Updates in Rhea: SPARQLing biochemical reaction data

Thierry Lombardot[1],[†], Anne Morgat[1],[†], Kristian B. Axelsen[1], Lucila Aimo[1],
Nevila Hyka-Nouspikel[1], Anne Niknejad[2], Alex Ignatchenko[3], Ioannis Xenarios[4],[5],
Elisabeth Coudert[1], Nicole Redaschi[1] and Alan Bridge[1],*

[1]Swiss-Prot Group, SIB Swiss Institute of Bioinformatics, CMU, 1 rue Michel-Servet, CH-1211 Geneva 4, Switzerland,
[2]Vital-IT Group, SIB Swiss Institute of Bioinformatics, Quartier Sorge, Bâtiment Génopode, CH-1015 Lausanne,
Switzerland, [3]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust
Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [4]Department of Biochemistry, University of Geneva, CH-1211
Geneva, Switzerland and [5]Center for Integrative Genomics, University of Lausanne, CH-1015 Lausanne, Switzerland

ABSTRACT

Rhea (http://www.rhea-db.org) is a comprehensive and non-redundant resource of over 11 000 expert-curated biochemical reactions that uses chemical entities from the ChEBI ontology to represent reaction participants. Originally designed as an annotation vocabulary for the UniProt Knowledgebase (UniProtKB), Rhea also provides reaction data for a range of other core knowledgebases and data repositories including ChEBI and MetaboLights. Here we describe recent developments in Rhea, focusing on a new resource description framework representation of Rhea reaction data and an SPARQL endpoint (https://sparql.rhea-db.org/sparql) that provides access to it. We demonstrate how federated queries that combine the Rhea SPARQL endpoint and other SPARQL endpoints such as that of UniProt can provide improved metabolite annotation and support integrative analyses that link the metabolome through the proteome to the transcriptome and genome. These developments will significantly boost the utility of Rhea as a means to link chemistry and biology for a more holistic understanding of biological systems and their function in health and disease.

INTRODUCTION

Rhea (http://www.rhea-db.org) is a comprehensive and non-redundant resource of expert-curated biochemical reactions that uses chemical entities from the ChEBI ontology
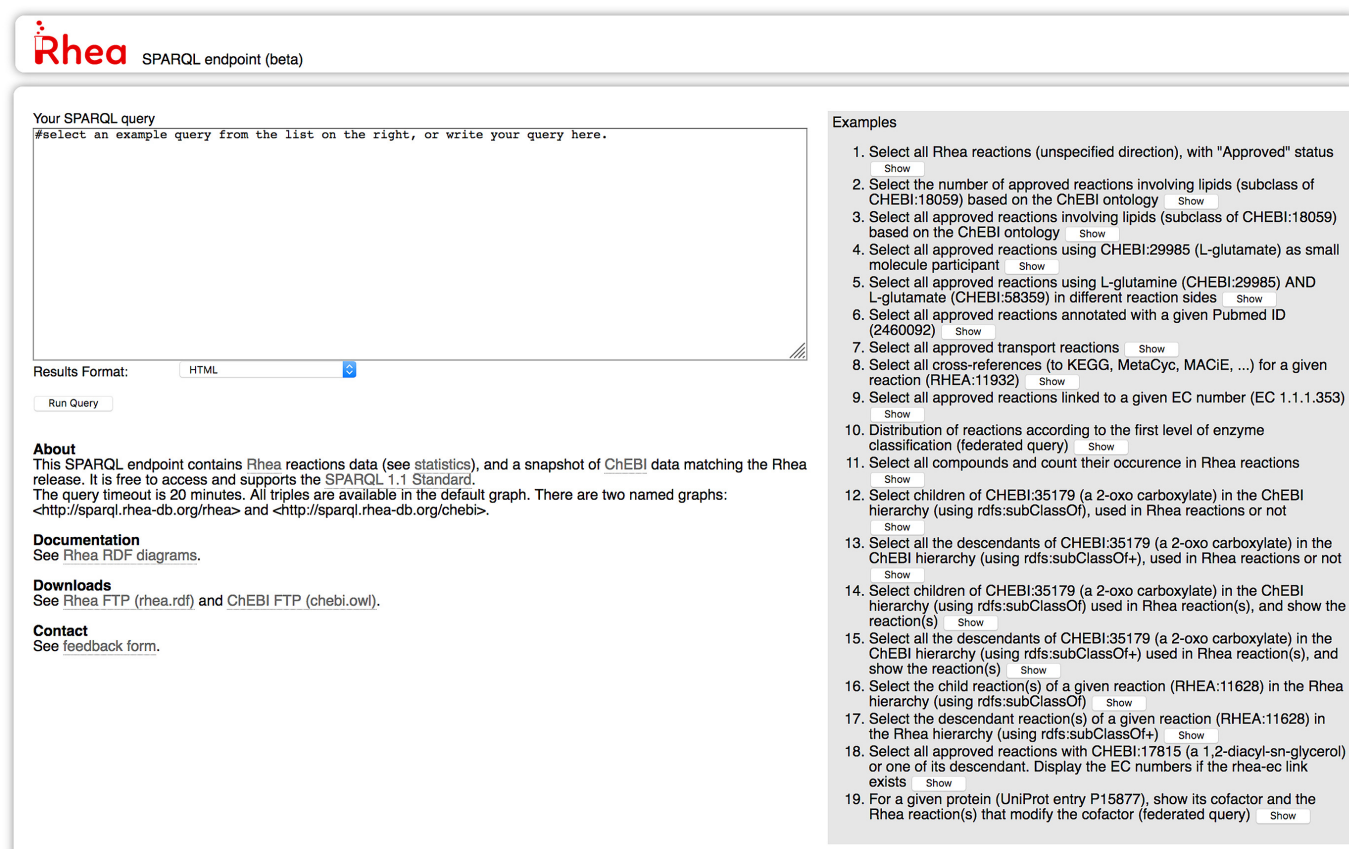
(1) to represent reaction participants. Rhea provides computationally tractable data on over 11 000 unique reactions curated from the scientific literature, covering reactions of the enzyme classification of the Nomenclature committee of the IUBMB (generally referred to as the Enzyme Classification, or 'EC') (2) as well as thousands of additional enzymatic reactions, transport reactions and spontaneously occurring reactions. Interested readers may find detailed information on Rhea reaction data in our previous publication in NAR (3).

Resources that use Rhea to describe enzymatic functions include IntEnz (4), the Enzyme Portal (5) and the Mechanism and Catalytic Site Atlas (M-CSA) (6), as well as platforms for genome scale metabolic models such as MetaNetX (7) and BiGG (8). Rhea is also currently linked to UniProtKB (9) via the enzyme classification of the IUBMB. Metabolite and metabolomics resources that use Rhea reaction data include the chemical ontology ChEBI, the SwissLipids knowledgebase for lipid biology (10) and the metabolomics repository MetaboLights (11). Rhea also links to (and is linked from) other reaction resources such as KEGG (12), MetaCyc (13) and Reactome (14), each of which also provides thousands of unique reactions.

Here, we describe recent developments in Rhea since our last publication (3), including the development of an RDF (resource description framework) representation of Rhea reaction data and a SPARQL endpoint to serve it. We also illustrate how to combine Rhea and UniProt RDF data through their respective SPARQL endpoints to generate new biological insights that combine chemical and biological knowledge from these distinct resources—a federated approach to data and knowledge mining.

*To whom correspondence should be addressed. Tel: +41 22 379 5059; Fax: +41 22 379 58 58; Email: alan.bridge@sib.swiss
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Figure 1. The Rhea SPARQL endpoint https://sparql.rhea-db.org/sparql. The Rhea SPARQL endpoint provides users with a portal to query Rhea RDF and other endpoints using the SPARQL 1.1 standards as well as a comprehensive set of sample queries and documentation on the Rhea RDF data model.

RESULTS

Rhea RDF data model and SPARQL endpoint

In order to facilitate the integration and reuse of Rhea reaction data we have developed an RDF representation of Rhea. RDF is a core semantic web technology for the World Wide Web Consortium that is well suited to applications in distributed and decentralized environments (see https://www.w3.org/RDF/ for more details).

Users can query Rhea RDF data using SPARQL (the SPARQL Protocol and RDF Query Language) at the Rhea SPARQL endpoint https://sparql.rhea-db.org/sparql (see Figure 1), which supports a range of complex and federated queries that merge data from other SPARQL endpoints. We provide a detailed description of the Rhea data model at our website https://www.rhea-db.org/rhea_rdf_documentation.pdf and invite interested readers to consult the documentation there. The Rhea SPARQL endpoint uses Virtuoso software (https://virtuoso.openlinksw.com/) and is hosted at the Vital-IT Center for high-performance computing (https://www.vital-it.ch/) of the SIB Swiss Institute of Bioinformatics. Rhea RDF data is also available to download at ftp://ftp.ebi.ac.uk/pub/databases/rhea/rdf/ serialized as RDF/XML.

Below we provide a small number of sample federated queries that illustrate how Rhea RDF data can be combined with UniProt RDF data (at https://sparql.uniprot.org/) to generate new biological insights that are not possible using either resource alone. Each of these queries utilizes a common mapping to enzyme classes of the IUBMB to link the two resources. The Rhea SPARQL endpoint provides many more sample queries designed to help new users familiarize themselves with the Rhea RDF data model and applications.

Sample Rhea SPARQL Query 1. Generate a reaction network for a specified microorganism of interest. The derivation of a list of candidate metabolic functions—in the form of a network of enzymes and reactions—is one of the first steps in the construction of draft genome scale metabolic models, popular tools to simulate and study metabolic systems (15). Such draft networks would normally be the subject of further iterative improvements and curation, including compartmentalization and the addition of biomass and hypothetical reactions necessary for the model to function.

This query demonstrates the use of Rhea to construct a network of enzymes and reactions for a specific organism of interest (in this case, *Escherichia coli* strain K12), returning a list of UniProtKB proteins and the Rhea reactions they catalyze.

Query 1.

```
PREFIX rh:<http://rdf.rhea-db.org/>

PREFIX ec:<http://purl.uniprot.org/enzyme/>

PREFIX up:<http://purl.uniprot.org/core/>

PREFIX taxon:<http://purl.uniprot.org/taxonomy/>

PREFIX keywords:<http://purl.uniprot.org/keywords/>

SELECT ?protein ?ecNumber ?reaction where {

SERVICE <https://sparql.uniprot.org/sparql/> {

        ?protein a up:Protein ;

        up:reviewed ?reviewed ;

        up:organism taxon:83333 ;

        up:classifiedWith keywords:1185 .

        {?protein up:enzyme ?ecNumber}

        UNION {?protein up:domain/up:enzyme ?ecNumber}

        UNION {?protein up:component/up:enzyme ?ecNumber} .

}

?reaction rdfs:subClassOf rh:Reaction .

?reaction rh:status rh:Approved .

?reaction rh:ec ?ecNumber .

}
```

Query 1 result. The query returns a network of ∼1600 protein-reaction links for *E. coli.* It could be easily adapted to generate a similar draft genome scale metabolic network model for any organism with complete proteome data in UniProtKB.

Sample Rhea SPARQL Query 2. Link human genes, transcripts and proteins to relevant metabolites. Integrated analyses that combine metabolomics and other types of 'omics data can advance our mechanistic understanding of disease, improve biomarker discovery and support the development personalized medicine programs (16–21).

This query demonstrates the use of Rhea to integrate knowledge of the metabolome, proteome, transcriptome and genome; it returns a list of identifiers for metabolites (ChEBI) mapped to the relevant gene and transcript (Ensembl) and protein sequences (UniProtKB/Swiss-Prot) of the enzymes that metabolize them in *Homo sapiens*. This federated query provides functionality similar to that of dedicated ID mapping tools such as MetaBridge (22).

Query 2.

```
PREFIX rh:<http://rdf.rhea-db.org/>

PREFIX taxon:<http://purl.uniprot.org/taxonomy/>

PREFIX ec:<http://purl.uniprot.org/enzyme/>

PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>

PREFIX up:<http://purl.uniprot.org/core/>

PREFIX keywords:<http://purl.uniprot.org/keywords/>

SELECT ?reaction ?chebi ?ecNumber ?protein

        ?ensemblTranscript ?ensemblGene WHERE {

SERVICE <https://sparql.uniprot.org/sparql/> {

        ?protein a up:Protein .

        ?protein up:reviewed true .

        ?protein up:organism taxon:9606 .

        ?protein up:classifiedWith keywords:1185 .

        ?protein rdfs:seeAlso ?ensemblTranscript .

        ?ensemblTranscript a up:Transcript_Resource .

        ?ensemblTranscript up:database

        <http://purl.uniprot.org/database/Ensembl> .

        ?ensemblTranscript up:transcribedFrom ?ensemblGene .

        {?protein up:enzyme ?ecNumber .}

        UNION {?protein up:domain/up:enzyme ?ecNumber .}

        UNION {?protein up:component/up:enzyme ?ecNumber .}

}

?reaction rdfs:subClassOf rh:Reaction .

?reaction rh:status rh:Approved .

?reaction rh:ec ?ecNumber .

?reaction rh:side ?reactionSide .

?reactionSide rh:contains ?participant .

?participant rh:compound ?compound .

?compound rh:chebi ?chebi .

}
```

Query 2 result. The query currently provides ∼40 000 links between metabolites (ChEBI) through their reactions to human enzymes (UniProtKB), transcripts and genes (Ensembl). Many of the metabolites identified by this query are actually chemical classes, rather than unique chemical structures; this SPARQL query could be extended to include members of these classes too if desired, thereby generating a mapping of genes, transcripts and proteins to 'plausible' metabolites (according to their chemical classification by ChEBI). We provide a further example of how to leverage the ChEBI classification in the next query.

Sample Rhea SPARQL Query 3. Identify putative enzymes for a specific metabolite. Metabolite databases such as LIPID MAPS (23), HMDB (24) and SwissLipids (10) in-

clude a large number of metabolites for which no enzyme is currently known. Chemical classifications and classifiers (25) provide a means to improve the annotation of these uncharacterized metabolites, in much the same way that protein classifications and classifiers (typically based on homology relations) can improve the annotation of uncharacterized proteins (26).

This query demonstrates how to combine the ChEBI classification with data from Rhea and UniProtKB in order to identify candidate enzymes for a specific metabolite of interest. The metabolite in question is Δ^1,Δ^7-dafachronic acid (CHEBI:83137), a potent ligand for DAF-12 which regulates aging in *Caenorhabditis elegans* (27). Δ^1,Δ^7-dafachronic acid does not feature in any Rhea reaction and is not linked to any known enzyme. The query uses the ChEBI parent/child ontology relations to retrieve all parent ChEBI classes for Δ^1,Δ^7-dafachronic acid, tracing back to the root of the ChEBI ontology and then searches for the candidate enzymes and reactions for these parent classes. This query effectively extends the annotation of experimentally characterized metabolite classes in UniProtKB/Swiss-Prot to currently unannotated members of the same chemical classes.

Query 3.

```
PREFIX rh:<http://rdf.rhea-db.org/>

PREFIX ch:<http://purl.obolibrary.org/obo/>

PREFIX ec:<http://purl.uniprot.org/enzyme/>

PREFIX up:<http://purl.uniprot.org/core/>

PREFIX taxon:<http://purl.uniprot.org/taxonomy/>

PREFIX keywords:<http://purl.uniprot.org/keywords/>

SELECT ?chebi ?ecNumber (count(?protein) as

      ?proteinCount) ?reaction ?equation WHERE {

      ?reaction rdfs:subClassOf rh:Reaction .

      ?reaction rh:status rh:Approved .

      ?reaction rh:equation ?equation .

      ?reaction rh:ec ?ecNumber .

      ?reaction rh:side ?reactionSide .

      ?reactionSide rh:contains ?participant .

      ?participant rh:compound ?compound .

      ?compound rh:chebi ?chebi .

      ch:CHEBI_83137 rdfs:subClassOf+ ?chebi.

SERVICE <https://sparql.uniprot.org/sparql/> {

      ?protein a up:Protein .

      ?protein up:reviewed ?reviewed .

      {?protein up:enzyme ?ecNumber}

      UNION {?protein up:domain/up:enzyme ?ecNumber}

      UNION {?protein up:component/up:enzyme ?ecNumber} .

      }

}
```

Query 3 result. The query proposes a total of 16 candidate enzyme classes (as defined by the enzyme classification of the IUBMB) for Δ^1,Δ^7-dafachronic acid. These sixteen enzyme classes act on those chemical classes of which Δ^1,Δ^7-dafachronic acid is a member, such as the 3-oxo-Δ^1 steroids (CHEBI:20156) and other parent classes of increasing generality such as the 3-oxo steroids (CHEBI:47788), and its parent classes the steroids (CHEBI:35341) and ketones (CHEBI:17087). Each of these enzyme classes are potential candidates to metabolize Δ^1,Δ^7-dafachronic acid. Known members of the most specific of these 16 enzyme classes, EC 1.3.99.4, which catalyzes the interconversion of 3-oxo-Δ^1 steroids and 3-oxo steroids, are currently restricted to bacteria. Members of other enzyme classes of lower specificity such as EC 1.1.1.184 (encoded by *dhrs-4* described in UniProtKB:G5EGA6), and EC 1.1.1.1 (encoded by *sodh-1*, *sodh-2*, *H24K24* and *dhs-3*, described in UniProtKB:Q17334, UniProtKB:O45687, UniProtKB:Q17335 and UniProtKB:A5JYX5) are found in *C. elegans*.

Other modes of Rhea access

In addition to now providing the Rhea SPARQL endpoint we also continue to maintain all the modes of access (interactive searches, programmatic access and data downloads) and data formats described in our previous publication (3) at www.rhea-db.org.

Rhea content

Rhea has continued to grow significantly since our last report through the expert curation of new chemical entities in ChEBI and reactions from peer-reviewed literature (see http://www.rhea-db.org/statistics for details). Rhea currently (release 96 of 13 July 2018) describes 11 173 unique reactions involving 9916 unique reaction participants and cites 12 611 unique literature references (PubMed identifiers). This represents an increase of ~1900 unique reactions, 1800 unique reaction participants and 3700 literature references since our last publication (3) (which described release 75 of 30 July 2016).

DISCUSSION

We have shown how federated SPARQL queries that combine Rhea reaction data with that from other SPARQL endpoints such as that of UniProt can facilitate a range of data integration and data mining tasks. These include the generation of draft genome-scale metabolic reaction networks and the identification of candidate enzymes, which are common use cases in systems biology applications such as metabolic modeling and engineering, and the integration of genome, transcriptome, proteome and metabolome data, which is of broad utility, including in the domain of personalized health and medicine.

The federated queries we describe currently exploit the mapping between Rhea reactions and the IUBMB enzyme classification to link Rhea and UniProtKB. In the near future UniProt will incorporate Rhea as an annotation vocabulary for enzymes in UniProtKB, and UniProt curators

will directly link Rhea reactions to UniProtKB/Swiss-Prot records as part of their normal curation workflow. This will significantly increase the coverage and specificity of enzyme annotation in UniProtKB, enhancing the utility of UniProtKB and Rhea for 'omics data integration and powering new search and analysis capabilities that combine protein sequence and function with chemical structure data.

REFERENCES

1. Hastings,J., Owen,G., Dekker,A., Ennis,M., Kale,N., Muthukrishnan,V., Turner,S., Swainston,N., Mendes,P. and Steinbeck,C. (2016) ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res.*, **44**, D1214–D1219.
2. McDonald,A.G., Boyce,S. and Tipton,K.F. (2009) ExplorEnz: the primary source of the IUBMB enzyme list. *Nucleic Acids Res.*, **37**, D593–D597.
3. Morgat,A., Lombardot,T., Axelsen,K.B., Aimo,L., Niknejad,A., Hyka-Nouspikel,N., Coudert,E., Pozzato,M., Pagni,M., Moretti,S. *et al.* (2017) Updates in Rhea - an expert curated resource of biochemical reactions. *Nucleic Acids Res.*, **45**, D415–D418.
4. Fleischmann,A., Darsow,M., Degtyarenko,K., Fleischmann,W., Boyce,S., Axelsen,K.B., Bairoch,A., Schomburg,D., Tipton,K.F. and Apweiler,R. (2004) IntEnz, the integrated relational enzyme database. *Nucleic Acids Res.*, **32**, D434–D437.
5. Pundir,S., Onwubiko,J., Zaru,R., Rosanoff,S., Antunes,R., Bingley,M., Watkins,X., O'Donovan,C. and Martin,M.J. (2017) An update on the Enzyme Portal: an integrative approach for exploring enzyme knowledge. *Protein Eng. Des. Sel.*, **30**, 245–251.
6. Ribeiro,A.J.M., Holliday,G.L., Furnham,N., Tyzack,J.D., Ferris,K. and Thornton,J.M. (2018) Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Res.*, **46**, D618–D623.
7. Moretti,S., Martin,O., Van Du Tran,T., Bridge,A., Morgat,A. and Pagni,M. (2016) MetaNetX/MNXref–reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic Acids Res.*, **44**, D523–D526.
8. King,Z.A., Lu,J., Drager,A., Miller,P., Federowicz,S., Lerman,J.A., Ebrahim,A., Palsson,B.O. and Lewis,N.E. (2016) BiGG models: a platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res.*, **44**, D515–D522.
9. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
10. Aimo,L., Liechti,R., Hyka-Nouspikel,N., Niknejad,A., Gleizes,A., Gotz,L., Kuznetsov,D., David,F.P., van der Goot,F.G., Riezman,H. *et al.* (2015) The SwissLipids knowledgebase for lipid biology. *Bioinformatics*, **31**, 2860–2866.
11. Kale,N.S., Haug,K., Conesa,P., Jayseelan,K., Moreno,P., Rocca-Serra,P., Nainala,V.C., Spicer,R.A., Williams,M., Li,X. *et al.* (2016) MetaboLights: an open-access database repository for metabolomics data. *Curr. Protoc. Bioinformatics*, **53**, doi:10.1002/0471250953.bi1413s53.
12. Kanehisa,M., Furumichi,M., Tanabe,M., Sato,Y. and Morishima,K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.
13. Caspi,R., Billington,R., Fulcher,C.A., Keseler,I.M., Kothari,A., Krummenacker,M., Latendresse,M., Midford,P.E., Ong,Q., Ong,W.K. *et al.* (2018) The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res.*, **46**, D633–D639.
14. Fabregat,A., Jupe,S., Matthews,L., Sidiropoulos,K., Gillespie,M., Garapati,P., Haw,R., Jassal,B., Korninger,F., May,B. *et al.* (2018) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **46**, D649–D655.
15. Thiele,I. and Palsson,B.O. (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.*, **5**, 93–121.
16. Beger,R.D., Dunn,W., Schmidt,M.A., Gross,S.S., Kirwan,J.A., Cascante,M., Brennan,L., Wishart,D.S., Oresic,M., Hankemeier,T. *et al.* (2016) Metabolomics enables precision medicine: "A White Paper, Community Perspective". *Metabolomics*, **12**, 149.
17. Suhre,K., Raffler,J. and Kastenmuller,G. (2016) Biochemical insights from population studies with genetics and metabolomics. *Arch. Biochem. Biophys.*, **589**, 168–176.
18. Karczewski,K.J. and Snyder,M.P. (2018) Integrative omics for health and disease. *Nat. Rev. Genet.*, **19**, 299–310.
19. Guijas,C., Montenegro-Burke,J.R., Warth,B., Spilker,M.E. and Siuzdak,G. (2018) Metabolomics activity screening for identifying metabolites that modulate phenotype. *Nat. Biotechnol.*, **36**, 316–320.
20. Kapono,C.A., Morton,J.T., Bouslimani,A., Melnik,A.V., Orlinsky,K., Knaan,T.L., Garg,N., Vazquez-Baeza,Y., Protsyuk,I., Janssen,S. *et al.* (2018) Creating a 3D microbial and chemical snapshot of a human habitat. *Sci. Rep.*, **8**, 3669.
21. Wigger,L., Cruciani-Guglielmacci,C., Nicolas,A., Denom,J., Fernandez,N., Fumeron,F., Marques-Vidal,P., Ktorza,A., Kramer,W., Schulte,A. *et al.* (2017) Plasma dihydroceramides are diabetes susceptibility biomarker candidates in mice and humans. *Cell Rep.*, **18**, 2269–2279.
22. Hinshaw,S.J., Lee,A.H.Y., Gill,E.E. and Hancock,R.E.W. (2018) MetaBridge: enabling network-based integrative analysis via direct protein interactors of metabolites. *Bioinformatics.* **34**, 3225–3227.
23. Fahy,E., Subramaniam,S., Murphy,R.C., Nishijima,M., Raetz,C.R., Shimizu,T., Spener,F., van Meer,G., Wakelam,M.J. and Dennis,E.A. (2009) Update of the LIPID MAPS comprehensive classification system for lipids. *J. Lipid Res.*, **50**, S9–S14.
24. Wishart,D.S., Feunang,Y.D., Marcu,A., Guo,A.C., Liang,K., Vazquez-Fresno,R., Sajed,T., Johnson,D., Li,C., Karu,N. *et al.* (2018) HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.*, **46**, D608–D617.
25. Djoumbou Feunang,Y., Eisner,R., Knox,C., Chepelev,L., Hastings,J., Owen,G., Fahy,E., Steinbeck,C., Subramanian,S., Bolton,E. *et al.* (2016) ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminform.*, **8**, 61
26. Pedruzzi,I., Rivoire,C., Auchincloss,A.H., Coudert,E., Keller,G., de Castro,E., Baratin,D., Cuche,B.A., Bougueleret,L., Poux,S. *et al.* (2015) HAMAP in 2015: updates to the protein family classification and annotation system. *Nucleic Acids Res.*, **43**, D1064–D1070.
27. Saini,R., Boland,S., Kataeva,O., Schmidt,A.W., Kurzchalia,T.V. and Knolker,H.J. (2012) Stereoselective synthesis and hormonal activity of novel dafachronic acids and naturally occurring steroids isolated from corals. *Org. Biomol. Chem.*, **10**, 4159–4163.