Vibhav Jha

BE 700 HW3

Clustering Gene Expression

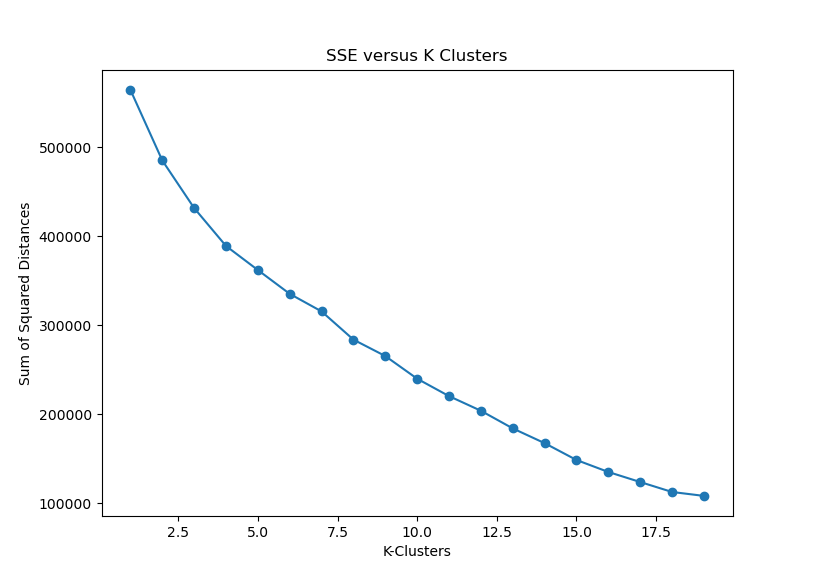
**Part 1.** Observations from all data using a heatmap

A heatmap for all human gene expression data was generated using the plotly library and Python 3.8. The data was scaled using sci kit learn’s min max scaler. **Figure 1** shows the full heatmap, however it is best viewed in its HTML format in a web browser. A copy named Figure1\_Fullheatmap.html is attached to the submission email. **Figure 1** is on a separate page attached to the end so as not to break up this document.

From **figure 1**, two things stand out. Certain subjects seem to be nearly opposite of the compared to the rest of the population in every gene (GSM624938, GSM624946). After scaling all the gene expression values, none of the genes especially standout as either all high or all low expression.

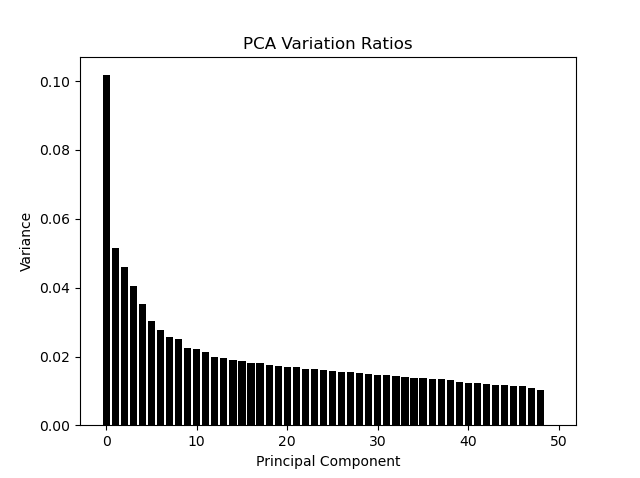
**Part 2.** K-means clustering and PCA

PCA was used to reduce dimensionality of the data set and done using sci kit learn and python 3.8. **Figure 2** shows the variances described by each principal component. The first principal component only was able to describe about 10% of the dataset. From the PCA, only the first two principal components were used to reduce the size of the dataset. This is not the ideal method; it is better to use a higher number of principal components in order to reduce the risk of underfitting. For this report, only the first two were chosen in order to simplify visualizing the results of the k-means clustering.

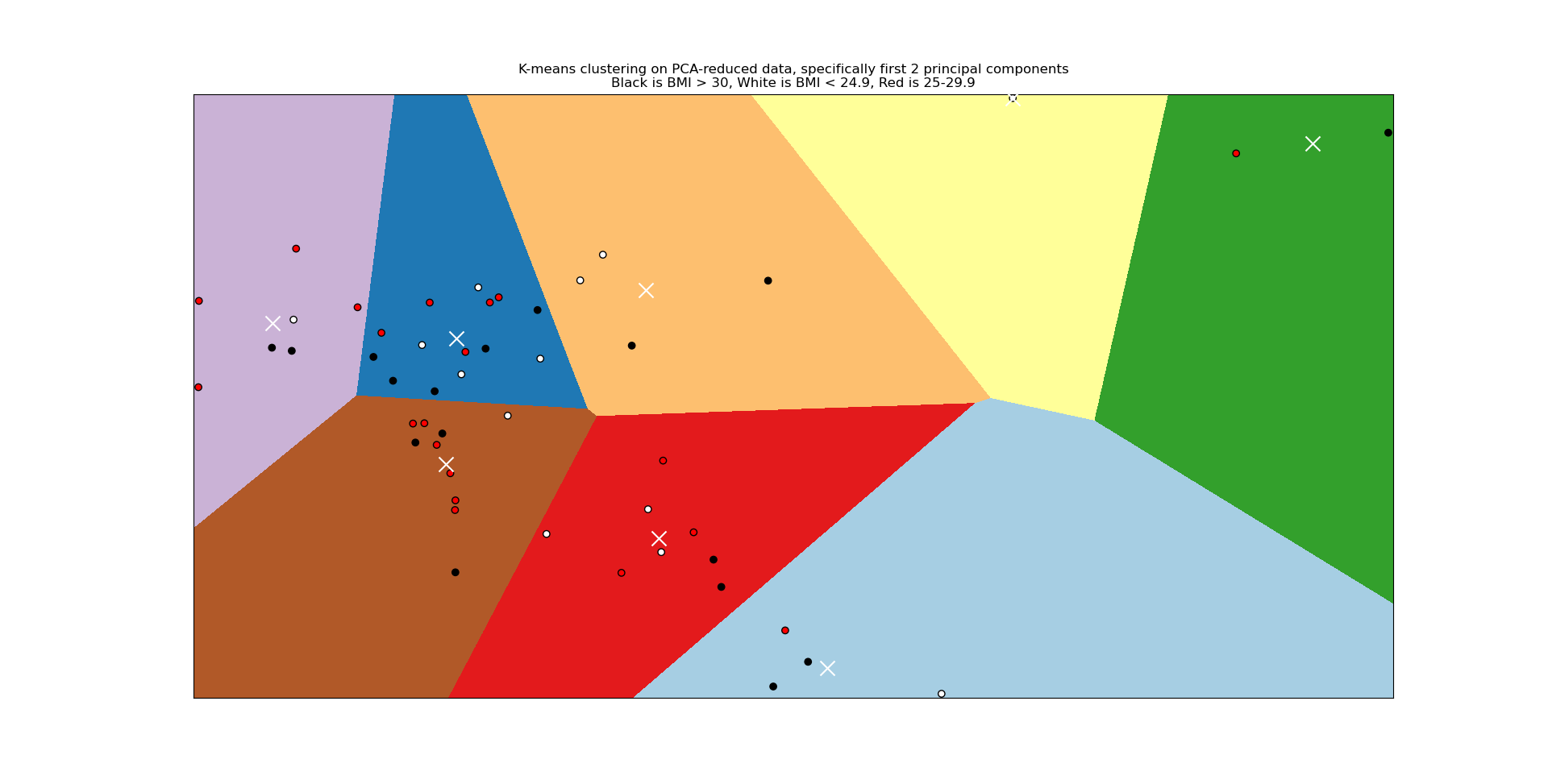


**Figure 2.** Sum of Squared Distances (or inertia) computed for K clusters up to 20. There is a slight dip at the k = 8 cluster, indicating that it may be the best amount of clusters for this data set.

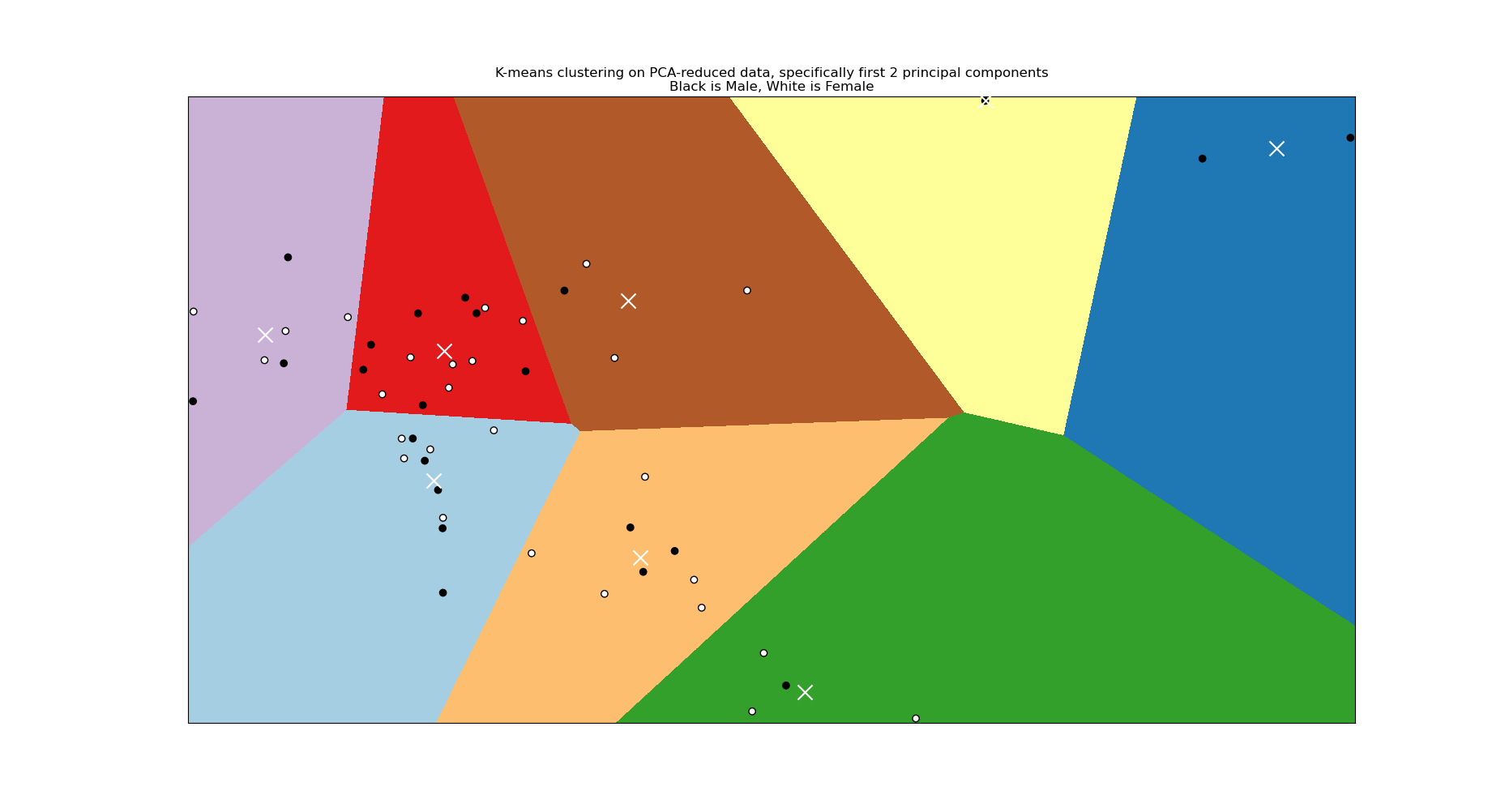
To choose the best number of centroids/clusters (or k) for k means, simple elbow analysis of the sum of squared distances of samples to the closest cluster center versus various cluster sizes was utilized. **Figure 3** shows a slight elbow at 8 clusters, though this is only slightly irregular compared to the neighboring values. **Figure 4** is the result of the 8-means clustering with BMI groups highlighted. BMI was separated into three categories, greater than 30 (obese), between 25 and 29.9 (overweight), and less than 24.9 (healthy weight until 18.5). There seemed to be no obvious clustering of BMIs, rather more overlapping in general. A similar conclusion can be found when overlaying genders for each subject, shown in **figure 5.** Overall clusters were not enriched with either gender.



**Figure 3.** PCA Variation Ratios for the entire set. There could only be a maximum of 49 principal components as it is limited by minimum(number of features, number of samples) – 1 within sci kit learn.



**Figure 4.** K-means clustering on PCA-reduced data, specifically the first 2 principal components. White crosses indicate the centroid centers, the background indicate decision boundaries. Also, BMI data is overlayed on the points with black pertaining to BMI > 30 being black, BMI < 24.9 being white and BMI between 25-29.9 being red.



**Figure 5.** K-means clustering on PCA-reduced data, specifically the first 2 principal components. White crosses indicate the centroid centers, the background indicate decision boundaries. Here, gender is overlayed with Black pertaining to male and white to female.

**Part 3.** DAVID Analysis

Considering the clustering done above shows clusters of subjects rather than cluster of genes, in order to do DAVID analysis 3 batches of randomly selected genes of random sizes was used (greater than 50, less than 500). A MATLAB script was used to generate the random sets (RandomGeneGroupSelections.m). Each set was saved as a text file and is attached to the submission email. In the DAVID tool the default gene ontology selections and the KEGG pathway selections were used.

**Using purely random gene selections should lead to non-significant results or high false discovery rates (FDR).**

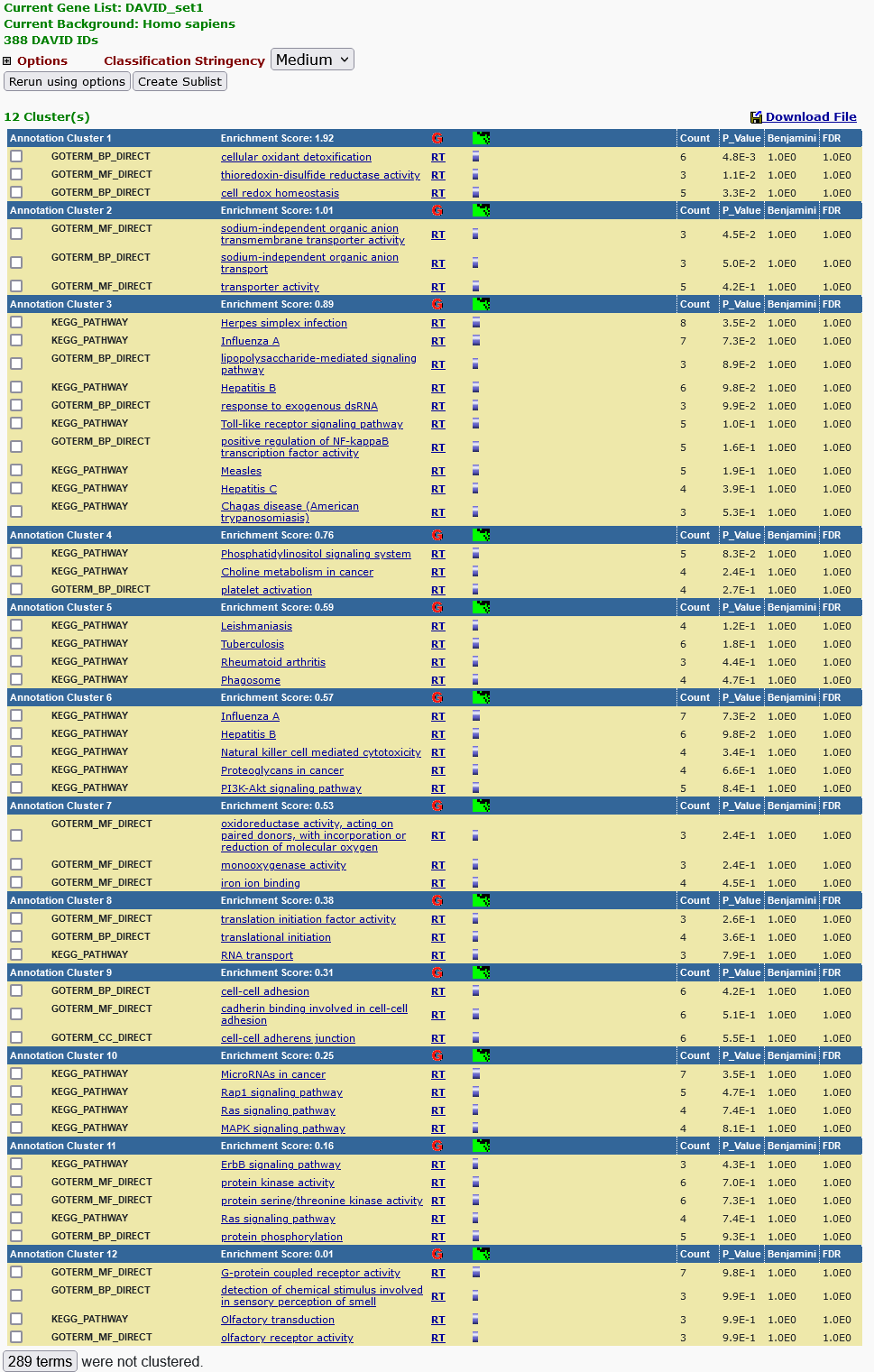
Set 1 of 417 genes provided 12 clusters with only one p-value of less than 0.05, however both the FDR and benjamini corrected p value were 1 for every value. This indicates that all features selected here are null.

Set 2 of 144 genes provided 4 clusters with two 3 p-values of less than 0.05. In cluster 1 and the KEGG\_pathway carbon metabolism and biosynthesis of antibiotics both had p-values less than 0.05 and FDRs or 0.88. This indicates that 88% of the significant features selected here are null. For the rest of the clusters the FDRs were 1, indicating no significance.

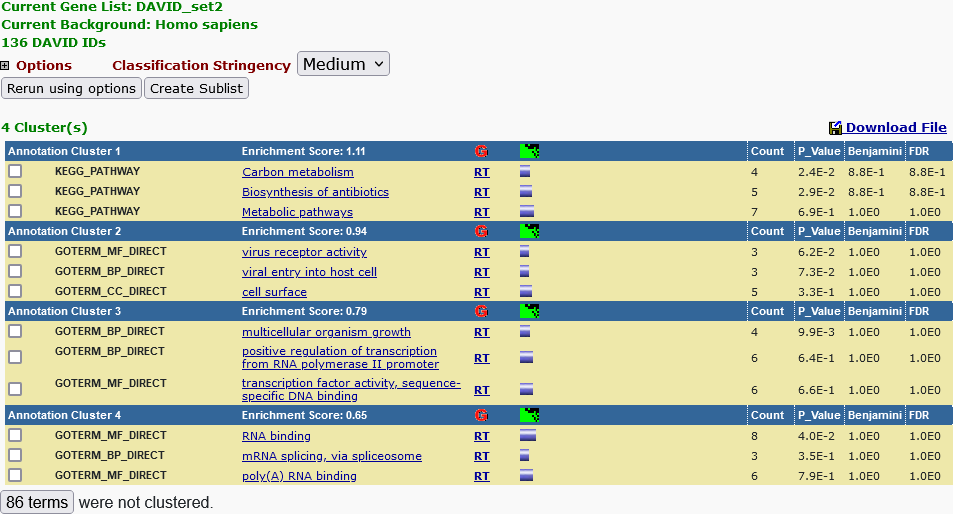
Set 3 of 183 genes provided 4 clusters. Cluster one and two, while having one p value of less than 0.05(GOTERM\_BP\_DIRECT, protein dephosphorylation), had an FDR of 1, thus indicating no significance. Cluster 2 however contained all KEGG\_pathway values where 3 of the four had p-values of less than 0.05. The lowest FDR was 0.65 and corresponded to the T cell receptor signaling pathway. While still a majority of the significance (65%) is null, intuitively there should be some correlation overall between the inhibition of T-cell pathways and type 1 diabetes. It was found in 2017 by Ge et al, that UBASH3A mediates risk for type 1 diabetes. The other two with FDR values less than 1 were HTLV-I infection (FDR = 0.79) and B cell receptor signaling pathway (FDR = 0.84).

Screenshots of the output table for each set’s DAVID analysis is below.

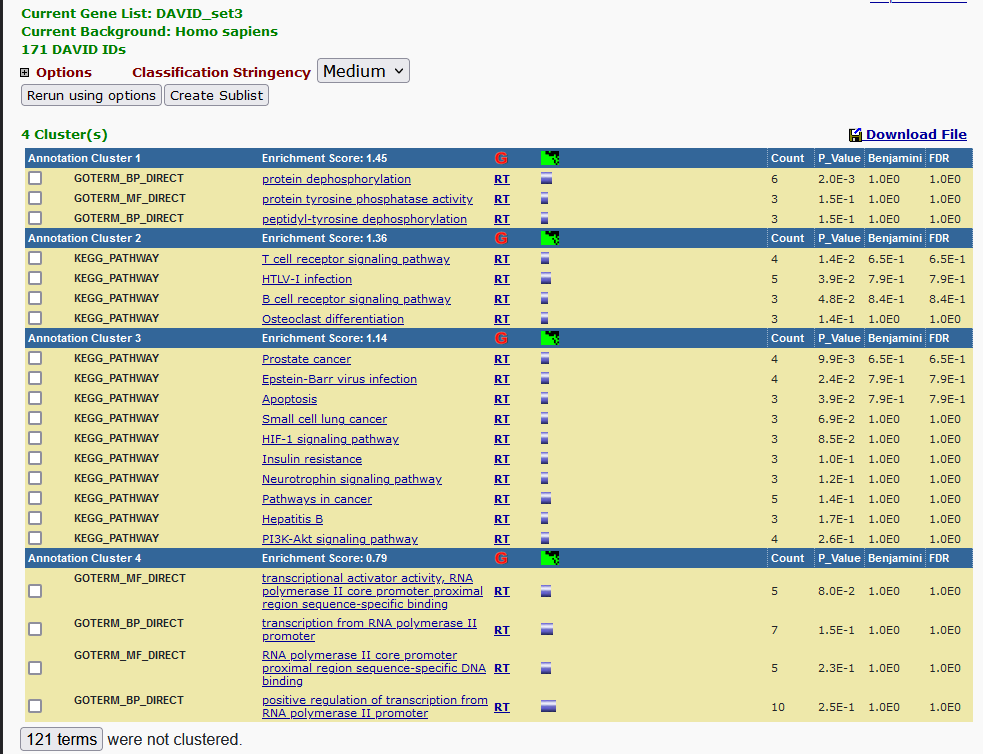
Ge Y, Paisie TK, Newman JRB, McIntyre LM, Concannon P. UBASH3A Mediates Risk for Type 1 Diabetes Through Inhibition of T-Cell Receptor-Induced NF-κB Signaling. Diabetes. 2017 Jul;66(7):2033-2043. doi: 10.2337/db16-1023. Epub 2017 Jun 12. PMID: 28607106; PMCID: PMC5482087.

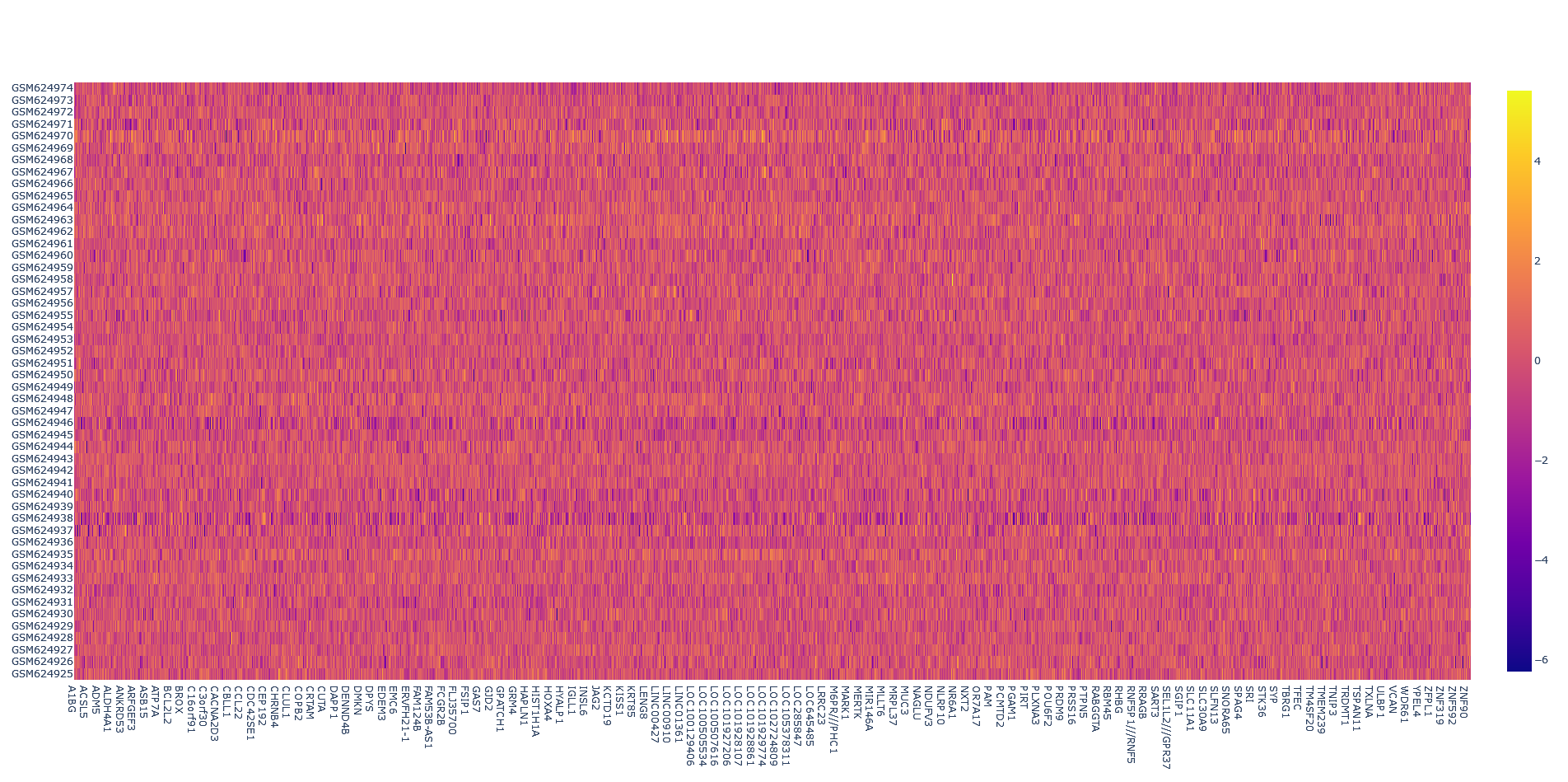
**Table 1.** DAVID analysis output on randomly selected gene set 1. the default gene ontology selections and the KEGG pathway selections were used. 

**Table 2.** DAVID analysis output on randomly selected gene set 2. the default gene ontology selections and the KEGG pathway selections were used.



**Table 3.** DAVID analysis output on randomly selected gene set 3. the default gene ontology selections and the KEGG pathway selections were used.





**Figure 1.** Heatmap of all gene expression data.

**Part 4**. Supplements

For Part 1.

1. Code can be found in HW3.py
2. Figure1\_Fullheatmap.html is attached to the email

For Part 2.

1. Code can be found in HW3.py

For Part 3.

1. DAVID sets copied below or in DAVID\_setX.txt (3 files)
2. Code can be found in RandomGeneGroupSelections.m

|  |  |  |
| --- | --- | --- |
| **SET 1** | **SET 2** | **SET 3** |
| |  | | --- | | TOPBP1 | | CDC25B | | TRIM31 | | NSL1 | | C7orf50 | | FOXR1 | | MAGEB2 | | WNT2 | | ZBTB21 | | CNTNAP3P2 | | ZFP37 | | WNK3 | | LOC101928847 | | SLAIN2 | | CHD8 | | LINC00629 | | TRIP10 | | SGCZ | | XCL2///XCL1 | | PALM | | ANTXR2 | | STAM-AS1 | | UBQLN2 | | PIGG | | RNF31///IRF9 | | RCHY1 | | KLC1 | | PAK4 | | CRYBB3 | | PRAMEF2///PRAMEF1 | | ANKRD11 | | FOXD4L1///FOXD4 | | ARID4A | | C7 | | SMTNL2 | | POLR1B | | GSTK1 | | VSTM4 | | ANKRD9 | | LIPC-AS1 | | KCNK7 | | RPS4X | | SHANK2-AS3 | | DAND5 | | LOC101929269 | | LOC100131170 | | OR51B2 | | PRKCA | | RNF113A | | FOLR2 | | PIK3C2A | | PAICS | | COPZ2 | | CCL16 | | LOC101930655///C7orf73///SLC13A4 | | XIRP1 | | HOXA10 | | MRAP | | EFS | | RIDA | | FAM47A | | LOC105370832 | | PPA2 | | TMEM155 | | XAGE2 | | MAGED1 | | CFAP44 | | CLDN19 | | FAM86DP///FAM86B2///FAM86FP///FAM86B1///FAM86C1 | | SPRR1B | | FAM27B///FAM27C///FAM182B///FAM182A | | SLC5A11 | | EZH2 | | TXNL1 | | IBA57-AS1 | | DERA | | FAM198A | | NEURL3 | | LOC101927660 | | IDS | | SNX5 | | MR1 | | MAP1A | | TRPA1 | | GABRP | | RNF214 | | RNASE4 | | KCNJ14 | | MIA3 | | BRD8 | | ATAD3B | | LOC728805 | | SCN5A | | UBQLN3 | | CDK11A///CDK11B | | MIEN1 | | LOC101927292 | | ADAM30 | | HMGCS1 | | COPS3 | | SH3GL1P2 | | GPR85 | | LOC646484 | | CPEB2-AS1 | | MYOM2 | | FBXO5 | | PACRG-AS2 | | PLPP6 | | RGPD6///RGPD8///RGPD3///RGPD4///RGPD5 | | LOC100294033///TCAF1 | | C14orf93 | | ELMSAN1 | | TRIM3 | | CLPS | | SNORA17B///SNORA17A///SNHG7 | | LRRC56 | | ZNF844 | | BTF3 | | LOC100129098 | | CAPNS2 | | YIF1A | | ABHD8 | | SAR1B | | SLCO2A1 | | TBL1Y | | C15orf65 | | KRT73 | | FAT4 | | SKP2 | | LINC01192 | | TRAPPC10 | | CYB5R4 | | FCF1 | | CHST6 | | CEP152 | | TBX2 | | MLYCD | | MAP1LC3B | | CHRNE | | STRA6 | | NLRP13 | | ID3 | | LOC149373 | | KRTAP5-AS1 | | BRE | | ESX1 | | CD2 | | CYP2J2 | | ESYT1 | | LINC00421 | | ARRDC4 | | TNFSF18 | | VANGL1 | | LOC101929454 | | LOC101929224 | | HMOX1 | | TMPRSS5 | | IQUB | | CBS | | SCUBE2 | | KIF23 | | EXD3 | | LAMA5-AS1 | | C6orf141 | | CDR2L | | USP32///USP6 | | WFDC9 | | MIR664B///SNORA56///DKC1 | | ATP8A1 | | EPHB6 | | IFNA17 | | SMG5 | | ADGRE3 | | ARHGAP11A | | CRH | | ORMDL3 | | RAB4B-EGLN2///EGLN2 | | OR7E12P | | LOC100422212 | | MAGEB6 | | GID8 | | REG3A | | DCAF4L1 | | PLEKHA4 | | CYP2B7P | | IQCF4 | | NOSIP | | SCUBE1 | | C11orf44 | | TXNRD1 | | SAV1 | | LOC101928973 | | LINC01500 | | LOC100132356 | | GPAA1 | | LOC105374325 | | LOC105377776 | | SLCO6A1 | | SH3RF3 | | OR1A2 | | KCNA6 | | SLC38A7 | | LOC91450 | | ID1 | | UPK3A | | TEX26-AS1 | | MAP2K2 | | NLRP8 | | MRPL17 | | DNAJC3 | | GM2A | | LOC101927446///GTPBP10 | | EMILIN1 | | SRRM2 | | DEFB108B | | EIF2B2 | | CRY2 | | EIF5B | | LINC01490 | | GPR82 | | TTC36 | | LINC01123///LINC01106 | | CYP4F3 | | TOLLIP-AS1 | | ZNF32-AS3 | | LIPF | | CBR4 | | FAM98B | | LEMD2 | | MTMR6 | | FBXO28 | | MZT1 | | PRND | | EDN3 | | CCDC8 | | GIMAP5 | | GTF2IRD1 | | LINC00845 | | LOC105373418 | | C17orf51 | | FBXO34 | | SLC12A4 | | AMMECR1L | | TXNDC11 | | RAB29 | | LOC101929154 | | MKRN3 | | ERI1 | | LOC100507403 | | YWHAE | | MAGEB1 | | LOC339468 | | ENDOV | | LOC101929174 | | NOC2L | | PIGR | | KLK15 | | IPCEF1 | | ZNF585B | | AP4E1 | | TLR2 | | TRIM29 | | SHISA7 | | C8orf49 | | FBXO18 | | HLA-DOB | | PIK3C2A | | CEP41 | | PTCSC1 | | CAPRIN2 | | PABPC1L2B///PABPC1L2A | | LOC101929787 | | SCN3B | | PRRC2B | | TNNT2 | | TMEM156 | | HIST1H4G///HIST1H4F | | PP12708 | | DGKB | | ANGPTL2 | | REC8 | | LOC102723831 | | LOC101928401 | | TOB2 | | NCOA4 | | NFKBIA | | SYNDIG1L | | SLC25A21-AS1 | | MIR6878///TARS2 | | CYP20A1 | | ESYT1 | | TMA7 | | AMDHD1 | | LOC101929280 | | CRAT | | ZNF280C | | PROX2 | | LOC102724009 | | LOC101927460 | | ATP6V1H | | PITRM1 | | ARFGAP2 | | BIRC3 | | LOC339978 | | C6orf226 | | SLIT2 | | SLCO5A1 | | PTGS1 | | CLDND2 | | PAX7 | | LOC284939///MAPK8IP2 | | ZKSCAN8 | | ORM2///ORM1 | | SLAMF6 | | LOC100506114 | | LINC01233 | | SNHG24 | | C14orf178 | | CEBPB | | CSNK2B | | KIR2DL2 | | SOCS4 | | SLC1A7 | | ATRNL1 | | KRT36 | | LOC642426 | | LINC00348 | | PAPOLB | | NPM1 | | GCHFR | | LINC01204 | | ADGRF1 | | ZNF485 | | CPT1C | | CAPN14 | | ITGAM | | DGKQ | | LOC101929268 | | HNRNPUL2-BSCL2///HNRNPUL2 | | WBP11 | | TSPAN17 | | ASMTL-AS1 | | RASGRP3 | | FHL5 | | LINC00668 | | MAGOH2P | | USP46 | | LINC00459 | | ZNF442 | | GMEB1 | | PPM1F | | PDCD6IP | | LRRC8A | | POU3F3 | | PDDC1 | | CTTNBP2 | | CDC7 | | ZSWIM3 | | CRYBB2P1///CRYBB2 | | ANKRD22 | | MEFV | | TICAM1 | | PDRG1 | | DCTN1 | | IQCK | | LOC100509814 | | ZNF396 | | CNP | | SUGP1 | | OR1J2 | | JPH4///AP1G2 | | DCUN1D3 | | LINC01007 | | LOC101928583 | | CCNG2 | | MRPS22 | | EIF2S2 | | KDM5C | | MORC4 | | FAM207A | | GATC///SRSF9 | | NFE2L2 | | FDX1L | | SNAPC2 | | ZNF431 | | RAB27B | | HRG | | MPHOSPH8 | | CARMIL2 | | TOX2 | | THAP7-AS1 | | SLFN11 | | FBRSL1 | | MTHFSD | | AIRE | | LINC00889 | | GRAMD4 | | COMMD6 | | CUL5 | | LINC00671 | | C3orf70 | | MYCNUT | | LOC101927446///GTPBP10 | | POLR3G | | PPFIA3 | | NXNL1 | | ANKRD37 | | BCORL1 | | GTSE1-AS1 | | LOC728868 | | PADI1 | | LDHA | | SMARCAL1 | | PSMB2 | | ZDHHC18 | | LOC729324 | | HBQ1 | | |  | | --- | | LOC101928551///ADAM29 | | LINC00574 | | SYN1 | | CRYBB3 | | HNRNPD | | FKBP8 | | PMP22 | | ECEL1 | | SLC39A2 | | ACTR8 | | MEG3 | | SLC4A5///DCTN1 | | RELT | | DCAF6 | | ACAT1 | | S100A2 | | WNT7B | | PPP1R13B | | GJA4 | | RRS1-AS1 | | PLIN1 | | KIAA1210 | | NELL2 | | LINC00965 | | MPND | | PPP1R21 | | CC2D2B | | TTC32 | | ZNF611 | | PHYHIP | | SHANK2-AS2 | | AMIGO2 | | CTSK | | TERT | | RELB | | LOC105373764 | | CDCA7L | | LOC101928751 | | LOC257396 | | GPR87 | | ARHGEF5 | | LINC01583 | | C5orf63 | | ARHGEF26-AS1 | | PLAG1 | | STAMBP | | CD55 | | CDH24 | | LOC101927648 | | DGCR6L | | SDHD | | NBAT1 | | LOC100129884 | | ZNF415 | | PRKAG1 | | DYRK1B | | CR1 | | LRRC70///IPO11 | | GRHL2 | | TRDMT1 | | C17orf99 | | GRIK2 | | DZIP1 | | RARA | | SNRNP70 | | REG1A | | CTSD | | SYAP1 | | ACACA | | IGLJ3///CKAP2///IGLV@///IGLC1 | | WDR35 | | USP3 | | AK9 | | PGAM1 | | COX20 | | MICAL1 | | EXOC3L4 | | COA3 | | LOC102725116 | | MIR3150B///MIR3150A///LOC286149 | | LOC101927346 | | LOC105373860 | | HAPLN2 | | TMA7 | | FAM208A | | TAB3 | | NUDT11 | | POP5 | | ANGPT4 | | PAQR7 | | ZSWIM8 | | LOC105371622 | | UBE2Q2 | | LMTK2 | | DIABLO | | SLC22A18 | | SLC27A1 | | ATP5G2P1 | | SF1 | | LOC100289283 | | STK16 | | IL36G | | DLAT | | TCP11L2 | | MZB1 | | SON | | AMELY | | PPP3CB-AS1 | | RPL39 | | KLKB1 | | DDX50 | | CRELD2 | | TMEM130 | | RAB39B | | TMCO1 | | FGF12 | | PCED1B-AS1 | | LINC01420 | | LOC158434 | | MFSD6 | | ANKS1B | | TCTEX1D1 | | IBTK | | LRRC73 | | ACOT4 | | ARPC4 | | UBE4B | | DPP4 | | PTPRJ | | KRIT1 | | RAP2B///RAP2A | | OTUD6B-AS1 | | TTN-AS1 | | NEDD4L | | AP1S2 | | CCDC25 | | RNF180 | | LOC100505622 | | NAE1 | | TAF7L | | LOC102723983///DAZAP2 | | SLC9A8 | | OS9 | | LOC100507387///FAM153A///FAM153B | | |  | | --- | | SLC30A3 | | ICAM2 | | SET | | MEGF11 | | FAM229A | | CAMK2N2 | | ATP6 | | PGBD1 | | POU2AF1 | | RABL2A///RABL2B | | SERPINB10 | | BRINP2 | | GSX1 | | RBM41 | | HELLS | | LOC101927359 | | CXXC1 | | HAND2 | | LOC100506371 | | LOC727820 | | GDF10 | | FOXE1 | | PICK1 | | EPB41 | | PHTF2 | | LINC01231 | | KIAA1468 | | HLA-DPA2 | | LOC149373 | | ZNF581 | | LOC728903///GLIDR | | PIK3R5 | | KIR2DS1 | | CCDC57 | | PRDX3 | | LINC00102 | | LOC100506563 | | HIPK1-AS1 | | C3orf22 | | LOC100507140 | | SLC25A27 | | DDX52 | | LOC101060835///LOC100996809///HLA-DRB5///HLA-DRB4///HLA-DRB1///HLA-DQB1 | | DCUN1D3 | | BRD9 | | ANKRD13C | | CLPTM1L | | FAM83D | | BEAN1 | | TUBGCP2 | | MAGEA6 | | ZC4H2 | | TMEM55B | | KIR2DL4 | | KCNAB3 | | OVGP1 | | SUMF2 | | LOC100129449 | | LOC90834///BRD1 | | COQ6 | | MB | | CDC16 | | VCPIP1 | | ZNF430 | | GMEB1 | | RTKN2 | | GPSM1 | | ZNF267 | | FSTL1 | | PGLS | | LINC01279 | | LAMTOR4 | | MMP27 | | MSI1 | | ATP5J2 | | LINC00595 | | POLN | | AHSA2 | | POLE | | FASTK | | LOC284950 | | MAGI1 | | IMPACT | | KTI12///TXNDC12 | | BBS5 | | ELOVL2 | | AAMP | | DEFB123 | | CCDC151 | | LINC00957 | | RPAP3 | | PSMC2 | | NFKBIA | | FEM1A | | PHACTR4 | | DCAKD | | PTPN9 | | NFIL3 | | NAP1L5 | | PMFBP1 | | LSG1 | | MPND | | NABP2 | | GPATCH2L | | SYNE2 | | BCAS3 | | SSH2 | | MGARP | | ELF4 | | TMEM14B///TMEM14C | | MKRN7P | | BCL2 | | SNORD8 | | TMEM52B | | LINC00560 | | MAPK1IP1L | | NLE1 | | ZDHHC13 | | PCYOX1L | | LOC100132686 | | EGOT | | TLX2 | | FBXO22 | | C1RL-AS1 | | PGLYRP1 | | ARF6 | | HIGD1A | | LINC01364 | | TMEM110 | | LINC01142 | | PTPN7 | | LINC00514 | | LOC100506123///LOC100506076 | | RC3H2 | | GALNT8///KCNA6 | | MIR4680///PDCD4 | | OR5P3 | | PLA2G4B///JMJD7-PLA2G4B | | C11orf1 | | LOC105371038 | | ITK | | LINC01204 | | KLF4 | | DCTPP1 | | SYNPR-AS1 | | LOC100505978 | | SLC25A32 | | MTOR | | IDI2-AS1 | | PPP3R1 | | STARD7-AS1 | | CCSER1 | | RPL39 | | LRRC8E | | MAGOH2P | | TGM3 | | LOC100630923 | | COMMD4 | | LOC101928907 | | C2orf69 | | LOC101928435 | | PLA2G4D | | TMEM62 | | HDLBP | | FLT3LG | | CEP131 | | ZMYM2 | | PPP2CB | | RTCB | | HIPK1-AS1 | | DNASE1L1 | | CDH12 | | MTA2 | | VPS53 | | DAZAP2 | | WFDC5 | | LOC105372695 | | BIN3-IT1 | | AP3M2 | | SUMF2 | | L3HYPDH | | UBFD1 | | MEAF6 | |