BioNumerics Tutorial:

Importing sequence assemblies from BAM and SAM files

1 Aim

With the BioNumerics BAM import routine, a sequence assembly in BAM or SAM format can be imported in BioNumerics. A BAM file (file extension `.bam`) is the binary version of a SAM (Sequence Alignment Map) file. A SAM file (file extension `.sam`) is a tab-delimited text file that contains large nucleotide sequence alignment data. Detailed format descriptions can be found on the SAM Tools web site: `http://samtools.sourceforge.net`. In this tutorial you will learn how to use the BAM import tool in BioNumerics and to inspect the imported assemblies.

2 Example data

As an example we will import three sequence assemblies in BAM format in a BioNumerics database. The example files can be found on the download page on our website (`http://www.applied-maths.com/download/sample-data`, "BAM files").

3 The Import wizard

1. Create a new database (see tutorial "Creating a new database") or open an existing database.
2. Select ***File*** > ***Import...*** (, **Ctrl+I**) to open the *Import* dialog box.
3. Choose the option ***Import sequence assemblies from BAM files*** under the ***Sequence type data*** item in the tree and click <***Import***>.
4. The import wizard allows you to browse for one or more files. Press <***Browse***>, navigate to the folder, select the BAM files and press <***Open***> (see Figure 1).
5. Press <***Next***>.

The next step of the import wizard lists the templates that are present to import sequence information in the database. The ***Default*** template will import the sequence assemblies in the database and link the sequences to new entries in the database (if the option ***Create x entries*** is checked in the final step). The keys are automatically created by the import routine.

As an exercise, we will create a new import template and link the file names of the BAM files to the ***Key*** field.

6. Click <***Create new***> to create a new import template.
7. Select the only row in the list, i.e. the file name, and click <***Edit destination***> or simply double-click on the row. Select "Key" and press <***OK***> (see Figure 2).
8. Press <***Next***>, make sure "Key" is checked in the ***Import links*** step and press <***Finish***>.

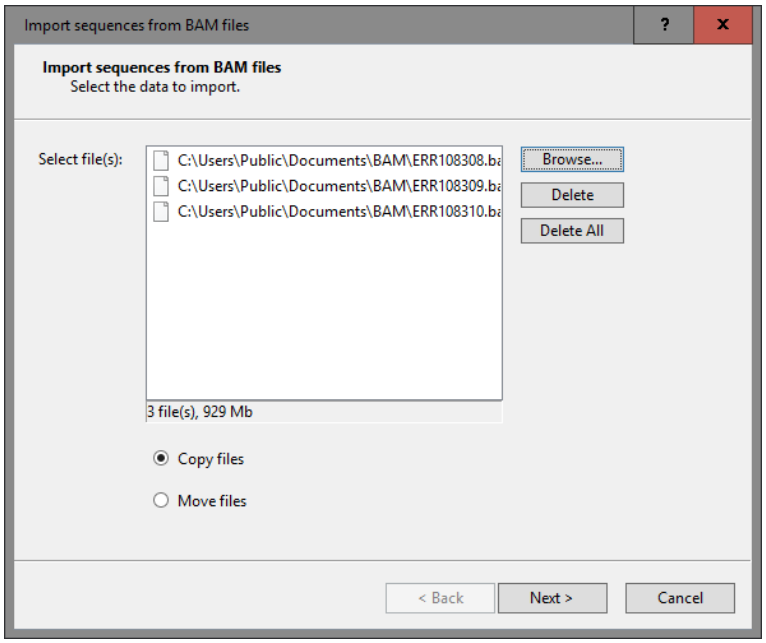

Figure 1: Select BAM files.

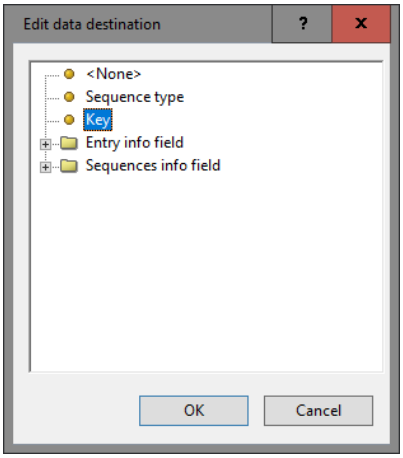

Figure 2: Link to Key field.

The import template can be saved to be able to use it again later on.

9. Enter a ***Name*** for the import template (e.g. "My BAM files") and optionally a ***Description***. Next, press <***OK***>.

10. The newly added template is automatically selected. Make sure <***Create new***> is selected from the ***Experiment type*** list or select an existing experiment and press <***Next***> (see Figure 3).

11. Specify a sequence type name if prompted for (e.g. **Assembly**) and press <***OK***> and confirm the action.

The last dialog will indicate that 3 new entries will be created during import (see Figure 4).

12. Press <***Finish***> to start the import into the database. The progress of the import is shown while the data is added to the BioNumerics database.

The new entries are created in the database and listed in the *Database entries* panel. The reads in the assembly files are sorted on position, if not already pre-calculated and available from the files, and the assembly file is indexed by genomic position to efficiently retrieve all reads aligning to a specific region. Both file manipulations use the SAM Tools.

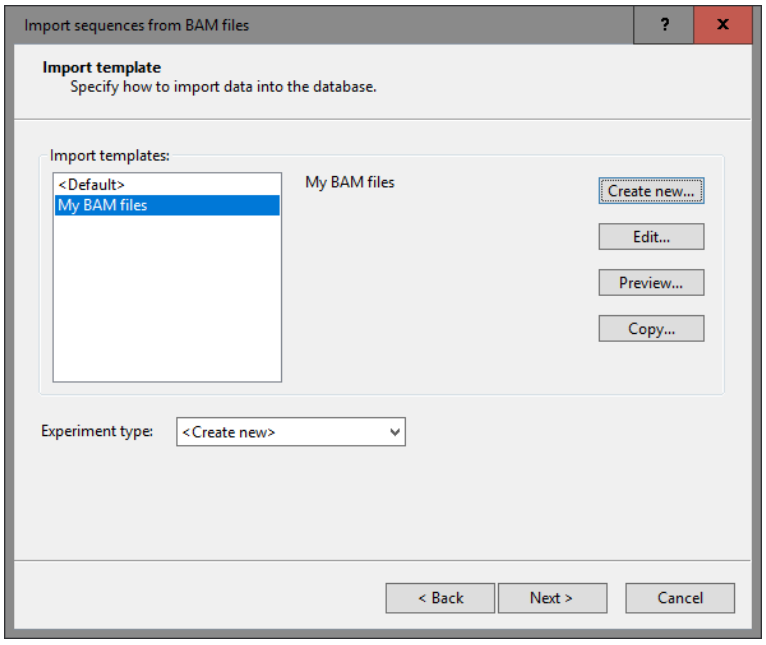

Figure 3: Import template.

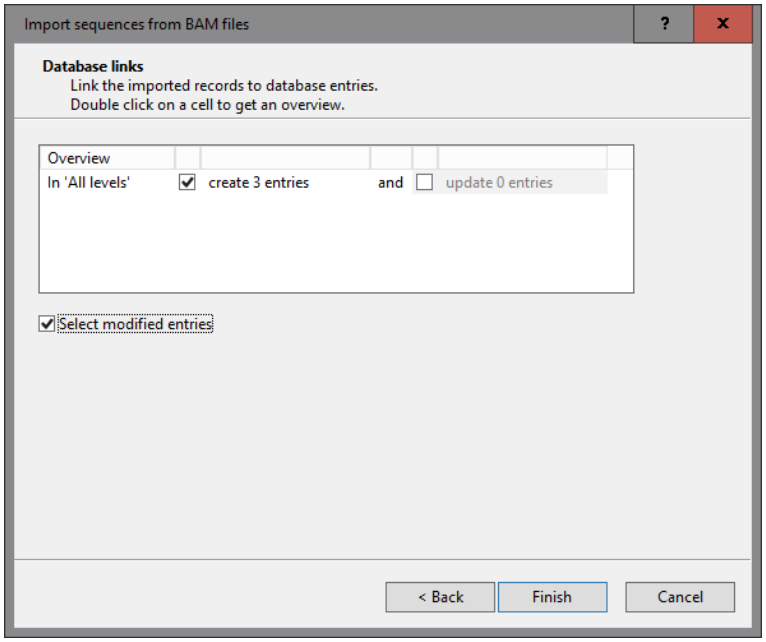

Figure 4: Import wizard: database links.

Once the import is completed, the consensus sequence, defined by the default coverage and base calling settings, is saved to the database together with the coverage information and can be viewed in the *Sequence editor* window (see 4).

4 Inspect assembly

1. Click on a green colored dot in the *Experiment presence* panel, corresponding to the sequence type experiment, to open the *Sequence editor* window for an entry.

The consensus sequence, defined by the default coverage and base calling settings, is displayed in the *Sequence Editor* panel. The details of the sequence assemblies can be viewed in the *BAM viewer* window.

2. Select ***File*** > ***Open assembler*** ( ) to launch the *BAM viewer* window (see Figure 5).

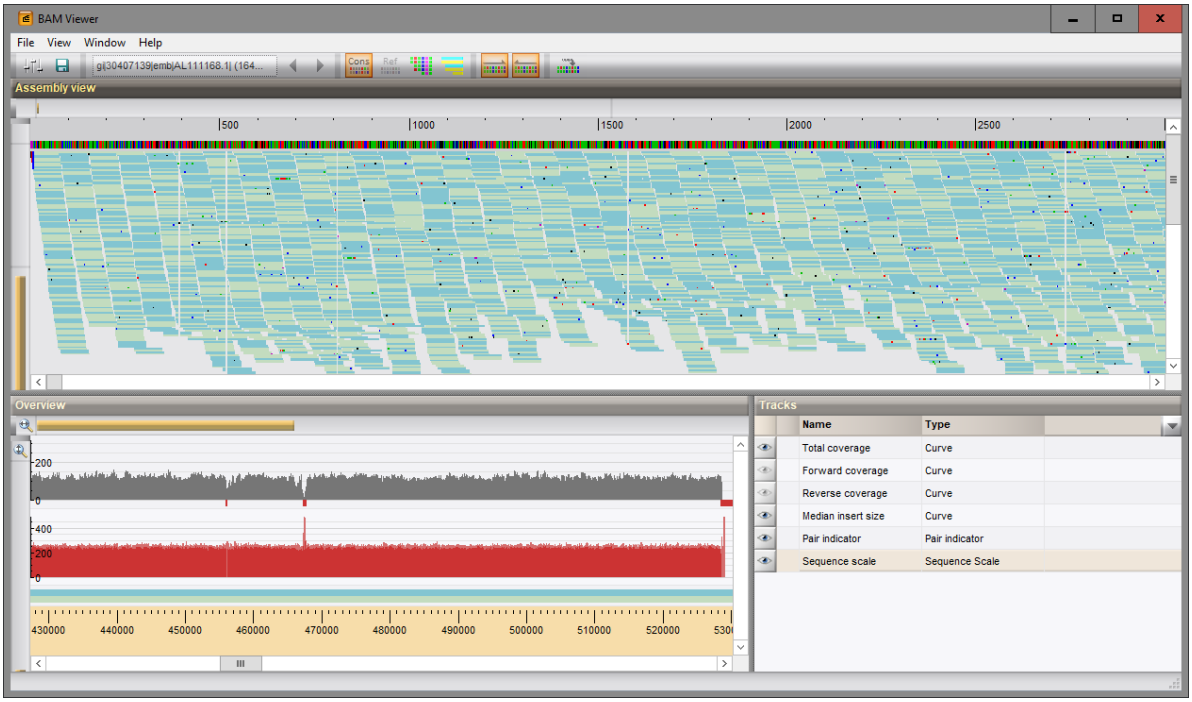

Figure 5: The *BAM viewer* window.

By default, the *Assembly view* panel displays the sequence scale based on assembly position, i.e. consensus positions including gaps, the reference sequence (not defined in this exercise), the consensus sequence and the alignment data, forward mapped reads indicated in blue and reverse mapped reads indicated in green.

Read bases that are identical to the consensus sequence are displayed in blue or green, depending on the mapping direction, whereas read bases that are different from consensus base call have their specific base color, i.e. green for `A`, blue for `C`, black for `G` and red for `T`.

The reads can be displayed at different levels of resolution:

3. Zooming is done by using the yellow zoom sliders or scrolling the mouse wheel while holding the **Ctrl**-button. This way, it is possible to zoom into individual read base level.

4. To display only the read outlines select ***View*** > ***Show only outlines*** ( ).

5. To display all bases use ***View*** > ***Color all bases*** ( ) (see Figure 6).

By default, both forward and reverse mapped reads are shown.

6. Select ***View*** > ***Show forward*** ( ) or ***View*** > ***Show reverse*** ( ) to only display the forward or reverse reads respetively. This allows to search for e.g. strand-specific sequencing artifacts leading to strand-specific SNPs.

7. To quickly access a specific position in the alignment, use ***View*** > ***Go to...*** ( ). This opens the *Go to assembly position* dialog box where the specific assembly position can be entered. After confirmation, the requested position is displayed in the center of the updated assembly view.

If multiple contigs were defined in the BAM or SAM file, all contigs will be saved as a concatenated sequence to the same sequence experiment type. One can jump between the different imported contigs by selecting any of the contigs from the drop-down list in the toolbar or via ***View*** > ***Show contig***.

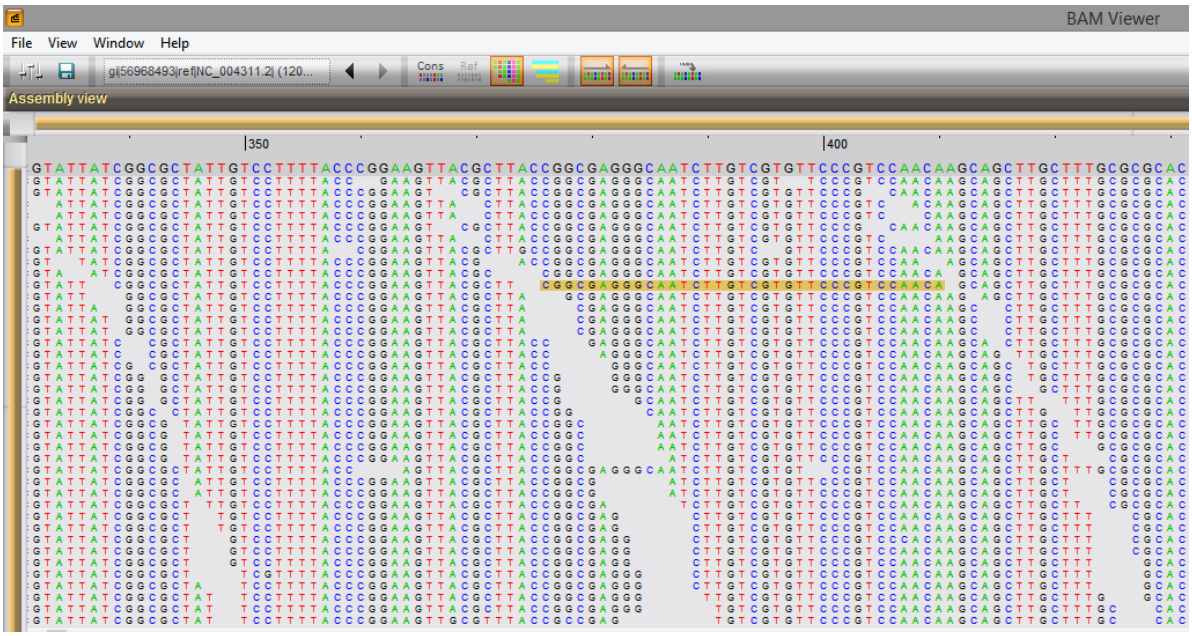

Figure 6: Detail of *BAM viewer* window with all bases color-coded.

In the *Tracks* panel, the different curves displayed in the *Overview* panel are listed. Clicking the icon in front of each track, displays/hides the tracks from the overview. By default, the total coverage, the median insert size of the paired-end reads, a pair indicator and the sequence scale are displayed.

Following tracks are available:

- The *Total coverage* contains the sum of forward and reverse read coverage information, as calculated over the consensus sequence. The red bars below the curve are visual indicators of the regions that do not satisfy the coverage criteria as defined in the consensus settings.

- The *Forward coverage* and *Reverse coverage* tracks contains similar information as the total coverage, but limited to forward and reverse mapped reads, respectively.

- The *Median insert size* visualizes the median absolute deviation of the insert sizes that cover the selected position. The median insert size is a robust measure of the variability in the insert size of all reads that cover a specific position. In contrast to the standard deviation, this measure is more resilient to outliers in insert sizes.

- The *Pair indicator* marks the mapping position of the selected reads and that of its corresponding paired read. For the selected forward mapped reads, the mapping positions are displayed in the blue channel, whereas for the selected reverse mapped reads, the corresponding mapping positions of the pairs are displayed in the green channel. Each selected read is indicated in orange, whereas its paired read is indicated in red (forward) and blue (reverse).

- The *Sequence scale* contains the base pair indication as imported from the BAM/SAM file.

Zooming in the *Overview* panel is managed by the yellow sliders or hovering over the channels and scrolling with the mouse wheel.

5 Conclusion

In this tutorial you have seen how easy it is to import BAM (and SAM) files in BioNumerics. The sequences can now be further analyzed in BioNumerics. More information can found in the analysis tutorials on our website. For detailed information about the *BAM viewer* window we refer to the BioNumerics reference manual.