biomodal software release notes April 2024

Contents

biomodal.com
info@biomodal.com

biomodal software release notes v1.0, April 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 v1.2.1

We have issued the following bug fixes and feature improvements in the current release:

Output Data Files

The following issues were resolved:

- **CURATE_QUANT process modifies upstream outputs on some executors**
    - The CURATE_QUANT process converts a one-indexed quantification tsv file generated by the EPIQUANT_QUANT process into a zero-indexed format for loading into a zarr store. It was identified that on some executors, this operation gets performed in place, so modifies the upstream output file in the EPIQUANT_QUANT process. If a pipeline is resumed, this can result in the indexing of the quantification tsv file getting shifted more than once. This can cause incorrect data in the quantification tsv file and in the zarr store. This bug only affected executors that directly access files in the scratch directory via hard links, rather than staging them locally, such as Google Batch, HPC executors (i.e. SLURM), and the local executor which runs the pipeline on a single node/machine. This bug is fixed in this release.
- **Targeted mode publishing error in EPIQUANT_QUANT**
    - An error in a file pattern in the publishing rules for epiquant_quant.nf meant that in targeted mode, no quantification files were published. This has now been corrected.

Support for Alternative Sequencing Platforms or Chemistry

The following issues were resolved:

- **Strategy for the resolution of quality scores with XLEAP chemistry**
    - The resolution of a pair of quality scores on R1 and R2 into a single quality score on a resolved read is currently performed via the use of empirical look-up tables that are specific to read length, instrument type, and assay type, but are not compatible with the latest quality-score binning strategy used by Illumina on-instrument control and RTA4 software for XLEAP flow cells. This issue can currently be circumvented using the following parameter override which will resolve quality scores by taking the minimum of the pair, rather than attempting to use empirical look-up tables:

      ```
      --additional-params couplet.phred=min
      ```

      This bug fix introduces the automatic detection of this issue and defaults to using the minimum quality score approach for resolving quality scores from XLEAP-SBS flow cells, avoiding the need to use the above parameter override in these circumstances.

Alternative Reference Genome Support

The following issues were resolved:

- **Silent BAM truncation error on large mouse samples**
  - A silent memory exhaustion error was identified when aligning deep sequence samples (> ~450M reads) to the GRCm38 (mouse) reference genome using the super_seq resource profile. Memory exhaustion resulted in truncated BAM files being passed downstream, resulting in reads lost at the alignment step and lower overall genome coverage.
    - The super_seq resource profile has been adjusted to improve the ratio of memory to CPUs when aligning to GRCm38 to avoid this memory exhaustion error.
    - Settings in the BWA_MEM2 module have been changed so that this class of failure in this module would be caught if it arose in future.
    - A minor optimisation has been introduced to the BWA_MEM2 module.
- **Infer data type for contig names in zarr stores to support any reference**
  - The process for generating zarr stores featured a hard-coded dependency that was only compatible with human and mouse reference genomes, resulting in an error generating zarr stores for other references. This is now fixed.
- **Error in COUPLET if a read category has no reads**
  - Previously, a division by zero error occurred in the COUPLET module if there were no reads at all in one of the following categories:
    - Acceptable reads (resolved with naïve alignment)
    - Rescued reads (resolved after pairwise alignment)
    - Discarded reads (do not resolve)

    This has been fixed.

The following new feature has been implemented:

- **Introduce genome config entry for *Bos taurus* and *Ovis aries***
  - This change introduces an entry in the genomes config file to support running the pipeline with *Bos taurus* (cow) and *Ovis aries* (sheep) reference genomes.

Other Bug Fixes

- **Error in ONE_STEP_BAMLET module if there are no long controls**
  - Previously, a division by zero error occurred in the ONE_STEP_BAMLET module if there were no reads mapping to the long controls (i.e. lambda or pUC19). This has been fixed.