



UNIVERSITÀ
DI TRENTO

Gene expression analysis of lung adenocarcinoma and matched adjacent non-tumor lung tissue

Daniele Capelli

Data Analysis and Exploration

July 18th, 2025





- 1 Problem Setting
- 2 Expression Profiling by Array Data
- 3 Functional Analysis
- 4 Multi-Omics Analysis
- 5 Discussion and Future Works



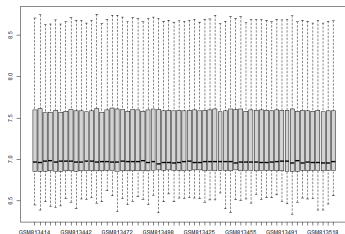
Problem Setting



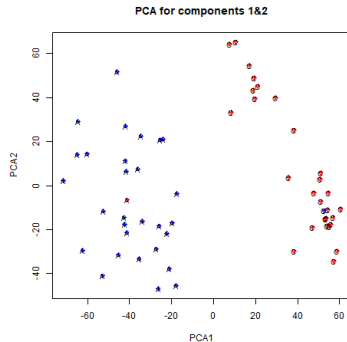
- **Aim:** Find differences between adenocarcinoma lung cancer and healthy tissues
- **Sample Data:**
 - 30 tumor samples and 29 non-tumor lung (NTL) samples
 - Highly heterogeneous population sample
- **Two omics:**
 - Expression profiling by array data
 - Methylation sites data
- Results compared to Selamat et al. (2012, [1])



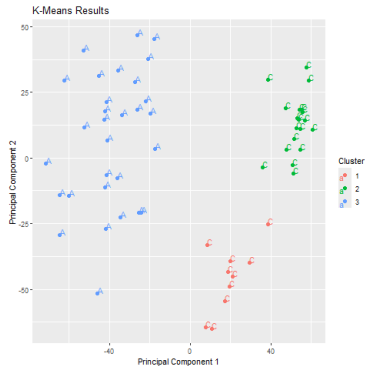
Expression Profiling by Array Data



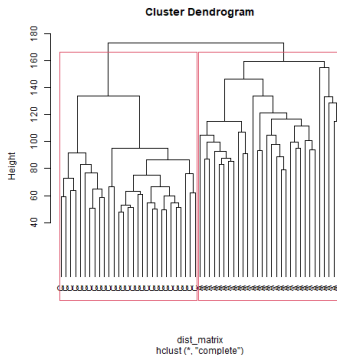
(a) Boxplot of the data



(b) First two PCA components
(tumor in blue, control in red)



(a) 3-Means Clustering

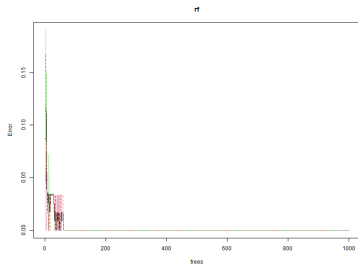


(b) Hierarchical Clustering

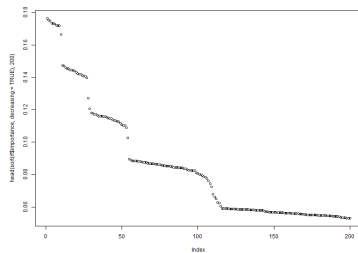


Fitted Models: Random Forest, Linear Discriminant Analysis, LASSO and SCUDO

- **Random Forest:** 1000 trees resulted in a perfect accuracy



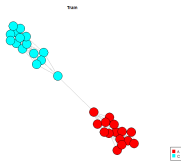
(a) Random Forest error



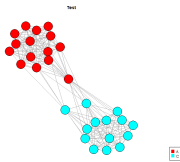
(b) Importance Plot (zoom)



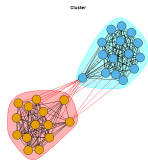
- **LDA**: resulted in a perfect accuracy
- **LASSO**: choice of 0.2 as regularization parameter resulted in a perfect accuracy score
- **SCUDO**: 250 keys selected both for the upper and the lower signature resulted in a perfect accuracy



(a) Scudo Network
(training data)



(b) Scudo Network
(validation data)



(c) Clustering of the
validation data



Functional Analysis



Figure: Functional Enrichment Analysis on the 400 most important genes from the random forest approach



- **Pathways identified by STRING:** "PDGF/VEGF domain, and Vascular Endothelial Growth Factor", "Constitutive Signaling by Aberrant PI3K in Cancer", "VEGF ligand–receptor interactions, and Tie signaling", "Leukocyte transendothelial migration", "Signaling by VEGF", "VEGFA–VEGFR2 signaling pathway", "VEGF signaling pathway" and "Hippo signaling regulation pathways" (linked to **angiogenesis** and **blood vessel formation**)
- **Pathways identified by EnrichNet:** "vasoconstriction", "mammary gland involution", "negative regulation of cytokine-mediated signaling pathway" and "vesicle organization"



Multi-Omics Analysis

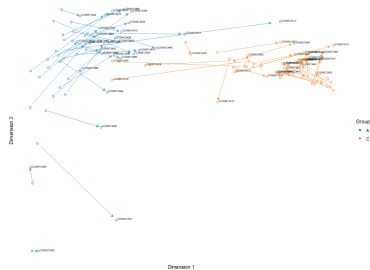
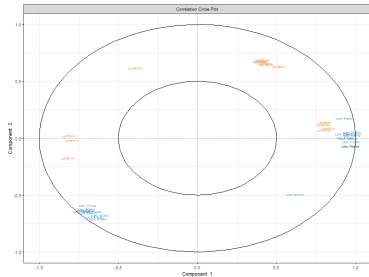


Figure: Results from the PLS analysis on the two omics.



Expression Genes	Gene linked to the correlated methylation site
S1PR1, STX11, TCF21, PEBP4, TM6SF1, CPA3, TCF21, LINC02538, FABP4, None	RPTOR, PAX8, CDKN1C, IGF2

Table 1: Gene expression and methylation sites variables negatively correlated from the sPLS technique (gene symbols)

Expression Genes	Gene linked to the correlated methylation site
CENPF, WDR62, TROAP, FOXM1, MELK, KIF15, None, KIF14, EPB41L3, FOXM1	IFNA17, KCNK16, CYP2E1, TCF7L2
S1PR1, STX11, TCF21, PEBP4, TM6SF1, CPA3, TCF21, LINC02538, FABP4, None	DLGAP2, RB1, GATA4
CENPF, WDR62, TROAP, FOXM1, MELK, KIF15, None, KIF14, EPB41L3, FOXM1	GABRA2, APOE

Table 2: Gene expression and methylation sites variables positively correlated from the sPLS technique (gene symbols)



Discussion and Future Works



- Supervised Learning Techniques showed perfect classification
- Most relevant pathways identified by our analysis and by Selamat et al. coincide
- We identified new significant pathways related to **angiogenesis** and **blood vessel formation**
- **Future work:**
 - Delve into the new identified pathways
 - Delve into the differences between smokers and non-smokers
 - Find new therapies using these results



Suhaida A. Selamat, Brian S. Chung, Luc Girard, Wei Zhang, Ying Zhang, Mihaela Campan, Kimberly D. Siegmund, Michael N. Koss, Jeffrey A. Hagen, Wan L. Lam, Stephen Lam, Adi F. Gazdar, and Ite A. Laird-Offringa.

Genome-scale analysis of dna methylation in lung adenocarcinoma and integration with mrna expression.

Genome Research, 22(7):1197–1211, 2012.



Thank you for the attention!



Figure: QR code to the GitHub repository with the report