If poly-A tails represent more than >=1% of the reads, does it mean this is a significant problem? Or what is the reason behind you rpoposing to test specifically for >=1% in option 2?

README.md:234

> Outlining steps for RNA-seq analysis...

Short answer first: 1% is a heuristic — a conservative flag to catch experiments where a non-trivial fraction of reads are dominated by long poly-A runs. It is not a hard rule; whether it's "significant" depends on your library prep, sequencing depth, and downstream goals.

Why 1% was suggested (intuition)

- It's small enough to ignore random noise but large enough to catch systematic problems.
- If only 0.1% of reads have long poly-A tails, that's usually random/background.
- If >=1% of reads do, that often indicates a library or sequencing artifact that could bias results (especially for quantification or low-input experiments).

Practical thresholds (guideline, not absolute)

- < 0.1% — negligible for most workflows.
- 0.1% – 1% — low; worth a quick check but often not critical.
- 1% – 5% — moderate; investigate (QC, sample reads) and consider trimming.
- 5% — likely a real problem (library prep or sequencing) that may require reprocessing or re-prepping libraries.

Why poly-A reads can be a problem

- They waste sequencing capacity (reads not informative).
- They can bias expression quantification if they map poorly or pile up at transcript ends.
- They may indicate adapter/overhang/fragmentation issues or contamination.

What to check next (short actionable plan)

1. Confirm these are artifactual, not biological (mRNA has poly-A tails but long terminal poly-A in reads at notable frequency is usually an artifact).
2. Inspect a sample of poly-A reads to see the pattern (terminal vs internal).
3. Run a QC tool (FastQC/fastp) to check per-base composition and overrepresented sequences.
4. If needed, trim poly-A tails and re-run QC and a small mapping test to check improvement.

Commands you can run (copyable) — quick checklist (I won't run them for you unless you ask):

- Show sample poly-A reads (example: reads ending with ≥10 A):
  zcat data/yourfile.fastq.gz | awk 'NR%4==1{h=$0} NR%4==2{seq=$0} NR%4==0{if(toupper(seq) ~ /A{10,}$/) {print h; print seq; getline; print $0}}' | head -n 12
- Run FastQC (nice overview):
  fastqc data/*.fastq.gz
- Trim poly-A tails with fastp (simple and robust):
  fastp -i data/ERR*_R1.fastq.gz -I data/ERR*_R2.fastq.gz -o trimmed_R1.fastq.gz -O trimmed_R2.fastq.gz --trim_poly_x -w 4 -h fastp_report.html
  (fastp's --trim_poly_x strips long A/T/G/C tails at read ends; it's a safe general option.)
- Re-check poly-A fraction after trimming with the small awk/zcat loop you already used.

When trimming is (and isn't) appropriate

- Trim when poly-A reads are artifactual and non-negligible (>= ~1%); trimming often improves mapping/quantification.
- Don't trim blindly if your experiment intentionally targets polyadenylated RNA in a way that naturally produces tail-containing reads — instead inspect a few reads to decide.

If you want, I can:

- Run a focused sample inspection on your `data/*.fastq.gz` to show example reads with poly-A tails, or
- Run the fastp trimming command on one file and show the before/after poly-A percentages.

Which follow-up would you like me to do?