Prokrastinator

Combining long and short sequencing reads for de novo genome assembly

Sarah von Löhneysen - Bioinf Leipzig

Prokrastinator | overview of the pipeline

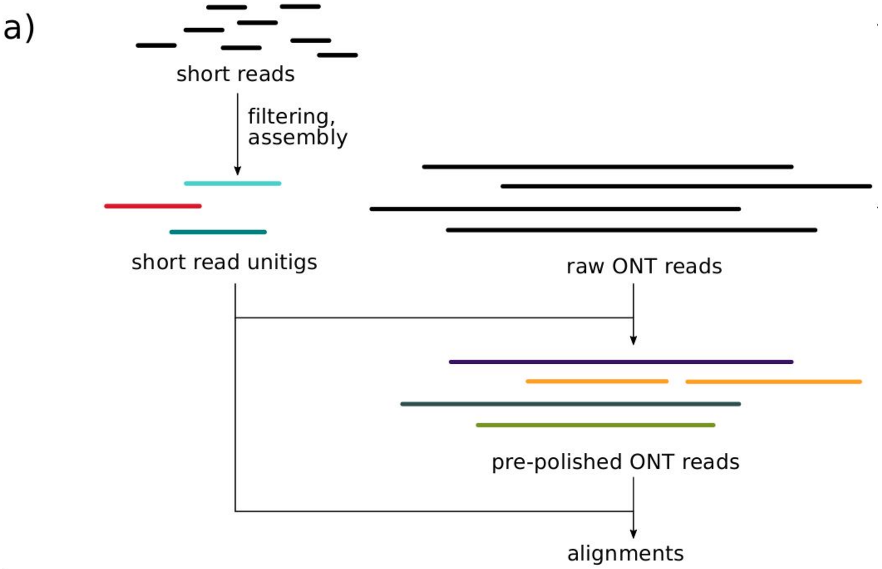
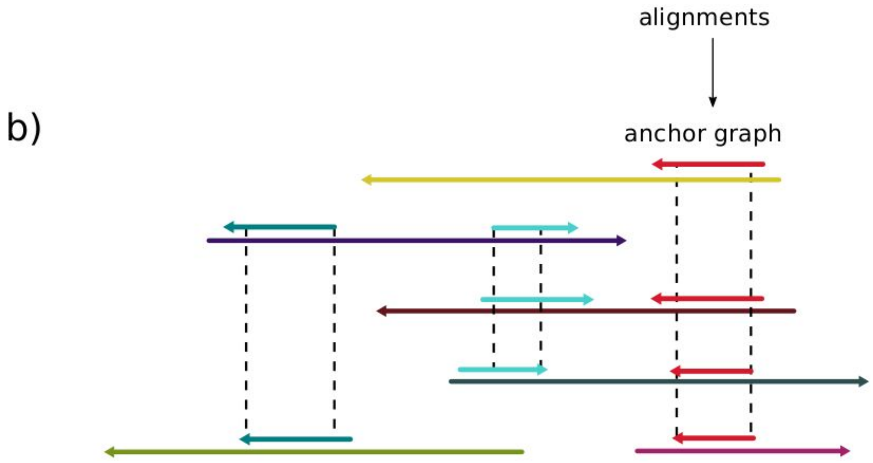


figure: adapted from T. Gatter et al.

→ assemble short reads and align them to long nanopore reads

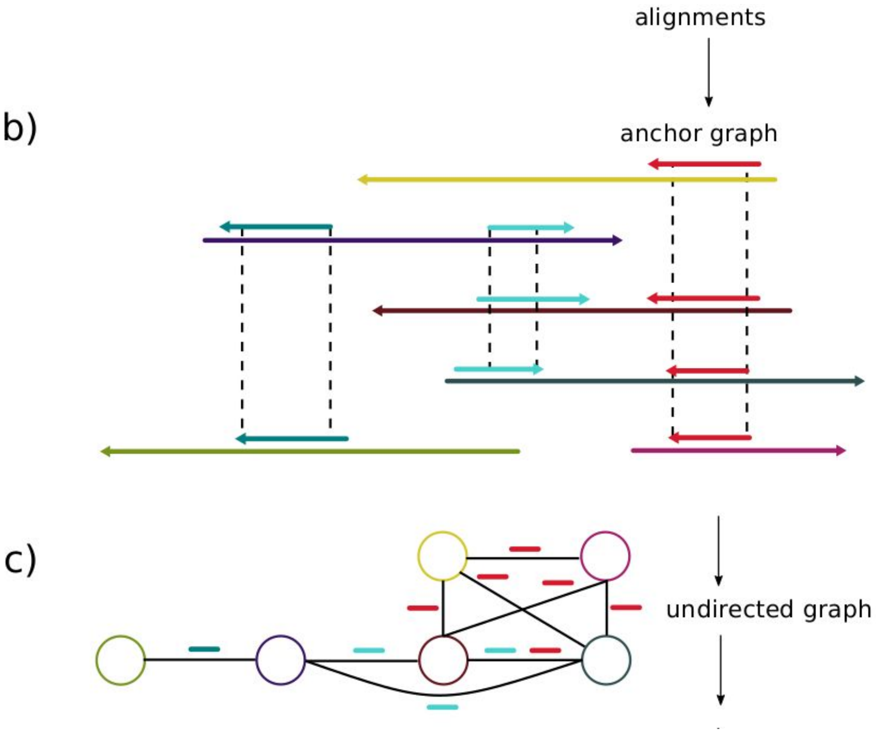
Prokrastinator | overview of the pipeline



→ short read unitigs connect and direct long nanopore reads as “anchors”

figure: adapted from T. Gatter et al.

Prokrastinator | overview of the pipeline

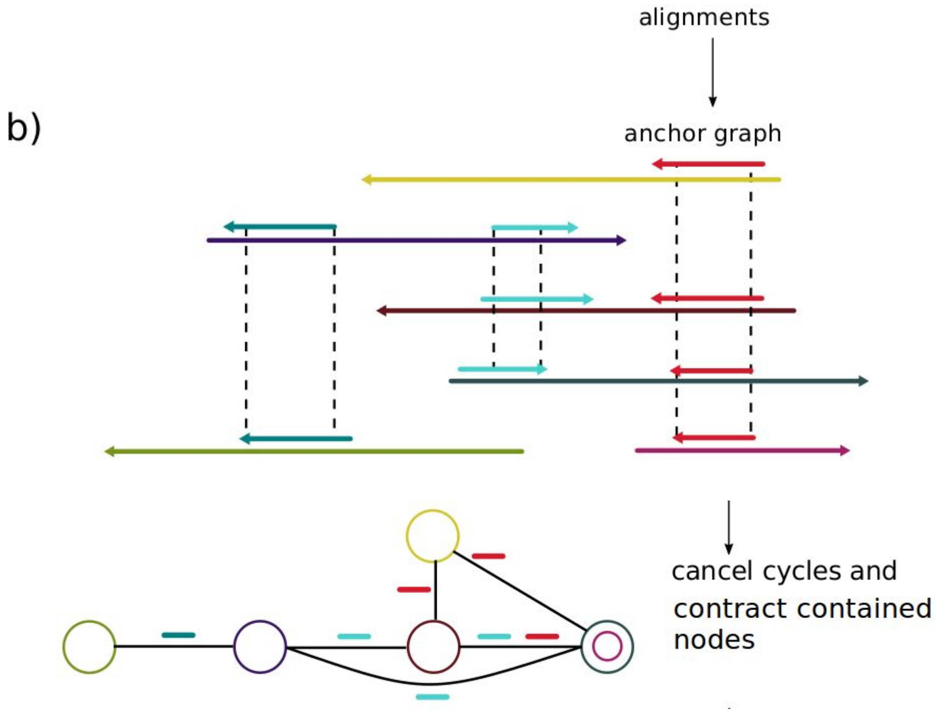


→ short read unitigs connect and direct long nanopore reads as “anchors”

→ nanopore reads are vertices; edge is drawn if anchor indicates an overlap

figure: adapted from T. Gatter et al.

Prokrastinator | overview of the pipeline

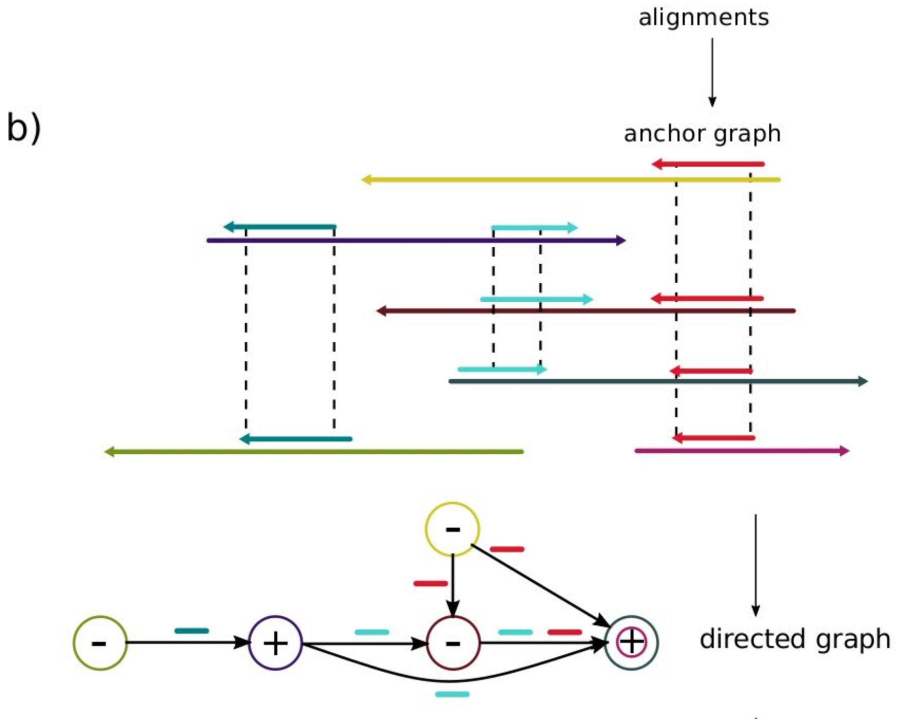


→ short read unitigs connect and direct long nanopore reads as “anchors”

→ cycles with contradiction in orientation are removed

figure: adapted from T. Gatter et al.

Prokrastinator | overview of the pipeline



→ short read unitigs connect and direct long nanopore reads as “anchors”

→ Graph is directed

→ best supported paths are translated into contigs

figure: adapted from T. Gatter et al.

The perfect anchor Set | requirements

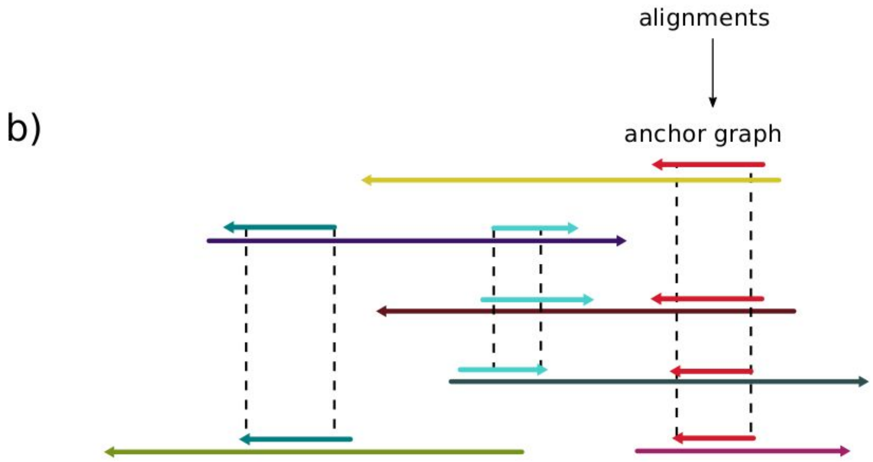


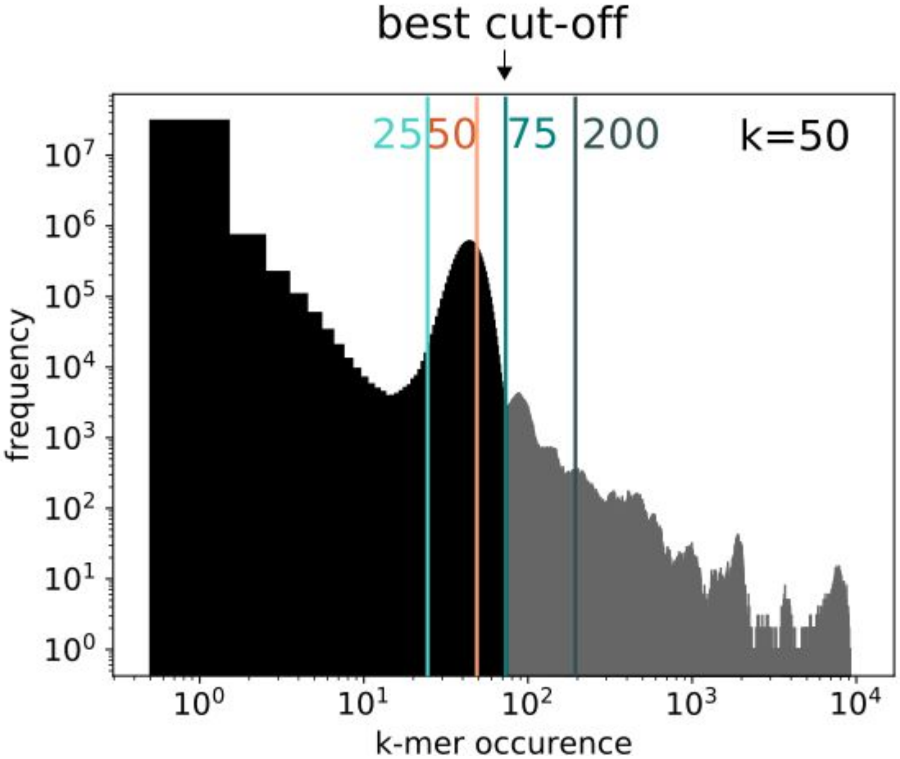
figure: adapted from T. Gatter et al.

→ correct

→ cover genome to sufficient degree

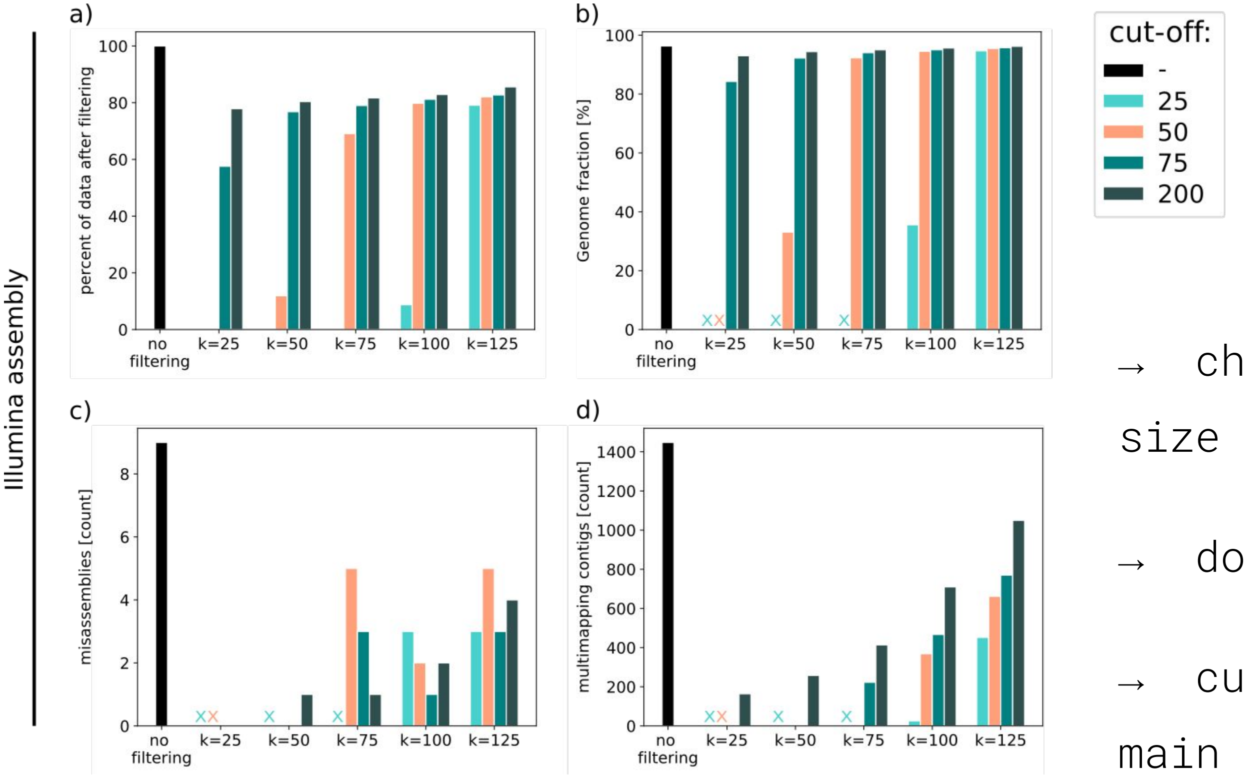
→ unique

unique anchors | k-mer filtering



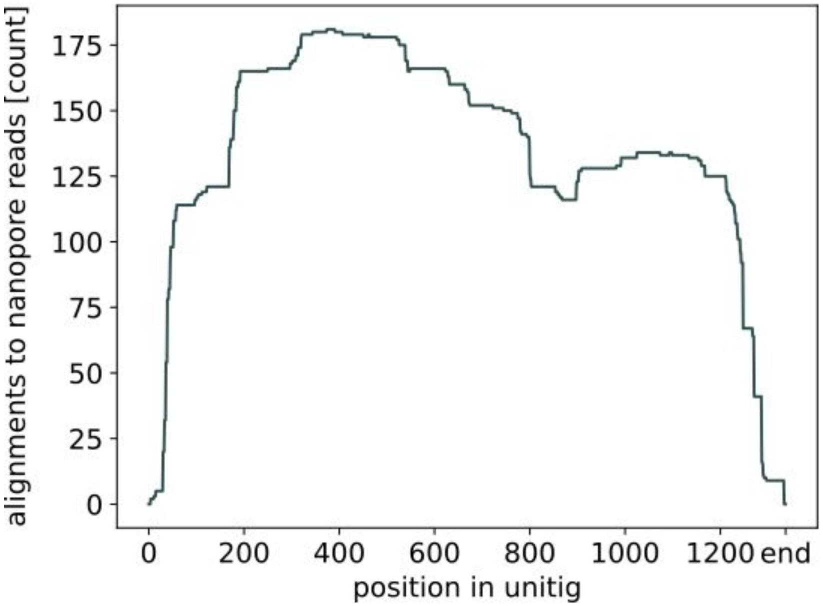
→ different combinations of k-mer size and cut-offs filter Data to different degree

unique anchors | k-mer filtering



- → choose medium k-mer size
- → don’t filter too much
- → cut-off right next to main peak

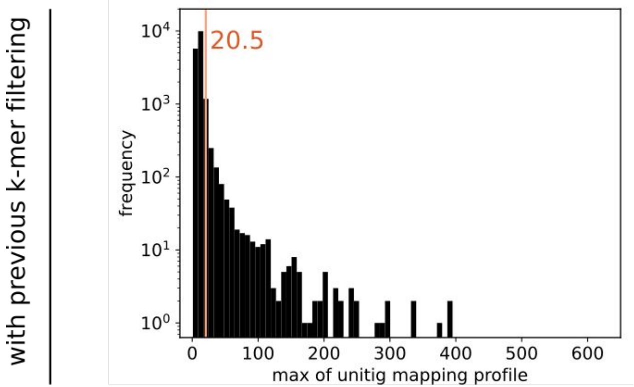
unique anchors | unitig filtering

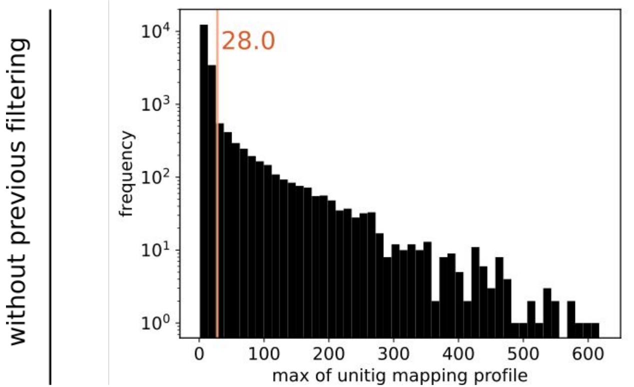


→ anchors are mapped to nanopore reads

→ calculate coverage profile for every unitig

unique anchors | unitig filtering

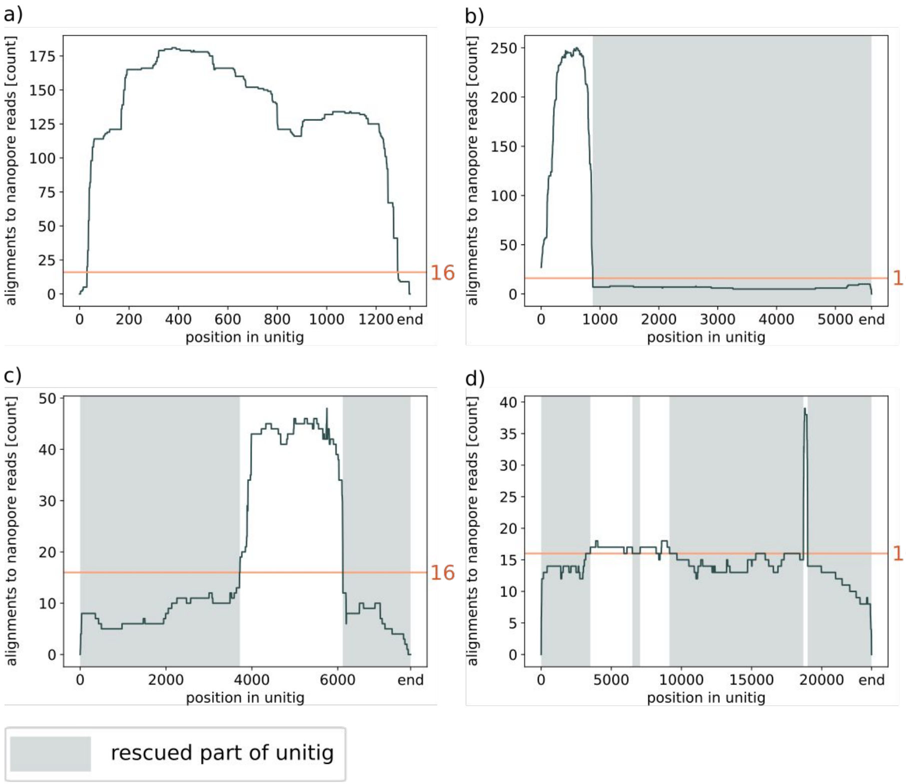




→ find outlier unitigs regarding the max of their coverage profile

→ exclude those from the anchor set

unique anchors | unitig filtering



→ cut peak regions from outlier unitigs

→ rescue regions >= 500 bp that fall under threshold and reinsert to anchor set

filter strategies | results

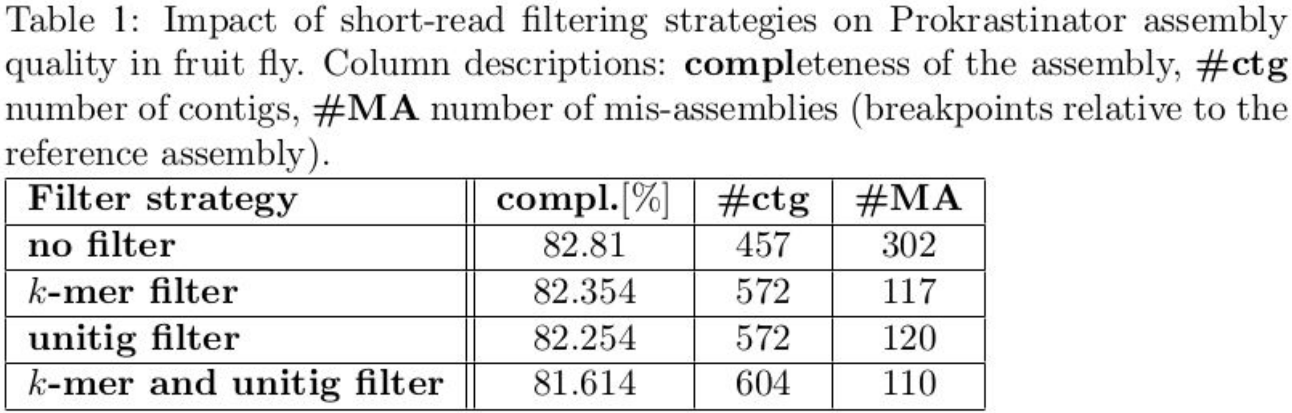
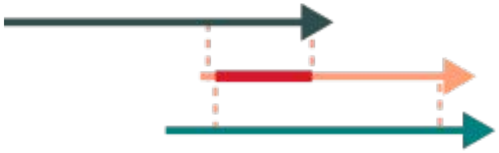


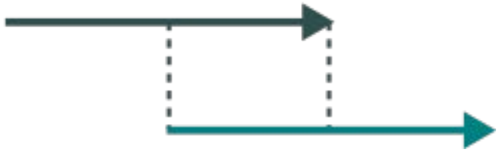
Table 1: Impact of short-read filtering strategies on Prokrastinator assembly quality in fruit fly. Column descriptions: **compl**eteness of the assembly, **#ctg** number of contigs, **#MA** number of mis-assemblies (breakpoints relative to the reference assembly).

Filter strategy	**compl.**[%]	**#ctg**	**#MA**
no filter	82.81	457	302
***k*-mer filter**	82.354	572	117
unitig filter	82.254	572	120
***k*-mer and unitig filter**	81.614	604	110

validation of anchor alignments

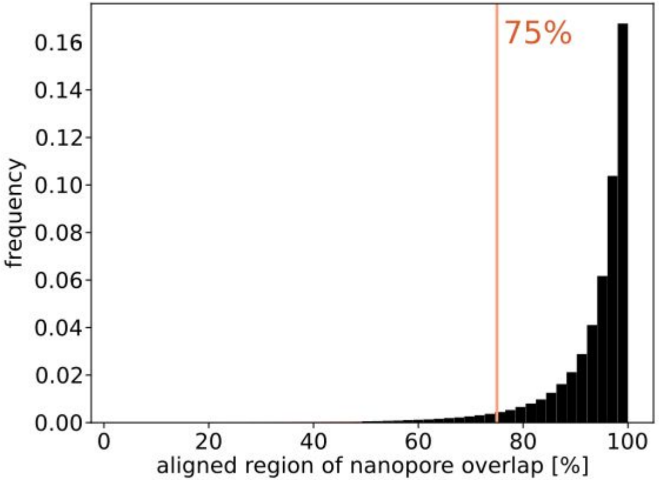


→ alignment with anchor



→ alignment from pairwise mapping

validation of anchor alignments



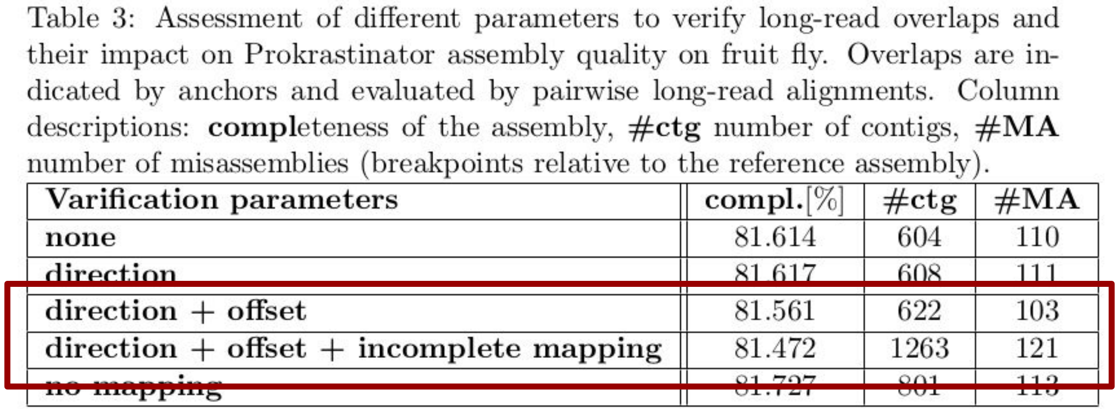


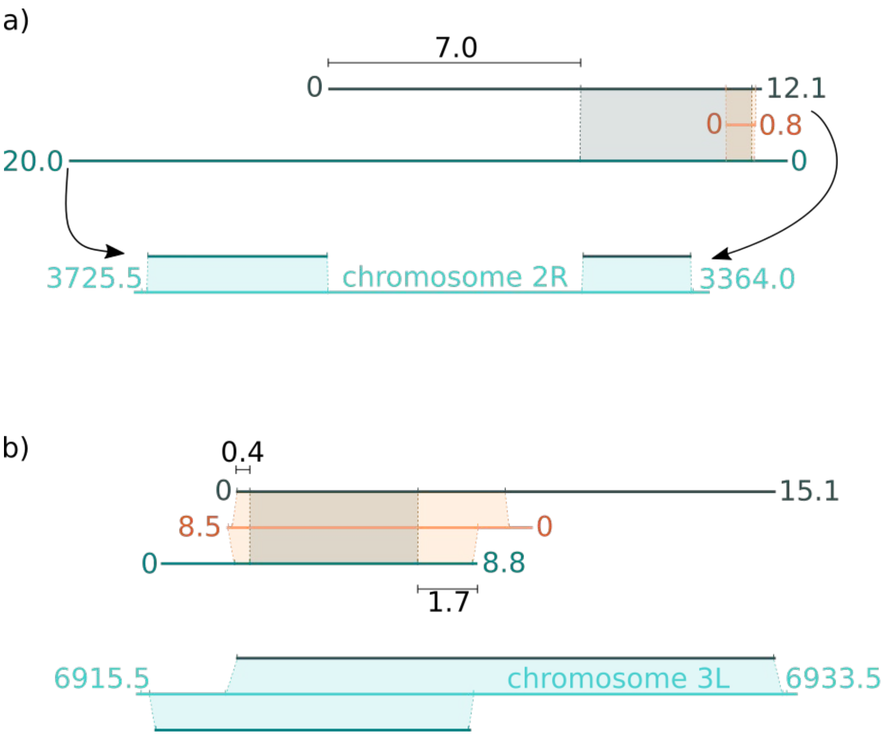
Table 3: Assessment of different parameters to verify long-read overlaps and their impact on Prokrastinator assembly quality on fruit fly. Overlaps are indicated by anchors and evaluated by pairwise long-read alignments. Column descriptions: **completeness** of the assembly, **#ctg** number of contigs, **#MA** number of misassemblies (breakpoints relative to the reference assembly).

Varification parameters	**compl.**[%]	**#ctg**	**#MA**
none	81.614	604	110
direction	81.617	608	111
direction + offset	81.561	622	103
direction + offset + incomplete mapping	81.472	1263	121
no mapping	81.727	801	113

→ 4.6% of the anchor links don’t fulfill the 75% requirement

→ Removing those has negative effect on the final Prokrastinator assembly and tends to break correct contigs apart

validation of anchor alignments



→ we can't distinguish between true and false negatives

→ alignment strategy should be improved

outlook

How can we reliably align long reads with high InDel counts?

Thank you for your attention