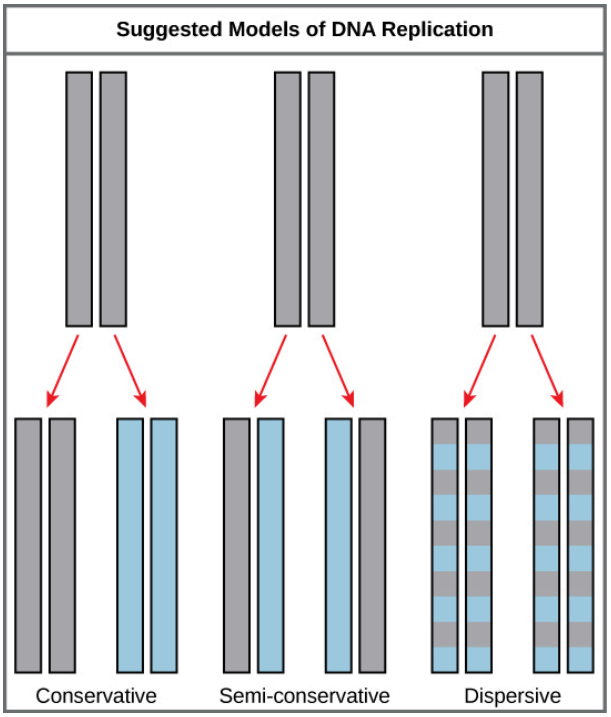
Basics of DNA Replication

DNA replication ensures that each daughter cell gets an exact copy of the DNA during cell division. The double-stranded structure of DNA suggested that the two strands could separate and each serve as a template for a new complementary strand, producing two identical DNA molecules.



Information Content in DNA, RNA, and Protein Sequences

Overview of Prokaryotic Transcription

Prokaryotic transcription is the process where messenger **RNA (mRNA) is made from DNA to produce proteins**. It happens in the **cytoplasm**, where **transcription and translation can occur simultaneously**. Unlike eukaryotes, which have **separate spaces for transcription (nucleus) and translation (cytoplasm), prokaryotes don't have a nucleus, so their genetic material is directly accessible to ribosomes in the cytoplasm.**

Transcription Process

- **Initiation**: RNA polymerase (RNAP) binds to a sigma factor to form a holoenzyme, which recognizes and binds to the promoter region on DNA. The DNA is unwound at the initiation site, creating an open complex.
- **Elongation**: RNA polymerase starts transcribing the DNA into RNA. Initially, some short, non-productive RNA segments are made, but once the sigma factor dissociates, transcription proceeds.
- **Termination**: Two mechanisms stop transcription:
    - **Intrinsic (Rho-independent)**: A palindromic sequence in the RNA forms a hairpin loop that causes RNA polymerase to dissociate.

- **Rho-dependent**: The Rho factor protein binds to the RNA and moves towards RNA polymerase, causing it to detach from the DNA.

Additional Details

- **Promoter Strength**: The strength of transcription is influenced by how tightly RNA polymerase binds to the promoter region, which depends on how similar the sequence is to the consensus sequence.
- **Transcription Factors**: These proteins can regulate how stable the transcription initiation process is and control the efficiency of RNA production.

Eukaryotic DNA Replication

DNA replication in eukaryotes is a complex process **involving many enzymes and proteins**. It occurs in three stages: **initiation, elongation, and termination.**

Initiation

DNA in eukaryotes is wrapped around histones to form nucleosomes. During initiation, the DNA must be made accessible to replication proteins. Replication starts at specific locations called **origins of replication**. In organisms like yeast, these origins have specific DNA sequences that attract replication proteins, while in humans, the origins may be identified through modifications in the nucleosomes.

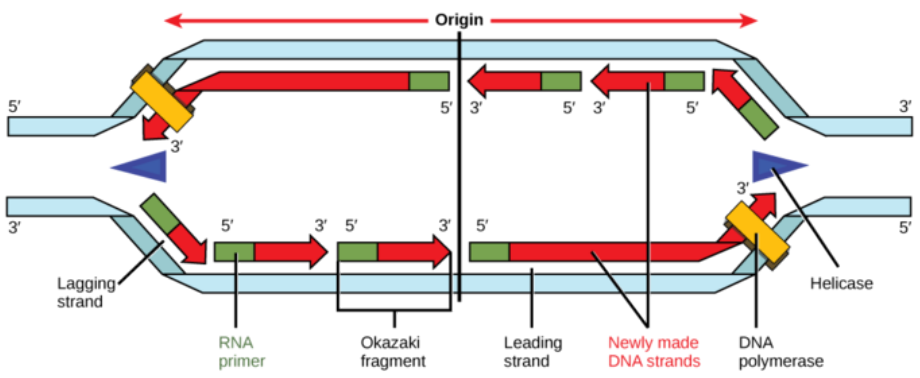
The process begins when proteins bind to the origin, "recruiting" others, including **helicase** enzymes that unwind the DNA. This creates two **replication forks** at the origin, forming a **replication bubble**. Multiple origins of replication allow DNA replication to happen in many places simultaneously.

Elongation

In elongation, **DNA polymerase** adds new DNA nucleotides to the 3′ end of a growing strand, matching them with the template DNA strand. DNA polymerase can't start a new strand, so **primase** first makes a short RNA primer. DNA polymerase then extends the new strand by adding nucleotides complementary to the template.

There are two strands in each replication fork:

- The **leading strand** is made continuously in the direction of the replication fork.
- The **lagging strand** is made in short pieces, called **Okazaki fragments**, because DNA polymerase can only add nucleotides in the 5′ to 3′ direction. Each Okazaki fragment starts with its own RNA primer.



Termination

Replication finishes when the leading strand of one replication bubble reaches the lagging strand of another bubble. The DNA polymerase stops when it encounters already-replicated DNA. However, there are **nicks** (gaps in the sugar-phosphate backbone) where the new strands are not fully connected.

The RNA primers are removed by **FEN1** and **RNase H**, and the gaps are filled by DNA polymerase. Finally, the enzyme **ligase** seals the nicks, completing the new DNA strand and finishing replication.

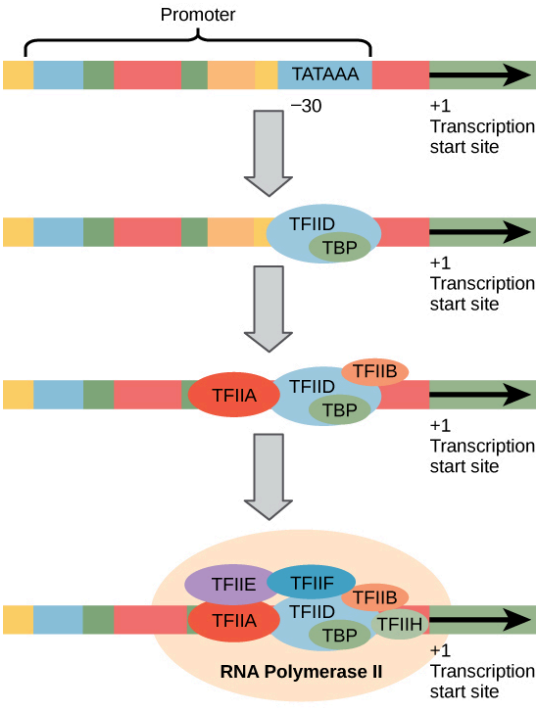
The Promoter and Transcription Machinery in Eukaryotes

Genes have a promoter region that helps control gene expression. The promoter is located just before the gene's coding sequence and can be short or long. A longer promoter allows for more protein binding, offering better control of gene expression. The size of the promoter and the level of control can vary widely between genes.

The promoter's main job is to bind **transcription factors** that help start the transcription process.

Within the promoter, there is a key part called the **TATA box**, which is made up of repeated thymine (T) and adenine (A) bases. This is where the **RNA polymerase** binds to start transcription.

To begin transcription, the first transcription factor, **TFIID**, binds to the TATA box. This attracts other transcription factors like **TFIIB**, **TFIIE**, **TFIIF**, and **TFIIH**. Once the transcription initiation complex is formed, RNA polymerase binds to the DNA and is activated by phosphorylation. This ensures that RNA polymerase is in the correct position to begin transcription.



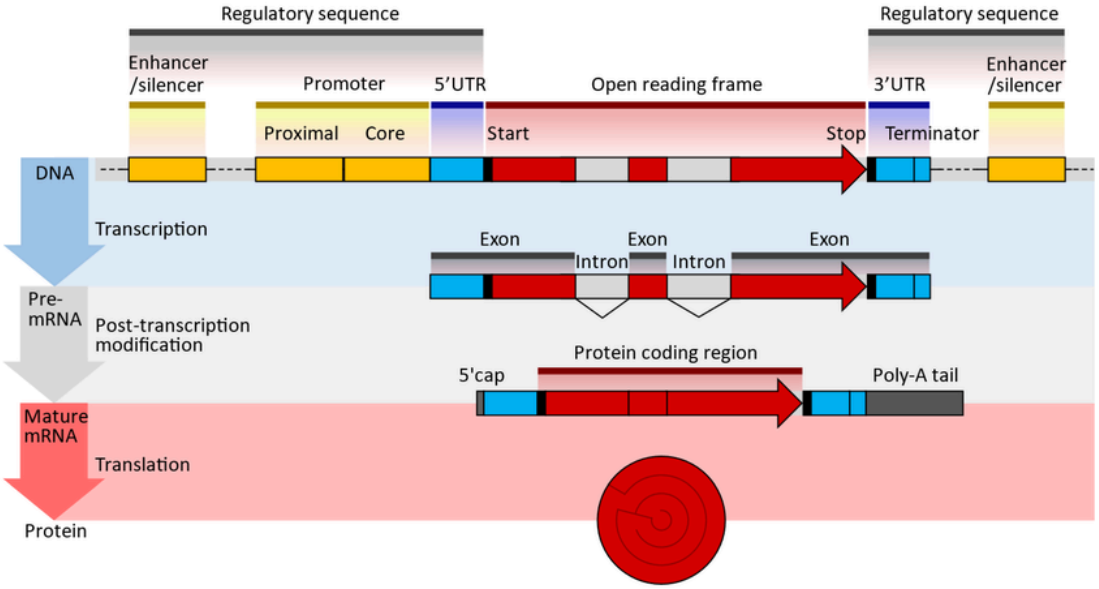
In addition to general transcription factors, there are **specific transcription factors** that bind to the promoter to regulate gene transcription. These factors are unique to certain genes and are activated by environmental signals. When these transcription factors bind to the promoter, they are called **cis-acting elements** because they are located on the same chromosome as the gene. These factors help turn genes on or off when needed.

Pre-mRNA Processing

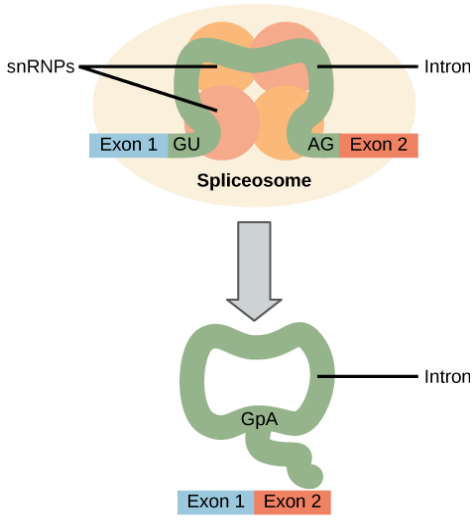
Eukaryotic pre-mRNA undergoes several processing steps before it can be translated into a protein. These steps help the mRNA last longer and are essential for its stability and proper functioning. Unlike prokaryotic mRNA, which has a very short lifespan (only about 5 seconds), eukaryotic mRNA lasts for several hours.

1. **5′ Capping**
   As the pre-mRNA is being made, a special cap **called 7-methylguanosine** is added to the 5′ end. This cap protects the mRNA from degradation and helps ribosomes recognize the mRNA for translation.

2. **3′ Poly-A Tail**
   After transcription, an enzyme cleaves the pre-mRNA and adds a string of about **200 adenine (A) nucleotides to the 3′ end, called the poly-A tail.** This tail protects the mRNA from degradation, helps it leave the nucleus, and plays a role in translation initiation.



3. **Pre-mRNA Splicing**
   Eukaryotic genes have **exons** (coding sequences) and **introns** (non-coding sequences). Introns need to be removed from the pre-mRNA before it can be used to make a protein. The process of removing introns is called **splicing**.



- **Discovery of Introns**: Introns were first discovered in the 1970s when scientists realized that some genes contain non-coding sequences. These sequences may regulate gene expression or be remnants of ancient genes.
- **Splicing Mechanism**: Splicing happens in the nucleus, carried out by a complex of proteins and RNA called the **spliceosome**. The

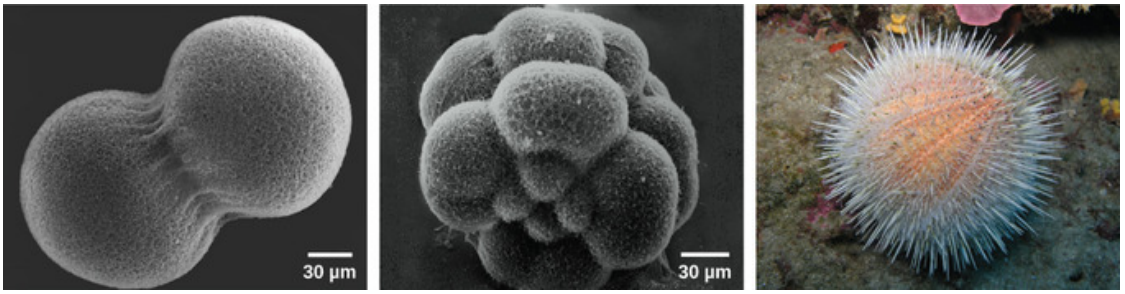
spliceosome identifies the start and end of introns and cuts them out, connecting the exons together.

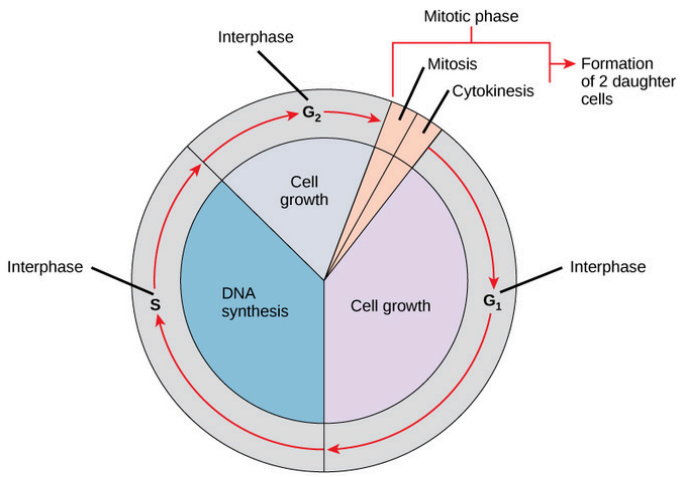
The result is a mature mRNA molecule with a 5′ cap, a 3′ poly-A tail, and only the exons left, ready to be translated into protein.

Introduction: Cell Division and Reproduction

Every human, like all sexually-reproducing organisms, starts life as a single fertilized egg or zygote. This one cell divides repeatedly to form the trillions of cells that make up the body. Even after reaching full growth, cell division continues to repair and regenerate tissues, such as producing new blood and skin cells. **All multicellular organisms rely on cell division for growth, maintenance, and repair.** Cell division is closely regulated, and mistakes in this process can be dangerous. Single-celled organisms also use cell division to reproduce.

Most cells in the body, called **somatic cells** (all body cells except egg and sperm cells), divide regularly. Somatic cells have two copies of each chromosome, one from each parent. These cells replace themselves throughout a person's life. For example, the cells in the digestive tract are constantly replaced.





Cell division occurs in a series of controlled steps known as the **cell cycle**, which includes cell growth, DNA replication, and division into two identical daughter cells. The cell cycle has two main phases:

1. **Interphase** - the cell grows and replicates its DNA.
2. **Mitotic phase** - the DNA and other cell contents are divided between two new cells.

Studying Gene Expression and Function

To understand how genes and the proteins they produce work, researchers often study **organisms that are missing a specific gene**. This approach helps reveal the function of the gene by observing what happens when it's absent.

In modern genetics, **gene function is often studied starting with DNA sequences from genome projects**. The challenge is to figure out what these sequences do. **One method is to compare the new gene's sequence with known proteins to predict its function.**

Classical Genetic Approach: Random Mutagenesis

Before **gene cloning technology, most genes were identified by observing what went wrong when the gene was mutated. This approach is most effective in organisms that reproduce quickly and can be genetically manipulated, like bacteria, yeast, and fruit flies. Mutations can be introduced using chemicals or radiation,** and then large numbers of organisms are screened to find mutants with interesting defects.

In humans, studying gene function is more challenging, but we can still learn about human genes through studies in model organisms.

The Gene Expression Omnibus (GEO) is a public database hosted by **NCBI,** designed to store **high-throughput genomic data,** such as **gene expression profiles from microarray and sequencing technologies.** It allows researchers to **submit and access data related to functional genomics. GEO supports various data types like mRNA expression, genomic DNA, and proteomic data.** Users can query data based on keywords, organisms, or other parameters and use tools to analyze and visualize gene expression patterns. **GEO also offers tools like GEO2R for differential gene expression analysis, ClueGO for functional network analysis, and FunRich for visualizing enrichment and network data. As of 2018, it contained over 2.4 million samples** and a vast number of gene expression profiles for numerous organisms. Researchers use GEO for data analysis and to explore experiments related to diseases, including cancer.

SPECIALIZED DATABASES

https://www.ncbi.nlm.nih.gov/geo/geo2r/
Expression data for HT29 cells treated with 5-aza-deoxy-cytidine [RNA-Seq]

Accession number: GSE41586
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE41586

The study compared RNA-Seq and microarray platforms using **HT-29 colon cancer cells treated with three concentrations of 5-aza-deoxy-cytidine.** Results showed strong correlation but some biases between the platforms. DESeq performed best for detecting differentially expressed genes, with high consistency between RNA-Seq and microarray methods.

Before Grouping

GEO accession GSE41586 Set Expression data for HT29 cells treated with 5-aza-deoxy-cytidine [RNA-Seq] RNA-seq Series analysis BETA

Samples Define groups Selected 0 out of 9 samples

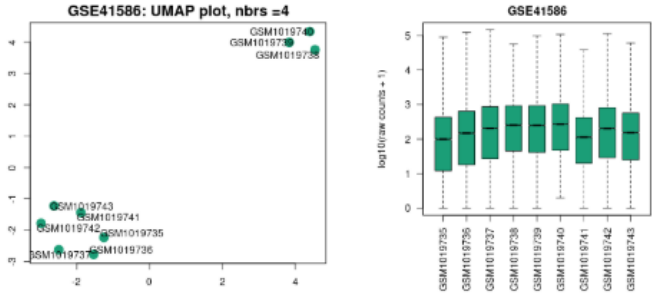
Columns Set

Group	Accession	Title	Source name	Cell line	Cell type	Treatment
-	GSM1019735	HT29 at 0 µM of 5-Aza, biological rep1	HT29 treated with 0 µM of 5-Aza	HT29	colon cancer	control
-	GSM1019736	HT29 at 0 µM of 5-Aza, biological rep2	HT29 treated with 0 µM of 5-Aza	HT29	colon cancer	control
-	GSM1019737	HT29 at 0 µM of 5-Aza, biological rep3	HT29 treated with 0 µM of 5-Aza	HT29	colon cancer	control
-	GSM1019738	HT29 at 5 µM of 5-Aza, biological rep1	HT29 treated with 5 µM of 5-Aza	HT29	colon cancer	5-Aza low
-	GSM1019739	HT29 at 5 µM of 5-Aza, biological rep2	HT29 treated with 5 µM of 5-Aza	HT29	colon cancer	5-Aza low
-	GSM1019740	HT29 at 5 µM of 5-Aza, biological rep3	HT29 treated with 5 µM of 5-Aza	HT29	colon cancer	5-Aza low
-	GSM1019741	HT29 at 10 µM of 5-Aza, biological rep1	HT29 treated with 10 µM of 5-Aza	HT29	colon cancer	5-Aza high
-	GSM1019742	HT29 at 10 µM of 5-Aza, biological rep2	HT29 treated with 10 µM of 5-Aza	HT29	colon cancer	5-Aza high
-	GSM1019743	HT29 at 10 µM of 5-Aza, biological rep3	HT29 treated with 10 µM of 5-Aza	HT29	colon cancer	5-Aza high

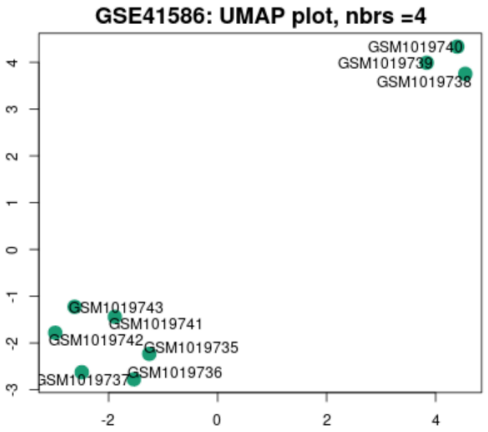
Analyze with GEO2R

Reanalyze if you changed any options.

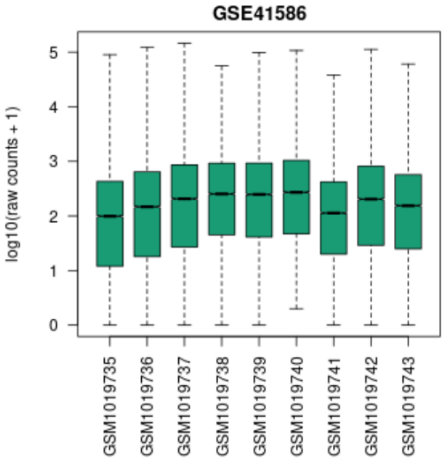
Visualization



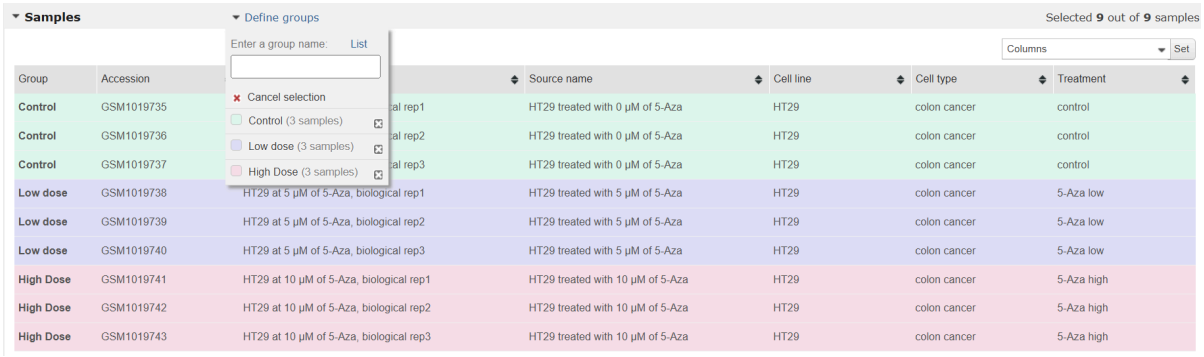
UMAP plot



Boxplot

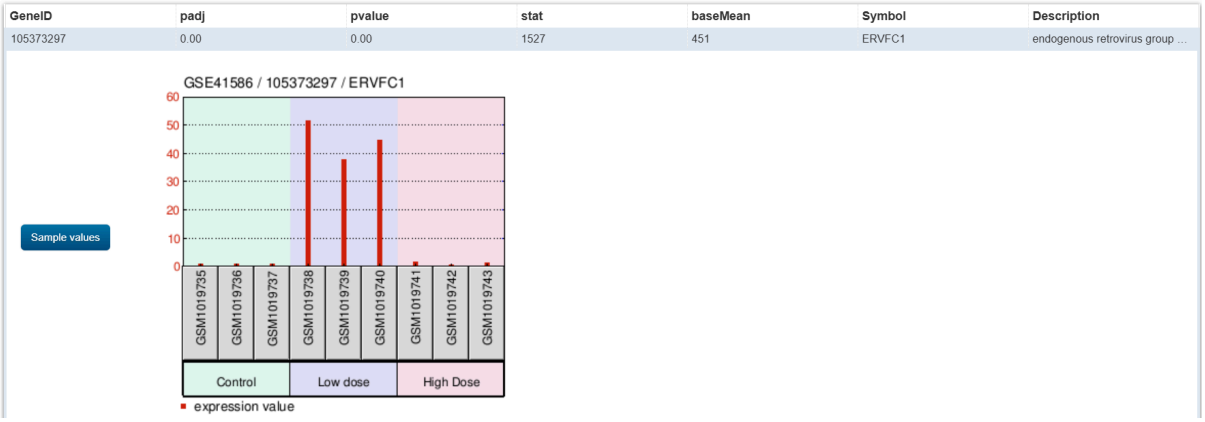


Analyze with GEO2R > Grouped under 3 concentrations

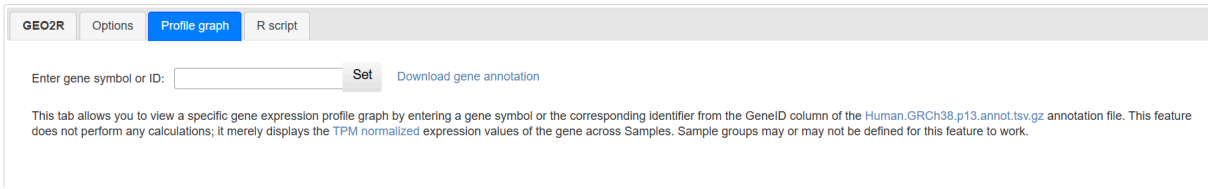


We can do this “define” upto 10 groups.

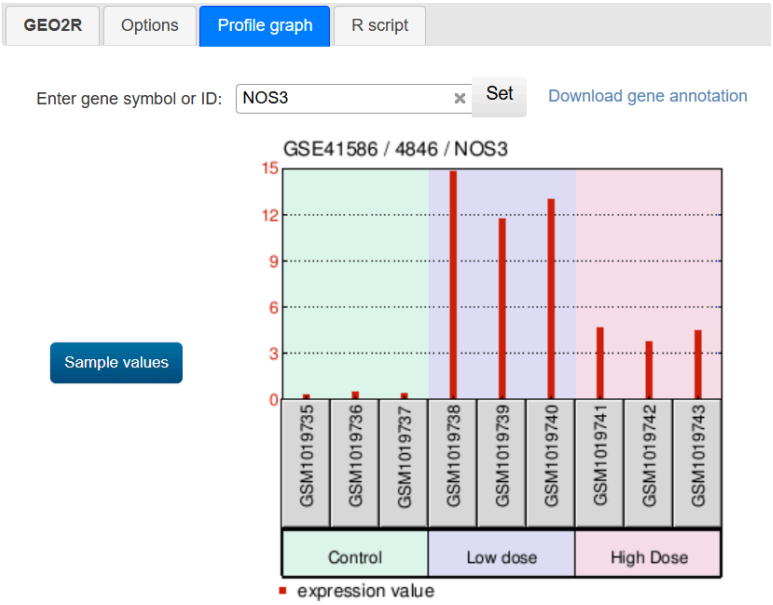
Analyze> We get a list of genes with there expression levels in each group.

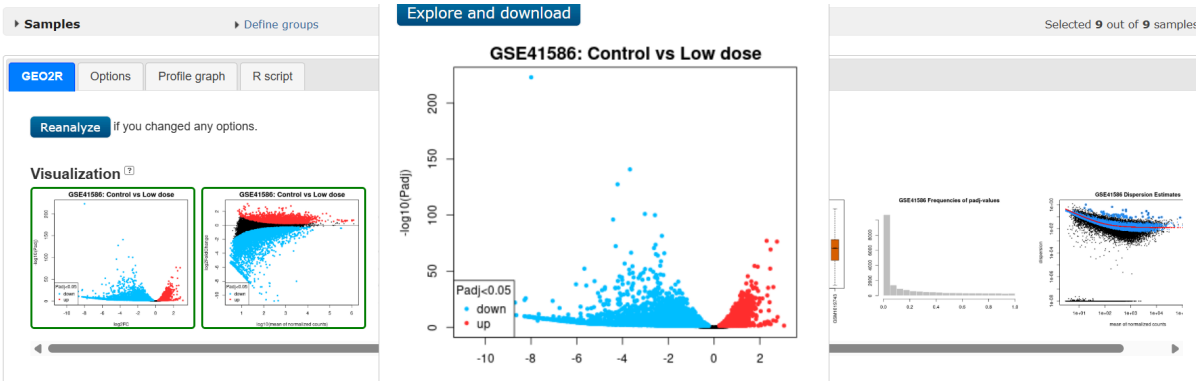


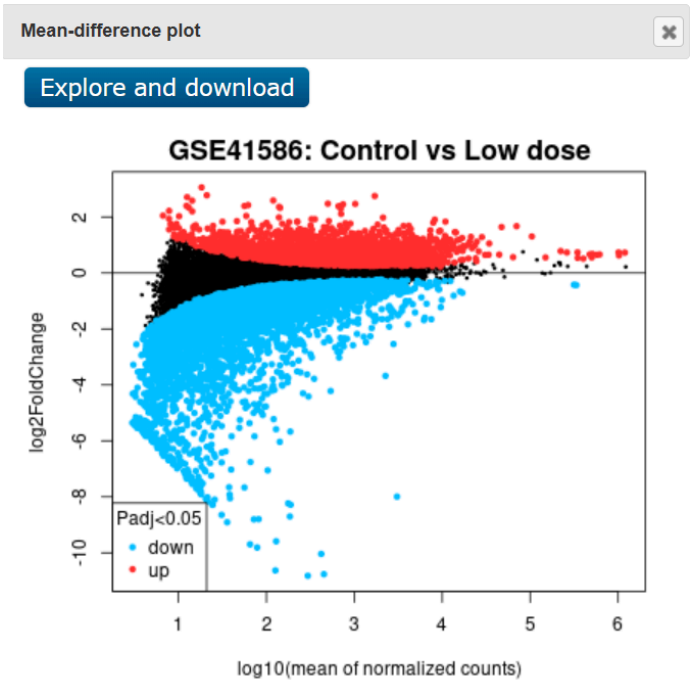
Red bars represent the expression in each group.



If I am interested in: **Nitric Oxide Synthase 3 (NOS3)**


GEO2R
Options
Profile graph
R script
Enter gene symbol or ID:
NOS3
Set
Download gene annotation
Sample values
GSE41586 / 4846 / NOS3
15
12
9
6
3
0
GSM1019735
GSM1019736
GSM1019737
GSM1019738
GSM1019739
GSM1019740
GSM1019741
GSM1019742
GSM1019743
Control
Low dose
High Dose
expression value

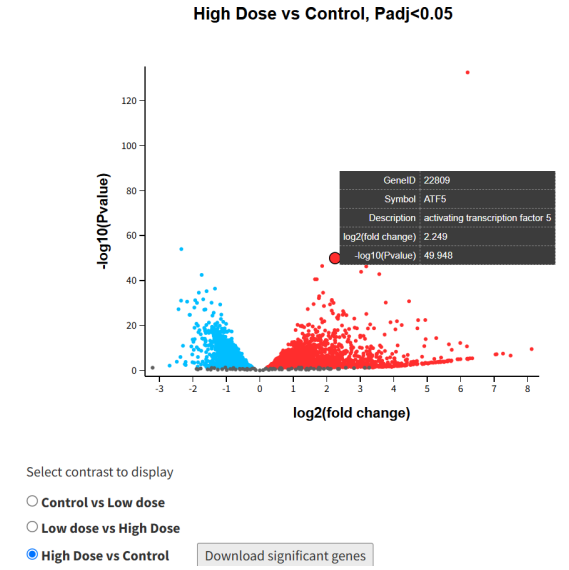

Samples
Define groups
GEO2R
Options
Profile graph
R script
Reanalyze
if you changed any options.
Visualization
Explore and download
GSE41586: Control vs Low dose
-log10(Padj)
Padj<0.05
down
up
Selected 9 out of 9 samples



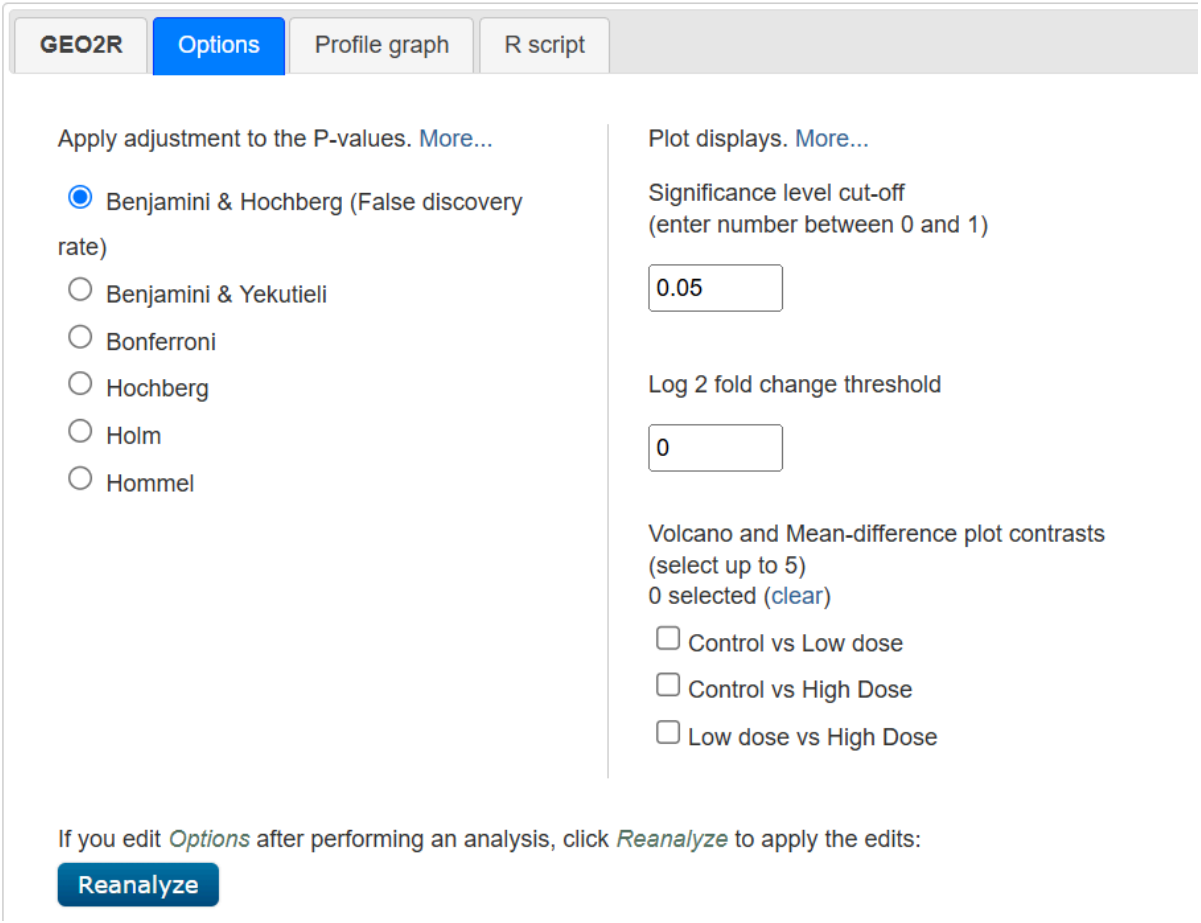
Red: Increased Expression

Blue: Decreased Gene Expression

Grey is threshold

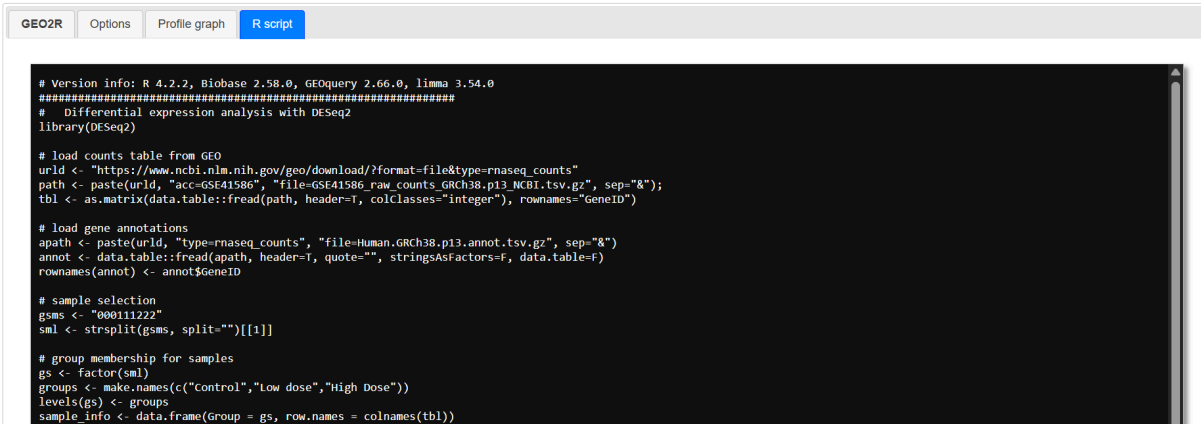


Explore and download: Click on each point to know the gene information, we can also download the significant genes in High/Low dose.



Options tab is available to change the parameters and reanalyse.

What ever changes we made in the analysis are automatically scripted and available to download:



https://blog.addgene.org/plasmids-101-five-popular-model-organisms

Genetic Screens and Identifying Mutants

Once a collection of mutants is created, researchers perform **genetic screens** to identify those with **interesting or abnormal traits.** These screens can detect simple or **complex phenotypes, such as a metabolic deficiency or behavior changes.**

Some mutants, like temperature-sensitive ones, are used to **study essential processes like DNA replication or cell cycle regulation**. These mutants function normally at one temperature but stop working at a higher or lower temperature, making it easier to study gene functions under controlled conditions.

Identifying Gene Locations

Once genes involved in a process are identified, scientists determine the order in which they work. **This helps to understand complex processes like metabolism or development**. By creating double mutants, researchers can figure out which genes act earlier or later in a process.

Gene Mapping with Linkage Analysis

To locate genes responsible for a mutant phenotype, **linkage analysis** is used. This method maps the gene's position by examining how genes close together on a chromosome tend to be inherited together. By looking at the frequency of recombination between genes, researchers can estimate their physical distance on a chromosome.

Linkage analysis also helps find genes responsible for human diseases by identifying markers that are closely linked to the disease gene. For example, the genes for cystic fibrosis and Huntington's disease were discovered using this technique.

https://flybase.org/

https://wiki.flybase.org/wiki/FlyBase:Downloads_Overview#Genes_data_.28Chado_XML.29

https://wiki.flybase.org/wiki/FlyBase:Tools_Overview

FlyBase has been a key resource for Drosophila research for over 30 years. I**t offers a wide range of curated data, including genetic information, phenotypes, and reagents for *Drosophila melanogaster*.**

FlyBase provides various **tools for data search and access**, including updates to its knowledge base and unique features developed through teamwork and individual initiatives. These tools help users find specific data and make the information more accessible.

Additionally, **FlyBase supports the Drosophila community with resources like an external resources wiki, a Fast-Track Your Paper tool for quicker data inclusion**, user support documentation, and **tutorials via Twitter.**

One key tool is QuickSearch, which helps users search data across different categories. It has evolved from a simple search box to a more sophisticated tool with various search options, including full-text searches, publications, and gene expression data. FlyBase also introduced a new data class to identify genetic tools with specific molecular properties.

Gene Groups and Pathways

FlyBase curates gene sets related by molecular function or biological role. These groups are organized into hierarchical reports for easy navigation. For example, gene sets like "OXIDOREDUCTASES" are further divided into smaller sets to provide detailed information.

Pathway reports, such as those for signaling pathways, also include visual representations and updated experimental data.

Human Disease Models (HDMs)
FlyBase has over 1,100 Drosophila models for human diseases. The HDM report serves as a **hub for disease information, linking to medical databases, publications, and gene reports.** It includes details on disease-related alleles, genetic interactions, and disease-implicated variants. This feature helps researchers study human diseases using Drosophila models, even for diseases not fully categorized in medical databases.

Experimental Tools
FlyBase has introduced a data class for experimental tools, which include genetic constructs like epitope tags, fluorescent proteins, and genome engineering tools. **Over 39,000 alleles representing these tools are cataloged, helping researchers find the right tool for their experiments. The tools are described using controlled vocabulary to make them easy to search and identify.**

2) Mouse Genome Informatics (https://www.informatics.jax.org/)

https://www.informatics.jax.org/marker/MGI:1349215

3) https://www.arabidopsis.org/