Project 04:
Aberrant expression in metastatic lung cancer

Biological Data Analysis course
SS 2020

Aurelien Dugourd
Nicolàs Palacio-Escat

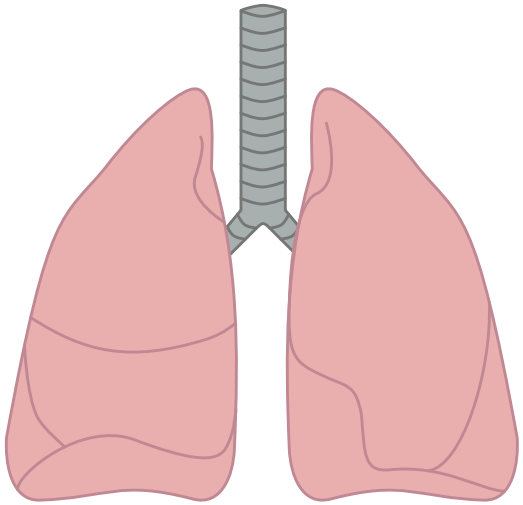
www.saezlab.org
@sysbiomed

Institute for Computational Biomedicine
Heidelberg University & RWTH Aachen

Overview

- Introduction
- Objectives
  - Exploratory analysis
  - Specific analysis
- Description of data set
- How to structure your project
  - Project proposal
  - Project must-have

Introduction

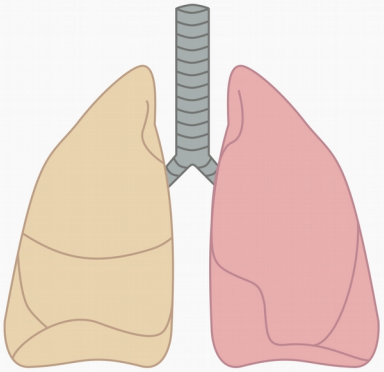


Lung cancer

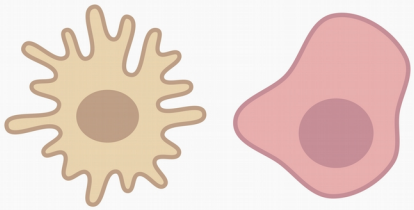
Lung cancer is when abnormal cells divide in an uncontrolled way to form a tumour in the lung.
The main symptoms are a cough, breathlessness and weight loss.
Treatments include surgery, chemotherapy and radiotherapy.

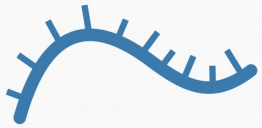
Objectives

Experimental data



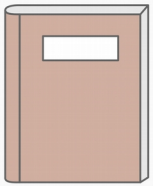
Cancer Healthy



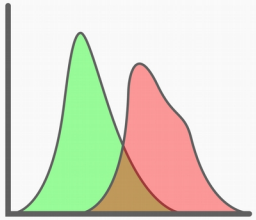


RNA expression

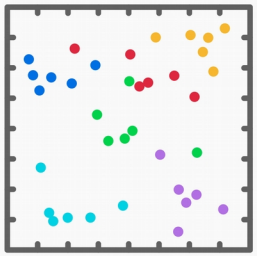
Exploratory analysis



Literature review

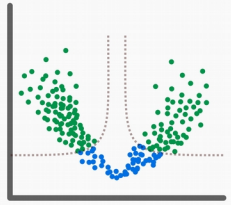


Descriptive statistics

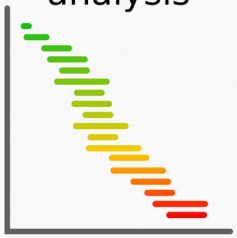


Dimensionality reduction

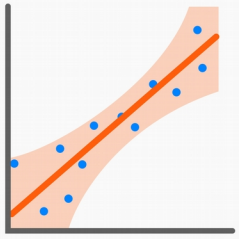
Specific analysis



Differential expression analysis



Signature extraction



Linear regression

Objectives – Exploratory analysis

- **Search literature about disease and type of data**
  - What can be analyzed, what to expect? …
- **Familiarize witht the data**
  - What units? How many genes/samples? …
- **Visualize the data**
  - How samples distribute? Does it need normalization? Are there technical bias (batch effect)? …
- **Apply dimensionality-reduction**
  - Are there patterns in the space of gene expression?

Objectives – Specific analysis

- **Choose one subtype of lung cancer to focus on**
- **Differential expression analysis**
  - Compare disease vs. healthy – how much expression changes? Is it significant?
- **Signature extraction**
  - Contextualize the changes from a molecular/functional perspective.
- **Modelling**
  - Can we explain/predict disease characteristics from the data analysis?

Description of the data set

Gene expression of 293 lung tumor and 14 non-tumoral lung samples.

Subclass	Sample
Non-tumoral lung	14
Adenocarcinoma	85
Squamous Cell Tumours	61
Large cell neuroendocrine	56
Basaloid tumours	39
Carcinoid tumours	24
Small cell carcinoma	21
Other histology	7
Total	307

Sample metadata:

- Gender
- Age
- Status (dead/alive)
- TNM stage
- Relapse
- Follow-up time
- Smoking
- Therapy received

How to structure your project - proposal

(10 minute presentation + 5 minutes discussion)

Define a project proposal, which should include:

- Summary of literature review
- Questions you want to address and how
- Approximate timetable.

How to structure your project – must-have

- Descriptive statistics about the data sets
- Graphical representations
- Dimension reduction analysis (PCA, clustering, k-means...)
- Statistical tests (t-test, proportion tests, etc)
- Linear regression analysis, either uni- or multivariate