Introduction

Given paired-end sequencing data of a Ty5_6p inserted *S. cerevisiae* (Strain: S288C) clone SacCer3, we aim to identify mapped positions, orientation, and consequences of transposon insertion with next-generation sequencing tools. We first verified the quality of sequencing data. Next, we obtained read pairs which lie at the region of transposon insertion through alignment on Ty5_6p and filtering reads of interest with appropriate FLAG tags. The final insert positions were identified with a second alignment of filtered reads on SacCer3. Finally, the orientation, as well as potential biological effects of these insertions were investigated by visualisation on a genome browser.

Implementation

A summary of the analysis pipeline is shown in Appendix 1. The methods for identifying transposon sites were inspired from the ISMapper tool[1].

Quality control

FASTQC was performed with *fastqc* command (FastQC v0.10.1)[2] on the paired sequencing reads files DNAseq_1.fq and DNAseq_2.fq. *grep* command on "FAIL" was used to detect any parameters which failed quality control. Quality control revealed that reads passed all checks. As such, pre-processing was not required, and reads were directly used in our analysis.

Identification of Insertion Sites

S. cerevisiae (SacCer3) genome and Ty5_6p transposon were indexed using *bowtie2-build* command (Bowtie2 v2.3.5.1)[3] with sacCer3.fa and ty5_6p.fa respectively. The indexes were exported as BOWTIE2_INDEXES at the directory of the genome.

Paired sequencing reads were first aligned onto the transposon sequence using *bowtie2*, with BOWTIE2_INDEXES set as directory of Ty5_6p and preset with "--very-fast" and "-p 4". DNAseq_1.fq and DNAseq_2.fq were set as the input arguments and alignments were outputted onto a SAM file using "-S" flag. *samtools view*, *samtools sort* and *samtools index* (Samtools v1.10)[4] were performed to convert SAM into BAM file format, sort alignments in BAM, and indexed. The initial alignment identifies reads which map to the junctions between SacCer3 and Ty5_6p. For these read pairs, we expect a read to map to Ty5_6p, and its mate to map to SacCer3. These reads are predicted to be aligned at the ends of Ty5_6p.

From the alignments, unmapped mates with mapped reads were filtered with their corresponding FLAG tags (Appendix 3.) using the *awk* command on the FLAG column. List of QNAME of the reads were obtained. Raw reads of these alignments were extracted using a bioinformatics tool, BBMap (v38.90) – *filterbyname.sh* script[5], with DNAseq_1.fq and DNAseq_2.fq as inputs and the list of QNAME provided as read names to be filtered. The parameter "substring=name" were indicated to allow substring matching of input read headers. These filtered unmapped mates were outputted as two fq files – filtered_1.fq and filtered_2.fq.

Since the mate-pairs had only one read aligned to Ty5_6p, mapping of the other read pair on the SacCer3 genome would correspond to the site of insertion of transposon. Unmapped mates were aligned onto SacCer3 (BOWTIE2_INDEXES set to directory of SacCer3), with filtered_1.fq and filtered_2.fq as input arguments in *bowtie2*. Output SAM file was converted to BAM, sorted, and indexed with the same commands and parameters as described earlier.

Visualisation

Ty5_6p inserted sites in the clone were investigated as follows. Firstly, chromosomes with highly mapped filtered mates were identified through enumerating frequencies in each chromosome. For closer examination of inserts position, the reads' mapping positions (POS column) were extracted and plotted as a histogram to determine its distribution on the chromosome. These analyses were performed using *awk* and *grep* ("-c" flag). Histogram was plotted in R (R v3.6.3)[6] using the *hist()* function.

All alignments were visualised with Integrated Genomics Viewer (IGV) (v2.9.2)[7]. Initial alignment of reads to Ty5_6p were performed with ty5_6p.fa as the reference. The subset of reads with unmapped mates were also loaded onto a separate track. Final alignment of filtered reads was visualised with sacCer3.fa as the reference genome. Gene annotation track were extracted from saccharomyces_cerevisiae.gff using *grep* on "gene" and loaded. Potential insertion sites were identified by querying regions corresponding to high frequency mappings from the histogram plots and assessed. Finally, the orientation of these inserts was determined by inspecting the read strand of its mate's alignment on Ty5_6p.

Code Availability

Code used in this study is provided as a shell script named main.sh.

Results and Discussion

Initial alignment of raw paired-end sequencing reads on Ty5_6p is shown in Figure 1. (Appendix 2.). Alignments revealed coverage of up to 97 reads per base. Since sequencing depth of 20X was performed, the high coverage could indicate the possibility of multiple inserts present in our clone. To further explore the inserted transposon(s), reads were coloured and categorised by first-in-pair strands, which revealed that both positive (F1R2) and negative (F2R1) pairs were present (Figure 1). This confirms that the transposon was indeed inserted more than once, and that both forward and reverse inserts are present in our clone.

Subset of unmapped mates were also visualised and found to be mapped to the ends of the transposon (Figure 1). This was expected, since these read-pairs is hypothesised to lie at the insertion point between the transposon and SacCer3. It was noted that a small number of mates belonging to Chr III, was identified at around 3,500 to 4,000 bp, and would be discussed later.

These unmapped mates were then aligned with the SacCer3 genome. In all, these mates were mapped highly to Chr III, IV, VII, and IX, and plotting its distributions showed clustering within a narrow range on the chromosomal region (Appendix 2; Figure 2), which may

correspond to a potential region of insertion. Chr VII had two distinct peaks: 940 to 980 bp, and 145,200 to 145,500 bp, and was plotted separately. From the five potential regions of insertion, direct query on IGV revealed three regions as true insertion points, with positive (pink) and negative (purple) strands flanking the insertion region (Appendix 2; Figure 3) at Chr IV (insert #1), IX (insert #2) and VII (145,374 - 145,413 bp) (insert #3). Inserts #1 and #2 were within the genes *TRP4* and *COX5B* respectively. Insert #3 did not affect any genes but was located close to the 3' end of *COX13*. The transposon at insert #2 was in the reverse direction. These results were summarised in Table 1 (Appendix 2).

Trp4 encodes for a transferase involved in the tryptophan biosynthetic pathway (Uniprot: P07285), while *COX5B* transcribes the subunit 5b of Cytochrome C Oxidase (Uniprot: P00425) and has roles in oxidative phosphorylation[8]. Therefore, it is likely that the metabolic pathways in the clone is impaired. Furthermore, the 3' ends of *COX13* (insert #3) was found to be enriched for several transcription factors (GAL4, HOG1, YFL044C) via CHIP-seq from *Harbison et. al* [9]. This insert may thereby affect regulation of downstream genes of these TFs.

Reads that mapped highly on Chr III were indeed not true insertions. Alignment to SacCer3 showed mapped mate pairs and of the correct insert size (not shown). These could have been the contribution from the small cluster at 3,500 to 4,000 bp of Ty5_6p, as described earlier. It is likely that SacCer3 and Ty5_6p might share similar sequences at these regions. The region at Chr VII from 940 to 980 bp was concluded to not be a true insertion site. Unlike the other three inserts, there was no clear point flanked by the positive and negative reads on either side (not shown). The alignments also had mapped mates on other chromosomes in SacCer3, and not within the insert size of 300 bp. Nonetheless, this region does not map to a protein coding gene and as such, further exploration might not be necessary.

Limitations

In our study, it is possible that other transposon insertions have not been identified due to insufficient coverage at these regions. There could be sequencing bias against the junctions between genome and transposon, leading to poor detection at these insertion sites.

Conclusion

With next-generation sequencing tools, we have successfully identified three potential sites of Ty5_6p transposon insertion: namely on chromosomes IV, IX and VII. Their orientation and the biological effects of these insertions has been investigated. Two inserts were indeed found to disrupt genes involved in metabolic pathways. The last insert, which lies on a regulatory sequence, could alter the expression of downstream genes. *In vitro* characterisation of the clone is recommended to elucidate the true biological effects of these inserts.

References

1. Hawkey J, Hamidian M, Wick RR, Edwards DJ, Billman-Jacobe H, Hall RM, et al. ISMapper: identifying transposase insertion sites in bacterial genomes from short read sequence data. BMC Genomics. 2015 Dec;16(1):667.

2. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010; Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

3. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Apr;9(4):357–9.

4. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.

5. Bushnell B. BBMap: filterbyname.sh [Internet]. Available from: https://github.com/BioInfoTools/BBMap/blob/master/sh/filterbyname.sh

6. R Core Team. R: A language and environment for statistical computing. [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2020. Available from: https://www.R-project.org/

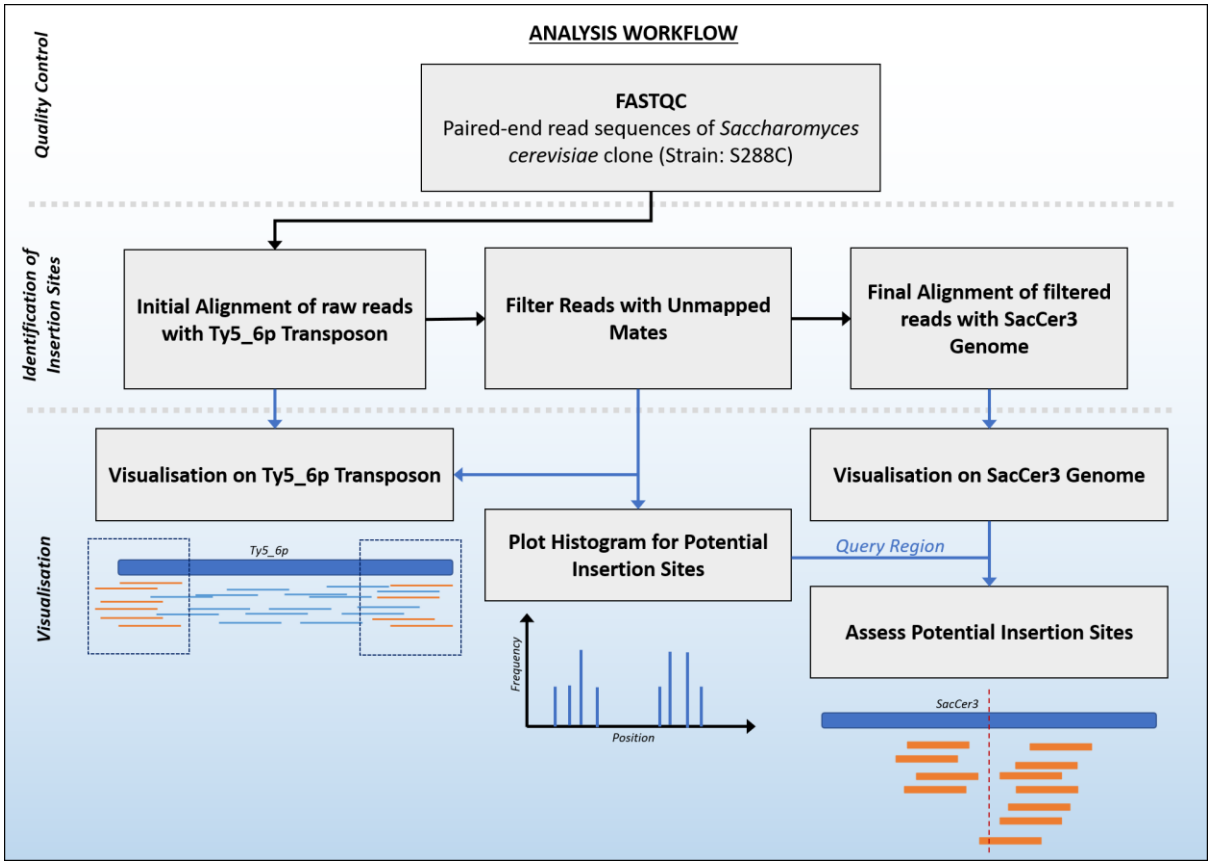
7. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011 Jan;29(1):24–6.

8. The UniProt Consortium, Bateman A, Martin M-J, Orchard S, Magrane M, Agivetova R, et al. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research. 2021 Jan 8;49(D1):D480–9.

9. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, et al. Transcriptional regulatory code of a eukaryotic genome. Nature. 2004 Sep;431(7004):99–104.

Appendices

Appendix 1: Analysis Workflow



Appendix 2: Figures and Tables of Results

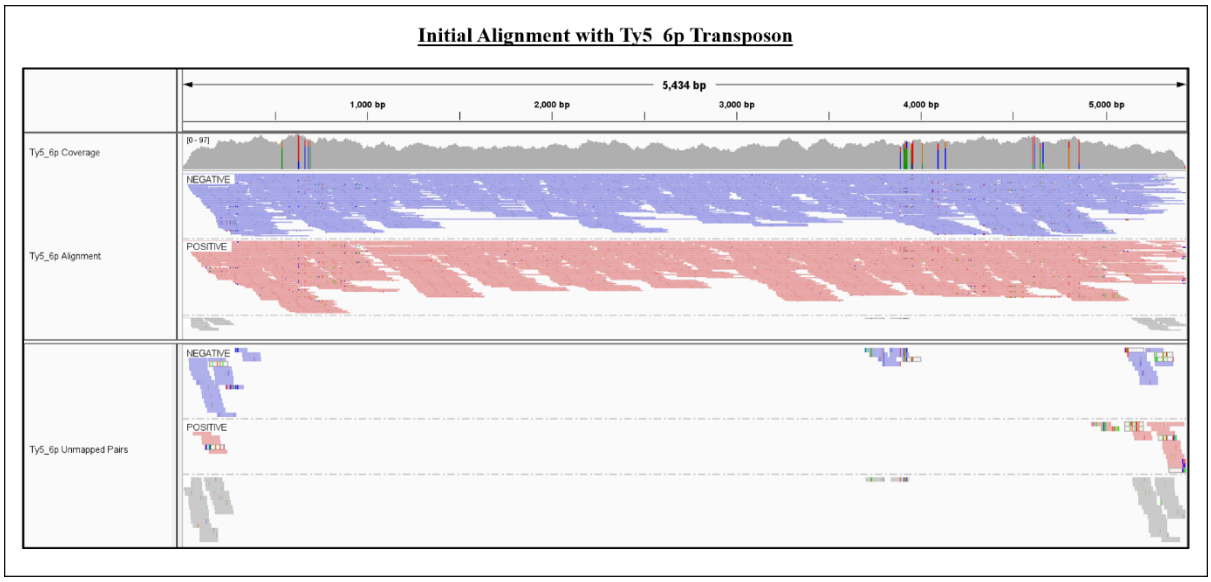


Figure 1. Initial alignment of raw paired-end sequencing reads on Ty5_6p. Reads were coloured and grouped by first-in-pair strand (negative: F2R1; positive: F1R2). Subset of reads with unmapped mates were mainly found at the ends of transposon sequence. Coverage of the transposon ranged from 0-97.

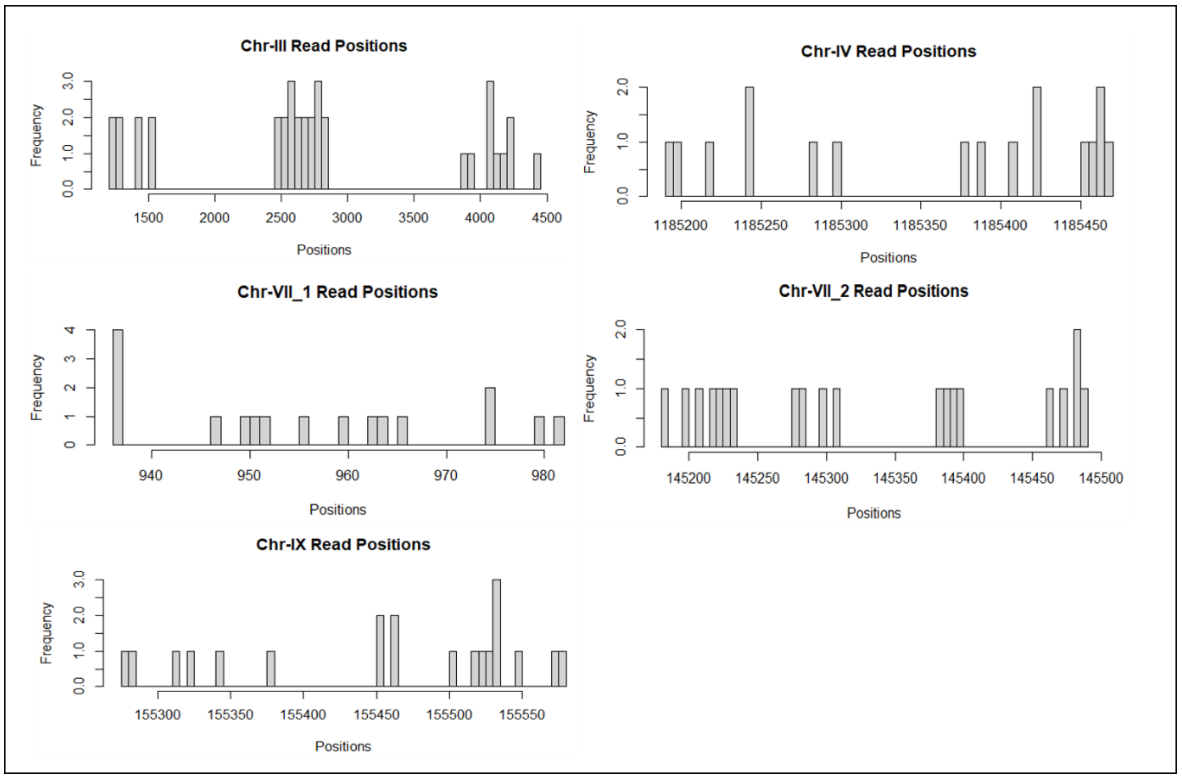


Figure 2. Histogram of positions in chromosomes with highly mapped reads. Alignments were mainly clustered within a narrow range on each chromosomal region, corresponding to a potential position of insertion. Two clusters were found in Chr-VII; 940 to 980 bp, and 145,200 to 145,500 bp.

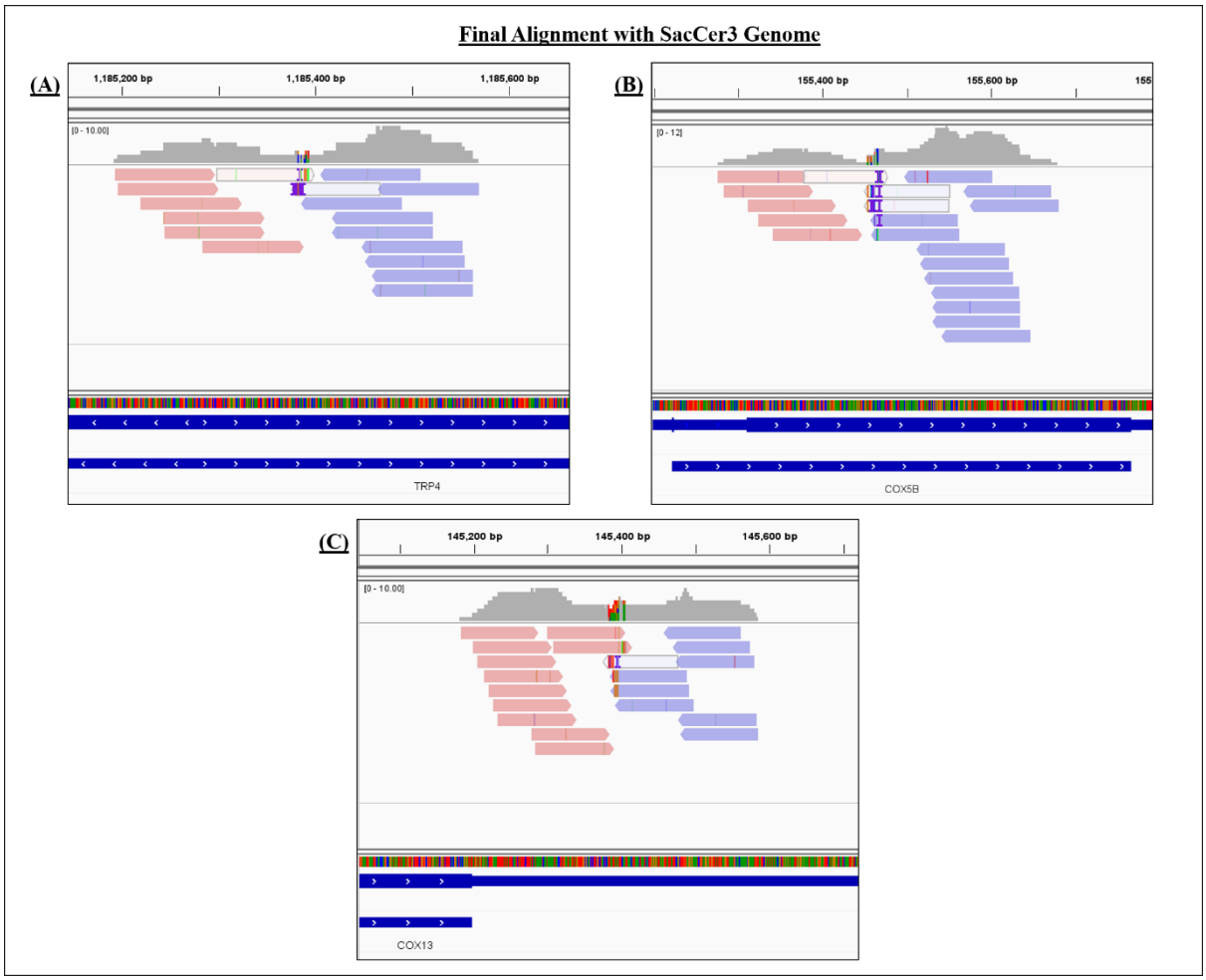


Figure 3. Final alignment of filtered unmapped mates on SacCer3 genome. Three inserts were identified: (A) ChrIV:1,185,367-1,185,406; (B) chrIX:155,442-155,481; and (C) chrVII:145,374-145,413. Reads were coloured by read stand (+: pink; -: purple). Midpoint of these alignments, with high mismatch bases is likely the point of insertion.

Table 1. Three inserts and their relative positions were identified. Of these, insert #1 and #2 were located within the sequences of *TRP4* and *COX5B* respectively. Insert #3 did not affect any genes but was located close to the 3' end of *COX13*.

Insert	Position	Orientation	Gene(s) Affected
#1	chrIV:1,185,367-1,185,406	Forward	*TRP4*
#2	chrIX:155,442-155,481	Reverse	*COX5B*
#3	chrVII:145,374-145,413	Forward	-

Appendix 3: FLAG Tags Filtered and their Associated Description

	FLAG Value			
Description (Decimal)	73	89	137	153
Read Paired (1)	✓	✓	✓	✓
Mate Unmapped (8)	✓	✓	✓	✓
Read Reverse Strand (16)		✓		✓
Mate Reverse Strand (32)				
First in Pair (64)	✓	✓		
Second in Pair (128)			✓	✓