**BENCHMARKING VIRAL INTEGRATION PIPELINE**

Susana Grigson, Suzanne Scott

[susie.grigson@flinders.edu.au](mailto:susie.grigson@flinders.edu.au)

CSIRO, Health and Biosecurity

**INTRODUCTION**

Gene therapies present a potential approach to treating a range of human diseases. Researchers at CMRI have developed a therapy for Ornithine transcarbamylase (OTC) deficiency using an adeno-associated virus (AAV) vector (Ginn, et al., 2020). The safety of AAV gene therapies has been questioned (Li, et al., 2011) after reports of AAV being implicated in hepatocellular carcinoma (Kay, 2007). This could possibly arise from integration of AAV disrupting a tumour suppressor gene or integrating in the promoter of an oncogene. An improved understanding of how, and where these vectors integrate would enable the safety of these potentially life-saving therapies to be resolved.

CSIRO have developed a bioinformatic pipeline to detect viral integrations in a host genome with the aim of investigating the integration of AAV vectors. This pipeline detects viral integrations in next generation sequencing reads (NGS) by aligning reads to viral and host sequences. This pipeline has been used to detect integrations of the OTC vector in FRG mouse and *Macaca fascicularis* (crab-eating macaque). Benchmarking the viral integration pipeline is necessary to determine the accuracy of the output. This can be achieved by assessing the ability of the pipeline to detect known integrations in a synthetic dataset. Furthermore, this process enables debugging and optimisation to improve accuracy of the pipeline.

**METHODS**

**Integration Simulation**

Synthetic data simulating integration of the OTC rAAV vector and wild-type AAV viruses in the human genome was simulated using human reference genome GRCH38. Data was generated using python scripts (<https://bitbucket.csiro.au/users/sco305/repos/intvi_pipeline/browse/benchmarking>). Integrations in chromosomes 20, 14 and X were performed separately. Simulations were completed for different types of integrations: whole, portion, rearrange and deletion, with different types of junctions: gap, whole and overlaps and with simulated viral episomes present. The integrated host sequences were saved as FASTA files and the details on the integrations performed (location in host, type, junction and sequence) recorded.

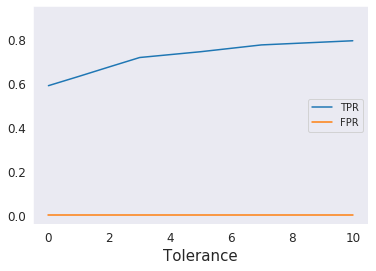
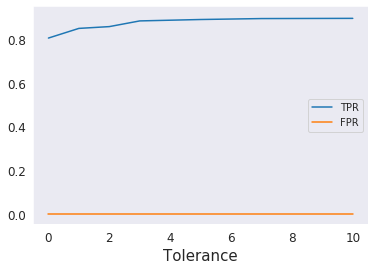
**NGS Simulation**

Paired-end NGS reads were simulated from the integrated host FASTA files using ART illumina (Huang, et al., 2012). Simulated reads had a read length of 150 bp, 20x coverage and a mean fragment size of 400 bp with a standard deviation of 15 bp. Reads containing viral integrations were identified using the output from the integration simulation. The amount of viral DNA and the location in the non-integrated host was also determined for all reads containing integrations.

**Pipeline Analysis**

FASTQ files of the simulated NGS reads were parsed to the viral integration pipeline. The pipeline attempted to identify reads containing viral integrations by aligning all reads to the integrated virus and host. Reads containing at least 20 soft-clipped bases in both the host and viral alignments explained by other alignment were considered to contain integrations. Integrations were considered in discordant read pairs where one read in an unmerged pair is mapped to the virus and the other mapped to the host. Additionally, short insertions were identified as reads matched to the host alignment at both ends with an inserted region between the matched regions. Comparisons were made between reads known to contain viral integrations and reads identified by the pipeline to contain integrations and the number of true positive, true negative, false positive and false negative predictions calculated. This process was repeated for synthetic datasets using different human chromosmes, integration types, integration junctions and viruses. Accuracy, precision and sensitivity was calculated for each test and statistically significant differences determined using one-way ANOVA and Tukey’s honesty significant difference post-hoc test in R. Tests were conducted with different tolerance levels for the number of soft-clipped bases allowed at the ends of integrated regions and with different numbers of viral episomes present in the synthetic data. The true positive rate (TPR) and false positive rate (FPR) was determined for different tolerance levels and different numbers of episomes. Three replicates were included for each set of conditions used to generate synthetic data.

**RESULTS AND DISCUSSION**

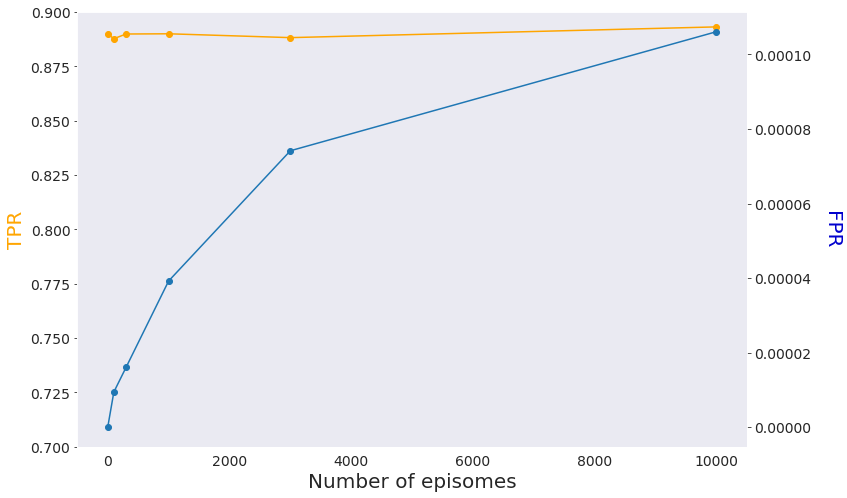


**A**

**B**

**Figure 1**: True positive rate (TPR) and false positive rate (FPR) of **A:** 300 clean whole integrations and **B:** 300 clean short (100bp) integrations of the OTC AAV in chromosome 20 at with tolerance levels of 0, 1, 2, 3, 5, 7 and 10 (n = 3). ART’s default error substitution rate was used.

Initial pipeline results returned few false positive predictions and many false negative predictions. Investigation of these false negatives revealed many of the reads containing integrations missed by the pipeline contained a few soft-clipped at the end of the matched end of the read. To enable detection of these reads, a tolerance was introduced for the number of soft-clipped bases allowed at the end of a read containing an integration. This reduced the number of false negative results for both whole and short, 100bp integrations of the OTC AAV vector (Fig. 1). The true positive rate (TPR) improved with increased tolerance and plateaued at a tolerance of 5 for whole integrations. This increased the TPR of clean whole integrations from 0.804 with 0 tolerance to 0.889 with a tolerance of 5 bp. Similarly, the TPR of short integrations increased from 0.589 with 0 tolerance to 0.744 with a tolerance of 5 bp. A tolerance of 5 bp was introduced into the pipeline. Whilst a higher tolerance could have been used, this would risk increasing the false positive rate with an experimental dataset.



**Figure 2**: True positive rate (TPR) and false positive rate (FPR) of 300 clean whole integrations of the OTC pAAV vector in chromosome 20 with 0, 100, 300, 1000 and 3000 episomes (n = 3).

Inclusion of episomes in the synthetic data increased the false positive rate (FPR) of the pipeline (Fig. 2). As the number of episomes in the synthetic data increased, the FPR increased. Addition of 10000 episomes resulted in a false positive rate of 0.000106 and caused a marginal decrease in true positive rate. Whilst a small FPR may not be of concern, a patient may be administered millions of copies of the OTC AAV vector, resulting in a much higher number of episomes in the nucleus of transduced cells. Experimentalists creating a library for analysis with the pipeline should take precautions to minimise the amount of episomal DNA to reduce the number of episomes sequenced and subsequently detected as integrations by the pipeline.

A screenshot of a cell phone

Description automatically generated

**Figure 3:** Mean accuracy, precision and sensitivity of the viral integration pipeline to detect integrations in synthetic data containing 300 clean whole integrations of the OTC AAV vector in chromosome 20 with clean, gapped and overlapping junctions (n = 3). ART’s default error substitution rate was used. Bars which do not share similar letters denote statistical significance (p < 0.05) using one-way ANOVA with Tukey’s post-hoc.

Detection of integrations with different junction types: clean, gap and overlap were compared (Fig. 3). Accuracy and precision of the pipeline was high for all junction types. No significant difference (p = 0.103) was observed between the sensitivity of gap junctions (0.823) and overlap junctions (0.855). The sensitivity of gap junctions (0.823) was significantly lower than clean and overlap junctions (p < 0.001). As reads with gap junctions contain random DNA, this results in the read being aligned to the incorrect region of the host sequence and failing to be identified as an integration. Regardless, these types of integrations can still be detected with a sensitivity greater than 80%. If a high coverage is used for sequencing, although some integrated reads may missed, it is likely that all integration events would be captured.

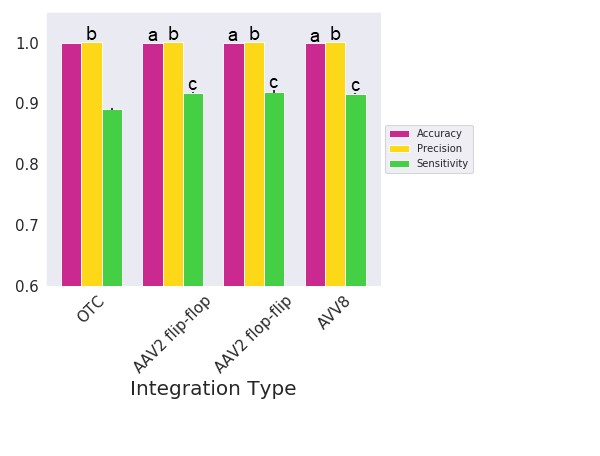
A screenshot of a cell phone

Description automatically generated

**Figure 4:** Mean accuracy, precision and sensitivity of the viral integration pipeline to detect integrations in synthetic data containing 300 whole, portion, rearrange and deletion integrations of the OTC AAV vector in chromosome 20 with clean junctions (n=3). ART’s default error substitution rate was used. Bars which do not share similar letters denote statistical significance (p < 0.05) using a one-way ANOVA with Tukey’s post-hoc.

Synthetic data containing different types of integrations were used to evaluate the pipeline (Fig. 4). High precision and accuracy were observed for whole, portion, rearrange and deletion integrations. Sensitivity was the greatest for whole integrations (0.889). The pipeline was significantly less sensitive to more complex types of integration with sensitivities of 0.838, 0.853 and 0.862 for portion, rearrange and deletion integrations.

AAV genomes, including the OTC AAV vector sequence, exist as strands of single-stranded DNA with inverse terminal repeats (ITRs) at each end (Le Bec and Douar, 2006). When a portion of the viral sequence is taken, rearranged or had a deletion applied, these ITRs may no longer be at the ends of the viral sequence. As the OTC AAV contains human DNA, the junctions between human sequence and complex integrations of the OTC AAV may be difficult to detect. These types of integrations may occur less frequently biologically as the ITRs assist with insertion of the virus into the host sequence (Weitzman, et al., 1994). Therefore, this shortcoming in the detection ability of the pipeline may have little influence on an experimental dataset.



**Figure 5:** Mean accuracy, precision and sensitivity of the viral integration pipeline to detect integrations in synthetic data containing 300 clean whole integrations of the OTC AAV vector, AAV2 flip-flop, AAV2 flop-flip and AAV8 in chromosome 20 (n=3). ART’s default error substitution rate was used. Bars which do not share similar letters denote statistical significance (p < 0.05) using one-way ANOVA with Tukey’s post-hoc.

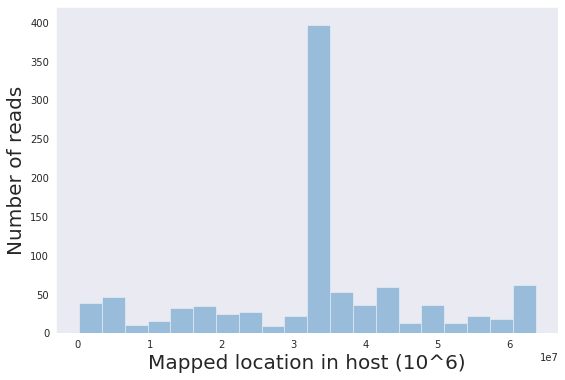
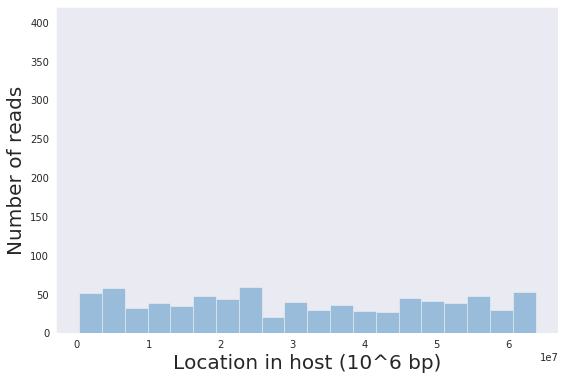
The detection of the OTC AAV vector was compared with wild-type AAV viruses: AAV2 flip-flop, AAV2 flop-flip and AAV8. High precision and accuracy was observed for all AAV viruses (Fig. 5). The sensitivity of the OTC AAV vector (0.889) was significantly lower than AAV2 flip-flop (0.917, p < 0.001), AAV2 flop-flip (0.919, p < 0.001) and AAV8 (0.916, p < 0.001). There was no significant difference in accuracy (p = 0.971), precision (p > 0.999) and sensitivity (p = 0.744) between AAV2 flip-flop and AAV2 flop-flip.

Improved detection capability of wild-type AAV viruses compared to the OTC AAV vector suggests the similarity of the OTC AAV vector to the human sequence makes it more difficult for the pipeline to detect.

A screenshot of a cell phone

Description automatically generated

**Figure 6:** Mean accuracy, precision and sensitivity of the viral integration pipeline to detect integrations in synthetic data containing 300 clean whole integrations of the OTC AAV vector in chromosomes 20, 14 and X (n = 3). ART’s default error substitution rate was used Bars which do not share similar letters denote statistical significance (p < 0.05) using a one-way ANOVA with Tukey’s post-hoc.



**B**

**A**

**Figure 7:** Actual left-most position (**A**) and left-most position when aligned to the host sequence (**B**) of false negative pipeline predictions from synthetic data containing 300 clean whole integrations of the OTC AAV vector in chromosome X.

**Table 1**: Read ID, left-most position (Pos), read sequence (Seq) and CIGAR of 5 false negative pipeline predictions from synthetic data containing 300 clean whole integrations of the OTC AAV vector incorrectly aligned at 31947145 bp on human chromosome X.

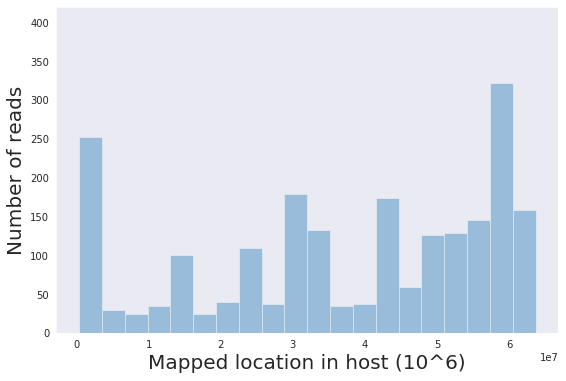
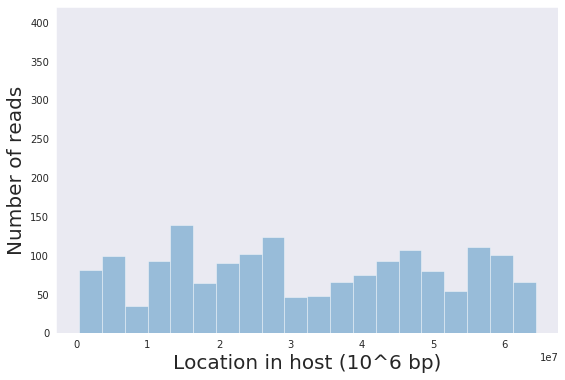
|  |  |  |  |
| --- | --- | --- | --- |
| ID | Pos | Seq | Cigar |
| chr20-2600252 | 31947146 | AACTGAAAGAGTACATGGCTGACTAATTTTTTTTATTTATGCAGAGGCCGAGGCCGCCTCGGCCTC  TGAGCTATTCCAGAAGTAGTGAGGAGGCTTTTTTGGAGGCCTAGGCTTTTGCAAAAAGTTGGCCA  CTCCCTCTCTGCGCGCTCG | 13S19M118S |
| chr20-1871236 | 31947146 | ACATCATGGCTGACTAATTTTTTTTATTTATGCAGAGGCCGAGGCCGCCTCGGCCTCTGAGCTATTC  CAGAAGTAGTGAGGAGGCTTTTTTGGAGGCCTAGGCTTTTGCAAAAAGTTGGCCACTCCCTCTCTG  CGCGCTCGCTCGCTCAC | 4S19M127S |
| chr20-8405998 | 31947146 | TGCAACCTCTCATGGCTGACTAATTTTTTTTATTTATGCAGAGGCCGAGGCCGCCTCGGCCTCTGAG  CTATTCCAGAAGTAGTGAGGAGGCTTTTTTGGAGGCCTAGGCTTTTGCAAAAAGTTGGCCACTCCC  TCTCTGCGCGCTCGCTC | 10S19M121S |
| chr20-1945602 | 31947146 | CTTCATGGCTGACTAATTTTTTTTATTTATGCAGAGGCCGAGGCCGCCTCGGCCTCTGAGCTATTCCA  GAAGTAGTGAGGAGGCTTTTTTGGAGGCCTAGGCTTTTGCAAAAAGTTGGCCACTCCCTCTCTGCGC  GCTCGCTCGCTCACT | 3S19M128S |
| chr20-2032236 | 31947146 | TAGCTGCGTATATTTTGCATGGCTGACTAATTTTTTTTATTTATGCAGAGGCCGAGGCCGCCTCGGCCT  CTGAGCTATTCCAGAAGTAGTGAGGAGGCTTTTTTGGAGGCCTAGGCTTTTGCAAAAAGTTGGCCAC  TCCCTCTCTGCGCG | 17S19M114S |

The OTC AAV vector contains the OTC gene found on human chromosome X and the SERPINA1 promoter on human chromosome 14. Therefore, we suspected the pipeline would be less accurate at detected the OTC AAV vector in chromosome X and 14 than other human chromosomes. The detection ability of the pipeline was compared for chromosomes 20, 14 and X (Fig. 6). High precision and accuracy was observed for all chromosomes tested, indicating a low false-positive rate. The sensitivity of chromosomes 20 (0.899) and 14 (0.903) was greater than the sensitivity of chromosome X (0.616). This demonstrates fewer integrations could be detected on chromosome X than chromosomes 20 and 14. Unexpectedly, the pipeline was more sensitive to integrations on chromosome 14 than chromosome X.

In an attempt to explain false negative results, the mapped location of false negative reads was compared with the actual location of false negative reads from clean whole integrations in chromosome X (Fig. 7). Whilst the actual location of false negative reads was across the chromosome, many of these reads were falsely aligned at 31947144 bp along human chromosome X. These reads falsely mapped to 31947144 contained the same 19 bp of DNA mapped to the host (Table 1). This matched region was identical to the first 19bp of the reverse complement of the OTC AAV vector. The aligner mapped these reads to 31947144 bp in chromosome X rather than to the viral sequence meaning these reads were not detected as containing viral integrations.

**B**

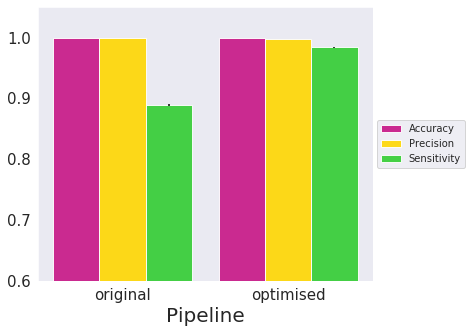
**A**



**Figure 8:** Actual start position (**A**) and start position when aligned to the host sequence (**B**) of false negative pipeline predictions from synthetic data containing 300 clean portion integrations of the OTC AAV vector in chromosome X.

Similarly, false negative reads from clean portion integrations in chromosome X were mapped to incorrect locations in the host sequence. Whilst the false negative reads were located across chromosome X, many of these reads were mapped to incorrect sites on the X chromosome (Fig. 8). These mis-mappings were also caused by short regions in the host sequence, homologous to the viral sequence. This occurred more frequently for portion integrations than whole integrations, as different parts of the viral sequence can appear in the integrated read rather than just the end of the viral sequence.

As an aside, even distribution of the integrated reads in the host indicates the simulation scripts correctly insert integrations in a random manner. The output of the scripts also demonstrates that the stored location of each read in the host is accurate.



**Figure 8:** Mean accuracy, precision and sensitivity of the original and optimised viral integration pipeline to detect integrations in synthetic data containing 300 clean whole integrations of the OTC AAV vector in chromosome 20 (n = 3). ART’s default substitution rate was used. Bars which do not share similar letters denote statistical significance (p < 0.05) using one-way ANOVA with Tukey’s post-hoc.

Analysis of the pipeline BAM files revealed that mis-mapped false negatives containing regions homologous between the virus and host resulted from the pipeline considering alignments which were not the primary alignment. Other bugs were also detected whereby the pipeline checked for one mapped and one unmapped read in discordant read pairs, prior to a later subroutine. Additionally, an issue was resolved where soft-clipped regions were required to meet the cut-off length for an integrations. Issues were also resolved after switching deduplication tools from Clumpify ((<https://sourceforge.net/projects/bbmap>) to deDup (<https://omictools.com/dedup-tool>). This changes significantly improved the accuracy and sensitivity of the pipeline (Figure 8). Whilst the optimised pipeline produces more false positive results than the original pipeline, this enables more integrations to be detected. Regardless, the precision of the optimised pipeline is above 99.8%. The remaining false positive and negative reads result from bases homologous to the virus and host at the site of integration and poor alignments.

Attempts were also made to compare the pipeline to existing viral integration pipelines such as VERSE (Wang et al., 2015). This pipeline is designed for testing whole genomes, thus, data from single chromosomes cannot be used. Significant changes would be required to test the simulated data with VERSE.

**CONCLUSION**

Benchmarking the viral integration pipeline enabled debugging and optimisation to improve the accuracy, precision and sensitivity of the pipeline. Whilst the pipeline detects partial integrations, integrations with rearrangements, integrations with deletions and integrations with gaps less accurately than clean whole integrations, a significant proportion of integrations can be captured. The viral integration pipeline can be effectively used to understand viral integrations in the human genome and to inform the safety of potentially beneficial AAV gene therapies.

**REFERENCES**

Beerenwinkel, N.*, et al.* Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Frontiers in microbiology* 2012;3:329-329.

Chatterjee, N. and Walker, G.C. Mechanisms of DNA damage, repair, and mutagenesis. *Environ Mol Mutagen* 2017;58(5):235-263.

Ginn, S.L.*, et al.* Efficient in vivo editing of OTC-deficient patient-derived primary human hepatocytes. *JHEP Reports* 2020;2(1):100065.

Huang, W.*, et al.* ART: a next-generation sequencing read simulator. *Bioinformatics* 2012;28(4):593-594. *Genome Med.* 2015

Kay, M.A. AAV vectors and tumorigenicity. *Nature biotechnology* 2007;25(10):1111-1113.

Le Bec, C. and Douar, A.M. Gene Therapy Progress and Prospects – Vectorology: design and production of expression cassettes in AAV vectors. *Gene Therapy* 2006;13(10):805-813.

Li, H.*, et al.* Assessing the potential for AAV vector genotoxicity in a murine model. *Blood, The Journal of the American Society of Hematology* 2011;117(12):3311-3319.

Wang, Q. *et al.* VERSE: a novel approach to detect virus integration in host genomes through refererence genome customization.

Wymant, C.*, et al.* Easy and accurate reconstruction of whole HIV genomes from short-read sequence data with shiver. *Virus Evolution* 2018;4(1).

Yamaguchi, S.*, et al.* Mutations and polymorphisms in the human ornithine transcarbamylase (OTC) gene. *Human mutation* 2006;27(7):626-632.