Computational analysis of adeno-associated virus vectors for gene editing applications

Author: Tatevik Jalatyan
BS in Data Science
American University of Armenia

Supervisor: Dr. Lilit Nersisyan
Armenian Bioinformatics Institute

Supervisor: Dr. Erik Aznauryan
Harvard University

***Abstract*—Adeno-associated viruses (AAV) are one of the most actively investigated gene therapy vehicles. However, several factors challenge their applications in humans, such as immune response and inability to reach target tissues. Various strategies are being applied to produce novel AAV variants with properties that overcome these challenges. One of them is directed evolution, which facilitates the engineering of proteins with desired features by applying mutagenesis and selective pressure. In particular, DNA shuffling is used to produce novel variants by fragmentation and reassembly of AAV capsid genes. However, systemic computational analysis of resulting variants is still limited. This paper introduces a new computational tool that enables comprehensive exploratory analysis of AAV chimeric libraries and identification of successful variants by extracting quantitative data from the sequence libraries.**

***Index Terms*—adeno-associated virus, directed evolution, DNA shuffling, chimera**

I. Introduction

Adeno-associated virus (AAV) vectors have become widely used for gene therapy applications, with several advantages over other viral vectors, including lower toxicity and the ability to express transgenes in various specific tissue types. Simply put, AAV is a non-pathogenic virus that can be engineered to deliver DNA to target cells [1]. It has been shown that AAV vectors are successful in early-stage clinical trials for the treatment of a wide variety of rare genetic disorders [2].

AAV is a protein shell (capsid) surrounding a 4.7 kilobase long single-stranded DNA genome which encodes non-structural (rep), structural (cap), assembly activating (aap), and membrane associated accessory (maap) proteins. AAV capsids consist of a mixture of three viral proteins (VPs): VP1, VP2, and VP3 encoded within the cap gene. VP3 is the major capsid protein, accounting for approximately 50 of the 60 capsid monomers [2]. VP3 contains a highly conserved core region and nine distinct variable regions (VRs), which are associated with functional roles in the AAV life cycle essential for successful gene delivery, including receptor binding, tissue transduction, and antigenic specificity [3]. The AAV capsid is the primary factor for targeting specific cell types and evading the pre-existing human immune response. Hence, numerous strategies have been developed to engineer novel capsid variants with the aim of improving the target delivery and immune system evasion properties. One such method is directed evolution, which involves subjecting the capsid genes to iterative rounds of mutagenesis and selection. DNA shuffling is one of the widely used techniques for creating a library of chimeric variants (CV) by random fragmentation and reassembly of capsid genes from naturally occurring AAV serotypes. The subsequent step is the iterative selection of the resulting AAV variants by expressing the variants in target cells and isolating selected variants (SV) with the desired features [4]. To identify the initial chimeric variants and the selected desired variants in a rapid and cost-effective manner, high-throughput sequencing technologies are used. However, computational methods and tools for the analysis of such datasets are limited [5] [6] [7].

Here, we report an R and bash based computational tool, Hafoe, to facilitate the automated exploratory analysis of the AAV chimeric libraries and identification of selected variants with desired features. We demonstrate the performance and applications of the tool using in silico generated datasets.

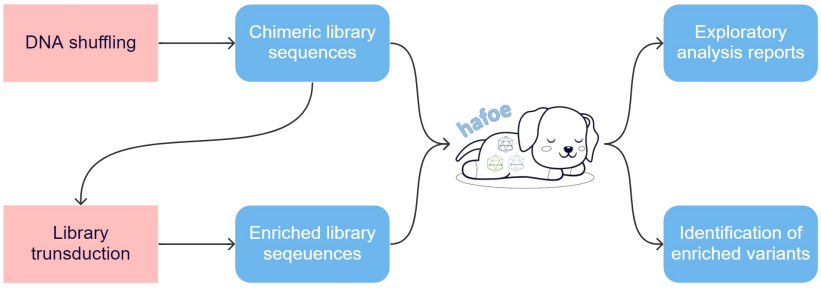
II. Results and Discussion

We have developed an R-based command-line tool, *hafoe*, to address the limitations of current tools to automatically analyze chimeric library sequencing data. *hafoe* was designed to perform two main functions: first, to explore the initial chimeric library sequencing data, reduce the redundancy by clustering, analyze parental contribution in variant sequences; and second, to identify the enriched variants which were able to enter the cells and get expressed.

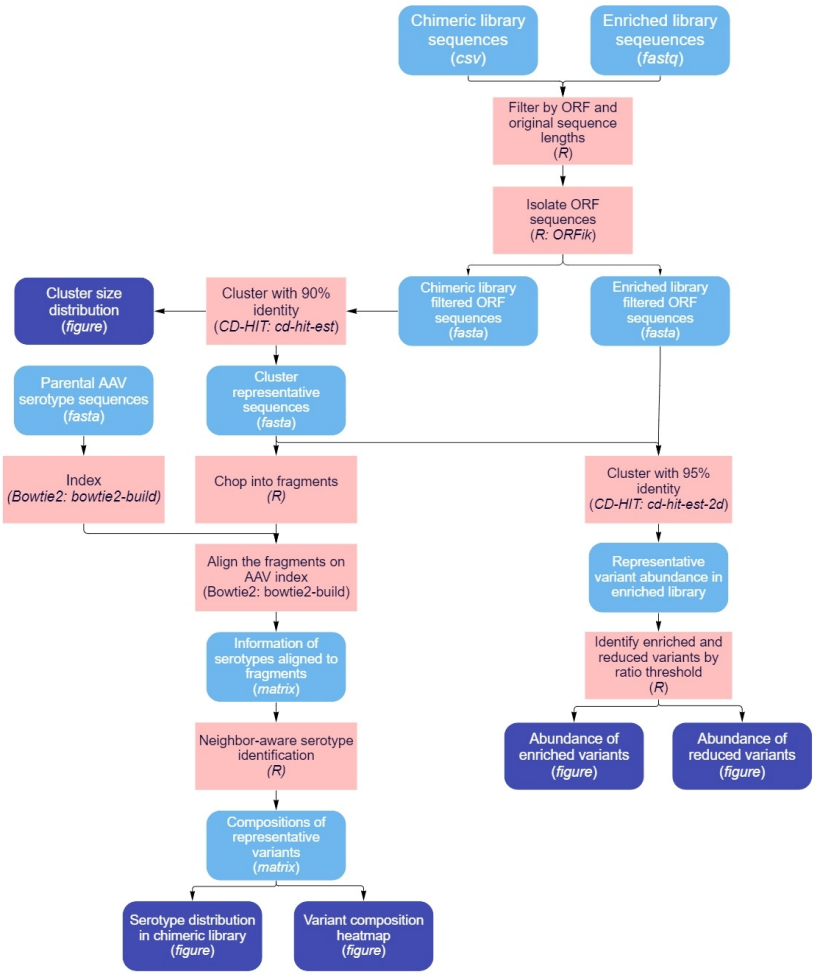
The user can provide only the chimeric library and specify the *-explore* option to perform only the first part of the analysis or provide both chimeric and enriched libraries by adding the *–identify* option to perform also the identification of enriched variants (Fig. 1).

A. Chimeric library exploratory analysis

In standard recombination based mutagenesis experiments, chimeric libraries are produced in the following way. The capsid genes of wild type AAV serotypes are cut into fragments with DNase I enzyme. These fragments are then filtered by length and reassembled by annealing to one another in homologous regions and subsequent extension of DNA strands with a DNA polymerase enzyme [reference needed]. The chimeric variants produced in this process are sequenced using PacBio sequencing technology, which enables long-read sequencing with reads reaching tens of kilobases in length and achieves 99.99% consensus accuracy [8].



(a) *Hafoe* workflow



(b) General pipeline

Fig. 1: Schematic representations of the *hafoe* workflow (a) and the general pipeline of the program.

In this study, this process was simulated, and the input chimeric library of 7,759 sequences was generated *in silico* (see the "Materials and Methods" section for details).

The preprocessing of the chimeric and enriched library datasets was performed by isolating ORF sequences and filtering based on ORF length and original sequence length (Materials and methods). This is required to ensure that all the variants used in the downstream analysis have the capsid genes of the required size to be biologically viable.

Then the 7,759 sequences of the chimeric library were clustered into 89 groups with at least 90% identity of sequences within a group and the representative sequences of the clusters were used in subsequent analysis (see the "Materials and Methods" section for details) (Fig. 2).

To study the parental serotype contribution and generally understand the distribution of AAV serotypes in the chimeric library, we have applied our neighbor-aware serotype identification method to the representative sequences. (Fig. 3b, Fig. 4b). Shortly, the sequences were cut into short fragments and aligned to parental serotypes, while multiple alignments were

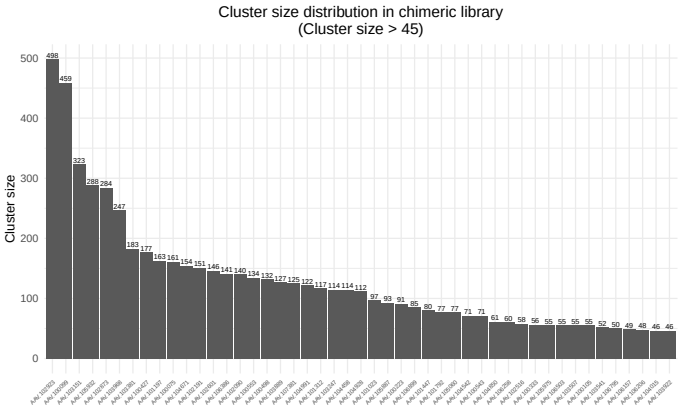


Fig. 2: Cluster size distribution in chimeric library. Representative sequences of each cluster are shown on x-axis.

resolved based on the alignment information of neighboring fragments (Materials and methods). Fig. 3a shows the actual AAV serotype distribution in the simulated chimeric library. We then assigned parental composition of the representative sequences using neighbor-aware serotype identification. Estimated frequencies of the parental AAV serotypes were then computed based on the cluster sizes of each representative (Fig. 3b). By comparing the actual distribution (Fig. 3a) with the estimated one (Fig. 3b), it is clear that the overall pattern of frequencies is very similar, with the most abundant serotypes being AAV2 and AAVrh8. However, we tended to underestimate the frequencies of the serotypes AAV7, AAV13, AAV1, and AAV3. The rest of the serotypes were almost entirely identified.

The actual (Fig. 4a) and predicted (Fig. 4b) compositions of each representative variant are visualized with heatmaps. Fragments with the "no alignment" label were mainly in the chimeric regions where two different serotypes are joined (Fig. 4a). This is an expected result as the fragments containing parts from more than one serotype may not align to a parental serotype if the original sequences are not homologous. In contrast, some regions tended to have multiple alignments (From Fig. 4b, columns 43-55 and 58-64), which is explained by conservation of the sequences in these regions across the parental AAV serotypes. In other words, as all serotypes in these regions have identical sequences, the program could not identify which serotypes these fragments are actually coming from. Overall, hafoe estimates parental serotype composition with 79% accuracy.

B. Enrichment identification

Usually the chimeric variants produced by directed random mutagenesis are used to transduce tissues of interest, and the variants capable of entering and expressing in those tissues are isolated. The resulting tissue-specific enriched variants are then sequenced and described.

Here we have simulated the process of enrichment with the assumption that having the desired features (entering the cell and expressing in it) is a rare event for a variant. Previously analyzed experimental data (not shown) also support this assumption. So, the abundance of the majority of variants should decrease in tissue-specific sequencing data (the enriched library) compared to initial DNA sequencing data (chimeric library). With this assumption we have simulated an enriched library with 3,024 sequences (see the "Materials and Methods" section for details).

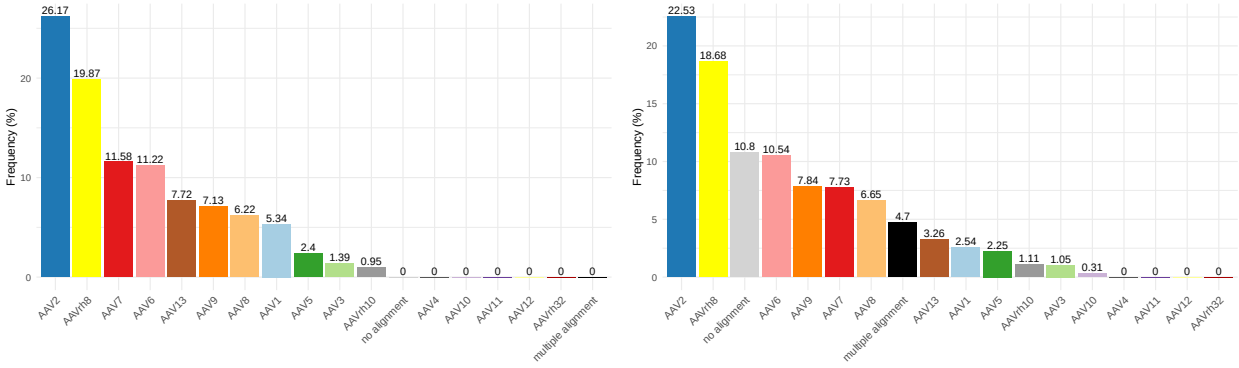
We have performed clustering with CD-HIT, identifying sequences in the enriched library that are similar to the representative sequences in the chimeric library at 95% identity threshold (Materials and methods). The sizes of the resulting 89 clusters were used as a proxy for the variants' abundance in the enriched library. The ratio of normalized fractions of representative variants in the chimeric library to the normalized number of representative variants in the enriched library was used as the final estimate of variant abundance (Fig. 5). The representative variants with a ratio greater than 1 were thought to be the enriched variants, and those with a ratio less than 1 were the reduced ones (Fig. 5).

From Fig. 10 it is obvious which variants are more successful in entering the cells and expressing in them. For example, the abundance of AAV.102159 increased from 0.58% in the chimeric library to 3.3%, and the abundance of AAV.102923 decreased from 6.87% to 3.88%. Even though both variants have a similar frequency in the enriched library, there is an almost 6-fold increase in AAV.102159 abundance and a 2-fold decrease in AAV.102923 abundance. So, not using the ratios would lead to inaccurate conclusions. This information can be easily used to isolate the successful variants from huge datasets of AAV variant sequences and test them in pre-clinical trials as gene therapy vectors.

To further analyze the sequence content of the enriched and reduced libraries, we have performed a multiple sequence alignment of the enriched and reduced variants separately (Fig. 6). We observe prevalence of AAV7 in the middle parts of the enriched variants, and prevalence of AAVrh8 in the reduced ones. Follow up analysis of parental serotype composition and their comparison between enriched and reduced variants will help elucidate which regions and of what origin are responsible for the desired features in the enriched variants.

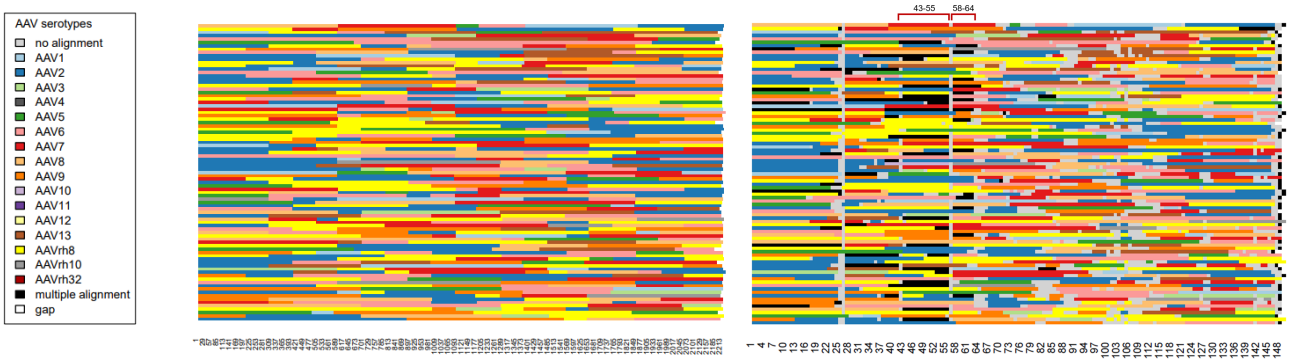
III. Conclusion

Design of AAV based delivery vehicles for gene therapy applications is an important aspect for treatment of a number of diseases. While experimental methods for obtaining novel AAV variants are taking a hit, computational means of analysis are still lagging behind. Here we introduce hafoe, a new computational tool, which enables comprehensive analysis of AAV chimeric variants obtained by directed random mutagenesis. Hafoe also enables analysis of the enrichment patterns of these variants in tissues of interest based on long-read sequencing data. The tool provides various visualizations and summary files that can be useful in exploring the produced chimeric variants and identifying those capable of entering the cells of interest and expressing in them. The program was tested on



(a) True (b) Predicted

Fig. 3: True distribution of AAV serotypes (a) and distribution of AAV serotypes obtained using neighbor-aware serotype identification method (b) in chimeric library. The rows represent the variants, and the columns represent the corresponding positions/fragments in the variants. The colors encode for the 16 parental AAV serotypes used to create the chimeric variants and for "no alignment" or "multiple alignments" if the fragment did not align to any serotype or aligned to multiple serotypes but remained unresolved by the program, respectively.



(a) True (b) Predicted

Fig. 4: Chimeric library representative variants' compositions based on true composition labels stored while generating the data (a) and based on the serotype alignment information obtained by neighbor-aware serotype identification method (b).

simulated datasets, leading to an accuracy of 79% in assessing parental composition of the chimeric variants. We also show how follow up studies on the results of enrichment analysis can be used to infer sequences responsible for the desired features.

IV. Materials and Methods

A. Implementation

Hafoe is written in R 4.1 and Bash 5.1 and can be executed in Unix operating systems. The input for the program are genome sequence files of parental AAV serotypes (in fasta format), the sequencing datasets from the chimeric library (in csv format), and the selected variant sequencing datasets from the enriched library (in fastq format) (optional).

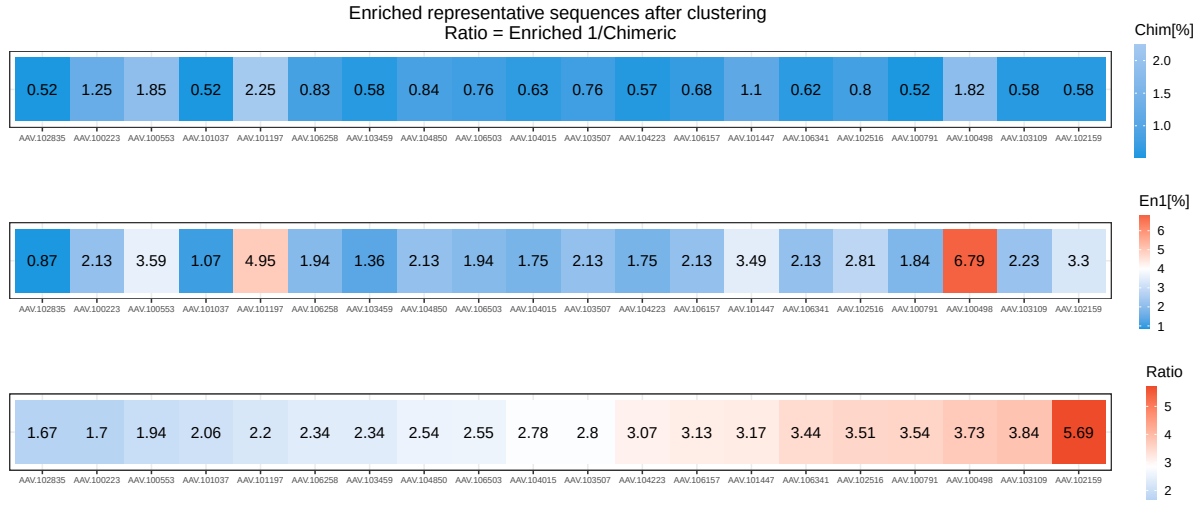
ORFik (v1.12.13), microseq (v2.1.4), seqinr (v4.2.8) are the main R packages used in the program to filter the sequences, extract open reading frame (ORF) sequences, read and write files in fasta and fastq formats, and read alignment files.

The indexing of parental AAV serotype sequences and the alignment of variant reads on it was performed by Bowtie2 (v2.4.2) program's *bowtie2-build* and *bowtie2* commands respectively [5]. CD-HIT (v4.8.1) program's *cd-hit-est* and *cd-hit-est-2d* commands were used for clustering of the chimeric sequences to reduce redundancy. [6]. Clustal Omega (v1.2.4) was used for multiple sequence alignment [7]. The overall pipeline is wrapped in bash scripts.

The R ggplot2 (v3.3.6) and gplots (v3.1.3) packages were used to visualize the obtained results.

B. Data simulation

To generate the chimeric library, first multiple sequence alignment of 16 parental AAV serotypes was performed using Clustal Omega (v1.2.4) program, which is used to identify regions of similarity and homology between multiple sequences of similar length [7]. This information was then used to simulate the DNA shuffling process. First, a serotype was randomly chosen from 16 parental AAV serotypes: AAV1-13, AAVrh8, AAVrh10, AAVrh32. Then a cut position was randomly chosen to produce a fragment of length 100-700 base pairs, as larger



(a) Enriched

(b) Reduced

Fig. 5: Enriched (a) and reduced (b) variants' abundances in chimeric and enriched libraries. The variants with ratio > 1 are enriched (a) and the variants with ratio < 1 are reduced (b) The top rows show the percentage of the variant in the chimeric library, the middle row shows percentage of the variant in the enriched library, and the bottom row shows the Ratio: % in the chimeric library / % in the enriched library.

fragments can reduce diversity and shorter fragments don't anneal properly [9]. The fragment from the start to cut position of the chosen serotype was used as a starting region of the new chimeric sequence. Next, another random cut position (downstream from the previous cut position) was chosen in a random serotype, and the fragment from the previous cut position to the current cut position was concatenated with the previous fragment. This was repeated until the cut position got too close (less than 100 base pairs) to the end position of alignment. In that case, the fragment was extended up to the end position. The information about parental compositions of the derived chimeric sequences was stored for measuring the accuracy of the program. This method was used to generate 300 distinct chimeric sequences. Random abundance counts from 1 to 50 were assigned to each sequence to demonstrate the redundancy of sequences in the chimeric library similar to experimental data. A CSV file was generated containing the chimeric library sequences and their counts.

Increase in the counts of a chimeric variant in the enriched library is a rare event. To illustrate this in the simulated data, the fraction change of variant counts in the enriched library compared to the chimeric library was modeled by a normal distribution with a mean of -1 and a standard deviation of 0.5. The values less than -1 were disregarded as the maximal decrease in the count can be 100% (Fig. 7).

Fig. 8 shows the distribution of variant counts in the enriched library (e) after applying the modeled fraction changes (f) on the chimeric library counts (c) by the following equa-

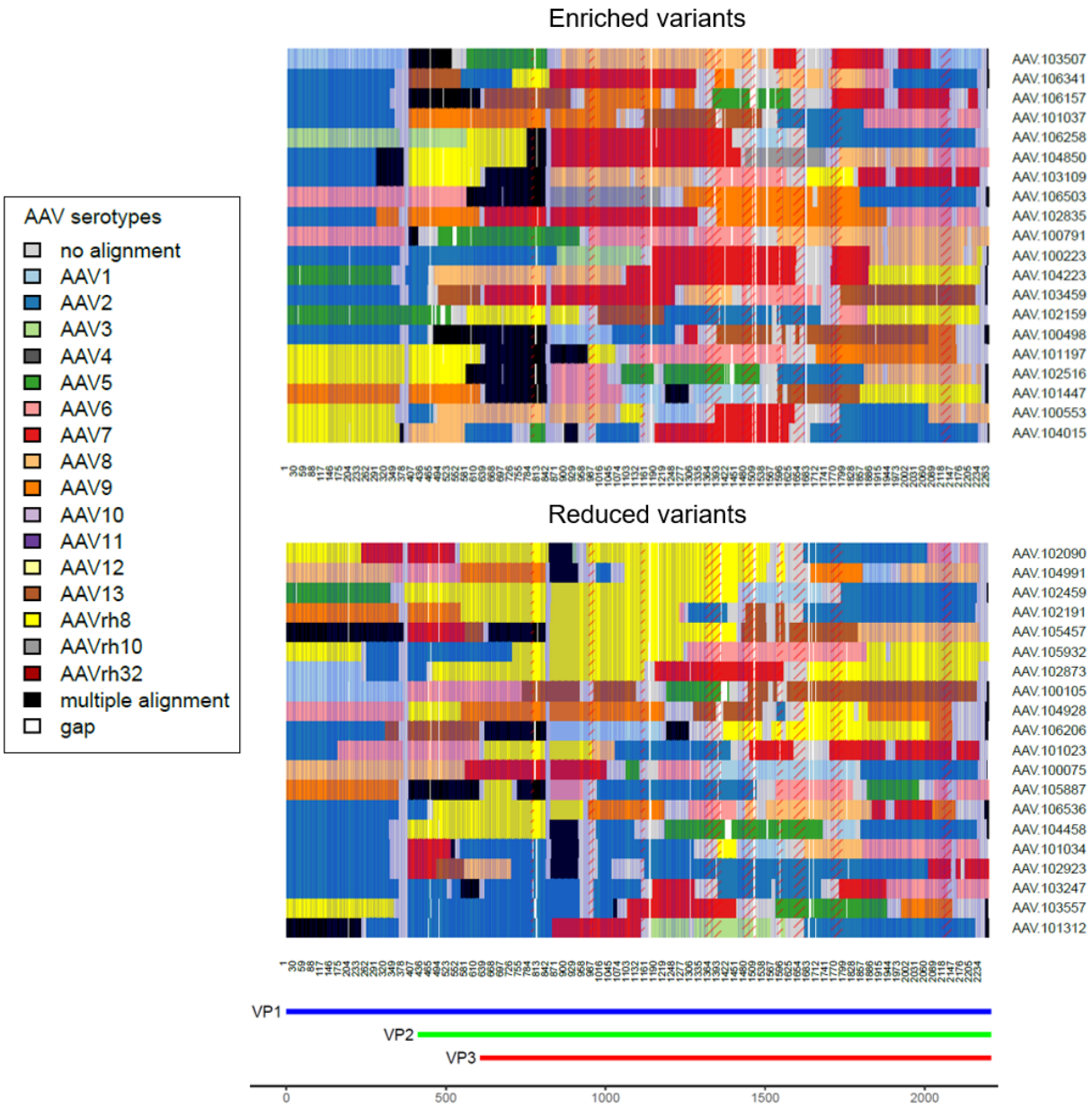


Fig. 6: Variant description of enriched and reduced variants. The gap positions generated by multiple sequence alignment in the variants are shown in white. The shaded regions represent conserved regions obtained from multiple sequence alignment of the variants, the red dashed lines represent positions of AAV2 variable regions [10].

tion: f was chosen randomly with sample function.

$$e = \lfloor (1 + f) * c \rfloor \tag{1}$$

Finally, a fastq file was generated containing the enriched library sequences replicated according to the enriched variant counts. Each base was assigned the highest sequencing quality score corresponding to the ASCII character tilde (˜) for PacBio sequencing.

C. Data preprocessing

Before applying the main methods of the program pipeline, both chimeric and enriched library datasets were filtered by the open reading frame (ORF) boundaries and and original sequence length. The ORFs were identified as the longest sequence regions starting from the start codon ATG, and ending with either of the three stop codons TAA, TAG, TGA in any of the reading frames, both on the original sequence, as well as its reverse complement. The variants having no ORF of size greater than 1.8 kilobases were filtered out. Of the remaining variants, only those with less than 3 kilobases of

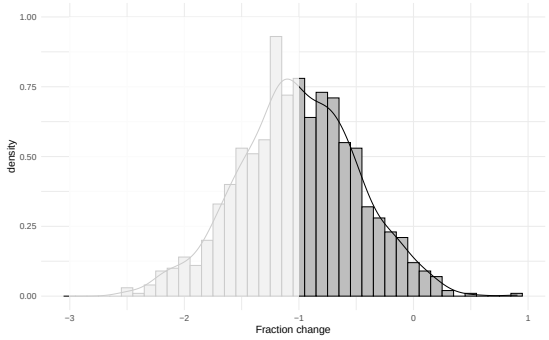


Fig. 7: Distribution of fraction change in variant counts from chimeric library to enriched library.

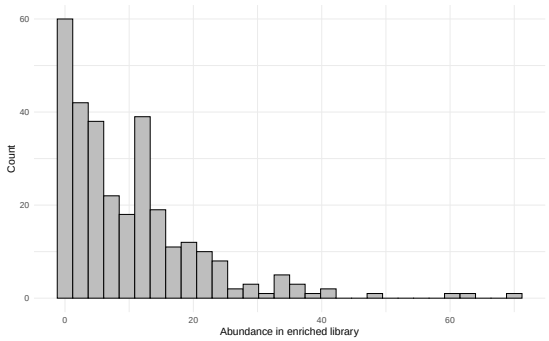


Fig. 8: Distribution of variant counts in enriched library.

the original sequence length were chosen. These thresholds may be adjusted by the user.

The library members were clustered using the CD-HIT program suitable for comparing and clustering nucleotide sequences to reduce sequence redundancy in the resulting chimeric library. Simply put, the program sorts the input sequences by length; the longest becomes the representative, the remaining sequences are compared with the representatives of existing clusters, and if the similarity is above a threshold, the sequence is grouped into that cluster; otherwise, the sequence becomes representative of a new cluster. fter performing the clustering *hafoe* assigns new representatives based on their abundance in the chimeric library.

D. Library size normalization

The percentages of each representative variant both in the chimeric and enriched libraries were used to normalize by the corresponding library size.

The variants with less than 0.5% abundance in the chimeric library were filtered out to avoid variants having a very high ratio because of the division of percentage in the enriched library by low percentage in the chimeric library. This does not lead to loss of information, as in experimental data, the variants with low abundance in the chimeric library most probably have poor packaging abilities, which is the capability to form the capsid and package the capsid DNA in it: a feature required for viable variants.

E. Neighbor-aware serotype identification

Neighbor-aware serotype identification is a method designed to describe the variant sequences and identify the variants' compositions in terms of parental AAV serotypes. For each input sequence it outputs a list of serotype numbers mapping fragments of the variant to the parental AAV serotypes from which they originate.

First, the variants' sequences are chopped into fragments of equal size (read length). The fragments can have overlapping regions if *–overlap* option and step size (the length of the region between starting positions of two consecutive fragments) are specified. The fragments are then aligned on AAV serotype genomes using the Bowtie 2 program. To perform strict alignment, the bowtie2 command is used with increased seed length (-L 30) and a reduced number of seed extension failed attempts (-D 2) options. For each variant, the numbers of serotypes which the fragments align to are stored in a list (Fig. 9).

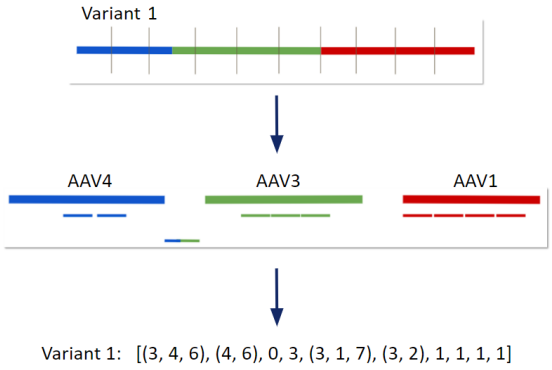


Fig. 9: Schematic representation of variant preprocessing: chopping the variant into fragments, aligning the fragments on AAV serotypes, storing the alignment results in a list.

This list can then be used for neighbor-aware serotype identification, which assigns the most probable serotype to each fragment based on its neighborhood.

The steps are described in Fig. 10: serotypes of neighboring positions are compared, and those that are not common in two neighboring positions are removed; if after this step there are still some positions with multiple serotypes, either a serotype is chosen randomly, and the first elimination step is repeated, or the number 17 is assigned to such positions to indicate that composition of the corresponding fragment is not resolved. In this study, the second option was used to avoid information loss due to random choice.

F. Parameter choice based on accuracy measurements

Neighbor-aware serotype identification was run on the chimeric library data with different combinations of read length and step size parameters. To choose the best option, the accuracy of the method was measured for each read length, and step size pair based on the true composition pattern stored during data generation with the following equation:

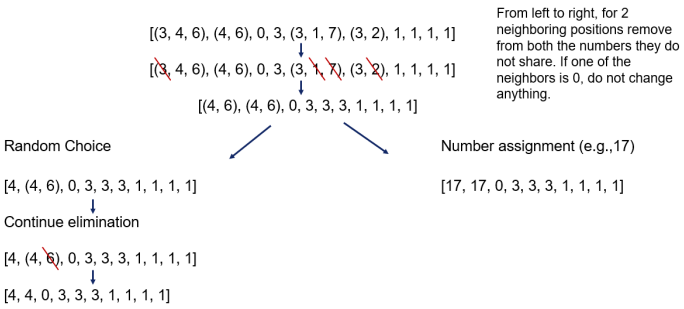


Fig. 10: Neighbor-aware serotype identification method performed on a list with the numbers of serotypes which the fragments aligned to. 1-16 values denote AAV serotypes which the fragments aligned to, 0 shows fragments which did not align to any serotype, 17 shows fragments which aligned to multiple serotypes and were not resolved by the method.

$$Avg.Accuracy = \sum_{i=1}^{n} \frac{correct_i}{len_i} \tag{2}$$

where $correct_i$ is the number of correctly described nucleotides in the i-th sequence, len_i is the length of the i-th sequence, and n is the number of sequences.

Read lengths of 100 nt and 150 nt had, in general, better accuracy. The best combination with read length 100 nt and step size 15 having 79% average accuracy was chosen for the downstream analysis (Fig. 11).

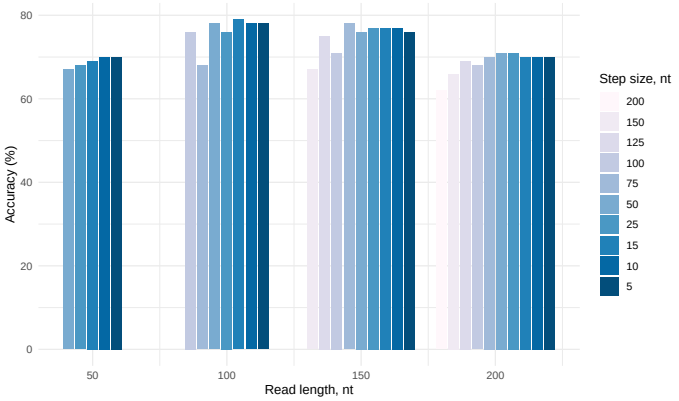


Fig. 11: Accuracy of neighbor-aware serotype identification method for different combinations of read length and step size parameters.

REFERENCES

[1] Naso, M. F., Tomkowicz, B., Perry, W. L., 3rd, Strohl, W. R. (2017). Adeno-Associated Virus (AAV) as a Vector for Gene Therapy. BioDrugs: clinical immunotherapeutics, biopharmaceuticals and gene therapy, 31(4), 317–334. https://doi.org/10.1007/s40259-017-0234-5

[2] Venkatakrishnan, B., Yarbrough, J., Domsic, J., Bennett, A., Bothner, B., Kozyreva, O. G., Samulski, R. J., Muzyczka, N., McKenna, R., Agbandje-McKenna, M. (2013). Structure and dynamics of adeno-associated virus serotype 1 VP1-unique N-terminal domain and its role in capsid trafficking. Journal of virology, 87(9), 4974–4984. https://doi.org/10.1128/JVI.02524-12

[3] Drouin, L. M., Agbandje-McKenna, M. (2013). Adeno-associated virus structural biology as a tool in vector development. Future virology, 8(12), 1183–1199. https://doi.org/10.2217/fvl.13.112

[4] Kienle E, Senís E, Börner K, Niopek D, Wiedtke E, Grosse S, Grimm D. Engineering and evolution of synthetic adeno-associated virus (AAV) gene therapy vectors via DNA family shuffling. J Vis Exp. 2012 Apr 2;(62):3819. doi: 10.3791/3819. PMID: 22491297; PMCID: PMC3460542.

[5] Huang W, Johnston WA, Boden M, Gillam EM. ReX: A suite of computational tools for the design, visualization, and analysis of chimeric protein libraries. Biotechniques. 2016 Feb 1;60(2):91-4. doi: 10.2144/000114381. PMID: 26842355.

[6] Morett, E. and A.G. Garciarrubio. 2004. Shuffled: a software suite that assists the analysis of recombinant products resulting from DNA shuffling. Biotechniques 37:354–358. https://doi.org/10.2144/04373BM03

[7] Schürmann, N., Trabuco, L., Bender, C. et al. Molecular dissection of human Argonaute proteins by DNA shuffling. Nat Struct Mol Biol 20, 818–826 (2013). https://doi.org/10.1038/nsmb.2607

[8] Langmead, B., Salzberg, S. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359 (2012). https://doi.org/10.1038/nmeth.1923

[9] Weizhong Li, Adam Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, BIOINFORMATICS, Volume 22, Issue 13, 1 July 2006, Pages 1658–1659, https://doi.org/10.1093/bioinformatics/btl158

[10] Sievers F, Wilm A, Dineen DG, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology 7:539 doi:10.1038/msb.2011.75

[11] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, S. Turner, Real-time DNA sequencing from single polymerase molecules. Science323, 133–138 (2009).

[12] Meyer, A. J., Ellefson, J. W., Ellington, A. D. (2014). Library generation by gene shuffling. Current protocols in molecular biology, 105, Unit–15.12.. https://doi.org/10.1002/0471142727.mb1512s105

[13] Marsic D, Govindasamy L, Currlin S, Markusic DM, Tseng YS, Herzog RW, Agbandje-McKenna M, Zolotukhin S. Vector design Tour de Force: integrating combinatorial and rational approaches to derive novel adeno-associated virus variants. Mol Ther. 2014 Nov;22(11):1900-9. doi: 10.1038/mt.2014.139. Epub 2014 Jul 22. PMID: 25048217; PMCID: PMC4429732.