Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

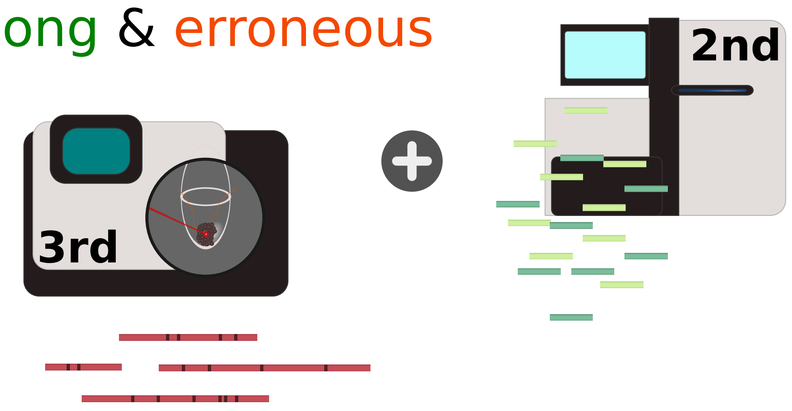
proovread: 3rd generation sequencing length with 2nd generation accuracy

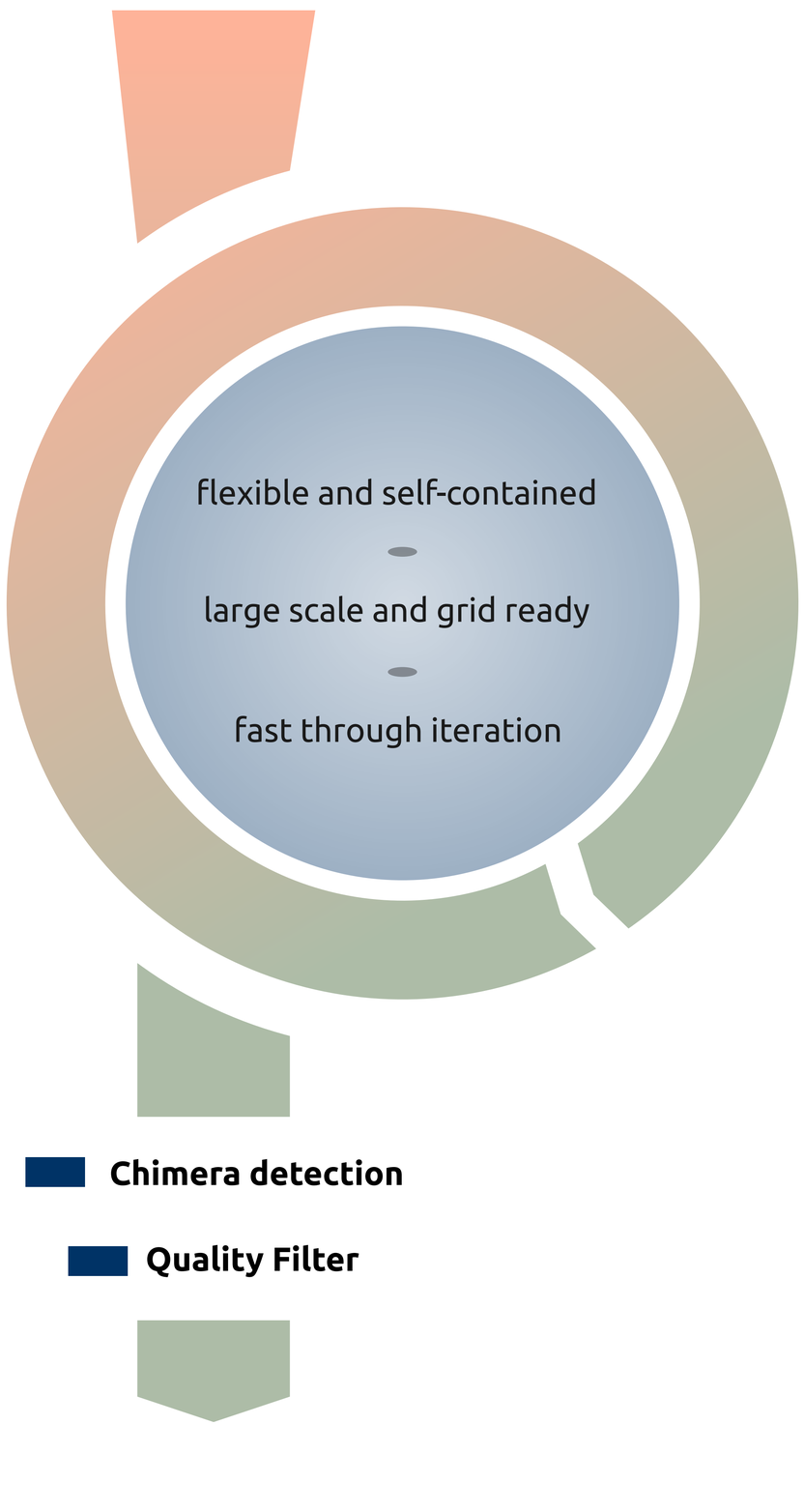
Thomas Hackl[1,2], Felix Bemm[1,2], Frank Förster[2]

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

short & accurate

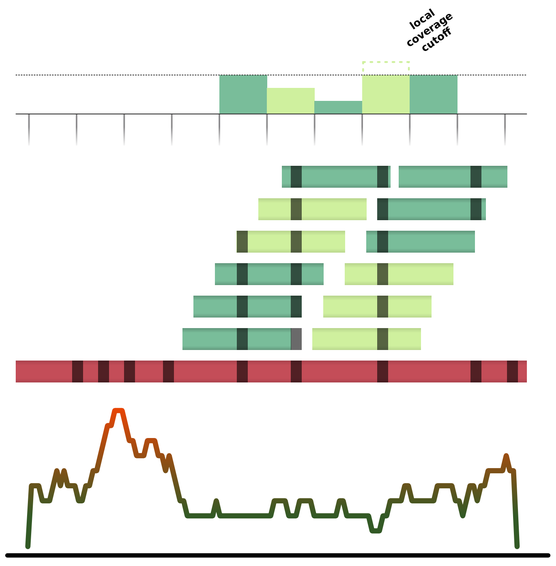
long & erroneous





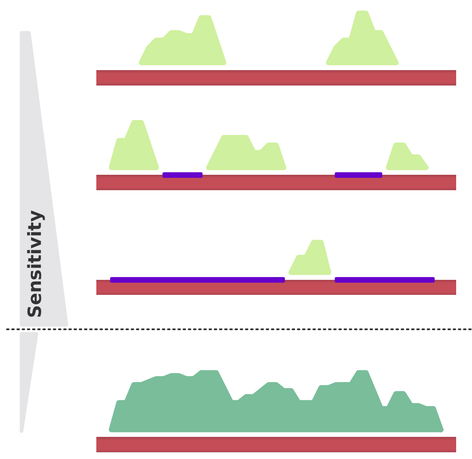
long & accurate

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.



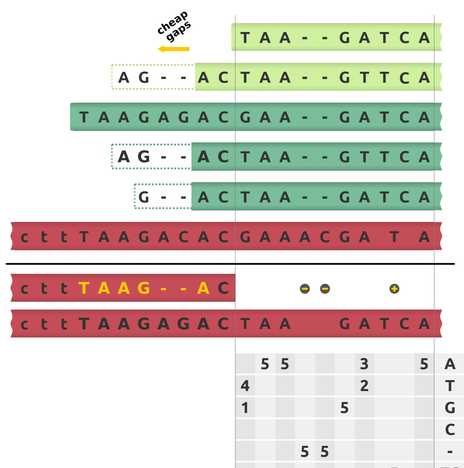
Mapping

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.



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

European Research Council
Established by the European Commission