Klk_Genes_Data_Analysis

Anouk Dupe, Maria Yemane, Dustin Schilling, David Eckey

06.07.2021

Contents

1. Introduction

KLKs are a family of 15 mammalian secreted serine proteases. Analysis has shown that the KLK locus is most likely located on chromosome 19 and forms the largest cluster of contiguous proteases in the entire genome. (Yousef et al. 2000).
All 15 kallikrein genes are proteolytic enzymes of steroid hormone regulation and are involved in the regulation of blood pressure, tissue remodeling, skin desquamation, and many other processes. The structure of KLK are similar with two beta-drums, two alpha-helices and a distinct loop involved in the regulation of activity and selectivity. Currently, the specific role of each kallikrein is unclear. It is known that they are involved in the complex regulatory processes, more specifically in those different signaling cascades.
Dysregulation of KLKs are frequently associated with cancer. Their expression in different tissues and their involvement in different physiological processes make them potential tumor expression markers (Fischer and Meyer-Hoffert, 2013).
Differential expression of different kallikrein genes has been found in different cancer types. While clear cell and papillary renal carcinomas have similar kallikrein expression profiles, chromophobe renal cell carcinoma has a unique expression profile (Tailor et al. 2018).

In the following, a set of in total 32 microarrays will be analysed. 20 of those originate from patients with breast cancer derived from the data set GSE65216 (Maire et al. 2013), 12 from patients with small cell lung cancer from GSE149507 (Cai et al. 2021). In this report, both cancer types are analyzed seperatively and both their results will be discussed later on.

2. Quality control

After reading in the data, the first step is to verify its quality by following the steps presented in "R Course Micoarray Analysis" by Dr. Maria Dinkelacker (2019). The goal of quality control is to identify samples for which the data characteristics are significantly different. These differences would be difficult to remove via variance stabilizing normalisation (vsn) and could interfere with the rest of the data. Samples that show odd characteristics will thereby be identified and replaced in the following quality control. The quality control is performed on the breast cancer microarray data set GSE65216 (Maire et al. 2013) and the small cell lung cancer microarray data set GSE149507 (Cai et al. 2021).

2.1 Quality control - GSE65216 breast cancer

Upon examination of each individual array, there are no scratches or lighter areas detectable, which means the arrays themselves are fine. Furthermore, the meanSd plot is being used to verify the variance stabilization. Here, the red line, which stands for the running median, should be horizontal. However for the breast cancer data, it follows a linear relationship.

In contrast to that, the quality of the breast cancer data is assured by the other plots, which is the reason why our group did not choose other chips for the data analysis. For the boxplots, it is clearly visible, that the differences in intensities between arrays is strongly reduced after normalisation. The boxplots only show little fluctuation in gene expression for the 20 arrays after normalisation. In addition to that, none of the current chips deviate strongly from each other for the density and RNA degradation plot (before and after normalisation). Also, the scatter plots show linear relationships between each chip. The quality control plots can be seen in the github repository.

2.2 Quality control - GSE149507 lung cancer

For the small cell lung cancer microarray, an abnormality was dectable in the scatter plots. The carcinoma tissue sample of patient number 5 did not show linear relationship to all of the other samples. Since the samples of dataset GSE149507 for normal and carcinoma tissue are linked to one patient each, we

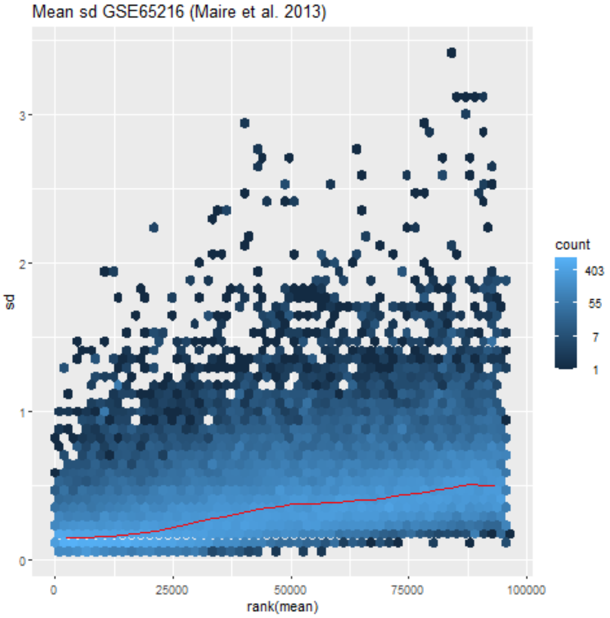


Figure 1: meanSD plot of breast cancer microarray GSE65216.

consequently replaced the two chips GSM4504109_SCLC_05_ca and GSM4504110_SCLC_05_n with new 2 new chips out of the gene expression omnibus. When performing the scatter plot control over again, there are no discrepancies.

Equivalent to the breast cancer microarray, there were no further abnormalities detectable for the lung cancer microarray.

3. TRA data

To distinguish between TRA KLK genes and non-TRA KLK genes, a total of 6 TRA data sets were utilized ((Su et al. 2002, 2004), (Roth et al. 2008), (Lattin et al. 2006), (human GTEX data 2015), (Uhlén et al. 2015)). These TRA data sets were than unified, which allowed the extraction of tissue-restricted KLKs after their transcription number.
To get an overview of the distribution of the Tissue Restricted Antigens, especially those that are Kallikrein genes, a pie chart was conducted. A pie chart in general allows a quick overview and a first assessment of numerical distribution values. In this pie chart, the distribution of Tissue Restricted Kallikrein-Antigens is displayed. Notable is the high prevalence of prostatic kallikrein genes, as well as an occurrence in esophagus, thyroid and salivary gland. Since six data sets were combined, annotations for the same tissue were different, which were fused manually.

4. Expression Analysis

4.1 Breast cancer GSE65216 (Maire et al. 2013)

The breast cancer microarray data GSE65216 (Maire et al. 2013) consists of 20 samples. The samples are all breast cancer tissue, differentiated by the four mutation types: Triple negative breast cancer (TNBC), Her2, Luminal A and Luminal B.

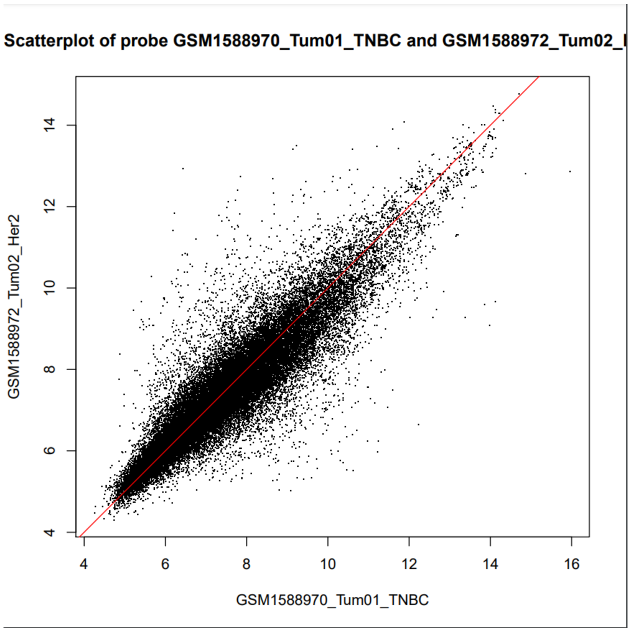


Figure 2: Example of scatter plot breast cancer GSE65216.

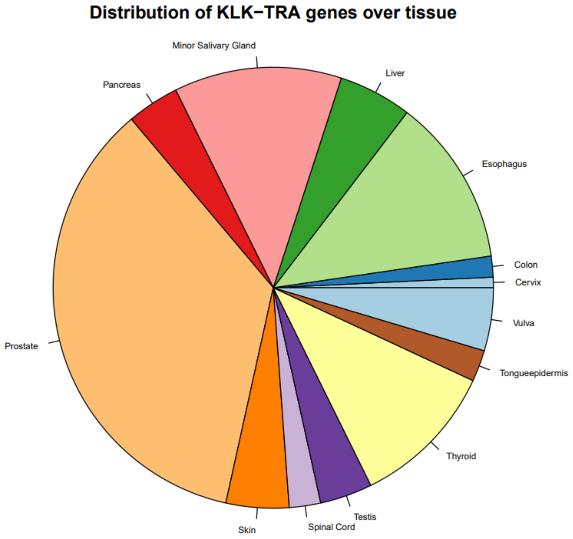


Figure 3: Tissue specificy of KLK genes - KLK genes from six TRA datasets are combined and sorted for tissue specificity

Clean up identical isoforms

Having a look onto the data set itself, it is noticeable, that some of the expression values of the KLK isoforms in the microarray are exactly identical. These identical isoforms should be cleared out to increase the real information value of the KLKs. This is easily done by conducting pairwise correlations, here the pearson correlation, between all KLK genes in a diagonal matrix. If the correlation yields a value of 1 then the two gene isoforms are the same and the latter one of both will be removed. Furthermore, the KLKs are sorted after their names in ascending order for the later visualization. In the end, 39 identical isoforms are removed, whereas a total of 73 KLK transcripts for the 15 KLK genes are being kept. Out of the 73 isoforms, 63 are TRAs, while only 10 are regarded as tissue restricted.

```
# transform dataset - genes on columns
df.TRA.KLK.breast <- data.frame(t(TRA.KLK.breast))

# Define function for calculating correlation
cor.genes <- function(df){
  df.cor <- cor(df, method = "pearson")
  diag(df.cor)=NA
  df.cor[upper.tri(df.cor)]=NA
  return(data.frame(df.cor))
}

# amount of identical columns
cor.TRA.KLK.breast <- cor.genes(df.TRA.KLK.breast)
length(which(cor.TRA.KLK.breast == 1))
```

```
## [1] 34
```

```
# cleanup function
genes.cleanup <- function(df){
  df[!duplicated(unclass(df))]
}

# remove identical columns
TRA.KLK.breast.clean <- genes.cleanup(df.TRA.KLK.breast)
dim(TRA.KLK.breast.clean)
```

```
## [1] 20 63
```

Overview gene expression

Since the data is vsnrma normalized, the median does not fluctuate too much for all samples. The lowest gene expression value of the first chip is roughly 6 while the highest gene expression value is around 16. Due to the logarithmic scale with a base of 2, gene expression is double as high between two samples if the log-fold change is +1.

```
gene.summary <- function(x){
 round(apply(x, 2, summary), digits = 2)
}
gene.summary(breastExprs)[,1]
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.26    6.28    7.24    7.50    8.43   15.96
```

The histogram represent the frequency of the present gene expression in breast cancer samples. It also shows, that the median gene expression of KLKs is much lower than the overall median gene expression. This means that most of the KLK gene expression is normally down-regulated in the perspective of the whole genome. (Yousef et al. 2004)

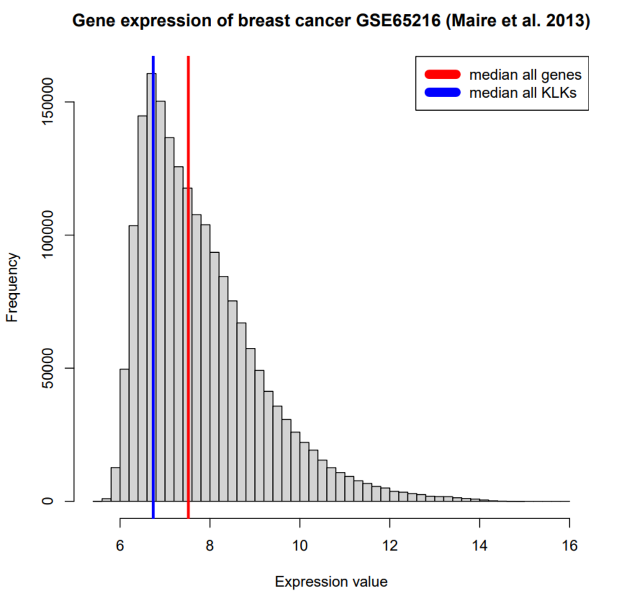


Figure 4: Histogram of breast cancer gene expression.

Boxplots

Here, the pattern for the fairly low gene expression of KLKs is also recognizable. Most of the boxplots of the single KLK gene isoforms are lower than the median expression of the whole breast cancer genome. Some KLK isoforms like KLK2.3 are even below a gene expression value of 6, so KLK isoforms like these are clearly down-regulated. This is compatible with the finding of Yousef et al. in which they state an overall down-regulation of KLK gene expression in breast cancer. There are only two isoforms that exceed the median of the whole genome expression of the breast cancer set. These are the isoforms KLK4.4 and KLK8.8.
KLK4 gene expression was found by Schmitt et al. to be up-regulated in breast cancer tissue as in comparison to healthy breast tissue. This seems to correspond with the finding shown in the boxplot, but only for the KLK4.4 isoform. Thereby, KLK4.4 needs to be looked on more carefully. In contrast to that, KLK8 seems to be higher expressed in both normal and cancer tissue.(Schmitt et al. 2013)

Heatmap

The dendrogram is the core for the emerging clustering in heatmaps.

Interesting in this respect is, that KLK4.4 forms its own branch independent of all the others. As already shown in the boxplots, KLK4.4 was distinctly up-regulated. Looking onto the other branches of the dendrogram, it is notable that besides KLK4.4 there are more possible clusters. To increase the clarity of the heatmap, KLKs are separated into 3 clusters. Optimal clustering via K-means will still be performed later on.

KLK4.4 clearly stands out (cluster 2) with an overall up-regulated gene expression across all sample. Furthermore, as it is annotated KLK4.4 belongs to the TRA group. Moreover, gene expression in cluster 1 is higher than in the third cluster. There are only few samples which seem to have up-regulated KLK isoforms.

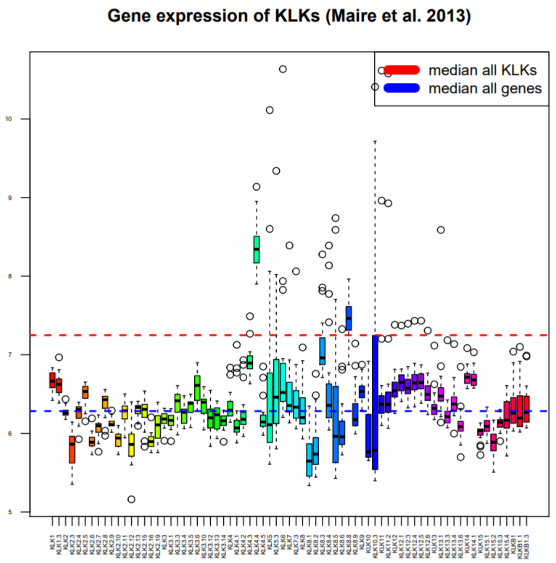


Figure 5: Boxplot of KLK gene expression in breast cancer.

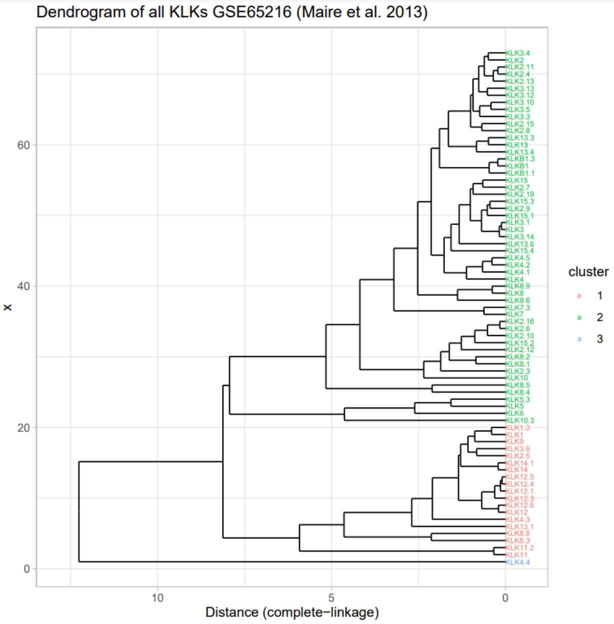


Figure 6: Dendrogram of KLK genes in breast cancer. Clustering is performed after the complete-linkage method. The genes are separated into 3 clusters.

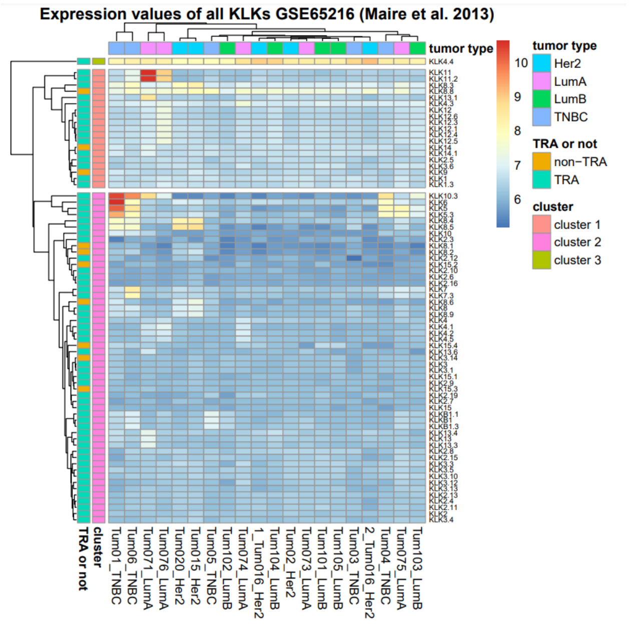
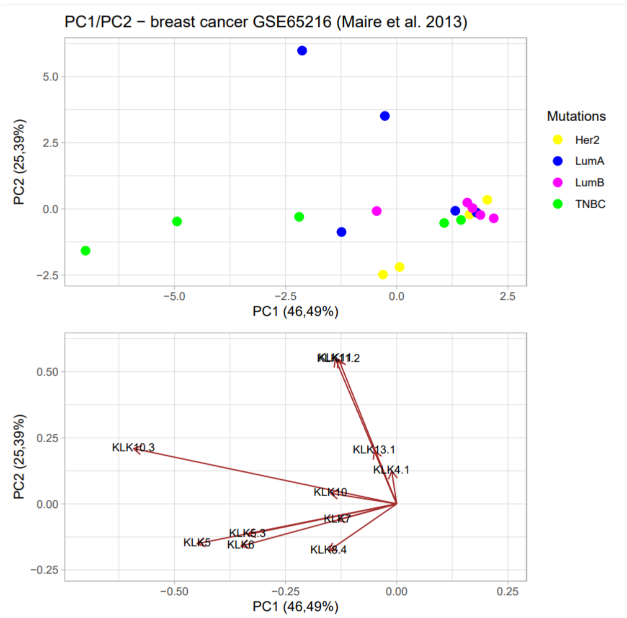


Figure 7: Heatmap of KLK gene expression in breast cancer. The samples are annotated corresponding to their mutation type. Additionally, the KLKs are differentiated by their cluster and potential tissue restriction.

For instance, the tumor sample number 1 and 6 (Tum01_TNBC and Tum01´6_TNBC) got one of the highest expression values across all the KLKs for KLK10.3, KLK6 and KLK5. As well as the tumor sample number 71 and 76 (Tum71_LumA, Tum76_LumA) for the transcripts KLK11 and KLK11.2.

PCA

The principal component analysis was conducted over the samples. Centering was enabled, while scaling was not included, due to the data being vsnrma normalized. The cumulative variance of the first two principal components (PCs) yield 72% of the total variance. Thereby, these two PCs explain 72% of the total information value. PC1 and PC2 are sufficient for the analysis. The breast cancer samples are distributed after their respective loadings of KLK gene expression.
\begin{figure}



{

}

\caption{PC1 (46,49%) is plotted against PC2 (25,39%). The upper part shows the distribution of the breast cancer samples annotated by their mutation type, while the lower part depicts the 12 highest loadings of the KLK genes.} \end{figure} The loadings consists of the the top 12 most differentiated KLK isoforms. This was conducted by adding the absolute values of the rotation matrix for each individual KLK isoform. As the PCA displays, some samples are more characterized by the expression of KLK11 and KLK11.2. This is mostly the case for two of the LumA samples. This was also observable in the heatmap by the higher expression of KLK11 and KLK11.2 for the tumor samples 71 and 76. Another finding of the PCA is that TNBC mutations are affected by KLK5 and KLK6 expression, also observable in the heatmap. Presumably KLK4.4 is not part of the top 12 loadings, since it is higher expressed across all tumor samples.

K-means clustering

K-means was performed to be able to draw conclusions on characteristics and the distribution of different Kallikrein genes. Here, the optimal number k of clusters was determined in doing a Within Cluster Sum of Squares – plot, also called the elbow method. A kink in the curve of a plot, in which the number k of clusters is plotted against the within cluster sum of squares, displays the optimal number of k clusters. Rising numbers of k will not cause a significant decline in the within cluster sum of squares anymore. For the breast cancer GSE65216 data set, the Within Cluster Sum of Squares – plot predicted an optimal number of $k = 6$, so k-means was performed using 6 clusters. The function of k-means automatically reduces dimensions to 2 if a dataset consists of 3 or more dimensions, so the k-means clustering does not directly cluster genes of interest according to their expression values, but cluster according to two dimensions that are influenced by expression values, as those explain most of the variance of the data. Due to this influence, k-means is still suited for clustering and thus comparing KLK expression patterns. Outstanding cluster is here marked as cluster 1 in green, containing KLKs4.4 and KLK8.8. Those genes already stood out in the heatmap analysis. Interestingly, cluster 4 is also very distinct but on the very other direction on the graph of two dimensions.

Hypothesis testing

WILCOX TEST WEITER AUSFÜHREN?! DATENSTRUKTUR ERKLÄREN?!

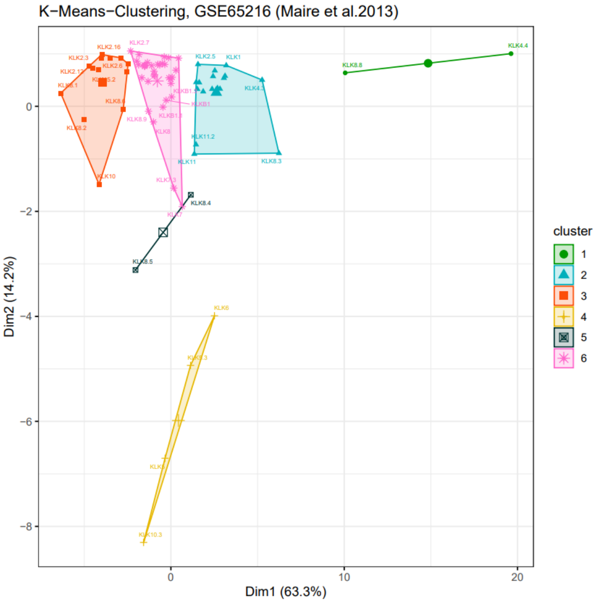


Figure 8: K-means cluster analysis with k = 6 clusters for the breast cancer data set

The expression values of the Kallikrein genes obtained from Marie et al. were not normally distributed. Therefore the non- parametric Wilcoxon-Mann-Whitney-Test was used.

(First, cluster 1 (KLK4.4 and KLK8.8) were tested for overexpression against all other KLKs individually from the dataset. KLK4.4 (TRA) was significantly higher expressed than all other KLK genes from the dataset. Likewise, KLK8.8 (non-TRA) was significantly higher expressed than all KLK genes, except KLK4.4. Those results conform with the observations from the heatmap and the k-means clustering. Cluster-4 (KLK5,5.3,6,10.3) was isolated in the k-means clustering. On the heatmap it was conspicuous that for some tumor types the genes of this cluster were higher expressed. Thus the upper-tail Wilcoxon-test was used. In contrast to Cluster-1, Cluster-4 could not be clearly identified as an overexpression cluster. KLK5.3 and KLK6 were higher expressed than two thirds of the other KLK genes, wereas KLK5,10.3 were not significantly higher expressed than most of the other KLKs.)

The main characteristic of the dataset from Marie et al. is the subdivision into the samples with different mutations (Her2, LumA, LumB, TNBC). In the heatmap four genes were identified, which were overexpressed in at least two of the five mutation specific mikrochips. KLK10.3,6 for Her2, KLK11,11.2 for LumA. Those four genes are also identified as one of the twelve main loadings of the PCA.

In Figure X, these genes are shown with the subdivision into the different mutation types. Significant expression differences between mutation samples are indicated with brackets.

In a previous study (Haritos et al. 2018) KLK6 expression was found to be generally downregulated in breast cancer tissue, but in HER2 and TNBC positive tumors KLK6 was overexpressed. Those findings are only reflected to a limited extend in this analysis. Only Her2 was found to be significantly higher expressed than LumA. TNBC was not significantly higher expressed compared to the other mutations.

Another study from Michaelidou et al. reported a higher expression of KLK8 in TNBC and Her2 positive tumors compared to LumA and LumB positive tumors. However, this analysis could only confirm significant TNBC overexpression for the isoform KLK8.5 compared to LumA and LumB. Nevertheless, the boxplots of KLK8.4 and KLK8.5 show the trend of Her2 and TNBC overexpression.

In contrast, a recurring pattern in figure X is the significant overexpression of TNBC in comparison to Her2. This observation includes KLK5, KLK5.3, KLK10, KLK10.3.

In summary, the conducted analysis could partially conform the findings from other research groups. The

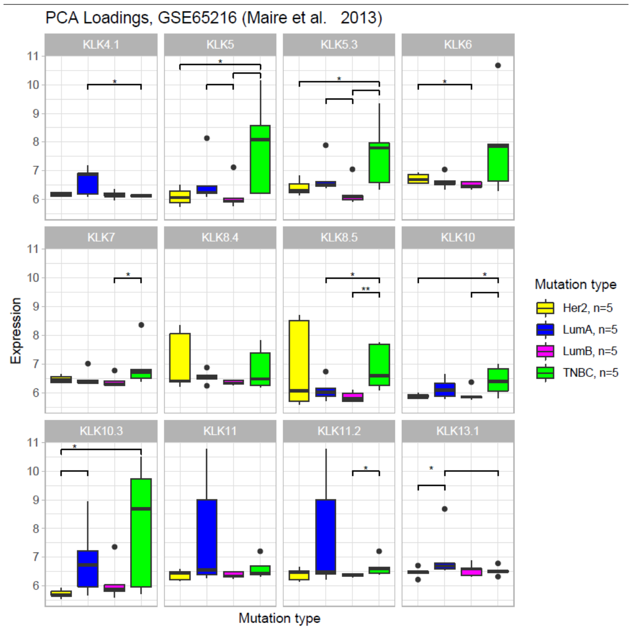


Figure 9: Panel plot of the PCA loading genes with significant bars. *: p-value <= 0.5, **: p-value <= 0.01

differences can probably be explained by the small amount of samples used in this analysis. In conclusion Kallikrein gene expression can be used for identifying tumor subtypes and even predict the outcome for a patient (Haritos et al. 2018).

4.2 Lung cancer GSE149507 (Cai et al. 2021)

The lung cancer microarray GSE149507 (Cai et al. 2021) derives from six patients with small cell lung cancer. The data set consists of a total of 12 samples. Carcinoma tissue and healthy lung tissue, which is adjacent to the carcinoma, make up 6 samples each.

Overview gene expression

Just as for the breast cancer data set, the median expression of the KLKs is beneath the the median of the overall gene expression, since KLKs are mostly down-regulated (Yousef et al. 2004). However for the lung cancer data set, it appears that the gene expression values are distributed more evenly, while the breast cancer histogram represents a right-skewed distribution.

Boxplots

Most of the KLK boxplots seem to be lower than the overall median gene expression and thereby are clearly down-regulated. Interestingly, KLK4.4 clearly stands out again as the highest expressed KLK gene. However, it needs to be considered that the expression of KLK4.4 is around the value 8, which does not really correspond with an over-expression in the context of the whole genome. Another interesting observation is that the boxplots of the KLK11 and KLK12 transcripts are big. With the whiskers of the boxplots, the gene expression spreads from a value of around 6 to about 9.5. This correlates with a high variety in gene expression for these gene transcripts. KLK11 and KLK12 gene expression therefore has a high information value that should be recognizable in the following methods. This could also be due to the fact that the lung cancer data set consists of of both normal and healthy tissue, as in comparison to the

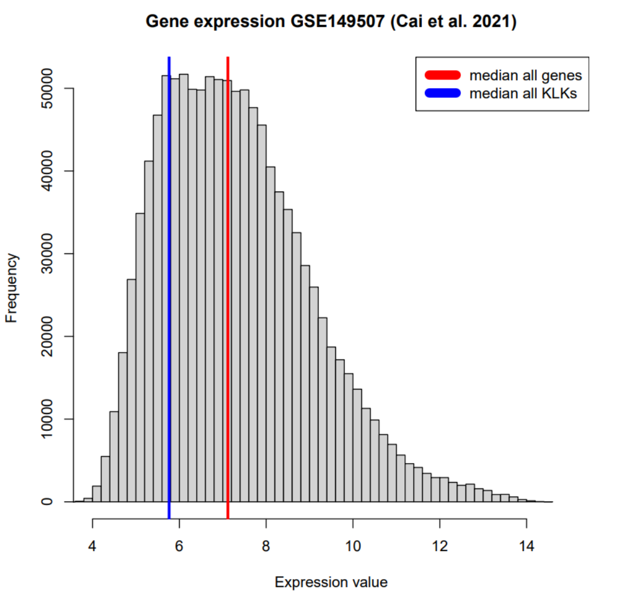


Figure 10: Histogram of lung cancer gene expression.

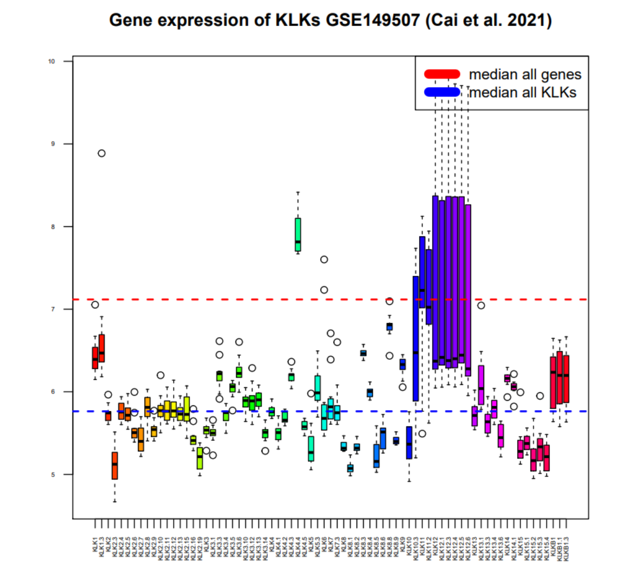


Figure 11: Boxplot of KLK gene expression in lung cancer.

breast cancer data set. Here, KLK11 and KLK12 are clear subject for further investigation in whether they are differently expressed between normal and carcinoma samples.

Heatmap

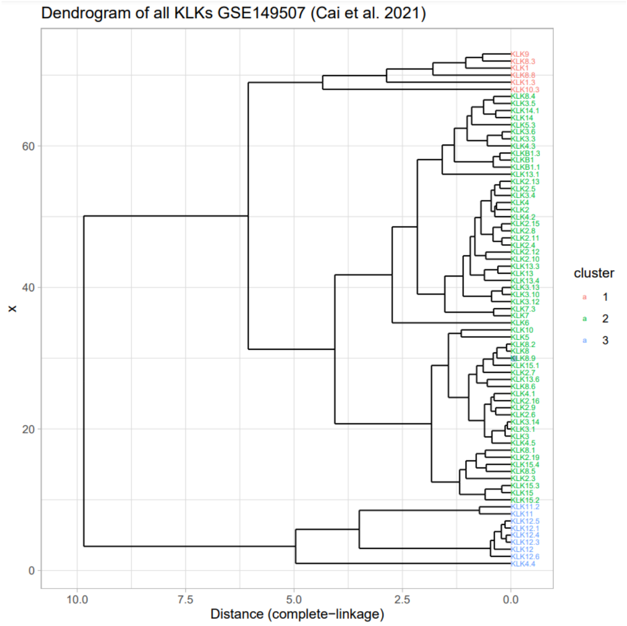


Figure 12: Dendrogram of KLK genes in lung cancer. Clustering is performed after the complete-linkage method. The genes are separated into 3 clusters.

The same strategy as for the breast cancer data set is applied to improve the quality of the heatmap. The dendrogram shows that multiple KLK11 and KLK12 isoforms, as well as KLK4.4 are part of one out of the three clusters.

Notably, the samples are clustered according to their tissue type being lung carcinoma or healthy tissue. As you can see in the dendrogram at the top, the normal samples are clustered into one group with additionally two more cancer samples. The other four cancer samples all form their own distinct group. This distribution by sample type clearly reflects itself in the KLK11 and KLK12 gene expression. While KLK4.4 is higher expressed for both normal and carcinoma samples, KLK11 and KLK12 isoforms are mainly higher expressed for the carcinoma sample. The only exception are the already mentioned carcinoma samples SCLC_01 and SCLC_03. Apart from this, the up-regulated gene expression for KLK11 and KLK12 are accompanied by the sample deriving from carcinoma tissue.

As stated by Borgoño et Diamandis, KLK11 up-regulation in lung cancer was found to have a unfavourable prognosis for the patient. A total of four out of the six cancer samples have slightly up-regulated KLK11 values. The significance will be tested. The two aforementioned carcinoma samples SCLC_01 and SCLC_03 even got down-regulated KLK11 expression, which thereby might indicate a good chance of treatment success. (Borgoño et Diamandis 2004)

PCA

\begin{figure}

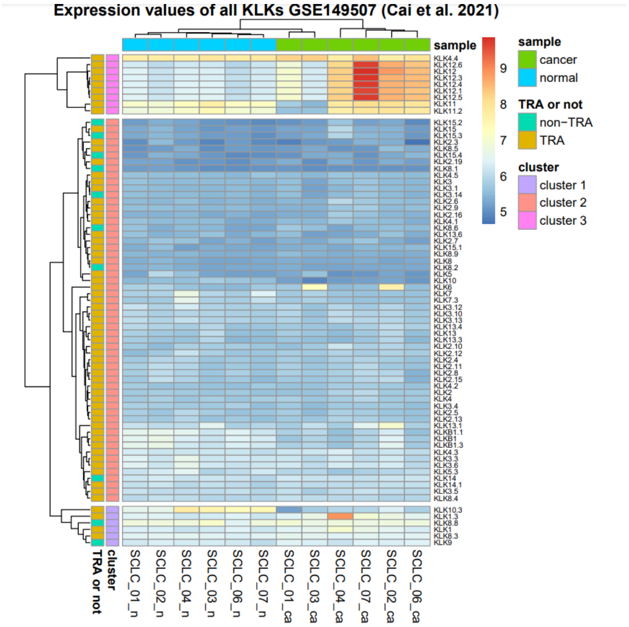
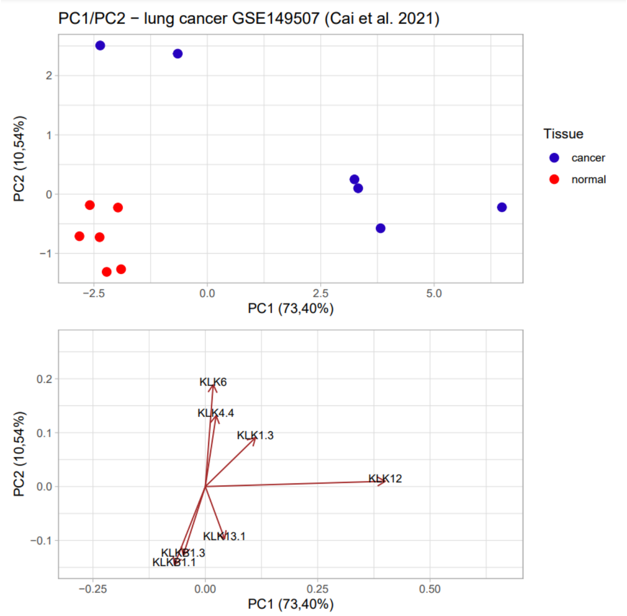


Figure 13: Heatmap of KLK gene expression in breast cancer. Carcinoma and normal samples are annotated. Additionally, the KLKs are differentiated by their cluster and potential tissue restriction.



{

}

\caption{PC1 (73,40%) is plotted against PC2 (10,54%). The upper part shows the distribution of the lung cancer samples annotated by their tissue type, while the lower part depicts the top 7 loadings of the KLKs.} \end{figure} The cumulative variance shows, that 84% of the total variance is explained by the first two PCs. These two PCs are more than sufficient for the analysis. However the first two PCs covering such a huge proportion of the whole variance corresponds with an overall low information value of the lung cancer data set. Considering the heatmap, in which most of expression values were down-regulated, there is only a low amount of differential gene expression going on. This explains the high cumulative variance. Nevertheless, the PCA still shows a clear separation between normal and carcinoma samples. Considering the loadings, four of the cancer samples are characterized by KLK12, while the other two tumor samples SCLC_01_ca and SCLC_03_ca are mainly represented by KLK4.4 and KLK6 expression. Just as shown in the heatmap, while four out of the six cancer samples have up-regulated KLK12 expression values, the two cancer samples SCLC_01_ca and SCLC_03_ca form an exception. These two do not go along with the KLK12 loading and are rather defined by KLK6 and KLK4.4 expression.

Clustering - kmeans

K-means performed for the lung cancer GSE149507 dataset showed an interesting and distinct cluster, which only consisted of KLK-subtypes of KLK12. Other genes that stood out in the heatmap analysis, for example KLK4.4 and KLK8.8, were found in the first cluster on the top left corner. The finding of optimal k clusters happened following the same procedure as for the breast cancer data set, described above. An optimal number of k = 5 clusters was determined using a Within Cluster Sum of Squares – plot.

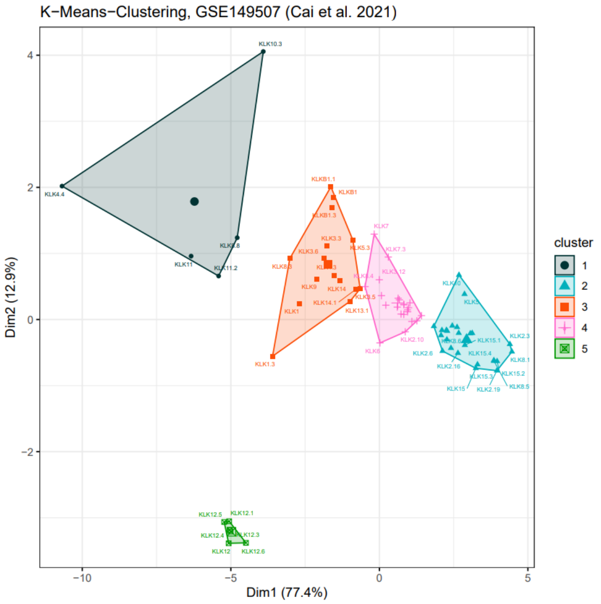


Figure 14: K-means cluster analysis with k = 5 clusters for the lung cancer data set

Hypothesis testing

The results of the PCA and the k-means indicate that for some KLKs the expression differs between the cancerous and normal tissue. Plotting the data did not indicate a normal distribution. Therefore, the Wilcoxon signed-rank test was applied.

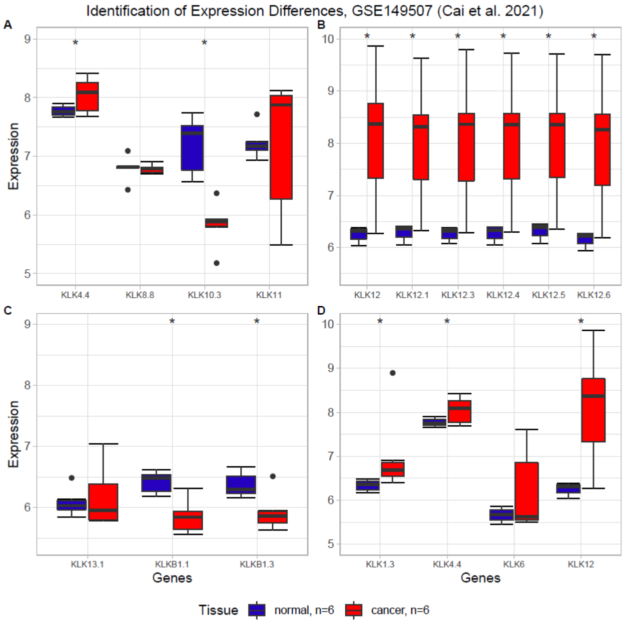
In figure X plot A, KLK4.4 was significantly higher expressed in cancer tissue. Unlike, KLK10.3 which was significantly higher expressed in normal tissue. The findings of decreased KLK11 expression in lung cancer from Sasaki et al. could not be confirmed. In fact, the median of the cancer tissue samples is clearly higher.

Plot B shows that KLK12 and its isoforms are significantly higher expressed in cancer tissue. The plot also visualizes the high similarity within isoforms, since in the cleanup only identical were removed. Documented KLK12 overexpression could not be found in the literature, but in functional studies KLK12 was identified as a pro-angiogentic factor.

First, the expression values from genes originating from cluster 1 in k-means analysis results were compared. For those genes, KLK10.3, KLK11, KLK4.4 and KLK8.8, expression values of cancerous tissue and normal tissue were compared through Wilcoxon signed-rank testing. (wäre das eher bildunterschrift?)

Here, KLK10.3 is cleary lower expressed in cancerous tissue compared to normal tissue, whileas KLK4.4 seems to be higher expressed in cancer tissue. For KLK11, expression values have a high variety for cancerous tissue, and only a low expression range in normal tissue. (-> for discussion: a deviation in expression values of KLK11 could maybe be an indication for lung cancer, since the variances are high, it could be that the KLK11 gene is mutated in cancerous tissue) For KLK8.8, no significant difference was detectable.

Wilcoxon signed-rank test for cluster 4 of k-means analysis, containing KLK12 and subtypes, showed a clear overexpression of all KLK12 derived KLKs in cancer tissue, compared to normal tissue.



Logistsic regression

KLKs are already used in cancer diagnostics to determine the tumor type of sample. Since expression patterns of many different genes are very complex and hard to interpret regression models are a helpful tool to determine the cancer type. For the lung cancer data used in this analysis a binary outcome for the mikrochips (cancer or normal tissue) is possible. Therefore logistic regression was chosen. The basic assumtions for logistsic regression are: 1. Independency of errors, every observation has to be separate from the others. 2. Linearity of the continuous variables in logit - the relationship between the variable and their logit transformed outcome should be linear. 3. Absence of multicollinearity or redundancy. 4. No outliners with a strong influence. 5. For every independent variable there should be at least ten outcomes (Jill C.Stoltzfus 2011).

These assumptions reveal the shortcomings of the used data and explain the experienced problems with logistic regression. First, main limitation of the used dataset is the low number of included mikrochips (only 12, 6 from cancer and 6 from normal tissue). Therefore expected problems of high standard error and large beta-coefficients for the independent variables were in encountered when including more then one independent variable. This phenomenon of unstable models is also called overfit-model. In fact, even for most individual genes, which had been identified ro have an significant expression difference, the described problems were encountered. It is worth noting that the data was split into a training dataset of 8 mikrochips and a test dataset of 4 mikrochips. But even with all the available mikrochips, the standard error of all of the individual tested genes, except KLK4.4 and KLK12, could not be decreased significantly. These observations are consistent with Feinstein, who recommends 20 outcomes per variable. KLKs are already used in cancer diagnostics to determine the tumor type of sample. Since expression patterns of many different genes are very complex and hard to interpret regression models are a helpful tool to determine the cancer type. For the lung cancer data used in this analysis a binary outcome for the mikrochips (cancer or normal tissue) is possible. Therefore logistic regression was chosen. The basic assumtions for logistsic regression are: 1. Independency of errors, every observation has to be separate from the others. 2. Linearity of the continuous variables in logit - the relationship between the variable and their logit transformed outcome should be linear. 3. Absence of multicollinearity or redundancy. 4. No outliners with a strong influence. 5. For every independent variable there should be at least ten outcomes (Jill C.Stoltzfus 2011). These assumptions reveal the shortcomings of the used data and explain the experienced problems with logistic regression.

First, main limitation of the used dataset is the low number of included mikrochips (only 12, 6 from cancer and 6 from normal tissue). Therefore expected problems of high standard error and large beta-coefficients for the independent variables were in encountered when including more then one independent variable. This phenomenon of unstable models is also called overfit-model. In fact, even for most individual genes, which had been identified ro have an significant expression difference, the described problems were encountered. It is worth noting that the data was split into a training dataset of 7 mikrochips and a test dataset of 3 mikrochips. But even with all the available mikrochips, the standard error of all of the individual tested genes, except KLK1.3 and KLK12, could not be decreased significantly.

Second, although genes with a correlation equal to one were removed, some genes are still highly correlating. This is primarily true for the different isoforms of the same gene, as visualized in figure x, plot B (hypothesis testing KLK12 and isoforms). Therefore, the effect of colliniarity would probably cause problems, even if more mikrochips were included. Hence a second cleanup, removing genes with high correlation (e.g. corr > 0.8), would be necessary.

As mentioned above KLK4.4 and KLK12 (and its isoforms) were the only gene where the standard error of the independent variable was not abnormaly high. But the p-value was, in both cases, not significant. In contrast, the prediction of these univariant models were surprisingly accurate. The model with KLK4.4 could predict the tissue type of 3 out of 4 mikrochips correctly, the model with KLK12 even predicted every tissue type right. On closer examination of the probabilities however, it becomes apparent that these models are far from reliable. The probabilities for normal to be cancer tissue were mostly over a quarter, indicating a high uncertainty.

In conclusion, the p-values of the independent variable in the models were not significant, the standard errors were still large and the predictions probabilities were not accurate. These results were not surprising considering the low sample size.

Discussion

References

Ardlie, K.G., Deluca, D.S., Segre, A.V., Sullivan, T.J., Young, T.R., Gelfand, E.T., Trowbridge, C.A., Maller, J.B., Tukiainen, T., Lek, M., et al. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science 348, 648-660.

Borgoño, C.A., and Diamandis, E.P. (2004). The emerging roles of human tissue kallikreins in cancer. Nature Reviews Cancer 4, 876-890.

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

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).

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

Dubey, A.K., Gupta, U., and Jain, S. (2016). Analysis of k-means clustering approach on the breast cancer Wisconsin dataset. International Journal of Computer Assisted Radiology and Surgery 11, 2033-2047.

Fischer, J., and Meyer-Hoffert, U. (2013). Regulation of kallikrein-related peptidases in the skin – from physiology to diseases to therapeutic options. Thromb Haemost 110, 442-449.

Haritos, C., Michaelidou, K., Mavridis, K., Missitzis, I., Ardavanis, A., Griniatsos, J., and Scorilas, A. (2018). Kallikrein-related peptidase 6 (KLK6) expression differentiates tumor subtypes and predicts clinical outcome in breast cancer patients. Clinical and Experimental Medicine 18, 203-213.

Kont, V., Laan, M., Kisand, K., Merits, A., Scott, H.S., and Peterson, P. (2008). Modulation of Aire regulates the expression of tissue-restricted antigens. Molecular Immunology 45, 25-33. 10.1016/j.molimm.2007.05.014.

Lattin JE, S.K., Su AI, Walker JR, Zhang J, Wiltshire T, Saijo K, Glass CK, Hume DA, Kellie S, Sweet MJ (2008). Expression analysis of G Protein-Coupled Receptors in mouse macrophages. Immunome Res. 4:5.

Lenga Ma Bonda, W., Iochmann, S., Magnen, M., Courty, Y., and Reverdiau, P. (2018). Kallikrein-related peptidases in lung diseases. Biol Chem 399, 959-971.

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

Maire, V., Némati, F., Richardson, M., Vincent-Salomon, A., Tesson, B., Rigaill, G., Gravier, E., Marty-Prouvost, B., De Koning, L., 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.

Michaelidou, K., Ardavanis, A., and Scorilas, A. (2015). Clinical relevance of the deregulated kallikrein-related peptidase 8 mRNA expression in breast cancer: a novel independent indicator of disease-free survival. Breast Cancer Research and Treatment 152, 323-336.

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.

Sasaki, H., Kawano, O., Endo, K., Suzuki, E., Haneda, H., Yukiue, H., Kobayashi, Y., Yano, M., and Fujii, Y. (2006). Decreased Kallikrein 11 Messenger RNA Expression in Lung Cancer. Clinical Lung Cancer 8, 45-48.

Schmitt, M., Magdolen, V., Yang, F., Kiechle, M., Bayani, J., Yousef, G.M., Scorilas, A., Diamandis, E.P., and Dorn, J. (2013). Emerging clinical importance of the cancer biomarkers kallikrein-related peptidases (KLK) in female and male reproductive organ malignancies. Radiology and Oncology 47, 319-329.

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., Moqrich, A., et al. (2002). Large-scale analysis of the human and mouse transcriptomes. Proceedings of the National Academy of Sciences 99, 4465-4470.

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

Tailor, P.D., Kodeboyina, S.K., Bai, S., Patel, N., Sharma, S., Ratnani, A., Copland, J.A., She, J.-X., and Sharma, A. (2018). Diagnostic and prognostic biomarker potential of kallikrein family genes in different cancer types. Oncotarget 9, 17876-17888. 10.18632/oncotarget.24947.

Uhlen, M., Fagerberg, L., Hallstrom, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, A., Kampf, C., Sjostedt, E., Asplund, A., et al. (2015). Tissue-based map of the human proteome. Science 347, 1260419-1260419.

Yousef, G.M., Chang, A., Scorilas, A., and Diamandis, E.P. (2000). Genomic Organization of the Human Kallikrein Gene Family on Chromosome 19q13.3–q13.4. Biochemical and Biophysical Research Communications 276, 125-133. 10.1006/bbrc.2000.3448.

Yousef, G.M., Magklara, A., and Diamandis, E.P. (2000). KLK12 Is a Novel Serine Protease and a New Member of the Human Kallikrein Gene Family—Differential Expression in Breast Cancer. Genomics 69, 331-341.

Yousef, G.M., Yacoub, G.M., Polymeris, M.E., Popalis, C., Soosaipillai, A., and Diamandis, E.P. (2004). Kallikrein gene downregulation in breast cancer. British Journal of Cancer 90, 167-172.

Zhang, Y., Bhat, I., Zeng, M., Jayal, G., Wazer, D.E., Band, H., and Band, V. (2006). Human kallikrein 10, a predictive marker for breast cancer. 387, 715-721.