biomodal software release notes duet pipeline 1.5.0 and CLI 2.0.0 December 2025

Contents

biomodal.com info@biomodal.com

+44 (0)1223 800 700

duet pipeline 1.5.0

We have issued the following feature improvements and bug fixes in the current pipeline 1.5.0 release (covering changes from 1.4.2):

Essential pipeline parameters

- **Pipeline mode parameter (5bp/6bp) now mandatory**
  - Previously, the default parameter was set to 5bp, which could lead to unintended processing of duet evoC data if the mode wasn't manually adjusted.
  - To prevent incorrect data processing the pipeline mode parameter now requires explicit user input:
    - 5bp for duet +modC data
    - 6bp for duet evoC data
  - If not set, the pipeline will fail validation and return a clear error message: ERROR ~ mode parameter must be set to either '5bp' (for duet +modC) or '6bp' (for duet evoC).

Summary reports and metrics

- **Improved summary reports: MultiQC report consolidation**
  - Previously, the pipeline generated both per-sample Data Quality System (DQS) reports and a MultiQC report, with some plots only available in one of the two reports. The per-sample DQS reports were static, making it difficult to compare plots across samples.
  - Per-sample HTML DQS reports have now been removed and unified into the HTML multi-sample MultiQC report:
    - All plots are now consolidated into one interactive HTML report.
    - Enables easier cross-sample comparisons and streamline review.
    - Reduces the need to open multiple static reports.
  - This change improves accessibility and interpretability of quality metrics across several sample sets.

- **Metric correction: mean methylation rates in the General Statistics table in MultiQC**
  - Previously, mean methylation rates in MultiQC could differ from those in the Excel Summary Report, which could cause confusion when comparing outputs.
  - MultiQC reported the mean across all resolved read positions, of the mean methylation rate at those read positions, without any adjustment for the differing quantity of bases at those read positions. In contrast, the Excel Summary Report provides autosome-

wide rates per methylation stage (e.g. mC) and context (e.g. CpG), such as “autosomal chromosomes rate of mC at CpG”.

- MultiQC has now been updated to match the autosome-wide methylation rates shown in the Excel Summary Report, ensuring consistency across outputs.

- **Minor metric name update: autosomal methylation rate metrics in csv Summary report**
  - In the Excel Summary Report, there are metrics that report autosome-wide rates per methylation state (e.g. mC) and context (e.g. CG), e.g. 'Autosomal chromosomes rate of mC at CpG'. The corresponding underlying metric field names in the csv Summary Report were previously of the format:
    - modality_summary_{context}_genome_{modification}, e.g. modality_summary_cg_genome_mc.
  - To make it clearer that the metric field names report autosomal rates only, the format has been changed to modality_summary_{context}_autosomes_{modification}, e.g. modality_summary_cg_autosomes_mc

Resolution algorithm

- **Generation and selection of q-tables during resolution**
  - Q-tables are empirically generated look-up tables used during read resolution to assign resolved Phred scores to resolved bases considering factors such as the assay (duet +modC or evoC), the sequencing chemistry (e.g. XLEAP), the sequencing instrument, and the Illumina base-calling software version. In this release, q-tables have been updated to improve their accuracy and to improve the selection logic that chooses which q-table to use. These changes include the following:
    - q-tables now include data generated from Illumina control software 1.3 for the NovaseqX platform.
    - Previously, if no compatible q-table was found, then the resolution algorithm would fall back to using the ‘minimum Phred’ approach to resolve quality scores. There is no longer an automated fallback approach; instead, a check is performed before commencing resolution and if no compatible q-table is found then the pipeline terminates advising the user that there is no compatible q-table, but that the pipeline can be relaunched setting the min Phred strategy via a parameter override.
    - By default, Q-tables are automatically selected based on inferring the instrument type from instrument ID in the FASTQ read headers. However, it's possible that your sequencing facility has renamed their sequencer instrument ID, or that the sequencing instrument has a non-standard ID. If this is the case, the automatic q-table selection step will exit reporting that no valid q-table was found. It is now possible to map non-standard instrument IDs to an expected instrument prefix by using the “--additional-params” CLI option to override the “prelude.instrument_override” parameter (see documentation for further details).

- Previously, if the empirical data used to generate the q-table had featured no miscalled bases for a given combination of settings then the q-table would record the maximum possible Phred score for that resolution option. This was considered unrealistic, so plus-one smoothing has now been introduced to the q-table generation process to ensure that the maximum possible q-score does not get assigned.

- Previously, NovaSeqX instruments were identified via an 'LH' prefix in the instrument ID, but we have identified instances where NovaSeqX instruments feature an 'LL' prefix. The rule for identifying a NovaSeqX instrument has been relaxed so that an 'L' prefix is sufficient to identify the instrument as a NovaSeqX.

- **Trimming of N's from the termini of the resolved reads**
  - The resolution step includes the masking of C's in the last three positions of resolved reads. This reduces the adverse impact of end repair on sensitivity. This can result in up to three consecutive N's at the termini resolved reads.
  - Previously, these were left on the resolved reads.
  - They are new trimmed off at the end of the resolution process.
  - Note that this might slightly decrease the value of the CpG-to-genome-wide coverage metric and in some cases the genetic accuracy metric compared to previous pipeline versions.

- **Optional post-resolution hard trimming**
  - The option to perform post-resolution hard-trimming has been introduced.
  - This can be used, if desired, to trim a fixed number of bases of either the beginning or the end of all resolved reads before they are passed on to alignment.
  - Users can now specify fixed-length trimming at the 5' or 3' ends of resolved reads by supplying a 5 or 3 integer with either of the following additional parameters:
    - prelude.rr_front_trim
    - prelude.rr_back_trim

    E.g. "--additional-params prelude.rr_back_trim=5"

Quantification

- **Plain-text quantification filename changes**
  - Previously, plain-text quantification files were gzipped.
  - Plain-text quantification files are now bgzipped, and the file extensions have been changed to better reflect the file contents and for better compatibility with Integrative Genomics Viewer (IGV).
  - New file extensions are:
    - *bismark.bed.gz for the bismark file
    - *bedgraph.bdg.gz for the bedgraph file
    - *bedmethyl.bed.gz for the bedmethyl file

- Note that the file extension for the cytosine report remains unchanged as cxreport.txt.gz.

Allele-Specific Methylation (ASM) output

- **Separate ASM calling for mC, hmC and modC**
  - The duet pipeline includes an ASM module that can be activated to generate ASM outputs.
  - Previously, when processing duet evoC data and performing ASM calling, a single output was generated based on calling ASM once using the modC category (i.e. the combination of mC and hmC).
  - The ASM module now produces separate files for mC, hmC, and modC when processing duet evoC data, offering more granular insights into methylation patterns.
  - Note that the modC ASM file will be equivalent to the modC ASM file generated from earlier pipeline versions.

DMR calling

- **DMR calling functionality has been removed and is not included in this release.**

Support for alternative pipeline and use cases

- **New human reference genome**
  - A new human reference genome, the T2T (telomere-to-telomere) CHM13v2, is now supported and can be selected for analysis.
  - This can be activated by setting the ref-genome parameter to T2T_CHM13v2, provided the associated reference package has been downloaded.

- **Reference file pipeline compatibility**
  - Where relevant, the reference file pipeline now uses bgzip, rather than gzip. This improves compatibility with some third-party tools.
  - As a result of this change, reference files have been regenerated.

Resource utilisation and runtime efficiency

- **Optimised plain-text methylation quantification writers**
  - Export of plain-text methylation quantification files has been optimised to reduce CPU and memory usage.

- **Increased resources for merging lanes in deeply sequenced mouse data**
  - The pipeline now allocates additional computational resources to the SAMTOOLS_MERGE_LANES step when processing mouse data with the super_seq profile.

- This update improves performance for users working with very deeply sequenced mouse samples.

- **Increased resources for joint genotyping of large sample sets**
  - If joint variant calling is activated, the pipeline now allocates additional computational resources to the GENOMICS_DB_IMPORT step. This enables smoother processing and improved scalability when working with large numbers of deeply sequencing samples.

Miscellaneous bug fixes

- **Resolved script error for Google Batch users**
  - An issue (missing shebang) that caused the export step to fail when running on Google Batch executors has been fixed.
  - This ensures that workflows using Google Batch now successfully complete without errors related to script execution.

- **Fixed error caused by third-party software update affecting quantification**
  - In some environments, users encountered an error when running the epiquant module built from duet pipeline release 1.4.* after 7 April 2025 (ImportError: cannot import name 'cbuffer_sizes' from 'numcodecs.blosc').
  - This was due to a compatibility issue between the version of the zarr used in the epiquant module of the pipeline and an updated version of its dependency, numcodecs.
  - The pipeline now uses an updated version of zarr (2.18.7), which resolved this compatibility issue and prevents the error from occurring.

- **Improved epiquant zarr store instantiation & logging**
  - The pipeline now uses absolute file paths (previously relative) when creating synchronization files for quantification in the epiquant module, which may reduce the risk of the process hanging during zarr store instantiation in certain environments or file systems.
  - Logging for this step has also been improved to aid troubleshooting if issues arise.

- **Consistent merging of sequencing lanes for reproducible results**
  - At the stage where the lane-wise aligned BAMs for a sample are merged, the order in which the lanes were merged was non-deterministic and was determined by race conditions associated with the upstream lane-wise alignments.
  - The pipeline now merges lane-wise aligned BAM files in a consistent order, eliminating run-to-run differences in read order that could previously cause small, non-significant variations in calculated accuracy metrics (such as sensitivity and specificity).

- This update ensures more reproducible results when processing the same input data multiple times.

- **Read category metrics are now reported even if the count is zero**
  - The bamlet module calculates read category count metrics, such as the number of reads that align to decoy contigs, the number of reads that are unmapped, or the number or reads that are marked as duplicates. These are passed on to downstream reports.
  - Previously, in the rare event that a category featured no reads, the associated read count metric would not be generated and would be absent from downstream reports.
  - This has been fixed, so that if a category features no reads, the associated read count metric is still generated and is set to zero.

- **Validation check for input files**
  - A bug has been fixed that was preventing the correct operation of validation checks that validate the presence of relevant input files prior to launching the pipeline.

- **Improved handling of temporary directories**
  - Local temporary directories are now instantiated when running the HAPLOTYPE_CALLER and MUTECT2 processes. This circumvents an error such as the following which occurred on some environment:
    - Error in tempfile() using template /tmp/parXXXXX.par: Parent directory does not exist Error in tempfile() using template /<your local tmp path>/parXXXXX.par: Parent directory (/<your local tmp path>/) does not exist at /venv/bin/parallel

CLI 2.0.0 Release

We are introducing a completely redesigned Command Line Interface (CLI) with this release. The new CLI 2.0.0 has been built from the ground up to offer a more robust, streamlined, and user-friendly experience.

Unlike previous CLI versions, CLI 2.0.0 eliminates the need for complex installation scripts and extensive software dependencies. It's now delivered as a single binary that runs locally, requiring only Java and a container orchestration tool (Docker, Apptainer, or Singularity) to execute biomodal pipelines on Linux platforms.

biomodal pipeline supported

- CLI 2.0.0 is only compatible with duet pipeline 1.5.0.

Key enhancements

The main enhancements over previous versions of the biomodal CLI are:

- Minimal software dependencies: the v1 CLI required 11 3rd party software dependencies. CLI 2.0.0 now only requires two: Java and a container runtime (Docker, Apptainer, or Singularity). All other functionality is built into the binary.
- Simplified installation: download a single binary - no more script bundles or additional setup tools.
- Modern architecture: rewritten using a modern programming language for improved performance and maintainability, replacing legacy Bash scripts.
- Intuitive interface: redesigned command structure with built-in help. For example:
    - "biomodal download" – for downloading software and reference data
    - "biomodal run" – for executing pipelines
- Diagnostics packaging: easily generate a ZIP archive of relevant settings and log files using the new "biomodal get diagnostics" command. This is particularly useful if you have any issues or queries to the biomodal Product Support & Training team.
- Optimised image downloads: direct download of pre-built module images for Apptainer and Singularity, eliminating the need for local Docker image conversion.
- Telemetry Control: you can now opt-in to share Metrics Reports independently of CLI event tracking.

Installing the new CLI

- **Important:** if you want to keep both CLI versions, we recommend installing CLI 2.0.0 in a separate location from v1 to allow both versions to coexist.

- **Note:** Ensure Java and a container runtime (Docker, Apptainer, or Singularity) are installed beforehand. Check the software requirements and permission section in the CLI 2.0.0 documentation to learn more about this.
- Next, to install the new CLI, simply run a single command to download the binary. Check the Installing the CLI section in our documentation.
- Once downloaded, the CLI will guide you through the configuration process and install itself in your chosen location.
- Container images and reference data can be shared across versions.

Platform support

CLI 2.0.0 supports:

- Local Linux servers
- HPC systems
- GCP Batch environments
- AWS Batch environment

We've also released a new section in CLI 2.0.0 documentation tailored for Platform/DevOps Engineers configuring GCP Batch. For advanced users, Terraform configurations are available to streamline AWS and GCP Batch setup.