biomodal software release notes November 2024

Contents

biomodal.com
info@biomodal.com

biomodal software release notes v1.0, November 2024

+44 (0)1223 800 700

biomodal Limited is registered in England and Wales, registered number: 08005377, registered address: The Trinity Building, Chesterford Research Park, Cambridge, CB10 1XL; VAT no: GB 0141 4564 31.

duet pipeline v1.4.0

We have issued the following updates in the current release:

Tertiary Analysis

- **New Differentially Methylated Regions (DMR) calling functionality**
    - A new DMR calling pipeline has been introduced as a follow-on from the main analysis pipeline. This new feature allows for the calling of both DMRs and DhMRs (Differentially Hydroxymethylated Regions) using the outputs from the main pipeline. The DMR calling feature is now available via the biomodal CLI (see CLI section below).
    - To run this, users need a multi-sample zarr store (a default output of the main pipeline) and a DMR sample sheet that maps sample IDs with conditions and covariates. Multiple zarr stores can also be used to combine outputs from different runs for a comprehensive DMR analysis.
    - The output of this functionality is a BED format, with further details on interpretation available in the data interpretation guide.
    - This enables users to easily produce DMR outputs directly through the biomodal CLI.

Alternative Reference Genome Support

- **New Mouse Reference Genome: GRCm38p6**
    - We have introduced a new mouse reference genome, GRCm38p6, which includes the final p6 revision of the GRCm38 assembly.
    - The previous GRCm38 reference contained multiple large fix patches, making it larger and with much of the primary assembly contigs soft-masked. The GRCm38p6 reference simplifies this by including only the final fix/revision of each chromosome.
    - We recommend that customers using the old GRCm38 reference switch to this updated version for more streamlined and accurate analysis.

Summary Reports & Metrics

- **New interactive QC reporting tool for an easier and faster QC review**
    - A new HTML pipeline summary report has been introduced, enabling users to plot any metrics from the default Excel Pipeline Summary Report across all samples on the run.
    - Previously, the pipeline generated a single multi-sample MultiQC HTML report, a set of per-sample HTML DQS Reports, and the Excel Pipeline Summary Report. This release includes an additional interactive HTML pipeline summary report, making it easier for users to compare, contrast and interpret summary report metrics.

- **Correction of Read Loss Categories in 'Resolution & Alignment Stats' Table**
    - In v1.3.0, certain categories of read loss, including those removed as duplicates, were missing from the 'Resolution & Alignment Stats' table in the DQSReport. This issue has been addressed in version 1.4.0, ensuring that all read loss categories are now correctly reflected.

- **Improvement of m-bias plots utility and interpretability**
    - M-bias plots, presented in both the multi-sample MultiQC report and per-sample DQS Reports, have been improved to enhance their utility and interpretability.
    - These plots present the modC, mC and hmC call-rates associated with read cycles. The two following changes have been implemented:
        - By default, the rate is now calculated at CpG sites only, rather than at all sites. This makes the units more interpretable because, for example, the units on the mC curve are now comparable to the CpG mC rate, rather than representing the rate of methylation at all C sites.
        - In duet evoC, the modC plot now represents all modifications (mC, hmC and undifferentiated modC), rather than representing just undifferentiated modC (where the modification couldn't be specified as mC or hmC).
            - This provides a more comprehensive view of all detected modifications, improving clarity and utility in the data analysis.

- **New coverage metrics including and excluding N's in the Pipeline Summary Report**
    - Previously, the Excel Pipeline Summary report featured a metric Mean coverage but it was not clear whether this metric included or excluded N's. We have now split the metric into two separate metrics for:
        - Mean coverage including N's
        - Mean coverage excluding N's
    - This provides more clarity of what the usable coverage is.

- **Improved Accuracy in GC Bias Evaluation using Primary Assembly Reference FASTA**
    - The Picard GC Bias module helps evaluate sequencing performance by generating GC Bias plots, comparing a BAM file against a reference FASTA file displayed in the MultiQC report.
    - Previously, there was an inconsistency as the BAM file only used primary assembly contigs, but the reference FASTA included decoy and alternative contigs, leading to potential inaccuracies.
    - Now, both the BAM file and the FASTA reference are aligned to primary assembly contigs, providing customers with a more accurate and consistent assessment of GC bias.

- **Removal of GC Bias Global Error metric from the Excel Pipeline Summary Report**
    - The metric GC Bias Global Error has been removed from the customer-facing Excel Pipeline Summary Report.
    - Customers are encouraged instead to refer to the CpG to genome-wide coverage ratio, and the TSS to non-TSS coverage ratio, and to the GC bias plots in the MultiQC report, as more useful and interpretable indicators of coverage bias.

Input Data Files

The following changes are introduced:

- **Pipeline metadata file is now optional**
    - In previous versions, a metadata file in csv format was required for the pipeline to run. Now, the metadata file is optional. If provided, the metadata will be included in the Excel and CSV reports generated at the end of the pipeline. However, if the metadata file is not supplied, the pipeline will still run and generate all expected reports at the end.
    - This provides users with greater flexibility to run the pipeline.

Output Data Files

The following changes are introduced:

- **Improved epigenetic quantification with a refined Bedmethyl file for duet evoC**
    - Previously, when Bedmethyl output files were requested for duet evoC data, three separate files were generated to showcase values for mC, hmC, and undifferentiated modC (where the modification could not be specified as mC or hmC).
    - With this release, the content of the modC Bedmethyl file has been updated. It now combines mC, hmC, and undifferentiated modC, making the output more consistent across the 5bp and 6bp pipeline modes and comparable with duet +modC data.
        - **Note**: if requesting Bedmethyl outputs when processing duet evoC data, three separate bedmethyl files are still generated for mC, hmC, and modC.

- **Improved epigenetic quantification with a refined cytosine report for duet evoC**
    - Previously, the duet evoC cytosine report had mC and hmC counts in separate columns, which made it incompatible with some third-party tools. Additionally, it excluded undifferentiated modC calls (where the modification couldn't be specified as mC or hmC).
    - The format has been updated to generate separate reports for mC and hmC, and a modC cytosine report that combines mC, hmC, and undifferentiated modC into a single column for easier integration with downstream tools.

- **New optional Bismark plain text quantification format**
    - An optional capability has been introduced to generate quantification outputs in a Bismark plain text format, providing compatibility for users who work with Bismark.

- **New optional Bedgraph plain text quantification format**
    - An optional capability has been introduced to generate quantification outputs in a Bedgraph plain text format, providing flexibility for users who need this format for their

analysis workflows. Methylation is presented in a 'track' format that can be loaded in IGV for easy visualisation.

Core Processing Logic

- **Prelude: added resolution category to reads during resolution**
    - Prelude now adds a tag during read resolution to indicate how a read was resolved:
        - XR:i:1 for reads resolved naively.
        - XR:i:2 for reads resolved via pairwise alignment.

- **Change to epigenetic quantification workflow for a faster and more robust pipeline**
    - A new version of epiquant was introduced which quantifies directly to a zarr format, resulting in a faster and more robust pipeline for epigenetic quantification.

Pipeline Resource Utilisation

- **Optimisation: updated resource profiles**
    - Resource profiles have been updated to improve resource allocation and reduce the risk of resource exhaustion, especially in the local and local_deep_seq profiles when processing mouse data.

- **Bug Fix: Prelude failure when number of threads exceeds number of chunks**
    - A bug has been fixed where Prelude would fail if the number of threads used in multiprocessing exceeded the number of chunks into which the FASTQ file was divided. This issue affected samples with fewer than:
        - 350,000 reads (default profile),
        - 700,000 reads (deep_seq profile),
        - 1,000,000 reads (super_seq profile).

Command Line Interface (CLI) v1.1.1

We have issued the following updates in the current release:

New functionality

- **Introducing “biomodal call_dmr” command to run DMR pipeline**
    - As mentioned in the section above, the DMR calling feature is now available via the biomodal CLI. The command lines and full instructions can be found in the duet software installation and running guide.
- **Introducing "--quantification-output" parameter to “biomodal analyse”**
    - A new command was introduced “--quantification-output" for users wishing to specify the output format for quantification files. Available formats include a combination of bedmethyl, cxreport, bismark, bedgraph, with cxreport set as the default.
- **Introducing recommended duet resource profiles when running "biomodal validate”**

Parameter updates

- Sharing events as well as metrics reports is now optional: during the “biomodal init” process, customers can choose their sharing preferences. In earlier versions, we asked only about sharing metrics reports; now, optional events are also restricted by preference.
- Running "biomodal validate" now recommends duet resource profiles.
- Parameter “--meta-file" is now optional when running "biomodal analyse".
- Reduced number of containers customers download when upgrading to duet 1.4.0.

Documentation updates

- A new section was added for the DMR pipeline.
- Details around optional events and metrics report sharing were added in the biomodal CLI.
- A description and link to “nf-core community pipelines” was added for reference.