Caracterització dirigida d'ample abast del metaboloma d'*Escherichia coli* mitjançant LC-MS1 i LC-MS2

Roger Giné Bertomeu

TREBALL FINAL DE GRAU BIOTECNOLOGIA

Tutor acadèmic: Ricardo Cordero Otero, Departament Bioquímica i Biotecnologia, ricardo.cordero@urv.cat
En cooperació amb: IISPV / Metabolomics Platform
Supervisor/s: Òscar Yanes Torrado, Departament Enginyeria Elèctrica, Electrònica i Automàtica, oscar.yanes@urv.cat

Data de convocatoria: juny 2021

Jo, Roger Giné Bertomeu , amb DNI 47937947-J sóc coneixedor de la guia de prevenció del plagi a la URV *Prevenció, detecció i tractament del plagi en la docència: guia per a estudiants* (aprovada el juliol 2017) (http://www.urv.cat/ca/vida-campus/serveis/crai/que-us-oferim/formacio-competencies-nuclears/plagi/) i afirmo que aquet TFG no constitueixen cap de les conductes considerades com a plagi per la URV.

Tarragona, 5 de juny de 2021

Índex

1. Dades del Centre

La Plataforma Metabolòmica és un centre de recerca conjunt creat per la URV (Universitat Rovira i Virgili, Tarragona, Espanya) i CIBERDEM (Xarxa Espanyola d'Investigació Biomèdica en Diabetis i Trastorns Metabòlics Associats). La Plataforma Metabolòmica forma part de l'IISPV (Institut d'Investigació Sanitària Pere Virgili), una important organització d'investigació mèdica que realitza nombroses iniciatives de recerca al país. L'objectiu principal de la Plataforma Metabolòmica és oferir serveis metabolòmics als grups de recerca biomèdica i clínica de CIBERDEM i URV.

La principal visió de la Plataforma Metabolòmica no és només actuar com un centre de mesura, sinó oferir assessorament metabòlic, intentant implicar-se plenament en els experiments relacionats amb el metabolisme proposats pels grups. Això significa que la col·laboració comença des del principi de l'estudi (definició d'objectius, dimensió i característiques del conjunt de mostres, disseny experimental metabòlic) fins al final (processament i interpretació de dades), ajudant a obtenir resultats clínics rellevants per als diferents grups de recerca als qui dona servei.

2. Resum

Un dels principals problemes de la metabolòmica no dirigida (*untargeted*) és la baixa cobertura de metabòlits identificats, en bona part deguda a la gran quantitat de senyals redundants i inespecífiques generades pels instruments. En aquest context presentem un software de metabolòmica semi-dirigida per l'anàlisi de dades LC-MS1 i LC-MS2 (RHermes), que permet la identificació de potencialment milers de compostos presents en una mostra. L'estratègia es basa en una anotació context-específica de les dades MS1 a partir d'una base de dades de formules moleculars i adductes, seguida d'una adquisició específica de dades LC-MS2 per identificar els compostos. Per validar l'especificitat biològica d'Hermes, s'han analitzat mostres procedents d'*Escherichia coli* cultivada en un medi marcat amb ^{13}C-glucosa i un medi control sense marcar. Els resulats suggereixen que Hermes és capaç de cobrir més eficaçment el metaboloma de *E. coli*, requereix un menor nombre de scans MS2 adquirits i permet identificar compostos molt poc abundants que no poden ser identificats pels mètodes d'adquicisició MS2 convencionals.

Paraules clau: Metabolomics, stable-isotope labelling, LC-MS1, DDA, Hermes,

3. Introducció

La metabolòmica no dirigida (*untargeted*) és una branca de la metabolòmica que es fonamenta en l'anotació i identificació de metabòlits sense assumir un coneixement previ de la seva naturalesa, és a dir, basant-se exclusivament en l'anàlisi de les senyals adquirides pels instruments sense informació externa (1). La metabolòmica *untargeted* es contraposa a la metabolòmica dirigida (*targeted*), on els experiments es dissenyen per la quantificació d'un conjunt de metabòlits ja coneguts (usualment <50) (1).

Les dues aproximacions han estat complementàriament utilitzades els darrers anys per al descobriment de biomarcadors i nous metabòlits així com en estudis que associen signatures metabòliques a patologies. En el camp de la biotecnologia, els estudis de metabolòmica dirigida han facilitat el procés d'optimització de vies metabòliques en microorganismes, ja que és possible quantificar les abundàncies de múltiples compostos d'una mateixa via i identificar possibles colls d'ampolla a optimitzar.

Per dur a terme experiments en metabolòmica *untargeted* és freqüent l'ús de la cromatografia líquida acoblada a espectrometria de masses (LC-MS), ja que presenta un elevat rang dinàmic i cobertura del metabolisme (2). Altres plataformes com la cromatografia de gasos acoblada a espectrometria de masses (GC-MS), *imaging* MS (i-MS) i ressonància magnètica nuclear (NMR) són també amplament utilitzades en el camp, complementant els resultats obtinguts per LC-MS. Els instruments GC-MS són utilitzats per l'anàlisi de compostos volàtils, com ara terpens; i-MS s'ha fet servir per monitoritzar la distribució espacial dels metabòlits, permetent ubicar-los a nivell d'orgànuls; NMR, per altra banda, permet l'anàlisi de la composició interna de lipoproteïnes, no és destructiva i permet la identificació inequívoca d'estereoisòmers. Aquestes tres tècniques, però, presenten limitacions com la dificultat en l'anàlisi de compostos termolàbils en GC-MS, el solapament de senyals en i-MS i la baixa sensibilitat en NMR. A la Taula 1 s'exposa una comparativa de les diferents tècniques, ressaltant les avantatges i inconvenients de cadascuna.

La LC-MS es basa en la separació de les molècules presents en una solució (fase mòbil) en funció de la seva polaritat al passar per una columna (fase estacionària). En metabolòmica s'utilitzen dos tipus de columnes en funció de quin tipus de metabòlits es volen aïllar: columnes d'interacció hidrofílica (HILIC - *Hydrophylic interaction*) i columnes de fase reversa (RP – *Reverse phase*). Les primeres disposen de compostos amb grups hidrofílics ancorats a la columna, els quals interaccionen amb els compostos polars de la fase mòbil i els alenteixen, augmentant el seu temps de retenció (RT). Les columnes de fase reversa, en canvi, contenen hidrocarburs saturats de cadena mitjana-llarga (entre 8 i 18 carbonis) que interaccionen amb els compostos apolars, invertint l'ordre en què elueixen els compostos.

Taula 1: Comparació de les diferents tècniques analítiques emprades en metabolòmica.

Mètode	Avantatges	Inconvenients
LC-MS	• Gran cobertura de metabòlits. • Alt rang dinàmic.	• No detecta compostos volàtils o que no ionitzen bé en fase líquida.
GC-MS	• Separació eficient (alta resolució) • Cicles de treball curts. • Anàlisi de metabòlits volàtils.	• Requereix estabilitat tèrmica dels analits.
iMS	• Aporta informació sobre la ubicació espacial dels metabòlits.	• No hi ha separació cromatogràfica. • Ions amb m/z semblants no es poden distingir bé.
NMR	• Anàlisi no destructiu. • Permet identificació inequívoca d'un analit. • Quantificació de lipoproteïnes.	• Solapament de senyals. • Processat de dades poc reproduïble. • Baixa sensibilitat per a analits poc abundants.

L'elecció de la columna òptima per dur a terme un experiment depèn de si es busquen separar preferentment compostos polars (columna HILIC) o apolars (columna de fase reversa). A banda de la tria de columna, hi ha altres factors a tenir en compte per optimitzar la separació dels compostos com ara els gradients d'elució, que són variacions incrementals en la composició de la fase mòbil que serveixen per millorar la separació entre compostos i accelerar la cromatografia.

Els instruments de LC-MS presenten generalment dos modes d'operació: MS1 i MS2 (també anomenat MS/MS o tàndem MS) (3). L'anàlisi MS1 consisteix en:

(i) Generar ions a partir de l'eluït de la columna cromatogràfica.

(ii) Separar els ions en funció del seu quocient massa/càrrega (m/z).

(iii) Detectar els ions i enregistrar el resultat.

L'anàlisi MS2 afegeix un pas extra al mode MS1 i, després de generar els ions, en selecciona només un conjunt i els sotmet a un procés de fragmentació. Aquest pas de fragmentació genera una sèrie d'ions que són característics de cada molècula (m/z i abundància relativa) i que permeten la identificació dels compostos.

Els fluxos de treball en LC-MS *untargeted* consisteixen en l'adquisició de dades LC-MS1 de la totalitat de les mostres, que serviran per a la quantificació dels compostos, seguida de l'adquisició de dades LC-MS2 de només un conjunt de mostres representatives, que serviran per a la identificació dels compostos. Dintre d'aquests fluxos de treball, podem fer una distinció entre els mètodes *one-step* i els *two-step*, en funció de si les adquisicions LC-MS1 i LC-MS2 es duen a terme consecutivament (*one-step*) o si existeix un anàlisi previ de les dades LC-MS1 abans de procedir amb l'adquisició de dades LC-MS2 (*two-step*).

Quan es duen a terme anàlisis *one-step,* es fan servir generalment dues estratègies per l'adquisició MS2: *data-dependent acquisition* (DDA) i *data-independent acquisition* (DIA). L'anàlisi DDA (Figura 1a) consisteix en:

(i) Adquirir un *scan* MS1 de tots els ions generats.
(ii) Buscar els ions més abundants i afegir-los a una llista d'objectius.
(iii) Aïllar cada ió en la llista individualment i fragmentar-lo, adquirint un *scan* MS2 per cada ió.
(iv) Repetir els passos (i)-(iii) fins al final de l'experiment.

Les anàlisis DDA han evolucionat amb el temps per augmentar la quantitat i qualitat de la informació generada: actualment són freqüents les anàlisis de DDA iteratiu, en les quals s'analitza una mostra múltiples vegades i es genera una llista d'exclusió per evitar fragmentar els mateixos ions en injeccions posteriors (4). L'anàlisi DIA (Figura 1b), en canvi, fragmenta tots els ions independentment de la seva intensitat, obtenint un espectre MS2 representatiu del conjunt. Els espectres adquirits per DIA són sovint convolucionats, és a dir, que presenten senyals provinents de múltiples compostos; el problema de la deconvolució en DIA és particularment difícil de solucionar i, donada la seva baixa sensibilitat, és menys freqüent que l'anàlisi DDA en aplicacions de metabolòmica. Un dels desenvolupaments en l'adqusició DIA ha estat l'estratègia SWATH (5) (*Sequential Windowed Acquisition of All THeoretical spectra*), que, en comptes d'aïllar la totalitat dels ions, restringeix l'interval d'adquisició a un interval de m/z ample (entre 25 i 50 Da), reduint la complexitat dels espectres obtinguts i facilitant la identificació dels compostos de manera no dirigida.

Les anàlisis *two-step* (Figura 1c) són l'evolució conceptual dels *one-step*: en comptes de deixar que l'aparell seleccioni els ions precursors més intensos (DDA) o que els fragmenti tots alhora (DIA), les anàlisis *two-step* permeten seleccionar quins precursors fragmentar en base a criteris derivats de l'anàlisi MS1; per exemple, es poden triar només aquelles senyals que tinguin abundàncies significativament diferents entre conjunts de mostres. D'aquesta manera, l'anàlisi MS2 queda restringit a senyals que es considerin rellevants pels investigadors i es poden focalitzar els recursos instrumentals a adquirir dades de major qualitat.

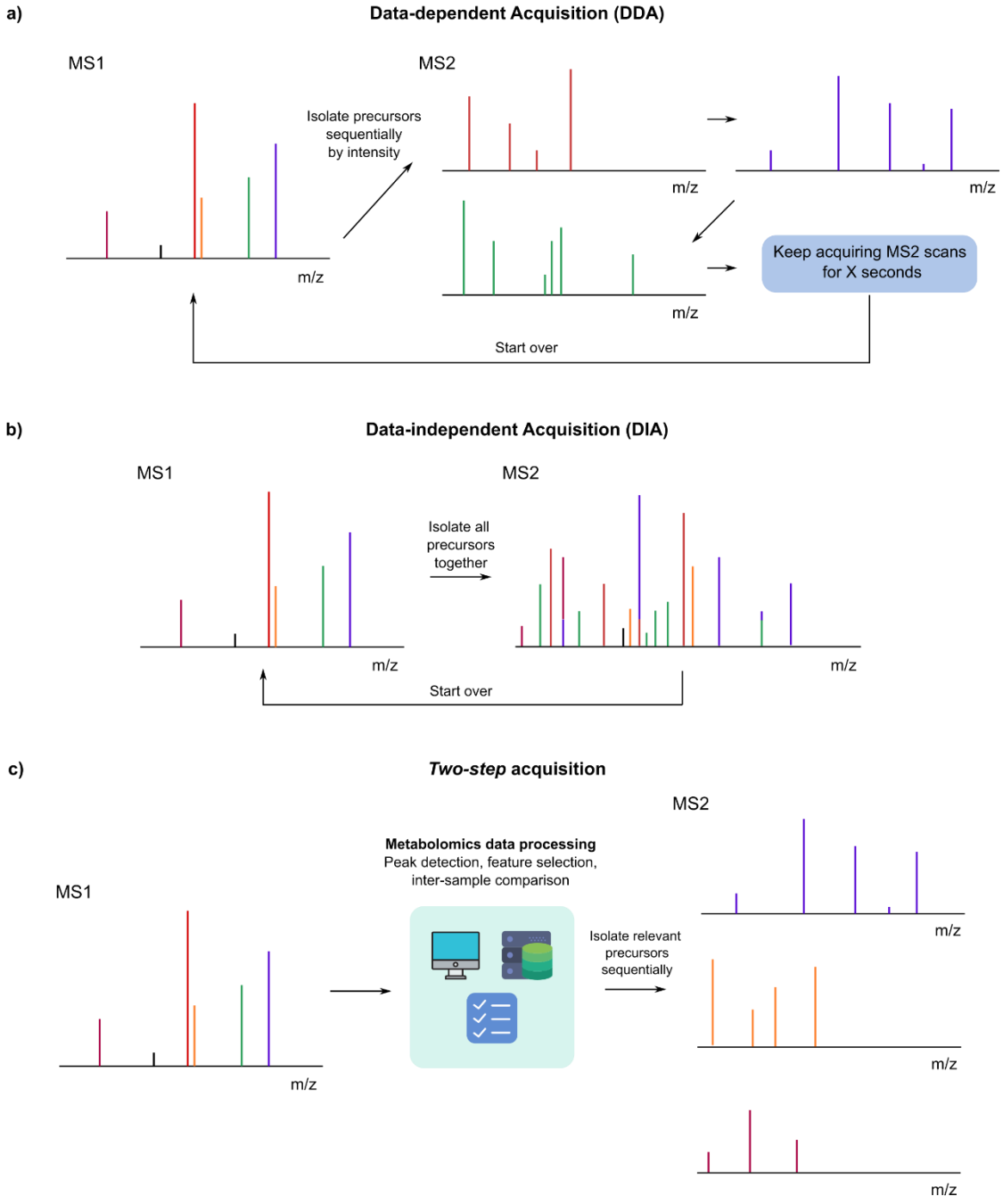

Figura 1: Comparació dels mètodes *data-dependent acquisition* (DDA) (a), *data-independent acquisition* (DIA) (b), i *two-step* (c). En el mètode DDA, després d'adquirir un scan MS1 s'aïllen els precursors més intensos de forma seqüencial i s'analitzen a nivell MS2. En el mètode DIA, en canvi, els precursors es fragmenten de forma conjunta, donant lloc a un espectre convolucionat on es mesclen els fragments de diferents ions precursors. En les aproximacions *two-step* hi ha un anàlisi de dades previ a l'adquisició MS2, de manera que només es seleccionen aquells ions precursors que siguin rellevants per a l'investigador.

Per tal de dur a terme l'anàlisi de dades de LC-MS1, existeixen diferents plataformes de software lliure disponibles. Una de les més emprades és XCMS (6,7), la qual permet la detecció de pics en múltiples mostres, l'agrupament d'aquests pics en *features* i l'anotació d'aquestes *features* basant-se en diferències de m/z pre-establertes.

Els experiments de marcatge isotòpic són extensament emprats en metabolòmica per estudiar el flux dels metabòlits en un sistema biològic, així com per identificar compostos i validar el seu origen biològic (2,8). El principal isòtop utilitzat és el ^{13}C, degut a la presència ubiqua del carboni en les biomolècules. Per denotar els nombre d'àtoms ^{13}C en una molècula s'utilitza la notació MX, on X és el nombre d'àtoms de ^{13}C presents. El ^{13}C és present a les molècules naturalment, establint un patró isotòpic que depèn del nombre d'àtoms de carboni en la molècula i que sovint s'empra per determinar experimentalment la formula molecular dels compostos. Altres isòtops estables que es poden observar en les biomolècules són el ^{15}N, ^{18}O, ^{2}H i el ^{34}S. En molècules que contenen halògens, com ara biomolècules generades per espècies marines, podem trobar els isòtops ^{37}Cl i ^{81}Br, els quals tenen una abundància natural molt elevada (24% i 49%, respectivament) i donen lloc a patrons isotòpics molt característics que permeten confirmar la seva presència.

A nivell pràctic, existeixen dos fenòmens complementaris que es poden aprofitar per quantificar el nivell de marcatge ^{13}C d'un metabòlit (Figura 2): (a) l'aparició de senyals marcades (Mx) que no poden ser explicades a partir de les abundàncies naturals i (b) la dilució isotòpica, que consisteix en la pèrdua de senyal del pic monoisotòpic (M0) en la mostra marcada en comparació a la mostra sense marcar. Per quantificar (a), Buescher *et al.* (9) defineixen la contribució fraccional (en anglès *fractional contribution*, FrC), que calcula la proporció d'àtoms marcats en el conjunt de senyals M0 fins a Mn, on n és el nombre total d'àtoms de carboni en la molècula:

$$FrC = \sum_{i=1}^{N} \frac{M_i \cdot i}{M0_{no\ marcat} \cdot n}$$

Per quantificar (b), Giné *et al.* (10) defineixen la puntuació del ràtio monoisotòpic (en anglès *monoisotopic ratio score*, MIRS), que és calcula el quocient entre la senyal M0 en la mostra marcada i en la no marcada:

$$MIRS = 1 - \frac{M0_{marcat}}{M0_{no\ marcat}}$$

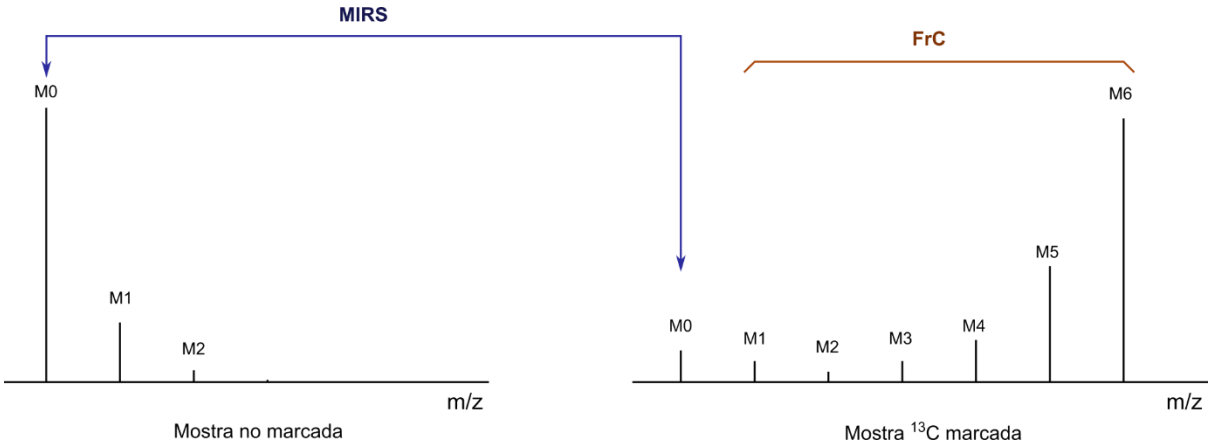

Figura 2: Representació del patró isotòpic natural de la glucosa ($C_6H_{12}O_6$) -esquerra- i del patró isòtopic del mateix compost que observem en una mostra on s'ha suplementat el medi de cultiu amb ^{13}C -dreta-. Es representen les dues mètriques de marcatge esmenades: contribució fraccional (FrC), basada en la presència de senyals marcades (M1-M6), i puntuació del ràtio monoisotòpic (MIRS), basada en l'absència de senyals no marcades (M0).

HERMES és una estratègia d'adquisició de dades de metabolòmica LC-MS1 i LC-MS2 creada per Giné *et al*. (10), que té com a objectiu caracteritzar el màxim possible una matriu biològica o ambiental. Per fer-ho, li dona la volta a les estratègies convencionals d'anàlisi de dades LC-MS: en comptes de dur a terme una detecció de pics i posterior anotació dels metabòlits, HERMES parteix de tota una llista de possibles ions que podrien estar presents a la mostra i anota les dades directament, sense necessitat de dur a terme cap procés de detecció de pics. Aquesta inversió en el paradigma del processat de les dades confereix a HERMES una alta sensibilitat (detecta compostos independentment del seu perfil d'elució) i especificitat (només detecta aquelles senyals que són desitjades pels investigadors). La possibilitat de triar quin conjunt de formules moleculars i adductes es volen buscar a les dades experimentals adquirides possibilita que es pugui fer servir HERMES en múltiples contexts, com són aplicacions biomèdiques, ambientals, alimentàries, etc.

El funcionament de l'estratègia HERMES (Figura 3) consisteix dels següents passos:

1) Generació d'una llista de les senyals que es pretén buscar a les dades (Figura 3a). Aquesta llista conté totes les combinacions de formules i adductes especificades per l'usuari (fórmules iòniques). Es duu a terme un procés de filtratge de les formules iòniques redundants, és a dir, aquelles que tenen un mateix nombre d'àtoms i la mateixa càrrega.
2) Càlcul dels patrons isotòpics adaptada a la resolució experimental. Per tal de poder validar les anotacions realitzades per HERMES, és sovint necessari comprovar que els perfils isotòpics observats quadrin amb els que teòricament se'n deriven de l'anotació. Per aquest motiu, el programa calcula quins isòtops podran ser distingits entre ells (i, per tant, detectats) donada la resolució de l'aparell utilitzat per adquirir les dades.

3) Anotació de les dades experimentals (Figura 3b). Per a cadascuna de les formules iòniques calculades en 1), es busquen quines senyals de les dades tenen un valor de massa/càrrega (m/z) que quadri amb el valor m/z teòric de la fórmula iònica que es busca dintre de l'error de massa de l'instrument. Si es troben senyals que coincideixen amb la senyal monoisotòpica, es busquen la resta d'isòtops calculats en 2).
4) Detecció de scans d'interès (SOI, Scans of Interest, Figura 3c). A partir de cadascuna de les anotacions obtingudes en 3), es busquen conjunts de senyals que es repeteixen en el temps (Annex 1 – Algorisme 1). Aquest pas substitueix el *peak-picking* que es fa servir en altres aproximacions.
5) Filtratge dels scans d'interès. Aquests conjunts de senyals són filtrats en els següents passos (Figura 3d):
    a. Subtracció del blanc experimental: es busca si existeixen senyals anàlogues a les detectades en un blanc experimental analitzat prèviament i s'eliminen. Una xarxa neuronal artificial (ANN, Artificial Neural Network) ha estat entrenada manualment per classificar si dos perfils de senyals (un de mostra, l'altra de blanc) són diferents entre ells o no.
    b. Fidelitat isotòpica: busca si les senyals d'isòtops associades a les SOIs tenen un patró d'intensitats compatible amb el patró d'intensitats teòric, el qual es pot calcular a partir de la fórmula molecular de l'anotació. L'algorisme té en compte la intensitat de les senyals i la possibilitat que determinats isòtops tinguin una intensitat per sota del llindar de detecció de l'instrument.
    c. Filtratge de fragments generats a la font: busca si la senyal anotada en cada SOI podria provenir de la fragmentació d'una altra SOI, el qual és un fenomen comú durant la ionització de metabòlits làbils a la font de ionització de l'instrument. A partir d'una base de dades fragmentació MS2 a baixes energies, HERMES s'inspira en Domingo *et al*. (11) per determinar si hi ha senyals compatibles amb fragments generats a la font i busca si existeixen similituds en els perfils d'elució entre SOIs per filtrar-les.
    d. Heurístiques de qualitat de senyal: es calculen diversos paràmetres de les SOIs, com són la intensitat màxima de les senyals, el rang entre la intensitat mínima i la màxima, la quantitat d'estructura (manca de soroll) present a les dades. L'aplicació de diferents filtres basats en aquests paràmetres permeten eliminar senyals sorolloses.
6) Generació d'una llista d'inclusió per a realitzar un experiment MS2. A partir d'una llista de SOIs, HERMES les agrupa en funció de la seva m/z per donar lloc a una llista d'inclusió. Aquesta llista té en compte la presència de SOIs amb valors de m/z molt semblants entre elles i que coelueixen en temps de retenció propers. Tot sovint, donat

que hi ha moltes entrades a monitoritzar de forma simultània, HERMES distribueix les entrades de la llista d'inclusió en múltiples injeccions que haurà de dur a terme l'instrument per separat.

7) Adquisició de les dades MS2. La llista d'inclusió s'importa al programari de l'espectròmetre de masses per crear un mètode instrumental únic per a cada injecció. L'instrument adquireix contínuament *scans* MS2 de cadascuna de les entrades (Figura 3e). Aquesta cobertura contínua de les entrades permet a HERMES dur a terme un procés de deconvolució de senyals que separa els fragments provinents de compostos que coelueixen (Figura 3f) i genera una llista d'espectres de fragmentació.
8) Identificació dels compostos. A partir dels espectres de fragmentació generats, es poden identificar els compostos a partir d'una comparació d'espectres amb una base de dades d'espectres de referència com ara METLIN (12) i MassBank (13). Alternativament, existeixen altres aproximacions *in silico* que permeten la identificació *de novo* de compostos pels quals no hi ha espectres de referència com ara CSI-FingerID (14), o permeten establir relacions de similitud estructural entre espectres MS2, com ara GNPS (15).

L'estratègia HERMES és integral a l'anàlisi d'una mostra biològica, és a dir, cobreix tant la generació de les dades experimentals com el seu anàlisi, fins arribar a una llista d'identificacions i quantificacions. Per poder dur a terme totes aquestes tasques d'una forma eficaç, organitzada i reproduible, he programat una implementació de HERMES en el llenguatge de programació R: RHermes.

El programa RHermes ha estat dissenyat amb la practicitat en ment: tots els mòduls i passos del processat poden executar-se mitançant comandaments de text senzills o fent servir una interfície gràfica (GUI), que permet analitzar i visualitzar les dades de forma dinàmica des d'un navegador web. Gràcies a aquesta doble implementació en text i GUI, usuaris de diferents camps de recerca i amb diversitat de coneixements de programació poden fer servir l'eina sense problema.

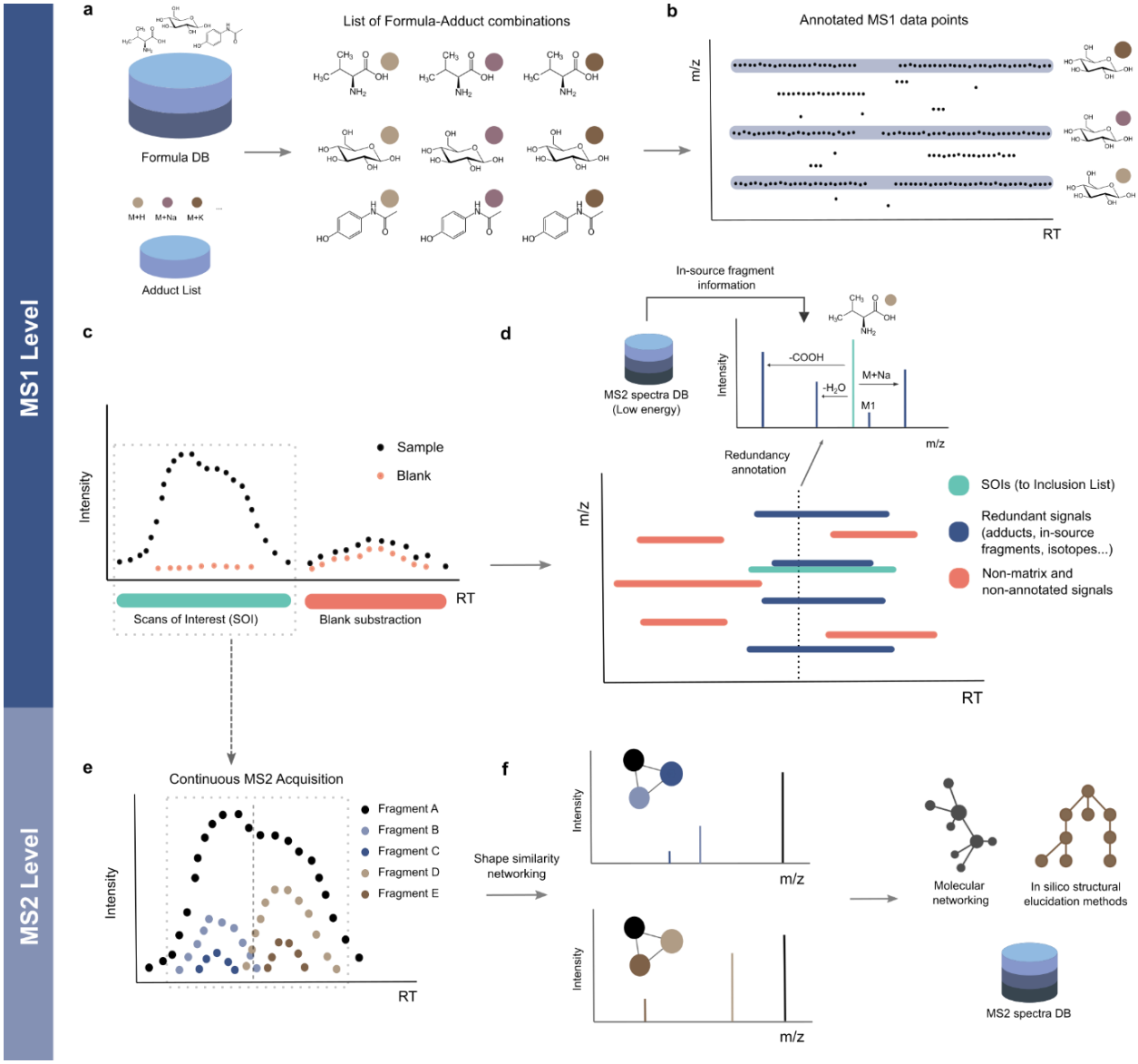

Figura 3: Descripció de l'estratègia HERMES per la caracterització del metaboloma. a) A partir d'una base de dades de formules moleculars i adductes, es calculen tots els possibles ions obtinguts a partir de les combinacions formula-adducte. b) Es busquen les m/z dels ions a les dades experimentals MS1 i s'anoten els scans. c) S'agrupen els conjunts de scans amb una mateixa anotació al llarg del temps. S'aplica un algorisme de substracció de blanc per eliminar senyals que no són pròpies de la mostra i que també aparèixen en un blanc analític. d) Es filtren les senyals redundants i mal anotades fent servir diferents filtres (fidelitat isotòpica, heurístiques de senyal, etc.). S'obté una llista d'inclusió per fer dur a terme un experiment MS2. e) S'adquereixen les dades MS2 de forma contínua, obtenint perfils d'elució per als diferents fragments obtinguts a partir d'un mateix ió precursor. f) Els fragments s'agrupen en funció de la seva similitud d'elució i donen lloc a una llista d'espectres de fragmentació, que poden ser utilitzats per identificar els metabòlits contra llibreries d'espectres de referència o emprats en altres estratègies computacionals. Figura adaptada de Giné *et al*. (10)

4. Hipòtesi de treball i objectiu/s

El software RHermes permet una millor identificació dels metabòlits presents en un mostra biològica en comparació amb tècniques convencionals en estudis de la metabolòmica com la *data-dependent acquisition* (DDA).

El present treball es proposa els següents objectius:

- Demostrar l'especificitat de l'estratègia d'identificació Hermes envers els metabòlits presents en una mostra de *Escherichia coli.*
- Validar els resultats obtinguts mitjançant una mostra de cultiu marcada completament amb ^{13}C.
- Comprovar la rellevància del temps d'injecció en LC-MS2 la qualitat dels espectres de fragmentació obtinguts.

5. Metodologia

5.1 Tractament de les mostres d'*Escherichia coli*[1]

Els extractes secs de *E. coli* (sense marcar i ^{13}C-marcat uniformement) es van reconsituir en 100 µL d'acetonitril-aigua (2:1), seguit de 30s de vortexat, 5 min de sonicació i 30s de vortexat, per tal d'alliberar els metabòlits homogèniament a la preparació.

L'anàlisi LC-MS es va realitzar en un sistema Thermo Scientific Vanquish Horizon UHPLC acoblat a un espectròmetre de masses Thermo Scientific ID-X Tribrid (Waltham, MA). L'anàlisi per cromatografia líquida d'interacció hidrofílica (HILIC) es va dur a terme fent servir una columna SeQuant ZIC-pHILIC (Merck Millipore, Burlington, MA) amb especificacions 150 mm x 2.1 mm, 5 µm. Els dissolvents de la fase mòbil estaven formats per A = 20mM bicarbonat d'amoni, 0.1% solució d'hidròxid d'amoni (25% d'amoníac en aigua) i 2.5 µM d'àcid medrònic en aigua:acetonitril (95:5) i B = 95% acetonitril, 5% aigua, 2.5 µM d'àcid medrònic. El compartiment de la columna es va mantindre termostatitzat a 40ºC durant els experiments. Es va aplicar el següent gradient amb un cabal de 250 µL/min: 0-1 min: 90% B, 1-12 min: 90-35% B, 12.5-14.5 min 25% B, 15 min 90% B seguit de 4 min de re-equilibrat amb un cabal de 400 µL/min i 2 min a 250 µL/min. El volum d'injecció va ser de 2 µL per a tots els experiments. Les dades es van adquirir amb els següents paràmetres: voltatge del esprai, 3.5 kV i -2.8kV en mode positiu i negatiu, respectivament; gas portador, 50; gas auxiliar, 10; gas oposat, 1; temperatura del tub de transferència iònica 300ºC; temperatura del vaporitzador, 200ºC; rang de massa 70-1000 Da; lents RF 60%; resolució, 120000 (MS1), 15000 (MS2); AGC target 200000 (MS1), 5e4 (MS2); màxim temps d'injecció, 200ms (MS1), 35ms (MS2 HERMES), 100ms (DDA iteratiu); finestra d'aïllament 1Da; energia de col·lisió 35% HCD.

[1] Adaptat íntegrament de Giné *et al.* (10)

5.2 Anàlisi de les dades amb RHermes

Totes les anàlisis descrites (excepte on s'indiqui) es van realitzar amb RHermes v0.99.0, executat en RStudio v1.4.1106 i el llenguatge de programació R v4.0.4.

Bases de dades de fórmules i adductes: es va dur a terme una fusió entre la base de dades *E. coli* Metabolome DataBase (ECMDB) i tots els metabòlits presents en els les rutes metabòliques associades a *E. coli* K12 en la Kyoto Encyclopedia of Genes and Genomes (KEGG). Els adductes utilitzats van ser, per a ionització positiva, $[M+H]^+$, $[M+Na]^+$, $[M+K]^+$, $[M+NH_4]^+$ i $[M]^+$, i, per a ionització negativa, $[M-H]^-$ i $[M+Cl]^-$.

Detecció de les SOI: es van fer servir dos filtres de densitat de scans de scans de 5s, intercalats amb 2.5s d'espai entre cadascun. Es va demanar un mínim valor d'integritat estructural en les SOI (ρ_{chaos}) de 0.5.

Filtratge de les SOI: Es va fer servir un filtre de fidelitat isotòpica demanant una similitud mínima de 0.8 entre el patró isotòpic teòric i l'observat experimentalment. La detecció de fragments generats a la font (*in-source fragments*) es va dur a terme fent servir una base de dades d'espectres de fragmentació d'ús intern, formada per espectres de les llibreries MassBankEU, MoNA, HMBD, Riken i NIST14. Es van seleccionar els espectres amb energies de col·lisió menors a 20 eV (ionització CID) o 20% nominal (ionització HCD).

Generació de la llista d'inclusió: es van generar dues llistes d'inclusió aplicant prioritització dels adductes $[M+H]^+$ i $[M+NH_4]^+$ (ionització positiva) i $[M-H]^-$ (ionització negativa). Els elements de les llistes d'inclusió es van organitzar en injeccions permetent un solapament de fins a 10 entrades a monitoritzar de forma simultània per l'instrument.

5.3 Obtenció dels *scans* per adquisició *data-dependent* (DDA) iteratiu

Es va dur a terme un total de 3 injeccions mitjançant un mètode amb llistes d'exclusió, per tal d'evitar fragmentar els ions que ja s'hagin analitzat en injeccions anteriors. Per fer-ho, es va convertir el primer fitxer DDA al format mzML fent servir MSConvert (16). A continuació, es va processar el fitxer mitjançant el programa IEomics (4) per obtenir la llista d'exclusió per a la següent injecció. Els paràmetres de processat utilitzats en el script en R van ser: RTWindow = 0.3 min, noiseCount = 25 i MZWindow = 0.001. La tolerància de massa considerada per a la generació de les llistes d'exclusió va ser de 5 ppm.

5.4 Identificació dels espectres MS2

Per dur a terme el *matching* dels espectres MS2 processats per HERMES es van ajustar els valors de m/z en intervals de 0.01 Da i es van normalitzar les intensitats dividint-les entre la suma d'aquestes. El mateix procés de normalització es va dur a terme amb els espectres de referència provinents de la base de dades d'espectres de fragmentació esmenada en l'apartat

anterior. La mètrica de similitud espectral emprada va ser el cosinus, calculada amb el paquet de R *philentropy*:

$$\cos(X,Y) = \frac{X \cdot Y}{|X| \cdot |Y|} = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2} \cdot \sqrt{\sum_i y_i^2}}$$

Altres mètriques com Topsöe, *fidelity* i *squared-chord* (Annex B) van ser també calculades amb *philentropy* i utilitzades per la validació dels resultats (Annex C – Figura S1).

5.5 Càlcul de les mètriques de marcatge isotòpic (FrC i MIRS)

Per a calcular la quantitat de carboni-13 (^{13}C) que s'ha incorporat en cada molècula estudiada es va implementar el càlcul de la contribució fraccional (FrC) i la puntuació del ràtio monoisotòpic (MIRS), les quals han estat definides a la introducció.

Donat que no hi ha cap software ni llibreria de programari en què aquest càlcul es trobi automatitzat en un format compatible amb HERMES, es va programar un script en el llenguatge de programació R que duu a terme els següents passos:

1. Aplicar l'annotació d'HERMES a les dades marcades, tenint en compte les senyals de tots els isòtops de carboni de cada molècula (des de M0, fins a Mn, on n és el nombre d'àtoms de carboni).
2. Fer servir la llista de SOIs de la mostra no marcada com a referència per a l'anàlisi.
3. Per a cada SOI (definida per una anotació i un interval de temps de retenció - RT):
    a. Buscar tots els punts que tinguin l'anotació de la SOI i que es trobin en l'interval de RT corresponent.
    b. Per a cada isòtop, buscar el *scan* amb la màxima intensitat (àpex).
    c. Calcular el FrC i el MIRS d'acord amb la seva definició.

5.6 Accessibilitat del codi i de les dades experimentals

El programari RHermes està disponible lliurement en un repositori de Github (www.github.com/RogerGinBer/RHermes) i inclou una documentació detallada de tota la seva funcionalitat, una guia d'usuari i exemples reals d'anàlisi de dades. RHermes es troba protegit per una llicència de software GPLv3, que permet el seu ús per a finalitats no comercials.

Les dades crues utilitzades en aquest estudi, així com els *scripts* que han donat lloc als resultats i les figures mostrades, es troben accessibles al repositori de Zenodo amb ID 4581662 (https://zenodo.org/record/4581662). Les dades es troben protegides per una llicència Creative Commons BY-4.0, de manera que poden ser utiltizades lliurement sempre que siguin degudament atribuïdes.

6. Resultats, discussió i relació amb els objectius plantejats

6.1 Resultats del processat de la mostra no marcada

Es van analitzar els dues rèpliques de la mostra de *E. coli* no marcada fent servir 12010 (ionització positiva) i 4876 (ionització negativa) fórmules iòniques derivades de 2463 formules moleculars extretes de la base de dades *Escherichia coli* Metabolome Database (ECMDB) i de Kyoto Encyclopedia of Genes and Genomes (KEGG). Per obtenir aquesta base de dades conjunta, es va dur a terme una integració de les dues, eliminant les entrades redundants. D'aquesta manera, ens vam assegurar que tot metabòlit reportat anteriorment en E. coli estigués cobert per la nostra base de dades i que, per tant, fos detectable per HERMES.

Les dues rèpliques analitzades van presentar una alta reproductibilitat, fet que es va comprovar a l'aplicar una subtracció de blanc d'una mostra respecte l'altra. Es va obtenir una llista amb únicament 52 i 21 SOIs (positiu i negatiu), la majoria de les quals degudes a petites derives de temps de retenció i problemes amb l'algorisme de subtracció de blanc. Validada la reproductibilitat del mètode analític, vaig centrar-me en una única rèplica. HERMES va obtenir un total de 2058 i 1081 SOIs en mode d'ionització positiu i negatiu, respectivament. Aquestes entrades van donar lloc a dues llistes d'inclusió per al seu anàlisi MS2 de 1251 i 661 entrades. La disparitat entre el nombre de SOIs i entrades a la llista d'inclusió en positiu i en negatiu és, en part, explicat pel fet que els espectres de ionització positiva són més rics en senyals redundants (adductes) que no els espectres adquirits en negatiu. A banda, el fet de considerar un major nombre de formules iòniques facilita la possibilitat de tenir falsos positius (senyals mal anotades degut a la seva proximitat a una m/z present a la base de dades), possiblement engrossint les llistes de SOIs.

En quant a l'anàlisi LC-MS2, es va analitzar la mateixa mostra no marcada de *E. coli* fent servir la llista d'inclusió d'HERMES i, com a control, es va analitzar també fent servir una anàlisi de *data-dependent acquisition* (DDA). Comparant els ions precursors aïllats pels dos mètodes (Figura 4a), s'observen clares diferències qualitatives en quant la selecció dels ions precursors. Comparant la m/z dels ions precursors a la base de dades de formules iòniques feta servir per HERMES, trobem que un 67.9% dels *scans* adquirits per DDA (Figura 4b) no poden ser associats a cap ió del metaboloma de *E. coli*. De la resta de *scans* que sí coincidien amb algun ió, vora la meitat no es trobaven associats a la llista d'inclusió i es podien associar a senyals també presents en el blanc experimental. Només un 16.4% dels *scans* totals adquirits per DDA (1 de cada 6) podien ser correctament associats a la llista d'inclusió obtinguda amb HERMES. D'altra banda, 591 entrades (47.2%) de la llista d'inclusió han estat monitoritzades únicament per HERMES.

És destacable en la Figura 4a l'existència d'una sèrie de franges verticals que han estat captades per DDA però que no corresponen a cap metabòlit de la base de dades. Crec que

és molt probable que siguin compostos pesants adherits a la columna que no provenen de la mostra i què han eluït en els punts en què s'ha modificat el gradient de la cromatografia.

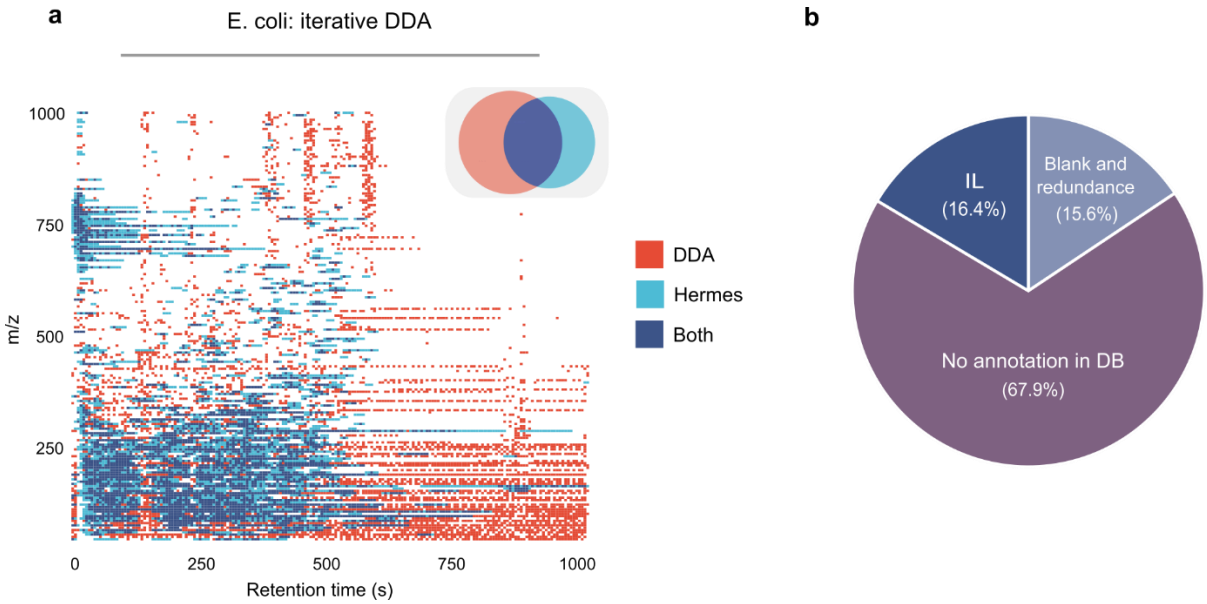

Figura 4: a) Distribució de les masses precursores dels scans MS2 adquirits per HERMES i per DDA. Els scans adquirits pels dos mètodes van ser agrupats en intervals de 5Da - 10s. El diagrama de Venn representa el solapament entre els dos mètodes a partir del nombre d'intervals omplerts per cadascun. **b)** Dintre dels scans adquirits per DDA, només un 16.4% d'aquests es troba associat a alguna entrada de la llista d'inclusió (*inclusion list*, IL) de HERMES. La resta o bé no van poder ser associats a cap anotació formula-adducte de la base de dades ECMDB+KEGG, o bé estaven associats a senyals del blanc. Figura adaptada de Giné *et al.* (10)

6.2 Resultats de la mostra ^{13}C-marcada

Per demostrar que HERMES detecta de forma específica compostos d'origen biològic, es va analitzar l'extracte de *E. coli* marcat en medi ^{13}C-glucosa en les mateixes condicions experimentals que la mostra no marcada. Per a cadascun dels ions de la mostra no marcada, es va calcular la contribució fraccional (FrC) i la puntuació del ràtio monoisotòpic (MIRS).

En primer lloc volia validar l'efectivitat de la subtracció de blanc per tal d'eliminar senyals inespecífiques de la mostra. Per fer-ho, es va calcular el marcatge amb les SOIs obtingudes sense fer eliminar el blanc (Figura 5a) i amb les SOIs obtingudes després d'eliminar-lo (Figura 5b). S'observa clarament com hi ha una reversió del patró de marcatge: més del 80% de les senyals que es consideren totalment no marcades (FrC i MIRS < 0.5) en les SOIs obtingudes sense eliminar el blanc desapareixen en aquest procés. En canvi, aquesta xifra és de només un 10% per a les senyals totalment marcades (FrC i MIRS > 0.5). Per tant, queda demostrat que el procés de subtracció del blanc és molt efectiu per tal d'obtenir una llista d'entrades específiques de la matriu biològica que s'estudia.

En quant a la generació de la llista d'inclusió, un 63% de les entrades de HERMES estan associades a metabòlits marcats, apuntant a què han estat sintetitzats en el propi

microorganisme (Figura 5c); En comparació, només un 20% dels *scans* adquirits per DDA s'associen amb metabòlits marcats, apuntant a què la resta de *scans* es troben associats a senyals provinents del blanc.

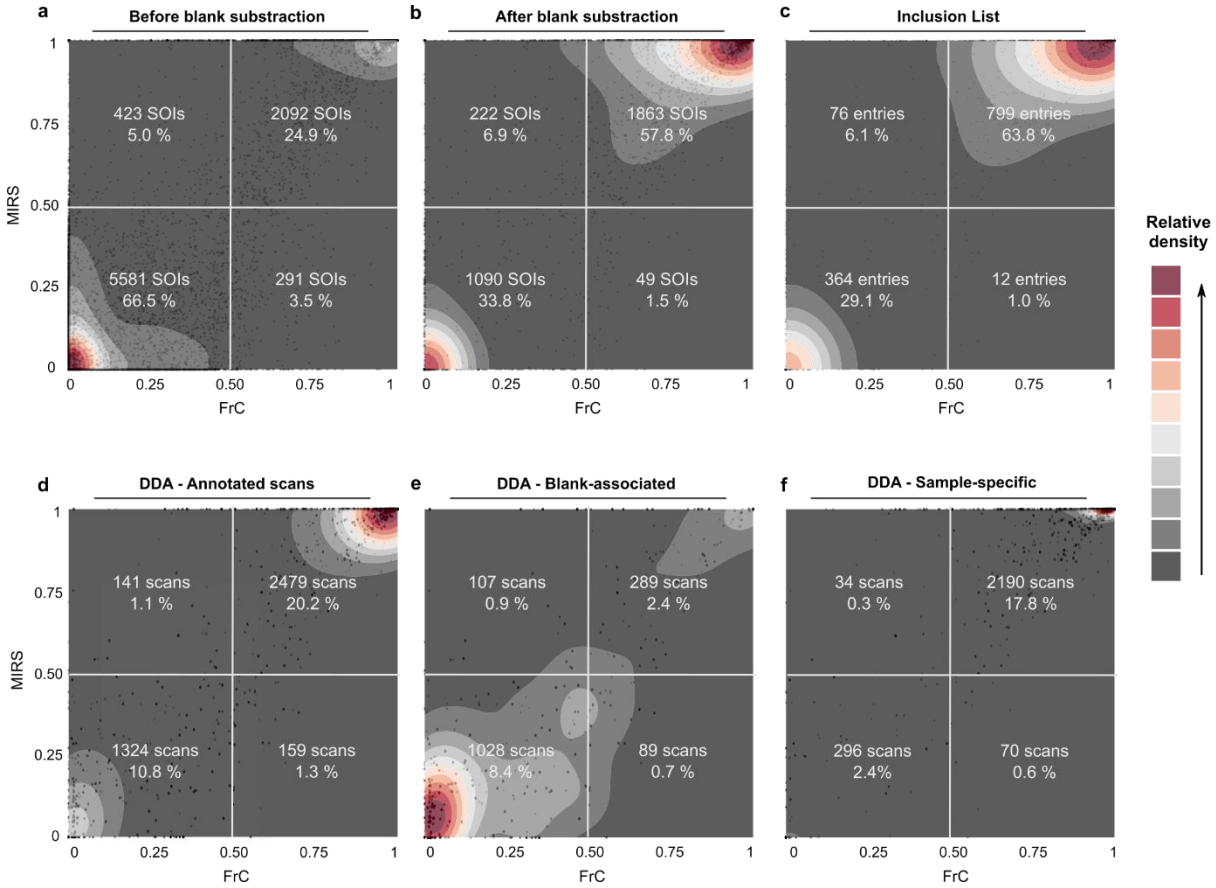

Figura 5: Representació de l'enriquiment isotòpic de les diferents SOIs definides per HERMES i dels scans adquirits per DDA. Es representa la contribució fraccional (FrC) i la puntuació de ràtio monoisotòpic (MIRS), així com el nombre d'entrades en cadascun dels quadrants. Es considera que una SOI / scan està marcat quan FrC i MIRS són superiors a 0.5. a) SOIs abans d'eliminar les senyals provinents del blanc, b) SOIs després d'eliminar el blanc, c) SOIs filtrades per fidelitat isotòpica (veure mètodes) i preparades per ser adquirides a nivell MS2, d) conjunt dels scans de DDA adquirits, e) subconjunt d'aquells scans que s'associen amb senyals provinents del blanc, f) subconjunt dels scans de DDA que s'associen amb senyals pròpies de la mostra. Figura adaptada de Giné *et al*. (10)

S'ha constatat que la majoria (>80%) de les entrades d'alta intensitat trobades per HERMES corresponen a metabòlits marcats amb ^{13}C, mentre que en entrades de menor intensitat hi havia major presència de compostos no marcats (35.6%, Figura 6-a i 6-b). Els compostos marcats que DDA va trobar són majoritàriament d'altes intensitats. És possible que algunes entrades de baixa intensitat que no presenten marcatge isotòpic observable per HERMES sí hagin estat marcades, però donat que es troben al límit de detecció de l'aparell, les senyals no es veuen i es confonen amb el soroll instrumental. Seria necessari fer un estudi dirigit MS1 cap a aquestes senyals per tal d'esbrinar definitivament si es troben marcades.

L'elevada especificitat biològica de les entrades d'HERMES es veu reflectida en un millor perfil de similituds espectrals contra les bases de dades d'espectres de referència, comparat

amb els resultats de DDA. Aquesta major similitud espectral s'ha validat fent servir mètriques de comparació d'espectres alternatives (Annex 2 i Annex 3 – Figura S1). En conjunt, HERMES ha aconseguit identificar quasi el doble de metabòlits respecte DDA, si bé a altes intensitats les diferències entre els dos mètodes es redueixen.

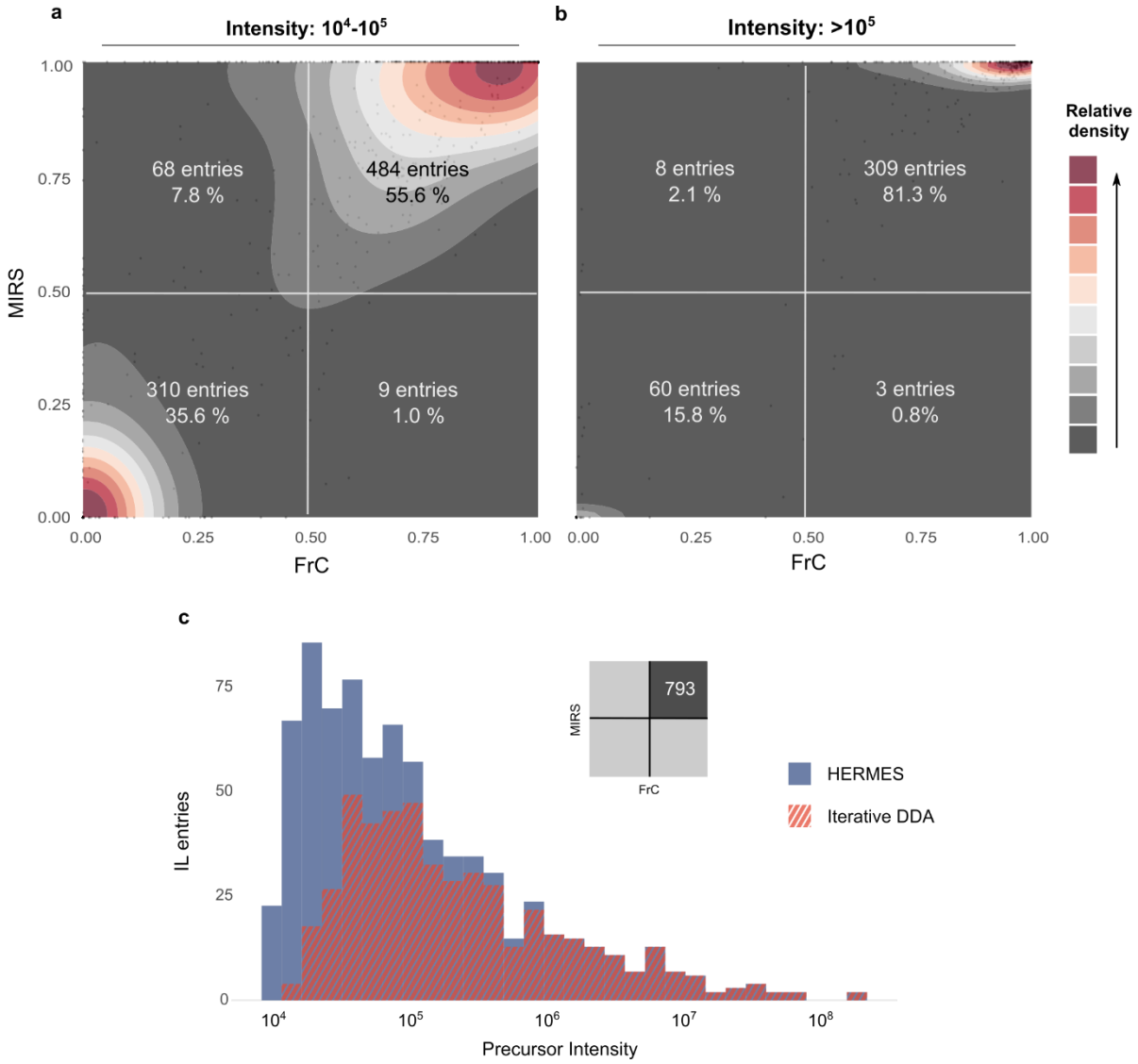

Figura 6: Distribució del marcatge isotòpic (FrC i MIRS) de les diferents entrades de la llista d'inclusió en funció de la seva intensitat. a) Baixa intensitat (10^4-10^5 ions); b) Alta intensitat (>10^5 ions); c) Representació de la cobertura d'aquelles entrades que es troben marcades (quadrant superior dret, representant FrC i MIRS > 0.5), per part de HERMES i de DDA. S'observa com un nombre significatiu d'entrades marcades a baixes intensitats no han estat monitoritzades per DDA. Figura adaptada de Giné *et al.* (10)

6.3 Efecte del temps d'injecció en la qualitat espectral MS2

Un dels principals problemes trobats en l'experiment va ser la manca de qualitat espectral en senyals poc intenses (< 10^5 ions) però que estaven marcades (FrC i MIRS > 0.5). En aquests casos, els espectres de fragmentació obtinguts fent servir un temps d'injecció de 35 ms no eren adequats per poder comparar-los contra les bases de dades degut a la manca de senyals consistents i la presència de soroll instrumental. Per solucionar el problema, es va seleccionar

una llista reduïda de senyals corresponents a metabòlits coneguts i poc abundants en la mostra no marcada (ex. NADH, NADPH, Uridina, etc.) i es van re-adquirir els seus espectres de fragmentació augmentant el temps d'injecció a 1500 ms.

Tal com s'observa en la figura 7, l'increment en el temps d'injecció enriqueix els espectres resultants i facilita la seva identificació contra bases de dades de referència.

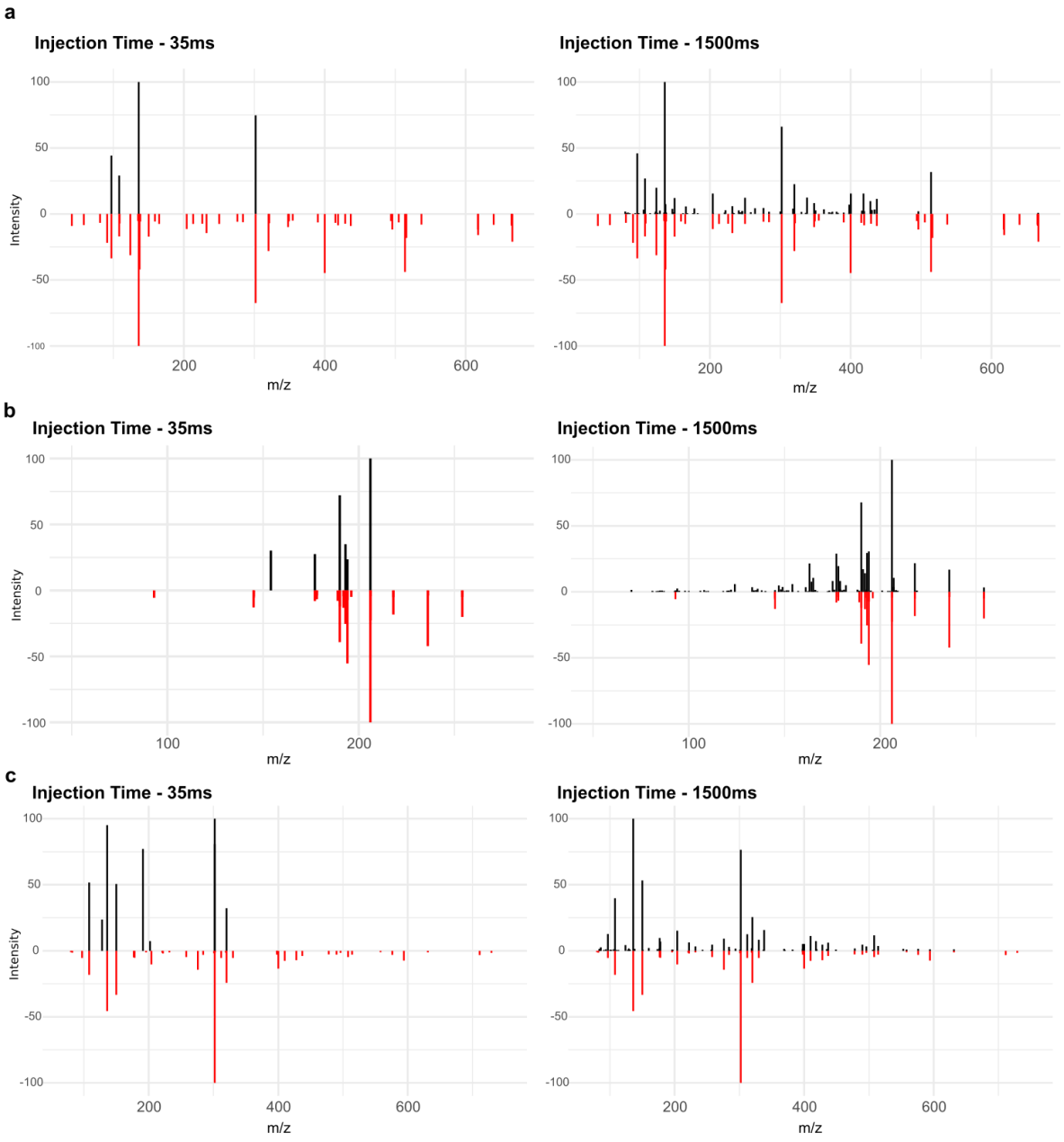

Figura 7: Efecte del temps d'injecció en la qualitat dels espectres MS2 obtinguts a partir de metabòlits de baixa intensitat (<10^5 ions). a) NADH, b) Uridina, c) NADPH. Els espectres en negre representen les dades adquirides experimentalment, mentre que els espectres en vermell representen l'espectre MS2 de referència adquirit a partir d'un estàndard pur del compost. Figura adaptada de Giné *et al.*(10)

La millora de la qualitat espectral MS2 observada a partir de l'augment del temps d'injecció és una troballa que, malgrat evident en els seus fonaments, és molt important per millorar les

taxes d'identificació de metabòlits. Actualment, els instruments operant sota un mètode DDA solen fer servir temps d'injecció propers a 100-200 ms. L'ús d'un temps d'injecció de 1500 ms de forma rutinària en un anàlisi no ha estat reportat en metabolòmica, si bé estudis molt recents en proteòmica comencen a comentar l'efecte del temps d'injecció en els ràtios d'identificació d'espectres de fragmentació de pèptids (17).

6.4 Discussió

L'anàlisi de dades de cromatografia líquida a nivell MS1 ha estat convenient per a la realització d'estudis amb grans nombres de mostres (>100), donat que cada mostra pot analitzar-se en un temps curt (<30 min) i els costos d'anàlisi són relativament baixos. Malgrat que la quantificació dels metabòlits sigui relativa en molts casos (donada la presència d'efecte matriu en moltes mostres i el cost prohibitiu dels estàndards purs de metabòlits), això no ha impedit que es facin servir quantificacions de *features* al llarg de les diferents mostres i s'apliquin tècniques d'estadística univariant i multivariant per trobar diferències significatives entre grups, candidats a biomarcadors, etc. Els resultats obtinguts a nivell MS2 amb HERMES suggerèixen que un nombre considerable d'entrades de la llista d'inclusió es troben composades de múltiples isòmers que coelueixen donant un perfil d'elució convolucionat. Aquests perfils no poden ser deconvolucionats pels programaris d'anàlisi de dades existents (7) i són erròniament agrupats en una única *feature*. Existeix la possibilitat que en molts estudis en què només es treballi a nivell de *features*, com és el cas d'eines en línia d'anàlisi de dades de LC-MS1 (18), s'estiguin integrant les intensitats de múltiples metabòlits com un de sol.

Una de les principals limitacions de l'estudi i de l'estratègia HERMES és la possibilitat de no considerar fórmules de metabòlits presents en *E. coli* que no hagin estat reportades a les bases de dades. Per solucionar el problema hem fet servir una conjunció de diferents bases de dades per evitar excloure metabòlits rellevants de la nostra anàlisi. A banda, s'estima que la possibilitat de trobar metabòlits amb fórmules no reportades en organismes tant caracteritzats com *E. coli* és baixa (19).

La valoració del marcatge isotòpic és una altra limitació de l'estudi, ja que és difícil mesurar els patrons isotòpics en metabòlits poc abundants a la mostra o que presenten dificultats per a ionitzar. Hipotetitzem que una fracció notable dels metabòlits que hem considerat com a no marcats (FrC i/o MIRS < 0.5) són en realitat compostos d'origen biològic en els que no s'ha pogut valorar correctament el marcatge degut a la seva baixa intensitat. Una solució per poder validar aquesta hipòtesi seria fer un estudi posterior de tipus targeted-MS1 (també conegut com *targeted selected ion monitoring* o t-SIM), en el qual només s'aïllen els ions provinents del compost d'interès durant un temps major per tal d'augmentar la sensibilitat de l'instrument.

D'aquesta manera, podríem estudiar els patrons isotòpics de totes les senyals amb independència de la seva intensitat. L'únic problema que presenta aquesta estratègia és la dificultat que té l'instrument per monitoritzar múltiples senyals de forma simultània, de manera que es requeriria un alt nombre d'injeccions i temps d'anàlisi per poder cobrir totes les senyals.

7. Conclusions

Ajudant-nos de les eines de marcatge isotòpic, hem comprovat que en el procés de substracció de blanc s'eliminen una gran quantitat de senyals que no estan marcades i que, per tant, no provenen del metabolisme de *E. coli*. L'alt percentatge d'entrades marcades a la llista d'inclusió (>63% d'entrades) ens demostra que HERMES és altament específic per detectar compostos d'origen biològic en la mostra. La tècnica DDA, considerada l'estat de l'art per a la identificació de metabòlits en experiments de metabolòmica LC-MS2, només cobreix metabòlits marcats en un 20% dels scans adquirits. Això implica que el 80% restant de temps de màquina està sent invertit en fragmentar ions que, malgrat ser intensos — ja que aquest és el criteri per adquirir senyals de DDA — no són biològicament rellevants sinó que provenen de contaminants del blanc experimental o són artefactes.

Els resultats es troben d'acord amb el què altres experiments han suggerit (19,20): una gran quantitat de les senyals detectades en els experiments de LC-MS1 no es troben associades a metabòlits d'interès, sinó a senyals exògenes. Aquest resultat és particularment rellevant per la comunitat metabolòmica, ja que molts estudis reporten *features* a nivell MS1 tractant-les com si fóssin metabòlits, sense mirar d'associar-les amb una formula molecular o sense dur a terme un *credentialing* (tal com suggereixen alguns autors (9,19,21)), la qual consisteix en analitzar una mostra que contingui una mescla de proporció coneguda marcada amb ^{13}C i sense marcar per determinar quins metabòlits són d'origen biològic.

Considerem, doncs, que s'han assolit tots els objectius proposats: s'ha vist clarament com HERMES pot trobar de forma específica els metabòlits marcats en la mostra d'*E. coli*, obtenint vora el triple de senyals marcades respecte a la tècnica de l'estat de l'art, DDA. A banda, s'ha pogut observar com es produeix una millora en la qualitat dels espectres de fragmentació MS2 a l'augmentar el temps d'injecció dels ions, la qual cosa ha permès identificar metabòlits que es troben en el límit de detecció de l'instrument.

HERMES és una estratègia prometedora per a la caracterització de mostres biològiques i ambientals, amb una alta versatilitat per detectar metabòlits a partir de qualsevol base de dades de formules moleculars. En el futur, és necessari seguir provant l'estratègia amb nous tipus de mostres, com ara extractes de plantes, aliments i begudes fermentades, on HERMES permetria tant la caracterització i quantificació de compostos amb impacte organolèptic així com la detecció de compostos no desitjats (pesticides, micotoxines...).

8. Bibliografia

1. Schrimpe-Rutledge AC, Codreanu SG, Sherrod SD, McLean JA. Untargeted Metabolomics Strategies—Challenges and Emerging Directions. Journal of the American Society for Mass Spectrometry. 2016 Dec 1; 27(12):1897–905.
2. Siuzdak G. The Expanding Role of Mass Spectrometry in Biotechnology. MCC Press; 2006.
3. Hoffmann E de, Stroobant V. Mass Spectrometry. Edition, T. Chichester: John Wiley & Sons, Ltd; 2012.
4. Koelmel JP, Kroeger NM, Gill EL, Ulmer CZ, Bowden JA, Patterson RE, et al. Expanding Lipidome Coverage Using LC-MS/MS Data-Dependent Acquisition with Automated Exclusion List Generation. J Am Soc Mass Spectrom. 2017;
5. Gillet LC, Navarro P, Tate S, Röst H, Selevsek N, Reiter L, et al. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: A new concept for consistent and accurate proteome analysis. Molecular and Cellular Proteomics. 2012;11(6).
6. Tautenhahn R, Böttcher C, Neumann S. Highly sensitive feature detection for high resolution LC / MS. 2008;16:1–16.
7. Smith CA, Want EJ, Maille GO, Abagyan R, Siuzdak G. XCMS : Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment , Matching , and Identification. 2006;78(3):779–87.
8. Capellades J, Navarro M, Samino S, Garcia-ramirez M, Hernandez C, Simo R, et al. geoRge: A Computational Tool To Detect the Presence of Stable Isotope Labeling in LC/MS-Based Untargeted Metabolomics. 2016;
9. Buescher JM, Antoniewicz MR, Boros LG, Burgess SC, Brunengraber H, Clish CB, et al. A roadmap for interpreting 13 C metabolite labeling patterns from cells. Current Opinion in Biotechnology. 2015;34:189–201.
10. Giné R, Capellades J, Badia JM, Vughs D, Schwaiger-Haber M, Vinaixa M, et al. HERMES: a molecular formula-oriented method to target the metabolome. bioRxiv. 2021 Mar 9;2021.03.08.434466. Available from: https://doi.org/10.1101/2021.03.08.434466
11. Domingo-Almenara X, Montenegro-Burke JR, Guijas C, Majumder ELW, Benton HP, Siuzdak G. Autonomous METLIN-Guided In-source Fragment Annotation for Untargeted Metabolomics. Analytical Chemistry. 2019 Mar 5; 91(5):3246–53.
12. Montenegro-Burke JR, Guijas C, Siuzdak G. METLIN: A Tandem Mass Spectral Library of Standards. In: Methods in Molecular Biology. 2020.
13. Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, et al. MassBank: A public repository for sharing mass spectral data for life sciences. Journal of Mass Spectrometry. 2010;45(7).
14. Dührkop K, Shen H, Meusel M, Rousu J, Böcker S. Searching molecular structure databases with tandem mass spectra using CSI : FingerID. 2015;
15. Aron AT, Gentry EC, Mcphail KL, Nothias L, Nothias-esposito M, Bouslimani A, et al. Reproducible molecular networking of untargeted mass spectrometry data using GNPS. Nature Protocols. 2012;

16. Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard : open source software for rapid proteomics tools development. 2008;24(21):2534–6.

17. Huang P, Liu C, Gao W, Chu B, Cai Z, Tian R. Synergistic optimization of Liquid Chromatography and Mass Spectrometry parameters on Orbitrap Tribrid mass spectrometer for high efficient data-dependent proteomics. Journal of Mass Spectrometry. 2020 Apr 1;56(4).

18. Tautenhahn R, Patti GJ, Rinehart D, Siuzdak G. XCMS online: A web-based platform to process untargeted metabolomic data. Analytical Chemistry. 2012 Jun 5; 84(11):5035–9.

19. Sindelar M, Patti GJ. Chemical Discovery in the Era of Metabolomics. Journal of the American Chemical Society. 2020;142(20):9097–105.

20. Duan L, Molnar I, Snyder JH, Shen G, Qi X. Discrimination and Quantification of True Biological Signals in Metabolomics Analysis Based on Liquid Chromatography-Mass Spectrometry. Molecular Plant. 2016;99(August):1217–20.

21. Jang C, Chen L, Rabinowitz JD. Metabolomics and Isotope Tracing. Vol. 173, Cell. Cell Press; 2018. p. 822–37.

9. Autoavaluació

L'elaboració d'aquest treball de final de grau ha estat una experiència molt constructiva que m'ha permès desenvolupar-me com a científic dins d'un grup de recerca punter en l'estudi de la metabolòmica.

L'estada de pràctiques a la Plataforma Metabolòmica, de la que se'n deriva aquest treball, m'ha ajudat a assolir múltiples competències transversals i específiques. D'entre elles, destaco la competència CT3 (resoldre problemes de forma creativa), ja que he hagut de buscar maneres per quantificar el marcatge dels metabòlits, donat que no hi havia eines preexistents per calcular el FrC i el MIRS. A més, vaig haver de pensar una manera innovadora d'adquirir espectres MS2 d'alta qualitat augmentant el temps d'injecció dels ions, la qual ha estat una estratègia molt exitosa.

Els coneixements que he adquirit al llarg de la doble titulació en Bioquímica, Biologia Molecular i Biotecnologia han estat essencials per poder dur a terme aquest treball, ja que m'han otorgat els fonaments teòrics per entendre la cromatografia líquida, l'anàlisi bioinformàtica de les dades de LC-MS i l'obtenció dels resultats i figures.

10. Annexos

10.1 Algorismes utilitzats

Algorisme 1: Detecció dels *Scans of Interest* (SOI)

Input: Una matriu de punts anotats $D = \{RT,\ intensity\}$, amplada de la finestra de temps de retenció bw, intensitat mínima I_{min}, metadata dels scans adquirits h, mínima densitat de scans (%) ρ, mínima puntació de caos C_{min}

Output: Llista de SOIs SOI_{list}

$D' = filter(D,\ intensity > I_{min})$

// Defineix intervals de temps de retenció

$B_{start} = [0,\ bw,\ 2 \cdot bw, \dots]\ ;\ B_{end} = [bw,\ 2 \cdot bw,\ 3 \cdot bw, \dots]$

// Compta el número de scans totals en cada interval de temps de retenció

$Thr = [\ B_{start} \leq\ h_{RT} \leq B_{end}]\ \cdot \rho$

Per a cada interval:

N_{bin} = $B_{start}[bin]\ \leq D'_{RT} \leq B_{end}[bin]$ // Compta nombre de punts anotats en l'interval

$\phi_{bin} =\ N_{bin} > Thr_{bin}$ // Determina si hi ha prou punts en l'interval

$\sigma =\ S(\phi_{bin})$ // Troba tots els conjunts punts connectats a ϕ_{bin}

Per a cada $conjunt$ en σ:

Calcula ρ_{chaos} amb el vector d'intensitats del $conjunt$

Si $\rho_{chaos} > C_{min}$:

Afegeix la informació del $conjunt$ a la SOI_{list} (inici, final, intensitat, anotació …)

Retorna SOI_{list}

Algorisme 1: L'algorisme de detecció de SOIs és una nova manera de detectar pics en les dades sense imposar cap forma d'elució cromatogràfica (a diferència de la resta de mètodes). La base de l'algorisme és la utilització d'una finestra mòbil per calcular la densitat de scans. Aquesta característica fa que l'algorisme sigui robust davant l'absència temporal de senyals, fet que és freqüent en senyals poc intenses i que altres algorismes no resolen (produeixen el que s'anomena *peak splitting*). Adaptat de Giné *et al.* (10)

10.2 Mètriques de similitud espectral

$$Topsoe(P,Q) = \sum_{i=1}^{d} \left(P_i \cdot \ln\left(\frac{2P_i}{P_i + Q_i}\right) + Q_i \cdot \ln\left(\frac{2Q_i}{P_i + Q_i}\right)\right)$$

$$Squared_{chord}(P,Q) = \sum_{i=1}^{d} \left(\sqrt{P_i} - \sqrt{Q_i}\right)^2$$

$$Fidelity(P,Q) = \sum_{i=1}^{d} \sqrt{P_i Q_i}$$

10.3 Figures suplementàries

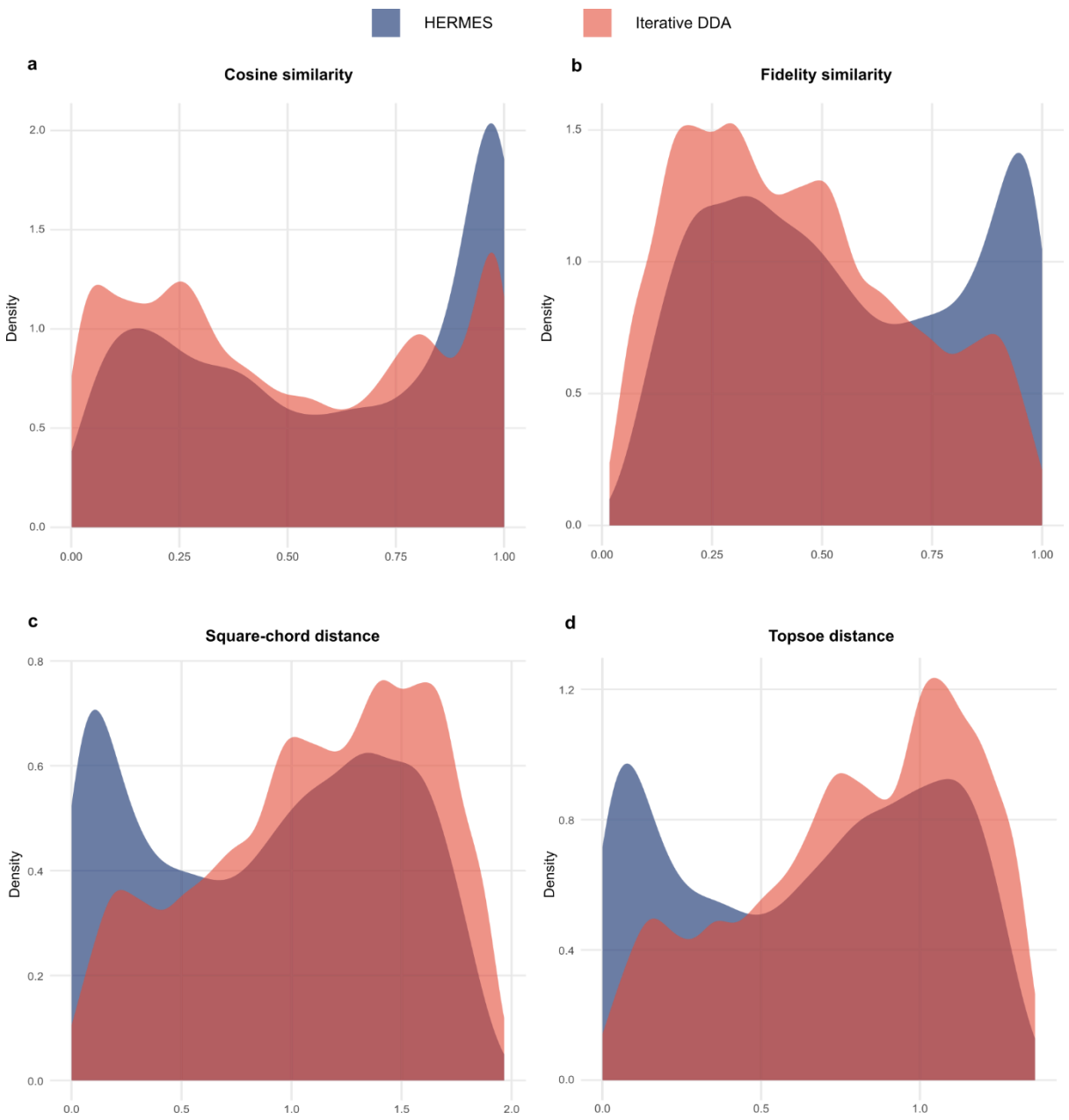

Figura suplementària S1. Comparació de mètriques alternatives per calcular la similitud entre espectres de fragmentació MS2. a) Similitud de cosinus b) Similitud de *Fidelity*. c) Distància *square-chord*. d) Distància *Topsoe*. Es va dur a terme una comparació entre les similituds espectrals obtingudes a partir dels espectres d'HERMES (blau) i els de DDA (vermell), contra les bases de dades d'espectres de referència. S'observa com HERMES obté majors similituds (a i b) i menors distàncies (c i d) en comparació a DDA. El millor perfil de mètriques apunta cap a una major qualitat dels espectres d'HERMES, ja que aquests són obtinguts de forma dirigida, buscant activament els compostos que es busca identificar. Adaptat de Giné *et al*. (10)