

Cite this: *Food Funct.*, 2025, **16**, 9408

## A computational reference for human genomics analysis in (poly)phenol research

 Mirko Treccani, \*<sup>a</sup> Lucia Ghiretti, <sup>a</sup> Nicole Tosi, <sup>a</sup> Cristiana Mignogna, <sup>a</sup> Anne-Marie Minihane, <sup>b,c</sup> Valeria Barili, <sup>d</sup> Daniele Del Rio, <sup>a</sup> Davide Martorana <sup>e,f</sup> and Pedro Mena <sup>a</sup>

(Poly)phenols are bioactive compounds found in plant-based food. They have been associated with numerous health-promoting features. To date, several research studies are investigating their contribution to human health and disease by applying different omics techniques. However, a standard reference for human genomics investigation is missing, limiting the current understanding and knowledge on the impact of (poly)phenols on humans when using omics approaches. In this study, we present a computational functional investigation of 121 candidate human genes involved in (poly)phenol absorption, distribution, metabolism, and excretion to propose a standard reference for genomics analyses in personalised nutrition and health research. Starting from genomics information, this reference framework includes gene networks, exploring their functional consequences and favouring the understanding of protein–protein interactions, thus paving the way for future multi-omics approaches.

Received 29th July 2025,  
Accepted 26th September 2025

DOI: 10.1039/d5fo03227j

rsc.li/food-function

### Introduction

(Poly)phenols are bioactive compounds found in plant-based food, such as fruits, vegetables, wine, and tea, which are characterised by the presence of one or more phenol groups.<sup>1</sup> Depending on their chemical structure, these compounds are classified into different groups, namely flavonoids, phenolic acids, stilbenes, lignans, and other (poly)phenols.<sup>2</sup> (Poly)phenols display a variety of effects, such as anti-inflammatory, indirect antioxidant, and antiproliferative actions, associated with a reduced risk of cardiometabolic diseases, some cancer types, and healthy ageing.<sup>3,4</sup>

Several human genes and their enzymatic and non-enzymatic protein products participate in the absorption, distribution, metabolism, and excretion (ADME) of (poly)phenols, influencing their bioavailability and bioactivity. After ingestion, dietary (poly)phenols reach the small intestine, where only a small fraction is absorbed. The bulk of (poly)phenols are catabolised by the gut microbiota to be widely absorbed at

the colonic level. Once absorbed, phenolic compounds mostly undergo phase II enzymatic metabolism, to increase their hydrophilicity and excretion through the urine; in these processes, they can be conjugated with one or moieties of glucuronic acid, sulphate and methyl groups.<sup>5,6</sup> Absorption comprises the action of several actors, such as the lactase phlorizin hydrolase (LPH) encoded by the lactase (LCT) gene, the cytosolic  $\beta$ -glucosidase (CBG) encoded by the glucosylceramidase beta 3 (GBA3), and sodium-dependent glucose transporter 1 (SGLT1) encoded by the solute carrier family 5 member 1 (SLC5A1) genes.<sup>7,8</sup> Conjugation with other molecules is catalysed by uridine-5'-diphosphate glucuronosyltransferases (UGTs), sulfotransferases (SULTs), and catechol-O-methyltransferases (COMT).<sup>4</sup> The action of other catalysers, such as the cytochrome P450 (CYP) gene family,<sup>8</sup> and transporters, like the adenosine triphosphate (ATP)-binding cassette (ABC)<sup>9</sup> and the solute carriers (SLC) transporter families,<sup>7,8</sup> may also determine (poly)phenol bioavailability and bioactivity.

An increasing number of studies are highlighting that ability and effect size of (poly)phenols on the risk of chronic disease depends on inter-individual variability in (poly)phenol bioavailability.<sup>10,11</sup> In recent years, various studies have employed omics approaches to elucidate the roles of genetics, epigenetics, microbiome, and environmental factors in contributing to inter-individual variability of different (poly)phenols,<sup>12–16</sup> aiming at proposing novel strategies for personalised health approaches.<sup>17–19</sup> However, despite these efforts, a reference for human genomics research in (poly)phenols is missing. To address this gap, we are proposing a compu-

<sup>a</sup>Human Nutrition Unit, Department of Food and Drug, University of Parma, 43125 Parma, Italy. E-mail: mirko.treccani@unipr.it<sup>b</sup>Nutrition and Preventive Medicine, Norwich Medical School, University of East Anglia, Norwich, UK<sup>c</sup>Norwich Institute of Healthy Ageing, University of East Anglia, Norwich, UK<sup>d</sup>Medical Genetics, Department of Medicine and Surgery, University of Parma, 43126 Parma, Italy<sup>e</sup>Medical Genetics Unit, Department of Onco-Hematology, University Hospital of Parma, 43126 Parma, Italy<sup>f</sup>CoreLab Unit, Research Center, University Hospital of Parma, 43126 Parma, Italy

tational investigation of human genes related to (poly)phenol ADME, using gene networks and computational functional analyses. In this study, we aim to provide an overview of the current resources available from public repositories, answering the question on what to expect when performing human genomics analysis in (poly)phenols research studies, hence proposing a reference for future omics investigations. These findings will help support multi-genomics and multi-omics investigations, advancing (poly)phenol-related personalised health approaches and potentially dietary guidance.

## Materials and methods

A set of candidate genes reported to be associated with (poly)phenols in terms of ADME, interaction with the intestinal microbial community, and modulation of gut physiology were considered.<sup>20</sup> These genes summarised the main processes involved in (poly)phenol ADME, such as glucuronidation (UGTs), sulfation (SULFs and SULTs), deglycosylation (LPHs), phase I and II metabolism (MUCs and CYPs), and transport (ABCs and SLCs).

Network analysis was performed using the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database version 12.0<sup>21</sup> (accessed on 15th January 2025), considering the following parameters: network type set to “full STRING network”; all the interaction sources were enabled (“Textmining”, “Experiments”, “Databases”, “Co-expression”, “Neighborhood”, “Gene Fusion”, and “Co-occurrence”); minimum interaction score set to “medium confidence (0.400)”. Clustering was performed using the STRING built-in *k*-means clustering algorithm. Functional enrichment analysis was performed using the following references: Gene Ontology (GO),<sup>22,23</sup> Kyoto Encyclopedia of Genes and Genomes (KEGG),<sup>24–26</sup> REACTOME,<sup>27</sup> WikiPathways,<sup>28,29</sup> MONARCH,<sup>30</sup> DISEASES,<sup>31</sup> TISSUES,<sup>32</sup> and COMPARTMENTS<sup>33</sup> databases. Significant enrichment was assessed by the false discovery rate (FDR) method.

Computational functional analysis was performed using the Functional Mapping and Annotation of Genome-Wide Association Studies (FUMAGWAS)<sup>34</sup> platform version 1.5.2 and its module GENE2FUNC, considering the following sources: Ensembl<sup>35</sup> version 110, the Molecular Signature Database (MSigDB)<sup>36</sup> version 7.0, WikiPathways<sup>28,29</sup> version 20191010, the GWAS Catalog<sup>37</sup> version e0\_r2022-11-29, and the Genotype-Tissue Expression (GTEx)<sup>38</sup> version 8. Differentially expressed gene sets and gene set enrichment analyses were performed using a hypergeometrical test adjusted by Bonferroni correction.

## Results

### Genes targeting (poly)phenolic metabolism

Among the 19 868 human protein-coding and the 42 160 human non-coding human genes, 121 genes were

selected since their reported involvement in (poly)phenol metabolism. These genes were located along 21 chromosomes, comprising all the autosomes but chromosomes 18 and 21, and the sex chromosome X. The most enriched chromosomes were three: 4 (18 genes), 2 (15 genes), and 11 (14 genes). Genes were almost equally distributed between the plus (71 genes) and minus (50 genes) strands of the human genome. All the genes were reported in Ensembl, except for mucin 3B (MUC3B). Finally, genes were categorised depending on their role in (poly)phenol ADME, with “Phase 2 metabolism” (31 genes), “Colonic metabolism and absorption” (25 genes), “Absorption, Distribution, Metabolism, and Elimination” (17 genes), and “Phase 1 metabolism” (16 genes) being the more represented. Of note, 16 genes fell into the “Additional” category. An overview is provided in Fig. 1.

### Gene network analysis

The 121 human target genes were used to build an interaction network. Of these, 5 genes (GBA3, MUC3B, MUC8, SLC22A20P and LDHB) were unavailable, due to the lack of their coded proteins, and thus were excluded. The network presented 116 nodes (*i.e.*, the surviving genes of interest) connected by 1098 edges, with an average node degree of 18.9 edges, and an enrichment *p*-value of  $p < 1 \times 10^{-16}$ , showing significant biological connection among genes/proteins. The full network is reported in Fig. 2.

Across all the tested databases, functional enrichment was observed at several levels, the complete results being reported in Table S1. Gene Ontology terms were enriched in biological processes related to different types of metabolisms, such as xenobiotics (GO:0006805, FDR =  $2.12 \times 10^{-48}$ ) and organic acids (GO:0006802, FDR =  $3.62 \times 10^{-31}$ ), and metabolic reactions, such as cellular glucuronidation (GO:0052695, FDR =  $8.29 \times 10^{-28}$ ); in molecular functions such as glucuronosyltransferase (GO:0015020, FDR =  $2.93 \times 10^{-24}$ ), UDP-glycosyltransferase (GO:0008194, FDR =  $1.08 \times 10^{-17}$ ), and monooxygenase (GO:0004497, FDR =  $6.27 \times 10^{-16}$ ) activities; in cellular components such as the endomembrane system (GO:0012505, FDR =  $8.37 \times 10^{-14}$ ) and the Golgi lumen (GO:0005796, FDR =  $1.21 \times 10^{-15}$ ) and different membrane compartments, such as the membrane (GO:0016020, FDR =  $8.37 \times 10^{-14}$ ), and intrinsic (GO:0031224, FDR =  $8.37 \times 10^{-14}$ ) and integral (GO:0016021, FDR =  $1.43 \times 10^{-13}$ ) components. In summary, significant terms comprised the metabolism of different phenolic compounds, such as xenobiotics, and associated molecular functions, mostly related but not limited to glucuronidation. An overview of the most significant results from GO is provided in Table 1.

Looking at pathways, enrichment was observed at several levels. The metabolism of different compounds was significant, such as for xenobiotics (hsa00980, FDR =  $6.63 \times 10^{-37}$ ; HSA-211981, FDR =  $2.54 \times 10^{-18}$ ), retinol (hsa00830, FDR =  $7.54 \times 10^{-36}$ ), and others. Moreover, several metabolic processes were enriched, such as pentose and glucuronate interconversion (hsa00040, FDR =  $2.54 \times 10^{-28}$ ), phase I and phase II metabolism (HSA-211945, FDR =  $6.85 \times 10^{-17}$ ; HSA-156580,



## ABSORPTION

LCT	MUC3B	MUC19	SLC22A20P
SLC5A1	MUC4	MUC20	SLC22A9
SLC2A2	MUC5AC	MUC21	SLC22A13
GBA3	MUC5B	TLR2	SLC22A12
ABCB1	MUC6	TLR4	SLC16A1
ABCC1	MUC7	TLR5	SLC16A2
ABCC2	MUC8	MYD88	SLC16A3
ABCC3	OVGP1	NOD2	SLC16A4
ABCC4	MUC12	SLC22A6	SLC16A5
ABCC5	MUC13	SLC22A7	SLC16A6
ABCG2	EMCN	SLC22A8	SLC16A7
MUC1	MUC15	SLC22A11	SLC16A8
MUC2	MUC16	SLC22A10	NFE2L2
MUC3A	MUC17		

## ADDITIONAL

ABCD3	LDHB	PPIA	ARNT
CYP2E1	CCL2	SLC22A5	CDKN1B
ITGA6	PDHA1	AKT1	HIF1A
LDHA	PDK3	AHR	KCNJ11

## EXCRETION

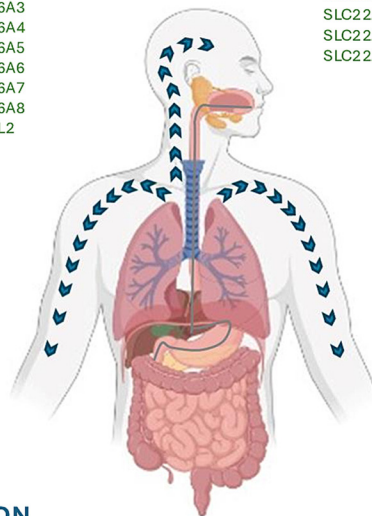
ABCB1	ABCG2	SLC22A20P	SLC16A3
ABCC1	SLC22A6	SLC22A9	SLC16A4
ABCC2	SLC22A7	SLC22A13	SLC16A5
ABCC3	SLC22A8	SLC22A12	SLC16A6
ABCC4	SLC22A11	SLC16A1	SLC16A7
ABCC5	SLC22A10	SLC16A2	SLC16A8

## DISTRIBUTION

SLC22A6	SLC22A20P	SLC16A1	SLC16A5
SLC22A7	SLC22A9	SLC16A2	SLC16A6
SLC22A8	SLC22A13	SLC16A3	SLC16A7
SLC22A11	SLC22A12	SLC16A4	SLC16A8
SLC22A10			

## METABOLISM

MUC1	MYD88	UGT2A2	SLC22A6
MUC2	NOD2	UGT2A3	SLC22A7
MUC3A	CYP1A1	UGT2B4	SLC22A8
MUC3B	CYP1A2	UGT2B7	SLC22A11
MUC4	CYP1B1	UGT2B10	SLC22A10
MUC5AC	CYP2C8	UGT2B11	SLC22A20P
MUC5B	CYP2C9	UGT2B15	SLC22A9
MUC6	CYP2C18	UGT2B17	SLC22A13
MUC7	CYP2C19	UGT2B28	SLC22A12
MUC8	CYP3A4	UGT8	SLC16A1
OVGP1	CYP3A5	GUSB	SLC16A2
MUC12	CYP3A7	SULT1A1	SLC16A3
MUC13	CYP3A43	SULT1A2	SLC16A4
EMCN	UGT1A1	SULT1A3	SLC16A5
MUC15	UGT1A3	SULT1B1	SLC16A6
MUC16	UGT1A4	SULT1C2	SLC16A7
MUC17	UGT1A5	SULT1C4	SLC16A8
MUC19	UGT1A6	SULT1E1	NFE2L2
MUC20	UGT1A7	SULT2A1	AKR7A2
MUC21	UGT1A8	SULT2B1	FMO1
TLR2	UGT1A9	SULT4A1	FMO2
TLR4	UGT1A10	SULF1	FMO3
TLR5	UGT2A1	SULF2	FMO4
		COMT	FMO5



**Fig. 1** Investigated genes involved in (poly)phenol ADME. The figure shows the 121 genes involved in absorption (A), distribution (D), metabolism (M), and excretion (E) of (poly)phenols that were investigated in this study. Additional genes result from previous enrichment analysis. Black genes: unique to the category; light blue: shared between A and M; green: shared among ADME.

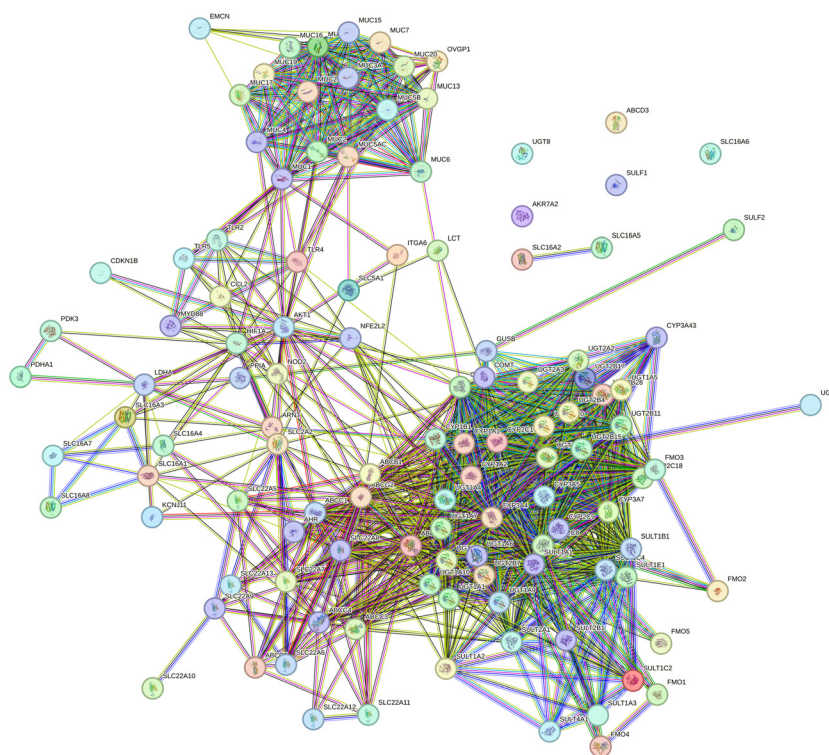
FDR =  $1.17 \times 10^{-33}$ ; WP702, FDR =  $2.14 \times 10^{-53}$ ), glucuronidation (HSA-156588, FDR =  $9.90 \times 10^{-26}$ ; WP698, FDR =  $2.30 \times 10^{-27}$ ), and sulfation (WP692, FDR =  $7.90 \times 10^{-13}$ ). Enrichment was also observed in diseases-related pathways, spanning from general “chemical carcinogenesis” (hsa05204, FDR =  $4.13 \times 10^{-48}$ ) to familial hyperphosphatemic tumoral calcinosis (HFTC; HSA-5083625, FDR =  $4.27 \times 10^{-25}$ ) and colorectal cancer (FDR =  $4.27 \times 10^{-25}$ ). Moreover, drug-related pathways showed significance at several levels, such as in drug metabolism (hsa00982, FDR =  $7.30 \times 10^{-43}$ ; hsa00983, FDR =  $2.81 \times 10^{-26}$ ) and ADME (HSA-98748784, FDR =  $3.54 \times 10^{-47}$ ), and for specific drugs like aspirin (HSA-9749641, FDR =  $1.01 \times 10^{-36}$ ), paracetamol (HSA-9753281, FDR =  $3.31 \times 10^{-23}$ ), and codeine and morphine (WP1604, FDR =  $9.08 \times 10^{-19}$ ). Overall, enriched terms were observed in different metabolic processes, such as phase I and phase II metabolism, for various phenolic metabolites, such as xenobiotics and retinol, and drugs, mostly of phenolic origin like aspirin and paracetamol; significant terms were also observed in cancer-related diseases. An overview of the most enriched terms is reported in Table 2.

Enrichment in genes associated with diseases and human phenotypes was also investigated. Significant signals were observed in metabolic disorders (DOID:0014667, FDR =  $4.47 \times 10^{-8}$ ; DOID:655, FDR =  $1.64 \times 10^{-7}$ ), mostly related to bilirubin,

bile and gallbladder (DOID:2741, FDR =  $1.57 \times 10^{-13}$ ; DOID:2739, FDR =  $2.79 \times 10^{-9}$ ; DOID:10211, FDR =  $6.70 \times 10^{-8}$ ). Moreover, metabolite (EFO:0005664, FDR =  $7.53 \times 10^{-14}$ ; EFO:0005653, FDR =  $7.53 \times 10^{-14}$ ) and vitamin measurements (EFO:0004729, FDR =  $4.33 \times 10^{-7}$ ; EFO:0004631, FDR =  $4.21 \times 10^{-6}$ ) and related abnormalities (HP:0001939, FDR =  $1.64 \times 10^{-5}$ ; HP:0100508, FDR =  $6.56 \times 10^{-5}$ ) turned out to be enriched. In summary, enriched terms for diseases pointed at metabolic condition and abnormalities, comprising bilirubin, metabolite and vitamin levels. An overview of the most significant findings for diseases and human phenotypes is provided in Tables 3 and 4, respectively.

Finally, enrichment in human tissues was also examined. Significant signals were observed in the visceral body cavity (BTO:0001491, FDR =  $2.52 \times 10^{-20}$ ), in organs related to ADME as the liver (BTO:0000759, FDR =  $2.47 \times 10^{-17}$ ), the gastrointestinal tract (BTO:0000511, FDR =  $4.25 \times 10^{-5}$ ), and the kidneys (BTO:0000671, FDR =  $1.35 \times 10^{-5}$ ), and in several glands, such as digestive, endocrine, and excretory (BTO:0000522, FDR =  $2.22 \times 10^{-14}$ ; BTO:0000345, FDR =  $1.88 \times 10^{-15}$ ; BTO:0001488, FDR =  $3.92 \times 10^{-7}$ ; BTO:0000431, FDR =  $1.12 \times 10^{-5}$ ). Globally, organs and tissues mostly involved in (poly)phenol ADME were enriched. The most significant results are reported in Table 5.





**Fig. 2** STRING network of genes related to (poly)phenols bioavailability. The figure shows the interaction network of 116 genes related to (poly)phenol bioavailability. Each node represents a gene/protein. Each edge represents an interaction between two nodes (i.e., genes/proteins); the increasing number of coloured bars on the same edge represents an increasing number of sources supporting the connection.

**Table 1** Most enriched terms from Gene Ontology. The table shows the top 5 most significant terms related to Gene Ontology Biological Process (lines 1–5), molecular function (lines 6–10), and cellular component (lines 11–15). For each term, the identifier, a description, the number of network genes over the total reported genes, the strength of the enrichment effect, the signal strength, and the Benjamini–Hochberg adjusted false discovery rate are reported. Strength is defined as  $\log_{10}(\text{observed genes in the network}/\text{expected genes in a random network})$ . Signal is defined as the weighted harmonic mean between observed over expected ratio and the  $-\log(\text{FDR})$ . FDR: false discovery rate; GO: Gene Ontology

GO-term	Description	Genes in network	Strength	Signal	FDR
GO:0006805	Xenobiotic metabolic process	38/115	1.75	9.17	$2.12 \times 10^{-48}$
GO:0071466	Cellular response to xenobiotic stimulus	39/181	1.56	7.33	$6.74 \times 10^{-44}$
GO:0009410	Response to xenobiotic stimulus	44/422	1.25	4.77	$6.74 \times 10^{-38}$
GO:0006082	Organic acid metabolic process	49/868	0.98	3.04	$3.62 \times 10^{-31}$
GO:0052695	Cellular glucuronidation	18/19	2.21	7.14	$8.29 \times 10^{-28}$
GO:0015020	Glucuronosyltransferase activity	18/34	1.95	5.9	$2.93 \times 10^{-24}$
GO:0008194	UDP-glycosyltransferase activity	20/143	1.38	3.45	$1.08 \times 10^{-17}$
GO:0004497	Monooxygenase activity	17/103	1.45	3.33	$6.27 \times 10^{-16}$
GO:0070330	Aromatase activity	12/24	1.93	3.91	$1.24 \times 10^{-15}$
GO:0016758	Hexosyltransferase activity	19/194	1.22	2.66	$2.01 \times 10^{-14}$
GO:0005796	Golgi lumen	17/106	1.44	3.26	$1.21 \times 10^{-15}$
GO:0012505	Endomembrane system	70/4721	0.4	0.92	$8.37 \times 10^{-14}$
GO:0031224	Intrinsic component of membrane	78/5828	0.36	0.85	$8.37 \times 10^{-14}$
GO:0016020	Membrane	99/9523	0.25	0.70	$8.37 \times 10^{-14}$
GO:0016021	Integral component of membrane	76/5670	0.36	0.85	$1.43 \times 10^{-13}$

Network nodes were clustered using the *k*-means clustering algorithm. To define the most appropriate number of clusters, an iterative approach was applied, starting from a minimum of 2 clusters and avoiding clusters comprising only one gene. The 2-means clustering identified two clusters, one of 109 genes and one of 2 genes, with 5 genes not belonging to any cluster.

Increasing the number of clusters, the 9-means clustering reported for the first time a cluster comprising only one gene. Therefore, the 8-means clustering was considered the most informative. The 8-cluster model comprised 57, 19, 14, 7, 5, 4, 3, and 2 genes. The smallest cluster was consistent across all the iterations. Moreover, it was possible to summarise each



**Table 2** Most enriched terms from KEGG, REACTOME and WikiPathways. The table shows the top 5 most significant terms related to the KEGG (lines 1–5), REACTOME (lines 6–10), and Wikipathways (lines 11–15) databases. For each term, the pathway identifier (hsa code for KEGG, HSA code for REACTOME, WP for WikiPathways), a description, the number of network genes over the total reported genes, the strength of the enrichment effect, the signal strength, and the Benjamini–Hochberg adjusted false discovery rate are reported. Strength is defined as  $\log_{10}(\text{observed genes in the network/expected genes in a random network})$ . Signal is defined as the weighted harmonic mean between observed over expected ratio and the  $-\log(\text{FDR})$ . ADME: absorption, distribution, metabolism, and excretion; FDR: false discovery rate; KEGG: Kyoto Encyclopedia of Genes and Genomes

Pathway	Description	Genes in network	Strength	Signal	FDR
hsa05204	Chemical carcinogenesis	34/76	1.88	10.01	$4.13 \times 10^{-48}$
hsa00982	Drug metabolism – cytochrome P450	30/64	1.90	9.28	$7.30 \times 10^{-43}$
hsa00140	Steroid hormone biosynthesis	28/60	1.90	8.78	$5.70 \times 10^{-40}$
hsa00980	Metabolism of xenobiotics by cytochrome P450	27/69	1.82	7.92	$6.63 \times 10^{-37}$
hsa04976	Bile secretion	28/88	1.73	7.41	$2.49 \times 10^{-36}$
HSA-211859	Biological oxidations	45/215	1.55	7.94	$1.33 \times 10^{-51}$
HSA-9748784	Drug ADME	36/105	1.77	9.12	$3.54 \times 10^{-47}$
HSA-9749641	Aspirin ADME	25/42	2.00	8.61	$1.01 \times 10^{-36}$
HSA-156580	Phase II – conjugation of compounds	28/104	1.66	6.67	$1.17 \times 10^{-33}$
HSA-1430728	Metabolism	63/2092	0.71	1.89	$4.66 \times 10^{-28}$
WP702	Metapathway biotransformation Phase I and II	44/180	1.62	8.68	$2.14 \times 10^{-53}$
WP691	Tamoxifen metabolism	19/21	2.19	7.66	$3.76 \times 10^{-30}$
WP698	Glucuronidation	18/25	2.09	6.84	$2.30 \times 10^{-27}$
WP2882	Nuclear receptors <i>meta</i> -pathway	29/312	1.2	3.54	$1.92 \times 10^{-23}$
WP5276	Estrogen metabolism	14/24	2.0	5.06	$3.32 \times 10^{-20}$

**Table 3** Most enriched terms from DISEASES. The table shows the top 5 most significant terms related to the disease-gene associations from the DISEASES database. For each term, the identifier, a description, the number of network genes over the total reported genes, the strength of the enrichment effect, the signal strength, and the Benjamini–Hochberg adjusted false discovery rate are reported. Strength is defined as  $\log_{10}(\text{observed genes in the network/expected genes in a random network})$ . Signal is defined as the weighted harmonic mean between observed over expected ratio and the  $-\log(\text{FDR})$ . FDR: false discovery rate

Disease	Description	Genes in network	Strength	Signal	FDR
DOID:2741	Bilirubin metabolic disorder	10/13	2.12	3.51	$1.57 \times 10^{-13}$
DOID:2739	Gilbert syndrome	7/8	2.17	2.4	$2.79 \times 10^{-9}$
DOID:2382	Kernicterus	6/6	2.23	2.08	$4.47 \times 10^{-8}$
DOID:3803	Crigler-Najjar syndrome	6/6	2.23	2.08	$4.47 \times 10^{-8}$
DOID:0014667	Disease of metabolism	28/1076	0.65	1.07	$4.47 \times 10^{-8}$

**Table 4** Most enriched terms from MONARCH. The table shows the top 5 most significant terms related to human phenotypes from the MONARCH database. For each term, the identifier, a description, the number of network genes over the total reported genes, the strength of the enrichment effect, the signal strength, and the Benjamini–Hochberg adjusted false discovery rate are reported. Strength is defined as  $\log_{10}(\text{observed genes in the network/expected genes in a random network})$ . Signal is defined as the weighted harmonic mean between observed over expected ratio and the  $-\log(\text{FDR})$ . FDR: false discovery rate

Phenotype	Description	Genes in network	Strength	Signal	FDR
EFO:0005664	Blood metabolite measurement	19/181	1.12	2.64	$7.53 \times 10^{-14}$
EFO:0005653	Serum metabolite measurement	22/287	1.11	2.37	$7.53 \times 10^{-14}$
EFO:0004739	Circulating cell free DNA measurement	9/9	2.23	3.33	$1.11 \times 10^{-12}$
EFO:0010551	Xanthurenate measurement	9/11	2.14	3.19	$2.86 \times 10^{-12}$
EFO:0004725	Metabolite measurement	33/1033	0.73	1.47	$2.86 \times 10^{-12}$

cluster but cluster 8 using a representative keywords: Cluster 1 was related to “Cellular glucuronidation”; Cluster 2 to “Defective GALNT3 causes HFTC”; Cluster 3 to “Xenobiotic transport” and “Recycling of bile acids and salt”; Cluster 4 to “Positive regulation of nitric-oxide synthase biosynthesis”; Cluster 5 to “Sulfation” and “Aryl sulfotransferase activity”; Cluster 6 to “Carbohydrate digestion and absorption”; and Cluster 7 “Proton-coupled monocarboxylate transport” and “Lactate transmembrane transporter activity” (Fig. 3). The clustered interaction network is reported in Fig. 4 and Table S2.

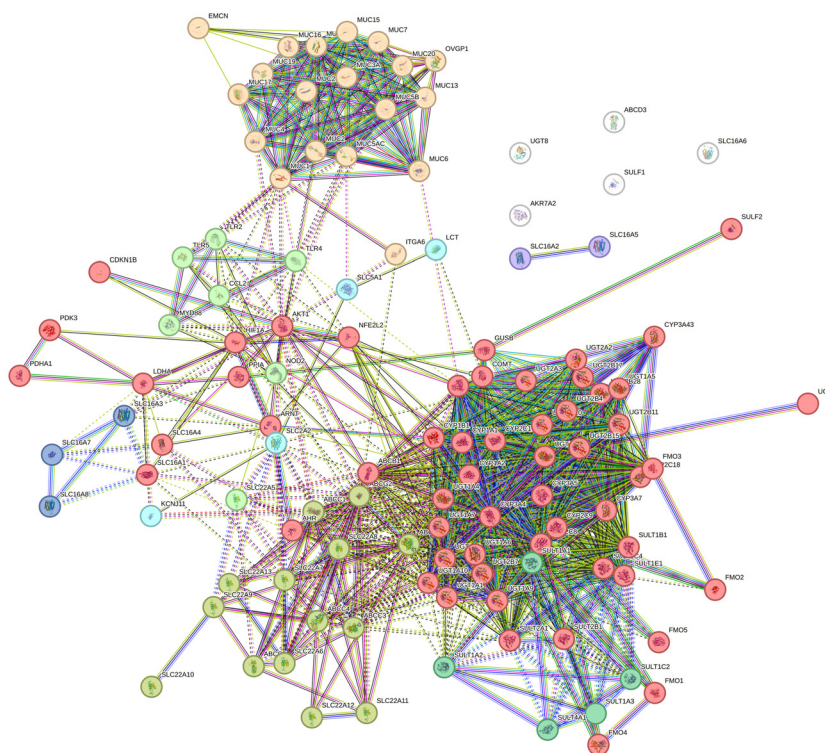
### Computational functional analysis

The selected genes were further investigated to gain additional knowledge regarding expression patterns across different tissue types, shared molecular functions among genes, and testing for enrichment analysis using pre-defined gene sets. Of the 121 input genes, 2 genes (MUC3B and SLC22A20P) were excluded since they did not present a recognised Ensembl ID, so 119 genes were investigated against 57 241 human coding and non-coding background genes.



**Table 5** Most enriched terms from TISSUES. The table shows the top 5 most significant terms related to human tissue expression from the TISSUES database. For each term, the identifier, a description, the number of network genes over the total reported genes, the strength of the enrichment effect, the signal strength, and the Benjamini–Hochberg adjusted false discovery rate are reported. Strength is defined as  $\log_{10}(\text{observed genes in the network}/\text{expected genes in a random network})$ . Signal is defined as the weighted harmonic mean between observed over expected ratio and the  $-\log(\text{FDR})$ . FDR: false discovery rate

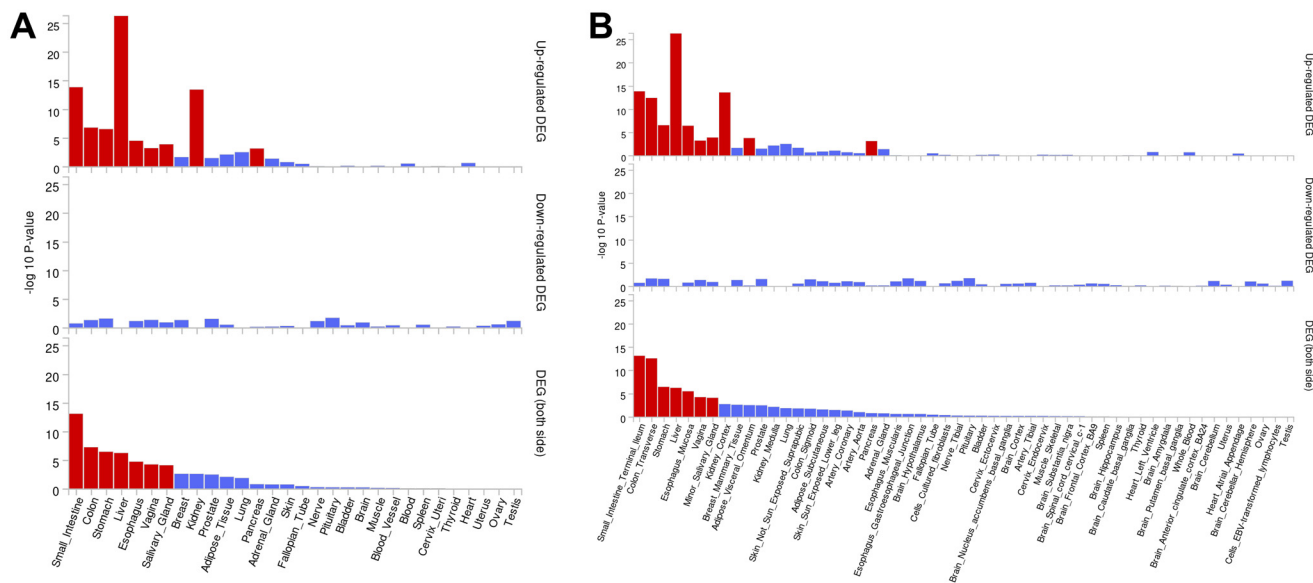
Tissue	Description	Genes in network	Strength	Signal	FDR
BTO:0001491	Viscus	84/5378	0.42	1.05	$2.52 \times 10^{-20}$
BTO:0000759	Liver	52/2125	0.62	1.42	$2.47 \times 10^{-17}$
BTO:0000345	Digestive gland	57/2881	0.53	1.18	$1.88 \times 10^{-15}$
BTO:0000522	Gland	86/7004	0.32	0.8	$2.22 \times 10^{-14}$
BTO:0001488	Endocrine gland	70/6403	0.27	0.61	$3.92 \times 10^{-7}$



color	cluster Id	gene count	description
●	Cluster 1	57	– 1. Chemical carcinogenesis 2. Cellular glucuronidation 3. Aspirin ADME
●	Cluster 2	19	Defective GALNT3 causes HFTC
●	Cluster 3	14	Mixed, incl. Xenobiotic transport, and Recycling of bile acids and salts
●	Cluster 4	7	– 1. Positive regulation of nitric-oxide synthase biosynthesis 2. Malaria 3. Diseases associated with the TLR signaling cascade
●	Cluster 5	5	– 1. Sulfation 2. Aryl sulfotransferase activity
●	Cluster 6	4	Carbohydrate digestion and absorption
●	Cluster 7	3	– 1. Proton-coupled monocarboxylate transport 2. Lactate transmembrane transporter activity
●	Cluster 8	2	SLC16A2, SLC16A5

**Fig. 3** *k*-Means clustering on the STRING network. The figure shows the STRING network divided into the eight clusters identified by *k*-means clustering. Each cluster is identified by a different colour. The legend reports the colour and the cluster identifier (Cluster Id), the number of genes (Gene count), and the description of the summarized pathway (description). A schematic version of this network is reported in Fig. S1.





**Fig. 4** General and specific tissue types enriched for differentially expressed genes from the GTEx catalog. The bar plots show the list of general (panel A, 30 tissues) and specific (panel B, 54 tissues) tissue types from the GTEx catalog version 8 and their enrichment for differentially expressed genes (DEG). Each panel is divided in three layers, showing up-regulated (top), down-regulated (middle) and both-sided regulated (bottom) DEG, as reported by the labels on the right side. Tissues are sorted by Bonferroni-corrected  $p$ -values (reported on the  $y$  axis by  $-\log_{10}(P\text{-value})$ ) on both-sided differential expression (bottom panel). Tissues presenting significant DEG sets (Bonferroni-corrected  $p$ -value  $< 0.05$ ) are highlighted in red.

Gene expression analysis was performed on both general and specific tissue types. Significant results (Bonferroni corrected  $p$ -value  $< 0.05$ ) were observed in up-regulated differentially expressed gene sets and no evidence of association resulted in down-regulated differentially expressed gene sets. As a consequence, both-sided differentially expressed gene sets turned out to be significant. Among the general tissue types ( $n = 30$ ), liver ( $p = 1.34 \times 10^{-25}$ ), small intestine ( $p = 3.46 \times 10^{-13}$ ), kidneys ( $p = 8.97 \times 10^{-13}$ ), and colon ( $p = 4.06 \times 10^{-6}$ ) showed the most significant results. This was reflected in the specific tissue types ( $n = 54$ ), where the liver ( $p = 2.41 \times 10^{-25}$ ), the small intestine terminal ileum ( $p = 6.23 \times 10^{-13}$ ), the kidney cortex ( $p = 1.11 \times 10^{-12}$ ), and the colon transverse ( $p = 1.70 \times 10^{-11}$ ) were significant. In summary, differential expression of these genes in tissues strictly related to (poly)phenol metabolism was significant. A complete overview is reported in Fig. 4.

An additional round of gene set enrichment analysis was performed. Significantly enriched terms from the Gene Ontology, KEGG, Reactome, and Wikipathways were concordant with previous findings from STRING. Further, functional analyses were performed on the gene sets from the Molecular Signature Database. Among the hallmark gene sets, significant results were observed in 3 gene sets related to metabolism, being xenobiotics ( $p = 2.70 \times 10^{-5}$ ), fatty acids ( $p = 3.20 \times 10^{-3}$ ), and peroxisome ( $p = 1.68 \times 10^{-2}$ ). Additionally, the inflammatory response was found to be significant ( $p = 3.84 \times 10^{-2}$ ). Enrichment was also observed in 4 positional gene sets on cytogenetic bands located on chromosomes 4 (chr4q13,  $p = 7.82 \times 10^{-14}$ ), 2 (chr2q37,  $p = 1.04 \times 10^{-4}$ ), 7 (chr7q22,  $p = 8.11$

$\times 10^{-3}$ ), and 1 (chr1q24,  $p = 4.21 \times 10^{-2}$ ). Overall, gene sets already computed for different dietary components were found to be significant together with specific genomic locations in human chromosomes. A complete overview is reported in Table 6.

Pathways from previously unexplored sources such as Biocarta were investigated. A total of 5 significantly enriched terms were observed in lipid metabolism (M16393,  $p = 9.53 \times 10^{-10}$ ), in hypoxia in the cardiovascular system (M13324,  $p = 4.86 \times 10^{-3}$ ; M5202,  $p = 8.96 \times 10^{-3}$ ), and in the activity and resistance to drugs (M22078,  $p = 3.31 \times 10^{-4}$ ; M22024,  $p = 1.38 \times 10^{-2}$ ).

Cancer-related gene sets were also explored. Enrichment was observed for 17 gene sets. Among the tested sets, cytochrome P450 (M5645,  $p = 5.46 \times 10^{-14}$ ) and heme biosynthesis (M2738,  $p = 2.11 \times 10^{-5}$ ), oxidoreductase activity (M10134,  $p = 9.29 \times 10^{-9}$ ), and transport of metabolites, such as carboxylic and organic acids (M13595,  $p = 6.68 \times 10^{-8}$ ), and of peptide and amino acids (M12507,  $p = 6.21 \times 10^{-5}$ ), were enriched. Moreover, xenobiotic metabolism, either general or in specific organs, such as the liver, resulted to be significant (M57,  $p = 1.69 \times 10^{-2}$ ; M1872,  $p = 2.11 \times 10^{-5}$ ; M14358,  $p = 2.76 \times 10^{-4}$ ). It can be highlighted that several biological functions and enzymatic activities associated with phenolic metabolites and the tissues where are happening were significant. A complete overview is provided in Table 7.

Different biological signatures were also explored. Oncogenic signatures showed enrichment in cellular pathways dysregulated in cancer and associated with NFE2L2 (M2870,  $p = 3.20 \times 10^{-5}$ ) and KRAS genes (M2880,  $p = 3.32 \times 10^{-2}$ ).



**Table 6** Enrichment analysis from the hallmark and positional gene sets in the Molecular Signature Database. The table shows the most enriched results from the hallmarks (lines 1–4) and positional (lines 5–8) gene sets. For each gene set, the name, the number of genes over the total number of genes in the set, the nominal and Bonferroni-adjusted  $p$ -values and the gene symbols are reported

Gene set	Number of genes	$P$ -Value	Adjusted $P$ -value	Genes
Xenobiotic metabolism	9/196	$5.41 \times 10^{-7}$	$2.70 \times 10^{-5}$	FMO3, FMO1, CYP2C18, ABCC2, CYP2E1, CYP1A1, CYP1A2, ABCC3, COMT
Peroxisome	5/99	$1.28 \times 10^{-4}$	$3.20 \times 10^{-3}$	ABCD3, ABCC5, UGT2B17, ABCB1, SULT2B1
Fatty acid metabolism	5/155	$1.01 \times 10^{-3}$	$1.68 \times 10^{-2}$	FMO1, SLC22A5, LDHA, CYP1A1, PDHA1
Inflammatory response	5/200	$3.07 \times 10^{-3}$	$3.84 \times 10^{-2}$	TLR2, AHR, HIF1A, NOD2, CCL2
chr4q13	13/85	$2.61 \times 10^{-16}$	$7.82 \times 10^{-14}$	UGT2B17, UGT2B15, UGT2B10, UGT2A3, UGT2B7, UGT2B11, UGT2B28, UGT2B4, UGT2A1, UGT2A2, UGT2A1, SULT1B1, SULT1E1, MUC7
chr2q37	8/148	$6.93 \times 10^{-7}$	$1.04 \times 10^{-4}$	UGT1A8, UGT1A10, UGT1A9, UGT1A7, UGT1A6, UGT1A5, UGT1A4, UGT1A3, UGT1A1
chr7q22	6/145	$8.11 \times 10^{-5}$	$8.11 \times 10^{-3}$	CYP3A5, CYP3A7, CYP3A4, CYP3A43, MUC12, MUC17, RN7SKP54
chr1q24	4/77	$5.61 \times 10^{-4}$	$4.21 \times 10^{-2}$	FMO3, FMO2, FMO1, FMO4

**Table 7** Enrichment analysis from the cancer gene modules from the Molecular Signature Database. The table shows the enriched terms from the cancer gene modules. For each gene set, the systematic name, the module identifier, the number of genes over the total number of genes in the set, the nominal and Bonferroni-adjusted  $p$ -values and the gene symbols are reported

Systematic name	Gene set (module)	Number of genes	$P$ -Value	Adjusted $P$ -value	Genes
M5646	M_106	8/12	$1.27 \times 10^{-16}$	$5.46 \times 10^{-14}$	CYP1B1, CYP3A4, CYP2C18, CYP2C19, CYP2C9, CYP2C8, CYP2E1, CYP1A1
M8592	M_135	9/21	$3.33 \times 10^{-16}$	$7.17 \times 10^{-14}$	FMO3, FMO4, CYP1B1, CYP3A4, CYP2C18, CYP2C19, CYP2C8, CYP2E1, CYP1A1
M10134	M_93	12/175	$6.47 \times 10^{-11}$	$9.29 \times 10^{-9}$	AKR7A2, FMO5, FMO3, FMO4, CYP1B1, CYP3A5, CYP3A4, CYP2C19, CYP2C8, LDHA, CYP1A1, COMT
M13595	M_71	6/21	$6.20 \times 10^{-10}$	$3.84 \times 10^{-7}$	SLC16A4, SLC16A1, ABCC3, SLC16A5, SLC16A8, SLC16A2
M2027	M_212	13/311	$4.46 \times 10^{-9}$	$3.84 \times 10^{-7}$	ABCD3, FMO3, FMO4, UGT2B15, UGT2B7, SULT1E1, CYP3A4, CYP2C8, LDHA, CYP1A1, ABCC1, ABCC3, COMT
M9415	M_186	5/18	$2.11 \times 10^{-8}$	$1.51 \times 10^{-6}$	SLC16A4, SLC16A1, ABCC3, SLC16A8, SLC16A2
M2738	M_505	4/13	$3.78 \times 10^{-7}$	$2.11 \times 10^{-5}$	UGT2B15, UGT2B7, UGT2B4, SULT1E1
M5974	M_43	7/94	$4.07 \times 10^{-7}$	$2.11 \times 10^{-5}$	FMO3, FMO4, CYP3A4, CYP2C19, CYP2C8, LDHA, CYP1A1
M1872	M_23	14/542	$4.42 \times 10^{-7}$	$2.11 \times 10^{-5}$	FMO5, MUC1, FMO3, FMO4, UGT2B15, UGT2B7, UGT2B4, CYP3A4, CYP2C19, CYP2C9, CYP2C8, CYP2E1, CCL2, ABCC3
M586	M_227	4/15	$7.15 \times 10^{-7}$	$3.08 \times 10^{-5}$	UGT1A6, UGT2B15, UGT2B7, UGT2B4
M12507	M_368	4/18	$1.59 \times 10^{-6}$	$6.21 \times 10^{-5}$	SLC16A4, SLC16A6, SLC16A8, SLC16A2
M17558	M_55	16/800	$1.87 \times 10^{-6}$	$6.63 \times 10^{-5}$	SLC16A4, FMO5, MUC1, FMO3, FMO4, ITGA6, UGT2B15, UGT2B7, UGT2B4, CYP2C19, CYP2C8, CYP2E1, MUC5B, SLC22A6, CCL2, ABCC3
M16754	M_218	4/19	$2.00 \times 10^{-6}$	$6.63 \times 10^{-5}$	SLC16A4, SLC16A7, SLC16A8, SLC16A2
M14358	M_88	15/802	$8.98 \times 10^{-6}$	$2.76 \times 10^{-4}$	SLC16A4, FMO5, MUC1, FMO3, FMO4, ITGA6, UGT2B15, UGT2B7, CYP2C19, CYP2C8, CYP2E1, MUC5B, SLC22A6, CCL2, ABCC3
M8950	M_117	13/697	$3.81 \times 10^{-5}$	$1.10 \times 10^{-3}$	SULT1C2, UGT1A4, SLC2A2, UGT2B17, UGT2B7, UGT2B4, UGT2A1, UGT2A1, SULT1B1, UGT8, CYP3A4, CYP2C8, CYP1A1, SLC16A2
M10538	M_162	3/19	$1.05 \times 10^{-4}$	$2.82 \times 10^{-3}$	SLC16A4, SLC16A7, SLC16A2
M57	M_247	3/35	$6.68 \times 10^{-4}$	$1.69 \times 10^{-2}$	FMO4, SULT1C2, UGT2B15

Moreover, cell type signatures resulting from single-cell studies in different human tissues proved to be significant. Among them, 34 terms were enriched, particularly in intestinal (M40028,  $p = 1.11 \times 10^{-14}$ ), hepatic (M39115,  $p = 3.37 \times 10^{-9}$ ), and pancreatic ( $6.44 \times 10^{-5}$ ) cell types.

Potential targets of regulation were also investigated, considering both microRNAs (miRNAs) and transcription factor (TF) targets. Six microRNAs (hsa-miR-6736-3p,  $p = 2.55 \times 10^{-5}$ ; hsa-miR-6507-5p,  $p = 7.93 \times 10^{-5}$ ; hsa-miR-4786-3p,  $p = 7.93 \times 10^{-5}$ ; hsa-miR-185-3p,  $p = 4.28 \times 10^{-4}$ ; hsa-miR-6839-3p,  $p = 4.41 \times 10^{-4}$ ; hsa-miR-4496,  $p = 5.90 \times 10^{-4}$ ) and one TF target

(M133 RYTAAWNNTGAY,  $p = 1.27 \times 10^{-2}$ ) were significantly enriched, mostly driven by the UGT1A complex together with additional genes, as reported in Table 8.

The GWAS Catalog was examined to retrieve associated phenotypes and diseases, resulting in a total of 42 enriched terms. Among them, different metabolites were significant, either general serum ( $p = 1.00 \times 10^{-21}$ ) and urinary ( $p = 5.67 \times 10^{-10}$ ) metabolites or specific compounds, such as bilirubin ( $p = 5.40 \times 10^{-12}$ ) and low-density lipoprotein cholesterol ( $p = 8.18 \times 10^{-5}$ ) levels. Of note, terms related to bilirubin were particularly enriched in terms of both significance and quantity.





**Table 8** Enrichment analysis from the microRNA and transcription factor targets gene sets in the Molecular Signature Database. The table shows the most enriched results from the microRNA targets (lines 1–6), and the transcription factor targets (line 7) gene sets. For each gene set, the name, the number of genes over the total number of genes in the set, the nominal and Bonferroni-adjusted *p*-values and the gene symbols are reported

Gene set	Number of genes	<i>P</i> -Value	Adjusted <i>P</i> -value	Genes
MIR6736_3P	9/123	$9.82 \times 10^{-9}$	$2.55 \times 10^{-5}$	UGT1A8, UGT1A10, UGT1A9, UGT1A7, UGT1A6, UGT1A5, UGT1A4, UGT1A3, UGT1A1, LDHA
MIR6507_3P	12/329	$8.07 \times 10^{-8}$	$7.93 \times 10^{-5}$	UGT1A8, UGT1A10, UGT1A9, UGT1A7, UGT1A6, UGT1A5, UGT1A4, UGT1A3, UGT1A1, ABCG2, AHR, HIF1A, CYP1A2
MIR4786_3P	9/159	$9.16 \times 10^{-8}$	$7.93 \times 10^{-5}$	UGT1A8, UGT1A10, UGT1A9, UGT1A7, UGT1A6, UGT1A5, UGT1A4, UGT1A3, UGT1A1, UGT2B17
MIR185_3P	8/147	$6.59 \times 10^{-7}$	$4.28 \times 10^{-4}$	UGT1A8, UGT1A10, UGT1A9, UGT1A7, UGT1A6, UGT1A5, UGT1A4, UGT1A3, UGT1A1
MIR6839_3P	8/152	$8.49 \times 10^{-7}$	$4.41 \times 10^{-4}$	UGT1A8, UGT1A10, UGT1A9, UGT1A7, UGT1A6, UGT1A5, UGT1A4, UGT1A3, UGT1A1
MIR4496	10/283	$1.36 \times 10^{-6}$	$5.90 \times 10^{-4}$	UGT1A8, UGT1A10, UGT1A9, UGT1A7, UGT1A6, UGT1A5, UGT1A4, UGT1A3, UGT1A1, AHR, CDKN1B
RYTAAWNNNTGAY_UNKOWN	5/60	$1.14 \times 10^{-5}$	$1.27 \times 10^{-2}$	UGT1A1, SLC22A6, SLC22A8, SLC22A12, SULF2

Additionally, some diseases, such as liver disease ( $p = 6.24 \times 10^{-14}$ ) and gout ( $p = 7.44 \times 10^{-4}$ ), as well as some drugs, such as acenocoumarol ( $p = 2.09 \times 10^{-2}$ ), were enriched. Significant dietary habits were also identified in terms of coffee ( $p = 2.61 \times 10^{-3}$ ), tea ( $p = 2.16 \times 10^{-4}$ ), and bitter beverages ( $p = 2.69 \times 10^{-2}$ ) consumption, with caffeine metabolism being also significant ( $p = 3.80 \times 10^{-2}$ ). Significant genetic associations previously reported comprised several urinary and plasma metabolites, different disease impacting metabolism and its tissues, and the consumption of (poly)phenol-rich foods. A representation of these results is reported in Fig. 5.

All the results from the enrichment analysis are reported in Table S3.

## Discussion

In this study, we present a computational functional analysis of 121 curated genes related to (poly)phenol ADME, providing a pooled collection of entries from the main genomic databases and resources to look for when performing genomics analysis in (poly)phenol research studies. Hence, this study provides a reference for human genomics and in general omics research on the bioavailability of (poly)phenols, their inter-individual variability, and, more widely, their effect on human health.

Our analyses confirmed the central role of these genes in biological processes strictly related to (poly)phenol metabolism and bioavailability. Numerous sources showed strong significant associations with different metabolites, like xenobiotics, xanthurenate, and organic acids, and related processes, spanning from transport to general metabolism, highlighting the crucial role of the selected genes. This was further supported by specific molecular functions characterising (poly)phenol metabolism, such as transport, glucuronidation and sulfation,<sup>4,7</sup> key processes in phase II metabolic reactions, and were also observed for specific phenolic-like drugs, like aspirin and paracetamol. Additionally, significant terms were also observed in the consumption of specific (poly)phenol-rich foods like tea and coffee. Supporting evidence was also identified at the tissue level, where enrichment and differential expression patterns were reported in the small intestine, the liver and the colon, core biological sites of (poly)phenol bioavailability and metabolism. The relevance of these genes was reinforced by the unsupervised *k*-means clustering, that confirmed their intricate biological relationships and highlighted the main biological processes and molecular functions<sup>4,7</sup> in (poly)phenol metabolism and bioavailability: these terms were the key representative functions (*i.e.*, the labels) of the eight clusters. Furthermore, this approach provided refined insights into the contribution of specific gene families. Some clusters were mostly identified by specific gene families, as in the case of mucins (MUCs); other clusters comprised complementary gene families, like the transport role carried out by ABC transporters (ABCs) and solute carriers (SLCs), and depicted the multifaceted interplay between biologically distant genes, such



**Fig. 5** Enrichment analysis on the GWAS catalog. The figure shows the enriched traits from the GWAS Catalog presenting common genes related to (poly)phenol bioavailability. Each trait is reported together with its name, the proportion of genes overlapping the total genes associated with the trait, the Bonferroni-adjusted enrichment  $p$ -value, and the list of associated genes.

as UDP-glucuronosyltransferases (UGTs), cytochromes P (CYPs), sulfatase (SULFs) and sulfotransferases (SULTs). This high level of interconnection translates into the impossibility of defining clear-cut clusters, demonstrating and supporting the strong interplay and biological complexity among these genes and their functions.

Beyond consolidating known associations, our findings revealed novel mechanistic insights and previously unreported associations into (poly)phenol metabolism. Among the 121 curated genes that were investigated, the gene network and  $k$ -means clustering analyses reported the lack of connections of these 5 genes, comprising UGT8, SULF1, SLC16A6, ABCD3, and Aldo-Keto Reductase Family 7 Member A2 (AKR7A2). These genes are involved in key metabolic processes, such as glucuronidation (UGT8), sulfation (SULF1), and transport (ABCD3 and SLC16A6), with AKR7A2 mostly involved in detoxification. Nevertheless, these genes belong to large and biologically relevant gene families, suggesting that their apparent isolation may reflect under-representation and slow curation of databases rather than a lack of functional relevance, hence opening to new research studies aiming at further confirming their involvement and contribution to (poly)phenol bioavailability.

Significant and enriched gene sets not only provided already known associations, but also suggested novel mechanisms and biological players. Among the metabolites, the most represented significant associations were observed in bilirubin and in bile acids. The preponderant enrichment of bilirubin may suggest possible unreported evidence of bilirubin-mediated disorders. An example is Gilbert's syndrome, a benign liver condition caused by a variant in the UGT1A1 pro-

moter gene and associated with high bilirubin levels and low risk of cardiovascular disease.<sup>39–41</sup> Other bilirubin-dependent associations like kernicterus and Crigler-Najjar syndrome were found to be associated, more likely due to their relationship with reported traits rather than to (poly)phenol metabolism and effects. In general, these findings suggest that variability in circulating phenolic concentrations may modulate the penetrance of metabolic syndrome, thereby prompting consideration of how differential phenolic exposure may contribute to the development of phenotypically related syndromes. Moreover, bile acids and salts were among the most represented metabolites, with regards to their recycling process. Notably, their significance is important since many (poly)phenols may be subjected to extended enterohepatic recirculation, hence highlighting novel mechanisms worth to be investigated.<sup>4,7</sup> Furthermore, associations of metabolites comprised also different drugs having phenolic-like motifs (benzene), like aspirin, paracetamol, morphine, tamoxifen, and codeine. These drugs were highlighted for their interactions with genes from UGT, SULT, MUC and CYP families, key contributors to core functions of (poly)phenol ADME, suggesting these genes as possible candidate pharmacogenes.<sup>42,43</sup> Hence, these results suggest novel and previously unreported mechanisms, paving the way for individualised pharmacological approaches mediated by food and nutrients, contributing to better understand the bioavailability of dietary secondary compounds and other xenobiotics.

Resulting associations with human diseases reinforced the role of (poly)phenols and their interaction with the human genome in modulating health. In relation to cancer, we identified that NFE2L2, normally involved in absorption and metab-



olism, was enriched with oncogenic signatures. Additionally, some genes belonging to the UGT2 and CYP2C families seemed to drive associations with the KRAS oncogene; these findings may suggest a pleiotropic effects between metabolism and oncogenic signalling, opening to possible investigation on the therapeutic effects of (poly)phenols in cancer, as recently observed in breast and colorectal cancers.<sup>44,45</sup> In parallel, several enriched terms supported the relationship between (poly)phenols and cardiometabolic health, identifying associations with genes involved in inflammation and inflammatory response, like TLR2 and CCL2,<sup>46</sup> and acting on risk factors like low-density lipoprotein, as reported by genes belonging to the UGT1A family.<sup>47,48</sup> These findings implicated the role of (poly)phenols as mediator of cardiometabolic risk, as investigated by ongoing research studies,<sup>17</sup> suggesting their modulatory effects in systemic metabolic and inflammatory conditions.

Finally, these novel findings proposed genetic regulation as a key mechanism shaping the biological effects of (poly)phenols. Different positional gene sets resulted to be significant, particularly in four chromosomal locations on chromosomes 1, 2, 4, and 7. Of note, these genomic regions were characterised by the presence of several members of key gene families together with apparently unrelated genes: chromosome 4 contained UGT2A and UGT2B families, and SULT1B1, SULT1E1 and MUC7 genes; chromosome 2 contained the UGT1A family; chromosome 7 contained members of the CYP3A family together with MUC12 and MUC17 genes; and chromosome 1 contained members of the FMO family. Despite the genomic position may be uninformative and only provide information on the location of the input genes, the enrichment of these gene sets suggests hints on unexplored biological features and additive effects, such as gene regulation, dosage compensation, epigenetic silencing and other regional effects.<sup>49</sup> Furthermore, the possible uncovered regulatory mechanisms were further supported by the enrichment for miRNAs and TF targets, highlighting the importance of regional effects and proposing novel candidates worth investigating.<sup>50</sup>

Nevertheless, this study presents several limitations that needs to be considered to contextualise its findings. Despite the curation of the 121 gene and the proposed reference, functional studies are needed to validate the role of these genes and their contribution to (poly)phenol metabolism and bioavailability. The limited resources regarding (poly)phenols on public available databases for human genomics reduced the power and resolution of this study, likely not fully exploiting the biological connection and relevance of these genes and their associations with human traits and diseases. In addition, the generalisation of our findings remains limited, emphasising the need for exploratory investigations and functional validation studies to support or denied current findings.

Looking forward, this study and the proposed reference favours for targeted applications in the context of human diseases and traits. The significant enrichment for bilirubin makes Gilbert's syndrome a promising case study. The role of UGT1A1 in the onset of this benign syndrome and its association with low glucuronidation ability is well recognised.

However, the possible role and implication of other metabolic processes have been poorly investigated: evidence comprising sulfation or other function are almost absent, making them worth to be investigated. At the same time, our investigation suggested the importance of genes related to (poly)phenol metabolism at the regulatory level. Functional validation will be essential to move beyond static genetic associations, expanding our current knowledge on the biological determinants shaping the inter-individual response to (poly)phenols and gaining a complete mechanistic understanding. Finally, the proposed reference implies its application to health. The observed associations and enrichment with numerous traits and diseases, spanning from cancer to complex disorders, underline the needs to explore the intersection between nutrition, metabolism, genetics and disease risk and progression, opening the way for precision nutrition approaches targeting a wide range of disorders. Overall, we might argue that these genes not only play a role in (poly)phenol ADME but also in the health status of the individual, highlighting the need for dedicated genomics and multi-omics investigations able to confirm previous associations and validate novel hypotheses. Genetic polymorphisms in genes interacting with or related to (poly)phenol ADME and their impact on health status may be key for future precision nutrition approaches of these dietary bioactive compounds.<sup>13,19</sup>

## Conclusion

In this study, we presented a computational functional analysis on 121 curated human genes related to (poly)phenol bioavailability. Our investigation confirmed the role of these candidate genes in various biological processes related to (poly)phenols in the human organism and provided novel insights that require further validation but need to be taken into consideration when performing genomics and omics analyses on human data. To sum up, we proposed an answer to the question "What to expect when performing human genomics investigation in (poly)phenol research studies?". Hence, we reported a new reference for omics investigation and protein-protein interactions in nutritional and (poly)phenol research that will boost our understanding of (poly)phenol ADME and will advance the available knowledge for personalised nutritional approaches. This framework may be transferred to other omics, such as metagenomics by integrating the current available knowledge at the intersection with (poly)phenols, as well as on other plant bioactive compounds, targeting the human genes that are mainly involved in their metabolism and bioavailability, to advance research on nutritionally relevant components contributing to human health.

## Author contributions

Conceptualization: M. T., D. M., P. M.; data curation: M. T., L. G., N. T.; formal analysis: M. T.; funding acquisition: P. M.; investigation: M. T.; methodology: M. T.; project adminis-



tration: P. M.; resources: M. T., A. M. M., D. D. R., P. M.; software: M. T.; supervision: D. M., P. M.; validation: M. T., P. M.; visualization: M. T., L. G.; writing – original draft: M. T., N. T.; writing – review & editing: M. T., L. G., N. T., C. M., A. M. M., V. B., D. D. R., D. M. P. M.

## Conflicts of interest

There are no conflicts to declare.

## Data availability

The data supporting this article have been included as part of the supplementary information (SI). Supplementary information is available. See DOI: <https://doi.org/10.1039/d5fo03227j>. Supplementary Fig. 1: Schematic representation of the *k*-means clustering on the STRING network. Supplementary Table 1: Complete list of enriched terms from the STRING network. Supplementary Table 2: STRING 8-means clustering. Supplementary Table 3: Gene-sets enrichment analysis from FUMAGWAS GENE2FUNC.

## Acknowledgements

The research was funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (PREDICT-CARE project, grant agreement No 950050) and the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.3 – Call for tender No. of 15 March 2022 of Italian Ministry of University and Research funded by the European Union – NextGenerationEU; Award Number: Project code PE00000003, Concession Decree No. 1550 of 11 October 2022 adopted by the Italian Ministry of University and Research, CUP D93C22000890001, Project title “ON Foods – Research and innovation network on food and nutrition Sustainability, Safety and Security –Working ON Foods”. This work was carried out in the frame of the ALIFAR project, funded by the Italian Ministry of University through the program ‘Dipartimenti di Eccellenza 2023–2027’.

## References

- C. Manach, A. Scalbert, C. Morand, C. Rémésy and L. Jiménez, Polyphenols: food sources and bioavailability, *Am. J. Clin. Nutr.*, 2004, **79**(5), 727–747.
- S. Quideau, D. Deffieux, C. Douat-Casassus and L. Pouységu, Plant Polyphenols: Chemical Properties, Biological Activities, and Synthesis, *Angew. Chem., Int. Ed.*, 2011, **50**(3), 586–621.
- A. Rana, M. Samtiya, T. Dhewa, V. Mishra and R. E. Aluko, Health benefits of polyphenols: A concise review, *J. Food Biochem.*, 2022, **46**(10), e14264.
- A. Rodriguez-Mateos, D. Vauzour, C. G. Krueger, D. Shanmuganayagam, J. Reed, L. Calani, *et al.*, Bioavailability, bioactivity and impact on health of dietary flavonoids and related compounds: an update, *Arch. Toxicol.*, 2014, **88**(10), 1803–1853.
- G. Di Pede, P. Mena, L. Bresciani, T. M. Almutairi, D. Del Rio and M. N. Clifford, *et al.*, Human colonic catabolism of dietary flavan-3-ol bioactives, in *Molecular Aspects of Medicine*, Elsevier Ltd, 2023, vol. 89.
- G. Di Pede, P. Mena, L. Bresciani, M. Achour, R. M. Lamuela-Raventós, R. Estruch, *et al.*, Revisiting the bioavailability of flavan-3-ols in humans: A systematic review and comprehensive data analysis, *Mol. Aspects Med.*, 2023, **89**, 101146.
- D. Del Rio, A. Rodriguez-Mateos, J. P. E. Spencer, M. Tognolini, G. Borges and A. Crozier, Dietary (Poly)phenolics in Human Health: Structures, Bioavailability, and Evidence of Protective Effects Against Chronic Diseases, *Antioxid. Redox Signal.*, 2013, **18**(14), 1818–1892.
- S. Liu, H. Zheng, R. Sun, H. Jiang, J. Chen, J. Yu, *et al.*, Disposition of Flavonoids for Personal Intake, *Curr. Pharmacol. Rep.*, 2017, **3**(4), 196–212.
- A. Crozier, D. Del Rio and M. N. Clifford, Bioavailability of dietary flavonoids and phenolic compounds, *Mol. Aspects Med.*, 2010, **31**(6), 446–467.
- D. Milenkovic, C. Morand, A. Cassidy, A. Konic-Ristic, F. Tomás-Barberán, J. M. Ordovas, *et al.*, Interindividual Variability in Biomarkers of Cardiometabolic Health after Consumption of Major Plant-Food Bioactive Compounds and the Determinants Involved, *Adv. Nutr.*, 2017, **8**(4), 558–570.
- C. D. Kay, N. Tejera, A. Jennings, S. Haldar, D. Bevan, L. C. Crossman, *et al.*, Effect of age and sex on the urinary elimination of a single dose of mixed flavonoids: results from a single-arm intervention in healthy United Kingdom adults, *Am. J. Clin. Nutr.*, 2025, **122**(1), 101–111.
- H. N. Rajha, A. Paule, G. Aragonès, M. Barbosa, C. Caddeo, E. Debs, *et al.*, Recent Advances in Research on Polyphenols: Effects on Microbiota, Metabolism, and Health, *Mol. Nutr. Food Res.*, 2022, **66**(1), e2100670.
- C. Favari, J. F. Rinaldi de Alvarenga, L. Sánchez-Martínez, N. Tosi, C. Mignogna, E. Cremonini, *et al.*, Factors driving the inter-individual variability in the metabolism and bioavailability of (poly)phenolic metabolites: A systematic review of human studies, *Redox Biol.*, 2024, **71**, 103095.
- L. Narduzzi, V. Agulló, C. Favari, N. Tosi, C. Mignogna, A. Crozier, *et al.*, (Poly)phenolic compounds and gut microbiome: new opportunities for personalized nutrition, *Microbiome Res. Rep.*, 2022, **1**(2), 16.
- W. Si, Y. Zhang, X. Li, Y. Du and Q. Xu, Understanding the Functional Activity of Polyphenols Using Omics-Based Approaches, *Nutrients*, 2021, **13**(11), 3953.
- J. Hu, R. Mesnage, K. Tuohy, C. Heiss and A. Rodriguez-Mateos, (Poly)phenol-related gut metabolites and human health: an update, *Food Funct.*, 2024, **15**(6), 2814–2835.



- 17 P. Mena, C. Mignogna, N. Tosi, E. Monica, V. Agulló, A. Rosi, *et al.*, Development of an Oral (Poly)Phenol Challenge Test (Opct) to Identify Aggregate Metabotypes for Dietary (Poly)Phenols and Their Drivers: A Study Protocol, *Curr. Dev. Nutr.*, 2022, **6**, 1148.
- 18 N. Tosi, C. Favari, L. Bresciani, E. Flanagan, M. Hornberger, A. Narbad, *et al.*, Unravelling phenolic metabotypes in the frame of the COMBAT study, a randomized, controlled trial with cranberry supplementation, *Food Res. Int.*, 2023, **172**, 113187.
- 19 M. A. Martínez-González, F. J. Planes, M. Ruiz-Canela, E. Toledo, R. Estruch, J. Salas-Salvadó, *et al.*, Recent advances in precision nutrition and cardiometabolic diseases, *Rev. Esp. Cardiol.*, 2025, **78**(3), 263–271.
- 20 S. Haldar, N. T. Hernandez, L. Ostertag, P. Curtis, A. Cassidy and A. M. Minihane, Genetic and phenotypic determinants of flavonoid absorption and metabolism: the COB study, *Arch. Public Health*, 2014, **72**(S1), O3.
- 21 D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, *et al.*, The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest, *Nucleic Acids Res.*, 2023, **51**(D1), D638–D646.
- 22 M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, *et al.*, Gene Ontology: tool for the unification of biology, *Nat. Genet.*, 2000, **25**(1), 25–29.
- 23 S. A. Aleksander, J. Balhoff, S. Carbon, J. M. Cherry, H. J. Drabkin, D. Ebert, *et al.*, The Gene Ontology knowledgebase in 2023, *Genetics*, 2023, **224**(1), iyad031.
- 24 M. Kanehisa, KEGG: Kyoto Encyclopedia of Genes and Genomes, *Nucleic Acids Res.*, 2000, **28**(1), 27–30.
- 25 M. Kanehisa, M. Furumichi, Y. Sato, Y. Matsuura and M. Ishiguro-Watanabe, KEGG: biological systems database as a model of the real world, *Nucleic Acids Res.*, 2025, **53**(D1), D672–D677.
- 26 M. Kanehisa, Toward understanding the origin and evolution of cellular organisms, *Protein Sci.*, 2019, **28**(11), 1947–1951.
- 27 M. Milacic, D. Beavers, P. Conley, C. Gong, M. Gillespie, J. Griss, *et al.*, The Reactome Pathway Knowledgebase 2024, *Nucleic Acids Res.*, 2024, **52**(D1), D672–D678.
- 28 M. Kutmon, A. Riutta, N. Nunes, K. Hanspers, E. L. Willighagen, A. Bohler, *et al.*, WikiPathways: capturing the full diversity of pathway knowledge, *Nucleic Acids Res.*, 2016, **44**(D1), D488–D494.
- 29 A. Agrawal, H. Balci, K. Hanspers, S. L. Coort, M. Martens, D. N. Slenter, *et al.*, WikiPathways 2024: next generation pathway database, *Nucleic Acids Res.*, 2024, **52**(D1), D679–D689.
- 30 T. E. Putman, K. Schaper, N. Matentzoglou, V. P. Rubinetti, F. S. Alquaddoomi, C. Cox, *et al.*, The Monarch Initiative in 2024: an analytic platform integrating phenotypes, genes and diseases across species, *Nucleic Acids Res.*, 2024, **52**(D1), D938–D949.
- 31 D. Grissa, A. Junge, T. I. Oprea and L. J. Jensen, Diseases 2.0: a weekly updated database of disease–gene associations from text mining and data integration, *Database*, 2022, **2022**, baac019.
- 32 O. Palasca, A. Santos, C. Stolte, J. Gorodkin and L. J. Jensen, TISSUES 2.0: an integrative web resource on mammalian tissue expression, *Database*, 2018, **2018**, bay003.
- 33 J. X. Binder, S. Pletscher-Frankild, K. Tsafou, C. Stolte, S. I. O'Donoghue, R. Schneider, *et al.*, COMPARTMENTS: unification and visualization of protein subcellular localization evidence, *Database*, 2014, **2014**(0), bau012–bau012.
- 34 K. Watanabe, E. Taskesen, A. van Bochoven and D. Posthuma, Functional mapping and annotation of genetic associations with FUMA, *Nat. Commun.*, 2017, **8**(1), 1826.
- 35 S. C. Dyer, O. Austine-Orimoloye, A. G. Azov, M. Barba, I. Barnes, V. P. Barrera-Enriquez, *et al.*, Ensembl 2025, *Nucleic Acids Res.*, 2025, **53**(D1), D948–D957.
- 36 A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo and J. P. Mesirov, Molecular signatures database (MSigDB) 3.0, *Bioinformatics*, 2011, **27**(12), 1739–1740.
- 37 J. MacArthur, E. Bowler, M. Cerezo, L. Gil, P. Hall, E. Hastings, *et al.*, The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog), *Nucleic Acids Res.*, 2017, **45**(D1), D896–D901.
- 38 K. G. Ardlie, D. S. Deluca, A. V. Segrè, T. J. Sullivan, T. R. Young, E. T. Gelfand, *et al.*, The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans, *Science*, 2015, **348**(6235), 648–660.
- 39 D. King and M. Armstrong, Overview of Gilbert's syndrome, *Drug Ther. Bull.*, 2019, **57**(2), 27–31.
- 40 L. Vitek and C. Tiribelli, Gilbert's syndrome revisited, *J. Hepatol.*, 2023, **79**(4), 1049–1055.
- 41 J. P. Miranda, A. Pereira, C. Corvalán, J. F. Miquel, G. Alberti, J. C. Gana, *et al.*, Genetic determinants of serum bilirubin using inferred native American gene variants in Chilean adolescents, *Front. Genet.*, 2024, **15**, 1382103.
- 42 V. Kastrinou Lampou, B. Poller, F. Huth, A. Fischer, G. A. Kullak-Ublick, M. Arand, *et al.*, Novel insights into bile acid detoxification via CYP, UGT and SULT enzymes, *Toxicol. in Vitro*, 2023, **87**, 105533.
- 43 M. Sidibe, A. Tazzite, H. Jouhadi and H. Dehbi, Impact of CYP2D6, CYP2C9/19, CYP3A4, UGT, and SULT Variability on Tamoxifen Metabolism in Breast Cancer Treatment, *J. Curr. Oncol.*, 2023, **6**(2), 61–67.
- 44 A. Tascioglu Aliyev, E. Panieri, V. Stepanić, H. Gurer-Orhan and L. Saso, Involvement of NRF2 in Breast Cancer and Possible Therapeutical Role of Polyphenols and Melatonin, *Molecules*, 2021, **26**(7), 1853.
- 45 S. Mehdinejad, M. Peymani, A. Salehzadeh and M. Zaefizadeh, Genetic insights and therapeutic potential for colorectal cancer: mutation analysis of KRAS gene and efficacy of Oleuropein-conjugated iron oxide nanoparticles, *Naunyn Schmiedeberg's Arch. Pharmacol.*, 2024, **397**(11), 8771–8783.
- 46 A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov and P. Tamayo, The Molecular Signatures



- Database Hallmark Gene Set Collection, *Cell Syst.*, 2015, **1**(6), 417–425.
- 47 M. S. Sandhu, D. M. Waterworth, S. L. Debenham, E. Wheeler, K. Papadakis, J. H. Zhao, *et al.*, LDL-cholesterol concentrations: a genome-wide association study, *Lancet*, 2008, **371**(9611), 483–491.
- 48 B. A. Ference, H. N. Ginsberg, I. Graham, K. K. Ray, C. J. Packard, E. Bruckert, *et al.*, Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. Evidence from genetic, epidemiologic, and clinical studies. A consensus statement from the European Atherosclerosis Society Consensus Panel, *Eur. Heart J.*, 2017, **38**(32), 2459–2472.
- 49 K. De Preter, R. Barriot, F. Speleman, J. Vandesompele and Y. Moreau, Positional gene enrichment analysis of gene sets for high-resolution identification of overrepresented chromosomal regions, *Nucleic Acids Res.*, 2008, **36**(7), e43–e43.
- 50 T. Ruskovska, I. Budić-Leto, K. F. Corral-Jara, V. Ajdžanović, A. Arola-Arnal, F. I. Bravo, *et al.*, Systematic analysis of nutrigenomic effects of polyphenols related to cardiometabolic health in humans – Evidence from untargeted mRNA and miRNA studies, *Ageing Res. Rev.*, 2022, **79**, 101649.

