biomodal software release notes December 2023

Contents

biomodal.com
info@biomodal.com

duet pipeline v1.1.2

Summary Reports & Metrics

- The SUMMARY_REPORT module that collates pipeline metrics into a single Excel file and accompanying csv file has been extensively re-written and has been renamed as PIPELINE_REPORT.
- New metrics have been introduced to the Excel Pipeline Summary Report which quantify the number of retained reads and bases at various stages of the pipeline as a proportion of the original number of reads and bases in the input FASTQ files:
    - Initial input reads fraction remaining after trimming and quality filtering
    - Initial input bases fraction remaining after trimming and quality filtering
    - Initial input reads fraction remaining after resolution
    - Initial input bases fraction remaining after resolution
    - Initial input reads fraction aligning to the genome
    - Initial input reads fraction aligning to the genome primary assembly after deduplication

  These metrics are further explained in the data interpretation guide.
- Trimming metrics reported via the Excel Pipeline Summary Report have been changed to more clearly categorise and quantify the types of trimming and filtering that have been performed. This includes reporting:
    - The number of 'internal' hairpins trimmed
    - The number of 'anchored' hairpins trimmed
    - The number of poly-G tails removed

  These metrics are further explained in the data interpretation guide.
- By default, FASTQC is now run on the raw (pre-processed) FASTQ files, instead of on the resolved FASTQ files.
- The names and descriptions of some metrics in the Excel Pipeline Summary Report have been updated to correct typos or to improve clarity/interpretability. Notably:
    - The metric called '*Genome mapped deduplicated reads*' has been renamed as '*Genome deduplicated reads mapping to primary assembly*' to clarify that it excludes any reads that map to alternate contigs or decoy sequences outside of the primary assembly.

Output Data Files

- A parameter has been corrected which was preventing the deduplicated BAM files from being published when the 'dedup_by_contigs' parameter was set to true.

Primary Processing: Trimming & Resolution

Reminder: Trimming is the step that removes the duet hairpin construct and other potential artefacts from reads. Resolution is the process that converts pairs of reads into resolved single-end reads with encoded methylation tags.

- The parallelisation strategy for the COUPLET module which performs read resolution has been changed, bringing the handling of all parallelisation inside the Python code, improving resilience and error-handling.
- When a fragment of double-stranded input DNA features a 5' overhang (sometimes referred to as a 'jagged end'), the end repair step in the assay uses a polymerase to extend the 3' end, but when this occurs the methylation pattern is lost from the repaired stretch. In duet pipeline v1.1.1, after resolution, a hard trim of 3nt from the tail of all resolved reads was performed to reduce the potential impact of end repair on methylation sensitivity. This approach increased methylation sensitivity, but reduced the number usable bases for downstream processing, reducing mean coverage. In pipeline version 1.1.2, no post-resolution trimming is performed; instead, cytosines in the last three bases of resolved reads are masked as N's, resulting in them getting excluded from methylation calling.
- Trimming settings have been refined to reduce excessive trimming of potential artefacts and to be stricter in the identification of hairpin sequences for removal; this reduces the number of bases discarded during trimming and increases the overall coverage achieved for a given number of input reads. Changes to trimming include the following:
    - A-tails are no longer removed from reads
    - G-tails are now only removed if there are 9 or more consecutive G's
    - Instead of accepting either protected or deaminated bases at the C positions of the hairpin (deamination-aware trimming), hairpin identification is now stricter, expecting protected bases
    - Instead of allowing a 25% error rate during hairpin identification, the error rate has been reduced to 3/28 = 10.7%
    - Instead of removing a hairpin on the basis of an overlap of only 2 nucleotides, an overlap of 3 nucleotides is required
- During the resolution step, for data generated on NextSeq2000 sequences, Phred scores are now resolved using empirical Q-tables, rather than being resolved by taking the lesser of the two Phred scores.

Support for Alternative Reference Genomes

- A MultiQC report rendering bug which produced an additional row per sample with empty columns on some non-human reference genomes is now fixed. The fix means the report shows only one row per sample and populates all metric columns.
- A bug has been fixed which caused a BAM file re-header step to fail when using some non-human reference genomes.
- A change has been introduced that increases the memory requested by the BWA_MEM2 process when aligning against mouse reference genomes.

Resource Utilisation & Multi-Platform Support

- The EPIQUANT_MBIAS module, which calculates whether there is any bias for methylation calling associated with sequencer read-cycle has been parallelised and optimised in order to improve runtime.
- An imprecise output directive in the SAMTOOLS_MERGE_LANES process caused the unnecessary copying of BAM files to the scratch directory during processing. This has been refined, reducing the disk requirements of the scratch directory.
- A misnamed resource directive in the deep_seq profile prevented the disk requirements for the SAMTOOLS_MERGE_LANES step from being appropriately increased. This could sometimes lead to disk exhaustion on cloud platforms.
- The memory requested for the DQSPY module has been increased to accommodate up to 192 samples on a single run.
- The strategy for refreshing the timestamp on index files has been improved so that this is only done in cases where the executor stages index files onto a virtual machine for processing. This resolves an error which could occur on non-cloud platforms when permissions to the reference directory were read-only.
- The ‘local’ and ‘local_deep_seq’ profiles intended for use when no cloud platform or HPC job scheduler were available, have been refined and improved in order to reduce, where possible, the resources requested by processes.
- A typo has been corrected which prevented the ‘local_deep_seq’ profile from being operational.
- A bug has been fixed associated with temporary Python paths used in the DQSPY module that affected specific platforms only.

Support for Targeted Sequencing

- In the targeted mode of the pipeline, after the calculation of on-target and enrichment metrics, a BAM file restricted to only the target/probe regions is generated for downstream steps, such as variant calling and ASM.
- In the targeted mode of the pipeline, additional metrics have been added to the Excel Pipeline Summary Report. These are described in detail in the Bioinformatics Pipeline Data Interpretation Guide.
- The Twist pan-cancer and Twist methylome panels are now supported via the 'target_panel' parameter.

Command Line Interface (CLI) v1.0.3

New functionality

- Targeted mode added as main parameters for biomodal analyse command

Parameter updates

- Updated software dependency from pcregrep to pcre2grep
- Added --work-dir as an optional parameter to the biomodal analyse command
- Renamed install scripts to better reflect use cases:
    - bootstrap-tool -> biomodal-cloud-utils
    - bootstrap-on-hpc -> biomodal-hpc-utils-admin
    - bootstrap-on-hpc-nosudo -> biomodal-hpc-utils-conda
- Added biomodal analyze as an alias to biomodal analyse
- Added queueSize parameter to limit the number of parallel pipeline processes
- Extended biomodal info command display additional information about the current installation

Cloud/HPC installation and running optimisation

- Added check and warning for insufficient initial free disk space to HPC install scripts
- Various HPC install scripts improvements and bug fixes
- Updated software dependency validation for orchestrator and worker nodes respectively
- Added Google SDK installation in biomodal-hpc-utils-conda
- Updated Terraform modules for AWS and GCP
- Replaced AWS bastion module for Terraform
- Added test option for Cloud install scripts
- Labels supported for Cloud platforms to enable resource cost tracking

biomodal.com
info@biomodal.com

Documentation updates

- Show link to online documentation directly in multiple CLI commands
- Restructured user documentation into sub-pages
- Added further information on installing software as modules on HPC
- Added LSF and SLURM scheduler recommendations
- Added decision flowchart to select correct installation script and options
- Added how to run the CLI as a HPC module
- Updated pipeline module hardware resource recommendations
- Amended new bootstrap/install script names
- Variant Associated Methylation (VAM) introduction
- Use of public IP on cloud bootstrapped VMs