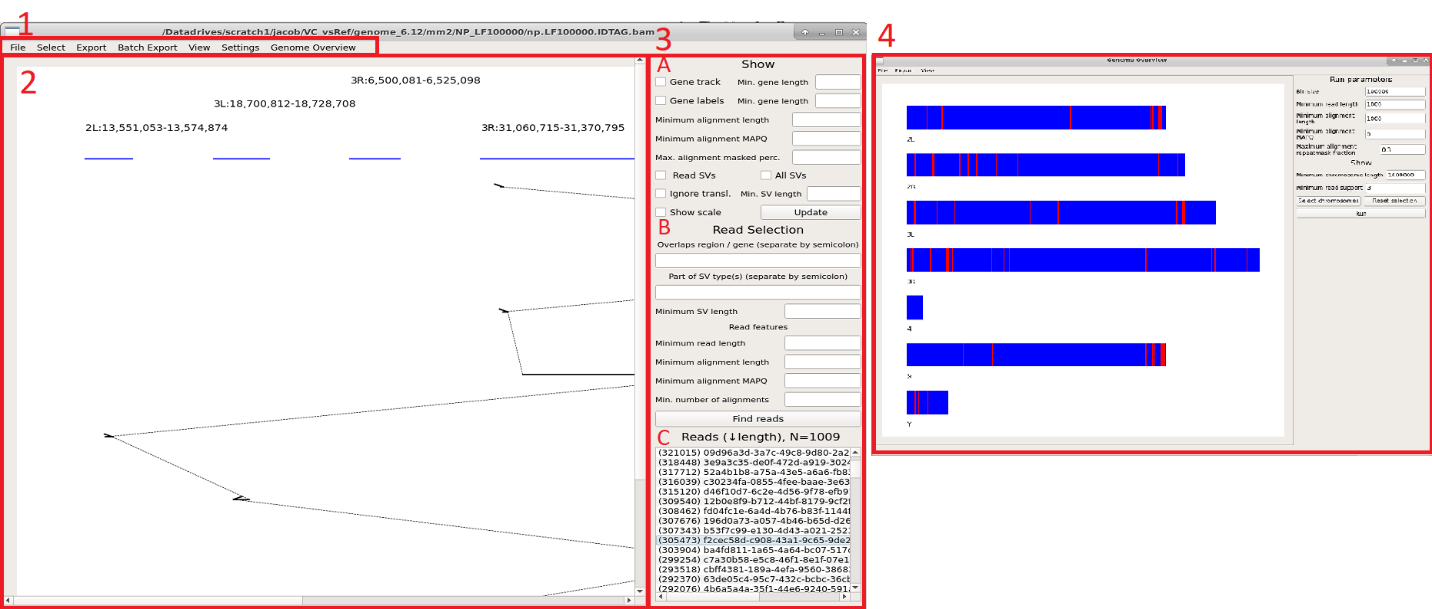
# Reference Manual

This is the manual for software X.

  
Figure 1. Software graphical user interface. The different panels of the user interface is indicated with red boxes and red text labels. The document is structured according to the sections.

## 1: Menu bar

### 1A: File

#### Import BAM

Prompts the user to import a read alignment file in BAM/SAM format. This file is required to use the software. If you have issues with memory upon import, you can specify the minimum length of reads to parse: See document section 1E (BAM import).

#### Import GTF

Prompts the user to import an annotation file in GTF/GFF format. This file is optional but improves the software experience: searchable regions via annotation name and visualization of annotation. The software parses data from columns 1, 3, 4, 5 and specific tags from the 9th column. For more details, including how to modify which keys to parse, see document section 1E (GTF/GFF import).

#### Import repeats

Prompts the user to import an annotation file in .out format, produced by the RepeatMasker software. This file is optional but improves the software experience: filters alignments by overlap in repeat-annotation regions. The software parses data from columns 5, 6 and 7. If your annotation is not in the RepeatMasker .out format, you can convert it to a file compatible with this software: Compile a tab-separated 15-column file with chromosome in column 5, reference start in column 6 and reference end in column 7.

#### Import VCF

Prompts the user to import a variant-call file in VCF format. This file is optional. The software can parse any VCF file which comply to the VCF specification. Since many variant-callers include custom data in the last column, the software was designed to parse Sniffles output. The data obtained from the last column include the second breakpoint reference name and position, variant type, the name of supportive reads and variant length. The identifier for these keys can be modified. For more information, see document section 1E (VCF import).

#### Exit

Terminates the software.

### 1B: Select

#### Input read name(s)

Populate the read selection with input read names. Multiple read-names can be pasted, separated by newline or comma.

#### Import read list

Prompts the user to import a file of read names to populate the read selection. Read-names must be separated by newline.

#### Clear selection

Clears the current selection and attempts to populate the default selection.

### 1C: Export

#### Current read

Opens a dialog holding the current highlighted read (left-click in read selection, see document section 3C).

#### Read highlight

Opens a dialog that holds all highlighted reads (left- and right-click in read selection, see document section 3C).

#### Read selection

Prompts the user to enter an output file to write all read names contained in read selection.

#### Regions

Opens a dialog with coordinates of currently plotted reference regions.

#### Alignments

Opens a dialog with alignment information in three columns of the currently drawn read: The first column holds the alignment coordinates on the reference, the second holds the alignment strand relative to the reference and the third holds the alignment coordinates on the read.

#### Image

Prompts the user to enter an output file to output the current plot. The output file name should include file format. Supported formats include pdf, png and svg. If no format is specified, the software will default to png. If the output file already exists, the user will be prompted to confirm overwrite.

### 1D: View

#### Reset zoom

Restores the zoom.

#### Zoom in

Zooms in on the plot, on **both** axes on or **X-axis** or **Y-axis** only. Zoom-in can also be invoked via “+” keyboard button, with modifiers shift and alt to zoom on X- and Y-axis individually.

#### Zoom out

Zooms out on the plot, on **both** axes on or **X-axis** or **Y-axis** only. Zoom-out can also be invoked via “-” keyboard button, with modifiers shift and alt to zoom on X- and Y-axis individually.

### 1E: Settings

To increase the likelihood that this software is compatible with hardware (memory) and formats (which keys to parse), there are some settings which can be set to determine how files are parsed.

#### Show

The distance to group alignments into the same displayed region can be adjusted in **Alignment chain distance**. The drawn region margin (number of base pairs to each side outside alignment hit) is controlled by **Region margin**.

#### BAM import

The minimum read length can be set to discard import of short read alignments. This option may be useful if you have issues with memory due to importing too many reads.

#### VCF import

The software was designed to parse standard VCF files and the VCF-file with custom data in the INFO-field produced by Sniffles. The keys parsed from the INFO-field can be modified, in case your VCF file use different keys (different software may label these keys differently). The information parsed from the INFO-field by the software is the second breakpoint reference name and position, variant type, the names of supportive reads and variant length.

#### GTF/GFF import

The data parsed from GTF/GFF annotation files are genes (for visualization) and additional tags (for read selection) in the attribute field. The delimiter of the attribute field can be modified for both formats and the tag holding the gene name. Since annotation files may contain information which is superfluous when only interested in the gene coordinates (multiple overlapping entries related to exons or coding sequence), the software can exclude parsing these lines to save computation time and memory. The filtered terms may be modified by the user to select specific information or further restrict the import.

### Genome overview

The Genome overview module is launched into a new window via the menu. For more information about the module, see below section.

## 2: Plot area

This area holds the canvas for paintings. Use plus (+) and negative (-) keys to zoom in the plot. To zoom on specific axis, use shift (X-axis) and alt (Y-axis) modifiers. Navigate via the scroll-bars or drags with the cursor. The minimal drawing shows read alignments (black horizontal bars, connected by a dotted line) in-order by alignment coordinates on the read. The alignment reference regions are shown as blue horizontal bars on the top with text label (formatted as “chromosome:start-end”). An arrow-head marker indicates the alignment strand relative to the reference. Red markings on the alignments show that this region overlaps on the read by one or more alignments.

Filtering of painted alignments can be controlled via settings on the right panel (for more information, see document section 3A). The right panel include controls to show gene tracks and structural variants, provided that this information is available via imported files.

Additional reads can be introduced to the plot (via right-clicking reads in the read selection list, for more details, see document section 3C) with a random color. The alignments of additional reads are only shown within the regions of the current read painting. This feature is useful to find if multiple reads are supportive of a rearrangement.

If the plot contains many elements, objects such as text labels may start to overlap. To alleviate overlapping elements, the user can zoom in to expand the plotting area. If needed, the user can re-draw the plot to re-position objects given the increased figure size.

## 3: Right panel

This panel consist of three sections, outlined below.

### 3A: Show

This section lets the user control what features to visualize into the plot area. The specified settings are applied upon drawing and a current plot can be re-drawn using the **Update** button located on the bottom-right of the section. Display of genes and associated labels are controlled via **Gene track** and **Gene label** checkboxes and are linked with the optional right-hand input fields to filter the display by length (**Min. gene length**). Requires gene annotation file to be imported. Alignments can be omitted from drawing based on length, MAPQ score, or overlap to repeat annotation (**Minimum alignment length, Minimum alignment MAPQ, Max. alignment masked perc.**). The cutoff for alignment repeat annotation is input as a percentage or fraction and requires a repeat-masking file to be imported. Structural variants can be shown by checking the boxes of **Read SVs** (shows variants associated with the selected read) and **All SVs**. Translocation variants can be omitted by checking the **Ignore transl.** box. Variants may be filtered by length specified into the **Min. SV length** input field. Finally, a scale indicator can be drawn controlled by the **Show scale** checkbox.

### 3B: Read Selection

The read selection enables the user to find reads to plot via the **Find Reads** button. Given that sequencing datasets most often are large and include noise in the form of technological artefacts and sample heterogeneity one major challenge for the user will be to filter interesting reads. Reads can be found in both generally and specifically. Into the top field a region (on format chr:start-end) or gene annotation can be specified (requires gene annotation file to be imported). Multiple requirements may be entered and separated by a semi-colon (e.g., “chr5:5000-100000;myFavoriteGene” will require reads to have alignments in the region 5000 to 100000 bp on chromosome five and overlap a gene named *myFavoriteGene*). Into the second field, variant types may be specified to select associated reads. Multiple types may be entered, separated with semi-colon (similar to above). Variants can be filtered on length as specified by the user in the **Minimum SV length** input field. Reads can also be selected based on read and alignment characteristics under **Read features**. Reads below input **Minimum read length** are filtered out. Only reads with alignments that pass the inputs **Minimum alignment length, Minimum MAPQ, Max. alignment masked** **perc.** (maximum alignment overlap to masked sequence, specified as percentage or fractional) and **Min. number of alignments** (minimum number of alignments of the aforementioned criteria) are kept.

### 3C: Read list

The read list is populated with selected reads sorted by length, shown within parenthesis. (If reads were selected via Genome Overview, the list is sorted first by rearrangement, indicated by **R**.) The total number of reads in the selection is indicated above the list. Reads are drawn upon a left-click and are marked in the list by a brown highlight (only if list holds <1000 reads due to performance). Additional reads can be plotted onto the regions of the current read by a right-click and is highlighted in green.

## 4: Genome overview module

The Genome overview module is launched in a separate window. This window is similar to the main window in many ways, including the general layout, zoom and navigation, and figure export. The module serves two major purposes: First, it provides the user with an overview of rearrangements indicated by split-alignments of reads. Second, it allows the selection of reads involved in a rearrangement. It is highly recommended to have repeat annotation imported when using this module to ignore alignments to repetitive sequence, but this is optional.

### Logic

The underlying logic in the Genome overview is the following:

1. The genome is divided into bins.
2. Read alignments are assigned to bins.
   1. Optionally, alignments can be filtered by length, MAPQ or the fraction of overlap to repeat annotated sequence.
3. Bins are investigated for abnormal connections by read alignments. A connection is deemed abnormal when a read has adjacent alignments which connect two non-adjacent genome bins, indicating a genomic rearrangement.
   1. Bin connections supported by a user-defined number of reads are considered.

### Control panel

A user can change Run parameters in the panel on the right side of the window. The size of bins is controlled by the input field **Bin size**. Smaller bin sizes increase the precision to group reads involved in the same event at the expense of compute time. Warning: A low bin size may lead to long computation. Reads can be discarded by length specified into **Minimum read length**. Alignments can be filtered out by length (**Minimum alignment length**), MAPQ (**Minimum alignment MAPQ**) or fraction overlap to repeat annotated sequence (**Maximum alignment masked fraction.** Note that this must not be a percentage. Any number >1 is interpreted as 1.).

Under the Show section of the panel, the user may filter what chromosomes to show: Chromosomes can be filtered out by length (**Minimum chromosome length**). Additional manual filtering of chromosomes and the order of appearance can be controlled in the **Select chromosomes** dialog. Any modifications can be restored by clicking the **Reset selection** button. The number of reads required for rearrangement flag is controlled by input field **Minimum read support**.

### Plot area

Chromosomes are drawn onto the plot area where bins that involve a rearrangement is shown in red. Red bins can be clicked to populate the Read selection list in the main window (see document section 3C). Reads are grouped in the list based on the pairwise bin connection they support (enumerated by each connection to bins other than the clicked one). Be wary of small bin sizes in large genomes as this may cause bins to not be visible until zoomed in sufficiently.

# Quick-start guide

The minimal requirement to use the software is an alignment file. To import the file, press **File** in the menu-bar then **Import BAM**. This software will now load alignments of reads. This action may take some time depending on the size of your alignment file, but enable fast display of reads once loaded. Optional import of annotation (RepeatMasker and GTF/GFF) and variant-call (VCF) files are loaded similarly.

Reads are first discovered (**Find reads** button in the right side panel) according to some selection criteria (**Read selection**) to populate the Read selection list where a read is drawn upon a left-click. Right-clicking reads will add it to the current plot and is useful to distinguish a putative rearrangement as heterogeneous or artefact. (Right-clicked reads are only drawn within the boundaries of the current left-clicked reads and is increased or decreased under **View -> Show settings**)

Drawn reads appear in the plotting area. The user can zoom in by “plus” or “negative” keyboard buttons. Shift/Alt modifiers control individual X and Y axis, respectively. Navigation is done by dragging inside the plot area. The view is resettable via menu bar **View -> Reset Zoom**.

To control what will be drawn into the plotting area, modify settings under the “Show” section in the right panel. New drawings follow these criteria; to update the current plot, click **Update** button. Genes and variants can optionally be displayed if these files have been loaded into the software.

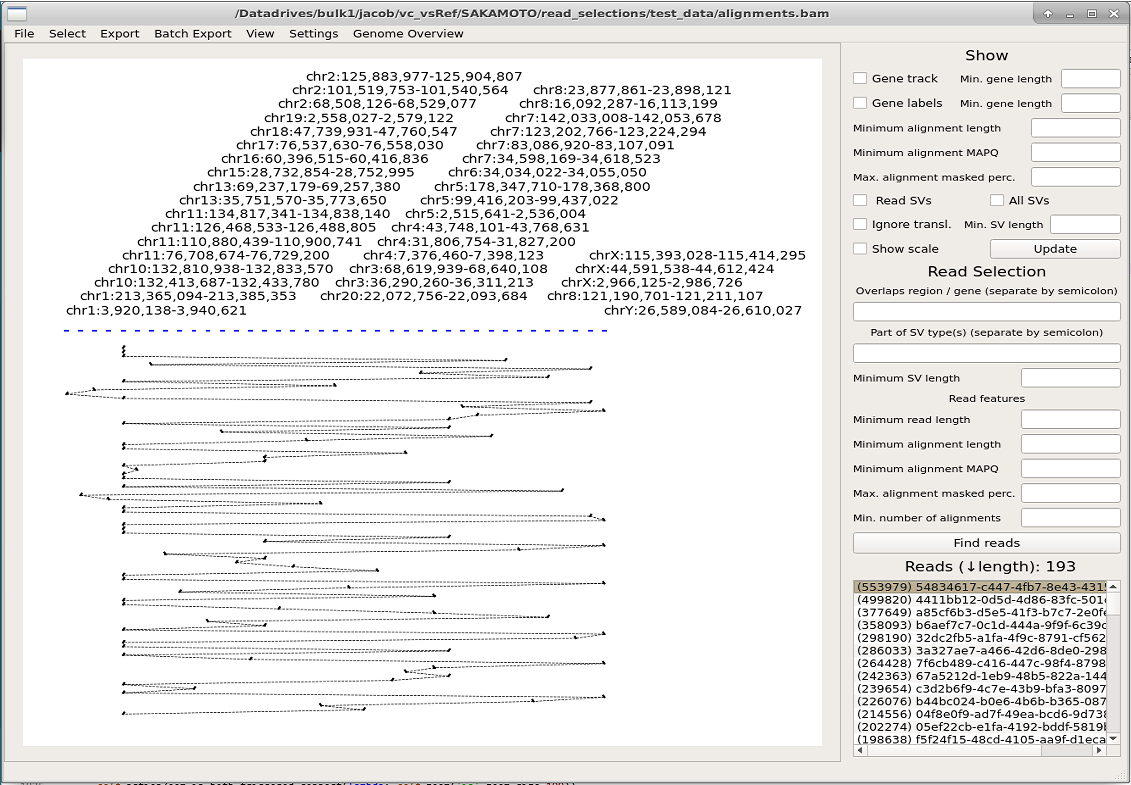
The Genome overview module is launched via menu bar **Genome Overview -> Launch module** and will open a secondary window. This module is useful to find reads involved in large-scale rearrangements in the genome. The module works by binning the genome and looking for alignments that are adjacent on the read but which overlap two non-adjacent bins.

# Examples

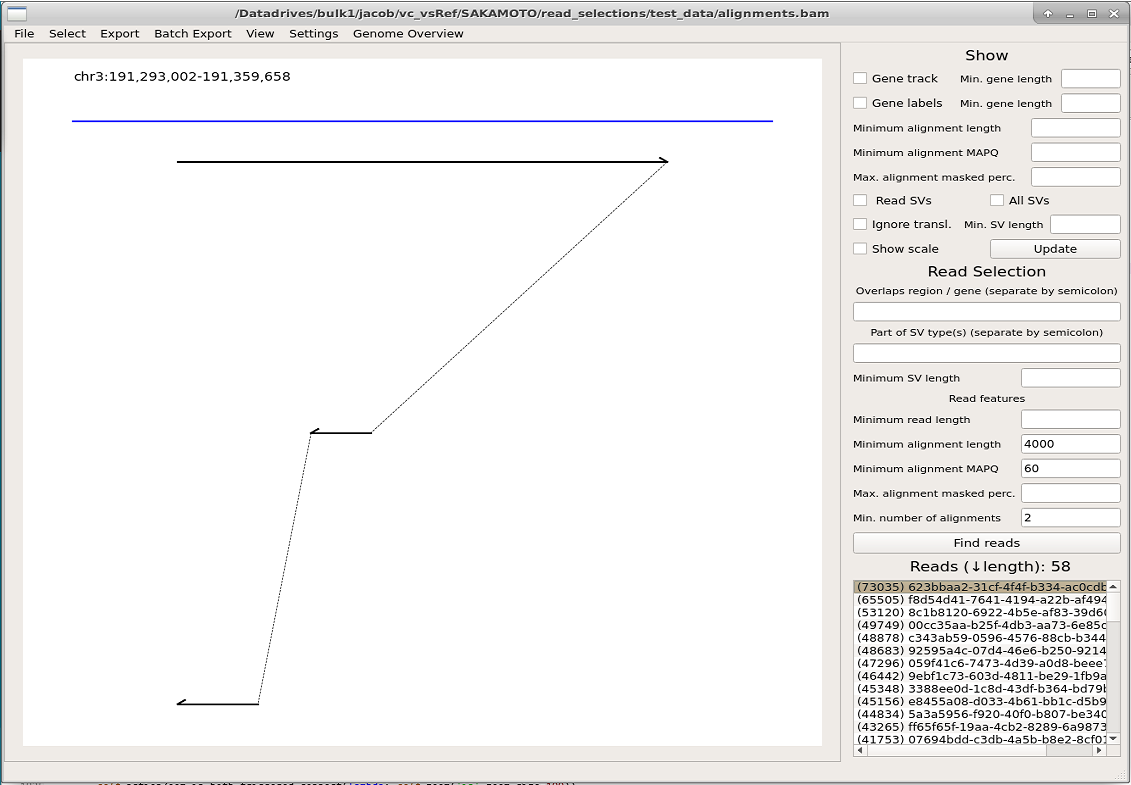
This section demonstrates the software using the included dataset, composed of selected reads from the long-read Nanopore dataset RERF-LC-KJ published by Sakamoto et. al. <http://www.genome.org/cgi/doi/10.1101/gr.261941.120>) and annotations for genes and repetitive sequence (extracted from Hg38 reference files). Variant calling was done using Sniffles with parameters –n -1 to output read names with calls on the full RERF-LC-KJ dataset and the variants were extracted for the selected reads.

## Example 1: First draw (import of reads and selection of alignments)

We start by importing the alignment file via **File->Import BAM** and pressing **Find reads** button in the right panel. This will display all reads found, in the Read selection list. Here we have 193 reads in the dataset and clicking the top read results in the draw below.

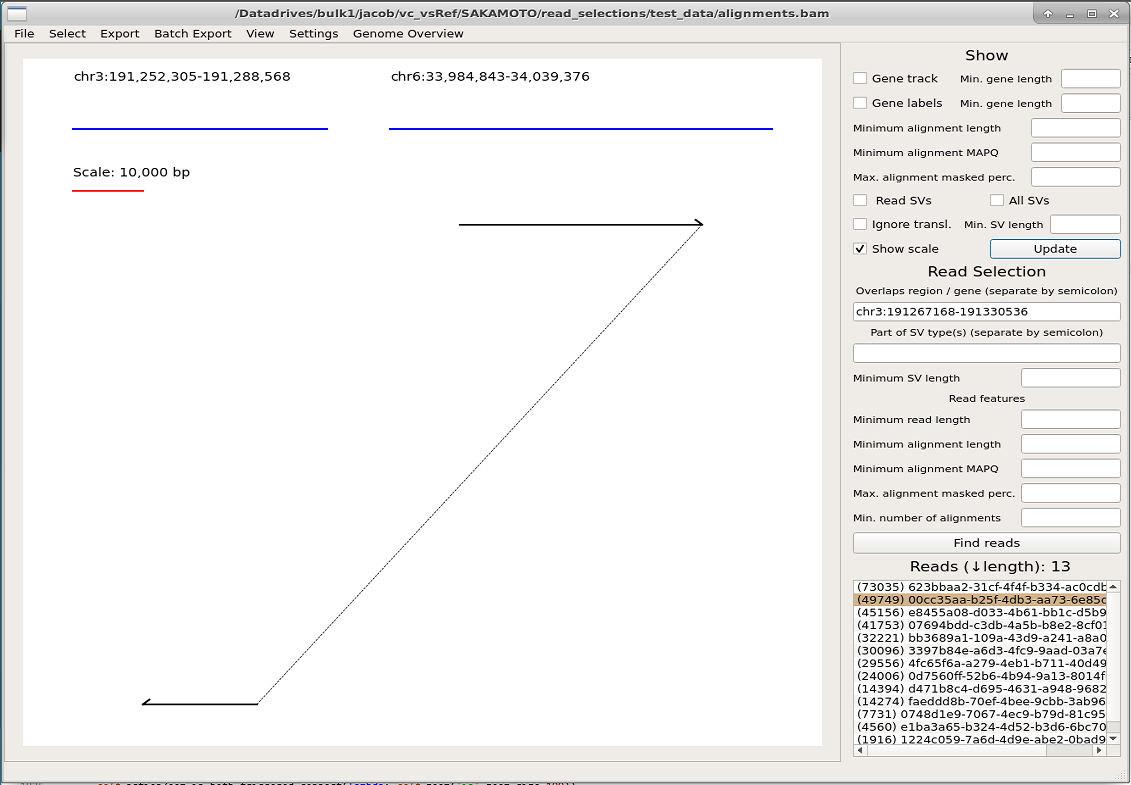
  
**Figure E1.1:** Drawing of the longest read in example dataset (This figure shows an bad read, see the next figure for a good example with explanations).

This read is likely an artefact from sequencing and, regardless of the cause, is composed of sequence short mappings to many loci in the reference genome. We can attempt to find more interesting reads by specifying a minimum alignment length of 4000 and MAPQ score 60 in at least two alignments. After finding reads by these criteria, the list return 58 reads and drawing the top read results in the image below.

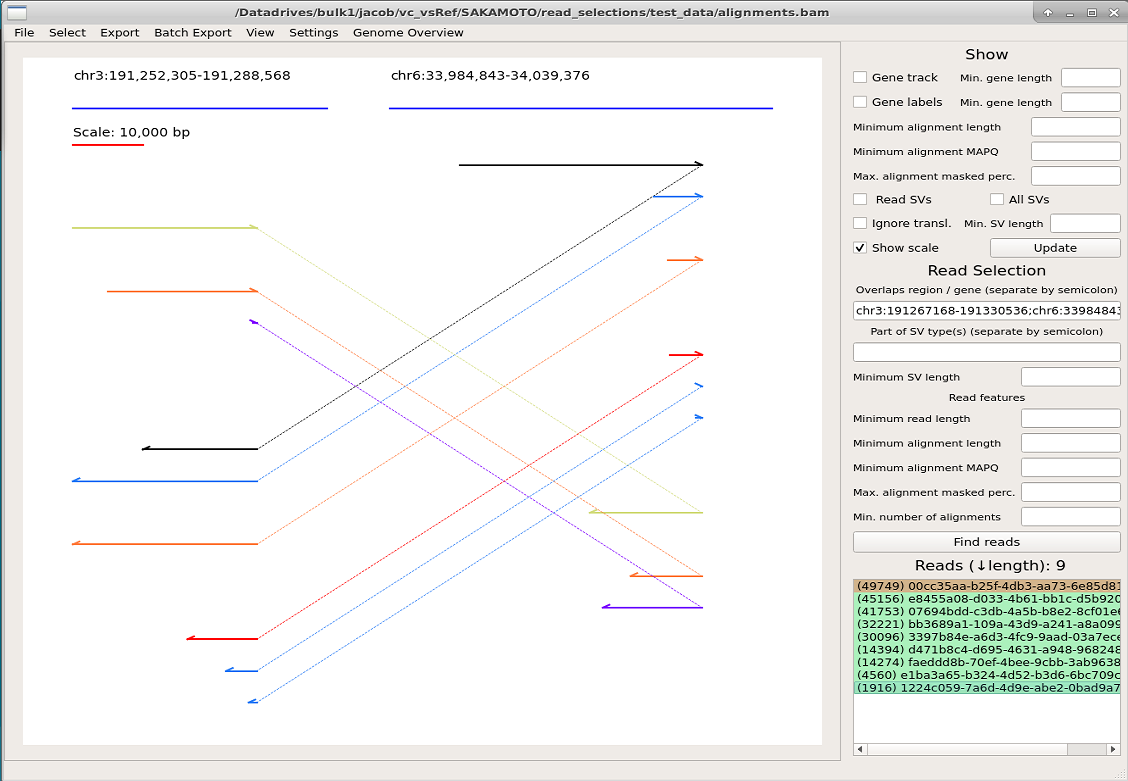
  
**Figure E1.2**: Drawing of the top read returned in read selection requiring at least two alignments with 4000 bp alignment length and MAPQ-score 60. Starting from the top, the reference region is labelled and drawn in blue. Read alignments (in-order, by alignment coordinates on the read) are painted as black lines with arrow-heads denoting the alignment strand (first alignment is forward mapped, the second and third are reverse mapped). The dotted line indicates that the alignments are connected.

## Example 2: Selecting reads in specific region(s)

Suppose we are interested in the region of UTS2B gene, located on chromosome 3 at coordinates 191,267,168 to 191,330,536. Reads overlapping this region can be selected by typing “chr3:191267168-191330536” into the first input field under the **Read Selection** section in the right panel. With no other specifications, the list now contains 13 reads (see image below). Upon drawing the reads, we note that the second read shows a remarkable alignment at chromosome 6. By checking the **Show scale** box and **Update** of the plot, we see by the scale indicator that this alignment large (>10 kbp in size) and likely reflective of a genomic rearrangement.

  
**Figure E2.1:** Drawing of read in selected region (chromosome 3) with scale indicator.

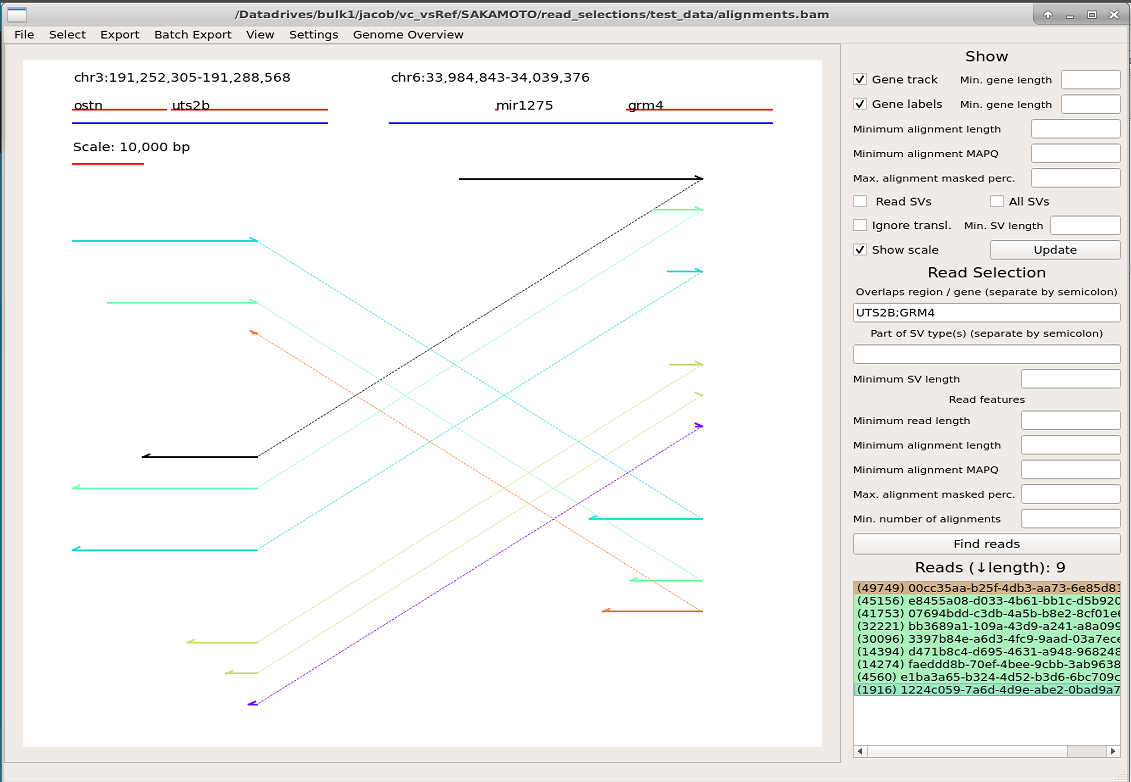
To verify the significance of this putative rearrangement, we can attempt to find reads mapping in both regions. Instead of typing the region on chromosome 6, we can copy it from **Export->Export regions** and paste it into the first field under **Read selection**, separating it from the previous region by a semi-colon. We **Find reads** and now have a list of 9 reads. We choose to draw the longest read (which is the one we have seen above) and since the list only has 9 entries, we add all entries to the plot by right-clicking them in the list, resulting in the image below.

  
**Figure E2.2:** Drawing of reads overlapping two regions. The selected read (left-clicked, highlighted brown in the list) is drawn in black and additional reads (right-clicked, highlighted green in the list) are drawn in other colors.

We conclude that this rearrangement is interesting and not a mere artefact based on the multiple reads that have breakpoints at the same position.

## Example 3: Selecting reads in specific gene(s) or region(s)

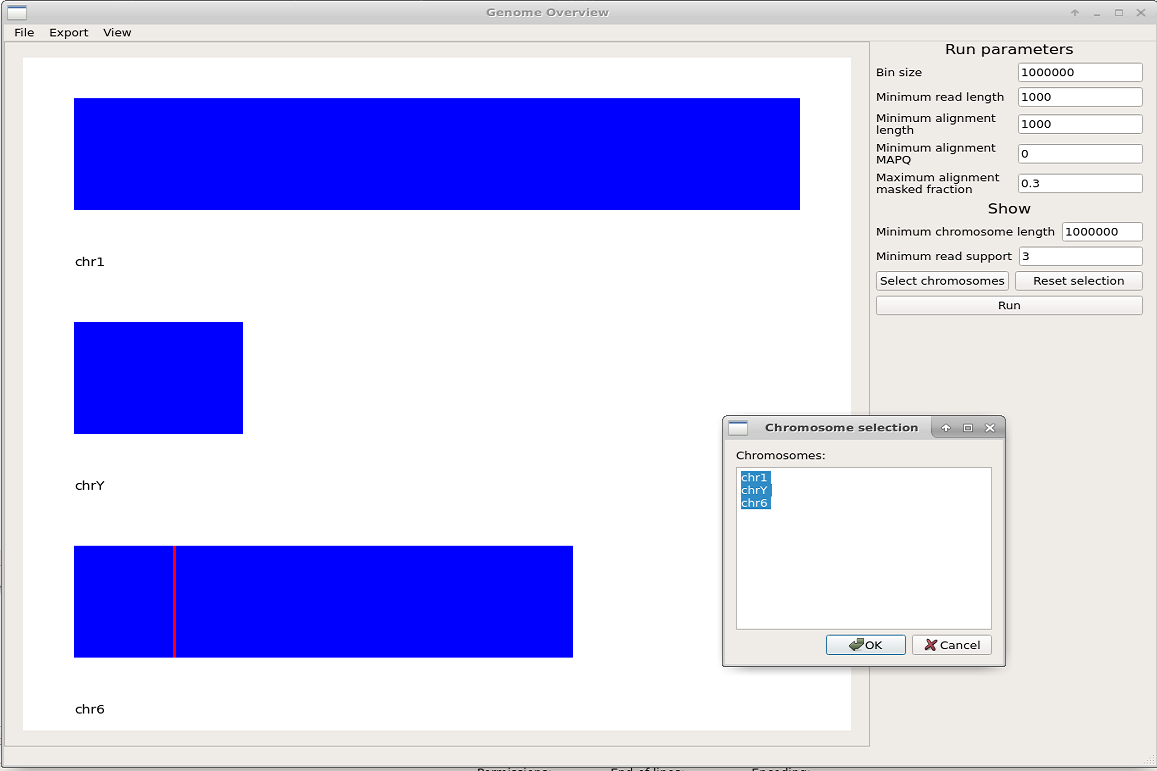
Similar to the example above, we are interested in the gene UTS2B (located on chromosome 3). After importing gene annotations via **File->Import GTF**, reads mapping to this gene can be found by entering UTS2B into the first field in the **Read Selection** section of the right panel. The resulting list shows 13 reads. We tick the checkbox **Gene track** and **Gene labels** to show the genes with names. We also tick the **Show scale** box to help us understand the size of drawn alignments. We again note reads mapping to chromosome 6, which we see overlap the gene GRM4. We could select reads mapping to both these genes by entering “UTS2B;chr6:33,984,843-34,039,376” via copy-paste from **Export->Export regions** as shown in the previous example, but instead we may enter UTS2B;GRM4 because it is more convenient.

  
**Figure E3.1:** Drawing of reads overlapping genes UTS2B and GRM4. The selected read (left-clicked, highlighted brown in the list) is drawn in black and additional reads (right-clicked, highlighted green in the list) are drawn in other colors. The gene track is shown (in red) between the reference regions (in blue) and their labels.

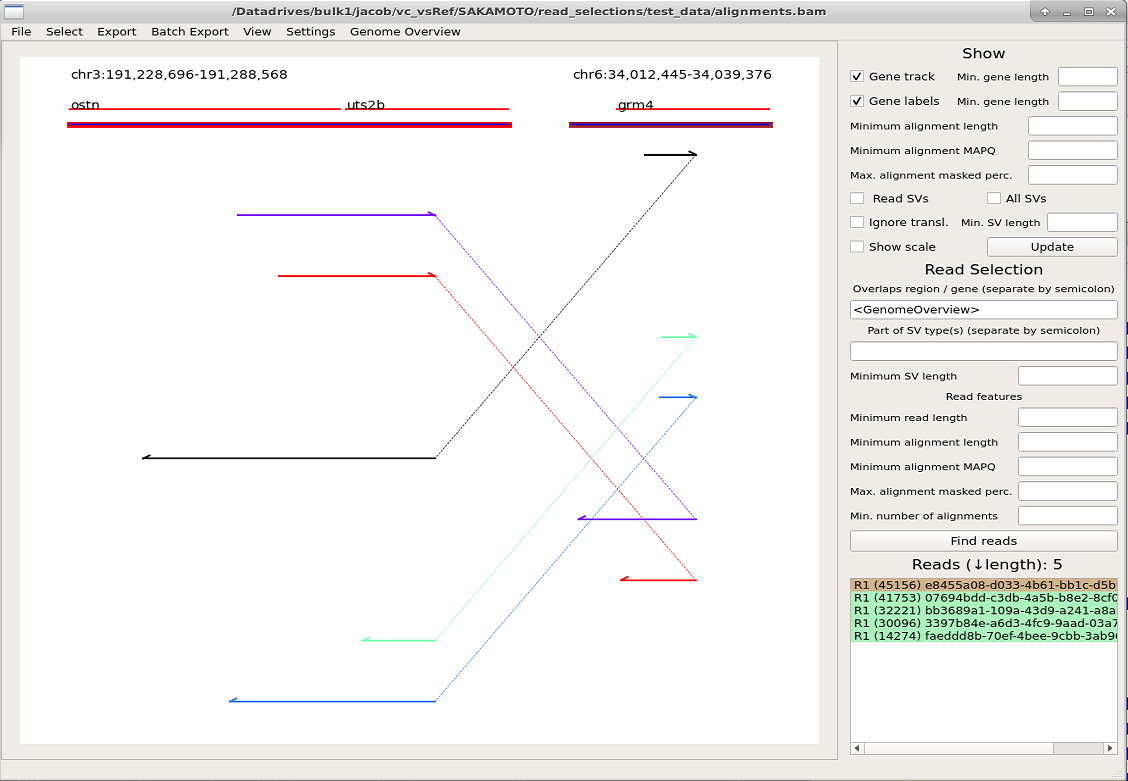
## Example 4: Identifying rearrangements via Genome overview module

We will now attempt to identify the rearrangement shown in previous examples via the Genome overview module. This module is useful to find large-scale (>hundreds of kbp or translocations, dependent on input bin size) rearrangements. We launch the module via **Genome Overview->Launch Module**. The software prompts us to include a RepeatMasker annotation file for repetitive sequence. Although we have it, we choose to ignore it at this time.

Since the human genome is large and we don’t want to zoom in on the drawing, we increase the **Bin size** from the default 100,000 to 1,000,000 (given the large human genome size, a drawing corresponding to 100 kbp may not appear visible and require zooming in on the plot). We also lower the **Minimum read support** from the default 10 to 3 reads, since this example dataset has low coverage. After starting the computation with the **Run** button (and waiting for the software to finish the compute, this take more time when datasets are larger, bin size is smaller and minimum read support is smaller) we see drawings of all chromosomes in our selection (by default these are chromosomes larger than **Minimum chromosome length**). We can modify the selection to chromosomes 1, Y and 6 (in this order) by pressing **Select chromosomes** and re-organizing the list. Press **Run** again to update the plot (given that no other settings were changed, this instantly re-draws the plot).

  
**Figure E4.1:** Genome overview of selected chromosomes. Chromosomes are painted in blue and non-adjacent bins which are connected by adjacent read alignments, indicative of a genomic rearrangement, are shown in red. The chromosome selection prompt from the **Select chromosomes** is shown. Rearranged (red) bins can be clicked to populate the supportive reads into the Read selection list of the main window.

We note the rearrangement on chromosome 6 shown in red and click on it. This populated the reads supportive of a rearrangement at this bin into the Read selection, grouped by what other bins they rearrange to. We find that the list is populated by rearrangements to various bins (R1, R2, up to R8; having some reads appear at multiple rearrangement locations (R1 and R5, displayed as R1+5). We begin to suspect that the sequence in this bin is high in repeats. As such, we close the Genome overview window, import annotation for repetitive sequence (**File->Import repeats**) and re-launch the module (**Genome overview->Launch module**). We re-set settings (**Bin size** to 1,000,000 and **Minimum read support** to 3) and also filter out alignments >70% (fraction 0.7 input to **Maximum alignment masked fraction.** Note that this number must be a fractional!). After clicking the red bin on chromosome 6 to populate the Read selection list, we now see this list contains 5 reads. Furthermore, the bin on chromosome 6 which we clicked now rearranges to one (R1) other bin. Inside the main window, we plot the first read via left-click in the Read selection list and include all other reads by right-clicking them, resulting in the below picture.

  
**Figure E4.2:** The picture shows reads selected via the Genome overview module (rearrangement bin on chromosome 6, shown in red in Figure E4.1). The reads are shown in the list grouped by rearrangement and enumerated. The gene track is shown in red between the reference region labels and drawings (by default shown as blue lines; here they are also highlighted by what rearrangement enumeration they represent).

We conclude that there is a fusion between genes UTS2B and GRM4, here shown supported by five reads.