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 with binaries of shrimp and blasr as well as bwa source code. If you want to employ other 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:

```
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 Input

long reads

short reads MiSeq/HiSeq

unitigs

nanopore

custom scenarios

5 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`.

6 Advanced Configuration

`-create-cfg`

7 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.

8 Workflow

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

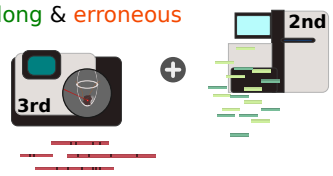
proovread: 3rd generation sequencing length with 2nd generation accuracy

Thomas Hackl[1,2], Felix Bemm[1,2], Frank Förster[2]

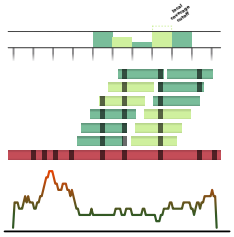
[1] Department of Molecular Plant Physiology and Biophysics, University of Wuerzburg
[2] Department of Bioinformatics, AG Genomics, University of Wuerzburg

short & accurate

long & erroneous

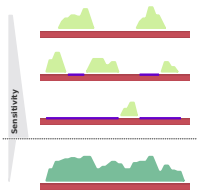


Pacific Bioscience's SMRT sequencing generates exceptionally long reads. But their length comes at the costs of an 15% error rate. Our correction pipeline proovread eliminates these errors in an iterative mapping-consensus approach using high accuracy short read data.



Mapping

Errors in raw single pass PacBio reads are randomly distributed. Common scoring schemes emulate evolutionary sequence changes. We devised a new model for the hybrid alignments reflecting the technical bias. Trusted short read alignments are selected by normalized scores in a local, coverage dependent context to account for the varying error distribution.

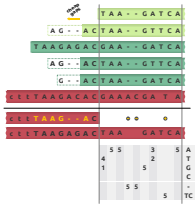


Iteration

Sensitive short read mapping on genomic scales is computationally expensive. In our iterative setup, reads are initially mapped at low sensitivity. Regions with sufficient coverage are precorrected and masked. The mapping and correction cycle is restarted with increased sensitivity on masked data. After three iterations, reads are realigned at high specificity. This procedure reduces runtime by more than ten fold compared to a single high sensitivity run.

Chimera detection

Quality Filter



Consensus

The gap favoring scoring model can cause frayed alignment ends rather than indicating mismatches. An apt trimming algorithm removes these artefacts. Subsequently, the high fidelity consensus of the piled up alignments is generated from a derived frequency matrix. In addition, we compute phred mimicking quality scores and encode positional confidence information in familiar FASTQ format.

long & accurate

Our work flow efficiently integrates 3rd generation read length and 2nd generation accuracy. On genomic and transcriptomic sample data we achieve over 99.95% base call accuracy at a recovery rate of more than 80%. Thus, proovread gives you the best from both worlds.

thomas.hackl@uni-wuerzburg.de

erc
European Research Council
Established by the European Commission

9 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?

10 Contact

If you have any questions or encounter problems or potential bugs, don't hesitate to contact us. Either report issues on github (https://github.com/BioInf-Wuerzburg/proovread/issues) or write an email to:

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