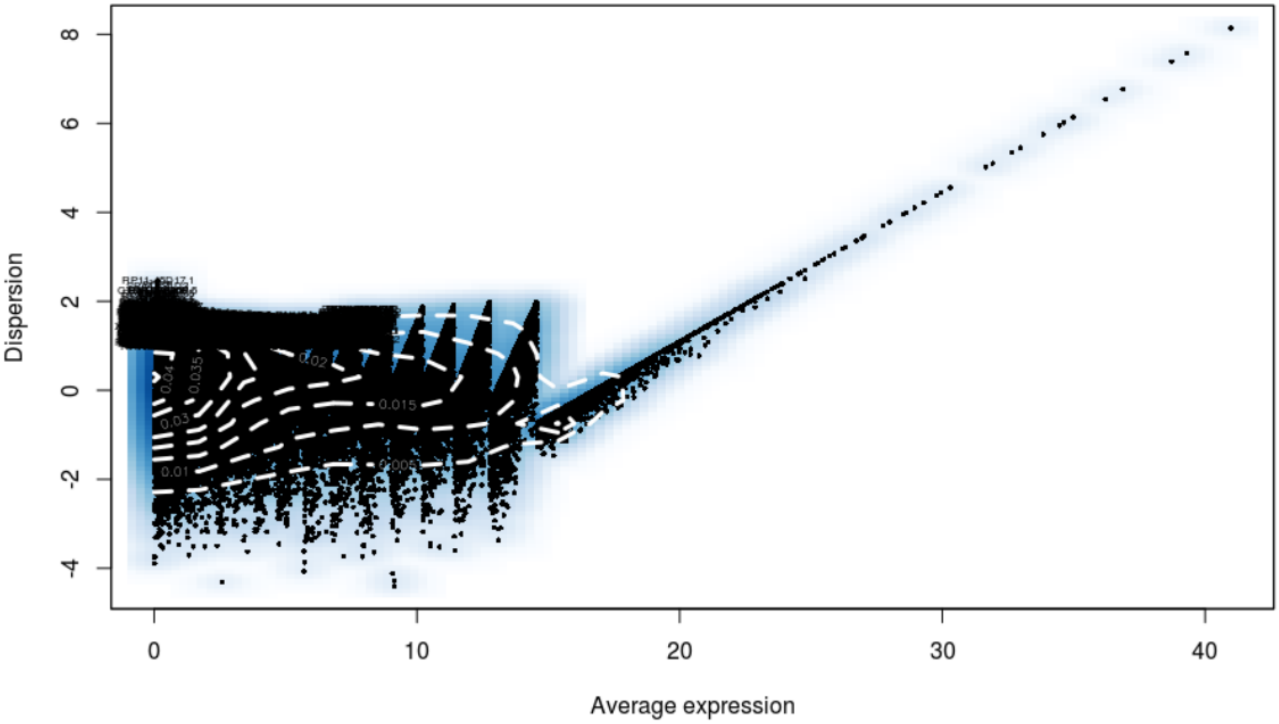
Analysis of scRNA-seq data on the examlpe of Li dataset

About dataset:
This dataset (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81861) is taken from the article about human colorectal tumors (https://www.nature.com/articles/ng.3818). However, these particular scRNA-seqs seem to be done on the model cell lines to check their new algorithm to analyse scRNA-seq data.

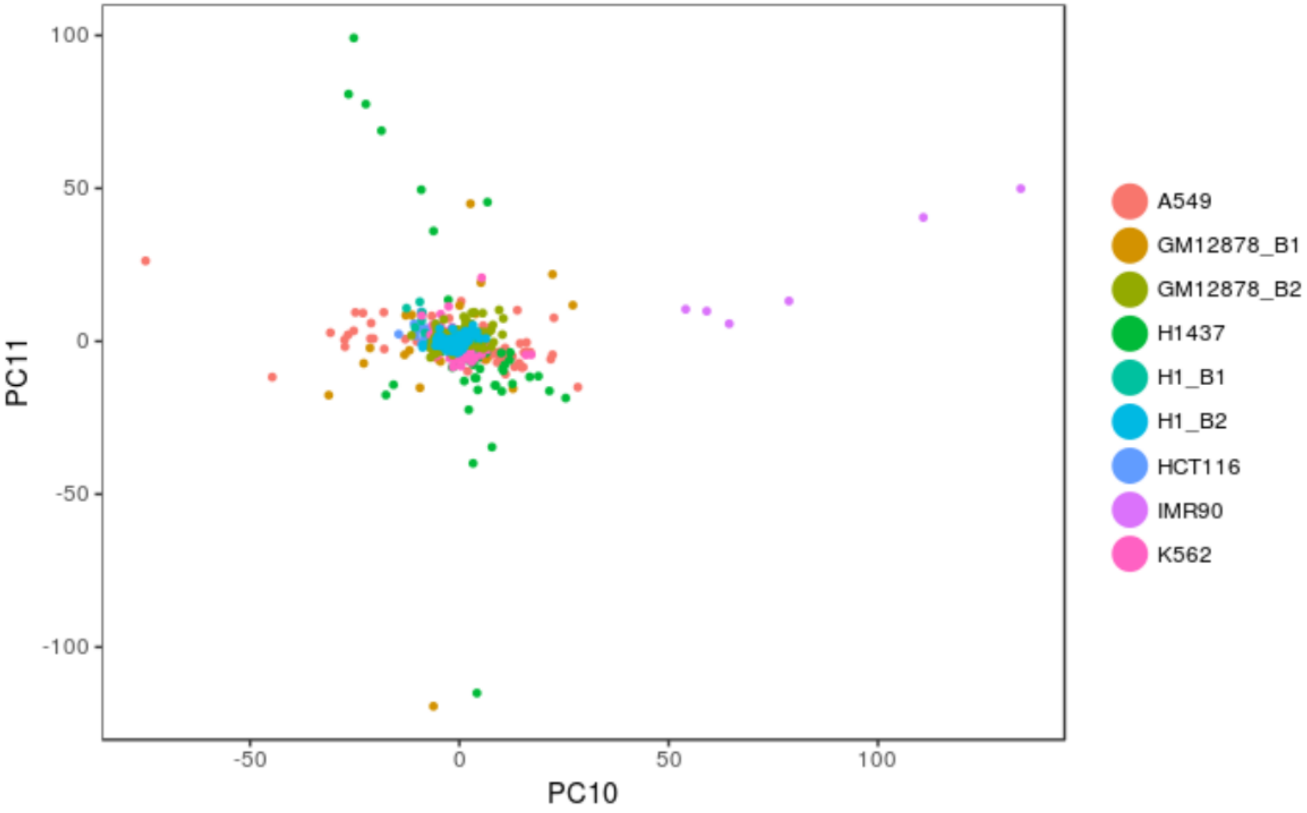
Analysis of quality

The 'mean variability plot':

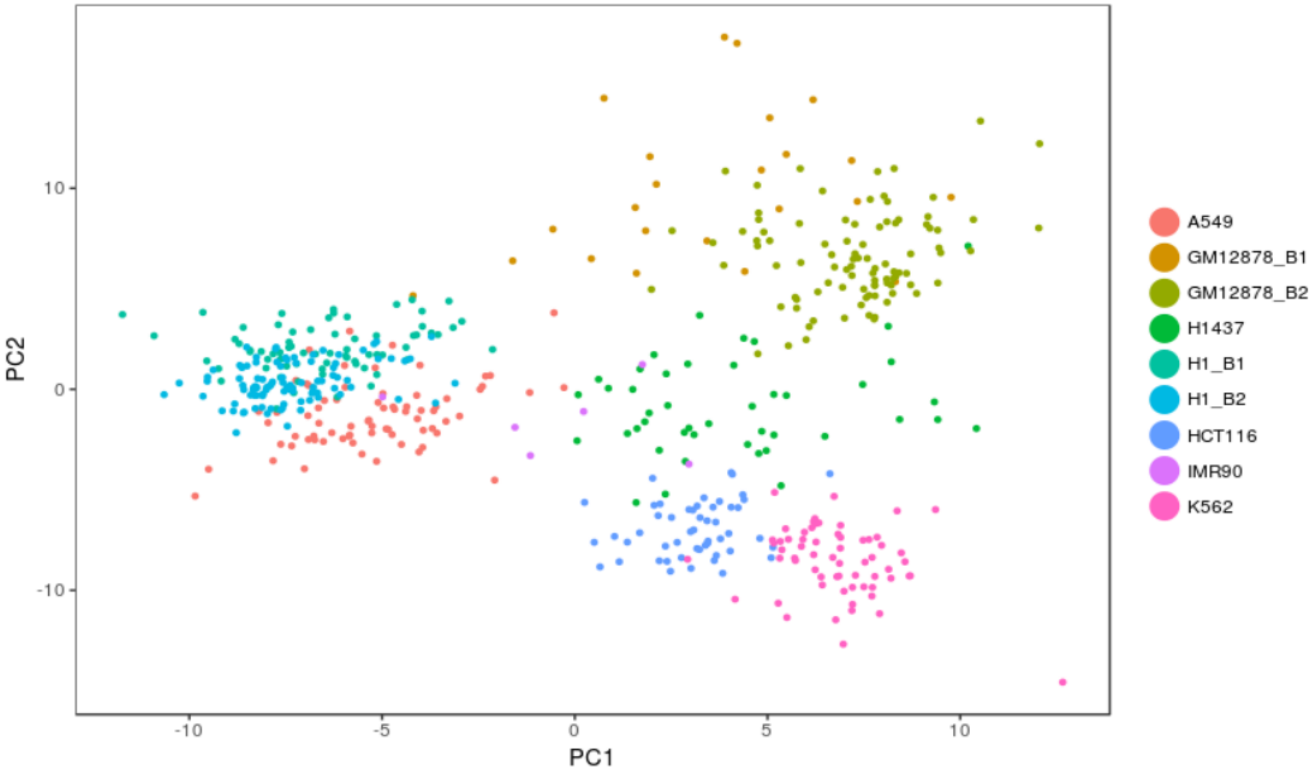


Analysis of clusters

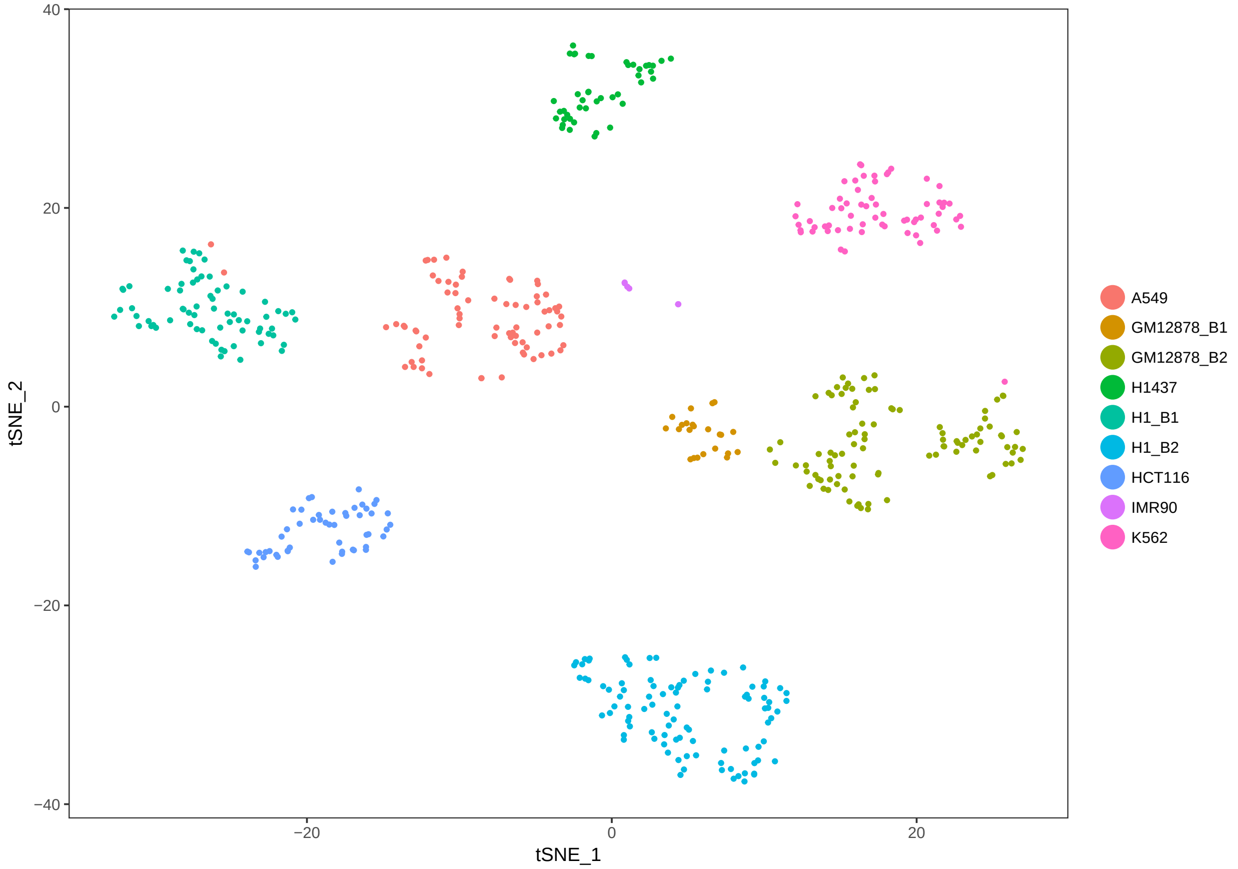
First we tried not to do any regression. The PCA plot looked like this:

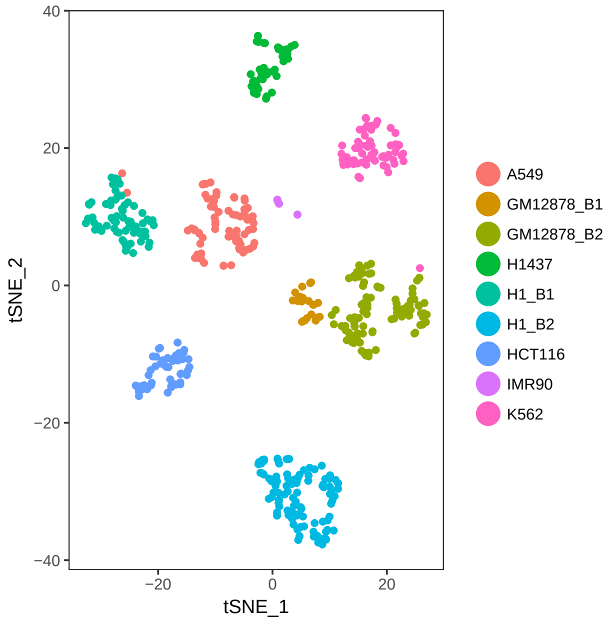


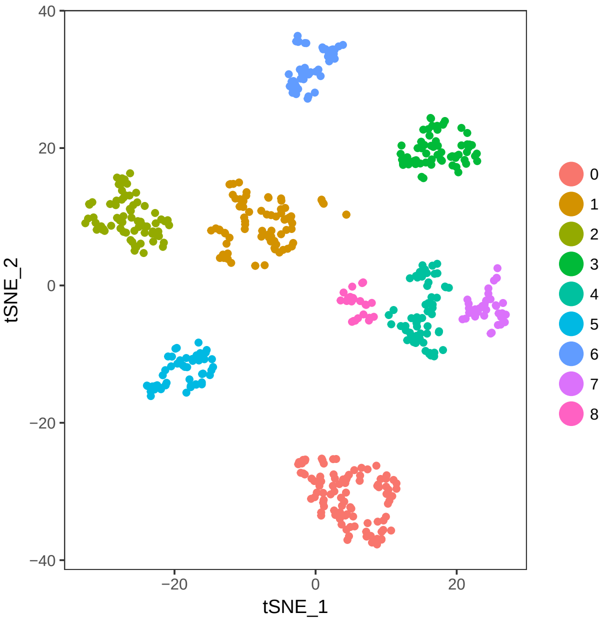
Then we used regression (regression variables were nUMI, G2M-score and S-score):



The clusters divided better on the second PCA, so we were using regression ever after. On the results of the PCA we performed t-SNE:







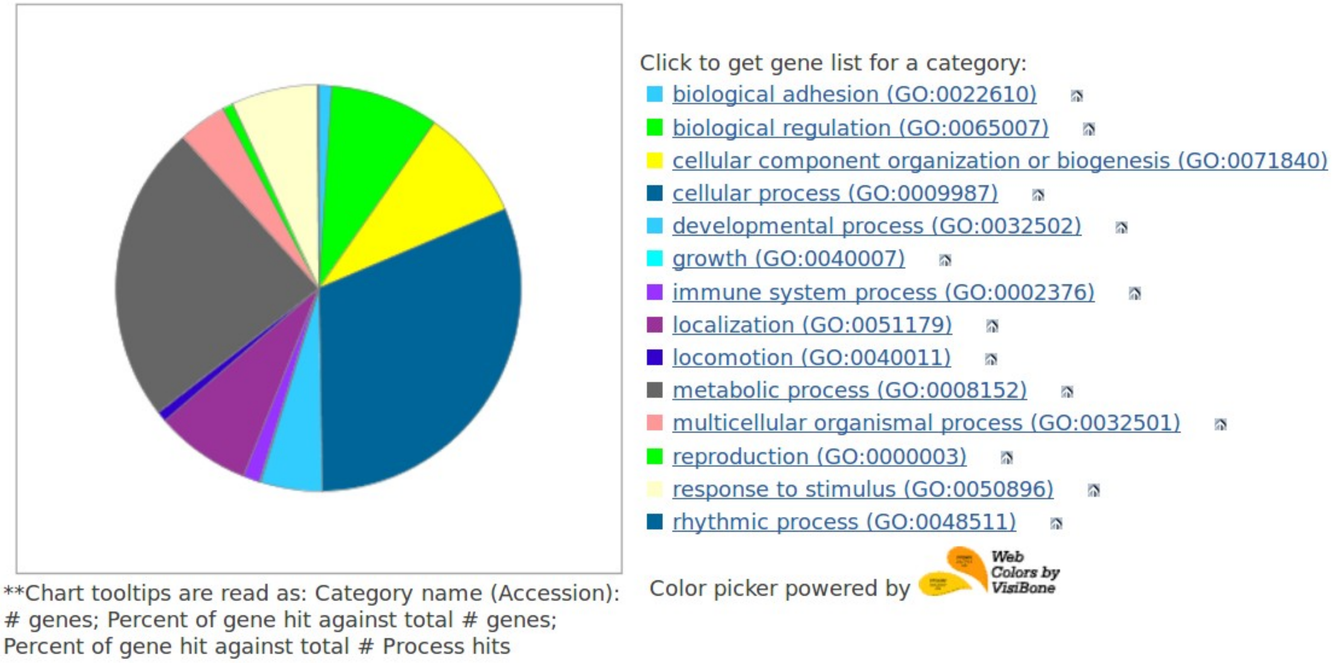
As we can see, the t-SNE plot divided the cells almost correctly. It is also interesting, that the batches for H1 and GM12878 are divided or our plots. The same results were received in the original article (they then made a correction for these batch effects).

Differential expression analysis

For differential expression analysis we chose 3 distinctive clusters, that were correctly distinguished on the t-SNE plot. The pie charts of enriched categories for these three clusters look practically the same (which is a bit weird). We also looked for several top DE genes for each cluster (which are different).

cluster 0 – H1_B2 (H1 cells – Human Embryonic Stem Cell Lines, batch 2)

Pie chart of enriched biological processes:



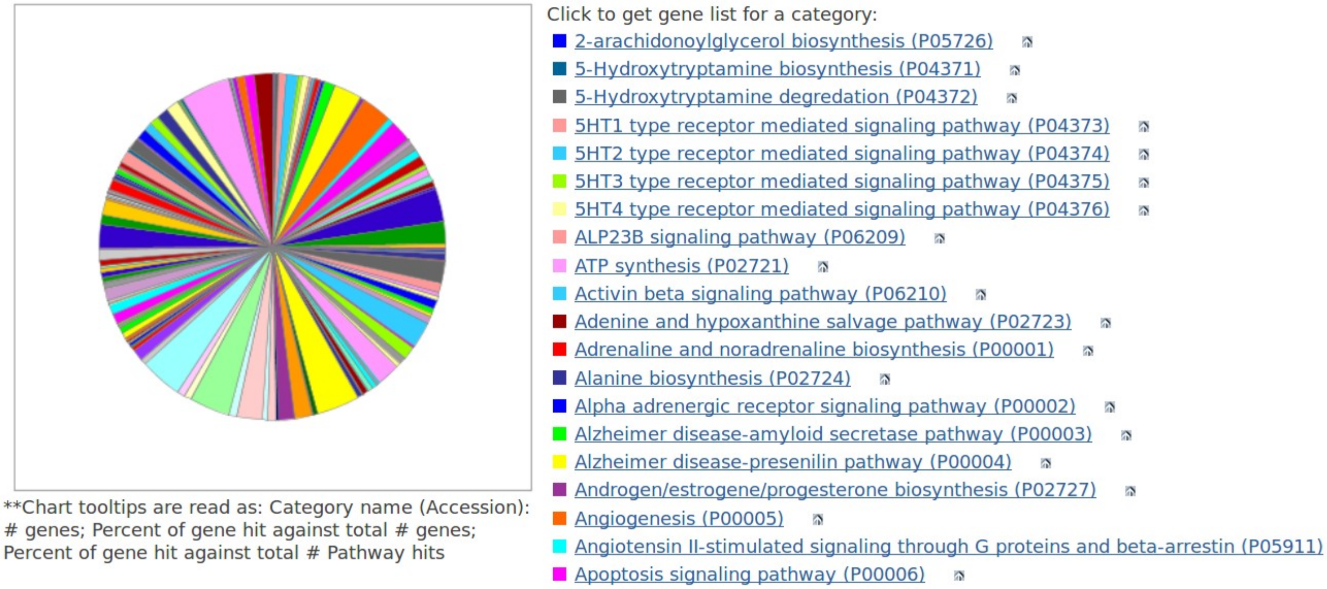
Click to get gene list for a category:

- ■ biological adhesion (GO:0022610)
- ■ biological regulation (GO:0065007)
- ■ cellular component organization or biogenesis (GO:0071840)
- ■ cellular process (GO:0009987)
- ■ developmental process (GO:0032502)
- ■ growth (GO:0040007)
- ■ immune system process (GO:0002376)
- ■ localization (GO:0051179)
- ■ locomotion (GO:0040011)
- ■ metabolic process (GO:0008152)
- ■ multicellular organismal process (GO:0032501)
- ■ reproduction (GO:0000003)
- ■ response to stimulus (GO:0050896)
- ■ rhythmic process (GO:0048511)

Color picker powered by Web Colors by VisiBone

**Chart tooltips are read as: Category name (Accession): # genes; Percent of gene hit against total # genes; Percent of gene hit against total # Process hits

Pie chart of enriched biological pathways:



Click to get gene list for a category:

- ■ 2-arachidonoylglycerol biosynthesis (P05726)
- ■ 5-Hydroxytryptamine biosynthesis (P04371)
- ■ 5-Hydroxytryptamine degredation (P04372)
- ■ 5HT1 type receptor mediated signaling pathway (P04373)
- ■ 5HT2 type receptor mediated signaling pathway (P04374)
- ■ 5HT3 type receptor mediated signaling pathway (P04375)
- ■ 5HT4 type receptor mediated signaling pathway (P04376)
- ■ ALP23B signaling pathway (P06209)
- ■ ATP synthesis (P02721)
- ■ Activin beta signaling pathway (P06210)
- ■ Adenine and hypoxanthine salvage pathway (P02723)
- ■ Adrenaline and noradrenaline biosynthesis (P00001)
- ■ Alanine biosynthesis (P02724)
- ■ Alpha adrenergic receptor signaling pathway (P00002)
- ■ Alzheimer disease-amyloid secretase pathway (P00003)
- ■ Alzheimer disease-presenilin pathway (P00004)
- ■ Androgen/estrogene/progesterone biosynthesis (P02727)
- ■ Angiogenesis (P00005)
- ■ Angiotensin II-stimulated signaling through G proteins and beta-arrestin (P05911)
- ■ Apoptosis signaling pathway (P00006)

**Chart tooltips are read as: Category name (Accession): # genes; Percent of gene hit against total # genes; Percent of gene hit against total # Pathway hits

Interesting top genes:

- **ESRG** – Embryonic Stem Cell Related (Non-Protein Coding).
- **TDGF1** – this gene encodes an epidermal growth factor-related protein that contains a cripto, FRL-1, and cryptic domain. The encoded protein is an extracellular, membrane-bound signaling protein that plays an essential role in embryonic development and tumor growth.
- **BCL11A** – gene, coding B-cell lymphoma/leukemia 11A protein.

cluster 3 – K562 cells (human immortalised myelogenous leukemia line)

Pie chart of enriched biological processes:

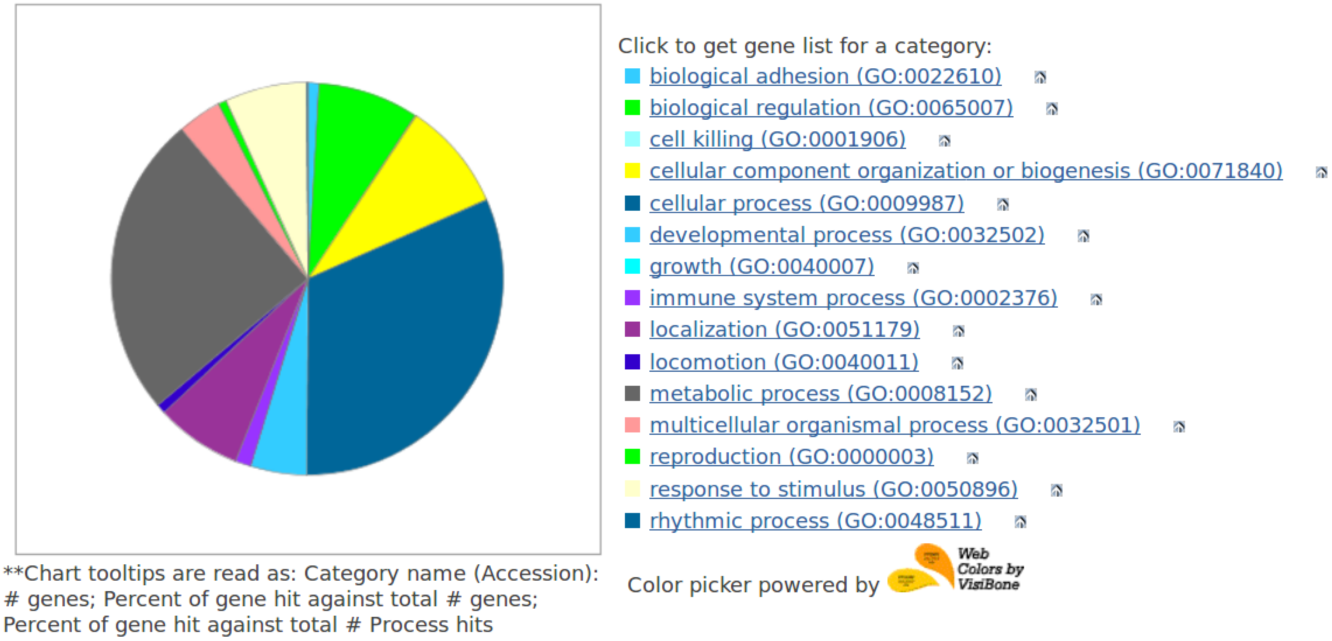


**Chart tooltips are read as: Category name (Accession): # genes; Percent of gene hit against total # genes; Percent of gene hit against total # Process hits

Click to get gene list for a category:

- ■ biological adhesion (GO:0022610)
- ■ biological regulation (GO:0065007)
- ■ cell killing (GO:0001906)
- ■ cellular component organization or biogenesis (GO:0071840)
- ■ cellular process (GO:0009987)
- ■ developmental process (GO:0032502)
- ■ growth (GO:0040007)
- ■ immune system process (GO:0002376)
- ■ localization (GO:0051179)
- ■ locomotion (GO:0040011)
- ■ metabolic process (GO:0008152)
- ■ multicellular organismal process (GO:0032501)
- ■ reproduction (GO:0000003)
- ■ response to stimulus (GO:0050896)
- ■ rhythmic process (GO:0048511)

Color picker powered by Web Colors by VisiBone

Pie chart of enriched biological pathways:

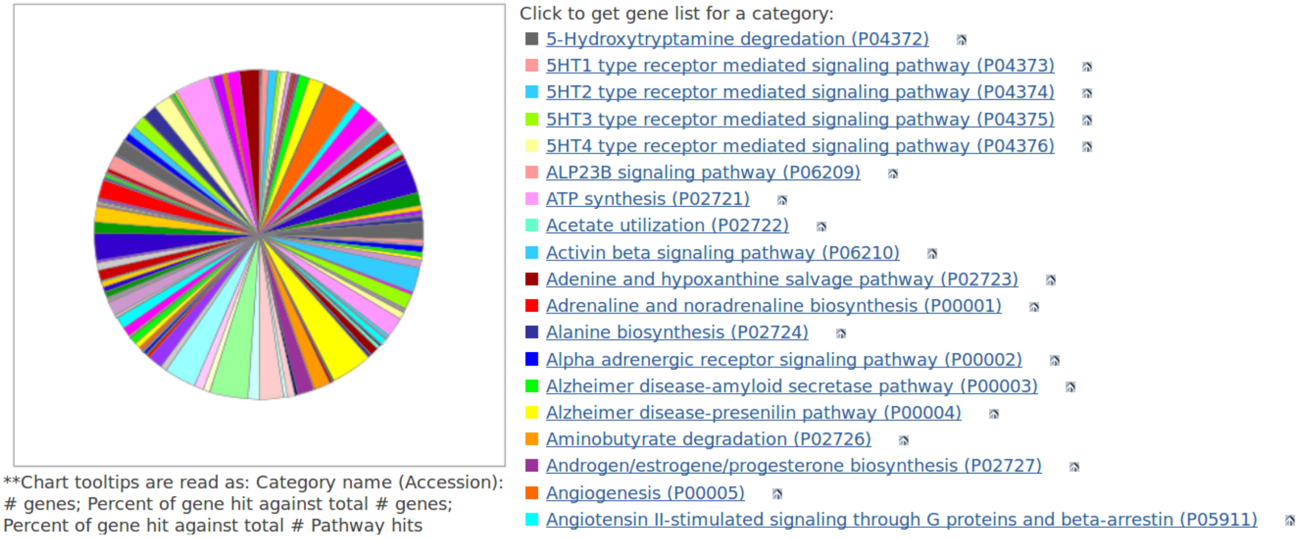


**Chart tooltips are read as: Category name (Accession): # genes; Percent of gene hit against total # genes; Percent of gene hit against total # Pathway hits

Click to get gene list for a category:

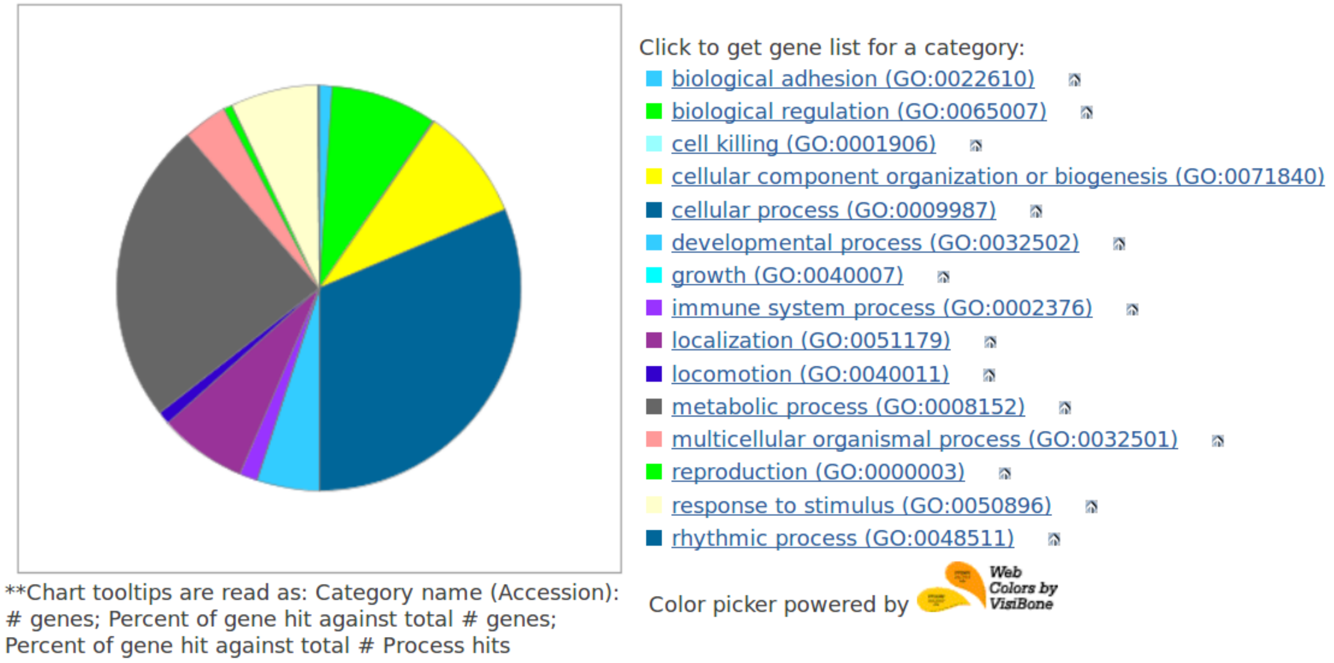
- ■ 5-Hydroxytryptamine degredation (P04372)
- ■ 5HT1 type receptor mediated signaling pathway (P04373)
- ■ 5HT2 type receptor mediated signaling pathway (P04374)
- ■ 5HT3 type receptor mediated signaling pathway (P04375)
- ■ 5HT4 type receptor mediated signaling pathway (P04376)
- ■ ALP23B signaling pathway (P06209)
- ■ ATP synthesis (P02721)
- ■ Acetate utilization (P02722)
- ■ Activin beta signaling pathway (P06210)
- ■ Adenine and hypoxanthine salvage pathway (P02723)
- ■ Adrenaline and noradrenaline biosynthesis (P00001)
- ■ Alanine biosynthesis (P02724)
- ■ Alpha adrenergic receptor signaling pathway (P00002)
- ■ Alzheimer disease-amyloid secretase pathway (P00003)
- ■ Alzheimer disease-presenilin pathway (P00004)
- ■ Aminobutyrate degradation (P02726)
- ■ Androgen/estrogene/progesterone biosynthesis (P02727)
- ■ Angiogenesis (P00005)
- ■ Angiotensin II-stimulated signaling through G proteins and beta-arrestin (P05911)

Interesting top genes:

- **CTCFL** (CCCTC-Binding Factor Like) is a Protein Coding gene. Seems to act as tumor suppressor.
- **GAGE** (GAGE12I, GAGE12F, GAGE12H, GAGE12C, GAGE2E, etc.) – a family of genes that are expressed in a variety of tumors but not in normal tissues, except for the testis.

cluster 5 – HCT116 cells (human colon cancer cell line)

Pie chart of enriched biological processes:



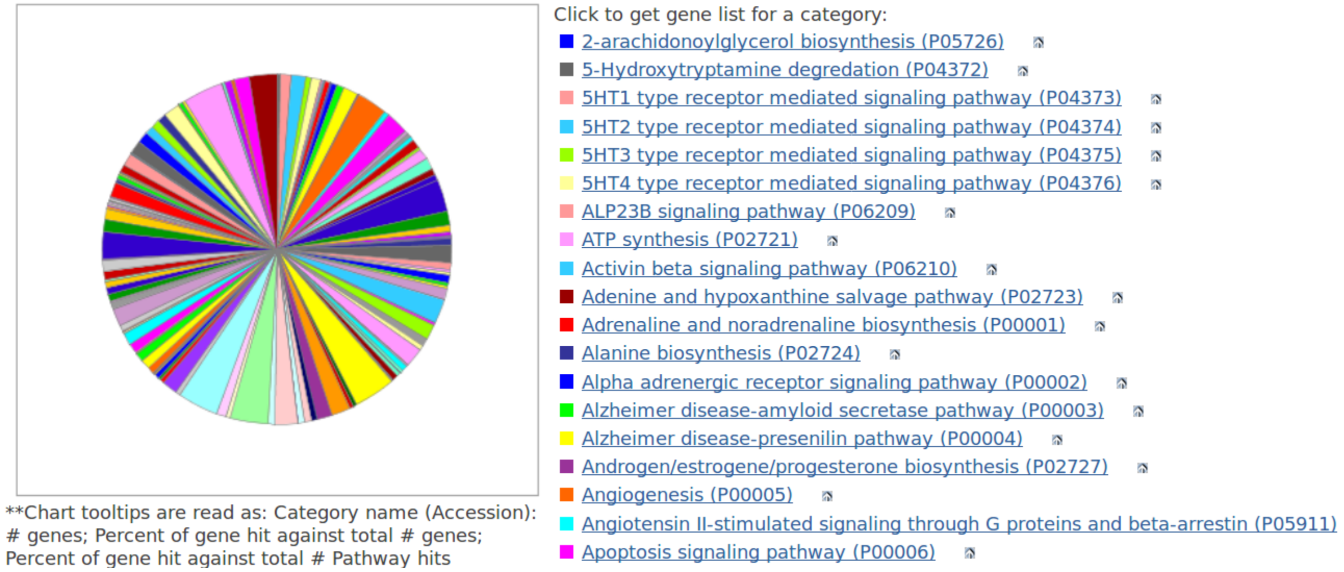
Click to get gene list for a category:

- biological adhesion (GO:0022610)
- biological regulation (GO:0065007)
- cell killing (GO:0001906)
- cellular component organization or biogenesis (GO:0071840)
- cellular process (GO:0009987)
- developmental process (GO:0032502)
- growth (GO:0040007)
- immune system process (GO:0002376)
- localization (GO:0051179)
- locomotion (GO:0040011)
- metabolic process (GO:0008152)
- multicellular organismal process (GO:0032501)
- reproduction (GO:0000003)
- response to stimulus (GO:0050896)
- rhythmic process (GO:0048511)

**Chart tooltips are read as: Category name (Accession): # genes; Percent of gene hit against total # genes; Percent of gene hit against total # Process hits

Color picker powered by Web Colors by VisiBone

Pie chart of enriched biological pathways:



Click to get gene list for a category:

- 2-arachidonoylglycerol biosynthesis (P05726)
- 5-Hydroxytryptamine degredation (P04372)
- 5HT1 type receptor mediated signaling pathway (P04373)
- 5HT2 type receptor mediated signaling pathway (P04374)
- 5HT3 type receptor mediated signaling pathway (P04375)
- 5HT4 type receptor mediated signaling pathway (P04376)
- ALP23B signaling pathway (P06209)
- ATP synthesis (P02721)
- Activin beta signaling pathway (P06210)
- Adenine and hypoxanthine salvage pathway (P02723)
- Adrenaline and noradrenaline biosynthesis (P00001)
- Alanine biosynthesis (P02724)
- Alpha adrenergic receptor signaling pathway (P00002)
- Alzheimer disease-amyloid secretase pathway (P00003)
- Alzheimer disease-presenilin pathway (P00004)
- Androgen/estrogene/progesterone biosynthesis (P02727)
- Angiogenesis (P00005)
- Angiotensin II-stimulated signaling through G proteins and beta-arrestin (P05911)
- Apoptosis signaling pathway (P00006)

**Chart tooltips are read as: Category name (Accession): # genes; Percent of gene hit against total # genes; Percent of gene hit against total # Pathway hits

Interesting top genes:

- **ALDH1A3**. This gene encodes an aldehyde dehydrogenase enzyme that uses retinal as a substrate. Mutations in this gene have been associated with microphthalmia, isolated 8, and expression changes have also been detected in tumor cells.

- **S100A14**. This gene encodes a member of the S100 protein family which contains an EF-hand motif and binds calcium. The gene is located in a cluster of S100 genes on chromosome 1. Levels of the encoded protein have been found to be lower (!) in cancerous tissue and associated with metastasis suggesting a tumor suppressor function.
- **ZBED2**, paralog of ZBED3. About ZBED3: this gene belongs to a class of genes that arose through hAT DNA transposition and that encode regulatory proteins. This gene is upregulated in lung cancer tissues, where the encoded protein causes an accumulation of beta-catenin and enhanced lung cancer cell invasion. In addition, the encoded protein can be secreted and be involved in resistance to insulin.