Comprehensive mutagenesis maps the effect of all single codon mutations in the AAV2 *rep* gene on AAV production

Nina K. Jain[1,2], Pierce J. Ogden*[1,2,3], George M. Church*[1,2]
*Corresponding author

Affiliations
1. Wyss Institute for Biologically Inspired Engineering, Boston, MA
2. Dept. of Genetics, Harvard Medical School, Boston, MA
3. Present address: Manifold Biotechnologies, Boston, MA

Email addresses:
Nina: ninajain@g.harvard.edu
Pierce: pierce.ogden@gmail.com
George: gchurch@genetics.med.harvard.edu

Abstract

Recombinant adeno-associated viruses (rAAVs) are the predominant gene therapy vector. Several rAAV vectored therapies have achieved regulatory approval, but production of sufficient rAAV quantities remains difficult. The AAV Rep proteins, which are essential for genome replication and packaging, represent a promising engineering target for improvement of rAAV production but remain underexplored. To gain a comprehensive understanding of the Rep proteins and their mutational landscape, we assayed the effects of all 39,297 possible single codon mutations to the AAV2 *rep* gene on AAV2 production. Most beneficial variants are not observed in nature, indicating that improved production may require synthetic mutations. Additionally, the effects of AAV2 *rep* mutations were largely consistent across capsid serotypes, suggesting that production benefits are capsid independent. Our results provide a detailed sequence-to-function map that enhances our understanding of Rep protein function and lays the groundwork for Rep engineering and enhancement of large-scale gene therapy production.

Introduction

rAAVs are a popular tool for the delivery of gene therapies as wild-type AAV is not pathogenic. Wild-type AAV consists of a 4.7 kb single-stranded DNA genome, which is packaged into an approximately 26 nm icosahedral capsid (Srivastava et al., 1983; Xie et al., 2002). The AAV genome is flanked on either end by 145 nucleotide inverted terminal repeats, which form hairpins, serve as the origins of viral replication, and are the only sequences required *in cis* for packaging of DNA into the capsid (Xiao et al., 1997). As such, the remainder of the AAV genome can be replaced with a gene of interest to generate rAAV vectors (Samulski & Muzyczka, 2014). The wild-type AAV genome consists of two genes, *rep* and *cap* (Srivastava et al., 1983). During rAAV production, these genes are supplied *in trans* to the inverted terminal repeat plasmid. The *cap* gene encodes three structural proteins, VP1, VP2, and VP3, which assemble in an approximately 1:1:10 ratio to form the 60-mer capsid (Cassinotti et al., 1988; Wörner et al., 2021). Engineering of the *cap* gene has enabled targeting of AAV vectors to specific tissues and cell populations. The *rep* gene encodes four proteins, Rep78, Rep68, Rep52, and Rep40, which are generated through the use of two promoters, p5 and p19, and alternative splicing (Davis et al., 2000).

The larger Rep proteins, Rep78 and Rep68, are required for genome replication while the smaller Rep proteins, Rep52 and Rep40, facilitate genome packaging. Notably, Rep78 and Rep68 alone are each sufficient for rAAV vector production (Hölscher et al., 1995). However, the presence of Rep52/40 enhances genome packaging and therefore rAAV titer (Chejanovsky & Carter, 1989; King et al., 2001). The *rep* gene encodes three protein domains, an origin-binding domain, a helicase domain, and a zinc-finger domain (Figure 1A) (Di Pasquale & Stacey, 1998; Im & Muzyczka, 1990; Smith & Kotin, 1998). All four Rep proteins contain the helicase domain as well as a nuclear localization signal (Cassell & Weitzman, 2004). Only Rep78 and Rep68 contain the origin-binding domain and only Rep78 and Rep52 contain the zinc-finger domain. Additionally, residues in the linker domain between the origin-binding and helicase domains cooperate with residues at the N-terminus of the helicase domain to facilitate oligomerization of the larger Rep proteins, which is required for AAV production (Zarate-Perez et al., 2012).

The origin-binding domain contains three separate DNA binding motifs that are important for genome replication (Hickman et al., 2002, 2004). Firstly, the origin-binding domain recognizes the Rep binding site, which consists of GCTC repeats present in the double stranded region of the inverted terminal repeats (Musayev, Zarate-Perez, Bardelli, et al., 2015). Unwinding of this double stranded DNA by the helicase domain enables the origin-binding domain to interact with single stranded DNA at its second motif, the active site pocket. Here, the origin-binding domain nicks the single stranded DNA at the terminal resolution site, an essential step in viral genome replication (Im & Muzyczka, 1990; Snyder et al., 1990). Finally, the origin-binding domain contains a single stranded DNA hairpin binding site, which interacts with one of the inverted terminal repeat hairpins. Hairpin binding, however, is not required for terminal resolution site nicking or genome replication (Wu et al., 1999). The origin-binding domain can also recognize GCTC repeats present in the p5 promoter; in the presence of a helper virus, such as adenovirus, the Rep proteins activate transcription from the endogenous promoters and in the

absence of helper virus the Rep proteins repress transcription (Labow et al., 1986; Murphy et al., 2007).

The helicase domain plays a definitive role in all steps of AAV production while the zinc-finger domain is likely dispensable for production. The helicase domain unwinds the double stranded inverted terminal repeats during genome replication and its activity has been shown to facilitate genome packaging (Brister & Muzyczka, 1999; King et al., 2001). There is also evidence that residues within the helicase domain mediate capsid interactions. Mutations to the helicase domain of Rep52/40, as well as mutations near the five-fold axis of symmetry in the capsid, reduced the interaction between the Rep proteins and capsid and resulted in lower rAAV titers (Bleker et al., 2006; King et al., 2001). The third Rep domain, the zinc-finger domain, has not been shown to bind DNA but is involved in binding to various host cell proteins (Di Pasquale & Stacey, 1998). Previous work indicated that premature stop codons introduced into the zinc-finger domain-encoding region of *rep* do not affect rAAV titers, suggesting that this domain is not required for rAAV production (Mietzsch et al., 2021).

As more rAAV vectored therapies are approved, manufacturing sufficient rAAV quantities to meet patient need is an increasingly important issue. Engineering the Rep proteins is a promising avenue for improving AAV production. The Rep proteins are involved in all steps of viral production, including regulation of VP expression, genome replication, and genome packaging. However, the Rep proteins are not a structural component of the final vector. As such, the Rep proteins can be designed to optimize viral production without affecting downstream processes, such as cell targeting, which are driven by the capsid. The sequence of the *rep* genes in naturally occurring serotypes is relatively well conserved (average of 78% amino acid sequence identity between AAV2 *rep* and AAV1-13 *rep*). While previous mutational studies of the Rep proteins have been conducted, these studies have focused primarily on identifying non-functional Rep variants to identify the location of key motifs (Davis et al., 2000; Gavin et al., 1999; Hörer et al., 1995; Yang et al., 1992). As such, Rep engineering efforts may benefit from the exploration of additional sequence diversity.

Towards this end, we generated a library of all possible single codon mutations of the AAV2 *rep* gene, the *rep* gene most commonly used for AAV production. We assayed the effect of these mutations on AAV2 production and generated a detailed sequence-to-function map of the AAV2 *rep* gene.

Results

Comprehensive mutagenesis assays the effect of all single codon mutations in the AAV2 *rep* gene on AAV2 production

To better understand how mutations to *rep* affect AAV production, we generated two *rep* libraries. The first library, referred to as pCMV-Rep78/68, consisted of a cytomegalovirus (CMV) promoter followed by the *rep* open reading frame with a M225G mutation introduced to prevent expression of the smaller Rep proteins. This library allowed us to assay the effect of mutating only Rep78 and Rep68. The second library, referred to as WT AAV2, was analogous to the sequence of the wild-type virus and contained the p5 promoter and *rep* and *cap* open reading

frames. Use of the endogenous promoters in the WT AAV2 library allowed us to capture the effect of *rep* mutations on Rep and VP expression. This library also enabled simultaneous mutation of all four Rep proteins. For each library, we generated all 39,297 possible single codon substitution and deletion variants spanning from the Rep78/68 start codon to the Rep78/52 stop codon (Figure 1A). All variants contained unique 20 bp barcodes at the 3' end of the *rep* gene (pCMV-Rep78/68 library) or *cap* gene (WT AAV2 library), and each codon variant was represented by a minimum of two barcode sequences. To assess the diversity of the plasmid libraries, we amplified and sequenced the barcodes from each of the plasmid pools (Figure 1B, Figure S1A, and Table S1). We observed 100 and 99.9% of the expected amino acid variants in the pCMV-Rep78/68 and WT AAV2 libraries, respectively.

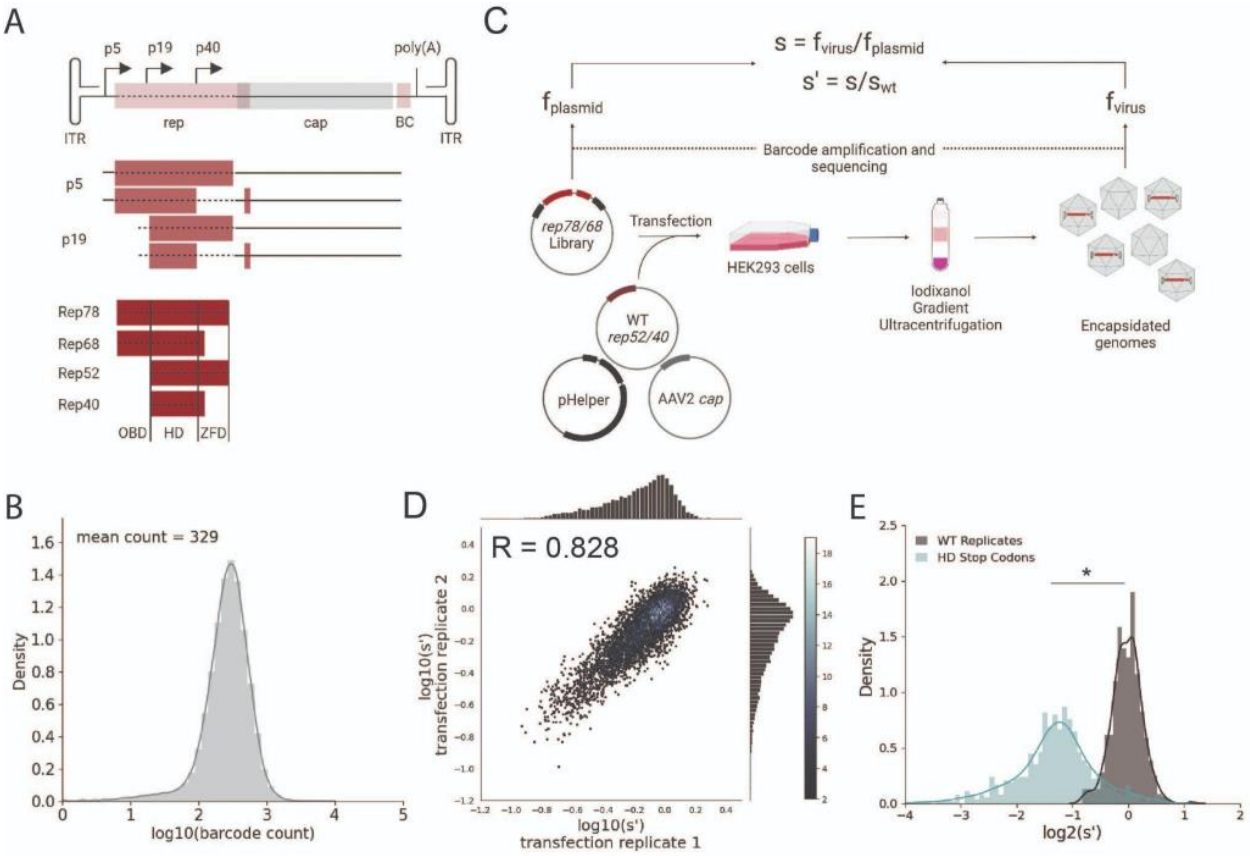

Figure 1. Comprehensive mutagenesis library design and production assay**.** (A) Organization of the AAV2 genome and Rep protein domains. Top: single-stranded DNA genome, middle: RNA transcripts, bottom: Rep proteins. Dotted lines indicate mutated regions. (B) Density plot of barcode counts in the pCMV-Rep78/68 plasmid library. (C) Overview of production assay for the pCMV-Rep78/68 library and calculation of wild-type normalized production fitness values (s'). (D) Amino acid level production fitness values from replicate transfections of the pCMV-Rep78/68 library. Pearson R correlation coefficient calculated after log transformation. (E) Density plot of production fitness values for wild-type (black) and premature stop codon (blue) controls for the pCMV-Rep78/68 library. *$p < 10^{-20}$ (Mann Whitney U test)

Next, we transfected the plasmid libraries into HEK293T cells to assay the effect of all single codon mutations on production of genome-containing viral particles (Figure 1C). Production fitness values for each variant were calculated and normalized to internal wild-type controls as shown in Figure 1C. For each library, we performed transfections in duplicate and compared the production fitness values calculated from each replicate (Figure 1D and Figure S1B). For both libraries, production fitness values were well-correlated across replicates, indicating that the genotype-phenotype linkage was maintained during viral production. After examining the correlation between biological replicates, we next looked at the distribution of fitness values for the wild-type and premature stop codon controls (Figure 1E and Figure S1C). As expected, premature stop codons had a deleterious effect on AAV2 production and fitness values for replicate wild-type controls clustered together.

Annotation of the AAV2 *rep* sequence-to-function map

To better understand the sequence-to-function map of the AAV2 Rep proteins, we visualized the data from our production assay in multiple ways. First, we generated heatmaps containing the wild-type normalized production fitness values for all single amino acid substitutions and deletions (Figure 2 and Figure S2). Additionally, we calculated the “mutability” of each residue by averaging the normalized production fitness values at each position across all substitutions. We mapped the mutability of each residue onto structures of the origin-binding domain in complex with the Rep binding site (PDB: 4ZQ9, Figures 3A-B), the origin-binding domain in complex with single-stranded DNA from the inverted terminal repeat hairpin (PDB: 6XB8, Figure 3C), the origin-binding domain alone (PDB: 5D6X, Figure S5A), and the helicase domain (PDB: 1S9H, Figure S5B) (James et al., 2003; Musayev, Zarate-Perez, Bardelli, et al., 2015; Musayev, Zarate-Perez, Bishop, et al., 2015; Santosh et al., 2020). Finally, we used the production fitness values calculated for individual barcodes to determine which amino acid changes resulted in production fitness values significantly different from wild-type (Figure S4A-B). We observed that the origin-binding and zinc-finger domains were more tolerant of mutation than the helicase domain. Interestingly, a stark boundary between amino acids D212 and A213 was observed; at residue A213 the larger Rep proteins become much less tolerant of mutation. This boundary is adjacent to linker domain residues P214-Y224, which were previously reported to be important for Rep78/68 oligomerization (Musayev, Zarate-Perez, Bardelli, et al., 2015; Zarate-Perez et al., 2012).

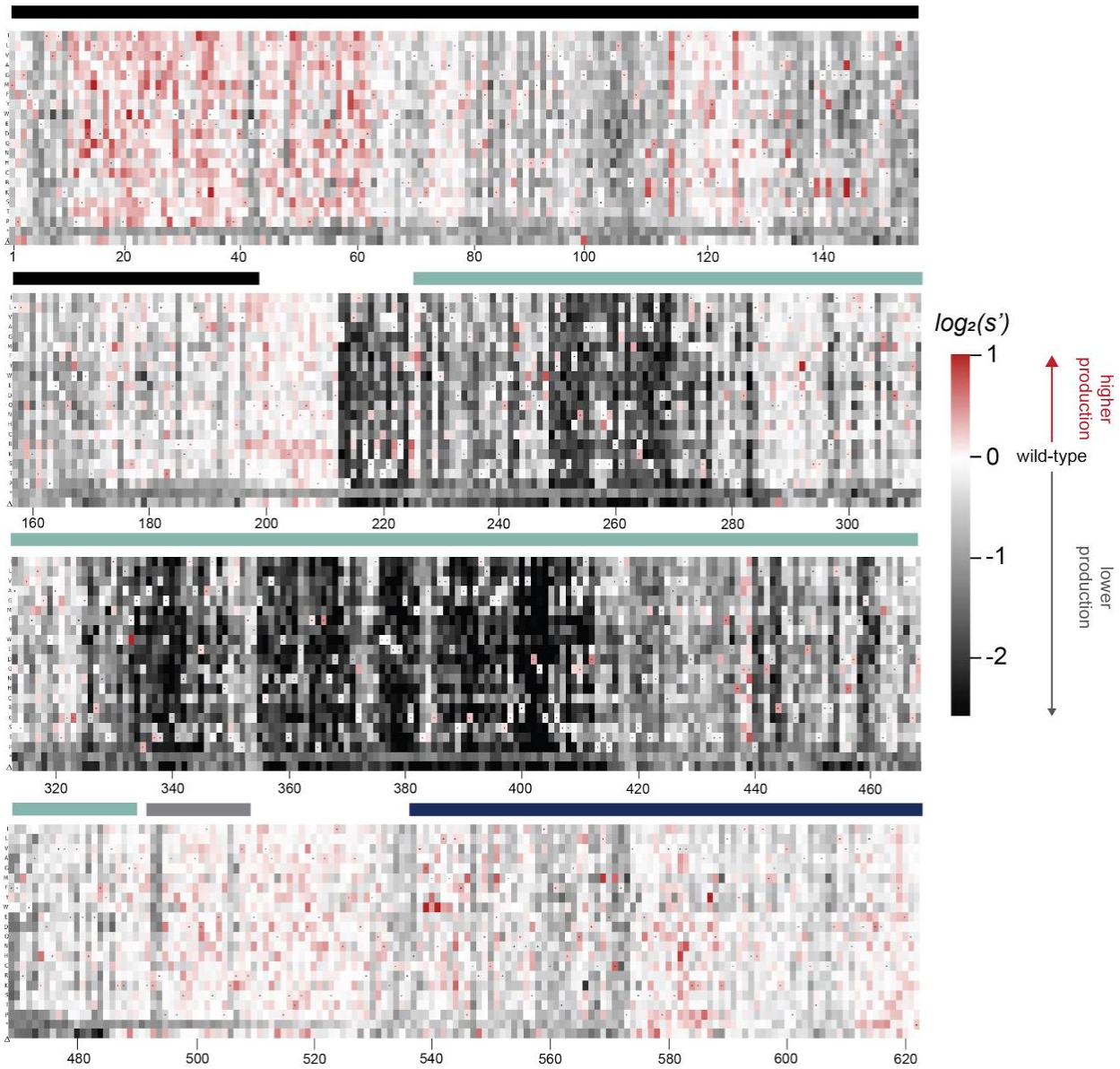

Figure 2. Effects of all single amino acid substitutions and deletions in the Rep78/68 proteins on AAV2 production. Amino acid level production fitness values from the pCMV-Rep78/68 production assay were calculated as in Figure 1C by summing barcode counts for synonymous mutations. Rectangles are colored by mutational effect on the production of genome-containing particles, with black indicating deleterious mutations and red indicating beneficial mutations. Colored bars above the heatmaps indicate protein domains. Black: origin-binding domain, light blue: helicase domain, gray: nuclear localization signal, and navy blue: zinc-finger domain. Black dots indicate wild-type amino acid identity.

When looking across all library members, the production fitness values determined for the pCMV-Rep78/68 and WT AAV2 libraries were well-correlated (Figure S3A). However, as expected, introduction of a methionine upstream of the Rep52/40 start codon was deleterious in the WT AAV2 format but not in the pCMV-Rep78/68 format, where Rep52/40 were expressed *in trans*.

The majority of beneficial substitutions clustered in the origin-binding domain. In particular, substitutions between residues V11 and D62 of the origin-binding domain enhanced AAV2 production relative to wild-type. These residues are involved in recognition of the inverted terminal repeat hairpin (Figure 3C) (Musayev, Zarate-Perez, Bardelli, et al., 2015). Origin-binding domain-hairpin interactions have been shown to enhance terminal resolution site nicking but are not required for nicking to occur (Wu et al., 1999). The majority of Rep binding site-interacting residues, on the other hand, were intolerant of mutation (Figure 3B). The origin-binding domain-Rep binding site interaction is important in mediating Rep-genome interactions during both genome replication and promoter regulation (Labow et al., 1986; Murphy et al., 2007; Musayev, Zarate-Perez, Bishop, et al., 2015). Interestingly, there are a handful of Rep binding site-interacting residues that were somewhat mutable, including S110 and N139, which form contacts with the phosphate backbone, and A141, which forms contacts with bases in the Rep binding site. Mutation of S110, N139, and A141, as well as G144 and V147, to positively charged residues had a beneficial effect on AAV2 production. Residues N139, A141, G144, and V147 are part of the loop that connects β-strand 4 to α-helix E (Musayev, Zarate-Perez, Bishop, et al., 2015). The sequence of this loop is highly variable across serotypes (Musayev, Zarate-Perez, Bardelli, et al., 2015). Notably, the asparagine at position 139 is positively charged in AAV5 (K139). However, no serotypes contain positively charged amino acids at positions 141, 144, or 147. Our data indicate that substitution of DNA-interacting residues in the origin-binding domain can enhance AAV production and identifies beneficial substitutions not observed in nature.

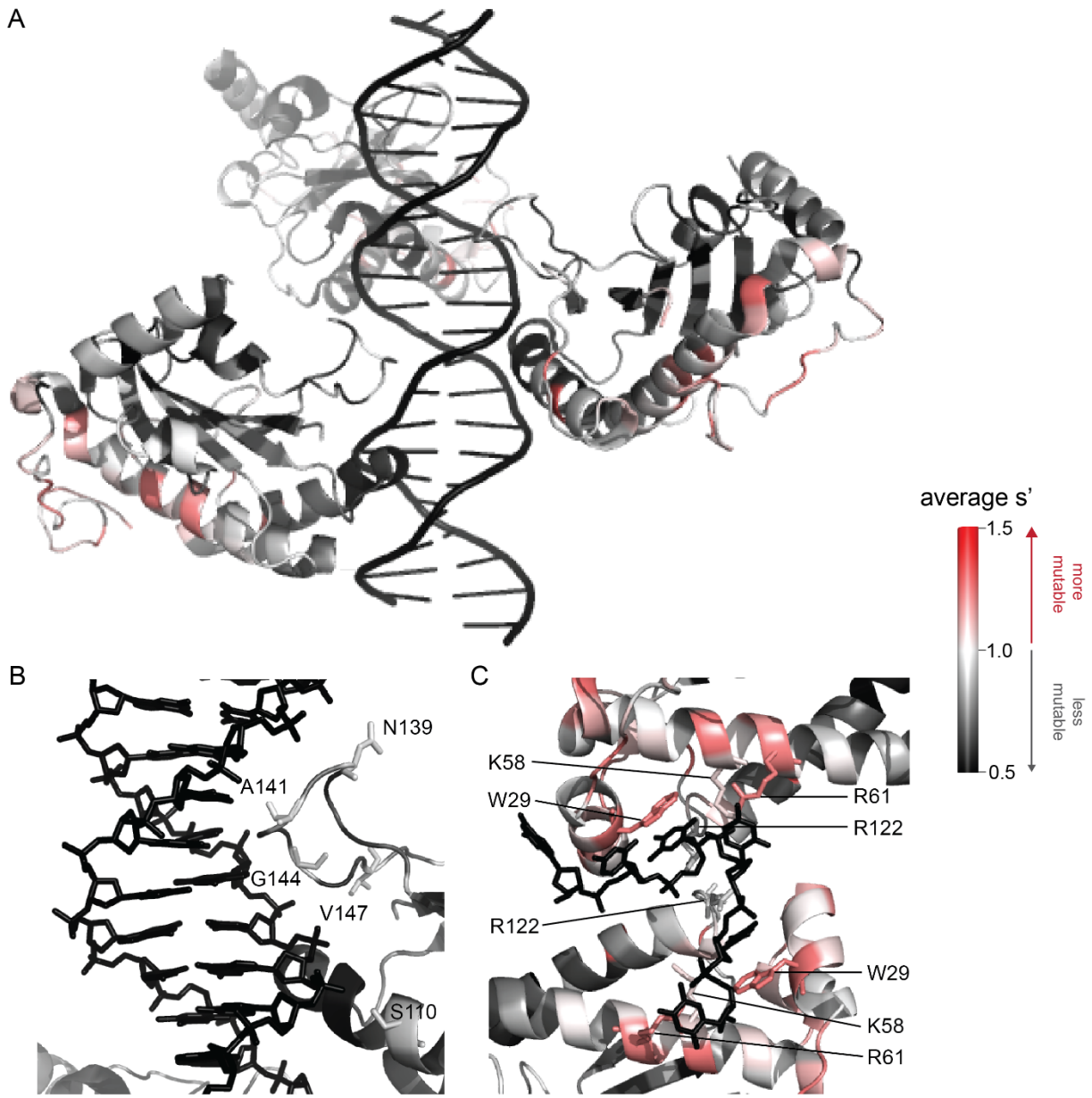

Figure 3. Beneficial substitutions cluster in DNA-interacting residues. (A) Average production fitness values for all substitutions at each position mapped onto the structure of the origin-binding domain in complex with the AAVS1 Rep binding site (PDB 4ZQ9). (B) Close up view of origin-binding domain-Rep binding site interactions. Residues where substitutions to positively charged residues are beneficial are shown as sticks. (C) Average production fitness values mapped onto the structure of the origin-binding domain in complex with the single stranded inverted terminal repeat hairpin (PDB 6XB8). DNA interacting residues are shown as sticks. Residues are colored by mutability, with red indicating higher mutability and black indicating lower mutability.

A clear pattern in the location of mutable residues in the origin-binding domain active site can also be observed (Figure S5A). This active site is responsible for cleaving single stranded DNA

and is formed by the origin-binding domain beta sheet and Y156, the active site nucleophile (Hickman et al., 2004). Residues with side chains directed towards the active site are less mutable than adjacent residues with their side chains directed away from the active site. It is more difficult to discern positional trends in mutability within the helicase domain as it is less tolerant of mutation than the origin-binding domain (Figure S5B). No structures of the ZFD are available. However, the majority of variants in this domain do not have production fitness values that are significantly different from wild-type (Figures S4A-B).

Generation of all single codon variants enables interrogation of nucleotide-level effects

As our libraries contain all possible single codon mutations, we were able to search for variants with nucleotide-level effects on production by comparing differences in production fitness values between synonymous codon variants (Figures S6-7). In general, there was good agreement in the fitness values for synonymous variants. The production fitness values for synonymous codon variants were more tightly distributed than the fitness values for nonsynonymous variants (Figures S8A-B). Comparison of fitness values between synonymous codons that do and do not introduce premature stop codons into alternate reading frames allowed us to search for possible frameshifted open reading frames. However, no evidence of frameshifted open reading frames was observed (Figures S9A-B).

We did observe an interesting pattern at amino acid Y283 (*rep* nucleotides 847-849). Variants with a c.849C>G mutation had lower production fitness values than synonymous variants without a c.849C>G ($p < 10^{-8}$, Mann-Whitney U-test). The deleterious effect of this mutation can also be observed by plotting the average fitness value at each nucleotide position for each of the four bases (Figures S9C-D). The negative effect of a c.849C>G mutation was consistent across the pCMV-Rep78/68 and WT AAV2 library formats and was also observed in experiments with alternate *cap* genes. In each library, there were fifteen codon variants with a c.849C>G mutation, thirteen of which had synonymous codons without a c.849C>G mutation for comparison. Our results indicate that the nucleotide sequence of the *rep* gene is optimized. However, we were not able to detect broad nucleotide-level effects in our data.

Comprehensive mutagenesis identifies beneficial variants not observed in nature

We observed that mutations to amino acids found in other naturally occurring serotypes were better tolerated than mutations to amino acids that are not found in nature (Figure 4A). However, a majority of the variants with production fitness values greater than that of wild-type are not observed in nature (Figure 4B). Only 245/2351 (6.17%) of variants with s' > 1 and only 4/115 (3.48%) of variants with s' > 1.5 are observed in AAV serotypes 1-13. These data emphasize the power of our comprehensive approach to identify novel functional sequence diversity.

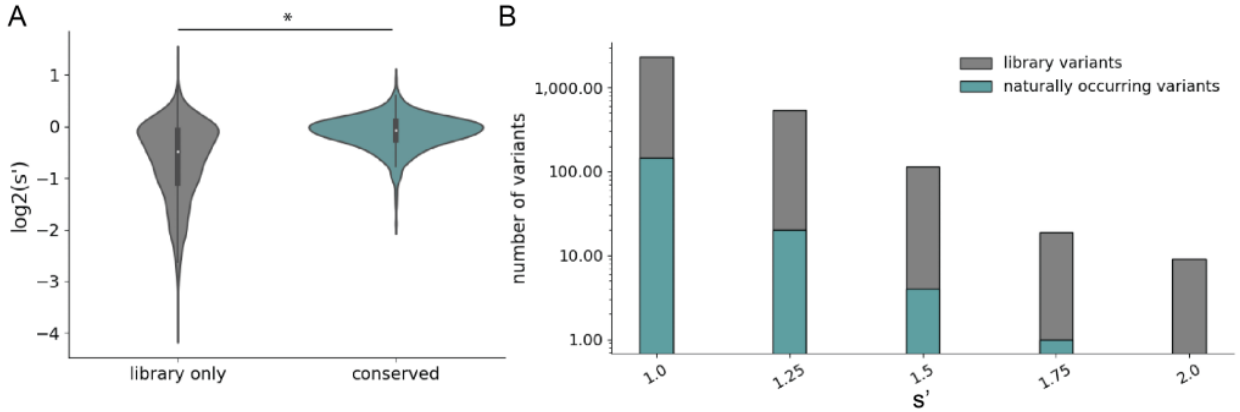

Figure 4. Comparison of comprehensive mutagenesis measurements to variation in nature. (A) Distribution of wild-type normalized production fitness values for conserved variants (blue) and variants found only in the library (gray). (B) Total number of variants and number of conserved variants with s' greater than wild-type (s' > 1). *$p < 10^{-20}$ (Mann Whitney U test)

Validation of multiplexed production assay results

To validate the results of our multiplexed production assay we selected fourteen variants, cloned them individually into the pCMV-Rep78/68 format, and determined their effect on production by measuring DNase-resistant particle titers (Figure 5A). We included H92A and K340H, mutations to the origin-binding domain and helicase domain active sites, respectively, as controls (Musayev, Zarate-Perez, Bardelli, et al., 2015; Smith & Kotin, 1998). Twelve variants with mutations in the origin-binding domain, linker domain, and helicase domain and fitness values greater than wild-type were also assayed. Finally, we included two deleterious variants identified in our multiplexed assay, N139G and A213V. The DNase-resistant particle titers determined from individual transfections are well-correlated with the fitness values determined from the multiplexed production assay (Figure 5B). We performed similar validation for a panel of WT AAV2 format variants (Figure S3B).

To enable multiplexed analysis of *rep* variants in our library assay, the mutant *rep* genes themselves are packaged into the AAV capsids. Relatively small amounts of *rep* library plasmids are also used in transfection to reduce the frequency at which multiple *rep* variant plasmids enter the same cell. These procedures maintain a genotype-phenotype linkage and allow us to identify the *rep* sequences that enable production of genome-containing particles. We sought to confirm that the effects observed in the library production assay would be conserved when the *rep* variants were supplied *in trans* to the inverted terminal repeat genome, as in the traditional triple plasmid transfection method used for rAAV production. To this end, we selected a small subset of *rep* variants and cloned them into an AAV2 pRepCap plasmid without inverted terminal repeats. Here, we assayed variants with single codon mutations in each of the Rep domains.

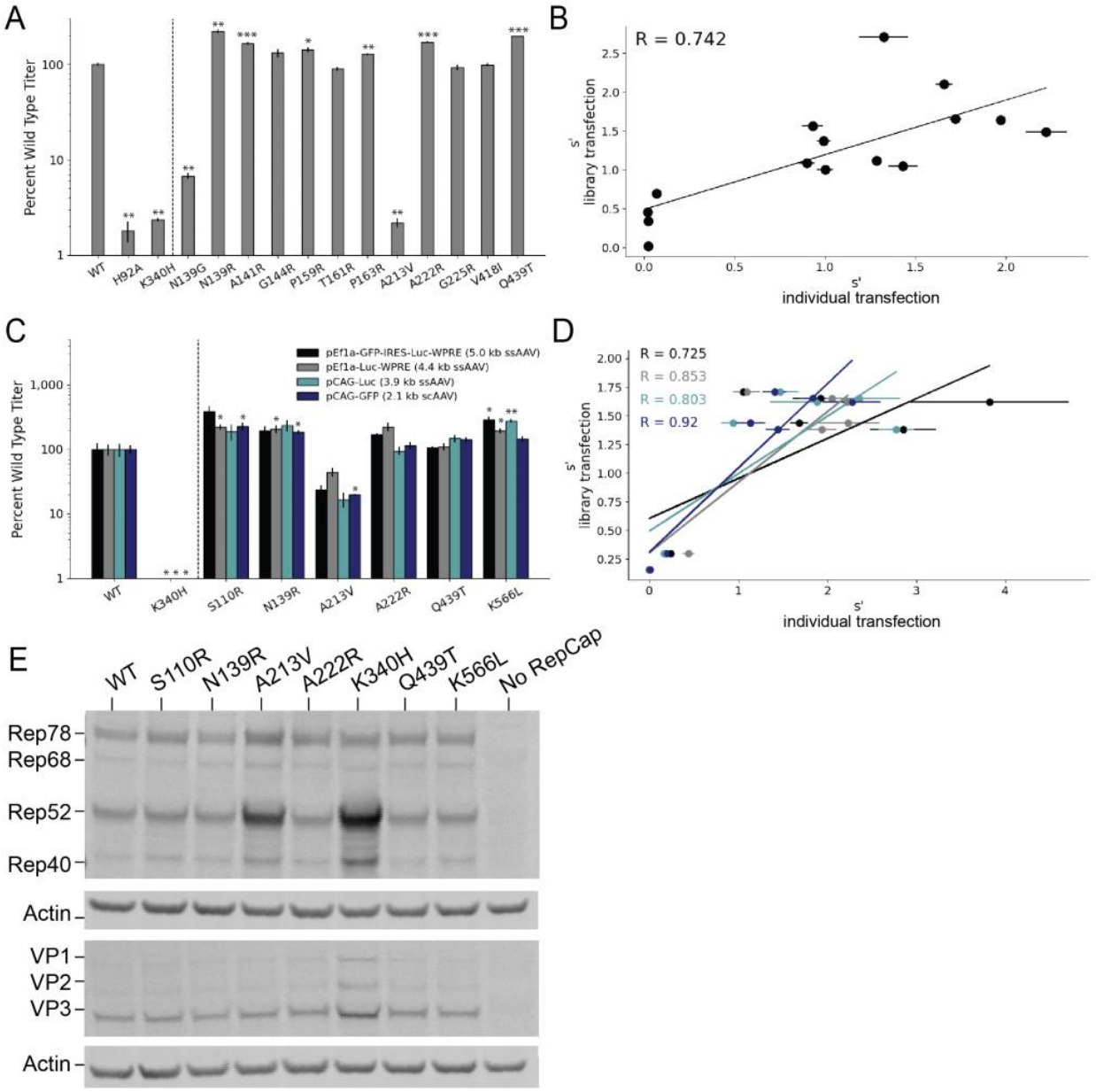

Figure 5. Validation of AAV2 library production assay results. (A) DNase-resistant particle titers for fourteen single amino acid pCMV-Rep78/68-inverted terminal repeat variants produced individually. Titers for previously characterized variants are plotted to the left of the dotted line. (B) Relationship between normalized production fitness values from library experiments and DNase-resistant particle titers from individual transfections. (C) DNase-resistant particle titers for four rAAV genomes produced with the indicated Rep variants. Titers for previously characterized variants are plotted to the left of the dotted line. (D) Relationship between rAAV DNase-resistant particle titers and normalized production fitness values from library experiments. (E) Expression of Rep and VP proteins from variant pRepCap plasmids by Western blot. For panels A and C, asterisks indicate significance of titer differences between the Rep variant and the relevant wild-type control. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$ (Welch's t-test)

We used these mutant pRepCap plasmids to produce rAAV vectors containing four different genomes, three single-stranded genomes and one self-complementary genome. The DNase-resistant particle titers for these pRepCap variants were relatively consistent across the different genomes (Figure 5C). Importantly, the DNase-resistant particle titers determined when the *rep* variants were supplied *in trans* correlated well with the normalized fitness values determined in our production assay (Figure 5D). Several single codon variants, including S110R, N139R, and K566L, showed 50-100% improvements in DNase-resistant particle titer over wild-type. Western blot analysis indicated that most of these variants had little to no effect on Rep protein or VP expression (Figure 5E). Interestingly, in the case of the deleterious variants, A213V and K340H, Rep52/40 protein levels were increased. Codon GCG213 is located just upstream of the Rep52/40 start codon. As such, mutations to this position likely affect the strength of the p19 promoter and therefore Rep52/40 expression levels. Notably, the effects of this mutation on the p19 promoter are separate from the deleterious effects of the A213V amino acid change; the negative effects of the A213V change are observed even when Rep52/40 are expressed *in trans* (Figure 5A).

The S110R variant increases transducing particle titer

We determined the relative transduction efficiency of rAAV2 produced with wild-type Rep and two different Rep variants, S110R and Q439T. We set up individual transfections with these Rep variants and affinity purified the resulting rAAV2 vectors. Use of the S110R Rep variant resulted in approximately 2-fold higher viral genome titers (Figure S10A). We also performed capsid ELISAs and determined that the S110R variant resulted in a similar increase in total particle titers; while viral genome titers were increased relative to wild-type Rep, the ratio of viral genomes: capsids was unchanged. We then used a constant volume of each rAAV2 prep to transduce HEK293T cells and quantified the relative amount of transducing particles by measuring transgene expression (luciferase activity, Figure S10B). Again, an approximately 2-fold increase in relative transduction was observed with rAAV2 produced with S110R Rep as compared to wild-type Rep.

Mutations in the AAV2 *rep* gene have similar effects on production of AAV2, AAV5, and AAV9 capsids

Many different AAV capsid serotypes are of clinical interest for their unique tissue targeting properties. Given the physical interaction between Rep proteins and the capsid, we hypothesized that some *rep* mutations would have unique effects on the production of different capsid serotypes. To identify *rep* variants with serotype-specific effects on production, we repeated our pCMV-Rep78/68 library production assay using AAV5 and AAV9 *cap* genes in place of AAV2 (Figures S11-14). The AAV5 capsid was chosen as it has low overall sequence identity with the AAV2 capsid (57.7% amino acid identity). The AAV9 capsid is more similar to the AAV2 capsid (81.7% amino acid identity) but is of particular clinical interest and differs from AAV2 at amino acid positions 329 and 330, which were previously identified as important for Rep-capsid interactions (AAV2: T329/T330, AAV9: V331/K332) (Bleker et al., 2006). Surprisingly, the correlation in fitness values across capsid serotypes was comparable to the correlation between biological replicates, indicating that the effect of AAV2 *rep* variants on production is largely consistent across capsid serotypes (Figures 6A-B). We individually

produced rAAV5 and rAAV9 vectors using a subset of our *rep* variants and confirmed that the titers with each *rep* variant are well correlated across capsid serotype (Figure S15).

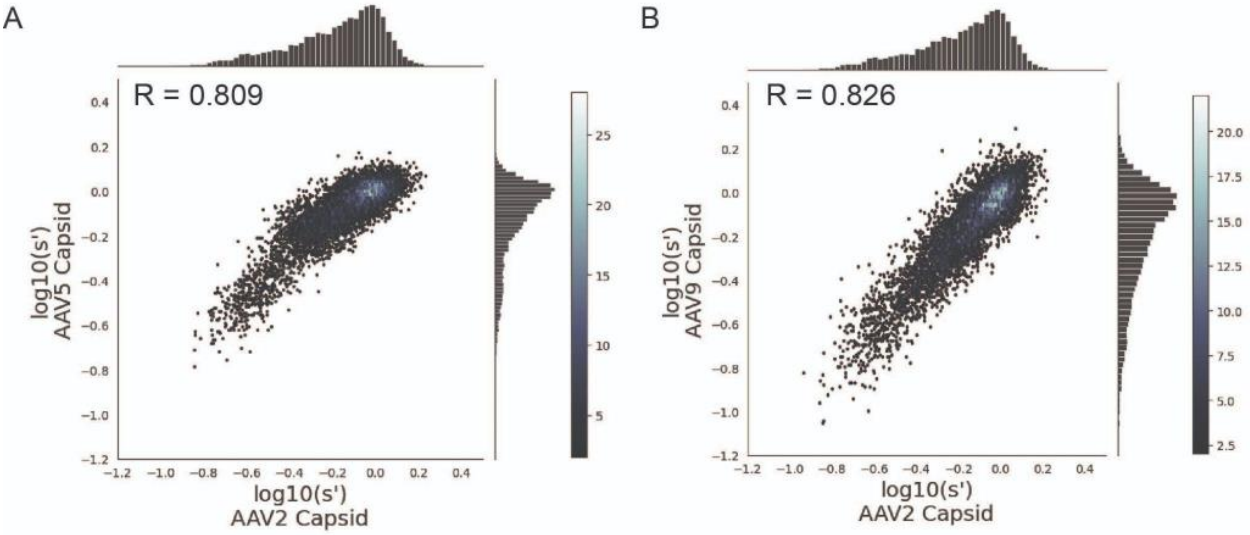

Figure 6. Mutations to AAV2 *rep* have similar effects on AAV2, AAV5, and AAV9 capsid production. Wild-type normalized production fitness values from the library production assay with (A) AAV5 and AAV2 capsids and (B) AAV9 and AAV2 capsids. Pearson R correlation coefficient calculated after log transformation.

Discussion

We have generated a comprehensive mutagenesis library of the AAV2 *rep* gene, containing all possible single codon substitutions and deletion variants. Multiplexed assay of this library allowed us to generate a sequence-to-function map, linking all variants to their effect on AAV production.

Many of the previous AAV mutagenesis studies have focused on the AAV *cap* gene. Previous work in our lab generated a library containing all possible single codon substitutions, deletions, and insertions of the AAV2 *cap* gene and assayed their effect on AAV2 production (Ogden et al., 2019). Notably, we observed a smaller range of fitness values in our production assay with the *rep* library than with the previously reported *cap* library. This observation aligns with our knowledge of Rep and capsid biology. The Rep proteins' primary function is to facilitate AAV production, while the capsid has evolved to not only enable assembly and genome packaging, but also to facilitate cell targeting, entry, and nuclear trafficking. Additionally, the Rep proteins possess multiple enzymatic activities that require the proteins to adopt specific and dynamic conformations. The capsid, in contrast, is not known to possess enzymatic activity outside of the phospholipase domain located at the N-terminus of VP1 (Stahnke et al., 2011). It follows that there are greater mutational constraints on *rep* than on *cap* in the context of AAV2 production.

Despite the smaller range of fitness values observed here, we identified numerous Rep variants with production fitness values greater than or equal to that of wild-type AAV2 Rep that are not observed in naturally occurring AAV serotypes. We attribute this discrepancy to the fact that optimal rAAV production may not equate to optimal fitness of AAV in the endogenous host.

Wild-type AAV has both a lytic and a latent cycle (Liu, 2014). The latent cycle occurs when no helper virus is present. Rep and VP expression are repressed and the AAV genome is integrated into the host cell genome in a process that also involves Rep (Pereira et al., 1997; Surosky et al., 1997). In the presence of a helper virus, such as adenovirus, the lytic cycle proceeds. In this instance, Rep78/68 activate transcription from the AAV promoters and genome replication and packaging occur (Pereira et al., 1997). rAAV production is somewhat analogous to the lytic cycle. However, naturally-occuring AAV must balance the effect of any mutations on fitness in both the lytic and latent contexts. An additional explanation for this finding is the relatively small number of AAV serotypes, for which rep sequences are available. In our analysis, we compared the beneficial variants identified in our library to the *rep* sequences from twelve different AAV serotypes (AAV1-13, no sequence for AAV9 *rep*). This is a small number of naturally occurring *rep* sequences when compared to the size of our library. While there have been efforts to identify additional naturally-occuring *cap* sequences, there has been relatively little effort to expand the number of naturally-occurring *rep* sequences (Wang et al., 2019). We expect that such work may identify *rep* sequences containing many of the beneficial variants identified in our library.

We observed that substitutions in the origin-binding and zinc-finger domains were better tolerated than substitutions in the linker and helicase domains. A sharp drop in mutability was observed between residues D212 and A213. Previous work has identified residues V215 through Y224 as the minimal linker sequence required to facilitate Rep78/68 oligomerization (Zarate-Perez et al., 2012). In a separate study, it was reported that mutation of P214 reduced the ability of Rep to oligomerize (Musayev, Zarate-Perez, Bardelli, et al., 2015). Our data indicate that residue A213 is intolerant of mutation. A213 likely plays an important role in AAV production and, given its position, may be involved in Rep78/68 oligomerization. Additional work is needed to confirm that mutations to A213 affect AAV production by interfering with Rep78/68 oligomerization. In both library formats, the zinc-finger domain was relatively tolerant of mutation and few substitutions in this domain resulted in production fitness values that were significantly different from wild-type. At some positions, even introduction of a premature stop codon was not deleterious. It was recently reported that premature stop codons introduced at positions 522 and 553 of Rep did not affect AAV2 production when supplied in a pRepCap plasmid (Mietzsch et al., 2021). Our results provide further evidence that the zinc-finger domain is dispensable for AAV production and identify A213 as a potential linker domain residue.

The majority of beneficial substitutions clustered in the origin-binding domain. In a recent investigation of *rep* hybrids, Mietzsch et al. reported that replacement of the entire AAV2 origin-binding domain with that of AAV1 or AAV8 improved the proportion of full capsids (Mietzsch et al., 2021). Our data supports the importance of the origin-binding domain for AAV production and identifies specific origin-binding domain regions where beneficial mutations cluster. Most substitutions of inverted terminal repeat-hairpin interacting residues and substitutions of residues in the Rep binding site-interacting loop to positively charged amino acids were beneficial. Previous studies have demonstrated that Rep binding to the inverted terminal repeat-hairpin improves the efficiency of terminal resolution site nicking but is not required for the nicking reaction to occur (Wu et al., 1999). The Rep binding site-interacting loop, on the other

hand, is important for recognition of double-stranded DNA during genome replication and promoter binding. Interestingly, further characterization of the S110R variant, which falls in the Rep binding site-interacting region, indicated that it increases viral genome and transducing particle titer relative to wild-type Rep but did not increase the ratio of viral genome: capsids, a proxy for the proportion of full capsids. Additional work is needed to understand which, if any, of the beneficial OBD variants identified here also improve the proportion of full capsids. Taken together, our results indicate that enhancement of Rep-DNA interactions may be a fruitful avenue for further improvement of AAV production.

Inclusion of all single codon variants in our libraries allowed us to investigate the effect of synonymous mutations on AAV production. Interestingly, we observed that introduction of a c.849C>G mutation resulted in lower production fitness values compared to synonymous variants. Interestingly, this position is located downstream of the p5 and p19 promoters. Additionally, this mutation had a negative effect on production even when the *cap* gene was supplied *in trans,* making the activity of the p40 promoter irrelevant for AAV production. The c.849C>G mutation also does not fall near any known AAV2 splice sites (Stutika et al., 2016). While the mechanism by which this mutation affects production remains unclear, our results emphasize that the *rep* gene is optimized for AAV production at both the amino acid and nucleotide levels.

While we were able to discern a trend in the effect of synonymous variants at nucleotide position 849 of *rep*, the selection values for synonymous codon variants in general were relatively closely distributed. This precluded further investigation of the effect of synonymous codons on AAV production. Synonymous variants likely have an effect on aspects of AAV production, such as genome replication, transcriptional regulation, mRNA stability, and protein expression. However, our assay measures the aggregate effect of *rep* variants on all steps in the AAV production process and is likely unable to detect the effects of synonymous variants on specific steps in this process if those steps are not rate-limiting.

In an attempt to identify Rep substitutions with a beneficial effect on the production of specific capsid serotypes, we repeated the pCMV-Rep78/68 library assay using AAV5 and AAV9 *cap* genes in place of AAV2 *cap*. Surprisingly, the correlation in fitness values across experiments was similar to the correlation between biological replicates, suggesting that the majority of AAV2 *rep* mutations have a similar effect on the production of AAV2, AAV5, and AAV9 capsids. These results suggest that either: *rep* mutations have a similar effect on Rep78/68-capsid interaction across serotypes, that the effect of perturbing Rep78/68-capsid interactions is obscured by the deleterious effects of the same mutations on other Rep activities, or that Rep78/68-capsid interactions are not a limiting factor in AAV production.

We have generated a comprehensive sequence-to-function map of the effect of all single codon AAV2 *rep* mutations on AAV2, AAV5, and AAV9 production. Our experiments identified thousands of functional Rep variants, laying the groundwork for further engineering of these proteins and enhancement of large-scale gene therapy production.

Materials and Methods

Library cloning

We generated *rep* variant libraries through pooled oligonucleotide (oligo) synthesis (Twist Biosciences) and subsequent Golden Gate Assembly using methods previously developed in our lab (Ogden et al., 2019). To begin, we generated two wild-type backbone plasmids, containing either the Rep78/68 open reading frame or the *rep* and *cap* open reading frames, and removed all BsaI, BbsI, EcoRV, SphI, and XbaI sites. Within the *rep* open reading frame we introduced a synonymous mutation at G339 (c.1017G>C). Within the VP open reading frame we introduced a synonymous mutation at V118 (c.354C>G) and a coding mutation F370Y (c.1109T>A). The *rep* gene was divided into eleven tiles and cloning for each tile was carried out separately. We designed 300-mer oligos to include 207 nucleotides of *rep*-coding sequence immediately followed by a BbsI site, an EcoRV site, another BbsI site, and a unique 20 nucleotide barcode sequence. All of these elements were flanked by BsaI and primer binding sites on either end of the synthesized sequence. Unique primer binding sites were used for each tile. To enable cloning of both the Rep78/68 and WT AAV2 format libraries, oligos containing the Rep52/40 start codon (M225) were synthesized with and without the M225G mutation. All possible single codon substitutions and deletions, including synonymous variants, were included in the library. All positions from the Rep78/68 start codon to the Rep78/52 stop codon (622 codons) were mutated. We did not include the eight codons at the end of the Rep68/40 open reading frame in our mutagenesis as these positions overlap with the VP1 open reading frame. Each codon variant was represented by at least two unique barcodes. For each tile, a minimum of ten uniquely barcoded wild-type controls were included. The Rep78/68 and WT AAV2 libraries each had a total of 81,116 uniquely barcoded variants.

Following synthesis, we amplified the oligos for each tile (Q5 Hot Start High-Fidelity 2X Master Mix, NEB). In parallel, we amplified the backbone vector using primers that introduced BsaI sites. Vector PCR products were digested with BsaI-HF v2, DpnI, and recombinant shrimp alkaline phosphatase (rSAP, NEB) overnight and PCR purified (Qiagen QIAQuick PCR Purification Kit) the following day. We then performed Golden Gate Assembly with the amplified oligos and vector PCR digest products (NEBridge Golden Gate Assembly Kit, BsaI-HF v2). Golden Gate Assembly reactions were cycled 100X (16°C for 5 minutes, 37°C for minutes) and then heat inactivated. Golden Gate Assembly products were PCR purified and eluted into 25 uL of Buffer EB (Qiagen). Eluates were drop-dialyzed against 30 mL of water for 1 hour and transformed (Lucigen E. cloni 10G ELITE electrocompetent cells). Transformed cells were recovered in 1 mL Lucigen recovery media at 37°C for 1 hour and the entire volume of outgrowth was used to inoculate 50 mL LB + Kanamycin cultures, which we grew at 30°C overnight. The following day, we midi-prepped these cultures *via* alkaline lysis (Qiagen Plasmid Plus Midi-Prep Kit). This first cloning step enabled generation of intermediate products, which contained any wild-type *rep* sequence upstream of the mutated oligo followed by the mutant oligo. In this step, all wild-type sequences downstream of the mutant oligo were removed.

To re-introduce these missing wild-type sequences, we used a second round of Golden Gate Assembly and the internal BbsI sites present in each oligo. Before performing Golden Gate Assembly, we ran rolling circle amplification on our intermediate cloning products. 10 ng of the

intermediate plasmid products were incubated with 10 uM random hexamer primers and 1X phi29 DNA polymerase buffer (NEB) at 95°C for 3 minutes and then cooled to room temperature. We then mixed these samples with rolling circle amplification solution (10 uM random hexamers, 1X phi29 DNA polymerase buffer, 5 U phi29 DNA polymerase, 1 mM dNTPs, 2 mg/mL recombinant albumin, 0.02 U inorganic pyrophosphatase) at a 1:1 ratio and incubated them at 30°C overnight. The resulting rolling circle amplification products were directly digested with BbsI-HF, DpnI, and recombinant shrimp alkaline phosphatase (all NEB) at 37°C overnight. The following day, we ran the digest products on 1% Tris-acetate EDTA gels and extracted and purified the correctly sized products (Qiagen QIAQuick Gel Extraction Kit). In parallel, we PCR amplified the missing downstream wild-type sequences for each tile from the backbone vectors; the primers used here also added BbsI sites. We ran Golden Gate Assembly with the rolling circle amplification digest products and vector PCR products, BbsI-HF (NEB), and T4 DNA Ligase (NEB), cycling as described above. Golden Gate Assembly products were transformed and midi-prepped as above. This second cloning step resulted in plasmids containing a full-length *rep* gene with a single codon mutation followed by a twenty nucleotide barcode. WT AAV2 library plasmids also contained a wild-type copy of the AAV2 *cap* gene between the variant *rep* genes and barcodes.

To enable packaging of variant *rep* sequences into AAV capsids, the step two cloning products were moved into inverted terminal repeat-containing vectors. Step two cloning products were subject to rolling circle amplification as above and digested with XbaI, SphI-HF, and EcoRV-HF. The *rep* and/or *cap* open reading frames in the backbone vectors were flanked by XbaI and SphI sites. EcoRV sites were part of the synthesized oligos and should only be present in step one cloning products. Digest products were gel extracted and ligated into inverted terminal repeat-containing vectors. In the case of the WT AAV2 format library, the inverted terminal repeat vector contained the p5 promoter and endogenous AAV2 polyA sequence. In the case of the Rep78/68 and Rep52/40 libraries, the inverted terminal repeat vector contained a CMV promoter, WPRE, and bGH polyA sequence. The integrity of the inverted terminal repeat destination plasmids was confirmed by SmaI digest and complete plasmid sequencing (MGH DNA Core).

Viral library production assay

HEK293T cells were cultured in DMEM supplemented with 10% FBS and seeded in five-layer cell culture multi-flasks (Corning 353144) at 8 x 10^7 cells/flask two days prior to transfection. We transfected HEK293T cells using polyethylenimine. For each library, replicate transfections were performed in separate cell stacks. For the pCMV-Rep78/68 library, transfections were performed with 1 ug of pCMV-Rep78/68 library plasmids, 50 ug of pHelper, 25 ug of pCMV-AAV2*cap*, and 1.5 ug of p19-Rep52/40 per cell stack. For the WT AAV2 library, transfections were performed with 1 ug of WT AAV2 library plasmids and 50 ug of pHelper. Additional plasmid ratios were tested for each library; conditions that resulted in the strongest correlation in production fitness values between replicate transfections were used; these are the conditions listed above. AAV5 and AAV9 capsid production assays were performed using the pCMV-Rep78/68 library as described above.

Three days post-transfection NaCl was added to a final concentration of 0.5 M and samples were incubated at 37°C for 3 hours to detach and lyse cells. Samples were transferred to fresh 500 mL bottles and incubated at 4°C overnight. The following day any precipitate was removed by pipetting and the remaining volume was passed through a sterile 0.22 um PES filter. PEG-8000 was added to a final concentration of 8% and samples were again incubated at 4°C overnight. The following day, we centrifuged samples at 3000 x g for 20 minutes to pellet the PEG-precipitated virus. Supernatants were discarded and pellets resuspended in 8 mL of DPBS. We then digested samples with a 1:10,000 dilution benzonase (Millipore Sigma 1.101695.0001) at 37°C for 45 minutes and subjected them to iodixanol gradient ultracentrifugation as previously described (Ogden et al., 2019; Zolotukhin et al., 1999). Briefly, benzonase-treated samples were underlaid with an iodixanol gradient (Millipore Sigma D1556) in polypropylene tubes (Beckman Coulter, 362183) and centrifuged at 242,000 x *g* for 1 hour at 16°C in a VTi50 rotor. Following ultracentrifugation, 500 uL fractions were collected from the 40% iodixanol layer and buffer-exchanged into DPBS using 100 kDa molecular weight cutoff centrifugal filter units (Millipore Sigma UFC910024).

Rep sequences in the purified pool represent variants capable of producing genome-containing viral particles. Barcodes from this viral pool, along with barcodes from the plasmid libraries, were amplified using flanking primer sequences. Illumina sequencing adapters and indices were added in a subsequent PCR. The resulting PCR amplicons were pooled and sequenced using the Illumina NextSeq platform (Biopolymers Facility, HMS). Barcode sequences were extracted from the resulting sequencing reads and barcode counts (c_v) from replicate transfections were summed. We calculated the frequency of each variant (f_v) in the viral library as $f_v = c/\Sigma c_v$ and then determined the production fitness for each variant as $s = f_{viral}/f_{plasmid}$. Finally, production fitness was normalized to that of the wild-type controls: $s = s/s_{WT}$.

Determination of DNase-resistant particle titers by qPCR

To validate the results of our multiplexed library production assay, fourteen single codon *rep* variants were selected and individually transfected into HEK293T cells. Transfections were performed in triplicate. We seeded cells at 4 x 10^5 cells/well in 6-well plates 24 hours prior to transfection. Small amounts of pCMV-Rep78/68 format variants, 50 pg per transfection, were used to recapitulate the low plasmid levels used in the library production assay. pCMV-Rep78/68 variants were transfected *via* PEI together with pHelper (2 ug), pCMV-AAV2*cap* (1 ug), and p19-Rep52/40 (2 ng) plasmids as in the library assay. Three days post-transfection, we lysed cells by 3X freeze-thaw. Samples were then clarified by centrifugation at 15,000 x *g* for 5 minutes. 5 uL of supernatant were subjected to DNase digest (ThermoFisher, EN0521) at 37°C for 30 minutes followed by heat inactivation at 65°C for 10 minutes. We then incubated samples at 98°C for 10 minutes to denature the capsids and measured DNase-resistant particle titers with qPCR (IDT PrimeTime Gene Expression Master Mix). Primers and a probe binding to the 5' region of the *rep* gene were used. The sequence of the forward primer was: 5'-GAGCATCTGCCCGGCATTTC-3', the sequence of the reverse primer was: 5'-ATCTGGCGGCAACTCCCATT-3', and the sequence of the probe was 5'-HEX-ACAGCTTTG-ZEN-TGAACTGGGTGGCCGA-3IABkFQ-3' (IDT).

Transfections with *rep* variants cloned into pRepCap plasmids were performed as described above, using the following amounts of plasmid for each transfection: 1 ug of pRepCap, 1 ug of inverted terminal repeat plasmid, and 2 ug of pHelper. The pEf1a-Luc-WPRE (ssAAV, Addgene plasmid #87951; http://n2t.net/addgene:87951; RRID:Addgene_87951), pCAG-Luc (ssAAV, Addgene plasmid #83281; http://n2t.net/addgene:83281; RRID:Addgene_83281), and pCAG-GFP (ssAAV, Addgene plasmid #83279; http://n2t.net/addgene:83279; RRID: Addgene_83279) Inverted terminal repeat plasmids were a gift from Mark Kay (Paulk et al., 2018; Pekrun et al., 2019). We used qPCR to determine plasmid concentration for transfection. The *rep* primer and probe sequences mentioned above were used to quantify pRepCap plasmids. To determine the concentration of GFP-expressing plasmids we used a forward primer with the sequence 5'-GAACCGCATCGAGCTGAA-3', a reverse primer with the sequence 5'-TGCTTGTCGGCCATGATATAG-3', and a probe with the sequence 5'-FAM-ATCGACTTC-ZEN-AAGGAGGACGGCAAC-3IABKFQ-3' (IDT). To determine the concentration of luciferase-expressing plasmids we used a forward primer with the sequence 5'-CATGTACCGCTTCGAGGAG-3', a reverse primer with the sequence 5'-GAAGCTAAATAGTGTGGGCACC3', and a probe with the sequence 5'-FAM-CTTGCGCAG-ZEN-CTTGCAAGACTATAAGATTC-3IABKFQ-3' (IDT). The same GFP- and luciferase-specific primers and probes were used to quantify the resulting rAAV vectors.

Western blot analysis of VP and Rep protein levels

Following transfection, media was aspirated from cell culture plates and cells were washed with DPBS. We then lysed cells with 50 uL of lysis buffer per well (6-well plates). The lysis buffer contained: 10 mM Tris-HCl (pH 8), 150 mM NaCl, 1% Triton X-100, and cOmplete mini protease inhibitor (1 tablet/10 mL lysis buffer). Lysates were transferred to fresh 1.5 mL tubes and centrifuged at 15,000 x *g* for 5 minutes to clarify. Total protein concentrations were determined by Bradford assay (Thermo Scientific Pierce Coomassie Bradford Protein Assay Kit). We loaded 75 ug of total protein from each sample on 4-12% Bis-Tris gels (ThermoFisher) and ran the gels for 45 minutes at 140V. We then transferred the proteins to PVDF membranes (ThermoFisher). Membranes were blocked with 5% milk in PBST for 1 hour at 4°C. Blots were then incubated with a 1:250 dilution of B1 anti-VP antibody or a 1:100 dilution of anti-Rep 303.9 antibody (both American Research Products) in blocking buffer overnight. Both primary antibody mixtures also included a 1:20,000 dilution of anti-β-actin antibody (Abcam, ab8227). The following day, blots were washed 3X with PBST and incubated with secondary antibodies for 1 hour in blocking buffer with 0.01% SDS. Goat anti-mouse IRDye 800CW (LiCor 925-32210) and donkey anti-rabbit IRDye 680RD (LiCor 925-68073) secondary antibodies were used. Blots were washed 3X with PBST and 1X with PBS, and the near-infrared fluorescence of the secondary antibodies was visualized using the Sapphire imager.

Acknowledgements

The authors would like to thank Siddharth Iyer, Anna Maurer, Kaia Mattioli, Aleksandra Prochera, Takeyuki Miyawaki, and Erik Aznauryan for their feedback on this manuscript. We would also like to acknowledge members of the Biopolymers Facility at Harvard Medical School, including Taylor Fennelly, Ashley Ciulla Hurst, and Baldwin Dilone, for their assistance with library sequencing. This work was funded by the Synthetic Biology Platform at the Wyss

Institute and the Wyss Institute faculty core allocation fund. Schematics were generated with BioRender.

Authors contributions

NKJ, PJO, and GMC conceived the study and designed experiments. NKJ performed experiments. NKJ performed data analysis with input from PJO. NKJ wrote the paper with input from all authors. GMC supervised the project and secured funding.

Competing interests

A full list of GMC's tech transfer, advisory roles, and funding sources can be found on the lab's website: http://arep.med.harvard.edu/gmc/tech.html. Harvard University has filed a patent application for inventions related to this work.

Code and data availability

Code and data for this paper can be accessed here: https://github.com/churchlab/aav_rep_scan.git. Illumina sequencing reads were uploaded to NCBI GEO (series accession number GSE226265).

Supplemental Materials

Materials and Methods

Analysis of stop codons in alternate reading frames

At each codon position, the average fitness value for all variants that introduced a premature stop codon into the +1 reading frame was determined. The average fitness value for all variants that did not introduce a stop codon into the +1 frame but were synonymous in the Rep frame to those that did was also calculated. Ten amino acid rolling averages for the +1 stops and +1 non-stops were calculated. These average fitness values were plotted and compared. The same procedure was used to determine the effect of premature stop codons in the +2 reading frame.

Comparison of effects on anonymous and non-synonymous codons on production fitness

For this analysis we used production fitness values (s') calculated at the codon level. For each fitness value, we calculated two mean centered fitness values. Positional mean centered fitness values were calculated as: $positional\ mean\ centered\ s' = s'_{codon} / average\ s'_{amino\ acid\ position}$. Synonymous codon mean centered fitness values were calculated as: $synonymous\ codon\ mean\ centered\ s' = s'_{codon} / average\ s'_{synonymous\ codon\ variants}$. Methionine and tryptophan codons were excluded as these amino acids are coded for by only a single codon. Mean centered fitness values for each codon variant were plotted versus amino acid position.

Affinity purification of AAV2 capsids

We purified rAAV2 vectors produced with variant pRepCap plasmids with AVB Sepharose (Cytiva) using a previously reported protocol, which we modified for use with gravity columns (Mietzsch et al., 2020). A 4.4 kb inverted terminal repeat genome (pAAV-EF1a-FLuc-WPRE-HGHpA) was used. pRepCap plasmids and the pAAV-EF1a-FLuc-WPRE-HGHpA inverted terminal repeat plasmid were quantified by qPCR as discussed above. HEK293T cells were harvested three days post-transfection and lysed by 3X freeze-thaw. We then subject samples to benzonase digest as described above and centrifuged samples at 6,000 x *g* for 30 minutes to pellet cell debris. Samples were then passed through 0.2 um PES filters, diluted 1:1 with TD Buffer (DPBS, 1 mM $MgCl_2$, 2.5 mL KCl), and incubated with 4 mL of AVB Sepharose slurry at 4°C overnight with shaking. The following day, we loaded the samples onto gravity columns, washed with 20 mL of 1X TD Buffer, and eluted samples with 5 mL 0.1 M glycine-HCl (pH 2.6). Eluates were immediately neutralized with 500 uL of 1 M Tris-HCl (pH 10). All elution fractions were pooled and buffer-exchanged as above and samples volumes were normalized to 400 uL. rAAV viral genome titers were determined *via* qPCR using the same luciferase primer and probe sequences discussed above. Total particle titers were determined *via* capsid ELISA (Progen PRATV) according to the manufacturer's protocol. Transfections, affinity purifications, and titer determinations were performed in duplicate.

Determination of relative transduction efficiency

Twenty-four hours prior to transduction, we seeded HEK293T cells at 2 x 10^4 cells/well in opaque white-bottom 96-well plates. Each rAAV prep was diluted 1:10,000 in serum-free media. 100 uL of this dilution was added to each well of a 96-well plate. We performed eight transductions with each rAAV prep. Forty-eight hours after transduction, we lysed the cells and measured luciferase activity (Promega One-Glo EX Luciferase Assay System, E8110).

Supplemental Figures

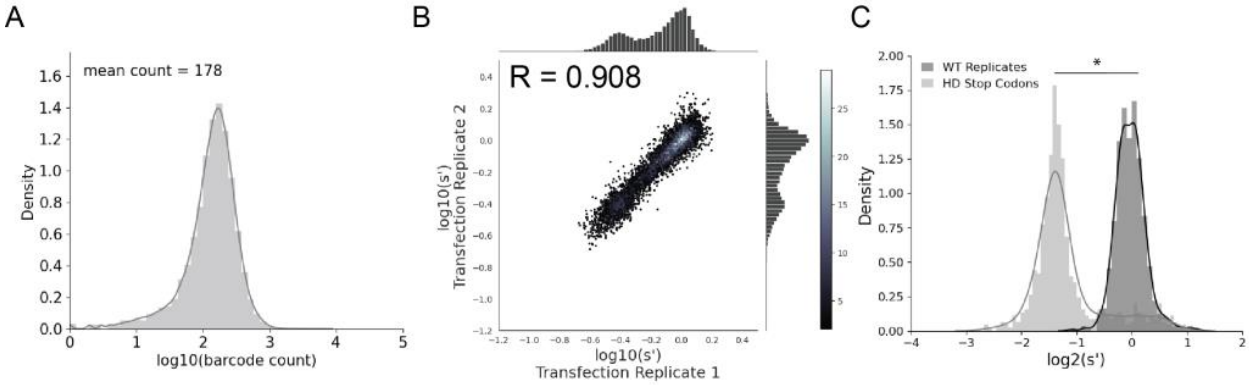

Figure S1. Comprehensive mutagenesis library design and production assay results for WT AAV2 format library. (A) Density plot of barcode counts in the WT AAV2 plasmid library. (B) Amino acid level production fitness values from replicate transfections for the WT AAV2 library. Pearson R correlation coefficient calculated after log transformation. (C) Density plot of wild-type (black) and premature stop codon (gray) controls for the WT AAV2 library. *$p < 10^{-20}$ (Mann Whitney U test)

Table S1. Percent of expected variants sequenced in plasmid and viral libraries.

	pCMV-Rep78/68		**WT AAV2**	
	Plasmid	**Viral**	**Plasmid**	**Viral**
Barcodes	99.8%	98.0%	98.3%	97.6%
Codon Variants	100%	99.9%	99.7%	99.7%
Amino Acid Variants	100%	100%	99.9%	99.9%

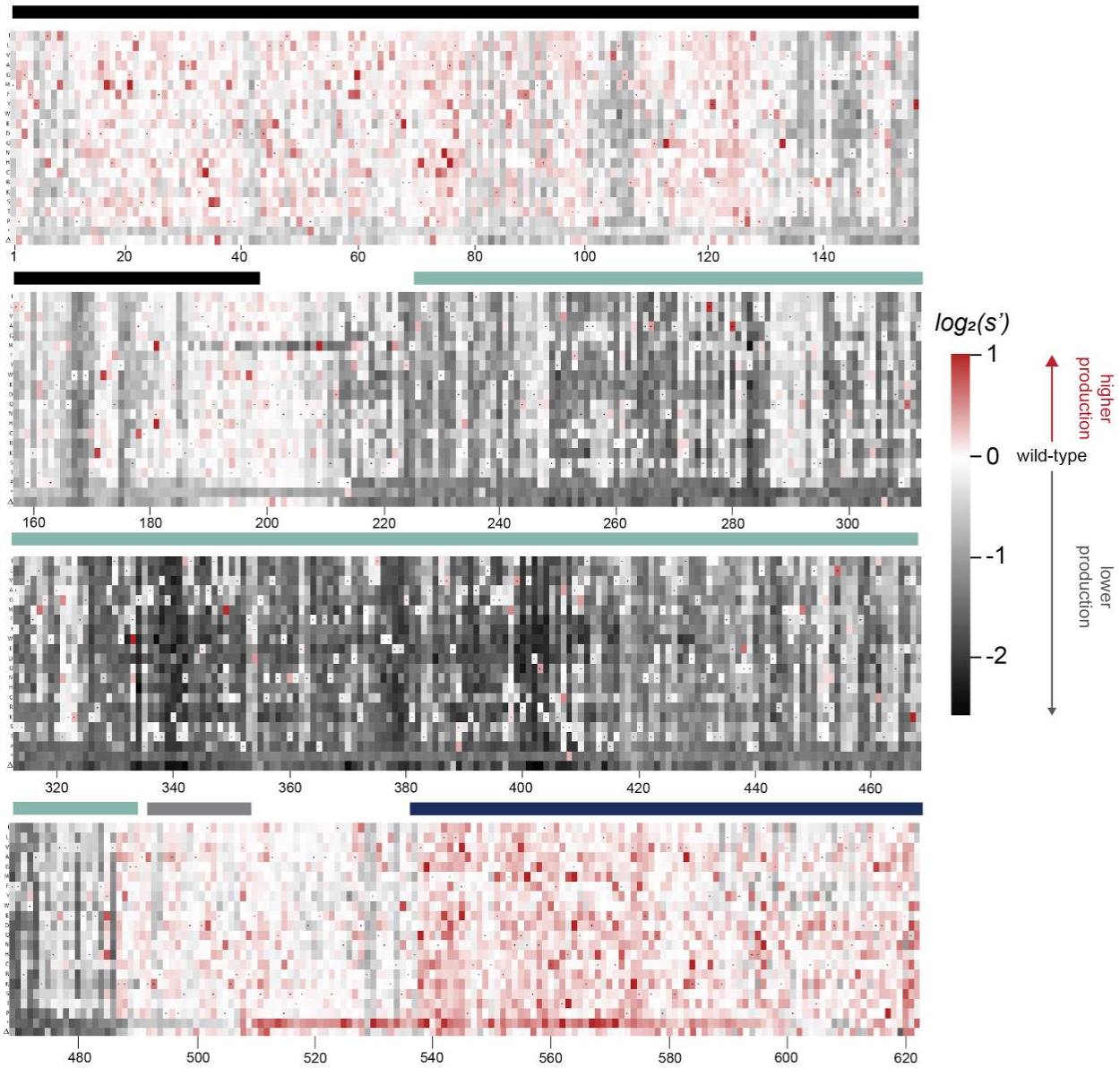

Figure S2. Effects of all single amino acid substitutions and deletions in Rep78, Rep68, Rep52, and Rep40 on AAV2 production. Amino-acid level production fitness values from the WT AAV2 library production assay were calculated as in Figure 1C by summing barcode counts for synonymous variants. Rectangles are colored by mutational effect on the production of genome-containing particles, with black indicating deleterious mutations and red indicating beneficial mutations. Colored bars above the heatmaps indicate protein domains. Black: origin-binding domain, light blue: helicase domain, gray: nuclear localization signal, and navy blue: zinc-finger domain. Black dots indicate wild-type amino acid identity.

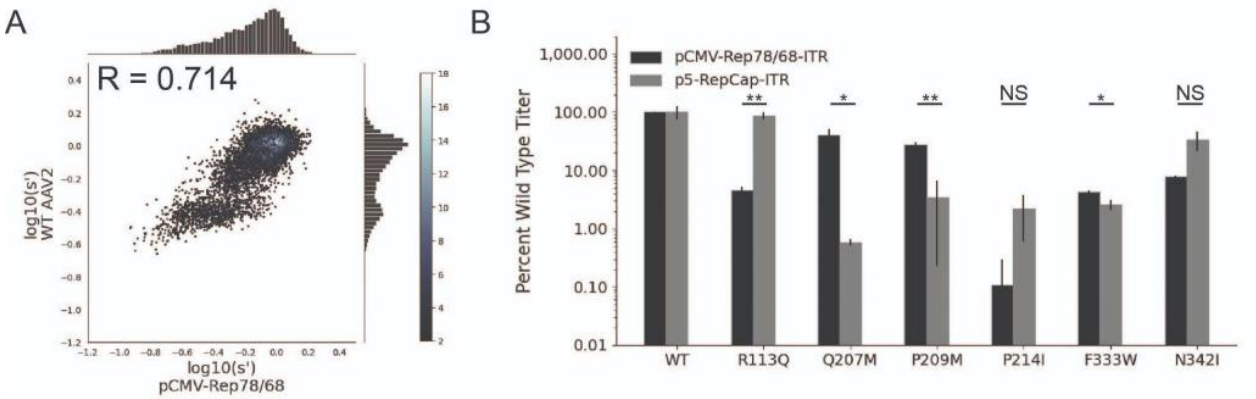

Figure S3. Mutations to AAV2 *rep* have similar effects on AAV2 production in pCMV-Rep78/68 and WT AAV2 format libraries. (A) Production fitness values for library production assay with pCMV-Rep78/68 and WT AAV2 format libraries, Pearson R correlation coefficient calculated after log transformation, (B) DNase-resistant particle titers for AAV2 capsids produced with pCMV-Rep78/68-inverted terminal repeat or p5-RepCap-inverted terminal repeat plasmids containing the indicated Rep substitutions. *$p < 0.05$, **$p < 0.01$ (Welch's t test)

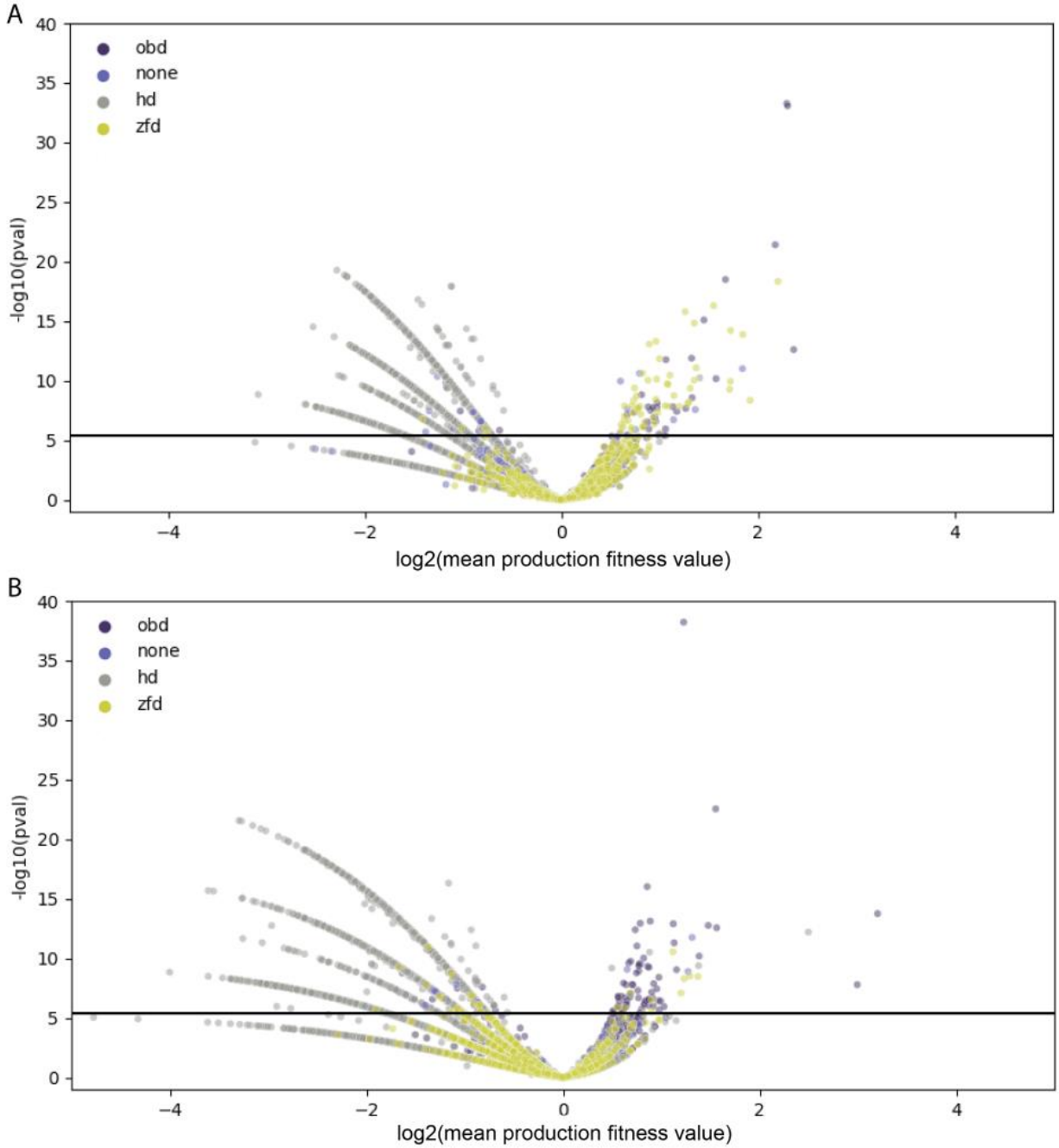

Figure S4. Comparison of Rep variant and wild-type production fitness values. For each amino acid variant, mean production fitness values were calculated by averaging the production fitness values for all uniquely barcoded variants coding for the relevant amino acid change. The mean production fitness value for each amino acid variant was compared to the mean production fitness value for the wild-type controls using a t-test. (A) Data for the WT AAV2 library. (B) Data for the pCMV-Rep78/68 library. The solid black line indicates the significance threshold after Bonferroni correction. Data points are colored according to their position in Rep where “none” refers to variants that fall outside of an annotated protein domain.

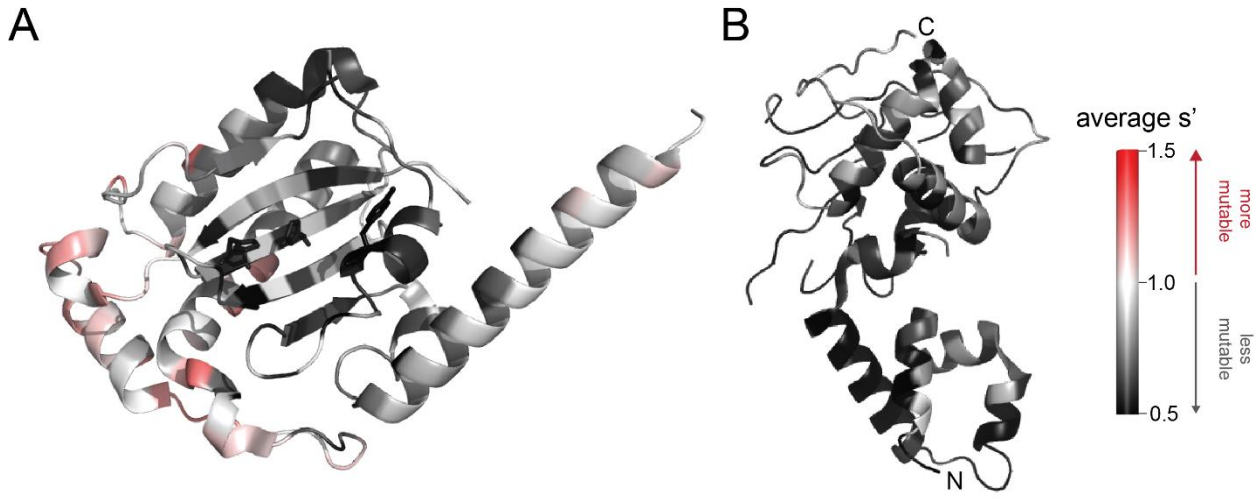

Figure S5. Average production fitness values from the pCMV-Rep78/68 library production assay mapped onto (A) the structure of the origin-binding domain active site (PDB 5DCX) and (B) the structure of the helicase domain (PDB 1S9H). H90, H92, and the active site nucleophile, Y156, are shown as sticks in (A). Residues are colored by mutability, with red indicating higher mutability and black indicating lower mutability.

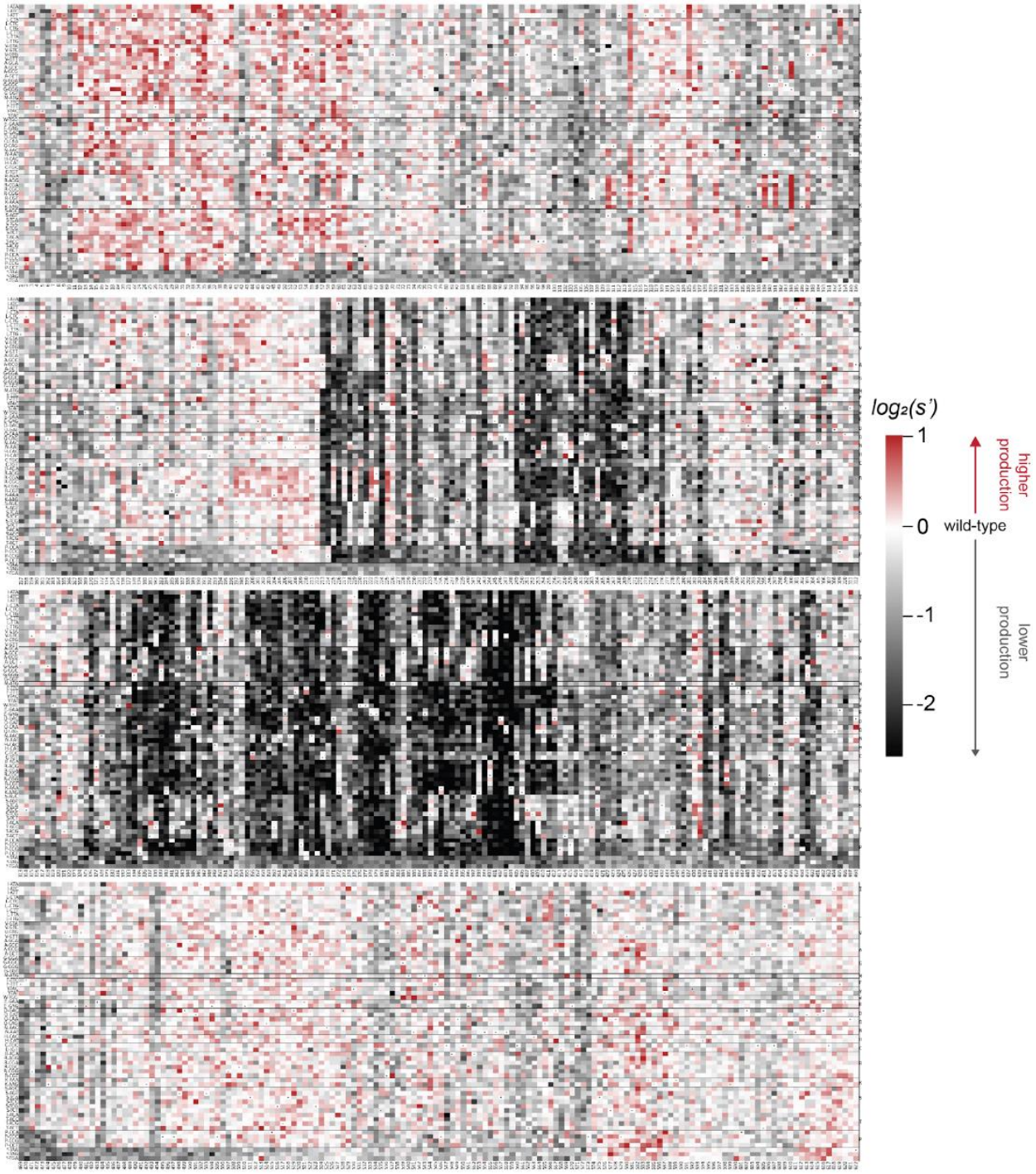

Figure S6. Codon level production fitness values for the pCMV-Rep78/68 format library. Codon level production fitness values were calculated as in Figure 1C by summing counts for all barcodes corresponding to a given codon variant. Rectangles are colored by mutational effect on the production of genome-containing particles, with black indicating deleterious mutations and red indicating beneficial mutations. Black dots indicate wild-type codon identity.

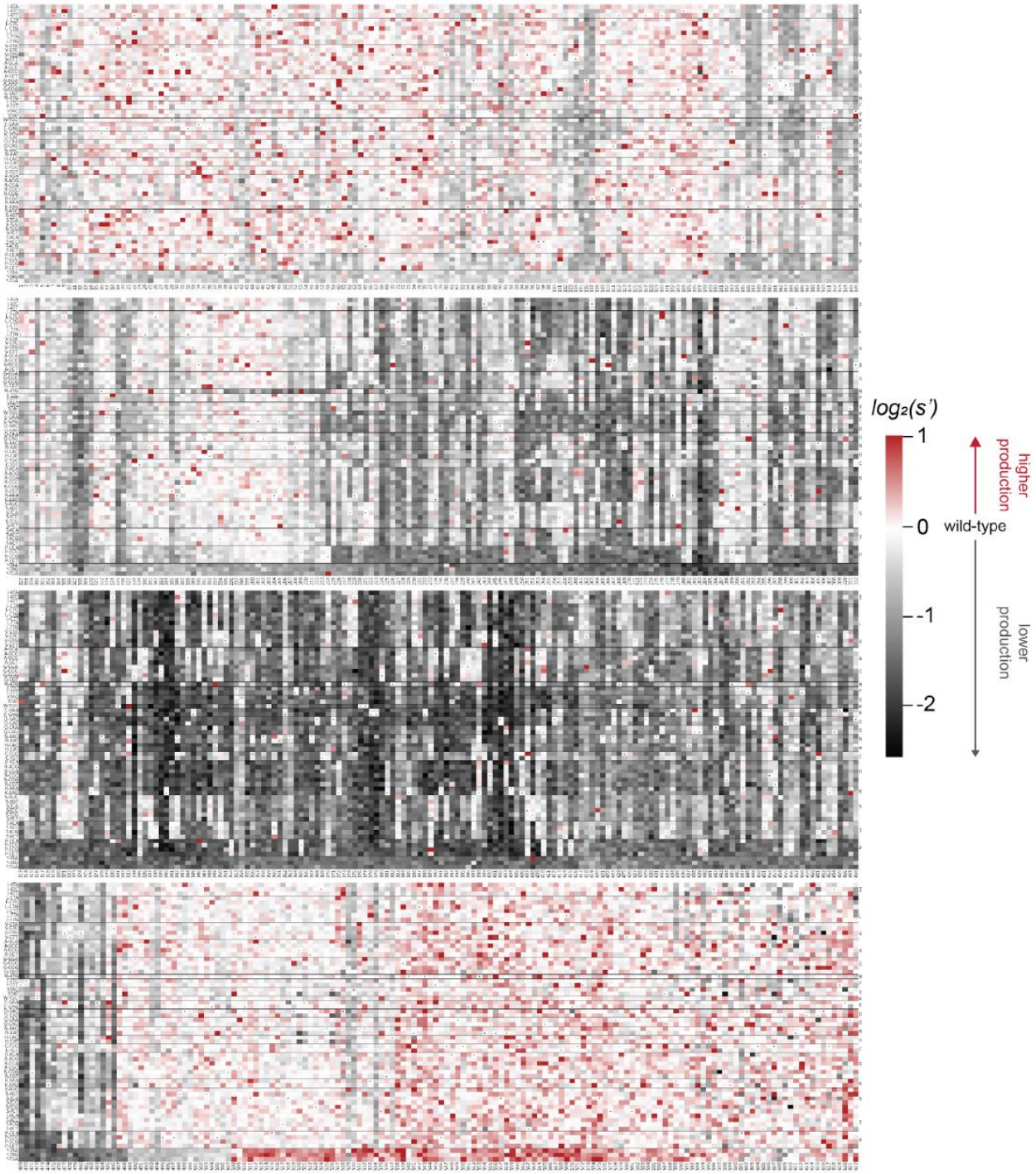

Figure S7. Codon level production fitness values for the WT AAV2 format library. Codon level production fitness values were calculated as in Figure 1C by summing counts for all barcodes corresponding to a given codon variant. Rectangles are colored by mutational effect on the production of genome-containing particles, with black indicating deleterious mutations and red indicating beneficial mutations. Black dots indicate wild-type codon identity.

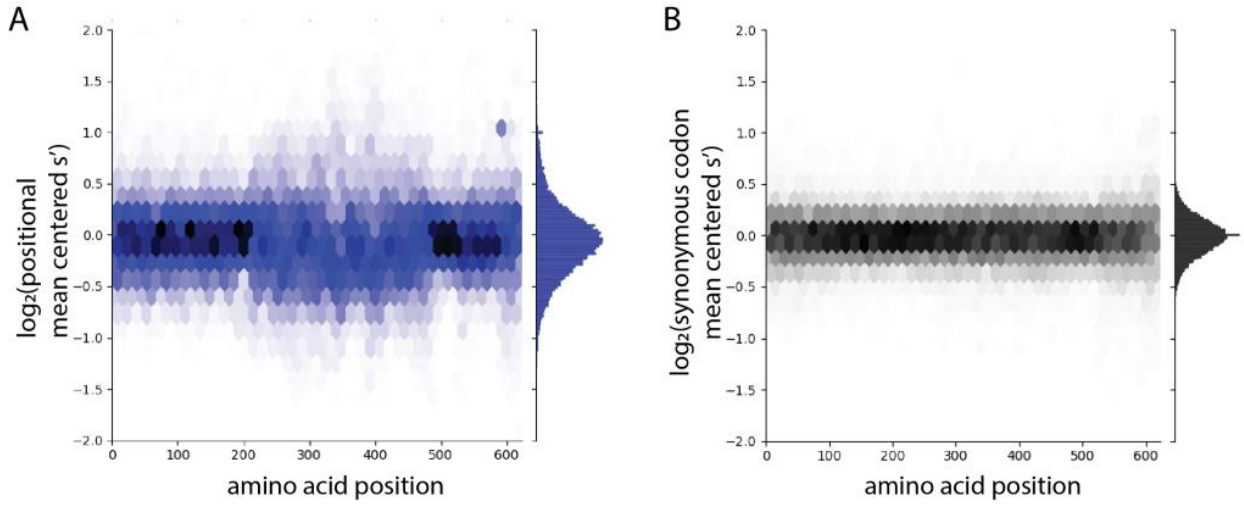

Figure S8. The distribution of production fitness values is narrower for synonymous variants than for nonsynonymous variants. (A) Positional mean centered fitness values for all codon variants in the WT AAV2 library and (B) Synonymous codon mean centered fitness values for all codon variants in the WT AAV2 library. For each codon variant, positional mean centered fitness values were calculated by normalizing to the average selection value of all codon variants at that position. Synonymous mean centered fitness values were calculated by normalizing the fitness values of each codon variant to the average fitness value of all its synonymous codons.

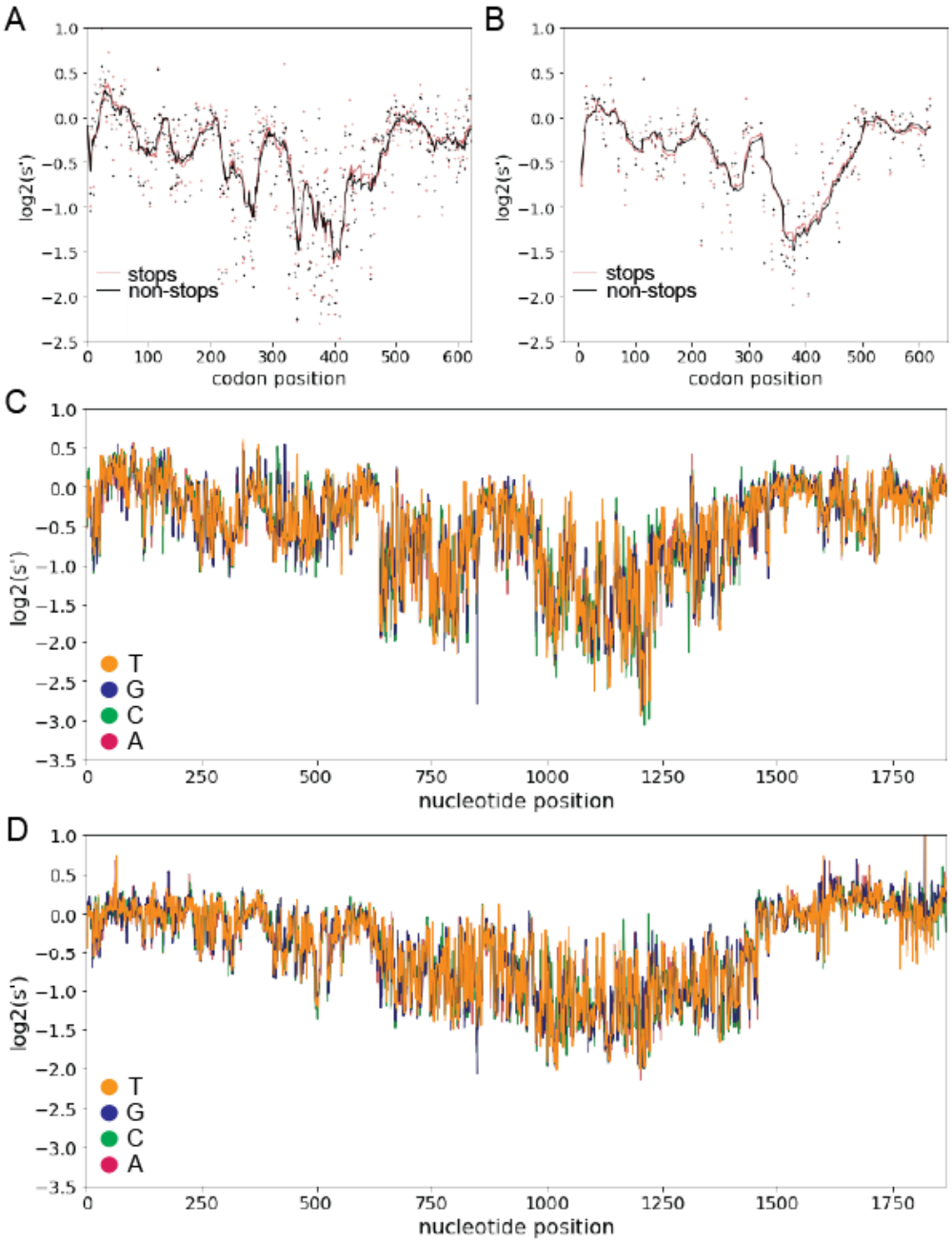

Figure S9. Inclusion of synonymous codon variants enables interrogation of nucleotide-level effects. Analysis of pCMV-Rep78/68 variants with premature stop codons in +1 (A) and +2 (B) reading frames does not identify any frameshifted open reading frames. Pink dots: average production fitness for mutations that introduce stop codons into +1 or +2 frame, black dots: average production fitness for mutations synonymous to +1 or +2 stop codon mutations in Rep open reading frame, lines indicate 10 bp rolling averages. (C) Average production fitness at each nucleotide position for variants that introduce the indicated nucleotide change in the pCMV-Rep78/68 library, orange: T, blue, G, green: C, red: A. (D) Average production fitness for variants that introduce the indicated nucleotide change in the WT AAV2 library.

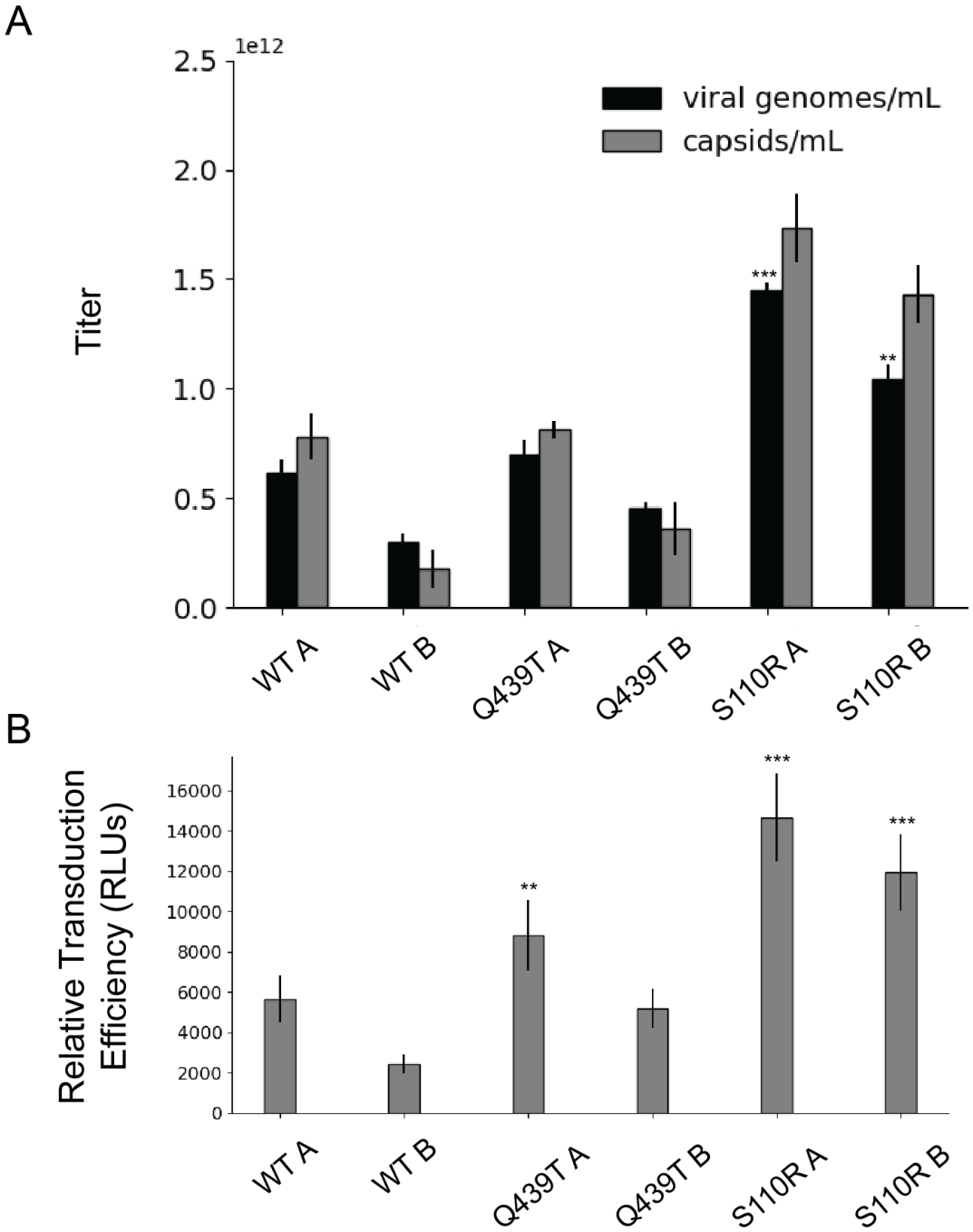

Figure S10. Effect of single amino acid Rep substitutions on the viral genome titer, physical particle titer, and relative transduction efficiency of affinity purified rAAV2. (A) Capsid (gray) and viral genome (black) titers for pAAV-EF1a-FLuc-WPRE-HGHpA rAAV2 produced with the indicated Rep variant and affinity purified with AVB Sepharose. For each variant, samples A and B represent replicate transfections and affinity purifications. (B) Relative transduction efficiency of purified rAAV preps as measured by luciferase signal. Asterisks indicate significance of difference between relevant prep and WT A prep. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$ (Welch's t-test)

Table S2. Viral genome and physical particle titers for rAAV2 produced with Rep variants

Rep variant	Replicate	qPCR titer (vg/mL)	qPCR std. dev. (vg/mL)	ELISA titer (capsids/mL)	ELISA titer std. dev. (capsids/mL)	Viral genome: capsid titer ratio
WT	**A**	6.18E+11	5.77E+10	7.82E+11	1.04E+11	0.79
	B	3.00E+11	3.85E+10	1.77E+11	8.95E+10	1.70
Q439T	**A**	7.01E+11	6.47E+10	8.14E+11	4.27E+10	0.86
	B	4.55E+11	2.33E+10	3.61E+11	1.21E+11	1.26
S110R	**A**	1.45E+12	3.24E+10	1.73E+12	1.56E+11	0.84
	B	1.04E+12	6.81E+10	1.43E+12	1.31E+11	0.73

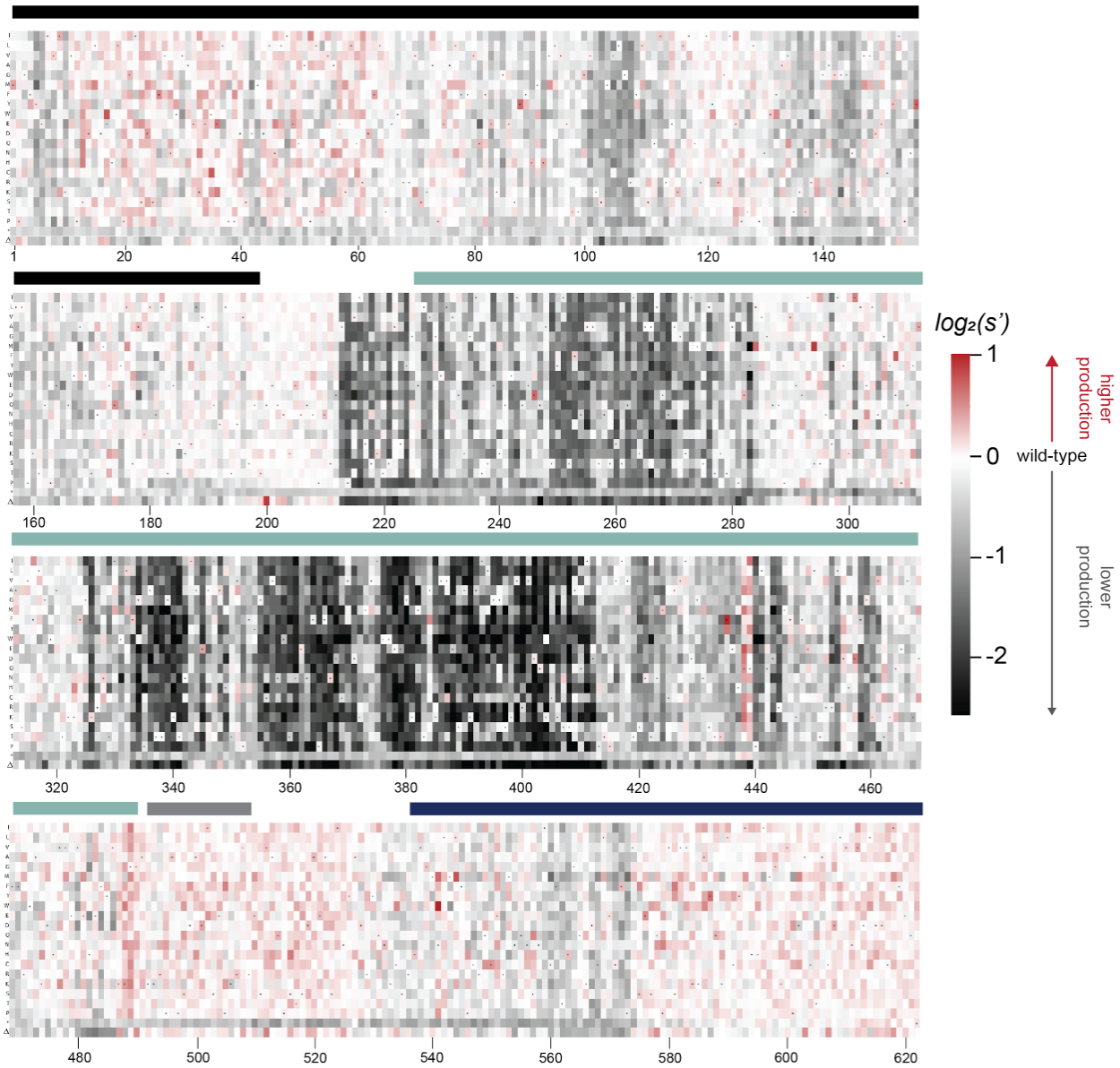

Figure S11. Effects of all single amino acid substitutions and deletions in the Rep78/68 proteins on AAV5 capsid production. Amino acid level production fitness values from the pCMV-Rep78/68 production assay were calculated as in Figure 1C by summing barcode counts for synonymous mutations. Rectangles are colored by mutational effect on the production of genome-containing particles, with black indicating deleterious mutations and red indicating beneficial mutations. Colored bars above the heatmaps indicate protein domains. Black: origin-binding domain, light blue: helicase domain, gray: nuclear localization signal, and navy blue: zinc-finger domain. Black dots indicate wild-type amino acid identity.

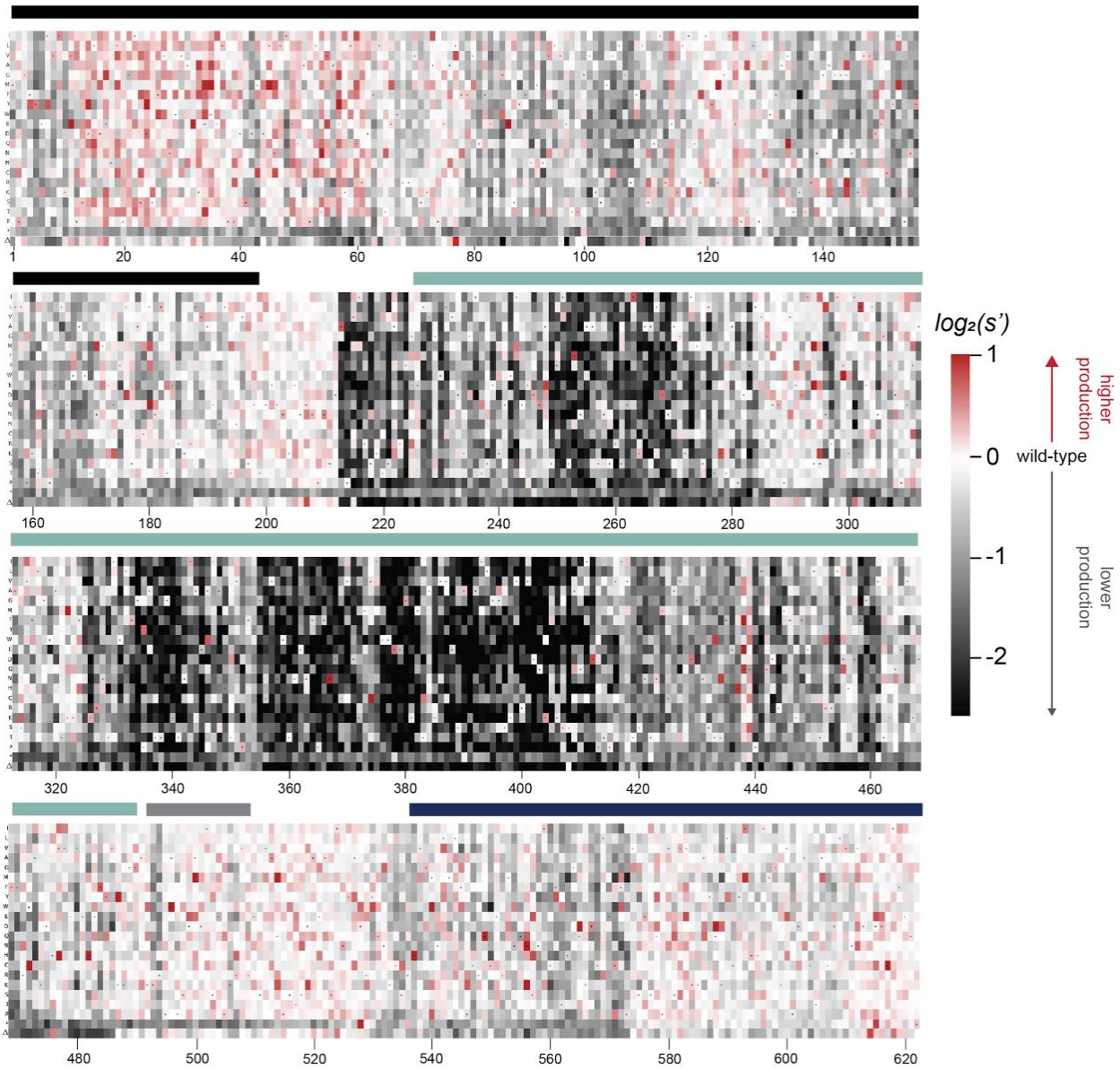

Figure S12. Effects of all single amino acid substitutions and deletions in the Rep78/68 proteins on AAV9 capsid production. Amino acid level production fitness values from the pCMV-Rep78/68 production assay were calculated as in Figure 1C by summing barcode counts for synonymous mutations. Rectangles are colored by mutational effect on the production of genome-containing particles, with black indicating deleterious mutations and red indicating beneficial mutations. Colored bars above the heatmaps indicate protein domains. Black: origin-binding domain, light blue: helicase domain, gray: nuclear localization signal, and navy blue: zinc-finger domain. Black dots indicate wild-type amino acid identity.

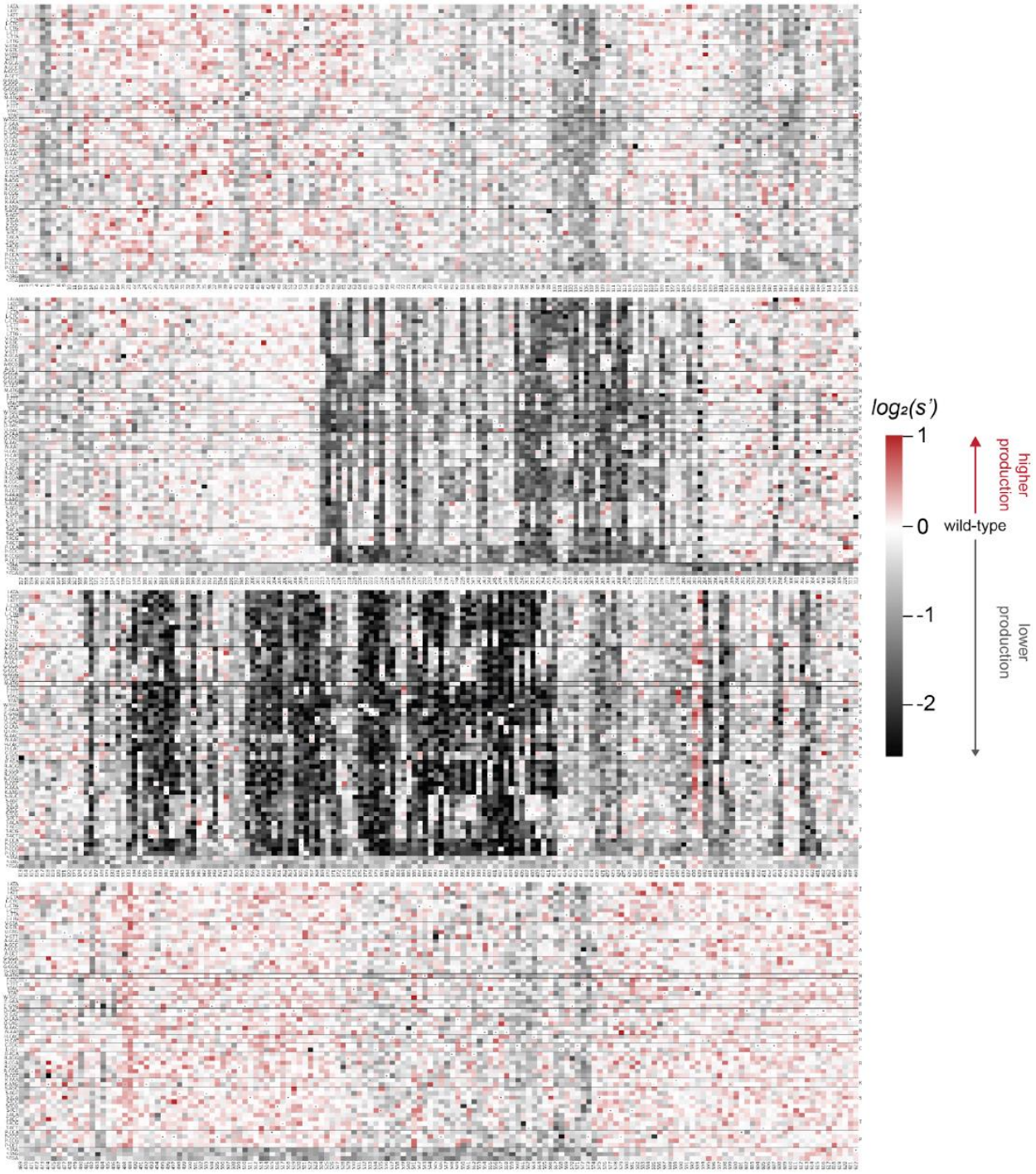

Figure S13. Codon level production fitness values for the AAV5 capsid production assay. Codon level production fitness values were calculated as in Figure 1C by summing counts for all barcodes corresponding to a given codon variant. Rectangles are colored by mutational effect on the production of genome-containing particles, with black indicating deleterious mutations and red indicating beneficial mutations. Black dots indicate wild-type codon identity.

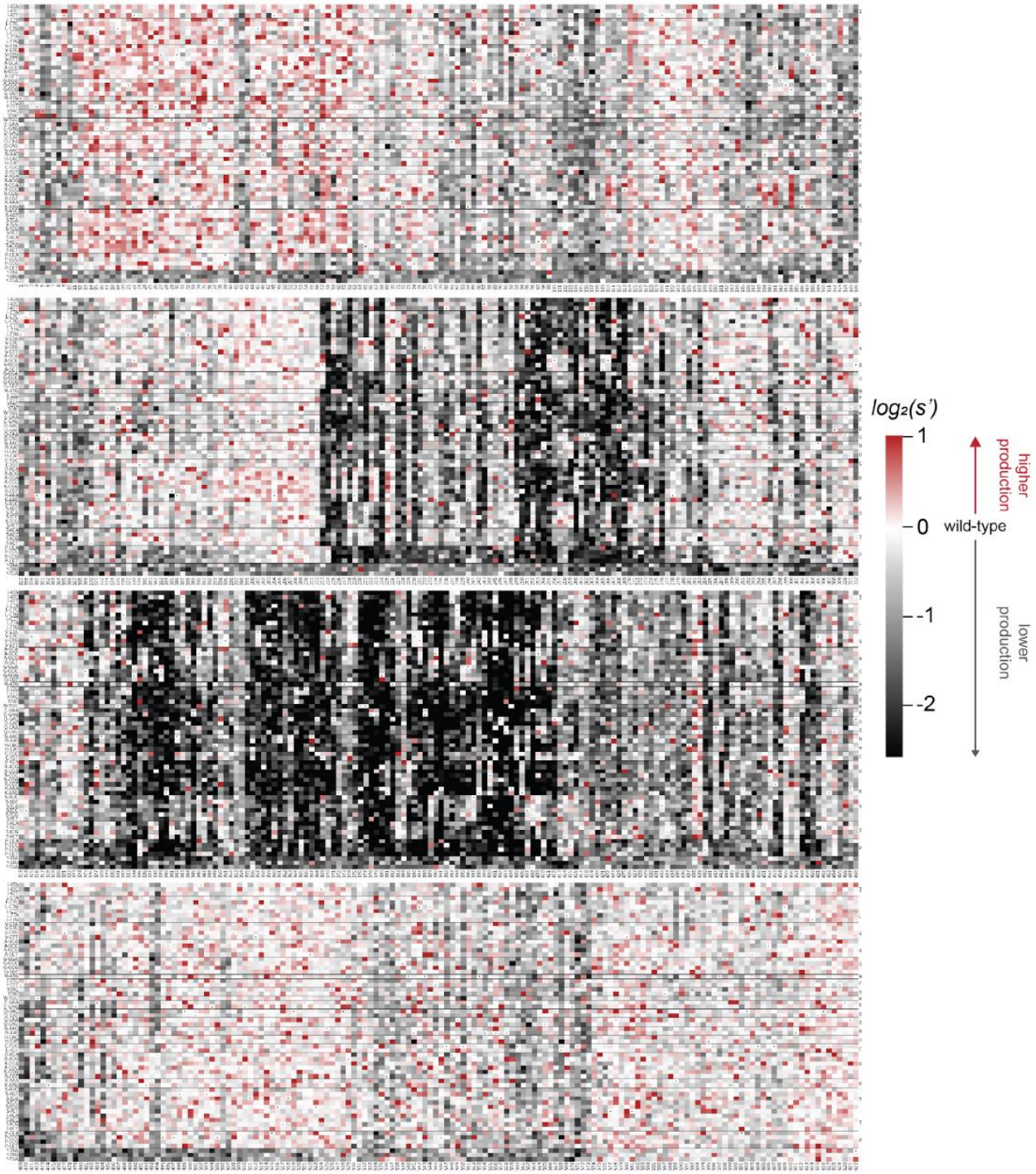

Figure S14. Codon level production fitness values for the AAV9 capsid production assay. Codon level production fitness values were calculated as in Figure 1C by summing counts for all barcodes corresponding to a given codon variant. Rectangles are colored by mutational effect on the production of genome-containing particles, with black indicating deleterious mutations and red indicating beneficial mutations. Black dots indicate wild-type codon identity.

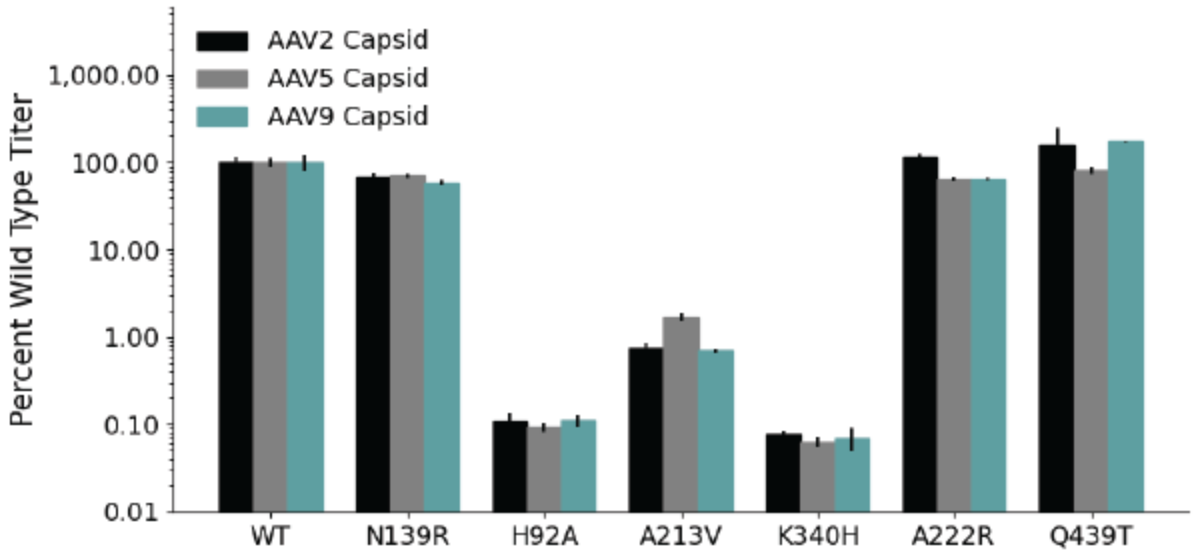

Figure S15. DNase-resistant particle titers for AAV2, AAV5, and AAV9 capsids produced individually with the indicated Rep variants.

References

Bleker, S., Pawlita, M., & Kleinschmidt, J. A. (2006). Impact of capsid conformation and Rep-capsid interactions on adeno-associated virus type 2 genome packaging. *Journal of Virology*, *80*(2), 810–820. https://doi.org/10.1128/JVI.80.2.810-820.2006

Brister, J. R., & Muzyczka, N. (1999). Rep-Mediated Nicking of the Adeno-Associated Virus Origin Requires Two Biochemical Activities, DNA Helicase Activity and Transesterification. *Journal of Virology*, *73*(11), 9325–9336. https://doi.org/10.1128/JVI.73.11.9325-9336.1999

Cassell, G. D., & Weitzman, M. D. (2004). Characterization of a nuclear localization signal in the C-terminus of the adeno-associated virus Rep68/78 proteins. *Virology*, *327*(2), 206–214. https://doi.org/10.1016/j.virol.2004.06.034

Cassinotti, P., Weitzand, M., & Tratschin, J.-D. (1988). Organization of the adeno-associated virus (AAV) capsid gene: Mapping of a minor spliced mRNA coding for virus capsid protein. *Virology*, *167*(1), 176–184. https://doi.org/10.1016/0042-6822(88)90067-0

Chejanovsky, N., & Carter, B. J. (1989). Mutagenesis of an AUG codon in the adeno-associated virus rep gene: Effects on viral DNA replication. *Virology*, *173*(1), 120–128. https://doi.org/10.1016/0042-6822(89)90227-4

Davis, M. D., Wu, J., & Owens, R. A. (2000). Mutational Analysis of Adeno-Associated Virus Type 2 Rep68 Protein Endonuclease Activity on Partially Single-Stranded Substrates. *Journal of Virology*, *74*(6), 2936–2942.

Di Pasquale, G., & Stacey, S. N. (1998). Adeno-Associated Virus Rep78 Protein Interacts with Protein Kinase A and Its Homolog PRKX and Inhibits CREB-Dependent Transcriptional Activation. *Journal of Virology*, *72*(10), 7916–7925. https://doi.org/10.1128/JVI.72.10.7916-7925.1998

Gavin, D. K., Young, S. M., Xiao, W., Temple, B., Abernathy, C. R., Pereira, D. J., Muzyczka, N., & Samulski, R. J. (1999). Charge-to-Alanine Mutagenesis of the Adeno-Associated Virus Type 2 Rep78/68 Proteins Yields Temperature-Sensitive and Magnesium-Dependent Variants. *Journal of Virology*, *73*(11), 9433–9445. https://doi.org/10.1128/JVI.73.11.9433-9445.1999

Hickman, A. B., Ronning, D. R., Kotin, R. M., & Dyda, F. (2002). Structural Unity among Viral Origin Binding Proteins: Crystal Structure of the Nuclease Domain of Adeno-Associated Virus Rep. *Molecular Cell*, *10*(2), 327–337. https://doi.org/10.1016/S1097-2765(02)00592-0

Hickman, A. B., Ronning, D. R., Perez, Z. N., Kotin, R. M., & Dyda, F. (2004). The Nuclease Domain of Adeno-Associated Virus Rep Coordinates Replication Initiation Using Two Distinct DNA Recognition Interfaces. *Molecular Cell*, *13*(3), 403–414. https://doi.org/10.1016/S1097-2765(04)00023-1

Hölscher, C., Kleinschmidt, J. A., & Bürkle, A. (1995). High-level expression of adeno-associated virus (AAV) Rep78 or Rep68 protein is sufficient for infectious-particle formation by a rep-negative AAV mutant. *Journal of Virology*, *69*(11), 6880–6885. https://doi.org/10.1128/jvi.69.11.6880-6885.1995

Hörer, M., Weger, S., Butz, K., Hoppe-Seyler, F., Geisen, C., & Kleinschmidt, J. A. (1995). Mutational analysis of adeno-associated virus Rep protein-mediated inhibition of heterologous and homologous promoters. *Journal of Virology*, *69*(9), 5485–5496. https://doi.org/10.1128/jvi.69.9.5485-5496.1995

Im, D.-S., & Muzyczka, N. (1990). The AAV origin binding protein Rep68 is an ATP-dependent site-specific endonuclease with DNA helicase activity. *Cell*, *61*(3), 447–457. https://doi.org/10.1016/0092-8674(90)90526-K

James, J. A., Escalante, C. R., Yoon-Robarts, M., Edwards, T. A., Linden, R. M., & Aggarwal, A. K. (2003). Crystal Structure of the SF3 Helicase from Adeno-Associated Virus Type 2.

Structure, *11*(8), 1025–1035. https://doi.org/10.1016/S0969-2126(03)00152-7

King, J. A., Dubielzig, R., Grimm, D., & Kleinschmidt, J. A. (2001). DNA helicase-mediated packaging of adeno-associated virus type 2 genomes into preformed capsids. *The EMBO Journal*, *20*(12), 3282–3291. https://doi.org/10.1093/emboj/20.12.3282

Labow, M. A., Hermonat, P. L., & Berns, K. I. (1986). Positive and negative autoregulation of the adeno-associated virus type 2 genome. *Journal of Virology*, *60*(1), 251–258. https://doi.org/10.1128/JVI.60.1.251-258.1986

Liu, L. (2014). Fields Virology, 6th Edition. *Clinical Infectious Diseases*, *59*(4), 613. https://doi.org/10.1093/cid/ciu346

Mietzsch, M., Eddington, C., Jose, A., Hsi, J., Chipman, P., Henley, T., Choudhry, M., McKenna, R., & Agbandje-McKenna, M. (2021). Improved Genome Packaging Efficiency of Adeno-associated Virus Vectors Using Rep Hybrids. *Journal of Virology*, *95*(19), e00773-21. https://doi.org/10.1128/JVI.00773-21

Mietzsch, M., Smith, J. K., Yu, J. C., Banala, V., Emmanuel, S. N., Jose, A., Chipman, P., Bhattacharya, N., McKenna, R., & Agbandje-McKenna, M. (2020). Characterization of AAV-Specific Affinity Ligands: Consequences for Vector Purification and Development Strategies. *Molecular Therapy - Methods & Clinical Development*, *19*, 362–373. https://doi.org/10.1016/j.omtm.2020.10.001

Murphy, M., Gomos-Klein, J., Stankic, M., & Falck-Pedersen, E. (2007). Adeno-Associated Virus Type 2 p5 Promoter: A Rep-Regulated DNA Switch Element Functioning in Transcription, Replication, and Site-Specific Integration. *Journal of Virology*, *81*(8), 3721–3730. https://doi.org/10.1128/JVI.02693-06

Musayev, F. N., Zarate-Perez, F., Bardelli, M., Bishop, C., Saniev, E. F., Linden, R. M., Henckaerts, E., & Escalante, C. R. (2015). Structural Studies of AAV2 Rep68 Reveal a Partially Structured Linker and Compact Domain Conformation. *Biochemistry*, *54*(38), 5907–5919. https://doi.org/10.1021/acs.biochem.5b00610

Musayev, F. N., Zarate-Perez, F., Bishop, C., Burgner, J. W., & Escalante, C. R. (2015). Structural Insights into the Assembly of the Adeno-associated Virus Type 2 Rep68 Protein on the Integration Site AAVS1*. *Journal of Biological Chemistry*, *290*(46), 27487–27499. https://doi.org/10.1074/jbc.M115.669960

Ogden, P. J., Kelsic, E. D., Sinai, S., & Church, G. M. (2019). Comprehensive AAV capsid fitness landscape reveals a viral gene and enables machine-guided design. *Science*, *366*(6469), 1139–1143. https://doi.org/10.1126/science.aaw2900

Paulk, N. K., Pekrun, K., Zhu, E., Nygaard, S., Li, B., Xu, J., Chu, K., Leborgne, C., Dane, A. P., Haft, A., Zhang, Y., Zhang, F., Morton, C., Valentine, M. B., Davidoff, A. M., Nathwani, A. C., Mingozzi, F., Grompe, M., Alexander, I. E., … Kay, M. A. (2018). Bioengineered AAV Capsids with Combined High Human Liver Transduction In Vivo and Unique Humoral Seroreactivity. *Molecular Therapy*, *26*(1), 289–303. https://doi.org/10.1016/j.ymthe.2017.09.021

Pekrun, K., Alencastro, G. D., Luo, Q.-J., Liu, J., Kim, Y., Nygaard, S., Galivo, F., Zhang, F., Song, R., Tiffany, M. R., Xu, J., Hebrok, M., Grompe, M., & Kay, M. A. (2019). Using a barcoded AAV capsid library to select for clinically relevant gene therapy vectors. *JCI Insight*, *4*(22). https://doi.org/10.1172/jci.insight.131610

Pereira, D. J., McCarty, D. M., & Muzyczka, N. (1997). The adeno-associated virus (AAV) Rep protein acts as both a repressor and an activator to regulate AAV transcription during a productive infection. *Journal of Virology*, *71*(2), 1079–1088. https://doi.org/10.1128/jvi.71.2.1079-1088.1997

Samulski, R. J., & Muzyczka, N. (2014). AAV-Mediated Gene Therapy for Research and Therapeutic Purposes. *Annual Review of Virology*, *1*(1), 427–451. https://doi.org/10.1146/annurev-virology-031413-085355

Santosh, V., Musayev, F. N., Jaiswal, R., Zárate-Pérez, F., Vandewinkel, B., Dierckx, C.,

Endicott, M., Sharifi, K., Dryden, K., Henckaerts, E., & Escalante, C. R. (2020). The Cryo-EM structure of AAV2 Rep68 in complex with ssDNA reveals a malleable AAA+ machine that can switch between oligomeric states. *Nucleic Acids Research*, *48*(22), 12983–12999. https://doi.org/10.1093/nar/gkaa1133

Smith, R. H., & Kotin, R. M. (1998). The Rep52 Gene Product of Adeno-Associated Virus Is a DNA Helicase with 3′-to-5′ Polarity. *Journal of Virology*, *72*(6), 4874–4881. https://doi.org/10.1128/JVI.72.6.4874-4881.1998

Snyder, R. O., Im, D. S., & Muzyczka, N. (1990). Evidence for covalent attachment of the adeno-associated virus (AAV) rep protein to the ends of the AAV genome. *Journal of Virology*, *64*(12), 6204–6213. https://doi.org/10.1128/jvi.64.12.6204-6213.1990

Srivastava, A., Lusby, E. W., & Berns, K. I. (1983). Nucleotide sequence and organization of the adeno-associated virus 2 genome. *Journal of Virology*, *45*(2), 555–564. https://doi.org/10.1128/jvi.45.2.555-564.1983

Stahnke, S., Lux, K., Uhrig, S., Kreppel, F., Hösel, M., Coutelle, O., Ogris, M., Hallek, M., & Büning, H. (2011). Intrinsic phospholipase A2 activity of adeno-associated virus is involved in endosomal escape of incoming particles. *Virology*, *409*(1), 77–83. https://doi.org/10.1016/j.virol.2010.09.025

Stutika, C., Gogol-Döring, A., Botschen, L., Mietzsch, M., Weger, S., Feldkamp, M., Chen, W., & Heilbronn, R. (2016). A Comprehensive RNA Sequencing Analysis of the Adeno-Associated Virus (AAV) Type 2 Transcriptome Reveals Novel AAV Transcripts, Splice Variants, and Derived Proteins. *Journal of Virology*, *90*(3), 1278–1289. https://doi.org/10.1128/JVI.02750-15

Surosky, R. T., Urabe, M., Godwin, S. G., McQuiston, S. A., Kurtzman, G. J., Ozawa, K., & Natsoulis, G. (1997). Adeno-associated virus Rep proteins target DNA sequences to a unique locus in the human genome. *Journal of Virology*, *71*(10), 7951–7959. https://doi.org/10.1128/jvi.71.10.7951-7959.1997

Wang, D., Tai, P. W. L., & Gao, G. (2019). Adeno-associated virus vector as a platform for gene therapy delivery. *Nature Reviews Drug Discovery*, *18*(5), Article 5. https://doi.org/10.1038/s41573-019-0012-9

Wörner, T. P., Bennett, A., Habka, S., Snijder, J., Friese, O., Powers, T., Agbandje-McKenna, M., & Heck, A. J. R. (2021). Adeno-associated virus capsid assembly is divergent and stochastic. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-21935-5

Wu, J., Davis, M. D., & Owens, R. A. (1999). Factors Affecting the Terminal Resolution Site Endonuclease, Helicase, and ATPase Activities of Adeno-Associated Virus Type 2 Rep Proteins. *Journal of Virology*, *73*(10), 8235–8244. https://doi.org/10.1128/JVI.73.10.8235-8244.1999

Xiao, X., Xiao, W., Li, J., & Samulski, R. J. (1997). A novel 165-base-pair terminal repeat sequence is the sole cis requirement for the adeno-associated virus life cycle. *Journal of Virology*, *71*(2), 941–948. https://doi.org/10.1128/jvi.71.2.941-948.1997

Xie, Q., Bu, W., Bhatia, S., Hare, J., Somasundaram, T., Azzi, A., & Chapman, M. S. (2002). The atomic structure of adeno-associated virus (AAV-2), a vector for human gene therapy. *Proceedings of the National Academy of Sciences*, *99*(16), 10405–10410. https://doi.org/10.1073/pnas.162250899

Yang, Q., Kadam, A., & Trempe, J. P. (1992). Mutational analysis of the adeno-associated virus rep gene. *Journal of Virology*, *66*(10), 6058–6069. https://doi.org/10.1128/jvi.66.10.6058-6069.1992

Zarate-Perez, F., Bardelli, M., Burgner, J. W., Villamil-Jarauta, M., Das, K., Kekilli, D., Mansilla-Soto, J., Linden, R. M., & Escalante, C. R. (2012). The interdomain linker of AAV-2 Rep68 is an integral part of its oligomerization domain: Role of a conserved SF3 helicase residue in oligomerization. *PLoS Pathogens*, *8*(6), e1002764.

https://doi.org/10.1371/journal.ppat.1002764

Zolotukhin, S., Byrne, B. J., Mason, E., Zolotukhin, I., Potter, M., Chesnut, K., Summerford, C., Samulski, R. J., & Muzyczka, N. (1999). Recombinant adeno-associated virus purification using novel methods improves infectious titer and yield. *Gene Therapy*, *6*(6), Article 6. https://doi.org/10.1038/sj.gt.3300938