medRxiv preprint doi: https://doi.org/10.1101/2023.08.04.23293659; this version posted June 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

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

Giovanni Maria De Filippis[3],**Maria Monticelli** [2], **Alessandra Pollice**[1], **Tiziana Angrisano**[1], **Bruno Hay Mele**[1], **Viola Calabrò**[1]

[1] Department of Electrical Engineering and Information Technology, University of Naples Federico II, Naples, via Claudio 21, 80125, Italy

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

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

Abstract

Objective This study aims to create a comprehensive and curated dataset of human genetic polymorphisms associated with nutrition by integrating data from multiple sources, including the LitVar database, PubMed, and the GWAS catalog. This consolidated resource is intended to facilitate research in nutrigenetics by providing a reliable foundation to explore genetic polymorphisms linked to nutrition-related traits.
Methods We developed a data integration pipeline to assemble and analyze the dataset. The pipeline performs data retrieval from LitVar and PubMed, data merging to build a unified dataset, definition of comprehensive MeSH lists, querying this dataset by MeSH to retrieve relevant genetic associations, and cross-referencing the output with the GWAS catalog.
Results The resulting dataset aggregates extensive information on genetic polymorphisms and nutrition-related traits. Through MeSH query, we identified key genes and SNPs associated with nutrition-related traits. Cross-referencing with the GWAS catalog provided insights on potential effects or risk alleles associated with this genetic polymorphisms. The co-occurrence analysis revealed meaningful gene-diet interactions, advancing personalized nutrition and nutrigenomics research.
Conclusion The dataset presented here consolidates and organizes information on genetic polymorphisms associated with nutrition, enabling detailed exploration of gene-diet interactions. This resource advances personalized nutrition interventions and nutrigenomics research by providing a standardized and comprehensive dataset. The flexible nature of the dataset allows its application to other genetic polymorphism investigations.

Keywords

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

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 the identification and categorizing of genetic polymorphisms associated with nutritional traits [4]. Thus, the field of nutritional genetics (nutrigenetics) was 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 bio-availability 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 [7].

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2023.08.04.23293659; this version posted June 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

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. Today, technologies are available to overcome these limitations: the use of ontologies for information retrieval (IR) is a well-known technique in the literature for semantic search [9], while Named Entity Recognition (NER) techniques are increasingly important in biomedical literature mining [10] to obtain key information on genomic variants for personalised medicine.

Here, we built a structured dataset of human genetic polymorphisms associated with nutrition by mining the LitVar database [11], 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.

This data where then queried employing MeSH ontology for retrieval of nutrition-related genetic data.

Specific sets of MeSH terms related to nutrition physiology, nutrition-related diseases, prevention through diet, and eating behavior were used to retrieve subsets of genes and their single-nucleotide polymorphisms (SNPs) potentially associated with nutrition-related traits. Cross-referencing with the GWAS catalog dataset [12] 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

To build an integrated dataset that aligns genomic data and scientific papers, we connected the LitVar and PubMed databases through shared PubMed IDs to enrich LitVar association data with Medical Subject Headings (MeSH).

MeSH ontology [13] 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 [14].

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 [11].

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 [15].

In order to build our curated dataset, we developed a Python pipeline that leverages the MeSH ontology as a pivotal framework to aggregate genetic polymorphism data for a topic of interest effectively. This pipeline is designed to streamline the information retrieval (IR), integration and analysis of genetic polymorphism data associated with a given biomedical field, such as nutrition. We named the resulting dataset GRPM, out of its main descriptors (Genes, RsIDs, PMIDs, MeSH). With the increasing importance of genetic factors in understanding nutrition-related traits, the GRPM dataset could help researchers and nutritionists explore and analyze these data efficiently.

The retrieval-integration pipeline is written in Python inside a Jupyter Notebook interactive environment [16], and comprises five modules designed for specific purposes (Figure 1). These modules support the following operations:

1. GRPM Dataset Building: operates 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 Selection for Retrieval: defines coherent MeSH ontology term sets for information retrieval over the whole GRPM Dataset.

3. GRPM Dataset MeSH Query: allows user to query the dateset using 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 metric enables the prioritization of genes related to a chosen topic, aiding further investigation or personalized nutrition interventions.

5. GWAS Data Integration: integrates the GWAS catalog dataset into the collection, providing insights

medRxiv preprint doi: https://doi.org/10.1101/2023.08.04.23293659; this version posted June 21, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

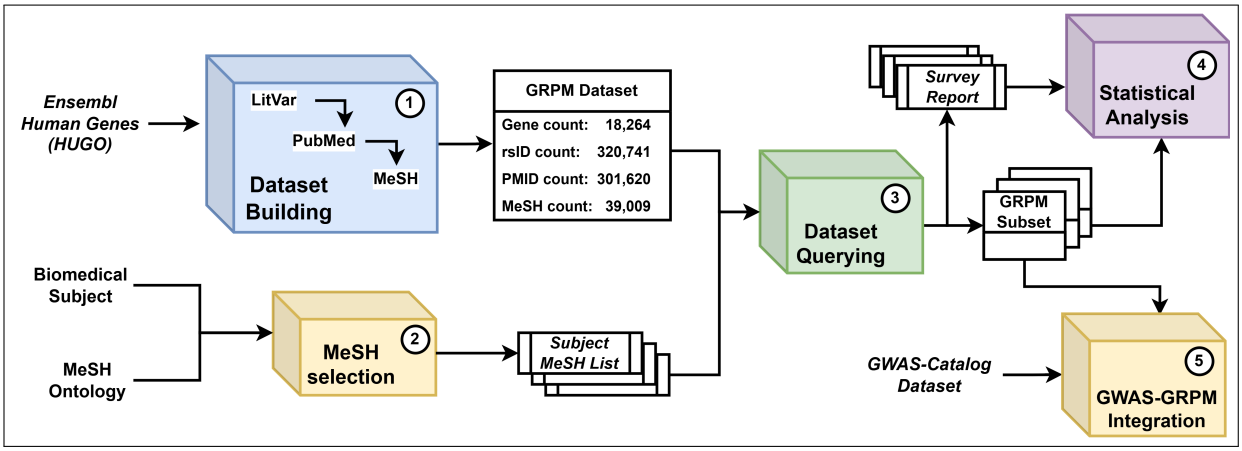

Figure 1: A graphical overview of the GRPM workflow showcasing the input data and interactions between the five modules.

into effect/risk alleles associated with identified genetic polymorphisms. This process enriches our nutrigenetic resource with supplemental data for further analysis.

The entire pipeline, including implementation details and usage instructions, can be openly accessed on GitHub[1].

2.1 GRPM Dataset Building

The first module uses the LitVar Application Programming Interface (API)[2] to retrieve all polymorphisms for each human gene within the LitVar database alongside all associated PubMed Identifiers (PMIDs). These PMIDs were subsequently employed as queries on PubMed to obtain the bibliographical data. We empoloyed an NBIB parser[3], to extract and structure this data in a machine readable format. The collected data were ultimately consolidated into a single CSV file ("GRPM dataset"), serving as the primary source against which MeSH term queries can be employed to retrieve genes and polymorphisms associated with specific contexts.

This work is based on the first version of LitVar, which is no longer available online and has been entirely replaced by LitVar2 [10]. This version was chosen based on several reasons. Firstly, the first version of LitVar possesses a higher level of reliability, a product of extensive examination and rectification of any discrepancies over its period of usage. Besides, the relatively simpler structure of the data in this version eschews unnecessary complexity posed by more recent data structures, thereby making data extraction and manipulation operations more straightforward. The decision to use LitVar1 was the result of a thorough cost-benefit analysis, weighing the potential superior data precision provided by LitVar2, which also comes with substantially larger datasets that could introduce additional noise, against the reliability and simplicity of the first version. The dataset produced here provides a faithful and historical archive of the first version of LitVar by collating the bibliographic references along with the genes and polymorphisms associated with them.

2.2 Dataset querying and retrieval

The retrieval system to get subsets of genes and polymorphisms from GRPM dataset employes 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 dataset.

The total set of MeSH describing the GRPM dataset comprises 21.705 terms related to LitVar publications retrieved out of the complete MeSH ontology (348.733 terms)[4]. Therefore, this subset collects ontology terms linked to papers exploring the associations between genetic variants and biomedical traits.

The second module is designed to extract from this collection of MeSH the tems that represent specific biomedical fields. Leveraging natural language processing (NLP) techniques, we generated topical words based on our domain knowledge in nutrigenetics. These topic words were then utilized to retrieve related MeSH ontology terms using the Natural Language Toolkit (NLTK Python package). Following retrieval, we filtered the remaining MeSH terms by their semantic types and performed a manual screening eliminating those that were ambiguous or potentially introduced bias. This filtering process ensures that only appropriate and meaningful terms are utilized for the subsequent full dataset screening.

This MeSH list are then used as query for retrieval in the third module, a procedure that employs about ten minutes to generates a comprehensive report and a curated *"Survey Dataset"* that captures the essential asso-

[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]nbib · PyPI: https://pypi.org/project/nbib/

[4]The complete Medical Subject Headings dataset can be downloaded at https://www.nlm.nih.gov/mesh/meshhome.html

medRxiv preprint doi: https://doi.org/10.1101/2023.08.04.23293659; this version posted June 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

ciation data. The reports generated from various surveys are subjected to individual and comparative analysis in the fourth module.

2.3 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 MeSH terms in query and if the ratio between these matching 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 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. P_{gi}: The total number of PMIDs associated with gene i;
2. $P_{mi,j}$: The number of i-related PMIDs containing at least one MeSH from the query j;
3. P_{mmax}: The highest $P_{mi,j}$ value across all the genes in L;
4. $P_{mscorei,j}$: the $P_{mi,j}$ value normalized P_{mmax};
5. $P_{mratioi,j}$: the ratio of P_m to P_g. 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 "P_m score" and "P_m 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.

$$GV_{ij} = \text{Pm index} \cdot \text{Pm ratio} = \frac{Pm_{ij}}{Pm_{\max}} \cdot \frac{Pm_{ij}}{Pg_i} \quad (1)$$

$$GI = \frac{GV}{GV_{\max}} \quad (2)$$

By integrating the P_m score and P_m 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 P_m ratio (green) and P_m 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).

In Section 3.2, we present the results obtained by applying the Gene Interest Index (GI) to ten nutritional MeSH queries results for genetic association retrieval on the GRPM dataset. We established a GI threshold of 0.0125, which corresponds to the mean value of the 95th percentile across all ten query results. This threshold encompasses the top 5% of the retrieval results on average, thereby accommodating the long-tail distribution characteristic of the data.

Most protein-coding genes had citations with at least one of the MeSH in the query, but not all are relevant. By setting a GI threshold, we prioritized genes that fit our tailored MeSH terms, focusing on those with higher relevance in nutrigenetic dietary advice. This helped eliminate noise and focus on genes likely to offer valuable insights into gene-diet interactions and personalized nutrition.

2.4 GWAS Catalog data integration

While examining every study retrieved 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[5] data [12].

To integrate the GWAS data within the GRPM dataset, we followed a specific workflow (fifth module). After retrieving GRPM association data by MeSH query, we applied a GI cut-off of 0.0125 to the results to prioritize the relevant genes. 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)[6] 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.

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

[6] NLTK - Natural Language Toolkit: https://www.nltk.org/

medRxiv preprint doi: https://doi.org/10.1101/2023.08.04.23293659; this version posted June 21, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

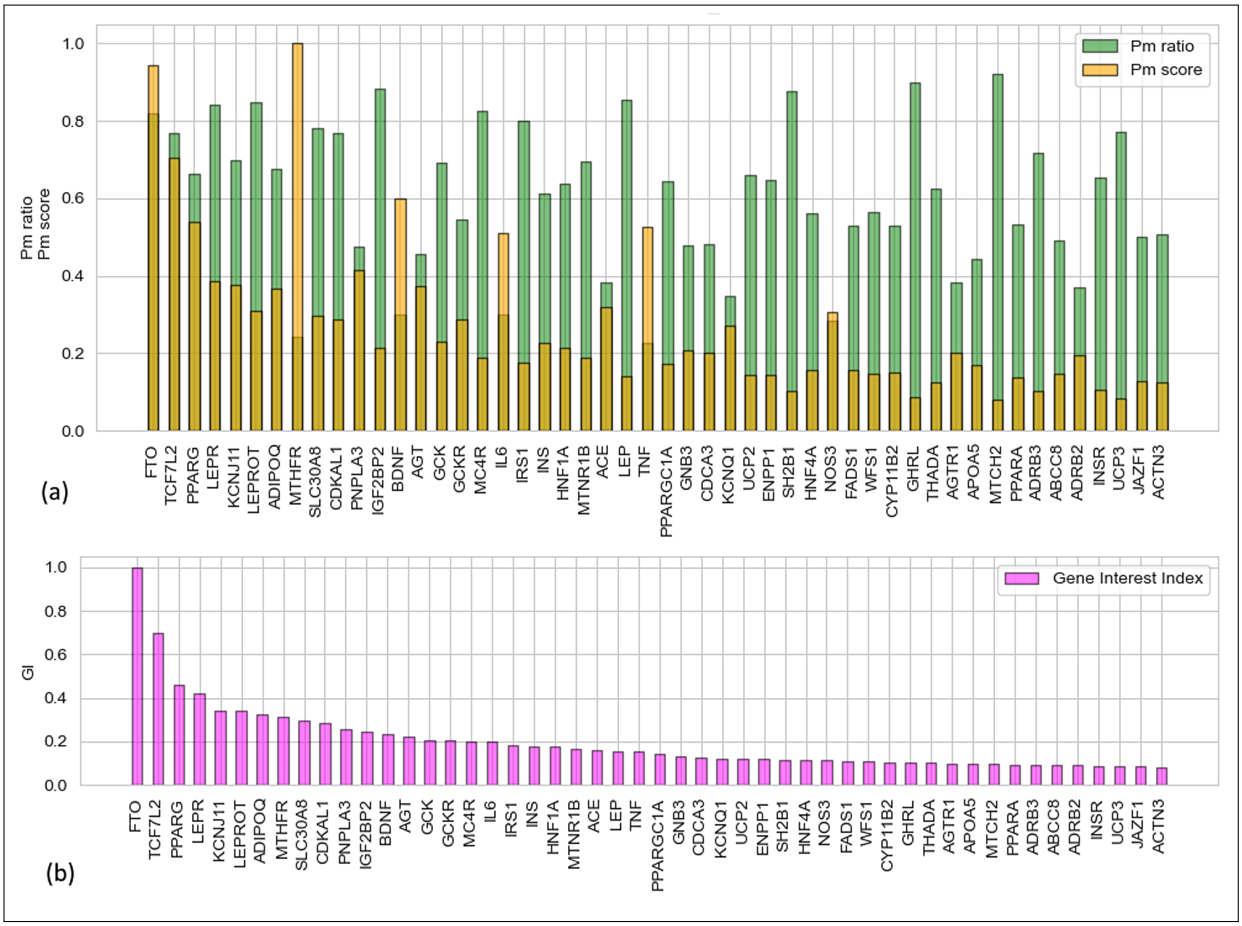

Figure 2: Visual representation of Gene Interest Index (GI) calculation though P_m score and P_m ratio, using as reference the results from *"Obesity and Weight Control"* MeSH query (see results section). Panel (a) shows the matching PMID ratio and overall matching PMID score. Panel (b) displays the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

3 Results

3.1 GRPM Dataset Statistics

When utilizing this 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 necessary to obtain a comprehensive understanding of the most represented genes within the dataset. A conspicuous portion of LitVar PMIDs extracted from LitVar is associated with MeSH terms (77%) , ensuring consistency and reliability for further investigations (Figure 3).

3.2 The 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.

Figure S2 presents an overview of the most interesting genes with their relative MeSH and PMID values on the ten nutrient lists used in our study. This figure offers a representative example of the analysis conducted and allows for a quick comparison of gene relevance across different nutrient-related traits. Figure S2 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 collects the feature metrics from the nutritional datasets extracted in our study. The table left section provides information on data retrieved with every nutritional MeSH query over the GRPM dataset. Given the extensive size of our dataset, the wide range of genes and associated MeSH terms makes challenging to identify the genes most pertinent to the specific research objectives. To address this, the right section of the table presents results filtered by 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

medRxiv preprint doi: https://doi.org/10.1101/2023.08.04.23293659; this version posted June 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

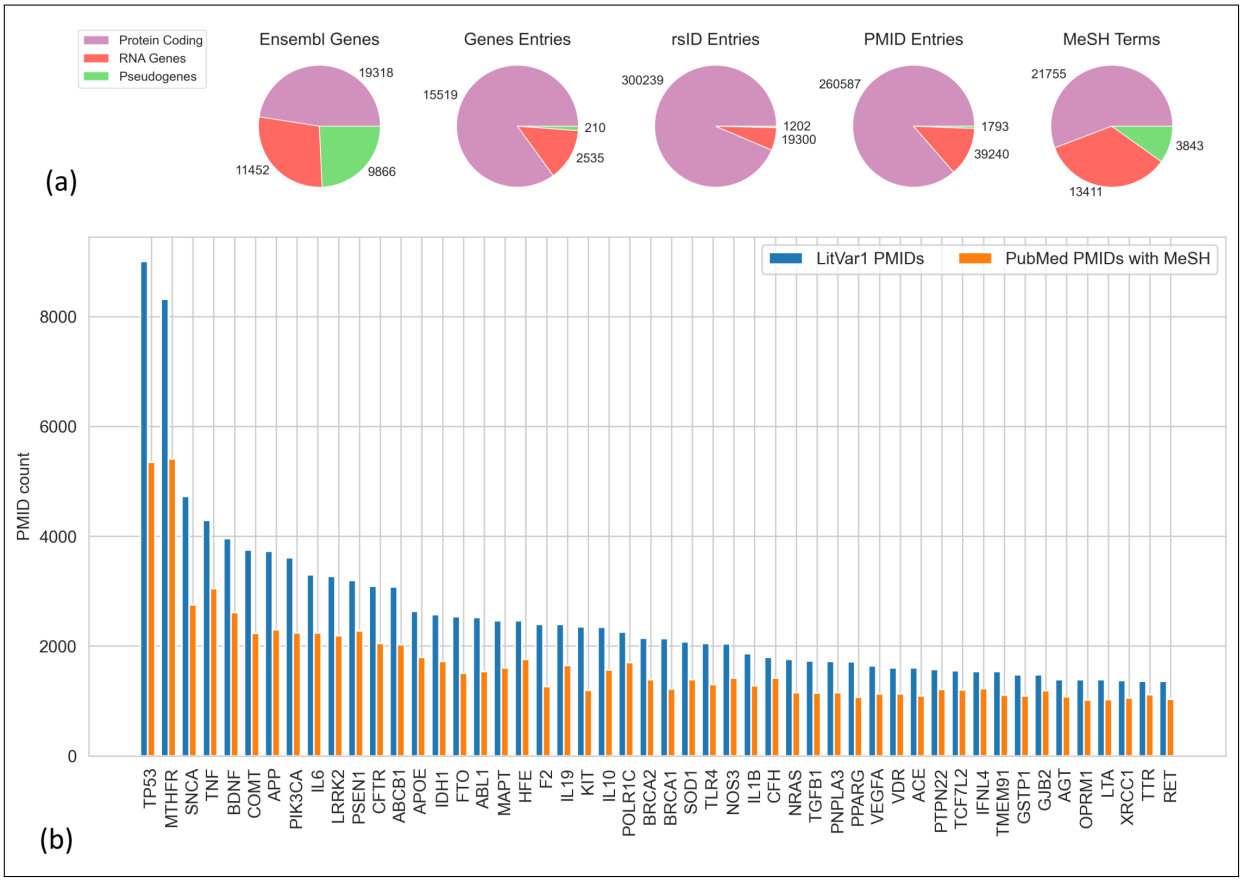

Figure 3: Statistics concerning the primary GRPM dataset. In (a), the diagram showcases the distribution of genes, rsID, PMID, and MeSH terms across the three dataset partitions. (b) depicts the top fifty most represented genes in the dataset, along with their occurrence in PMID counts and PMID associated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

most relevant and meaningful search results.

We constructed co-occurrence matrices to explore the extent of overlap within the data obtained from the ten nutritional MeSH queries. Figure 4 shows the co-occurrence patterns between genes, rsID, PMID, and MeSH terms among filtered results by GI > 0.0125. The correlation matrix[7] 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.

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 lists. The genes extracted using the our method show specificity for the list of MeSH terms used as a reference model. This behavior suggests that our method can identify genes related to that particular nutritional aspect. It can be seen in Figure 5 (a) that some of the genes has higher GI in other MeSH lists than the one taken as a reference in the plot, 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 queries with those obtained with twenty random MeSH lists[8] of the same size. Figure 5 (b) provides an example of the comparison results on the *"General Nutrition"* list. Another example is shown in Supplementary Materials (S4).

[7]In this correlation analysis, 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.

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

medRxiv preprint doi: https://doi.org/10.1101/2023.08.04.23293659; this version posted June 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

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, the identification of allergenic components, and strategies to manage 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

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.

4 Discussion

Understanding how genetic variations influence individual nutritional requirements, metabolism, and health outcomes is crucial for developing personalized nutrition interventions [17]. 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 [18], diabetes [19], or cardiovascular diseases [20].

Personalized approaches in healthcare and prevention are at the forefront of translational bioinformatics [21]. 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 [22, 23].

The GRPM dataset is an integrated resource for diverse data sources related to genetic polymorphisms associated with nutrition. This resource enables efficient retrieval,

medRxiv preprint doi: https://doi.org/10.1101/2023.08.04.23293659; this version posted June 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Table 2: Categories of nutrition-related MeSH terms used to build the nutrigenetic database.

	All MeSH matching in DB				Interesting entries based on GI (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

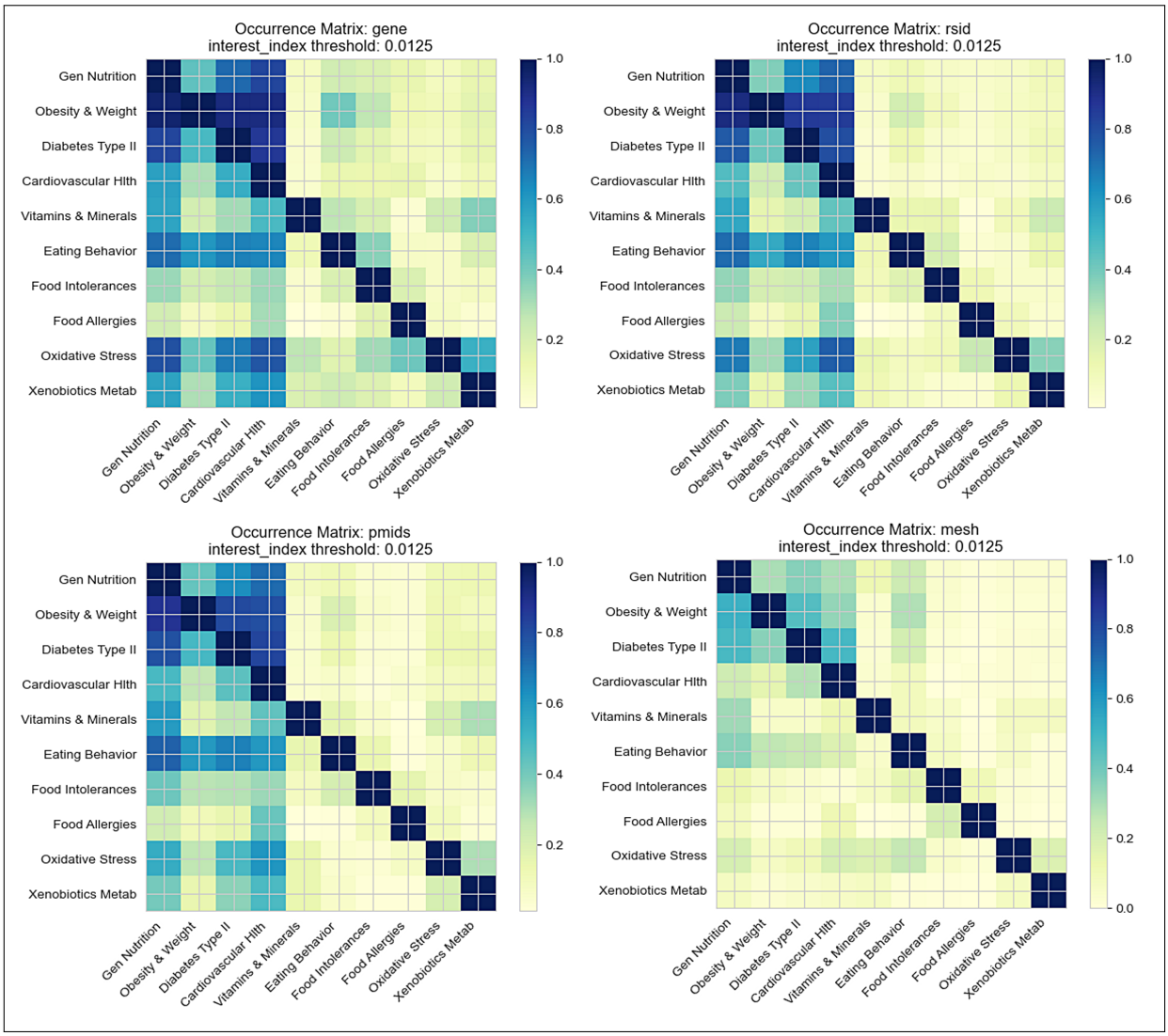

Figure 4: Co-occurrence pattern among genes, rsID, PMID, and MeSH terms observed in the filtered results (GI $>$ 0.0125) from the ten nutritional MeSH lists used for screening. The color key represents the amount of overlap between datasets scaled over the total size of each dataset.

merging, and analysis, facilitating comprehensive research in nutrigenomics. Our approach leverages data mining and data integration techniques to identify relevant studies on nutrition-associated genetic polymorphisms based on specific MeSH sets. The pipeline allowed us to extract subsets of genes and associated SNPs linked to nutrition-

medRxiv preprint doi: https://doi.org/10.1101/2023.08.04.23293659; this version posted June 21, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC 4.0 International license .

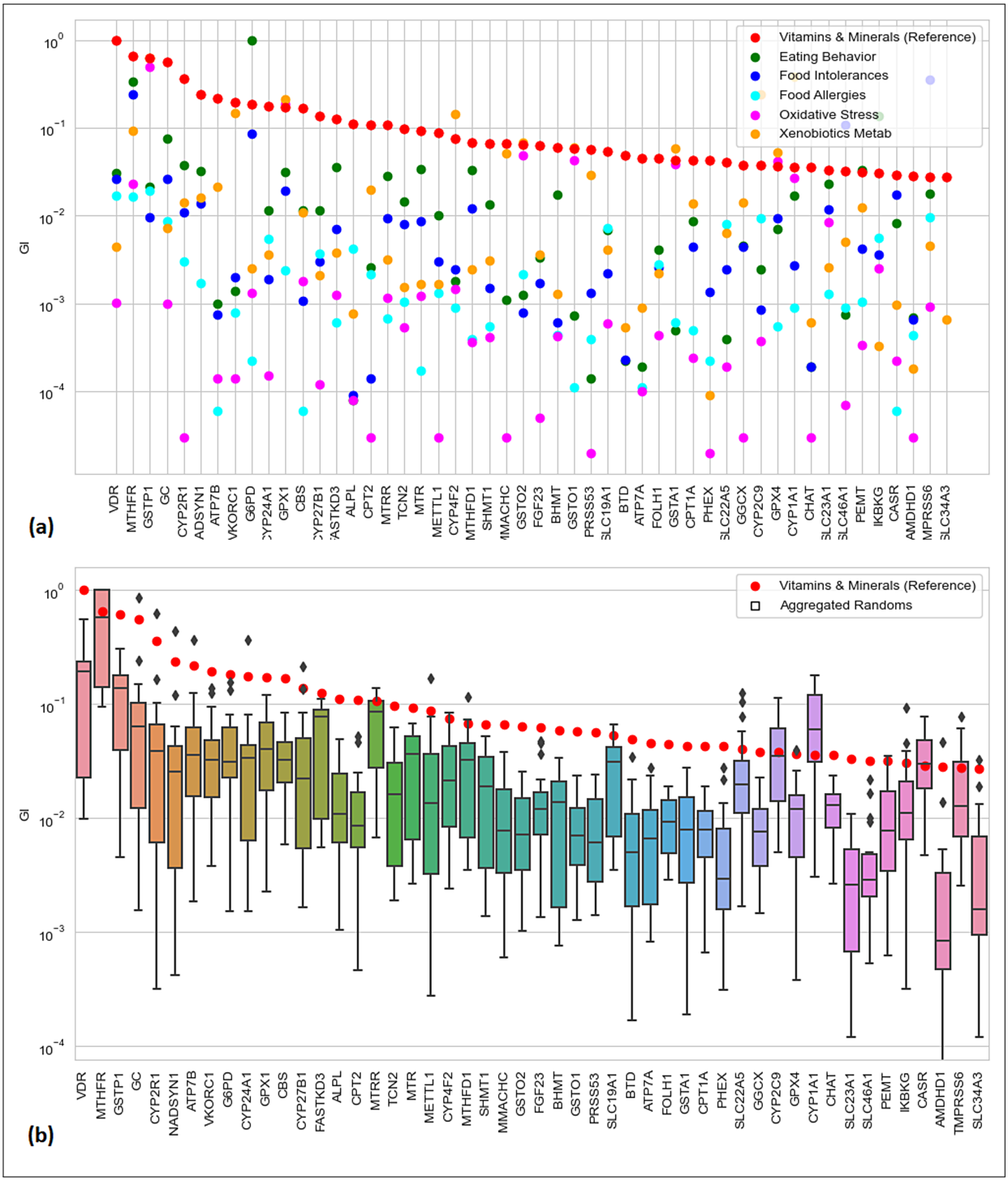

Figure 5: (a) Comparison of Gene Interest Index (GI) for the top fifty genes on the "Vitamin and Micronutrient Metabolism" MeSH list with GI obtained from five different lists of nutrition MeSH terms. (b) Comparison between GI obtained from the *"General Nutrition"* list and GI obtained using twenty 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.

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 [24]. 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 [24]. 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

medRxiv preprint doi: https://doi.org/10.1101/2023.08.04.23293659; this version posted June 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

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

GRPM Data				GWAS Data		
LITVAR GENE	**LIVAR RSID**	**LITVARPUBMED PMID**	**MeSH**	**MAPPED TRAIT**	**DISEASE-TRAIT**	**STRONGEST 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	24502231	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

specific genetic polymorphisms associated with various health or biological dysfunctions, empowering healthcare practitioners to tailor interventions based on an individual's genetic profile [25].

As remarked by Floris et al., 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 [26]. 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 [26]. 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 dataset, 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

medRxiv preprint doi: https://doi.org/10.1101/2023.08.04.23293659; this version posted June 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

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 [27]. 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 [28]. 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 [29]. 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 [30].

Furthermore, it is essential to acknowledge gene-environment interactions' complex and multifactorial nature, including interactions with dietary factors [31]. 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 [32]. 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 et al.[33] to ensure the reliability of individual sources.

5 Conclusion

Our study presents a comprehensive nutrigenetic dataset, constructed by integrating data from multiple sources using the MeSH ontology. This dataset is a valuable resource for exploring genetic polymorphisms associated with nutrition-related traits. By consolidating and standardizing genetic polymorphism data, our work aims to advance personalized nutrition interventions and contribute to the field of nutrigenomics.

The curated dataset fills a significant gap in the existing literature, providing a reliable and unified resource for investigating gene-diet interactions. It underscores the importance of standardized curation processes and highlights the role of translational bioinformatics in merging and analyzing information from diverse sources. By doing so, it facilitates comprehensive research in nutrition and genetics, offering a practical tool for researchers and nutritionists alike.

We hope this dataset will serve as a foundational resource for future nutrigenetic studies and help in the development of personalized nutrition strategies based on genetic insights.

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 &

medRxiv preprint doi: https://doi.org/10.1101/2023.08.04.23293659; this version posted June 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

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.

Acknowledgements

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

References

[1] Chen Y, Michalak M, Agellon LB. Importance of Nutrients and Nutrient Metabolism on Human Health. The Yale Journal of Biology and Medicine. 2018 Jun;91(2):95-103. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6020734/.

[2] Kiani AK, Bonetti G, Donato K, Kaftalli J, Herbst KL, Stuppia L, et al. Polymorphisms, diet and nutrigenomics. Journal of preventive medicine and hygiene. 2022;63(2):E125 E141. Type: Review. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85143558205&doi=10.15167%2f2421-4248%2fjpmh2022.63.2S3.2754&partnerID=40&md5=e98acce37426a90d596889f1ca8ae0a2.

[3] Loktionov A. Common gene polymorphisms and nutrition: Emerging links with pathogenesis of multifactorial chronic diseases (review). Journal of Nutritional Biochemistry. 2003;14(8):426 451. Type: Review. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0042476482&doi=10.1016%2fS0955-2863%2803%2900032-9&partnerID=40&md5=1cb80db054e82501f4ef0c2b542996fa.

[4] Mathers JC. Nutrigenomics in the modern era. Proceedings of the Nutrition Society. 2017;76(3):265 275. Type: Conference paper. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84994148988&doi=10.1017%2fS002966511600080X&partnerID=40&md5=8fd5734c6ed6cdaedc3b15e9dfa9c783.

[5] Kaput J. Nutrigenomics research for personalized nutrition and medicine. Current Opinion in Biotechnology. 2008;19(2):110 120. Type: Review. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-42249110345&doi=10.1016%2fj.copbio.2008.02.005&partnerID=40&md5=944c2f160a182ba0323cd7080add1c67.

[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. Journal of Nutrigenetics and Nutrigenomics. 2011;4(2):69 89. Type: Review. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-79957575438&doi=10.1159%2f000327772&partnerID=40&md5=1d407bb099d501ab47fa89a194817258.

[7] Comerford KB, Pasin G. Gene–dairy food interactions and health outcomes: A review of nutrigenetic studies. Nutrients. 2017;9(7). Type: Review. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85021946993&doi=10.3390%2fnu9070710&partnerID=40&md5=c9ec03abc1e1f1846c300f32e6f146de.

[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;110. Type: Review. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85150292199&doi=10.1016%2fj.nut.2023.112002&partnerID=40&md5=da66de1f8c8fed275c9f20c5c37a3577.

[9] Rinaldi AM. An ontology-driven approach for semantic information retrieval on the Web. ACM Transactions on Internet Technology. 2009;9(3). Type: Article. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-69149101349&doi=10.1145%2f1552291.1552293&partnerID=40&md5=e1caaf058622b6021b99da269e1eb5f3.

[10] Lee K, Wei CH, Lu Z. Recent advances of automated methods for searching and extracting genomic variant information from biomedical literature. Briefings in Bioinformatics. 2021;22(3). Type: Review. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85107087851&doi=10.1093%2fbib%2fbbaa142&partnerID=40&md5=09219fc8d2da02c74d32d05bcb6cab17.

[11] 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 Research. 2018;46(W1):W530 W536. Type: Article. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85050875921&doi=10.1093%2fnar%2fgky355&partnerID=40&md5=33e94302dc9be940dd29a33b080d1860.

[12] 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 Research. 2017;45(D1):D896 D901. Type: Article. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85016161935&doi=10.1093%2fnar%2fgkw1133&partnerID=40&md5=7f7ed903fb90e2d5af8d14cb0cf7b649.

[13] NCBI. Use of MeSH in Cataloging. nlm.nih.gov; 2020. Available from: https://wayback.archive-it.org/org-350/20200228165129/https://www.nlm.nih.gov/tsd/cataloging/MeSH_CatPractices.html.

[14] 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 2091. Type: Article. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84990978870&doi=10.1007%2fs11192-016-2119-7&partnerID=40&md5=e5dec3162ad346859121cbc56b840acc.

[15] Khare R, Leaman R, Lu Z. Accessing biomedical literature in the current information landscape. Methods in Molecular Biology. 2014;1159:11 31. Type: Article. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84927517824&doi=10.1007%2f978-1-4939-0709-0_2&partnerID=40&md5=f7a12d5bdf88f755a03a3e2dd421745f.

[16] Pimentel JF, Murta L, Braganholo V, Freire J. Understanding and improving the quality and reproducibility of Jupyter notebooks. Empirical Software Engineering. 2021;26(4). Type: Article. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85105602640&doi=10.1007%2fs10664-021-09961-9&partnerID=40&md5=dbcd40cf499e324efbc1f19e7d1bf586.

[17] Reddy VS, Palika R, Ismail A, Pullakhandam R, Reddy GB. Nutrigenomics: Opportunities & challenges for public health nutrition. Indian Journal of Medical Research. 2018;148(5):632 641. Type: Review. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85060205670&doi=10.4103%2fijmr.IJMR_1738_18&partnerID=40&md5=9ed75a00c8b589d390182954717b3ae2.

medRxiv preprint doi: https://doi.org/10.1101/2023.08.04.23293659; this version posted June 21, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

[18] O'Rahilly S, Farooqi IS. Human obesity as a heritable disorder of the central control of energy balance. International Journal of Obesity. 2008;32:S55 S61. Type: Review. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-58149485817&doi=10.1038%2fijo.2008.239&partnerID=40&md5=ef9c31008a2b70bf602d180b47c379d0.

[19] Wang DD, Hu FB. Precision nutrition for prevention and management of type 2 diabetes. The Lancet Diabetes and Endocrinology. 2018;6(5):416 426. Type: Review. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85044924586&doi=10.1016%2fS2213-8587%2818%2930037-8&partnerID=40&md5=e9a3477086610a0e05b59d4e34cae1e0.

[20] Peña-Romero AC, Navas-Carrillo D, Marín F, Orenes-Piñero E. The future of nutrition: Nutrigenomics and nutrigenetics in obesity and cardiovascular diseases. Critical Reviews in Food Science and Nutrition. 2018;58(17):3030 3041. Type: Review. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85061264857&doi=10.1080%2f10408398.2017.1349731&partnerID=40&md5=548508ce9560494c22a132fb9316f1f2.

[21] Di Camillo B, Giugno R. From translational bioinformatics computational methodologies to personalized medicine. Journal of Biomedical Informatics. 2022;133. Type: Editorial. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85136517464&doi=10.1016%2fj.jbi.2022.104170&partnerID=40&md5=cfd15b8444deba15ca6d5c8fa6a69702.

[22] Lee KH, Kim JH. Evolution of Translational Bioinformatics: Lessons learned from TBC 2016. BMC Medical Genomics. 2017;10. Type: Review. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85019545424&doi=10.1186%2fs12920-017-0262-5&partnerID=40&md5=2d404b3e9b41ae64f041fd3fd93e5081.

[23] Tenenbaum JD. Translational Bioinformatics: Past, Present, and Future. Genomics, Proteomics and Bioinformatics. 2016;14(1):31 41. Type: Review. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84969443677&doi=10.1016%2fj.gpb.2016.01.003&partnerID=40&md5=0e0b08f0ee43ab4c64042291e5f5b2a5.

[24] Zhang Y, Zhu Q, Liu H. Next generation informatics for big data in precision medicine era. BioData Mining. 2015;8(1). Type: Review. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84946400116&doi=10.1186%2fs13040-015-0064-2&partnerID=40&md5=69660314fb7287609b213e2cc4bbf847.

[25] Carrasco-Ramiro F, Peiró-Pastor R, Aguado B. Human genomics projects and precision medicine. Gene Therapy. 2017;24(9):551 561. Type: Review. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85032492708&doi=10.1038%2fgt.2017.77&partnerID=40&md5=46bbdba11ea3ba93950a33ae1fd2097e.

[26] Floris M, Cano A, Porru L, Addis R, Cambedda A, Idda ML, et al. Direct-to-consumer nutrigenetics testing: An overview. Nutrients. 2020;12(2). Type: Review. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85079874650&doi=10.3390%2fnu12020566&partnerID=40&md5=184519f71b694c531f5f1613056b5ba8.

[27] Carter AO, Griffin GH, Carter TP. A survey identified publication bias in the secondary literature. Journal of Clinical Epidemiology. 2006;59(3):241 245. Type: Article. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-32644451903&doi=10.1016%2fj.jclinepi.2005.08.011&partnerID=40&md5=1ada5474382f671ffa59fb8640ce4de5.

[28] Lachance J, Tishkoff SA. SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. BioEssays. 2013;35(9):780 786. Type: Article. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84881616630&doi=10.1002%2fbies.201300014&partnerID=40&md5=a38c848268be465d7f5747ce5300ee0c.

[29] Williams WG. Uses and limitations of registry and academic databases. Seminars in Thoracic and Cardiovascular Surgery: Pediatric Cardiac Surgery Annual. 2010;13(1):66 70. Type: Article. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-77949612067&doi=10.1053%2fj.pcsu.2010.02.007&partnerID=40&md5=38bd56ddf29efef51fdc5dd92c76e85c.

[30] Nilsson PD, Newsome JM, Santos HM, Schiller MR. Prioritization of Variants for Investigation of Genotype-Directed Nutrition in Human Superpopulations. International Journal of Molecular Sciences. 2019;20(14). Type: Article. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85071875349&doi=10.3390%2fijms20143516&partnerID=40&md5=6201b77e51ace4325543afd40c07d9a9.

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

[32] Cole BS, Hall MA, Urbanowicz RJ, Gilbert-Diamond D, Moore JH. Analysis of Gene-Gene Interactions. Current Protocols in Human Genetics. 2017;95(1):1.14.1 1.14.10. Type: Review. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85050423617&doi=10.1002%2fcphg.45&partnerID=40&md5=4fd9f9e95a157f00ecf47e9364e2ba48.

[33] 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 and Nutrition. 2017;12(1). Type: Review. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85038113677&doi=10.1186%2fs12263-017-0584-0&partnerID=40&md5=cc62bdd389b7ac3408c3ac4dc939d84e.