Distinguishing between cancer subtypes based on transcriptomics and epigenomics using the example of kidney cancer and larynx cancer

Mia Anscheit, Matanat Mammadli, Friederike Wohlfarth

Outline

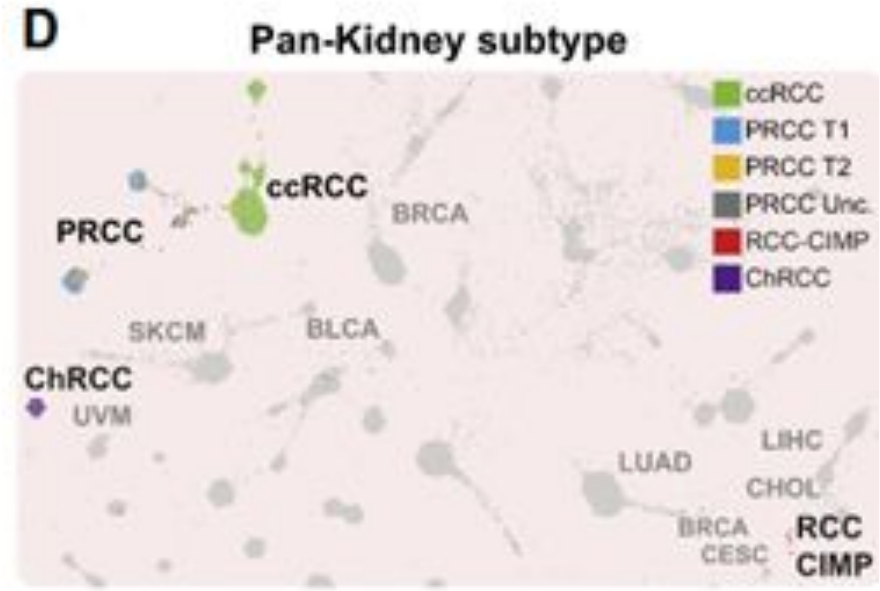
1. Background / Motivation
2. Data
3. Project plan
4. Methods / Software

1. Background / Motivation

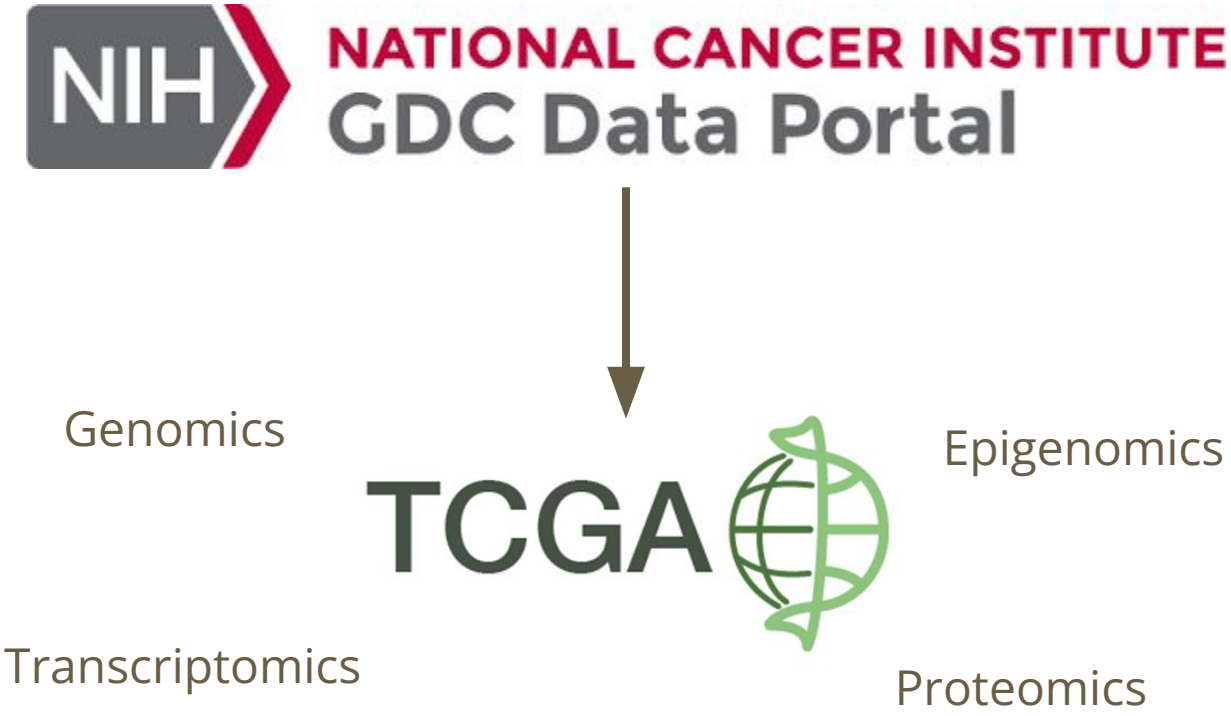
- cancer types and affected tissues show different transcriptome and methylation profiles
- clustering is primarily organized by histology, tissue type, or anatomic origin
- samples from related organ systems group together
- within one subtype different tissues are differently affected by smoking / abundant between sexes (laryngeal cancer)

1. Background / Motivation

- pan-kidney subtype:



2. Data: Database



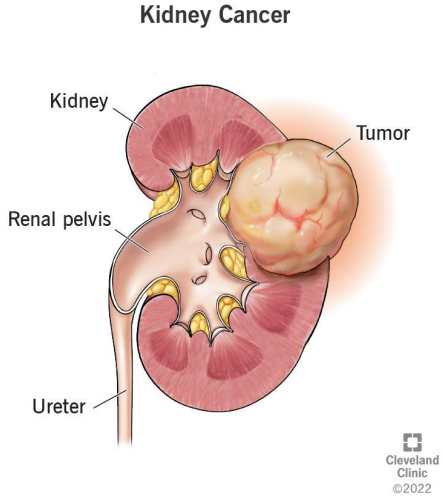
2. Data: Samples

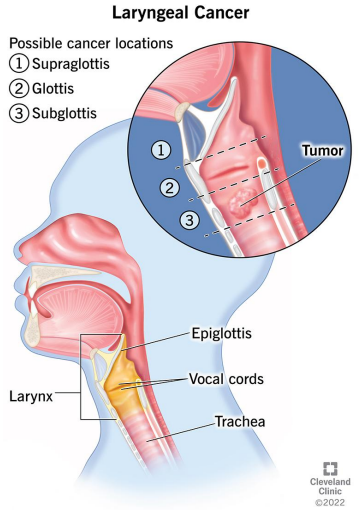
Kidney cancer

- TCGA-KIRC (521 samples)
- TCGA-KIRP (291 samples)
- TCGA-KICH (66 samples)

Larynx cancer

- TCGA-HNSC (116 samples)





2. Data: Omics

Transcriptomics
differential gene expression analysis
→ based on RNA-Seq count data
(tsv)

Epigenomics
differential methylated regions analysis
→ based on DNA methylation data
(idat)

3. Project plan

- Differential expression and methylation analysis (on omics data)

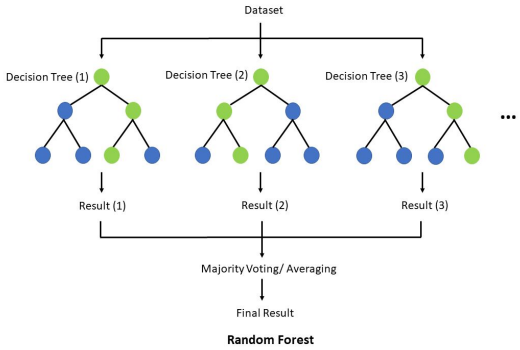
- Feature selection (most differential or significant features)

- Machine Learning (compare different types of models)

- Feature importance analysis (features with the most impact on the predictive power of a model)

- Gene annotation (biological roles, associated pathways)

4. Methods / Software



- R (Bioconductor) or Python for computation
- TCGAbiolinks package to load files into R
- DESeq2 for differential gene expression analysis
- methylationArrayAnalysis for differential methylation analysis
- scikit-learn (VarianceThreshold, SelectKBest) for feature selection
- the caret package and scikit-learn for machine learning (Random forest, clustering methods etc.)
- Random Forest Classifier (or caret) and Cross Validation for hyperparameter tuning
- Random forest or XGBoost for feature importance analysis
- GO enrichment for gene annotation

GENEONTOLOGY
Unifying Biology

DESeq2

References

Joshua D. Campbell et al. “Genomic, Pathway Network, and Immunologic Features Distinguishing Squamous Carcinomas”. In: Cell Reports 23.1 (Apr. 3, 2018), 194–212.

Katherine A. Hoadley et al. “Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer”. In: Cell 173.2 (Apr. 5, 2018), 291–304.

Alana Sorgini et al. “Analysis of the TCGA Dataset Reveals that Subsites of Laryngeal Squamous Cell Carcinoma are Molecularly Distinct”. In: Cancers 13.1 (Dec. 31, 2020), p. 105.