Using *fanyi* to assist research communities in retrieving and interpreting information

Guangchuang Yu[1,*]

[1] Department of Bioinformatics, School of Basic Medical Sciences, Southern Medical University, Guangzhou, 510515, China

[*] Correspondence: Guangchuang Yu <gcyu1@smu.edu.cn>

Abstract

Manually retrieving gene information is tedious and professional vocabulary hinders researchers from quickly understanding the information. To solve these issues, we implemented the fanyi package to help query gene information from NCBI, and with the help of several online translation services powered by AI, an accurate translated version can be obtained to help users quickly capture the information. Although the examples presented here are for biomedical research, the translation utilities can be applied to any other domain.

Introduction

English, as the first language of scientific research, plays an important role in disseminating scientific research results. However, for a large number of graduate students from non-English speaking countries, there are still more or less obstacles to using English. This is also why many local media and social platforms use their language to interpret scientific research results, which is popular. With the development of artificial intelligence, text translation has become quite accurate. A large number of students use translation software to convert English materials into their language when reading scientific research papers and also use translation software to translate their domestic language into English when writing papers.

In biomedical research, a common scenario is for researchers to search for one term or gene after another, obtain explanations for the term or gene, and then use translation tools to translate this information into the local language, to quickly capture relevant information for interpretation. This is usually cumbersome and requires a lot of time. For example, when conducting single-cell analysis, the interpretation of marker genes in different cell clusters often requires the aforementioned method. To interpret these genes, as mentioned above, repeating operations for each gene requires a significant amount of time investment. Many researchers choose not to do this and instead directly use the genes for functional enrichment analysis. Only the results of enrichment analysis were used to selectively search for certain genes. Although it saves time, it is also very likely to miss some important genes.

In response to this situation, we have developed the R package, fanyi, which means “translation”

in Chinese Pinyin. It can utilize online AI translation services to translate between multiple languages. At present, this package can support users to use Baidu, Bing, and Youdao translation services. With this package, users can translate the text using R language, which is widely used in multiple fields. For example, R is a necessary Swiss army knife for omics data analysis and visualization. With the help of fanyi, analysis results can be directly translated into the local language, and descriptions of functions, pathways, phenotypes, and diseases will become easier to understand.

In the field of biomedicine, retrieving gene information can be considered a daily task, such as searching for marker genes to better interpret cell types. To better serve researchers in the biomedical field with this translation tool, we have implemented automatic retrieval of gene descriptions. This allows for batch retrieval of analysis results, such as a list of differentially expressed genes, genes with high mutations in tumors, and marker genes for different cell types, etc. By automatically retrieving descriptions and summaries of genes from previous studies, it will greatly save the user time. Moreover, through translation, language barriers can be eliminated. Enable researchers to spend more time thinking about scientific questions.

Results

Showcase 1:

Here, we use the famous PBMC data set as an example. After identifying the marker genes of all the cell clusters, we converted the gene symbols to ENTREZ gene IDs using clusterProfiler (Wu *et al.*, 2021) and then queried gene information using the gene_summary function provided in the fanyi package.

```
library(Seurat)
pbmc.markers <- FindAllMarkers(pbmc, only.pos = TRUE)

library(dplyr)
top2 <- pbmc.markers |>
    group_by(cluster) |>
    filter(avg_log2FC > 1) |>
    slice_head(n = 2) |>
    ungroup()

g <- clusterProfiler::bitr(top2$gene,
            fromType = 'SYMBOL',
            toType = 'ENTREZID',
            OrgDb = 'org.Hs.eg.db')

library(fanyi)
gs <- gene_summary(g$ENTREZ)
```

As demonstrated in Figure 1, the information output by FindAllMarkers can be integrated with the output of gene_summary, and all these pieces of information can be exported as a table to help the user understand the marker genes.

gene	avg_log2FC	p_val_adj	description	summary
CD8 T				
CCL5	3.329135	7.696099e-190	C-C motif chemokine ligand 5	This gene is one of several chemokine genes clustered on the q-arm of chromosome 17. Chemokines form a superfamily of secreted proteins involved in immunoregulatory and inflammatory processes. The superfamily is divided into four subfamilies based on the arrangement of the N-terminal cysteine residues of the mature peptide. This chemokine, a member of the CC subfamily, functions as a chemoattractant for blood monocytes, memory T helper cells and eosinophils. It causes the release of histamine from basophils and activates eosinophils. This cytokine is one of the major HIV-suppressive factors produced by CD8+ cells. It functions as one of the natural ligands for the chemokine receptor chemokine (C-C motif) receptor 5 (CCR5), and it suppresses in vitro replication of the R5 strains of HIV-1, which use CCR5 as a coreceptor. Alternative splicing results in multiple transcript variants that encode different isoforms. [provided by RefSeq, Jul 2013]
NKG7	2.459665	4.334551e-177	natural killer cell granule protein 7	Predicted to be integral component of plasma membrane. Predicted to be active in plasma membrane. [provided by Alliance of Genome Resources, Apr 2022]
Naive CD4 T				
CCR7	2.397366	1.312702e-83	C-C motif chemokine receptor 7	The protein encoded by this gene is a member of the G protein-coupled receptor family. This receptor was identified as a gene induced by the Epstein-Barr virus (EBV), and is thought to be a mediator of EBV effects on B lymphocytes. This receptor is expressed in various lymphoid tissues and activates B and T lymphocytes. It has been shown to control the migration of memory T cells to inflamed tissues, as well as stimulate dendritic cell maturation. The chemokine (C-C motif) ligand 19 (CCL19/ELC) has been reported to be a specific ligand of this receptor. Signals mediated by this receptor regulate T cell homeostasis in lymph nodes, and may also function in the activation and polarization of T cells, and in chronic inflammation pathogenesis. Alternative splicing of this gene results in multiple transcript variants. [provided by RefSeq, Sep 2014]
LDHB	1.206019	5.137444e-108	lactate dehydrogenase B	This gene encodes the B subunit of lactate dehydrogenase enzyme, which catalyzes the interconversion of pyruvate and lactate with concomitant interconversion of NADH and NAD+ in a post-glycolysis process. Alternatively spliced transcript variants have been found for this gene. Recent studies have shown that a C-terminally extended isoform is produced by use of an alternative in-frame translation termination codon via a stop codon readthrough mechanism, and that this isoform is localized in the peroxisomes. Mutations in this gene are associated with lactate dehydrogenase B deficiency. Pseudogenes have been identified on chromosomes X, 5 and 13. [provided by RefSeq, Feb 2016]
B				
CD79A	6.911221	0.000000e+00	CD79a molecule	The B lymphocyte antigen receptor is a multimeric complex that includes the antigen-specific component, surface immunoglobulin (Ig). Surface Ig non-covalently associates with two other proteins, Ig-alpha and Ig-beta, which are necessary for expression and function of the B-cell antigen receptor. This gene encodes the Ig-alpha protein of the B-cell antigen component. Alternatively spliced transcript variants encoding different isoforms have been described. [provided by RefSeq, Jul 2008]
MS4A1	5.710520	0.000000e+00	membrane spanning 4-domains A1	This gene encodes a member of the membrane-spanning 4A gene family. Members of this nascent protein family are characterized by common structural features and similar intron/exon splice boundaries and display unique expression patterns among hematopoietic cells and nonlymphoid tissues. This gene encodes a B lymphocyte surface molecule which plays a role in the development and differentiation of B-cells into plasma cells. This family member is localized to 11q12, among a cluster of family members. Alternative splicing of this gene results in two transcript variants which encode the same protein. [provided by RefSeq, Jul 2008]
FCGR3A+ Mono				
CDKN1C	5.428812	2.316582e-208	cyclin dependent kinase inhibitor 1C	This gene is imprinted, with preferential expression of the maternal allele. The encoded protein is a tight-binding, strong inhibitor of several G1 cyclin/Cdk complexes and a negative regulator of cell proliferation. Mutations in this gene are implicated in sporadic cancers and Beckwith-Wiedemann syndrome, suggesting that this gene is a tumor suppressor candidate. Three transcript variants encoding two different isoforms have been found for this gene. [provided by RefSeq, Oct 2010]
HES4	4.993200	2.079606e-203	hes family bHLH transcription factor 4	Predicted to enable DNA-binding transcription factor activity, RNA polymerase II-specific and RNA polymerase II cis-regulatory region sequence-specific DNA binding activity. Predicted to be involved in anterior/posterior pattern specification and regulation of transcription by RNA polymerase II. Predicted to be part of chromatin. Predicted to be active in nucleus. [provided by Alliance of Genome Resources, Apr 2022]
DC				
FCER1A	7.633833	2.030720e-216	Fc epsilon receptor Ia	The immunoglobulin epsilon receptor (IgE receptor) is the initiator of the allergic response. When two or more high-affinity IgE receptors are brought together by allergen-bound IgE molecules, mediators such as histamine that are responsible for allergy symptoms are released. This receptor is comprised of an alpha subunit, a beta subunit, and two gamma subunits. The protein encoded by this gene represents the alpha subunit. [provided by RefSeq, Aug 2011]
SERPINF1	8.026573	1.997116e-203	serpin family F member 1	This gene encodes a member of the serpin family that does not display the serine protease inhibitory activity shown by many of the other serpin proteins. The encoded protein is secreted and strongly inhibits angiogenesis. In addition, this protein is a neurotrophic factor involved in neuronal differentiation in retinoblastoma cells. Mutations in this gene were found in individuals with osteogenesis imperfecta, type VI. [provided by RefSeq, Aug 2016]
NK				
FGFBP2	5.404509	4.673698e-233	fibroblast growth factor binding protein 2	This gene encodes a member of the fibroblast growth factor binding protein family. The encoded protein is a serum protein that is selectively secreted by cytotoxic lymphocytes and may be involved in cytotoxic lymphocyte-mediated immunity. An increase in the amount of gene product may be associated with atopic asthma and mild extrinsic asthma.[provided by RefSeq Staff, Oct 2006]
GZMB	5.977437	1.090260e-264	granzyme B	This gene encodes a member of the granzyme subfamily of proteins, part of the peptidase S1 family of serine proteases. The encoded preproprotein is secreted by natural killer (NK) cells and cytotoxic T lymphocytes (CTLs) and proteolytically processed to generate the active protease, which induces target cell apoptosis. This protein also processes cytokines and degrades extracellular matrix proteins, and these roles are implicated in chronic inflammation and wound healing. Expression of this gene may be elevated in human patients with cardiac fibrosis. [provided by RefSeq, Sep 2016]
Platelet				
GP9	11.484216	0.000000e+00	glycoprotein IX platelet	This gene encodes a small membrane glycoprotein found on the surface of human platelets. It forms a 1-to-1 noncovalent complex with glycoprotein Ib, a platelet surface membrane glycoprotein complex that functions as a receptor for von Willebrand factor. The complete receptor complex includes noncovalent association of the alpha and beta subunits with the protein encoded by this gene and platelet glycoprotein V. Defects in this gene are a cause of Bernard-Soulier syndrome, also known as giant platelet disease. These patients have unusually large platelets and have a clinical bleeding tendency. [provided by RefSeq, Oct 2008]
Memory CD4 T				
IL32	1.307077	3.966555e-86	interleukin 32	This gene encodes a member of the cytokine family. The protein contains a tyrosine sulfation site, 3 potential N-myristoylation sites, multiple putative phosphorylation sites, and an RGD cell-attachment sequence. Expression of this protein is increased after the activation of T-cells by mitogens or the activation of NK cells by IL-2. This protein induces the production of TNFalpha from macrophage cells. Alternate transcriptional splice variants, encoding different isoforms, have been characterized. [provided by RefSeq, Jul 2008]
LTB	1.331267	1.453850e-82	lymphotoxin beta	Lymphotoxin beta is a type II membrane protein of the TNF family. It anchors lymphotoxin-alpha to the cell surface through heterotrimer formation. The predominant form on the lymphocyte surface is the lymphotoxin-alpha 1/beta 2 complex (e.g. 1 molecule alpha/2 molecules beta) and this complex is the primary ligand for the lymphotoxin-beta receptor. The minor complex is lymphotoxin-alpha 2/beta 1. LTB is an inducer of the inflammatory response system and involved in normal development of lymphoid tissue. Lymphotoxin-beta isoform b is unable to complex with lymphotoxin-alpha suggesting a function for lymphotoxin-beta which is independent of lymphotoxin-alpha. Alternative splicing results in multiple transcript variants encoding different isoforms. [provided by RefSeq, Jul 2008]
CD14+ Mono				
LGALS2	3.651790	0.000000e+00	galectin 2	The protein encoded by this gene is a soluble beta-galactoside binding lectin. The encoded protein is found as a homodimer and can bind to lymphotoxin-alpha. A single nucleotide polymorphism in an intron of this gene can alter the transcriptional level of the protein, with a resultant increased risk of myocardial infarction. [provided by RefSeq, Jul 2008]
S100A8	6.645056	0.000000e+00	S100 calcium binding protein A8	The protein encoded by this gene is a member of the S100 family of proteins containing 2 EF-hand calcium-binding motifs. S100 proteins are localized in the cytoplasm and/or nucleus of a wide range of cells, and involved in the regulation of a number of cellular processes such as cell cycle progression and differentiation. S100 genes include at least 12 members which are located as a cluster on chromosome 1q21. This protein may function in the inhibition of casein kinase and as a cytokine. Altered expression of this protein is associated with the disease cystic fibrosis. Multiple transcript variants encoding different isoforms have been found for this gene. [provided by RefSeq, Jan 2016]

Figure 1. Table of marker gene information.

More importantly, the gene description and summary information query by gene_summary can be used in data visualization to provide a more informative graph.

```
library(ggsc)
library(ggplot2)
library(aplot)

d <- gs[1:4, ]
d$name <- factor(d$name, levels = d$name)
p <- ggplot(d, aes(x=1, y=3)) +
    geom_text(aes(label=str_wrap(summary, 75)), hjust=0) +
    xlim(1,2) +
    facet_wrap(~name, ncol=1) +
    theme_noaxis()+
    xlab(NULL) + ylab(NULL)
```

```
p2 <- sc_violin(pbmc, features[1:4], ncol=1) +
     theme(legend.position='none',
          axis.text.x = element_text(angle=60, hjust=1)) +
     xlab(NULL) + ylab(NULL)

insert_right(p2, p, width=1.5)
```

The above source codes employ the ggsc package to visualize the expression distribution of the selected marker genes among different cell clusters as violin plots and display the gene summary side by side to help the user interpret the results as demonstrated in Figure 2A.

```
pb <- sc_dim(pbmc) +
   sc_dim_geom_label(geom = ggrepel::geom_text_repel, color = "black",
                     bg.color = "white") +
     theme(legend.position = "none")

pc <- sc_feature(pbmc, features='CD3E', density = TRUE) +
     labs(subtitle = gs$description,
          caption=str_wrap(gs$summary, 180)) +
     scale_color_viridis_c() +
     theme(legend.position = c(.95, .2))
```

We further use ggsc to visualize the pbmc dataset. Figure 2B illustrates the cell clustering results and Figure 2C visualizes the expression of the CD3E gene in different cells as a density plot. With the gene information provided, we can display the gene description as a plot subtitle and the gene summary information as a plot caption. The information provided in Figure 2C helps us to capture the idea of how the CD3E gene functions and to interpret the expression distribution among different cell clusters.

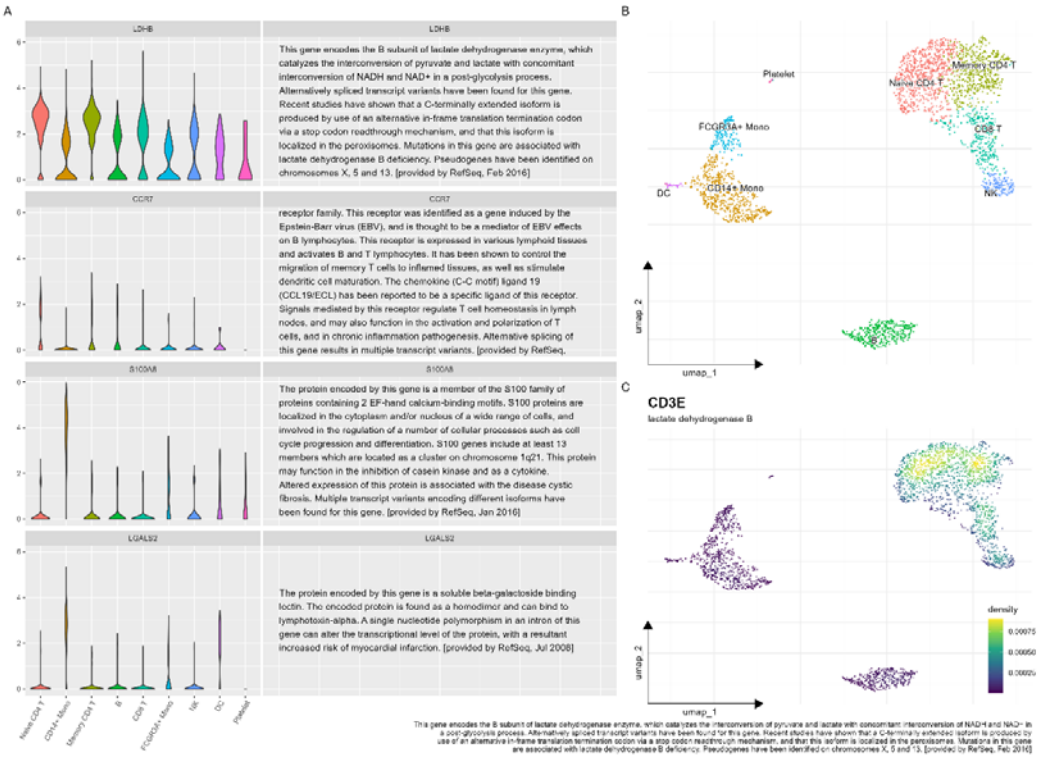


Figure 2. Use gene information to create an informative graph with self-explanation.

Showcase 2:

```
library(DESeq2)
library(dplyr)
library(clusterProfiler)
library(fanyi)
library(ggtree)

data(gse, package='airway')

dds <- DESeqDataSet(gse, design = ~donor + condition)
dds <- DESeq(dds)
res <- results(dds)

res2 <- as.data.frame(res)
res2$ID <- rownames(res)
res2 <- res2[!is.na(res2$log2FoldChange), ]
de <- arrange(res2, padj) |>
    group_by(sign(log2FoldChange)) |>
    slice(1:20) |>
    pull(ID)
```

```
mat <- assay(rlog(dds))[de, ]
mat <- mat - rowMeans(mat)

de <- sub("\\.\\d+", "", de)
rownames(mat) <- de
d <- as.data.frame(mat)
k <- kmeans(t(mat), 2)
d <- d[, names(sort(k$cluster))]
```

We used an example dataset provided in the airway package and employed DESeq2 to identify differentially expressed genes. We extracted 20 genes from up-regulated and down-regulated genes respectively. These 40 genes were further used to demonstrate how fanyi can help interpret the results.

```
g <- bitr(de, 'ENSEMBL', c('ENTREZID', 'SYMBOL', 'GENENAME'), 'org.Hs.eg.
db')
g$gn2 <- translate(g$GENENAME)

hc <- hclust(dist(mat))
p <- ggtree(hc)

p <- p %<+% g
p <- p + geom_tiplab(aes(label=SYMBOL), align=T)
p2 <- gheatmap(p, d, offset=3, colnames_angle=90, colnames_position="top",
 colnames_offset_y=2)

pg <- p2 +
    geom_tiplab(aes(label = gn2),
                align=TRUE, offset=13, linetype=NA) +
    hexpand(.6) + vexpand(.05) +
    theme(legend.position = c(.1, .9))
```

We used bitr function to convert gene IDs and used the translate function to convert the full gene name into Chinese. We then applied the hierarchical clustering method to cluster the genes based on the expression profiles. The clustering tree was visualized using ggtree (Yu, 2020) with gene symbols and then gheatmap was utilized to visualize the tree with the expression matrix. Finally, we displayed the Chinese version of the full gene names to help interpret the result (Figure 3A). With fanyi, users can translate information into their local language and display the information with the English version side by side. We no longer need to translate professional vocabulary manually. It saves a lot of time and helps users to quickly capture the information when looking at the figure.

As demonstrated in Figure 3A, the genes can be divided into two clades, we applied cuttree function to separate the genes into two clusters and then employed clusterProfiler to characterize biological processes regulated by these two gene clusters as demonstrated in Figure 3B. In fanyi, we provide a helper function, translate_ggplot, to help users translate words and sentences in a ggplot graph. As demonstrated in Figure 3C, the biological process terms presented in Figure 3B are translated into Chinese with only one line of function call.

```
x <- cutree(hc, k=2)
gc  = split(g$ENTREZID, x[g$ENSEMBL])
bp <- compareCluster(gc, fun=enrichGO, ont="BP", OrgDb='org.Hs.eg.db')

dp1 <- dotplot(bp)
dp2 <- translate_ggplot(p1, 'y')
```

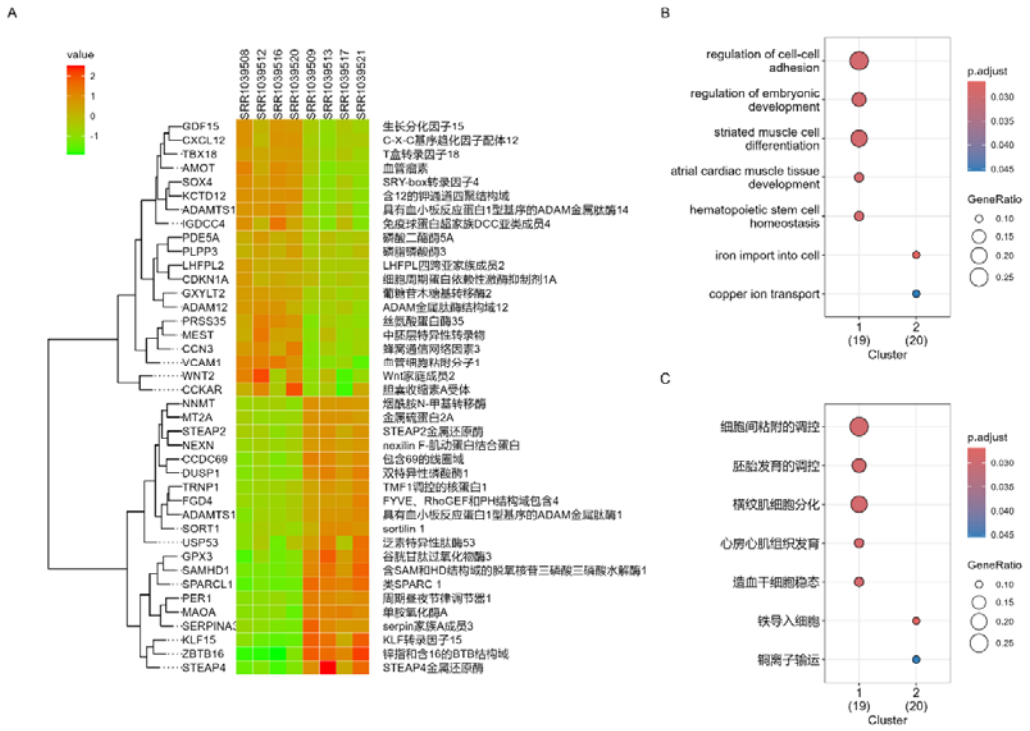


Figure 3. Translated information in local language help users quickly capture and understand the information presented.

Discussion

The fanyi package presented here can help retrieve and interpret information and thus save researchers a lot of time. It has a least three advantages:

1. Querying gene information to quickly create a summarized table with rich information for marker genes, differential expression genes, highly mutated genes, etc., to help users digest how the genes work.

2. The information query by the package can be utilized to annotate the plot to create an informative graph with more or less self-explanation.
3. Professional vocabulary hinders non-English researchers from understanding the results. Translating into the local language helps quickly digest the information.

Future plan:

1. Query information from pubmed to support more diverse searches.
2. Maybe OpenAI to help summarize information.
3. Maybe query information from database outside of biomedical fields.

Acknowledgments

This study was supported by the National Natural Science Foundation of China [32270677]. The manuscript was drafted with the help of the fanyi package. We appreciate the feedback and support from fanyi users.

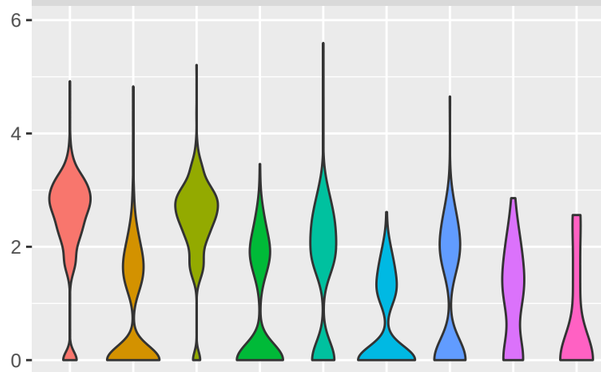
References

Wu,T. *et al.* (2021) clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, 100141.

Yu,G. (2020) Using ggtree to Visualize Data on Tree-Like Structures. *Curr. Protoc. Bioinforma.*, **69**, e96.

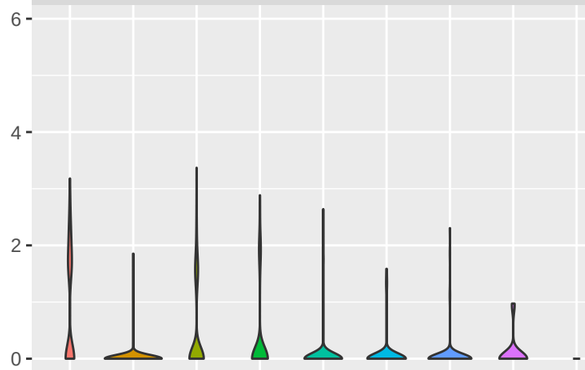
gene	avg_log2FC	p_val_adj	description	summary
CD8 T				
CCL5	3.326135	7.696099e-198	C-C motif chemokine ligand 5	This gene is one of several chemokine genes clustered on the q-arm of chromosome 17. Chemokines form a superfamily of secreted proteins involved in immunoregulatory and inflammatory processes. The superfamily is divided into four subfamilies based on the arrangement of the N-terminal cysteine residues of the mature peptide. This chemokine, a member of the CC subfamily, functions as a chemoattractant for blood monocytes, memory T helper cells and eosinophils. It causes the release of histamine from basophils and activates eosinophils. This cytokine is one of the major HIV-suppressive factors produced by CD8+ cells. It functions as one of the natural ligands for the chemokine receptor chemokine (C-C motif) receptor 5 (CCR5), and it suppresses in vitro replication of the R5 strains of HIV-1, which use CCR5 as a coreceptor. Alternative splicing results in multiple transcript variants that encode different isoforms. [provided by RefSeq, Jul 2013]
NKG7	2.459665	4.334551e-177	natural killer cell granule protein 7	Predicted to be integral component of plasma membrane. Predicted to be active in plasma membrane. [provided by Alliance of Genome Resources, Apr 2022]
Naive CD4 T				
CCR7	2.397366	1.312702e-83	C-C motif chemokine receptor 7	The protein encoded by this gene is a member of the G protein-coupled receptor family. This receptor was identified as a gene induced by the Epstein-Barr virus (EBV), and is thought to be a mediator of EBV effects on B lymphocytes. This receptor is expressed in various lymphoid tissues and activates B and T lymphocytes. It has been shown to control the migration of memory T cells to inflamed tissues, as well as stimulate dendritic cell maturation. The chemokine (C-C motif) ligand 19 (CCL19/ECL) has been reported to be a specific ligand of this receptor. Signals mediated by this receptor regulate T cell homeostasis in lymph nodes, and may also function in the activation and polarization of T cells, and in chronic inflammation pathogenesis. Alternative splicing of this gene results in multiple transcript variants. [provided by RefSeq, Sep 2014]
LDHB	1.206019	5.137444e-108	lactate dehydrogenase B	This gene encodes the B subunit of lactate dehydrogenase enzyme, which catalyzes the interconversion of pyruvate and lactate with concomitant interconversion of NADH and NAD+ in a post-glycolysis process. Alternatively spliced transcript variants have been found for this gene. Recent studies have shown that a C-terminally extended isoform is produced by use of an alternative in-frame translation termination codon via a stop codon readthrough mechanism, and that this isoform is localized in the peroxisomes. Mutations in this gene are associated with lactate dehydrogenase B deficiency. Pseudogenes have been identified on chromosomes X, 5 and 13. [provided by RefSeq, Feb 2016]
B				
CD79A	6.911221	0.000000e+00	CD79a molecule	The B lymphocyte antigen receptor is a multimeric complex that includes the antigen-specific component, surface immunoglobulin (Ig). Surface Ig non-covalently associates with two other proteins, Ig-alpha and Ig-beta, which are necessary for expression and function of the B-cell antigen receptor. This gene encodes the Ig-alpha protein of the B-cell antigen component. Alternatively spliced transcript variants encoding different isoforms have been described. [provided by RefSeq, Jul 2008]
MS4A1	5.718520	0.000000e+00	membrane spanning 4-domains A1	This gene encodes a member of the membrane-spanning 4A gene family. Members of this nascent protein family are characterized by common structural features and similar intron/exon splice boundaries and display unique expression patterns among hematopoietic cells and nonlymphoid tissues. This gene encodes a B-lymphocyte surface molecule which plays a role in the development and differentiation of B-cells into plasma cells. This family member is localized to 11q12, among a cluster of family members. Alternative splicing of this gene results in two transcript variants which encode the same protein. [provided by RefSeq, Jul 2008]
FCGR3A+ Mono				
CDKN1C	5.428812	2.316382e-208	cyclin dependent kinase inhibitor 1C	This gene is imprinted, with preferential expression of the maternal allele. The encoded protein is a tight-binding, strong inhibitor of several G1 cyclin/Cdk complexes and a negative regulator of cell proliferation. Mutations in this gene are implicated in sporadic cancers and Beckwith-Wiedemann syndorome, suggesting that this gene is a tumor suppressor candidate. Three transcript variants encoding two different isoforms have been found for this gene. [provided by RefSeq, Oct 2010]
HES4	4.993200	2.079608e-203	hes family bHLH transcription factor 4	Predicted to enable DNA-binding transcription factor activity, RNA polymerase II-specific and RNA polymerase II cis-regulatory region sequence-specific DNA binding activity. Predicted to be involved in anterior/posterior pattern specification and regulation of transcription by RNA polymerase II. Predicted to be part of chromatin. Predicted to be active in nucleus. [provided by Alliance of Genome Resources, Apr 2022]
DC				
FCER1A	7.633833	2.030720e-216	Fc epsilon receptor Ia	The immunoglobulin epsilon receptor (IgE receptor) is the initiator of the allergic response. When two or more high-affinity IgE receptors are brought together by allergen-bound IgE molecules, mediators such as histamine that are responsible for allergy symptoms are released. This receptor is comprised of an alpha subunit, a beta subunit, and two gamma subunits. The protein encoded by this gene represents the alpha subunit. [provided by RefSeq, Aug 2011]
SERPINF1	8.026573	1.997116e-203	serpin family F member 1	This gene encodes a member of the serpin family that does not display the serine protease inhibitory activity shown by many of the other serpin proteins. The encoded protein is secreted and strongly inhibits angiogenesis. In addition, this protein is a neurotrophic factor involved in neuronal differentiation in retinoblastoma cells. Mutations in this gene were found in individuals with osteogenesis imperfecta, type VI. [provided by RefSeq, Aug 2016]
NK				
FGFBP2	5.404509	4.673698e-233	fibroblast growth factor binding protein 2	This gene encodes a member of the fibroblast growth factor binding protein family. The encoded protein is a serum protein that is selectively secreted by cytotoxic lymphocytes and may be involved in cytotoxic lymphocyte-mediated immunity. An increase in the amount of gene product may be associated with atopic asthma and mild extrinsic asthma.[provided by RefSeq Staff, Oct 2008]
GZMB	5.977437	1.090260e-264	granzyme B	This gene encodes a member of the granzyme subfamily of proteins, part of the peptidase S1 family of serine proteases. The encoded preproprotein is secreted by natural killer (NK) cells and cytotoxic T lymphocytes (CTLs) and proteolytically processed to generate the active protease, which induces target cell apoptosis. This protein also processes cytokines and degrades extracellular matrix proteins, and these roles are implicated in chronic inflammation and wound healing. Expression of this gene may be elevated in human patients with cardiac fibrosis. [provided by RefSeq, Sep 2016]
Platelet				
GP9	11.484216	0.000000e+00	glycoprotein IX platelet	This gene encodes a small membrane glycoprotein found on the surface of human platelets. It forms a 1-to-1 noncovalent complex with glycoprotein Ib, a platelet surface membrane glycoprotein complex that functions as a receptor for von Willebrand factor. The complete receptor complex includes noncovalent association of the alpha and beta subunits with the protein encoded by this gene and platelet glycoprotein V. Defects in this gene are a cause of Bernard-Soulier syndrome, also known as giant platelet disease. These patients have unusually large platelets and have a clinical bleeding tendency. [provided by RefSeq, Oct 2008]
Memory CD4 T				
IL32	1.307077	3.966555e-86	interleukin 32	This gene encodes a member of the cytokine family. The protein contains a tyrosine sulfation site, 3 potential N-myristoylation sites, multiple putative phosphorylation sites, and an RGD cell-attachment sequence. Expression of this protein is increased after the activation of T-cells by mitogens or the activation of NK cells by IL-2. This protein induces the production of TNFalpha from macrophage cells. Alternate transcriptional splice variants, encoding different isoforms, have been characterized. [provided by RefSeq, Jul 2008]
LTB	1.331267	1.453850e-82	lymphotoxin beta	Lymphotoxin beta is a type II membrane protein of the TNF family. It anchors lymphotoxin-alpha to the cell surface through heterotrimer formation. The predominant form on the lymphocyte surface is the lymphotoxin-alpha 1/beta 2 complex (e.g. 1 molecule alpha/2 molecules beta) and this complex is the primary ligand for the lymphotoxin-beta receptor. The minor complex is lymphotoxin-alpha 2/beta 1. LTB is an inducer of the inflammatory response system and involved in normal development of lymphoid tissue. Lymphotoxin-beta isoform b is unable to complex with lymphotoxin-alpha suggesting a function for lymphotoxin-beta which is independent of lympyhotoxin-alpha. Alternative splicing results in multiple transcript variants encoding different isoforms. [provided by RefSeq, Jul 2008]
CD14+ Mono				
LGALS2	5.651790	0.000000e+00	galectin 2	The protein encoded by this gene is a soluble beta-galactoside binding lectin. The encoded protein is found as a homodimer and can bind to lymphotoxin-alpha. A single nucleotide polymorphism in an intron of this gene can alter the transcriptional level of the protein, with a resultant increased risk of myocardial infarction. [provided by RefSeq, Jul 2008]
S100A8	6.645056	0.000000e+00	S100 calcium binding protein A8	The protein encoded by this gene is a member of the S100 family of proteins containing 2 EF-hand calcium-binding motifs. S100 proteins are localized in the cytoplasm and/or nucleus of a wide range of cells, and involved in the regulation of a number of cellular processes such as cell cycle progression and differentiation. S100 genes include at least 13 members which are located as a cluster on chromosome 1q21. This protein may function in the inhibition of casein kinase and as a cytokine. Altered expression of this protein is associated with the disease cystic fibrosis. Multiple transcript variants encoding different isoforms have been found for this gene. [provided by RefSeq, Jan 2016]

A



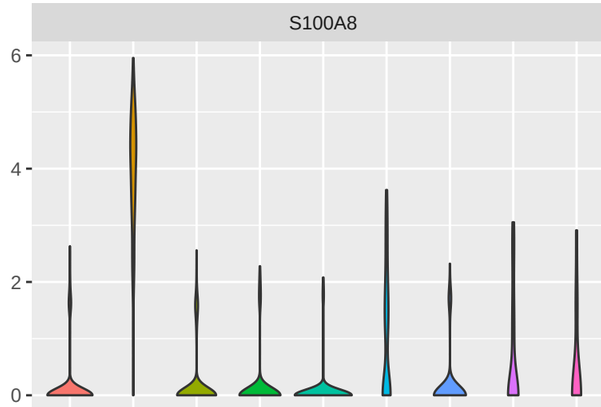
LDHB

This gene encodes the B subunit of lactate dehydrogenase enzyme, which catalyzes the interconversion of pyruvate and lactate with concomitant interconversion of NADH and NAD+ in a post−glycolysis process. Alternatively spliced transcript variants have been found for this gene. Recent studies have shown that a C−terminally extended isoform is produced by use of an alternative in−frame translation termination codon via a stop codon readthrough mechanism, and that this isoform is localized in the peroxisomes. Mutations in this gene are associated with lactate dehydrogenase B deficiency. Pseudogenes have been identified on chromosomes X, 5 and 13. [provided by RefSeq, Feb 2016]



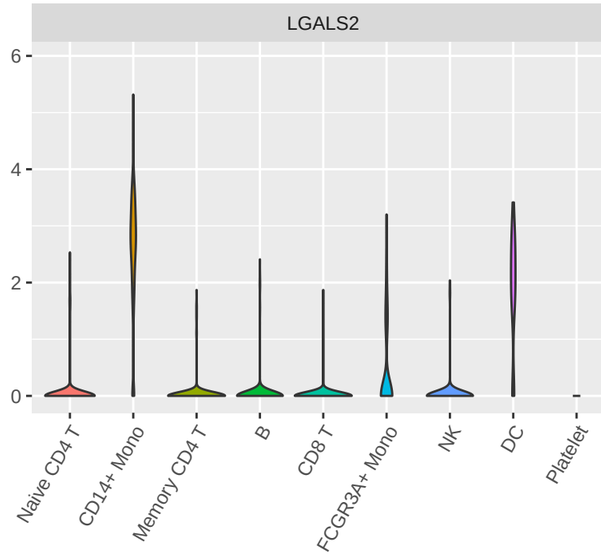
CCR7

receptor family. This receptor was identified as a gene induced by the Epstein−Barr virus (EBV), and is thought to be a mediator of EBV effects on B lymphocytes. This receptor is expressed in various lymphoid tissues and activates B and T lymphocytes. It has been shown to control the migration of memory T cells to inflamed tissues, as well as stimulate dendritic cell maturation. The chemokine (C−C motif) ligand 19 (CCL19/ECL) has been reported to be a specific ligand of this receptor. Signals mediated by this receptor regulate T cell homeostasis in lymph nodes, and may also function in the activation and polarization of T cells, and in chronic inflammation pathogenesis. Alternative splicing of this gene results in multiple transcript variants. [provided by RefSeq,



S100A8

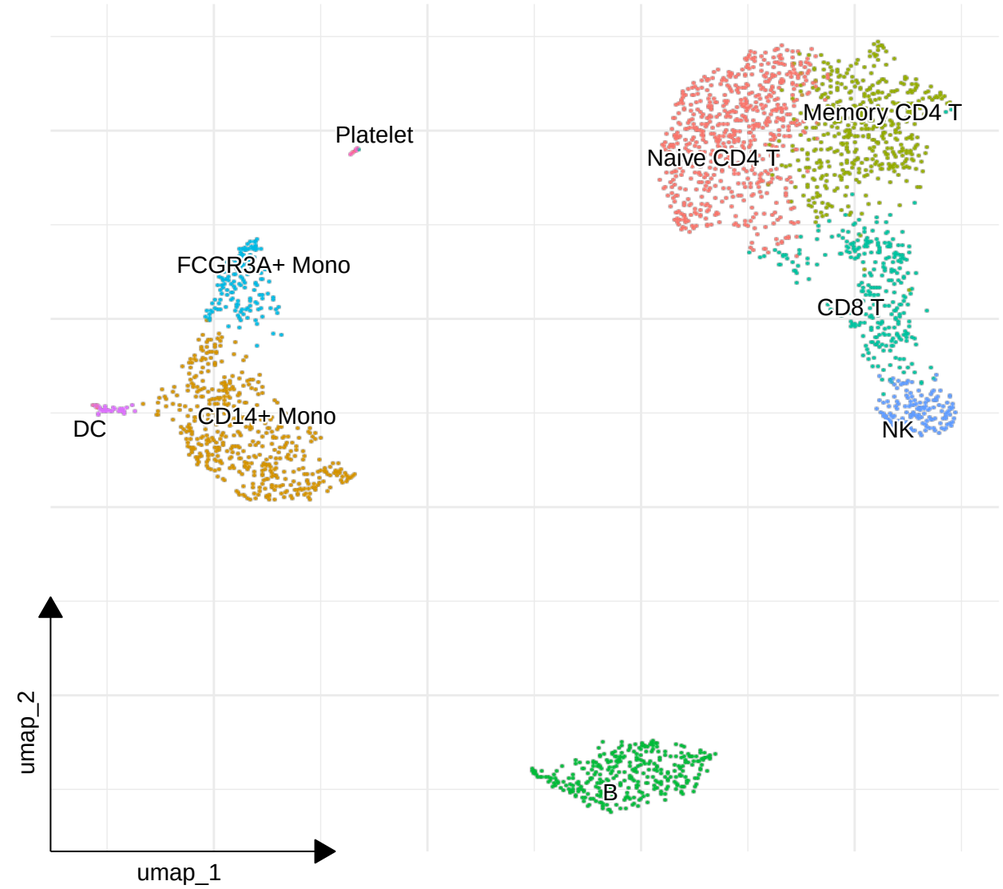
The protein encoded by this gene is a member of the S100 family of proteins containing 2 EF−hand calcium−binding motifs. S100 proteins are localized in the cytoplasm and/or nucleus of a wide range of cells, and involved in the regulation of a number of cellular processes such as cell cycle progression and differentiation. S100 genes include at least 13 members which are located as a cluster on chromosome 1q21. This protein may function in the inhibition of casein kinase and as a cytokine. Altered expression of this protein is associated with the disease cystic fibrosis. Multiple transcript variants encoding different isoforms have been found for this gene. [provided by RefSeq, Jan 2016]



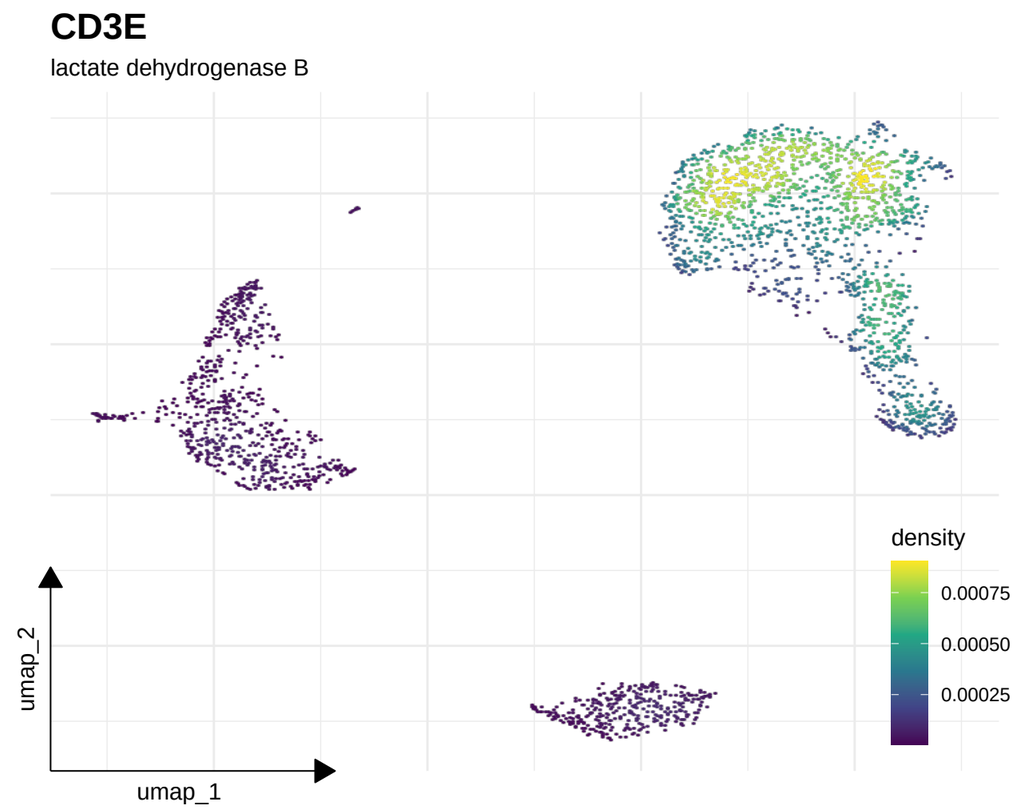
LGALS2

The protein encoded by this gene is a soluble beta−galactoside binding lectin. The encoded protein is found as a homodimer and can bind to lymphotoxin−alpha. A single nucleotide polymorphism in an intron of this gene can alter the transcriptional level of the protein, with a resultant increased risk of myocardial infarction. [provided by RefSeq, Jul 2008]

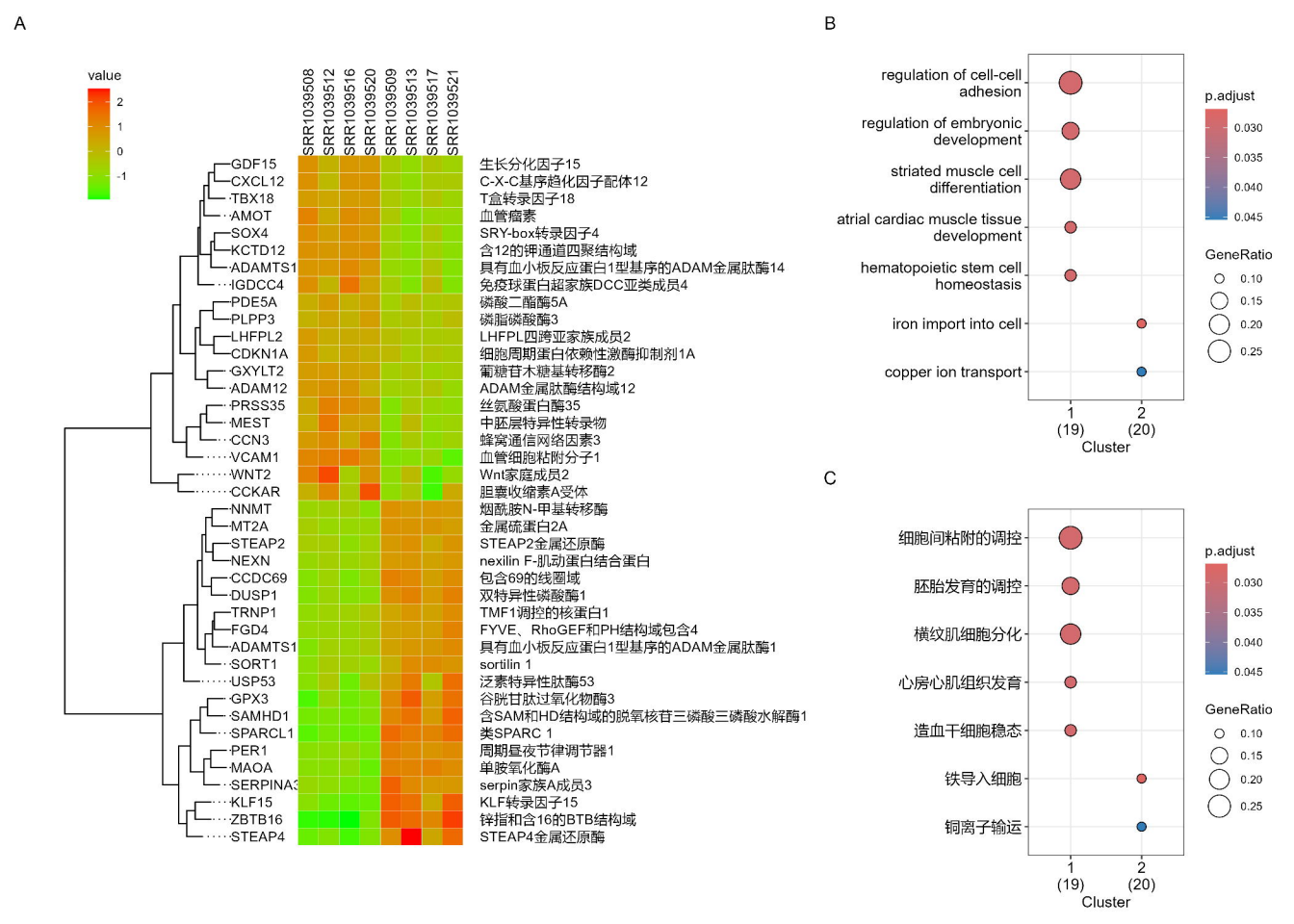
B



C



This gene encodes the B subunit of lactate dehydrogenase enzyme, which catalyzes the interconversion of pyruvate and lactate with concomitant interconversion of NADH and NAD+ in a post−glycolysis process. Alternatively spliced transcript variants have been found for this gene. Recent studies have shown that a C−terminally extended isoform is produced by use of an alternative in−frame translation termination codon via a stop codon readthrough mechanism, and that this isoform is localized in the peroxisomes. Mutations in this gene are associated with lactate dehydrogenase B deficiency. Pseudogenes have been identified on chromosomes X, 5 and 13. [provided by RefSeq, Feb 2016]


A
value
2
1
0
-1
SRR1039508
SRR1039512
SRR1039516
SRR1039520
SRR1039509
SRR1039513
SRR1039517
SRR1039521
GDF15 生长分化因子15
CXCL12 C-X-C基序趋化因子配体12
TBX18 T盒转录因子18
AMOT 血管瘤素
SOX4 SRY-box转录因子4
KCTD12 含12的钾通道四聚结构域
ADAMTS1 具有血小板反应蛋白1型基序的ADAM金属肽酶14
IGDCC4 免疫球蛋白超家族DCC亚类成员4
PDE5A 磷酸二酯酶5A
PLPP3 磷脂磷酸酶3
LHFPL2 LHFPL四跨亚家族成员2
CDKN1A 细胞周期蛋白依赖性激酶抑制剂1A
GXYLT2 葡糖苷木糖基转移酶2
ADAM12 ADAM金属肽酶结构域12
PRSS35 丝氨酸蛋白酶35
MEST 中胚层特异性转录物
CCN3 蜂窝通信网络因素3
VCAM1 血管细胞粘附分子1
WNT2 Wnt家庭成员2
CCKAR 胆囊收缩素A受体
NNMT 烟酰胺N-甲基转移酶
MT2A 金属硫蛋白2A
STEAP2 STEAP2金属还原酶
NEXN nexilin F-肌动蛋白结合蛋白
CCDC69 包含69的线圈域
DUSP1 双特异性磷酸酶1
TRNP1 TMF1调控的核蛋白1
FGD4 FYVE、RhoGEF和PH结构域包含4
ADAMTS1 具有血小板反应蛋白1型基序的ADAM金属肽酶1
SORT1 sortilin 1
USP53 泛素特异性肽酶53
GPX3 谷胱甘肽过氧化物酶3
SAMHD1 含SAM和HD结构域的脱氧核苷三磷酸三磷酸水解酶1
SPARCL1 类SPARC 1
PER1 周期昼夜节律调节器1
MAOA 单胺氧化酶A
SERPINA3 serpin家族A成员3
KLF15 KLF转录因子15
ZBTB16 锌指和含16的BTB结构域
STEAP4 STEAP4金属还原酶
B
regulation of cell-cell adhesion
regulation of embryonic development
striated muscle cell differentiation
atrial cardiac muscle tissue development
hematopoietic stem cell homeostasis
iron import into cell
copper ion transport
1 (19)
2 (20)
Cluster
p.adjust
0.030
0.035
0.040
0.045
GeneRatio
0.10
0.15
0.20
0.25
C
细胞间粘附的调控
胚胎发育的调控
横纹肌细胞分化
心房心肌组织发育
造血干细胞稳态
铁导入细胞
铜离子输运
1 (19)
2 (20)
Cluster
p.adjust
0.030
0.035
0.040
0.045
GeneRatio
0.10
0.15
0.20
0.25