Umi-pipeline-nf: Accurate consensus sequence creation for UMI-tagged nanopore data

Stephan Amstler[1], Lukas Forer[1], Sebastian Schönherr[1], Stefan Coassin[1]
[1] Institute of Genetic Epidemiology, Medical University of Innsbruck, Innsbruck, Austria

Motivation

Long-range **nanopore sequencing allows single-molecule sequencing**, but still faces challenges due to its high error rate (≈Q20).

Umi-pipeline-nf creates highly accurate single-molecule consensus sequences for unique molecular identifier (UMI)-tagged amplicons from nanopore sequencing data. **This reduces error rates by over 100-fold and achieves consensus sequence quality >Q40.**

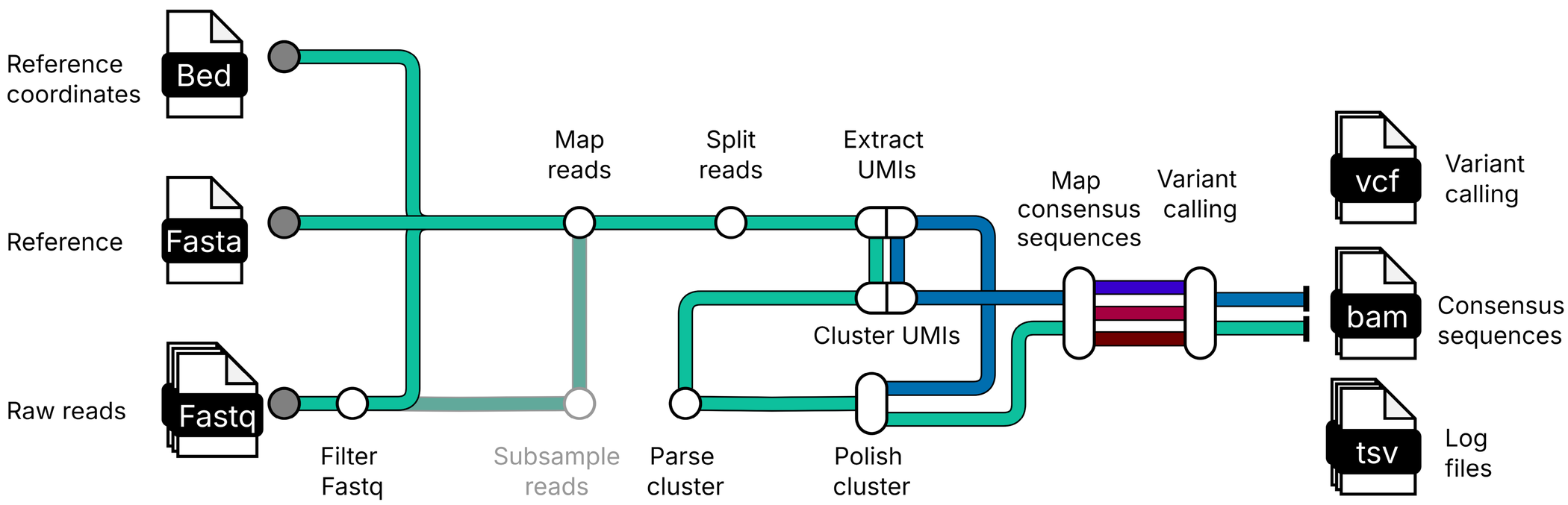
Of note, umi-pipeline-nf originates from a Snakemake-based pipeline (nanoporetech/pipeline-umi-amplicon; original workflow developed by Karst et al. *Nat Methods* 18, 2021). We migrated the pipeline to Nextflow and included several optimizations and additional functionalities.

Use cases

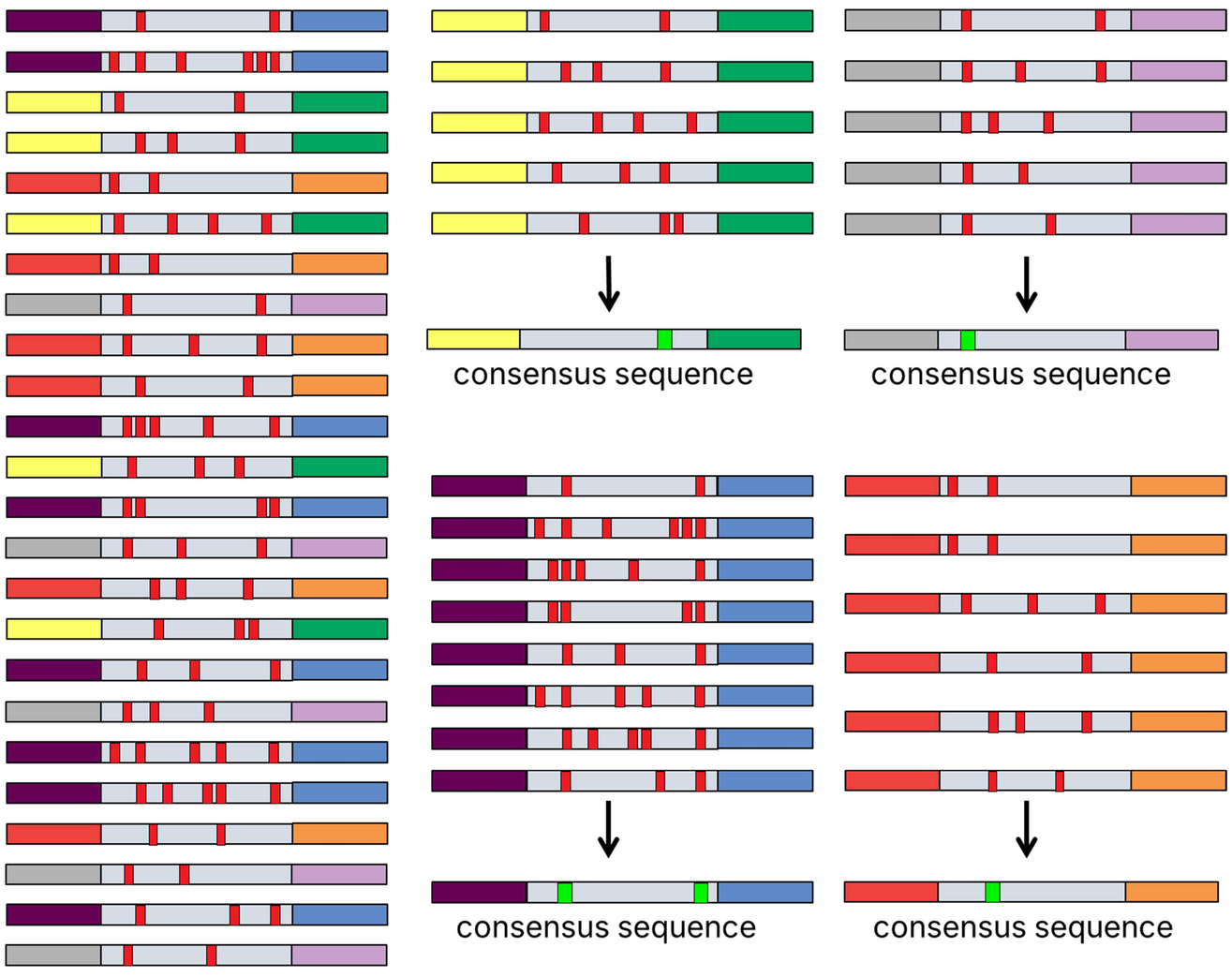
Umi-pipeline-nf is particularly useful in applications requiring **virtually error-free sequencing with clonal, respectively single-molecule resolution**, such as sequencing **repetitive genome regions**, studying **intra-host viral evolution**, investigating **cancer clonal evolution**, or determining detailed **metagenomics profiles.** This allows using an amplicon-based approach to generate reference data for complex regions with clonal resolution at scale.

We recently used umi-pipeline-nf to generate highly accurate, full-length haplotypes with single repeat resolution of a long, complex, and highly polymorphic human repeat element, the *LPA* KIV-2 VNTR (Amstler et al. *Genome Med* 16, 2024).

Umi-pipeline-nf



Principle of UMI nanopore sequencing (UMI-ONT-Seq)



Workflow

- **Filter fastq** - Quality control of input Fastq files
- **Subsample reads** - Optional subsampling of the reads
- **Map reads and Split reads** - Filtering for full-length reads by alignment against a reference sequence
- **Extract UMIs** - Extraction of terminal UMIs
- **Cluster UMIs** - Clustering of extracted UMIs
- **Parse Cluster** - Parsing and filtering of UMI clusters
- **Polish cluster** - Producing highly accurate consensus sequences for each UMI cluster with Medaka
- **Variant calling** - Optional low-frequency variant calling of the obtained consensus sequences

Key features

- **Detailed quality control of the UMI clusters to remove potentially admixed clusters**
- **Support of multiple variant callers to provide flexibility in data analysis**
- **Docker and Singularity containers to ensure reproducibility and portability to HPC clusters**
- **UMI cluster filtering preserves the highest quality reads**
- **Optional GPU accelerated cluster polishing (Beta feature)**
- **Real-time monitoring of the number of clusters per sample during sequencing (Beta feature)**

Funding: This research was funded in whole or in part by the Austrian Science Fund (FWF) [DOI: 10.55776/PAT5152823].

https://github.com/genepi/umi-pipeline-nf