SAHLGRENSKA ACADEMY
CORE FACILITIES

UNIVERSITY OF GOTHENBURG

TC-Hunter: Identifying insertion sites of a transgenic construct within its host

Marcela Davila[1], Jelena Milosevic[2], John Inge Johnsen[2] and Per Kogner[2]

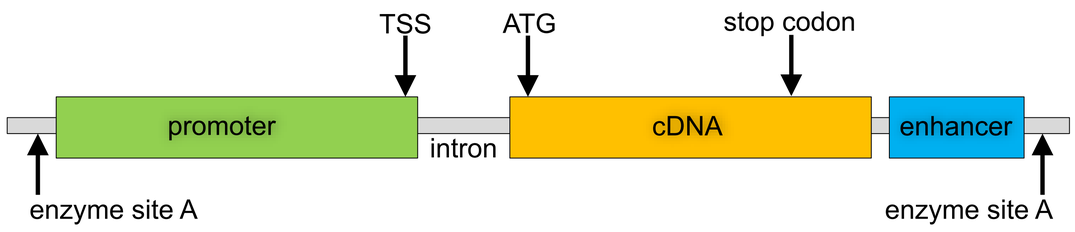
[1]Bioinformatics Core Facility, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

[2]Department of Women's and Children's Health, Karolinska Institute, Stockholm, Sweden

Introduction

Genetically manipulated animal models are considered essential for studying gene functions in whole animals.

Typically a construct, containing critical elements for gene expression (such as promoters, introns, the protein coding sequence of interest and a poly-A site), is microinjected into a host. It is common to evaluate the presence of the transgene by using Polymerase Chain Reaction (PCR) and Southern blotting.



Today, thanks to whole genome sequencing data we are now capable to identify the exact insertion site of the construct in the host (see references).

Here we present **TC-Hunter** (Transgenic Construct Hunter tool), an algorithm that takes aligned pair-end data and extracts candidate positions where the transgenic construct may have been incorporated.

The method

TC-Hunter accepts alignment files of massive parallel sequencing. Anchors[1] and chimeric[2] reads are identified and combined to identify common regions that are plotted for manual inspection. The location of the precise insertion site is identified using the chimeric reads in a second alignment round. An html report harboring candidates and graphical visualizations is also created.

[1] Read pairs where one read aligns to the host and the other aligns to the construct.

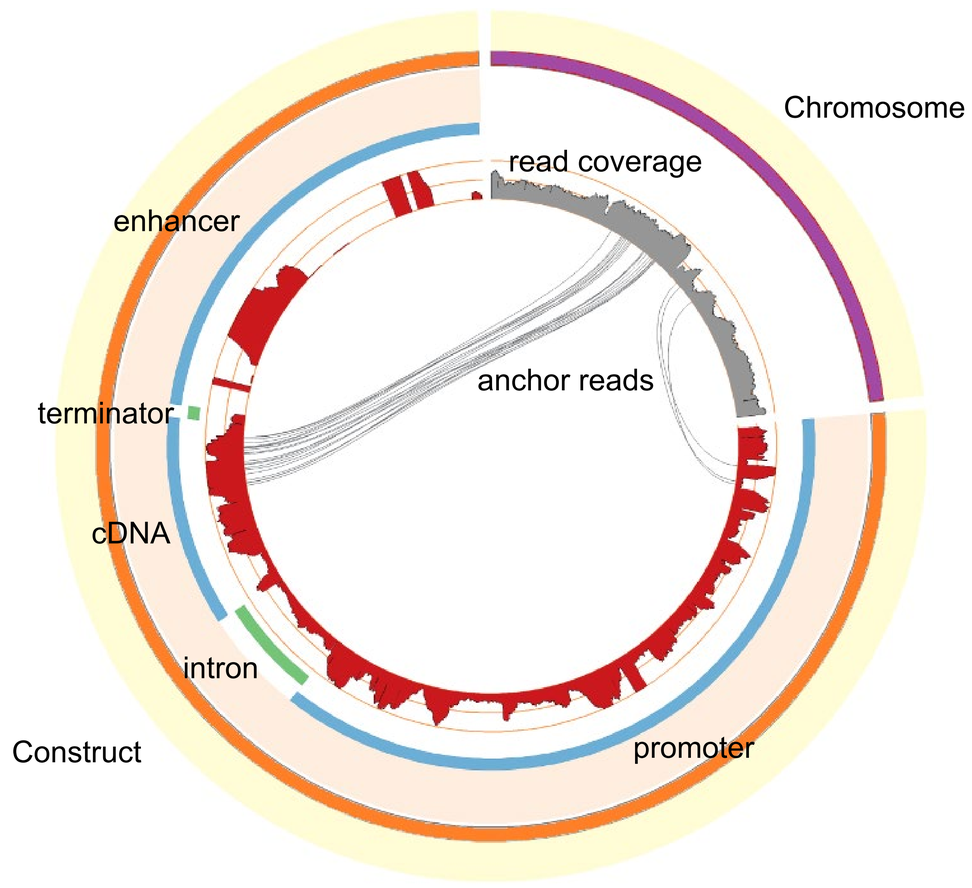
[2] Read that partially aligns to both the host and the construct (soft clipped)

This project is financially supported by the Swedish Foundation for Strategeic Research

Contact: marcela.davila@gu.se

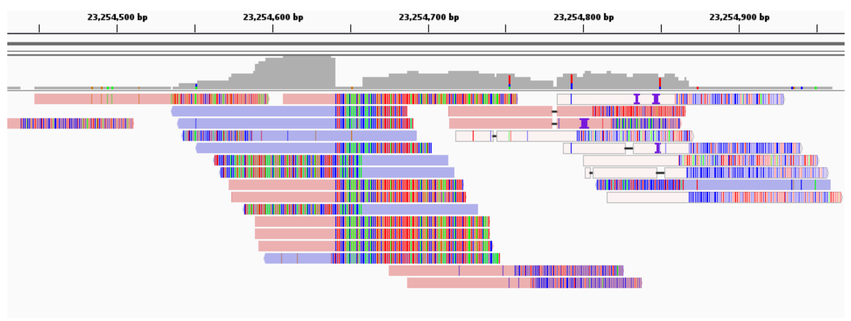
Graphical representation

The anchor reads are plotted in a circular graph, showing both the insertion site in the target genome (purple region) and the piece of the construct that is inserted (orange region). The coverage for the insertion region (gray histogram) and the construct (red histogram) are shown together with the distribution of the different elements of the construct (blue/green boxes).



Ongoing work

We are implementing the extraction of the insertion sites, using the soft clipped reads:



We are on the validation step of the candidate insertion sites of our samples.

We are setting a benchmarking protocol to evaluate TC-Hunter against some known algorithms (see references)

References

Guo, B., et al. (2016). "Identification of Genomic Insertion and Flanking Sequence of G2-EPSPS and GAT Transgenes in Soybean Using Whole Genome Sequencing Method." Frontiers in Plant Science **7**(1009).

Lambirth, Kevin C. et al. "CONTRAILS: A Tool for Rapid Identification of Transgene Integration Sites in Complex, Repetitive Genomes Using Low-Coverage Paired-End Sequencing." *Genomics Data* 6 (2015): 175–181. *PMC*. Web. 28 Aug. 2018.

(And references within)