Primo set di analisi sui dati di lipidomica

Davide Biganzoli

2/5/23

Introduzione

La lipidomica è una disciplina emergente che si concentra sull'analisi quantitativa dei lipidi all'interno di un campione biologico. I lipidi sono una classe di biomolecole complessa e eterogenea, che svolgono un ruolo chiave in una vasta gamma di processi biologici, tra cui la regolazione del metabolismo energetico, la membrana cellulare e la segnalazione cellulare.

Negli ultimi decenni, la lipidomica ha fatto passi da gigante nella comprensione del ruolo dei lipidi nella fisiologia e nella patologia umana, grazie anche alle tecnologie di analisi avanzate, come la spettrometria di massa che permettono di quantificare migliaia di lipidi in un'unica analisi.

In questo primo report, ho prodotto una prima analisi descrittiva che include dei boxplot che mostrano come ogni classe di lipidi distribuisce in termini di intensità. Inoltre, per rappresentare i campioni, ho prodotto dei boxplot che riportano la distribuzione dei valori per ogni sample.
In seguito, ho effettuato due distinte analisi delle componenti principali:

1. L'analisi della PCA sui dati non trasposti, che vuole esaminare la variabilità tra i campioni rispetto ai lipidi. La PCA viene eseguita su tutti i lipidi e fornisce una rappresentazione delle relazioni tra i campioni in base alla quantità dei lipidi.
2. L'analisi della PCA sui dati trasposti che invece esamina la variabilità tra i lipidi rispetto ai campioni. La PCA viene eseguita su tutti i campioni e fornisce una rappresentazione delle relazioni tra i lipidi in base alle espressioni nei campioni.

In entrambi i casi, l'obiettivo della PCA è di ridurre la dimensionalità del dataset mantenendo al contempo le informazioni più significative. Tuttavia, la disposizione dei dati influenzerà la forma e l'interpretazione del risultato della PCA.

Ho infine effettuato una prima analisi univariata producendo dei modelli di regressione lineare, così da poter identificare le classi di lipidi significativamente differenti tra le condizioni biologiche.

Table 1: Dove: NW = Normoweight (o CTRL, n=5); OW = Overweight (n=3); OB = Obese (n=6); SV = Severe obesity (n=2).

ID	Gruppo	mg campione
Sample_5	NW	87.4
Sample_16	NW	53.9
Sample_18	NW	80.0
Sample_20	NW	73.1
Sample_25	NW	51.6
Sample_1	OW	89.3
Sample_4	OW	88.4
Sample_21	OW	68.3
Sample_2	OB	81.8
Sample_3	OB	69.9
Sample_6	OB	60.9
Sample_7	OB	65.5
Sample_10	OB	85.7
Sample_26	OB	72.0
Sample_8	SV	79.6
Sample_9	SV	105.5

Analisi descrittiva

Ai dati era stata già effettuata una normalizzazione "Internal Standard" (IS), che in generale permette di correggere le variazioni introdotte dalla variabilità della matrice biologica, dalla variabilità degli strumenti di analisi e dalla variabilità dell'efficienza di ionizzazione. In seguito la media di ciascun analita è stata normalizzata per i mg di campione (Tabella 1).

Ai dati ho applicato la trasformata logaritmica per far fronte alla forte dispersione dei valori di picco per alcune classi di citochine.

Inizialmente vi erano 43 differenti classi di lipidi, ma ho preferito raggruppare alcune subclassi: come ad esempio le Ceramidi, che sono state raggruppate tutte sotto l'identificativo "Cer", seguendo l'ontologia proposta da LIPID MAPS (che risulta essere quella più comune ed implementata nelle pipeline di analisi). Ovviamente se vogliamo fare analisi più approfondite sulle sub-classi, non è un problema riconvertire le classi secondo l'ontologia precedente.

Qui sotto ho riportato tutte le 29 classi di lipidi incluse nel dataset:

Class	Freq
CAR	5
Cer	50
CL	1
DG	76
DGCC	1
DGO	15
DGTS	1
FA	32
Hex2Cer	8
Hex3Cer	9
HexCer	6
LPC	12
LPCO	6
LPE	8
LPEO	4
MG	9
NAE	6
PC	126
PE	103
PEtOH	1
PG	4
PI	11
PS	17
PSO	6
SM	49
SPB	3
ST	6
TG	248
TGO	44

In totale ci sono 870 analiti, tenendo conto dei 24 lipidi che si ripetono nel nome ma che ricadono in una subclasse differente.

Come è possibile vedere, ci sono alcune classi che contengono un solo lipide riportato. Seppur non riportata, vi è una classe NA che riporta al suo interno solo 3 lipidi (Cer 18:1(O2)/34:3, CoQ5, SL 15:2(O)/36:0).

Ho voluto visualizzare anche come distribuisce ogni classe di lipidi per i valori di area di picco:

Per quanto riguarda i 16 sample posso produrre dei boxplot informativi che valutino la distribuzione dei valori per ognuno:

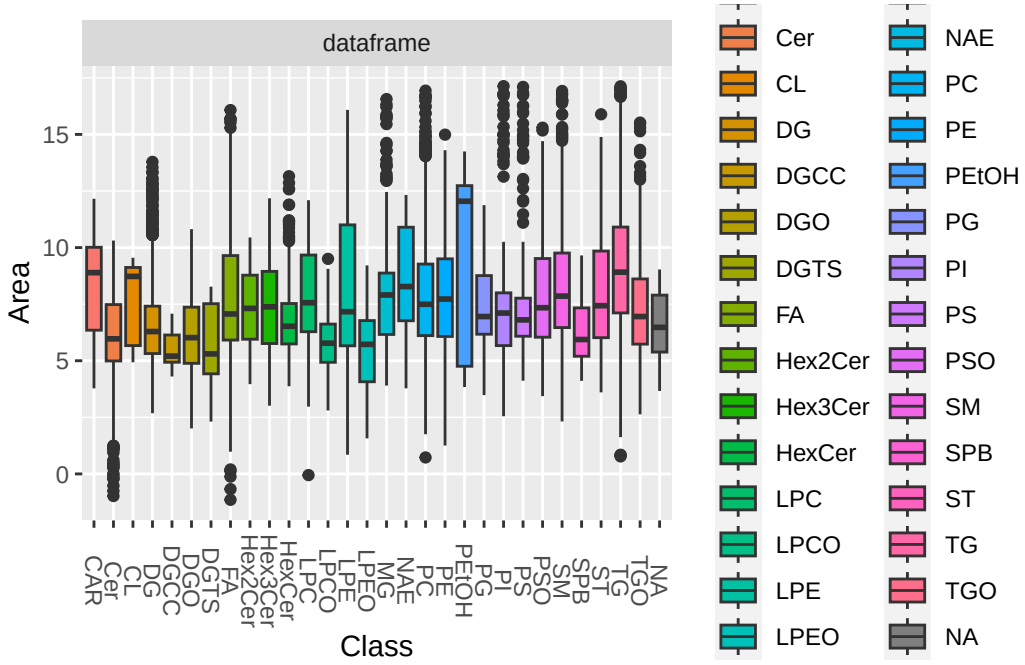


Figure 1: Boxplot chart per esaminare la distribuzione dei valori di area di picco per classe di lipidi

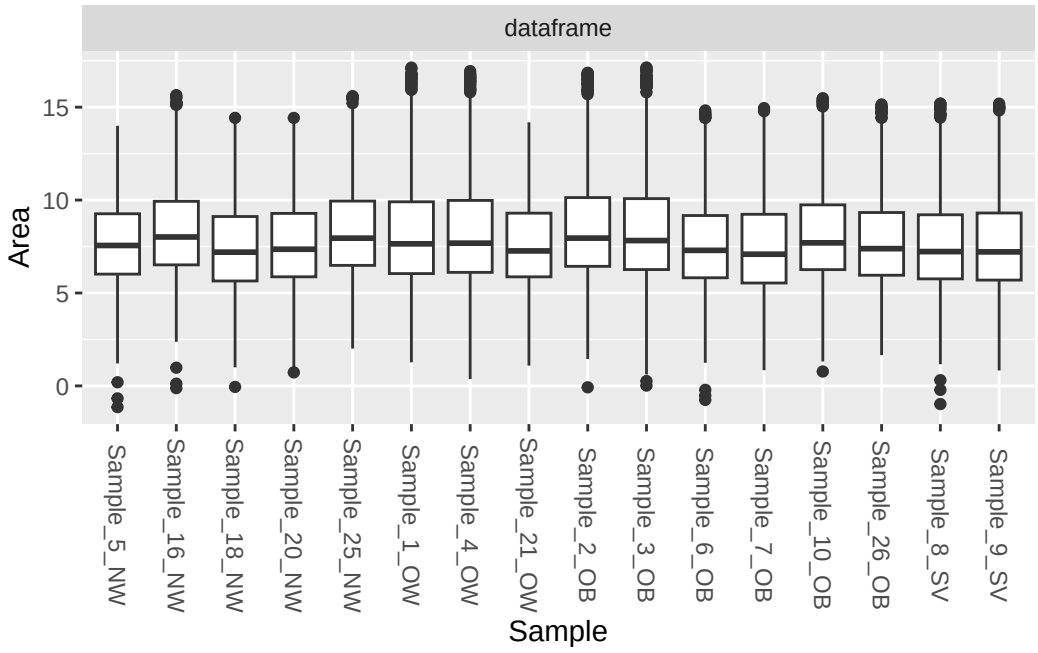


Figure 2: Boxplot chart per esaminare la distribuzione dei valori di area di picco per sample

Vediamo che le distribuzioni sono uniformi, tenendo conto del fatto che i dati erano stati già pre-processati da Unitech OMICs. Infatti, di tutti i lipidi sono stati considerati quelli che, **nei campioni Pool**, presentavano un valore di area con CV% inferiore del 30%.

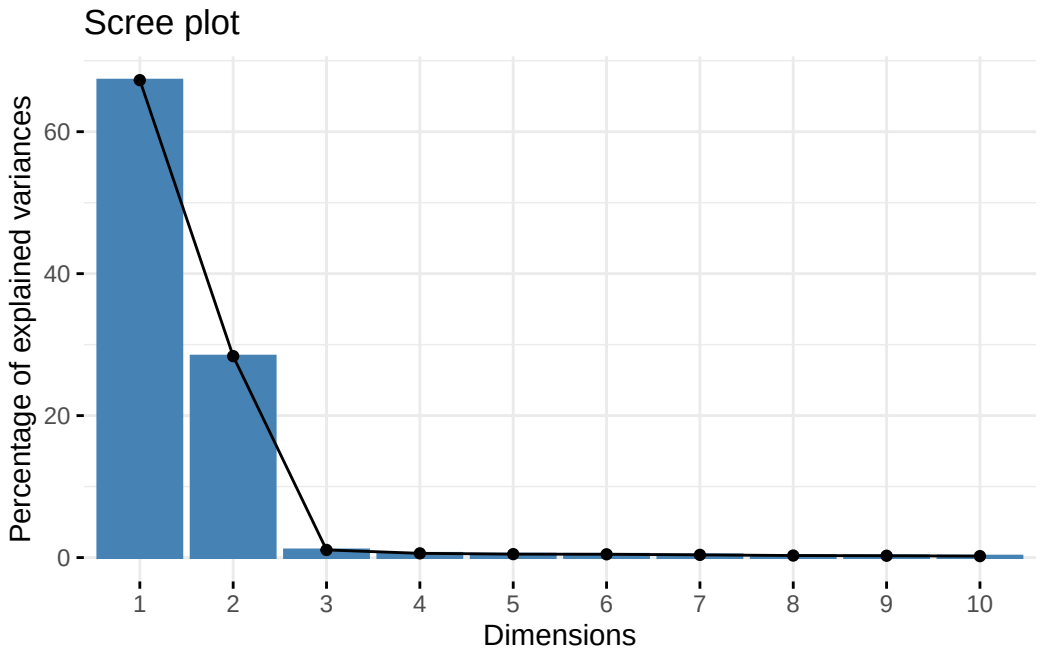
Analisi multivariata

Ho quindi deciso di produrre una prima analisi delle componenti principali, come specificato nell'introduzione:

In generale, se si è interessati a comprendere le relazioni tra i campioni, si dovrebbe eseguire la PCA come indicato al punto 1. Se invece si è interessati a comprendere le relazioni tra i gruppi di lipidi, si dovrebbe eseguire la PCA come indicato al punto 2.

PCA (1): dati non trasposti

Di nuovo, questo tipo di setting dell'analisi delle componenti principali vuole esaminare la variabilità tra i campioni rispetto ai lipidi:



Nella prima immagine ho riportato lo screeplot che visualizza quanto ogni componente spiega la varianza dei campioni. Nella seconda ho voluto rappresentare come si dispongono sul piano bidimensionale tutti i punti, corrispondenti ai singoli lipidi; inoltre ho sovrapposto un biplot riportante i vettori, i quali con la direzione e la lunghezza indicano la correlazione tra i sample

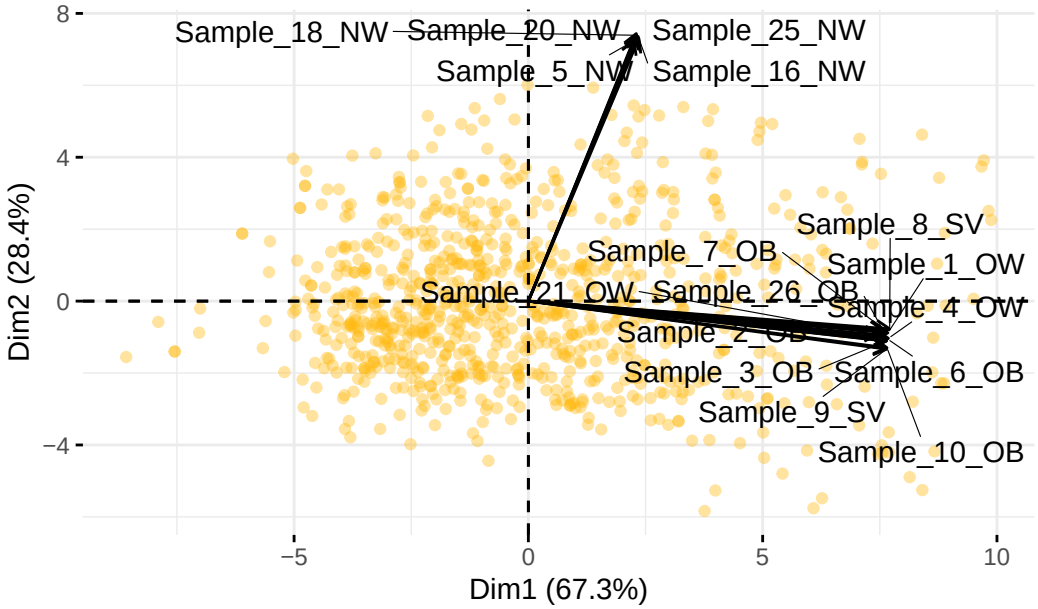
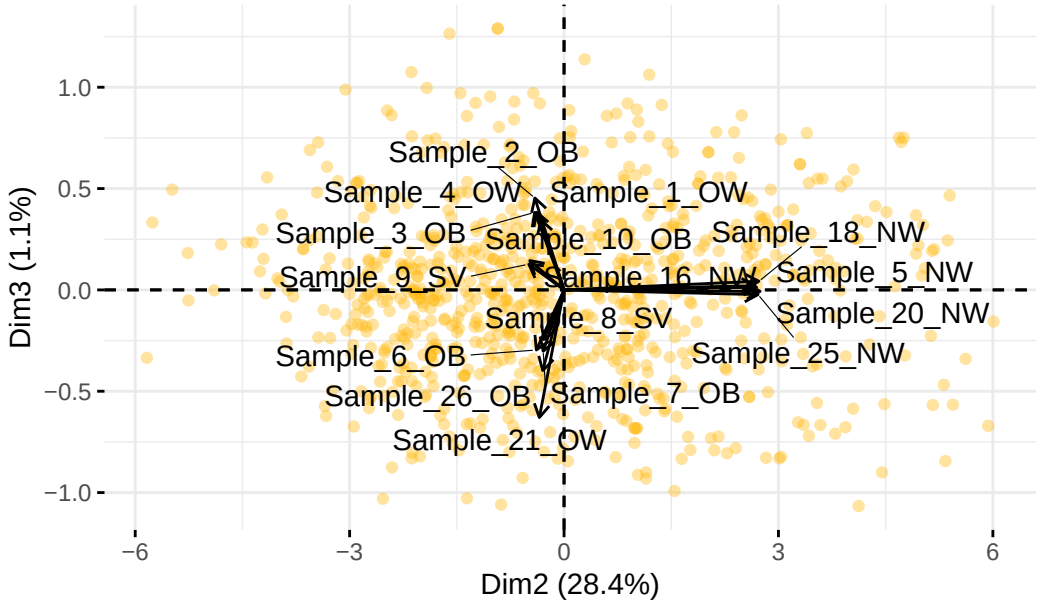


Figure 3: Screeplot e biplot di PCA (1)

e la loro importanza nella descrizione della varianza dei dati. Si evidenziano una forte correlazione tra i campioni del gruppo di controllo (NW); inoltre si verifica la medesima condizione tra gli altri gruppi e, intersecandosi perpendicolarmente con i vettori di controllo si evidenzia maggiormente l'assenza di correlazione tra i due cluster di vettori.

Riporto qui sotto la rappresentazione della seconda e della terza componente:



PCA (2): dati trasposti

Nuovamente, questo setting delle analisi permette di esaminare la variabilità tra i lipidi rispetto ai campioni, fornendo una rappresentazione delle relazioni tra i lipidi in base alle espressioni nei campioni.

Come sopra riporto lo screeplot e la rappresentazione sul piano bidimensionale, con biplot, della prima e della seconda componente:

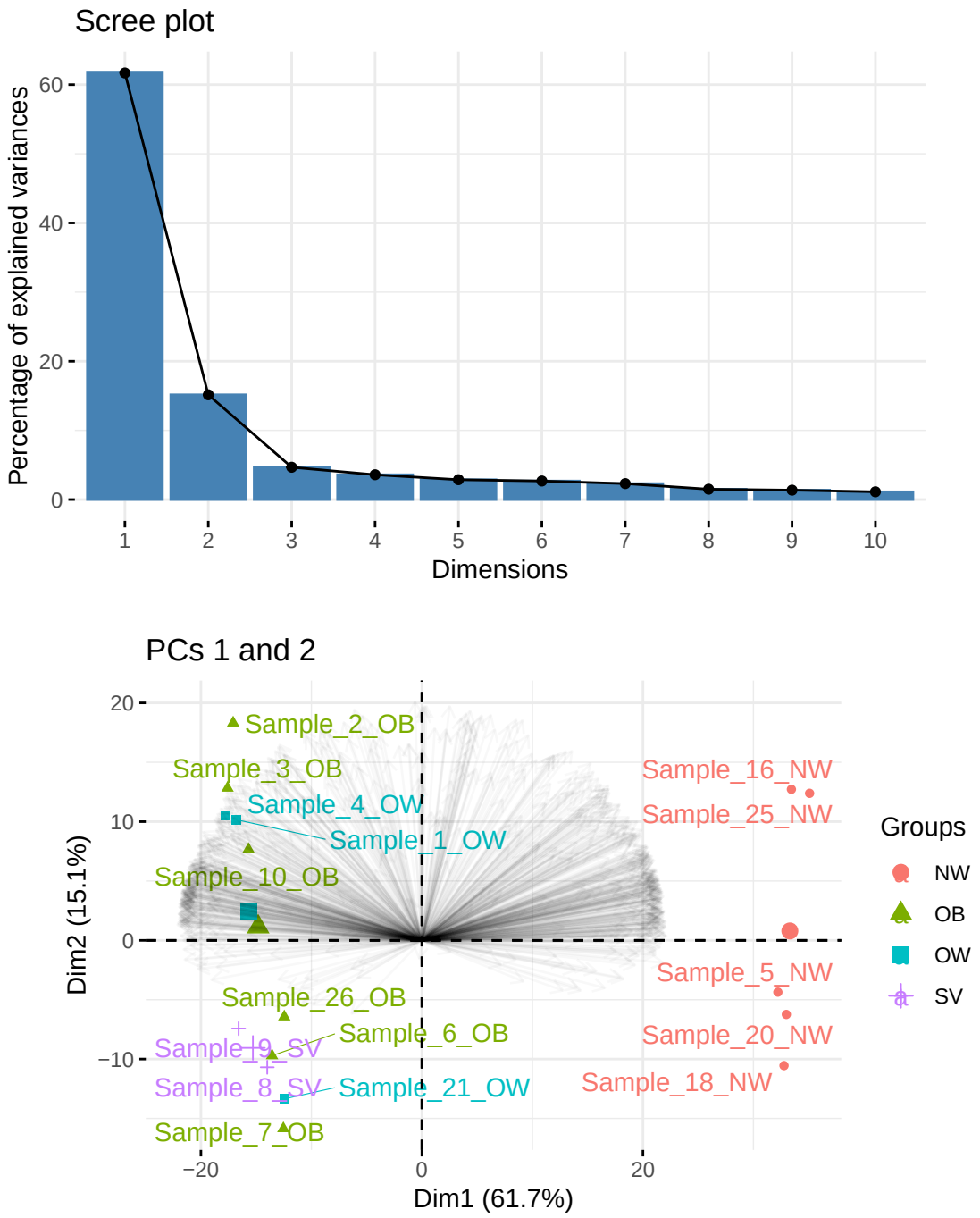


Figure 4: Screeplot e biplot di PCA (2)

E aggiungo anche la rappresentazione della seconda e della terza:

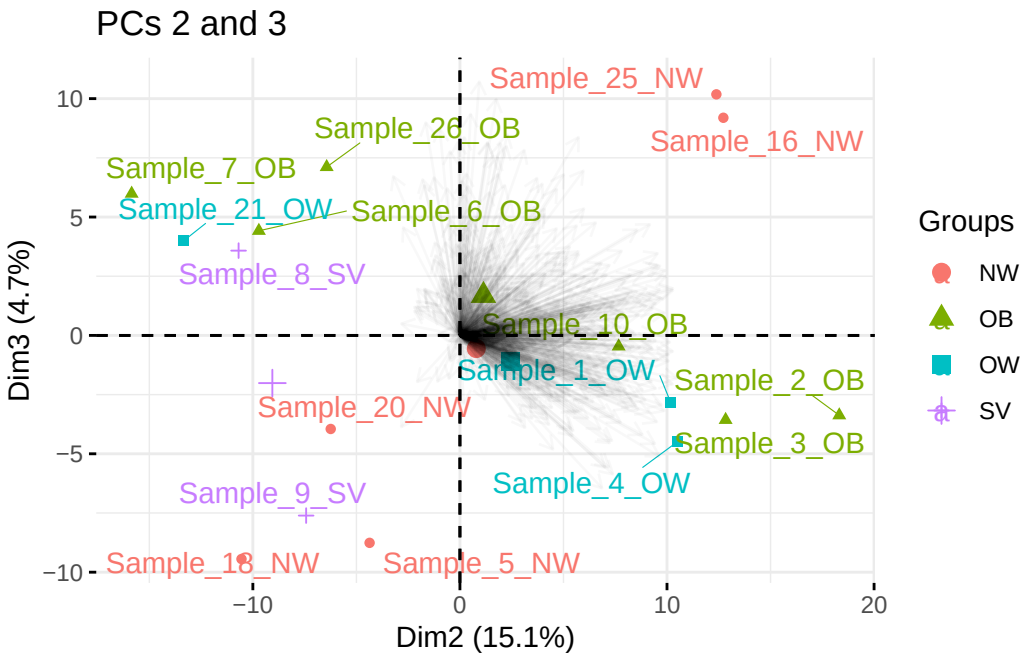


Figure 5: Seconda e terza componente proiettate per la PCA (2)

Analisi univariata

L'analisi univariata può essere svolta fra due gruppi, può essere multi-gruppi o anche multi-fattoriale, tuttavia ho preferito partire dal confronto tra il gruppo NW (che rappresenta il nostro controllo) contro tutti gli altri gruppi (OW, OB e SV).

L'obiettivo è quello di identificare eventuali differenze significative nei livelli di regolazione dei lipidi, che possono essere attribuite ai vari gruppi (ossia, le condizioni sperimentali che confrontiamo).

Per fare ciò, ho adattato un modello lineare per modellare una relazione tra variabile dipendente (in questo caso, i valori di area di picco) e una o più variabili indipendenti (nel nostro caso le condizioni sperimentali). Ho così calcolato una statistica t moderata, che tiene conto della variabilità dei dati e si adatta ai confronti multipli, aiutando a ridurre il rischio di falsi positivi (ad esempio, identificare differenze che non sono effettivamente significative).

Il "volcano plot" è un tipo di grafico comunemente utilizzato negli studi -omici per visualizzare i dati delle analisi delle differenze di espressione dei lipidi in diversi campioni.

I tre volcano plot riportano i tre confronti, come stabilito sopra. L'asse delle ordinate rappresenta la misura della significatività statistica dell'effetto di un trattamento o di una condizione su un certo lipide, mentre l'asse delle ascisse rappresenta il valore del logaritmo del fold change di quel lipide. Il p-value indica la significatività statistica dell'associazione tra un gene e un

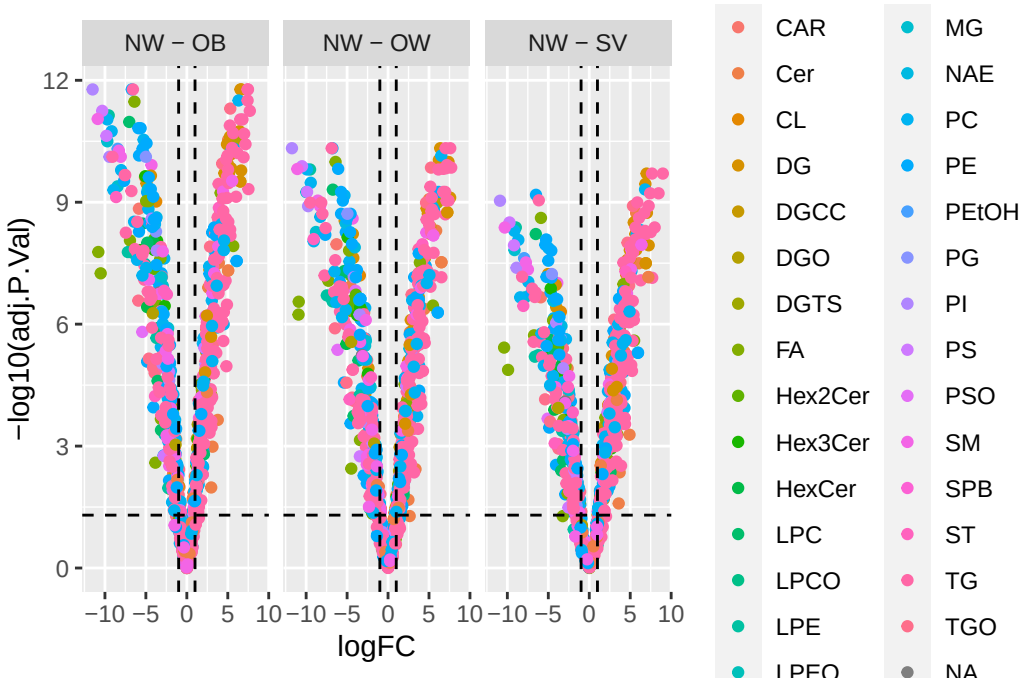


Figure 6: Volcano Plots coi tre confronti

trattamento o condizione, e in questo setting è stato calcolato con un aggiustamento per Benjamini-Hochberg. Questo metodo controlla l'expected false discovery rate (FDR) sotto la soglia (specificata) di 0.05.

Com'è possibile vedere dai volcano plots, a fronte di un elevato numero di lipidi facenti parte della famiglia dei trigliceridi (TG, con 248 lipidi), delle fosfatidilcoline (PC, con 126 lipidi) e fosfatidiletanolamine (PE, con 103 lipidi) non è facile interpretare questo risultato. Tuttavia, quello che emerge è che prendendo per esempio la classe dei diacilgliceroli (DG), si evidenzia una up-regulation di questi nei pazienti normopeso.

Software usati per le analisi

We used R version 4.3.0 (R Core Team 2023) and the following R packages: factoextra v. 1.0.7 (Kassambara and Mundt 2020), FactoMineR v. 2.8 (Lê, Josse, and Husson 2008), kableExtra v. 1.3.4 (Zhu 2021), knitr v. 1.42 (Xie 2014, 2015, 2023), lipidr v. 2.13.0 (Mohamed and Molendijk 2022), rmarkdown v. 2.21 (Xie, Allaire, and Grolemund 2018; Xie, Dervieux, and Riederer 2020; Allaire et al. 2023), tidyverse v. 2.0.0 (Wickham et al. 2019).

Allaire, JJ, Yihui Xie, Christophe Dervieux, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, et al. 2023. *rmarkdown: Dynamic Documents for r*. https://github.com/rstudio/rmarkdown.

Kassambara, Alboukadel, and Fabian Mundt. 2020. *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. https://CRAN.R-project.org/package=factoextra.

Lê, Sébastien, Julie Josse, and François Husson. 2008. "FactoMineR: A Package for Multivariate Analysis." *Journal of Statistical Software* 25 (1): 1–18. https://doi.org/10.18637/jss.v025.i01.

Mohamed, Ahmed, and Jeffrey Molendijk. 2022. *lipidr: Data Mining and Analysis of Lipidomics Datasets*. https://doi.org/10.18129/B9.bioc.lipidr.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2014. "knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.

———. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.

———. 2023. *knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Xie, Yihui, J. J. Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown.

Xie, Yihui, Christophe Dervieux, and Emily Riederer. 2020. *R Markdown Cookbook*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown-cookbook.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with "kable" and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.