Artefacts in Hu et al data - substantiating details for reviewers

A) Amplification Bias and Controls

Duplicate reads and bottlenecking

Viewing the alignment of the reads to hg19 shows that there has been considerable over-amplification of the original DNA fragments such that there are typically 10 copies of the same original fragment that have been sequenced (Figure 1). This shows that care will be needed to ensure that the presence of reads from duplicated fragments does not cause misinterpretation of the data. The methods section of the Hu et al. paper states

> *"Briefly, low quality reads and duplicate reads, as well as adaptor contamination reads, were removed to obtain clean reads for subsequent analysis."*

suggesting that this potential problem was recognised.

The PCR Bottleneck Coefficient (PBC) (http://genome.ucsc.edu/ENCODE/qualityMetrics.html#definitions) is a measure of amplification bias, where a value less than 0.5 represents severe bottlenecking. We calculated the PBC values for two samples (T4826, T3157) which were 0.2256 and 0.2450. The read count distribution for one sample is plotted in Figure 1. This underlines the magnitude of over-amplification in these data and the importance of removing its biasing effects completely.

Mitochondrial inserts

In order to better understand the HPV insertion data we used as a control fragments which were in part mitochondrial sequence and in part nuclear genome sequence. Reads associated with known segments of mitochondrial DNA in the nuclear genome were excluded. 6797 different instances of hybrid fragments were found. All but six (0.1%) of these reads were duplicates of a single original fragment (representative example in Figure 2).

The six exceptions were examined and found to group as follows:

Number of instances	Conclusion	Example
3	Reads are from multiple copies of a single fragment. Our duplicate read removal wrongly identified one of the reads as being from an independent fragment.	Figure 3
1	Isolated repetitive sequence that maps to mitochondrial and nuclear genomes and is incorrectly identified as an insertion.	Figure 4
2	Highly repetitive regions where some sequences containing similar mitochondrial sequences are aligned, and these are incorrectly interpreted as insertions.	Figure 5

The absence of any mitochondrial integration site that is validated by more than one independent fragment indicates that, as expected, there is no evidence of additional mitochondrial integrations within the samples. Consequently, the 6791 cases that were identified were artefacts associated with single fragments consisting partly of mitochondrial DNA and partly of nuclear DNA. A likely cause of these artefacts is ligation of a very small fraction of the fragments during the fragmentation stage. Even if a scenario of an average of about 50 mitochondrial insertions per sample was seriously considered then it would not be conceivable that not a single one of these cases would be represented by more than one original DNA fragment. These artefacts demonstrate that single fragments cannot be considered as evidence of insertions.

Hybrid fragment rates for all samples are provided in sheet “Mitochondrial hybrids” in file “Validated insertions and mitochondrial control.xlsb”. This includes rates for mitochondrial hybrids as well as rates of single-fragment HPV hybrids.

6535 of the fragments that were mixed nuclear and mitochondrial DNA were represented by reads which spanned the location of ligation of the two types of DNA. Of these 5209 had a short homologous sequence (such as shown in Figure 2) and 963 had additional nucleotides inserted between the mitochondrial and nuclear genome sequences. This indicates that the presence of homologous sequences is a significant catalyst for the appearance of this artefact.

B) Reported HPV insertions

The raw read data was used to better understand the 3546 integration sites associated with tumour samples (i.e. excluding cell lines) from Supplementary Table S5. These can be divided up into four groups as follows:

1) HPV integration evidence based on single fragments

2449 of the sites were found to be associated with evidence from multiple copies of reads from a single fragment, indicating that the duplicate read removal process had not been fully effective. Therefore, for these sites there is insufficient evidence to show that an insertion has occurred (see A above).

In 1156 of the sites one or both of the paired end reads covers the join between the HPV and the human genome sequence, such as shown in Figure 6 to Figure 8.

In the remaining 1293 sites one of the paired-end reads is host genome and the other is HPV, indicating that the join between the HPV and the human genome sequence is between the two reads, such as shown in Figure 9 to Figure 12. Closer inspection of these cases shows that in just under 1000 of the 1293 sites, our alignment of the reads to the genomes clearly shows that the (artefactual) integration site is not at the position shown in Table S5. Figure 9, Figure 11 and Figure 12 examine the alignment of the reads to the genomes in three of these cases and shows that the reported insertion site is consistent with a misassembly of the two reads based on a short homologous sequence shared by the two reads. This results in an incorrect identification of the (artefactual) integration position as well as an incorrect identification of a homologous sequence as being a basis of the “integration”.

Figure 11 and Figure 12 consider cases where Table S6 shows a sequence that was allegedly obtained by Sanger sequencing. Oddly, the Sanger sequence is consistent with what appears to be the sequence obtained by misassembling the paired-end reads.

2) Additional integrations associated with a validated insert

There are 484 integrations where our algorithm can find no evidence for an integration but where there is another integration site very closely nearby. Our analysis shows that these reads are associated with and fully explained by the nearby integration site. It appears that the incorrect fragment assembly has incorrectly identified additional integrations associated with these reads. Figure 13 shows one such example with five incorrectly assigned integrations alongside a single validated integration.

Such false identification of additional integration sites will give the incorrect impression that there is an integration hot spot at this location.

3a) Integrations where our algorithm confirms the integration

There are 464 integrations where our algorithm validates the integration listed in Table S5. In each case our analysis nevertheless finds fewer independent fragments are associated with the integration than are shown in Table S5. Figure 14 shows one example.

3b) Evidence of cross contamination

However, within the integration sites in Table S5 we identified 53 integrations (out of which some but not all are single-fragment loci) at 20 locations where the exact same location appears in multiple samples, normally with SRR sample numbers that are very close to each other (for full details refer to the correspondence with the authors). This strongly suggests contamination between the samples. Figure 15 shows one such example where the same integration site appears in eight different samples, with SRR numbers 1610999, 1611001 to 1611003 and 1611104 to 1611108. Six of these eight cases appear in Table S5.

In many cases (such as the case shown) there are significant numbers of reads associated with the integration site in more than one sample, such that it is not possible clearly to identify one of the samples as being the source of the contamination. This suggests the possibility that in these cases there could be be some other source of contamination for all of the samples.

At minimum all but one of each of the cross contamination loci should be excluded (leaving 442 sites remaining), with there being some argument for excluding the evidence from all of these sites.

4) Cases not reproduced by our pipeline

Finally, there are 149 reported integration sites in Table S5 where our computational pipeline failed to identify any reads supporting insertions. The 13 locations associated with sample T3157 were examined "by hand" (Figure 23 to Figure 35) and in each case it could be shown that the location was not associated with an integration that was validated by reads from independent fragments. This suggests that the vast majority, if not all of the 149 locations that our software failed to identify are also artefacts. The figure of 87 per cent artefacts stated in the article has been calculated relative to the set of 3546-149=3397 reported insertions where our pipeline finds relevant associated reads.

C) Microhomology

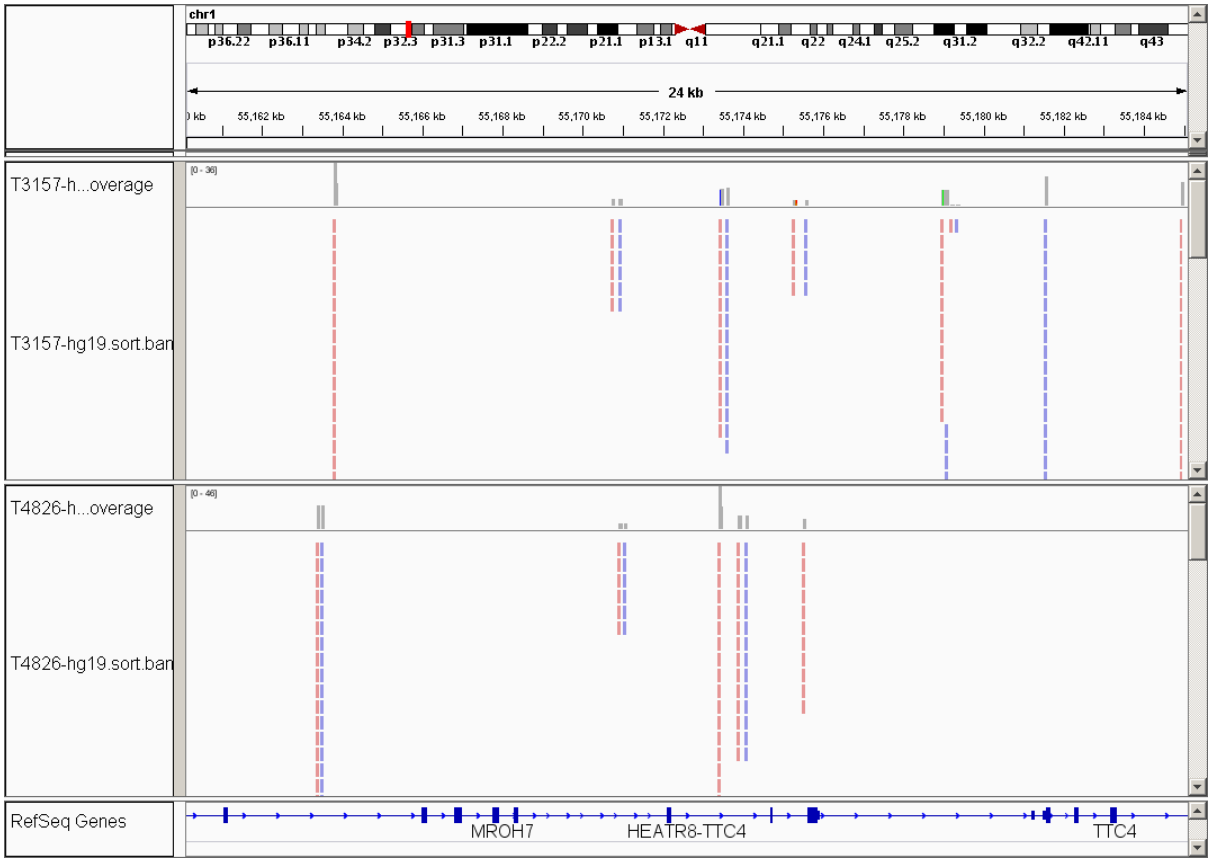
Of the 442 sites that remain (provided in sheet "Validated insertions" of file "Validated insertions and mitochondrial control"), our analysis suggests that there are only 380 sites where there are reads that cover the junction between the HPV sequence and the host genome. Of these sites there are 10 sites that show clear indications for other types of artefacts having occurred rather than actual insertions (details not provided and these 10 sites excluded in the following).

The remaining 370 sites divide up as follows:

Clean join between HPV and host genome	28
Single bp insert	28
More than 1 bp insert	131
Single bp micro-homology	33
More than 1 bp micro-homology	150

This suggests that less than 50% of the integrations are as a result of micro-homologous sequences, and that there are comparable numbers of integrations with additional sequence inserted between the HPV and the host sequence. Figure 16 to Figure 22 show some examples of HPV integrations with additional intervening sequence.

This indicates that, while micro-homology is probably the main driver behind the ligation artefacts, for genuine insertion sites the addition of nucleotides occurs with a similar frequency as micro-homology.



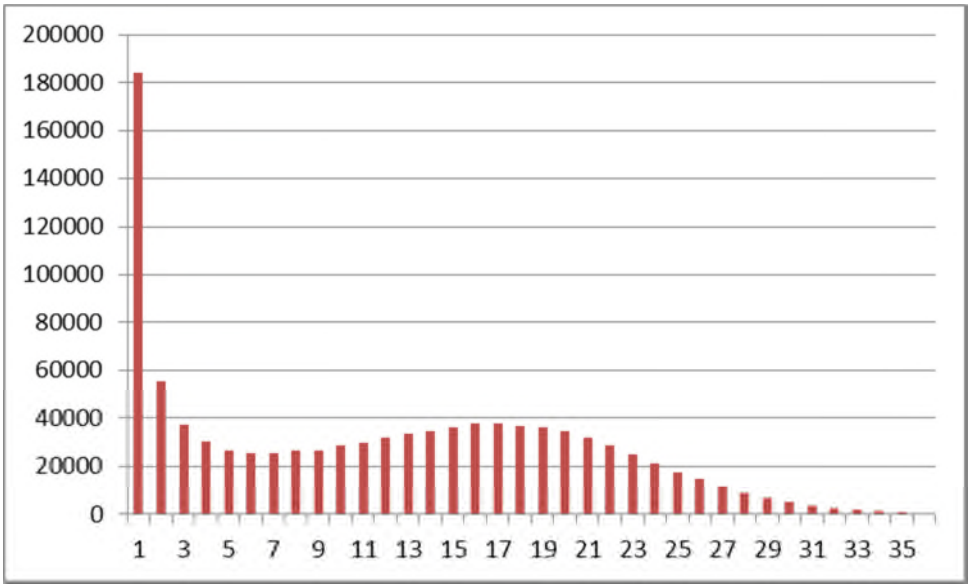
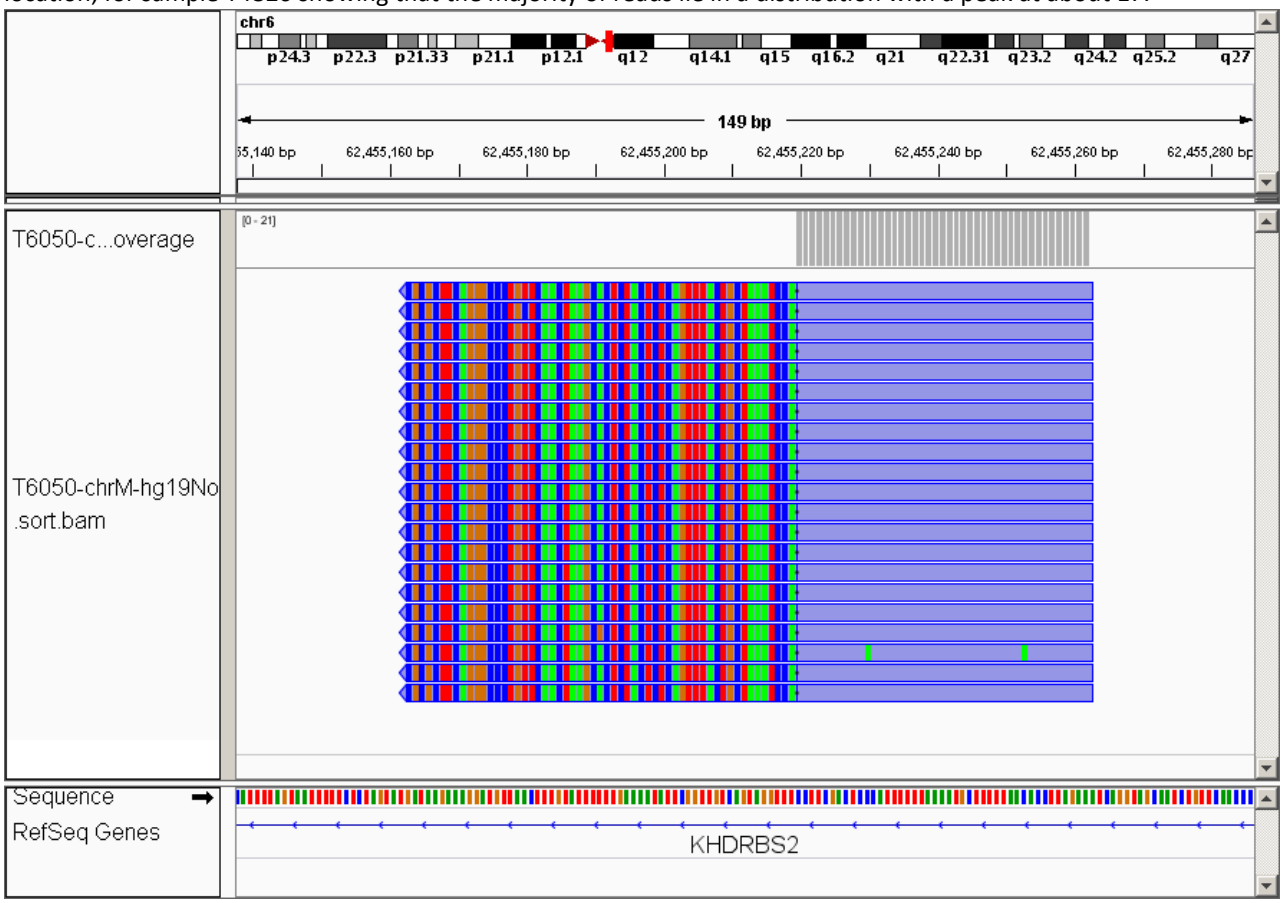


Figure 1 **Duplicate reads and bottlenecking**

The upper graph shows region chr1:55,160,098-55,185,133 for samples T3158 and T4826 showing that there are of the order of 10 reads per locus. The lower graph shows the distribution of numbers of reads at a given genomic location, for sample T4826 showing that the majority of reads lie in a distribution with a peak at about 17.




```
chr6:62455148-62455275     ATTTTTCTCTAGTATGATAAGAAAGGATGTTAACTTTGTATTTTTTGAAAATATTCGGTTGT
Read 1                  GGTGACGGGCGGTGTGTACGCGCTTCAGGGCCCTGTTCAACTAAGCACTCTACTCTCAGTTTACT
Read 2                             CGCGCTTCAGGGCCCTGTTCAACTAAGCACTCTACTCTCAGTTTACT
chrM:1504-1370 TTGAGGAGGGTGACGGGCGGTGTGTACGCGCTTCAGGGCCCTGTTCAACTAAGCACTCTACTCTCAGTTTACT
                        <-------Reads match ChrM sequence ------------------------------>
```

```
chr6 cont            CAGTAGGTTTCCTTCGACTCCCATTTTTTAAAATGCTTTTTAACACCTTAGAAATCAGGTAGCAAT
Read 1               GCTAAATCCACCTTCGACTCCCATTTTTTAAAATG
Read 2               GCTAAATCCACCTTCGACTCCCATTTTTTAAAATGCTTTTTAACACCTTAGAA
chrM:1504-1370       GCTAAATCCACCTTCGACCCTTAAGTTTCATAAGGGCTATCGTAGTTTTTCTGGGGTAGAAAA
                     <--match ChrM---->
                               <-----------Reads match chr6-------------->
Homologous sequence            CCTTCGAC
```

Figure 2 **One of the 6797 Mitochondrial 'integrations' that are based on evidence from a single fragment** SRR10997 = T6050 chr6:62,455,138-62,455,286. Only part of the 21 duplicate reads matches the chromosome 6 sequence. The lower section of the diagram shows the two aligned read sequences (Read 1 and 2) showing a substantial overlap region. The assembled fragment is then aligned to the chromosome M (blue shaded) and chromosome 6 sequence (red shaded), showing that it is a ligation of fragments from the two chromosomes. There is an eight nucleotide micro homologous sequence at the site of the ligation.

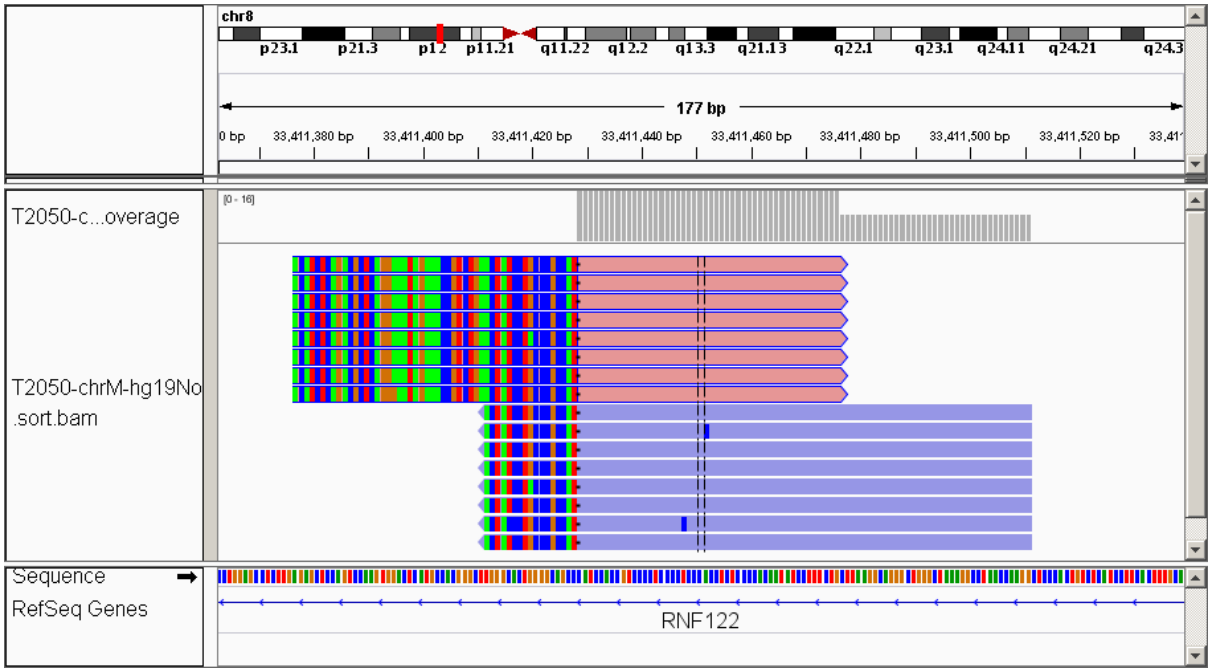


Figure 3 **Mitochondrial integrations that are artefacts**
Sample SRR1611100 = T2050 Eight copies of a single original fragment which consists of a section of mitochondrial DNA ligated to a section of nuclear DNA. Our algorithm still incorrectly identified the data as coming from two independent fragments.

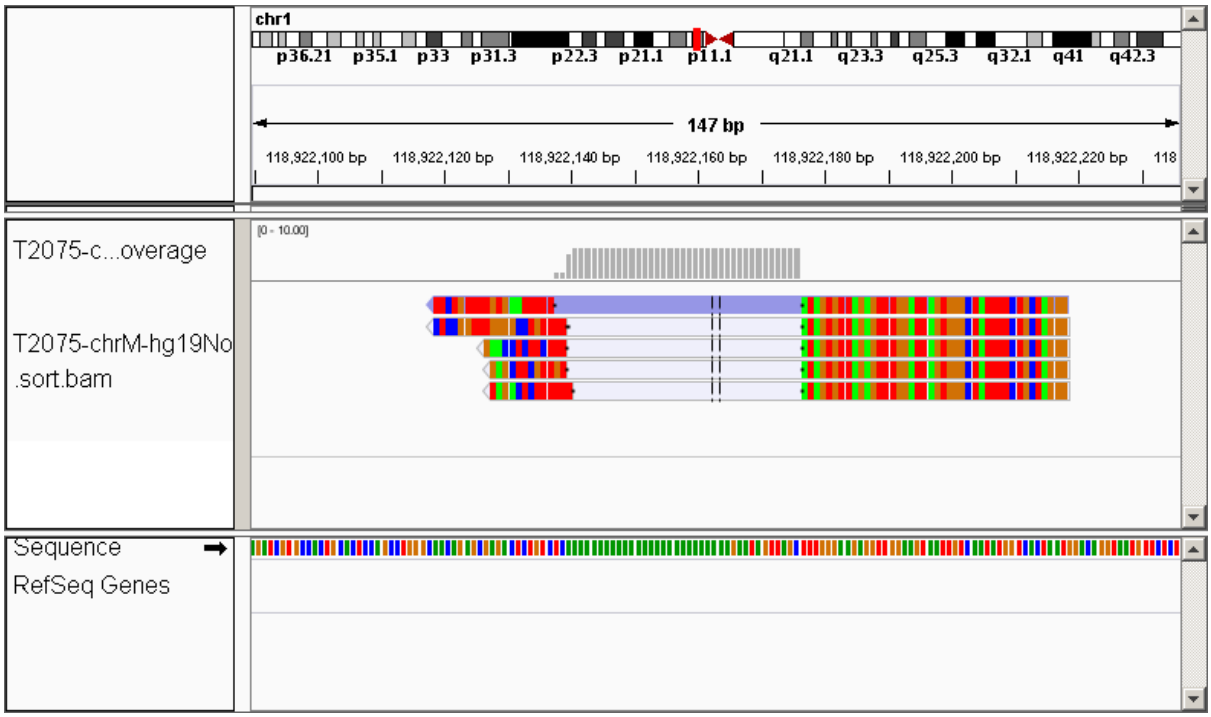


Figure 4 **Mitochondrial integrations that are artefacts**
Sample SRR1611107 = T2075. A repetitive region from the mitochondrial genome is incorrectly mapped to the nuclear genome giving a false positive indication of mitochondrial integration into the human genome.

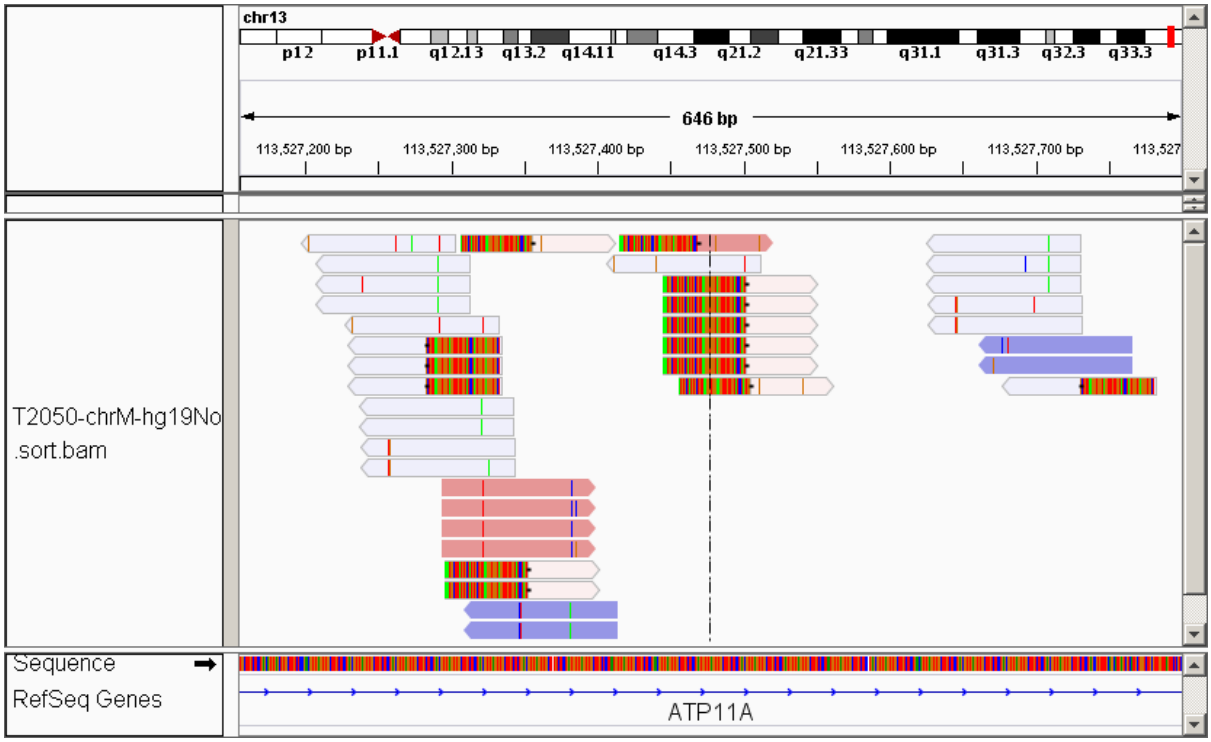


Figure 5 **Mitochondrial integrations that are artefacts**
SRR1611111 = T2079. chr13:113,527,155-113,527,801 A highly repetitive region where some mitochondrial fragments with similar sequences are mapped. The current software incorrectly interprets this as a mitochondrial integration corroborated by multiple independent fragments.

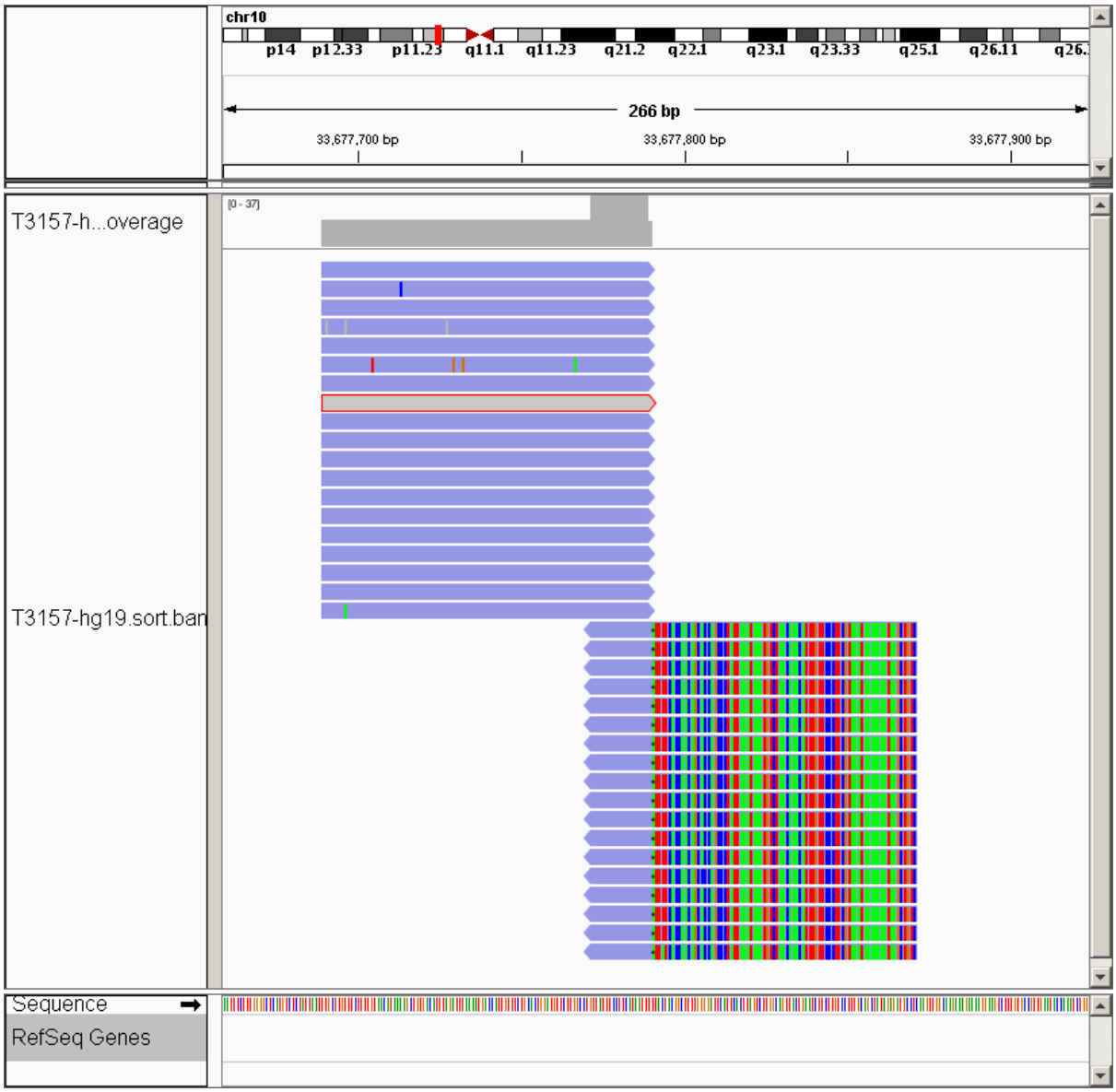


Figure 6 **HPV integration evidence based on single fragments: Read spans join**
chr10:33677790 T3157 Integration No 52. Paired end reads come from 19 copies of a single original fragment, Table S5 identifies 6 non-identical reads.

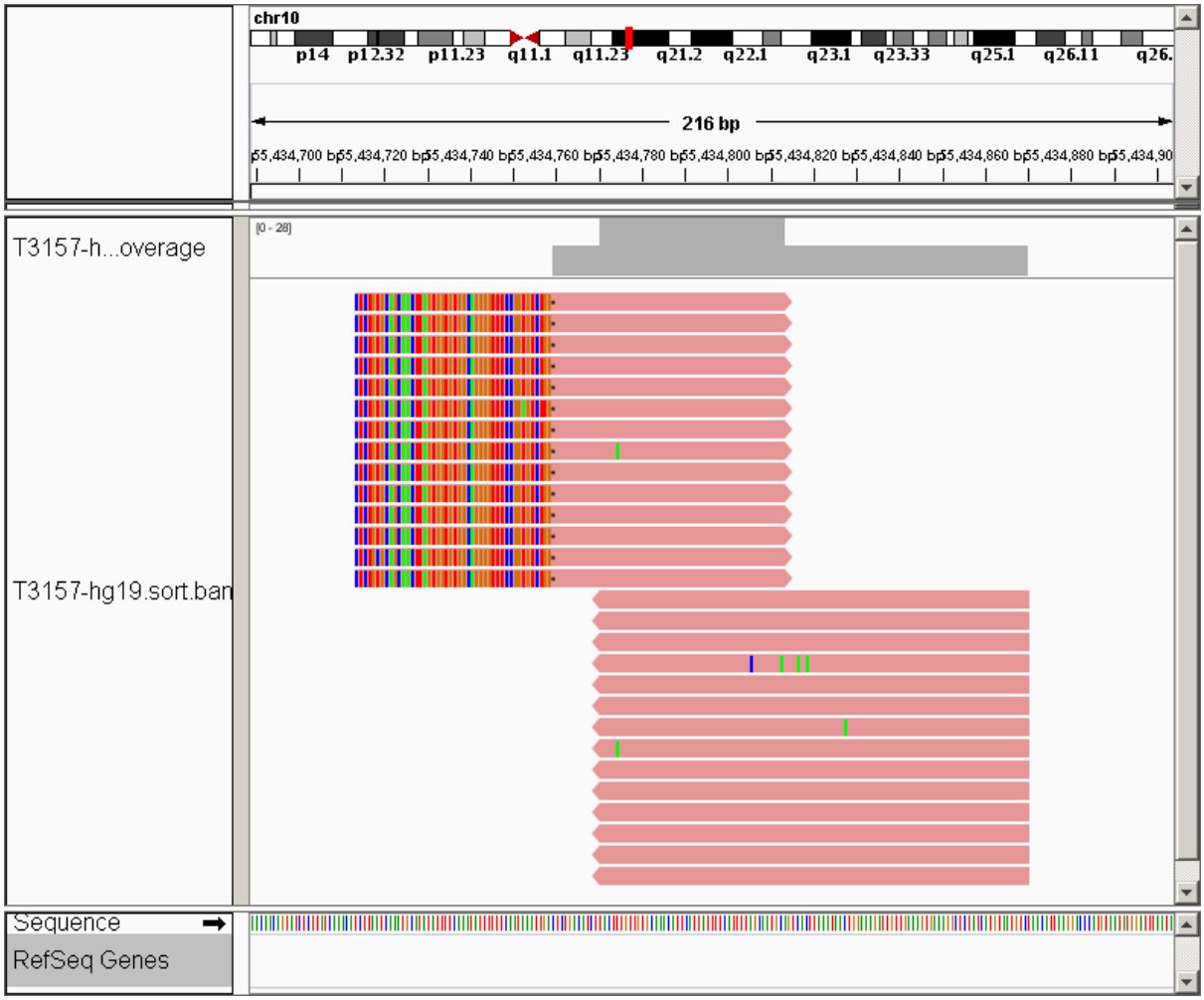


Figure 7 **HPV integration evidence based on single fragments: Read spans join**
chr10:55434760 T3157 Integration No 54. Table S5 identifies six independent reads.

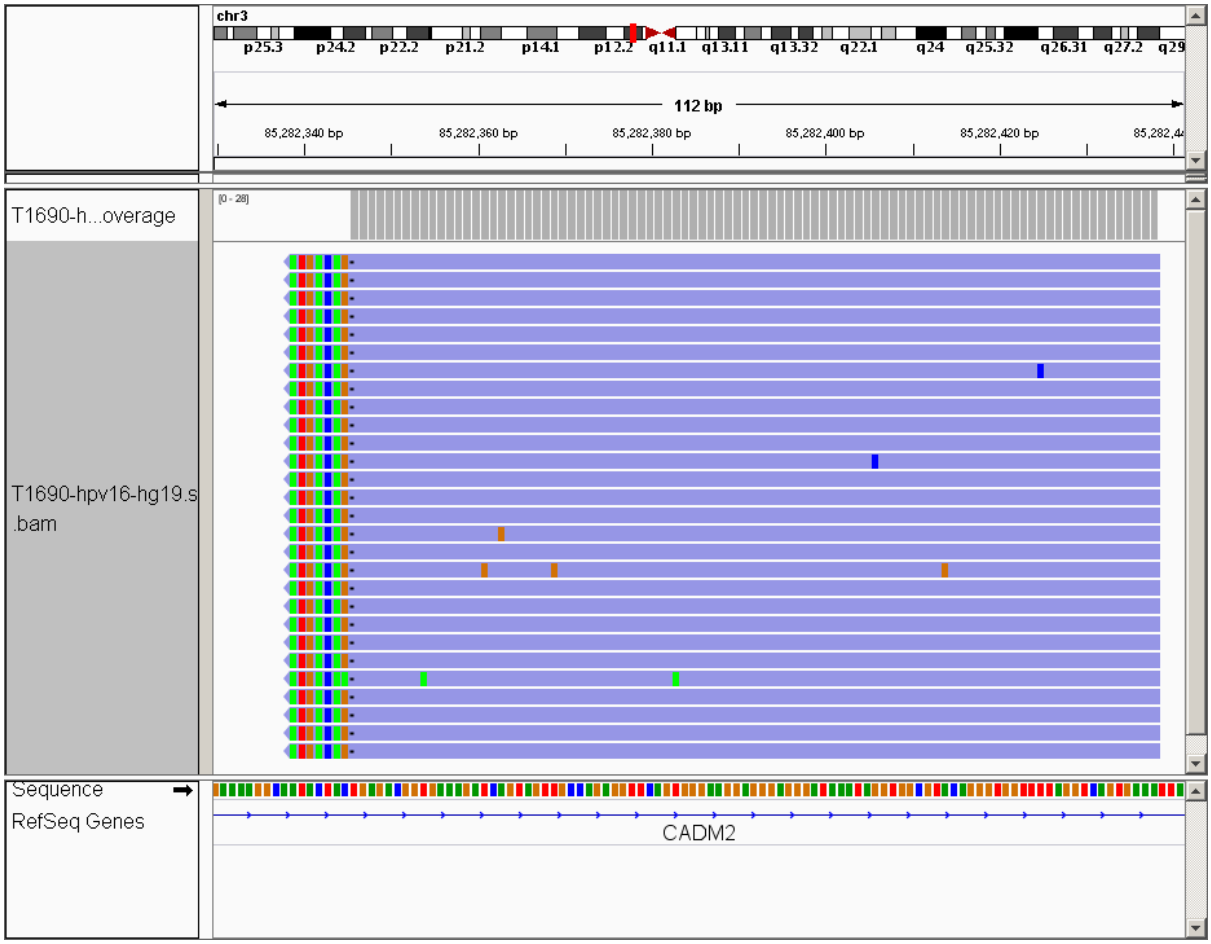
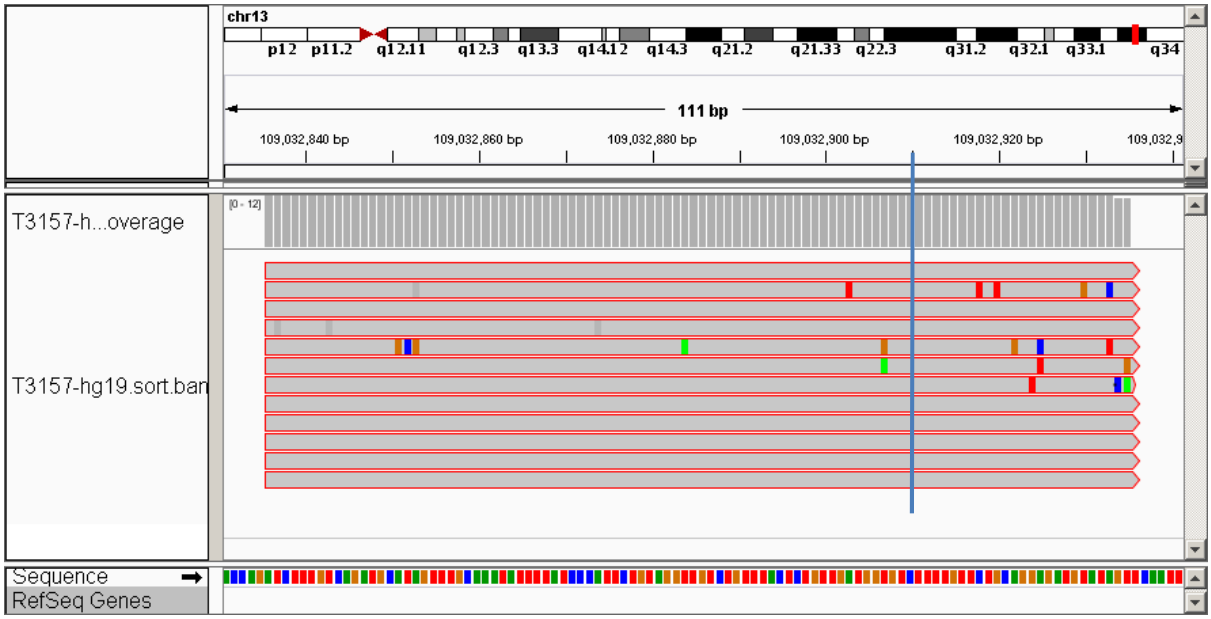


Figure 8 **HPV integration evidence based on single fragments: Read spans join**
T1690 = SRR1611116 Insert No 2. The 28 reads appear to be duplicates of a single original fragment. Taking PCR and sequencing errors into account, there is no evidence for the nine independent reads which are claimed providing evidence for this integration.



Alignment of Read2 to hg19

```
Hg19          TACCAGATCTTTGTCAGATGCATAGTTTGCAAATATTTTATCCCATTCTGTAGGTTGTCTGTTTAC
Read 2              ATCTTTGTCAGATGCATAGTTTGCAAATATTTTATCCCATTCTGTAGGTTGTCTGTTTAC
Read 2              ATCTTTGTCAGATGCATAGTTTGCAAATATTTTATCCCATTCTGTAGGTTGTCTGTTTAC
Read 1                                                      GAGGTAGGTCGTGGT
                    <------hg19 alignment from Table S5----------------------->

Hg19          TCTGTTGATGGTGTCTTTTGTTCTGCAGGAGATGTATAGTTCAATTAG
Read 2        TCTGTTGATGGTGTCTTTTGTTCTGCAGGAGATGTATAGT
Read 2        TCTGTTGATGGTGTCTTTTGTTCTGCAGGAGATGTATAGT
Read 1        CAGCCATTAGGTGTGGGCATTAGTGGCCATCCTTTATTAAATAA ATTGGATGACACAGAAAATGCT
              < hg19 align->
'homology'              GGTGT
```

Alignment of Read1 to HPV

```
HPV           GTAGGTGTTGAGGTAGGTCGTGGTCAGCCATTAGGTGTGGGCATTAGTGGCCATCCTTTATTAAATAA
Read 1              GAGGTAGGTCGTGGTCAGCCATTAGGTGTGGGCATTAGTGGCCATCCTTTATTAAATAA

HPV           ATTGGATGACACAGAAAATGCTAGTGCTTATGCAGCAAATGCAGGTGTGGATAATAGAGAATGTATAT
Read 1        ATTGGATGACACAGAAAATGCTAGTGCTTATGCAGCAAATG
```

Figure 9 **HPV integration evidence based on single fragments: Join is between the two reads**
T3157 Int No 107. Table S6 identifies 6 independent reads. The table also identifies HPV as being spliced in at 109,032,910 (the position indicated by the blue line) whereas the fragment matches hg19 on both sides of this position. The alignment of the two reads to hg19 and the hpv16 sequences (boxed) confirms that Read 2 aligns perfectly with hg19 (blue shading) and Read 1 aligns perfectly with hpv16 (red shading). However, if the two reads are misassembled based on the short homologous sequence GGTGT that is present in both reads (and the orange and yellow shaded sequences discarded) then this gives the insertion point listed in Table S5 suggesting that the insertion point is based on this misassembly.
In doing this the sequence GGTGT is incorrectly identified as the micro-homologous sequence that is the basis of the integration.

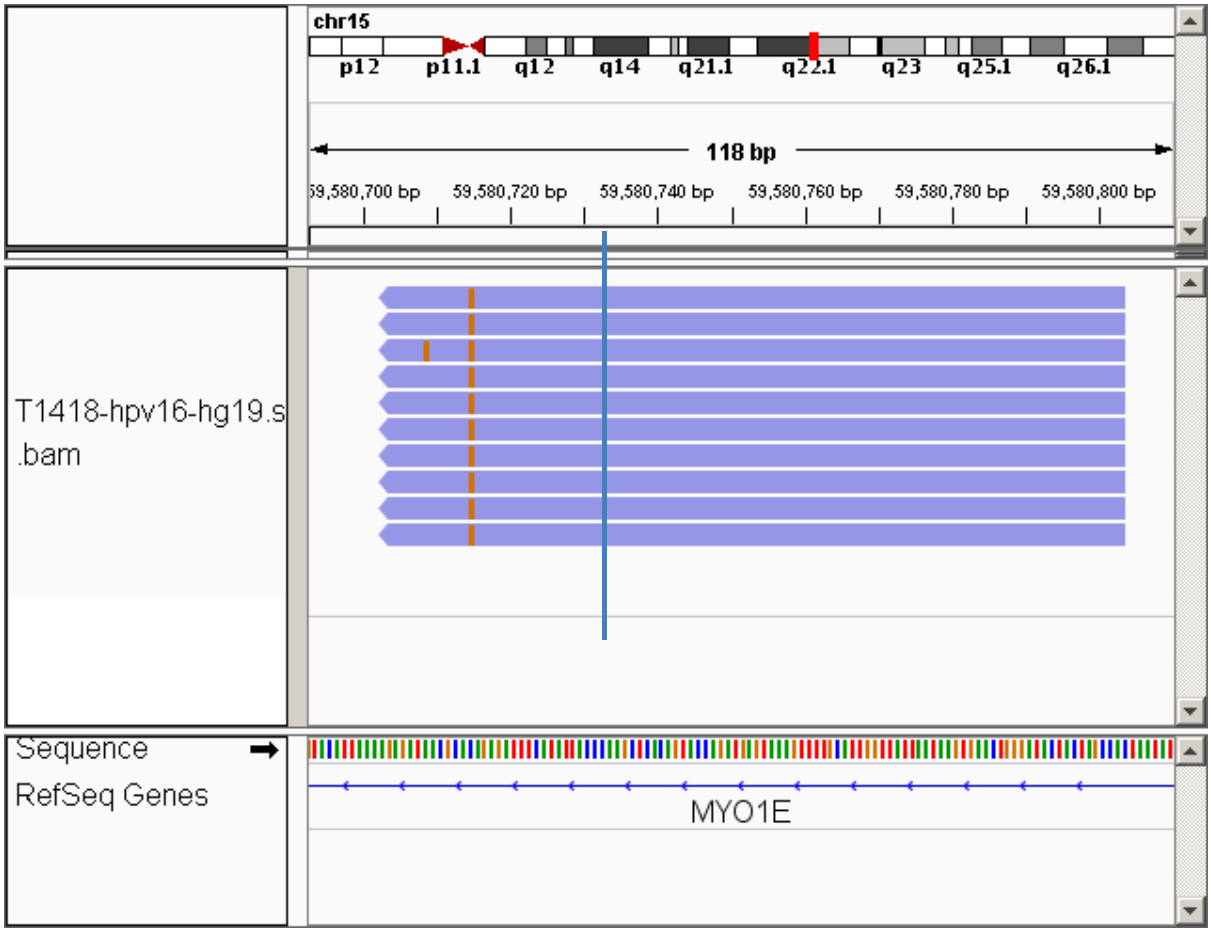
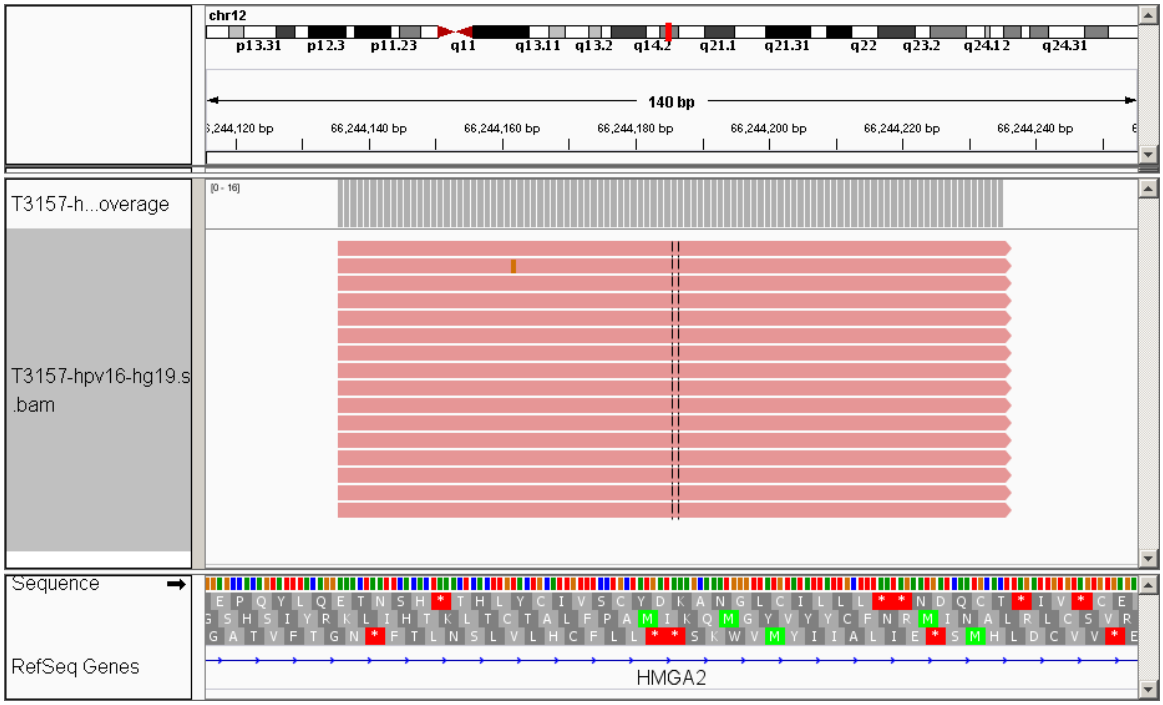


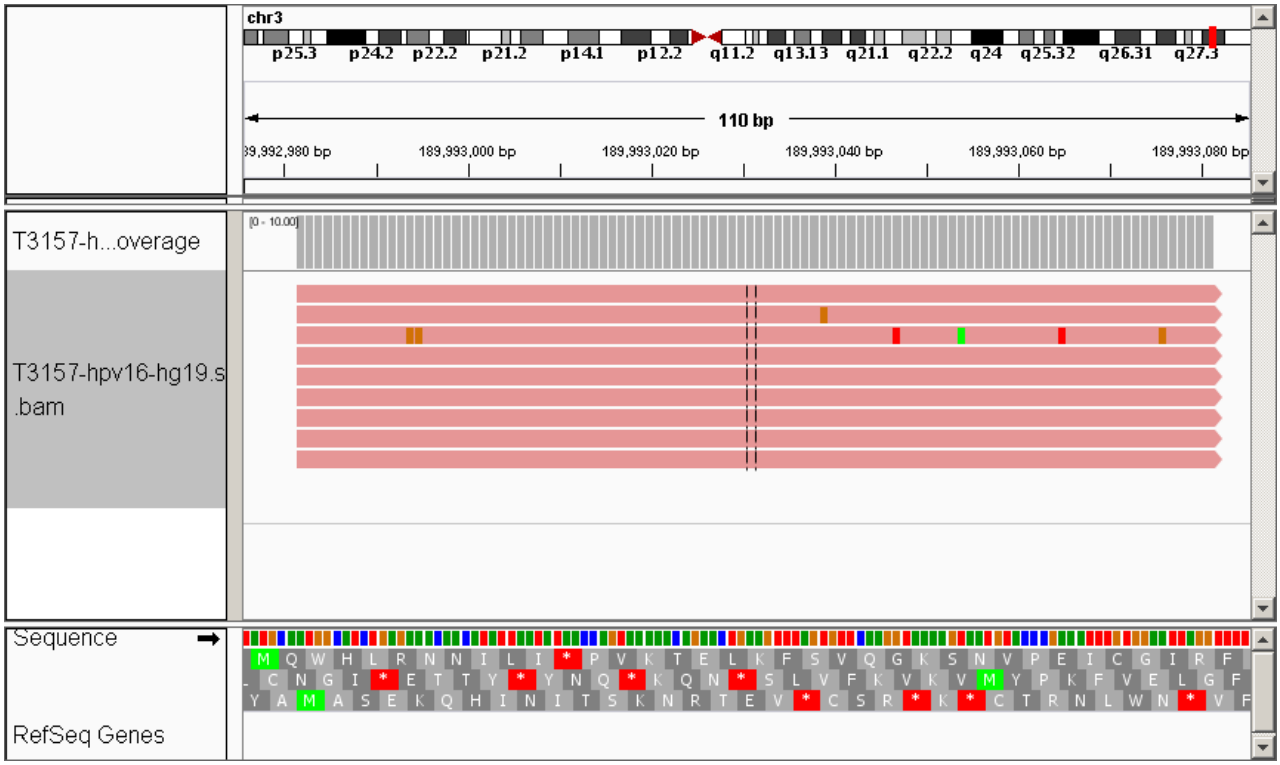
Figure 10 **HPV integration evidence based on single fragments: Join is between the two reads**
T1418 Integration number 10. Table S5 identifies there as being 4 reads associated with this integration. The integration site is identified as being at location chr15:59,580,734 (blue line). Again, the read data contradicts this as being an integration site.




```
Hg19      CAGGAAACTAATTCACACTAAACTCACTTGTACTGCATTGTTTCCTGCTATGATAAAGCAAATGGGTTATGTATATT
Read 1        AAACTAATTCACACTAAACTCACTTGTACTGCATTGTTTCCTGCTATGATAAAGCAAATGGGTTATGTATATT
Read 1        AAACTAATTCACACTAAACTCACTTGTACTGCATTGTTTCCTGCTATGATAAAGCAAATGGGTTATGTATATT
Read 2 reversed                                CTGGTTGCAAATCTAACATATATTCATGCAATGTAGGTG
Read 2 reversed                                CTGGTTGCAAATCTAACATATATTCATGCAATGTAGGTG
Sanger        AAACTAATTCACACTAAACTCACTTGTACTGCATTGTTTCCTGCTATGATAAAGCAAATGGGTTATGTAGGTG
HPV16         TTTAATTGCTCATAACAGTAGAGATCAGTTGTCTCTGGTTGCAAATCTAACATATATTCATGCAATGTAGGTG

Hg19                  ATTGCTTTAATAGAATGATCAATGCACTTAGATTGTGTAGTGTGAGA
Read 1                ATTGCTTTAATAGAATGATCAATGCAC
Read 1                ATTGCTTTAATAGAATGATCAATGCAC
Read 2 reversed       TATCTCCATGCATGATTACAGCTGGGTTTCTCTACGTGTTCTTGATGATCTGCAACAAGAC
Read 2 reversed       TATCTCCATGCATGATTACAGCTGGGTTTCTCTACGTGTTCTTGATGATCTGCAACAAGAC
Sanger                TATCTCCATGCATGATTACAGCTGGGTTTCTCTACGTGTTCTTGATGATCTGCAACAAGAC
HPV16                 TATCTCCATGCATGATTACAGCTGGGTTTCTCTACGTGTTCTTGATGATCTGCAACAAGACATACATCG
```

Figure 11 **HPV insertion evidence based on single fragments: Join is between the two reads**
T3157= SRR1610995 . chr12:66,244,116-66,244,255 Integration no 92.
As in Figure 9 Read 1 maps perfectly to hg19 (blue shading) and read 2 maps perfectly to hpv16 (pink shading) suggesting that the join lies between the two reads. However, the integration site listed in Table S5 can be explained by misaligning the two fragments (boxed section) based on the short homologous region (blue nucleotides) and discarding the two end sections (yellow and orange).
Oddly, the Sanger sequence in Table S6 covering this insertion (that the paper claims was independently obtained) is based on what appears to be a misassembly of the fragments. We have also identified in red the region of the Sanger sequence that they identify as being viral (shown in red here).




```
hg19 (reversed)     TGATTTGTGTAGGATCATACCTCTTTTAATAGGAAAAACTAAGATAAAAACCTAATTCCACAAATTT
Read 2 reversed                                                    CCTAATTCCACAAATTT
Read 2 reversed                                                    CCTAATTCCACAAATTT
Read 1         CAAGCTCCTTCATTACTTCCTATAGTTCCAGGGGCTCCACAATATACAATTATTGCTGATGCAGGTGACTTT
Read 1         CAAGCTCCTTCATTACTTCCTATAGTTCCAGGGGCTCCACAATATACAATTATTGCTGATGCAGGTGACTTT
Sanger                                AGTTCCAGGGTCTCCACAATATACAATTATTGCTGATGCAGGTGACTTT
Hpv16      TGACCAAGCTCCTTCATTAATTCCTATAGTTCCAGGGTCTCCACAATATACAATTATTGCTGATGCAGGTGACTTT

hg19         CGGGTACATTACTTTTACCTTGAACACTAAACTTCAGTTCTGTTTTTACTGGTTATATTAATATGTTGTTTCTCAGATGCCATTGC
Read 2       CGGGTACATTACTTTTACCTTGAACACTAAACTTCAGTTCTGTTTTTACTGGTTATATTAATATGTTGTTTCTCAGATGCCAT
Read 2       CGGGTACATTACTTTTACCTTGAACACTAAACTTCAGTTCTGTTTTTACTGGTTATATTAATATGTTGTTTCTCAGATGCCAT
Read 1       TATTTACATCCTAGTTATTACATGTTAC
Read 1       TATTTACATCCTAGTTATTACATGTTAC
Sanger       TATGTACATTACTTTTACCTTGAACCCTAAACTTCAGTTCTGTTTTTACTGGTTATATTAATATGTTGTTTCTCAGATGCCAT
Hpv16        TATTTACATCCTAGTTATTACATGTTACGAAAACGACGTAAACGTTTACCATATTTTTTTTCAGATGTCTCTTTGGCTGC
```

Figure 12 **HPV integration evidence based on single fragments: Join is between the two reads**
T3157 chr3:189993061 Integration site 239. chr3:189,992,976-189,993,085
In this case read 2 maps perfectly to hg19 (blue shading) and read 1 maps perfectly to hpv16 (pink shading) suggesting that the join lies between the two reads. Again, the integration site listed in Table S5 can be explained by misaligning the two fragments (boxed section) based on the short homologous region (blue TACAT nucleotides) and discarding the two end sections.
Also again, the Sanger sequence covering this insertion that the paper claims was independently obtained is consistent with what appears to be a misassembly of the fragments.

Figure 13 **Additional inserts associated with a validated insert**
T7413 Integration Nos 164 to 169. Blue arrows indicate the six integration sites identified in Table S5. The (leftmost) integration at chr6:90,350,635 is fully supported by the data. There are no hybrid reads consistent with the integration sites identified at the other locations, and there are no paired end reads that could not be explained by the integration at chr6:90,350,635 and therefore would require the existence of these additional integrations.

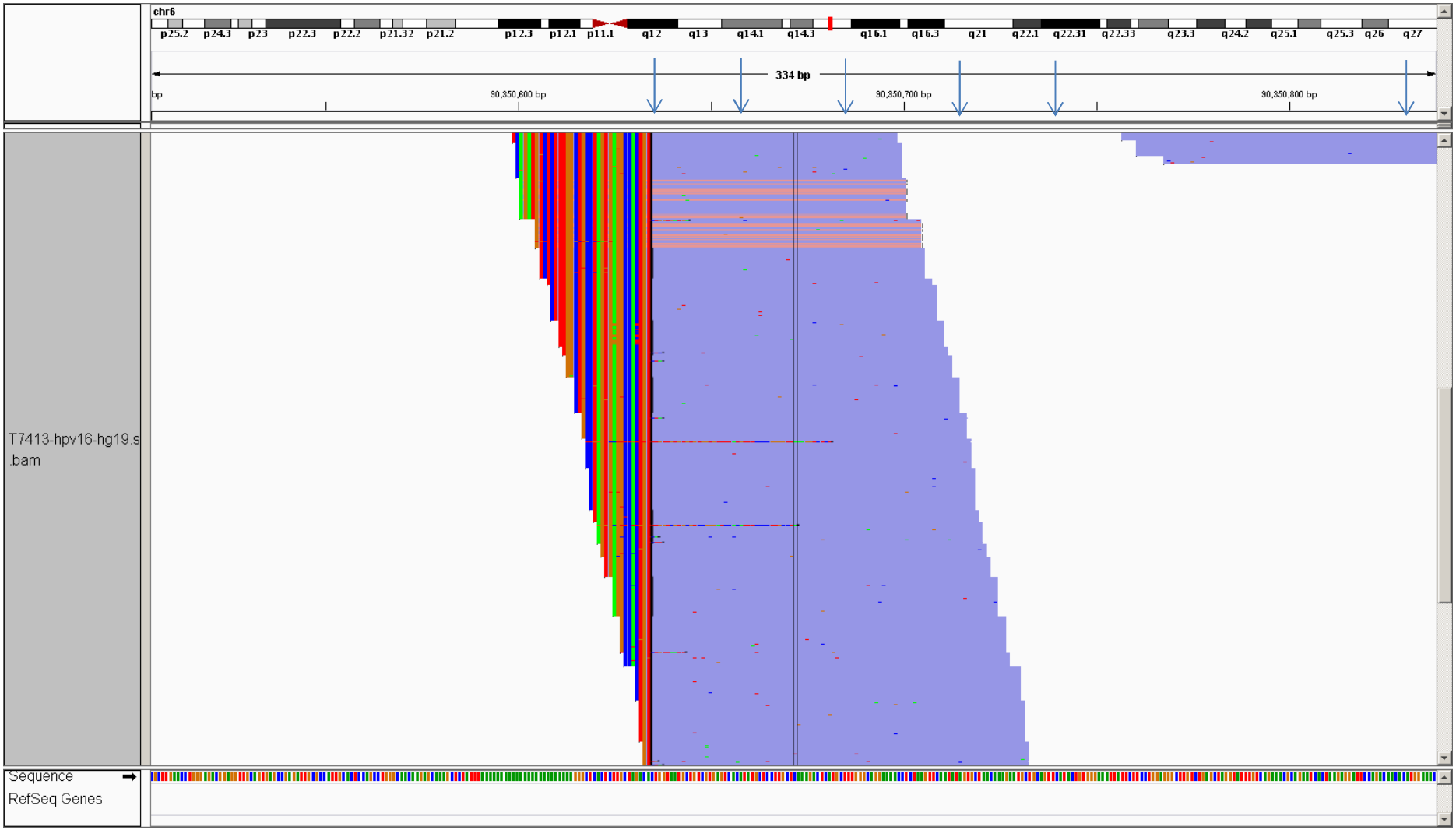


Figure 13 Cont.

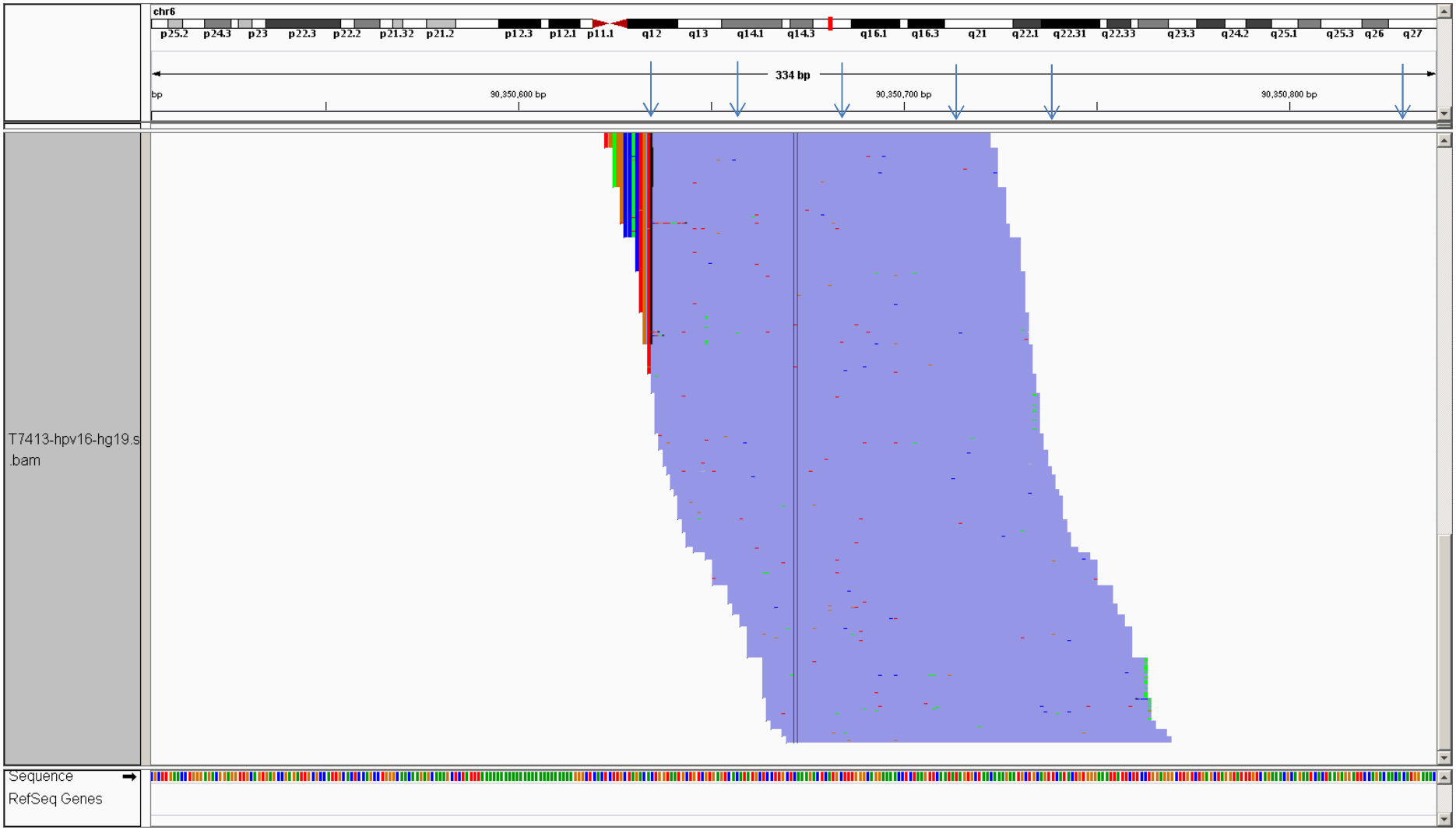


Figure 13 Cont.

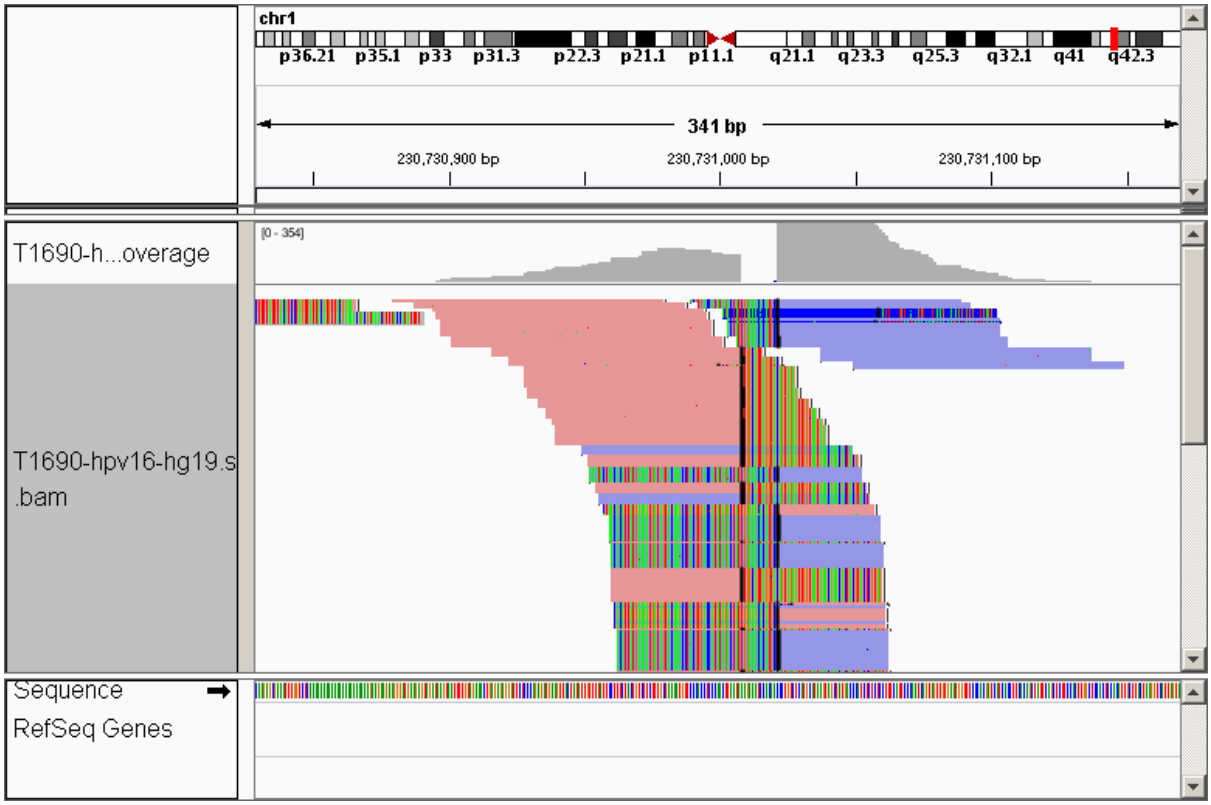


Figure 14 **integrations where our algorithm confirms the integration**

T1690 = SRR1611116. integration No 1. An integration is identified at chr1:230731022 in Table S5. The reads provide confirmation of the integration. 31 independent reads are claimed for the transition to HPV and 52 for the transition back to hg19. The new analysis indicates that there are probably only 13 truly independent fragments.

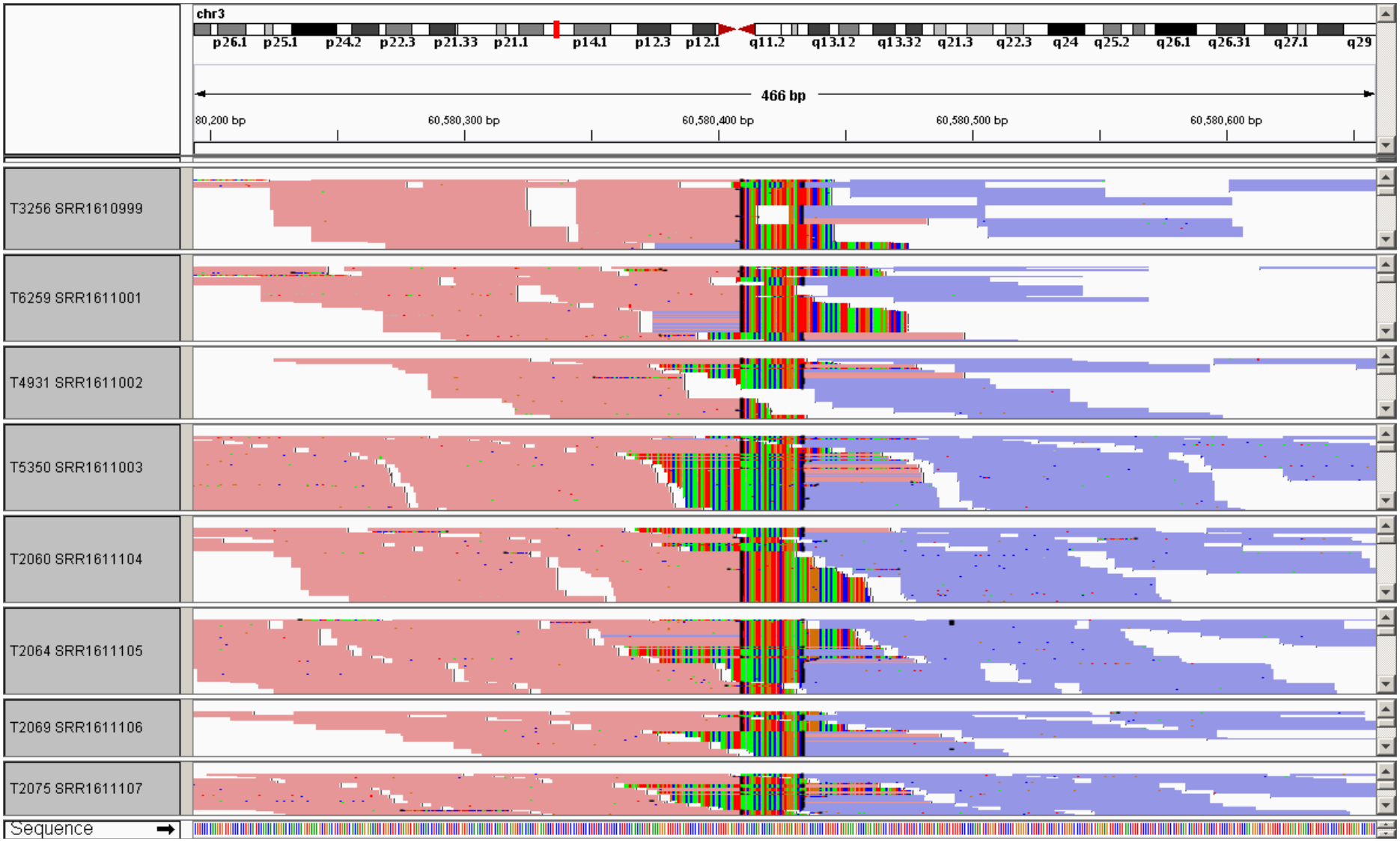
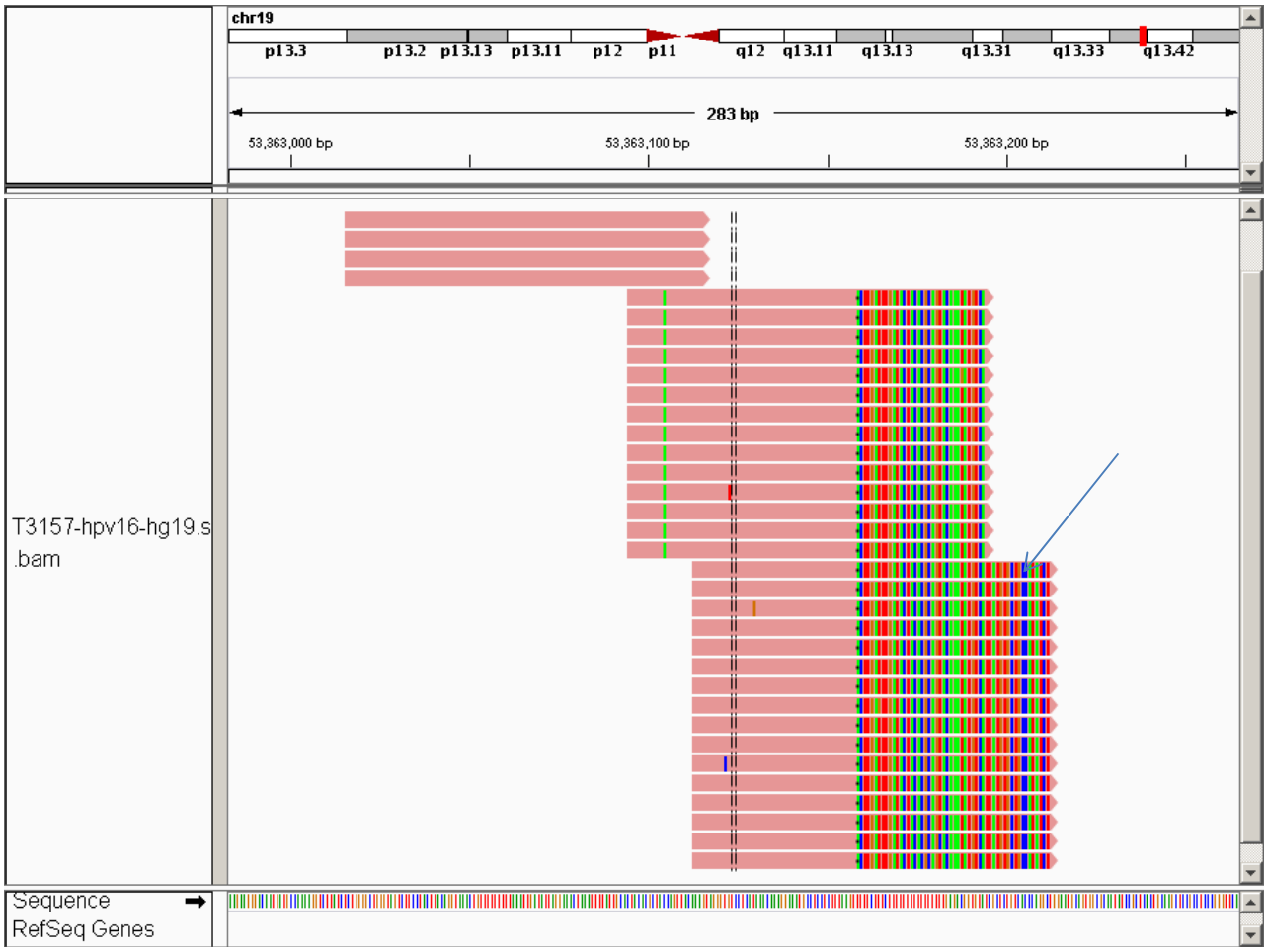


Figure 15 **Evidence of cross contamination**

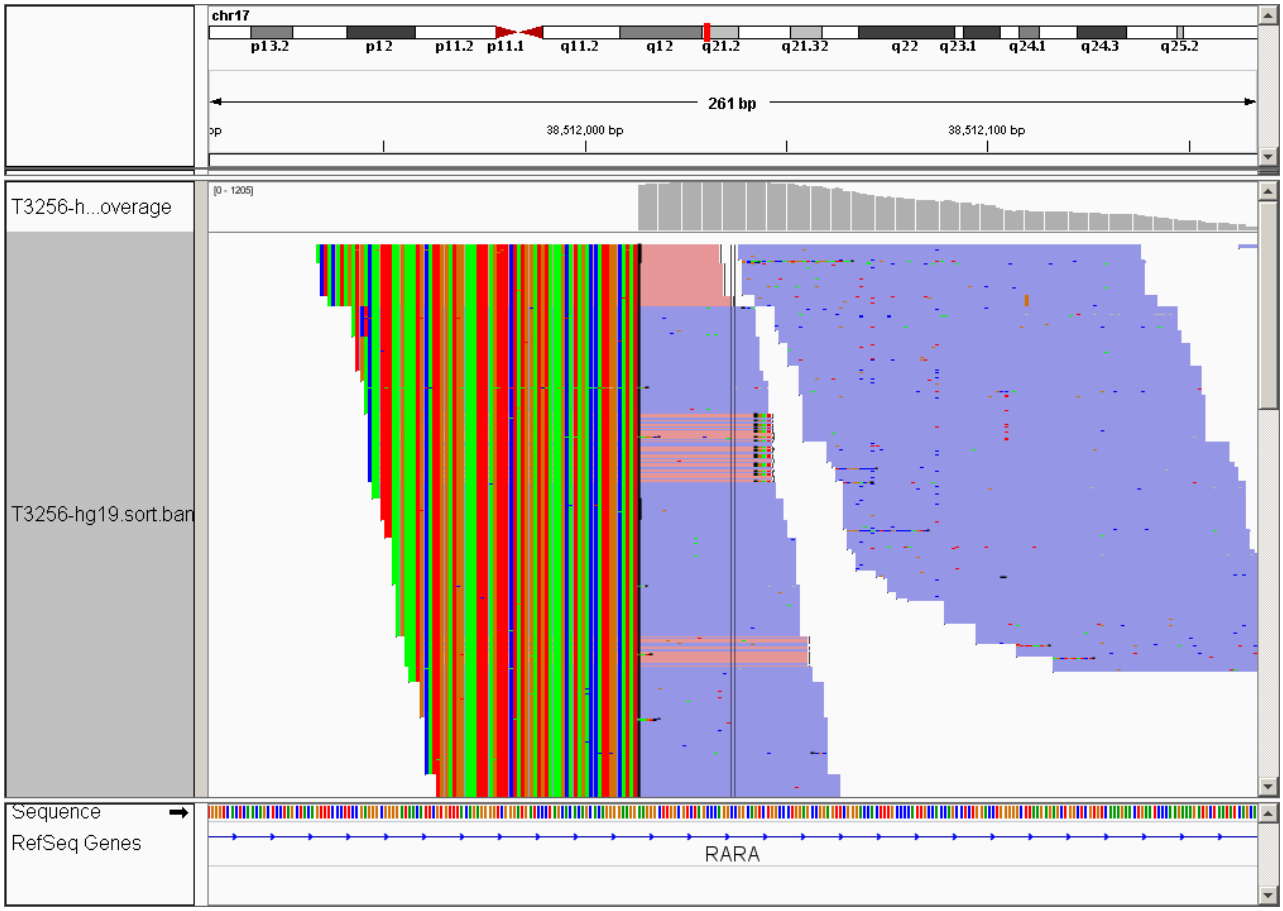
chr3:60,580,194-60,580,659 Eight samples in two groups of sequential submission identities show evidence for exactly the same HPV integrations. Six of the eight (ie all samples other than T4931=SRR1611002 and T2069=SRR1611106) are reported as independent integrations in Table S5. Other samples show no sign of integrations at this location.




```
hg19          GCTACCAGTTACAAAGTAGGGTGTCCTCATAAACAAAGGAAATGCCTCCTCTTTAAGTTTTTCTTTCTTTTTT
Read 1                  AAAGTAGGGTGTCCTCATAAACAAAGGAAATGCCTCCTCTTTAAGTACTTGATTTGATACT
Hpv16         TACAACGAGCACAGGGCCACAATAATGGCATTTGTTGGGGTAACCAACTATTTGTTACTGTTGTTGATACT

Hg19          TTTTTTTTTTGAGACGGTGTATCGCTCTGTCACCCAGGCTGGAA
Read 1        ACACGCAGTACAAATATGTCATTATGTGCTGCCATATCT
Hpv16         ACACGCAGTACAAATATGTCATTATGTGCTGCCATATCTACTTCAGAAACTACATATAAAAATACTAACTTTAA
```

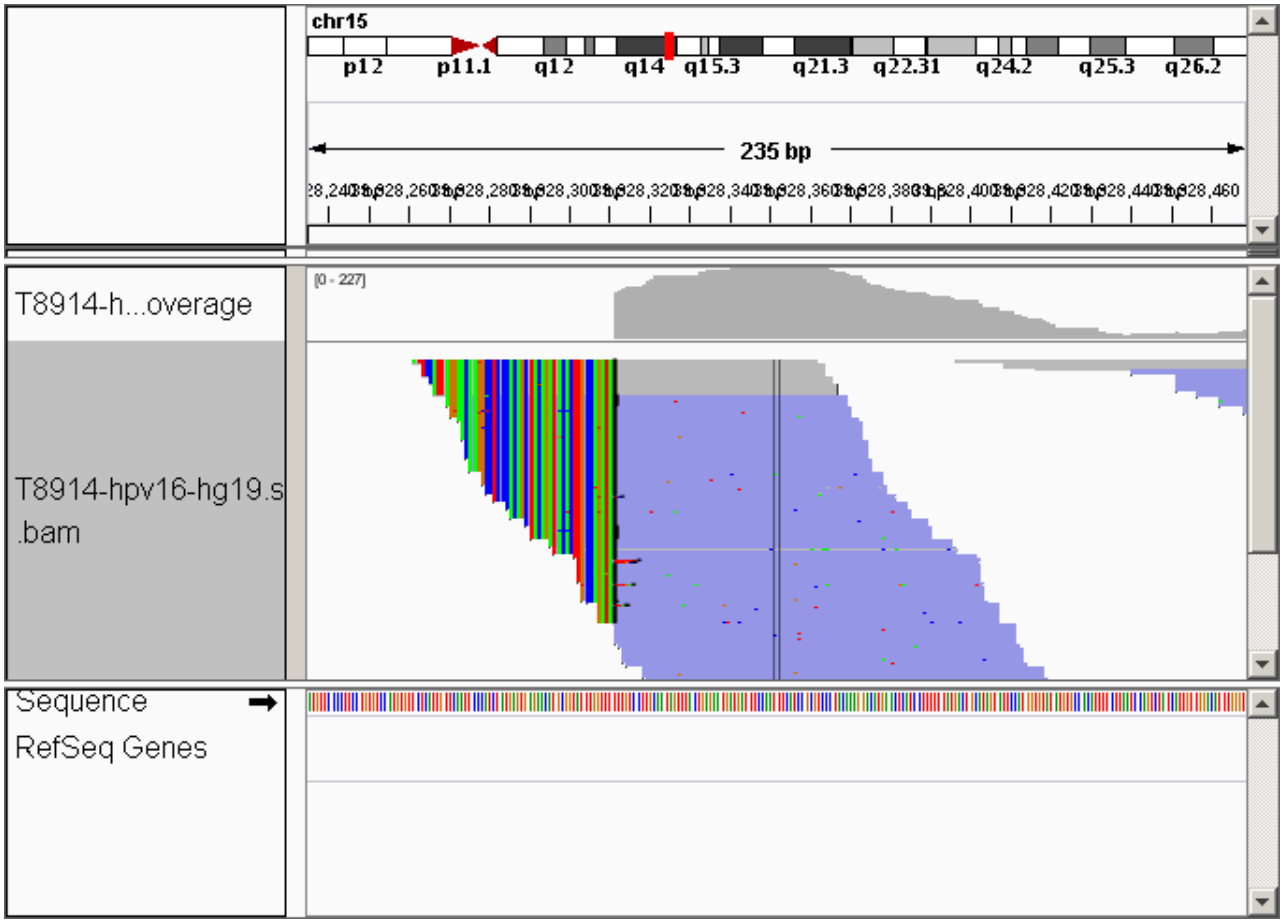
Figure 16 **An example HPV integration with additional sequence inserted**
SRR1610995 T3157. Insert No 177. This integration is confirmed by reads from three independent fragments The lower portion takes read SRR1610995.1577879 (arrowed) and shows how it is part hg19 (shaded blue) and part hpv16 (shaded green) with the sequence ACTTGAT inserted in between.




```
Hg19        CTCCTAGCTACAGTTTCCCTTCCGAGGGCGGGGATAACATTCGTGTTTACAGAGGGGTCGGGATGATCCCTAGC
Read 1                    ACTACATAGATGACAATTTAAGAAATGCATTGGATGGAAATTTAGTTTCTATGGATGTAAAG
              CCCTGTTGGAACTACATAGATGACAATTTAAGAAATGCATTGGATGGAAATTTAGTTTCTATGGATGTAAAG

hg19        ACACAGCACAGGGGAAGGAAGGGCTTGGCGTCTAGCCCAGGCCGGCAGTCTGGCCCT
Read 1      CATAGACCATTGGCATATAAGGGCTTGGCGTCTAGCCC
Hpv16       CATAGACCATTGGTACAACTAAAATGCCCTCCATTATTAATTACATCTAACATTAATGCTGGTACAGATTCTAGGT
```

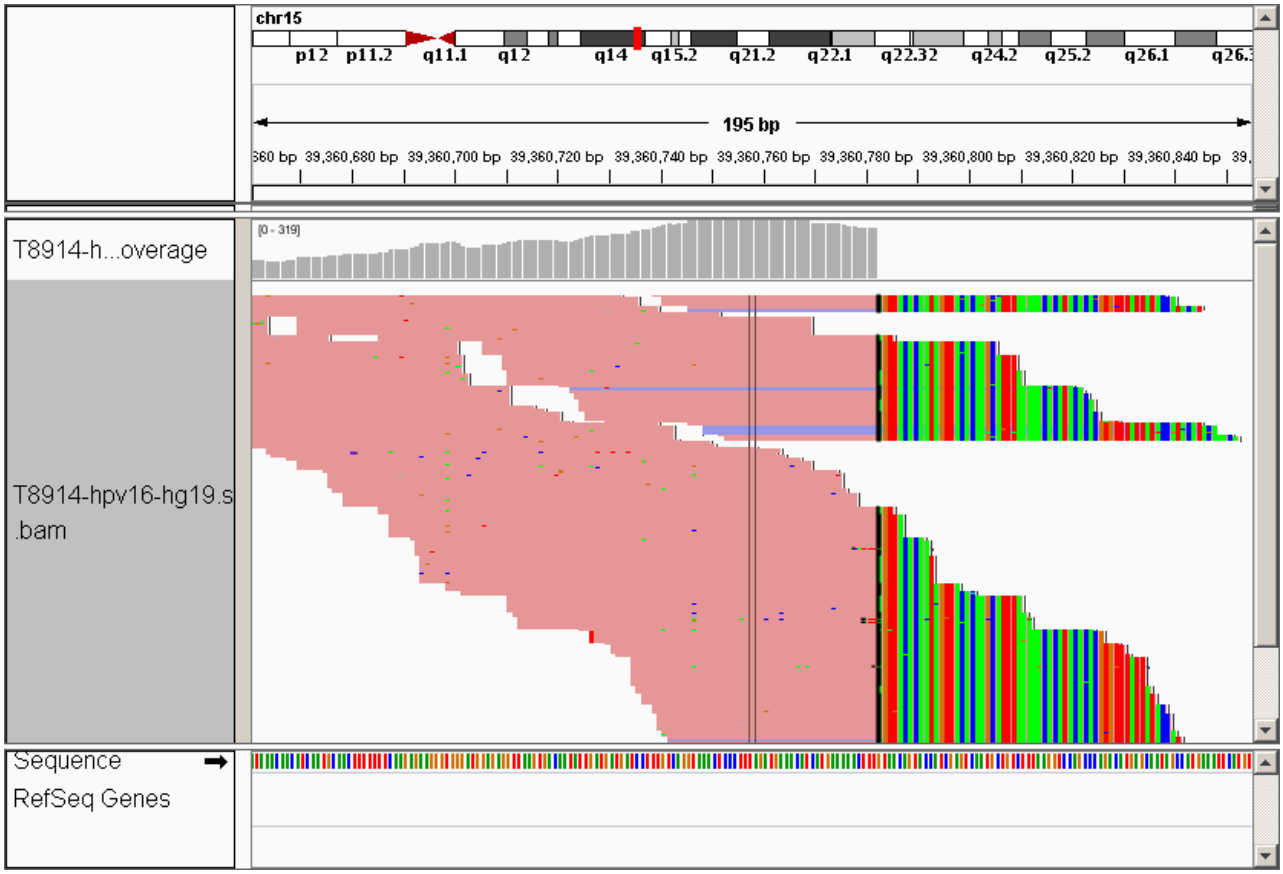
Figure 17 **An example HPV integration with additional sequence inserted**
SRR1610999 T3256 Insert No 15. The lower section shows an analysis of read SRR1610999.422614. Colour key as Figure 16




```
Hg19        TGTGTTCTCATTGTTCAATTCCCATCTATGAGTGAGAACATGCGGTATTTGGTTTTTTGTCC
Read              ATTCCATTAGGAACAAGGCCTCCCACAGCTACAGATACACTTGCCAGATATTTGTCC
Hpv16       GGTATATTCCATTGGGAACAAGGCCTCCCACAGCTACAGATACACTTGCTCCTGTAAGACCC

Hg19        TTGCAATAGTTTACTGAGAATGATGATTTCCAATTTCATCCGCGTCCCT
Read        TTGCAATAGTTTACTGAGAATGATGATTTCCAATTTCATCCGC
Hpv16       CCTTTAACAGTAGATCCTGTGGGCCCTTCTGATCCTTCTATAGTTTCTT
```

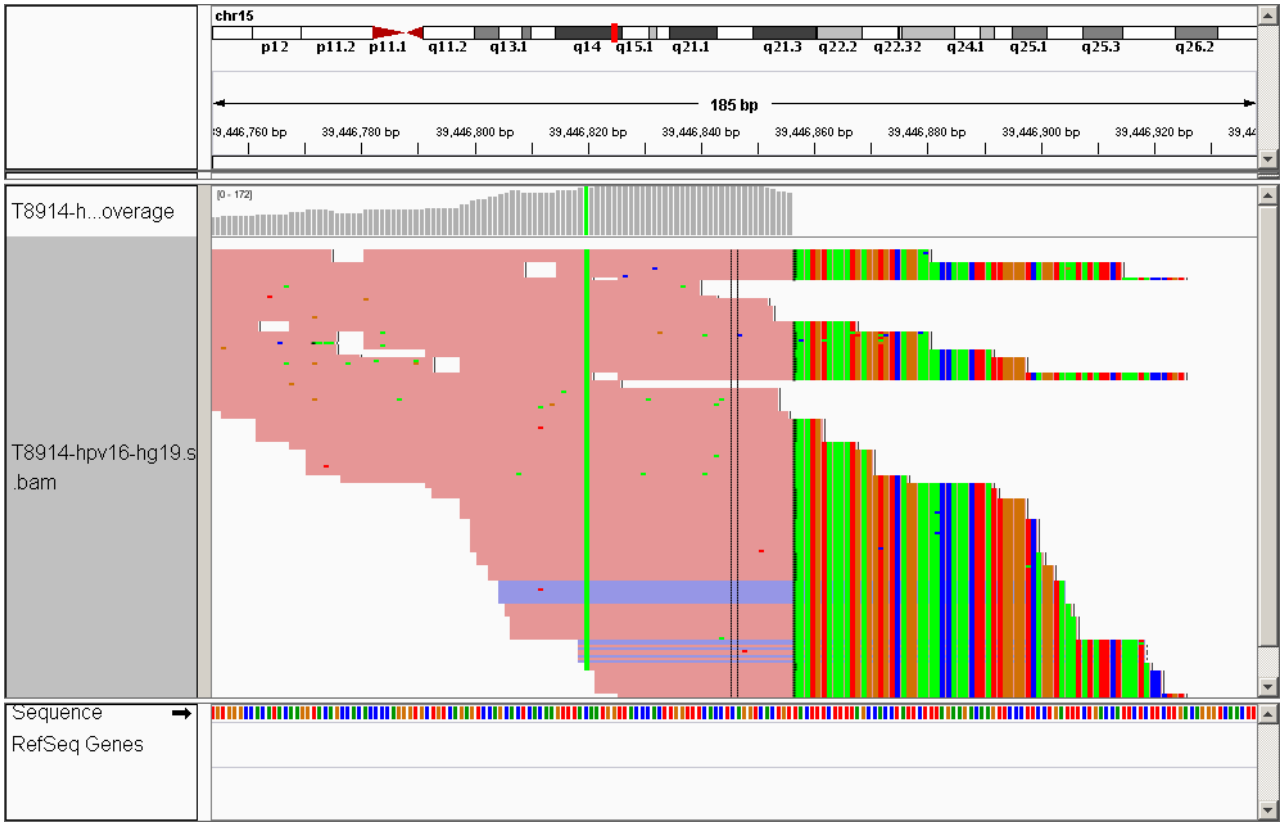
Figure 18 **An example HPV integration with additional sequence inserted**
SRR1611024 T8914 Insert No 1. Colour key as Figure 16




```
Hg19                    TGATCCTTTGTCAGGTAAACTCCCTTTAGATGAAACATCATGAAAAACTTTGTAACAGTACCATCG
Read                          TTGTCAGGTAAACTCCCTTTAGATGAAACATCATGAAAAACTTTAGTTACACAATAGTT
Hpv16 (rev)       ATAACCACAACACAATTAGTAGGTGTTGAAACAATAAGTTTATTTATGTTGCATGACACAATAGTT

Hg19              AGTCAATGCTGTCTTTAAGGTCCAGCCCAGACATGAGCAATAGCCATGAAATT
Read              ACACAAGCCTTTAAAAACACATACACACGTGTTTATTATAC
Hpv16 (rev)       ACACAAGCATTTAAAAACACATACACACGTGTTTATTATACCATACATACAAA
```

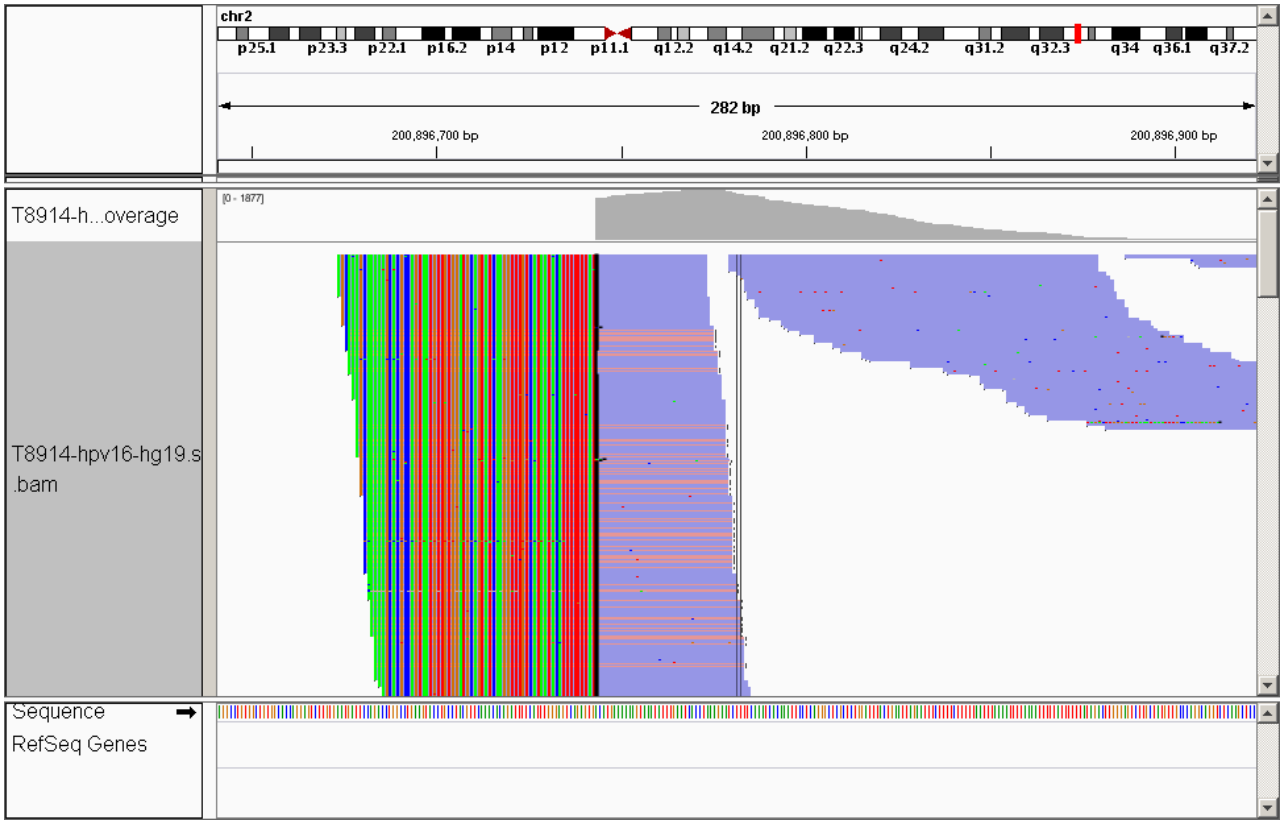
Figure 19 **An example HPV integration with additional sequence inserted**
SRR1611024 T8914 Insert No 9. Analysis of read SRR1611024.994064. Colour key as Figure 16




```
Hg19                    CTCACAAGTTTACAATGGTTACCCACTTGGTTTCACCTCAGGTCACTGTTACTTTCTT
Read                          GTTTAAAATGGTTACCCACTTGGTTTCACCTCAGGTCACTGTAAATGTAAA
Hpv 16                  CGACCTACCTCAACACCTACACAGGCCCAAACCAGCCGCTGTGTATCTGGATTATAAA

Hg19                TTAGCCACCTATTCTTTAGAAGACAAGTTCCCTTCCTGATTTCTGACCATTCTCTTCTTG
Read                ATGAGGTGTCAGGAAAACCAAACTTATTGGGGTCAGGTAAATATATTCT
Hpv16               ATGAGGTGTCAGGAAAACCAAACTTATTGGGGTCAGGTAAATGTATTCTAAATACCCTGT
```

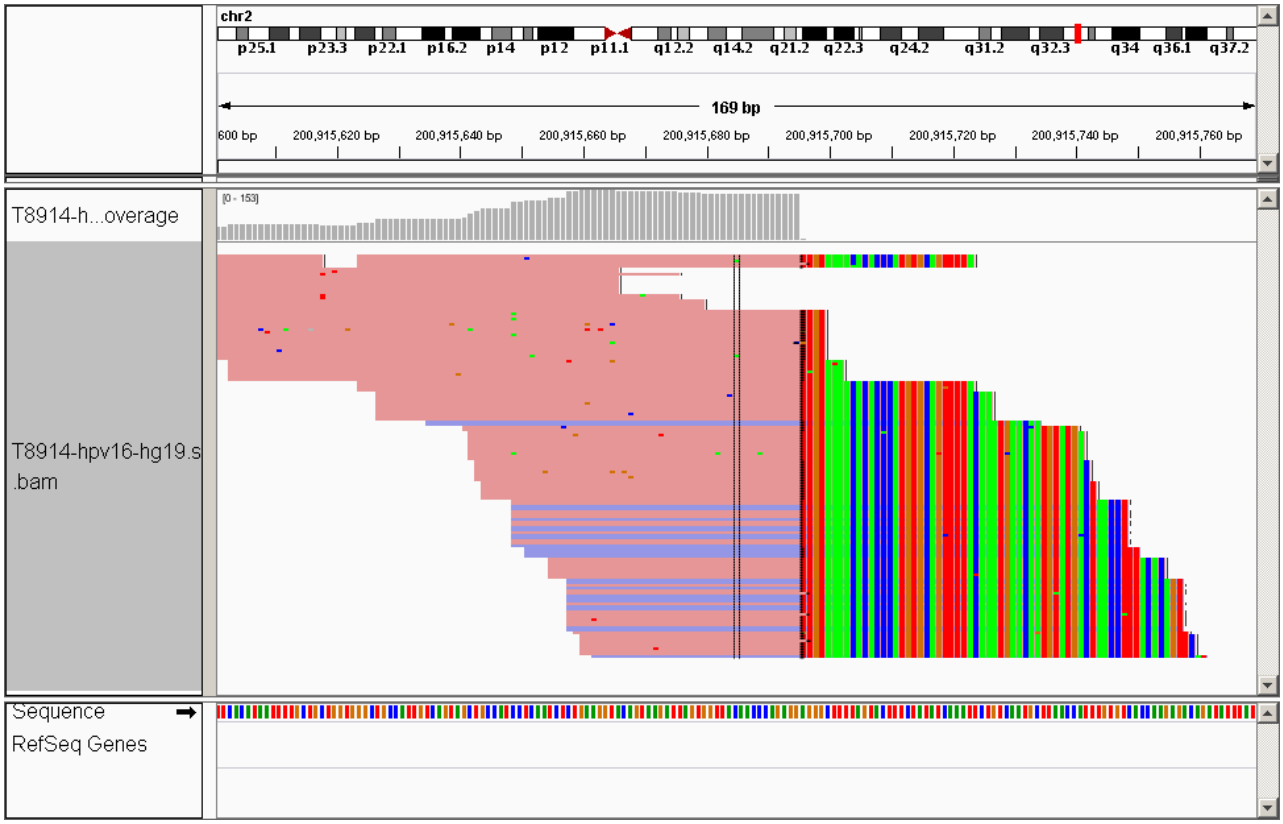
Figure 20 **An example HPV integration with additional sequence inserted**
SRR1611024=T8914, Insert No 9. Analysis of read SRR1611024.2619489. Colour key as Figure 16




```
Hg19          CCAGGAGTCTTGTGAATGATTCCAGACCTGGCCTGAAGTACCTGCCTACGTTCAAACACAGATT
Read                        AGCAAAGCAAAAAGCACGCCAGTAATGTTGTGGATGCAGTATCAAGGTTT
Hpv16         CAAAAGCACACAAAAGCAAAGCAAAAAGCACGCCAGTAATGTTGTGGATGCAGTATCAAGATTT

Hg19          TTCATTTGGGAGTCAGAGTGACATGAAAAAAGTAAATATTTGAATCACCTAAAGAATC
Read          GTCATAATACATTTTTTTTATACATGAAAAAAGTAAATATTTGAATCACCT
Hpv16         GTCATATAGACATAAATCCAGTAGACACTGTAATAGTTTTTGGTATTTTAACTTGAGACAA
```

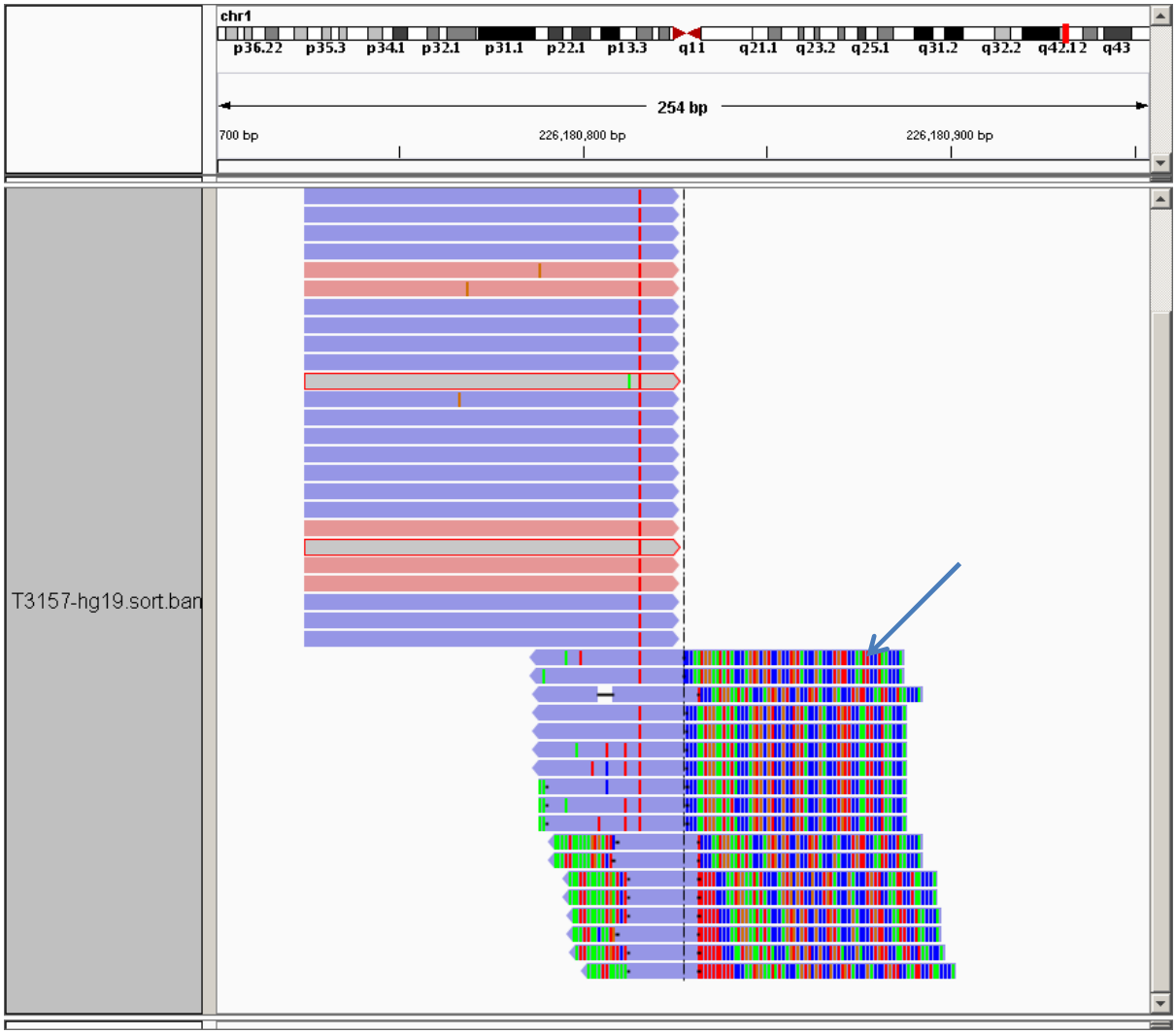
Figure 21 **An example HPV integration with additional sequence inserted**
Insert No 15 Analysis of read SRR1611024.13051199. Colour key as Figure 16




```
Hg19          TCCTCAATCTCTGAAGGTACAGTAAGATTAGTGGTTACAACCCAGAGGAGGGCTTTT
Read                    CTGAAGGTACAGTAAGATTAGTGGTTACAACCCAGAGGTTGTAAAAC
Hpv16           TTAGTATTATTATATAAGTTGCTTGTAAATGTGTAACCCAAAATCGGTTTGCAC

Hg19        AGTACTTTCTATATCAAATTCTGTCAATGCTTAACCATTTGTGTCACCAAGAGACAGAT
Read        ACACCCATGTGCAGTTTTACAAATGAACAATGTATGACTAACCTTTACACAGT
Hpv16       ACACCCATGTGCAGTTTTACAAATGAACAATGTATGACTAACCTTTACACAGTTCATGTAT
```

Figure 22 **An example HPV integration with additional sequence inserted**
SRR1611024=T8914, Insert No 11. Analysis of read SRR1611024.12139823. Colour key as Figure 16




```
Chr1          TGTATCACAAGATTAACAATGATCTCTACTTTCTTTCTTTTTTTTTTTTTTTTAAGAAAAGATGAACAT
Read 1              AAGATTAAAAATTATCTCTACTTTCTTTTTTTTTTTTTTCCCAATGGAATATACCCAGT
HPV 4490-4390             AGCAAGTGTATCTGTAGCTGTGGGAGGCCTTGTTCCCAATGGAATATACCCAGT
                    <-----Matches hg19 -------------------->
                                                    <---- Matches HPV--------
Homologous sequence                                 TTTTT


Read 1        GCGTCCGCCTGTACCCGACCCTGTTCCAATTCCTAACCCA
HPV cont      GCGTCCGCCTGTACCCGACCCTGTTCCAATTCCTAACCCACCAAAAAATACACCCATACTTCCATA
              <---- Matches HPV---------------------->
```

Figure 23 **Table S5 integrations not identified in our analysis**
T3157 Insert No 28. The hybrid reads are all copies of the same fragment that have been aligned at different locations because of misreads associated with the polyT region. They are consistent with being multiple copies of a single fragment where there was a ligation between HPV and the human genome.

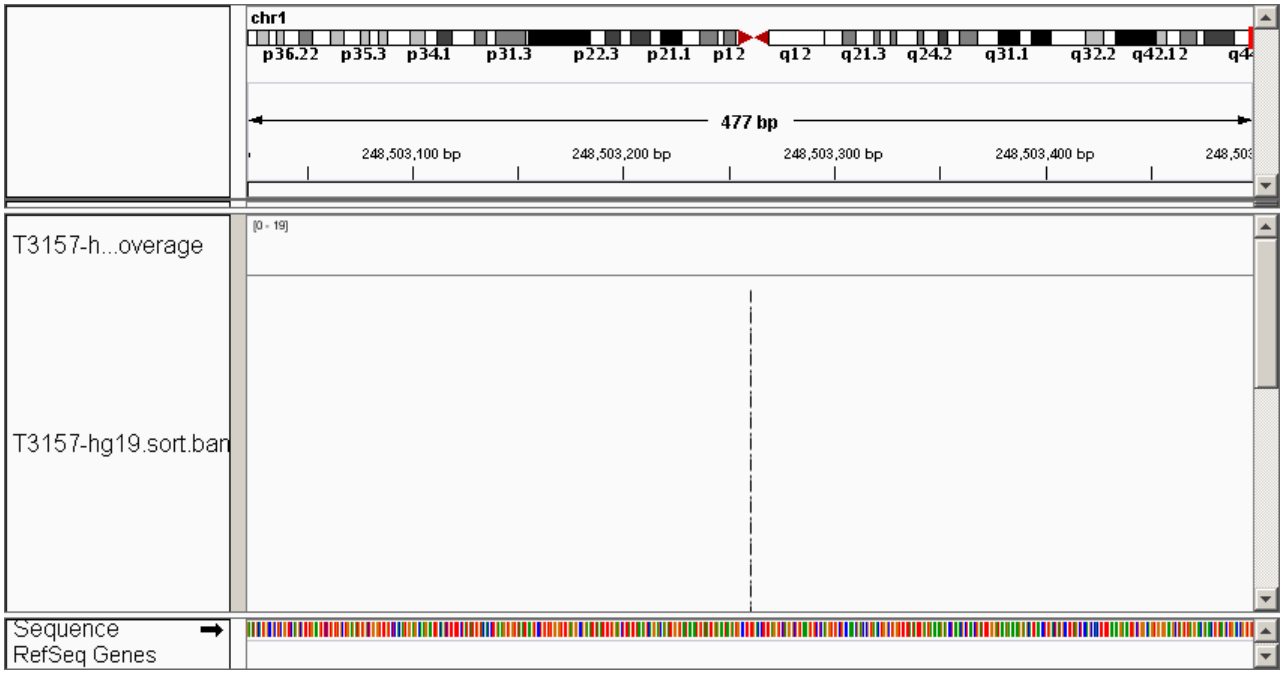


Figure 24 **Table S5 integrations not identified in our analysis**
T3157 Insert No 32 There are no fragments that map to this position with the bwa aligner we used. We suspect that slight differences in the mapping algorithm mean that the fragments have been mapped to another location in the genome. If there were multiple independent fragments then it is likely that one of them would have mapped here so this suggests that the fragments are not associated with a genuine HPV integration

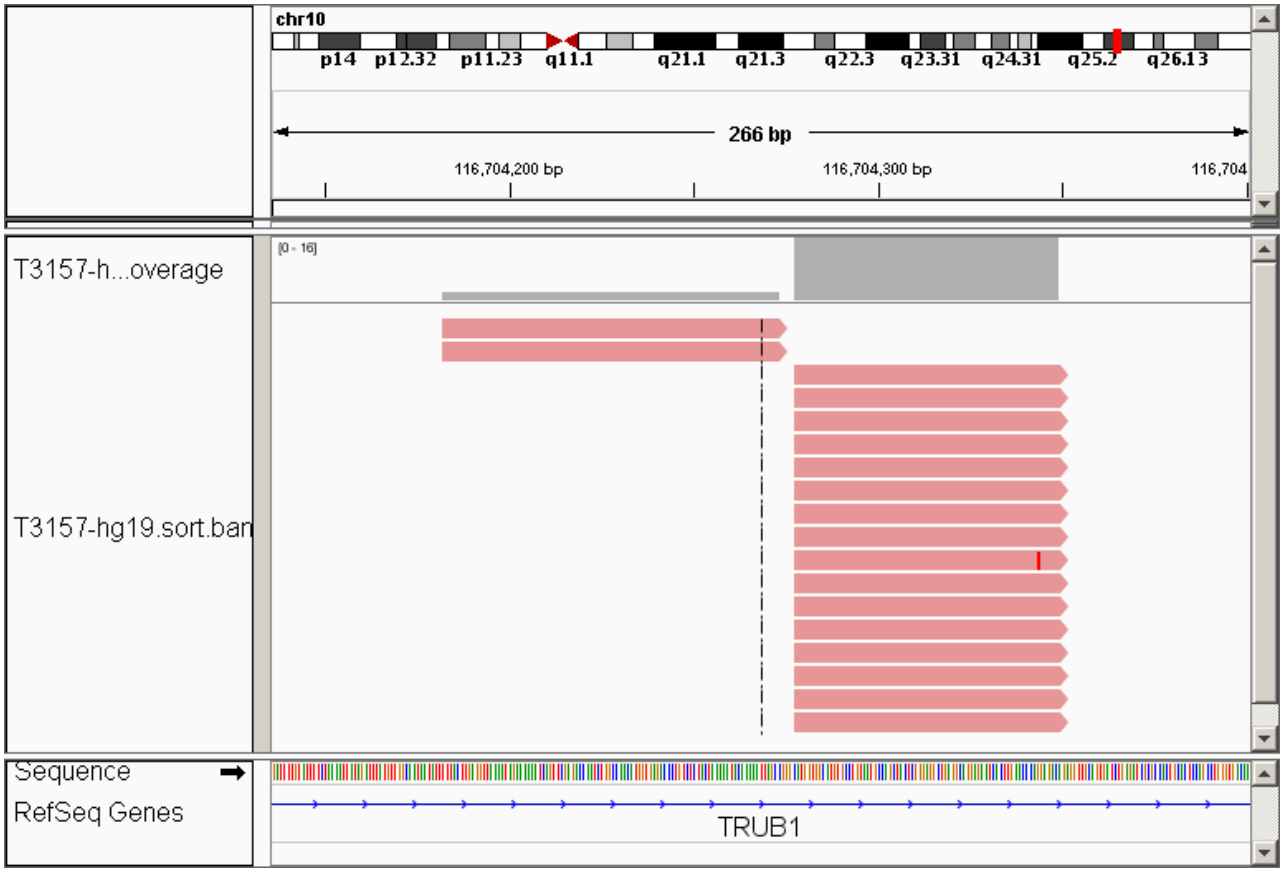


Figure 25 **Table S5 integrations not identified in our analysis**
T3157 Insert 60. There are 16 copies of a fragment that aligns slightly to the left of the integration site listed in Table S5. Our adaptor trimming identified the two paired reads as fully overlapping each other. After removal of the 28bp adaptor sequence in each fragment this leaves the 72bp sequence shown in the figure. This sequence fully aligns to the human genome so provides no evidence for an insertion.

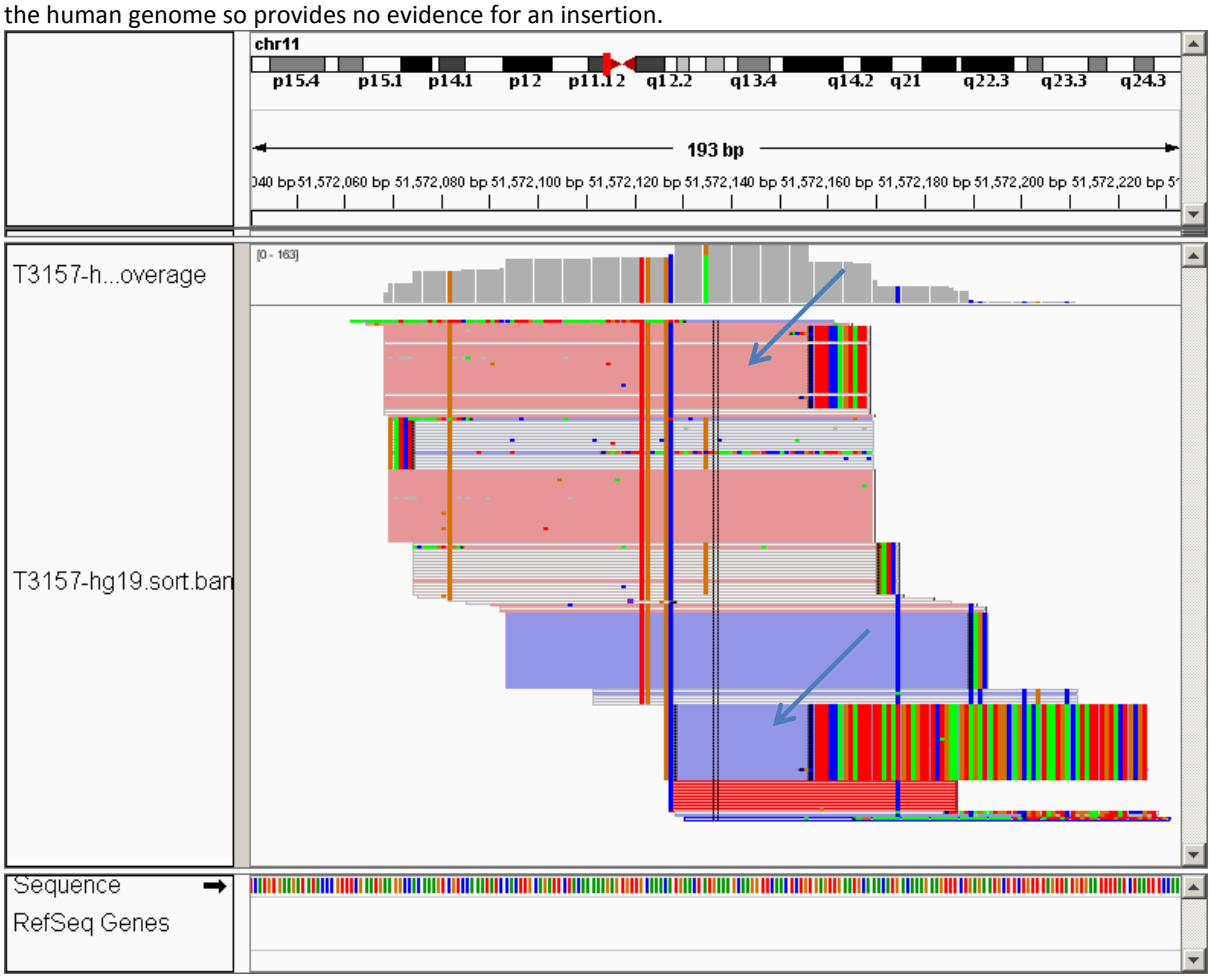


Figure 26 **Table S5 integrations not identified in our analysis**
T3157 Insert No 73. While there are a number of different blocks of reads at this location, there are only two blocks (blue arrows) which are consistent with the integration site listed in Table S5. These are forward and reverse reads associated with multiple copies of a single original fragment. The fragment is a hybrid fragment containing both human and HPV genomic sequence, but as a single fragment should be rejected as almost certainly being an artefact.

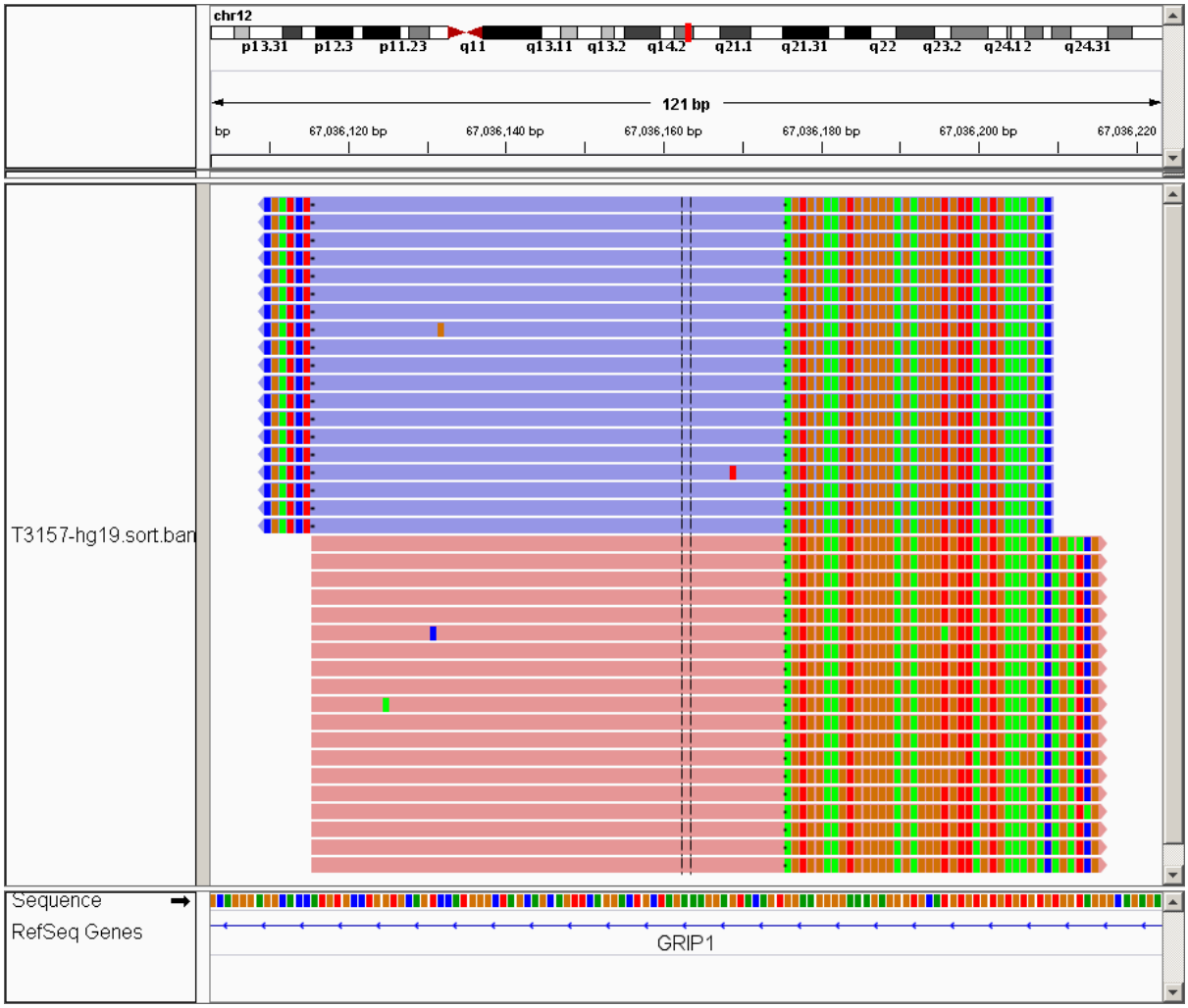


Figure 27 **Table S5 integrations not identified in our analysis**

T3157 Insert No 93 The reads are forward and reverse reads of multiple copies of the same original fragment which is a hybrid of HPV and the human genome. Our insertion detection software did not identify this insertion because of the presence of the additional sequence on the left which also does not match the human genome sequence.

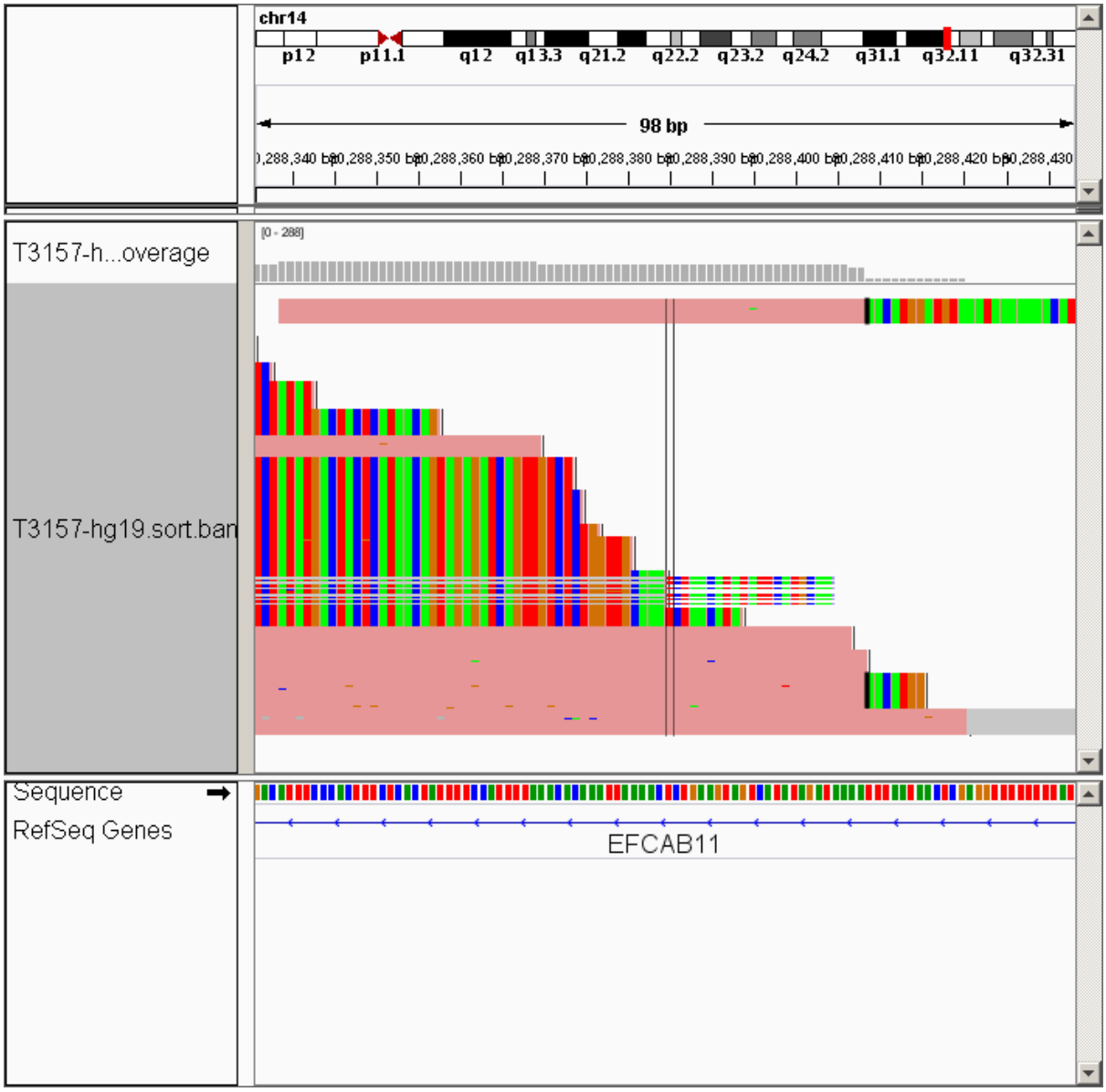


Figure 28 **Table S5 integrations not identified in our analysis**
T3157 Insert 122. Table S5 identifies two inserts very close together at chr14:90,288,385 and chr14:90,288,408. Our analysis confirms that chr14:90,288,408 is supported by two independent reads, but finds no support for an additional insert at chr14:90,288,385 (black vertical line).

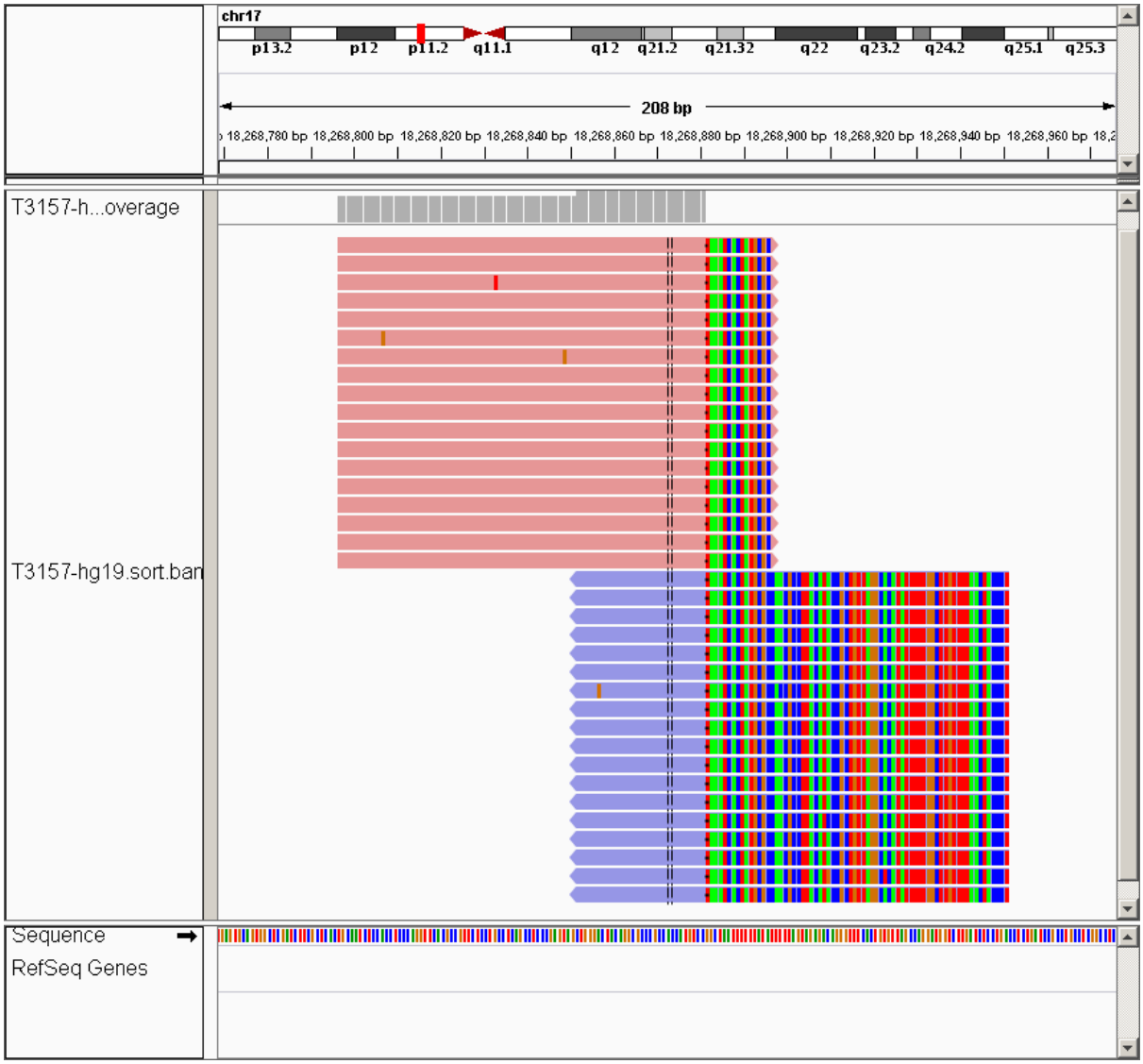


Figure 29 **Table S5 integrations not identified in our analysis**

T3157 Insert 150. chr17:18268881 This integration was not identified by our algorithm, but is associated with multiple copies of a single fragment so does not represent an independently verified integration.

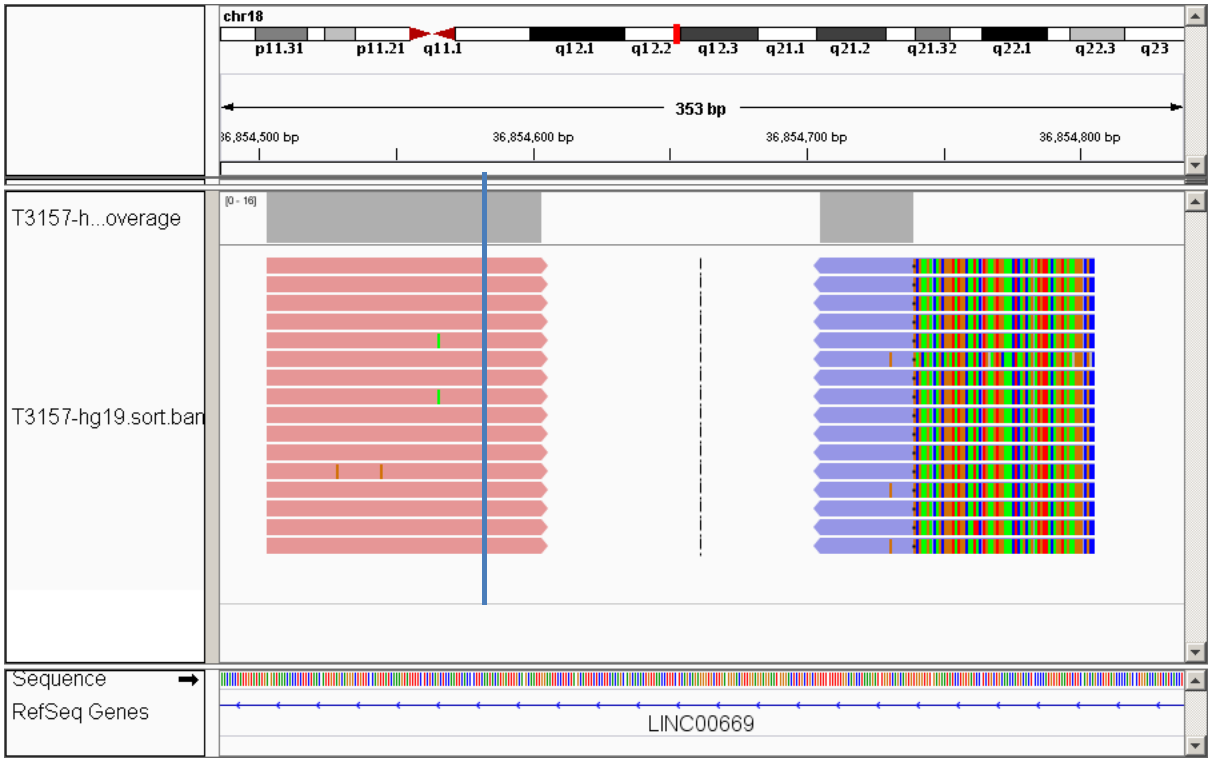


Figure 30 **Table S5 integrations not identified in our analysis**
T3157 Insert No 164. There are multiple copies of the reads from either end of a long paired end fragment. Table S5 incorrectly identifies the insertion at chr18:36,854,576 (vertical blue line), whereas we correctly identify the integration approximately 300bp further on. Neither site represents an integration corroborated by independent reads.

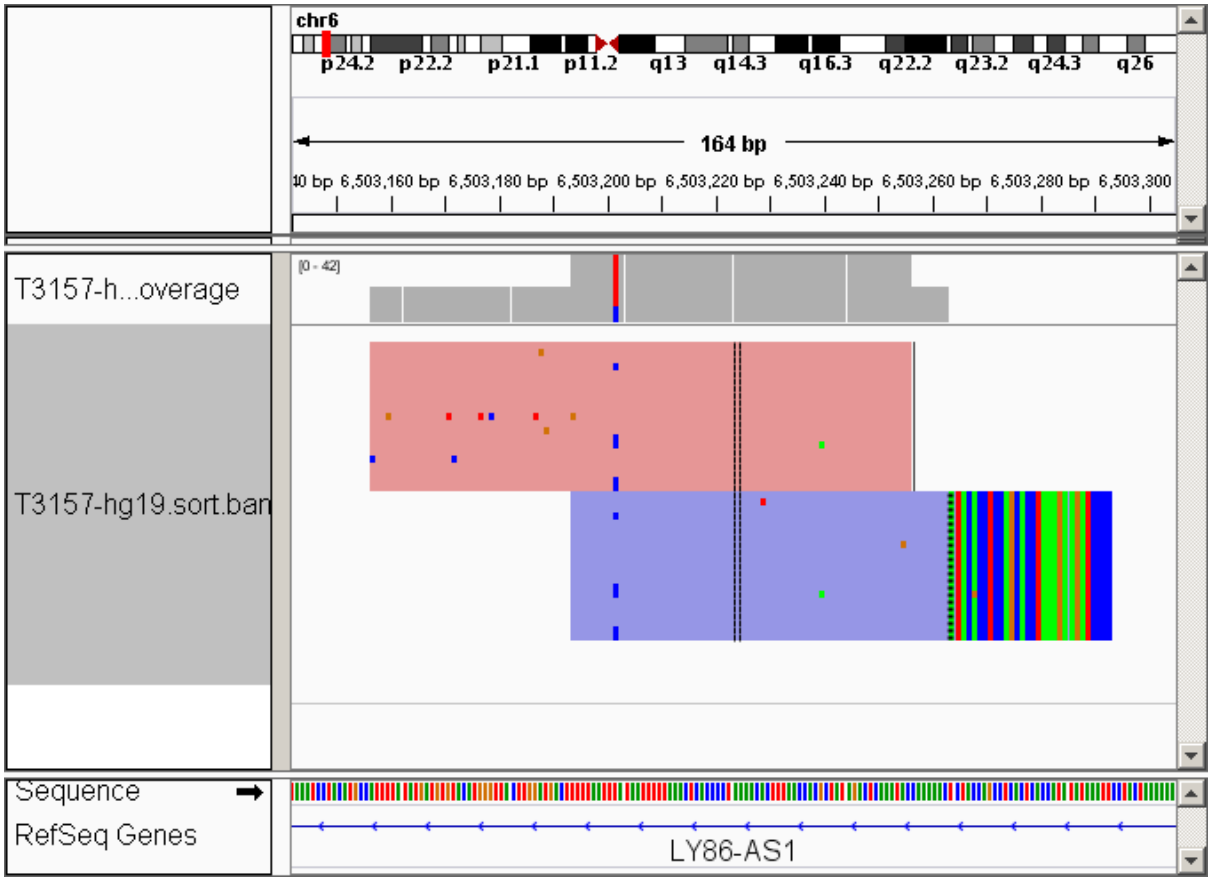


Figure 31 **Table S5 integrations not identified in our analysis**
T3157 Insert No 286. This insert was not identified by our algorithm, but is associated with a multiple copies of a single fragment so does not represent an independently verified integration.

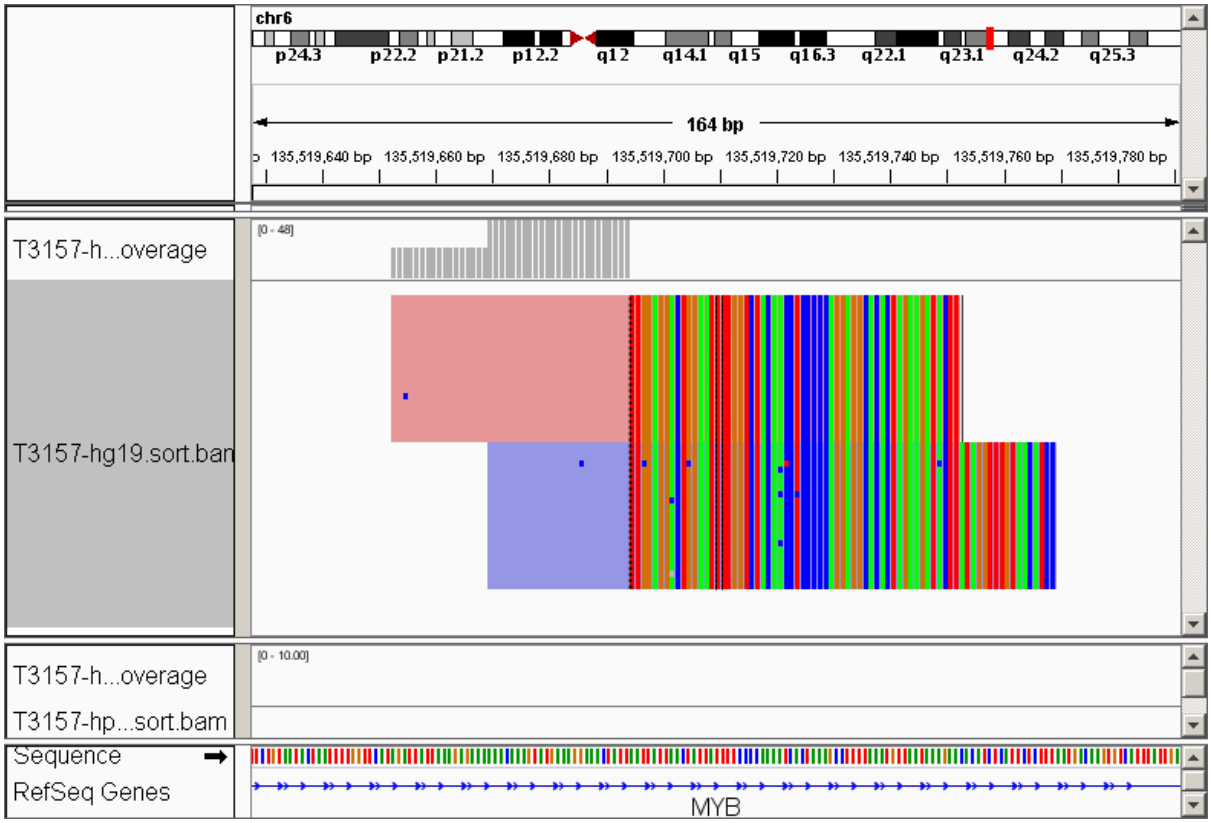


Figure 32 **Table S5 integrations not identified in our analysis**
T3157 Insert No 304 This integration was not identified by our algorithm, but is associated with multiple copies of a single fragment so does not represent an independently verified integration.

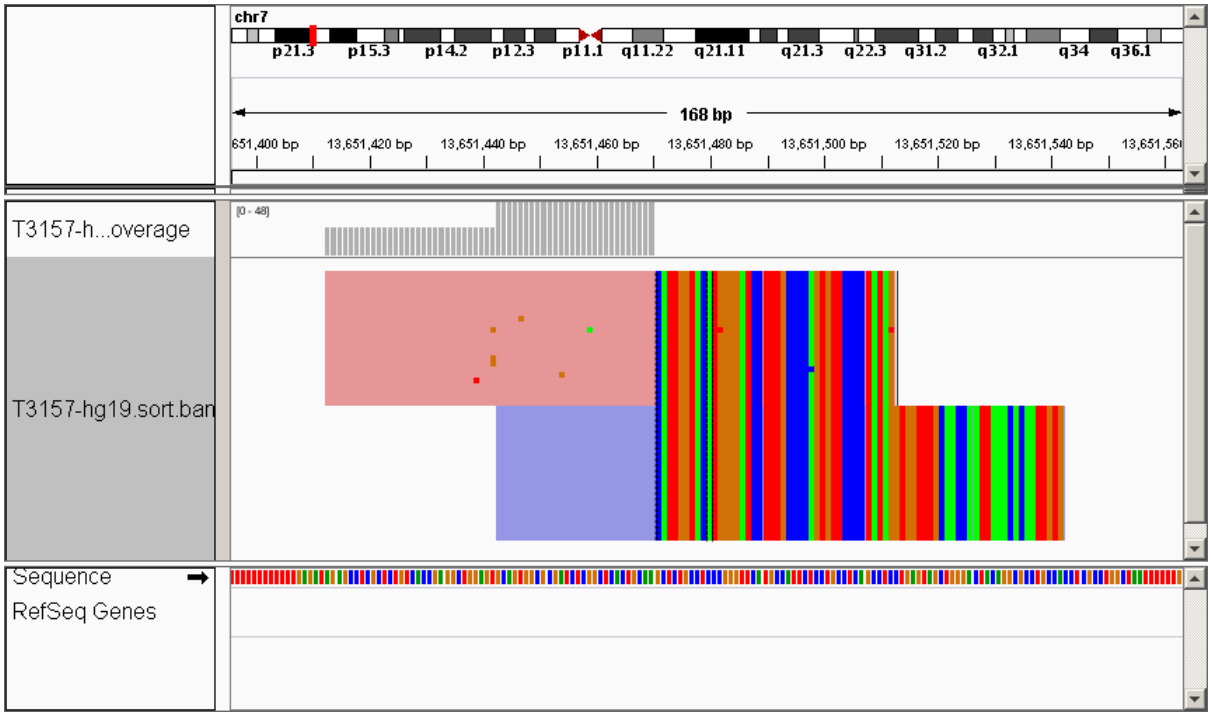


Figure 33 **Table S5 integrations not identified in our analysis**
T3157 Insert No 312 This integration was not identified by our algorithm, but is associated with multiple copies of a single fragment so does not represent an independently verified integration.

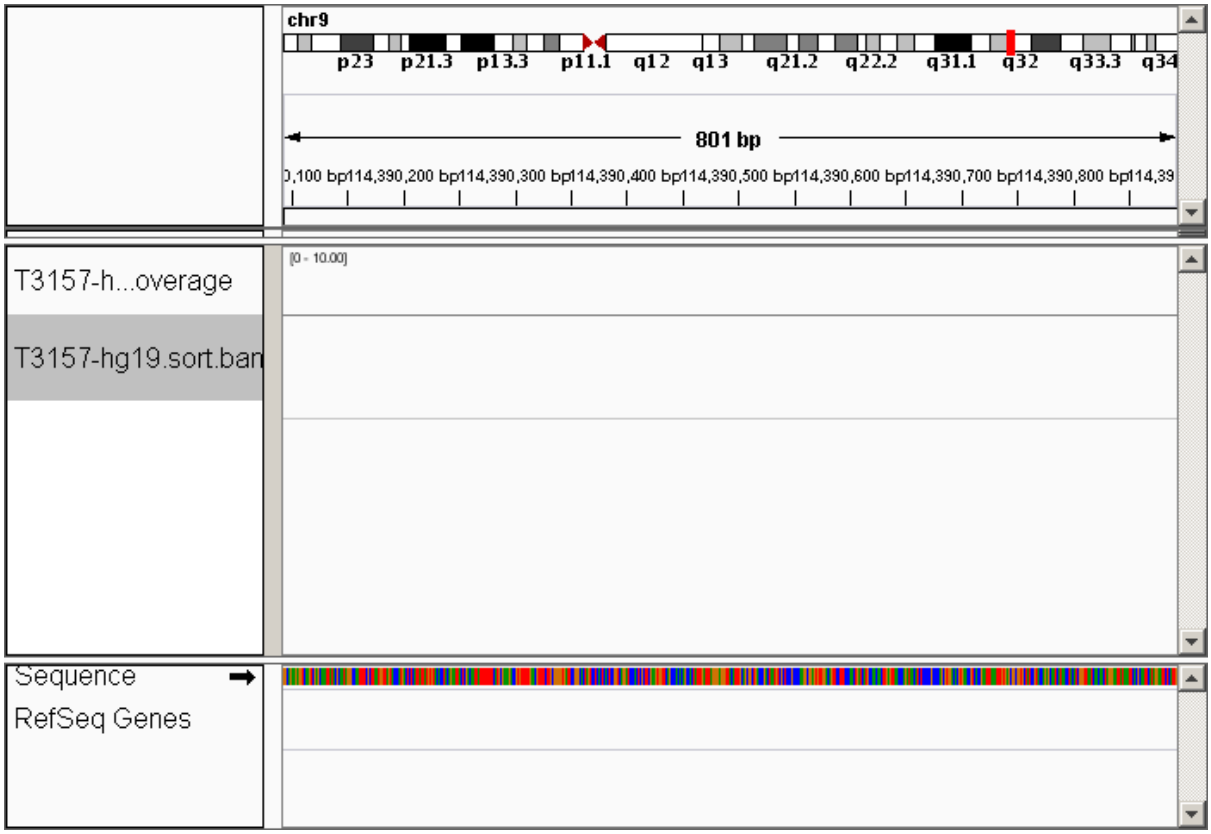


Figure 34 **Table S5 integrations not identified in our analysis**
T3157 Insert No 345 There are no fragments that map to this position with the bwa aligner we used. This is a similar case to that shown in Figure 24

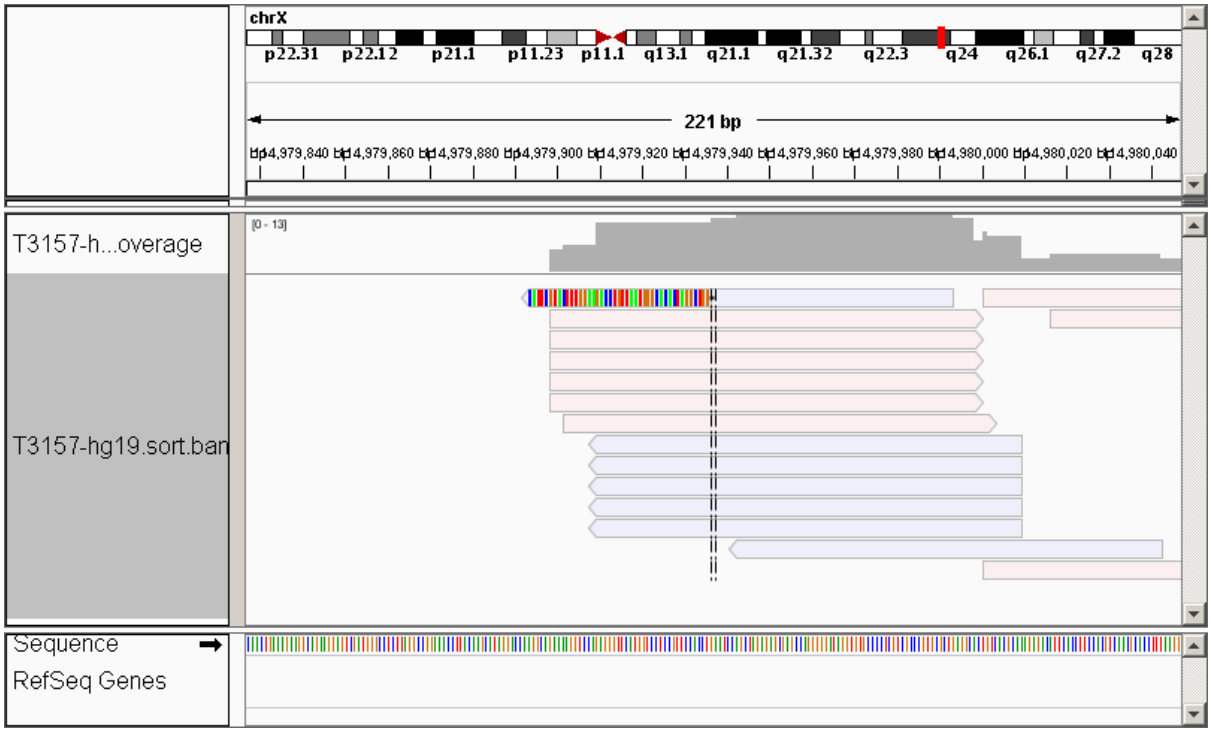


Figure 35 **Table S5 integrations not identified in our analysis**
T3157 Insert No 366. This is a region where the mappings are non-unique, so different decisions will be made by different aligners as to the location of the fragments. Our algorithm identifies a single read that is part HPV and part hg19, which is below the threshold for inclusion as a valid integration.