Informe

REPOSITORI:

https://github.com/msantacreuv/Santacreu-Vilaseca-Marta-PAC1

1 ELECCIÓ DEL DATASET

He triat el conjunt de dades *Cachexia*, del repositori de *Github* [1] facilitat en l'enunciat, per diversos motius.

Primer de tot, per la meva formació, tan acadèmica com professional. Com a bioquímica i biòloga molecular amb experiència en investigació de estats metabòlics en situacions fisiopatològiques, el meu interès gravita especialment cap a l'estudi dels processos metabòlics, i les seves alteracions. Es per això que considero que l'estudi de la caquèxia, sent aquesta un estat d'alteració i deterioració metabòlica profunda, és quelcom essencial per comprendre millor els mecanismes subjacents de diverses patologies així com identificar en el camí possibles dianes terapèutiques i/o biomarcadors.

D'altra banda, després de cercar diversos *datasets*, l'estructura d'aquest m'ha semblat especialment adequat per a aquesta activitat ja que no només és utilitzat en estudis de metabòlica sinó que a més presenta una estructura senzilla que facilita la seva interpretació i adaptació al exercici proposat.

Per aquests motius he considerat aquest *dataset* com una opció idònia per a realitzar l'activitat, tant per la seva rellevància biològica com per les seves característiques tècniques.

2 INCORPORACIÓ DEL DATA SET AL SUMMARIZEDEXPERIMENT

Per començar, importem les dades que hem descarregat prèviament en format csv, utilitzem la funció **read.csv.** Per fer-nos una idea de com es troben distribuïdes les dades mostrem les primeres files amb **head()**. Tenim un *dataframe* amb 77 observacions (files) i 65 variables (columnes). La primera i segona columna contenen el identificador del pacient (*Patient.ID)* i el grup al que pertanyen (*Muscle.loss)* respectivament. Les altres 63 columnes comprenen valors numèrics de les concentracions dels diferents metabòlits. A l'hora d'incorporar les dades en l'objecte *summarizedExperiment* ho haurem de tenir en compte.

Ara que ja tenim les dades del *dataset* carregades, procedim a crear l'objecte de la classe *SummarizedExperiment*. De forma similar al *ExpressionSet*, aquest objecte inclou tant les dades en si (en aquest cas valors de metabòlits) com les metadades associades. Per crear-lo hem de considerar:

1. Les dades dels metabòlits seran la matriu *d'assays*: Per tant en el nostre *dataset* englobarà de la tercera columna de les dades fins la última.

2. Les metadades de les mostres es crearan com a *dataframe* i seran les dues primeres columnes.

3. No tenim metadades sobre les *features* (metabòlits).

Primer crearem la matriu de metabòlits que, com hem dit, contindrà els valors de metabòlits per cada pacient, les columnes seran cada mostra (*sample*) i les files seran els metabòlits (*features*). És important considerar que en el *dataset* que hem descarregat tenim les dades al reves, és a dir, les mostres són les files i els metabòlits les columnes, haurem de transposar la matriu per a que quedi com la necessitem per l'objecte ja que per a aquesta classe és molt important que les columnes de *l'assay* coincideixin amb la informació de les files de les metadades.

Seguidament dins el mateix objecte guardarem les metadades (la informació, en aquest cas, de cada mostra), on les files han de ser els pacients i les columnes seran el ID d'aquests i el grup al que pertanyen (caquèxia o control). Farem anar colData ja que tenim metadades de les *samples,* si en tinguéssim sobre les *features* hauríem també de guardar dins l'objecte aquesta informació com a rowData [2,3,4].

2.1 DIFERÈNCIES ENTRE EXPRESSIONSET I SUMMARIZEDEXPERIMENT

Tot i que les classes *d'ExpressionSet* i *SummarizedExperiment* tenen una funció molt similar (emmagatzemar dades experimentals i metadades associades), tenen certes diferències clau que determinen l'elecció de una o altra classe per cada experiment.

D'una banda *ExpressionSet* esta més orientada a dades d'expressió genètica que haguem obtingut a partir de tècniques com *microarrays* o seqüenciació, esta pensada per gestionar dades d'expressió de gens i les anotacions i metadades d'aquests. D'altra banda SummarizedExperiment és una classe més genèrica, podem gestionar dades d'expressió però també dades de metabolòmica, proteòmica....

Quelcom interessant és la capacitat de *SummarizedExperiments* de contenir múltiples conjunts de dades diferents alhora. Els *assays* són llistes de matrius i per tant podem anar emmagatzemant diferents *assays* en el mateix objecte (per exemple metabòlits i gens a la vegada), mentre que els *ExpressionSets* són més rígids i guarden una sola matriu d'expressió i no es tan senzill gestionar múltiples dades a la vegada.

També es diferent com es gestionen les metadades, mentre que amb els *ExpressionSets* les metadades són informació sobre els gens i es guarda en *phenoData*, en *SummarizedExperiment* podem guardar tant dades associades a les files (*features*) com a les columnes (*samples*) depenent de si ho guardem a *colData* o *rowData*. Per aquest motiu, tot i que els *ExpressionSet* són una estructura molt emprada i estandarditzada, la classe *SummarizedExperiment* ens aporta flexibilitat per adaptar-nos a qualsevol tipus de dades òmiques així com per treballar amb múltiples dades a la vegada. [4,5]

3 ANÀLISI EXPLORATÒRIA DE LES DADES

3.1 OBSERVACIÓ INICIAL DE LES DADES

Comencem fent una exploració prèvia on comprovem si l'objecte s'ha creat correctament i fem un cop d'ull a les metadades.

Les dimensions de l'objecte són de 63x77 és a dir 63 files (metabòlits) i 77 columnes (pacients) les quals coincideixen amb les dades que teníem al *dataset* inicial (un cop transposat). Tenim la matriu *d'assays* amb els identificadors com a nom de les columnes (*colnames*) i com a *rownames* trobem els noms dels metabòlits. *ColData* es comporta tal com esperàvem i conté la variable *Muscle.loss* que ens indica si el pacient pertany al grup caquèxia o control.

3.2 ESTADÍSTIQUES DESCRIPTIVES

Per continuar analitzant les dades és important que ens fem una idea de com es comporten aquestes, tan per veure si hi ha *missing values* (NA) (que podrien alterar el seu anàlisis estadístic futur) com per veure el comportament genèric de les dades. Per fer això hem fet servir la funció **skim()**.

Les dades no presenten NA i al analitzar els histogrames dels metabòlits s'aprecia que la major part de les mesures d'aquests es troben concentrades en rangs baixos amb alguns valors extrems, suggerint possibles *outliers*. Per tal de observar la conducta d'aquests farem un *boxplot* molt senzill (*fig. 1*), simplement per veure de forma ràpida valors que es comportin de forma curiosa.

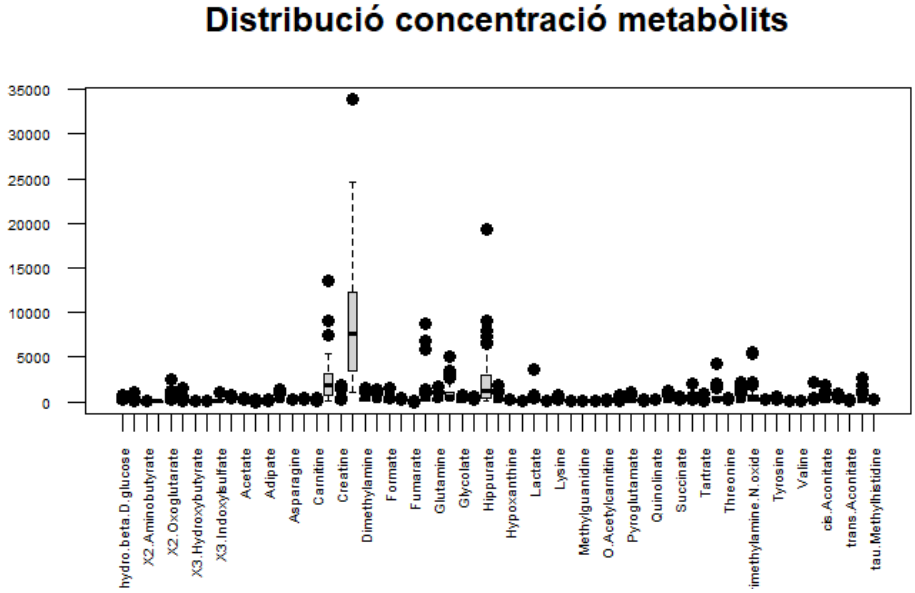

Figura 1. Boxplot metabòlits

Confirmem que efectivament hi ha alguns valors molt alts. Tot i que amb tants metabòlits no podem veure clarament els noms dels que tenen valors elevats observem que el metabòlit que presenta un valor màxim i més variabilitat coincideix amb el que té el valor més alt en el resum que hem fet prèviament, pertany a la Creatinina. Els altres amb valors

especialment alts són: *Hippurate*, *Citrate* i *Glucose*. Com que ens ha cridat especialment l'atenció la creatinina l'investigarem més en detall fent un histograma per veure la seva distribució (*fig. 2*).

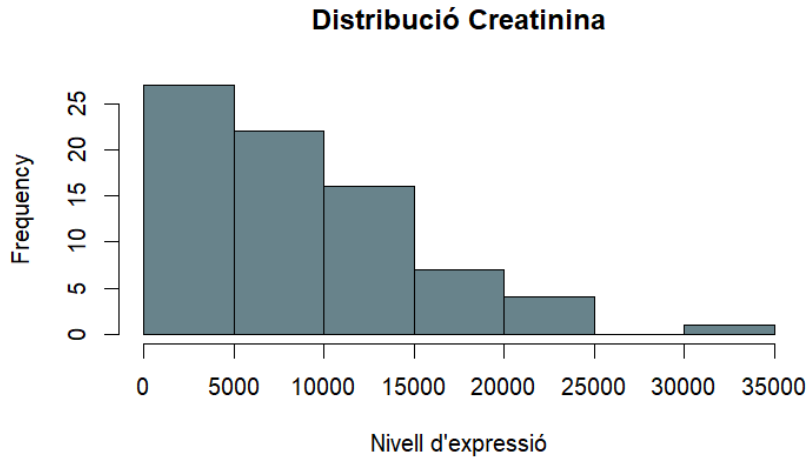

Figura 2. Distribució Creatinina

Tot i que la major part de les dades es concentren en valors de 0 a 5000, veiem clarament que hi ha un nombre molt baix de valors entre 30000-35000 i cap entre 25000 -30000 suggerint que es tracta, efectivament, de un *outlier*. Com que es tracta d'un anàlisis exploratori no eliminarem res però per a posteriors anàlisis amb les dades seria interessant repetir aquest procés així com mirar el rang interquartílic dels diferents metabòlits per determinar si hi ha *outliers* que s'hagin d'eliminar (o si són biològicament rellevants mantenir-los i fer anàlisis addicionals amb aquests).

Fins ara només hem mirat les dades en conjunt, però no podem oblidar que a les metadades (*colData*) tenim la informació sobre a quin grup pertanyen les nostres mostres (control o caquexia) per tant anem a repetir el procés separant per grup.

Quan mirem el resum del grup control (que hem fet amb skim()) veiem que altre cop el metabòlit amb nivells més alts es la creatinina (valor màxim de ~15000), factor que es repeteix en el grup caquèxia i que és en aquest grup on tenim el possible *outlier* de ~33000. Repetim el *boxplot* de tots els metabòlits que hem fet abans però separant pels dos grups (*fig 3*)

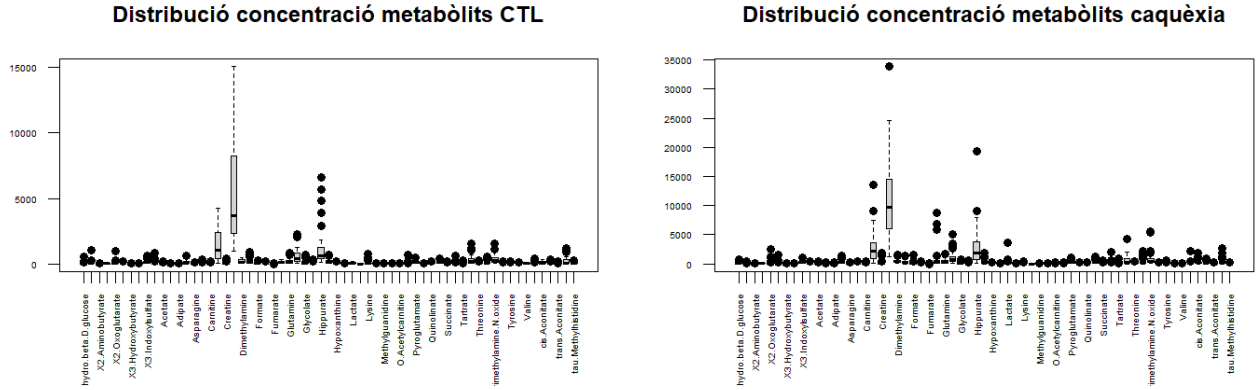

Figura 3. Distribució metabòlits segons els grups

En principi sembla que tenen un comportament similar tot i que hi ha valors especialment extrems en el grup caquexia. Podem trobar quins metabòlits son més abundants en ambdós grups amb ajuda del resum que hem fet anteriorment:

- **Control**: Creatinina, hipurat, citrat i glicina.
- **Caquexia**: Creatinina, hipurat, citrat i glucosa.

Fem un altre *boxplot,* aquest cop però més concret centrant-nos amb aquests 5 metabòlits i en les comparacions entre grups (*fig. 4*)

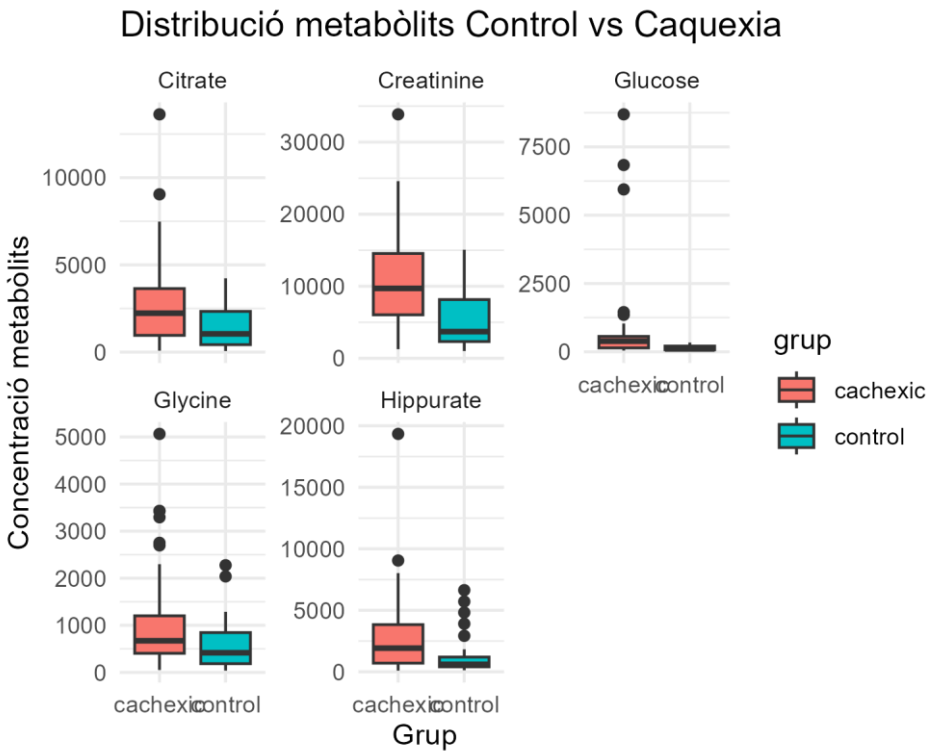

Figura 4. Comparació metabòlits més alts per grup

Els cinc metabòlits en el grup caquèxia presenten valors més alts respecte el control així com més valors extrems, coincidint amb el que s'ha observat en la *figura 3*.

3.3 Interaccions i Correlacions

Per seguir amb el nostre anàlisis exploratori, explorarem les correlacions entre els metabòlits per tal d'observar si hi ha patrons especialment interessants a tenir en compte en futurs anàlisis. Per començar fem un PCA (anàlisis de components principals) que ens permet reduir la quantitat de variables (dimensions) per a visualitzar patrons o agrupaments en les nostres dades, en aquest cas ho fem separant per els dos grups.

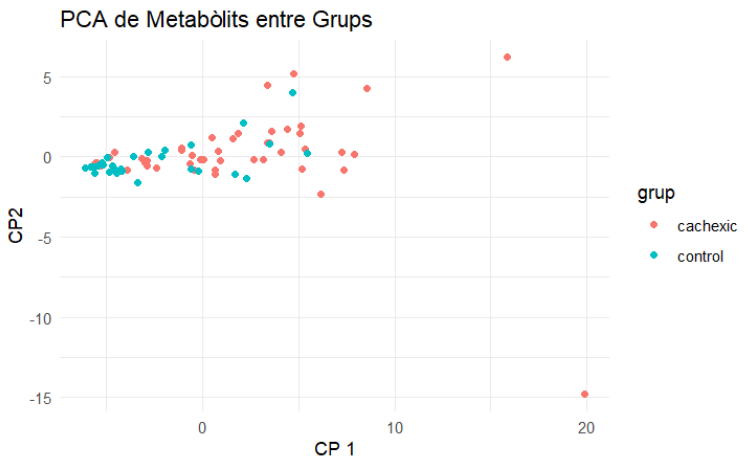

Figura 5. PCA dels metabòlits separant per grups

Tot i que els dos grups no s'acaben de separar si que s'aprecia una zona més concentrada de punts en *control* mentre que els del grup *cachexic* presenten una major dispersió. Aquest fet ens suggereix major variabilitat en els metabòlits d'aquest grup. Per seguir explorant aquesta idea de patrons diferents segons els grups podem prosseguir amb un *heatmap* que ens permetrà estudiar el comportament de les correlacions dels metabòlits segons els grups *(fig.6)*

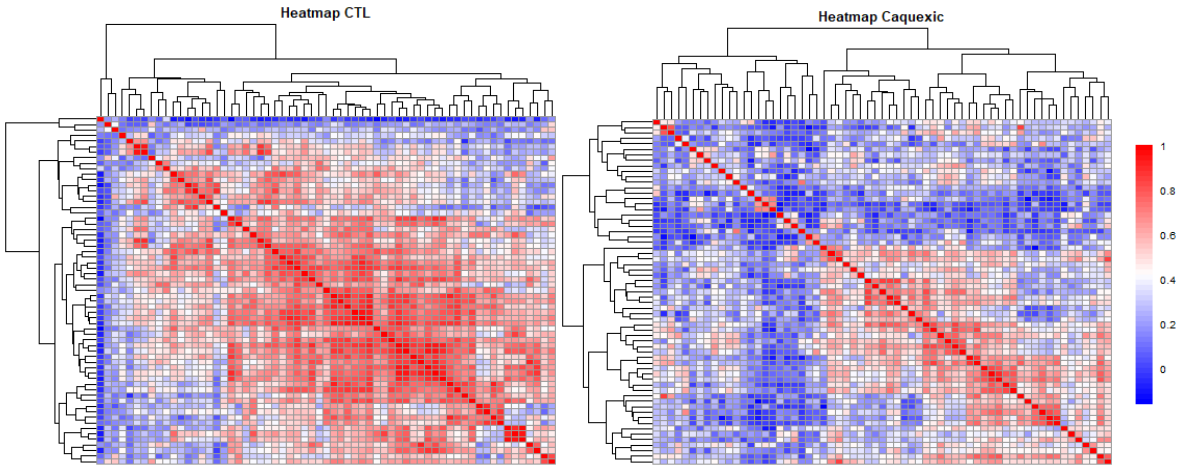

Figura 6. Heatmaps dels metabòlits separant per grups

A simple vista podem ressaltar com en el cas del grup control hi ha un nombre molt més abundant de correlacions positives, mentre que en el cas del grup *caquexic* la zona del quadrant de baix a la dreta presenta correlacions especialment positives i tot el demès negatives.

4 INTERPRETACIÓ BIOLÒGICA DELS RESULTATS

Un dels primers resultats que hem vist durant la anàlisis exploratòria és que el metabòlit més abundant, especialment en el grup caquèxia, és la creatinina. Aquest metabòlit deriva de la descomposició de la creatina, la qual trobem principalment en músculs. La concentració especialment alta d'aquest metabòlit sembla suggerir una major taxa de descomposició muscular en el grup caquèxia, quelcom lògic si considerem que una de les característiques d'aquest estat patològic és la pèrdua de massa. Tan el grup control com el grup caquèxia presenten valors elevats d'hipurat, valors alts d'aquest metabòlit s'associen a major diversitat en el microbioma, altre cop destaca especialment els nivells d'aquest en el grup patològic, això podria deures a un desequilibri de la microbiota intestinal en aquests pacients (relacionat amb malnutrició o pèrdua de massa muscular) provocant una disbiosi intestinal i afectant per tant la producció de certs metabòlits [6,7].

Podem destacar també els nivells de citrat, metabòlit fonamental en el cicle de Krebs, que es clau per la producció d'energ. En el grup caquèxia aquest metabòlit es presenta amb major dispersió i nivells més elevats (respecte el control) el qual sembla indicar una variabilitat més elevada. Aquesta dispersió podria estar causada per alteracions del metabolisme energètic com a resposta a l'estrès cònic lligat a aquesta condició, aquest augment podria reflectir l'adaptació del metabolisme energètic per intentar mantenir els nivells d'energia en un entorn amb pèrdua de massa muscular.

Pel que fa al a glucosa, tot i que en control manté uns rangs considerablement estable (comportament pròpi en individus sans), en el grup caquèxia són considerablement alts i amb bastanta dispersió, això podria ser causat per la inflamació crònica associada a la caquèxia que pot interferir en la capacitat del cos per utilitzar eficientment la glucosa indicant un metabolisme energètic alterat (com ja em vist amb el citrat).

Al fer el PCA s'ha observat que els metabòlits del grup control es presenten més compactes suggerint major homogeneïtat, mentre que pel contrari el grup caquèxia presenta una dispersió considerablement elevada. A nivell biològic això ens suggereix que en el cas del grup patològic hi ha una major diversitat metabòlica (relacionada amb els efectes de la condició) que produeixen patrons metabòlics dispars entre els pacients. També sembla indicar que hi ha un comportament bastant heterogeni en les dades de l'estat caquèxia i per tant del progres d'aquest, cal dir que no sabem l'origen de aquesta caquèxia i si els pacients presenten una patologia subjacent diferent això podria afectar al comportament d'aquests patrons metabòlics, el qual també es reflecteix en els *heatmaps* dels grups. En el cas de control hi ha una gran quantitat de correlacions positives, el qual podria indicar un metabolisme més coordinat i equilibrat, pel contrari en el grup patològic aquest quadre canvia mostrant principalment correlacions negatives que suggereixen una disrupció metabòlica de certes vies així com una conducta allunyada del estat fisiològic (control).

En conclusió, amb el nostre anàlisis exploratori hem pogut observar clarament un comportament diferencial entre els dos grups, agafant el control com a grup saludable es evident les alteracions en el grup caquèxia, destacant especialment una alteració en el metabolisme energètic però també en la microbiota i equilibri metabòlic general.

5 REFERÈNCIES

1. *metaboData/Datasets/2024-Cachexia/description.md at main · nutrimetabolomics/metaboData*. (s.f.). GitHub. https://github.com/nutrimetabolomics/metaboData/blob/main/Datasets/2024-Cachexia/description.md
2. LiquidBrain Bioinformatics. (2021, 8 de mayo). *Summarized Experiment (se) Object from Bioconductor* [Video]. YouTube. https://www.youtube.com/watch?v=ObaFEq4U3DQ
3. (s.f.). Bioconductor - Home. https://bioconductor.org/packages/release/bioc/manuals/SummarizedExperiment/man/SummarizedExperiment.pdf
4. Zanfardino, Mario & Franzese, M. & Pane, Katia & Cavaliere, Carlo & Monti, Serena & Esposito, Giuseppina & Salvatore, Marco & Aiello, Marco. (2019). Bringing radiomics into a multi-omics framework for a comprehensive genotype–phenotype characterization of oncological diseases. Journal of Translational Medicine. 17. 10.1186/s12967-019-2073-2.
5. *ExpressionSet and SummarizedExperiment - Easy Guides - Wiki - STHDA*. (s.f.). STHDA - Accueil. https://www.sthda.com/english/wiki/expressionset-and-summarizedexperiment
6. Groothof D, Shehab NBN, Erler NS, Post A, Kremer D, Polinder-Bos HA, Gansevoort RT, Groen H, Pol RA, Gans ROB, Bakker SJL. Creatinine, cystatin C, muscle mass, and mortality: Findings from a primary and replication population-based cohort. J Cachexia Sarcopenia Muscle. 2024 Aug;15(4):1528-1538. doi: 10.1002/jcsm.13511. Epub 2024 Jun 19. PMID: 38898741; PMCID: PMC11294032.
7. Pallister T, Jackson MA, Martin TC, Zierer J, Jennings A, Mohney RP, MacGregor A, Steves CJ, Cassidy A, Spector TD, Menni C. Hippurate as a metabolomic marker of gut microbiome diversity: Modulation by diet and relationship to metabolic syndrome. Sci Rep. 2017 Oct 20;7(1):13670. doi: 10.1038/s41598-017-13722-4. PMID: 29057986; PMCID: PMC5651863.