**Trimming recommendations (conclusions at end)**

* *We currently used adapter trimming, then LEADING:20 TRAILING:20 SLIDINGWINDOW:4:15 MINLEN:20*
  + *= Phread Quality 20 (15 for sliding window)*
  + *= Min length 20*
* **MacManes 2014 (treated as the God reference)**
  + tested PHRED = 0 (adapter trimming only), ≤2, ≤5, ≤10, and ≤20
  + Other parameters constant: MINLEN = 25, ILLUMINACLIP = barcodes.fa:2:40:15, SLIDINGWINDOW size = 4
    - no PHRED given for sliding window, I assume he used the respective PHRED scored for both leading, trailing, and the sliding window
  + 🡪 no big bias difference between PHRED=5 and 20
  + 🡪 less harsh trimming keeps more info at same bias
* LEADING:2 TRAILING:2 SLIDINGWINDOW:4:2 MINLEN:25, as recommended in MacManes
* SLIDINGWINDOW:4:5 LEADING:5 TRAILING:5 MINLEN:25, as recommended in MacManes
* No trimming at all
* No or modest trimming results in the most biologically accurate gene expression estimates
* Increased trimming 🡪 decrease in number of matching paired reads mapped; decrease in number of ORFs that can be identified; decrease in number of distinct transcripts detected through *de novo* assembly; increase in number of false positive variant calls
* Minimum length of 20 reduces bias, bigger no difference
* [**https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3871669/#**](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3871669/)**:**
  + Trimming increases the quality and reliability of the analysis  
    Gene expression analysis: between Q=20 and Q=30 for Trimmomatic  
    SNP identification: loss of noise with any trimmer with a Q threshold equal to or above 20
  + De novo genome assembly: Negative effects for high quality values (e.g. Q>30)
  + 🡪 Trimming is beneficial in RNA-Seq, SNP identification and genome assembly procedures, with the best effects evident for intermediate quality thresholds (Q between 20 and 30)
* **De novo assembly of transcriptome from next-generation sequencing data, Xuan Li**
  + Aggressive quality filtering can result in discarding a substantial portion of sequence data, thus disproportionately affecting some transcripts with biased nucleotide content or lower expression level
  + MacManes and Eisen studies indicate that significant improvement on assembly accuracy was achieved by applying the **error correction process** (will look into that)
  + Stringent trimming of nucleotides with quality scores >20 produced poorer transcriptome assembly
  + Advice to use gentler or no trimming
* **Guidelines for RNA-seq projects: applications and opportunities in non-model decapod crustacean species,** [**Tuan Viet Nguyen**](https://link.springer.com/article/10.1007/s10750-018-3682-0#auth-1)
  + Quality trimming with a PHRED score ranging from 20 to 30 is normal for most RNA-seq experiments
  + PHRED score threshold of 30 or above is usually required for variant calling experiments
  + However, MacManes highlighted that although strict trimming is usually applied, in some cases, a more gentle trimming (PHRED score < 2 or < 5) might be more optimal (because short- and low-expressed transcripts suffer from heavy negative bias when using harsh trimming)
  + Therefore, lowering the PHRED score threshold in the quality-control step can result in a greater transcript discovery rate
  + 🡪 Suggest gentle trimming initially as suggested by MacManes
* [**https://www.biorxiv.org/content/biorxiv/early/2019/04/30/585745.full.pdf**](https://www.biorxiv.org/content/biorxiv/early/2019/04/30/585745.full.pdf)
  + Phred score = 5, read length = 36, based on MacManes
* [**https://www.basepairtech.com/blog/trimming-for-rna-seq-data/**](https://www.basepairtech.com/blog/trimming-for-rna-seq-data/)
  + Our own findings and that of the research community motivate the incorporation of a light amount of trimming in RNA-seq data (we use a Q threshold of 10)

# Trimming of sequence reads alters RNA-Seq gene expression estimates, [Claire R. Williams](https://link.springer.com/article/10.1186/s12859-016-0956-2#auth-1)

* + Before length filtering, multi-hit reads mapped to over 99 % of detected genes, indicating that expression estimates were broadly influenced by short reads aligning to multiple locations (bad)
  + After imposing a minimum read length requirement of 36 bases: only 1.8 % of correctable genes contained any multi-hit reads
  + 🡪 length = 36

# A robust (re-)annotation approach to generate unbiased mapping references for RNA-seq-based analyses of differential expression across closely related species

* + Most datasets had bases with Phred quality score > Q20
  + Therefore, we did not trim these bases but instead relied on the aligner software to make the quality call (following recently published guidelines)
* FastQC evaluation: measured by different metrics, such as the average PHRED error score, GC content biases and position-specific quality variations
* *Final conclusion: trying range from PHRED scored 5 to 20 with read length 25*
  + *LEADING:5 TRAILING:5 SLIDINGWINDOW:4:5 MINLEN:25*
  + *LEADING:10 TRAILING:10 SLIDINGWINDOW:4:10 MINLEN:25*
  + *LEADING:15 TRAILING:15 SLIDINGWINDOW:4:15 MINLEN:25*
  + *LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:25*