rna_fq2bam

Table of contents

This tool is the equivalent of fq2bam for RNA-Seq samples, receiving inputs in FASTQ format, performing alignment with the splice-aware STAR algorithm, optionally marking of duplicate reads, and outputting an aligned BAM file ready for variant and fusion calling.

Quick Start

```
# This command assumes all the inputs are in INPUT_DIR and all the outputs go to OUTPUT_DIR. docker run --rm --gpus all --volume INPUT_DIR:/workdir --volume OUTPUT_DIR:/outputdir \ --workdir /workdir \ nvcr.io/nvidia/clara/clara-parabricks:4.3.1-1 \ pbrun rna_fq2bam \ --in-fq /workdir/${INPUT_FASTQ_1} /workdir/${INPUT_FASTQ_2} \ --genome-lib-dir /workdir/${PATH_TO_GENOME_LIBRARY}/ \ --output-dir /outputdir/${PATH_TO_OUTPUT_DIRECTORY} \ --ref /workdir/${REFERENCE_FILE} \ --out-bam /outputdir/${OUTPUT_BAM} \ --read-files-command zcat
```

Compatible CPU Command

The output from these commands will be identical to the output from the above command. See the Output Comparison page for comparing the results.

```
# STAR Alignment $ ./STAR \ --genomeDir <INPUT_DIR>/${PATH_TO_GENOME_LIBRARY} \ --readFilesIn <INPUT_DIR>/${INPUT_FASTQ_1} <INPUT_DIR>/${INPUT_FASTQ_2} \ --outFileNamePrefix <OUTPUT_DIR>/${PATH_TO_OUTPUT_DIRECTORY}/ \ --outSAMtype BAM SortedByCoordinate \ --readFilesCommand zcat # Mark Duplicates $ gatk MarkDuplicates \ --java-options -Xmx30g \ -I Aligned.sortedByCoord.out.bam \# This filename is determined by STAR. -O <OUTPUT_DIR>/${NAME_OF_OUTPUT_BAM_FILE} \ -M metrics.txt
```

Note

Make sure you have the same version of STAR installed that was used to build the genome index.

The Parabricks version of STAR is compatible with the 2.7.2a CPU-only version of STAR.

rna_fq2bam Reference

Run RNA-seq data through the fq2bam pipeline. It will run STAR aligner, co- ordinate sorting and mark duplicates.

Input/Output file options

--ref REF

Path to the reference file. (default: None)

Option is required.

--in-fq [IN_FQ ...]

Path to the pair-ended FASTQ files followed by optional read groups with quotes (Example: "@RG\tID:foo\tLB:lib1\tPL:bar\tSM:sample\tPU:foo"). The files must be in fastq or fastq.gz format. All sets of inputs should have a read group; otherwise, none should have a read group, and it will be automatically added by the pipeline. This option can be repeated multiple times. Example 1: --in-fq sampleX_1_1.fastq.gz sampleX_1_2.fastq.gz --in-fq sampleX_2_1.fastq.gz sampleX_2_2.fastq.gz. Example 2: --in-fq sampleX_1_1.fastq.gz sampleX_1_2.fastq.gz "@RG\tID:foo\tLB:lib1\tPL:bar\tSM:sample\tPU:unit1" --in-fq sampleX_2_1.fastq.gz sampleX_2_2.fastq.gz "@RG\tID:foo2\tLB:lib1\tPL:bar\tSM:sample\tPU:unit2". For the same sample, Read Groups should have the same sample name (SM) and a different ID and PU. (default: None)

--in-se-fq [IN_SE_FQ ...]

Path to the single-ended FASTQ file followed by optional read group with quotes (Example: "@RG\tID:foo\tLB:lib1\tPL:bar\tSM:sample\tPU:foo"). The file must be in fastq or fastq.gz format. Either all sets of inputs have a read group, or none should have one, and it will be automatically added by the pipeline. This option can be repeated multiple times. Example 1: --in-se-fq sampleX_1.fastq.gz --in-se-fq sampleX_2.fastq.gz . Example 2: --in-se-fq sampleX_1.fastq.gz "@RG\tID:foo\tLB:lib1\tPL:bar\tSM:sample\tPU:unit1" --in-se-

fq sampleX_2.fastq.gz "@RG\tID:foo2\tLB:lib1\tPL:bar\tSM:sample\tPU:unit2" . For the same sample, Read Groups should have the same sample name (SM) and a different ID and PU. (default: None)
--genome-lib-dir GENOME_LIB_DIR

Path to a genome resource library directory. The indexing required to run STAR should be completed by the user beforehand. (default: None)

Option is required.

--output-dir OUTPUT_DIR

Path to the directory that will contain all of the generated files. (default: None)

Option is required.

--out-bam OUT_BAM

Path of the output BAM file. (default: None)

Option is required.

Tool Options:

--out-prefix OUT_PREFIX

Prefix filename for output data. (default: None)

--read-files-command READ_FILES_COMMAND

Command line to execute for each of the input files. This command should generate FASTA or FASTQ text and send it to stdout: For example, zcat to uncompress .gz files, bzcat to uncompress .bz2 files, etc. (default: None)

--read-group-sm READ_GROUP_SM

SM tag for read groups in this run. (default: None)

--read-group-lb READ_GROUP_LB

LB tag for read groups in this run. (default: None)

--read-group-pl READ_GROUP_PL

PL tag for read groups in this run. (default: None)

--read-group-id-prefix READ_GROUP_ID_PREFIX

Prefix for the ID and PU tags for read groups in this run. This prefix will be used for all pairs of FASTQ files in this run. The ID and PU tags will consist of this prefix and an identifier that will be unique for a pair of FASTQ files. (default: None)

--num-sa-bases NUM_SA_BASES

Length (bases) of the SA pre-indexing string. Longer strings will use more memory, but allow for faster searches. A value between 10 and 15 is recommended. For small genomes, the parameter must be scaled down to min(14, log2(GenomeLength)/2 - 1). (default: 14)

--max-intron-size MAX_INTRON_SIZE

Maximum align intron size. If this value is 0, the maximum size will be determined by (2^winBinNbits)*winAnchorDistNbins. (default: 0)

--min-intron-size MIN_INTRON_SIZE

Minimum align intron size. Genomic gap is considered intron if its length is greater than or equal to this value, otherwise it is considered Deletion. (default: 21)

--min-match-filter MIN_MATCH_FILTER

Minimum number of matched bases required for alignment output. (default: 0)

--min-match-filter-normalized MIN_MATCH_FILTER_NORMALIZED

Same as --min-match-filter, but normalized to the read length (sum of the mate lengths for paired-end reads). (default: 0.66)

--out-filter-intron-motifs OUT_FILTER_INTRON_MOTIFS

Type of filter alignment using its motifs. This string can be "None" for no filtering, "RemoveNoncanonical" for filtering out alignments that contain non-canonical junctions, or "RemoveNoncanonicalUnannotated" for filtering out alignments that contain non-

canonical unannotated junctions when using the annotated splice junctions database. The annotated non-canonical junctions will be kept. (default: None)
--max-out-filter-mismatch MAX_OUT_FILTER_MISMATCH

Maximum number of mismatches allowed for an alignment to be output. (default: 10)

--max-out-filter-mismatch-ratio MAX_OUT_FILTER_MISMATCH_RATIO

Maximum ratio of mismatches to mapped length allowed for an alignment to be output. (default: 0.3)

--max-out-filter-multimap MAX_OUT_FILTER_MULTIMAP

Maximum number of loci the read is allowed to map to for all alignments to be output. Otherwise, no alignments will be output and the read will be counted as "mapped to too many loci" in the Log.final.out. (default: 10)

--out-reads-unmapped OUT_READS_UNMAPPED

Type of output of unmapped and partially mapped (i.e. mapped only one mate of a paired-end read) reads in separate file(s). This string can be "None" for no output or "Fastx" for output in separate FASTA/FASTQ files, Unmapped.out.mate1/2. (default: None)

--out-sam-unmapped OUT_SAM_UNMAPPED

Type of output of unmapped reads in SAM format. The string can be "None" to produce no output, "Within" to output unmapped reads within the main SAM file, "KeepPairs" to produce no output (with unmapped mates will be recorded for each alignment), or "Within_KeepPairs" to output unmapped reads within the main SAM file (with unmapped mates recorded for each alignment). (default: None)

--out-sam-attributes OUT_SAM_ATTRIBUTES [OUT_SAM_ATTRIBUTES ...]

A string of SAM attributes in the order desired for the output SAM. The string can contain any combination of the following attributes: {NH, HI, AS, nM, NM, MD, jM, jI, XS, MC, ch}. Alternatively, the string can be "None" for no attributes, "Standard" for the attributes {NH, HI, AS, nM}, or "All" for the attributes {NH, HI, AS, nM, NM, MD, jM, jI, MC, ch} (e.g. "--outSAMattributes NH nM jI XS ch"). (default: Standard)

--out-sam-strand-field OUT_SAM_STRAND_FIELD

Cufflinks-like strand field flag. The string can be "None" for no flag or "intronMotif" for the strand derived from the intron motif. Reads with inconsistent and/or non-canonical introns will be filtered out. (default: None)

--out-sam-mode OUT_SAM_MODE

SAM output mode. The string can be "None" for no SAM output, "Full" for full SAM output, or "NoQS" for full SAM output without quality scores. (default: Full)

--out-sam-mapq-unique OUT_SAM_MAPQ_UNIQUE

The MAPQ value for unique mappers. Must be in the range [0, 255]. (default: 255)

--min-score-filter MIN_SCORE_FILTER

Minimum score required for alignment output, normalized to the read length (i.e. the sum of mate lengths for paired-end reads). (default: 0.66)

--min-spliced-mate-length MIN_SPLICED_MATE_LENGTH

Minimum mapped length for a read mate that is spliced and normalized to the mate length. Must be greater than 0. (default: 0.66)

--max-junction-mismatches MAX_JUNCTION_MISMATCHES MAX_JUNCTION_MISMATCHES MAX_JUNCTION_MISMATCHES MAX_JUNCTION_MISMATCHES

Maximum number of mismatches for stitching of the splice junctions. A limit must be specified for each of the following: (1) non-canonical motifs, (2) GT/AG and CT/AC motif, (3) GC/AG and CT/GC motif, (4) AT/AC and GT/AT motif. To indicate no limit for any of the four options, use -1. (default: [0, -1, 0, 0])

--max-out-read-size MAX_OUT_READ_SIZE

Maximum size of the SAM record (bytes) for one read. Recommended value: >(2*(LengthMate1+LengthMate2+100)*o utFilterMultimapNmax. Must be greater than 0. (default: 100000)

--max-alignments-per-read MAX_ALIGNMENTS_PER_READ

Maximum number of different alignments per read to consider. Must be greater than 0. (default: 10000)

--score-gap SCORE_GAP

Splice junction penalty (independent of intron motif). (default: 0)

--seed-search-start SEED_SEARCH_START

Defines the search start point through the read. The read split pieces will not be longer than this value. Must be greater than 0. (default: 50)

--max-bam-sort-memory MAX_BAM_SORT_MEMORY

Maximum available RAM (bytes) for sorting BAM. If this value is 0, it will be set to the genome index size. Must be greater than or equal to 0. (default: 0)

--align-ends-type ALIGN_ENDS_TYPE

Type of read ends alignment. Can be one of two options: "Local" will perform a standard local alignment with soft-clipping allowed; "EndToEnd" will force an end-to-end read alignment with no soft-clipping. (default: Local)

--align-insertion-flush ALIGN_INSERTION_FLUSH

Flush ambiguous insertion positions. The string can be "None" to not flush insertions or "Right" to flush insertions to the right. (default: None)

--max-align-mates-gap MAX_ALIGN_MATES_GAP

Maximum gap between two mates. If 0, the max intron gap will be determined by (2^winBinNbits)*winAnchorDistNbins. (default: 0)

--min-align-spliced-mate-map MIN_ALIGN_SPLICED_MATE_MAP

Minimum mapped length for a read mate that is spliced. Must be greater than or equal to 0. (default: 0)

--max-collapsed-junctions MAX_COLLAPSED_JUNCTIONS

Maximum number of collapsed junctions. Must be greater than 0. (default: 1000000)

--min-align-sj-overhang MIN_ALIGN_SJ_OVERHANG

Minimum overhang (i.e. block size) for spliced alignments. Must be greater than 0. (default: 5)

--min-align-sjdb-overhang MIN_ALIGN_SJDB_OVERHANG

Minimum overhang (i.e. block size) for annotated (sjdb) spliced alignments. Must be greater than 0. (default: 3)

--sjdb-overhang SJDB_OVERHANG

Length of the donor/acceptor sequence on each side of the junctions. Ideally, this value should be equal to mate_length - 1. Must be greater than 0. (default: 100)

--min-chim-overhang MIN_CHIM_OVERHANG

Minimum overhang for the Chimeric.out.junction file. Must be greater than or equal to 0. (default: 20)

--min-chim-segment MIN_CHIM_SEGMENT

Minimum chimeric segment length. If it is set to 0, there will be no chimeric output. Must be greater than or equal to 0. (default: 0)

--max-chim-multimap MAX_CHIM_MULTIMAP

Maximum number of chimeric multi-alignments. If it is set to 0, the old scheme for chimeric detection, which only considered unique alignments, will be used. Must be greater than or equal to 0. (default: 0)

--chim-multimap-score-range CHIM_MULTIMAP_SCORE_RANGE

The score range for multi-mapping chimeras below the best chimeric score. This option only works with --max-chim-multimap > 1. Must be greater than or equal to 0. (default: 1)

--chim-score-non-gtag CHIM_SCORE_NON_GTAG

The penalty for a non-GT/AG chimeric junction. (default: -1)

--min-non-chim-score-drop MIN_NON_CHIM_SCORE_DROP

To trigger chimeric detection, the drop in the best non-chimeric alignment score with respect to the read length has to be smaller than this value. Must be greater than or

equal to 0. (default: 20)
--out-chim-format OUT_CHIM_FORMAT

Formatting type for the Chimeric.out.junction file. Possible types are {0, 1}. If type 0, there will be no comment lines/headers. If type 1, there will be comment lines at the end of the file: command line and Nreads: total, unique, multi. (default: 0)

--two-pass-mode TWO_PASS_MODE

Two-pass mapping mode. The string can be "None" for one-pass mapping or "Basic" for basic two-pass mapping, with all first pass junctions inserted into the genome indices on the fly. (default: None)

--out-chim-type OUT_CHIM_TYPE

Type of chimeric output. This string can be "Junctions" for Chimeric.out.junction, "WithinBAM" for main aligned BAM files (Aligned.*.bam), "WithinBAM_HardClip" for hard-clipping in the CIGAR for supplemental chimeric alignments, or "WithinBAM_SoftClip" for soft-clipping in the CIGAR for supplemental chimeric alignments. (default: None)

--no-markdups

Do not perform the Mark Duplicates step. Return BAM after sorting. (default: None)

--read-name-separator READ_NAME_SEPARATOR [READ_NAME_SEPARATOR ...]

Character(s) separating the part of the read names that will be trimmed in output (read name after space is always trimmed). (default: /)

Performance Options:

--num-threads NUM_THREADS

Number of running worker threads per GPU. (default: 4)

Common options:

--logfile LOGFILE

Path to the log file. If not specified, messages will only be written to the standard error output. (default: None)

--tmp-dir TMP_DIR

Full path to the directory where temporary files will be stored.

--with-petagene-dir WITH_PETAGENE_DIR

Full path to the PetaGene installation directory. By default, this should have been installed at /opt/petagene. Use of this option also requires that the PetaLink library has been preloaded by setting the LD_PRELOAD environment variable. Optionally set the PETASUITE_REFPATH and PGCLOUD_CREDPATH environment variables that are used for data and credentials (default: None)

--keep-tmp

Do not delete the directory storing temporary files after completion.

--no-seccomp-override

Do not override seccomp options for docker (default: None).

--version

View compatible software versions.

GPU options:

--num-gpus NUM_GPUS

Number of GPUs to use for a run. GPUs 0..(NUM_GPUS-1) will be used.

> ⓘ **Note**
>
> The *--in-fq* option takes the names of two FASTQ files, optionally followed by a quoted read group. The FASTQ filenames must not start

with a hyphen.