Molzym GmbH & Co. KG — Mary-Astell-Str. 10 — 28359 Bremen — Germany

Molzym GmbH & Co. KG
Mary-Astell-Str. 10
28359 Bremen
Germany

Tel.: +49 (0) 421 69 61 62 0
E-Mail: info@molzym.com
Internet: www.molzym.com

Important Note for NGS analyses in connection with Molzym's IEC included in SepsiTest™-UMD, UMD-SelectNA™, Micro-Dx™ and MolYsis-SN*plus*™ IVD

An internal extraction control (IEC)—required in some countries for extraction kits used in clinical settings—is implemented in SepsiTest™-UMD (U-010-0xy), UMD-SelectNA™ (U-050-0xy), Micro-Dx™ (U-200-0xy), and MolYsis-SN*plus*™ IVD (U-300-0xy) kits. This allows to verify the DNA extraction process by co-extracting the "Control DNA" together with the microbial DNA from the sample, followed by detection in a separate assay (MA Control). In procedures involving a complete DNA identification, such as shotgun metagenomic analysis or whole genome sequencing (WGS) using next-generation sequencing (NGS), the presence of the IEC in the eluate may need to be taken into consideration.
The co-extracted "Control DNA", in the following called IEC, consists of a ~500 bp fragment derived from lambda phage DNA (sequence data shown in Figure 1). This IEC is typically present in the eluate at a concentration of approximately 0.2–10 fg/µl.
The presence of the IEC sequence can be utilized for validation and quantification purposes in NGS workflows. However, special attention is required if the reference database (DB) used for analysis contains prophage sequences, such as those from *E. coli*. In such cases, assignment of the IEC sequence to bacterial taxa may occur.
To avoid misclassification, please use the hereafter provided IEC sequence (Figure 1) within your BioIT analyses:

```
 aagcagacg acatctggaa tctgcgcaag gatgattatt ttgttaacga tgaagcgcgg gcgcgttact gggatgatcg
tgaaaaggcc cgtcttgcgc ttgaagccgc ccgaaagaag gctgagcagc agactcaaca ggacaaaaat gcgcagcagc
agagcgatac cgaagcgtca cggctgaaat ataccgaaga ggcgcagaag gcttacgaac ggctgcagac gccgctggag
aaatataccg cccgtcagga agaactgaac aaggcactga aagacgggaa aatcctgcag gcggattaca acacgctgat
ggcggcggcg aaaaaggatt atgaagcgac gctgaaaaag ccgaaacagt ccagcgtgaa ggtgtctgcg ggcgatcgtc
aggaagacag tgctcatgct gccctgctga cgcttcaggc agaactccgg acgctggaga agcatgccgg agcaaatgag
```

Figure 1: DNA sequence data of Molzym's IEC reference: fragment of lambda phage DNA

For Oxford Nanopore Sequencing (ONT), one of the following two approaches can be carried out to bioinformatically remove the IEC matching reads in the analyses

Required Open-Source software tools:

- Mapper (for short read data, e.g. Alignment-Tool **BWA**, or for long read data, e.g. **minimap2**)
- **bedtools**
- **Samtools**
- **SeqKit**
- **Seqtk** (to prepare the IEC reference if necessary)

Approach 1:

Within this approach the FASTQ-file (Input) is mapped against the IEC sequence to identify matching reads within the FASTQ-file (Input). The matching reads are listed in a text file and a modified FASTQ-file (Input without IEC reads) is created without the IEC matching reads. Both files, the original FASTQ-file (Input all reads) and the modified FASTQ-file (Input without IEC reads) can be used for further analysis (see Figure 2).

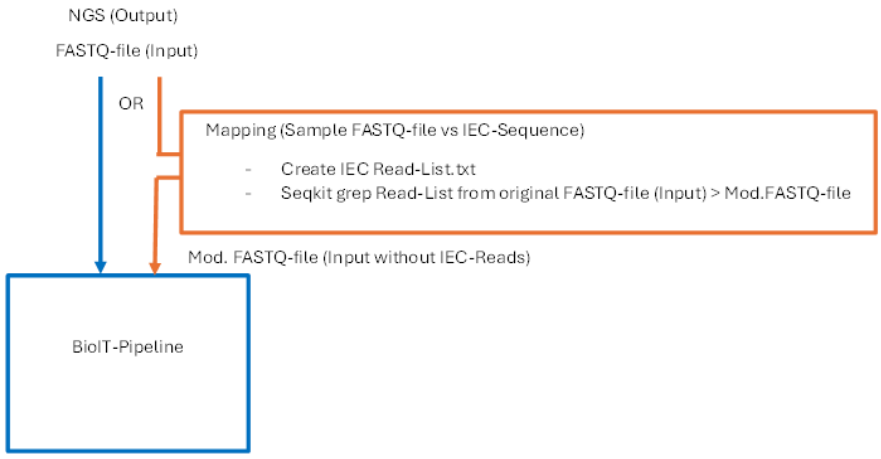

Figure 2: Flow diagram of data analysis with optional, prior modified run data (FASTQ-Input, orange)

Script:

- Prepare the IEC sequence (Figure 3)

```
 aagcagacg acatctggaa tctgcgcaag gatgattatt ttgttaacga tgaagcgcgg gcgcgttact gggatgatcg
tgaaaaggcc cgtcttgcgc ttgaagccgc ccgaaagaag gctgagcagc agactcaaca ggacaaaaat gcgcagcagc
agagcgatac cgaagcgtca cggctgaaat ataccgaaga ggcgcagaag gcttacgaac ggctgcagac gccgctggag
aaatataccg cccgtcagga agaactgaac aaggcactga aagacgggaa aatcctgcag gcggattaca acacgctgat
ggcggcggcg aaaaaggatt atgaagcgac gctgaaaaag ccgaaacagt ccagcgtgaa ggtgtctgcg ggcgatcgtc
aggaagacag tgctcatgct gccctgctga cgcttcaggc agaactccgg acgctggaga agcatgccgg agcaaatgag
```

Figure 3: DNA sequence data of Molzym's IEC reference: fragment of lambda phage DNA

- Using **minimap2**, create alignment of Input reads (FASTQ-file input) and IEC reference (IEC-Ref.fasta).

```
bash
~$ minimap2 -ax map-ont IEC-Ref.fasta FASTQ-file.fastq | samtools view -bS -o IEC-
Read.bam
```

- Aligned IEC reads are located within the IEC-Read.bam.
- Using **Samtools** separate aligned and unaligned reads and create an FASTQ-file only including IEC-Reads.
- Create a list (IEC-aligned.txt) of read headers for all IEC aligned reads.

```
~$ samtools view -F 4 IEC-Read.bam | cut -f1 | sort | uniq > IEC-aligned.txt
```

- Remove all IEC-Read from the original FASTQ-file (Input) using Seqkit and create a new modified FASTQ-file without IEC reads.

```
~$ seqkit grep -v -f IEC-aligned.txt FASTQ-file.fastq > IEC-free_Input.fastq
```

- The modified FASTQ-file (IEC-free_Input.fastq) is ready for use in your BioIT Pipeline.

Approach 2:

In an initial step, required only once per reference DB version, the IEC is aligned against the microbial reference DB to identify matching regions. Those regions are listed within a BED-file for every reference genome with start and stop position. After the first alignment within your BioIT pipeline, the created BAM-file can be filtered using ***bedtools*** and the BED-file to remove all IEC region matching read. Overlapping reads are spared and remain within the modified BAM-file (see Figure 4).

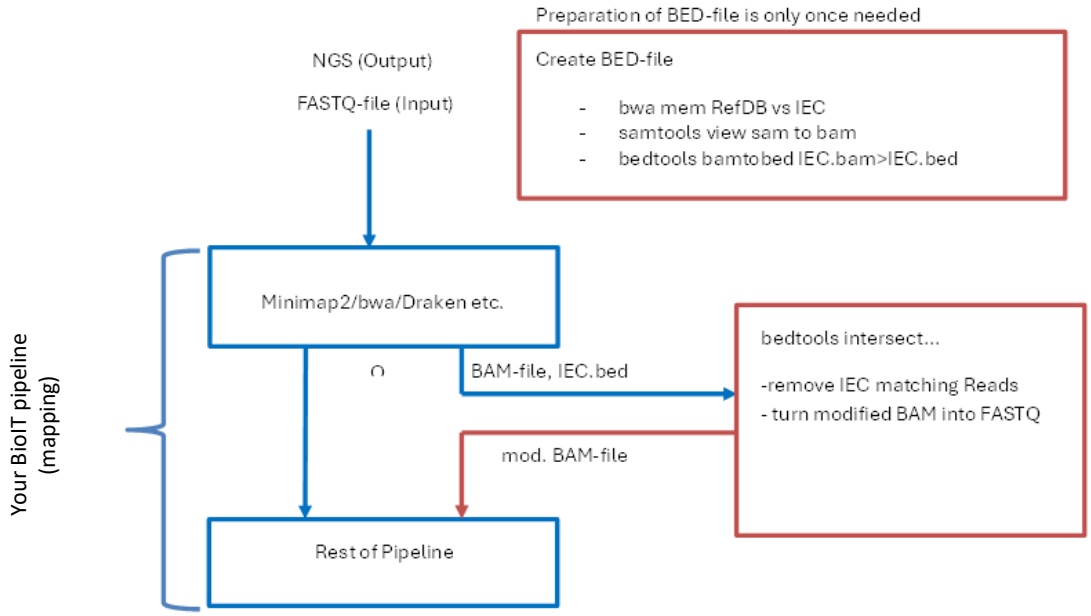

Figure 4: Flow diagram of modifying data (BAM-file) following the mapping step

Script:

- Once per reference DB version used in BioIT pipeline, identify all regions matching the IEC sequence by comparing the IEC sequence with the microbial reference DB. The names of the reference genomes and the start and stop position of the matching regions are listed in a BED-file (IEC-bed). This BED-file can be used for all further analyses. If it is necessary to add further controls, the BED-file can be adapted.

```
bash
~$ bwa mem microbialRefDB.fasta IEC.fasta > IEC.sam
~$ samtools view -Sb IEC.sam > IEC.bam
~$ bedtools bamtobed -i S1_IEC.bam > IEC.bed
~$ cat IEC.bed
eg. Output: IEC.bad
CP040643.1926310 926788
CP040643.1927310 927788
```

- Once the Sample_BAM-file (output alignment) has been generated, all reads matching the IEC sequence can be removed with the BED-file.

```
~$ bedtools intersect -v -a Sample.bam -b IEC.bed -f 1.0 > modified_Sample.bam
```

- If a FASTQ file is required again for further examinations, the modified BAM-file can be converted back into a FASTQ file using ***Samtools*** or ***bedtools***.

Please contact us at support@molzym.com if you have any questions.