ReadXplorer User Manual

For ReadXplorer Version 2.2.3

Manual Author:
Dr. Rolf Hilker

Bioinformatics and Systems Biology
Faculty of Biology and Chemistry
Justus-Liebig University Gießen, Germany
readxplorer@computational.bio.uni-giessen.de
http://www.readxplorer.org

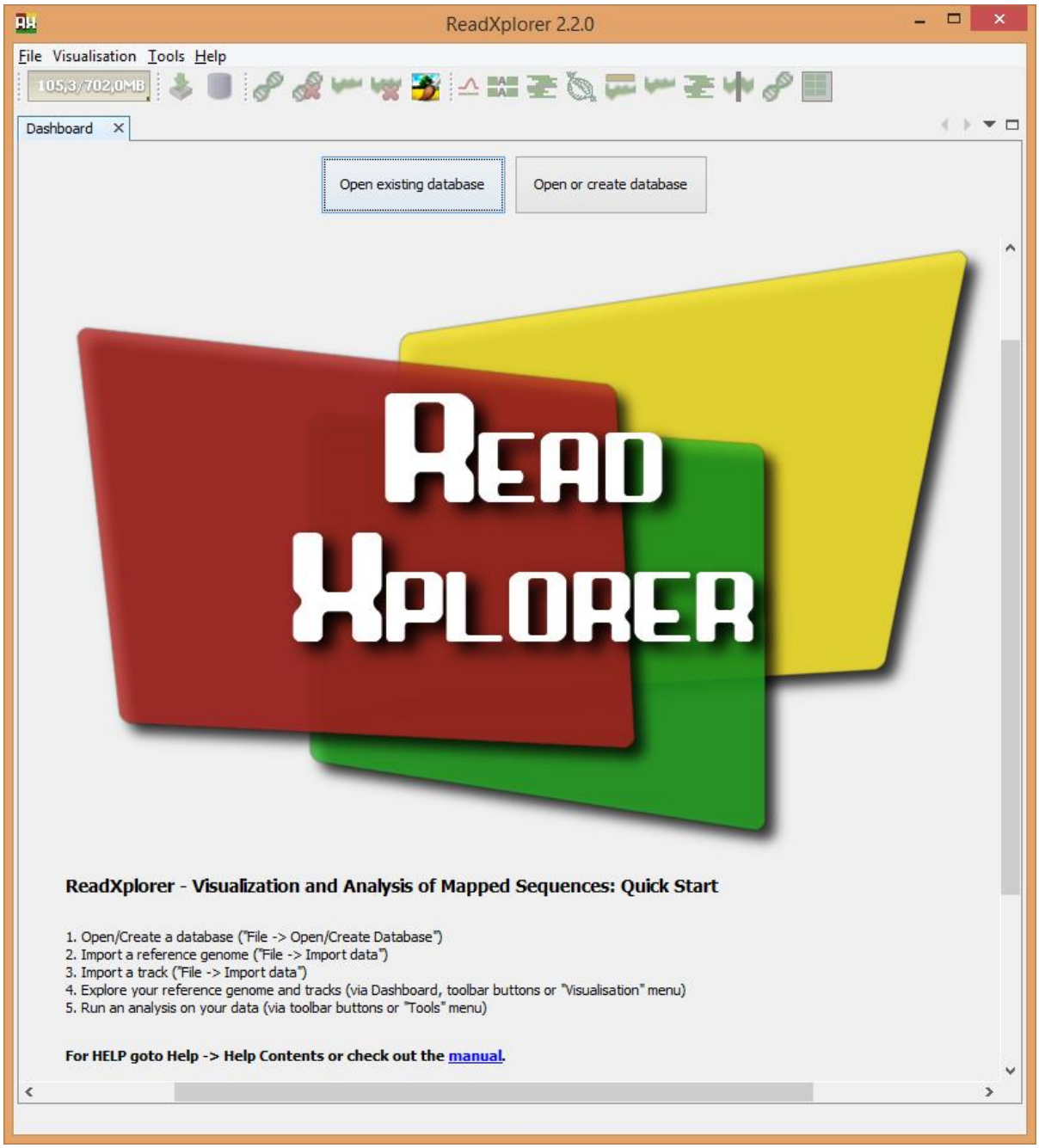
2016-10-13

This document is meant to explain all functionality provided by ReadXplorer and will be continuously expanded when new ReadXplorer versions are released. If something is unclear or you are missing the explanation for a feature, please feel free to write me an email.

Content

ReadXplorer

ReadXplorer is a client software. The project data for each ReadXplorer project is stored in a database (h2 or MySQL, we recommend h2 here). The database (DB) contains the reference genome and the basic track (= mapping data set) information. Therefore, the reference sequence and annotations are stored in the DB for the reference. For the tracks only the basic information, such as name and path to the corresponding BAM file, is stored in the DB.



ReadXplorer – Command Line Interface

To start the command line version of ReadXplorer, the *readxplorer-cli* start scripts or exe files from the *bin* directory have to be used. The help of the command line interface is available via the command "*readxplorer-cli --help*" and below. This version of ReadXplorer enables the use of ReadXplorer within automatic bioinformatics pipelines. It automatically creates a database, imports references and tracks and subsequently starts automatic analysis functions of choice – all with a single command.

Examples:

readxplorer-cli --ref ./Escherichia_coli.gb **--reads** ./se-mappings/ **--db** my-first-rx-project

This command starts the ReadXplorer CLI version importing the Escherichia coli reference in GenBank format as well as all single-end read mapping files (.bam) in the folder "se-mappings". The database file of the resulting project will be named "my-first-rx-project".

readxplorer-cli --ref ./Escherichia_coli.gb **--read**s ./pe-mappings/ **--paired-end**

In contrast to the first example, this time ReadXplorer imports all bam files in "./pe-mappings" as interlaced paired-end read mappings as indicated by the *–paired-end option.*

readxplorer-cli --ref ./Escherichia_coli.gb **--reads** ./pe-1-mappings/ **--per** ./pe-2-mappings/

In this third example ReadXplorer behaves exactly the same way as in the second but this time paired-end mappings are provided as non-interlaced files split into forward and reverse reads located in "pe-1-mappings" and "pe-2-mappings", respectively.

readxplorer-cli --ref ./Escherichia_coli.gb **--reads** ./se-mapping **--db** my-second-rx-project **--threads** 10 **--verbose --snp --tss**

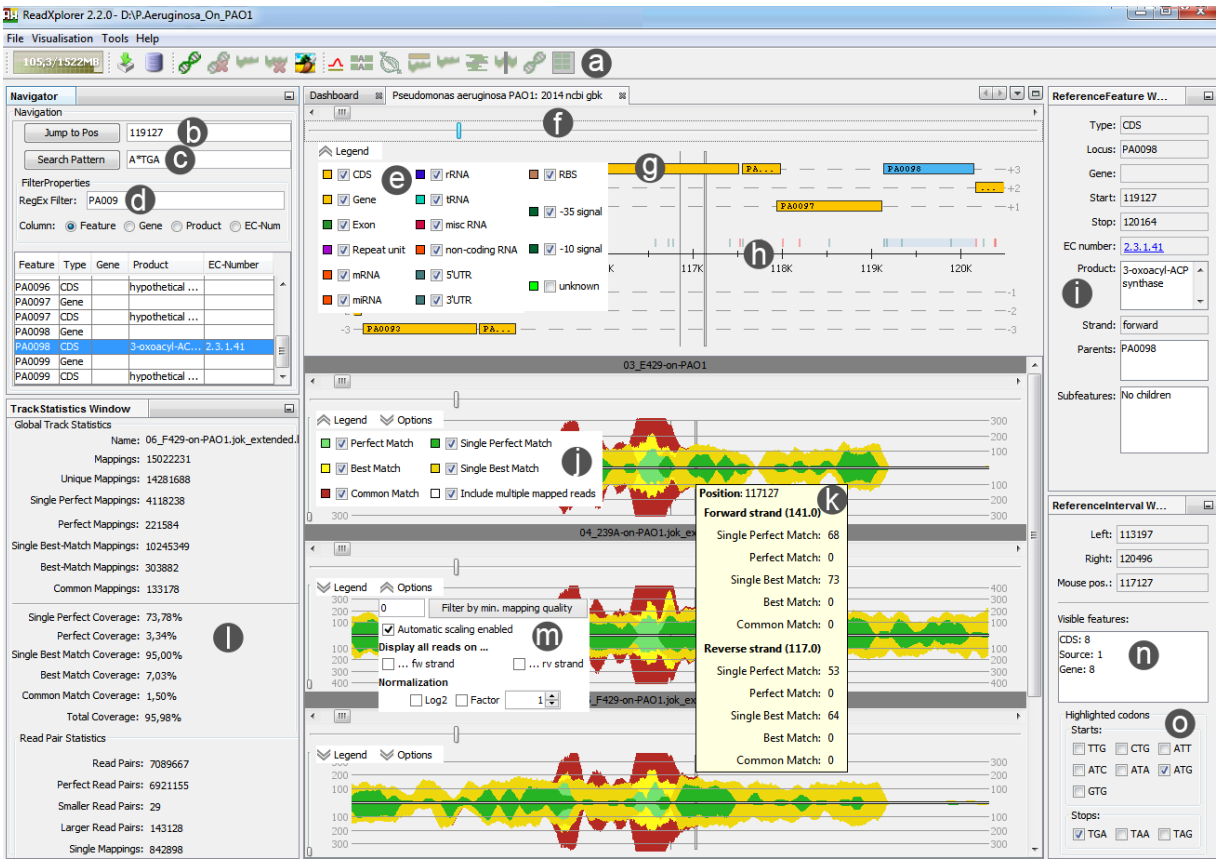
By this command ReadXplorer imports single-end mappings in "se-mapping" and names the database file of this new project "my-second-rx-project". Additionally, ReadXplorer uses up to 10 threads prints more detailed information and performs SNP detection analyses as well as transcription start site analyses. For each kind of analysis a separate spreadsheet file will be stored containing a sheet for each mapping file.

Available options:

Short Option	Long Option	Description
	--ref	Path to reference genome file. Supported formats: FASTA, GFF2/3, GenBank, EMBL
	--reads	Path to directory with SAM/BAM read files to import and analyze. Supported read files: single-end and forward or interlaced paired-end
	--per	Path to directory with SAM/BAM reverse paired-end read files to import and analyze. Use this parameter if paired-end reads are not interlaced but separated into forward and reverse reads.
	--threads	Number of available worker threads, e.g. with 10 threads one can import/analyze 10 read files in parallel.
	--db	Database name to persistently store imported data (default: readxplorer). Existing databases will be deleted!
	--properties	Path to a custom property file. This file can be used to set custom analyses properties.
-p	--paired-end	Flag indicating <reads> as interlaced paired-end reads. Use this flag only if <per> is not set.
	--snp	Perform single nucleotide polymorphism (SNP) analyses on all imported read files.
	--tss	Perform transcription start site (TSS) analyses on all imported read files.
-v	--verbose	Print detailed messages.
-h	--help	Print the readxplorer-cli usage.

ReadXplorer - Main Window

This is the main window of ReadXplorer with one opened reference and three opened track data sets. The different items are explained below.



a) Global Toolbar and context dependent toolbar buttons. From left to right:

1. RAM monitor (available via right-click on the toolbar -> Performance)
2. Import data into the opened DB
3. Delete data from the opened DB
4. Open reference
5. Close the currently selected/viewed reference
6. Open track(s) (also in Multiple Track Viewer)
7. Close all opened tracks for the currently selected/viewed reference
8. Open save screenshot wizard
9. Start a *Differential Gene Expression Analysis* via a wizard
10. Start a *SNP and DIP Detection* via a wizard
11. Start a *Genome Rearrangement Detection* for read pair data using GASV
12. Start different *Transcription Analyses* via a wizard
13. Start a *Genomic Feature Coverage Analysis* via a wizard
14. Start a *General Coverage Analysis* via a wizard
15. Open the *Detailed Viewer* for a track belonging to the currently viewed reference (*Histogram Viewer, Alignment Viewer, Read Pair Viewer*)
16. Open the double track viewer for selected tracks belonging to the currently viewed reference
17. Open the reference sequence editor
18. Open the *Thumbnail Viewer*

b) Position navigator: By clicking "Jump to Pos", the entered position is centered for the currently viewed reference. If the reference contains multiple chromosomes/contigs/other sequences, a selection box is displayed as well. The position is then chosen from the selected chromosome.

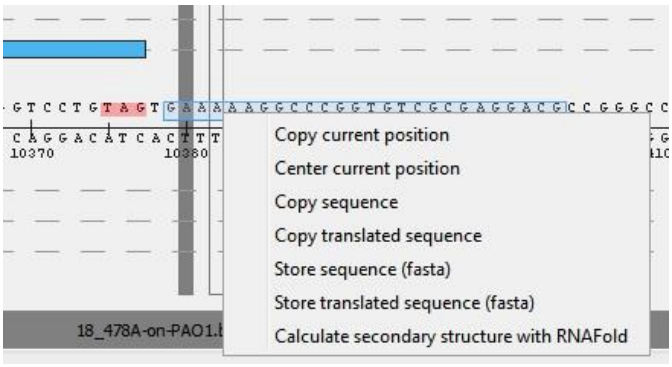
c) Pattern filter: Searches and centers the next occurrence of the entered pattern/sequence and highlights all occurrences in the currently visible interval on the sequence bar of the Reference Viewer in blue. By clicking the "Search Pattern" button a second time, the next occurrence of the pattern, which is currently beyond the viewed reference interval, is centered.

d) Reference feature filter: Filters the currently selected column of the reference feature table below for the regular expression or plain text entered in the "RegEx Filter" field. By clicking any of the features in the table, its start position is centered in the viewers.

e) Reference Viewer and its legend: It displays the reference features in a six frame view, with a sequence bar in the middle. The legend displays the coloring of each available feature type and contains a checkbox to enable or disable viewing of the respective feature type. In the example all features are shown, except the features of the type "unknown". Next to the legend, a chromosome selection box is shown, if multiple chromosomes/contigs/other sequences are contained in the current reference.

f) Scrollbar and zoom slider: They are included in each viewer. The scrollbar allows scrolling to each reference position and the zoom slider allows adjusting the zoom level of the currently viewed reference interval.

g) Selected reference feature: A feature selected by a left click is highlighted in blue and its details are shown in the ReferenceFeature Window (*i)*). If start and stop codons are currently shown, then only the codons in the reading frame of the selected feature are shown. Only one feature can be selected at a time.



h) Sequence bar and interactive start codons: In the middle of the reference viewer, the sequence bar is shown. When zoomed in completely, the reference sequence of both strands is visible. The sequence bar also offers to select a sequence of interest by holding the left mouse button on the sequence bar and selecting the sequence of interest. After selecting a sequence, the right-click menu offers the options shown in the screenshot below. E.g. copying or storing (as fasta) the sequence of interest and predicting the RNA secondary structure (see *RNA Secondary Structure Prediction*). Furthermore, the sequence bar displays start and stop codons, if they are selected in (*o)*). By clicking on a highlighted start codon (red) the reference

interval from the start to the next in-frame stop codon is highlighted in light blue. By clicking the start codon again, the highlighting disappears. For more details see *General Options - Genetic Code*.

i) Reference Feature Window: Displays the details of the currently selected reference feature. For features including an EC-number, it shows the link to an enzyme database of choice. This database link can be configured via "Tools->Options->Miscellaneous->Locations" Here, the list of "Parents" is the list of reference features placed directly upward in the feature hierarchy and the list of "Subfeatures" contains all features placed directly downward in the feature hierarchy. The hierarchy is based on the attribution in the original reference file (e.g. GFF3 feature hierarchy). When other import formats are used (Genbank, EMBL), a gene is the top parent feature and can contain exons, CDS and different RNAs. An exon can again contain CDS or RNA annotations.

j) Track Viewer and its Legend: Displays the coloring of the five mapping classes used by ReadXplorer. The green area ((Single) Perfect) depicts reads that perfectly map to the reference sequence with either exactly one (Single Perfect Match, green) or multiple perfect mappings (Perfect Match, light green), the yellow area ((Single) Best Match) depicts reads that map to the respective location of the reference with the least possible number of mismatches in the whole reference. If exactly one Best Match mapping exists for the read, it is a Single Best Match (dark yellow), but there can be more than one mapping for one read on the reference with e.g. 2 mismatches. In this case both belong to the Best Match class (bright yellow). A red area (Common Match) depicts reads that still match to the red colored region of the reference with less than the maximum allowed mismatches during the mapping process, but there is another position in the reference, where they fit with less mismatches. For further details see *Read Classification.* The options menu of the Track Viewer is described in *m)*.

k) Track Viewer tooltip: The tooltip shows the coverage of the hovered reference position for all five supported mapping classes separated by strand.

l) Track Statistics Window: Displays the global data set statistics of the currently selected track.

1. Mappings: The total number of mappings in the whole data set, which were mapped to this reference.
2. Unique mappings: The total number of mappings among the number of "Mappings", which where only mapped once to the reference.
3. Single Perfect mappings: The total number of mappings belonging to the Single Perfect Match class in the whole track data set.
4. Perfect mappings: The total number of mappings belonging to the Perfect Match class in the whole track data set.
5. Single Best Match mappings: The total number of mappings belonging to the Single Best Match class in the whole track data set.
6. Best Match mappings: The total number of mappings belonging to the best match class in the whole track data set.
7. Single Perfect Match coverage: The percentage of the reference sequence, covered by at least one mapping of the Single Perfect Match mapping class.
8. Perfect Match coverage: The percentage of the reference sequence, covered by at least one mapping of the Perfect Match mapping class.

9. Single Best Match coverage: The percentage of the reference sequence, covered by at least one mapping of the Single Best Match mapping class.
10. Best Match coverage: The percentage of the reference sequence, covered by at least one mapping of the Best Match mapping class.
11. Common Match coverage: The percentage of the reference sequence, covered by at least one mapping of the Common Match mapping class.

If the currently viewed track is a read pair (paired-end or mate pair) data set, the read pair statistic is displayed:

1. Read Pairs: The total number of read pairs successfully (= at least one read of the pair) mapped to the reference sequence in the whole track data set.
2. Perfect read pairs: The number of perfect read pairs, whose distance is within a given range and whose orientation is as expected.
3. Smaller read pairs: The number of read pairs whose distance is smaller than the minimum allowed distance of a perfect pair.
4. Larger read pairs: The number of read pairs whose distance is larger than the maximum allowed distance of a perfect pair.
5. Single mappings: The number of mappings whose partner either did not map to the reference at all, or is already contained in a perfect or smaller distance pair with the read of this mapping and its partner. The latter can occur, when e.g. one read of a read pair maps to two or more positions in the reference, but in one position it can form a perfect or smaller distance pair with its partner. Then all other mappings of the multi mapped read are classified as single mappings.

m) Track Viewer Options: Here, the coverage can be filtered by mapping quality (if the quality is available for this data set, otherwise nothing is changed). Automatic scaling of the coverage can be enabled to adapt the coverage plot to fit the viewer area. For non-strand specific data, all reads can be displayed on the forward or on the reverse strand. Besides that, a normalization method can be applied to the coverage of the track: either by log2 or by a chosen factor x with $0 < x < 10$.

n) Reference Interval Window: It summarizes information about the currently viewed reference interval.

o) Start and stop codon selection: Allows the selection of start and stop codons, which are then highlighted in the sequence bar of the reference viewer (h). If you want to change the start and stop codons to fit another organism (see *General Options - Genetic Code*).

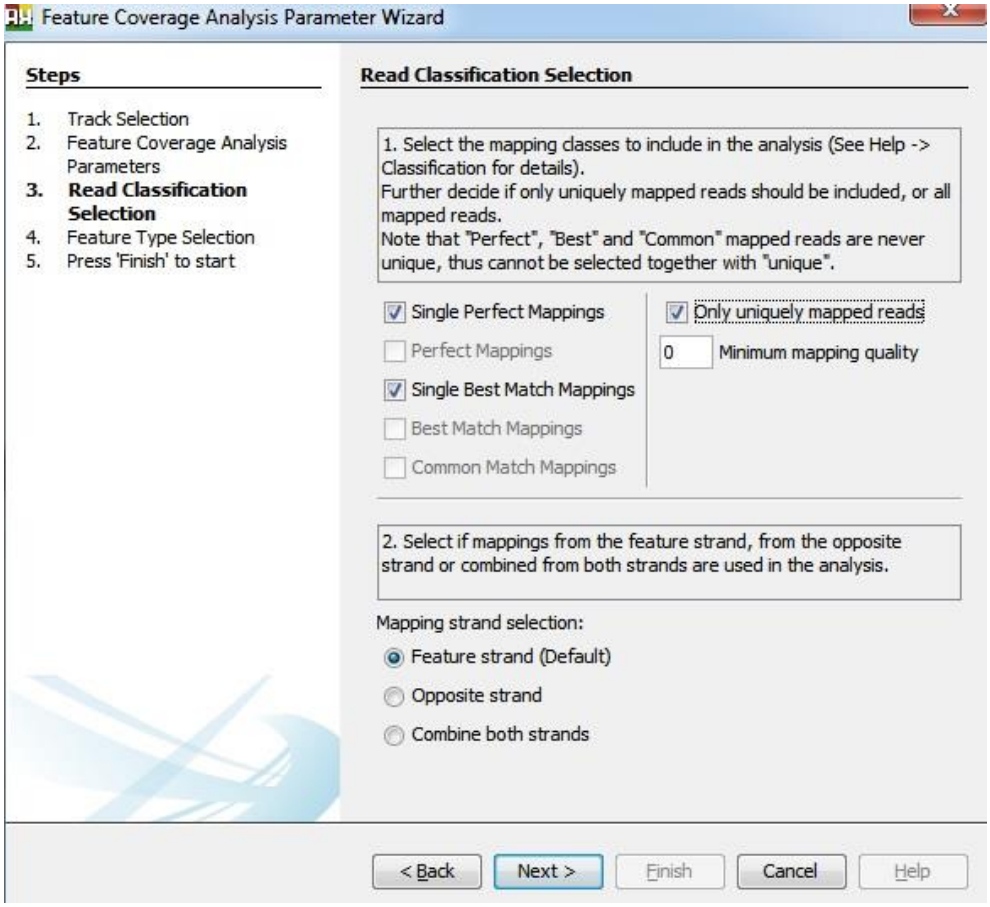
Read Classification

Read mapping data has to be provided in SAM, BAM or SARUMAN (JOK) output format. Thus, the alignment has to be performed beforehand with a mapping tool of choice. ReadXplorer classifies the reads during the import process by mapping quality and read pair concordance. A read mapped to a certain reference position is called "mapping". **For each read, the number of mappings** on the reference is counted along with the **lowest number of mismatches** found among the various mappings of that read. **Uniquely mapped reads** or reads with a certain

amount of mismatches can thus easily be queried.
Note that the same read can map to different reference positions with the same or a varying number of mismatches. This is important for the read mapping classification. **Mappings** are **classified into five classes**:

1. **Single Perfect Match**: Contains all reads which have only one mapping without any mismatches (**green**). Other mappings with mismatches are allowed to exist for a mapping classified as Single Perfect Match.
2. **Perfect Match**: Contains all reads which have more than one mapping without any mismatches (light **green**).
3. **Single Best Match**: Analogue to the Single Perfect Match class. A mapping that maps nowhere else on the reference with fewer mismatches than at the current position is called co-optimal. All reads with only one co-optimal mapping belong to the Single Best Match class (**dark yellow**). Other mappings with more mismatches are allowed to exist for a mapping classified as Single Best Match.
4. **Best Match**: Contains all reads with more than one co-optimal mappings (**bright yellow**).
5. **Common Match**: All remaining mappings belong to the Common Match class (**red**). These are reads, which have a better mapping elsewhere in the genome than at the current position.

Note, that Single Perfect and Single Best Match mappings do not need to be uniquely mapped. **Unique mappings** can easily be accessed by the **mapping count** stored in each mapping for each read. The mapping classes are displayed in the different data viewers (see *ReadXplorer - Main Window j)*) and can be selected explicitly for each analysis as shown in the screenshot below for the Feature Coverage Analysis (see *Genomic Feature Coverage Analysis*). Additionally to the classification, reads can be filtered by minimum PHRED scaled mapping quality. The mapping quality is set by many mapping tools. If the mapping quality is not available, this parameter is dismissed.



Read Pair Classification

For paired end or mate pair mappings a read pair classification algorithm takes into account all occurrences of each read during the import process. In the following we distinguish between **all mappings of the two reads of a read pair, which we call "mapping pair",** and **two mappings of a mapping pair classified as "read pair"**. In general, the algorithm classifies the mappings into three different classes: "Perfect" read pairs with correct orientation and a pair distance within a given range defined by the user (see *Import one or more paired end or mate pair track(s)*), "Distorted" read pairs, whose distance deviates from the perfect distance interval and/or whose orientation is incorrect, and "Single Mappings", whose partner could not be mapped on the reference. A special case of Single Mappings occurs when one or both reads of a mapping pair also map to other regions of the reference, but they cannot be associated to a Perfect or Distorted read pair. There might be more than one Perfect read pair for the same mapping pair. Further, the mapping classes (Single) Perfect, (Single) Best Match and Common are considered in the read pair identification algorithm to improve the accuracy of correct predictions. For each mapping pair, Perfect mapping read pairs are preferred to Best Match and both are preferred to Common mapping read pairs. The read pairs are visualized in a special Read Pair Viewer (see *Read Pair Viewer*).

Create and open/close a Database (DB)

The connection to a ReadXplorer project database (a portable H2 database) can be handled either from the main menu or from the dashboard. All necessary steps are described below.

1. Either click on one of the dashboard buttons "Open existing database" or "Create new database" or choose "File -> Open/Create Database" in the main window.
2. Click "Choose" to select an existing DB or navigate to a folder in which you want to create a new DB. In the latter case: Enter a name for the DB behind the path of the selected folder in the "Select Database" text field. **The DB is then created automatically by ReadXplorer.**
3. Check "Save data" to make ReadXplorer remember your last DB selection.
4. Click finish to open the existing or create and open the new selected DB.
5. To close the currently opened DB, just click "File -> Close Database Connection" or on one of the two dashboard buttons "Close this and open another database" or "Create new database". If you want to directly switch to another DB, click "File -> Open/Create Database".

Import Data

1. Either choose the respective toolbar icon (see *ReadXplorer - Main Window a)*) or go to "File -> Import data".
2. Choose the tab of the data type you want to import. Note that at first a reference has to be imported and then tracks can be imported on that reference.
3. If a reference or track, which has been added to the import list, shall not be imported, just select in the list and click "Remove" in the respective tab.

Import a reference sequence/genome

1. **Make sure that all references have the same locus name as used by the reference file during the mapping process of the tracks, which shall be imported afterwards.**
2. Select the "References" tab and click "Add".
3. Select the file type to import: Genbank, EMBL, GFF3, GFF2 or plain FASTA.
4. Choose a file of the selected file type.
5. Choose a name for the reference.
6. Enter a description of the reference, by default, it is the file name.
7. Click "OK" to add the reference to the import list.

Import mapping data sets

When we talk about "importing" track data (= read mapping data set) here, it means that ReadXplorer reads the corresponding SAM/BAM or SARUMAN (JOK) input file and adds read classification information (see *Read Classification* and *Read Pair Classification*) into a new "copy" of the original input file. The copy contains all original information plus the read classification, carried out by ReadXplorer. During the import process your original SAM/BAM file is never altered. ReadXplorer creates intermediate files during the import process next to the original SAM/BAM file. The temporary directory for these files can be configured via

"Tools->Options->Miscellaneous" (see *General Options*). At the end an extended version of the original file is created in the same directory. Besides the mapping, an index file is created for quicker access of the data.
When a track is opened, all data is read directly from the corresponding extended mapping file. If this file is moved, then ReadXplorer asks for the new path to that mapping file. If the index file should be missing, ReadXplorer creates a new index file.

Mismatching reference identifiers in mappings and reference

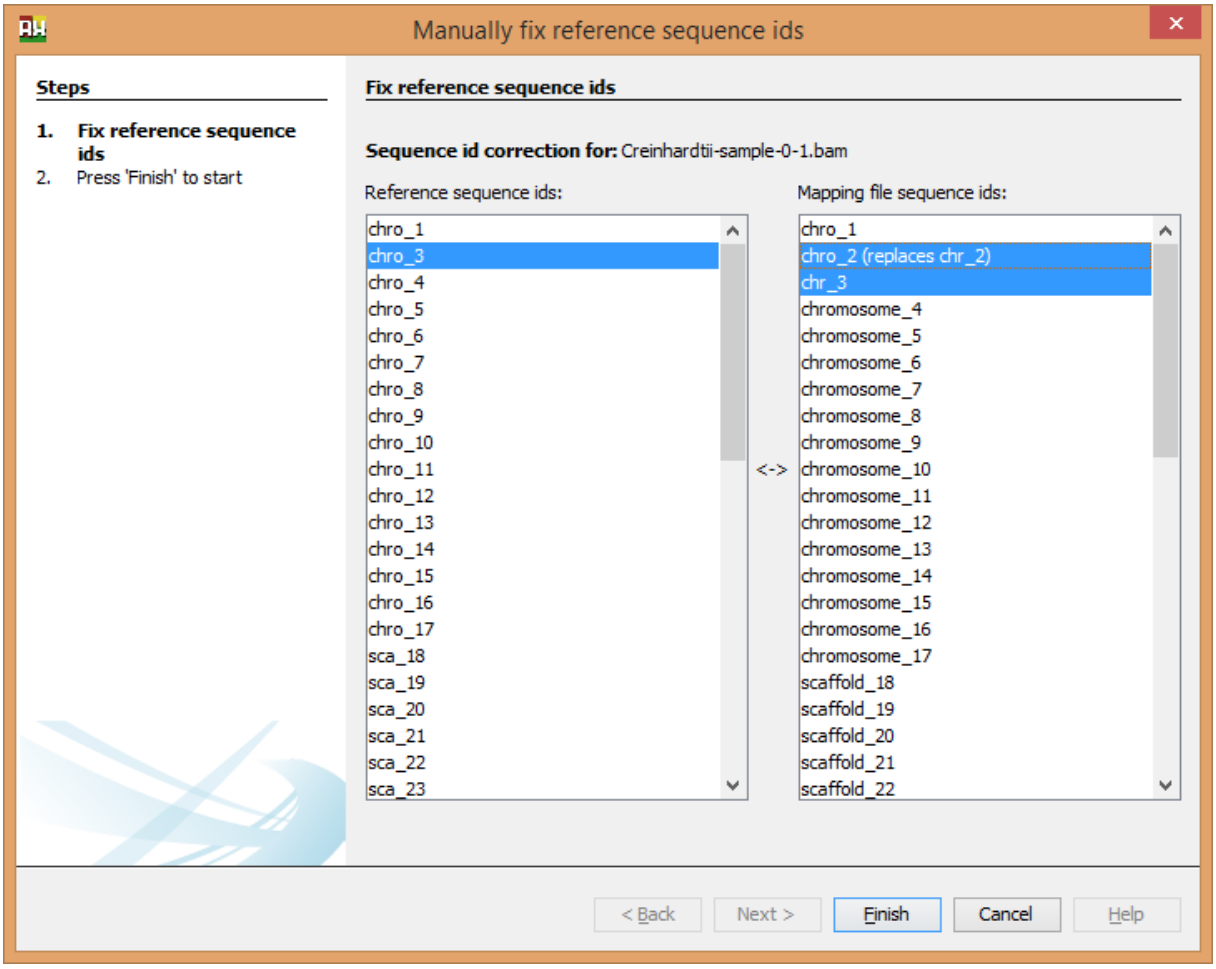
Since version 2.2, ReadXplorer supports automatic and manual correction of mismatching reference identifiers between mapping files and the chosen reference. This can happen for example when downloading the Genbank and Fasta reference files from NCBI. Both files have a different locus tag and sequence identifier respectively. When the identifiers in the mapping file are not adapted, no data will be shown after importing it into ReadXplorer – simply because no data is found for the reference chromosome. The mapping is typically performed on the Fasta file and for downstream analysis in ReadXplorer the Genbank reference is used to have the genome annotations at hand.

To avoid any problems, ALWAYS MAKE SURE THAT BOTH REFERENCE FILES (the annotated and the Fasta file) CONTAIN THE SAME SEQUENCE IDENTIFIERS.

If this is not the case for some reason, the procedure for automatically correcting the sequence identifiers is as follows:

1. The REFERENCE SEQUENCE IDENTIFIERS are NEVER ALTERED, since this would also affect other tracks imported for that reference.
2. For references with only one chromosome, the identifier from the reference is written into the BAM mapping file.
3. For the mapping file ids, several special characters (:,/,\,*,?,|,<,>,") are replaced by '-'.
4. For the mapping file ids, a check is performed if replacement of "chr_x", "chrx", "chromosome_x" or "chromosomex" with each other leads to a match to one of the reference identifiers.
5. Finally, "chr" or "chromosome" prefixes are removed and only the chromosome number is retained and checked against the reference identifiers.

When the automatic correction still fails for some of the references identifiers, a wizard for manually correction opens automatically (see below). When multiple files are imported, all mapping files containing exactly the same sequence identifiers are handled on a single wizard page. If their sequence dictionary is different, a new wizard page is added for each mapping file with a unique dictionary.



The wizard lists all identifiers present in the reference and in the mapping file(s) and enables replacement of ids of the mapping file (right column) by ids present in the reference (left column) via drag and drop. Note that the order of mapping file ids cannot be altered, since this would compromise the assignment between mappings and their ids. Multiple columns can be dragged at once.

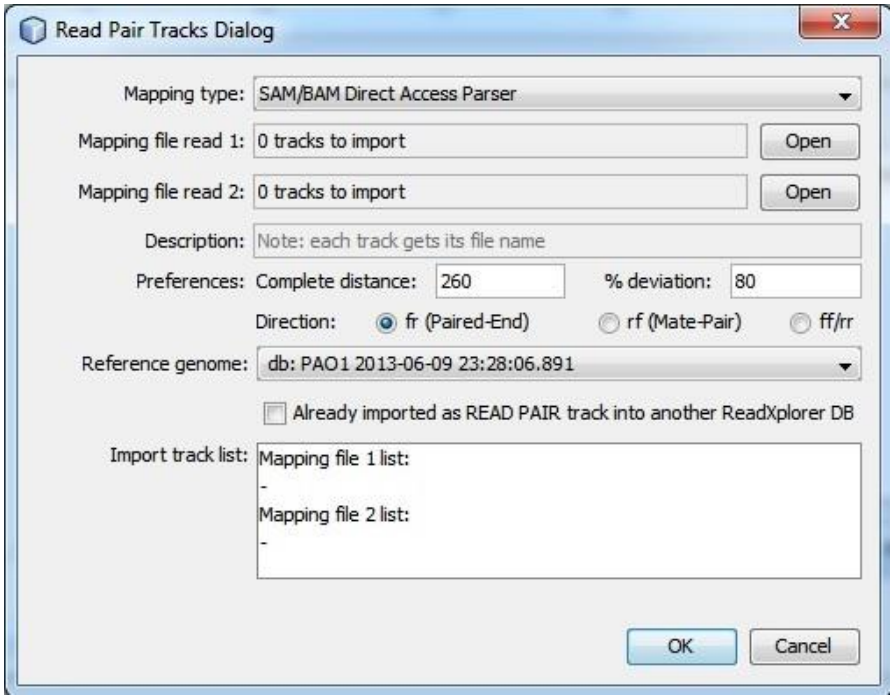
Import one or more single end track(s)

1. Select the "Tracks" tab and click "Add".
2. Select the reference genome, for which the single end track shall be imported.
3. Select the mapping parser to use: For SAM/BAM files use the "SAM/BAM Direct Access Parser" and for SARUMAN mapping output (JOK) use the "Jok to BAM Direct Access Parser".

4. Click "Open" to select the mapping file(s) to import. When importing multiple files, they have to be contained in the same directory. Chosen multiple files are then listed in the "Import track list".
5. Change the description for a single track import.
6. If the specified track(s) have already been used in another ReadXplorer DB, check the box "Was already imported in another ReadXplorer DB". In this case the read classification is already stored in the bam file and does not need to be recalculated. This speeds up the import process.
7. Click "OK" to add the track(s) to the import list.
8. If more tracks from another directory shall be imported, start with 1. again.

Import one or more paired end or mate pair track(s)

1. Select the "Read Pair Tracks" tab and click "Add".
2. Select the mapping parser to use: For SAM/BAM files use the "SAM/BAM Direct Access Parser" and for SARUMAN mapping output (JOK) use the "Jok to BAM Direct Access Parser".



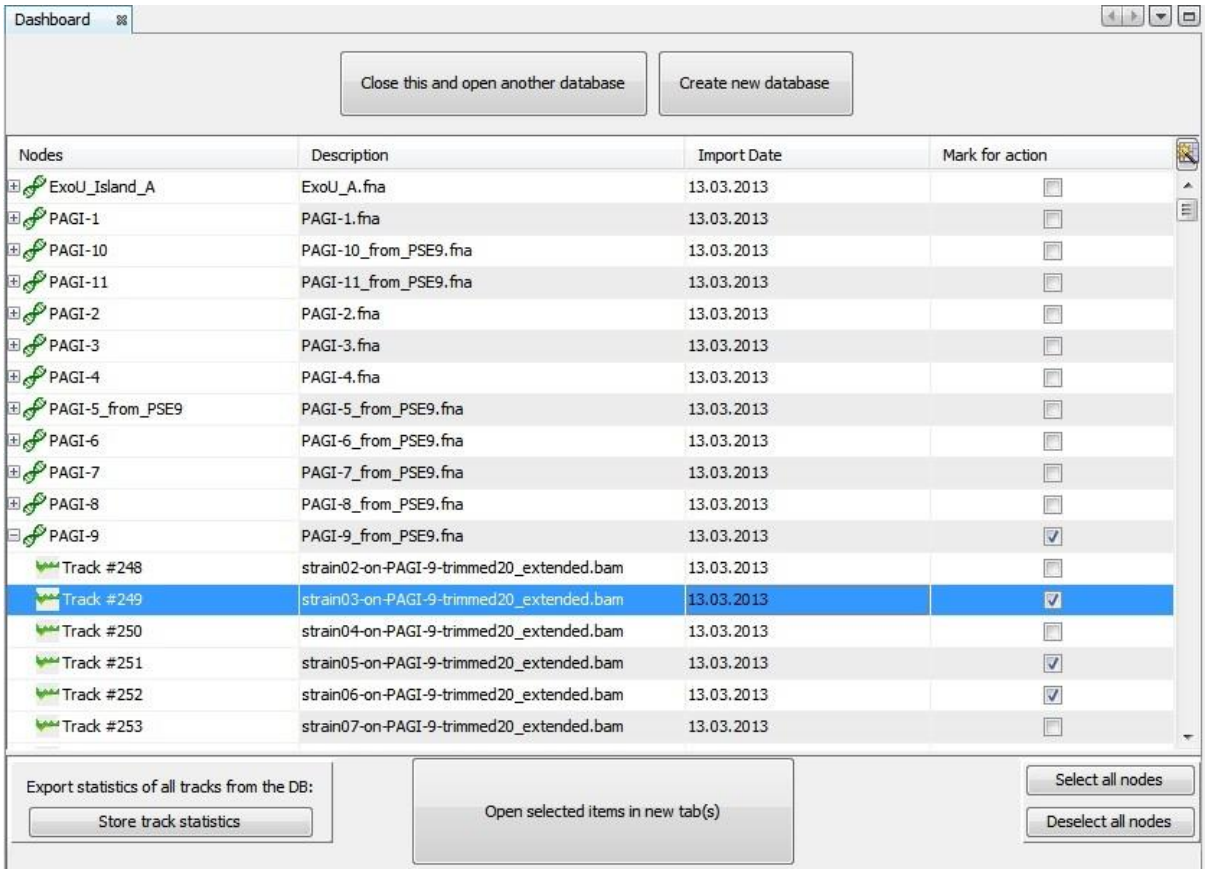
3. Click "Open" to select the mapping file(s) to import. Since read pair mappings are sometimes contained in two separate files (for read 1 and read 2), it is possible to add two mapping files. But only the "Mapping file read 1" has to be selected. The "Mapping file read 2" can be blank if all reads are contained in one file. When importing multiple files, they have to be contained in the same directory. The multiple files are then listed in the "Import track list". **Multiple files for read 1 and read 2 have to be in the same alphabetical order in order to be combined.**
4. Change the description for a single read pair track import.
5. Select the "Complete distance", which is the expected distance of the read pairs.
6. Select the maximum allowed "% deviation" from the "Complete distance" to classify a read pair as perfect.
7. Select the direction of the read pairs: forward-reverse for paired-end, reverse-forward for mate pair or forward-forward/reverse-reverse for other libraries.
8. Select the reference genome, for which the read pair track shall be imported.

9. If the specified read pair track(s) have already been used in another ReadXplorer DB as read pair tracks, check the box "Already imported as READ PAIR track into another ReadXplorer DB".
10. Click "OK" to add the track(s) to the import list.
11. If more read pair tracks from another directory shall be imported, start with 1. again.

Open References and Tracks

There are three ways to open a reference or a track: Either via the Dashboard ("Tools -> Dashboard"), via the first and third toolbar buttons (see *ReadXplorer - Main Window a)*) or via "Visualization -> Open new reference" and then "Visualization -> Open track".

The Dashboard offers to mark each reference and track separately by the "Mark for action" checkbox and also to highlight a consecutive list of tracks and references with a left mouse click on the top item to open, and then, while holding shift, again a left mouse click on the lowest item to open. Holding "CTRL" and then left clicking on items is another way to select multiple items. To open the current selection click the button "Open selected items in new tab(s)". The “Store track statistics”-button (bottom left) allows to store all read mapping statistics for all tracks stored in the current database in an xls or csv file for further use.



Navigation

There are different possibilities to navigate through the reference sequence.

1. Use the scroll bar (see *ReadXplorer - Main Window f)*, which is contained in every viewer, to scroll to the desired position.
2. Use the position navigator: By clicking "Jump to Pos" in the Navigator window (see *ReadXplorer - Main Window b)*), the entered position is then centered for the currently viewed reference.
3. Use panning by left clicking and holding the left mouse button on a viewer and then moving the mouse left or right.
4. Click on a reference feature from the reference feature table (see *ReadXplorer - Main Window d)*).
5. Click on a row in any analysis result in order to center the selected position.
6. Use "Search Pattern" to navigate to the next occurrence of the entered pattern (see *ReadXplorer - Main Window c)*).
7. Right click in a viewer and choose "Center current position" in order to center the position at which the mouse currently hovers.

Besides the navigation, the viewed reference interval can be adapted using the zoom slider (see *ReadXplorer - Main Window f)* , which is contained in every viewer, except the histogram and alignment viewer. Another way of zooming is to move the mouse wheel while hovering the viewer. In this case, when zoomed in, the viewer centers on the currently hovered position.

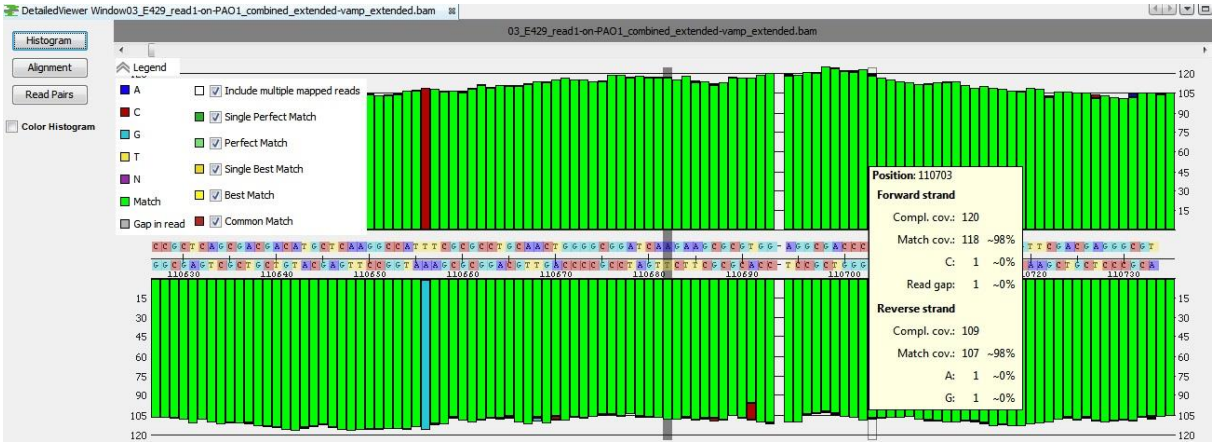
All viewers for one reference genome are always synchronized for easy navigation and comparison of data.
Additionally, each viewer allows copying the current mouse position via right click.

Detailed Viewer

The Detailed Viewer (see *ReadXplorer - Main Window a)*) can be accessed via the respective toolbar icon or "Tools -> Detailed Viewer". It offers three different views on the mapping data: the Histogram Viewer, the Alignment Viewer and the Read Pair Viewer. This viewer can also be used to combine multiple data sets. For details about these viewers have a look below.

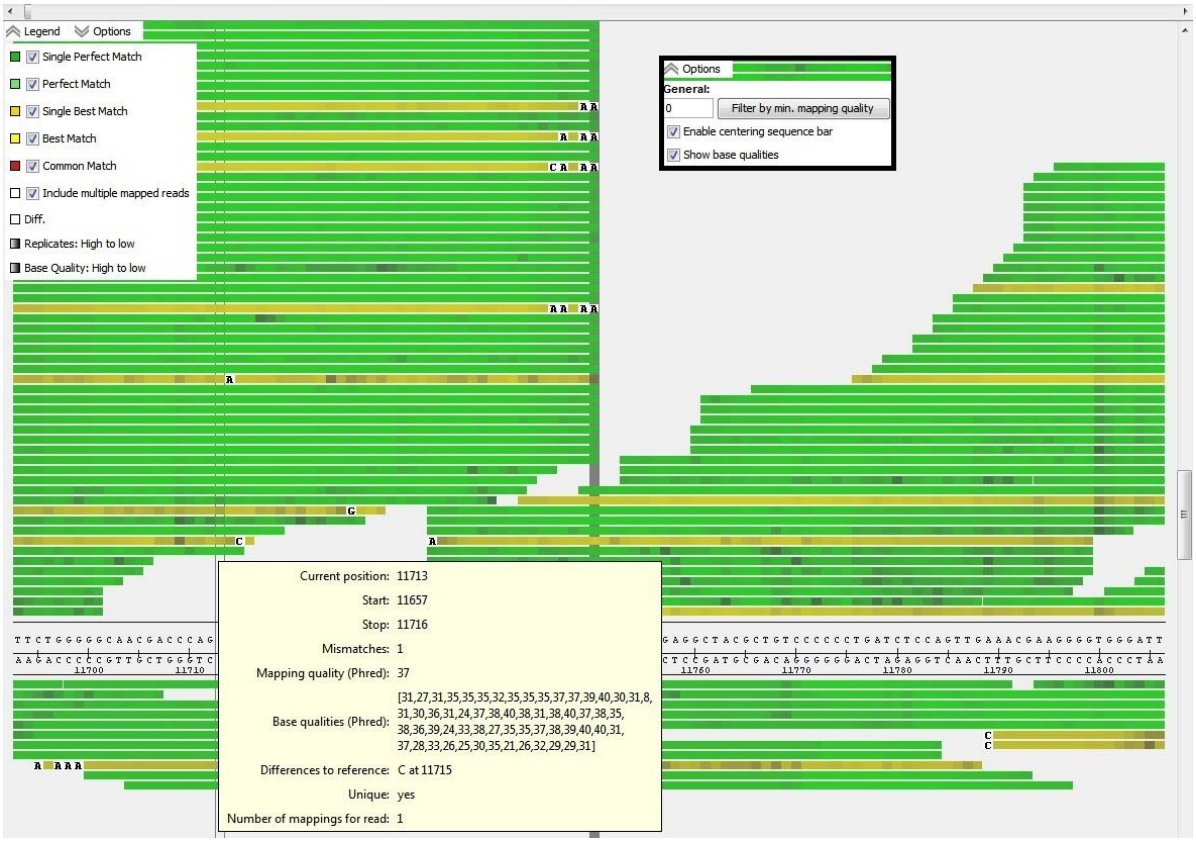
Histogram Viewer

The Histogram Viewer is part of the Detailed Viewer (see *Detailed Viewer*) and supplies intuitive exploration of position specific coverage information as histogram. The match coverage is shown in green and all mismatches and gaps are displayed in a base specific color, which is explained in the legend. In this viewer, the bases on the sequence bar in the middle are also colored according to their respective base. The checkbox "Color Histogram" switches the coloring of the match coverage from green to the base specific color. The tooltip displays the actual base count values at the hovered position.



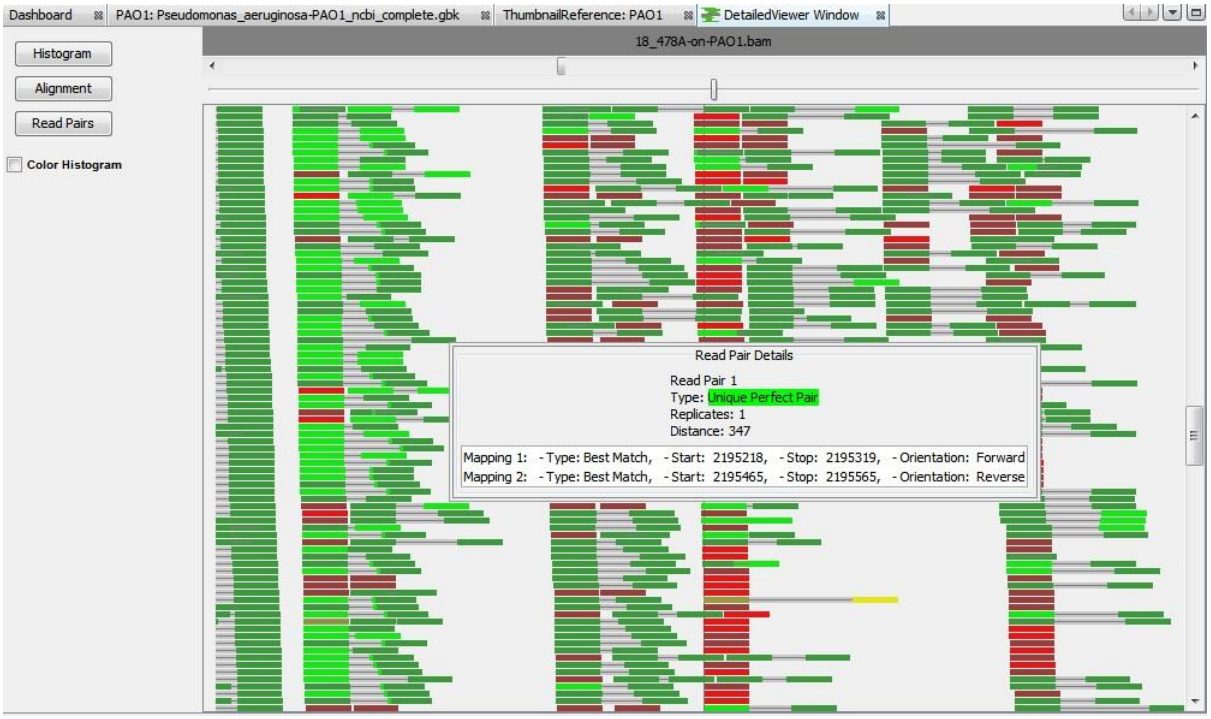
Alignment Viewer

The interactive alignment viewer is part of the Detailed Viewer (see *Detailed Viewer*) and displays each computed read alignment and colors the mappings according to their mapping quality. The coloring is the same as in the Track Viewer (see *Read Classification*). All mismatches and gaps are displayed for each alignment and reference gaps are inserted in the sequence bar in the middle, if a mapping contains insertions. The tooltip always displays the properties of the currently hovered alignment. With a right click on an alignment, the corresponding reference sequence can be copied or the secondary structure of the corresponding sequence can be calculated. The **options panel** (inlet) enables centering the viewer on the sequence bar when scrolling and coloring the alignments by base qualities (Good quality bases are shown in bright colors, while darker positions indicate worse base quality). The height of the alignments in pixels can also be adjusted in this panel. Characters of mismatching bases are only shown when using at least 8 pixels per alignment.

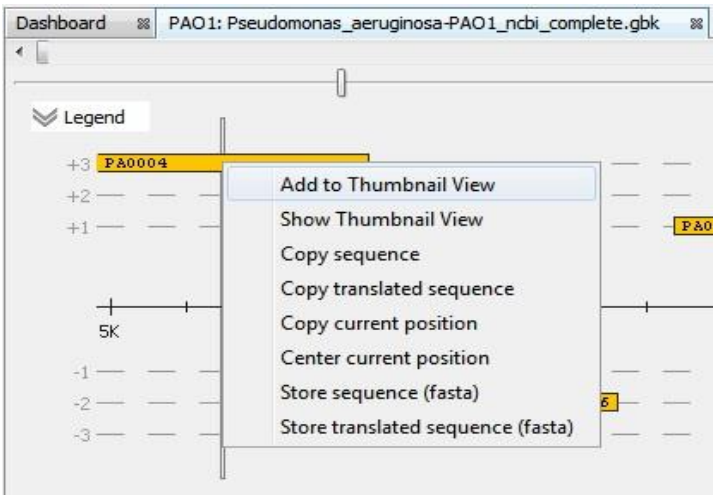


Read Pair Viewer

For paired end or mate pair data, the Read Pair Viewer shows the pair configuration of all aligned reads (see *Read Pair Classification*). It is part of the Detailed Viewer (see *Detailed Viewer*). The reads of a pair are connected via a dark grey line on a light grey background. Perfect pairs are displayed in green, while distorted pairs are yellow and single mappings are red. This coloring allows for easy identification of possible rearrangements, repetitive regions or other variation in the data. Detailed information about each read pair can be obtained by clicking on one of the reads of the pair, as shown on the screenshot below.



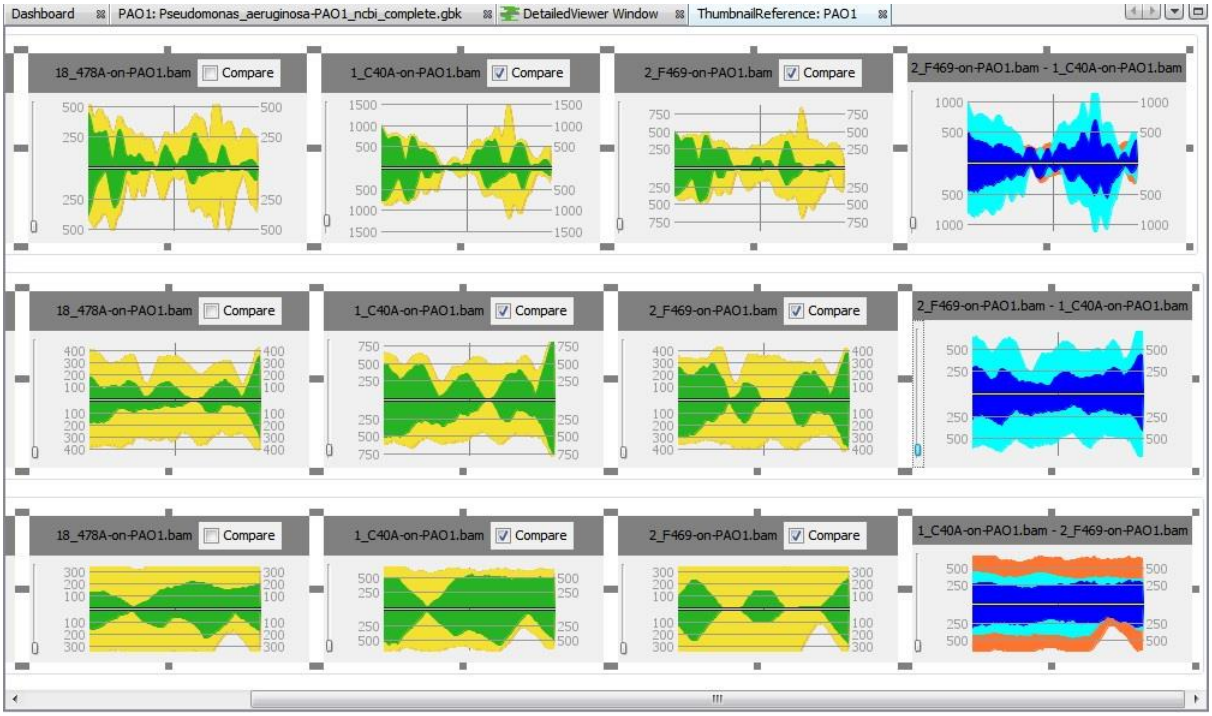
Thumbnail Viewer



The Tumbnail Viewer allows inspecting the coverage of selected features (e.g. genes) among selected tracks. Thus, you can easily compare the coverage between different experiments (different tracks) and draw conclusions much easier, as you get a direct overview of all selected tracks for the selected feature. Of course you can select and view more than only one feature at once.

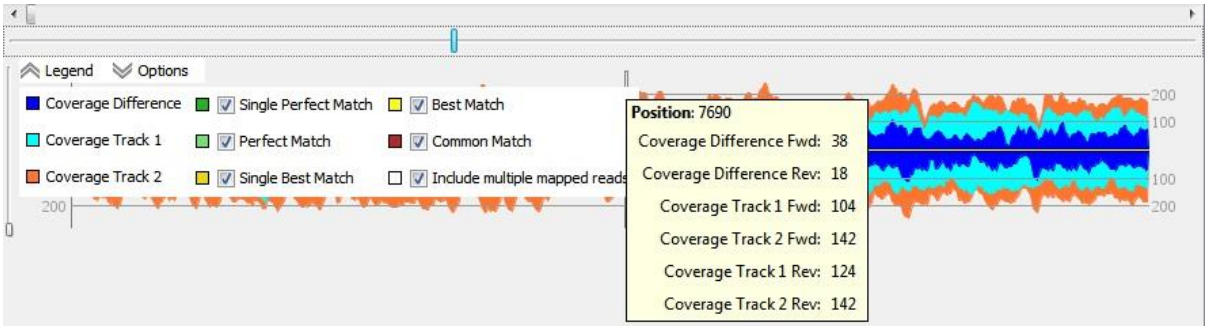
Adding features is done by right-clicking the feature in the Reference Viewer (see *ReadXplorer - Main Window e)*) and choosing "Add to Thumbnail View" as shown below.

The Thumbnail Viewer can be accessed either by right-clicking a feature or choosing the thumbnail view button in the icon bar or the "Tools" menu.
To adjust the size of a Thumbnail Viewer row just click on the grey border squares placed around each small Track Viewer.
By selecting "compare" for two tracks in the Thumbnail Viewer, a Double Track Viewer is opened. This viewer allows to simply compare the coverage of both tracks for the selected reference feature (see *Double Track Viewer*).



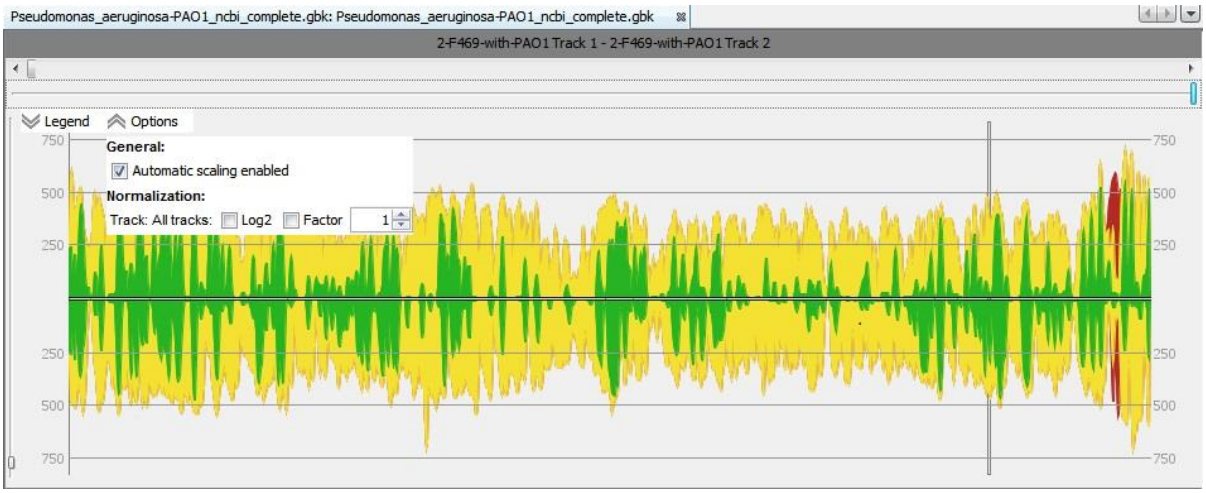
Double Track Viewer

The Double Track Viewer visualizes the coverage differences between exactly two tracks. This viewer uses the total coverage (see *Read Classification*) of the activated read classes for both tracks. It can be accessed via the respective toolbar icon (see *ReadXplorer - Main Window a)*) or "Tools -> Double Track Viewer". In the background the coverage of track one is painted in orange, then the coverage of track two is painted on the middle layer in cyan. On the front layer, the coverage difference is painted in dark blue. In the example, the coverage difference between both analyzed tracks is quite consistent. The tooltip shows the total coverage of both tracks and the coverage difference for the hovered position.



Multiple Track Viewer

The Multiple Track Viewer combines the coverage of an arbitrary number of selected tracks in one data set. Combinations of tracks can also be used for any of ReadXplorer's automatic analysis functions. The Multiple Track Viewer can be accessed via the "Open Track" toolbar icon (see *ReadXplorer - Main Window a)*) or "Visualization -> Open Track". By checking "Combine selected tracks", the Multiple Track Viewer is used instead of the ordinary Track Viewer. Also the *Detailed Viewer* supports track combination.



General Analysis Framework

All integrated analysis methods rely on the mapping classification (see *Read Classification*). All parameters are user adjustable and configured via simple wizards. For references with multiple chromosomes/contigs/other sequences, the complete reference genome is analyzed.
Analyses supported by ReadXplorer are single nucleotide and insertion-deletion polymorphism **(SNP and DIP**) **detection**, a reference **feature coverage analysis,** a general **coverage analysis**, a **genome rearrangement detection**, and **RNA secondary structure prediction**. Especially for **RNA-seq** experiments it offers **differential gene expression**, **transcription start site** (TSS), **novel transcript** and **operon detection** as well as **read count and read count normalization calculations (RPKM and TPM**) per reference feature. All analyses can operate on **combined data sets**. Just select "Combine selected tracks" in the track selection for the analysis.
The results of all these analyses are displayed in form of **tables** which can be **sorted** by each column, **filtered** by column values. Additionally, the reference position of the currently selected result is **centered** in each corresponding data viewer. Below the result table the used **parameters** of the current analysis are shown and a small **statistic** can be viewed by clicking "Show Statistics".
All analysis result tables can directly be **exported** into **Excel (xls)** or **CSV** files. These include an extra sheet with the search parameters, the ReadXplorer version and a small statistic for the performed analysis. When the results are too long for one data sheet, another data sheet is added to the file, until all data is stored in the file.
NOTE: For cells with multi-line entries in Excel:
If the line break is not used by default, you have to turn it on:
Mark all data -> right click data -> Format cells -> Alignment -> Check line break -> Click OK

SNP and DIP Detection

It becomes available after opening a reference. Then it can be carried out for any or multiple tracks, belonging to that reference. In short we just call it SNP detection from here on. An example result is shown in the screenshot below.

1. Click the corresponding toolbar icon (see *ReadXplorer - Main Window a)*) or choose "Tools -> SNP detection".
2. Select the track(s) to analyze with a left click.
3. Select the SNP detection parameters. They are explained in the text field above the parameter selection.
4. Select the mapping classes (see *Read Classification*) to include in the analysis.
5. Select the feature types, for which the codon translation shall be activated. The other features types are ignored and the SNP or DIP is marked as "intergenic".
6. Click "Finish" to start the SNP and DIP detection

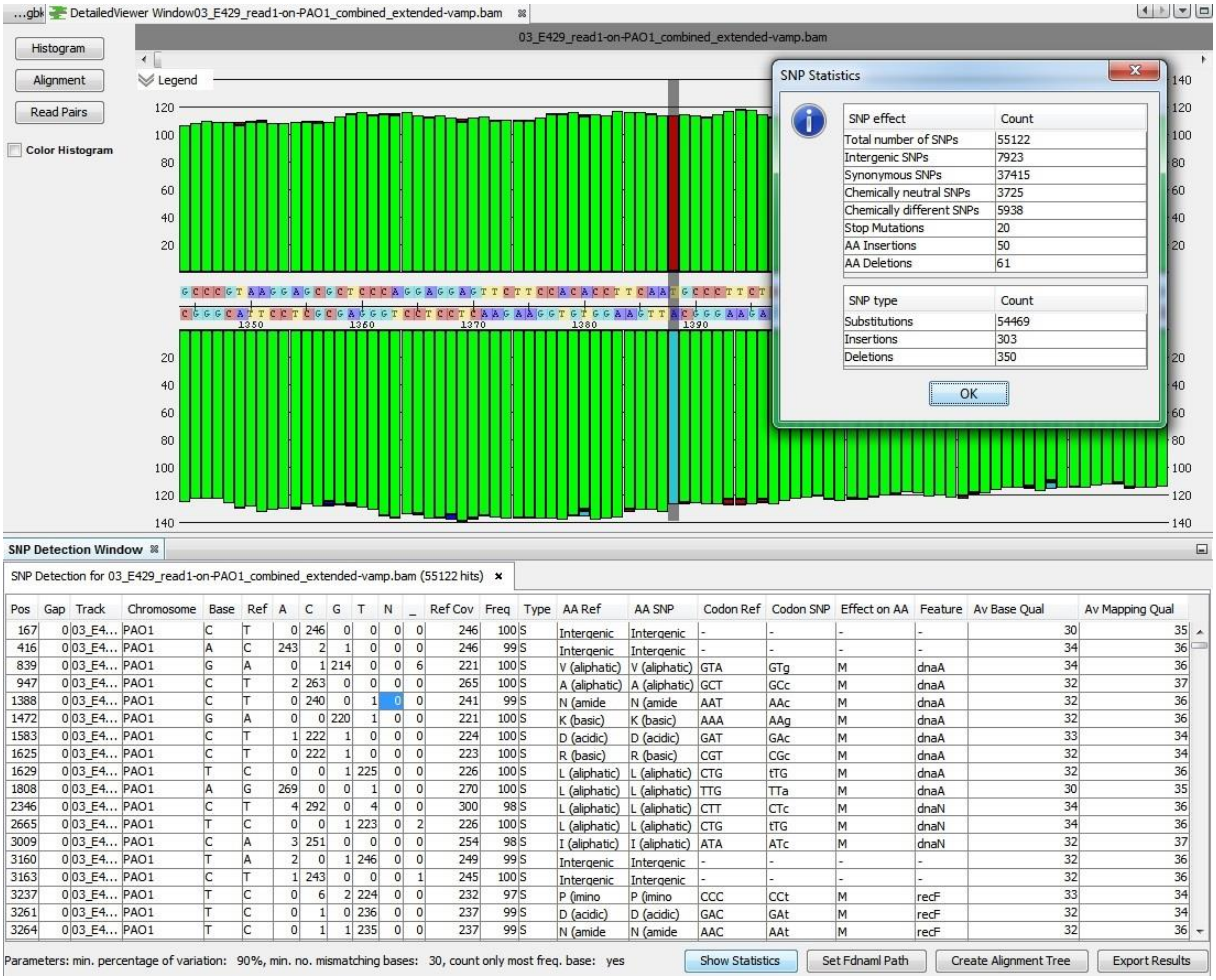
Then the whole bam file(s) is/are analyzed for SNPs and DIPs with the given parameters.

Features:

- Multiple consecutive insertions or deletions are each listed as one result.
- Exon, intron structure is taken into account: If a feature has subfeatures a SNP is only translated, if it is located in an exon or CDS. The translation uses the length of all corresponding subfeatures of the parent feature to determine the correct triplet around a SNP.
- If a feature (e.g. gene) has no subfeatures, then all SNPs in that feature are translated with respect to reading frame and strand.

The Result table contains the following content:

1. Pos: The **absolute position** of the SNP in the reference genome.
2. Gap Index: Value >0 for insertions. Describes the **position of the insertion** in the read for consecutive insertions.
3. Track: The **name** of the track, this SNP is belonging to.
4. Base: The **consensus base** found in the mappings at the current position. Meaning the base with the highest coverage.
5. Ref: The **reference base** at the current position.
6. A/C/G/T/N/ : **Number of occurrences** of each of these bases or a gap at the current position.
7. RefCov: **Complete coverage** of the reference genome by this track at this position. For an insertion the coverage of the position directly after is taken into account.
8. Freq: The **frequency** of all mismatching bases among the complete coverage at the current position. For an insertion the coverage of the position directly after is taken into account.
9. Type: **Effect type** of the SNP. **M** = Match, **S** = Substitution, **I** = Insertion, **D** = Deletion of the reference base.
10. Amino SNP: If the current position lies within a feature (e.g. gene) and is not an insertion or deletion the amino acid is calculated for the triplet at the current position including the SNP base. The **amino acid** is represented in **one-character-code** together with its main **chemical property**.
11. Amino Ref: If the current position lies within a feature (e.g. gene) and is not an insertion or deletion the amino acid is calculated for the reference triplet including the current position. The **amino acid** is represented in **one-character-code** together with its main **chemical property**.
12. Codon Ref: If the current position lies within a feature (e.g. gene) and is not an insertion or deletion the corresponding amino acid codon of the reference is shown here.
13. Codon SNP: If the current position lies within a feature (e.g. gene) and is not an insertion or deletion the corresponding amino acid codon emerging from the SNP is shown here. the SNP position within the codon is depicted by a lowercase letter.
14. Effect on AA: Constitutes the **effect** of the SNP on the amino acid sequence of its protein. **M** = match, **I** = frame shift +1, **D** = frame shift -1, **N** = chemically neutral substitution, **E** = chemically different substitution.
15. Feature (Gene): Either the **name** of the underlying gene, or, if the name is not available, its **locus tag** or **EC number**. Always shown, if the SNP is located in a gene.
16. Average Base Quality: The average PHRED scaled **base quality** for the SNP base in all reads mapped to this SNP position.
17. Average Mapping Quality: The average PHRED scaled **mapping quality** of all reads mapped at the current SNP position.



Pos	Gap	Track	Chromosome	Base	Ref	A	C	G	T	N	_	Ref Cov	Freq	Type	AA Ref	AA SNP	Codon Ref	Codon SNP	Effect on AA	Feature	Av Base Qual	Av Mapping Qual
167	0	03_E4...	PAO1	C	T	0	246	0	0	0	0	246	100	S	Intergenic	Intergenic	-	-	-	-	30	35
416	0	03_E4...	PAO1	A	C	243	2	1	0	0	0	246	99	S	Intergenic	Intergenic	-	-	-	-	34	36
839	0	03_E4...	PAO1	G	A	0	1	214	0	0	6	221	100	S	V (aliphatic)	V (aliphatic)	GTA	GTg	M	dnaA	34	36
947	0	03_E4...	PAO1	C	T	2	263	0	0	0	0	265	100	S	A (aliphatic)	A (aliphatic)	GCT	GCc	M	dnaA	32	37
1388	0	03_E4...	PAO1	C	T	0	240	0	1	0	0	241	99	S	N (amide	N (amide	AAT	AAc	M	dnaA	32	36
1472	0	03_E4...	PAO1	G	A	0	0	220	1	0	0	221	100	S	K (basic)	K (basic)	AAA	AAg	M	dnaA	32	36
1583	0	03_E4...	PAO1	C	T	1	222	1	0	0	0	224	100	S	D (acidic)	D (acidic)	GAT	GAc	M	dnaA	33	34
1625	0	03_E4...	PAO1	C	T	0	222	1	0	0	0	223	100	S	R (basic)	R (basic)	CGT	CGc	M	dnaA	32	34
1629	0	03_E4...	PAO1	T	C	0	0	1	225	0	0	226	100	S	L (aliphatic)	L (aliphatic)	CTG	tTG	M	dnaA	32	36
1808	0	03_E4...	PAO1	A	G	269	0	0	1	0	0	270	100	S	L (aliphatic)	L (aliphatic)	TTG	TTa	M	dnaA	30	35
2346	0	03_E4...	PAO1	C	T	4	292	0	4	0	0	300	98	S	L (aliphatic)	L (aliphatic)	CTT	CTc	M	dnaN	34	36
2665	0	03_E4...	PAO1	T	C	0	0	1	223	0	2	226	100	S	L (aliphatic)	L (aliphatic)	CTG	tTG	M	dnaN	34	36
3009	0	03_E4...	PAO1	C	A	3	251	0	0	0	0	254	98	S	I (aliphatic)	I (aliphatic)	ATA	ATc	M	dnaN	32	37
3160	0	03_E4...	PAO1	T	A	2	0	1	246	0	0	249	99	S	Intergenic	Intergenic	-	-	-	-	32	36
3163	0	03_E4...	PAO1	C	T	1	243	0	0	0	1	245	100	S	Intergenic	Intergenic	-	-	-	-	32	36
3237	0	03_E4...	PAO1	T	C	0	6	2	224	0	0	232	97	S	P (imino	P (imino	CCC	CCt	M	recF	33	34
3261	0	03_E4...	PAO1	T	C	0	1	0	236	0	0	237	99	S	D (acidic)	D (acidic)	GAC	GAt	M	recF	32	34
3264	0	03_E4...	PAO1	T	C	0	1	1	235	0	0	237	99	S	N (amide	N (amide	AAC	AAt	M	recF	32	36

In this example we have centered the marked insertion position in the table by clicking on it. We examine the position in the histogram viewer. We have a look at the statistics window. It shows the number of SNPs and DIPs in the different categories and gives a comprehensive overview for the result.

Transcription Analyses

The transcription analyses module enables you to carry out different automated analyses for transcriptome data sets: **transcription start site detection**, **operon detection**, **Read Count and Normalization (TPM and RPKM) calculations**. These analyses can be carried out in parallel. After deciding which analyses to run, you have to select the included mapping classes (see *ReadXplorer - Main Window a)*) becomes available after opening a reference. Alternatively you can start it via "Tools -> Transcription Analyses".

Transcription Start Site (TSS) Detection

The TSS detection (enabled by checking the "TSS detection" box in the wizard) has two default parameters.

A) The minimal number of read starts at the investigated position (default = 50 reads).

B) The minimal coverage increase in percent from one position to the neighboring one (default = 50%).

In order to output currently unannotated transcript information along with the detected TSS, you can choose the "Novel transcript detection". For detected TSS which do not have neighboring features on their respective strand in the given "Max distance of features to assign to TSS" distance, it adds the information that this is a novel transcript. Further, a transcript stop is inferred ("Cov. Transcript Stop" in the result table) by the coverage threshold "Min transcript extension coverage", which has to be set if the novel transcript detection checkbox is selected. Additionally, the position of the first start codon on the respective strand is calculated and the corresponding CDS up to the next in-frame stop codon is calculated and listed in the table. These entries aid distinguishing real CDS from other small RNAs. The corresponding table columns are:

- "Start Codon Pos"
- "Leader Length": Bp offset from the TSS to the start codon start position
- "Stop Codon Pos"
- "Codon Transcript Length": Bp length of the transcript deduced from the two codons

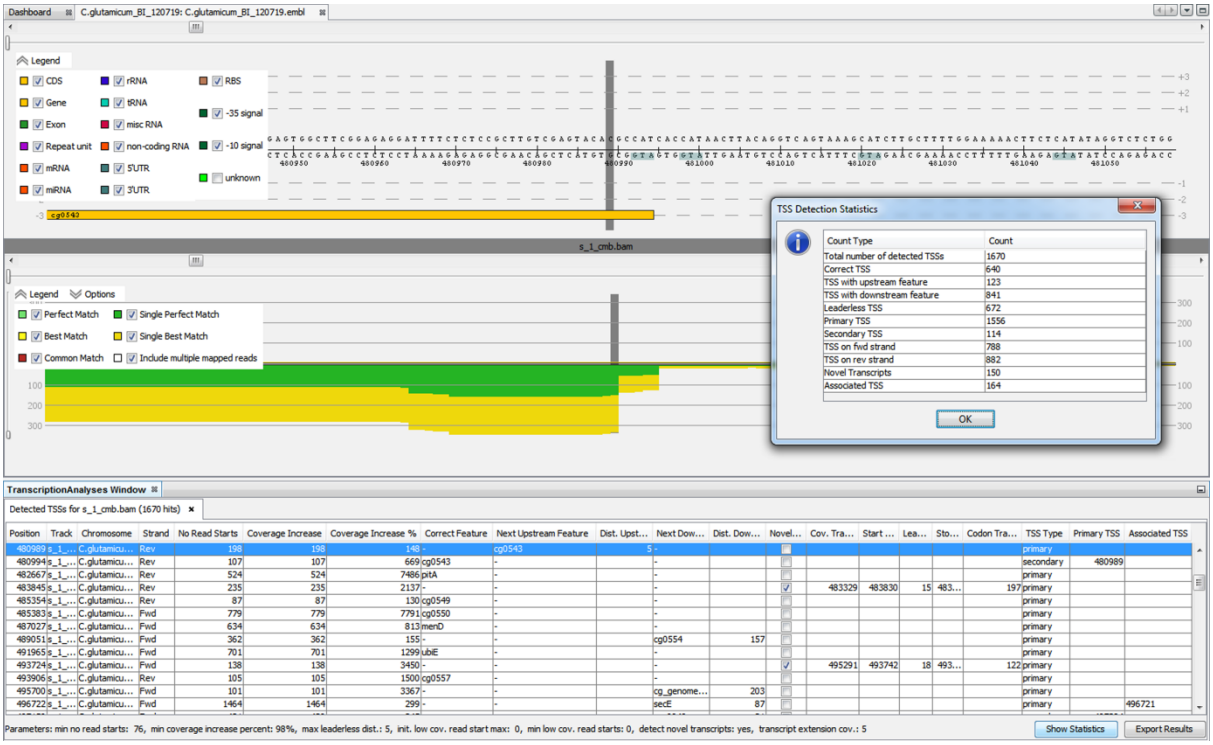
TSSs are also classified as "primary" or "secondary" TSS. This is achieved by flagging the most significant TSS in the chosen maximum distance to genomic features (s. above) as primary and all other valid TSS as secondary.

Additionally, two more parameters can be set for TSS detection in low coverage regions:
C) The Max low coverage read starts (default = 20 reads) can be set
D) The Min number read starts (default = 10 reads) for predicted TSS with less than x (in the default case 20) reads at the TSS position has to be set. It can be the same as the mandatory first parameter, but it is recommended to choose a lower value, depending on your analysis and data quality. If one of these two additional parameters is > 0 the other one also has to be set correctly.

By checking the Automatic parameter estimation box, the parameters are estimated from the data in the track. In this case only the mandatory parameters are set.

Transcripts can be classified as leaderless, if their distance to the next genomic feature is within the base pair window configured via the "Max distance to feature for leaderless transcripts" field.

To prevent receiving multiple TSS for the same TSS area, the check box "Associate all TSS with most significant TSS in this window of bps" has been added. Only the most significant TSS in the entered base pair window (default = 3bp) is kept. All others are associated to this TSS and deleted from the result list. They still appear in the "Associated TSS" column of the result table, thus **ARE NOT LOST**!

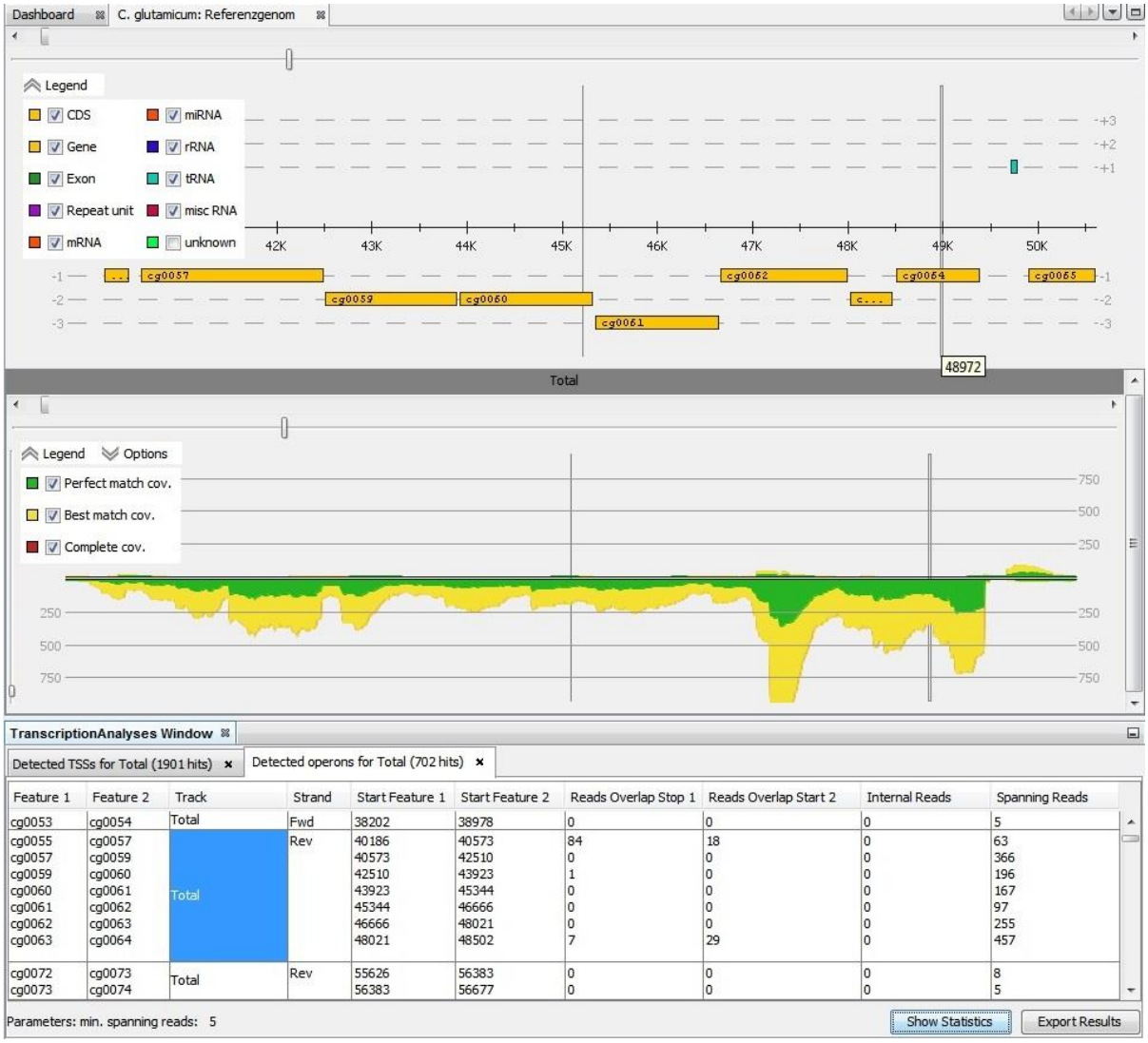


This is an exemplary result of a TSS detection. The highlighted position in the table is centered in the Reference and Track Viewers. The secondary TSS (480994) at the start of the gene has reads (transcripts) starting at exactly that position, so it is annotated correctly and we have a leaderless transcript at the inspected position (a "Leaderless" column only appears in exported tables). But the selected primary TSS (480989) has many more reads starting there. One possibility is that an alternative transcript can be generated from this position on.
The statistics window displays the overall statistics of this TSS detection. TSS with a downstream feature can either have a wrong annotated CDS start, or we can observe a 5'UTR in this experiment.

When exporting TSS tables, the upstream and downstream region of each TSS can be stored. The length of this region is configurable during export.

Operon Detection

The Operon detection allows to predict operons for prokaryotic genomes on the basis of: A) Minimal number of spanning reads (default = 5 reads) between two annotated genes on the same strand. Reads are called "spanning", if they overlap both neighboring features.



This is an exemplary result of an operon detection on *Corynebacterium glutamicum*. The highlighted operon is centered in the Reference and Track Viewers. Here, we can observe, that the coverage continues through all genes from cg0064 to cg0055 on the reverse strand. Therefore, we can presume that they are transcribed in an operon. The table lists each neigboring pair of annotations in one row as "Feature 1" and "Feature 2". In the last four columns we get the information how many reads overlap, span or are positioned between these two annotations.

Read Count & Normalization (TPM & RPKM) Calculations

The read count normalization analysis calculates three values:

- **TPM = Transcripts per million** according to: Li *et al.* 2010.
  The formula is:

  $$\mathbf{TPM} = 10^6 * \left(\frac{c}{l}\right) * \left(1 \Big/ \sum_i \frac{c_i}{l_i}\right)$$ where

  c = number of mappable reads for gene (or genomic feature)
  l = effective length (or length) of gene (or genomic feature)
  i = 1 - #genes (or genomic features of the same type)

- **RPKM = Reads per kilobase per million mapped reads** according to: Mortazavi *et al.* 2008.
  The formula is:
  RPKM = $\mathbf{10^9} * \frac{C}{N*L}$ where
  C = number of mappable reads for feature (e.g. gene)
  N = total number of mappable reads for experiment/data set
  L = sum of feature (e.g. gene) base pairs

- Raw read count values: All values overlapping the feature and used for the formula calculations.

for all annotated features of the user-selected feature types (e.g. CDS or gene), which have a raw read count within the given range by:

1. The minimum number of reads (default = 1 read)
2. The maximum number of reads (default = 100000000 reads)

mapped within the bounds of a feature.

When the **effective length** option is chosen instead of the total feature length (default), the length of each genomic feature is shortened by the mean read length of reads mapping to that feature. Therefore, the effective length represents the number of bases of the feature, where reads can start. Mainly used for eukaryotes. The formula for the effective length is as defined by: Trapnell *et al.* 2010

$$\tilde{l}_i = \sum_{x \leq l_i} \lambda_F(x) * (l_i - x + 1),$$

where x is one of the observed read length values in transcript i of length l_i, and λ_F is the fraction of reads for i with length x.

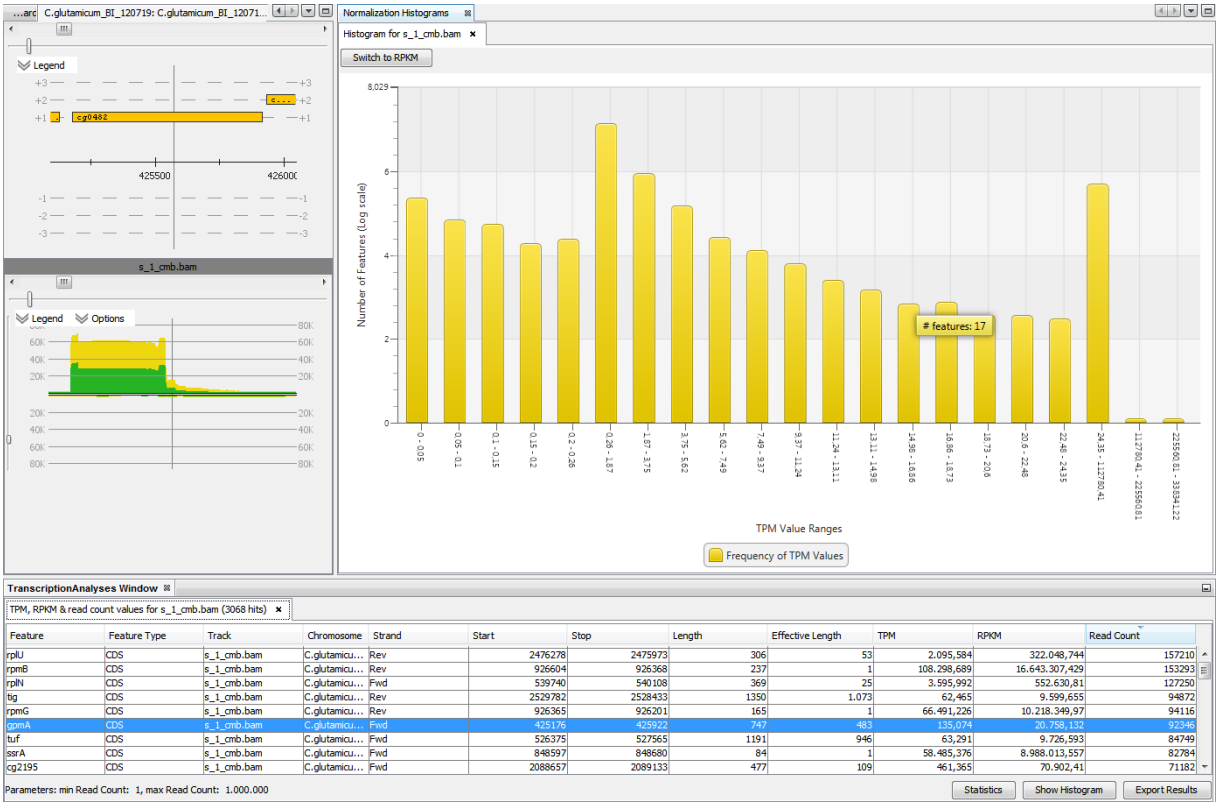
The implemented **assignment model for reads** to features counts a read for a genomic feature if it at least overlaps with the feature. If a read overlaps multiple features:

- If they are sub-features of a larger feature (i.e. CDS and exons of the same gene), they are counted once per gene
- If the overlapping features do not belong together (e.g. a read overlaps two neighboring genes), the read is counted for the first of these features on the respective strand (on the reverse strand this is the last, instead of the first feature).
- If the read is completely contained in multiple features of the same feature type (e.g. for overlapping genes), the fraction of the read is counted for both features (a read completely contained in two genes is thus counted with 0.5 for each of the genes).

This model including the fraction has been chosen, because according to the literature (Li *et al.* 2010), models including read mappings overlapping and contained in multiple features show a better correlation to microarray data than models neglecting such read mappings - thus throwing away real data.

Wizard steps:

1. Click the corresponding toolbar icon (see *ReadXplorer - Main Window a)*) or choose "Tools -> Transcription Analyses".
2. Select the track(s) to analyze with a left click.
3. Select the "Read Count & Normalization calculation".
4. Select the mapping classes (see *Read Classification*) to include in the analysis.
5. Select the Normalization and Read Count Calculation parameters.
6. Select the feature types, for which the Read Count & Normalization Calculation shall be carried out and configure an offset (if desired) for both feature boundaries - increasing the number of bases of a feature. NOTE for eukaryotes: If e.g. "Gene" is selected, but not "Exon" the RPKM and Read Count values still correspond to the exons of that gene and NOT to all reads mapped within the gene's boundaries. If "Exon" or "CDS" are selected e.g. together with "Gene", it means that an extra row for each "Exon" or "CDS" satisfying the given parameters will be shown in the result next to the "Gene" result row. If no "Exon" information is available for a "Gene" it is screened for mRNA/rRNA/tRNA, and only if they are also not present, all reads mapped within the gene's boundaries are counted.
7. Click "Finish" to start the Read Count & Normalization Calculation.



The screenshot shows an Read Count & Normalization Calculation result with calculated TPM, RPKM and Read Count values in the last three columns of the table. If the **effective length** option is enabled, this length is shown in the fourth last column. It represents the calculated length of a feature, at which reads may start. This is estimated in connection with the **mean read length** observed for the feature reads (see above). When the effective length option is disabled, "-1" is printed in this column and "-" is printed in exported tables. The selected gene and its coverage is centered on the left hand side. The histogram shows the distribution of TPM values on the number of genes (log scale). The tooltip displays the actual count of

features which belong to a histogram bar. The histogram can be switched between TPM and RPKM values by the "Switch to ..."-button at the top.
The **total number of assignable mappings** in the analysis statistics corresponds to all mappings which could be assigned to any of the selected feature types. Thus when multiple feature types (e.g. genes and CDS) are selected in one analysis, this value can exceed the number of mappings in the data set. This is due to the fact that feature types on different hierarchy levels (like genes and CDS) are treated separately - each read mapping can be assigned to both a gene and a CDS during such an analysis.

Correlation Analysis

A correlation coefficient can be used to identify regions of two tracks mapped on the same reference showing very similar or completely different coverage progression. The coefficient allows to identify communalities and differences in the data sets automatically.
The correlation of an interval c_i is defined as real number from $-1 \leq c_i \leq 1$, where a $c_i > 0$ indicates a positive correlation between a pairwise measurement x_i, y_i (two coverage values from two data sets for the same interval), whereas a $c_i < 0$ indicates a negative correlation. To give two examples of many possible applications: 2 RNA-seq data sets (e.g. a whole transcriptome and a 5' enriched one) can be compared to identify all TSS regions correlating in both tracks and all genomic regions with correlating expression levels can be retrieved from two whole transcriptome RNA-seq data sets. Two correlation analysis methods are implemented in ReadXplorer:

The **Bravais-Pearson coefficient** is defined as

$$\rho_e(x,y) := \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2\,(y_i - \bar{y})^2}}$$

with $\bar{x}$ and $\bar{y}$ representing the empirical mean. This coefficient requires normally distributed random variables with a linear relationship between each other. If these constraints are not met, the correlation can be underestimated.

Spearman's rank coefficient is defined as

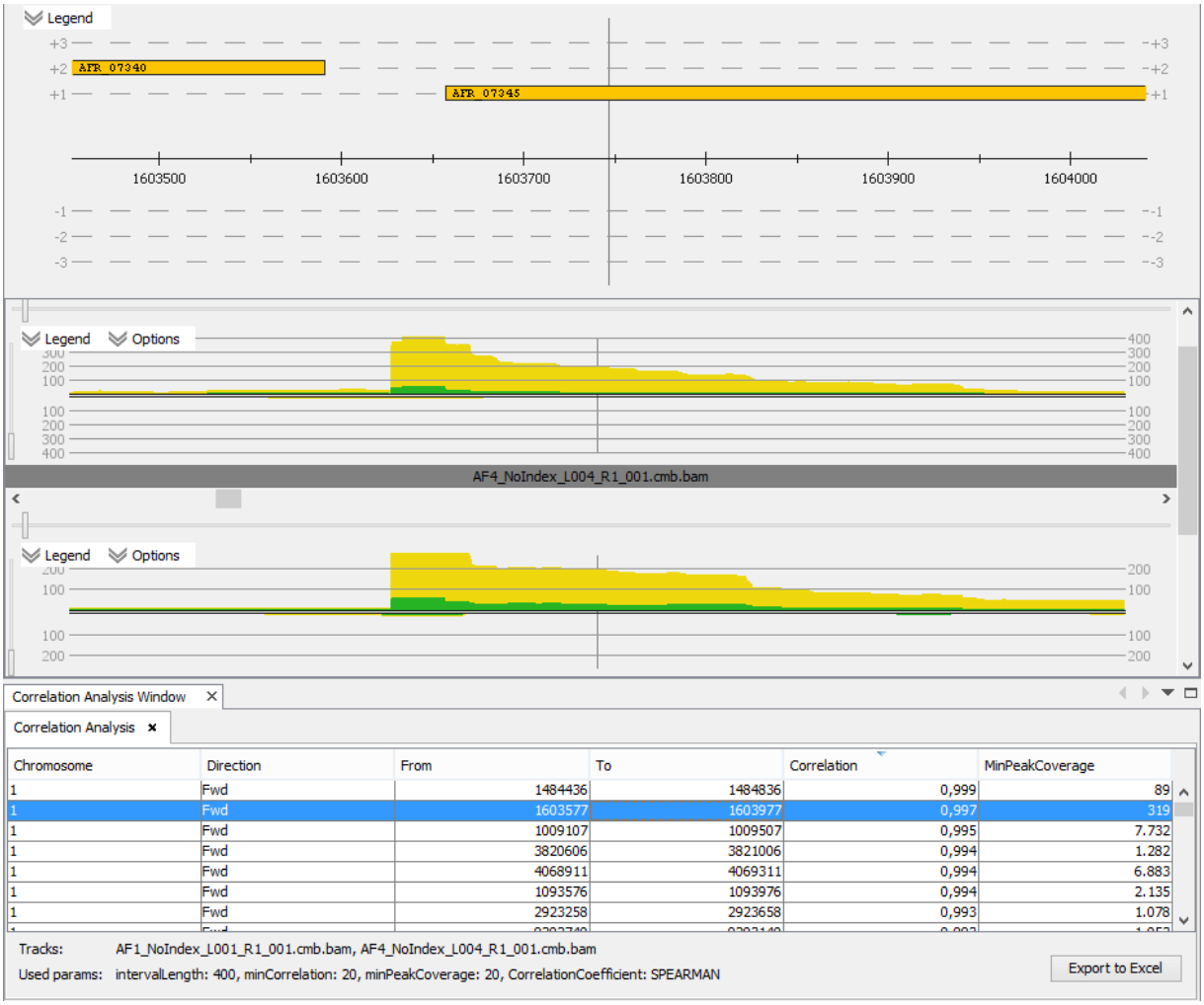
$$\rho_s(x,y) := \frac{\sum_{i=1}^{n}(rg(x_i) - \overline{rg}_x) * (rg(y_i) - \overline{rg}_y}{\sqrt{\sum_{i=1}^{n}(rg(x_i) - \overline{rg}_x)^2} * \sqrt{\sum_{i=1}^{n}\left(rg(y_i) - \overline{rg}_y\right)^2}}$$

With $rg(x_i)$ denoting the rank of x_i. and $\overline{rg}_x$ designating the mean of all ranks of x. Spearman's rank correlation coefficient has the advantage that no assumptions on the underlying probability distribution have to be made and no linear relationship is required between the random variables.

Wizard steps:

1. Select exactly 2 tracks (the track combination will be ignored).
2. Select the correlation coefficient (PEARSON or SPEARMAN), the minimum percent of correlation to list an interval in the result, the minimum peak coverage for both tracks to consider an interval for the result (For each correlated interval the software will determine the peak coverage for both tracks. The minimum of these two values will be compared with the value selected here) and the length of the correlated intervals between both tracks (e.g. mean read length).
3. Select the read classification (see *Read Classification*) to include in the analysis.

An exemplary result is shown below:



The example shows a correlation analysis result sorted by the correlation column. The selected interval shown in the corresponding track viewers has very high positive correlation between both bacterial RNA-seq data sets.

Differential Gene Expression Analysis

The Differential Gene Expression Analysis enables you to automatically collect all necessary count data for an experiment and run a differential gene expression analysis tool of your choice. ReadXplorer offers DESeq, DESeq2, baySeq and the Express Test. Additionally, the same wizard offers export of count data tables. Exemplary results are shown in the screenshots below. For further details about the methods, manual and plots generated by the different tools, please visit the respective tool homepage linked above.

1. Choose "Tools -> Differential Gene Expression Analysis".
2. Select the tool for the analysis (DESeq, baySeq or Express Test).
3. Setup the analysis design. The different steps are briefly explained on each wizard page and in detail in the corresponding list below.
4. General setup:
    - Select the genomic feature types, to take into consideration for the analysis. The other features types are ignored.
    - Set the offset for the start and stop positions of each genomic feature. Very useful to include otherwise ignored reads for genomic features whose annotation starts further downstream of the actual start position observed in the RNA-Seq data.
    - Decide, if the read orientation is taken into account (only reads on the feature strand are taken into consideration), or if all reads from both strands within range of a genomic feature are used.
    - Only for DESeq and baySeq: Decide, if you want to store the parameters and results generated from the used R tool for further processing.
5. Select the mapping classes (see *Read Classification*) to include in the analysis.
6. Click "Finish" to start the differential gene expression analysis.

Wizard steps DESeq (two condition design):

1. Define the experiment design: standard two condition design or multifactor design (for further information see DESeq manual).
2. First select the reference to use in the upper field and afterwards all tracks to include in the analysis in the second field.
3. Assign each track to one of the two conditions by using the arrows next to the two condition fields. ReadXplorer automatically identifies the best mode to estimate the dispersion values of the data for the given setup (For details see DESeq manual).
4. Continue at "4. General setup" in the general instruction list above.

Wizard steps DESeq2:

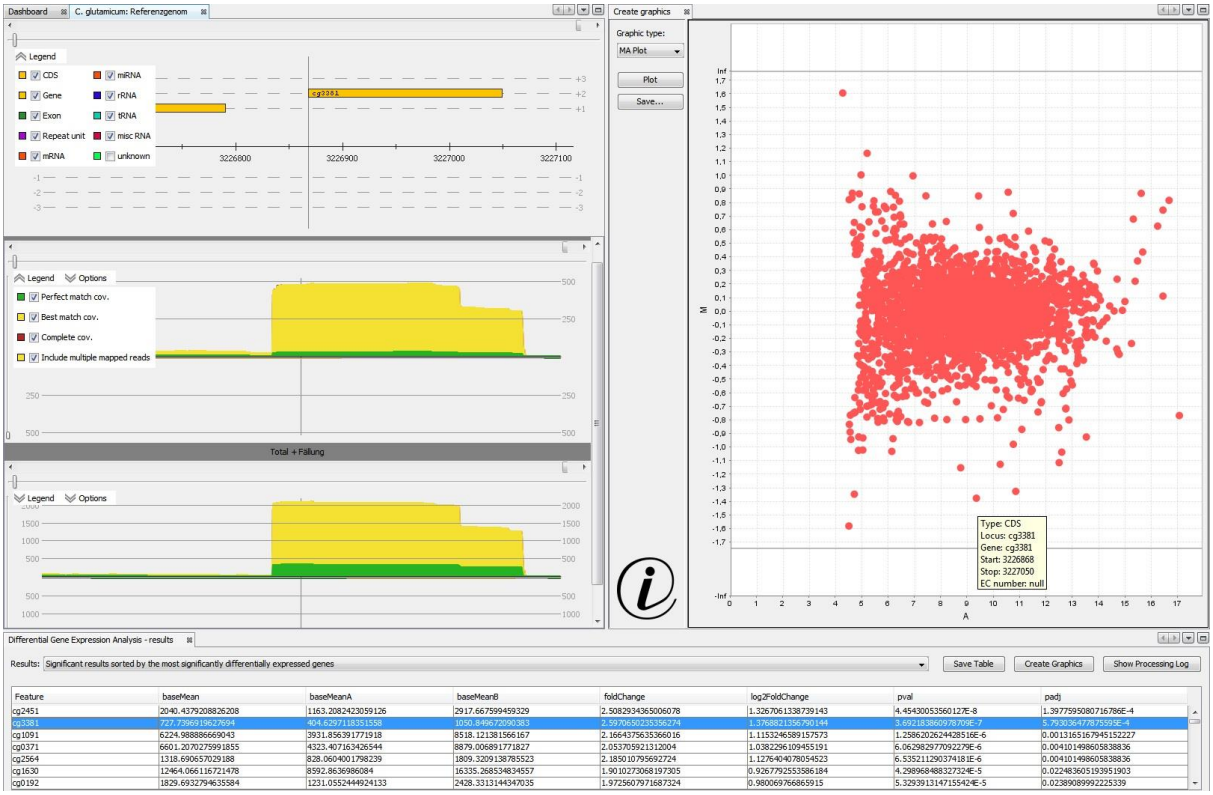
1. First select the reference to use in the upper field and afterwards all tracks to include in the analysis in the second field.
2. Assign each track to one of the two conditions by using the arrows next to the two condition fields.
3. Continue at "4. General setup" in the general instruction list above.

Wizard steps baySeq:

1. First select the reference to use in the upper field and afterwards all tracks to include in the analysis in the second field.
2. Define replicate structure: At first select all tracks, which are replicates of the first condition. Then click "Add as replicates". Continue with the next conditions in the same way until all tracks are assigned into the replicate structure. If only one track is available per condition, just select one per condition.
3. Create baySeq expression models, which represent the experiment design (for further details see baySeq manual). For analysis of a standard two condition experiment, only one model separating the two conditions is needed. Similarly to step 2, first select all tracks belonging to the first experimental condition of your first expression model, then click the arrow button and continue with all other conditions of the model for all other tracks in the same way until all tracks are assigned to this first model. Then click "Add model". Now you can create more models, depending on the experiment design. A model containing all tracks (control, which assumes that no differential expression is present in the data set) is created automatically.
4. Continue at "4. General setup" in the general instruction list above.

Wizard steps Express Test:

1. First select the reference to use in the upper field and afterwards all tracks to include in the analysis in the second field.
2. Assign each track to one of the two conditions by using the arrows next to the two condition fields.
3. First continue at "4. General setup" in the general instruction list above. Afterwards continue here.
4. Select normalization of the read count values: The normalized read count can be calculated automatically across all genes or house keeping genes can be selected for normalization (hold CTRL while clicking a gene).
5. Continue at step 5 in the general instruction list above.



This example shows a differential gene expression result using DESeq. The result table shows all values calculated by DESeq for all genomic features of interest. The table can be sorted by the most significantly differential, up- or down-regulated genes. The "Save Table" button stores the table as an Excel sheet. The "Show processing log" button opens another tab with the log of the DESeq run. The "Create Graphics" button opens the graphics tab, which allows to generate, view and store different plots, depending on the selected analysis tool. Here, an interactive (denoted by the "i"-icon at the bottom left) M/A plot of the results is shown. All interactive plots allow centering a gene of interest in all synchronized viewers by simply clicking on it. This function enables direct close inspection of the actual underlying data. Each plot can also be saved for further use by clicking "Save...".

A brief explanation of the different plots:

1. DESeq:
    - M/A-plot: An interactive M/A plot similar to DESeq's M/A plot (Log 2 fold change vs. base means) created by ReadXplorer. On the x-axis, the plot visualizes A (normalized mean expression (0.5 * log2 (baseMeanA * baseMeanB))), while the y-axis plots M (log2 fold change). Each dot in the plot can be hovered and selected by left clicking. This action centers the synchronized viewers on the selected gene of interest, enabling visual inspection of the underlying data. A gene is more likely to be differentially expressed, if it is located farther away from 0 on the y-axsis and has a higher normalized mean expression.
    - Gene dispersion vs. normalized mean expression: This plot contains the empirical per gene dispersion values (black dots) on the y-axis and the corresponding fitted dispersion values as a red line plotted against the mean expression strength of each gene on the x-axis. The plot is doubly logarithmic.

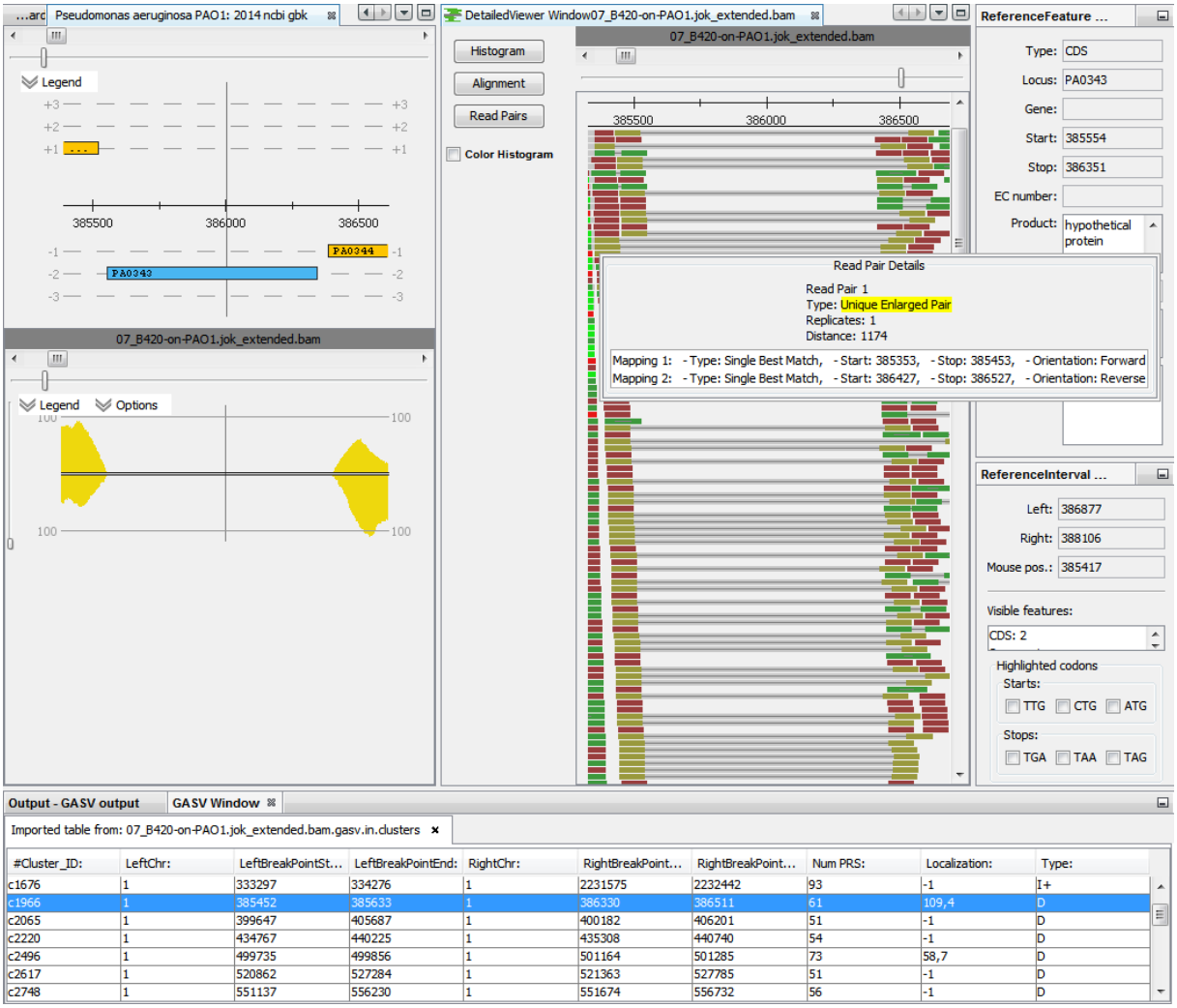
For details on gene dispersion please read the DESeq publication and the DESeq manual.

- Log 2 fold change vs. base means: DESeq's M/A plot. It contains normalized mean expression (A) against log2 fold change (M) for each gene.
- Histogram of p-values: This plot visualizes the probability of genes not to be differentially expressed against its frequency in the experiment.

2. DESeq2:
    - M/A-plot: See the description above for DESeq.
    - Per gene estimates against normalized mean expression: This plot contains the empirical per gene dispersion estimates (black dots) on the y-axis, the corresponding fitted dispersion values as a red line and the final maximum *a posteriori* estimates used in testing (blue) plotted against the mean expression strength of each gene on the x-axis. The plot is doubly logarithmic. For details on gene dispersion please read the DESeq2 publication and the DESeq2 manual.
    - Histogram of p-values: This plot visualizes the probability of genes not to be differentially expressed against its frequency in the experiment.
3. baySeq:
    - Priors: The log prior probability means (see Prior probability on Wikipedia) for a selected sample group.
    - M/A plot for count data: Where the log-ratio would be infinite, because the data in one of the sample groups consists entirely of zeros, we plot instead the log-values of the other group.
    - Posterior likelihoods of differential expression against lot-ratio: Posterior likelihoods of differential gene expression against log-ratio (where this would be non-infinite) or log values (where all data in the other sample group consists of zeros).
4. Express Test:
    - M/A plot: Shows the same data as the interactive M/A-plot of DESeq.
    - Ratio plots: These two plots visualize the gene ratio of AB or BA on the x-axis, while the y-axis shows the calculated confidence value. A higher confidence (top) depicts a higher probability, that the result is trustworthy.

Genome Rearrangement Detection

An integrated version of the command line tool GASV 2.0 enables the detection of genome rearrangements / structural variants for read pair data. All options of GASV are accessible via the analysis wizard (see http://gasv.googlecode.com/svn/trunk/doc/GASV_UserGuide.pdf for available options). The analysis is configured to only use Single Perfect and Single Best Match mappings to minimize false predictions.



#Cluster_ID:	LeftChr:	LeftBreakPointSt...	LeftBreakPointEnd:	RightChr:	RightBreakPoint...	RightBreakPoint...	Num PRS:	Localization:	Type:
c1676	1	333297	334276	1	2231575	2232442	93	-1	I+
c1966	1	385452	385633	1	386330	386511	61	109,4	D
c2065	1	399647	405687	1	400182	406201	51	-1	D
c2220	1	434767	440225	1	435308	440740	54	-1	D
c2496	1	499735	499856	1	501164	501285	73	58,7	D
c2617	1	520862	527284	1	521363	527785	51	-1	D
c2748	1	551137	556230	1	551674	556732	56	-1	D

Since the tool assumes that chromosome designations are numbers and this is not the case for prokaryotes, ReadXplorer utilizes the GASV option to set a chromosome naming file to use arbitrary chromosome designations. GASV writes its results into a tab separated file in CSV format on the hard disk. The content of this file is immediately visualized in ReadXplorer after finishing an analysis. The only difference is that the left and right breakpoint borders have been split into two columns to increase readability and simplify filtering and sorting of the result table.

The figure shows an exemplary result of a GASV analysis of *P. aeruginosa* read pair data. The region selected in the GASV result table is automatically centered and reveals that this particular region from the reference PAO1 is deleted in the strain B420. The Read Pair Viewer enables a detailed inspection of the region.

Genomic Feature Coverage Analysis

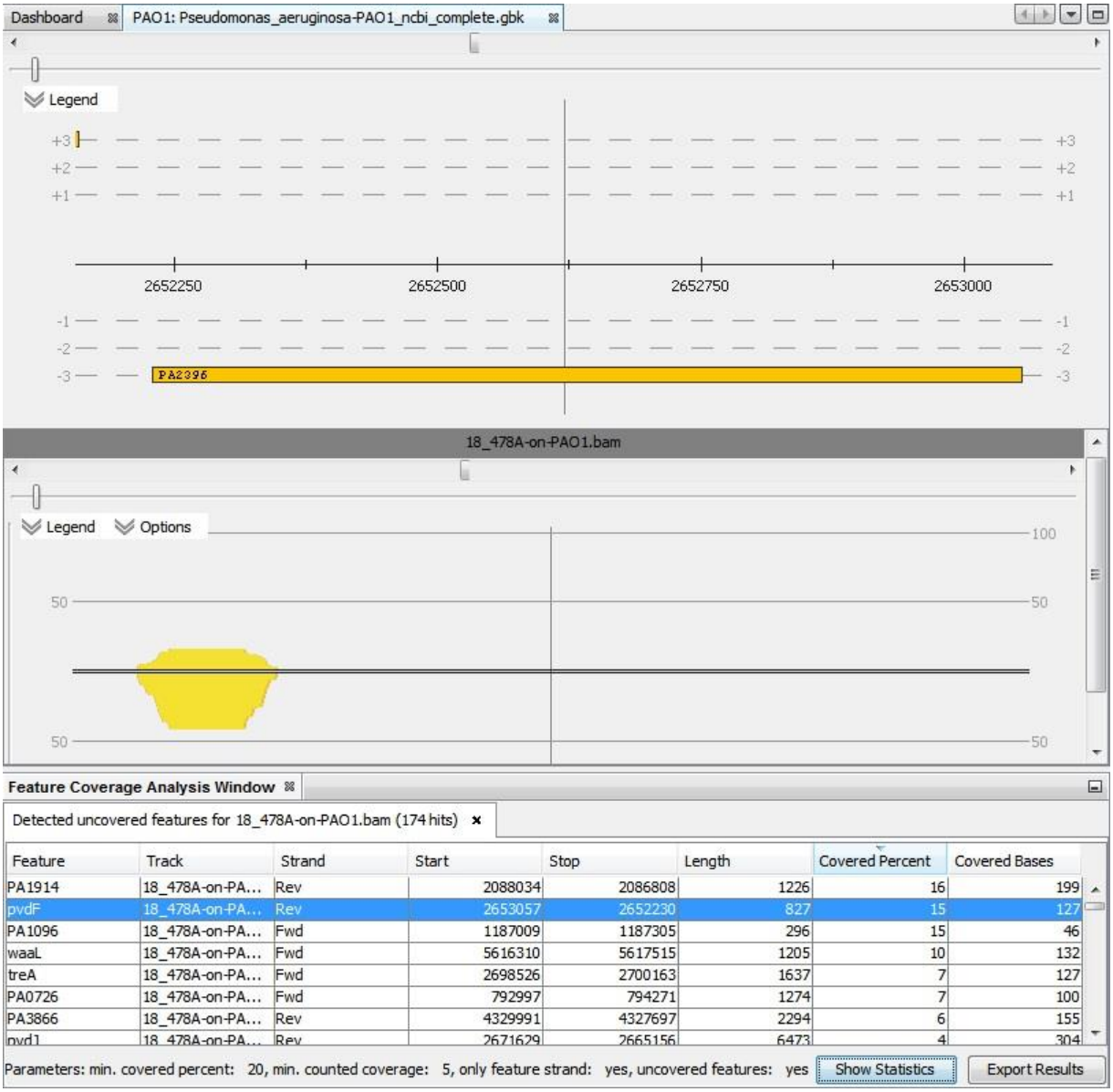
The Feature Coverage Analysis empowers you to detect all reference features that show predefined characteristics in terms of their coverage. It becomes available after opening a reference. Then it can be carried out for any or multiple tracks, belonging to that reference. An example result is shown in the screenshot below.

1. Click the corresponding toolbar icon (see *ReadXplorer - Main Window a)*) or choose "Tools -> Feature Coverage Analysis".
2. Select the track(s) to analyze with a left click.
3. Select the analysis parameters. They are explained in the text field above the parameter selection and in the paragraph below.
4. Select the mapping classes (see *Read Classification*) to include in the analysis.
5. Select the feature types, for which the feature coverage analysis shall be carried out. The other features types are ignored.
6. Click "Finish" to start the feature coverage analysis.

The first parameter defines how many percent of a feature have to be covered with at least a coverage larger or equal to the second parameter value to be detected as "covered feature". The "Only count mappings on the feature strand" checkbox decides whether only the mappings on the strand of a feature are counted, or all mappings within the range of the feature.
The "Detect uncovered instead of covered features" checkbox switches the mode from detecting features which have a satisfactory amount of reads to the contrary. Then all reference features are listed, which have a lower coverage than defined by the percentage threshold.

This analysis is useful for studies of both, RNA-Seq and resequencing data sets. For RNA-Seq experiments, it facilitates the identification of genes which exhibit a certain minimum coverage or which are not expressed at all. For resequencing experiments it allows to explore the set of common or divergent features between the reference and its tracks and the tracks among one another.

In the following example, features with coverage of less than 20% were queried. The minimum coverage to count for a covered position was 5 and only mappings on the strand of the analyzed feature were taken into consideration. The used parameters are also shown below the result table.



General Coverage Analysis

The Coverage Analysis empowers you to detect all reference intervals that show predefined characteristics in terms of their coverage. It becomes available after opening a reference. Then it can be carried out for any or multiple tracks, belonging to that reference. An example result is shown in the screenshot below.

1. Click the corresponding toolbar icon (see *ReadXplorer - Main Window a)*) or choose "Tools -> Coverage Analysis".
2. Select the track(s) to analyze with a left click.
3. Select the analysis parameters. They are explained in the text field above the parameter selection and in the paragraph below.

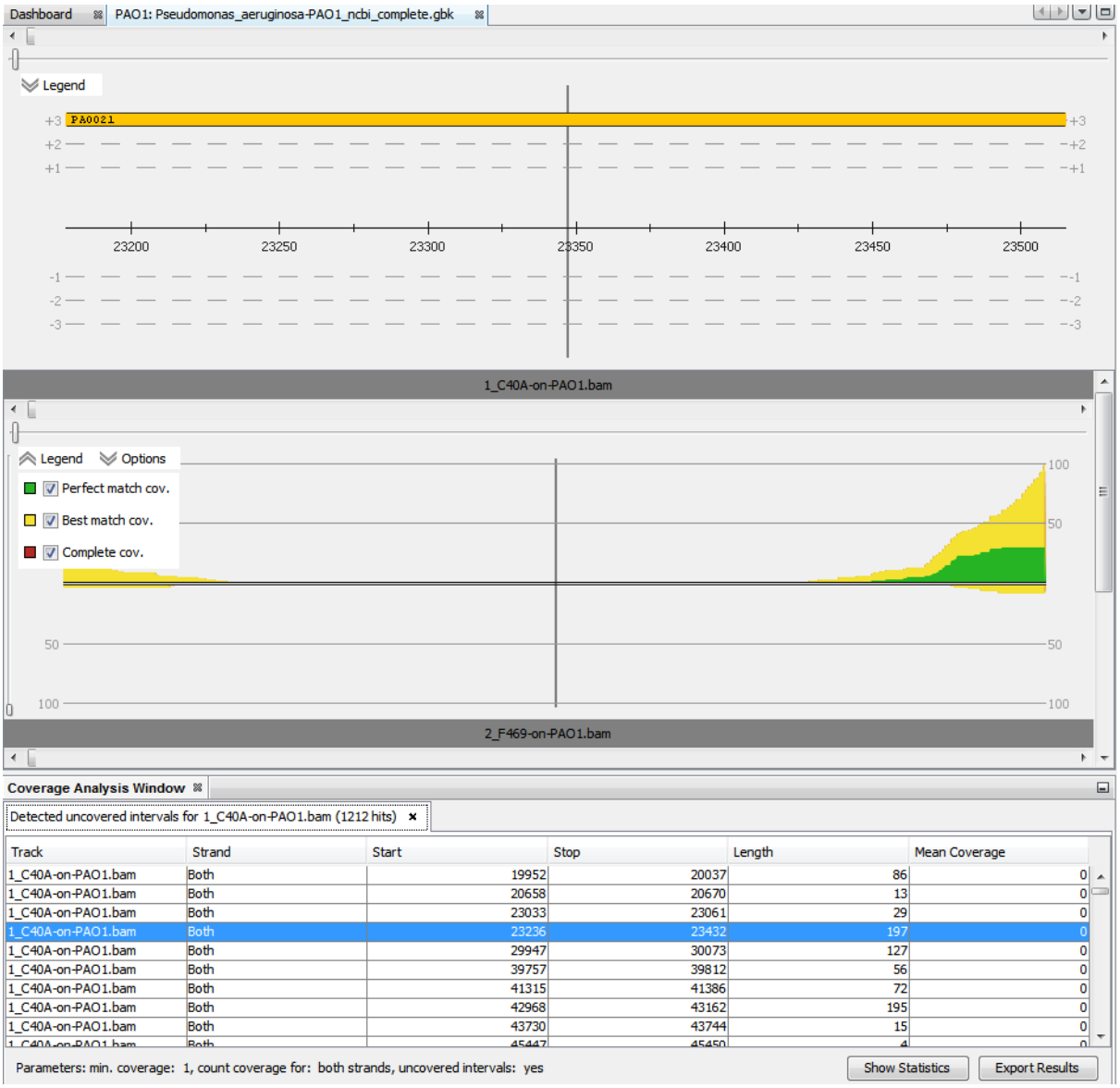
4. Select the mapping classes (see *Read Classification*) to include in the analysis.
5. Click "Finish" to start the Coverage Analysis.

The first parameter defines the minimum coverage at each position within a "covered" reference interval.
The two radio buttons below define the counting method: either the coverage of both strands is summed, or each strand is analyzed separately.

The "Detect uncovered instead of covered intervals" checkbox switches the mode from detecting intervals which have a satisfactory amount of reads to the contrary. Then all reference intervals are listed, which have a lower coverage than defined by the minimum coverage threshold.

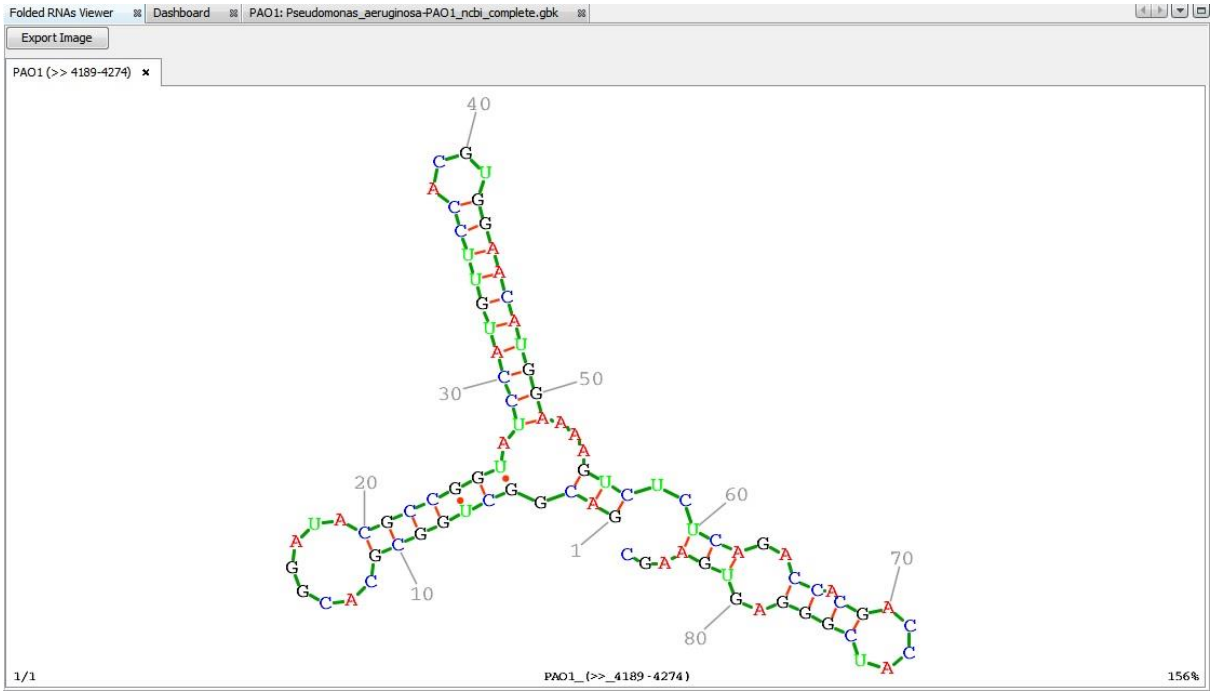
This analysis allows to explore the set of covered or uncovered intervals for the tracks and the tracks among one another.

In the following example, intervals with no coverage (coverage < 1) were queried and the coverage of both strands was merged. The used parameters are also shown below the result table.



RNA Secondary Structure Prediction

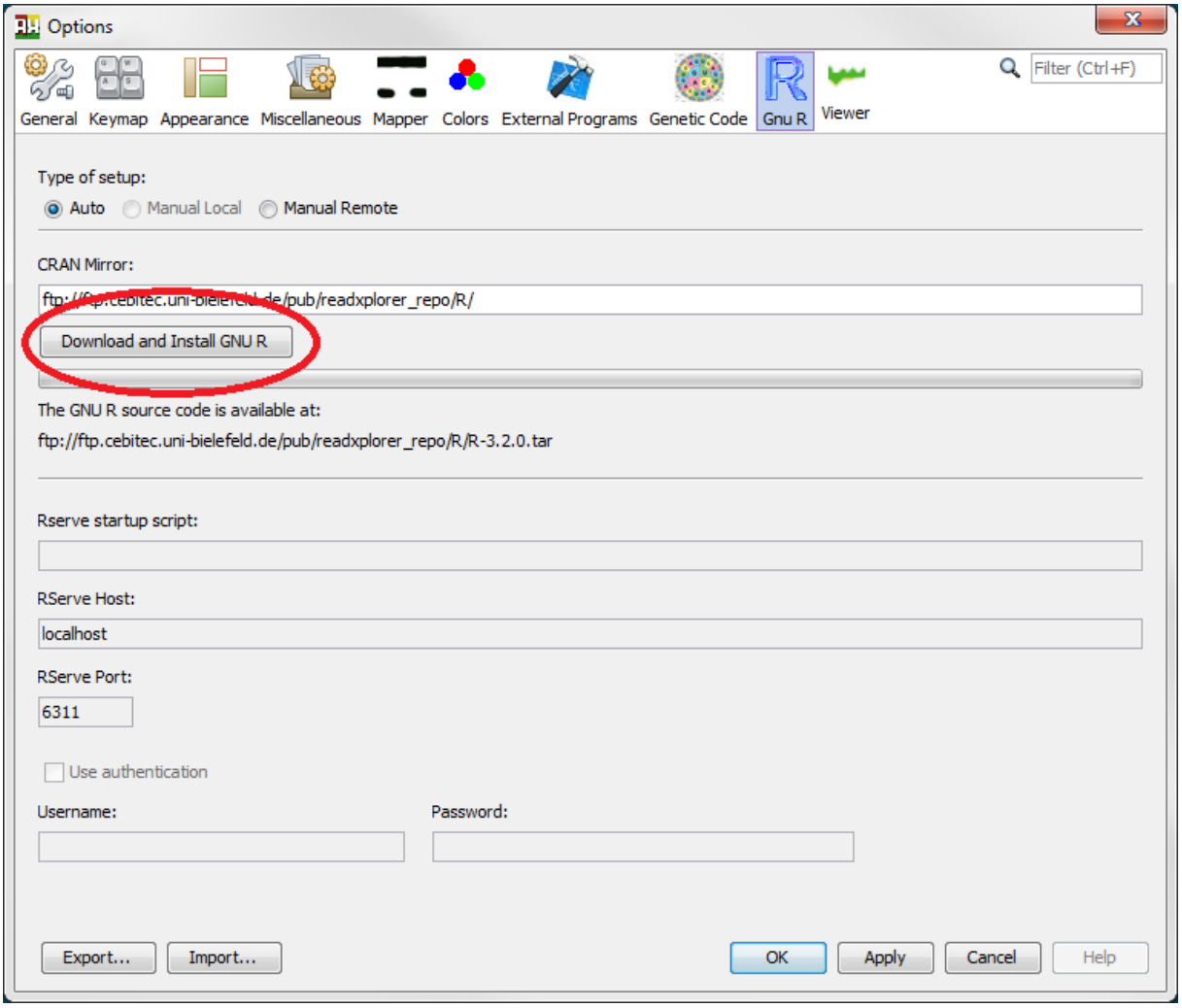
The RNA secondary structure prediction can be used either to calculate the most probable secondary structure of a selected sequence (see *ReadXplorer - Main Window h)*) from the reference or from a selected alignment in the *Alignment Viewer*. It uses the webservice of RNAfold coupled with an integrated version of the software RNAMovies for the visualization of the predicted secondary structure. The result is displayed in a new tab as shown below. It can be exported as image by pressing the "Export Image" button.



R Installation for Differential Gene Expression Analysis

Since Version 2.1 the connection from ReadXplorer to GNU R is realized using Rserve instead of rJava.

Windows Setup:
We offer a pre-build GNU R package for Windows that contains all necessary R packages. Within ReadXplorer go to "Tools"->"Options"->"GNU R" and select "Download and Install GNU R".
After accepting the License the download will start automatically and a progress indicator will appear. When the installation is finished you can use the differential expression analysis located in the "Tool" menu.



Mac OS Setup:

We do not offer an OS X version with a bundled GNU R installation anymore. However, the GNU R configuration page located within ReadXplorer ("ReadXplorer" → "Preferences" → "GNU R") offers the option to configure a connection to a user installed GNU R installation. It is possible to run the R instance on your local machine as well as connecting to a GNU R instance running on a remote host. Due to security constrains in OS X the usage of a start-up script is not possible. You must start your GNU R Rserve instance manually. The setup procedure is the same as described in the section "Installing a GNU R instance" in the "Linux Setup" chapter.

Linux Setup:

As there are so many different versions of Linux we do not offer an automatic installation on Linux systems. However, the GNU R configuration page located within ReadXplorer ("Tools"->"Options"->"GNU R") offers advanced setup options that allow different setup scenarios on Linux system. It is possible to run the R instance on your local machine as well as connecting to a GNU R instance running on a remote host. A few setup examples are given below.

Installing a GNU R instance:

For each scenario the machine that should run GNU R needs the required Software at first. Obviously, the first thing that needs to be installed is GNU R itself. The GNU R versions we use for Mac OS and Windows is 3.2.0, other 3.X versions should work as well, but we did not test them. There are GNU R packages included in most Linux distribution you can use or you can download the source-code at https://www.r-project.org/ . Which way is the best to go depends on your personal preferences and the Linux distribution you are using. After installing GNU R you also need to download the required, additional R packages used by ReadXplorer. For the connection between ReadXplorer and GNU R we need "Rserve", which can be installed in an open R session by typing:

```
install.packages("Rserve")
```

We also need the packages for "baySeq", "DESeq" and "DESeq2". These packages are hosted on http://bioconductor.org/ . The installation is done in R by typing:

```
## try http:// if https:// URLs are not supported
source("https://bioconductor.org/biocLite.R")
biocLite("baySeq")

source("https://bioconductor.org/biocLite.R")
biocLite("DESeq")

source("https://bioconductor.org/biocLite.R")
biocLite("DESeq2")
```

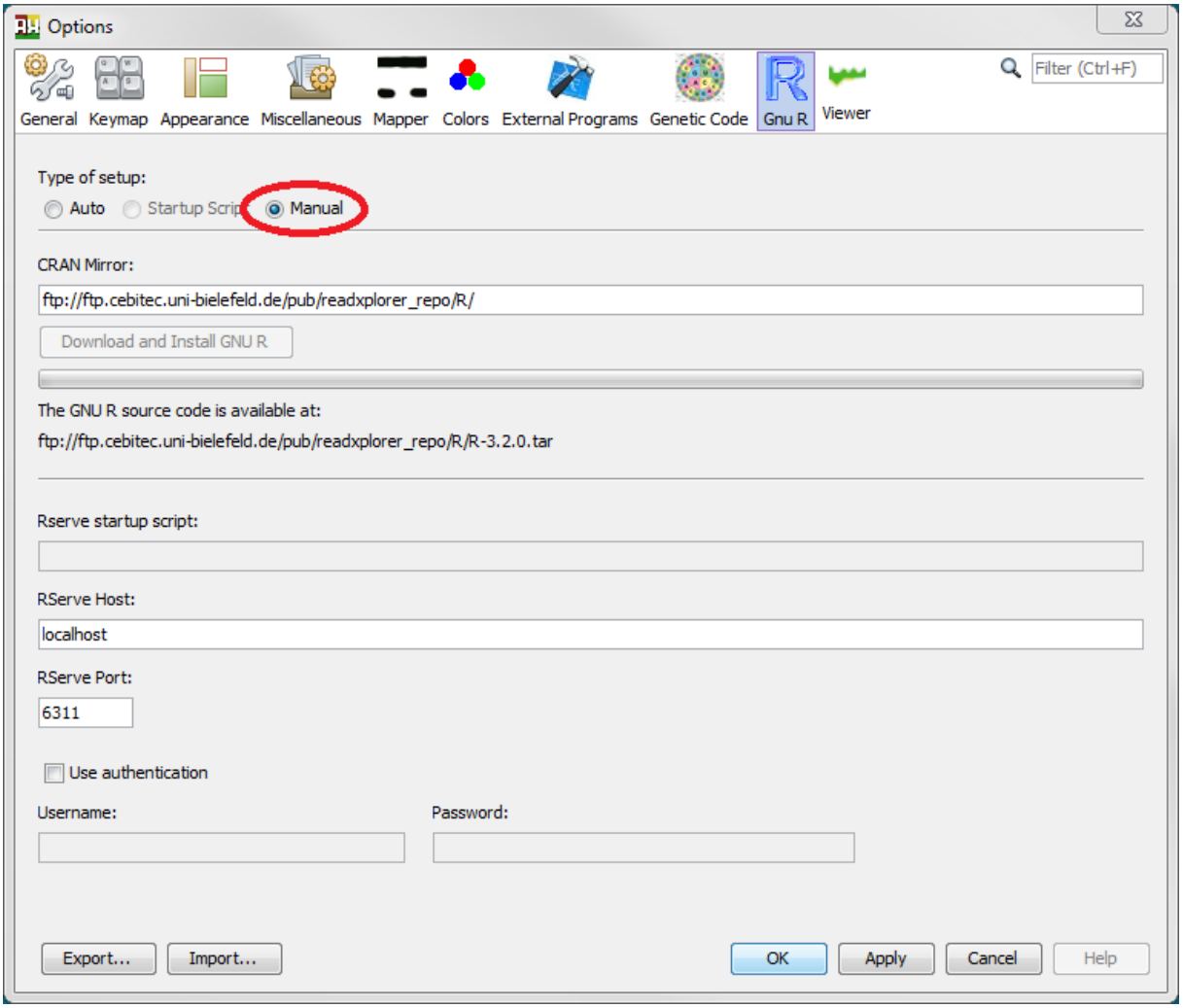
If that fails, please find further installation instructions on the http://bioconductor.org/ homepage.
Once everything is installed we are ready to setup the connection in one of the ways described below.

Connect to a manually started GNU R instance (local or remote)
Once GNU R and all the necessary packages are installed you can start the Rserve instance in different ways. When within a running GNU R session type

```
library(Rserve)
Rserve()
```

to start the Rserve Server instance. You can also start the Rserve server directly from the command line by executing "`R CMD Rserve`" provided the R is correctly included in your PATH. After executing one of these commands you will have a running Rserve instance listening on its default port 6311 for connections. Please be aware that if you are on a multi user system every one with access to the machine can now connect to this Rserve instance which is running with your user privileges. Because GNU R is capable of accessing the local file system this also means that this person can gain access to your files. So if you are on a multiuser system please check the Rserve manual on how to setup a password protected connection. The manual also describes how to setup Rserve to allow remote connections. This is especially useful if you want to provide one central R instance in your network which can then be used by different users at the same time.
In a final step the ReadXplorer settings need to be adjusted to connect to the local (or remote) running Rserve instance. Within ReadXplorer go to "Tools"->"Options"->"GNU R" and select

"Manual". Your must now enter the IP Address or Hostname of the machine running Rserve as well as the port. If the Rserve instance requires authentication check the "Use authentication" checkbox and enter a valid "Username" and "Password" combination. ReadXplorer will store the credentials in the key storage provided by you operating system. If your Rserve instance is running on your local machine (the same machine you are running ReadXplorer on) and you haven't changed the default port you can just use the default settings ("localhost" as "Rserve Host" and "6311" as "Rserve Port"). You should now be able to run the differential gene expression analysis located in the "Tools" menu.



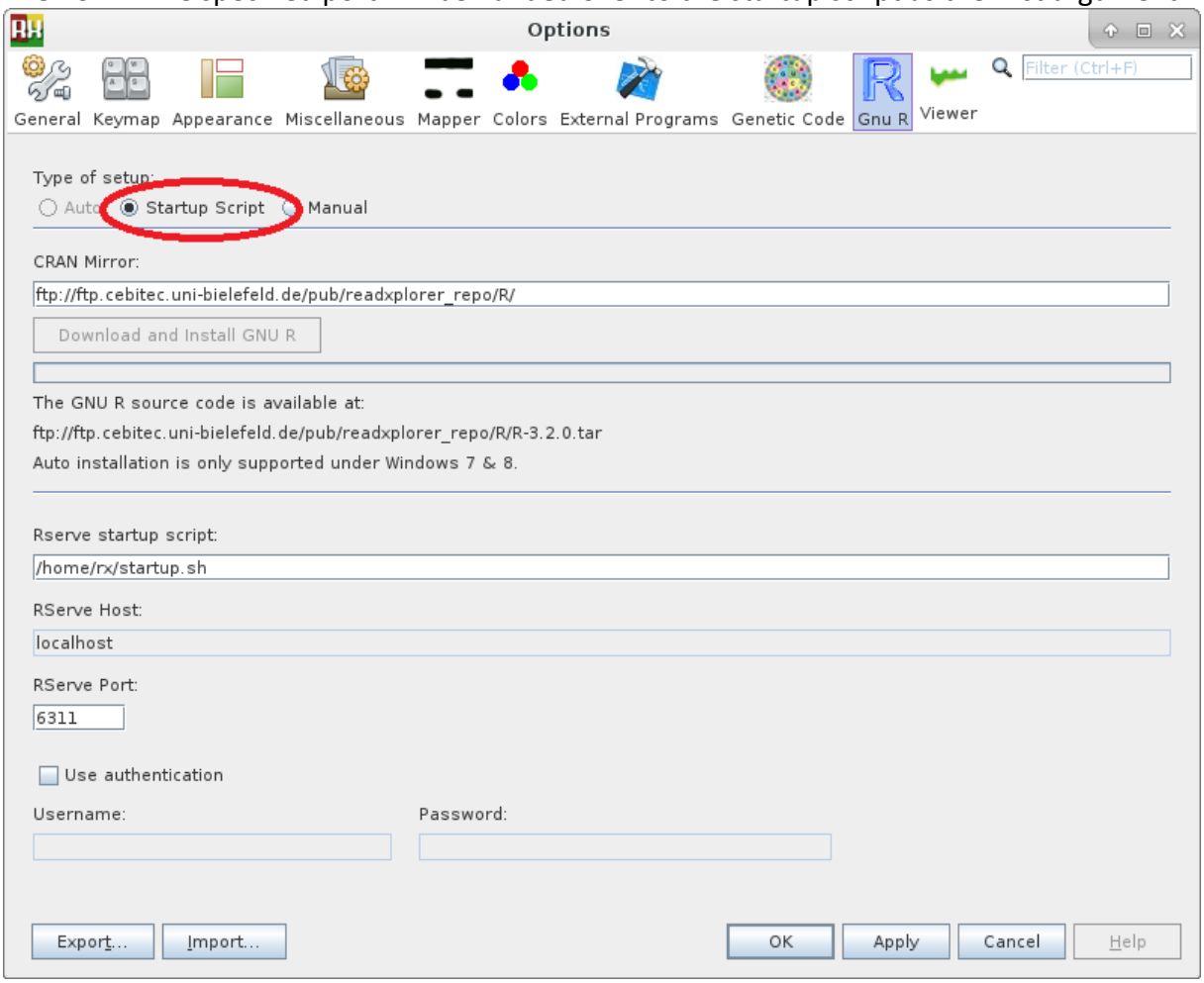
Use a startup script (local)

ReadXplorer also offers you the usage of a startup script. This allows ReadXplorer to start up the R instance when needed and to shut it down after the computation is complete. A start-up script could for example look like this:

```
#!/bin/bash
R CMD Rserve --RS-port $1 --vanilla
```

ReadXplorer will execute the startup script each time a differential gene expression analysis is started unless there is already a running instance. In this case the already running instance will

be used. On Linux hosts multiple connections to one Rserve instance should cause no problems. The path to the startup script can be set in the options menu at "Tools"->"Options"->"GNU R". The specified port will be handed over to the startup script as the first argument.

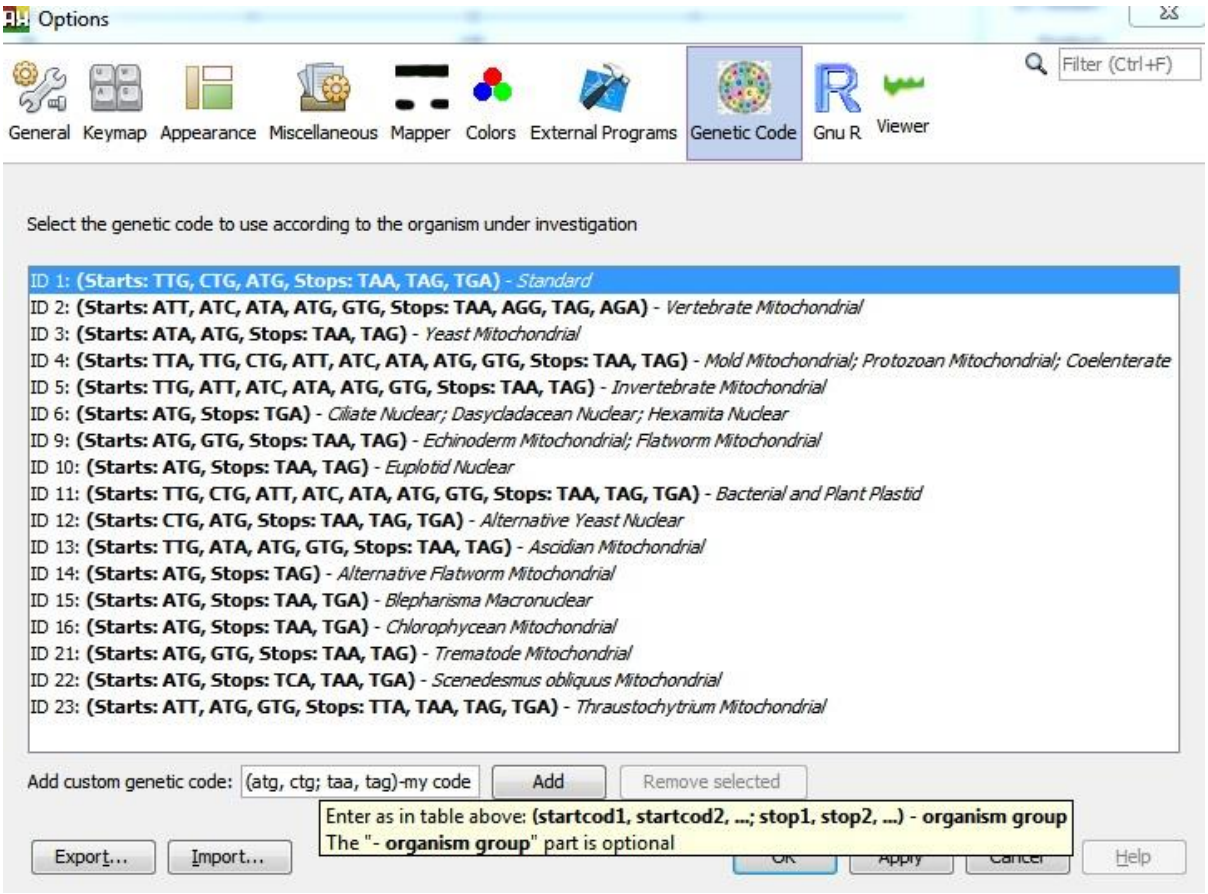


General Options

Via "Tools -> Options" all available general options can be adjusted.

- The "General" tab allows setting a proxy.
- The "Keymap" tab allows adjusting the keyboard shortcuts.
- The "Miscellaneous" tab allows editing "Files", “Locations”, and "Output" settings. The **"Locations"** tab allows configuring the **temporary directory** for track imports (SAM/BAM/JOK). This is important when your system drive has limited space and you want to switch the temporary directory to another disk. Further it enables configuring the enzyme DB to be used to create EC number links.
- The "Mapper" tab allows to set the path to an own script, calling a read mapping software (see *Mapping* (coming soon)). If you want to use the default script (starting bwa) supplied with ReadXplorer, do not change this setting.
- The "Colors" tab allows changing the colors of the five mapping classes (see *Read Classification*), the Double Track Viewer and the background color of the viewers (useful e.g. for creating publication screenshots).

- The "Genetic Code" tab allows to select another genetic code for highlighting of start and stop codons. If the required genetic code is not in the list, it can be added by the user (see screenshot below). **Note:** No translation table is available if a custom genetic code is selected. In this case the standard genetic code (index 1) is applied for translations (e.g. in *SNP and DIP Detection*).
- The "Gnu R" tab allows configuring the GNU R installation to be used for *Differential Gene Expression Analysis* as described above in section *R Installation for Differential Gene Expression Analysis*.
- The "Viewer" tab allows changing general viewer options. Currently, the height of all (double/multiple) track viewers can be chosen and automatic scaling of the coverage in these viewers can be de/activated. The maximum zoom level of the *Alignment Viewer* can be adjusted here, too.



The figure shows how to add a **custom genetic code**. One custom code named "Other standard" has already been added to the list and a second one is prepared in the "Add custom genetic code:"-field. The format of a custom genetic code has to adhere to the example in the tooltip below the text field. **Note** that start and stop codons have to be separated by a **";"**.

Help

The internal help for ReadXplorer can be found under "Help -> Help Contents".