Short- and Long-read Sequencing for Integration Site Analysis (ISA) of Viral Vectors for Gene Therapy

Andrea O'Hara, Yang Han, Yongjun Fan, David Corney, Chris Mozdzierz, Haythem Latif, Ginger Zhou

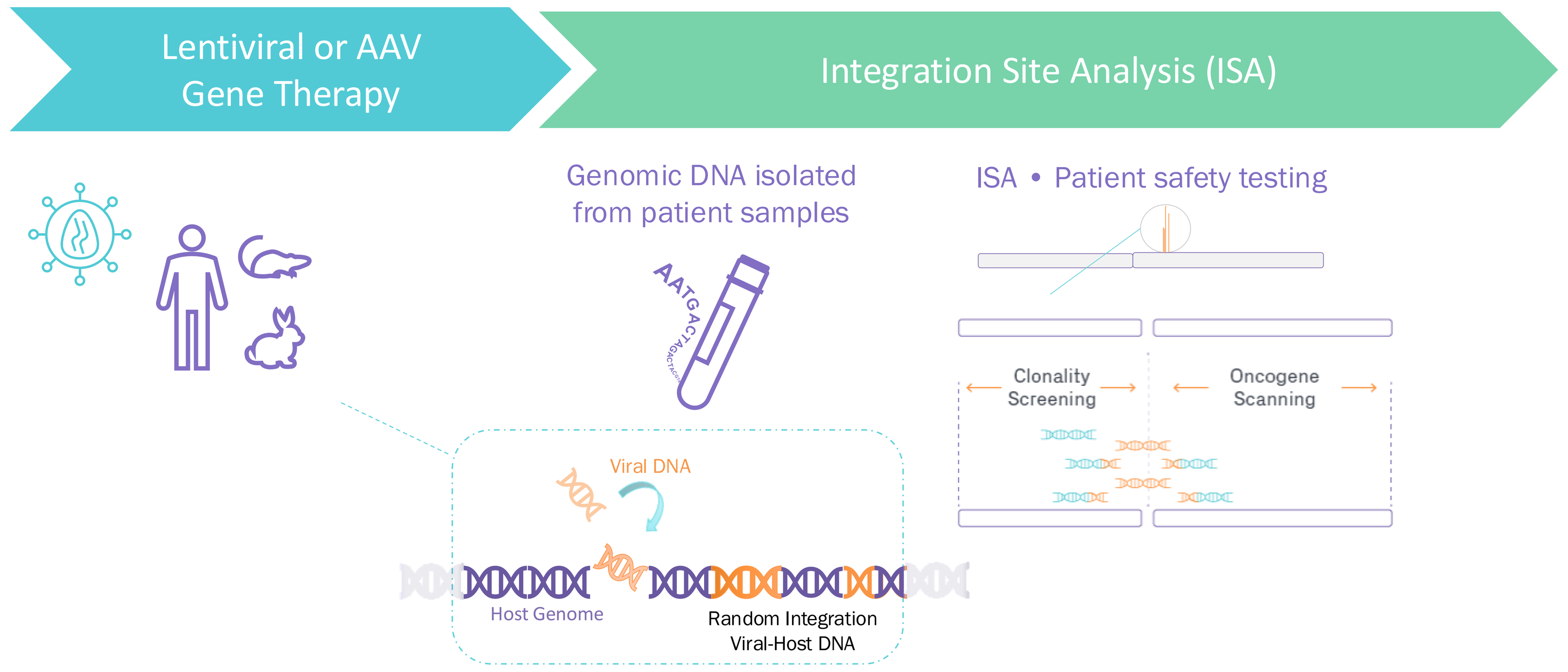
GENEWIZ Multiomics & Synthesis Solutions From Azenta Life Sciences, South Plainfield, NJ

GENEWIZ® From Azenta Life Sciences

Abstract

While lentivirus has historically been the viral vector of choice for gene therapy, recombinant adeno-associated viral (rAAV) vectors have seen widespread application as a gene therapy delivery platform in preclinical and clinical studies. Unlike lentiviral vectors, which are fundamentally integrated into the host genome, recombinant AAV largely remains episomal after transduction into the nucleus of the host cells; however, data has shown AAV may integrate into the host genome in a random, non-homologous manner. The frequency of AAV integration has been estimated at 0.1% to 10% in hepatocytes. In consideration of the theoretical risk of tumorigenesis associated with lentiviral and AAV integration in humans, assessment of viral vector integration in *in vitro* studies, animal models, and patient samples are recommended by the FDA for preclinical and clinical studies and long-term follow-up studies (LTFU). AAV integration analysis presents several challenges, including the random nature of AAV integration sites and low expected integration frequency. To address this, Next Generation Sequencing (NGS) for viral vector integration analysis has been adopted as a highly sensitive assay. Traditional approaches utilize targeted enrichment sequencing with hybrid-capture or linear amplification PCR in combination with short-read sequencing technology. The addition of long-read sequencing offers the advantage of being able to identify chimeric reads with clear sequence structure as genome-insert-genome. Here we describe multiple mixed ISA approaches in lentiviral and AAV studies, with improved sensitivity and specificity for ideal detection of viral vector integration.

Validated Process for ISA Detection



Conclusions

- Gene therapy has extraordinary potential as a therapy for a wide range of diseases and disorders but has potential for off-target insertional mutagenesis; for this reason, regulatory agencies require long-term surveillance.
- Recent ASGCT recommendations suggest risk-based monitoring for clonal expansion; random rAAV integration patterns are expected within some tissue-specific biopsies based on previous animal studies.
- A validated enrichment-based approach enables detection of truncated insertions which may occur with AAV-based therapies but requires higher input amounts and has reduced specificity.
- A validated PCR-based approach confers high sensitivity and specificity, as well as having low input requirements.
- Long-read whole genome sequencing approaches are ideal for determining both location of insertion, as well as fidelity of insertion sequence within the host genome.
- Integration site analysis (ISA) includes insertions and frequencies, clone characterization, oncogene analysis, integration hotspots, longitudinal profiling, and vector integrity.
- It is recommended to consider the relative advantages and disadvantages of each approach during study design.

Results

A. Target enrichment sequencing / hybridization - capture (TES)

Enzymatically fragment gDNA → End-repair, dA-tailing, indexed primer amplification → Hybridization-capture baits → Universal primer amplification → Illumina short-read sequencing → FASTQ data → Data QC and adapter trimming → Align reads to host and vector combined genome → Extract insertion sites → Annotate insertion sites

B. Chromosome-wide distribution of insertion sites

Chr	Position	Strand	Abundance*	Gene	Oncogene(s)**
chr2	32916229	-	4.47	LINC00486	NON_ONCO
chr11	61391254	-	4.39	TMEM216	NON_ONCO
chr7	106101469	-	3.07	SYPL1	NON_ONCO
chr2	32916557	+	2.90	LINC00486	NON_ONCO
chr8	124172004	-	2.87	FER1L6-AS2	NON_ONCO
chr10	31970049	+	2.22	KIF5B	ONCOGENE
chr1	204412048	+	2.00	PPP1R15B	NON_ONCO
chr2	32916234	+	1.98	LINC00486	NON_ONCO
chr2	32916485	+	1.47	LINC00486	NON_ONCO
chr2	222866777	-	1.47	ACSL3	ONCOGENE

*Abundance is the percentage of reads supporting the insertion site over all reads across all insertion sites.
**If the insertion site falls within 100,000 bases up or downstream of an oncogene, the oncogene will be reported in this field.

C. Linear amplification mediated - PCR (qsLAM-PCR)

Mechanically fragment gDNA → End-repair, dA-tailing, primer ligation → Linear amplification with biotin labelled primers based on ROI → 2 round nested PCR with Illumina compatible primers → Illumina short-read sequencing → FASTQ data → Extract UMIs → QC, alignment & deduplication → Extract insertion sites → Annotate insertion sites

D. Top insertion sites identified

Chromosome	Position	Strand	Abundance*	Oncogene(s)**
chr2	25,844,480	+	1.24%	-
chr2	137,787,042	-	1.23%	-
chr16	30,883,571	-	1.19%	-
chr22	31,423,859	+	1.05%	-
chr14	96,836,571	+	0.98%	-
chr12	81,678,426	+	0.98%	-
chr22	30,114,959	+	0.94%	-
chr19	5,630,057	-	0.87%	-
chr12	15,548,426	+	0.87%	-

E. Long-read whole genome sequencing

Mechanically shear gDNA ~10-15kb → End-repair and ligate barcoded adapters → Pool libraries and purify library pool → Anneal primers and bind polymerase → PacBio long-read sequencing → Bam data output → Combine host & insert genome → QC & alignment to combined genome → Extract insertion sites → Annotate insertion sites

F. Insertion site identification results

A. Schematic of TES-validated wet lab/bioinformatics workflows for ISA. Following initial viral vector transduction, samples (typically whole blood and/or enriched populations) are collected and transported to Azenta for genomic DNA isolation and ISA testing. Several ISA workflows were developed and benchmarked. Following library preparation and sequencing, the resulting data was processed through a pipeline to annotate the breakpoint/integration site based on proximity to known oncogenes. The target capture/enrichment sequencing (TES)-based approach uses baits to target a region of interest (ROI) of the designed gene therapy (LTR and/or transgene/insert), starting with a traditional whole genome sequencing (WGS) library. This approach does not require intact ends, thus enabling detection of truncated insertions which may occur with AAV-based therapies. **B. Example TES results.** Results from an animal sample treated with AAV. Genome map of all insertion sites. Green areas in the plot show insertion site locations on chromosomes. Black segments are stretches of the genome where no insertion sites were observed. Top 10 insertion sites identified with abundance shown as the percentage of reads supporting the insertion site over all reads across all insertion sites. Oncogenes within 100kb of insertion site are indicated. **C. Schematic of Quantitative Shearing Linear Amplification Mediated - PCR (qsLAM-PCR) validated workflow.** qsLAM-PCR targets the LTR/ITR, followed by 2 rounds of nested PCR for high sensitivity. **D. Example LAM-PCR results.** Top insertion sites identified by LAM-PCR in lentiviral (LV)-treated sample. **E. Schematic of long-read whole genome sequencing validated workflow for ISA.** Traditional long-read library preparation is followed by a custom bioinformatics workflow to identify the insertion sites based on high-confidence integrated reads. **F. Example WGS ISA results.** Breakpoint of host genome shown; full-length insert was confirmed with high confidence at the breakpoint.