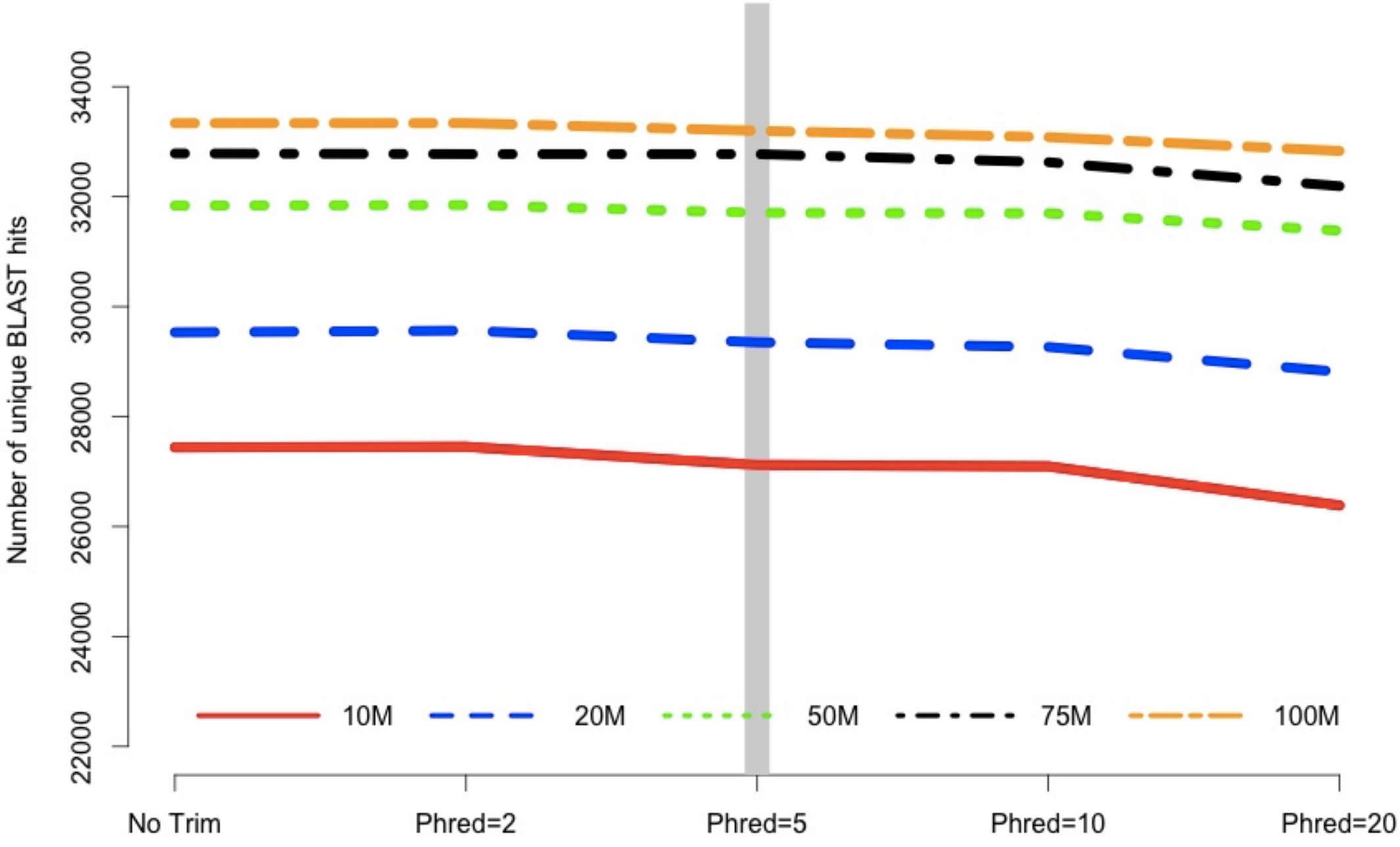
How many reads/reps/samples

- For *de novo* transcriptome assembly, diff expression
  - SNP calling
- For reference generation?? (20-30M? Francis et al 2013)

How many reads/reps/samples



How many reads/reps/samples

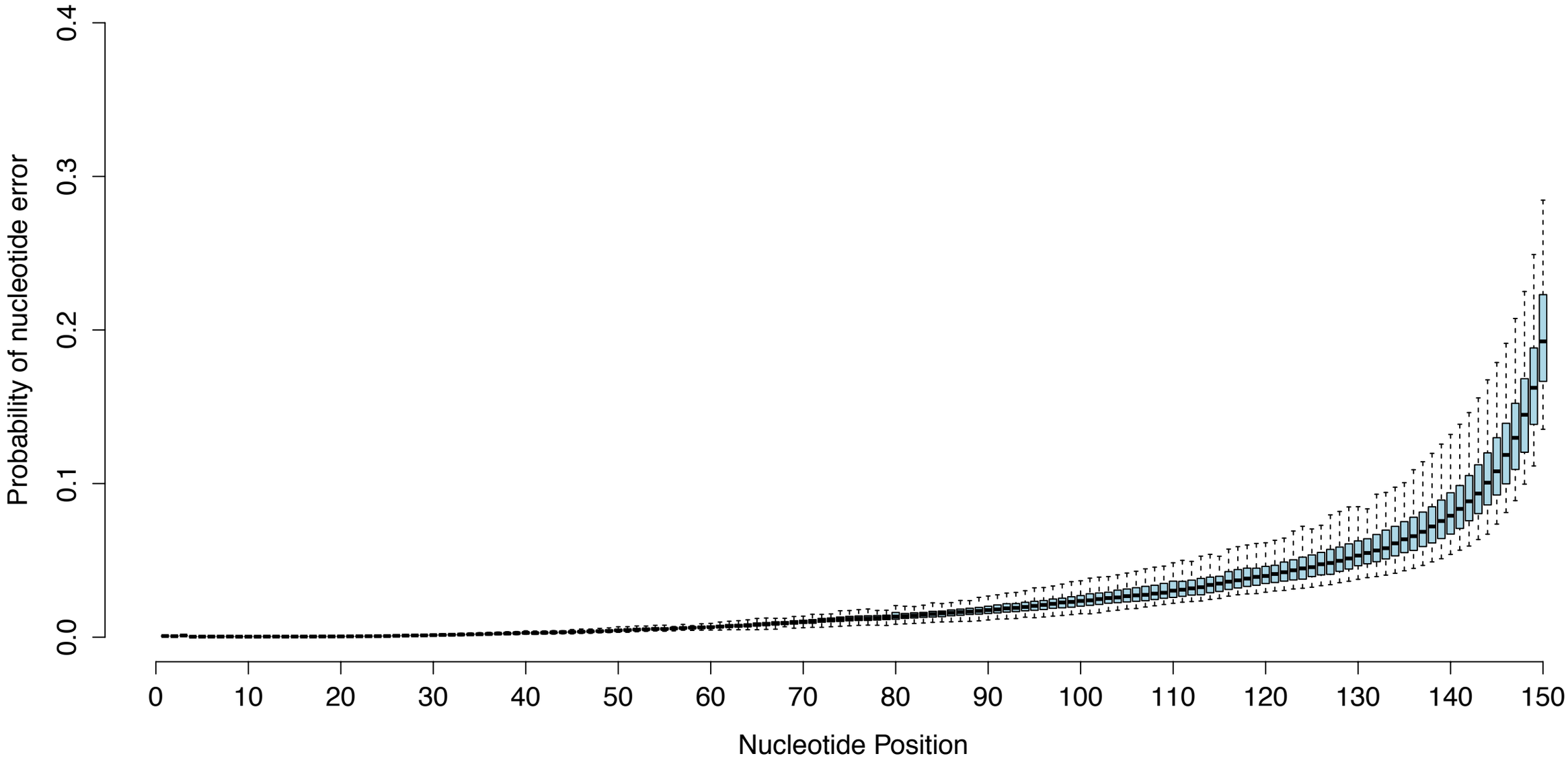
- For *de novo* transcriptome assembly, diff expression
  - SNP calling
- For reference generation??
- What is the n treatments are very different?
- Now you have a reference, how many reads for replicates?
- How to sequence these samples?

Now we have sequence reads…

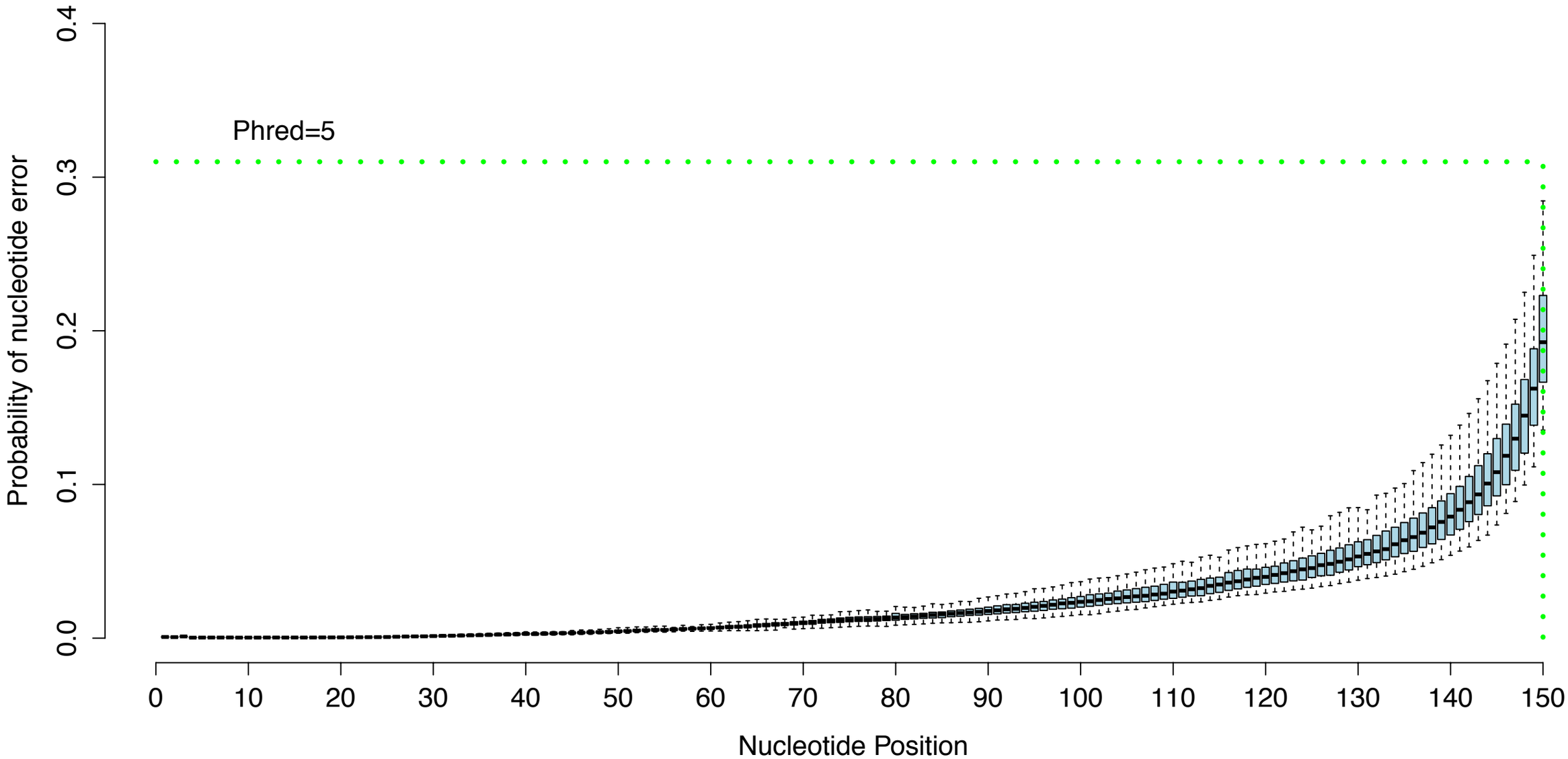
- Adapter trimming (Trimmomatic)
- Quality trimming

Quality trimming of NGS data

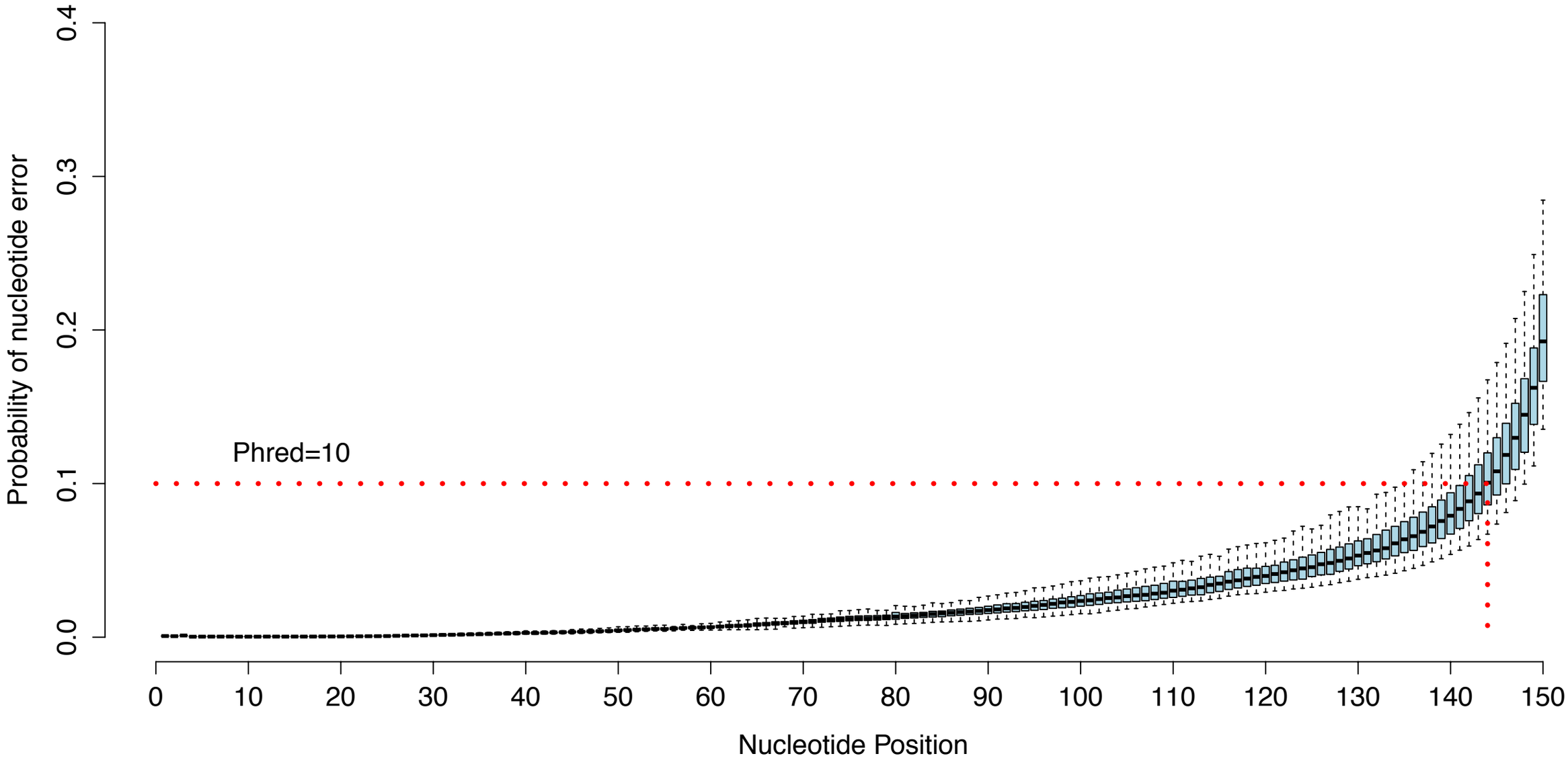
- Universal practice



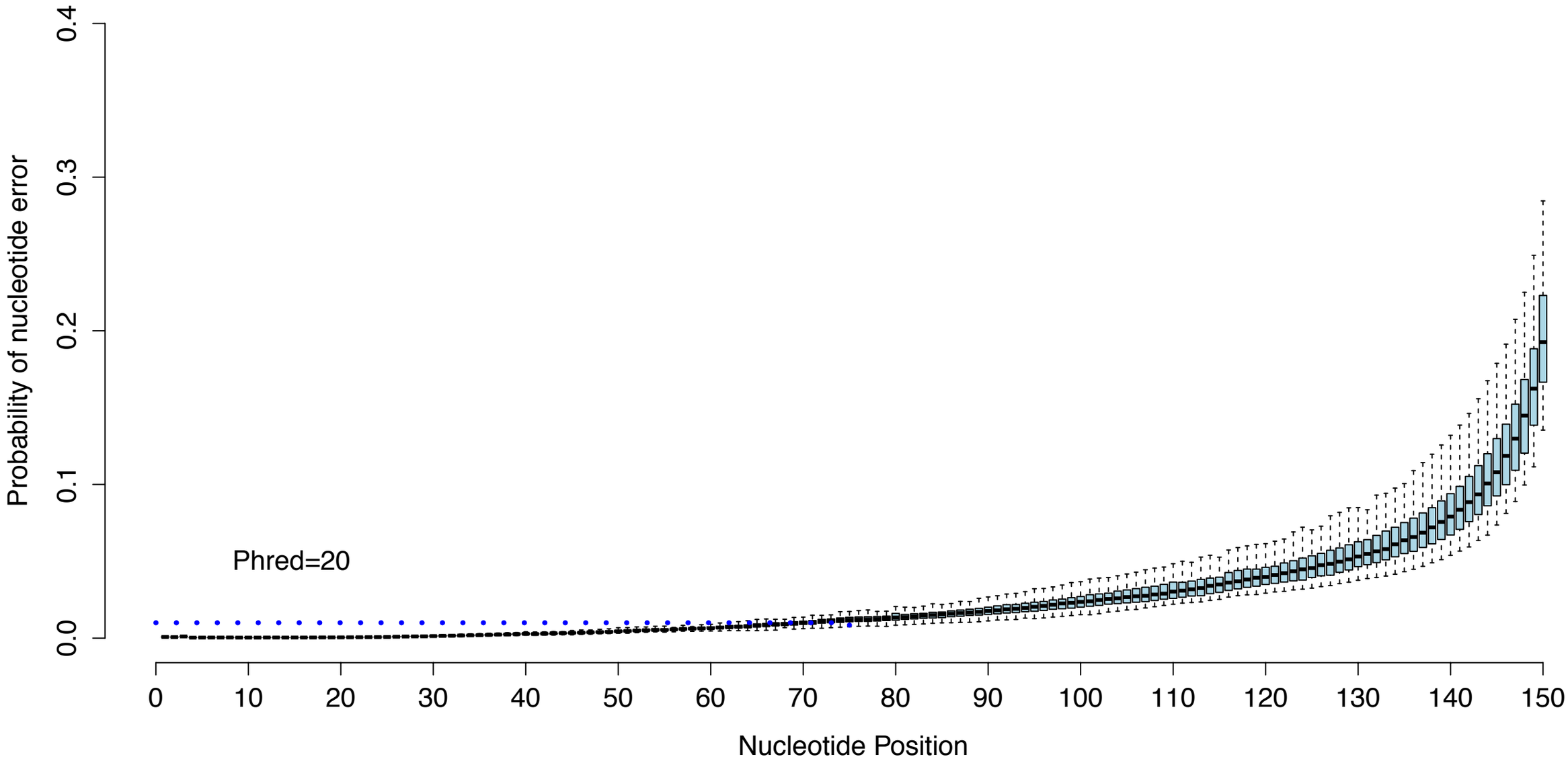
Quality trimming of NGS data



Quality trimming of NGS data



Quality trimming of NGS data

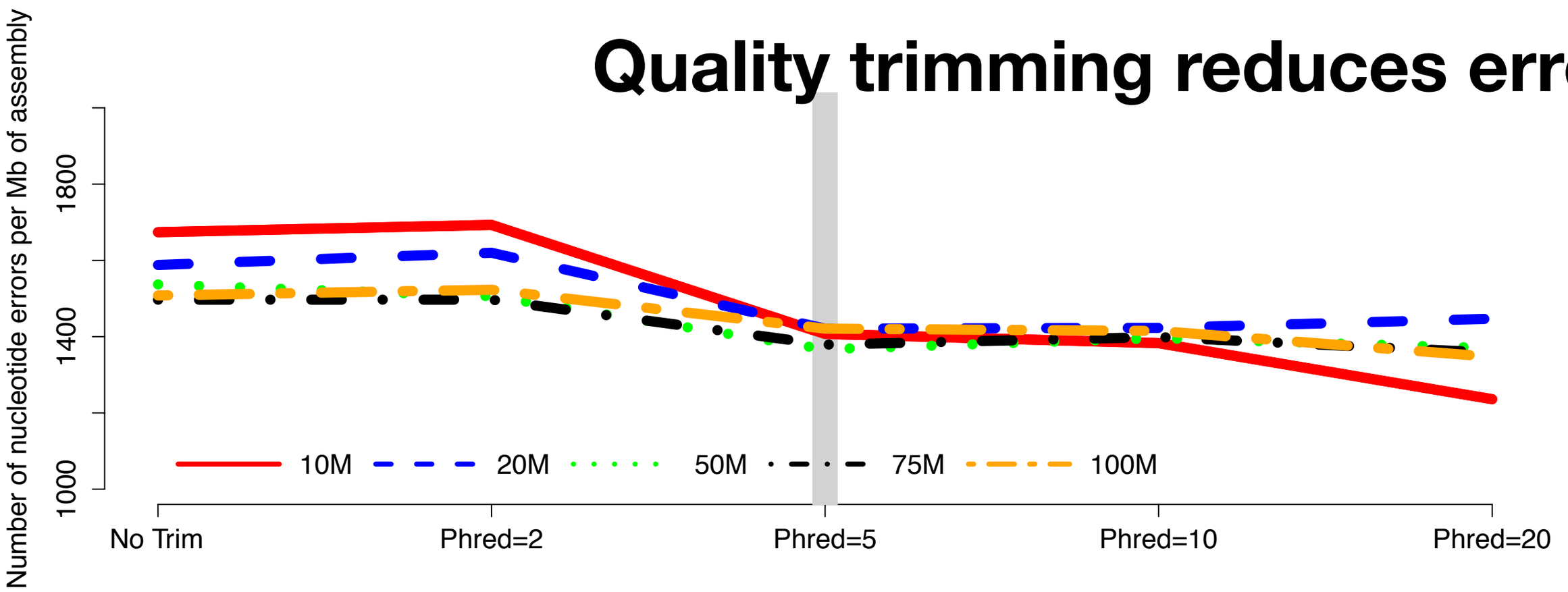


Trimming Experiment

- 2 Illumina datasets > adapter trimmed.
- Subsampled to 10M, 20M, 50M, 75M, 100M PE reads.
- Trimmed at Phred 0,2,5,10,20
- Assembled using Trinity and SOAPdenovo-Trans
- Developed metrics for evaluating transcriptome assemblies.

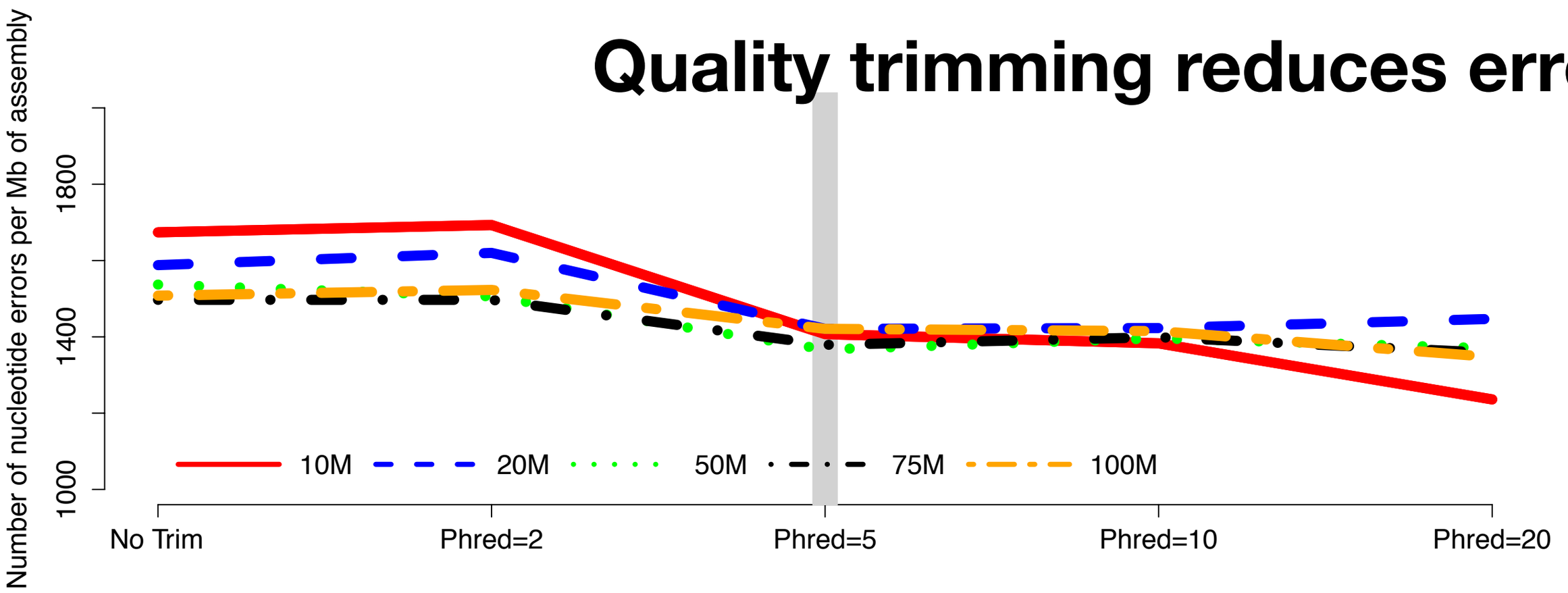
MacManes, Frontiers in Genetics 2014

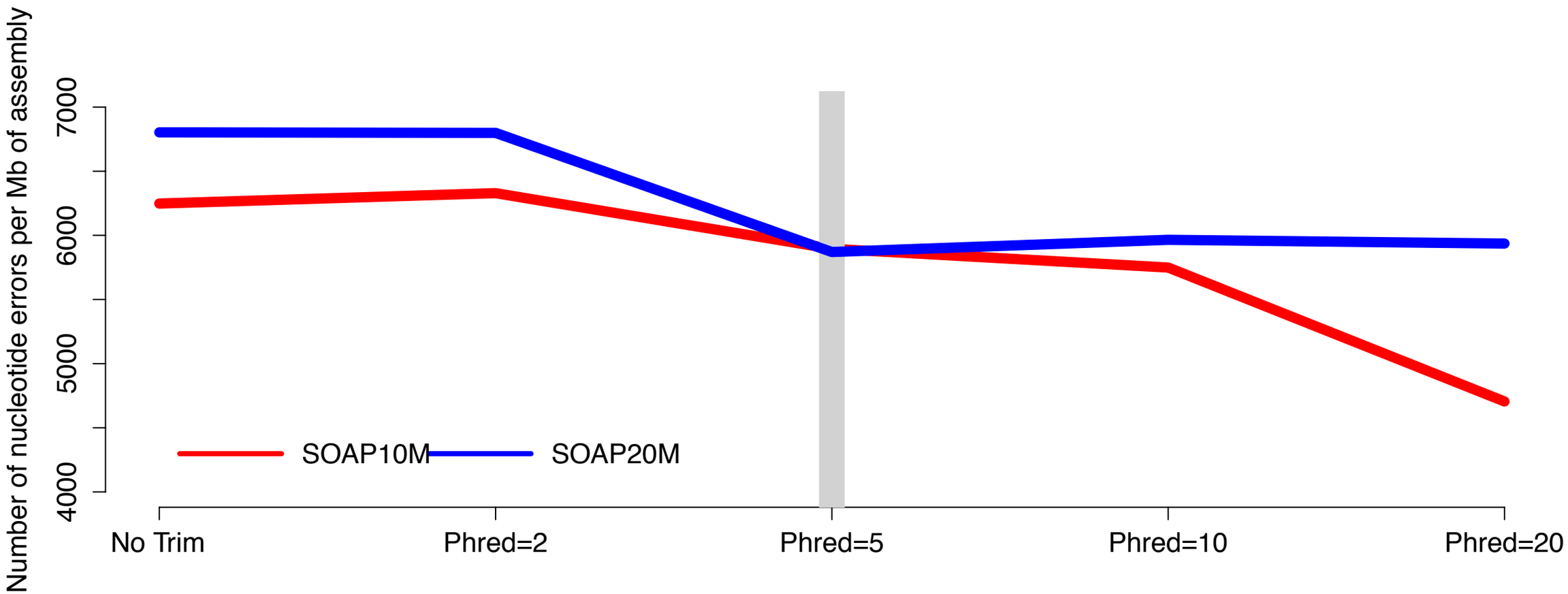
Quality trimming reduces error



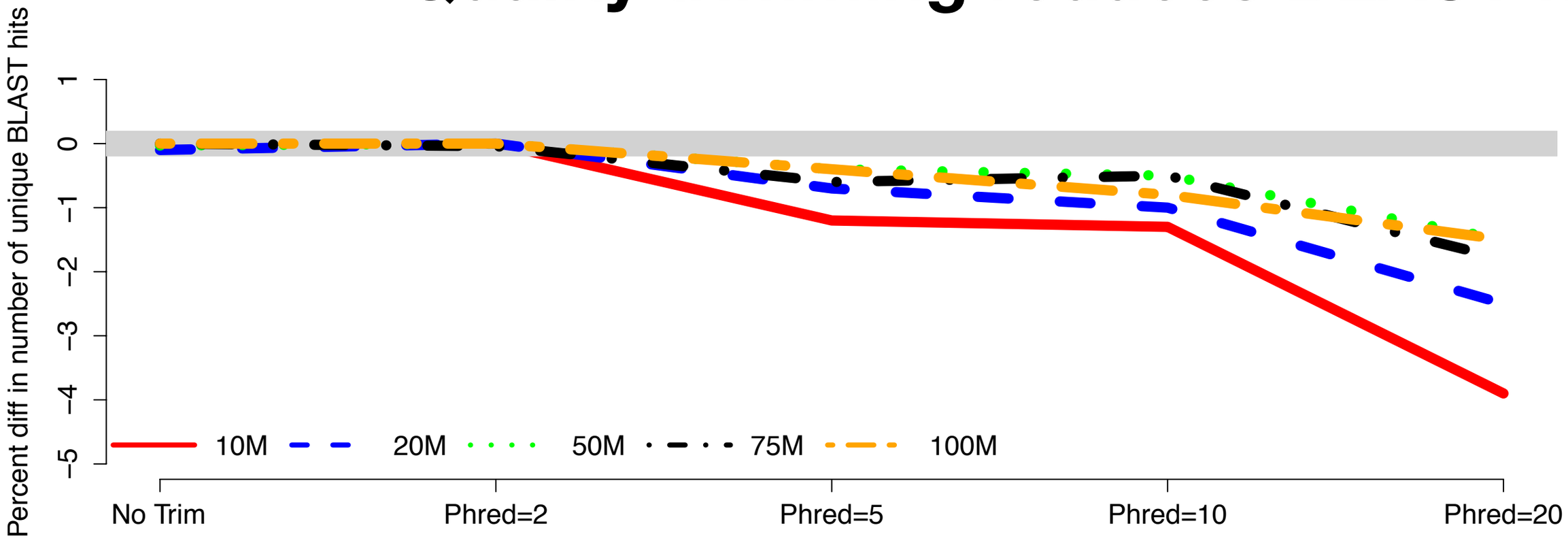
MacManes, Frontiers in Genetics 2014

Quality trimming reduces error



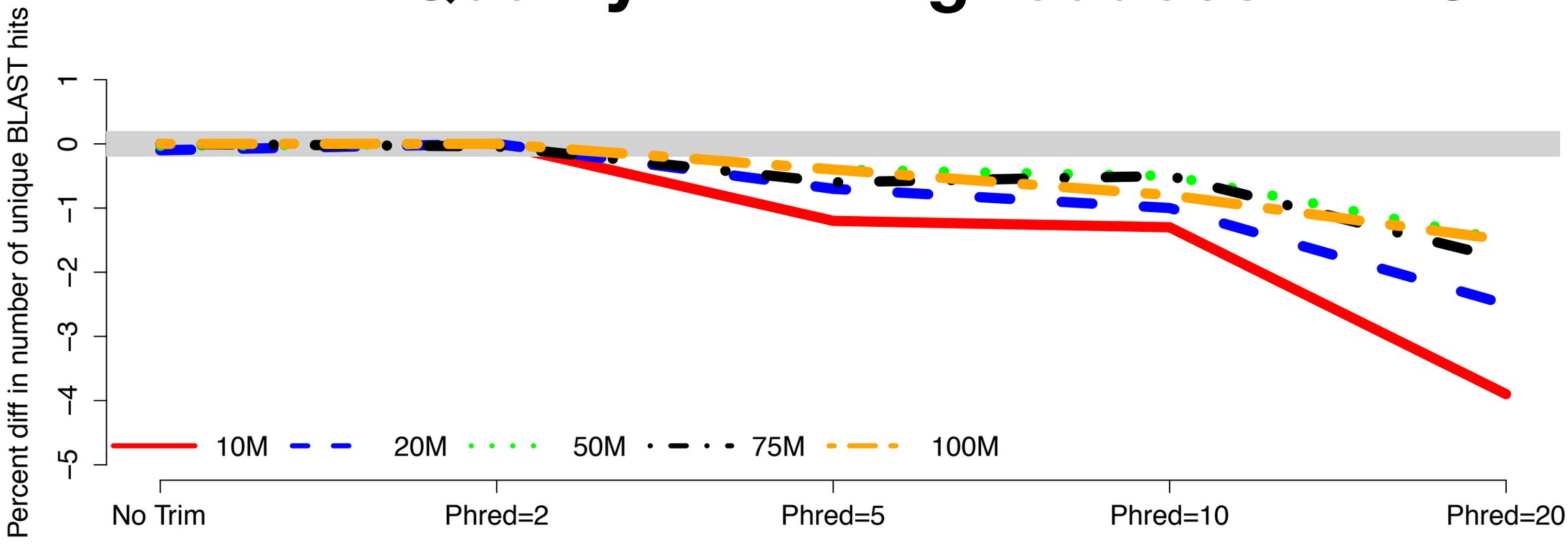


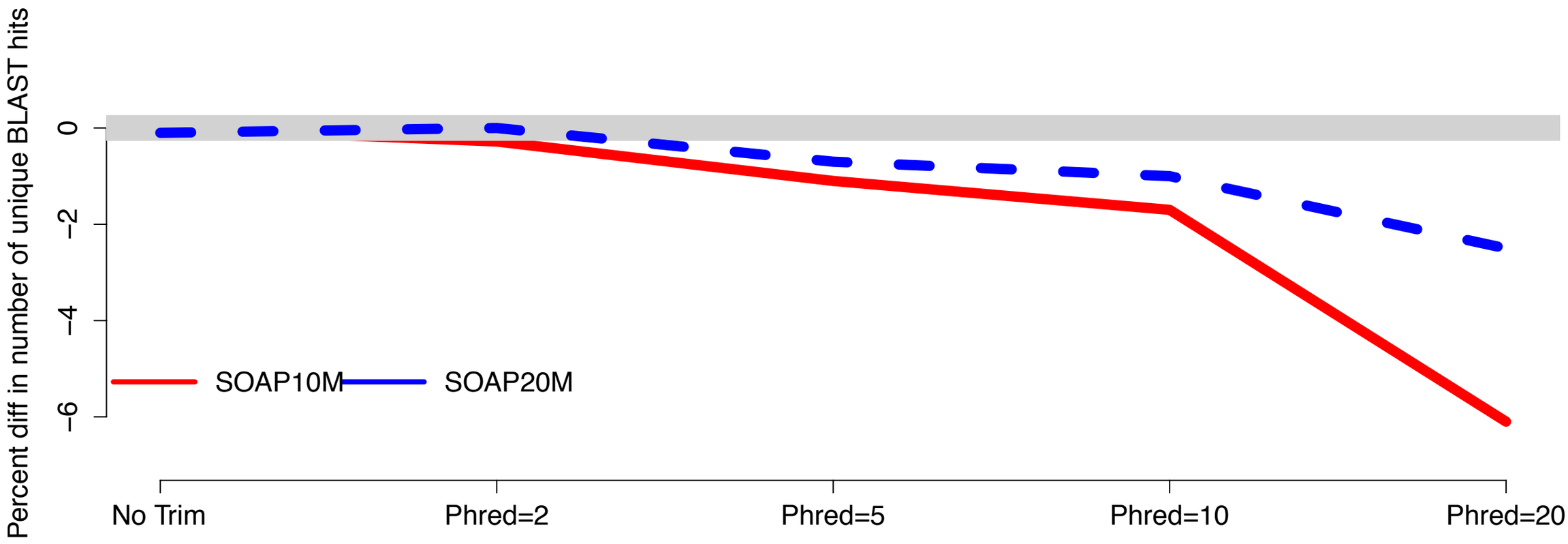
Quality trimming reduces BLAST hits



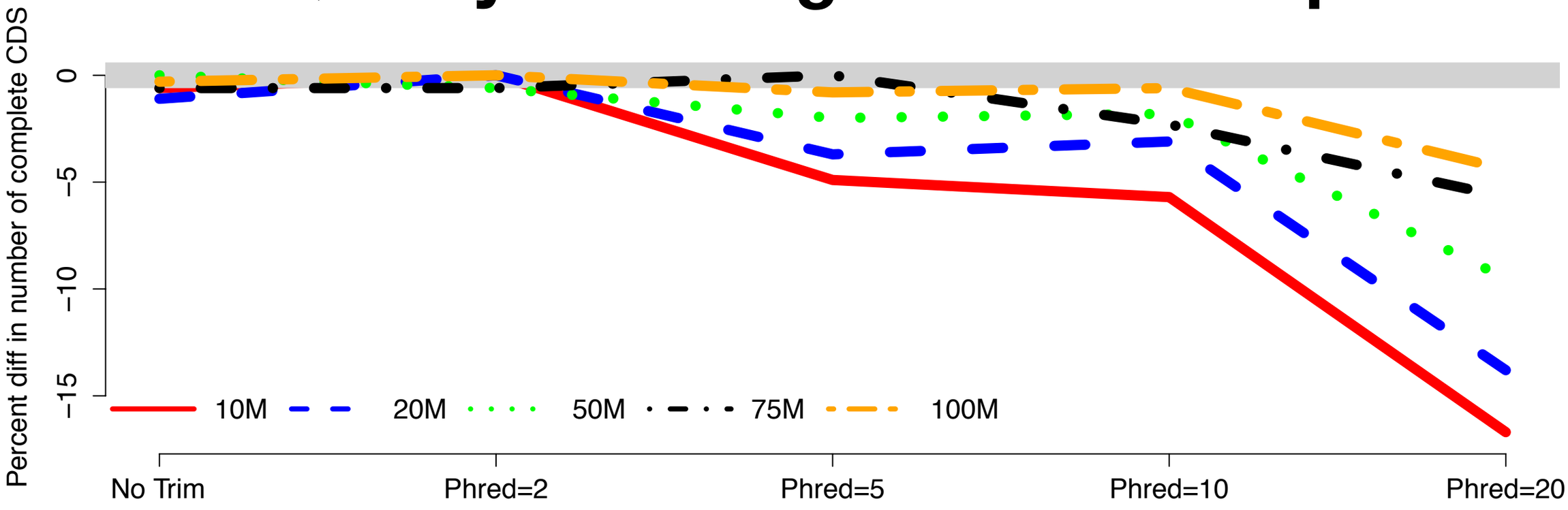
MacManes, Frontiers in Genetics 2014

Quality trimming reduces BLAST hits



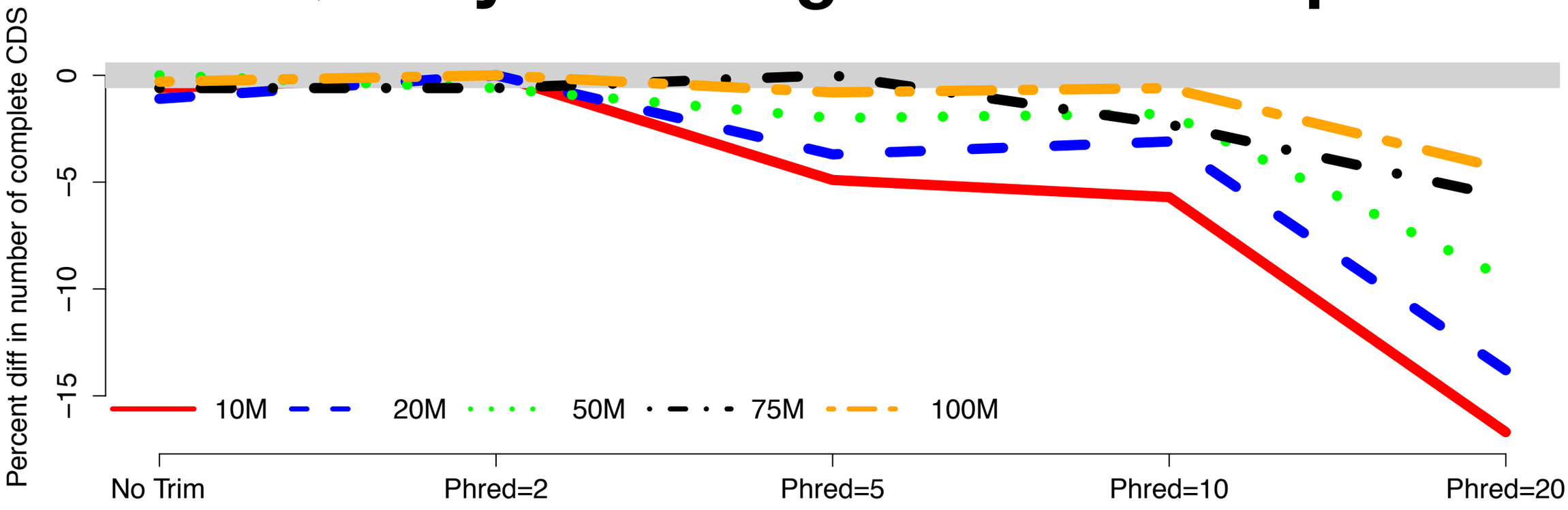


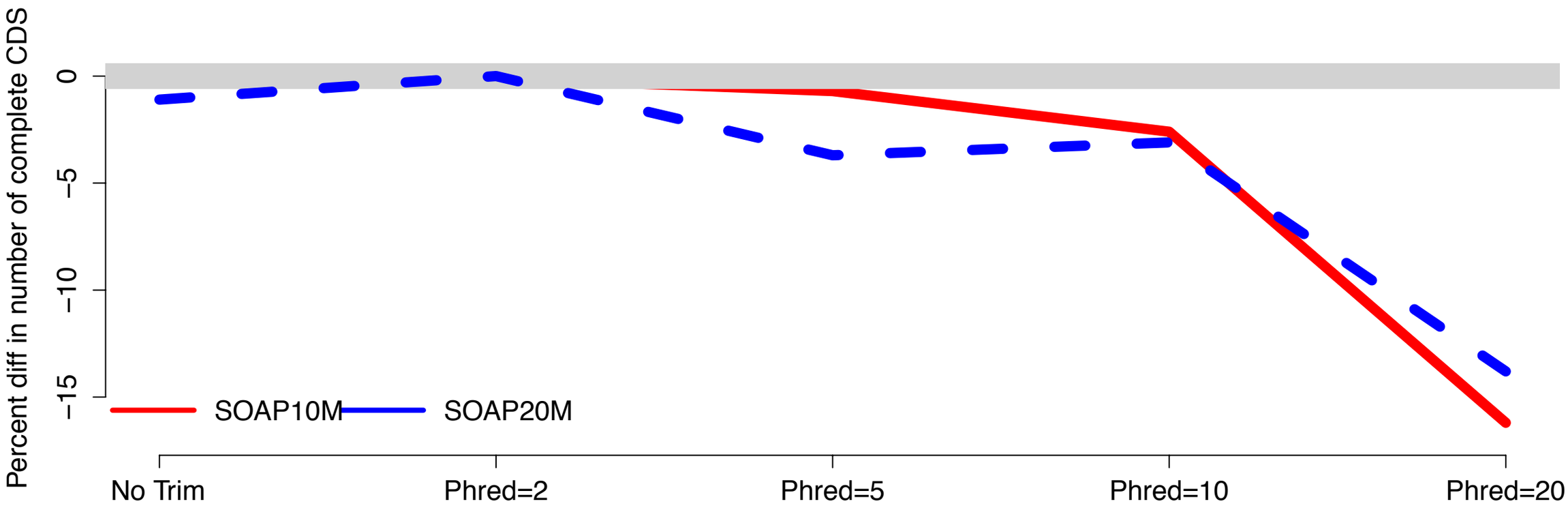
Quality trimming reduces complete CDS



MacManes, Frontiers in Genetics 2014

Quality trimming reduces complete CDS





Summary

- Trimming does reduce assembly error, but at the cost of content & contiguity.
- **<u>Proposed guidelines.</u>**
    1. To max assembly content and contiguity ➠ Trim at 0 or 2ish
    2. If concerned about error ➠ Trim at Phred=5
    3. Usually probably never trim at Phred ≥ 10

MacManes, Frontiers in Genetics 2014

Links

- https://www.youtube.com/watch?v=5NiFibnbE8o
- http://liorpachter.wordpress.com/
  - http://liorpachter.wordpress.com/2014/04/30/estimating-number-of-transcripts-from-rna-seq-measurements-and-why-i-believe-in-paywall/
  - http://liorpachter.wordpress.com/2014/03/01/using-statistical-methods-to-estimate-and-take-into-account-experimental-measurement-errors-a-case-study-using-high-throughput-proteomics-data/
  - http://liorpachter.wordpress.com/2013/10/17/non-uniform-coverage-of-transcripts-in-rna-seq-experiments/

MacManes, Frontiers in Genetics 2014