biomodal software release notes June 2024

Contents

biomodal.com
info@biomodal.com

+44 (0)1223 800 700

duet pipeline v1.3.0

We have issued the following updates in the current release:

Summary Reports & Metrics

- **Standardised metric names for ONE_STEP_BAMLET**
    - There have been some small changes to the underlying metric field names calculated by One-Step-Bamlet to better standardise their format. This includes standardising the format of the underlying metric names to conform with the pattern bamlet_{metric}_{ref_group}. It also includes replacing a duplication rate metric calculated on bases to one calculated on reads. This mostly affects the field names in the csv files.
    - The metric affected in the Excel report is bamlet_genome_reads_inc_dups, now renamed to bamlet_inc_duplicates_reads_genome.
    - The metrics affected in the csv report are:
        - bamlet_genome_reads_inc_dups, now bamlet_inc_duplicates_reads_genome
        - bamlet_discarded_genome_duplicate_ reads, now bamlet_duplicate_reads_genome
        - bamlet_prop_duplicated_bases_genome, now bamlet_prop_duplicated_reads_genome; note the change from duplication rate based on bases to now based on reads
        - bamlet_read_counts_genome, now bamlet_mapped_reads_genome

- **Change to autosomal and mitochondrial modC rate calculation**
    - The pipeline calculates rates of modified C and unmodified C in the autosomes and in the mitochondria. These are calculated relative to CpG sites from the reference. Previously, the denominator for these metrics included **modC** calls, **C** calls and '**other**' calls, where the 'other' calls include:
        - Other bases called (G, T, A) that align to the C in a reference CpG – these might be SNPs or mis-called bases.
        - N bases - these may be from the sequencer, or from error suppression during read resolution, or from the masking of Cs in the tail of reads which is performed to lessen the impact of end repair on sensitivity.
    - The inclusion of these 'other' calls in the denominator made the autosomal modC rate and the mitochondrial unmodified C rate artificially low, especially when compared to the outputs from alternative technologies and tools (e.g. methyldackel), which don't feature these non-C calls in the denominator – therefore, the 'other' calls are now excluded from the denominator, which we believe makes the metrics more accurate, interpretable, and comparable.

- **New ASM metrics in the Pipeline Summary Report**
    - The ASM module calls allele-specific methylation at heterozygous SNPs (sites) and these are output to a file, but previously there were no metrics associated with ASM presented in the Pipeline Summary Report. The ASM output file categorises sites based on whether they:
        - Pass the necessary filters for ASM to be called
        - Pass the necessary filters for ASM to be called, but exhibit a low methylation difference between the two alleles
        - Do not pass the necessary filters for methylation to be called for one of several reasons (e.g. low coverage)
    - A new set of ASM metrics has been introduced to the Excel pipeline report which summarise the quantity of sites in each ASM filter category, providing a high level summary of the ASM output. These metrics are:
        - Site passing filters
        - Sites with low methylation difference
        - Sites with low CpG coverage at both alleles
        - Sites with low CpG coverage at allele 1
        - Sites with low CpG coverage at allele 2
        - Sites with no CpG coverage at both alleles
        - Sites with no CpG coverage at allele 1
        - Sites with no CpG coverage at allele 2
        - Sites with no reads

Output Data Files

The following changes are introduced:

- **Mutect2: Introduce GATK recommended filtering**
    - GATK best practice workflows for somatic variant calling recommend running somatic variants through FilterMutectCalls to remove false positives. This change introduces this post hoc filtering step in the MUTECT2 module.

- **Optionally publish CRAM files**
    - By setting publish_crams = true, the pipeline will publish CRAM files instead of BAM files. These are more heavily compressed and use less storage than BAM files because the data is converted into a difference from the reference.
    - To read CRAMs, it is necessary to have the reference FASTA that was used to perform the original alignment.
    - Note that in this mode the reference FASTA is also published to the output directory.

+44 (0)1223 800 700

Primary Processing: Trimming & Resolution

The following changes are introduced:

- **Trimming and resolution are now performed by a biomodal-bespoke module named Prelude**
    - A new pipeline module has been introduced which performs trimming and resolution of reads in a single step. This replaces the previous Cutadapt and Couplet modules which performed trimming and resolution separately.
    - Prelude is faster to run and reduces the quantity of data written to the scratch directory.

Alternative Reference Genome Support

- **Introduce support for a Hemp (Cannabis sativa) reference**
    - Cannabis sativa has been added to the genomes.config file and reference files have been created, so this species is now supported via the pipeline and the CLI.

The following new feature has been implemented:

- **Prepare own reference genome**
    - With CLI release 1.1.0, we have made a reference-generating pipeline to prepare references outside of what we provide – all that is needed is a gzipped fasta file. Please see CLI section later in this document for more information.

Pipeline Resource Utilisation

- **Reduce scratch footprint of SAMTOOLS_MERGE_LANES**
    - The scratch directory footprint of the SAMTOOLS_MERGE_LANES process has been reduced substantially by avoiding the unnecessary saving to the scratch directory of a large BAM file.

- **ONE-STEP-BAMLET: parallelise stats collection**
    - The collection of statistics about the aligned BAM file has been parallelised and combined with the filtering of the BAM file.

- **Deduplication: pipe directly to samtools markdup**
    - When samples feature in only one lane, the BWA_MEM module now streams the aligned BAM file directory into samtools markdup for the marking of duplicate reads.
    - When samples feature multiple lanes, the SAMTOOLS MERGE_LANES module now streams the merged BAM file directory into samtools markdup for the marking of duplicate reads. This avoids unnecessarily writing out the interim pre-dedup BAM file to the scratch directory and then reading it into a separate process to perform duplicate marking.

- The duplicate marking tool is also changed from the single-threaded tool Picard MarkDuplicates to the multi-threaded tool samtools markdup, which reduces the wall clock time. This also means that the dedup_by_contigs parameter no longer needs to be manually set.
- These changes do not take effect when UMIs are in use.

+44 (0)1223 800 700

Command Line Interface (CLI) v1.1.0

We have issued the following updates in the current release:

New functionality

- **Reference pipeline available to install and run for CLI users**
- **Introducing pre-run check via biomodal validate command**
    - The new 'validate' command is used with the same parameters used in the 'analyse' command. This command will not run the duet pipeline, but rather inform the user if there is a problem with any of the parameters used.
    - 'biomodal validate' will check:
        - All input, output and work/temp directories referenced, and files are present
        - Read-write access to paths
        - Consistency of meta file
        - FASTQ filename patterns
        - Parameter combination and format/logic of optional '--additional-parameters'
        - duet software present in expected locations
        - Sufficient free space on local filesystems
        - Internet and biomodal API access
- **HTML report for runtime metrics**
    - We have included a new Nextflow report in the 'diagnostics' folder of the run output folder, called 'nf_report.html'. This report will detail all pipeline options used and contain useful interactive plots that gives an overview of the distribution of resource usage for each process and task in the duet pipeline workflow.

Parameter updates

- To enable use of any new reference genomes customers generate using the new reference genome pipeline, we have included a new parameter supporting inclusion of the genome configuration file created.
    - This new parameter is called 'reference-genome-profile' and is used with the 'analyse' command. This parameter is the complete file name for the new configuration file and should only be used when running the pipeline with your new reference genome.

Documentation updates

- Updated resource requirements for each module in the duet pipeline, in relation to running with standard settings or when using the two additional profiles ‘deep_seq’ or ‘super_seq’.
- Updated reference to new online password reset link.
- Added a section for resource requirements for each module in the new reference genome pipeline.
- Added a section in the HPC recommendations to help with setting additional memory settings for Sun Grid Engine (SGE), Oracle Grid Engine (OGE) or Univa Grid Engine (UGE) schedulers.
- Added section for manually updating the CLI from 1.0.x to 1.1.0, without running one of the installers.
- Added section detailing how to run duet pipeline with support for somatic variant calling.

+44 (0)1223 800 700