Discovering genes and molecular mechanisms beyond differential gene expression, using machine learning k-means clustering

|  |  |
| --- | --- |
| The objective of the study: | The aim of this analysis is to study the molecular mechanisms beyond differential gene expression using K-means Cluster Analysis in machine learning. This approach could be beneficial for therapeutics discovery or personalized treatment. |
| Dataset: | The dataset used in this analysis was GSE191037 containing squamous carcinoma A431 cells treated with itraconazole.  The cells in the control group includes 3 pairs of A431 cancer cells while 3 treatment groups were treated with itraconazole (1μM, 48h). A431 cells are a model human cell line (epidermoid carcinoma). |
| Description: | According to the DrugBank, Itraconazole was used for the treatment of various fungal infections in immunocompromised and non-immunocompromised patients, such as pulmonary and extrapulmonary blastomycosis, histoplasmosis, and onychomycosis. [Wishart et al].  Understanding cellular mechanisms are important in drug discoveries and personalized treatment. The interaction between the cellular function and gene interaction could give us insight on the mechanisms of action for tailoring drugs to personalise treatment.  The analysis was performed using DESeq2 and machine learning K-means Cluster analysis. R programming language was used.  Fold change is used in this analysis to understand the mechanisms of differentially expressed genes and DESeq2 algorithm tests for differential expression based on a model using the negative binomial distribution.  Itraconazole is not only used to treat fungal infections, it is also used in the treatment of cancers such as prostate, breast, lung and BCC (skin cancer). Cellular mechanisms and genetic variations could respond either negatively or positively based on the individual person.  The analysis was conducted to understand cell mechanism of action and expression of genes. There could be multiple contributions that could lead to responses to the treatment of itraconazole. Hence, using machine learning K-means Cluster Analysis might assist in finding related genes which are groups together based on significant p-values due to the treatment of itraconazole.  The data are segmented into *k* clusters using the k-means algorithm. Therefore, each cluster has a center often called a centroid and terminates at local optimum. However, this algorithm is sensitive to outliers. The k-means technique was applied after the significant result from DESeq2 algorithm. Hence, outliers in the dataset seems not an issue.  The assumption in k-means, is that we consider “all clusters are the same size” ((Winn et al., n.d.) |
| Method: | The dataset consists of Track\_id, Gene Name, Locus, 3 groups of control and 3 groups of treatment groups.  The data were cleaned up and pre-processed using R programming before applying DESeq2 algorithm.  The following libraries were loaded to do further analysis and visualization.  library(tidyverse) # data manipulation  library(cluster) # clustering algorithms  library(factoextra) # clustering algorithms & visualization  The cut of point p-value was less than 0.05 and the result contains 767 rows of genes that were of below p-value 0.05. Generally, further analysis was done using the top 10 genes. However, I conducted further analysis utilizing the significant p-values (< 0.05) in the machine learning K-Means Cluster Analysis.  Machine learning k-means cluster analysis was applied to k=2, k=3, k=4, k=5 and used the method below to find optimal k cluster for further analysis. The analysis includes k up to 15 for optimal k clusters.    Fig 1: The Elbow method used for selection of optimal k cluster |
| Data Visualization: | Fig 2: Visualization of original dataset between control and Itraconazole treatment groups  Chart, box and whisker chart  Description automatically generated  Fig 3: Boxplot with means between control and Itraconazole treatment groups    Fig 4: Violin plot normal distribution with means between control and Itraconazole treatment groups  The variance of the data set using sigma and mean of genes across samples between control and treatment groups were:    Fig 5: The green line indicates in Poisson distribution the variance equal means. Which indicates the dataset was not homoscedastic. Hence, the data were transformed using a log.    Fig 6: Principal Component Analysis (PCA), comparison between control groups (CTR) and treatment groups (TRM). |
| Results/  Finding: | The result from DESeq2:  *Out of 11780 with nonzero total read count*  *adjusted p-value < 0.1*  *LFC > 0 (up) : 184, 1.6%*  *LFC < 0 (down) : 155, 1.3%*  *outliers [1] : 0, 0%*  *low counts [2] : 8447, 72%*  *(mean count < 11)*  Visualization of resulting K-Means Cluster Analysis k=2, k=3, k=4 and k=5 are as follows:    Fig 7: Visualization of K-Means Cluster Analysis k=2, k=3, k=4 and k=5  The Above visualization and silhouette method indicated that k=4 is optimal cluster to do further analysis.  The cut-off point for the k cluster was 4 (k=4). The result is as follows:  *Cluster means: k=4*  *log2FoldChange pvalue*  *1 1.3316765 0.01452048*  *2 4.1112117 0.00989127*  *3 -0.9994537 0.01467252*  *4 -3.3163967 0.01286101*  The average size of each cluster was (363, 48, 290, 66) respectively. The result indicated negatively and positively express genes with the treatment of the Itraconazole drug at (1μM, 48h).  The significate results including k=4 clusters and genes (vector) are as follows:   |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | Cluster 1 | | | | | | | | CYP51A1  PRSS22  IBTK  TNFRSF12A  CYTH3  CLCN6  PLAUR  ANLN  TMEM159  ELOVL5  GPRC5A  CYP24A1  SNAPC1  ARNTL2  ADSS  SCML1  HDAC9  LIMA1  PRSS8  MSMO1  LAMC2  SBNO2  ZC3H15  ME1  GOLGA5  PFKP  IDI1  KLF6  ACSL4  MAPK6  CYBRD1  RDH11  ACTN1  IGF2BP2  LMAN1  ACTR6  NFKB2  ITCH  FDFT1  WDFY1  PPP1R15A  TPX2  FXYD5  FLNA  ZNF165  SERPINB2  TAF13  ERO1A  TFAP2A  HMGA1  IER3 TMED7.TICAM2 | ICAM1  CMTM6  CDV3  EZR  TGFB2  HSP90AB1  ITPR3  DSP  SCD  PIK3IP1  XBP1  BIK  HMOX1  TTLL12  RANGAP1  PYGL  HIF1A  NFKBIA  PRELID3B  CDC25B  MYL12A  MIB1  STAG2  PLS3  SMS  EEA1  LACTB  RIPK2  SQLE  RELB  PPP1R13L  LSR  MET  SERPINE1  DVL1  EIF3A  CCSER2  SMC3  KPNB1  SYNGR2  ALDOC  NMU  AREG  CD2AP  PIM3  STK39  SMURF1  PJA2  CENPW  FGFBP1  SDCBP  LYN | KLHL5  OSBP  CCND1  DSE  SERINC1  HINT3  NCOA7  PTP4A1  HECA  SLC39A7  PHF1  TTK  ERBIN  HMGCR  AMOTL2  GALNT3  FHL2  RND3  ERRFI1  RNF19B  GADD45A  CD46  F3  IRF6  KLF7  TNFAIP3  RAB32  FOXO3  SPP1  UBE2B  PTBP3  CTNNAL1  SET  ATL2  IFIT3  IFIT2  AVPI1  DUSP1  PANK3  ACAT2  DUSP4  TMEM156  PLAU  P4HA1  OPTN  PFKFB2  ACSL3  CHPF  CSE1L  PMEPA1  FAM72D  CDKN2AIPNL | CDKN1A  SOX4  SOX9  TNFSF9  C3  GPCPD1  RRBP1  EIF2S2  NR1D1  SBDS  GLIS2  PTPN12  CDC42EP1  RAC2  LIF  KRT17  DNAJB9  CALU  MINDY2  COPB1  SLC44A2  KLF16  SAT1  LDLR  GDF15  THEMIS2  ULBP2  ACSS2  ACLY  PPFIA1  TOP2A  WBP2  OSER1  ATP13A3  TTC9  CDK7  ARL8B  SPIRE1  SLC43A3  COL4A2  EHF  KRT7  TMBIM1  CKAP4  FLNB  LCP1  SKIL  LRRC8A  SLC31A1  PIM1  DUSP5  CEP55 | SEC24B  AMIGO2  PHLDA1  DUSP6  CDR2  OSBPL1A  GATA6  C18orf8  NPC1  CBLC  EPHA2  CYR61  TINAGL1  ATP1B1  PFDN2  MCL1  DUSP10  MBOAT2  GOLGA4  TMF1  PHLDB2  CCNA2  OSMR  PLK2  TNIP1  TNFRSF21  PPP1R18  NFKBIE  EGFR  CCT6A  MAL2  UGCG  LCN2  PROSER2  GLUD1  CELF1  ADAM8  ASAP2  CACUL1  MBNL1  HNRNPDL  PTPRK  WNT7A  CCT8  ZFAND3  B3GNT7  EIF4A2  MMP14  BCAR3  RABGGTB  ATAD1  ARFGAP3 | TSC22D3  ETS2  FMNL2  NCK1  PPP1R15B  CDA  GNE  FDPS  SQSTM1  ITGA5  RBM15  RPL22L1  TPRA1  LIPH  GNL3  RNF168  MB21D1  OSGIN2  PDP1  UBAP1  VDAC2  DDX21  ARL5B  GPR176  SLFN5  NAV2  C15orf48  GOLGA2  MVD  NLRP7  KRT80  SRRM2  DDIT4  ARF4  TAP1  FAM84B  FEM1B  GDI1  PHACTR4  ZDHHC18  MICB  MICA  HLA.E  MZT1  IPO7  KRT6A  KRT81  PPIG  ITGAV  USO1  AREGB  HACD2 | PCSK9  CLIC4  CKAP2L  FASN  CNBP  HNRNPF  CSGALNACT2  IFFO2  ALCAM  KRT86  KIF5B  SOX7  JUNB  ENC1  DHCR7  MYEOV  TRIB1  DAG1  NET1  NR1D2  MARCKSL1  CKAP5  FOSL1  CCDC85B  SFN  ETV4  MEX3C  ANKLE2  PAWR  JUN  THBD  FGD6  GPR157  RTKN2  ACBD3  GBP6  MACC1  ACTG1  TACSTD2  CSF1  SOCS3  KPNA7  KRT6B  ZFP36L1  KPNA4  UBE2H  KRT16  ARL4C  CD55  LAMB3  PPP1CB |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | Cluster 2 | | | | | | | | IL32  KRT31  HAS3  KRT23  VNN1  HMGCS1  IL1A | PI3  EREG  IL1B  FFAR2  BHLHE40  LPIN1  SAA2 | CXCL11  CXCL8  SAA1  PLK3  CREG2  ZNF114  ARID3B | SEMA7A  ANXA3  RHCG  S100A7  TMEM171  ATF3  PIGR | AMTN  CGB5  C6orf222  CGB8  C10orf105  CXorf49B  YOD1 | INSIG1  TMEFF1  SAA2.SAA4  TET1  C7orf55.LUC7L2  FAM47E.STBD1 | NEURL3  CXCL1  STARD4  IL36G  ABTB2 CXorf49  FST |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | Cluster 3 | | | | | | | | AOC1  ARF5  TRAPPC6A  RPS20  EHD2  HMGB3  AKR7A2  SZRD1  SEC61A1  RPL18  BCAS1  TNFRSF1A  IFI35  CAMK2A  DHRS9  ENO1  FBLN1 RPL13A  RPL11  SYTL1  HSPG2  RPS8  LMO4  CELSR2  CRABP2  ANXA9  SLC39A1  EPHX1  IGFBP3  MAGEA4  RPL10  RPL7A  PTGES  CALR  ARL14  HIST1H2AC  UBA7  RPS17  RPL35A  RIPK4  COA3  UQCR10 | BCKDHB  RPS5  GSTP1  CTTN  EPB41L1  DYNLL1  RPL6  RPLP0  PEBP1  PSME1  PHGDH  TMEM14A  NDUFB7  TIMP3  RPL3  NDRG3  ADNP  CTSZ  PRPF6  EPS8  ADGRF1  GTF3C6  MARVELD1  RPL30  HIST1H2BD  LAD1  CBR1  PSMB4  BTG2  AMFR  CALM3  RPL8  CAMK2N1  H3F3A  S100A11  OLFML2A  KRT5  RXRA  FOXD2  ANXA8L2  ZNF395  BCAM | PLP2  VGLL1  CTSH  NUCB1  AES  RPS19  PTPRS  CAV1  EPHB6  CHCHD2  EDF1  UNC5B  VSIR  RPL28  RPL19  LGALS3BP  RAB5C  SLC9A3R1  RPL34  RPL9  CITED2  CTSB  NUDT2  SYK  AQP3  C10orf10  MOAP1  RPL27A  BEX3  PPIB  STAT6  RPL13  MT.ND2  MT.ND5  MT.CO1  HMGN2  MT.ND3  MT.ND1  MT.ATP6  RPL39  MT.CO3  TUBB4B | RPS13  ARHGDIB  TPI1  SASH1  MCM3  RPS12  PERP  GMDS  SEMA5A  REEP6  ATP6V1B1  TMEM59  HDAC1  ACADM  NENF  CTSD  RPS25  HMGN3  MT1E  SF3B5  SLC16A5  UBB  FOS  GSTA4  KRT15  KRT13  DLK2  RPS21  RPS7  SNCG  H1F0  PNRC2  AKR1C3  S100A4  PLXNB2  ANXA4  RPS26  S100A10  RPL37A  HIST1H2BK RPL17  AGRN | VAMP8  NDUFA8  CLU  CSTA  CAT  RPL5  CDKN2C  FAM210B  SLPI  NCOA5  RPS10  KLHDC3  MEA1  AHNAK  TM9SF2  PPDPF  RPL23  SNRPD2  SNRPB  ITPA  ID1  ERGIC3  CAPNS1  RPL4  IL20RB  RPL15  TP53I11  MUC20  ACAD9  RPS27  CALML3  CALML5  FUCA1  TCEA3  PSMB8  HSPA1B  HSPA1A  SAMD9 POLR2J2.1  MARCKS  HIST1H1C | PRDX5  IFI6  SGPP1  FGD3  ECHS1  DGCR6L  ATP6V1F  RPL36  BST2  LAMA5  ASS1  RPL27  NDUFA2  ANO1  TNS4  LGALS3  TRIM22  FLOT2  CYP2S1  C19orf33  PSCA  LY6D  EEF2  IGFBP6  MRPL58  FTH1  FADD  SCARA3  MAT2A  SLC25A6  HINT1  S100A6  RPS4X  TFDP1  RPL23A  TPM2  MT.ND6  MT.CO2  MT.CYB  RPL10A TXNIP | PCNA  RARRES3  GSTM1  PSAT1  GLUL  HS6ST1  RPL35  RPS6  RPLP1  CYP1B1  IDH1  GLTP  RPS2  NDUFB10  RPS11  ADIRF  EIF4EBP2  ANXA8L1  NDUFC2  PMVK  PBXIP1  CLDN1  ITPRIPL2  SELENOH  CRIP1  EMP2  MT.ATP8  DHFR  RPL41  RPS18  GPX1  RPS28  ATP5J2  ARPC1A  DDOST  UBE2V1  EEF1G  RPL36A  HNRNPH2  CNPY2  ANXA8 |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | Cluster 4 | | | | | | | | IL20RA  CLCA4  CD74  SLAMF7  NTN1  MPPED2  FGFR3  ST6GALNAC1  ABLIM1  PSAPL1 | VAV3  GCHFR  RAB15  CTDSPL  PLIN2  LYPD1  ABI3BP  ITGB2  IL6R | SCNN1B  CAVIN2  KRT4  TMEM37  ID4  MFSD4A  ABO  GPX2  ZBED2 | ALG1L  HLA.DRB1  LDLRAD1  HLA.DMA  HLA.DRA  JPT2  UGT1A3  UGT1A4  SELENOP | ALDH1A1  FBP1  SCNN1G  KLK11  KLK13 ALPP  ALPPL2  HPGD  IVL  MXRA5 | TNFSF10  OVOL2  BPIFB1  LSP1  IDO1 TMPRSS2  TMEM173  PBX1 CDH5  INHBB | UPK3BL1  HIST2H4B  HIST2H4A DOCK8  ALDH3A1  FOLR1  EPPK1  MMP12  AGR2 |   As you can see from the above K-means cluster analysis result, k=4, cluster 2 and cluster4 Log2FC were (4.11, -3.32) respectively. Hence, further analysis will perform using the genes from (k=4), clusters 2 and 4. |
| K-Means Cluster Validation | The silhouette coefficient was used in this analysis to identify the goodness of clusters.    Fig 6: K-Means Cluster Analysis: Silhouette Method, identify the optimal cluster    Fig 7: Visualization of Clusters silhouette plot |
| Epigenomic Roadmap: | H3K27me3 is required in silent gene expression during the histone modifications.  k=4, cluster 2 Analysis of genes associated with epigenomics are as follows:   1. H3K27me3 Stomach Mucosa 2. H3K27me Pancreatic Islets 3. H3K27me Stomach Smooth Muscle 4. H3K27me Penis Foreskin Fibroblast Primary Cells 5. H3K27me Rectal Mucosa     Fig 8: Epigenomic Road map of k= 4, cluster 2(log2FC 4.1112117, pvalue 0.00989127) genes  k=4, cluster4, the analysis of genes associated with epigenomics are as follows:   1. H3K27me3 CD4+ CD25+ CD127- Treg Primary Cells 2. H3K27me Neurosphere Cultured Cells Cortex Derived 3. H3K4me1 Penis Foreskin Fibroblast Primary Cells 4. H3K27me iPS DF 19.11 5. H3K27me Stomach Smooth Muscle     Fig 9: Epigenomic Road map of k= 4, cluster 4 (log2FC -3.3163967, pvalue 0.01286101) |
| Transcription Factor Protein to Protein Interaction (PPI) | ***k=4, cluster 2***    ***k=4, cluster 4***  No significant transcription factor protein to protein interaction (PPI). |
| Kinase: | Kinase perturbation extracted from GEO database are as follows:  ***k=4, cluster 2***  *(Up) kinase perturbations:*  IGF1R drug inhibition 47 GSE14024  *(Down) kinase perturbations:*  CDK8 knockdown 163 GSE38061  ***k=4, cluster 4***  *(Up) kinase perturbations:*  FGFR3 drug inhibition 36 GDS5023  TGFBR2 knockout 296 GSE22989  TGFBR1 active mutant 291 GSE14523  CDK8 knockdown 161 GSE38061 |
| Molecular Function: | The molecular function from GO.  ***k=4, cluster 2***    ***k=4, cluster 4*** |
| Cells Types: | Significant result from Cancer cell line Encyclopedia are as follows:  ***k=4, cluster 2***   1. BICR56 UPPER AERODIGESTIVE TRACT 2. SCC15 UPPER AERODIGESTIVE TRACT 3. BHY UPPER AERODIGESTIVE TRACT 4. NCIH1666 LUNG 5. BICR22 UPPER AERODIGESTIVE TRACT   ***k=4, cluster 4***   1. KMBC2 UNINARY TRACT 2. TE9 OESOPHAGUS 3. CAL33 UPPER AERODIGESTIVE TRACT 4. BICR6 UPPER AERODIGESTIVE TRACT 5. RT11284 URINARY TRACT |
| Pathway: | Top 5 partway from KEGG pathway database are as follows:  ***k=4, cluster 2***   1. Cytokine-cytokine receptor interaction 2. Rheumatoid arthritis 3. IL-17 signaling pathway 4. Legionellosis 5. Pertussis   ***k=4, cluster 4***   1. Retinol metabolism 2. Metabolism of xenobiotics by cytochrome P450 3. Ascorbate and aldarate metabolism 4. Viral protein interaction with cytokine and cytokine receptor 5. Drug metabolism |
| PhenGen Association: | ***k=4, cluster 2***    ***k=4, cluster 4*** |
| Drug Interaction: | The drug interaction varies depending on the individual. |
| Drug Mechanisms: | INHIBITOR |
| Summary: | In summary, the clusters are of not similar size. The significant genes/molecular mechanisms were negative and positive fold changes. They are cluster 2 (Log2FC = 4.1112117, pvalue = 0.00989127) and cluster 4 (Log2FC = -3.3163967, pvalue = 0.01286101). The molecular mechanisms in cluster 2 were high in interaction with the drug Itraconazole while cluster 4 molecular mechanisms were expressed negatively. |

***References:***

*Compound Report Card*. (n.d.). Www.ebi.ac.uk. Retrieved February 26, 2022, from <https://www.ebi.ac.uk/chembl/compound_report_card/CHEMBL64391/>

Edward Y. Chen. (n.d.). *Enrichr*. Maayanlab.cloud. https://maayanlab.cloud/Enrichr/

‌

Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2017 Nov 8. doi: 10.1093/nar/gkx1037.

Winn, J., Bishop, C. M., Diethe, T., Guiver, J., & Zaykov, Y. (n.d.). *Model-Based Machine Learning (Early Access): an online book*. Mbmlbook.com. Retrieved February 26, 2022, from https://mbmlbook.com/

‌