Project
“Integrative database for single-cell RNA transcript tissue expression data and their gene, disease-variant associations”

Group members :- Prachi Sardana,
Matthew Runyan and Xenlei Hu

Database Management Systems
Project
CS5200
Group 30

Research Question ❓

How do variations in tissue-specific transcript expressions influenced by alternative splicing contribute to the susceptibility and progression of specific diseases, as mediated through known gene-disease associations and variant impacts?

Project Aim:

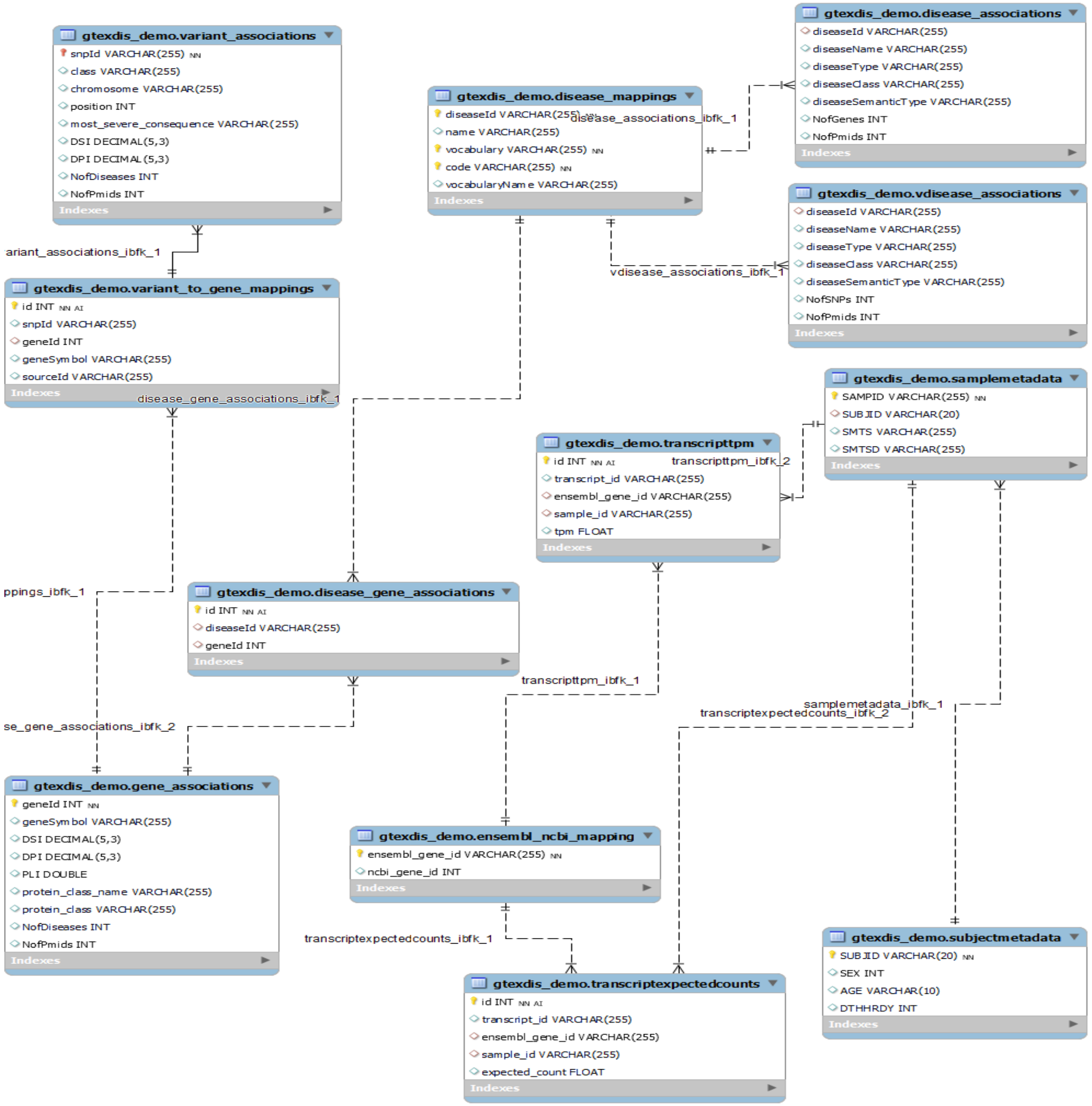
- To integrate transcript-level RNA-seq data from GTEx with gene and variant-disease associations from DisGeNET. This integration efficiently identifies differentially spliced genes associated with specific diseases or variants. Such insights enhance understanding of gene function in disease contexts, aiding biomedical research and therapeutic development.

Abstract

- Alternative splicing is a critical mechanism that diversifies gene expression and impacts cellular functions. This research leverages single-cell RNA sequencing (scRNA-seq) data, focusing on tissue-specific expression profiles from the GTEx portal, integrated with genetic variant and disease information from DisGeNET. We developed a database that enables users to query and analyze at gene and transcript levels. This integration of multidimensional biological data allows for efficient identification of genetic variants and gene-disease associations, providing a valuable tool for researchers in genomics, transcriptomics, and bioinformatics. The project aims to facilitate multi-level transcriptomic analysis to identify key genetic drivers correlated with disease phenotypes. By associating genetic variants with gene expression profiles across various biological contexts, we seek to uncover molecular signatures indicative of disease states and progression. Our database integration aims to advance the field of precision medicine, offering insights that could influence prognosis and therapeutic interventions. This initiative not only enhances our understanding of the genetic bases of diseases but in the future would also support the development of treatment strategies, marking a significant contribution to personalized healthcare.

Citation:- Jeon, J., Kim, K. T., Choi, J., Cheong, K., Ko, J., Choi, G., ... & Lee, Y. H. (2022). Alternative splicing diversifies the transcriptome and proteome of the rice blast fungus during host infection. RNA biology, 19(1), 373-386.

DATABASE DESIGN



gtexdis_demo.variant_associations
snpId VARCHAR(255) NN
class VARCHAR(255)
chromosome VARCHAR(255)
position INT
most_severe_consequence VARCHAR(255)
DSI DECIMAL(5,3)
DPI DECIMAL(5,3)
NofDiseases INT
NofPmids INT
Indexes

gtexdis_demo.disease_mappings
diseaseId VARCHAR(255) NN
name VARCHAR(255)
vocabulary VARCHAR(255) NN
code VARCHAR(255) NN
vocabularyName VARCHAR(255)
Indexes

gtexdis_demo.disease_associations
diseaseId VARCHAR(255)
diseaseName VARCHAR(255)
diseaseType VARCHAR(255)
diseaseClass VARCHAR(255)
diseaseSemanticType VARCHAR(255)
NofGenes INT
NofPmids INT
Indexes

gtexdis_demo.vdisease_associations
diseaseId VARCHAR(255)
diseaseName VARCHAR(255)
diseaseType VARCHAR(255)
diseaseClass VARCHAR(255)
diseaseSemanticType VARCHAR(255)
NofSNPs INT
NofPmids INT
Indexes

gtexdis_demo.variant_to_gene_mappings
id INT NN AI
snpId VARCHAR(255)
geneId INT
geneSymbol VARCHAR(255)
sourceId VARCHAR(255)
Indexes

gtexdis_demo.samplemetadata
SAMPID VARCHAR(255) NN
SUBJID VARCHAR(20)
SMTS VARCHAR(255)
SMTSD VARCHAR(255)
Indexes

gtexdis_demo.transcripttpm
id INT NN AI
transcript_id VARCHAR(255)
ensembl_gene_id VARCHAR(255)
sample_id VARCHAR(255)
tpm FLOAT
Indexes

gtexdis_demo.disease_gene_associations
id INT NN AI
diseaseId VARCHAR(255)
geneId INT
Indexes

gtexdis_demo.gene_associations
geneId INT NN
geneSymbol VARCHAR(255)
DSI DECIMAL(5,3)
DPI DECIMAL(5,3)
PLI DOUBLE
protein_class_name VARCHAR(255)
protein_class VARCHAR(255)
NofDiseases INT
NofPmids INT
Indexes

gtexdis_demo.ensembl_ncbi_mapping
ensembl_gene_id VARCHAR(255) NN
ncbi_gene_id INT
Indexes

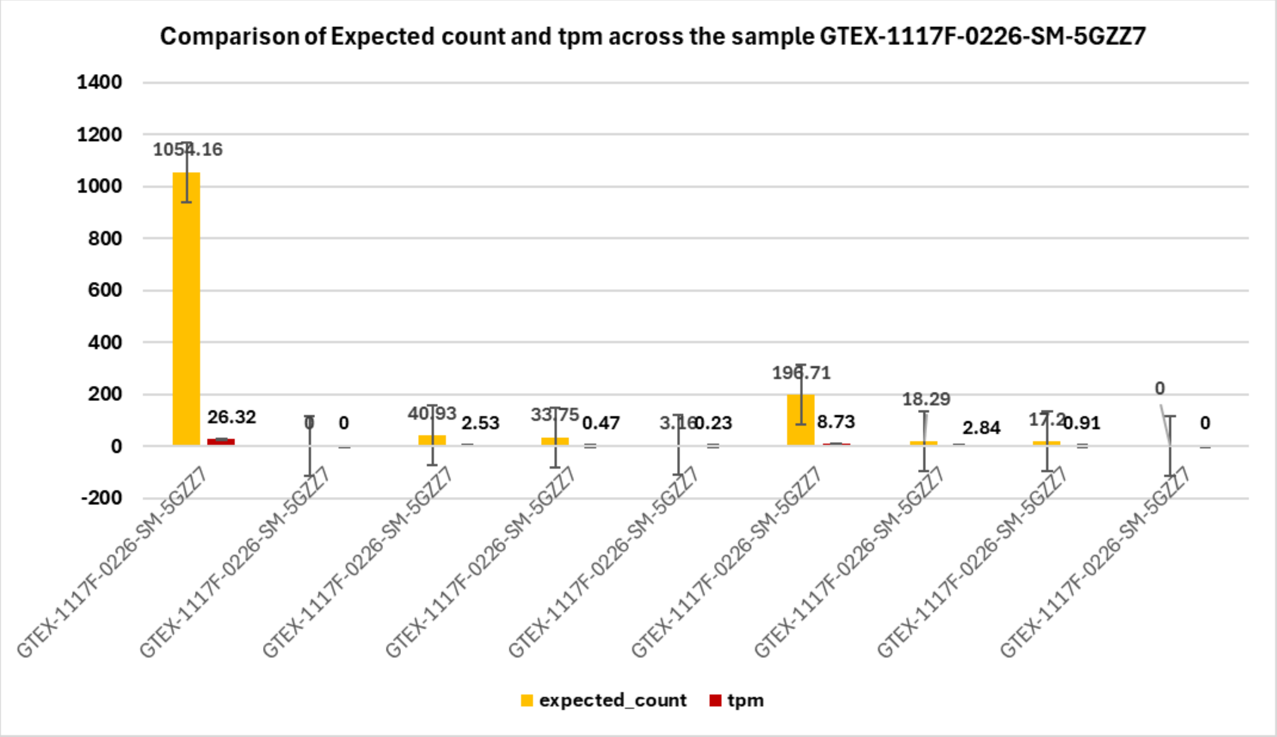
gtexdis_demo.subjectmetadata
SUBJID VARCHAR(20) NN
SEX INT
AGE VARCHAR(10)
DTHHRDY INT
Indexes

gtexdis_demo.transcriptexpectedcounts
id INT NN AI
transcript_id VARCHAR(255)
ensembl_gene_id VARCHAR(255)
sample_id VARCHAR(255)
expected_count FLOAT
Indexes

ariant_associations_ibfk_1
disease_associations_ibfk_1
vdisease_associations_ibfk_1
disease_gene_associations_ibfk_1
transcripttpm_ibfk_2
ppings_ibfk_1
transcripttpm_ibfk_1
samplemetadata_ibfk_1
transcriptexpectedcounts_ibfk_2
se_gene_associations_ibfk_2
transcriptexpectedcounts_ibfk_1

User Cases

Query 1 :- Comparison of Expected count and transcript per million (tpm records) across the sample :



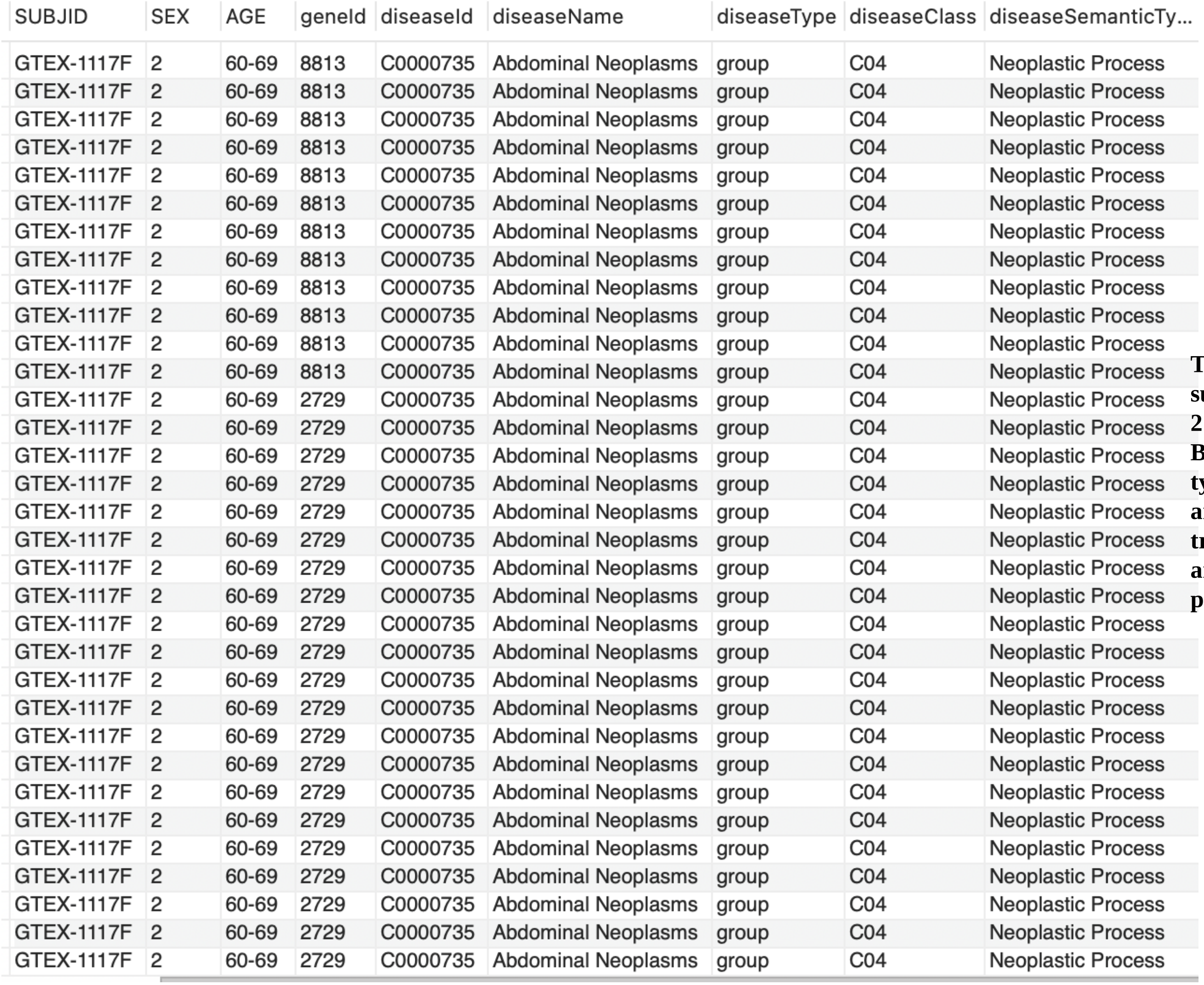
•Variability and Relationship Between Metrics: The graph highlights the alignment between expected counts and TPM across most transcripts, suggesting a general consistency in how these two metrics represent transcript abundance despite being on different scales. However, the presence of outliers indicates that some transcripts might be influenced by factors like post-transcriptional modifications or experimental artifacts, affecting their expected count disproportionately compared to their normalized TPM value.

•Significance of Outliers: One transcript shows an exceptionally high expected count relative to its TPM, hinting at possible post-transcriptional events that affect the transcript's stability or availability for translation, or potentially an artifact in data collection or processing. This discrepancy could have significant implications for biological interpretation and further analysis.

•Challenges in RNA-seq Data Interpretation: Zero values for expected counts in some transcripts suggest technical challenges in RNA-seq such as sequencing depth, capture efficiency, or biological phenomena like tissue-specific expression. This emphasizes the need for robust data processing and normalization techniques to ensure that the results reflect true biological variations rather than technical biases.

User Cases

Query 2 Identified diseases associated with a specific subject, revealing genetic markers linked to diseases and supporting personalized treatment plans.



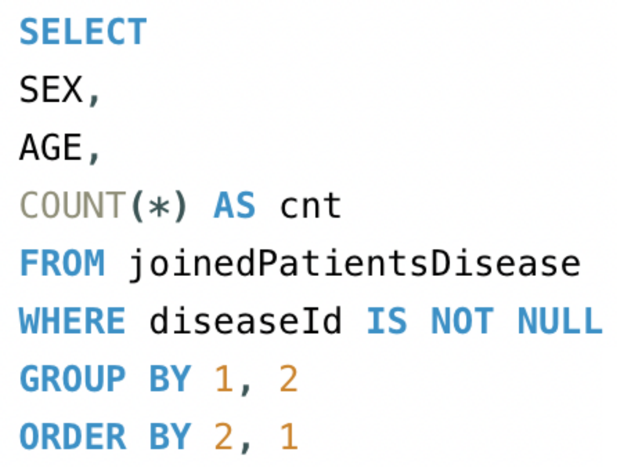
SUBJID	SEX	AGE	geneId	diseaseId	diseaseName	diseaseType	diseaseClass	diseaseSemanticTy...
GTEX-1117F	2	60-69	8813	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	8813	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	8813	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	8813	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	8813	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	8813	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	8813	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	8813	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	8813	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	8813	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	8813	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	8813	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	2729	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	2729	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	2729	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	2729	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	2729	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	2729	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	2729	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	2729	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	2729	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	2729	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	2729	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	2729	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	2729	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	2729	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	2729	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	2729	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	2729	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	2729	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	2729	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	2729	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process
GTEX-1117F	2	60-69	2729	C0000735	Abdominal Neoplasms	group	C04	Neoplastic Process

The subject GTEX-1117F is identified as female (SEX = 2). The age range of the subject is 60-69. The subject is associated with two genes: geneId 8813 and geneId 2729.

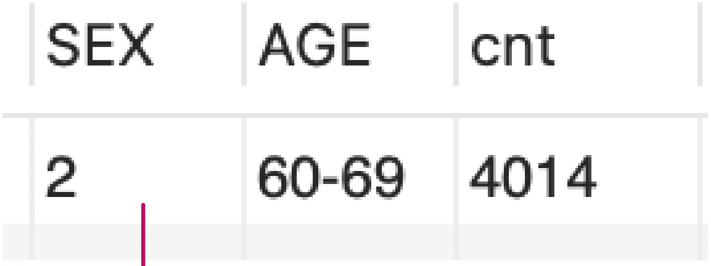
Both genes are linked to the disease Abdominal Neoplasms (diseased C0000735), a type of Neoplastic Process. The output contains multiple rows for the same disease and genes, indicating redundancy in the data. This may be due to multiple transcripts or variations being associated with the same gene-disease pair. This analysis can support research efforts in identifying genetic markers and developing personalized treatment plans.

User Cases

Query 3 analyzed the sampling group, highlighting significant demographics such as females aged 60-69, crucial for targeted healthcare strategies.



```
SELECT
SEX,
AGE,
COUNT(*) AS cnt
FROM joinedPatientsDisease
WHERE diseaseId IS NOT NULL
GROUP BY 1, 2
ORDER BY 2, 1
```

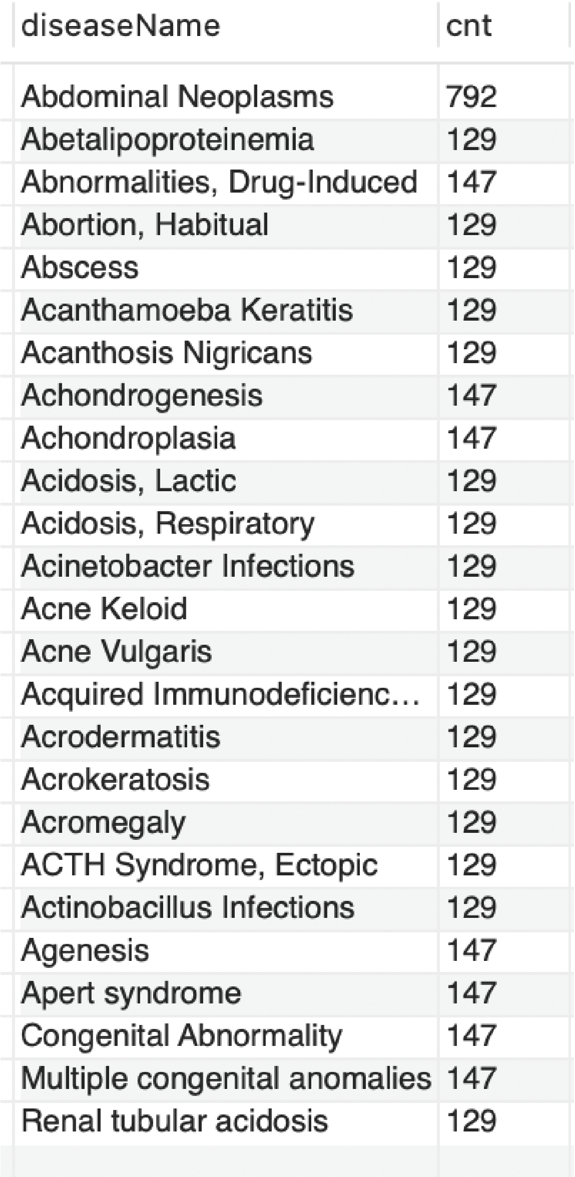


SEX	AGE	cnt
2	60-69	4014

There are 4014 records for females aged 60-69 with a non-null diseaseId. This result highlights that females aged 60-69 are a significant group in the dataset, which could be important for targeted healthcare strategies. Further analysis could reveal which diseases are most prevalent in this demographic. This can help in resource allocation and developing preventive measures for the most common conditions.

User Cases

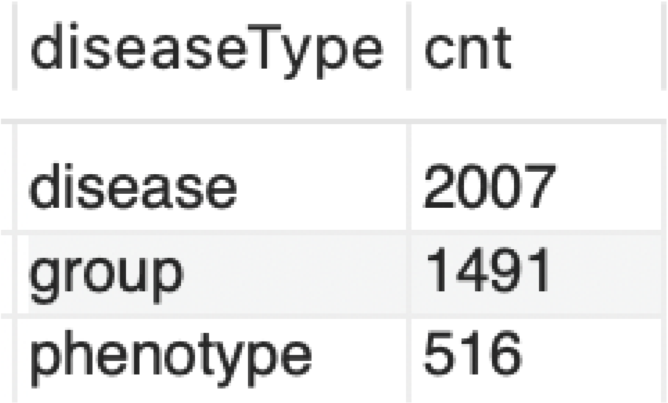
Query 4 Identifying the most prevalent disease distributions by disease name



diseaseName	cnt
Abdominal Neoplasms	792
Abetalipoproteinemia	129
Abnormalities, Drug-Induced	147
Abortion, Habitual	129
Abscess	129
Acanthamoeba Keratitis	129
Acanthosis Nigricans	129
Achondrogenesis	147
Achondroplasia	147
Acidosis, Lactic	129
Acidosis, Respiratory	129
Acinetobacter Infections	129
Acne Keloid	129
Acne Vulgaris	129
Acquired Immunodeficienc…	129
Acrodermatitis	129
Acrokeratosis	129
Acromegaly	129
ACTH Syndrome, Ectopic	129
Actinobacillus Infections	129
Agenesis	147
Apert syndrome	147
Congenital Abnormality	147
Multiple congenital anomalies	147
Renal tubular acidosis	129

- **The disease Abdominal Neoplasms has the highest count with 792 occurrences. This indicates it is the most frequent disease in the dataset. Many diseases, such as Abetalipoproteinemia, Abscess, Acidosis, Lactic, and Acquired Immunodeficiency, have a uniform count of 129 occurrences each.**
- **A few other diseases, such as Abnormalities, Drug-Induced, Achondroplasia, and Acromegaly, have a count of 147. The uniform counts (e.g., 129, 147) across many diseases suggest there might be a pattern in data collection or preprocessing. This uniformity could indicate grouped or batched data entries. The high prevalence of Abdominal Neoplasms suggests a need for further investigation. Researchers could focus on this disease to understand its genetic markers, risk factors, and potential interventions.**

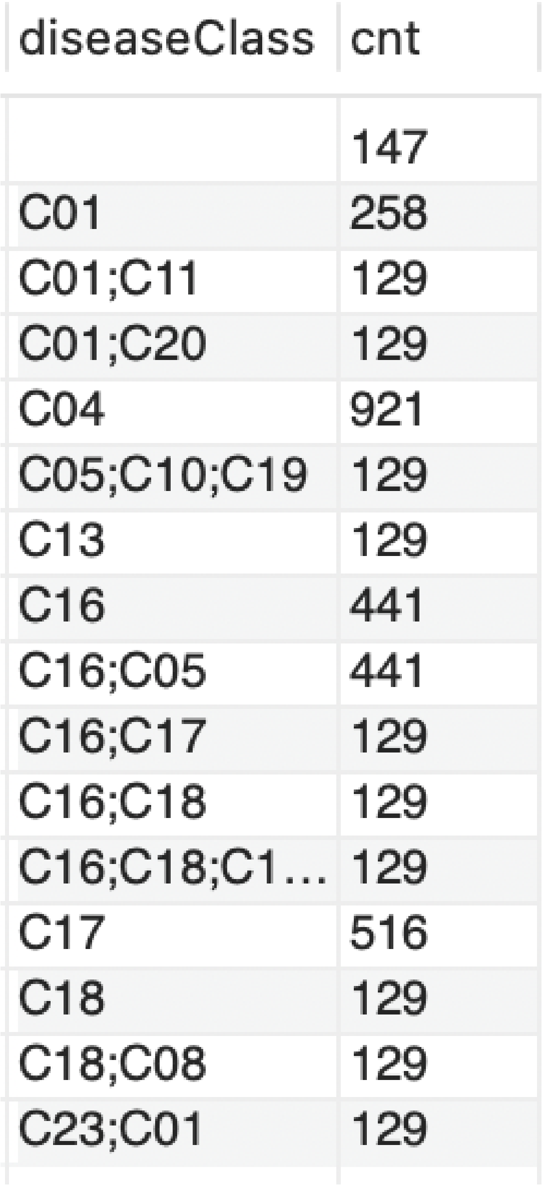
Query 5 analyzed the sampling group, highlighting significant demographics such as females aged 60-69, crucial for targeted healthcare strategies.



diseaseType	cnt
disease	2007
group	1491
phenotype	516

The most common disease type is labeled as disease, with 2007 occurrences. This indicates that the majority of the entries in the dataset are classified under this type.
The group type has 1491 occurrences, making it the second most common.
The phenotype type has 516 occurrences, indicating a smaller but still significant presence. The categorization of diseases into disease, group, and phenotype provides a way to understand how diseases are classified in the dataset. This can be useful for further analysis and understanding of the scope of the data.

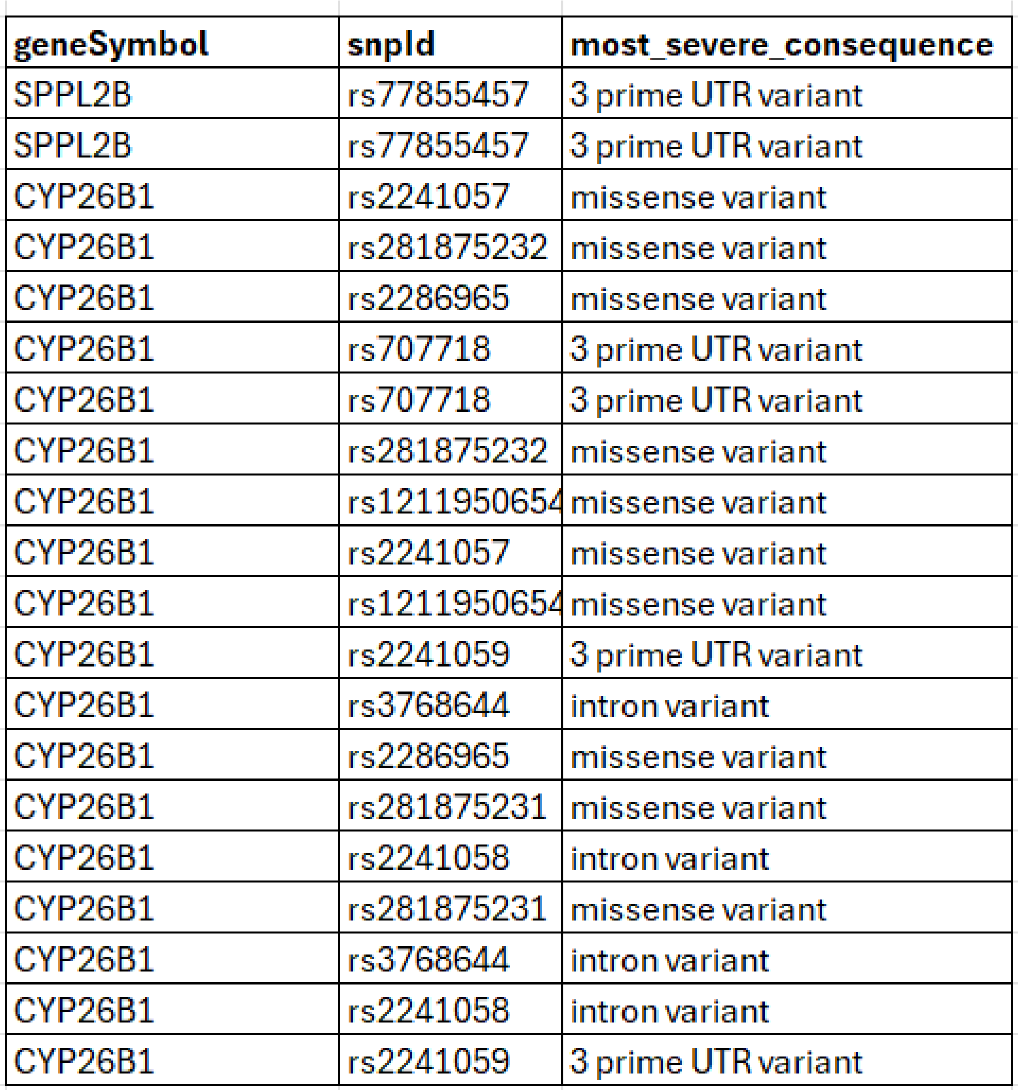
Query 6 examined disease distribution by disease class, uncovering prevalent disease classes and combination classes for further investigation.



diseaseClass	cnt
	147
C01	258
C01;C11	129
C01;C20	129
C04	921
C05;C10;C19	129
C13	129
C16	441
C16;C05	441
C16;C17	129
C16;C18	129
C16;C18;C1…	129
C17	516
C18	129
C18;C08	129
C23;C01	129

The disease class C04 has the highest count with 921 occurrences. This indicates it is the most common disease class in the dataset.
Other frequent disease classes include C16 with 441 occurrences and C17 with 516 occurrences. The high count of C04 indicates a significant prevalence of diseases in this class. Investigating the specific conditions included in this class can provide insights into common health issues within the dataset.
Similarly, the high counts for C16 and C17 suggest these classes are also prevalent and warrant further investigation. The presence of combination classes (e.g., C01; C11) suggests that some diseases may be classified under multiple categories. Understanding the criteria for these combinations can help in refining the classification system. By conducting additional analyses and validating the data against external sources, researchers can gain a comprehensive understanding of disease prevalence and characteristics, informing healthcare strategies and resource allocation.

Query 7 To Find the top 20 records of genes with their Snp id and most severe consequence type.



geneSymbol	snpId	most_severe_consequence
SPPL2B	rs77855457	3 prime UTR variant
SPPL2B	rs77855457	3 prime UTR variant
CYP26B1	rs2241057	missense variant
CYP26B1	rs281875232	missense variant
CYP26B1	rs2286965	missense variant
CYP26B1	rs707718	3 prime UTR variant
CYP26B1	rs707718	3 prime UTR variant
CYP26B1	rs281875232	missense variant
CYP26B1	rs1211950654	missense variant
CYP26B1	rs2241057	missense variant
CYP26B1	rs1211950654	missense variant
CYP26B1	rs2241059	3 prime UTR variant
CYP26B1	rs3768644	intron variant
CYP26B1	rs2286965	missense variant
CYP26B1	rs281875231	missense variant
CYP26B1	rs2241058	intron variant
CYP26B1	rs281875231	missense variant
CYP26B1	rs3768644	intron variant
CYP26B1	rs2241058	intron variant
CYP26B1	rs2241059	3 prime UTR variant

Missense variants are the changes in the DNA that result in the substitution of one amino acid for another in the protein made by a gene. Listed several times associated with the CYP26B1 gene (e.g., rs281875232, rs2286965). These are typically more significant than UTR variants as they directly affect the protein structure and function which is significant from the results.

Summary/Future work

- In this project, we successfully developed and analyzed a comprehensive database integrating tissue-specific expression profiles from the GTEx portal with genetic variant and disease information from DisGeNET.
- Our database enables multi-level transcriptomic analysis, facilitating the identification of genetic variants and gene-disease associations. Here, we summarize the key accomplishments and limitations of our work.
- Implement Data Normalization: Advance data de-duplication and normalization techniques to enhance database integrity.
- Expand Demographic Factors: Include broader demographic variables for richer, personalized gene-disease insights.
- Refine Methodologies: Review and standardize data collection and preprocessing to ensure data quality.
- Conduct External Validation: Validate findings against external datasets for broader applicability and reliability.
- Develop Splicing Variant Tools: Create tools for detailed analysis of alternative splicing and its impacts on diseases.

Acknowledgment

"We extend our deepest gratitude to Professor John Rachlin for his invaluable guidance and support throughout the coursework. We also thank the dedicated teaching assistants for their assistance and insights during the coursework, enabling us to successfully navigate and complete this complex endeavor."

THANK YOU