

Direct-ligation nanopore sequencing reveals snapback genome species in rAAV vector preparations

M Wachowius¹, H Martins¹, N Marquardt¹, G Pasha¹, F Leseigneur¹, G Ganjam¹, E Ayuso¹

[1] Siegfried DINAMIQS AG, Schlieren, 8952, Switzerland

Contact: nicoletta.marquardt@dinamiqs.com



INTRODUCTION

Recombinant adeno-associated virus (rAAV) vectors are widely used for in vivo gene delivery due to their efficient transduction, durable transgene expression, and flexible vector design. As rAAV manufacturing advances toward clinical applications, packaged genome integrity has emerged as an important critical quality attribute (CQA). In addition to full-length genomes, rAAV preparations may contain truncated, incomplete, and snapback genome (SBG) species. SBGs are aberrant genome species containing inverted, complementary sequences that can fold back on themselves to form double-stranded hairpin-like structures. Such genome heterogeneity may affect vector quality, potency, and safety, yet remains challenging to characterize using conventional analytical methods. Here, we applied ligation-based nanopore sequencing and computational analysis to characterize genome heterogeneity in an AAV9 vector preparation. Clustering based on shared mismatch patterns, followed by secondary-structure prediction, revealed highly self-complementary sequences that suggest the presence of SBG species. Orthogonal analyses using alkaline gel electrophoresis and duplex digital PCR (dPCR) provided complementary evidence of genome heterogeneity.

RESULTS AND DISCUSSION

1 Vector genome heterogeneity identified by nanopore sequencing and orthogonal analysis

Nanopore sequencing using a ligation-based library preparation identified a dominant read-length peak at approximately 2'712 bp, close to the expected vector genome length of 2'695 bp based on the GOI plasmid reference, alongside a heterogeneous shorter-read population with a local peak at approximately 738 bp. Consistent with genome heterogeneity, alkaline gel electrophoresis showed a prominent band migrating at approximately 3 kb, together with additional faint lower-molecular-weight bands.

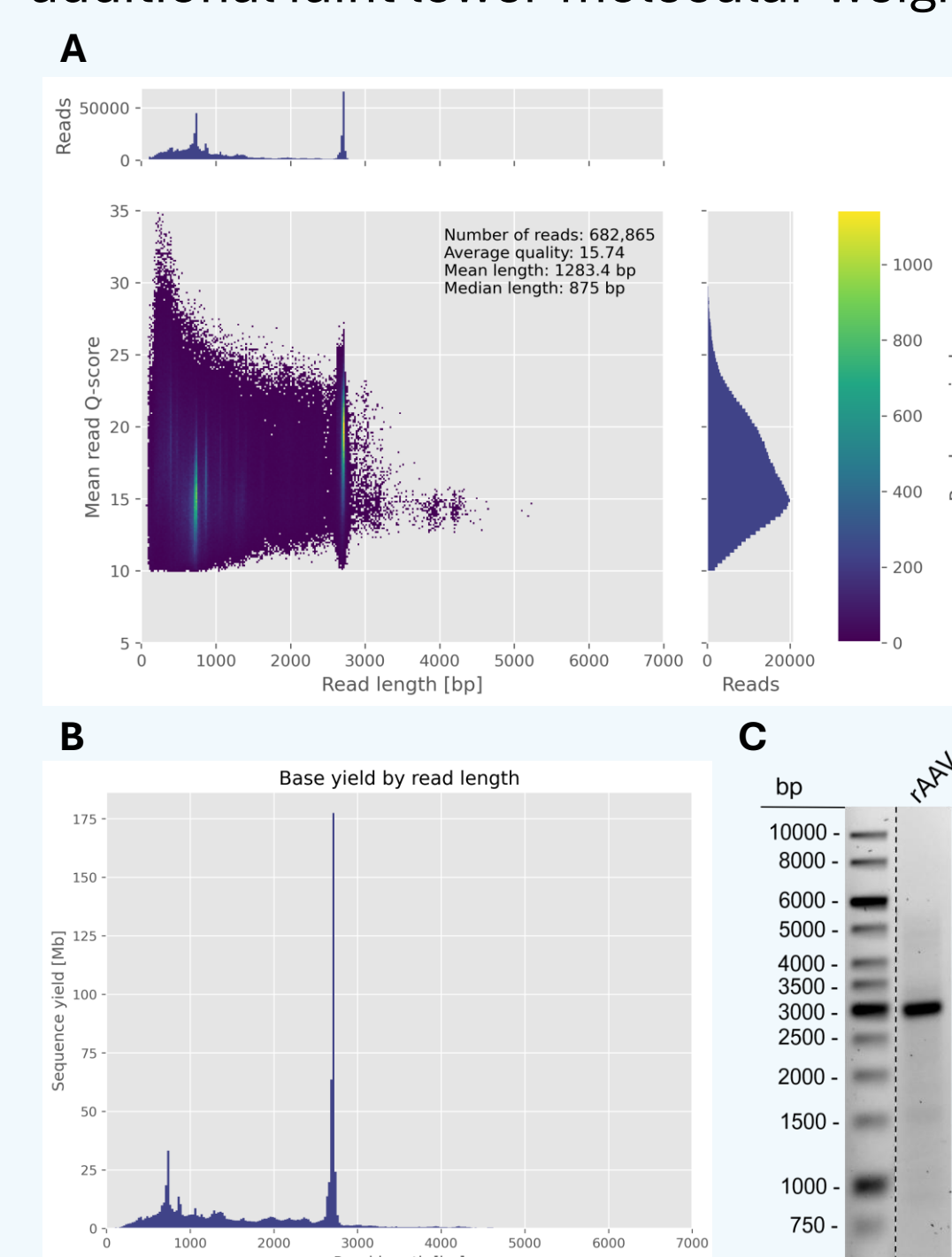


Figure 1. Read-length profiles and orthogonal assessment of packaged DNA size. (A) Read length versus read Q-score and (B) sequence yield versus read length, showing a dominant read-length population at 2,712 bp and a shorter population centred at approximately 738 bp. The dominant peak is close to the expected vector genome length of 2'695 bp based on the GOI plasmid reference. (C) Alkaline gel electrophoresis shows a prominent band migrating slightly below 3 kb, alongside faint lower-molecular-weight DNA species, indicating genome heterogeneity.

2 Duplex dPCR reveals reduced linkage between vector genome targets

Duplex dPCR targeting internal eGFP and terminal ITR2 sequences showed an estimated linkage of 85% in the plasmid control versus 49% in the rAAV sample. ITR2 abundance (~600 copies/ μ L) exceeded eGFP abundance (~400 copies/ μ L), indicating unequal target representation and reduced co-detection. Together, these findings provide complementary evidence of potential vector genome heterogeneity.

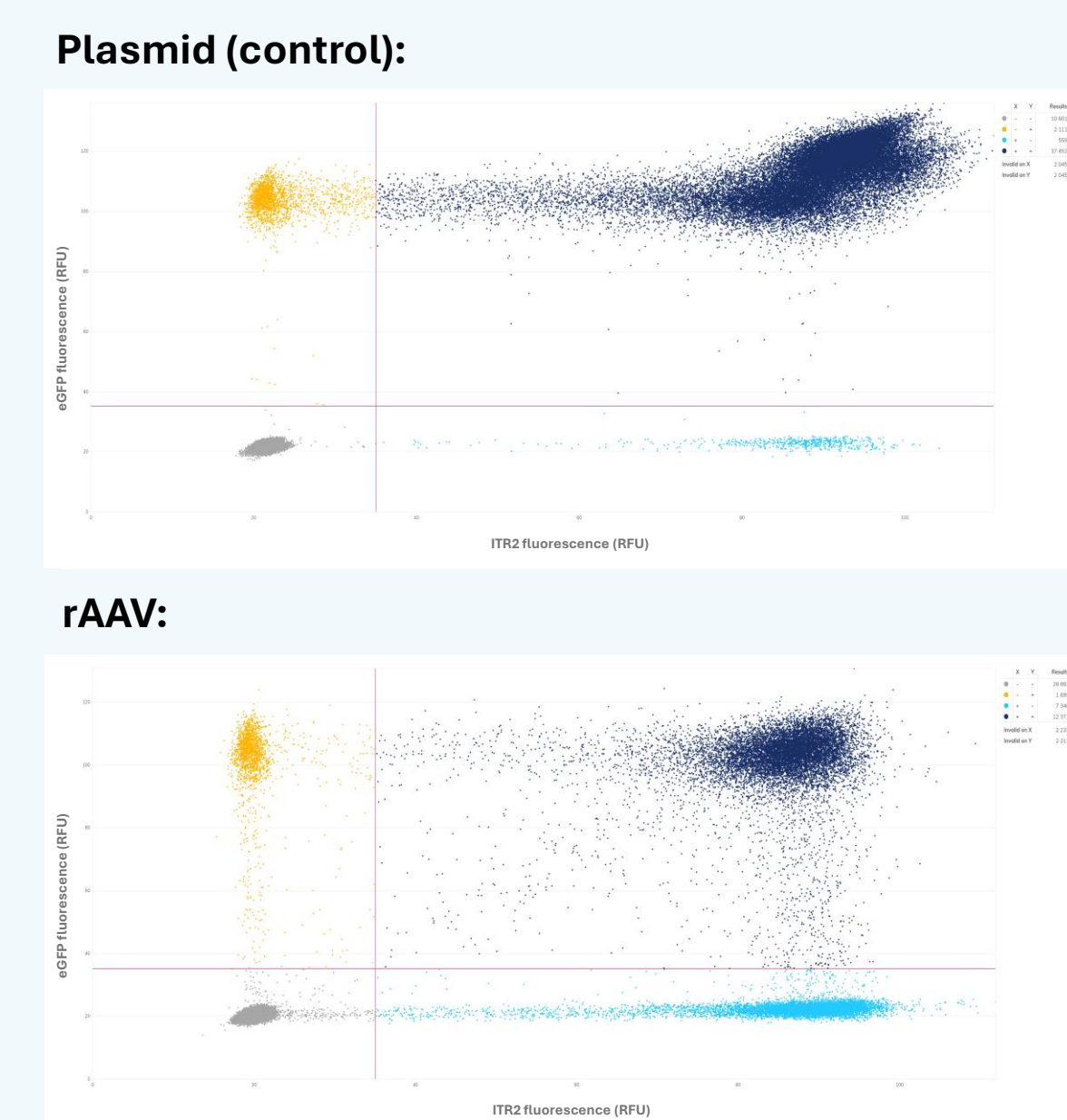


Figure 2. Duplex dPCR analysis of eGFP-ITR2 target linkage using the QIAGEN QIAcuity platform. Two-dimensional scatter plots show a more prominent partition double-positive population in the plasmid control than in the rAAV sample, which exhibits more single-positive partitions and higher ITR2 than eGFP copy numbers, consistent with reduced linkage between the two targets.

METHODS

Ligation-based nanopore sequencing: AAV9 vector DNA was prepared using the Oxford Nanopore Ligation Sequencing workflow with native barcoding (SQK-NBD114.24) and sequenced on an Oxford Nanopore platform. Reads were quality-filtered, aligned to the reference sequence using minimap2, and analysed for read-length distributions, strand-specific coverage, read endpoints, and mismatch patterns. Alignments were visualised in IGV. Reads sharing mismatch profiles were clustered, and cluster consensus sequences were assessed for self-complementarity using ViennaRNA with DNA-specific thermodynamic parameters.

Alkaline gel electrophoresis: Packaged AAV9 DNA was analysed under alkaline denaturing conditions to assess genome size distribution and identify additional lower-molecular-weight DNA species.

Duplex digital PCR: Duplex dPCR targeting internal eGFP and terminal ITR2 sequences was performed using the QIAGEN QIAcuity platform. Target abundance and linkage between the two targets were assessed to investigate vector genome heterogeneity and potential discrepancies between internal target quantification and intact genome representation.

3 Reproducible nanopore coverage patterns reveal distinct subgenomic populations

Strand-specific coverage profiles revealed pronounced drops in sequence coverage, predominantly in the promoter region, accompanied by clustering of read endpoints at defined genomic positions corresponding to the observed coverage transitions (Figure 3). To assess the reproducibility of these patterns across different sample treatments, relative genome coverage was compared between three differently processed rAAV samples (Experiments 1–3). All three samples exhibited similar reductions in coverage within the 5' region, supporting the presence of reproducible subgenomic populations with defined genomic boundaries (Figure 4).

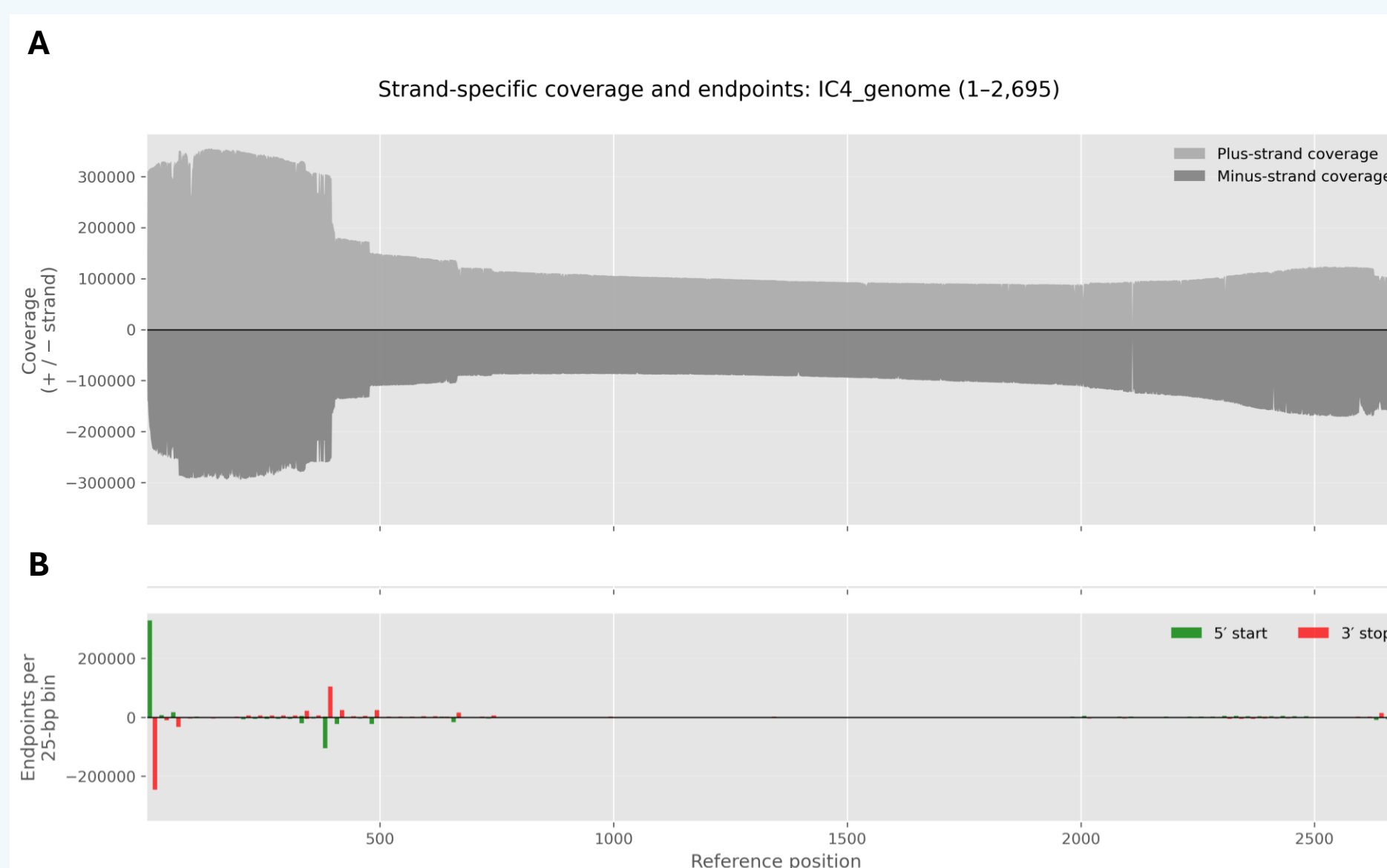


Figure 3. Strand-specific coverage and read start/stop positions indicate distinct subgenomic species. (A) Strand-specific minimap2 alignment coverage across the vector genome shows stepwise drops in coverage, predominantly in the promoter region. (B) Distribution of read start and stop positions across the vector genome shows clustering at defined genomic positions, corresponding to the observed coverage transitions and suggesting the presence of distinct truncated or rearranged subgenomic species.

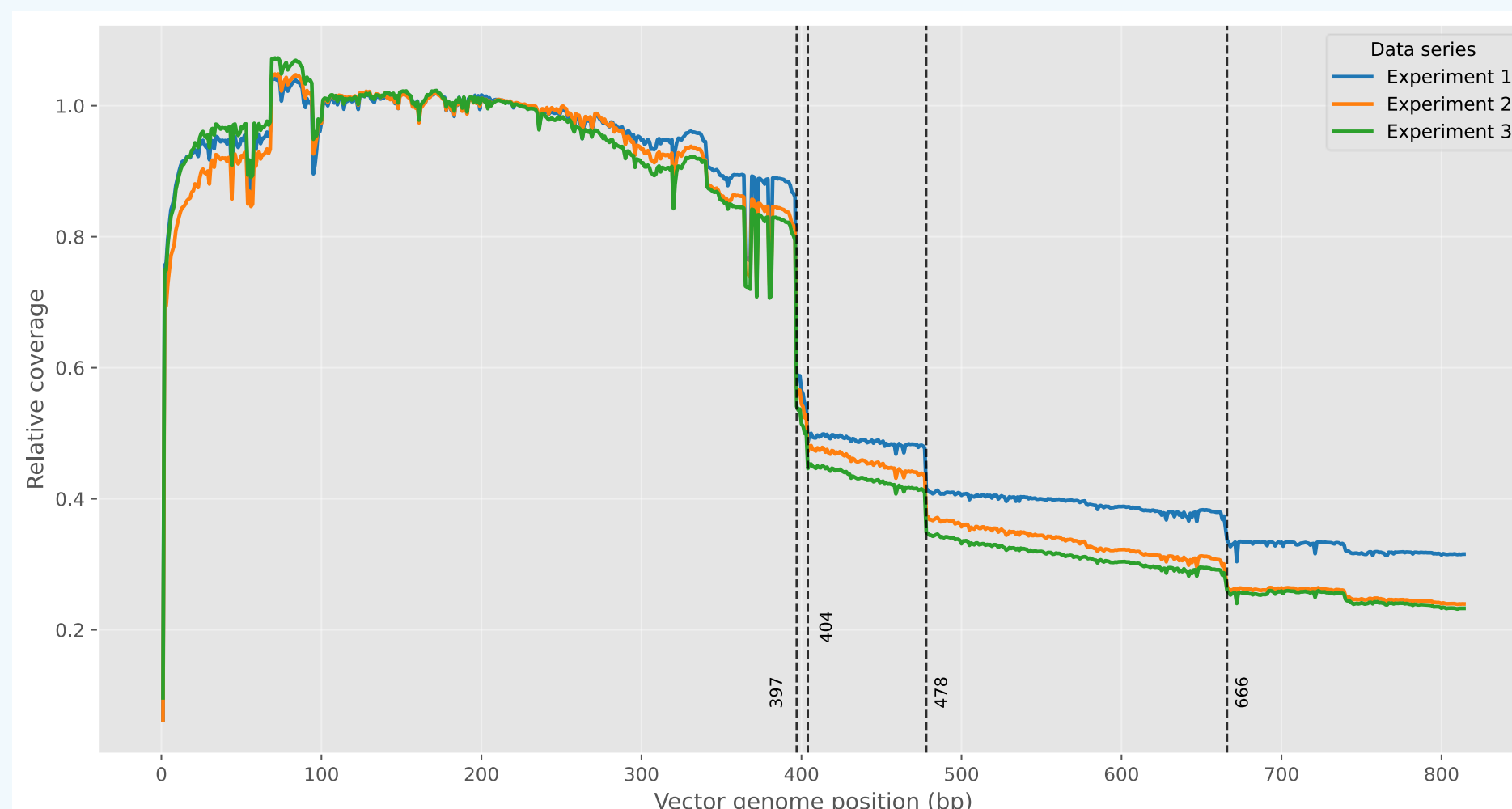


Figure 4. Comparison of relative genome coverage across three differently processed rAAV samples. Relative coverage profiles across positions 0–800 bp of the GOI plasmid reference sequence show similar reductions in coverage, predominantly within the promoter region, across all three experiments. Dotted lines indicate the specific vector genome positions at which pronounced drops in coverage occur, highlighting the reproducibility of these patterns despite differences in sample processing.

The reproducibility of these coverage patterns and their corresponding read endpoint clusters suggested that the subgenomic populations represented recurrent molecular species rather than random sequencing artefacts. This prompted further investigation of their composition and potential structural characteristics.

4 Sequence clustering and structural analysis support the presence of snapback genomes

Alignment analysis identified a subpopulation of reads with high sequence identity to the reference at the 5' end, followed by pronounced mismatches predominantly in the promoter region (Figure 5 B). The recurrence of these patterns across multiple reads suggested the presence of distinct subgenomic populations. To investigate these patterns, reference-aligned reads were clustered based on their mismatch profiles, enabling the generation of cluster-level consensus sequences (Figure 5 C).

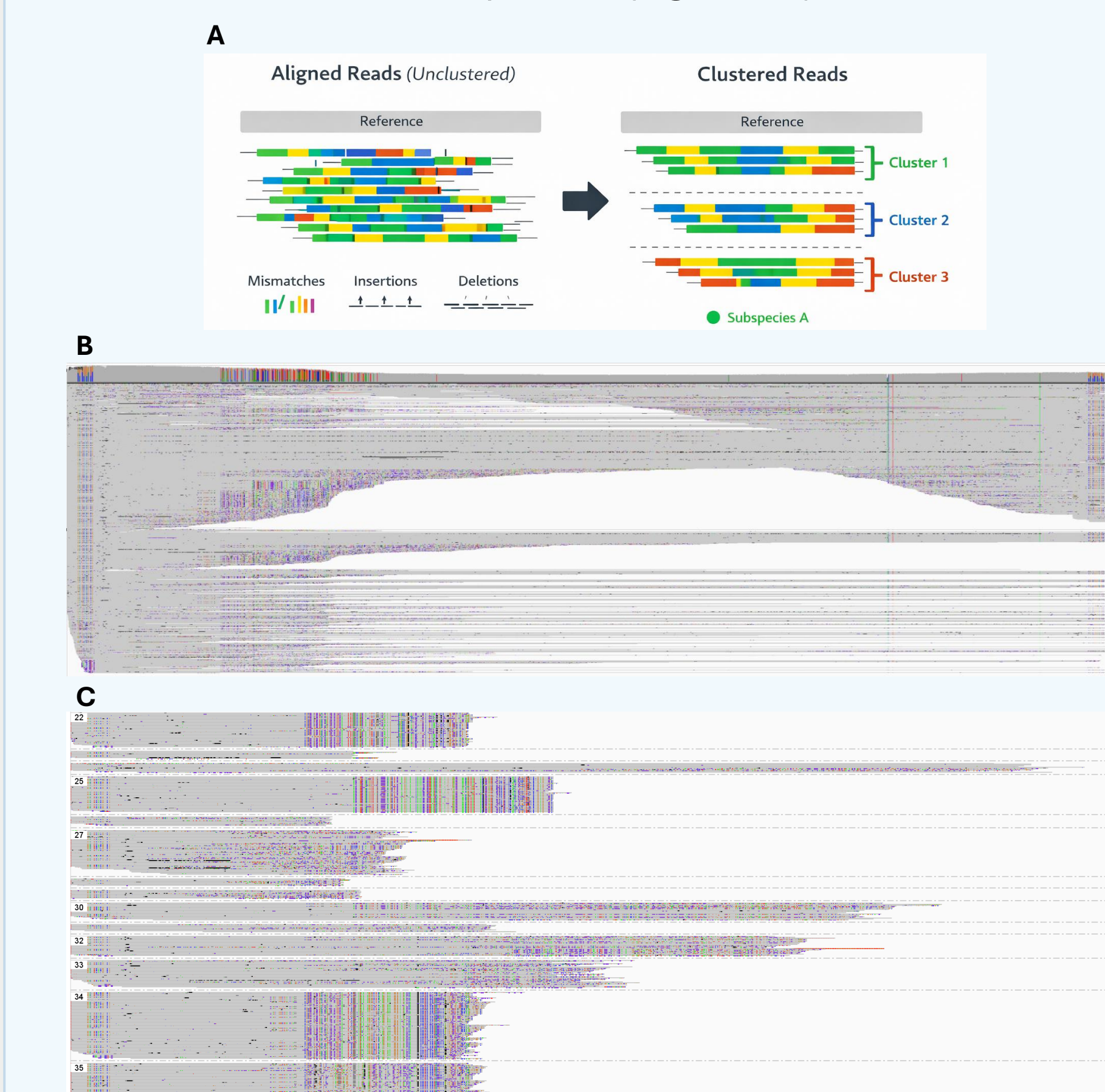


Figure 5. Identification of recurring sequence patterns by read clustering. (A) Schematic workflow of unsupervised clustering of reference-aligned reads based on shared mismatch patterns, followed by generation of cluster-level consensus sequences. (B) IGV alignment reveals a subpopulation of reads with high sequence identity to the reference at the 5' end, followed by pronounced mismatches predominantly in the promoter region. Matching bases are shown in grey, while mismatches are highlighted in colour. (C) Clustering groups reads with similar mismatch profiles, revealing recurrent sequence patterns within the heterogeneous read population.

Secondary-structure prediction of these consensus sequences revealed extensive intramolecular self-complementarity in most clusters exhibiting the characteristic mismatch pattern (Figure 6). These structural features are consistent with the expected architecture of snapback genomes, in which inverted, self-complementary regions may account for the apparent mismatches observed upon alignment to the reference.

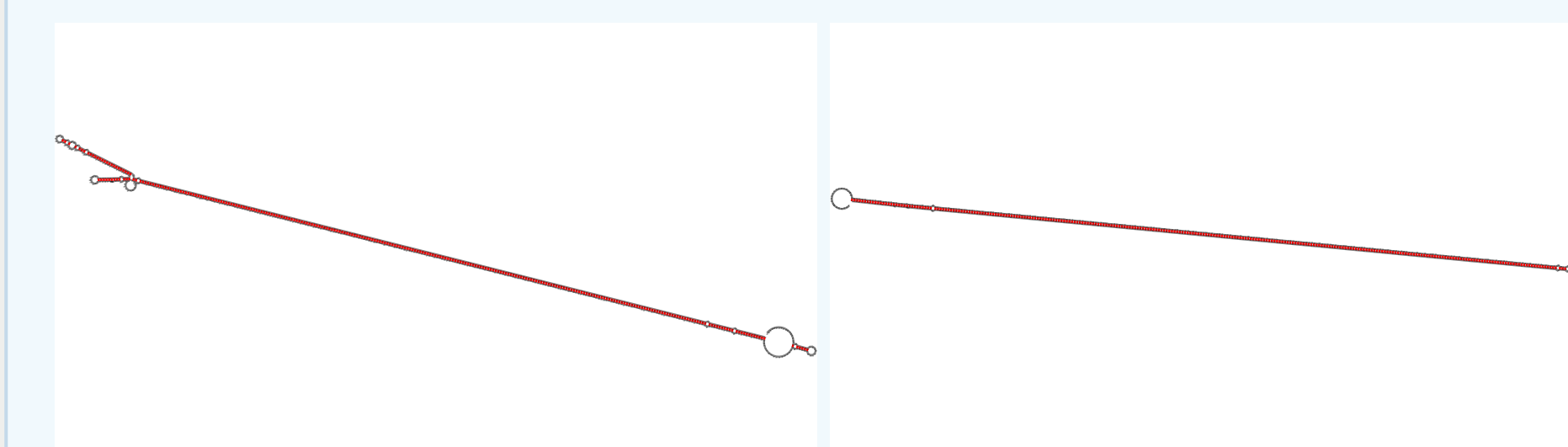


Figure 6. Secondary-structure prediction reveals self-complementary genome species. Predicted secondary structures of representative cluster consensus sequences, generated using ViennaRNA with DNA-specific thermodynamic parameters, reveal extensive intramolecular self-complementarity across multiple clusters, supporting the presence of snapback genome species within the rAAV preparation.

CONCLUSION

Ligation-based nanopore sequencing, combined with computational analysis and unsupervised machine learning, enables high-resolution characterization of rAAV genome heterogeneity and can reveal previously unrecognized genome species.

- rAAV preparations contain heterogeneous genome populations, including full-length and shorter species.
- Orthogonal analyses indicate potential overestimation of intact vector genomes by conventional ITR2-based quantification.
- Direct-ligation sequencing and computational analysis identified reproducible sequence patterns and distinct subgenomic read populations.
- Clustering and structural analysis provided evidence for snapback genome species, demonstrating the potential of this workflow to uncover previously unrecognized vector genome structures.

References

1. Lorenz et al. (2011). ViennaRNA Package 2.0. Algorithms for Molecular Biology, 6, 26.
2. Turner & Mathews (2010). NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic Acids Research, 38, D280–D282.
3. Zhang et al. (2021). Satellite subgenomic particles are key regulators of adeno-associated virus life cycle. Viruses, 13(6), 1185
4. Wang, Tai & Gao (2019). Adeno-associated virus vector as a platform for gene therapy delivery. Nature Reviews Drug Discovery, 18, 358–378.

Acknowledgements

1. This work is funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or HaDEA. Neither the European Union nor the granting authority can be held responsible for them. 2. This work is funded by the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee grant numbers 10080927 (The Francis Crick Institute Limited), 10078726 (King's College London), 10082354 (Muscular Dystrophy Group of Great Britain and Northern Ireland) and 10078461 (University College London). 3. This work has received funding from the Swiss state Secretariat for Education, Research and Innovation (SERI).

Project funded by

