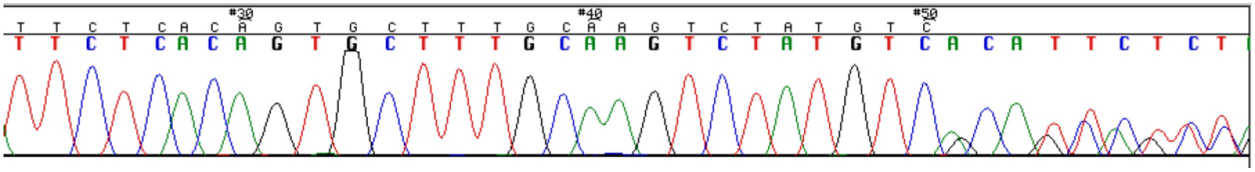
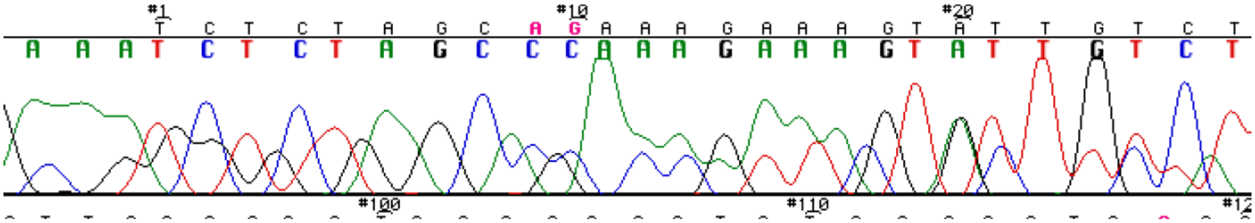
These notes were written using Sequencher but the overall concepts can be applied to any gene-editing software. For those who are using Geneious: contig is just another name for an alignment of like sequences, which you can do in Geneious. Doing so allows you to 1) find all the seqs that are HIV (they will align with HXB2 or whatever HIV reference strand you are using) and 2) batch edit multiple sequences with the same IS all at once. It can save you a bit of time, but editing them one by one is also an acceptable approach.

Steps:

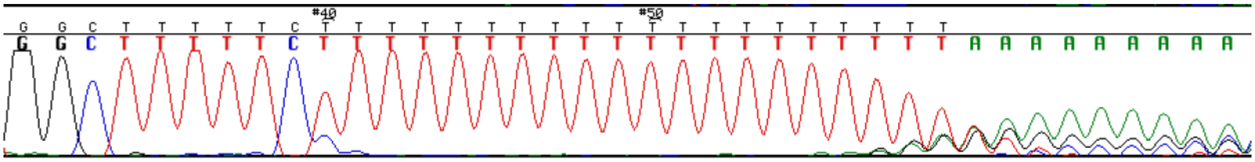
1) Import all abi files into Sequencher
2) Select all files and contig
3) Open all files (including contigs)
4) Visually inspect each chromatogram/contig:
    a. Trim to TCTCTAGCA or GCCCTTCCA at the beginning of the sequence.
        i. The tool only identifies consensus LTR. If your LTR sequence is not consensus, you will need to trim the entire LTR sequence before submitting to the tool.
    b. Trim obviously low quality sequence from the 3' end
    c. Look for recombination in the body of the sequence (single peaks become 2 peaks). Delete double-peaked region.



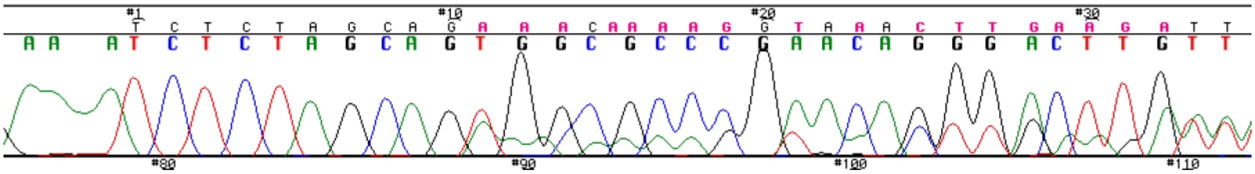
    d. Look for recombination within the LTR (LTR region has double peaks but you can see the AAAATCTCTAGCA). Add "recomb" to the name of the sequence.



    e. Look for homopolyA or homopolyT runs. All sequence following these runs will become unreadable. Trim the unreadable sequence.



    f. Look for mixed sequences → LTR will be single peaks but the human sequence will have double peaks.



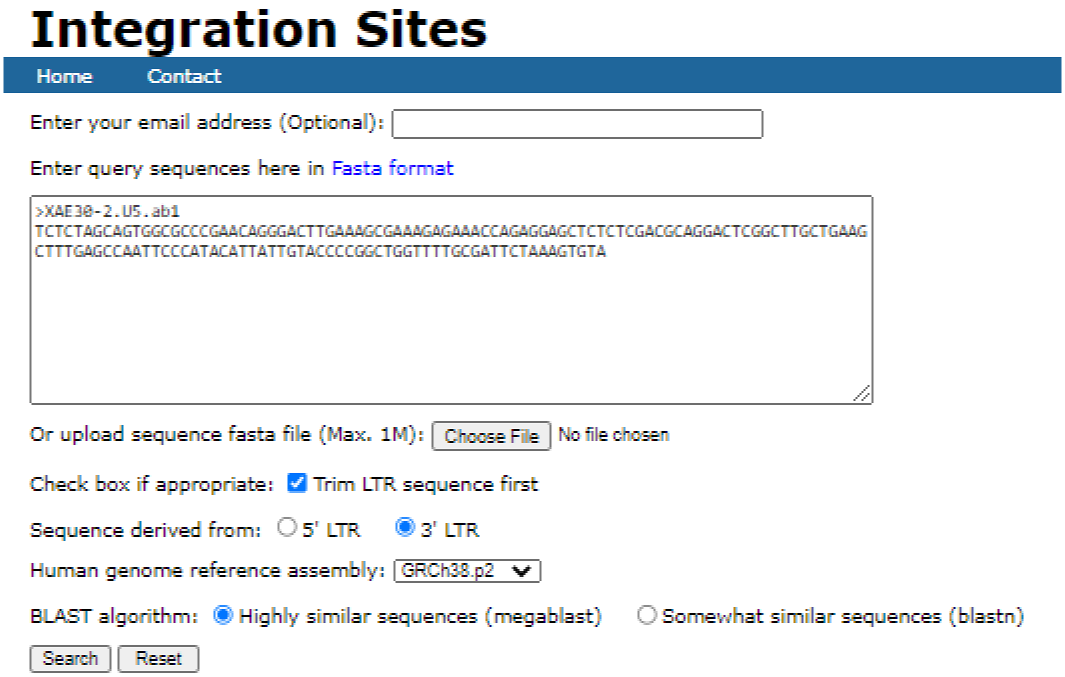
i. Add “mix” to the sequence name.
ii. Remove mixed sequences from contigs.
iii. Look to see if one of the sequences is a 5’loop by teasing out the UTR sequence GTGGCGCCC immediately following TCTCTAGCA (3’). Or, the GTCCCCCTTTT sequence immediately following GCCCTTCCA (5’). Add “mix5’loop” or “mixnefloop” to the sequence name.

g. Look for sequences with no recognizable LTR. If need be, try contig’ing a reference LTR sequence to see if there is a partial deletion (you will usually need to edit the full LTR sequence at the beginning of the sequence to create a region long enough to contig).
h. If the sequence is an obvious FAIL, put an ‘x’ in front of the sequence name so that it sorts to the bottom of the list.
i. Some sequences will look like they kinda worked, but in fact are low quality cross-contamination from another strong, clean sequencing reaction (who knows whether physical or digital?). If you have a strong, clean reaction for one of your reads, you will likely see one or more identical low quality sequences in the same set (usually from faint bands that would have otherwise been a FAIL). They should be considered FAILS.

5) At this point all single sequences and contigs should be edited. Now go back and tease apart the mixed and recomb seqs (adding the identifiers to the name of the seq allows you to easily find them). This is something that requires practice. Usually a BLAT search of the raw sequence will identify one of the sequences or one of the sequences will be a 5’ loop or nef loop. Once you have the first sequence identified and edited to match reference, you can identify the second sequence by editing each peak in the chromatogram with the alternate peak at that position. Once you have 30 bases or so, you can try BLATing the sequence and often times you will get a match. Then you can use the reference sequence to edit additional bases if necessary. Sometimes blastn will do a better job than BLAT at finding short, non-exact matches (make sure to select “Somewhat similar sequences (blastn)” under program selection on the BLAST page). If one of your mixed seqs happens to be in a repetitive element that hasn’t been found before, it may be impossible to find the correct location.

NOTE BLAT is much faster and easier to use than BLAST for creating reference sequences. With BLAT, you can directly copy the matching reference sequence AND a hundred bases both before and after your match. The copied sequence can be directly pasted into a new sequence file in Sequencher (and I would assume in Geneious) and then contig’d to your query (as a reference for editing your sequence). BLAST gives you an alignment as output, which cannot be directly copied and pasted.

6) Once everything is edited, dissolve all contigs
7) Select all edited seqs and export as a concatenated fasta file (do not export the ‘fails’ or mixes that you were not able to tease apart).
8) **Open the fasta file in the online IS tool (https://indra.mullins.microbiol.washington.edu/integrationsites/). Select 5’ or 3’ LTR and ‘trim LTR sequence first’. Click the ‘search’ button.**



Integration Sites

Enter your email address (Optional):

Enter query sequences here in Fasta format

```
>XAE30-2.U5.ab1
TCTCTAGCAGTGGCGCCCGAACAGGGACTTGAAAGCGAAAGAGAAACCAGAGGAGCTCTCTCGACGCAGGACTCGGCTTGCTGAAG
CTTTGAGCCAATTCCCATACATTATTGTACCCCGGCTGGTTTTGCGATTCTAAAGTGTA
```

Or upload sequence fasta file (Max. 1M): No file chosen

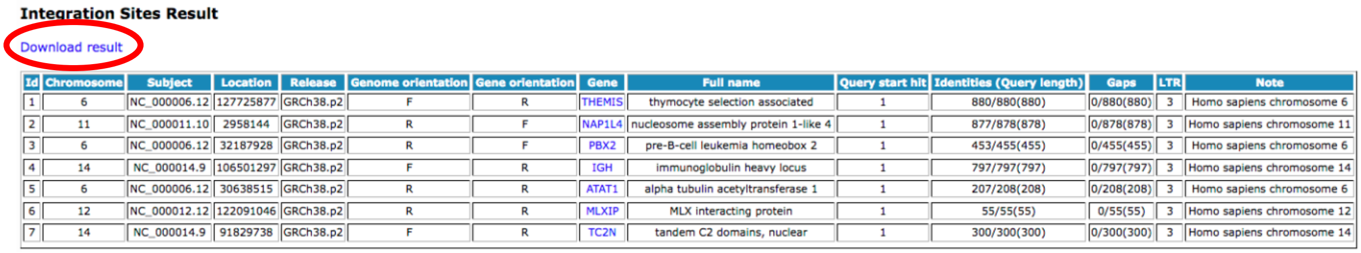
Check box if appropriate: ☑ Trim LTR sequence first

Sequence derived from: ○ 5' LTR ● 3' LTR

Human genome reference assembly: GRCh38.p2

BLAST algorithm: ● Highly similar sequences (megablast) ○ Somewhat similar sequences (blastn)

9) When the search is complete, click the download link and open the file in excel. Name/save the text file as an excel workbook.



Integration Sites Result

Id	Chromosome	Subject	Location	Release	Genome orientation	Gene orientation	Gene	Full name	Query start hit	Identities (Query length)	Gaps	LTR	Note
1	6	NC_000006.12	127725877	GRCh38.p2	F	R	THEMIS	thymocyte selection associated	1	880/880(880)	0/880(880)	3	Homo sapiens chromosome 6
2	11	NC_000011.10	2958144	GRCh38.p2	R	F	NAP1L4	nucleosome assembly protein 1-like 4	1	877/878(878)	0/878(878)	3	Homo sapiens chromosome 11
3	6	NC_000006.12	32187928	GRCh38.p2	R	F	PBX2	pre-B-cell leukemia homeobox 2	1	453/455(455)	0/455(455)	3	Homo sapiens chromosome 6
4	14	NC_000014.9	106501297	GRCh38.p2	F	R	IGH	immunoglobulin heavy locus	1	797/797(797)	0/797(797)	3	Homo sapiens chromosome 14
5	6	NC_000006.12	30638515	GRCh38.p2	R	R	ATAT1	alpha tubulin acetyltransferase 1	1	207/208(208)	0/208(208)	3	Homo sapiens chromosome 6
6	12	NC_000012.12	122091046	GRCh38.p2	R	R	MLXIP	MLX interacting protein	1	55/55(55)	0/55(55)	3	Homo sapiens chromosome 12
7	14	NC_000014.9	91829738	GRCh38.p2	F	R	TC2N	tandem C2 domains, nuclear	1	300/300(300)	0/300(300)	3	Homo sapiens chromosome 14

10) Sort the seqs by chromosome and cut/paste those that match to HIV into a separate tab (= “HIV” tab).

11) Sort the seqs by ID. Identify those that have multiple hits.
    a. Determine whether the multiple hits are actually from the same site (ie, overlapping genes) or are from multiple different sites because the IS is in a repetitive element.
    b. If HIV is integrated in overlapping genes (same site), cut/paste one of the data lines to a new tab (= “overlapping” tab).
    c. If the IS has multiple different sites, go back to the edited abi file (in Sequencher) to see if you can add any additional length back to the edited sequence. If the sequence is not long enough to get a single identity, then cut/paste the seqs into a separate tab in your excel file (= “not identifiable” tab). If they are not identifiable they shouldn’t go in the database, but the data should still be kept for future reference.

12) There may be some sequences that the IS tool will classify as ‘no significant match’. These are usually shorter sequences, or sequences with a lot of As and Ts (low complexity). Try re-submitting the sequence through his tool, but click the blastn option (less stringent). This will usually locate the site.

13) Sort the seqs by qstart – larger to smaller. qstart should be ‘1’.
    a. Some qstart that are larger than ‘1’ will be reads off of the 5’ LTR→UTR→human (or 3’ LTR→nef→human). These may be real, but could also be PCR artifacts. I cut/paste them into a separate tab. Those that are found more than once are proliferating and thus are

real. Those that are found only once are suspect. Add to another tab to separate from other data.

b. Some qstart that are larger than ‘1’ will be due to recombination within the sequence or fusion with other sequences. If you BLAT the full sequence, you should be able to tell where the recombination occurs. Trim the offending sequence off of the 3’ end until you get a ~100% match with a qstart of 1. Replace the incorrect data in your spreadsheet.

Sequence with a qstart of 87 – blue matches to reference:

```
tctctagcac acgctgcgta actccactta cacgaagatt agagggagac  50
aacagactga agaaacaggc acacgctgtg taactccact tacatgAAGA  100
TTAGAGGGAG ACAAcAGACT GAAGAAACAG GCACACGCTG CGTAACTCCA  150
CTTACACGAA GATTAGAGGG AGACAAcAGA CTGAAGAAAC AGGCgCACGC  200
TGCGTAACTC CACTTAcACG AAGATTAGAG GGAGACAACA GACTGAAGAA  250
ACAGGCaCAC GCTGtGTAAC TCCACTTACA CGAAGATTAG AGGGAGACAA  300
CgGACTGAAG AAcCAGGCAC ACGCTGTGTA ACTCCgCTTA CATGAAGGtT  350
TTAGAGCAac CAGGTTCACA GACTCGGAGC GGAGTGGTTG CTACCAGGGG  400
CTGCAGaGAG GGGAATGGGG AGCTGGT
```

Delete blue sequence from above and reBLAT. Now has qstart of 1 and 100% identity:

```
tctctagcaC ACGCTGCGTA ACTCCACTTA CACGAAGATT AGAGGGAGAC  50
AACAGACTGA AGAAACAGGC ACACGCTGTG TAACTCCACT TACATG
```

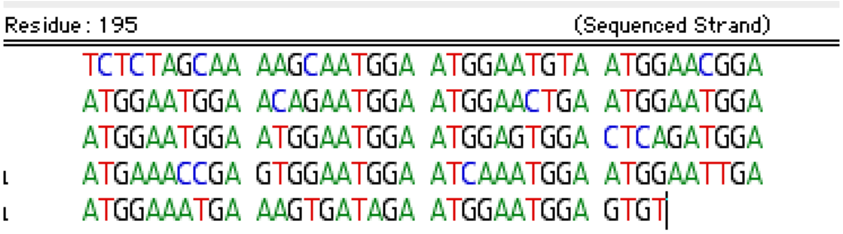
c. Some qstart will be ‘3’. This is usually due to a mismatch at position 2 with reference. Trimming off the mismatch is a function of BLAST (which is what the IS tool uses). If you BLAT, it will give you a qstart of 1, and show you the mismatch. What I usually do is go ahead and edit the mismatch to match reference so that I can get a correct location (replacing the data in my spreadsheet). The mismatch may be a real polymorphism, or it may be a PCR error.

d. Infrequently there are extra bases between the LTR and human for which a source cannot be identified. Check the original sequence for errors and if none, then leave a note that it was checked.

e. Infrequently the first base off of the LTR is a mismatch, again, check the original sequence for errors, and leave a note that the sequence was checked. Unless the sequence is from a proliferating clone, there is no way to know whether it’s a true polymorphism or PCR error.

f. Infrequently there are integration sites in areas for which there is a high degree of polymorphism (ie, telomeres, centromeres, repetitive sequences etc). Often times they won’t have a good match with reference and in some cases the closest hits will not have a qstart of 1. I handle these on a case-by-case basis as to whether they should go into the database.

IS with repetitive sequence:

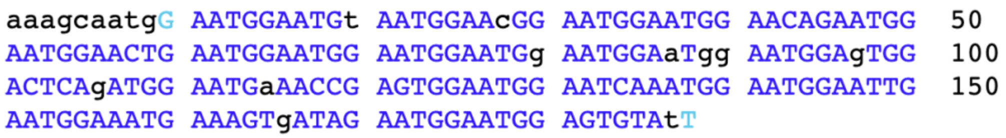


```
Residue : 195                                   (Sequenced Strand)

TCTCTAGCAA AAGCAATGGA ATGGAATGTA ATGGAACGGA
ATGGAATGGA ACAGAATGGA ATGGAACTGA ATGGAATGGA
ATGGAATGGA ATGGAATGGA ATGGAGTGGA CTCAGATGGA
ATGAAACCGA GTGGAATGGA ATCAAATGGA ATGGAATTGA
ATGGAAATGA AAGTGATAGA ATGGAATGGA GTGT
```

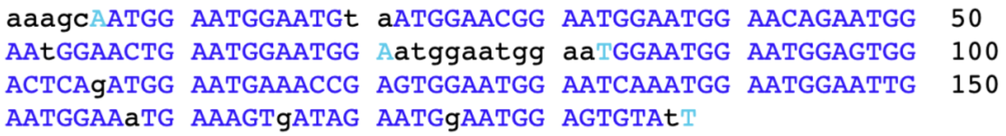
Place this sequence in the ‘not identifiable’ tab of your spreadsheet.

Closest match is ‘unplaced’ – ie, it hasn’t been mapped:



```
aaagcaatgG AATGGAATGt AATGGAAcGG AATGGAATGG AACAGAATGG   50
AATGGAACTG AATGGAATGG AATGGAATGg AATGGAaTgg AATGGAgTGG   100
ACTCAgATGG AATGaAACCG AGTGGAATGG AATCAAATGG AATGGAATTG   150
AATGGAAATG AAAGTgATAG AATGGAATGG AGTGTAtT
```

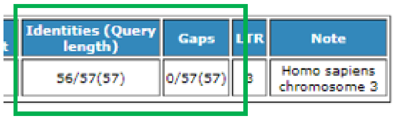
Second closest match is also ‘unplaced’:



```
aaagcAATGG AATGGAATGt aATGGAACGG AATGGAATGG AACAGAATGG   50
AAtGGAACTG AATGGAATGG Aatggaatgg aaTGGAATGG AATGGAGTGG   100
ACTCAgATGG AATGAAACCG AGTGGAATGG AATCAAATGG AATGGAATTG   150
AATGGAAaTG AAAGTgATAG AATGgAATGG AGTGTAtT
```

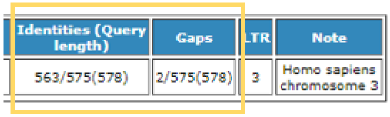
14) Recheck all sequences for which the IS tool did not find an LTR (TCTCTAGCA or GCCCTTCCA). Correct errors where found, or if no error, but there is an LTR with a mismatch or small deletion, delete the full LTR sequence and re-submit to the IS tool to get a correct location (replacing the data in the spreadsheet). TCTCTAGCA is very well conserved, but GCCCTTCCA is not!

15) As a final check, look to see if there are more than 2 mismatches with reference. If there are more, then check the sequence to make sure that there are no errors/recombinations which could affect identification. **Often times only a few mismatches will distinguish repetitive elements located in different sites.** If trimming the sequence from the 3’ end changes the IS location, put the trimmed version through the IS tool and replace the correct data in your spreadsheet.



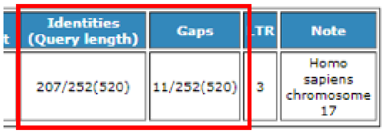
Identities (Query length)	Gaps	LTR	Note
56/57(57)	0/57(57)	3	Homo sapiens chromosome 3

OK



Identities (Query length)	Gaps	LTR	Note
563/575(578)	2/575(578)	3	Homo sapiens chromosome 3

Questionable (check sequence quality, if ok, then suspect repetitive region).



Identities (Query length)	Gaps	LTR	Note
207/252(520)	11/252(520)	3	Homo sapiens chromosome 17

Bad (suspect fusion)

16) Check for cross-contamination. Tip: keep a master file of all final, QC’d IS data so that you can check for cross contamination. You can sort by chromosome and then by location and look for the same integration site in different patients.