# Reference Manual

Interactive Long-read-Visualization Tool (ILVT) is a graphical user interface software that allows users of different (or none) bioinformatics expertise to directly interact with their dataset. It takes a few minutes to load the read alignments of a 10X coverage human cancer sample sequenced with Nanopore (2.3 million reads of 30 gigabases sequence), after which reads can be interacted with instantly. Genomic rearrangements are discovered in another few minutes. ILVT is useful to:

* Overview a new dataset and guide analysis design
* Investigate specific regions (e.g., rearrangement in cancer-related genes or structural variant calls from a VCF file)
* Unravel the structure of regions with multiple rearrangements that are spanned by long-reads and not resolved with a variant calling software

The software and installation instructions are available on GitHub (<https://github.com/jaclew/ILVT>).

# Table of contents

[Reference Manual 1](#_Toc922666614)

[Table of contents 1](#_Toc1872470447)

[Installation 3](#_Toc427878521)

[Software interface 3](#_Toc1395373276)

[1: Menu bar 3](#_Toc827098893)

[1A: File 3](#_Toc1517755627)

[Import BAM 3](#_Toc1908497348)

[Import GTF 3](#_Toc278073033)

[Import repeats 3](#_Toc643721816)

[Import VCF 3](#_Toc906716413)

[Exit 3](#_Toc248499355)

[1B: Select 3](#_Toc1410416301)

[Input read name(s) 3](#_Toc1767123333)

[Import read list 3](#_Toc641742765)

[Clear selection 3](#_Toc1818613479)

[1C: Export 3](#_Toc1785337277)

[Current read 3](#_Toc446917334)

[Read highlight 3](#_Toc945124438)

[Read selection 3](#_Toc1558898098)

[Regions 3](#_Toc891203466)

[Alignments 3](#_Toc292146170)

[Image 3](#_Toc877481271)

[1D: View 3](#_Toc1205137958)

[Reset zoom 3](#_Toc766527795)

[Zoom in 3](#_Toc479624412)

[Zoom out 3](#_Toc1385161855)

[1E: Settings 3](#_Toc1831785952)

[Show 3](#_Toc139302328)

[BAM import 3](#_Toc1136266977)

[VCF import 3](#_Toc82647796)

[GTF/GFF import 3](#_Toc2086719382)

[Genome overview 3](#_Toc1249659148)

[2: Plot area 3](#_Toc795582061)

[3: Right panel 3](#_Toc820462861)

[3A: Show 3](#_Toc1058160372)

[3B: Read Selection 3](#_Toc635667839)

[3C: Read list 3](#_Toc187969990)

[4: Genome overview module 3](#_Toc855215190)

[Logic 3](#_Toc1938191851)

[Control panel 3](#_Toc879166899)

[Plot area 3](#_Toc864921508)

[Quick-start guide 3](#_Toc1384605967)

[Examples 3](#_Toc1356815441)

[Example 1: First draw (import of reads and selection of alignments) 3](#_Toc1941502192)

[Example 2: Selecting reads in specific region(s) 3](#_Toc516948262)

[Example 3: Selecting reads in specific gene(s) or region(s) 3](#_Toc1468466391)

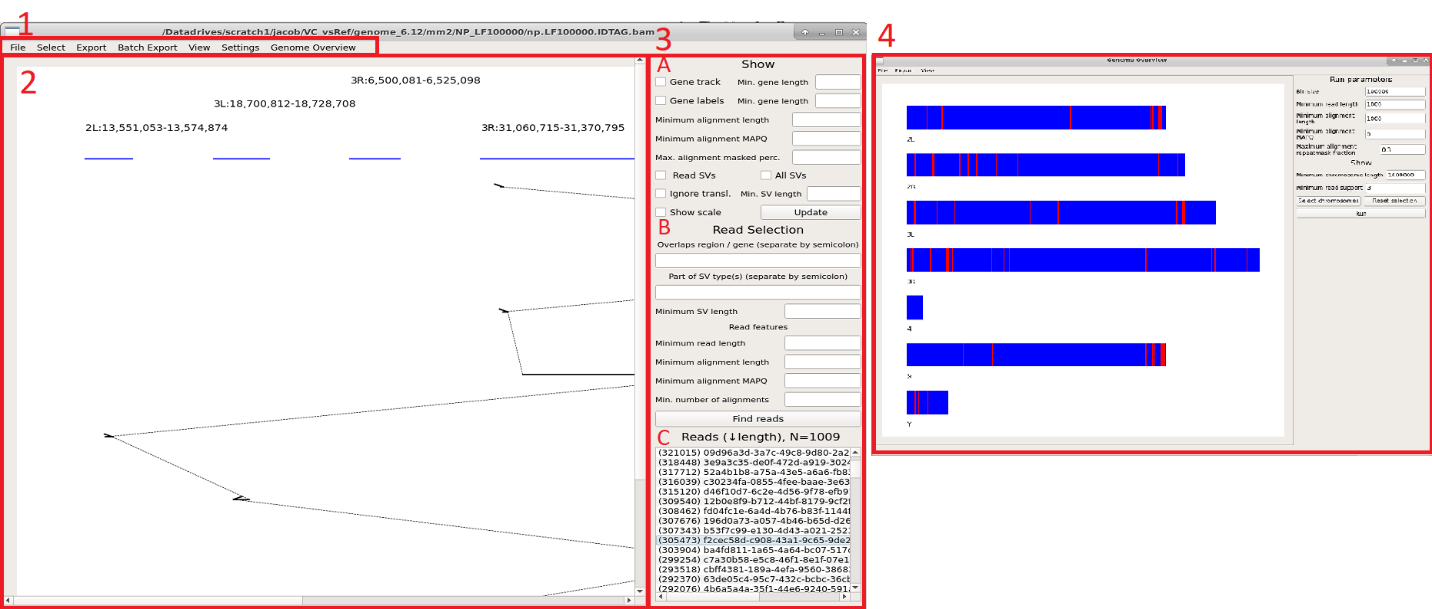
[Example 4: Identifying rearrangements via Genome overview module 3](#_Toc2119919500)

# Installation

Please refer to the ILVT Github repository at <https://github.com/jaclew/ILVT> for detailed installation instructions on Linux, Windows, and macOS.

ILVT is written in Python version 3 and depends on Python libraries PyQt5, Matplotlib, and Pysam. The libraries can be installed via pip (the Package Installer for Python). Download the software files to your location (from Github or using Git) and launch the software with Python.

# Software interface

  
Figure 1. Software graphical user interface. The user interface panels are indicated with red boxes and red text labels.

The software user interface is presented in Figure 1 and the sections below are structured according to the marked regions in the figure.

## 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 memory runs out upon import, the minimum length of reads to parse can be specified: 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 by enabling filter of alignments based on overlap in repeat-annotated regions. The software parses data from columns 5, 6 and 7. If the annotation is not in the RepeatMasker .out format, it can be converted 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 to parse the output of another variant caller. For more information, see document section 1E (VCF import).

#### Exit

Terminates the software.

### 1B: Select

#### Input read name(s)

Populates the read selection with input read names. Multiple read names can be pasted, separated by a new line 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 a new line.

#### Clear selection

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

### 1C: Export

#### Current read

Opens a dialog that holds the current highlighted read (selected by a left click in read selection, see document section 3C).

#### Read highlight

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

#### Read selection

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

#### 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 the alignment) is controlled by **Region margin**.

#### BAM import

The minimum read length can be set to discard import of short read alignments and can be increased if the user experiences issues with memory due to importing too many reads.

#### VCF import

The software was designed to parse standard VCF files and the VCF file produced by Sniffles, which holds additional data in the INFO-field. The keys parsed from the INFO field can be modified to allow the parsing of a VCF file produced by another software that use different keys. The information parsed from the INFO field by the software is the secondary breakpoint reference name and position, the variant type, the names of supportive reads and the 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 and the tag holding the gene name can be modified for both formats. Since annotation files can 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 such entries to save computation time and memory. The terms to be filtered can be modified by the user to select specific information from, or increase the filtering of, the annotation file import.

### Genome overview

The Genome overview module is launched into a new window via the menu. For more information about the module, see section 4 below.

## 2: Plot area

This area holds the canvas for paintings. Zooming in the plot can be done via plus (+) and negative (-) keys. Shift (X-axis) and alt (Y-axis) modifiers can be used to zoom on specific axis. Navigation in the plot is done using the scroll bars or by dragging movements 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”). The alignment strand relative to the reference is indicated by an arrowhead. Red markings on the alignments denote overlaps by multiple alignments on the read (the sequence of the red marking is involved in multiple alignments).

Painted alignments can be filtered and is controlled by 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 (painted upon right clicking reads in the read selection list, for more details, see document section 3C) are introduced to the plot with a random color. The alignments of additional reads are only shown within the regions of the current read painting. Plotting additional reads is useful to find if multiple reads support a rearrangement.

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

## 3: Right panel

The panel consist of three sections: Show, Read selection and Read list.

### 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 redrawn 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**). This requires a gene annotation file to be imported.

Alignments can be omitted from drawing based on length, MAPQ score, or overlap to sequence annotated as repetitive (**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 RepeatMasker .out file to be imported (For more information, see document section 1A: Import repeats).

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. In the bottom of the section, the **Show scale** checkbox controls the drawing of a scale indicator.

### 3B: Read Selection

The read selection enables the user to find reads to plot via the **Find Reads** button. Given that sequencing datasets often are large and include noise in the form of technical artefacts and sample heterogeneity, it is a challenge to filter the interesting reads. Reads can be found generally (e.g., by alignment characteristics) and specifically (e.g., in specific regions). Into the top field a region (formatted chr:start-end) or gene annotation can be specified (requires gene annotation file to be imported). Multiple requirements can be entered, separated by a semi-colon (e.g., “chr5:5000-100000;myFavoriteGene” finds reads that have alignments in the region 5000 to 100000 bp on chromosome five and that overlaps the gene *myFavoriteGene*, located elsewhere). Into the second field, variant types can be specified to select the reads that are associated with the specified variant type. Multiple variant types may be entered, separated by a semicolon (similar to above). Variant types can be filtered by length specified 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 and 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 fraction) 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 and sorted by length (shown within parenthesis before the read name). If reads were selected via Genome overview, the list is sorted by rearrangement enumeration, 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 (the highlight apply if the list holds <1000 reads, due to interface performance). Additional reads can be plotted (within 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 has the same layout and commands to zoom, navigate, and export figures as the main window. The module enables an overview of rearrangements (inferred from split alignments of reads) in the genome and to select the reads that are involved in a rearrangement. To ignore alignments associated with repetitive sequence that may confound the rearrangement discovery, it is highly recommended that repeat annotations are imported into the software, but this is optional.

### Logic

The logic by which rearrangements are called in the Genome overview is described below:

1. The genome is divided into bins.
2. Read alignments are assigned to bins.
   1. Optionally, alignments are filtered by length, MAPQ or the fraction of overlap to repeat annotated sequence (see below section).
3. Bins are queried for abnormal connections (alignments that are adjacent on the read and that connect two non-adjacent genome bins).
   1. A rearrangement is called between bin connections that are supported by a user-defined number of reads.

### Control panel

The parameters of the Genome overview can be modified in the panel on the right side of the window. The genome bin size is controlled by the input field **Bin size**, where a smaller bin size number increase the precision to group reads involved in the same event but at the expense of computation time. A higher bin size number decrease rearrangement call precision but reduces compute time. *Warning: A low bin size number may lead to very long computation time.* Reads can be discarded by length, specified into the **Minimum read length** field. 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 number must be entered as a fraction 0-1, any number >1 is interpreted as 1).

Under the Show section of the panel, the user can filter which 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 are 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). *Warning: small bin sizes in large genomes may cause bins to not be visible (the bin is less than a pixel) and requires the user to zoom in*.

# Quick-start guide

The minimal requirement to use the software is an alignment file. To import the alignment file, press **File** in the menu bar then **Import BAM**. The software will now load alignments of reads. This action may take some time depending on the size of the imported alignment file but enable fast display of reads once loaded. Optional import of annotation (RepeatMasker and GTF/GFF) and variant-call files (VCF) 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 on reads will add these to the current plot and this operation is useful to distinguish a putative rearrangement as heterogeneous or an 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 and alt modifiers control individual X and Y axis, respectively. Navigation is done by dragging the cursor inside the plot area. The view is resettable via menu bar **View -> Reset Zoom**.

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

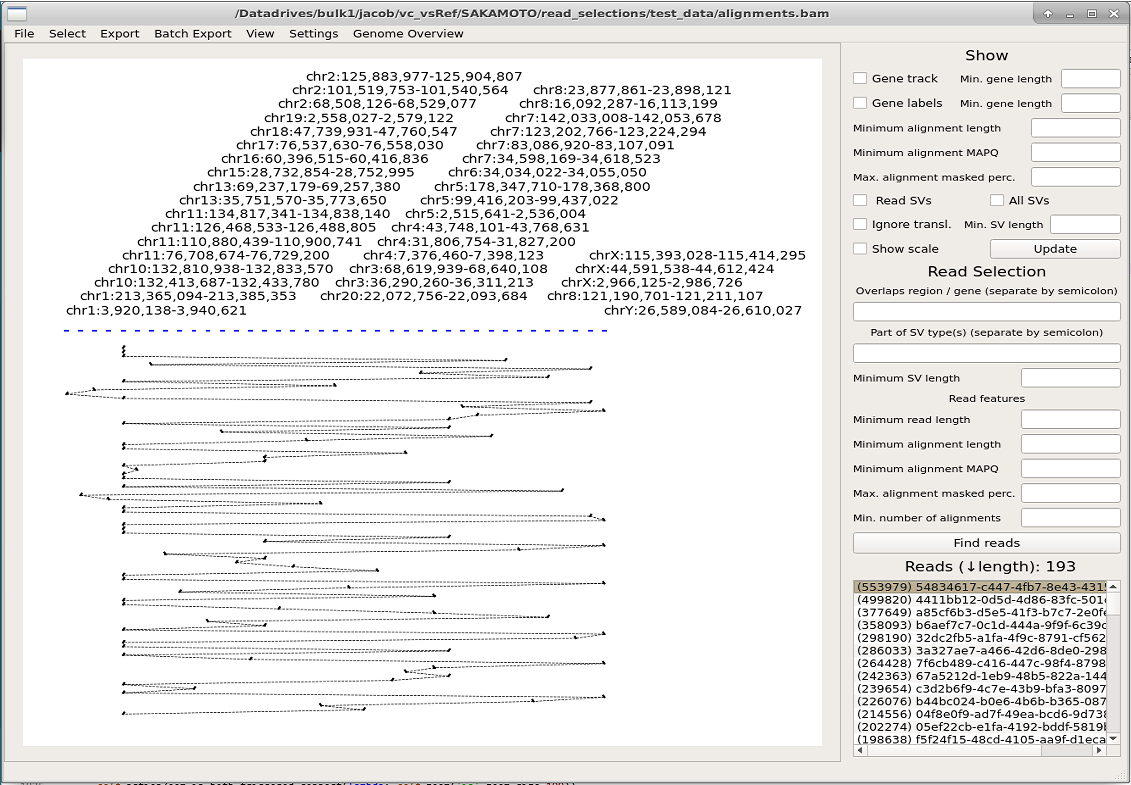
The Genome overview module is launched via the menu bar, **Genome Overview -> Launch module** and opens a secondary window. This module is useful to visualize all large-scale rearrangements in the genome and find the reads involved. The module works by binning the genome and querying alignments that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

# Examples

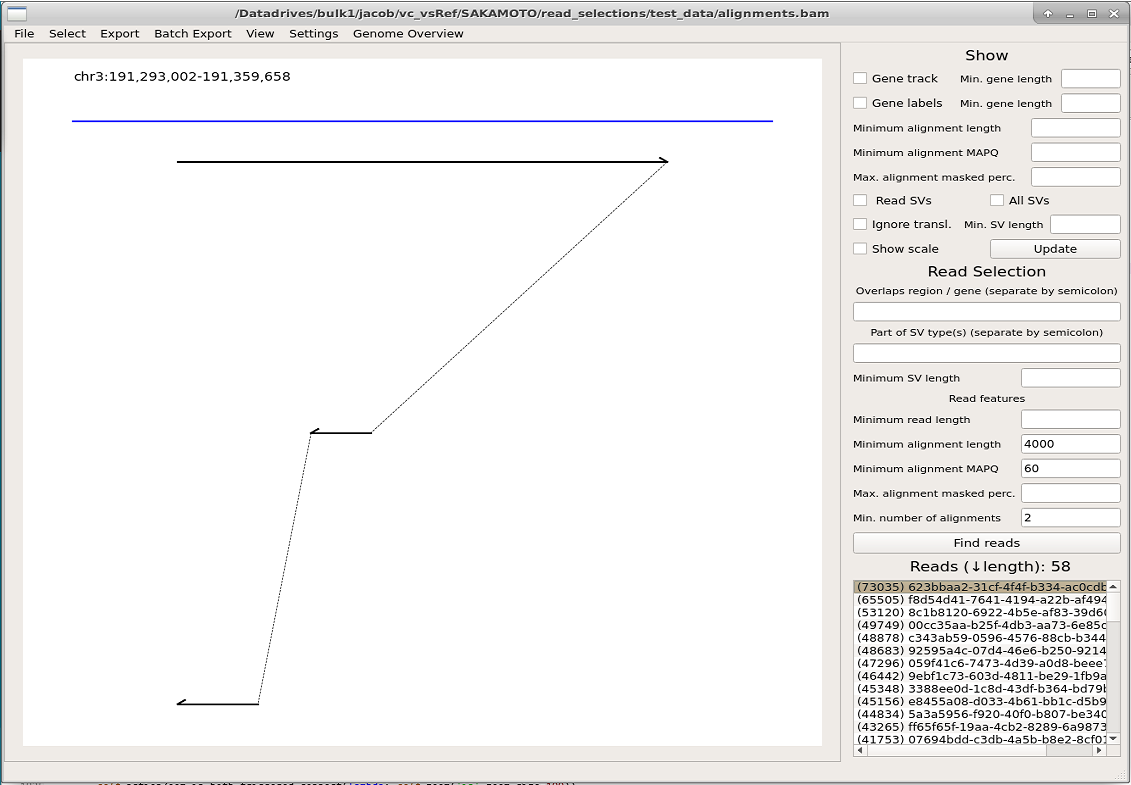
This section demonstrates the software using the included dataset, which is composed of selected reads from the long-read Nanopore dataset RERF-LC-KJ published by Sakamoto et. al.[[1]](#footnote-2) and annotations for genes and repetitive sequence (subsets of Hg38 reference files). Variant calling was done using Sniffles[[2]](#footnote-3) with parameters –n -1 to output read names and was run on the full RERF-LC-KJ dataset. Variants were kept if any of the selected reads were reported in the call.

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

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

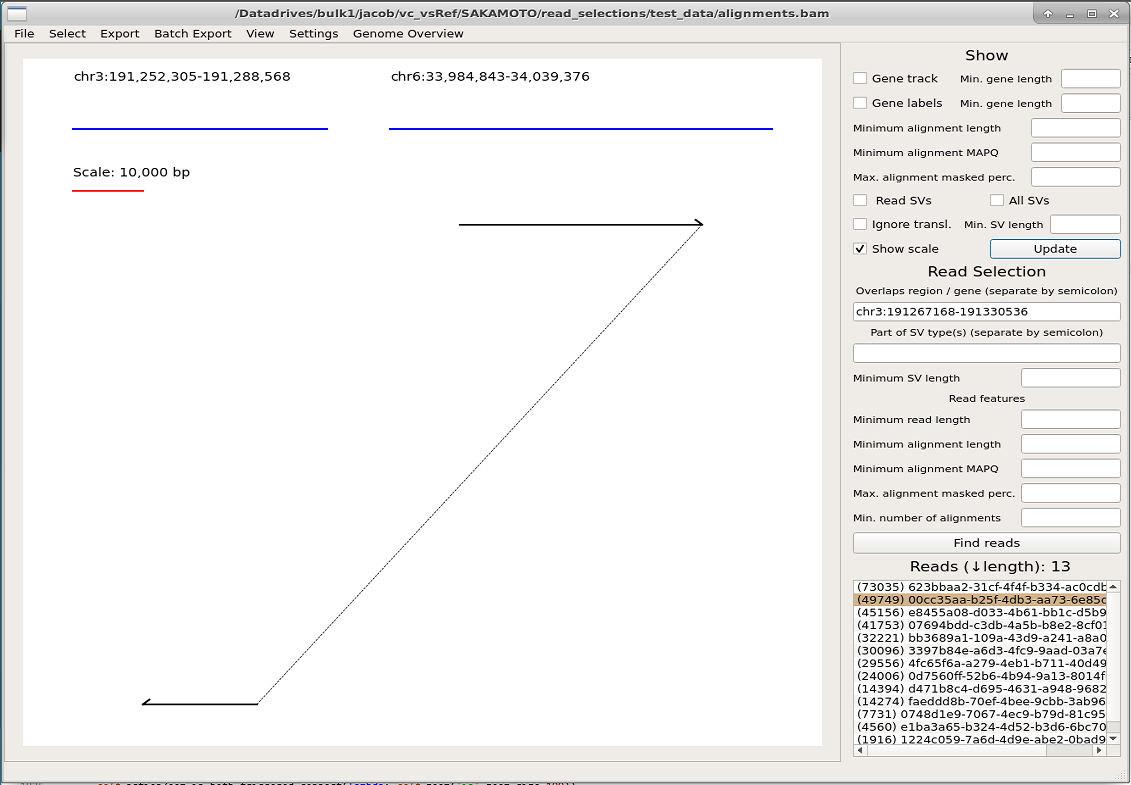
  
**Figure E1.1:** Drawing of the longest read in example dataset (This figure shows a “bad” read, see the next figure for a good example with explanations).

This read is likely an artefact from the sequencing and (regardless of the cause) is composed of short sequence mappings to many loci, scattered 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 returns 58 reads and drawing the top read results in the image below.

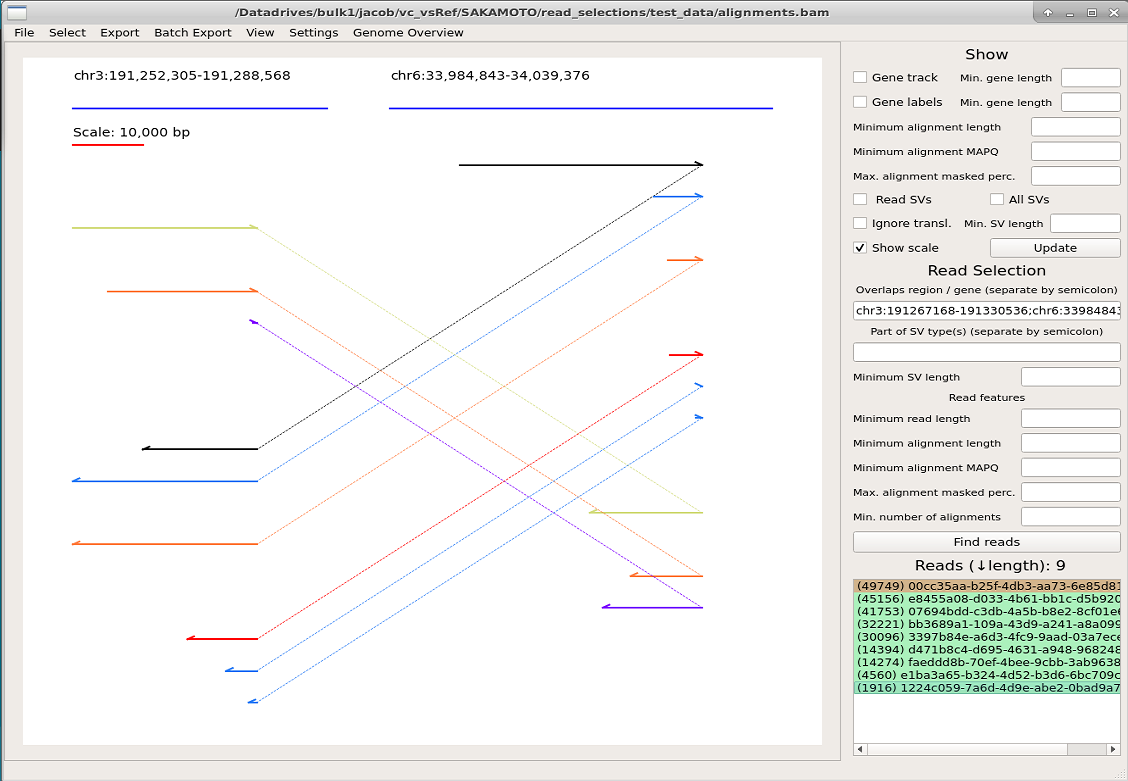
  
**Figure E1.2**: Drawing of the top read returned in read selection requiring at least two alignments with 4000 bp alignment length and MAPQ-score 60. Starting from the top, the reference region is labelled and drawn in blue. Read alignments (in order, by alignment coordinates on the read) are painted as black lines with arrowheads 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 the *UTS2B* gene, located on chromosome 3 at coordinates 191,267,168 to 191,330,536 bp. 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 on chromosome 6. By checking the **Show scale** box and **Update** of the plot, we see by the scale indicator that this alignment is 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.

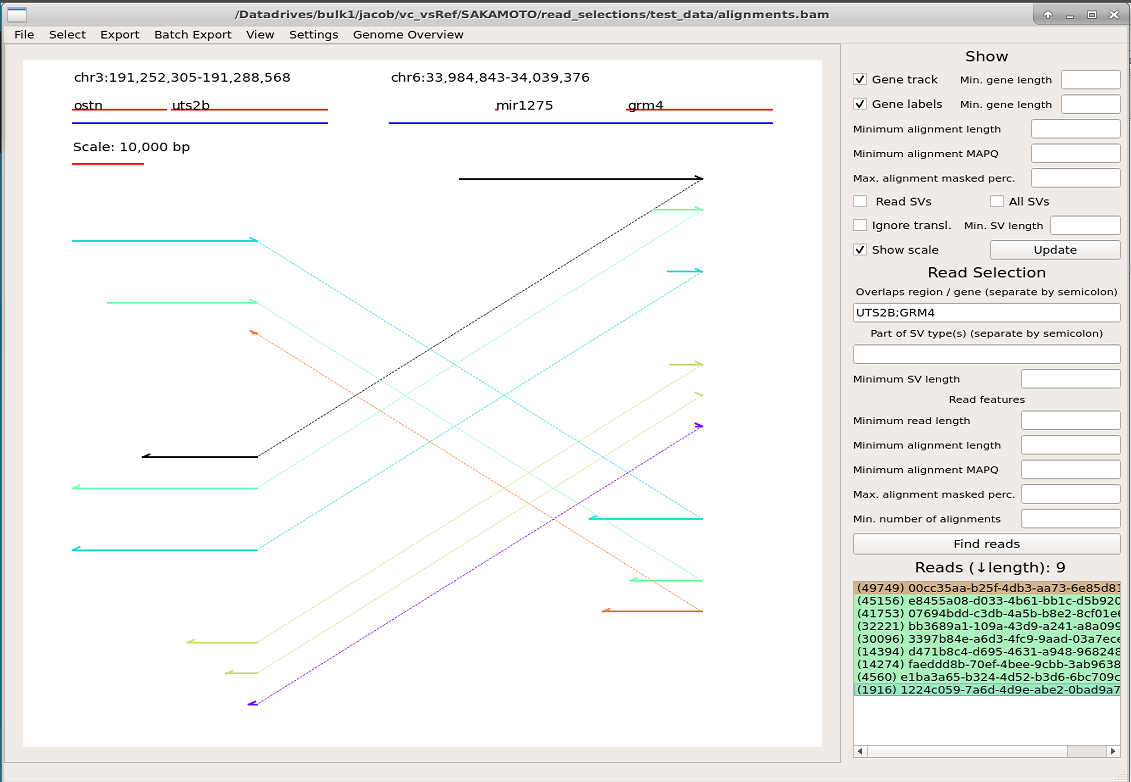
To verify the significance of this putative rearrangement, we can attempt to find more reads mapping in the regions of the currently drawn read. Instead of typing the region on chromosome 6, we can copy it from **Export->Export regions** and paste it into the first field under **Read selection**, separating 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 shown above) and since the list only has 9 entries, we draw all reads 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 since there are multiple reads that have breakpoints at the same position.

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

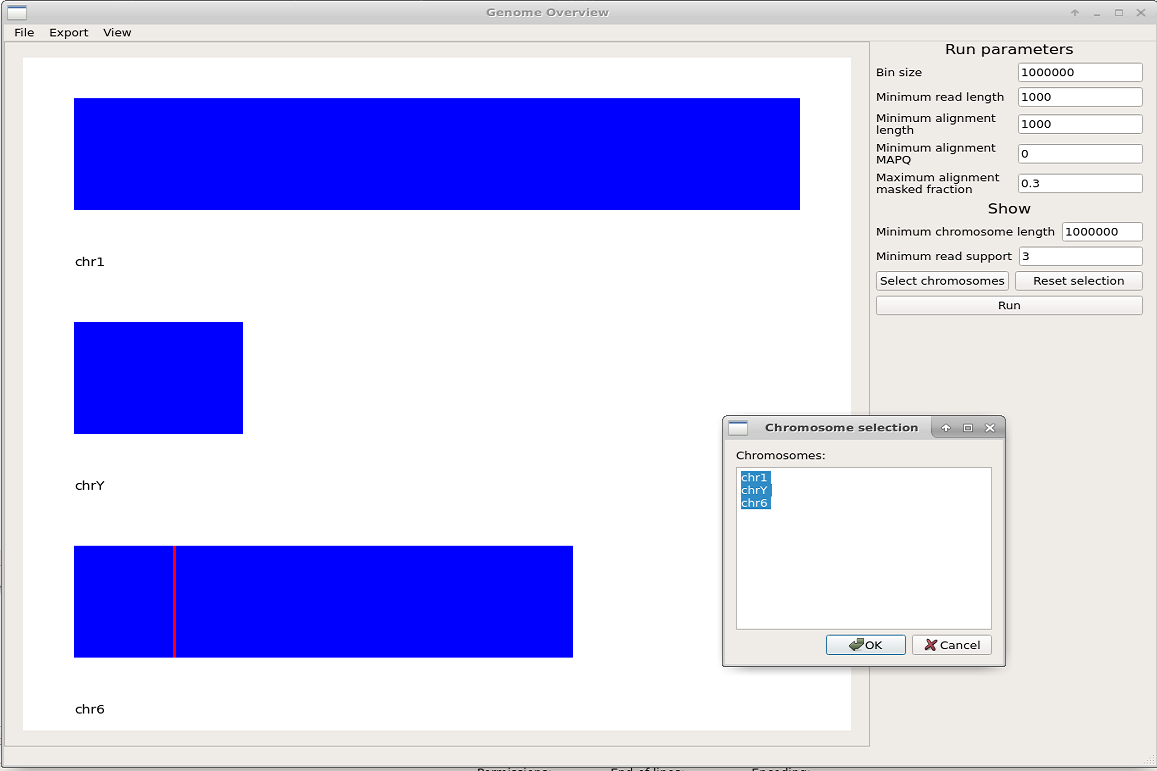
Similar to the example above, we are interested in the gene *UTS2B* (located on chromosome 3). After importing gene annotations via **File->Import GTF**, reads mapping to this gene can be found by entering UTS2B into the first field in the **Read Selection** section of the right panel. The resulting 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 realize the size of the drawn alignments. We note the reads that map to chromosome 6, and that overlaps 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 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.

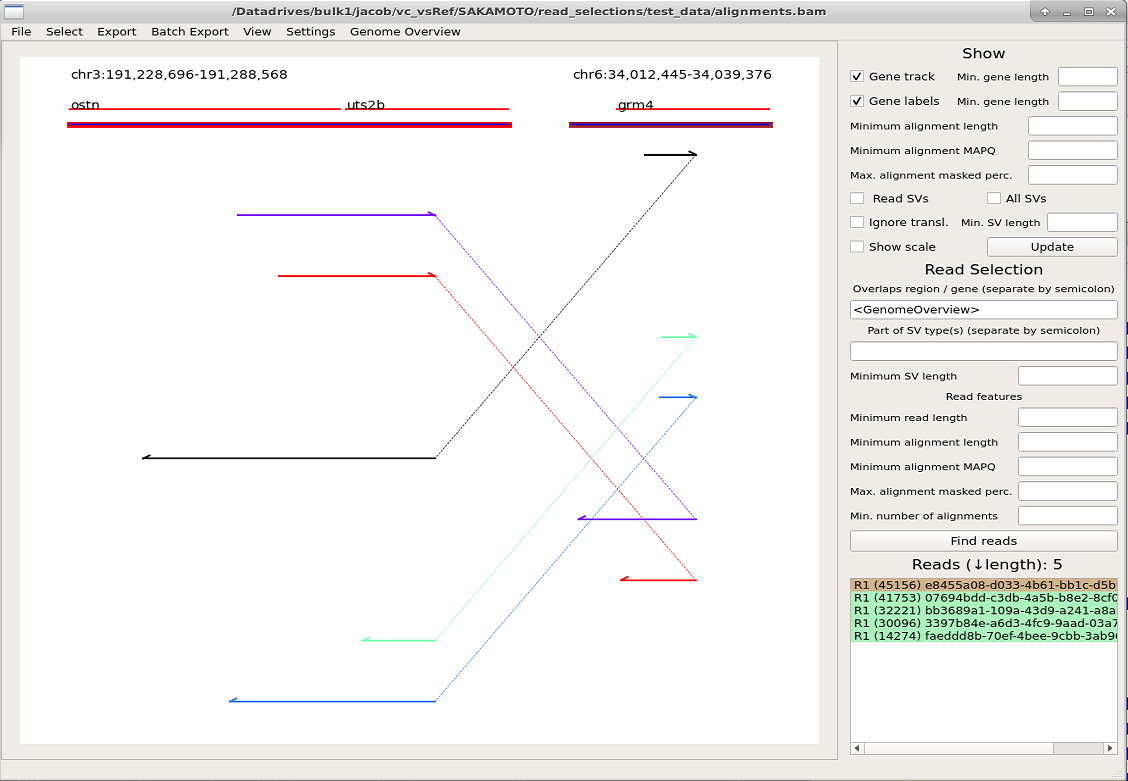
## Example 4: Identifying rearrangements via Genome overview module

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 rearrangements (>hundreds of kbp or translocations, dependent on the input bin size). We launch the module via the menu bar **Genome Overview->Launch Module**. The software prompts us to include a RepeatMasker annotation file for repetitive sequence. Although we have this annotation file, 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 be 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 computing, 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 reorganizing the list. Press **Run** again to update the plot (given that no other settings were changed, this instantly redraws the plot).

  
**Figure E4.1:** Genome overview of selected chromosomes. Chromosomes are painted in blue and non-adjacent bins that 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 the other bins that they support a rearrangement to. We find that the list is populated by rearrangements to various bins (R1, R2, up to R8; where some reads appear at multiple rearrangement locations (R1 and R5, displayed as R1+5). We suspect that the sequence in this bin is high in repeats (explaining the multiple rearrangement locations associated with the reads in this bin). As such, we close the Genome overview window, import annotation for repetitive sequence (**File->Import repeats**) and relaunch the module (**Genome Overview->Launch module**). We reset the settings (**Bin size** to 1,000,000 and **Minimum read support** to 3) and then filter out alignments >70% (fraction 0.7 input to **Maximum alignment masked fraction.** Note that this number must be a fraction!). 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 that we clicked on now has a rearrangement to one other bin (R1). Inside the main window, we plot the first read by a left click in the Read selection list and add 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 also highlighted by what rearrangement enumeration they represent from the Genome overview).

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

1. Sakamoto, Y., Xu, L., Seki, M., Yokoyama, T. T., Kasahara, M., Kashima, Y., Ohashi, A., Shimada, Y., Motoi, N., Tsuchihara, K., Kobayashi, S. S., Kohno, T., Shiraishi, Y., Suzuki, A., & Suzuki, Y. (2020). Long-read sequencing for non-small-cell lung cancer genomes. *Genome research*, *30*(9), 1243–1257. https://doi.org/10.1101/gr.261941.120 [↑](#footnote-ref-2)
2. Sedlazeck, F.J., Rescheneder, P., Smolka, M. et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods 15, 461–468 (2018). https://doi.org/10.1038/s41592-018-0001-7 [↑](#footnote-ref-3)