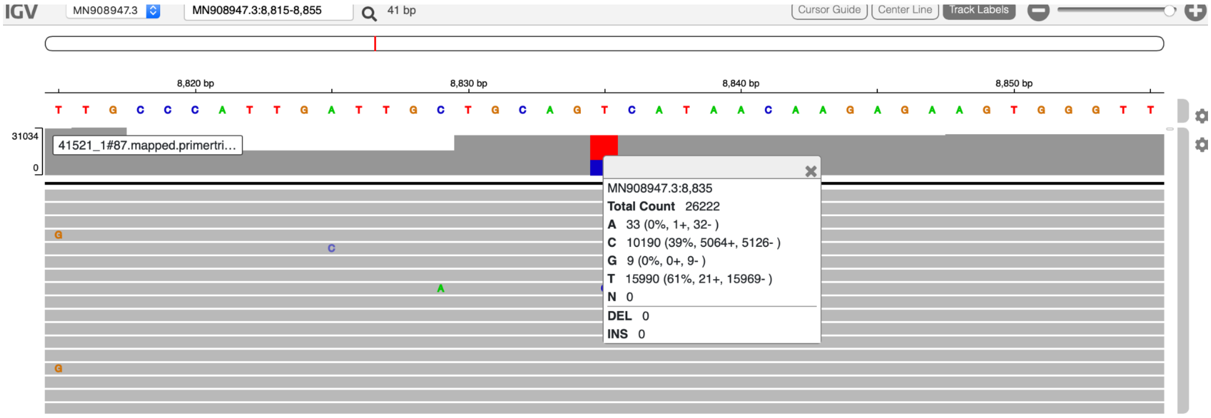
Justification for template length filtering

Description of Issue

Unexpected SNPs in samples where primer panel V4 was used and this was linked to unexpected PCR products. In particular SNPs at positions 8835 (mutation orf1ab:V2857A, T -> C SNP) and 15521 (orf1ab:F5086Y)

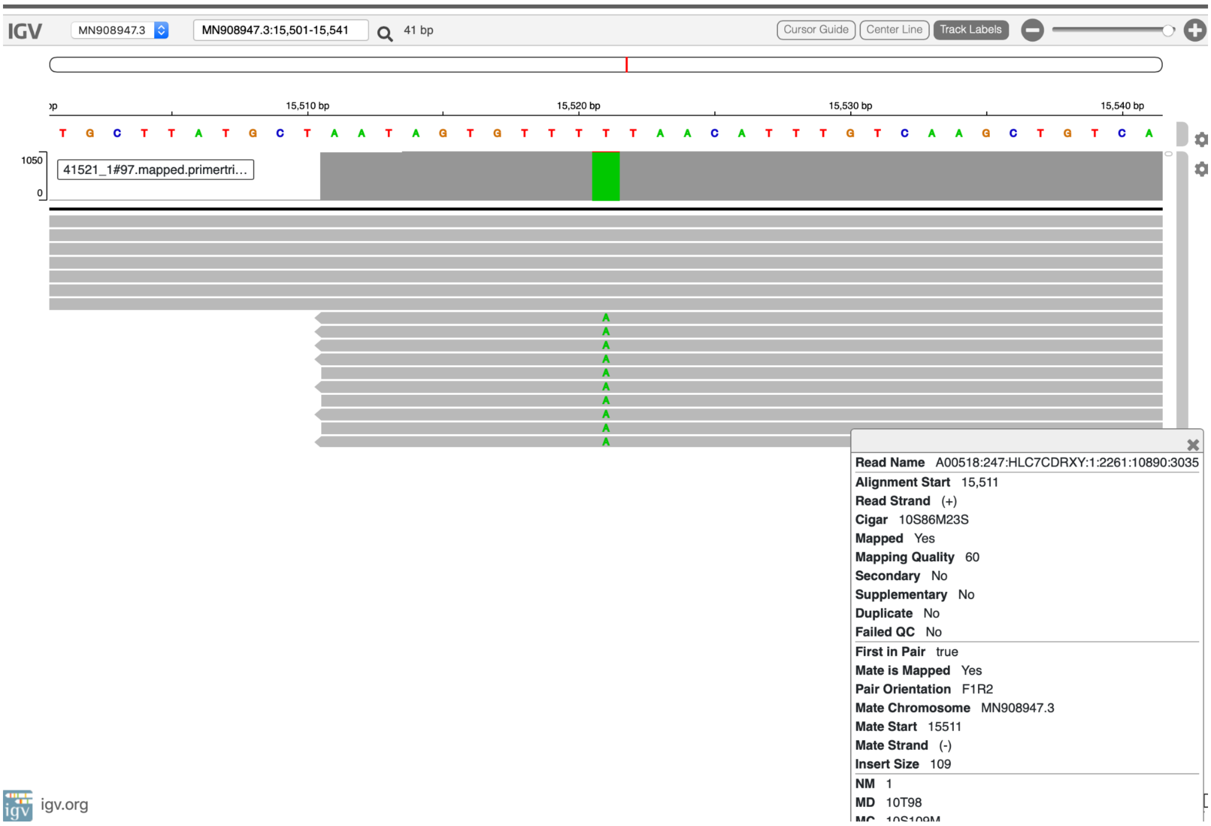
Example 1 from IGV, sample 41521_1#87 has a **T -> C** SNP at position 8835 in the variant calling file . This variant was called as it occured in 39% of reads, exceeding the 25% cutoff.

```
REGION  POS     REF     ALT     REF_DP  REF_RV  REF_QUAL        ALT_DP  ALT_RV  ALT_QUAL
ALT_FREQ        TOTAL_DPVAL     PASS    GFF_FEATURE     REF_CODON       REF_AA  ALT_CODON
ALT_AA
8835    T       C       15396   15375   35      9976    4945    36      0.393189  25372    0
TRUE    NA      NA      NA      NA      NA
```



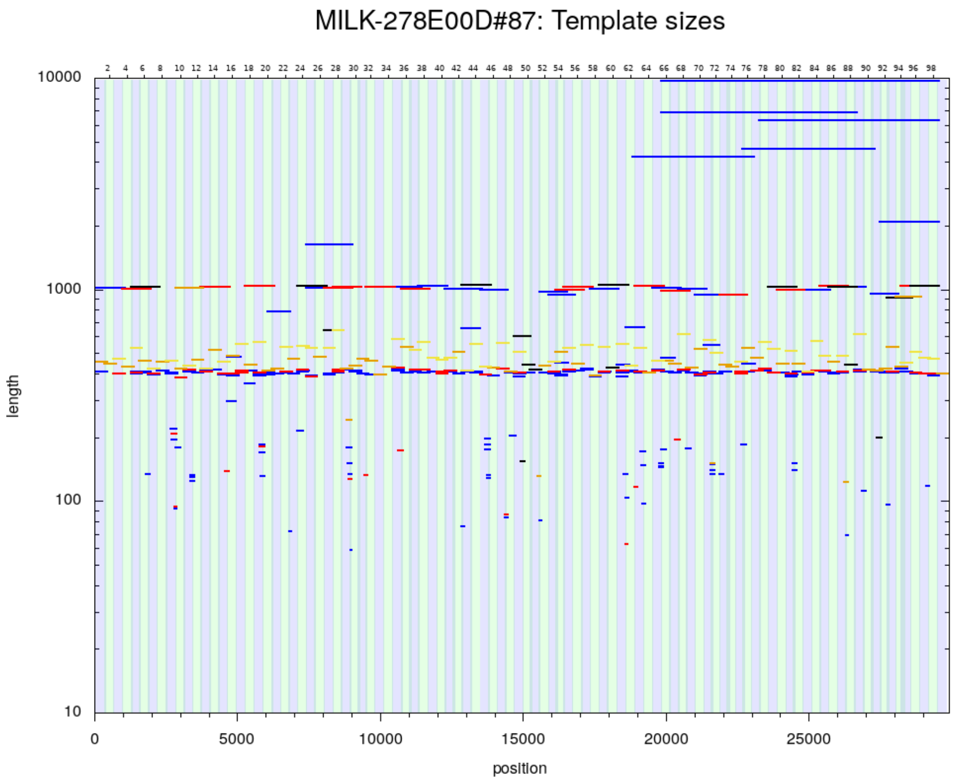
MN908947.3:8,835
Total Count 26222
A 33 (0%, 1+, 32-)
C 10190 (39%, 5064+, 5126-)
G 9 (0%, 0+, 9-)
T 15990 (61%, 21+, 15969-)
N 0
DEL 0
INS 0

Example 2 from IGV , 41521_1#97.mapped.primertrimmed.sorted.bam,
T -> A SNP at 15521 in the short reads (mapping to reference position 15511 to 15596) is incorrect.



Read Name A00518:247:HLC7CDRXY:1:2261:10890:3035
Alignment Start 15,511
Read Strand (+)
Cigar 10S86M23S
Mapped Yes
Mapping Quality 60
Secondary No
Supplementary No
Duplicate No
Failed QC No
First in Pair true
Mate is Mapped Yes
Pair Orientation F1R2
Mate Chromosome MN908947.3
Mate Start 15511
Mate Strand (-)
Insert Size 109
NM 1
MD 10T98
MC 10S109M

Image showing presence of short length templates



Proposed solution[1]

Filtering of short products less than 300 bp

```
samtools view -o ${sampleName}.filtered.mapped.bam -U ${sampleName}.bad.mapped.bam -e
'!flag.munmap && (tlen > 300 || tlen < -300)' -F 2048
```

This should not remove genuine data as our minimum valid template is 400 bp

Results

A comparison of the consensus and variant calling pipeline results for run 41521 was done.

Consensus file comparison

Comparing the 384 consensus files from run 41521 lane 1 before and after tlen filtering, 71 files were unchanged and 313 differed. The majority of the changes were an N in the filtered version[2] . Total positional changes across all samples were 4656 (1533 unique positions, 626 with >1 occurance). 158 of these were at reference position 8835. 284 at position 15521

Variant calling file comparison

Comparing the 384 ncovIlluminaCram_ncovIllumina_sequenceAnalysis_callVariants tsv files from 41521 lane 1, 324/384 files had some changes in variants called.

There were 170 occurrences of T -> C at reference position 8835 in the original and none of these were in the filtered files. There were 301 occurrences of T -> A at reference position 15521, 3 of these were still present in the filtered file (low depth and/or FAIL)

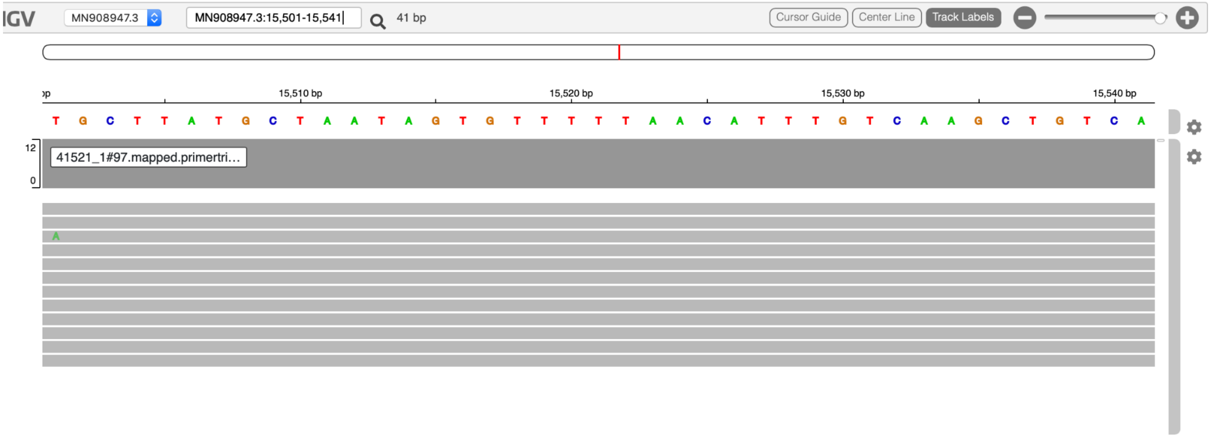
The majority of the other variant changes were due to SNPs removed after filtering. Of these, there were 37 across 32 samples. 33 were singleton SNPs. The 4 non-singletons were at position 71 (4 samples), 72 (11 samples), 14552 (2 samples) and 15451 (3 samples).

6 samples acquired a SNP after filtering. 5 of these were at position 28271 which was already widely observed in other samples. 1 from sample 41521_1#383 was at 24825 and was low depth.

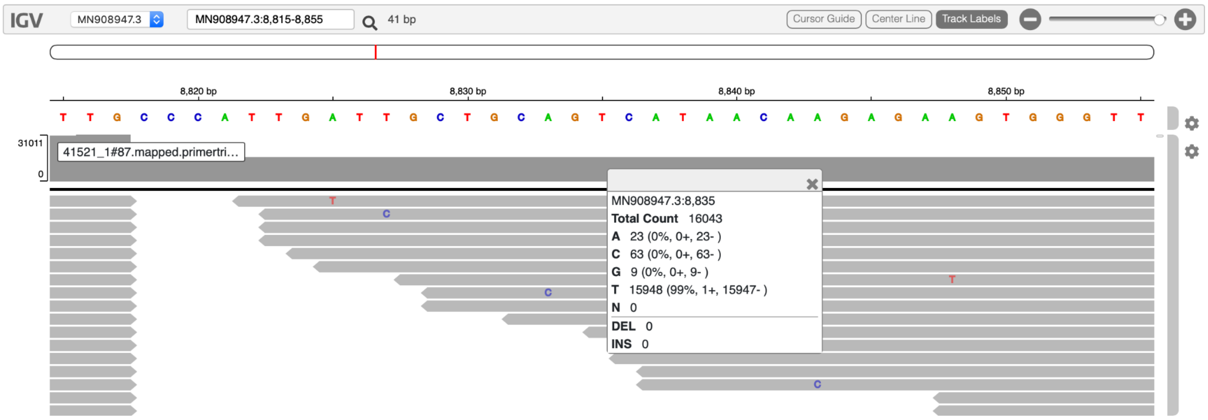
[1] Proposed by Rob Davies, WSI

[2] An exception was 41521_1#106, which had a 35 bp deletion in the original file

Example from IGV post filtering, 41521_1#97.mapped.primertrimmed.sorted.bam, reads containing the variant have been removed.

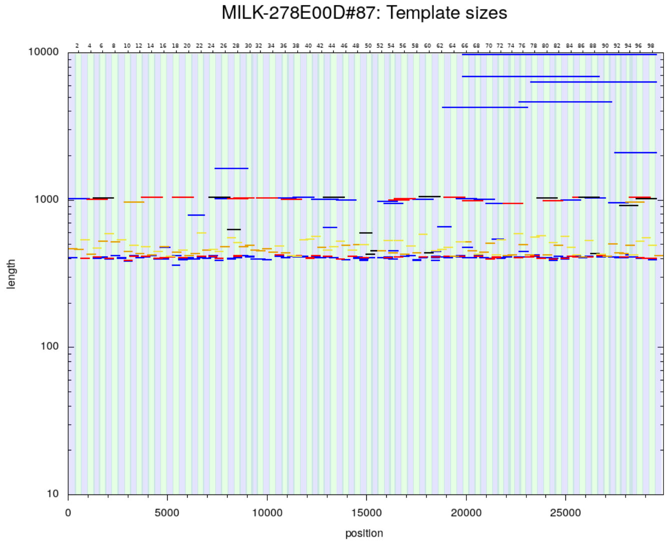


Example from IGV post filtering, 41521_1#87.mapped.primertrimmed.sorted.filtered.bam, over 99% of remaining reads match the reference at position 8835



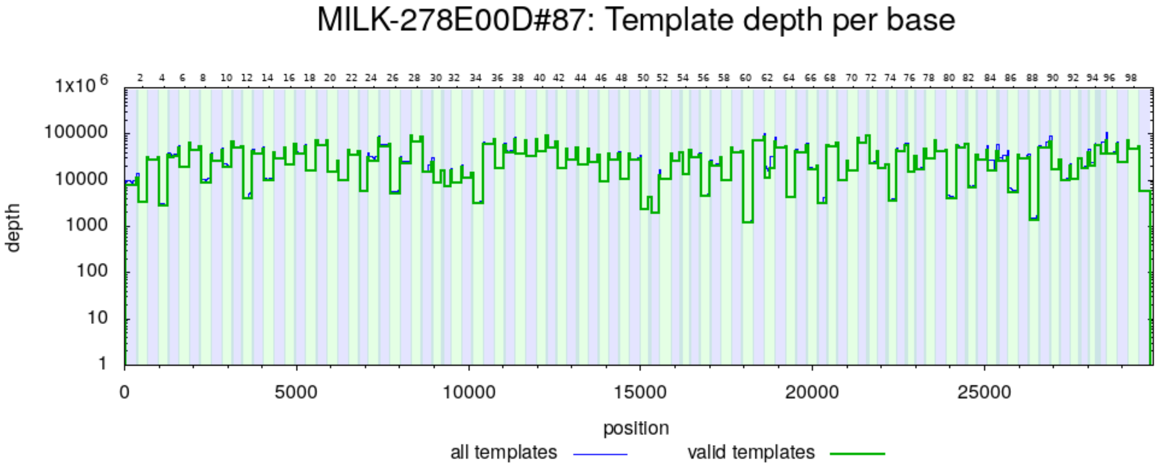
MN908947.3:8,835
Total Count 16043
A 23 (0%, 0+, 23-)
C 63 (0%, 0+, 63-)
G 9 (0%, 0+, 9-)
T 15948 (99%, 1+, 15947-)
N 0
DEL 0
INS 0

<u>Image showing short templates having being removed.</u>

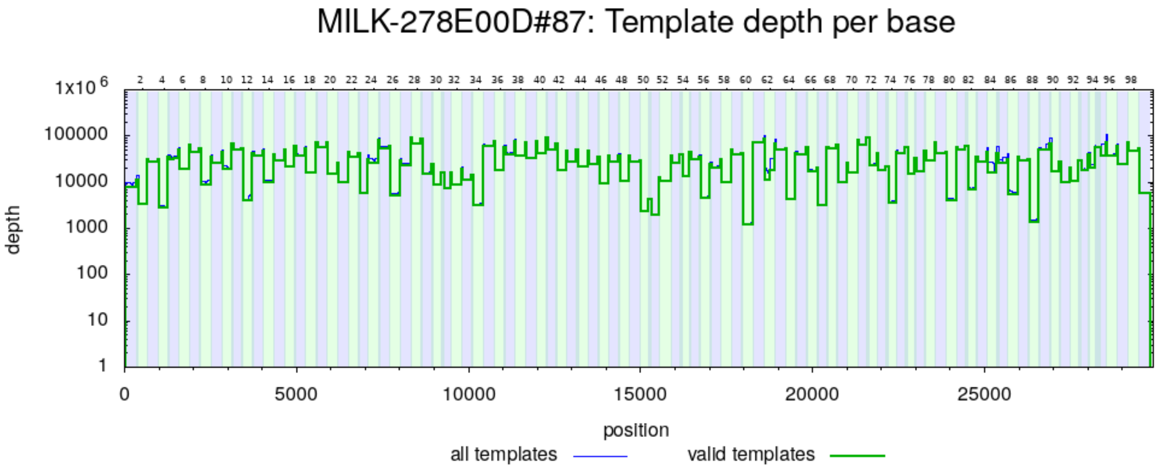


Valid templates appear unchanged

Original



Filtered



Artic pipeline duration

The time for the artic pipeline decreased for the filtered run. (Times in minutes)

	Min	Max	Mean	Median	Mode
Original:	1.07	34.73	14.04	14.48	16.28
Filtered:	1.02	30.83	10.48	10.83	9.69