BaseSpace - MiSeq Reporter Software v2.4

Release Notes

For MiSeq Systems Connected to BaseSpace

June 2, 2014

Revision History

Revision	Date	Description of Change
A	May 22, 2014	Initial Version

Introduction

These Release Notes detail the key changes to software components for the MiSeq Reporter workflows on BaseSpace between the specific versions listed in the table below.

Software Application	Prior Version	New Version
MiSeq Reporter	2.2	2.4

This software update takes place in the BaseSpace Cloud environment and will affect all customers running the automatic launch MiSeq Reporter workflows in BaseSPace.

The MiSeq Reporter v2.4 user guides available on illumina.com are up to date for this updated version of the workflows in BaseSpace.

As this update includes two versions of new features and increased functionality, these release notes detail the changes for both.

I. MiSeq Reporter v2.4

NEW FEATURES FROM V2.4:

- The 16S Metagenomics workflow now includes an additional HTML report in the analysis subfolder of the run output folder. This report uses the same data as the standard 16S reports, but includes additional visualizations and summaries. The MiSeq Reporter Metagenomics Workflow Reference Guide has been updated with details on how to use this report. Note for Internet Explorer users: IE9.0 or later is required to view all of the plots.
- The default value for the file copy timeout has been increased. This enhancement was made to improve the management of the final copy step, which transfers results to the MiSeq Output directory. The default time limit is now 30 minutes, an increase from the prior limit of 15 minutes in previous versions of MiSeq Reporter.
- The reporting of algorithm versions and parameter settings applied during analysis has been improved for all workflows, with this information now being included in the BAM and vcf/gVCF results files. The BAM header portion of BAM files now includes the aligner name and version. VCF files now include the dbSNP database version if available

DEFECT REPAIRS FROM V2.4:

- TruSeqAmplicon Workflow:

There was a visualization bug where the 'Details' tab in the MiSeq Reporter user interface contained no information on Sample and Variants. The bug is now fixed.

- Somatic Variant Caller option:
    - Memory usage has now been adjusted to handle gVCF output files properly. This improvement was made to resolve an issue where using the Somatic Variant Caller with gVCF output could result in out-of-memory errors.

  An issue leading to improper “low variant frequency” filtering of reference calls when using the Somatic Variant Caller with gVCF outputs has been corrected.

An issue leading to a 1bp shift in the intervals in which variant calling was applied has been corrected. This issue only occurred with use of the Somatic Variant caller. With previous versions, variants at the start or end of the targeted region could be missed: This is now fixed.

Previously, when using the Enrichment workflow with the Somatic Variant Caller, variant calling would be performed over the entire genome: a fix has been made so that now, when using the Enrichment workflow with the Somatic Variant Caller, variants are now called only inside the targeted region as intended.

- Metagenomics 16S Workflow:
  - The percent abundance table displayed in the detailed view in MiSeq Reporter properly handles species with very low abundance. With this change, all low abundance species (defined as species with less than 0.25% abundance each) are combined into one entry in table, named Other.
- StitchReads setting:
  - A fix was made to the StitchReads function so that now reads are stitched even if the read lengths differ after adapter trimming, or differ simply due to the run configuration. Previously an "index array out of bounds" error would occur in this situation. With this fix, reads with different lengths can now be stitched.

NEW FEATURES FROM V2.3:

- MSR will now generate gVCF files as an additional output file for TruSeqAmplicon, Enrichment, and PCR Amplicon workflows. This can be manually configured by adding OutputGenomeVCF,1 to the [Settings] section of the sample sheet. Note IEM 1.6 does not provide the option to add this setting. Functionally, the gVCF files replace the sites.txt file. For more information on gVCF files see https://sites.google.com/site/gvcftools/home/about-gvcf
- The Enrichment workflow has been updated:
  - For the Enrichment workflow only, added a new optional Sample Sheet setting ManifestPaddingSize which allows the user to specify the size (in basepairs) of padding to be applied to the genomic coordinates of the targeted regions. Prior versions of MSR did not apply any padding.
  - The default padding setting is 150 bp added upstream and downstream of the targeted regions
  - To specify 100 bp of padding upstream and downstream of the targeted regions, the entry `ManifestPaddingSize,100` must be manually added to the [Settings] section of the sample sheet. Note IEM 1.7 does not provide the option to add this setting.
  - If padding is applied, read and base alignment statistics are now reported for the targeted regions with and without padding. Variants are only called within the target region.
  - Changes to the enrichment summary csv output file:
    - An additional line is now included which states the value of the padding size applied (150bp padding by default; e.g.,"Padding size:,150").
    - If padding is applied, there will be 4 additional lines reporting padded base and read enrichment statistics.
    - All notes are now found in the header section of the file
    - Data lines are reorganized by type in the following order: run, read, base, coverage and variant statistics

- The Metagenomics workflow has been updated:
    - A new, faster algorithm has been implemented, resulting in a more than 2-fold reduction in the time required for analysis. This new algorithm is built on the RDP classifier algorithm published by Wang et al, 2007, with Illumina-proprietary modifications to allow analysis of Illumina paired end reads and species-level classification. For more information on the RDP algorithm see http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1950982/.
    - Updated to an Illumina-curated version of the Green Genes 13.5 (May 2013) taxonomy database.
    - Genus-level or species-level classification are now both enabled. Previous version of MSR only provided genus level classifications. For genus-level classification (faster but less granular) you must manually add TaxonomyFile, gg_13_5_genus_32bp.dat to the [Settings] section of the sample sheet. Note IEM 1.7 provides the option to add this setting.
    - Compared to previous versions of the Metagenomics workflow, the genus level results will be highly similar but not exactly the same. Results will be more similar at higher taxonomic levels.
- Added the StitchReads setting for the GenerateFASTQ and TruSeqAmplicon workflows. This must be manually configured by adding StitchReads,1 to the [Settings] section of the Sample Sheet. Note IEM does not provide the option to add this setting.
    - When this setting is added, MSR will combine data from read pairs into a single fastq entry for both reads. For any read pairs for which stitching fails, the FASTQ file will contain both read 1 and read 2. This stitched FASTQ file is not compatible with downstream secondary analysis workflows from Illumina other than as described below for TruSeq Amplicon, but may be appropriate as input to some third-party tools.
    - For read stitching to be applied to a given read pair, the reads must overlap by a minimum of 10 bases.
    - At each overlap position, the consensus stitched read has the base call (and quality score) of the read with higher q-score.
    - Stitched reads can only be analyzed with the TruSeq Amplicon workflow in MSR. When the StitchReads setting is used for this workflow, the stitched read and each individual read will be aligned, and this information is used in variant calling. A bam file entry will be written out for the stitched read, read 1 and read 2. In some cases, the use of this setting may improve accuracy of variant calling.
        - Further details of the StitchReads setting implementation are as follows:
            - For a pair of reads, each possible overlap between read 1 and 2 of 10 bases of overlap or more is considered. (This minimum threshold of 10 helps minimize how many reads are stitched incorrectly because of a chance match)
            - For each possible overlap a score of 1 – MismatchRate is calculated. Perfectly matched overlaps have a MismatchRate of 0, resulting in a score of 1; random sequences have an expected score of 0.25.
            - If the best overlap has a score of at least 0.9, and if the best overlap has score at least 0.1 greater than any runner-up, then the reads are stitched together at this overlap.
- Updated the abbreviation for TruSeq Amplicon runs from "C" to "TA" in the analysis fly-out.
- Starling updated to version 2.0.3. Starling is an optional variant caller that can be used with the TruSeq Amplicon, Resequencing and PCR Amplicon workflows. The default variant caller for these workflows remains GATK version 1.6.

- New and improved warning message for poor indexing schemes are now logged in the AnalysisLog.txt file. Example: Warning: Index sequences differ by < 3 bases
- Clearer logging for failed processes will now be included in AnalysisLog.txt. MSR now logs the non-zero exit code and trace back.
- The default MaximumHoursPerProcess has been increased to 72.
- The dbSNP version used by MiSeq Reporter remains version 131.

DEFECT REPAIRS FROM V2.3:

- Trimmed read length is now properly displayed in the histogram for the Small RNA workflow: the information in this graph was incomplete in previous versions of MSR. The actual adapter trimming function and trimming results were never affected by this issue.
- Warnings or error messages returned from the bcl to fastq generation process are now properly captured in the AnalysisError.txt file.
- For the de novo assembly workflow, a dot-plot can now be generated from either a *.fasta or a *.fa reference file.
- The unit of computer memory (bytes versus gigabytes) is now properly logged when MSR kills a process for using too much memory. Previously the unit of memory was not correct.
- Sample Sheet settings used by MCS but not by MSR will no longer trigger warning messages in MSR.
- The bam headers generated by bwa will now properly include the PL:ILLUMINA entry in the @RG tag.
- A fix was made to the adapter trimming function so that it now properly handles N characters in reads when matching to the specified adapter sequence for trimming. Previously it was treating N characters as A.
- For the TruSeq Amplicon workflow, an improvement was made to the set of off-target sequences considered during alignment. These off target sequences now include sequences that could occur in the extremely rare instance that ULSO and DLSO probes from different probe sets produce an amplicon. This change would reduce the incidence of false positive variants in rare incidences where this occurs.
- For the TruSeq Amplicon workflow, an issue leading to a failure to display graphical results for runs with over 96 samples has been corrected.

KNOWN ISSUES IN V2.4:

- The 16S Metagenomics workflow generally matches the top 3 hits to classify each read. In rare cases, species level identification can lead to misclassification, rather than stepping back to a higher taxonomic level to identify the reads. This particular error can occur when the top hit is not fully classified to a species level match in GreenGenes, but the second best hit is classified.