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

Prachi Sardana/ Prachi Sardana, Matthew Runyan, Xinlei Hu
cs5200 Summer 2024 Final Project
Northeastern University, Boston, MA, USA

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 both 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.

Introduction

The goal of the project is to integrate data from two biological databases DisGeNET and GTEx. The DisGeNET database contains genes to disease associations and variants to disease associations while the GTEx database contains various types of sequencing data for various tissues. Specifically, we used transcript-level RNA-seq data for various tissues and integrated this transcript expression data with the DisGeNET data. The goal for integrating these two biological databases was to enable the user to query the integrated database for a disease or variant and receive an order list of the most differentially spliced genes that are associated with the given gene.

Motivation

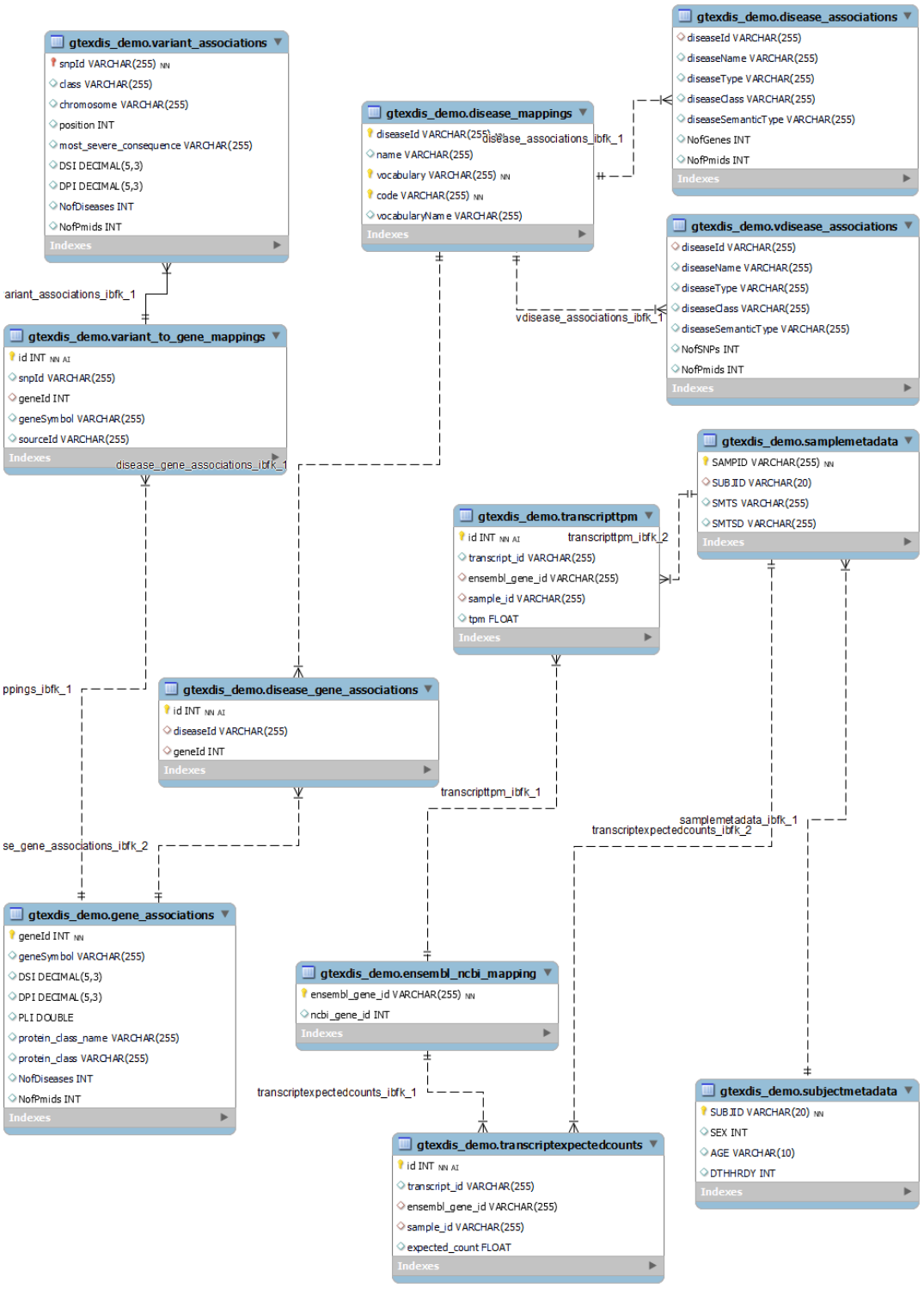
The integration of these databases addresses a significant gap in current biomedical research tools, where researchers often need to consult separate resources to obtain disease-gene association and gene expression data. This disjointed approach can be cumbersome and time-consuming. Moreover, understanding the expression of genes at the transcript level in different tissues can provide deeper insights into the mechanisms by which genetic variations influence disease phenotypes, which is crucial for advancing personalized medicine.

Significance

This project is significant as it enables researchers to:

- Efficiently query combined genetic and expression data.
- Gain insights into the tissue-specific expression patterns of genes associated with particular diseases or variants.
- Identify potential targets for therapeutic intervention and contribute to the development of more effective, personalized treatment strategies.

Database Design



The database design constitutes 2 sections – attributes of each entity and their relationships.
The first section includes the description of attributes and the second section includes the attributes and their relationships.

Entities:

Central to this schema are the gtexdis_demo.samplemetadata and gtexdis_demo.subjectmetadata tables, which store comprehensive metadata about biological samples and the subjects from whom they were collected. The gtexdis_demo.samplemetadata table includes identifiers like SAMPID (sample ID) and SUBJID (subject ID), along with tissue type (SMTS), ensuring each sample is uniquely identifiable and traceable to its source.

The gtexdis_demo.subjectmetadata table provides demographic and clinical information about subjects, including sex (SEX), age (AGE), and cause of death (DTHHRDY), thereby enriching the context for the biological samples and facilitating demographic and clinical correlations in genetic studies.

The schema's capability to handle transcript data is encapsulated in the gtexdis_demo.transcripttpm and gtexdis_demo.transcriptexpectedcounts tables. These tables record transcript abundance and expected count data for each sample, respectively, using Ensembl gene IDs to link specific transcripts to genes.

The gtexdis_demo.transcripttpm table logs transcript abundance in terms of Transcripts Per Million (TPM), while the gtexdis_demo.transcriptexpectedcounts table records expected counts of transcripts, both of which are critical metrics for gene expression analysis. By linking this data to samples via sample_id, these tables enable detailed examination of gene expression profiles across different tissues and conditions.

Genetic variant data is meticulously detailed in the gtexdis_demo.variant_associations table, which includes attributes such as SNP ID (rsnpid), dbSNP Variant ID (dbsnpVariantId), chromosome (chromosome), position (position), and the most severe consequence of the variant (most_severe_consequence).

This table provides a comprehensive overview of genetic variants, including DSI (The Disease Specificity Index) is the specificity of a gene to certain diseases, DPI (Disease Pleiotropy index for the gene) which measures the gene's involvement across multiple diseases, the number of SNPs and PubMed IDs associated with each variant.

The gtexdis_demo.variant_to_gene_mappings table bridges these variants to genes, indicating which genes are affected by specific SNPs through attributes like geneId, geneSymbol, and sourceId.

Disease-related data is systematically organized in the gtexdis_demo.disease_mappings and gtexdis_demo.disease_associations tables. The gtexdis_demo.disease_mappings table includes disease identifiers (diseaseId), names (name), and associated vocabularies and codes (vocabulary, code, vocabularyName).

The gtexdis_demo.disease_associations table enriches this information by providing attributes like disease type (diseaseType), disease class (diseaseClass), and the number of SNPs and PubMed IDs associated with each disease. These tables collectively offer a detailed taxonomy of diseases and their genetic underpinnings.

Crucially, the gtexdis_demo.disease_gene_associations table establishes the critical link between diseases and genes, identifying which genes are implicated in specific diseases. This table is essential for understanding the genetic basis of diseases and their molecular mechanisms.
The gtexdis_demo.ensembl_ncbi_mapping table facilitates the integration of data from different genomic databases by providing mappings between Ensembl gene IDs and NCBI gene IDs.

By integrating sample data with genetic and disease information through these interrelated tables, the schema enables researchers to perform detailed analyses of gene expression, investigate the effects of genetic variants, and explore disease associations.

Relationships:

gtexdis_demo.samplemetadata and gtexdis_demo.subjectmetadata tables, which are linked in a one-to-many relationship, where each subject (one) can provide multiple samples (many). The gtexdis_demo.transcripttpm and gtexdis_demo.transcriptexpectedcounts tables, which store transcript data, are connected to gtexdis_demo.samplemetadata through a many-to-one relationship, where many transcript records can be linked to one sample. The gtexdis_demo.variant_associations table, detailing genetic variants, is indirectly connected to the transcript tables via gtexdis_demo.variant_to_gene_mappings, establishing a many-to-many relationship as one variant can affect multiple genes and vice versa. Disease data is organized in gtexdis_demo.disease_mappings and gtexdis_demo.disease_associations tables, which are in a one-to-one relationship as each disease ID corresponds to detailed disease attributes. The gtexdis_demo.disease_gene_associations table creates a many-to-many relationship between diseases and genes, indicating that multiple diseases can be associated with multiple genes. The gtexdis_demo.ensembl_ncbi_mapping table provides a one-to-one mapping between Ensembl and NCBI gene IDs, facilitating

cross-database integration. These interrelated tables and their respective one-to-one, one-to-many, and many-to-many relationships enable a robust framework for linking sample metadata with genetic variants, transcript data, and disease information, thereby supporting detailed analyses of gene expression, variant effects, and disease associations.

Data Sources and Methods

The data for this project was obtained from DisGeNET [1] and GTEx Portal [2]. Initially, the database schema was created by defining tables for subject metadata, sample metadata, gene associations, variant associations, disease mappings, and transcript data. To begin, subject and sample metadata were loaded from GTEx datasets, linking each sample to its corresponding subject through unique identifiers. The gtexdis_demo.samplemetadata and gtexdis_demo.subjectmetadata tables store detailed metadata about biological samples and subjects, ensuring traceability and facilitating demographic analysis.

Gene and disease data were then incorporated from DisGeNET datasets, involving the loading of disease mappings and gene associations into their respective tables. Temporary tables were employed to ensure only valid data entries were transferred to the primary tables, maintaining data quality. Genetic variant data were subsequently integrated by mapping variants to genes using the variant_to_gene_mappings table and associating variants with their characteristics in the variant_associations table. The ensembl_ncbi_mapping table was populated to map Ensembl gene IDs to NCBI gene IDs, ensuring seamless integration of data.

Given the original data volume of 200 GB, a subset of the data was created to demonstrate the database functionality. Melted test data files, each containing 1,000 lines of transcript expected counts and TPM values, were loaded into the TranscriptExpectedCounts and TranscriptTPM tables, respectively. This subset allows for efficient testing and validation of the database schema while the complete datasets can be accessed and downloaded from GTEx [1]and DisGeNET [2] portals.

Finally, disease-gene associations were established by linking diseases to genes based on specific criteria, ensuring the database could facilitate detailed bioinformatics analyses. This comprehensive approach ensures the database is robust, scalable, and capable of supporting complex queries, thereby enabling researchers to explore the intricate relationships between genes, genetic variants, and diseases effectively.

Analysis

Samples population stats

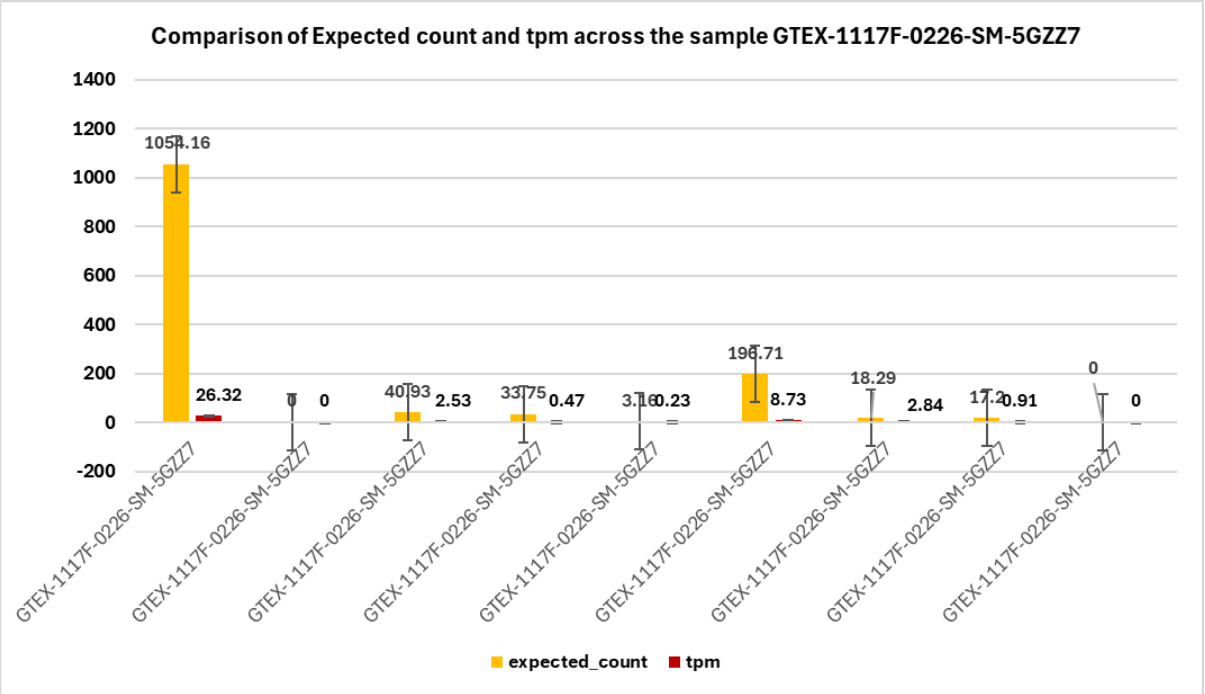
```
406 •  WITH
407        patientsSamples AS (
408            SELECT * FROM SubjectMetadata
409            LEFT JOIN SampleMetadata USING (SUBJID))
410    SELECT SEX, AGE, count(*) AS cnt FROM patientsSamples
411    GROUP BY 1, 2
412    order by 2, 1;
```

SEX	AGE	cnt
1	20-29	1115
2	20-29	666
1	30-39	1172
2	30-39	510
1	40-49	2073
2	40-49	1531
1	50-59	5134
2	50-59	2487
1	60-69	5031
2	60-69	2490
1	70-79	521
2	70-79	221

Query 1 Compare Expected Counts and TPM of Samples

```
SELECT
   sm.SAMPID,
   sm.SUBJID,
   tec.transcript_id,
   tec.ensembl_gene_id,
   tec.expected_count,
   tt.tpm
FROM
   TranscriptExpectedCounts tec
JOIN
   TranscriptTPM tt ON tec.transcript_id = tt.transcript_id
                AND tec.ensembl_gene_id = tt.ensembl_gene_id
                AND tec.sample_id = tt.sample_id
JOIN
   SampleMetadata sm ON tec.sample_id = sm.SAMPID
ORDER BY
```

sm.SAMPID, tec.transcript_id;



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: The occurrence of zero values for expected counts in some transcripts suggests 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.

Query 2 Find the subject's disease

```
SELECT
SUBJID,
SEX,
AGE,
geneId,
diseaseId,
diseaseName,
diseaseType,
diseaseClass,
diseaseSemanticType
FROM joinedPatientsDisease
WHERE diseaseId IS NOT NULL
    AND SUBJID = 'GTEX-1117F';
```

Output:

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
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 (diseaseId 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.

Query 3 Stats of sampling group

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

Output:

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.

Query 4 Disease distribution by name

```
SELECT
    diseaseName,
    count(*) AS cnt
    FROM joinedPatientsDisease
    WHERE diseaseId IS NOT NULL
GROUP BY 1
ORDER BY 1
```

Output:

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 Disease distribution by diseaseType

```
SELECT
    diseaseType,
    count(*) AS cnt
    FROM joinedPatientsDisease
    WHERE diseaseId IS NOT NULL
GROUP BY 1
ORDER BY 1;
```

Output:

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 the scope of the data.

Query 6 Disease distribution by diseaseClass

```
SELECT
    diseaseClass,
    count(*) AS cnt
    FROM joinedPatientsDisease
    WHERE diseaseId IS NOT NULL
GROUP BY 1
ORDER BY 1;
```

Output:

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 Gene Symbols, SNP IDs, and Severe Consequences in Disease-Linked Genes

```
SELECT
    g.geneSymbol,
    v.snpId,
    v.most_severe_consequence
FROM gene_associations g
JOIN variant_to_gene_mappings vg ON g.geneId = vg.geneId
JOIN variant_associations v ON vg.snpId = v.snpId
JOIN disease_gene_associations dg ON g.geneId = dg.geneId
JOIN disease_mappings dm ON dg.diseaseId = dm.diseaseId
LIMIT 20;
```

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

Finding the top 20 records of genes with their Snp id and most severe consequence type. The repeated listing of specific SNPs with severe consequences like "missense variant" suggests these are critical areas for genetic research and could be

significant in studies related to the functions and disorders associated with the CYP26B1 gene.Missense Variant 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.

Conclusions

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.

Accomplishments

We created a robust database schema that integrates genetic variants, gene expression data, and disease information. This schema supports complex queries and detailed analyses at both gene and transcript levels. We also developed several key queries to extract meaningful insights from the data: Query 1 compared expected counts and TPM of samples, providing a comprehensive view of transcript data. Query 2 identified diseases associated with a specific subject, revealing genetic markers linked to diseases and supporting personalized treatment plans. Query 3 analyzed the sampling group, highlighting significant demographics such as females aged 60-69, crucial for targeted healthcare strategies. Query 4 and Query 5 analyzed disease distribution by name and type, respectively, identifying common diseases and their classifications. Query 6 examined disease distribution by disease class, uncovering prevalent disease classes and combination classes for further investigation. Query 7 shows the top 20 records of genes with their Snp id and most severe consequence type.

The analysis revealed that Abdominal Neoplasms is the most frequent disease, suggesting a focus for further research on genetic markers and risk factors. The demographic analysis emphasized the significant presence of females aged 60-69, a key group for developing preventive measures and healthcare strategies. The uniform counts observed in many diseases indicated potential patterns in data collection or preprocessing, highlighting areas for methodological review and improvement. The categorization of diseases into types and classes provided a structured understanding of disease distribution, useful for further targeted analyses.

Limitations

The presence of multiple entries for the same disease and genes, indicating redundancy, points to a need for data normalization and de-duplication to ensure data accuracy and efficiency. The uniform counts across several diseases suggest a potential issue with data batching or grouped data entries, necessitating a review of data collection methods. While we identified significant demographic groups, the analysis could be expanded to include a wider range of demographic factors for a more comprehensive understanding of disease prevalence and characteristics. The insights gained from our database would benefit from validation against external epidemiological data to ensure the reliability and applicability of our findings.

Future Work

To build on the successes and address the limitations of this project, future work should focus on:

- Data Normalization and De-duplication: Implementing techniques to reduce redundancy and ensure data integrity.
- Enhanced Demographic Analysis: Expanding the scope to include more demographic factors for deeper insights.
- Methodological Review: Reviewing data collection and preprocessing methods to address uniform counts and improve data quality.
- External Validation: Cross-referencing findings with external datasets to validate insights and ensure broader applicability.

By advancing these areas, we can further enhance our understanding of genetic variants, gene expression profiles, and their associations with diseases, ultimately contributing to the field of precision medicine and personalized healthcare.

Author Contribution

The success of our project was the result of a collaborative effort where each team member played a critical role. Here's how each member contributed:

Matthew Runya:

- **Data Acquisition**: Matthew was responsible for finding and collecting the necessary data for our project. This foundational work ensured that we had the appropriate datasets to integrate and analyze.
- **Project Direction**: He decided on the subject for our team, setting the overall focus and objectives of the project.
- **Initial Design**: Matthew provided the drafted ER diagram, which served as the blueprint for our database schema. This initial design was crucial in guiding the subsequent development and integration efforts.

Prachi Sardana:

- **Database Creation and Data Loading**: Prachi was instrumental in creating the database and loading the data into it. Her efforts ensured that the data was accurately and efficiently stored, making it ready for analysis.
- **Problem Solving**: Prachi put a lot of time and effort into helping to troubleshoot the database setup.
- **Report Writing**: Prachi authored the 'Abstract', 'Introduction', and 'Data Source and Method sections of the final report. These sections provided a clear and concise overview of the project, its objectives, and the methodology used, setting the stage for the detailed analysis that followed.

Xinlei Hu:

- **Problem Solving**: Xinlei was troubleshooting and resolving issues that arose during the database setup and integration, ensuring that the tables were correctly linked.
- **Query Development and Analysis**: She wrote the queries to extract meaningful insights from the database and performed the analysis on the output.
- **Report Writing**: Xinlei wrote the 'Analysis' part of the final report, detailing the findings from the queries and the implications of these results. This section provided an in-depth look at the data and highlighted key insights and patterns.

Conclusion

Each team member's contributions were vital to the overall success of our project. Matthew's groundwork in data acquisition and initial design provided a strong foundation. Prachi's work in database creation and her clear, informative sections of the report ensured that the project was well-structured and comprehensible. Xinlei's problem-solving abilities and detailed analysis brought the project full circle, transforming our data into meaningful conclusions and insights. Together, these efforts created a comprehensive and valuable resource for understanding genetic variants and their associations with diseases.

References

1. DisGeNET - a database of gene-disease associations. (n.d.). https://www.disgenet.org/downloads#

2. GTEX Portal. (n.d.). https://gtexportal.org/home/downloads/adult-gtex/bulk_tissue_expression

3. GTEX Portal. (n.d.). https://gtexportal.org/home/downloads/adult-gtex/metadata

4. 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.