Università degli studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in

Scienze Statistiche

Relazione Finale

Confronto di metodi statistici in uno studio di associazione genomica per la dipendenza da nicotina

Relatore Prof. Davide Risso
Dipartimento di Scienze Statistiche

Laureanda Elisabetta Dargenio
Matricola N 1236614

Anno Accademico 2020/2021

A mia nonna Teresa

Indice

Elenco delle figure

Introduzione

Con il completamento dello "Human Genome Project" nel 2001 si è ottenuta una mappa quasi completa del genoma umano. Con l'avvento delle tecnologie che hanno permesso di fare una genotipizzazione su larga scala, si sono sviluppati anche i *Genome Wide Association Studies* (GWAS). I GWAS nascono con l'obiettivo di cercare associazioni tra le variazioni genetiche - note come Polimorfismi a Singolo Nucleotide (SNP) - e una malattia complessa (fenotipo), come ad esempio il diabete o le diverse forme di cancro.
Solo molti anni dopo si è iniziato a condurre GWAS studiando tratti più complessi di interesse sociologico, come ad esempio il livello di educazione o l'uso di alcol e tabacco (Tanksley et al., 2019). Le analisi di associazione fatte soprattutto in questi nuovi GWAS sono semplicistiche: dato il numero molto elevato di SNP da testare, le analisi vengono svolte solitamente utilizzando modelli univariati, ma questo dal punto di vista statistico comporta dei problemi.
Il primo problema è la scelta della soglia utilizzata per ritenere un'associazione statisticamente significativa. In genere si applica la correzione di Bonferroni, che impone una soglia troppo stringente, e rischia di non cogliere le reali associazioni presenti nei dati. Spesso inoltre si assume indipendenza tra i test, ma in realtà è noto che gli SNP siano tra loro correlati, sia per il *Linkage Disequilibrium* (LD), sia perché SNP adiacenti sono di norma coinvolti negli stessi percorsi biologici.
Il secondo problema riguarda la classe di variabili usata nei modelli. Prima di adattare i modelli, i genotipi vengono solitamente ricodificati in numeri, poiché questo rende la modellazione più agevole, risultando in un solo p-value da valutare per ogni modello adattato. L'utilizzo di variabili categoriali come

continue, però, implica l'assegnazione di grandezze ordinate ai livelli della variabile, che significa attribuire ad un genotipo valore più elevato di un altro, in modo arbitrario.
Il terzo ed ultimo problema evidenziato in questo elaborato, è l'assenza di variabili confondenti nei modelli. Generalmente l'unico aspetto che viene preso in considerazione è il controllo per gruppi etnici, eseguendo i test di associazione separatamente per ogni gruppo e poi ricombinando i risultati. Ma è noto che negli studi di espressione genica su larga scala possono esserci diversi fattori che influenzano l'espressione di un gene, come ad esempio fattori tecnici o ambientali, che sarebbe bene inserire nell'analisi laddove possibile.
Per superare le lacune statistiche delle analisi classiche dei GWAS, in questo elaborato vengono presentati metodi statistici differenti rispetto a quelli abitualmente utilizzati, introducendo anche metodi per dati ad alta dimensionalità sviluppati per altri contesti, ma che trovano applicazione anche nei dati genomici. Si mostra l'applicazione di un modello che ricalca quello classico univariato, ma con delle correzioni per risolvere alcuni dei deficit statistici, e in seguito di modelli logistici multivariati con metodi di regolarizzazione di tipo lasso, elastic net e fused lasso. Questi metodi vengono applicati a uno studio di associazione *genome-wide* sulla dipendenza da nicotina. I dati analizzati sono stati estratti da openSNP, un progetto open source che consente ai partecipanti dei test genetici *Direct-To-Consumer* di pubblicare gratuitamente i propri dati genetici insieme alla informazioni fenotipiche, in modo da renderli disponibili per la ricerca.

Con l'obiettivo di presentare metodi statistici più corretti per i GWAS, l'elaborato è così strutturato: nel Capitolo 1 vengono mostrati i metodi classici usati per i GWAS, evidenziandone i limiti; nel Capitolo 2 si presentano i metodi statici per dati ad alta dimensionalità; nel Capitolo 3 si riporta un esempio di applicazione a dati reali sulla dipendenza da nicotina, dei metodi classici e di quelli proposti in questo elaborato; nel Capitolo 4 si mostrano i risultati dell'analisi.

Capitolo 1

Studi di associazione genomica

1.1 Genotipizzazione

Ogni specie è definita da un insieme distinto di caratteristiche comuni, ma anche all'interno di una specie vi sono differenze tra gli individui. Col termine "variazione" si descrive la caratteristica che è diversa tra individui distinti all'interno di una singola specie. La genotipizzazione è la procedura sperimentale che identifica le differenze nella sequenza del DNA tra individui o popolazioni. Questa viene utilizzata per comprendere la connessione tra genotipi e fenotipi. Un genoma presenta una variante quando differisce da una sequenza di riferimento, che è derivata dalla popolazione generale o da un sottogruppo definito. Una sequenza variante può differire dalla sequenza di riferimento in molti modi. I tipi di variazione genetica includono varianti a singolo nucleotide (SNV), polimorfismi a singolo nucleotide (SNP), inserimenti e delezioni (indel) e variazione del numero di copie (CNV).
I polimorfismi a singolo nucleotide sono il tipo più comune di variante di sequenza studiata dai ricercatori; sono tipicamente definiti come SNV che si verificano in almeno l'1% della popolazione.
Confrontando le variazioni genetiche tra gli individui di una specie, i ricercatori possono identificare "firme genetiche" ereditabili, o marcatori rilevanti per tratti specifici.
Gli studi di associazione a livello di genoma, o GWAS, confrontano le diffe-

renze genetiche tra interi genomi di due individui o popolazioni. Ad esempio, i genomi di un gruppo di persone che hanno una malattia possono essere paragonati a sequenze genomiche di un gruppo simile di persone senza la malattia. Qualsiasi SNP o aplotipo (serie di SNP adiacenti) che è più prevalente in quelli con la malattia è chiamato marcatore genetico associato («What is genotyping? | IDT», 2021).
Negli esseri umani, il 99,9% di tutte le basi nel genoma, da individuo a individuo, sono le stesse. Il restante 0,1% rende unica una persona.
Le variazioni presenti nel genoma possono essere di tre tipologie: innocue, che non provocano alcun cambiamento nel fenotipo (la maggior parte degli SNP); dannose, che causano malattie, come ad esempio diabete, cancro, malattie cardiache; latenti, riscontrate nelle regioni codificanti e regolatorie del genoma che di per sé non sono dannose, il loro cambiamento nella sequenza diventa evidente solo in determinate condizioni, come la suscettibilità ai tumori o la risposta ai farmaci (Prediger, 2019).

1.2 Genome Wide Association Studies (GWAS)

Nel 2001 è stato completato lo "Human Genome Project", riportando una mappa quasi completa del genoma umano. Questo progetto ha fornito una mappa, ma non le informazioni su quante variazioni ci fossero tra i genomi e dove queste si trovassero. Quelle informazioni, che sono cruciali per eseguire i GWAS, provengono da database in cui i dati sul polimorfismo sono stati raccolti nel corso degli anni ("dbSNP"); e dagli sforzi di risequenziamento di panel di genomi di riferimento raccolti nell'ambito del progetto "HapMap" (Uitterlinden, 2016). Parallelamente e alimentate da questi progetti, anche le tecnologie per l'analisi del DNA hanno fatto grandi progressi, portando alla tecnologia dei microarray che hanno permesso di genotipizzare il DNA per centinaia di migliaia di SNP.
In particolare due società, Affymetrix e Illumina, hanno realizzato prodotti commerciali che hanno aperto la possibilità di fare una genotipizzazione su larga scala, e valutare centinaia di migliaia di SNP in relazione a fenotipi e malattie, in centinaia di soggetti umani. Gli scienziati si sono resi conto che i progetti di studi epidemiologici sarebbero stati ottimali per trovare

varianti genetiche causali per malattie complesse, sono stati creati database su larga scala di variazioni genetiche tra gli esseri umani e la tecnologia di analisi del DNA ha fornito strumenti per eseguire tali esperimenti. É in questo contesto che nel 2004 nascono i *Genome Wide Association Studies* (GWAS) (Uitterlinden, 2016).

1.3 Analisi classica dei dati GWAS

Un passaggio iniziale, prima della vera e propria analisi, è un controllo di qualità (QC) appropriato. Errori nei dati possono sorgere per numerose ragioni, ad esempio, a causa della scarsa qualità dei campioni di DNA, della scarsa ibridazione del DNA nell'array, delle sonde genotipiche con prestazioni scadenti e della contaminazione o confusione dei campioni.

1.3.1 Controllo della qualità

Gli step che vengono solitamente svolti per ottenere un campione di genotipo affidabile indicati da Marees et al., 2018 sono sette:

- **Controllo dei valori mancanti** sui singoli SNP e sugli individui: rimuovere dal campione i soggetti e gli SNP con una percentuale di valori mancanti superiore a una soglia scelta
- **Discrepanza tra i sessi**: i maschi dovrebbero avere una stima dell'omozigosi del cromosoma X >0.8 e le femmine dovrebbero avere un valore <0.2; se molti soggetti hanno questa discrepanza, i dati dovrebbero essere controllati con attenzione.
- **Soglia di frequenza allelica minore (MAF)**: includere solo gli SNP con MAF superiore alla soglia scelta; gli SNP con un MAF basso sono rari, quindi manca la potenza per rilevare le associazioni SNP-fenotipo.
- **Deviazioni dall'equilibrio di Hardy-Weinberg (HWE)**: l'HWE è un modello della genetica delle popolazioni che postula che all'interno di una popolazione (ideale), vi è equilibrio delle frequenze alleliche e genotipiche da una generazione all'altra.

- **Tasso di eterozigosi**: escludere gli individui con tassi di eterozigosi alti o bassi; rimuovere gli individui che hanno ±3 deviazione standard dalla media del tasso di eterozigosi dei campioni.

- **Parentela**: impostare una soglia e creare un elenco di individui con parentela al di sopra della soglia scelta (ad esempio secondo grado di parentela).

- **Controllo per gruppi etnici**: nel caso in cui nel campione siano presenti diversi gruppi etnici (ad es. americani, asiatici, europei) è raccomandato eseguire i test di associazione separatamente per ogni gruppo e poi combinare i risultati, ad esempio attraverso una meta-analisi.

1.3.2 Test di associazione statistica

In seguito al controllo della qualità, i file risultati dalla genotipizzazione tramite microarray, vengono sottoposti a una valutazione statistica dell'associazione del genotipo con il fenotipo di interesse. I genotipi possono essere codificati come 0, 1 e 2, rispettivamente per i soggetti omozigoti per l'allele di riferimento, eterozigoti e omozigoti per l'allele variante. Per le analisi si possono utilizzare la regressione logistica, in caso di fenotipo dicotomico, o la regressione lineare nel caso di tratti quantitativi. Il tipo di analisi è relativamente semplice, ma i dataset sono molto grandi e richiedono una potenza di calcolo e capacità di archiviazione sufficienti. I risultati di un'analisi GWAS sono tipicamente i p-value per ogni SNP, riportati sotto forma di Manhattan plot oppure in un QQ plot come p-value osservati contro p-value attesi. Poiché viene effettuato un numero enorme di confronti statistici, è necessaria una correzione per il livello di significatività. Quello che di solito si fa è cercare una soglia inferiore al classico 0.05, sotto cui poter considerare l'associazione tra SNP e fenotipo significativa, in modo da tenere conto di tutte le associazioni risultate significative per caso. La soglia viene scelta principalmente in due modi: utilizzando la correzione di Bonferroni, dunque dividendo 0.05 per il numero di test indipendenti eseguiti; oppure viene scelta pari a $5 \cdot 10^{-8}$. Quest'ultimo valore deriva dalla divisione di 0.05 per 1 milione di test indipendenti in base al numero di blocchi LD, quindi significa

avere una significatività del 5% tenendo conto del fatto che all'interno dei blocchi LD non ci sia indipendenza tra i test. Con LD si intende *Linkage Disequilibrium*: nel progetto "HapMap" è stata stimata l'esistenza di un milione di blocchi di *Linkage Disequilibrium*, tratti di diversi milioni di paia di basi in cui le varianti mostrano forti correlazioni tra loro (Uitterlinden, 2016).

Consideriamo ad esempio due loci collegati: Locus 1 ha alleli A_1, A_2, ... , A_m, che si verifica alle frequenze p_1, p_2, ... , p_m e Locus 2 ha alleli B_1, B_2, ... , B_n che si verificano alle frequenze q_1, q_2, ... , q_n nella popolazione. I possibili aplotipi possono essere indicati come A_1B_1, A_1B_2, ... , A_mB_n con frequenze h_{11}, h_{12}, ... , h_{mn}. I due loci collegati sono detti in *Linkage Equilibrium* (LE), se l'occorrenza dell'allele A_i e l'occorrenza dell'allele B_j in un aplotipo sono eventi indipendenti. Al contrario, gli alleli sono in *Linkage Disequilibrium* (LD) quando non si verificano casualmente. Sotto *Linkage Disequilibrium*, gli aplotipi non si verificano alle frequenze previste nel caso di indipendenza tra gli alleli. Si ha LD positivo quando due alleli si presentano insieme sullo stesso aplotipo più spesso del previsto, e LD negativo quando gli alleli si presentano insieme sullo stesso aplotipo meno spesso del previsto (Calabrese, 2019).

Una volta che uno SNP è stato identificato come significativo per l'intero genoma, viene eseguita un'ulteriore analisi bioinformatica dettagliata per esaminare la regione locus in cui si trova lo SNP e determinare l'esatta struttura del blocco LD e quali SNP potenzialmente funzionali possono essere identificati nel blocco LD (Uitterlinden, 2016).
Poiché i risultati dei GWAS hanno mostrato che le dimensioni dell'effetto dei singoli SNP sono piccole, i ricercatori hanno sviluppato un interesse per i metodi che aggregano l'effetto degli SNP. Uno dei più usati è il punteggio di rischio poligenico (PRS), relativamente facile da calcolare e che può essere applicato a campioni con dimensioni modeste. Il PRS combina le dimensioni dell'effetto di più SNP in un singolo punteggio individuale aggregato che può essere utilizzato per prevedere il rischio di malattia. Viene calcolato in base al numero di varianti di rischio che una persona possiede, pesato per le dimensioni dell'effetto degli SNP, ricavati da un GWAS. In quanto tale, il punteggio è un'indicazione del rischio genetico totale di un individuo per un

particolare tratto, che può essere utilizzato per la previsione clinica o per lo screening. (Marees et al., 2018)

1.4 Problemi con i metodi classici

Come abbiamo visto, le analisi classiche che vengono fatte nei GWAS mostrano particolare attenzione ai dati a livello biologico, andando ad esaminare con accuratezza l'affidabilità del campione genotipizzato, ma evidenziano anche alcune lacune a livello di analisi statistica dei dati. Le problematiche, nell'analisi classica dei GWAS sono principalmente tre: la soglia di significatività utilizzata, l'utilizzo di variabili categoriali come continue e l'omissione di variabili di confondimento.

1.4.1 Soglia di significatività

Il p-value dei singoli test non garantisce quanti falsi positivi ci siano nei risultati complessivi, per questo motivo nella gran parte dei GWAS si controlla un'altra quantità: il Family Wise Error Rate (FWER). Il FWER è definito come la probabilità di osservare almeno un falso positivo tra i test fatti.

$$FWER = P(V > 0) = 1 - (1 - \alpha)^{m_0} \tag{1.1}$$

con m_0=numero di ipotesi nulle testate, V=numero di falsi positivi.
Di solito non conosciamo m_0, ma conosciamo il numero totale di ipotesi testate (m), che è un limite superiore per m_0 dato che $m_0 \leq$ m.
Il metodo più utilizzato, per controllare il FWER è la correzione di Bonferroni, che consiste - se si vuole controllare il FWER ad un livello α_{FWER} - nello scegliere come soglia $\alpha = \alpha_{\text{FWER}}/\text{m}$. Nel contesto dei GWAS questo viene applicato dividendo 0.05 per il numero di SNP, assumendo - in modo erroneo - indipendenza tra i polimorfismi a singolo nucleotide; oppure, tenendo conto del fatto che alcuni di questi siano tra loro correlati, dividendo per il numero di blocchi LD indipendenti tra loro, circa un milione.
Un potenziale svantaggio di questo metodo, tuttavia, è che se m_0 è grande, la soglia di rifiuto è molto piccola. Ciò significa che i test individuali devono essere molto potenti se vogliamo avere qualche possibilità di rilevare qualcosa,

e questo nella realtà spesso non è possibile (Holmes & Huber, 2019, cap.6). L'utilizzo della correzione di Bonferroni porta quindi a selezionare una soglia di significatività troppo stringente, che implica una perdita di potenza dell'analisi, e rischia di non riuscire a identificare le associazioni presenti nei dati.

1.4.2 Tipologia di covariate

Prima di eseguire l'analisi statistica, una pratica comune nei GWAS è ricodificare il genotipo, trasformando le basi azotate in numeri. Questo passaggio rende la modellazione più agevole, ma comporta una trasformazione della classe delle variabili analizzate: da categoriali a continue.
L'utilizzo di variabili continue - al posto delle categoriali - come covariate in un modello, implica un'assunzione molto forte sui valori che la variabile SNP assume. In particolare, si assegna una grandezza ordinata ai tre livelli (0, 1 e 2), con distanza di un'unità - tra un livello e il successivo - assegnata arbitrariamente. Solitamente il genotipo codificato come 2 corrisponde a quello con entrambi gli alleli mutati, e quello come 1 al caso in cui solo uno dei due alleli risulti mutato. In questo caso ad esempio, si assume che il valore per un dato SNP del soggetto omozigote per l'allele variante, sia esattamente due volte quello del soggetto eterozigote. Questo in generale non è vero, ma per agevolare le analisi statistiche i modelli vengono spesso costruiti utilizzando le covariate continue.

1.4.3 Assenza di confondenti

L'aumento del rischio di una malattia in presenza di un'esposizione, non implica necessariamente una relazione causale tra l'esposizione e la malattia. Una ragione per tali associazioni non causali è la presenza di una terza variabile chiamata confondente o variabile confondente (Kamangar, 2012).
Negli studi di espressione genica su larga scala ci possono essere un certo numero di fattori non misurati o non modellati che possono influenzare l'espressione di un determinato gene. Questi fattori influenti rappresentano fonti di variazione comune nell'espressione genica che possono essere osservate tra più geni.

Le principali fonti di variazione dell'espressione sono dovute a fattori tecnici, ambientali, demografici, genetici o dovute al disegno sperimentale. A volte, anche dopo l'applicazione della normalizzazione, questi confondenti non vengono rimossi. Questo accade perché le tecniche di normalizzazione sono comunemente utilizzate per rilevare e regolare la variazione sistematica dell'espressione, dovuta a fonti tecniche e di laboratorio ben caratterizzate. Spesso nei GWAS non si utilizzano approcci per identificare e tener conto di tutte le fonti di variazione sistematica dell'espressione, inclusa la variazione dovuta a fattori non misurati o non modellati di fonti sia biologiche che tecniche (Leek & Storey, 2007). Per questo motivo il controllo per gruppi etnici nel controllo della qualità dei dati come unica soluzione al confondimento, risulta insufficiente nelle analisi classiche dei GWAS.

Capitolo 2

Metodi statistici per dati ad alta dimensionalità

Il termine "alta dimensionalità" si riferisce al caso in cui il numero di parametri incogniti da stimare p è molto più grande del numero di osservazioni n, cioè p$\gg$n (Gauraha, 2018). La dimensionalità - il numero di variabili nei dati - ha un effetto complesso, che comporta sia un aumento del peso computazionale, sia un rapido aumento della complessità concettuale dei modelli utilizzati, e di conseguenza della loro interpretazione e utilizzo operativo. Il costo computazionale connesso ai dati ad alta dimensionalità ha delle ripercussioni sul modo di lavorare con questi dati, con l'aumentare della dimensionalità, i metodi ad alto costo computazionale diventano meno adoperabili. Questo effetto impedisce l'utilizzo di alcuni strumenti, o quanto meno li rende meno pratici, favorendone altri di minor costo computazionale. Quando si modella una risposta utilizzando un numero elevato di covariate, le stime dei minimi quadrati di un modello lineare hanno spesso una distorsione bassa ma una varianza elevata, rispetto a modelli con un numero inferiore di variabili. Per questo motivo è sconsigliabile utilizzare i minimi quadrati quando p è molto elevato. Una possibile soluzione sono i metodi di regolarizzazione, che consistono nel modificare il metodo di stima introducendo la possibilità di utilizzare uno stimatore distorto, che può dare il vantaggio di avere una varianza minore. L'idea sottostante a questi metodi è di ottenere uno *shrinkage* dei parametri verso la media, in modo che l'intercetta non

venga penalizzata (Azzalini & Scarpa, 2012).
Sono descritti in seguito tre metodi di regolarizzazione applicati in questo elaborato: lasso, elastic net e fused lasso.

2.1 Lasso

Lasso è l'acronimo di *Least Absolute Shrinkage and Selection Operator*, ovvero un metodo per fare simultaneamente selezione delle variabili e *shrinkage* dei coefficienti.
L'obiettivo della regressione lasso è quello di identificare le variabili e i rispettivi coefficienti di regressione, che minimizzino la somma dei residui quadrati, ponendo un vincolo sui coefficienti (Ranstam & Cook, 2018; Tibshirani, 1996). Il vincolo utilizzato dal lasso è la norma l_1, definita come

$$||\beta||_1 = \sum_{j=1}^{p} |\beta_j| \tag{2.1}$$

La norma l_1 dei coefficienti del modello di regressione è forzata a rimanere sotto a un valore fissato t.
La funzione di regressione lasso risolve quindi il problema:

$$\min_{\beta} ||y - X\beta||_2^2 + \lambda||\beta||_1 = \min_{\beta} \sum_{i=1}^{N}(y_i - \beta x_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2.2}$$

A causa della geometria della penalità della norma l_1, il lasso riduce alcuni dei coefficienti di regressione esattamente a zero (Gauraha, 2018).
Per spiegare questo concetto si riporta di seguito l'esempio del caso p=2 come in Tibshirani, 1996.

Il criterio dei minimi quadrati equivale alla funzione quadratica

$$(\beta - \hat{\beta}^0)^T X^T X(\beta - \hat{\beta}^0) \tag{2.3}$$

I contorni ellittici di questa funzione, centrati sulla stima OLS sono riportati in Figura 2.1. Il quadrato ruotato e il cerchio sono le zone di vincolo rispettivamente del metodo lasso e del metodo ridge. La soluzione per il lasso è il

primo punto in cui i contorni toccano il quadrato, e quando questo accade in uno degli angoli del quadrato, corrisponde a un coefficiente posto a zero. Come si può notare invece nel caso del ridge, i contorni non possono toccare nessun angolo, e quindi i coefficienti saranno raramente posti esattamente uguali a zero. Per questo motivo il lasso può anche essere visto come un'alternativa ai metodi di selezione variabili, mentre il ridge riduce i coefficienti di regressione, in modo che le variabili con un contributo minore al risultato abbiano i loro coefficienti vicini allo zero, ma mantiene nel modello tutte le variabili (Pereira et al., 2016).

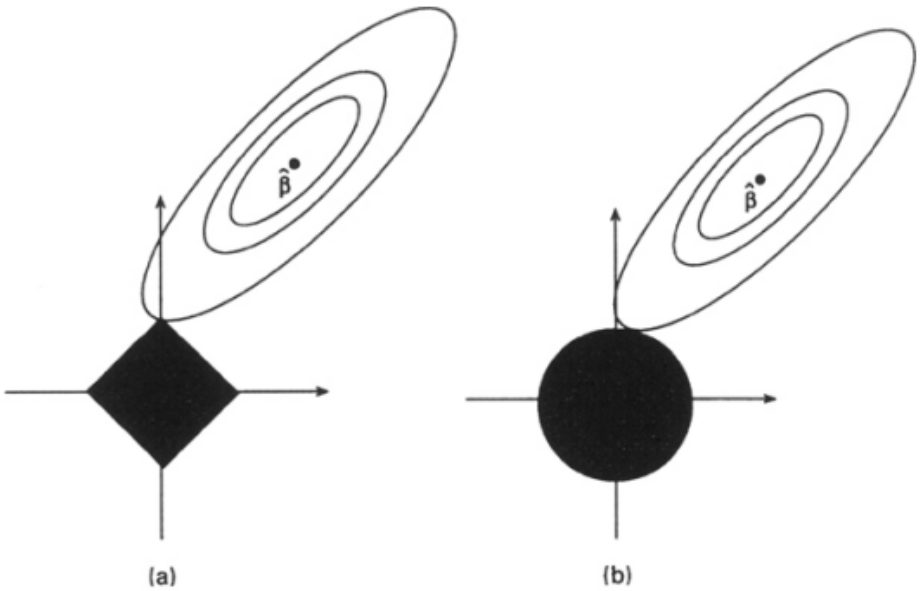

Figura 2.1: Vincoli di lasso e ridge nel caso bidimensionale. Immagine tratta da Tibshirani, 1996.

Quando confrontiamo le stime dei coefficienti ottenute dalle regressioni ridge e lasso, osserviamo che se gli input sono ortogonali, i coefficienti di regressione ridge sono ottenuti dalla moltiplicazione dei coefficienti dei minimi quadrati per una costante tra 0 e 1, mentre il lasso fa una traslazione - uniforme rispetto ai minimi quadrati - della stima dei coefficienti verso lo 0. In Figura 2.2 è riportato il caso semplice in cui le colonne della matrice X sono ortonormali; si noti che la regressione stepwise tronca a zero i coefficienti stimati piccoli (Azzalini & Scarpa, 2012).

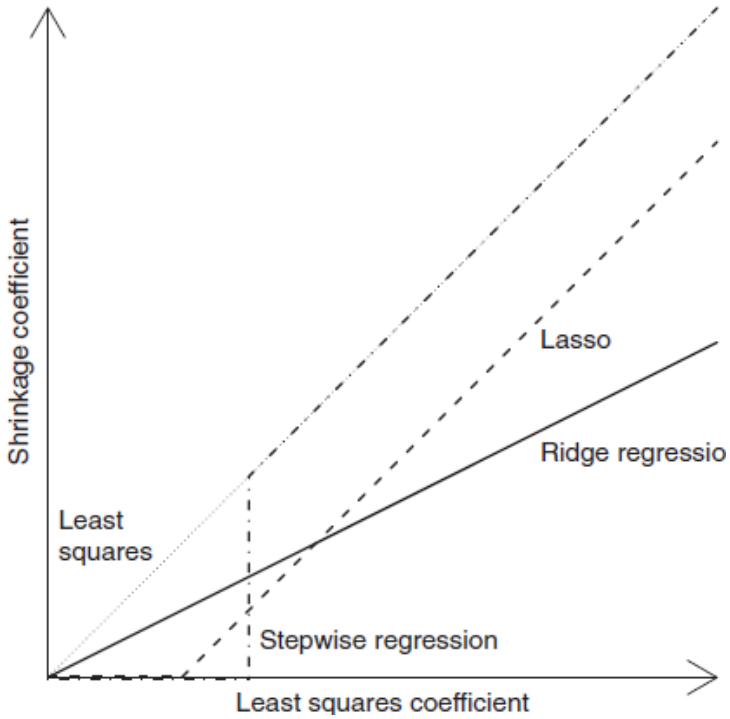

Figura 2.2: Coefficiente trasformato rispetto al coefficiente dei minimi quadrati per regressione ridge, lasso e stepwise per il caso ortonormale. Immagine tratta da Azzalini e Scarpa, 2012

La scelta di λ viene spesso effettuata utilizzando un approccio automatizzato di convalida incrociata *k-fold*. In questo approccio i campioni vengono suddivisi in k gruppi di uguale dimensione, ad ogni step della procedura, k-1 gruppi vengono usati per allenare il modello, e il gruppo rimanente per testarlo. La procedura viene eseguita k volte cambiando sempre il gruppo "lasciato fuori" e usando gli altri sempre per l'apprendimento del modello. Un risultato complessivo viene prodotto combinando i k risultati di convalida separati per un intervallo di valori di λ, e scegliendo il λ ottimale, che viene quindi utilizzato per determinare il modello finale (Ranstam & Cook, 2018).

2.2 Elastic Net

Nonostante il successo in molte applicazioni, il metodo lasso ha qualche inconveniente nei problemi di selezione delle variabili in cui vi sono variabili altamente correlate. In generale, il lasso potrebbe funzionare meglio in una situazione in cui alcuni predittori hanno coefficienti elevati e i restanti predittori hanno coefficienti molto piccoli, mentre il ridge è più performante quando il risultato è una funzione di molti predittori, tutti con coefficienti di dimensioni approssimativamente uguali (Pereira et al., 2016).
Motivati dall'analisi dei dati di microarray, in cui solitamente si hanno molte migliaia di predittori, spesso correlati tra loro, Zou e Hastie, 2005b hanno proposto l'uso di una penalità che è una via intermedia tra le due tecniche di regolarizzazione lasso e ridge: la regressione elastic net.
L'elastic net combina le proprietà della regressione di ridge e lasso usando come penalizzazione una somma pesata della norma l_1 e del quadrato della norma l_2 del vettore β dei coefficienti (De Mol et al., 2009)

$$(1-\alpha)||\beta||_1 + \alpha||\beta||_2^2 \tag{2.4}$$

Il primo termine impone la sparsità della soluzione, mentre il secondo induce un effetto di raggruppamento, in cui i predittori fortemente correlati tendono ad essere dentro o fuori dal modello insieme. Quando α=1, l'elastic net si riduce a una regressione ridge, mentre con α=0 si ottiene la regressione lasso, con una penalizzazione convessa ma non strettamente convessa. Questi argomenti si possono vedere nell'esempio del caso con p=2 riportato in Figura 2.3
In modo simile al lasso, l'elastic net esegue contemporaneamente la selezione automatica delle variabili e lo *shrinkage* continuo, ma a differenza del lasso può selezionare gruppi di variabili correlate.
La funzione di costo che la regressione elastic net si prefigge di minimizzare è

$$\min_{\beta} \frac{1}{2N}\sum_{i=1}^{N}(y_i - x_i\beta)^2 + \lambda(\frac{1-\alpha}{2}\sum_{j=1}^{p}\beta_j^2 + \alpha\sum_{j=1}^{p}|\beta_j|) \tag{2.5}$$

Per la scelta dei parametri λ ed α anche in questo caso si utilizzano metodi adattivi, come ad esempio la convalida incrociata. Avendo due parametri di

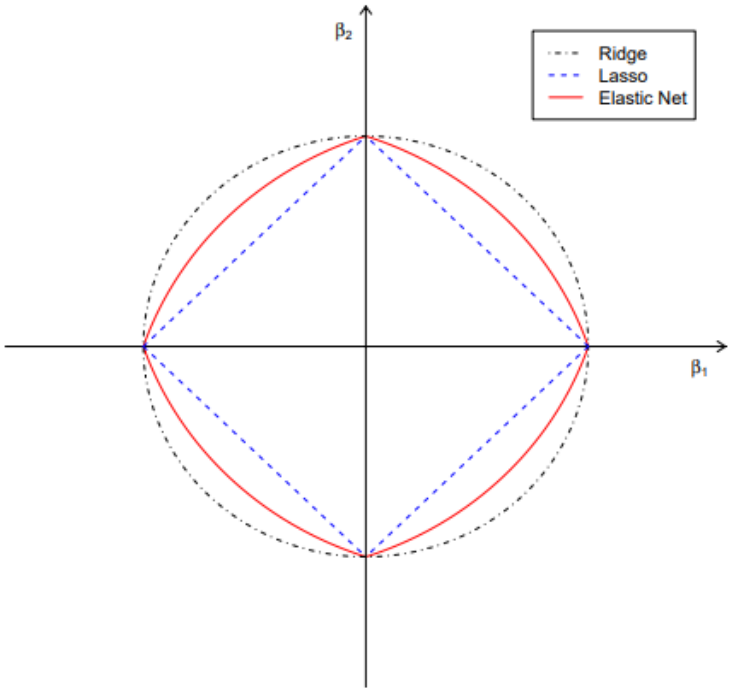

Figura 2.3: Vincoli di ridge, lasso ed elastic net nel caso bidimensionale. Immagine tratta da Zou e Hastie, 2005a

tuning, tipicamente si sceglie prima una griglia relativamente piccola di valori per α. Poi, per ogni α, si produce l'intero percorso di soluzioni per l'elastic net. L'altro parametro di tuning è scelto tramite *k-fold* (di solito k=10) cross-validation, l'α scelto è quello che da il minore errore cross-validato.

2.3 Fused Lasso

Sono state sviluppate molte varianti della regressione lasso di Tibshirani, 1996, per tenere conto delle informazioni strutturali aggiuntive sulle covariate. Una tra queste varianti è l'elastic net di Zou e Hastie, 2005b vista nel paragrafo precedente, progettata per selezionare simultaneamente variabili fortemente correlate. La variante del fused lasso di Tibshirani et al., 2005 invece, assume che le variabili possano essere ordinate secondo qualche criterio, e penalizza una combinazione convessa della norma l_1 delle differenze di

coefficienti adiacenti e della norma l_1 del vettore dei coefficienti stesso.

$$\sum_{j=1}^{p} |\beta_j| \quad e \quad \sum_{j=2}^{p} |\beta_j - \beta_{j-1}| \tag{2.6}$$

Il primo vincolo incoraggia la sparsità dei coefficienti, il secondo incoraggia la sparsità nelle loro differenze (Tibshirani et al., 2005). Questi vincoli vengono mostrati nell'esempio con p=2 in Figura 2.4, in cui la penalizzazione sui coefficienti è rappresentata dal quadrato ruotato, e la penalizzazione sulla differenza di coefficienti adiacenti è rappresentata dal rettangolo nero. La soluzione del fused lasso si ottiene, quando i contorni ellittici della funzione di perdita dei minimi quadrati, toccano sia il quadrato che il rettangolo.

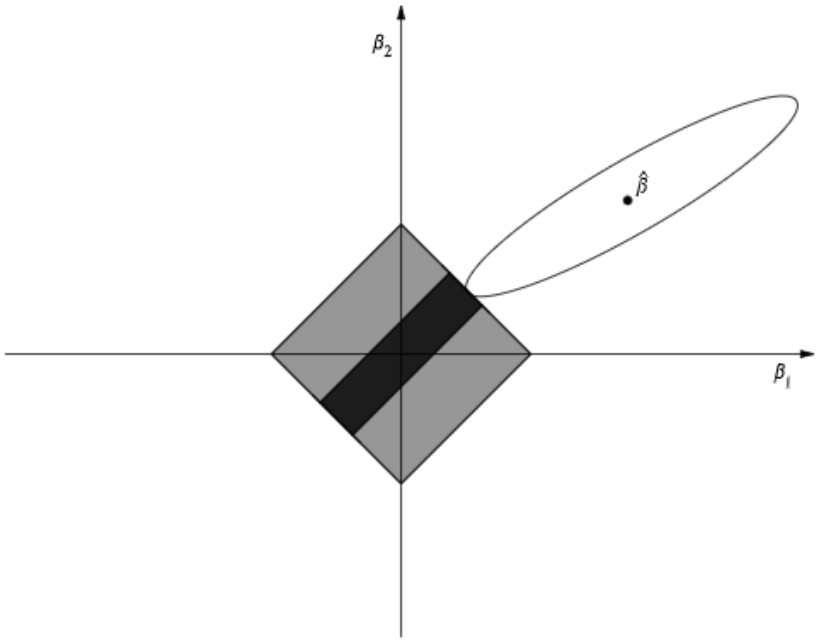

Figura 2.4: Vincoli di fused lasso nel caso bidimensionale. Immagine tratta da Tibshirani et al., 2005

In modo simile all'elastic net, il fused lasso seleziona insieme variabili correlate, ma a differenza dell'elastic net rende i coefficienti localmente uguali tra loro (Lee et al., 2014).
Da Tibshirani et al., 2005 viene riportato di seguito un esempio di questi concetti su un campione simulato, i risultati sono riportati in Figura 2.5.

Consideriamo un esempio con p=100 predittori e N=20 campioni. I dati sono stati generati da un modello $y_i = \sum_j x_{ij}\beta_j + \varepsilon_i$, dove le x_{ij} sono Normali Standard, $\varepsilon_i \sim N(0,\sigma^2)$ con σ=0.05, e β ha due aree diverse da zero.
Con il termine "fusion" si intende una regressione in cui si applica una penalizzazione solo sulle differenze dei coefficienti adiacenti.
La Figura 2.5(a) mostra la soluzione per il lasso, usando s_1=4.2 e s_2=∞. La Figura 2.5(b) mostra la stima "fusion" usando s_1=∞ e s_2=5.2. Infine la Figura 2.5(c) mostra il fused lasso, usando s_1=$\sum_j |\beta_j|$=56.5 e s_2=$\sum_j |\beta_j - \beta_{j-1}|$=13.
I vincoli s_1 e s_2 sono stati scelti in ogni modello minimizzando l'errore di previsione (Tibshirani et al., 2005).

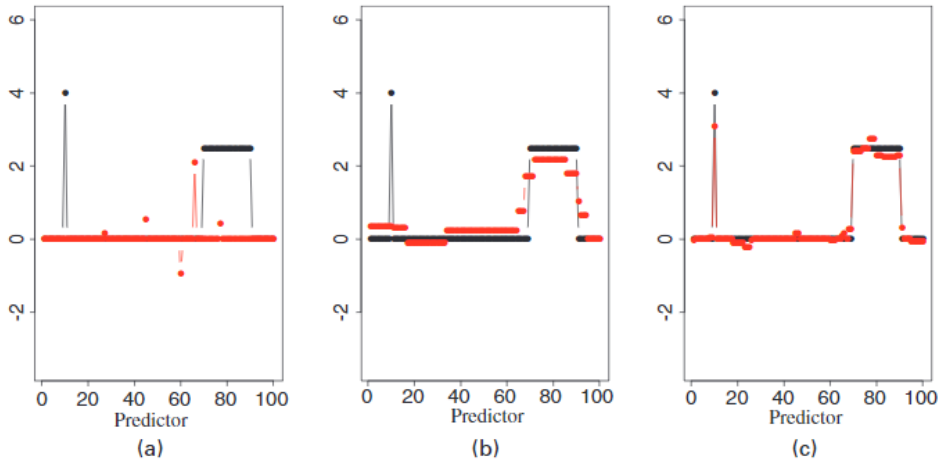

Figura 2.5: Esempio simulato con p=100 predittori. Punti neri coefficienti reali, punti rossi coefficienti stimati da ciascun metodo: (a) lasso, s_1=4.2; (b) fusion, s_2=5.2; (c) fused lasso, s_1=56.5, s_2=13. Immagine tratta da Tibshirani et al., 2005

Il lasso ha una scarsa performance, non riesce ad identificare nessuna delle due aree; il "fusion" cattura il plateau ma non riesce ad isolare il picco; il fused lasso fa un buon lavoro nel complesso, individuando correttamente entrambe le zone di coefficienti non zero.

In conclusione si riporta l'equazione di stima dei coefficienti col metodo fused

lasso:

$$\hat{\beta} = \underset{\beta}{\text{argmax}}\{l(\beta) - \lambda_1 \sum_{j=1}^{p} |\beta_j| - \lambda_2 \sum_{j=1}^{p-1} |\beta_{j+1} - \beta_j|\} \tag{2.7}$$

con $l(\beta)$ funzione di log-verosimiglianza. I parametri di tuning λ_1 e λ_2 vengono scelti in modo da minimizzare l'errore di previsione, calcolato utilizzando la divisione in *train* e *test* del dataset oppure tramite convalida incrociata, come visto per i coefficienti di penalizzazione dell'elastic net.

Capitolo 3

Un GWAS sulla dipendenza da nicotina

3.1 Letteratura GWAS sulla dipendenza da nicotina

Negli ultimi decenni sono stati svolti numerosi GWAS per indagare una possibile predisposizione genetica negli individui affetti da dipendenza da nicotina. La maggior parte di questi studi dispongono degli stessi metodi per svolgere l'analisi statistica. Quello che si fa solitamente è adattare un modello lineare, o logistico, in base alla tipologia del fenotipo preso in considerazione, studiando l'associazione tra dipendenza da nicotina ed i singoli SNP, con modelli di regressione univariata.

La dipendenza da nicotina è definita in diversi modi, in alcuni studi si considera il numero di sigarette al giorno consumate (Thorgeirsson et al., 2010; Thorgeirsson et al., 2008) e il numero massimo di sigarette fumate in 24 ore (Hällfors et al., 2019), in altri se il soggetto fuma o non fuma sigarette (Pandey et al., 2017), in Bergen et al., 1999 si considera un fenotipo quantitativo misurato come pacchetti di sigarette fumati all'anno, e uno dicotomico con soggetti che non hanno mai fumato contro chi invece l'ha fatto. Un altro metodo per definire il fenotipo in questa tipologia di studi è il Fagerström Test for Nicotine Dependence (FTND) (Bierut et al., 2007; Hancock et al.,

2018; Loukola et al., 2013). Il FTND è uno strumento standard per valutare l'intensità della dipendenza fisica dalla nicotina. Il test è stato progettato per fornire una misura ordinale della dipendenza da nicotina correlata al fumo di sigaretta. Contiene sei voci che valutano la quantità di consumo di sigarette, la compulsione all'uso e la dipendenza. («Instrument: Fagerstrom Test for Nicotine Dependence (FTND)», 2014).
In Lind et al., 2009 invece viene usato il DSM-IV, una misura che esamina sette criteri atti a valutare le caratteristiche cliniche della dipendenza. Il DSM-IV consente una diagnosi categorica della presenza di dipendenza da nicotina in base al fatto che i fumatori soddisfino almeno tre dei sette criteri (Rose & Dierker, 2010).

Per tenere conto della dipendenza tra i vari SNP, si aggiusta l'analisi imponendo una soglia per il p-value come visto nel paragrafo 1.3.2. Nella maggior parte degli studi si sceglie 10^{-8}, ma a volte si sceglie la soglia dividendo 0.05 per il numero di SNP inclusi nell'analisi (Thorgeirsson et al., 2008).

3.2 Dati analizzati

Centinaia di migliaia di persone vengono genotipizzate attraverso società di test genetici *Direct-To-Consumer*. Tuttavia, questi dati non sono pubblici per motivi di privacy e quindi non possono essere utilizzati nella ricerca. Per questo motivo si è ritenuto opportuno creare una repository di open data per questa tipologia di dati. É così che nasce openSNP, un progetto open source che consente ai partecipanti dei test genetici *Direct-To-Consumer* di pubblicare gratuitamente i propri dati genetici insieme alle informazioni fenotipiche, così da renderli disponibili per la ricerca. Gli utenti di openSNP possono creare un profilo personale e aggiungere nuovi fenotipi che non sono ancora elencati dagli altri utenti. Attualmente gli utenti possono caricare i loro risultati di genotipizzazione dalle aziende 23andMe, deCODEme e FamilyTreeDNA tramite un'interfaccia web al progetto openSNP (Greshake et al., 2014).
Tramite il pacchetto *rsnps* è possibile accedere ai dati direttamente dal software statistico R. Con la funzione *fetch_genotypes()* è possibile estrarre il

genoma del singolo utente, inserendo come opzioni l'url che riporta direttamente al file presente sul sito openSNP per quel particolare utente, e indicando con "-1" che si vogliono estrarre tutte le righe del file. Per estrarre i fenotipi invece si utilizza la funzione *phenotypes_byid()*, utilizzando l'id assegnato a ciascun fenotipo per selezionare quello di interesse, ed indicando con l'opzione "*return_ ='users'*" che il formato di output richiesto è un dataframe con due colonne: l'user id e l'espressione del fenotipo scelto.
La risposta cui si è interessati per questo elaborato è la dipendenza da nicotina (id=20), in particolare i soggetti vengono classificati in due categorie: fumatori e non fumatori. I soggetti sono stati ricodificati in queste due categorie sulla base del fenotipo caricato sul portale dall'utente.
Nello specifico, vengono inclusi nei fumatori:

- tutti gli attuali fumatori di sigarette, con frequenza che va da 2 sigarette al giorno, fino a 3 pacchetti al giorno;
- i fumatori di sigarette elettroniche con nicotina;
- gli utilizzatori di tabacco da immersione (dipping tobacco);
- gli ex fumatori di sigarette con frequenza da 2 sigarette al giorno, fino a 60 sigarette al giorno;
- gli ex fumatori che ora utilizzano la sigaretta elettronica con nicotina;
- gli ex fumatori che ora usano gomme da masticare composte da nicotina.

Nella categoria "non fumatori" vengono inclusi:

- coloro che non fumano e non hanno mai fumato;
- chi ha provato sigarette ma dichiara di non avere dipendenza;
- coloro che fumano saltuariamente sigarette ma non hanno dipendenza;
- chi fuma il sigaro occasionalmente.

La variabile risposta così codificata consta di 114 fumatori e 184 non fumatori. Il secondo fenotipo utilizzato nell'analisi, come confondente, è il sesso (id=60). Anche per questa variabile è stata necessaria una ricodifica dei livelli, poiché alcuni utenti erano persone transgender o non binarie. Per questi soggetti si è tenuto conto del sesso biologico, così da avere in conclusione solo le categorie "Male" e "Female". Sarebbe stato interessante inserire come ulteriore possibile confondente la variabile relativa alla nazionalità, ma questa presentava troppi valori mancanti, e per questo motivo si è deciso di non includerla nell'analisi.
Non per tutti gli utenti presenti nel pacchetto *rsnps* è possibile estrarre il genoma, e non tutti hanno fornito indicazioni riguardo alla dipendenza da nicotina. Il numero di soggetti cui risulta estraibile direttamente da R il genoma e che hanno dichiarato se fumano o meno sono 455. Tra questi, vi sono alcuni il cui sequenziamento deriva dalla società Ancestry, alcuni da FamilyTreeDNA e altri da 23andMe. I formati dei file di output della genotipizzazione variano col variare della società, per questo motivo non è possibile un'unica analisi per tutti i dati. Si sceglie quindi di selezionare solo i dati provenienti dalla società 23andMe poiché più numerosi. Gli utenti così selezionati risultano 327, in seguito ad ulteriori operazioni sui dati, emerge che 16 di questi soggetti presentano file criptati, 2 mostrano genotipi errati in diversi loci, 1 ha una sola base sequenziata per ogni SNP. Dopo aver eliminato i soggetti con genotipizzazione problematica, il numero finale di utenti inclusi nell'analisi è 308.
Tutti i dati sequenziati e caricati sul portale, indipendentemente dalla società che li ha prodotti, sono stati estratti tramite la tecnologia microarray.

3.2.1 Tecnologia dei Microarray

I microarray sono un gruppo di tecnologie progettate per identificare l'espressione genica in un campione, esponendo quantità di materiale biologico su un vetrino di plastica o vetro dove sono state precedentemente depositate sequenze di DNA note o proteine che fungono da rivelatori. Ogni punto su un microarray contiene più filamenti identici di DNA, ed ogni punto rappresenta un gene.

Le sequenze ancorate sono immobilizzate in punti specifici in modo ordinato e fisso, sono comunemente chiamate sonde (probes) del microarray.

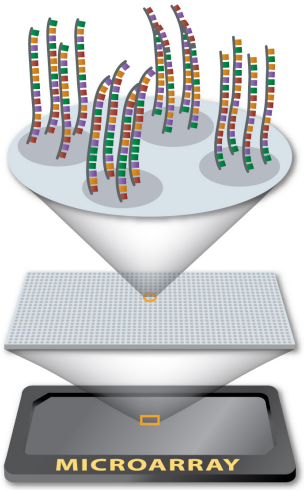

Figura 3.1: Sonde sul microarray. Immagine tratta da «DNA Microarray», n.d.

Le sequenze di DNA del campione biologico, generalmente note come bersagli (target), vengono trasformate in una forma stabile di DNA, come il DNA complementare (cDNA), e nel processo vengono etichettate con un colorante fluorescente. La lunghezza del bersaglio varia da brevi oligonucleotidi a grandi frammenti cromosomici. Le molecole bersaglio etichettate sono esposte a tutte le sonde sul vetrino, ci si aspetta che durante questa esposizione, le molecole bersaglio abbiano la possibilità di ibridarsi con le loro sequenze complementari sul vetrino. Durante l'ibridazione, il target si legherà alle sonde sull'array per sequenza in modo complementare, e il campione in eccesso, non ibridato, sarà rimosso tramite una procedura di lavaggio. A questo punto, ogni sonda del microarray dovrebbe essere legata ad una quantità di target etichettato proporzionale al livello di espressione del gene rappresentato da quella sonda. La quantità di emissione fluorescente su ciascuna sonda sarà utilizzata per generare un'intensità di segnale, che sarà successivamente elaborata da strumenti bioinformatici e fornirà informazioni sul livello di espressione di tutti i geni corrispondenti (Ginsberg, 2010; Guindalini & Pellegrino, 2016). Quest'ultimo processo si chiama "imaging", viene utilizzato uno scanner laser per le sonde con etichetta fluorescente, o un "imager" al

fosforo per sonde etichettate con radioattività.

I microarray possono essere classificati in base alla loro applicazione, quella più comune è l'analisi dell'espressione genica, ma ce ne sono molti altri.
Gli array SNP sono un tipo di array di DNA che si basa sugli stessi principi degli array di espressione, ma la differenza principale è che le sonde devono essere in grado di distinguere tra diverse variazioni alleliche. Cioè, per ogni polimorfismo da rilevare, l'array contiene le diverse possibili variazioni nel sito specifico. Questi tipi di array possono essere utilizzati per la genotipizzazione, ovvero per rilevare gli SNP all'interno delle popolazioni (Simó et al., 2014, cap. 1).

3.3 Prepocessing

Come prima operazione sui dati, viene fatto un controllo della qualità prendendo come riferimento gli step classici dei Genome Wide Association Studies. Tuttavia, avendo come focus dell'elaborato i metodi statistici, alcuni step, di interesse principalmente biologico, non sono stati eseguiti.
Primo step è la ricodifica dei dati, alcuni SNP su cui non è stato rilevato il genotipo sono stati codificati con "- -". Questi vengono trasformati in "NA", in modo che si possa calcolare la percentuale di dati mancanti per ogni SNP. Nel dataset è presente un numero elevato di valori mancanti per alcuni SNP, questo perché la matrice su cui si svolgono le analisi deriva dall'unione di genomi sequenziati di 308 individui in periodi differenti. A seconda del periodo, i genomi sono stati sequenziati con versioni del "Human Reference Assembly" differenti. Questi ultimi sono genomi di riferimento, utilizzati come esempio rappresentativo dell'insieme di geni in un singolo organismo idealizzato di una specie. In particolare, per i dati analizzati, sono state usate le versioni GRCh37 e NCBI36. Quando si aggiorna il genoma di riferimento umano, si ottiene un sequenziamento del DNA più accurato e alcune delle zone (loci) che prima non erano state sequenziate, vengono rilevate. Per questo si verifica che gli SNP rilevati nei soggetti con cui è stata usata la versione 37, presentino "NA" nei soggetti in cui è stata utilizzata la versione 36. Questo motiva la grande riduzione di dimensioni del dataset in seguito alla rimozione

degli SNP che presentavano più dell'80% di valori mancanti.
Dopodiché, per eliminare gli SNP su cui era espresso un solo genotipo, o quasi, viene utilizzata la funzione *nearZeroVar()* del pacchetto *caret*.
Infine si rimuovono anche alcune colonne di introni codificati con lettere diverse dalle basi azotate.
In seguito a questi primi passaggi, il dataset passa da 8 959 838 colonne a 857 909.
Viene poi eseguito un controllo per dati mancanti anche sui campioni, vengono rimossi dal dataset i soggetti con più del 90% di NA.
Il dataset finale comprende 298 righe, un soggetto per ogni riga, e 857 909 colonne, di cui le prime tre rispettivamente "user_id", "sesso", "fenotipo" e nelle successive uno SNP per ogni colonna.
Si procede poi con la ricodifica delle basi azotate in numeri, tramite la funzione *raw.data()* del pacchetto *snpReady*. Questa funzione, ponendo l'opzione "outfile=‘012’" restituisce in output una matrice codificata con "2" per il genotipo con entrambi gli alleli di riferimento (AA), "1" per quello con una sola mutazione (Aa) e "0" per quello con entrambi gli alleli mutati (aa) (Granato & Fritsche-Neto, 2017). La selezione dell'allele di riferimento per ogni SNP viene fatto in ordine alfabetico, per cui se in un dato SNP gli alleli espressi sono "G" e "T", la base "G" sarà considerata di riferimento e "T" sarà considerata mutazione (Granato, 2018).

3.4 Analisi esplorativa

Data la complessità dei dati, si sceglie di visualizzarli tramite due metodi di riduzione della dimensionalità: la Principal Component Analysis (PCA) e il Multidimensional Scaling (MDS).
La PCA è una tecnica per ridurre la dimensionalità di grandi dataset, partendo dai dati originali. Tramite questa riduzione, la PCA è in grado di fornire un'interpretazione dei dati, minimizzando la perdita di informazione. Per preservare la variabilità dei dati, e dunque l'informazione contenuta in essi, la PCA trova nuove variabili che sono combinazioni lineari di quelle originali, incorrelate tra loro e che massimizzano progressivamente la varianza (Jolliffe & Cadima, 2016).

Il Multidimensional Scaling è un'altra metodologia utilizzata nel caso di dataset ad alte dimensionalità, che però a differenza della PCA, prende in input la matrice delle distanze tra i punti del dataset originale, cioè le similarità o dissimilarità osservate tra tutte le coppie di osservazioni. La matrice delle distanze è una matrice simmetrica, con i valori sulla diagonale principale tutti pari a 0, e valori fuori dalla diagonale tutti non negativi.
L'obiettivo del MDS è cercare di riprodurre quanto possibile le distanze tra i punti originari ma in uno spazio di dimensionalità ridotta, solitamente con 2 o 3 dimensioni (Pegoraro, 2019).
Per svolgere l'analisi esplorativa con entrambi i metodi, si è utilizzato il pacchetto *SNPRelate*, che prende in input per le sue funzioni una classe di oggetti specifica denominata *SNPGDSFileClass*, creabile direttamente con il pacchetto stesso tramite la funzione *snpgdsCreateGeno()*. Per svolgere l'Analisi delle Componenti Principali si è applicata la funzione *snpgdsPCA()*, che prende in input l'oggetto della classe *SNPGDSFileClass* e chiede di indicare il nome con cui vengono identificati gli SNP e il numero di core (CPU) utilizzati.
Per il MDS invece si calcola prima la distanza tra le osservazioni, tramite la funzione *snpgdsIBS()*, che calcola la frazione di "Identity By State" per ogni coppia di campioni. L'"IBS" è una misura di distanza tra i genomi di due individui, si dice che due individui sono "Identical By State" se sullo stesso SNP presentano gli stessi alleli. Nello specifico, questa funzione crea una matrice i cui elementi vanno da 0 a 1, e si calcolano come la media su tutto il genoma di $1 - |g_{1,i} - g_{2,i}|/2$, dove $g_{1,i}$ e $g_{2,i}$ sono rispettivamente il genotipo del primo e del secondo individuo sullo SNP i (Zheng, 2020).
Infine, il Multidimensional Scaling si ottiene utilizzando la funzione *cmdscale()*, che prende in input la matrice delle IBS, e restituisce un insieme di punti tale che le distanze tra i punti siano approssimativamente uguali alle dissimilarità.

3.5 Modelli adattati

La prima analisi statistica segue le linee indicate dalla letteratura classica, dunque per modellare il fenotipo dicotomico (fumatore/non fumatore), viene adattato un modello logistico per ogni SNP, tenendo le covariate come variabili continue. Se P è la probabilità di essere un fumatore:

$$log(\frac{P}{1-P}) = \beta_0 + \beta_1 SNP \tag{3.1}$$

I modelli successivi si pongono come obiettivo quello di partire dall'analisi classica dei GWAS, e introducendo delle modifiche, andare a colmare alcune delle lacune statistiche menzionate nel paragrafo 1.4.

Nella seconda analisi si modella il fenotipo con uno SNP alla volta, utilizzato come variabile categoriale, inserendo come confondenti - sulla base di ciò che è emerso dalle analisi esplorative - il sesso e le prime due dimensioni del Multidimensional Scaling.

$$log(\frac{P}{1-P}) = \beta_0 + \beta_1 SNP + \beta_2 sex + \beta_3 MDS1 + \beta_4 MDS2 \tag{3.2}$$

La procedura per ottenere un singolo p-value per ogni SNP, è quella di adattare il modello completo in Equazione 3.2 e confrontarlo con il modello con le sole variabili confondenti in Equazione 3.3, facendo un test del rapporto di verosimiglianza tramite la funzione *anova()*, con l'opzione "test='LRT'".

$$log(\frac{P}{1-P}) = \beta_0 + \beta_2 sex + \beta_3 MDS1 + \beta_4 MDS2 \tag{3.3}$$

Se indichiamo con $l(x,\beta)$ la funzione di log-verosimiglianza, e vogliamo testare H_0: $\beta_1=0$ contro H_1: $\beta_1 \neq 0$ allora il test del rapporto di verosimiglianza risulterà:

$$LR = 2(l(x,\hat{\beta}) - l(x,\hat{\beta}_0)) \tag{3.4}$$

La statistica test risulta asintoticamente distribuita come un χ^2 con 1 grado di libertà.

Si prosegue poi con le analisi introducendo i modelli per i dati ad alta dimensionalità discussi nel capitolo 2.
Per poter adattare un modello multivariato studiando gli SNP congiuntamente, è necessaria una selezione delle variabili per ridurre il peso computazionale. Si decide quindi di fare un primo filtraggio rimuovendo tutti gli SNP con una percentuale di valori mancanti superiore al 50%, e successivamente di includere nel modello i primi mille SNP che hanno ottenuto, col test del rapporto di verosimiglianza (3.4), i pvalue minori, e che quindi possiamo considerare come quelli con più probabilità di avere un'associazione con il fenotipo.
In questo nuovo dataset ridotto sono ancora presenti degli NA, che vanno trattati perchè le funzioni che si utilizzeranno accettano in input solo matrici piene. Si sceglie quindi di imputare i valori mancanti tramite la funzione *impute.knn()* del pacchetto bioconductor *impute*.
Per ogni SNP con valori mancanti, questa funzione trova i k vicini più vicini usando una distanza euclidea, tra i soggetti che per quello SNP non hanno valore mancante. Dopo aver trovato i k vicini più vicini, imputiamo gli elementi mancanti facendo la media dei valori (non mancanti) dei suoi vicini. Nel caso in cui tutti i k *nearest neighbor* avessero valori mancanti per quello SNP, si utilizzerà la media su tutti i soggetti per quello SNP come valore da imputare (Hastie, Tibshirani et al., 2021).
Si procede dunque con il modello di regressione logistica con penalizzazione di tipo lasso, che viene adattato con la funzione *glmnet()* del pacchetto omonimo. Il valore di λ ottimale è calcolato tramite *10-fold* cross-validation con la funzione *cv.glmnet()*.
La funzione risolve il problema

$$\min_{\beta_0,\beta} - \left[\frac{1}{N} \sum_{i=1}^{N} y_i \cdot (\beta_0 + x_i^T \beta) - log(1 + e^{(\beta_0 + x_i^T \beta)}) \right] + \lambda ||\beta||_1 \qquad (3.5)$$

(Hastie, Qian et al., 2021).
Dove N è il numero di osservazioni pari a 298, β è il vettore dei coefficienti degli SNP di lunghezza p=1 000, e il λ usato è pari a 0.0311.

Il modello successivo è il modello logistico con l'approccio dell'elastic net, che viene adattato anche in questo caso con la funzione *glmnet()*. Per questo modello i parametri di *tuning* sono stati calcolati usando il pacchetto *glmnetUtils*, il quale contiene la funzione *cva.glmnet()*, che permette di fare cross-validation simultaneamente per i parametri α e λ, ed adattare una regolarizzazione di elastic net per ogni combinazione di parametri.
La funzione risolve il problema

$$\min_{\beta_0,\beta} - \left[\frac{1}{N}\sum_{i=1}^{N} y_i \cdot (\beta_0 + x_i^T\beta) - log(1 + e^{(\beta_0 + x_i^T\beta)})\right] + \lambda\left[\frac{(1-\alpha)}{2}||\beta||_2^2 + \alpha||\beta||_1\right] \tag{3.6}$$

(Hastie, Qian et al., 2021).
Con α scelto pari a 0.512 e λ ottimale pari a 0.04.

L'ultimo modello è la regressione logistica con il metodo fused lasso, l'analisi viene svolta su un solo cromosoma alla volta, in modo da poter utilizzare la posizione sul cromosoma come ordinamento delle variabili. Per motivi computazionali, si sceglie di svolgere l'analisi solo sul cromosoma 2, che risulta quello con più SNP annotati tra i primi mille SNP più significativi risultati dal test del rapporto di verosimiglianza (3.4), il processo sarebbe analogo per tutti gli altri cromosomi.
La dimensionalità delle variabili va ridotta per problemi computazionali anche in questo caso, quindi si selezionano solo gli SNP sul cromosoma 2 che hanno percentuale di valori mancanti inferiore al 50%, e che hanno ottenuto un p-value inferiore a 0.05 nel test del rapporto di verosimiglianza (3.4). I valori mancanti vengono imputati con la funzione *impute.knn()*. Il dataset finale su cui viene adattato il fused lasso è composto da 2 648 variabili.
Per adattare il modello e trovare i parametri di *tuning* si usa il pacchetto *penalized*. Per trovare i parametri λ_1 e λ_2 da inserire nel modello si procede in due step: profilazione e ottimizzazione della log-verosimiglianza cross-validata.
La profilazione si trova con le funzioni *profL1()* e *profL2()*, che servono per esaminare l'effetto di λ_1 e λ_2 sulla log-verosimiglianza cross-validata, rispettivamente facendo variare λ_1 tenendo fisso λ_2, e facendo variare λ_2 tenendo fisso λ_1. L'output della profilazione, viene dato in input alle funzioni *optL1()* e *optL2()* per ottenere i λ_1 e λ_2 ottimali (Goeman et al., 2018).

Per adattare il modello di regressione con penalizzazione di tipo fused lasso si utilizza la funzione *penalized()*, fornendo come opzioni i due parametri ottimali calcolati precedentemente, e "*fusedl=TRUE*". I parametri stimati soddisfano

$$\hat{\beta} = \underset{\beta}{\text{argmax}}\{l(\beta) - \lambda_1 \sum_{j=1}^{p} |\beta_j| - \lambda_2 \sum_{j=1}^{p-1} |\beta_{j+1} - \beta_j|\} \tag{3.7}$$

(Tibshirani et al., 2005).
Con $l(\beta)$ funzione di log-verosimiglianza della binomiale, p pari a 2 648, λ_1 pari a 2.5207 e λ_2 pari a 23.3807.

Capitolo 4

Risultati

4.1 Analisi esplorativa

Per vedere qual è la distribuzione degli SNP sul genoma, e trovare se ci sono zone con concentrazione maggiore di SNP, si riporta il grafico di densità degli SNP in Figura 4.1 Quello che emerge è che sul cromosoma 6 c'è una

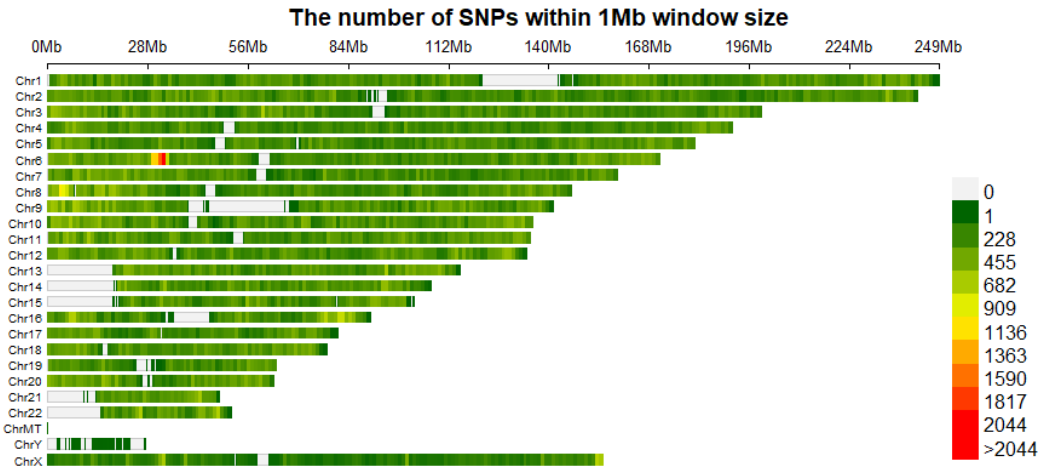

Figura 4.1: Grafico della densità degli SNP

zona con un'alta concentrazione di SNP, mentre sugli altri cromosomi non emergono altre zone simili.

Il grafico con le prime due componenti principali calcolate sui dati analizzati è riportato in Figura 4.2.

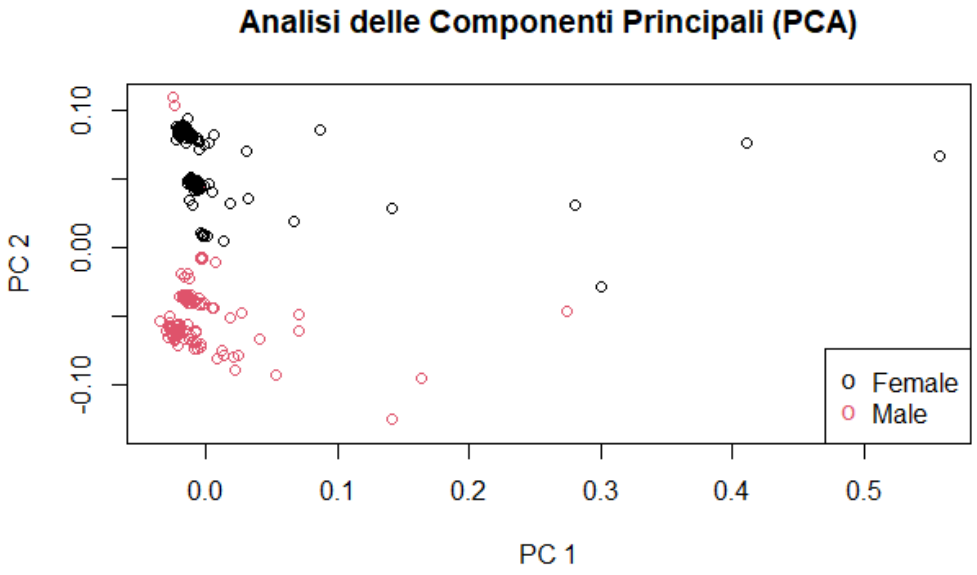

Figura 4.2: Prime due componenti principali, colorate per sesso

I soggetti sono stati rappresentati sul grafico come punti colorati in base al sesso biologico, si nota che la seconda componente principale separa quasi completamente i maschi dalle femmine. Inoltre emergono dei gruppi di punti che potrebbero essere determinati da uno o più possibili confondenti. Purtroppo non essendo presenti altre variabili nel dataset che possano spiegare tale raggruppamento, sarà difficile tenerne conto in fase inferenziale.

Il risultato del Multidimensional Scaling è riportato in Figura 4.3.
Si nota che anche in questo caso emergono dei gruppi distinti, e allo stesso modo della PCA, una delle due componenti del Multidimensional Scaling separa quasi perfettamente i maschi dalle femmine. Per tener conto della struttura della popolazione, non disponendo di altre variabili, si inseriscono nel modello le prime due dimensioni del MDS.

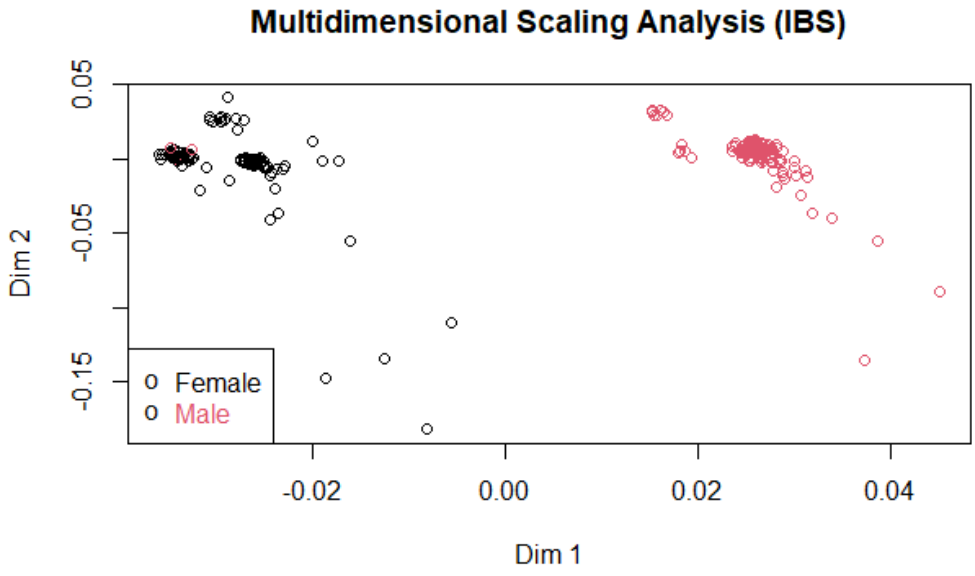

Figura 4.3: Multidimensional Scaling sulla distanza IBS

Per capire in che modo la variabile sesso si distribuisca rispetto alla variabile risposta, si riporta la Tabella 4.1 in cui è calcolata la distribuzione percentuale dei due sessi rispetto alla dipendenza da nicotina.
Quello che emerge è che tra le femmine la percentuale di fumatori è del 45%

Tabella 4.1: Distribuzione dei due sessi rispetto alla dipendenza da nicotina

	Fumatori	Non fumatori
Femmine	0.4545	0.5455
Maschi	0.2835	0.7165

circa, mentre tra i maschi è del 28% circa. Quindi si può affermare che tra le femmine ci sia una percentuale di fumatori che è 1.6 volte quella dei maschi. Per questo motivo un modello senza la covariata sesso potrebbe risultare in associazioni spurie tra alcuni SNP e la dipendenza da nicotina.

4.2 Risultati GWAS

4.2.1 Regressione logistica

I p-value risultanti dal modello 3.1 e dal test del rapporto di verosimiglianza 3.4, sono riportati sotto forma di QQ-plot e di Manhattan plot.
Il QQ-plot è una rappresentazione grafica della deviazione dei p-value osservati dall'ipotesi nulla: i p-value osservati per ogni SNP sono ordinati dal più grande al più piccolo e tracciati rispetto ai valori attesi da una distribuzione Uniforme in [0,1]. Se i valori osservati corrispondono ai valori attesi sotto l'ipotesi nulla, tutti i punti si trovano sulla linea bisettrice tra l'asse x e l'asse y. I Manhattan plot rappresentano i p-value dell'intero GWAS su scala genomica, e sono rappresentati sull'asse x in ordine genomico per cromosoma e posizione sul cromosoma, mentre il valore sull'asse y rappresenta il $-log_{10}$ del p-value (Ehret, 2010).

In Figura 4.4 è riportato il QQ-plot con i risultati del modello 3.1, in questo caso i p-value osservati sono compatibili con l'ipotesi nulla e si dispongono principalmente sulla bisettrice con poche deviazioni.
I risultati sotto forma di Manhattan plot sono invece riportati in Figura 4.5. Si nota che nessun p-value arriva alla soglia tipicamente utilizzata di 10^{-8}, ma il p-value più piccolo raggiunto è dell'ordine di 10^{-7}. Inoltre ciò che emerge è che i p-value più piccoli sono stati rilevati tutti sugli SNP posizionati sul cromosoma X, che è il cromosoma che caratterizza i due sessi, e questo va a rimarcare la necessità di inserire la variabile sesso nell'analisi in quanto possibile fonte di confondimento.

Il test del rapporto di verosimiglianza (3.4), testa la significatività dei singoli SNP inseriti all'interno del modello come variabili categoriali, tenendo conto dei possibili confondenti con la variabile sesso e con le prime due dimensioni del Multidimensional Scaling.
Il QQ-plot con i risultati di questo modello è riportato in Figura 4.6, in questo caso i p-value sembrano distribuirsi ancora di più - rispetto a quelli del modello precedente - sulla bisettrice.

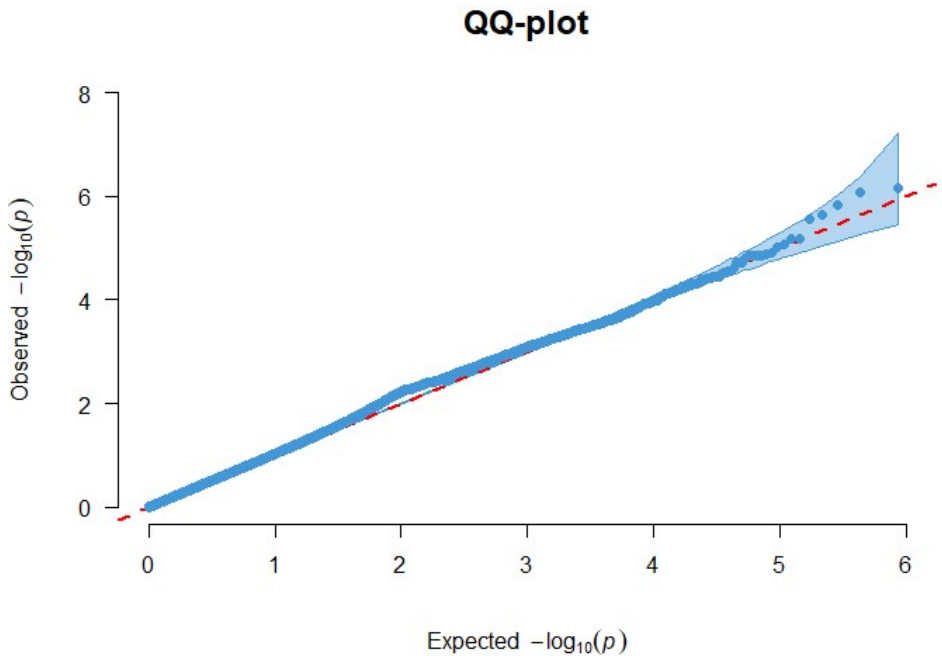

Figura 4.4: $-log_{10}$ dei valori osservati dei p-value, ottenuti con il modello 3.1, vs $-log_{10}$ dei valori attesi dei p-value

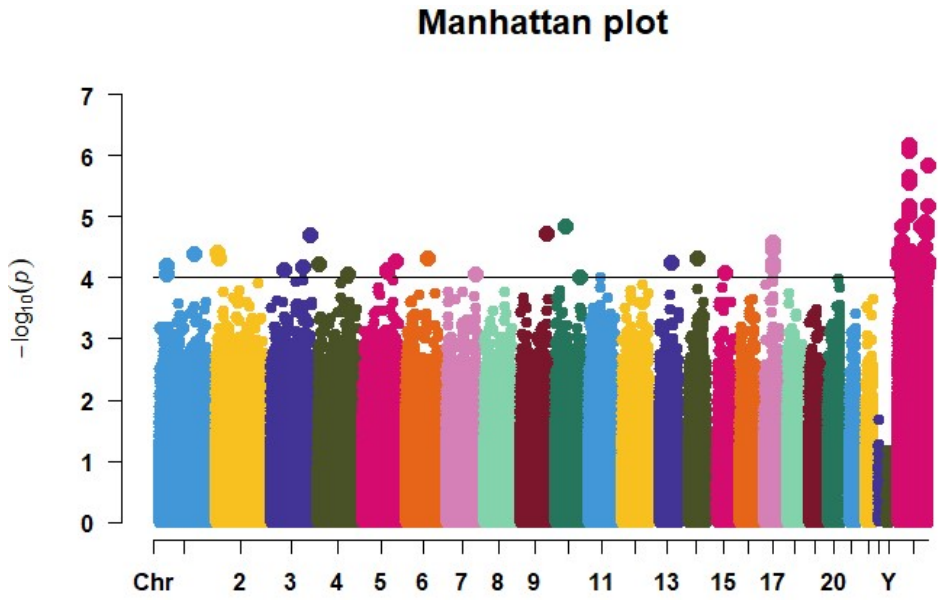

Figura 4.5: Manhattan plot dei p-value ottenuti con il modello 3.1

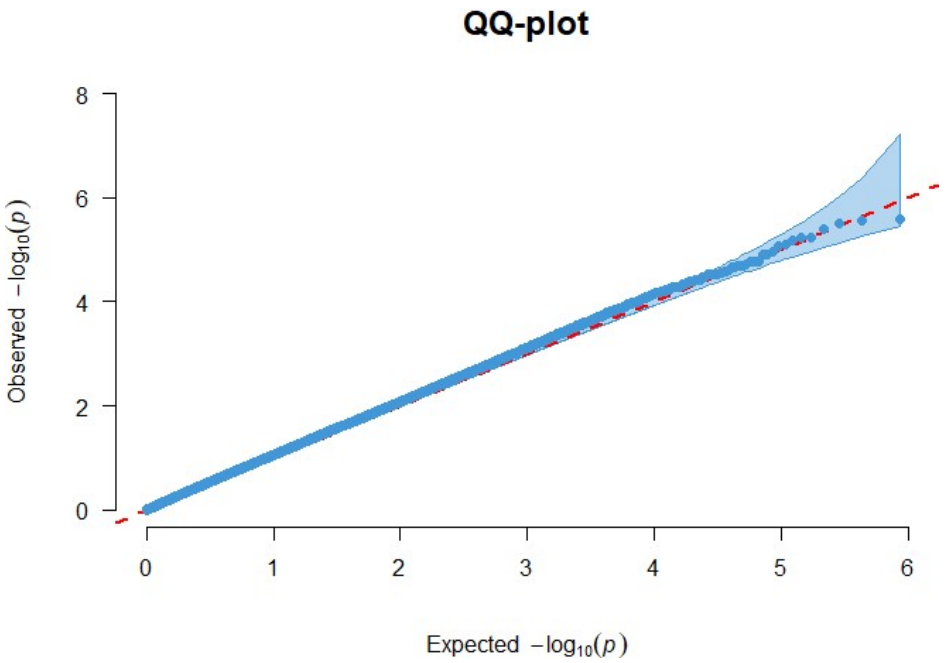

Figura 4.6: $-log_{10}$ dei valori osservati dei p-value, ottenuti con il test 3.4, vs $-log_{10}$ dei valori attesi dei p-value

Il Manhattan plot dei p-value ottenuti col test (3.4) è riportato in Figura 4.7. Anche in questo caso nessuno SNP ha p-value inferiore alla soglia classica, e inoltre il p-value più piccolo in questo caso ha un ordine di grandezza pari a 10^{-6}, quindi più grande rispetto al più piccolo p-value del modello precedente. Un'altra differenza tra i risultati di questi primi due modelli emerge chiaramente dal confronto dei due Manhattan plot: nel secondo caso gli SNP con p-value più piccoli sono distribuiti quasi in modo uniforme su tutti i cromosomi. Questo sembra mostrare ancora una volta che la variabile sesso agisse effettivamente da confondente, polarizzando i p-value più piccoli sul cromosoma X nei risultati del primo modello in cui non si consideravano variabili confondenti.

Se l'analisi di associazione genome-wide si fermasse a questi primi due modelli, il risultato sarebbe che nessuno SNP risulta significativamente associato alla dipendenza da fumo. Ma come abbiamo visto, la correzione di Bonferroni è una metodologia molto conservativa, che rischia di escludere anche le

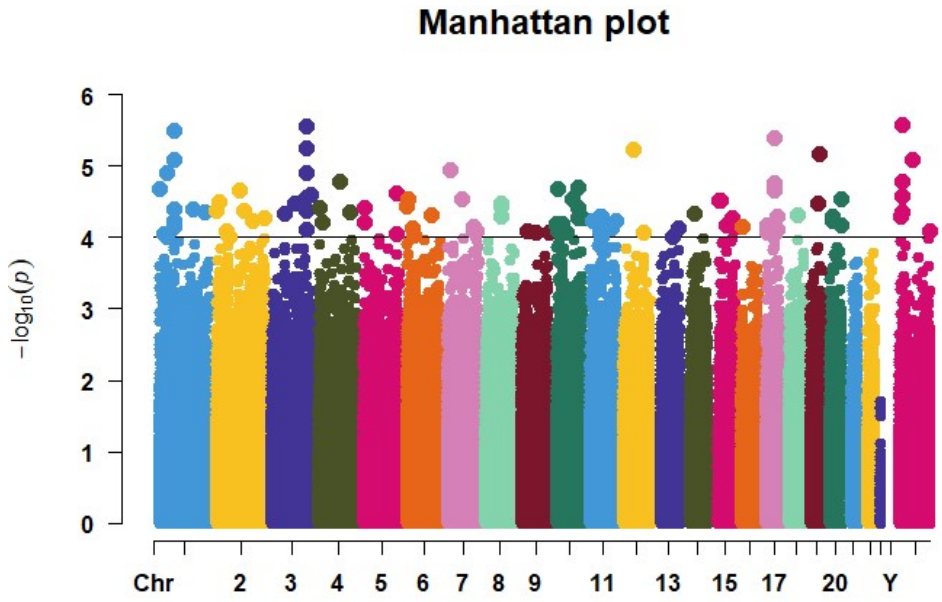

Figura 4.7: Manhattan plot dei p-value ottenuti con il test 3.4

associazioni realmente significative. Per questo motivo, si decide di partire dai risultati del secondo modello - quello più corretto a livello statistico - per condurre le successive analisi con i modelli per dati ad alta dimensionalità.

4.2.2 Regressione penalizzata

Per il modello lasso ed elastic net si selezionano i primi mille SNP con i p-value più piccoli risultanti dal test (3.4), tra gli SNP con percentuale di valori mancanti inferiore al 50%.

La selezione del parametro di *tuning* λ per il modello lasso viene fatta tramite cross-validation, il risultato della devianza residua calcolata con la convalida incrociata al variare di λ è riportato in Figura 4.8.
Si sceglie come valore di λ quello che nell'output della funzione *cv.glmnet()* viene indicato con *lambda.1se*, ovvero il valore massimo di λ tale che l'errore cross-validato sia entro un errore standard del minimo. In questo caso λ risulta pari a 0.0311, con questo valore fissato, il modello lasso seleziona 104 predittori.

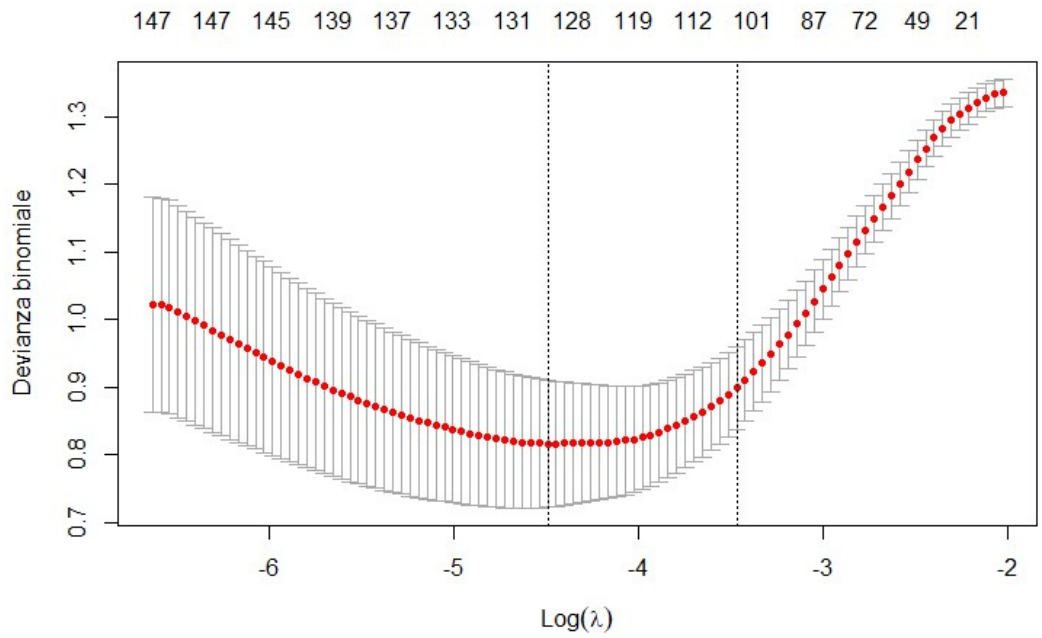

Figura 4.8: Devianza binomiale al variare di λ in scala logaritmica nel modello lasso

Per il modello elastic net viene fatta una convalida incrociata per selezionare α e λ facendoli variare insieme. Il risultato della devianza binomiale al variare dei due parametri di *tuning* è riportato in Figura 4.9. Si sceglie il valore di α pari a 0.512, in modo che l'elastic net risulti una via di mezzo tra la regressione ridge e lasso.
Per α uguale a 0.512 si mostra poi in Figura 4.10 il grafico della devianza binomiale al variare di λ. Il valore di λ viene scelto come per il metodo lasso, in questo caso risulta pari a 0.04. L'elastic net con α e λ scelti come appena descritto, effettua una selezione di 159 predittori, tra cui sono inclusi tutti i 104 estratti anche dal modello lasso.

Tramite SNPnexus, uno strumento di annotazione *web-based* per l'analisi e l'interpretazione di variazioni di sequenziamento sia note che nuove, è possibile mostrare le varianti con un'associazione fenotipica nota, sulla base dei dati di COSMIC e ClinVar, due archivi pubblici di interpretazioni di varianti clinicamente rilevanti (Oscanoa et al., 2020).
Dando in input i 159 predittori selezionati dall'elastic net, emerge che 37 SNP

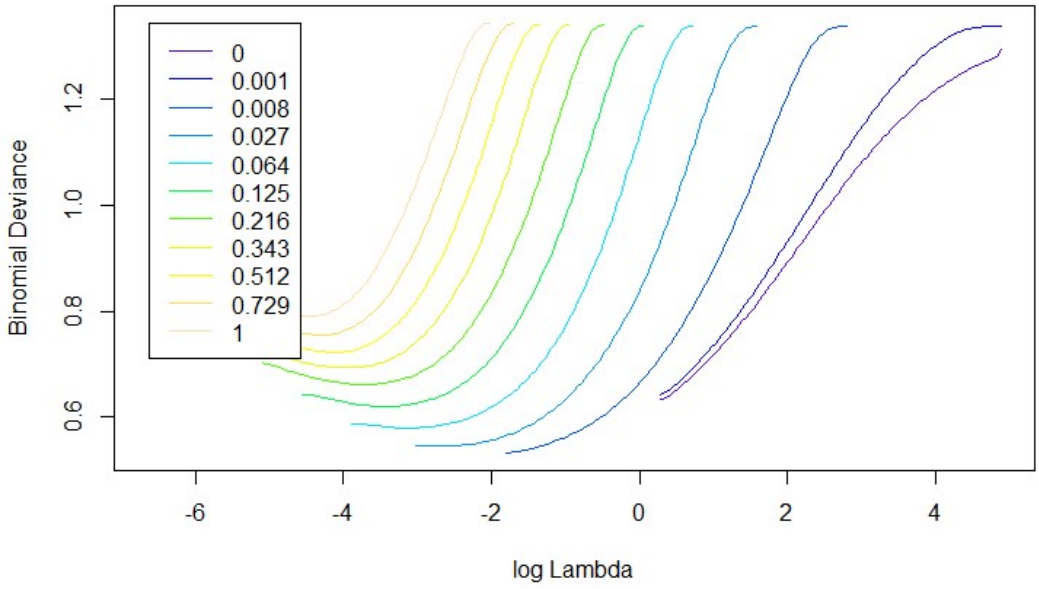

Figura 4.9: Devianza binomiale al variare di α e λ in scala logaritmica nel modello elastic net

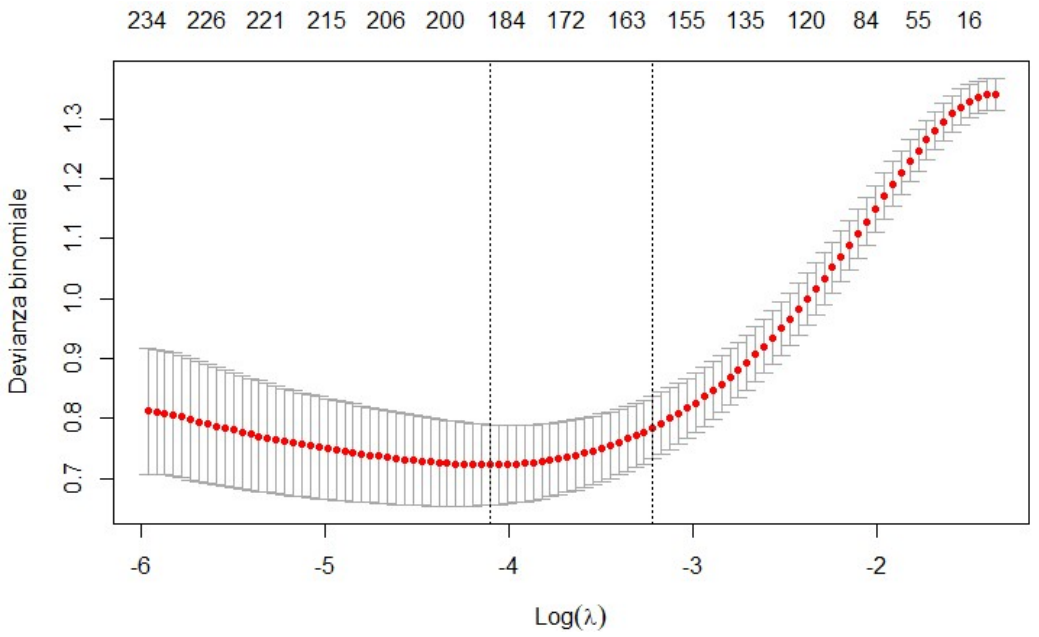

Figura 4.10: Devianza binomiale al variare di λ in scala logaritmica con α=0.512 nel modello elastic net

sono stati trovati associati al fenotipo "*Tobacco use disorder*", 7 sono stati trovati associati al fenotipo "*Smoking/smoking behavior*", e 4 in entrambi i fenotipi. In particolare lo SNP rs8071667 è quello risultato più volte associato all'uso di tabacco o all'essere fumatore, rilevato in totale in 21 studi. Il nome del gene che si sovrappone a questo SNP è SLC6A4. Questo gene codifica per una proteina integrale di membrana che trasporta il neurotrasmettitore serotonina dagli spazi sinaptici ai neuroni presinaptici. La proteina codificata interrompe l'azione della serotonina e la ricicla in una maniera sodio-dipendente. Questa proteina è un bersaglio di stimolanti psicomotori, come anfetamine e cocaina («SLC6A4 [Homo sapiens (human)]», 2021). Si potrebbe approfondire il funzionamento di questa proteina per capire se la nicotina possa avere lo stesso ruolo di anfetamine e cocaina, e se la dipendenza possa essere una conseguenza della variazione di questi meccanismi biologici.

Infine per il modello fused lasso vengono analizzati 2 648 SNP posizionati sul cromosoma 2. Per selezionare λ_1 e λ_2 da inserire nel modello si utilizza anche in questo caso la cross-validation.
Il grafico che mostra la verosimiglianza cross-validata al variare di λ_1 è riportato in Figura 4.11. Si nota che la funzione raggiunge l'unico massimo tra 0 e 5, e poi inizia a calare fino a stabilizzarsi dopo 15. In Figura 4.12 si mostra invece lo stesso grafico ma al variare di λ_2, in questo caso la funzione è strettamente concava, e il massimo si raggiune tra 10 e 100. Si mostra infine in Figura 4.13 il percorso del lasso adattato al variare di λ_2 che mostra quante e quali variabili vengono poste pari a zero per ogni valore del parametro di *tuning*. I valori ottimali, ovvero i valori per cui la log-verosimiglianza cross-validata risulta massima, sono pari a 2.5208 per λ_1 e 23.3807 per λ_2. Il modello fused lasso con i parametri di *shrinkage* così calcolati, seleziona un totale di 218 predittori.
Per avere un'idea spaziale di come i predittori vengano estratti si riporta in Figura 4.14 il grafico con i coefficienti calcolati dal fused lasso per ogni SNP analizzato sul cromosoma 2, ordinati per posizione.
Come ci si aspetta dal fused lasso, gli SNP vicini sul cromosoma vengono selezionati insieme, e il loro coefficiente risulta uguale o simile.

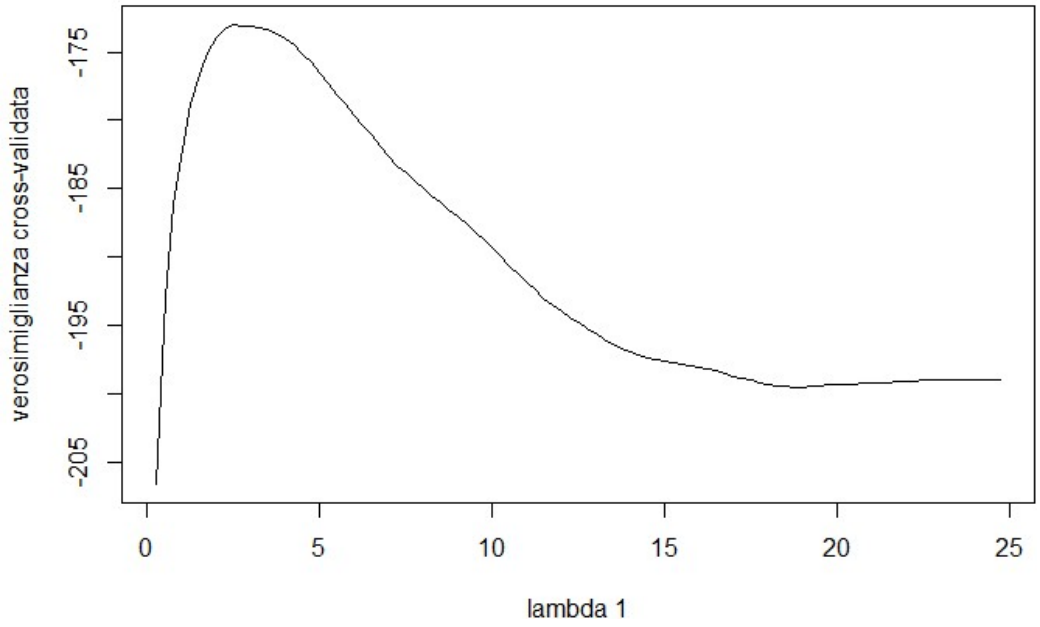

Figura 4.11: Verosimiglianza cross-validata al variare di λ_1

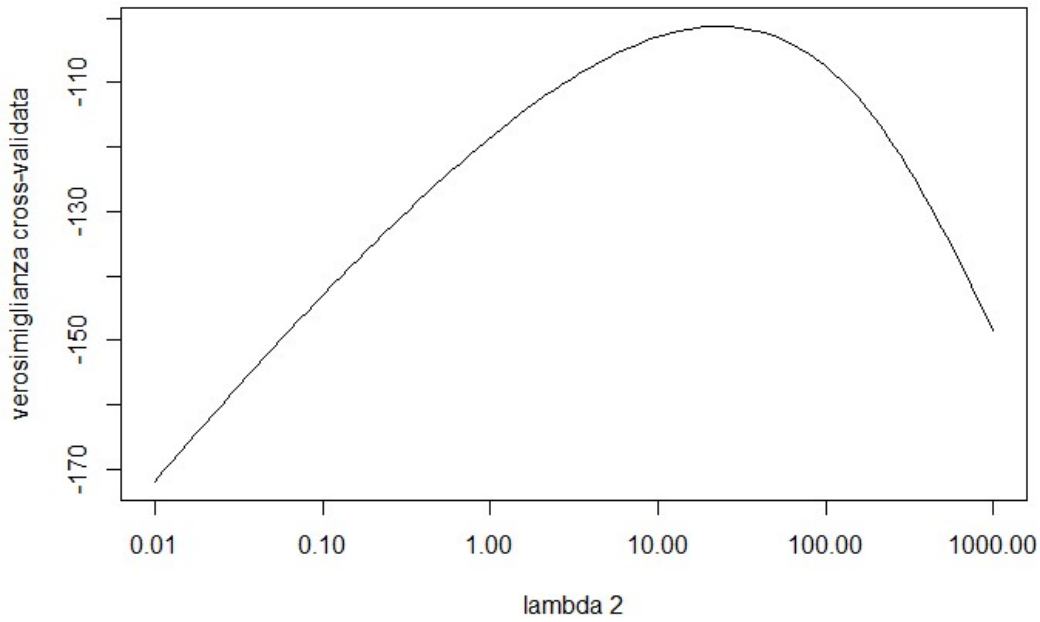

Figura 4.12: Verosimiglianza cross-validata al variare di λ_2

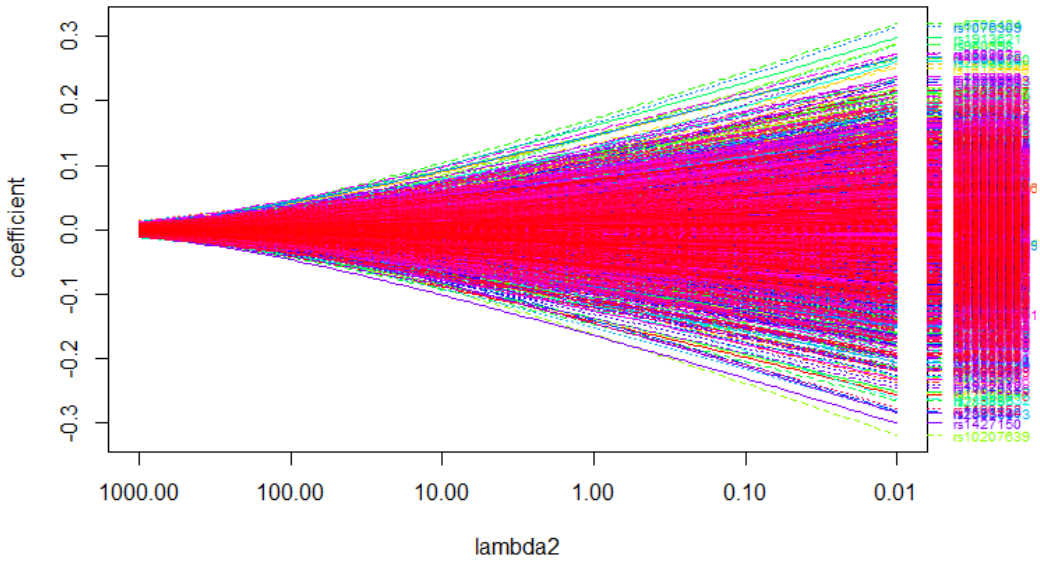

Figura 4.13: Percorso del lasso adattato al variare di λ_2

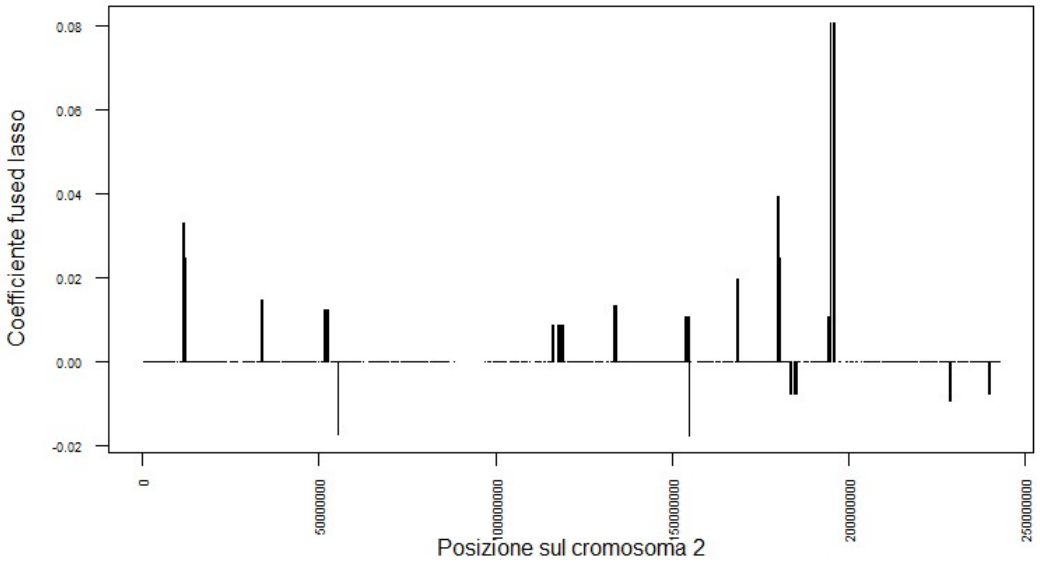

Figura 4.14: Coefficienti degli SNP calcolati dal modello fused lasso ordinati per posizione sul cromosoma 2

Per poter fare un confronto diretto tra i tre metodi per dati ad alta dimensionalità utilizzati, si adattano i modelli lasso ed elastic net allo stesso dataset usato per il fused lasso, ovvero 2 648 SNP sul cromosoma 2.
I parametri di *tuning* vengono scelti con gli stessi criteri usati in precedenza. Si mostrano i risultati dei coefficienti calcolati dal metodo lasso con λ uguale a 0.0585 e dal metodo elastic net con α uguale a 0.512 e λ pari a 0.0411 sugli SNP posizionati sul cromosoma 2, rispettivamente in Figura 4.15 e Figura 4.16.
Si fa notare che il metodo fused lasso ed elastic net estraggono molti più predittori rispetto al lasso, in particolare vengono selezionati 218 SNP dal fused lasso, 207 dall'elastic net e 66 dal lasso. Questa prima evidenza va a confermare il fatto che la regressione lasso non sia in grado di tenere conto della correlazione tra predittori e quindi probabilmente riesca a selezionarne solo alcuni tra quelli importanti. Inoltre si fa notare che i 207 predittori selezionati dall'elastic net si distribuiscono su tutto il cromosoma e hanno coefficienti molto variabili, mentre il fused lasso tende a selezionare solo SNP vicini e posizionati su poche zone del cromosoma, attribuendo ai predittori adiacenti coefficienti uguali o simili.

Nonostante il confronto tra metodi su dati reali sia complesso per la mancanza di "*ground truth*", i risultati del fused lasso appaiono più corretti se si tiene conto della natura della correlazione tra gli SNP, ovvero che SNP vicini sul cromosoma spesso sono coinvolti negli stessi percorsi biologici.
Facendo un'intersezione dei risultati ottenuti con i tre modelli, emergono 14 SNP che vengono selezionati dal lasso, dall'elastic net e dal fused lasso. Fra questi, da SNPnexus emerge che 4 sono stati associati all'utilizzo di tabacco. Nei 218 predittori selezionati in totale dal fused lasso, ne risultano invece 72 associati all'utilizzo di tabacco.

Per questi risultati così diversi graficamente tra i tre metodi applicati sul cromosoma 2, potrebbe risultare interessante svolgere l'intera analisi di associazione Genome Wide anche tenendo conto della disposizione spaziale dei predittori, adattando il modello fused lasso sull'intero genoma.

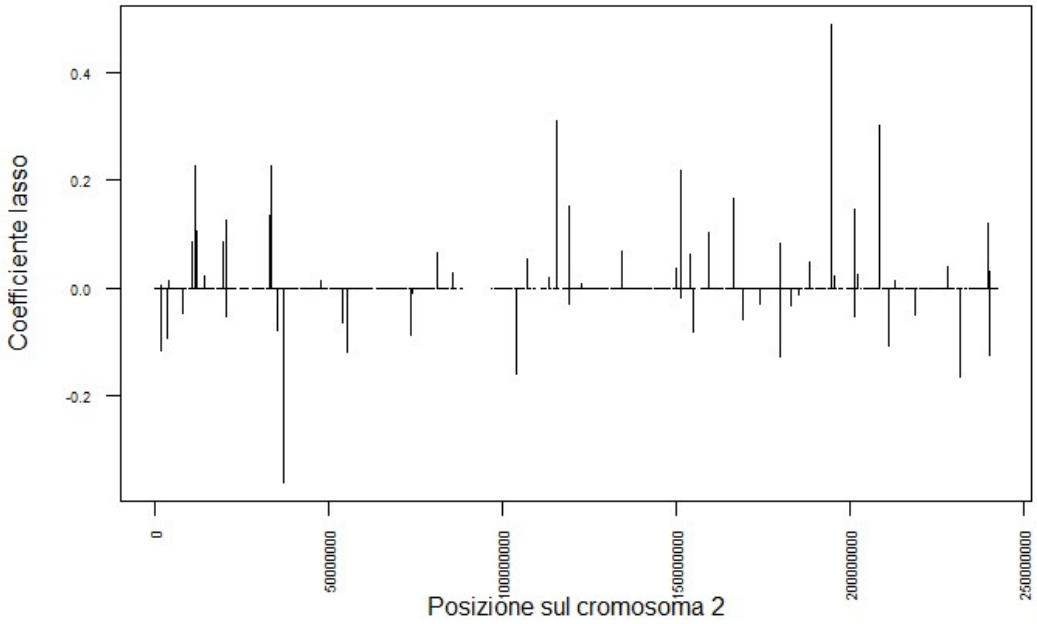

Figura 4.15: Coefficienti degli SNP calcolati dal modello lasso ordinati per posizione sul cromosoma 2

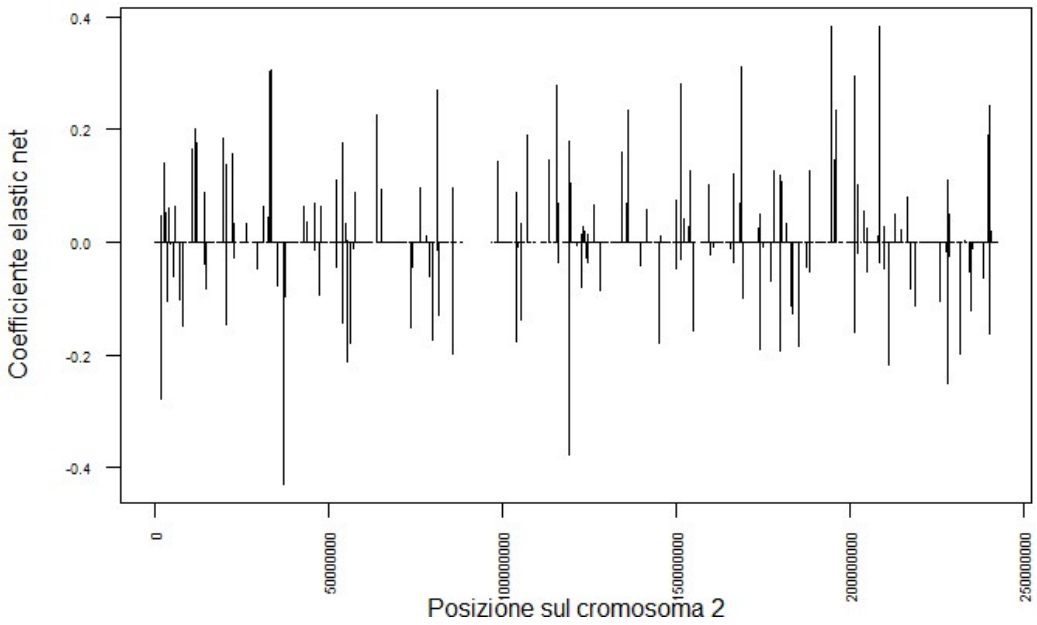

Figura 4.16: Coefficienti degli SNP calcolati dal modello elastic net ordinati per posizione sul cromosoma 2

Conclusioni

Per migliorare i metodi statistici tipicamente utilizzati nei GWAS, che presentano dei problemi a livello teorico, sono stati proposti dei modelli alternativi. Tra questi, sono stati introdotti tre metodi per dati ad alta dimensionalità: lasso, elastic net e fused lasso. I modelli sono stati applicati a dati genomici reali provenienti da utenti dei test genetici *Direct-To-Consumer*.
I risultati ottenuti da questa applicazione, sembrano mostrare che in generale modelli multivariati che applicano regolarizzazione funzionino meglio dei classici modelli univariati con correzione di Bonferroni.
Nei risultati di questi ultimi infatti, nessuno SNP riesce a raggiungere la soglia richiesta dalla correzione di Bonferroni, concludendo l'analisi con nessuna associazione rilevata.
I metodi ad alta dimensionalità invece, selezionano predittori in tutti i modelli adattati, sia sull'intero genoma, che sul solo cromosoma 2. Tra i modelli per dati ad alta dimensionalità, il fused lasso sembra essere quello che funziona meglio, se si tiene conto della natura dei dati. Con questo metodo, sul solo cromosoma 2, vengono selezionati 218 predittori, tra cui 72 che sono risultati precedentemente associati con l'utilizzo di tabacco.
Anche l'elastic net applicata sull'intero genoma sembra andare piuttosto bene, selezionando, tra i primi mille SNP che hanno ottenuto p-value minore sull'intero genoma con il modello univariato, 159 predittori, di cui 37 già risultati associati all'utilizzo di tabacco o all'essere fumatore.

Uno dei possibili sviluppi di questo elaborato è motivato dai risultati ottenuti sul cromosoma 2, che sembrano indicare come vantaggiosa l'inclusione dell'informazione sulla disposizione spaziale dei predittori nel modello, e dun-

que l'utilizzo del modello fused lasso sui dati dell'intero genoma.
Inoltre si fa notare che uno dei principali limiti dei GWAS, è che i risultati di questi studi danno informazioni solamente riguardo ad una associazione tra il fenotipo e la variazione, ma non permettono di inferire causalità.
Inferire relazioni causali tra fenotipi è una sfida importante e ha importanti implicazioni per la comprensione dell'eziologia dei processi patologici. Per questo in questi ultimi anni sono stati sviluppati metodi statistici per l'inferenza causale che sfruttano i principi della randomizzazione mendeliana (MR) utilizzando dati risultanti dai GWAS (Hemani et al., 2018).
Un possibile proseguimento di questo elaborato potrebbe essere proprio l'applicazione di questi metodi ai risultati ottenuti in questa analisi, per comprendere più a fondo l'eziologia della dipendenza da nicotina.

Bibliografia

Azzalini, A. & Scarpa, B. (2012). *Data analysis and data mining : an Introduction*. Oxford University Press.

Bergen, A. W., Korczak, J. F., Weissbecker, K. A., Goldstein, A. M. & Branch, G. E. (1999). A Genome-Wide Search for Loci Contributing to Smoking and Alcoholism. *Genetic Epidemiology*, *17*(1), 55–60.

Bierut, L. J., Madden, P. A., Breslau, N., Johnson, E. O., Hatsukami, D., Pomerleau, O. F., Swan, G. E., Rutter, J., Bertelsen, S., Fox, L., Fugman, D., Goate, A. M., Hinrichs, A. L., Konvicka, K., Martin, N. G., Montgomery, G. W., Saccone, N. L., Saccone, S. F., Wang, J. C., ... Ballinger, D. G. (2007). Novel genes identified in a high-density genome wide association study for nicotine dependence. *Human Molecular Genetics*, *16*(1), 24–35.

Calabrese, B. (2019). Linkage Disequilibrium. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, *1-3*, 763–765.

De Mol, C., De Vito, E. & Rosasco, L. (2009). Elastic-net regularization in learning theory. *Journal of Complexity*, *25*(2), 201–230.

DNA Microarray. (n.d.). https://learn.genetics.utah.edu/content/labs/microarray/

Ehret, G. B. (2010). Genome-Wide Association Studies: Contribution of Genomics to Understanding Blood Pressure and Essential Hypertension. *Current hypertension reports*, *12*(1), 17.

Gauraha, N. (2018). Introduction to the LASSO. A Convex Optimization Approach for High-dimensional Problems. *Resonance*, *23*(4), 439–464.

Ginsberg, S. D. (2010). Gene Microarrays. *Encyclopedia of Movement Disorders*, 538–540.

Goeman, J., Meijer, R. & Chaturvedi, N. (2018). L1 and L2 Penalized Regression Models. *Biometrical Journal*, *52*(1), 70–84.

Granato, I. (2018). *snpReady and BGGE: R packages to prepare genomic datasets and perform genome-enabled predictions*. University of Sao Paulo.

Granato, I. & Fritsche-Neto, R. (2017). *snpReady: a helper tool to run genomic analysis*. https://cran.r-project.org/web/packages/snpReady/vignettes/snpReady-vignette.html

Greshake, B., Bayer, P. E., Rausch, H. & Reda, J. (2014). openSNP–A Crowdsourced Web Resource for Personal Genomics. *PLOS ONE*, *9*(3), e89204.

Guindalini, C. & Pellegrino, R. (2016). Gene Expression Studies Using Microarrays. *Rodent Model as Tools in Ethical Biomedical Research* (pp. 203–216). Springer.

Hällfors, J., Palviainen, T., Surakka, I., Gupta, R., Buchwald, J., Raevuori, A., Ripatti, S., Korhonen, T., Jousilahti, P., Madden, P. A., Kaprio, J. & Loukola, A. (2019). Genome-wide association study in Finnish twins highlights the connection between nicotine addiction and neurotrophin signaling pathway. *Addiction Biology*, *24*(3), 549–561.

Hancock, D. B., Guo, Y., Reginsson, G. W., Gaddis, N. C., Lutz, S. M., Sherva, R., Loukola, A., Minica, C. C., Markunas, C. A., Han, Y., Young, K. A., Gudbjartsson, D. F., Gu, F., McNeil, D. W., Qaiser, B., Glasheen, C., Olson, S., Landi, M. T., Madden, P. A. F., ... Johnson, E. O. (2018). Genome-wide association study across European and African American ancestries identifies a SNP in DNMT3B contributing to nicotine dependence. *Molecular Psychiatry*, *23*(9), 1911–1919.

Hastie, T., Qian, J. & Tay, K. (2021). *An Introduction to glmnet*. https://cran.us.r-project.org

Hastie, T., Tibshirani, R., Narasimhan, B. & Chu, G. (2021). *impute: Imputation for microarray data*. http://www-stat.stanford.edu/~hastie/Papers/missing.pdf

Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J., Langdon, R., Tan, V. Y.,

Yarmolinsky, J., Shihab, H. A., Timpson, N. J., Evans, D. M., Relton, C., Martin, R. M., Smith, G. D., Gaunt, T. R. & Haycock, P. C. (2018). The MR-base platform supports systematic causal inference across the human phenome. *eLife*, *7*, e34408.

Holmes, S. & Huber, W. (2019). Testing. *Modern Statistics for Modern Biology*. Cambridge University Press.

Instrument: Fagerstrom Test for Nicotine Dependence (FTND). (2014). Recuperato ottobre 4, 2021, da https://cde.drugabuse.gov/instrument/d7c0b0f5-b865-e4de-e040-bb89ad43202b

Jolliffe, I. T. & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202.

Kamangar, F. (2012). Confounding Variables in Epidemiologic Studies: Basics and Beyond. *Archives of Iranian Medicine*, *15*(8), 508–516.

Lee, S. H., Yu, D., Bachman, A. H., Lim, J. & Ardekani, B. A. (2014). Application of fused lasso logistic regression to the study of corpus callosum thickness in early Alzheimer's disease. *Journal of neuroscience methods*, *221*, 78.

Leek, J. T. & Storey, J. D. (2007). Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLOS Genetics*, *3*(9), e161.

Lind, P. A., Macgregor, S., Vink, J. M., Pergadia, M. L., Hansell, N. K., Moor, M. H. M. D., Smit, A. B., Hottenga, J.-J., Richter, M. M., Heath, A. C., Martin, N. G., Willemsen, G., Geus, E. C. D., Vogelzangs, N., Penninx, B. W., Whitfield, J. B., Montgomery, G. W., Boomsma, D. I. & Madden, P. A. F. (2009). A Genomewide Association Study of Nicotine and Alcohol Dependence in Australian and Dutch Populations. *Twin Research and Human Genetics*, *13*, 10–29.

Loukola, A., Wedenoja, J., Keskitalo-Vuokko, K., Broms, U., Korhonen, T., Ripatti, S., Sarin, A.-P., Pitkäniemi, J., He, L., Häppölä, A., Heikkilä, K., Chou, Y.-L., Pergadia, M. L., Heath, A. C., Montgomery, G. W., Martin, N. G., Madden, P. A. F. & Kaprio, J. (2013). Genome-wide

association study on detailed profiles of smoking behavior and nicotine dependence in a twin sample. *Molecular Psychiatry*, *19*(5), 615–624.

Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C. & Derks, E. M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*, *27*(2), e1608.

Oscanoa, J., Sivapalan, L., Gadaleta, E., Ullah, A. Z. D., Lemoine, N. R. & Chelala, C. (2020). SNPnexus: a web server for functional annotation of human genome sequence variation (2020 update). *Nucleic Acids Research*, *48*(W1), W185–W192.

Pandey, N., Pal, S., Sharma, L. K., Guleria, R., Mohan, A. & Srivastava, T. (2017). SNP rs16969968 as a strong predictor of nicotine dependence and lung cancer risk in a North Indian population. *Asian Pacific Journal of Cancer Prevention*, *18*(11), 3073–3079.

Pegoraro, E. (2019). *Statistica per Data Science con R*. http://www.r-project.it/_book/multidimensional-scaling-mds.html

Pereira, J. M., Basto, M. & da Silva, A. F. (2016). The Logistic Lasso and Ridge Regression in Predicting Corporate Failure. *Procedia Economics and Finance*, *39*, 634–641.

Prediger, E. (2019). Genotyping: Terms to know. https://eu.idtdna.com/pages/education/decoded/article/genotyping-terms-to-know

Ranstam, J. & Cook, J. A. (2018). LASSO regression. *British Journal of Surgery*, *105*(10), 1348–1348.

Rose, J. S. & Dierker, L. C. (2010). DSM-IV nicotine dependence symptom characteristics for recent-onset smokers. *Nicotine & Tobacco Research*, *12*(3), 278.

Simó, C., Cifuentes, A. & García-Cañas, V. (2014). *Fundamentals of Advanced Omics Technologies: From Genes to Metabolites*. Elsevier.

SLC6A4 [Homo sapiens (human)]. (2021). https://www.ncbi.nlm.nih.gov/gene#phenotypes

Tanksley, P. T., Motz, R. T., Kail, R. M., Barnes, J. C. & Liu, H. (2019). The Genome-Wide Study of Human Social Behavior and Its Application in Sociology. *Frontiers in Sociology*, *4*, 53.

Thorgeirsson, T. E., Gudbjartsson, D. F., Surakka, I., Vink, J. M., Amin, N., Geller, F., Sulem, P., Rafnar, T., Esko, T., Walter, S., Gieger, C., Rawal, R., Mangino, M., Prokopenko, I., Mägi, R., Keskitalo, K., Gudjonsdottir, I. H., Gretarsdottir, S., Stefansson, H., ... Stefansson, K. (2010). Sequence variants at CHRNB3–CHRNA6 and CYP2A6 affect smoking behavior. *Nature Genetics*, *42*(5), 448–453.

Thorgeirsson, T. E., Geller, F., Sulem, P., Rafnar, T., Wiste, A., Magnusson, K. P., Manolescu, A., Thorleifsson, G., Stefansson, H., Ingason, A., Stacey, S. N., Bergthorsson, J. T., Thorlacius, S., Gudmundsson, J., Jonsson, T., Jakobsdottir, M., Saemundsdottir, J., Olafsdottir, O., Gudmundsson, L. J., ... Stefansson, K. (2008). A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature*, *452*(7187), 638–642.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *67*(1), 91–108.

Uitterlinden, A. G. (2016). An Introduction to Genome-Wide Association Studies: GWAS for Dummies. *Seminars in Reproductive Medicine*, *34*(4), 196–204.

What is genotyping? | IDT. (2021). Recuperato settembre 10, 2021, da https://eu.idtdna.com/pages/applications/genotyping

Zheng, X. (2020). *snpgdsIBS: Identity-By-State (IBS) proportion*. https://rdrr.io/bioc/SNPRelate/man/snpgdsIBS.html

Zou, H. & Hastie, T. (2005a). *Regularization and Variable Selection via the Elastic Net*. https://web.stanford.edu/~hastie/TALKS/enet_talk.pdf

Zou, H. & Hastie, T. (2005b). Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, *67*(2), 301–320.

Ringraziamenti

Vorrei dedicare qualche riga per ringraziare tutti coloro che, con il loro sostegno, mi hanno permesso di arrivare fino a qui.

Ringrazio innanzitutto il mio relatore, il professor Risso, che mi ha seguito con infinita disponibilità nella stesura dell'elaborato, e mi ha trasmesso la passione per la bioinformatica.
Ringrazio mia nonna Teresa, a cui dedico la tesi, mia sostenitrice numero uno, che mi è stata accanto per tutti questi due lunghi e difficili anni, gioendo con me ad ogni mio successo, e ricordandomi ogni giorno quanto fosse fiera di me. Senza il suo supporto non sarei arrivata dove sono ora.
Ringrazio Margherita, per avermi insegnato che il valore di una persona non si misura in voti, e per aver condiviso con me gioie e dolori di questi cinque anni di università.
Un ringraziamento speciale va ai miei genitori, che da sempre mi sostengono nella realizzazione dei miei progetti. Non finirò mai di ringraziarvi per avermi permesso di arrivare fin qui ed aver sempre creduto in me e nelle mie capacità.
Ringrazio tutti gli amici, i parenti e tutte le persone che hanno fatto parte della mia vita in questi ultimi due anni, ognuno di voi ha contribuito al raggiungimento di questo risultato.
Infine ringrazio me stessa, per non essermi mai arresa, per aver scoperto una forza che non credevo di avere ed aver affrontato le più grandi difficoltà, ricavandone sempre qualche insegnamento.