minimap2 (Beta)

Table of contents

Run a GPU-accelerated minimap2.

This tool aligns long read sequences against a large reference database using an accelerated KSW2 to convert FASTQ to BAM/CRAM.

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 minimap2 \ --ref /workdir/${REFERENCE_FILE} \ --in-fq /workdir/${INPUT_FASTQ} \ --out-bam /outputdir/${OUTPUT_BAM}
```

Compatible CPU-based minimap2, GATK4 Commands

The commands below are the minimap2-v2.26 and GATK4 counterpart of the Clara Parabricks command above. The output from these commands will be identical to the output from the above command. See the Output Comparison page for comparing the results. You may need to increase the Java heap size based on your dataset, or decrease the number of --MAX_RECORDS_IN_RAM.

```
# Run minimap2 and pipe the output to create a sorted BAM. $ minimap2 -ax map-pbmm2 \ <INPUT_DIR>/${REFERENCE_FILE} \ <INPUT_DIR>/${INPUT_FASTQ} | \ gatk SortSam \ --java-options -Xmx30g \ --MAX_RECORDS_IN_RAM 5000000 \ -I /dev/stdin \ -O cpu.bam \ --SORT_ORDER coordinate
```

Please note that two changes must be made to the baseline minimap2 code in order to match the results exactly:

Firstly, a new preset must be made in `options.c` in the `mm_set_opt` function that tries to replicate the preset of pbmm2 by setting these parameters as a new preset named "map-pbmm2":

```
io->k = 19; io->w = 10; io->batch_size = 0x7fffffffffffffffL; // always build a uni-part index mo->flag |= MM_I_HPC; mo->flag |= MM_F_CIGAR; mo->flag |= MM_F_LONG_CIGAR; mo->flag |= MM_F_EQX; mo->flag |= MM_F_SOFTCLIP; mo->flag |= MM_F_NO_PRINT_2ND; mo->flag |= MM_F_HARD_MLEVEL; mo->mask_level = 0; mo->e2 = 1; mo->zdrop = 400; mo->a = 2; mo->b = 5; mo->q = 5; mo->q2 = 56; mo->e = 4; mo->zdrop_inv = 50; mo->bw = 2000;
```

Secondly, a fix must be made to the baseline KSW2 code to round the loop fission start and end points by changing them to `st` and `en` respectively. If the start point (`st0`) is a number below 16, but greater than 0, its scoring values will not be initialized correctly, but will still be used later when computing the actual alignment. This can be fixed by rounding the start and end points to multiples of 16.

To make this fix, change the following code in `ksw2_extd2_sse.c`:

```
// loop fission: set scores first if (!(flag & KSW_EZ_GENERIC_SC)) { for (t = st0; t <= en0; t += 16) { __m128i sq, st, tmp, mask; sq = _mm_loadu_si128((__m128i*)&sf[t]); st = _mm_loadu_si128((__m128i*)&qrr[t]); mask = _mm_or_si128(_mm_cmpeq_epi8(sq, m1_), _mm_cmpeq_epi8(st, m1_)); tmp = _mm_cmpeq_epi8(sq, st); #ifdef __SSE4_1__ tmp = _mm_blendv_epi8(sc_mis_, sc_mch_, tmp); tmp = _mm_blendv_epi8(tmp, sc_N_, mask); #else tmp = _mm_or_si128(_mm_andnot_si128(tmp, sc_mis_), _mm_and_si128(tmp, sc_mch_)); tmp = _mm_or_si128(_mm_andnot_si128(mask,
```

```
tmp), _mm_and_si128(mask, sc_N_)); #endif _mm_storeu_si128((__m128i*)((int8_t*)s + t), tmp); } } else { for (t = st0; t <= en0; ++t) ((uint8_t*)s)[t] = mat[sf[t] * m + qrr[t]]; }
```

Fixed version that uses `lf_start` and `lf_en`:

```
// loop fission: set scores first int lf_start = st, lf_en = en; if (!(flag & KSW_EZ_GENERIC_SC)) { for (t = lf_start; t <= lf_en; t += 16) { __m128i sq, st, tmp, mask; sq = _mm_loadu_si128((__m128i*)&sf[t]); st = _mm_loadu_si128((__m128i*)&qrr[t]); mask = _mm_or_si128(_mm_cmpeq_epi8(sq, m1_), _mm_cmpeq_epi8(st, m1_)); tmp = _mm_cmpeq_epi8(sq, st); #ifdef __SSE4_1__ tmp = _mm_blendv_epi8(sc_mis_, sc_mch_, tmp); tmp = _mm_blendv_epi8(tmp, sc_N_, mask); #else tmp = _mm_or_si128(_mm_andnot_si128(tmp, sc_mis_), _mm_and_si128(tmp, sc_mch_)); tmp = _mm_or_si128(_mm_andnot_si128(mask, tmp), _mm_and_si128(mask, sc_N_)); #endif _mm_storeu_si128((__m128i*)((int8_t*)s + t), tmp); } } else { for (t = lf_start; t <= lf_en; ++t) ((uint8_t*)s)[t] = mat[sf[t] * m + qrr[t]]; }
```

minimap2 Reference

Align long read sequences against a large reference database to convert FASTQ to BAM/CRAM.

Input/Output file options

--ref REF

Path to the reference file. (default: None)

Option is required.

--index INDEX

Path to a minimizer index file generated by vanilla minimap2 to reduce indexing time. (default: None)

--in-fq IN_FQ

Path to a query sequence file in fastq or fastq.gz format. (default: None)

Option is required.

--knownSites KNOWNSITES

Path to a known indels file. The file must be in vcf.gz format. This option can be used multiple times. (default: None)

--interval-file INTERVAL_FILE

Path to an interval file in one of these formats: Picard-style (.interval_list or .picard), GATK-style (.list or .intervals), or BED file (.bed). This option can be used multiple times. (default: None)

--out-recal-file OUT_RECAL_FILE

Path of a report file after Base Quality Score Recalibration. (default: None)

--out-bam OUT_BAM

Path of a BAM/CRAM file after sorting. (default: None)

Option is required.

--out-duplicate-metrics OUT_DUPLICATE_METRICS

Path of duplicate metrics file after Marking Duplicates. (default: None)

--out-qc-metrics-dir OUT_QC_METRICS_DIR

Path of the directory where QC metrics will be generated. (default: None)

Tool Options:

--preset PRESET

Which preset to apply. Possible values are {map-pbmm2,map-hifi,map-ont}. (default: map-pbmm2)

--eqx

Write =/X CIGAR operators. (default: None)

-L INTERVAL, --interval INTERVAL

Interval within which to call bqsr from the input reads. All intervals will have a padding of 100 to get read records, and overlapping intervals will be combined. Interval files should be passed using the --interval-file option. This option can be used multiple times (e.g. "-L chr1 -L chr2:10000 -L chr3:20000+ -L chr4:10000-20000"). (default: None)

-ip INTERVAL_PADDING, --interval-padding INTERVAL_PADDING

Amount of padding (in base pairs) to add to each interval you are including. (default: None)

--standalone-bqsr

Run standalone BQSR after generating sorted BAM. This option requires both --knownSites and --out-recal-file input parameters. (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)

Performance Options:

--num-threads NUM_THREADS

Number of processing threads. (default: 128)

--gpuwrite

Use one GPU to accelerate writing final BAM. (default: None)

--gpuwrite-deflate-algo GPUWRITE_DEFLATE_ALGO

Choose the nvCOMP DEFLATE algorithm to use with --gpuwrite. Note these options do not correspond to CPU DEFLATE options. Valid options are 0 and 3. Option 0 is faster while option 3 provides a better compression ratio. (default=0) (default: None)

--gpusort

Use GPUs to accelerate sorting. (default: None)

--use-gds

Use GPUDirect Storage (GDS) to enable a direct data path for direct memory access (DMA) transfers between GPU memory and storage. Must be used concurrently with *--gpuwrite*. Please refer to Parabricks Documentation > Best Performance for information on how to set up and use GPUDirect Storage. (default: None)

--low-memory

Use low memory mode (default: None)

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.