proovread manual

Thomas Hackl

Contents

1 Installation

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

NOTE: proovread comes with its own, slightly modified version of bwa. Using it with a standard bwa built will fail.

2 Dependencies

- Log::Log4perl
- NCBI Blast-2.2.24+ or later
- samtools-1.1 or later

proovread is distributed ready with binaries of SHRiMP2 and BLASR. 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.

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 (BAM) and other intermediary files from iterations have a look at `-keep-temporary`.

The phred scores produced by proovread derive from short read support of each base during correction. The values are scaled to realistically mimic sequencing phred accuracies:

Phred	Accuracy	p33
40	99.99	I
30	99.90	?
20	99.00	5
10	90.00	+

5 Input

5.1 `-long-reads`

Primarily proovread has been designed to correct *PacBio subreads*. You get these reads either from PacBio's SMRT-Portal or by using `dextract` from Gene Myers PacBio assembler DAZZLER, which I would recommend.

In general, reads can be provided in FASTQ or FASTA format. Quality information is used, but only has minor advantages. More valuable are subread information given in default PacBio IDs, which if available are utilized by proovreads `ccseq` module to improve correction performance. Reads shorter then 2x the mean short read length will be ignored.

It is also possible to feed other types of erroneous sequences to proovread, e.g. contigs, 454 reads, ... However, keep in mind that the alignment model for mappings has been optimized for PacBio reads and may produce artifacts in other scenarios. We are currently working on a version optimized for *Oxford Nanopore* data.

5.2 `-short-reads`

For correction of long reads, proovread needs high coverage short read data. Typically these are HiSeq (75-150bp) and MiSeq reads (200-300bp), but also 454 or PacBio CCS reads can be used. Reads need to have FASTQ/A format and may differ in length. Pairing information are not used. Use of quality trimmed or error corrected reads can improve results.

The recommended coverage for short reads data is 50X. If you have less coverage, it is definitely still worth running proovread. However, you will have to have a look at the short reads sampling parameter (sr-sampling) in proovread's Advanced Configuration and adjust them accordingly. If you are having trouble with the adjustment, just write me a quick email. With lower coverage, decreased contiguity is to be expected.

5.3 `-unitigs`

In addition to short reads, unitigs can/should be used for correction in particular for large data sets (eukaryotes). Unitigs are high-confidence assembly fragments produced by for example ALLPATHS or the Celera Assembler. In contrast to contigs, unitigs don't extend past any conflict in the underlying short read data, making them highly reliable.

There are two huge advantages of using pre-computed unitigs:

1. Contiguity: unitigs are longer then corresponding short reads, which makes them easier to align and give better chances to also correct difficult regions.
2. Speed: During unitig computation, all redundancy is removed from the data, creating a minimal set which can be aligned much faster.

However, unitigs only cover regions without conflicts in short read data space. To correct PacBio reads in full length these gaps need to be corrected with primary short read data.

6 Advanced Configuration

proovread comes with a comprehensive configuration, which allows tuning down to the algorithms core parameters. A custom configuration template can be generated with `-create-cfg`. Instructions on format etc. can be found inside the template file.

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 provided. With chunks of less than 20 Mbp it easily runs on a 8 GB RAM machine.

In theory, proovread can be simply parallelized by increasing `-threads`. However, there are single thread steps and other bottlenecks, which at some point render it more efficient, to run e.g. 4 instances at 8 threads in parallel to make full use of a 32 CPU machine.

8 Algorithm and Implementation

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

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

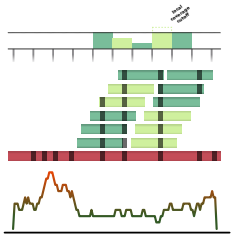
Thomas Hackl[1,2], Felix Bemm[1,2], Frank Förster[2]

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

short & accurate

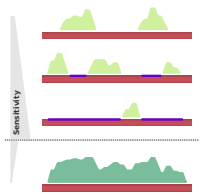
long & erroneous

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.

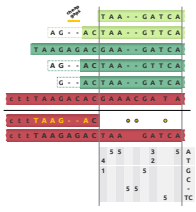


Iteration

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

If you use proovread, please cite:

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

Please, also recognize the authors of software packages, employed by proovread:

Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. Li H. (2012) (bwa)

Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. Mark J Chaisson; Glenn Tesler. (2012)

SHRiMP: Accurate Mapping of Short Color-space Reads. Stephen M Rumble; Phil Lacroute; Adrian V. Dalca; Marc Fiume; Arend Sidow; Michael Brudno. (2009)

10 Contact

If you have any questions, 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