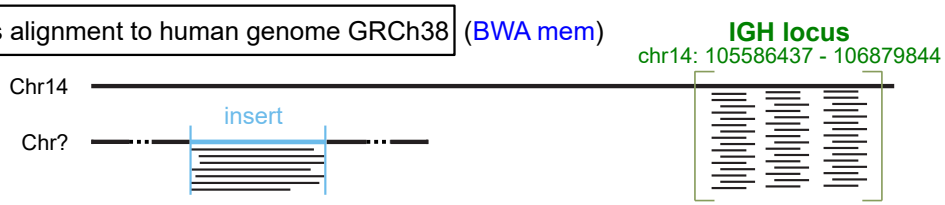
FASTQ (raw) = targeted amplicon reads (300 bp PE)

1. Quality-based trimming and adapter removal. Remove reads < 100 bp (Trim Galore, FastQC)

FASTQ (trimmed)

2. Reads alignment to human genome GRCh38 (BWA mem)

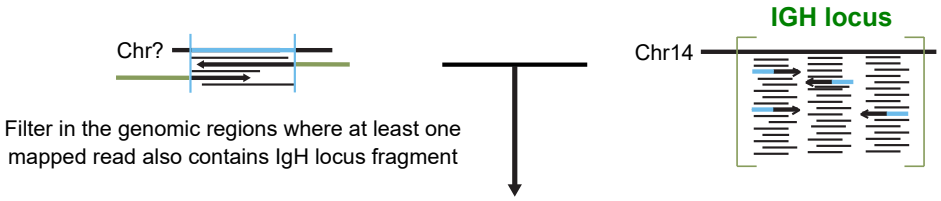


BAM

3. Select genomic ranges where coverage ⩾ 10 reads and 20 bp ⩽ length < 2000 bp (bedtools)

4. Annotate inserts with GENCODE v29 (Java, BEDOPS)

5. *De novo* assembly for each potential insert (samtools, pysam)



FASTQ

De novo assembly (Trinity)

Minimum contig length = 150 bp

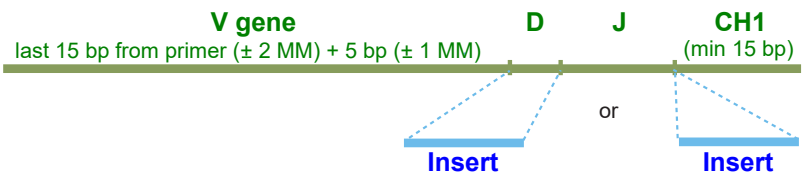
FASTA

6. Validate the contig sequence

BLAST all inserts against each contig
- identify insert coordinates within the contig (min length = 200 bp)
- identify "fused" inserts
- remove insert(s) from contig sequence

Identify V,D,J and CH1 genes (IgBLAST, Java) and V-gene **ORF** to translate into protein

Validate contig where the insert is between V gene and CH1 (Java)



+ Add information about putative functional domain and/or exon-exon junction
UniProt, JAGuaR, BLAST, Java

Annotated contigs