# Boost that metabolomic confidence

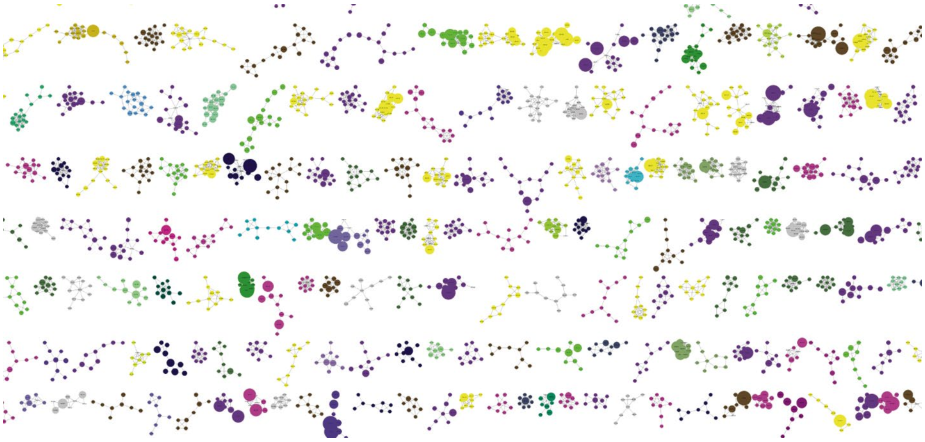
Change is a constant in the burgeoning field of metabolomics. That includes data analysis tools and repositories.

Vivien Marx

Mutations can change genomes. Regulation shifts gene expression. Given that it's the biochemical output of metabolic processes, "an individual's metabolome is always changing," says Arthur Castle, who directs programs in metabolomics at the National Institute of Health (NIH) and is program officer for metabolomics at the NIH Common Fund. Metabolomics can, for example, capture how circadian rhythms shift the types and levels of metabolites an organism produces, reveal transitions between health and disease in people or model organisms, or uncover how drought affects crop yield or how biochemical communication between microbes and soil shape an ecosystem. Computational methods help find patterns in metabolomic data troves, but it's challenging terrain. Cellular metabolism involves a few thousand endogenous metabolites, says Castle. Add to that the output of microbes metabolizing in and on us, our ingested medications and over-the-counter substances, plants and chemicals we are exposed to. "There certainly could be tens of thousands, if not millions, of potential metabolites that you can detect if you had unlimited sensitivity," he says. "The theoretical chemical space is certainly well into the millions." With human blood, urine or feces, researchers can't be sure how many metabolites to expect, says Justin van der Hooft, a metabolomics researcher at Wageningen University, but "what we in practice see is probably something from 1,200, 1,300 to maybe 2,000, in a very good setup and if you have the time."

Labs want to identify metabolites with confidence, but metabolites don't exactly wear name tags. Meaningful spectral peaks have to be detected, associated metabolites have to be named and, eventually, the pathways they play a role in have to be determined. To validate, labs check online repositories, but they don't always find high quality there, says van der Hooft.

In his Northwestern University synthetic biology lab, Michael Jewett and team want to pick pathways or know which fluxes to enable for desired molecular transformations. "Metabolomics is hard," he says. The literature does not yet have all the answers. The many data analysis tools can feel overwhelming, says van der Hooft. The data contain artifacts and unknowns. Standards and sharing practices are still evolving. Metabolomics researchers from a number of European research institutions point to a lack of depth and breadth in metabolite databases[1]. Computational tools, data and minimum information standards exist, but navigation can be tough for the average researcher. That's why labs and funders want to keep growing metabolomics and make it more navigable.



Networks of metabolites from multiple human gut samples can be computationally grouped into families of chemical classes. Each color represents a class. The process applies a natural language processing technique called topic modeling. Credit: van der Hooft lab, Wageningen Univ.

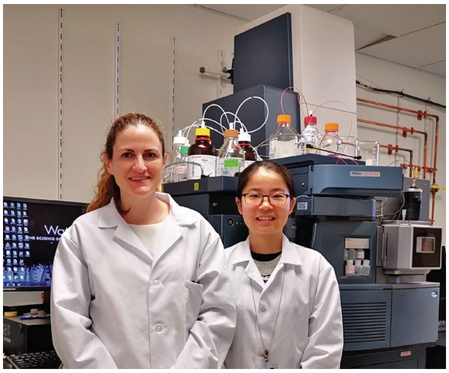
## Joining up

Traditionally, data from different instruments "have been considered incomparable," which hampers the burgeoning field's expansion, says Masanori Arita, a RIKEN metabolomics researcher. Just as the International Nucleotide Sequence Database Collaboration enabled genomic sequence data sharing by mirroring GenBank, the DNA DataBank of Japan and the European Nucleotide Archive, metabolomics resources are beginning to join up to ease computational workflows, says Claire O'Donovan, who leads metabolomics at the European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL-EBI) and who runs the database MetaboLights. MetabolomeXchange is a portal for accessing datasets; there are repositories such as Metabolomics Workbench, Metabolomics Repository Bordeaux — a plant resource with NMR data — and MetaboLights.

MetaboLights holds metabolomics experiments and associated information such as mass spectra. Its 'cousins' include MassBank, the mass spectral database run under the auspices of the Mass Spectrometry Society of Japan; Metabolomics Workbench in the U.S.; and the Human Metabolome Database in Canada. Early in 2020, a Japanese repository called MetaboBank will go online and exchange metadata with MetaboLights, says Arita. It's ambitious, says O'Donovan. But as with genomics and proteomics, resources must be amply fed with studies and data to be powerful. At the annual Metabolomics Society meeting in July 2020, says O'Donovan, repository managers from North America, Japan and Europe will discuss how to foster data exchange and ask "what does the community want to see in the repositories going forward?"

Around 80% of metabolomics studies in MetaboLights involve mass spectrometry and an estimated 20% use nuclear magnetic resonance (NMR) spectroscopy, says O'Donovan. Mass spec is the metabolomics workhorse due to its sensitivity, says Castle. People had been moving away from NMR, but now use it to refine their analysis.

For single-compound identification, NMR can help when mass spec is insufficient,



There is no one-size-fits-all when it comes to data analysis in metabolomics, says Caroline Johnson (left). She and postdoctoral fellow Yuping Cai find plenty of unknowns — new metabolites — in their data. Credit: P. Zhang, Yale School of Public Health

says Christoph Steinbeck from Friedrich Schiller University Jena, who founded MetaboLights as an EBI researcher.

## Cautious confidence

Labs want to be sure an identified metabolite is not a spurious signal. "Even collecting blood, for example, into a different type of tube will massively change the metabolome," says Caroline Johnson from Yale School of Public Health. The sample can degrade during extraction, it can fragment in the mass spec, and solvent and plastic tubing can add background noise, says Yuping Cai, postdoctoral fellow in Johnson's lab. Qualitative and quantitative confidence is almost always limited when labs 'make calls' and identify hundreds of metabolites in a sample, says Castle. They might have mass, mobility, chromatography pattern and mass fragmentation data. But the sample can contain isomers, molecules with the same mass but different atomic arrangements. A measured metabolite might be well-known to a team who has validated it with an authenticated standard, which renders their confidence "very, very high," he says. But stereochemistry can dent that confidence. There are enantiomers — molecular mirror images — to contend with. "Ideally you'd like to say I'm 99% confident this is alanine," he says. But it might be L- or D-alanine. With lipid metabolites, labs are just "scratching the surface" of side-chain diversity and lipid combinations. Such quandaries need further analysis, which computational methods help explore.

## Find a tool, use a tool

'Peak-picking' to assign molecular identities to chromatographic peaks was long done manually, says van der Hooft. Five files took around a week. To do so reproducibly with dozens or hundreds of files takes computational tools. "There's no lack of tools anymore," he says, but labs need to select wisely.

Many tools can be downloaded or used in online portals and serve different purposes, such as identification of metabolites from raw mass spectra or pathway analysis. XCMS[2] is "a killer app," says Johnson about software from the lab of her postdoctoral advisor Gary Siuzdak at Scripps Research Institute. There's a standalone R-based version and a browser-based version for extracting metabolic features from raw mass spec data and for statistical analysis. Johnson, who added new ideas to the online version, provided experimental data for software testing and identified user needs, says that XCMS lets users stack metabolomic data onto metabolic pathways and integrate results with genomic and proteomic data. Her lab uses the browser-based version but switches to the R-based tool to stay in R for correcting batch effects and signal drift. The tool's "instant putative metabolite and metabolic pathway identification is a win-win situation for anyone doing untargeted metabolomics analysis," says Johnson. In addition to XCMS, she and Cai say labs can consider MZmine, MetAlign, MAVEN and MS-DIAL. To ensure reproducibility, Cai recommends investigators keep track of the software version they use and avoid switching versions mid-analysis.

"There's no lack of tools anymore," says Justin van der Hooft about computational metabolomics.

Credit: E. Burrillon

SIRIUS 4[3], for example, is in its fourth iteration, with, according to its developers, improved identification rates and accelerated processing. It can be used standalone or as a web-based service for identifying a metabolite using isotope pattern and fragmentation analysis. SIRIUS includes CSI-FingerID, a web-service that searches molecular structure databases.

Several tools let labs assign mass spec fragmentation spectra to chemical structures and match them to known data in databases, such as MassBank or The Human Metabolome Database. But mass alone can be insufficient for metabolite identification, says van der Hooft. Liquid chromatography helps but might not always resolve isomers. Not all metabolites have reference spectra, he says. He and colleagues at the University of Glasgow built MS2LDA, which uses a natural language processing technique, topic modeling, to pull out mass fragment motifs and group co-occurring fragment peaks related to metabolite substructures, such as flavonoid cores. In his view, substructure analysis can help address the bottleneck in metabolite annotation and identification. The team built MS2LDA into a processing workflow called MolNetEnhancer[4] with annotation and chemical classification tools. To let labs try it or use it regularly, they've added it to Global Natural Products Social Molecular Networking (GNPS), a cloud-based ecosystem based at the Dorrestein lab at the University of California, San Diego. Tool documentation, says van der Hooft, makes it easier for users to gauge a tool's best uses. Among other developments, he points to the community effort ReDU, the Reanalysis of Data User interface for finding and reusing data in GNPS. The organizers note it's desirable to more quickly query repository-scale public data and avoid issues of proprietary file types, with, as the authors point out, "inconsistent metadata formats." ReDU[5] offers tandem mass spec data that are "uniformly formatted" and metadata backed by ontology terms. As van der Hooft explains, users can, for example, select data such as 'urine samples from men under 40'. Another new development, he says, is Mass Spectrometry Search Tool[6] (MASST), a search tool for querying public small-molecule tandem mass spec data across metabolomics repositories.

O'Donovan sees "a big push" in the community toward cloud-based analysis environments. She, Steinbeck and others across Europe co-developed the PhenoMeNal[7] Gateway, which stands for Phenome and Metabolome Analysis. It lets researchers launch analytical workflows on a cloud of their choice, says Steinbeck. As he waits for the next round of funding, he reaches out to potential commercial users, and PhenoMeNal is now integrated into the European Open Science Cloud. ELIXIR, which includes research organizations in 23 European countries that offer data repositories or data analysis tools, is working on 'use cases' to help labs see which instances are best addressed by which tools and in which order those are best used, says O'Donovan. One ELIXIR study is looking at metabolite identification tools and the other is for 'fluxomics', a metabolomics subfield in which labs chronicle small-molecule-based changes in networks in vivo, such as changes as a person eats. The EBI held a hackathon to explore how metabolomics tools can run in Galaxy, a cloud-based genomics analysis environment. "In Galaxy you can set up workflows, which is ideal," she says, given how many processing steps metabolomics data analysis entails.

Metabolomic analysis yields a bounty of as-of-yet unknown metabolites. Here, some 'unknowns' from the Johnson lab at Yale School of Public Health, the Steinbeck lab at Friedrich Schiller University Jena and the van der Hooft lab at Wageningen University. Credit: N. Carol Photogr./Moment/Getty, A. Vartanian, E. Dewalt/Springer Nature

## Unknowns

Over the last decade, says Johnson, databases have evolved such that it no longer takes months to identify the list of metabolites that matter for a biological question. In her lab, they used to try plenty: synthesize new metabolites to confirm identification; apply triple quadruple mass spec to confirm abundance; "and I rarely see either of these being performed alongside untargeted metabolomics now." The expansion of tools and databases leads to greater use by researchers beyond analytical chemists. Yet some important but unidentifiable metabolites aren't curated in databases, says Johnson. They are pursued less vigorously now that it's generally easier to identify a larger number of metabolites. "It's a pity!" she says.

Metabolomic analysis yields a bounty of unknown metabolites. "I think it's becoming more and more common," says Castle, partially due to microbiome findings. Food metabolites may be well-studied, but the same cannot be said for microbial metabolites in soil or the body, says van der Hooft. Unknowns that a lab encounters "should not simply be dismissed with an assumption," says Johnson. In a recent analysis of colorectal cancer tissues, her lab found unknown metabolites "highly significant in tissues coming from women." These might be microbial metabolites, a hunch the team is now investigating. Databases of mass spec fragmentation are useful resources for identifying unknown metabolites. "MassBank is the biggest of these," says Castle. And it offers tools, too. Also helpful: knowledge bases such as those part of The Human Metabolome Database, which offers descriptions of where metabolites are found and their functions. The NIH resource PubChem also contains metabolite information.

## Metabolome business

To accelerate metabolomic pathway analysis, Jewett sets up collaborations, such as with Oak Ridge National Laboratory for analytical measurements and with companies such as Lockheed Martin for computational analysis and deep learning. For synthetic biology projects, thousands of metabolic reactions might be relevant, says Jewett. "If we could study more pathways we could eventually push designs to be more efficient," he says. As metabolomics has grown, so have commercial analysis-as-a-service offerings, such as those from Biocrates Life Sciences, Nightingale Health and Metabolon. What helps the field grow, says Metabolon's research and development director Annie Evans, is the way it can "unpack biology" with systems understanding. The company does metabolomic analyses for commercial and academic labs of all sizes. The field will keep growing, she says, and microbiome research is a big contributor. Discovery of the microbiome is akin to discovering "a new organ" in the human body.

Metabolon has long curated a proprietary library of metabolites and pathways for its analysis of human samples, microbiome or plant tissue, some examples of which are analyzing tryptophan metabolism in mouse models of human disease or exploring ancient metabolism through dental scrapings from mummies, says Evans. The company has its own data mining and software tools because "when we started there was absolutely no software available," she says. In a human blood sample, her teams can identify between 2,000 and 2,500 metabolites. She is confident in the ability of the proprietary pipeline to identify molecules and name the entities against a backdrop of noisy mass spec data. A single compound produces many signatures, and mishandling data can lead to misidentification. "You have to have your eyes open for those novel molecules that you didn't know were there," she says. Something might look like a new metabolite, "but it's not, it's just an adduct of phenylalanine."

Annie Evans hopes to see more tools to address the question "How do you make sense of what this means for you biologically?"

Credit: Metabolon

In the Metabolomics Quality Assurance and Quality Control Consortium[8] (mQACC), Evans and O'Donovan co-chair a working group on experimental processes. In the past, scientists drawing on Metabolon's services were not allowed to submit the full raw mass spectra to a public repository. Evans sees the risk that "people will take our data and make poor conclusions with it and actually end up hurting the industry." But, she says, "this is an evolving space." The company's latest policies permit investigators to, for example, include these data in papers or grant submissions. "We can provide it in certain circumstances," she says.

O'Donovan is glad to work with Evans. First interactions with Metabolon were "'no, we don't share,'" she says. "We've moved on a lot since then." The company recently agreed to use the open source mass spec data format called mzML. "If people submit to us in mzML, that means the whole of the metabolomics community can exploit that data," says O'Donovan. The format does not capture all underlying raw data; that would require using proprietary software, which not every lab has. Getting data from Metabolon and other organizations in this open format is "a major improvement," says O'Donovan. In NMR, the open data standard nmrML also has a chance, says Steinbeck, especially when the community pushes for its use. In a given paper, labs want to see the raw data available and provided in a shareable format, he says.

Ongoing discussions with Metabolon and other companies are shaping standards discussions in the research community, says O'Donovan. The field changes fast, and so must the requirements of what repositories should capture. For now, there is no gold standard for submission: "only the community can tell us that," she says.

David Foster, vice president of data and informatics at Metabolon, says that in any given metabolomic study thousands of metabolites might be changing, so he and his team have "thousands of variables to parse through" and map to their biological meaning. The company works with scientists at many project stages, from early pilot phase to scaled-up projects, says Greg Michelotti, scientific director of biology at Metabolon. He and his colleagues see metabolomics becoming a "central modality" for deconvoluting other -omics streams.

Metabolomic data can expand insights from massive genomics data for biological systems because of how it can also identify non-genomic influences on biology from the environment or the microbiome. The presence or absence of a single nucleotide polymorphism can alter a biological system, says Foster, and metabolomics captures other, more subtle, changes. In a multi-omics context, metabolomics is "the other bookend" to genomics, he says, given that it's "closest to the action" in a biological system.

Foster sees a need for computational tools that explore the dense network of relationships between genes, proteins and metabolites. Metabolomics needs the approach taken in genomics, with well-constructed, interoperable tools. Evans see much tool development in academia focused on compound identification. "What I really want is that next step," she says. She hopes to see more tools to address the question "how do you make sense of what this means for you biologically?" given that "we are so far from having really comprehensive annotation tools." Such tools could highlight molecules associated with, say, inflammation or liver toxicity and capture a bigger picture showing biological changes. Another need she sees is for better ways to map metabolites onto pathways and capture interplay between pathways, since "one molecule can be involved in 17 different pathways."

At Metabolon, automation is applied to process samples from large-scale studies and ready them for metabololomic analysis. Credit: Metabolon

## Remember the metadata

"It would be interesting to hear how often the data is accessed and used actually," says Johnson of existing resources such as those accessible via MetabolomeXchange. Quite often she finds data without preprocessing information, which hinders replicability of analysis. In genomics, it's mandatory to make data publicly available. "It should be the same for metabolomics," she says. It would be useful to have the data in multiple standardized formats, such as the raw mass spectra, tandem mass spec data, standards run, mzML format, and the final processed spreadsheet data matrix for further statistical analysis. "Otherwise it's pretty difficult to get anything meaningful from the data," she says.

Resources need data and metadata, says O'Donovan, such as data about the organism from which the sample is drawn. With human data, metadata might cover sex and weight; with plants, it could be species and growing conditions — it's information about the instrument with which data was generated, including temperature or types of controls. Some preliminary reports indicate widely differing practices and the omission of information on whether instruments were tested before an experiment starts. mQACC can give labs guidance on experimental and instrument setup for studies, says O'Donovan, such as whether samples should be pooled. Teaching in metabolomics is evolving, but "it's not evolving as fast as the field is," she says. This matters given that metabolomics is drawing in many new entrants with little or no chemistry background. The Metabolomics Workbench was developed as a repository for data and statistical analysis in a cloud environment. In its second phase of development, now underway, says Castle, the team is expanding the workbench for its use in large-scale studies that might involve access-controlled data. Metabolomic data is not as identifiable as genomic data, but "the metadata might actually be very identifiable," he says.

## Next tools

Among the tools needed in metabolomics, Castle hopes for better ways to assess mass spec data in terms of biology. A Common Fund project underway at the University of Michigan at Ann Arbor focuses on interrogating large datasets and linking them with other types of -omics data; others involve tools for normalization and pathway analysis at the University of Colorado Denver and tools to battle artifacts at Washington University in St. Louis.

"One of the greatest needs, of course, is identifying unknowns," says Castle, even when there are no standards. In the works is a database of 'hypothesized compounds' including simulated mass spec signatures and ion mobilities. The idea is to build "in silico reference databases" with potentially millions of compounds "that could be reasonably accurate to what a true authenticated standard is," he says. Such a resource could cover a broad chemistry spectrum, which "would really go to making metabolomic experiments much more interpretable." For their analyses, labs can start with known metabolites and use them to establish some biologically driven reactions, and then expand on those by building on anticipated reactions of these metabolites. This approach is "an expansion from the biologically known" to chemoinformatically generate millions of compounds randomly. "Some of them would make sense and some of them may be thermodynamically unlikely," he says, but it could be possible to establish a huge reference database.

At the moment, says O'Donovan, in metabolomics "we're still at the baby-step level, but there is a huge amount of goodwill." In her observation, funders "are very, very interested in making sure that metabolomics is as powerful as it can be." At the EBI, experience with genomics and proteomics data and resources deliver useful lessons for metabolomics. "It's critical to collaborate right across the world," she says, "because we can only do this together." ❒

Vivien Marx
*Technology editor for Nature Methods.*
*e-mail: v.marx@us.nature.com*

Published online: 20 December 2019
https://doi.org/10.1038/s41592-019-0694-2

References
1. Rijswijk, M. et al. *F1000Res* **6**, 1649 (2017).
2. Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R. & Siuzdak, G. *Anal. Chem.* **78**, 779–787 (2006).
3. Dührkop, K. et al. *Nat. Methods* **16**, 299–302 (2019).
4. Ernst, M. et al. *Metabolites* **9**, 144 (2019).
5. Jarmusch, A. et al. Preprint at *bioRxiv* https://doi.org/10.1101/750471 (2019).
6. Wang, M. et al. Preprint at *bioRxiv* https://doi.org/10.1101/591016 (2019).
7. Peters, K. et al. *Gigascience* **8**, giy149 (2019).
8. Beger, R. D. et al. *Metabolomics* **15**, 4 (2019).