Advanced characterization of DNA molecules in rAAV vector preparations by single-stranded virus next-generation sequencing

Emilie Lecomte*, Benoît Tournaire*, Benjamin Cogné, Jean-Baptiste Dupont, Pierre Lindenbaum, Mélanie Martin-Fontaine, Frédéric Broucque, Cécile Robin, Matthias Hebben, Otto-Wilhelm Merten, Véronique Blouin, Achille François, Richard Redon, Philippe Moullier# and Adrien Léger#

This file includes:

SUPPLEMENTARY RESULTS

I) Development and validation of SSV-Seq

To differentiate DNA outside and inside the viral capsid, we compared the ability of several DNases, to digest DNA fragments mixed with rAAV particles. Quantification of residual DNA after DNases treatment showed that the combination of the Baseline-ZERO endonuclease and the Plasmid-Safe exonuclease led to reduction of the amount of DNA exceeding $1x10^5$ fold (**Fig.S1**). In parallel, the amount of rAAV genome was not affected by the treatment. In addition, for each sample analysed by SSV-Seq, a spike-in control consisting of exogenous DNA (Lambda phage DNA) was used to validate the efficiency of the DNAses treatment by analysing mapped reads after NGS (Table S5). Since the exact nature of the rAAV outer capsid and the DNA contaminant is not known, we do not exclude that our spike-in controls would not reproduce DNA strongly tighten to the capsid surface. Even though the DNA remaining after DNases treatment is likely to be almost entirely encapsidated, as a precautionary measure we preferred to use the term "DNase protected. Regarding DNA extraction from rAAV productions, we decided to avoid the use of silica column, known to induce size selection bias and possible DNA contaminations (Evans GE et al, J Clin Microbiol. 2003 Jul;41(7):3452-3.). Instead, we took advantage of a simple procedure based on a modified salting-out precipitation method (Gentra-Puregene blood kit, Qiagen, Venlo, Limburg, Netherlands) and obtained an average yield of rAAV DNA extraction between 50% and 90%.

Since current NGS technologies require a double stranded DNA (dsDNA) input for library preparation, the single stranded rAAV genome had to be converted into dsDNA in vitro (**Fig.1c**). We developed a robust method relying on random priming and DNA Polymerase I with inherent 3'→5' and 5'→3' exonucleases activities. The efficiency of second strand synthesis was controlled by incorporating fluorescein-labelled dUTP (**Fig.S2**). Although a smear of labelled DNA can be observed on Southern immunoblotting, a strong band is visible at the expected rAAV genome size (**Fig.S2a**). The smear indicates a partial fragmentation during the second strand synthesis. Then, the fluorescence intensity was quantified using a fluorescence plate reader (VictorX3, Perkin Elmer, Waltham, MA, USA) and normalized by the mass of DNA in each sample (**Fig.S2b**). All the samples were detected above the negative control, with intensities comparable to the positive control.

Since this protocol might skew relative sequence representation, ratios of rAAV genome copies per plasmid backbone copies were determined using qPCR before and after the second strand synthesis (**Fig.S3, Table S9**). They did not show significant sequence enrichment during this critical step of the protocol.

Lastly, we performed an adapted procedure for NGS library preparation from Kozarewa et al (Kozarewa and Turner, J Clin Microbiol. 2003 Jul;41(7):3452-3). The samples were sheared by sonication in fragments with a median size of 300 bp, end repaired, A-tailed and adapters were ligated. One of the 2 adapters contains a short DNA barcode, also called index, which is different for each experimental sample. These fragments were amplified by PCR using the PfuUltra II Fusion polymerase (**Fig.1d**). To remove the smallest fragments and the dimers of adapters, each step was followed by a gel free size-selection (SPRIselect, Beckman Coulter, Indianapolis IN, USA). Libraries were controlled after sonication and PCR steps by electrophoresis on Agilent 2100 Bioanalyzer (**Fig.S4**). We obtained PCR products centered on 400 bp (range 250 to 1000 bp) confirming that the upstream steps were correctly performed. Finally, samples were quantified and pooled in equimolar quantities. We also add 1 to 5% of Phi-X DNA in the mix to increase sequence diversity for Illumina Sequencing to compensate the high redundancy of the dataset. High throughput sequencing was achieved on an Illumina HiSeq platform (Rapid Run, 2 × 101pb).

II) Production, purification and characterization of rAAV vectors

We produced an rAAV 2/8 CMVp-eGFP-hygroTK-bGHpA raw batch by transient transfection of adherent HEK-293 cells and purified it either by cesium chloride density gradient (CsCl), affinity chromatography (AVB) or ion exchange chromatography (IEX) (**Fig.S5**). Recombinant AAV preparations were characterized with standard quality controls methods, including extensive qPCR analyses (**Table S9)**.

To determine the protein purity of rAAV productions, SDS-PAGE followed by Coomassie blue and silver staining were performed. For all batches, protein purity was higher than 90% as determined using GeneTools software from Coomassie blue staining gel (**Fig.S6a**). However, particles purified by AVB chromatography contained more contaminants (lipids, nucleic acids, glycans or proteins) as shown by silver staining (**Fig.S6b**). In addition, the Coomassie blue staining suggests that rAAV purified by AVB contains more empty particles, as confirmed by an AAV8 titration ELISA (**Fig.S6c**). This is likely to be due to the inability of AVB columns to differentiate between full, malformed and empty particles, while the IEX and CsCl purifications used in this work can enrich in full rAAV particles.

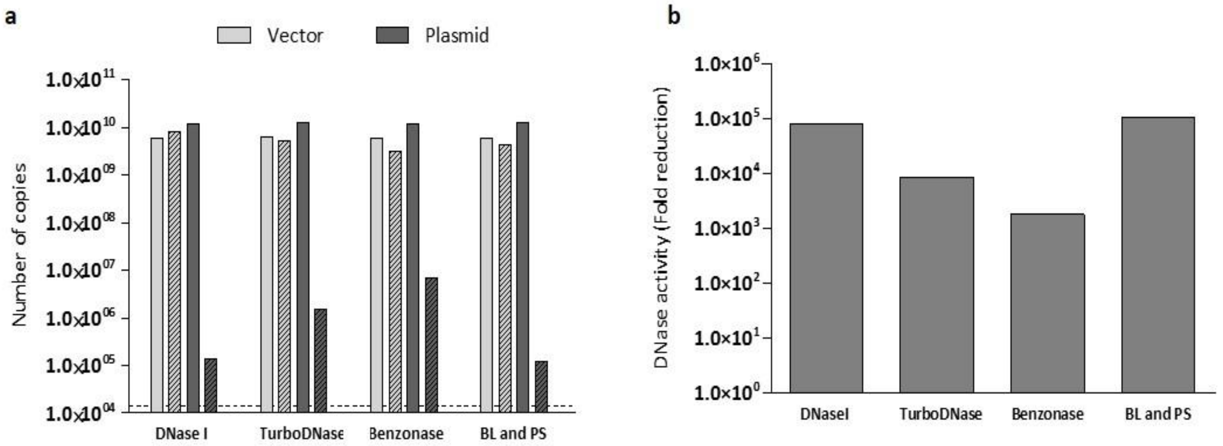
Quantification of the DNA composition by qPCR (**Table S1**) showed that rAAV particles purified by CsCl were less contaminated by the vector plasmid and, to a lesser extent, by the helper plasmid, compared with both chromatographic methods. Altogether, these results emphasize the impact of the purification process on the quality of rAAV batches.

III) Development and validation of ContaVect

A dedicated bioinformatics pipeline (ContaVect) was developed to attribute each sequencing reads to its most likely original sequence (**Fig.1e**). Users can provide several reference sequences for the rAAV genome and the possible DNA contaminants (vector plasmid, helper plasmid, human genome…). If several reference sequences have homologies, it can results in the misattribution of sequencing reads, and/or bad quality mapping, as shown thanks to an artificial dataset mimicking an rAAV production (CMVp-eGFP-hygroTK-bGHpA) in **Table S3**. Indeed, without any preprocessing of the reference sequences we obtained low sensitivity (true positive rate) and specificity (true negative rate) scores for some of the references (< 50%). We developed a fast reference preprocessing module integrated into ContaVect to identify homologies with Blast algorithm and mask them in the less frequent reference, based on a ranking of estimated relative abundance provided by the user. With the preprocessing module, we obtained a much better sensitivity (90.9%, to 100%) and specificity (99.7% to 100%) (**Table S4**). The software is still under active development but the version used for the analyses performed in this study (v0.2) is freely available with an extensive user and developer documentation at https://github.com/a-slide/ContaVect/tree/v0.2.

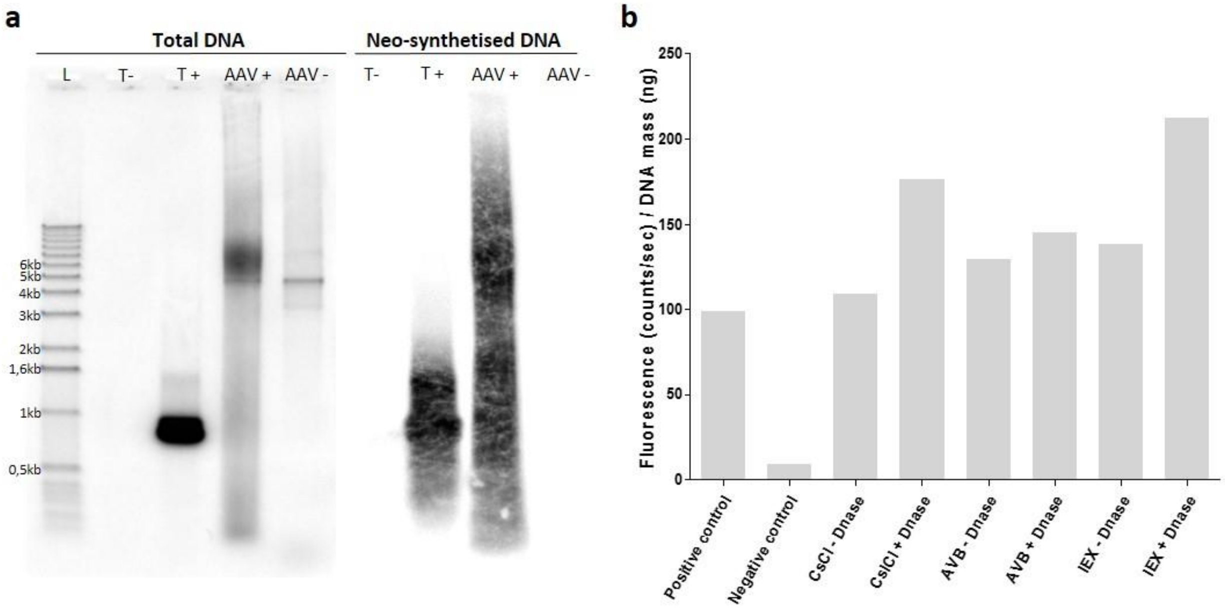
SUPPLEMENTARY FIGURES

Figure S1. Comparative efficiency of DNase on DNA spiked in rAAV production



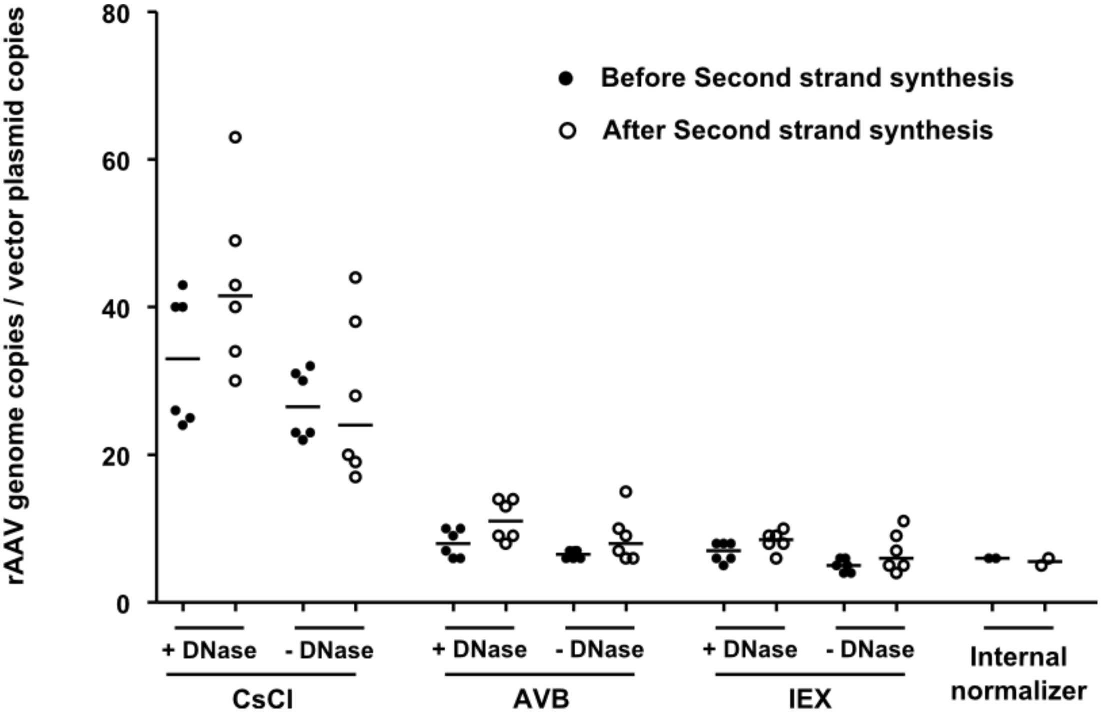
DNase I, TurboDNase, Benzonase or Baseline-ZERO/Plasmid-safe mix (BL and PS) were used to digest 1x10^{10} copies of linearized plasmid mixed with 1x10^{10} rAAV vector particles. After purification, DNase activity was evaluated by qPCR targeting the plasmid and the vector sequences. **(a)** For each DNase condition, the number of copies was represented before and after treatment. Plain bars represent the undigested samples and hatched bars the conditions where the DNase treatment was performed. The dotted line shows the LOQ (1.4x10^4 copies). **(b)** DNase activity represented as fold reduction of plasmid after DNase treatment.

Figure S2. Control of efficient second strand synthesis



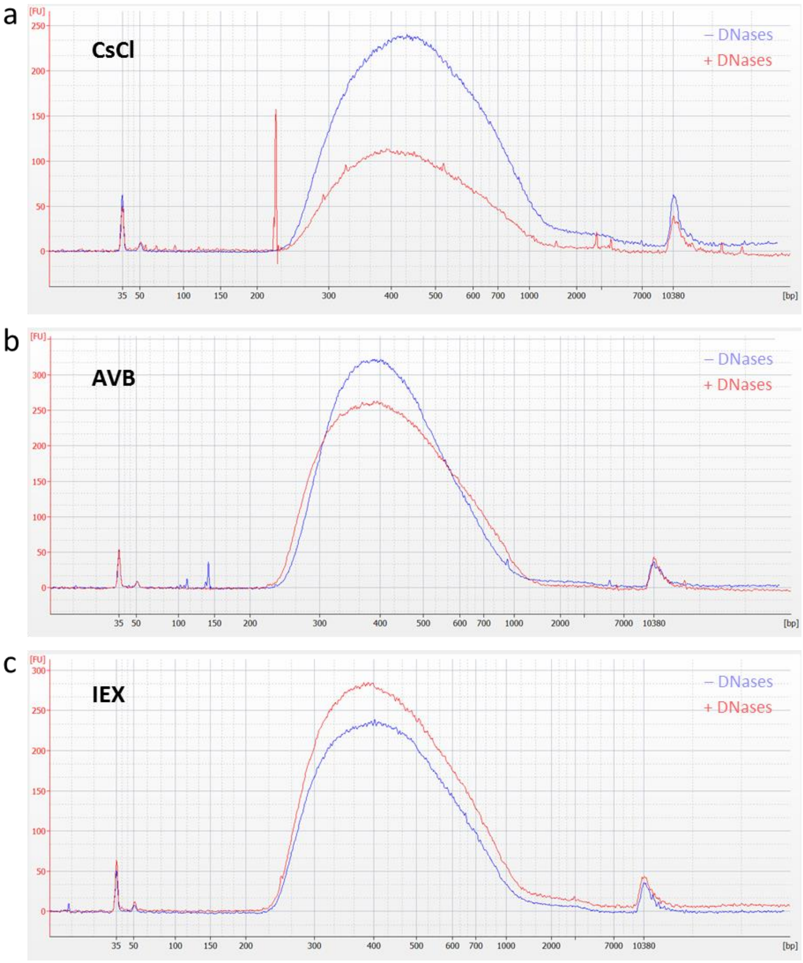
The efficiency of the second strand synthesis was controlled by adding a fluorescein-12 labelled dUTP, then evaluating its incorporation in the neo-synthesized DNA qualitatively (Southern Immunoblot) and quantitatively (fluorimetry). **(a)** Anti-fluorescein immunoblot performed from the rAAV2/8-CMVp-eGFP-hygroTK-bGHpA preparation purified by AVB chromatography, following SSV-Seq protocol until the second strand synthesis with fluorescein-12-dUTP. Total DNA was detected using GelRed staining. The negative control (T-) contained only water and the positive control (T+) is a 780 bp DNA fragment labelled by PCR with fluorescein-12-dUTP. The experimental samples AAV+ and AAV- were processed according to the same protocol except that no DNA Pol I was added during the second strand synthesis of the AAV-. L: 1kb Ladder (Life technologies). **(b)** Fluorescence intensities of rAAV preparations quantified using Perkin Elmer Victor X3 in the 6 experimental samples described in the manuscript. Results were normalised by converting the number of copies after second strand synthesis obtained by qPCR into total DNA mass. The positive control is the same than in **(a)** and the negative control is a plasmid fragment containing the rAAV genome that underwent the protocol but without DNA Pol I.

Figure S3. Selection bias induced by second strand synthesis



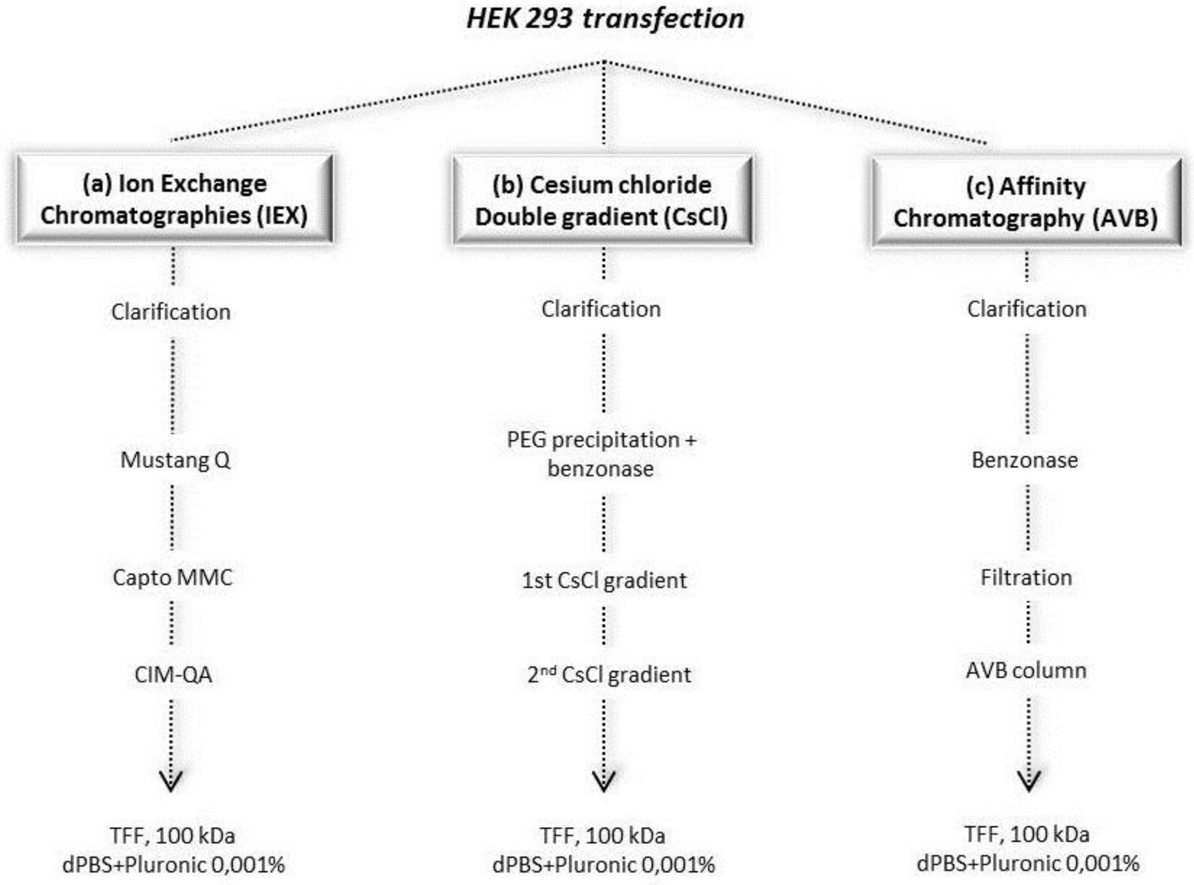
To evaluate the possible selection/amplification bias due to the second strand synthesis, we calculated a ratio between the rAAV genome copy number and vector plasmid copy number, in process before and after this critical step. The rAAV genome and vector plasmid copy numbers were determined using qPCR targeting the bGH polyadenylation signal (**Table S9,** BGH pA) and the kanamycin resistance gene (**Table S9,** Kana R), respectively. No significant difference was found for all experimental samples when comparing the ratio before and after second strand synthesis. Horizontal lines represent the medians. n=6 for each rAAV2/8-CMVp-eGFP-hygroTK-bGHpA samples. n=2 for the internal normalizer and the control. Statistics: two tailed Mann-Withney’s U-test, Confidence interval = 95%.

Figure S4. Distribution of DNA fragment sizes after NGS library preparation



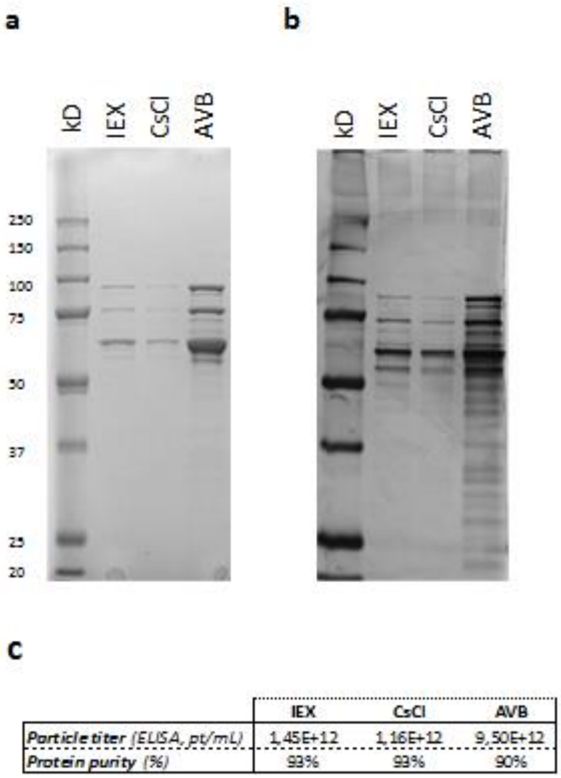
The distribution of DNA fragment sizes was determine at the end of the NGS library preparation protocol by the Agilent 2100 Bioanalyzer system using High sensitivity DNA chip. As showed for each of the 6 rAAV sample libraries prepared for this study, fragments below 250 bp were eliminated by the successive washing steps.

Figure S5. Overview of AAV 2/8-CMV-GFP-hTK-BGHpA vectors purification



After HEK-293 cells transient transfection, production was splitted in three parts. Batch **(a)** was clarified by filtration (Merck-Millipore), purified by three different ion exchange chromatography (IEX) steps: anionic membrane (Pall Corporation), cationic multimodal (GE Healthcare Life sciences) and anionic monolith (Bia Separations) columns. Batch **(b)** was clarified by centrifugation, precipitated with a polyethylene glycol (PEG), treated with benzonase, and purified by double Cesium Chloride gradient (CsCl). Batch **(c)** was clarified by centrifugation, treated with benzonase, and purified by immune-affinity chromatography (GE Healthcare Life sciences). All of the vectors were concentrated and formulated in DPBS containing 0.001% Pluronic F-68 by tangential flow filtration (TFF) using a fibre cartridge of 100 kDa (GE Healthcare Life sciences).

Figure S6. Characterization of rAAV productions purity and titer



	IEX	CsCl	AVB
Particle titer (ELISA, pt/mL)	1,45E+12	1,16E+12	9,50E+12
Protein purity (%)	93%	93%	90%

The protein purity of rAAV preparation purified by IEX, CsCl and AVB methods was evaluated by running SDS-PAGE gels and staining with (a) Coomassie blue (1×10^{11} vector genomes/sample) or (b) silver staining (2×10^{10} vector genomes/sample). kD: BenchMark protein ladder in kiloDalton (Invitrogen) (c) The overall protein purity was determined using Coomassie blue staining and the vector particle titer was evaluated using an anti-AAV8 capsid ELISA assay.

Figure S7. Overview of the protocol followed in this study

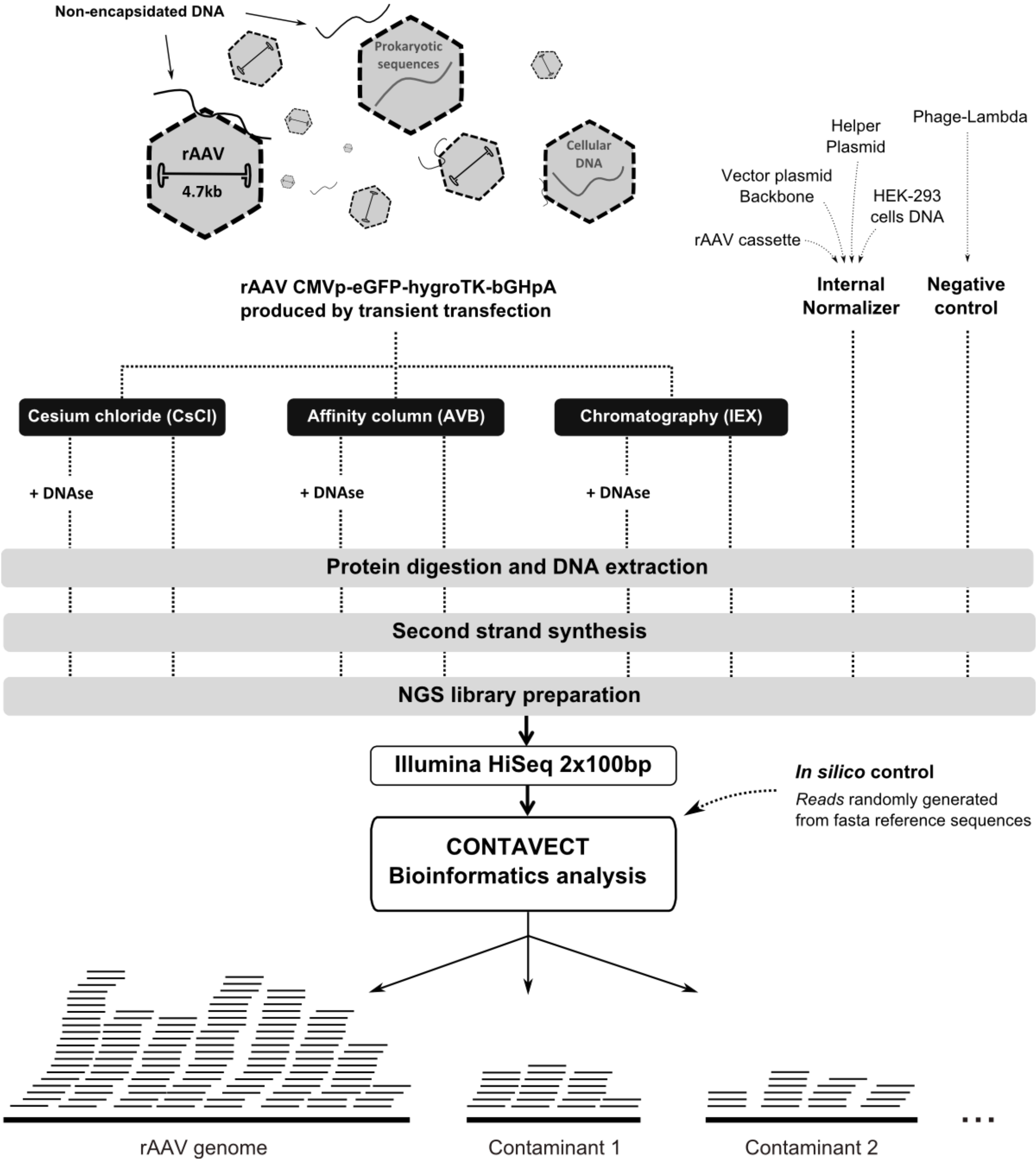
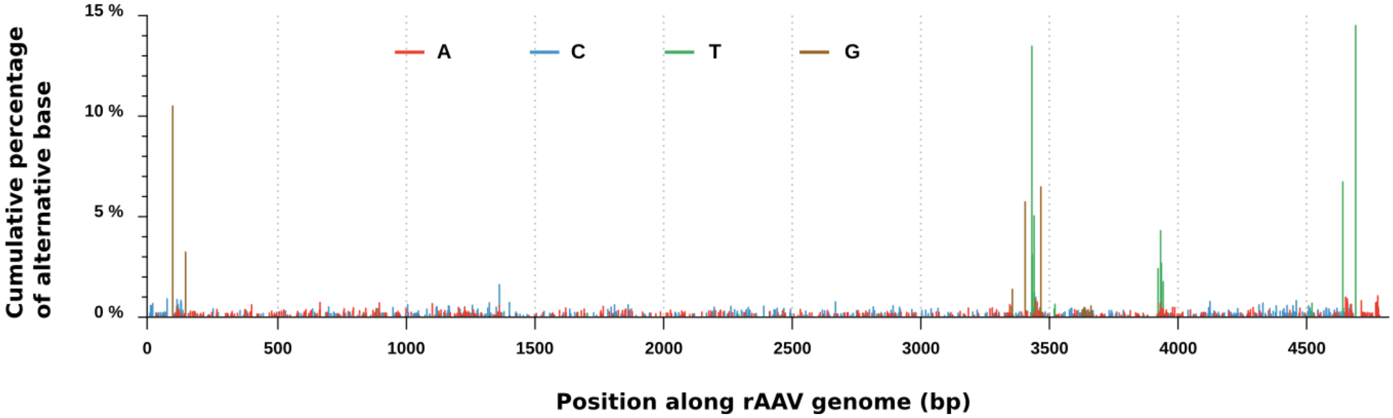


Figure S8. Percentage of single nucleotide variants along rAAV genome for the plasmid control from the internal normalizer



Cumulative percentage of alternative base A (red), C (blue), T (green) and G (brown) compared with the reference sequence for each nucleotide position along rAAV genome in base pair. When several variants were found at the same nucleotide position, variant contributions were stacked. SNVs are represented on the graph if they were found in at least 1 of the 2 technical replicates of the internal normalizer control.

SUPPLEMENTARY TABLES

Table S1. qPCR titration of rAAV and DNA contaminants

Reference	Reference length (pb)	PCR target	Target length (pb)	CsCl - DNAse (cp/ml)	CsCl + DNAse (cp/ml)	AVB - DNAse (cp/ml)	AVB + DNAse (cp/ml)	IEX - DNAse (cp/ml)	IEX + DNAse (cp/ml)
rAAV genome	4794	ITR2	62	2.61E+12	2.67E+12	4.49E+12	5.20E+12	2.07E+12	2.19E+12
		CMVp	124	7.05E+11	5.41E+11	9.49E+11	8.74E+11	4.77E+11	3.88E+11
		BGHpA	69	5.00E+11	4.47E+11	6.79E+11	6.71E+11	2.96E+11	2.96E+11
		GFP 1	74	5.81E+11	5.17E+11	7.62E+11	6.83E+11	3.42E+11	3.45E+11
		GFP 2	66	6.62E+11	6.04E+11	9.21E+11	8.41E+11	3.78E+11	3.79E+11
Vector plasmid backbone	2209	KanaR	147	1.14E+10	9.13E+09	6.33E+10	6.09E+10	3.71E+10	3.17E+10
Helper plasmid	22757	Cap8	81	7.31E+06	5.97E+06	3.83E+07	3.74E+07	1.35E+07	1.35E+07
		Rep2	75	8.72E+06	8.73E+06	4.62E+07	3.05E+07	1.72E+07	1.69E+07
		E4	67	7.69E+06	6.67E+06	3.56E+07	3.51E+07	1.58E+07	1.29E+07
Human genome	3099750718	human ALB1	100	<LOQ	<LOQ	<LOQ	<LOQ	<LOQ	<LOQ

rAAV productions were titered by qPCR targeting different sequences present in the rAAV genome. DNA contaminants from the vector plasmid backbone, the helper plasmid and the human genome were quantified using sequence specific targets. Similar to the SSV-Seq protocol, encapsidated and non-encapsidated targets were differentiated by our DNases treatment. The quantity of each target was expressed in copy number per mL of final rAAV product. For human genome, the copy number was below our limit of quantification in all of the conditions ($3x10^3$ copies/mL).

Table S2. Description of the samples analyzed by SSV-Seq.

Sample Name	Composition (estimated number of copies)	DNase treatment	Mean read quality	Number of reads
AAV CsCl+	AAV preparation CsCl purification (2×10^{11} cp of rAAV = 484 ng) + phage λ DNA (24.2ng)	+	34.74	9 658 441
			35.18	7 674 203
AAV CsCl-		-	34.9	7 904 503
			35.36	8 047 430
AAV AVB+	AAV preparation AVB purification (2×10^{11} cp of rAAV = 484 ng) + phage λ DNA (24.2ng)	+	34.68	6 658 586
			35.51	8 046 244
AAV AVB-		-	34.99	7 906 618
			35.74	6 340 719
AAV IEX+	AAV preparation IEX purification (2×10^{11} cp of rAAV = 484 ng) + phage λ DNA (24.2ng)	+	34.79	8 810 295
			35.45	6 699 569
AAV IEX-		-	34.91	9 028 587
			35.47	7 826 906
Negative control	phage λ DNA (484 ng)	+	35.72	8 523 711
			36.52	6 717 691
Internal normalizer	rAAV genome (2×10^{11} cp) + vector plasmid backbone (1×10^{10} cp) + Helper plasmid (4×10^{9} cp) + sonicated HEK-293 cells DNA (1×10^{2} cp)	-	35.04	7 574 472
			35.65	6 786 931

Each sample was analyzed in 2 technical replicates. The estimated composition of the samples corresponds to the number of copies (cp) determined by qPCR titrations (when available) or by microspectrophotometry. The number of reads and mean PHRED quality indicates values for raw data after sample de-multiplexing but before any quality filtering.

Table S3. Confusion matrix and mapping prediction rate of ContaVect determined without pre-processing of references

		EXPECTED						
		AAV	**Backbone**	**Helper**	**Ad5**	**Human**	**Unmapped**	Total
OBSERVED	**AAV**	14000000	0	0	0	0	0	14000000
	Backbone	0	206120	2	0	0	0	206122
	Helper	0	204	4752	0	0	0	4956
	Ad5	0	0	0	10	0	0	10
	Human	0	0	0	0	28952	0	28952
	Unmapped	0	213676	246	0	572	200000	414494
	Total	14000000	420000	5000	10	29524	200000	

Percentages	**AAV**	**Backbone**	**Helper**	**Ad5**	**Human**	**Unmapped**
True positive (sensitivity)	100.00%	49.08%	95.04%	100.00%	98.06%	100.00%
False negative	0.00%	50.92%	4.96%	0.00%	1.94%	0.00%
True negative (specificity)	100.00%	100.00%	95.88%	100.00%	100.00%	48.25%
False positive	0.00%	0.00%	4.12%	0.00%	0.00%	51.75%

Confusion Matrix and mapping prediction rates were obtained with an artificial dataset mimicking an rAAV production CMVp-eGFP-hygroTK-bGHpA generated using Fastq Control Sampler.

Upper table. Reads from rAAV, Backbone, Helper, Ad5, Human references were generated from the reference fasta sequences (see material and method section) and the read from the "unmapped" reference from a randomly generated sequence. The numbers of reads generated per reference are indicated in the lower row in the corresponding columns. The last columns summarize the number of read attributed to each reference after mapping with ContaVect, without the reference pre-processing module. A green cell indicates a correct assignation, while a red cell corresponds to mapping errors.

Lower table. The percentages of true positives, true negatives, false positives and false negatives were calculated from results obtained in the upper table for each reference.

Table S4. Confusion matrix and mapping prediction rate of ContaVect determined with a pre-processing of references

		EXPECTED						
		AAV	Backbone	Helper	Ad5	Human	Unmapped	Total
OBSERVED	AAV	14000000	0	0	0	0	0	14000000
	Backbone	0	420000	455	0	0	0	420455
	Helper	0	0	4543	0	0	0	4543
	Ad5	0	0	0	10	0	0	10
	Human	0	0	0	0	28952	0	28952
	Unmapped	0	0	2	0	572	200000	200574
	Total	14000000	420000	5000	10	29524	200000	

Percentages	AAV	Backbone	Helper	Ad5	Human	Unmapped
True positive (sensitivity)	100.00%	100.00%	90.86%	100.00%	98.06%	100.00%
False negative	0.00%	0.00%	9.14%	0.00%	1.94%	0.00%
True negative (specificity)	100.00%	99.89%	100.00%	100.00%	100.00%	99.71%
False positive	0.00%	0.11%	0.00%	0.00%	0.00%	0.29%

Confusion Matrix and mapping prediction rates were obtained with an artificial dataset mimicking an rAAV production CMVp-eGFP-hygroTK-bGHpA generated using Fastq Control Sampler.

Upper table. Reads from rAAV, Backbone, Helper, Ad5, Human references were generated from the reference fasta sequences (see material and method section) and the read from the "unmapped" reference from a randomly generated sequence. The numbers of reads generated per reference are indicated in the lower row in the corresponding columns. The last columns summarize the number of read attributed to each reference after mapping with ContaVect, using the reference pre-processing module. A green cell indicates a correct assignation, while a red cell corresponds to mapping errors.

Lower table. The percentages of true positives, true negatives, false positives and false negatives were calculated from results obtained in the upper table for each reference.

Table S5. Distribution of contaminants in absolute number of reads

Reference name	Length (bp)	Negative control	Internal Normalizer	CsCl - DNAse	CsCl + DNAse	AVB - DNAse	AVB + DNAse	IEX - DNAse	IEX + DNAse
Phi X174	5 386	155 022 18 440	131 849 21 624	157 572 25 431	214 214 26 479	137 587 22 447	127 001 31 976	174 238 30 181	201 407 20 004
λ phage	48 502	15 820 832 12 524 254	1 993 838	1 317 316 1 118 877	509 357	1 567 573 1 074 966	270 515	1 863 260 1 284 740	631 401
rAAV genome	4 794	17 217 1 205	8 014 026 7 996 633	12 513 125 12 848 053	16 768 399 13 281 337	11 820 182 9 729 696	11 207 153 13 446 419	13 302 866 11 698 973	14 536 121 11 248 707
Vector plasmid backbone	2 209	913 366	1 138 694 1 031 197	148 331 167 598	142 147 116 191	533 912 460 698	348 283 485 013	737 458 745 067	815 318 491 676
Helper plasmid	22 757	54 22	3 937 978 628 280	1 472 1 622	1 763 1 348	8 119 6 671	5 267 8 028	7 167 6 601	7 807 9 321
Ad5 in 293	4 344	0 2	2 4	0 0	0 0	0 2	0 2	0 0	0 0
human genome	3 099 750 718	2 292 3140	196 882 263 076	13 210 17 749	6 361 5 804	36 122 30 671	24 708 39 382	28 946 28 913	26 333 19 637
Unmaped		68 642 203 077	374 962 2 414 472	99 746 260 792	101 967 210 724	213 741 287 211	143 356 537 789	182 021 319 873	171 241 292 122
Total		16 064 972 12 750 506	13 796 386 12 356 124	14 250 772 14 440 122	17 235 360 13 642 240	14 317 236 11 612 362	11 856 038 14 549 124	16 295 956 14 114 348	15 758 858 12 081 868

The number of reads assigned by ContaVect to each reference, including technical decoy references and unmapped reads, is indicated for the 2 technical replicates of each sample. The size of references in base pairs (bp) is indicated in the second column of the table.

Table S6. Distribution of reads in a specific locus of chr15 and in the D-loop of mtDNA

	Read count		% per human genome read		% per total mapped read	
Sample name	**chr15 gene**	**d-loop**	**chr15 gene**	**d-loop**	**chr15 gene**	**d-loop**
Cont Neg	0	0	0.00%	0.000%	0.000%	0.0000%
	0	0	0.00%	0.000%	0.000%	0.0000%
Internal Normalizer	0	41	0.00%	0.021%	0.000%	0.0003%
	89	29	0.03%	0.011%	0.001%	0.0003%
CsCl - DNAse	1	237	0.01%	1.794%	0.000%	0.0019%
	0	154	0.00%	0.868%	0.000%	0.0012%
CsCl + DNAse	2	14	0.03%	0.220%	0.000%	0.0001%
	0	4	0.00%	0.069%	0.000%	0.0000%
AVB - DNAse	6763	1	18.72%	0.003%	0.055%	0.0000%
	5332	2	17.38%	0.007%	0.052%	0.0000%
AVB + DNAse	4466	3	18.08%	0.012%	0.039%	0.0000%
	7094	2	18.01%	0.005%	0.051%	0.0000%
IEX - DNAse	0	4	0.00%	0.014%	0.000%	0.0000%
	0	1	0.00%	0.003%	0.000%	0.0000%
IEX + DNAse	0	0	0.00%	0.000%	0.000%	0.0000%
	0	1	0.00%	0.005%	0.000%	0.0000%

The number of reads found in a specific locus of chr15 (not disclosed due to confidentiality concerns) and in the D-loop of mtDNA (human genome GRCh38 MT: 16,078 - 16,561), is indicated in the 2nd and 3rd columns of the table for the 2 technical replicates of each sample. The 4th and 5th columns contain the percentages of the reads compared with all of the reads found in the human genome, and the 2 last columns the percentage compared with all of the reads mapped, regardless of the reference. Red cells represent samples for which there is a higher contamination.

Table S7. Comparative distribution of reads in AAV ITR extremities with separated of merged AAV and vector backbone references

	Separated AAV and Backbone references			Merged AAV and Backbone references		
	reads in ITR	reads in AAV	% reads in ITR	reads in ITR	reads in AAV	% reads in ITR
Cont Neg	612	17217	3.55%	903	17901	5.04%
	46	1205	3.82%	63	1240	5.08%
Internal Normalizer	373242	8014026	4.66%	516768	8229781	6.28%
	377024	7996633	4.71%	531094	8228037	6.45%
CsCl - DNAse	393995	12513125	3.15%	542707	12751338	4.26%
	497911	12848053	3.88%	666227	13096780	5.09%
CsCl + DNAse	495941	16768399	2.96%	655128	17183806	3.81%
	519122	13281337	3.91%	664758	13624315	4.88%
AVB - DNAse	456315	11820182	3.86%	691500	12198959	5.67%
	507244	9729696	5.21%	719640	10056082	7.16%
AVB + DNAse	464122	11207153	4.14%	660816	11606326	5.69%
	943264	13446419	7.01%	1351071	14122346	9.57%
IEX - DNAse	437581	13302866	3.29%	709229	13744228	5.16%
	598961	11698973	5.12%	900234	12151130	7.41%
IEX + DNAse	559802	14536121	3.85%	821117	15116144	5.43%
	559588	11248707	4.97%	790512	11712476	6.75%

We assess the efficiency of mapping of read overlapping AAV ITR extremities when aligning with ContaVect, either with separated AAV and plasmid backbone references (one fasta for each), or with a unique merged reference corresponding to the complete vector plasmid. To do so, reads overlapping both left and right ITR were counted and compared with reads overlapping the rAAV genome. This was done directly from BAM files using a script based on pysam 0.8.1 (htslib interface for python). For the separated AAV and Backbone references, the coordinates of left ITR, right ITR and rAAV genome in the reference "Cassette-AAV-CMV-GFP-hTK" are [1,130], [4665,4794] and [1,4794], respectively. For the fused AAV and Backbone reference, the coordinates of left ITR, right ITR and rAAV genome in the reference "SSV9K2-CMV-GFP-HygroTK-bGHpA" are [1192,1321], [5856,5985] and [1192,5985], respectively.

Table S8. Index sequences

Sample id	Index Sequences	
	Replicate 1	Replicate 2
CsCl -DNase	TGACCA	CGATGT
CsCl +DNase	GCCAAT	TGACCA
AVB -DNase	CTTGTA	ACAGTG
AVB +DNase	GTGAAA	GCCAAT
IEX -DNase	ACAGTG	CAGATC
IEX +DNase	CAGATC	CTTGTA
Negative control	CGATGT	AGTCAA
Internal normalizer	AGTCAA	GTGAAA

Indexed “all in one” adapters compatible with Illumina TrueSeq protocol were obtained from Sigma-Aldrich (Oligonucleotide sequences 2007-2014 Illumina, Inc. All rights reserved):

P5 adapter 5’_AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T_3’

P7 adapter 5’_P-GATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNATCTCGTATGCCG TCTTCTGCTTG_3’

5’: 5’ DNA end, 3’: 3’ DNA end, *: phosphorothioate linkage, P: phosphorylated end.

P7 adapter contains a six bases index used to identify samples after multiplexing (NNNNNN in P7 sequence), as described by Illumina. The table summarizes the indexes used for the 2 replicates of each samples. Indexes were randomized between the two SSV-Seq runs to avoid eventual index sequence bias.

SUPPLEMENTARY MATERIAL AND METHODS

rAAV vectors production and purification

All of the vectors were manufactured and characterized at INSERM UMR 1089 Vectors Production Core (Nantes, France), except for the Immune affinity chromatography step, which was performed at Genethon (Evry, France).

Vectors production

The rAAV 2/8 CMVp-eGFP-hygroTK-bGHpA was produced by transfecting HEK-293 cells from the laboratory working bank with the pSSV9-CMV-GFP-ires-HTK-bGHpA vector plasmid and the pDP8-Kana helper plasmid (containing AAV2 rep, AAV8 cap, and adenovirus E2A, VA RNA and E4 genes) as described in Ayuso et al, 2014, *Hum. Gene Ther.* **25**: 977–987. Similar to GMP production protocol, HEK-293 cells were cultured in DMEM medium w/o red phenol, with 4,5g/l glucose and 10% gamma-irradiated FBS (HyClone). Plasmids transfection was performed with jetPEI (PolyPlus Transfection Illkirch, France) on a total of three 10-chambers CellBIND Surface CellSTACK (Corning Life Sciences), in DMEM medium w/o red phenol, with 4,5g/l glucose (HyClone) and w/o FBS. After 72 h, the cells and the culture supernatant were collected and the crude bulk was split in three equal volumes for subsequent purifications.

Cesium Chloride (CsCl) gradients ultracentrifugation purification

The crude bulk was clarified by centrifugation, precipitated with a polyethylene glycol (PEG), treated with Benzonase Nuclease (Merck-Millipore) and purified by double Cesium Chloride gradient (CsCl) for enrichment in full particles, as extensively described in Ayuso et al, 2010, *Gene Ther.* **17**: 503–510.

Ion-exchange chromatography (IEX) purification

The crude bulk was clarified by filtration using Millipore Opticap XL2 Polysep disposable capsule filters (Merck-Millipore), and purified by three successive ion exchange chromatography steps:

- An anionic membrane, Mustang® Q XT Ion Exchange Chromatography Capsules (Pall Corporation). The clarified bulk was loaded with 20mM Tris buffer to reduce salinity to 40 mM and adjusted to pH 8.0 to capture the rAAV vector particles and then eluted with increasing concentration of 300 mM NaCl
- A multimodal weak cation exchanger XK26/20 column (GE Healthcare Life sciences). The previously eluted volume was loaded with 50mM MES buffer to reduce salinity to 30 mM and adjusted to pH 6,0 to capture the rAAV vector particles and eluted with increasing concentration of 600 mM NH4Cl
- An anionic monolith columns CIMmultus QA 80ml (Bia Separations). The previously eluted volume was loaded with 20mM Tris buffer to reduce salinity to 40 mM and adjusted to pH 8.5 to capture the rAAV vector particles and then eluted with increasing concentration of 300 mM NaCl

All three columns were already used several times for rAAV purification before this study, but were sanitized prior usage with the following protocol: 2 to 3 column volumes (CV) of water, incubation with 1N NaOH for 30 to 60 min, 2 to 3 CV of water, 5 to 10 CV of Tris NaCl buffer, 2 to 3 CV of water, 5 to 10 CV of EtOH 20%.

Immune affinity chromatography purification

The crude bulk was clarified by centrifugation, treated with Benzonase Nuclease (Merck-Millipore) and purified by immune affinity chromatography with a single AVB Sepharose High Performance column (GE Healthcare Life sciences) as described in Smith et al, 2009, *Mol. Ther. J. Am. Soc. Gene Ther.* **17**: 1888–1896.

Similar to IEX chromatography, the column was already used several times for rAAV purification before this study, but was sanitized prior usage with the following protocol: 2 to 3 column volumes (CV) of water, incubation with 0.1M H3PO4 (phosphoric acid) + 1M NaCl for 20 min, 2 to 3 CV of water, 5 to 10 CV of PBS buffer, 2 to 3 CV of water, 5 to 10 CV of EtOH 20%.

Concentration and formulation

The three rAAV batches obtained by CsCl, IEX and AVB purifications were concentrated by tangential flow filtration (TFF) with a 100 kDa molecular weight cut-off (GE Healthcare). The concentrated vectors were formulated in Dulbecco's phosphate-buffered saline (Lonza, Verviers, Belgium) containing 0.001% Pluronic F-68 (Gibco/Life Technologies).

Comparative efficacy of DNases

DNase treatment

Samples were treated with 20U of DNase I (Roche, Basel, Switzerland), 10U of TurboDNase (Life Technologies, Carlsbad, CA, USA), 250U of Benzonase (Merck, Billerica, MA, USA) or a mix of 4U of Plasmid-Safe (PS) and 10U of Baseline-ZERO (BL). They were incubated 2 hours at 37°C in a final volume of 200µL containing the buffers recommended by each manufacturers or an optimized buffer for our DNases mix (Baseline ZERO buffer, supplemented with 1mM of ATP). The reaction was stopped with 3mM of EDTA 30min at 75°C.

Quant-iT PicoGreen quantification

A working solution of the Quant-iT PicoGreen dsDNA reagent (Life technologies) was prepared by diluting the concentrated DMSO solution 400-fold in TE. Samples and phage-λ DNA stock solution were diluted in TE buffer (10mM Tris-HCl, 1mM EDTA, pH 7.5) to a final volume of 10 µL before the addition of 90 µL of the working solution. After 5 minutes at room temperature, protected from light, fluorescein intensities were quantified using a fluorescence plate reader (VictorX3, Perkin Elmer). DNA concentration of each sample was determined from the standard curve generated by phage-λ dilutions.

Quality controls of rAAV productions

Particle titer

The particles concentration was determined using the Progen AAV8 Titration ELISA kit (Progen Biotechnik GmbH, Heidelberg, Germany) according to the manufacturer's instructions.

Vector genome and DNA contaminants quantification

Each rAAV vector was treated with 4U of Plasmid-Safe and 10U of Baseline-ZERO for 2h at 37°C. The reaction was stopped with 3mM of EDTA for 30 min at 75°C. DNA extraction was carried out using the High Pure Viral Nucleic Acid kit (Roche). Then, the vector genome concentration was determined by Taqman qPCR targeting the ITRs (ITR2), the transgene (GFP1 and GFP2), the promoter (CMVp) or the polyadenylation signal (bGHpA). DNA contaminants were also quantified by qPCR targeting plasmid backbone (KanaR), helper plasmid (cap8, rep2 and E4) and human genome (human ALB1). qPCR reactions were performed with the StepOne Plus Real-time PCR system (Life Technologies) in a final volume of 20µL using the Premix Ex Taq kit (Takara Bio Inc., Otsu, Japan). Conditions for each reaction are detailed in **Table S9**, following MIQE guidelines (Bustin,S.A. et al. (2009). Clin. Chem., 55, 611–622.).

Protein purity and identity

The purity and identity were evaluated by SDS-PAGE, using Coomassie Blue (Imperial^{TM} Protein Stain, Thermo Fisher Scientific Inc., Waltham, MA, USA) and silver staining (PlusOneTM Silver Stain kit, Protein, GE Healthcare, Little Chalfont, United Kingdom) according to the manufacturer's recommendations. The purity relative to non-vector impurities visible on stained gels was determined using GeneTools (Syngene, Frederick, MD, USA).

In process NGS libraries preparation controls

Controls of second strand synthesis

A reaction containing ¼ fluorescein-12-dUTP and ¾ dTTP (Thermo Fisher Scientific Inc.), instead of dTTP alone, was performed similar to the protocol described in the manuscript material and methods. The mix was then purified using NucleoSpin®Gel and PCR Clean-up kit (Macherey-Nagel) and the fluorescence intensity was quantified using Victor X3 (Perkin Elmer).

DNA quantification

Aliquots were sampled after each major step of SSV-Seq protocol, (1) after the proteinase K treatment, (2) after the rAAV DNA extraction and (3) after the double strand synthesis purification. From these samples, rAAV genomes and plasmid backbone contaminants were quantified by Taqman qPCR targeting BGHpA sequence and KanaR sequence, respectively.

Control of NGS library preparation

The size of fragments was verified after sonication and after NGS library preparation by the Agilent 2100 Bioanalyzer system using High sensitivity DNA chips (Agilent Technologies, Santa Clara, CA, USA), according to the manufacturer’s guidelines.