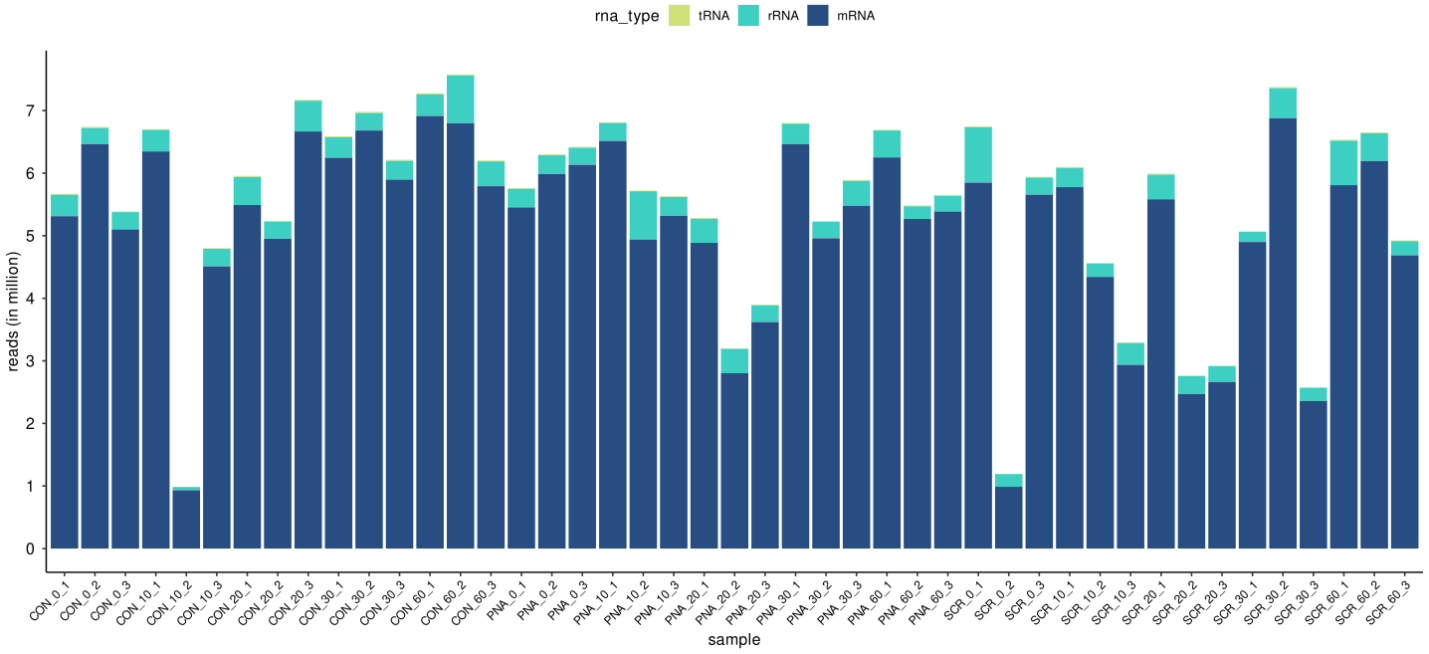
Prevotella project RNA-Seq data analysis

1. **Samples used for RNA-Seq**

- We treated *P. copri* with our PNA/control samples. We have 45 samples in total:
  - 5 timepoints (0,10,20,30,60 min) → 5
  - Three conditions: KFF-*acpP* , KFF-Scrambled, PNA and water control → 5*3=15
  - All with biological triplicates: 15*3 = **45 samples**

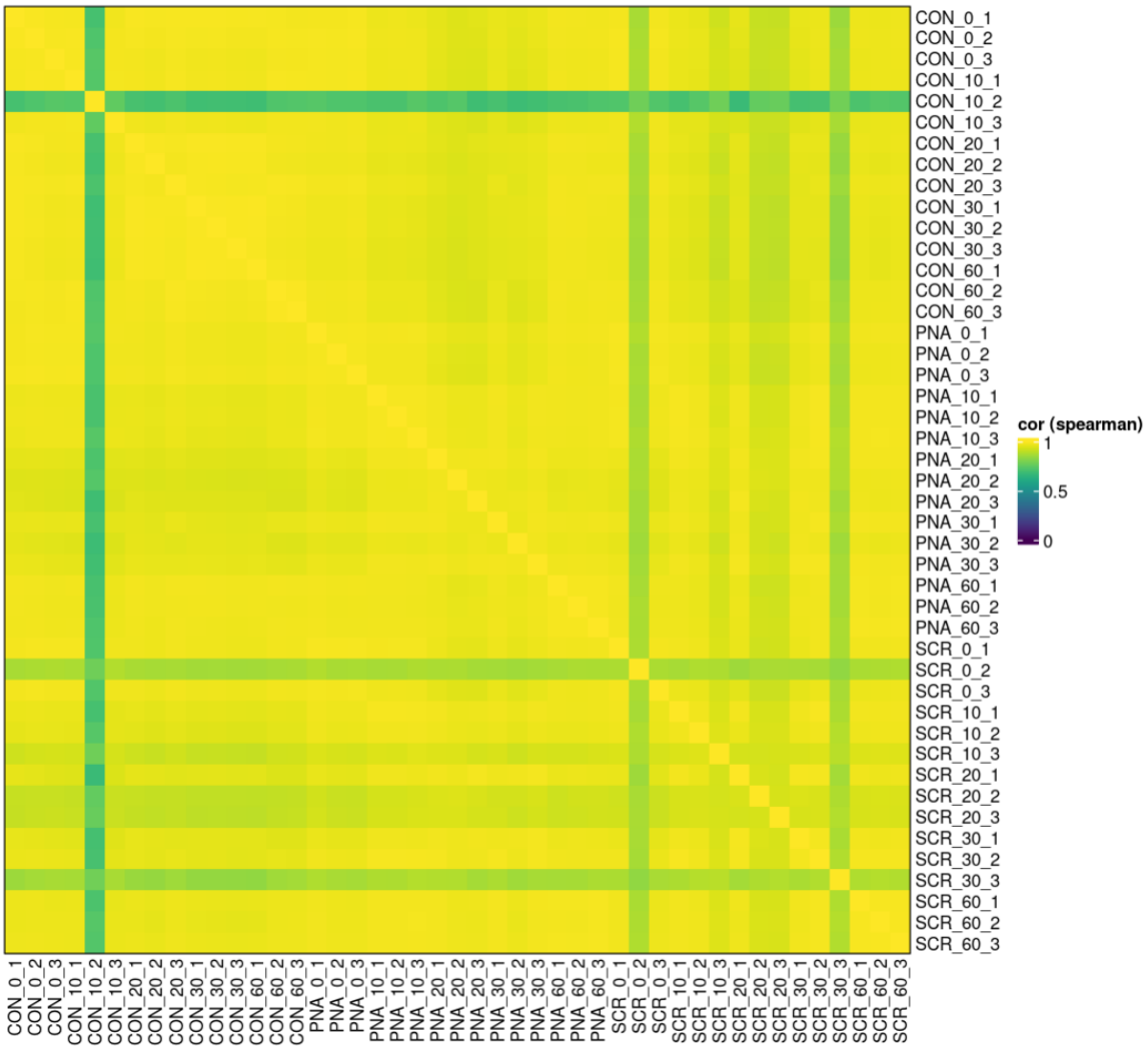
2. **Mapping statistics**

- I ran an initial analysis and mapped/counted the reads. The mapped read counts, sorted by RNA type look like this:



- Overall, it looks very good as most samples have 3-7 mio reads mapping to *Prevotella* mRNA, and the rRNA depletion seems to have worked very well.

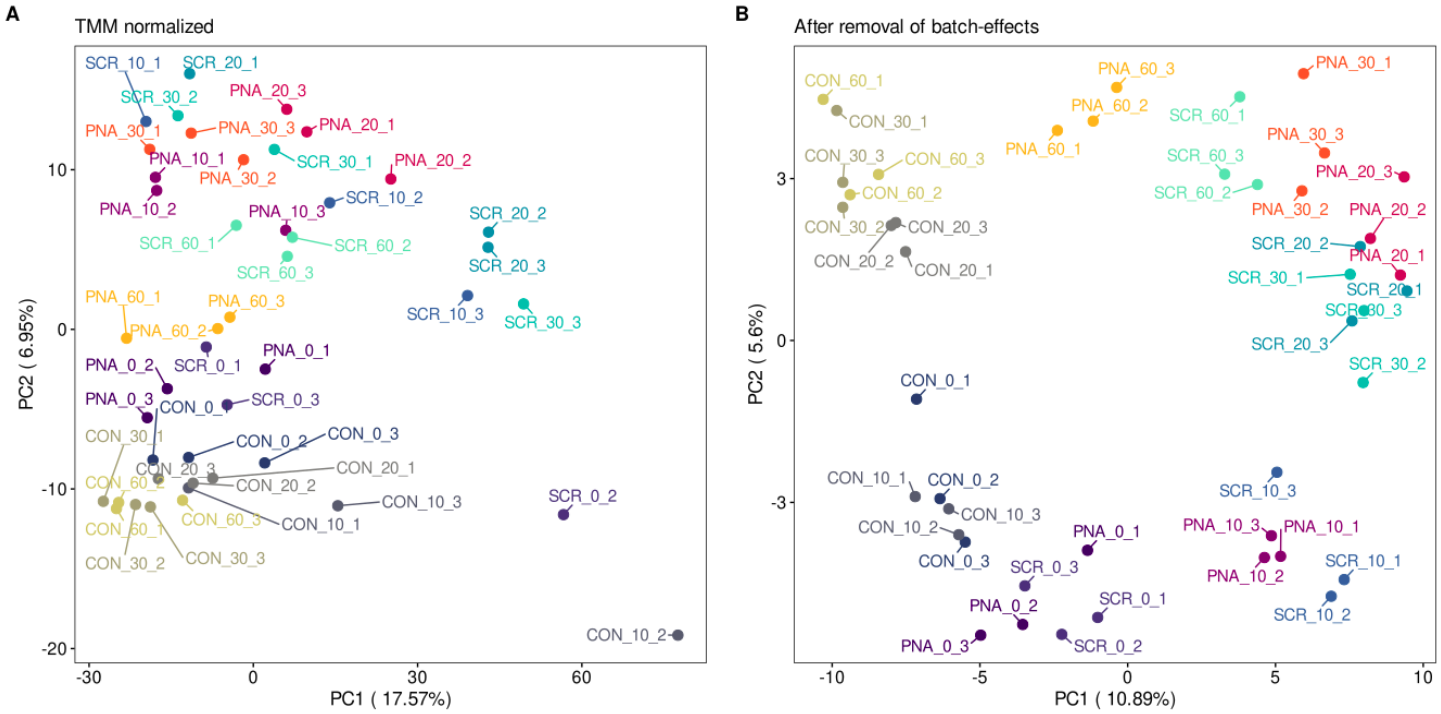
- CON_10_2 only has around 1 million mapped reads, and the correlation between this sample and others is not that great (see plot below, closer to 1 → better correlation). Also, SCR_0_2 has few reads and rather low correlation (plot shows pearson correlation of normalized reads). Did you expect this/ does it make sense that there were a few with lower mRNA? Anyway, it is not a huge problem, as we have triplicates, and we can account for this in the Normalization step.



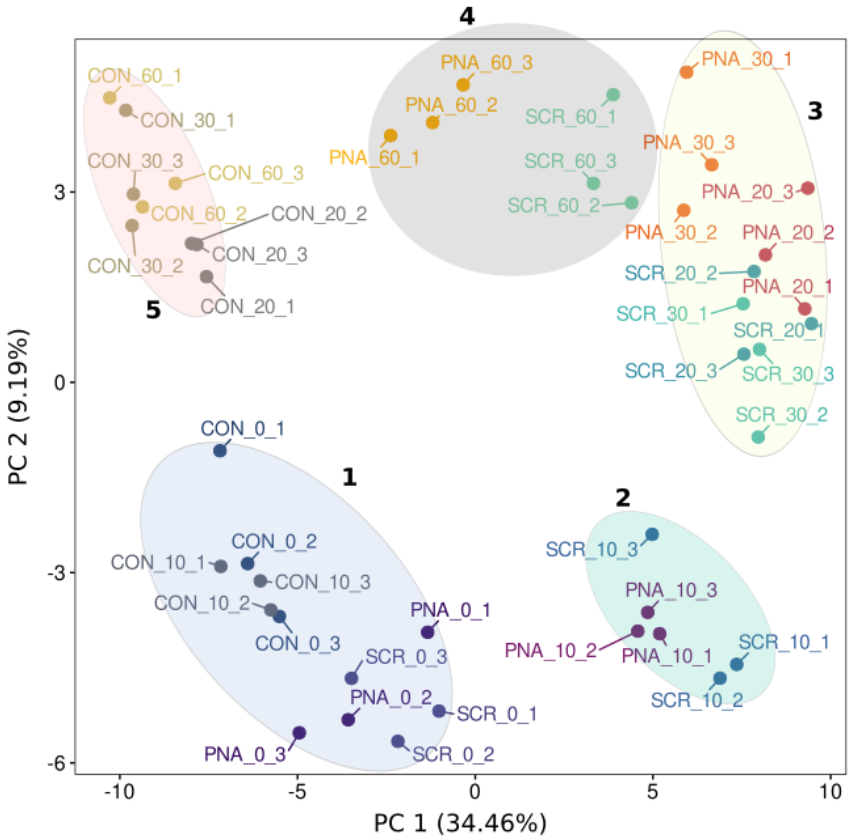
3. Principal component analysis (PCA)

- Next, I did a PCA (left plot below, A), and it looks quite good. Most of the replicates cluster together, except again for the aforementioned 2 samples.

- To remove the unwanted variation of these samples, I ran RUVs Normalization (see https://www.nature.com/articles/nbt.2931) to remove unwanted variation, below is the plot before (A) and after (B) the removal of unwanted variation:



- Below, I made some clusters to interpret the PCA results:



- The separation in the PCA with clusters is quite nice and it shows that all samples at timepoint 0, as well as the control after 10 min, cluster together (Cluster 1). This

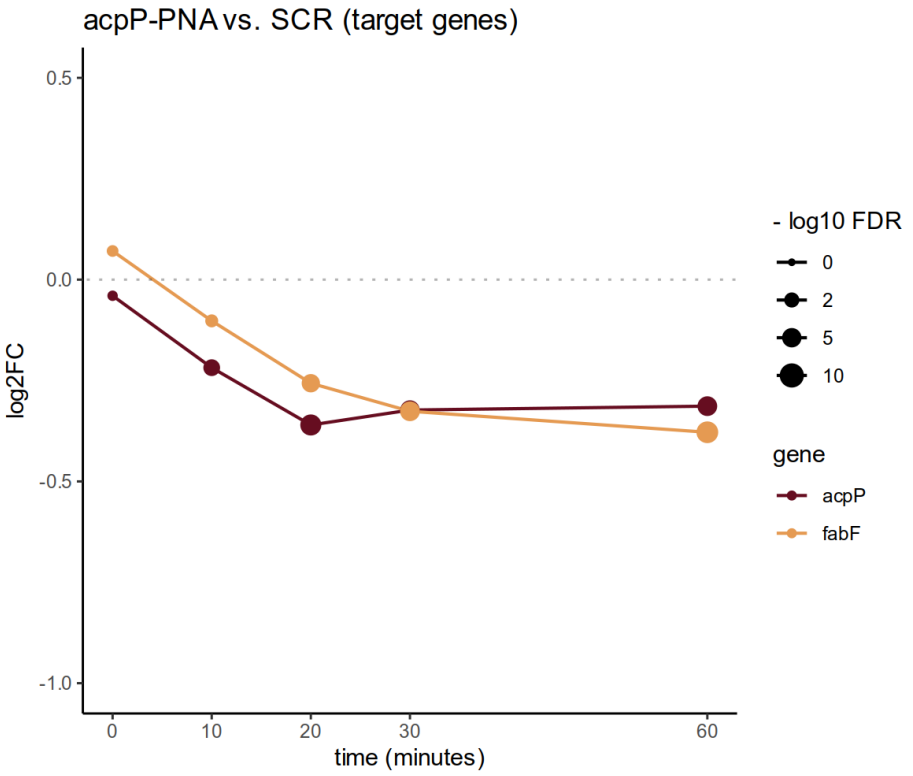
means that at T0 and 10min, not much is happening yet, even though the PNA-treated samples (cluster 2) show difference in the transcriptome compared to the untreated control (cluster 1).

- PNA and PNA_SCR also stay close after 10 (Cluster 2) 20 and 60 minutes (Cluster 3). At time-point 60, the PNA/ SCR samples (Cluster 4) seem to get closer to the control group (Cluster 5) again. Maybe this shows that, while at 20 and 30 minutes the cells are quite stressed (--> very different expression profile compared to the control), but after 60 minutes, the bacteria start going back to normal again (maybe the PNA is less active then).

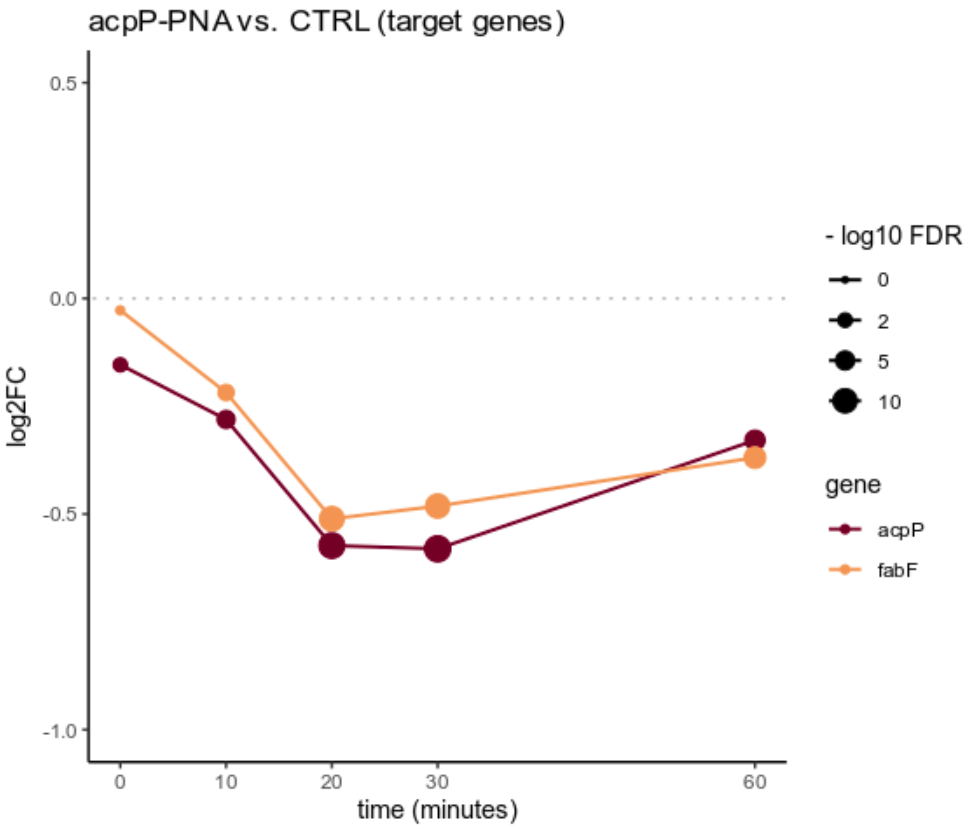
4. Differential Expression analysis:

- Next, I did edgeR differential expression (DE) analysis, comparing different time points etc. I think the most interesting points are:
    - How do the targeted genes change (acpP and fabF) - how do mRNA levels change between acpP-PNA at different time points compared to T0 of acpP-PNA? How do the acpP-PNA samples compare to scrambled-PNA/water control? Depletion?
    - Are there any interesting pathways that change with PNA treatment compared to untreated samples?

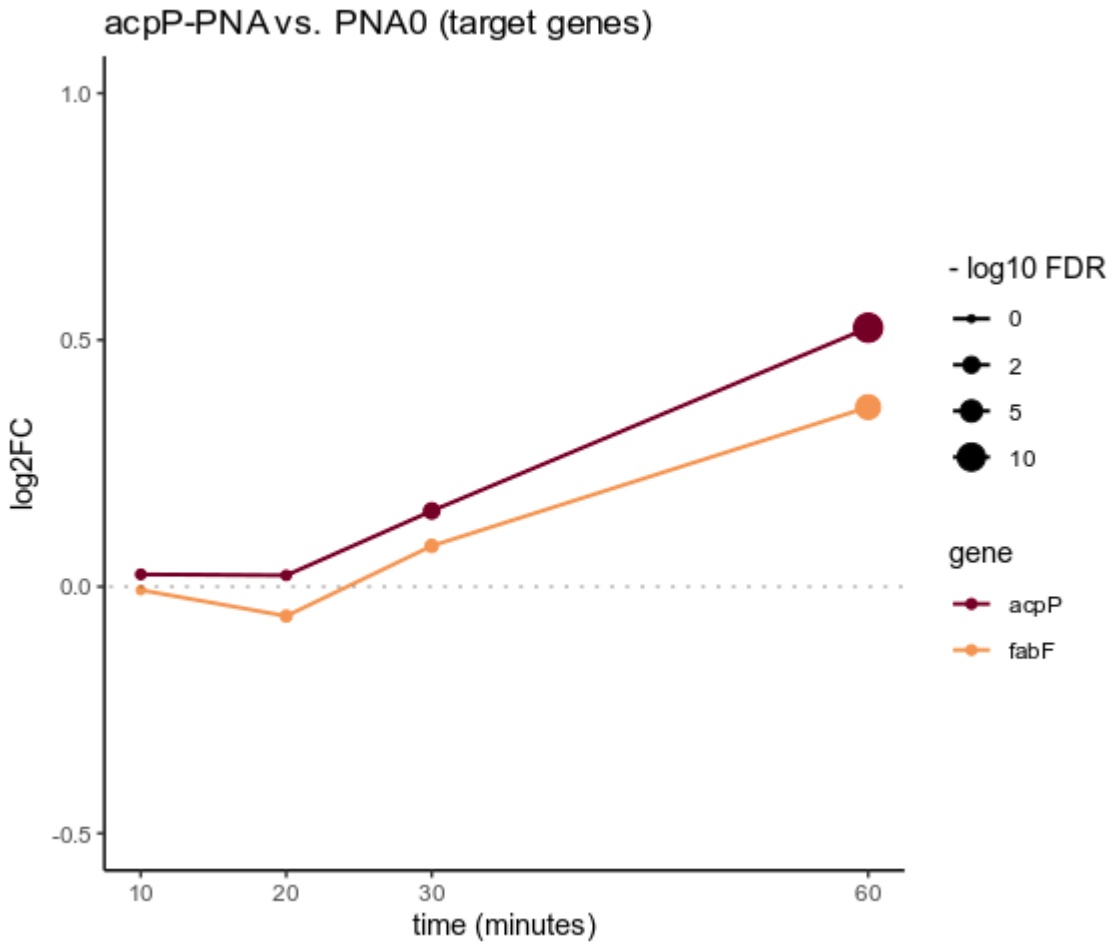
- Below, I show a plot comparing acpP-PNA to scr-PNA at each time point. I show fold changes of acpP and fabF. Point size denotes significance (bigger size → smaller FDR-p-value)
- Below, I show the log2FC of acpP-PNA vs. scrambled control at different time points for acpP and fabF. The downregulation of acpP is quite low (around -0.3) compared to the scrambled control. But it is consistent. And the strongest downregulation is at 20-30 minutes.



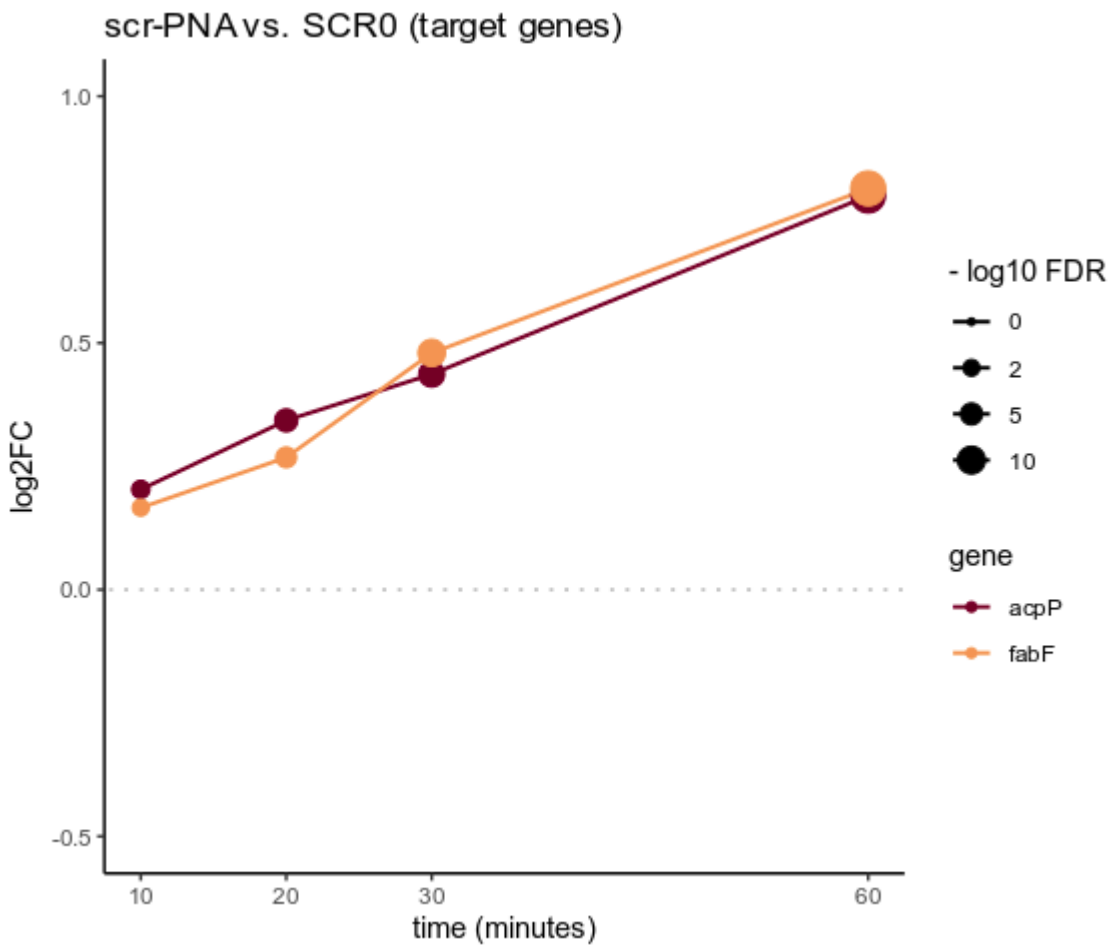
- We can see similar results for the comparison to water control samples. Only at 60 minutes, the downregulation of the target genes seems to get less strong:



- Below, I compared acpP-PNA at different time points to T0 at 0 minutes. We can see no significant change till 30 minutes. But at 60 minutes, acpP and fabF both show higher mRNA levels compared to T0.
- The fact that there seems to be no change until 30 minutes, might be because without the PNA the genes would be upregulated. Therefore I also check whether mRNA levels of acpP and fabF increase over time in the scrambled control sample in the next plot.



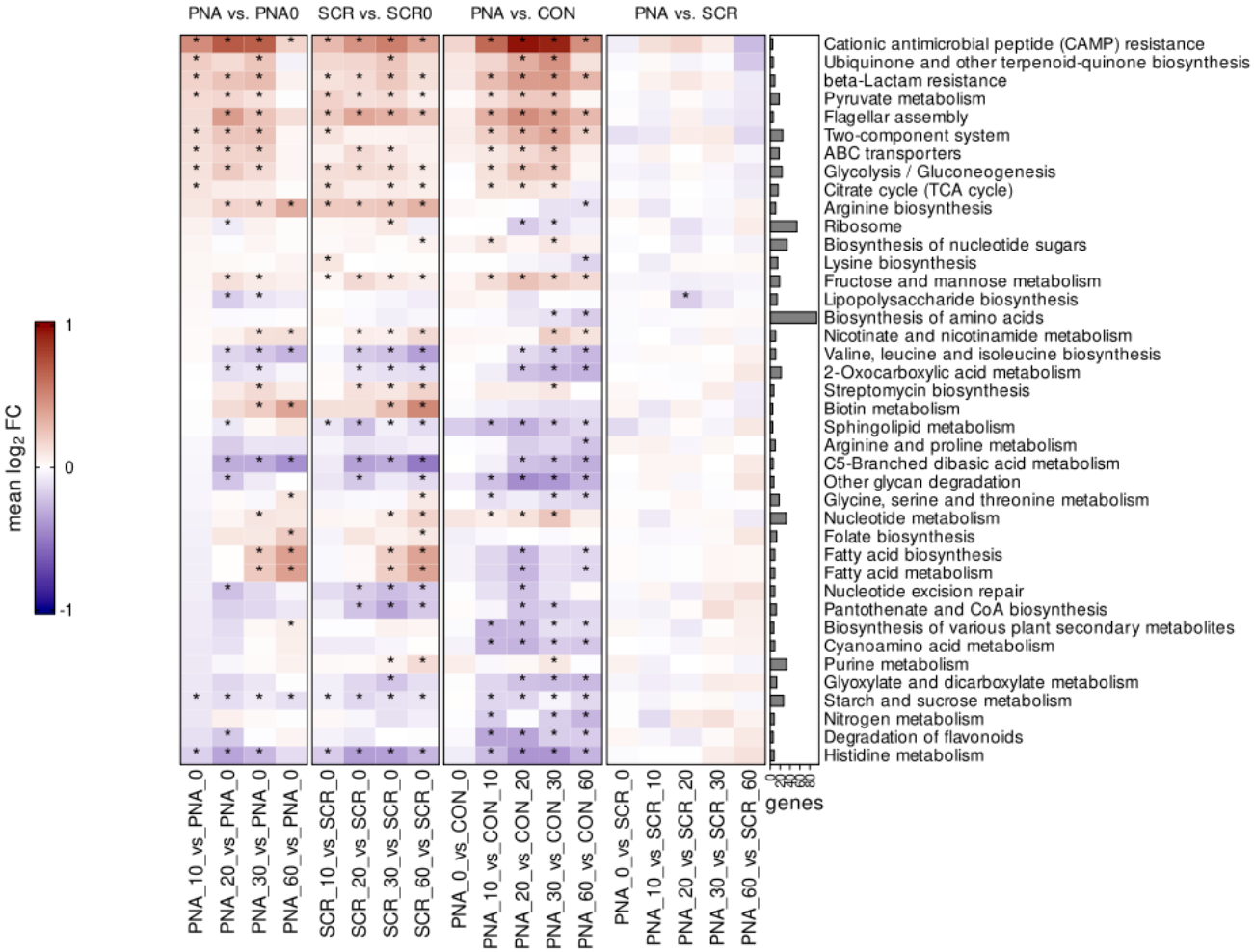
- Below, I did the same plot for scrambled control. As expected, we can see an increase in acpP and fabF level already at the early time ponts. This means that in general these genes are upregulated over time (related to growth phase?), when no PNA is targeting acpP. When acpP is targeted, this mRNA increase is suppressed, until the time point of 60 minutes.



5. KEGG-pathway analysis:

- I performed a KEGG-pathway enrichment analysis. Below I show a heatmap, showing all the differential upregulation/depletion of KEGG-pathways. The heatmap shows, per sample, the most significantly enriched/depleted pathways. The color denotes average log2FC, i.e. if blue, a pathway is downregulated, if red, upregulated. The barplot on the right denotes the number of genes belonging to the respective pathway.
- In general, the acpP-PNA shows almost identical pathway regulation as the scrambled control. This makes sense as they probably both have the same effect on the membrane, independent of the exact PNA sequence. That is also likely is the reason why between PNA and SCR there are almost no significantly different pathways (last columns)
- Cationic antimicrobial peptide response & 2 component systems are upregulated upon PNA exposure, with a peak at time points 20-30. This is similar to our results in *Salmonella* and UPEC and probably due to membrane disruption by the CPP.

- Probably also interesting is the fatty acid metabolism pathway, of which acpP is a member.



Experiments with metatranscriptomics

- Another experiment she did was RNA-Seq on a community (three strains: *P. copri*, *B. theta*, and *B. vulgatus*). 2 different protocols were tried, and 2 time points of KFF-acpP exposure (+control), after 0 and 2 hours were measured. As this is a metatranscriptomic dataset, and I expect some of the reads between *vulgatus* and *theta* to be similar, so the analysis is a bit more complex. I'll update you once I have done more.