Metabolite reporting in large-scale studies within different metabolomics communities: DO WE SPEAK THE SAME LANGUAGE?

Ghina Hajjar, David Benaben, Nils Paulhe, Christophe Duperier, Olivier Filangi, Franck Giacomoni, Blandine Comte, Estelle Pujos-Guillot

Metabolite reporting in large-scale studies within different metabolomics communities:

DO WE SPEAK THE SAME LANGUAGE?

Ghina Hajjar [a], David Benaben [b,c], Nils Paulhe [a,c], Christophe Duperier [a,c], Olivier Filangi [c,d], Franck Giacomoni [a,c], Blandine Comte [a], Estelle Pujos-Guillot [a]

[a] *Université Clermont Auvergne, INRAE, UNH, Plateforme d'Exploration du Métabolisme, MetaboHUB Clermont, Clermont-Ferrand, France*
[b] *UMR Biologie du Fruit et Pathologie, Université de Bordeaux, INRAE, Bordeaux Metabolome, MetaboHUB, PHENOME-EMPHASIS Villenave d'Ornon, France*
[c] *INRAE "Metabolomics Semantic DataLake" team, EMPREINTE & PROSODIe CATI, France*
[d] *IGEPP, INRAE, Institut Agro, Université Rennes, P2M2, Le Rheu, France*

INTRODUCTION

High throughput metabolomic studies are increasing within **various scientific communities** from analytical chemistry to epidemiology. Therefore, **standardized reporting** is crucial for data sharing. To date, there are no established standards for **metabolite reporting**. Our **objective** was to review the existing practices in terms of metabolite reporting in different scientific communities both in **published results** [1-4] and across **databases** [HMDB V5.0 / PubChem (June 2022) / ChEBI (July 2022) / KNApSAcK (August 2022)].

MATERIALS & METHODS

We considered **plasma metabolites** reported in **human large-scale** studies from different communities, namely analytical chemistry, medicine and epidemiology [1-4]. We focused only on metabolites reported as **level 1** identification according to the Metabolomics Standard Initiative (MSI) [5].

Published study	Scientific community	Number of metabolites
[1] Gonzalez-Dominguez *et al.* **2020** *Analytical Chemistry*, 92: 13767-13775	Analytical chemistry	677
[2] Liu *et al.* **2020** *Analytical Chemistry*, 92: 8836-8844	Analytical chemistry	487
[3] Pietzner *et al*. **2021** *Nature Medicine*, 27: 471-479	Epidemiology & medicine	585
[4] Yu *et al*. **2019** *American Journal of Epidemiology*, 188: 991-1012	Epidemiology & medicine	588

➢ Evaluation of metabolite reporting in publications

We applied a data curation workflow on the published lists of annotated metabolites. The workflow consisted of a combination of manual and automatic steps (***Figure 1***).

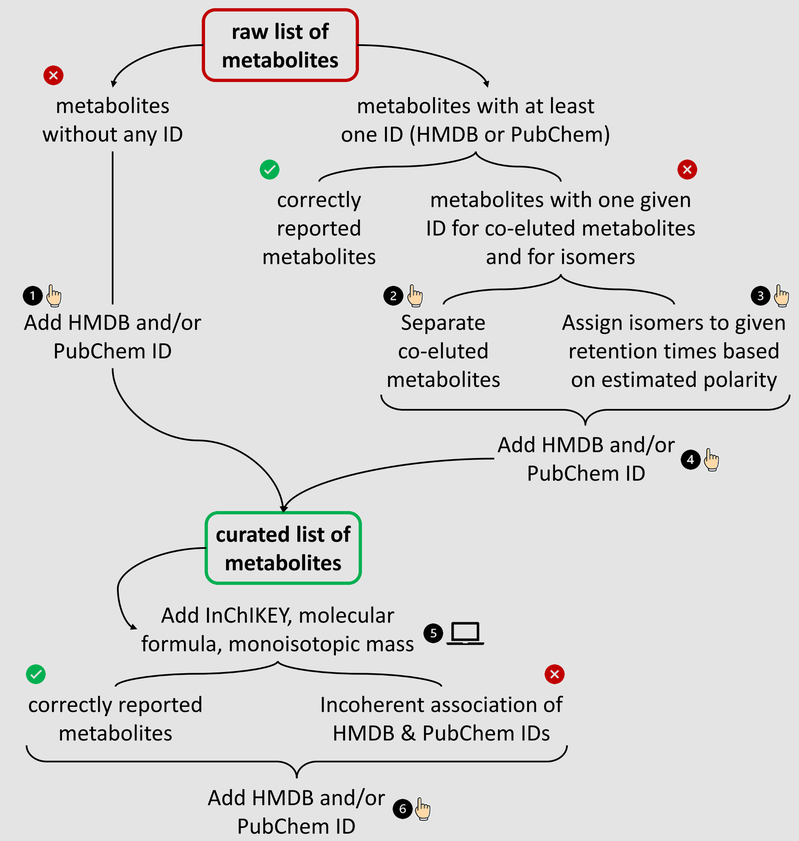

Figure 1. *Data curation workflow applied to metabolites reported in published studies [1-4].*

➢ Evaluation of cross-linking of metabolites across databases

Information was calculated under a Big Data infrastructure (Apache Spark) and Scala programming language with the help of the Metabolomics Semantic DataLake team (***Figure 2***).

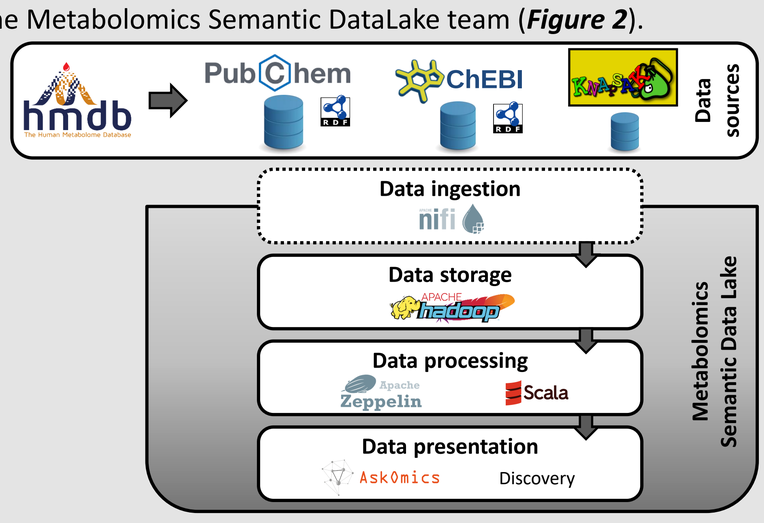

Figure 2. *Data flow in a 'Metabolomics Semantic DataLake' infrastructure: Application to assess the cross-linking between metabolites reported in HMDB V5.0 and their reference to PubChem, ChEBI or KNApSAcK.*

RESULTS

➢ Ambiguities observed in published results

Metabolite reporting was not standardized within scientific communities.

- Metabolites were reported using IDs referring to both biological & chemical databases and with identification levels according to the MSI [5]
- Reported common names were different between communities and/or studies
- Available metadata (e.g. MS annotations and other analytical data) were dependent on the scope of the study

Some metabolites were reported with either missing or incoherent information. After data curation (***see Figure 1***), metabolites were compared between studies using either the names given by the authors or the InChIKeys (***Figure 3***).

The overall length of an InChIKey is fixed at 27 characters (dashes included)

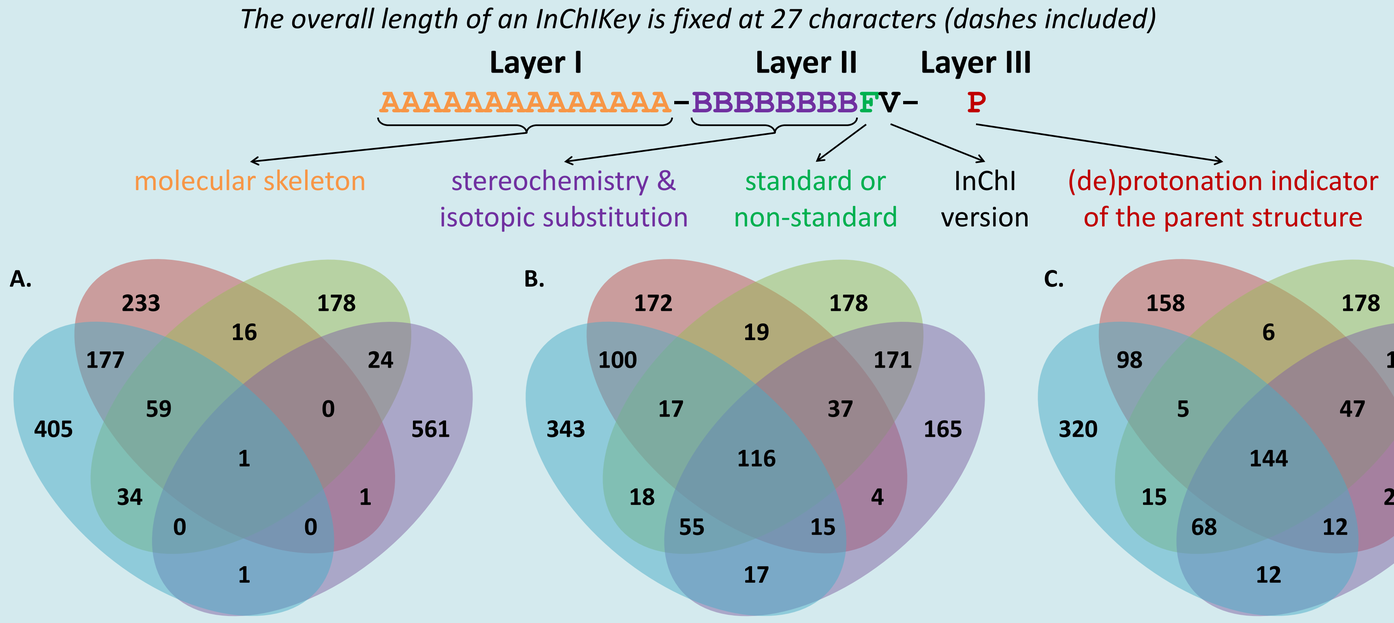

Figure 3. *Venn diagrams used to compare metabolites reported in Gonzalez-Dominguez et al. 2020 (●), Liu et al. 2020 (●), Pietzner et al. 2021 (●) & Yu et al. 2019 (●) and using **A.** reported metabolite names, **B.** metabolite complete InChIKEYs, & **C.** layer I of the InChIKEYs.*

As shown in ***Figure 3***, using the reported metabolite **names**, we could spot **only 1** metabolite reported by all 4 studies. However, using the **InChIKey**, the common metabolites were **above 100**. ***Figures 3B*** & ***3C*** show the importance of the **isomers'** identification in metabolite reporting. In this case, the InChIKey was the **most suitable identifier** for data i**nter-comparison**.

➢ Ambiguities observed across databases

Some metabolites' nomenclatures were incoherent across databases

e.g. $C_5H_{12}NO_2^+$: HMDB: **Betaine** (HMDB0000043); PubChem: **Trimethyl glycine** (CID 248); ChEBI: **N,N,N-trimethylglycinium** (CHEBI:41139)

All HMDB entries with ambiguous cross-links to ChEBI, PubChem and KNApSAcK were identified. The percentage of mismatches was independent of the total number of cross-links provided.

- ChEBI: 13 198 HMDB entries with external reference to ChEBI → 4% mismatch (***Figure 4A***)
- PubChem: 103 593 HMDB entries with external reference to PubChem → 2% mismatch (***Figure 4B***)
- KNApSAcK: 7 674 HMDB entries with external reference to KNApSAcK → 66% mismatch

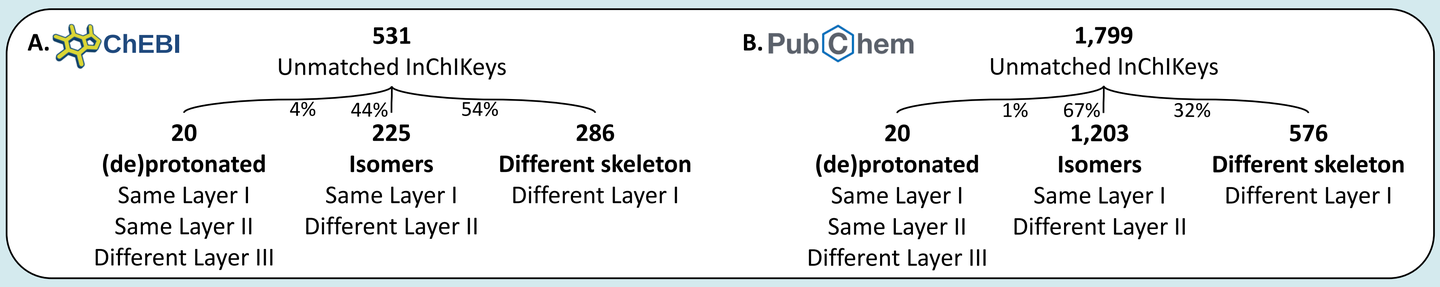

Figure 4*: Distribution of HMDB entries with external references to (**A**) ChEBI and (**B**) PubChem having unmatched InChIKeys.*

Mismatchings were classified according to the unmatched InChIKey layer(s)

- (De)Protonation of the core parent structure
  - *e.g.* N6,N6,N6-Trimethyl-L-lysine: HMDB0001325 (a) linked to ChEBI:17311 (b)
    - (a) MXNRLFUSFKVQSK-QMMMGPOBSA-**N (parent structure)**
    - (b) MXNRLFUSFKVQSK-QMMMGPOBSA-**O (protonated structure)**
- Isomerism, stereochemistry
  - *e.g.1.* D-Xylitol: HMDB0002917 (c) linked to PubChem CID 6912 (d)
    - (c) InChI=1S/C5H12O5/c6-1-3(8)5(10)4(9)2-7/h3-10H,1-2H2/t3-,4+,**5+** → HEBKCHPVOIAQTA-**SCDXWVJYSA**-N
    - (d) InChI=1S/C5H12O5/c6-1-3(8)5(10)4(9)2-7/h3-10H,1-2H2/t3-,4+,**5?** → HEBKCHPVOIAQTA-**NGQZWQHPSA**-N
  - *e.g.2.* Methylcysteine: HMDB0002108 (e) linked to PubChem CID 24417 (f)
    - (e) IDIDJDIHTAOVLG-**GSVOUGTGSA**-N → **D-isomer**
    - (f) IDIDJDIHTAOVLG-**VKHMYHEASA**-N → **L-isomer**
- Mismatch between structurally different compounds
  - *e.g.* 4,4-Dimethyl-5a-cholesta-8-en-3b-ol (HMDB0006840) linked to 5,6,7,8-tetrahydrofolylglutamic acid (CHEBI:27650)

CONCLUSION

This work will allow providing guidelines for a more effective and reproducible metabolomics data sharing (e.g. use common identifiers such as InChIKey, perform a deep data curation, etc.).

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).

References: [1] Gonzalez-Dominguez *et al.* **2020** *Anal. Chem.*, 92: 13767-13775; [2] Liu *et al.* **2020** *Anal. Chem.*, 92: 8836-8844; [3] Pietzner *et al*. **2021** *Nat. Med.*, 27: 471-479; [4] Yu *et al.* **2019** *Am. J. Epidemiol.*, 188: 991-1012; [5] Sumner *et al.* **2007** *Metabolomics*, 3: 211-221.
Databases: HMDB V5.0; PubChem (June 2022); ChEBI (July 2022); KNApSAcK (August 2022).

ghina.hajjar@inrae.fr