AgileStructure user guide

Table of contents generated with markdown-toc

`AgileStructure` is composed of three components: AgileStructure.exe, AgileStructure.dll and AgileStructure.runtimeconfig.json, to work, all the files need to be in the same folder.

Data requirements

Prior knowledge of the likely location of the break point

`AgileStructure` is designed to identify break points with user assistance rather than scan the whole alignment for possible break points, consequently its expected that the user will have some prior knowledge as to where the break point is such has a cytogenetics and/or karyotyping report, a list of known disease genes for the patients condition or a single pathogenic variant in a patient with a recessive disease for whom a second pathogenic variant can not be found.

Aligned data format

`AgileStructure` is designed to visualise aligned long read data formatted as indexed `bam` files. It's expected that the index file will have the same name as the `bam` file with the `*.bai` extension appended to the bam file's name, for instance the bam file:
`CNTNAP2.srt.mm2.bam`
will have a index file named:
`CNTNAP2.srt.mm2.bam.bai`
which will be in the same directory as the bam file.
The `bam` file must contain the header section which lists the name and size of each reference sequence in the reference genome.

Preferred long read sequence aligners

Long reads that span a break point will appear to consist of two regions of homology, mapping to different locations in the genome. How these chimeric alignments are reported are aligner specific. Some aligners such as minimap2 (github, paper), treat the two regions as different alignments, but will report the secondary alignment as a condensed CIGAR string in the primary alignments tag section, while others, report the read as two separate alignments, but not directly reference the other alignment's position and CIGAR string. However, for shorter indels both types of aligner may report them in the CIGAR string (see this section).
`AgileStructure` is only able to analyse reads in which the indel is reported in the primary alignment's CIGAR string or the secondary alignment is reported in the tag section: The reporting of the secondary alignment in the tag section is the most flexible method and will allow more complex break points to be processed. Consequently, it is recommended to align data using an aligner such as minimap2.

Optional data

To aid the analysis, it is possible to view the putative break points with reference to the location of repeat and gene sequences. This data can be obtain from the USCS genome browser as described here.

Importing alignment data

Data is imported as a pair of files, the pre-aligned bam file and its index file, by either pressing the `BAM file` button (Figure 1a) or by selecting the `Analysis` > `Open BAM file` menu option (Figure 1b) and selecting the required `bam` file.

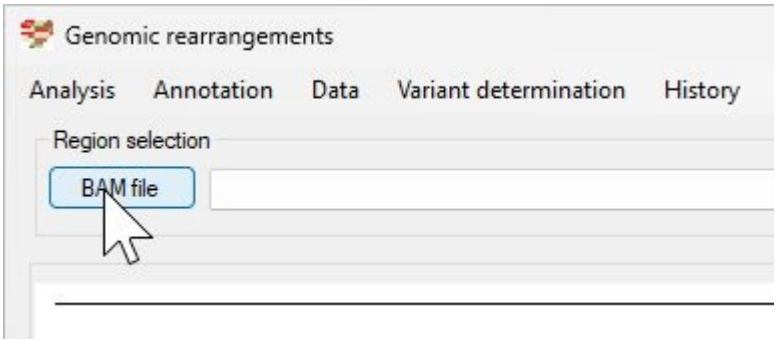

Figure 1a

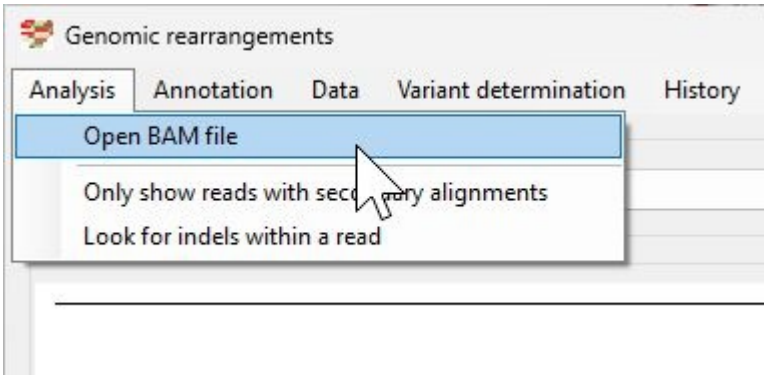

Figure 1b

`AgileStructure` will read the header section of the bam file and populate the dropdown list box next to the `BAM file` button, with the name of the reference sequences in the bam file (Figure 2).

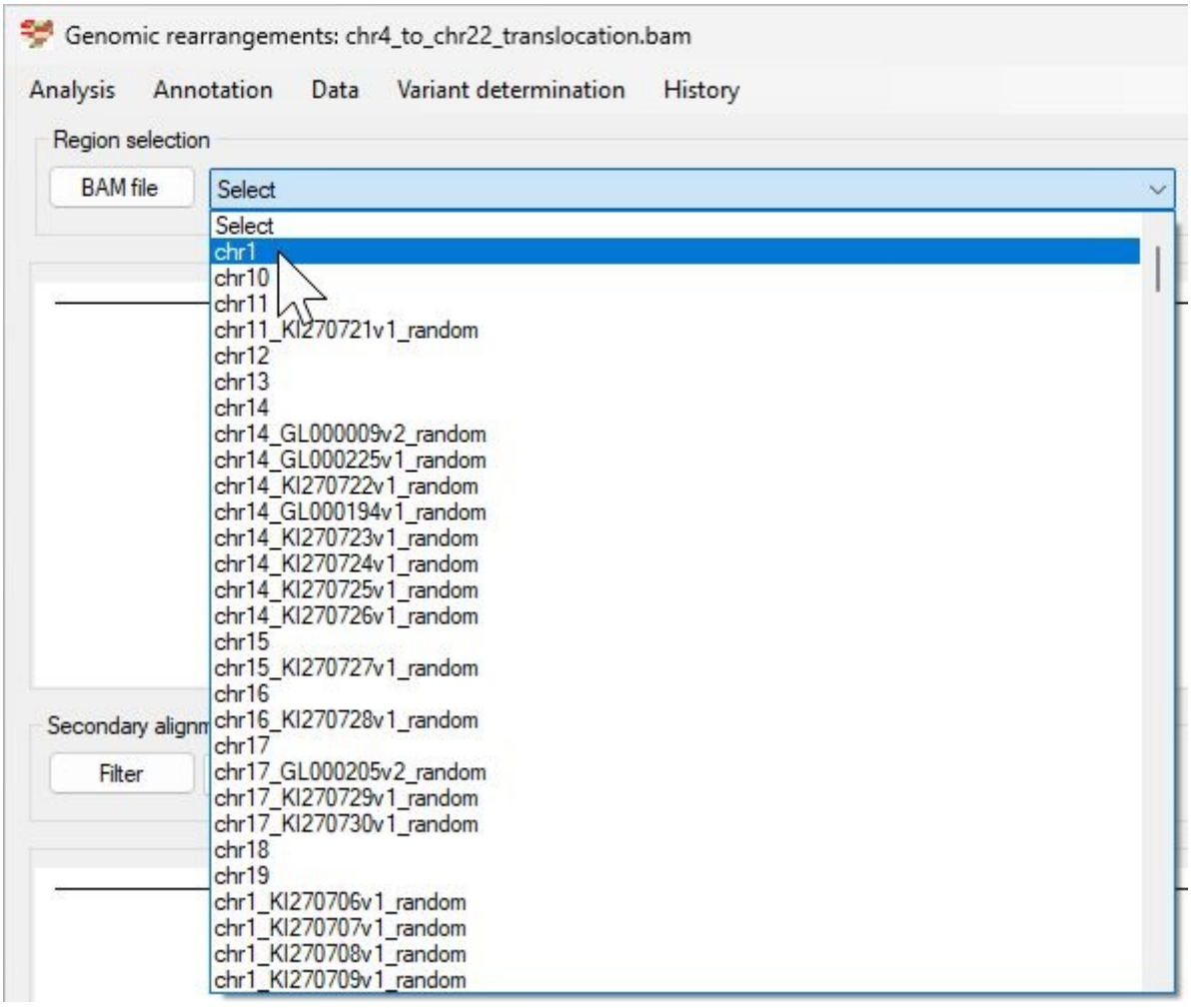

Figure 2

How indexed bam files are processed

Before indexing a bam file, reads are ordered with reference to the read's genomic coordinates with the data then compressed with the qzip algorithm to form discrete chunks of compressed data containing reads that start in a small region. For example the first chunk in a file may contain reads with alignments starting on chr 1 between 1 bp and 16,000 bp with the next chunk containing reads whose alignment starts on chr 1 between 16,000 bp and 32,000 bp. The indexing of a file creates a 2nd file (.bam.bai) that lists the start point of each chunk in the bam file and the start position of the first read in that chunk.
When a program has to find reads mapping to a certain region, it looks in the bam.bai files to find the chunks that contain the start positions of the reads mapped to the region. Once it's found the chunks mapping to the data, it looks up in the index file were that data starts in the bam file and then reads the data at that point in the bam file until it comes to the end of the compressed chunk. It repeats the process until it has read all the data for the region of interest. This works fine for short read data since all the reads in a chunk start between two well defined points (the genomic start site of the chunk and the end of the chunk plus the length of the read i.e. 16,000 bp to 32,000 bp plus 150 bp). However, with long read data, a read may be longer than the size of a chunk's region, so a read 20 kb long may start in one chunk while its end may be in the next or the next but one chunk. This causes an issue when you select a small region, as reads that over lap the region may be listed in a chunk much further upstream and so may not be found when reading the bam file. Consequently, `AgileStructure` reads the chunk that ends just before the start of the region of interest, but if a read is particularly long it may be missed as it starts even further upstream. Therefore when selecting regions, initially select a slightly larger (16 kb added to each side) region than required and note if a long read spanning the break point disappears when you shrink the region.

Selecting the region to view

Select the required chromosome (reference sequence) from the upper dropdown list box and enter the region's coordinates in the two text boxes to the right of this drop down list box and press the `Get reads` button (Figure 3). The coordinates are checked to make sure they are not greater than the chromosomes length as reported in the `bam` file. If no chromosome has been selected these values will be limited to '1'.

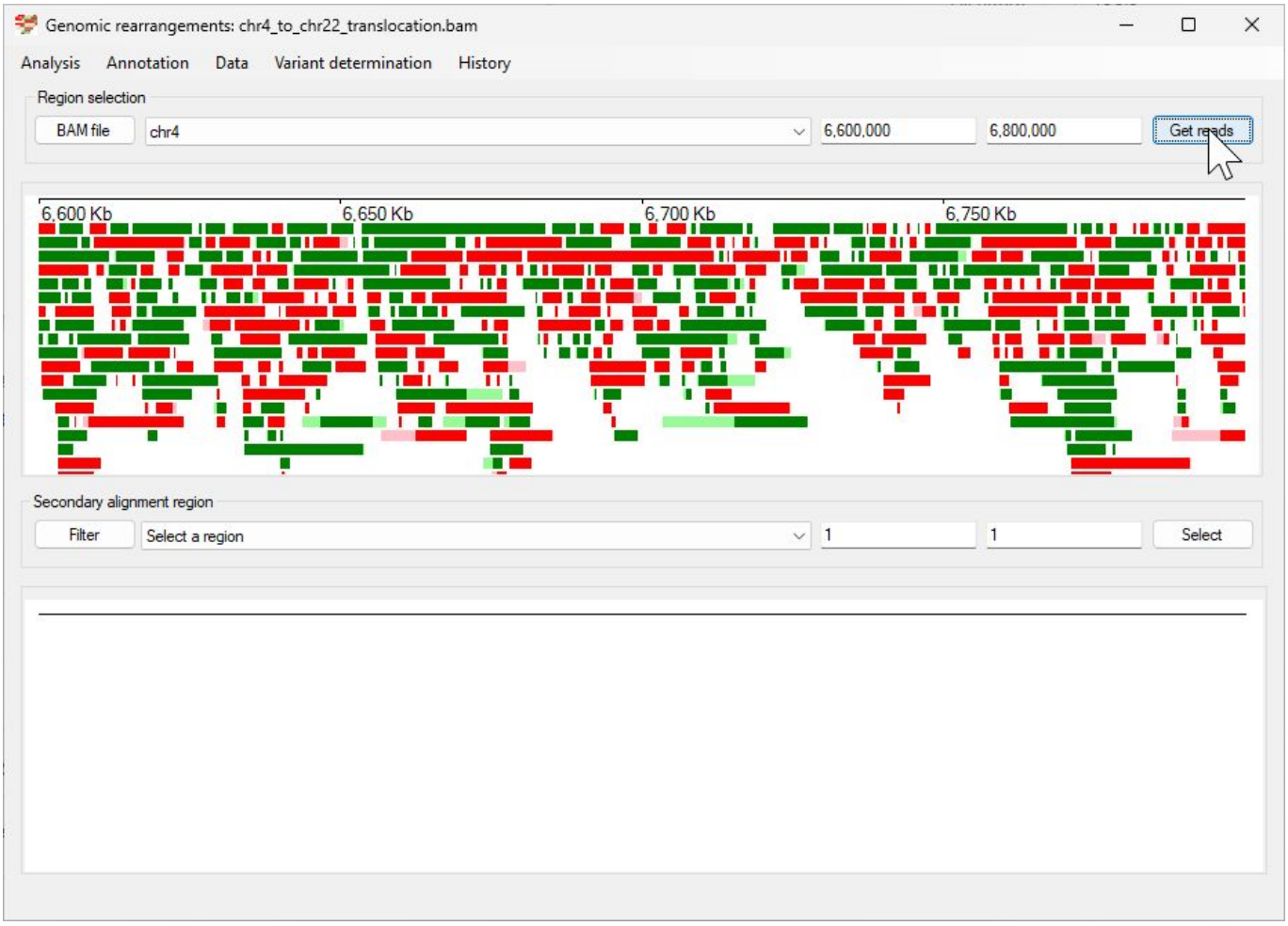

Figure 3

The position of reads mapping to the region are shown as green (aligned to the forward strand) and red (aligned to the reverse strand) rectangles scaled to the length of the read. Soft clipped sequences are identified as pale green or pale red extensions to the darker green/red aligned data. The size of the pale rectangles is proportionate to the length of the unaligned sequence and their location only indicates whether they are on the 5' or 3' of the aligned sequence.
It is important to note that in the default view, reads are drawn as a solid box spanning the length of the alignment, if a read as a large deletion this will not be shown, however they can be visualised by selecting the `Analysis` > `Look for indels within a read` menu option (see section Identifying Indels using the primary alignments CIGAR string).
`AgileStructure` does not have an upper limit on the size of the region or number of reads it will process and will attempt to read the requested data until it as processed the region or the computer runs out of memory. While there is no upper limit, you should try to limit the amount of data it reads as reading the underlying bam file can be a slow process due to its size.

Hiding reads without a soft clipped segment

When adding reads to the display, they are stacked so as little space as possible is used, however for alignments with a high read depth, the stacks may be to too tall to fit in the image. Since, reads that have a soft clipped region are more important in break point detection it is possible to hide those without an unaligned fragment by selecting the `Analysis` > `Only show reads with secondary alignments` menu option (Figure 4).

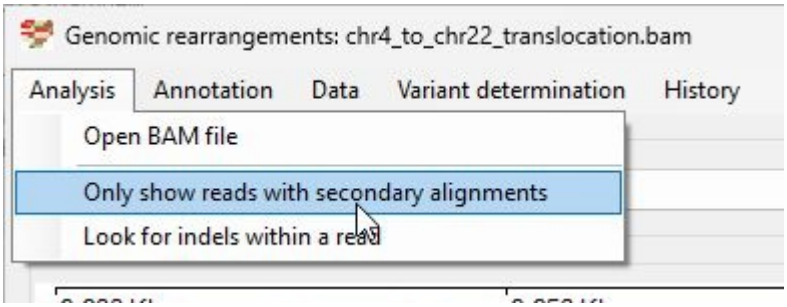

Figure 4

The filtered image will contain fewer reads, making those at the break point more apparent.

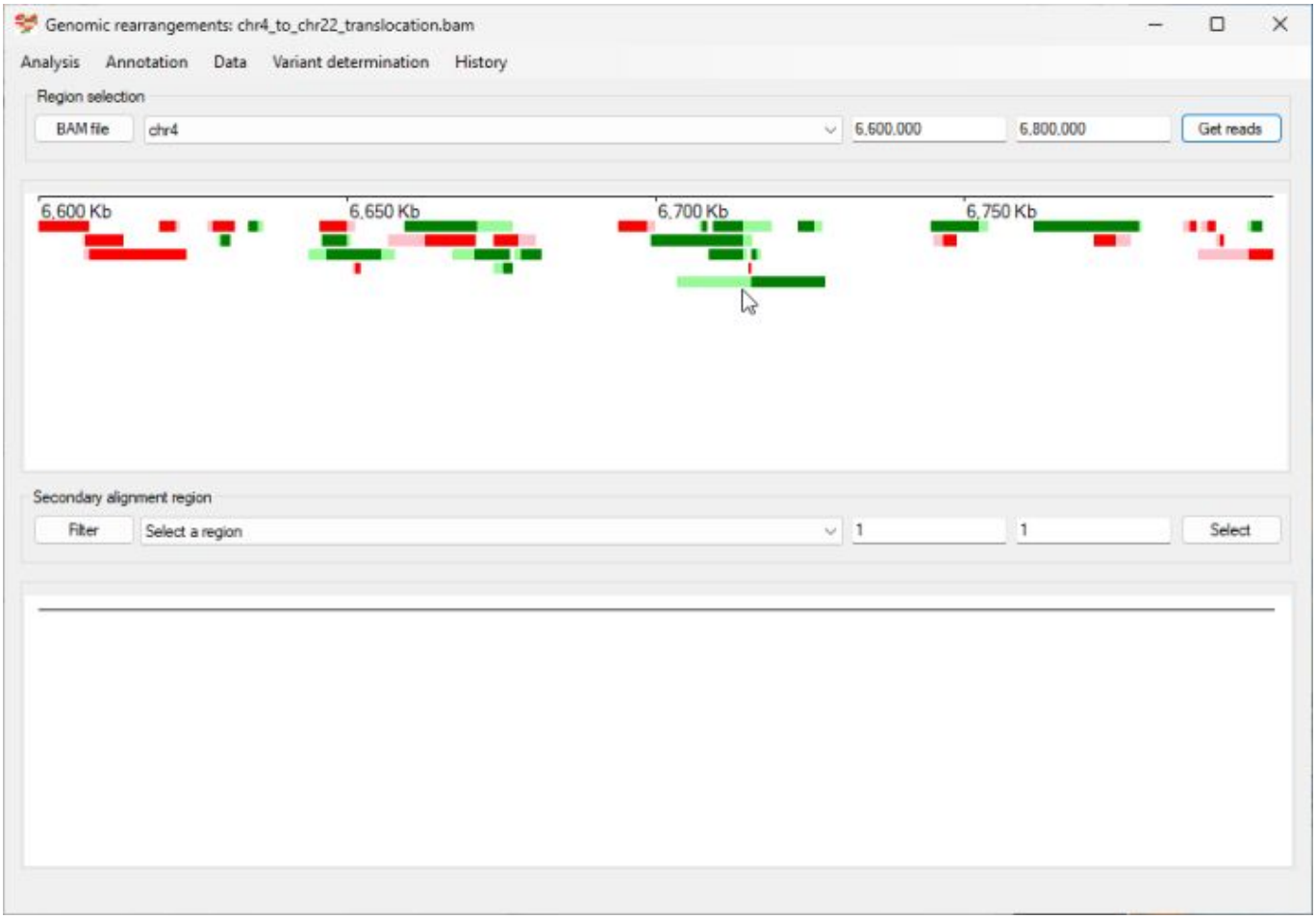

Figure 5: When reads with no secondary alignment are removed the break point is more apparent: see mouse cursor in Figure 5 and compare to same region in Figure 3

Looking for putative break points in the selected region.

It may be possible to simply identify the the break point at this point, especially for large homozygous deletions, but in many situations particularly for heterozygous break points they may not standout. Consequently, `AgileStructure` scans the displayed reads, looking for 250 bp regions in which multiple read alignments prematurely terminate and the remaining soft clipped sequences all maps to the same secondary location. These regions are then noted and entered in to the lower drop down list box (Figure 6).

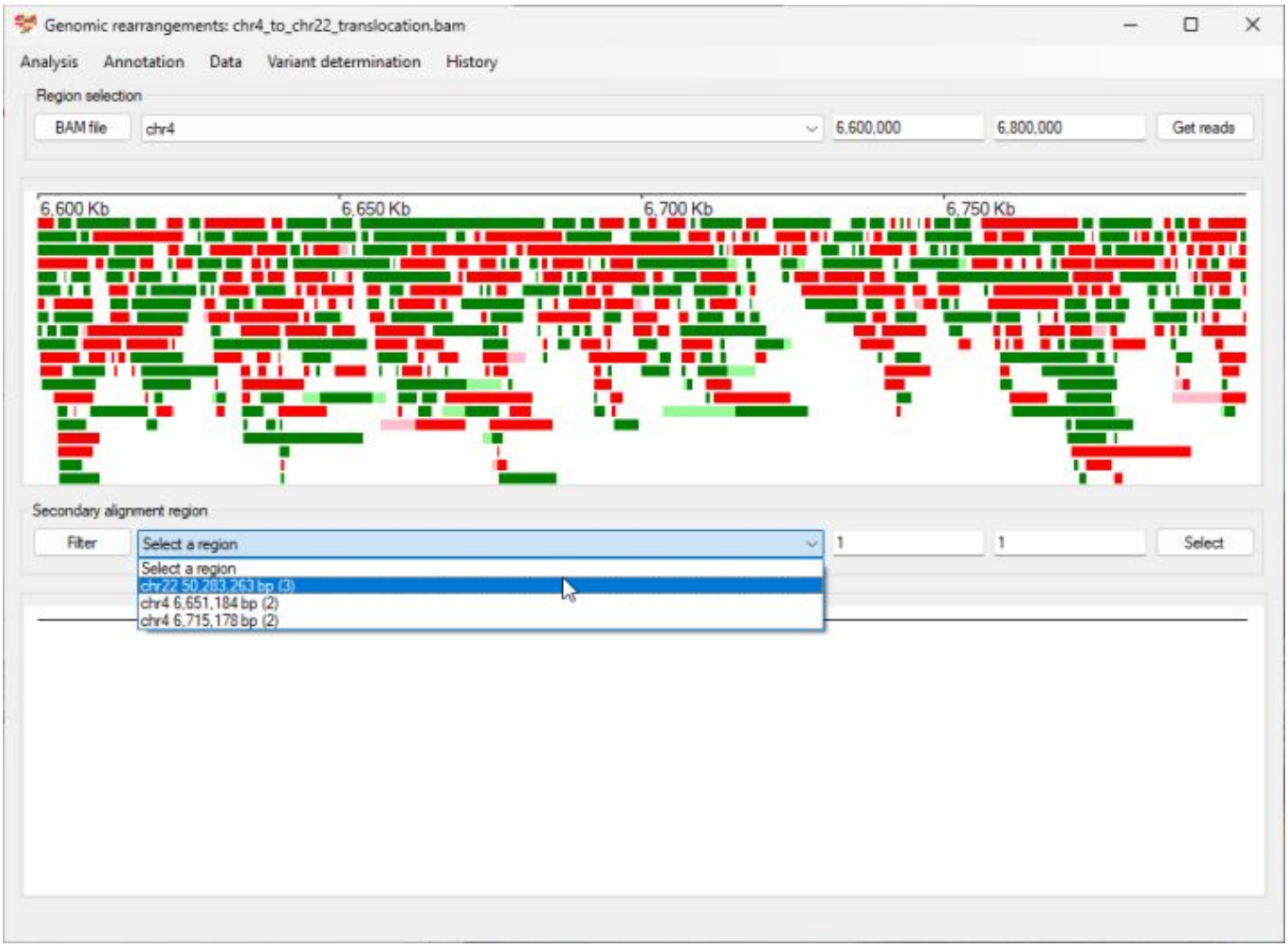

Figure 6

For extended regions and/or alignments with a high read depth, this list may contain a large number of entries. To filter these regions, press the `Filter` button to the left of the lower drop down list box. This will open the `Filter possible break points` window (Figure 7a and 7b). The upper drop down list box allows the break points to be filtered by the chromosome that the secondary alignments are mapped too (Figure 7a), while the lower number select box will filter the results by the number of reads linked to each putative break point.

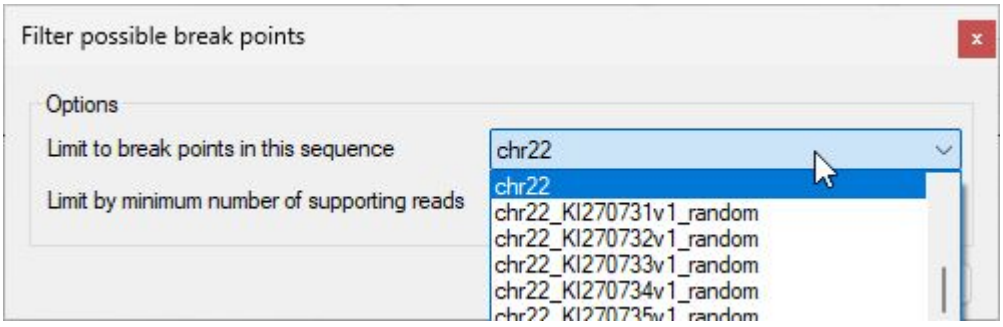

Figure 7a

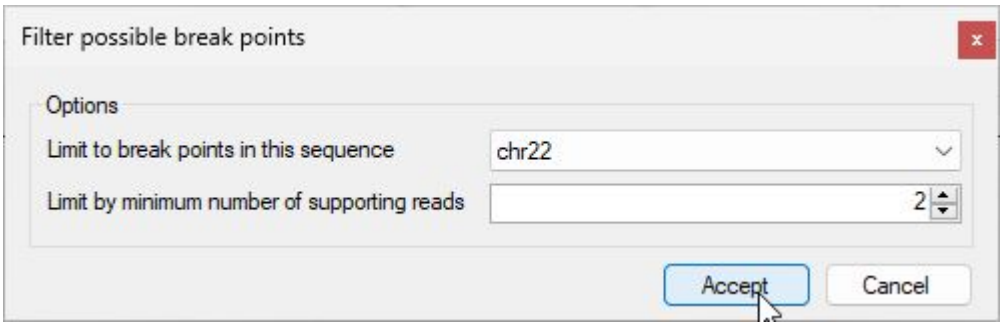

Figure 7b

Pressing the `Accept` button will remove all break points that do not match the criteria (Figure 8), while pressing the `Cancel` will remove all filtering.

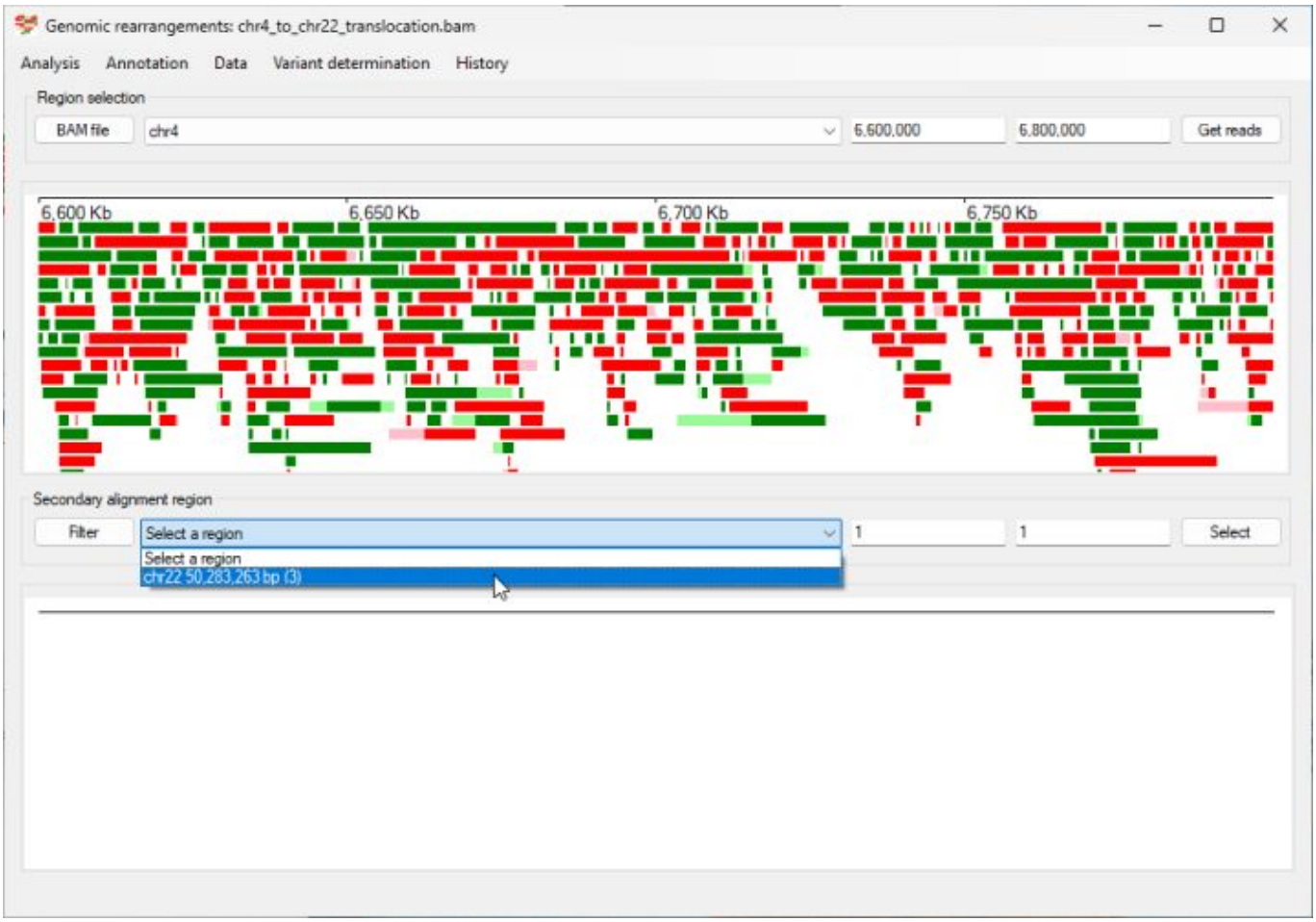

Figure 8

Selecting a break point from this list will cause the reads with soft clipped sequences mapped to the second break point's flanking regions to be displayed in the lower panel (Figure 9). As before, reads are drawn in green or red for those mapping to the forward and reverse strands, with aligned sequences darker than the unaligned soft clipped sequences. It is important to note that only reads that are present in the upper image are shown in the lower image and that sequences that were aligned to the reference sequence in the upper image will be unaligned, soft clipped sequences in the lower image.

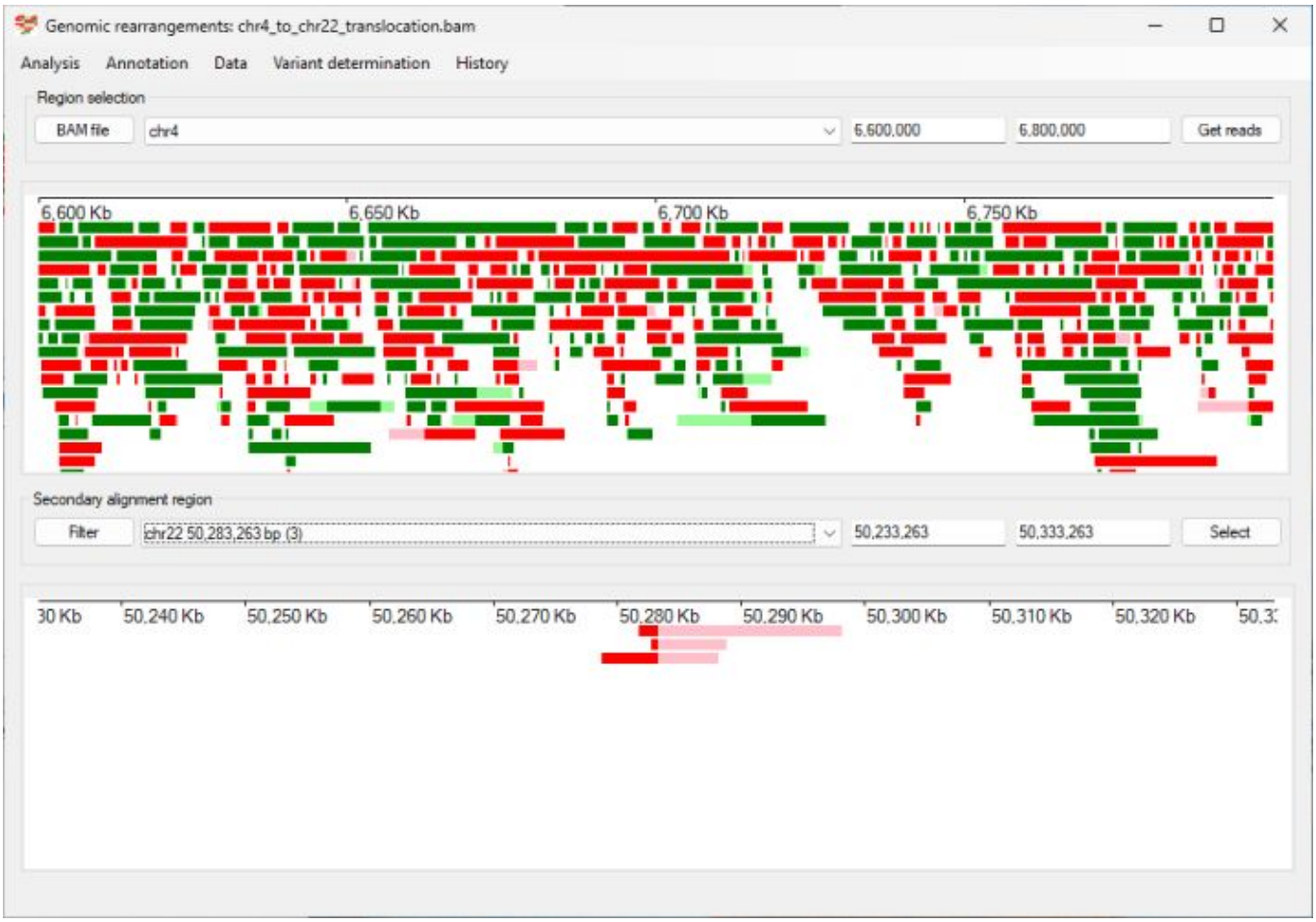

Figure 9

Viewing read alignment information

Selecting the `Data` > `View read data` (Figure 10a) will cause a resizable window to appear that consists solely of a text area. If the mouse cursor is held over a read, its underlying data will be written to the text area (Figure 10b). For a sequence to be shown, the cursor has to hover over the read for a little while. This makes it possible to select a read and then quickly move the cursor to the new window and copy the data to paste in a document etc.

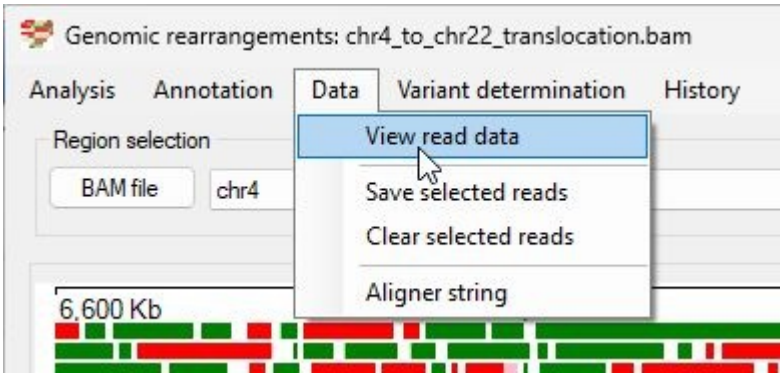

Figure 10a

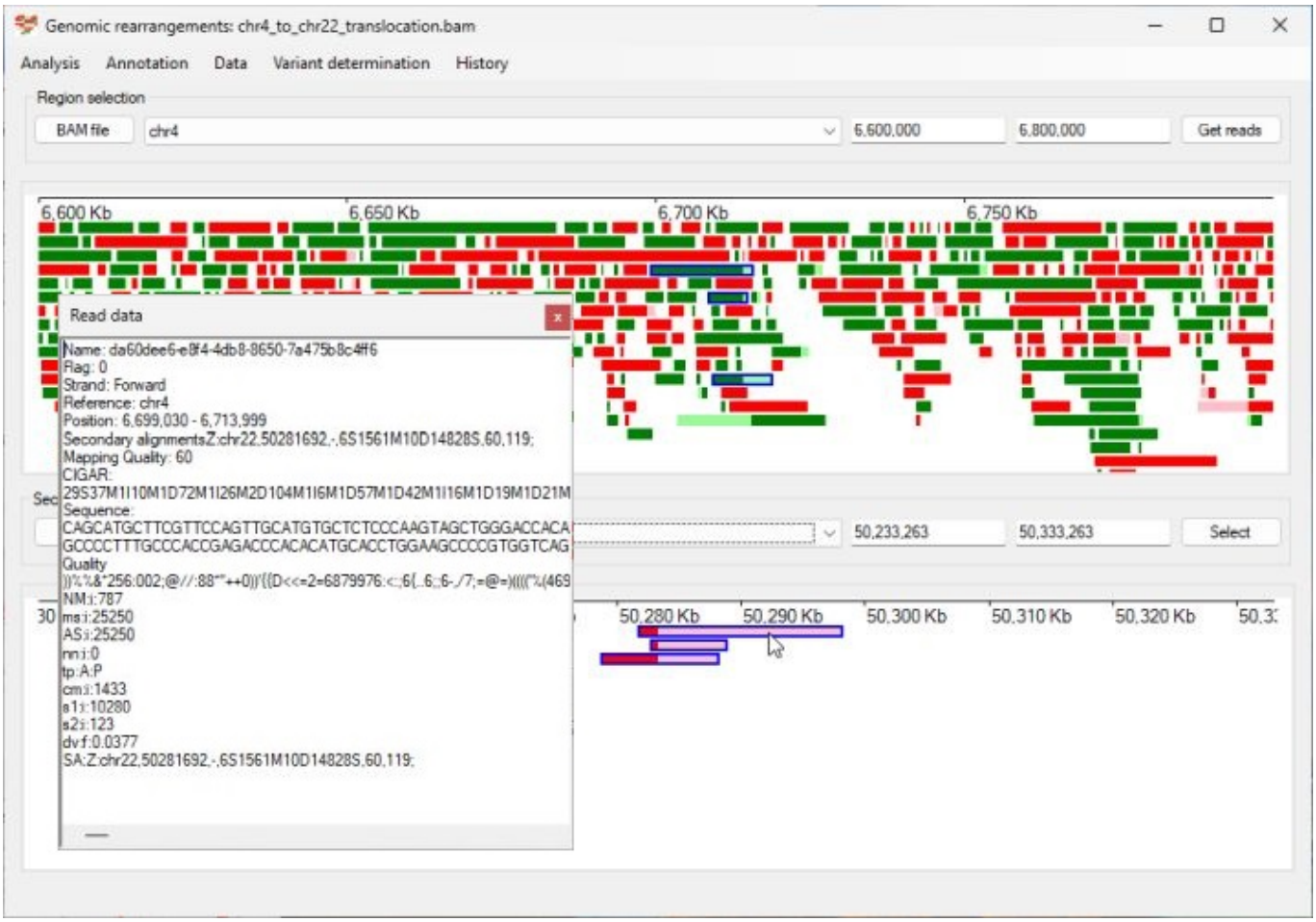

Figure 10b

Since the read, quality score string and CIGAR string can be several thousand characters long, the text area doesn't word wrap text (unless it's very long) and so if you want to read the end of a CIGAR string you must use the horizontal scroll bar.
As well as the sequence and quality string, this information contains the primary and secondary alignment location's, as well all the tags added by the aligner. The format of the tag data can be aligner specific with the aligner's documentation giving a full description of each tag's meaning.

Selecting reads linked to a break point

When the upper image contains a large number of reads, it may not be possible to identify the reads associated with the selected break point, however clicking on a read in either image will cause it to be selected and drawn with a blue boarder. Clicking on all the reads linked to a break point in the lower image will help to identify the location of the break point in the upper image (Figure 11)

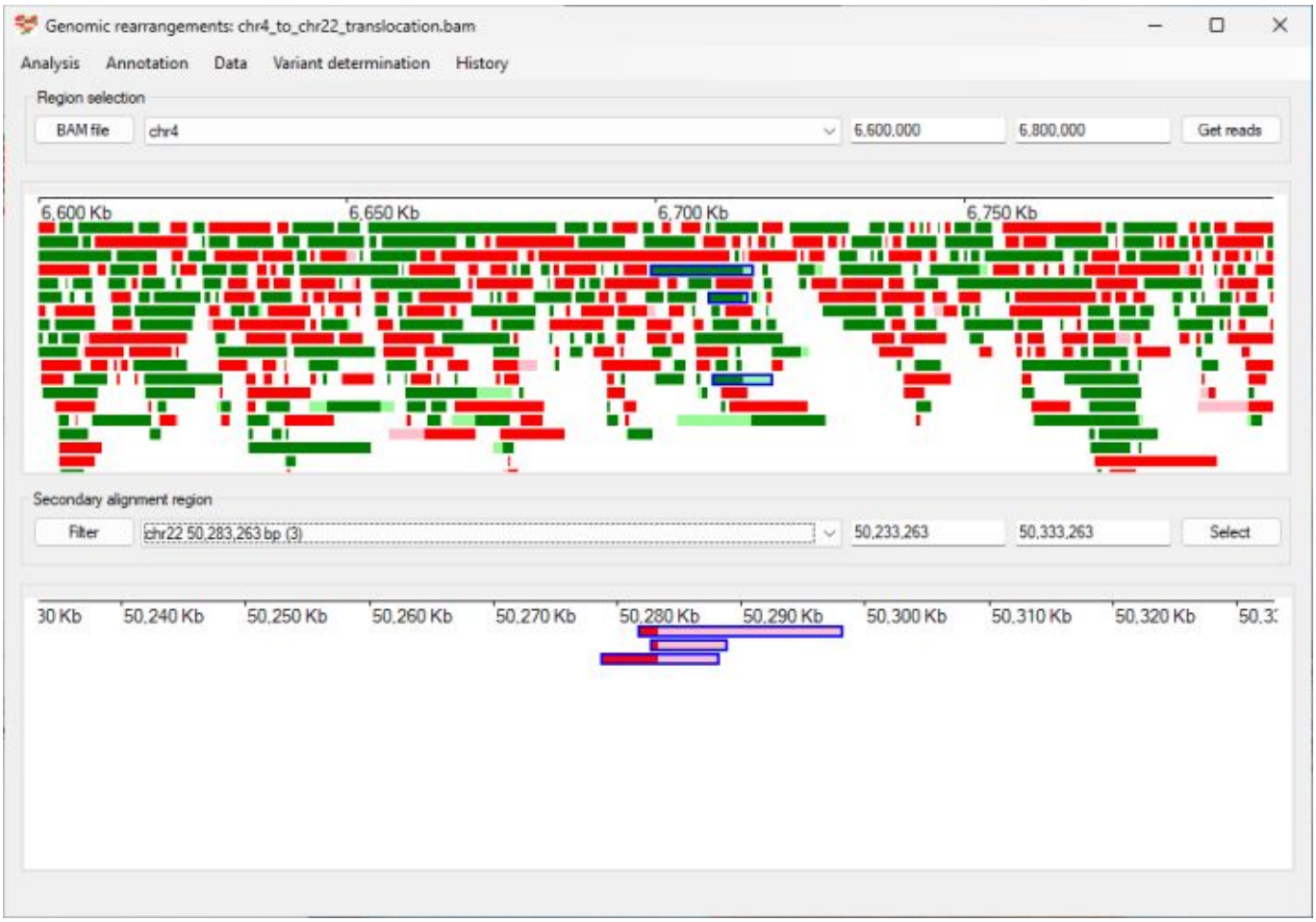

Figure 11

If you click on a selected read, it will be deselected, while selecting the `Data` > `Clear selected reads` option will deselect all selected reads (Figure 12). Finally, if new data is imported from the bam file (i.e. the `Get reads` button is pressed) the selection will be cleared.

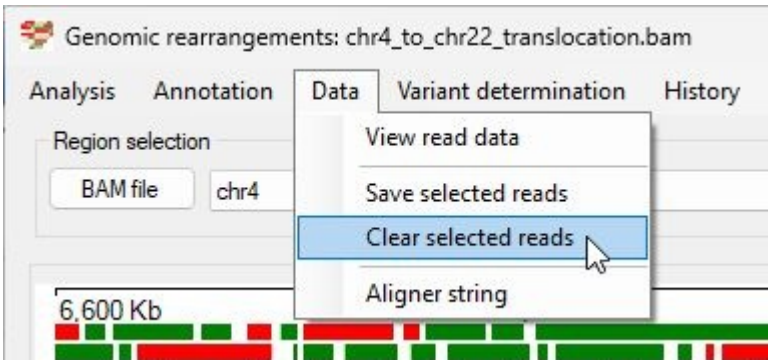

Figure 12

Saving alignment information for selected reads

Rather than manually saving the data for a series of read alignments, its possible to save the data of selected reads to a text file using the `Data` > `Save selected reads` (Figure 13).

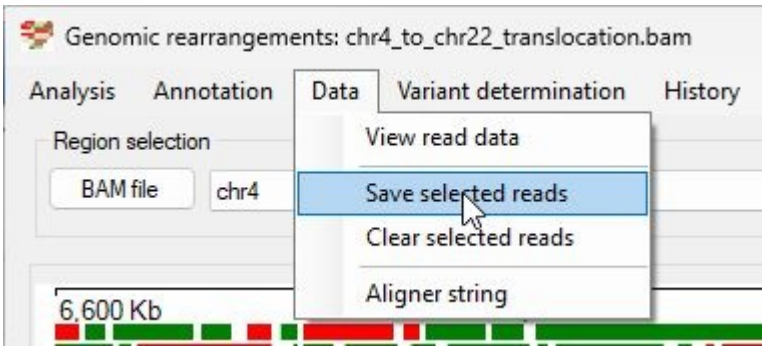

Figure 13

Annotating break points using soft clipped data

Once the reads spanning a break point have been selected, is is possible to get `AgileStructure` to attempt to identify the type of variant: deletion, duplication, insertion, inversion or translocation. To identify what type of mutation the break point represents, select the `Variant determination` > `Use soft clip data` > `Variant type` menu option (Figure 14a). `AgileStructure` will then scan the orientation of the primary and secondary alignments of the selected reads to determine what type of mutation it is. This is reported in a message box with the possible answers of "Deletion", "Insertion", "Inversion", "Duplication" or "Translocation" as well as messages indicating any error processing the data or user data selection issues (Figure 14b).
For this feature to work a region must be selected in the lower panel.

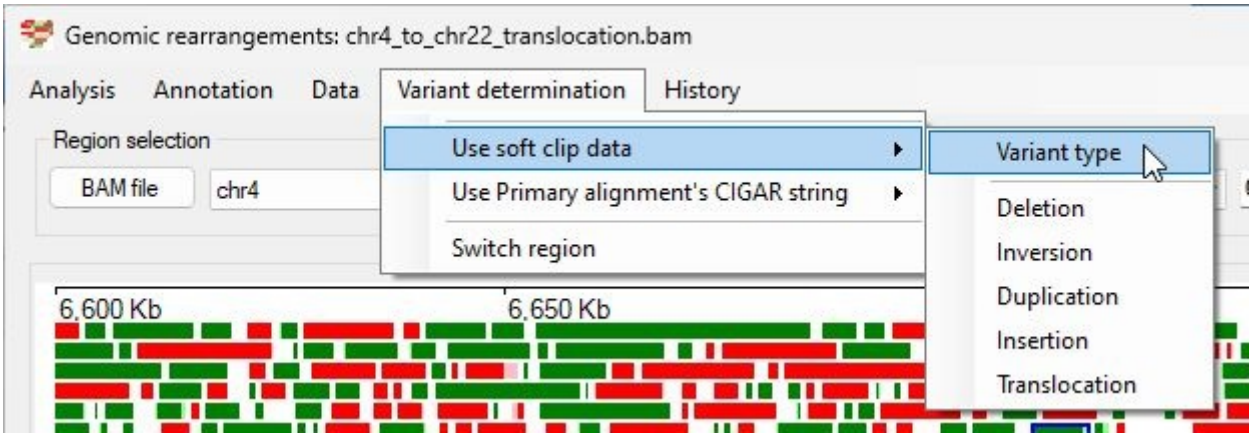

Figure 14a

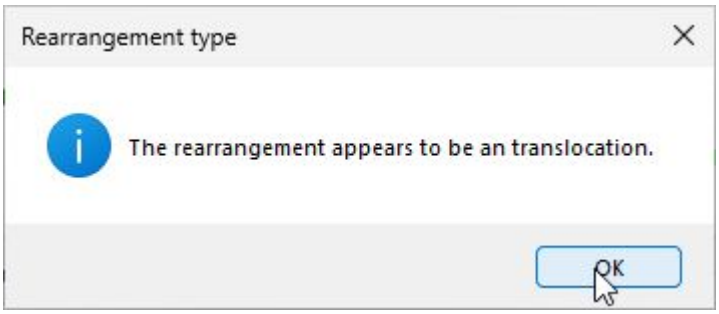

Figure 14b

Once the variant type is determined it is then possible to annotate the break point by selecting the appropriate option. The links below show examples of analysing each type of mutation

Deletion

A worked example is here.

Duplication

A worked example is here.

Insertion

A worked example is here.

Inversion

A worked example is here.

Translocation

A worked example is here.

Identifying Indels using the primary alignments CIGAR string

`AgileStructure` is primarily designed to identify chromosomal break points by looking for sets of reads whose alignment is broken in two, such that their primary alignment aligns at one location and their secondary alignments are all located to a more distant common region possibly on a different chromosome. However, it is also able to identify insertions and deletions that do not cause the alignment to be fragmented, but whose presence is noted in the primary alignment's CIGAR string.
Selecting the `Analysis` > `Look for indels within a read` menu option (Figure 15) causes the reads to be redrawn with deletions shown as a horizontal black line linking two blocks of aligned sequences while an insertion is shown as a vertical line projecting above and below the aligned sequence. Since ONT data contains numerous short indels, only insertions/deletions longer than 10 bp are shown.

Figure 15

When redrawn using the the CIGAR string to identify insertions and deletions their presence becomes apparent. For example in Figure 16 the large deletion spanning 1,495,000 bp to 1,534,000 bp of chromosome 1 and the insert at 1,586,000 bp (above the cursor) are easily identified.

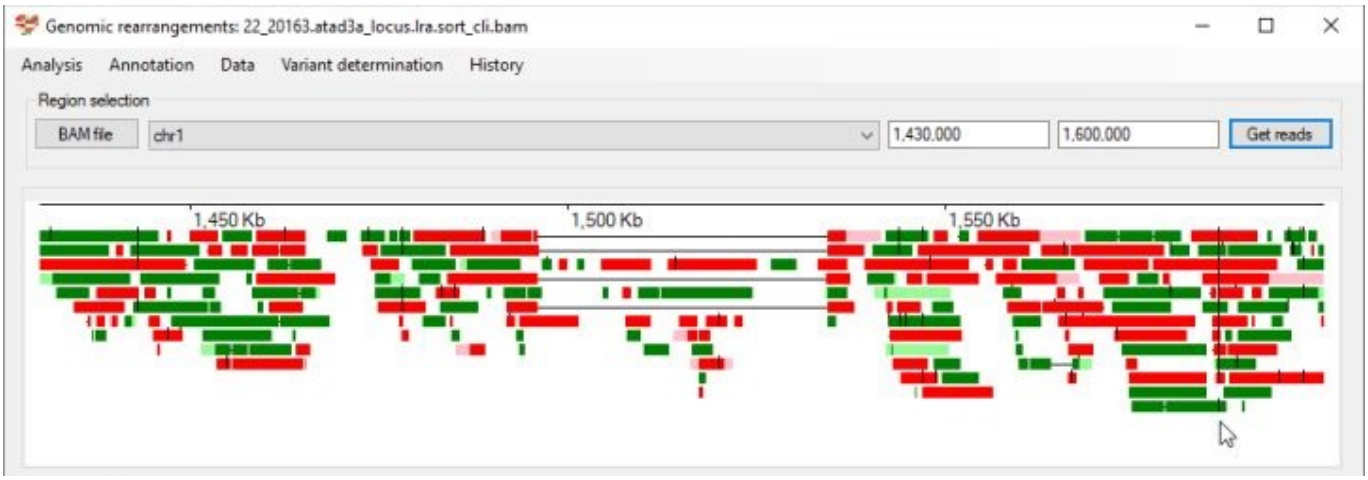

Figure 16

Important note

Since ONT data is very noisy the exact point of the break point may appear to vary by a number of base pairs between different reads, while artefactual indels may also be present in the reads. Consequently `AgileStructure` scans the beginning and ends of the indels, sorts them by position and then reports the median values in the reported variant. Using the median value rather than the average reduces the chance an artifactual indel unduly influencing the annotation, but it is important to ensure that the individual indels are checked to make sure a 2nd, possibly artifactual, indel is not somehow disrupting the annotation.

Identifying insertions using the primary alignments CIGAR string

To annotate an insert, select the reads the containing variant of interest and select the `Variant determination` > `Use primary alignment's CIGAR string` > `Insertion` menu option (Figure 17a). This will display a message box, listing any insertions over 10 bps followed by the read's name and the annotation of the variant. The analysis may require reads with multiple insertions to be deselected.

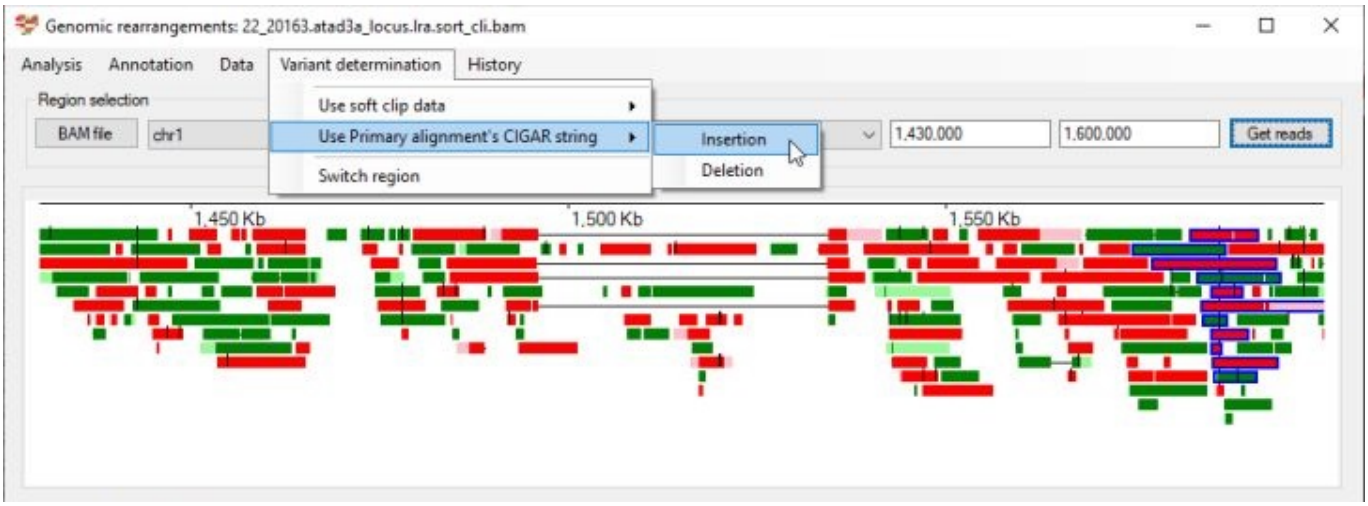

Figure 17

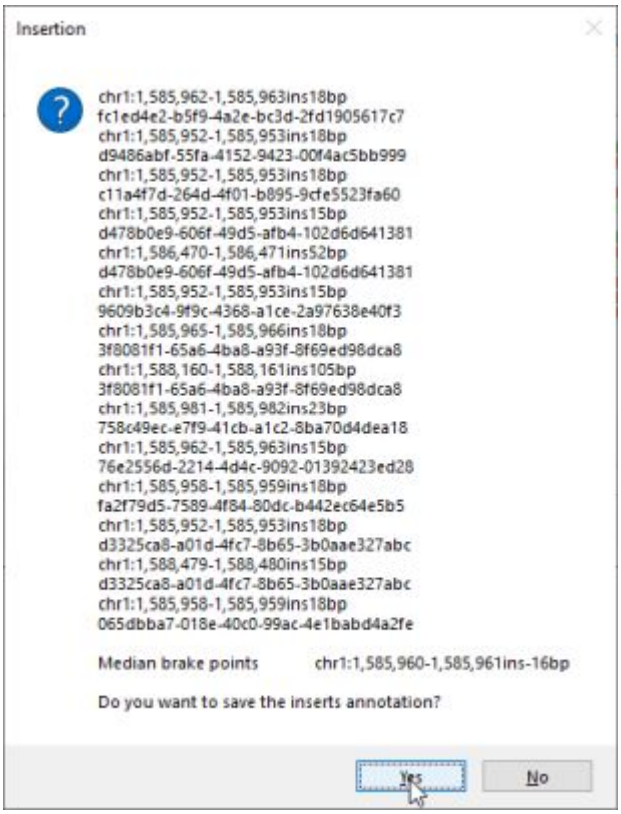

Figure 17b: The order of the reads in the message box is ordered by chromosomal position and so is not in the same order as the reads are displayed

Identifying deletions using the primary alignments CIGAR string

To annotate a deletion, select the reads containing the variant of interest and then select the `Variant determination` > `Use primary alignment's CIGAR string` > `Deletion` menu option (Figure 18a). This will open a message box, listing the deletions over 10 bps followed by the reads and finally the variant's annotation (Figure 18b).

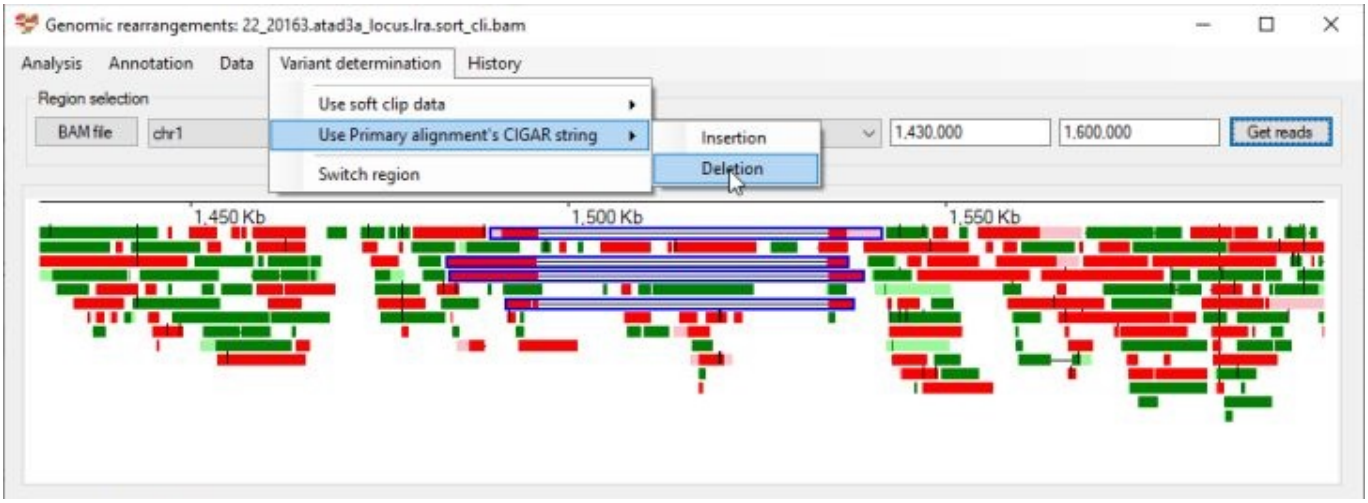

Figure 18

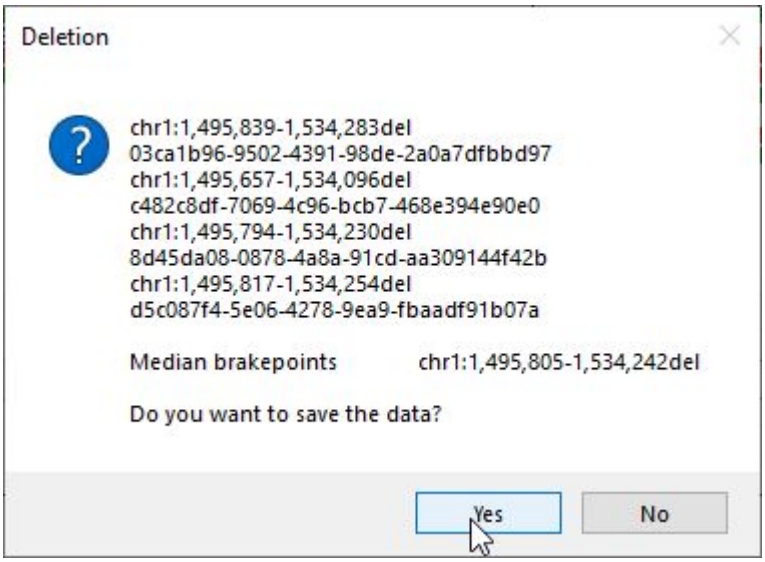

Figure 18b: The order of the reads in the message box is ordered by chromosomal position and so is not in the same order as the reads are displayed

Navigating the read data

Changing the region by typing the coordinates

As previously mentioned, `AgileStructure` displays the primary and secondary alignments in two panels, above each display are two text areas where the start and end points of the displayed data can be changed. Since the primary read data is retrieved from the bam file which can be slow, changes to the primary alignment region are only made when the `Get reads` button is pressed. However, changes to the coordinates of secondary alignment image are displayed instantly.

Moving the region to the left and right with the left and right arrow keys

Rather than typing in new locations in to the text areas, its possible to move the region to the left or right by clicking on one of the text areas so that the text area becomes active (i.e. you could edit the value by typing) and then pressing the `Ctrl` + `left arrow` or `Ctrl` + `right arrow` keys. This will shift the region in the appropriate direction to an adjacent, none overlapping region of the same length as the original.

Changing the width of the region with the Up and Down arrow keys

In a similar manner to moving the region to the left or right, it's possible to double or half the width of the region: Activate a text area and then pressing the `Ctrl` + `Up arrow` or `Ctrl` + `Down arrow` keys. While the size of the region changes, it remains centered on the same point in the reference sequence.

Changing the region by selecting a region with the mouse

The mouse can be used to select a sub-region of the current display in either panel by moving the cursor to the desired start point and then moving the mouse to the end point while holding the right mouse button down (Figure 19a), when the mouse button is released the display is redrawn (Figure 19b). This allows a feature to be more closely observed, for instance in Figure 19a, four reads appear to have an insert in the same location, zooming to the region (Figure 19b) suggests that they may not be genuine as the locations differ. However, by selecting the `Variant determination` > `Use primary alignment's CIGAR string` > `Insertion` menu option, it appears all the reads have a 134 to 135 bp insertion suggesting its position is

inaccurately located possibly due to variable alignments as a consequence of sequencing errors and/or alignment to low complexity sequence (Figure 19c).

Figure 19a:

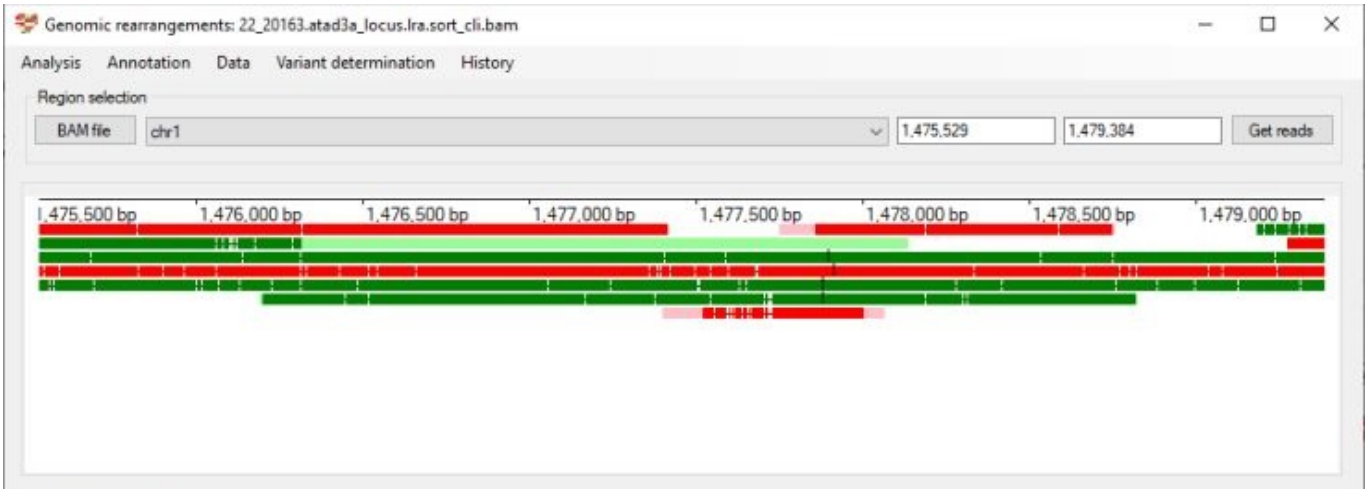

Figure 19b

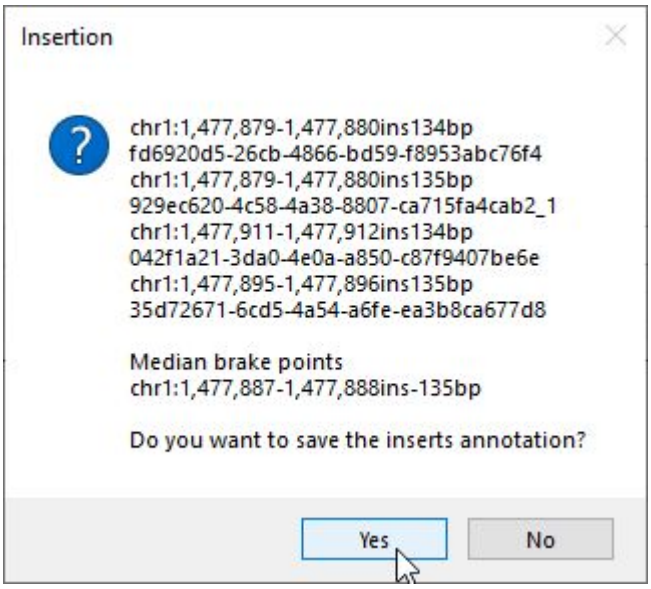

Figure 19c

Changing the regions using the History menu options

As each change in either the primary and secondary display coordinates is made, the old positions are saved, allowing the views to be recreated by selecting the appropriate coordinates from the lists in `History` > `Primary alignments` or `History` > `Secondary alignments` menu options (Figure 20).

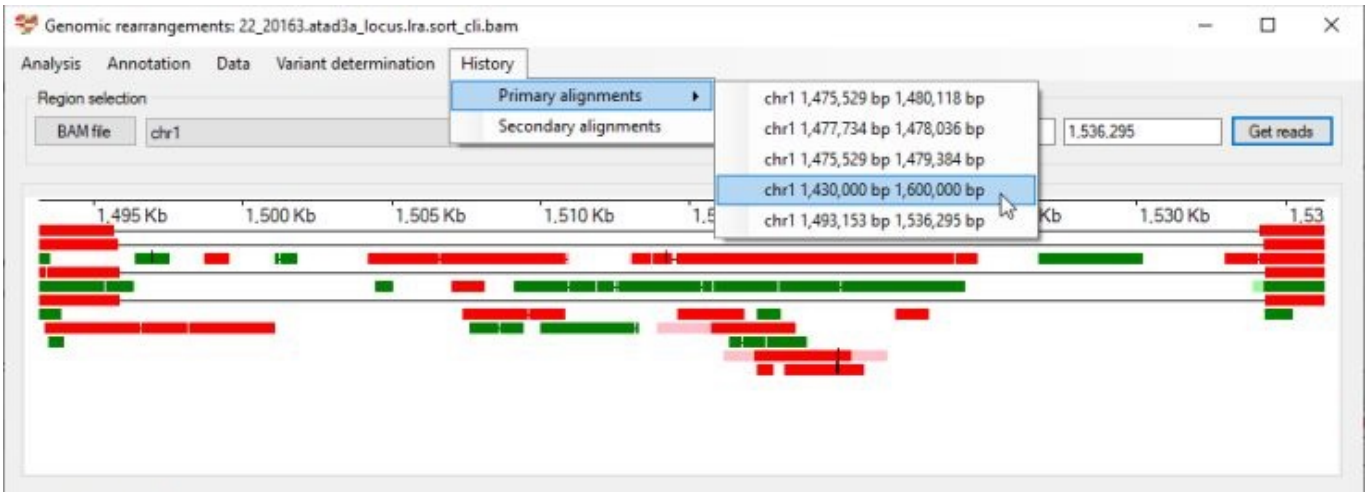

Figure 20

Selecting an area that contains a specific gene

The Displaying gene positions section explains how to import gene locations, once imported, it is possible to navigate to a region that contains a specific gene by selecting the `Annotation` > `Gene coordinates` menu option (Figure 21a).

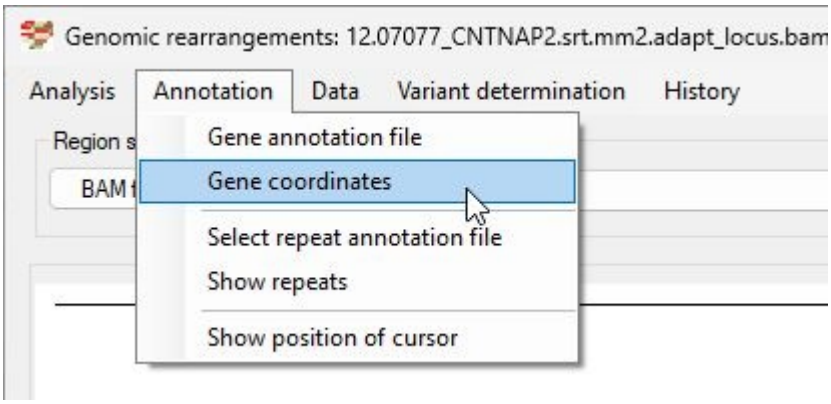

Figure 21a

This will open the `Gene coordinates` window which consists of two text areas, type the gene symbol for the gene of interest in the upper text area (Figure 21b).

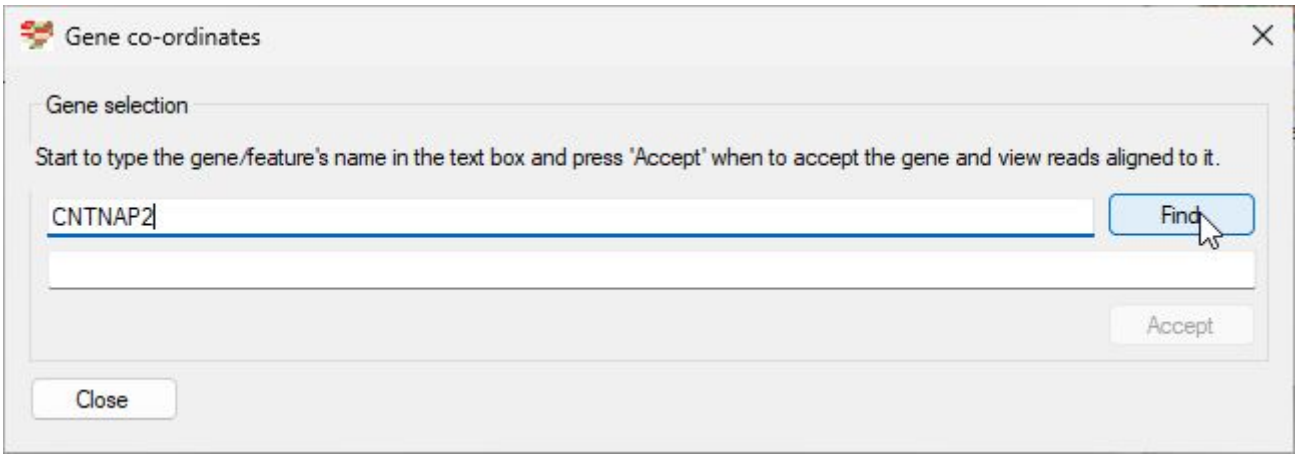

Figure 21b

Press the `Find` button and if the gene symbol is present in the imported gene coordinate data, it's coordinates will be displayed in the lower text area (Figure 21c).

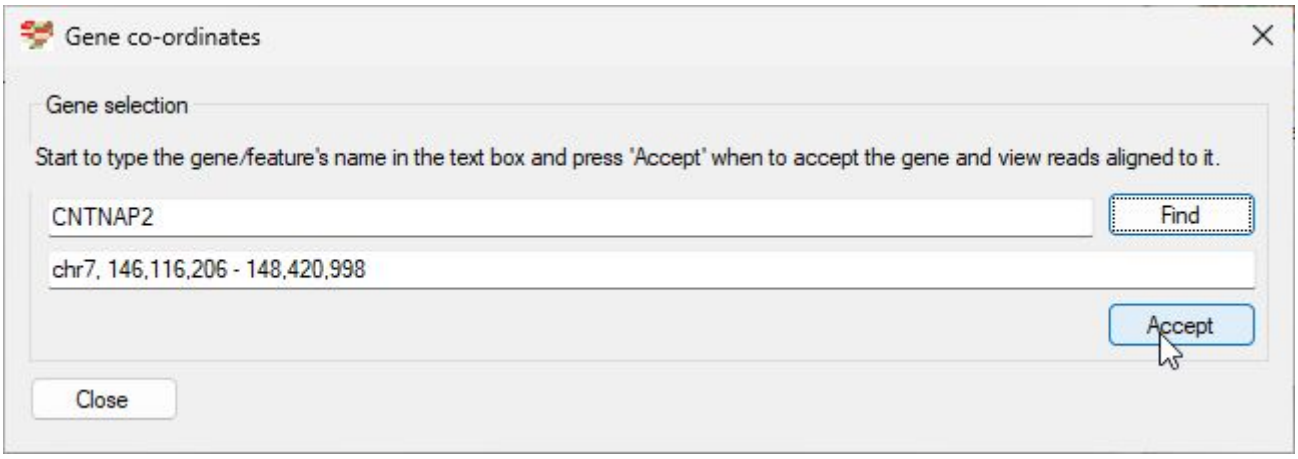

Figure 21c

Pressing `Accept` will reset the the coordinates in `AgileStructure` main window, and pressing the `Get reads` button will update the Primary alignment window (Figure 21d).

Figure 21d

Viewing data with reference to genomic features

It is possible to view the read data with reference to the genes and repeats around the break point. The chromosomal locations of the genes and repeats can be downloaded from the UCSC Genome Browser as described here.

Displaying gene positions

Gene coordinate data can be imported by selecting the `Annotation` > `Gene annotation file` menu option (Figure 22).

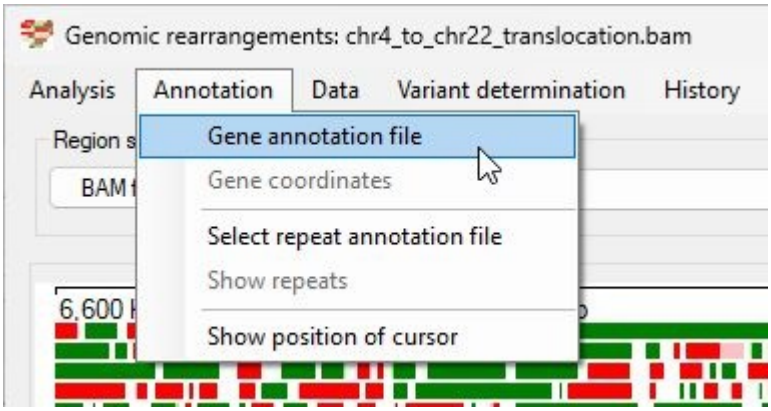

Figure 22

Genes are displayed as black rectangles with their exons drawn as either green (gene on forward strand) or yellow (gene on reverse strand) at the bottom of the displays (Figure 23).

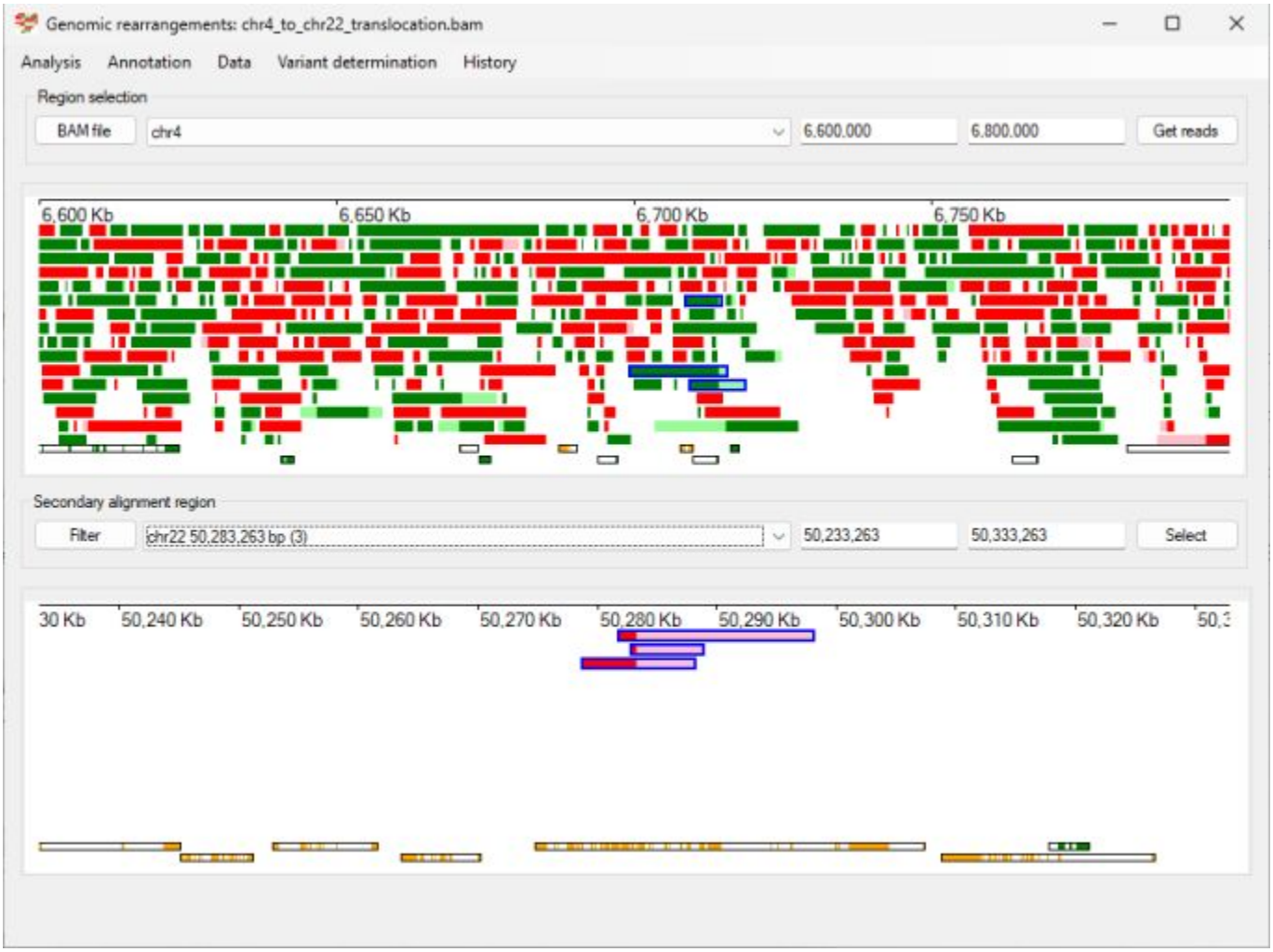

Figure 23

Clicking on a gene will cause is name to be displayed to the top left of the appropriate display, for instance in Figure 24 the genes near the break point (Primary alignment display: LOC105378240 and secondary alignment display: PLXNB2) have been selected.

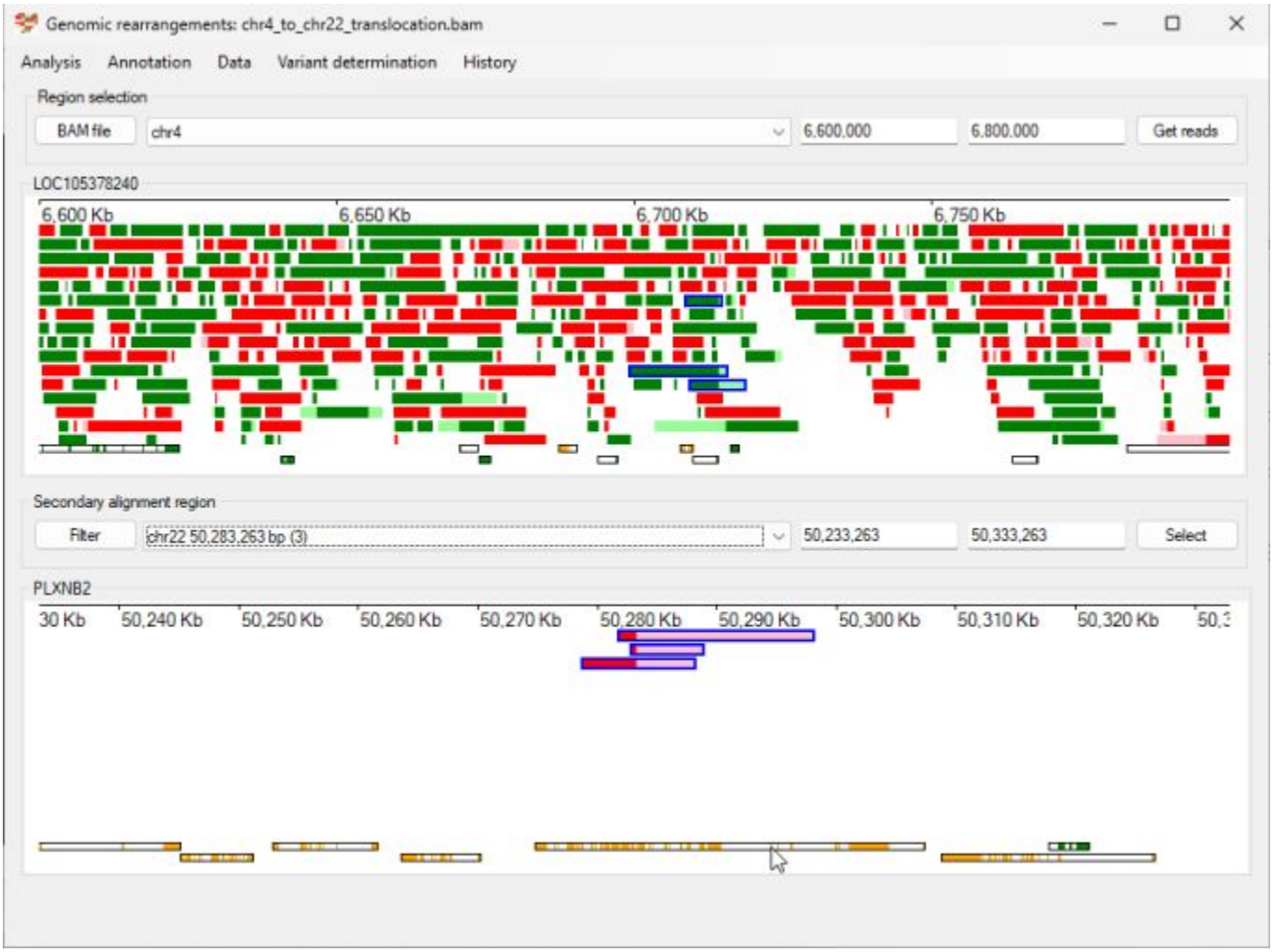

Figure 24

Displaying repeat positions

Repeat coordinates are imported by selecting the `Annotation` > `Select repeat annotation file` option (Figure 25). Unlike the gene positions, repeats are only drawn when the `Annotation` > `Show repeats` option is selected (Figure J). This is due to the large number of repeats requiring an excessive amount of memory to store and then slow to draw across large regions. Consequently, `AgileStructure` will only retain the repeat file's filename and reads the file each time it is required to draw them.

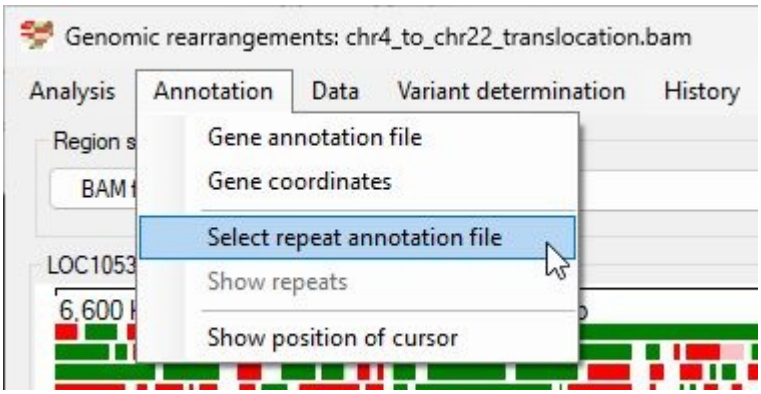

Figure 25

The repeats are drawn as black rectangles filled in pale blue (forward strand) or pale yellow (reverse strand) across a single row at the very bottom of the displays. As with genes, clicking on a repeat will cause it's name, class and family to be displayed at the top left of the display. For example in Figure 26, the repeats close to the break point (Primary alignment: AluSz, SINE, Alu and Secondary alignment: (CCCACC)n, Simple repeat, Simple repeat) have been selected.

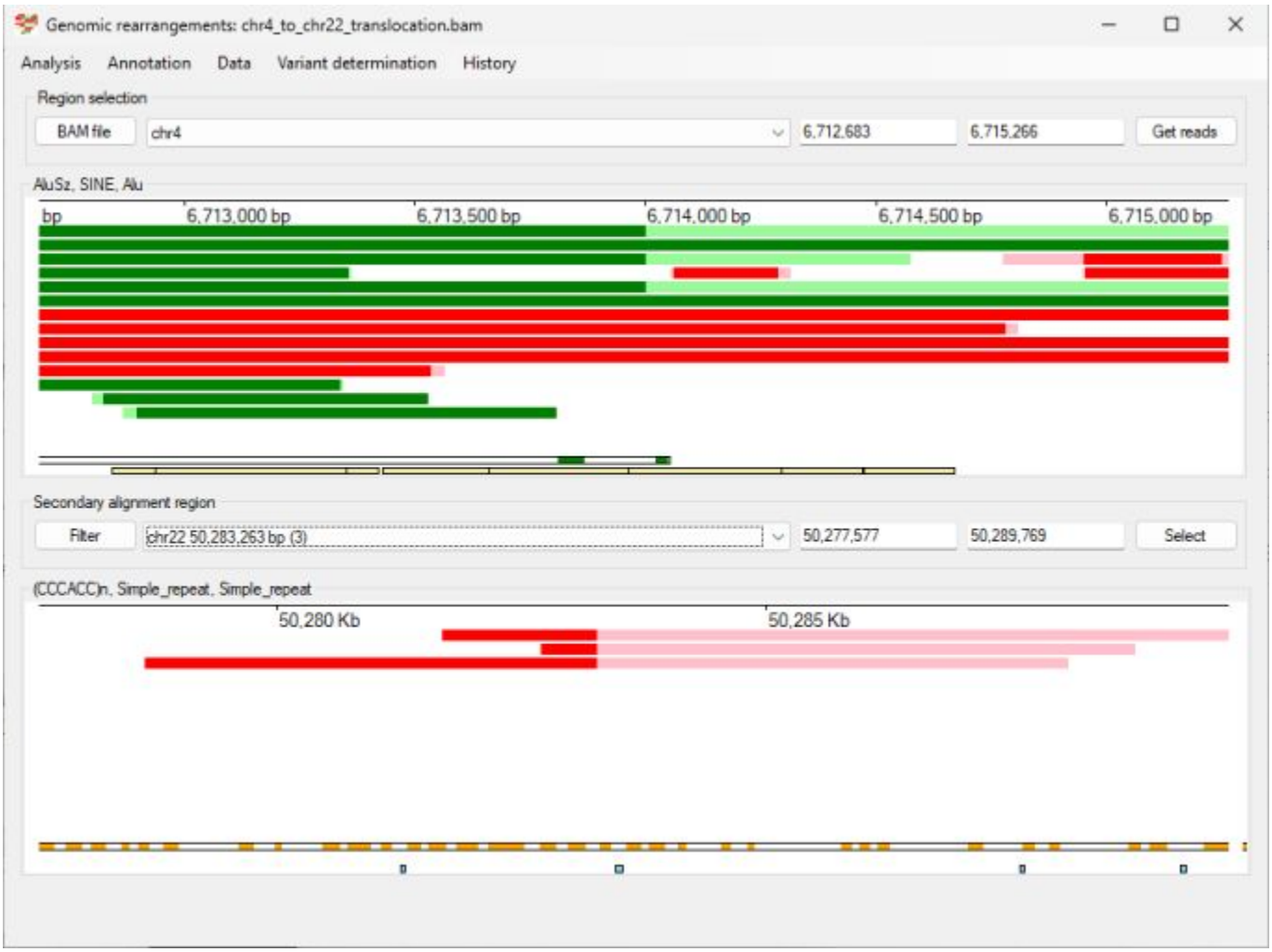

Figure 26

Miscellaneous functions

Cursor location

The `Annotation` > `Show position of cursor` menu option displays the genomic coordinates of cursor's position (Figure 27b). It should be remembered that this is inaccurate as a region 1 Mb wide, drawn on an image 860 pixels wide will have 1,162.8 bps mapped to each pixel consequently, this is only shown as an aid to understanding the region.

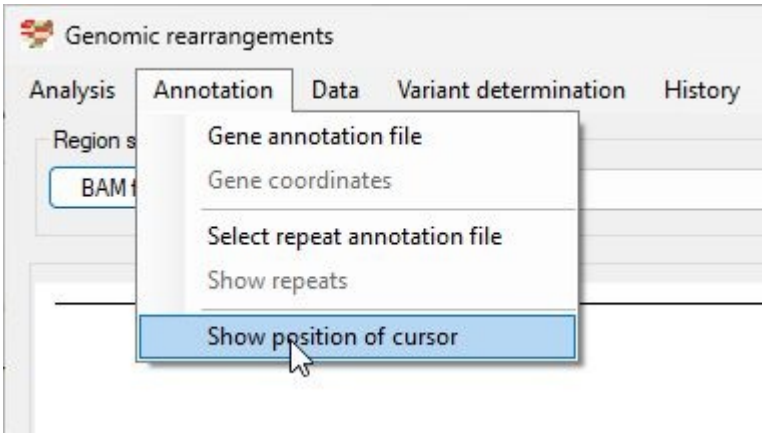

Figure 27a

Figure 27b

Aligner string

Typically, the aligner used to map the reads to the reference genome will include the command line arguments used in the alignment in the BAM file's header section. This information can be viewed by selecting the `Data` > `Aligner string` menu option (Figure 28a and 28b). This may prove useful when for instance, you need to be certain which reference genome was used in the alignment.

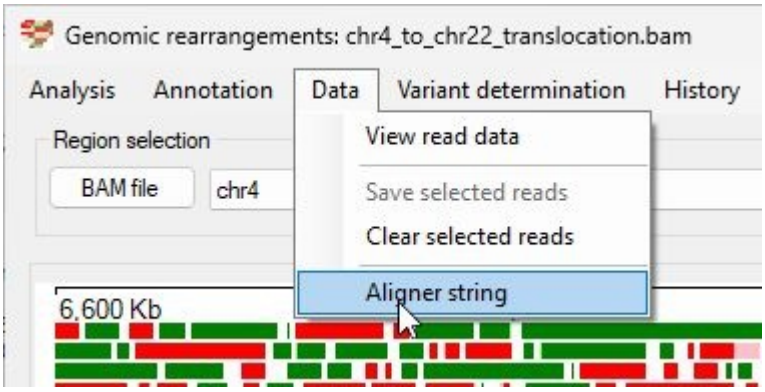

Figure 28a

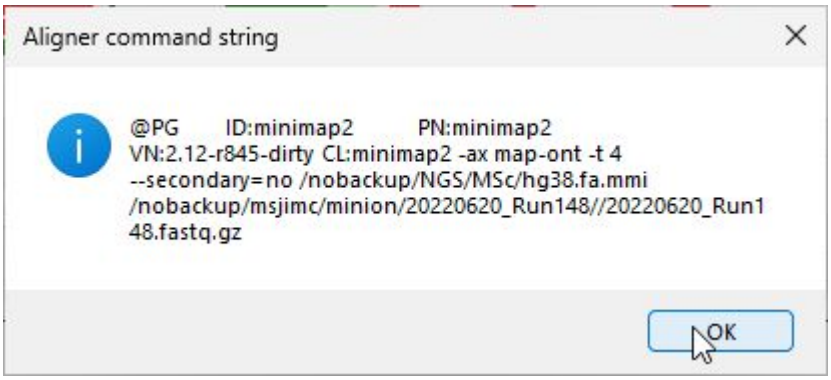

Figure 28b: The command string identifies the data in the '20220620_Run148.fastq.gz' file was aligned to the 'hg38.fa.mmi' minimap2 index (human genome : hg38).