biomodal software release notes March 2024

Contents

duet pipeline v1.2.0

Summary Reports & Metrics

- The order of sections in the Excel Pipeline Summary Report has been changed so that the more important metrics, such as the ones calculated on the controls, are at the top of the report, and less important metrics are towards the bottom.
- Previously, all possible metrics calculated by the pipeline were output in the Pipeline Summary Report csv file, including metrics of low importance or relevance. Now, the summary report csv

biomodal.com
info@biomodal.com

+44 (0)1223 800 700

output from the Pipeline Report module features only those metrics that are also output in the Excel file. An accompanying mapping from field names, to friendly names and descriptions is also output. A separate all metrics csv containing the full set of metrics is output into the diagnostics subdirectory.

- Previously, the pipeline reported a metric associated with the mean length of reads of particular classes ('Genome reads mean length'), but this metric included soft-clipped bases and was rounded to the nearest whole number. The calculation of this metric has been improved so that soft-clipped bases are excluded and the values are no longer rounded.
- The calculation of all metrics related to the retention or loss of reads or bases has been reviewed and improved to create more accurate information with better visibility. These metrics are used to populate a more comprehensive table and bar chart in the 'Resolution and Alignment Stats' section of the sample-specific Data Quality Summary Reports (dqsreport) generated by the pipeline.
- If any sample IDs were entirely numeric, a failure occurred in the DQSREPORT module and sample-level reports were not generated. This has now been fixed.
- The Excel pipeline summary report features a coverage bias metric called 'TSS to non-TSS coverage ratio' which was intended to report the ratio of coverage at TSS (transcription start site) loci to coverage at non-TSS loci. It was identified that this metric was incorrectly reporting the ratio of bases covered at 1x or above at TSS loci to bases covered at 1x or above at non-TSS loci, rather than the ratio of coverage. This calculation has been corrected. The metric has also been disabled in targeted mode where such a ratio is likely to be confounded by the choice of target panel.
- A bug has been fixed which caused some additional superfluous plots with an unclear title to appear in the MultiQC Report if any sample IDs featured a capital 'R' followed by a digit.
- A hmC sensitivity metric has been introduced in the customer-facing Excel pipeline summary report.

Output Data Files

- A new filetype (zarr) has been introduced to enable the use of the duet data analysis suite for cohort analysis studies at scale.
- The output directory has been extensively reorganised to group outputs into the following categories:
    - Controls
    - Diagnostics
    - Reports
    - Sample outputs
- The default format for MM tags which encode the modification status of cytosines in reads in FASTQ and BAM files has been changed for both the +modC (5-base) and duet evoC (6-base)

modes of the pipeline. The changes are described in the Data Interpretation Guide and improve the compatibility of these file types with IGV and with pysam.

- Capability to generate an evoC (6-base) Cytosine Report has been introduced. The format of the 6L Cytosine Report is described in the Data Interpretation Guide.
    - Note that undifferentiated modC calls, where a modification is called but it is not possible to differentiate whether it is mC or hmC, are excluded from the mC, hmC, and C columns, but feature in the 'Coverage' column. The 'Coverage' column additionally includes SNPs where the call is not a C, and cases where there is an N in the read at the given position.
    - Additionally, when quantifying evoC (6-base) data in CHG and CHH contexts, no Cytosine Report is produced.
    - Although the Cytosine Report is an externally defined format, this 6L version is a locally-defined modification of the externally defined format. It's compatibility with third-party tools has not been evaluated.
- The generation of bedMethyl files for duet evoC has been disabled by default. This can be enabled if desired and is described in the CLI documentation.

Primary Processing: Trimming & Resolution

- Poly-G trimming settings have been revised to improve the identification and removal of polyG artefacts and to better accommodate differences in the properties of polyG artefacts on NextSeq and NovaSeq instruments. These changes primarily affect libraries that feature IDT UDI-UMIs and primarily influence whether reads are discarded at the trimming stage or the resolution stage.
- The ability to detect NovaSeqX reads has been introduced; empirical Q-tables generated from samples sequenced on a NovaSeqX have been introduced for use in resolving Phred scores when processing NovaSeqX reads.

Resource Utilisation & Multi-Platform Support

- The usage of bgzip to compress files in the EPIQUANT_QUANT module has been changed to avoid the unnecessary retention of the original unzipped after completion of the compression process.

- A more efficient method of writing output files has been introduced for the COUPLET module.
- Couplet memory profiling optimisations have been introduced to improve efficiency.
- Multiple modules associated with the filtering of BAM files and the collation of stats associated with BAM files have been consolidated into a new module which simplifies BAM filtering and stats collection. In this release, the module is run twice – once for the collection of stats and once to perform filtering of BAMs for downstream processing.
- The resource overrides in the super_seq profile have been reviewed and where they were unnecessarily excessive, they have been reduced.
- Some optimisations have been introduced to cause the cutadapt module to run quicker.

Support for Targeted Sequencing

- Previously, when the pipeline was run in targeted mode, the following metrics reported in the Excel Pipeline Report were inaccurate because they were calculated without actually removing the duplicate reads. This has now been corrected for the following metrics:
    - Genome mapped deduplicated reads
    - Initial input reads fraction aligning to the genome after deduplication
- 'TSS to non-TSS coverage ratio' has been disabled in targeted mode where the ratio of coverage at TSS (transcription start site) loci to coverage at non-TSS loci is likely to be confounded by the choice of target panel.

Command Line Interface (CLI) v1.0.4

Parameter updates

- Using orientation-reversing UDI-UMI fork-heads previously required the **r1_r2_switch** parameter. Now, when using orientation-reversing UDI-UMI fork-heads, it is no longer necessary to set the **r1_r2_switch** parameter; whether or not an orientation switch is required will be automatically detected and evoked based on the nucleotide content of the first 10,000 reads.

Cloud/HPC installation and running optimisation

- HPC queue setting recommendations, override per module
- Ability to test pipeline using GIAB data
- Removed old references when upgrading from 1.0.1 to 1.0.2
- Continuous Integration automation added for `auth` and `init` stages

- Added information on how to enable generation of the FASTQC report and bedMethyl files for duet evoC
- Tweaks and typo fixing
- New duet output folder structure described

Documentation updates

- Update hardware requirements for duet pipeline 1.2.0 modules
- Recommendation for nextflow queue settings
- Instructions for analysing GIAB data
- Added examples for analyse command
- Clarified location of configuration files

+44 (0)1223 800 700