BITS :: Call for Abstracts 2021 - Oral communication

Type	Oral communication
Session	Algorithms for Bioinformatics
Title	rAAVen: a novel bioinformatic pipeline for recombinant AAV integration analysis and viral rearrangement study in vivo
All Authors	Cipriani C(1), Spinozzi G(1), Esposito S(1), Benedicenti F(1), Dalvadi D(2), Grompe M(3), Montini E(1), Cesana D (1), Calabria A(1)

Affiliation

(1) San Raffaele Telethon Institute for Gene Therapy, Ospedale San Raffaele, Milano
(2) Institute for Healthy Aging, Center for Neuroscience Discovery, Department of Pharmacology and Neuroscience, University of North Texas Health Science Center
(3) Department of Molecular and Medical Genetics, Oregon Health Sciences University

Motivation

Recombinant adeno-associated viral (rAAV) vectors are considered promising tools for gene therapy in many applications, such as liver directed gene therapy or including gene editing strategies for Chimeric-Antigen receptor T (CAR-T) based approaches. rAAV is mainly maintained in its episomal form but several studies demonstrated its intra and inter -molecular recombination leading to rearrangements and integration into the host cell genome either at non-homologous sites where DNA damage may have occurred or by homologous recombination. High-throughput technologies allow amplification and sequencing of DNA fragments containing rAAV thus enabling the analysis of rAAV integration and recombinations in vivo, both in preclinical models and in clinical settings. The assessment of safety and efficacy of the treatment become crucial for any drug approval, and in gene therapy is also required to provide clonal studies by analyzing vector integration sites (IS). A bioinformatics tool that can both precisely detect vector integration sites and reconstruct viral rearrangements is still missing. Indeed, we developed a novel bioinformatics tool, rAAVen, to support vector integration, mutation, and recombination studies from both long sequencing reads and short paired-end reads.

Methods

To study and characterize rAAV IS and rearrangement events, we designed a new pipeline for both long and short paired-end sequencing reads, called rAAven. The first steps in rAAven are data quality analysis and filtering (using FastqC and MultiQC), trimming of adapter portions and demultiplexing (using Flexbar). Then passing filter reads are aligned with BWA MEM on AAV genome to identify only reads containing vector sequences and filtering results with Samtools. rAAVen aligns all reads containing AAV on a custom mixed genome composed by both the target genome (such as human hg19) and the vector genome. Aligning the reads to the mixed genome gives us information not only on the primary alignment (which can be both on AAV genome or on the target genome) but also information on the secondary (for short reads) or supplementary alignment (for long reads). Combining alignment results and their respective cigar strings we are able to characterise the different sub pieces of the same read. Parsing the CIGAR string of each alignment, rAAVen recognizes the junction position spacing the vector and cellular genome (labeling as chimera) to identify the IS from chimeric DNA fragments. Similarly, we used CIGAR string to identify any eventual rearrangements and we can also characterize the span between the two genomes (insertions or deletions). CIGAR strings are parsed using a custom script written in Python for the short reads and R for the long reads.
Although long and short reads may present similar components (such as AAV portions), they are designed as independent branches due to the different parameters in place. Those parameters are optimised for the different lengths and custom filters are applied for short reads (such as for repetitive elements). Genomic annotation for the closest gene is run by BedTools whereas the quantification of each IS abundance is performed accounting the number of distinct fragment lengths.

Results

We developed rAAVen as new bioinformatic pipeline and tested its accuracy and recall using simulated datasets. We run simulations on the long reads pipeline obtaining a precision and recall >0.94 on IS identification (with a span of +-3 bp). Simulation studies on rearrangements showed results giving us a precision >0.7 using the most stringent parameters (going up to 0.9 allowing a false positive for read). Similar statistical assessments are ongoing for the short reads. We successfully used rAAVen in a liver directed AAV gene therapy study with PacBio long sequencing reads (range 800-15,000 bp, median 4,500 bp) to study both AAV IS and viral rearrangements in vivo and in vitro. Further applications on studies involving short paired-end reads are ongoing.

With rAAVen (available on GitHub) all laboratories interested on tracking AAV IS and rearrangements will be able to perform clonal trackings studies, safety and long term efficacy analyses, and rearrangement discovery.

Info
-
Figure
-
Availability -

Corresponding Author	
Name, Surname	Carlo, Cipriani
Email	cipriani.carlo@hsr.it
Submitted on	09.05.2021