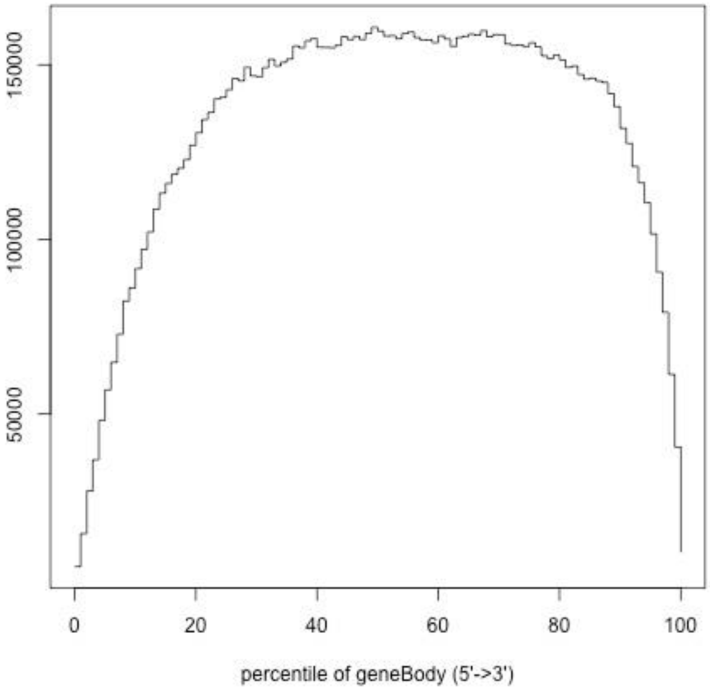
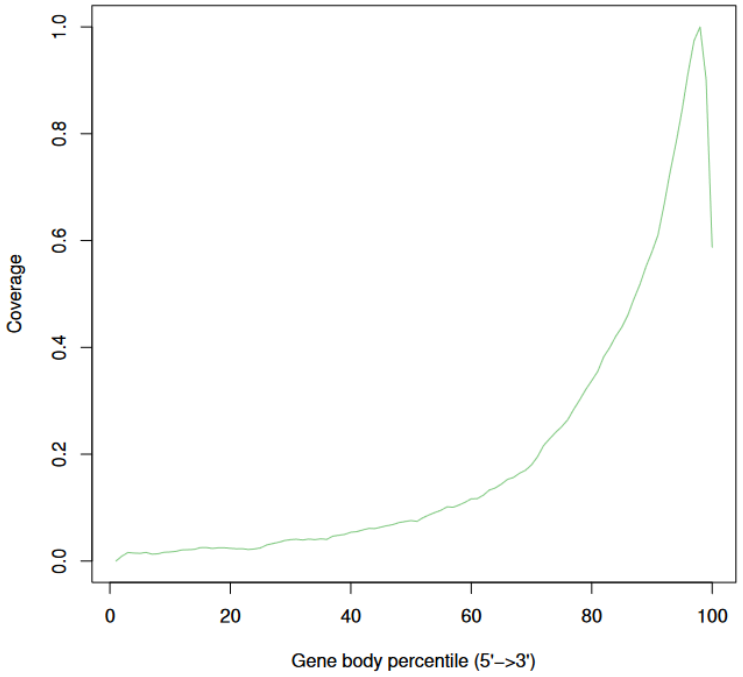
Analysis of
QuantSeq FWD UMI 3’ RNA-seq data

Full length RNA-seq vs 3’ RNA-seq





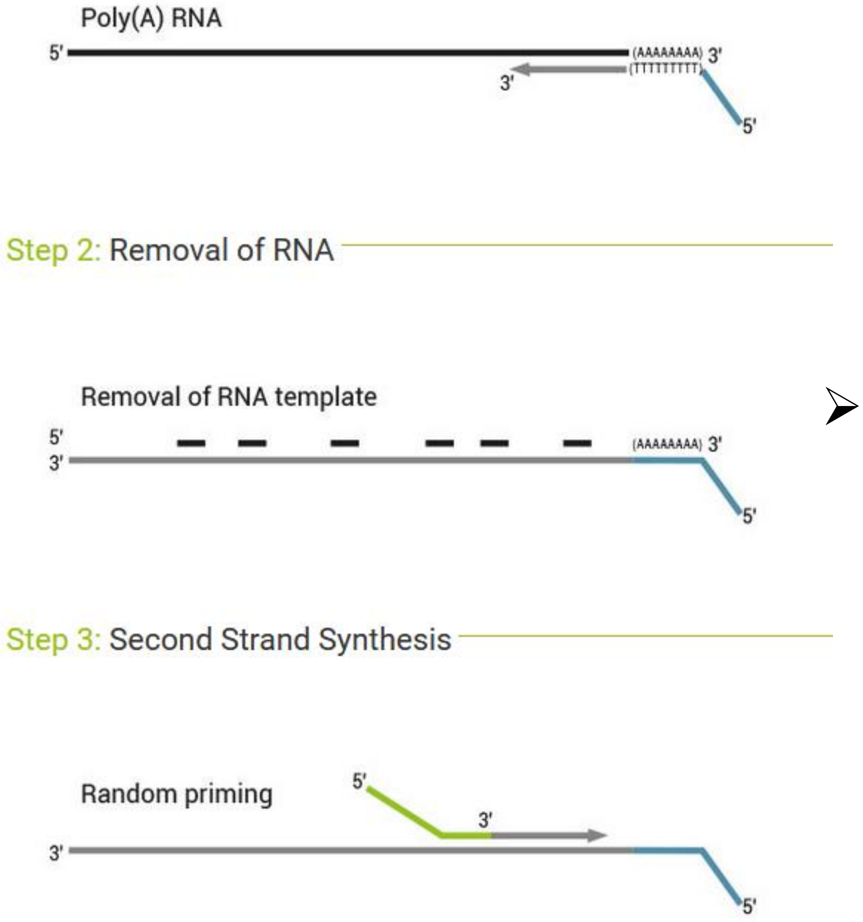
- **Full length RNA-seq: reads cover whole transcripts**
- **3’ RNA-seq: reads cover only the 3’ ends of transcripts**
  - Not possible to detect transcript isoforms
  - Sufficient for gene-level quantitation

QuantSeq 3’ mRNA-seq data

- **Reads come from the 3’ end, near polyA**
  - polyA read-through to adapters is common, needs to be trimmed
- **Just one fragment per transcript is produced**
  - Transcript length does not affect read counts
- **Use only R1 reads**
  - R2 reads start with poly(T) and have low quality
- **Option to remove PCR duplicates using unique molecular identifiers (UMIs)**
  - Useful for low-input and formalin-fixed, paraffin-embedded (FFPE) samples*

*Jang et al (2021) BMC Genomics 22:759

QuantSeq workflow



www.lexogen.com/quantseq-workflow/

- **Reverse transcription**
  - oligodT priming
  - contains the R2 linker, so R2 reads start with poly(T)

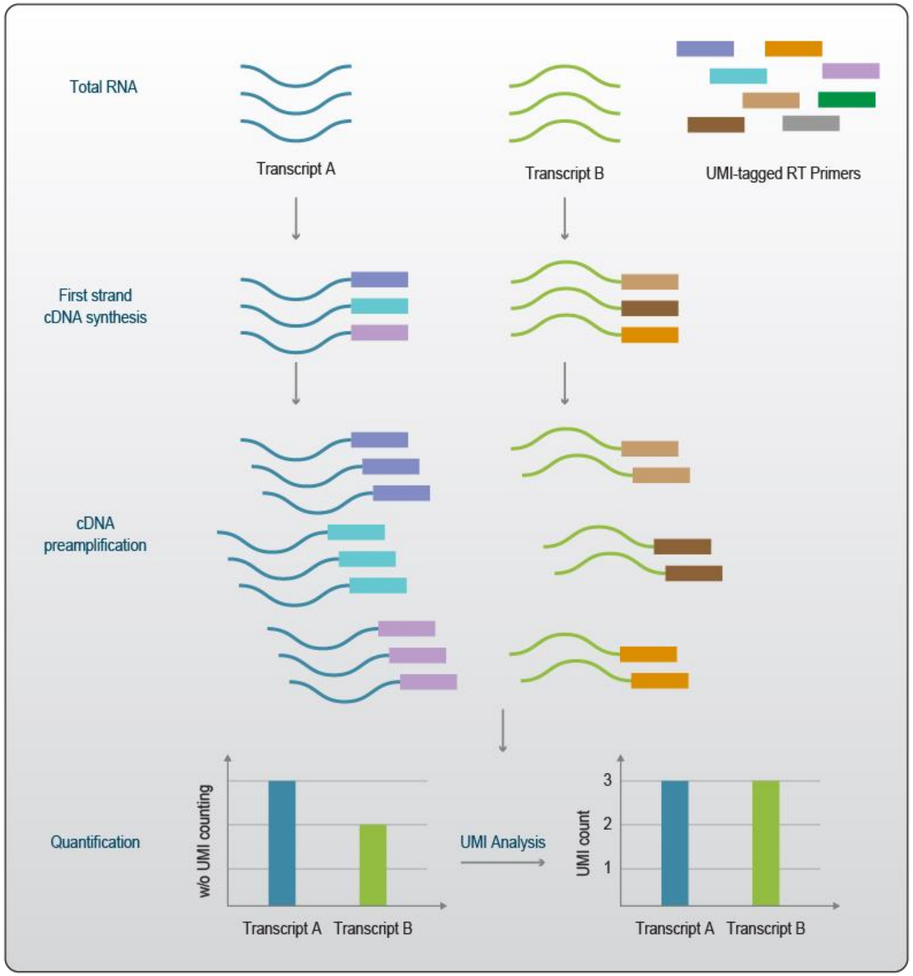
  → R2 reads have low quality

  → use only R1 reads
- **Second strand synthesis with random priming**
  - Errors in first nucleotides due to non-specific hybridization of the random primer to the cDNA template.

  → Use aligner with soft-clipping

Unique molecular identifiers (UMIs)

- **Reduce amplification bias → more accurate quantitation**



www.lexogen.com/rna-lexicon-what-are-unique-molecular-identifiers-umis-and-why-do-we-need-them/

QuantSeq FWD UMI 3’ mRNA-seq data

- **Reads contain 6 nt unique molecular identifiers (UMIs)**
- **Located at the start of the read, need to be removed and stored in the read name before alignment to genome**
- **Deduplication: Reads which map to the same genomic location and have the same UMI are grouped together, and only one representative read is kept**

QuantSeq 3’ mRNA-seq data analysis steps

- **Quality control / Read quality with MultiQC** → html report
- **Preprocessing / Extract UMIs from QuantSeq reads** → FASTQ
- **Preprocessing / Trim QuantSeq reads with BBDuk** → FASTQ
- **Alignment / STAR or HISAT2** → BAM
- **Preprocessing / Deduplicate aligned QuantSeq reads** → BAM
- **Quality control / RNA-seq quality metrics with RseQC** → pdf
- **RNA-seq / Count aligned reads per genes with HTSeq** → tsv
- **Utilities / Define NGS experiment** → tsv
- **Quality control / PCA and heatmap of samples with DESeq2** → pdf
- **RNA-seq / Differential expression using DESeq2** → tsv
- **Utilities / Annotate Ensembl identifiers** → tsv

UMIs from sequence to read names

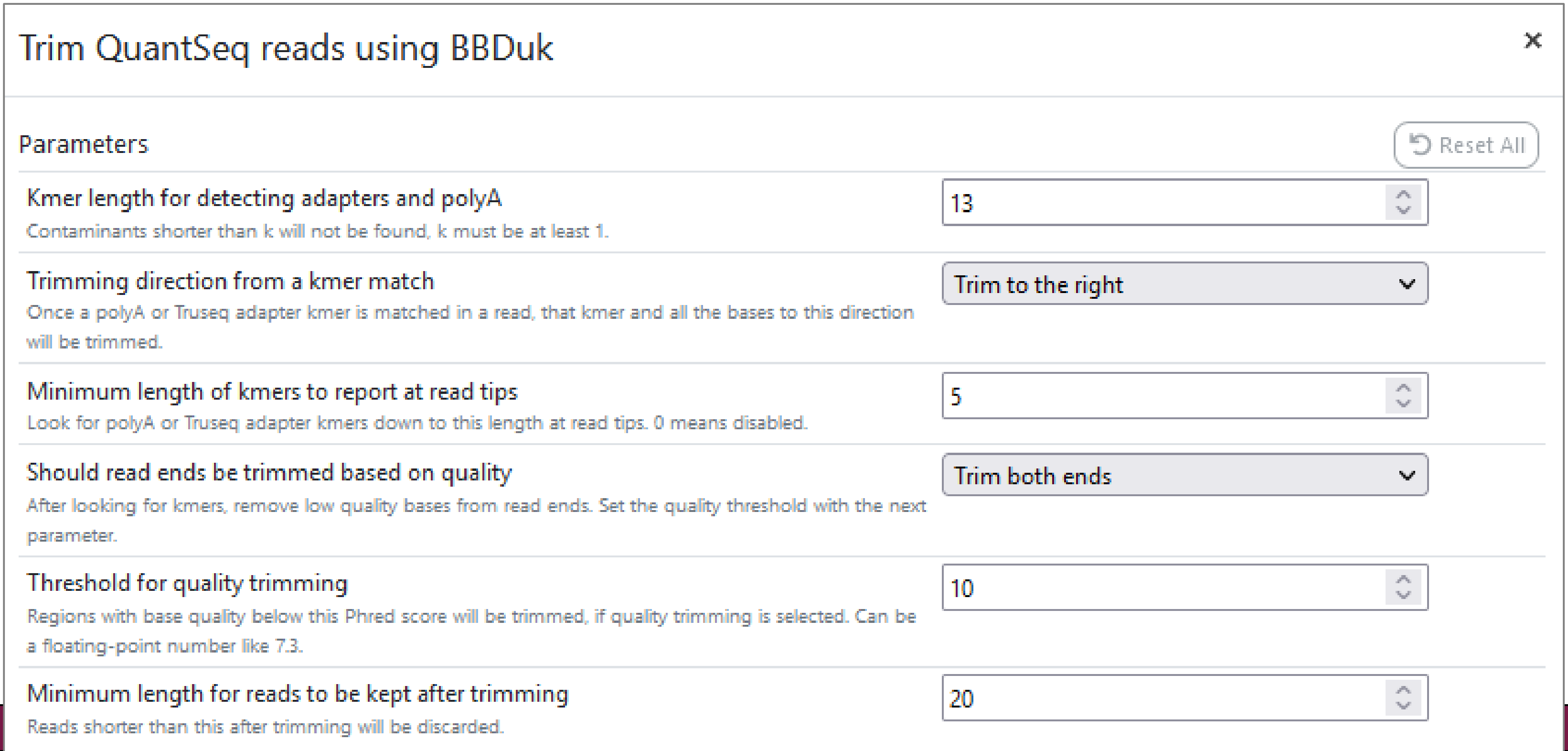
- **Use the tool Preprocessing / Extract UMIs from QuantSeq reads**
  - Extracts the 6-base UMI and stores it in read's name
  - Removes the TATA spacer in position 7-10
  - log.txt file tells how many reads contained TATA and were processed
  - Based on the extract tool of the UMI-Tools package

```
@A00464:250:HW3NWDRXX:1:2236:5565:26412 1:N:0:GCATGG+CTAACT
AGCGGGTATAGTCTGTGAGTGCAACGTGTAAACAGCTGAGCCAAGAACTAATGGAAAAAAAAAAAAAAAAAAAGATCGGAAGAGCACACGTCTGAACTCCA
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,:FFFFFF:FFFF:FFF::FFF:,F:FFFFFFFFFF:FFFFFFFFFFFFFFF
@A00464:250:HW3NWDRXX:1:1173:21947:24158 1:N:0:GCATGG+CTAACT
CAGGGCTATAGGAAGAGCACACGTCTGAACTCCAGTCACGCATGGATCTCGTATGCCGTCTTCTGCTTGAAACTTGGGGGGGGGGGGGGGGGGGGGGGGGG
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,FF:FFFFFFFFFFFFFF:FFFF:,,,,FFFFFFFFFFFFFFFFFFFFFFFFF
```

```
@A00464:250:HW3NWDRXX:1:2236:5565:26412_AGCGGG 1:N:0:GCATGG+CTAACT
GTCTGTGAGTGCAACGTGTAAACAGCTGAGCCAAGAACTAATGGAAAAAAAAAAAAAAAAAAAGATCGGAAGAGCACACGTCTGAACTCCA
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,:FFFFFF:FFFF:FFF::FFF:,F:FFFFFFFFFF:FFFFFFFFFFFFFFF
@A00464:250:HW3NWDRXX:1:1173:21947:24158_CAGGGC 1:N:0:GCATGG+CTAACT
GGAAGAGCACACGTCTGAACTCCAGTCACGCATGGATCTCGTATGCCGTCTTCTGCTTGAAACTTGGGGGGGGGGGGGGGGGGGGGGGGGG
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,FF:FFFFFFFFFFFFFF:FFFF:,,,,FFFFFFFFFFFFFFFFFFFFFFFFF
```

Trim polyA, adapters and low quality ends

- **Tool Preprocessing / Trim QuantSeq reads with BBDuk**
  - Detects and removes polyA tails and Illumina TruSeq adapters
  - Trims low-quality bases from read ends
  - Removes reads that are too short after trimming
  - Based on the BBDuk tool of the BBTools package



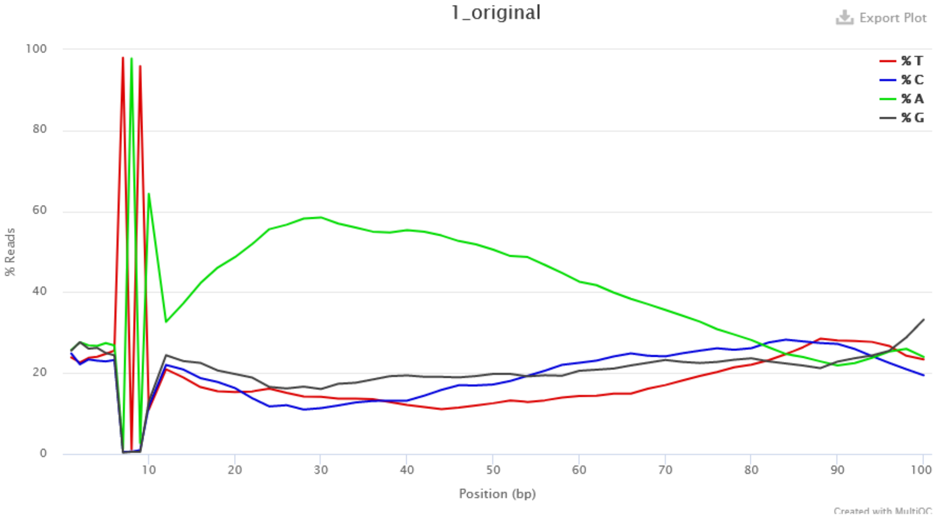
Trim QuantSeq reads using BBDuk

Parameters

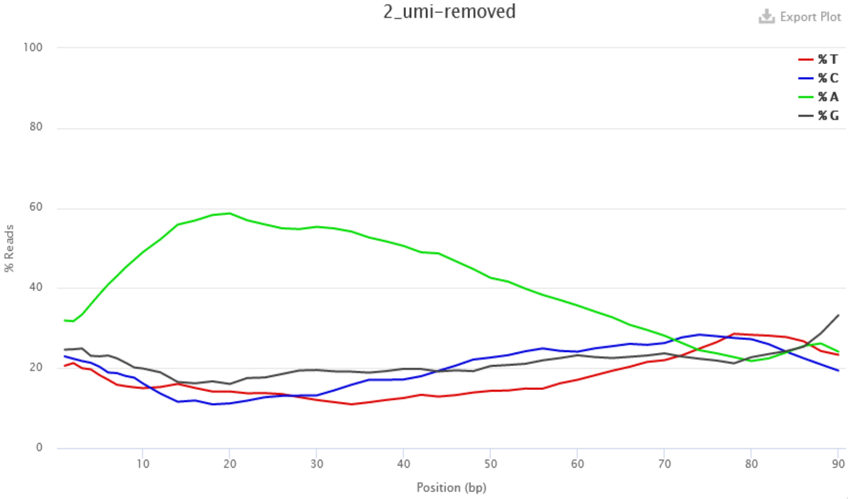
Kmer length for detecting adapters and polyA Contaminants shorter than k will not be found. k must be at least 1.	13
Trimming direction from a kmer match Once a polyA or Truseq adapter kmer is matched in a read, that kmer and all the bases to this direction will be trimmed.	Trim to the right
Minimum length of kmers to report at read tips Look for polyA or Truseq adapter kmers down to this length at read tips. 0 means disabled.	5
Should read ends be trimmed based on quality After looking for kmers, remove low quality bases from read ends. Set the quality threshold with the next parameter.	Trim both ends
Threshold for quality trimming Regions with base quality below this Phred score will be trimmed, if quality trimming is selected. Can be a floating-point number like 7.3.	10
Minimum length for reads to be kept after trimming Reads shorter than this after trimming will be discarded.	20

Base composition plot

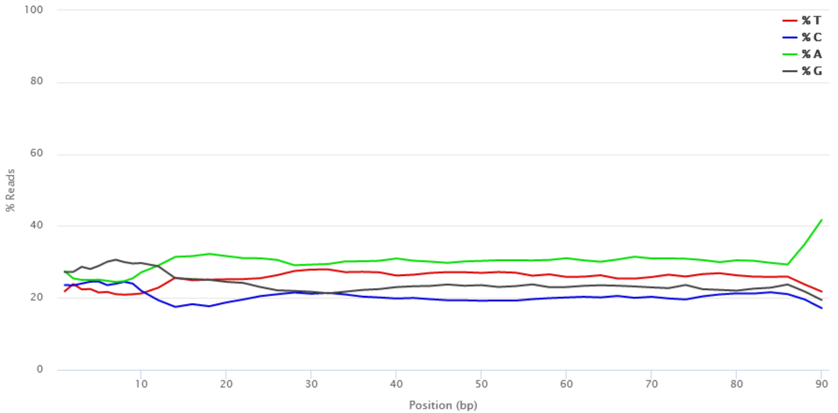
- ➢ **Raw reads**



- ➢ **After extracting UMIs and TATA**



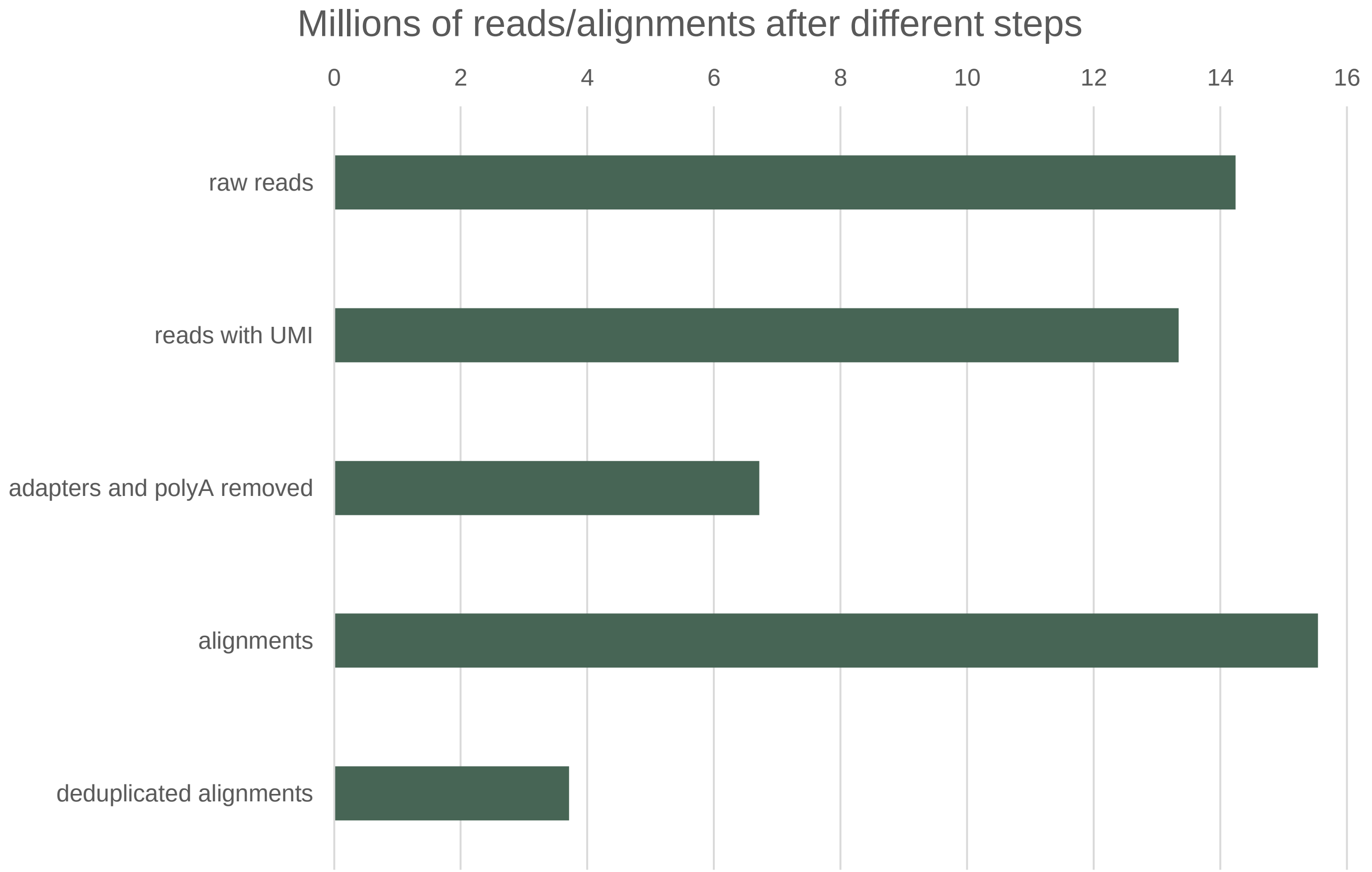
- ➢ **After trimming polyA and adapters**



Remove amplification bias using UMIs

- **Tool Preprocessing / Deduplicate aligned QuantSeq reads**
  - Identifies which reads have the same mapping position.
  - Groups those reads which have the same/similar UMI.
  - Two grouping methods
    - **Unique:** reads must have exactly the same UMI sequence. Fast but doesn't allow for sequencing errors.
    - **Directional:** builds networks where nodes are UMIs and edges connect UMIs with an edit distance </= 1. Identifies clusters of UMIs. Slow, allows for errors.
  - Keeps a single representative read
    - lowest number of mapping coordinates
    - highest mapping quality. Note that base quality is not considered
  - Output is a deduplicated BAM file and optional statistics files
    - average edit distance between the UMIs at each position
    - counts for unique combinations of UMI and position
    - UMI-level summary statistics
  - Based on the dedup tool of the UMI-Tools package
    - https://umi-tools.readthedocs.io/en/latest/reference/dedup.html

Adapter/polyA removal and deduplication reduce the number of reads



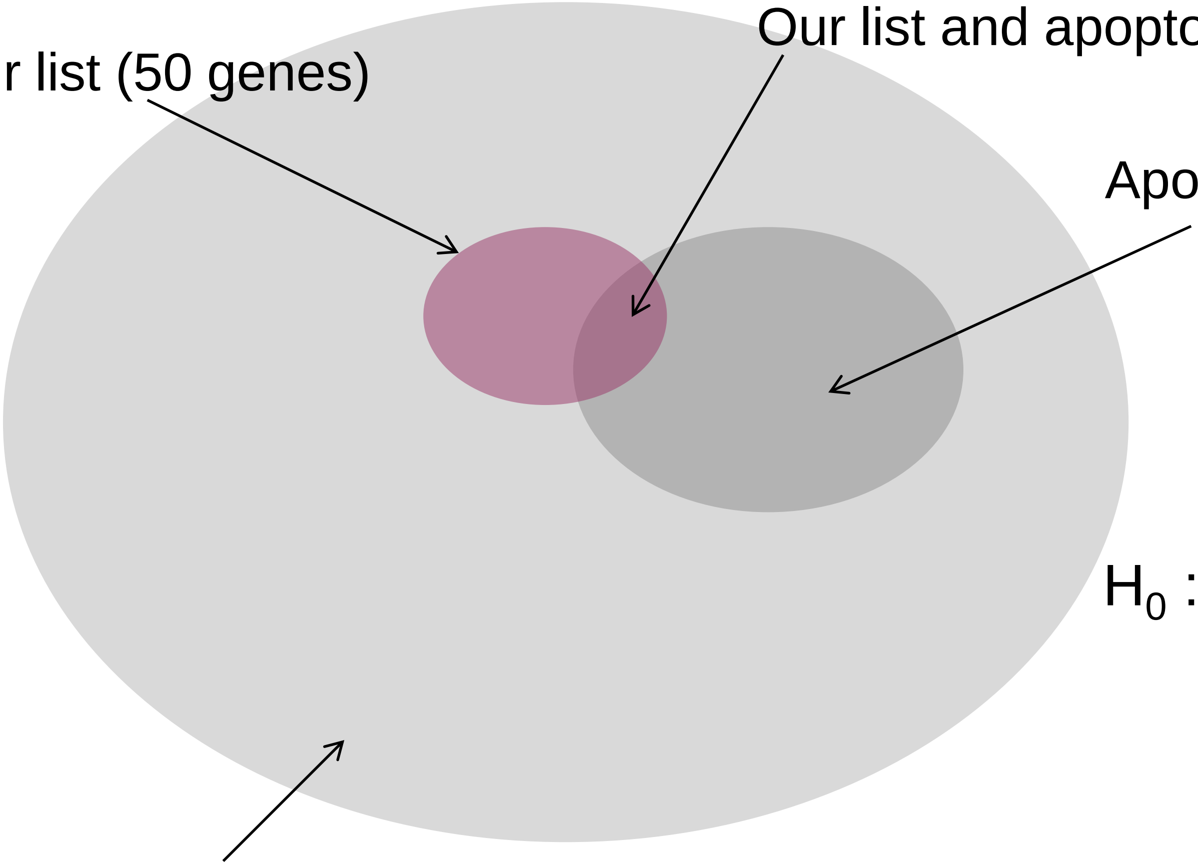
Pathway analysis

Pathway analysis – why?

- **Statistical tests can yield thousands of differentially expressed genes**
- **It is difficult to make ”biological” sense out of the result list**
- **Looking at the bigger picture can be helpful, e.g. which pathways are differentially expressed between the experimental groups**
- **Databases such as KEGG, GO, Reactome and ConsensusPathDB provide grouping of genes to pathways, biological processes, molecular functions, etc**

Gene set enrichment analysis

1. **Perform a statistical test to find differentially expressed genes**
2. **Check if the list of differentially expressed genes is ”enriched” for some pathways**



$$H_0 : \frac{\frac{10}{50}}{\frac{200}{30000}} = 30 >> 1$$

ConsensusPathDB

- **One-stop shop: Integrates pathway information from 32 databases covering**
  - biochemical pathways
  - protein-protein, genetic, metabolic, signaling, gene regulatory and drug-target interactions

- **Developed by Ralf Herwig's group at the Max-Planck Institute in Berlin**

- **ConsensusPathDB over-representation analysis tool is integrated in Chipster**
  - runs on the MPI server in Berlin