**Cancer hallmarks and tumour metabolism**

What cellular characteristics make a cell a cancer cell?

Do all tumour types share the same features? (e.g., carcinoma vs. liquid tumours)

Can we identify patterns of cellular modifications in cancer samples?

* Cancer cells must adapt their metabolism to new physiological conditions.

**Main questions:**

* Which genes / genesets have a modified expression in cancer cells compared with their normal counterparts?
* Are these patterns universal over all cancer types, or can we recognize patterns?
* Can we group cancer types according to their altered pathway activities?
* How can we quantify the activity of a pathway?

**DATA:**

Tcga\_tumor\_log2TPM dataset:

* 60,489 x 9,741
* RNA-seq (bulk) -> Count matrix -> Normalisation -> log2(TPM)
* Rows: genes, columns: samples, values: expression values

Genes:

* Character with 60,489 elements
* Contains the names of all 60,489 genes which constitute the row names of the tcga\_tumor\_log2 matrix

Tcga\_tumor\_annotations:

* Dataframe: 9,741 entries, 37 total columns
* Clinical information about the 9,741 samples
* Many empty cells (no NAs)

Tcga\_tumor\_normal:

* List with 5 elements:
  + BRCA, KIRC, LUAD, THCA
  + PRAD (list with 3 elements)
    - Tumor (data.frame with 19,624 rows/genes and 52 colums/samples)
    - Normal (data.frame with 19,624 rows/genes and 52 colums/samples)
    - Clinical (data.frame with 52 rows/samples and 37 colums/clinical information)
* 2 samples from each gene: 1st from tumor, 2nd from normal tissue of same patient
* Comparison between tumor and normal cell in same patient possible
* We are going to take a closer look at prostate adenocarcinoma

Signalling / metabolic pathways can be characterised using a list of relevant genes:

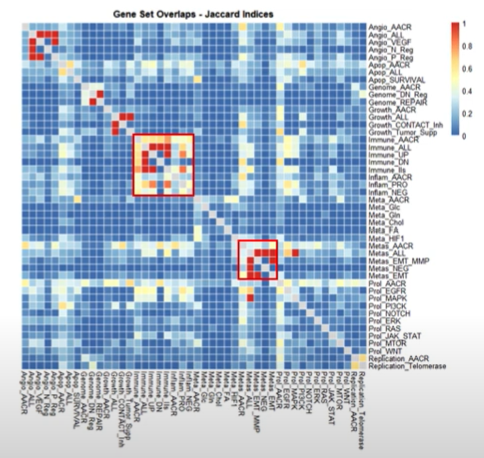
Hallmarks\_genesets:

* Some genesets are given by supervisor
* List with 2 elements:
  + List with 46 genesets
  + Description/names of the 46 genesets

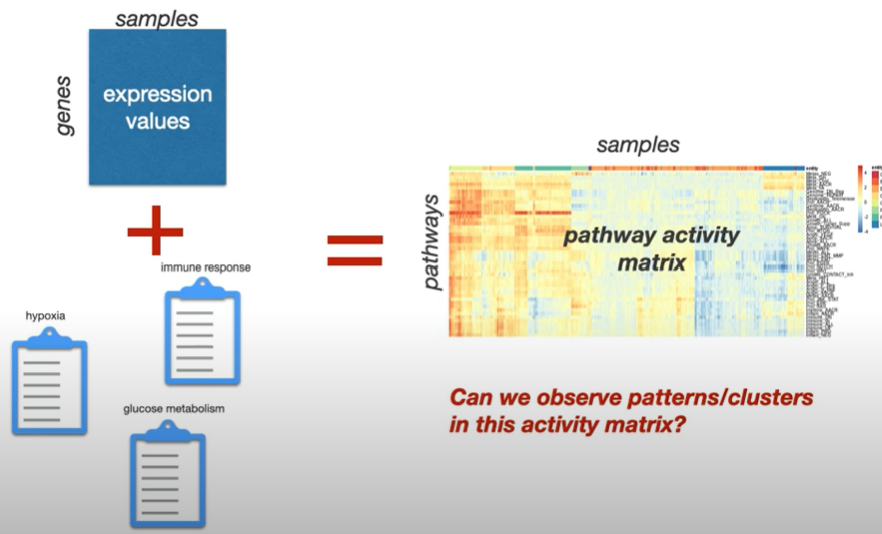
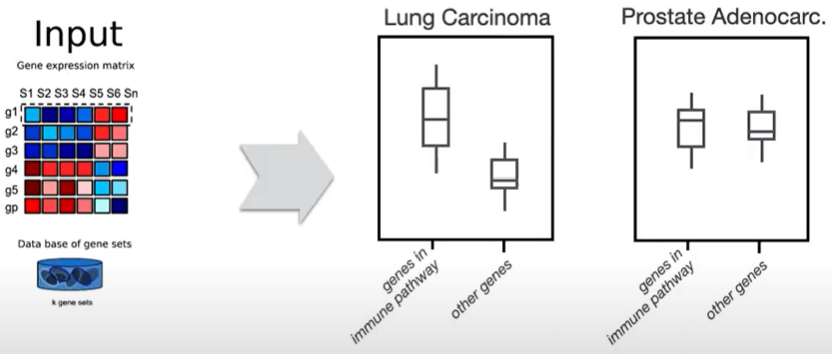
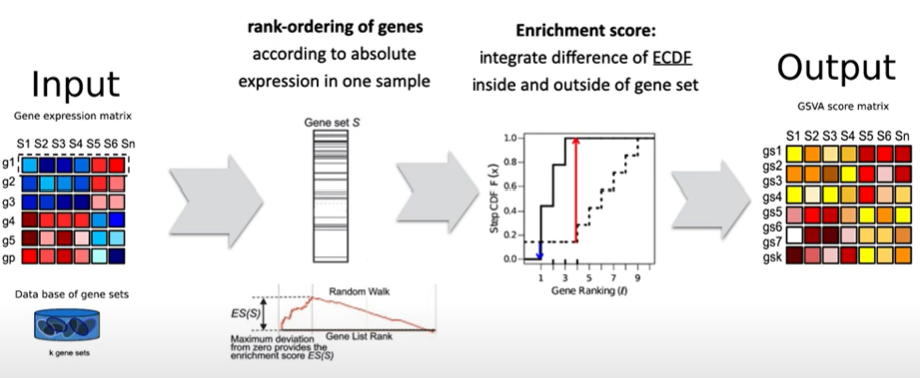
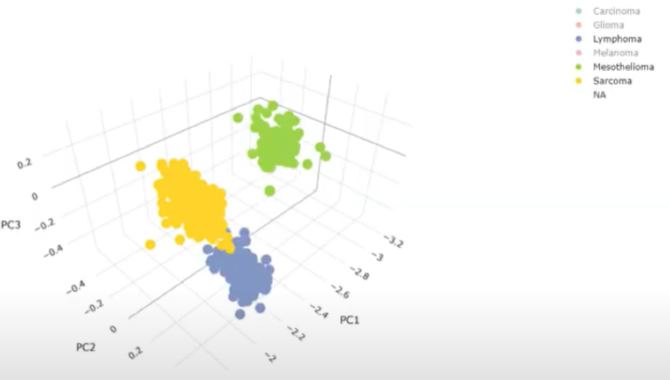
**GOAL1: IDENTIFY CANCER HALLMARK PATHWAYS**

* Identify **additional** genesets from the literature which might be interesting to look at/ are relevant for cancer
  + E.g.: metabolic pathways
* List of dozens or hundreds of genes
* Collect those gene sets 🡪 MSigDB (R package + biomart)

**GOAL2: COMPARE GENE SETS**

* Comparison of Hallmarks\_genesets with additional genesets
  + How many shared genes are there between the gene sets?
* Compare all datasets to **find groups of datasets** with shared genes
  + Find a suitable metric to quantify the distance/similarity between 2 sets
  + Graphical representation
    - Heatmap: genesets x genesets
      * Similarity matrix
      * Find overlaps between genesets
      * **Jaccard Indices**
      * Which of the genesets have a lot of shared genes?
* 

**GOAL3: PAN-CANCER ANALYSIS OF CANCER HALLMARKS AND METABOLIC PATHWAYS**

* Find a way to score pathway activity
* Combine Tcga\_tumor\_log2TPM with (Hallmark\_genesets+additional genesets) to a pathway activity matrix
  + Rows: pathways
  + Columns: samples
  + Are there patterns/clusters visible? (-> Heatmap)
    - Which pathways?
    - Which tumors?
  + 
* How to define numerical values for pathway activity matrix?
  + Input: Gene expression matrix and data base of gene sets
  + **Start with one pathway and later compute for every pathway**
  + Compare Boxplots for each tumor type, e.g.:
    - Genes in path x vs. other genes
    - t-Test, t-Values (basic)
    - 
  + rank-ordering of genes
    - according to absolute expression in one sample
    - take 1 column
    - order all genes from most expressed to least expressed
    - where are our genes of interest (e.g., immune pathway dataset) in this list?
      * At the top, at the bottom, randomly distributed?
      * Compute/quantify using cumulative distribution, area under curve
      * **Yield 1 number (Enrichment score)**
        + Integrate difference of ECDF (?) inside and outside geneset
        + 🡪 matrix with samples x gensets
  + 
  + Graphical display
    - Pathway activity matrix
      * PCA
      * Clustering
      * Identify patterns and clusters of samples/pathways
        + Can we identify clusters corresponding to different types of tumors?
        + Compare with clinical information (Tcga\_tumor\_annotations)
      * 

**GOAL4: SPECIFIC ANALYSIS TO COMPARE NORMAL TO CANCER CELLS IN PROSTATE ADENOCARCINOMA**

* Tcga\_tumor\_normal
* PRAD (list with 3 elements)
  + Tumor (data.frame with 19,624 rows/genes and 52 colums/samples)
  + Normal (data.frame with 19,624 rows/genes and 52 colums/samples)
  + Clinical (data.frame with 52 rows/samples and 37 colums/clinical information)
* Similar analysis as in GOAL3 for
  + Tumor samples
  + Normal samples
* Test for significant different activity between normal and tumor samples
  + Paired t-test?
* Wang: ***Was üblicherweise gemacht wird:***
  + Foldchange berechnen (condition1/condition2)
  + Mittelwert eines genes (einer condition)
  + Foldchange berechnen
  + Mittelwerte zwischen den beiden conditions vergleichen -> signifikant? t-test (paired)
  + Für jedes Gen habt ihr einen Foldchange (FC) und den dazugehörigen p-value (!!! multiple testing problem !!! -> p-value muss korrigiert werden z.B. Benjamini-Hochberg/Bonferroni)
  + p-values per gene (based on Foldchange) -> Ranking for GSEA!! ???
  + Make a Volcano Plot (Standard plot in RNA-seq Analyse)

* + Später könnt ihr denselben Workflow für Pathways machen -> Visualisierung im Volcano Plot zeigt euch die upregulated bzw. die downregulated pathways

**GOAL5: PREDICT ACTIVITIES USING A REGRESSION MODEL**

* Choose one metabolic pathway of interest (one of the additional pathways?)
* Build a linear regression model to predict the chosen pathway from the other pathways
* Use the results from GOAL2 (Innuendo on Similarity/Correlation? -> PCA?)