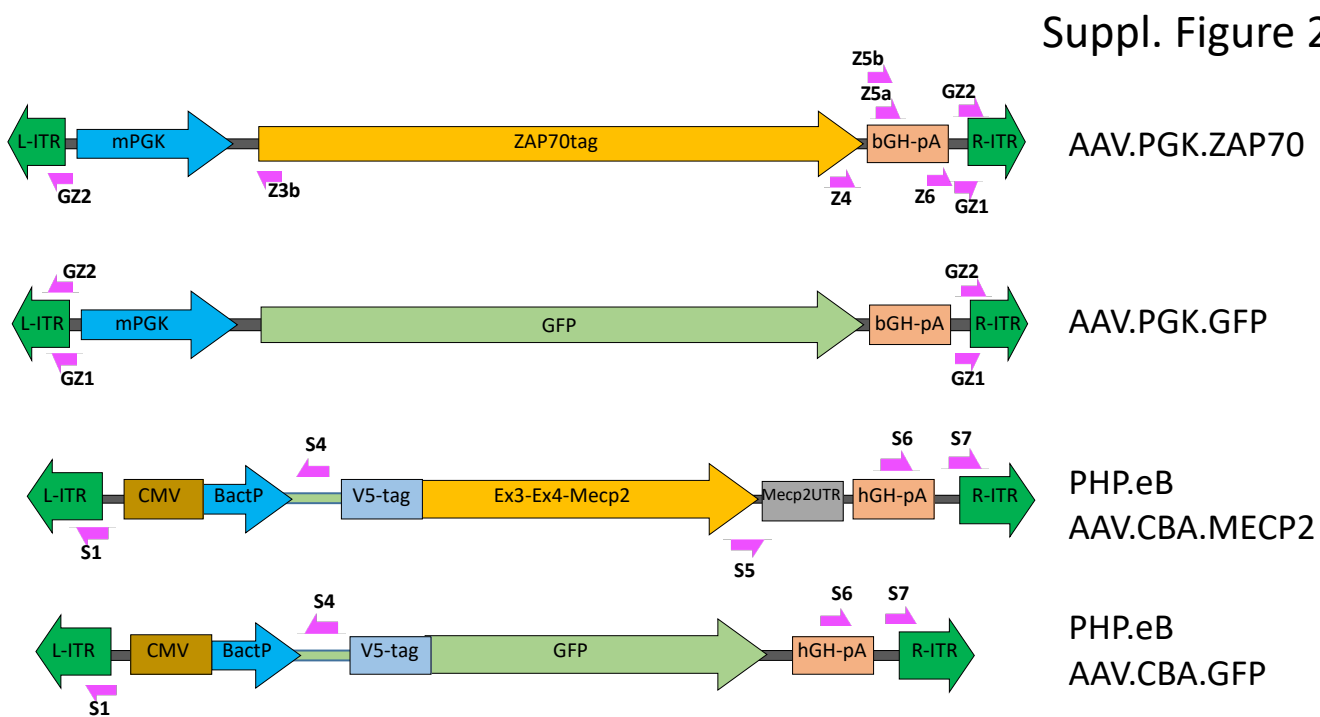
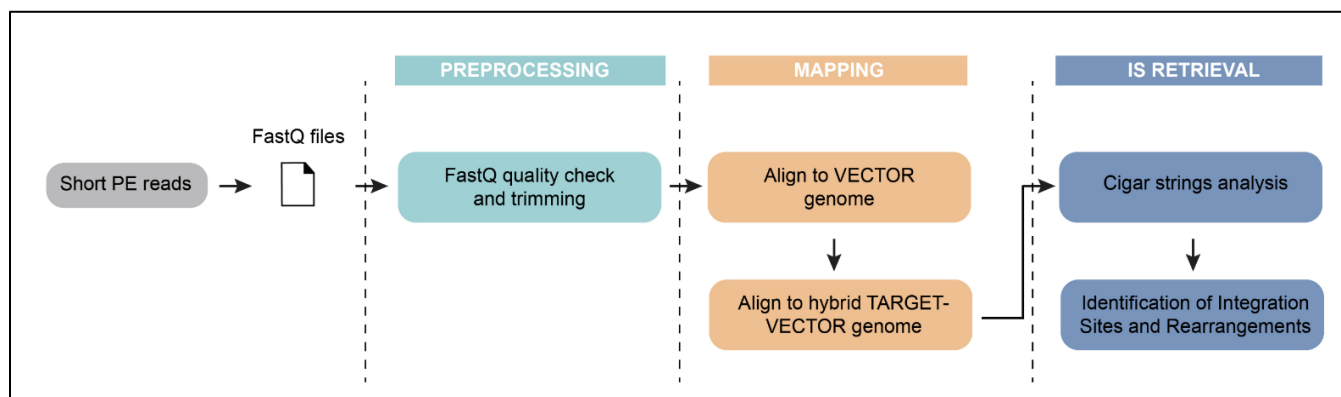


**Supplementary Figure 1: AAV vector copy in liver of Zap70-deficient mice treated with AAV8.** AAV genome copy numbers in liver samples assessed by qPCR in intrathymic AAV8–ZAP-70–transduced mice at the indicated time points. Vector genomes per diploid genome (Vg/Dg) were quantified relative to the albumin gene. White dots and white triangles represent non-injected control mice. The gray area corresponds to the limit of quantification of PCR detection:  $0.000411$  vg/dg.

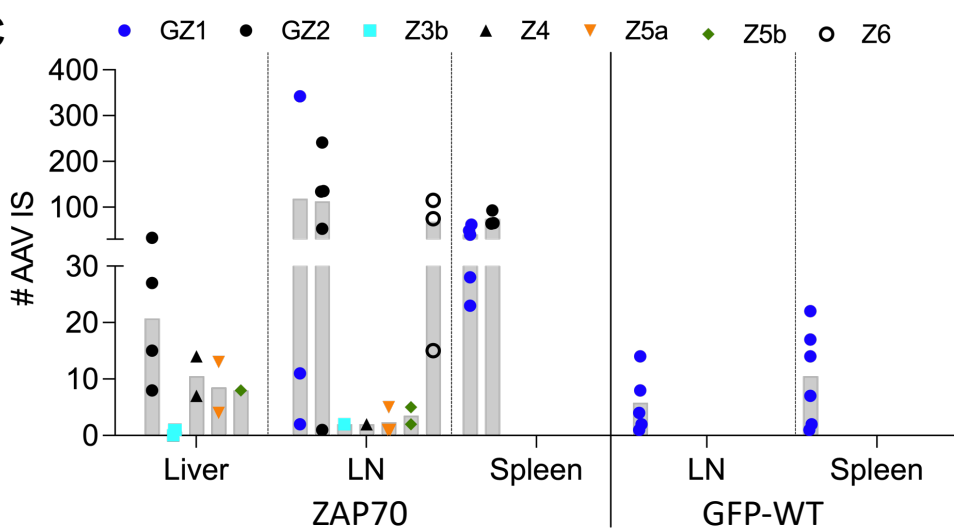
A



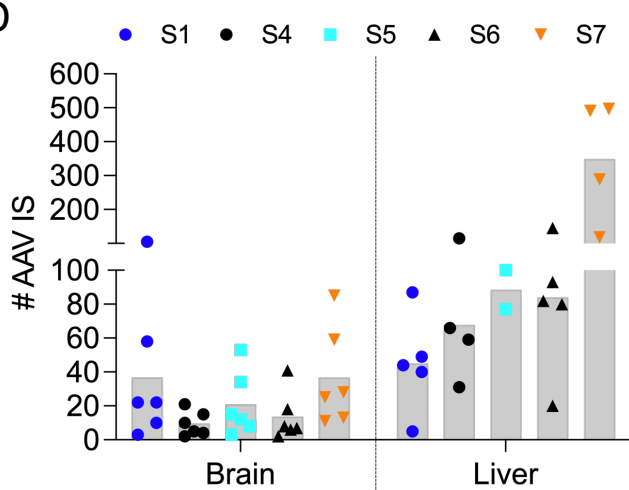
B



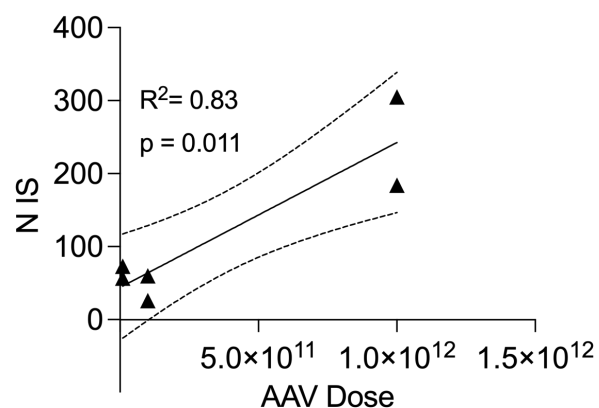
C



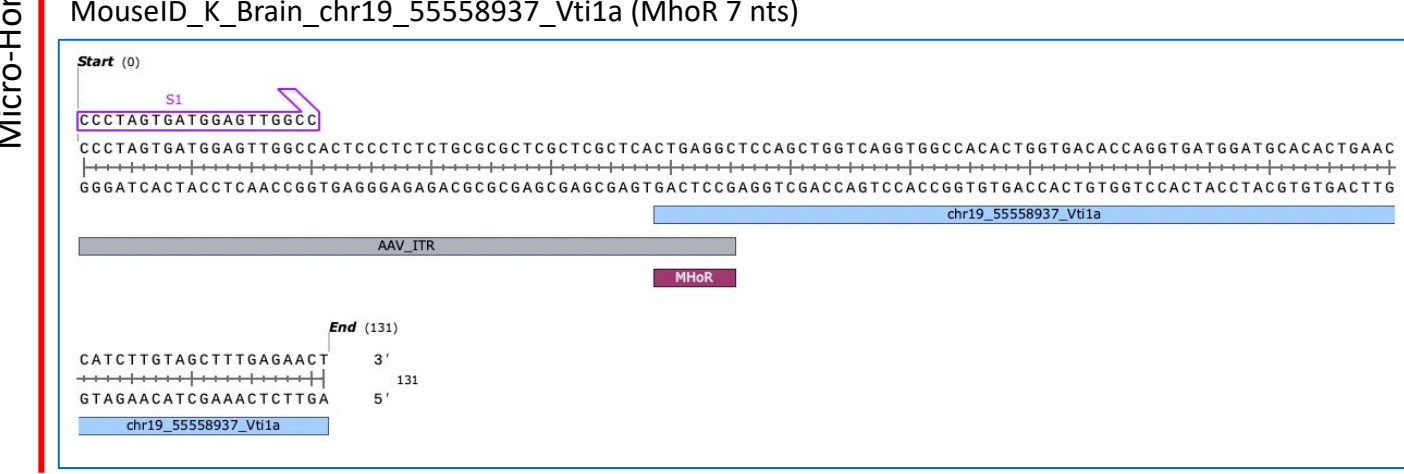
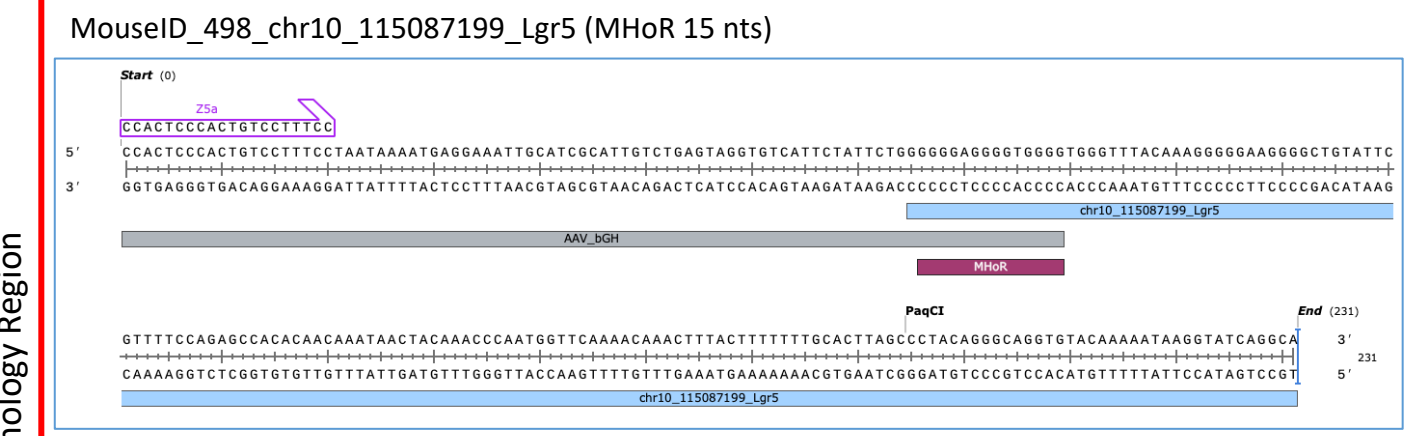
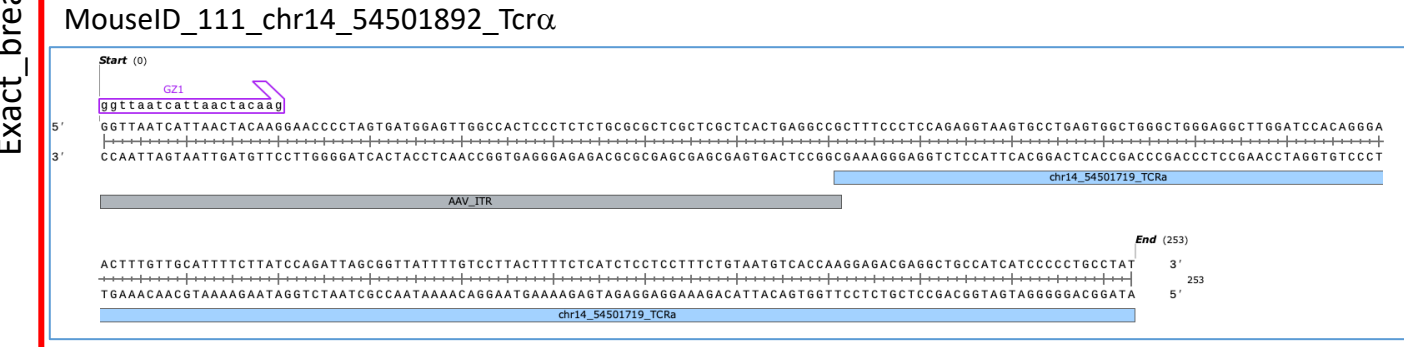
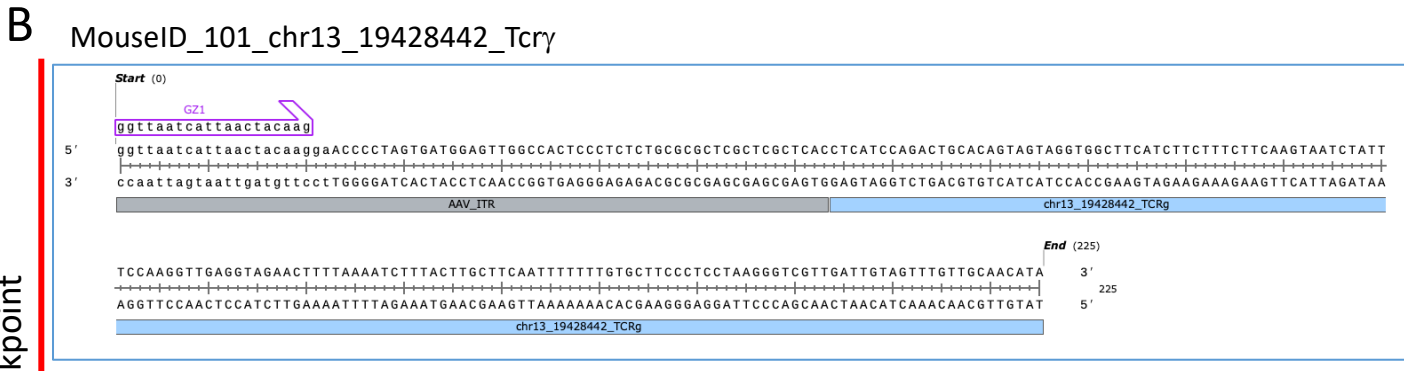
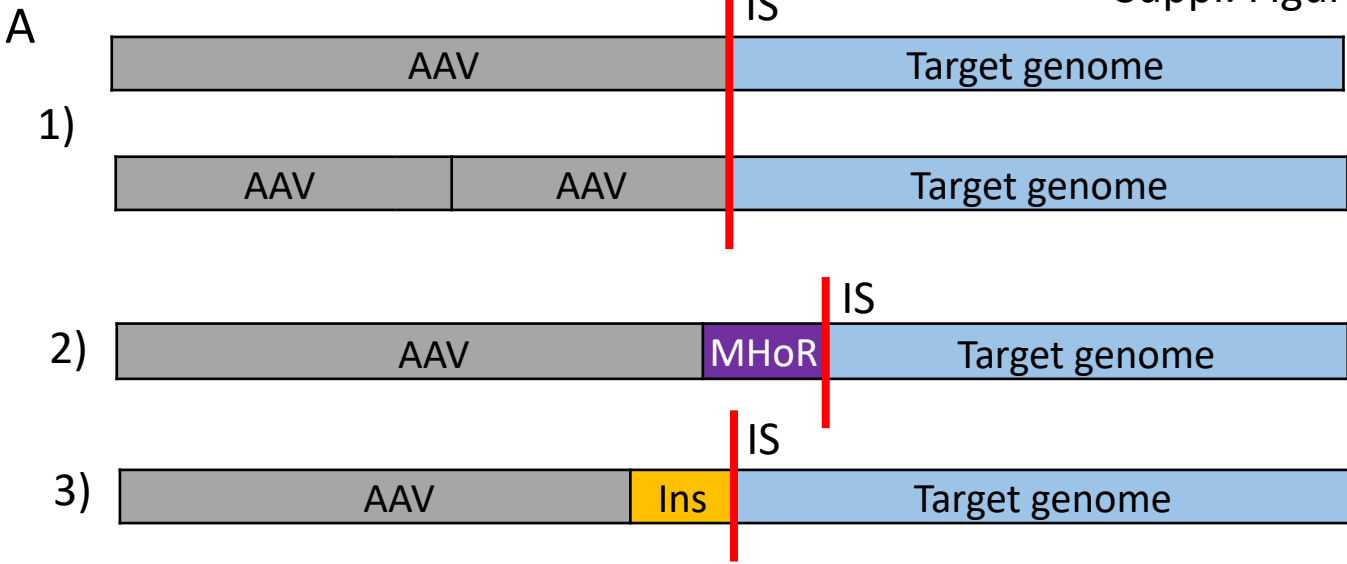
D



E



**Supplementary Figure 2: Retrieval of AAV IS.** A) Schematic representation of the genomic map of the different AAV used in the study and relative position of PCR primer sets adopted for the retrieval of IS from AAV vectors; B) Schematic workflow of the bioinformatics procedures adopted for the identification and characterization of AAV IS; C, D) Number of AAV IS retrieved by each PCR systems in intrathymic injected mice; D) Number of AAV IS retrieved by each PCR systems in *Mecp2*-deficient mice systemically injected with AAV.PHP.eB; E) Positive correlation between the number of AAV IS retrieved in the brain and the dose of vector injected in *Mecp2*-deficient mice ( $1 \times 10^{10}$ ;  $1 \times 10^{11}$  and  $1 \times 10^{12}$  vg/mouse.). Statistics was performed usign Pearson correlation test.



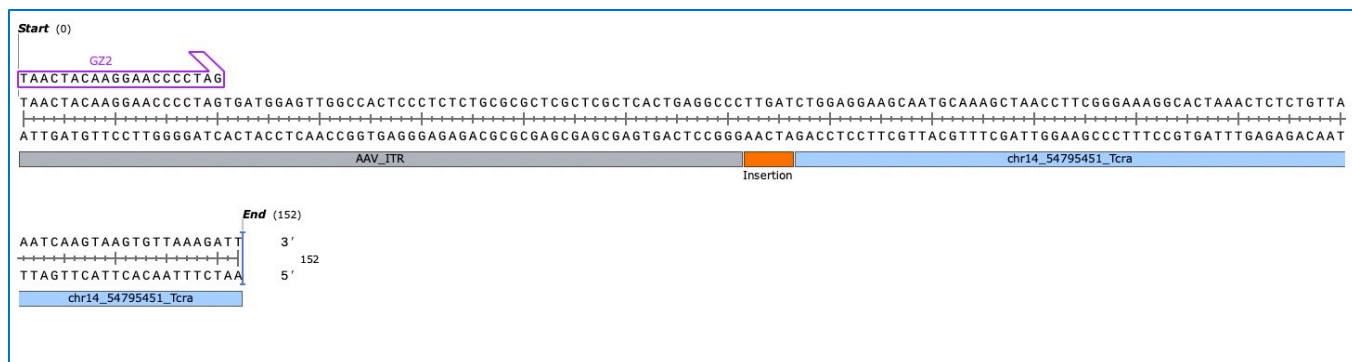
Exact\_breakpoint

Micro-Homology Region

### MouseID\_101\_LN\_chr14\_54794419\_TCRa (12 nts Ins)



### MouseID\_101\_LN\_chr14\_54795451\_TCRa (5 nts Ins)

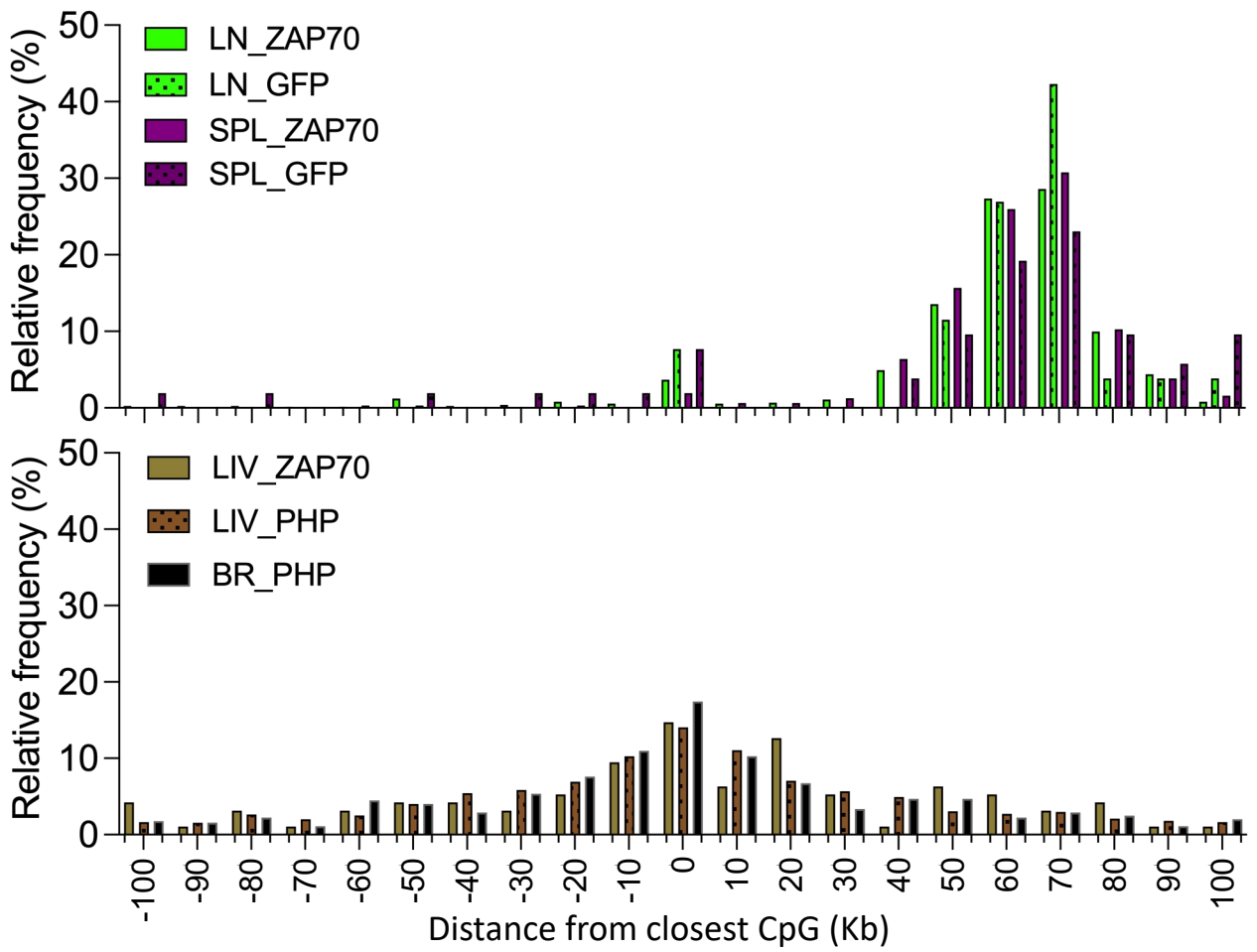


**Supplementary Figure 3: Schema and real examples of AAV/host genome chimeric reads identified by RAAVloli pipeline.** A, B) Schema of the 3 different types of AAV-containing reads identified as AAV IS by RAAVloli pipeline. Nucleotide portion aligning to the AAV genome is indicated in grey, while nucleotide portion aligning on the mouse genome is indicated in azul

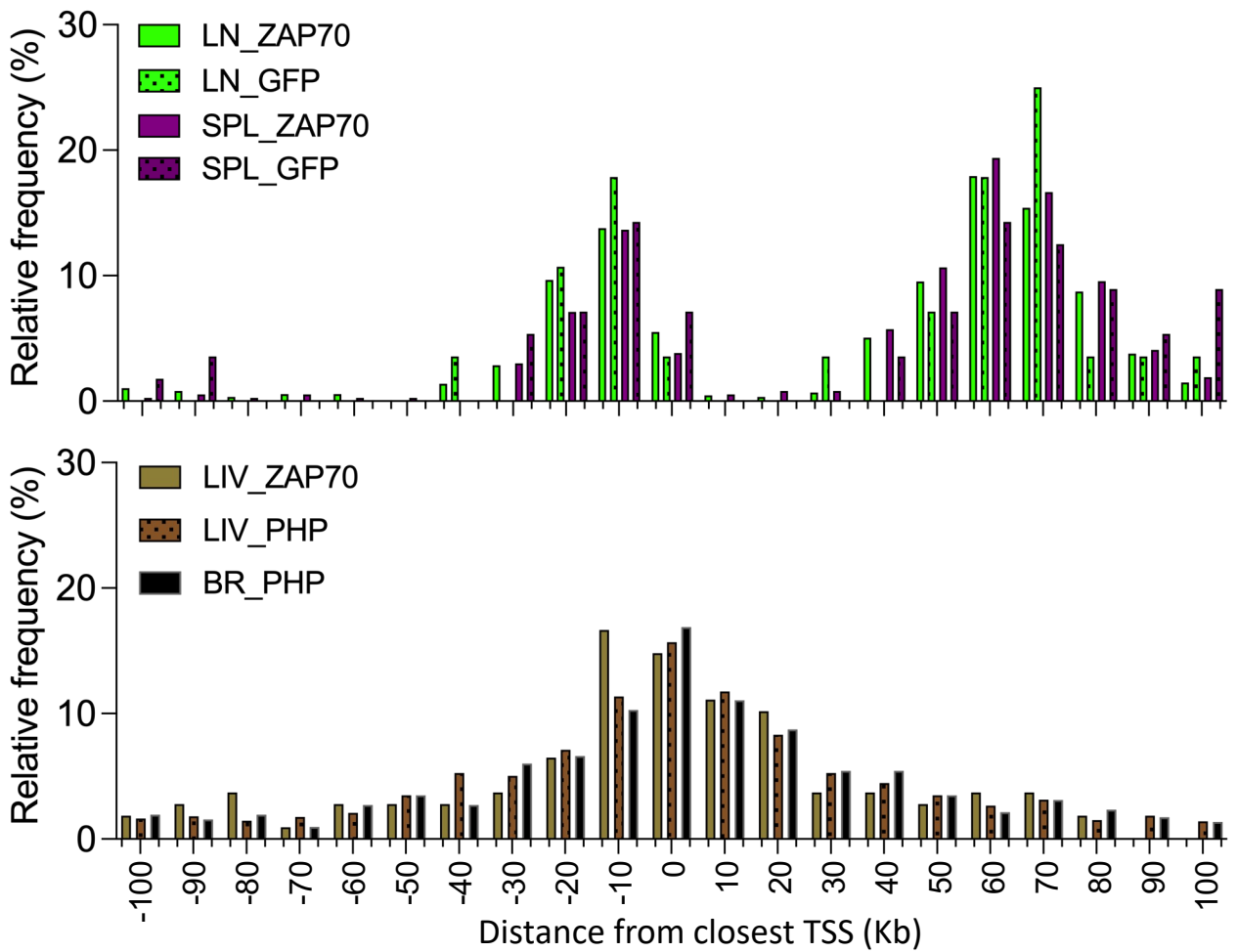
1. AAV IS characterized by a precise homology breakpoint between the vector and the host chromosomal sequences, here referred as (Exact\_breakpoint);
2. AAV IS characterized by micro-homology areas between the vector and the host chromosomal sequences, here referred as (MHoR\_breakpoint), and indicated by the purple rectangle;
3. AAV IS characterized by random nucleotide insertion between the vector and the host chromosomal sequences, here referred as (Insertion\_breakpoint), and indicated by the orange rectangle.

.B) Real sequence examples of chimeric reads where AAV IS were identified. Color code is the same as previously reported.

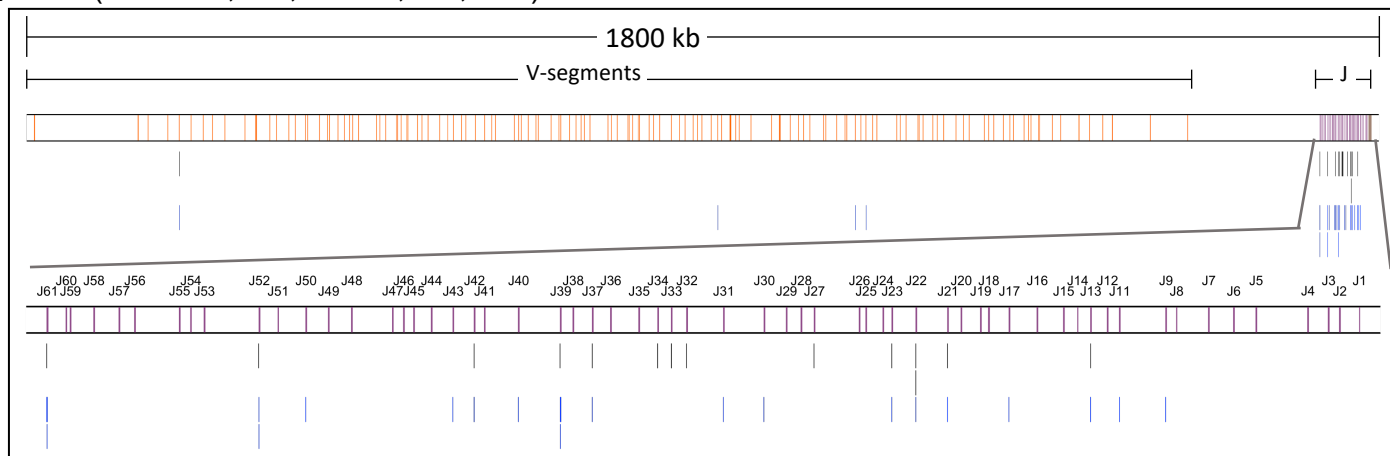
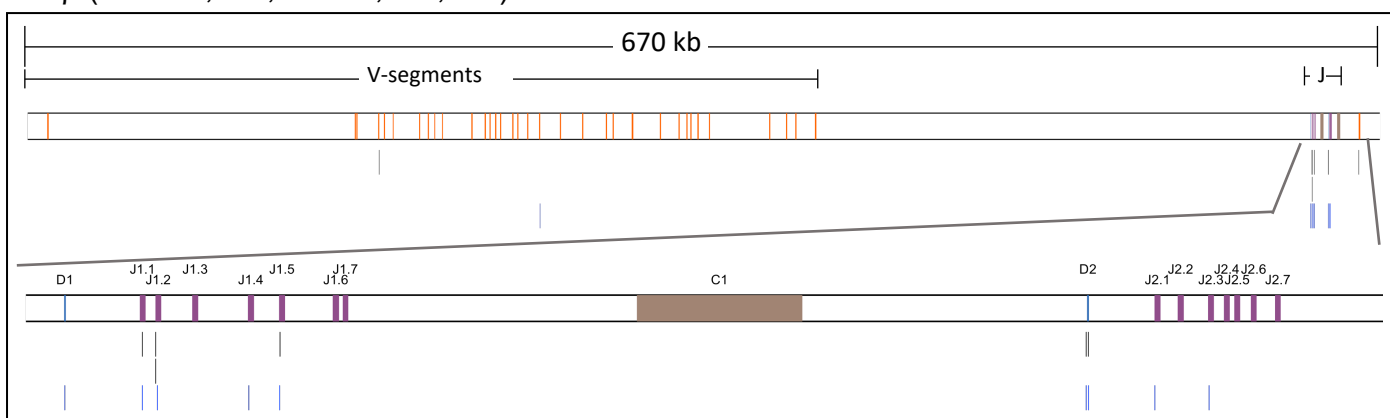
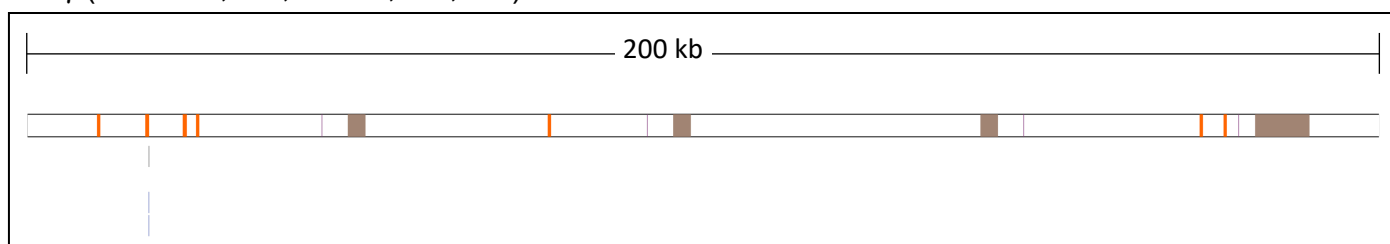
A



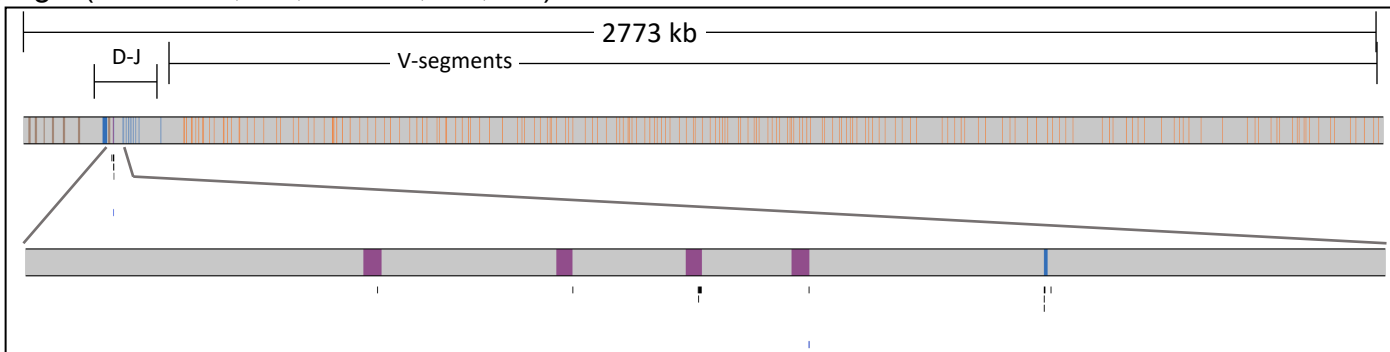
B



**Supplementary Figure 4: Integration site distribution in the host genome. A, B** Integration site distribution around CpG islands (A) and Transcriptional Start site (B) of RefSeq genes for the different IS datasets as indicated.

A Tcr $\alpha$  (chr14:53,047,642-54,843,873)B Tcr $\beta$  (chr6:40,841,295-41,508,370)C Tcr $\gamma$  (chr13:19,269,911-19,444,212)

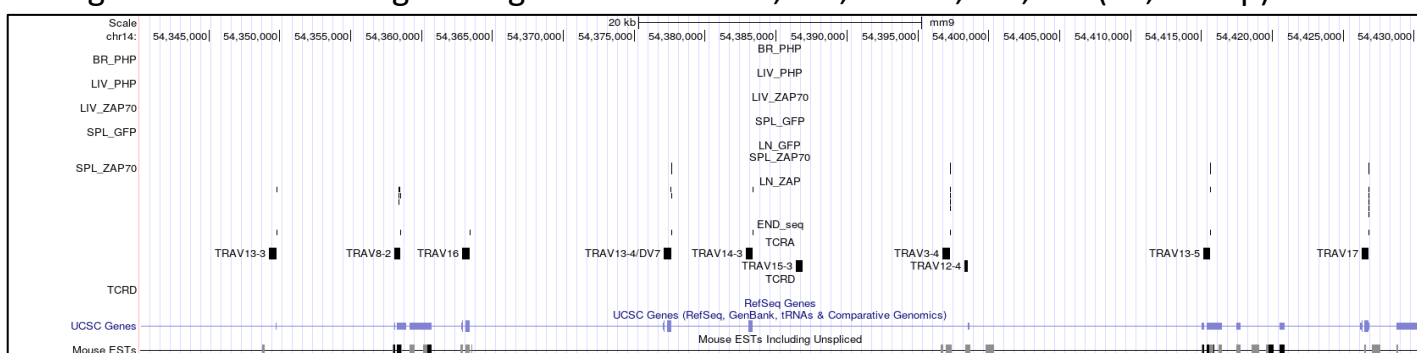
## D IgH (chr12:114,484,415-117,258,165)



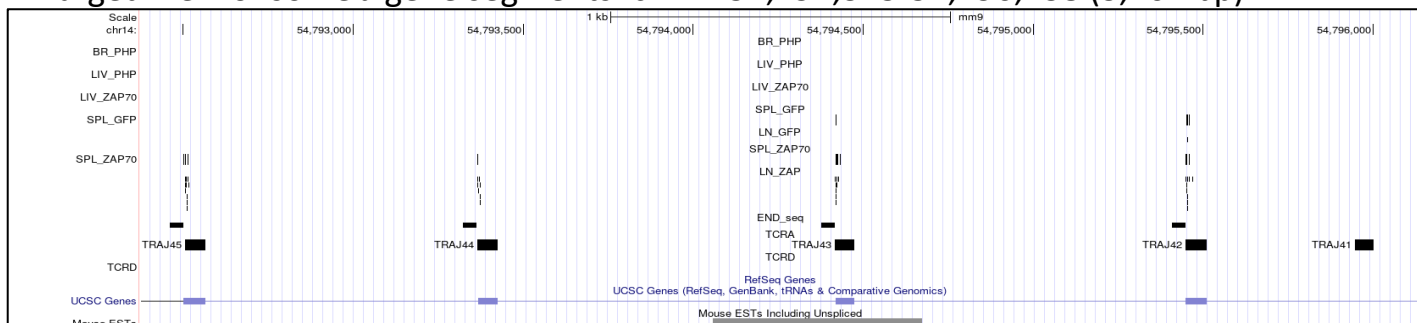
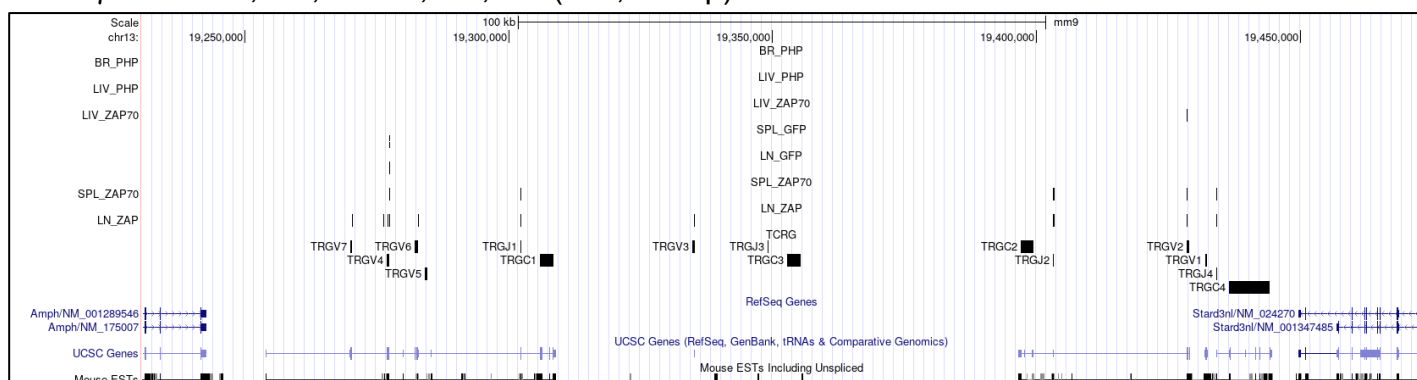
**Supplementary Figure 5: Distribution of AAV IS within Tcr loci.** A-C) Genomic distribution of AAV IS in the LN and SPL datasets of GFP-treated mice targeting TCR genes: TCR $\alpha$  (A), TCR $\beta$  (B) and TCR $\gamma$  (C), as indicated. Genomic coordinates and scale are indicated in each panel. Black and blue lines indicate the position of the AAV IS from the LN (black) and SPL (blue) datasets. Variable (V), diversity (D), and joining (J) segments are indicated by yellow, blue and purple rectangles, respectively. Constant (C) regions are represented as brown rectangles. Clusters of AAV IS are identified in TCR genes, especially in the J- segment region, whose genomic area is enlarged below the Tcr $\alpha$  and Tcr $\beta$  loci; D) Genomic distribution of AAV IS in the LN and Spleen datasets of ZAP70-treated mice targeting IgH gene. Genomic coordinates and scale are indicated in each panel. Black lines indicate the position of the AAV IS. Variable (V), diversity (D), and joining (J) segments are indicated by yellow, blue and purple rectangles, respectively. Constant (C) regions are represented as brown rectangles. Few AAV IS are identified in IgH gene especially in the J- segment region, whose genomic area is enlarged below the IgH locus. Gene segment genomic coordinates were retrieved from the IMGT®, the international ImMunoGeneTics information system® (<https://www.imgt.org>);

A. *Tcr $\alpha$* : chr14:53,037,891-54,894,005 (1,856,115 bp)

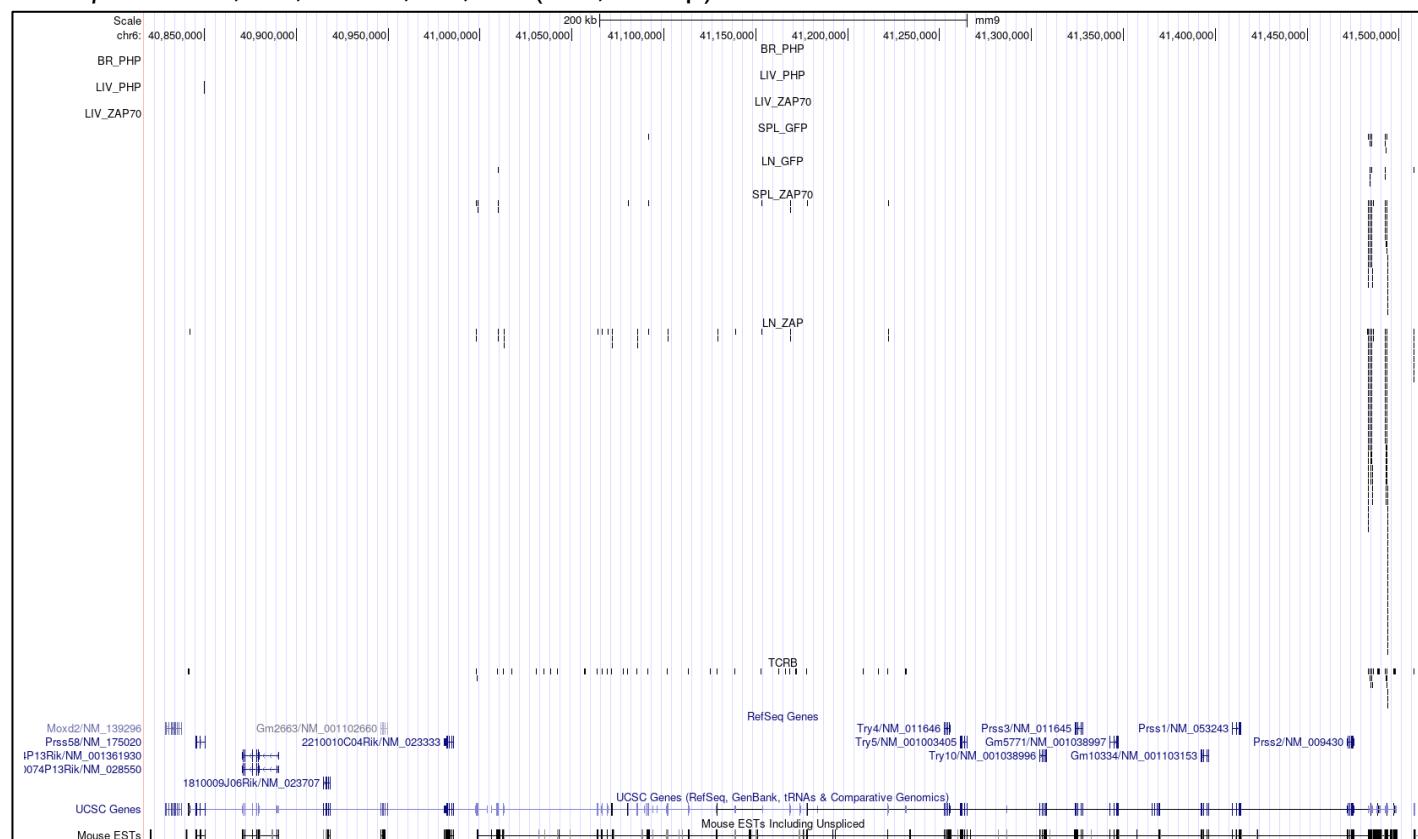
## Enlarged view of some V-gene segments: chr14:54,340,144-54,430,435 (90,292 bp)



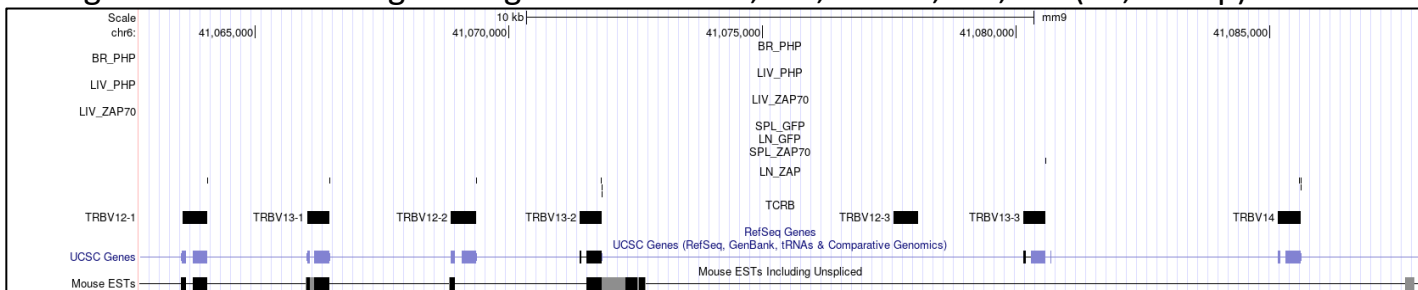
## Enlarged view of some J-gene segments: chr14:54,792,375-54,796,138 (3,764 bp)

B. *Tcr $\gamma$* : chr13:19,230,535-19,472,838 (242,304 bp)

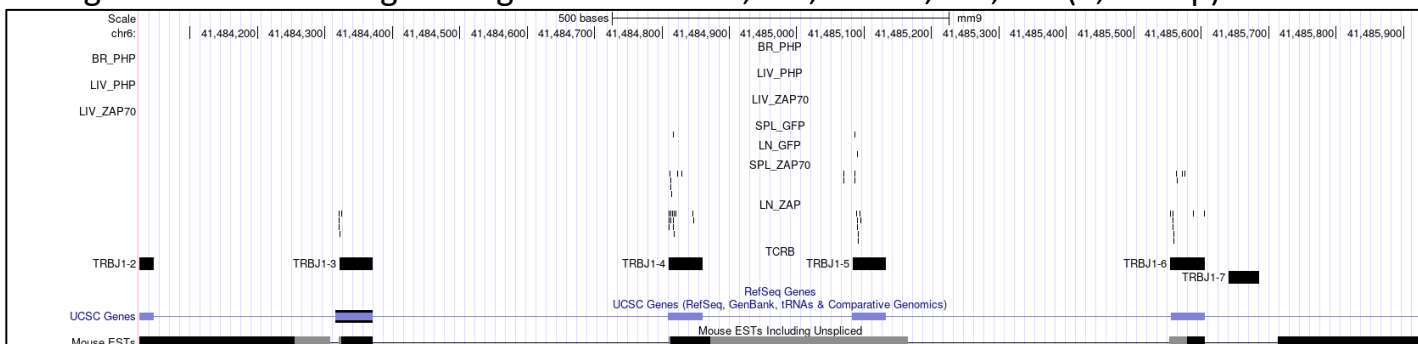


C. Tcr $\beta$ : chr6:40,817,634-41,511,995 (694,362 bp)

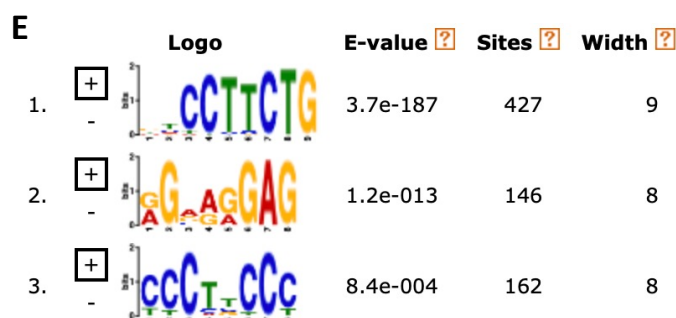
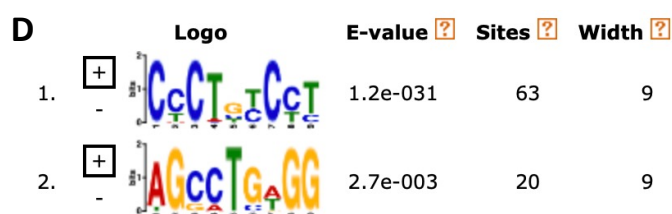
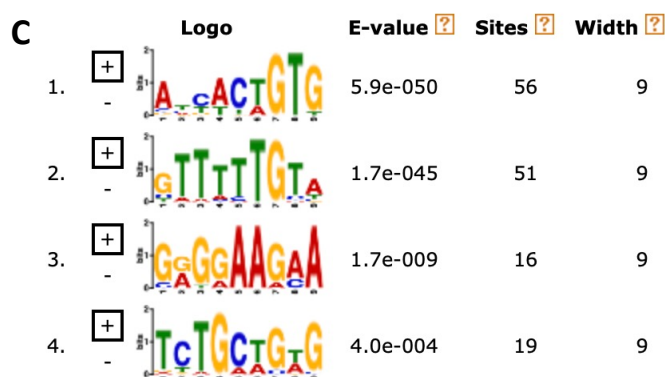
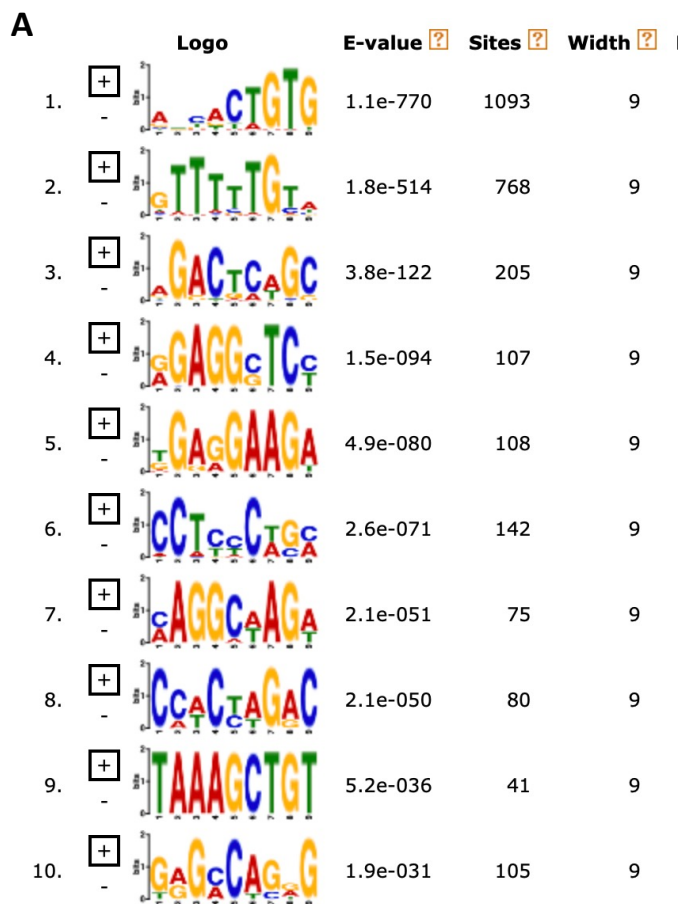
## Enlarged view of some V-gene segments: chr6:41,062,722-41,087,986 (25,265 bp)



## Enlarged view of some J-gene segments: chr6:41,484,026-41,485,926 (1,901 bp)



**Supplementary Figure 6: Genomic representation of mouse TCR loci and distribution of AAV IS** A-C) Genomic distribution of AAV IS observed in LN, Liver and Brain dataset and targeting TCR genes: TCR $\alpha$  (A) TCR $\gamma$  (B) and TCR $\beta$  (C), as indicated. Genomic coordinates and scale are indicated in each panel. Black lines in Brain, Liver and LN datasets indicate the position of AAV IS; Black lines below End-seq data indicate the position of breaks identified in thymocyte by End-seq, Black lines below TCRA/D, TCRB and TCRG indicate the gene segment positions identified using genomic coordinate from IMGT®, the international ImMunoGeneTics information system® (<https://www.imgt.org>). Representative enlarged genomic region containing some V and J gene segments from TCR $\alpha$  and TCR $\beta$  loci. AAV IS within TCRs clustered within the 3' region of V segments and the 5' region of J segments. Furthermore, many AAV IS in TCR $\alpha$  locus are positioned in genomic regions identified in mouse thymocyte by End-seq.



**Supplementary Figure 7: Motif logo identified in the different datasets.** Motif logo for the top 10 most significant recurrent motifs identified in the Zap70 (A) LN and spleen dataset, (B) liver, and (C) GFP LN and SPL dataset; (D) Mecp2 brain and (E) liver

**Supplementary Table 1. Sequencing reads identified by the RAAVIoli bioinformatics pipeline**

Genotype	Group	Vector	Tissue	Total reads	Reads aligning only on AAV	Chimeric Reads aligning on AAV and mouse genome	%AAV_only	%chimeric
Zap70-KO	ZAP70	AAV.PGK.ZAP70	LN	17,311,492	8,466,993	8,844,055	48.91	51.09
			LIV	13,393,962	13,345,530	48,430	99.64	0.36
			SPL	1,802,621	270,836	1,531,763	15.03	84.97
WT	WT-GFP	AAV.PGK.GFP	LN	4,177,580	4,167,532	10,043	99.76	0.24
			SPL	1,452,426	1,327,069	125,338	91.37	8.63
Meep2-KO	PHP.eB	AAV.CB A.GFP	LIV	30,013,851	29,479,228	532,438	98.22	1.78
		AAV.CBA .Meep2	BR	25,573,177	25,468,622	104,359	99.59	0.41
			Total	93,725,109	82,525,810	11,196,426	88.05	11.95

## Non-redundant AAV Integration Site identified in each sample

Genotype	Group	Vector	Dose	MouseID	TP	Tissue	NumIS	Vg
WT	GFP-WT	AAV.PGK.GFP	4.4x10exp10	N10	10	LN	1	nd
			4.4x10exp10	N17	105	LN	8	nd
			4.4x10exp10	N27	10	LN	2	nd
			4.4x10exp10	N28	10	LN	4	nd
			4.4x10exp10	N29	10	LN	14	nd
			5.9x10exp10	SP1	30	SPL	7	nd
			5.9x10exp10	SP4	11	SPL	17	0.004
			5.9x10exp10	SP8	11	SPL	16	0.003
			5.9x10exp10	SP9	11	SPL	22	nd
			5.9x10exp10	SP13	11	SPL	1	nd
ZAP70_KO	ZAP70	AAV.PGK.ZAP70	4.6x10exp11	101	21	LN	195	nd
			4.6x10exp11			SPL	28	nd
			4.6x10exp11	103	301	LIV	27	nd
			4.6x10exp11			SPL	113	nd
			4.6x10exp11	104	301	LIV	33	nd
			4.6x10exp11			LN	225	nd
			4.6x10exp11	106	301	LN	11	nd
			4.6x10exp11			LN	70	nd
			4.6x10exp11	109	21	LN	2	nd
			4.6x10exp11			SPL	88	nd
			4.6x10exp11	110	21	LN	1	nd
			4.6x10exp11			SPL	103	nd
			4.6x10exp11	111	21	LN	468	nd
			4.6x10exp11			SPL	62	nd
			4.6x10exp11	488	21	LIV	51	nd
			4.6x10exp11	498	21	LIV	20	nd
MECP2_KO	PHP.eB	AAV.CBA.GFP	10exp12	L	39	LIV	492	nd
			10exp12	M	45		269	nd
		10exp11	P	75	25		0.01	
		10exp12	S	19	764		23.26	
		10exp12	T	19	938	35.26		
		10exp10	H	56	BR	57	9.03	
		10exp10	I	56		73	13.26	
		10exp11	D	112		26	1.75	
		10exp11	E	224		60	16.73	
		10exp12	J	21		184	45.41	
		10exp12	K	21		305	2.39	

**Supplementary Table 2. Analyses of the AAV vector-host genome junctions  
N IS characterized by MHoR, Insertion or Exact breakpoint at the vector-host genome junctions**

Group (N)	LN_ZAP70	LIV_ZAP70	SPL_ZAP70	LN_GFP	SPL_GFP	LIV_PHP	BR_PHP
MOHR	55	36	19	1	11	1369	163
Exact	218	66	114	8	18	841	436
Inser	699	29	261	20	34	278	106
MHoR/Exact	754	65	280	21	45	1647	269
Tot	972	131	394	29	63	2488	705

**% of IS characterized by MHoR, Insertion or Exact breakpoint at the vector-host genome junctions**

Group (N)	LN_ZAP70	LIV_ZAP70	SPL_ZAP70	LN_GFP	SPL_GFP	LIV_PHP	BR_PHP	Av.	SEM
MOHR	5.66	27.48	4.82	3.45	17.46	55.02	23.12	19.57	6.91
Exact	22.43	50.38	28.93	27.59	28.57	33.80	61.84	36.22	5.43
Inser	71.91	22.14	66.24	68.97	53.97	11.17	15.04	44.21	10.22
MHoR/Exact	77.57	49.62	71.07	72.41	71.43	66.20	38.16	63.78	5.43

**Stats summary: p-value defined by Fisher exact test. P-value lowered that 0.0001 are represented.**

G1	G2	P-value	Exact-pvalue
LN_ZAP	LIV_ZAP70	<0.0001	4.7997E-28
LN_ZAP	LIV_PHP	<0.0001	2.126E-267
LN_ZAP	BR_PHP	<0.0001	8.35E-126
SPL_ZAP70	LIV_ZAP70	<0.0001	6.9185E-19
SPL_ZAP70	LIV_PHP	<0.0001	2.401E-117
SPL_ZAP70	BR_PHP	<0.0001	1.2087E-66
LN_GFP	LIV_ZAP70	<0.0001	2.5731E-06
LN_GFP	BR_PHP	<0.0001	3.2942E-10
LN_GFP	LIV_PHP	<0.0001	6.0082E-13
SPL_GFP	LIV_ZAP70	<0.0001	1.6348E-05
SPL_GFP	LIV_PHP	<0.0001	5.0548E-16
SPL_GFP	BR_PHP	<0.0001	1.7053E-11

**Supplementary Table 3. AAV IS targeting gene transcriptional unit****N IS characterized by MHoR, Insertion or Exact breakpoint at the vector-host genome junctions**

Bin Center	LN_ZAP70	LIV_ZAP70	SPL_ZAP70	LN_GFP	SPL_GFP	LIV_PHP	BR_PHP
-30	4	6	1	0	1	90	23
-10	16	0	1	0	2	252	71
10	85	13	29	3	6	222	61
30	36	10	11	0	3	202	60
50	30	6	20	1	1	221	70
70	80	11	21	0	3	216	72
90	670	28	302	23	39	240	60
+10	21	17	6	1	1	249	76
+30	3	5	2	1	1	88	26

**% of IS characterized by MHoR, Insertion or Exact breakpoint at the vector-host genome junctions**

Bin Center	LN_ZAP70	LIV_ZAP70	SPL_ZAP70	LN_GFP	SPL_GFP	LIV_PHP	BR_PHP
-30	0.41	4.58	0.25	0.00	1.59	3.62	3.26
-10	1.65	0.00	0.25	0.00	3.17	10.13	10.07
10	8.74	9.92	7.36	10.34	9.52	8.92	8.65
30	3.70	7.63	2.79	0.00	4.76	8.12	8.51
50	3.09	4.58	5.08	3.45	1.59	8.88	9.93
70	8.23	8.40	5.33	0.00	4.76	8.68	10.21
90	68.93	21.37	76.65	79.31	61.90	9.65	8.51
+10	2.16	12.98	1.52	3.45	1.59	10.01	10.78
+30	0.31	3.82	0.51	3.45	1.59	3.54	3.69

**Stats summary: data for Fisher exact test**

	LN_ZAP70	LIV_ZAP70	SPL_ZAP70	LN_GFP	SPL_GFP	LIV_PHP	BR_PHP
In gene	901	68	383	27	52	1101	323
Out	71	63	11	2	11	1387	382

**Stats summary: p-value defined by Fisher exact test. P-value lowered that 0.0001 are represented**

G1	G2	P-value	Exact-pvalue
LN_ZAP70	LIV_ZAP70	<0.0001	3.8397E-29
LN_ZAP70	LIV_PHP	<0.0001	1.516E-172
LN_ZAP70	BR_PHP	<0.0001	3.602E-105
SPL_ZAP70	LIV_ZAP70	<0.0001	5.1537E-33
SPL_ZAP70	LIV_PHP	<0.0001	8.17E-105
SPL_ZAP70	BR_PHP	<0.0001	1.5935E-79
LN_GFP	LIV_ZAP70	<0.0001	1.717E-05
LN_GFP	LIV_PHP	<0.0001	7.2081E-08
LN_GFP	BR_PHP	<0.0001	2.7351E-07
SPL_GFP	LIV_ZAP70	<0.0001	3.4144E-05
SPL_GFP	LIV_PHP	<0.0001	1.3108E-09
SPL_GFP	BR_PHP	<0.0001	1.1109E-08

**Supplementary Table 4. Genes targeted by AAV IS in the different dataset**

Dataset	GeneID	N IS	Targeting (%)
LN_ZAP70 N=972	Tcr $\alpha$	567	58.33
	Tcr $\beta$	192	19.75
	Tcry	87	8.95
	IgH	15	1.54
	Satb1	12	1.23
SPL_ZAP70 N=394	Tcr $\alpha$	257	65.23
	Tcr $\beta$	73	18.53
	Tcry	38	9.64
	IgH	5	1.27
	Mir684-1	2	0.51
LIV_ZAP70 N=131	Tcr $\alpha$	8	6.15
	Alb	4	3.08
	Itgam	2	1.54
	McpH1	2	1.54
LN_GFP N=29	Tcr $\alpha$	17	58.62
	Tcr $\beta$	8	27.59
	Tcry	1	3.45
SLN_GFP N=63	Tcr $\alpha$	31	49.21
	Tcr $\beta$	10	15.87
	Tcry	2	3.17
	IgK	2	3.17
PHP_LIV N=2488	Alb	9	0.36
	Diras2	8	0.32
	Mtarch2	6	0.24
	Msi2	6	0.24
	Ofcc1	5	0.20
	Tmem114	5	0.20
	Pakap	5	0.20
PHP_BR N=705	Ncam2	3	0.43
	Prrc2b	3	0.43
	Fars2	3	0.43
	Atp2c2	3	0.43
	C330004P14Rik	2	0.28
	Abca13	2	0.28
	Specc1	2	0.28



**Supplementary Table 5. AAV IS close to RSS site**

**N IS close to RSS site ( $\pm 50$  bp)**

	LN_ZAP70	SPL_ZAP70	LIV_ZAP70	LN_GFP	SPL_GFP	BR_PHP	LIV_PHP
In	877	370	37	26	50	108	414
Out	95	24	94	3	13	597	2074
% In	90.2	93.9	28.2	89.7	79.4	15.3	16.6

**Stats summary: p-value defined by Fisher exact test. P-value lowered that 0.0001 are represented**

G1	G2	P-value	Exact-pvalue
LN_ZAP	LIV_ZAP	<0.0001	4.1186E-52
LN_ZAP	BR_PHP	<0.0001	6.729E-229
LN_ZAP	LIV_PHP	<0.0001	0
SPL_ZAP	BR_PHP	<0.0001	5.696E-158
SPL_ZAP	LIV_PHP	<0.0001	1.36E-208
SPL_ZAP	LIV_ZAP	<0.0001	6.8564E-50
LN_GFP	BR_PHP	<0.0001	1.6573E-17
LN_GFP	LIV_PHP	<0.0001	2.3519E-17
LN_GFP	LIV_ZAP	<0.0001	8.7384E-10
SPL_GFP	BR_PHP	<0.0001	4.9813E-26
SPL_GFP	LIV_PHP	<0.0001	1.0682E-26
LIV_ZAP	SPL_GFP	<0.0001	2.1665E-11