Gene Therapy (2015) **22,** 934–946

www.nature.com/gt

ORIGINAL ARTICLE

AAV ancestral reconstruction library enables selection of broadly infectious viral variants

J Santiago-Ortiz[1,5], DS Ojala[1,5], O Westesson[2], JR Weinstein[3], SY Wong[2], A Steinsapir[1], S Kumar[2], I Holmes[2] and DV Schaffer[1,2,4]

Adeno-associated virus (AAV) vectors have achieved clinical efficacy in treating several diseases. However, enhanced vectors are required to extend these landmark successes to other indications and protein engineering approaches may provide the necessary vector improvements to address such unmet medical needs. To generate new capsid variants with potentially enhanced infectious properties and to gain insights into AAV's evolutionary history, we computationally designed and experimentally constructed a putative ancestral AAV library. Combinatorial variations at 32 amino acid sites were introduced to account for uncertainty in their identities. We then analyzed the evolutionary flexibility of these residues, the majority of which have not been previously studied, by subjecting the library to iterative selection on a representative cell line panel. The resulting variants exhibited transduction efficiencies comparable to the most efficient extant serotypes and, in general, ancestral libraries were broadly infectious across the cell line panel, indicating that they favored promiscuity over specificity. Interestingly, putative ancestral AAVs were more thermostable than modern serotypes and did not use sialic acids, galactose or heparan sulfate proteoglycans for cellular entry. Finally, variants mediated 19- to 31-fold higher gene expression in the muscle compared with AAV1, a clinically used serotype for muscle delivery, highlighting their promise for gene therapy.

Gene Therapy (2015) **22,** 934–946; doi:10.1038/gt.2015.74

INTRODUCTION

Advances in DNA sequencing, synthesis and computational phylogenetic analyses are enabling the computational reconstruction and experimental investigation of ancestral protein variants. Following the first ancestral reconstruction study, which resurrected a functional, ancestral digestive ribonuclease from an extinct bovid ruminant using the parsimony principle,[1] reconstructions and functional analyses have been carried out on inferred ancestral proteins belonging to eubacteria, bony vertebrates, mammals and the least common ancestor of higher primates using several inference methods including the parsimony, consensus, Bayesian distance and maximum likelihood methods.[2] Such ancestral reconstructions and subsequent analysis of resurrected variants have yielded insights into the conditions that led to protein evolution as well as the continuous adaption of organisms to changing environmental conditions.[3]

Ancestral reconstructions have also been harnessed to incorporate additional sequence diversity into genetic libraries for protein engineering. For instance, small libraries of resurrected ancestral variants were used in evolutionary studies of protein diversification[3–5] and to generate variants that are more tolerant to deleterious mutations. Moreover, inferred ancestral sequences have been combined with extant sequences by swapping residues of interest (for example, residues in or close to an enzyme's catalytic site) in modern sequences with those of the inferred ancestor. This residue swapping approach was used in basic evolutionary studies[6] as well as to screen for variants with properties such as increased thermostability,[7] improved catalytic activity,[8] novel substrate binding[9] and higher solubility.[10] Ancestral reconstruction is thus a versatile approach to explore new sequence space for engineering proteins with novel or enhanced properties and it may likewise offer potential for gene therapy.

This approach has recently been extended to more complex, multimeric proteins including viruses. The evolutionary history of viruses is an especially interesting application given their rapid mutational rates, importance to public health and promise for gene therapy. For example, ancestral reconstructions of viral proteins have been generated with the goal of developing vaccine candidates against HIV-1 and influenza virus,[11,12] and to study the functionality and properties of the resurrected variants of HIV-1, influenza and coxsackievirus.[13,14] These studies demonstrated that viral reconstructions could recapitulate properties of modern variants, including immunogenicity, packaging, tropism and cell receptor dependencies. These properties are key to the viral life cycle and they are also important properties for viruses used as gene therapy vectors.

Adeno-associated virus (AAV) vectors are highly promising for gene therapy. AAVs are non-pathogenic[15] and can transduce numerous dividing and non-dividing cell types, leading to long-term expression in the latter.[16] AAV vectors have accordingly been used for gene therapy in various tissues including the liver, lung, brain, eye and muscle.[17,18] Furthermore, Glybera, the first gene therapy product approved in the European Union in 2012, employs an AAV1 vector.[19] The amino acid composition of the viral capsid, encoded by the *cap* gene, affects AAV tropism, cell receptor usage and susceptibility to anti-AAV-neutralizing antibodies.[20] These key properties in turn have an impact on

[1]Department of Chemical and Biomolecular Engineering, University of California, Berkeley, CA, USA; [2]Department of Bioengineering, University of California, Berkeley, CA, USA; [3]Department of Chemistry, University of California, Berkeley, CA, USA and [4]The Helen Wills Neuroscience Institute, University of California, Berkeley, CA, USA. Correspondence: Professor DV Schaffer, Department of Chemical and Biomolecular Engineering, University of California, 274 Stanley Hall, Berkeley, CA 94720-1462, USA.
E-mail: schaffer@berkeley.edu
[5]These authors contributed equally to this work.
Received 7 April 2015; revised 22 June 2015; accepted 8 July 2015; accepted article preview online 17 July 2015; advance online publication, 13 August 2015

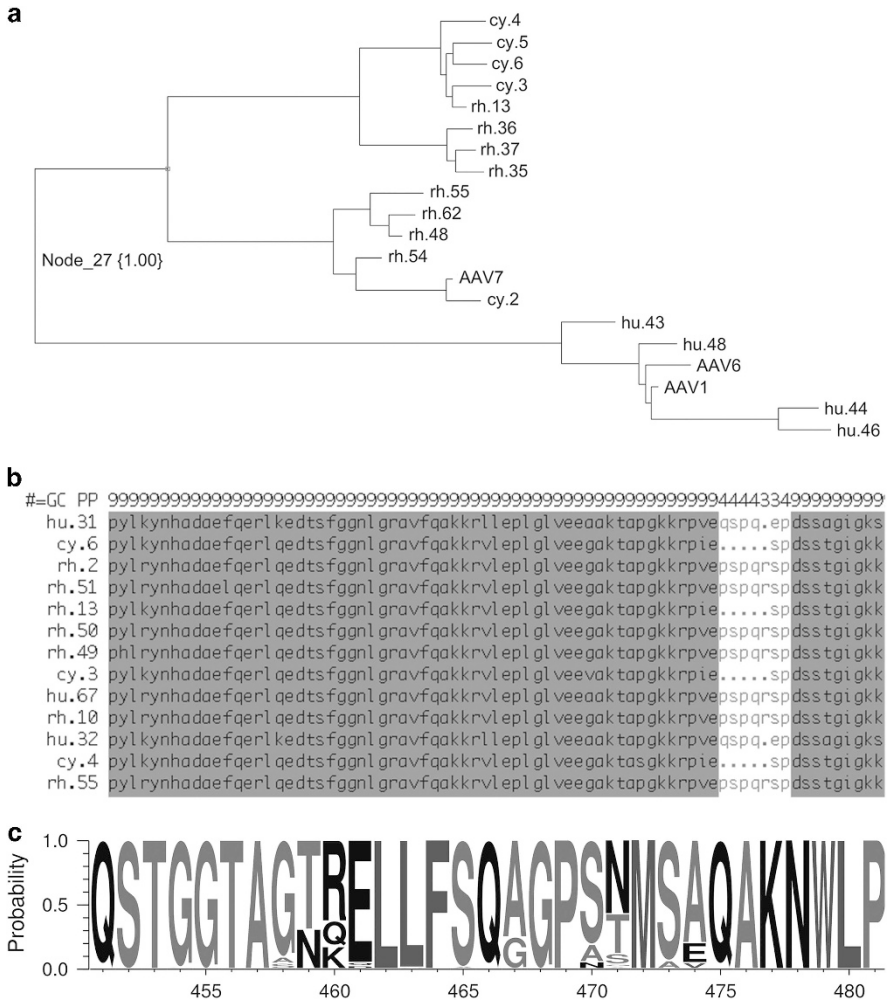

Figure 1. Ancestral AAV sequence reconstruction. (**a**) A phylogenetic tree relating a subset of extant AAV variants at node 27. Curly braced numbers indicate clade posterior probabilities.[23] The phylogenetic tree graphic was generated in Dendroscope.[62] (**b**) A multiple sequence alignment of a subset of AAV variants with column-specific confidence annotated along the top with single digits. Confidence ranges from above 0.9 (shaded grey) to 0.3–0.4 (shaded white). (**c**) A distribution of predicted ancestral amino acid sequences for node 27, residues 451–481. The character height of each amino acid is proportional to its posterior probability.

efficacy in therapeutic gene delivery, which is often limited by poor transduction of numerous cell types, off-target transduction, difficulties with biological transport barriers and neutralization by pre-existing anti-AAV antibodies.[18] However, extensive engineering of the AAV capsid, via modification of the *cap* gene, promises to improve numerous clinically relevant properties.[18]

Given the functional diversity of natural AAV serotypes, availability of numerous genetic sequences and demonstrated clinical efficacy of recombinant vectors, AAV is an intriguing candidate for ancestral reconstruction, which could further our understanding of its evolutionary history and plasticity. Interesting questions include whether reconstructed variants exhibit higher or lower infectivity on a range of cell types and whether they are relatively specific for particular cells—an attractive feature for many clinical applications—or are instead promiscuous, as are many extant serotypes. Finally, ancestral sequences and libraries may be useful starting materials for directed evolution studies,[8,21] especially considering that such AAVs likely gave rise to the modern serotypes with their divergent biological properties and tropism.

Motivated by these questions, we conducted an ancestral reconstruction of the AAV capsid. Acknowledging and taking advantage of the inherent ambiguity in reconstructing sequences containing highly divergent residues, we synthesized the inferred ancestral capsid not as a single 'best guess' sequence but rather as a large combinatorial library of candidate sequences incorporating degenerate residues at positions of low confidence. We then explored whether phenotypic selection of this ancestral sequence space using five cell lines representative of different tissues would lead to highly infectious variants and whether these would be promiscuous—that is, broadly infectious particles—or exhibit specific tropisms. The ancestral library was found to be fit, with packaging and transduction efficiencies that were on par with extant serotypes, and genetically selected variants were found to be broadly infectious on different cell lines. Furthermore, putative ancestral clones exhibited strong *in-vivo* gene delivery efficiency, underscoring the potential of such vectors for gene therapy applications.

RESULTS

Ancestral AAV sequence reconstruction

The goals of ancestral sequence reconstruction are, given a set of extant DNA sequences, to generate a phylogenetic tree and sequence alignment that relates these sequences and to infer the sequences of ancestral variants at different ancestral nodes. Accurate sequence reconstruction is challenging owing to ambiguity in the evolutionary relationships between extant

variants (which affects the phylogenetic tree-building step) as well as sequence divergence at highly variable residues (which affects the sequence alignment and ancestral reconstruction steps). As a starting point, we reconstructed the phylogeny of human, macaque and rhesus monkey AAV *cap* sequences retrieved from Genbank ($n = 52$).[22] We used MrBayes,[23] which conducts Bayesian Markov chain Monte Carlo simulation of tree space, to estimate the confidence values at each internal node (shown in curly braces in Figure 1a and Supplementary Figure S1). This approach generated a phylogenetic tree relating extant sequences, which is essentially a hypothesis concerning the evolutionary history of AAVs. Each branch on this tree depicts the evolutionary direction that diversified the sequences and each internal node represents a 'splitting' event where two AAV lineages diverged. With many ancestral nodes to choose from (full tree in Supplementary Figure S1), we selected node 27 (Figure 1a) based on its high confidence value (1.00), which minimizes one potential source of uncertainty (at the level of phylogenetic relationships between entire sequences) and thus improves confidence in the finer-grained downstream reconstruction of individual amino acids' evolutionary histories. This node is also the ancestor of serotypes with demonstrated clinical efficacy (AAV1 and Glybera), biomedical interest (AAV6)[24] or relative resistance to neutralizing antibodies (AAV7).[25]

We then used a Markov chain Monte Carlo alignment sampler, HandAlign,[26] to explore alignment space and predict the ancestral sequence of the most likely alignment at node 27. HandAlign generates a multiple sequence alignment, arranging the sequences of different variants in aligned 'columns' such that residues grouped in a column share a common ancestor (Figure 1b). It also performs the ancestral reconstruction simultaneously with the alignment and accounts for sequence insertions, deletions and character substitutions. Figure 1c shows the distribution of predicted amino acids as a sequence logo, with character heights proportional to posterior probabilities. The majority of amino acid positions could be predicted with high confidence ($\geqslant 0.90$) and thus represented residues highly conserved during evolution. However, as is common in ancestral reconstruction, other positions were less evolutionarily conserved and were thus predicted with lower confidence.

A DNA library was designed based on these results and residues above the 0.90 confidence value were fixed, whereas those below this confidence level were varied by introducing the two or three most likely amino acids (above a threshold value of 0.08), such that the fraction of library members containing each amino acid at a given position reflects the probability of that amino acid appearing in the sequence reconstructions. The locations, identities and synthesis frequencies of the 32 variable residues are presented in Table 1 and the most likely full ancestral *cap* amino acid sequence is shown in Supplementary Figure S2 and aligned with extant serotypes in Supplementary Figure S3. The ancestral *cap* library was synthesized (GeneArt, Life Technologies, Grand Island, NY, USA) and analysis of 61 sequenced clones from this library revealed that the amino acid frequencies at variable positions were not significantly different from the theoretical probabilities from the library ($P < 0.001$, see Materials and Methods), highlighting the correctness of the library synthesis.

Phenotypic selection of ancestral AAV library

Given the inherent probabilistic uncertainty of ancestral reconstruction, rather than investigating many possible candidate ancestral sequences one by one, we selected the library as a whole for functional clones. Specifically, after validating the

Table 1. Variable positions synthesized in ancestral AAV library

Position	*Residue 1*	*% Frequency*	*Residue 2*	*% Frequency*	*Residue 3*	*% Frequency*
264	T	55	Q	25	A	20
266	A	63	S	37		
268	S	70	A	30		
448	S	71	A	29		
459	T	69	N	31		
460	R	63	Q	20	K	17
467	A	75	G	25		
470	S	85	A	15		
471	N	60	T	32	S	8
474	A	83	E	16		
495	S	75	T	25		
516	D	91	N	9		
533	D	86	E	14		
547	Q	81	E	11	T	8
551	A	50	K	50		
555	T	54	A	46		
557	E	86	D	14		
561	M	62	L	28	I	10
563	S	80	N	19		
577	E	50	Q	50		
583	S	86	D	8	A	6
593	A	45	Q	39	V	16
596	A	81	T	19		
661	A	71	E	19	T	10
662	V	53	T	26	A	22
664	T	66	S	34		
665	P	64	A	26	Q	10
710	T	87	A	13		
717	N	69	D	31		
718	N	60	S	40		
719	E	79	D	21		
723	S	68	T	32		

Abbreviation: AAV, adeno-associated virus.

initial synthesized distribution of amino acids at the 32 variable positions, we probed how those positions would change when subjected to selective pressure for packaging and infectivity, which are key factors for successful viral replicative fitness during the natural evolution of AAV. The ancestral library was cloned into an AAV packaging plasmid and viral particles were produced by transfection into human embryonic kidney 293T cells as previously described.[27] The viral genomic titer was comparable to levels obtained when packaging libraries based on extant AAV serotypes (data not shown), indicating that the ancestral library can support robust packaging titers. The amino acid distribution at variable positions was only slightly altered by one round of packaging and we hypothesized that additional selective pressure for infectivity could reveal more about the significance of each variable position. We chose five cell lines representative of different tissues to conduct rounds of selection: C2C12 mouse myoblast cells, IB3-1 human lung epithelial cells, B16-F10 mouse skin melanoma cells, human embryonic kidney 293T cells and L0 human glioblastoma (GBM) tumor-initiating cells. Briefly, for each round 1×10^5 cells of each type were infected with iodixanol-purified, replication-competent AAV libraries at an initial genomic multiplicity of infection (MOI) of 5000 and successful virions were recovered by superinfecting the cells with adenovirus type 5 two days later. Six rounds of selection were conducted on each cell line, resulting in five independently selected pools, and the stringency of selection was increased during subsequent rounds by decreasing the genomic MOI (Supplementary Table S1).

To assess the progression of selection at each variable position, clones were sequenced ($n = 14$) from each library after initial viral packaging (hereafter referred to as post packaging), after three rounds of selection and after six rounds of selection. This analysis revealed a range of outcomes for each variable position across the different cell lines. Figure 2 shows the positions of the variable amino acids mapped onto the crystal structure of AAV1 (the most homologous serotype with a solved structure) and Figure 3 depicts the dominant amino acid at each of these positions after six rounds of selection (data after three rounds shown in Supplementary Figure S4) for each selected pool as a heat map, with darker shades representing higher convergence. As expected, selection for infection of cell lines led to increased convergence and Figure 4 shows the percentage change in amino acid frequency in round 6 relative to post packaging (increases in blue, decreases in red and change of amino acid in yellow); Supplementary Figure S5 shows the percentage change in amino acid frequency in round 6 relative to round 3. Some amino acid positions approached full convergence to the same residue across all cell lines; other positions were divergent or even acquired specific identities unique to only one cell line. The majority of residues unique to one cell line are located on the surface of the capsid and they could, for example, have a role in altering the affinity of capsid interactions with cell surface receptors.

To determine whether the changes in amino acid frequencies imparted by phenotypic selection were statistically significantly different from the initial synthesized distribution, we conducted Bayesian Dirichlet-multinomial model comparison tests (as described in Materials and Methods), to calculate the posterior probability that the two sets of variable amino acids come from different distributions. This analysis identified several amino acid positions that are significantly different after selection ($P < 0.05$, shown in green) and many more that are moderately different ($P < 0.5$, shown in yellow) (Figure 5).

Transduction efficiency of evolved ancestral libraries

Phenotypic selection could conceivably lead to specific infectivity on a given cell line or may alternatively increase overall infectivity in a promiscuous manner across all cell types. We investigated these possibilities by evaluating the transduction efficiency of evolved ancestral libraries on the cell line panel. The degree of convergence for each amino acid position after six rounds of selection is shown in Figure 3. Selection did not drive full convergence to a single sequence, potentially owing to the presence of neutral positions that conferred no selective advantage. Therefore, rather than packaging individual clones, the libraries selected on each cell line were each packaged as a pool of recombinant virus (at a low ratio of AAV helper plasmid per producer cell to minimize mosaic capsids), resulting in five distinct round-6 ancestral libraries; results thus represent overall or average library infectivities. High titer, iodixanol-purified recombinant AAV (rAAV) encoding the green fluorescent protein (GFP) was produced for the ancestral libraries as well as for natural serotypes AAV1-6, 8 and 9, for comparison of transduction efficiency and tropism. Infection at a genomic MOI of 2000 (or 32 000 for C2C12s) revealed a range of properties (Figure 6). Functionally selected ancestral libraries mediated high delivery efficiencies most comparable to AAV1 and AAV6, and generally superior to AAV4, AAV5, AAV8 and AAV9. Ancestral libraries were especially successful in infecting C2C12 and GBM cell lines relative to natural serotypes. Importantly, we observed a large increase in infectivity when comparing the synthesized vs. the round 6 ancestral libraries, suggesting phenotypic selection of advantageous amino acids at the variable positions. Interestingly, the libraries in general displayed broad infectivity across all cell lines, indicating that this reconstructed ancestral pool contains promiscuous AAVs, a property known to be advantageous for natural evolutionary adaptability.[28,29]

Characterization of the thermostability of candidate ancestral AAV variants

High thermostability and enhanced tolerance to mutations are also properties that could confer an evolutionary advantage to ancestral viral capsids.[3,7,30] We benchmarked the thermostability of AAV variants selected from our reconstructed pool against the natural serotypes AAV1, AAV2, AAV5 and AAV6 by assaying their transduction efficiency after heat treatment. Specifically, for initial analysis we chose the ancestral library selected on C2C12 cells and a representative variant from this library, C7. Virions packaged with self-complementary CMV (cytomegalovirus promoter)-GFP were treated for 10 min at different temperatures using a thermal gradient before being cooled down to 37 °C and used to infect 293T cells. We normalized the resulting fraction of GFP-expressing cells after treatment at each temperature to the sample incubated at 37 °C (Figure 7). Ancestral variants displayed higher thermostability than natural serotypes and showed moderate transduction levels even at the highest treatment temperature, 78 °C, which ablated transduction by natural serotypes. The obtained thermostabilities confirm those previously reported for natural serotypes,[31] which showed that AAV5 is more stable than AAV1, and that AAV2 is less stable than both. Enhanced thermostability of the ancestral variants in general could enable an increased tolerance to destabilizing mutations and consequently a higher evolutionary adaptability.

Characterization of ancestral AAV glycan dependencies and susceptibility to neutralizing antibodies

Our *in-vitro* transduction experiments demonstrated the broad infectivity of reconstructed variants. Given that ancestral node 27 gave rise to AAV1 and AAV6, we were interested in determining whether the candidate ancestral clones shared the same glycan dependencies, or whether those evolved later. AAV1 and AAV6 use both α-2,3 and α-2,6 *N*-linked sialic acids as their primary receptor and AAV6 has moderate affinity for heparan sulfate proteoglycans (HSPGs).[24] To probe HSPG usage, we transduced parental CHO-K1 cells and the pgsA CHO variant line deficient in HSPG. To examine sialic acid dependence, we transduced parental

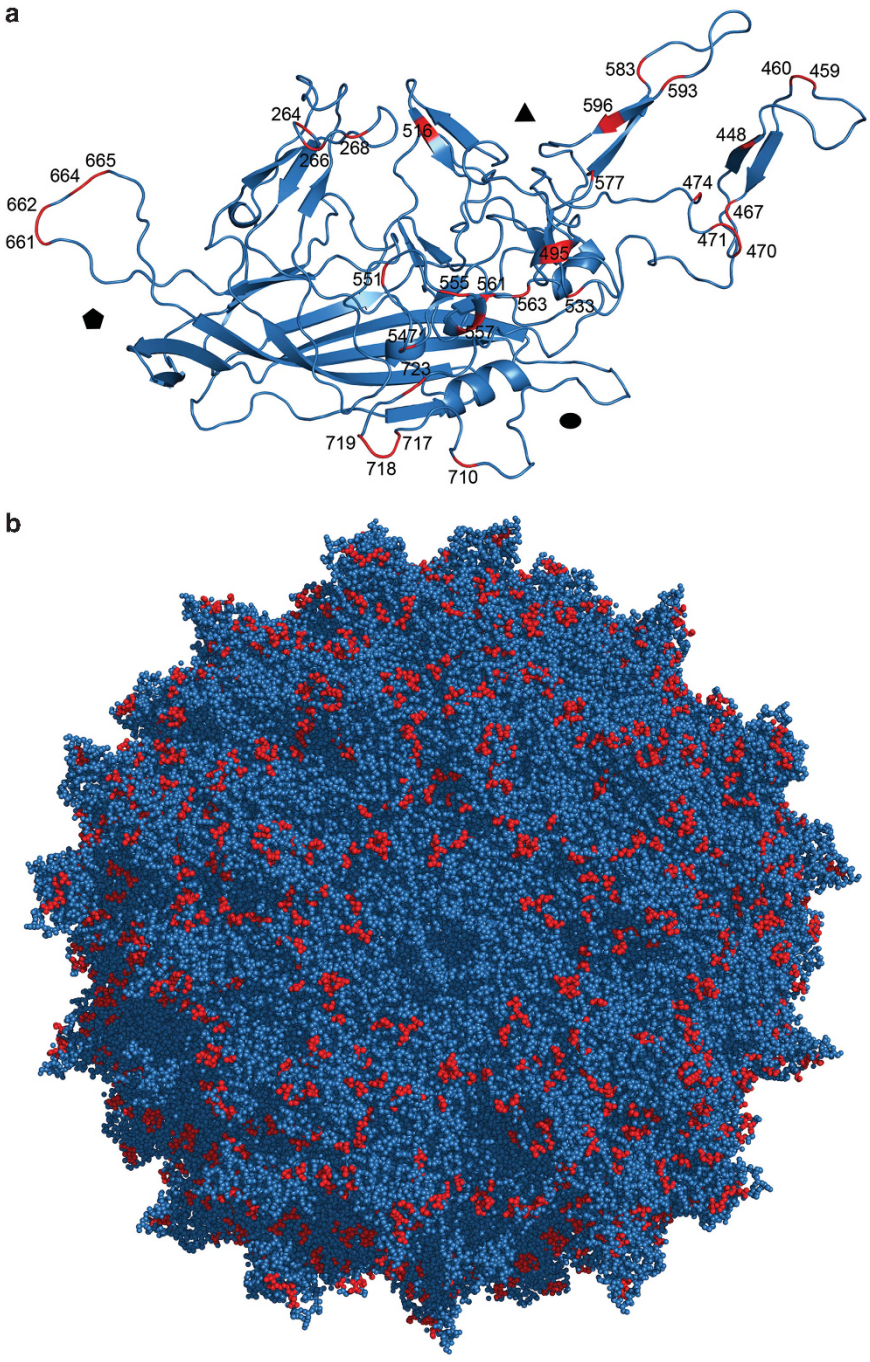

Figure 2. Variable residues mapped to the crystal structure of homologous AAV1, the closest AAV relative with an available structure. A three-dimensional molecular model of the AAV1 capsid was generated in PyMOL.[63] An amino acid alignment of the ancestral AAV sequence with AAV1 was used to map the highlighted residues to the (**a**) individual asymmetric unit and (**b**) full biological assembly.

Pro5 CHO cells presenting glycans with both *N*- and *O*-linked sialic acids, a Lec2 CHO variant cell line deficient in all *N*- and *O*-linked sialic acids and a Lec1 line deficient in complex and hybrid type *N*-glycans including sialic acids[32] (Figure 8b). Interestingly, candidate ancestral AAVs exhibited no dependence on HSPG or *N*- and *O*-linked sialic acids (Figure 8a). We also verified that selected individual clones exhibited similar transduction behavior as the evolved libraries (Supplementary Figure S6).

We next examined whether ancestral AAVs were neutralized by antibodies against a broad range of contemporary AAVs, in particular human intravenous immunoglobulin that contains polyclonal antibodies against extant serotypes due to natural exposure across the human population. *In-vitro* incubation with intravenous immunoglobulin strongly reduced transduction of ancestral libraries and the AAV1 control (Supplementary Figure S7), indicating that this ancestral pool is not highly serologically distinct from its progeny. Additional capsid engineering may be necessary to address this clinically relevant problem.

Characterization of ancestral variants *in vivo* in mouse gastrocnemius muscle

On finding that the ancestral AAV libraries exhibited efficiencies comparable to or in some cases higher than extant serotypes on a panel of cell lines from representative tissues, we next probed

Amino acid	Theoretical distribution	Synthesized library	Post-packaging	C2C12 round 6	293T round 6	IB3 round 6	GBM round 6	B16 round 6
264	T, 55%	T, 53%	A, 52%	Q, 86%	A, 79%	A, 71%	A, 79%	Q, 50%
266	A, 63%	A, 54%	S, 57%	S, 100%	S, 86%	S, 100%	S, 86%	A, 56%
268	S, 70%	S, 72%	S, 87%	S, 100%	S, 100%	S, 100%	S, 100%	S, 100%
448	S, 71%	S, 56%	A, 52%	S, 79%	S, 64%	S, 57%	A, 57%	S, 56%
459	T, 69%	T, 79%	T, 61%	N, 100%	T, 100%	T, 100%	T, 100%	T, 88%
460	R, 63%	R, 79%	R, 78%	R, 86%	R, 93%	R, 100%	R, 93%	R, 88%
467	A, 75%	A, 79%	A, 61%	G, 86%	A, 93%	G, 79%	A, 64%	A, 75%
470	S, 85%	S, 96%	S, 87%	A, 79%	S, 77%	S, 93%	S, 79%	S, 94%
471	N, 60%	N, 67%	N, 83%	N, 100%	N, 79%	T, 71%	N, 86%	N, 50%
474	A, 83%	A, 94%	A, 87%	A, 100%	A, 100%	A, 100%	A, 100%	A, 100%
495	S, 75%	S, 89%	S, 87%	S, 100%	S, 67%	S, 100%	S, 50%	S, 81%
516	D, 91%	D, 98%	D, 100%	D, 100%	D, 100%	D, 100%	D, 100%	D, 88%
533	D, 86%	D, 90%	D, 87%	E, 79%	D, 93%	D, 86%	D, 100%	D, 88%
547	Q, 81%	Q, 79%	Q, 83%	Q, 93%	Q, 93%	Q, 100%	Q, 100%	Q, 100%
551	A, 50%	A, 58%	A, 64%	A, 100%	A, 67%	A, 100%	A, 100%	A, 100%
555	T, 54%	T, 52%	T, 59%	T, 79%	A, 60%	A, 57%	A, 71%	A, 56%
557	E, 86%	E, 80%	E, 86%	D, 79%	E, 67%	E, 64%	E, 57%	E, 75%
561	M, 62%	M, 75%	M, 68%	M, 93%	M, 93%	M, 64%	M, 57%	M, 75%
563	S, 80%	S, 65%	S, 73%	S, 79%	S, 87%	N, 86%	S, 50%	S, 69%
577	E, 50%	E, 55%	Q, 59%	Q, 100%	Q, 60%	Q, 100%	Q, 100%	Q, 88%
583	S, 86%	S, 80%	S, 77%	S, 100%	S, 73%	S, 100%	S, 100%	S, 81%
593	A, 45%	Q, 49%	Q, 45%	Q, 79%	A, 60%	V, 86%	A, 71%	Q, 44%
596	A, 81%	A, 69%	A, 68%	T, 93%	A, 80%	T, 64%	T, 79%	T, 56%
661	A, 71%	A, 82%	A, 82%	A, 100%	A, 64%	A, 64%	A, 57%	A, 100%
662	V, 53%	V, 69%	V, 68%	V, 93%	V, 57%	T, 71%	V, 64%	V, 69%
664	T, 66%	T, 87%	T, 82%	S, 71%	S, 50%	T, 93%	T, 86%	T, 88%
665	P, 64%	P, 65%	P, 77%	P, 100%	P, 79%	P, 93%	P, 71%	P, 100%
710	T, 87%	T, 90%	T, 100%	T, 100%	T, 100%	T, 100%	T, 100%	T, 100%
717	N, 69%	N, 81%	N, 77%	N, 100%	N, 100%	N, 79%	N, 79%	N, 100%
718	N, 60%	N, 83%	N, 59%	S, 79%	S, 93%	N, 57%	S, 57%	S, 94%
719	E, 79%	E, 63%	E, 82%	E, 100%	E, 64%	E, 100%	E, 100%	E, 100%
723	S, 68%	S, 79%	S, 77%	S, 100%	T, 57%	S, 86%	T, 71%	S, 81%

Figure 3. Dominant amino acids at variable positions after six rounds of selection. A heat map was generated based on the frequency of the most common amino acid at each position in the different libraries. The dominant amino acid and frequency at each position were determined based on sequencing results from individual clones $n = 61$ (synthesized library), $n = 23$ (post packaging) and $n = 14$ (for each ancestral library after selection on respective cell lines).

Amino acid	Synthesized library → Post-packaging	Post-packaging → C2C12 round 6	Post-packaging → 293T round 6	Post-packaging → IB3 round 6	Post-packaging → GBM round 6	Post-packaging → B16 round 6
264	T, 53%. → A, 52%.	A, 52%. → Q, 86%.	26 A	19 A	26 A	A, 52%. → Q, 50%.
266	A, 54%. → S, 57%.	43 S	29 S	43 S	29 S	S, 57%. → A, 56%.
268	15 S	13 S	13 S	13 S	13 S	13 S
448	S, 56%. → A, 52%.	A, 52%. → S, 79%.	A, 52%. → S, 64%.	A, 52%. → S, 57%.	5 A	A, 52%. → S, 56%.
459	-18 T	T, 61%. → N, 100%.	39 T	39 T	39 T	27 T
460	-1 R	7 R	15 R	22 R	15 R	9 R
467	-18 A	A, 61%. → G, 86%.	32 A	A, 61%. → G, 79%.	3 A	14 A
470	-9 S	S, 87%. → A, 79%.	-10 S	6 S	-8 S	7 S
471	16 N	17 N	-4 N	N, 83%. → T, 71%.	3 N	-33 N
474	-7 A	13 A	13 A	13 A	13 A	13 A
495	-2 S	13 S	-20 S	13 S	-37 S	-6 S
516	2 D	0 D	0 D	0 D	0 D	-13 D
533	-3 D	D, 87%. → E, 79%.	6 D	-1 D	13 D	1 D
547	4 Q	10 Q	11 Q	17 Q	17 Q	17 Q
551	6 A	36 A	3 A	36 A	36 A	36 A
555	7 T	19 T	T, 59%. → A, 60%.	T, 59%. → A, 57%.	T, 59%. → A, 71%.	T, 59%. → A, 56%.
557	6 E	E, 86%. → D, 79%.	-20 E	-22 E	-29 E	-11 E
561	-7 M	25 M	25 M	-4 M	-11 M	7 M
563	8 S	6 S	14 S	S, 73%. → N, 86%.	-23 S	-4 S
577	E, 55%. → Q, 59%.	41 Q	1 Q	41 Q	41 Q	28 Q
583	-3 S	23 S	-4 S	23 S	23 S	4 S
593	-4 Q	33 Q	Q, 45%. → A, 60%.	Q, 45%. → V, 86%.	Q, 45%. → A, 71%.	-2 Q
596	-1 A	A, 68%. → T, 93%.	12 A	A, 68%. → T, 64%.	A, 68%. → T, 79%.	A, 68%. → T, 56%.
661	0 A	18 A	-18 A	-18 A	-25 A	18 A
662	-1 V	25 V	-11 V	V, 68%. → T, 71%.	-4 V	1 V
664	-5 T	T, 82%. → S, 71%.	T, 82%. → S, 50%.	11 T	4 T	6 T
665	12 P	23 P	1 P	16 P	-6 P	23 P
710	10 T	0 T	0 T	0 T	0 T	0 T
717	-4 N	23 N	23 N	1 N	1 N	23 N
718	-24 N	N, 59%. → S, 79%.	N, 59%. → S, 93%.	-2 N	N, 59%. → S, 57%.	N, 59%. → S, 94%.
719	19 E	18 E	-18 E	18 E	18 E	18 E
723	-2 S	23 S	S, 77%. → T, 57%.	8 S	S, 77%. → T, 71%.	4 S

Figure 4. Change in amino acid frequency at variable positions after six rounds of selection. The percent change in amino acid frequency between the post-packaging library and evolved libraries after six rounds of selection on each cell line was calculated. If the identity of the dominant amino acid did not change, the increase (blue) or decrease (red) in frequency is displayed. If selection resulted in a change in amino acid identity at that position, the new amino acid and frequency is shown (yellow).

			C2C12			293T			IB3			B16			GBM		
Amino acid	Theoretical vs. Synth.	Synth. vs. PP	PP vs. R3	PP vs. R6	R3 vs. R6	PP vs. R3	PP vs. R6	R3 vs. R6	PP vs. R3	PP vs. R6	R3 vs. R6	PP vs. R3	PP vs. R6	R3 vs. R6	PP vs. R3	PP vs. R6	R3 vs. R6
264	0.000	7.6	99.8	100.0	0.2	14.6	0.5	0.4	57.9	54.4	0.1	91.5	95.5	0.1	0.3	37.4	1.3
266	0.000	0.0	81.4	71.8	0.8	0.1	0.5	0.4	0.6	71.8	3.1	0.1	0.1	0.1	0.2	0.5	0.1
268	0.000	0.1	3.2	2.9	0.8	0.1	2.9	9.9	3.1	2.9	0.8	2.9	3.2	0.8	0.1	2.9	1.5
448	0.000	0.0	0.1	0.4	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.3	0.1	0.1	0.1
459	0.000	0.1	0.1	98.7	64.5	0.5	55.1	2.6	60.9	55.1	0.8	1.4	0.4	0.2	0.2	55.1	6.5
460	0.000	0.0	0.1	0.1	0.1	0.2	0.3	0.1	0.3	8.5	1.5	0.3	0.2	0.2	0.1	0.3	0.2
467	0.000	0.1	0.1	5.3	12.5	0.1	1.4	3.3	0.1	1.3	2.2	0.1	0.1	0.1	0.1	0.1	0.1
470	0.035	0.3	0.2	81.4	3.2	0.3	39.4	0.5	7.1	7.1	0.2	0.1	0.1	0.2	0.1	0.9	0.8
471	0.000	0.1	0.1	5.0	2.8	7.6	0.1	1.8	0.1	17.5	0.8	0.3	1.7	0.3	0.4	0.1	0.6
474	0.000	0.1	3.2	2.9	0.8	3.4	2.9	0.7	0.1	2.9	1.5	0.1	3.2	3.8	3.1	2.9	0.8
495	0.000	0.0	0.1	2.9	1.4	0.2	0.2	0.1	0.1	2.9	3.1	0.1	0.1	0.1	0.1	1.5	1.1
516	0.001	0.5	0.6	0.6	0.8	0.6	0.6	0.7	0.6	0.6	0.8	0.6	3.9	2.8	0.6	0.6	0.8
533	0.000	0.0	1.7	81.4	0.3	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	3.1	2.9	0.8
547	0.000	0.0	5.8	0.6	1.7	1.2	0.2	10.2	5.4	5.0	0.8	0.0	5.8	8.8	0.6	5.0	1.5
551	0.000	0.0	0.1	43.3	11.6	0.2	0.1	0.1	0.3	43.3	3.1	1.0	53.7	1.7	0.1	43.3	6.5
555	0.000	0.0	0.1	0.2	0.3	2.4	0.1	0.2	0.4	0.1	0.1	0.1	0.1	0.1	0.1	0.4	0.8
557	0.000	0.0	0.3	75.9	1.3	0.4	0.2	0.1	0.1	0.2	0.1	0.1	0.1	0.1	3.3	0.5	62.7
561	0.000	0.0	1.1	0.8	0.2	0.0	0.9	0.3	31.3	0.1	37.1	0.0	2.7	0.5	0.2	0.0	3.0
563	0.000	0.0	0.1	0.1	0.1	0.1	0.1	0.4	18.6	35.0	100.0	0.1	0.1	0.1	0.5	0.2	0.1
577	0.000	0.1	72.1	61.3	0.8	0.6	0.1	0.5	66.9	61.3	0.8	0.4	0.5	0.1	66.9	61.3	0.8
583	0.000	0.0	11.7	9.6	0.8	0.0	0.0	0.0	10.6	9.6	0.8	0.2	0.1	0.1	10.6	9.6	0.8
593	0.000	0.0	3.2	3.5	0.3	0.1	0.1	0.0	0.1	57.0	98.5	0.0	0.0	0.0	0.0	11.5	0.4
596	0.000	0.0	0.8	66.4	0.5	0.1	0.1	0.1	0.1	0.4	0.8	0.1	0.2	0.3	0.1	3.9	2.2
661	0.000	0.0	0.1	5.5	22.9	0.1	0.0	0.0	0.0	7.0	8.8	0.0	6.4	3.8	0.0	0.1	0.0
662	0.000	0.0	0.3	0.7	0.2	0.1	0.2	0.1	0.0	3.0	1.1	0.0	0.1	0.2	0.1	0.0	0.2
664	0.000	0.0	0.3	13.9	0.3	0.1	2.1	0.5	0.1	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1
665	0.000	0.0	0.3	9.6	2.8	0.0	0.6	1.8	0.3	0.4	0.4	0.0	11.7	19.9	0.3	0.4	0.1
710	0.000	2.6	0.6	0.7	0.8	11.1	0.7	6.5	0.6	0.7	0.8	0.7	0.6	0.8	4.2	0.7	3.1
717	0.000	0.0	0.1	9.6	22.9	0.1	9.6	6.5	0.3	0.1	0.3	0.2	11.7	1.7	0.1	0.1	0.1
718	0.000	0.3	0.1	1.0	0.3	0.4	23.4	21.3	0.2	0.1	0.2	0.4	39.3	0.5	0.1	0.1	0.1
719	0.000	0.1	6.4	5.5	0.8	0.5	0.1	0.5	0.2	5.5	1.5	0.2	6.4	1.7	0.2	5.5	1.6
723	0.000	0.0	0.3	9.6	1.4	1.1	89.8	3.5	0.1	0.1	0.4	0.2	0.1	0.2	0.2	5.4	5.7

Figure 5. Identification of key variable residues by Bayesian Dirichlet-multinomial model comparison tests. A comparison of the two sets of variable amino acids was conducted to identify positions that changed significantly during selection. The posterior probability that the two sets of amino acids come from two different probability distributions was calculated assuming probability parameters that are Dirichlet distributed with low pseudocounts to reflect sparse observed sequences. Results colored green indicate a >95% chance that the sets came from different distributions, yellow a >50% chance, red a >5% chance and no color a <5% chance. Synth, synthesized library; PP, post packaging; R3, round 3 of selection; R6, round 6 of selection.

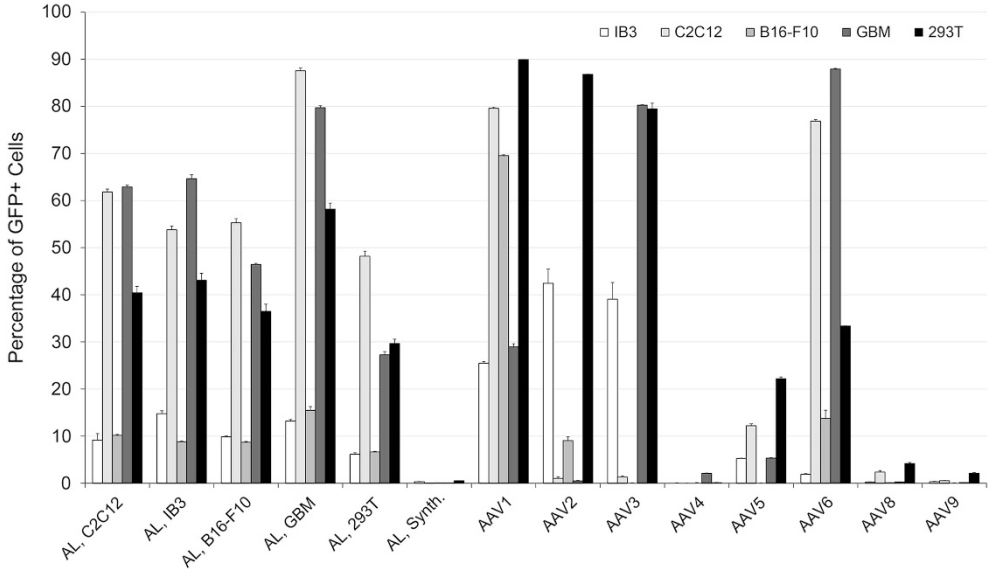

Figure 6. Transduction efficiency of ancestral libraries benchmarked against natural AAV serotypes. After six rounds of selection, viral genomic DNA was recovered from ancestral libraries and packaged as rAAV scCMV-GFP along with wild-type AAV1–6, 8 and 9. Cell lines were infected at a genomic MOI of 2000 (293T, IB3, B16-F10 and GBM) or 32 000 (C2C12). The fraction of GFP-expressing cells was quantified by flow cytometry 72 h later. Data are presented as mean ± s.e.m., $n = 3$. AL, ancestral library.

in-vivo infectivity. Based on the high transduction efficiencies of candidate ancestral AAVs on the most nonpermissive cell line (C2C12 mouse myoblasts), we chose to evaluate *in-vivo* transduction of mouse gastrocnemius muscle. In particular, individual ancestral variant clones from the selected viral pools (Supplementary Table S2) that were closest to the consensus sequences of libraries evolved on C2C12 (clones C4 and C7) and GBM cells (clone G4) were chosen, based on the efficiency of these two libraries in transducing C2C12 myoblasts *in vitro*. In addition, these variants were benchmarked against AAV1, given its clinical efficacy in muscle-targeted gene therapy.[33] We generated rAAV vectors expressing firefly luciferase under the control of the hybrid CAG (CMV early enhancer/chicken β-actin/splice acceptor of *β-globin* gene) promoter. A volume of 30-µl DNase-resistant genomic particles (5×10^{10} vg) was injected into each gastrocnemius muscle of BALB/c mice, and after 6 weeks mice were killed and tissue luciferase activities analyzed (Figure 9). Ancestral reconstruction variants yielded 19- to 31-fold higher transgene expression than AAV1 in gastrocnemius muscle, with variant C7 yielding the highest expression. Interestingly, variant C7 was the

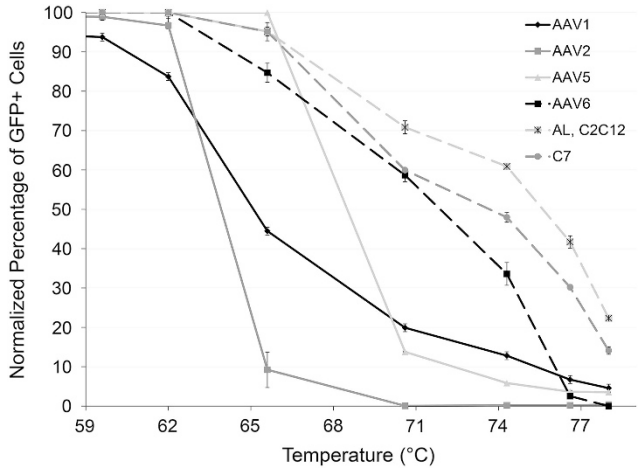

Figure 7. Candidate ancestral variants display higher thermostability than natural serotypes. The thermostability of the ancestral library selected on C2C12 cells and of the representative ancestral variant C7 was characterized and compared with that of natural serotypes 1, 2, 5 and 6. Virions packaged with scCMV-GFP were incubated at temperatures ranging from 59.6 °C to 78 °C for 10 min before being cooled down to 37 °C and used to infect 293T cells. The fraction of GFP-expressing cells was quantified by flow cytometry 72 h later. Data are presented, after being normalized to the fraction of GFP-expressing cells after incubation at 37 °C, as mean ± s.e.m., $n = 3$.

most abundant sequence (71%) in the round 6 ancestral library selected on C2C12 cells. These results demonstrate that candidate ancestral AAVs also exhibit high infectivity *in vivo* and even offer the potential to exceed the performance of the best contemporary natural serotypes in gene therapy applications.

DISCUSSION

Ancestral sequence reconstruction offers unique opportunities to study fundamental biological questions of virus evolution and fitness, including the characterization of ancestral sequence space relative to extant serotypes, the importance of mutational tolerance and evolutionary conservation, and the comparative advantages of promiscuous versus selective tropism. The primary challenge of ancestral reconstruction is to accurately infer an ancestral sequence despite uncertainty arising from sequence divergence within hypervariable regions of extant variants. We have combined sophisticated computational and library synthesis approaches to address this uncertainty and thereby generate a functional ancestral AAV library. We then studied the biological properties of this library to learn more about the evolutionary behavior of AAV and the gene therapy potential of reconstructed ancestral variants.

The posterior probability that an AAV ancestral sequence accurately reflects the actual ancestral virus is the product of the probabilities that each of the amino acids in the capsid protein is correctly predicted. At positions of high evolutionary convergence, the posterior probability nears 1.0; yet, there are many sites that diverged during evolution and thus cannot be predicted with such high confidence. Our library synthesis approach addressed this concern by introducing the 2 or 3 most likely amino acids at the 32 lowest confidence positions in the AAV *cap* protein. Interestingly, the majority of positions varied in our ancestral library have not been previously described in studies of the functional importance of single mutations to the AAV capsid.[34,35] Unlike previous ancestral reconstructions of enzymes and other proteins, which used single best guess ancestral sequences,[11,36] or which sampled only a small fraction of library variants owing to the low throughput of enzymatic assays,[8,37] our massively parallel phenotypic selection enabled screening of a large library (and is limited only by the transformation efficiency of electrocompetent bacteria).

The selection strategy applied pressure for efficient packaging and transduction of cell types representing a variety of tissues. By comparing the frequencies of amino acids selected at variable positions to the theoretical ancestral sequence prediction, one can gain insights into both the accuracy of our sequence reconstruction as well as the functional role of each residue in AAV biology. Comparison of sequences from the synthesized library with those recovered after initial library packaging suggested that one round of packaging imposed no statistically significant changes on the amino acid distribution at variable positions (except for a low 0.076 probability change in preference from a threonine to an alanine at residue 264). However, with the selection for infectivity on a range of cell types, specific positions begin to diverge and differences between round 6 and post-packaging sequences were more significant than between round 3 and post-packaging sequences, likely because six rounds enabled a larger cumulative effect of positive selection. Genetic drift may also have a role but is unlikely to be the main driving force given that the time to fixation by genetic drift increases with population size[38] and a large number of virions ($>10^8$) was used in each sequential round of selection.

By comparing the ancestral libraries after six rounds of selection with the post-packaging library, we identified several trends in the level of convergence of the amino acid residues, suggesting these positions may have potential roles in modulating properties such as capsid stability and infectivity. Some amino acid positions approached full convergence to the same residue across all cell lines (268, 460, 474, 516, 547, 583, 665, 710, 717 and 719); these positions are distributed throughout the capsid and may, for example, be important for core viral functions such as capsid stability, uncoating or endosomal escape. Others showed more divergent outcomes across different cell lines (264, 467, 593, 664 and 723) and may be neutral with respect to overall fitness. Finally, some positions (459, 470, 471, 533, 555, 596, 662 and 718) acquired identities specific to a given cell line and may confer an infectious advantage on each respective cell line.

Positions 264 and 459 showed the strongest evidence of change due to selection ($P < 0.05$). Position 459 is prominently exposed on loop IV of the AAV capsid surface. Position 264 is positioned on loop I of the capsid and has been identified as a key determinant of muscle tropism in the rationally engineered variant AAV2.5.[39] There is also suggestive evidence of selection at positions 266, 470, 533, 551, 557, 577, 596 and 723 in various libraries ($P < 0.5$). Position 533 has been described as a key contributor to infectivity and glycan dependence in our previously evolved variant ShH10, a vector differing by only four amino acids from AAV6 but exhibiting unique tropism in the retina.[32] In addition, Lochrie *et al.*[40] examined several other of these positions in their thorough mutational analysis of the AAV2 serotype, although AAV2 lies in a different phylogenetic clade than ancestral node 27. The characterization of these variable positions is therefore novel and lessons learned may inform targeted mutagenesis efforts to improve the fitness of extant variants.

The assembly-activating protein, which is involved in directing capsid proteins to the nucleolus and in assembly of the viral capsid in this organelle,[20] is translated from an alternative open reading frame with a non-canonical CTG start codon present within the *cap* gene. This alternate open reading frame is also present in the ancestral reconstruction of the AAV capsid at node 27 (Supplementary Figure S2b). Three of the variable residues (positions 264, 266 and 268) are present within the assembly-activating protein open reading frame and the putative ancestral assembly-activating protein sequence is otherwise conserved across the reconstruction. As discussed above, residue 264 is among the positions that showed strong and statistically

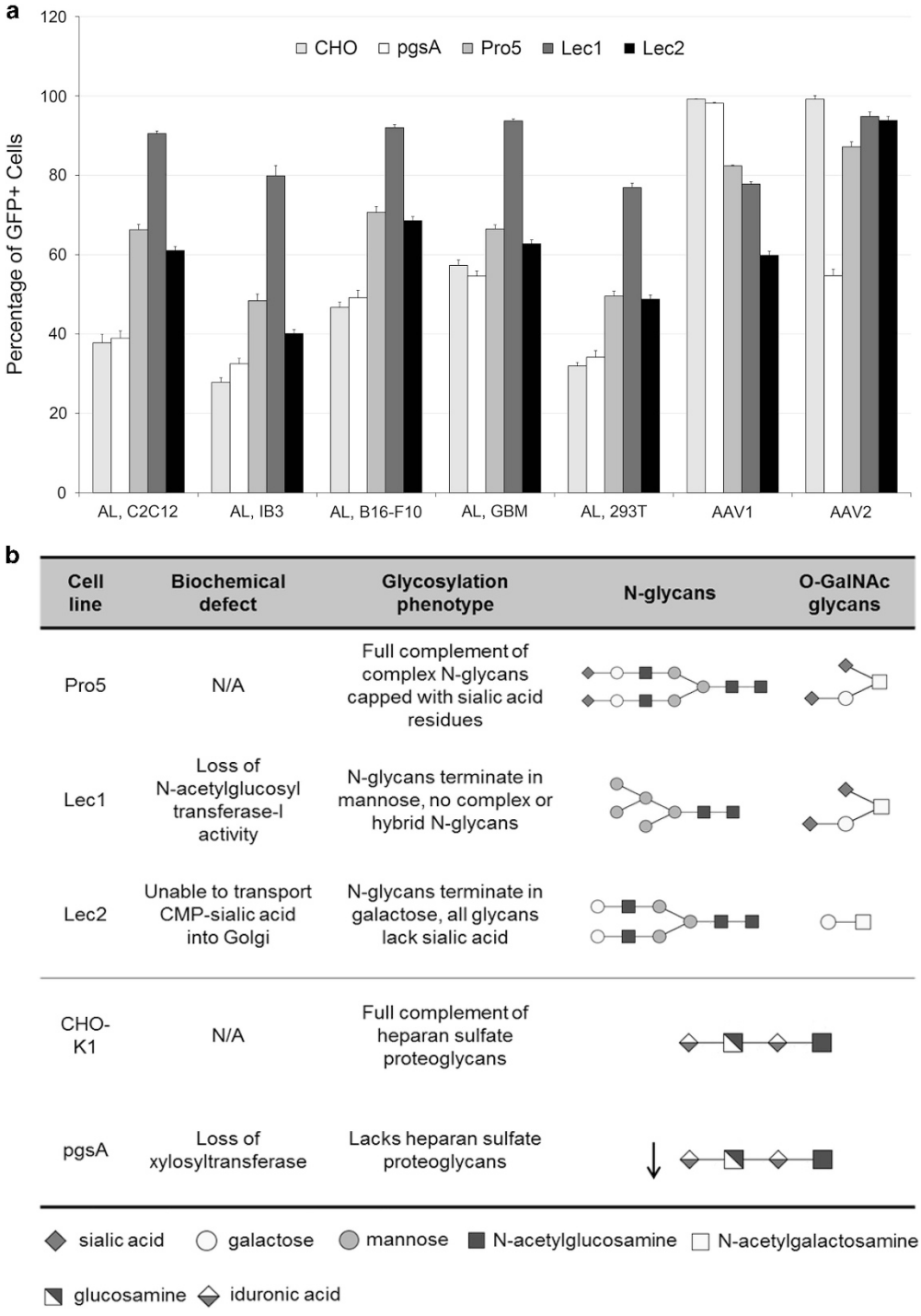

Figure 8. Glycan dependency of candidate ancestral AAV variants. (**a**) After six rounds of selection, the transduction efficiency of ancestral libraries carrying scCMV-GFP was quantified by flow cytometry 72 h after infection at a genomic MOI of 2000 (Pro5, Lec1 and Lec2) and 50 000 (CHO-K1 and pgsA). The CHO-K1/pgsA comparison examines HSPG dependence, whereas Pro5/Lec1 and Pro5/Lec2 probe sialic acid dependence. Data are presented as mean ± s.e.m., $n = 3$. (**b**) Glycans present on CHO glycosylation mutants. AL, ancestral library.

significant changes after six rounds of packaging and infection on the cell line panel and it is possible that both capsid and assembly-activating protein may have undergone functional selective pressure during this process.

The *in-vitro* transduction results also demonstrate the importance of using a library approach coupled with selection. A single or small number of best guess sequences could likely include deleterious amino acids that significantly impact fitness. Indeed, our data show that the synthetized ancestral AAV library evaluated before rounds of selection reproducibly exhibited dramatically lower infectivity than libraries subjected to selective pressure. This is not surprising, given that numerous directed evolution studies demonstrate that even single point mutations can significantly alter enzyme activity or virus infectivity by several orders of magnitude.[34,41–43]

Interestingly, despite differences in amino acid composition at variable positions, the ancestral libraries selected for infecting a number of individual cell lines subsequently demonstrated broad

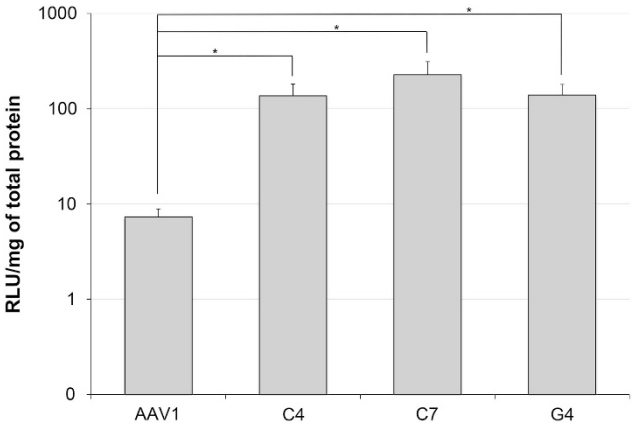

Figure 9. Evaluation of gastrocnemius muscle transduction. Luciferase activity measured in relative light units (RLUs) per mg protein was determined in gastrocnemius tissue homogenate 48 days after intramuscular administration of 5×10^{10} viral particles of ancestral clones C4, C7, G4 or AAV1 in adult mice. Controls injected with phosphate-buffered saline displayed no activity (data not shown). *Statistical difference of $P < 0.05$ by two-tailed Student's *t*-test.

tropism across all of these cell lines. Such promiscuity may have been rewarded during the natural evolution of AAVs, as the ability to replicate in different cell and tissue types enhances virus spread. In fact, most natural AAV serotypes exhibit broad tropism,[44,45] indicating that promiscuity continues to be a valued trait for natural evolution.

Such broad tropism, however, has important implications for gene therapy. In cases where disease pathologies affect multiple tissues and cell types (for example, lysosomal storage disorders), broader infectivity could be an advantageous trait. Expanded tropism may also be useful for infecting cell types refractory to infection by most AAV serotypes or for *ex-vivo* treatments of homogeneous cell populations where off-target infectivity is not a concern. However, in the majority of gene therapy applications, it is desirable to limit transgene expression to a target tissue for several important reasons, including risks associated with off-target expression of the transgene, off-target transduction leading to higher immune presentation and reaction, and higher overall dosages needed to overcome vector dilution into multiple tissues. This is true not only when vector is delivered via routes that lead to intentional exposure to multiple tissues (for example, intravascular delivery) but also for local injection into multiple tissues in which vector leakage into circulation can lead to widespread distribution to multiple organs. For example, biodistribution studies have shown the spread of AAV vectors to sites distant from the target tissue after injecting viral particles through the hepatic artery, intramuscularly or into the putamen of the brain.[46–48]

To address concerns with off-target transgene expression, strategies for controlling gene and protein expression including cell-type-specific promoters[49] and microRNA elements[50,51] are being explored to restrict expression to target cells. These approaches are promising, but would not address immune presentation of the capsid protein. Therefore, the optimal scenario is one in which selective AAV tropism is engineered through modification of the capsid protein. Directed evolution can generate vectors capable of targeted gene delivery[32] and evolution for enhanced AAV infectivity on a given target cell in general can enable a reduction in vector dose and thereby reduce the level of off-target transduction. Ancestral variants may be promising starting points for such directed evolution efforts given their high infectivity and representation of a capsid protein sequence space that is different from and complementary to extant serotypes.

High thermostability may also be an advantageous property for AAV engineering. Ancestral sequences have been correlated with increased thermostability in multiple studies[3,7,30] and, in fact, enriching for seemingly neutral mutations that resemble an ancestral sequence has been shown to increase protein kinetic and thermodynamic stability, and to improve the probability of acquiring new function mutations.[52] This work lends additional evidence of the correlation between ancestral sequences and thermostability by demonstrating that candidate ancestral AAV variants are more thermostable than contemporary serotypes.

We also characterized the glycan dependencies of ancestral variants and found that previously studied AAV glycan dependencies including *N*- and *O*-linked sialic acids, HSPGs and galactose were not used. It is conceivable that these dependences may have arisen more recently in the evolution along these AAV lineages.

In addition, we found that ancestral libraries were as susceptible to neutralizing antibodies as AAV1, suggesting that this ancestral reconstruction pool exhibits immunogenic properties similar to current serotypes. Multiple antigenic regions have been mapped on natural AAV serotypes, including AAV1, AAV2, AAV5 and AAV8.[53] Given that AAV1 is a descendant of the node 27 ancestral reconstruction, we aligned known AAV1 epitopes[54] with the ancestral reconstruction sequence. Mapped antigenic regions corresponding to AAV1 residues 496–499, 583, 588–591 and 597 were conserved in the ancestral reconstruction. In addition, the ancestral sequence is identical to several known AAV2 antigenic regions including residues 272–281, 369–378 and 562–573.[55] Such conserved regions may contribute to the observed susceptibility of ancestral variants to neutralizing antibodies.

Interestingly, previous studies have also demonstrated cross-seroreactivity between ancestral and extant viral capsids. In particular, antiserum against extant viruses has been shown to neutralize reconstructed ancestral variants[14] and ancestral viruses can elicit neutralizing antibodies that protect against currently circulating strains, a property that has been exploited for the development of vaccine candidates.[11,12]. Neutralizing antibodies may therefore pose a significant clinical challenge for ancestral vectors. Further capsid engineering under a strong selective pressure for evading neutralizing antibodies may enable selection of combinations of mutations that promote antibody evasion.[56] For example, there are variable residues in the ancestral reconstruction that map to antigenic regions corresponding to AAV1 residues 456–459, 494, 582 and 593–595, and to antigenic regions in other serotypes.[54] Mutations in these regions could disrupt the binding of antibodies to capsid epitopes and could potentially be combined with other mutagenesis strategies to engineer variants with enhanced antibody evasion properties.[56]

Ancestral AAVs demonstrated efficient *in-vitro* gene transfer to C2C12 mouse myoblast cells comparable to AAV1, a current gold standard for muscle transduction, yet used a different receptor for cell entry. This distinction may contribute to their efficient *in-vivo* infectivity, which impressively reached 19- to 31-fold higher levels of expression than AAV1 in mouse gastrocnemius muscle. If the improved expression observed with ancestral reconstruction vectors is reproducible in human muscle tissue, ancestral variants will be auspicious candidates for clinical translation.

In summary, our results indicate that a library of AAV variants representing sequence space around a key ancestral node is rich in broadly infectious variants with potential in gene therapy applications. We have taken initial steps in characterizing this sequence space by varying the amino acids at the lowest confidence positions identified by ancestral sequence reconstruction, followed by phenotypic selection to yield highly functional sets of amino acids at these locations. Sequence analysis of variable residues revealed a variety of outcomes ranging from

highly conserved residues to more neutral positions that are pliable to change. Selected variants were promiscuous in their infectivity but showed promise as recombinant vectors *in vitro* and *in vivo*, and the putative mutational tolerance and evolvability of this library could be further harnessed in directed evolution studies to overcome gene therapy challenges such as targeted gene delivery and immune evasion.

MATERIAL AND METHODS

Ancestral reconstruction

AAV *cap* sequences (n = 52) from Genbank,[22] including those from human and non-human primate origin, were incorporated in this analysis, starting from lists of AAV sequences published in previous phylogenetic analyses.[57,58] The MrBayes package[23] was used to perform Bayesian Markov chain Monte Carlo simulation of tree space and estimate the confidence values at each internal node. We then used the Markov chain Monte Carlo alignment sampler HandAlign[26] to explore alignment space and estimate regional confidence for the most likely alignment at node 27, discarding all but the sequences descended from this node. HandAlign generates a multiple sequence alignment, arranging the sequences of different variants in aligned 'columns' such that residues grouped in a column share a common ancestor. Each alignment column was modeled as a realization of the standard phylogenetic continuous-time Markov process of character evolution, using amino acid and empirical codon substitution rate matrices that were estimated from databases of aligned protein-coding sequence.[59] HandAlign performs the reconstruction simultaneously with the alignment and accounts for sequence insertions, deletions and character substitutions. The codon-level model was used to account for the possibility of synonymous substitutions with a phenotype at the DNA level; we also checked for the possibility of dual selection in overlapping reading frames ('overprinted' genes), by reconstructing both ancestral reading frames at the codon level. Neither of these subtle effects appeared significant enough to warrant prioritizing synonymous (silent, DNA level) variants over the many nonsynonymous amino acid variants.

Library construction and vector packaging

The reconstructed ancestral AAV *cap* sequence was synthesized (GeneArt, Life Technologies) with a library size of 5.6×10^{11}, greater than the theoretical diversity of 2.5×10^{11}. The library was digested with *Hin*dIII and *Not*I, and ligated into the replication competent AAV packaging plasmid pSub2. The resulting ligation reaction was electroporated into *Escherichia coli* for plasmid production and purification. Replication competent AAV was then packaged and purified by iodixanol density centrifugation as previously described.[56,60] DNase-resistant genomic titers were obtained via quantitative real-time PCR using a Biorad iCycler (Bio-Rad, Hercules, CA, USA) and Taqman probe (Biosearch Technologies, Novato, CA, USA).[60]

Cell culture

C2C12 mouse myoblast, B16-F10 skin melanoma cells, CHO-K1, pgsA, Pro5, Lec1 and Lec2 cells were obtained from the Tissue Culture Facility at the University of California, Berkeley, CA, USA. IB3-1 lung epithelial and human embryonic kidney 293T cells were obtained from American Type Culture Collection (Manassas, VA, USA). Unless otherwise noted, all cell lines were cultured in Dulbecco's modified Eagle's medium (Life Technologies) at 37 °C and 5% CO_2. L0 human GBM tumor-initiating cells were kindly provided by Dr Brent Reynolds (University of Florida, Gainesville, FL, USA) and propagated in neurosphere assay growth conditions[61] with serum-free media (Neurocult NS-A Proliferation kit, Stem Cell Technologies, Vancouver, BC, Canada) that contained epidermal growth factor (20 ng ml^{-1}, R&D, Minneapolis, MN, USA), basic fibroblast growth factor (10 ng ml^{-1}, R&D) and heparin (0.2% diluted in phosphate-buffered saline, Sigma-Aldrich, St Louis, MO, USA). IB3-1 cells were cultured in Dulbecco's modified Eagle's medium/F-12 (1:1) (Invitrogen, Carlsbad, CA, USA). CHO-K1 and pgsA cells were cultured in F-12K medium (American Type Culture Collection), and Pro5, Lec1 and Lec2 cells were cultured in minimal essential medium α-nucleosides (Life Technologies). Except for GBM culture, all media were supplemented with 10% fetal bovine serum (Invitrogen) and 1% penicillin/streptomycin (Invitrogen).

Library selection and evolution

All cell lines were seeded in six-well tissue culture plates at a density of 1×10^5 cells per well. One day after seeding, cells were infected with replication-competent AAV libraries. After 24 h of exposure, cells were superinfected with adenovirus serotype 5. Approximately 48 h later, cytopathic effect was observed and virions were collected by three freeze/thaw steps followed by treatment with Benzonase nuclease (1 U ml^{-1}) (Sigma-Aldrich) at 37 °C for 30 min. Viral lysates were then incubated at 56 °C for 30 min to inactivate adenovirus serotype 5. The viral genomic titer was determined as described above. To analyze *cap* sequences, AAV viral genomes were extracted after packaging and rounds 3 and 6 of selection, amplified by PCR and sequenced at the UC Berkeley DNA Sequencing Facility.

Statistical analysis of variable positions in evolved ancestral libraries

A comparison of the two sets of amino acids at each variable amino acid position was conducted to identify variable positions whose library proportions had changed significantly during selection. The posterior probability that the two sets of variable amino acids come from two different probability distributions was calculated assuming probability parameters that are Dirichlet distributed with low pseudocounts, to reflect sparse observed counts. For comparison of the synthesized and theoretical library, post-synthesis amino acid frequencies distributed via a Dirichlet multinomial were compared with the theoretical probabilities from the library distributed by a multinomial.

In vitro transduction analysis

After six rounds of selection, ancestral library viral genomes were cloned into the pXX2 rAAV packaging plasmid. To benchmark the infectivity of rAAV ancestral libraries against a panel of natural AAV serotypes, vectors were packaged with a self-complementary CMV-GFP cassette using the transient transfection method previously described.[56,60] Cell lines (293T, C2C12, IB3-1, B16-F10, CHO-K1, pgsA, Pro5, Lec1 and Lec2) were seeded in 96-well plates at a density of 15 000 cells per well. One day after seeding, cells were infected with rAAV at a genomic MOI of 2000 (293T, C2C12, IB3-1, B16-F10 and GBM), 10000 (Pro5, Lec1 and Lec2), 32 000 (C2C12) or 50 000 (CHO-K1, pgsA) ($n = 3$). To analyze antibody evasion properties, ancestral rAAV libraries were incubated at 37 °C for 1 h with serial dilutions of heat-inactivated intravenous immunoglobulin (Gammagard, Baxter Healthcare, Hayward, CA, USA) and then used to infect HEK293T cells at a genomic MOI of 2000 ($n = 3$).

To characterize thermostability, virions packaged with self-complementary CMV-GFP were diluted with Dulbecco's modified Eagle's medium supplemented with 2% fetal bovine serum and incubated at temperatures ranging from 59.6 °C to 78 °C for 10 min in a thermocycler (Bio-Rad) before being cooled down to 37 °C and used to infect 293T cells at genomic MOIs ranging from 1500 to 16 000; MOIs were adjusted to ensure an adequate number of GFP-positive cells for analysis. For all studies, the fraction of GFP-expressing cells 72 h post infection was quantified with a Guava EasyCyte 6HT flow cytometer (UC Berkeley Stem Cell Center, Berkeley, CA, USA).

In vivo animal imaging and quantification of luciferase expression

High-titer rAAV CAG-Luciferase vectors were purified by iodixanol gradient and then concentrated and exchanged into phosphate-buffered saline using Amicon Ultra-15 centrifugal filter units (Millipore). To study skeletal muscle transduction, 5×10^{10} rAAV-Luc DNase-resistant genomic particles were injected in a volume of 30 µl into each gastrocnemius muscle of 7-week-old female BALB/c mice (Jackson Laboratories, $n = 3$) as previously described.[56] Six weeks after injection, animals were killed and the gastrocnemius muscle was collected and frozen. Luciferase activity was determined and normalized to total protein as previously described.[60] All animal procedures were approved by the Office of Laboratory Animal Care at the University of California and conducted in accordance with NIH guidelines on laboratory animal care.

CONFLICT OF INTEREST

DVS is an inventor on patents involving AAV directed evolution.

ACKNOWLEDGEMENTS

We are grateful to Professor Brent Reynolds (University of Florida) for kindly providing the L0 human GBM tumor-initiating cells. This work was supported by the National Institutes of Health grant (R01EY022975). DSO is supported by a National Science Foundation Graduate Fellowship and JSO is supported by a National Science Foundation Graduate Fellowship and a UC Berkeley Graduate Division Fellowship. IH and OW were supported by the National Human Genome Research Institute grant (HG004483).

AUTHOR CONTRIBUTIONS

DSO and JSO designed the project and carried out the experimental work, analyzed and interpreted the data, and wrote and edited the manuscript. OW and JRW designed the project and edited the manuscript. SYW carried out GBM cell culture and edited the manuscript. AS assisted with molecular cloning. SK edited the manuscript. IH designed the project, analyzed and interpreted the data, supervised the project and edited the manuscript. DVS designed the project, interpreted the data, supervised the project through all stages and edited the manuscript.

REFERENCES

1 Stackhouse J, Presnell SR, McGeehan GM, Nambiar KP, Benner SA. The ribonuclease from an extinct bovid ruminant. *FEBS Lett* 1990; **262**: 104–106.
2 Cole MF, Gaucher EA. Utilizing natural diversity to evolve protein function: applications towards thermostability. *Curr Opin Chem Biol* 2011; **15**: 399–406.
3 Gaucher EA, Govindarajan S, Ganesh OK. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* 2008; **451**: 704–707.
4 Ortlund EA, Bridgham JT, Redinbo MR, Thornton JW. Crystal structure of an ancient protein: evolution by conformational epistasis. *Science* 2007; **317**: 1544–1548.
5 Ugalde JA, Chang BS, Matz MV. Evolution of coral pigments recreated. *Science* 2004; **305**: 1433.
6 Afriat-Jurnou L, Jackson CJ, Tawfik DS. Reconstructing a missing link in the evolution of a recently diverged phosphotriesterase by active-site loop remodeling. *Biochemistry* 2012; **51**: 6047–6055.
7 Watanabe K, Ohkuri T, Yokobori S, Yamagishi A. Designing thermostable proteins: ancestral mutants of 3-isopropylmalate dehydrogenase designed by using a phylogenetic tree. *J Mol Biol* 2006; **355**: 664–674.
8 Alcolombri U, Elias M, Tawfik DS. Directed evolution of sulfotransferases and paraoxonases by ancestral libraries. *J Mol Biol* 2011; **411**: 837–853.
9 Chen F, Gaucher EA, Leal NA, Hutter D, Havemann SA, Govindarajan S *et al.* Reconstructed evolutionary adaptive paths give polymerases accepting reversible terminators for sequencing and SNP detection. *Proc Natl Acad Sci USA* 2010; **107**: 1948–1953.
10 Gonzalez D, Hiblot J, Darbinian N, Miller JC, Gotthard G, Amini S *et al.* Ancestral mutations as a tool for solubilizing proteins: the case of a hydrophobic phosphate-binding protein. *FEBS Open Bio* 2014; **4**: 121–127.
11 Kothe DL, Li Y, Decker JM, Bibollet-Ruche F, Zammit KP, Salazar MG *et al.* Ancestral and consensus envelope immunogens for HIV-1 subtype C. *Virology* 2006; **352**: 438–449.
12 Ducatez MF, Bahl J, Griffin Y, Stigger-Rosser E, Franks J, Barman S *et al.* Feasibility of reconstructed ancestral H5N1 influenza viruses for cross-clade protective vaccine development. *Proc Natl Acad Sci USA* 2011; **108**: 349–354.
13 Rolland M, Jensen MA, Nickle DC, Yan J, Learn GH, Heath L *et al.* Reconstruction and function of ancestral center-of-tree human immunodeficiency virus type 1 proteins. *J Virol* 2007; **81**: 8507–8514.
14 Gullberg M, Tolf C, Jonsson N, Mulders MN, Savolainen-Kopra C, Hovi T *et al.* Characterization of a putative ancestor of coxsackievirus B5. *J Virol* 2010; **84**: 9695–9708.
15 Berns KI, Linden RM. The cryptic life style of adeno-associated virus. *BioEssays* 1995; **17**: 237–245.
16 Ellis BL, Hirsch ML, Barker JC, Connelly JP, Steininger RJ 3rd, Porteus MH. A survey of ex vivo/in vitro transduction efficiency of mammalian primary cells and cell lines with Nine natural adeno-associated virus (AAV1-9) and one engineered adeno-associated virus serotype. *Virol J* 2013; **10**: 74.
17 Asokan A, Schaffer DV, Samulski RJ. The AAV vector toolkit: poised at the clinical crossroads. *Mol Ther* 2012; **20**: 699–708.
18 Kotterman MA, Schaffer DV. Engineering adeno-associated viruses for clinical gene therapy. *Nat Rev Genet* 2014; **15**: 445–451.
19 Gaudet D, Methot J, Dery S, Brisson D, Essiembre C, Tremblay G *et al.* Efficacy and long-term safety of alipogene tiparvovec (AAV1-LPLS447X) gene therapy for lipoprotein lipase deficiency: an open-label trial. *Gene Therapy* 2013; **20**: 361–369.
20 Sonntag F, Schmidt K, Kleinschmidt JA. A viral assembly factor promotes AAV2 capsid formation in the nucleolus. *Proc Natl Acad Sci USA* 2010; **107**: 10220–10225.
21 Khersonsky O, Rosenblat M, Toker L, Yacobson S, Hugenmatter A, Silman I *et al.* Directed evolution of serum paraoxonase PON3 by family shuffling and ancestor/consensus mutagenesis, and its biochemical characterization. *Biochemistry* 2009; **48**: 6644–6654.
22 Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucleic Acids Res* 2014; **42**: D32–D37.
23 Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* 2001; **17**: 754–755.
24 Wu Z, Miller E, Agbandje-McKenna M, Samulski RJ. Alpha2,3 and alpha2,6 N-linked sialic acids facilitate efficient binding and transduction by adeno-associated virus types 1 and 6. *J Virol* 2006; **80**: 9093–9103.
25 Gao GP, Alvira MR, Wang L, Calcedo R, Johnston J, Wilson JM. Novel adeno-associated viruses from rhesus monkeys as vectors for human gene therapy. *Proc Natl Acad Sci USA* 2002; **99**: 11854–11859.
26 Westesson O, Barquist L, Holmes I. HandAlign: Bayesian multiple sequence alignment, phylogeny and ancestral reconstruction. *Bioinformatics* 2012; **28**: 1170–1171.
27 Koerber JT, Maheshri N, Kaspar BK, Schaffer DV. Construction of diverse adeno-associated viral libraries for directed evolution of enhanced gene delivery vehicles. *Nat Protoc* 2006; **1**: 701–706.
28 Bloom JD, Romero PA, Lu Z, Arnold FH. Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. *Biol Direct* 2007; **2**: 17.
29 Aharoni A, Gaidukov L, Khersonsky O, Mc QGS, Roodveldt C, Tawfik DS. The 'evolvability' of promiscuous protein functions. *Nat Genet* 2005; **37**: 73–76.
30 Thornton JW. Resurrecting ancient genes: experimental analysis of extinct molecules. *Nat Rev Genet* 2004; **5**: 366–375.
31 Rayaprolu V, Kruse S, Kant R, Venkatakrishnan B, Movahed N, Brooke D *et al.* Comparative analysis of adeno-associated virus capsid stability and dynamics. *J Virol* 2013; **87**: 13150–13160.
32 Klimczak RR, Koerber JT, Dalkara D, Flannery JG, Schaffer DV. A novel adeno-associated viral variant for efficient and selective intravitreal transduction of rat Muller cells. *PLoS One* 2009; **4**: e7467.
33 Hauck B, Xiao W. Characterization of tissue tropism determinants of adeno-associated virus type 1. *J Virol* 2003; **77**: 2768–2774.
34 Vandenberghe LH, Breous E, Nam HJ, Gao G, Xiao R, Sandhu A *et al.* Naturally occurring singleton residues in AAV capsid impact vector performance and illustrate structural constraints. *Gene Therapy* 2009; **16**: 1416–1428.
35 Wu Z, Asokan A, Grieger JC, Govindasamy L, Agbandje-McKenna M, Samulski RJ. Single amino acid changes can influence titer, heparin binding, and tissue tropism in different adeno-associated virus serotypes. *J Virol* 2006; **80**: 11393–11397.
36 Thornton JW, Need E, Crews D. Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. *Science* 2003; **301**: 1714–1717.
37 Voordeckers K, Brown CA, Vanneste K, van der Zande E, Voet A, Maere S *et al.* Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication. *PLoS Biol* 2012; **10**: e1001446.
38 Lanfear R, Kokko H, Eyre-Walker A. Population size and the rate of evolution. *Trends Ecol Evol* 2014; **29**: 33–41.
39 Bowles DE, McPhee SW, Li C, Gray SJ, Samulski JJ, Camp AS *et al.* Phase 1 gene therapy for Duchenne muscular dystrophy using a translational optimized AAV vector. *Mol Ther* 2012; **20**: 443–455.
40 Lochrie MA, Tatsuno GP, Christie B, McDonnell JW, Zhou S, Surosky R *et al.* Mutations on the external surfaces of adeno-associated virus type 2 capsids that affect transduction and neutralization. *J Virol* 2006; **80**: 821–834.
41 Schmidt DM, Mundorff EC, Dojka M, Bermudez E, Ness JE, Govindarajan S *et al.* Evolutionary potential of (beta/alpha)8-barrels: functional promiscuity produced by single substitutions in the enolase superfamily. *Biochemistry* 2003; **42**: 8387–8393.
42 Asuri P, Bartel MA, Vazin T, Jang JH, Wong TB, Schaffer DV. Directed evolution of adeno-associated virus for enhanced gene delivery and gene targeting in human pluripotent stem cells. *Mol Ther* 2012; **20**: 329–338.
43 Varadarajan N, Gam J, Olsen MJ, Georgiou G, Iverson BL. Engineering of protease variants exhibiting high catalytic activity and exquisite substrate selectivity. *Proc Natl Acad Sci USA* 2005; **102**: 6855–6860.
44 Zincarelli C, Soltys S, Rengo G, Rabinowitz JE. Analysis of AAV serotypes 1-9 mediated gene expression and tropism in mice after systemic injection. *Mol Ther* 2008; **16**: 1073–1080.
45 Wu Z, Asokan A, Samulski RJ. Adeno-associated virus serotypes: vector toolkit for human gene therapy. *Mol Ther* 2006; **14**: 316–327.
46 Schenk-Braat EA, van Mierlo MM, Wagemaker G, Bangma CH, Kaptein LC. An inventory of shedding data from clinical gene therapy trials. *J Gene Med* 2007; **9**: 910–921.

47 Toromanoff A, Cherel Y, Guilbaud M, Penaud-Budloo M, Snyder RO, Haskins ME *et al.* Safety and efficacy of regional intravenous (r.i.) versus intramuscular (i.m.) delivery of rAAV1 and rAAV8 to nonhuman primate skeletal muscle. *Mol Ther* 2008; **16**: 1291–1299.
48 Ciron C, Cressant A, Roux F, Raoul S, Cherel Y, Hantraye P *et al.* Human alpha-iduronidase gene transfer mediated by adeno-associated virus types 1, 2, and 5 in the brain of nonhuman primates: vector diffusion and biodistribution. *Hum Gene Ther* 2009; **20**: 350–360.
49 Gray SJ, Foti SB, Schwartz JW, Bachaboina L, Taylor-Blake B, Coleman J *et al.* Optimizing promoters for recombinant adeno-associated virus-mediated gene expression in the peripheral and central nervous system using self-complementary vectors. *Hum Gene Ther* 2011; **22**: 1143–1153.
50 Qiao C, Yuan Z, Li J, He B, Zheng H, Mayer C *et al.* Liver-specific microRNA-122 target sequences incorporated in AAV vectors efficiently inhibits transgene expression in the liver. *Gene Therapy* 2011; **18**: 403–410.
51 Geisler A, Jungmann A, Kurreck J, Poller W, Katus HA, Vetter R *et al.* microRNA122-regulated transgene expression increases specificity of cardiac gene transfer upon intravenous delivery of AAV9 vectors. *Gene Therapy* 2011; **18**: 199–209.
52 Bershtein S, Goldin K, Tawfik DS. Intense neutral drifts yield robust and evolvable consensus proteins. *J Mol Biol* 2008; **379**: 1029–1044.
53 Tseng YS, Agbandje-McKenna M. Mapping the AAV capsid host antibody response toward the development of second generation gene delivery vectors. *Front Immunol* 2014; **5**: 9.
54 Gurda BL, DiMattia MA, Miller EB, Bennett A, McKenna R, Weichert WS *et al.* Capsid antibodies to different adeno-associated virus serotypes bind common regions. *J Virol* 2013; **87**: 9111–9124.
55 Wobus CE, Hugle-Dorr B, Girod A, Petersen G, Hallek M, Kleinschmidt JA. Monoclonal antibodies against the adeno-associated virus type 2 (AAV-2) capsid: epitope mapping and identification of capsid domains involved in AAV-2-cell interaction and neutralization of AAV-2 infection. *J Virol* 2000; **74**: 9281–9293.
56 Maheshri N, Koerber JT, Kaspar BK, Schaffer DV. Directed evolution of adeno-associated virus yields enhanced gene delivery vectors. *Nat Biotechnol* 2006; **24**: 198–204.
57 Gao G, Vandenberghe LH, Alvira MR, Lu Y, Calcedo R, Zhou X *et al.* Clades of adeno-associated viruses are widely disseminated in human tissues. *J Virol* 2004; **78**: 6381–6388.
58 Takeuchi Y, Myers R, Danos O. Recombination and population mosaic of a multifunctional viral gene, adeno-associated virus cap. *PLoS One* 2008; **3**: e1634.
59 Westesson O, Holmes I. Developing and applying heterogeneous phylogenetic models with XRate. *PLoS One* 2012; **7**: e36898.
60 Koerber JT, Jang JH, Schaffer DV. DNA shuffling of adeno-associated virus yields functionally diverse viral progeny. *Mol Ther* 2008; **16**: 1703–1709.
61 Deleyrolle LP, Harding A, Cato K, Siebzehnrubl FA, Rahman M, Azari H *et al.* Evidence for label-retaining tumour-initiating cells in human glioblastoma. *Brain* 2011; **134**: 1331–1343.
62 Huson DH, Scornavacca C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst Biol* 2012; **61**: 1061–1067.
63 Schrodinger, LLC. The PyMOL Molecular Graphics System, Version 1.3r1 2010.

Supplementary Information accompanies this paper on Gene Therapy website (http://www.nature.com/gt)