Benchmarking NGS Aligners with empirically derived sequencing error profiles - Release 1.10.6

Overview

A new method is described whereby empirically derived sequenced read error profiles are used when simulating reads with a known ground truth instead of applying modelled error profiles to simulated reads. Modelled error profiles are reliant on inherent understanding of model parameterisations and these understandings generally do not account for the various factors which may influence the error profiles of individual reads other than general base calling error distributions and known sequence library preparation artefacts.

Method

Actual sequenced reads are aligned against a known targeted genome assembly or set of target sequences. Individual reads claimed as being aligned are processed against the claimed alignment loci and the base error profile distribution along the length of the read is obtained by iterating along the read length and recording the read offsets at which there are base mismatches. This is further extended such that if the aligned read contains any insertions or deletions then the offset and length of these InDels is also recorded. Additionally, the strand to which the read was aligned is recorded as part of the sequencing error profile. If paired end alignments are being processed, then the individual read profiles and insert size for a given aligned pair is used as if a single extended profile when generating simulated read pairs.

When generating the simulated reads to be used in benchmarking, read sequences are randomly sampled from the targeted genome assembly, and each read sequence will individually have a empirically derived sequencing error profile applied such that the simulated reads will have a known ground truth originating loci with an error profile consistent with those observed in the original sequenced read. This consistency extends to insertions and deletions, strand, and if paired end simulation then the insert size as well as the individual mate read error profiles.

The simulated reads with error profiles applied are then aligned back to the target by each aligner being benchmarked, and scored at both the read and individual base levels with Fbeta-measure weightings specified according to benchmarking objectives. Default weightings (beta=0.1) are applied for bases which are correctly aligned, misaligned, unaligned and for a special case in which reads have been silently trimmed but the base lies within the ground truth loci boundaries for that read. For both read and base levels, F-scoring is generated for two cases – against the background of all putative alignments allowing for inclusion of base weightings into the score, and for alignments only which excludes unaligned weightings.

Discussion

Utilising empirically derived sequencing error profiles bypasses the limitations of modelled error profiles but does have an inherent issue in that there is circularity – error profiles used in the simulated reads are derived from reads with errors which the aligner was able to originally align. So it would be expected that simulated reads would be aligned with very high scores by that aligner. To minimise the inherent circularity, it is intended that multiple aligners be used as controls with each aligners' empirically derived error profiles being used to generate a set of independent ground truth simulated reads. Each aligner then independently aligns set members with scores of that aligners set member alignment retained (matrix all vs. all) and scores later compared when determining which aligner is best suited for meeting experimental objectives.

Instead of multiple aligners, it could be the same aligner but with different parameterisation values. This is very useful when doing a parameterisation sweep when exploring parameterisation for those values which are likely to maximise achieving experimental objectives.

In addition to overall scoring the generated results, in CSV format, contain all raw base and read counts for the various alignment classification as well as histogram bins for number of correctly aligned bases in reads allowing choice of downstream characterisation methods for determining appropriate cut off thresholds to classify reads as being correctly aligned.

Implementation

A multiphase approach has been adopted and implemented in the ‘ngskit4b benchmark’ subprocess with the currently selected phase determined through a parameterisation value. ‘ngskit4b benchmark’ is aligner agnostic, providing that the aligner generates well formed SAM alignments and retains simulated read names as read identifiers in the SAM alignments then that aligner can be benchmarked.

Phase 0 is optional, it enables copying a limited number of raw reads from one readset into a second readset so as to reduce the number of reads requiring alignment in the subsequent mandatory phase 1 processing.

In phase 1 an existing SAM alignment by the aligner being benchmarked is processed and alignment details including SAM CIGAR is recorded. The SAM read sequence is then processed using the CIGAR and initial alignment loci by iterating over each base, obeying CIGAR operators, and recording each base at which there is a mismatch when that base is aligned against the target genome assembly or target sequences. For each read the original alignment CIGAR and an extended CIGAR containing exact ‘+’ and mismatch ‘S’ SAM operators plus alignment sense, and mate CIGARs with insert size if paired end alignments, is written to a CSV formatted file. The contents of this CSV file constitute the empirically derived sequencing error profiles for the combination of aligner and readset which were originally aligned.

In phase 2, a parameter specified number of read sequence start loci are randomly sampled without replacement from a targeted genome or sequences. Each sampled start loci individually has a empirically derived error profile applied such that the fully sampled read will have the same alignment sense, length, mismatch errors at the same offset, insertions and deletions of original size, as were present in the original alignment but now the ground truth of originating loci is known.
Additionally, if paired end simulation then the pairs of reads will have the original insert size with mate read error profiles identical to those in the original alignments.

The simulated reads are written to file in Fasta format with ground truth and error profile CIGAR contained in the Fasta descriptor with sufficient detail to enable individual base ground truth alignment validation when the simulated read has been subsequently aligned provided the Fasta read name at the start of descriptor identifying this read is retained in the resultant SAM alignment.

In phase 3, following alignment by the aligner being benchmarked, SAM alignments of the simulated reads are processed with the SAM read identifiers used as a key into the simulated reads file. The ground truth and simulated CIGARs extracted from the simulated reads file are then compared with the SAM alignment loci and CIGAR enabling characterisation as to if individual bases have been aligned to their target ground truth originating loci.

When scoring the benchmarked aligners, F-measures (beta = 1.0 and default beta = 0.1) are generated for both base and read level counts.

Scripting

Example python scripts are provided which provide a framework for managing a complete benchmark processing workflow starting from the original sequenced read alignments through to generation of the benchmarking results. These scripts, for single end and paired end bench marking, only require minimal editing to meet specific benchmarking objectives.