A Section 508–conformant HTML version of this article is available at https://doi.org/10.1289/EHP4713.

Generating the Blood Exposome Database Using a Comprehensive Text Mining and Database Fusion Approach

Dinesh Kumar Barupal[1] *and Oliver Fiehn*[1]

[1]National Institutes of Health (NIH) West Coast Metabolomics Center, Genome Center, University of California, Davis, Davis, California, USA

Background: Blood chemicals are routinely measured in clinical or preclinical research studies to diagnose diseases, assess risks in epidemiological research, or use metabolomic phenotyping in response to treatments. A vast volume of blood-related literature is available via the PubMed database for data mining.

Objectives: We aimed to generate a comprehensive blood exposome database of endogenous and exogenous chemicals associated with the mammalian circulating system through text mining and database fusion.

Methods: Using NCBI resources, we retrieved PubMed abstracts, PubChem chemical synonyms, and PMC supplementary tables. We then employed text mining and PubChem crowdsourcing to associate phrases relating to blood with PubChem chemicals. False positives were removed by a phrase pattern and a compound exclusion list.

Results: A query to identify blood-related publications in the PubMed database yielded 1.1 million papers. Matching a total of 15 million synonyms from 6.5 million relevant PubChem chemicals against all blood-related publications yielded 37,514 chemicals and 851,999 publications records. Mapping PubChem compound identifiers to the PubMed database yielded 49,940 unique chemicals linked to 676,643 papers. Analysis of open-access metabolomics papers related to blood phrases in the PMC database yielded 4,039 unique compounds and 204 papers. Consolidating these three approaches summed up to a total of 41,474 achiral structures that were linked to 65,957 PubChem CIDs and to over 878,966 PubMed articles. We mapped these compounds to 50 databases such as those covering metabolites and pathways, governmental and toxicological databases, pharmacology resources, and bioassay repositories. In comparison, HMDB, the Human Metabolome Database, links 1,075 compounds to blood-related primary publications.

Conclusion: This new Blood Exposome Database can be used for prioritizing chemicals for systematic reviews, developing target assays in exposome research, identifying compounds in untargeted mass spectrometry, and biological interpretation in metabolomics data. The database is available at http://bloodexposome.org. https://doi.org/10.1289/EHP4713

Introduction

Human blood is the most commonly used sample matrix in clinical as well as in epidemiological studies. In this context, factors measured in blood are investigated that indicate a subject's health status (Chaleckis et al. 2016) or risk for chronic diseases (Miranti et al. 2017). Typically, such factors are built into the study design along with smoking status, food frequency questionnaires, sex, race, body mass index, and age. Yet these factors do not reflect exposures to exogenous chemicals and phenotypes of human health such as endogenous metabolites. Thousands of chemicals from food, drugs, household chemicals, and environmental pollutants enter the bloodstream. Information on these compounds could stratify subject cohorts to actual exposures (instead of relying on questionnaires) and could also be used to estimate risks of specific compounds impacting health. Furthermore, exposure chemicals may be biochemically transformed by cellular enzymes, for example, in the liver (Gu and Manautou 2012) or by gut microbes (Koppel et al. 2017). The dynamics of chemical intake, metabolism, and transport govern the human blood's chemical composition and concentrations (Bray et al. 2018). Therefore, the overall dynamics of chemical exposures and bodily metabolism (metabolomics) must be studied by comprehensive blood chemical analysis. For example, increased blood homocysteine was shown to modify stroke risk in a subset of hypertensive Chinese adults (Zhao et al. 2017), while serum 4-hydroxynonenal was shown to be an oxidative stress biomarker in rats (Kim et al. 2015). The term exposome is defined by the totality of exposures during a lifespan (Wild 2005). The blood exposome concept includes performing association modeling on all the chemicals present in human blood (Rappaport et al. 2014; Vlaanderen et al. 2017). It requires measuring all these compounds using analytical assays, analyzing the levels of those compounds, and then interpreting them in the contexts of phenotypes and physiology (Dennis et al. 2017). Since only a small portion of risk for chronic diseases can be explained by genetic factors, a large portion of the remaining risk can be attributed to these chemicals that are present in the blood. It has been strongly suggested that studies of exposomics and genomics will provide a better estimate of risk and causes of chronic diseases (Wild 2005).

Comprehensive chemical analysis in blood has tremendously improved over the past 20 y. Single-molecule assays are being replaced with multitargets assays or untargeted chemical profiling offered by mass spectrometry–based targeted and untargeted metabolomics assays (Marksteiner et al. 2018; Cohen et al. 2018). The increased breadth of the analysis has the potential to provide quantitative data for up to 900 compounds (Barupal et al. 2019; Hu et al. 2019; Long et al. 2017; Price et al. 2017), as exemplified in Karl et al. (2017), where untargeted metabolomics was used to compare 737 chemicals in the plasma of soldiers before and after rigorous training exercises. Untargeted plasma metabolomics has been routinely used in clinical and epidemiological settings to identify exposure biomarkers (Rothwell et al. 2014) and chronic disease risk factors (Li et al. 2018). Two major challenges still remain. How can we interpret changes in the level of hundreds of compounds (Karl et al. 2017; Miller et al. 2016), and how can we identify the chemical structure of detected

Address correspondence to Dinesh Kumar Barupal, Genome Center, University of California, Davis, 451 East Health Science Dr., Davis, CA 95616 USA. Telephone: (530) 979-4354. Email: Dinkumar@ucdavis.edu; Oliver Fiehn, Genome Center, University of California, Davis, 451 East Health Science Dr., Davis, CA 95616 USA. Telephone: (530) 754-8258. Email: Ofiehn@ucdavis.edu

The authors declare that they have no actual or potential competing financial interests.

Supplemental Material is available online (https://doi.org/10.1289/EHP4713).

Received 6 November 2018; Revised 9 September 2019; Accepted 11 September 2019; Published 26 September 2019.

signals in mass spectrometry profiling? The most logical step is to map all detected signals to a list of all known chemicals that have been reported in the literature in mammalian blood. Yet such a comprehensive database does not exist. Many compounds are found in the literature that are not covered by existing repositories such as the Human Metabolome Database (HMDB) (Wishart et al. 2018), which was manually curated. Yet such repositories are necessarily incomplete due to the large volume of literature. Instead, text mining has been used in other important areas of biomedical research such as tagging genes (Funk et al. 2014) and diseases (Jimeno et al. 2008), mapping complex biological relationships (Ananiadou et al. 2010; Pyysalo et al. 2012), and understanding mechanisms and disease contexts (Jensen et al. 2006; Zhu et al. 2013). However, to our knowledge, text mining has not been used to date for exposome analysis of blood-associated chemicals. We show here, through text mining and database fusion, how to construct a complete list of all chemicals currently reported to be associated with the mammalian circulatory system.

Methods

PubMed Data

Blood-related (relevant publications that reported measurements of chemicals in blood) papers were identified by a PubMed query using the combined keywords (blood[Title/Abstract] OR serum [title/abstract] OR plasma[title/abstract] OR circulating[title/abstract]) AND (level* OR concentration* OR content OR value*) AND has_abstract[filt] AND eng[language] NOT review[pt]". Query results were downloaded in the XML format from the PubMed website interface. Two-word-pair phrases ("shingle" in text mining), for example, "plasma glucose" or "blood pressure," were extracted from abstracts and titles for "blood OR sera OR serum OR circulating OR plasma." Shingles were manually reviewed by the authors, and those that did not reflect blood chemical analysis were marked as false positives and consequently compiled into a phrase exclusion list (Excel Table S1). PubMed abstracts that only had these exclusion phrases were removed from the list of literature reports on blood chemicals.

PubChem Data

The following four mapping files were downloaded from the PubChem FTP server (ftp://ftp-private.ncbi.nlm.nih.gov/pubchem/Compound/) on 24 April 2019 and saved in a tab-separated text format: *a*) PubChem CID to chemical synonyms, *b*) PubChem CID to Biosystems, *c*) PubChem CID to PubMed identifiers, and *d*) PubChem CID to InChIKey.

We manually defined PubChem structure depositors that were relevant to exposome research using mission descriptions of the associated organizations (Excel Table S2). A composite query of terms and depositors was compiled to retrieve relevant structures from PubChem (https://www.ncbi.nlm.nih.gov/pccompound) using the following search strategy: "Tox21"[SourceName] OR "NIST"[SourceName] OR "NIAID"[SourceName] OR "NCBI Structure"[SourceName] OR "DTP/NCI"[SourceName] OR Nikkaji[SourceName] OR "Broad Institute"[SourceName] OR "NCGC"[SourceName] OR "MLSMR"[SourceName] OR "MetabolomicsWorkbench"[SourceName] OR "EPA Substance Registry Services"[SourceName] OR "ChemIDplus"[SourceName] OR "Human Metabolome Database"[SourceName] OR "Springer Nature"[SourceName] OR "NCBI Structure"[SourceName] OR "Comparative Toxicogenomics Database"[SourceName] OR "KEGG"[SourceName] OR "ChEBI"[SourceName] OR "EPA DSSTox"[SourceName] OR "FDA/SPL Indexing Data"[SourceName] OR "pccompound pubmed"[Filter] OR "pccompound pmc"[Filter] OR "has mesh"[Filter] OR "pccompound omim"[Filter] OR "pccompound gene"[Filter] OR "pccompound gds"[Filter] OR "pccompound biosystems"[Filter] OR "has dailymed"[Filter] OR pccompound_pcassay_active[filt] OR "has src nih mlp"[Filter]. One limitation of this approach is that the list covers major exposome-related databases and sources; however, it does not cover databases that have not deposited their chemical structures to the PubChem database.

The PubChem synonym file was used for retrieving synonyms for the CIDs returned by the PubChem search. A manually curated list of synonyms (Excel Table S3) that did not refer to chemical names was prepared, and synonyms from this list were removed from the PubChem synonym list. The filtered list of synonyms was searched against the PubMed abstracts to identify relevant publications that reported measurements of chemicals in blood.

PMC Data

Blood metabolomics papers were identified by a PMC query of the syntax ("metabolomics"[Body - All Words] OR "metabolomic"[Body - All Words] OR "metabolite profiling"[Body - All Words] OR "metabolic profiling"[Body - All Words] OR ("metabolome"[Body - All Words] OR "metabolome"[Body - All Words])) AND (blood[Body - All Words] OR serum[Body - All Words] OR plasma[Body - All Words] OR circulating[Body - All Words] OR blood[Body - All Words] OR serum[Body - All Words] OR plasma[Body - All Words] OR circulating[Body - All Words]) AND "open access"[filter] AND "has suppdata"[Filter]. Supplementary data tables (CSV, XLS, or DOC) for metabolomic papers were retrieved using the PMC open-access web service. These tables were extracted using an R script (version 3.4.1; R Development Core Team) by identifying chemical names by matching table entries against the mapping file "PubChem CID to chemical synonyms." The PMC database provides 5.6 million full-text articles (as of September 2019) for reading online at their website; however, only a subset of these articles are available for download and for subsequent computational text mining. The PMC copyright policy is available at https://www.ncbi.nlm.nih.gov/pmc/tools/textmining/.

Human Metabolome Database Data

Serum metabolite annotations were downloaded from the HMDB website in an XML format. Using an R script, biofluid and PubMed annotations were extracted for each metabolite in the database. To focus on primary publications instead of generic databases or reviews, the top 10 publications (based on total counts of chemicals and citation accuracy; see Excel Table S4) and their associated compounds were removed from the list.

Merging Data

A master list of unique PubChem CIDs was created by combining the results of PubChem, PubMed, and PMC blood metabolomics queries (Excel Table S5). A hashed version of an International Chemical Identifier (InChI) and Simplified Molecular Input Line Entry System for PubChem CIDs were downloaded from the PubChem identifiers exchange service. Molecular structures have unique chemical descriptors of the three-dimensional structure defined by InChIKeys. While, in many cases, researchers may truly identify specific metabolic stereoisomers (such as distinguishing glucose and galactose), in other cases, scientists may have used ambiguous names. We therefore used the first block of the InChIKeys, describing the two-dimensional chemical structure, to use a conservative approach for analyzing the blood exposome. Also, structures with noncovalent bonds, such as those found in a salt of a chemical, for instance, "Octenidine dihydrochloride" and

"Octenidine," were flagged in the database. Then we removed structures that were labeled with stable isotopes (such as ^{2}D or ^{13}C).

Using the final list of PubChem CIDs, the structure definition files (SDFs) of all blood exposome compounds were downloaded from the PubChem structure download utility. Information of exact monoisotopic mass and lipophilicity (XlogP) was extracted from the SDF. To test the presence of chemicals in databases relevant to exposome research, we collected and cross-referenced PubChem CIDs of chemicals in 50 different depositors that we manually defined as relevant to exposome research to include environmental pollutants, occupational hazards, drugs, dietary compounds, endogenous compounds, and food additives (Table 1).

Computer Hardware and Software

Calculations were performed using a 4-core (3.7 GHz) computer with 64 GB of memory on 2 TB of solid state drive hard drives. R (version 3.4.1) was used for processing all the files.

Code and Data Availability

The R script is available at https://github.com/barupal/exposome and in Supplemental Material, "The Blood Exposome Database - R script." Web addresses for input information sources are provided in Excel Table S6.

Results

An overview of the workflow as described in the "Methods" section is provided in Figure 1. In the following sections, we describe the general approach and results from each step of the process.

Identifying Publications Using Blood Measurements

The most comprehensive resources for literature-based searches are PubMed with abstracts of over 29 million scientific articles and the PMC full-text repository for about 5.5 million papers which were searched in April 2019. Using synonyms for "blood," we therefore exploited the PubMed database to search for all papers related to "measurements in blood," using the four terms "level," "concentration," "content," and "value," as described in the "Methods" section. In a broad literature search of all terms related to blood, we found a total of 1.4 million papers—too many to be manually assessed. We kept the query terms generic enough to cover both humans and other mammals. In this way, we yielded dedicated exposure studies (e.g., Wang et al. 2012) that were conducted only on animal models such as rats or pigs, showing that the inclusion of animal studies was a valid and important aspect of constructing a full exposome database. However, an intrinsic problem of generic query terms is the identification of many false-positive publication records. For example, queries returned reports that used specific chemicals for treatments in study designs without measurement of chemicals in blood. To remove such publication records, we created a phrase exclusion list (Excel Table S1) of about 10,000 phrases that included phrases such "blood pressure" or "blood transfusion." Such phrases built a context in which chemicals were likely used as actors in clinical or preclinical trials, but not chemicals that were reported as measured in blood. If phrases from the exclusion list occurred only once in an abstract together with a chemical, we removed the corresponding paper from the blood exposome publication list. We removed 350,133 publications in this way, retaining a final list of 1,085,023 literature abstracts with PMIDs from PubMed. This final list of all included PMIDs is available on the Blood Exposome Database website (https://exposome1.fiehnlab.ucdavis.edu/download/pmid_title_abstract_sb.zip).

Querying Chemicals by Synonym Lists

Inspired by the construction of the medical subject headings (MeSH) (Coletti and Bleich 2001) ontology, we used a chemical synonym entity lookup approach. Chemical synonyms were extracted from the PubChem database, which stores over 100 million structures from 634 sources. With our existing resources, it was impractical to query 100 million structures and their synonyms. We therefore prioritized chemicals from 19 sources within the PubChem database that indicated a direct relationship with biology, exposures, or medicine (Excel Table S2) and were potentially relevant to the exposome research. We subsequently created a subset of chemicals from these databases using 10 built-in PubChem search filters that mapped directly to biologically relevant contexts, such as "bioassay" or "present in PubMed," biological pathways, genetic diseases, and drugs (Excel Table S2). Overall, we compiled a total of 7.5 million unique chemicals to be cross-referenced with our list of 1.1 million blood-related associated publications. The full list of included PubChem structures can be retrieved from the PubChem database using the search strategy described in "PubChem Data" in the "Methods" section.

In research papers, scientists use different synonyms for the same chemical. We therefore had to query all synonyms that mapped to the 7.5 million PubChem CID structures in the PubChem compound synonym file that links 147 million chemical names to 100 million PubChem CIDs. PubChem regularly updates this file and provides it through their FTP server (see "PubChem Data" in the "Methods" section). From this file, we retrieved 15 million synonyms for our list of 7.5 million chemicals that we used for all further queries of blood-related chemical reports by use of text mining of PubMed abstracts and PMC supplement tables.

Linking Literature to Chemicals

We used three approaches (Figure 1) to link literature to chemicals: *a*) Chemical synonym name lookups against PubMed abstracts, *b*) direct links between PubChem and PubMed, and *c*) mining blood-related PMC supplement tables for chemical synonyms.

First, we used PubMed directly. The PubMed database can be queried by use of a web interface for a limited number of queries. To run a few hundred queries, NCBI Eutils utilities (https://www.ncbi.nlm.nih.gov/books/NBK25500/) can be used, but not for millions of queries. We therefore downloaded our final list of over 1 million blood-related PubMed abstracts and queried them locally in R. When querying all 15 million synonyms against the 1.1 million blood-related research papers, a total 851,999 PMIDs were returned with 37,514 unique chemicals and 62,809 synonyms. This query also yielded the frequency with which chemicals were reported in our list of included blood-related PubMed abstracts.

Secondly, we used literature information that is directly mapped in PubChem. PubChem maintains a compound/literature file that is regularly updated (ftp://ftp-private.ncbi.nlm.nih.gov/pubchem/Compound/Extras/CID-PMID.gz). New annotations of chemicals with literature are submitted by any new depositor, including new annotations built within the MeSH database. The corresponding PubChem literature mapping file linked 8,565,681 research papers to 1,838,374 unique PubChem CID chemicals. Indeed, the largest contributing depositor originated from the MeSH database with 7,614,881 papers and 119,287 unique chemicals. A total of 282,741 unique compounds mapping to

Table 1. Coverage of blood-related structures in different databases and sources relevant for exposome research.

Source category	Source name and description	Website	Blood chemicals	Application areas
Literature data	PubChem, PubMed	pubchem.ncbi.nlm.nih.gov/	49,542	Blood exposome
	PubMed Abstract	ncbi.nlm.nih.gov/pubmed/	37,070	Blood exposome
	PMC: Blood Metabolomics	ncbi.nlm.nih.gov/pmc/	4,036	Blood exposome
Metabolite databases	Metabolomics Workbench	www.metabolomicsworkbench.org/	11,194	Metabolism, general
	MassBank of North America (MoNA)	https://massbank.us/	7,238	Metabolomics
	Human Metabolome Database	www.hmdb.ca/	7,039	Human metabolism
	LipidMaps database	www.lipidmaps.org/	3,243	Lipid metabolism
Ontology	National Institutes of Health (NIH), NCBI: Medical Subject Headings	www.ncbi.nlm.nih.gov/mesh	39,285	Biological relevance
	Chemical Entities of Biological Interest (ChEBI)	www.ebi.ac.uk/chebi/	15,646	Biological relevance
Pathway databases	Kyoto Encyclopedia of Genes and Genomes (KEGG)	www.genome.jp/kegg/	11,440	Biochemical pathways
	NIH, NCBI: Gene (human)	www.ncbi.nlm.nih.gov/gene	9,190	Precision medicine
	BioCyc	biocyc.org/	5,257	Biochemical pathways
	NIH, NCBI: Structure (Protein)	ncbi.nlm.nih.gov/Structure/	3,878	Precision medicine
	NIH, NCBI: BioSystems Database	www.ncbi.nlm.nih.gov/biosystems/	3,813	Biochemical pathways
	NIH: Online Mendelian Inheritance in Man	www.ncbi.nlm.nih.gov/omim/	2,731	Genetic disorders
Governmentd atabases	Japan Chemical Substance Dictionary (NIKKAJI)	jglobal.jst.go.jp/en/	30,871	Biomonitoring
	U.S. Food and Drug Administration (FDA): Structured Product Labeling	labels.fda.gov/	17,362	Biomonitoring
	European Chemical Agency (ECHA)	echa.europa.eu/	12,368	Biomonitoring
	U.S. National Institute of Standards and Technology: Mass Spectrometry Data Center	chemdata.nist.gov	10,480	Biomonitoring
	U.S. Environmental Agency (EPA): Substance Registry Services	ofmpub.epa.gov/sor_internet/registry/substreg/LandingPage.do	678	Biomonitoring
	U.S. FDA: Food Additive database	www.fda.gov/food/ingredientspackaginglabeling/foodadditivesingredients/default.htm	1,207	Biomonitoring
	U.S. FDA: Center for Food Safety and Applied Nutrition	www.fda.gov/about-fda/office-foods-and-veterinary-medicine/center-food-safety-and-applied-nutrition-cfsan	83	Biomonitoring
Pharmacology	NIH, National Library of Medicine (NLM): DailyMed	dailymed.nlm.nih.gov/dailymed/	4,483	Drugs
	U.S. Department of Agriculture (USDA): Dr. Duke's Phytochemical and Ethnobotanical Database	phytochem.nal.usda.gov/phytochem/search/list	4,135	Food biomarkers
	World Health Organization (WHO): Anatomical Therapeutic Chemical Classification System	www.who.int/medicines/regulation/medicines-safety/toolkit_atc/en/	3,754	Drugs
	Logical Observation Identifiers Names and Codes	loinc.org/	1,812	Clinical assays
	U.S. FDA: Endocrine Disruptor Knowledge Base	www.fda.gov/science-research/bioinformatics-tools/endocrine-disruptor-knowledge-base	821	Endocrine disrupters
Toxicological databases	U.S. EPA: Distributed Structure-Searchable Toxicity (DSSTOX)	www.epa.gov/chemical-research/distributed-structure-searchable-toxicity-dsstox-database	21,427	Exposome: toxicants
	Comparative Toxicogenomics Database	ctdbase.org/	9,878	Exposome: toxicants
	NIH: Toxicology in the 21st Century	ncats.nih.gov/tox21	6,899	Exposome: toxicants
	U.S. EPA: Toxic Substances Control Act	www.epa.gov/laws-regulations/summary-toxic-substances-control-act	6,515	Exposome: toxicants
	NIH, NLM: Chemical Carcinogenesis Research Information System	toxnet.nlm.nih.gov/newtoxnet/ccris.htm	4,607	Exposome: toxicants
	NIH, NLM: Hazardous Substances Data Bank	toxnet.nlm.nih.gov/newtoxnet/hsdb.htm	4,512	Exposome: toxicants
	NIH, NLM: Information on Hazardous Chemicals and Occupational Diseases	hazmap.nlm.nih.gov/	2,669	Exposome: occupational
	U.S. EPA: Pesticides	www.epa.gov/pesticides	1,851	Exposome: toxicants
	The Organization for Economic Co-operation and Development: Existing Chemicals Database	hpvchemicals.oecd.org/ui/Default.aspx	1,690	Exposome: daily
	NIH, NLM: Household Products Database	householdproducts.nlm.nih.gov/	1,601	Exposome: daily
	International Labor Organization (ILO): International Chemical Safety Cards (ICSC)	www.ilo.org/safework/info/publications/WCMS_113134/lang–en/index.htm	1,311	Exposome: occupational

Note: Descriptions and web addresses for these sources and databases are provided in Table S6. PubChem CIDs from each database and sources were cross-referenced against the master list of PubChem CIDs in the Blood Exposome Database.

Table 1. (*Continued.*)

Source category	Source name and description	Website	Blood chemicals	Application areas
	New Jersey Right to Know: Hazardous Substance List	www.nj.gov/health/workplacehealthandsafety/right-to-know/hazardous-substances/	1,271	Exposome: toxicants
	California Office of Environmental Health Hazard Assessment	oehha.ca.gov/	1,013	Exposome: toxicants
	U.S. Centers for Disease Control and Prevention (CDC), National Institute for Occupational Safety and Health (NIOSH)	www.cdc.gov/niosh/index.htm	828	Exposome: occupational
	California Preposition 65: Safe Drinking Water and Toxic Enforcement Act of 1986	oehha.ca.gov/proposition-65/law/proposition-65-law-and-regulations	787	Exposome: toxicants
	U.S. CDC, Agency for Toxic Substances and Disease Registry	www.atsdr.cdc.gov/	746	Exposome: toxicants
	WHO: International Agency for Research on Cancer (IARC) Monographs	monographs.iarc.fr/	580	Carcinogens
	U.S. EPA: Integrated Risk Information System	www.epa.gov/iris	447	Carcinogens
	USDA: Pesticide Data Program	www.ams.usda.gov/datasets/pdp	340	Exposome: toxicants
	WHO: Joint Food and Agriculture Organization (FAO)/WHO Expert Committee on Food Additives	www.who.int/foodsafety/areas_work/chemical-risks/jecfa/en/	259	Food additives
BioAssay databases	NIH: Molecular Libraries and Imaging	commonfund.nih.gov/molecularlibraries/index	18,748	Pharmaceuticals
	NIH, National Cancer Institute (NCI): Developmental Therapeutics Program	dtp.cancer.gov/	9,896	Pharmaceuticals
	NIH, National Institute of Allergy and Infectious Diseases (NIAID): screening program	www.niaid.nih.gov	7,508	Pharmaceuticals
	NIH, National Center for Advancing Translational Sciences (NCATS): Chemical Genomics Center	ncats.nih.gov/ncgc	8,788	Pharmaceuticals
	NIH, Common Fund: Molecular Libraries and Imaging program	commonfund.nih.gov/molecularlibraries/index	5,152	Pharmaceuticals
	The Broad Institute of Massachusetts Institute of Technology (MIT) and Harvard	www.broadinstitute.org	5,154	Pharmaceuticals

2,649,881 papers were submitted by various structure depositors, including publishers and governmental agencies, and a total 1,565,939 unique compounds and associated 69,401 papers were submitted by the NCBI BioAssay database. Many pairs of compound–paper associations were submitted by more than one contributor. We cross-referenced the list of 1.1 million publications of blood measurement–related studies with the PubChem literature mapping file, yielding a subset file of 676,643 publications with 49,940 chemicals (Figure 2). This approach led us to retrieve compounds that were not mentioned in the publications abstracts but that were directly provided by different depositors. Those compounds could be mentioned in the main text or supplementary tables of a paper. As an example, N-lactoyl-leucine was reported for human plasma (Jansen et al. 2015) in the main text but not mentioned in the corresponding literature abstract.

Thirdly, we used the 6 million papers deposited in PMC. Because of copyright restrictions, not all papers from the PMC database can be downloaded for text mining. Instead, we

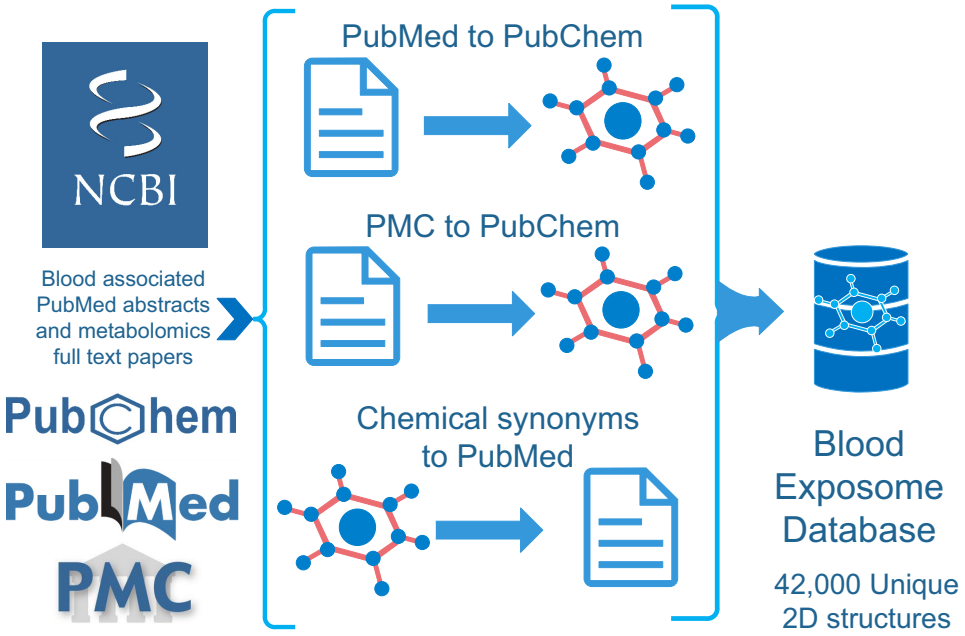

Figure 1. Overview schema for constructing the Blood Exposome Database. Three NCBI hosted databases were used as inputs for the workflow that yielded 42,000 two dimensional structures for blood specimens.

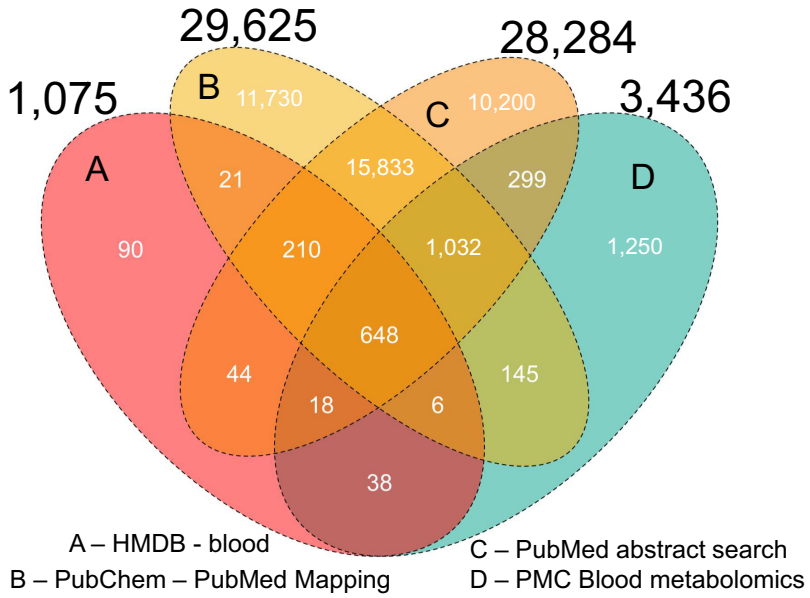

Figure 2. Overlap analysis of the origin of 41,474 achiral blood chemicals. PubChem to PubMed mapping provided the most comprehensive overview of the blood related compounds.

performed a literature search using keywords associated with blood metabolomics publications (see "PMC Data" in the "Methods" section) against the PMC database and retrieved 7,683 open-access papers. For these papers, we downloaded 1,706 supplementary data tables and 7,617 supplementary text files. We then focused on retrieving chemical names reported in the supplementary tables. This approach covered compounds that were neither reported in literature abstracts or directly mapped to the information in the PubChem database (Cohen et al. 2018). Extraction of chemical names from these supplementary tables and mapping them to PubChem database compound CIDs yielded 4,039 chemical compounds linked to 204 papers, providing a consolidated list of chemicals detected by metabolomics assays in blood.

Merging of Compound Lists and Comparative Analysis of Text-Mining Approaches

When inspecting the file contents for the three approaches to link literature to chemicals (chemical synonym name lookups, direct PubChem/PubMed links, and PMC supplement tables), we found that each approach yielded unique new literature-based blood chemicals. We therefore merged the contents of these three mapping lists to obtain a consolidated list of 66,691 chemicals linked to 878,966 publications (Figure 2). For each chemical compound, we obtained metadata such as InChIKey, XlogP, exact mass, molecular formula, and SMILES codes from the PubChem database. Chemicals can have structure variants such as salts, neutral structures, and stereoisomers, leading to multiple CID entries in the PubChem database. These structure variants overestimate the total number of unique chemicals in the Blood Exposome Database. The exposome list was constrained using the structure variant filtering approach (see "Merging Data" in the "Methods" section) to a total of 41,474 achiral two-dimensional structures linked to 65,957 PubChem CIDs and 878,966 publications. The database is provided in Excel Table S5. We investigated chemicals composing the Blood Exposome Database with respect to classic chemical parameters. Median lipophilicity, as indicated by the partition coefficient XlogP, was 2.1 (such as those found for penicillin, estrone sulfate, or resolvin E1) (Figure 3A). Surprisingly, the distribution of reported exposome compounds in blood is normally distributed between hydrophilic compounds (at negative XlogP values, such as phenylalanine) and hydrophobic chemicals [above XlogP 5, such as benz(a)pyrene]. The median mass was found at 318 Da, with a 98% range from 10 to 1,500 Da (Figure 3B). Interestingly, we observed a skewed distribution towards small molecules, likely due to an underrepresentation of blood lipids.

Text mining for lipid names and their transformation into InChIKeys is challenging because researchers often use abbreviations and leave parts of the chemical structure ambiguous, such as TG (54:4). Over 3,000 compounds were reported in more than 100 papers, whereas 20,896 chemicals were linked to a single literature report (Figure 3C). If a compound was frequently detected and reported in blood, we had a high confidence that this chemical is a genuine, bona fide blood exposome compound. Yet we could not simply exclude all compounds that had been rarely reported, but we certainly expected that rarely reported blood chemicals may have had higher false-positives rate. As an example of a false positive, we identified "N-Nitrosopyrrolidine" as a blood chemical in one paper (Tan-ariya et al. 1998), based on the appearance of the synonym "no-pyr" in the abstract; however, on further inspection, we found that the reference to "no-pyr" in the abstract actually referred to the absence of pyrimethamine activity as "no PYR" (Tan-ariya et al. 1998). On the other hand, "hydrazobenzene" was only reported once (Dodd et al. 2012) in the literature as well, but it is also found in the U.S. Food and Drug Administration (FDA) and Environmental Protection Agency (EPA) databases, increasing the confidence that this compound was positively detected in blood.

Next, we investigated the coverage of the blood exposome with respect to the extent at which any single text-mining approach would have already yielded a comprehensive overview on the blood-associated chemicals (Figure 2). When comparing our text mining results directly, we found that 98% of all retrieved blood chemicals were covered by combining results of PubChem–PubMed mapping with results of PubMed abstracts queries. Importantly, both PubChem–PubMed mapping and PubMed abstracts queries resulted in more than 15,000 unique blood compounds each, showing that both approaches were

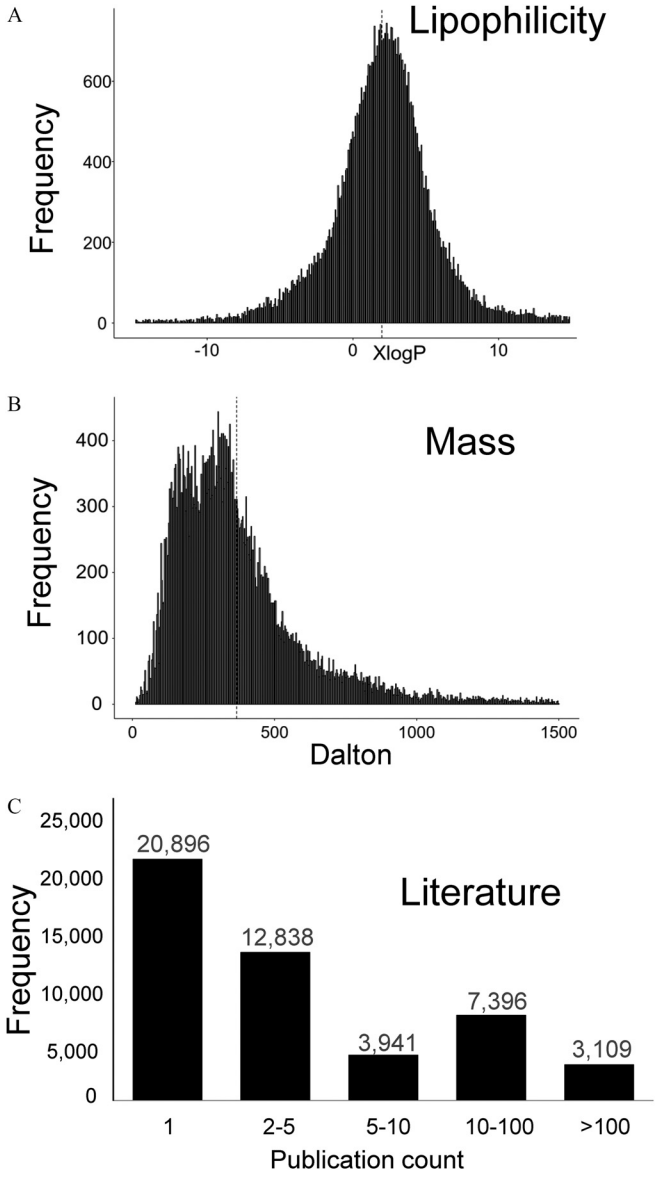

Figure 3. Distribution of lipophilicity (A), molecular weight (B), and publication count (C) in the Blood Exposome Database. The *y*-axis shows the frequency of chemicals. Xlogp is a unitless measurement for lipophilicity, in which negative values indicate more polar compounds.

necessary and complemented each other. In comparison, the PMC blood metabolomics search added fewer additional unique structures. As validation for our approach, we found only 90 unique compounds, with most being measurements of endogenous compounds rather than exogenous exposures when we searched the HMDB (Figure 2). HMDB is a very popular tool that is created through a manual literature search, and it contains exposome compounds (such as food compounds) in addition to classic metabolites. Up to 92% of HMDB's blood-associated compounds were covered by our workflow. Overall, HMDB only contains 1,075 achiral chemical structures with blood annotation with primary literature citations. In comparison, our three text mining approaches (PubChem–PubMed mapping, a PubMed abstracts search, and a PMC blood metabolomics search) yielded 46-fold more annotated blood chemicals. Therefore, our analysis suggests that the current HMDB content greatly underreports the human blood exposome. We successfully annotated 7,039 HMDB compounds with primary literature on blood measurements by use of our text-mining methods. For example, xanthohumol has been reported to be present in the human blood (Legette et al. 2014), information that is currently missing in the HMDB database.

To attain functional and biological contexts and enriching the set of available metadata, we cross-referenced all blood-associated chemicals with chemicals in 50 prioritized databases for the exposome research. Because we did not intend to perform overlap analysis but did plan to perform direct mapping of specific compounds to databases, we used individual PubChem CID entries. Table 1 gives a summary of how many compounds were successfully mapped to metabolite databases, pathway and enzyme repositories, or databases focusing on ontologies, bioassays, governmental agencies, pharmacology, or toxicology. The largest contribution was found through the MeSH, the Kyoto Encyclopedia of Genes and Genomes (KEGG), the Japan Chemical Substance Dictionary (NIKKAJI), NIH Molecular Libraries Program (NIH_MLP), Distributed Structure-Searchable Toxicity (DSSTOX), and the National Library of Medicine DailyMed databases.

The purpose of these database assignments is to enable a prioritization for specific investigations. For example, the Tox21 database of the U.S. EPA contained 6,899 known toxicants and pollutants that we found as detected in blood samples. Such compounds might be used as candidates for expanding biomonitoring programs such as the National Health and Nutrition Examination Survey (NHANES) (Sobus et al. 2015) or the California Biomonitoring program (Mann 2018). Secondly, we investigated food exposure biomarkers such as those retrieved from the U.S. Department of Agriculture database of phytochemicals commonly found in foods and plants. Our analysis suggests that 4,135 phytochemicals were reported in blood-related papers. Such compounds could be used as inclusion lists for untargeted metabolomics investigations of epidemiological cohort studies to complement food frequency questionnaires. Thirdly, 7,238 blood exposome compounds have associated spectra in the MassBank of North America database (https://massbank.us), enabling their annotation in untargeted metabolomics assays. Conversely, over 30,000 published blood chemicals lack public experimental mass spectra but might be detected in blood metabolomics experiments. These compounds can be used as target structures for MS-FINDER software (RIKEN Center for Sustainable Resource Science) (Tsugawa et al. 2016)–based substructure annotation.

To explore the overall chemical diversity of blood exposome chemicals, the database utilizes the MeSH ontology. By mapping MeSH to chemical compounds from the PubChem database, 1,161 chemical ontology classes were obtained with a class size of at least 50 chemicals. For example, PubChem lists 88 compounds as annotated by the MeSH term "polychlorinated biphenyls." We used these MeSH terms as "classes" in the Blood Exposome Database for user queries. Fifty-eight polychlorobiphenyls (PCBs) were found in the Blood Exposome Database. Here, PCB-153 and Aroclor 1,254 were found to be the most reported compounds in the class of PCBs. Apart from the main entry MeSH term, synonyms can be queried (such as "polychlorinated biphenyls" or "PCBs," but not "PCB" because that is not a synonym in MeSH). The compound list, identifiers, chemical and physical properties, literature data, and coverage in different biomedical databases are provided as a single data file at bloodexposome.org. Users can query and retrieve the database content using a range of options that are highlighted in the online instructions page for the database. The Blood Exposome Database will be updated on a quarterly basis to include new entries from the PubMed and PubChem databases.

Discussion

Blood chemicals were analyzed in routine clinical assays and in metabolomics, yielding a large volume of biomedical literature. A systematic compilation of blood-related chemicals will assist in the design of future studies ranging from epidemiology to nutrition research or environmental hazard assessments. We have developed such a comprehensive blood exposome database in a novel way by use of text mining and merging query results, adopting concepts that have been used before in genomic research (Cañada et al. 2017). Our database supports ongoing efforts to combine informatics resources for exposomics research (Gabb and Blake 2016; Manrai et al. 2017; Rappaport et al. 2014), but is by far larger than existing repositories, including the suite of chemicals listed in the HMDB blood database (Wishart et al. 2018). The database can be used for a wide range of applications, such as compound identification in untargeted metabolomics, development mass spectral libraries, meta-analysis of chemicals, and risk for chronic diseases, prioritizing chemicals for toxicity evaluation or interpreting the biological implications of metabolomics studies.

Similar work on chemical text mining is pursued by MeSH (Lowe and Barnett 1994), NutriChem (Ni et al. 2017), and PolySearch (Liu et al. 2015), but is less comprehensive. Among these tools, MeSH covers a large part of the exposome database presented here but misses compounds that were deposited directly by submitters or that were present in the supplementary tables of open-access articles. However, the utility of systematic use of ontologies as demonstrated by MeSH is shown by subsidiary tools like Meta2MeSH that provide statistical significance to linking chemicals with genes, diseases, and phenotypes (Sartor et al. 2012). A similar tool based on MeSH ontologies was developed as PolySearch, which associates biological concepts with biomedical literature and various databases (Liu et al. 2015). Similarly, the NutriChem database associates phytochemicals, food, and diseases using text mining of PubMed abstracts (Ni et al. 2017). However, these approaches do not appropriately associate the presence of chemicals in blood with PubMed abstracts, and they miss annotations of compounds with literature if these chemicals were directly submitted to the PubChem database. For this reason, our approach added 15,015 compounds through PubMed abstract searches using the PubChem synonym list. We recommend these synonyms be added to the MeSH ontology.

In contrary, our approach is more comprehensive because we combined three approaches, including PubMed abstract search, PMC supplementary tables, and crowdsourced information in PubChem to annotate literature with blood and chemical names. Our approach greatly accelerates and extends manual curation efforts to create a blood exposome database. We found almost 1 million papers reporting blood chemical measurement. It is not possible to manually extract chemical details from that many papers. In comparison, HMDB only used 1,278 papers for associating HMDB entries with blood levels. It would be interesting to automatically extract levels of blood chemicals directly from supplementary tables (or public databases); however, not all reports provide such tables or deposit data in public databases.

At Mayo Clinic, medical doctors can order tests for 900 chemicals in blood using a set of targeted analytical assays, showing that many compounds have direct medical relevance (https://www.mayocliniclabs.com/test-catalog/). Over 10,000 achiral blood chemicals were reported more than 10 times in the literature, while an even larger number of compounds were reported less than 10 times, and only 3,109 compounds were frequently analyzed with more than 100 citations. We think that these figures reflect a bias in clinical and epidemiological studies toward a few metabolites when many more compounds could be relatively easily determined with modern technologies. While we cannot enumerate the number of false-positive chemical annotations, we are the first to use a very large exclusion list of 10,000 phrases to control this error rate.

We suggest that authors should directly submit compound lists from their studies to the PubChem database to improve the coverage of compound to literature mapping. Supplementary tables should be submitted in CSV formats including InChIKeys for each measured compound. While authors will continue using chemical names and abbreviations in the main text of scientific reports, it is increasingly important to limit the ambiguity for chemical synonyms by using InChIKeys (Wohlgemuth et al. 2010). We likely underestimated the report on plasma lipids because very few authors use full chemical names or even InChIKeys for detected lipids. It appears highly unlikely that the chemical community will all adhere to one reference list to name particular structures. Therefore, we strongly urge chemists to use PubChem CIDs and InChIKeys to promote standardization in the fields so they can be found in full-text searches.

Although we here give evidence for 41,474 achiral compounds measured in blood, high-resolution mass spectrometry detects many more signals of unidentified origins in blood specimens (Andra et al. 2017). Annotation of chemical structures to these signals remains difficult. MS-FINDER software (Tsugawa et al. 2016) and other *in silico* methods (Allen et al. 2014; Blaženović et al. 2017; Lai et al. 2018; Psychogios et al. 2011; Ruttkies et al. 2016) propose the most probable identifications for unknown chemicals using chemical fragmentation rules. Using *a priori* knowledge has been another successful method in annotating compounds in untargeted metabolomics surveys (Edmands et al. 2015; O'Sullivan et al. 2017). The Blood Exposome Database opens a new opportunity for *in silico* spectra prediction and correlation mining to (re)analyze acquired high-resolution mass spectrometry data sets to predict high-confidence annotation for unidentified peaks.

The Blood Exposome Database will also serve as a knowledge base to guide investigations in chronic disease risk, metabolomics, metabolic regulation, precision medicine, exposure biomarkers, and chemical biomonitoring. For environmental epidemiology, links to IARC monographs and Tox21 and NLM HazMap databases can now be used to prioritize blood chemicals to be studied in prospective human cohorts. Previously, we have used the NCBI Eutils web services to identify cancer epidemiology–related publications for pesticides (Guha et al. 2016), and we have now extended this approach using R-based text mining to also include measurements of 422 pesticides in blood.

Conclusion

Manually curated databases cannot keep up with the pace and the wealth of information that is presented in peer-reviewed publications. We here show that text mining can be efficiently used to retrieve actionable data from the public NCBI database. From almost 1 million publications. we retrieved more than 41,474 achiral chemical structures that were associated with mammalian blood, and we expanded this repository to 65,957 unique isomer structures and their salts. We argue that such a database is needed as a baseline for exposome studies that aim at finding all exposome chemicals in mass spectrometry–based untargeted assays. The database can be used in epidemiology research, cheminformatics, and related areas.

Acknowledgments

This work was funded through NIH awards U54 AI138370, U19 AG023122, and U2C ES030158. D.K.B. and O.F. conceptualized the study, performed the analysis, and wrote the manuscripts.

References

Allen F, Pon A, Wilson M, Greiner R, Wishart D. 2014. CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. Nucleic Acids Res 42(Web Server issue):W94–W99, PMID: 24895432, https://doi.org/10.1093/nar/gku436.

Ananiadou S, Pyysalo S, Tsujii J, Kell DB. 2010. Event extraction for systems biology by text mining the literature. Trends Biotechnol 28(7):381–390, PMID: 20570001, https://doi.org/10.1016/j.tibtech.2010.04.005.

Andra SS, Austin C, Patel D, Dolios G, Awawda M, Arora M. 2017. Trends in the application of high-resolution mass spectrometry for human biomonitoring: an analytical primer to studying the environmental chemical space of the human exposome. Environ Int 100:32–61, PMID: 28062070, https://doi.org/10.1016/j.envint.2016.11.026.

Barupal DK, Zhang Y, Shen T, Fan S, Roberts BS, Fitzgerald P, et al. 2019. A comprehensive plasma metabolomics dataset for a cohort of mouse knockouts within the international mouse phenotyping consortium. Metabolites 9(5):E101, PMID: 31121816, https://doi.org/10.3390/metabo9050101.

Blaženović I, Kind T, Torbašinović H, Obrenović S, Mehta SS, Tsugawa H, et al. 2017. Comprehensive comparison of in silico MS/MS fragmentation tools of the CASMI contest: Database boosting is needed to achieve 93% accuracy. J Cheminform 9(1):32, PMID: 29086039, https://doi.org/10.1186/s13321-017-0219-x.

Bray GA, Redman LM, de Jonge L, Rood J, Sutton EF, Smith SR. 2018. Plasma amino acids during 8 weeks of overfeeding: relation to diet body composition and fat cell size in the proof study. Obesity (Silver Spring) 26(2):324–331, PMID: 29280309, https://doi.org/10.1002/oby.22087.

Cañada A, Capella-Gutierrez S, Rabal O, Oyarzabal J, Valencia A, Krallinger M. 2017. LimTox: a web tool for applied text mining of adverse event and toxicity associations of compounds, drugs and genes. Nucleic Acids Res 45(W1):W484–W489, PMID: 28531339, https://doi.org/10.1093/nar/gkx462.

Chaleckis R, Murakami I, Takada J, Kondoh H, Yanagida M. 2016. Individual variability in human blood metabolites identifies age-related differences. Proc Natl Acad Sci USA 113(16):4252–4259, PMID: 27036001, https://doi.org/10.1073/pnas.1603023113.

Cohen IV, Cirulli ET, Mitchell MW, Jonsson TJ, Yu J, Shah N, et al. 2018. Acetaminophen (paracetamol) use modifies the sulfation of sex hormones. EBioMedicine 28:316–323, PMID: 29398597, https://doi.org/10.1016/j.ebiom.2018.01.033.

Coletti MH, Bleich HL. 2001. Medical subject headings used to search the biomedical literature. J Am Med Inform Assoc 8(4):317–323, PMID: 11418538, https://doi.org/10.1136/jamia.2001.0080317.

Dennis KK, Marder E, Balshaw DM, Cui Y, Lynes MA, Patti GJ, et al. 2017. Biomonitoring in the era of the exposome. Environ Health Perspect 125(4):502–510, PMID: 27385067, https://doi.org/10.1289/EHP474.

Dodd DE, Pluta LJ, Sochaski MA, Wall HG, Thomas RS. 2012. Subchronic hepatotoxicity evaluation of hydrazobenzene in Fischer 344 rats. Int J Toxicol 31(6):564–571, PMID: 23134713, https://doi.org/10.1177/1091581812465322.

Edmands WM, Ferrari P, Rothwell JA, Rinaldi S, Slimani N, Barupal DK, et al. 2015. Polyphenol metabolome in human urine and its association with intake of polyphenol-rich foods across European countries. Am J Clin Nutr 102(4):905–913, PMID: 26269369, https://doi.org/10.3945/ajcn.114.101881.

Funk C, Baumgartner W Jr, Garcia B, Roeder C, Bada M, Cohen KB, et al. 2014. Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. BMC Bioinformatics 15:59, PMID: 24571547, https://doi.org/10.1186/1471-2105-15-59.

Gabb HA, Blake C. 2016. An informatics approach to evaluating combined chemical exposures from consumer products: a case study of asthma-associated chemicals and potential endocrine disruptors. Environ Health Perspect 124(8):1155–1165, PMID: 26955064, https://doi.org/10.1289/ehp.1510529.

Gu X, Manautou JE. 2012. Molecular mechanisms underlying chemical liver injury. Expert Rev Mol Med 14:e4, PMID: 22306029, https://doi.org/10.1017/S1462399411002110.

Guha N, Guyton KZ, Loomis D, Barupal DK. 2016. Prioritizing chemicals for risk assessment using chemoinformatics: examples from the IARC monographs on pesticides. Environ Health Perspect 124(12):1823–1829, PMID: 27164621, https://doi.org/10.1289/EHP186.

Hu JR, Grams ME, Coresh J, Hwang S, Kovesdy CP, Guallar E, et al. 2019. Serum metabolites and cardiac death in patients on hemodialysis. Clin J Am Soc Nephrol 14(5):747–749, PMID: 30962187, https://doi.org/10.2215/CJN.12691018.

Jansen RS, Addie R, Merkx R, Fish A, Mahakena S, Bleijerveld OB, et al. 2015. N-lactoyl-amino acids are ubiquitous metabolites that originate from CNDP2-mediated reverse proteolysis of lactate and amino acids. Proc Natl Acad Sci USA 112(21):6601–6606, PMID: 25964343, https://doi.org/10.1073/pnas.1424638112.

Jensen LJ, Saric J, Bork P. 2006. Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet 7(2):119–129, PMID: 16418747, https://doi.org/10.1038/nrg1768.

Jimeno A, Jimenez-Ruiz E, Lee V, Gaudan S, Berlanga R, Rebholz-Schuhmann D. 2008. Assessment of disease named entity recognition on a corpus of annotated sentences. BMC Bioinformatics 9(Suppl 3):S3, PMID: 18426548, https://doi.org/10.1186/1471-2105-9-S3-S3.

Karl JP, Margolis LM, Murphy NE, Carrigan CT, Castellani JW, Madslien EH, et al. 2017. Military training elicits marked increases in plasma metabolomic signatures of energy metabolism, lipolysis, fatty acid oxidation, and ketogenesis. Physiol Rep 5(17):e13407, PMID: 28899914, https://doi.org/10.14814/phy2.13407.

Kim DH, Kwack SJ, Yoon KS, Choi JS, Lee BM. 2015. 4-hydroxynonenal: a superior oxidative biomarker compared to malondialdehyde and carbonyl content induced by carbon tetrachloride in rats. J Toxicol Environ Health Part A 78(16):1051–1062, PMID: 26252470, https://doi.org/10.1080/15287394.2015.1067505.

Koppel N, Maini Rekdal V, Balskus EP. 2017. Chemical transformation of xenobiotics by the human gut microbiota. Science 356(6344):eaag2770, PMID: 28642381, https://doi.org/10.1126/science.aag2770.

Lai Z, Tsugawa H, Wohlgemuth G, Mehta S, Mueller M, Zheng Y, et al. 2018. Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. Nat Methods 15(1):53–56, PMID: 29176591, https://doi.org/10.1038/nmeth.4512.

Legette L, Karnpracha C, Reed RL, Choi J, Bobe G, Christensen JM, et al. 2014. Human pharmacokinetics of xanthohumol, an antihyperglycemic flavonoid from hops. Mol Nutr Food Res 58(2):248–255, PMID: 24038952, https://doi.org/10.1002/mnfr.201300333.

Li XS, Wang Z, Cajka T, Buffa JA, Nemet I, Hurd AG, et al. 2018. Untargeted metabolomics identifies trimethyllysine, a TMAO-producing nutrient precursor, as a predictor of incident cardiovascular disease risk. JCI Insight 3(6):99096, PMID: 29563342, https://doi.org/10.1172/jci.insight.99096.

Liu Y, Liang Y, Wishart D. 2015. PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more. Nucleic Acids Res 43(W1):W535–W542, PMID: 25925572, https://doi.org/10.1093/nar/gkv383.

Long T, Hicks M, Yu HC, Biggs WH, Kirkness EF, Menni C, et al. 2017. Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. Nat Genet 49(4):568–578, PMID: 28263315, https://doi.org/10.1038/ng.3809.

Lowe HJ, Barnett GO. 1994. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. JAMA 271(14):1103–1108, PMID: 8151853, https://doi.org/10.1001/jama.1994.03510380059038.

Mann J. 2018. *Biomonitoring California findings on perfluoroalkyl and polyfluoroalkyl substances (PFASs). Presentation to the Scientific Guidance Panel Meeting*. Oakland, CA: Biomonitoring California.

Manrai AK, Cui Y, Bushel PR, Hall M, Karakitsios S, Mattingly CJ, et al. 2017. Informatics and data analytics to support exposome-based discovery for public health. Annu Rev Public Health 38:279–294, PMID: 28068484, https://doi.org/10.1146/annurev-publhealth-082516-012737.

Marksteiner J, Blasko I, Kemmler G, Koal T, Humpel C. 2018. Bile acid quantification of 20 plasma metabolites identifies lithocholic acid as a putative biomarker in Alzheimer's disease. Metabolomics 14(1):1, PMID: 29249916, https://doi.org/10.1007/s11306-017-1297-5.

Miller DB, Ghio AJ, Karoly ED, Bell LN, Snow SJ, Madden MC, et al. 2016. Ozone exposure increases circulating stress hormones and lipid metabolites in humans. Am J Respir Crit Care Med 193(12):1382–1391, PMID: 26745856, https://doi.org/10.1164/rccm.201508-1599OC.

Miranti EH, Stolzenberg-Solomon R, Weinstein SJ, Selhub J, Mannisto S, Taylor PR, et al. 2017. Low vitamin B12 increases risk of gastric cancer: a prospective study of one-carbon metabolism nutrients and risk of upper gastrointestinal tract cancer. Int J Cancer 141(6):1120–1129, PMID: 28568053, https://doi.org/10.1002/ijc.30809.

Ni Y, Jensen K, Kouskoumvekaki I, Panagiotou G. 2017. NutriChem 2.0: exploring the effect of plant-based foods on human health and drug efficacy. Database (Oxford) 2017, PMID: 29220436, https://doi.org/10.1093/database/bax044.

O'Sullivan JF, Morningstar JE, Yang Q, Zheng B, Gao Y, Jeanfavre S, et al. 2017. Dimethylguanidino valeric acid is a marker of liver fat and predicts diabetes. J Clin Invest 127(12):4394–4402, PMID: 29083323, https://doi.org/10.1172/JCI95995.

Price ND, Magis AT, Earls JC, Glusman G, Levy R, Lausted C, et al. 2017. A wellness study of 108 individuals using personal, dense, dynamic data clouds. Nat Biotechnol 35(8):747–756, PMID: 28714965, https://doi.org/10.1038/nbt.3870.

Psychogios N, Hau DD, Peng J, Guo AC, Mandal R, Bouatra S, et al. 2011. The human serum metabolome. PLoS One 6(2):e16957, PMID: 21359215, https://doi.org/10.1371/journal.pone.0016957.

Pyysalo S, Ohta T, Miwa M, Cho HC, Tsujii J, Ananiadou S. 2012. Event extraction across multiple levels of biological organization. Bioinformatics 28(18):i575–i581, PMID: 22962484, https://doi.org/10.1093/bioinformatics/bts407.

Rappaport SM, Barupal DK, Wishart D, Vineis P, Scalbert A. 2014. The blood exposome and its role in discovering causes of disease. Environ Health Perspect 122(8):769–774, PMID: 24659601, https://doi.org/10.1289/ehp.1308015.

Rothwell JA, Fillâtre Y, Martin JF, Lyan B, Pujos-Guillot E, Fezeu L, et al. 2014. New biomarkers of coffee consumption identified by the non-targeted metabolomic profiling of cohort study subjects. PLoS One 9(4):e93474, PMID: 24713823, https://doi.org/10.1371/journal.pone.0093474.

Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S. 2016. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. J Cheminform 8:3, PMID: 26834843, https://doi.org/10.1186/s13321-016-0115-9.

Sartor MA, Ade A, Wright Z, States D, Omenn GS, Athey B, et al. 2012. Metab2MeSH: annotating compounds with medical subject headings. Bioinformatics 28(10):1408–1410, PMID: 22492643, https://doi.org/10.1093/bioinformatics/bts156.

Sobus JR, DeWoskin RS, Tan YM, Pleil JD, Phillips MB, George BJ, et al. 2015. Uses of NHANES biomarker data for chemical risk assessment: trends, challenges, and opportunities. Environ Health Perspect 123(10):919–927, PMID: 25859901, https://doi.org/10.1289/ehp.1409177.

Tan-Ariya P, Ubalee R, Na-Bangchang K, Karbwang J. 1998. Plasma containing artemether-pyrimethamine has ex vivo blood schizonticidal activity against Plasmodium falciparum. Southeast Asian J Trop Med Public Health 29(2):213–224, PMID: 9886101.

Tsugawa H, Kind T, Nakabayashi R, Yukihira D, Tanaka W, Cajka T, et al. 2016. Hydrogen rearrangement rules: computational MS/MS fragmentation and structure elucidation using MS-FINDER software. Anal Chem 88(16):7946–7958, PMID: 27419259, https://doi.org/10.1021/acs.analchem.6b00770.

Vlaanderen JJ, Janssen NA, Hoek G, Keski-Rahkonen P, Barupal DK, Cassee FR, et al. 2017. The impact of ambient air pollution on the human blood metabolome. Environ Res 156:341–348, PMID: 28391173, https://doi.org/10.1016/j.envres.2017.03.042.

Wang X, Zhou H, Zeng S. 2012. Identification and assay of 3'-O-methyltaxifolin by UPLC-MS in rat plasma. J Chromatogr B Analyt Technol Biomed Life Sci 911:34–42, PMID: 23217303, https://doi.org/10.1016/j.jchromb.2012.09.006.

Wild CP. 2005. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol Biomarkers Prev 14(8):1847–1850, PMID: 16103423, https://doi.org/10.1158/1055-9965.EPI-05-0456.

Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vázquez-Fresno R, et al. 2018. HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res 46(D1):D608–D617, PMID: 29140435, https://doi.org/10.1093/nar/gkx1089.

Wohlgemuth G, Haldiya PK, Willighagen E, Kind T, Fiehn O. 2010. The chemical translation service–a web-based tool to improve standardization of metabolomic reports. Bioinformatics 26(20):2647–2648, PMID: 20829444, https://doi.org/10.1093/bioinformatics/btq476.

Zhao M, Wang X, He M, Qin X, Tang G, Huo Y, et al. 2017. Homocysteine and stroke risk: modifying effect of methylenetetrahydrofolate reductase C677T polymorphism and folic acid intervention. Stroke 48(5):1183–1190, PMID: 28360116, https://doi.org/10.1161/STROKEAHA.116.015324.

Zhu F, Patumcharoenpol P, Zhang C, Yang Y, Chan J, Meechai A, et al. 2013. Biomedical text mining and its applications in cancer research. J Biomed Inform 46(2):200–211, PMID: 23159498, https://doi.org/10.1016/j.jbi.2012.10.007.