Crowd-Sourced Chemistry: Considerations for Building a Standardized Database to Improve Omic Analyses

Jaqueline A. Picache,* Jody C. May,* and John A. McLean*

Department of Chemistry, Center for Innovative Technology, Vanderbilt Institute of Chemical Biology, Vanderbilt Institute for Integrative Biosystems Research and Education, Vanderbilt University, Nashville, Tennessee 37235, United States

ABSTRACT: Mass spectrometry (MS) is used in multiple omics disciplines to generate large collections of data. This data enables advancements in biomedical research by providing global profiles of a given system. One of the main barriers to generating these profiles is the inability to accurately annotate omics data, especially small molecules. To complement pre-existing large databases that are not quite complete, research groups devote efforts to generating personal libraries to annotate their data. Scientific progress is impeded during the generation of these personal libraries because the data contained within them is often redundant and/or incompatible with other databases. To overcome these redundancies and incompatibilities, we propose that communal, crowd-sourced databases be curated in a standardized fashion. A small number of groups have shown this model is feasible and successful. While the needs of a specific field will dictate the functionality of a communal database, we discuss some features to consider during database development. Special emphasis is made on standardization of terminology, documentation, format, reference materials, and quality assurance practices. These standardization procedures enable a field to have higher confidence in the quality of the data within a given database. We also discuss the three conceptual pillars of database design as well as how crowd-sourcing is practiced. Generating open-source databases requires front-end effort, but the result is a well curated, high quality data set that all can use. Having a resource such as this fosters collaboration and scientific advancement.

1. INTRODUCTION

Mass spectrometry serves as a foundational analytical technology in untargeted omics experiments.1 In recent decades, MS has enabled the collection of big data in biomedical research. As more data is collected and the age of big data matures, many opportunities arise to gain insightful knowledge about biomedical systems that were not previously accessible.2

While these terms are often used interchangeably, their definitions have nuanced differences as described in Table 1.6 Annotation of omics data relies heavily on matches from database queries.4 Success in the annotation process is contingent upon the quality of the database being queried as well as the amount of unique information known about the omic compound in question. A few prominent, large-scale databases include the Human Metabolome Database, PubChem, and UniProt.7−9

All three of these databases rely on crowd-sourced information. Generally, crowd-sourcing is an active solicitation of content, ideas, or services from a large community. When performed by scientific database curators, crowd-sourcing involves active parsing of the scientific literature to update and addend contents in an automated fashion. This automated crowd-sourcing process is necessary given that there are reports of >290,000 proteins and >25,000 endogenous metabolites in humans.7,10 While databases such as those previously mentioned provide an important service to the biomedical research community, they remain incomplete, and in some cases, it is challenging to recognize where they are incomplete. As a result, research groups end up developing their own data libraries or databases. Consequences of building personalized libraries and databases include a loss of time and resources due to a redundancy of data acquisition and curation, limited scientific collaboration due to incompatibilities (e.g., informatics, jargon, etc.), and research opacity because raw data is often not referenced or otherwise available.11 To alleviate these consequences, we propose that field experts build a crowd-sourced database that integrates into successful pre-existing workflows. It should be noted that contributors to the database (i.e., the crowd) will most likely also be field experts. Since database developing is an iterative process, open dialogue...
Table 1. Data Collection Terms

| term        | description                                                                                       |
|-------------|--------------------------------------------------------------------------------------------------|
| data set    | a collection of data                                                                            |
| database    | an organized collection of records that is standardized to enable searching and retrieval of content |
| data library| a collection of data materials in various formats with the capability of providing information to a target group |
| repository  | a collection of digital documents stored for preservation and public access                       |

between the developer and crowd is encouraged to meet field specific needs. Two examples of successful crowd-sourced databases are the MassBank of North America (MONA) and the Unified Collision Cross Section Compendium. MONA was the first public repository for small molecule mass spectral data. The Unified CCS Compendium is a database of drift-tube ion mobility mass spectrometry data of omic compounds. Here, we discuss a model to create a crowd-sourced omics database including five pillars of database features that need to be considered. Further, we discuss design concepts and how crowd-sourcing is currently done within the research community.

2. DATABASE FEATURES

A generalized schematic of how an omic database is developed is shown in Figure 1. Specifically, an initial data set is processed via data curation, standardization, and annotation with metadata. Next, the curated data set is compiled into a database that gets disseminated to others within a field. These end users utilize the information within the database to gain knowledge about their own experiments, which leads to novel scientific conclusions and the formulation of future questions. These new conclusions become newly generated data sets which then undergo dereplication, validation, and standardization to be added to the pre-existing database. Even though this process only contains five general stages, much should be considered along each step. It is recommended that the following features be considered before data acquisition and curation as well as development of a database begins.

**Standardization Requirements.** The overall goal of a database is to create a collection of data that end users can use with as few barriers as possible. One way to minimize barriers that end users will face is to create a standardized system which includes a standard data type, reporting format, terminology, quality control process, metadata inclusion, and/or reference material information, as shown in Figure 2. Data type refers to what kind of data the database will contain. Will it be data from one specific technology or technique? Will this data be in the primary (raw) or secondary (processed) form? Primary data is preferred for scientific transparency. However, it is often larger and will require more computational storage space and data management resources. Secondary data is more common due to their smaller storage requirements and ease of use. Most end users prefer to look at conclusive or summative data.

Reporting format and standard terminology must be considered. If a database contains primary data, how will that data be uploaded by the user? The database management system will need the capabilities to handle large data file transfers as well as automated indexing of addended data. Database management systems are discussed further in section 3. Databases that contain secondary data are easier to manage in terms of indexing and storage needs. However, the database developers need to create a standard format that is both informative and facile enough for end-users to comply with. Database specific terminology must be defined from the beginning so that users understand what is required of them and how to use the data within. Such terminology should unambiguously convey experimental design, data acquisition, and data processing parameters. Furthermore, any information needed to provide a context for the reported results should also be included. This enables other users to fully understand the stated conclusions and compare studies from different research groups. One crucial aspect of databases is that each record within the database needs to have a unique identifier. In metabolomics, this can be a compound’s InChI Key or molecular structure. In genomics, this could be a specific gene locus. This unique identifier enables universal indexing of records without ambiguity and quick data import and export from the database.

**Metadata Documentation.** Metadata is defined as “minimum information needed to ensure that submitted data are sufficient for clear interpretation and querying by other scientists.” As previously mentioned, inclusion of contextual information as well as experimental procedures is imperative. Providing metadata maximizes a data set’s utility by allowing others to understand, reproduce, and build off of reported work.

Figure 1. General Database Development. Databases start with an initial data set that undergoes standardization. This standardized database is disseminated through the research–peer review cycle. Subsequently, new data is added to the existing database and the cycle begins again.

Figure 2. Database Features. To maximize the utility of their database, developers should consider the data type, standard terminology, included metadata, reference materials, and data management systems when designing their database.
Database developers should provide guidelines about what type of information is needed for interpretation and querying for a specific field. Alternatively, a database can contain primary references for a given data set such that end users can obtain the metadata elsewhere. The Minimum Information for Biological and Biomedical Investigations (MIBBI) is a useful resource when deciding what metadata should be included.\textsuperscript{16} MIBBI contains registries of reporting efforts for biological/biomedical studies as well as field specific recommendations which include the Minimum Information about a Genome Sequence (MIGS) and Minimum Information about a Metagenomic Sequence (MIMS) for genomic and transcriptomic data, the Minimum Information about a Proteomics Experiment (MIAPE) for proteomics data, and the Core Information for Metabolomics Reporting (CIMR) for metabolomics data.\textsuperscript{16−19}

**Reference Materials.** The standardization process goes beyond informatics and reporting. It is recommended that database guidelines include a physical standard such that the reagent can be added as a control in experiments. This standard would serve as a stable reference point for data quality when compared to known experimental values for said reference standard.\textsuperscript{19} Having a reference standard that is accessible to a breadth of end users enables data comparisons and quality checks between experiments, across platforms, and between research groups. This is particularly important when used in omics experiments in clinical/diagnostic settings. The reference standard choice is often decided by the users within the field, but standard materials and associated measurements are provided by the National Institute of Standards Technology (https://www.nist.gov/services-resources/standards-and-measurements) in the United States and the Laboratory of the Government Chemist (http://lgc.co.uk) in the United Kingdom.

**Quality Assurance.** The success of any database is contingent upon its quality assurance (QA). QA for a database is the process that ensures that data and informatic tools within meet a certain standard as dictated by the database design model and specifications.\textsuperscript{20} QA is a twofold process: The first is during initial development of the database. It begins with the standardization procedures previously described as well as developing tools that can audit the database intermittently. These audits should ensure that all of the data is represented accurately and as planned, and that all of the database functionality is operating properly.\textsuperscript{20} The second process is when new data is added to the database. Procedures should be in place to vet the quality of the incoming data such that it meets the standardization requirements previously set forth.\textsuperscript{20} This ensures that the integrity of the database is maintained.

### 3. DESIGN CONCEPTS

**Conceptual Design.** The conceptual design of a database defines the data requirements of and the application of the database in question.\textsuperscript{15} This includes the metadata requirements as well as the standardized data type and format as previously discussed. This pillar of the database design is field and data specific. Discussions about the aforementioned database features from section 2 should be included during the conceptual design phase.

**Logical Design.** The logical design of a database involves the implementation and management systems of the database.\textsuperscript{15} It can be thought of as the “back-end” design of a database. Particular attention should be paid to determining how and when data will be normalized and background/noise corrected and which, if any, further transformations will occur. Often, genomic and transcriptomic databases contain primary data that is normalized by the database infrastructure during the data submission process. Proteomic and metabolomic data is usually presented as secondary data that includes the larger context of the experiment performed.\textsuperscript{15} Additionally, proteomic and metabolomic data have more variety in potential output, in terms of content and size, when compared to genomic and transcriptomic data. As a result, this type of data is normalized and background subtracted before submission to a database. Developers need to also consider if data sets are to be kept separate or merged. Data sets can be merged to save space and represent a crowd-sourced conclusion. Keeping them separate enables study comparisons. Both options are used in bioinformatics, and the overall aim of a given database will determine which is more suitable.

Once data is added into a database, it needs to be maintained. Several options exist to retain order and search capabilities of a database. For large data sets, SQL Server can be used. It works well with relational data, especially if individual records have many attributes associated with them.\textsuperscript{21} SQL data sets can be transformed into the XML data format which is amenable to many informatic solutions and coding languages. For smaller data sets, developers can utilize spreadsheet-based solutions which are easily hosted online and can be transferred via CSV data format. Automated maintenance and quality checks are recommended for both options and should be determined before development begins. However, maintenance is an iterative process and should be adjusted as needed. These tendencies are general and individual developers should choose the appropriate logistical design for their specific type of data.

**Physical Design.** The physical design of a database involves determining the hardware necessary to support the database as well as the design of any graphical interfaces needed.\textsuperscript{15} It can be thought of as the front-end design pillar in which the ease of use by experts in the field is the top priority. Developers should determine if their database will be hosted via an application or online. Additionally, developers should decide if and what to archive (i.e., should outdated results be kept?) as well as design tools that can query live and archived data.\textsuperscript{20} Informatic tools such as statistical models and data visualization graphics are also designed during this phase.\textsuperscript{15} This state of the design process is the most open-ended, and graphical output can vary widely. Furthermore, it is the most iterative stage, as databases are likely to change depending on their contents and new tools being added. Figure 3 summarizes the three design concepts of planning a database.

### 4. CROWD-SOURCING DATA

The past decade has seen a push for data sharing and crowd-sourcing research.\textsuperscript{11,22−24} The age of big data has matured alongside the ongoing improvements in computational power and data storage capacity. These concurrent movements allow researchers to gather more data than that which they could have collected independently and perform wide-scale studies not previously possible. There are two main crowd-sourcing techniques being used: (1) data-mining from publicly available large data sets and (2) crowd-sourcing data acquisition and/or analysis.\textsuperscript{24} The first technique is often used in public health studies where large numbers of data points are needed and through which patient histories are sifted.\textsuperscript{24} The second type of crowd-sourcing is used in multiple disciplines within biomedical research including computational chemistry, genomics, medic-
5. CONCLUSIONS AND OUTLOOK

Biomedical research is moving toward using big data that is often crowd-sourced in order to make more general conclusions of observed phenomena. Using crowd-sourced data has many benefits including a large sample pool, increased transparency throughout the scientific method, and more constructive discourse within a field or project. However, crowd-sourced data has a variable level of quality which can compromise results. By creating databases with crowd-sourced data sets, quality assurance procedures can be put into place. While this process is laborious, the end result is a highly curated database that can be used for the foreseeable future.

While there is no one metric of success for a database, one gauge can be how widely disseminated and utilized the database is. The Unified CCS Compendium is an example of a successful database given that it is used by field experts internationally in a variety of studies, both fundamental and biomedical based investigations. This success can be attributed to the Compendium being user-friendly and user-focused. Metadata standards as well as inclusion guidelines are explicitly provided to contributors. Furthermore, a standardized spreadsheet-based reporting format is provided along with guidelines about the quality control process. Further discussion and specific details on these attributes have been previously reported.13

Two final considerations pertain to (1) funding crowd-sourced databases and (2) practical considerations for the longevity and ongoing maintenance of these databases once they are developed. Resolutions to both of these considerations are ongoing discussions within the informatics community, and there is no one solution. One potential funding resource is collaboration with other research groups, the private sector, and/or a government agency. However, we propose a governmental/private sector alliance to retain the open-access databases post development while accepting contributions from academic researchers. This provides for any practical resources needed to maintain these large databases as well as continued open dialogue between all three sectors. Ultimately, this facilitated collaboration will enable biomedical researchers to take advantage of the opportunities and discoveries that this wealth of data presents.

AUTHOR INFORMATION

Corresponding Author
*E-mail: john.a.mclean@vanderbilt.edu.

ORCID
Jaqueline A. Picache: 0000-0002-1948-9243
Jody C. May: 0000-0003-4871-5024
John A. McLean: 0000-0001-8918-6419

Author Contributions
J.A.P. provided conceptual input for this work, created all of the figures within, and is the primary writer of this manuscript. J.C.M. and J.A.M. provided conceptual input for this work. All authors have given approval to the final version of the manuscript.

Notes
The authors declare no competing financial interest.

Biographies

Jaqueline A. Picache received her B.S. at the University of Notre Dame before becoming an NIH postbaccalaureate Intramural Research Training Award fellow under Dr. Forbes D. Porter. She then matriculated into the Chemistry Ph.D. program at Vanderbilt.
University. She has been a graduate student under John A. McLean since 2017. Her research interests include developing bioinformatic solutions and utilizing ion-mobility mass spectrometry to perform global, molecular profiling experiments.

Jody C. May is the Associate Director of the Center for Innovative Technology and a Research Assistant Professor in Chemistry at Vanderbilt University. He completed his B.S. in Chemistry at the University of Central Arkansas and his Ph.D. in Analytical Chemistry from Texas A&M University. His research interests focus on the development of next generation ion mobility-mass spectrometry techniques for untargeted molecular discovery.

John A. McLean is the Chair of the Department of Chemistry, the Stevenson Professor of Chemistry, and the Director of the Center for Innovative Technology. He received a B.S. in Chemistry from the University of Michigan, his Ph.D. at the George Washington University, and completed postdoctoral training at Texas A&M University prior to joining the faculty at Vanderbilt University. His research focuses on the intersection of instrumentamation and bioinformatics in support of systems, synthetic, and chemical biology.

**ACKNOWLEDGMENTS**

Financial support for this research was provided by the National Institutes of Health (NIH NIGMS R01GM092218 and NIH NCI 1R03CA222452-01) and the NIH supported Vanderbilt Chemical Biology Interface training program (ST32GM065086-16).

**REFERENCES**

(1) May, J. C.; McLean, J. A. Annu. Rev. Anal. Chem. 2016, 9, 387.
(2) Costa, F. F. Drug Discovery Today 2014, 19 (4), 433.
(3) Baker, E. S.; Patti, G. J. Am. Soc. Mass Spectrom. 2019, 30 (10), 2031.
(4) Scheubert, K.; Hufsky, F.; Petras, D.; Wang, M.; Nothias, L.-F.; Duehrkop, K.; Bandeira, N.; Dorrestein, P.; Boecker, S. Nat. Commun. 2017, 8, 1494.
(5) Waldman, S. A.; Terzic, A. Clin. Pharmacol. Ther. 2016, 99 (3), 250.
(6) Levine-Clark, M.; Carter, T. M. ALA Glossary of Library and Information Science; American Library Association: Chicago, IL, 2013; pp 86–87, 163, 230.
(7) Wishart, D. S.; Feunang, Y. D.; Marcu, A.; Guo, A. C.; Liang, K.; Vázquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N.; Sayeeda, Z.; Lo, E.; Assempropor, N.; Berjanski, M.; Singhal, S.; Arndt, D.; Liang, Y.; Badran, H.; Grant, J.; Serra-Cayuela, A.; Liu, Y.; Mandal, R.; Neveu, V.; Pon, A.; Knox, C.; Wilson, M.; Manach, C.; Scalbert, A. Nucleic Acids Res. 2018, 46 (D1), D608.
(8) Kim, S.; Chen, J.; Cheng, T.; Gindyuley, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. Nucleic Acids Res. 2019, 47 (D1), D1102.
(9) Bateman, A. Nucleic Acids Res. 2019, 47 (D1), D506.
(10) Kim, M. S.; Pinto, S. M.; Getnet, D.; Niranjo, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kellar, D. S.; Isselrin, R.; Jain, S.; Thomas, J. K.; Muthusamy, B.; Leal-Rojas, P.; Kumar, P.; Sahasrabuddhe, N. A.; Balakrishnan, L.; Advani, J.; George, B.; Renuse, S.; Selvan, L. D. N.; Patil, A. H.; Nanjappa, V.; Radhakrishnan, A.; Prasad, S.; Subbannayya, T.; Raju, R.; Kumar, M.; Sreenivasamurthy, S. K.; Marimuthu, A.; Sathe, G. J.; Chavan, S.; Datta, K. K.; Subbannayya, Y.; Sahu, A.; Yelamanchi, S. D.; Jayaram, S.; Rajagopalan, P.; Sharma, J.; Murthy, K. R.; Syed, N.; Goel, R.; Khan, A. A.; Ahmad, S.; Dey, G.; Mudgal, K.; Chatterjee, A.; Huang, T. C.; Zhong, J.; Wu, X.; Shaw, F. P.; Freed, D.; Zahari, M. S.; Mukherjee, K. K.; Shankar, S.; Mahadevan, A.; Iam, H.; Mitchell, C. J.; Shankar, S. K.; Satishchandra, P.; Schroeder, J. T.; Srideshmukh, R.; Maitra, A.; Leach, S. D.; Drake, C. G.; Halushka, M. K.; Prasad, T. S. K.; Hruban, R. H.; Kerr, C. L.; Bader, G. D.; Iacobuzio-Donahue, C. A.; Gowda, H.; Pandey, A. Nature 2014, 509 (7502), 575.
(11) Riccardi, E.; Pantano, S.; Potestoi, R. Interface Focus 2019, 9 (3).
(12) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, M. Y.; Nakamachi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioh, T. J. Mass Spectrom. 2010, 45 (?), 703.
(13) Picache, J. A.; Rose, B. S.; Balinski, A.; Leaptrot, K. L.; Sherrod, S. D.; May, J. C.; McLean, J. A. Chem. Sci. 2019, 10 (4), 983.
(14) Chervitz, S. A.; Deutsch, E. W.; Field, D.; Parkinson, H.; Quackenbush, J.; Rocca-Serra, P.; Sansone, S.-A.; Stoeckert, C. J., Jr.; Taylor, C. F.; Taylor, R.; Ball, C. Bioinformatics for Omics Data; Humana Press: New York, 2011; pp 31–69.
(15) Schneider, M. V.; Orchard, S. Bioinformatics for Omics Data; Humana Press: New York, 2011; pp 3–30.
(16) Taylor, C. F.; Field, D.; Sansone, S. A.; Aerts, J.; Apweiler, R.; Ashburner, M.; Ball, C. A.; Binz, P. A.; Bogue, M.; Booth, T.; Brazma, A.; Brinkman, R. B.; Michael Clark, A.; Deutsch, E. W.; Fiehn, O.; Foster, J.; Ghazal, P.; Gibson, F.; Gray, T.; Grimes, G.; Hancock, J. M.; Hardy, N. W.; Hermjakob, H.; Julian, R. K.; Kane, M.; Kettner, C.; Kinsinger, C.; Kolker, E.; Kuiper, M.; Nove, N. L.; Leebens-Mack, J.; Lewis, S. E.; Lord, P.; Mallon, A. M.; Marthandan, N.; Masuya, H.; McNally, R.; Mehrle, A.; Morrison, N.; Orchard, S.; Quackenbush, J.; Recey, J. M.; Robertson, D. G.; Rocca-Serra, P.; Rodriguez, H.; Rosenfelder, H.; Santoyo-Lopez, J.; Scheuermann, R. H.; Scholer, D.; Smith, B.; Snape, J.; Stoeckert, C. J.; Tipton, K.; Sterk, P.; Untergasser, A.; Vandesompele, J.; Wiemann, S. Nat. Biotechnol. 2008, 26 (8), 889.
(17) Field, D.; Garrity, G.; Gray, T.; Morrison, N.; Selengut, J.; Sterk, P.; Tatusova, T.; Thomson, N.; Allen, M. J.; Angiuoli, S. V.; Ashburner, M.; Axelrod, N.; Baldauf, S.; Ballard, S.; Boore, J.; Cochrane, G.; Cole, J.; Dawyndt, P.; O.; Yen, P. DePamphilis, C.; Edwards, R.; Furhake, N.; Feldman, R.; Gilbert, J.; Gilna, P.; Glöcker, F. O.; Goldstein, P.; Guralnici, R.; Haft, D.; Hancock, D.; Hermjakob, H.; Hertz-Fowler, C.; Hugenholtz, P.; Joint, I.; Kagan, L.; Kane, M.; Kennedy, J.; Kowalchuk,
G.; Kottmann, R.; Kolker, E.; Kravitz, S.; Kyprides, N.; Leebens-Mack, J.; Lewis, S. E.; Li, K.; Lister, A. L.; Lord, P.; Maltsev, N.; Markowitz, V.; Martiny, J.; Methe, B.; Mizrahi, I.; Moxon, R.; Nelson, K.; Parkhill, J.; Proctor, L.; White, O.; Sansone, S.-A.; Spiers, A.; Stevens, R.; Swift, P.; Taylor, C.; Tateno, Y.; Tett, A.; Turner, S.; Ussery, D.; Vaughan, B.; Ward, N.; Whetzel, T.; Gil, I. S.; Wilson, G.; Wipat, A. Nat. Biotechnol. 2008, 26 (5), 541.

(18) Martínez-Bartolomé, S.; Binz, P.-A.; Albar, J. P. Plant Proteomics: Methods and Protocols; Humana Press: New York, 2014; Vol. 1072, pp 765–780.

(19) Spicer, R. A.; Salek, R.; Steinbeck, C. Sci. Data 2017, 4, 170137.

(20) Harel, A.; Dalah, I.; Pietrokovski, S.; Safran, M.; Lancet, D. Bioinformatics for Omics Data; Humana Press: New York, 2011; Vol. 719, pp 71–96.

(21) Sander, A.; Wauer, R. J. Biomed. Semantics 2019, 10, 7.

(22) Björnmalm, M.; Caruso, F. Angew. Chem., Int. Ed. 2018, 57 (5), 1122.

(23) Silberzahn, R.; Uhlmann, E. L. Nature 2015, 526 (7572), 189.

(24) Khare, R.; Good, B. M.; Leaman, R.; Su, A. I.; Lu, Z. Briefings Bioinf. 2016, 17 (1), 23.

(25) Budin-Ljøsne, I.; Isaeva, J.; Knoppers, B. M.; Tassé, A. M.; Shen, H. Y.; McCarthy, M. I.; Harris, J. R. Eur. J. Hum. Genet. 2014, 22 (3), 317.

(26) Cooper, S.; Khatib, F.; Treuille, A.; Barbero, J.; Lee, J.; Beerli, M.; Leaver-Fay, A.; Baker, D.; Popovic, Z.; Players, F. Nature 2010, 466 (7307), 756.

(27) Poussin, C.; Belcastro, V.; Martin, F.; Boué, S.; Peitsch, M. C.; Hoang, J. Chem. Res. Toxicol. 2017, 30 (4), 934.

(28) Goodwin, C. R.; Sherrod, S. D.; Marasco, C. C.; Bachmann, B. O.; Schramm-Sapyta, N.; Wikswo, J. P.; McLean, J. A. Anal. Chem. 2014, 86, 6563.

(29) Plante, P.-L.; Francovic-Fontaine, É.; May, J. C.; McLean, J. A.; Baker, E. S.; Laviolette, F.; Marchand, M.; Corbeil, J. Anal. Chem. 2019, 91 (8), 5191.

(30) Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; Greiner, R.; Wishart, D. S. J. Cheminf. 2016, 8, 61.