UNIVERSITÀ DEGLI STUDI FIRENZE

Scuola di
Scienze Matematiche
Fisiche e Naturali

Corso di Laurea in
Chimica

# STUDIO DELLE VARIAZIONI INTER ED INTRAINDIVIDUALI DEL FINGERPRINT METABOLICO ATTRAVERSO L'ANALISI NMR DI CAMPIONI DI SALIVA

# STUDY OF INTER AND INTRAINDIVIDUAL CHANGES IN METABOLIC FINGERPRINT THROUGH NMR ANALYSIS OF SALIVA SAMPLES

**Relatore**
Prof. Claudio Luchinat

**Correlatore**
Dott. Leonardo Tenori

**Candidato**
Antonio Mazzoleni

Anno Accademico **2012/2013**

# INDICE

# 1 INTRODUZIONE

## 1.1 Scienze Omiche e Metabolomica

Nell'ambito della biologia dei sistemi[1], hanno avuto un intenso sviluppo nell'ultimo ventennio le cosiddette "scienze omiche": l'avvento del sequenziamento dell'intero genoma e altre tecnologie sperimentali *high-throughput* hanno trasformato la ricerca biologica da una disciplina relativamente povera di dati ad un insieme molto ricco e complesso di informazioni[2]. L'obiettivo fondamentale delle scienze omiche sta nell'interpretare questo vasto insieme di dati in un approccio olistico e globale per ricavarne un'informazione a più livelli sugli ambienti cellulari e infine su sistemi viventi evoluti.

Le scienze definite omiche sono varie e consentono di ottenere diversi tipi di informazione [Figura 1]. Le prime e le più sviluppate, come la genomica, la trascrittomica e la proteomica, identificano e caratterizzano i componenti molecolari di una cellula. Nello specifico, la genomica è lo studio della sequenza dell'intero genoma e dell'informazione contenuta in esso, ed è sicuramente la più matura e la più sviluppata tra le scienze omiche. La trascrittomica fornisce informazioni sia sulla presenza che sull'abbondanza relativa di trascritti di RNA, indicando in tal modo i componenti attivi all'interno della cellula. La proteomica mira ad individuare e quantificare i livelli cellulari di ogni proteina codificata dal genoma.[2]

Nel vasto campo delle scienze omiche si collocano anche la metabonomica e la metabolomica, che sono definite rispettivamente come "la misura quantitativa della risposta metabolica dinamica multiparametrica dei sistemi viventi a stimoli patofisiologici o mutazioni genetiche"[3] e come "l'analisi globale e quantitativa di tutti i metaboliti"[4]. Anche se effettivamente esiste una differenza tra le due, esse vengono considerate in definitiva quasi equivalenti dalla comunità scientifica[5], in quanto le procedure analitiche sono esattamente le stesse.

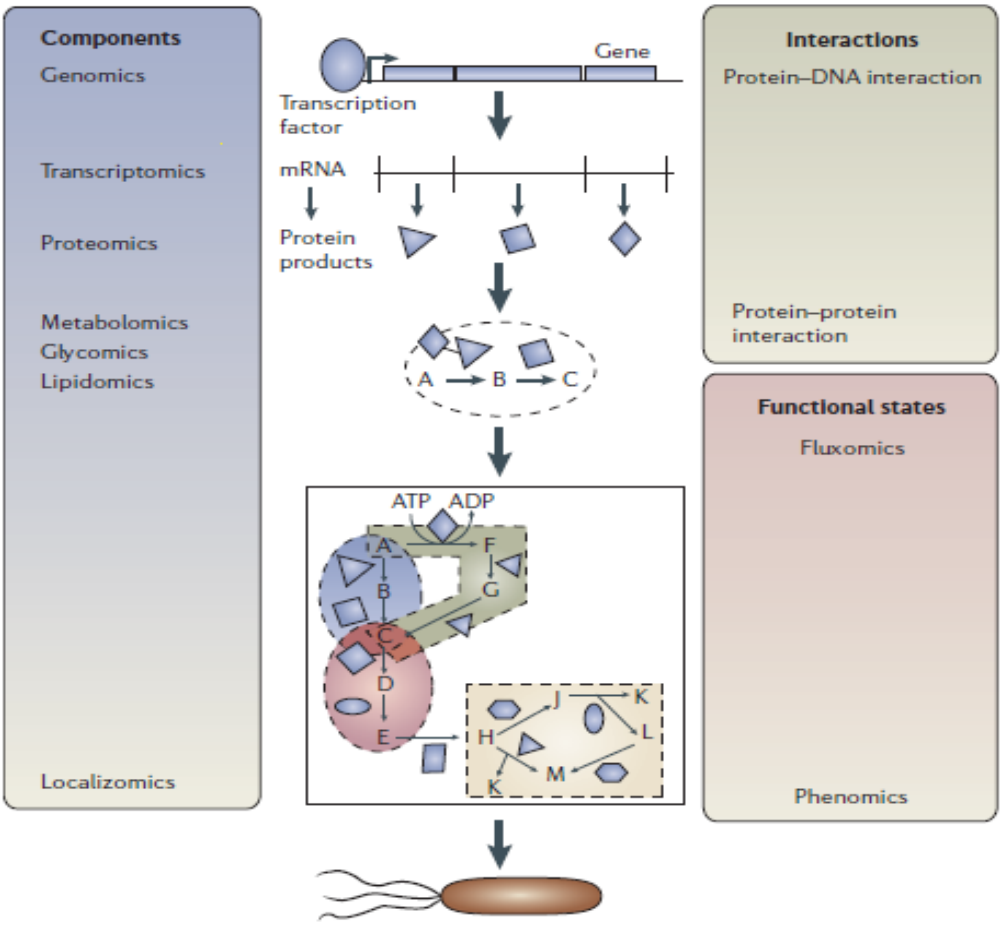



Figura 1- Alcune tra le principali scienze omiche (Adattata da [2])

Fino a pochi anni fa, il paradigma dominante nella rappresentazione del funzionamento di un sistema biologico era dettato da una visione gerarchica dall'alto verso il basso, secondo un'interpretazione unidirezionale del flusso di informazioni, dai geni, ai trascritti, alle proteine, le quali vanno infine ad incidere sulle vie metaboliche e dunque conducono a cambiamenti nel fenotipo dell'organismo [Figura 2(A)]. Da questo postulato si poteva dedurre che l'identificazione della sequenza genica di un sistema biologico sarebbe stata da sola sufficiente per prevedere le principali caratteristiche funzionali.[6]

Tuttavia questa visione lineare dell'informazione biologica non è più valida, in quanto è stato ampiamente compreso che i processi cellulari sono in realtà strettamente connessi tra loro attraverso meccanismi retroattivi [Figura 2(B)].

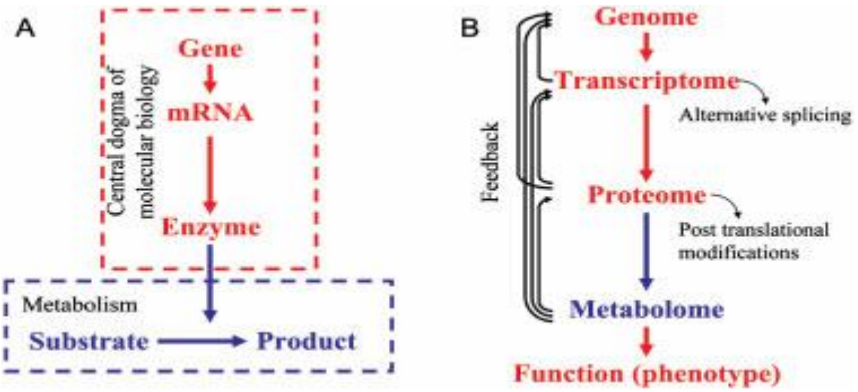



**Figura 2-(A) Dogma centrale della biologia molecolare, in cui il flusso di informazioni procede in maniera gerarchica; (B) schema generale secondo le scienze omiche. (Adattata da [6])**

Questo appare ovvio dal momento che un sistema vivente si modifica continuamente in relazione allo stato fisiologico e alle interazioni con l'ambiente esterno, in risposta a stimoli chimici e fisici e a modifiche geniche, determinando la variazione o l'insorgere di nuovi meccanismi. Un sistema biologico, come sistema complesso, può trovare vie alternative per i processi che determinano i flussi di materia ed energia, ovvero esercita un controllo sul suo funzionamento attraverso non una struttura gerarchica rigida, ma attraverso una struttura organizzata con interconnessioni tra il patrimonio genetico, proteico e metabolico, in maniera globale e flessibile.

Grazie alla metabolomica, è possibile dunque descrivere il profilo chimico in termini delle piccole molecole presenti in cellule, tessuti, organi e fluidi biologici. Le sue componenti (i metaboliti, che vanno a costituire il metaboloma) possono essere viste come il prodotto finale dell'espressione genica o dell'attività proteica (enzimi)[7], definendo così il fenotipo biochimico di un sistema biologico nel suo insieme, compreso l'uomo.

Il metaboloma è l'insieme quantitativo di tutte le molecole a basso peso molecolare (tipicamente inferiore a 3000 Da) presenti nelle cellule in un particolare stato fisiologico o evolutivo.[6] Sebbene il metaboloma e quindi la metabolomica siano certamente complementari alla trascrittomica e alla proteomica, gli studi metabolomici hanno dimostrato alcuni importanti vantaggi che hanno reso questa scienza omica, nata molto dopo le altre, sempre più popolare, come è visibile dal sorprendente aumento di pubblicazioni su dati di metabolomica avvenuto nel primo decennio del ventunesimo secolo[8]. I motivi di questo crescente successo

sono diversi, a partire innanzitutto dalla possibilità di riconoscere e quantificare moltissimi metaboliti direttamente da sistemi biologici complessi con accuratezza e precisione. In più, il metaboloma è l'ultimo nella scala biologica che va dai geni alla loro funzione ed è quindi il più adatto a descrivere le attività della cellula a livello funzionale: qualsiasi variazione dovuta a stimoli fisiopatologici viene amplificata nel metaboloma, il che comporta un notevole aumento di sensibilità, oltre al fatto che la risposta a questi stimoli avviene in maniera estremamente rapida, molto più velocemente rispetto al genoma o al proteoma. Questo perchè i flussi metabolici, e quindi anche le quantità stesse di metaboliti, sono regolati non solo dall'espressione genica ma anche da meccanismi post-trascrizionali e post-traduzionali influenzati da fattori ambientali, stati patofisiologici, microbioma intestinale, xenobiotici, perciò il metaboloma può essere considerato il più vicino al fenotipo[6].

In altre parole, mentre la genomica e la proteomica suggeriscono un possibile modo di funzionamento del sistema, la metabolomica dà la rappresentazione reale del sistema, ed è quindi indispensabile per monitorare la risposta biochimica degli organismi dovuta a fattori esterni, quali farmaci, malattie, dieta, e così via.

La maggior parte degli studi di metabolomica viene condotta utilizzando biofluidi comuni come urine, siero e plasma, facilmente ottenibili da mammiferi, in particolare dagli umani, in una maniera non invasiva essendo questi facilmente reperibili e utilizzati in molte altre analisi biologiche.

Fluidi secreti o escreti da un organismo vivente forniscono una panoramica unica del suo stato biochimico poiché la composizione di un determinato biofluido è diretta conseguenza della funzione delle cellule che sono riservate alla produzione di esso. Numerose informazioni e utili dettagli biochimici ad esempio sulla disfunzione di un organo, su una malattia o sulla tossicità di un farmaco, possono essere infatti ricavati dalla composizione di un particolare liquido.

Possono essere impiegati per le analisi anche molti altri fluidi, come saliva, sudore, condensato di respiro, bile, fluido cerebrospinale e così via. [9] Tuttavia di norma non sono utilizzati in quanto o le tecniche di estrazione sono molto invasive, oppure non si riescono ad ottenere informazioni soddisfacenti dai metaboliti contenuti.

Il numero di metaboliti presenti in questi biofluidi umani non è esattamente conosciuto, ma si stima essere comunque dell'ordine di poche migliaia, che confrontato con i circa 24000 geni e le centinaia di migliaia di proteine stimate apporta un altro notevole vantaggio alla metabolomica, riducendo drasticamente il numero di elementi da determinare.

## 1.2 Tecniche Analitiche in Metabolomica

Alla base di ogni studio di metabolomica che abbia successo c'è un insieme di dati di alta qualità che produce una fotografia a livello biochimico, la quale riflette lo stato di un organismo attraverso le sue piccole molecole endogene o esogene in quel determinato momento. Il tipo di informazione che la metabolomica cerca di raggiungere può essere suddiviso in due livelli, differenti ma complementari. Il primo di questi è fondamentalmente un risultato spettrale o cromatografico che riflette le variazioni complessive nella concentrazione dei metaboliti presenti, senza necessariamente identificare i componenti effettivi che stanno cambiando.[10] Questo tipo di dati può essere pensato come un'impronta digitale metabolica (*fingerprint*)[8] ed è semplicemente una rappresentazione numerica della risposta analitica derivante dai componenti che costituiscono il campione. In teoria qualsiasi metodo analitico potrebbe fornire un *fingerprint* metabolico, ma sono pochi quelli che si rivelano altamente riproducibili e contemporaneamente riescono a garantire un elevato livello di informazioni.[11] Il secondo livello di interpretazione, frequentemente usato soprattutto nell'ambito della *drug discovery*, è dato dall'identificazione e la quantificazione di un elenco predefinito di metaboliti che possono essere collegati a vie metaboliche specifiche e, quindi, fornire *biomarkers* e/o informazioni meccanicistiche di un processo, approccio che viene definito *metabolic profiling*.[4,6,8,10]

Ci sono relativamente poche tecniche che sono in grado di fornire in maniera esauriente entrambi questi livelli di dettaglio , e indubbiamente predominanti tra queste sono i metodi analitici basati sulla spettroscopia di risonanza magnetica nucleare (NMR) e sulla spettrometria di massa (MS). Il grosso vantaggio di NMR e

MS è fondamentalmente dovuto alla capacità di entrambe queste tecniche di generare *patterns* spettrali riproducibili e ricchi di informazioni e di identificare direttamente componenti molecolari all'interno di campioni biologici complessi, oltre a poter determinare la struttura di metaboliti di interesse e l'abbondanza relativa e assoluta delle molecole.

La spettroscopia NMR è una tecnica non distruttiva che offre una risposta analitica riproducibile e lineare con un elevato *range* dinamico, e costituisce dunque un ottimo strumento sia per il *fingerprinting* che per il *profiling*. Un grande vantaggio degli approcci basati sull'NMR è che i biofluidi possono essere analizzati direttamente con una minima preparazione del campione, ed il costo richiesto per l'acquisizione di ciascuno spettro è piuttosto basso, tuttavia risulta una tecnica abbastanza insensibile.

Al contrario, la spettrometria di massa è una tecnica distruttiva ampiamente diffusa in metabolomica grazie alla sua elevata sensibilità. Per questo motivo, a differenza dell'NMR, usato principalmente per il *fingerprinting* di biofluidi, la MS viene più spesso impiegata come rivelatore altamente sensibile e selettivo per l'identificazione e la quantificazione di determinati metaboliti, normalmente utilizzata in combinazione con tecniche cromatografiche. I sistemi accoppiati con la MS più utilizzati impiegano gas cromatografia (GC), cromatografia liquida ad alta prestazione (HPLC), o elettroforesi capillare (CE) come tecniche di separazione iniziali.[10]

Nonostante l'ormai incontrastata prevalenza di NMR e MS, anche altri approcci possono essere utilizzati per eseguire analisi metabolomiche. Tra questi, importanti sono sicuramente le tecniche di spettroscopia vibrazionale come IR e Raman, che sono state rivalutate come utili strumenti per il *fingerprinting* metabolico.[6,8,10]

## 1.3 NMR in Metabolomica

Lo spin è una proprietà intrinseca introdotta dalla meccanica quantistica che caratterizza qualsiasi particella in grado di ruotare su se stessa, ed è pensabile quindi come una sorta di momento angolare, in analogia alla meccanica classica.

Ad esso è associato un numero quantico I, che è un numero intero per i bosoni e semintero per i fermioni: protoni, neutroni ed elettroni assumono dunque valori di spin uguali a ± ½.

Particelle con spin I presentano (2I+1) sottolivelli energetici normalmente degeneri detti stati di spin.

Nel caso di un nucleo, le proprietà di spin dei protoni e dei neutroni che lo compongono si combinano per definire lo spin totale del nucleo, il quale, essendo carico, ruotando genera un campo magnetico a cui è associato un momento magnetico $\mu = \gamma \hbar \sqrt{[I(I+1)]}$, dove γ è il rapporto giromagnetico caratteristico del nucleo in esame e ħ è la costante di Planck.

Quando il momento magnetico nucleare associato ad uno spin nucleare è posto in un campo magnetico esterno, i diversi stati di spin forniscono diverse energie potenziali magnetiche [Figura 3(A)]. In presenza di un campo magnetico statico che produce una piccola quantità di polarizzazione di spin, un segnale a radiofrequenza (RF) di frequenza appropriata può indurre una transizione tra stati di spin. Questo salto pone alcuni degli spin nel loro stato energetico più elevato [Figura 3 (B)].

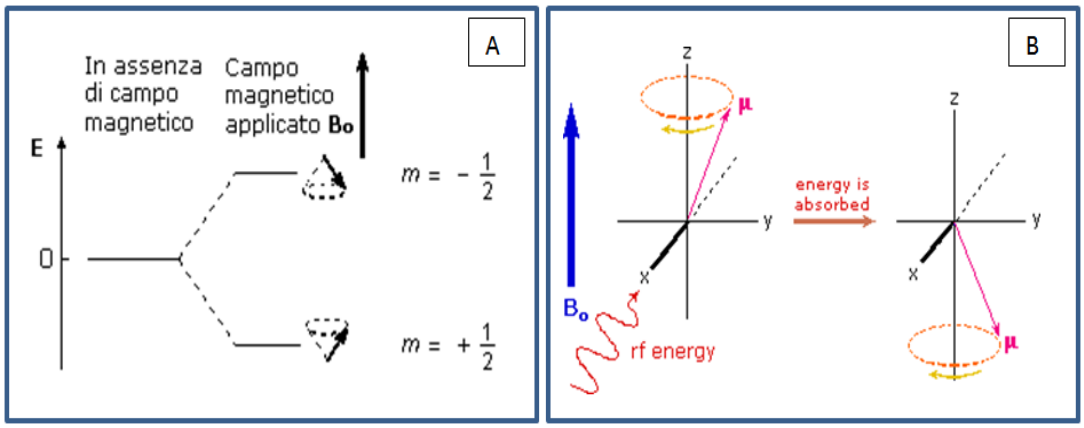



**Figura 3-Processo di assorbimento energetico che porta ad una transizione tra stati di spin nucleare, quando il nucleo in questione è sottoposto ad un campo magnetico esterno e ad un impulso di radiofrequenza.**

Se il segnale a radiofrequenza viene quindi spento, il rilassamento degli spin allo stato inferiore produce una quantità misurabile di segnale RF alla frequenza di risonanza associata con la transizione di spin nucleare. Questo processo è chiamato Risonanza Magnetica Nucleare (NMR).

Un momento magnetico μ posto in un campo magnetico esterno avrà una energia potenziale legata al suo orientamento rispetto a tale campo e tenderà a compiere un moto di rotazione intorno alla direzione del campo stesso con una frequenza tradizionalmente chiamata frequenza di Larmor, che dipende sia dall'intensità del campo sia dal rapporto giromagnetico γ.

La frequenza di Larmor può essere visualizzata classicamente in termini di precessione del momento magnetico intorno al campo magnetico, analoga alla precessione di una trottola intorno al campo gravitazionale. Può anche essere visualizzata quanto-meccanicamente in termini di energia quantica di transizione tra i due possibili stati di spin per spin=1/2 . Questa può essere espressa come energia fotonica secondo la relazione di Planck, per cui la differenza di energia potenziale magnetica è $h\nu_{Larmor} = 2\ \mu B$.

Ogni molecola contenente uno o più atomi con momento magnetico di spin nucleare μ non nullo è potenzialmente rilevabile mediante spettroscopia NMR e, poiché gli isotopi con momenti magnetici diversi da zero includono $^{1}H$, $^{13}C$, $^{14}N$, $^{15}N$, e $^{31}P$, tutte le molecole biologicamente importanti hanno almeno un segnale NMR. Questi segnali sono caratterizzati da frequenza (*chemical shift*), intensità, struttura fine, e proprietà di rilassamento magnetico, tutte caratteristiche che riflettono il preciso intorno chimico del nucleo rilevato. Perciò gli spettri NMR contengono molte informazioni circa l'identità delle molecole nel campione, ed è proprio per questo che l'NMR può essere usato per identificare e quantificare metaboliti in campioni di differente origine biologica [Figura 4].

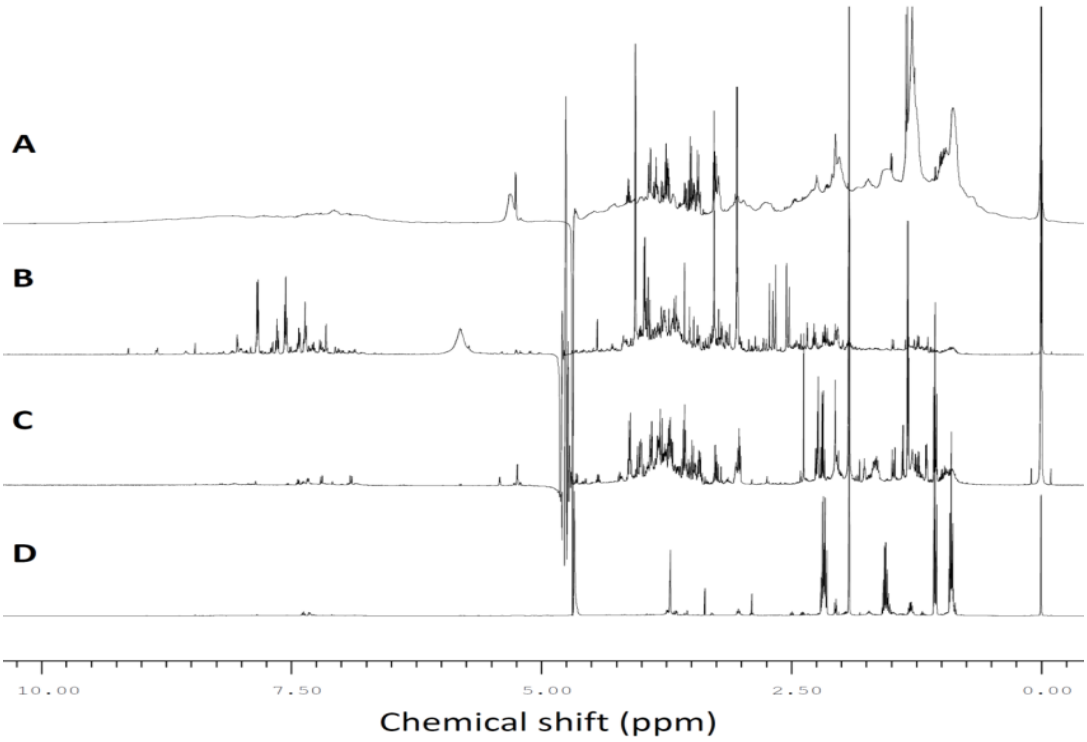



**Figura 4- Spettri $^{1}$H-NMR ottenuti dall'analisi di siero (A), urine (B), saliva (C), ed estratti fecali (D).**

Oltre a ciò, l'NMR è una tecnica molto versatile che presenta numerosi vantaggi. Per prima cosa, è un metodo non distruttivo, e possono essere acquisiti spettri di sospensioni cellulari, tessuti, piante, di metaboliti estratti e purificati. In particolare, lo sviluppo della tecnica NMR HR-MAS, per cui la rapida rotazione del campione ad un angolo di 54.7° (il cosiddetto "angolo magico") rispetto al campo magnetico applicato provoca la riduzione dello slargamento dei segnali e della perdita di informazione associata, ha reso sperimentalmente realizzabile l'analisi su interi pezzi di tessuto senza trattamento preliminare.

È inoltre possibile mediante NMR determinare la struttura di un nuovo metabolita, dimostrare l'esistenza di una particolare via metabolica in vivo, e localizzare la distribuzione di un metabolita in un tessuto. Infine, l'abbondanza naturale di alcuni isotopi magnetici biologicamente rilevanti è bassa e questo consente di utilizzare come *labels* questi isotopi, in particolare $^{2}$H, $^{13}$C, $^{15}$N, introducendoli nel sistema metabolico prima dell'analisi NMR. Ciò permette l'esplorazione di vie metaboliche, portando a informazioni qualitative sui legami tra precursori marcati e i loro prodotti e informazioni quantitative sui flussi metabolici.

L'isotopo magneticamente attivo indubbiamente più utilizzato nella maggior parte delle applicazioni di NMR per il *fingerprinting* e il *profiling* metabolico è l'$^{1}$H.

In $^{1}$H-NMR viene misurata la precessione dello spin protonico quando è sottoposto ad un campo magnetico esterno. Dal punto di vista pratico, un campione contenente protoni (nuclei di idrogeno) viene posto all'interno di un intenso campo magnetico per produrre polarizzazione parziale dei protoni. Viene anche imposto un forte impulso RF sul campione per eccitare alcuni degli spin nucleari nel loro stato di energia superiore. Quando questo forte segnale RF viene spento, gli spin tendono a tornare al loro stato inferiore, producendo una piccola quantità di radiazione alla frequenza di Larmor associata a tale campo. L'emissione di radiazione è associata al rilassamento degli spin dei protoni dal loro stato eccitato . Si induce un segnale a radiofrequenza in una bobina del rivelatore che viene amplificato per visualizzare il segnale NMR. Il segnale raccolto è oscillante con frequenza v, la frequenza di Larmor per il nucleo in esame, si smorza nel tempo e viene detto FID (Free Induction Decay), libero decadimento dell'induzione.

Se il campione contiene nuclei con differenti frequenze di risonanza, questi vengono tutti eccitati contemporaneamente dall'impulso RF e quindi il segnale raccolto sarà una curva complessa, chiamata interferogramma, data dalla combinazione di più FID, uno per ogni frequenza assorbita dai nuclei.

Per poter risalire alle singole frequenze che combinandosi tra loro hanno generato il segnale complesso, è necessario applicare una procedura matematica detta Trasformata di Fourier che permette di passare dal grafico in funzione del tempo, il FID, al grafico in funzione delle frequenze, lo spettro NMR [Figura 5].

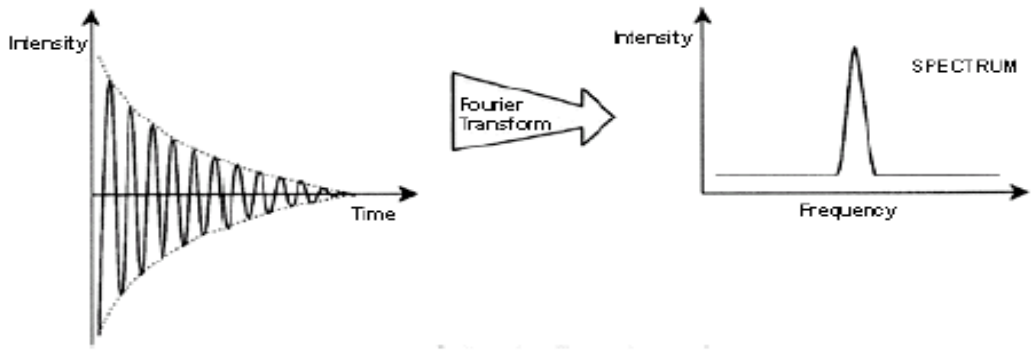



**Figura 5- Passaggio dal dominio del tempo (FID) a quello delle frequenze (spettro NMR) grazie alla Trasformata di Fourier**

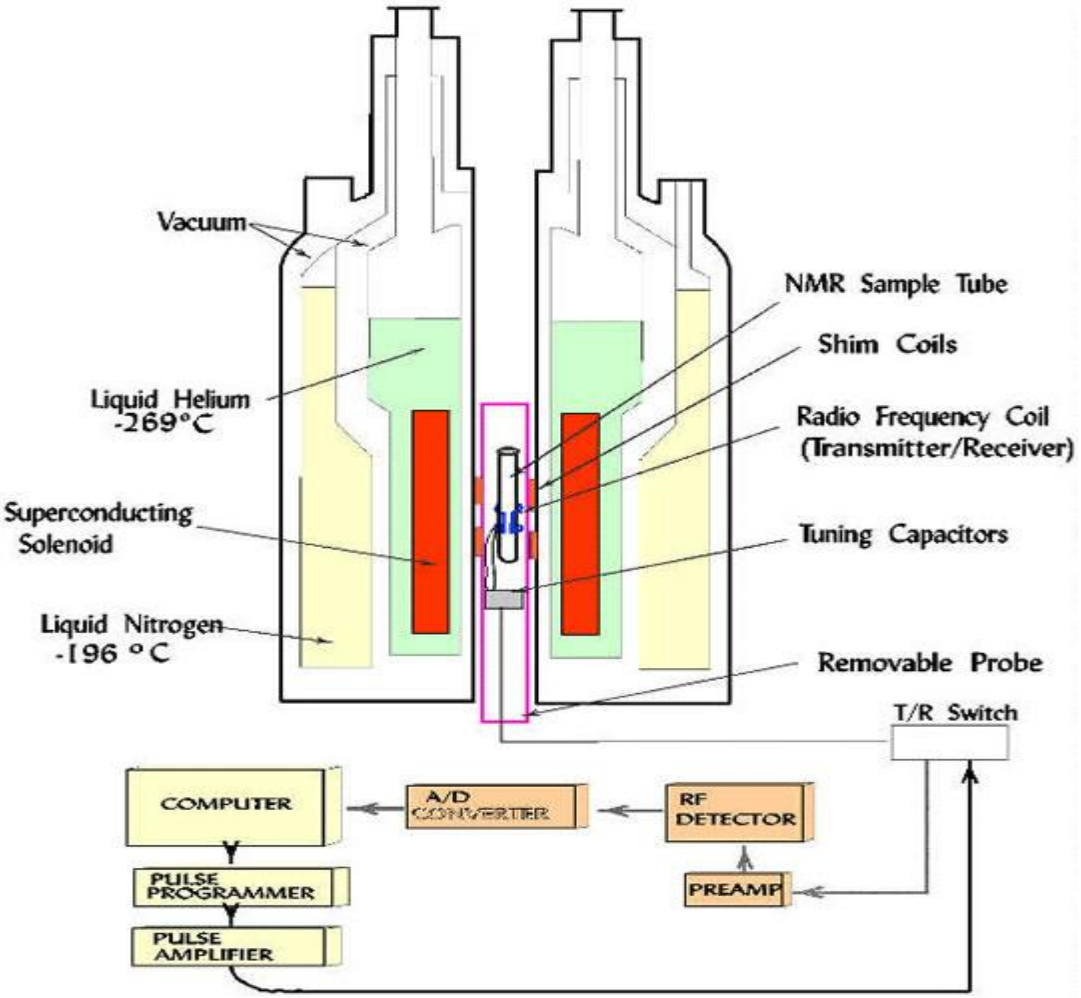



Figura 6- Sezione di uno spettrometro NMR

Per $^{1}$H-NMR , la soglia di concentrazione per il rilevamento di un metabolita in un estratto utilizzando un moderno spettrometro ad alto campo magnetico è circa 10 µM , corrispondente ad una quantità di 5 nmol nel volume di campione (tipicamente 500 µL). In pratica, la sensibilità ottenibile è fortemente dipendente dalla forza del campo magnetico e dalla progettazione del *probehead* che permette di rilevare i segnali. Sono disponibili spettrometri NMR con intensità di campo fino a 21 Tesla, corrispondente ad una frequenza $^{1}$H-NMR di 900 MHz, ma l'analisi metabolomica viene fatta con strumenti più comunemente disponibili che operano nella gamma 300-600 MHz . Poiché l'incremento della forza di campo aumenta anche la risoluzione spettrale, riducendo il numero di segnali sovrapposti nello spettro, sono gli spettrometri all'estremità superiore di questa gamma di frequenze i più efficaci e i più usati per il *profiling* metabolico mediante $^{1}$H-NMR.[10]

Inoltre, recenti sviluppi tecnologici, in particolare l'introduzione di *cryoprobes*, dove la bobina del rivelatore e il pre-amplificatore sono raffreddati a 20K, e di *microcoil probes*[12], hanno migliorato notevolmente la sensibilità della tecnica[10].

Oltre alla sensibilità non elevata, l'$^{1}$H-NMR è affetta dal problema che la dispersione dei segnali nello spettro è piuttosto piccola, e questo comporta un'estesa sovrapposizione dei picchi in molte regioni spettrali. Questo svantaggio è poco importante se l'approccio seguito è quello di indagare le variazioni del *fingerprint* complessivo, mentre risulta determinante nel caso si voglia riconoscere e quantificare specifici metaboliti.

Una pur parziale soluzione consiste nell'utilizzare tecniche basate su esperimenti NMR bidimensionali, che aumentano la risoluzione spettrale distribuendo i segnali lungo due assi di frequenza . Questi esperimenti sfruttano le interazioni tra gli isotopi rilevabili in una molecola, ed è possibile ottenere sia correlazione omonucleare, dove i due assi di frequenza dello spettro corrispondono al medesimo nucleo, di solito $^{1}$H, sia correlazione eteronucleare, se un'asse di frequenza corrisponde all'$^{1}$H e l'altro corrisponde a $^{13}$C, $^{15}$N o, occasionalmente, $^{31}$P. Esempi di esperimenti bidimensionali omonucleari $^{1}$H-$^{1}$H sono il J-res, utile per attenuare i segnali di macromolecole e per ottenere informazioni ulteriori su molteplicità e costanti di accoppiamento, il COSY e il TOCSY, dai quali si ricavano indicazioni sulle relazioni tra i protoni; tra quelli eteronucleari invece il più usato è il $^{13}$C-$^{1}$H HSQC, che permette di identificare i protoni che sono legati ad un certo carbonio e viceversa.

Manipolare i segnali NMR per produrre uno spettro bidimensionale richiede più tempo ed elaborazioni più complesse rispetto ad un esperimento monodimensionale semplice, per questo in metabolomica l'utilizzo di spettri a due dimensioni è poco diffuso e riservato solamente a particolari necessità.

Infine, la dipendenza dal pH del *chemical shift* di protoni vicini a gruppi ionizzabili può costituire una limitazione per l'NMR come strumento di *fingerprinting*. L'utilizzo di strumenti informatici, come algoritmi di allineamento oppure la procedura del *bucketing*, minimizza le differenze dovute a piccole variazioni di pH o di forza ionica.

## 1.4 Metodi Chemiometrici

Da uno studio di metabolomica viene ottenuta una grossa quantità di dati che viene analizzata tramite tecniche di analisi chemiometrica.

La chemiometria è una branca della chimica analitica rivolta all'applicazione di metodi matematici e statistici per gestire, interpretare e predire dati chimici.[13] In particolare, i modelli chemiometrici più usati nell'indagine metabolomica sono quelli di classificazione, di *modeling* e di regressione dell'analisi statistica multivariata, in quanto è necessario utilizzare più variabili per caratterizzare i sistemi in esame. Le misurazioni possono essere disposte in una matrice di dati, in cui ogni riga costituisce un'osservazione (ad esempio ciascun campione, un esperimento, un intervallo di tempo) e le colonne rappresentano le variabili misurate (ad esempio lunghezza d'onda, numero di massa, c*hemical shift*, con le relative intensità). Questo processo genera un insieme di dati enorme e complesso, che è difficile da riassumere e manipolare senza gli strumenti adeguati.[14]

Tali strumenti sono forniti proprio dalla statistica multivariata che, a partire da questo set multidimensionale di coordinate metaboliche, permette di realizzare con ottimi risultati i principali scopi dell'analisi metabolomica, ovvero:

- esaminare complessivamente le differenze globali, le tendenze nelle variazioni e le relazioni che intercorrono tra campioni e variabili;
- determinare se i campioni esaminati tendono o meno a dividersi in *clusters*, in gruppi evidentemente distinti (ad esempio sani/malati);
- mettere in evidenza i metaboliti maggiormente responsabili di tali differenze;
- realizzare modelli predittivi per nuovi campioni.

I metodi di visualizzazione per evidenziare le differenze tra campioni e/o tra variabili possono essere divisi in due gruppi: *unsupervised* e *supervised*. I metodi *unsupervised* sono utilizzati per l'esplorazione preliminare dei dati e il loro scopo è quello di fornire una visualizzazione complessiva dei dati, riducendo le variabili e cercando di massimizzare la varianza tra di essi, senza però fornire informazioni basate su una conoscenza a priori dei dati per guidare l'analisi.

Il metodo *unsupervised* più utilizzato è sicuramente l'analisi delle componenti principali (PCA) che consente di valutare le correlazioni tra le variabili e la loro rilevanza, visualizzare gli oggetti, individuando l'eventuale presenza di *outliers* e di *clusters*, sintetizzare la descrizione dei dati, eliminando rumore o informazione spuria, e ridurne la dimensionalità. La PCA consiste in un processo di rotazione dei dati originali definiti da una matrice *X* di dimensione *n* x *p*, effettuato in modo che il primo nuovo asse sia orientato nella direzione di massima varianza dei dati, il secondo sia perpendicolare al primo e sia nella direzione della successiva massima varianza dei dati, e così di seguito per tutti i *p* nuovi assi. Visto che le prime componenti sono le più significative in quanto spiegano la maggiore varianza tra i dati, ogni campione può essere rappresentato da relativamente poche componenti, che sono combinazioni lineari delle variabili originali, piuttosto che da migliaia di variabili. Questo processo dà origine a due nuove matrici: la matrice dei *loadings* e quella degli *scores*. Quella dei *loadings* è la matrice le cui colonne rappresentano gli autovettori della matrice di covarianza (o di correlazione); le righe rappresentano le variabili originali: ciò significa che, selezionato un autovettore, in ciascuna riga troviamo i coefficienti numerici che rappresentano l'importanza di ciascuna variabile originale in quell'autovettore. In altre parole, il grafico dei *loadings* consente di analizzare il ruolo di ciascuna variabile nelle diverse componenti, le loro correlazioni dirette e inverse, la loro importanza. Gli *scores* esprimono le coordinate per ciascun campione nel nuovo sistema di riferimento, ed il relativo grafico permette di visualizzare il comportamento degli oggetti nelle diverse componenti e le loro similarità, ovvero di individuare raggruppamenti di oggetti simili (*clusters*), la presenza di oggetti particolari (*outliers*), il manifestarsi di particolari regolarità e distribuzioni, ed è perciò il più importante per l'indagine preliminare dei dati ottenuti. [Figura 7]

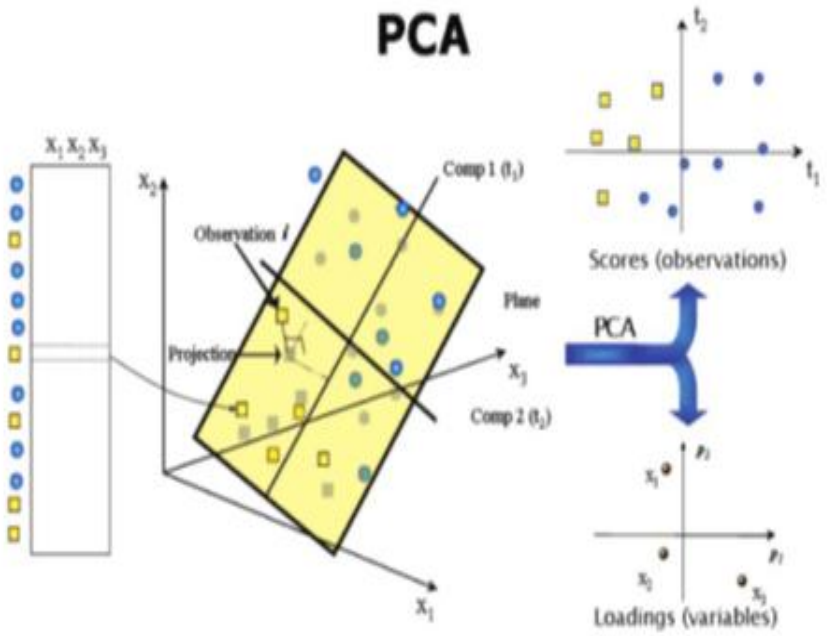



**Figura 7- Esempio di analisi delle componenti principali in un set tridimensionale di dati (tre variabili), con relativi grafici degli scores e dei loadings.**

Anche i metodi *supervised* vengono utilizzati per visualizzare l'esistenza di raggruppamenti e similarità tra dati, in questo caso però il sistema viene istruito con informazioni aggiuntive, ad esempio il numero e il tipo di classi che si vogliono individuare. Il più usato tra questi metodi è quello di regressione parziale con minimi quadrati (PLS), che correla la matrice *X* delle variabili indipendenti (la matrice dei dati) con una matrice *Y* che contiene le variabili dipendenti, ad esempio delle informazioni sulla natura dei campioni (sani/malati, maschi/femmine, e così via). In altre parole la PLS ricava delle nuove variabili (analoghe alle componenti principali, in questo caso più spesso chiamate variabili latenti) che massimizzano la covarianza tra i dati in *X* e le informazioni contenute in *Y*.

Inoltre, sia i metodi *supervised* che quelli *unsupervised* sono frequentemente applicati in combinazione con altre procedure tipiche dell'analisi multivariata, con lo scopo di aumentare la separazione tra i gruppi di campioni. In particolare, un metodo molto usato è quello dell'analisi canonica (CA) che consente di studiare le correlazioni tra i due blocchi di variabili *X* e *Y*, ma a differenza dei metodi di regressione non presume alcuna relazione causa-effetto tra la matrice X e la matrice Y.

A questo punto uno dei punti cruciali dell'analisi metabolomica è riuscire a costruire modelli robusti che valgano anche per nuovi dati e consentano cioè di predire se un campione appartenga o meno ad un determinato gruppo: per questo scopo sono necessari metodi di classificazione. Il più semplice di questi metodi è il k-NN (k-Nearest Neighbors), che si basa sulla scelta di una distanza, generalmente la distanza euclidea, e sulla selezione di numero intero di k intorni (gli oggetti più vicini ad ogni oggetto da classificare) ai quali si estende la valutazione delle classi cui essi appartengono al fine di collocare l'oggetto considerato con un criterio di assegnazione maggioritario. Come si può vedere anche dalla Figura 8, la scelta del numero k di intorni è sempre fondamentale poiché spesso si rivela cruciale nell'assegnare un campione (contrassegnato da una stella) ad una classe o ad un'altra.

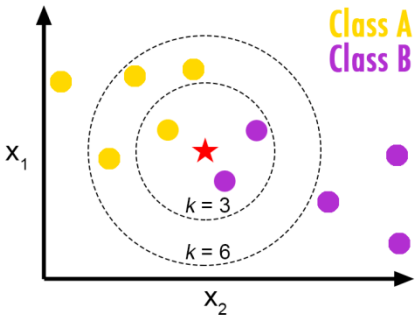


Figura 8- Metodo di Classificazione k-NN

## 1.5 Applicazioni della Metabolomica

La metabolomica ha un grande numero di applicazioni differenti, in quanto, grazie all'utilizzo di strumenti statistici (per l'analisi di profili metabolici) e bioinformatici (per la creazione di *database* e per l'annotazione e il trattamento di dati), può essere utilizzata come potente strumento per la caratterizzazione dei sistemi viventi a livello biochimico. Prima di tutto, essendo incentrata sullo studio delle alterazioni metaboliche e dell'omeostasi a livello dell'intero organismo, essa svolge un ruolo fondamentale nel comprendere l'integrità funzionale del sistema, consentendo di individuare l'insorgere di processi patologici, osservare la presenza di mutazioni genetiche, verificare la tossicità di un farmaco[15].

Analizzando le differenze metaboliche tra sistemi perturbati e non perturbati, ad esempio tra volontari sani e pazienti con una malattia, si può arrivare ad avere una comprensione molto più profonda di quest'ultima[16], risalendo persino al

meccanismo di sviluppo, ai flussi metabolici coinvolti, all'eventuale scoperta di specifici *biomarkers*.

Finora la metabolomica è stata in grado di fornire informazioni significative su una vasta gamma di patologie come cancro, diabete, celiachia, malattie cardiovascolari, disturbi neurologici, respiratori, intestinali.[17]

Applicata nel campo alimentare, la metabolomica consente di creare un'impronta molecolare in grado di rappresentare fedelmente la varietà di un prodotto alimentare che può avere valore diagnostico o comunque di classificazione. Per questo motivo, questa scienza offre un'eccezionale opportunità di studiare molti aspetti legati agli alimenti, tra cui l'analisi delle componenti molecolari, la qualità e la rilevazione di autenticità.

## 1.6 Il Fenotipo Metabolico

Una delle applicazioni più importanti e più rivoluzionarie della metabolomica è quella che riguarda la determinazione dell'esistenza di un fenotipo metabolico individuale, basato sulla teoria per cui l'insieme dei metaboliti presenti all'interno di uno stesso biofluido fornisce un'impronta digitale (*fingerprint*), che è caratteristica per ciasciun soggetto e procura informazioni sullo stato fisiologico del soggetto stesso. Tra gli obiettivi più interessanti della metabolomica c'è infatti quello di arrivare a spiegare completamente la natura della relazione che intercorre tra i polimorfismi genetici dei diversi individui e il loro *fingerprint* metabolico, allo scopo di comprendere chiaramente la risposta dei differenti organismi agli stimoli esterni.

In passato sono state osservate differenze sperimentali in profili metabolici dovute a differenze di ceppi genetici in due tipi di ratto, cosa che ha suggerito l'esistenza di diversi "*metabotypes*", definiti come "la descrizione multiparametrica di un organismo in un determinato stato fisiologico basato su dati metabolomici"[18]. Essi possono perciò essere visti come una fotografia degli stati stazionari che definiscono l'omeostasi, il cui mantenimento, che è alla base della vita di ogni essere vivente, è regolato da numerosi cicli biochimici e meccanismi cellulari e molecolari: l'immenso potenziale della metabolomica sta quindi nel fornire tali

immagini ed infine dare informazioni sulle deviazioni dalle condizioni ottimali, che possono essere direttamente collegate alla presenza di uno stato patofisiologico.
La presenza di un fenotipo metabolico unico per ciascun individuo o gruppo di individui è stata dunque da tempo ipotizzata, ma solo nell'ultimo decennio sono state raccolte evidenze sperimentali grazie ad analisi complete e sistematiche di biofluidi, in particolare utilizzando urine [19,20,21].
L'identificazione di un *metabotype* caratteristico per ogni soggetto potrebbe rivelarsi estremamente importante in molti campi, come la nutrigenomica e la farmacologia, introducendo innovativi miglioramenti nella pianificazione di terapie e diete personalizzate, nella predizione e nella valutazione sull'efficacia o sulla tossicità di farmaci, nella prognosi e nella diagnosi di malattie. Ovviamente, una condizione essenziale per la loro utilità è che questi si mantengano praticamente inalterati nel tempo; inoltre, il fatto che i profili metabolici sperimentali siano differenti non solo per motivi genetici ma anche per fattori legati all'ambiente esterno, allo stile di vita, all'età, allo stato nutrizionale e di salute, fa sì che i continui cambiamenti giornalieri, siano essi casuali (per esempio legati alla dieta) oppure dovuti a variazioni biochimiche cicliche, costituiscano una fonte di rumore non trascurabile che incide sulla ricerca sistematica di un *fingerprint* metabolico invariante per un dato soggetto. Grazie alla raccolta di numerosi campioni e all'analisi metabolomica è stato possibile dimostrare che questo esiste e che si mantiene relativamente stabile nel tempo, anche a distanza di anni[20], ma ulteriori studi devono necessariamente essere compiuti affinché si raggiunga una caratterizzazione completa e definitiva dei *metabotypes*.

## 1.7 Analisi Metabolomica di Saliva

La saliva è un importante e complesso fluido biologico dalle molte funzioni, come la lubrificazione della cavità orale, la digestione preliminare del cibo, e la protezione da microorganismi, e viene prodotta da molteplici ghiandole salivari, in particolare dalle tre coppie di ghiandole salivari maggiori (parotide, sottomandibolare e sottolinguale).

La secrezione di saliva è disciplinata dal sistema nervoso autonomo, la composizione è molto variabile e dipende in particolare dalla presenza o meno di stimolazione nella produzione.[22]

La saliva è costituita da quasi il 99% di acqua, mentre la restante parte è composta da elettroliti (in particolare cloruri, carbonati, bicarbonati, fosfati, solfati, tiocianati e cationi come $Na^+$ e $K^+$), muco, acidi nucleici, metaboliti, glicoproteine e proteine tra cui numerosi enzimi. Inoltre contiene fluido crevicolare gengivale e tutti i componenti derivanti dalla cavità orale, compresi batteri e residui di cibo.[23]

La saliva umana costituisce una fonte molto interessante di *biomarkers,* grazie alla sua facilità di campionamento e al fatto che il suo contenuto riflette alcune delle molecole presenti nel sangue, per cui viene considerata utile per la diagnosi di varie malattie, disturbi endocrini e per la valutazione consumo di droga.[24]

Ad oggi, la composizione proteica della saliva è stata ampiamente studiata, ma solo nel corso degli ultimi anni è stata posta maggiore attenzione su quella metabolica ed alcuni tentativi sono stati fatti per collegare il *fingerprint* metabolico della saliva con situazioni patologiche.[25,26] Diversi metaboliti della saliva sono stati identificati con successo[27], ed è stata studiata la variabilità tra diversi soggetti[28], tuttavia le potenzialità della saliva per gli studi di metabolomica non sono ancora state esplorate completamente e la letteratura pubblicata a riguardo è molto limitata.

# 2 ORGANIZZAZIONE E SCOPO DEL LAVORO

Il presente lavoro si inquadra all'interno di un più ampio progetto di studio condotto dal gruppo di ricerca di Metabolomica afferente al Professor Claudio Luchinat presso il Centro di Risonanze Magnetiche – CERM dell'Università degli Studi di Firenze, in collaborazione con la Professoressa Sandra Wallner-Liebmann del Centro di Medicina Molecolare presso l'Istituto di Fisiopatologia e Immunologia dell'Università Medica di Graz. Tale progetto prevede l'analisi di numerosi fluidi corporei (urine, saliva, sudore, condensato di respiro, siero, plasma, feci), tutti raccolti da un gruppo di volontari, sia maschi che femmine, in un arco di tempo di 10 giorni, organizzato come riportato in Figura 9.

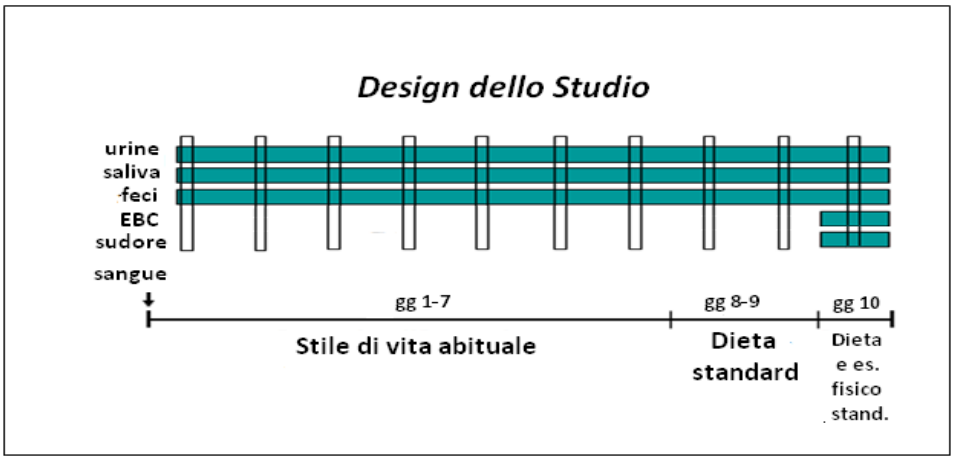



Figura 9 - Organizzazione del progetto

Lo scopo del progetto è quello di esaminare le variazioni inter ed intraindividuali in profili metabolici umani, con particolare attenzione agli effetti della dieta standardizzata e dell'esercizio fisico, per approfondire e scoprire aspetti ancora inesplorati del metaboloma umano, delle sue interazioni con l'ambiente esterno e delle cause delle sue alterazioni, attraverso l'analisi dettagliata di biofluidi diversi, che riflettono differentemente queste variazioni.

In questo lavoro di tesi presentiamo i risultati ottenuti dallo studio dei campioni di saliva, analizzati mediante spettroscopia NMR, con l'obiettivo di rivalutare la saliva come oggetto di interesse ricco di informazioni metaboliche, con uno studio

sistematico e approfondito che aiuti a chiarire le potenzialità ancora inesplorate di tale biofluido per la metabolomica.

L'individuazione di fenotipi metabolici ("*metabotypes*") è già stata ampiamente oggetto di studio su biofluidi come le urine[19]. Con questo lavoro si vuole dimostrare prima di tutto che si può osservare la presenza di un fenotipo metabolico caratteristico per ogni soggetto anche in campioni di saliva umana, e quindi, una volta stabilito ciò, valutarne le variazioni inter- e intra- individuali in un arco di tempo di pochi giorni e secondo diverse condizioni esterne, come la differente risposta ai cambiamenti nell'alimentazione o all'introduzione di sforzi fisici.

La pesante influenza della dieta sui *metabotypes* è già stata analizzata su campioni di urine[29], mentre ancora non risulta ben approfondita la risposta metabolica in seguito all'attività fisica.

Le informazioni contenute negli spettri NMR ottenuti (1230 in tutto) sono state analizzate da tre diversi punti di vista:

- *Fingerprinting* interindividuale: attraverso metodi chemiometrici si cerca di determinare la presenza di un riconoscimento individuale, o comunque legato a fattori caratteristici di un gruppo di soggetti (ad esempio il genere), che possono influenzare e definire il fenotipo metabolico.
- *Fingerprinting* intraindividuale: con gli stessi metodi vengono analizzate anche le variazioni nel tempo, cioè quelle più strettamente connesse a fattori esterni come dieta ed esercizio fisico.
- *Profiling* metabolico: un cospicuo gruppo di metaboliti viene identificato negli spettri NMR, ed i relativi segnali vengono analizzati al variare di tutti i parametri esaminati precedentemente, allo scopo di desumere eventuali molecole che abbiano differenze significative in concentrazione o tra i diversi soggetti o nel tempo.

L'intento generale è dunque quello di affermare l'esistenza di un *fingerprint* metabolico individuale, influenzato da fattori esterni quali la dieta o lo stile di vita, ma intrinsecamente presente nei campioni di saliva a prescindere da questi, ed infine mettere in luce alcuni tra i metaboliti responsabili delle variazioni inter ed intraindividuali tra i profili metabolici.

# 3 MATERIALI E METODI

## 3.1 Design dello Studio

Per questo studio sono stati reclutati 23 volontari sani, 17 di sesso femminile e 6 di sesso maschile, di età compresa tra i 22 e i 57 anni (media 32.1 anni), con un indice di massa corporea tra 18 e 30 $kg/m^2$. I criteri di esclusione utilizzati nella selezione dei soggetti includono l'uso corrente di qualsiasi farmaco o regolare terapia, la partecipazione ad un altro studio entro i 3 mesi precedenti, una malattia acuta entro le 2 settimane precedenti l'inizio dello studio. Inoltre sono stati esclusi i soggetti qualora presentassero alterazioni significative dal punto di vista clinico, ed è stato proibito l'eccessivo consumo di alcol, con un apporto massimo settimanale non superiore a 28 unità.

Ciascun volontario ha fornito tra i 4 e i 6 campioni al giorno per un totale di dieci giorni. Dal punto di vista dell'alimentazione, per i primi 7 giorni ciascun soggetto ha potuto scegliere liberamente la propria, mentre una dieta standardizzata è stata introdotta per gli ultimi 3; il decimo e ultimo giorno è stato previsto anche lo svolgimento di esercizio fisico, monitorato con l'uso di contapassi e standardizzato secondo protocolli pubblicati.

## 3.2 Raccolta e Conservazione dei Campioni

Sono stati raccolti 1230 campioni (in media 55 campioni per ciascun individuo) di circa 1 mL di saliva direttamente in *vial* criogeniche, ed immediatamente congelati a -80°C.

Ogni volontario è stato invitato a non forzare la salivazione, mentre per quanto riguarda la raccolta del primo campione della giornata, sono stati vietati sia l'assunzione di cibo e bevande sia lavaggi o risciacqui per l'igiene orale.

## 3.3 Preparazione dei Campioni per l'Analisi NMR

I campioni sono stati scongelati a temperatura ambiente e centrifugati (14000 rpm per 30 minuti a 4°C) per rimuovere cellule, muco e altre macromolecole. Sono stati quindi prelevati 540 µL di surnatante ed immediatamente aggiunti a 60 µL di buffer fosfato (composizione: 0.2 M $Na_2HPO_4$, 0.2 M $NaH_2PO_4$, 30 mM $NaN_3$, e 10 mM (TSP sodio trimetilsilil [2,2,3,3-$^2H_4$] propionato) in 100% $^2H_2O$; pH 7.0).

$NaN_3$ (sodio azide) è stato aggiunto come conservante per assicurare che non venissero generati o consumati metaboliti a causa di batteri presenti nella saliva durante il tempo di preparazione dei campioni o di acquisizione degli spettri NMR.

Da ognuna di queste miscele è stata prelevata un'aliquota di 450 µL e trasferita in un tubo NMR (diametro esterno: 4.25 mm) per l'analisi.

## 3.4 Acquisizione degli Spettri NMR

Per ogni campione di saliva è stato acquisito uno spettro monodimensionale $^1H$-NMR con sequenza NOESY-presat (terminologia Bruker: noesygppr1d.comp), 128 scansioni, *receiver gain* 11.3, 65536 *data points*, larghezza spettrale 12019 Hz, tempo di rilassamento 4 s e *mixing time* 0.1 s.

Tutti gli spettri sono stati acquisiti utilizzando uno spettrometro Bruker 600 MHz operante alla frequenza di Larmor per il protone (600.13 MHz) e dotato di *cryoprobe* CPTCI, un'unità di *tuning-matching* automatica (ATM) e un campionatore automatico. Tramite una termocoppia PT 100 viene stabilizzata la temperatura a livello di approssimativamente ±0.1 K al campione. Prima dell'acquisizione, i campioni vengono conservati per almeno 3 minuti all'interno del *probehead* dello strumento, per equilibrare la temperatura a

**Figura 10- Spettrometro NMR Brucker 600 MHz operante a 14.1 T**

quella di esercizio, che nel caso di campioni di saliva è 300.0 K.

## 3.5 Processing degli Spettri NMR

Le Free Induction Decays (FIDs) sono state moltiplicate per una funzione esponenziale equivalente ad un allargamento di 1 Hz, prima dell'applicazione della trasformata di Fourier. Gli spettri trasformati sono stati corretti manualmente per quanto riguarda la fase e le distorsioni della linea di base e calibrati rispetto al picco a 0.00 ppm del TSP usando *TopSpin* (*Bruker Biospin*).

Ogni spettro monodimensionale nella regione compresa fra 0.15 e 4.5 ppm e in quella tra 6 e 10 ppm (escludendo quindi la zona intorno al segnale dell'acqua) è stato suddiviso in *bins* (ovvero intervalli di *chemical shift*) di 0.02 ppm e le corrispondenti aree spettrali sono state integrate usando il software AMIX (versione 3.8.4; *Bruker BioSpin*).

Per scalare i dati è stato utilizzato l'algoritmo PQN (*Probabilistic Quotient Normalization*) che rappresenta un'alternativa più accurata e robusta rispetto al ben più usato *scaling* sull'area totale, poiché risulta in grado di interfacciarsi bene anche con importanti variazioni di diluizione e concentrazione di metaboliti[30], come nel caso della saliva.

Tutti i metaboliti di interesse sono stati assegnati basandosi su spettri monodimensionali già assegnati in letteratura e sui *database* di riferimento, in particolare HMDB (*Human Metabolome DataBase*) e BIOREFCODE (*Bruker BioSpin*).

## 3.6 Analisi Statistica

La matrice di dati ottenuta è stata analizzata mediante metodi chemiometrici, applicati utilizzando R, un *software open source* costituito da un insieme di macro, librerie, oggetti che possono essere utilizzati per la gestione e l'analisi dei dati e la produzione di grafici.[31]

Per ottenere una comprensione generale della varianza dei profili NMR, il primo approccio ai dati è stato quello dell'analisi delle componenti principali (PCA),

attraverso la quale si è cercato una proprietà che fosse più discriminante di altre tra quelle note (ad esempio il genere, il giorno di raccolta e così via).

Quindi sono state effettuate analisi *supervised* per cercare di massimizzare la discriminazione tra i vari gruppi per ciascuna proprietà: la PLS è stata realizzata utilizzando la funzione "pls.regression", implementata nella libreria di R "plsgenomics"; la CA è stata eseguita usando la funzione di R standard "cancor" e a seconda dell'analisi utilizzando "pls" o "pca" come metodo di riduzione delle variabili.

La capacità predittiva dei modelli statistici ottenuti (accuratezza) è stata stimata mediante cross-validazione (CV) secondo l'approccio Monte Carlo con classificatore k-NN (k=5).

Infine le regioni spettrali relative ai metaboliti assegnati negli spettri NMR sono state allineate mediante algoritmi di allineamento, quindi integrate così da poter ottenere le concentrazioni dei metaboliti in unità di intensità relative. Tali intensità sono state analizzate a seconda delle diverse proprietà intrinseche che caratterizzano i campioni, così da ottenere una visione complessiva dell'andamento delle concentrazioni metaboliche in funzione dei vari parametri esaminati, ed eventualmente determinare i metaboliti che hanno cambiamenti significativi. La significatività statistica è stata assegnata utilizzando il test univariato non parametrico Wilcoxon. Un p-valore $<0.05$ è considerato statisticamente significativo.

# 4 RISULTATI E DISCUSSIONE

## 4.1 Spettri NMR di saliva umana

Dalle analisi sui campioni sono stati ottenuti 1230 spettri NMR monodimensionali, visualizzabili in Figura 11 e in Figura 12.

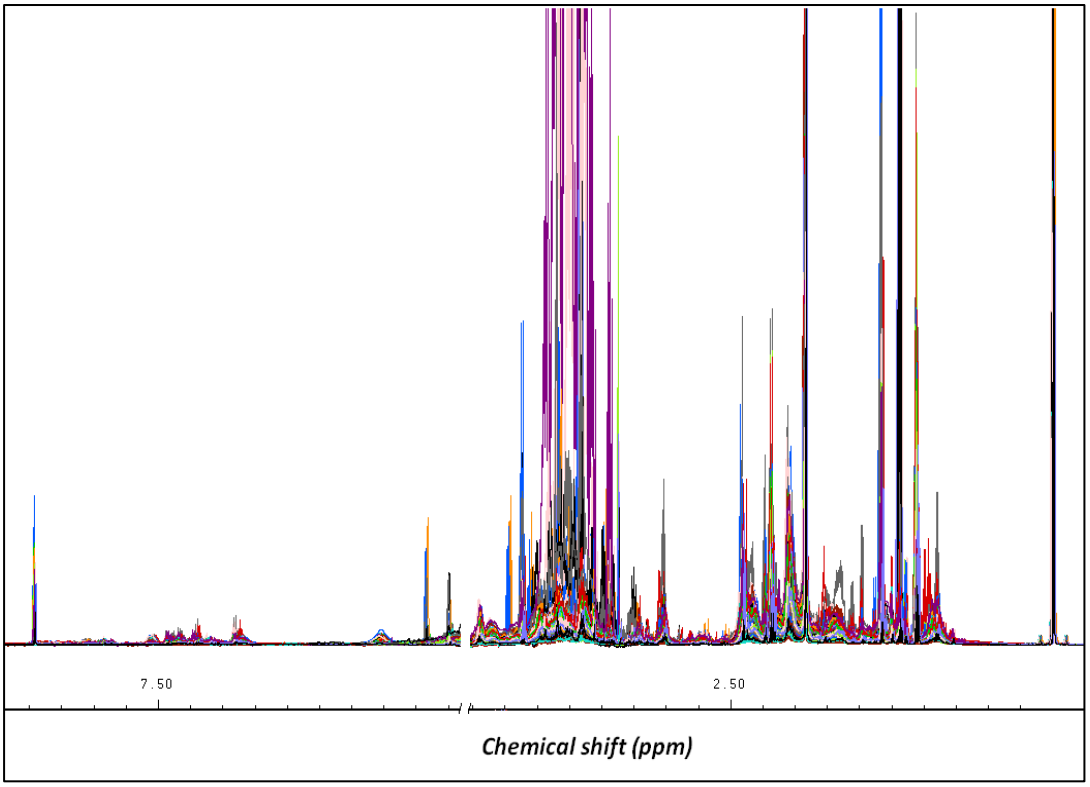



**Figura 11- Sovrapposizione di alcuni degli spettri NMR acquisiti. La zona contenente il segnale dell'acqua (4.50-6.00 ppm) è esclusa.**

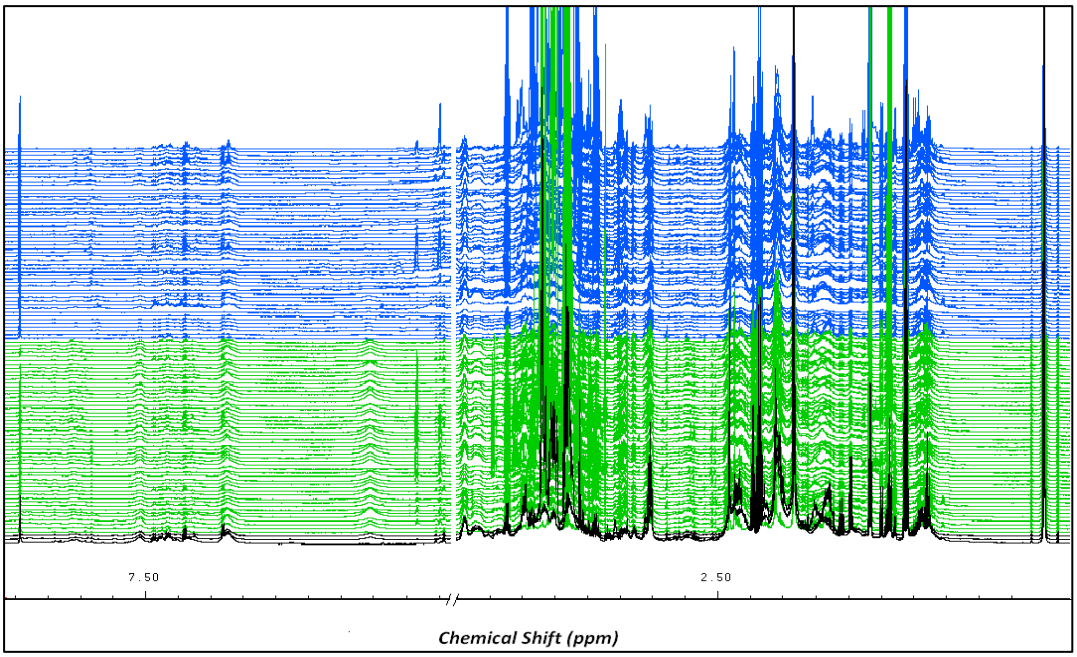



**Figura 12- Alcuni degli spettri NMR ottenuti, colorati in relazione al soggetto che ha fornito il campione. La zona contenente il segnale dell'acqua è esclusa.**

In Tabella 1 sono riportati i 23 metaboliti identificati ed il corrispettivo segnale utilizzato per le analisi successive. Quattro esempi di picchi assegnati vengono mostrati in Figura 13.

**Tabella 1 – Metaboliti identificati per questo studio con rispettivo $\delta_H$ in ppm; tra parentesi sono riportati la molteplicità e l'assegnamento.**

| | | | |
|---|---|---|---|
| **Propionato** | 1,06 (t, $CH_3$) | **Formiato** | 8,46 (s, CH) |
| **Etanolo** | 1,19 (t, $CH_3$) | **Isobutirrato** | 1,22 (d,$CH_3$) |
| **Lattato** | 1,33 (d, $CH_3$) | **Piruvato** | 2,38 (s,$CH_3$) |
| **Alanina** | 1,48 (d, $CH_3$) | **Succinato** | 2,41 (s,$CH_2$) |
| **n-Butirrato** | 1,56 (m, $CH_2$) | **Sarcosina** | 2,74 (s,$CH_3$) |
| **Acetato** | 1,93 (s, $CH_3$) | **Trimetilammina** | 2,89 (s,$CH_3$) |
| **Gruppi N-Acetilici** | 2,04 (m, $CH_3$) | **Colina** | 3,20 (s,$CH_3$) |
| **Citrato** | 2,55 (dd, $CH_2$) | **Isoleucina** | 1,02 (d,$CH_3$) |
| **Metilammina** | 2,61 (s, $CH_3$) | **Trimetilammina N-Ossido (TMAO)** | 3,22 (s,$CH_3$) |
| **Metanolo** | 3,37 (s,$CH_3$) | **Glicole Propilenico** | 1,14 (d,$CH_3$) |
| **Glicina** | 3,57 (s,$CH_2$) | **Valina** | 0,99 (d,$CH_3$) |
| **Fenilalanina** | 7,37 (m, CH) | | |

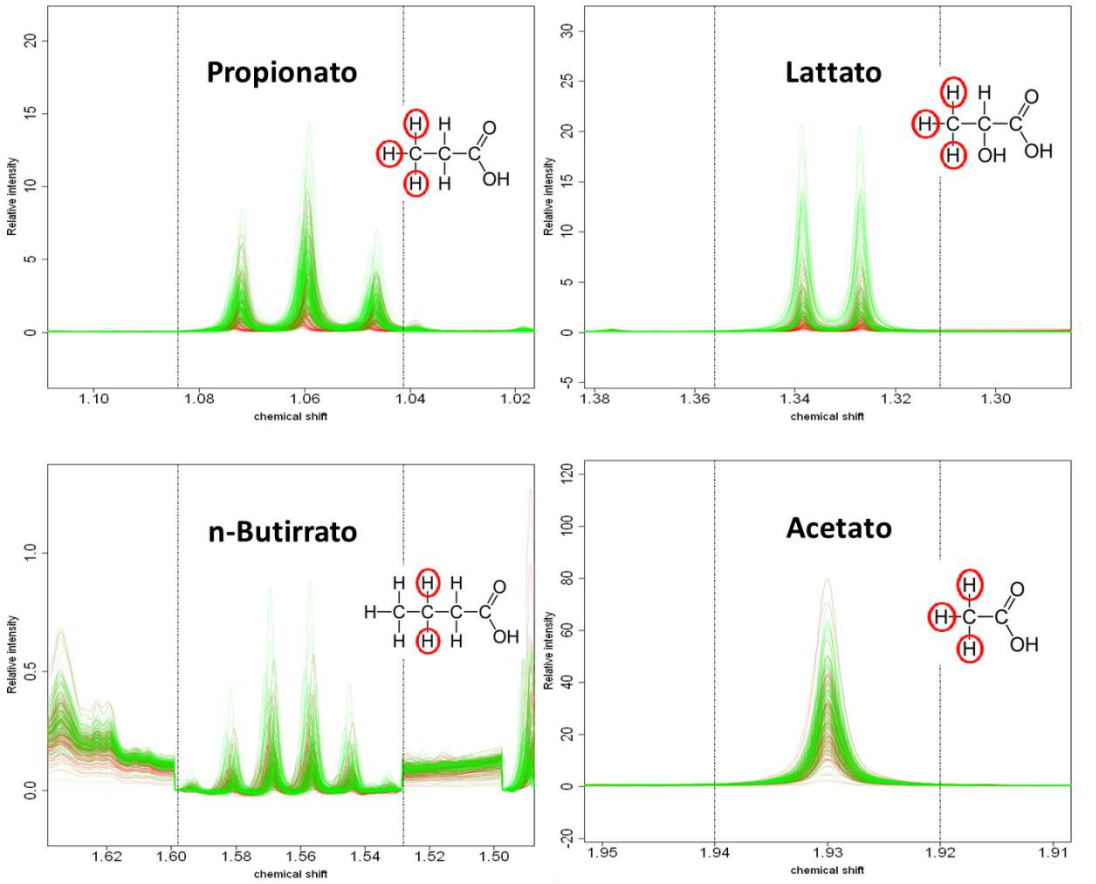



**Figura 13- Quattro esempi di segnali presenti negli spettri NMR utilizzati per l'assegnamento dei metaboliti.**

## 4.2 Ricoscimento degli individui: studio del *fingerprint* metabolico individuale

Per verificare l'esistenza di un *metabotype* individuale, l'approccio iniziale ai dati da un punto di vista descrittivo è quello del metodo *unsupervised* della PCA. L'analisi delle componenti principali viene applicata per prima cosa all'intera matrice dei *buckets* (1230 campioni). Per visualizzare i dati si possono mettere in grafico gli *scores* delle prime due componenti (che insieme spiegano circa il 77% della varianza) [Figura 14]. Da questo si evince il fatto che non sono presenti naturali clusterizzazioni o distribuzioni regolari: la maggior parte dei dati si trova infatti in una zona abbastanza contenuta di spazio, seppur con numerosi *outliers* (campioni che appaiono molto diversi dagli altri e quindi più lontani nel grafico), a volte anche molto distanti (questi ultimi sono stati esclusi in figura).

Da questo tipo di analisi preliminare risulta quindi evidente che le differenze intraindividuali sono considerevoli, mentre quelle interindividuali non appaiono ancora del tutto chiare.

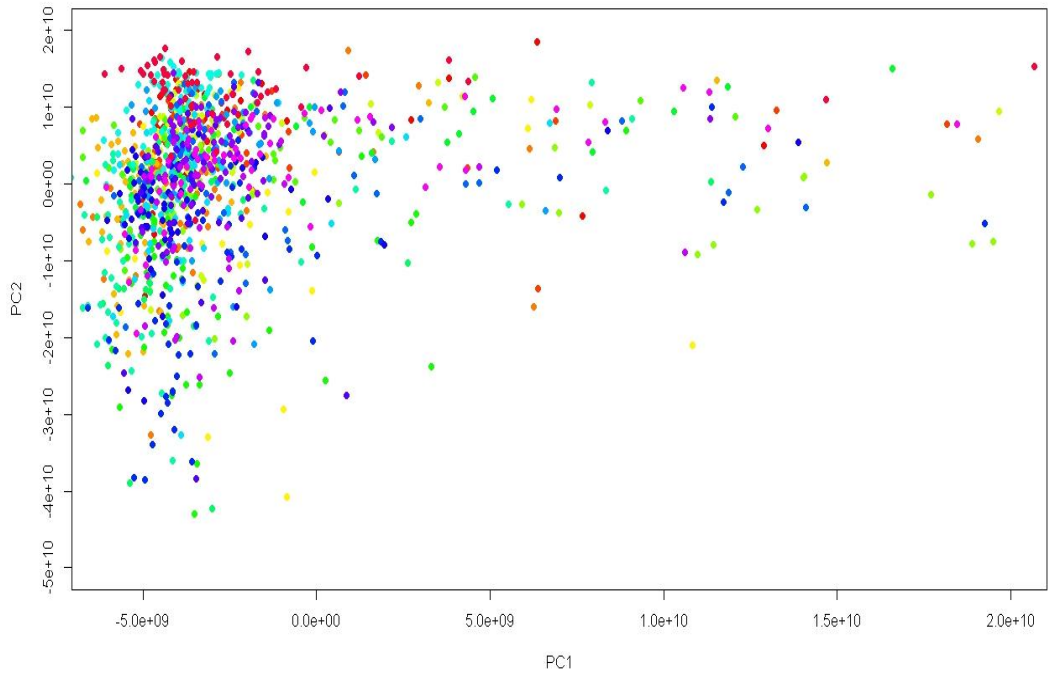
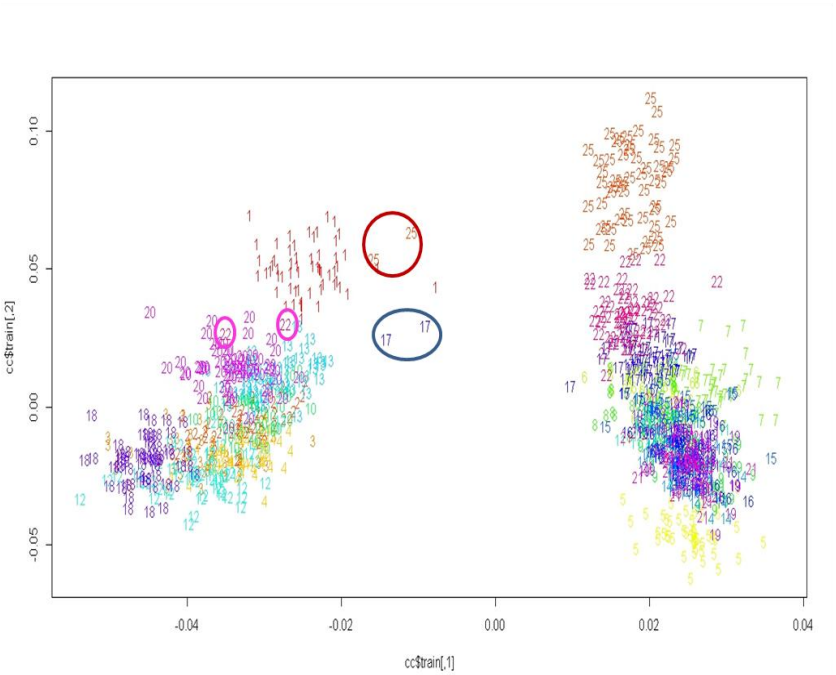



**Figura 14- Grafico degli scores della PCA su tutti i campioni acquisiti.**
**Ogni punto rappresenta un campione, colorato a seconda degli individui.**
**(alcuni *outliers* sono stati esclusi limitando gli assi per migliorare la visualizzazione complessiva)**

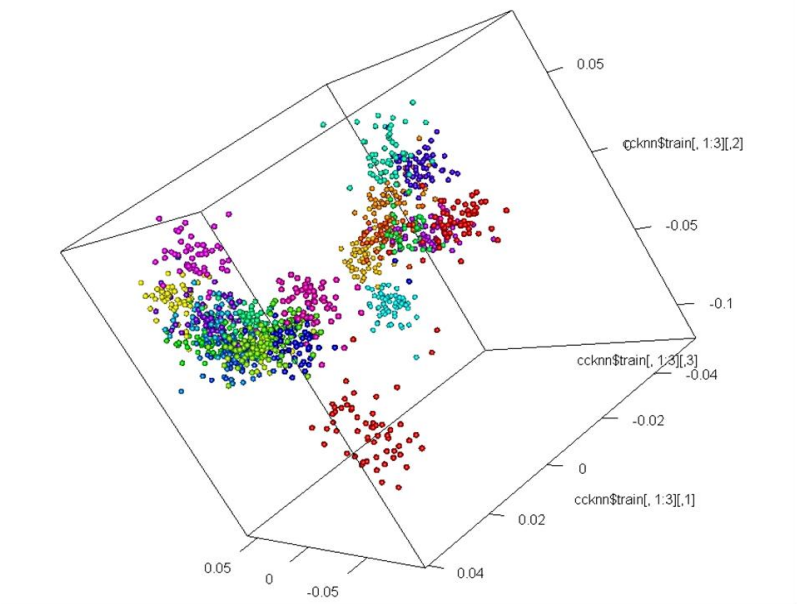
La PCA viene applicata anche a numerosi sottogruppi di dati per approfondire e visualizzare in maniera più immediata differenze e similarità: sono stati presi in considerazione soltanto i primi campioni della giornata, i campioni raccolti nei giorni di dieta libera, quelli in dieta standardizzata, quelli della giornata in cui era previsto l'esercizio fisico, e infine quelli appartenenti solo ad alcuni individui. In tutti i casi i risultati sono stati simili al precedente e poco significativi.

Passando a metodi *supervised* come PLS/CA [Figura 15] e PCA/CA/KNN [Figura 16] si è notata subito una clusterizzazione in due macrogruppi all'interno dei quali è comunque possibile scorgere una suddivisione in piccoli raggruppamenti composti dagli individui. Indagando sulle ragioni di tale separazione sono state individuate come cause sia il diverso periodo di raccolta (per motivi logistici infatti lo studio si è svolto per i due gruppi di individui in momenti diversi) sia la differenza di *buffer* utilizzati nell'analisi. Questi *buffer*, in teoria aventi la medesima composizione, provengono da due stock diversi, ed uno dei due possiede minore capacità tampone e differente forza ionica, per cui molti segnali risultano leggermente *shiftati*, soprattutto nella zona degli aromatici.
Tuttavia il riconoscimento individuale verificabile tramite cross-validazione è molto alto, come si può vedere dalle percentuali sulla diagonale della matrice di confusione, così come l'accuratezza.




| INDIV | DIAGONAL |
|---|---|
| 1M | 98.3% |
| 2F | 93.7% |
| 3F | 93.8% |
| 4F | 98.3% |
| 5F | 83.1& |
| 6F | 90.7% |
| 7F | 97.6 % |
| 8M | 82.9 % |
| 9F | 90.3% |
| 10M | 89.4% |
| 11F | 91.5% |
| 12M | 83.9 % |
| 13F | 100% |
| 14F | 89.1 % |
| 15F | 85.2 % |
| 16F | 93.5 % |
| 17F | 86.4 % |
| 18F | 90.7 % |
| 19M | 93.1 % |
| 20F | 94.8 % |
| 21M | 90.5 % |
| 22F | 88.1% |
| 25F | 86.8 % |
| (Acc. | 90.7%) |

**Figura 15- A sinistra, il grafico delle prime due componenti principali della PLS/CA su tutti i dati (1230 spettri). A destra, le percentuali presenti sulla diagonale della matrice di confusione, da cui si valuta l'accuratezza nel riconoscimento di ogni singolo soggetto.**




| INDS | DIAGONAL |
|---|---|
| 1M | 97.5% |
| 2F | 96.1% |
| 3F | 92.9% |
| 4F | 100% |
| 5F | 82.9 % |
| 6F | 96.8 % |
| 7F | 92.0 % |
| 8M | 89.4 % |
| 9F | 91.8 % |
| 10M | 96.3 % |
| 11F | 91.0 % |
| 12M | 92.5 % |
| 13F | 100% |
| 14F | 91.4 % |
| 15F | 82.2 % |
| 16F | 94.8 % |
| 17F | 89.8 % |
| 18F | 94.4 % |
| 19M | 92.7 % |
| 20F | 96.8 % |
| 21M | 97.6 % |
| 22F | 92.6 % |
| 25F | 95.4 % |
| (Acc. | 93.5%) |

**Figura 16- A sinistra, il grafico tridimensionale delle prime tre componenti principali della PCA/CA/KNN su tutti i dati. A destra, le percentuali presenti sulla diagonale della matrice di confusione.**

Inoltre la discriminazione tra gruppi di *buffer* è presente esclusivamente sulla prima componente principale, come si vede se si mettono in grafico gli *scores* relativi alla seconda, alla terza e alla quarta componente [Figura 17].

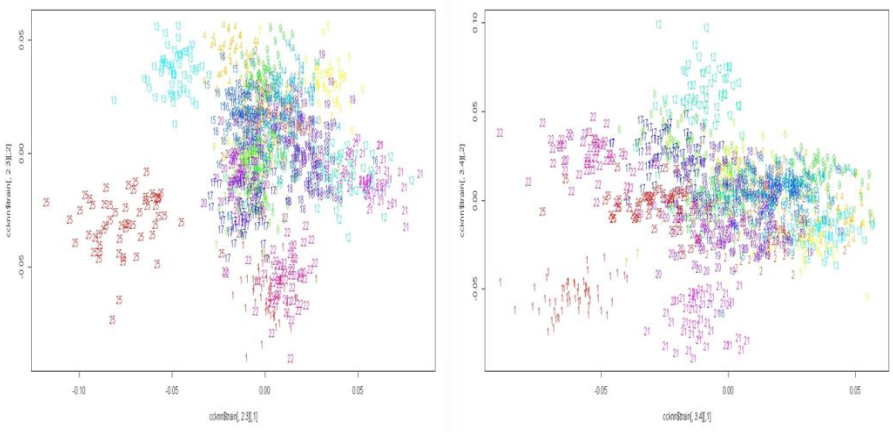
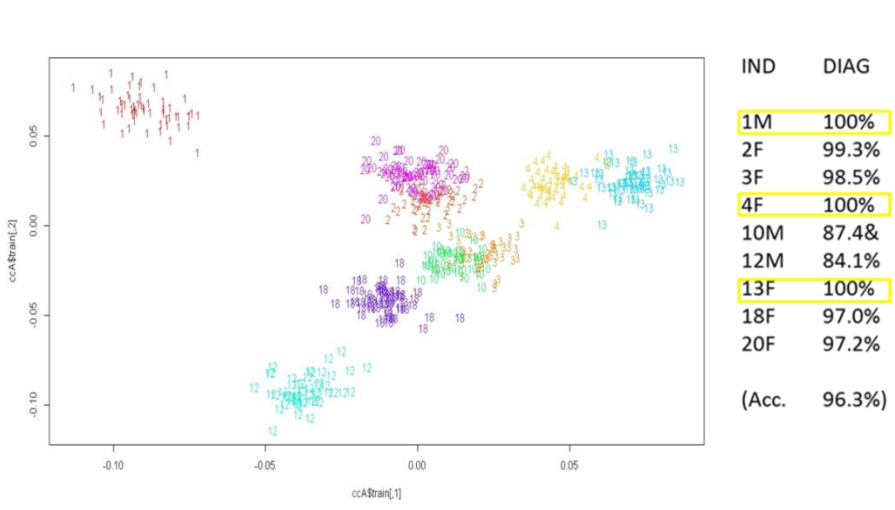


**Figura 17- 2° vs 3° componente principale (a sinistra), e 3° vs 4° componente principale (a destra) di PCA/CA/KNN.**

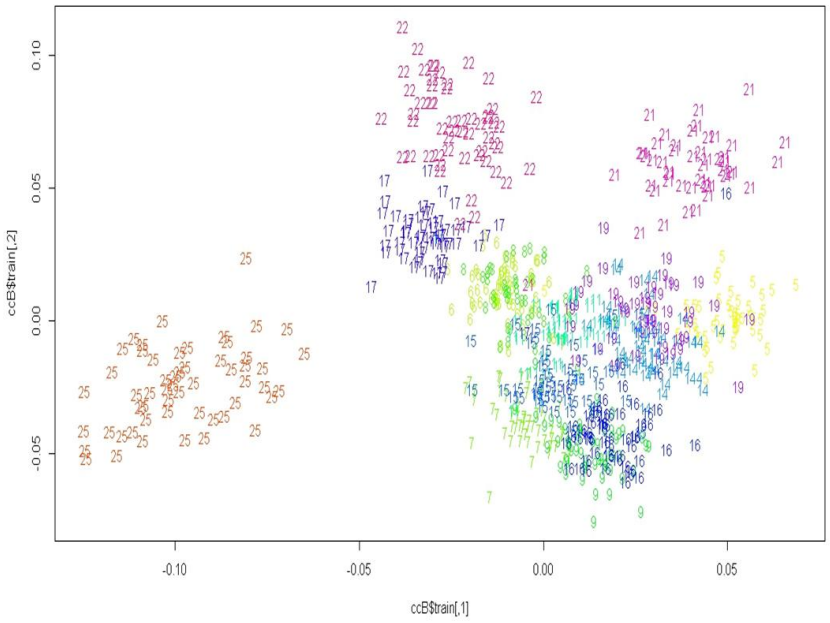
A conferma della robustezza dei risultati e dell'irrilevanza del *buffer* nell'ottenimento di un ottimo riconoscimento individuale, ulteriori studi sono stati fatti per verificare che l'accuratezza nel discernere ciascun soggetto da un altro si mantenesse elevata.

Per prima cosa, il set completo di dati è stato suddiviso in due sottogruppi, corrispondenti ai campioni analizzati con l'uno o con l'altro *buffer*. In entrambi i casi i risultati sono stati ottimi, con accuratezze molto alte e riconoscimento individuale estremamente evidente. [Figura 18 e Figura 19]



| IND | DIAG |
|---|---|
| 1M | 100% |
| 2F | 99.3% |
| 3F | 98.5% |
| 4F | 100% |
| 10M | 87.4& |
| 12M | 84.1% |
| 13F | 100% |
| 18F | 97.0% |
| 20F | 97.2% |
| (Acc. | 96.3%) |

**Figura 18- PCA/CA/KNN effettuata su campioni analizzati tutti con uno stesso *buffer*. A destra, le percentuali di riconoscimento individuale presenti sulla diagonale della matrice di confusione.**




| INDS | DIAGONAL |
|---|---|
| 5F | 86.6% |
| 6F | 92.9% |
| 7F | 91.8% |
| 8M | 95.0% |
| 9F | 93.3% |
| 11F | 91.6% |
| 14F | 95.9% |
| 15F | 80.7% |
| 16F | 95.4% |
| 17F | 88.5% |
| 19M | 97.6% |
| 21M | 99.0% |
| 22F | 92.0% |
| 25F | 96.0% |
| (Acc. | 92.5%) |

**Figura 19- PCA/CA/KNN effettuata sui campioni analizzati con l'altro *buffer*. A destra, le percentuali di riconoscimento individuale presenti sulla diagonale della matrice di confusione.**

Inoltre sono stati studiati alcuni sottogruppi di dati "misti", ovvero provenienti sia da un gruppo che dall'altro. In tutti i casi l'accuratezza dei risultati si è mantenuta estremamente alta, in alcuni casi superando addirittura il 98%.

Infine, si è scelto di eliminare le differenze dovute al buffer per via statistica. A questo scopo è stata fatta dapprima un'analisi supervised (PLS/CA) per massimizzare la varianza tra i due gruppi di dati. L'accuratezza per tale discriminazione risulta essere del 100%. In seguito la componente principale, che raccoglie in sé le informazioni più marcatamente determinanti per questa separazione, è stata eliminata, ed è stata ottenuta dunque un nuova matrice di dati corretti che, se divisi nuovamente per *buffer*, non risultano più discriminati (accuratezza <40%). [Figura 20]

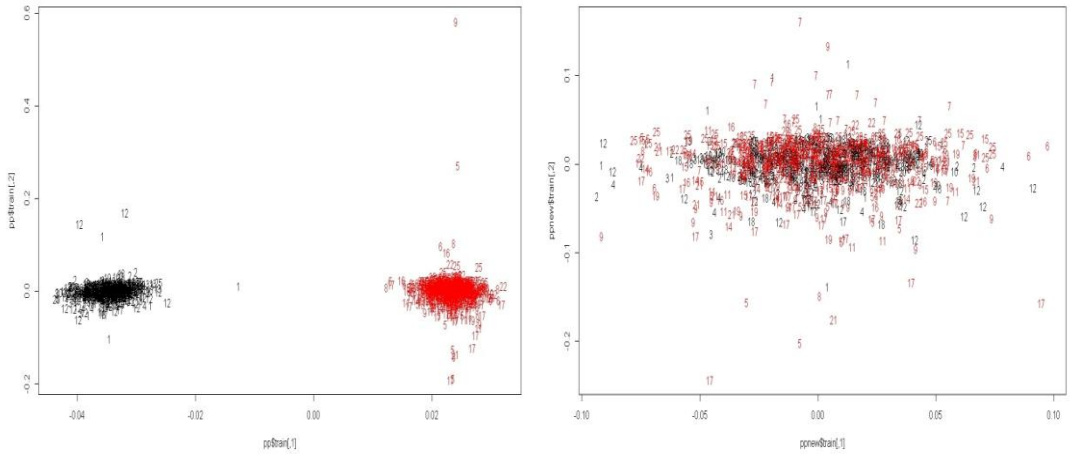


**Figura 20- PLS/CA effettuata su tutti i campioni suddivisi per *buffer,* prima e dopo la correzione tramite metodi statistici.**

Sulla nuova matrice ottenuta sono state eseguite le analisi viste in precedenza, con risultati sorprendenti: mentre l'analisi *unsupervised* risulta pressoché invariata, le analisi *supervised* mostrano che il riconoscimento individuale non solo è mantenuto, ma è addirittura aumentato, sia per i singoli sottogruppi di dati e che per il set completo di campioni.

| INDS | PLS/CA |
|---|---|
| 1M | 98.8% |
| 2F | 91.4% |
| 3F | 89.5% |
| 4F | 99.7% |
| 5F | 83.1 % |
| 6F | 94.1 % |
| 7F | 95.8 % |
| 8M | 92.4 % |
| 9F | 83.9 % |
| 10M | 87.1 % |
| 11F | 94.3 % |
| 12M | 96.5 % |
| 13F | 100% |
| 14F | 98.2 % |
| 15F | 94.2 % |
| 16F | 96.7 % |
| 17F | 90.2 % |
| 18F | 95.7 % |
| 19M | 96.8 % |
| 20F | 99.2% |
| 21M | 97.9 % |
| 22F | 95.0 % |
| 25F | 92.9 % |
| (Acc. | 93.8%) |

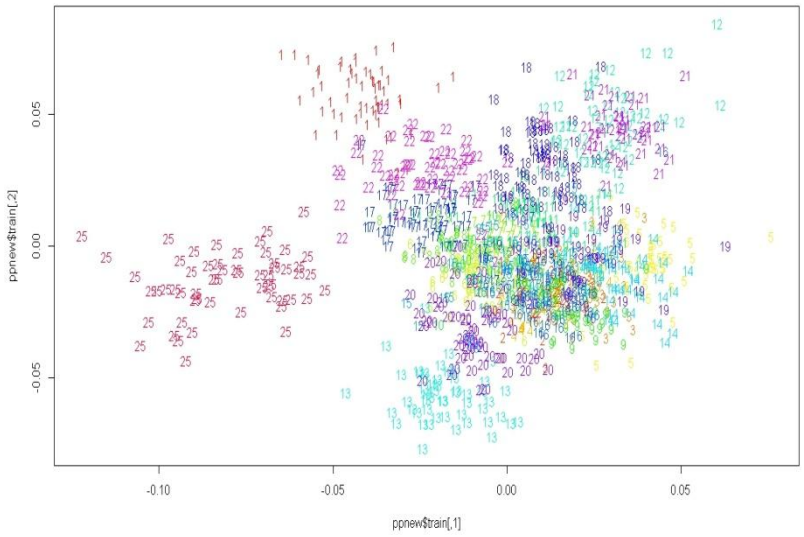



**Figura 21- Riconoscimento individuale del modello basato sulla PLS/CA effettuata sulla nuova matrice di dati. A destra, il relativo grafico in cui vengono proiettate le prime due componenti principali.**

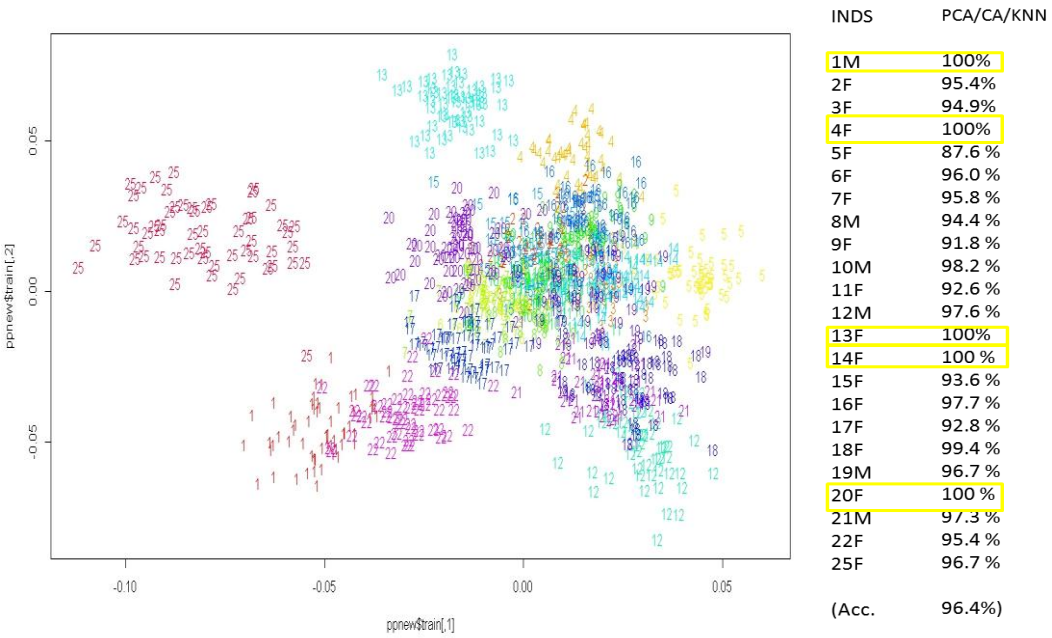



**Figura 22- Riconoscimento individuale del modello basato sulla PCA/CA/KNN effettuata sulla nuova matrice di dati. A sinistra, il relativo grafico in cui vengono proiettate le prime due componenti principali.**

A questo punto sono stati utilizzati sottogruppi di campioni per osservare eventuali correlazioni tra il riconoscimento individuale e altri fattori, in particolare la dieta. Gli stessi metodi *supervised* sono stati dunque eseguiti prendendo in considerazione:

(I) solo i primi campioni della giornata, raccolti durante i primi 8 giorni; (dieta libera)

(II) tutti i campioni dei primi 7 giorni più il primo campione dell'ottavo giorno; (dieta libera)

(III) tutti i campioni raccolti negli ultimi 3 giorni, escluso il primo campione dell'ottavo giorno; (dieta standardizzata, incluso l'esercizio fisico del decimo giorno)

(IV) tutti i campioni ad esclusione del primo raccolti l'ottavo giorno, e tutti quelli raccolti il nono giorno (dieta standardizzata escluso l'esercizio fisico)

La distinzione tra i vari soggetti è rimasta ottima seppur con qualche sostanziale diminuzione, dovuta presumibilmente alla minore quantità di campioni disponibili per la costruzione dei modelli statistici di riconoscimento.

**Tabella 2- Percentuali di riconoscimento individuali presenti sulle matrici di confusione, ottenute da modelli basati su PCA/CA/KNN, effettuata sui quattro diversi sottogruppi di campioni.**

| INDIVIDUO | PCA/CA/KNN (I) | PCA/CA/KNN (II) | PCA/CA/KNN (III) | PCA/CA/KNN (IV) |
|---|---|---|---|---|
| **1 M** | 86.7 % | 100 % | 93.9 % | 95.2 % |
| **2 F** | 36.4 % | 90.4 % | 91.0 % | 94.0 % |
| **3 F** | 87.7 % | 100 % | 81.2 % | 68.4 % |
| **4 F** | 89.5 % | 100 % | 100 % | 100 % |
| **5 F** | 39.7 % | 88.2 % | 84.5 % | 68.1 % |
| **6 F** | 82.5 % | 85.7 % | 87.7 % | 79.4 % |
| **7 F** | 77.0 % | 98.5 % | 92.6 % | 75.0 % |
| **8 M** | 87.5 % | 87.0 % | 100 % | 93.2 % |
| **9 F** | 67.9 % | 89.9 % | 87.3 % | 91.7 % |
| **10 M** | 81.2 % | 96.3 % | 91.7 % | 37.5 % |
| **11 F** | 95.8 % | 91.7 % | 92.8 % | 98.5 % |
| **12 M** | 77.5 % | 89.1 % | 95.5 % | 100 % |
| **13 F** | 97.9 % | 100 % | 100 % | 100 % |
| **14 F** | 100 % | 93.8 % | 100 % | 86.1 % |
| **15 F** | 85.2 % | 88.4 % | 54.8 % | 61.9 % |
| **16 F** | 96.9 % | 94.1 % | 98.4 % | 93.0 % |
| **17 F** | 60.4 % | 83.8 % | 58.7 % | 80.2 % |
| **18 F** | 85.0 % | 89.9 % | 82.1 % | 97.4 % |
| **19 M** | 100 % | 91.6 % | 89.9 % | 78.3 % |
| **20 F** | 86.3 % | 97.0 % | 100 % | 87.3 % |
| **21 M** | 100 % | 98.0 % | 81.9 % | 74.5 % |
| **22 F** | 100 % | 92.9 % | 94.7 % | 90.7 % |
| **25 F** | 83.3 % | 94.1 % | 87.4 % | 93.6 % |
| **ACCURAT.** | **83.1 %** | **93.1 %** | **89.1 %** | **86.0 %** |

Come ultima analisi per quanto riguarda il fenotipo metabolico legato a caratteristiche intrinseche degli individui, si è valutata l'entità del riconoscimento di genere. Prima di tutto andando ad indagare se nei raggruppamenti trovati per l'identificazione individuale sia possibile osservare una discriminazione legata al sesso del soggetto. Appare immediatamente evidente che tale caratteristica non è

determinante nel distinguere un individuo dall'altro, ma solo con un'analisi *supervised* si riesce ad ottenere una buona accuratezza (85.2%) con cui predire il genere della persona.

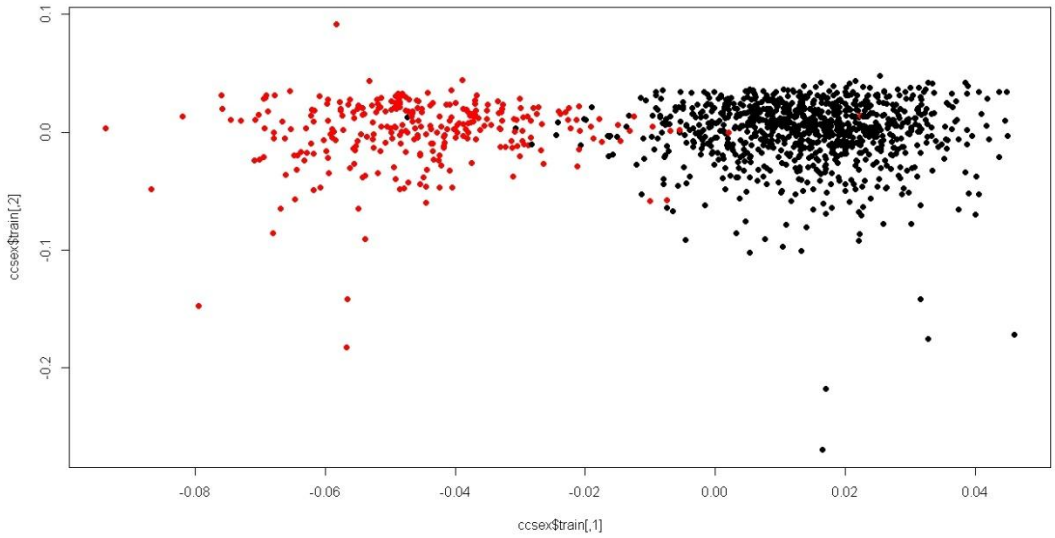



**Figura 23 - PCA/CA/KNN che mette in evidenza la distinzione tra individui di sesso maschile (in rosso) e femminile (in nero)**

## 4.3 *Fingerprinting* giornaliero: analisi delle variazioni dei dati nel tempo

Per esaminare i cambiamenti del *fingerprint* metabolico giorno per giorno, e quindi andare ad indagare se l'influenza della dieta, dell'esercizio fisico, o del momento della giornata in cui è stato raccolto il campione, è determinante rispetto alla presenza di un profilo metabolico intrinseco per un dato soggetto, le analisi eseguite sono state le stesse viste in precedenza. Quello che cambia è l'approccio di studio per cui le variazioni che si vanno ad evidenziare sono quelle intraindividuali nel corso del tempo.

Dalla PCA precedentemente eseguita non si ottiene alcuna informazione su una differenziazione in tal senso, al contrario, se colorati a seconda del giorno o del momento della giornata in cui i campioni sono stati ottenuti, i punti del grafico risultano ancora più eterogenei rispetto al riconoscimento individuale visto sopra.

Anche stavolta è solo con analisi *supervised* che si ottengono risultati significativi.

Da una PCA/CA/KNN effettuata su tutti i campioni prendendo come parametro di discriminazione i giorni di raccolta, si nota subito che i punti corrispondenti ai campioni ottenuti negli ultimi tre giorni di studio (quindi quelli relativi al periodo di dieta standardizzata) sono tutti molto vicini tra loro e separati dal resto dei punti sulla prima componente principale. Sebbene questa separazione non sia netta, rimane evidente come la dieta influisca sul profilo metabolico portando ad un riconoscimento per lo meno visivo dei corrispettivi campioni.

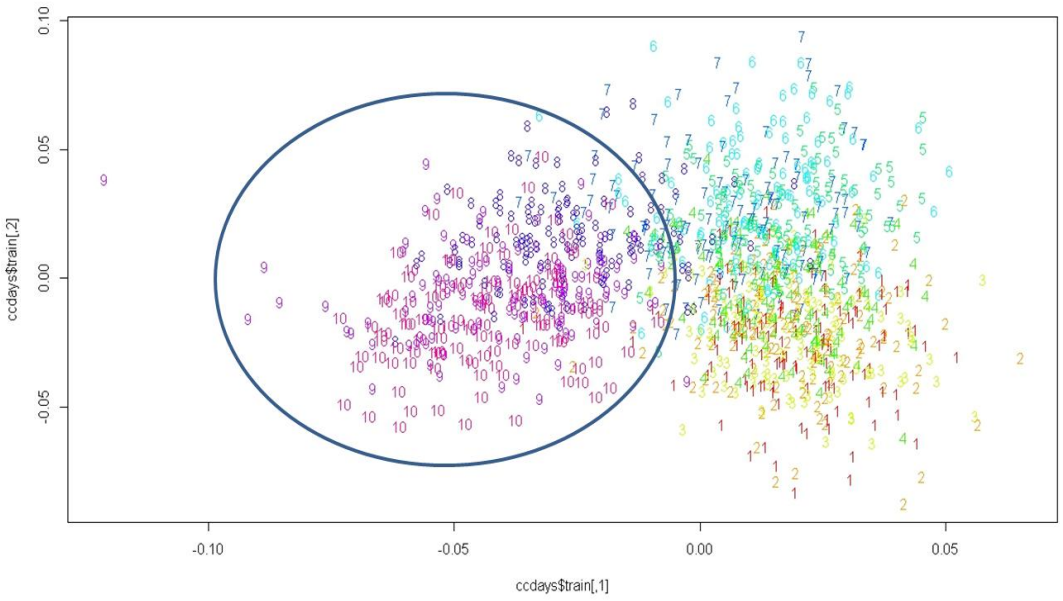



**Figura 24- PCA/CA/KNN per il riconoscimento del giorno di raccolta dei campioni. Nel cerchio sono per lo più presenti i punti relativi agli ultimi tre giorni.**

Per quanto riguarda i modelli di riconoscimento, le cross-validazioni effettuate usando di volta in volta svariati gruppi di campioni possono essere riassunte in due risultati fondamentali. Il primo, prevedibile, mette in evidenza il fatto che il primo campione di saliva ottenuto da ciascun soggetto l'ottavo giorno è più "simile" ai campioni raccolti nei primi sette giorni, ovvero viene statisticamente riconosciuto come appartenente al gruppo di campioni relativi alla dieta libera. Questa considerazione può sembrare un'ovvia banalità dal momento che è il primo campione della mattina, raccolto prima che la dieta standardizzata avesse ufficialmente inizio, tuttavia è utile per verificare la validità del metodo e l'accuratezza dei risultati che esso fornisce.

Il secondo, meno immediato, è che i campioni relativi al decimo giorno, sia considerando che escludendo il primo campione della giornata fornito da ciascun volontario, non risultano essere significativamente diversi da quelli raccolti nei due giorni precedenti, in quanto dalle matrici di confusione si ottengono basse percentuali sulla diagonale, con gran parte delle interferenze dovuta ai giorni 8 e 9, e l'accuratezza nel riconoscimento risulta sempre piuttosto scarsa. È difficile dire se ciò sia dovuto al fatto che l'esercizio fisico abbia effetti minimi sul metaboloma salivare e prevalga invece l'influenza della dieta, oppure se la piccola quantità di campioni a disposizione relativi a quest'unico giorno non sia sufficiente a rilevare l'introduzione di attività fisica, sicuramente è però possibile affermare che se essa apporta modifiche al *fingerprint* metabolico complessivo, queste non si riscontrano nell'immediato.

Queste considerazioni suggeriscono di poter dividere i campioni in due gruppi a seconda della presenza o meno della dieta standardizzata, sui quali viene effettuata una PCA/CA/KNN [Figura 25]. L'accuratezza nel riconoscimento si aggira intorno all'88%, ed è quindi discretamente elevata anche se significativamente più bassa rispetto a quanto avviene per il riconoscimento individuale.

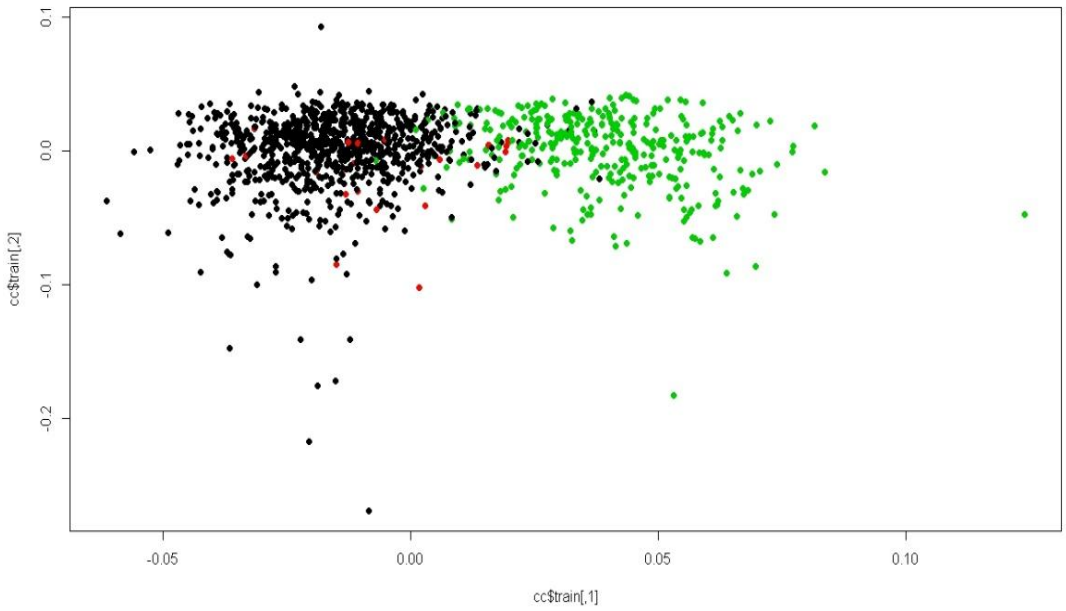



**Figura 25- PCA/CA/KNN per il riconoscimento tra i campioni raccolti durante il periodo di dieta libera (colorati in nero), e quelli raccolti durante i tre giorni di dieta standardizzata (colorati in verde).**

Ciò che è emerso con maggior evidenza dallo studio del *fingerprint* del tempo, consiste però nella netta differenza che intercorre tra il primo campione raccolto alla mattina e gli altri accumulati durante tutto il resto di una giornata, piuttosto che tra campioni ottenuti in giorni diversi. Questo dato risulta chiaro prima di tutto dal punto di vista visivo, in quanto dal grafico della PCA/CA/KNN [Figura 26], effettuata istruendo il sistema con un vettore che indica il momento della giornata (da 1 a 6) in cui è stato raccolto un campione, si nota che i punti relativi ai primi campioni della mattina sono distintamente separati da tutti gli altri, i quali sono invece a stento riconoscibili tra di loro.

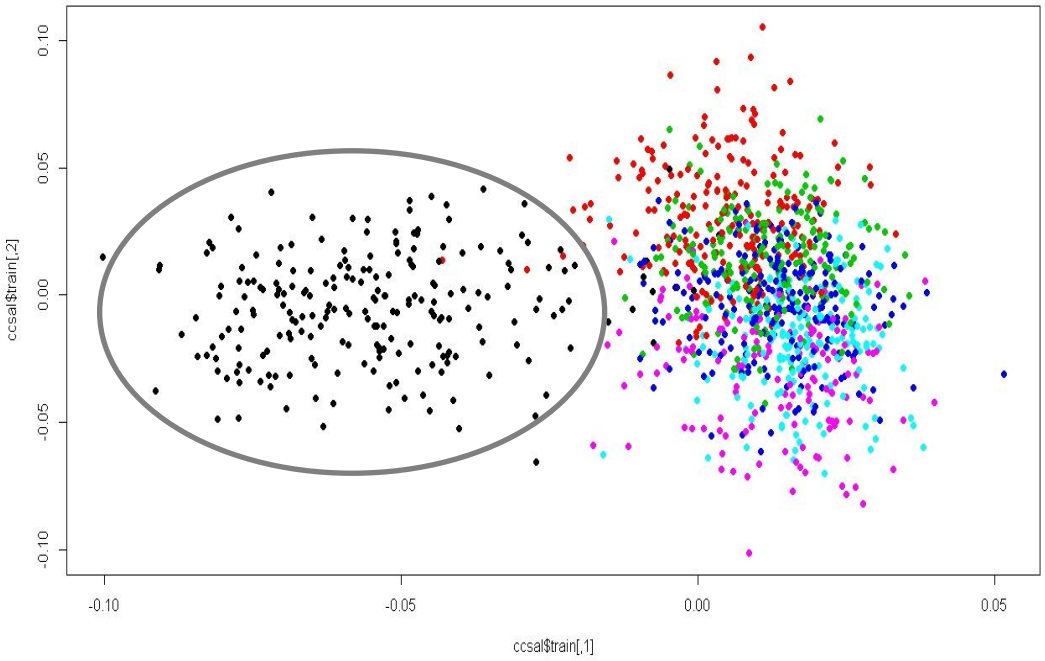



**Figura 26- PCA/CA/KNN per il riconoscimento dei campioni in base al momento di raccolta durante la giornata. I primi campioni del giorno sono tutti contenuti all'interno del cerchio e nettamente separati dal resto.**

In secondo luogo, questa diversità si rivela anche nella matrice di confusione che si ottiene dal metodo della cross-validazione [

Tabella 3], in cui il riconoscimento del momento di raccolta è estremamente basso per tutti i campioni eccetto che per il primo.

**Tabella 3- Matrice di confusione ottenuta dalla cross-validazione per valutare il modello di riconoscimento tra i campioni, in relazione al momento di raccolta durante la giornata.**

| | 1° | 2° | 3° | 4° | 5° | 6° |
|---|---|---|---|---|---|---|
| 1° | **68.9** | 5.1 | 6.2 | 10.8 | 6.1 | 2.9 |
| 2° | 3.9 | **31.9** | 23.8 | 23.0 | 9.5 | 8.0 |
| 3° | 0.6 | 24.9 | **18.4** | 35.6 | 11.8 | 8.6 |
| 4° | 1.1 | 13.0 | 27.2 | **24.1** | 26.2 | 8.4 |
| 5° | 0.8 | 9.0 | 14.0 | 33.3 | **24.2** | 18.6 |
| 6° | 4.8 | 5.8 | 13.5 | 22.9 | 31.9 | **21.1** |

Analogamente a quanto visto in precedenza con la suddivisione dei campioni in base ai giorni di dieta libera e standardizzata, in questo caso risulta naturale la divisione dei campioni in due gruppi, uno costituito dai soli primi campioni della giornata e l'altro da tutto il resto dei campioni. L'accuratezza nell'assegnamento di un campione all'uno o all'altro gruppo nei modelli di riconoscimento è in questo caso estremamente elevata (≈ 95%), e questo risulta assai interessante nell'ottica di studi futuri.

Quanto riportato finora si colloca nell'ambito dello studio delle variazioni temporali complessive del *fingerprint* metabolico, ma, volendo ottenere dettagli circoscritti alla risposta di ciascun individuo, è necessario analizzare singolarmente il set di campioni appartenenti a ciascun volontario. Per questo scopo sono state effettuate le stesse analisi *supervised* (PCA/CA/KNN e PLS/CA) su ogni sottogruppo di dati, istruendo il sistema prima con un vettore indicante il momento della giornata in cui era stato raccolto il campione e poi con uno relativo ai dieci giorni di raccolta.
I risultati sulle diversità intraindividuali nell'arco di tempo di una giornata ricalcano esattamente quelli già visti con lo studio globale su tutti i campioni: per tutti i soggetti appare infatti evidente che i campioni raccolti al mattino sono sempre separati dall'insieme eterogeneo costituito da tutti gli altri campioni. A volte questa separazione è meno netta, ed i punti relativi al primo campione del giorno sono solo di poco discostati dal resto, tuttavia la distinzione resta comunque visibile.

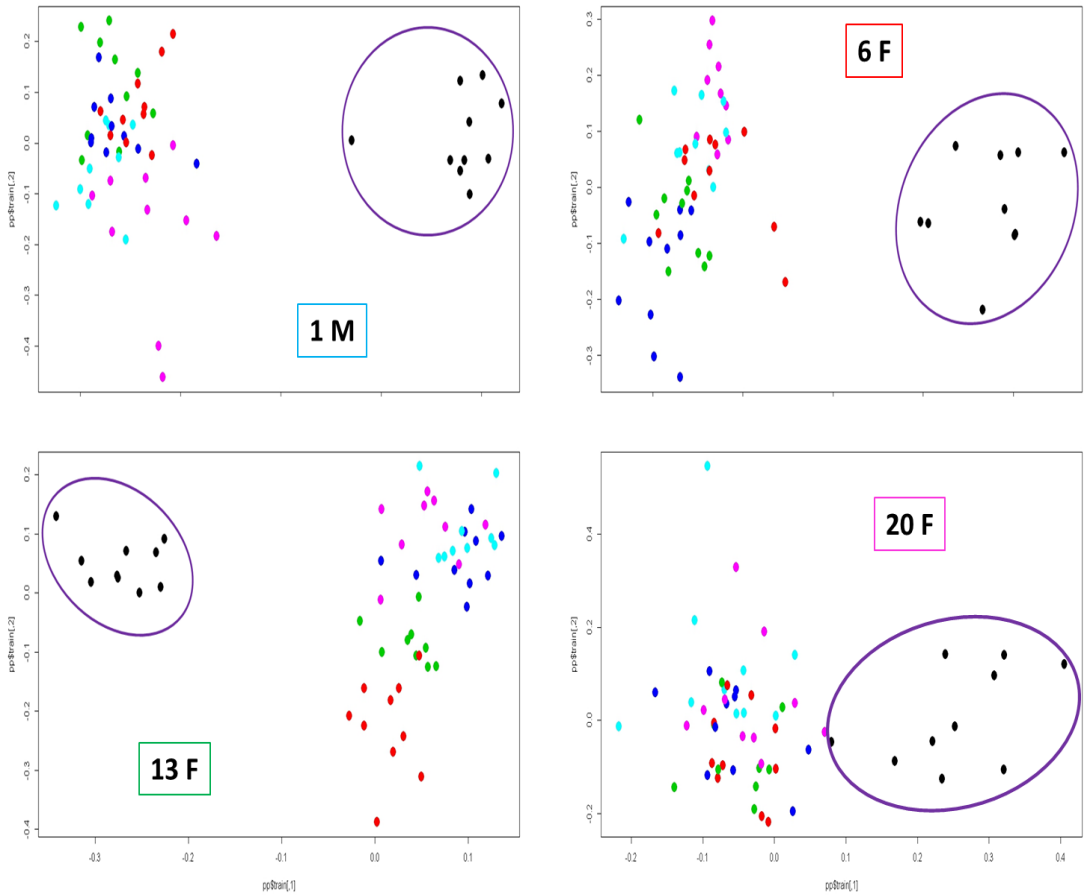



**Figura 27- Quattro esempi di PCA/CA/KNN effettuata su un set di campioni appartenenti tutti ad uno stesso individuo per il riconoscimento dei campioni in base al momento di raccolta durante la giornata. I primi campioni della giornata sono cerchiati in viola.**

Conclusioni meno chiare si possono trarre invece dall'analisi delle variazioni negli spettri NMR tra un giorno e l'altro nei dieci giorni di studio. L'influenza di dieta ed esercizio fisico non appare infatti così evidente per tutti gli individui allo stesso modo, ma anzi la risposta metabolica a tali fattori spesso risulta ritardata, e solo i giorni 9 e 10 (ovvero gli ultimi) risultano spostati dal resto. È necessario inoltre precisare che salvo poche eccezioni la separazione tra i due gruppi non è mai netta, pertanto la variabilità intraindividuale associata all'introduzione di una differente alimentazione è molto meno significativa di quella che intercorre tra il primo campione della giornata e il resto delle salive raccolte.

Tenendo presenti queste premesse, gli individui che hanno partecipato al progetto si possono dividere in quattro gruppi a seconda del cambiamento del *fingerprint* giornaliero in relazione ai parametri previsti dallo studio.

Un primo gruppo, che è comunque il più ampio (11 individui su 23), è quello costituito da coloro che mostrano una separazione, seppur non sempre ben definita, tra i primi sette giorni di studio e gli ultimi tre.

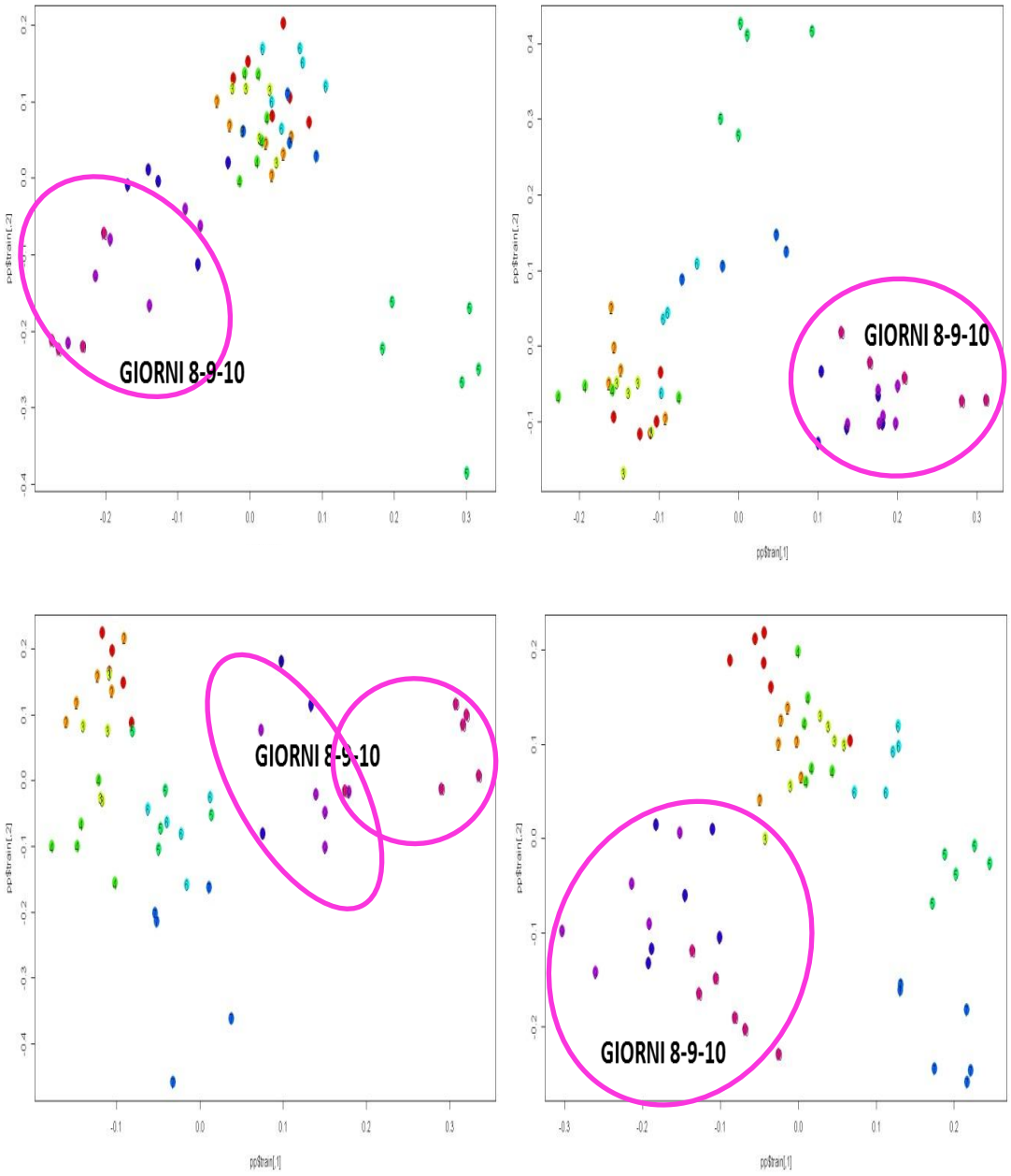



**Figura 28 - Quattro esempi di PCA/CA/KNN effettuata su un set di campioni appartenenti tutti ad uno stesso individuo per il riconoscimento dei campioni in base al giorno di raccolta.**

Un secondo gruppo è costituito dai quattro individui che sembrano avere una risposta "ritardata" alla dieta, ovvero da quelli per i quali i punti relativi ai soli giorni 9 e 10 sono separati dal resto.

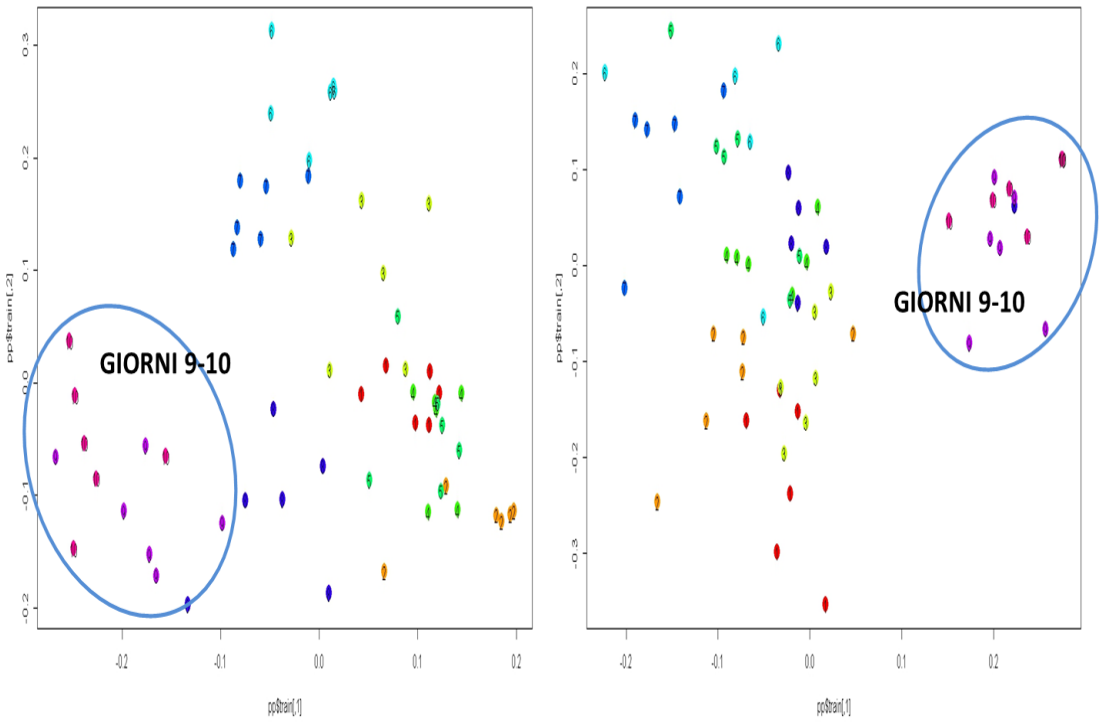



**Figura 29- Due esempi di PCA/CA/KNN effettuata su un set di campioni appartenenti tutti ad uno stesso individuo per il riconoscimento dei campioni in base al giorno di raccolta.**

Nel terzo gruppo (4 individui) sono solo i punti relativi al decimo giorno ad essere separati dal resto, ma è difficile affermare se il motivo sta negli effetti dell'attività fisica che "superano" quelli della dieta, oppure se l'introduzione di un'alimentazione standardizzata risulta visibile per alcuni soggetti solo dopo qualche giorno.

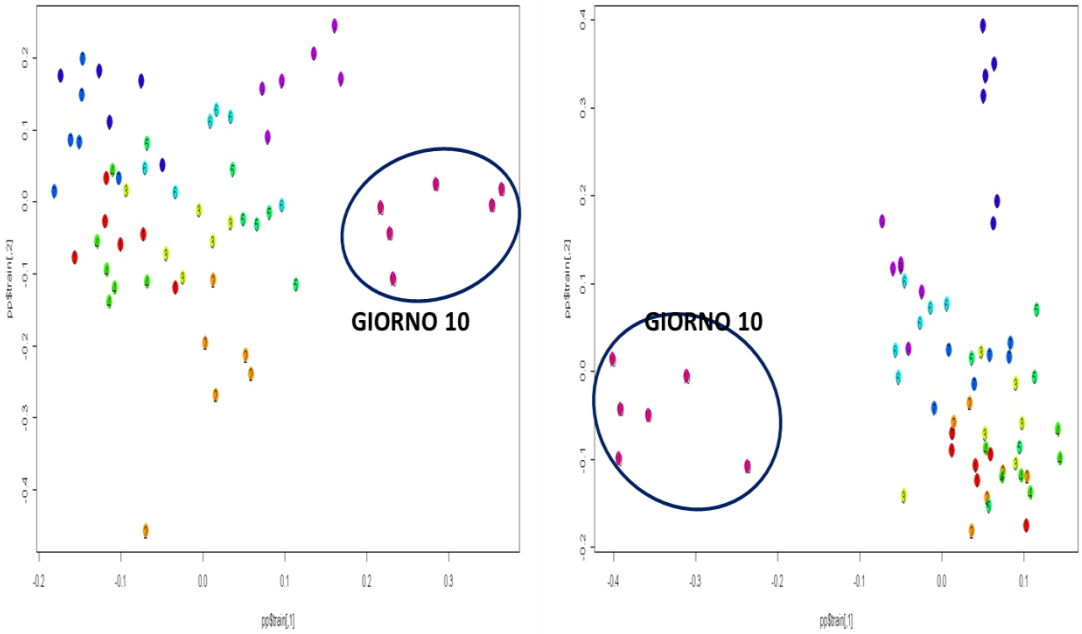



**Figura 30- Due esempi di PCA/CA/KNN effettuata su un set di campioni appartenenti tutti ad uno stesso individuo per il riconoscimento dei campioni in base al giorno di raccolta.**

Infine, per i pochi individui rimasti non si riesce ad osservare una distinzione chiara tra i giorni relativi alla dieta e gli altri, in quanto non sembra essere presente una relazione tra i giorni di raccolta e la distribuzione dei punti corrispondenti a ciascun campione.

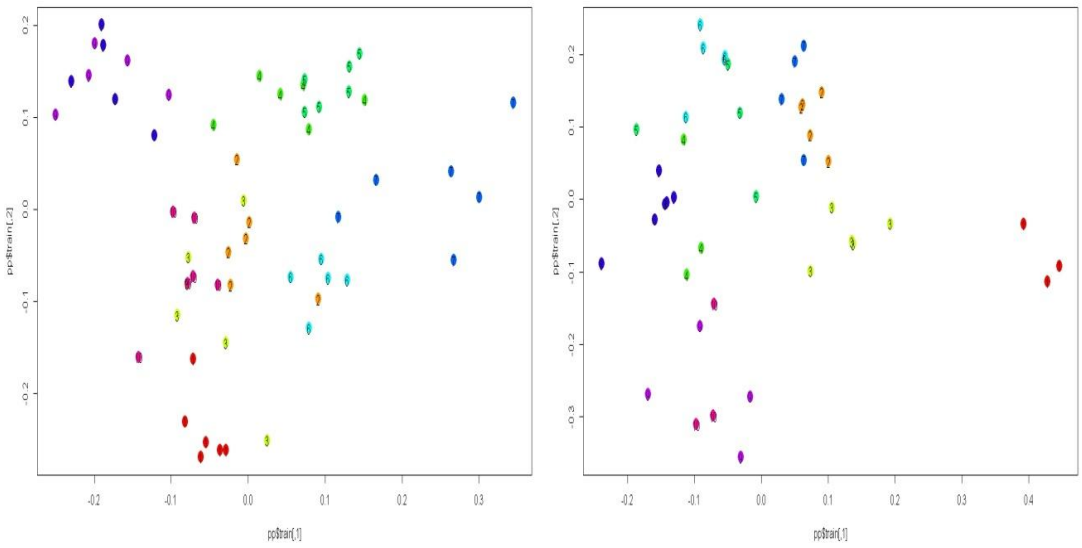


**Figura 31 - Due esempi di PCA/CA/KNN effettuata su un set di campioni appartenenti tutti ad uno stesso individuo per il riconoscimento dei campioni in base al giorno di raccolta.**

## 4.4 *Profiling* metabolico: analisi delle variazioni di singoli metaboliti

Nell'ultima parte del lavoro ci siamo occupati dello studio di selezionati metaboliti [Tabella 1], i cui valori in concentrazioni relative sono stati calcolati attraverso l'integrazione dei rispettivi segnali direttamente sugli spettri NMR non suddivisi in *buckets*. La quantificazione assoluta non è stata effettuata, per cui le concentrazioni sono riportate in unità arbitrarie e sono state valutate esclusivamente le differenze relative tra i vari gruppi in esame.

Per prima cosa siamo andati ad analizzare le differenze tra i livelli dei metaboliti in base al genere. I metaboliti con p-valore<0.005 e quindi considerati significativamente discriminanti, sono risultati essere solamente tre, tutti in concentrazione maggiore nei soggetti di sesso maschile: lattato, trimetilammina e colina. [Figura 32]

Il fatto che la concentrazione media relativa della maggior parte dei metaboliti in esame sia pressoché la stessa tra maschi e femmine sottolinea nuovamente, come era stato visto dalle precedenti analisi sul *fingerprint* metabolico, che non esiste una chiara differenziazione legata al genere, come invece è stata riscontrata in campioni di urine.

Questi risultati sono però parzialmente in disaccordo con quelli riportati in lavori precedenti[28], dove era stata invece rilevata una distinzione più evidente ed erano emersi molti metaboliti, anche tra quelli presi in considerazione in questo studio, statisticamente significativi nel discernimento tra maschi e femmine, per cui ulteriori verifiche si rendono necessarie.

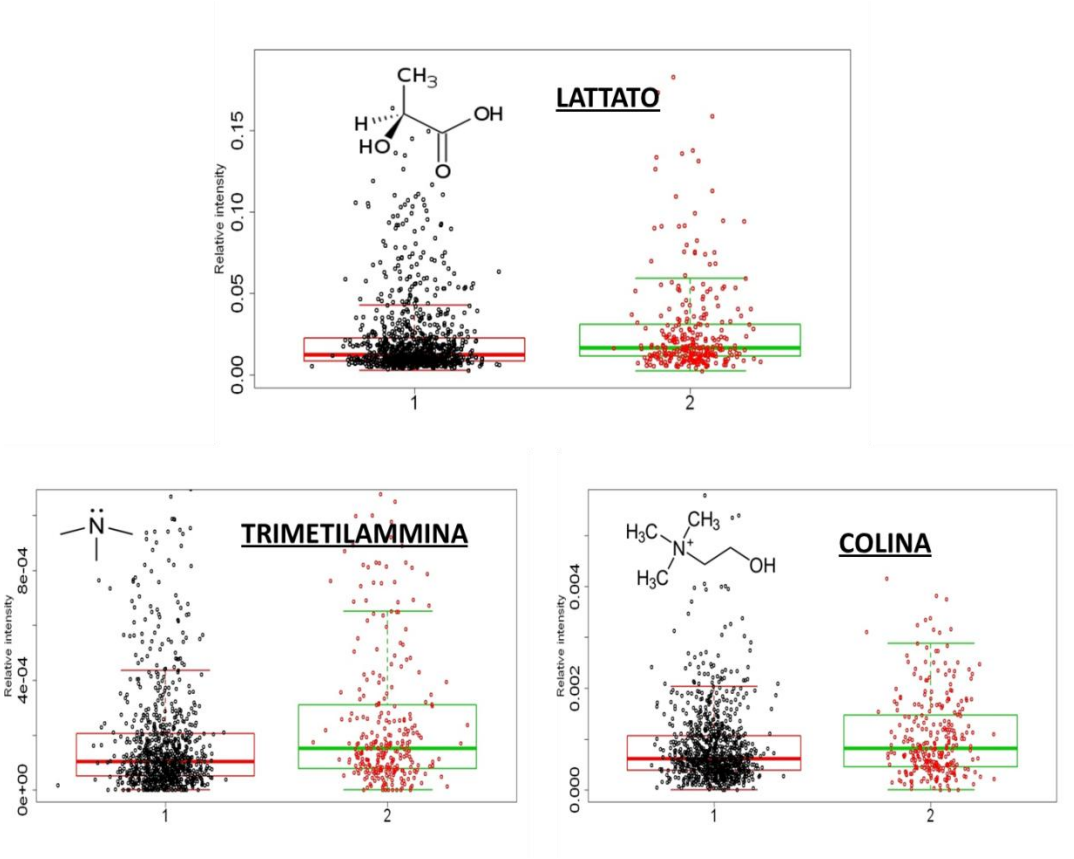



**Figura 32 – *Boxplot* in cui sono riportate le intensità relative dei segnali corrispondenti ai metaboliti più significativi nel confronto tra i soggetti di sesso femminile (1) e quelli di sesso maschile (2).**

In secondo luogo, si è voluto indagare se l'effetto di stimoli esterni come la variazione della dieta o dell'attività fisica comportasse un cambiamento rilevante nella concentrazione delle molecole in esame.

Raccogliendo perciò gli spettri NMR in tre gruppi (dieta libera, dieta standardizzata, esercizio fisico) e andando a vedere come varia la concentrazione di ciascun metabolita in ognuno di essi, 12 dei 23 metaboliti in esame sono risultati essere statisticamente significativi nella discriminazione tra i gruppi. Nello specifico, le concentrazioni relative di butirrato, succinato e tirosina aumentano nel corso degli ultimi 3 giorni, mentre lattato, formiato e TMAO hanno una leggera diminuzione. Propionato e glicole propilenico hanno invece un debole aumento di concentrazione media soltanto nei giorni 8 e 9, mentre il decimo giorno torna ai livelli iniziali, dunque non si può escludere che sia un effetto casuale piuttosto che legato all'introduzione di un'alimentazione standardizzata.

Un aumento durante il decimo giorno, quindi presumibilmente legato all'introduzione di sforzo fisico [Figura 33], si osserva per amminoacidi come alanina, fenilalanina, tirosina e isoleucina.

La molecola che sembra invece essere maggiormente influenzata dalla dieta standardizzata è il metanolo, la cui concentrazione subisce una drastica riduzione nel corso degli ultimi 3 giorni [Figura 34].

**Tabella 4 – Elenco dei metaboliti e delle rispettive intensità relative (mediana ± deviazione standard sulla mediana) nei tre gruppi di giornate, ed infine i p-valori relativi ai confronti tra i gruppi (in rosso sono evidenziati quelli statisticamente significativi).**

| Metabolita | 1 - Dieta libera | 2 - Dieta standard | 3 - Esercizio Fisico | p-valori |
|---|---|---|---|---|
| **Propionato** | 0.021723 ± 0.012718 | 0.028188 ± 0.017421 | 0.023916 ± 0.016387 | **1vs2 =0.003**<br>2vs3 =10.6<br>1vs3 =2.23 |
| **Lattato** | 0.015435 ± 0.010083 | 0.011803 ± 0.006142 | 0.010814 ± 0.006563 | **1vs2 =0.002**<br>2vs3 =2.75<br>**1vs3 <<0.005** |
| **Alanina** | 0.001233 ± 0.000672 | 0.001317 ± 0.000750 | 0.001751 ± 0.000919 | 1vs2 =17.3<br>2vs3 =0.007<br>**1vs3 <<0.005** |
| **n-Butirrato** | 0.000531 ± 0.000787 | 0.001193 ± 0.001518 | 0.001124 ± 0.001659 | **1vs2 <<0.005**<br>2vs3 =24.4<br>**1vs3 =0.001** |
| **Metanolo** | 0.001106 ± 0.000728 | 0.000483 ± 0.000273 | 0.000426 ± 0.000256 | **1vs2 <<0.005**<br>2vs3 =0.35<br>**1vs3 <<0.005** |
| **Tirosina** | 0.006306 ± 0.002600 | 0.007713 ± 0.002582 | 0.008379 ± 0.002812 | **1vs2 <<0.005**<br>2vs3 =2.09<br>**1vs3 <<0.005** |
| **Fenilalanina** | 0.000376 ± 0.000304 | 0.000455 ± 0.000332 | 0.000610 ± 0.000307 | 1vs2 =2.14<br>**2vs3 <<0.005**<br>**1vs3 <<0.005** |
| **Formiato** | 0.001602 ± 0.001303 | 0.001335 ± 0.000974 | 0.000935 ± 0.000803 | 1vs2 =1.62<br>2vs3 =0.011<br>**1vs3 <<0.005** |
| **Succinato** | 0.005985 ± 0.002190 | 0.006965 ± 0.002836 | 0.007155 ± 0.002640 | **1vs2 <<0.005**<br>2vs3 =20.6<br>**1vs3 <<0.005** |
| **Isoleucina** | 0.000103 ± 0.000096 | 0.000112 ± 0.000112 | 0.000162 ± 0.000124 | 1vs2 =9.39<br>2vs3 =0.014<br>**1vs3<<0.005** |
| **TMAO** | 0.000348 ± 0.000310 | 0.000321 ± 0.000306 | 0.000187 ± 0.000215 | 1vs2 =25.0<br>2vs3 =0.014<br>**1vs3 <<0.005** |
| **Glicole propilenico** | 0.003180 ± 0.001206 | 0.004598 ± 0.001822 | 0.003770 ± 0.001766 | **1vs2 <<0.005**<br>**2vs3 <<0.005**<br>1vs3 =0.035 |

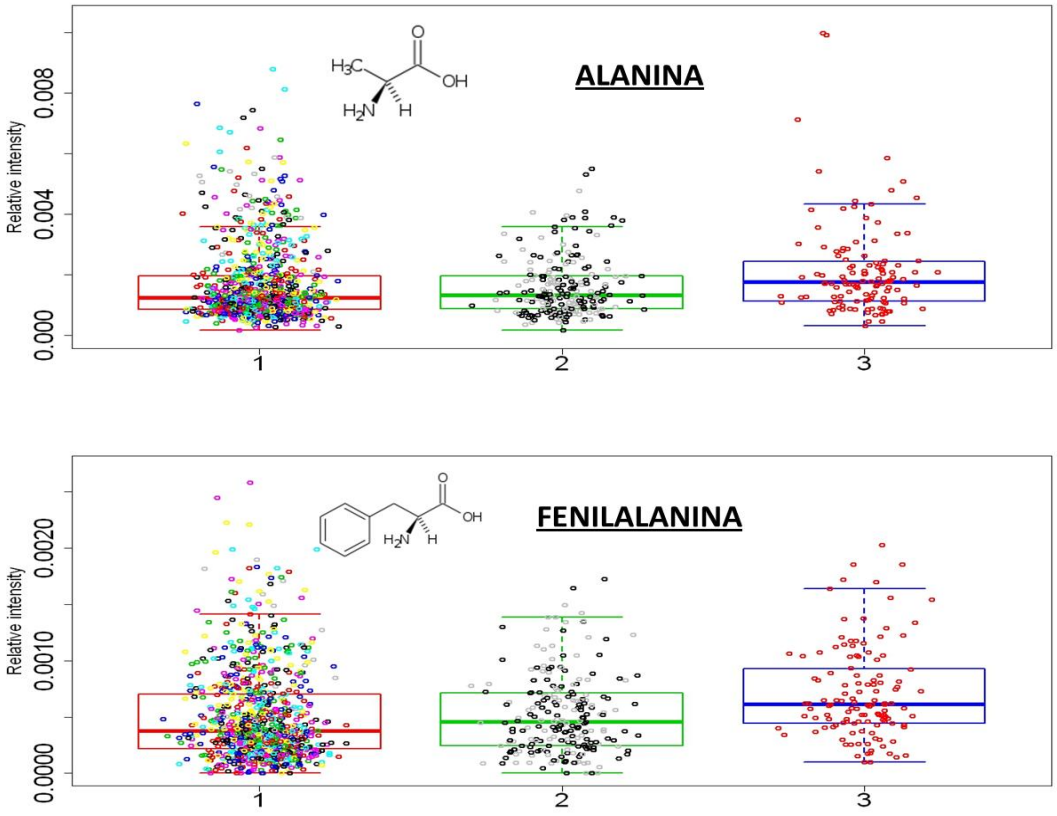



**Figura 33 - *Boxplot* in cui sono riportate le intensità relative dei segnali corrispondenti ad alcuni tra i metaboliti più significativi nel confronto tra i campioni ottenuti nei giorni di dieta libera (1), di dieta standardizzata (2) e di sforzo fisico (3).**

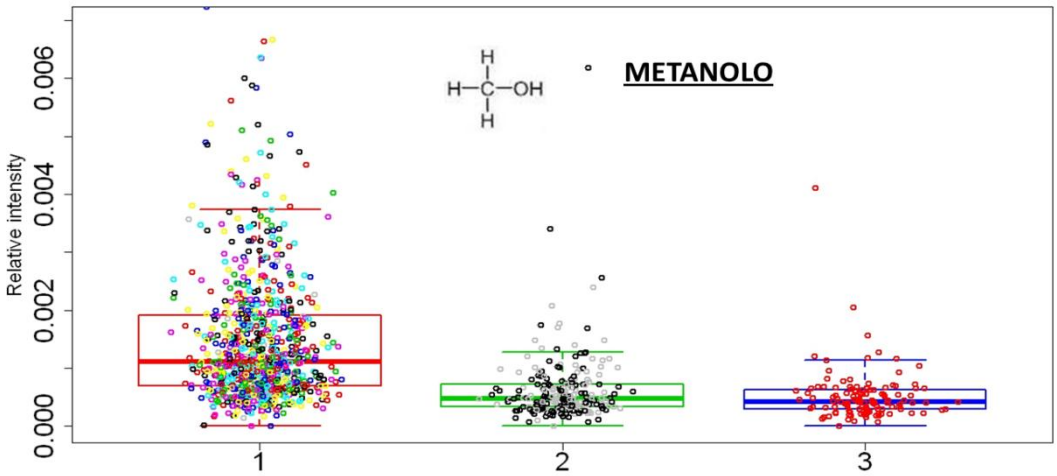



**Figura 34 - *Boxplot* in cui sono riportate le intensità relative del segnale del metanolo nei giorni di dieta libera (1), di dieta standardizzata (2) e di sforzo fisico (3).**

Infine, per quanto riguarda la variazione tra i metaboliti in relazione al momento di raccolta, dai grafici ottenuti è chiaramente visibile la grande differenza del contenuto della prima saliva rispetto al resto, dove la concentrazione rimane praticamente costante.

Dai p-valori ottenuti dal confronto tra la concentrazione di metabolita nel primo campione della giornata e quella media in tutti gli altri campioni, si può vedere come la quasi totalità delle molecole in esame risulti significativamente diversa.

**Tabella 5 - Elenco dei metaboliti e delle rispettive intensità relative (mediana ± deviazione standard sulla mediana) calcolate per i due gruppi in base al momento di raccolta, ed infine i p-valori relativi al confronto statistico tra i due gruppi.**

| Metabolita | 1° campione della giornata | Altri campioni | p-value |
|---|---|---|---|
| **Propionato** | 0.038718 ± 0.016 | 0.020540 ± 0.011 | **<<0.005** |
| Etanolo | 0.008327 ± 0.0027 | 0.008147 ± 0.0026 | 3.98 |
| **Lattato** | 0.018417 ± 0.011 | 0.012506 ± 0.0076 | **<<0.005** |
| **Alanina** | 0.003134 ± 0.0017 | 0.001154 ± 0.00057 | **<<0.005** |
| **n-Butirrato** | 0.001804 ± 0.0017 | 0.000502 ± 0.00074 | **<<0.005** |
| **Acetato** | 0.139804 ± 0.042 | 0.090854 ± 0.041 | **<<0.005** |
| **Gruppi N-acetilici** | 0.034873 ± 0.015 | 0.054833 ± 0.016 | **<<0.005** |
| **Citrato** | 0.000149 ± 0.00018 | 0.000333 ± 0.00032 | **<<0.005** |
| **Metilammina** | 0.000241 ± 0.00011 | 0.000068 ± 0.000043 | **<<0.005** |
| Metanolo | 0.000861 ± 0.00054 | 0.000879 ± 0.00065 | 17.7 |
| **Glicina** | 0.004335 ± 0.0016 | 0.003789 ± 0.0014 | **<<0.005** |
| **Tirosina** | 0.005735 ± 0.0070 | 0.007041 ± 0.0026 | **<<0.005** |
| **Fenilalanina** | 0.000730 ± 0.00049 | 0.000375 ± 0.00029 | **<<0.005** |
| **Formiato** | 0.000963 ± 0.00082 | 0.001543 ± 0.0012 | **<0.005** |
| **Isobutirrato** | 0.000479 ± 0.00039 | 0.000295 ± 0.00026 | **<<0.005** |
| **Piruvato** | 0.001752 ± 0.0012 | 0.000804 ± 0.00062 | **<<0.005** |
| **Succinate** | 0.004317 ± 0.0017 | 0.006653 ± 0.0023 | **<<0.005** |
| **Sarcosina** | 0.000178 ± 0.00011 | 0.000088 ± 0.000052 | **<<0.005** |
| **Trimetilammina** | 0.000565 ± 0.00043 | 0.000094 ± 0.000067 | **<<0.005** |

| | | | |
|---|---|---|---|
| **Colina** | 0.001781 ± 0.00081 | 0.000570 ± 0.00036 | **<<0.005** |
| **Isoleucina** | 0.000252 ± 0.00019 | 0.000092 ± 0.000081 | **<<0.005** |
| TMAO | 0.000275 ± 0.00035 | 0.000343 ± 0.00029 | 6.24 |
| Glicole propilenico | 0.003710 ± 0.0012 | 0.003373 ± 0.0015 | 0.053 |
| **Valina** | 0.000897 ± 0.00081 | 0.000057 ± 0.000085 | **<<0.005** |

In particolare, quasi tutti i metaboliti (15 sui 23 in esame) si trovano in concentrazione maggiore nel primo campione, mentre diminuiscono nettamente durante il resto del giorno. Tra questi quelli che hanno la diminuzione più evidente sono le ammine come metilammina, colina e trimetilammina, quest'ultima caratterizzata dal più piccolo p-valore ottenuto.

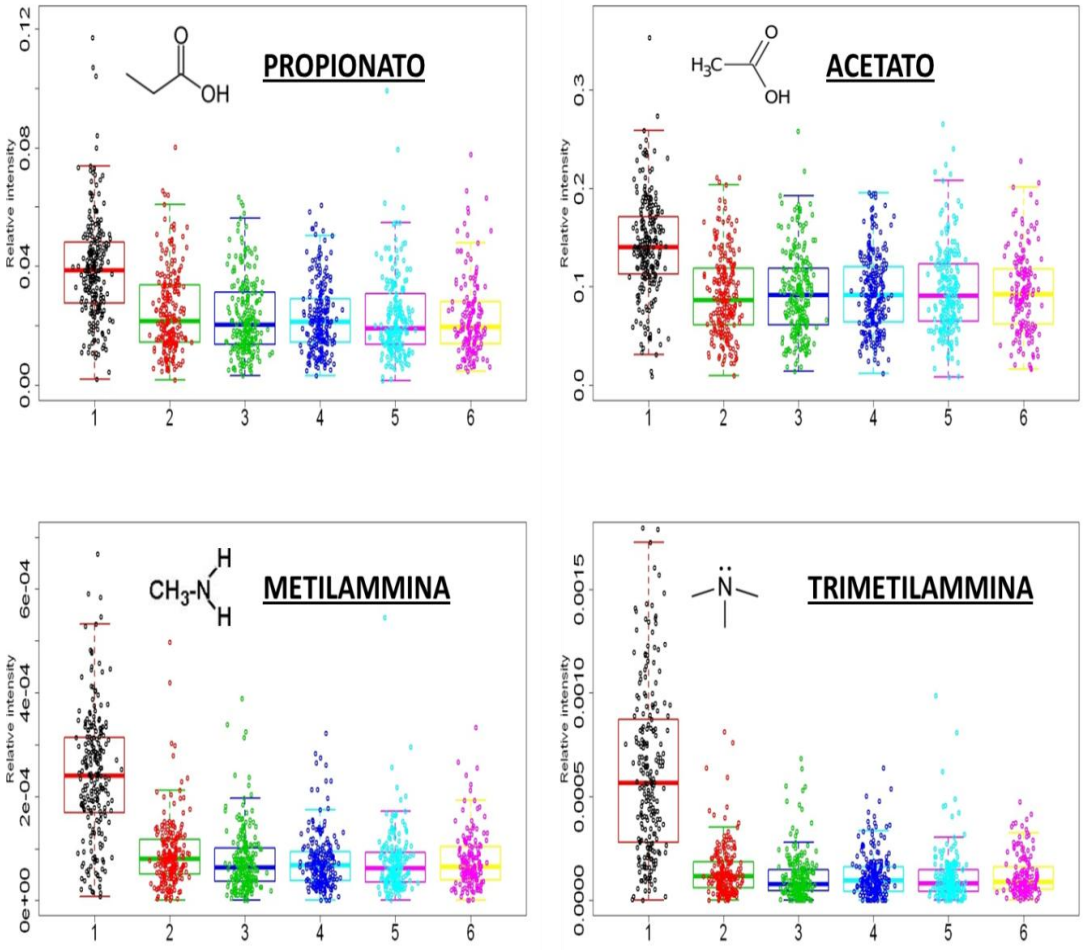



**Figura 35 - *Boxplot* in cui sono riportate le intensità relative dei segnali corrispondenti ad alcuni tra i metaboliti significativi nel confronto basato sul momento della giornata in cui sono stati raccolti i campioni.**

I pochi metaboliti per cui si ottiene un aumento di concentrazione durante la giornata sono citrato, succinato e, seppur di poco, il formiato, ma l'incremento più evidente si osserva per i segnali dei gruppi N-acetilici.

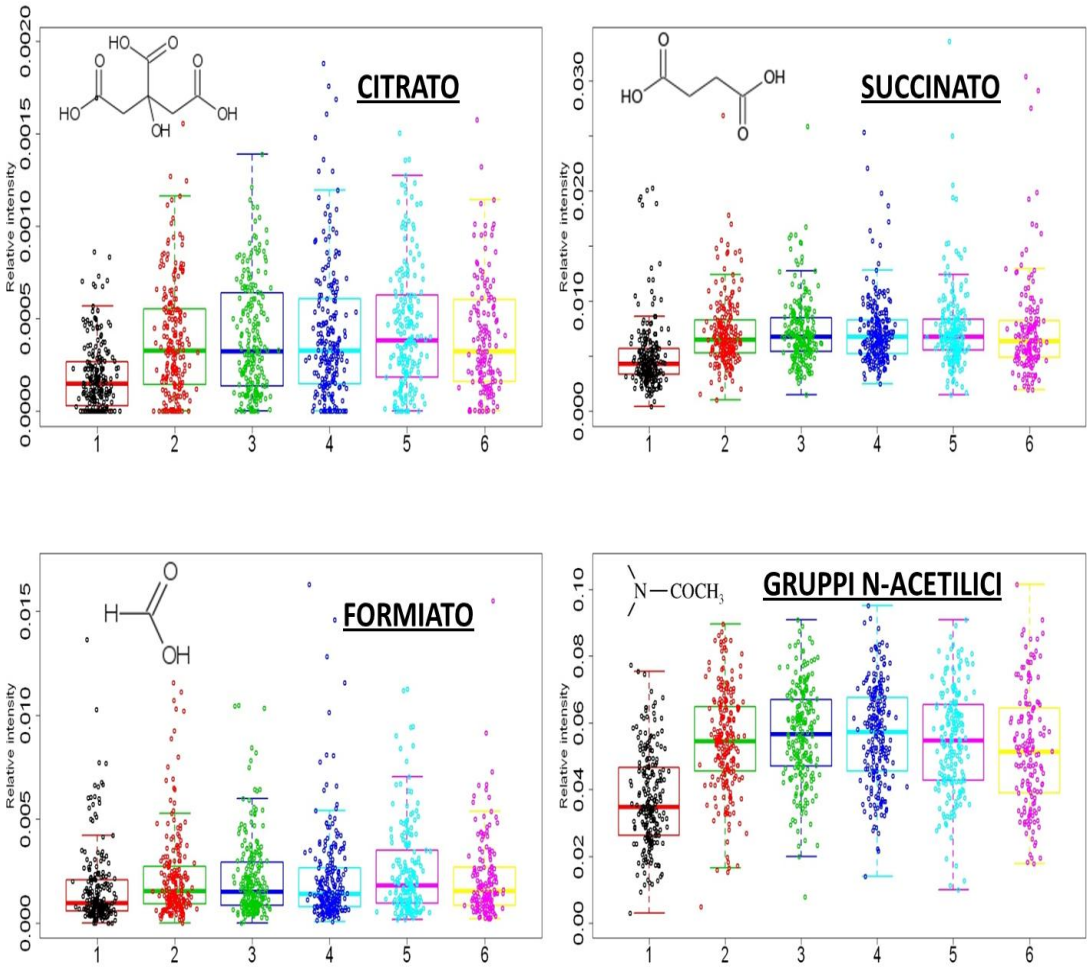



**Figura 36 - *Boxplot* in cui sono riportate le intensità relative dei segnali corrispondenti ad alcuni tra i metaboliti significativi nel confronto basato sul momento della giornata in cui sono stati raccolti i campioni.**

Infine, alcoli come etanolo, metanolo e glicole propilenico non subiscono variazioni sostanziali durante la giornata, ma il loro livello si mantiene pressoché costante, così come quello della trimetilammina N-ossido, per la quale è visibile una leggera diminuzione, che però non risulta statisticamente significativa.

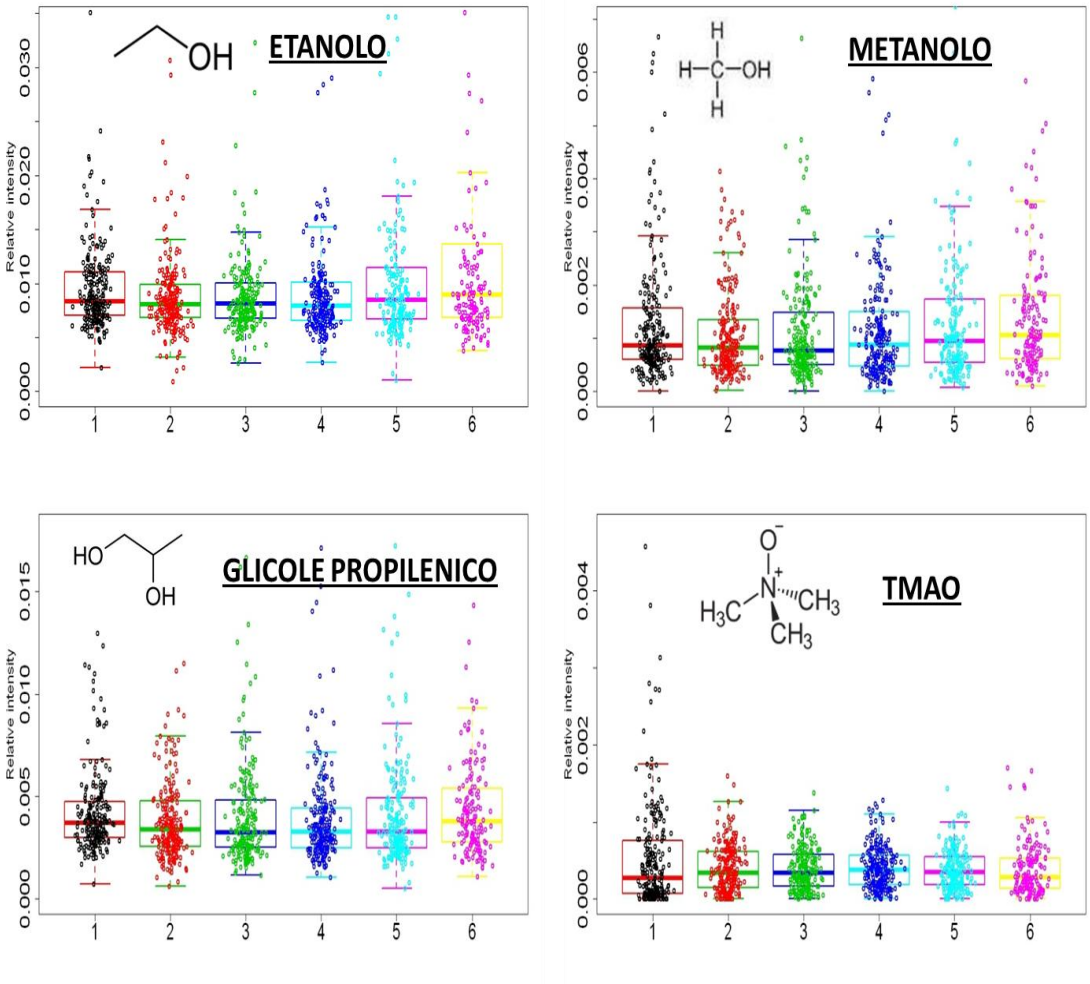



**Figura 37 - *Boxplot* in cui sono riportate le intensità relative dei segnali corrispondenti ai metaboliti che non risultano statisticamente significativi nel confronto basato sul momento della giornata in cui sono stati raccolti i campioni.**

Per concludere, i risultati mostrati, seppur riguardanti solo una piccola parte delle molecole che costituiscono il metaboloma salivare, sottolineano ulteriormente che l'effetto principale che influenza la composizione e la concentrazione di un campione è legato al momento di raccolta durante la giornata, seguito da quello dovuto all'introduzione di fattori esterni come la variazione della dieta e solo in minima parte può essere considerato correlato al sesso dell'individuo che ha fornito il campione.

# 5 CONCLUSIONI

Fino ad ora il potenziale della saliva come biofluido di indagine negli studi di metabolomica non era ancora stato completamente approfondito. Con questo lavoro di tesi si dimostra che, al pari di urine e sangue, ma con il grande vantaggio di poter essere raccolta in maniera ancora più semplice e meno invasiva, la saliva si rivela una fonte estremamente ricca di informazioni metaboliche. I risultati riportati rimarcano il valore e la rilevanza dell'analisi dei profili metabolici applicata a spettri NMR di saliva.

La più interessante scoperta consiste nel fatto che anche per essa viene verificata la possibilità di eliminare il rumore dovuto alle variazioni giornaliere casuali grazie alla collezione di numerosi campioni nell'arco di poco tempo per uno stesso individuo, e di conseguenza viene attestata la presenza di un *fingerprint* metabolico individuale costituito dalla parte invariante dei campioni appartenenti ad un singolo soggetto. Come si evince dalle numerose analisi e dalle elevate accuratezze dei relativi modelli di riconoscimento, anche per la saliva viene confermata sperimentalmente l'esistenza di *metabotypes* specifici per una determinata persona, in quanto è effettivamente possibile riconoscere un soggetto in un gruppo di individui con il 100% di probabilità. Si dimostra inoltre che questi continuano ad essere perfettamente visibili anche in condizioni di dieta standardizzata.

Questa distinzione così accurata sembra dipendere però solo in minima parte dal genere del soggetto, osservazione che appare confermata anche dall'analisi di singoli metaboliti, dalla quale è emerso che solo 3 molecole (lattato, trimetilammina e colina) tra quelle esaminate risultano essere significativamente maggiori nei soggetti di sesso maschile.

Anche l'esplorazione delle differenze intraindividuali e delle variazioni negli spettri NMR collegabili alla dieta o all'esercizio fisico ha condotto a importanti risultati.

Per prima cosa, la sostanziale differenza che intercorre tra il primo campione prelevato alla mattina e tutti gli altri raccolti durante il resto della giornata, i quali invece non risultano essere distinguibili tra loro, è stata rilevata sia dall'analisi chemiometrica su tutti i campioni, sia da quella da quella effettuata su ciascun

individuo, sia infine dall'investigazione condotta su singoli metaboliti. Questa diversità appare fondamentale nell'ottica di futuri studi su campioni di saliva, e utile nella comprensione del processo di secrezione da parte delle varie ghiandole, che ne influenzano differentemente la composizione. Determinante in tal senso è l'indagine sui metaboliti, che rivela come la concentrazione relativa di moltissime molecole sia più elevata nella saliva raccolta al risveglio rispetto alle altre.

In secondo luogo, lo stesso tipo di procedura analitica mette in luce come le differenze dovute all'introduzione della dieta o dell'attività fisica siano molto meno evidenti, per quanto comunque rilevabili. Quello che appare più interessante sono i risultati ottenuti dall'analisi dei campioni appartenenti a ciascun individuo, i quali dimostrano che la risposta metabolica agli stimoli esterni varia molto da soggetto a soggetto: un profilo metabolico quotidiano è fondamentale per monitorare il responso individuale all'esposizione alimentare e agli effetti di una dieta, e questo sottolinea come la nutrizionistica dovrebbe far affidamento sempre maggiore sugli approcci metabolomici. Infine, l'esercizio fisico non sembra avere un contributo particolarmente rilevante, e ciò è deducibile sia dall'analisi globale degli spettri, sia da quella sui singoli metaboliti, da cui emerge che solo amminoacidi come alanina e fenilalanina subiscono un leggero ma comunque statisticamente significativo aumento durante l'ultimo giorno di raccolta.

Tutti questi risultati, integrati con quelli ottenuti dall'analisi degli altri biofluidi, costituiscono un grande passo in avanti nella comprensione del metaboloma umano e delle sue interazioni con l'ambiente esterno.

Inoltre, l'identificazione di un fenotipo metabolico salivare specifico per ciascun individuo apre le porte ad una nuova, promettente linea di ricerca per cui la saliva, grazie ad i suoi preziosi vantaggi, potrebbe costituire un biofluido fondamentale nella costruzione di *database* a cui fare riferimento nella prognosi e della diagnosi di malattie, nella progettazione di terapie o diete personalizzate, nel seguire gli sviluppi di un trattamento farmacologico per un dato soggetto, monitorandone la risposta metabolica così da ottenere la massima efficacia. Per raggiungere questi obiettivi, ulteriori studi si rendono necessari; in particolare, sarebbe auspicabile verificare la stabilità del *fingerprint* salivare in un tempo molto più lungo e chiarire ancor più nel dettaglio le sue reazioni agli stimoli esterni, tuttavia, in accordo con

quanto emerso da questa ricerca, non sembra immotivato immaginare la saliva come futuro, utile oggetto di indagine nello *screening high throughput* e su larga scala.

# 6 ABBREVIAZIONI

RNA = Ribonucleic Acid;

NMR = Nuclear Magnetic Resonance;

MS = Mass Spectrometry;

GC = Gas Cromatography;

HPLC = High Pressure (o Performance) Liquid Cromatography;

CE = Capillary Electrophoresis;

IR = Infra-Red;

RF = Radio Frequency;

FID = Free Induction Decay;

J-res = J-resolved;

COSY = Correlation Spectroscopy;

TOCSY = Total Correlation Spectroscopy;

HSQC = Heteronuclear Single Quantum Coherence;

PCA = Principal Component Analysis;

PLS = Partial Least Square;

CA = Canonical Analysis;

KNN = k-Nearest Neighbour;

EBC = Exhaled Breath Condensate;

ATM = Automatic Tuning Matching;

PQN = Probabilistic Quotient Normalization;

CV = Cross-Validation;

TMAO = Trimethylamine N-Oxide.

# 7 BIBLIOGRAFIA


1. Nicholson, J. K. & Wilson, I. D. Opinion: understanding 'global' systems biology: metabonomics and the continuum of metabolism. *NatRevDrug Discov* **2,** 668–676 (2003).
2. Joyce, A. R. & Palsson, B. Ø. The model organism as a system: integrating 'omics' data sets. *Nat. Rev. Mol. Cell Biol.* **7,** 198–210 (2006).
3. Nicholson, J. K., Lindon, J. C. & Holmes, E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica Fate Foreign Compd. Biol. Syst.* **29,** 1181–1189 (1999).
4. Fiehn, O. Metabolomics--the link between genotypes and phenotypes. *Plant MolBiol* **48,** 155–171 (2002).
5. Lindon, C. J., Nicholson, J. K. & Holmes, E. *The Handbook of Metabonomics and Metabolomics*. (Elsevier, 2007).
6. Hollywood, K., Brison, D. R. & Goodacre, R. Metabolomics: Current technologies and future trends. *PROTEOMICS* **6,** 4716–4723 (2006).
7. Nielsen, J. & Oliver, S. The next wave in metabolome analysis. *Trends Biotechnol.* **23,** 544–546 (2005).
8. Ellis, D. I., Dunn, W. B., Griffin, J. L., Allwood, J. W. & Goodacre, R. Metabolic fingerprinting as a diagnostic tool. *Pharmacogenomics* **8,** 1243–1266 (2007).
9. Lindon, C. J., Nicholson, J. K. & Everett, J. R. NMR Spectroscopy of Biofluids. *Annu Rep NMR Spectro* **38,** 1–88 (1999).
10. Robertson, D. G., Reily, M. D. & Baker, J. D. Metabonomics in Pharmaceutical Discovery and Development. *J. Proteome Res.* **6,** 526–539 (2007).
11. Robertson, D. G., Reily, M. D. & Baker, J. D. Metabonomics in preclinical drug development. *Expert Opin. Drug Metab. Toxicol.* **1,** 363–376 (2005).
12. Hopson, R. E. & Peti, W. Microcoil NMR spectroscopy: a novel tool for biological high throughput NMR spectroscopy. *Methods Mol. Biol. Clifton NJ* **426,** 447–458 (2008).
13. Pack, S. Factor analysis in chemistry, (2nd edition), E. R. Malinowski, Wiley-Interscience, 1991. ISBN 0-471-53009-3. Price £43.70. *J. Chemom.* **5,** 545–545 (1991).
14. Trygg, J., Holmes, E. & Lundstedt, T. Chemometrics in metabonomics. *J. Proteome Res.* **6,** 469–479 (2007).
15. Nicholson, J. K., Connelly, J., Lindon, J. C. & Holmes, E. Metabonomics: a platform for studying drug toxicity and gene function. *Nat. Rev. Drug Discov.* **1,** 153–161 (2002).
16. Madsen, R., Lundstedt, T. & Trygg, J. Chemometrics in metabolomics—A review in human disease diagnosis. *Anal. Chim. Acta* **659,** 23–33 (2010).
17. Calabrò *et al.* A Metabolomic Perspective on Coeliac Disease. *Autoimmune Dis.* **2014,** (2014).
18. Gavaghan, C. L., Holmes, E., Lenz, E., Wilson, I. D. & Nicholson, J. K. An NMR-based metabonomic approach to investigate the biochemical consequences of genetic strain differences: application to the C57BL10J and Alpk:ApfCD mouse. *FEBS Lett.* **484,** 169–174 (2000).

19. Assfalg, M. *et al.* Evidence of different metabolic phenotypes in humans. *Proc.Natl.Acad.Sci.U.S.A* **105,** 1420–1424 (2008).
20. Bernini, P. *et al.* Individual human phenotypes in metabolic space and time. *J ProteomeRes* **8,** 4264–4271 (2009).
21. Holmes, E. *et al.* Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature* **453,** 396–400 (2008).
22. Relationships between the metabolome and the fatty acid composition of human saliva; effects of stimulation - Springer. at <http://link.springer.com/article/10.1007%2Fs11306-012-0440-6/fulltext.html#Fig1>
23. Metabolic profiling of human saliva before and after induced physiological stress by ultra-high performance liquid chromatography–ion mobility–mass spectrometry - Springer. at <http://link.springer.com/article/10.1007%2Fs11306-013-0541-x/fulltext.html>
24. Kaufman, E. & Lamster, I. B. The Diagnostic Applications of Saliva— A Review. *Crit. Rev. Oral Biol. Med.* **13,** 197–212 (2002).
25. Aimetti, M., Cacciatore, S., Graziano, A. & Tenori, L. Metabonomic analysis of saliva reveals generalized chronic periodontitis signature. *Metabolomics* **8,** 465–474 (2012).
26. Bertram, H. C., Eggers, N. & Eller, N. Potential of human saliva for nuclear magnetic resonance-based metabolomics and for health-related biomarker identification. *Anal. Chem.* **81,** 9188–9193 (2009).
27. Silwood, C. J. L., Lynch, E., Claxson, A. W. D. & Grootveld, M. C. 1H and 13C NMR Spectroscopic Analysis of Human Saliva. *J. Dent. Res.* **81,** 422–427 (2002).
28. Takeda, I. *et al.* Understanding the human salivary metabolome. *NMR Biomed.* **22,** 577–584 (2009).
29. Stella, C. *et al.* Susceptibility of human metabolic phenotypes to dietary modulation. *J Proteome Res* **5,** 2780–2788 (2006).
30. Bertini, I., Luchinat, C., Miniati, M., Monti, S. & Tenori, L. Phenotyping COPD by 1H NMR metabolomics of exhaled breath condensate. *Metabolomics* 1–10 doi:10.1007/s11306-013-0572-3
31. Ihaka, R. & Gentleman, R. R: A Language for Data Analysis and Graphics. *J Comput Stat Graph* **5,** 299–314 (1996).