Computational Strategies in Nutrigenetics: Constructing a Reference Dataset of Nutrition-Associated Genetic Polymorphisms

Giovanni Maria De Filippis[a] (0009-0002-8395-0724), Maria Monticelli[a,b] (0000-0003-3136-2138), Alessandra Pollice[a] (0000-0003-1619-5925), Tiziana Angrisano[a] (0000-0002-5940-5278), Bruno Hay Mele[a*] (0000-0001-5579-183X *), Viola Calabro`[a] (0000-0002-6508-8889)

[a] Department of Biology, University of Napoli "Federico II", Complesso Universitario Monte Sant'Angelo, Via Cinthia, 80126 Napoli, Italy

[b] Institute of Biomolecular Chemistry (ICB), National Research Council (CNR), Via Campi Flegrei 34, 80078 Pozzuoli, Italy

Abstract

Objective: This study aimed to build a comprehensive dataset of human genetic polymorphisms associated with nutrition by integrating data from multiple sources, including the LitVar database, PubMed, and the GWAS catalog. Such a resource could facilitate the exploration of genetic polymorphisms associated with nutrition-related traits.

Methods: We developed a Python pipeline to streamline the integration and analysis of genetic polymorphism data associated with nutrition. We employed the MeSH ontology as a framework to aggregate relevant genetic data. The pipeline comprises five distinct modules that go through the following steps: data extraction from LitVar and PubMed articles, generation of a joint dataset by data merging, generation of comprehensive MeSH term lists, filtering of the joint dataset using the selected MeSH sets, lexical analysis and augmentation of the dataset with data from of the GWAS catalog dataset.

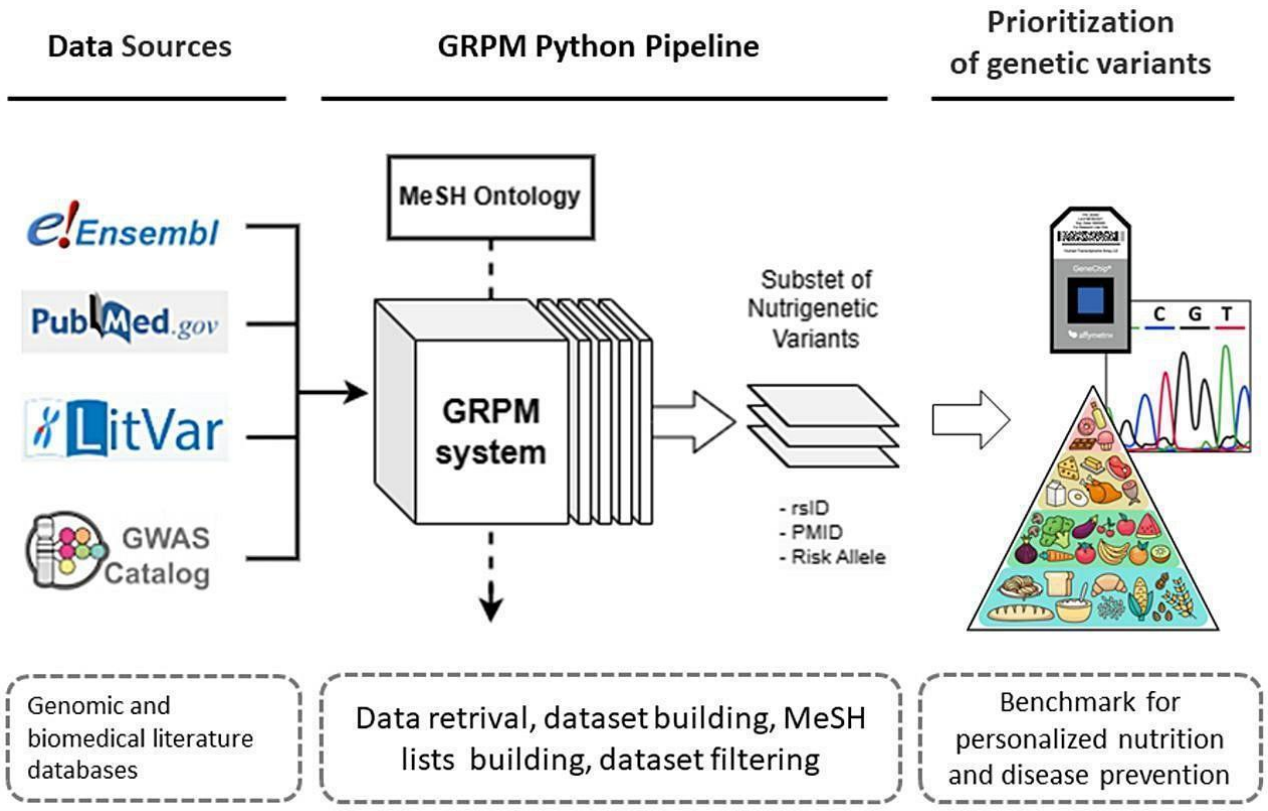
Results: We successfully aggregated a wide range of papers and data on genetic polymorphism and nutrition-related traits into a single dataset. Cross-referencing with the GWAS catalog dataset provided information about possible effects or risk alleles associated with the identified genetic polymorphisms. The nutrigenetic dataset we developed is a tool for nutritionists and researchers, serving as a preliminary benchmark for personalized nutrition interventions based on genetic testing.

Conclusion: The pipeline presented here consolidates and organizes information on genetic polymorphisms associated with nutrition, enabling comprehensive analysis and exploration of gene-diet interactions.

Overall, the method contributes to advancing personalized nutrition interventions and nutrigenomics research. The flexible nature of the system allows its application to other investigations related to genetic polymorphisms.

Keywords: Nutrigenetics, Genetic polymorphisms, Personalized nutrition, Gene-diet interactions, Data integration, MeSH ontology

Graphical Abstract



1. Introduction

Nutrition is critical to health and disease (1). Emerging evidence suggests that genetic polymorphisms significantly impact an individual's response to different nutrients and dietary patterns by affecting nutrient bioavailability and metabolism (2). Moreover, it has been demonstrated that common gene variations are linked to complex chronic health issues significantly affected by nutritional factors (3). Advancements in genomics technologies and the subsequent availability of large-scale genetic data have fueled interest in identifying and categorizing genetic polymorphisms associated with nutrition-related traits (4). The field of nutrigenetics was thus born to comprehend how genetic variations influence an individual's nutritional requirements, metabolism, and health outcomes (5). By considering an individual's genetic profile, healthcare professionals and nutritionists can provide tailored dietary advice and interventions that optimize nutrient bioavailability and promote better health outcomes in that individual (6). Nutrigenetic associations imply that specific genetic polymorphisms can induce susceptibility to chronic diseases. The response to specific nutrients or dietary patterns may be crucial in determining health outcomes (6).

Recent literature contains extensive data on nutrition-associated genetic polymorphisms (2,7). However, these data are often scattered, diverse in format, and lack a standardized curation process. Such complications hinder data integration, limit information extraction and synthesis, and pose a barrier to data utilization in decision support systems (8). Integrating available data and overcoming the limits of self-reported methods in research is crucial for accurate omics data integration, nutrigenetics, and nutrigenomics research, especially in clinical settings (8). Therefore, there is a need to develop curated and consolidated resources that integrate nutrition-associated genetic polymorphism data, along with omics data, to advance personalized nutrition interventions and clinical decision-making.

Here, we built a structured dataset of human genetic polymorphisms associated with nutrition by integrating data from different established sources: the LitVar database (9), which contains curated information on genetic variations and their functional effects; the Pubmed-Medline database, which provides structured MeSH ontology annotations; and the GWAS catalog dataset, which reports human variant-traits associations. Our dataset includes data from Medline studies associated with nutrition-related genetic polymorphism. Specific sets of MeSH terms related to nutrition physiology, nutrition-related diseases, prevention through diet, and eating behavior were used to extract subsets of genes and their single-nucleotide polymorphisms (SNPs) potentially associated with nutrition-related traits. Cross-referencing with the GWAS catalog dataset (10) provided information about effect/risk alleles associated with the collected studies. The database was curated to ensure data quality, consistency, and relevance to nutrition and nutrigenomics research, thus providing a valuable resource to investigate the intricate interplay between genetics and nutrition.

2. Methods

2.1 General Presentation

This study uses Medical Subject Headings (MeSH) to connect genetic polymorphism data from various resources. *MeSH ontology* (11) is a standardized and controlled vocabulary that offers descriptors utilized in biomedicine and informatics to classify and categorize biomedical literature and data. The standardized and controlled nature of MeSH terms makes them highly adaptable for broader utilization in analyzing scientific indicators. MeSH terms associated with a document can be thought of as references to a collection of knowledge stored as documents in a database (12).

To build a structured dataset that aligns genomic data and scientific papers, we connected the LitVar and PubMed databases through shared MeSH terms. The LitVar database is a comprehensive and publicly accessible resource that collects information on genetic variations and their associated scientific literature. It aims to bridge the gap between genomic data and the relevant literature by aggregating and organizing information on genetic variants from a wide range of sources (9). PubMed serves as the principal repository of biomedical literature. Extracting data from PubMed is crucial for various research purposes, such as literature reviews, data mining, and knowledge discovery (13).

We developed a Python pipeline that leverages the MeSH ontology as a crucial framework to aggregate genetic polymorphism data for a topic of interest effectively. We named the tool "GRPM system" (Gene-RsID-PMID-MeSH) and specifically designed it to streamline the integration and analysis of genetic polymorphism data associated with a given biomedical field, such as nutrition. With the increasing importance of genetic factors in understanding nutrition-related traits, the GRPM system could help researchers and nutritionists explore and analyze such data efficiently.

The system is written in Python, exploits the Jupyter Notebook format (14), and comprises five modules designed for specific purposes (Figure 1). These modules support the following operations:

1. Data Retrieval and Merging: facilitates data extraction, integration, and consolidation from source databases, including LitVar and PubMed, ensuring a comprehensive collection of genetic polymorphisms associated with topic-related traits (GRPM dataset).
2. MeSH Term List Creation: generates coherent MeSH term lists and links them to the collection, enabling efficient exploration of the GRPM dataset. This enrichment empowers users to access specific genetic information relevant to the chosen topic easily.
3. Dataset Filtering with Selected MeSH Terms: enables the user to screen collected data using selected MeSH terms. This way, users can refine their search and focus on specific areas of interest within the GRPM dataset.

4. Statistical Analysis: assigns each gene a relative measure of interest based on the number and proportion of associated findings gathered in the GRPM dataset. This calculation enables the prioritization of genes related to a chosen topic, aiding further investigation or personalized nutrition interventions.
5. GWAS Dataset Incorporation: integrates the GWAS catalog dataset into the collection, providing insights into effect/risk alleles associated with identified genetic polymorphisms. This process enriches our nutrigenetic resource with supplemental data for further analysis.

More detailed information about the GRPM system, including implementation details and usage instructions, is available on GitHub[1].

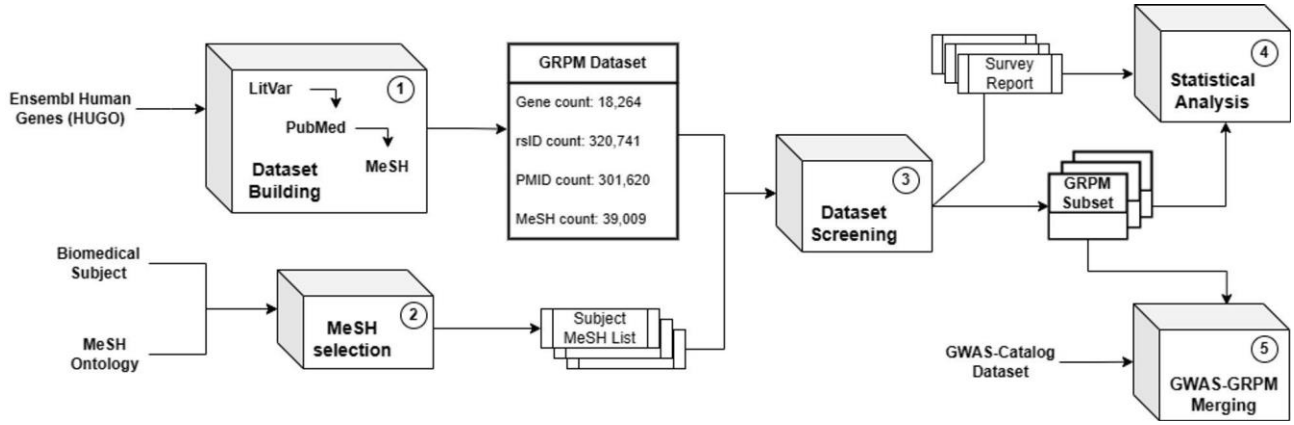


Figure 1: A graphical overview of the GRPM system workflow showcasing the input data and interactions between the five modules. A brief quantitative description of the GRPM Dataset is also shown.

2.2 GRPM Dataset Building

The first module uses the LitVar Application Programming Interface (API)[2] to retrieve all polymorphisms for each human gene[3] within the LitVar database alongside all associated PubMed Identifiers (PMIDs). These PMIDs were subsequently employed as queries on PubMed to obtain MEDLINE data. We utilized the NBIB parser[4], a Python package designed explicitly for parsing MEDLINE-PubMed data, to streamline the data collection. The collected data were ultimately consolidated into a single CSV file (from now on called "GRPM ds"), serving as the primary source against which MeSH term queries can be launched to retrieve genes and polymorphisms associated with specific contexts.

[1] GRPM_system (github.com): https://github.com/johndef64/GRPM_system
[2] LitVar API Docs: https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/LitVar/api.html
[3] The HGCN names lists of human protein-coding genes, RNA genes, and pseudogenes were retrieved from the Ensembl database through the BioMart application.
[4] nbib · PyPI: https://pypi.org/project/nbib/

2.3 Dataset filtering and screening

The GRPM system is designed to retrieve subsets of genes and polymorphisms from GRPM ds, employing a user-defined list of MeSH terms as a hook. Careful selection of the MeSHs is crucial at this stage: the list must represent the chosen search field out of the total complex of terms in GRPM ds. For this purpose, we referred to the subset of 21.705 MeSH terms related to LitVar publications retrieved instead of the complete MeSH ontology (348.733 terms)[5]. This subset collects MeSH terms linked to papers exploring the associations between genetic variants and biomedical traits.

The second module is built to subset this extensive MeSH collection with representatives of a particular biomedical field. We utilized the natural language processing capabilities of Generative Pre-trained Transformer (GPT) provided by OpenAI[6] through its API. The procedure involved generating simple lists of words through one or more biologically relevant prompts (see Supplementary Materials). These lists are used to extract the real MeSH terms related to the subject from our dataset (70-400 MeSHs extracted for each query). Subsequently, the extracted MeSH terms were manually screened to eliminate ambiguous and bias-generating terms. This filtering process ensures that only appropriate and meaningful terms are utilized for the subsequent full dataset screening.

The screening of the entire dataset using the selected MeSH set (our *query*) is obtained by running our third Jupyter module. It takes approximately three to four hours on an average workstation using MeSH lists ranging in size from 100 to 400 terms. When the procedure ends, the system generates a comprehensive report and a curated "Survey Dataset" that captures the essential association data. The reports generated from various surveys are subjected to individual analysis and comparative examination in the fourth module.

2.4 Gene Interest Index (GI)

We consider a gene "interesting" if its related SNPs are associated with a substantial number of PMIDs (*i.e.*, scientific papers) that include relevant MeSH terms and if the ratio between these relevant papers and the total number of papers associated with the gene is sufficiently high.

To assess the relevance of the gene set retrieved for the chosen topic, it is crucial to consider the MeSH set employed as a single entity rather than independently, given the difference in the relative importance of terms. To define a gene as “interesting” based on its associated MeSH terms from related LitVar studies, we propose scaling the number of detected PMIDs (PubMed IDs) by all the PMIDs associated with that gene in

[5] The complete Medical Subject Headings dataset can be downloaded at https://www.nlm.nih.gov/mesh/meshhome.html
[6] OpenAI. (2023). ChatGPT (Mar 14 version) [Large language model]. https://chat.openai.com/chat

LitVar. This approach helps minimize selection bias caused by extensively studied genes associated with more MeSH terms than others — these terms could not be directly correlated with the query topic.

Given the set of genes L(i) retrieved with the query (*j*), we introduce the following indices:

1. **Pg_i**: The total number of PMIDs associated with gene *i*;
2. **P$m_{i,j}$:** The number of *i*-related PMIDs containing at least one MeSH from the query *j*;
3. **Pm_{max}**: The highest **P$m_{i,j}$** value across all the genes in L;
4. **P*m* *score*$_{i,j}$**: the P$m_{i,j}$ value normalized Pm_{max};
5. **P*m* *ratio*$_{i,j}$**: the ratio of Pm to Pg. It measures the proportion of matching PMIDs to the total PMIDs associated with the gene.

Based on these indices, we introduce the "**Gene Interest Value**" **(GV)**, calculated as the product of "Pm score" and "Pm ratio" and its normalized form, the "**Gene Interest Index**" **(GI**), which is adjusted relative to the maximum value obtained in the survey. The ratio serves as a modifier in determining the level of interest for each gene.

$$GV_{ij} = \text{Pm index} \cdot \text{Pm ratio} = \frac{Pm_{ij}}{Pm_{max}} \cdot \frac{Pm_{ij}}{Pg_i} \quad (1.1)$$

$$GI = \frac{GV}{GV_{max}} \quad (1.2)$$

By integrating the Pm score and Pm ratio, the GI method acts as a coherent measure of gene relevance. Figure 2 visually represents an example of gene prioritization obtained through the Index using the "Obesity and Weight Control" MeSH list as a reference. Panel (a) shows the Pm ratio (green) and Pm score (yellow). It highlights the importance of considering both indexes, which produce different orders. In Panel (b), the gene relevance-based sorting achieved with the GI is presented, and it is possible to appreciate the highest prioritization performance versus the other two. The integrated assessment provided by the GI method allows for more accurate gene prioritization, leading to a deeper understanding of gene-gene interactions and potential therapeutic targets in obesity and weight control management. Another example of gene prioritization through GI is presented in Supplementary Materials (Figure S1).

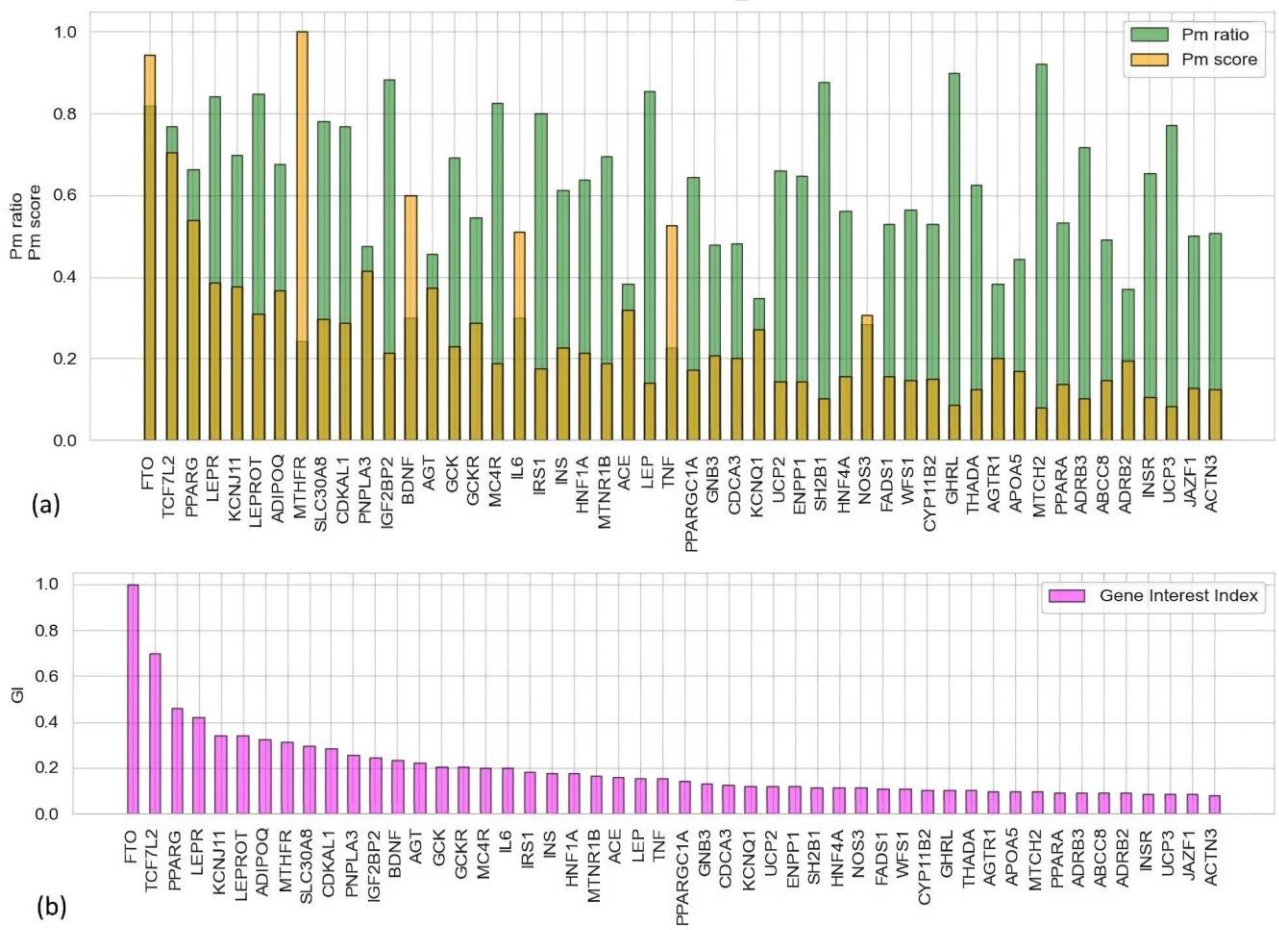


Figure 2: Visual representation of Gene Interest Index (GI) calculation though Pm score and Pm ratio, using as reference the results from "Obesity and Weight Control" MeSH query (see results section). Panel (a) shows the matching PMID ratio and overall matching PMID score. Panel (b) displays the gene relevance sort achieved with the GI.

2.5 GWAS Catalog data integration

While examining every paper to unravel the associated effect allele for each SNP ID (rsID) can be time-consuming, an initial indication of the potential effect allele is valuable for conducting preliminary studies. To address this issue, we leveraged Ensembl GWAS Catalog[7] data (10).

To integrate the GWAS data with the GRPM dataset, we followed a specific workflow (fifth module). First, we retrieved the GRPM Survey data. Then, we applied a Gene Interest (GI) cut-off of 0.0125 to the GRPM Survey data to prioritize the relevant genes. Most protein-coding genes have cited works with at least one of the MeSHs of the lists used, but this does not imply that it is relevant to consider them all. By setting a GI threshold value, we aimed to prioritize the genes that most fit with our tailored MeSH terms, allowing us to focus on genes and their SNPs that demonstrated a higher degree of interest and relevance in the field of nutrigenetic dietary advice. This approach helped remove noise or irrelevant results from the search process,

[7] GWAS Catalog (ebi.ac.uk): https://www.ebi.ac.uk/gwas/docs/file-downloads

allowing us to focus on genes more likely to provide valuable insights into gene-diet interactions and personalized nutrition interventions. At this point, the filtered GRPM Survey and GWAS dataset were merged based on the rsIDs. The merge was efficiently aligned with the GRPM MeSH terms through a correspondence dictionary. We subsequently utilized the Natural Language Toolkit (NLTK)[8] to tokenize the MeSH terms (and all their possible synonyms) and GWAS-mapped traits to perform the alignment. Finally, we retrieved the strongest SNP-risk allele for each rsID using the correspondence dictionary. This information serves as an initial indication and can be beneficial for conducting further studies, whether in a clinical or in-silico setting, based on the identified associations.

3. Results

3.1 GRPM Dataset Statistics

When utilizing the dataset, it is essential to account for each gene's relative richness in PMID. The genes that garnered the most attention in research are associated with a higher number of PMIDs, resulting in more MeSH annotations. Hence, to ensure data normalization, it is imperative to obtain a comprehensive understanding of the most represented genes within the dataset.

We divided the GRPM dataset[9] into three distinct partitions based on the gene type included: Protein Coding Genes, RNA Genes, and Pseudogenes. Since some genes are absent in LitVar, the dataset covers 80% of human protein-coding genes. Figure 3 (a) shows the statistics for each partition, including the number of genes, rsID (reference SNP ID number) entries, PMID entries, and MeSH terms associated with each gene type. These genes were considered for further analysis and integration with other data sources. It is possible to appreciate how pseudogenes are associated with a negligible relative fraction of variants and publications (0.003% and 0.005%, respectively), although they account for ⅓ of Ensembl Genes entries. Figure 3 (b) shows the analysis of the fifty most extensively studied and represented genes within the GRPM dataset. Among them, TP53 and MTHFR concentrate the highest number of PMIDs, with and without MeSHs. This deviation is expected since both encode proteins extremely relevant to human metabolism and health — TP53 (ENSG00000141510) encodes the Cellular tumor antigen p53 (P53_HUMAN), a tumor suppressor that monitors DNA integrity and initiates cellular responses to prevent tumor formation; MTHFR (ENSG00000177000) encodes the methylenetetrahydrofolate reductase (MTHR_HUMAN), that takes part in homocysteine metabolism and is essential for methylation reactions.

[8] NLTK - Natural Language Toolkit: https://www.nltk.org/
[9] Available on Zenodo at https://zenodo.org/record/8205724 DOI: 10.5281/zenodo.8205724

A conspicuous portion of LitVar PMIDs (~77%) PMIDs extracted from LitVar is associated with MeSH terms, ensuring consistency and reliability for further investigations.

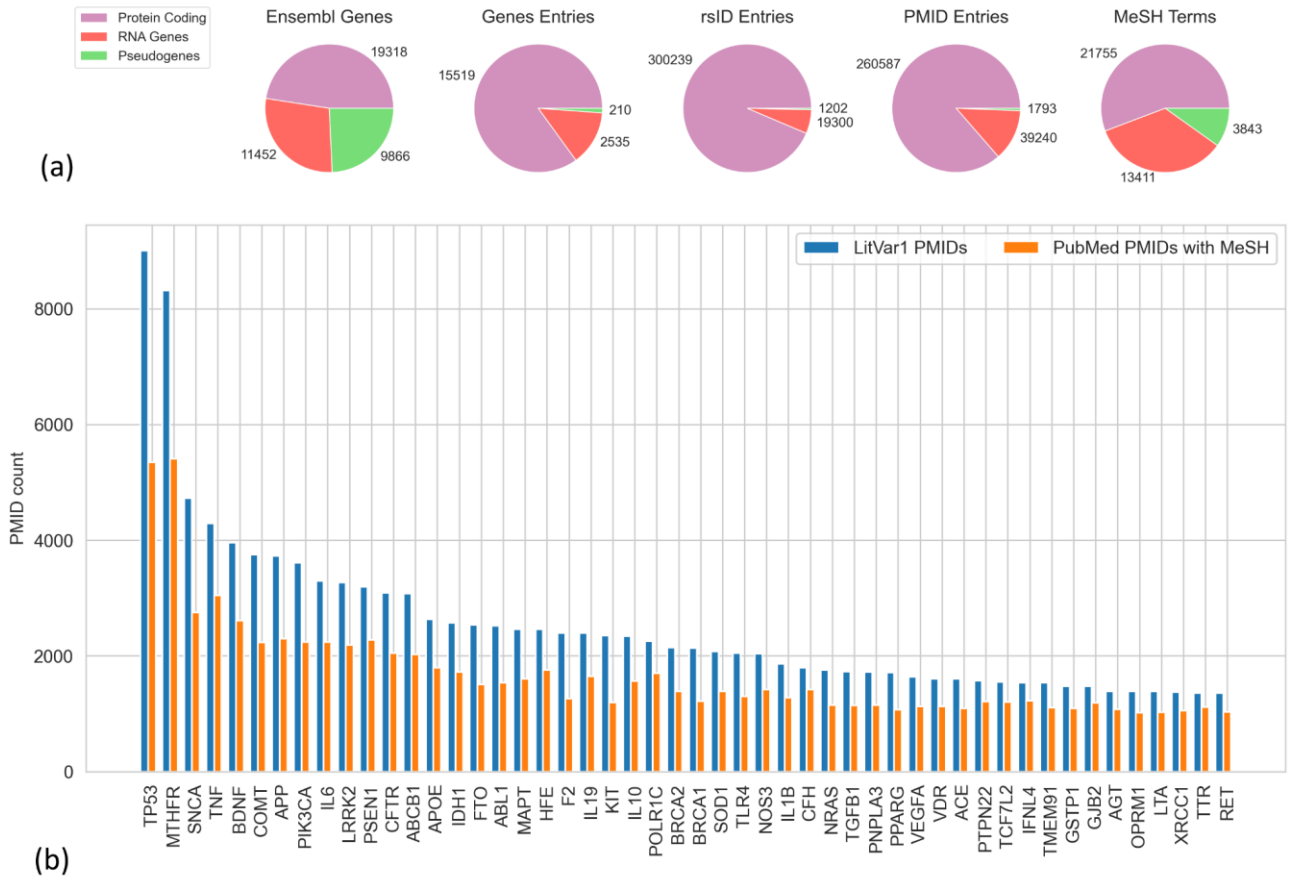


Figure 3 Statistics concerning the primary GRPM dataset. In (a), the diagram showcases the distribution of genes, rsID, PMID, and MeSH terms across the three dataset partitions. (b) depicts the top fifty most represented genes in the dataset, along with their occurrence in PMID counts and PMID associated with MeSH terms.

3.2 Nutrigenetic Dataset

Our study aimed to create a nutrigenetic dataset using a collection of MeSH terms related to different aspects of nutrition. We utilized specific MeSH terms that covered nutrition physiology, nutrition-related diseases, disease prevention through diet, and eating behavior. This approach allowed us to process LitVar-PubMed data in a way relevant to personalized nutritional approaches. In Table 1, we describe each selected field of interest in the context of a personalized nutritional approach.

227 **Table 1:** Categories of nutrition-related MeSH terms used to build the nutrigenetic database.

Category	Description	MeSH count
General Nutrition	A broad range of topics related to nutrition, including dietary patterns, nutrient requirements, nutritional status, and the impact of nutrition on overall health and well-being.	413
Obesity, Weight Control, and Compulsive Eating	Terms related to weight management, including obesity, weight loss strategies, and disorders such as binge eating or compulsive overeating.	243
Cardiovascular Health and Lipid Metabolism	Terms related to nutrition and cardiovascular health, including the impact of dietary factors on lipid metabolism, cholesterol levels, and the prevention of cardiovascular diseases.	319
Diabetes Mellitus Type II and Metabolic Syndrome	Terms related to type II diabetes and metabolic syndrome. Including dietary interventions, glucose metabolism, insulin resistance, and related complications.	528
Vitamin and Micronutrients Metabolism and Deficiency-Related Diseases	Terms related to the metabolism of essential vitamins and micronutrients, the impact of deficiencies on health, and the development of associated diseases.	175
Eating Behavior and Taste Sensation	Terms related to individual eating behaviors, including factors influencing food choices, taste preferences, satiety, and appetite regulation.	292
Food Intolerances	Terms related to adverse reactions to specific foods, such as lactose intolerance or gluten sensitivity. Explores the genetic and physiological factors underlying food intolerances and their impact on dietary choices.	145
Food Allergies	Examines the genetic basis of food allergies, identification of allergenic components, and strategies for managing allergic reactions through diet.	65
Diet-induced Oxidative Stress	Explores the relationship between dietary factors and oxidative stress and investigates the impact of diet on oxidative stress levels and its health implications.	77
Xenobiotics Metabolism	Focuses on the metabolism of foreign substances (xenobiotics) in the body, including drugs, environmental toxins, and dietary components.	170

228

229 Figure S2 presents an overview of the most interesting genes with their relative MeSH and PMID values on
230 the ten nutrient lists used in our study. This figure offers a representative example of the analysis conducted
231 and allows for a quick comparison of gene relevance across different nutrient-related traits. The Figure allows

for a quick comparison of gene relevance across different nutrient-related traits, showing the relative richness of the associated MeSH terms and the papers associated with each gene.

Table 2 provides statistical data on the ten nutritional MeSH term lists employed. The table left section provides information on data retrieved with every nutritional MeSH term list scanning the GRPM dataset. In a large dataset like this, there can be a wide range of genes and associated MeSH terms, making it challenging to identify the genes of highest interest and relevance to the specific research focus. For this reason, the right section presents the same data filtered Gene Interest (GI) < 0.0125, representing the most significant matches available for further investigation and nutrigenetic applications. Applying a GI cut-off was necessary to select the most relevant and meaningful search results.

Table 2: Statistical data associated with the ten nutritional MeSH lists applied to the build dataset.

	All MeSH matching in DB				Interesting entries based on GI threshold (0.0125)			
label	**#gene**	**#rsID**	**#PMID**	**#MeSH**	**#gene**	**#rsID**	**#PMID**	**#MeSH**
General Nutrition	11,560	83,288	62,473	413	686	26,456	44,859	397
Obesity & Weight Control	9,713	53,879	35,563	243	317	10,842	22,123	230
Diabetes Type II & MS	10,717	68,844	49,896	319	603	22,270	36,198	297
Cardiovascular Health	12,368	105,598	85,065	528	975	41,931	66,113	521
Vitamins & Minerals	4,045	16,857	11,941	175	89	3,525	6,882	147
Eating Behavior	5,525	20,607	13,734	292	211	4,252	7,241	256
Food Intolerances	4,040	14,117	7,416	145	392	5,008	4,726	125
Food Allergies	4,681	16,777	11,032	65	451	6,289	7,762	64
Oxidative Stress	5,156	20,919	19,295	77	75	2,559	10,058	60
Xenobiotics Metab	7,115	35,686	27,237	170	173	7,159	14,171	151

We constructed co-occurrence matrices to explore the extent of overlap within the data obtained from the ten nutritional MeSH lists. Figure 4 shows the co-occurrence patterns between genes, rsID, PMID, and MeSH terms among filtered results by GI > 0.0125. The heatmaps[10] in Figure 4 show interesting overlap patterns among the different MeSH lists. For example, the results obtained from the "obesity" lists show a high overlap with the "diabetes" and "cardiovascular health" lists, sharing approximately 80% of the items. The latter shows a 40-60% overlap in this first group of three.

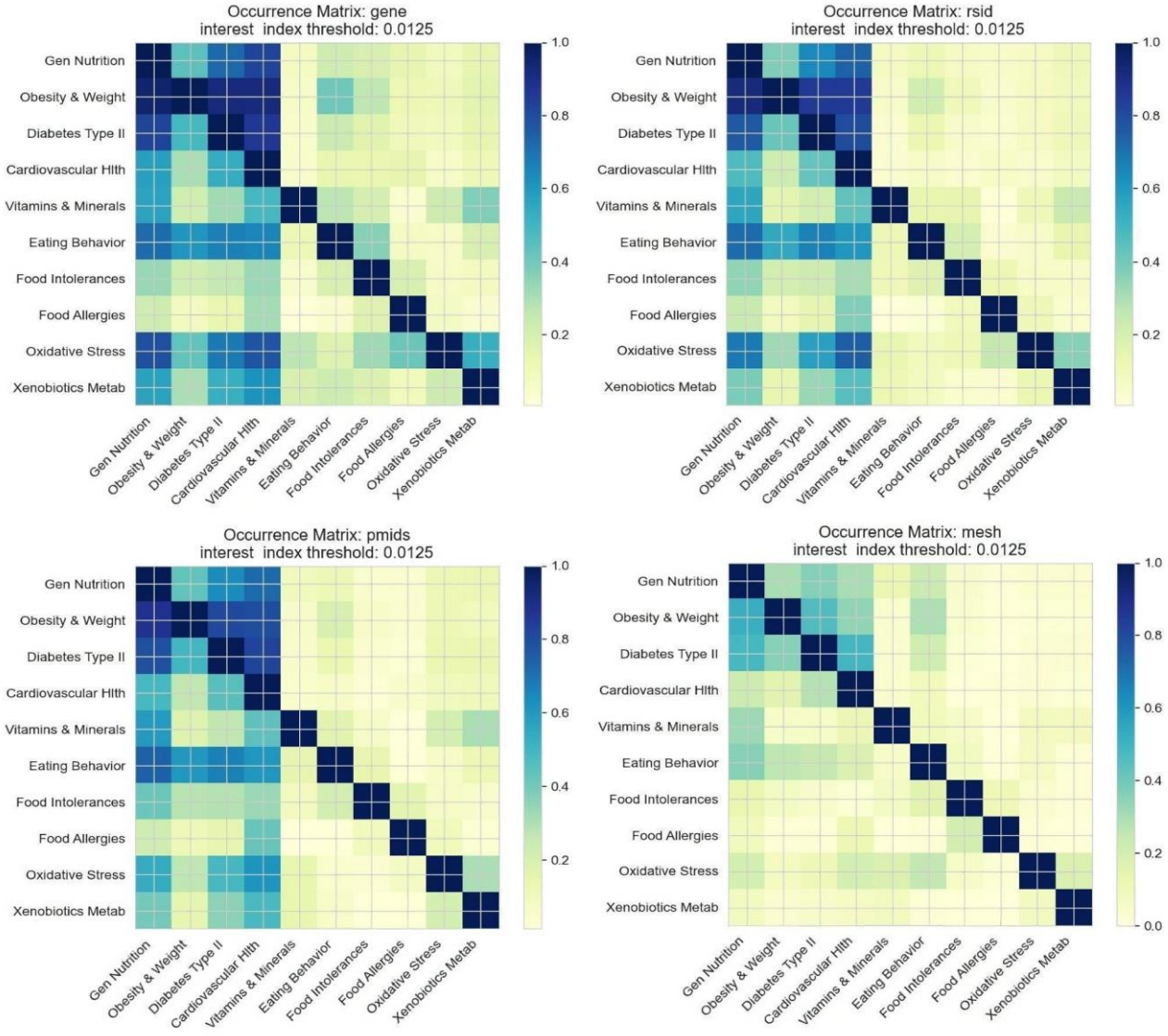


Figure 4: Co-occurrence patterns among genes, rsID, PMID, and MeSH terms observed in the filtered results (GI > 0.0125) from the ten nutritional MeSH lists used for screening. The color key represents the amount of overlap between datasets scaled over the total size of each dataset.

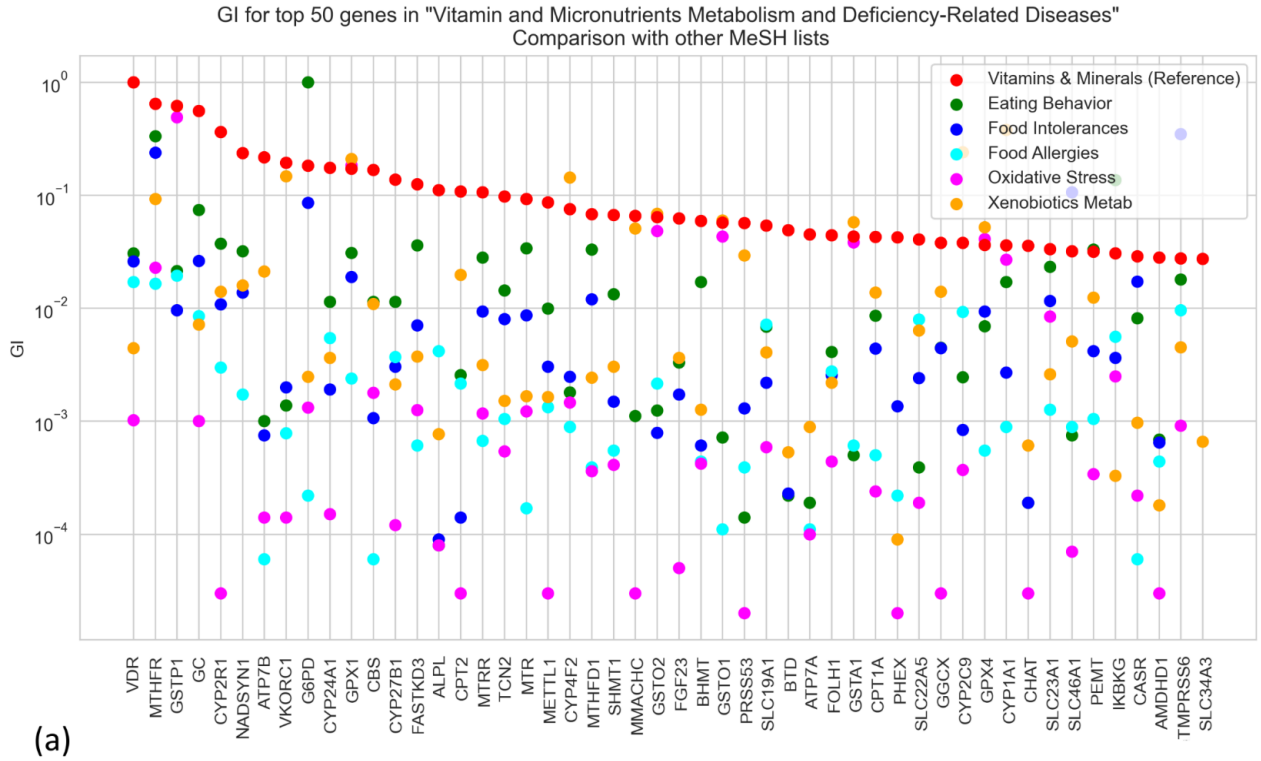
[10] In the heatmap, we calculated row-wise (i.e., dataset) relative values as the ratio between number of shared entities and total number of entities of the row. This approach allows for a focused examination of co-occurrence within specific categories, providing a granular understanding of the relationships and interactions at the row level.

3.3 Method Validation

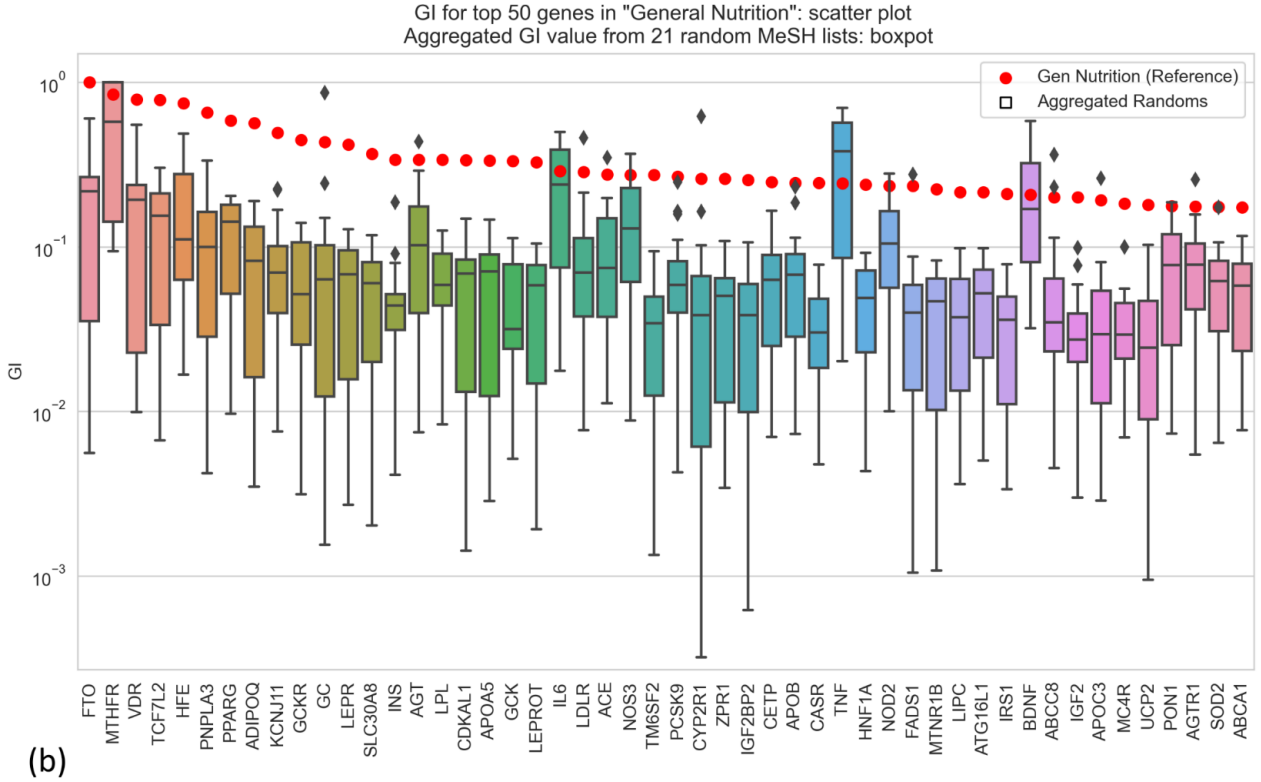
Figure 5 (a) shows the fifty most interesting genes in the "Vitamin and Micronutrient Metabolism" list compared to the values of interest on results obtained from 5 other nutritional MeSH term lists. The genes extracted using the GRPM system show specificity for the list of MeSH terms used as a reference model. This behavior suggests the GRPM system can identify genes related to that particular nutritional aspect. However, some of them are more interesting in other MeSH term lists than the one taken as a reference, meaning they are more closely associated with other nutritional features or specific biological processes. This observation suggests the complexity of gene regulation in nutrient metabolism and underscores the importance of considering a broader range of nutritional MeSH terms to gain a comprehensive understanding of the biological system under consideration. Supplementary Materials provide additional GI comparison results, showcasing the comparison among another MeSH list utilized in our study (Figure S3).

To ensure the accuracy and reliability of the data collected, we compared the results obtained with biologically consistent MeSH lists with those obtained with twenty random MeSH lists[11] of the same size. Figure 5 (b) provides an example of the comparison results on the "General Nutrition" list. Another reference is shown in Supplementary Materials (S4).

[11] Containing 450 random terms each, based on 21,705 LitVar MeSH in the GRPM dataset.



(a)



(b)

Figure 5: (a) Comparison of Gene Interest Index (GI) for the top fifty genes on the "Vitamin and Micronutrient Metabolism" MeSH list with GI obtained from five different lists of nutrition MeSH terms. (b) Comparison between GI obtained from the "General Nutrition" list and GI obtained using twenty randomly generated MeSH lists of the same size represented as boxplots. In both graphs the y-axis is logarithmic; genes are ordered by decreasing interest relative to the reference list.

3.4 GWAS data integration

Cross-referencing data between the GRPM dataset and the GWAS catalog dataset provides indicative information about possible risk alleles associated with the collected studies. **Table 3** shows a sample of this integration on the GRPM "General Nutrition" dataset, along with the corresponding GWAS catalog information, such as Mapped Trait and Strongest Risk Allele. The preliminary results obtained through NLTK show congruence between the MeSH associated with the PMID and the mapped GWAS trait.

The merging process on the "General Nutrition" dataset, with a GI cut-off of 0.0125, resulted in the following statistics: The number of genes from the LitVar database was 365, while the number of genes with mapping information was 359. There were 186 MeSH terms, 467 mapped traits, 1155 disease/traits, and 1678 identified strongest SNP-risk alleles.

Table 3: Sample of fifteen GRPM associations from the "General Nutrition" list merged on rsID with the GWAS Catalog dataset.

GRPM Data				GWAS Data		
LITVAR GENE	LIVAR RSID	LITVAR PMID	PUBMED MeSH	MAPPED TRAIT	DISEASE/TRAIT	STRONGEST SNP-RISK ALLELE
SEM1	rs7781370	22698912	Body Mass Index	body height	Height	rs7781370-T
MTHFR	rs9651118	33213085	Hypertension	red blood cell distribution width	Red cell distribution width	rs9651118-T
CADM2	rs13078807	25893265	Pediatric Obesity	obesity	Obesity	rs13078807-G
C1QTNF6	rs229533	25751624	Diabetes Mellitus, Type 1	type 2 diabetes mellitus	Type 1 diabetes	rs229533-C
GHR	rs6184	34074802	Body Mass Index	body height	Height	rs6184-A
TMEM258	rs102275	31636271	Lipoproteins, LDL	high density lipoprotein cholesterol measurement	Fasting total cholesterol in large HDL	rs102275-C
SHROOM3	rs56281442	34502231	Diabetes Mellitus, Type 2	type 2 diabetes mellitus	Type 2 diabetes	rs56281442-G

LDLR	rs6511720	27973560	Coronary Artery Disease	coronary artery disease	Coronary artery disease	rs6511720-T
TP53INP1	rs10097617	30054458	Diabetes Mellitus, Type 2	type 2 diabetes mellitus	Type 2 diabetes	rs10097617-T
PROX1	rs2075423	32390949	Diabetes, Gestational	type 2 diabetes mellitus	Type 2 diabetes	rs2075423-G
HFE	rs1799945	11473047	Blood Glucose	systolic blood pressure	Systolic blood pressure	rs1799945-G
ESR1	rs6902771	31213659	Body Mass Index	body weight	Weight	rs6902771-T
LEP	rs17151919	33631239	Leptin	leptin measurement	circulating leptin levels	rs17151919-A
CYP2R1	rs7129781	23456391	Vitamin D3 24-Hydroxylase	vitamin D measurement	Vitamin D levels	rs7129781-C
CYP2C19	rs4494250	27618448	Hypertension	diastolic blood pressure	Diastolic blood pressure	rs4494250-A

4. Discussion

Understanding how genetic variations influence individual nutritional requirements, metabolism, and health outcomes is crucial for developing personalized nutrition interventions (15). By considering an individual's genetic profile, healthcare operators and nutritionists can provide tailored dietary advice, optimizing nutrient bioavailability, and promoting better health outcomes, thus preventing chronic diseases such as obesity (16), diabetes (17), or cardiovascular diseases (18).

Personalized approaches in healthcare and prevention are at the forefront of translational bioinformatics (19). The development of computational methods and tools for consolidating and analyzing information from multiple sources enhances data integration, enabling the translation of findings into personalized nutrition interventions and disease prevention strategies (20,21).

The GRPM resource is an integrated platform for diverse data sources related to genetic polymorphisms associated with nutrition. This resource enables efficient retrieval, merging, and analysis, facilitating comprehensive research in nutrigenomics. Our approach leverages data mining and merging techniques to

identify relevant studies on nutrition-associated genetic polymorphisms based on specific MeSH term sets. The pipeline allowed us to extract subsets of genes and associated SNPs linked to nutrition-related traits. Machine learning algorithms could further enhance the analysis of these data to uncover intricate gene-diet interactions by enabling the discovery of patterns and correlations and by producing predictive models (22). Indeed, although focused primarily on genetic polymorphism and associated literature, integrating our study with multi-omics data could provide further insight into the interplay between genetics and nutrition, thus providing a more holistic understanding of such a complex biological field (22).

It is essential to highlight that this method's potential applications extend beyond the scope of our study. Our approach can be employed to gather specific genetic polymorphisms associated with various health or biological dysfunctions, empowering healthcare practitioners to tailor interventions based on an individual's genetic profile (23).

As remarked by Floris and co-workers, the current approach adopted by many companies in nutrigenetic counseling still relies on a limited set of genes and polymorphisms for genetic testing and counseling (24). However, as sequencing costs continue to decrease and sequencing technologies become more accessible than before, it is no longer justifiable to base nutrigenetic panels on a small number of genetic markers (24). Using a limited set of genes and polymorphisms may overlook significant genetic variations that affect an individual's response to nutrients and dietary patterns. Our study addresses the limitations of current approaches in nutrigenetics by consolidating and standardizing information on genetic polymorphisms associated with nutrition. Our nutrigenetic dataset offers a broader scope and coverage, improving the global understanding of the interplay between genetics and nutrition-related traits.

Within GRPM ds, MeSH term richness across nutrigenetic categories is quite heterogeneous, confirming that some categories may have a broader range of MeSH terms, covering various aspects of nutrition, while others may have a narrower focus, addressing specific subtopics within nutrition.

Regarding our nutrigenetic dataset, particularly concerning the genes found on the MeSH lists described in Table 2, it is worth mentioning that the most interesting genes identified in our study are well-known in the field of nutrigenetics. These genes have been extensively studied and represented in literature. The fact that these genes consistently appear in multiple MeSH lists further supports their relevance concerning nutrition-related traits.

From a broader standpoint, the analysis revealed a high degree of overlap between genes associated with specific nutritional aspects (Figure 4). For example, the "obesity" lists exhibit high overlap with the "diabetes" and "cardiovascular health" lists, indicating shared genetic polymorphisms and pathways. This finding is not surprising given the close relationship between obesity, diabetes, and cardiovascular health, as these

conditions often coexist and share common genetic and physiological factors. The high overlap suggests that shared genetic polymorphisms and pathways may be involved in these conditions.

Conversely, the lower degree of overlap between specific lists could be attributed to the specificity of these lists, focusing on more specific biological processes or conditions with distinct genetic underpinnings. This behavior can be attributed to several factors. Firstly, these lists may be described by fewer MeSH terms, leading to a narrower focus and less overlap with other lists. Secondly, these lists may pertain to more specific biological processes or conditions with distinct genetic underpinnings than the broader conditions captured by the other lists.

The results obtained from cross-referencing the GRPM dataset with the GWAS catalog dataset further validated the findings, providing information about potential risk alleles associated with the identified genetic polymorphisms. The preliminary results obtained through NLTK show congruence between the MeSH associated with the PMID and the mapped GWAS trait (Table 3).

We thoroughly validated our method to ensure the validity and reliability of our results. Firstly, we compared the most interesting genes associated with each MeSH list to the results obtained from other nutritional MeSH term lists (Figure 5, A). This validation step demonstrated the effectiveness of the GRPM system in identifying genes specifically related to the chosen nutritional aspect. However, we also observed that specific genes showed higher interest in other MeSH term lists, highlighting the complexity of gene regulation in nutrient metabolism and the need to consider multiple aspects of nutrition.

Furthermore, to ensure our data's accuracy and reliability, we compared the results obtained with biologically consistent MeSH term lists and randomly aggregated MeSH terms (Figure 5, B). This comparison was crucial in assessing the significance of our findings. Our results demonstrated that using biologically consistent MeSH lists led to meaningful results. The genes and variants identified using these lists were relevant to the specific nutritional context studied. On the other hand, when we adopted randomly aggregated MeSH terms lacking biological cohesion, the results were non-significant and lacked meaningful associations. This comparison highlights the importance of utilizing biologically relevant MeSH lists to retrieve and prioritize genuinely relevant to the chosen biomedical context.

However, it is essential to acknowledge the limitations of this approach. One limitation of our approach is the reliance on available literature and databases. The accuracy and reliability of the build resource depend on the quality and completeness of the data retrieved from various sources, as well as the accuracy of the data structuring and integration process. Our method relies on data from Medline studies, which may be subject to publication bias (25). The data quality and consistency of retrieved data heavily depend on the quality of the original studies and the curation process. Despite efforts to ensure data quality, the original studies' inconsistencies, errors, and biases may still be present in the constructed dataset.

Moreover, our dataset is limited to the data available in the LitVar database, GWAS-Catalog, and other sources used in our study (26). As a result, it may not encompass all potential genetic polymorphisms associated with nutrition-related traits. Relying on available literature and data collection databases has limitations (27). Despite our efforts to minimize MeSH attribution bias, the dataset could not contain all the relevant literature. Inconsistencies, errors, and biases in the original studies may be transferred to the constructed dataset. Finally, the dataset may cover only some populations and ethnicities, which could limit its applicability to diverse populations with different genetic backgrounds (28).

Furthermore, it is essential to acknowledge gene-environment interactions' complex and multifactorial nature, including interactions with dietary factors (29). While our dataset captures a subset of the possible interactions, it may not encompass their full complexity. In interpreting the associations between genetic polymorphisms and nutrition-related traits, it is crucial to consider other factors, such as environmental influences, epigenetic modifications, and gene-gene interactions (30). Therefore, the complexity of gene-environment interactions, including interactions with dietary factors, requires further investigation beyond the scope of this research.

Finally, to improve the validity of the study's results, it is essential to assess the quality and scientific validity of the literature sources through established criteria. Future research could follow the scientific validity assessment criteria described by Grimaldi and co-workers (31) to ensure the reliability of individual sources.

5. Conclusion

Our study presents a comprehensive approach for building a nutrigenetic dataset by integrating data from multiple sources through the MeSH ontology. In conclusion, the study presents a methodological framework and a valuable resource for exploring genetic polymorphisms associated with nutrition-related traits. The integration and analysis of genetic polymorphism data in the field of nutrition holds great promise for advancing personalized nutrition interventions. This study highlights the importance of utilizing standardized curation processes and comprehensive datasets in nutrigenomics research. The curated dataset presented here fills a gap in the existing literature by providing a consolidated and standardized resource for investigating genetic polymorphisms associated with nutrition-related traits.

This study exemplifies the importance of translational bioinformatics in nutrition and nutrigenomics, underscoring the significance of computational approaches in consolidating and analyzing information from multiple sources, thereby facilitating comprehensive research in the field of nutrigenomics.

Bibliography

1. Chen Y, Michalak M, Agellon LB. Importance of Nutrients and Nutrient Metabolism on Human Health. Yale J Biol Med. 2018 Jun 28;91(2):95–103.

2. KIANI AK, BONETTI G, DONATO K, KAFTALLI J, HERBST KL, STUPPIA L, et al. Polymorphisms, diet and nutrigenomics. J Prev Med Hyg. 2022 Oct 17;63(2 Suppl 3):E125–41.

3. Loktionov A. Common gene polymorphisms and nutrition: emerging links with pathogenesis of multifactorial chronic diseases (review). J Nutr Biochem. 2003 Aug;14(8):426–51.

4. Mathers JC. Nutrigenomics in the modern era. Proc Nutr Soc. 2017 Aug;76(3):265–75.

5. Kaput J. Nutrigenomics research for personalized nutrition and medicine. Curr Opin Biotechnol. 2008 Apr;19(2):110–20.

6. Fenech M, El-Sohemy A, Cahill L, Ferguson LR, French TAC, Tai ES, et al. Nutrigenetics and nutrigenomics: viewpoints on the current status and applications in nutrition research and practice. J Nutr Nutr. 2011;4(2):69–89.

7. Comerford KB, Pasin G. Gene–Dairy Food Interactions and Health Outcomes: A Review of Nutrigenetic Studies. Nutrients. 2017 Jul 6;9(7):710.

8. Singh V. Current challenges and future implications of exploiting the omics data into nutrigenetics and nutrigenomics for personalized diagnosis and nutrition-based care. Nutrition. 2023 Jun 1;110:112002.

9. Allot A, Peng Y, Wei CH, Lee K, Phan L, Lu Z. LitVar: a semantic search engine for linking genomic variant data in PubMed and PMC. Nucleic Acids Res. 2018 Jul 2;46(Web Server issue):W530–6.

10. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017 Jan 4;45(Database issue):D896–901.

11. Use of MeSH in Cataloging [Internet]. [cited 2023 Aug 3]. Available from: https://wayback.archive-it.org/org-350/20200228165129/https://www.nlm.nih.gov/tsd/cataloging/MeSH_CatPractices.html

12. Leydesdorff L, Comins JA, Sorensen AA, Bornmann L, Hellsten I. Cited references and Medical Subject Headings (MeSH) as two different knowledge representations: clustering and mappings at the paper level. Scientometrics. 2016;109(3):2077–91.

13. Khare R, Leaman R, Lu Z. Accessing Biomedical Literature in the Current Information Landscape. Methods Mol Biol Clifton NJ. 2014;1159:11–31.

14. Pimentel JF, Murta L, Braganholo V, Freire J. Understanding and improving the quality and reproducibility of Jupyter notebooks. Empir Softw Eng. 2021;26(4):65.

15. Reddy VS, Palika R, Ismail A, Pullakhandam R, Reddy GB. Nutrigenomics: Opportunities & challenges for public health nutrition. Indian J Med Res. 2018 Nov;148(5):632–41.

16. O'Rahilly S, Farooqi IS. Human obesity as a heritable disorder of the central control of energy balance. Int J Obes 2005. 2008 Dec;32 Suppl 7:S55-61.

17. Wang DD, Hu FB. Precision nutrition for prevention and management of type 2 diabetes. Lancet Diabetes Endocrinol. 2018 May;6(5):416–26.

18. Peña-Romero AC, Navas-Carrillo D, Marín F, Orenes-Piñero E. The future of nutrition: Nutrigenomics and nutrigenetics in obesity and cardiovascular diseases. Crit Rev Food Sci Nutr. 2018 Nov 22;58(17):3030–41.

19. Di Camillo B, Giugno R. From translational bioinformatics computational methodologies to personalized medicine. J Biomed Inform. 2022 Sep 1;133:104170.

20. Lee KH, Kim JH. Evolution of Translational Bioinformatics: lessons learned from TBC 2016. BMC Med Genomics. 2017 May 24;10(Suppl 1):32.

21. Tenenbaum JD. Translational Bioinformatics: Past, Present, and Future. Genomics Proteomics Bioinformatics. 2016 Feb;14(1):31–41.

22. Zhang Y, Zhu Q, Liu H. Next generation informatics for big data in precision medicine era. BioData Min. 2015;8:34.

23. Carrasco-Ramiro F, Peiró-Pastor R, Aguado B. Human genomics projects and precision medicine. Gene Ther. 2017 Sep;24(9):551–61.

24. Floris M, Cano A, Porru L, Addis R, Cambedda A, Idda ML, et al. Direct-to-Consumer Nutrigenetics Testing: An Overview. Nutrients. 2020 Feb 21;12(2):566.

25. Carter AO, Griffin GH, Carter TP. A survey identified publication bias in the secondary literature. J Clin Epidemiol. 2006 Mar;59(3):241–5.

26. Lachance J, Tishkoff SA. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. BioEssays News Rev Mol Cell Dev Biol. 2013 Sep;35(9):780–6.

27. Williams WG. Uses and limitations of registry and academic databases. Semin Thorac Cardiovasc Surg Pediatr Card Surg Annu. 2010;13(1):66–70.

28. Nilsson PD, Newsome JM, Santos HM, Schiller MR. Prioritization of Variants for Investigation of Genotype-Directed Nutrition in Human Superpopulations. Int J Mol Sci. 2019 Jul 18;20(14):3516.

29. Virolainen SJ, VonHandorf A, Viel KCMF, Weirauch MT, Kottyan LC. Gene-environment interactions and their impact on human health. Genes Immun. 2023 Feb;24(1):1–11.

30. Cole BS, Hall MA, Urbanowicz RJ, Gilbert-Diamond D, Moore JH. Analysis of Gene-Gene Interactions. Curr Protoc Hum Genet. 2017 Oct 18;95:1.14.1-1.14.10.

31. Grimaldi KA, van Ommen B, Ordovas JM, Parnell LD, Mathers JC, Bendik I, et al. Proposed guidelines to evaluate scientific validity and evidence for genotype-based dietary advice. Genes Nutr. 2017;12:35.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used the Generative Pre-trained Transformer (GPT) provided by OpenAI API to facilitate the subsetting of the extensive MeSH collection with representatives of a particular biomedical field. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

Author contributions

GMDF: Writing - original draft, Conceptualization, Data curation, Software, Visualization, Writing – review & editing; **MM**: Writing - original draft, supervision, Investigation, Writing – review & editing; **AP**: Supervision, Writing – review & editing, Validation; **TA**: Supervision, Writing – review & editing, Validation; **BHM**: Writing - original draft, Conceptualization, Formal analysis, Validation, Writing – review & editing; **VC**: Funding acquisition, Conceptualization, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

Our research was supported by Federazione Nazionale Degli Ordini dei Biologi.