# PCA

Principal Component Analysis (PCA) is a statistical method, which can be used for reducing the number of dimensions within a dataset. The goal of dimensionality reduction is to describe the properties of the data in a lower number of dimensions than present in the original dataset, while retaining as much information as possible.

PCA achieves this through feature extraction. This means that new features are generated by combining the original features in a specific way. This approach has the advantage that it does not require the upfront knowledge about which features are more related to the outcome of interest, and which could be left out. On the other hand, the extracted features are much less interpretable, as they are composed of a high number of different variables.

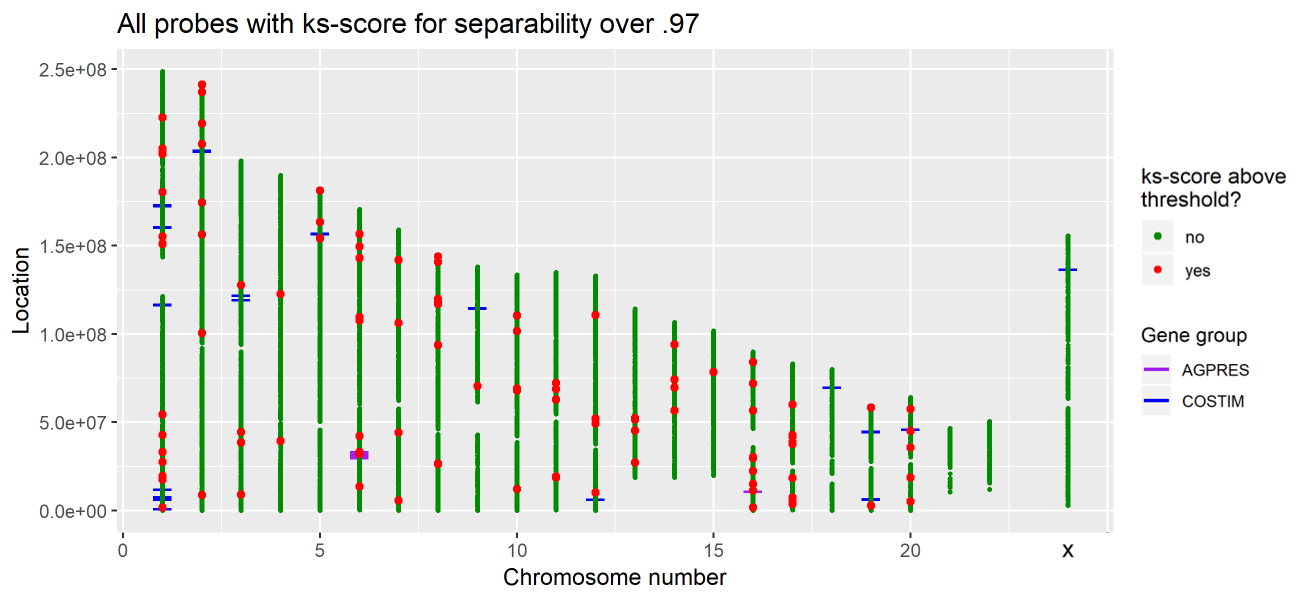
The extracted features, or principle components, have the following properties:

1. For p-dimensional data, a principle component is a linear combination of the original variables, hence, where.
2. For principle component, the loadings vector is obtained by finding the linear projection that maximizes the total amount of variance within the dataset.
3. Each new generated principle component is orthogonal to all of the previous principle components. Hence, for the kth principle component, we have for each.

By definition of these properties, from a p-dimensional dataset that consist of n observations, at most principle components can be extracted.

The optimal number of principle components to select for the analysis is subjective to the application. Sometimes a specific number of dimensions is desired, which means the above-described process can be stopped after a fixed number of iterations. A common alternative is looking at the total explained variance. As, by definition of point 2 an 3 above, all principle components are both independent of each other and decreasingly ordered in amount of variance they explain, the desired number of dimensions can be selected once the accumulated variance explained exceeds a certain threshold.

# Picture description (separability on chromosomes)



* The location of each probe is visualized above. The x-axis shows the chromosome on which the probe lies, where the y-axis represents the probe’s location on that specific chromosome.
* Initially, each probe is marked with a green dot. If the ks-score for the separability of the two tumor types exceeds a certain pre-defined threshold, the color of the marker for that probe is changed to red.
* For both the COSTIM and the AGPRES gene-groups, each probe attached to a gene belonging to either groups gets flagged with a blue or purple line, respectively.

# Kolmogorov-Smirnov test (KS score)

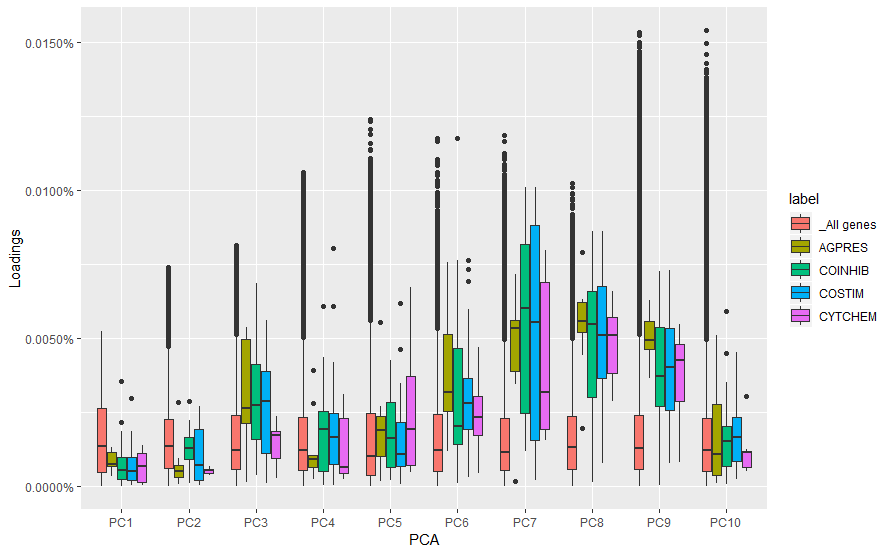
The Kolmogorov-Smirnov test, in short the KS test, is a nonparametric test that compares two data samples. The goal of the KS test is to determine if two data samples come from the same distribution, noting that it is not specified what that common distribution is.

The KS score quantifies a distance between the empirical distribution functions of two samples. The KS score is mathematically defined by:

,

where and are the empirical distribution functions of the first and the second sample respectively, and the supremum function. If both samples comes from the same distribution, then converges to almost surely in the limit. To conclude, the KS score lays in the interval , where a score closer to zero indicates that both samples are more likely to be drawn from the same distribution.

# Picture description (PCA genegroups)



We distinguish 4 main gene groups: antigen presentation (agpres), T cell co-inhibitory (coinhib), T cell co-stimulator (costim) and cytokines/chemokines (cytchem). After applying PCA analysis, the influence of these gene groups were investigated on the individual principle components.

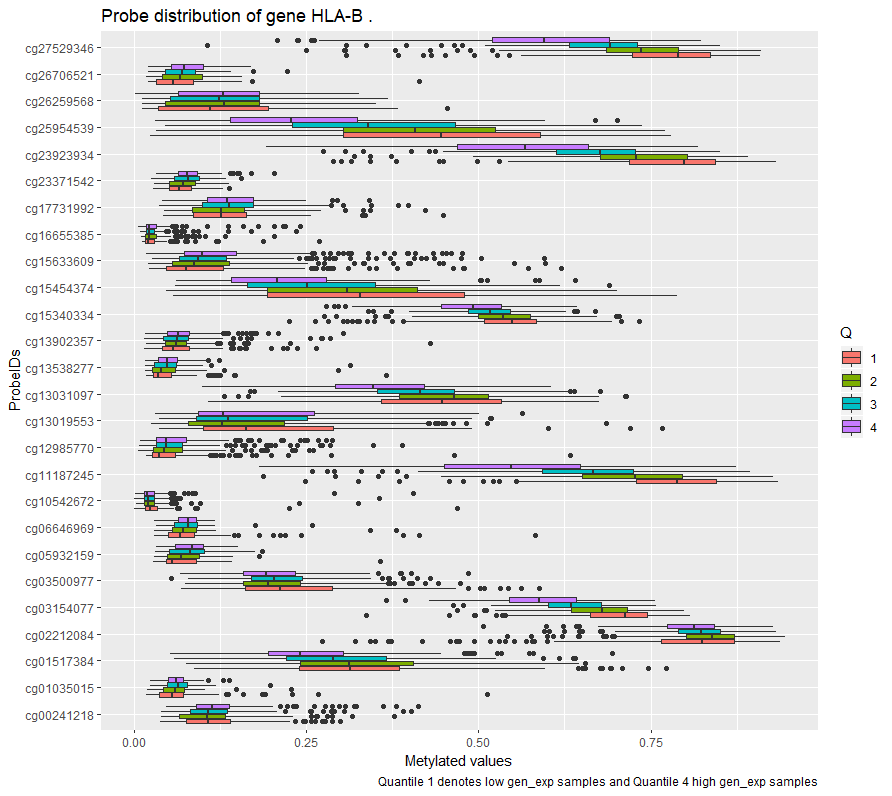
* The y-axis represents the loading of a gene in the RNA expression dataset to a principle component.
* The red boxes indicate all genes in the RNA expression dataset, whereas (for example) the yellow boxes only represents the genes in the agpres gene group.
* We see that all gene groups are most influential in PC7, PC8 and PC9.
* *The above boxplot can be seen as following: the midline is the median of your data, with the lower and upper limits of the box being the first and third quartile (25th and 75th percentile) respectively. By default, the whiskers will extend up to 1.5 times the interquartile range from the top (bottom) of the box to the furthest datum within that distance. If there are any data beyond that distance, they are represented individually as points ('outliers').*

# Data prep

Something about prepping the data we used for Methylation (which is done at Hackathon?).

**To do.**

# Picture description (Bridge between RNA and Methylation)



For all genes in the RNA expression data set corresponding methylation data can be found. Interesting is to see how a increase/decrease of gene expression influences the methylation values of the genes. Following is a bullet-point description on how the above image is constructed:

* Consider a gene in the RNA expression dataset *(example: HLA-B)*
* Patient samples are divided into 4 quantiles according to their RNA expression. Samples in quantile 1 have the lowest RNA expression and samples in quantile 4 have the highest RNA expression.
* For the gene in the RNA expression dataset corresponding ProbeIDs are found in the methylation dataset. (*in the example of HLA-B a list of 26 probeIDs)*
* For each quantile and probeID the corresponding methylation distribution out of the methylation dataset is visualized by a boxplot.

The most important visualization here, is that a shift of the quantiles over the methylation values indicate high correlation between the RNA expression and the methylation of a probe *(example: cg27529346).* This can also be expressed in a general correlation value, only the main advantage in this type of image is that the height of the methylated values is visualized as well.