final

Zhuoyang Chen

February 25, 2020

Final project

Differential Regulation Network-based Quantification and Prioritization of Key Genes Underlying Cancer Drug Resistence Based on Time-course RNA-Seq Data

https://doi.org/10.1371/journal.pcbi.1007435 (https://doi.org/10.1371/journal.pcbi.1007435)

Introduction

This project is based on the article from Zhang et.al. Different glioma cell lines are treated with an cAMP activator dbcAMP. Through experiments, cell line LN18 shows resistance to the drug and DBTRG is sensitive. Samples are collected at 0h, 6h, 12h, 24h and 48h after treatment.

The hypothesis is do gene expression profiles differ based on the two time-course RNA-seq data between drug-sensitive and resistant cell lines.

In the article, untested cell line U87 is used as test set after analyzing the difference between the two cell lines. In this project, I mainly focus on the tested DBTRG and LN18 cells.

Method

Library Preparation

The RNA-Seq data is generated by Dr. Xingcheng Liu, who help culture cell, prepare the RNA library and get sequencing data. The name can be found on the NCBI RNA data website and also in the Ackwonledge section.

Beads containing oligo (dT) were used to isolate poly(A) mRNA from total RNA. Purified mRNA was then fragmented in fragmentation buffer. Using these short fragments as templates, random hexamer-primers ere used to synthesize the first-strand cDNA. The second-strand cDNA was synthesized using buffer, dNTPs, RNase H and DNA polymerase I. Short double-stranded cDNA fragments were purified with a QIAquick PCR extraction kit (vendor) and eluted with EB buffer for end repair and the addition of an 'A' base.

RNA libraries were prepared for sequencing using standard Illumina protocols. The short fragments were ligated to Illumina sequencing adaptors. DNA fragments of a selected size were gel-purified and amplified by PCR.

Download RNA-Seq Data

Cell lines are epithelial brain cancer glioma cells: LN-18, U87 and DBTRG-05MG. They are treated with 1mM cAMP activator dbcAMP and data collected at 0, 12, 24, 36, 48h.
Accession number: **GSE128722**

```
##       Samples Cells Time
## 1  SRR8769935 DBTRG    0
## 2  SRR8769936 DBTRG    6
## 3  SRR8769937 DBTRG   12
## 4  SRR8769938 DBTRG   24
## 5  SRR8769939 DBTRG   48
## 6  SRR8769940   U87    0
## 7  SRR8769941   U87    6
## 8  SRR8769942   U87   12
## 9  SRR8769943   U87   24
## 10 SRR8769944   U87   48
## 11 SRR8769945  LN18    0
## 12 SRR8769946  LN18    6
## 13 SRR8769947  LN18   12
## 14 SRR8769948  LN18   24
## 15 SRR8769949  LN18   48
```

Download a summary file that contains all the *SRR* files called **SRR_Acc_List.txt** and retrieve all the *SRR* files via SRAtoolkits. All the *SRR* files are split into 2 files that are pair-end.

Quality Control

Run FastQC on all samples and use multiqc to visualize all the results.

Problems encounted:

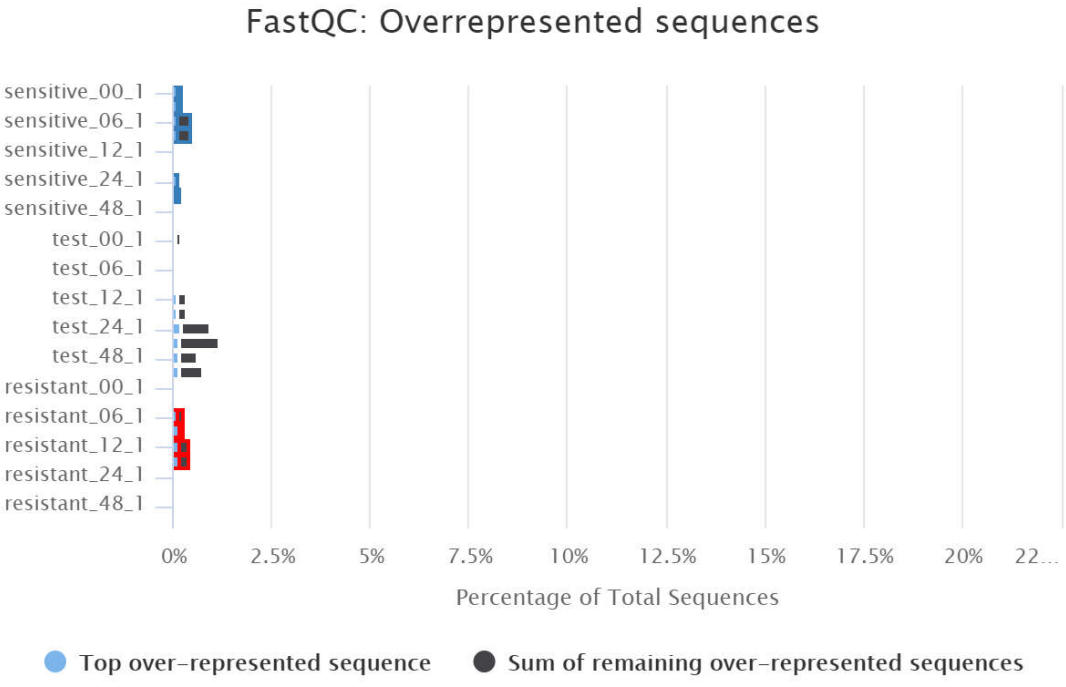
1. When use SRA Analysis to see **Taxonomy Analysis**, found that there is a strong signal of Mycoplasma hyorhinis contamination. Although the percentage is not high, it has a high coverage over 50. Other samples also have different extent of contamination, but not with Mycoplasm highest.
2. Check the FastQC report of the first sample and found that there is a very high percentage of duplicated sequence up over 50%. Use NCBI Blast to check the origin of those highly overrepresented reads, results show that they are all from human Mitochondria ATP8 (ATP synthase F0 subunit 8) genes.

By looking for references, Mycoplasmas are notoriously common contamination for cell culture becuase it is so flexible to pass through most filter membranes and can reach a high concentration without cause any disturbance. It has a small genome size (0.6 Mbp) but lack genes for precursors synthesis and energy metabolism. Thus may alter the host's cell biology. In this case, Mycoplasma hyorhinis endonucleases can degrade host cell DNA, providing DNA precursors for the parasite.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4357728/ (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4357728/)

In terms of PolyA selection, it is very likely to enrich mt-rRNA and mt-mRNA which is also rich of AT.

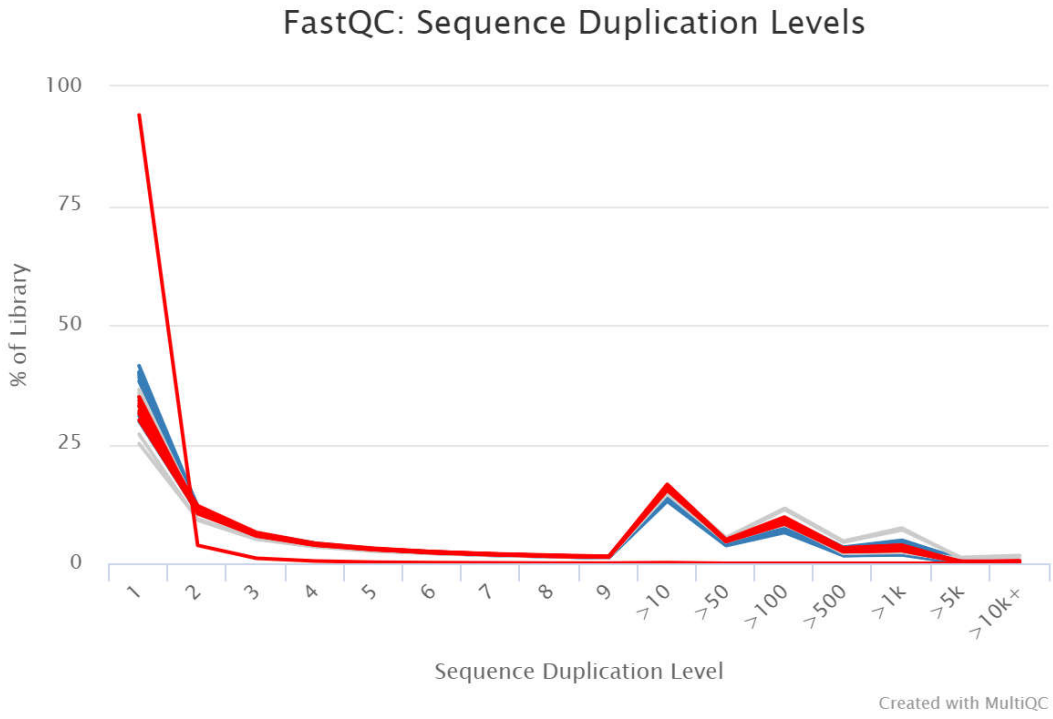
Mutltiqc Results

- From overrepresented reads results, there are in fact very low overrepresented sequences.



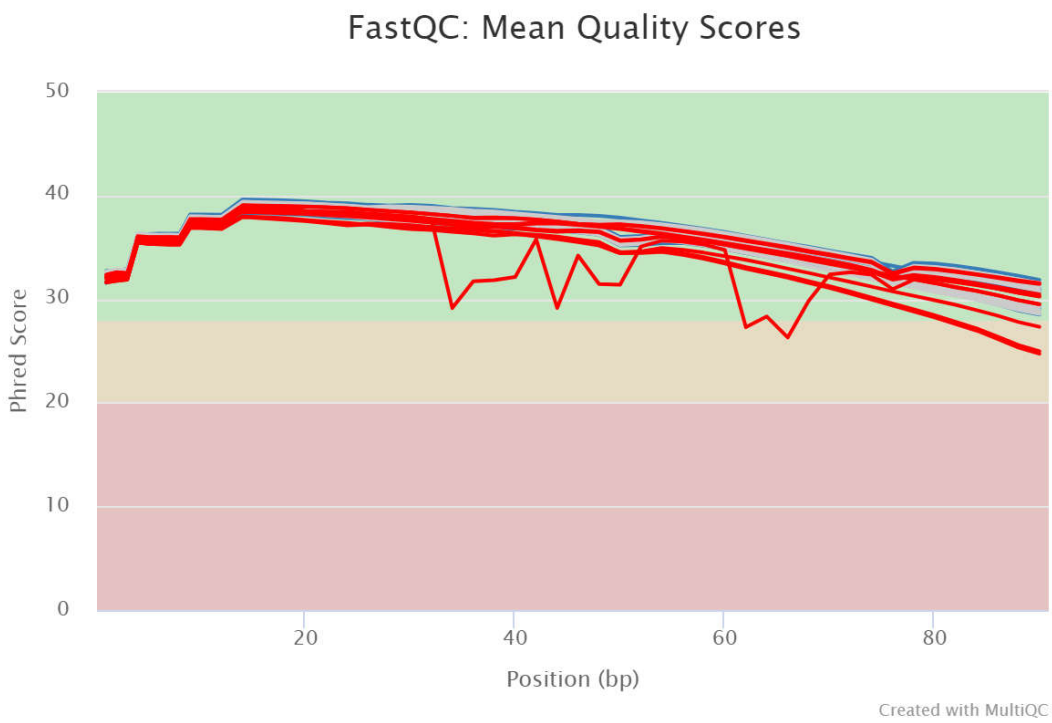
Overrepreseted sequence level

- From *Seqence Counts*, all the samples are nearly of the same level, must all of them have a very high duplication level, over 50%, except one sample **SRR8769949_1.** High duplication level in RNA-seq data is expected because we usually over amplify samples in order to detect low expression genes.



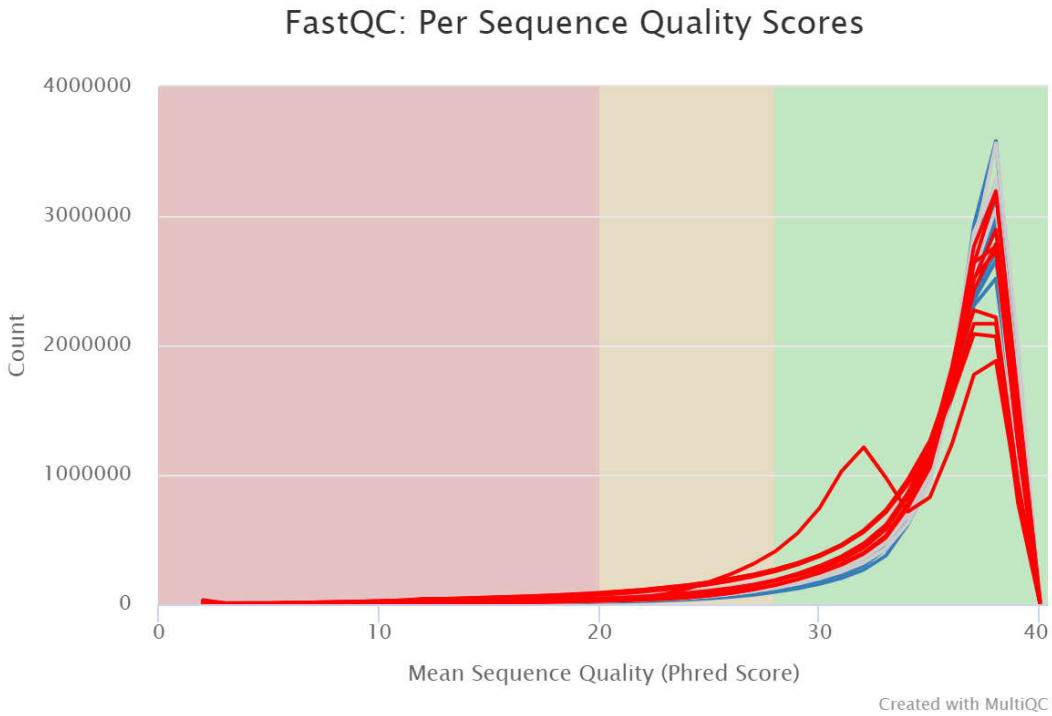
Duplication level. Blue is sensitive cell lines, red is resistant.

- However from *Mean Quality Scores* I found that **SRR8769949_1** has lower and more fluctuating scores.



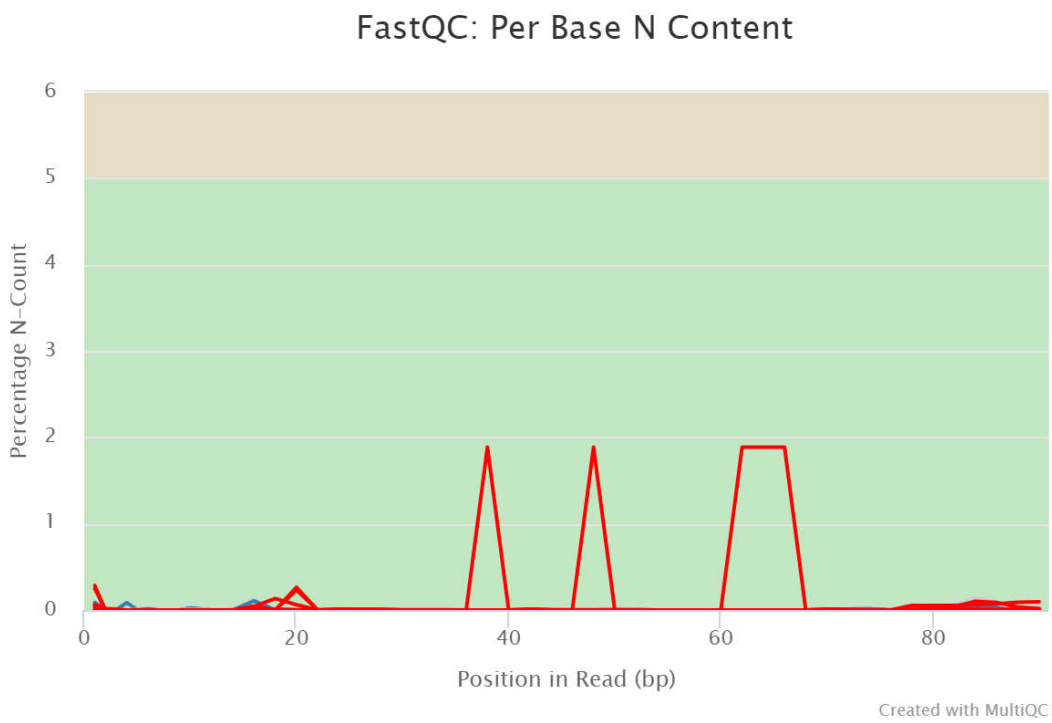
Mean Quality Scores. Blue is sensitive cell lines, red is resistant.

- Also from *Per Sequence Quality Scores* it has minor peak at Phred 33, while others just have one major peak.



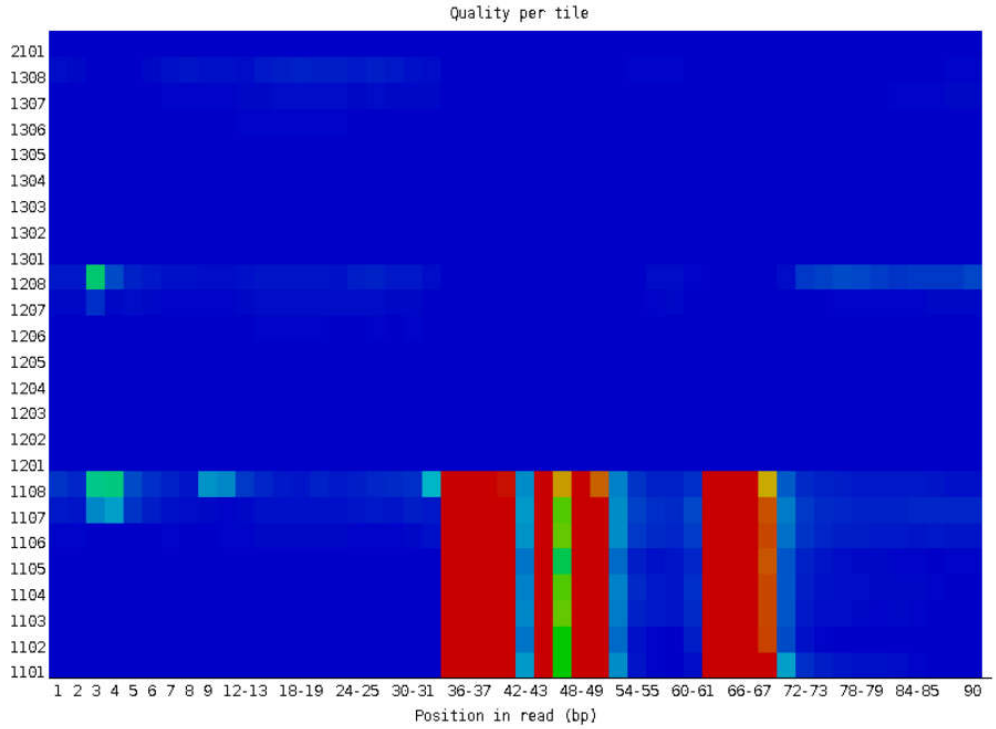
Per Sequence Quality Scores. Blue is sensitive cell lines, red is resistant.

- From *Per Base N content* I saw that sample has tree peaks at 38, 48, 62-64 bp with a percentage of 1.88%. From this aspect, the low quality in some position of **SRR8769949_1** may be caused by the corresponding N content. The minor peak at Phred 33 further verify this, as N base would be assigned a Phred scores as 33.



Per Base N content. Blue is sensitive cell lines, red is resistant.

- When checking the tile quality, I obseverd that a section of the flow cells are abnormal, showing a continuous area of red colors, and the area is corresponding to the N bases regions. This indicates that something wrong is on the lanes or sequencer.



Tile quality

Preprocessing

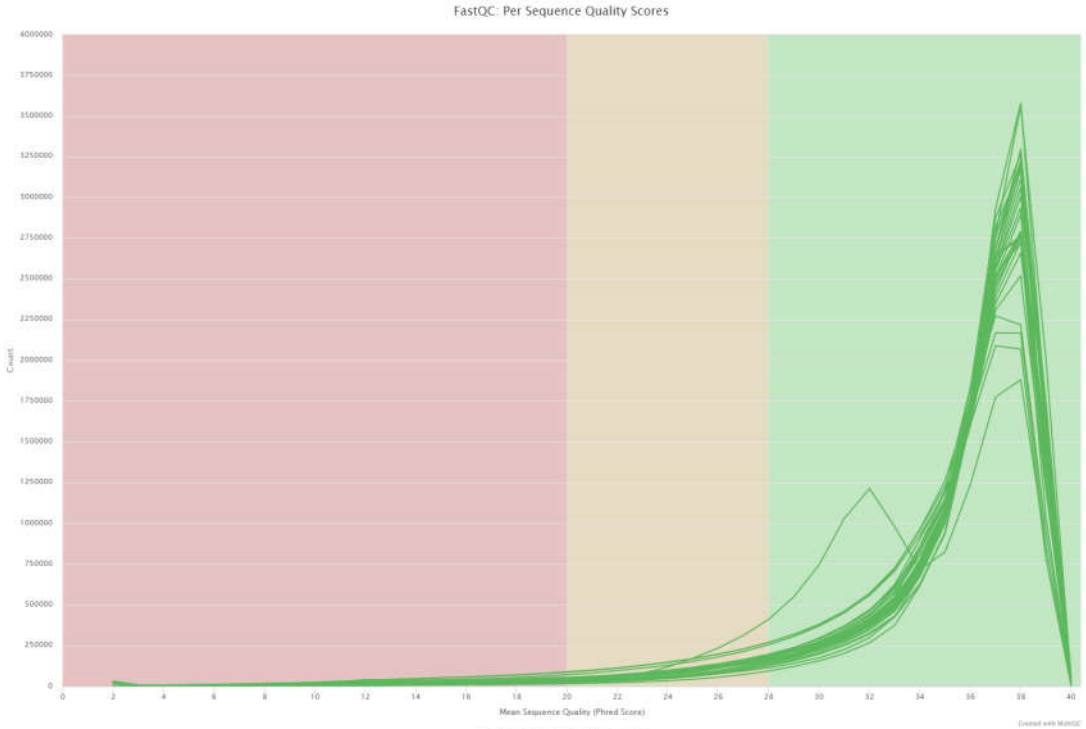
Use trim_galore to filter out all the low quality reads. If a read have a low mean quality, the mate read would also be dropped.

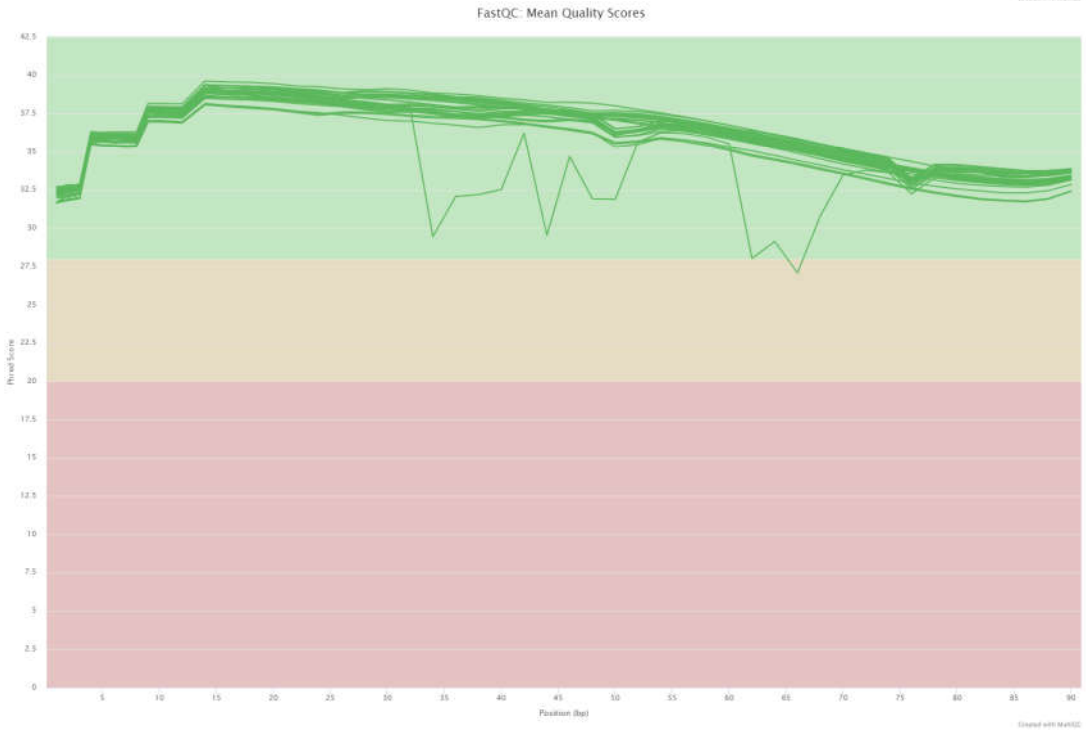
Here *–paired* indicates trimming sequence in paired mode, which enables trim_galore to check the two paired reads at the same time and if one read are abandoned, the other would be abandoned as well. *–retained* indicates retaining the bandoned reads. *stringency* is set as 10 to only consider a sequence as adaptor when more than 10 bases are overlapped.

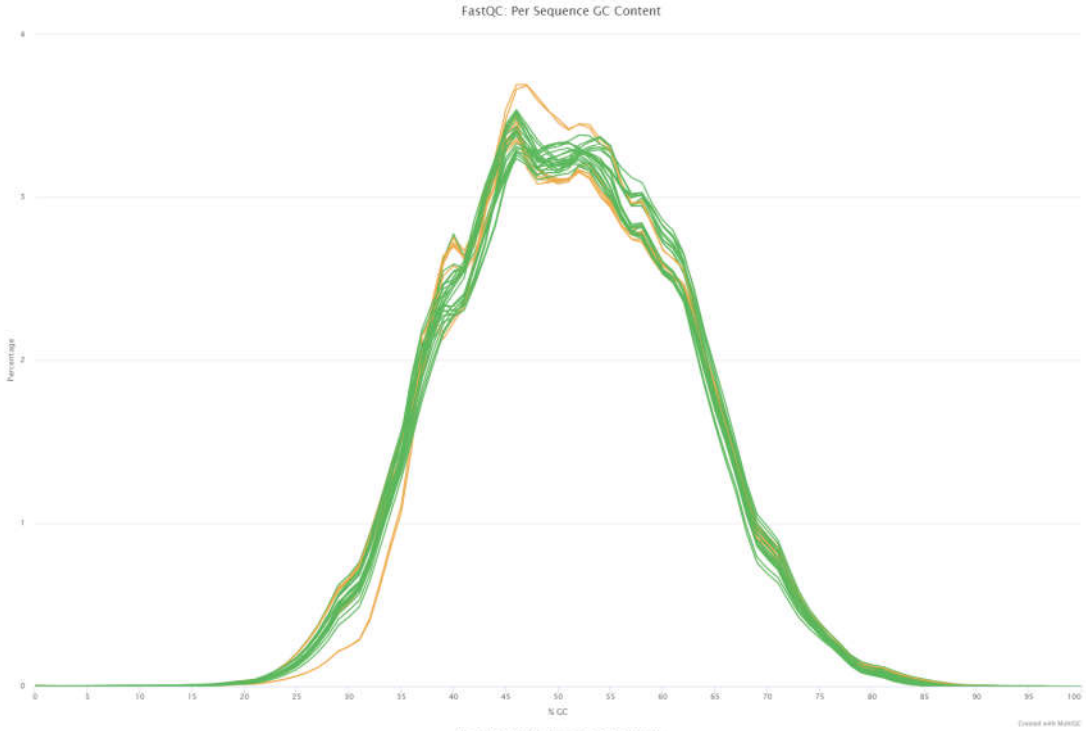
To check results after trimming, run FastQC on the trimmed data. Then run multiqc.

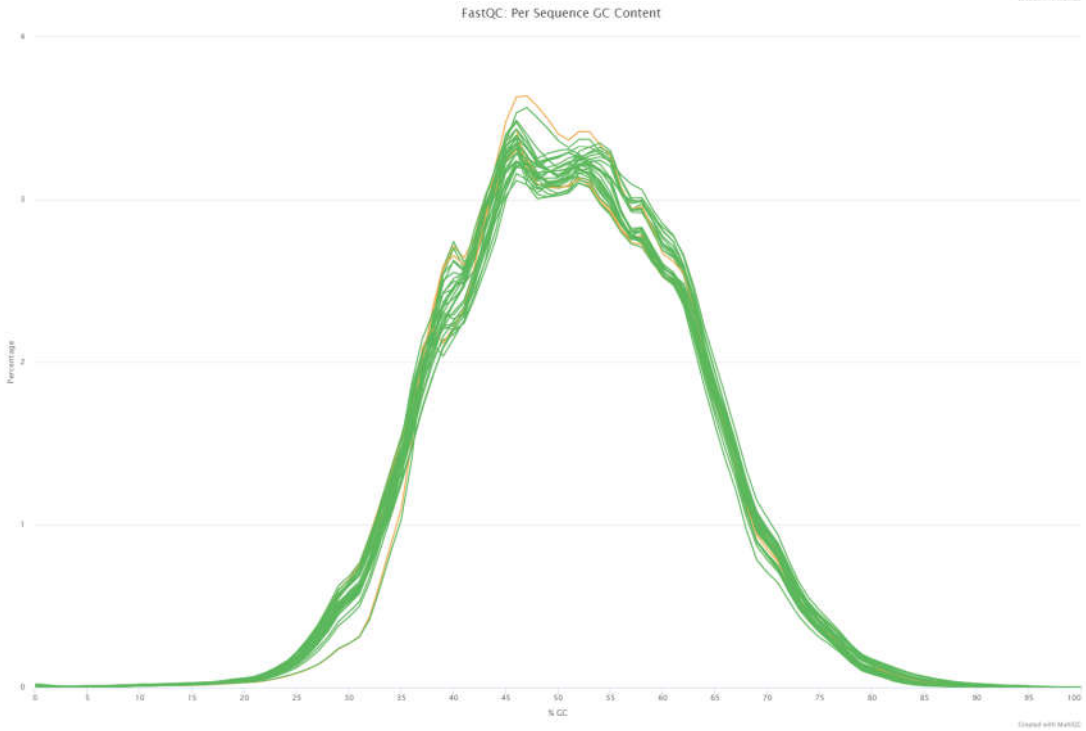
Process results

By a simple trimming step, the contamination and duplicate issue are still there, since we just improve the overall quality of reads. There are improvements on **Per Base Sequence Quality** and **GC content**. It is unnecessary for these samples to perform a trimming step, because there is no adapter detected and the overall quality is high.


FastQC: Per Sequence Quality Scores
Count
Mean Sequence Quality (Phred Score)


FastQC: Mean Quality Scores
Phred Score
Position (bp)



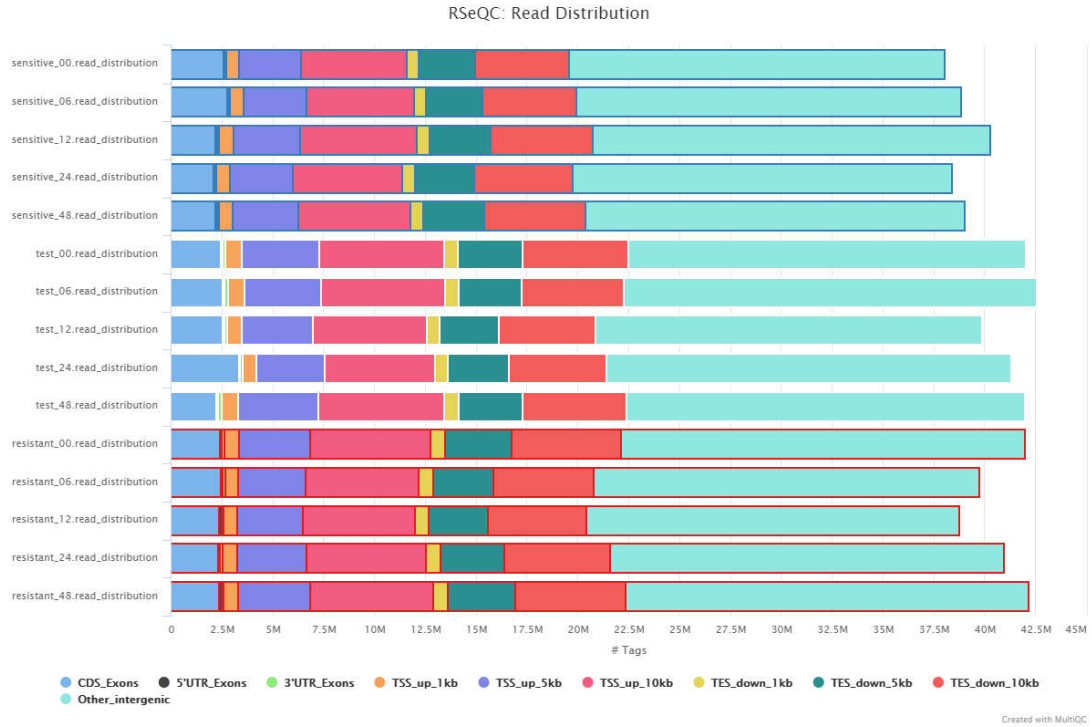


alignment

Alignemt step is implemented by STAR using default parameters. After alignment, perform alignment QC by RSeQC. Since there is no corresponding BED file in Ensembl, so I generate my own BED file and from gtf annotation and .bai file from .bam file. Then perform read_distribution and geneBody_coverage analysis using RSeQC.

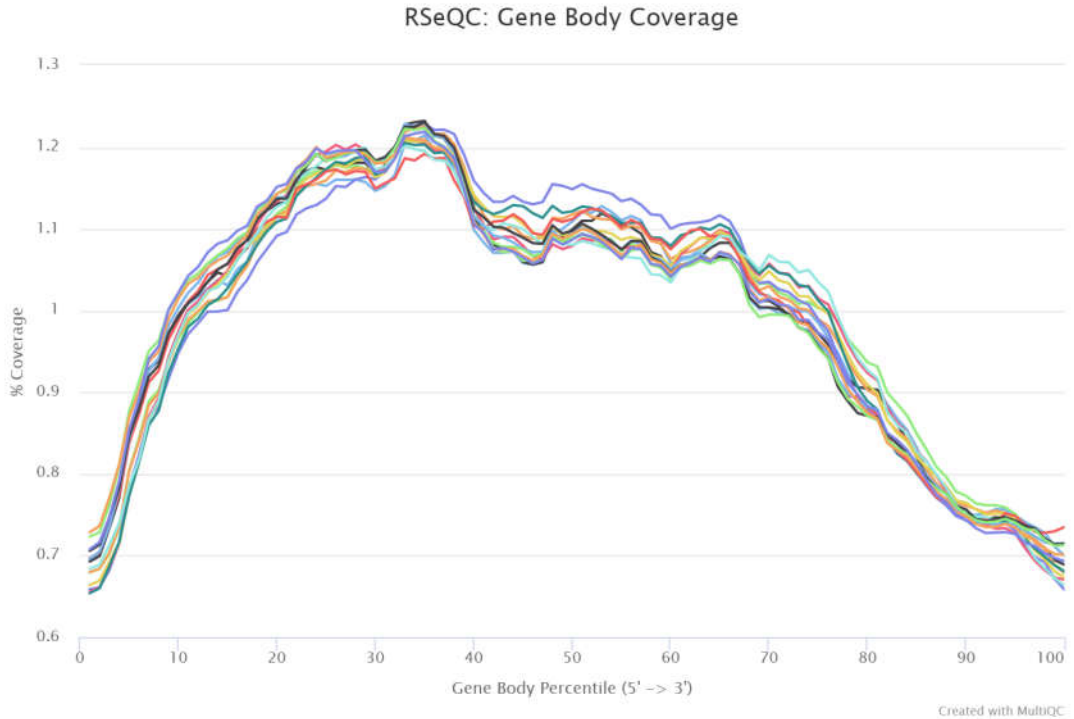
Use multiqc again to inspect the alignment results. Results are shown as below.

We can see from the figure that there are around 2.5M reads assigned to CDS, a small portion compared to the total reads. I think this is enough for downstream analysis thanks to a sequencing depth of 13M for each samples.



Read distribution

From the gene body coverage figure, there is no 5' and 3' bias, which is great, no severe degredation.

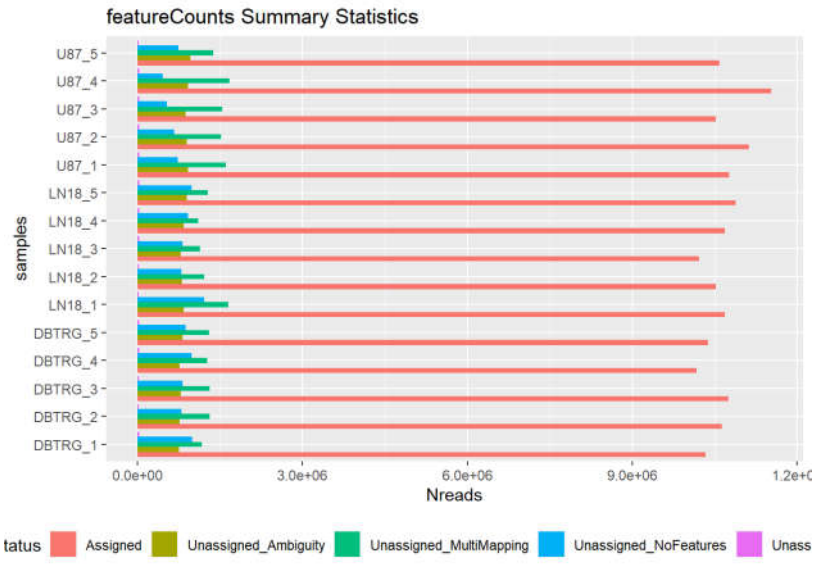


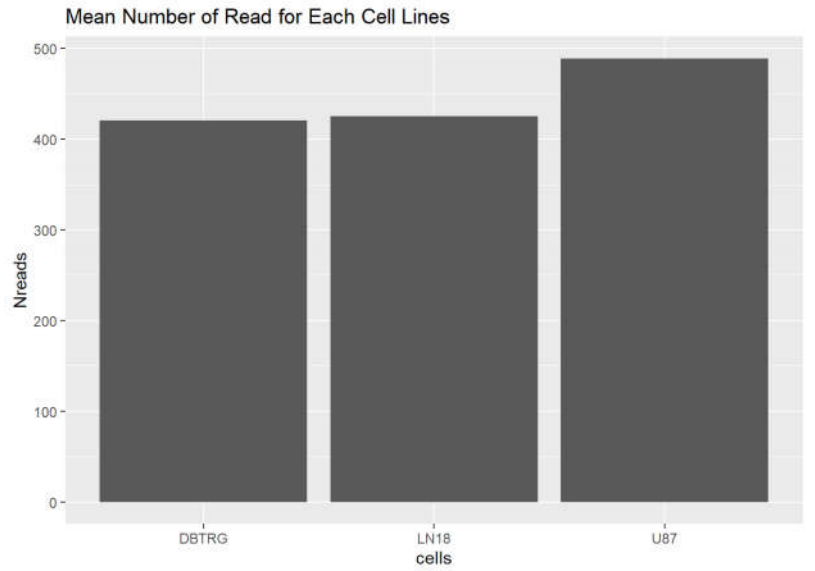
geneBody coverage

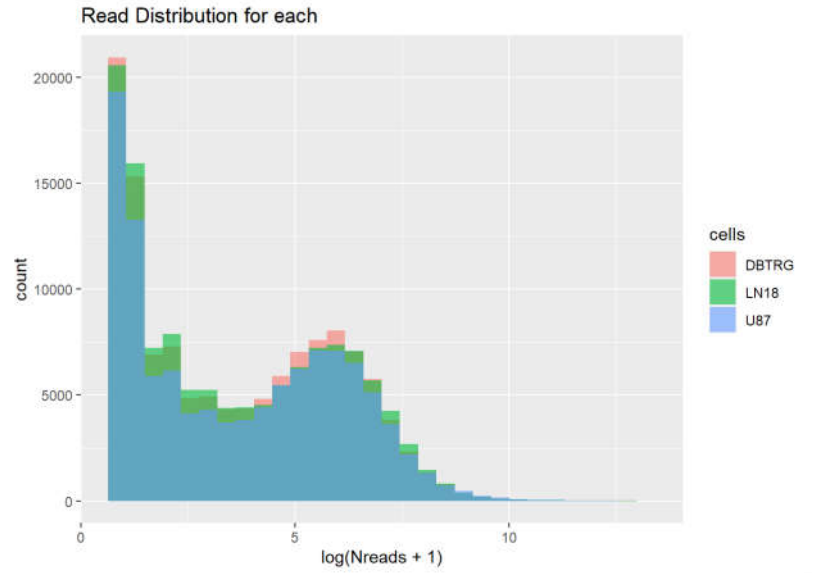
Feature counts

Use featureCounts to summarize the mapped reads information.

Results shown that all samples has an assigned rate over 76%.



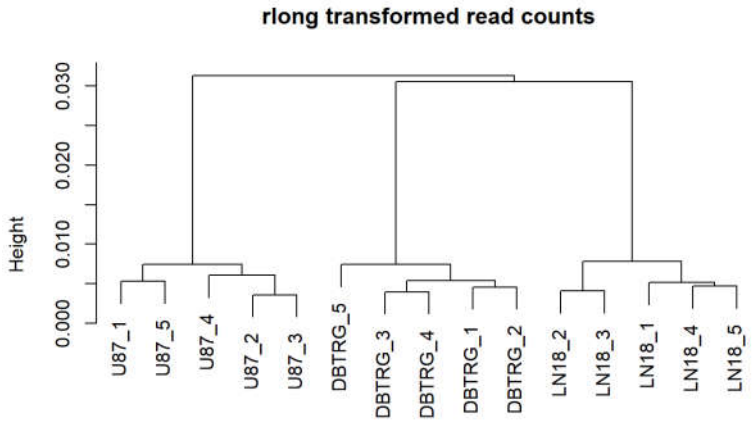


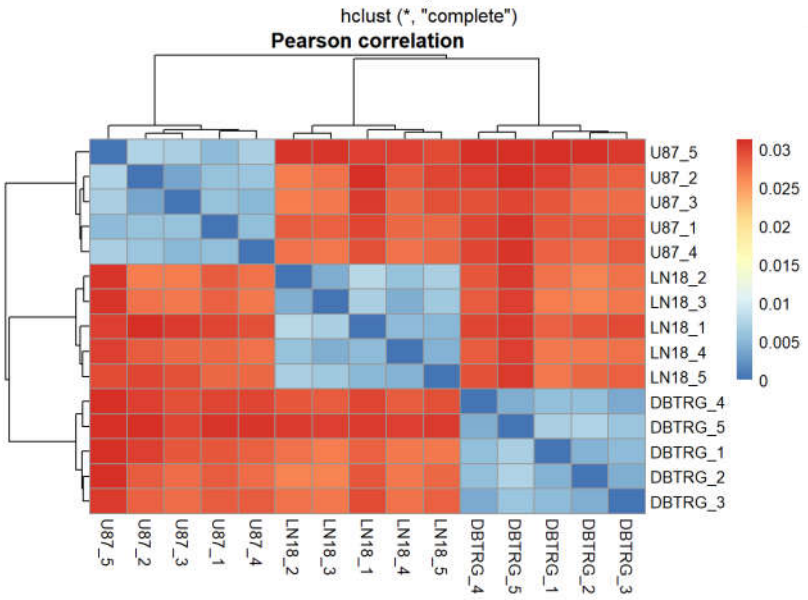


Results

Heatmap and dendrogram

Get data into DESeq2.

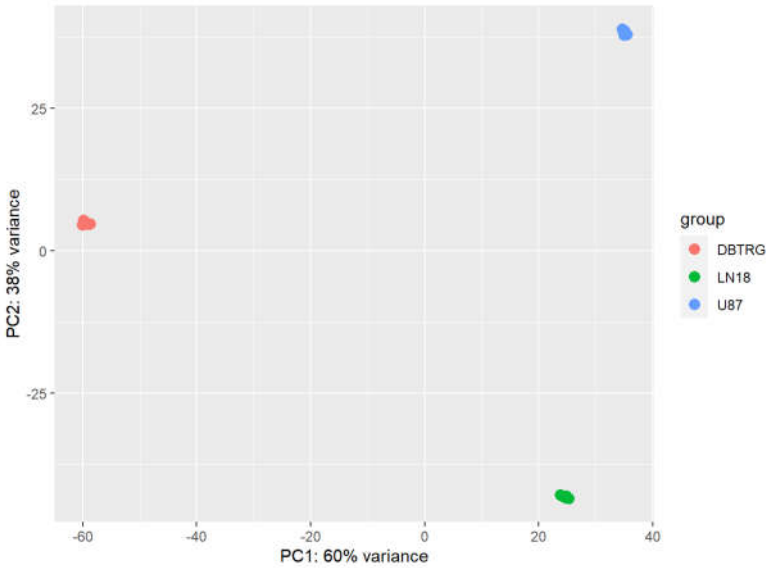


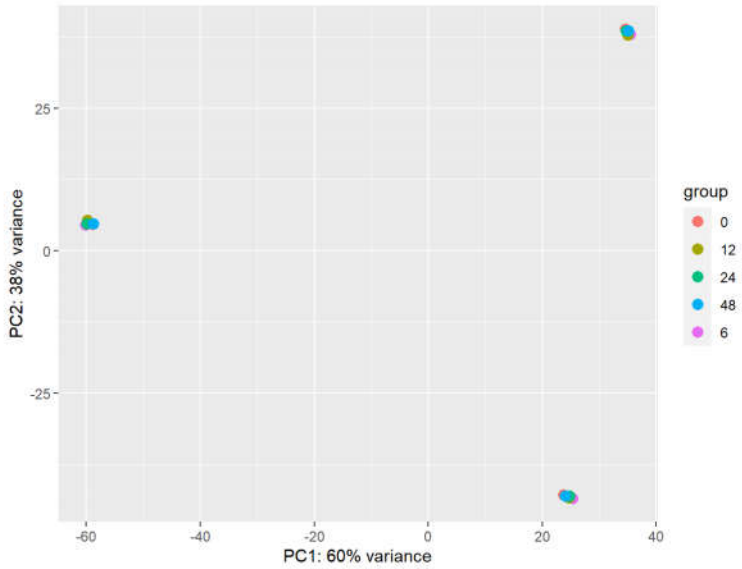


From the cluster results, we see that the three different cell lines are clustered as three groups correctly, but what is unexpected is in the article, DBTRG and LN18 is the experimental groups, while U87 si the test group. We known that DBTRG is the drug-sensitive cell and the other is the resistant one. From the cluster plot, we see DBTRG and LN18 are clustered together rather than U87 with one of them. Since we hope to see that if U87 is clustered with one of the two experimental cell lines, then we can predict whether the test cell lines U87 is sensitive or resistant. From the plot, the overall expression depends more on cell lines than their drug sensitivity properties.

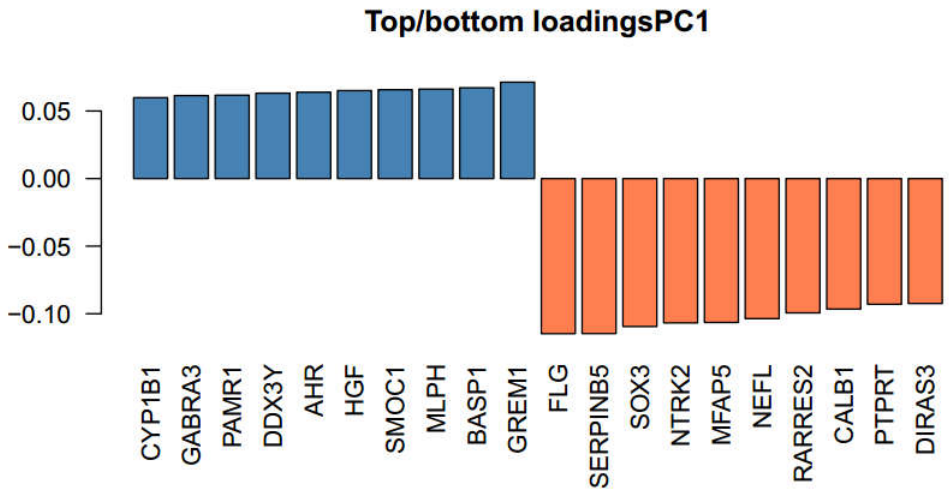
PCA

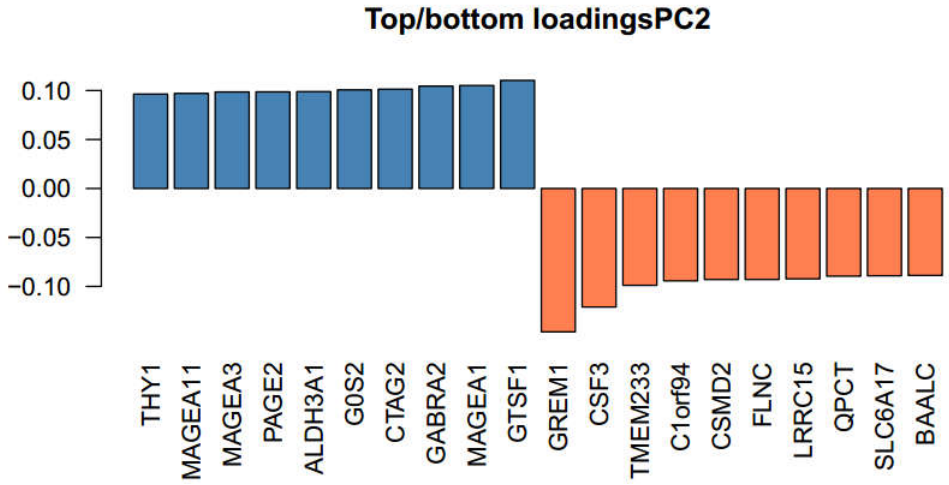
Perform PCA analysis.





From the PCA plots we can see that cell lines are the determined factors that seperate samples but not time. Again, since PCA is performed on overall expression, this tells us difference of overall expression is largerly explained by cell lines.





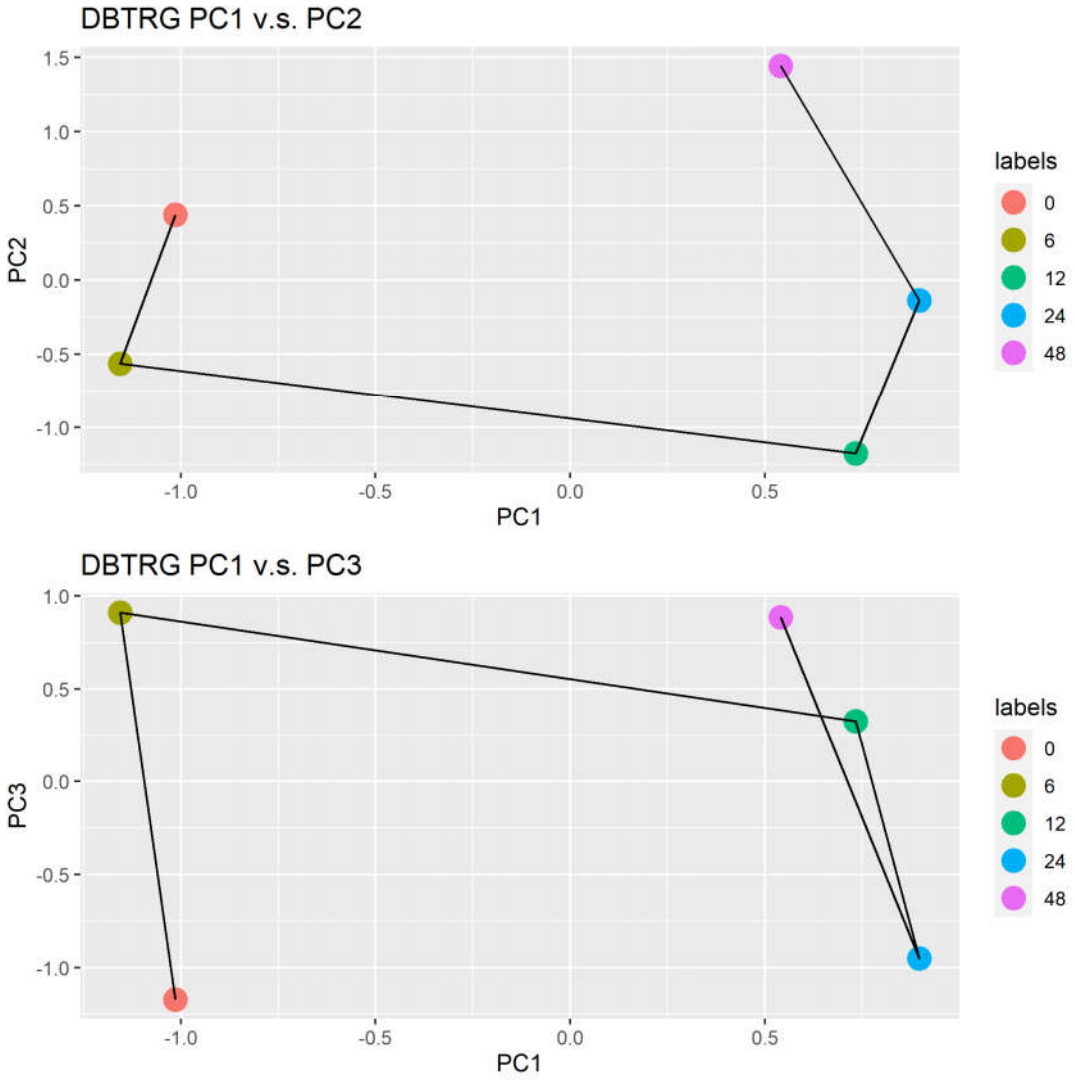
Top Loading of PCA

The top loading genes are GREM1 (BMP antagonist) for positive coefficient and FLG (filaggrin) for negative.

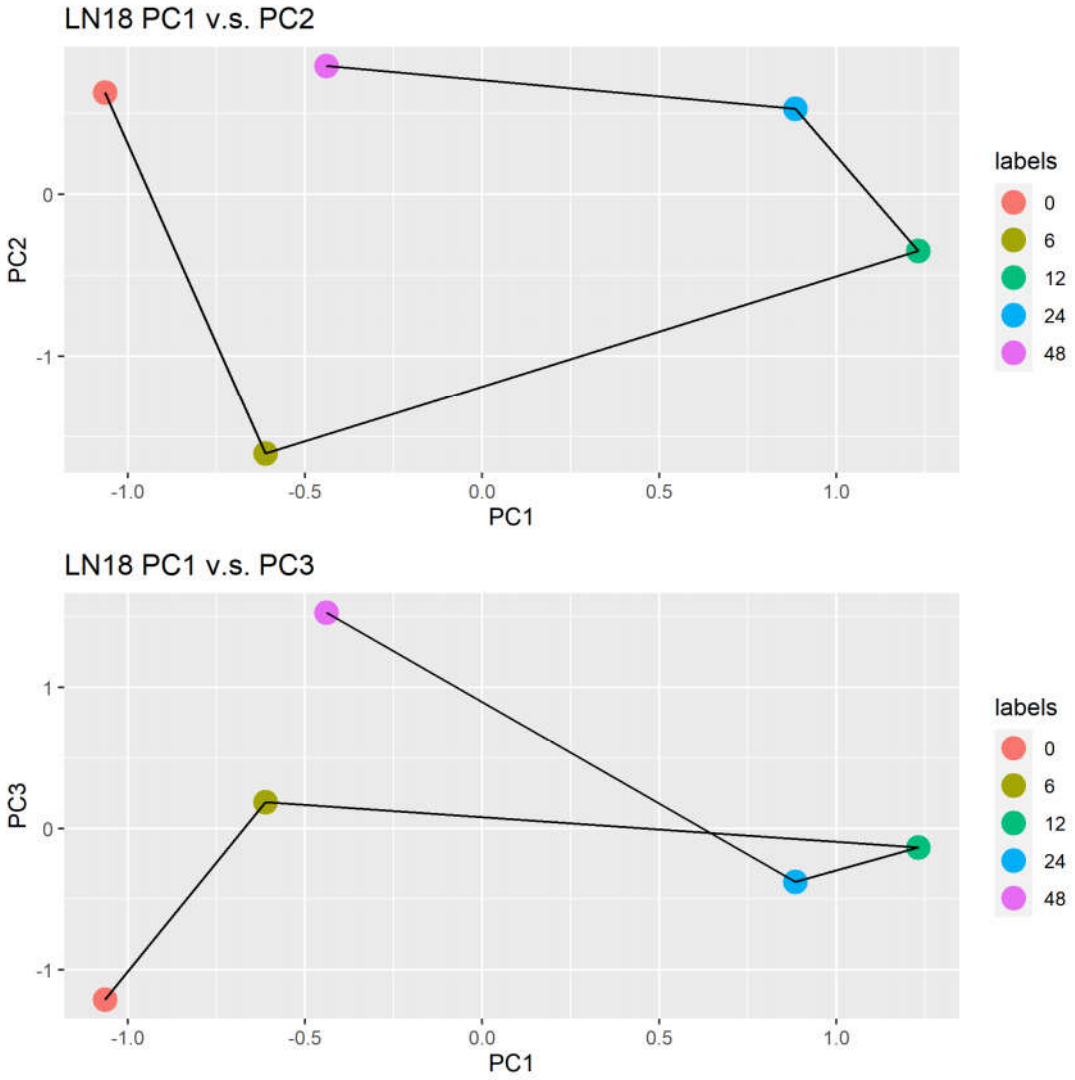
Gremlin primarily inhibits bone morphogenesis and is implicated in disorders of increased bone formation and several cancers.

Filaggrin is essential for the regulation of epidermal homeostasis.

Implement PCA on each cell line, I noticed that the drug resistance can kine of be captured. We can see that as time increasing, the cell status of DBTRG is moving forward, while that of LN18 is moving approach its original position (Time 0).

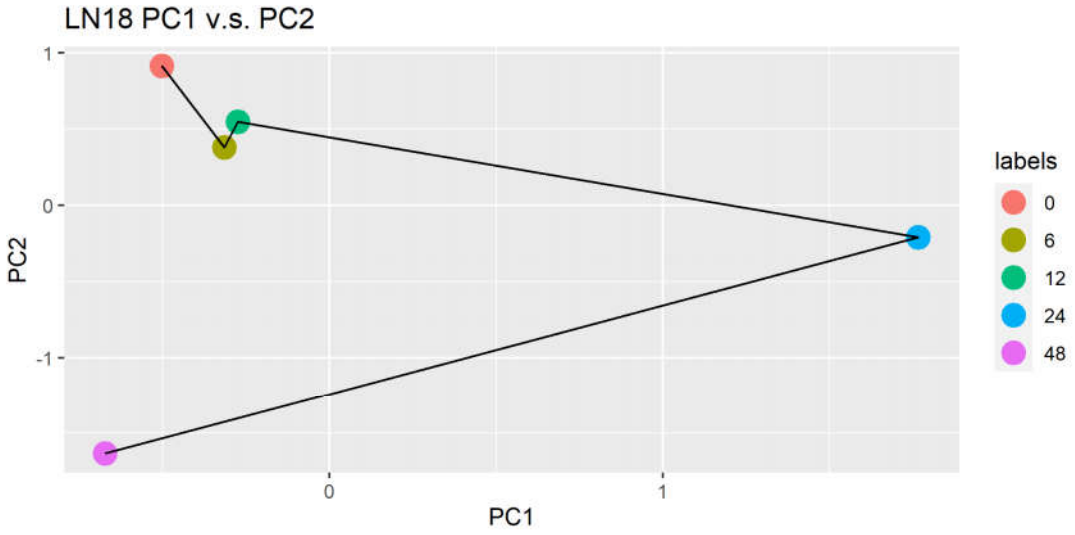


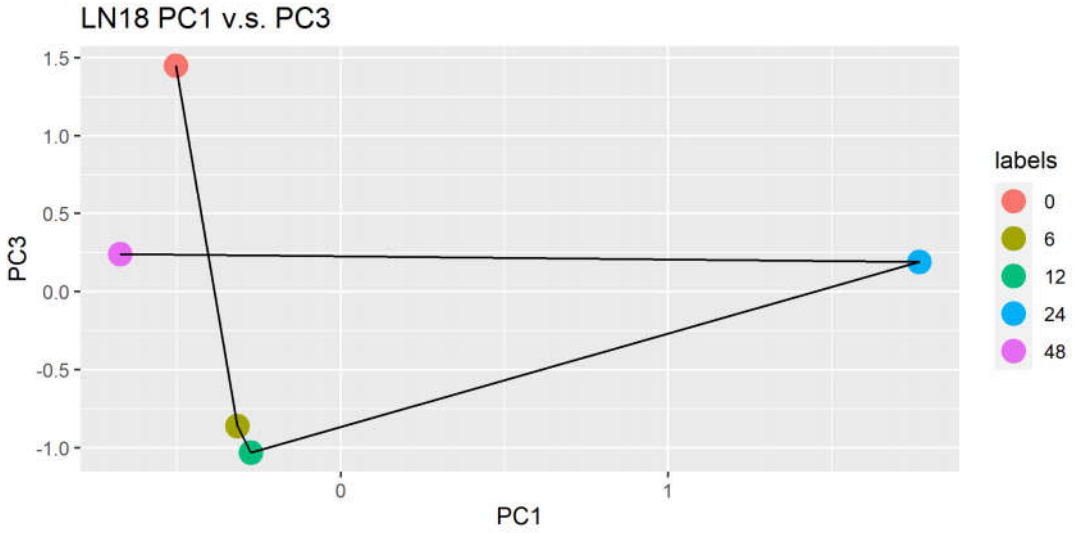
PCA of DBTRG (drug-sensitive)



PCA of LN18 (drug-sensitive)

A similar trend of going closer to the original position can be observed in the test cell line U87.





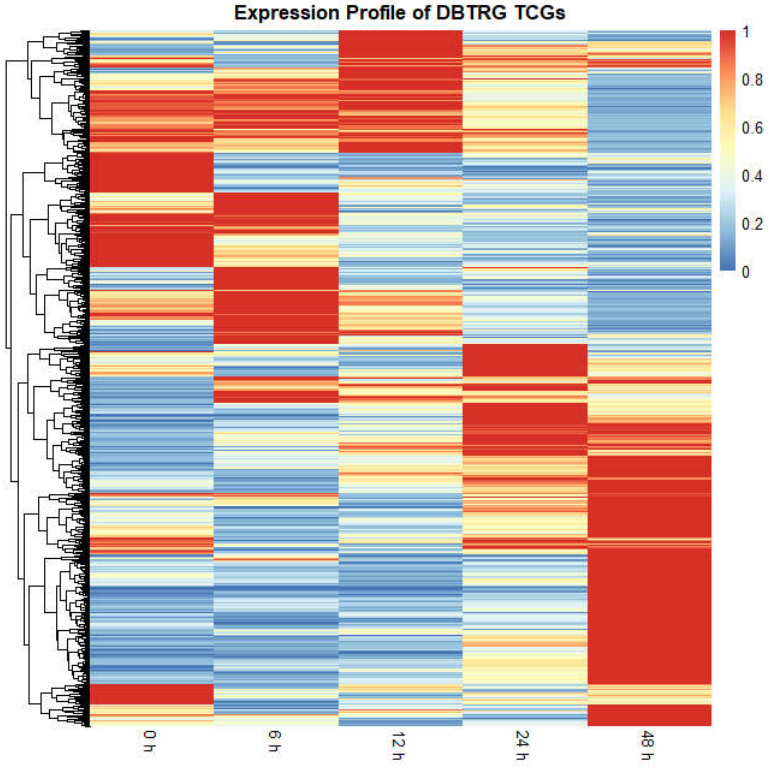
PCA of U87 (test)

TCGs (Temporally Changed Genes) Identifications

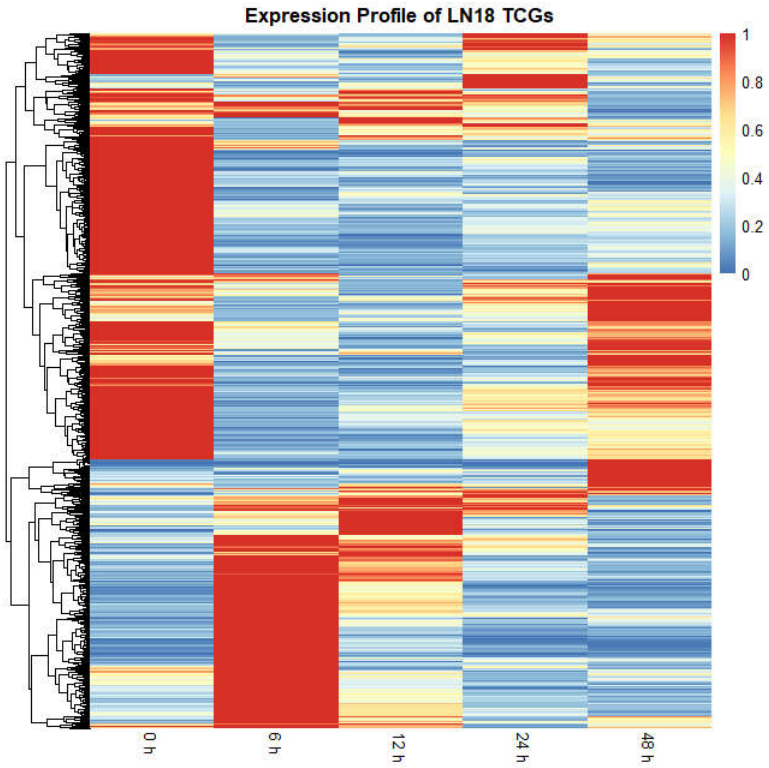
Here use the same method as stated in the article. Since the threshold for maximum expression of a gene across time and the minimum fold change between points are designed empirically, and the original article use FPKM as measurement while we use gene counts, we should define them for this analysis. Given that there are totally 20000 genes and 100 - 200 genes are identified as TCG (1%-2%), we try different threshold to gain a similar portion of TCGs.

As I test the parameters in the article, which threshold for maximum expression is 10, and the minimum fold change is 5. Out of 39000 expressed genes, 1100 genes are identified as TCGs for DBTRG, and 1300 genes for LN18, which are near 3%. Since it is not far away from the expected ratio, I adapt this set of parameter for further analysis.

From the expression profiles, we can see that they are quite different between sensitive and resistant cells.

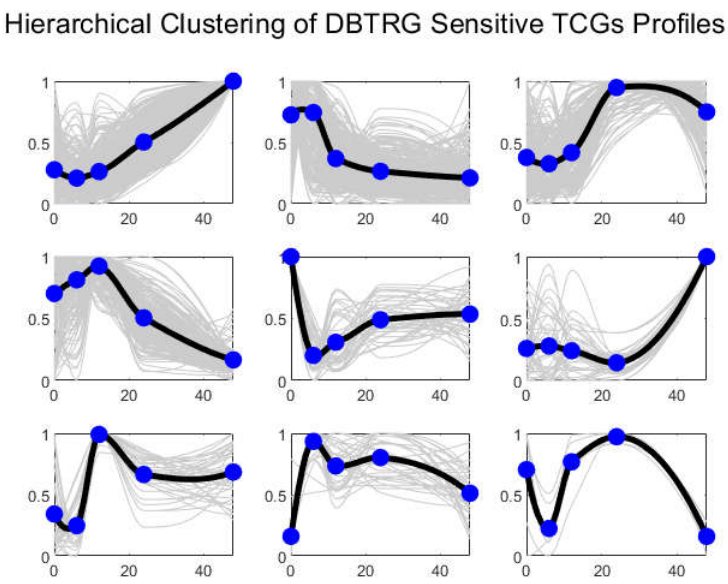


Expression Profile of DBTRG (drug-sensitive)

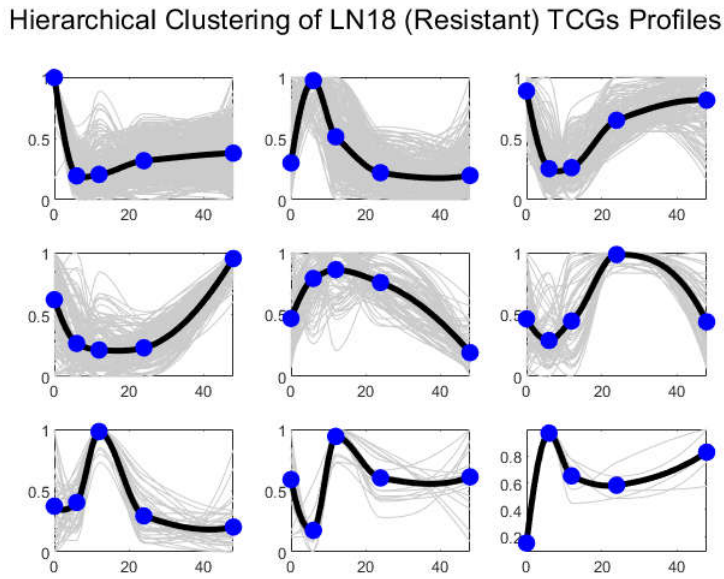


Expression Profile of LN18 (drug-resistant)

I cluster TCGs using the hierarchical clustering result above for both sensitive and resistant cells. We can see that the effect of going back to the original expression level could be observed in resistant cells, while these effect is less obvious in sensitive cells. Although there is kind of bounding back phenomenon in sensitive cells, the final expression level is more away from the starting point than those of resistant cells.



Expression across time of sensitive TCG clusters

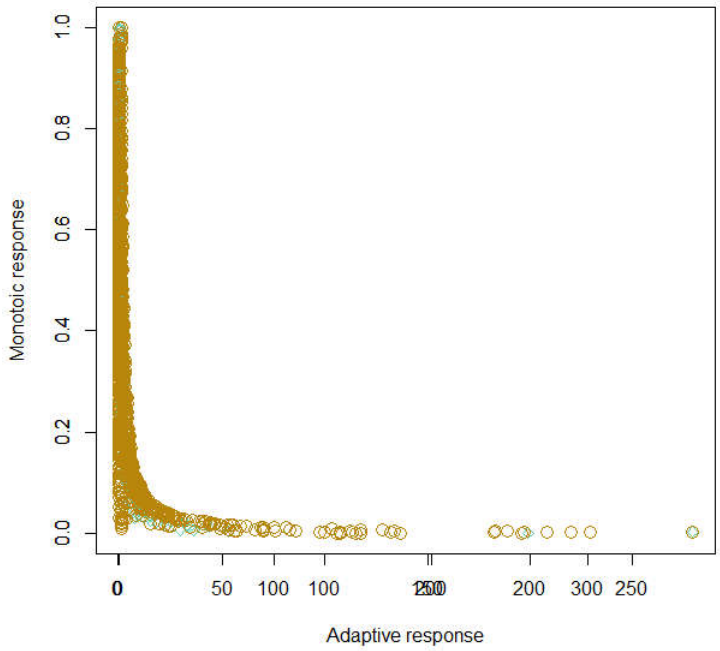


Expression across time of resistant TCG clusters

Measurement of Monotonicity and Adaption

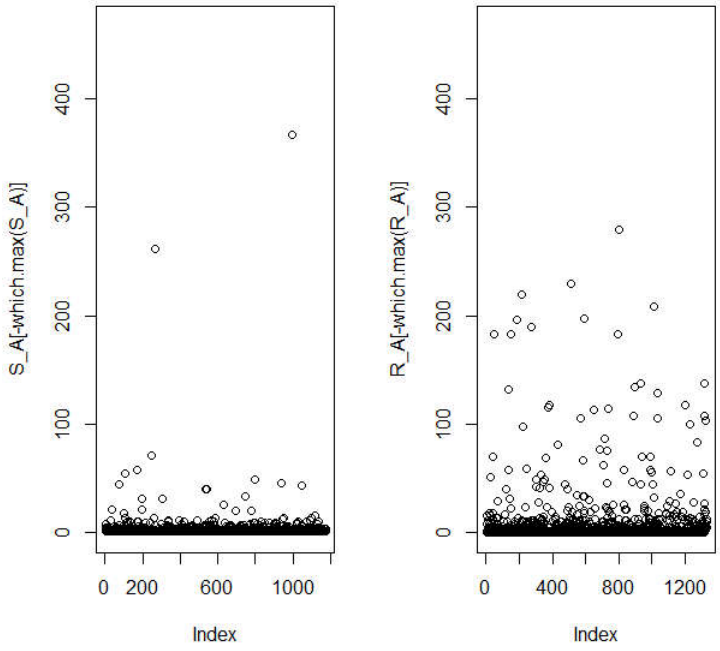
To more clearly see the effect bounding back of resistant cells (adaption), we define monotonicity which measures how the ending point is compared to the max point, and adaption which measures how the overall change (difference between start point and end point) is compared to the max deviation from start point.

From the figure, we do see there are more points with higher adaptive scores for resistant cells, while most of them show strong monotonicity.



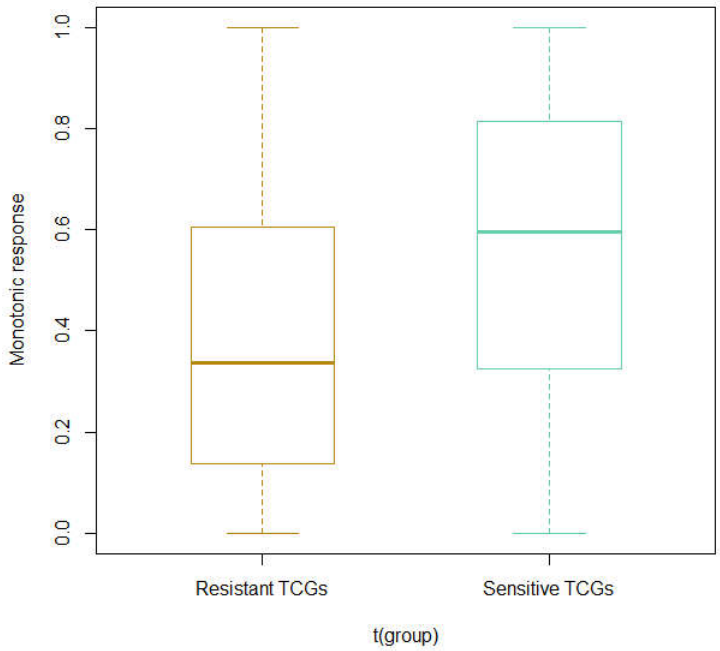
Monotonicity response and Adaptive response. Green is sensitive cells and brown is resistant cells.

This could be clearly see from the plots below.

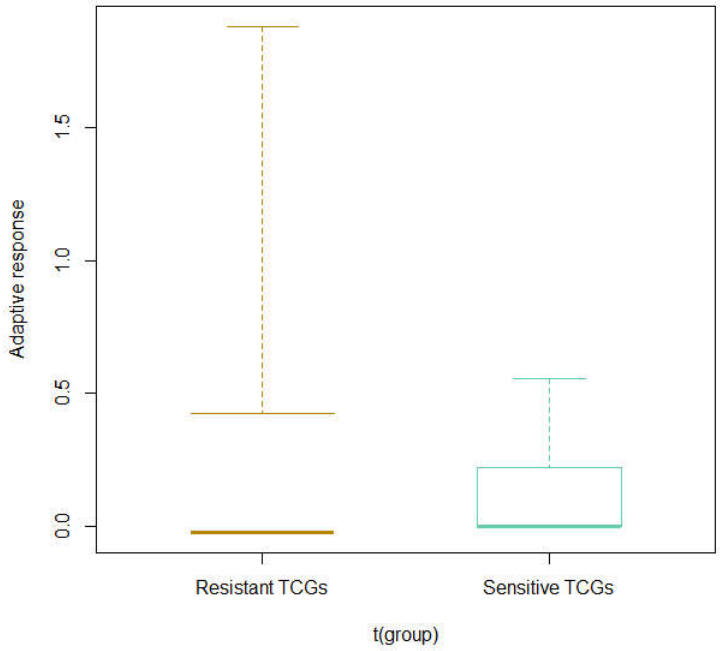


Adaptive Response. S represents sensitive cells, R represents resistant cells

However the difference on adaptive scores between the two cell lines could not be captured when performaing Wilcox test. No significance is shown on adaptive scores but on monotonic scores.



Mono response which shows significance



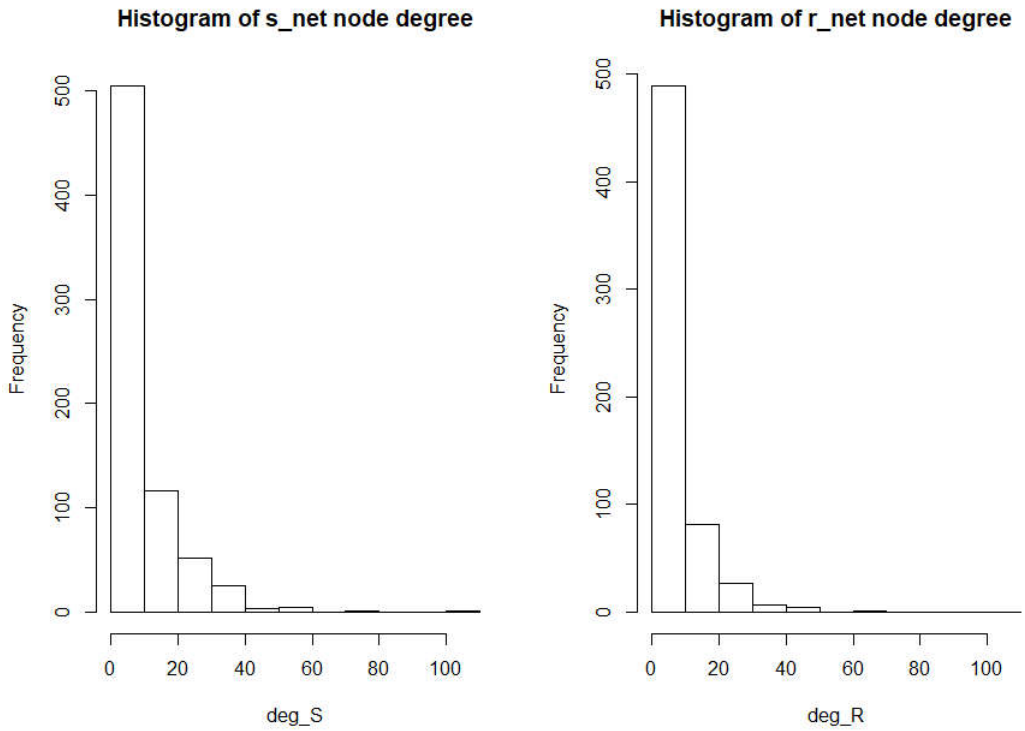
Adaptive Response which has no significance

Network Modeling

Implement gene-wise correlation test to each gene pair in both sensitive and resistant cells.

Transform TCGs gene ID into HGNC symbols, use the list of these symbols (proteins) to gain protein-protein interaction on STRING website. Based on the PPI data, correlation coefficients and pvals, construct gene interaction network.

After network building, degree of node are calculated for both network and plot histogram to show their distributions. From the figure, we can't see that there is more high degree nodes in resistant network.



Histogram of degree of nodes

Discussion

Although at first the contamination and overrepresented reads bothered me, when inspecting the alignment summary, there are in fact high mapping rate across all samples. Those 3-4% unmapped may due to the contaminous Mycoplasma and the duplicates only take up around 1%. A BWA mapping is better implemented on the Mycoplasma reference genome to see the influence of contamination.

The problemetic resistant_48_1 (SRR8769949_1) has a relative high N content at ceratin positions, that why it has a low duplicate level and overall lower quality. However this problem is produced by the machine itself and there is something wrong with the lanes. It doesn't not influence the mapping when comparing the mapping rate, read depth and coverage with others. STAR would treat this N bases as mismatch and could still be able to deal with those, according to STAR default setting that minimum overlap is 3 bases.

From the PCA on the whole samples, cells are correctly grouped by their cell lines but not by time course. This can be explained by the the fact that different cell lines act differently to the drug treatment, that surely they would not clustered together. More importantly, the expression profile is like a characteristic of cell lines, each cell line has its own profile due to the micro-environment around, and only a small portion of genes would be greatly affected by the treatment. This effect could be obscured by tiny change of other genes. That is why we should try to find out genes that most likely be responsible to the drug resistance.

From PCA on each cell lines, we can see how state changes across time and observe the bounding back tendency of resistant cells. The PCA step resembles picking up the first few genes that can represent the whole gene set. While it somewhat has an effect of finding the responsible genes, it still represents the whole gene set and thus, more rubust than the exact resistant genes. This is the easier way we observe different pattern between sensitive and resistant cells, but it is still not obvious.

The idea of finding TCGs is based on knowledge that genes targeted by the drug and genes that regulate them are among the most likely to be changed. So through setting expression threshold and fold change filtering conditions, TCGs are selected and clustering method followed to show clearer profiles of the rubustness in resistant cells.

The author defines two formula to measure monotinicity and adaption (robustness) of the two cell lines. I tried the same method and didn't get a good figure as that in the article. It turns out that they made a mistake in their codes by mixing up sensitive and resistant cells. In fact, my results show that sensitive cells do show a stronger monotinicity however the adaptive effect is not significantly different between the two groups. By simple plotting, we do see resistant cells tend to have more genes that have higher adaption scores, but is just a little higher and only take up a small portion. I think this may due to the definition of adaption score, which would be extremely large when there is almost no change across 48h. A better measurement or scaling strategy should be considered.

In the original article, the author continues to construct networks and measure the complexity of them. They did a good job by integrating PPI knowledge with statistical test. However, they claim that they define positive coefficient as enhancing effect and negative one as inhibiting effect, which seems reasonable. I don't think the regulatory relationship could be decided by this way, as the protein interaction doesn't necessarily means that their genes are regulated as the way that coefficients indicate, not to mention they further use this reconstruct network to count how many feedbacks are in the network and thus measure the robustness.

So I build the network and ignore the sign of coefficient, every gene pair that has a coefficient larger than 0.75, smaller than a 0.05 cutoff and show interaction in PPI would be linked together. Degree of nodes in both network are calculated but I don't expect that results show sensitive network have a higher number of high degree nodes than resistant. This happened when those PPI genes are not well correlated with each other, and the reason may be that performing linear regression on time-course data is not a good idea, since the expression of a gene across time would always not be normal.

Conclusion

Gene expression profiles are different between drug-sensitive and -resistant cells. There is tendency of bounding back to the orignial expression level in resistant cells.

In this project, I performed quality control on all samples and tried to assess their availability. Think carefully about the influence of contamination and overrepreseted sequence. Use exploratory methods to inspect data including PCA, clustering and heatmap. Design methods to evaluate the effect of resistance, including PCA, clustering and time-course expression profiles.

Although so far not successful, try to use Monotonicity scores and adaptive scores to distinguish the two cell lines. Build networks on both cells and evaluate their degree of nodes.