k)}.$$

We now select the length k such that the probability of seeing at least one common k-mer between G and all human V genes does not exceed 0.01. Using the value of $E(15)=6.1\times10^{-3}$, we get the probability of seeing at least one match of length k=15 between a random string G and all human V genes. This probability is much lower than the accepted significance cutoff of 0.01. However, if we consider k = 14, we get a P-value of 0.024, which does not satisfy our P-value threshold of 0.01.

Since for J genes, $num(J)$=6, the average length of human J genes equal to 52, and the $MinLength_J$ equal to 70, we get a P-value=3.598×10^{-3}. These significance values back the empirical result noticed by sampling a thousand randomly generated strings, satisfactorily allowing us to retain the $MinLength_V$ ($MinLength_J$) threshold at 15 (11) nt.

Supplemental Method: Aligning candidate IG genes

We consider the local alignment of the RSSV flanking regions of length $MaxLength_V$ containing a candidate V gene for all twenty target species. We note that the parameters selected for aligning these flanking regions against the closest canonical human V gene results in the local alignment length comparable to the length of the canonical human V gene (**Supplemental Figure S18**). This observation reinforces our choice of alignment parameters since we do not see short spurious local alignments spanning only a small part of the canonical human V genes.

During the extraction of the candidate V gene from the region flanking the candidate RSSV in IterativeIGDetective, we extend the start of alignment of the candidate gene to the start of the closest human V gene. This is done to prevent incorrect translation in candidate V genes. In practice, we note that most alignments start perfectly with the start of the human V gene. In most of the other cases, we had to extend the alignment in the 5' direction by 1 or 2 nucleotides to make it coincide with the start of the human V gene. Alignments in 818 out of 971 candidate V genes (84%) in twenty target species did not need an extension. 135 and 16 alignments required an extension by 1 and 2 nucleotides respectively. The remaining two alignments required extensions by 4 and 16 nt.

Supplemental Method: Alternate similarity thresholds for detection of V genes

We considered an alternative to the iterative approach by attempting to set a non-iterative mode lower percent identity threshold of $pi_{alt} \leq pi_{min}$. We recollect that the default pi_{min} threshold was set to 70% and noted that the lowest percent identity of any target-like V gene shared with a canonical human V gene was 55% (**Supplemental Figure S19**). We therefore benchmarked IterativeIGDetective running in the iterative mode against IterativeIGDetective running in the non-iterative mode with a lenient pi_{alt} threshold of 55%, as well as a slightly more moderate threshold equal to 60 % (which was our original pi'_{min} threshold). **Supplemental Table S18** shows us that running IterativeIGDetective with lower non-iterative pi_{alt} results in decrease in F1 score in human and mouse references. However, in the iterative mode, the F1 score reduces only slightly for mouse while remaining constant for human and cow. An additional drawback of using pi_{alt} would be difficulties in determining its value. Unlike the pi_{min} value which was determined using empirical and theoretical data described in **Supplemental Method "Parameter selection for identifying candidate IG genes"**, it is not clear how to set pi_{alt}. An arbitrarily selected pi_{alt} could lead to predicting many false positive genes in target species. On the other hand, the iterative approach utilizes and adapts to the variation in genes detected in a target species to help identify new V genes in the target species.

Supplemental References

Avnir Y, Tallarico AS, Zhu Q, Bennett AS, Connelly G, Sheehan J, Sui J, Fahmy A, Huang CY, Cadwell G, et al., 2014. Molecular signatures of hemagglutinin stem-directed heterosubtypic human neutralizing antibodies against influenza A viruses. *PLoS pathogens* **10**(5):e1004103.

Krumsiek J, Arnold R, Rattei T. 2007. Gepard: a rapid and sensitive tool for creating dotplots on genome scale, *Bioinformatics* **23**(8): 1026–1028

Nagawa F, Ishiguro KI, Tsuboi A, Yoshida T, Ishikawa A, Takemori T, Otsuka AJ, Sakano H. 1998. Footprint analysis of the RAG protein recombination signal sequence complex for V(D)J type recombination. *Mol Cell Biol* **18**: 655–663.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al., 2011. Fast, scalable generation of high quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**(1): 539.