MacVector 17.5

for Mac OS X

Using NGS Data to Map Transposon Insertion Locations

Tor los

Software for Scientists

Copyright statement

Copyright MacVector, Inc, 2020. All rights reserved.

This document contains proprietary information of **MacVector**, **Inc** and its licensors. It is their exclusive property. It may not be reproduced or transmitted, in whole or in part, without written agreement from **MacVector**, **Inc**.

The software described in this document is furnished under a license agreement, a copy of which is packaged with the software. The software may not be used or copied except as provided in the license agreement.

MacVector, Inc reserves the right to make changes, without notice, both to this publication and to the product it describes. Information concerning products not manufactured or distributed by **MacVector, Inc** is provided without warranty or representation of any kind, and **MacVector, Inc** will not be liable for any damages.

This version of the Transposon Insertion Mapping tutorial was published in January 2020.

Contents

CONTENTS	3
INTRODUCTION	4
SAMPLE FILES	4
TUTORIAL	4
Filter for Transposon-Containing Reads	4
Assembling a Transposon Consensus	8
Identify and Save Reads for Each Genomic Insertion	11
Assemble Filtered Sets of Reads	16
Identifying Transposition Sites on the Genome	19
What Genes Have the Transpositions Interrupted?	22
Shortcuts!	24

Introduction

One common approach to investigating gene function is to randomly generate knockout mutations and screen for desired phenotypes. In bacteria, this is frequently done using transposons, typically carrying a selectable marker so that only those clones that take up the transposon can survive. With the advent of high throughput Next Generation Sequencing (NGS), it is often easiest and most cost-effective to simply sequence the entire genome of each isolate and then determine via software where in the genome the transposon insertion has occurred.

One trick to make it even cheaper to perform this analysis is to pool multiple clones into a single sequencing reaction and then use software to identify the different insertion locations. This tutorial demonstrates how you can use MacVector to identify the insertion locations of five individual transposition events pooled into a single NGS experiment.

Sample Files

This tutorial uses four files;

- (a) A complete annotated bacterial genome (Genome.nucl)
- (b) The sequence of the transposon used in the mutagenesis (Transposon.nucl)
- (c) A pair of Illumina MiSeq gzipped NGS data files resulting from the sequencing of 5 individual transposon isolate clones (Run_R1_001.fastq.gz and Run_R2_001.fastq.gz)

For the purposes of this tutorial, we have organized these into a minimal folder hierarchy as follows;

🚞 Insertior	AnalysisTutorial		
		😻 🗸 Q S	Search
Name	Date Modified	Size	Kind
genome.nucl	Today, 5:06 PM	2.5 MB	MacVector NA Sequence
🔻 📄 NGSData	Today, 6:22 PM		Folder
🖹 Run_R1_001.fastq.gz	Today, 11:01 AM	228.7 MB	gzip compressed archive
🖹 Run_R2_001.fastq.gz	Today, 11:01 AM	262.7 MB	gzip compressed archive
👔 Transposon.nucl	Today, 4:06 PM	37 KB	MacVector NA Sequence

The files can be downloaded using this link;

https://macvector.net/insertionanalysistutorialdata.zip

Once downloaded, extract the data and move the enclosing folder to a suitable location on your hard drive.

Tutorial

Filter for Transposon-Containing Reads

There is no need for us to assemble entire genomes for this analysis (though MacVector can do that if necessary). All we need are those reads that cross the junction between transposon and genome. If we can filter for those reads that match the transposon, then some of them will be at the ends and will cross into genomic sequence. We **could** search specifically for reads that match the ends of the transposon, but realistically there is no need to do that.

We can accomplish this using the **Database | Align to Folder** function in MacVector. This will scan folders on your file system containing sequences (including any reads in gzipped fasta and fastq files) searching for matches to an input sequence. You can then retrieve the matching sequences/reads of interest into a much smaller fasta/fastq file. One really nice thing about the function is that it is paired-end read aware, so if one read of a pair matches the query sequence, the other read of the pair will also be retrieved. For this experiment, if one of the pair lies within the transposon and the other lies outside, both will be retrieved, allowing us to pull out adjacent genome sequence even if neither read crossed the junction.

The first step is to open the Transposon.nucl sequence and invoke Database | Align to Folder.

Search Folder
Search in enclosed folders (recursively) Folder contains paired-end reads
Options
Hash Value: 12 Processing: Align
Scoring Matrix
DNA identity with penalties matrix.nmat
Open Choose
Region
1 to 1894
Defaults Cancel OK

There are a number of settings here that you need to change from the defaults;

Search Folder: click on Choose and select the folder containing your data.

Folder contains paired-end reads: this NGS data is paired-end, and selecting this helps tremendously with the analysis.

Hash Value: Its usually good to set this to the maximum unless you have less than 4 GB RAM or if your NGS data is particularly noisy (e.g. PacBio or Nanopore). Setting it to 12 means there must be at least one run of 12 perfect matches between the query and a read before MacVector will initiate an alignment.

Scores to Keep: You want to keep as many hits as you can, so set it to at least 10,000.

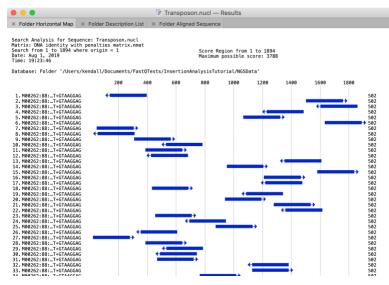
Scoring Matrix: For this type of analysis, you would expect very high identity between reads and the transposon. By using the DNA Identity with Penalties.nmat matrix, this will give priority to reads with short perfect matches at the ends of the query sequence (i.e. the junction reads we really want) versus those that have weak end-to-end similarity elsewhere in the query.

When you click **OK**, the search will run. The length of time this takes is dependent primarily on the size of your NGS data set and to some extent on the size of the query sequence and the speed of your machine. However, this is one case where the amount of RAM on your machine makes little difference as the reads are streamed into memory a few at a time and thus use very little RAM. In this experiment, the fastq files each contain 1.28 million Illumina MiSeq reads of

251 nt each and the search transposon is just under 2,000 bp. On a relatively high end 2018 i9 MacBook Pro with 32 GB RAM, the search completes in under 15 minutes;

ummary		Filter Options
Entries:	29973856 2509856 13:46.70	Entries to show: 1 to 1404
Processing Time: 00:	00:00.25	Score Region: 1 to 1894
Matches Saved: Matches Trimmed: Lowest Score Retained:	1404 0 76	Display Region: 1 to 1894
Significant: Probable:	0 0	Display Options
Possible:	967	Description list
Matches Aligned:	1404	V Horizontal map
		Aligned sequences
		Defaults Cancel OK

We want to see all of the **Display Options**, so check all the boxes.



We don't care about the Folder Horizontal Map tab too much - it gives an overview of the locations of the hits. As expected, the best hits land squarely in the middle of the transposon.

You can scroll through the **Folder Aligned Sequences** tab – it shows the text alignment of each read with the transposon. This can be a lot of text. As you explore you will see a number of reads that match the 5' end of the transposon and clearly extend to the left of the sequence. Those are the reads that we really want, but we don't have to specifically search them out at this point. What is most interesting is the **Folder Description List** tab;

• • •	🕅 Transp	oson.nucl — Results			
imes Folder Horizontal Map Description List	imes Folder Description List	imes Folder Aligned Sequence			
Search Analysis for Sequ Matrix: DNA identity wit Search from 1 to 1894 wh Date: Aug 1, 2019 19:23:46	h penalties matrix.nmat	Score Region from 1 t Maximum possible scor			
Database: Folder '/Users	/kendall/Documents/FastQ1	ests/InsertionAnalysisTutor	ial/NGS	Data'	
		Database Retrieve To Disk o a single fasta or fasto fi		the mate	hing sequences,
or Database Retrieve	to File to save them into	a single fasta or fastq fi	le)		
Sequence			Opt.	Init.	Description
		2 1:N:0:GGACTCCT+GTAAGGAG	502	502	
		55 1:N:0:GGACTCCT+GTAAGGAG 55 2:N:0:GGACTCCT+GTAAGGAG	502 502	502 502	
		1 1:N:0:GGACTCCT+GTAAGGAG	502	502	
		1 2:N:0:GGACTCCT+GTAAGGAG	502	502	
		<pre>I9 1:N:0:GGACTCCT+GTAAGGAG</pre>	502	502	
	00-CH79D:1:1101:14404:5897		502	502	
	00-CH79D:1:1101:14404:5897		502	502	
		5 1:N:0:GGACTCCT+GTAAGGAG 5 2:N:0:GGACTCCT+GTAAGGAG	502 502	502 502	
		3 1:N:0:GGACTCCT+GTAAGGAG	502	502	
		3 2:N:0:GGACTCCT+GTAAGGAG	502	502	
	0-CH79D:1:1101:16169:6395		502	502	
	0-CH79D:1:1101:16169:6395		502	502	
		<pre>'9 1:N:0:GGACTCCT+GTAAGGAG</pre>	502	502	
		<pre>1:N:0:GGACTCCT+GTAAGGAG</pre>	502	502	
		08 2:N:0:GGACTCCT+GTAAGGAG 4 1:N:0:GGACTCCT+GTAAGGAG	502 502	502 502	
		4 1:N:0:GGACTCCT+GTAAGGAG 6 1:N:0:GGACTCCT+GTAAGGAG	502 502	502	
		6 2:N:0:GGACTCCT+GTAAGGAG	502	502	
		8 1:N:0:GGACTCCT+GTAAGGAG	502	502	
33 NOO3C3.00.0000000		0 3.N.O.CCACTCCT.CTAACCAC	500	505	

While this looks like a plain text window, it is quite interactive. If you select one or more lines (even just a partial line as shown above) MacVector knows that you are interested in those sequences and will activate a number of menu items in the **Database** menu;

Retrieve and Open – this will retrieve the individual sequences or reads and open each in a separate window. Not recommended for more than 10 or so sequences!

Retrieve to Disk – retrieves and saves each sequence or read as a separate file in a folder you can designate.

Retrieve to File – retrieves and saves each sequence or read into a single fasta or fastq file.

The great thing about **Retrieve to File** is that it is paired-end read aware, so in cases like this it will actually save the reads into a pair of files that have "-1" and "-2" added to the filename.

You can optionally carefully scroll down the **Description List**, finding just those files that you want to retrieve, but for our purposes, the easiest thing is to choose **Edit | Select All**, followed by **Database | Retrieve to File** and give the file(s) a suitable format and name;

	📒 InsertionAna	IlysisTutorial 🗘 🔨	Q Search	
Favorites	Name	Date Modified	 Size 	Kind
Stropbox	NGSData	Today, 6:22 PM		Folder
÷ .	Genome.nucl			
A Google Drive	🔯 Transposon.nucl		37 KB	
MelissaMadsenNGS				
Europe Cordner				

When you press **Save**, you'll find that two files have been created – these are a matched pair of files where each entry in "-1" has its pair in "-2". Many algorithms (*Bowtie*, *Velvet*, *SPAdes*) can take advantage of the paired reads in the files to generate better alignments.

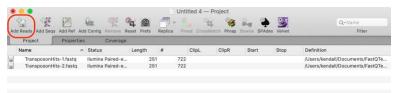
	rtionAnalysisTutorial	🕏 👻 🔍 Sear	ch
Name	Date Modified	Size	Kind
Genome.nucl	Today, 5:06 PM	2.5 MB	MacVector NA Sequence
🔻 🛅 NGSData	Today, 6:22 PM		Folder
Run_R1_001.fastq.gz	Today, 11:01 AM	228.7 MB	gzip compressed archive
Run_R2_001.fastq.gz	Today, 11:01 AM	262.7 MB	gzip compressed archive
👔 Transposon.nucl	Today, 4:06 PM	37 KB	MacVector NA Sequence
📄 TransposonHits-1.fastq	Today, 8:21 PM	464 KB	FASTQ Sequence
TransposonHits-2.fastq	Today, 8:21 PM	464 KB	FASTQ Sequence

Assembling a Transposon Consensus

Now that we have a filtered collection of reads that match the transposon, we need a strategy to identify the reads that overlap the insertion locations in the genome. The easiest way to do this is to simply assemble all of the reads – the transposon should assemble very cleanly, but adjacent to the ends should be a mix of the five genome insertion locations. Those areas should have a lot of misaligned sequences which we can use to identify the genomic flanking sequences where the transposon inserted.

First, we create a File | New | Assembly Project (you'll need to have the *Assembler* module for this). Then we have two options for adding the Transposon Hits;

(a) Add Reads: If you have a reasonable number of hits (this data set has over 700 reads in each file) you should click on the Add Reads toolbar button – the reads will be added as references to disk files;



NGS-based assembly algorithms such as *Velvet* and *SPAdes* work better with pairs of files, rather than individual sequences. However, they do require a reasonable coverage level (10x or more) so, if you have less than that, you may prefer to import the data as individual sequences;

(b) Add Seqs: This will import the reads as individual sequences, as long as there are less than 5,000 sequences in each file, otherwise they will appear as disk-based references. When you have very few reads/sequences, it's typically better to use *phrap* as an assembly algorithm. While the MacVector implementation of *phrap* can use references to fasta/fastq files, its often easier and cleaner to see the individual reads, particularly when you have less than 50 reads to assemble (more on that later) or if you have a mix of long contigs and short "patch" sequences to assemble.

		97 G.	05 Q	a =.	d 5 — Project		8 9		20	0~	Name	
		as Add Ref Add Cont	tig Remove Reset	Prefs Replica	Phred CrossM	atob Dhran	Bowtie SP	Ades V	elvet	~	Filter	
Auu	Project	Properties	Coverage	Preis Replica	Place Clossin	aton Phrap	Bowne 3P	ues y	circu	-	Filter	-
	Name					Status	Length	#	ClipL	^	ClipR	Sta
24	M00262:88	3:000000000-CH79D	:1:1101:2521:14656	1:N:0:GGACTCC	T+GTAAGGAG			251		1	25	
60		3:000000000-CH79D						251		1	25	
2		3:000000000-CH79D						251		1	25	
20		3:000000000-CH79D						251		1	25	
60		3:000000000-CH79D						251		1	25	
6		3:000000000-CH79D						251		1	25	
24		3:000000000-CH79D						251		1	25	
50		3:000000000-CH79D						251		1	25	
64		3:000000000-CH79D						251		1	25	
64		3:000000000-CH79D						251		1	25	
64	M00262:88	3:000000000-CH79D	:1:1101:8262:16215	1:N:0:GGACTCC	T+GTAAGGAG			251		1	25	i i
20	M00262:88	3:000000000-CH79D	:1:1101:8262:16215	2:N:0:GGACTCO	CT+GTAAGGAG			251		1	25	È
50	M00262:88	3:000000000-CH79D	:1:1101:8701:24218	1:N:0:GGACTCC	T+GTAAGGAG			251		1	25	1
64	M00262:88	3:000000000-CH79D	:1:1101:8701:24218	2:N:0:GGACTCO	CT+GTAAGGAG			251		1	25	E.
64	M00262:88	3:000000000-CH79D	:1:1101:10034:1002	8 1:N:0:GGACTC	CT+GTAAGGAG			251		1	25	1
50	M00262:88	3:000000000-CH79D	:1:1101:10034:1002	8 2:N:0:GGACTO	CCT+GTAAGGAG	3	1	251		1	25	i.
20	M00262:88	3:000000000-CH79D	:1:1101:10413:2727	5 1:N:0:GGACTC	CT+GTAAGGAG			251		1	25	È
24	M00262:88	3:000000000-CH79D	:1:1101:10413:2727	5 2:N:0:GGACTC	CT+GTAAGGAG	ř.		251		1	25	É
2	M00262:88	3:000000000-CH79D	:1:1101:10829:1932	2 1:N:0:GGACTC	CT+GTAAGGAG			251		1	25	1
24	M00262:88	3:00000000-CH79D	:1:1101:10829:1932	2_2:N:0:GGACTO	CCT+GTAAGGAG	3		251		1	25	Ú.
2	M00262:88	3:000000000-CH79D	:1:1101:10907:1315	5_1:N:0:GGACTC	CT+GTAAGGAG			251		1	25	i l
M	M00262:88	3:00000000-CH79D	:1:1101:10907:1315	5_2:N:0:GGACTC	CT+GTAAGGAG		1	251		1	25	É.
36	M00262:88	3:00000000-CH79D	:1:1101:11317:2467	B_1:N:0:GGACTCO	CT+GTAAGGAG		1	251		1	25	i.
24	M00262:88	3:00000000-CH79D	:1:1101:11317:2467	B_2:N:0:GGACTC	CT+GTAAGGAG		1	251		1	25	6
24	M00262:88	3:000000000-CH79D	:1:1101:11419:20121	1:N:0:GGACTCO	CT+GTAAGGAG		1	251		1	25	i i
14	M00262:88	3:00000000-CH79D	:1:1101:11419:20121	2:N:0:GGACTC	CT+GTAAGGAG		1	251		1	25	É
24	M00262:88	3:000000000-CH79D	:1:1101:13263:1324	9_1:N:0:GGACTC	CT+GTAAGGAG		1	251		1	25	i i
W.	M00262:88	3:000000000-CH79D	:1:1101:13263:1324	9_2:N:0:GGACTO	CCT+GTAAGGAG	5	1	251		1	25	Ê
¥	M00262:88	3:000000000-CH79D	:1:1101:13527:1876	2_1:N:0:GGACTC	CT+GTAAGGAG		1	251		1	25	i.
1	M00262:88	3:00000000-CH79D	:1:1101:13527:1876	2_2:N:0:GGACTC	CT+GTAAGGAG	1	4	251		1	25	Ē
24	M00262:88	3:00000000-CH79D	:1:1101:13971:6683	1:N:0:GGACTCC	T+GTAAGGAG		2	251		1	25	i.
	1100000000		4.4404.40074.0000					154		4	05	ř.

For this data, with 2 x 700 reads, we can import as "Reads" and assemble using *SPAdes*. So, we select the two fastq.gz files and click on the **SPAdes** toolbar item.

Read pre-processing
Discard reads less than 33 nt
Trim ends with quality less than
Trim N's from ends
Discard short reads that contain any N's
SPAdes Options
Override coverage cutoff: 5
Use custom K-MER: 31,41,51,61,71,81,91,101,
Threads: 11 C Enter odd values less than 128 in ascending order, separated by commas (e.g. 21,33,55).
Use "careful" mode with MismatchCorrector (slow)
Generate Alignments Using Bowtie
Threads: 11
? Defaults Cancel OK

The defaults are basically fine, but click on the **Generate Alignments Using Bowtie** checkbox so we get to see the actual reads aligned to the consensus. When you click **OK**, the *SPAdes* alignment will run. With a relatively small number of reads like this (2×700) that assemble into a short (~3 kb) sequence, the assembly only takes about 30 seconds. When complete, a new *SPAdes* job appears in the *Assembly Project*;

	e eads Add Segs Add Ref Add Co	- (- - - -	Prefs Reg] ~ 🔍	led 4 — Pr	-	Bowtie	SPAdes	Velvet		Q~Name Filter
uu	Project Properties	Coverage	FIEI3 NO	Plied	Crossinater	- Fillap	Downe	JF Ades	VOIVOL		T Intel
	Name	Status	Length	#	ClipL /	 ClipR 		Start	Stop		Definition
1	TransposonHits-1.fastq	Ilumina Paired-e	251	722							/Users/kendall/Documents/FastQTests/
	TransposonHits-2.fastq	Ilumina Paired-e	251	722							/Users/kendall/Documents/FastQTests/
	▼ SPAdes TransposonHits-x										
	Unaligned_Reads_1_1.fq.gz	Ilumina Paired-e	251	164							/var/folders/z5/djh03ch97tvdycgx_v95
	Unaligned_Reads_1_2.fq.gz	Ilumina Paired-e	251	164							/var/folders/z5/djh03ch97tvdycgx_v95
	NODE_1_length_2737_cov		2737	1413		1	2737		1	2737	

SPAdes contigs are always named using "NODE_X_...." nomenclature. In this case we have a single contig. If we double-click on the NODE_1_... item, a contig window appears;

			NODE	_1_length_2737_c	ov_56.92911	9 Contig 1	— Editor				
A 1		-	0	16CA 16CA	AGC						
Locked Text View	Prefs	Replica '	Translations W	//XX							
Editor		Мар	Features	Annotations	Summa						
	-				1	,					
	Conser	376 ISUS TGAATTA		390 400 ITAAGTGCTTTATCACA	410 ATTTTCAGACAAO	420 ACTCGTGCAT	430 AATAACAGGT	440 TGGCTGATA	450 AGTCCCCGGTC	460 TAAGCTTGC	470 ATGCCTGC
M00262:88:0008:21	1499:823	74 🕨	ATAT	TTAAGTGCTTTATCACA	ATTTTCAGACAAG	ACTCGTGCAT	AATAACAGGT	TGGCTGATA	AGTCCCCGGTC	TAAGCTTGC	ATGCCTGC
M00262:88:005:224	475:218	19 4	TAT	TTAAGTGCTTTATCACA	ATTTTCAGACAAG	ACTCGTGCAT	AATAACAGGT	TGGCTGATA	AGTCCCCGGTC	TAAGCTTGC	ATGCCTGC
M00262:88:0011:28	8165:975	54 🕨		GTGCTTTATCACA	ATTTTCAGACAAG	ACTCGTGCAT	AATAACAGGT	TGGCTGATA	AGTCCCCGGTC	TAAGCTTGC	АТБССТБС
M00262:88:00102:6	6834:502	26 4		TCACA	ATTTTCAGACAAG	ACTCGTGCAT	AATAACAGGT	TGGCTGATA	AGTCCCCGGTC	TAAGCTTGC	ATGCCTGC
M00262:88:003:209	923:1490	07 🔺		TCACA	ATTTTCAGACAAG	ACTCGTGCAT	AATAACAGGT	TGGCTGATA	AGTCCCCGGTC	TAAGCTTGC	АТБССТБС
M00262:88:000:125	577:142	23 ┥			ATTTTCAGACAAG	ACTCGTGCAT	AATAACAGGT	TGGCTGATA	AGTCCCCGGTC	TAAGCTTGC	АТБССТБС
M00262:88:00103:	9229:79	90 b			GACAAG	ACTCGTGCAT	AATAACAGGT	TGGCTGATA	AGTCCCCGGTC	TAAGCTTGC	АТБССТБС
M00262:88:004:179	954:183	54 🕨			GACAAG	ACTCGTGCAT	AATAACAGGT	TGGCTGATA	AGTCCCCGGTC	TAAGCTTGC	АТБССТБС
M00262:88:004:179	954:183	54 📢 GATAGGA	AGTCCGTCTCGTGGGG	CTCGGAGATGTGTATAA	GAGACAG GACAAG	ACTCGTGCAT	AATAACAGGT	TGGCTGATA	AGTCCCCGGTC	TAAGCTTGC	ATGCCTGC
M00262:88:0012:82	290:232	50 🕨				GTGCAT	AATAACAGGT	TGGCTGATA	AGTCCCCGGTC	TAAGCTTGC	ATGCCTGC
M00262:88:0012:82	290:232	50 🖌 🛛 CATA	ACGAGATAGGAGTCCO	STCTCGTGGGCTCGGAG	ATGTGTATAAGAG	ACAGGTGCAT	AATAACAGGT	TGGCTGATA	AGTCCCCGGTC	TAAGCTTGC	ATGCCTGC
M00262:88:008:237	765:180	52 🕨				1	AATAACAGGT	TGGCTGATA	AGTCCCCGGTC	TAAGCTTGC	ATGCCTGC
M00262:88:008:237	765:180	52 📢 CGCCCAC	CGAGAGCTACACGTA	AAGGAGTCGTCGGCAGC	GTCAGATGTGTAT	AAGAGACAG	AATAACAGGT	TGGCTGATA	AGTCCCCGGTC	TAAGCTTGC	ATGCCTGC
M00262:88:003:254	496:1740	08 ▶	ATAT	FATGAGTCGGTTAGTTA	TGAGAATTTAGCA	TCAGAAAAAA	TATAACAGGT	TGGCTGATA	AGTCCCCGGTC	TAAGCTTGC	ATGCCTGC
M00262:88:004:19	588:105	72 🕨				ACC	TATAACAGGT	TGGCTGATA	AGTCCCCGGTC	TAAGCTTGC	ATGCCTGC
M00262:88:007:14	548:1860	03 ▶	CTATAT	FATGAGTCGGTTAGTTA	TGAGAATTTAGCA	TCAGAAAAAA	TATAACAGGT	TGGCTGATA	AGTCCCCGGTC	TAAGCTTGC	ATGCCTGC
M00262:88:003:213	356:130	32 ▶ ACTCTAT	TTTTAGAATTTTTTAT	TGTCCACTTCTCCTCA	AGTGTTTTACTTA	CTTTTTTATG	TATAACAGGT	TGGCTGATA	AGTCCCCGGTC	TAAGCTTGC	ATGCCTGC
M00262:88:0003:77	717:193	51 🕨 АСТСТАТ	TTTTAGAATTTTTTAT	TGTCCACTTCTCCTCA	AGTGTTTTACTTA	CTTTTTTATG	TATAACAGGT	TGGCTGATA	AGTCCCCGGTC	TAAGCTTGC	ATGCCTGC
M00262:88:003:126	634:2324	15 ▶ AGACCTT	TTATAAAGGTAAGAG	TAGCAGTACTTACTGA	GAGTTTACTTTT	ATAATATACO	TATAACAGGT	TGGCTGATA	AGTCCCCGGTC	TAAGCTTGC	ATGCCTGC
M00262:88:004:270	080:1419	32 ▶ ACTCTAT	TTTTAGAATTTTTTAT	TGTCCACTTCTCCTCA	AGTGTTTTACTTA	CTTTTTTATO	TATAACAGGT	TGGCTGATA	AGTCCCCGGTC	TAAGCTTGC	ATGCCTGC
M00262:88:0005:21	198:134:	29 🕨	CTATAT	TATGAGTCGGTTAGTTA	TGAGAATTTAGCA	TCAGAAAAAA	TATAACAGGT	TGGCTGATA	AGTCCCCGGTC	TAAGCTTGC	ATGCCTGC
M00262:88:004:141	133:146	59 🕨				ACC	TATAACAGGT	TGGCTGATA	AGTCCCCGGTC	TAAGCTTGC	ATGCCTGC
M00262:88:0012:10	0319:93	75 🕨 AGACCTI	TTATAAAGGTAAGAG	TAGCAGTACTTACTGA	GAGTTTACTTTTC	ATAATATACO	TATAACAGGT	TGGCTGATA	AGTCCCCGGTC	TAAGCTTGC	ATGCCTGC
M00262:88:0002:23	316:172	38 ▶ TTATCTA	νΑΤΤΤΑΤΟΤΑΟΤΑΤΑΙ	TATGAGTCGGTTAGTTA	TGAGAATTTAGCA	TCAGAAAAAA	TATAACAGGT	TGGCTGATA	AGTCCCCGGTC	TAAGCTTGC	ATGCCTGC
M00262:88:007:230					GAGTTTACTTTTC	ATAATATACO	TATAACAGGT	TGGCTGATA	AGTCCCCGGTC	TAAGCTTGC	ATGCCTGC
M00262:88:001:27			TTATAAAGGTAAGAG	TAGCAGTACTTACTGA	GAGTTTACTTTTC						
M00262:88:0001:85						ACC	TATAACAGGT	TGGCTGATA	AGTCCCCGGTC	TAAGCTTGC	ATGCCTGC
M00262:88:006:139							ATAACAGGT	TGGCTGATA	AGTCCCCGGTC	TAAGCTTGC	ATGCCTGC
M00262:88:006:129									AGTCCCCGGTC		
M00262:88:0013:14				CAGTACTTACTGA	GAGTTTACTTTTC						
M00262:88:0014:15	5341:518	37 🕨				TAATATACO	TATAACAGGT	TGGCTGATA	AGTCCCCGGTC	TAAGCTTGC	ATGCCTGC

If you scroll through the contig, you will eventually reach a region where many of the reads have greyed out areas to the left of a cutoff point, where to the right they are in bold text, as shown above. This is the left end of the contig - the greyed-out sequences should represent the flanking regions of the other insertion sites. The bold sequences at the top of the screenshot to the left of the inflection location also represent genomic flanking sequence – the *SPAdes* assembler will randomly choose one of the flanking sequences (typically the one with the most reads) and use that as the consensus.

While we could directly use this contig to determine the various flanking genomic regions of the transposon insertions, we will save the consensus sequence, then use MacVector's **Align to Reference** functions for further analysis. If you right-click (<ctrl>-click on a trackpad) a popup menu appears;



Choose **Export Consensus without Gaps** and choose a suitable filename to save the consensus sequence (e.g. Consensus.nucl).

Identify and Save Reads for Each Genomic Insertion

For the next step we will align all of the transposon-containing reads (and their pairs) to the saved consensus sequence, then examine the junction between the transposon and genomic sequence to pull out reads representing each of the five insertion sites.

First, open the saved Consensus.nucl then choose Analyze | Align to Reference. Click on the Add Seqs toolbar button and select the TransposonHits-1.fastq and TransposonHits-2.fastq files. The Editor tab should look something like this;

					Consense	us Aligi	nment — E	ditor					
0 5	6	-	2.	122	073	周幻	0 ~	AGCT	ACGTAC	ACETAC		-	
Unlocked Text View	1000CL	Replica	Topology	Add Soor	Remove Seq	Alian	Translations		Eiret Mismatek	Next Mismatch	n Width		
		Constant Sector and					1001000000000	DUIS			i widen		_
Editor	M	ap	Featu	ires	Annotation	s	Text	1	SNPs	Problems			
Sort			10	20	30	49	50	68		80	90	100	1.
	Consensu		CGITICCGC	CIATATAT	TIGCCGGAATA	AAGCTAC	TITTACICC	IIGCIIC	AAAGAAGAAAATA	ACTGTTCCGCTAT	AATTACTATT	TTATCATGAT	I CAA
M00262:88:00+GTAA			ATCAAGCCI	TACGGTCA	CGTAACCAGCA	AATCAAT	ATCACTGTGT	GGCTTCA	GECCECCATCCA	TGCGGAGCCGTAC	AAATGTACGG	CCAGCAACG	TCG
M00262:88:00+GTAA										CCGCAGTGGATGG			
M00262:88:00+GTAA										TATACGAAGTTACG			
M00262:88:00+GTAA	GGAG									TCCCTGATGTTAC			
M00262:88:00+GTAA	GGAG									TAGTGCTTACGTT			
M00262:88:00+GTAA	GGAG	GTGGTTA	TTTAAGAAA	GGAACCCA	TAAGCTCACAAG	CCCCATA	AAAAGTGCAA	CGTAAAA	TAAAATATATAAA	GCAGATTTATAAG	CCCAATCTTG	TGTTGTCTT	TTC
M00262:88:00+GTAA	GGAG	CTCATGA	CCATTAACO	CTTCCTTC	ATGTGCACCTT	AAATCTO	ATAAACTCTT	TGATGAT	CGCCATATTATC	TCCTCTCCCTTGC	TGACCATTAT	AATATCCCT	TATO
M00262:88:00+GTAA	GGAG	CGTATAA	AGTATCCTA	TACGAAGT	TACGGCTCCTCC	TAGAACO	ATCGCCGCAT	GCTAGCA	TAAAACGCGCGCG	CTAACAAGTGCCC	AGCGCGCTGC	СТСАЛАЛАЛ	тсто
M00262:88:00+GTA/	GGAG	GTGATAT	TGATTTGCT	GGTTACGG	TGACCGTAAGGC	TTGATGA	AACAACGCGG	CGAGCTT	TGATCAACGACCT	TTTGGAAACTTCG	GCTTCCCCTG	GAGAGAGCGA	AGAT
M00262:88:00+GTAA	GGAG	ATCGTGG	CTGGCTCGA	AGATACCT	SCAAGAATGTCA	TTGCGCT	GCCATTCTCC	AAATTGC	AGTTCGCGCTTA	CTGGATAACGCCA	CGGAATGATG	тсетсетес	ACA
M00262:88:00+GTAA	GGAG	AAGGAAG	ACCGTATGA	AGGCACAC	AGACGGCTAAAC	TTAAGGT	CACAAAAGGC	GGACCGC	TTCCATTCGCGTG	GGATATTCTTTCA	CCGCAATTTA	TGTATGGTTO	CTA
M00262:88:00+GTAA	GGAG	GGCCTCC	ATCTTTCAG	CTTCAGGC	STTGCTTGATTT	CGCCTTI	AAGTGCTCCG	тсстсса	GGTACATTCTCTC	GCTGCTGGCTTCC	CAGCCCATCG	тттсттст	GCAT
M00262:88:00+GTAA	GGAG	ATCATAC	ATAATGTTA	ATGCAACA	STTATTTAATGT	ATGGTTG	TAATAATAAA	ACTGTAA	ATGGTATGCTAGA	ТТАААААТААААТ	CATCAATAAA	TATCCGGGA	AACT
M00262:88:00+GTAA	GGAG	ATTCTAT	TGCTCCTT	TTGGGACG	TAACGAAACTT	GGTTAGT	ACTTGGCGGC	GGTGGTT	TATTTGCGGCTT	CCCGTTGGCGTAT	TCGCTATTAA	TGCCGGCTT	TATA
M00262:88:00+GTAA	GGAG	ACTITAT	ACGAAGTTA	TCAGACCG	SGGACTTATCAG	CCAACCT	GTTATTATT	ATAATTA	TACTAACGATGT	ATCTGCACATCGT	CTTTCAATGA	GATATAATA	TAT
M00262:88:00+GTAA	GGAG	ATTCTGT	ACCGAGTTI	AGCACCGG	ACGAAGATTTAA	ATGATCA	TATTGTAGGA	CATCTAG	GTCGTAGAGATT	TGTATTAATTAAT	CCTTTAACAA	GCAGTCTTG	ATG
M00262:88:00+GTAA	GGAG	CCTTCAT	ACGGTCTTC	CTTCGCCT	TCGCCTTCGATT	TCAAACT	CATGACCATT	AACGCTT	CCTTCCATGTGCA	CCTTAAATCTCAT	AAACTCTTTG	ATGATCGCCA	ATAT
M00262:88:00+GTAA	GGAG	GCATACO	ATTTACAGI	TITATTAT	TACAACCATACA	TTAAATA	ACTGTTGCAT	TAACATT	ATGTATGATTTA	CCTAAGTTATCTG	AGTAACATAA	GGGATATTA	TAAT
M00262:88:00+GTAA	GGAG	CTTCGTA	TAGTACACA	TTATACGA	AGTTATCTCTAG	ATTATT	GCCGACTACC	TTGGTGA	TCTCGCCTTTCAC	GTAGTGAACAAAT	TCTTCCAACT	GATCTGCGCG	GCG
M00262:88:00+GTAA	GGAG	GTCCCGC	ATTTGGTAC	AGCGCAGT	ACCEGECAMAAT	cacacca	AAGGATGTCG	CTGCCGA	CTGGGCAATGGAG	CGCCTGCCGGCCC	AGTATCAGCC	CGTCATACT	TGA
M00262:88:00+GTAA	GGAG	CGTAAGO	ACTACATTI	CGCTCATCO	SCCAGCCCAGTC	GGGCGGG	GAGTTCCATA	GCGTTAA	GGTTTCATTTAG	GCCTCAAATAGAT	CCTGTTCAGG	AACCGGATCA	AAA
M00262:88:00+GTAA	GGAG	TTTGTAC	GGCTCCGCA	GTGGATGG	GGCCTGAAGCC	ACACAGT	GATATTGATT	TGCTGGT	TACGGTGACCGTA	AGGCTTGATGAAA	CAACGCGGCG	AGCTTTGAT	CAA
M00262:88:00+GTAA	GGAG	AATAGAT	CCTGTTCAG	GAACCGGA	TCAAAGAGTTCC	TCCGCCG	CTGGACCTAC	CAAGGCA	ACGCTATGTTCTC	TTGCTTTTGTCAG	CAAGATAGCC	AGATCAATG	TCG
M00262:88:00+GTAA	GGAG	ATCAACG	ACCTTTTGG	AAACTTCG	GCTTCCCCTGGA	GAGAGCO	AGATTCTCCG	CGCTGTA	GAAGTCACCATTO	TTGTGCACGACGA	CATCATTCCG	TGGCGTTATC	CCA
M00262:88:00+GTAA	GGAG	ATATTAT	AATGGTCAG	CAAGGGAG	AGGAAGATAATA	TGGCGAT	CATCAAAGAG	TTTATGA	GATTTAAGGTGCA	CATGGAAGGAAGC	GTTAATGGTC	ATGAGTTTGA	AAA
M00262:88:00+GTAA	AGGAG	ATCGTAT	AGTETTEAT	TATGTGAC	STAATATCCAGC	TTGATAT	TGACGTTGTA	TGCTCCA	GGTAATTGTACCO	GTTTCTTAGCTTT	GTATGTCGTC	TTGACCTCG	GCA
M00262:88:00+GTAA	AGGAG	CCCTTGC	TGACCATTA	TAATATCC	CTTATGTTACTC	AGATAAC	TTAGGATAAA	TCATACA	TAATGTTAATGCA	ACAGTTATTTAAT	GTATGGTTGT	AATAATAAAA	ACTO
M00262:88:00+GTAA	AGGAG	CGTTATG	AGAATGCAG	AAGAGGAT	GCATATCTCAC	TTTCTTG	AGCATATGGC	TTTTAAA	GGTACAAAAACTA	GAACGGCAAAACA	GATAGCGGAA	GCGTTTGAC	тстл
M00262:88:00+GTAA	GGAG	GTTTAAC	TTTATAAAI	AAATTCTC	CATCTTGCAGTG	ATGAATO	CTGCGTCACC	GTAACAA	CGCCGCCATCTTC	GAAGTTCATGACG	CGTTCCCATT	TAAATCCTTC	ccaa
M00262:88:00+GTAA	GGAG	GTATGAA	GGCACACAG	ACGGCTAA	ACTTAAGGTCAC	AAAAGGO	GGACCGCTTC	CATTCGC	GTGGGATATTCT	TCACCGCAATTTA	TGTATGGTTC	TAAAGCCTA	TGTO

Now click on the Align toolbar button and set up the alignment parameters something like this – the only parameter I've changed is the Hash Value to help speed up the alignment;

Alignment Type:	Seq	uence Con	nfirmation ᅌ
Residue Scoring			Alignment Parameters
Mat	ch:	2	Hash Value: 12 ᅌ
Mismat	ch:	-3	Sensitivity: 6 🗘
Ambiguous Mat	ch:	0	Score Threshold: 50
Gap Pena	lty:	4	X Dropoff: 25
		Defaults	Cancel OK

After you click **OK**, the alignment should take just a few seconds. If you scroll through the alignment, you should eventually reach an area where many of the reads abruptly do not match the reference and the left ends of the reads are shown greyed out (you may have to scroll downwards as well as rightwards to see this);

					Consensu	is Alig	Inment — I	ditor				
_ ^		-	~		0B	23	0 ~	AGCT	ACGTAC	ACGTAC		
Unlocked Text View	Prefs	Replica	Topology		Remove Seqs	Align	Translation	5 Dots	_ •	Next Mismatch	Width	
Editor	N	lap	Featu	ires	Annotation		Text	-	SNPs	Problems		
Sort				,				1				
	onsensi	360 US AGAATTT	370 Igaattata	380 TCGGCAAGT	390 TATATTAAGTGCT			10 GACAAG	420 ACTCGTGCA-TAATA	430 44 A-ACAGGTTGGCTG		460 STCTAAGCTTO
		us AGAATTT	IGAATTATA	TCGGCAAGT	TATATTAAGTGCT	TTATC	ACAATTTTCA-	GACAAG	ACTCGTGCA-TAATA	A-ACAGGTTGGCTG	ATAAGTCCCCG	TCTAAGCTTO
M00262:88:00+GTAA		•							ACTCGTGCA-TAAT/			
M00262:88:00+GTAA		•		GT					ACTCGTGCA-TAATA			
M00262:88:00+GTAA		•							ACTCGTGCA-TAATA			
M00262:88:00+GTAA		•							ACTCGTGCA-TAATA			
M00262:88:00+GTAA		P			GTGCT				ACTCGTGCA-TAATA			
M00262:88:00+GTAA		4							ACTCGTGCA-TAAT/ ACTCGTGCA-TAAT/			
M00262:88:00+GTAA		4				10			ACTEGTGEA-TAATA			
M00262:88:00+GTAA		A CONTACC	LCATACCAC.						ACTEGTGEA-TAAT/			
M00262:88:00+GTAA		SCATACG							ACTEGTGEA-TAAT/			
M00262:88:00+GTAA									ACTCGTGCA-TAAT/			
M00262:88:00+GTAA					ACCTACACCO	AACCA			A-T-GTGTA-TAAG			
M00262:88:00+GTAA			CATAC	CACATACCI					ACAGGTGCA-TAAT/			
M00262:88:00+GTAA		•						100000		A-ACAGGTTGGCTG		
M00262:88:00+GTAA			SECECCAC	CGAGAGCT			SCAGOGTOAGA	TGTGTA	TAAGAGAGACAGTAATA			
M00262:88:00+GTAA)								A-ACAGGTTGGCTG		
M00262:88:00+GTAA				ATTTATCT/	ACTATATATGAGT		AGTTATGAGAA	TTTAGC	ATCAGAAAAAATAT			
M00262:88:00+GTAA	GGAG	b			ATATATGAGT		AGTTATGAGAA	TTTAGC		A-ACAGGTTGGCTG	ATAAGTCCCCG	TCTAAGCTTO
M00262:88:00+GTAA	GGAG	•							ACCTATA	A-ACAGGTTGGCTG	ATAAGTCCCCG	TCTAAGCTTO
M00262:88:00+GTAA	GGAG	•							GTATA	A-ACAGGTTGGCTG	ATAAGTCCCCG	TCTAAGCTTO
M00262:88:00+GTAA	GGAG		GCGACCAC	AGAGATCT/	ACACGTAAGGAGT	CGTCG	GCAGCGTCAGA	TGTGTA	TAAGAGACAGGT ATA	A-ACAGGTTGGCTG	-ATAAGTCCCCG	GTCTAAGCTTO
M00262:88:00+GTAA	GGAG			ATTTATCT/	ACTATATATGAGT		AGTTATGAGAA	TTTAGC		A-ACAGGTTGGCTG	-ATAAGTCCCCG	GTCTAAGCTTO
M00262:88:00+GTAA	GGAG	•							ATA	A-ACAGGTTGGCTG	- ATAAGTCCCCG	STCTAAGCTTO
M00262:88:00+GTAA	GGAG	4			ACACGTACGGA		GGCAGCGTCA	GATGTG	TATAAGAGACAG AT	A-ACAGGTTGGCTG	ATAAGTCCCCG	STCTAAGCTTO
M00262:88:00+GTAA	GGAG	CATGGCA	АТАСТСТАТ	TTTAGAATI	TTTTTATTGTCCA	сттст	CTCAAGTGTT	TTACTT	ACTTTTTTATGT AT	A-ACAGGTTGGCTG	-ATAAGTCCCCG	STCTAAGCTTO
M00262:88:00+GTAA	GGAG	•							ACCTATA	A-ACAGGTTGGCTG	ATAAGTCCCCG	STCTAAGCTTO
M00262:88:00+GTAA	GGAG	•					AGCGTCAGATG	TGTATA	AGAGACAGACCT AT	A-ACAGGTTGGCTG	-ATAAGTCCCCG	STCTAAGCTTO
M00262:88:00+GTAA	GGAG	•			CTATATATGAGT	CGGTT	AGTTATGAGAA	TTTAGC	ATCAGAAAAAAT ATA	A-ACAGGTTGGCTG	-ATAAGTCCCCG	STCTAAGCTTO
M00262:88:00+GTAA	GGAG	•					GAGTTT	ACTITI	CATAATATACCT ATA	A-ACAGGTTGGCTG	-ATAAGTCCCCG	STCTAAGCTTO
M00262:88:00+GTAA	GGAG	▶ GGGTTTT	GAAGACCTT	TTATAAAGG	STAAGACTAGCAG	TACTT	ACTGAGAGTTT	ACTTTT	САТААТАТАССТ АТ	A-ACAGGTTGGCTG	-ATAAGTCCCCG	STCTAAGCTTC

Before proceeding further, you may want to save the alignment so that you can go back to it later. Choose File | Save As... and save the alignment with a suitable filename.

The reads towards the top of the screenshot are those that SPAdes used to create its preferred consensus and so align perfectly with the reference. These likely represent the insertion event that was most populous in the original pool of five, or that sequenced the most efficiently. In any event, we would like to select just those reads and save them into a separate file. Luckily, MacVector provides a simple way of doing this.

First, choose one of the reads that has perfect identity across the junction and select a few residues, ending one or two residues into the junction e.g.

```
TTTATCACAATTTTCA-GACAAGACTCGTGCA-TAATA-ACAGGTTGGCTG-ATA
TTTATCACAATTTTCA-GACAAGACTCGTGCA-TAATA-ACAGGTTGGCTG-ATA
   TCACAATTTTCA-GACAAGACTCGTGCA-TAATA-ACAGGTTGGCTG-ATA
   TCACAATTTTCA-GACAAGACTCGTGCA-TAATA-ACAGGTTGGCTG-ATA
        ATTTTCA-GACAAGACTCGTGCA-TAATA-ACAGGTTGGCTG-ATA
ATGTGTATAAGAGACAGGACAAGACTCGTGCA-TAATA-ACAGGTTGGCTG-ATA
                GACAAGACTCGTGCA-TAATA-ACAGGTTGGCTG-ATA
                GACAAGACTCGTGCA-TAATA-ACAGGTTGGCTG-ATA
CAAGGAGTCGTCGGCAGCGTCAGA-T-GTGTA-TAAGAGACAG---GGCTG-ATA
GGGCTCGGAGATGTGTATAAGAGACAGGTGCA-TAATA-ACAGGTTGGCTG-ATA
                          GTGCA-TAATA-ACAGGTTGGCTG-ATA
TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTAATA-ACAGGTTGGCTG-ATA
                                TAATA-ACAGGTTGGCTG-ATA
TCGGTTAGTTATGAGAATTTAGCATCAGAAAAAATATA-ACAGGTTGGCTG-ATA
TCGGTTAGTTATGAGAATTTAGCATCAGAAAAAATATA-ACAGGTTGGCTG-ATA
                              ACCTATA-ACAGGTTGGCTG-ATA
                                GTATA-ACAGGTTGGCTG-ATA
```

In this case I have selected a total of 7 residues (ignoring the gap) where the last two are the "AT" that appear to represent the left end of the transposon. Now we'd like to select all of the other reads that contain the same sequence aligned in the same position. Right-click (<ctrl>-click with a trackpad) and choose the Select Overlapping Reads Containing Selected Sequence in the context-sensitive menu;

Export Consensus with Gaps Export Consensus without Gaps
Align Selected Reads Delete Selected Reads Reset (un-align) Selected Reads
Export Selected Reads as FASTA Export Selected Reads as FASTQ Select Matching Pairs
Select Overlapping Reads Containing Selected Sequence
Extend Reference with Selected Read

Immediately, all of the reads containing that sequence at that position become selected;

					Consensu	s Alig	gnment — Ec	ditor				
		-	~		03	23	0 ~	AGCT	ACGTAC	ACGTAC		
Unlocked Text View	Prefs	Replica	Topology	Add Seqs	Remove Seqs	Align	Translations	Dots	_ · · •	Next Mismatch	Width	
Editor		Мар	Featu	ires	Annotations		Text	5	SNPs	Problems		
Sort		360	370	380	390	4	<i>00 41</i>	0	420	430 44	0 450	460
(Consens	GAATTT	TGAATTATA	TCGGCAAG	TATATTAAGTGCT	TATC	ACAATTTTCA-G	ACAAGA	CTCGTGCA-TAATA	A-ACAGGTTGGCTG	-ATAAGTCCCCGG	TCTAAGCTTO
M00262:88:00+GTA		SUS AGAATTT	TGAATTATA						CTCGTGCA-TAAT/			
M00262:88:00+GTA	AGGAG	•		G	TATATTAAGTGCT	TATC	ACAATTTTCA-G	ACAAGA	CTCGTGCA-TAATA	-ACAGGTTGGCTG	ATAAGTCCCCGG	TCTAAGCTTC
M00262:88:00+GTA	AGGAG	•			ATATTAAGTGCT	ТАТС	ACAATTTTCA-G	ACAAGA	CTCGTGCA-TAATA	-ACAGGTTGGCTG	ATAAGTCCCCGG	TCTAAGCTT
M00262:88:00+GTA	AGGAG	-			TATTAAGTGCT	ТАТС	ACAATTTTCA-G	ACAAGA	CTCGTGCA-TAATA	A-ACAGGTTGGCTG	-ATAAGTCCCCGG	TCTAAGCTTC
M00262:88:00+GTA	AGGAG	•			GTGCT	TATC	ACAATTTTCA-G	ACAAGA	CTCGTGCA-TAAT/	-ACAGGTTGGCTG	-ATAAGTCCCCGG	TCTAAGCTT
M00262:88:00+GTA	AGGAG	-				TC	ACAATTTTCA-G	ACAAGA	CTCGTGCA-TAAT/	-ACAGGTTGGCTG	-ATAAGTCCCCGG	TCTAAGCTTC
M00262:88:00+GTA	AGGAG	-				TC	ACAATTTTCA-G	ACAAGA	CTCGTGCA-TAAT/	A-ACAGGTTGGCTG	-ATAAGTCCCCGG	TCTAAGCTTC
M00262:88:00+GTA	AGGAG	4					ATTTTCA-G	ACAAGA	CTCGTGCA-TAAT/	-ACAGGTTGGCTG	-ATAAGTCCCCGG	TCTAAGCTTC
M00262:88:00+GTA	AGGAG	GCATACG					ATAAGAGACAGG	ACAAGA	CTCGTGCA-TAATA	A-ACAGGTTGGCTG	-ATAAGTCCCCGG	TCTAAGCTTC
M00262:88:00+GTA	AGGAG	•					Gi	ACAAGA	CTCGTGCA-TAAT/	A-ACAGGTTGGCTG	-ATAAGTCCCCGG	TCTAAGCTTC
M00262:88:00+GTA	AGGAG	•					Gi	ACAAGA	CTCGTGCA-TAAT/	A-ACAGGTTGGCTG	-ATAAGTCCCCGG	TCTAAGCTTC
M00262:88:00+GTA	AGGAG	•			ACCTACACGC/	AGGA	GTCGTCGGCAGC	GTCAGA	-T-GTGTA-TAAGA	GACAGGGCTG	-ATAAGTCCCCGG	TCTAAGCTTC
M00262:88:00+GTA		4	CATAC					AAG AGA	CAGGTGCA-TAATA			
M00262:88:00+GTA	AGGAG	•							GTGCA-TAAT/	A-ACAGGTTGGCTG	-ATAAGTCCCCGG	TCTAAGCTTC
M00262:88:00+GTA	AGGAG		GGCGCCCAC	CGAGAGCT/	ACACGTAAGGAGT	GTCG	GCAGCGTCAGAT	GTGTAT	AAGAGACAGTAAT/	A-ACAGGTTGGCTG	-ATAAGTCCCCGG	TCTAAGCTTC
M00262:88:00+GTA	AGGAG	•							TAAT	A-ACAGGTTGGCTG	-ATAAGTCCCCGG	TCTAAGCTTC
M00262:88:00+GTA	AGGAG	► ACACAGC	CATTATCTA	ATTTATCT/	ACTATATATGAGT		AGTTATGAGAAT	TTAGCA	TCAGAAAAAAT AT	A-ACAGGTTGGCTG	-ATAAGTCCCCGG	TCTAAGCTTC
M00262:88:00+GTA	AGGAG	•			ATATATGAGT		AGTTATGAGAAT	TTAGCA	TCAGAAAAAAT AT	A-ACAGGTTGGCTG	-ATAAGTCCCCGG	TCTAAGCTTC
M00262:88:00+GTA	AGGAG	•							ACCTATA	A-ACAGGTTGGCTG	-ATAAGTCCCCGG	TCTAAGCTTC
M00262:88:00+GTA	AGGAG	•							GTATA	A-ACAGGTTGGCTG	-ATAAGTCCCCGG	TCTAAGCTTC
M00262:88:00+GTA	AGGAG	ACGACGC	GGCGACCAC	AGAGATCT/	ACACGTAAGGAGT		GCAGCGTCAGAT	GTGTAT	AAGAGACAGGTATA	A-ACAGGTTGGCTG	-ATAAGTCCCCGG	TCTAAGCTTC
M00262:88:00+GTA	AGGAG		CATTATCTA	ATTTATCT/	ACTATATATGAGT		AGTTATGAGAAT	TTAGCA	TCAGAAAAAAT AT	A-ACAGGTTGGCTG	-ATAAGTCCCCGG	TCTAAGCTTC
M00262:88:00+GTA		•							AT/	A-ACAGGTTGGCTG	-ATAAGTCCCCGG	TCTAAGCTTC
M00262:88:00+GTA	AGGAG	•			ACACGTACGGA	STCGT	CGGCAGCGTCAG	ATGTGT	ATAAGAGACAG AT	A-ACAGGTTGGCTG	-ATAAGTCCCCGG	TCTAAGCTTC
M00262:88:00+GTA	AGGAG	CATGGCA	ATACTCTAT	TTTAGAAT	TTTTTATTGTCCA		CCTCAAGTGTTT	ТАСТТА	CTTTTTTATGTATA	-ACAGGTTGGCTG	-ATAAGTCCCCGG	TCTAAGCTTC
M00262:88:00+GTA	AGGAG								ACCTAT	-ACAGGTTGGCTG	-ATAAGTCCCCGG	TCTAAGCTTC
M00262:88:00+GTA	AGGAG	•					AGCGTCAGATGT	GTATAA	GAGACAGACCTAT	-ACAGGTTGGCTG	-ATAAGTCCCCGG	TCTAAGCTTC
M00262:88:00+GTA	AGGAG				CTATATATGAGT	GGTT	AGTTATGAGAAT	TTAGCA	TCAGAAAAAAT AT	-ACAGGTTGGCTG	-ATAAGTCCCCGG	TCTAAGCTTC
M00262:88:00+GTA	AGGAG						GAGTTTA		ATAATATACCTAT	-ACAGGTTGGCTG	-ATAAGTCCCCGG	TCTAAGCTTC
M00262:88:00+GTA	AGGAG	▶ SGGTTTT	GAAGACCTT	TTATAAAG	STAAGACTAGCAG	TACTT	ACTGAGAGTTTA		ATAATATACCTATA	A-ACAGGTTGGCTG	-ATAAGTCCCCGG	TCTAAGCTTC

Now, remember, we started with paired-end reads. Depending on orientation, some of the reads will have a pair lying further within the transposon, while others will have a partner lying further into the genomic sequence. It would be nice to select those too. MacVector to the rescue! Being careful not to change the selections, again right-click to bring up the context-sensitive menu. This time choose **Select Matching Pairs**.



You may not see a visible change on screen if none of the matching pairs are currently visible. However, the appropriate pairs will be selected, even if they are not visible and even if they did not align to the reference.

Finally, we would like to save the selected reads to a fasta or fastq file. Generally, we want to save in the same format we started with, in this case fastq. There are two ways to do this. One way is to use the context sensitive menu again, this time choosing **Export Selected Reads as FASTQ**.

Export Consensus with Gaps Export Consensus without Gaps
Align Selected Reads
Delete Selected Reads
Reset (un-align) Selected Reads
Export Selected Reads as FASTA
Export Selected Reads as FASTQ
Select Matching Pairs
Select Overlapping Reads Containing Selected Sequence
Extend Reference with Selected Read

Choose a suitable name for the fastq file. Here, I've created a new Ends folder and have called the file Left-1.fastq as this is the first set of junction reads from the left end of the transposon;

		Save As: Tags:	Left-1.fastq						
			Ends	0	•	Qs	earch		
Favorites Dropbox Google Drive MelissaMadsenNGS	Name			Date Mo	odified	~	Size	Kin	d
		Format:	FASTQ Multiple Sequence		0				
New Folder							С	ancel	Save

You can also save the reads using the main File | Export Selected Reads As... menu item and choosing the file format in the export dialog.

Finally, now that we have saved the reads, we are really not interested in those particular anymore. To avoid confusion when we start looking at the other junction reads, let's just delete them. Assuming you still have them selected, this can also be done with the context-sensitive menu;

Export Consensus with Gaps Export Consensus without Gaps
Align Selected Reads
Delete Selected Reads
Reset (un-align) Selected Reads Export Selected Reads as FASTA Export Selected Reads as FASTQ Select Matching Pairs Select Overlapping Reads Containing Selected Sequence Extend Reference with Selected Read

You can, of course, just press the <delete> key to accomplish the same thing.

Now we are ready to repeat this with the other sets of reads. If you look carefully at the greyedout residues, you will see certain patterns repeated. For example, the following reads clearly fall into two groups

```
3ΑΤΤΤΑGCATCAGAAAAAATATA-A(

ΓΤΑCTTTTCATAATATACCTATA-A(

ΓΤΑCTTTTCATAATATACCTATA-A(

3ATTTAGCATCAGAAAAAATATA-A(

3ATTTAGCATCAGAAAAAATATA-A(

ΓΤΑCTTTTCATAATATACCTATA-A(

ΓΤΑCTTTTCATAATATACCTATA-A(

3ATTTAGCATCAGAAAAAATATA-A(

3ATTTAGCATCAGAAAAAATATA-A(
```

One set has the sequence ATTTAGCATCAGAAAAAAT before the **ATA-A** bold residues and the other has TACTTTTCATAATATACC. We can use the same principle as before to select the two distinct sets of reads. Note that MacVector doesn't care about the fact that some of the residues are greyed out, it will treat them just the same as the bold residues. So, lets select a few residues across the junction of the first of the sequence motifs. The exact length of selection is up to you. You want to select enough residues to ensure that only those reads that have a unique sequence are selected, but not so many that you miss some of the reads that just have short extensions across the junction. In this case I selected 8 residues, including the final **AT**.

Again, we can use the context-sensitive menu items Select Overlapping Reads Containing Selected Sequence, Select Matching Pairs, Export Selected Reads as FASTQ and Delete Selected Reads to save the pairs of reads containing the selection and to remove them from the alignment. In this case, I called the resulting file Left-2.fastq.

Next we repeat the procedure with the TACTTTTCATAATATACC sequence, calling those hits Left-3.fastq.

That should, in theory, leave us with just two remaining sets of insertion reads. In fact, there do appear to be two distinct sets of reads remaining;

```
СААĞТĞТТТАСТТАСТТТТТТАТĞТАТА

СААĞТĞТТТАСТТАСТТТТТТАТĞТАТА

СААĞТĞТТТАСТТАСТТТТТТАТĞТАТА

АТА

АААĞАТА

АААĞАТА

ССААĞСАТАААССТТĞАСААААААСАĞТА

СААĞCATAAACCTTĞACAAAAAACAĞTA

СААĞCATAAACCTTĞACAAAAAACAĞTA
```

Again, we can repeat the same procedure as above, saving the selected reads as Left-4.fastq and Left-5.fastq. When those reads have been removed from the alignment, our junction looks like this;

```
GAGTCGTCGGCACGTCAGA-T-GTGTA-TAAGAGACAG---GGCTG-AT.
CGGCAGCGTCAGTGTGTATAAGAGACAGTAATA-ACAGGTTGGCTG-AT.
                            TAATA-ACAGGTTGGCTG-AT
                          ACCTATA-ACAGGTTGGCTG-AT.
                            GTATA-ACAGGTTGGCTG-AT.
CGGCAGCGTCAGTGTGTATAAGAGACAGGTATA-ACAGGTTGGCTG-AT.
                              ATA-ACAGGTTGGCTG-AT
GTCGGCAGCGTCGATGTGTATAAGAGACAGATA-ACAGGTTGGCTG-AT.
                          ACCTATA-ACAGGTTGGCTG-AT.
  AGCGTCAGATTGTATAAGAGACAGACCTATA-ACAGGTTGGCTG-AT.
                          ACCTATA-ACAGGTTGGCTG-AT
                              ATA-ACAGGTTGGCTG-AT
                          AAAGATA-ACAGGTTGGCTG-AT
                          ACCTATA-ACAGGTTGGCTG-AT.
                              GTA-ACAGGTTGGCTG-AT
                                          GGCTG-AT
```

Finally, we can see there is no significant commonality in the remaining reads. Some have short greyed out regions and thus were not selected by the **Select Overlapping Reads Containing Selected Sequence**, whereas the others are likely sequencing errors. At this stage we can be confident we have found the 5 sets of junction reads at this end.

Now we need to turn our attention to the right end. Again, after scrolling around, we can clearly see the greyed out reads at the junction. Note, that to help you view reads when scrolling around, if you click on a residue in the reference or consensus sequences, the display will scroll vertically to bring the nearest overlapping read into view.

			Consensus Alig	nment.axml –	- Edito	r			
0 5 6	a 🚽 -	A. 3	CI 8	» 0 ·	AGCT	ACGTAC	ACETAC		
	refs Replica	Tonology Add Se	gs Remove Segs Ali	n Translations		~ ~	Next Mismatch		
Editor	Мар	Features	Annotations	Text	Dots	SNPs	Problems		
Sort	мар	reatures	Annotations	lext		SNPS	Problems		
	229		2310 2320 AT-CAGCCAACCTGTTAT	2330	2346			2370 231	
00000			AT-CAGCCAACCTGTTA						
MUU262:88:00+G IAAGG			AT-CAGCCAACCIGITA		GALCO	GTATAGGE	CACGAATACTAG	MACGAMATIA	CUTATACCATI
M00262:88:00+GTAAGG	AG > CGAAGTT	ATCAGACCGGGGGACTT	AT-CAGCCAACCTGTTAT	-ACCGTATTCTCA	TGACCG	GTATAGGCATT			
M00262:88:00+GTAAGG	AG 4 CGAAGTT	ATCAGACCGGGGGACTT	AT-CAGCCAACCTGTTAT	GATAAGCGTAATC	GTTTAA	CGGAACAAGTAG			
M00262:88:00+GTAAGG	AG 4 CGAAGTT	ATCAGACCGGGGA							
M00262:88:00+GTAAGG	AG > CGAAGTT	ATCAGACCGGGGGACTT	AT-CAGCCAACCTGTTAT	-ACCGTATTCTCA	TGACCG	GTATAGGCATTAA	ATGTC		
M00262:88:00+GTAAGG	GAG > CGAAGTT	ATCAGACCGGGGGACTG	TETETTATACACATETG	CGCTGCCGACGAC	TCCTTA	CGTGGAGATCTCG	TGGTCGCCCT		
M00262:88:00+GTAAGG	GAG > CGAAGTT	ATCAGACCGGGGGACTT	AT-CAGCCAACCTGTTAT	-ACCETATTCTCA	TGACCG	GTATAGGCATTAA	ATGTCCGCCTAT		
M00262:88:00+GTAAGG	GAG > CGAAGTT	ATCAGACCGGGGGACTT	AT-CAGCCAACCTGTTA	TCATAGAACCGAA	ATGAAA	AGCGTAGTCCCAAJ	TATAACGACTGT		
M00262:88:00+GTAAGG	GAG 4 CGAAGTT	ATCAGACCGGGGGACTT	AT-CAGCCAACCTGTTA	TCATAGAACCGAA	ATGAAA	AGCGTAGTCCCAA	TATAACGA		
M00262:88:00+GTAAGG	GAG 4 CGAAGTT	ATCAGACCGGGGGACTT	AT-CAGCCAACCTGTTAT	-ACCGTATTCTCA	TGACCG	GTATAGGCA			
M00262:88:00+GTAAGG	GAG > CGAAGTT	ATCAGACCGGGGGACTT	AT-CAGCCAACCTGTTAT	-ACCETATTCTCA	TGACEG	GTATAGGCACTGTO	TETTATACACATO	TGAEGET	
M00262:88:00+GTAAGG	GAG 📲 CGAAGTT	ATCAGACCGGGGGACTT	AT-CAGCCAACCTGTTAT	TATTTATAATTAT	ACTAAC	GATGTTATCTGCAG	ATCGTCTTTCAAT	GAGATATAATAT	ATTCATAT
M00262:88:00+GTAAGG	GAG > CGAAGTT	ATCAGACCGGGGGACTT	AT-CAGCCAACCTGTTAT	-AAGACAGAGCAA	CCTAAG	CCATCTATAAGTTA	CACGAATACTAG	TAAACGAAAATTA	CCTATACCATT
M00262:88:00+GTAAGG	GAG 4 CGAAGTT	ATCAGACCGGGGGACTT	AT-CAGCCAACCTGTTAT	TATTTATAATTAT	ACTAAC	GATGTTATCTGCA	ATCGTCTTTCAAT	GAGATATAATAT	ATTCATATTT/
M00262:88:00+GTAAGG	GAG > CGAAGTT	ATCAGACCGGGGGACTT	AT-CAGCCAACCTGTTA	TCATAGAACCGAA	ATGAAA	AGCGTAGTCCCAA	TATAACGATGACO	TATATGAGCCTT	AAAACGGAATT
M00262:88:00 + GTAAGG	AG 4 CGAAGTT	ATCAGACCGGGGGACTT	AT-CAGCCAACCTGTTAT	-AAGACAGAGCAA	CCTAAG	CCATCTATAAGTTA	CACGAATACTAG	AAACGAAAATTA	CCTATACCATI
M00262:88:00+GTAAGG	GAG > CGAAGTT	ATCAGACCGGGGGACTT	AT-CAGCCAACCTGTTAT	-AAGACAGAGCAA	CCTAAG	CCATCTATAAGTTA	CACGAATACTAGT	AAACGAAAATTA	CCTATACCATI
M00262:88:00+GTAAGG	GAG 4 CGAAGTT	ATCAGACCGGGGGACTT	AT-CAGCCAACCTGTTAT	-ACCGTATTCTCA	TGACCO	GTATAGGCATTAA	ATGTCCGCCTATA	GAGTCAAACGCT	TECOLTATET
M00262:88:00+GTAAGG	GAG 4 CGAAGTT	ATCAGACCGGGGGACTT	AT-CAGCCAACCTGTTA	TCATAGAACCGAA	ATGAAA	AGCGTAGTCCCAAJ	TATAACGATGACO	TATATGAGCCTT	AAAACGGAATT
M00262:88:00+GTAAGG	AG 4 CGAAGTT	ATCAGACCGGGGGACTT	AT-CAGCCAACCTGTTAT	-ACCGTATTCTCA	TGACCG	GTATAGGCATTAA	ATGTCCGCCTAT	GAGTCAAACGCT	TECGETATETE
M00262:88:00+GTAAGG	AG CGAAGTT	ATCAGACCGGGGGACTT	AT-CAGCCAACCTGTTA	TCATAGAACCGAA	ATGAAA	AGCGTAGTCCCAA	TATAACGATGACO	TATATGAGCCTT	AAAACGGAATT
M00262:88:00+GTAAGG	AG > CGAAGTT	ATCAGACCGGGGGACTT	AT-CAGCCAACCTGTTA	TCATAGAACCGAA	ATGAAA	AGCGTAGTCCCAA	TATAACGATGAC	TATATGAGCCTI	AAAACGGAATT
M00262:88:00+GTAAGG	AG > CGAAGTT	ATCAGACCGGGGGACTT	AT-CAGCCAACCTGTTA	TCATAGAACCGAA	ATGAAA	AGCGTAGTCCCAA	TATAACGATGACO	TATATGAGCCTT	AAAACGGAATT
M00262:88:00+GTAAGG	AG 4 CGAAGTT	ATCAGACCGGGGGCCTT	AT-CAGCCAACCTGTTA	TCATAGAACCGAA	ATGAAA	AGCGTAGTCACAAA	TATAACGATGAAA	TATATGAGCCTT	AAAACGGAATT
M00262:88:00+GTAAGG	AG > CGAAGTT	ATCAGACCGGGGGACTT	AT-CAGCCAACCTGTTA	TCATAGAACCGAA	ATGAAA	AGCGTAGTCCCAAA	TATAACGATGAC	TATATGAGCCTI	AAAACGGAATI
M00262:88:00+GTAAGG	AG & CGAAGTT	ATCAGACCGGGGGACTT	AT-CAGC						
M00262:88:00+GTAAGG	AG > CGAAGTT	ATCAGACCGGGGGACTT	AT-CAGCCAACCTGTTAT	-ACCGTATTCTCA	TGACCG	GTATAGGCATTAA	ATGTCCGCCTATA	GAGTCAAACGCT	TECECTATETE
M00262:88:00+GTAAGG	AG 4 CGAAGTT	ATCAGACCGGGGGACTT	AT-CAGCCAACCTGTTA	-ACCGTATTCTCA	TGACCG	GTATAGGCATTAA	ATGTCCGCCTAT	GAGTCAAACGCT	TECECTATETE
M00262:88:00+GTAAGG	AG > CGAAGTT	ATCAGACCGGGGGACTT	AT-CAGCCTGTCTCTTA	-ACACATETEEGA	GCCCAC	GAGACGGACTECT	TETEGTATGEEG	CTTCTGCTTGA	
M00262:88:00+GTAAGG	A CONTRACTOR OF THE OWNER OWNER OF THE OWNER OWNE								
	201010111								

Now we just have to repeat the same analysis as we did for the left end, but calling our read collections Right-1.fastq through Right-5.fastq. When finished, close the alignment window, but don't save – we don't want to replace the original that had all the reads aligned before we started removing reads we had already saved to disk.

Assemble Filtered Sets of Reads

The most comprehensive way to proceed now is to separately assemble each of the 10 sets of reads we have created and then save the consensus from each of those to be aligned to the reference genome to reveal the exact locations of the transposon insertions.

Create a new File | New | Assembly Project, click on the Add Reads toolbar button and select the 10 sets of saved fastq files.

•	•						Untitl	ed 5	— P	roject					
4		P	Ŵ	Œ	° 4	0	- -			्म	`	2	۹	25	Q~Name
Add	Reads Add Seqs	Add Ref	Add Contig	Remove	Reset	Prefs	Replica	Phre	ed Cr	ossMatch	Phrap	Bowtie	SPAdes	Velvet	Filter
	Project	Proper	ties	Coverag	le										
	Name		^ Status		Ler	ngth	#		ClipL	ClipR	Start	Stop	Definit	tion	
	Left-1.fastq		llumina	Unpaired	d	251		48					/Users	/kendall,	/Documents/FastQTests/Insert
	Left-2.fastq		llumina	Unpaired	b	250		18					/Users	/kendall,	/Documents/FastQTests/Insert
	Left-3.fastq		llumina	Unpaired	b	250		12					/Users	/kendall,	/Documents/FastQTests/Insert
	Left-4.fastq		llumina	Unpaired	b	250		8					/Users	/kendall,	/Documents/FastQTests/Insert
	Left-5.fastq		llumina	Unpaired	ł	250		32					/Users	/kendall,	/Documents/FastQTests/Insert
	Right-1.fastq		llumina	Unpaired	b	251		18					/Users	/kendall,	/Documents/FastQTests/Insert
	Right-2.fastq		llumina	Unpaired	b	250		16					/Users	/kendall,	/Documents/FastQTests/Insert
	Right-3.fastq		llumina	Unpaired	b	251		40					/Users	/kendall,	/Documents/FastQTests/Insert
	Right-4.fastq		llumina	Unpaired	b	250		16					/Users	/kendall,	/Documents/FastQTests/Insert
	Right-5.fastq		llumina	Unpaired	b	249		4					/Users	/kendall,	/Documents/FastQTests/Insert

Note that some of the fastq files have very few sequences in them - just 8 for Left-4.fastq and only 4 for Right-5.fastq. These are far too few for typical NGS short-read assemblers such as *SPAdes* and *Velvet* to assemble. However, *phrap* can easily assemble small numbers of reads as long as they share some overlap and, in addition, will take into account the quality scores in the fastq files to generate better alignments. We want to assemble each set of reads independently, so first select Left-1.fastq, then click on the **Phrap** toolbar button.

Basic	Advanced	Miscellaneous		
Pairwise Alignments		Filtering		
Mismatch penalty:	-2	Minimum alignme	ent score:	25
Gap initiation penalty:	-4	Potential vector b	ases:	0
Gap extension penalty:	-3	Assembly		
Banded search		Stringency:		5
Minimum match length:	14	Maximum gap:		5
Maximum match length:	30	Repeat stringenc	y:	0.95
Consensus				
Minimum segment size:	9	Node spacing:		8
Short Read Defaults De	faults		Cancel	ОК

Click on the **Short Read Defaults** button to set up optimal parameters for this alignment, then click **OK**. After a short delay (2-3 seconds) a new *Phrap* job object appears in the project window.

۲	• •					U	ntitle	ed 5 —	Proj	ect							
Add	Reads Add Seqs	Add Ref Add	Contig	Remove	Reset	Prefs	Repl	lica P	hred	CrossMat	ich Ph	rap I	Bowti	e SP/	Q~Nam	Filter	>>
	Project	Properties		Coverag	e												
	Name		^ Sta	itus		Length		#		ClipL	ClipR	St	art	Stop	Definition		
	Left-1.fastq		llun	nina Unpa	ired		251		48						/Users/kenda	II/Documents,	/FastQ
	Left-2.fastq		llun	nina Unpa	ired		250		18						/Users/kenda	II/Documents,	/FastQ
	Left-3.fastq		llun	nina Unpa	ired		250		12						/Users/kenda	II/Documents,	/FastQ
	Left-4.fastq		llun	nina Unpa	ired		250		8						/Users/kenda	II/Documents,	/FastQ
	Left-5.fastq		llun	nina Unpa	ired		250		32						/Users/kenda	II/Documents,	/FastQ
	Right-1.fastq		llun	nina Unpa	ired		251		18						/Users/kenda	II/Documents,	/FastQ
	Right-2.fastq		llun	nina Unpa	ired		250		16						/Users/kenda	II/Documents,	/FastQ
	Right-3.fastq		llun	nina Unpa	ired		251		40						/Users/kenda	II/Documents,	/FastQ
	Right-4.fastq		llun	nina Unpa	ired		250		16						/Users/kenda	II/Documents,	/FastQ
	Right-5.fastq		llun	nina Unpa	ired		249		4						/Users/kenda	II/Documents,	FastQ
-982	V Phrap Left-1 -	13:54 - Aug															
ee.	Contig 2						251		1	1	25	1	1	251			
-986	Contig 3					1	109		42	1	110	9	1	1109			

You can see that two contigs were created, but one only contains a single read (the # column) and the other is much larger and contains most of the input reads. Double-click on Contig 3 to open up the *Contig Editor*.

•••							Contig	3 — Edit						
	6	P.	~ 9	0 ~	\mathcal{D}	-		AGC	F 😩 .					
Locked Text View	Prefs	Replica	Trans	lations	Width	Baseca	alls Qua	lities Do	s Create					
Editor	1	Мар		Features	[Annotat	ons	Sumn	nary					
			380	390		00	410	420	430	440	450	460	470	480
						GCCGCCCG	ACTGGGC	TGGCGATG	AGCGAAATGT	AGTGCTTACO	STTGTCCCGCA	TTTGGTACAGC	SCAGTAACCG	SCAAAATCGC
M00262:88:00+GTA	AGGAG	AAT	GAAACCT	TAACGCTAT	GAA									
M00262:88:00+GTA	AGGAG	AAT	GACACCT	TAACGCTAT	GAACCC	GCCGCCCG	ACTGGGC	TGGCGATG						
M00262:88:00+GTA	AGGAG	► AAT	GAAACCT	TAACGCTAT	GAACTC	GCCGCCCG	ACTGGGC	TGGCGATG	AGCGAAATGT	AGTGCTTACO	STTGTCCCGC#	TTTGGTACAGC	GCAGTAACCG	GCAAAATCGC
M00262:88:00+GTA	AGGAG	► AAT	GAAACCT	TAACGCTAT	GAACTC	GCCGCCCG	ACTGGGC	TGGCGATG	GCGAAATGT	AGTGCTTACO	STTTTCCCGC/	TTTGGTACAGC	SCCGTAACCG	GCAAAATCCO
M00262:88:00+GTA	AGGAG	► AAT	GAAACCT	TAACGCTAT	GAACTC	GCCGGCCG	ACTGGGC	TGGCGATG	GCGAAAAGT	AGTGCTTACO	TTGTCCCGCA	TTTTGTACAGC	SCATGAAACG	STCAAATCGC
M00262:88:00+GTA	AGGAG	► AAT	GAAACCT	TAACGCTAT	GAACTC	GCCGCCCG	ACTGGGC	TGGCGATG	GCGAAATGT	AGTGCTTACO	TTGTCCCGCA	TTTGGTACAGC	GCAGTAACCG	GCAAAATCGC
M00262:88:00+GTA	AGGAG	► AAT	GAAACCT	TAACGCTAT	GAACTC	GCCGCCCG	ACTGGGC	TGGCGATG	GCGAAATGT	AGTGCTTACO	TTGTCCCGCA	TTTGGTACAGC	GCAGTAACCG	GCAAAATCGC
M00262:88:00+GTA	AGGAG								GCGAAATGT	AGTGCTTACO	TTGTCCCGCA	TTTGGTACAGC	GCAGTAACCG	GCAAAATCGC
M00262:88:00+GTA	AGGAG								ATGT	AGGGCTTACO	TTGTCCCGCA	TTTGGGACAGC	SCAGTAACCG	GCAAAATCGC
M00262:88:00+GTA	AGGAG											GC	GCAGTAACCG	SCAAAATCGC

As you scroll through you can see that the reads have assembled very nicely. You can save the consensus by selecting the File | Export Consensus As... menu item or using the right-click context menu.

Export Consensus with Gaps
Export Consensus without Gaps
Export Selected Reads as FASTA Export Selected Reads as FASTQ Select Matching Pairs Select Overlapping Reads Containing Selected Sequence
Cannot Circularize Consensus

I saved the consensus as Left-1.nucl.

Now close the *Contig Editor* window and repeat for each of the .fastq input files. You may find it easier to run all the *phrap* assemblies first and to then go through each *phrap* job in turn, exporting the consensus.

Not all of the phrap assemblies generate clear "major" contigs. Phrap Left-4 generates two contigs of similar size, one with 3 reads and one with 5 reads. We'll save both of these contigs, with the names Left-4a.nucl and Left4b.nucl

•	• •			Unti	tled 5 –	Project						
		P 4	-		6	-		୍ଷିୟ	Q.	Name		>>
Ad		Add Ref Add Con	tig Remove	Reset	Prefs	Replica	Phred	CrossMa	tch	Fi	lter	
	Project	Properties	Coverag	ge								
	Name	~ 5	Status		Length	#		ClipL	ClipR	Start	Stop	Definition
C.	▼ Phrap Left-1 -	13:54 - Aug										
₩.	Contig 2				:	251	1	1	251	1	251	
₩	Contig 3				11	09	42	1	1109	1	1109	
ŒE	▼ Phrap Left-2 -	14:11 - Aug 2										
₩.	Contig 5				3	327	2	1	327	1	327	
WE .	Contig 6				3	324	2	1	324	1	324	
¶E.	Contig 7				7	'30	12	1	730	1	730	
¶E.	▼ Phrap Left-3 -	14:11 - Aug 2										
₩.	Contig 9				8	39	12	1	839	1	839	
-91E	▼ Phrap Left-4 -	14:11 - Aug 2										
₫ E.	Contig 11					571	3	1	571	1	571	
¶E.	Contig 12				!	512	5	1	512	1	512	
€E.	▼ Phrap Left-5 -	14:11 - Aug 2										
¶E.	Contig 14					311	2	1	311	1	311	
₩.	Contig 15				5	46	28	1	546	1	546	
₩	▼ Phrap Right-1	- 14:12 - Aug										
em	Contia 17				6	10	17	1	610	1	610	

Note that the procedure above represents the slow and methodical way of generating the consensus sequence for each filtered set of reads. In reality, you can save a lot of time by selecting all of the phrap result objects...

• •			U	ntitle	d 5 — F	Pro	ject						
	P :	ĥ.	ØŽ	0	0	F	7 ~		Q~1	lame			>>
Add Reads Add Seqs	Add Ref Add	Contig R	emove	Reset	t Prefs	R	eplica	Phre	d	Filter			
Project	Properties	С	overag	e									
Name	^	Status		L	ength		#		ClipL	ClipR	Start	Stop	Defin
Right-4.fastq		Ilumina L	Inpaire	d	25	50		16					/User
Right-5.fastq		Ilumina L	Inpaire	d	24	19		4					/User
🔹 🔻 Phrap Left-1 - 13	3:54 - Aug												
Contig 2					2	51		1	1	251	1	251	
Contig 3					11(9		42	1	1109	1	1109	
🔹 🔻 Phrap Left-2 - 14	4:11 - Aug 2												
Contig 5					3:	27		2	1	327	1	327	
Contig 6					32	24		2	1	324	1	324	
Contig 7					73	30		12	1	730	1	730	
🔹 🔻 Phrap Left-3 - 14	4:11 - Aug 2												
Contig 9					83	39		12	1	839	1	839	
🔹 🔻 Phrap Left-4 - 14	4:11 - Aug 2												
Contig 11					5	71		3	1	571	1	571	
Contig 12					5	12		5	1	l 512	1	512	
🔹 🔻 Phrap Left-5 - 14	4:11 - Aug 2												
Contig 14					3	11		2	1	I 311	1	311	
Contig 15					54	16		28	1	546	1	546	
▼ Phrap Right-1 - 1	14:12 - Aug												
Contig 17					64	8		17	1	648	1	648	
▼ Phrap Right-2 - 1	14:12 - Aug												
Contig 19					25	50		1	1	250	1	250	
Contig 20					25	50		1	1	250	1	250	
Contig 21					6	12		13	1	l 612	1	612	
🔹 🔻 Phrap Right-3 - 1	14:12 - Aug												
Contig 23					84	17		35	1	847	1	847	
🔹 🔻 Phrap Right-4 - 1	14:12 - Aug												
Contig 25					88	32		16	1	882	1	882	
🔹 🔻 Phrap Right-5 - 1	14:12 - Aug												
Contig 27					3	72		4	1	372	1	372	

... and then choosing File | Export Selected Contigs To... which will prompt you for a folder to save the contigs into (I created a new empty folder), and will then save the consensus from each contig into that folder using the name of the contig e.g. Contig 2.nucl, Contig 3.nucl etc. You will see in the Shortcuts section, where we align the consensus sequences to the genome to identify the transposition locations, that either way will generate the same results.

Identifying Transposition Sites on the Genome

Now we can use the *Align to Reference* function to align the left and right consensus sequences to the target genome. In theory, only the section of each consensus representing the genome sequence should align, with any transposon sequence remaining unaligned and greyed out. In addition, if all has worked correctly, each alignment site should have a pair of consensus sequences aligning, one for each side of the transposition location.

First, we should open Genome.nucl, then choose Analyze | Align to Reference. Make sure you have the Editor tab selected, click on the Add Seqs toolbar button and select all the consensus sequences you created in the previous section.

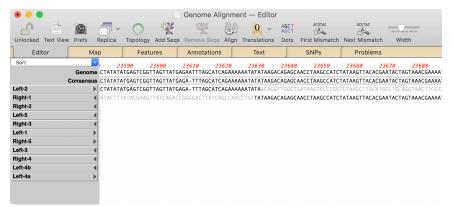
• • •						Genome Ali	gnme	ent — Editor					
-0	10-1- 10000	6	-	0		CH	BA	0 ~	AGCT	ACGTAC	ACGTAC		
Unlocked	Text View	Prefs	Replica	Topology	Add Seqs	(Align	Translations	Dots	First Mismatch	Next Mismatch	Width	
Edi	itor 🥤	N	ар	Featu	ures	Annotations	. [Text	5	SNPs	Problems		
Sort			2	10	20	30	40	50	60	70	80	90	100
				TAGGAATGT	CCAGAGCT	FAAAAGCAATGCT	TTAGTI	TTAATTAATCGG	GAGTAA	CTACATATCCAAC	CAAACTAACAAGAG	CCTACCTGA	ACTCGTAG
	c	Consensu											
Left-1			AAGCCA	aCAGTGATA	TTGATTTG	CTGGTTACGgTGA	CCGTA	AGGCTTGATGAA	ACAACG	CGGCGAGCTTTGA	TCaACGACCTTTTG	GAAACTTCG	GCtTCCCC
Left-2			ATTTaA	CCAAATGGG	GGAAATAG	TCATTCTATGCCG	acTGC	TAATATGGATAG	CTTACC	AAAAAATTTAGCT	gTAACCCCACCTAT	AAATGTAGG	GTCACTATA
Left-3			aGATTCO	ACAGCTCTT	AAAGCATA	TAATATCTGTTCT	CTTGT	TGCTTGAATAGA	AACATC	ATCTGAAGCCATA	CCGTAAGCTTTTTC	AAGCTCTTG	АТСТТТАА
Left-4a			GATATTO	ATTTGCTGG	TTACGGTG	ACCGTAAGGCTTG	ATGAA	ACAACGCGGCGa	GCTTTG	ATCAACGACCTTT	TGGAAACTTCGGCT	тсссстбба	GaGAGCGA
Left-4b			gTTATA	ACGGTATCA	AATGGGGG	тттаадстстсдт	ATTTT	CCATTAAGCCG	AGCATG	ATTTTACCATCGC	CGCTTAGTTTAGAA	GACTGGCAA	GAGGATTG
Left-5			-								TAAAAAGTGCAACG		
Right-1			GATTCA	CTTGGCAAT	AGCTTGCT	TTTCCTTAACTAT	TCTAG	TTCCTCTACTT	CTAATA	GTTTATCTTCTAA	АТССАЛАТАСТСТА	ATTTCTTGT	CAAGTTCT
Right-2											GGTATGCTAGATTA		
Right-3											AACTCATGACCATI		
Right-4											TCCCTTATGTTACI		
Right-5											ΤϹϾΑΤΑΤΑΑΤΑΑΘΟ		
night-5			GLLGTA	AATAACAAT	TCTAAGGT	MAAGCAAAAAA 1 1 1	AILLI	GAAGATGATGA	TAACGG	LAGIAAIGLIIGI	I C GA I A I AA I AAGU	CIATICITA	GCTAAAAG

Click on the Align toolbar button – the default parameters are fine for this, but if you find the alignment taking some time, you can try increasing the Hash Value to 12.

When the alignment completes, you should see that all of the sequences aligned except for Left-4a. Chances are this was created from internal transposon sequence that is not present in the genome.

• • •)					Genome Ali	gnme	ent — Editor					
n		6	-	0	1	QZ	3	0 ~	AGCT AGCT	ACGTAC	ACGTAC		-
Unlocked	Text View	Prefs	Replica	Topology		Remove Seqs		Translations	Dots	First Mismatch	Next Mismatch	Width	
Edi	tor	N	/ap	Featu	ires	Annotations		Text		SNPs	Problems		
Sort			~	10	20	30	40	50	60	70	80	90	100
		Genor	ne GCTATAA	TAGGAATGT	CCAGAGCTT	AAAAGCAATGCT	TTAGT	TTAATTAATCGG	GAGTAA	CTACATATCCAAC	CAAACTAACAAGAG	CCTACCTGA	ACTCGTAG
	c	consens	us										
Left-2													
Right-1			•										
Right-2			•										
Left-5			•										
Right-3			•										
Left-1			•										
Right-5			•										
Left-3			•										
Right-4			4										
Left-4b			•										
Left-4a			▶ GATATTG	ATTTGCTGG	TTACGGTGA	ACCGTAAGGCTTG	ATGAA.	ACAACGCGGCGA	<i>GCTTTG</i> .	ATCAACGACCTTT	TGGAAACTTCGGCT	ТССССТБСА	GAGAGCGA

If you click on the Left-2 name button, the display will scroll to show where that sequence aligns. If you continue scrolling, you should eventually reach the location where Left-2 and Right-1 diverge.



You can see that the two sequences represent the two genomic flanking regions of this insertion site. Not they both share a **TA** overlap - a feature of the transposition characteristic of this transposon.

Select the shared **TA** residues in the Genome reference sequence. Let's create a feature to highlight this insertion site. By default, the *Align to Reference* Editor tab does not have a Create toolbar button, but we can add one easily enough. Right-click (or <ctrl>-click) on any grey space on the toolbar and a context sensitive menu will appear. Choose Customize Toolbar... to bring up a sheet with options;

ACGTAC A t Mismatch	Lock Text View	ACGTAC Next Mismatch	Prefs	Remove Seqs	Qualities Space
<u>0</u> ~	Ē			g	
	Text View		++ Flexible Space	Print	Space
anslations	Text View	Width	Flexible Space	Print	Space
o the toolbar.					
Replica Topo	ology Add Seqs R	emove Seqs Align	O ~ AGCT AGCT Translations Dots I	(A)	ACGTAC ACGTAC Mismatch Wid
	o the toolbar.	D · O 😤			

Drag the Create button (outlined above) onto the Editor tab toolbar and click Done.

Now, with the **TA** residues selected, we can click on the **Create** toolbar button to bring up the feature creation/editor dialog.

					Fe	ature	Symbol					
			5.			_	-					
			Fe	eature	кеуwo	ra: m	obile_elen	nent				
Location:												
Starts	5	Start Base		Stops		Stop B	lase	Operation		GenBar	nk	
At	\$ 2	23628		,	At ≎	23629		Continuo	us () 🗘	23628.	.23629	
+-					0.0				0.000	ration:	Join	
					Cor	nplemer	ntary Strar	nd	Ope	ation:	Join	×
Location in												
		ank Format:										
2362823		ank Format:										
2362823		ank Format:			Quali	fiers	Free-Forn	n				
	3629	ank Format:				fiers	Free-Forn	n				
Qualifi	3629 ier		С	Commer		fiers	Free-Forn	n				
Qualifi	3629 ier bile_ele	ank Format: ement_type	C	Commer	nts	fiers	Free-Forn	n				
Qualifi	3629 ier bile_ele		C		nts	fiers	Free-Forn	n				
Qualifi	3629 ier bile_ele		C		nts	fiers	Free-Forn	n				
Qualifi	3629 ier bile_ele		C		nts	fiers	Free-Forn	n				
Qualifi	3629 ier bile_ele		C		nts	fiers	Free-Forn	n				
Qualifi	3629 ier bile_ele		C		nts	fiers	Free-Forn	n				
Qualifi	3629 ier bile_ele		C		nts	fiers	Free-Forn	n				
Qualifi	3629 ier bile_ele		C		nts	fiers	Free-Forn	n				
Qualifi	3629 ier bile_ele		C		nts	fiers	Free-Forn	n		Can		OK

Normally, I would recommend that you use <code>misc_feature</code> for annotating something like this, but the <code>Genome.nucl</code> sequence already has quite a lot of <code>misc_feature</code> annotations, so, to avoid our new feature getting lost in the midst of so many annotations, let's use mobile element instead.

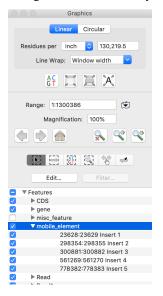
Also click on the "+" button (circled) to add a /note with a suitable name for our transposition location (e.g. "Insert 1").

Click on the Symbol tab and let's add some eye-catching graphics for this;

Genome Alignment
Feature Symbol
► mobile_element
V Fill: Color: Pattern: V
Pen: Color: Thickness: Hairline 🗘
Show label
<description></description>
Helvetica 9.0 pt Color:
Orientation: Outside Center ᅌ Concentric ᅌ
Visible
Show as a graphic
Level: Outside Sequence 📚
Defaults Revert Cancel OK

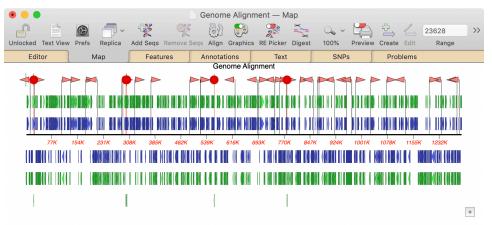
This will create a nice bright red "lollipop" at the insertion location.

Repeat with the other 4 insertion locations. As expected, at least for this data set, the 5 pairs nicely flank the 5 insertion sites. Note that to save some effort setting the symbols, you can create all mobile_element features without changing the default symbol, switch to the Map tab, then double-click on the mobile_element entry in the Features tree list so that you can change all of the feature symbols at one time.



What Genes Have the Transpositions Interrupted?

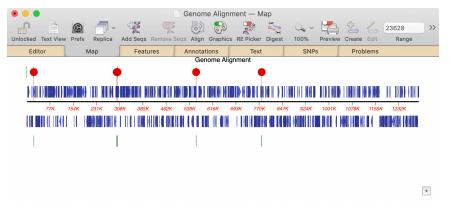
If we look at the **Map** tab, by default it's a bit of a mess with all of the different layers of features turned on;



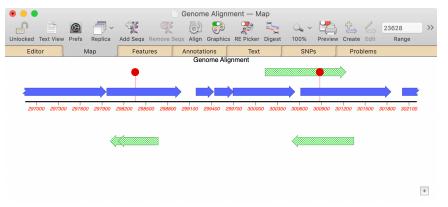
You can just about make out the red lollipops we added as flags for the insertion locations. This genome was downloaded directly from GenBank and has the standard default MacVector feature appearance – CDS features in blue, gene features in green and various RNA features as lines with arrowheads. In general, CDS features have the most extensive annotation when GenBank submissions are run through the Prokaryotic Genome Annotation Pipeline, so let's turn off the gene and RNA features, just leaving the CDS, mobile_element and Read features using the *Graphics Palette*;

Graphics
Linear Circular
Residues per inch ᅌ 130,219.5
Line Wrap: Window width
AF I I X
Range: 1:1300386
Magnification: 100%
Edit Filter
Features
∠ ► CDS
▶ gene
misc_feature
> mobile_element
✓ ► Read
Read*
► rRNA
► source
► Results
P Results

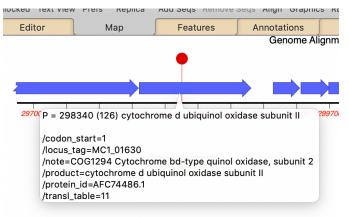
This cleans up the display tremendously and we can now clearly see our "lollipops" along with the green lines that are actually the locations of our aligned consensus sequences.



The second "lollipop" is actually two sites very close to each other – we can see that more clearly if we click to the left of the graphic, hold down the mouse and drag across a few pixels to "zoom" into the sequence. Note that at any time when zooming, you can double-click to get back to the "fit to window" scaling. You can also use the up and down arrow keys to zoom in/out 2x with each press and use the left/right arrow keys to nudge the zoomed region left or right.



You can clearly see that both of these transpositions have inserted into a CDS feature. If you mouse over the blue arrow representing the feature, a tooltip appears with additional information about the feature.



In addition, if you click on the blue graphic to select it, then switch to the **Features** tab, the feature is selected and scrolls into view;

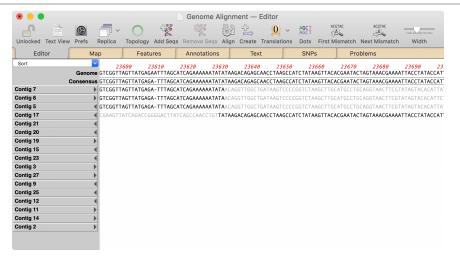
• • •				Genome Alignm	nent — Features		
n l		-	- 👰 🖞	, 💪 💁 🗉	€⇒		Q~ Description
Unlocked Tex	t View Prefs	Replica	Remove Seqs Cre	ate Edit Delete J	loin		Filter
Editor		Map	Features	Annotations	Text	SNPs	Problems
Type CDS	Start 296587	^ Stop 2979	C Descr	•	ubiquinol oxidase su	bunit I	
			/trans	in_id=AFC74485.1 _table=11 ation=MEFDQGLLS	SRIQFAFTISFHIIFPA	TIGLASFLAVIEGLV	VLKTKKPVYQEIYKFWVKIFAVTFG
gene	296587	2979	957 /locus	_tag=MC1_01625			
CDS	297965	2985	/locus /note: 984 /produ /prote /trans	ict=cytochrome d u in_id=AFC74486.1 _table=11	ome bd-type quinol ubiquinol oxidase su LPLVWGGLIATAICLY	bunit II	PFAPTDDCRHKMINSIAPFWDGN
gene	297965	2989	,	_tag=MC1_01630			
Read	298011	2986		Right-2			
Read	298122	2986		Left-5			
mobile_elem	298354	2983		e_element_type= Insert 2			

Shortcuts!

The strategy outlined in the previous sections is very thorough and walks you carefully through each step. It should give you a good idea of the different functions in MacVector that can help you analyze this sort of data and is particularly applicable for small data sets where there may be very few reads that actually cross the transposon/genome boundaries. But, once you understand the concepts, you can often take shortcuts and still end up with the exact same results for less effort. Here are a couple of examples.

Export All Contigs at Once

As described above, once you have used *phrap* to assemble contigs for each of the filtered read sets, rather than export each consensus separately, you can just save them all into a folder. Then, you can run an *Align to Reference* of all the contigs against the genome in the same way as we did with the individually named contigs;



In this case, there are more contigs that do not align to the genome (likely pure transposon sequences) and some locations, such as above, where several contigs align either side of an insertion site. But the end result is the same – greyed out transposon sequences either side of an **AT** insertion point.

Directly Align Raw Reads

An even simpler approach is to completely bypass the read selection and contig assembly steps and just align all of the transposon containing reads from the *Align to Folder* search results directly against the genome. In some cases, this can work well – however it is reliant on having reads that cross the transposon junctions with enough adjacent target sequence present to align to the reference genome. Accordingly, you may want to adjust the alignment parameters a little. Here I've added the *Align to Folder* hits from the initial transposon search to a Genome.nucl *Align to Reference* and clicked on the **Align** toolbar button.

Alignment Type: Sec	uence Conf	irmation ᅌ	
Residue Scoring		Alignment Parameters	
Match:	2	Hash Value:	12 文
Mismatch:	-3	Sensitivity:	6 ᅌ
Ambiguous Match:	0	Score Threshold:	35
Gap Penalty:	4	X Dropoff:	25
	Defaults	Cancel	ОК

The main change from the defaults is to lower the **Score Threshold** to 35. As a **Match** is worth a score of 2, this means you need 18 perfect matches, for a total score of 36, in order for the read to align. You can try dropping this down to an even lower value to handle even shorter overlaps, but anything below 20 (i.e. 10 perfect matches) is likely to accept spurious matches. Note that I also increased the hash value to 12 for speed, though unless you have many thousands of reads to process, this is likely unnecessary.

• • •				Genom	e Align	ment -	Editor						
	- -	0	ž	Ø	(2)	÷	0	~ 1	AGCT	ACGTA		ACGTAC	
Unlocked Text View Prefs	Replica	Topology Ad		Remove Ser	ڪري. Alian د	Create	Translat			~ ~	natch Next	~ ~	Width
Editor	lap	Features		Annotatio		Tex			SNPs		Problem	s	
Sort		1)										
		23580 235		23600 GTTAGTTATG	23610 GAATTT	2362		23630 ATAT-A		640 AGCAACC	23650 TAAGCCATC	23660 TATAAGTT	23670 23 ACACGAATACTAGTAA
													ACACGAATACTAGTAA
M00262:88:00+GTAAGGAG	TAATTTA	TCTACTATATAT	GAGTCG	G									
M00262:88:00+GTAAGGAG	▶ TAATTTA	тстастататат	GAGTCG	GTTAGTTATG/	GAATTT	AGCATCAG	лалалат	ATAACA	GGTTGG	СТБАТАА	GTCCCCGGT	TAAGCTT	GCATGCCTGCAGGTA
M00262:88:00+GTAAGGAG		тстастататат	GAGTCG	GTTAGTTATG/	GAATTT	AGCATCAG	АААААТ	ATAAC/	GGTTGG	СТБАТАА	GTCCCCGGT	TAAGCTT	GCATGCCTGCAGGTAA
M00262:88:00+GTAAGGAG	◀ TAATTTA	тстастататат	GAGTCG	GTTAGTTATG/	GAATTT	AGCATCAG	АААААТ	ATAACA	GGTTGG	СТБАТАА	GTCCCCGGT	TAAGCTT	GCATGCCTGCAGGTAA
M00262:88:00+GTAAGGAG	▶ TAATTTA	тстастататат	GAGTCG	GTTAGTTATG/	GAATTT	AGCATCAG	АААААТ	ATAACA	GGTTGG	СТБАТАА	GTCCCCGGT	TAAGCTT	GCATGCCTGCAGGTAA
M00262:88:00+GTAAGGAG	TAATTTA	тстастататат	GAGTCG	GTTAGTTATG/	GAATTT	AGCATCAG	АААААТ	ATAACA	GGTTGG	СТБАТАА	GTCCCCGGT	TAAGCTT	GCATGCCTGCAGGTAA
M00262:88:00+GTAAGGAG	▶ TAATTTA	тстастататат	GAGTCG	GTTAGTTATG	GAATTT	AGCATCAG	АААААТ	ATAACA	GGTTGG	СТБАТАА	GTCCCCGGT	TAAGCTT	GCATGCCTGCAGGTAA
M00262:88:00+GTAAGGAG	TAATTTA	ТСТАСТАТАТАТ	GAGTCG	GTTAGTTATG	GAATTT	AGCATCAG	AAAAAT	ATAAC/	GGTTGG	СТБАТАА	GTCCCCGGT	TAAGCTT	GCATGCCTGCAGGTAA
M00262:88:00+GTAAGGAG	▶ TAATTTA	ТСТАСТАТАТАТ	GAGTCG	GTTAGTTATG	GAATTT	AGCATCAG	AAAAAT	ATAAC/	GGTTGG	СТБАТАА	GTCCCCGGT	TAAGCTT	GCATGCCTGCAGGTAA
M00262:88:00+GTAAGGAG	•	CTATATAT	GAGTCG	GTTAGTTATG	GAATTT	AGCATCAG	АААААТ	ATAAC/	GGTTGG	СТБАТАЛ	GTCCCCGGT	TAAGCTT	GCATGCCTGCAGGTAA
M00262:88:00+GTAAGGAG		CTATATAT	GAGTCG	GTTAGTTATG	GAATTT	AGCATCAG	АААААТ	ATAAC/	GGTTGG	СТБАТАЛ	GTCCCCGGT	TAAGCTT	GCATGCCTGCAGGTAA
M00262:88:00+GTAAGGAG	•	ΑΤΑΤΑΤ	GAGTCG	GTTAGTTATG/	GAATTT	AGCATCAG	АААААТ	ATAAC/	GGTTGG	СТБАТАЛ	GTCCCCGGT	TAAGCTT	GCATGCCTGCAGGTAA
M00262:88:00+GTAAGGAG	◀ GTATAAG	AGACAGATATAT	GAGTCG	GTTAGTTATG/	GAATTT	AGCATCAG	АААААТ	ATAAC/	GGTTGG	СТБАТАА	GTCCCCGGT	TAAGCTT	GCATGCCTGCAGGTAA
M00262:88:00+GTAAGGAG	•	ATATA	GAGTCG	GTTAGTTATG/	GAATTT	AGCATCAG	АААААТ	ATAAC/	GGTTGG	СТБАТАА	GTCCCCGGT	TAAGCTT	GCATGCCTGCAGGTAA
M00262:88:00+GTAAGGAG	•					P	AACCG AT	A-ATA/	AGA-AG	AGAAACO	TAAACCAAC	ГААААСАТ	AAGCGAAAACTAGTAA
M00262:88:00+GTAAGGAG	4						CCTG	TAT-A	AGACAG	AGCAACO	TAAGCCATC	TATAAGTT	ACACGAATACTAGTAA
M00262:88:00+GTAAGGAG		CTTTA	TACGAA	GTTATCAGACO	GGGGAC	TTATCAGO	CAACCTG	TAT-	AGACAG	AGCAACO	TAAGCCATC	TATAAGTT	ACACGAATACTAGTAA
M00262:88:00+GTAAGGAG	CTTCGTA	TAGGATACTTTA	TACGAA	GTTATCAGACO	GGGGAC	TTATCAGO	CAACCTG	TAT-A	AGACAG	AGCAACO	TAAGCCATC	TATAAGTT	ACACGAATACTAGTAA
M00262:88:00+GTAAGGAG		ACTTTA	TACGAA	GTTATCAGACO	GGGGAC	TTATCAGO	CAACCTG	TAT-A	AGACAG	AGCAACO	TAAGCCATC	TATAAGTT	ACACGAATACTAGTAA
M00262:88:00+GTAAGGAG	▶ CTTCGTA	TAGGATACTTTA	TACGAA	GTTATCAGACO	GGGGAC	TTATCAGO	CAACCTG	TAT-A	AGACAG	AGCAACO	TAAGCCATC	ΓA	
M00262:88:00+GTAAGGAG	•		CGAA	GTTATCAGACO	GGGGAC	TTATCAGO	CAACCTG	TAT-A	AGACAG	AGCAACO	TAAGCCATC	TATAAGTT	ACACGAATACTAGTAA
M00262:88:00+GTAAGGAG	CTTCGTA	TAGGATACTTTA	TACGAA	GTTATCAGACO	GGGGAC	TTATCAGO	CAACCTG	TTAT-A	AGACAG	AGCAACO	TAAGCCATC	TATAAGTT	ACACGAATACTAGTAA
M00262:88:00+GTAAGGAG	▶ CTTCGTA	TAGGATACTTTA	TACGAA	GTTATCAGACO	GGGGAC	TTATCAGO	CAACCTG	TAT-	AGACAG	AGCAACO	TAAGCCATC	TATAAGTT	ACACGAATACTAGTAA
M00262:88:00+GTAAGGAG	•	CTTTA	TACGAA	GTTATCAGACO	GGGGAC	TTATCAGO	CAACCTG	TAT-/	AGACAG	AGCAACO	TAAGCCATC	TATAAGTT	ACACGAATACTAGTAA
M00262:88:00+GTAAGGAG	•				AC	TTATCAGO	CAACCTG	TAT-A	AGACAG	AGCAACO	TAAGCCATC	TATAAGTT	ACACGAATACTAGTAA
M00262:88:00+GTAAGGAG	▶ CTTCGTA	TAGGATACTTTA	TACGAA	GTTATCAGACO	GGGGAC	TTATCAGO	CAACCTG	TAT-A	AGACAG	AGCAACO	TAAGCCATC	TATAAGTT	ACACGAATACTAGTAA
M00262:88:00+GTAAGGAG	▶ CTTCGTA	TAGGATACTTTA	TACGAA	GTTATCAGACO	GGGGAC	TTATCAGO	CAACCTG	TTAT-	AGACAG	AGCAACO	TAAGCCATC	TATAAGTT	ACACGAATACTAGTAA
M00262:88:00+GTAAGGAG	•				C	TTATCAGO	CAACCTG	TTAT-	AGACAG	AGCAACO	TAAGCCATC	TATAAGTT	ACACGAATACTAGTAA
M00262:88:00+GTAAGGAG	•												
M00262:88:00+GTAAGGAG	•												

After clicking on the first aligned read and scrolling along, you can clearly see that this simple approach works extremely well to identify the first insertion location, avoiding the need to work through all of the steps outlined above. But this may not always be the case. There can also be confusing areas like this one below, which is nowhere near any of the insertion sites but presumably reflects a region that has random short identity with the transposon sequence;

• • •	● ● Genome Alignment — Editor							
_n ==	la	0 👻	Ø3 R	》 순. 🛛 🛈		ACGTAC ACGTAC		
Unlocked Text View	v Prefs Replica		نتين 🚌 Remove Seas Alic	in Create Transla		Mismatch Next Mismatc		
Editor	Map	Features	Annotations	Text	SNPs	Problems		
Sort					1	1		
USIN	Genome ATCAGAA	81760 817 AAACGTAAAAAAATATT		81790 818 ACATAGAAAAAAT		81820 81830 CCTTGTCTTTAACTACTAC	81840 81850 STAAACTTGCTAAAATGC	
Consensus TAGATTAAAAATCATCATCAATAAAATCATCC								
M00262:88:00+GTA	AGGAG 4							
M00262:88:00+GTA		ATGTATGGTTGTAATAA	TAAAACTGTAAATGGTA	TGCTAGATTAAAAAT	ΑΑΑΑΤΓΑΤΓΑΑΤΑΑΑΤ	TCCGGGAAACTTCCGGAGAT	TTTTTGAGGCAGCGCGC	
M00262:88:00+GTA						TCCGGGAAACTTCCGGAGAT		
M00262:88:00+GTA						TCCGGGAAACTTCCGGAGAT		
M00262:88:00+GTA		ATGTATGGTTGTAATAA	ТААААСТGТАААТGGTA	TGCTAGATTAAAAAT	AAAATCATCAATAAAT	TCCGGGAAACTTCCGGAGAT	TTTTTGAGGCAGCGCGC	
M00262:88:00+GTA	AGGAG > ITATTTA	ATGTATGGTTGTAATAA	ГААААСТGТАААТGGTA	TGCTAGATTAAAAAT	АЛААТСАТСААТАААТ	TCCGGGAAACTTCCGGAGAT	TTTTTGAGGCAGCGCGC	
M00262:88:00+GTA	AGGAG ┥ ITATTTA	ATGTATGGTTGTAATAA	ГААААСТGТАААТGGTA	TGCTAGATTAAAAAT	ААААТСАТСААТАААТ/	TCCGGGAAACTTCCGGAGAT	TTTTTGAGGCAGCGCGC	
M00262:88:00+GTA	AGGAG		ACTGTAAATGGTA	TGCTAGATTAAAAAT	ААААТСАТСААТАААТ/	TCCGGGAAACTTCCGGAGAT	TTTTTGAGGCAGCGCGC	
M00262:88:00+GTA	AGGAG 🕨 ITATTTA	ATGTATGGTTGTAATAA	ТААААСТGТАААТGGTA	TGCTAGATTAAAAAT	ААААТСАТСААТАААТ	TCCGGGAAACTTCCGGAGAT	TTTTTGAGGCAGCGCGC	
M00262:88:00+GTA	AGGAG 📢 ITATTTA	ATGTATGGTTGTAATAA	ТААААСТGТАААТGGTA	TGCTAGATTAAAAAT	ААААТСАТСААТАААТ	TCCGGGAAACTTCCGGAGAT	TTTTTGAGGCAGCGCGC	
M00262:88:00+GTA	AGGAG ┥ ITATTTA	ATGTATGGTTGTAATAA	ТААААСТGТАААТGGTA	TGCTAGATTAAAAAT	ААААТСАТСААТАААТ	TCCGGGAAACTTCCGGAGAT	TTTTTGAGGCAGCGCGC	
M00262:88:00+GTA	AGGAG 🕨 ITATTTA	ATGTATGGTTGTAATAA	ТААААСТGТАААТGGTA	TGCTAGATTAAAAAT	ААААТСАТСААТАААТ	TCCGGGAAACTTCCGGAGAT	TTTTTGAGGCAGCGCGC	
M00262:88:00+GTA	AGGAG ┥ ITATITA	ATGTATGGTTGTAATAA	ТААААСТGТАААТGGTA	TGCTAGATTAAAAAT	ААААТСАТСААТАААТ	TCCGGGAAACTTCCGGAGAT	TTTTTGAGGCAGCGCGC	
M00262:88:00+GTA	AGGAG 📢 ITATTTA	ATGTATGGTTGTAATAA	ГААААСТGТАААТGGTA	TGCTAGATTAAAAAT	ААААТСАТСААТАААТ	ATC .		
M00262:88:00+GTA	AGGAG 📢 ITATTTA	ATGTATGGTTGTAATAA	ГААААСТGТАААТGGTA	TGCTAGATTAAAAAT	ΑΑΑΑΤCΑΤCΑΑΤΑΑΑΤ	TCC GGGAAACTTCCGGAGAT	TTTTTGAGGCAGCGCGC	
M00262:88:00+GTA	AGGAG 🕨 ITATTTA	ATGTATGGTTGTAATAA	ГААААСТGТАААТGGTA	TGCTAGATTAAAAAT	ΑΑΑΑΤCΑΤCΑΑΤΑΑΑΤ	TCC GGGAAACTTCCGGAGAT	TTTTTGAGGCAGCGCGC	
M00262:88:00+GTA	AGGAG 🕨 ITATTTA	ATGTATGGTTGTAATAA	ГААААСТGТАААТGGTA	TGCTAGATTAAAAAT	AAAATCATCAATAAAT	TCC GGGAAACTTCCGGAGAT	TTTTTGAGGCAGCGCGC	
M00262:88:00+GTA	AGGAG 📢 ITATTTA	ATGTATGGTTGTAATAA	ТААААСТGТАААТGGTA	TGCTAGATTAAAAAT	AAAATCATCAATAAAT	TCC GGGAAACTTCCGGAGAT	TTTTTGAGGCAGCGCGC	
M00262:88:00+GTA	AGGAG 🕨 ITATTTA	ATGTATGGTTGTAATAA	ТААААСТGТАААТGGTA	TGCTAGATTAAAAAT	ΑΑΑΑΤCΑΤCΑΑΤΑΑΑΤ	TCC GGGCTGTCTCTTATAC/	ACATCTCCGAGCCCACGA	
M00262:88:00+GTA		ATGTATGGTTGTAATAA	ТААААСТGТАААТGGTA	TGCTAGATTAAAAAT	ΑΑΑΑΤCΑΤCΑΑΤΑΑΑΤ	TCC GGGAAACTTCCGGAGAT	TTTTTTGAGGCAGCGCGC	
M00262:88:00+GTA	AGGAG 🖣 ITATITA	ATGTATGGTTGTAATAA	ТААААСТGТАААТGGTA	TGCTAGATTAAAAAT	ΑΑΑΑΤCΑΤCΑΑΤΑΑΑΤ	TCC GGGAAACTTCCGGAGAT	TTTTTTGAGGCAGCGCGC	
M00262:88:00+GTA	AGGAG 🖣 ITATITA	ATGTATGGTTGTAATAA	ТААААСТGТАААТGGTA	TGCTAGATTAAAAAT	ΑΑΑΑΤCΑΤCΑΑΤΑΑΑΤ	TCC GGGAAACTTCCGGAGAT	TTTTTTGAGGCAGCGCGC	
M00262:88:00+GTA	AGGAG 🕨 ITATITA	ATGTATGGTTGTAATAA	ГААААСТGТАААТGGTA	TGCTAGATTAAAAAT	ΑΑΑΑΤCΑΤCΑΑΤΑΑΑΤ	TCC GGGAAACTTCCGGAGAT	TTTTTTGAGGCAGCGCGC	
M00262:88:00+GTA	AGGAG 🖣 ITATITA	ATGTATGGTTGTAATAA	ГААААСТGТАААТGGTA	TGCTAGATTAAAAAT	ΑΑΑΑΤCΑΤCΑΑΤΑΑΑΤ	TCC GGGAAACTTCCGGAGAT	TTTTTTGAGGCAGCGCGC	
M00262:88:00+GTA						TCC GGGAAACTTCCGGAGAT		
M00262:88:00+GTA						TCC GGGAAACTTCCGGAGAT		
M00262:88:00+GTA		ATGTATGGTTGTAATAA	ТААААСТGТАААТGGTA	TGCTAGATTAAAAAT	ΑΑΑΑΤCΑΤCΑΑΤΑΑΑΤ	TCC GGGAAACTTCCGGAGAT	TTTTTGAGGCAGCGCGC	
M00262:88:00+GTA		ATGTATGGTTGTAATAA	ТААААСТGТАААТGGTA	TGCTAGATTAAAAAT	ΑΑΑΑΤCΑΤCΑΑΤΑΑΑΤ	TCCGGGAAACTTCCGGAGAT	TTTTTTGAGGCAGCGCGC	
M00262:88:00+GTA								
M00262-88-00 +GTA	AGGAG 4							

However, with this relatively clean set of data, all five insertion sites show up clearly and you just have to walk through a number of obviously incorrect spurious alignments to find the important alignments.