LSM3241 CA2

HUNTING FOR GENOMIC INSERTIONS AND THEIR CONSEQUENCES

CHER WEI YUAN (A0155266B)

Abstract

In the yeast *Saccharomyces cerevisiae*, there are five families of transposons. A transposon, Ty5-6p of family TY5 was inserted into the genome of yeast strain S288C. Here, I show the use of bowtie2 alignment and simulation techniques to find the positions and orientation of transposon insert in the yeast strain S288C. The developed protocol I used found two transposons inserted in forward orientation in chromosome III and IV in position 11537 and 1331781 respectively. Further analysis revealed that the insertions were in genes YCL068C and YDR434W, which codes for a pseudogene and GPI17 respectively. I hope my work can be valuable to others who desire to find transposable elements in their line of work.

Introduction

Transposable elements, commonly known as 'jumping genes', are repetitive and mobile DNA sequences (Bleykasten-Grosshans et al., 2013). The yeast *Saccharomyces cerevisiae* has five families of transposons named TY1 to TY5 (Zou et al., 1996). Although TY5 preferentially inserts itself in telomeric regions, it remains difficult to predict the exact position of insertion (Zou et al., 1996). In the Zachary O'Mices laboratory, Ty5-6p of family TY5 was randomly inserted into the yeast genome (strain S288C). Subsequently, next-generation sequencing (NGS) was conducted with Illumina 1.9 with 20-fold coverage and a paired-end 2x100 bp library with an average insert size of 300 bp. Here, with the data generated from NGS, I used bowtie2 alignment and simulation techniques to find the site and orientation of insertion. The methodology developed here can be valuable to future work studying the preferential insertion sites (in base-pair resolution) of other transposable elements.

Methods and Materials

I have provided an overview of the workflow in Figure 1. To achieve the project aims, I used bash in the Windows version of Ubuntu 18.04 (Canonical Group Limited, 2019). In addition, Python 3.7.2 (Python Software Foundation, 2018) and R (R Core Team, 2018) in RStudio (Core RStudio Team, 2018) were also used in the downstream analysis. The scripts documenting the codes used in the workflow are provided in GitHub (https://github.com/CherWeiYuan/LSM3241-TY5-insertion-in-SacCer3).

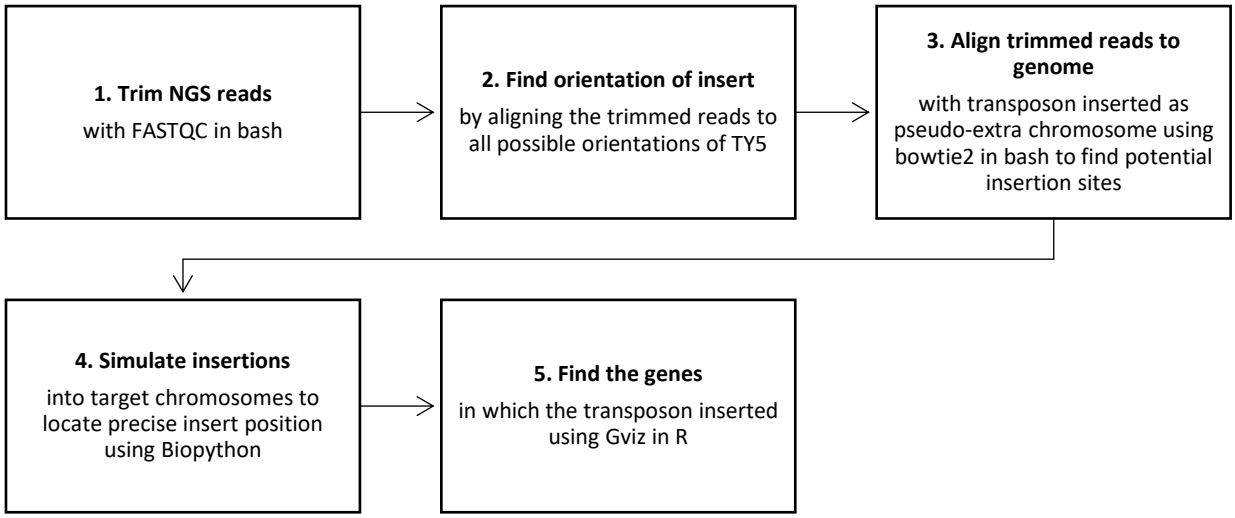


Figure 1. Workflow for identifying the number of Ty5-6p (here on referred to as TY5, genebank accession U19263.1) transposon inserts in the *Saccharomyces cerevisiae* genome (SacCer3) and the position and orientation of insertion. After the position of the insertion is determined, the biological effect of the insertion was predicted by identifying and reviewing the genes in which the inserts were located.

1. **Trimming NGS reads with FASTQC:** I assessed the quality of the reads in FASTQC (Andrews, 2018) to check if the NGS was successful in providing high quality reads for further analysis. The median base call quality score for each position in the reads were all within the high-quality region with score 28 to 40 (the equivalent of less than 0.2% error rate of base call). Overall, none of the positions had base calls that fell under the medium quality score of 20. This means that all base calls had at least 99.0% chance of being correct and thus the NGS succeeded. Nonetheless, for one of the mate pair reads, I noticed there was a problem with overrepresented 7-mers of ATACCGC around 6th position and CAGCGAT around 10-11th position. In other words, the 7-mers were preferentially located at the start of the reads. Thus, I suspect these Kmers were sequences of primers used to amplify the fragments of known sizes during library preparation. As these random primers are not recognized by FASTQC, they were identified as statistically significant spikes of enrichment.

   Hence, I needed to trim the reads before aligning them to the genome. This was done using Trimmomatic in bash (Bolger et al., 2014). From read quality encoding, under 'Encoding', I noticed that Illumina 1.9 was used. Hence, Phred-33 was used to calculate the quality score. Following which, only 121 reads were removed due to bad quality. I checked the quality of the trimmed reads using FASTQC again and the overrepresentation of Kmers problem was resolved. However, FASTQC flagged the sequence length distribution as a problem instead. Upon closer inspection, I realized the issue was due to the reads mostly having 96-100 bp. As the paired-end sequencing sequenced the 100 bp ends on each library fragment, the distribution of read lengths was expected to be near the 100 bp range. Thus, this warning by FASTQC can be safely ignored.

As the trimmed reads are of high quality and were in the expected range of read length with no important problem flagged by FASTQC, I am assured that they can be used for downstream analysis.

2. **Finding the orientation of insert:** As the transposon can be inserted in forward or reverse direction, I can figure out the way they were inserted by aligning the trimmed reads to each of the orientation. First, I generated the different orientations (reverse, reverse complement and complement) of the TY5 sequence using the programme Reverse Complement (https://www.bioinformatics.org/sms/rev_comp.html). Next, I aligned the trimmed reads to each of the file using bowtie2 (Langmead & Slazberg, 2013) in bash. As the reads only align to the forward sequence, I know that all my transposon inserts in the genome were in that orientation.

3. **Aligning the trimmed reads to the genome to find a general location of insert:** With the awareness of TY5 inserted in the forward direction, I concatenated TY5 in that specific orientation as a pseudo-extrachromosome in the SacCer3 genome in bash. Using bowtie2 again, I aligned the trimmed reads to the genome-transposon reference genome. I picked bowtie2 over HISAT as my aligner because the former align reads more precisely (Keel & Snelling, 2018). Although HISAT was faster (Keel & Snelling, 2018), I had only a genome to align and thus speed is not my concern. In contrast, precision was important because I needed the reads to be mapped correctly in order to determine the position of insertion. Next, in bash I used samtools (Wysoker et al., 2009) to convert the SAM output from bowtie2 into BAM files and to sort and index the latter files. The BAM files were then sent to IGV (Robinson et al., 2012) for visualization of the reads aligned to the genome. Then, I coloured the mates of reads that were mapped to TY5. Thus, I was able to visualize these reads on the chromosome flanking the insert region (for example, Figure 2). The flanked region marked several positions in which the transposon could be inserted.

4. **Simulation of transposon insertion events into target chromosome to find the precise location of insertion:** Upon narrowing down the general region the transposon was in, I wrote a Biopython script (Wilczynski et al., 2009) to iteratively insert the TY5 sequence into each position of the candidate site, thus generating duplicate target chromosomes with TY5 inserted into unique sites. This simulation step was necessary to predict the exact position of the transposon insert. Next, all trimmed reads were aligned to the simulated chromosome with bowtie2 and the output files were manipulated with samtools (see point 3 above). The resulting BAM files were viewed under IGV to identify the insertion point.

5. **Predicting biological effects of insertion:** The genes in which the transposon inserted were identified in R using Gviz and Bioconductor (Hahne & Ivanek, 2016) and GenomicRanges (Lawrence et al., 2013) packages. They were picked over the ggbio package (Lawrence, 2012) because they support retrieval of updated data from UCSC and offer better graphical representations. For example, I was able directly download the yeast genome and annotation from Bioconductor (Carlson, 2015; Team TBD, 2014). After Gviz, I did a literature review on the identified genes to predict the phenotype of the yeast caused by the insertion.

Results and Discussion

After the reads were trimmed, I aligned them to the transposon in different orientations (forward, reverse, reverse complement and complement. The trimmed reads aligned only to the forward read, suggesting that the transposon was inserted in a forward fashion. Hence, I concatenated the forward TY5 to the yeast genome (SacCer3) as a pseudo-extrachromosome and allowed the trimmed reads to align to it. The resulting alignment was viewed in IGV where mates (of TY5 reads) mapped to the same region of a chromosome were coloured (for example, Figure 2). Thus, the general region of transposon insertion can be identified visually. There were three general regions in the chromosome to which the mate of the TY5 reads was mapped (chrIII: 11500s, chrIII: 198600s and chrIV: 1331700s). I will begin the discussion with region 11500s in chromosome III.

As seen in Figure 2, I deduced that the insertion must be within the red box of the region of 11530 and 11554. This region is flanked by reads where those on the left of the red box had mates were mapped to the 5' end of the transposon and those on the right had mates mapped to the 3' end of TY5. For the blue, orange and brown reads within the red box, mismatches occurred often in the 3' end of the reads because the 3' ends consist of transposon sequences which mapped to chromosome (this was possible because these reads had most of their sequence mapped to the chromosome and hence could tolerate some mismatches near the ends). Hence, the region of mismatches in Figure 2 indicate a general region the insertion was located. To confirm that this region has the transposon, I searched for mates of reads from the opposite ends of the transposon that was the closest to together on the chromosome: they were the purple and light blue reads which indicated the transposon was in the red box region in Figure 2 (between position 11530 and 11554).

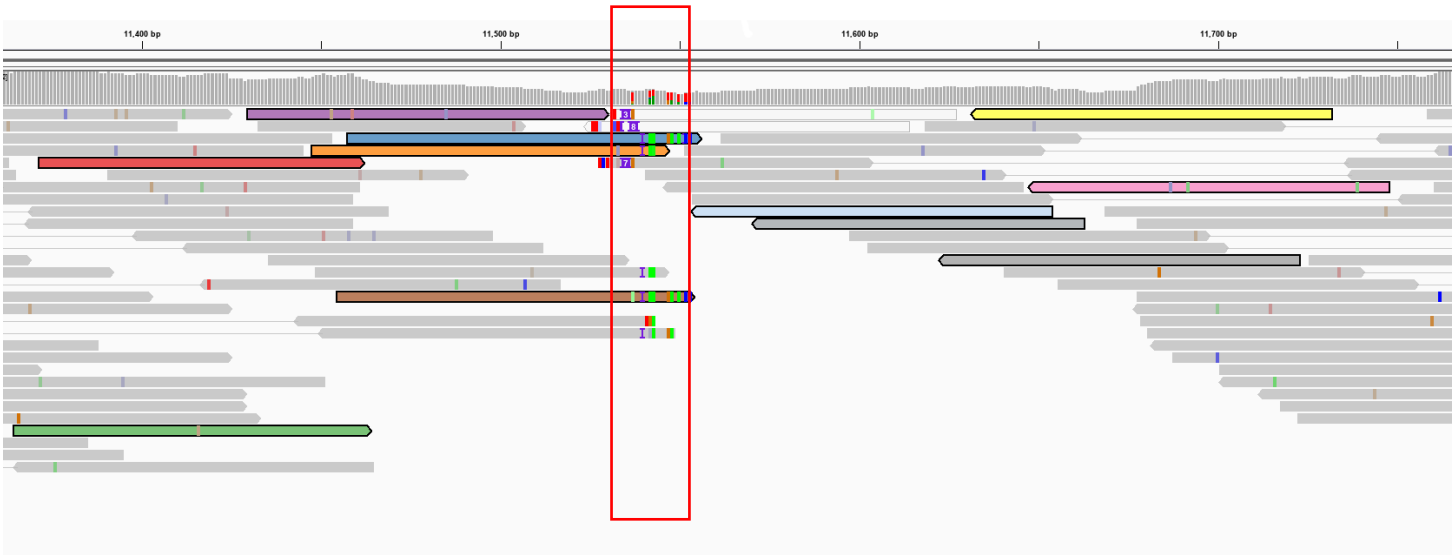


Figure 2. The region where the transposon could be inserted in chromosome III of SacCer3. Coloured reads have mate read mapped to the pseudo-extrachromosome TY5. The red box highlights a region I suspect the TY5 insert to be located (between position 11530 and 11554). Image is taken from IGV (Robinson et al., 2012).

Next, I wrote a script using Biopython to iteratively insert TY5 in each of the position from 11530 to 11554. These simulated chromosomes, each with TY5 inserted in a different position, were aligned against the trimmed reads. As the insertion at position 11552 yielded the greatest number of reads aligned to the chromosome, the chromosome was viewed in IGV first (Figure 3). The results show many mapped reads with a 15 bp gap at the 5' end of the transposon, suggesting that the actual insertion should be located 15 bp downstream of position 11553 of the chromosome in position 11537. In the chromosome with insertion at position 11537 (Figure 4), the reads were mostly aligned well with no mismatches. To ensure the read was in the exact position, I checked the simulated chromosome with insertion at position 11536 and 11538. The former had reads mapped to the region with an extra base while the latter had reads lacking a base (Supplementary Figure 1 and 2), thus confirming the transposon insert at position 11537 in the forward direction.

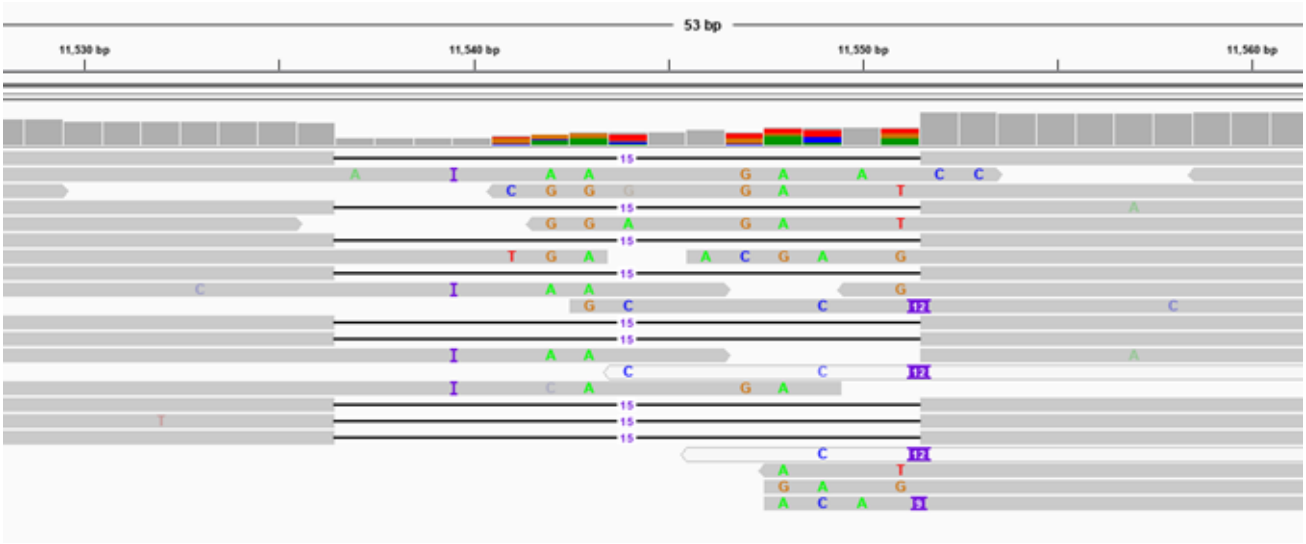


Figure 3. Chromosome III with the simulated insertion of TY5 at position 11552. Image is taken from IGV (Robinson et al., 2012).

Next, I examined the potential site of insertion in chromosome IV in the ~1331700s region. Again, I viewed the mate (of TY5 reads) mapped to the chromosome (Figure 5). On the left of the red box, the reads were mapped to the 5' end of transposon and on the right, the reads were mapped to the 3' end of TY5. Oddly, the red and blue reads in the red box were mapped to the 3' end and 5' end of TY5 respectively. As these two reads were anomalous compared to the rest of the highlighted reads, they were excluded from my analysis. Thus, I suspect the region of transposon insert to be between the closest reads with the mate on opposite ends of the transposon, which is between yellow and green read as demarcated with the red box (Figure 5).

Following which, I conducted the same simulation experiment to produce chromosomes with TY5 insertion in position 1331743 to 1331783. The chromosomes were aligned against the trimmed reads and the result at position 1331764 is shown in Figure 6. At the end of the transposon insert, there was a 17 bp gap amongst the reads mapped to the region. This suggests that the insert should be 17 bp

upstream at position 1331781. By viewing the alignment for position 1331781 insertion (Figure 7), I confirmed that the TY5 insert was in that exact position in forward orientation.

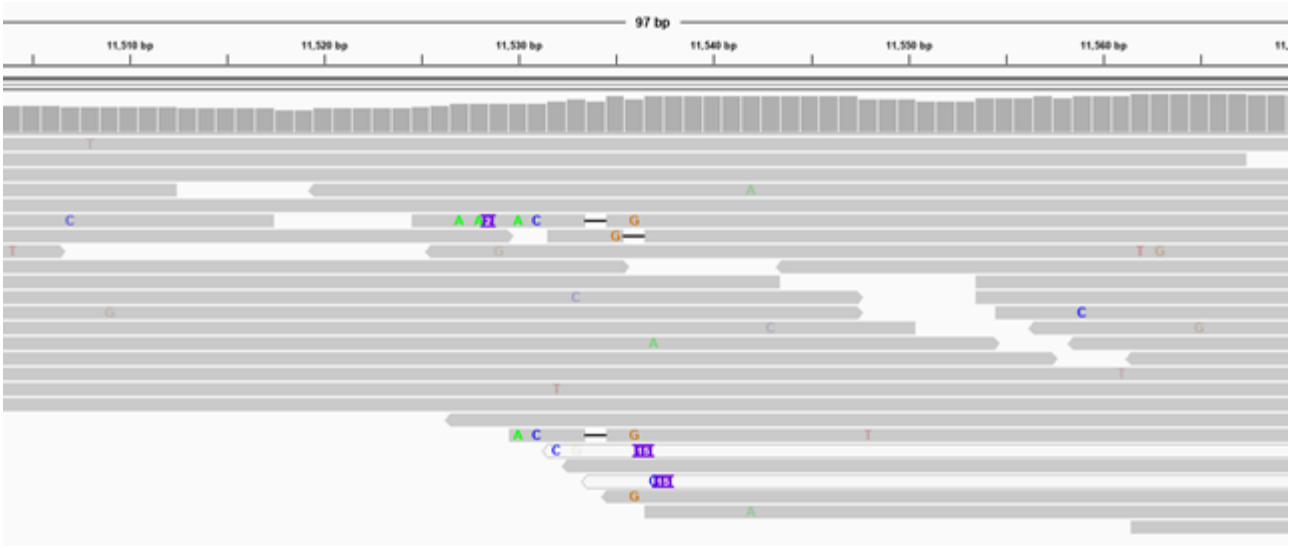


Figure 4. Chromosome III with the simulated insertion of TY5 at position 11537. Image is taken from IGV (Robinson et al., 2012).

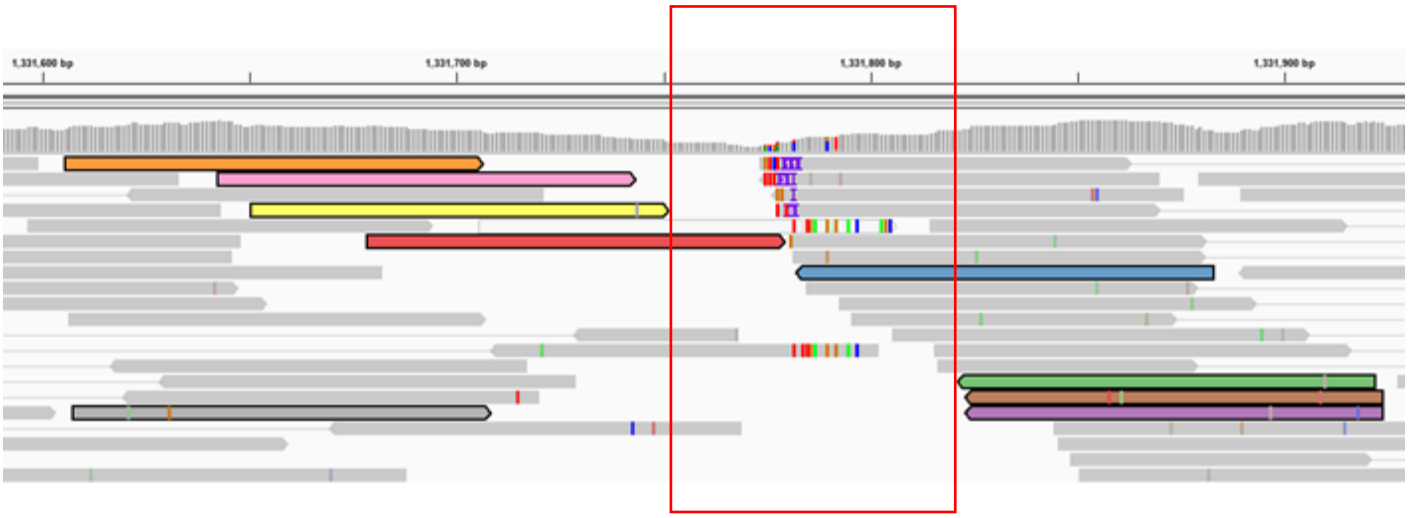


Figure 5. The region where the transposon could be inserted in chromosome IV of SacCer3. Coloured reads have mate read mapped to the pseudo-extrachromosome TY5. The red box highlights a region I suspect the TY5 insert to be located (between position 1331743 and 1331783). Image is taken from IGV (Robinson et al., 2012).

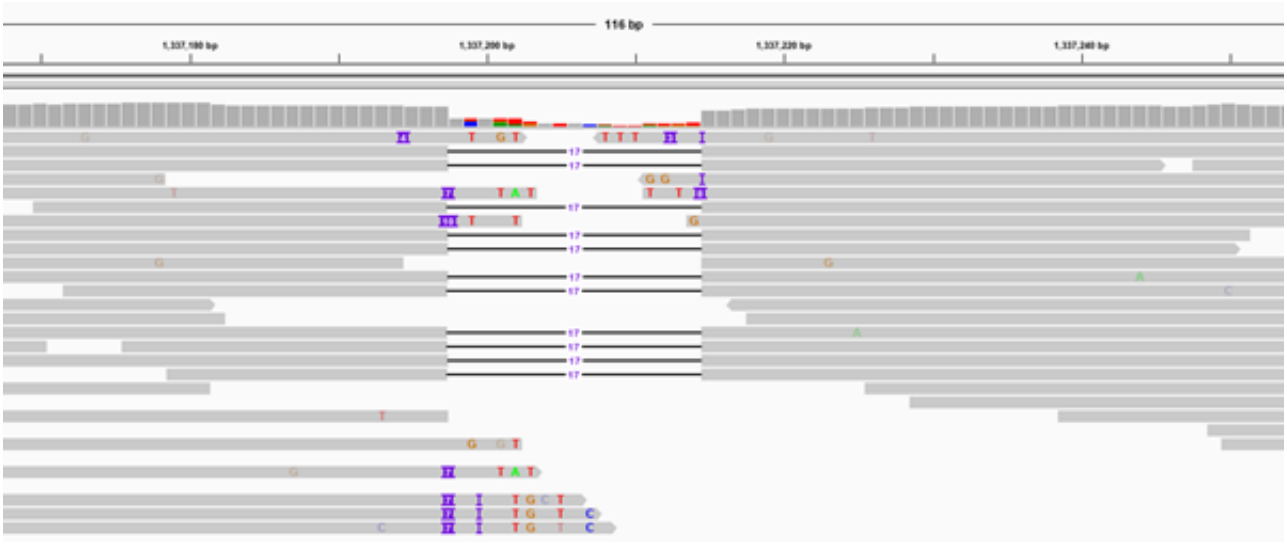


Figure 6. Chromosome IV with the simulated insertion of TY5 at position 1331764. Image is taken from IGV (Robinson et al., 2012).

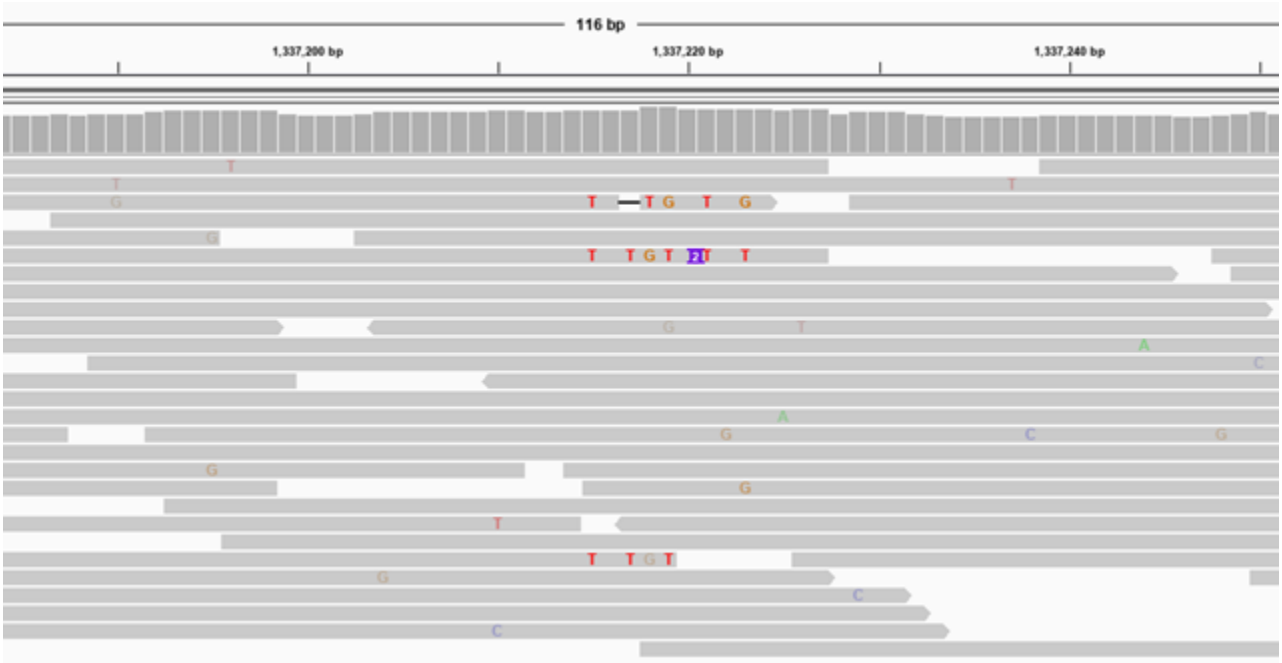


Figure 7. Chromosome IV with the simulated insertion of TY5 at position 1331781. Image is taken from IGV (Robinson et al., 2012).

The last suspected insertion indicated by the pseudo-extrachromosome TY5 was in chromosome III in the area of ~198600s. However, by viewing the region in IGV (Figure 8), there were no clear signs of an insertion in the area. If there was an insertion, then I would expect to observe an imaginary vertical column in which few reads would map (as seen in the red boxes in Figure 2 and 5) because bowtie usually does not map reads with both chromosomal and transposon sequences to the reference genome. Such a column was not observed in Figure 8. Instead, the reads in the region covered the entire area with the expected sequencing depth of 20-fold coverage. Also, as observed

from my previous examples, the region with insertions have high rates of mismatches (Figure 2 and 5) due to the reads with a small proportion of transposon sequence mapping onto the chromosome. However, the reads in Figure 8 were mapped almost perfectly.

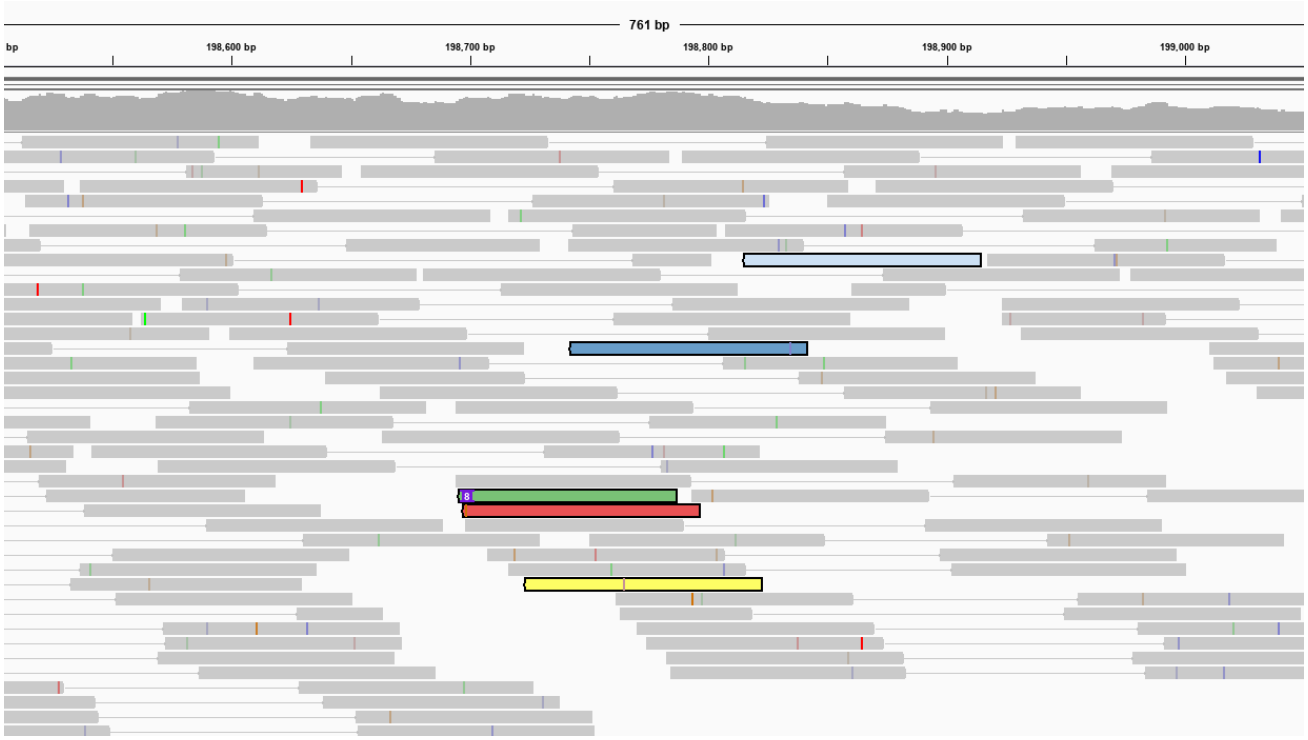


Figure 8. The region where the transposon could be inserted in chromosome IV of SacCer3. Coloured reads have mate read mapped to the pseudo-extrachromosome TY5. Image is taken from IGV (Robinson et al., 2012).

In addition, only one read mapped to the 3' end of the pseudo-extrachromosome TY5. This was unusual given the sequencing depth of 20-fold coverage. I would expect more reads to be found as per the previous examples, where more than four reads were found on each end of pseudo-extrachromosome TY5. Obtaining a low number of reads is possible but unlikely if the transposon was inserted in the region. However, by corroborating with previous pieces of evidence, it seems that the low number of reads was likely due to the transposon not existing in the region. Also, the only read from the 3' end of the pseudo-extrachromosome TY5 was not informative to indicate the position of a possible insert (see light blue read in Figure 8). This is because the read overlapped with another read (dark blue) from the 5' end of the pseudo-extrachromosome TY5. Taken together, the evidence suggests that the four reads mapped to chromosome III (Figure 8) with mate mapped to TY5 were likely to be artefacts. In conclusion, there were only two transposon inserts in the yeast genome, one in position 11537 of chromosome III and the other in position 1331781 of chromosome IV.

After obtaining the insertion positions, I plotted a gene region track using Gviz and Bioconductor (Hahne & Ivanek, 2016) for both insertions and discovered that the transposon was found at the 5' end of the YCL068C gene (Figure 9) and near the middle of the YDR434W gene (Figure 10). YCL068C is classified as a pseudogene found in the intergenic region of the *S. cerevisiae* genome (Lafontaine et al., 2004). YCL068C was identified as it shared high sequence similarity with YCR038C, an active gene involved in cell budding (Chant et al., 1991). YCR038C consist of a GDP-GTP exchange factor (Chant et al., 1991) and YCL068C was electronically predicted by Gene Ontology to have guanyl-nucleotide exchange factor activity (NCBI, 2019) which supports the view that the two genes were related by an ancestral gene with GDP-GTP exchange activity.

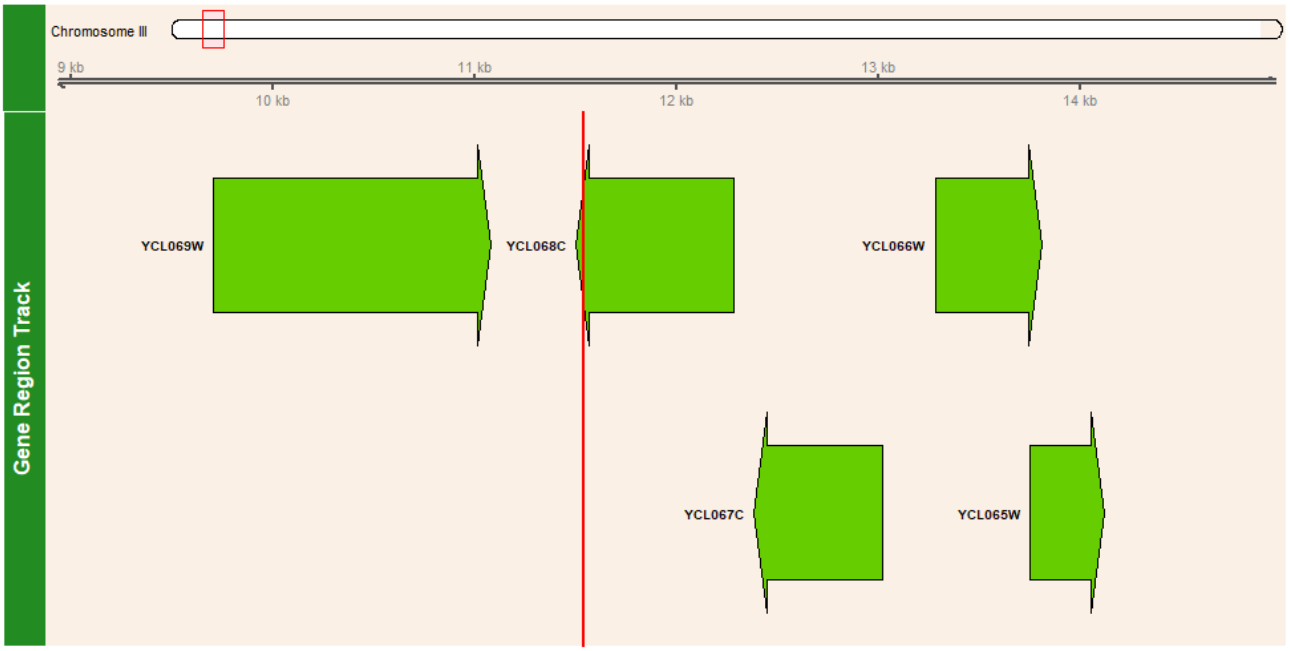


Figure 9. Chromosome III and the gene (YCL068C) in which the transposon was inserted. The red box along the chromosome ideogram shows the region where the genes in the gene region track below were located. A red vertical indicates the position 11537 where TY5 was inserted. Image is generated from Gviz and Bioconductor (Hahne & Ivanek, 2016).

Interestingly, the YDR434W gene codes for a glycosylphosphatidylinositol, GPI17. The family of GPI proteins anchor target proteins on the endoplasmic reticulum (ER), Golgi apparatus and the cell surface (Fujita et al., 2011). GPIs were first synthesized on the ER on the cytosolic side and the GPI must be flipped towards the ER lumen before it matures and attaches to the target proteins (Kajiwara et al., 2008). YDR434W/ GPI17 is part of a five-member enzyme complex called GPIT which conducts a transamidation step in the ER lumen to attach target proteins to other GPIs (Komath et al., 2018). Functionally, a YDR434W-null mutant cannot survive (Giaever et al., 2002). Yet, site-directed and random mutants of YDR434W were functional, which suggests that the protein does not contain domains that interact with the other components of GPIT (Zhu et al., 2005). To date, YDR434W is known to have only its cytosolic N- and C- terminals with a non-cytoplasmic region of no known function (InterPro, 2019; Zhu et al., 2005). Hence, as the protein has no known functional domains in the insert site, and no experiment had inserted a 5433-long sequence such as a TY5 in the

centre of YDR434W (Figure 10), it is difficult to predict if the transposon would disrupt GPIT function. To speculate, if the position of the N- and C- terminals in the cytosol is important to the function of GPI17, then the transposon may have caused the GPI17 protein to misfold and lose the ability to anchor in the ER lumen properly. If so, then GPIT function may be affected and could cause the yeast to die (Giaever et al., 2002). Notably, YDR434W/ GPI17 performs other functions such as lipid remodelling, ceramide and sterol transport (Kajiwara et al., 2008). Hence, the transposon insert may potentially disrupt these functions too but as before, it is difficult to pinpoint the disruption of functions given that no functional domains in GPI17 had been identified (Giaever et al., 2002). In such cases, it is also possible that GPI17 does not need any domains to perform its function and only require hydrophobic sequences to anchor itself on organelle or cell surfaces and its terminals to attach to proteins. If so, then the insertion may not disrupt GPI17 function at all.

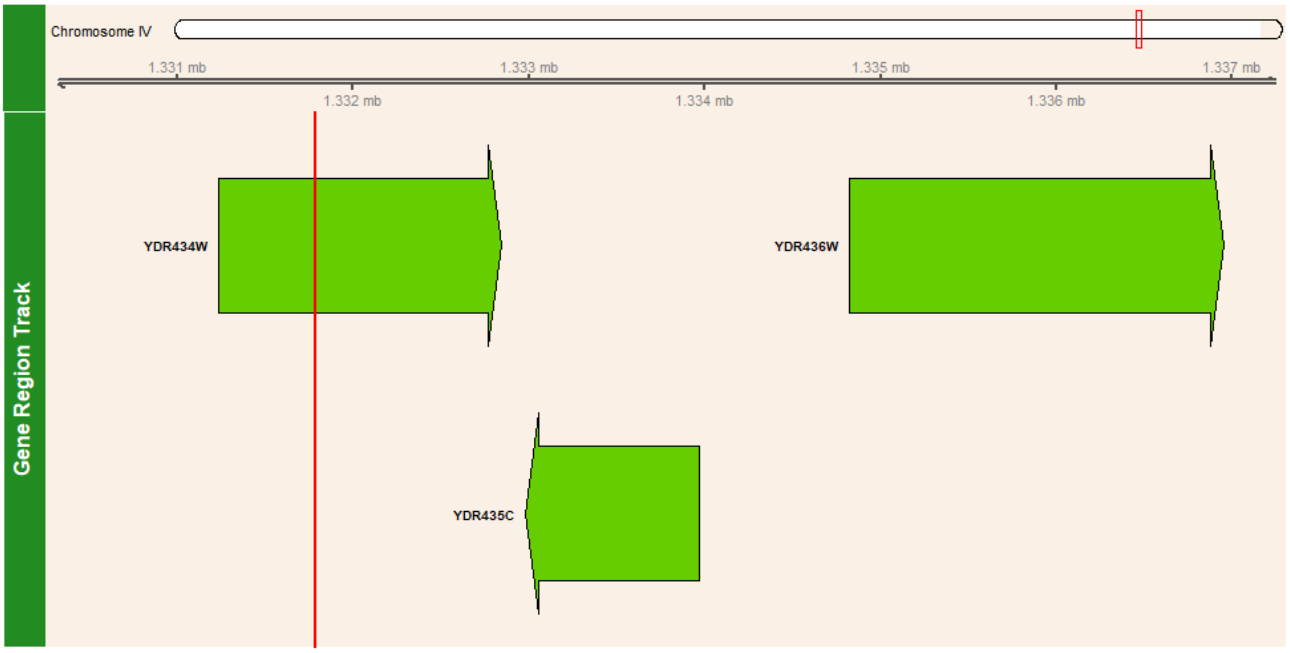


Figure 10. Chromosome IV and the gene (YDR434W) in which the transposon was inserted. The red box along the chromosome ideogram shows the region where the genes in the gene region track below were located. A red vertical indicates the position 1331781 where TY5 was inserted. Image is generated from Gviz and Bioconductor (Hahne & Ivanek, 2016).

Conclusion

In summary, TY5 was inserted into SacCer3 in chromosome III and IV in position 11537 and 1331781 respectively in forward orientation. The former position is located within a pseudogene YCL068C with a possible GDP-GTP exchange factor while the latter is found in YDR434W, a gene coding for GPI17 which is involved in protein anchoring to cell or organelle surfaces. Future experiments could explore whether GPI17 function can be disrupted by the transposon insert and if so, how the disruption was caused (e.g. by protein misfolding or domain interruption). The results can help elucidate the components of GPI17 that is important to its function.

References

Andrews, S. (2018). FastQC: A Quality Control tool for High Throughput Sequence Data. Retrieved from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Bleykasten-Grosshans, C., Friedrich, A., & Schacherer, J. (2013). Genome-wide analysis of intraspecific transposon diversity in yeast. *BMC Genomics*, *14*(1). https://doi.org/10.1186/1471-2164-14-399

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Canonical Group Limited. (2019). Ubuntu. Retrieved from https://www.microsoft.com/en-sg/p/ubuntu/9nblggh4msv6?rtc=1&activetab=pivot:overviewtab

Carlson, M. (2015). Bioconductor - TxDb.Scerevisiae.UCSC.sacCer3.sgdGene: Annotation package for TxDb object(s). Retrieved from https://bioconductor.org/packages/release/data/annotation/html/TxDb.Scerevisiae.UCSC.sacCer3.sgdGene.html

Chant, J., Corrado, K., Pringle, J. R., & Herskowitz, I. (1991). Yeast BUD5, encoding a putative GDP-GTP exchange factor, is necessary for bud site selection and interacts with bud formation gene BEM1. *Cell*, *65*(7), 1213–1224. https://doi.org/10.1016/0092-8674(91)90016-R

Core RStudio Team. (2018). RStudio: Open source and enterprise-ready professional software for R. Boston. Retrieved from https://www.rstudio.com/

Fujita, M., Watanabe, R., Jaensch, N., Romanova-Michaelides, M., Satoh, T., Kato, M., … Kinoshita, T. (2011). Sorting of GPI-anchored proteins into ER exit sites by p24 proteins is dependent on remodeled GPI. *Journal of Cell Biology*, *194*(1), 61–75. https://doi.org/10.1083/jcb.201012074

Giaever, G., Dow, S., Lucau-danila, A., Anderson, K., Arkin, A. P., Astromoff, A., … Ko, P. (2002). Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, *418*, 1–5. https://doi.org/10.1038/nature00935

Hahne, F., & Ivanek, R. (2016). Visualizing Genomic Data Using Gviz and Bioconductor (pp. 335–351). https://doi.org/10.1007/978-1-4939-3578-9_16

InterPro. (2019). GPI transamidase component GPI17. Retrieved March 30, 2019, from https://www.ebi.ac.uk/interpro/protein/Q04080

Kajiwara, K., Watanabe, R., Pichler, H., Ihara, K., Murakami, S., Riezman, H., & Funato, K. (2008). Yeast ARV1 Is Required for Efficient Delivery of an Early GPI Intermediate to the First Mannosyltransferase during GPI Assembly and Controls Lipid Flow from the Endoplasmic Reticulum. *Molecular Biology of the Cell*, *19*(1), 308–317. https://doi.org/10.1091/mbc.E07

Keel, B. N., & Snelling, W. M. (2018). Comparison of Burrows-Wheeler transform-based mapping algorithms used in high-throughput whole-genome sequencing: Application to illumina data for livestock genomes 1. *Frontiers in Genetics*, *9*(FEB), 1–6. https://doi.org/10.3389/fgene.2018.00035

Komath, S. S., Singh, S. L., Pratyusha, V. A., & Sah, S. K. (2018). Generating anchors only to lose them: The unusual story of glycosylphosphatidylinositol anchor biosynthesis and remodeling in yeast and fungi. *IUBMB Life*, *70*(5), 356–383. https://doi.org/10.1002/iub.1734

Lafontaine, I., Fischer, G., Talla, E., & Dujon, B. (2004). Gene relics in the genome of the yeast *Saccharomyces cerevisiae*. *Gene*, *335*(1–2), 1–17. https://doi.org/10.1016/j.gene.2004.03.028

Langmead, B., & Slazberg, S. L. (2013). Fast gapped-read alignmnet with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923.Fast

Lawrence, M. (2012). ggbio. Retrieved March 31, 2019, from https://bioconductor.org/packages/release/bioc/html/ggbio.html

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., … Carey, V. J. (2013). Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology*, *9*(8), e1003118. https://doi.org/10.1371/journal.pcbi.1003118

NCBI. (2019). YCL068C hypothetical protein [*Saccharomyces cerevisiae* S288C]. Retrieved March 30, 2019, from https://www.ncbi.nlm.nih.gov/gene/?term=YCL068C%5Bgene name%5D AND *Saccharomyces cerevisiae*%5Borganism%5D

Python Software Foundation. (2018). Python Language Reference. Retrieved from https://www.python.org/downloads/release/python-372/

R Core Team. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://cran.r-project.org/bin/windows/base/

Robinson, J., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2012). Integrative Genomics Viewer. *Nat Biotechnol.*, *29*(1), 24–26. https://doi.org/10.1038/nbt.1754.Integrative

Team TBD. (2014). BSgenome.Scerevisiae.UCSC.sacCer3: *Saccharomyces cerevisiae* (Yeast) full genome. Retrieved from http://bioconductor.org/packages/release/data/annotation/html/BSgenome.Scerevisiae.UCSC.sacCer3.html

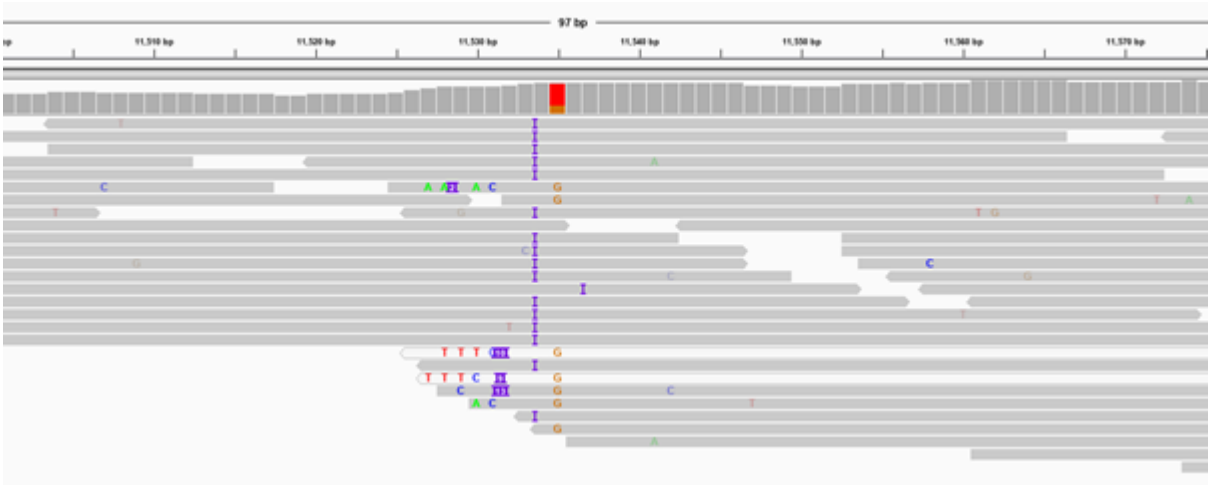
Wilczynski, B., Cox, C. J., Friedberg, I., de Hoon, M. J. L., Cock, P. J. A., Chapman, B. A., … Hamelryck, T. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Wysoker, A., Fennell, T., Marth, G., Abecasis, G., Ruan, J., Li, H., … Handsaker, B. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

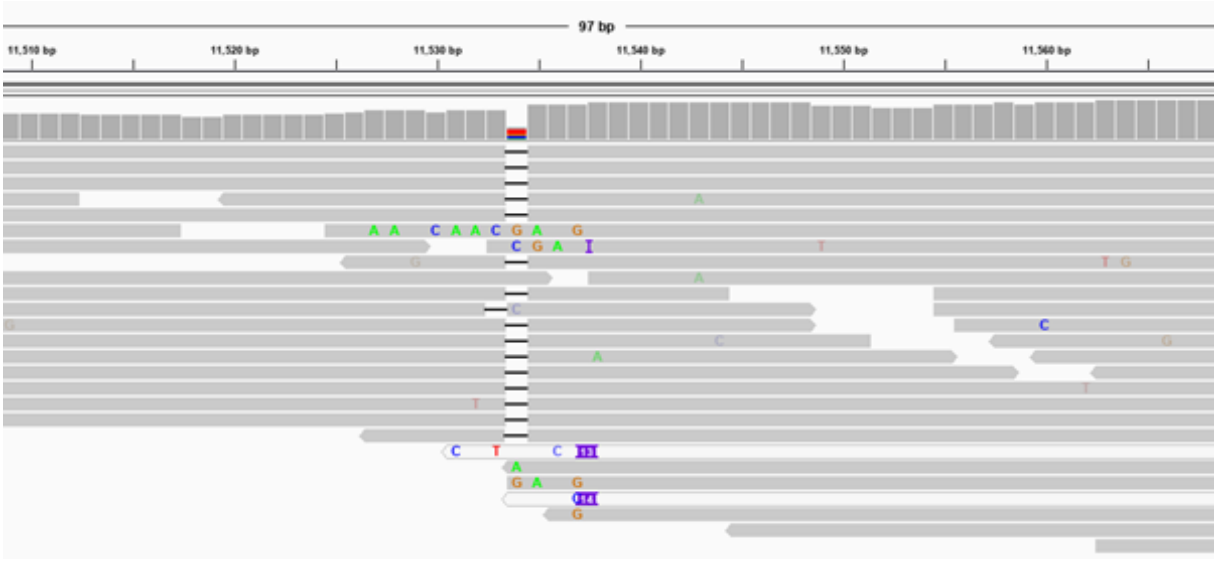
Zhu, Y., Fraering, P., Vionnet, C., & Conzelmann, A. (2005). Gpi17p does not stably interact with other subunits of glycosylphosphatidylinositol transamidase in *Saccharomyces cerevisiae*. *Biochimica et Biophysica Acta - Molecular and Cell Biology of Lipids*, *1735*(1), 79–88. https://doi.org/10.1016/j.bbalip.2005.05.001

Zou S, Ke N, Kim Jm, & Voytas Df. (1996). The Saccharomyces retrotransposon Ty5 integrates preferentially into regions of silent chromatin at the telomeres and mating loci. *Genes and Development*, *10*(5), 634–645.

Appendix



Supplementary Figure 1. Chromosome III with simulated insertion of TY5 at position 11536. Reads mapped to the region seems to have an extra base, suggesting that the insertion should be one base upstream at position 11537.



Supplementary Figure 2. Chromosome III with simulated insertion of TY5 at position 11538. Reads mapped to the region seems to have a missing base, suggesting that the insertion should be one base downstream at position 11537.