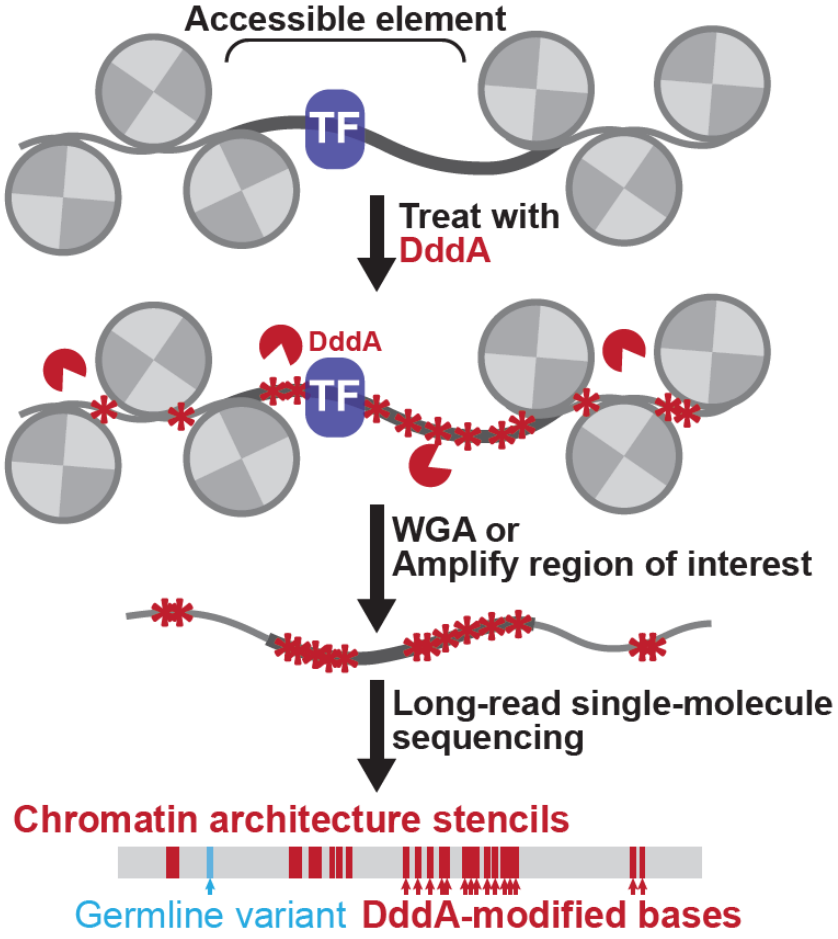
Targeted DAF-seq QC Pipeline

8/7/2025

Stephanie Bohaczuk

Mapping single-molecule chromatin architectures using DAF-seq


Accessible element
TF
Treat with DddA
DddA
TF
WGA or Amplify region of interest
Long-read single-molecule sequencing
Chromatin architecture stencils
Germline variant
DddA-modified bases

Using a non-specific cytidine deaminase (SsDddA) as spray paint

dsDNA cytosine deaminases (DddA) ‘stencil’ chromatin architecture onto DNA molecules via C->U mutations, which are preserved as C->T mutations upon amplification

Swanson, Mao et al., ***bioRxiv*** (2024)

Goals

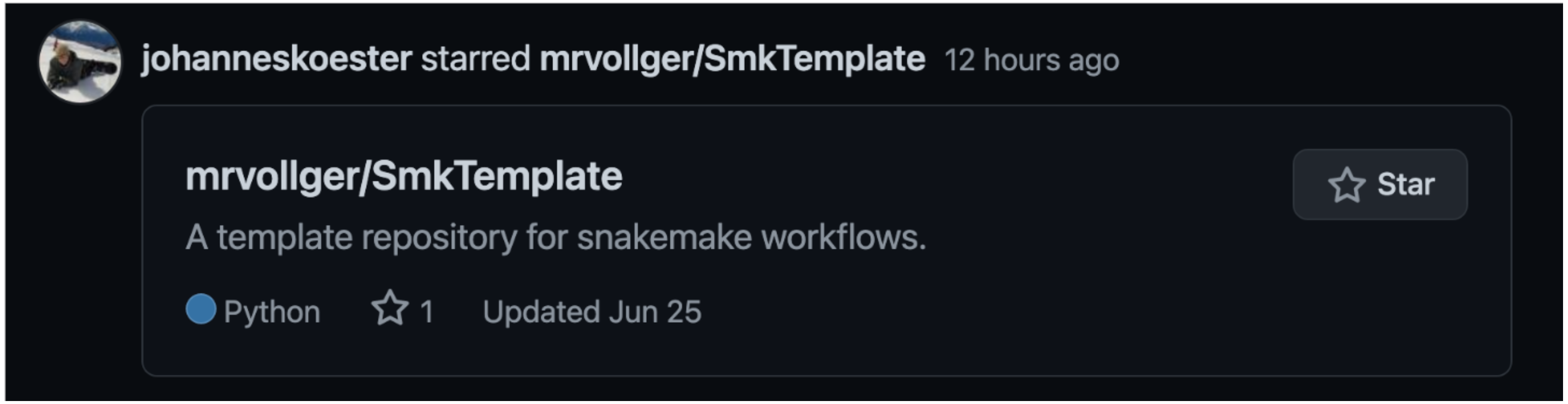
- Standardize file processing
- Automatic QC plots, prod reporter integration
- Flexibility for sample/target specific needs
- Suitable for collaboration (easy to install, sharable results)
- Familiarity for Stergachis lab users
- Minimize storage without sacrificing flexibility

Input		ONT BAM or FASTA (e.g. Plasmidsaurus file)	Unaligned or primary read only PacbBio BAM
Outputs	qc	Data tables • Detailed/summary targeting metrics • Sequence metrics (reads) Read QC plots • Targeting • Per region: • Deamination rate • Mutation rate • Strand type • Enzyme bias	Data tables • Detailed/summary targeting metrics • Sequence metrics (reads) • **Sequence metrics (consensus)** • **Deduplication metrics** Read QC plots • Targeting • Per region: • Deamination rate • Mutation rate • Strand type • Enzyme bias • **Deduplication** **Consensus QC plots (per region):** • **Deamination rate** • **Mutation rate** • **Strand type** • **Enzyme bias**
	align	Aligned bam (all reads) “Decorated” reads (filtered)	Aligned bam (all reads) “Decorated” reads (filtered) **Consensus reads**

Implementation

- Snakemake

- Mitchell’s template (https://github.com/mrvollger/SmkTemplate)
  - Follows familiar design for Stergachis lab/collaborators
  - Modeled on snakemake best practices

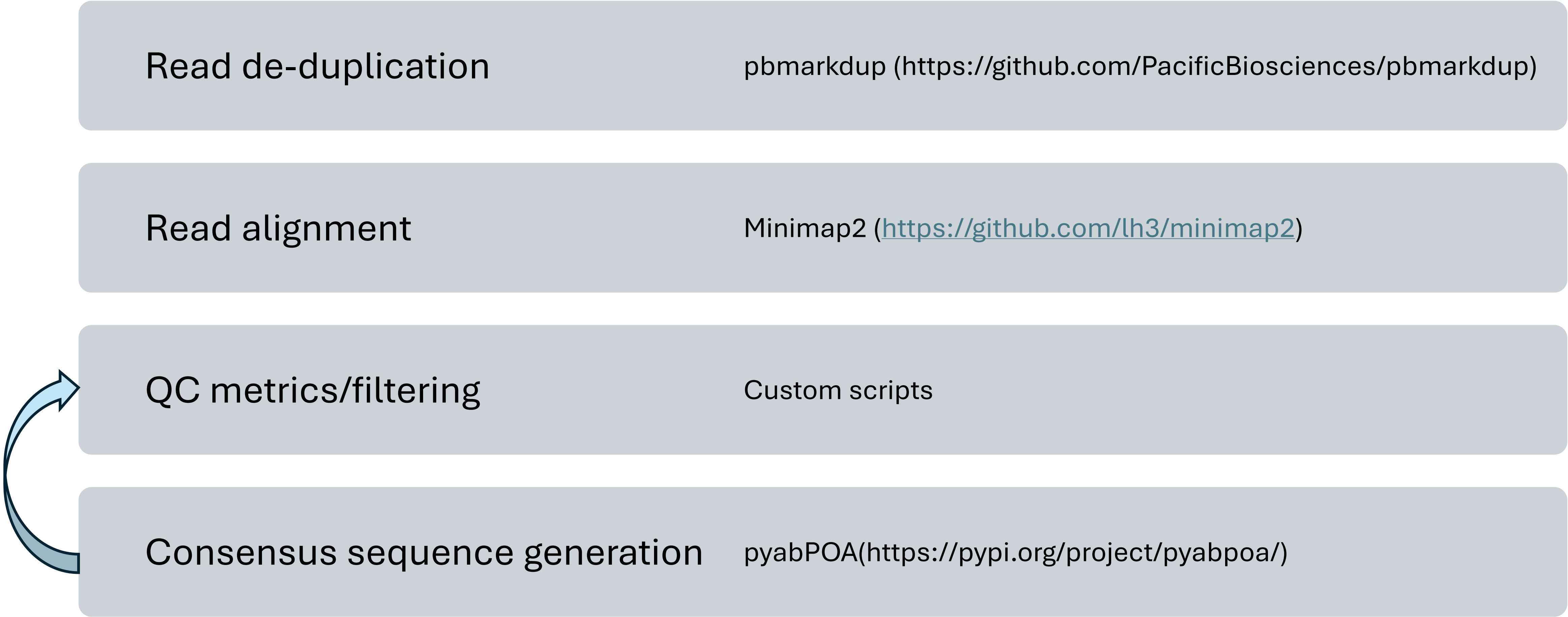


johanneskoester starred mrvollger/SmkTemplate 12 hours ago

mrvollger/SmkTemplate

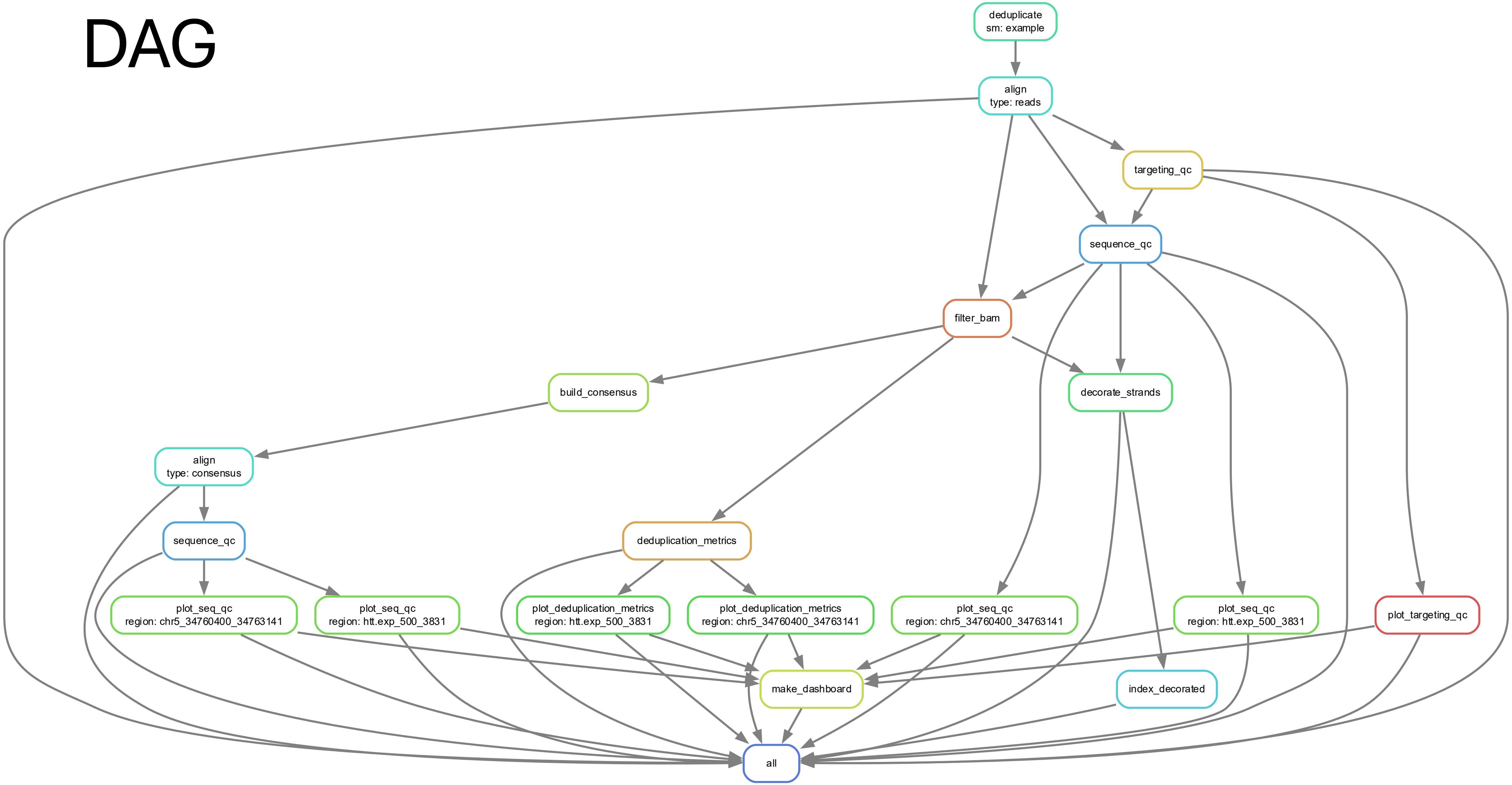
A template repository for snakemake workflows.

Simplified overview of the pipeline



Read de-duplication	pbmarkdup (https://github.com/PacificBiosciences/pbmarkdup)
Read alignment	Minimap2 (https://github.com/lh3/minimap2)
QC metrics/filtering	Custom scripts
Consensus sequence generation	pyabPOA(https://pypi.org/project/pyabpoa/)

DAG

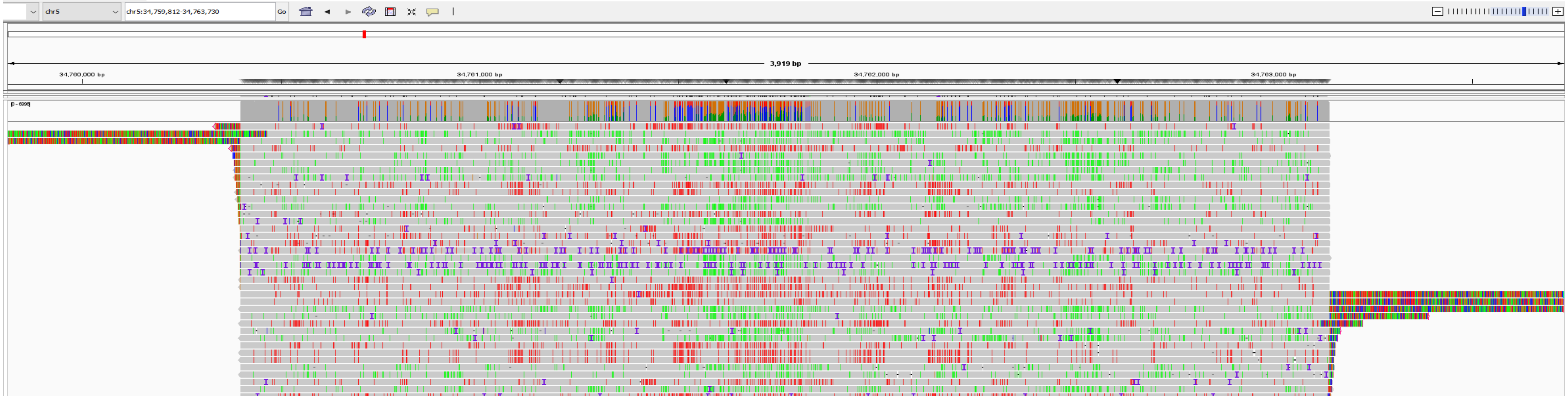

deduplicate
sm: example
align
type: reads
targeting_qc
sequence_qc
filter_bam
build_consensus
decorate_strands
align
type: consensus
sequence_qc
deduplication_metrics
plot_seq_qc
region: chr5_34760400_34763141
plot_seq_qc
region: htt.exp_500_3831
plot_deduplication_metrics
region: htt.exp_500_3831
plot_deduplication_metrics
region: chr5_34760400_34763141
plot_seq_qc
region: chr5_34760400_34763141
plot_seq_qc
region: htt.exp_500_3831
plot_targeting_qc
make_dashboard
index_decorated
all

Outputs: Aligned bam

results/{sample_name}/align/{sample_name}.mapped.reads.bam

- All reads from the input file (incl. unmapped, supplementary)
- du/ds tags from pbmarkdup (PacBio)

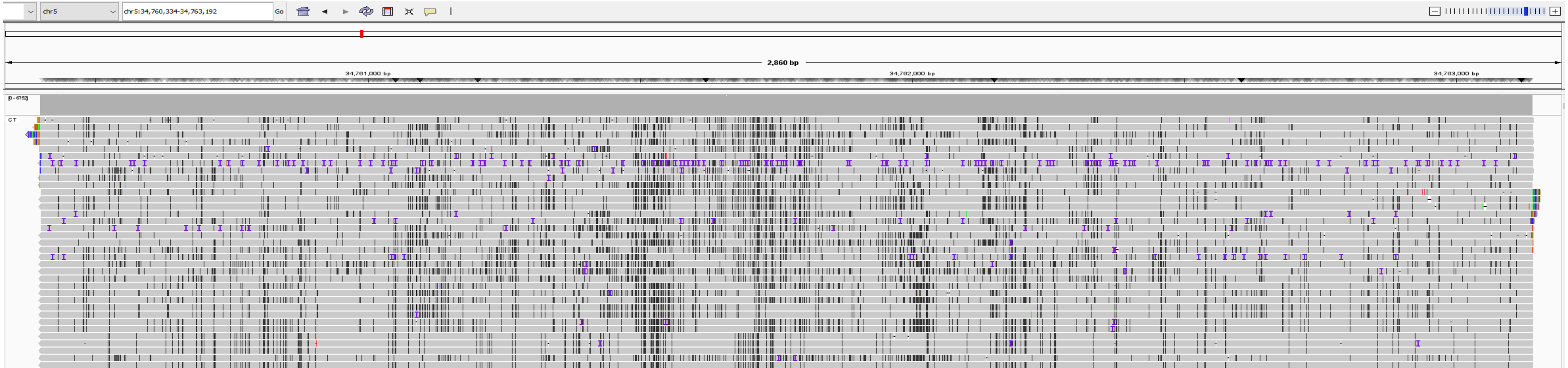
⚠ Caution: pbmarkdup does not include the du tag for the group’s namesake read



Outputs: “Decorated” reads bam

results/{sample_name}/align/{sample_name}.decorated.bam

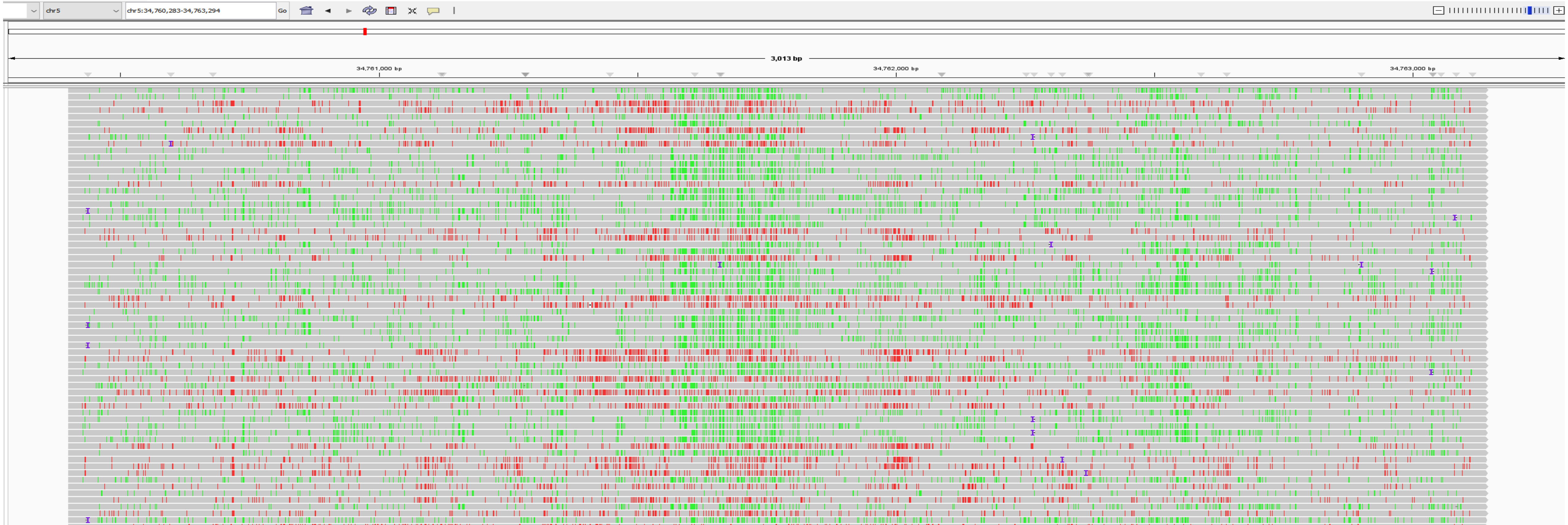
- Sample (5000 default) full-length, on-target reads for visualization
- Deaminations replaced with ambiguous base (CT=Y, GA=R)
- Tags: ‘st’-strand



Outputs: Consensus bam

results/{sample_name}/align/{sample_name}.mapped.consensus.bam

- Consensus sequences from multiple sequence alignment (abPOA)



Outputs: Detailed targeting metrics table

results/{sample_name}/qc/reads/{sample_name}.detailed_targeting_metrics.tbl.gz

Provides detailed information for custom analysis

Column	Description
chrom	Target chromosome
start	Target start
end	Target end
full_length_reads	Comma-separated string with names of reads with primary alignment spanning the whole target, +/- end_tolerance
non_full_length_reads	Comma-separated string with names of reads with primary alignments at the target that fail to span the whole target, +/- end_tolerance
non_primary_reads	Comma-separated string with names of reads with supplementary alignments at the target (span not considered). The same read names could be included in other columns

Outputs: Summary targeting metrics table

results/{sample_name}/qc/{sample_name}.summary_targeting_metrics.tbl
Provides summary information for plotting

Column	Description
chrom	Target chromosome
start	Target start
end	Target end
#_full_length_reads	Count of reads with primary alignment spanning the whole target, +/- end_tolerance
#_non_full_length_reads	Count of reads with primary alignments at the target that fail to span the whole target, +/- end_tolerance
#_non_primary_reads	Count of reads with supplementary alignments at the target (span not considered)
total_fibers in bam(primary+unmapped)	Count of the total number of fibers in the BAM, calculated as the sum of the number of primary aligned reads plus unmapped

Outputs: Detailed sequence metrics table

results/{sample_name}/qc/{type}/{sample_name}.detailed_seq_metrics.{type}.tbl.gz, where type is ‘read’ or ‘consensus’

Each entry is an individual read

Column	Description
read_name	Name of read
chrom	Target chromosome
reg_start	Target start
reg_end	Target end
strand_start	Strand start in genomic coordinates
strand_end	Strand end in genomic coordinates
length	Full length of sequencing read
strand	Strand designation, can be ‘CT’, ‘GA’, ‘none’, ‘undetermined’, or ‘chimera’
duplicate	‘du’ tag value, if present, else None
…	…

Outputs: Detailed sequence metrics table

Column	Description
…	…
mutation_count	Number of missense mutations, does not include mutations consistent with deamination for that strand. None for non ‘CT’ or ’GA’ strands.
deamination_positions	Comma-separated list of deamination-like mutations, ‘C>T’ for ‘CT’ strands and ‘G>A’ for ‘GA’ strands. None for non ‘CT’ or ’GA’ strands.
{N}C_count , where N=A,C,G,T,O	Count of all 2bp {N}C reference positions contained in the region overlapping the read. O is used at read ends and indels. Complements are counted for ‘GA’ strands
{N}C_deam , where N=A,C,G,T,O	Count of deaminated 2bp {N}C reference positions contained in the region overlapped by the read. O is used at read ends and indels. Complements are counted for ‘GA’ strands
{N}C_deam_rate , where N=A,C,G,T,O	Quotient of {N}C_deam /{N}C_count , representing the frequency at which that 2bp context is deaminated. Complements are counted for ‘GA’ strands
total_count	Sum of {N}C_count columns, total count of C reference positions in the region overlapped by the read.
total_deam	Sum of {N}C_deam columns, total count of C reference positions in the region overlapped by the read.
all_deam_rate	Quotient of total_deam/total_count, representing the overall deamination rate
mutation_rate	Quotient of mutation_count/length. Does not consider indels.

Outputs: Summary sequence metrics table

results/{sample_name}/qc/{type}/{sample_name}.summary_seq_metrics.{type}.tbl.gz, where type is ‘read’ or ‘consensus’

Provides summary information for plotting

Column	Description
chrom	Target chromosome
reg_start	Target start
reg_end	Target end
strand	Strand designation, can be ‘CT’, ‘GA’, ‘none’, ‘undetermined’, or ‘chimera’
count	Count of fibers represented
mutation_rate	Comma-separated list of individual fiber mutation rates. Only for ‘CT’ and ‘GA’ strands
all_deam_rate	Comma-separated list of individual fiber overall deamination rates. Only for ‘CT’ and ‘GA’ strands
{N}C_deam_rate, where N=A,C,G,T,O	Comma-separated list of individual fiber deamination rates at the {N}C base pair. Only for ‘CT’ and ‘GA’ strands.

Outputs: Deduplication metrics table (PacBio only)

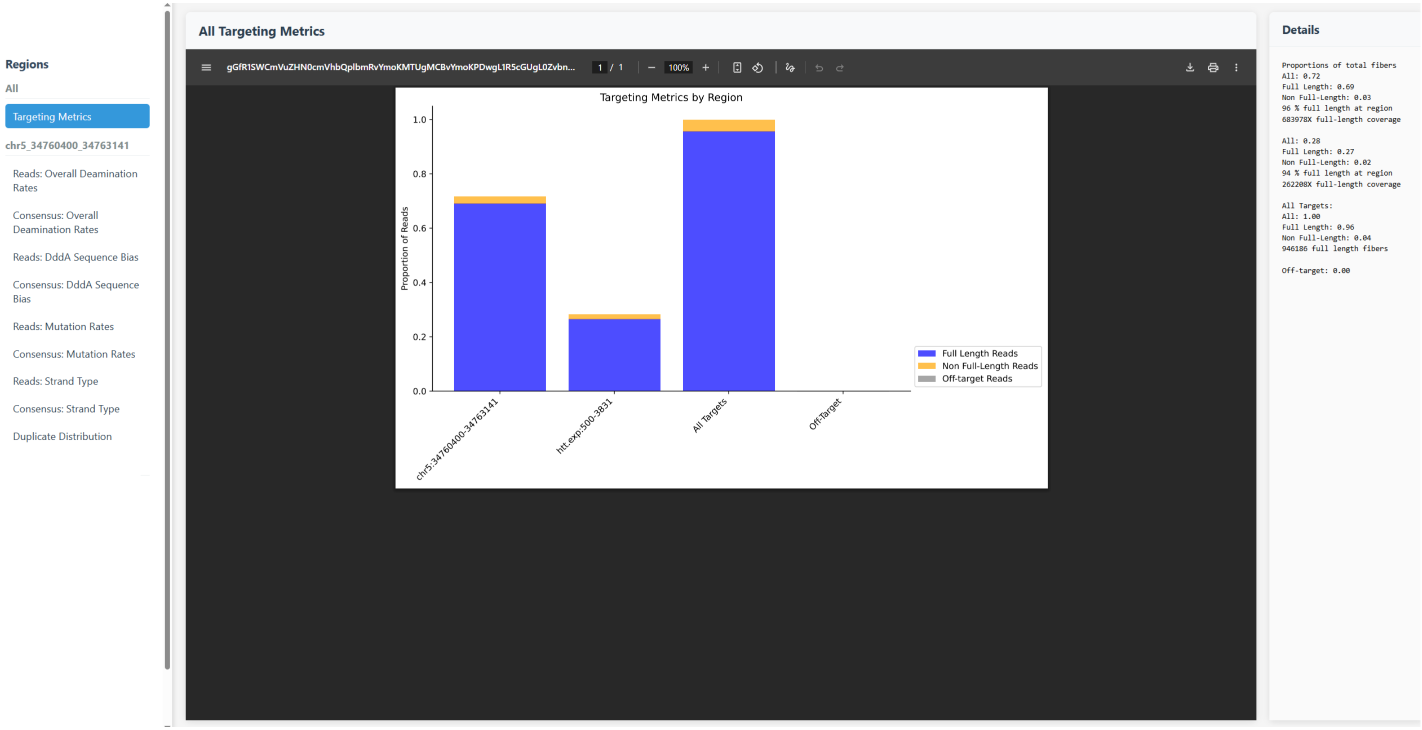
results/{sample_name}/qc/reads/{sample_name}.deduplication_metrics.tbl.gz

Only considers full-length, ‘CT’ or ‘GA’ strand reads

Column	Description
chrom	Target chromosome
start	Target start
end	Target end
du_tags	Comma-separated string of pbmarkdup ‘du’ tags, or read names for unique reads
counts	Comma-separated string with count of full-length, ‘CT’ or ‘GA’ strand reads in each ‘du’ tag group

Outputs: QC Plots Dashboard (Demo)

results/{sample_name}/qc/{sample_name}.dashboard.html



```
Proportions of total fibers
All: 0.72
Full Length: 0.69
Non Full-Length: 0.03
96 % full length at region
683978X full-length coverage

All: 0.28
Full Length: 0.27
Non Full-Length: 0.02
94 % full length at region
262208X full-length coverage

All Targets:
All: 1.00
Full Length: 0.96
Non Full-Length: 0.04
946186 full length fibers

Off-target: 0.00
```

Can also access individual plot pdfs in results/{sample_name}/qc/{type}/plots/, where type is "read" or "consensus"

Quick guide to running pipeline

Setup: Install pixi and clone repository from github

1. Create manifest .tbl file
2. Create config .yaml file
3. Run the pipeline (optional dry run first with –n)
   - pixi run snakemake --configfile config/config.yaml
4. Copy the QC HTML locally to view

Manifest (.tbl) file

- Three columns, sample, file and regs
- Regions should be comma separated, formatted as chr:xx-xx
- No explicit limit on number of regions

```
sample  file    regs
sample1 /path/to/sample/sample1.bam  chr4:3073138-3075853,chr3:179228176-179236561
sample2 /path/to/sample/sample1.bam  chr4:3073138-3075853
```

Config (.yaml) file

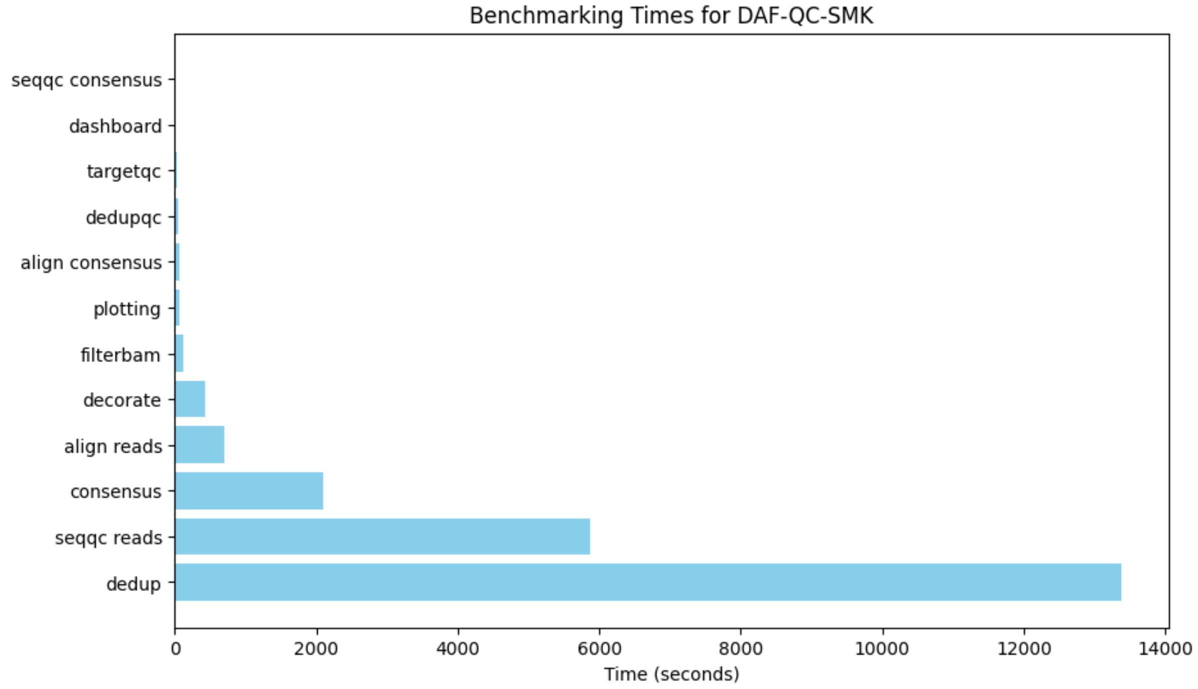
- Requires reference, manifest, and platform designation

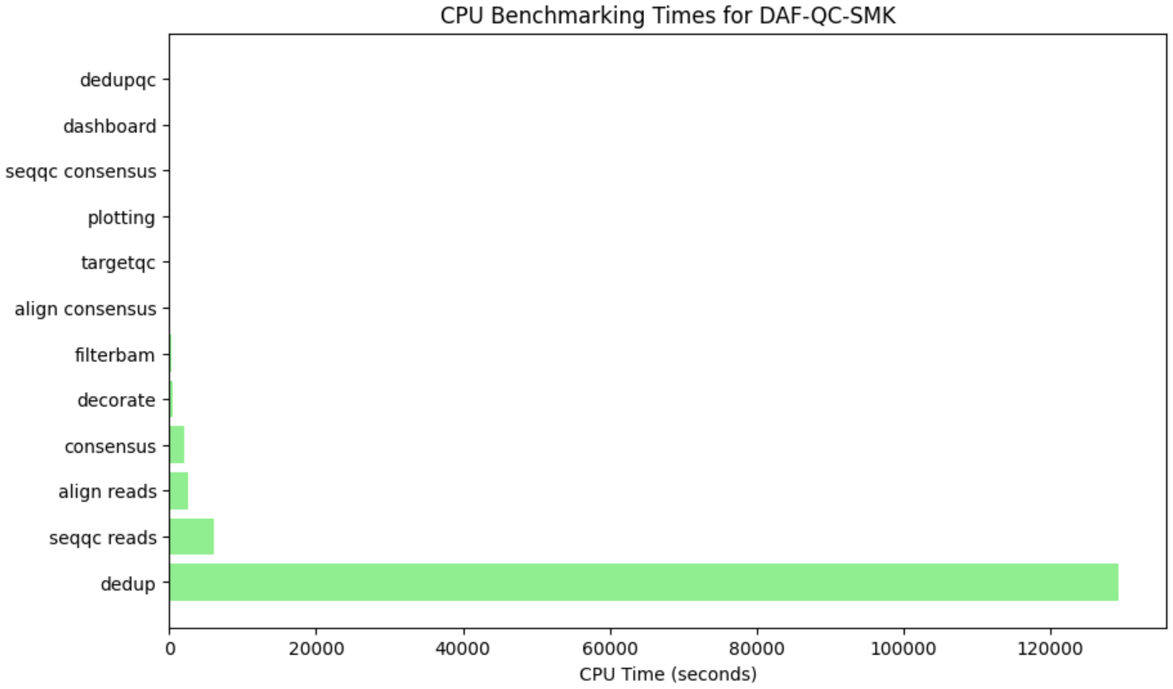
```
ref: /path/to/ref.fa # path to a reference fasta
manifest: config/config.tbl # table with samples to process

platform: pacbio # sequencing platform, either 'pacbio' or 'ont'. Default is pacbio


# Optional (both platforms)
chimera_cutoff: 0.9 # minimum fraction of (C->T|G->A)/(C->T+G->A) for ‘CT’ or ‘GA’ strand designation
min_deamination_count: 50 # minimum number of deaminations for ‘CT’ or ‘GA’ strand designation
end_tolerance: 30 # +/- end length tolerance (in bp) for classifying full-length reads
decorated_samplesize: 5000 # Approximate number of reads to output as decorated reads bam for visualization
```

Config (.yaml) file continued

```
# PacBio-specific options. Input should always be bam files
dup_end_length: 0 # length to consider for deduplication. 0 will consider the whole read
dup_min_id_perc: 99.2 # minimum identity percentage required to mark duplicates
consensus: True # whether to generate consensus sequences. Default is True
consensus_min_reads: 3 # minimum number of reads required to generate a consensus
sequence

# ONT-specific options
is_fastq: False # for ONT files only, specifies fastq input file, otherwise bam is
expected. Default is False
```

Bottlenecks





Run on compute-ultramem, c -10, MEM=100G

Useful tips

- For one sample, a single node with 100 GB RAM, 10-20 cores is sufficient
- Grab full length read names, strand designation, deamination proportions, and deamination positions from results/{sample_name}/qc/{type}/{sample_name}.detailed_seq_metrics.{type}.tbl.gz
- Set “decorated_samplesize” to a really big number e.g. 100000000 if you want every read processed (this could take a long time)
- Pipeline tolerates empty regions, but at least one region must have a full length read (check targeting metrics files if pipeline fails)

Acknowledgements

Stergachis lab members, especially:

- Mitchell Vollger
- Elliott Swanson
- Shane Neph
- Chris Oliveira
- Annelise Mah-Som