Final report

Andreas Breuß, Anastassia Fink, Leon Hornich, Yuan Sun

19.07.2021

Introduction

Tissue-restricted antigens (TRAs) are antigens which are highly expressed in a specific tissue and are otherwise mostly absent in different body tissues. The only other tissue where all TRAs are expressed in the body are the medullary thymic epithelial cells (mTECs) during so called promiscuous gene expression (pGE). pGE is an important step in the process of negative selection of T-Cells. The negative selection of autoreactive T-Cells is a vital step to prevent autoimmune reactions in humans. (Kyewski and Derbinski, 2004)

Some TRAs are overexpressed in cancer for unknown reasons. Because they are otherwise strictly regulated and only expressed in specific tissues, this makes them great targets for cancer immunotherapy.(Kyewski and Derbinski, 2004; Rosenberg, 1999)

In this context we want to look at the casein gene family. The casein genes (CSN) are a gene family located in the casein gene region on the longer arm of chromosome 4 in humans. (Rijnkels, 2002)

The casein Gene family contains the five genes *Csna*, *Csnb*, *Csng*, *Csnd* and *Csnk*. The CSN genes are regulated by hormones to be temporarily expressed during late pregnancy and the postpartum lactation period. The expression of CSN genes is spatially restricted to mammary gland epithelial cells (MECs). (Derbinski et al., 2007)

In the MECs of lactating female mice the expression of CSN genes is correlated with other functionally related genes.These functionally related and coexpressed genes are the milk protein genes lacalbumin-α (*Lalba*), whey acidic protein (*WAP*) and the transcription factor *Elf5* (E74-like factor 5). *Elf5* is involved in the regulation of *WAP* expression. In mice the CSN genes are located in the CSN gene region on chromosome 5. However, *Lalba* is located on chromosome 15 and *WAP* on chromosome 11. Thus, these genes are coexpressed, but not colocalized. The genes *Sult1d1* (sulfotransferase family 1D, member 1) and *Odam* (odontogenic ameloblast associated) are colocalized with the CSN genes in the CSN gene region on chromosome 5 in mice. However, the expression of these colocalized, but not functionally related genes *Sult1d1* and *Odam* is not correlated with the expression of CSN.(Derbinski et al., 2007)

In this project our primary goal was it to investigate the expression of CSN genes in tumorous tissue of lung cancer and four different types of breast cancer. In this sense we wanted to determine if the CSN genes are up-regulated in breast or lung cancer, which would make them potentially good targets for cancer immunotherapy. Furthermore we investigated if the expression of casein genes is correlated with other functionally related genes like *Lalba*, *WAP* and *Elf5* in breast or lung cancer, similarly to their expression in MECs, or if it may show a correlated expression to colocalized but functionally unrelated genes like *Sult1d1* and *Odam*. If CSN genes are up-regulated in one of the cancer types this would give us a first hint about the way how they are up-regulated in the cancer cells.

We analyzed two datasets. The first one contained the gene expression data of cancerous and non-cancerous tissue from small cell lung cancer patients. The second dataset was build from the gene expression data from patients of the four different breast cancer types HER2, LumA, LumB and TNBC.

Small cell lung cancer (SCLC) is classified as a high-grade neuroendocrine tumor. SCLC usually occurs in heavy smokers, it has a very bad prognosis and currently no targeted therapy is available. It is assumed that the inactivation of *TP53* and *RB1* are the initiating mutations which lead to SCLC.(Gazdar et al., 2017; Cai et al., 2021)

HER2 positive breast cancer is characterized by the overexpression of human epidermal growth factor receptor-2 (HER2) on breast cancer cells. Because of the overexpression of HER2, HER2 positive breast cancer can be treated with antibodies against HER2. (Maubant et al., 2015)

In luminal A (LumA) and luminal B (LumB) breast cancer the estrogen receptor and the progesterone receptor is expressed, as well as other hormone receptor-related genes. This expression allows for targeted therapy with tamoxifen. (Maubant et al., 2015; Maire et al., 2013)

Triple negative breast cancer (TNBC) is characterized by its lack of estrogen receptor and progesterone receptor expression and no overexpression of HER2. TNBC describes a group of different breast cancers with these characteristics, which usually have a poor prognosis because of its high proliferation rate and genetic instability. Additionally, currently no targeted therapy is available for TNBC, which makes it even more interesting for exploration of possible therapeutic targets. (Maubant et al., 2015; Maire et al., 2013, Maire et al., 2013)

Methods and results

The breast cancer dataset includes 20 Chips of Breast cancer samples with 5 for each different subtype, which were mentioned before. The lung cancer dataset consists of 6 patients of lung cancer samples which bare 2 chips per patient, and as mentioned: one cancerous and one non-cancerous sample and therefore 12 chips in total. First, we started with implementing the raw data from the CEL files into our r file. For this task we had to know which specific microarray chip we are using (HGU133PLUS2_Hs_ENST) since they have different alignments. So, we had implemented the referring libraries hgu133plus2hsenstcdf and hgu133plus2hsenstprobe after we installed the BiocManager package which includes the affy, vsn and AnnotationDbi package. These packages are used for the normalization of the data into a matrix and are also used for exchanging the affy ID´s with the relating gen name or so-called gen symbol. After displaying the CEL files we did not discover any abnormal patterns at first sight and continued after saving the output as pdf to ensure the scale stays in a 1:1 ratio since r studio tends to automatically review it rectangular. In the next step a short data clean-up was made, generally looking for NA´s or corrupted data. As far as the data clean up shows, no missing values were found and the vsn normalization of the data was performed and a matrix was created for the later use.

```
#normalization

#breastcancer
data_b.norm <- vsnrma(data_b)
data_b_vsnrma_matrix <- exprs(data_b.norm)
```

Before the date is used for statistics the quality of the data must be ensured and therefore, we ran several quality controls at first. Starting with a meanSd plot looking for outliers. We also did compare the expression values of the different chip's trough boxplots, before and after normalization to have a comparison since running only one of them could lead to missing some information. Out of these 2 plots we assumed that the data from chip 9 of the lung cancer dataset could be of lower quality.

For further proof of quality of the rtPCR, two rna degradation plots each were applied, one shifted and scaled and the other one just scaled. Here again chip 9 as mentioned before showing abnormal behaviour. Last we compared the different chips of each dataset trough scatterplots. After observing the scatterplots we finally decided to get rid of chip 9.

Subsequential the expression values were extracted, the x_at suffix was removed and the ensembl.103.txt file was read into our r file to provide the missing gen symbols trough the ensemble ID as mentioned before.

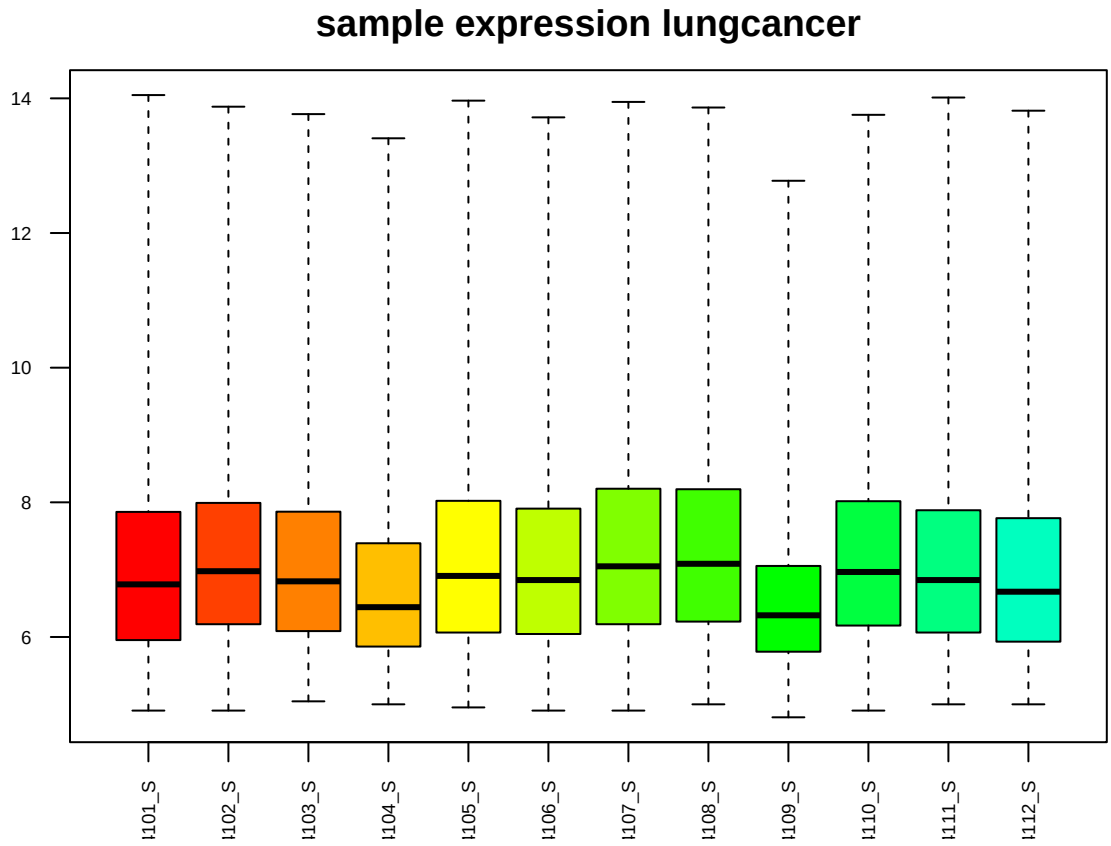


Figure 1: Qc Boxplots

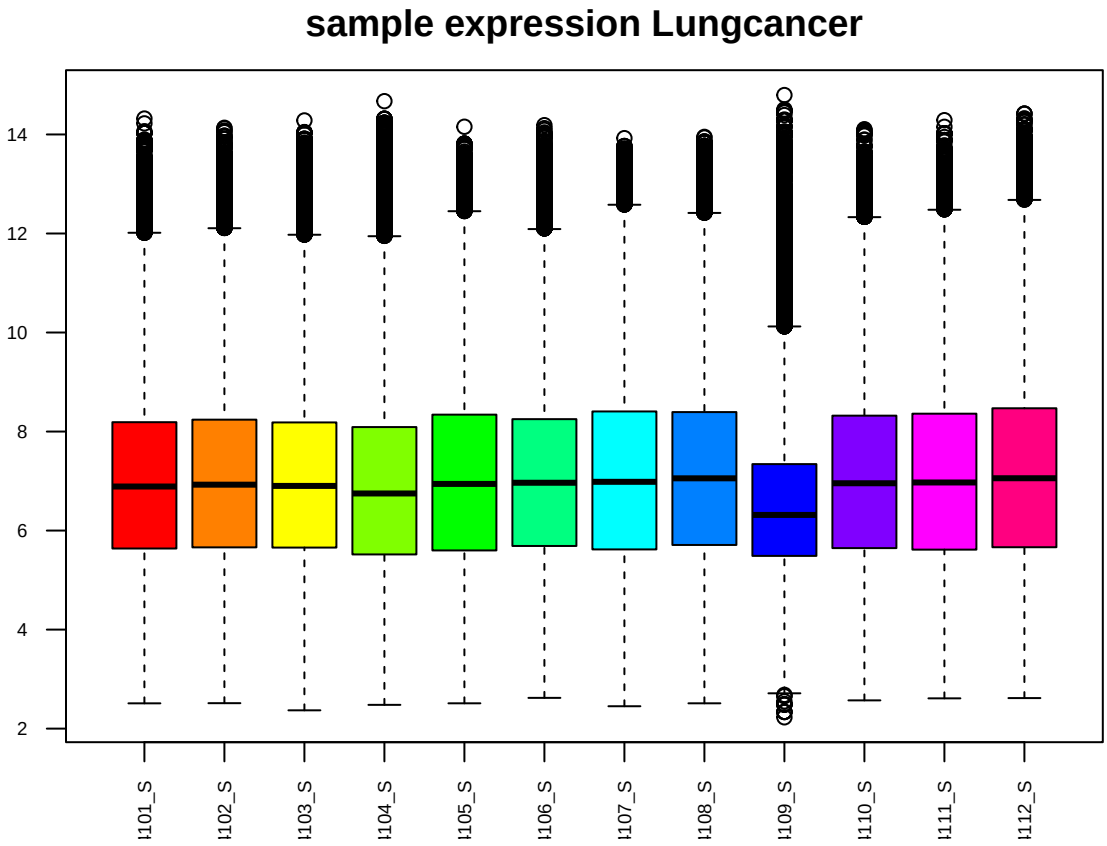


Figure 2: Qc Boxplots

We also observed the tissue distribution of csn genes and did find out that most of the transcripts we found in testis tissue in humans, which is kind of unusual and what we will go further into it in the end of the report.

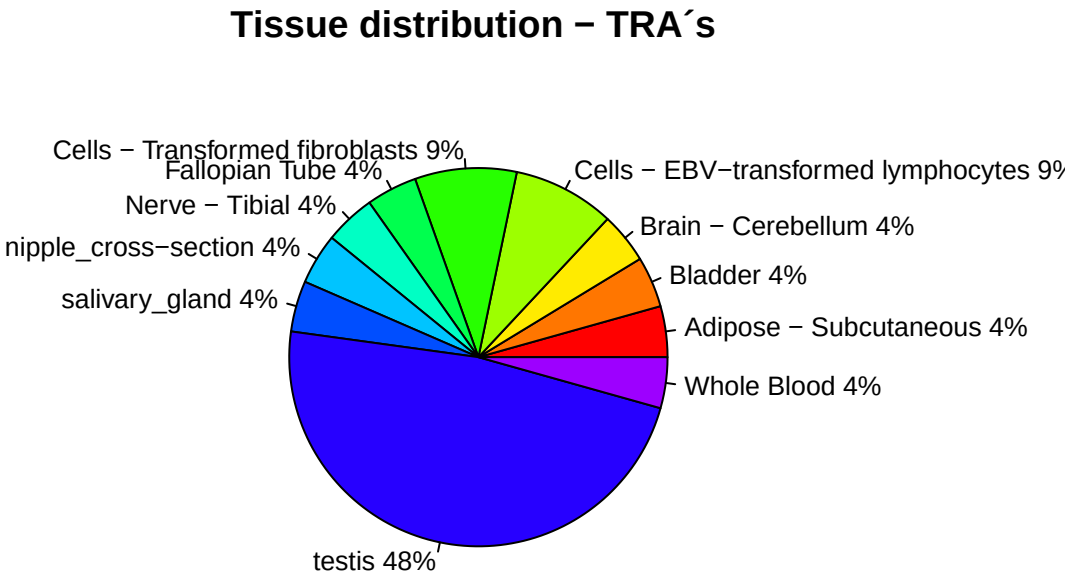
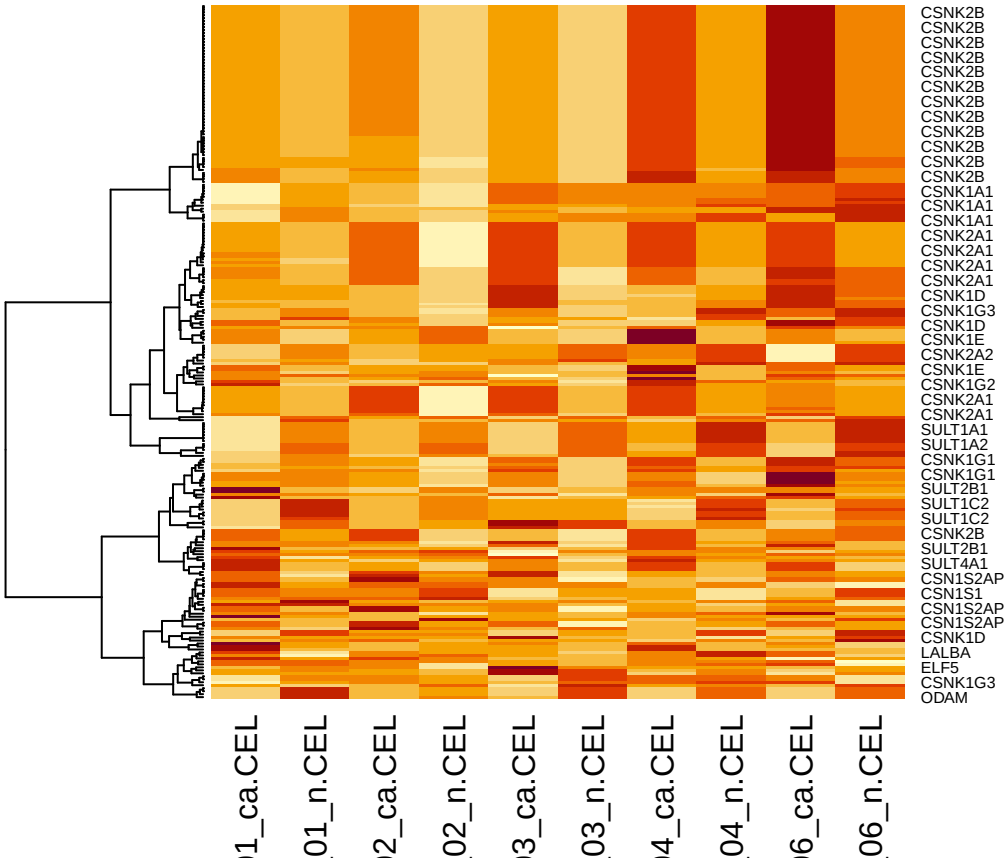


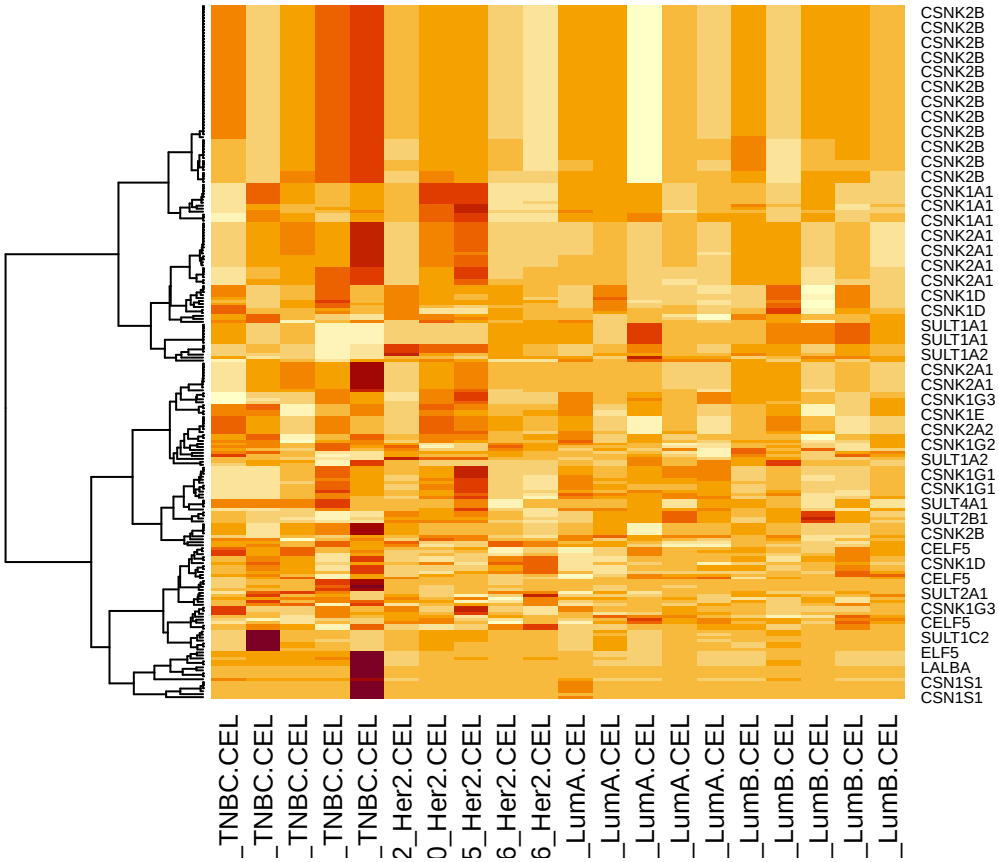
Figure 3: TRA tissue distribution

Heatmap

In order to find some structure in the gene expression value we made a heatmap. Due to the high amount of transcripts it is not possible to show the row names. Here we can see the distribution of the gene expression value, it is clearly visible that in the heatmap from lung cancer dataset that the gene CSNK2B in the cancerous group has a clearly higher expression value in cancerous group as the non-cancerous group. Beside that we also find out that the gen CSNK2A1 have also a higher expression value in cancerous group as in the non-cancerous group. With this information we can continue with the t-test to validate our thought whether this genes are really higher expressed in cancerous group as in the non- cancerous group or we can investigate another CSN genes , so we can get an overview on the difference between the cancerous group and non-cancerous group regarding the CSN genes. Genes like ELF5, ODAM, SULTI show no significant difference in the expression pattern between the cancerous and non-cancerous group. So, we will discard these three genes in the following t-test

Note: GSM1588972_Tum02_Her2 was changed to GSM1588978_Tum02_Her2!!! as solution, because windows doesnt allow to disable autosorting. We decided not to cluster the x axis to have a better comparison between the cancerous and the non cancerous samples and also between the different subtypes of cancer.


CSNK2B
CSNK2B
CSNK2B
CSNK2B
CSNK2B
CSNK2B
CSNK2B
CSNK2B
CSNK2B
CSNK2B
CSNK2B
CSNK1A1
CSNK1A1
CSNK1A1
CSNK2A1
CSNK2A1
CSNK2A1
CSNK2A1
CSNK1D
CSNK1G3
CSNK1D
CSNK1E
CSNK2A2
CSNK1E
CSNK1G2
CSNK2A1
CSNK2A1
SULT1A1
SULT1A2
CSNK1G1
CSNK1G1
SULT2B1
SULT1C2
SULT1C2
CSNK2B
SULT2B1
SULT4A1
CSN1S2AP
CSN1S1
CSN1S2AP
CSN1S2AP
CSNK1D
LALBA
ELF5
CSNK1G3
ODAM
1_ca.CEL
01_n.CEL
2_ca.CEL
02_n.CEL
3_ca.CEL
03_n.CEL
4_ca.CEL
04_n.CEL
6_ca.CEL
06_n.CEL



Principal component analysis (PCA)

Principal component analysis (PCA) was performed for the lung cancer dataset and the breast cancer dataset. However before the PCA could be performed highly correlated variables, more specifically transcripts with highly correlated expression were filtered out.The threshold for correlation was set at 0,8.

After filtering of initially 181 CSN transcripts only 20 remained in the lung cancer dataset. With these remaining 20 transcripts the PCA was performed.

If we would want to perform a dimension reduction we have to look at the Eigenvalues and cumulative variance of the principal components (PCs). In this case we have 12 PCs. For the first 6 PCs the Eigenvalue is above 1 and the cumulative variance adds up to over 90%. This means, that with only 6 PCs we could explain over 90% of the cumulative variance of our lung cancre dataset.

However in this case we will only look at the PCA biplot, where the two PCs which explain the most variance are plotted against each other.

The biplot shows how the transcripts are correlated. Samples which group together are correlated. As we can see, the ellipses which mark the different groups are strongly overlapping and overall all data points are grouped together. This means that there is no significant difference between the group of healthy and cancerous cells in the lung cancer dataset.

Following the same steps we performed a PCA on the breast cancer dataset.

After filtering of the transcripts with a correlation above 0.8 19 transcripts from originally 181 transcripts remain. On these remaining 19 CSN transcripts PCA is performed.

In the breast cancer dataset the Eigenvalue for the first 6 PCs is above 1. However,the first 9 PCs are needed to explain over 90% of the cumulative variance.

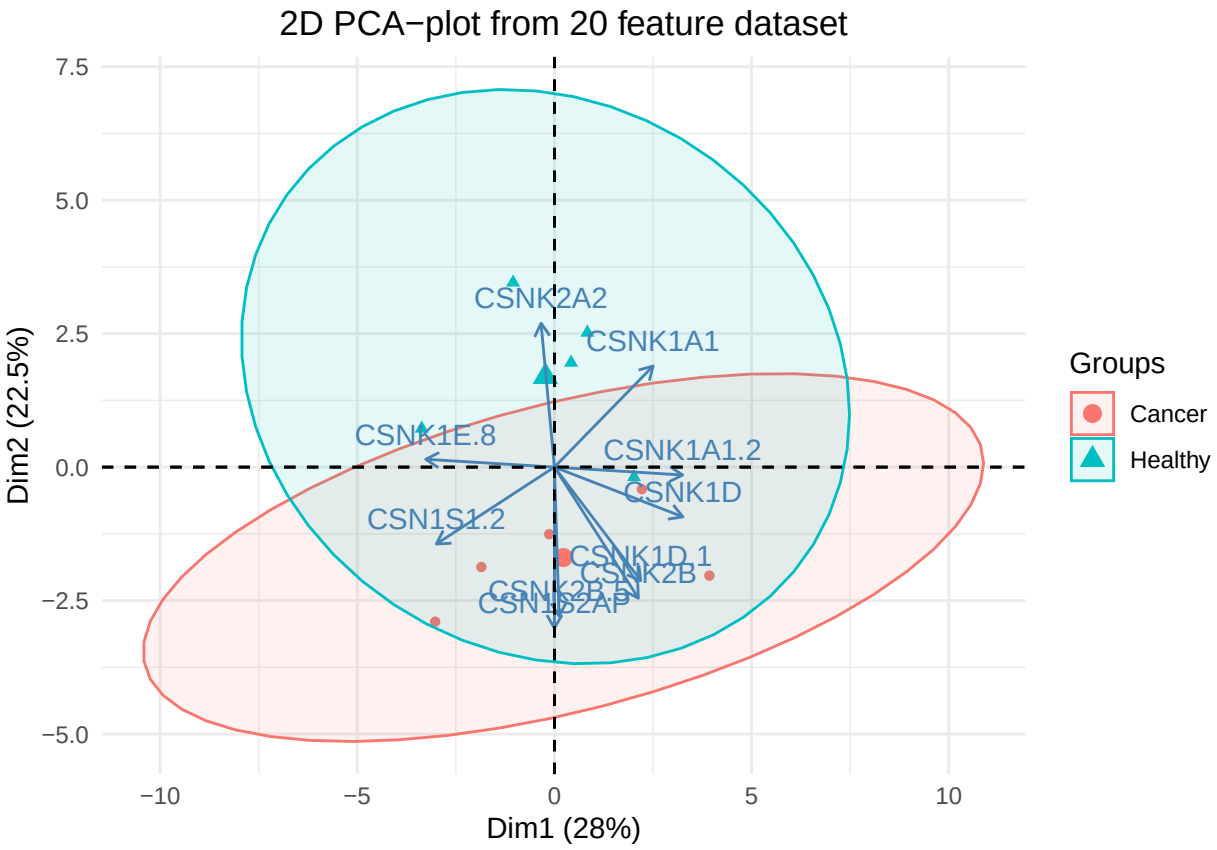


Figure 4: PCA biplot

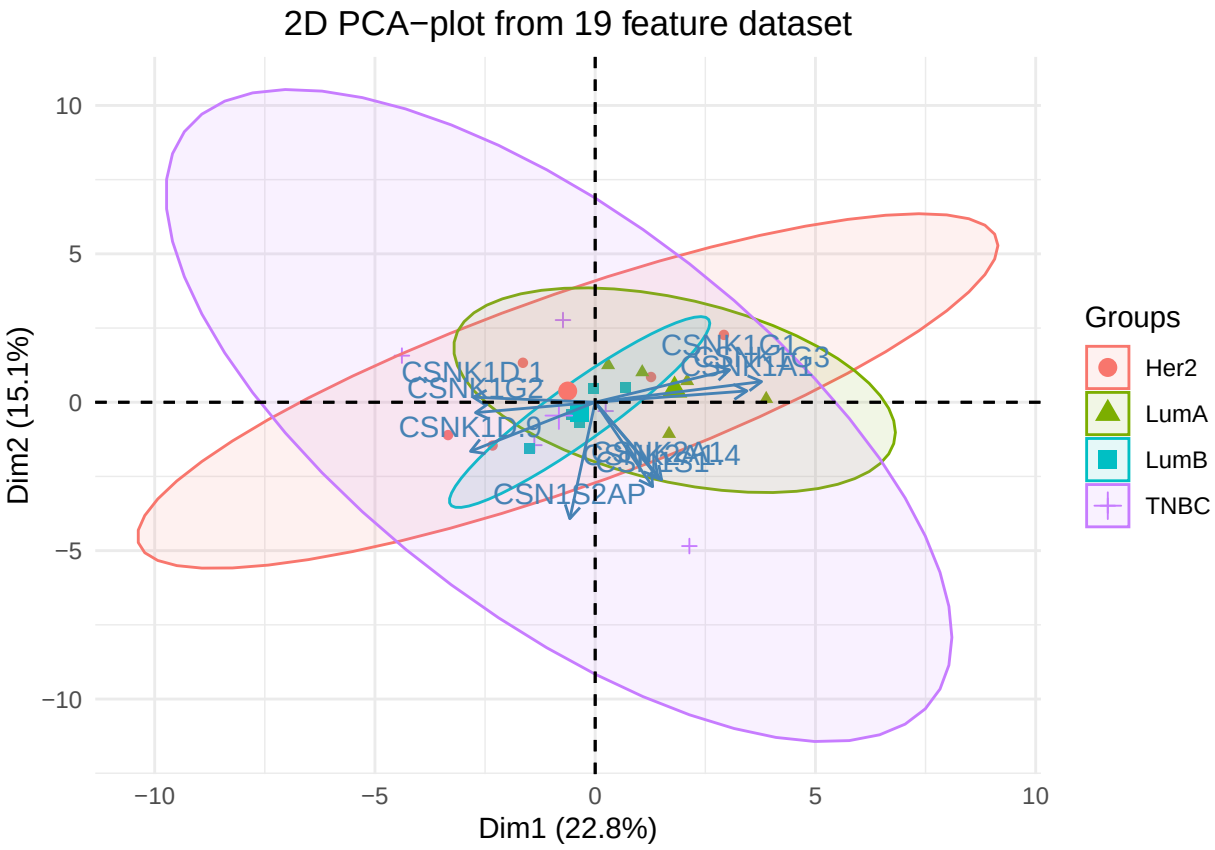


Figure 5: PCA biplot

For the breast cancer dataset the PCA biplot, similarly to the lung cancer PCA biplot, shows no clear distinction between the different groups. However, in this case we differentiate between the four different cancer types HER2, LumA, LumB and TNBC. As a result of our PCA we could not find any significant correlation effects between the different groups, neither in the lung cancer dataset nor in the breast cancer one. However, a possible effect cannot be ruled out entirely, because while all of the groups overlapped mostly, they did not entirely overlap. That means with a bigger dataset the groups could be more distinct.

T-test

In order to find out some relationship in the lung cancer data type, we use t-test to prove whether there is significant difference between the different genes. we set the ho hypothesis: The cancerous group has overall the same expression value in comparison to the non-cancerous group. To verify this hypothesis we divide the lung cancer data type in cancerous and non-cancerous chips. so we have two group of chips, in one group there are cancerous chips and in another group there are only non-cancerous chips. There are ten CSN genes which we have interest to find out the relationship between the cancerous and non-cancerous group .The expression value of every single CSN are extracted , and we compare directly the expression level between the two groups. We perform the t-test using this code.

The results show that gens CSNK2A1,CSN1S1,CSNK1G2,CSNk1E,CSNK1G1,CSNK2B has a p value under 0,05, so that we can say that this gen in cancerous group are higher expressed than gen in non-cancerous group. The results are also acceptable because out of 10 gens 6 of them have a higher expression value than the non-cancerous group. The other genes have higher p value as 0,05 meaning we can not criticize the h0 hypothesis that the expression level of gen between this two groups is same. so it could be important for the researchers to look closer into the expression pattern of CSNK2A1 gen ,the reason why this gen is higher expressed in lung cancer patient.

looking at the results from t-test we can get to the conclusion that the hypothesis that the gene in cancerous group should have a higher expression value than in non-cancerous group is valid. The question is why there are still four genes left which manifest the same expression pattern both in cancerous group and in non-cancerous group. it may be caused by the calculation mechanism by the t-test where the mean value are compared with each other.In this case a abnormal value which should have led to difference could be ignored if they are all averaged.

The other data type is breast cancer. there are 20 chips with 4 different type from breast cancer, so we want to compare the expression pattern in these 4 different types from breast cancer. In order to do this, we use one way anova test. in this case, we find out that the 10 genes we have fast the same expression level because the p value are all higher as 0,05. It is not astonishing because we are comparing the different breast cancer types.

Regression model

The last method used is a linear regression model to evaluate if there are significant correlations between the casein genes and the genes Sult1A1, ODAM, ELF5. In this last step every csn transcript was compared with one of the before mentioned genes. The F-test was used to determine if the thesis is valid or not. In a boxplot diagram we compared if there is a visible tendency for the different genes. Earlier we saw that a lot of transcripts have almost equal expression values this has to be considered. The problem is if on gene has more equal transcripts than every other. That could mean these transcripts could outweigh the other genes easily if an effect occurs. A solution would have been to use a median value out of this data, but we were told to work with all transcripts. The lung cancer dataset does show some correlation between CSN/ODAM but the other genes dont. On the other hand-side the breast cancer dataset shows some significant tendencies towards CSN/LALBA and CSN/ODAM. Every time about 50% of the p-values are lower then 0,05. Since this method uses all values of every chip to compare this could probably let us assume that the CSN/LALBA correlation is specific to breast cancer and is probably found in all 4 subtypes. Very interesting is the CSN/ODAM correlation which occurs in both cancer datasets. This is very

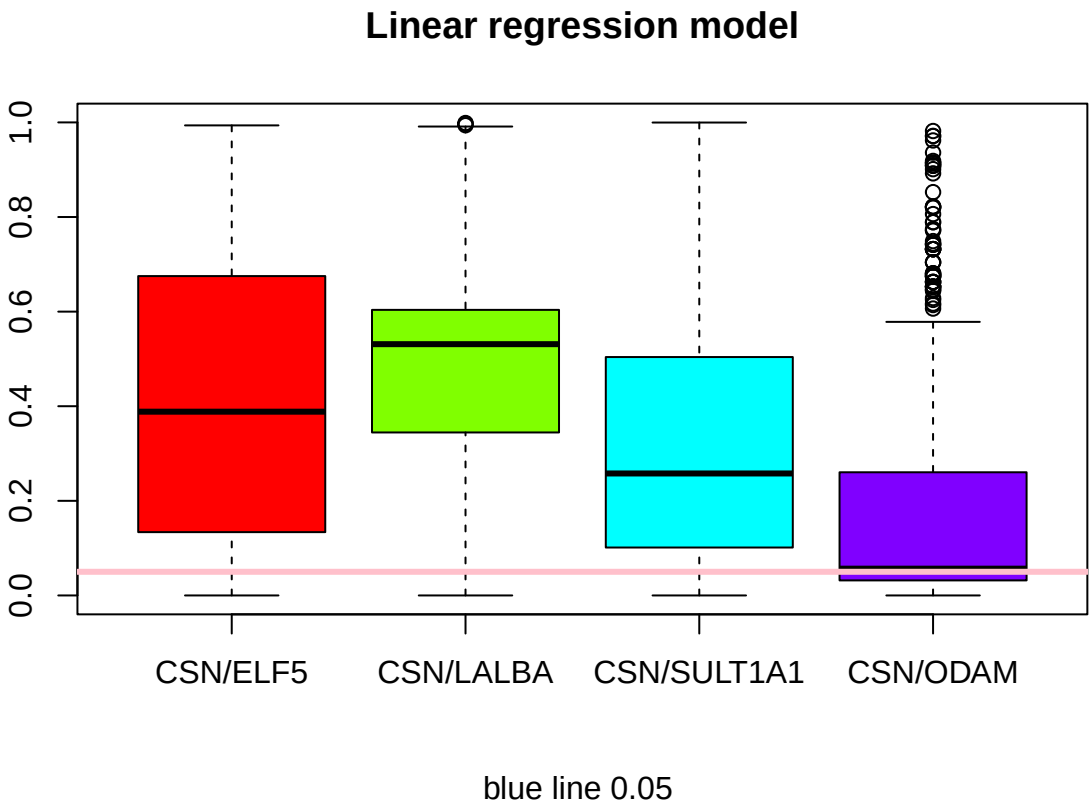


Figure 6: Regression lung cancer

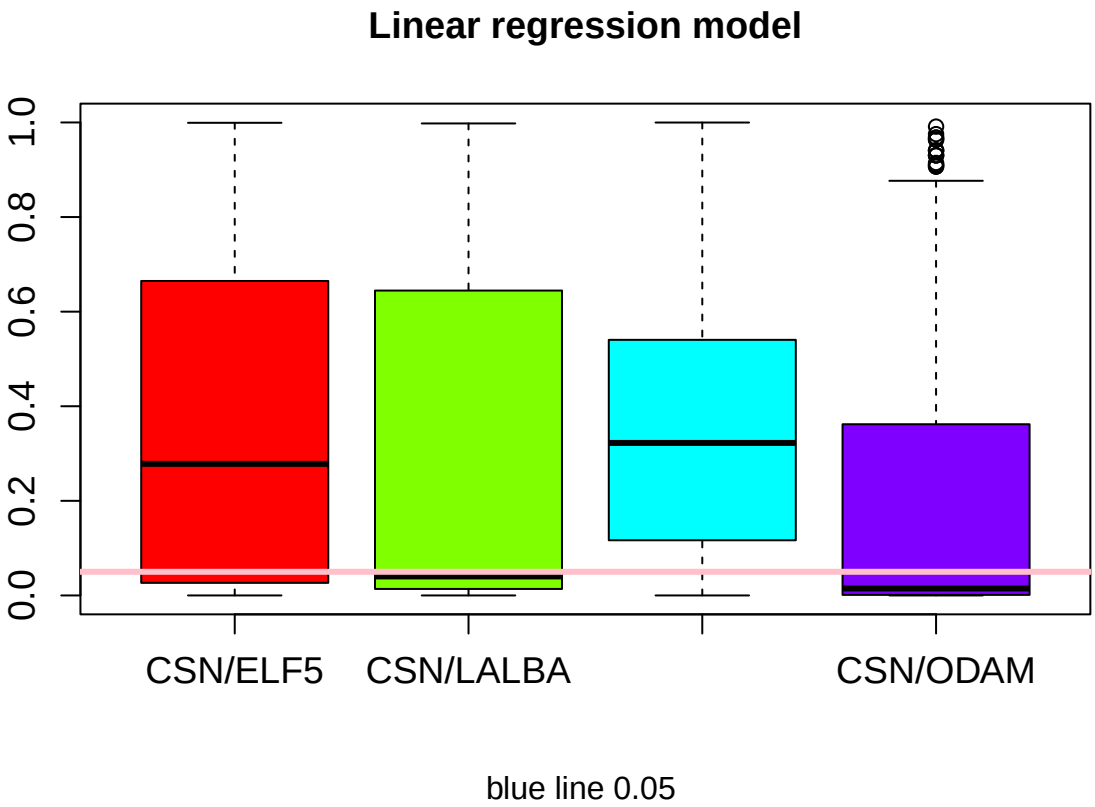


Figure 7: Regression breast cancer

interesting, because this one is a colocalized gene and usually dont show correlation to casein gene expression as mentioned in the introduction.

References

Cai, L., Liu, H., Huang, F., Fujimoto, J., Girard, L., Chen, J., Li, Y., Zhang, Y.-A., Deb, D., and Stastny, V., et al. (2021). Cell-autonomous immune gene expression is repressed in pulmonary neuroendocrine cells and small cell lung cancer. Commun Biol *4*, 314.

Derbinski, J., Pinto, S., Rösch, S., Hexel, K., & Kyewski, B. (2008). Promiscuous gene expression patterns in single medullary thymic epithelial cells argue for a stochastic mechanism. Proc. Natl. Acad. Sci., *105*, 657-662.

Dinkelacker, M. (2019). Chromosomal clustering of tissue restricted antigens, Dissertation, University Heidelberg, Germany.

Dinkelacker, M. (2007). A database of genes that are expressed in a tissue-restricted manner to analyse promiscous gene expression in medullary thymic epithelial cells. Diplomarbeit, Albert-Ludwigs-Universitaet, Freiburg, Germany.

Gazdar, A.F., Bunn, P.A., and Minna, J.D. (2017). Small-cell lung cancer: what we know, what we need to know and the path forward. Nature reviews. Cancer *17*, 725-737.

GTEx Consortium. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science, *348*, 648-660.

Lattin, J.E., Schroder, K., Su, A.I., Walker, J.R., Zhang, J., Wiltshire, T., Saijo, K., Glass, C.K., Hume, D.A., and Kellie, S., et al. (2008). Expression analysis of G Protein-Coupled Receptors in mouse macrophages. Immunome Research *4*, 5.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., and Young, N., et al. (2013). The Genotype-Tissue Expression (GTEx) project. Nature Genetics *45*, 580-585.

Maire, V., Baldeyron, C., Richardson, M., Tesson, B., Vincent-Salomon, A., Gravier, E., Marty-Prouvost, B., Koning, L. de, Rigaill, G., and Dumont, A., et al. (2013). TTK/hMPS1 is an attractive therapeutic target for triple-negative breast cancer. PloS one *8*, e63712.

Maire, V., Némati, F., Richardson, M., Vincent-Salomon, A., Tesson, B., Rigaill, G., Gravier, E., Marty-Prouvost, B., Koning, L. de, and Lang, G., et al. (2013). Polo-like kinase 1: a potential therapeutic option in combination with conventional chemotherapy for the management of patients with triple-negative breast cancer. Cancer research *73*, 813-823.

Maubant, S., Tesson, B., Maire, V., Ye, M., Rigaill, G., Gentien, D., Cruzalegui, F., Tucker, G.C., Roman-Roman, S., and Dubois, T. (2015). Transcriptome analysis of Wnt3a-treated triple-negative breast cancer cells. PloS one *10*, e0122333.

Rijnkels, M. (2002). Multispecies comparison of the casein gene loci and evolution of casein gene family. J Mammary Gland Biol Neoplasia *7*, 327-345.

Rosenberg, S. A. (1999). A new era for cancer immunotherapy based on the genes that encode cancer antigens. Immunity *10*, 281-287.

Roth, R.B., Hevezi, P., Lee, J., Willhite, D., Lechner, S.M., Foster, A.C., and Zlotnik, A. (2006). Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. Neurogenetics *7*, 67-80.

Su, A.I., Cooke, M.P., Ching, K.A., Hakak, Y., Walker, J.R., Wiltshire, T., Orth, A.P., Vega, R.G., Sapinoso, L.M., and Moqrich, A., et al. (2002). Large-scale analysis of the human and mouse transcriptomes. Proceedings of the National Academy of Sciences of the United States of America *99*, 4465-4470.

Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., and Kreiman, G., et al. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. Proceedings of the National Academy of Sciences of the United States of America *101*, 6062-6067.

Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., and Asplund, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. Science (New York, N.Y.) *347*, 1260419.