Alignment-free characterization of Inverted Terminal Repeats (ITR) in wild-type and recombinant AAV genomes

Alignment-free characterization of Inverted Terminal Repeats (ITR) in wild-type and recombinant AAV genomes

1st Sofia Ostellino
DAUIN
Politecnico di Torino
Turin, Italy
sofia.ostellino@polito.it

2nd Raffaele Fronza
Director of Bioinformatics
ProtaGene GmbH
Heidelberg, Germany
raffaele.fronza@protagene.com

3rd Alfredo Benso
DAUIN
Politecnico di Torino
Turin, Italy
alfredo.benso@polito.it

***Abstract*—This article proposes the development of a novel tool for analyzing the structure and characteristics of recombinant adeno-associated virus (rAAV) vectors during vector production and quality assessment. The tool utilizes dotplots, a graphical method for comparing sequences, and a deep learning-based image classification approach. The focus is on the inverted terminal repeats (ITRs) sequences, which play a critical role in identifying and differentiating AAV types. The tool aims to infer the ITR origin, and improve vector analysis and quality control. The dataset creation process involves generating dotplots of wild-type AAV ITRs and introducing small mutations to simulate biological noise. Future work includes addressing the impact of mutations on vector characteristics to detect major structural anomalies, as well as further analyzing pair-dotplots for vector characterization.**

***Index Terms*—AAV, gene therapy, deep learning, ITR, dotplot, bioinformatics**

I. Introduction

A. Gene Therapy and AAVs

Adeno-associated virus (AAV) vectors have garnered significant attention in recent years as gene delivery platforms. By replacing the *rep* and *cap* genes with a transgene of interest, recombinant AAVs (rAAVs) have emerged and achieved considerable success in gene therapy. Notably, Gylbera was approved by the European Medicine Agency (EMA) in 2012 for the treatment of lipoprotein lipase deficiency, followed by the approval of Luxturna in the United States [1].

Despite their success, the use of AAVs as gene delivery vehicles presents several challenges associated with manufacturing and delivery mechanisms. AAVs are composed of a protein capsid and a single-stranded DNA genome of approximately 4.7 kb. The genome is flanked by two T-shaped inverted terminal repeats (ITRs), which are shared components between wild-type AAVs and rAAVs. The therapeutic gene of interest is encapsulated between these ITRs, replacing the viral genome. There are two types of rAAVs commonly used: single-stranded AAV (ssAAV) and self-complementary AAV (scAAV). ssAAVs are transcriptionally inert upon reaching the nucleus and require conversion to double-stranded DNA through host DNA polymerase activity or strand annealing of the plus and minus strands. In contrast, scAAVs can undergo transcription without the need for additional steps. AAV2 is the most frequently used serotype, while AAV8 is commonly employed in blood disorders, AAV9 in lysosomal storage disorders (LSDs), and AAV1 and AAV9 in neuromuscular disorders [2].

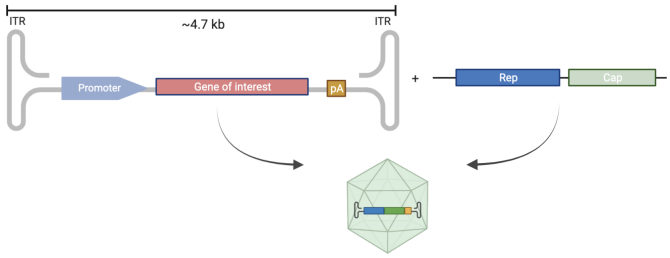

Fig. 1. rAAV production.

B. The ITRs

The inverted terminal repeats (ITRs) are crucial components of AAVs, spanning 145 bases in length. They contain internal sequences that serve various regulatory and priming functions [9], including the Rep binding elements (RBE) and terminal resolution site (trs) [12]. The integration of the engineered AAV genome relies heavily on the ITRs. ITRs are asymmetric and can adopt two different configurations known as flip and flop. In the flip configuration, the B-B' arm is closest to the 3' end, while in the flop configuration, the C-C' arm is closest to the 3' end, following the reference nomenclature [11], as illustrated in Fig.2. By possessing distinct configurations, the ITRs play a critical role in genome packaging and proper viral replication.

1) Challenges: Gene therapy faces several challenges, particularly concerning vector integrity, vector heterogeneity, and vector safety. Ensuring vector integrity is essential and should be assessed during vector production. Vector heterogeneity,

978-1-6654-5992-1/23/$31.00 ©2023 IEEE

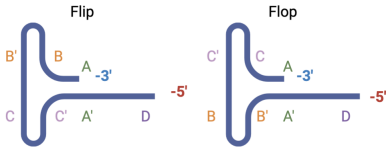

Fig. 2. Flip-flop ITRs configurations.

which is associated with the ITRs, poses a significant challenge, as it can compromise vector functionality and safety.

To address these challenges, quality monitoring in rAAV manufacturing is particularly important. Single molecule real-time (SMRT) sequencing allows for the sequencing of vectors from ITR to ITR, facilitating the characterization of ITRs. Next-generation sequencing (NGS) methods have emerged as a novel technique to quantify and investigate residual DNA in vectors, detect contaminants, and identify production mutations. Recent studies [4] have demonstrated the feasibility of obtaining full-length resolution with NGS, covering ITR-to-ITR without pre-fragmentation. However, NGS has limitations, including longer waiting times for sequencing and data analysis, the ability to quantify ITR heterogeneity [4], [10], and reliance on alignment algorithms.

Vector heterogeneity, influenced by vector design, vector transgene, and the production platform, compromises vector functionality. High heterogeneity necessitates higher vector doses during clinical use, increasing the risk of tissue or cellular toxicity [1]. Assessing the heterogeneity of packaged vectors is challenging due to the limitations of standard sequencing and realignment techniques [6]. Instability of the ITRs not only affects genome stability but also the identification and quantification of species such as truncated forms or chimeric genomes. These can be transduced into cells and pose potential hazards. Investigating the composition of ITRs in packaged genomes can provide insights into ITR integrity, as there is a strong correlation between ITR integrity and vector heterogeneity [3]. Further research is needed to determine the extent to which ITR heterogeneity impacts vector safety and efficacy. Additionally, the presence of mutant ITRs in vector preparations and their impact on transgene expression and stability in cells remains a question for future exploration [3].

C. State-of-the-art

The current state-of-the-art literature reveals a lack of comprehensive discussion regarding the applied knowledge about ITRs in vector production and quality assessment [10]. Safety assessment of AAV vectors is a critical aspect, and the instability and high degree of heterogeneity observed in ITRs, along with mutations, can significantly impact packaging yields [3]. While the direct impact of ITRs on vector production and transgene expression is acknowledged, their function upon entering the cell nucleus remains poorly understood [8]. The presence of inverted repeat sequences in ITRs can lead to genome instability due to their potential to form secondary structures such as hairpins or cruciform structures, which can hinder rAAV genome replication [13].

Noteworthy examples from recent literature include the work by Tran et al. [6] and Zhang et al. [15]. Tran et al. [6] analyzed a vector preparation, differentiating six classes of ITRs (flip, flop, trident structure, unresolved, missing B-arm, missing C-arm). This approach allowed them to identify the percentage of flip and flop configurations in the preparation, ideally in a 1:1 ratio, and visualize the results using coverage-based representations. On the other hand, Zhang et al. [15] conducted a comprehensive analysis of the entire vector and highlighted the presence of four types of non-canonical genome particles alongside the standard rAAV packed genome.

These studies emphasize the importance of analyzing the heterogeneity of the entire vector, thus identifying various ITR types making use of informative result representations, such as dotplots, rather than relying on coverage analysis.

II. Methods

We propose the development of a novel, efficient and user-friendly tool that utilizes an alignment-free method based on dotplots to analyse the structure and characteristics of vectors during vector production and quality assessment. This approach eliminates the reliance on alignment techniques, overcoming the limitations of NGS alignment algorithms.

Our proposed tool relies on a deep learning framework for image classification, trained to analyse dotplots. By applying this framework, the tool will be able to analyze the structure of rAAV vectors, inferring the ITR origin and detect major structural anomalies within a population of vectors, such as missing branches or incomplete ITR structures.

The primary objective of our tool is to provide practical utility during vector production for vector analysis and quality assessment, relying on the straightforward visualization method of dotplots, which has not yet been extensively explored for this type of application. Vectors ITRs play a critical role in identifying and differentiating AAV types. Therefore, our method aligns with the current interests in state-of-the-art literature. The aim of the tool is to enhance the efficiency and effectiveness of vector analysis in the field of bioinformatics.

In our work, we have chosen to utilize dotplots as a graphical method for comparing two sequences and identifying similarities between them. Dotplots provide an intuitive visualization that can immediately reveal important information about the sequences being analyzed. The x-axis and y-axis of the dotplots represent the two sequences, and the graph is filled based on nucleotide matching between the sequences.

After considering several available tools for dotplot creation, we selected Flexidot [5] due to its flexibility and range of functionalities. Flexidot allows us to create both self-dotplots, which involve a single sequence, and pair-dotplots, which involve multiple sequences. Additionally, Flexidot provides the ability to generate dotplots using different K-values. K-values refer to the word size used by the algorithm during the creation of the dotplot. Dotplots with varying levels of detail can be obtained adjusting the K-value, as illustrated in Figure 3.

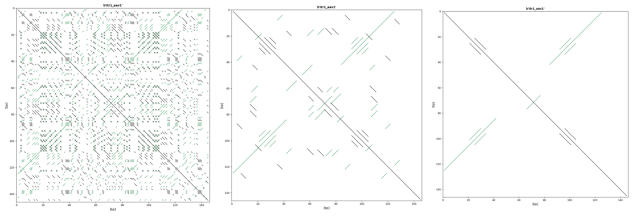

Fig. 3. Example of dotplots of the ITR 3' of AAV1 with different k-values.

This approach will enable efficient and accurate analysis of rAAV vector structures, involving Flexidot and a deep learning framework for image classification trained to recognize dotplots. By applying our tool to dotplots of rAAV vectors, we aim to infer the ITR origin and detect major structural anomalies, such as missing branches or incomplete ITR structures.

Through our work, we strive to contribute a practical and effective tool that enhances vector analysis and quality control in the field of bioinformatics. The combination of dotplots and deep learning-based analysis will enable a comprehensive understanding of rAAV vector structures, improving the efficiency and reliability of vector production processes.

The first implementation steps for creating the infrastructure of the tool are summarized in Figure 4. A preliminary data analysis phase focused on the ITRs of wild-type AAVs, and on the automatization of Flexidot execution. The reference wild-type AAV sequences were downloaded from the NCBI dataset [1], and the ITR sequences where extracted from the following AAV sequences: AAV1, AAV2, AAV3, AAV4, AAV5, AAV6, AAV7, AAV8.

ITRs dotplots were realized starting from these reference sequences, and part of the dataset for the image classification was implemented.

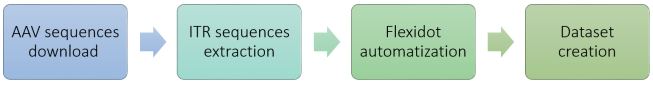

Fig. 4. Implementation steps.

A. Dataset Creation

The first self-dotplots were created for each ITR of every wild-type AAV, using different values of K (2, 3, 5, and 7). The generated dotplots' axes and labels were cropped, and the images were resized to a reference dimension, as shown in Figure 5 and 6. These dotplots were annotated and used as the base for the classifier's dataset.

The resulting dataset, consisting of annotated dotplot images representing the ITRs of wild-type AAVs, forms the foundation for training the deep learning classifier. This dataset will enable the model to learn the characteristics and patterns associated with different ITR origins and structural anomalies, allowing it to accurately classify and analyze rAAV vector structures in future steps of our work.

Fig. 5. Image processing pipeline

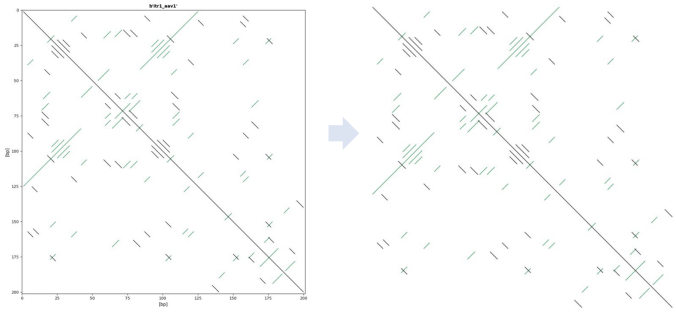

Fig. 6. Example of processed dotplot images

To introduce variability and increase the robustness of the classifier, new ITR sequences were created by introducing small mutations. This step was crucial in increasing the dataset's size and simulating the biological noise typical of real data. The process involved introducing random mutations into each ITR reference sequence. The mutations were chosen based on a uniformly distributed mutation rate and aimed to represent small variations in the nucleotide sequences. Specifically, a random number between 1 and 10 was selected to determine the number of mutations to be introduced. These mutations included deletions, insertions, and substitutions, each involving fewer than 5 nucleotides. Dotplots were generated for each mutated ITR sequence, using the same process described earlier to create dotplot images representing the ITRs with introduced variations. By incorporating these mutated sequences and their corresponding dotplots into the dataset, we introduced additional diversity and biological noise, making the classifier more robust and capable of handling variations and mutations commonly found in rAAV vector structures. The expanded dataset, consisting of annotated dotplot images representing both wild-type and mutated ITRs, provides a comprehensive set of examples for training the deep learning classifier and further improving its accuracy in analyzing rAAV vector structures.

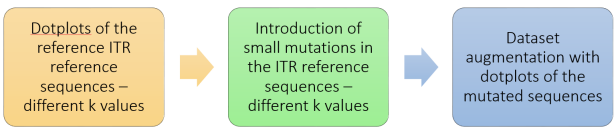

Fig. 7. Dataset creation

The dataset creation process is summarized in Figure 7 and dataset numerosity is reported in Table I.

[1]https://www.ncbi.nlm.nih.gov/genome/viruses/

TABLE I
DATASET DETAILS (EX: INTRODUCTION OF MUTATIONS FOR 32 RUNS ON ALL SEQUENCES).

Dataset numerosity without mutations	65 images
Dataset numerosity with mutations	2113 images

As shown in Figure 8, the effect of introducing small mutations in the ITR sequences can be observed in the dotplot, resulting in visibly altered patterns. The figure illustrates a comparison between the self-dotplot of the 3' ITR of AAV2 (on the left) and the dotplot of the same ITR sequence after the introduction of small mutations, including deletions, insertions, and substitutions. It is important to note that the details of each mutation run were saved separately to accurately track the mutations for each sequence.

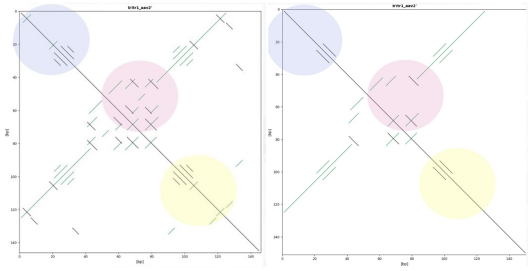

Fig. 8. Example of the effects of small mutations of the dotplots

III. DISCUSSION

The classifier was implemented using Python and the Fastai library [14], which is specifically designed for deep learning applications. The initial dataset consisted of 2113 images that underwent the preprocessing steps including image processing, cropping, and labeling.

The labeling process involved assigning labels to each image to indicate the corresponding ITR (either 3' or 5') and the specific AAV of origin. This labeling scheme allowed for accurate classification and identification of the ITR sequences in the subsequent analysis.

Figure 9 provides an example of the dataset used for training the classifier, illustrating the cropped and labeled dotplot images. These images serve as input to the deep learning model, enabling it to learn and recognize patterns and features associated with different ITR sequences and AAV types.

The implementation of the classifier using the Fastai library facilitates the training and evaluation of the deep learning model, enabling accurate classification of dotplot images and providing a valuable tool for analyzing rAAV vector structures.

The deep learning classifier was implemented as a Resnet 18 network, fine tuned for the task of single-label classification, reaching an accuracy of 80%. The Resnet18 network is a convolutional neural network that consists in 18 deep-layers: we used a pretrained Resnet18 newtwork, as the size of our dataset did not allow a training from scratch.

In future work, we plan to address the challenges associated with ITR heterogeneity and the impact of mutations on vector

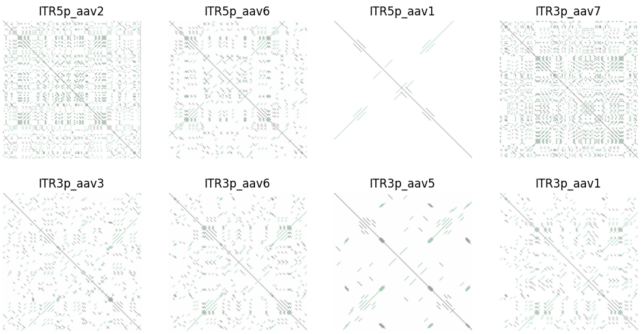

Fig. 9. Example of the dataset for the DL network

characteristics. The use of next-generation sequencing (NGS) for vector characterization is a promising technique, but it has limitations in quantifying ITR heterogeneity. To overcome this challenge, we propose utilizing pair-dotplots between the ITR sequence and the mutated sequence to identify the characteristics of the ITRs in the vector [10] [4] [3]. We also plan to expand the size of the dataset, and to test other deep learning architectures.

The full implementation of our tool will follow these steps (see Figure 10):

1) Input: The tool will take as input a FASTA file containing N reads of the vector.
2) Selection: A subset of M reads will be selected from the input file to create self-dotplots, which compare each sequence against itself.
3) Classification: The self-dotplots will be classified to determine if they represent ITRs. Using a majority voting approach, the tool will identify the viral origin of the vector by detecting the presence of ITR3' and ITR5' and determining the corresponding AAV type.
4) Pair-Dotplot Generation: Pair-dotplots will be generated by comparing the ITR sequences of the vector with the ITR sequences of reference wild-type AAVs.
5) ITR Structure Identification: The pair-dotplots will be analyzed to identify specific ITR structure characteristics, such as mutations, missing branches, or incomplete ITR structures. These insights will help assess the impact of ITR heterogeneity on vector properties and functionality.

Fig. 10. Example of the practical usage of the tool.

By implementing these steps, we aim to enhance the analysis of rAAV vector structures and improve our understanding of the ITR characteristics that affect vector production and quality. Future steps include a detailed analysis of the pair-dotplots for ITRs vector characterization, the fine-tuning of

the classificator, the testing of the tool and the comparison with literature and literature's data.

IV. Acknowledgments

The research work was conducted by Sofia Ostellino while at ProtaGene GmbH (Heidelberg, DE) during an internship.

Funding: This work was partially supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, with particular reference to the partnership on the "Research and innovation on future telecommunications systems and networks, to make Italy more smart (RESTART)" program (PE00000001, CUP: D93C22000910001).

"Digital Driven Diagnostics, prognostics and therapeutics for sustainable Health care" – Acronimo D3 4 HEALTH, Codice programma PNC0000001, CUP B53C22005980001, Avviso n. 931 del 06/06/2022 – Piano nazionale per gli investimenti complementari al Piano nazionale di ripresa e resilienza (PNC), finanziato dal Ministero dell'Università.

V. Author Contributions

Sofia Ostellino performed the experiments/data collection/analysis, interpreted the results, and wrote the manuscript. Raffaele Fronza conceptualized and designed the study, supervised all aspects of the work, and contributed to writing the manuscript. Alfredo Benso supervised the manuscript and the analysis of data.

References

[1] Wang, Dan, et al, "Adeno-associated virus vector as a platform for gene therapy delivery, " Nature reviews. Drug discovery, vol. 18, fasc. 5, maggio 2019, pp. 358–78. PubMed Central, https://doi.org/10.1038/s41573-019-0012-9.

[2] Au, Hau Kiu Edna, et al. "Gene Therapy Advances: A Meta-Analysis of AAV Usage in Clinical Settings". Frontiers in Medicine, vol. 8, 2022. Frontiers, https://www.frontiersin.org/articles/10.3389/fmed.2021.809118.

[3] Tran, Ngoc Tam, et al. "Human and Insect Cell-Produced Recombinant Adeno-Associated Viruses Show Differences in Genome Heterogeneity". Human Gene Therapy, vol. 33, fasc. 7–8, aprile 2022, pp. 371–88. PubMed, https://doi.org/10.1089/hum.2022.050.

[4] Namkung, Suk, et al. "Direct ITR-to-ITR Nanopore Sequencing of AAV Vector Genomes". Human Gene Therapy, vol. 33, fasc. 21–22, novembre 2022, pp. 1187–96. PubMed, https://doi.org/10.1089/hum.2022.143.

[5] Seibt, Kathrin M., et al. "FlexiDot: Highly Customizable, Ambiguity-Aware Dotplots for Visual Sequence Analyses". Bioinformatics (Oxford, England), vol. 34, fasc. 20, ottobre 2018, pp. 3575–77. PubMed, https://doi.org/10.1093/bioinformatics/bty395.

[6] Tran, Ngoc Tam, et al., "AAV-Genome Population Sequencing of Vectors Packaging CRISPR Components Reveals Design-Influenced Heterogeneity". Molecular Therapy. Methods & Clinical Development, vol. 18, settembre 2020, pp. 639–51. PubMed, https://doi.org/10.1016/j.omtm.2020.07.007.

[7] Mahmoud, Medhat, et al., "Structural variant calling: the long and the short of it". Genome Biology, vol. 20, fasc. 1, novembre 2019, p. 246. BioMed Central, https://doi.org/10.1186/s13059-019-1828-7.

[8] Tai, Phillip W. L. "ITRs: The Terminal Frontier," Human Gene Therapy, vol. 31, fasc. 3–4, febbraio 2020, pp. 143–44. PubMed, https://doi.org/10.1089/hum.2020.29108.pwt.

[9] Berns, Kenneth I., "The Unusual Properties of the AAV Inverted Terminal Repeat". Human Gene Therapy, vol. 31, fasc. 9–10, maggio 2020, pp. 518–23. liebertpub.com (Atypon), https://doi.org/10.1089/hum.2020.017.

[10] Wilmott, Patrick, et al. "A User's Guide to the Inverted Terminal Repeats of Adeno-Associated Virus". Human Gene Therapy Methods, vol. 30, fasc. 6, dicembre 2019, pp. 206–13. liebertpub.com (Atypon), https://doi.org/10.1089/hgtb.2019.276.

[11] Lusby, E., et al., "Nucleotide Sequence of the Inverted Terminal Repetition in Adeno-Associated Virus DNA". Journal of Virology, vol. 34, fasc. 2, maggio 1980, pp. 402–09. DOI.org (Crossref), https://doi.org/10.1128/jvi.34.2.402-409.1980.

[12] Pan, Xiufang, et al. "Rational Engineering of a Functional CpG-Free ITR for AAV Gene Therapy, " Gene Therapy, vol. 29, fasc. 6, giugno 2022, pp. 333–45. www.nature.com, https://doi.org/10.1038/s41434-021-00296-0.

[13] Xie, Jun, et al., "Short DNA Hairpins Compromise Recombinant Adeno-Associated Virus Genome Homogeneity, " Molecular Therapy, vol. 25, fasc. 6, giugno 2017, pp. 1363–74. DOI.org (Crossref), https://doi.org/10.1016/j.ymthe.2017.03.028.

[14] Howard, Jeremy et al., https://github.com/fastai/fastai

[15] Zhang, Junping, et al., "Subgenomic Particles in rAAV Vectors Result from DNA Lesion/Break and Non-Homologous End Joining of Vector Genomes,". Molecular Therapy - Nucleic Acids, vol. 29, settembre 2022, pp. 852–61., https://doi.org/10.1016/j.omtn.2022.08.027.