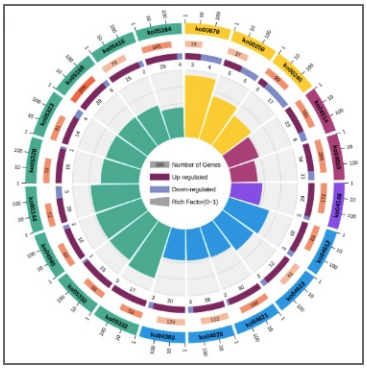


DEC 01, 2023

OPEN ACCESS

DOI:
dx.doi.org/10.17504/protocols.io.eq2lyjqpqlx9/v1

Protocol Citation: Hussain Zubair 2023. Enrichment Analysis: Fundamentals and Visualization of Enrichment Analysis in Genomic Data Interpretation. **protocols.io** https://dx.doi.org/10.17504/protocols.io.eq2lyjqpqlx9/v1

Protocol status: Working
We use this protocol and it's working

Created: Dec 01, 2023

Enrichment Analysis: Fundamentals and Visualization of Enrichment Analysis in Genomic Data Interpretation

Hussain Zubair[1]

[1]Zhengzhou University

Hussain Zubair

DISCLAIMER

The content provided herein is for educational and informational purposes only and is not intended for use as a professional or commercial guide. While efforts have been made to ensure accuracy and completeness, the author and publisher disclaim any liability for any errors, omissions, or discrepancies that may be present. Furthermore, the content does not represent any endorsement of specific methodologies or software, nor does it constitute a definitive treatment of the topic of enrichment analysis. Any use of the information provided in this publication is at the reader's own risk, and it is recommended that users validate the information herein with authoritative sources before use in any academic publication or research. This material is published without warranty, and the author and publisher shall not be liable for any damages incurred as a result of its use.

ABSTRACT

This article encapsulates a dialogue on enrichment analysis and its applications in genomics, specifically focusing on the analytical and visualization techniques for interpreting high-throughput experimental data. Key concepts such as Gene Set Enrichment Analysis, hypergeometric distribution, and Fisher's Exact Test were explicated, along with their relevance in determining the significance of gene representation in biological pathways. Additionally, the discussion traversed the realm of data visualization, highlighting the creation of basic graphical representations using R programming, without pre-built software packages. The intent was to illuminate the principles of analysis and visualization in a foundational manner, setting the stage for future discussions on advanced graphical techniques in enrichment analysis.

Last Modified: Dec 01, 2023

PROTOCOL integer ID: 91671

Content of this article

1 This article will exemplify the application of KEGG enrichment analysis to elucidate:
1) The conceptual framework of enrichment analysis;
2) The underlying principles governing enrichment analysis; and
3) The methodologies for visual representation of enrichment analysis findings.

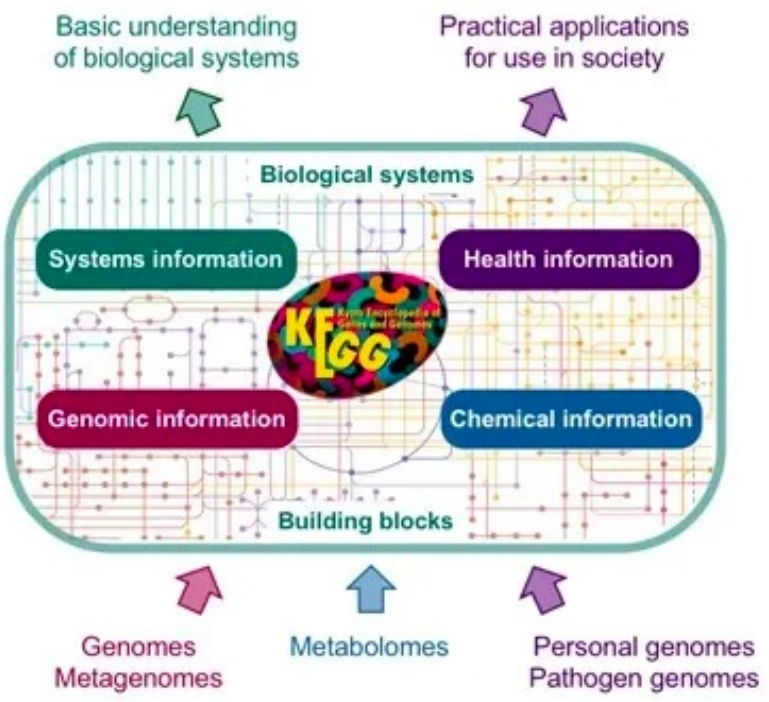


Fig 1: Illustration of KEGG applications

Introduction

2 **Enrichment Analysis** represents a sophisticated approach for interrogating high-throughput experimental datasets. It is predominantly employed to elucidate the extent of enrichment of gene sets or biological entities under specific experimental conditions, with a focus on functions, pathways, or distinct biological processes. The primary objective of enrichment analysis is to assess whether genes or entities identified in an experimental setup are disproportionately represented in certain functions or pathways. This assessment facilitates the inference of significant enrichment of these functions or pathways in the context of the experimental conditions.

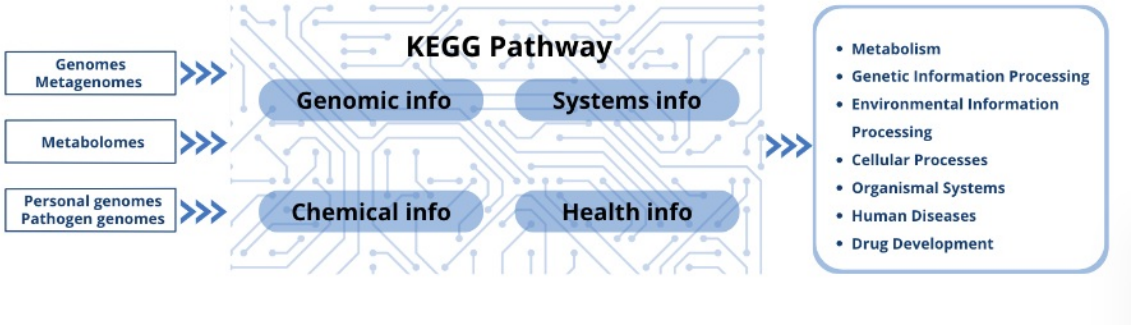


Fig 2: KEGG flowchart

Principal Modalities of Enrichment Analysis

3

1. **Gene Set Enrichment Analysis (GSEA):** This technique involves the categorization of genes based on specific attributes, such as expression levels or regulatory statuses. Subsequently, it assesses whether genes within a predefined set are disproportionately represented in certain segments of the categorized list. GSEA represents a categorization-based approach to enrichment analysis.
2. **Pathway Enrichment Analysis:** This modality focuses on determining whether genes/metabolites within a set are overrepresented in specific biological pathways. Resources such as the KEGG and Reactome databases, which house comprehensive biological pathway information, are instrumental in facilitating this analysis.
3. **Gene Ontology Enrichment Analysis:** This analysis involves scrutinizing whether genes within a set show overrepresentation across various levels of the Gene Ontology (GO). Such an analysis aids in comprehending the roles of these genes in cellular composition, biological processes, and molecular functions.
4. **Disease Enrichment Analysis:** This approach is applied to analyze whether gene sets correlated with specific diseases or disease categories show overrepresentation in the experimental data.
5. **Compound Enrichment Analysis:** This involves analyzing the overrepresentation of genes or gene sets associated with specific compounds within the experimental data.

Output Metrics in Enrichment Analysis

4

Typically, the outputs of enrichment analysis include the Enrichment Score, p-value, adjusted p-value (to account for multiple testing corrections), and enrichment plots. These metrics are pivotal in aiding researchers to identify and characterize biological attributes that are significantly enriched under specific experimental conditions.

Gene Ontology (GO) Framework and Ontological Classification...

5

Gene Ontology (GO) is an established framework that offers structured vocabularies for annotating gene

and gene product attributes across species. This ontology encompasses a comprehensive set of concepts and categories that detail gene functions and the interrelations among these concepts. Functionally, GO organizes these gene functionalities into three primary domains:

- **Molecular Function (MF):** This pertains to the specific molecular activities executed by gene products, such as binding or catalytic activities.
- **Cellular Component (CC):** This domain refers to the precise cellular locales where gene products exert their effects, encompassing various cellular structures and locations.
- **Biological Process (BP):** This encompasses broader biological pathways and processes that are orchestrated by the collective activities of multiple gene products.

The Kyoto Encyclopedia of Genes and Genomes (KEGG), in contrast, is a resource comprising an array of meticulously curated pathway maps. These maps symbolize the networks of molecular interactions and reactions, encapsulating a diverse array of biochemical and cellular processes. The pathways delineated within KEGG are broadly segmented into seven overarching categories, reflecting various aspects of biological systems and processes:

- **Metabolism:** Encompassing the array of chemical reactions and pathways responsible for maintaining cellular function.
- **Genetic Information Processing:** Covering mechanisms related to the transmission and expression of genetic information.
- **Environmental Information Processing:** Involving pathways and networks that enable cellular response and adaptation to environmental changes.
- **Cellular Processes:** Addressing various cellular functions and structural dynamics.
- **Organismal Systems:** Concerning the complex biological processes and systems operating at the organism level.
- **Human Diseases:** Focusing on the pathways and mechanisms underlying human diseases.
- **Drug Development:** Targeting the pathways relevant to pharmacology and therapeutic intervention development.

Preparation of Input Data for Enrichment Analysis

6 Focusing on the KEGG gene pathway enrichment analysis (applicable analogously to metabolites), consider a scenario where differentially expressed genes have been identified between control and treatment groups within a specific experimental sample. Prior to conducting enrichment analysis, it is imperative to compile a comprehensive background dataset. Essentially, this entails a KEGG database file, encompassing an exhaustive compilation of pathway data within KEGG, along with detailed information about the genes incorporated in these pathways.

Methodology for Background File Compilation:

The composition of the background file is intricately linked to the species from which the experimental sample is derived. KEGG maintains an extensive repository of reference pathway gene annotations for a variety of common species. The spectrum of species supported by KEGG, along with their respective gene annotations, can be accessed at KEGG Organism List, which provides a crucial resource for accurate enrichment analysis. (https://www.genome.jp/kegg/catalog/org_list.html)

KEGG Organisms: Complete Genomes

Eukaryotes: 977 Bacteria: 8119 Archaea: 423

[Genomes | Species | Genus | Meta]

Eukaryotes

Category			Organisms			Source
			hsa	KGB	Homo sapiens (human)	RefSeq
			ptr	KGB	Pan troglodytes (chimpanzee)	RefSeq
			pps	KGB	Pan paniscus (bonobo)	RefSeq
			ggo	KGB	Gorilla gorilla gorilla (western lowland gorilla)	RefSeq
			pon	KGB	Pongo abelii (Sumatran orangutan)	RefSeq
			nle	KGB	Nomascus leucogenys (northern white-cheeked gibbon)	RefSeq
			hmh		Hylobates moloch (silvery gibbon)	RefSeq
			mcc	KGB	Macaca mulatta (rhesus monkey)	RefSeq
			mcf	KGB	Macaca fascicularis (crab-eating macaque)	RefSeq
			mthb	KGB	Macaca thibetana thibetana (Pere David's macaque)	RefSeq
			mni		Macaca nemestrina (pig-tailed macaque)	RefSeq
			csab	KGB	Chlorocebus sabaeus (green monkey)	RefSeq
			caty		Cercocebus atys (sooty mangabey)	RefSeq
			panu	KGB	Papio anubis (olive baboon)	RefSeq
			tge	KGB	Theropithecus gelada (gelada)	RefSeq
			mleu		Mandrillus leucophaeus (drill)	RefSeq
			rro	KGB	Rhinopithecus roxellana (golden snub-nosed monkey)	RefSeq
			rbb		Rhinopithecus bieti (black snub-nosed monkey)	RefSeq
			tfn		Trachypithecus francoisi (Francois's langur)	RefSeq
			pteh	KGB	Piliocolobus tephrosceles (Ugandan red Colobus)	RefSeq
			cang		Colobus angolensis palliatus (Angola colobus)	RefSeq
			cjc	KGB	Callithrix jacchus (white-tufted-ear marmoset)	RefSeq
			sbq		Saimiri boliviensis boliviensis (Bolivian squirrel monkey)	RefSeq
			cimi		Cebus imitator (Panamanian white-faced capuchin)	RefSeq
			csyr		Carlito syrichta (Philippine tarsier)	RefSeq
			mmur	KGB	Microcebus murinus (gray mouse lemur)	RefSeq
			lcat	KGB	Lemur catta (Ring-tailed lemur)	RefSeq
			pcoq		Propithecus coquereli (Coquerel's sifaka)	RefSeq

KEGG Organisms: Complete Genomes

The JSON files pertinent to the species of interest can be downloaded and subsequently parsed to retrieve the necessary background data. An alternative approach involves employing web scraping techniques to systematically collect comprehensive data on all genes and metabolites present within each pathway. In instances involving species that lack reference pathway gene annotations, or in the case of entirely novel species, the prediction of gene function annotations necessitates comparative analysis with existing gene sequences. This process can be facilitated through specialized websites or bioinformatics software. However, this aspect falls outside the primary scope of this discussion and hence will not be explored in detail.

With these steps, you are now equipped with the essential dataset for enrichment analysis: a set of differentially expressed genes pertinent to your study and the relevant KEGG background file tailored to your specific species.

Methodology and Theoretical Framework for Enrichment Anal...

7 1. Consider a scenario with 11 genes identified as differentially expressed (n),

2. Within this subset, 3 genes are associated with Pathway A (k),
3. The comprehensive KEGG background dataset comprises 8000 genes (M),
4. Of these, Pathway A encompasses 120 genes (N).
5. The formula to compute the Fold Enrichment, also known as the Enrichment Score, is articulated as:

$$\text{Enrichment Score} = \left(\frac{k}{n}\right) / \left(\frac{N}{M}\right)$$

Applying this formula yields:

$$\text{Enrichment Score} = \left(\frac{3}{11}\right) / \left(\frac{120}{8000}\right) = 18.18181818$$

To evaluate the statistical significance of the observed enrichment, one computes the P-value utilizing statistical tests such as the hypergeometric distribution test and Fisher's exact test, which are standard in the field.

The hypergeometric distribution test employs the following Probability Mass Function (PMF):

$$p(k; M, n, N) = \frac{\binom{N}{k}\binom{M-N}{n-k}}{\binom{M}{n}}$$

This is calculated for k within the range of $\max(0, N - M + n)$ to $\min(n, N)$. The binomial coefficients within the formula are defined by the expression:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Here, $n!$ represents the factorial of n, and $\binom{n}{k}$ quantifies the total number of distinct combinations of k elements that can be selected from a larger set of n elements.

- N represents the total number of items in the population (the total number of all genes),
- K is the number of successful items in the population (the number of genes associated with a specific function or pathway),
- n is the number of samples drawn from the population (the size of the gene set in the experiment),
- k is the number of successful items in the sample (the number of genes associated with a specific function or pathway in the experiment).

The probability P value for gene enrichment is the probability $P(X \geq k)$ of drawing 11 genes from a total of 8000, with at least three belonging to Pathway A, which is equivalent to $1 - P(X \leq k - 1)$.

Use Python's scipy to calculate the hypergeometric distribution:

pythonCopy code:

```
from scipy.stats import hypergeom

# Calculate the P-value for k or more successes
p_value = 1 - hypergeom.cdf(3 - 1, 8000, 11, 120)

# Alternatively, calculate the survival function for k or more successes
p_value = hypergeom.sf(3 - 1, 8000, 11, 120)

# Output the P-valueprint(f'p-value: {p_value}')
```

Methods:

	Method	Parameters	Description
	rvs	M, n, N, loc=0, size=1, random_state=None	Generates random variates from a hypergeometric distribution.
	pmf	k, M, n, N, loc=0	Computes the probability mass function for a hypergeometric distribution.
	logpmf	k, M, n, N, loc=0	Computes the log of the probability mass function for a hypergeometric distribution.
	cdf	k, M, n, N, loc=0	Calculates the cumulative distribution function for a hypergeometric distribution.
	logcdf	k, M, n, N, loc=0	Computes the log of the cumulative distribution function for a hypergeometric distribution.
	sf	k, M, n, N, loc=0	Computes the survival function (also defined as 1 - cdf, but the sf is sometimes more accurate).

Fisher's Exact Test:

Within the SciPy library, the execution of Fisher's Exact Test is facilitated by the scipy.stats.fisher_exact function. This function is adept at determining the statistical association present between two categorical variables.

pythonCopy code:

```
from scipy.stats import fisher_exact

def calculate_fisher_exact_p_value(contingency_table):
    """
    Calculate the p-value from Fisher's Exact Test.

    Parameters:
    contingency_table: A 2x2 contingency table in the form of a list of
lists or a NumPy array.

    Returns:
    p_value: The p-value calculated from Fisher's Exact Test.
    """
    odds_ratio, p_value = fisher_exact(contingency_table,
alternative='greater')
    return p_value

# Example usage
genes_in_pathway = 120  # Total number of genes in Pathway A
genes_not_in_pathway = 8000 - 120  # Total number of genes not in Pathway
A
observed_genes_in_pathway = 3  # Number of observed genes in Pathway A
observed_genes_not_in_pathway = 11 - 3  # Number of observed genes not in
Pathway A

contingency_table = [
    [observed_genes_in_pathway, genes_in_pathway -
observed_genes_in_pathway],
    [observed_genes_not_in_pathway, genes_not_in_pathway -
observed_genes_not_in_pathway]
]

p_value = calculate_fisher_exact_p_value(contingency_table)

print(f"P-value: {p_value}")
```

In the given instance, the parameter **"alternative='greater"** within the **"fisher_exact"** function delineates the hypothesis testing direction. This parameterization indicates a focused interest in scenarios where the observed gene count within the pathway exceeds the anticipated count. Consequently, the derived p-value quantifies the likelihood that the observed gene frequency within the specified pathway surpasses the expected frequency. A diminutive p-value signifies that the observed gene count is of statistical significance, potentially implying enrichment of the pathway in the context of the experimental conditions.

Visualization of Enrichment Analysis Outcomes

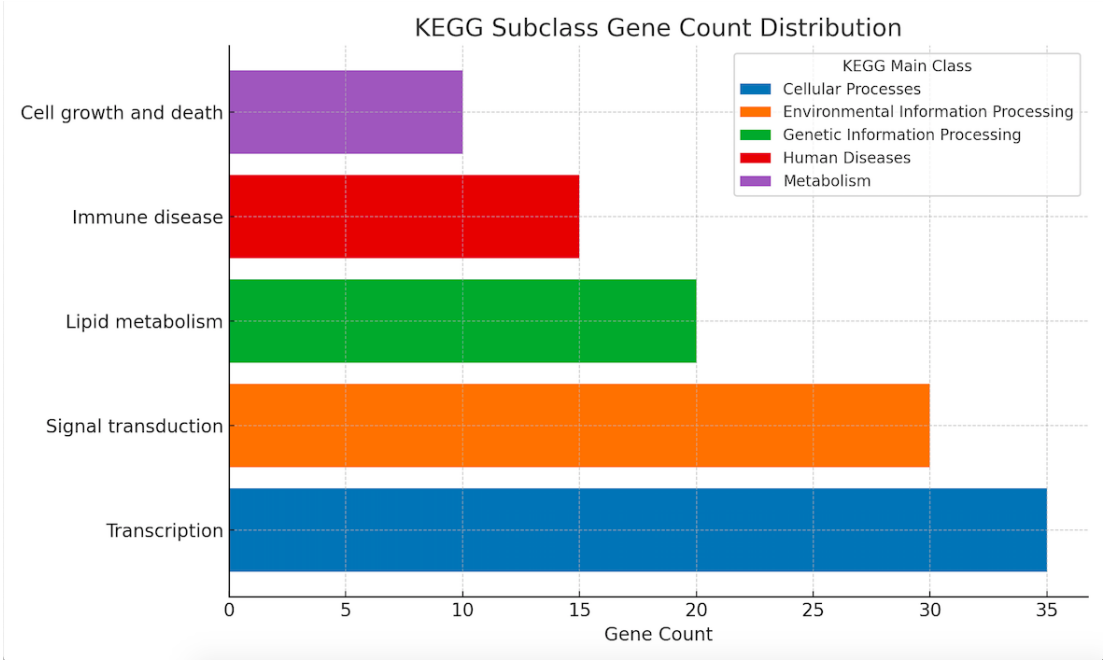
8 A plethora of graphical representations are employed to elucidate the outcomes of enrichment analyses. Notably, bar graphs, bubble plots, enrichment network diagrams, and chord diagrams are among the frequently utilized visual formats.

In the ensuing discourse, we will fabricate a dataset, ostensibly representing results from an enrichment analysis. The primary objective is to illustrate the plotting techniques, with R being the chosen statistical programming environment for this illustrative exposition.
Here's the rewritten version of the code for clarity:

rCopy code:

```
library(ggplot2)# Constructing a mock dataset
data <- data.frame(
  KEGGMainClass = c('Metabolism',
                    'Genetic Information Processing',
                    'Environmental Information Processing',
                    'Cellular Processes',
                    'Human Diseases'),
  EnrichmentScore = c(12, 15, 30, 24, 20),
  GeneCount = c(15, 40, 30, 18, 20),
  KEGGSubClass = c('Lipid metabolism',
                   'Transcription',
                   'Signal transduction',
                   'Cell growth and death',
                   'Immune disease'),
  KEGGTerm = c('Fatty acid biosynthesis',
               'RNA polymerase',
               'MAPK signaling pathway',
               'Cell cycle',
               'Inflammatory bowel disease'),
  pValue = c(0.0049, 0.001, 0.05, 0.02, 0.1))
```

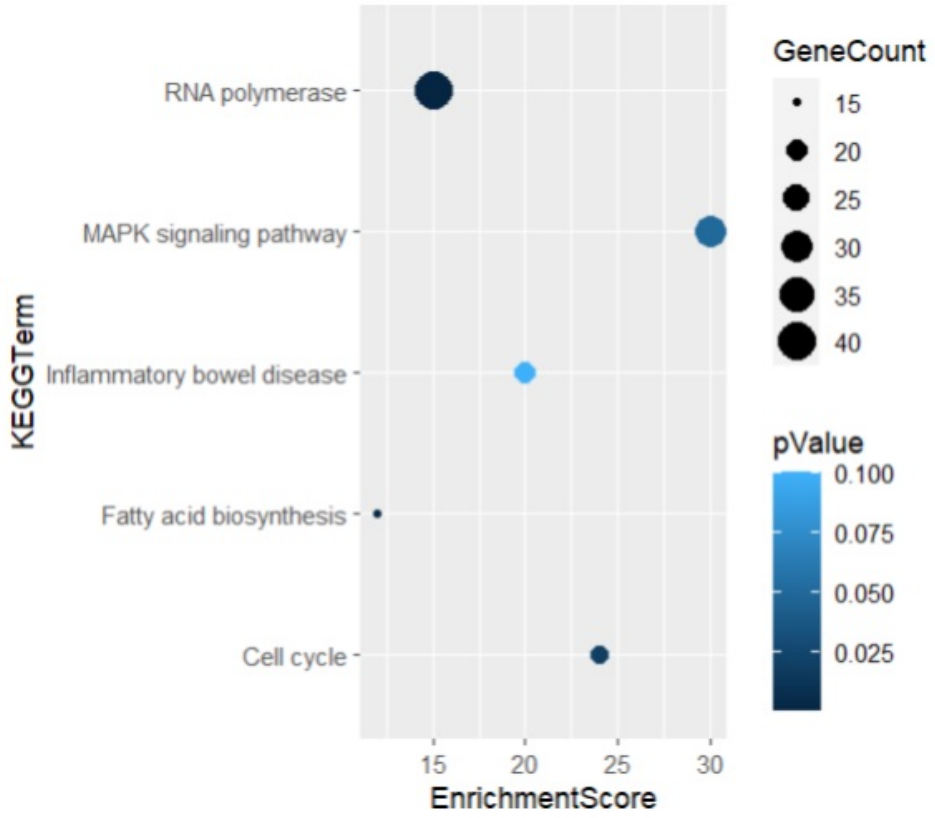
Draw a Bar Plot:



rCopy code:

```
# Load the ggplot2 package
library(ggplot2)# Create a bar plot
ggplot(data=data) +
  geom_bar(mapping = aes(x=KEGGSubClass, y=GeneCount,
fill=KEGGMainClass), stat='identity') +
  coord_flip() +
  theme(aspect.ratio = 1/3)
```

Draw a Bubble Plot:



rCopy code:

```
# Load ggplot2 for data visualization
library(ggplot2)# Constructing the bubble plot
ggplot(data=data) +
  geom_point(mapping = aes(x=EnrichmentScore, y=KEGGTerm,
                           size=GeneCount, color=pValue),
stat='identity') +
  theme(aspect.ratio = 2)
```

Note: The illustrations provided are rudimentary sketches, created with basic coding techniques and without the application of specialized enrichment analysis software packages. The objective was to expound upon the foundational concepts of analytical processing and graphical representation. Consequently, the visualizations have not been fine-tuned or enhanced for aesthetic appeal. Those with an interest in more sophisticated graphical representations may anticipate further developments in subsequent publications.