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

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

FASTQ (trimmed)

b. Align reads to human genome GRCh37 (BWA mem)

```
Chr14
insert
Chr?
```

(c) Select genomic ranges where coverage >= 2 (or 40) reads and 50bp <= length < 2000bp

(d) Identify potential inserts like the following: (pysam)

```
Chr14
insert
```

One read in 3' is chimaeric with Switch region

One read in 5' is chimaeric with Switch region

(e) Merge inserts coming from 2 and 40 reads coverage and annotate them with GENCODE v19

For overlapping inserts, if difference between 2 inserts <= 10bp, we keep the shortest one, otherwise we keep the longest one.

(f) De novo assembly for each potential insert (samtools, pysam)

```
Chr?
```

Filter in paired reads containing the insert (chimaeric or non-chimaeric ones): encompassing mate pairs and spanning mate pairs

```
Chr14
```

Filter in the spanning mate pairs in the Switch region (proper paired with a chimaeric read)

De novo assembly (Trinity)

Validate insert (BLAST, Java)

- insert sequence not homologous with Switch region
- insert sequence complete in the contig
- two flanking regions of minimum 50bp that match with Switch
  - select the shortest contig sequence that fulfills the criteria

switch region insert switch region