

1 **Prediction of best features in heterogeneous Lung adenocarcinoma samples**  
2 **using Least Absolute Shrinking and Selection Operator.**

3

## 4 **Abstract**

5 This study aims to create a tumor heterogeneity-based model for predicting the best features of  
6 lung adenocarcinoma (LUAD) in multiple cancer subtypes using the Least Absolute Shrinking  
7 and Selection Operator (LASSO). The RNA-Seq raw count data of 533 LUAD samples and 59  
8 normal samples were downloaded from the TCGA data portal. Based on consensus clustering  
9 method samples was divided into two subtypes, and clusters were validated using silhouette  
10 width. Furthermore, we estimated subtypes for the abundance of immune and non-immune  
11 stromal cell populations which infiltrated cancer tissue. We established the LASSO model for  
12 predicting each subtype's best features. Enrichment pathway analysis was then carried out.  
13 Finally, the validity of the LASSO model for identifying features was established by the survival  
14 analysis. Our study suggests that the unsupervised clustering and Machine learning methods such  
15 as LASSO model-based feature selection can be effectively used to predict relevant genes which  
16 might play an essential role in cancer diagnosis.

## 17 **Keywords**

18 LASSO, Feature Selection, Heterogeneity, Biomarkers, Survival analysis, Expression patterns

19

## 20 **1. Introduction**

21 Lung cancer is reported to be the most deadly cancer [1]. Its shows the worst survival rate when  
22 compared with colon, breast, and pancreatic cancers combined. According to the American  
23 cancer society estimates both small cell and non-small cell lung cancer is the second most  
24 common cancer in both men and women. About 13% of all cancers are lung cancers. Lung  
25 Squamous cell carcinoma (LUSC) and Adenocarcinoma (LUAD) account for 15% and 85% of

26 all lung cancer, respectively [2]. Lung cancer is a highly heterogeneous disease, and the  
27 identification of cancer subtypes is decisive for clinicians. Genetic mutations, cancer  
28 microenvironment, immune, and therapeutic selection pressures all dynamically contribute to  
29 tumor heterogeneity. Heterogeneity may lead to cells with a differential molecular signature  
30 within single tumor tissue, and in some cases, it may contribute to therapy resistance [3].

31 Therefore, deciphering LUAD heterogeneity will have a significant impact in designing  
32 precision medicine strategy. Heterogeneous data suffers from a large number of covariates, and  
33 identification of variable selection is necessary to obtain more accurate predictions with a large  
34 number of covariates.

35 Many computer-based diagnostic and predictive models have been used for predicting the risk of  
36 a variety of cancers, such as logistic regression, Cox proportional hazard model, artificial neural  
37 networks, decision trees and support vector machines. Previous studies indicate standard  
38 stepwise selection approaches which are not best for regression models with a vast number of  
39 covariates [4]. Alternatively, least absolute shrinkage and selection operator (LASSO), has  
40 received much attention for identification and selection of best variables. LASSO was introduced  
41 by Robert Tibshirani in 1996 [5]. Regularisation and feature selection are the two critical tasks  
42 LASSO performs. LASSO estimates the regression coefficients by maximising the log-  
43 likelihood function with the constraint that the sum of the absolute values of the regression  
44 coefficients,  $\sum_{j=1}^k \beta_j$ , is less than or equal to a positive constant.

45 In this study, we selected the best features using LASSO established model. We downloaded the  
46 RNASeq data for LUAD samples from The Cancer Genome Atlas (TCGA) database. We  
47 differentiated the samples based on clusters into two subtypes to study the tumor heterogeneity.  
48 Differentially expressed genes (DEGs) were identified between two subtypes and normal groups,

49 followed by predicting relevant variables that are associated with the response variable using the  
50 LASSO model and validating the variables using survival analysis. We also estimated the  
51 population abundance of tissue-infiltrating immune and stromal cell populations in each subtype  
52 to decipher the inflammatory, antigenic, and desmoplastic reactions occurring in cancer tissue.  
53 Our study provides new insight into tumor heterogeneity and its importance in sample  
54 classification for predicting biomarkers of LUAD.

## 55 **2. Materials and Methods**

### 56 **2.1. Data source.**

57 The RNASeq data of LUAD, including 533 LUAD samples, and 59 normal samples were  
58 downloaded from the TCGA database (<https://portal.gdc.cancer.gov/>) in June 2019. All the raw,  
59 preprocessed data, images, supplemental tables and supporting files can be accessed at  
60 <https://github.com/AteeqKhaliq/LUAD/>.

### 61 **2.2. Data preprocessing and grouping.**

62 533 Primary solid Tumor samples and 59 Solid Tissue Normal samples were downloaded from  
63 the TCGA database. We calculated a variance stabilising transformation (VST) from the raw  
64 count data and transformed the counts yielding a matrix of values approximately homoskedastic.

### 65 **2.3. Molecular subtyping analysis.**

66 Feature dimension reduction was needed to remove irrelevant features and to reduce noises, and  
67 we used median absolute deviation (MAD) method and the features with  $MAD > 0.5$  were  
68 selected from set 2 groups. Consensus clustering (CC) [6] was used for the identification of  
69 subtypes on tumor samples. Silhouette width [7] was used to validate sample clustering.

## 70 **2.4. Differential gene expression analysis.**

71 Differential gene expression (DGE) was assessed by using the DESeq2 package [8] (Version  
72 1.24.0, <https://bioconductor.org/packages/release/bioc/html/DESeq2.html>) on Subtype-1 and  
73 Subtype-2 samples when compared with normal samples. Log2 fold change  $\geq$  +/- 2 and P-  
74 value  $< 0.05$  were used as the cut-off values to identify the DEGs.

## 75 **2.5. Construction of the LASSO model.**

76 Glmnet Package [9](Version 2.0-18, <https://cran.r-project.org/web/packages/glmnet/index.html>)  
77 was used to fit a generalised linear model via penalised maximum likelihood, LASSO model was  
78 established (Least Absolute Shrinkage and Selection Operator) on the DEGs from individual  
79 Subtype-1 and Subtype-2 cancer samples. We built a single pass (single fold) lasso-penalised  
80 model and performed 10-fold cross-validation to identify the best predictor.

## 81 **2.6. Survival Analysis.**

82 To find clinically or biologically meaningful biomarkers Kaplan-Meier survival curves [10] were  
83 generated by selecting the best predictors from individual subtypes. Kaplan-Meier curves were  
84 generated using the TRGAted [11] (<https://github.com/ncborcherding/TRGAted>) package  
85 implemented in R.

## 86 **2.7. Quantification of the abundance of the immune and stromal cell population in Cancer** 87 **Subtypes.**

88 We estimated the abundance of tissue-infiltrating immune and non-immune stromal cell  
89 populations in Subtype-1 and Subtype-2 samples. MCP-counter [12]  
90 (<https://github.com/ebecht/MCPcounter>) Package was used to estimate the Microenvironment

91 Cell Populations. VST normalised gene expression matrix was used for the estimation of an  
92 immune and stromal cell population in tumor samples.

### 93 **2.8. Gene classification and enrichment analyses.**

94 clusterProfiler [13](Version 3.12.0,  
95 <http://bioconductor.org/packages/release/bioc/html/clusterProfiler.html>) was used to annotate the  
96 DEGs from Subtype-1 and Subtype-2 groups to biological processes, molecular functions, and  
97 cellular components in a directed acyclic graph structure with a q-value cutoff of 0.2, Kyoto  
98 Encyclopedia of Genes and Genomes (KEGG) [14] was utilized to annotate genes to pathways,  
99 and Disease Ontologies.

100

## 101 **3. Results**

### 102 **3.1. Cancer subtype identification in LUAD samples**

103 We used Consensus clustering (CC) method, an unsupervised clustering method for grouping  
104 subtypes in LUAD. CC method is the most widely used for subtype discovery in high  
105 dimensional datasets. We used settings of the agglomerative hierarchical clustering algorithm  
106 using Pearson correlation distance. Two distinct clusters were discovered in our datasets, 89 and  
107 444 samples were clustered in Subtype-1 and Subtype-2 respectively (Fig.1.A). We have  
108 validated consistency within clusters of data using Silhouette width (Fig.1.B).

### 109 **3.2. Identification of DEGs in Subtype-1 and Subtype-2 LUAD samples.**

110 We compared the subtype-1 and subtype-2 with the normal samples and based on the p-value  
111 cutoff  $< 0.05$  and  $\log_2$  fold change  $> \pm 2$  we identified significant DEGs. 2033 genes were

112 upregulated, and 505 were downregulated in case of subtype-1 (Fig.2.A), and 5309 genes were  
113 upregulated, and 1219 were downregulated in case of subtype-2 (Fig.2.B) shows differential  
114 expression pattern in subtype-1 and subtype-2. The DEGs in both subtypes were used for  
115 building the LASSO predictive model and for the identification of best predictor genes in LUAD  
116 heterogeneous cancer data.

### 117 **3.3. LASSO model for identification of best predictive genes.**

118 LASSO was first described by Robert Tibshirani in 1996. Regularisation and feature selection  
119 are the two crucial task performed by LASSO. RNASeq datasets are high dimensional datasets,  
120 with smaller sample size and a large number of features also called small-n-large-p datasets ( $p$   
121  $\gg n$ ). High dimensional data will be sparse, and only a few features affect the response variable,  
122 and LASSO is known to identify the best features that affect the response variable. We deal with  
123 a  $p \gg n$  situation for feature selection in our Subtype-1 and Subtype-2 datasets, thus probably  
124 not all DEGs are relevant for the identification of features which affect the response variable.  
125 The purpose of our analysis is to identify the feature selection task and underline which genes are  
126 more relevant to predict and to classify them as biomarkers, to do so, we have used the LASSO  
127 model.

128 The result shows the trends of the 40 and 43 most relevant features selected by our model in  
129 subtype-1 and subtype-2 LUAD, respectively. The next step would be to find the most  
130 appropriate values for  $\lambda$  for our LASSO model. We analysed the  $\lambda$  value using the 10 fold cross-  
131 validation, between  $\lambda$  min that gives minimum mean cross-validated error or  $\lambda 1se$ , that gives a  
132 model such that error is within one standard error of the minimum. Using this analysis, we  
133 obtained the most relevant genes which are unique to subtype-1 and subtype-2 in the detection of  
134 a LUAD. A list of best-predicted genes available for each cancer subtypes is shown in Table 1.

### 135 **3.4. Analysis of the microenvironment of Subtype-1 and Subtype-2 LUAD samples.**

136 The abundance of tissue-infiltrating immune and non-immune stromal cell populations is highly  
137 informative. It has been shown that the extent of infiltrating immune cells is associated with  
138 disease prognosis. T-cell infiltrates, endothelial cells and fibroblasts are associated with a  
139 favorable outcome and also poor prognosis in some cancer types [15]. To understand the  
140 immunological microenvironment in our expression subset-1 and subset-2 we used MCP-counter  
141 method as described by Becht et al.[12]. The estimations consist of single sample scores which  
142 are computed on each sample independently in two subtypes. The heatmap shown in Figure 3  
143 clearly distinguish our subtype-1 and subtype-2 into two different categories based on tissue-  
144 infiltrating immune and non-immune stromal cell populations. Subtype-1 shows apparent  
145 increase in B lineage cells, monocytic cells, Cytotoxic lymphocytes, Natural killer cells, and CD  
146 8 T cells and Subtype-2 shows decreased levels of T-cells, macrophages, B cells, and natural  
147 killer (NK) cells, as well as endothelial cells and fibroblasts. Our study clearly distinguishes  
148 LUAD subtypes based on their inflammatory and stromal profiles, and Subtype-1 LUAD  
149 samples show increased expression of immunological markers than Subtype 2 samples.

### 150 **3.5. Pathway analysis.**

151 Subtype-1 and Subtype-2 showed distinct and unique features which are involved in cancer  
152 progression. Genes such as IL22RA2, PLA2G2C identified in subtype1 are upregulated and  
153 found to be involved in canonical cancer regulatory pathways such as the JAK-STAT signalling  
154 pathway and RAS signalling pathway (Fig. 4). The PLA2G2C gene is associated with alpha-  
155 Linolenic acid as well as ether lipid metabolism, and both are known to play a role in cancer  
156 progression [16]. Our model identified ONECUT1 gene in subtype1, which is associated with  
157 regulating pluripotency of stem cells and DEFA3 gene which is associated with Transcriptional



158 misregulation in cancer. The gene AWAT2 identified in subtype1 is found to be involved in  
159 Retinol metabolism, and studies suggest that retinoid signalling triggers tumor development [17].  
160 The NECTIN4 gene highlighted in our study is associated with Adherens junction, which plays  
161 an essential role in cancer initiation and progression [18]. PYCR1 gene involved in proline  
162 metabolism which is highly correlated with cancer is also found to be differentially expressed in  
163 subtype1. EFNA3 is upregulated in Subtype-2, and previous studies show that it contributes to  
164 the metastatic spread of breast cancer [19]. Our model predicted that increased expression of  
165 LGR4, TESC, TOP2A, and ZNF695 is suggestive of increased invasive and metastatic activity in  
166 Subtype-2 [20]. Decreased expression of FGF10 is seen in Subtype-2 and predicted by our model  
167 is suggestive of dysplasia in LUAD subtype-2 samples [21].

### 168 **3.6. Validation with survival analysis.**

169 The LUAD samples were classified into two subtypes based on the consensus clustering method.  
170 Overall survival analysis for the most predictive genes identified by our model in Subtype-1 and  
171 Subtype-2 groups was conducted. The genes predicted by our LASSO model accurately  
172 predicted the outcome of a patient's survival using gene expression data. Genes such as LAD1,  
173 NECTIN4, SLC25A48, LYVE1, EFNA3, FGF10, HELT, HTR3C, and LGI3 yielded accurate  
174 predictions for the risk of LUAD and can be used in cancer prediction. Survival plots and its p-  
175 value is shown in Supplemental Figures.

## 176 **4. Discussion**

177 In this study, we developed a LASSO based model for accurate feature selection in LUAD. Our  
178 model removed variables that are redundant and removed features which do not add any valuable  
179 information in disease prediction. Analysis using the survival data for the predicted genes

180 showed that the model could effectively predict genes responsible for disease prognosis in high  
181 dimensional datasets. Deciphering cancer heterogeneity is very critical in understanding cancer  
182 dynamics and also for the development of personalised cancer treatment [22]. We used  
183 Consensus clustering method to determine the number of clusters in our samples, and we  
184 clustered the samples into two groups which produced optimal silhouette width for the  
185 determined clusters. Differential gene expression analysis showed distinct expression patterns in  
186 both Subtype-1 and Subtype-2.

187 The number of differentially expressed genes was very high, and in these situations, it is difficult  
188 to predict the relevant variables. LASSO model was established around DE genes in Subtype-1  
189 and Subtype-2 groups. Not all the expressed genes were relevant, and our model predicted the  
190 most relevant genes which were involved in disease progression. Decreased expression of  
191 LYVE1 and MED28P8 in Subtype-1 and FGF10, HELT, HTR3C, LIG3, PACRG, PLAC9P1,  
192 and STX11 in Subtype-2 showed worse overall survival in LUAD samples. Whereas increased  
193 expression of genes such as CDC37P2, DCST1, IL22RA2, LAD1, NECTIN4, SLC25A48,  
194 TMEM51.AS1, TMPRSS4, VPS9D1-AS1, AWAT2, BCL9P1 and CD5L in Subtype-1 and  
195 EFNA3, LGR4, TESC, TOP2A and ZNF695 in subtype-2 showed decreased overall survival in  
196 LUAD samples.

197 Long intervening noncoding RNAs (lncRNAs) are known to be critical regulators of numerous  
198 biological processes, and substantial evidence supports that lncRNA expression plays a  
199 significant role in tumorigenesis and tumor progression. Increased expression of LINC00862 in  
200 Subtype-1 samples correlates with worse survival in LUAD subtypes. Whereas, decreased  
201 expression of LINC00211 in Subtype-1 and LINC01506, LINC01785 and LINC01996 in  
202 Subtype-2 showed worse survival in LUAD samples. Our LASSO model predicts the most

203 relevant and distinct genes from Subtype-1 and Subtype-2 samples which might be an important  
204 factor in cancer diagnosis and management.

205 The best predictors for subtype-1 and subtype-2 from the LASSO model were found to be  
206 involved in several regulatory pathways. Subtype-1 gene such as S100A12 is a vital serum  
207 inflammatory marker and has been illustrated in several cancer types such as oropharyngeal  
208 squamous cell carcinoma and gastric cancers [23]. Subtype-1 samples show increased expression  
209 of CD5L and TMPRSS4, which induces cancer stem cell-like properties and promotes malignant  
210 transformation by limiting lung epithelial cell apoptosis and promoting immune escape in  
211 NSCLC patients [24]. Long noncoding RNA VPS9D1-AS1 overexpression in subtype-1 predicts  
212 poor prognosis and serves as a biomarker to predict the prognosis of NSCLC [25].

213 Overexpression of nectin-4 oncoprotein, LYVE-1/PCAB, PROM2, and LAD1 are associated  
214 with poor overall survival in subtype-1 samples and can be considered as candidate serum and  
215 tissue biomarker as well as therapeutic target [26].

216 Overexpression of Lgr4 in Subtype-2 samples promotes tumor aggressiveness may potentially  
217 become a novel target for cancer therapy [27]. Subtype-2 samples show upregulation of TOP2A  
218 and ZNF695 is associated with worse prognosis and induces overrepresentation of growth and  
219 proliferation pathways and can act as prognostic and predictive markers [28]. Hypoxia-inducible  
220 oncogene EFNA3 is overexpressed in subtype-2 samples may play a critical role in the focal  
221 adhesion kinase (FAK) signalling and VEGF-associated tumor angiogenesis pathway [19].

222 Downregulation of tumor suppressor gene STX11 in Subtype-2 samples predicts poor prognosis.  
223 Various studies indicate the role of STX11 expression in suppressing the proliferation of T-cell  
224 [29].

225 Our model predicts lncRNAs such as LINC01506, LINC01785, LINC01996, LINC00862, and  
226 LINC02014 were expressed in subtype-1 and subtype-2 samples which can be considered as  
227 potential biomarkers and shows poor overall survival in LUAD. LncRNAs might be used as  
228 biomarkers and drug targets for early diagnosis, prognosis and personalised treatment of LUAD  
229 patients.

230 Our study suggests that Consensus Clustering methods and LASSO combined will help us to  
231 develop a model with the most appropriate characteristics. Consistent with these finding,  
232 different subtypes showed distinct unique features which underscore the importance of sample  
233 grouping and assessment. Furthermore, Survival analyses validate that the survival time of the  
234 predicted genes correlates with gene expression pattern, which is recognisably different in both  
235 the Subtypes, indicating that LASSO model could effectively be used to overcome the feature  
236 selection problem and can be used for accurate prediction of risk in LUAD.

237

238 **Funding:** This research did not receive any specific grant from funding agencies in the public,  
239 commercial, or not-for-profit sectors.

240

241 **Table description:**

242 **Table 1:** List of best-predicted genes by LASSO Model for cancer Subtype-1 and Subtype-2

243 **Figures description:**

244 **Figure 1**

245 **Fig. 1.** LUAD sample clustering. (A) CC plot shows the clustering of samples into two distinct  
246 subtypes. (B) Silhouette plot for validating the sample clustering. (C) PCA plot indicates distinct  
247 sample groups.

248 **Figure 2:**

249 **Fig. 2.** DGE analysis at standard cutoff of  $\pm 2$  fold change at p value of  $<0.05$ . (A) DGE  
250 volcano plot for Subtype-1 samples. (B) DGE volcano plot for Subtype-2 samples.

251 **Figure 3:**

252 **Fig. 3.** The heatmap distinguish subtype-1 and subtype-2 into two different categories based on  
253 tissue-infiltrating immune and non-immune stromal cell population.

254 **Figure 4:**

255 **Fig.4. Fig.4. Pathway analysis for the best predicted genes by LASSO model. (A) Subtype-1**  
256 **LASSO predicted genes pathway analysis. (B) Subtype-2 LASSO predicted genes pathway**  
257 **analysis.**

258

259

260

261 **References**

- 262 [1] R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, 2015., CA. Cancer J. Clin. 65 (2017)  
263 5–29. doi:10.3322/caac.21254.
- 264 [2] K. Inamura, Lung Cancer: Understanding Its Molecular Pathology and the 2015 WHO  
265 Classification., Front. Oncol. 7 (2017) 193. doi:10.3389/fonc.2017.00193.
- 266 [3] H.A. Bolck, C. Corrà, A. Kahraman, A. von Teichman, N.C. Toussaint, J. Kuipers, F.  
267 Chiovaro, V.H. Koelzer, C. Pauli, W. Moritz, P.K. Bode, M. Rechsteiner, N.  
268 Beerenwinkel, P. Schraml, H. Moch, Tracing Clonal Dynamics Reveals that Two- and  
269 Three-dimensional Patient-derived Cell Models Capture Tumor Heterogeneity of Clear  
270 Cell Renal Cell Carcinoma., Eur. Urol. Focus. (2019). doi:10.1016/j.euf.2019.06.009.
- 271 [4] N. Houssami, L. Irwig, J.M. Simpson, M. McKessar, S. Blome, J. Noakes, The influence  
272 of clinical information on the accuracy of diagnostic mammography, Breast Cancer Res.  
273 Treat. 85 (2004) 223–228. doi:10.1023/B:BREA.0000025416.66632.84.
- 274 [5] R. Tibshirani, Regression Shrinkage and Selection Via the Lasso, J. R. Stat. Soc. Ser. B.  
275 58 (1996) 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x.
- 276 [6] S. Monti, P. Tamayo, J. Mesirov, T. Golub, Consensus clustering: A resampling-based  
277 method for class discovery and visualization of gene expression microarray data, Mach.  
278 Learn. (2003). doi:10.1023/A:102394950.
- 279 [7] P.J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster  
280 analysis, J. Comput. Appl. Math. 20 (1987) 53–65. doi:10.1016/0377-0427(87)90125-7.
- 281 [8] M.I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for

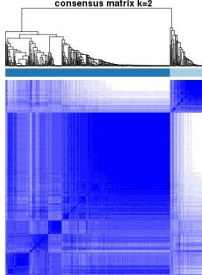
- 282 RNA-seq data with DESeq2, *Genome Biol.* 15 (2014) 550. doi:10.1186/s13059-014-0550-  
283 8.
- 284 [9] J. Friedman, T. Hastie, R. Tibshirani, Regularization Paths for Generalized Linear Models  
285 via Coordinate Descent., *J. Stat. Softw.* 33 (2010) 1–22.  
286 <http://www.ncbi.nlm.nih.gov/pubmed/20808728>.
- 287 [10] E.L. Kaplan, P. Meier, Nonparametric Estimation from Incomplete Observations, *J. Am.*  
288 *Stat. Assoc.* 53 (1958) 457–481. doi:10.1080/01621459.1958.10501452.
- 289 [11] N. Borchherding, N.L. Bormann, A.P. Voigt, W. Zhang, TRGAted: A web tool for survival  
290 analysis using protein data in the Cancer Genome Atlas., *F1000Research.* 7 (2018) 1235.  
291 doi:10.12688/f1000research.15789.2.
- 292 [12] E. Becht, N.A. Giraldo, C. Germain, A. de Reyniès, P. Laurent-Puig, J. Zucman-Rossi,  
293 M.-C. Dieu-Nosjean, C. Sautès-Fridman, W.H. Fridman, Immune Contexture,  
294 Immunoscore, and Malignant Cell Molecular Subgroups for Prognostic and Theranostic  
295 Classifications of Cancers., *Adv. Immunol.* 130 (2016) 95–190.  
296 doi:10.1016/bs.ai.2015.12.002.
- 297 [13] G. Yu, L.-G. Wang, Y. Han, Q.-Y. He, clusterProfiler: an R package for comparing  
298 biological themes among gene clusters., *OMICS.* 16 (2012) 284–7.  
299 doi:10.1089/omi.2011.0118.
- 300 [14] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, M. Hirakawa, KEGG for representation  
301 and analysis of molecular networks involving diseases and drugs., *Nucleic Acids Res.* 38  
302 (2010) D355-60. doi:10.1093/nar/gkp896.

- 303 [15] N.A. Giraldo, E. Becht, F. Pagès, G. Skliris, V. Verkarre, Y. Vano, A. Mejean, N. Saint-  
304 Aubert, L. Lacroix, I. Natario, A. Lupo, M. Alifano, D. Damotte, A. Cazes, F. Triebel,  
305 G.J. Freeman, M.-C. Dieu-Nosjean, S. Oudard, W.H. Fridman, C. Sautès-Fridman,  
306 Orchestration and Prognostic Significance of Immune Checkpoints in the  
307 Microenvironment of Primary and Metastatic Renal Cell Cancer., *Clin. Cancer Res.* 21  
308 (2015) 3031–40. doi:10.1158/1078-0432.CCR-14-2926.
- 309 [16] J.P. Chamberland, H.-S. Moon, Down-regulation of malignant potential by alpha linolenic  
310 acid in human and mouse colon cancer cells, *Fam. Cancer.* 14 (2015) 25–30.  
311 doi:10.1007/s10689-014-9762-z.
- 312 [17] X.-H. Tang, L.J. Gudas, Retinoids, Retinoic Acid Receptors, and Cancer, *Annu. Rev.*  
313 *Pathol. Mech. Dis.* 6 (2011) 345–364. doi:10.1146/annurev-pathol-011110-130303.
- 314 [18] V. Vasioukhin, Adherens Junctions and Cancer, in: *Subcell. Biochem.*, 2012: pp. 379–  
315 414. doi:10.1007/978-94-007-4186-7\_16.
- 316 [19] L. Gómez-Maldonado, M. Tiana, O. Roche, A. Prado-Cabrero, L. Jensen, A. Fernandez-  
317 Barral, I. Guijarro-Muñoz, E. Favaro, G. Moreno-Bueno, L. Sanz, J. Aragonés, A. Harris,  
318 O. Volpert, B. Jimenez, L. del Peso, EFNA3 long noncoding RNAs induced by hypoxia  
319 promote metastatic dissemination., *Oncogene.* 34 (2015) 2609–20.  
320 doi:10.1038/onc.2014.200.
- 321 [20] X. An, F. Xu, R. Luo, Q. Zheng, J. Lu, Y. Yang, T. Qin, Z. Yuan, Y. Shi, W. Jiang, S.  
322 Wang, The prognostic significance of topoisomerase II alpha protein in early stage  
323 luminal breast cancer, *BMC Cancer.* 18 (2018) 331. doi:10.1186/s12885-018-4170-7.
- 324 [21] E. El Agha, S. Bellusci, Walking along the Fibroblast Growth Factor 10 Route: A Key



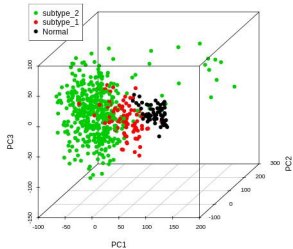
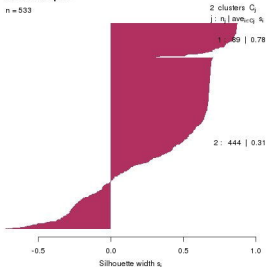
- 325 Pathway to Understand the Control and Regulation of Epithelial and Mesenchymal Cell-  
326 Lineage Formation during Lung Development and Repair after Injury, *Scientifica* (Cairo).  
327 2014 (2014) 1–20. doi:10.1155/2014/538379.
- 328 [22] I. Dagogo-Jack, A.T. Shaw, Tumour heterogeneity and resistance to cancer therapies, *Nat.*  
329 *Rev. Clin. Oncol.* 15 (2018) 81–94. doi:10.1038/nrclinonc.2017.166.
- 330 [23] G.-Q. Huang, Z.-P. Ke, H.-B. Hu, B. Gu, Co-expression network analysis of long  
331 noncoding RNAs (lncRNAs) and cancer genes reveals SFTA1P and CASC2 abnormalities  
332 in lung squamous cell carcinoma., *Cancer Biol. Ther.* 18 (2017) 115–122.  
333 doi:10.1080/15384047.2017.1281494.
- 334 [24] A.L. de Aberasturi, M. Redrado, M. Villalba, L. Larzabal, M.J. Pajares, J. Garcia, S.R.  
335 Evans, D. Garcia-Ros, M.E. Bodegas, L. Lopez, L. Montuenga, A. Calvo, TMPRSS4  
336 induces cancer stem cell-like properties in lung cancer cells and correlates with ALDH  
337 expression in NSCLC patients., *Cancer Lett.* 370 (2016) 165–76.  
338 doi:10.1016/j.canlet.2015.10.012.
- 339 [25] J. Tan, L. Yang, Long noncoding RNA VPS9D1-AS1 overexpression predicts a poor  
340 prognosis in non-small cell lung cancer, *Biomed. Pharmacother.* 106 (2018) 1600–1606.  
341 doi:10.1016/j.biopha.2018.07.113.
- 342 [26] S. Huber, N. Gagliani, L.A. Zenewicz, F.J. Huber, L. Bosurgi, B. Hu, M. Hedl, W. Zhang,  
343 W. O'Connor, A.J. Murphy, D.M. Valenzuela, G.D. Yancopoulos, C.J. Booth, J.H. Cho,  
344 W. Ouyang, C. Abraham, R.A. Flavell, IL-22BP is regulated by the inflammasome and  
345 modulates tumorigenesis in the intestine, *Nature.* 491 (2012) 259–263.  
346 doi:10.1038/nature11535.

- 347 [27] B. Tan, X. Shi, J. Zhang, J. Qin, N. Zhang, H. Ren, M. Qian, S. Siwko, K. Carmon, Q.  
348 Liu, H. Han, B. Du, M. Liu, Inhibition of Rspo-Lgr4 Facilitates Checkpoint Blockade  
349 Therapy by Switching Macrophage Polarization, *Cancer Res.* 78 (2018) 4929–4942.  
350 doi:10.1158/0008-5472.CAN-18-0152.
- 351 [28] G.-X. Hou, P. Liu, J. Yang, S. Wen, Mining expression and prognosis of topoisomerase  
352 isoforms in non-small-cell lung cancer by using Oncomine and Kaplan–Meier plotter,  
353 *PLoS One.* 12 (2017) e0174515. doi:10.1371/journal.pone.0174515.
- 354 [29] N. Yoshida, S. Tsuzuki, K. Karube, T. Takahara, M. Suguro, H. Miyoshi, M. Nishikori,  
355 M. Shimoyama, K. Tsukasaki, K. Ohshima, M. Seto, STX11 functions as a novel tumor  
356 suppressor gene in peripheral T-cell lymphomas, *Cancer Sci.* 106 (2015) 1455–1462.  
357 doi:10.1111/cas.12742.
- 358



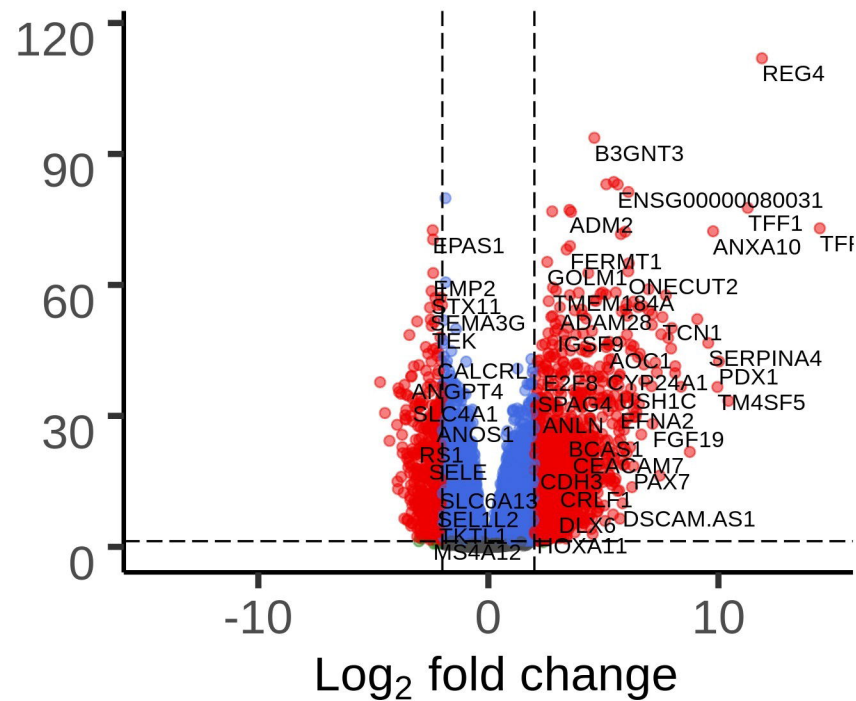
### Silhouette plot

n = 533



## subtype\_1\_vs\_Normal

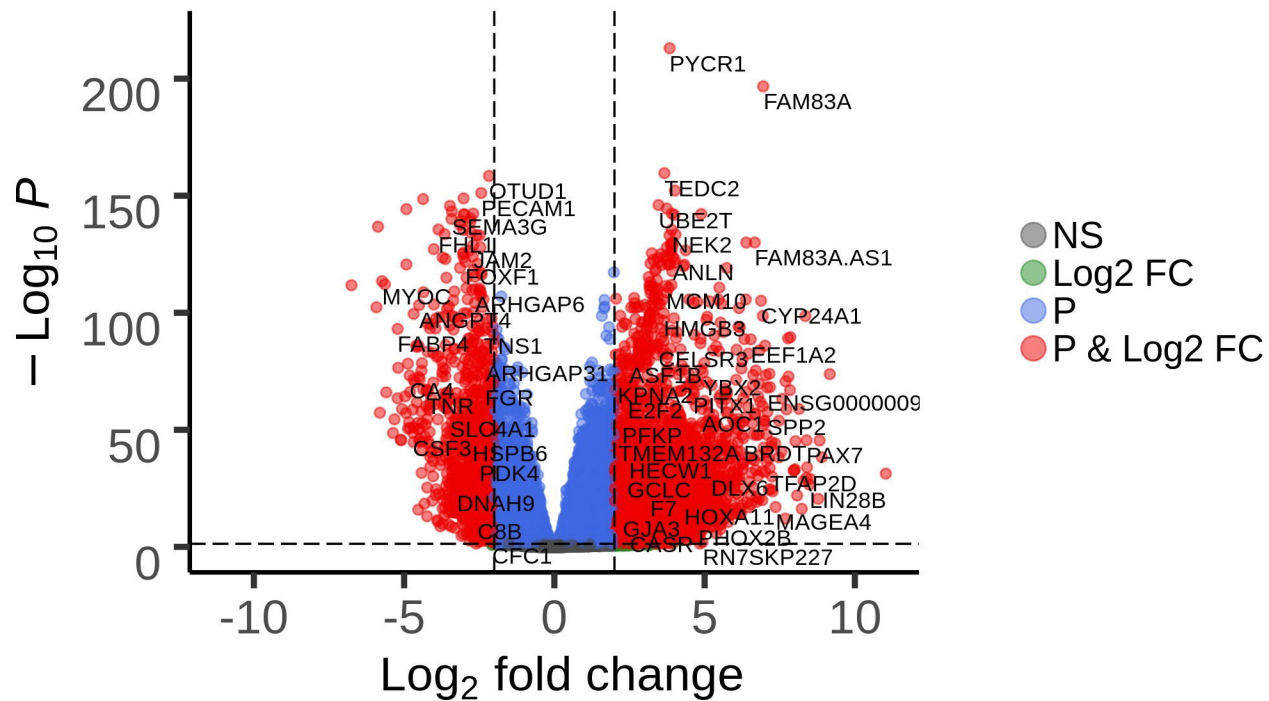
differential expression



Total = 51936 variables

## subtype\_2\_vs\_Normal

differential expression



Total = 51936 variables

