# Impact of NGS data trimming on differential gene expression analysis in two groups of bees

Mateusz Kołomański[1], Magdalena Frąszczak[1], Magda Mielczarek[1,2]

[1] Biostatistics group, Department of Genetics,
Wroclaw University of Environmental and Life Sciences

[2] National Research Institute of Animal Production

## Objectives

**Determine the impact of NGS data trimming on the results of differential gene expression analysis.**
**Establish the best practices in regards to NGS data handling and filtering.**

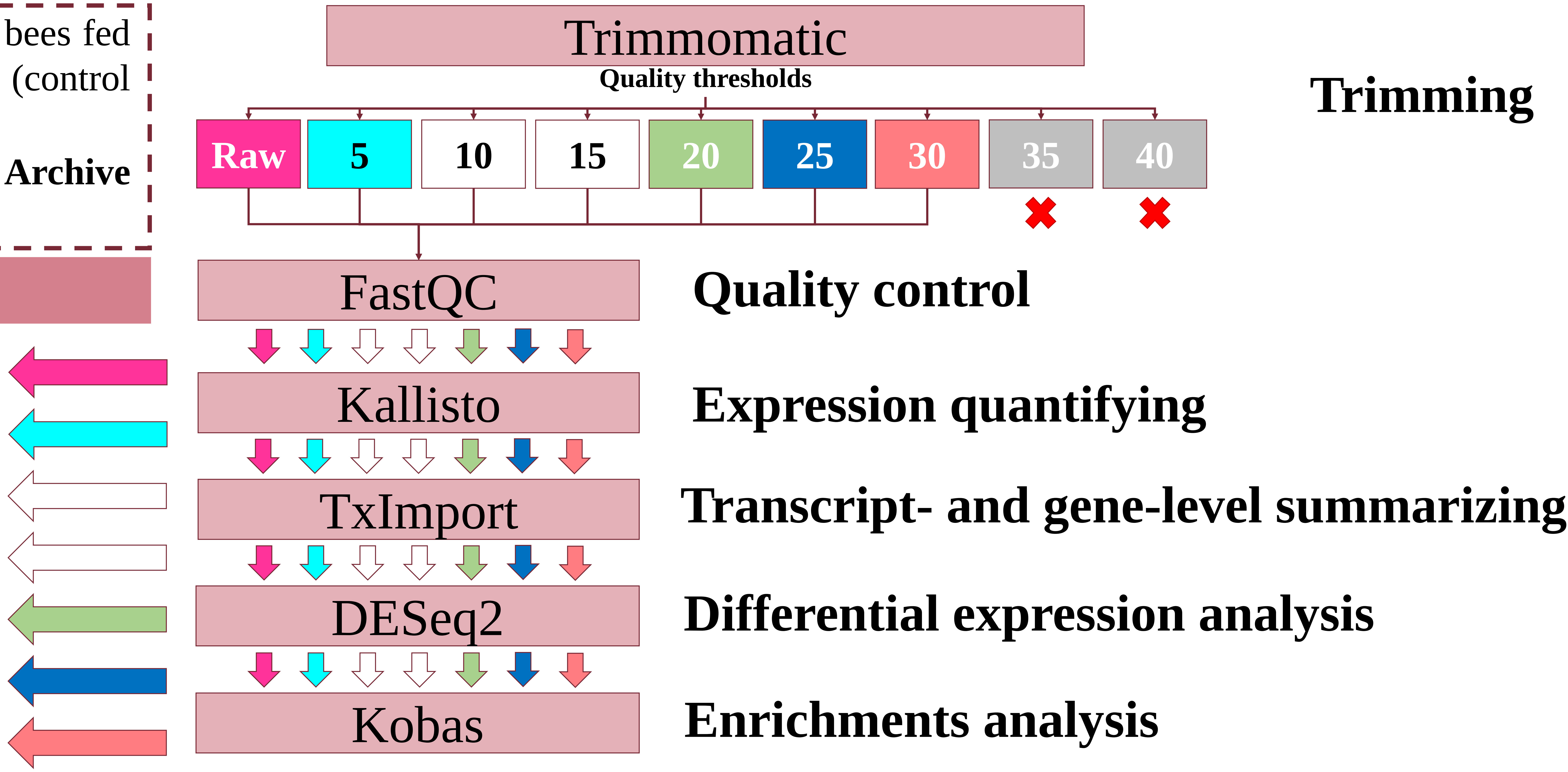
## Dataset

- RNAseq data of two groups of 8-days-old bees fed honey (test group) and honey with pollen (control group);
- Data aquired from **European Nucleotide Archive** (PRJNA175445).

## Statistical analysis

- **Spearman correlation** and **regression** for number of reads that survived;
- **Venn diagrams** for differentialy expressed transcripts and genes;
- **Q-Cochran** test for statistically significant KEGG pathways.

## Bioinformatics Pipeline



Trimmomatic

Quality thresholds

Raw 5 10 15 20 25 30 35 40

✖ ✖

Trimming

FastQC

Quality control

Kallisto

Expression quantifying

TxImport

Transcript- and gene-level summarizing

DESeq2

Differential expression analysis

Kobas

Enrichments analysis

# Results

## 1. Sequence number analysis

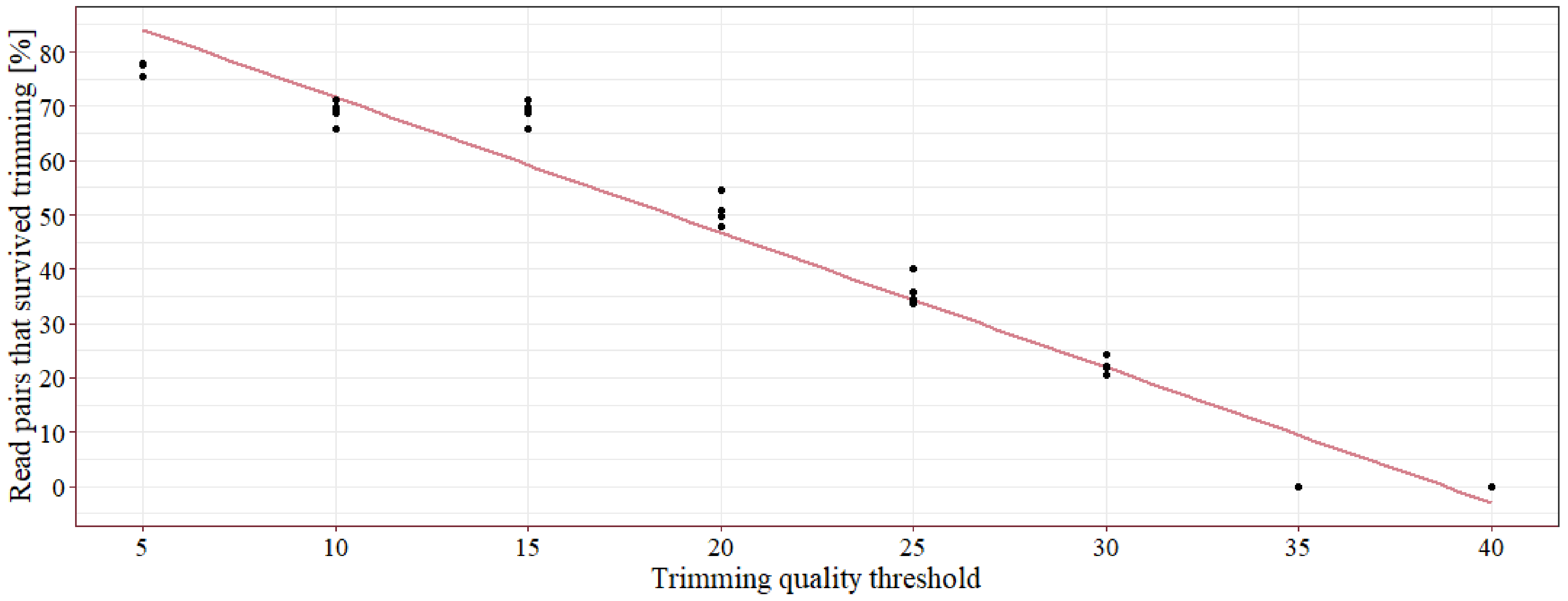


Figure 1. Spearman test showed significant negative correlation between trimming quality threshold and percentage of read pairs that survived trimming (p-value = 2.58595 * $10^{-32}$, R = -0.9765135).

## 2. Impact on differentialy expressed transcripts and genes

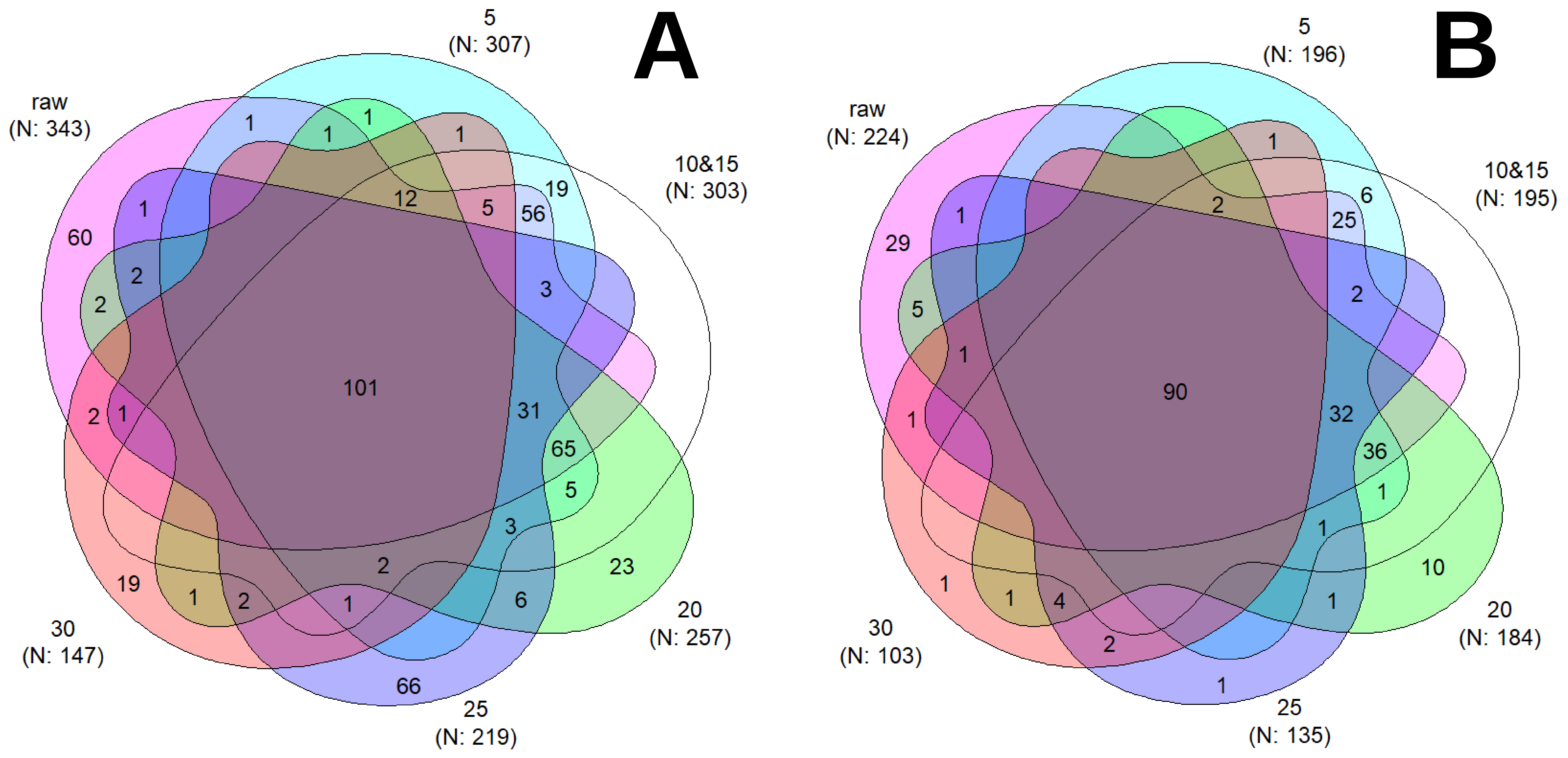


Figure 2. Transcripts (A) and genes (B) detected as having a significant ($p < 0.05$) differences between two groups of bees (test and control) across datasets with different quality thresholds. Most datasets consist of unique transcripts and genes.

## 3. KEGG Pathway analysis

- KEGG pathways analysis revealed between 2 and 6 statistically significant pathways;
- Varying trimming quality thresholds result in significant differences in KEGG pathways (p-value = 0.00143);
- Generally, sets that were trimmed more strictly showed higher number of statistically significant pathways.

Table 1. Significant KEGG pathways

| TRANSCRIPT-LEVEL | | | | |
|---|---|---|---|---|
| QT\Pathway | ame03010 | ame04146 | ame04213 | ame04141 |
| Raw | + | + | | |
| 5 | + | + | | |
| 10 & 15 | + | + | | |
| 20 | + | + | + | |
| 25 | + | + | | |
| 30 | + | + | + | + |

| GENE-LEVEL | | | | | | | |
|---|---|---|---|---|---|---|---|
| QT\Pathway | ame03010 | ame04146 | ame04213 | ame00603 | ame00790 | ame00190 | ame04141 |
| Raw | + | + | + | | | | |
| 5 | + | + | + | | | | |
| 10 & 15 | + | + | + | + | + | | |
| 20 | + | + | + | | + | + | + |
| 25 | + | + | + | | + | | + |
| 30 | + | + | + | | + | | + |

# Conclusions

- NGS data filtering impacts the differential expression analysis on gene- and transcript-levels.
- The threshold quality value impacts not only the amount of available data for analysis, but also particular transcripts/genes that are detected in differential gene expression analysis, as well as significant KEGG pathways, which might lead to different biological conclusions.
- More studies need to be done on the impact of NGS data trimming on analysis using data from various model organisms and artificially generated datasets.

## Acknowledgements

This project was supported by Poznań Supercompouting and Networking Centre