# Identifying a Duplication

The case study involves the ***Agmo*** knock out mouse also discussed in Case 2 in the Identification of insertions read me file.

This transgenic mouse is described here:

> Sailer S, Coassin S, Lackner K, Fischer C, McNeill E, Streiter G, Kremser C, Maglione M, Green CM, Moralli D, Moschen AR, Keller MA, Golderer G, Werner-Felmayer G, Tegeder I, Channon KM, Davies B, Werner ER, Watschinger K. When the genome bluffs: a tandem duplication event during generation of a novel ***Agmo*** knockout mouse model fools routine genotyping. Cell Biosci. 2021 Mar 16;11(1):54. doi: 10.1186/s13578-021-00566-9. PMID: 33726865; PMCID: PMC7962373.

**Background**

A transgenic ***Agmo*** knockout mouse was created by inserting a lacZ-neoR cassette close to exon 2 of the ***Agmo*** gene. However, difficulties in genotyping ***Agmo***-deficient mice led to the sequencing of the mouse and the identification of a 94 Kb tandem duplication of the 5' end of the ***Agmo*** gene.

The data for this experiment is hosted on the NCBI SRA site as SRR12783028.

Reads which mapped to the start of the ***Agmo*** gene and contained extended unaligned data were used to search the NCBI blast database to obtain sequences homologous to the lacZ-NeoR cassette. This identified the sequence JN960306.1 from which the sequence for the lacZ-NeoR cassette was extracted (15041 to 22164 bp) and added to mm10 mouse genome reference sequence as a separate reference sequence called transgene. The long read sequencing data was then aligned to this extended reference sequence and used in this guide.

While the detection of the cassette is separate from the detection of the duplication, its identification is shown below and in the insertion walk through.

The RefSeq gene data for the mm10 genome reference was downloaded from the Genome Browser's Table Browser as describe here.

## Analysis

Import the aligned data by pressing the `BAM file` button. While it's possible to determine the location of the ***Agmo*** gene from a number of sources, in this example we'll get `AgileStructure` to identify the region using the RefSeq gene data set. First, download the data set as described here. Then, select the `Annotation` > `Gene annotation file` menu option and choose the downloades file (Figure 11). The annotation file must correspond to the reference build used to align the data.

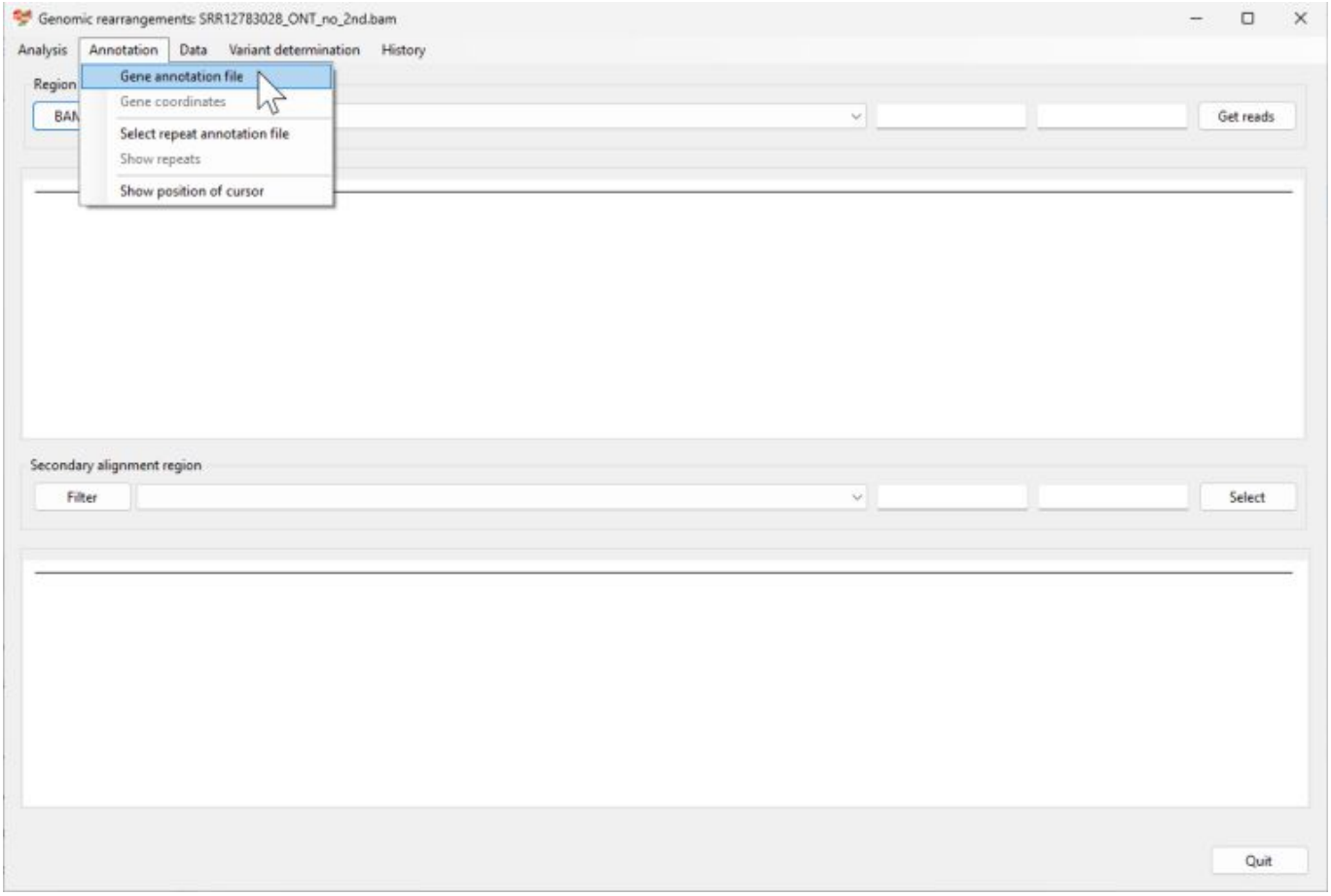


Figure 11

The file will take a couple of seconds to load before you can select the `Annotation` > `Gene coordinates` menu option, which will open the `Gene coordinates` window (if no bam file has been selected, this window will not appear). Enter ***Agmo*** into the upper text area and press the `Find` button. The coordinates for ***AMGO*** will then appear in the lower text area. (Figure 12)

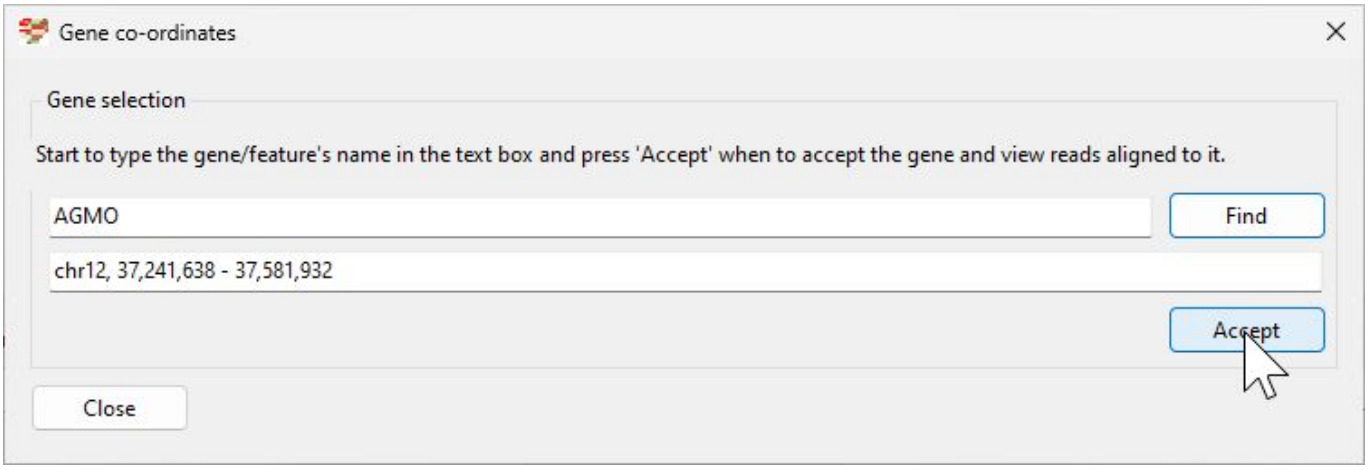


Figure 12

Pressing the `Accept` button will then cause the gene's coordinates to appear in the dropdown list and text areas of the upper panel. Since the duplication affects the 5' end of the gene, change the start of the display region from 37,241,638 bp to 37,100,000 bp and press the `Get reads` button to display reads mapping to ***Agmo*** and 5' upstream sequences. Since a RefSeq annotation file was entered, the ***Agmo***'s exons will be displayed at the bottom of the display panel (Figure 13).

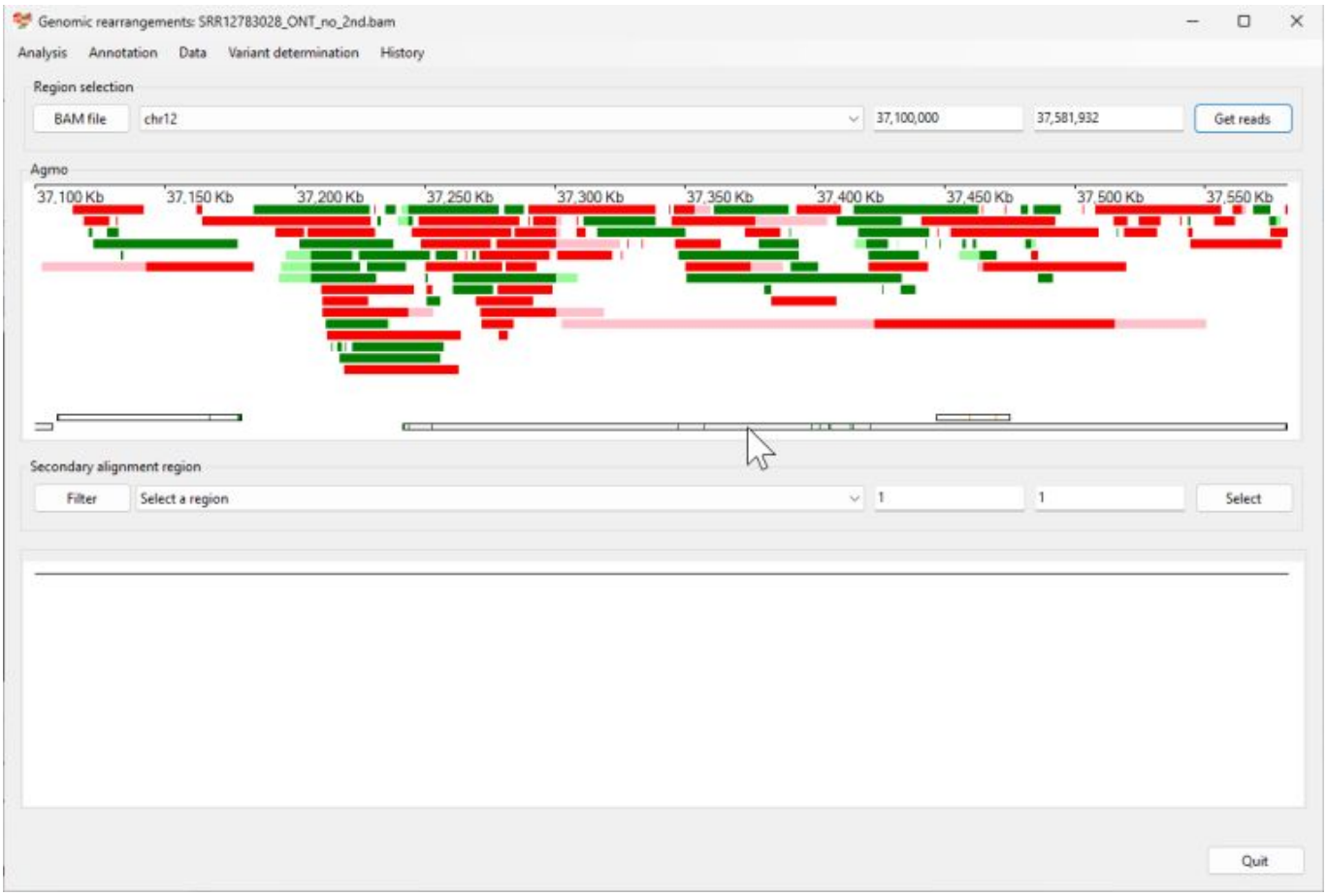


Figure 13

To view the secondary alignments in the upper panel, select a region on chromosome 12 from the dropdown list in the lower panel. Adjust the display limits to match the upper panel, then choose the reads that span the breakpoints (Figure 14).

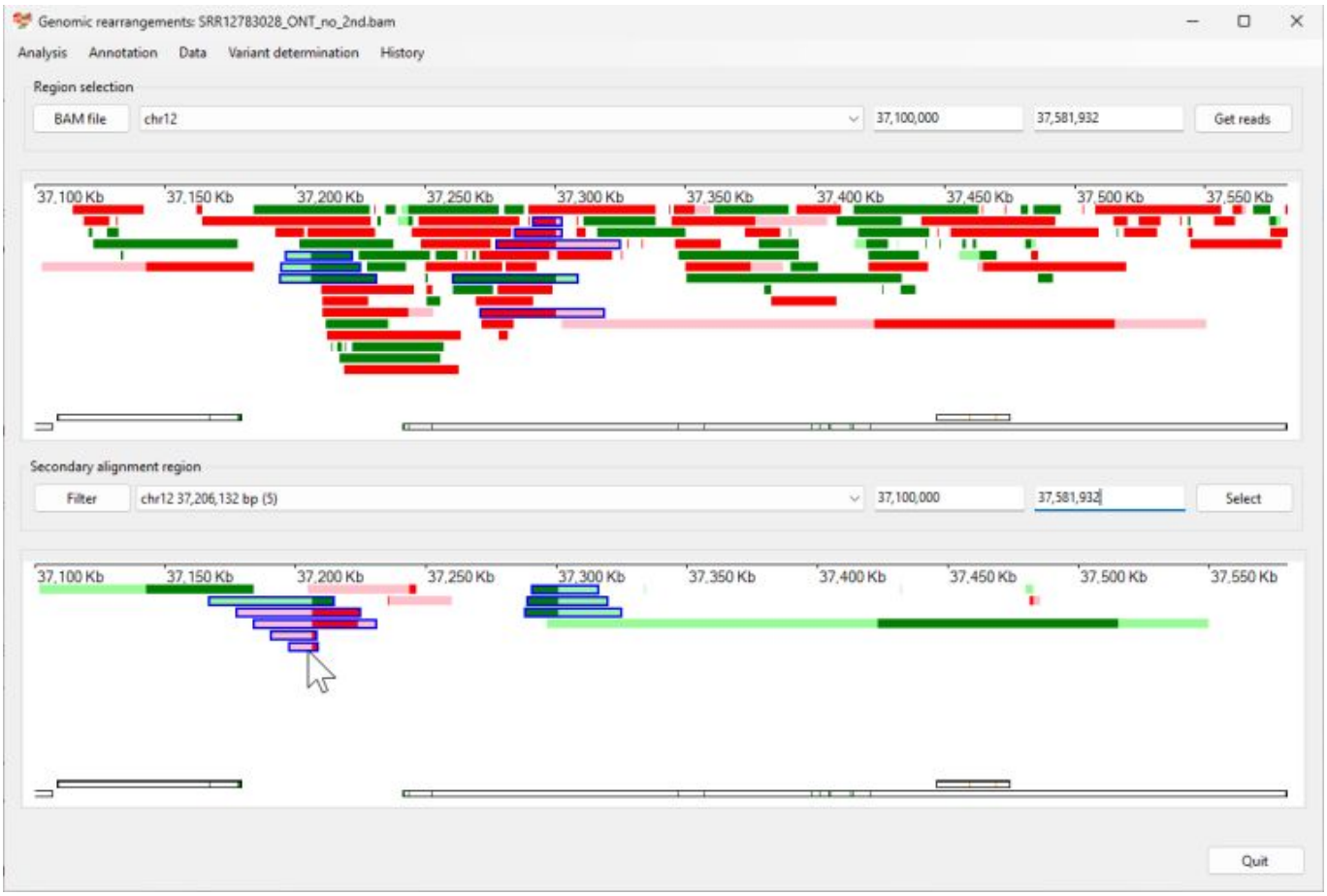


Figure 14

Selecting the `Variant determination` > `Use soft clip data` > `Duplication` menu option, prompts `AgileStructure` to annotate the variant (Figure 15)

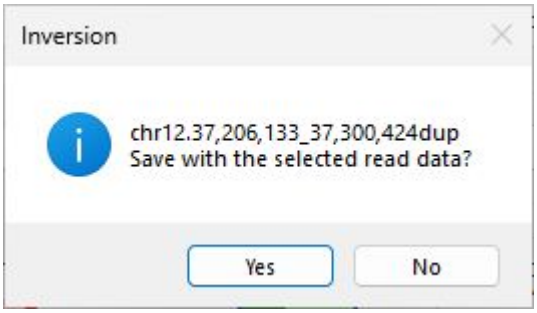


Figure 15

The variant **chr12.37,206,133_37,300,424dup** closely matches the region published duplicated region: **chr12:37,206,133–37,300,425**.

## Identification of the lacZ-NeoR cassette

As stated above the transgenic mouse also included a lacZ-NeoR cassette. This sequence has been added to the mouse mm10 reference sequence and called 'transgene'.

Following on from the detection of the duplication, all selected reads were deselected by using the `Data` > `Clear selected reads` menu option (Figure 16)

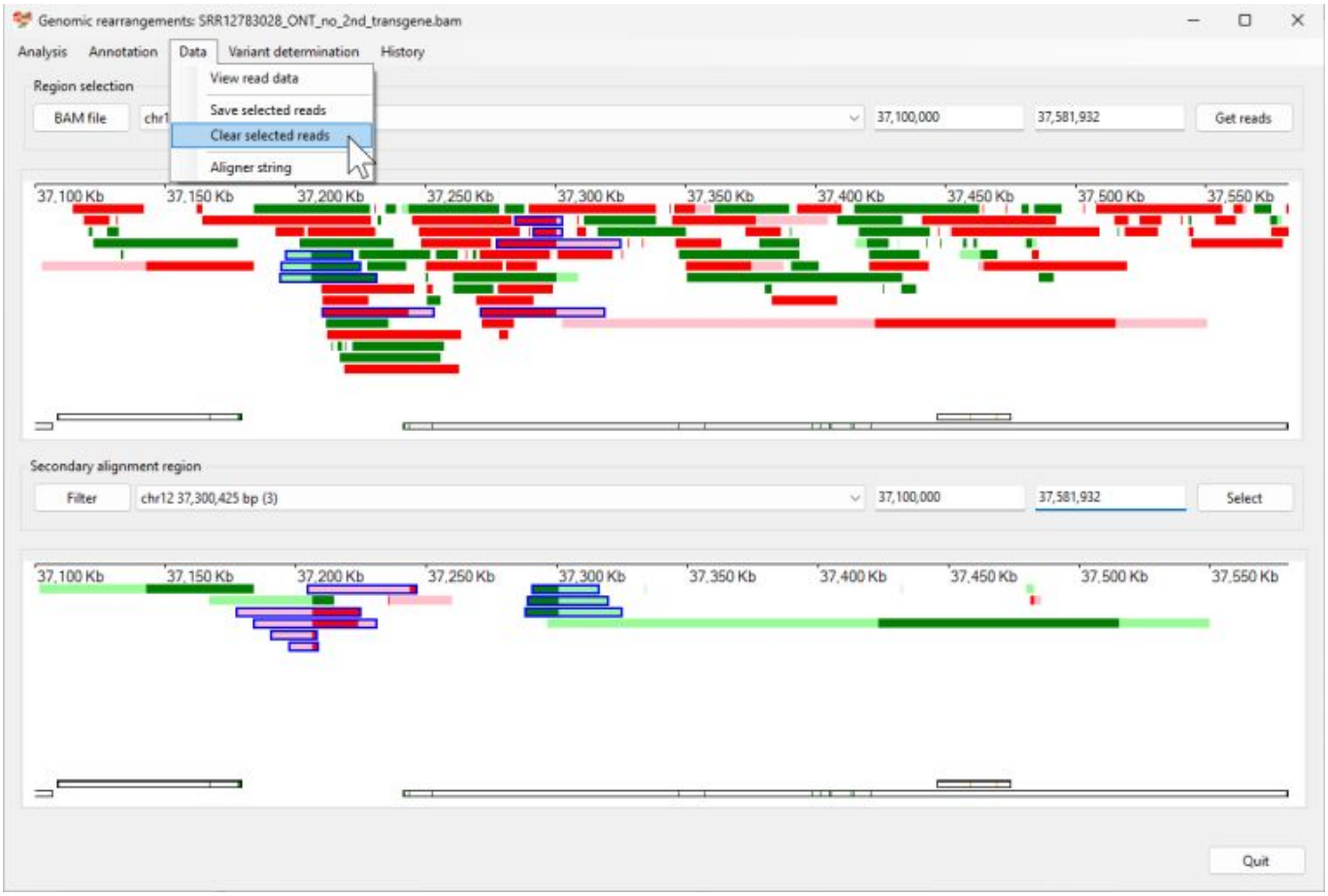


Figure 16

From the lower dropdown list, select the region linked to the 'transgene' references sequence. Then, in the lower panel select the two reads mapping to the transgene sequence by clicking on them (Figure 17)

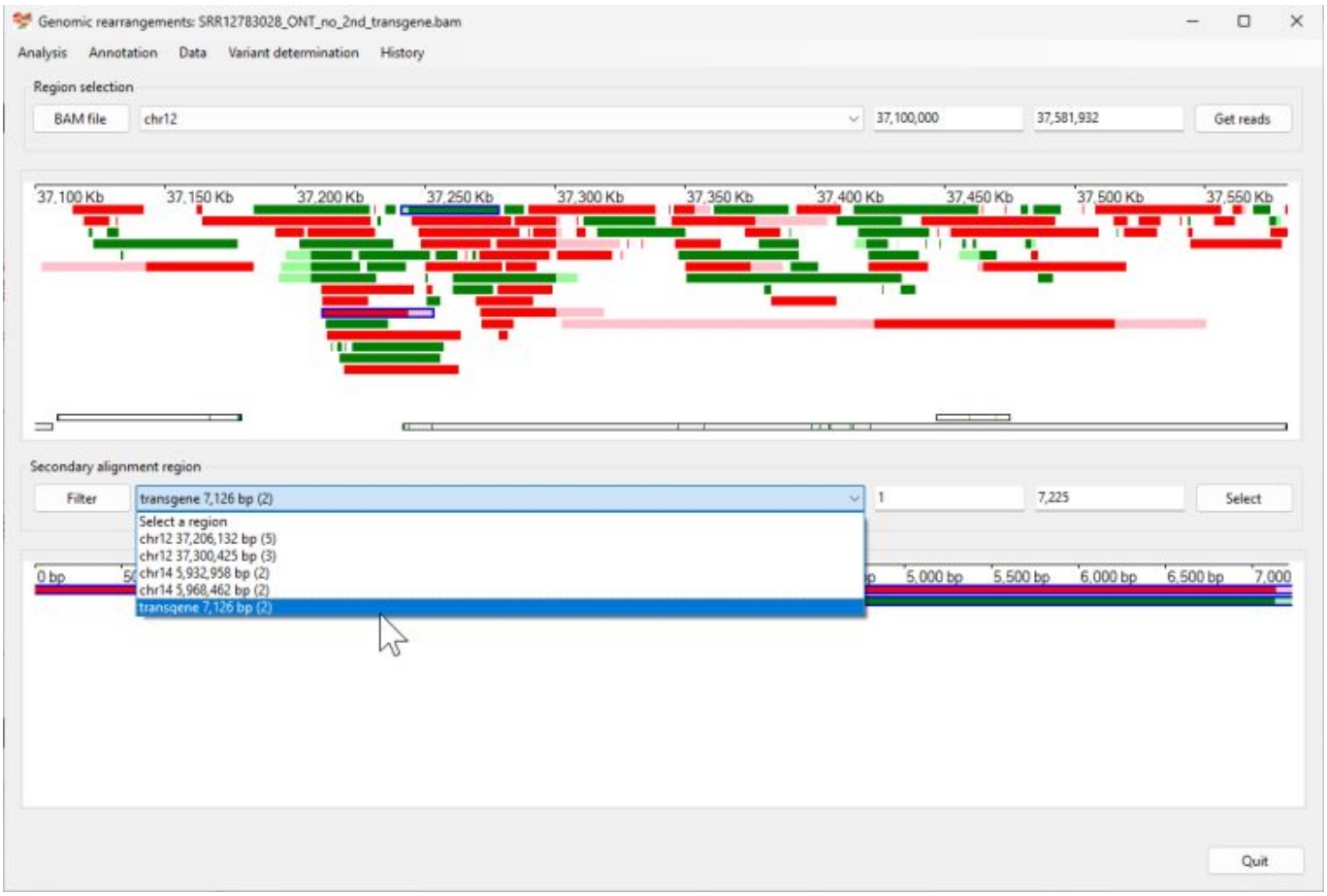


Figure 17

The primary alignment for these reads can be seen at approximately 37,243,000 bp on chromosome 12 in intron 1 of the ***Agmo*** gene as expected (Figure 18). To annotate the insertion, go to **Variant determination** > **Use soft clip data** > **Insertion**. AgileStructure will then identify the insertion as `chr12:37,243,310ins transgene.4,567,7126` (Figure 19), which is near exon 2, the known location of the cassette.

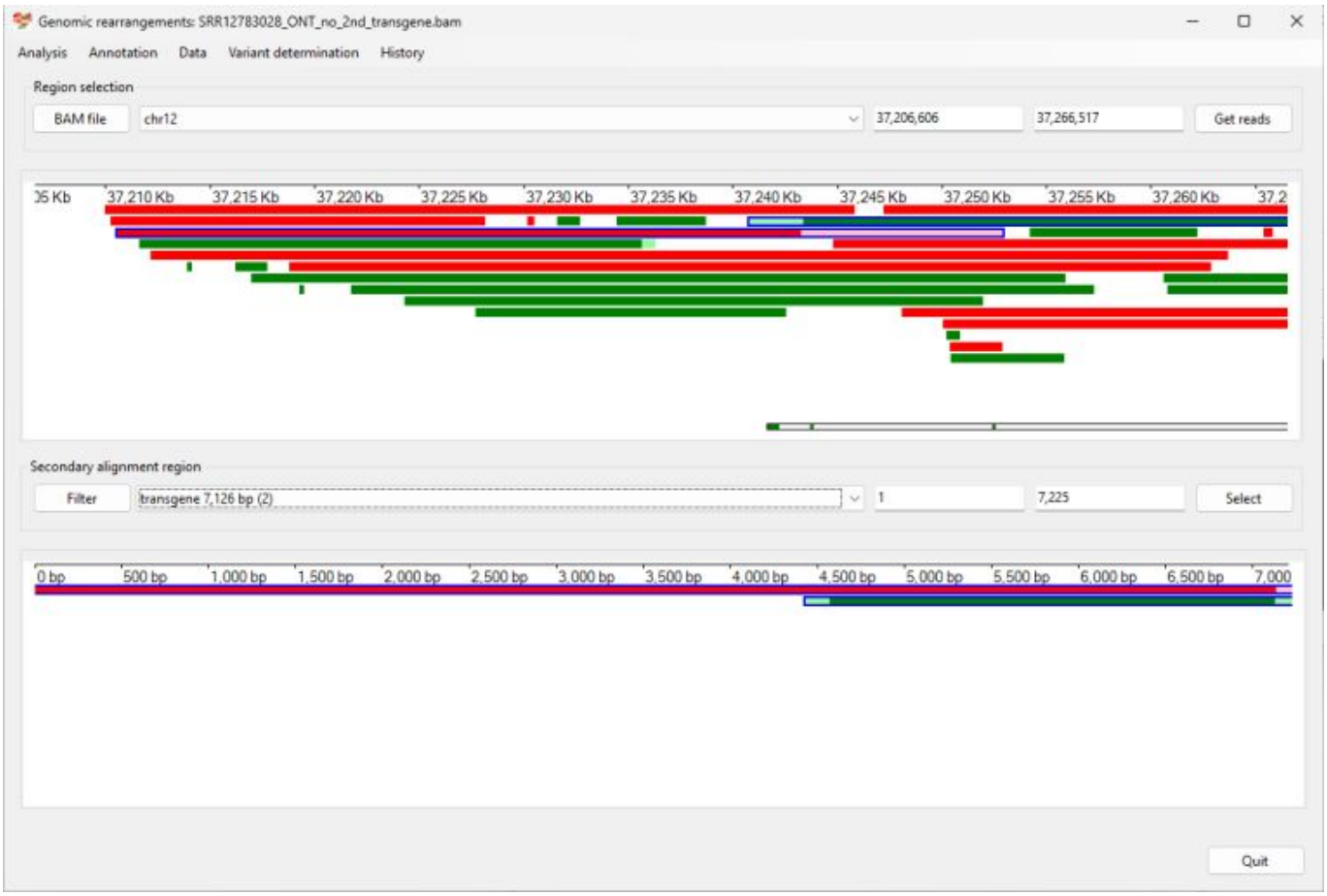


Figure 18

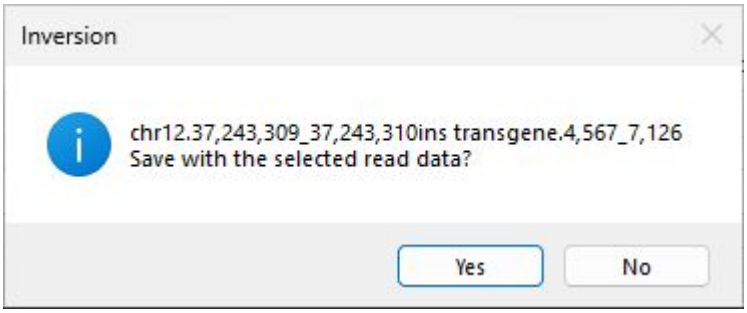


Figure 19