BioNumerics Tutorial:

wgMLST typing in BioNumerics: detailed exploration of results

1 Introduction

This tutorial further elaborates on the wgMLST results obtained after job submission. The step-by-step procedure to submit and fetch wgMLST jobs in your BioNumerics database can be found in the tutorial: "wgMLST typing in BioNumerics: routine workflow".

2 De novo assembly

The results from the de novo assembly algorithm, i.e. concatenated de novo contig sequences are stored in the sequence experiment type **denovo**.

1. Click on the green colored dot for one of the entries in the ***denovo*** column in the *Experiment presence* panel.

The *Sequence editor* window opens, containing the results from the de novo assembly algorithm (see Figure 1). The concatenated de novo contig sequences are displayed in the *Sequence Editor* panel and are separated by pipes (|). Details on the different contigs can be inspected in the *Contigs* panel.

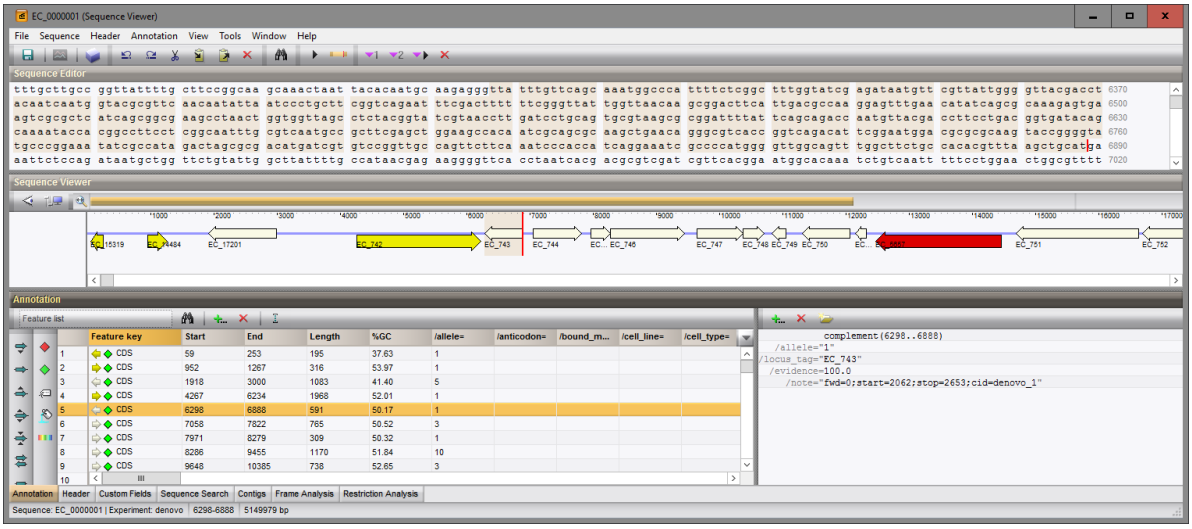

Figure 1: The Sequence editor.

When an assembly-based calling was performed, the detected loci are listed in the *Annotation* panel (see 3.3 for more information about this job and the results):

- Loci of which the sequence has a 100% match with an existing allele in the nomenclature allele database, or when the sequence was new and passed the automatic submission criteria are indicated in **white**. The allele number is displayed in the ***allele*** column.

- Loci that do not have a 100% match with an allele in the nomenclature allele database and that do not fulfill the automatic submission criteria are displayed in **red** (when IUPAC code is present in the sequence) or **yellow** (when the sequence only consists of non-ambiguous bases). The best matching reference allele is listed in the ***allele*** column.

2. Close the *Sequence editor* window.

The results of the de novo assembly can also be consulted in the *wgMLST quality assessment* window, giving you a nicer overview of the assembly combined with the results of the allele calling (if performed).

3. Select some entries in the *Database entries* panel.

4. Select ***WGS tools*** > ***wgMLST quality assessment...*** ( ) to open the *wgMLST quality assessment* window.

The *Genome* panel (bottom left) shows the graphical representation of the sequence of the currently selected entry in the *Entries* panel (see Figure 2).

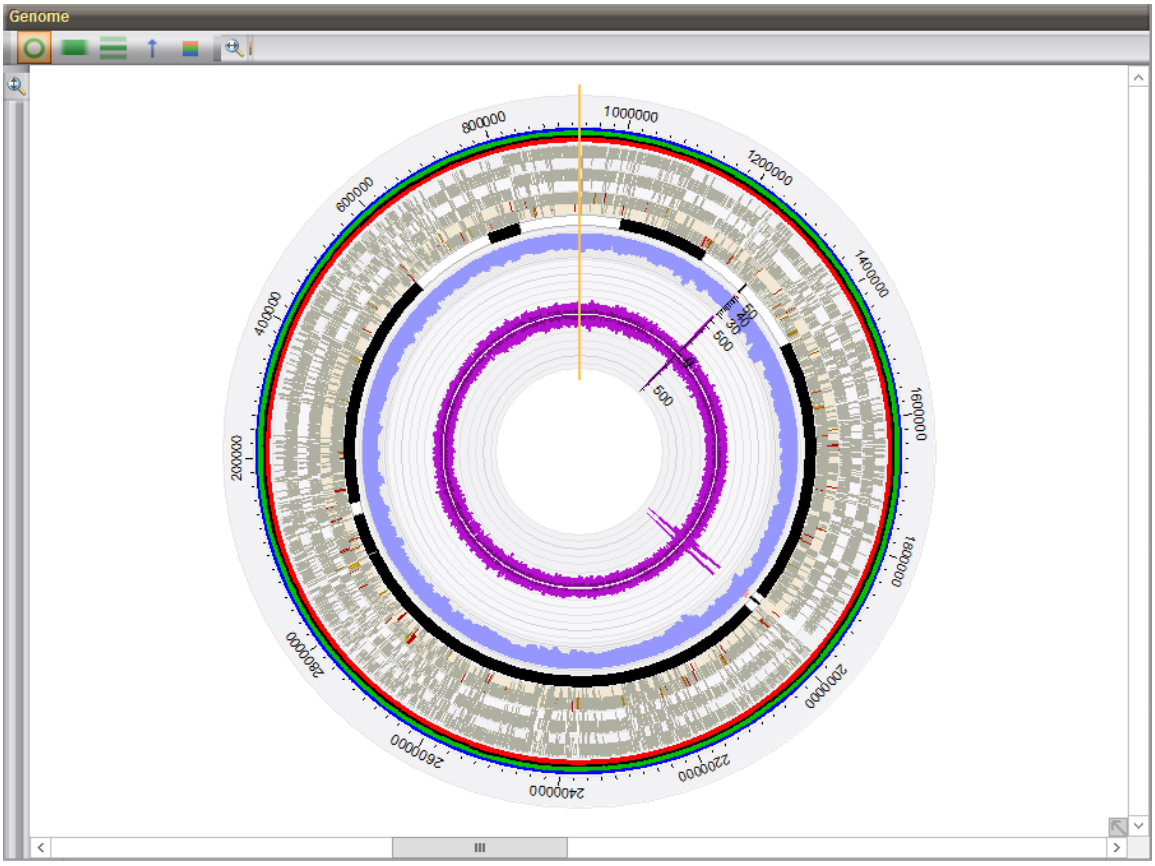

Figure 2: Graphical overview.

5. Use the zoom slider next to the toolbar in the *Genome* panel to zoom in on the sequence. Zooming is done on the upper area of the circular sequence and can be done up to base level (see Figure 3).

The bases are colored based on following color scheme: green - `A`, blue - `C`, red - `T`, black - `G`, and gray for any IUPAC code denoting ambiguous positions.

The de novo contigs are separated by pipes (|). The contigs are also graphically represented in the "Contigs" track (a few tracks below the sequence) with alternating white and black blocks, denoting different contigs (see Figure 3). When the Velvet de novo assembly algorithm was used, the contigs are randomly ordered; when the Spades algorithm was selected the contigs are ordered based on the contig size (see Figure 2: randomly ordered black and white contig blocks).

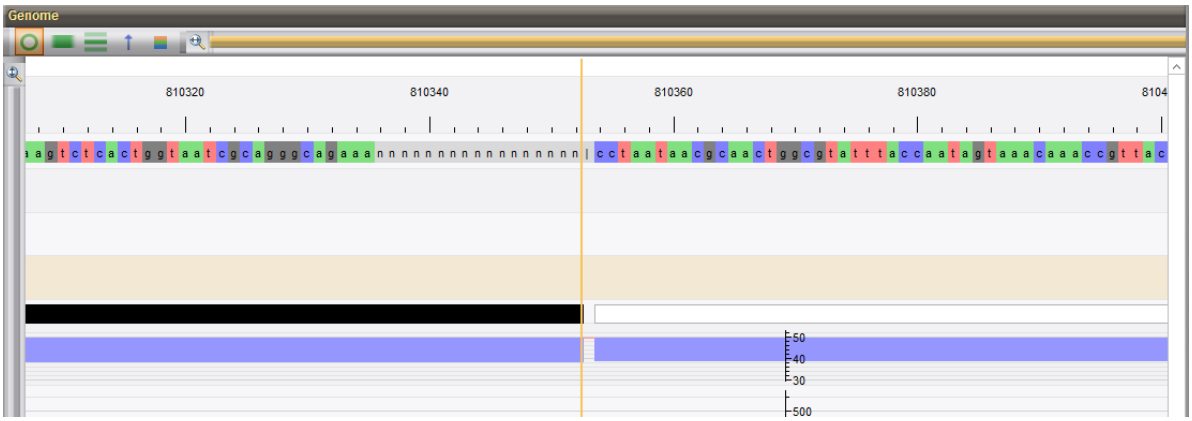

Figure 3: Zooming in on the sequence.

6. Close the *wgMLST quality assessment* window.

3 Allele calling

3.1 Introduction

The **wgMLST** experiment contains the allele calls for the detected loci.

1. Click on the green colored dot for one of the entries in the ***wgMLST*** column in the *Experiment presence* panel to open a character card (see Figure 4).

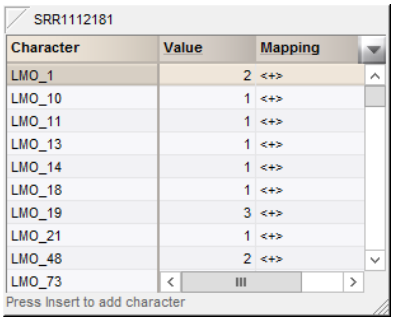

Figure 4: The character experiment card for an entry.

The Locus identifiers are listed in the ***Character*** column and the allele calls are listed in the ***Value*** column.

2. Close the character experiment card by clicking on the triangle in the top left corner.

A detailed overview of the allele calling results can be consulted in the *wgMLST quality assessment* window.

3. Select some entries in the *Database entries* panel.

4. Select ***WGS tools*** > ***wgMLST quality assessment...*** () to open the *wgMLST quality assessment* window.

The *Alleles* panel displays the allelic assignments of the currently selected entry in the *Entries* panel.

5. Select an entry in the *Entries* panel. The *Genome* panel and the *Alleles* panel are now updated with the information for this entry.

Default, the allele calling results of ***All loci*** are displayed. Another subscheme of the **wgMLST** character experiment type can be selected from the drop-down list, restricting the view to only those characters contained in the selected view (see Figure 5).

In most reference databases following views have been defined at the curator level and are synchronized upon installation: the default view **All loci**, the **Core loci**, the **MLST** view for the traditional seven housekeeping loci, and the **wgMLST loci** view containing all loci except the ones present in the **MLST** view. User-defined views - if defined - can also be selected from the list.

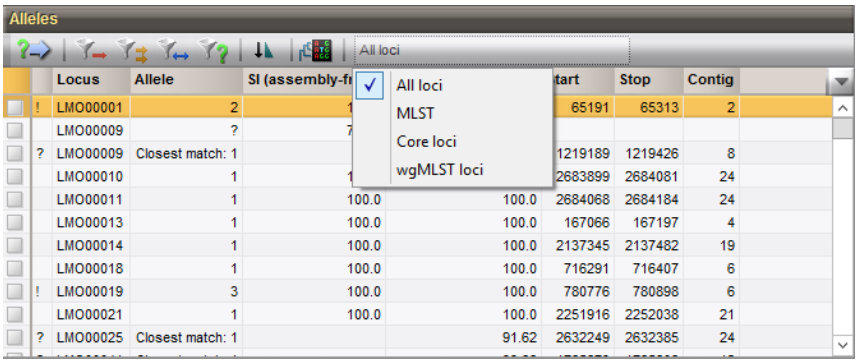

Figure 5: Filter based on subschemes.

Let us take a closer look at the allele calls present in the *Alleles* panel. The results of the different allele calling algorithms (assembly-free versus assembly-based) are split up in the next sections for ease of interpretation.

3.2 Assembly-free calls

Starting from the sequence read sets, this algorithm uses a k-mer based approach to check which loci are present from the organism-specific wgMLST scheme in the reads. The settings are listed in the *Find alleles* dialog box (see Figure 6) and can be called with ***WGS tools*** > ***Submit jobs...*** ( ) and click <***Settings***>. The default settings are specified by the curator and are imported upon installation of the plugin. Normally no changes are required.

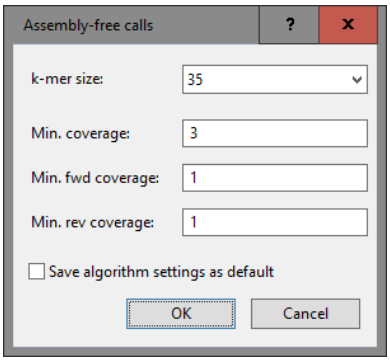

Figure 6: Assembly-free settings.

All loci that passed the assembly-free criteria are listed in the *Alleles* panel (see Figure 7 for an example). The locus identifier is displayed in the ***Locus*** column. The result of the matching of the allelic sequences against the nomenclature allele database records are listed in the ***Allele*** and ***SI (assembly-free)*** columns:

- When a 100% match is found with an allele in the allele database, the allele number is indicated in the ***Allele*** column and the similarity value (100%) is indicated in the ***SI (assembly-free)*** column.

- Matches with a similarity below 100% are also listed, but are not further considered. A question mark is displayed in the ***Allele*** column and the similarity value with the best matching reference allele is indicated in the ***SI (assembly-free)*** column.

Details of the selected assembly-free calling are shown in the *Details* panel below: the ***Sequence identity*** between the allelic sequence and the best matching reference in the allele database and the ***keyword coverage*** are listed.

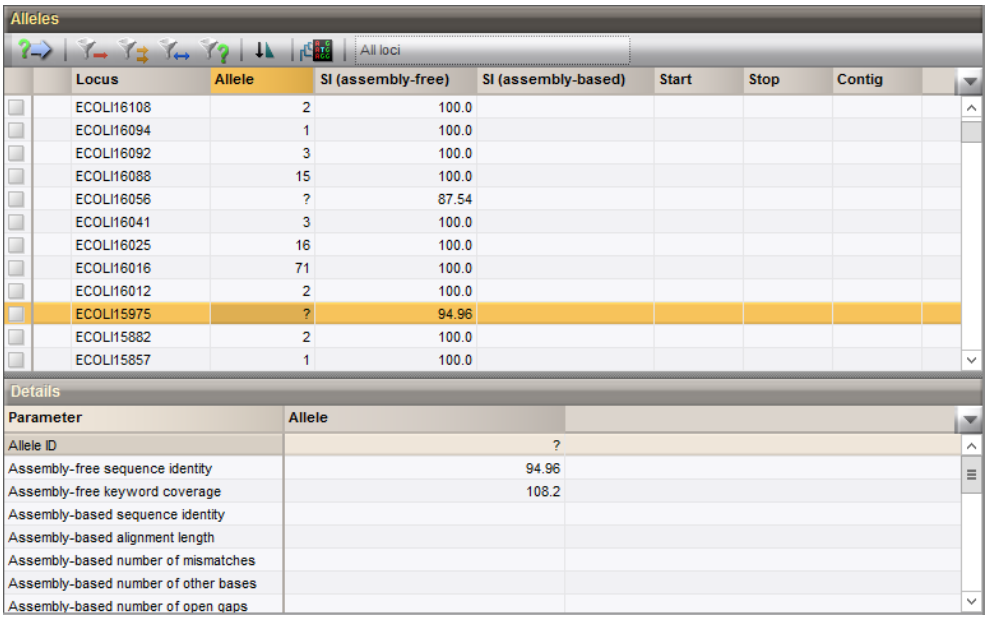

Figure 7: Assembly-free results: perfect and non-perfect matches.

Loci that were only detected based on the assembly-free algorithm will not be plotted on the sequence in the *Genome* panel since no contig position information can be derived from the assembly-free algorithm. If the locus is also detected by the assembly-based approach (see 3.3), the locus will be plotted both on the ***Assembly-free calls*** and ***Assembly-based calls*** track (see 3.4).

3.3 Assembly-based calls

This algorithm performs a BLAST-based allele detection on the de novo assembled contigs. The settings are listed in the *Perform BLAST on assemblies* dialog box (see Figure 8) and can be called with ***WGS tools*** > ***Submit jobs...*** ( ) and click <***Settings***>. The default settings are specified by the curator and are imported upon installation of the plugin. Normally no changes are required.

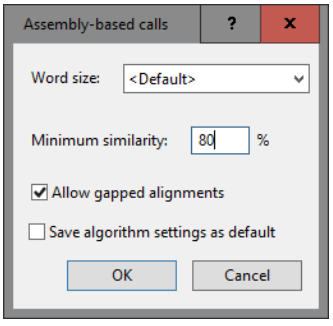

Figure 8: Assembly-based settings.

Only the detected alleles that passed the ***Minimum similarity*** threshold, i.e. the minimum BLAST similarity between the allele sequence and (one of) the reference sequence(s) in the allele database are retained and are listed in the *Alleles* panel. The locus identifier is displayed in the ***Locus*** column.

The results of the exact matching of the allelic sequence against the reference and accepted alleles in the allele database are listed in the ***Allele*** and ***SI (assembly-based)*** columns.

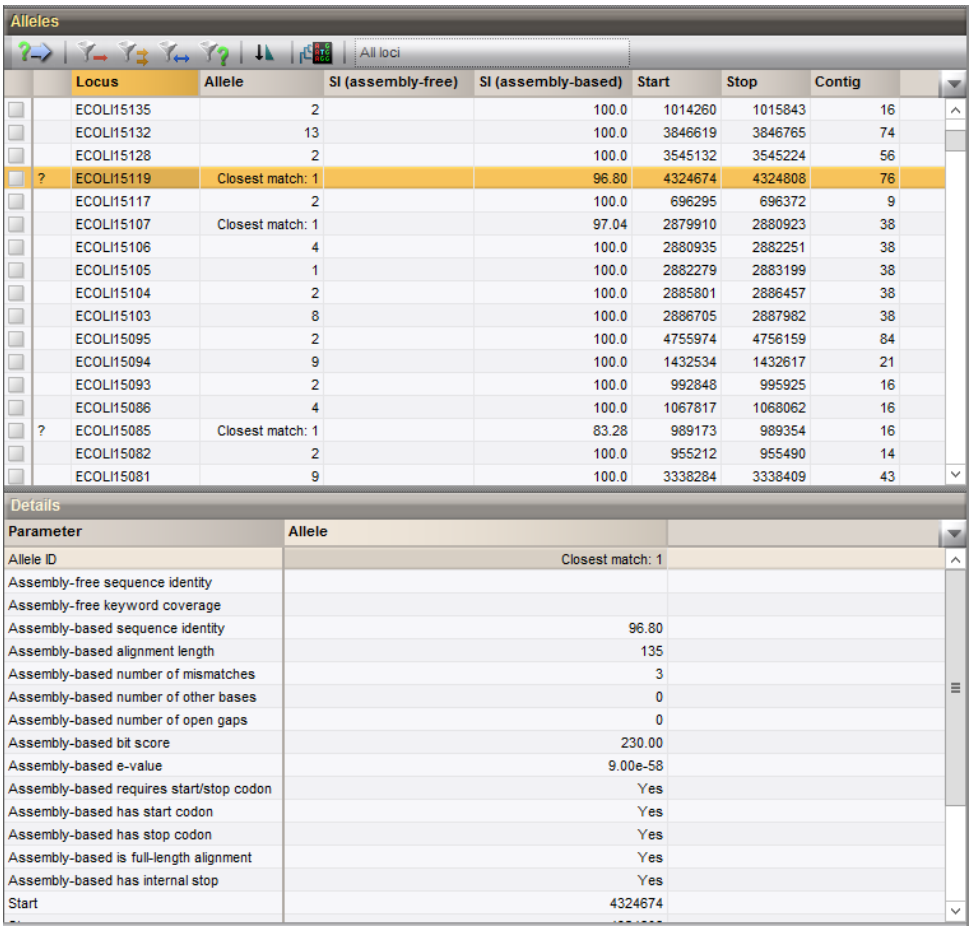

Figure 9: Assembly-based results.

- When a 100% match (***SI (assembly-based)***) is found with a reference or accepted allele sequence for a locus, the allele number is indicated in the ***Allele*** column.
- Matches that do not have a 100% match with an allele in the allele database but fulfill all specified automatic submission criteria (see below) are automatically submitted and receive the ”tentative” status until approved by the curator. This is indicated with an ”!” in the first column. An automatic curation process is followed instantly: when the ”tentative” allele passes the curator settings, the status is automatically converted to ”accepted”. All accepted alleles are updated each night.
- When a 100% match (***SI (assembly-based)***) is found with a tentative allele sequence for a locus, an ”!” is indicated in the first column, the (tentative) allele number is indicated in the ***Allele*** column.
- Matches that do not have a 100% match with an allele in the allele database and that do not fulfill the automatic submission criteria are indicated with the text ***Closest match: x***. The best matching reference allele is listed (x) together with the similarity with this reference sequence (see ***SI (assembly-based)*** column). When the sequence consists of non-ambiguous bases a ”?” is indicated in the first column (eligible for manual submission); when IUPAC code is present, nothing is indicated in the first column.

The automatic submission criteria can be called with ***WGS tools*** > ***Settings...***: click the *wgMLST* tab and the <***Auto submission criteria***> button. By default, the ***Use nomenclature acceptance criteria*** option will be checked, meaning that the automatic submission settings are used that are defined by the curator of the allele database. By default a start and stop codon are required in case of CDS loci, internal stops are not allowed, and a minimum homology with the reference allele(s) is required for automatic submission.

6. Click on a locus in the *Alleles* panel that was detected by the assembly-based algorithm.

Details are shown in the *Details* panel below. The selection in the *Genome* panel is updated: the locus is selected and is now located in the upper area of the circular sequence.

7. Zoom in on the sequence using the zoom slider.

The locus is plotted on the map (based on the ***Start***, ***Stop*** and ***Contig*** information of the locus) on the ***Assembly-based calls*** track (see Figure 10). The locus identifier and allele sequence number (between brackets) are indicated. Matches that do not have a 100% match (see ***SI (assembly-based)*** column) are colored based on the similarity value: yellow over red (lowest similarity). When the locus was also detected by the assembly-free algorithm, the locus is also plotted on the ***Assembly-free calls*** track.

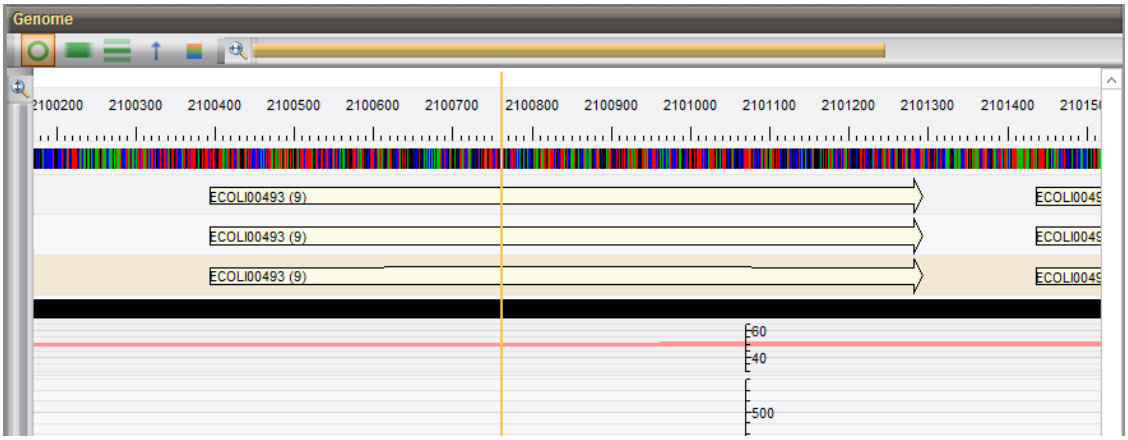

Figure 10: Loci mapped on the sequence.

3.4 Summary calls

When both algorithms (assembly-free and assembly-based) were run, all available data from the two allele identification algorithms are ”summarized” into a single set of allele assignments and stored in the ***wgMLST*** character experiment. The way the data is ”summarized” depends on the calls that were obtained for each locus and on the settings defined in the *wgMLST tab* in the *Calculation engine settings* dialog box (see Figure 11):

- If there is no overlap between the perfect (100%) matches between both algorithms for a locus, the summary calls will have no results as the allele calls were discrepant for that locus.
- If both methods found one perfect (100%) match for a locus corresponding to the same allele, this allele call is included in the summary for this locus.
- If only the Assembly-free method found a single perfect (100%) match for a locus, the allele call is included in the summary for this locus.
- If only the Assembly-based method found a single perfect (100%) match for a locus, the allele call is included in the summary for this locus.

- If one method found multiple perfect matches (100%) for a locus, the lowest allele ID is default retained for this locus in the summary (***Store lowest common allele ID*** is default checked). When the option ***Store as absent value*** is checked, no consensus call is retained.
- If both methods found multiple perfect matches (100%) for a locus, the lowest common allele ID is default retained for this locus in the summary (***Store lowest common allele ID*** is default checked). When the option ***Store as absent value*** is checked, no consensus call is retained.

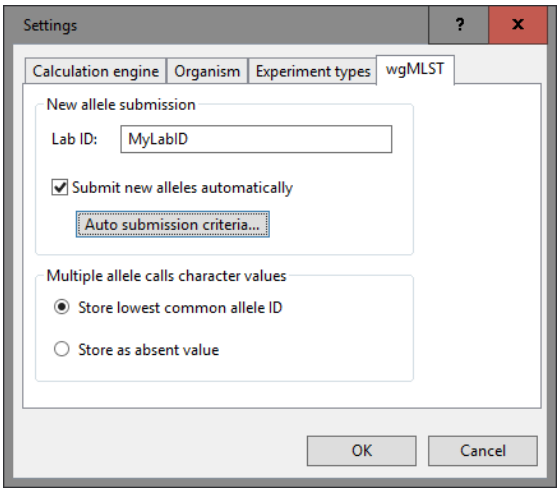

Figure 11: Multiple calls.

Assembly-free calls with a similarity below 100% ("?" in ***Allele*** column) are never considered for allele calling.

4 Quality results

4.1 Character card

The character experiment type **quality** provides insight in the quality of the reads and the results obtained for the different submitted jobs.

1. Click on the green colored dot in the ***quality*** column to open the character card for an entry in the database.

The **quality** character card contains quality statistics for the raw data, the trimmed data, the de novo assembly and the different allele identification algorithms (see Figure 12).

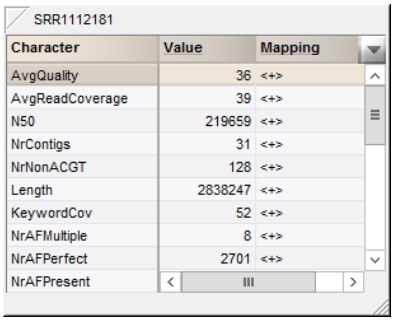

Figure 12: The character experiment card for an entry.

Based on the values stored in this experiment possible presence of low quality input data can be checked before launching jobs on the calculation engine and the results of the different jobs can be checked for the presence of contamination and bad assembly and calling results. This can be done in the *wgMLST quality assessment* window (see 4.2) and in the *Comparison* window (see 4.3).

2. Close the character experiment card by clicking on the triangle in the top left corner.

4.2 Quality assessment window

The quality parameters can also be consulted in the *wgMLST quality assessment* window.

3. Select some entries in the *Database entries* panel.

4. Select ***WGS tools*** > ***wgMLST quality assessment...*** ( ) to open the *wgMLST quality assessment* window.

A summary of the quality assessment is shown for each of the selected entries in the *Entries* panel (see Figure 13).

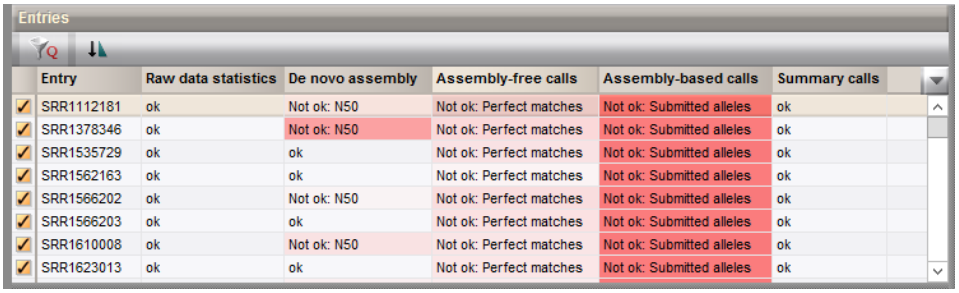

Figure 13: The *Entries* panel.

5. Double-click an entry in the *Entries* panel to show the detailed quality control parameters (see Figure 14).

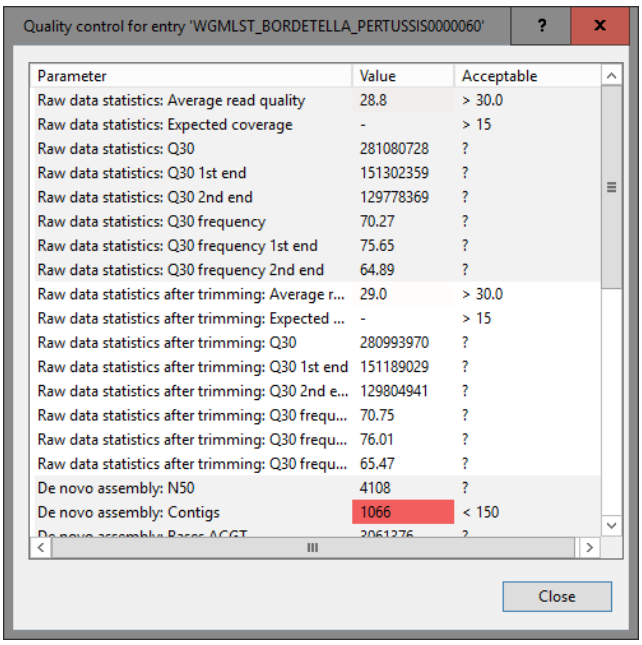

Figure 14: Quality control.

The quality parameters are grouped based on the data sets and algorithms that were launched: ***Raw data statistics***, ***Raw data statistics after trimming***, ***De novo assembly***, ***Assembly-free calls***, ***Assembly-based calls***, and ***Summary calls***.

The values of the selected entry are listed in the ***Value*** column. A number of quality criteria are evaluated against the accepted thresholds, as defined by the curator (see ***Acceptable*** column). The intensity of the red color in the ***Value*** column is an indication of the magnitude of deviation.

If all criteria of a group of parameters are within acceptable bounds, ”OK” is printed in the corresponding column in the *Entries* panel. If this is not the case, the parameter which deviates most is the final value that is reported. Note that one or more parameters failing to meet the required threshold does not per definition indicate a failed analysis, just that the calculated statistics do not fall within the interval specified as acceptable by the allele database curator.

6. Click on the ’?’ in the right upper corner of the *Quality control* dialog box for a detailed description of all the parameters displayed.

Some parameters are more informative and important than others. The most important ones are highlighted in 5.

7. Close the *Quality control* dialog box and *wgMLST quality assessment* window.

4.3 Comparison window

The quality parameters can also be consulted in a very quick and easy way in the *Comparison* window.

8. In the *Main* window, select the entries that you want to analyze using the check-boxes next to the entries or with the **Ctrl**- or **Shift**-keys.

9. Highlight the *Comparisons* panel in the *Main* window and select ***Edit*** > ***Create new object...*** () to create a new comparison for the selected entries.

10. Click on the  next to the experiment name **quality** in the *Experiments* panel to display the quality data in the *Experiment data* panel.

11. Select ***Characters*** > ***Show values*** () to show the corresponding character values for all entries in the comparison.

12. Click on the drop-down list next to the **quality** experiment in the *Experiments* panel to display the default defined character views (see Figure 15).

The quality parameters are grouped based on the data sets and algorithms and the view can be restricted to each of these groups: Raw data statistics (after trimming), De novo assembly, Assembly-free calls (***NrAF***), Assembly-based calls (***NrBAF***), and Summary calls (***NrConsensus***). If user-defined character views have been defined, these are also listed.

Some parameters are more informative and important than others. The most important ones are covered in 5.

5 Quality parameters

5.1 Read quality parameters

AvgQuality: the average quality depends on the sequencing technology used. For Illumina reads, the average read quality should be above 30.

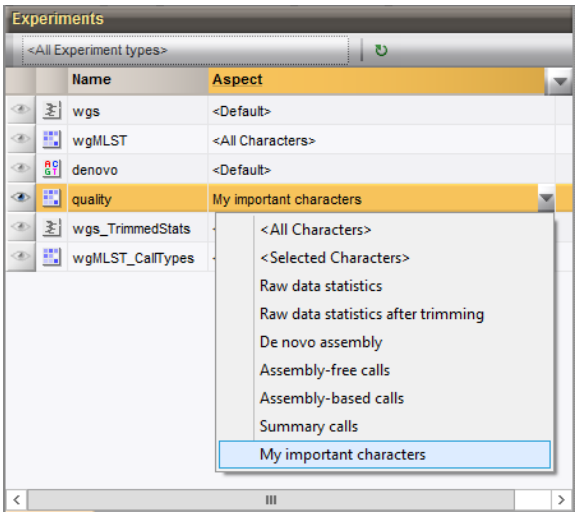

Figure 15: Character views.

AvgQuality_Trimmed: this is the average quality of the reads that were retained after trimming. This value is always slightly higher than the ***AvgQuality*** since bad quality reads are removed in the trimming step, increasing the average overall quality score.

AvgReadCoverage: the expected coverage for each base is calculated based on the number of bases in the reads and the expected sequence length. Samples with coverages below 10 should be removed from the analysis. Ideally this number should be above 30.

5.2 De novo assembly parameters

Length: this length should be close to the length you expect for your organism. Assemblies that are a lot smaller than expected, can be removed from the analysis. For larger lengths, this can be explained by the presence of a plasmid or contamination (see 5.5).

N50: this is the length of the median contig. In general a length above 100 000 is acceptable.

NrACGT: this number should ideally be close to the genome size you expect for your organism.

NrContigs: this number depends on the organism you are working with. In general a value below 400 is acceptable.

5.3 Assembly-free allele calls

NrAFMultiple: some loci might have multiple allele hits so a low number is acceptable. If a very high number of multiple allele hits is observed, this indicates a presence of contamination (see 5.5).

NrAFPerfect: all assembly-free calls that have a perfect (100%) match with an allele in the allele database.

NrAFPresent: all assembly-free calls (= perfect (100%) matches and non-perfect matches).

5.4 Assembly-based allele calls

NrBAFMultiple: some loci might have multiple allele hits so a low number is acceptable.

NrBAFPerfect: all assembly-based calls that have a perfect (100%) match with an allele in the allele database.

NrBAFPresent: all assembly-based calls (= perfect (100%) matches and non-perfect matches). This number should be within an acceptable range you expect for your organism. A very low number should be removed from the analysis. A much higher number than expected can be the result of a mix of two isolates (see 5.5).

Alleles to submit: all hits that do not have a 100% match with an allele in the allele database, and that can (but are not yet) submitted to the allele database (= consisting only of non-ambiguous bases).

Submitted alleles: all hits that were submitted to the allele database.

5.5 Contamination indicators

5.5.1 Contamination with an isolate of a different genus

- The ***Length*** of the de novo assembly will be much higher than expected since the set of contigs of both organisms are concatenated into one large single sequence.
- The number of contigs (***NrContigs***) will typically be much higher than expected since it includes the sum of contigs of both organisms.

Contamination with an isolate of a different genus does not have a large effect on the wgMLST calling, as none of the loci of the contaminating isolate will be recognized as these are not present in the scheme of the organism of interest. Only the loci of the isolate of interest are recognized. If the allele recovery (***NrBAFPresent*** and - if available - the ***CorePercent***) is acceptable, the entry can be included in the analysis. The isolate cannot however be used for the detection of virulence and/or resistance genes - as you cannot be sure from which organism the gene comes from - or as reference for a SNP analysis.

Contamination with an isolate of a different genus can also be observed in the *wgMLST quality assessment* window (select ***WGS tools*** $>$ ***wgMLST quality assessment...*** () from the *Main* window to open the *wgMLST quality assessment* window). When decent sized contigs with detected loci are alternated with other contigs with no loci detected, this is an indication of contamination with an isolate of a different genus. Typically the contigs will also have a different %GC and coverage.

To trace back the contaminated genus, open the genome sequence in the *Sequence editor* window (click on the green colored dot in the ***denovo*** column in the *Experiment presence* panel in the *Main* window), select a sequence of a contig with no loci detected and blast it (***Tools*** $>$ ***BLAST analysis...***). It is recommended not to select a sequence with more than 1000 bp as this will slow down the blast analysis.

5.5.2 Contamination with an isolate of the same species

- The ***NrAFMultiple*** will be very high and represents the number of core genes that are different between the isolates.
- The ***Length*** of the de novo assembly will be higher than expected since it consists of the concatenated shared (core) genome and the (pan) non-shared genomes. The more similar the strains the smaller the pan genomes.
- The ***NrBAFPresent*** is typically higher than expected, since two pan genomes are present in the de novo assembled sequence.

A contamination with an isolate of the same species is a very complex case and typically you cannot use these samples for wgMLST or wgSNP analysis. Note that the exact effect on the quality parameters depends on many factors, including the relatedness of the mixed strains and their relative concentration.