medRxiv preprint doi: https://doi.org/10.1101/2023.08.04.23293659; this version posted November 13, 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 The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.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[1], **Maria Monticelli**[2], **Alessandra Pollice**[3], **Tiziana Angrisano**[3], **Bruno Hay Mele**[3], **Viola Calabrò**[3]

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

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

[3] Department of Biology, University of Napoli "Federico II", Complesso Universitario Monte Sant'Angelo, Via Cinthia 26, 80126 Napoli, Italy

Abstract

Objective This study aims to create 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. 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 queries in order to retrieve relevant genetic associations, and cross-referencing the results with the GWAS data.
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 GWAS data 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 in this study consolidates and organizes information on genetic polymorphisms associated with nutrition, facilitating detailed exploration of gene-diet interactions. This resource advances personalized nutrition interventions and nutrigenomics research. The dataset is publicly accessible at `https://zenodo.org/records/14052302`, its adaptable structure ensures applicability in a broad range of genetic investigations.

Keywords

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

Correspondence to:

Bruno Hay Mele
Biology Dept., University of Naples Federico II
Via Cinthia, 26 Naples, Italy
bruno.haymele@unina.it

EJBI ...
recieved:
accepted:
published:

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

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 November 13, 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 .

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]. 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 comprehensive and structured 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 PubMed 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 resulting dataset was validated 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

We developed an integrated dataset by cross-referencing genomic data with scientific literature using shared PubMed IDs to link the LitVar and PubMed databases. This linkage allowed us to enrich LitVar's association data by incorporating Medical Subject Headings (MeSH).

The *MeSH Ontology* is a structured and controlled vocabulary that supports the annotation and indexing of biomedical literature and datasets. It provides standardized descriptors that facilitate the organization and retrieval of scientific information within the domains of biomedicine and healthcare informatics. The structured nature of MeSH terms enhances their applicability and robustness for comprehensive information retrieval and analysis of scientific indicators. In a database context, MeSH terms function as metadata references, enabling precise literature categorization and search [13].

The LitVar database is an extensive and publicly accessible repository that aggregates information on genetic variants and links them with corresponding scientific literature. Its purpose is to address the challenge of connecting genomic data with the relevant literature by synthesizing information on genetic variants from diverse sources, employing NER techniques [11]. PubMed operates as the primary digital library for biomedical literature, offering an invaluable resource for scientific inquiry. Data extraction from PubMed is essential for various research activities, including systematic reviews, data mining, and knowledge discovery [14].

To construct our dataset, we implemented a Python-based data processing pipeline designed to integrate heterogeneous data sources by utilizing the MeSH ontology as a central schema for data harmonization. This pipeline allows extraction, integration, and querying of genetic polymorphism information pertinent to specific research domains, such as nutrigenetics. The output dataset, referred to as GRPM dataset, encapsulates primary identifiers, including Genes, Reference SNP IDs (RsIDs), PubMed IDs (PMIDs), and MeSH terms. Recognizing the pivotal role of genetic influences in nutrition-related traits, the GRPM dataset facilitates comprehensive exploration and analysis, thereby supporting researchers and nutritionists in advancing personalized nutrition studies.

The data retrieval and integration pipeline is implemented in a Jupyter Notebook environment [15] and comprises five distinct modules, each tailored to perform specific functions essential for the assembly and utilization of a comprehensive genetic polymorphism dataset (as illustrated in Figure 1). These modules are outlined as follows:

1. *GRPM Dataset Building*: This module orchestrates the extraction, integration, and consolidation of data from source databases, including LitVar and PubMed. It ensures the comprehensive collection of genetic polymorphisms associated with topic-related traits (GRPM dataset).

 ©2011 EuroMISE s.r.o.

medRxiv preprint doi: https://doi.org/10.1101/2023.08.04.23293659; this version posted November 13, 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 .

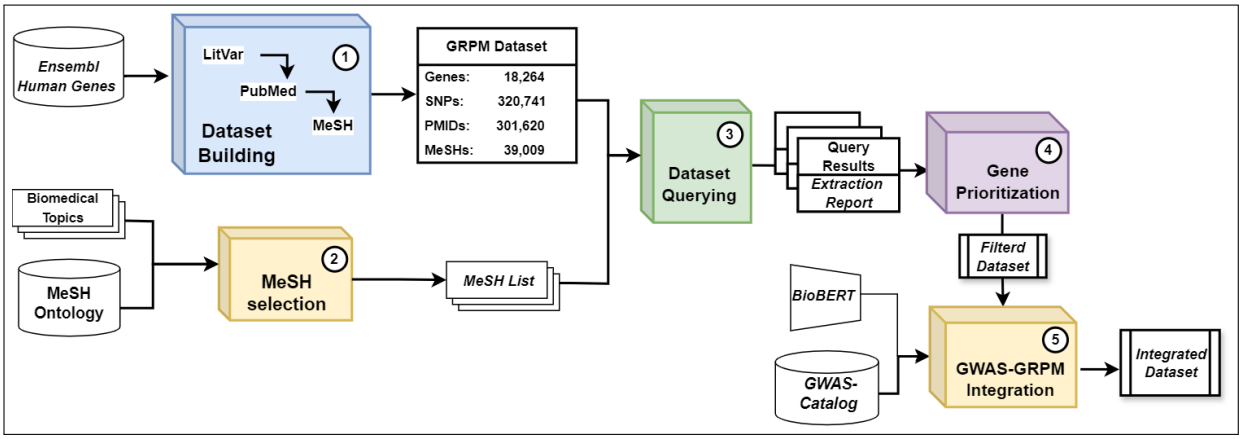


Figure 1: Graphical overview of our workflow showcasing the input data and interactions between the five modules.

2. *MeSH Term Selection*: This module facilitates the definition of MeSH ontology term sets. These are employed to query the GRPM dataset.

3. *GRPM Dataset Querying*: This module allows users to query the dateset using predefined MeSH terms. Users can thus refine their searches to focus on specific areas of interest, ensuring targeted data retrieval pertinent to their research needs within the GRPM dataset.

4. *Gene Prioritization*: This module assigns a relative importance score to each gene based on the frequency and proportion of associated findings. This metric enables the prioritization of genes pertinent to specific research topics, aiding further investigation.

5. *GWAS Data Integration*: The integration of the GWAS catalog provides additional insights into effect and risk alleles associated with the identified genetic polymorphisms. This process enriches our nutrigenetic resource with supplemental data for further analysis.

All implementation details, along with usage instructions, are available in the open-access repository 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 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 LitVar 2.0 [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 LitVar first version was the result of a thorough cost-benefit analysis, weighing the potential superior data precision provided by LitVar 2.0, 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 Query definition

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 from the complete MeSH ontology[4] (348,733 terms). Therefore, this subset collects ontology terms linked to papers exploring the associations between genetic variants and biomedical traits.

The second module is designed to select from the MeSH ontology a set of terms that represent a specific biomed-

[1] `https://github.com/johndef64/GRPM_system`
[2] `https://www.ncbi.nlm.nih.gov/research/litvar2/api`
[3] `https://pypi.org/project/nbib/`
[4] The complete MeSH 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 November 13, 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 .

ical fields. Initially, we constructed a collection of topic-specific terms based on our domain knowledge in nutrigenetics. These topic-relevant terms were then processed to identify corresponding MeSH terms using a straightforward Natural Language Processing (NLP) approach.Subsequently, the extracted MeSH terms underwent a filtering process based on their defined "semantic types". This step was followed by a manual screening to remove any ambiguous or potentially biased terms. This rigorous filtration ensures that only the most relevant terms are employed for dataset querying.

2.3 GRPM Dataset querying

In the third module, the predefined MeSH list is utilized to execute targeted queries on the GRPM Dataset, aimed at retrieving association data pertinent to genetic research (*"Survey Dataset"*).

Upon execution of the queries, an extraction report is compiled, aggregating the occurrence statistics from various data sources (NCBI, LitVar and PubMed) for each retrieved gene. These reports are subsequently subjected to both individual and comparative analyses in the fourth module.

2.4 Nutrigenetic Dataset Building

In order to build a nutrigenetic dataset, we defined ten major topic of interest covering nutrition physiology, nutrition-related diseases, disease prevention through diet, and eating behavior. For each of these topics, following the procedure described in the 2.2 section, we prepared MeSH queries to submit to our dataset. The topics delineated include:

1. *General Nutrition*: Encompasses a broad range of issues concerning dietary patterns, nutritional status, and overall health.
2. *Obesity, Weight Control, and Compulsive Eating*: Focuses on weight management and related disorders.
3. *Cardiovascular Health and Lipid Metabolism*: Related to the impact of diet on heart health and lipid levels.
4. *Diabetes Mellitus Type II and Metabolic Syndrome*: Covers dietary interventions and metabolic complications.
5. *Vitamin and Micronutrients Metabolism and Deficiency-Related Diseases*: Involves the metabolism and health impact of vitamins and micronutrients.
6. *Eating Behavior and Taste Sensation*: Related to food choices, taste preferences, and appetite regulation.
7. *Food Intolerances*: Addresses adverse reactions to specific foods and their genetic basis.
8. *Food Allergies*: Explores genetic aspects and dietary management strategies for allergies.
9. *Diet-induced Oxidative Stress*: Investigates how diet influences oxidative stress and health outcomes.
10. *Xenobiotics Metabolism*: Focuses on how the body processes foreign substances like drugs and toxins.

For more detailed descriptions and the MeSH count associated with each topic, please refer to Table S1 in the Supplementary Materials.

The nutrigent dataset was then constructed by merging the results of the 10 queries and filtering them using a metric based on the number and specificity of publications associated with each gene. Then, we calculated for each gene a score (**Gene Interest Index (GI)**), considering potentially "interesting" a gene if its related SNPs are associated with a substantial number of PMIDs (PubMed IDs) that include MeSH terms in the query and if the ratio between these PMIDs count and the total number of gene-associated PMIDs is sufficiently high.

To evaluate the pertinence of the retrieved gene set for the specified topic, it is essential to treat the employed MeSH set as a cohesive unit rather than analyzing the terms in isolation. This approach acknowledges the varying significance of each term within the context of the overall ontology. To determine if a gene is "interesting" based on its linked MeSH terms in LitVar-annotated studies, we propose normalizing the count of identified PMIDs by the total PMID count associated with that gene within the LitVar database. This methodology aids in reducing selection bias that may arise from genes that are extensively researched and thus associated with a larger number of MeSH terms, which might not be directly relevant to the specific query at hand.

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.

 ©2011 EuroMISE s.r.o.

medRxiv preprint doi: https://doi.org/10.1101/2023.08.04.23293659; this version posted November 13, 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 .

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

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

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, 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 cut-off 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.

2.5 GWAS data integration

The exploration of each study to discern the associated effect allele for every SNP ID presents a challenge due to its time-intensive nature. An initial assessment of the potential effect allele is instrumental for conducting preliminary investigations. To streamline this process, we integrated the Ensembl GWAS Catalog data[5] [12] into our nutrigenetics dataset, in the fifth module in our data pipeline.

To incorporate GWAS data, we leveraged the BioBERT language model, a state-of-the-art tool specifically tailored for biomedical text mining [16]. BioBERT, an adaptation of BERT, is designed to capture and represent complex biomedical concepts within a high-dimensional vector space. This enhanced representation is instrumental in discerning semantic relationships within biomedical terminology. This approach for semantic annotation and entity linkage draws on methodologies similar to those evaluated in Tutubalina et al.'s work on information extraction using BERT models for biomedical search engines [17]. We employed BioBERT-generated embeddings for both MeSH terms and GWAS trait descriptions to assess semantic similarity. This was achieved by calculating the cosine similarity between these numerical vectors, allowing for the identification of closely related entities. Only associations that scored above a 90% similarity threshold were retained, following a thorough manual verification of their correctness. This step provided us with a reliable correlation map between MeSH concepts and GWAS traits or diseases. Subsequently, we merged the enriched GWAS Catalog data with our nutrigenetic dataset, aligning them based on SNP identifiers. The integration was further processed using the correlation map, ensuring that only highly relevant associations were included. As result, the enhanced dataset includes information on the most significant SNP-risk alleles linked to each Reference SNP ID (rsID).

Additionally, the alignment of GWAS associations acts as a quality validation mechanism for the proposed methodology based on MeSH terms, counteracting potential biases often introduced by text mining. The inclusion of a MeSH term does not inherently verify that the cited SNP is biologically associated with the queried phenotype; rather, it ensures citation within the same study, securing more accurate phenotype-genotype connections. By incorporating GWAS data, we can mitigate these biases, ensuring that the associations identified through MeSH-driven queries align with validated genetic correlations within the domain of personalized nutrition research.

3 Results

As result, using the outlined workflow, we have constructed a primary dataset, termed GRPM, from which we have extracted a nutrigenetics-specific subset using custom-designed MeSH-based queries. This dataset was filtered using the GI metric to select the most relevant results. Subsequently, it was augmented with GWAS data through semantic integration techniques. Consequently, the outcomes are three datasets of progressively smaller size:

1. *Primary GRPM Dataset*: Dimensions (16,610,132 rows, 6 columns)
2. *Nutrigenetic Subset*: Dimensions (1,171,249 rows, 6 columns)
3. *Nutrigenetic Dataset Augmented with GWAS*: Dimensions (179,664 rows, 15 columns)

In Tables 1 through 2 and 3, examples of entries for the three datasets along with their respective data dictionaries are provided. These datasets are stored on Zenodo (`https://zenodo.org/records/14052302`) and can

[5] `https://www.ebi.ac.uk/gwas/docs/file-downloads`

medRxiv preprint doi: https://doi.org/10.1101/2023.08.04.23293659; this version posted November 13, 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 .

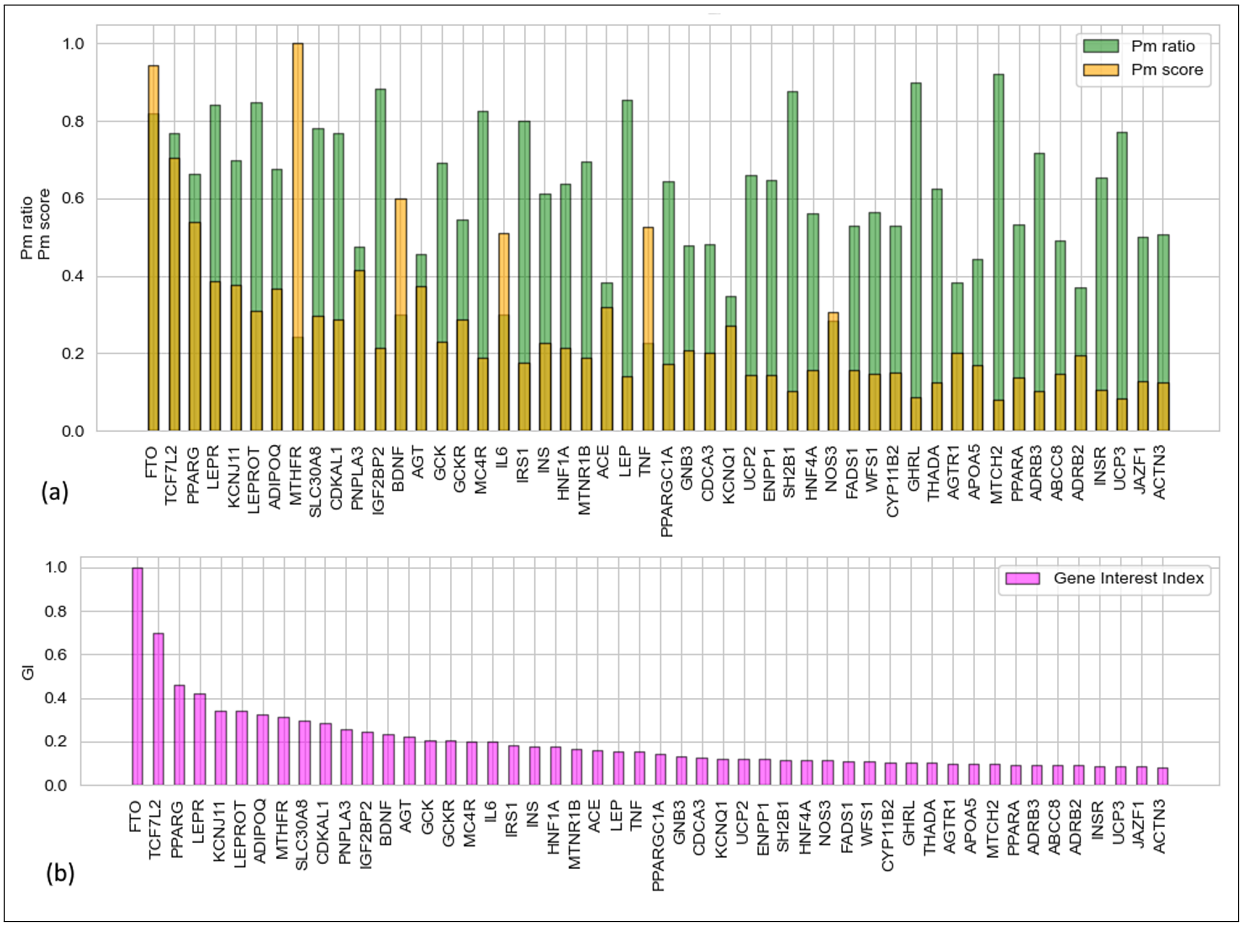


Figure 2: Visual representation of Gene Interest Index (GI) calculation though P_m score and P_m ratio, using as reference the results from *"Obesity and Weight Control"* MeSH query (see results section). Panel (a) shows the matching PMID ratio and overall matching PMID score. Panel (b) displays the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

be easily accessed using the provided Jupyter Notebooks or by installing the Python package (`https://github.com/johndef64/GRPM_system`). Query examples can be found in the *'test'* directory of the package.

```
pip install git+https://github.com/johndef64/
    GRPM_system.git
```

Listing 1: Python package installation

LitVar is a significant resource developed through extensive NER of genetic literature. However, the LitVar API (`https://www.ncbi.nlm.nih.gov/research/litvar2/api`) currently supports queries using gene symbols or variants (rsID), but does not facilitate retrieval based on phenotypic effects. Our research addresses this limitation by enhancing the LitVar dataset with MeSH ontology and GWAS phenotypes. This integration facilitates targeted retrieval of variants from indexed literature, enabling the construction of a specialized nutrigenetics dataset.

GRPM Dataset Figure 3 presents a compositional and comparative overview of the primary GRPM dataset. This dataset comprises approximately 77% of the PMIDs retrieved from the LitVar database (as of June 2023). These PMIDs have been associated with MeSH terms sourced from PubMed, facilitating enhanced semantic annotation and interoperability.

Employing this dataset, it is crucial to consider the PMID count associated with each gene. Genes with higher research prominence are linked to a greater number of PMIDs, resulting in an increased number of MeSH annotations. Therefore, to ensure proper data normalization and analysis, it is vital to comprehensively identify and understand the most frequently represented genes within the dataset (Figure 3b).

Nutrigenetic dataset Our study aimed to create a *nutrigenetic dataset* using a collection of MeSH terms related to different aspects of nutrition.

Table 2 provides a statistical overview of the data sourced from nutritional MeSH-based queries executed on the GRPM dataset. Given the extensive scope of our dataset, which includes a wide range of genes and associated MeSH terms, identifying genes that are most pertinent to specific research objectives presents a notable challenge. To address this complexity, this primary results are refined using a Gene Interest (GI) metric threshold of 0.0125. This filtering criterion is crucial for isolating

medRxiv preprint doi: https://doi.org/10.1101/2023.08.04.23293659; this version posted November 13, 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: GRPM Dataset description

Data Dictionary

Field	Description	Type	Example
gene	Gene symbol	string	GLIS3
type	Gene type*	string	PCG
rsid	Reference SNP ID	string	rs779079998
pmid	PubMed ID for associated literature	string	17076841
mesh	Medical Subject Headings	string	"Diabetes Mellitus, Type 2"
qualifier	Qualifier for MeSH term	string	genetics
major	Indicates if the MeSH term is a major topic	boolean	True

Dataset Sample entry

gene	type	rsid	pmid	mesh	qualifier	major
DNMT3A	PCG	rs377577594	25315154	DNA Methylation	drug effects	True
GLIS3	PCG	rs7020673	24060607	Diabetes Mellitus, Type 2	genetics	True
CFB	PCG	rs779079998	25222132	NIH 3T3 Cells		False
EYA1	PCG	rs3779747	23840632	Protein Tyrosine Phosphatases	metabolism	False
ABCB1	PCG	rs2032582	33326171	Ovariectomy		False

* PCG: protein coding genes; RNA: RNA genes; PSD: pseudogenes

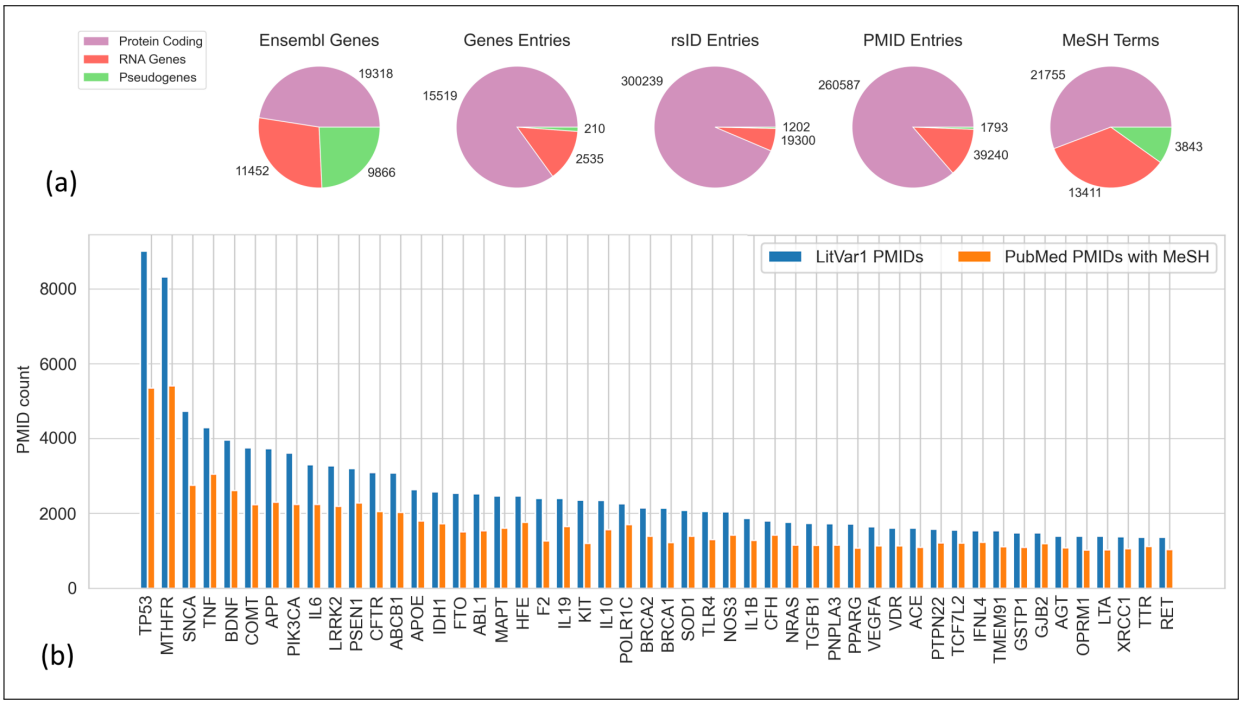


Figure 3: Overview of the Primary GRPM Dataset. Panel (a) shows the distribution patterns of genes, rsID, PMID, and MeSH terms for each entry type. Panel (b) identifies the 50 most frequent genes within the dataset, alongside their referenced PMID counts and correlated MeSH terms.

the most significant genetic associations, thereby enhancing their utility in nutrigenetics research. By leveraging this data-driven approach, we have effectively prioritized the extraction of the most pertinent and impactful data, facilitating the development of our final, refined dataset.

Nutrigenetic GWAS dataset Cross-referencing data between our refined nutrigentic dataset and GWAS catalog provides indicative information about possible risk alleles associated with the collected studies. Table 3 shows the data dictionary and sample entries of this cross-dataset, displaying data from the genetic literature alongside the GWAS data associated to the same rsID, such as Mapped Trait and Strongest Risk Allele. The resulting dataset comprises 385 unique genes with mapped data. Additionally, it includes 126 distinct MeSH terms, 316 mapped GWAS traits, and identifies 1,211 strongest SNP-risk alleles. This structured dataset aligns with our objective to facilitate precise and targeted retrieval in the nutrigenetics domain.

medRxiv preprint doi: https://doi.org/10.1101/2023.08.04.23293659; this version posted November 13, 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: Nutrigenetic Datasest structure and statistics

Data Dictionary

Attribute	Description	Type	Example
gene	Gene symbol	string	TCF7L2
rsid	Reference SNP ID	string	rs7901695
pmid	PubMed ID	string	17311858
mesh	Medical Subject Headings	string	Body Height
topic	Nutrigenetic topic of interest	string	Obesity, Weight Control and Compulsive Eating
interest_index	Gene Interest Index, normalized score	float	0.69794

Nutrigenetic Dataset Entry Sample

gene	rsid	pmid	mesh	topic	interest_index
PCSK9	rs970575319	24115837	Receptors, LDL	Cardiovascular Health and Lipid Metabolism	0.41714
NOS3	rs1799983	30738311	Nitric Oxide	Diet-induced Oxidative Stress	0.6236
IL18	rs1946518	31835423	Insulin Resistance	General Nutrition	0.03126
THADA	rs7578597	25667308	Diabetes Mellitus, Type 2	Cardiovascular Health and Lipid Metabolism	0.08387
MIA3	rs17465637	28686695	Inflammation	Diabetes Mellitus Type II and Metabolic Syndrome	0.06364

Statistics

	MeSH query results				Filtered values (GI > 0.0125)			
topic	**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
Unique values	**13,955**	**160,578**	**135,737**	**1,593**	**1,773**	**63,531**	**99,759**	**1,486**

3.1 Data Analysis

To elucidate the relationships within our dataset, we analyzed the prevalence of key genes and their associated MeSH terms and PMIDs across the ten nutrigenetic topics investigated in our research. This analysis, detailed in more detail in the Supplementary Material (see Figure S2), provides a comparative framework to assess the relevance of genes for various nutrition-related traits. In Figure S2, the richness of associated MeSH terms and PMIDs for the top 50 genes is highlighted, enabling an enhanced understanding of gene relevance within each topic.

To evaluate the extent of data overlap from the ten nutritional topics, we constructed co-occurrence matrices. Figure 4 illustrates the co-occurrence patterns among genes, rsID, PMID, and MeSH terms from the resulting nutrigenetic dataset. In this correlation analysis, we computed relative values for each topic by evaluating the ratio of shared entities to the total number of entities. This methodology allows for a detailed examination of co-occurrence within specific categories, facilitating a nuanced understanding of relationships and interactions. Remarkably, the correlation matrix in Figure 4 reveals substantial overlap between various MeSH term lists, such as the topic of 'obesity', which shares approximately 80% of its items with the topic of 'diabetes' and 'cardiovascular health', demonstrating a significant 40-60% overlap in this cluster of three categories.

3.2 Method Validation

Figure 5a shows the 50 most interesting genes for the topic *"Vitamin and Micronutrient Metabolism"* compared to the values of interest on results obtained from 5 other nutrigentic topics. The genes extracted using our method show specificity for the topic used as a reference. This behavior suggests that our method can identify genes related to that particular nutritional aspect. It can be seen in Figure 5a that some of the genes have higher GI in other topics 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 (Figures S3, S4). To ensure the accuracy and reliability of the data collected, we compared the results obtained with biologically consistent MeSH queries with those obtained with 20 random MeSH queries[6] of the same size. Figure 5b provides an example of the comparison results in the *"General Nutrition"* topics. Another example is shown in Supplementary Materials (S4).

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

 ©2011 EuroMISE s.r.o.

medRxiv preprint doi: https://doi.org/10.1101/2023.08.04.23293659; this version posted November 13, 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: Nutrigenetic-GWAS Cross-Dataset description and statistics

Data Dictionary

Field	Description	Type	Example
GRPM_GENE	Gene symbol	string	PCSK9
GRPM_RSID	Reference SNP ID	string	rs11591147
GRPM_PMID	PubMed ID	string	29083407
GRPM_MESH	Medical Subject Headings (MeSH) term	string	Coronary Artery Disease
GRPM_TOPIC	Research topic or category	string	General Nutrition
GRPM_GI	Gene Interest Index (GI)	float	0.2677
SEMANTIC_SIMILARITY	Semantic similarity score	float	0.9777
GWAS_DISEASE/TRAIT	GWAS reported disease or trait	string	Coronary artery disease
GWAS_CONTEXT	GWAS context such as variant type	string	missense_variant
GWAS_STRONGEST_SNP-RISK_ALLELE	Strongest SNP-risk allele identified in GWAS	string	rs11591147-T
GWAS_OR-BETA	Odds ratio or beta value from GWAS	float	0.2406
GWAS_STUDY_ID	Identifier for the GWAS study	string	GCST005195

Dataset Entry Sample

Gene	RSID	PMID	MESH	Topic	GI	Semantic Similarity	GWAS Disease /Trait	GWAS Context	GWAS SNP-Risk Allele	GWAS OR/-Beta	GWAS Study ID
PCSK9	rs11591147	29083407	Coronary Artery Disease	General Nutrition	0.2677	0.9777	Coronary artery disease	missense variant	rs11591147-T	0.2406	GCST005195
TM6SF2	rs58542926	26409295	Non-alcoholic Fatty Liver Disease	Cardiovascular Health	0.3265	0.9218	Nonalcoholic fatty liver disease	stop gained	rs58542926-T	0.3900	GCST90104598
CNNM2	rs11191548	27251080	Hypertension	General Nutrition	0.0569	1.0000	Hypertension	3 prime UTR variant	rs11191548-C	0.8600	GCST011141
PCSK9	rs11591147	23226021	Cholesterol, LDL	Cardiovascular Health	0.4171	0.9110	Total cholesterol to lipids ratio in IDL	missense variant	rs11591147-T	0.2576	GCST90301976
LPA	rs3798220	23881580	Triglycerides	Cardiovascular Health	0.1575	0.9508	Triglyceride levels	missense variant	rs3798220-C	0.0921	GCST90019523
ALDH2	rs671	20518787	Alcohol Drinking	Eating Behavior and Taste Sensation	0.3146	0.9136	Alcohol consumption	missense variant	rs671-A	1.3877	GCST90104202
F2	rs1799963	26554832	Venous Thrombosis	Cardiovascular Health	0.4423	0.9343	Venous thromboembolism	3 prime UTR variant	rs1799963-A	0.7972	GCST90399745

Statistics

Topic	GRPM_GENE	GRPM_RSID	GRPM_PMID	GRPM_MESH	GWAS_DISEASE/TRAIT	GWAS_STUDY_ID
General Nutrition	214	551	3978	40	139	402
Obesity & Weight Control	97	257	1672	26	38	169
Diabetes Type II & MS	197	526	3477	46	333	683
Cardiovascular Health	242	610	4585	63	497	899
Vitamins & Minerals	16	31	1259	9	14	19
Eating Behavior	35	51	518	7	11	41
Food Intolerances	43	65	99	7	8	24
Food Allergies	44	55	90	11	17	30
Oxidative Stress	8	11	655	5	9	12
Xenobiotics Metab	10	14	1399	10	13	28
Unique Values	385	944	7278	126	588	1140

4 Discussion

Understanding how genetic variations influence individual nutritional requirements, metabolism, and health outcomes is crucial for developing personalized nutrition interventions [18]. 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 [19], diabetes [20], or cardiovascular diseases [21].

Personalized healthcare and prevention strategies are central to advancements in translational bioinformatics [22]. The creation of sophisticated computational methodologies and tools for aggregating and analyzing data from diverse sources significantly enhances data integration, enabling the transformation of biomedical findings into tailored nutritional guidance and precision nutrition strategies for disease prevention [23, 24].

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 [25]. 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 [25]. 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.

The resulting dataset is an integrated resource for diverse data sources related to genetic polymorphisms associated with nutrition. This resource enables efficient retrieval and analysis, facilitating comprehensive research

medRxiv preprint doi: https://doi.org/10.1101/2023.08.04.23293659; this version posted November 13, 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 .

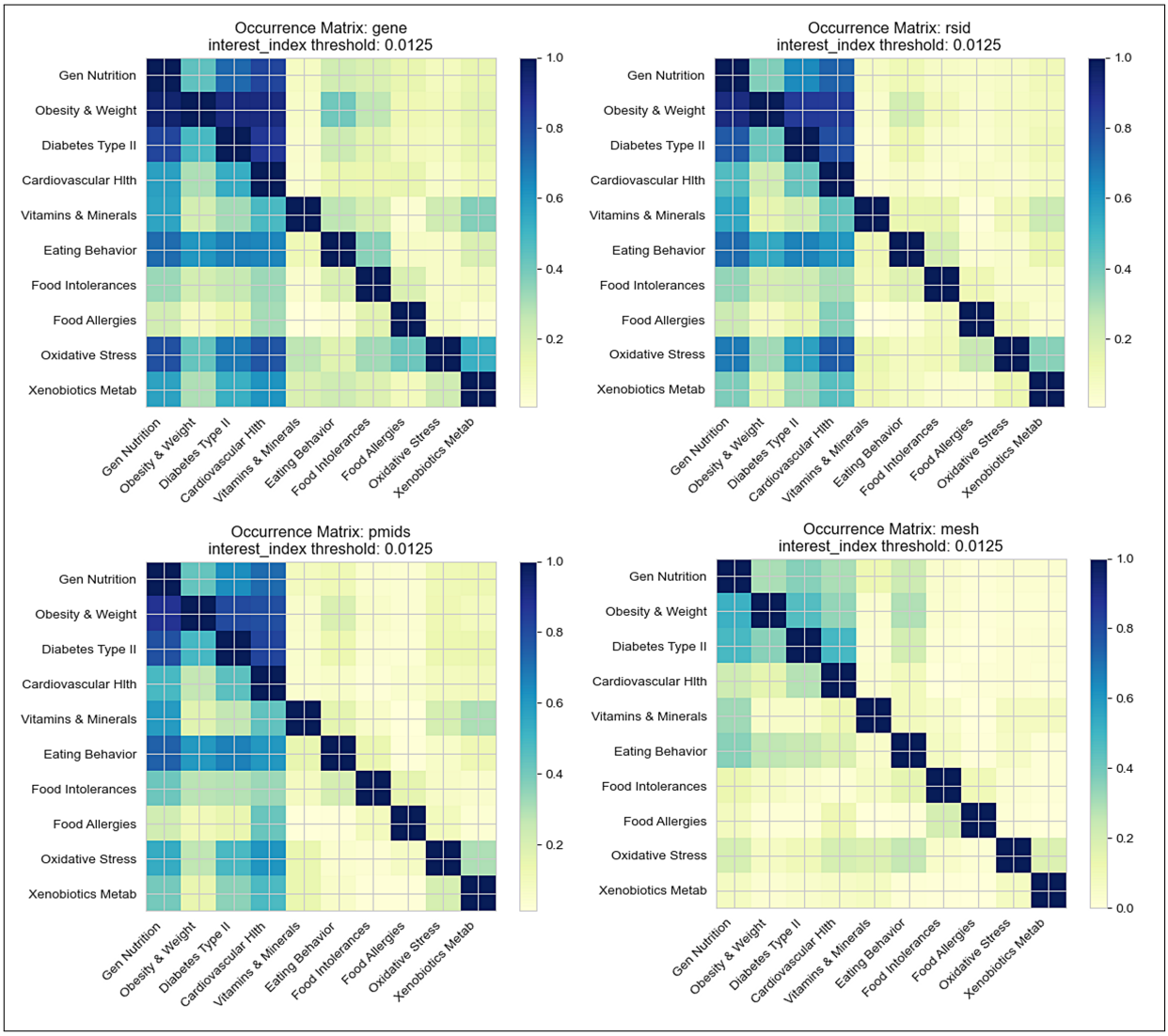


Figure 4: Co-occurrence patterns observed among genes, rsIDs, PMIDs, and MeSH terms in the refined dataset (GI > 0.0125) across ten nutritional topics. The color coding indicates the degree of overlap between datasets, normalized by the total size of each dataset.

in nutrigenomics. 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 [26].

In our nutrigenetic dataset, the heterogeneity of MeSH term representation across nutrigenetic topics suggests that some topic encompass a wider array of aspects within nutrition, while others are more specialized with a focused scope. Notably, the genes highlighted in our study, as derived from MeSH queries (as detailed in Table 2), are prominent in nutrigenetics research. These genes, frequently cited in the literature, consistently appear across multiple MeSH queries, underscoring their significance regarding nutrition-related traits.

From a broader standpoint, the analysis revealed a high degree of overlap between genes associated with specific nutritional topics (Figure 4). For example, 'obesity' exhibit high overlap with the 'diabetes' and 'cardiovascular health', 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 topics could be attributed to their specificity, focusing on more specific biological processes or conditions with distinct genetic underpinnings. This behavior can be attributed to several factors. Firstly, these topics may be described by fewer MeSH terms, leading to a narrower focus and less overlap with others. Secondly, these topics may pertain to more specific biological processes or conditions with distinct

 ©2011 EuroMISE s.r.o.

medRxiv preprint doi: https://doi.org/10.1101/2023.08.04.23293659; this version posted November 13, 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 .

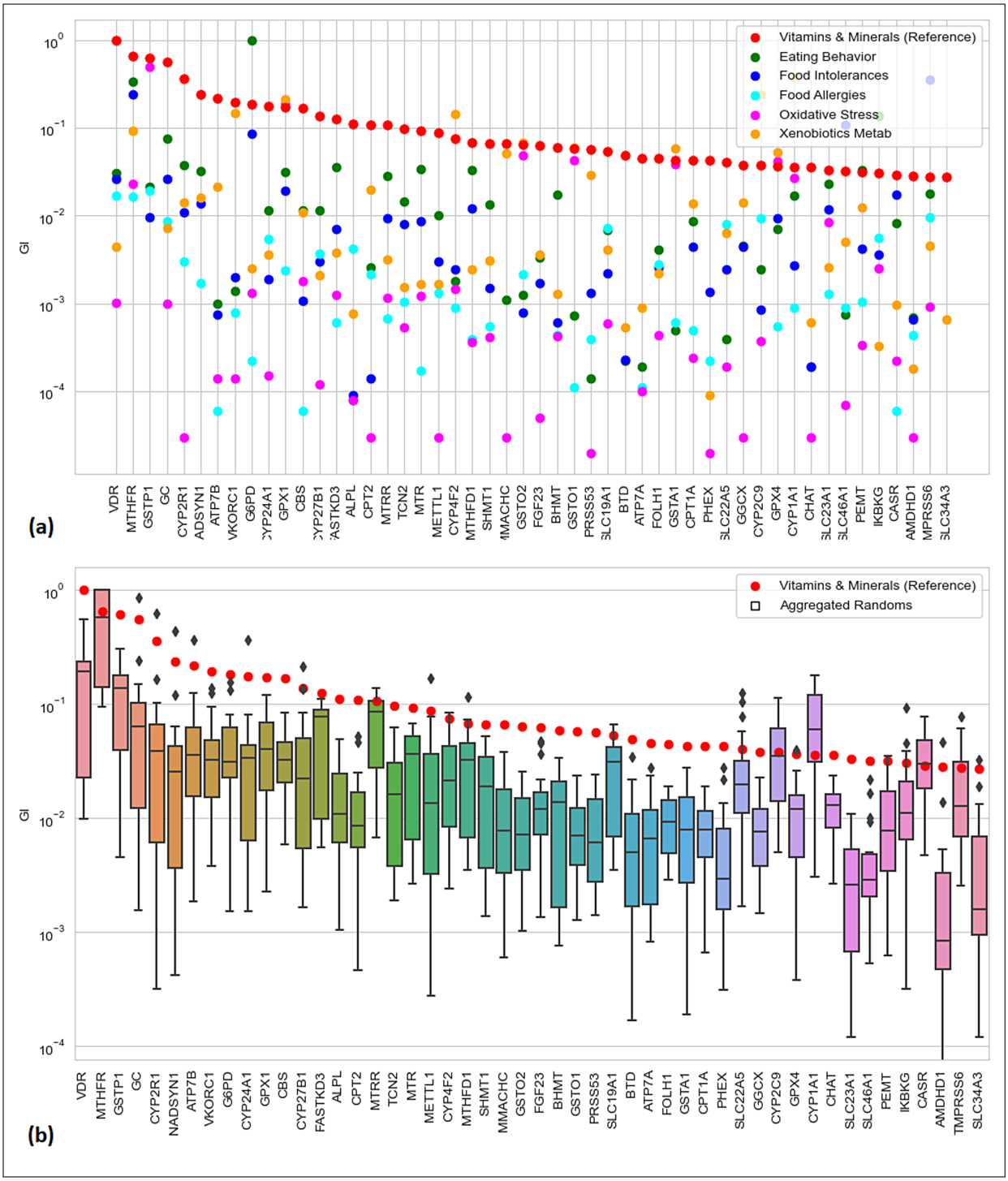


Figure 5: (a) Comparison of GI for the top 50 genes on the *"Vitamin and Micronutrient Metabolism"* topic with GI obtained from 5 different nutritional topics. (b) Comparison between GI obtained from the *"General Nutrition"* list and GI obtained using 20 randomly generated MeSH queries of the same size represented as boxplots. In both graphs the y-axis is logarithmic; genes are ordered by decreasing interest relative to the reference list.

genetic underpinnings than the broader conditions captured by the others. The integration of the nutrigenetic dataset with the GWAS data corroborated our findings by identifying potential risk alleles linked to specific genetic polymorphisms. Utilizing semantic similarity techniques, powered by the BioBERT language model, revealed a consistent alignment between MeSH terms associated with each PMID and the corresponding GWAS traits (Table 3). This congruence affirms that genetic associations identified through literature-based MeSH queries are reinforced by empirical evidence from GWAS studies.

To evaluate the validity and reliability of our results, we

medRxiv preprint doi: https://doi.org/10.1101/2023.08.04.23293659; this version posted November 13, 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 .

conducted a comparative analysis of the top-scoring genes across different nutritional queries, as depicted in Figure 5a. This validation step effectively demonstrated the capability of our method to accurately pinpoint genes pertinent to distinct nutritional topics. Additionally, we observed some genes exhibiting high relevance scores across multiple nutritional contexts, underscoring the intricate nature of gene regulation in nutrient metabolism and the necessity of considering various nutritional dimensions.

To validate the accuracy and reliability of our dataset, we assessed the outcomes generated from biologically coherent MeSH queries against those derived from random, non-cohesive queries (Figure 5b). Our analysis underscored that biologically consistent MeSH queries yielded significant and contextually relevant results for the nutritional aspects under investigation. In contrast, random MeSH queries produced non-significant outcomes with no meaningful associations. This comparison underscores the necessity of employing biologically pertinent MeSH queries for extracting and prioritizing data that is genuinely relevant to the specified biomedical domain.

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, inconsistencies, errors, and biases in the original studies may still be present in the constructed dataset.

Moreover, our dataset is limited to the data available in the LitVar database, GWAS-Catalog, and the 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 the complex and multifactorial nature of gene-environment interactions, including 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. Moreover, to improve the validity of the study 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.

In considering the dataset's future evolution, the framework is designed to accommodate ongoing advancements in nutrigenetics. The pipeline's modular architecture allows for systematic updates with new data sources and literature, ensuring the dataset remains at the cutting edge of scientific discovery. However, future updates must be meticulously curated and validated to preserve the dataset's integrity, acknowledging that emerging research may revise previously accepted insights. Implementing robust data governance structures is essential to mitigate inconsistencies and safeguard the dataset's overall reliability.

In future work, this dataset can be harnessed for the application of advanced NLP techniques, such as more recent semantic analysis methodologies. Currently, this dataset has already been emloyed for data-driven topic modeling and graph-based semantic analysis, which helps to uncover underlying themes within the nutrigenetics literature [34]. By employing text mining techniques, particularly those utilizing pretrained language models, the analysis of these datasets can be significantly enhanced. This will facilitate the identification of complex interactions between genes and dietary factors, allowing for the detection of patterns and correlations as well as the development of predictive models [35].

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 dataset, openly available, fills a significant gap in the existing resources in the field, 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.

 ©2011 EuroMISE s.r.o.

medRxiv preprint doi: https://doi.org/10.1101/2023.08.04.23293659; this version posted November 13, 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 .

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.

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.

[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.

[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.

[4] Mathers JC. Nutrigenomics in the modern era. Proceedings of the Nutrition Society. 2017;76(3):265 275. Type: Conference paper.

[5] Kaput J. Nutrigenomics research for personalized nutrition and medicine. Current Opinion in Biotechnology. 2008;19(2):110 120. Type: Review.

[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.

[7] Comerford KB, Pasin G. Gene–dairy food interactions and health outcomes: A review of nutrigenetic studies. Nutrients. 2017;9(7). Type: Review.

[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.

[9] Rinaldi AM. An ontology-driven approach for semantic information retrieval on the Web. ACM Transactions on Internet Technology. 2009;9(3). Type: Article.

[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.

[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.

[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.

[13] 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.

[14] Khare R, Leaman R, Lu Z. Accessing biomedical literature in the current information landscape. Methods in Molecular Biology. 2014;1159:11 31. Type: Article.

[15] 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.

[16] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234-40. Available from: `https://doi.org/10.1093/bioinformatics/btz682`.

[17] Tutubalina E, Miftahutdinov Z, Muravlev V, Shneyderman A. A Comprehensive Evaluation of Biomedical Entity-centric Search. In: Li Y, Lazaridou A, editors. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track. Abu Dhabi, UAE: Association for Computational Linguistics; 2022. p. 596-605. Available from: `https://aclanthology.org/2022.emnlp-industry.61`.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] Di Camillo B, Giugno R. From translational bioinformatics computational methodologies to personalized medicine. Journal of Biomedical Informatics. 2022;133. Type: Editorial.

[23] Lee KH, Kim JH. Evolution of Translational Bioinformatics: Lessons learned from TBC 2016. BMC Medical Genomics. 2017;10. Type: Review.

[24] Tenenbaum JD. Translational Bioinformatics: Past, Present, and Future. Genomics, Proteomics and Bioinformatics. 2016;14(1):31 41. Type: Review.

[25] 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.

medRxiv preprint doi: https://doi.org/10.1101/2023.08.04.23293659; this version posted November 13, 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 .

[26] Carrasco-Ramiro F, Peiró-Pastor R, Aguado B. Human genomics projects and precision medicine. Gene Therapy. 2017;24(9):551 561. Type: Review.

[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.

[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.

[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.

[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.

[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.

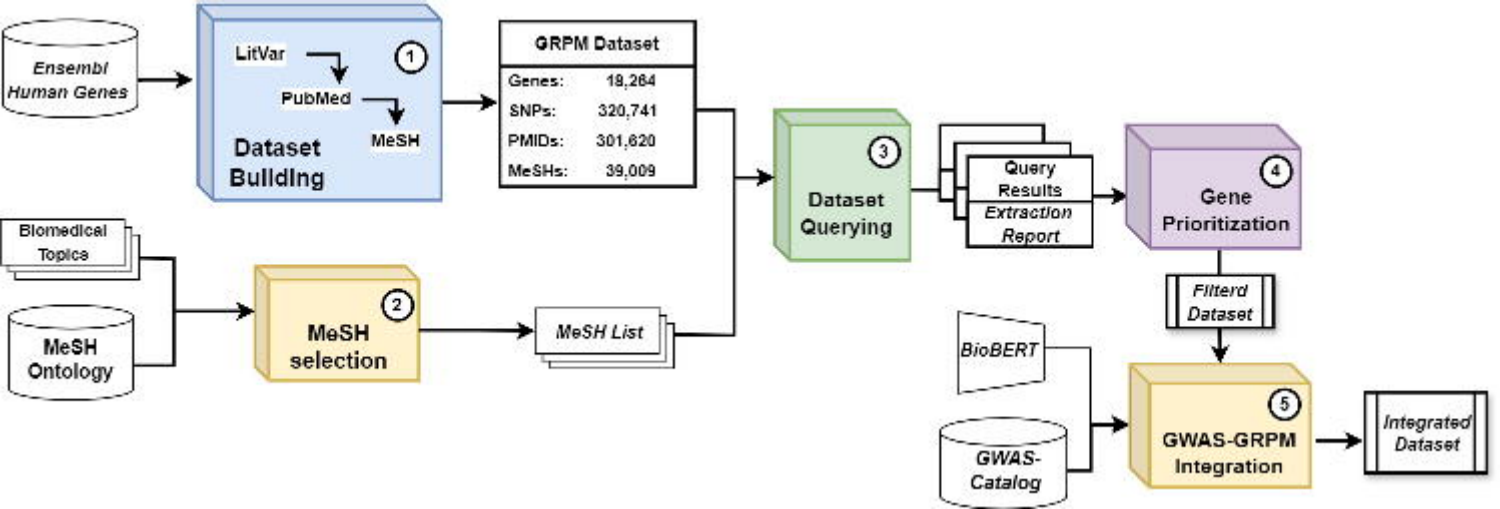
[32] Cole BS, Hall MA, Urbanowicz RJ, Gilbert-Diamond D, Moore JH. Analysis of Gene-Gene Interactions. Current Protocols in Human Genetics. 2017;95(1):1.14.1 1.14.10. Type: Review.

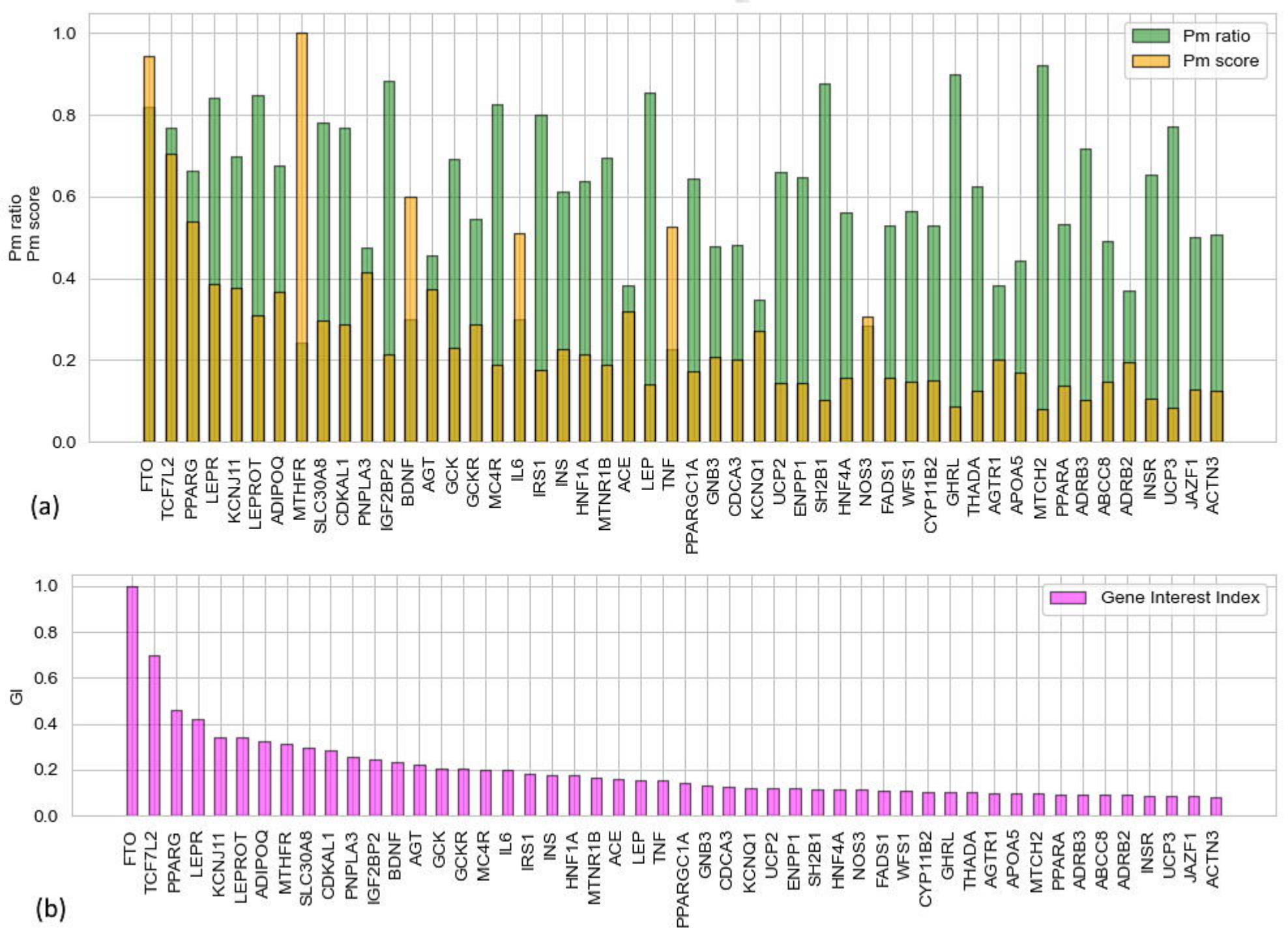
[33] Grimaldi KA, van Ommen B, Ordovas JM, Parnell LD, Mathers JC, Bendik I, et al. Proposed guidelines to evaluate scientific validity and evidence for genotype-based dietary advice. Genes and Nutrition. 2017;12(1). Type: Review.

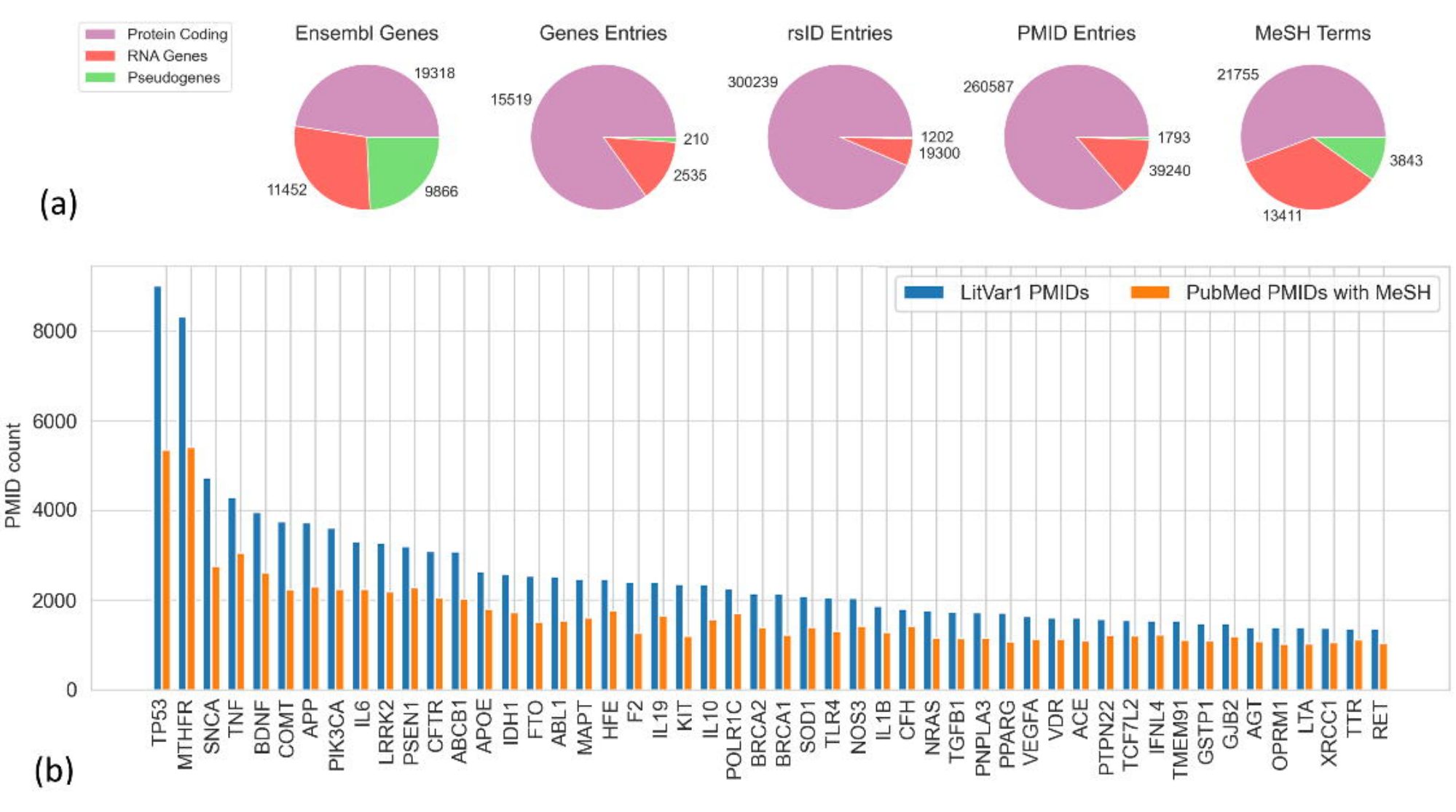
[34] De Filippis GM, Rinaldi AM, Russo C, Tommasino C. Enhanced Semantic Understanding with Graph-Based Information Retrieval. In: Boratto L, Malitesta D, Marras M, Medda G, Musto C, Purificato E, editors. Advances on Graph-Based Approaches in Information Retrieval. Cham: Springer Nature Switzerland; 2025. p. 11-24.

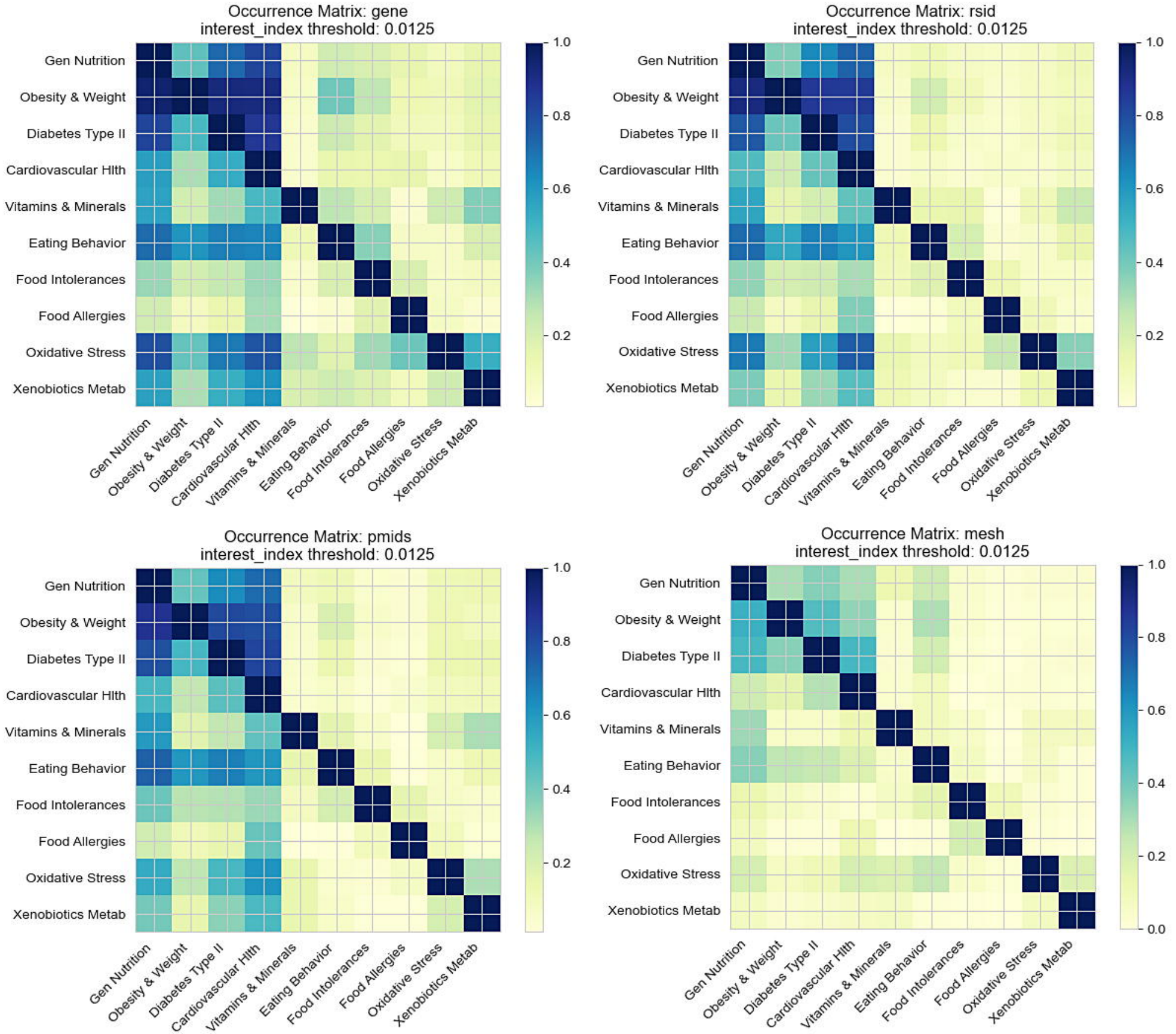
[35] Zhang Y, Zhu Q, Liu H. Next generation informatics for big data in precision medicine era. BioData Mining. 2015;8(1). Type: Review.

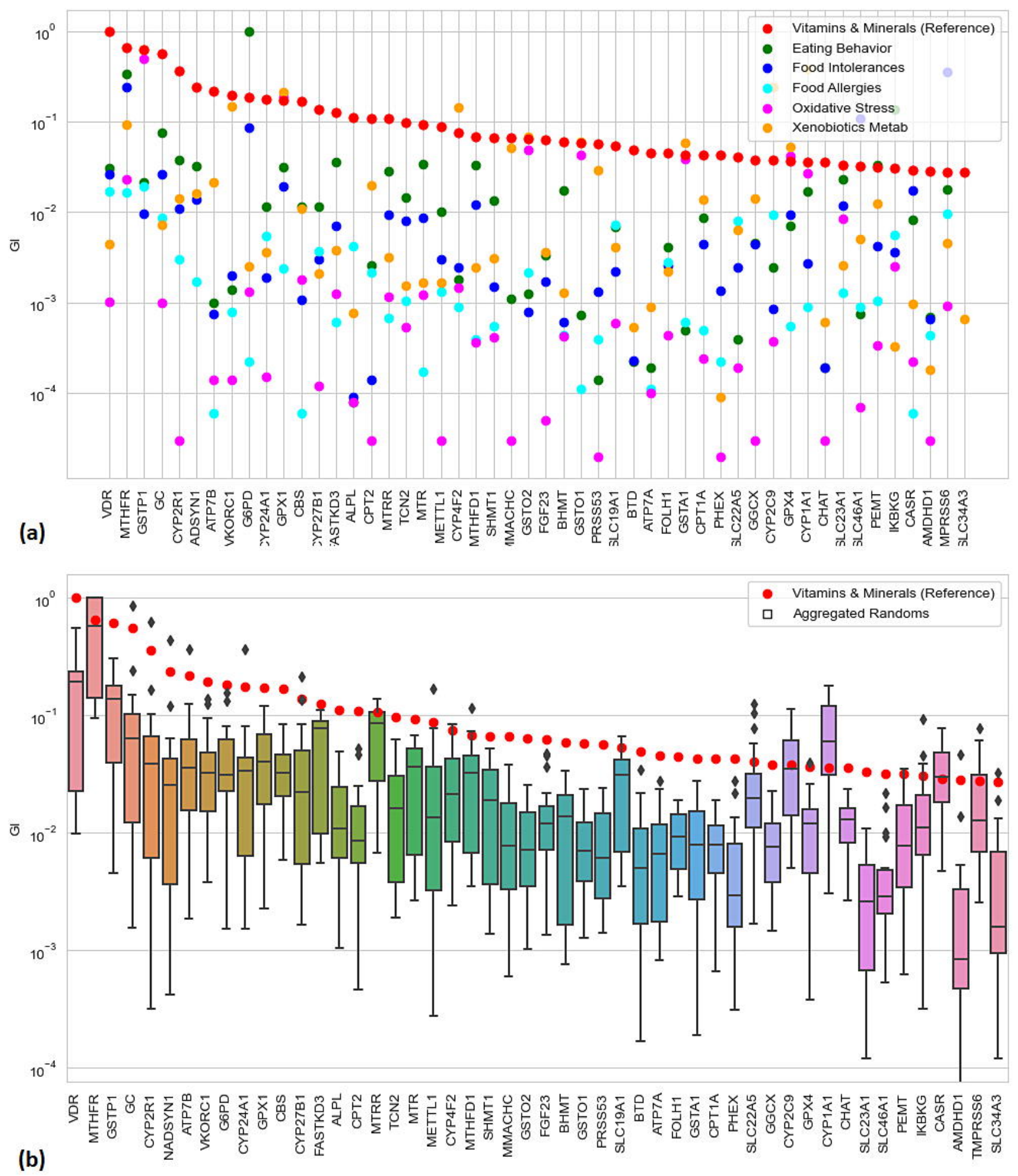
 ©2011 EuroMISE s.r.o.


Ensembl Human Genes
LitVar
PubMed
MeSH
1
Dataset Building
GRPM Dataset
Genes: 19,264
SNPs: 320,741
PMIDs: 301,620
MeSHs: 39,009
Biomedical Topics
MeSH Ontology
2
MeSH selection
MeSH List
3
Dataset Querying
Query Results
Extraction Report
4
Gene Prioritization
Filterd Dataset
BioBERT
GWAS-Catalog
5
GWAS-GRPM Integration
Integrated Dataset


Pm ratio
Pm score
1.0
0.8
0.6
0.4
0.2
0.0
FTO
TCF7L2
PPARG
LEPR
KCNJ11
LEPROT
ADIPOQ
MTHFR
SLC30A8
CDKAL1
PNPLA3
IGF2BP2
BDNF
AGT
GCK
GCKR
MC4R
IL6
IRS1
INS
HNF1A
MTNR1B
ACE
LEP
TNF
PPARGC1A
GNB3
CDCA3
KCNQ1
UCP2
ENPP1
SH2B1
HNF4A
NOS3
FADS1
WFS1
CYP11B2
GHRL
THADA
AGTR1
APOA5
MTCH2
PPARA
ADRB3
ABCC8
ADRB2
INSR
UCP3
JAZF1
ACTN3
(a)
Gene Interest Index
GI
(b)


Protein Coding
RNA Genes
Pseudogenes
Ensembl Genes
19318
11452
9866
Genes Entries
15519
210
2535
rsID Entries
300239
1202
19300
PMID Entries
260587
1793
39240
MeSH Terms
21755
3843
13411
(a)
LitVar1 PMIDs
PubMed PMIDs with MeSH
PMID count
0
2000
4000
6000
8000
TP53
MTHFR
SNCA
TNF
BDNF
COMT
APP
PIK3CA
IL6
LRRK2
PSEN1
CFTR
ABCB1
APOE
IDH1
FTO
ABL1
MAPT
HFE
F2
IL19
KIT
IL10
POLR1C
BRCA2
BRCA1
SOD1
TLR4
NOS3
IL1B
CFH
NRAS
TGFB1
PNPLA3
PPARG
VEGFA
VDR
ACE
PTPN22
TCF7L2
IFNL4
TMEM91
GSTP1
GJB2
AGT
OPRM1
LTA
XRCC1
TTR
RET
(b)


Occurrence Matrix: gene
interest_index threshold: 0.0125
Occurrence Matrix: rsid
interest_index threshold: 0.0125
Occurrence Matrix: pmids
interest_index threshold: 0.0125
Occurrence Matrix: mesh
interest_index threshold: 0.0125
Gen Nutrition
Obesity & Weight
Diabetes Type II
Cardiovascular Hlth
Vitamins & Minerals
Eating Behavior
Food Intolerances
Food Allergies
Oxidative Stress
Xenobiotics Metab
1.0
0.8
0.6
0.4
0.2
0.0


Vitamins & Minerals (Reference)
Eating Behavior
Food Intolerances
Food Allergies
Oxidative Stress
Xenobiotics Metab
GI
(a)
VDR
MTHFR
GSTP1
GC
CYP2R1
ADSYN1
ATP7B
VKORC1
G6PD
CYP24A1
GPX1
CBS
CYP27B1
FASTKD3
ALPL
CPT2
MTRR
TCN2
MTR
METTL1
CYP4F2
MTHFD1
SHMT1
MMACHC
GSTO2
FGF23
BHMT
GSTO1
PRSS53
SLC19A1
BTD
ATP7A
FOLH1
GSTA1
CPT1A
PHEX
SLC22A5
GGCX
CYP2C9
GPX4
CYP1A1
CHAT
SLC23A1
SLC46A1
PEMT
IKBKG
CASR
AMDHD1
MPRSS6
SLC34A3
Vitamins & Minerals (Reference)
Aggregated Randoms
GI
(b)
VDR
MTHFR
GSTP1
GC
CYP2R1
NADSYN1
ATP7B
VKORC1
G6PD
CYP24A1
GPX1
CBS
CYP27B1
FASTKD3
ALPL
CPT2
MTRR
TCN2
MTR
METTL1
CYP4F2
MTHFD1
SHMT1
MMACHC
GSTO2
FGF23
BHMT
GSTO1
PRSS53
SLC19A1
BTD
ATP7A
FOLH1
GSTA1
CPT1A
PHEX
SLC22A5
GGCX
CYP2C9
GPX4
CYP1A1
CHAT
SLC23A1
SLC46A1
PEMT
IKBKG
CASR
AMDHD1
TMPRSS6
SLC34A3