Data analysis for identification of insertion sites in host genomes

Background

Several genomic applications need to deal with the integration of exogenous DNA into the host genomes:

- Generation of transgenic organisms, i.e. (random) integration of transgenes
  - E.g. transgenic plants generated by DNA transfer methods
- Viral integration into the host
  - E.g. insertions by virial vectors for gene therapy in human cells
- Integration of genes or constructs using homologous recombination, CRISPR or other gene editing techniques

In some cases, it is unknown where the exogenous DNA has integrated into the host genome, whereas in others you might know the region you would like to validate. Using Samplix Xdrop® enrichment technology it is possible to design a Detection Sequence within the insert or next to the expected insertion site to enrich for long fragments spanning the insert. Using Xdrop you will enrich for all those DNA fragments containing the detection sequence, hence the insert and the genomic context will be retrieved. Using long read sequencing (e.g. PacBio or Oxford Nanopore), it is then possible to obtain reads that matches the insert and run into the genomic location of the inserted construct.

Aim

In this document we describe the procedure for identifying the insertion sites in host genomes using Xdrop enrichment and Oxford nanopore long read sequencing.
In brief, we identify those reads that match the insert or constructs, and then check where these reads map to the genome, which will reveal the border between the insert and host genome.

Procedure

We recommend to start out masking any region homologous to the construct in the host genome. First map the construct to the genome with *Minimap2* and create a bed file corresponding to the mapping regions using for example *bedtools bamtobed*. After, masking those regions using *bedtools maskfasta* create a new reference genome, by adding the construct as an extra chromosome.

We recommend starting by mapping all reads to the new masked genome containing the construct used to generate the transgenic cells or organisms using *Minimap2* with default settings for oxford nanopore reads (-ax map-ont) (*1*). One can also be less strict in the analysis and instead map to the construct alone. Next, check the coverage profile for the entire construct for instance in a sequence viewer (e.g. IGV) (*2*). This way you will be able to visualize which parts of the construct have inserted into the genome.
Afterwards, extract both primary and supplementary mapping reads that map to the construct for example using *samtools view,* specifying which reads to extract by providing the construct coordinates (*3*), *and seqkit grep.* These will be the reads of interest for finding insertions, and it is important to both get primary and supplementary mapped reads if using the strict approach.

Once the reference is ready and the reads from the construct extracted, use *Minimap2* to map the reads from the construct to the masked genome which includes the construct.
We recommend toadd a TAG containing the read name to the bam file, this can be done using *samtools view* and *samtools index*.
Next use *bedtools genomecov* to generate of list of areas with coverage in the bam file. Using bedtools merge the regions can be collapsed and a coverage from above one may be usefull.
You can use a sequence viewer (e.g. IGV) to look at potential insertion borders in the regions found. Open the bam file in IGV, activate the Region Navigator (from the Regions top menu) and select the regions of interest: the construct and the insertion site. In this way, you will be able to visualize the insert site and construct at the same time and color and group the reads based on the TAG (right click and select color reads). This will show how reads span from the genome into the construct.

In case the analysis is done on samples containing several insertions, it can be challenging to find reads for one specific insertion. In this case, we recommend to extract both primary and supplementary mapping reads to a potential insertion sites in the genome and map those reads back to the genome with the construct. Next, one can easily find reads that span genome and construct borders, when visualizing the bam file in IGV.

Once the insert has been located in the genome, it is possible to reconstruct a new host genome including the insert at the position(s) identified and then remap all the reads to the reconstructed genome using *Minimap2*. When visualizing the result with IGV, it should provide a good overview of the reads spanning both the insert left and right borders and into the host genome. In case you already know the exact position of the insert in the genome this may be your starting point for analysis.

Further validation of the insertion site may be achieved by designing PCR assays amplifying across the borders between the construct and the host genome. The PCR product and size of this will validate the insertion site, and can be sequenced (e.g. Sanger) to further validate the sequence around the border.

Note: It is important to know whether the organism is a clone or a heterozygous sample as for the latter insertions will appear more rare and have less reads spanning the borders with the host genome. This will make finding the location of the insertion more challenging.

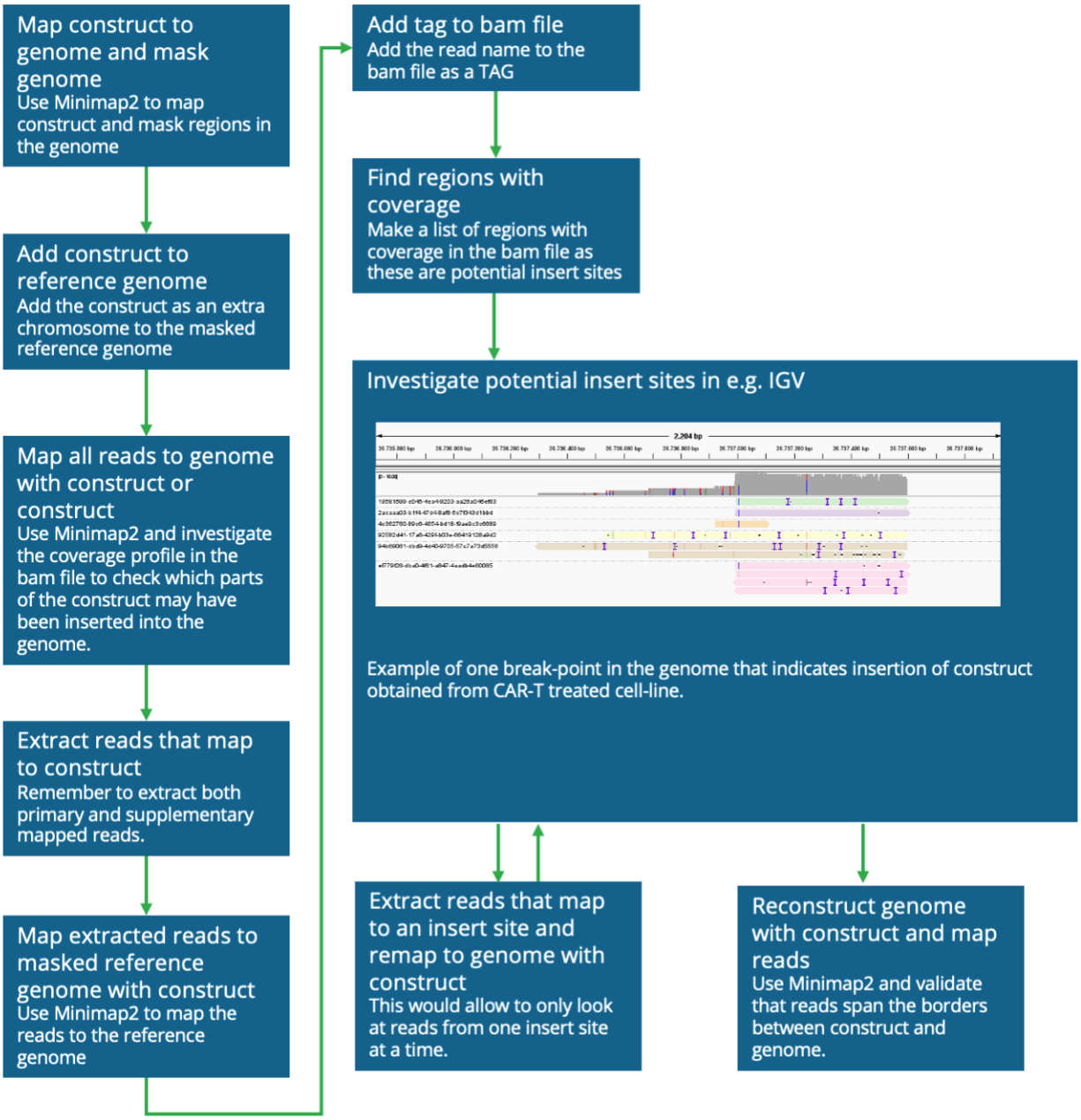

Fig. 1. The workflow for identification of insert sites in genomes after Xdrop enrichment and Oxford Nanopore Long read sequencing.