DNA-Seq of algae transfected with 527 Construct

Raw data statistics and quality graph:

FastQC Analysis of raw fastq showing quality of all bases

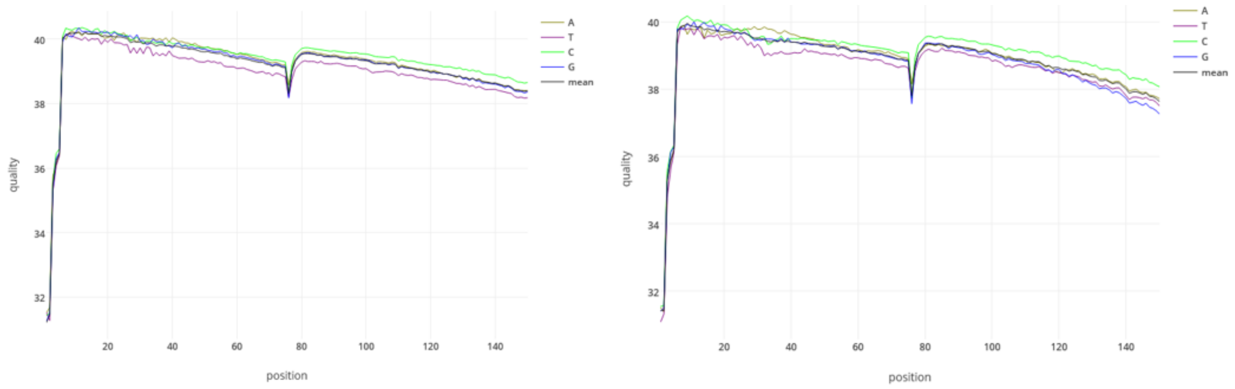


Figure 1 - Quality graph for all reads. Left R1, Right R2

Quality Control Preprocessing:
Fastp analysis on the raw fastq files removed 18 starting 5' bases

Endogenous regions in 527 construct

Region Name	Construct Start - Stop	Algae Chromosome	Algae Start	Algae Stop	E-value	bitscore
FCPA/LHCF4/CLP_promoter	902-1342	NC_011670.1	138139	137700	0	769
Vac	1347-1412	NC_011680.1	187955	187891	4.09E-28	121
FCPA-terminator	2145-2474	NC_011670.1	137195	136859	5.27E-134	475
FCPB/LHCF4_promoter	2475-2719	NC_011670.1	136804	136560	6.78E-87	318
FCPA-terminator	3095-3336	NC_011670.1	137105	136859	4.03E-84	309

Table 1: Table showing 527 endogenous region overlap with the Phaeodactylum tricornutum genome

We extracted sequences from the regions marked as endogenous in the construct and performed BLAST on the algae genome to find matches. After masking of these regions from the algae genome, we further mapped all reads to it for downstream analysis.

After pre-processing and performing the alignment we received the following samples (in yellow the reads we continue with from each stage):

Genome:	527 Construct			Phaeodactylum_tricornutum From IR1				Phaeodactylum_tricornutum Raw Files			
Aligner:	trimmer: fastp / aligner: bwa-mem [default]			trimmer: fastp / aligner: bwa-mem [default]				trimmer: fastp / aligner: bwa-mem [default]			
Read Set:	IR1			IR2							
Total Reads(R1+R2)	**mapped**	**unmapped**	**% mapped**	**TOTAL**	**mapped**	**unmapped**	**% mapped**	**TOTAL**	**mapped**	**unmapped**	**% mapped**
17,156,776	19,290	17,137,486	0.11%	19,290	666	18,624	3.56%	17,156,776	16,279,824	876,952	94.8%

Table 2: Table showing alignment statistics

We initially aligned all reads from the raw data to the 527 construct. 0.11% of all raw reads map to the plasmid. We then filter the IR1 set by read quality and fragment length and map the resulting reads to the algae. We remain with 666 which are shared between the genomes.

We aligned the raw reads to the algae genome only for control reasons, no further analysis was implemented.

An alignment coverage graph for the 527-construct genome

We initially observe the alignment for the construct. A global view of this alignment can be seen below (upper pane fig. 2). We see multiple breakpoints in the plasmid, easily recognized by discordant reads and by the clipped reads signaling the breakpoint in the genome.

To increase possible detection of insertion points in the algae overlooked by TDNAscan, we manually extracted all CLR and DIR reads using JVARKIT[11] from the IR1 read group (lower pane fig. 4) and aligned this group to the algae genome. We than scanned regions of coverage to detect possible insertion point.

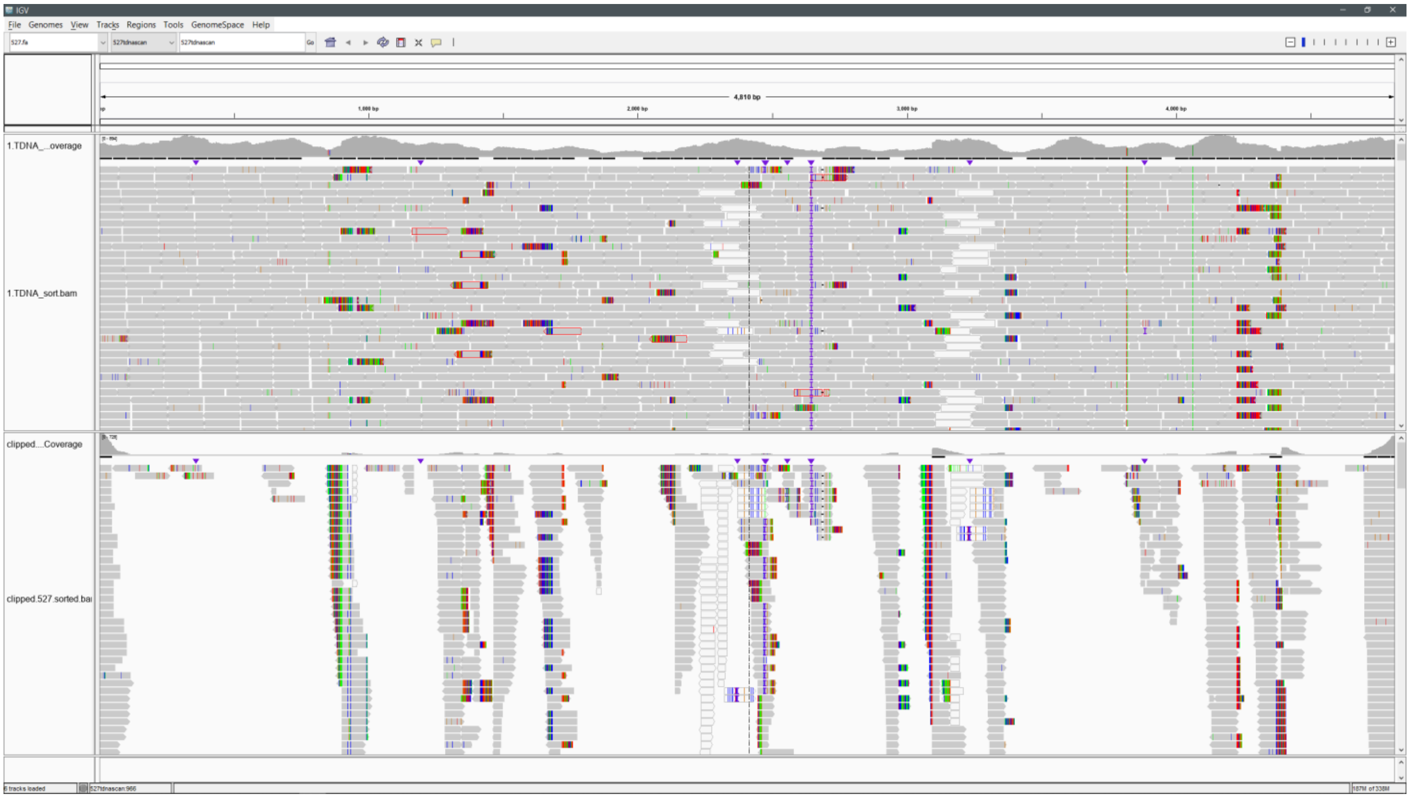


Figure 2 - Upper pane: Global view of alignment to 527 construct. Lower pane: extracted CLR and DIR from initial alignment

Detection of DNA construct insertion location and orientation in the Algae genome

Combined TDNAscan analysis and manual analysis identified three regions of possible insertion of truncated 527 construct into the algae genome. Each region is identified based on the number of clipped reads (CLR) and discordant reads (DIR) identified around the breakpoint.

#	Algae Chr	Algae Breakpoint	Number of CLR and DIR in the region	DNA Construct Start and Stop	Orientation	Insertion Freq	Algae Insertion Site
1	NC_011694.1	234748	CLR:30 DIR:1	tdna_st:1686 tdna_end:-	+	0.267857142857	PHATRDRAFT_50043
2	NW_002238037.1	42944	CLR:59 DIR:0	tdna_st:- tdna_end:4227	-	0.093949044586	Upstream to PHATRDRAFT_bd1488
3	NC_011675.1	192081	CLR:2 DIR:0	tdna_st:- tdna_end:-	-	Unknown	FRE4

Table 3 - TDNAscan identified insertion locations by 527 construct

Each row in the table above displays an insertion position in the algae genome called by TDNAscan.

- Algae Chr – The algae chromosome on which an insertion has been identified
- Algae Breakpoint – The position of insertion
- Number of CLR and DIR in the region - The amount of CLR and DIR reads used to call this insertion
- DNA Construct Start and Stop - The start and stop positions of the truncated plasmid, in case the region could not be established from the data, a minus sign is given
- Orientation - The orientation on the inserting plasmid
- Insertion Freq - Zygosity Estimation of the inserting plasmid
- Algae Insertion Site – Annotates if a gene exists in this position

Region 1: (Insert location in Algae - NC_011694.1: 234748)

Insertion position of the truncated construct DNA is clearly seen by 30 CLR and 1 DIR reads aligned to the 527 construct with positive orientation.

Non distinct information regarding PHATRDRAFT_50043 found on JGI

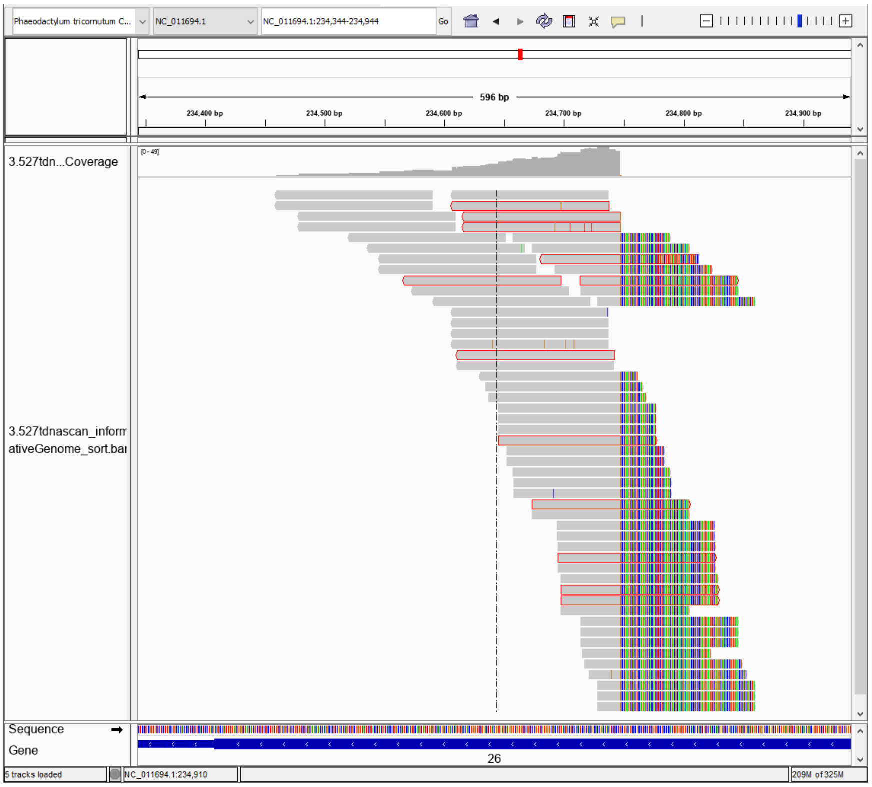


Figure 3 - Region1, truncated clipped and discordant reads showing position of tDNA breakpoint locations on algae genome

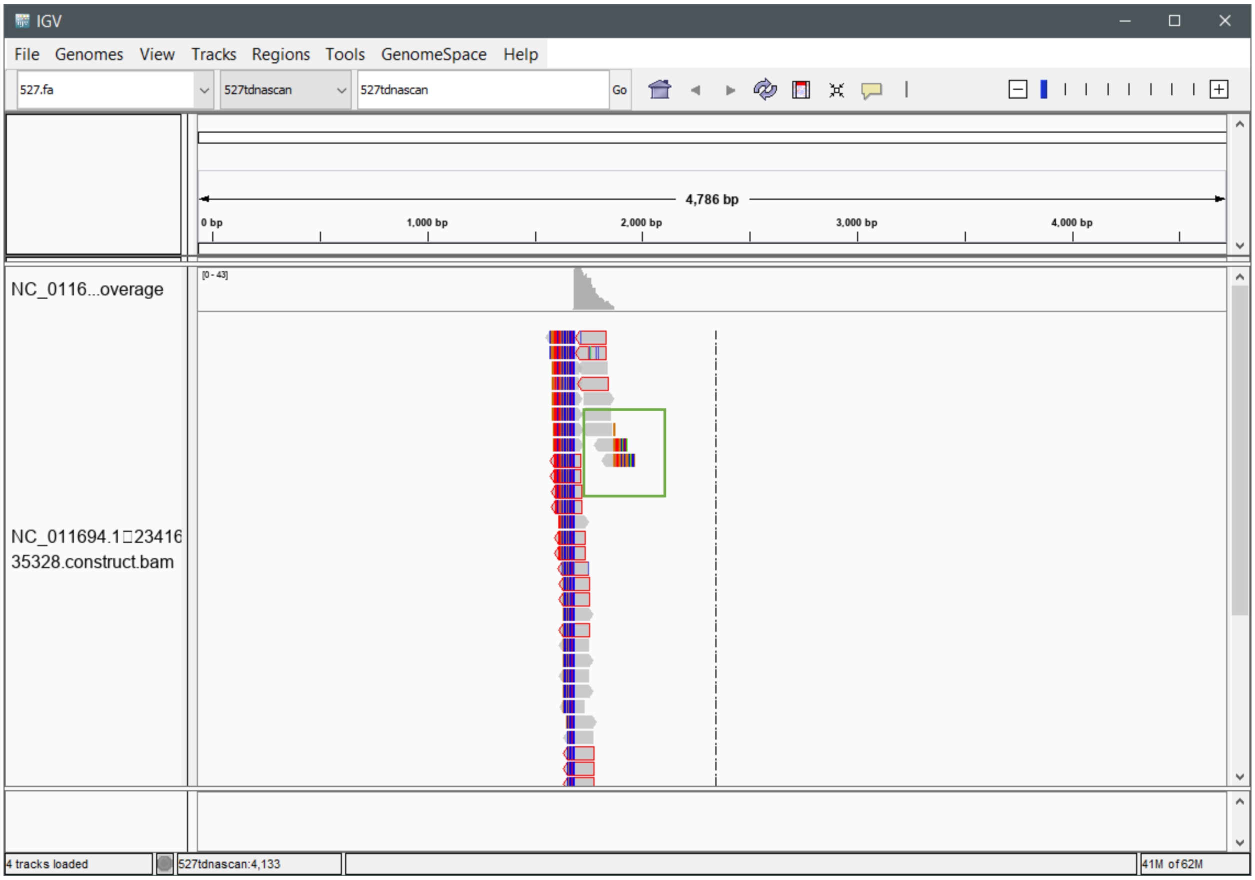


Figure 4 - Global view of truncated soft clipped reads showing start position of tDNA breakpoint locations on 527 construct genome for region 1. Reads in green box indicate supplemental alignments and might indicate truncation construct stop

This region is located at the middle of GFP2 and might be responsible for one of the southern blot bands detected.

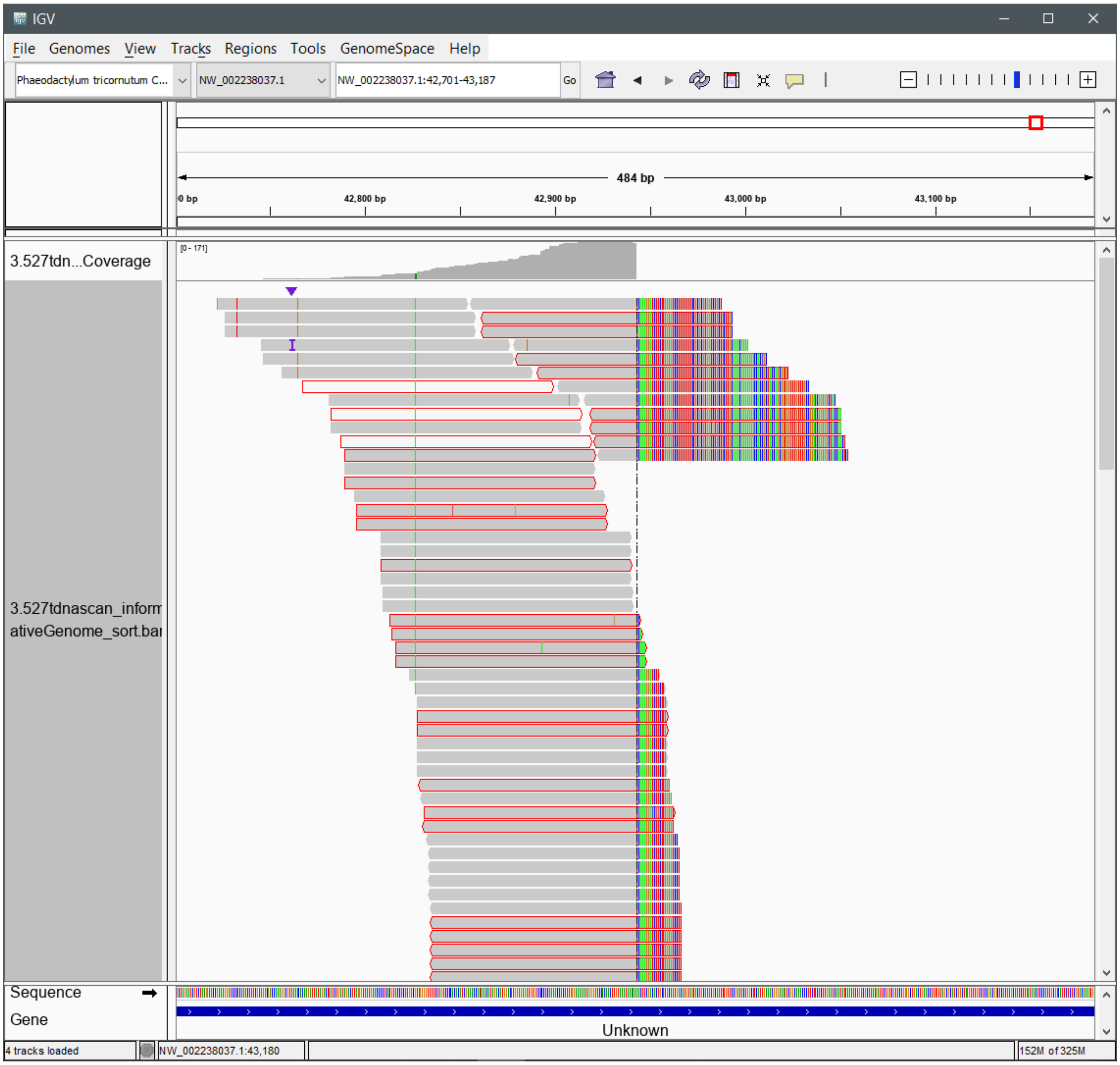
>Region1

ACGACCGTGGTGTTCTGGACTCTGACGGGTTGATGCGAGGGTGCGGTGGTCTCGTTCCGCGGTGTCGTGGGGGCTAGCGCTT
GACTGCCTTCCGCACGTCCTCTACCTGCCACCGGACCGACGTCTCCCCTgcccgaaggctacgtccaggagcgcaccatcttcttcaaggacga
cggcaactacaagacccgcgccgaggtgaagttcgagggcgacaccctggtgaaccgcatcgagctga

Region 2: (Insert location in Algae - NW_002238037.1: 42944)

Insertion position of the truncated construct DNA is clearly seen by 59 primary aligned CLR reads to the 527 construct with negative orientation.

PHATRDRAFT_bd1488 is an hypothetical protein according to BioCyc



Phaeodactylum tricornutum C...

NW_002238037.1

NW_002238037.1:42,701-43,187

4 tracks loaded

NW_002238037.1:43,180

152M of 325M

Figure 5 – Region2, truncated clipped reads showing position of tDNA breakpoint locations on algae genome

>Region2

GACATAGGGGTTTTCTAGTCCCCACACATCGAAACCATAAACATAGACATGGACATGGAGAGATGTGTCCCCTCAAAACCACT
ATGGGAAAAATAGGGATCCACAACTTGTAAATAAAAAtcaaaggatcttcttgagatcctttttttctgcgcgtaatctgctgcttgcaaacaaaaaa
accaccgctaccagcggtggtttgtttgccggatcaagagctaccaactct

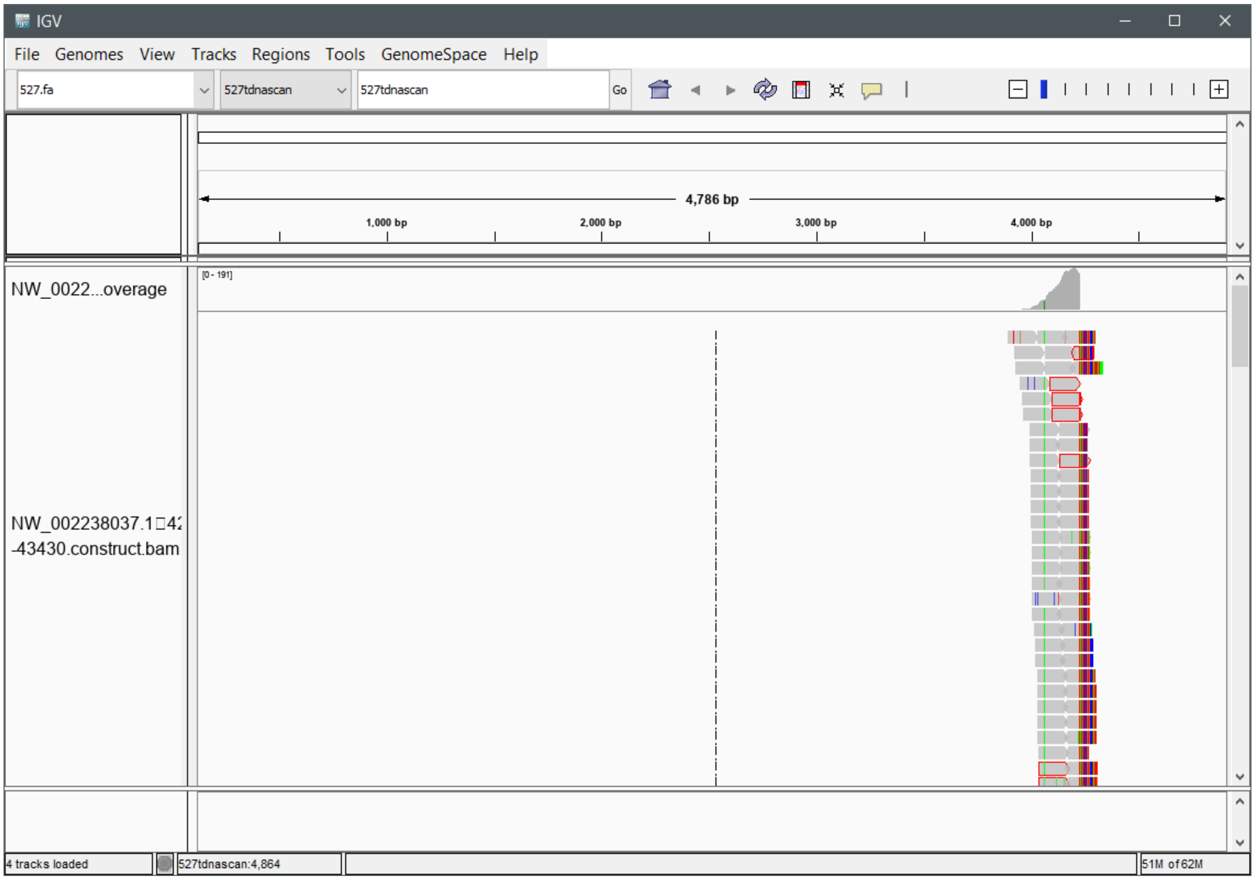


Figure 6 - Global view of truncated soft clipped reads showing end position of tDNA breakpoint locations on 527 construct genome for region 2.

This region ends far from the GFP2 and could possibly contain it.

Region 3: (Insert location in Algae - NC_011675.1: 192081)

Insertion position of the truncated construct DNA is clearly seen by 2 primary aligned CLR reads to the 527 construct with negative orientation.

Gene FRE4 Ferric reductase, NADH/NADPH oxidase and related proteins JGI

Low chances for GFP2 present at this insertion as it seems that the truncated DNA ends at position 900 but it could not be confirmed.

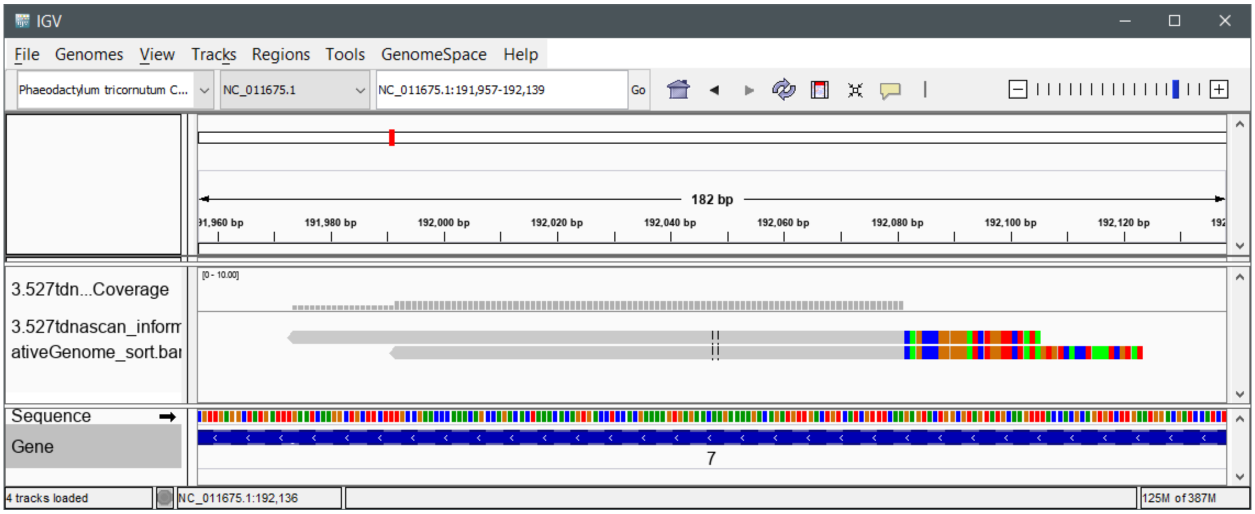
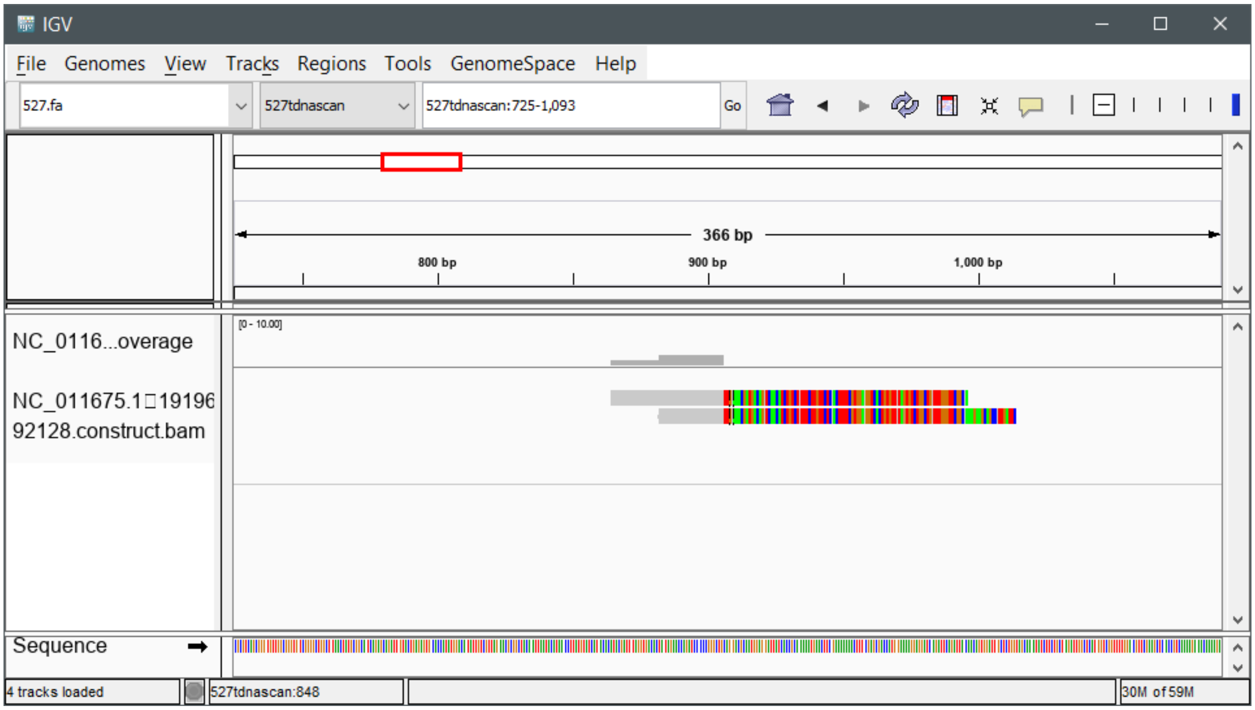


Figure 7 – Region3, truncated clipped reads showing position of tDNA breakpoint locations on algae genome

>Region3

GAATCAAGGCTGCTTCTTTGCCGAACCCAAACAGCCAGACACTAAACACAGTGACCTCCACGAAAAGTGAGAACAAGAAAAG
ACAGATGTTGACTGTCATCGTTTCAAcagcccgggggatctggttctatagtgtcacctaaatcgtat



527.fa

527tdnascan

527tdnascan:725-1,093

4 tracks loaded

527tdnascan:848

30M of 59M

Figure 8 - Global view of truncated soft clipped reads showing end position of tDNA breakpoint locations on 527 construct genome for region 3.

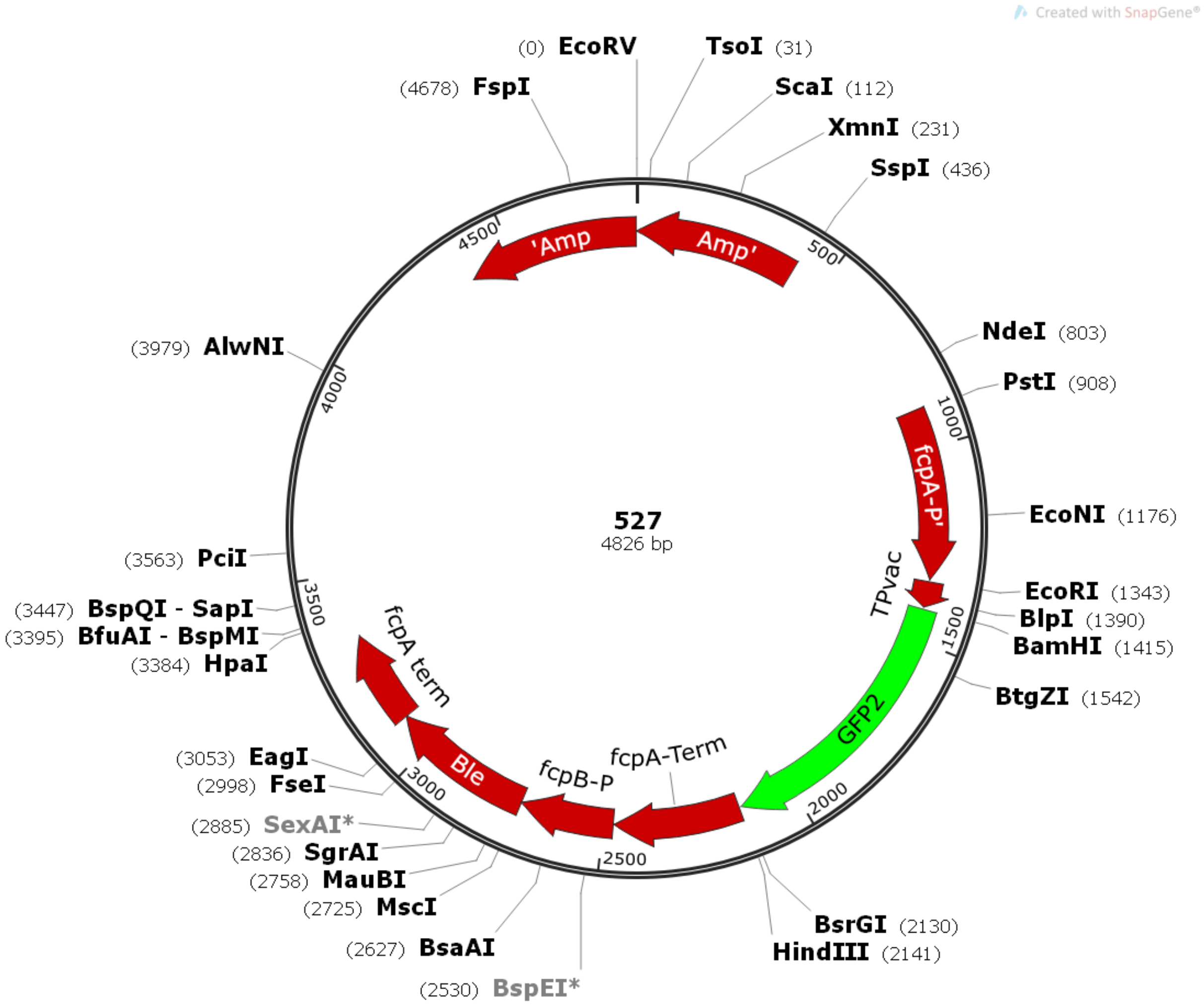


Figure 9 - Global view of 527 Plasmid

Southern Blot experiment revealed 3 copies of insertions and as per this analysis, 3 insert position has been detected

Supplemental:

Analysis procedure outline:

1. Pre-processing – fastqc (v0.11.9)[1] review in addition to the trimming of adaptors and low quality bases removal using fastp (v0.20.1)[2].
2. Removal of reads mapping to endogenous region-
    a. We converted cm5 formatted snapgene[7] files to fasta sequence and indexed using bwa index (0.7.17-r1188)[5]
    b. We extracted endogenous region of 833 plasmid and run BLAST to identify overlapping region in the Phaeodactylum tricornutum genome.
    c. Then we masked these regions in the algae genome using Bedtools(2.26.0)[10].
3. Map reads (aligning) –
    a. Building index: Obtained genome annotation file for Phaeodactylum tricornutum CCAP 1055/1 in GFF format on NCBI[3], converted into GTF format using gffread (v0.11.7)[4].
    b. We use BWA-MEM (0.7.17-r1188)[5] to align the reads coming from step 2 to Phaeodactylum tricornutum.

    We use samtools (1.9)[6] to convert and manipulate the aligned reads file.
4. Detection of the corresponding DNA construct insertion location and orientation –
    a. Building fasta reference genome and gff3 annotation file from provided SnapGene[7] cm5 file.
    b. We use TDNAscan[8] to initially align all reads to the DNA construct and filtering the non-relevant reads. The reads mapped to the construct are then mapped to Phaeodactylum tricornutum genome. The discordant reads (DIR) are defined as one read of a pair successfully mapped to the algae reference genome and the other read of the same pair mapped to part of the inserted DNA construct. The soft-clipped reads (CLR) are reads where one partial read of a single read perfectly mapped to the algae reference genome and the other partial read of the same single read perfectly mapped to the inserted DNA construct.
5. Post-processing mapped data –

    Zygosity Estimation – The reads with minimal mapping quality (MAPQ ≥ 30), which can span the DNA construct insertion sites at least 3 bp at both sides, will be considered as having the presence of a reference allele. The DNA construct insertion frequency was determined by the following equation:

$$insertion\ frequency = \frac{N_{clr}}{N_{clr} + N_{span}}$$

N_{clr} is the number of CLR at the insertion region.

N_{span} is the number of reads spanning 3 bp upstream and downstream of the insertion site.

Manipulation of alignment files using samtools[6] and Picard[9] packages

References:

[1] https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[2] Shifu Chen, Yanqing Zhou, Yaru Chen, Jia Gu; fastp: an ultra-fast all-in-one FASTQ preprocessor, Bioinformatics, Volume 34, Issue 17, 1 September 2018, Pages i884–i890, https://doi.org/10.1093/bioinformatics/bty560.

[3] https://www.ncbi.nlm.nih.gov/genome/418?genome_assembly_id=29101

[4] http://ccb.jhu.edu/software/stringtie/gff.shtml

[5] Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754-1760. [PMID: 19451168]

[6] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, and 1000 Genome Project Data Processing Subgroup, The Sequence alignment/map (SAM) format and SAMtools, Bioinformatics (2009) 25(16) 2078-9 [19505943]

[7] SnapGene® software (from GSL Biotech; available at snapgene.com).

[8] Sun, Liang, et al. "TDNAscan: A Software to Identify Complete and Truncated T-DNA Insertions." Frontiers in Genetics (2019),doi: 10.3389/fgene.2019.00685

[9] "Picard Toolkit." 2019. Broad Institute, GitHub Repository. http://broadinstitute.github.io/picard/; Broad Institute

[10] Quinlan, Aaron R., and Ira M. Hall. "BEDTools: a flexible suite of utilities for comparing genomic features." Bioinformatics 26.6 (2010): 841-842.

[11] Java utilities for Bioinformatics - http://lindenb.github.io/jvarkit/