Systematic evaluation of spliced alignment programs for RNA-seq data

Pär G Engström[1,13], Tamara Steijger[1], Botond Sipos[1], Gregory R Grant[2,3], André Kahles[4,5], The RGASP Consortium[6], Gunnar Rätsch[4,5], Nick Goldman[1], Tim J Hubbard[7], Jennifer Harrow[7], Roderic Guigó[8,9] & Paul Bertone[1,10–12]

High-throughput RNA sequencing is an increasingly accessible method for studying gene structure and activity on a genome-wide scale. A critical step in RNA-seq data analysis is the alignment of partial transcript reads to a reference genome sequence. To assess the performance of current mapping software, we invited developers of RNA-seq aligners to process four large human and mouse RNA-seq data sets. In total, we compared 26 mapping protocols based on 11 programs and pipelines and found major performance differences between methods on numerous benchmarks, including alignment yield, basewise accuracy, mismatch and gap placement, exon junction discovery and suitability of alignments for transcript reconstruction. We observed concordant results on real and simulated RNA-seq data, confirming the relevance of the metrics employed. Future developments in RNA-seq alignment methods would benefit from improved placement of multimapped reads, balanced utilization of existing gene annotation and a reduced false discovery rate for splice junctions.

examines the density of independent reads at those loci. Many algorithms also consider base-call quality scores and use sophisticated indexing schemes to decrease runtime.

Here we assess the performance of 26 RNA-seq alignment protocols on real and simulated human and mouse transcriptomes. We adopted a competitive evaluation model applied in other areas of bioinformatics[11–14]. Developers were invited to run their software and submit results for evaluation as part of the RNA-seq Genome Annotation Assessment Project (RGASP). Programs included six spliced aligners GSNAP[7], MapSplice[4], PALMapper[8], ReadsMap, STAR[9] and TopHat[5,6]) and four alignment pipelines (GEM[3], PASS[15], GSTRUCT and BAGET). GSTRUCT is based on GSNAP, whereas BAGET uses a contiguous DNA aligner to map reads to the genome as well as to exon junction sequences derived from reference gene annotation. For comparison, the contiguous aligner SMALT was also tested. SMALT can map reads in a split manner, but it lacks several features of dedicated spliced aligners, such as precise determination of exon-intron boundaries. We demonstrate that choice of align-