proovread manual

Thomas Hackl

Contents

1 Installation

```
git clone --recursive https://github.com/BioInf-Wuerzburg/proovread
cd proovread/util/bwa
make
```

2 Dependencies

- Log::Log4perl
- blastn (NCBI Blast-2.2.24+ or later)

proovread is distributed ready with binaries of SHRiMP2 and blasr, as well as bwa source code. If you want to employ your own installed version of these mappers, have a look at the Advanced Configuration section.

3 Usage

Test your installation by running proovread on the included sample data set.

```
proovread --sample --pre /tmp/pr-sample
```

Don't run proovread on entire SMRT cells directly, it will only blast your memory and take forever. Split your data in handy chunks of a few Mbp first:

```
# located in /path/to/proovread/bin
SeqChunker -s 20M -o pb-%03d.fq pb-subreads.fq
```

Run proovread on one chunk first.

```
proovread -l pb-001.fq -s reads.fq [-u unitigs.fa] --pre pb-001
```

If things go smoothly, submit the rest.

```
xargs
qsub
sbatch
```

4 Output

By default, proovread generates six files in the output folder:

.trimmed.f[aq]	high accuracy pacbio reads, trimmed for uncorrected/low quality regions
.untrimmed.fq	complete corrected pacbio reads including un-/ poorly corrected regions
.ignored.tsv	ids of reads and the reason for excluding them from correction
.chim.tsv	annotations of potential chimeric joints clipped during trimming
.parameter.log	the parameter set used for this run

If you are interested in mappings (SAM) and other intermediary files from iterations have a look at `-keep-temporary`.

5 Input

long reads PacBio long reads FASTQ/A format. Reads shorter then 2x the mean short read length are ignored

MiSeq/HiSeq Illumina short reads: FASTQ/A format. Pairing information are not used. The recommended read length is 75bp to 150bp. Reads may differ in length. Use of quality trimmed of even error corrected reads can improve results. The recommended coverage is 30-50X. On large, repetitive dataset we also have had good experiences with normalized read sets.

unitigs

nanopore

custom scenarios

6 Hardware and Parallelization

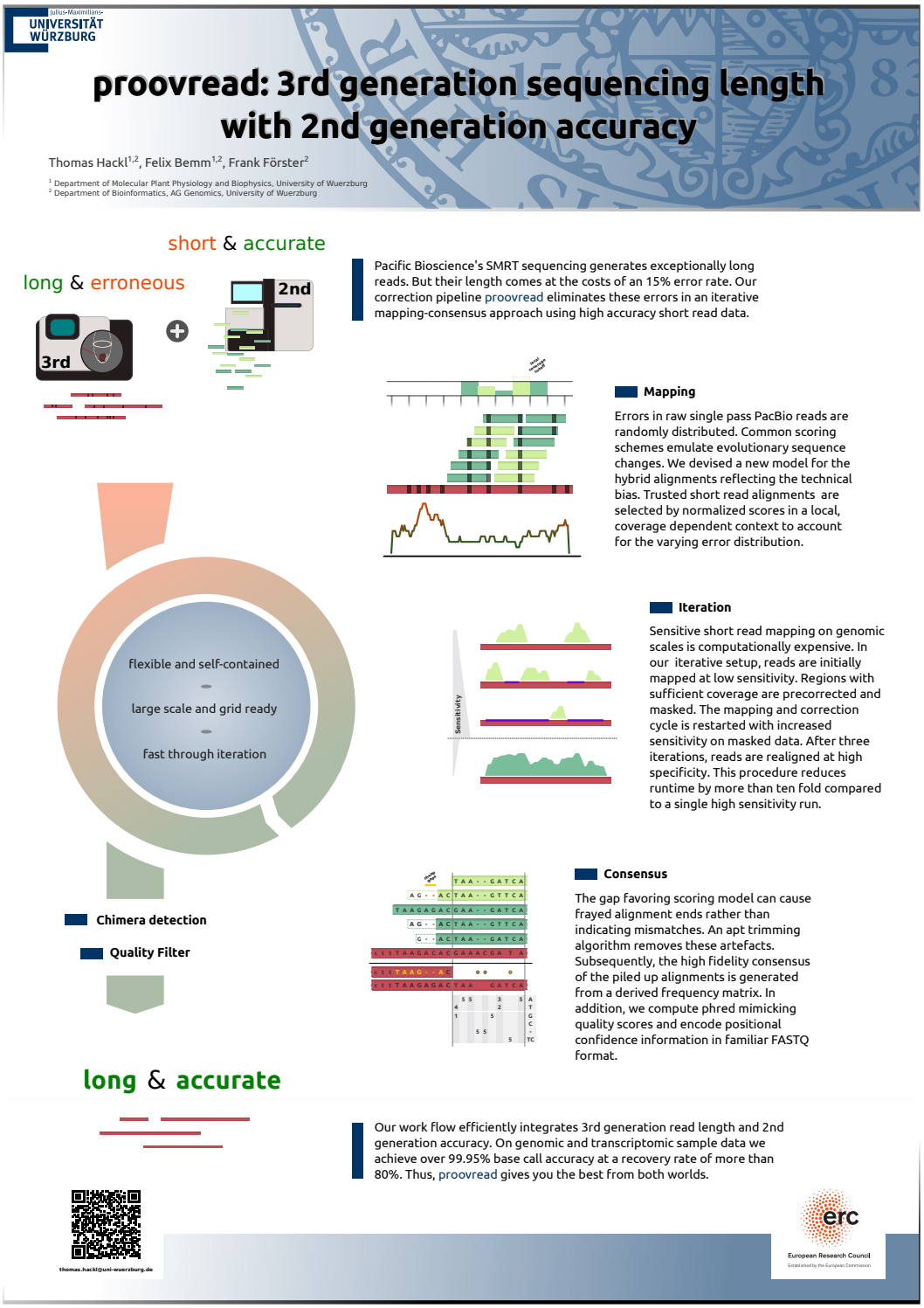
proovread has been designed with low memory node cluster architectures in mind. Peek memory is mainly controlled by the amount of input long reads. With chunks of less than 20 Mbp it easily runs on a 8 GB RAM machine.

Efficient parallelization in most cases is only possible for up to 4 or 8 threads in one instance.

7 Advanced Configuration

`-create-cfg`

8 How proovread work



9 Assembly of proovread read data

10 Citing proovread

proovread: large-scale high accuracy PacBio correction through iterative short read consensus. Hackl, T.; Hedrich, R.; Schultz, J.; Foerster, F. (2014).

`http://dx.doi.org/10.1093/bioinformatics/btu392`?

shrimp

bwa

blasr

11 Contact

If you have any questions or encounter problems or potential bugs, don't hesitate to contact us. Either report issues on github or write an email to:

- Thomas Hackl - thomas.hackl@uni.wuerzburg.de
- Frank Foerster - frank.foerster@biozentrum.uni-wuerzburg.de