Metabolite reporting within large-scale studies: where do we stand?

Ghina Hajjar, Franck Giacomoni, Blandine Comte, Estelle Pujos-Guillot

Metabolite reporting within large-scale studies

WHERE DO WE STAND?

Ghina Hajjar [a], Franck Giacomoni [a], Blandine Comte [a], Estelle Pujos-Guillot [a]

[a] Université Clermont Auvergne, INRAE, UNH, Plateforme d'Exploration du Métabolisme, MetaboHUB Clermont, Clermont-Ferrand, France

Introduction

Metabolomic Epidemiology, aiming to study population-based variation in the human metabolome with respect to health-related outcomes or exposures [1], is expanding within the metabolomics research community. Therefore, data sharing is crucial to enable the delivery of its scientific potential. However, despite the existence of repositories, as well as minimum reporting standards for chemical analyses, there is no established standards for metabolite data reporting. In this context, the aim of this work was to review the existing practices in terms of metabolite reporting in different scientific communities by taking into consideration the metabolite identification levels [2] and recent nomenclature guidelines, particularly for lipids [3] and ion annotation [4].

Evaluation of metabolite reporting: Metabolites were collected from human large-scale studies published in peer-reviewed journals covering several specialties (*see table below*). Reported plasma metabolites were compared between studies before and after data curation (*see figure 1*) in order to evaluate the current state of metabolite reporting in published studies.

Published study	Scientific community	Number of metabolites
[5] Gonzalez-Dominguez *et al.* **2020** *Analytical Chemistry*, 92: 13767-13775	Analytical chemistry	677
[6] Liu *et al.* **2020** *Analytical Chemistry*, 92: 8836-8844	Analytical chemistry	487
[7] Pietzner *et al.* **2021** *Nature Medicine*, 27: 471-479	Epidemiology & medicine	585

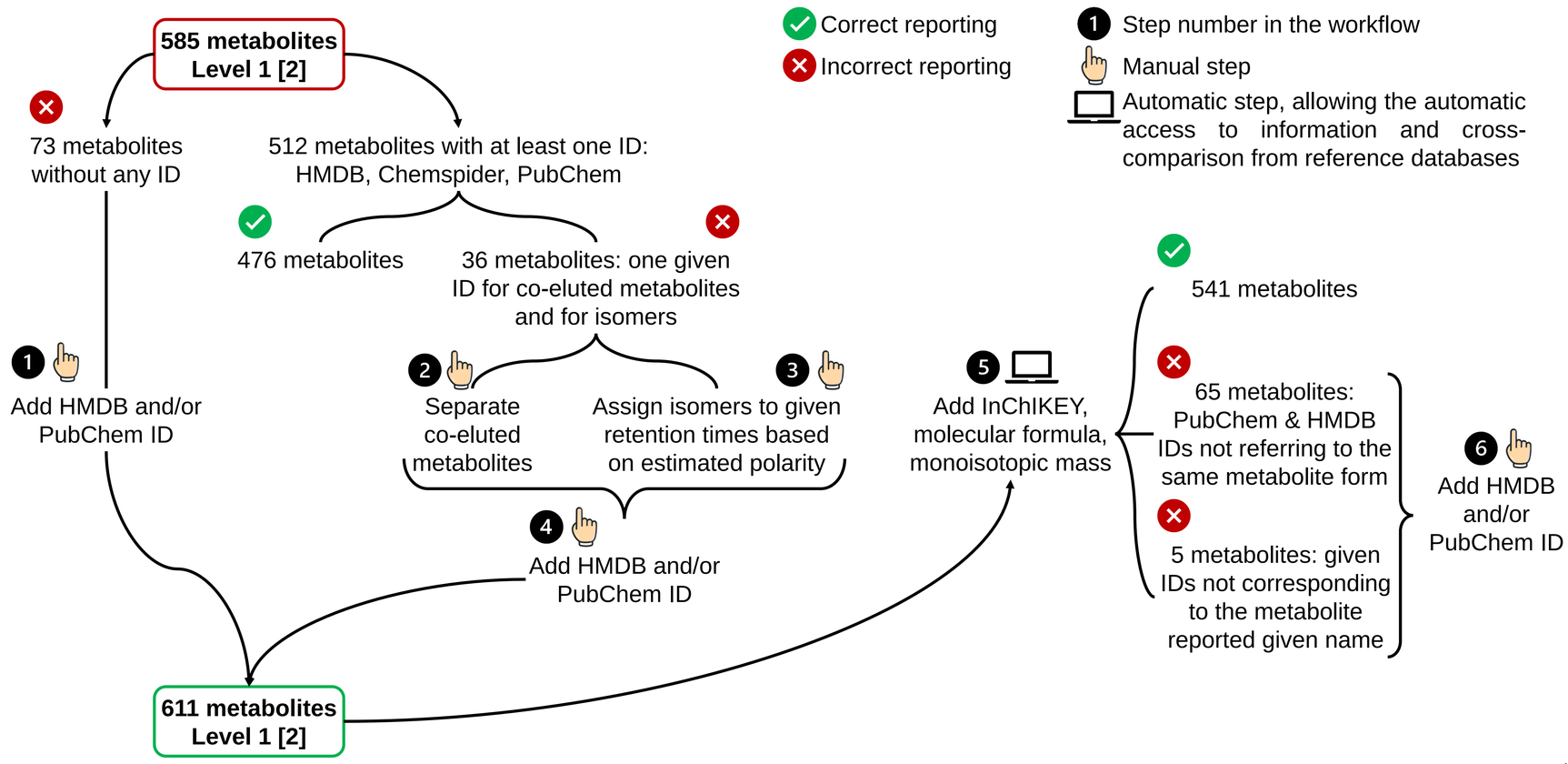

Figure 1. ► Data curation workflow (example with metabolites reported in [7])

Comparison of metabolite reporting between studies

Ambiguities observed in published results:

1 to 31% of metabolites [5-7] were reported with either missing or incoherent information

- Given IDs not referring to the same isomer
- Metabolite name not corresponding to the molecular formula
- Isomers were listed with their corresponding retention times, yet without any indication of the isomer's identity

Metabolite reporting depending on the scope of the study:

In both scientific communities:

- Identification levels were reported according to [2]
- Metabolites were reported using multiple IDs, referring to both biological and chemical databases

Specific to scientific community and/or study:

- Reported common names were different between the communities and/or studies (*e.g.* acid or basic form), with various analytical metadata (*e.g.* MS annotations).

Ambiguities observed across databases: 11% of metabolites reported in [7] presented incoherent information across HMDB and PubChem databases

- Nomenclatures
  e.g. Betaine reported by HMDB is $C_5H_{12}NO_2^+$ (HMDB0000043) linked to PubChem CID 248 named trimethyl glycine whereas Betaine reported by PubChem is $C_5H_{11}NO_2$ (CID 247)
- Optical isomerism / Stereochemistry of asymmetric carbons
  e.g. D-Xylitol: HMDB0002917 (a) *vs* PubChem CID 6912 (b)
  (a) InChI=1S/C5H12O5/c6-1-3(8)5(10)4(9)2-7/h3-10H,1-2H2/t3-,4+,**5+**
  (b) InChI=1S/C5H12O5/c6-1-3(8)5(10)4(9)2-7/h3-10H,1-2H2/t3-,4+,**5?**
- Incoherent links (acid/base; zwitterionic or canonical forms, molecules with a permanent charge)
  e.g. L-Carnitine: HMDB0000062 $C_7H_{16}NO_3^+$ linked to PubChem CID 10917 $C_7H_{15}NO_3$

Comparing level 1 metabolites between studies

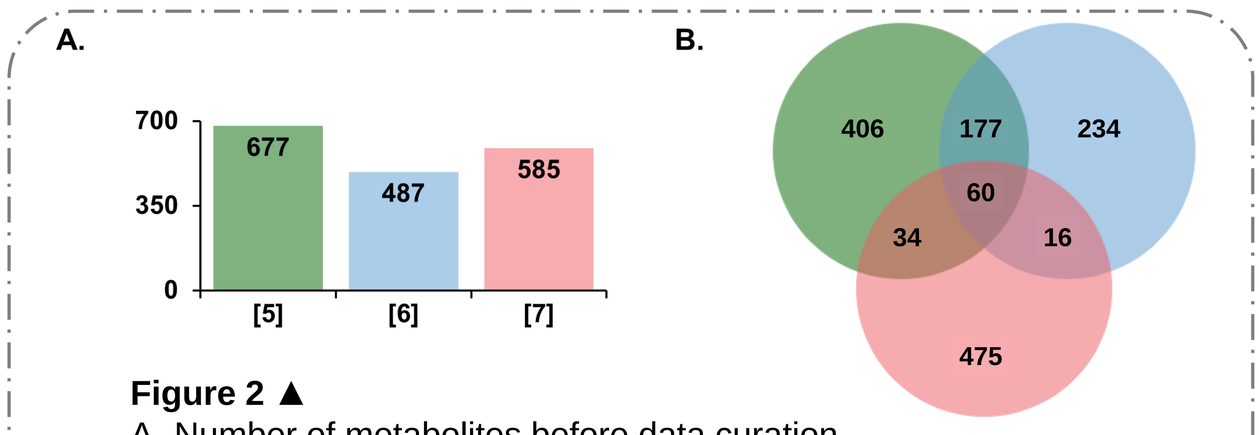

Figure 2 ▲

A. Number of metabolites before data curation.

B. Comparison of metabolites reported in [5], [6] & [7] using given reported metabolite names.

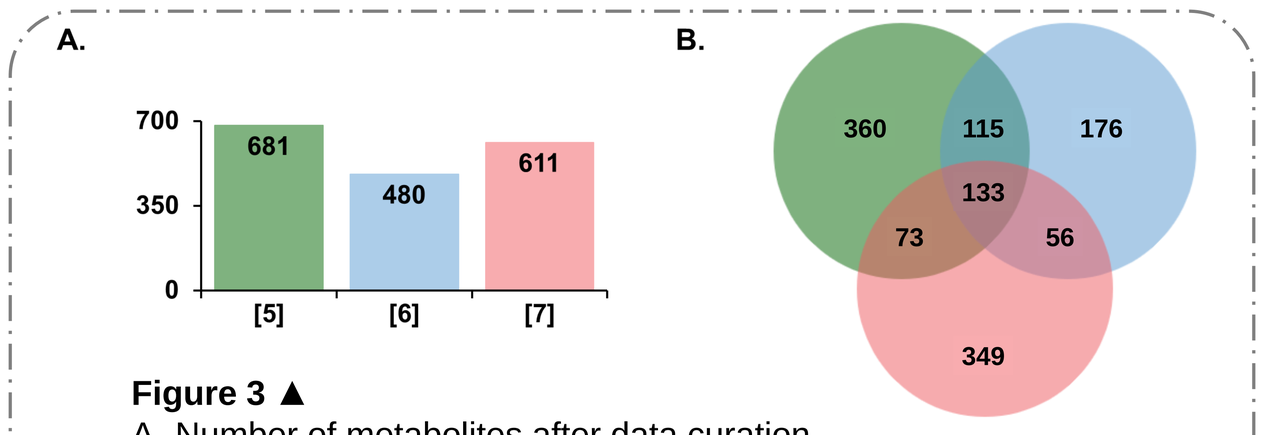

Figure 3 ▲

A. Number of metabolites after data curation.

B. Comparison of metabolites reported in [5], [6] & [7] using metabolite InChIKEYs.

Conclusions & Perspectives

Although not required, the compact hash code of the IUPAC International Chemical Identifier "InChIKey" was found to be the most suitable identifier for comparing reported metabolites between studies. It is therefore recommended to use either this identifier, or perform curation of the association between IDs and exact known structure of reported metabolites. In order to provide guidelines for a more effective and reproducible metabolomics data sharing, other metabolite reporting within large-scale studies will be explored for a wider coverage of metabolite classes in large human cohorts such as studies published by the COnsortium of METabolomics Studies (COMETS) [8].

References

[1] Lasky-Su *et al.* 2021 *Metabolomics*, 17:45; [2] Sumner *et al.* **2007** *Metabolomics*, 3: 211-221; [3] Liebisch *et al.* **2020** *J. Lip. Res.*, 61: 1539-1555; [4] Damont *et al.* **2019** *J. Mass Spec.*, 54: 567-582; [5] Gonzalez-Dominguez *et al.* **2020** *Anal. Chem.*, 92: 13767-13775; [6] Liu *et al.* **2020** *Anal. Chem.*, 92: 8836-8844; [7] Pietzner *et al.* **2021** *Nat. Med.*, 27: 471-479; [8] Bardou *et al.* **2014** *BMC Bioinfo.*, 15:293; [8] Yu *et al.* **2019** *Am. J. Epidemiol.*, 188: 991-1012.

Acknowledgments

This work is supported by the French Ministry of Research and National Research Agency as part of the French metabolomics and fluxomics infrastructure (MetaboHUB-ANR-INBS-0010). We would like to thank the French-Speaking Network of Metabolomics and Fluxomics (RFMF) for the travel grant awarded to G.H.

ghina.hajjar@inrae.fr