UNIVERSITA' DEGLI STUDI DI PADOVA

FACOLTA' DI INGEGNERIA

Corso di laurea magistrale in Bioignegneria

Studio di associazione Genome Wide:

Preprocessing e Selezione SNPs

LAUREANDA : Giulia Bruscagin

RELATORE: prof Barbara di Camillo

CORRELATORE: dott. Francesco Sambo

Anno Accademico 2010/2011

RINGRAZIAMENTI

Colgo l'occasione per ringraziare i dottori Barbara di Camillo e Francesco Sambo, per avermi guidato nella realizzazione di tutto il lavoro, per i consigli e per la disponibilità.

Ringrazio anche il dottor Emanuele Trifoglio per avermi aiutato a svolgere gran parte del lavoro di questa tesi e per la disponibilità dimostrata.

Un ringraziamento alla mia famiglia che mi ha sostenuto e incoraggiato in tutti gli anni di studio, e alle mie compagne Emanuela, Sara, Marta e Valentina per i cinque bellissimi anni passati insieme.

This study makes use of data generated by Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is avaible from www.wtccc.org.uk. Funding for the project was provided by Wellcome Trust under award 076113 and 085475'.

Indice

Sommario

Gli studi di associazione intervengono nello studio di dataset casi/controllo caratterizzati da un'importante mole di dati. Tipicamente un dataset è costituito da un numero di soggetti per classe dell'ordine delle migliaia e da un numero di variabili, i polimorfismi a singolo nucleotide, nell'ordine di 10^8 È necessario operare una riduzione sensata del numero di variabili iniziali per poter operare efficientemente una classificazione. Dopo un'attenta analisi dello stato dell'arte, e dopo aver evidenziato i limiti principali dovuti essenzialmente alla feature selection e alla riduzione di informazione che ne consegue, in questa tesi si propone un nuovo approccio basato sulla definizione di mutua informazione per la definizione di metavariabili, da utilizzarsi poi nella classificazione. La metodologia viene applicata con successo su due distinti dataset, isolando in prima battuta il pathway dell'insulina e procedendo alla classificazione delle metavariabili ricostruite. L'applicazione del metodo nella sua completezza prevede la costruzione di un classificatore aggregato dei singoli classificatori costruiti su diversi pathway biologici.

Introduzione

I recenti sviluppi nelle tecnologie del sequenziamento del genoma hanno permesso l'applicazione a livello di popolazione su larga scala dei test di associazione genetica sfruttando in particolare uno specifico tipo di marcatori: i polimorfismi a singolo nucleotide, o *SNPs,* i quali contribuiscono a regolare la predisposizione a una determinata patologia. I test in questione vengono denominati Test di Associazione *Genome Wide*, in quanto, come base di partenza, considerano la totalità di dati a disposizione dal sequenziamento degli SNPs (circa 10 milioni nel genoma umano). L'obiettivo che questi studi si propongono è quello di effettuare una classificazione tra campioni di soggetti che si dividono in controlli (soggetti sani) e casi (soggetti affetti da una specifica patologia che ha base genica) sulla base delle differenze di frequenza allelica con cui gli SNPs si presentano. Nella prima parte della tesi, dopo un breve inquadramento della tematica da un punto di vista biologico, viene proposto lo stato dell'arte che riguarda i test di associazione facendo particolare attenzione a come i ricercatori si pongono davanti alla grande mole di dati che si hanno a disposizione in questo genere di studi. Se da una parte un dataset estremamente informativo può costituire un enorme vantaggio, dall'altra, un enorme numero di variabili costituisce un limite all'applicazione di qualsiasi algoritmo di classificazione. È' necessario ridurre il numero di variabili in gioco. Dallo stato dell'arte, si osservano due procedure alternative (individuazione di tagSNPs, o selezione univariata e successivamente multivariata) che sicuramente portano a una riduzione delle variabili da considerare, ma allo stesso tempo riducono pesantemente il contenuto informativo del dataset iniziale. L'idea della tesi nasce proprio da questa osservazione e si pone come obiettivo quello di ridurre il numero di varaibili ma mantenendo la quantità globale di informazione intatta. Per farlo, si ricorre alle definizioni di entropia e mutua informazione che andranno a stabilire il criterio con cui costruire le metavariabili. La tesi, rispetto allo stato dell'arte, propone un'altra novità: infatti, a differenza delle nomali procedure che considerano separatamente i cromosomi nello studio di determinate regioni geniche, qui viene considerato un pathway nel suo insieme. Si ritiene infatti che, in questo modo, sia possibile individuare, se esiste, una rete di regolazione tra diversi geni di un unico pathway che hanno la caratteristica di intervenire nella regolazione di un unico prodotto.

Questa strategia è stata applicata a due dataset, rilasciati dalla WTCCC, costituiti da controlli e casi di diabete di tipo 1 e 2 rispettivamente.

La tesi si compone di 8 capitoli. Nel primo capitolo vengono brevemente forniti alcuni concetti base relativamente alla biologia degli argomenti trattati. Nel secondo capitolo si approfondisce il fenomeno del Linkage Disequilibrium, le misure ad esso associato e lo stato dell'arte dei principali algoritmi disponibili per la definizione di aplotipi e tag SNP. Nel terzo capitolo viene presentato in generale il problema della stratificazione di popolazione, ricorrente nella maggior parte degli studi di associazione, e vengono riportati i principali metodi per la sua risoluzione. Il quarto capitolo è dedicato alla descrizione di uno dei più importanti database disponibile gratuitamente in rete per la consultazione degli SNPs: HapMap. Viene brevemente riportata la storia del progetto, le sue finalità e i metodi di consultazione. Il quinto capitolo descrive il programma PLINK utilizzato per l'analisi preliminare sui dati WTCCC, per la mappatura dei marcatori e il preprocessing. Ne vengono brevemente illustrate le principali funzionalità. Il sesto, settimo e ottavo capitolo sono infine dedicati alla descrizione dei dati disponibili, al metodo implementato e ai risultati ottenuti dalla mappatura della rete, ricostruita con le metavariabili, e dalla classificazione ottenuta.

Capitolo 1

Polimorfismi a singolo nucleotide e test di associazione

In questo capitolo si introducono alcuni concetti fondamentali oggetto dello studio della tesi. L'intenzione di questa sezione non è di presentare in modo esaustivo tutti gli argomenti sulla genetica e sulla biologia, per la quale si rimanda ai testi specifici, ma richiamare i fondamenti di biologia cellulare e genetica necessari per la comprensione degli argomenti di cui si discuterà in seguito.

1.1 DNA e Polimorfismi di un singolo nucleotide

Il DNA è il veicolo per l'immagazzinamento e la trasmissione dell'informazione genetica, codificata nella sequenza lineare dei nucleotidi, utile alla sopravvivenza della cellula nell'organismo. Questa informazione è trasferita in maniera opportuna alla cellula, affinché questa possa essere operativa. La molecola deputata al trasferimento di questa informazione è l'RNA e il segmento codificante per un prodotto biologico operativo è il *gene*. Il gene rappresenta un carattere ereditario (mendeliano o monofattoriale) definito da un segmento di DNA. In ciascun cromosoma, i geni hanno un ordine preciso e ciascuno occupa una posizione specifica detta locus. Le forme alternative di un gene di dicono *alleli*; in un individuo, i due alleli occupano sui cromosomi omologhi lo stesso locus. È da sottolineare che i geni rappresentano la porzione codificante del genoma e corrispondono solo a circa il 2% dell'intera sequenza. Il resto è costituito da sequenze non codificanti (ripetizioni, sequenze introniche, regioni intrageniche,..) e per gran parte di esse non è ancora chiara la funzione. Il gene è l'unità fondamentale, fisica e funzionale, dell'informazione genetica e contiene le istruzioni per l'assemblaggio di proteine e RNA funzionali. L'espressione dell'informazione codificata in un gene avviene in due stadi: la trascrizione e la

traduzione. Nella trascrizione il filamento di DNA fa da stampo per produrre RNA messaggero (mRNA) che viene poi opportunamente modificato. L'mRNA servirà poi a sua volta da stampo per l'assemblaggio di proteine.

La salvaguardia del materiale genetico richiede meccanismi estremamente precisi sia di duplicazione che di riparazione. Nonostante ciò, possono avvenire nel DNA di una cellula delle variazioni casuali della normale sequenza nucleotidica (mutazioni). Le malattie genetiche sono causate proprio da mutazioni del genoma. Si parla di *polimorfismo* quando una variazione genetica ha una frequenza maggiore dell'1% nella popolazione, dove con variazione si intende sostituzione, aggiunta o delezione di un nucleotide. Il fenomeno può avvenire lungo l'intera sequenza, quindi sia in corrispondenza di regioni codificanti che non codificanti. Il più semplice dei polimorfismi è il risultato della mutazione di una singola base in cui il nucleotide viene sostituito con un altro. I polimorfismi di un singolo nucleotide (*Single Nucleotide Polimorphism*, SNP, Figura 1.1.a, [1]) vengono definiti come differenze di una singola base in sedi specifiche. Gli SNP costituiscono fino al 90% delle differenze delle sequenze tra individui umani e nel genoma umano si verificano a diversi intervalli di basi (mediamente ogni 200 basi nucleotidiche). La presenza degli SNP, e degli alleli con cui si presentano, è verificata attraverso il *sequenziamento* di un campione di DNA. Lo specifico set di alleli osservato su un singolo cromosoma, o su parte di esso, è chiamato *aplotipo* (Figura 1.1.b [1]). Nuovi aplotipi sono formati da ulteriori mutazioni o da eventi di ricombinazione che hanno luogo nel momento in cui i cromosomi materno e paterno si scambiano rispettivamente i corrispondenti segmenti di DNA per dare un cromosoma che risulta un "mosaico" degli aplotipi parentali. È stato empiricamente osservato che il numero degli aplotipi nella popolazione è limitato, in quanto gli SNP e i loro alleli mostrano un forte livello di associazione (fenomeno conosciuto come linkage disequilibrium che sarà descritto in dettaglio in seguito).

L'elevata frequenza di questi marker nel genoma li rende uno strumento ideale per la precisa mappatura genica di differenti malattie. Sebbene la maggior parte di essi non abbia effetti diretti sulla funzione cellulare (per esempio non cambiano la sequenza proteica, né alterano la trascrizione), molti sono strettamente associati, attraverso il linkage disequilibrium (paragrafo 1.2), con alleli o geni causa di malattie e spesso controllano il modo in cui il soggetto risponde a specifici farmaci [1].

Figura 1.1: SNPs e aplotipi. ***(a):*** *Sono mostrati quattro segmenti di DNA da quattro versioni della stessa regione cromosomica di persone diverse. La maggior parte della sequenza è identica in questi cromosomi, ma si evidenziano tre basi in cui si verifica una variazione. Ogni SNP ha due possibili alleli; il primo allele mostrato nel pannello* ***a*** *presenta gli alleli C e T.* ***(b)****: Aplotipi. Un aplotipo è composto da una particolare combinazione di alleli di SNPs vicini. Sono mostrati qui i genotipi osservati per 20 SNPs che si estendono per 6,000 basi di DNA. Vengono mostrate solo le basi variabili, inclusi i tre SNPs che sono mostrati nel pannello* ***a****.****(c)****: sono evidenziati i tagSNPs che sono gli SNPs più rappresentativi dell'aplotipo, fortemente correlati con gli altri.*

1.1.1 Frequenza degli SNPs e Minor Allele Frequency

La frequenza di un allele *a* in una popolazione è definita come

$$f_a = \frac{numero\ di\ copie\ dell'allele}{numero\ totale\ degli\ alleli\ di\ quel\ locus}$$

Si propone di seguito un esempio di calcolo della frequenza di un allele. Si suppone di avere una popolazione costituita da 8 marker che presentano le seguenti coppie alleiche {AA, Aa, AA, aa, Aa, AA, AA, Aa}. Si vuole ora calcolare la frequenza dell'allele A; considerando che ognuno dei marker ha una coppia di alleli, quindi si ha un totale di 16 alleli nel locus. Risulta:

$$f_A = \frac{2+1+2+0+1+2+2+1}{16} = 0.6875$$

Sono verificate anche le seguenti relazioni:

$$p = f_{AA} + \frac{1}{2} f_{Aa} = f_A$$

$$q = f_{aa} + \frac{1}{2} f_{Aa} = f_a = 1 - p$$

All'intero di una popolazione è possibile determinare una minor frequenza allelica (*minor allele frequency*, MAF), definita come il rapporto tra la frequenza della variante più rara e quella più comune di un determinato SNP. Lo studio della MAF consente di distinguere gli SNP monomorfici (MAF<5%) da quelli polimorfici (MAF≥5%). Una MAF troppo bassa, tipicamente minore del 5%, indica generalmente errori nel sequenziamento o polimorfismi troppo rari (MAF<0.01). Un aspetto importante è che la MAF influenza la quantità di falsi positivi nei test statistici, quindi la scelta di una soglia, al di sotto della quale scartare gli SNP con MAF minore, non è banale. Generalmente la maggior parte degli studi non tiene conto di marker con MAF<5-10% in quanto hanno scarsa capacità di individuare l'effetto genetico su una determinata malattia. La distribuzione della MAF è variabile da popolazione a popolazione: uno SNP molto comune in un gruppo etnico può invece essere molto raro in un altro.

1.2 Linkage disequilibrium (LD)

Con il termine *linkage* intendiamo la presenza di geni in loci vicini sullo stesso cromosoma. Si verifica la tendenza di loci (geni e/o marcatori) vicini ad essere trasmessi assieme, come un'unità, attraverso la meiosi. Con *linkage disequilibrium* si specifica la combinazione di alleli in fase a due o più loci che si verifica più spesso di quanto non accadrebbe per puro caso. Due marker, siano essi entrambi genetici, o un marker genetico e il marker malattia, si dicono in LD quando si presentano insieme in uno stesso individuo più frequentemente di quanto ci si attenderebbe per caso. La presenza di un LD indica dunque cosegregazione di due marker e, nel caso di un marker genetico e del marker malattia, questo indica la presenza di associazione del polimorfismo studiato con un aumentato rischio di insorgenza della malattia. In generale, il LD tra due SNPs decresce con la distanza fisica e l'estensione del LD varia enormemente in dipendenza della regione del genoma considerata. Per questo capitolo, si fa riferimento a una descrizione puramente qualitativa del fenomeno, distinguendo con "alto" e "basso grado" la presenza o meno del LD, in quanto si intende valutare il

LD solo da un punto di vista biologico. Per una valutazione quantitativa si rimanda al capitolo successivo.

Per completezza, si riporta anche la definizione di *linkage equilibrium* con la quale si fa riferimento alla segregazione completamente indipendente dei marker sui cromosomi. È quindi una condizione tale per cui il LD è assente.

1.2.1 Variazioni di LD nel genoma

Il LD è fortemente influenzato dal tasso di ricombinazione locale, dalla distanza genica considerata tra i marcatori, ed è correlato con altri fattori, a loro volta associati al tasso di ricombinazione, come il contenuto di basi G e C, la densità genica e la presenza di SINE o ripetizioni *Alu*[1] [1]. Tra le varie popolazioni si osserva un certo grado di similarità nelle regioni classificate ad alto o basso linkage disequilibrium. Questo deriva non solo da un passato storico condiviso tra le popolazioni, ma anche dal fatto che i fenomeni che modulano il linkage disequilibrium possono essere influenzati dalle caratteristiche delle sequenze locali.

È generalmente confermato che mentre i colli di bottiglia della popolazione, le suddivisioni geografiche, e la selezione naturale tendono ad aumentare l'estensione del linkage disequilibrium nella singola popolazione, la crescita della popolazione e l'accoppiamento casuale tendono a diminuire l'estensione del linkage disequilibrium nel genoma. In aggiunta a questi fattori di genetica di popolazione, l'estensione del linkage disequilibrium in una particolare regione del genoma può anche essere influenzata da caratteristiche fisiche delle sequenze di DNA circostanti. Alcuni tipi di sequenze, come ad esempio le sequenze ricche di basi G e C, possono essere associate ad alti tassi di ricombinazione e/o mutazione, due fenomeni che possono direttamente abbassare i livelli circostanti di LD. Il linkage disequilibrium è generalmente più debole vicino alla terminazione dei cromosomi, probabilmente dovuto all'alto tasso di ricombinazione nell'intorno di queste regioni, ed è più forte attorno ai centromeri e nelle altre porzioni interne di ogni cromosoma (ad esempio eterocromatina e larghe porzioni ripetute del genoma), dove i tassi di ricombinazione sono in media più bassi. Inoltre, il LD è generalmente più forte nei cromosomi grandi, che hanno un tasso di ricombinazione più

[1] Le SINE sono un tipo di sequenze intersperse ripetute, generalmente brevi e presenti nelle zone altamente codificanti. Le SINE più comuni presenti nei mammiferi (e quindi anche nell'uomo) sono quelle della famiglia *Alu*. Presenti nel genoma umano in oltre 1 milione di copie, costituiscono circa l'11% del patrimonio genetico totale. Vengono utilizzate come marker genici.

basso, e più debole nei cromosomi piccoli, che hanno una ricombinazione più alta. La presenza di sequenze ripetute in ogni regione è fortemente associata al linkage disequilibrium, e questa correlazione appare significativamente aumentata quando si considerano finestre di genoma più grandi. La presenza di ripetizioni LINE è associata ad un aumento del livello di LD, mentre altri tipi di ripetizione, in particolare le SINE e le *Alu*, sono fortemente associate a un abbassamento del livello di LD.

Per caratterizzare ulteriormente la relazione tra linkage disequilibrium e le variazioni di sequenza nel genoma, Smith AV. *et a*l [2] hanno diviso il genoma in quartili secondo i livelli stimati di linkage disequilibrium in ogni finestra di 100-kb. La minor frequency allele (MAF) degli SNPs in ognuno dei quattro quartili (corrispondenti alle regioni di LD di livello alto, basso e intermedio) era molto simile. Dallo studio è risultato che il contenuto di G e C decresce gradualmente in corrispondenza dell'aumento del grado di LD (da 4349 nucleotidi di GC su 10,000 basi nel quartile con il disequilibrium più basso, a 3904 nucleotidi GC per 10,000 basi nel quartile con il disequilibrium più alto). Si è notato inoltre che i geni sono significativamente più concentrati nei quartili con il livello più alto e più basso di LD rispetto ai quartili del livello intermedio. Questo suggerisce che mentre per alcuni geni può essere vantaggioso essere localizzati in regioni di forte linkage disequilibrium, dove avvengono poche ricombinazioni alleliche, altri geni possono essere favoriti dalla diversità di aplotipi presenti. Smith *et al* [1] hanno inoltre provveduto ad una annotazione funzionale di ogni gene, in dipendenza dalla loro localizzazione nei quartili a più alto e più basso LD, mediante il database di Gene Ontology (GO). Geni associati alla risposta immunitaria (inclusi geni coinvolti nella risposta infiammatoria, umorale, e nella risposta ad agenti patogeni e parassiti) e ai processi neurofisiologici (inclusa la percezione sensoriale) sono spesso localizzati in regioni con basso grado di LD. Al contrario, geni associati alla risposta al danneggiamento del DNA, al ciclo cellulare, o al metabolismo del DNA e RNA sembrano più spesso localizzati in regioni a forte linkage disequilibrium. Questi ultimi geni infatti rappresentano i processi biologici conservati dove la ricombinazione e la mutazione si trasformano in cambiamenti deleteri che vengono generalmente rimossi attraverso la selezione naturale.

1.2.2 Aplotipi e blocchi di aplotipi

Un aplotipo, come già anticipato precedentemente, è una particolare combinazione di alleli a diversi siti polimorfici sullo stesso cromosoma osservati in una popolazione a forte LD. Gli aplotipi rappresentano regioni ereditate senza sostanziale ricombinazione negli antenati della popolazione attuale, quindi, alla luce di quanto visto, sono regioni in cui il livello di LD è molto alto. In prima analisi, si può notare che la variazione genetica è organizzata in segmenti relativamente corti a forte linkage disequilibrium, individuati come *blocchi di aplotipi*, ognuno dei quali contenente pochi aplotipi comuni separati da punti caldi (*hotspots*) di ricombinazione.

La maggior attrattiva nell'uso degli aplotipi è l'idea che gli aplotipi comuni riescono a catturare la maggior parte delle variazioni genetiche in regioni del genoma abbastanza grandi, e possono essere testati utilizzando un numero ridotto di SNP (*tagSNP*) che, all'interno dell'aplotipo considerato, risultano essere più significativi e in correlazione con gli altri marcatori dello stesso aplotipo.

Nonostante il LD costituisca un mezzo potente per l'analisi di associazione, ci sono tuttavia delle forti limitazioni di cui va tenuto conto. Il fenomeno del LD è infatti influenzato da altri fenomeni che fanno soprattutto riferimento agli eventi evolutivi, come ad esempio la storia della popolazione (la sua struttura geografica e eventuali cambiamenti nella dimensione della popolazione stessa), mutazioni, selezione naturale e deriva genetica. Di conseguenza, modelli deterministici che mettano in relazione il tasso di ricombinazione e il LD possono fallire nel tentativo di catturare l'enorme fenomeno stocastico che è alla base del processo evolutivo, e può generare risultati fuorvianti nella ricostruzione dei pattern di variazione genetica

1.3 Equilibrio Hardy-Weinberg

La legge di Hardy-Weinberg (HWE) descrive ciò che accade ad alleli e genotipi in una popolazione "ideale" infinitamente grande, con accoppiamenti casuali e non soggetta ad alcuna forza evolutiva come mutazione, migrazione e selezione. In tali condizioni il modello di Hardy-Weinberg predice che:

1. Le frequenze degli alleli in un pool genetico non variano nel tempo;

2. Se si considerano due alleli, *A* e *a*, a un dato locus, dopo una sola generazione di incroci casuali, le frequenze genotipiche *AA*:*Aa*:*aa* nella popolazione possono essere espresse come

$$p^2 + 2pq + q^2 = 1$$

dove p = frequenza dell'allele *A* e q = frequenza dell'allele *a*.

La relazione è verificata considerando le relazioni riportate nel paragrafo 1.1.1 per p e q. Risultano le relazioni (graficate in Figura 1.3):

$$p = 1 - q$$

$$f_{AA} = p^2 \quad f_{aa} = q^2 \quad f_{Aa} = 2qp$$

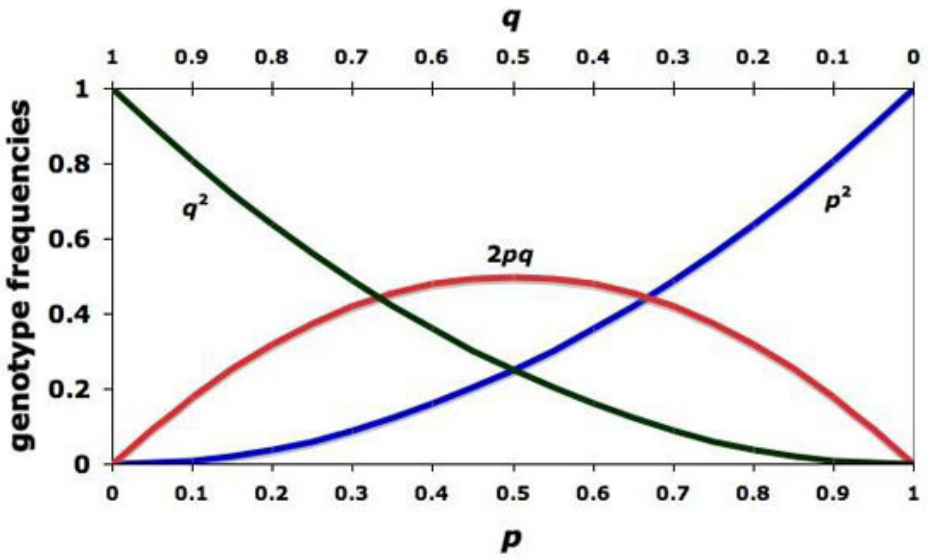


Figura 1.3: il principio di Hardy-Weinberg per due alleli. L'asse orizzontale mostra le frequenze dei due alleli p e q, e l'asse verticale mostra le frequenze del genotipo. Ogni curva mostra uno dei tre genotipi.

Una popolazione che rispetta questi criteri, e in cui le frequenze p e q dei due alleli a un dato locus danno luogo alle frequenze genotipiche attese, si dice in equilibrio di Hardy-Weinberg. Il modello utilizza i principi mendeliani della segregazione e semplici calcoli probabilistici per spiegare la relazione esistente tra frequenze alleliche e genotipiche in una popolazione. La legge stabilisce che le frequenze alleliche nella popolazione non cambiano da una generazione alla successiva e le frequenze genotipiche, dopo una generazione di incroci casuali, possono essere dedotte dalle frequenze alleliche. In altri termini, questa popolazione non cambia o si evolve rispetto al locus che abbiamo preso

in considerazione. Si ricordi tuttavia che, perché ciò si verifichi, devono essere rispettate le assunzioni relative alla popolazione teorica descritta dal modello di Hardy-Weinberg.

1. Tutti gli individui con qualsiasi genotipo hanno lo stesso tasso di sopravvivenza e uguale successo riproduttivo, cioè non c'è selezione.
2. Nessun nuovo allele viene creato o modificato nella popolazione per mutazione.
3. Non c'è migrazione di individui verso l'interno o verso l'esterno della popolazione.
4. La popolazione è infinitamente grande. In termini pratici, cioè significa che una popolazione è sufficientemente grande per cui errori di campionamento e altri effetti casuali risultano irrilevanti.
5. Gli accoppiamenti nella popolazione sono casuali.

Queste assunzioni rendono la legge di Hardy-Weinberg particolarmente utile per la ricerca nel campo della genetica di popolazione. Specificando le condizioni in cui una popolazione non può evolversi il modello di Hardy-Weinberg identifica le forze che nel mondo reale causano il cambiamento delle frequenze alleliche e quantifica le diverse forze evolutive. Ci sono altre importanti conseguenze della legge: la prima è la dimostrazione che un carattere dominante non aumenta necessariamente da una generazione alla successiva; la seconda è la dimostrazione che la variabilità genetica, una volta stabilitasi in una popolazione ideale, può essere mantenuta sino a quando le frequenze alleliche non cambiano; la terza è che, se sono valide le assunzioni di Hardy-Weinberg, una volta conosciuta la frequenza di un solo genotipo è possibile calcolare le frequenze di tutti gli altri genotipi per quel locus. Questo aspetto è particolarmente utile, in quanto permette di calcolare la frequenza dei portatori eterozigoti di patologie genetiche recessive conoscendo solo la frequenza degli individui affetti.

In realtà, è difficile che una popolazione reale sia totalmente conforme al modello di Hardy-Weinberg e che tutte le frequenze alleliche e genotipiche rimangano immutate generazione dopo generazione. Una violazione delle assunzioni di Hardy-Weimberg (sopra elencate), sotto forma di selezione, mutazione, migrazione, deriva genetica e incrocio non casuale possono causare un cambiamento delle frequenze alleliche e genotipiche [4].

1.4 Studi di associazione

Gli studi di associazione sull'intero genoma (*whole-genome association studies*, GWAS) sono stati recentemente proposti come un approccio molto efficace per l'individuazione di molte cause e fattori genetici che costituiscono la causa fondamentale di malattie comuni. A differenza degli studi di linkage, che considerano il fenomeno dell' ereditarietà di regioni cromosomiche legate alla presenza di malattie all'interno di una famiglia, gli studi di associazione considerano invece la differenza tra la frequenza di varianti genetiche rilevata in individui non imparentati tra loro e affetti da un disturbo (casi) e la stessa frequenza misurata in soggetti sani (controlli).

Gli studi di associazione possono avvenire mediante due approcci: diretto ed indiretto. Uno studio di associazione diretto consiste nel catalogare e testare una ad una tutte le possibili varianti causali. Le frequenze di queste varianti sono quindi confrontate tra casi (pazienti) e controlli; il risultato che si attende è che la variante che conferisce la predisposizione al disturbo sia più frequente nei casi. Tuttavia, risulta chiaro fin dall'inizio che l'approccio diretto presenta non pochi problemi pratici. Un'applicazione di questa strategia sull'intero genoma implica l'identificazione di tutti i geni umani (fino a 30,000 geni) così come di tutte le loro varianti. Per di più, è difficile identificare varianti che si collocano in regioni introniche e polimorfismi. Per queste ragioni l'utilizzo del metodo diretto è limitato a pochi casi e quasi sempre sostituito con l'applicazione del metodo indiretto. La strategia indiretta evita la necessità di catalogare tutte le varianti che potenzialmente potrebbero dare predisposizione ad un dato disturbo, e si basa invece sull'associazione tra malattia e polimorfismi (marker) localizzati vicino a locus strategici. Tali associazioni derivano da studi di linkage disequilibrium tra marker e loci coinvolti nella predisposizione al distubo in esame. La strategia indiretta quindi impiega una densa mappa di marker polimorfici per esplorare il genoma in maniera sistematica. La scelta dei marker differenzia ulteriormente l'approccio indiretto in due diverse strategie. Nella prima i marker vengono scelti molto vicini alle regioni esoniche di geni conosciuti. La seconda impiega anche marker localizzati in regioni più ampie, praticamente in tutto il genoma, considerando quindi i cromosomi nella loro interezza, incluse regioni introniche e regioni comprese tra un gene e l'altro. La scelta dei marker ricade in ogni caso sugli SNPs biallelici a causa della loro alta frequenza con cui compaiono nel genoma umano, per il basso tasso di mutazione a cui vanno incontro e per la facilità con cui possono essere analizzati.

Capitolo 2

Linkage disequilibrium

La disponibilità di vasti data set provvede a dare un'informazione sul linkage disequilibrium (LD) per oltre un milione di markers. L'analisi del LD aiuta nell'interpretazione degli studi di associazione genome-wide, che hanno come base di partenza proprio i risultati di questa analisi, e facilita l'identificazione degli alleli che dimostrano una predisposizione a determinate malattie.[1] Il vantaggio è dato dalla possibilità di indagare sulla relazione esistente tra sequenze locali, in particolare si fa riferimento ai polimorfismi a singolo nucleotide biallelici (SNPs) e il locus malattia, basandosi sui pattern di LD. La conoscenza della distribuzione del LD è necessaria per tradurre in termini quantitativi l'associazione tra SNPs e il fenotipo osservato e per identificare la possibile collocazione nel genoma della variante indagata.

2.1 Misure di Linkage Disequilibrium

Sono state proposte numerose misure di LD: la scelta della misura può avere un sostanziale impatto sull' accuratezza e sull'interpretabilità del risultato ottenuto [3].

2.1.1 Misure di LD Pairwise

Le misure più comuni, di seguito presentate, sono limitate a un confronto tra due loci, per questo motivo vengono denominate *Pair-Wise LD Measures*. Si considerino quindi due loci, ogni locus avente due alleli. Un allele malattia e un allele normale segregano nel primo locus, e due alleli marker, A1 e A2 rispettivamente, segregano nell'altro locus. Si ricava, da un campione di popolazione, una tabella 2x2 riportata in Tabella 2.1.

	Allele malattia	Allele normale	
Allele A1	n_{11}	n_{12}	n_{1+}
Allele A2	n_{21}	n_{22}	n_{2+}
	n_{+1}	n_{+2}	n

Tabella 2.1: layout di un campione di frequenze in una tabella 2x2

In Tabella 2.1, n_{11} è il numero di aplotipi nel campione che portano l'allele malattia e l'allele marker A1, n_{1+} è il numero di aplotipi che portano l'allele A1, n_{1+} è il numero di aplotipi che potano l'allele malattia e n è il numero totale di aplotipi campionati. Dividendo queste quantità per n, ci si riporta alle probabilità p (Tabella 2.2).

	Allele malattia	Allele normale	
Allele A1	p_{11}	p_{12}	p_{1+}
Allele A2	p_{21}	p_{22}	p_{2+}
	p_{+1}	p_{+2}	1

Tabella 2.2: layout di un campione con le relative probabilità in una tabella 2x2

Da qui è possibile calcolare le probabilità condizionate: ad esempio, la probabilità di avere l'allele A1 nell'aplotipo, sapendo che l'allele malattia è presente, è denotata con $p_{1|+1} = p_{11} / p_{+1}$. In maniera analoga, la probabilità di avere l'allele normale nell'aplotipo, sapendo che è presente l'allele marker A2, è data da $p_{2|2+} = p_{22} / p_{2+}$. Naturalmente, le p sono stime campionarie dei parametri reali non noti, indicati con π. Si userà π per le definizioni che seguono, con l'idea che queste quantità non note vengano poi stimate dalle osservazioni sul campione. Vengono di seguito riportate le misure più frequentemente utilizzate per la quantificazione del linkage disequilibrium [2]. La componente fondamentale di molte misure di disequilibrium è la differenza tra il numero di aplotipi osservato e atteso che portano l'allele malattia e l'allele A1, o la sua espressione equivalente:

$$D = \pi_{11} - \pi_{1+}\pi_{+1} = \pi_{22} - \pi_{2+}\pi_{+2} =$$

$$= \pi_{11}\,\pi_{22} - \pi_{12}\,\pi_{21}$$

Quando la differenza tra i due prodotti è grande, allora ho alto LD, mentre risulta basso quando ho una differenza minima prossima al valore nullo. Se D=0 ritrovo la

condizione di linkage equilibrium. Questa misura è in cui varia dipende dalle frequenze alleliche, rendendo difficile il confronto tra marker diversi. di difficile interpretazione in quanto il segno che assume è arbitrario e il range Si ricorre più spesso alla versione normalizzata di *D*, definita come *D'*:

$$D' = \begin{cases} \dfrac{D}{\min(\pi_{1+}\pi_{+2}, \pi_{+1}\pi_{2+})}, & D > 0 \\ \dfrac{D}{\min(\pi_{1+}\pi_{+1}, \pi_{+2}\pi_{2+})}, & D < 0 \end{cases}$$

che varia in un range compreso tra ± 1. Quando assume valore ±1 significa che ho un massimo valore di LD misurato e risulta praticamente assente la ricombinazione tra i marker. L'uso di *D'* comporta principalmente due svantaggi. Il primo è che il suo valore risulta sovrastimato se misurato su campioni di piccole dimensioni. Il secondo svantaggio nell'uso di questa misura è che può assumere un valore prossimo a ±1 (indicando quindi un alto grado di LD) anche quando un allele è molto raro, il quale solitamente risulta di scarso interesse pratico [4,5]. Si può per ultimo osservare che, nel tempo, *D'* tende a decrescere con andamento esponenziale tra loci in LD.

La misura usata più frequentemente è il quadrato della misura standardizzata, Δ o, più spesso indicata con *r*:

$$r^2 = \frac{D^2}{(\pi_{1+}\pi_{+1}\pi_{2+}\pi_{+2})}$$

Questa misura è solitamente elevata al quadrato in quanto si vuole eliminare la possibile arbitrarietà di segno introdotta nel momento in cui gli alleli sono classificati. Di fatto, r^2 è il coefficiente di correlazione di una tabella 2x2, del tipo di Tabella 2.1 o 2.2. Se π_{1+} è uguale a π_{+1}, allora r^2 varia da un valore minimo pari a 0, a un valore massimo pari a 1. In particolare, quando assume valore unitario, significa che i markers forniscono la stessa informazione. Se $\pi_{1+} \neq \pi_{+1}$, allora varia da un minimo pari a 0, e risulta $r^2<1$. Vengono di seguito riportati gli andamenti del valore di LD, espresso mediante le misure appena introdotte, in funzione della distanza fisica tra i marker e del tempo. Si valuta inizialmente il comportamento del LD in funzione della distanza fisica. In Figura 2.1 è mostrato l'andamento della misura r^2 per una finestra del genoma compresa tra 2 e 3 Mbp sul cromosoma 3. L'approccio è basato su un modello (di Otha e Kimura, 1986,

[1]) che predice il valore del disequilibrium atteso tra due alleli, posizionati rispettivamente ai loci *i* e *j* , come:

$$E(r^2{}_{ij}) = \frac{1}{1 + R_{ij}}$$

dove R_{ij} è il tasso di ricombinazione della popolazione dato a sua volta da $R_{ij} = 4Nc_{ij}$, con c_{ij} frazione di ricombinazione tra i markers[2] *i* e *j*, e *N* dimensione della popolazione.

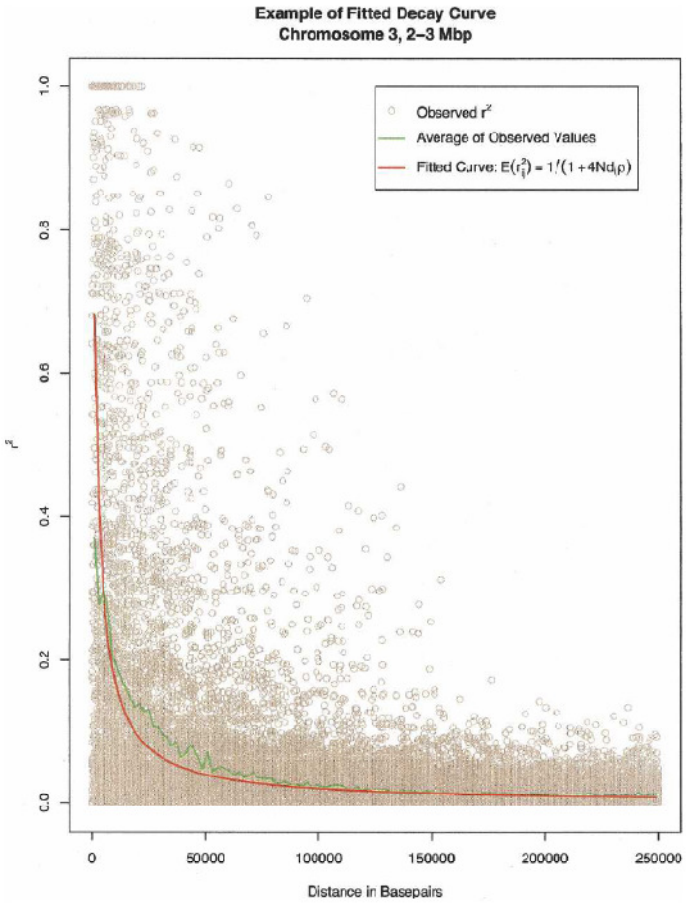


Figura 2.1: coefficiente pairwise di disequilibrium r² per una finestra di genoma. In ascissa si ha la distanza misurata in coppie di basi, in ordinata il valore di r². I cerchi indicano i valori osservati. La curva verde indica la media dei valori osservati. La curva rossa indica la curva risultante dal fit del modello, che riproduce il decadimento del LD come funzione del tasso di ricombinazione e della distanza tra markers. L'esempio fa riferimento alla finestra tra 2 e 3 Mb sul cromosoma 3

[2] La frazione di ricombinazione è la probabilità che avvenga un evento di ricombinazione tra due loci ed è data dal rapporto (numero meiosi ricombinanti)/(numero totale meiosi). Può assumere valore compreso tra 0 e 0.5. Loci molto vicini avranno frazione di ricombinazione nulla, mentre loci molto lontani o su cromosomi diversi avranno frazione di ricombinazione pari a 0.5.

Il parametro fittato definisce una curva per il decadimento del LD (in Figura 2.1 è evidenziata in rosso).

L'andamento conferma quanto già in parte detto nel paragrafo 1.2.1 riguardo le variazioni del LD nel genoma. Il modello di Otha e Kimura riportato è stato utilizzato per fittare il valore di r^2 su un genoma completo. Smith et al.[1] hanno utilizzato i dati del database HapMap per la popolazione CEU (si veda capitolo 4.2) suddiviso il genoma in finestre mobili, e in ogni finestra fittato i valori del coefficiente di disequilibrium. Il LD è calcolato tra marker separati di circa 30 kb l'uno dall'altro; i coefficienti sono stati poi calcolati e fittati all'interno di finestre di 100 kb distribuite su tutto il genoma.

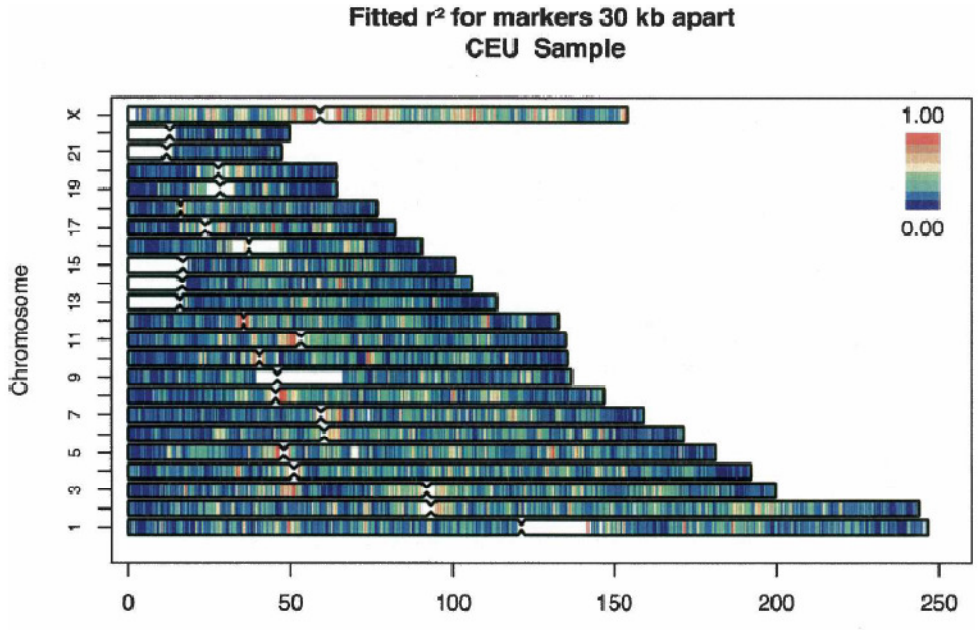


Figura 2.2: riepilogo per tutti i cromosomi dei valori fittati di LD sull'intero genoma (coefficienti di LD per i markers separati di 30 kb). In ascissa è riportata la distanza in Mb, in ordinata il numero identificativo del cromosoma. Il codice a colori indica l'intensità del valore di LD.[Albert VS, Sequence features in regions of weak and strong linkage disequilibrium, Genome Res 2005]

In Figura 2.2 mostra una struttura del LD che vede alternate regioni ad alto LD nel genoma inframezzate da regioni a basso LD, il che può essere dovuto sia ad alti tassi di ricombinazione sia a conversione genica [4].

In merito al secondo aspetto, cioè la dipendenza del LD dal fattore temporale, si osserva che, con il passare delle generazioni, si tende a raggiungere asintoticamente l'equilibrio

(o in altre parole, la condizione di linkage equilibrium), per cui la misura D tende al valore nullo. Il fenomeno che provoca la diminuzione del disequilibrium è appunto la ricombinazione tra i due geni. Dati D_0 il valore iniziale del disequilibrium, D_n il valore del disequilibrium dopo n generazioni, e R il tasso di ricombinazione tra i due marcatori, si ha:

$$D_n = D_0(1 - R)^n$$

L'andamento è graficato in Figura 2.3 da cui risulta evidente, così come dalla formula, che il valore del disequilibrium alla n-esima generazione sarà inferiore a quello della generazione 0, inoltre maggiore è il tasso di ricombinazione, più ripido è il decadimento della curva, quindi più veloce è il raggiungimento dell'equilibrio.

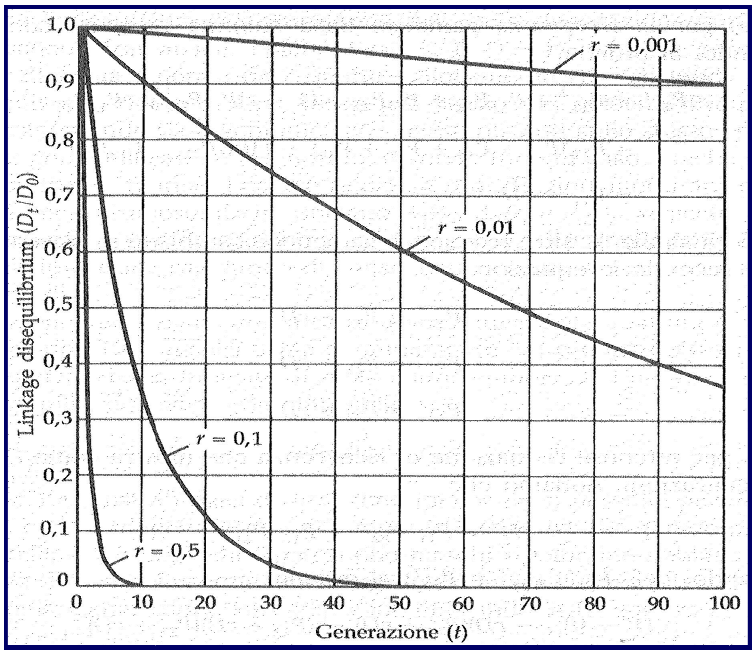


Figura 2.3: andamento del LD nel tempo. In ascissa è riportata la generazione t-esima, mentre in ordinata, il valore percentuale di LD dato dal rapporto $^{D_t}/_{D_0}$ *quindi dal valore D alla t-esima generazione e il valore di D alla generazione 0. La curva decade tanto più rapidamente quanto più alto è il valore di R, cioè del valore di ricombinazione.*

2.1.2 Misure di LD Multilocus

Sia r^2 che *D'* sono misure tra due loci; tuttavia, se si ha a disposizione una densità di marker molto alta, è interessante riassumere il LD su tutta la regione. Sebbene le misure Pairwise siano molto utili nella stima di LD tra coppie, non possono però considerare più di due loci contemporaneamente e quindi non riescono a individuare associazioni simultanee tra più loci.[4] Per superare il limite delle misure pairwise proposte, si ricorre a misure che considerano più loci contemporaneamente, classificate anche come *Multilocus LD Measures*. Si è inizialmente proposto di calcolare i valori di LD tra tutte le coppie di SNPs, costruire una matrice di LD e considerare successivamente le diverse sottomatrici. Ma questo approccio non risolve il problema di come combinare l'informazione dalle diverse sottomatrici per la descrizione del LD tra più loci. Altre soluzioni comprendono l'uso di un modello di Markov nascosto per stimare il tasso di ricombinazione storico tra aplotipi, che è direttamente correlato a *D'*, oppure la suddivisione in blocchi andando a guardare i pattern di ricombinazione usando intervalli di confidenza per *D'*. Sfortunatamente, queste sono misure specifiche per l'aplotipo osservato, non estendibili ad tutti i possibili aplotipi. È stata recentemente introdotta una nuova misura multilocus in grado di descrive direttamente la forza del valore di LD tra più loci [4]. La *Normalized Entropy Difference* ε è basata sul concetto di entropia *H* normalmente definita come:

$$H = -\sum_i p_i \log p_i$$

dove p_i rappresenta la probabilità degli stati che il sistema può assumere e la sommatoria è estesa a tutti gli stati del sistema. Uno stato mancante (p_i=0) non contribuisce alla misura di entropia, dal momento che 0 log0 è settato a 0 per definizione (quindi *H*=0). L'entropia risulta massima quando tutti gli stati sono equamente probabili: ogni osservazione provvederà a dare il massimo dell'informazione. Al contrario, l'entropia è nulla quando c'è solo uno stato, si conosce quindi in modo esatto il sistema e nessuna osservazione può incrementare ulteriormente l'informazione già disponibile. Una sequenza di due o più loci è ora vista come un sistema dove gli aplotipi possibili sono considerati gli stati di questo sistema. Si considerino ora *m* loci bi-allelici. Questa sequenza può assumere 2^m aplotipi di quali *n* sono assunti presenti. L'entropia è usata per misurare la quantità:

$$H_B = -\sum_{i=1}^{n} p_i \log p_i$$

dove p_i è la frequenza dell'aplotipo *i*. Sotto l'ipotesi di *linkage equilibrium*, p_i può essere espressa come il prodotto delle frequenze alleliche ai loci:

$$q_i = q_{a_1 \ldots a_m} = \prod_{k=1}^{m} p_{(k)}(1 - p_{(k)})$$

dove q_i è la frequenza dell'*i*-esimo aplotipo, a_k (k=1,…m) è l'allele del *k*-esimo SNP nell'aplotipo *i*, $p_{(k)}$ è la frequenza dell'allele del *k*-esimo SNP. L'entropia nel caso di linkage equilibrium, riscritta utilizzando le frequenze così definite è:

$$H_E = -\sum_{i=1}^{2^m} q_i \log q_i.$$

Una deviazione dallo stato di equilibrio, data ad esempio da una riduzione del numero di aplotipi presenti e una diversa frequenza aplotipica rispetto a quelli attesi in equilibrium, rappresenta un incremento di informazione sul sistema. Queste deviazioni daranno una diminuzione dell'entropia rispetto al caso in cui lo stato è in equilibrio. Forze come la ricombinazione e la conversione genica tendono a distruggere i pattern di LD, e nel tempo le sequenze tendono a ritornare allo stato di linkage equilibrium. La differenza:

$$\Delta H = H_E - H_B$$

è quindi la misura della deviazione della sequenza osservata dallo stato di linkage equilibrium atteso. Per permettere confronti tra diversi set di loci si è proposta la versione scalata di ΔH compresa tra [0,1) indicata con ε:

$$\varepsilon = \frac{\Delta H}{H_E} = 1 - \frac{H_B}{H_E}$$

ε è la misura per il LD denominata *Nomalized Entropy Difference*. Per definizione, la misura ε permette di considerare un numero illimitato di loci contemporaneamente. Comunque, nel caso pratico, il numero è limitato dalle dimensioni del campione. Alcuni aplotipi che possono essere presenti con frequenze rilevanti in alcune popolazioni, potrebbero non essere presenti nei campioni osservati, dovuto anche semplicemente al

fatto che il campione è troppo piccolo per poterli osservare. In questo caso, la ε può portare a una sovrastima del LD. La misura è sensibile sia al numero di aplotipi osservati che alle loro frequenze. ε è pari a 0 se e solo se una sequenza è in stato di equilibrium, aumenta al diminuire del numero di aplotipi presenti e aumenta all'aumentare della deviazione delle frequenze dalla condizione di equilibrium. La misura distingue anche tra vari gradi di LD oltre che l'assenza di più di un aplotipo.

Sia le misure Pair-Wise che Multilocus, sono influenzate fortemente nella loro accuratezza della dimensione del campione su cui viene fatto il calcolo di LD. In particolare, campioni più grandi possono minimizzare l'errore di campionamento e produrre una valutazione più adeguata del LD. Inoltre i soggetti appartenenti a gruppi famigliari possiedono più informazione rispetto a campioni di individui scorrelati tra loro. Segue che la valutazione di LD su nuclei famigliari è più accurata.

Risulta evidente, da quanto detto, che i pattern di LD osservati nella popolazione sono il risultato di una complessa interazione tra i fattori genetici e la storia demografica della popolazione stessa. In particolare, la ricombinazione gioca un ruolo chiave nel formare e modellare i pattern di LD in una popolazione. Quando avviene un evento di ricombinazione tra due loci, questo tende a ridurre la dipendenza tra gli alleli che quegli stessi loci portano e quindi riduce il LD. Anche se gli eventi di ricombinazione in una singola meiosi sono relativamente rari, se si considerano piccole regioni, il numero totale di meiosi che hanno luogo ad ogni generazione ha un sostanziale effetto cumulativo sui pattern di LD. A causa della complessità del fenomeno, molti modelli per l'interpretazione e l'analisi del LD non tengono conto di questo aspetto.

2.2 Definizione degli aplotipi mediante misure di LD

Gli aplotipi, o blocchi di aplotipi, rappresentano regioni ereditate senza sostanziale ricombinazione dagli antenati nella popolazione moderna. La storia della ricombinazione tra le coppie di SNPs può essere stimata con l'uso delle misure definite precedentemente. Se si utilizza la misura *D'*, gli aplotipi vengono quindi definiti come regioni in cui la misura standard del disequilibrium *D'* è considerata pari a 1 (o comunque molto vicina al valore unitario, facendo opportunamente riferimento a degli intervalli di confidenza), in assenza di ricombinazione, per tutte (o quasi) le coppie di marker nella regione. Poiché i valori di *D'* tendono ad essere sovrastimati, nel caso in

cui si abbiano a disposizione pochi campioni o si lavori con alleli rari, si preferisce fare riferimento a intervalli di confidenza piuttosto che alle stime puntuali della misura.

A separare gli aplotipi, ossia regioni in cui la ricombinazione è al livelli considerevolmente ridotti, sono le variazioni locali, posizionate lungo il cromosoma, denominate *punti caldi*, o *hotspots*, dove al contrario si registra un tasso di ricombinazione elevato e una caduta significativa del valore di LD. Si possono definire hotspots quei segmenti di lunghezza mediamente inferiore alle 10 kb, dove la misura di r^2 tra due marker vicini non eccede mai il valore di 0.10 . I punti caldi, in sostanza, rappresentano interruzioni del LD, costituiscono meno del 10% delle sequenze, e si è stimato che in queste brevi regioni avvenga circa il 50% degli eventi di ricombinazione. Queste variazioni nel tasso di ricombinazione spiegano almeno in parte le recenti osservazioni del grado di eterogeneità nei valori di LD tra i marker SNPs lungo il genoma. Considerando nuovamente la popolazione CEU rappresentata prima (in Figura 2.1, sul LD), si propone ora la mappa dei punti caldi lungo l'intero genoma in Figura 2.4. Gli hotspots sono individuati mediante la colorazione rossa.

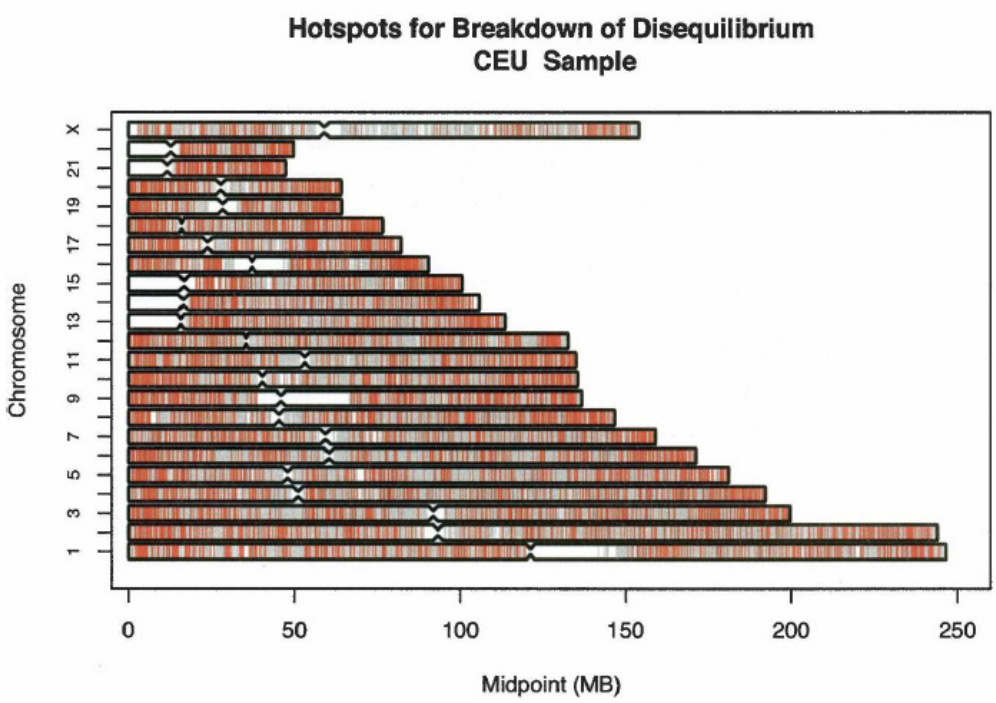


Figura 2.4: riepilogo per ogni cromosoma degli intervalli tra i marker (in rosso) dove il LD decade molto rapidamente. Intervalli in cui il LD decade in modo meno pronunciato sono evidenziate in azzurro. In ascissa è riportata la distanza in Mb, in ordinata il numero identificativo del cromosoma [Albert VS, Sequence features in regions of weak and strong linkage disequilibrium, Genome Res 2005].

2.3 Algoritmi di LD e inferenza di aplotipi

Gli algoritmi per il calcolo del linkage sono derivati dall'iniziale problema del mappaggio genico in base alla frazione di ricombinazione, così da risalire alla posizione del locus malattia nel genoma rispetto al marcatore. In molti casi gli aplotipi non sono "letti" direttamente nel momento in cui il DNA umano viene sequenziato (fatta eccezione di alcuni casi come il sequenziamento del cromosoma Y o tecniche di conversione diploide-aploide), ma vengono dedotti dai dati di genotipo. Un problema fondamentale con la ricostruzione di mappe di linkage nell'uomo è che alcuni dati importanti spesso mancano. In questo modo non è possibile semplicemente contare le ricombinazioni negli incroci, proprio perché la mancanza di tutte le informazioni non permette di capire senza ambiguità dove queste sono avvenute.

Nonostante l'indiscussa necessità di capire i pattern di LD nel genoma, soprattutto a causa dell'impatto che questo aspetto ha nel design e nell'analisi del mappaggio dei geni "malattia" nell'uomo, i metodi più comunemente usati per l'interpretazione e l'analisi del LD soffrono almeno di una delle seguenti limitazioni:

1. Sono basati su una valutazione delle misure di LD definite solo per coppie di loci, piuttosto che considerare tutti i loci contemporaneamente;
2. Assumono una struttura a blocchi per schematizzare i pattern di LD, il che potrebbe non essere appropriato per tutti i loci;
3. Non mettono direttamente in relazione i pattern di LD con i meccanismi biologici di interesse, come ad esempio il tasso di ricombinazione.

In letteratura si possono trovare un'ampia gamma di algoritmi che permettono di risolvere il problema di inferenza di aplotipi. È importante ricordare che questi algoritmi sono pensati per applicazioni su dati di genoma diploide, quindi con patrimonio genetico che presenta due copie di ogni cromosoma. Mentre gli aplotipi rappresentano le informazioni degli alleli di SNPs su *un* cromosoma, i genotipi rappresentano le informazioni combinate di alleli di SNPs su *due* cromosomi. Se ho n set di genotipi, dove ognuno riporta l'informazione di m SNPs, il numero di aplotipi (massimo) sarà 2^m. È necessario sottolineare che il problema di inferenza di aplotipi non è di immediata e diretta soluzione, dovuto all'ambiguità di risoluzione. L'ambiguità nasce dalla presenza di SNPs eterozigoti: in particolare, quando si è in presenza di c (dove $c > 1$) SNPs eterozigoti nel genotipo, ci sono 2^{c-1} coppie di aplotipi che possono risolvere il genotipo.

Quindi, il genotipo non può essere univocamente risolto senza l'aggiunta di vincoli o considerazioni di tipo biologico.

Un tipo di approccio, uno tra i primi implementato, fa riferimento a *modelli di parsimonia*. Questi metodi assumono che una data popolazione target condivida un numero relativamente ridotto di aplotipi in comune dovuto al linkage disequilibrium. Gli algoritmi si basano sulla costruzione incrementale della soluzione mediante una applicazione iterata di una regola di inferenza, detta regola di Clark (dal nome del ricercatore che per primo ha introdotto il principio nel 1990). L'algoritmo identifica inizialmente i genotipi che contengono solo alleli omozigoti o al più un unico allele eterozigote. Questi genotipi possono essere risolti in maniera univoca e la corrispondente coppia di aplotipi viene memorizzata in un set di aplotipi già identificati, denotato con *I*. Per i rimanenti genotipi "ambigui", si esamina l'insieme *I* per vedere se in esso sono contenuti aplotipi che siano compatibili con il dato genotipo. Se viene trovato un aplotipo tale da soddisfare questa condizione, il rispettivo genotipo viene etichettato come "risolto". Il procedimento è iterato fintanto che tutti i genotipi sono stati risolti o non sono trovati nuovi aplotipi. In questi modelli, l'algoritmo ricerca per ogni passo una soluzione localmente ottima (algoritmo di greedy), cercando una soluzione locale che incrementi il meno possibile il numero di aploitipi, senza vincoli relativi alla soluzione globale. L'algoritmo di Clark è semplice e intuitivo, tuttavia presenta numerosi svantaggi: molti dei genotipi "ambigui" possono rimanere non risolti e un ordine differente di iterazione può portare alla costruzione di aplotipi diversi. La mancanza di un modello di soluzione globale implica che la soluzione trovata può essere diversa, ad ogni applicazione dell'algoritmo, a seconda delle scelte effettuate ad ogni passo, quindi non tutte le sequenze di applicazione della regola di inferenza consentono di raggiungere una soluzione al problema. Un'altra formulazione del problema si basa sul principio di pura parsimonia, proposto da Gusfield. In questo caso si ricerca l'insieme di aplotipi di minima cardinalità che risolve l'insieme dei genotipi dati. Tutti i metodi basati sui modelli di parsimonia assumono che il numero di aplotipi distinti in una popolazione sia più piccolo del numero possibile di aplotipi distinti che si avrebbero in condizione di linkage equilibrium (cioè in totale assenza di linkage disequilibrium). Perciò, quando un dataset non soddisfa questa condizione, le performance di questi modelli diventano insoddisfacenti.

I *modelli coalescenti*, invece, si basano su assunzioni biologiche relative alla storia evolutiva delle mutazioni. In particolare, secondo questi modelli, le mutazioni (cioè le transizioni da un allele all'altro in alcuni polimorfismi) avvengono una sola volta e, per questo, sono condivise da un insieme di individui con il medesimo antenato. Ne segue che un cromosoma sprovvisto di tale mutazione, non può essere discendente del medesimo antenato che invece presenta quella mutazione. Inoltre è assunta l'assenza di ricombinazione; segue da ciò che una sequenza target non viene considerata come il risultato di un evento di ricombinazione tra i due cromosomi parentali, ma è come se venisse ereditata da un singolo progenitore. Un approccio di questo tipo assume che gli aplotipi di una popolazione evolvano secondo un modello genetico identificato da un grafo ad albero che descrive la storia evolutiva di un set di sequenze di DNA. Il termine con cui si identifica l'albero è *filogenesi perfetta*. Se si hanno *2n* aplotipi, dove ogni aplotipo è costituito da *m* SNPs, l'albero sarà composto da *2n* foglie, esattamente una per aplotipo, e ognuno degli *m* SNP andrà a costituire gli *m* archi. In Figura x.5 è riportato un esempio di filogenesi perfetta per un set di 4 aplotipi.

	SNP_1	SNP_2	SNP_3
$Haplotype_1$	1	1	0
$Haplotype_2$	1	0	0
$Haplotype_3$	1	0	1
$Haplotype_4$	0	0	0

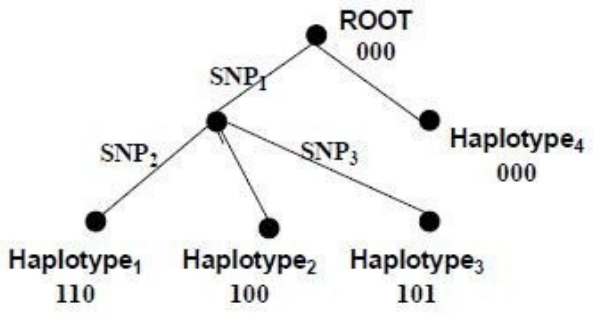


Figura x.5: esempio di filogenesi perfetta. In alto è proposto lo spettro dei 4 aplotipi composti dai 3 SNPs. Mentre la figura sottostante è la rappresentazione mediante grafo ad albero della filogenesi perfetta.

Anche se le perfomance dei metodi basati sulla filogenesi perfetta sono state migliorate, questi soffrono ancora della stretta conformità al modello coalescente: è possibile che non esista filogenesi perfetta per un dato genotipo. Nel caso reale, spesso i dati non soddisfano i requisiti del modello coalescente e violano le assunzioni biologiche alla

base del modello stesso. Le principali ragioni possono essere dovute a errori nel sequenziamento o assenza dell'informazione relativa alla ricombinazione. Per mantenere la filogenesi perfetta si è proposto di eliminare un numero di genotipi dai dati originali in modo che i restanti potessero soddisfare i requisiti di filogenesi perfetta o assegnare un valore arbitrario agli alleli mancanti in modo che i genotipi così ricostruiti potessero ancora soddisfare la filogenesi perfetta. Le soluzioni proposte ancora non hanno dato risultati adeguati e in molti casi manca una soluzione euristica. Un approccio più realistico è offerto dai metodi di *filogenesi imperfetta*: questi assumono che la maggior parte dei genotipi (non tutti) soddisfino il modello della filogenesi perfetta. Quindi, considerano un modello meno "rigido" del precedente che permetta di inserire un certo numero di mutazioni ricorrenti e di ricombinazioni. Tra più soluzioni candidate che soddisfano questo modelli, quella che soddisfa la maximum-likelihood, dato il genotipo, è scelta come soluzione. Tuttavia, la gestione di un numero esponenziale di soluzioni candidato possibili rimane ancora un problema non risolto.

I metodi basati sui modelli di parsimonia e sui modelli coalescenti propongono un approccio diretto nella risoluzione del genotipo con ogni coppia di aplotipo. Viceversa, i metodi basati su *modello statistico* si basano su un approccio più indiretto. Gli algoritmi proposti nei modelli precedenti richiedono un insieme iniziale di aplotipi risolti e la loro soluzione dipende fortemente dall'ordine in cui vengono risolti. L'approccio proposto dai modelli statistici diminuisce l'importanza di tale dipendenza, rendendo la procedura più affidabile. L'idea principale è che gli aplotipi hanno una distribuzione di probabilità non nota nella popolazione in esame e che i genotipi osservati di ogni individuo sono semplicemente combinazione di due aplotipi presi a caso dalla popolazione. L'obiettivo dell'approccio statistico è quindi di stimare la frequenza degli aplotipi in modo che sia possibile calcolare facilmente gli aplotipi di ogni individuo, basandosi sulla distribuzione di probabilità e su considerazioni di tipo biologico. I metodi di risoluzione sono di due tipologie: la *Maximum Likelihood Iinference* e il *Bayesian Frequencies Haplotype Inference problem.*

Gli elementi principali di un modello di tipo statistico per generare dati relativi agli aplotipi e ai genotipi sono:

- $G = (g_1, g_2, \ldots, g_m)$ è l'insieme dei genotipi osservati;

- $H = (H_1, H_2, \ldots, H_m)$ sono le corrispondenti coppie di aplotipi incognite, dove $H_i = (h_{i1}, h_{i2})$ per $i = 1,\ldots,m$;
- $\Theta = (\theta_1, \ldots, \theta_v)$ indica il valore delle v frequenze non note degli aplotipi (dove v è la cardinalità dell'insieme di tutti gli aplotipi);

Comunemente, la stima delle frequenze aplotipiche di una popolazione è fatta partendo da sequenze geniche di individui non correlati tra di loro, campionati casualmente, sotto l'assunzione dell'equilibrio Hardy-Weinberg. La formulazione *Maximum Likelihood Inference* (o formulazione ML) ha come obiettivo quello di massimizzare la distribuzione degli aplotipi nella popolazione, massimizzando la verosimiglianza (*likelihood*) dei dati di genotipo. In input viene dato l'insieme dei G genotipi: a differenza delle frequenze del genotipo, che possono essere direttamente calcolate dal dataset, le frequenze aplotipiche sono non note e vanno stimate. In uscita si ha l'insieme delle frequenze degli aplotipi $\{h_1, \ldots, h_v\}$ (dove v è il numero di tutti i possibili aplotipi) che massimizza la funzione di somiglianza (*likelihood*) per l'insieme di genotipi osservato. Usando le frequenze stimate, ogni genotipo può essere risolto dalla coppia di aplotipi con la frequenza massima tra tutte le coppie compatibili con il dato genotipo. L'algoritmo implementato più diffusamente e proposto per la risoluzione di un problema di *Maximum Likelihood* è l'*Expectation Maximization (E-M) algorithm* di Excoffier e Slatkin (1995), e viene proposto per stimare gli aplotipi e le frequenze che massimizzano la funzione di somiglianza dell'insieme dei genotipi. Sia Θ_{Ht} l'insieme delle frequenze degli aplotipi e G_t l'insieme delle probabilità di tutti i genotipi al passo t. L'algoritmo EM assegna un valore iniziale alle frequenze aplotipiche Θ_{H0} (un insieme iniziale possibile di frequenze è quello che corrisponde all'assunzione che tutti i possibili aplotipi siano equiprobabili). Basandosi su Θ_{H0} , può essere facilmente calcolato il valore atteso di un genotipo di G e può quindi essere calcolato il valore di G_1 (*Expectation* step) I valori attesi dei genotipi in G_1 vengono usati successivamente per stimare nuovamente le frequenze degli aplotipi, attraverso lo step di massimizzazione, il che porta a calcolare Θ_{H1}. L'algoritmo consiste nell'iterare i due passi (l'*expectation* e il *maximization*) fino alla convergenza, cioè finché la differenza tra Θ_{Ht} e Θ_{Ht+1} non sia più piccola di un valore predefinito. Ad ogni iterazione, il valore di Θ_{Ht} viene migliorato massimizzando la funzione di somiglianza dell'insieme G. Diversi autori riportano il passo di *Expectation* dell'algoritmo mediante la seguente formulazione:

$$E\{\delta_h(H_i)|G_i)\} = \frac{\sum_{H-Gi} \delta_h(H) p_{h1} p_{h2}}{\sum_{H-Gi} p_{h1} p_{h2}}$$

Dove $\delta_h(H)$ = 0,1, o 2 è il "dosaggio aplotipico" ossia il conteggio del numero di copie di h contenute nel vero (ma in generale non nota) coppia di aplotipi H portata dall'individuo i-esimo. La stima di $\delta_h(H)$ è calcolata condizionata ai dati di genotipo G per ogni soggetto e considerando il set delle frequenze p come se fossero note. La sommatoria $\sum_{H-Gi}$ indica la sommatoria sulle coppie di aplotipi ordinate H=(h_1, h_2), con frequenze p_{h1} e p_{h2} rispettivamente, che sono compatibili con i dati di genotipo osservato. Poiché si assume l'equilibrio Hardy-Weinberg, la coppia H ha probabilità pari al prodotto p_{h1} p_{h2}. Questo algoritmo risulta di utile applicazione nei modelli con *variabili nascoste*: tipici esempi di variabili nascoste sono i dati mancanti o non osservabili. Si procede, infatti, andando ad assegnare un valore atteso alle variabili nascoste come se queste fossero note. La complessità temporale dell'algoritmo EM è $O(m2^k)$ dove m è il numero di genotipi e k è il numero massimo di SNPs eterozigoti nei genotipi. Il limite maggiore dell'algoritmo risiede nell'incremento esponenziale del numero di aplotipi all'aumentare del numero di SNPs eterozigoti. La soluzione più comune al problema è la divisione dell'intero set di SNPs in sottoinsiemi più piccoli e contigui, definiti anche pseudo-blocchi, procedendo poi alla combinazione degli aplotipi selezionati dai singoli sottoinsiemi attraverso un approccio bottom-up. Poiché l'algoritmo si basa sul'assunzione dell'equilibrio H-W, e dato che questa condizione è quasi sempre soddisfatta se si considerano grandi gruppi di individui, allora l'algoritmo può essere applicato con risultati soddisfacenti su campioni di dati molto grandi. Studi di simulazione hanno comunque dimostrato la validità dell'algoritmo anche quando l'assunzione dell'HWE è violata. Sono state inoltre valutate le performance dell'algoritmo quando si è in presenza di errori di sequenziamento e/o dati mancanti: se si è in condizioni di moderato o forte LD tra gli SNPs, l'assenza fino al 30% di dati non sembra influire sull'accuratezza del risultato. Tuttavia, se si è in condizioni di basso LD, gli errori di sequenziamento influiscono pesantemente sul risultato diminuendone l'accuratezza, pertanto in tale situazione è preferibile considerare i dati incerti come non noti. Come già anticipato, l'algoritmo EM è diffusamente impiegato e si può trovare implementato anche nel software *Haploview* (2003-2006 Broad Institute of MIT and Harvard).

Una seconda formulazione è la *Bayesian Frequencies Haplotype Inference problem*: come la *Maximum Likelihood*, si propende anche in questo caso per un approccio di tipo statistico. Mentre nel Maximum Likelihood si andava a trovare un set di parametri del modello che andassero a massimizzare la probabilità dei dati di genotipo *G* dato il modello, l'approccio Bayesiano va a calcolare la distribuzione a posteriori dei parametri del modello dati i dati di genotipo *G*, ossia $P(\Theta|G)$. In input sono passati un insieme *G* di genotipi e una distribuzione a priori delle frequenze dei genotipi; in uscita è riportata la distribuzione a posteriori delle frequenze degli aplotipi dato *G*. Rispetto ai modelli ML, questi richiedono più tempo computazionale ed è più difficile raggiungere la convergenza del risultato.

Gli approcci statistici sono i metodi più popolari e più largamente implementati. L'accuratezza di questi metodi è infatti in qualche modo migliore di quella risultante dall'applicazione delle altre soluzioni. Inoltre, i metodi di parsimonia e i modelli coalescenti spesso presentano soluzioni multiple, rendendo difficile il confronto delle loro performance con altri metodi. Per ultimo, i metodi statistici risultano sicuramente più robusti perché applicabili anche nel caso di genotipi mancanti o ambigui. Questa condizione non è infrequente, in quanto dovuta ad errori di sequenziamento o alleli mancanti. La mancanza dell'informazione sull'allele aumenta notevolmente la complessità computazionale dell'algoritmo, mentre l'errore di sequenziamento rende più difficile la risoluzione del problema, dal momento che non si conosce quale allele sia sbagliato. È da sottolineare inoltre che, nonostante i risultati di questi approcci siano molto promettenti, possono comunque cadere in difetto quando i livelli di LD decrescono. Questa scarsa accuratezza si presenta soprattutto quando consideriamo un numero consistente di SNPs, dal momento che il LD tende a decrescere con l'aumentare della distanza tra SNPs. Scarsa accuratezza si registra anche in presenza di rari aplotipi, in quanto il razionale alla base di molti algoritmi è la condivisione del maggior numero di aplotipi possibile nella popolazione. In conclusione, alla luce delle questioni illustrate, la ricerca sugli algoritmi di ricostruzione di LD e aplotipi dovrebbe concentrarsi sul miglioramento delle performance di questi in dataset con basso LD, in cui sono presenti errori di sequenziamento, alleli mancanti e aplotipi rari.

2.4 Tag SNP

In regioni ad alto LD, si può selezionare un ridotto set di SNP per identificare efficientemente gli aplotipi comuni. Un ridotto sottoinsieme di SNP è preferibile sia per ridurre le esigenze di sequenziamento del genoma, sia per eliminare la ridondanza di informazione data dal considerare gli SNPs nella loro totalità. Quindi, per ovviare alla grande quantità di dati che è necessario per risolvere problemi legati al DNA, molti scienziati si sono dedicati alla ricerca di sottoinsiemi minimi di SNP che permettano comunque di rappresentare le associazioni tra malattie e geni. Questo processo viene chiamato *identificazione di tagSNP* e in generale gli SNP selezionati vengono indicati come *tagSNP* e quelli non selezionati come *taggedSNP*.

Per risolvere un problema di selezione di TagSNP è necessario trovare un sottoinsieme ottimale di SNPs, di dimensione minore rispetto l'originale, valutando come questo rappresenti bene i dati di genotipo rispetto a tutti gli altri possibili sottoinsiemi ottenibili. La motivazione della selezione di TagSNP ha origine, come già accennato sopra, nel linkage disequilibrium, concetto che peraltro è alla base dei test di associazione gene-malattia. Quando è presente un alto LD tra due SNPs, l'informazione portata dai loro alleli è pressoché identica. Quindi, possiamo selezionare uno tra gli SNPs ridondanti in modo che, anche un sottoinsieme dell'originale gruppo di SNPs, mantenga la stessa informazione. Quale sia la migliore strategia che permetta la selezione del sottoinsieme ottimale, rimane ancora un problema aperto. Vengono di seguito proposte vari approcci possibili al problema finora sviluppati.

Come già osservato nel capitolo introduttivo, la struttura a blocchi del genoma umano dimostra che il genoma può essere partizionato in blocchi discreti tali che, all'interno di ciascuno, la maggior parte della popolazione condivide un piccolo numero di aplotipi comuni (all'incirca 3.5). Basandosi su questa assunzione, i primi algoritmi sviluppati miravano a trovare il sottoinsieme di SNPs che meglio catturava la maggior parte della limitata diversità tra aplotipi dai dati originali. Questi metodi sono definiti *Haplotype Diversity-based Methods*. Per definire e quantificare la diversità tra gli aplotipi, sono state proposte diverse misure. Alcune usano, come misura di diversità, il numero di aplotipi che sono distinguibili in modo univoco dal sottoinsieme *T'* di TagSNPs. Si sceglie quindi il sottoinsieme con la misura di diversità più alta. Un'altra soluzione è definire la diversità *non* catturata dal sottoinsieme *T'* di tagSNPs come il numero di alleli diversi tra ogni coppia di aplotipi nello stesso gruppo basato su *T'*. Se il

sottoinsieme *T'* partiziona correttamente tutti i distinti aplotipi in gruppi diversi, allora la diversità aplotipica residua sarà nulla. Si andrà a selezionare in questo caso il sottoinsieme *T'* che mostra diversità residua più piccola o nulla. Un'altra misura popolare della diversità aplotipica è basata sull'entropia di Shannon. Sia *n'* il numero di aplotipi distinti nel dataset *D* di aplotipi, e p_i siano le frequenze relative dell'*i*-esimo aplotipo distinto. La diversità di *D* può essere calcolata come entropia *H*:

$$H(D) = -\sum_{i=1}^{n'} p_i \log_2 p_i$$

Come negli altri metodi, gli aplotipi sono partizionati in gruppi in modo che quelli nello stesso gruppo condividano gli stessi alleli. L'entropia del dataset *D* è misurata basandosi su questa partizione. Gli aplotipi che sono posizionati nello stesso gruppo sono considerati identici. Maggiore è il numero dei sottoinsiemi candidato *T'*, maggiore sarà l'entropia del dataset basato su quella partizione. Il sottoinsieme con la misura di entropia più alta sarà selezionato come soluzione. I metodi finora introdotti esaminano esaustivamente tutti i sottoinsiemi del set originale di SNP, limitando la loro applicazione solo a piccoli insiemi di SNPs. Per ovviare a questa limitazione sono state proposte diverse soluzioni di tipo euristico o algoritmi greedy. I metodi basati sulla misura della diversità sono intuitivi e diretti. Tuttavia, per assicurarsi che la diversità tra aplotipi sia effettivamente limitata e per velocizzare la performance dell'algoritmo, si esegue una divisione in blocchi adiacenti della regione in esame, e la selezione dei TagSNP viene fatta blocco per blocco. Una limitazione a questo modo di operare in blocchi sta nella possibilità che l'unione dei set di TagSNP definiti sui blocchi possa non essere il sottoinsieme ottimale di TagSNP per l'intera regione. Per di più, tra i blocchi possono esistere, con alta probabilità, regioni a basso LD che vanno ad aumentare il numero di aplotipi diversi rendendo inapplicabili tali metodi. Risulta evidente quindi che, con questo approccio, il particolare partizionamento della regione target influenza pesantemente la selezione dei TagSNP.

E' possibile effettuare la scelta dei TagSNP affidandosi anche all'algoritmo EM descritto nei paragrafi precedenti in merito all'inferenza degli aplotipi. Infatti la stessa procedura può essere applicata alle situazioni in cui dobbiamo scegliere il particolare set di TagSNP. Si utilizza nuovamente lo stimatore, $E\{\delta_h(H)|G_k\}$, del numero di copie $\delta_h(H)$ con l'osservato genotipo *G*, assumendo di nuovo che le frequenze aplotipiche

siano note e valido l'equilibrio Hardy-Weinberg. Si procede quindi con il calcolo della correlazione R^2_h tra $E\{\delta_h(H)|G_k\}$ e $\delta_h(H)$ come :

$$R_h^2 = \frac{Var[E\{\delta_h(H)|G\}]}{Var\delta_h(H)} = \frac{Var[E\{\delta_h(H)|G\}]}{2p_h(1-p_h)}$$

Dove p_h è la frequenza dell'alotipo h, e la varianza dell'aspettazione a nominatore è calcolata come sommatoria su tutti i possibili valori del genotipo G come:

$$Var[E\{\delta_k(H)|G\}] = \sum_G E\{\delta_k(H)|G\}^2 p(G) - 4p_k^2$$

Per ogni sottoinsieme di SNPs si può formalmente calcolare R^2_h usando la formula riportata. Per il ridotto set di SNP ci saranno più coppie aplotipiche che saranno compatibili per il genotipo basato solamente sul ridotto set di SNP piuttosto che sul genotipo basato sul data set con tutti gli SNPs. Ne consegue una varianza minore e quindi un minor valore per R^2_h. Il calcolo dei TagSNP prevede innanzitutto l'applicazione dell'algoritmo EM per identificare gli aplotipi comuni all'interno di regioni ad alto linkage disequilibrium. Per un blocco comprendente n SNPs ad alto LD, si definisce il miglior set di m TagSNP ($m < n$) come quegli SNPs che massimizzano il minimo valore di R^2_k calcolato per ogni aplotipo comune. Il calcolo di R^2_k per ogni aplotipo richiede la generazione dell'intero set di copie di aplotipi, H, per un dato set di frequenze aplotipiche non nulle e un valore di $\delta_h(H)$ compatibile con ognuno dei possibili SNP sequenziati. Questo deve quindi essere fatto per ogni aplotipo comune h, e per ogni set candidato di TagSNP. In dipendenza del numero di SNPs, il procedimento può risultare piuttosto oneroso in molti casi. Per ottimizzare la scelta, invece di optare per una ricerca esaustiva di tutte le

$$\frac{n!}{(n-m)!\,m!}$$

scelte dei m TagSNPs è stato implementato un metodo di inclusione passo per passo: si sceglie come TagSNP candidato, da un set di k SNP, quello che contribuisce al maggior incremento del valore di R^2_h calcolato sui rimanenti k-1 SNPs. Il processo viene ripetuto per vedere se R^2_h può essere aumentato ulteriormente per sostituzione con ognuno degli altri k-1 SNP con ogni altri SNP non selezionato come Tag.

Una seconda categoria di metodi proposti sono i Pairwise Association-based Methods. Questo approccio si basa sull'idea che un set di TagSNP dovrebbe essere il più piccolo sottoinsieme di SNPs disponibili in grado di predire il locus malattia di un aplotipo. Il locus malattia è l'obiettivo degli studi di associazione, quindi generalmente non è noto a priori e si deve procedere con una stima. Il set di tagSNP è selezionato in modo tale che tutti gli SNPs di un aplotipo siano altamente associati con uno dei tagSNP selezionati. In questo modo, anche se uno SNP che svolge un ruolo rilevante in una malattia non viene selezionato come tag, la sua associazione alla malattia può essere indirettamente dedotta dai tagSNP, in quanto sono scelti secondo un criterio di forte associazione con esso. Nella maggior parte degli studi, il LD è utilizzato come misura per la stima dell'associazione pairwise. Sono stati proposti due algoritmi per la soluzione del problema. Il primo ricorre a una partizione del set originale di SNPs mediante clustering gerarchico, dove gli SNPs all'interno dello stesso cluster hanno un valore di LD con almeno uno degli altri SNPs maggiore di un livello scelto a priori. Inizialmente, ogni SNP costituisce un cluster a sé stante. Le fusioni tra i cluster C_i e C_j si basano sulla seguente definizione di distanza:

$$D(C_i, C_j) = \min_{\forall s \in (C_i \cup C_j)} (D_{max}(s))$$

dove $D_{max}(s)$ è la massima distanza, definita dal valore di LD (ad esempio r^2), tra lo SNP s e tutti gli altri SNPs nei due cluster. I due cluster più vicini, basati sulla distanza $D(C_i,C_j)$, sono quindi uniti iterativamente. L'unione dei cluster si ferma quando la più piccola distanza tra due cluster è più grande di una certa soglia. Lo SNP s che definisce la distanza di ogni cluster unito, è scelto come rappresentativo del cluster. Generalmente, viene scelto uno SNP da ogni cluster basandosi su considerazioni di tipo pratico come la facilità di sequenziamento, l'importanza della collocazione del locus o significatività portata dalla mutazione dello SNP. Il secondo algoritmo proposto è un algoritmo di *greedy*: questo innanzitutto esamina tutte le relazioni di LD tra tutte le coppie di SNPs, e per ogni SNP conta il numero di SNPs con i quali ha un valore di LD maggiore di una certa soglia. Lo SNP con il maggior numero di conteggi è clusterizzato insieme agli SNPs associati ad esso e diventa il tagSNP del cluster. Questa procedura è iterata con i rimanenti SNPs fintanto che tutti gli SNPs non sono stati clusterizzati. Gli SNP che invece presentano livello di LD minore della soglia fissata, sono clusterizzati singolarmente. Alternativamente, è stata proposta per i metodi di *Pairwise Association*

la scelta di tagSNP tali che il loro valore di LD sia come prima maggiore di un certo livello fissato, ma in questo caso il valore di LD deve essere soddisfatto non solo per uno, ma per tutti gli SNP nel cluster. Anche per questo criterio sono applicabili i due algoritmi appena descritti. Tutti i metodi di associazione pairwise hanno complessità $O(cnm^2)$, dove il numero di cluster è c, il numero di aplotipi è n, e il numero di SNPs è m. Quindi, in generale, sono veloci dal punto di vista computazionale rispetto ai metodi basati sulla definizione di distanza, e non richiedono una partizione in blocchi della regione target. Il principale svantaggio è rappresentato dal fatto che non riescono a identificare dipendenze multiple tra più di due SNPs; tendono inoltre a selezionare più tagSNP rispetto agli altri metodi.

Una possibile soluzione agli aspetti negativi presentati per i metodi Pairwise è rappresentata dai metodi *Tagged SNP Prediction-based Methods*. Questi metodi considerano la selezione di TagSNP come un problema di ricostruzione dell'aplotipo originale usando solo un sottoinsieme di SNPs. Il loro obiettivo è quindi selezionare un set di SNPs che possano predire gli SNPs non selezionati (i tagged) con il minimo errore. È stato inizialmente proposto di selezionare i tagSNP sulla base della loro accuratezza nel predire i tagged. Sono state quindi definite misure di *informatività* per trovare il sottoinsieme ottimale di SNP, utilizzate poi in programmazione dinamica. A differenza dei metodi pairwise, questo approccio permette di identificare associazioni multiple tra SNP e il numero di TagSNP è generalmente minore rispetto a quelli selezionati dai precedenti. Inoltre, il ricorso a tecniche di programmazione dinamica garantisce di trovare una soluzione che sia un ottimo globale. Tuttavia, i maggiori limiti di queste tecniche sono il tempo computazionale di ordine esponenziale e il confinamento che viene imposto alla regione di ricerca dei tagSNP nell'intorno di un locus fisico.

L'ultima categoria è rappresentata dai metodi *Phenotype Association-based Methods*. Questo approccio assume la disponibilità dell'informazione del fenotipo e, sulla base di questa, cercano di trovare un set di SNPs che possa distinguere gli individui portatori di malattia (i casi) dagli individui sani (controlli). Sotto questa prospettiva, la selezione dei tagSNP è una sorta di feature selection, che mira a selezionare un set di features che distingue tra due classi (casi/controlli) con un piccolo errore. Uno dei classificatori più implementati è il classificatore Bayesiano. Assume che l'allele di uno SNP è indipendente da quello degli altri fenotipi dati, e classifica ogni aplotipo come caso o

controllo basandosi sulla sua probabilità di appartenere a una di queste classi. Il sottoinsieme *T'* di SNPs con la miglior accuratezza di classificazione è selezionato come set di tagSNP. La limitazione più forte sta nell'assunzione di indipendenza tra SNP alla base del classificatore. Nel caso reale, infatti, tra gli SNPs esiste un'associazione non casuale, data dal linkage disequilibrium. È stato successivamente proposto un altro metodo che non solo classifica correttamente i dati, ma garantisce anche che le sue performance siano statisticamente significative. Questo algoritmo è basato su una tecnica di bootstrap nel quale si eseguono circa 1000 permutazioni su *n* coppie di aplotipo-fenotipo che costituiscono il dataset originale. Questi metodi sono direttamente correlati all'obiettivo principale dell'analisi di aplotipi e dei testi di associazione gene-malattia. la principale limitazione di questo approccio consiste nel la necessità dell'informazione sull'aplotipo, che potrebbe non essere disponibile in anticipo. Inoltre, generalmente, il numero di aplotipi usati per la selezione di tagSNP è relativamente piccolo, quindi i tag selezionati che classificano un campione ridotto molto bene, potrebbero non rappresentare altrettanto bene un campione più grande. Questo ha conseguenze che possono condizionare i test di associazione gene-malattia.

La praticità della selezione dei TagSNP è stata empiricamente dimostrata in numerosi studi di simulazione. I risultati dimostrano che lavorare con i tagSNP porta a una perdita minima della potenza dei test di associazione (dove con potenza dei test di associazione intendiamo a probabilità che il test rifiuti l'ipotesi nulla). Tuttavia, rimangono numerosi problemi aperti:

1. È necessario assicurarsi della qualità dei database di SNP e la loro applicabilità a una popolazione più ampia di quella campionata.
2. La maggior parte degli algoritmi di selezione dei tagSNP si concentrano su aplotipi comuni o SNPs comuni piuttosto che quelli rari. Le variazioni comuni sono sicuramente di grande interesse in quanto molti disturbi e malattie comuni possono essere spiegate proprio da variazioni comuni di DNA piuttosto che da quelle rare. Tuttavia, è ancora aperta la questione su quali variazioni, se quelle comuni o quelle rare, influenzino la predisposizione a malattie comuni e complesse.
3. Diversi algoritmi richiedono dati di aplotipo piuttosto che di genotipo. Quando sono disponibili solo dati di genotipo, deve essere calcolato l'aplotipo da questi dati e lavorare su questa informazione. Si sa però che molti algoritmi sviluppati

per il calcolo degli aplotipi possono produrre più di una soluzione e dare quindi un certo grado di incertezza. Finora, nessun metodo di selezione di TagSNP considera l'incertezza derivante dalla ricostruzione di aplotipi.

4. Tutti gli algoritmi presentati assumono che il set di tagSNP selezionato da un campione possa dare buoni risultati anche per un altro campione proveniente dalla stessa popolazione. Per assicurarsi che le performance possano essere generalizzate, si dovrebbero campionare un numero sufficiente di individui ed evitare situazioni di overfitting.
5. È importante definire i limiti o confini dei blocchi di LD all'interno dei quali concentrare lo spazio di lavoro per la selezione dei tagSNP. Sebbene l'idea di trovare TagSNP non dipenda in sé dall'esistenza o dalla qualità dei blocchi di LD, il pattern di LD influenza la proprietà dei TagSNP, incluso il vantaggio che può derivare dal loro uso e il grado di trasferibilità su un campione di popolazione più ampio.
6. I metodi presentati non tengono in alcun modo conto del tasso di ricombinazione e degli aspetti demografici (ad esempio collo di bottiglia) che possono modificare la struttura dei pattern di LD.

Capitolo 3

Stratificazione di popolazione

Le recenti proposte tra le tecnologie del sequenziamento e l'aumentata disponibilità di marker genici hanno aperto la strada agli studi di associazione genome wide su larga scala [8]. Un potenziale problema che nasce da ogni studio basato su popolazione è la presenza non identificata di una stratificazione di popolazione che può simulare un segnale di associazione spurio e di conseguenza portare a falsi positivi e/o alla non individuazione dei reali effetti [5,8,9,10]. Quando casi e controlli hanno frequenze alleliche diverse attribuibili a diversi background di popolazione, e non ad eventuali espressioni fenotipiche, allora si dice che lo studio è caratterizzato da stratificazione di popolazione.

La stratificazione di popolazione è probabilmente la ragione più spesso citata della mancata riproducibilità dei risultati dei test di associazione.

3.1 Cause della stratificazione

La stratificazione può essere dovuta principalmente a tre cause che sono: la presenza di una struttura (cioè di sottogruppi) all'interno di una popolazione (*population structure*, Figura 3.1), presenza di legami famigliari tra i soggetti campionati (*family structure*), e presenza di correlazioni, anche lontane, tra i soggetti con legami famigliari di grado superiore al primo (*cryptic relatedness*) [9].

Tipicamente, gli studi di associazione genome-wide evitano di campionare, all'interno dello stesso studio, individui provenienti da diverse popolazioni con l'intenzione di evitare la stratificazione di popolazione. In generale, per ogni studio caso-controllo, il pool di popolazione da cui i casi vengono campionati dovrebbe essere lo stesso dal quale anche i controlli vengono campionati.

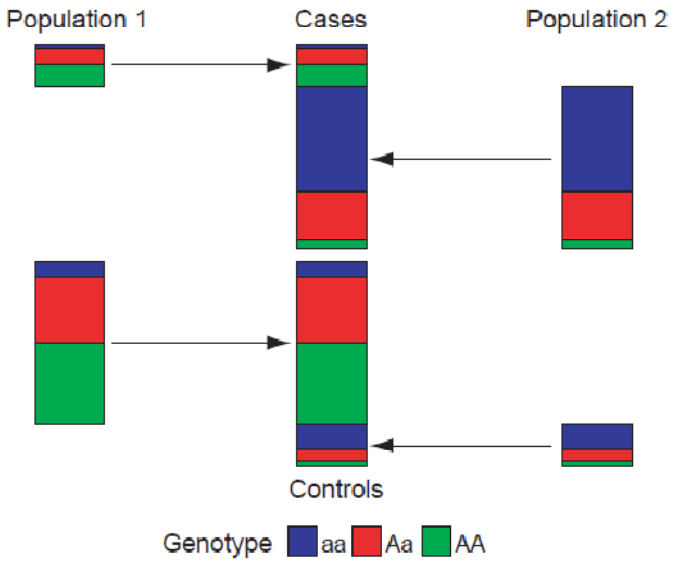


Figura 3.1: Effetti della struttura di popolazione su un locus di uno SNP. Se lo studio di popolazione coinvolge sottopopolazioni che differiscono geneticamente, e se anche la prevalenza della malattia differisce tra queste sottopopolazioni, allora le proporzioni di casi e controlli campionati da ogni sottopopolazione tenderà ad essere diversa, dal momento che le frequenze alleliche o genotipiche tra casi e controlli ad ogni locus saranno divere per ogni sottopopolazione. La Figura mostra un esempio di questo scenario con due popolazioni dove i casi sono costituiti da un eccesso di individui provenienti dalla popolazione 2 e la popolazione 2 ha una frequenza minore dell'allele A rispetto a quella rilevata nella popolazione 1. In questo esempio, la struttura simula il segnale di associazione data dalla significativa differenza nelle frequenze alleliche e genotipiche tra casi e controlli causata non dal fenotipo, ma dalle caratteristiche della popolazione stessa.

3.2 Identificazione della stratificazione

I metodi, di seguito brevemente esposti, propongono una possibile soluzione al problema di stratificazione. Questi richiedono un numero sufficiente di SNPs (preferibilmente maggiore di 100) che siano stati sequenziati nei casi e nei controlli in aggiunta agli SNPs candidati nel test di associazione [5].

3.2.1 Genomic control (GC)

Il metodo più comunemente utilizzato, soprattutto nell'identificazione della presenza di stratificazione, è il Genomic Control (GC). È basato sull'osservazione che la population

structure cambia la distribuzione nulla della statistica χ^2 di un fattore moltiplicativo λ, che viene calcolato basandosi su una collezione di marker come il rapporto:

$$\lambda = \frac{median(S_i)}{0.456}$$

a numeratore viene calcolata la mediana della statistica $S^2 = \lambda \chi^2$, ed è divisa per la media teorica calcolata in ipotesi nulla. In assenza di stratificazione, l'associazione tra le varianti genetiche e la malattia dovrebbe seguire una distribuzione χ^2 a 1 grado di libertà (1 df), pertanto il valore di λ risulta circa uguale a 1. In presenza di stratificazione, il valore di λ è maggiore di 1, condizione che si verifica in quanto che la stratificazione tende a incrementare la statistica di un fattore costante. Generalmente, valori di λ minori di 1.05 vengono ancora tollerati come non critici, e si può considerare pressochè assente la stratificazione [9]. Il valore di λ varia al variare della dimensione del campione, sia in funzione del numero di loci considerati nel calcolo del GC che del numero di individui campionati [8,9,10,11]. La dipendenza di λ dalle dimensioni del campione (intese come numero di individui) è mostrata in Figura 3.2.

Risulta evidente che, a parità di condizioni di stratificazione, all'aumentare delle dimensioni del campione, aumenta anche il valore di λ. Dalla Figura si nota anche che il valore di λ risulta unitario in assenza di stratificazione e tende invece ad aumentare man mano la presenza di stratificazione diventa più marcata, in accordo con quanto detto prima. Infine, quando per identificare la stratificazione vengono usati solo un piccolo numero di loci (da 50 a 100) allora il test GC è spesso non conservativo, il p-value calcolato è più grande di quello dato dalla distribuzione χ^2 e ancora risultano falsi positivi. E' stato dimostrato che il GC non corregge adeguatamente la stratificazione di popolazione se per stimare il fattore di correzione sono utilizzati troppo pochi loci (quindi pochi marker). Al contrario, quando il numero di loci è molto grande, il test è molto conservativo e tende ad eliminare i falsi positivi dal risultato, ma ne consegue una perdita di potere statistico in alcune applicazioni [8]. La scelta del numero di SNPs necessari per la corretta soluzione al problema di stratificazione dovrebbe tenere conto anche dell'intensità degli effetti genetici causati dal marker in studio.

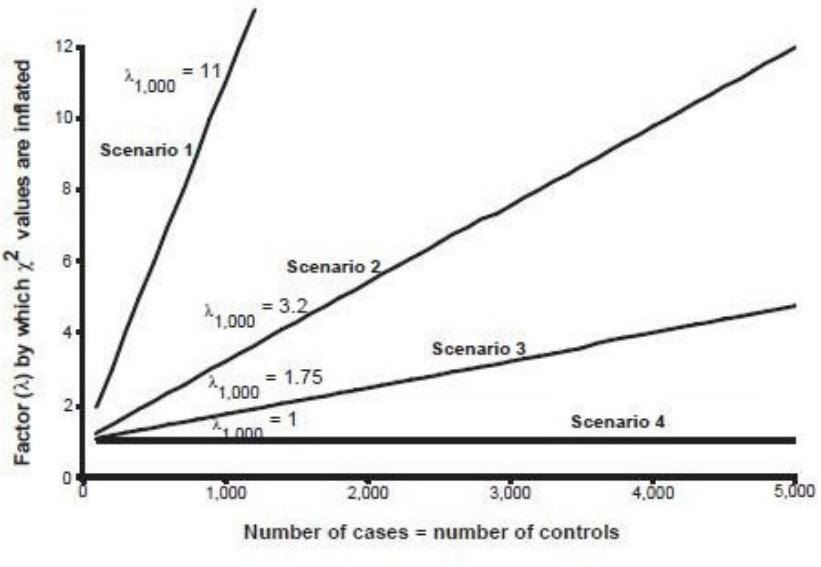


Figura 3.2: effetti della stratificazione sugli studi di associazione. La stratificazione influisce sulla statistica χ^2 di un fattore λ, che cambia in dipendenza delle dimensioni del campione. Lo scenario 1 corrisponde a una condizione di evidente stratificazione, lo scenario 4 corrisponde a assenza di stratificazione, gli scenari 2 e 3 corrispondono a condizioni intermedie, dove la stratificazione è rispettivamente più e meno marcata. In ordinata si ha il valore di λ, mentre in ascissa si ha la dimensione del campione per uno studio in cui si suppone che il numero dei casi sia uguale al numero dei controlli. λ_{1000} fa riferimento alla situazione in cui il numero di casi e controlli è uguale a 1000.

3.2.2 Principal component analysis (PCA)

La Principal Component Analisys (PCA) rappresenta un approccio multivariato che permette di rilevare differenze in termini di strutture genetiche tra sottogruppi di individui. Se applicata a dataset con dati campionati da individui provenienti da diverse sottopopolazioni, allora l'analisi è in grado di dare una interpretazione geografica ai dati, individuando graficamente in modo distinto i gruppi relativi alle sottopopolazioni. I soggetti sono caratterizzati in uno spazio vettoriale come punti distinti, in modo tale che gli individui geneticamente simili formino "nuvole" compatte (cluster), mentre gli individui geneticamente più dissimili si posizionino in regioni più distanti dello spazio (Figura x.3). La presenza di cluster può aiutare a individuare sottopopolazioni di individui omogenei tra loro che possono essere considerati separatamente nelle analisi di associazione aumentandone il potere statistico.

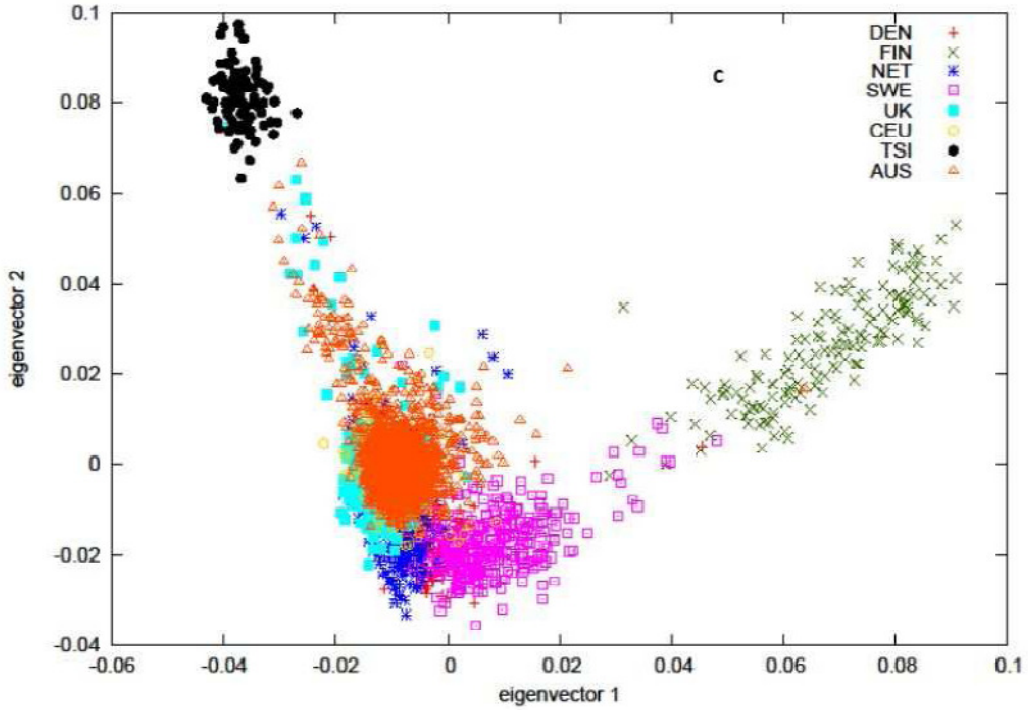


Figura 3.3: rappresentazione delle prime due componenti principali dell'analisi PCA (rispettivamente in ascissa e ordinata) che mettono a confronto le caratteristiche di background genetico dei diversi soggetti con diversi progenitori. Ogni punto rappresenta un individuo analizzato, mentre il colore fa riferimento al diverso gruppo etnico.

Il metodo della PCA utilizza le componenti principali come covariate per correggere la stratificazione negli studi GWA: un modello semplice lineare rappresenta il fenotipo Y come funzione degli effetti fissi X:

$$Y = XB + \varepsilon$$

Qui, X=[G_i PC_{1i} …PC_{mi}] indica il genotipo del marker candidato (G_i) in aggiunta a covariate opzionali, come ad esempio il sesso o l'età, B denota invece i coefficienti degli effetti fissi e ε è il termine di rumore che tiene conto della variazione di Y non spiegata. La PCA applica una correzione più grande ai marker che presentano forti differenze alleliche. A differenza delle implementazioni iniziali, la PCA è diventata un'analisi computazionalmente trattabile anche su grandi dataset. Approcci correlati a questo, come il multidimensional scaling (MDS) e il matching genetico sono stati implementati da diversi software come PLINK [riferimenti]. Una limitazione del metodo è l'incapacità di tener conto nel modello della presenza di family structure e cryptic relatedness [9].

3.2.3 Mixed models

I modelli misti (mixed models) possono modellare tutte e tre le cause della stratificazione di popolazione, cioè population structure, family structure e cryptic relatedness. L'approccio alla base è la modellizzazione dei fenotipi (*Y*) considerando contemporaneamente effetti fissi (*X*) e effetti random (*u*). Il modello lineare è dato dal

$$Y = XB + u + \varepsilon$$

Gli effetti fissi comprendono lo SNP candidato e covariate opzionali (*B*), come sesso o età, mentre gli effetti random sono basati su una matrice delle covarianze fenotipiche, che a sua volta è modellata come la somma di variazione random ereditabile e non ereditabile. Nel modello, *u* denota una componente della varianza del rumore $u+\varepsilon$ che è distribuita secondo la matrice *K* di Kinship.

$$Var(u) = \sigma^2 K$$

Qui, *u* rappresenta la componente ereditabile della variazione random mentre ε rappresenta la componente non ereditabile. La matrice di Kinship *K* è definita secondo la similarità genotipica tra coppie di individui, quindi la sua struttura è sicuramente influenzata dalla struttura di popolazione, family structure e cryptic relatedness. Il parametro σ^2 esprime quanto gli individui, geneticamente simili, sono anche fenotipicamente simili. Più sono simili i due individui, più è alto il valore della correlazione.

Capitolo 4

Il progetto internazionale HapMap

4.1 Cos'è HapMap

HapMap (Haplotype - Map) è un database, disponibile gratuitamente in internet, catalogo delle variazioni genetiche comuni che si osservano negli esseri umani e ne descrive i pattern comuni . Il progetto ha permesso l'emergere e lo sviluppo degli studi di associazione *genome wide*, e si ritiene possa essere una risorsa chiave per i ricercatori nel tentativo di identificare geni che possono avere ripercussioni sulla salute, malattie, risposte ai farmaci e fattori ambientali. Il progetto nasce da una collaborazione cominciata nell'Ottobre del 2002 tra diversi paesi quali: Giappone, Inghilterra, Canada, Cina Nigeria, e Stati Uniti. L'obiettivo che i ricercatori si sono proposti con questo progetto è determinare i pattern comuni nelle variazioni della sequenza di DNA nel genoma umano attraverso la caratterizzazione di queste stesse varianti (individuate negli SNPs), la stima della frequenza con cui compaiono e il grado di correlazione tra queste, utilizzando campioni di DNA provenienti da popolazioni di discendenza africana, asiatica ed europea.

Malattie comuni come le patologie cardiovascolari, cancro, obesità, diabete, malattie psichiatriche o infiammatorie sono causate dall'azione combinata di fattori genetici e ambientali. Le cause di queste malattie non sono imputabili all'azione di un unico gene, bensì viene riconosciuta una predisposizione genetica, a carico degli aplotipi, che si combina poi con fattori ambientali. Le variazioni del genoma possono quindi servire da marker genici per determinare l'associazione tra una particolare regione del genoma e la malattia. Un approccio di questo tipo può essere velocemente applicato grazie ad HapMap a qualsiasi gene candidato nel genoma, o a qualsiasi regione che può essere individuata da studi di linkage familiare, o, in ultimo, all'intero genoma per l'individuazione di fattori di rischio. HapMap propone una importante scorciatoia

nell'individuazione dei geni candidati e negli studi di linkage e di associazione basati su l'intero genoma. Nel suo scopo, l'International HapMap Project ha molto in comune con il *Human Genome Project*, il quale sequenzia tutto il genoma umano. Ma mentre quest'ultimo progetto si occupa di sequenziale l'intero genoma, incluso il 99.9% di genoma che tutti abbiamo in comune, il progetto HapMap si occupa di caratterizzare i pattern comuni all'interno del 0.1% che ci differenzia l'uno dall'altro.

4.2 Realizzazione del progetto HapMap

Il progetto è di fatto divenuto possibile grazie al confluire delle seguenti conoscenze: la disponibilità della sequenza del genoma umano (si è fatto riferimento a diversi database già esistenti, come ENCODE), un database di SNPs comuni (dbSNP, successivamente arricchito dal progetto stesso), l'intuizione sul linkage disequilibrium (LD) e lo sviluppo di tecnologie accurate e relativamente a basso costo per il sequenziamento high-throughtput degli SNPs. Il lavoro ha avuto effettivamente inizio nel 2002 con la realizzazione di una prima fase che ha interessato 269 campioni di DNA prelevati da individui provenienti da Nigeria, Utah (USA), Cina e Giappone, dai quali sono stati sequenziati 1.007.329 SNPs. La seconda fase ha consentito di sequenziale complessivamente 4.6 milioni di SNPs in ognuno dei campioni HapMap. Attualmente è in corso una terza fase che coinvolge un numero di campioni ancora maggiore, proveniente da diversi gruppi etnici (inclusi paesi come il Messico, altri stati degli USA, il Kenya e anche l'Italia). L'elenco completo delle popolazioni coinvolte nel progetto è riportato in Tabella 4.1. I campioni hanno solo un'etichetta che ne identifica la provenienza etnica e il sesso del donatore; non è possibile in nessun caso risalire, attraverso le informazioni disponibili, all'individuo. Le varie fasi si differenziano non solo per la quantità di campioni e donatori coinvolti nel progetto, ma anche nella densità con cui vengono sequenziati gli SNP, la stima della minor allele frequency (MAF), e i pattern di linkage disequilibrium (LD). Infatti nella fase due, gli SNPs sono sequenziati più densamente e sono stati inclusi anche quelli che presentano in media una MAF minore rispetto a quella della fase 1: questo ha contribuito a un significativo miglioramento nella rappresentazione di variazioni rare rispetto a quanto fatto inizialmente. Le fasi del progetto e le relative caratteristiche vengono riassunte schematicamente in Tabella 4.2 .

Etichetta	Campione di popolazione	# campioni	QC campioni
ASW (A)*	America (USA) del sud ovest con antenati Africani	90	71
CEU (C)*	Residenti in Utah (USA) con antenati dal nord e ovest europa	180	162
CHB (H)	pop. Han di Pechino, Cina;	90	82
CHD (D)	cinesi in Metropolitan Denver, Colorado	100	70
GIH (G)	indiani Gujarati residenti a Houston, Texas	100	83
JPT (J)	giapponesi di Tokyo, Japan	91	82
LWK (L)	Luhya in Webuye, Kenya	100	83
MEX(M)*	Pop. con antenati messicani a Los Angeles (Ca)	90	71
MKK(K)*	Maasai in Kinyawa, Kenya	180	171
TSI (T)	Toscani, Italia	100	77
YRI (Y)*	Yoruban in Ibadan, Nigeria	180	163
		1,301	1,115

Tabella 4.1: elenco delle popolazioni coinvolte nel progetto(aggiornato alla fase 3). La colonna Etichetta si riferisce all'abbreviazione utilizzata come riferimento alla popolazione. La presenza dell'asterisco indica che i campioni sono stati relativi a nuclei famigliari composti da 3 individui (madre, padre e figlio). La colonna QC riporta il numero di campioni che vengono mantenuti dopo il filtraggio Quality Control- Le cifre nell'ultima riga si riferiscono rispettivamente al totale dei campioni e dei campioni filtrati.

Fasi	Descrizione Fasi
Fase 1	Raccolta campioni da 4 popolazioni (CEU, YRI, CHB, JPT) per un totale di 269 individui; #SNPs: circa 1 milione, approssimativamente circa 1 ogni 5 kb, MAF>0.05
Fase 2	Raccolta dati intensificata nelle stesse popolazioni, #SNPs: vengono aggiunti più di 3.1 milioni di SNPs per un totale di 4.6 milioni, 1 ogni 1 kb approssimativamente, MAF>0.05
Fase 3	Raccolta dati nelle popolazioni che in questa fase arrivano a 11. Si contano più di 1.6 milioni di SNPs nuovi.

Tabella 4.2: in questo schema vengono riportate le fasi principali del progetto con una breve descrizione nella colonna a destra

Con l'avanzare del progetto non solo vengono aggiunti al database nuovi SNPs, ma vengono anche ridefiniti SNP la cui posizione e/o funzione prima risultava inesatta.

I gruppi di ricerca che si occupano del sequenziamento del DNA seguono un uguale e rigido protocollo per il controllo della qualità dei dati e per l'identificazione degli SNPs. Si stima che l'accuratezza per fase di sequenziamento sia in media del 99.5% per i vari gruppi; tuttavia c'è un alto tasso di dati mancanti e discrepanze nel sequenziamento stesso.

4.3 Pubblicazione e consultazione dei dati

Il progetto è impegnato in un rilascio rapido e completo dei dati, e si assicura che i risultati siano di dominio pubblico a nessun costo per l'utenza. Tutti i dati relativi ai nuovi SNPs, alle frequenze alleliche e genotipiche sono consultabili e scaricabili dal sito dell' HapMap Data Coordination Centre (http://www.hapmap.org).

La ricerca dei contributi genetici alle malattie umane generalmente si concentra su geni candidati identificati sulla base di studi di linkage e/o associazione, o sulla base di pathway che si presume siano coinvolti in un particolare aspetto della malattia in esame. Nello studio dei geni candidati, un ricercatore vorrà conoscere se ci sono SNPs comuni nelle immediate vicinanze, quali alleli hanno, e quali sono le relative frequenze alleliche nella popolazione. La sezione *Data* del sito consente un accesso interattivo al database attraverso un browser (*Generic Genome Browser*) che permette agli utenti di ricercare un particolare gene, o una regione di interesse di piccole o medie dimensioni, all'interno del genoma e quindi di visualizzare la distribuzione di SNPs e modelli di variazione comune nella regione stessa. La ricerca viene effettuata utilizzando il nome di una sequenza, di un gene, locus o altri punti di riferimento. I principali risultati su cui HapMap permette agevolmente di indagare sono:

- Identificazione SNPs e stima delle loro proprietà;
- Distribuzione della frequenza allelica dei campioni provenienti da un gruppo etnico;
- Frequenza allelica degli SNPs tra le varie popolazioni;
- Aplotipi condivisi dai diversi gruppi etnici;
- Valutazione del LD nel genoma umano e variazioni nel tasso di ricombinazione;
- Selezione di tagSNP per studi di associazione.

Viene di seguito presentato un esempio di ricerca che mostra un possibile utilizzo del Genomic Browser. Dopo l'accesso al sito www.hapmap.org e alla sezione Data, si entra nel browser vero e proprio. La query si effettua andando a digitare, nel campo *Cerca*, il termine selezionato. Può essere usato indifferentemente uno di questi termini:

- il nome del cromosoma (ad esempio "Chr10");
- la posizione in intervallo di basi sul cromosoma nel formato Cromosoma:start…stop (ad esempio "Chr10: 114,700,201…114,916,051");
- il nome dello SNP utilizzando l'identificativo "rs" (ad esempio "rs081062");
- il nome del gene secondo la nomenclatura del *NCBI RefSeq* (ad esempio "NM 153254");
- il nome comune del gene (ad esempio "BRCA2");
- la banda cromosomica (ad esempio "5q31").

Una volta eseguita la ricerca, verrà mostrata la pagina con i risultati. Se si hanno più corrispondenze alla query effettuata, allora la pagina mostrerà tutti i risultati specificandone la locazione sul genoma. Per default, il browser propone i risultati aggiornati alla data più recente; in alternativa è possibile selezionare l'origine dei dati, aggiornati in corrispondenza delle varie fasi del progetto. In particolare il menù consente di consultare i dati relativi alla fase 2, alla fase 3 o entrambe, con differenti date di rilascio dei dati stessi. Questa scelta non è banale, in quanto condiziona le opzioni di scaricamento e ricerca.

Si vuole indagare, a scopo illustrativo, sul fattore di trascrizione *Trascription Factor 7-like2* ,TCF7L2[3].

4.3.1 Esempio: Ricerca di TCF7L2

Nel caso in esame è sufficiente scrivere il nome TCF7L2; in alternativa si può indicare il numero del cromosoma e l'intervallo di basi che ne individuano la posizione. In questo esempio vengono consultati i dati relativi alla fase 3 aggiornati al febbraio 2009. Una volta inviata la richiesta, il DB individua automaticamente la posizione del fattore di trascrizione, riportando il numero del cromosoma in cui si trova e l'intervallo di basi

[3] TCF7L2 è un fattore di trascrizione comune noto anche come TCF4. Influenza la trascrizione di diversi geni, quindi svolge molteplici di funzioni all'interno della cellula. È anche implicato in diverse malattie; diversi SNPs sono associati in particolar modo al diabete di tipo 2. Nella popolazione europea è stato identificato come il principale fattore di rischio per questa patologia.

in cui il fattore è codificato. Il progetto permette di applicare metodi di analisi nuovi o già esistenti per l'analisi e la visualizzazione dei dati. Di seguito sono presentati i risultati grafici e sperimentali di tale ricerca.

Panoramica

Il DB indica che il fattore si trova sul cromosoma 10. La regione occupata dal fattore di trascrizione in esame è evidenziata in giallo in Figura 4.1 e interessa la posizione 114,700,201 - 114,916,051 individuabile sul righello che suddivide l'intero cromosoma in Mb. Il grafico gt'd SNPs/500Kb riporta il numero di SNPs sequenziati da HapMap ogni 500Kb sul cromosoma.

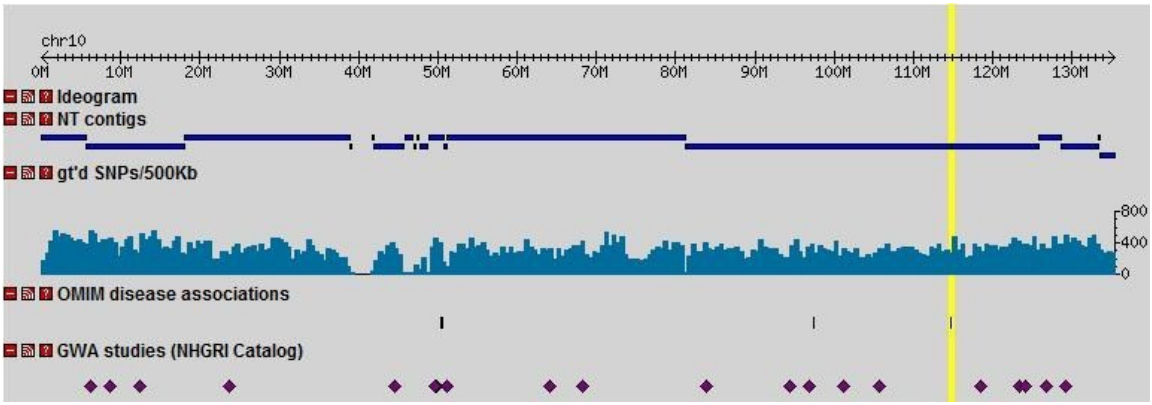


Figura 4.1: panoramica della regione TCF7L2

Regione

Il riquadro sottostante (Figura 4.2) ripresenta in scala diversa la porzione evidenziata in Figura 4.1, concentrandosi sulla zona che interessa il fattore di trascrizione scelto. La sezione del grafico **gt'd SNPs/20Kb** riporta il numero di SNPs individuati ogni 20 Kb.

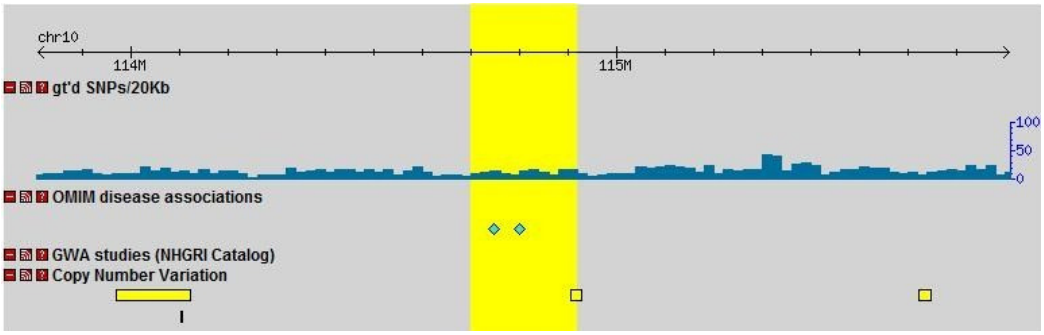


Figura 4.2: regione del cromosoma 10 occupata dal fattore di trascrizione

Dettagli

Nella sezione Dettagli vengono proposti diversi tipi di analisi; per default inizialmente viene mostrata solo una parte di tutta l'informazione disponibile. La sezione più importante è la *Genotype SNPs* (Figura 4.3). I vari SNPs sequenziati dal progetto sono indicati con un triangolo equilatero in corrispondenza della loro posizione sul cromosoma (Figura 4.3). Il grado di dettaglio con cui vengono visualizzati cambiano a seconda dello zoom imposto.

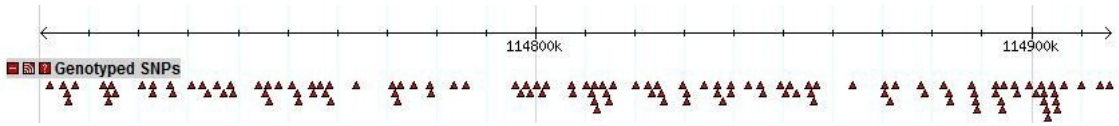


Figura 4.3: visualizzazione grafica della posizione degli SNPs individuabile per mezzo del righello posizionato sopra

Andando a cambiare scala è possibile individuare con maggior dettaglio gli SNPs e leggerne il nome e le frequenze alleliche stimate per ogni popolazione. Se si sceglie ad esempio di indagare sull'intervallo di posizione 114,730,000 - 114,800,000 troviamo il seguente risultato (Figura 4.4). Posizionando il cursore su ogni singolo SNP o cliccando su esso, si apre una tabella che riporta le frequenze genotipiche e alleliche. I dati sono suddivisi in base alla popolazione di provenienza per le quali vengono utilizzate le seguenti sigle:

Selezionando, ad esempio, lo SNP rs081062 si ottengono le informazioni nella tabella riportata in Figura 4.5.

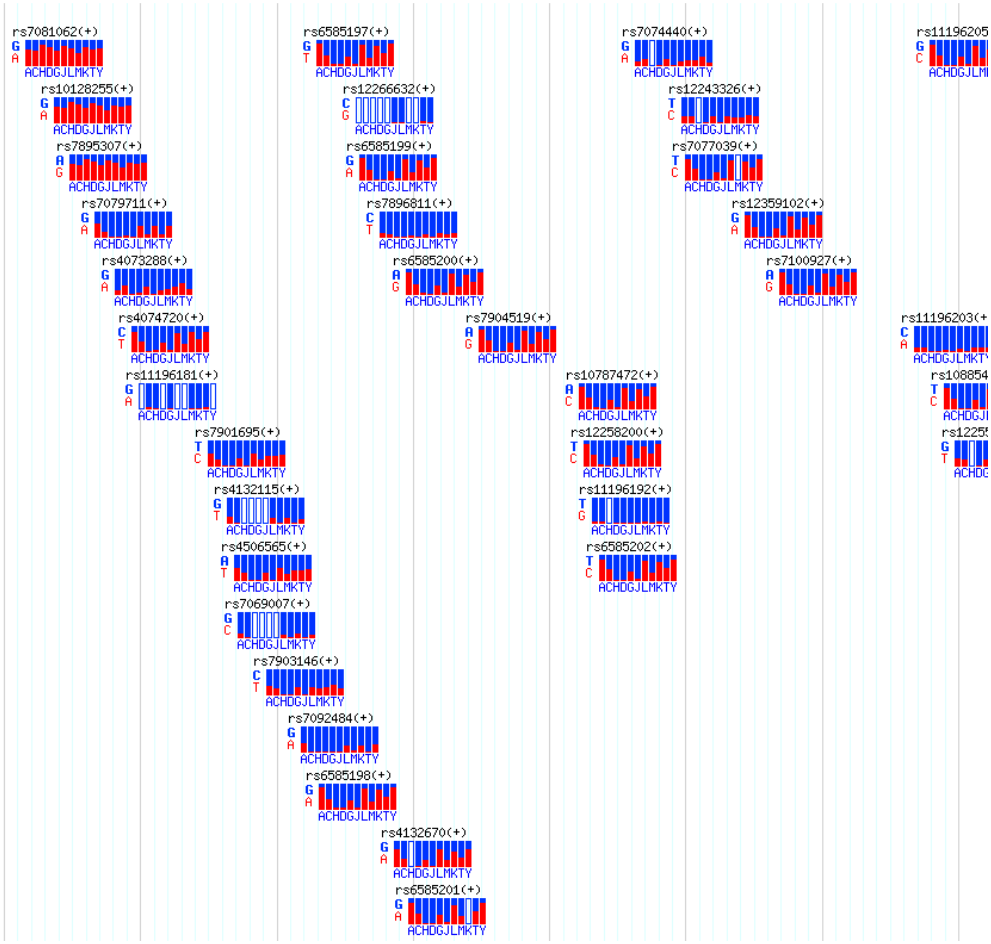


Figura 4.4: visualizzazione più dettagliata degli SNPs compresi all'interno della regione 114,730,000-114,800,000

SNP info: refSNP rs7081062 with alleles A/G in dbSNP b126 (dbSNP report | Ensembl SNPview)

Genomic location: chr10:114730735..114730735, (+) strand relative to the human reference sequence

Frequency report:

Population	Genotype frequencies									Allele frequencies						
	genotype	freq	count	genotype	freq	count	genotype	freq	count	Total	Ref-allele allele	freq	count	Other-allele allele	freq	count	Total
ASW (A)	G/G	0.075	4	A/G	0.547	29	A/A	0.377	20	53	G	0.349	37	A	0.651	69	106
CEU (C)	G/G	0.150	17	A/G	0.522	59	A/A	0.327	37	113	G	0.412	93	A	0.588	133	226
CHB (H)	G/G	0.024	2	A/G	0.321	27	A/A	0.655	55	84	G	0.185	31	A	0.815	137	168
CHD (D)	G/G	0.059	5	A/G	0.412	35	A/A	0.529	45	85	G	0.265	45	A	0.735	125	170
GIH (G)	G/G	0.182	16	A/G	0.409	36	A/A	0.409	36	88	G	0.386	68	A	0.614	108	176
JPT (J)	G/G	0.023	2	A/G	0.395	34	A/A	0.581	50	86	G	0.221	38	A	0.779	134	172
LWK (L)	G/G	0.111	10	A/G	0.378	34	A/A	0.511	46	90	G	0.300	54	A	0.700	126	180
MEX (M)	G/G	0.260	13	A/G	0.440	22	A/A	0.300	15	50	G	0.480	48	A	0.520	52	100
MKK (K)	G/G	0.056	8	A/G	0.451	64	A/A	0.493	70	142	G	0.282	80	A	0.718	204	284
TSI (T)	G/G	0.102	9	A/G	0.466	41	A/A	0.432	38	88	G	0.335	59	A	0.665	117	176
YRI (Y)	G/G	0.080	9	A/G	0.434	49	A/A	0.487	55	113	G	0.296	67	A	0.704	159	226

Note: the 'reference' allele is the base observed in the reference genome sequence at this location

Figura x.5: tabella che si visualizza selezionando lo SNP rs7981062

Si può visualizzare la percentuale di contenuto di G/C nel DNA nella regione selezione (Figura 4.6).

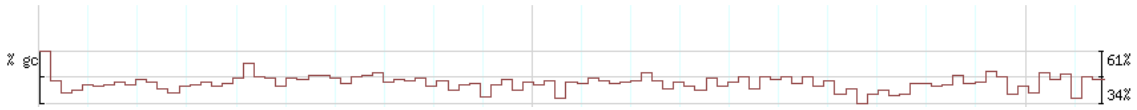


Figura 4.6: contenuto di G e C nella sequenza del fattore di trascrizione

Nella pianificazione di uno studio di associazione è essenziale la conoscenza dell'estensione del linkage disequilibrium (LD) nella regione target per la riduzione del numero di SNPs da sequenziare. La determinazione dei pattern di LD è stato uno dei principali obiettivi del progetto HapMap. I dati possono essere scaricati in blocco dal sito o consultati interattivamente utilizzando il browser. Per quanto riguarda la seconda opzione, nella sezione Dettagli è possibile vedere graficamente la distribuzione dei pattern di LD nella regione selezionata. I parametri chiave sono il tipo di misura che si intende usare per il calcolo del LD (parametri D', r^2 e il LOD score) [ref capitolo], l'orientazione del triangolo con cui si visualizza il pattern (se con il vertice in alto o in basso), lo schema di colori., e se le dimensioni dei box nel plot devono essere proporzionali alla distanza genomica tra i marker o di dimensione uniforme. I pattern di LD possono essere visualizzati per una o più popolazioni. In Figura x.7 si visualizza il Plot del LD ricavato con i soli dati dei campioni della popolazione CEU e utilizzando le due misure più comuni: D' e r^2. In alternativa è possibile visualizzare il plot per ognuna delle altre popolazioni.

Il passo successivo alla determinazione dei pattern di LD, è la scelta dei tagSNP. Per piccole regioni è possibile selezionare i tagSNPs "a mano", utilizzando i supporti grafici e numerici generati sopra. Tuttavia il miglior risultato è garantito sempre dall'utilizzo di un algoritmo che scegla i tagSNPs massimizzando formalmente il numero di SNPs in LD catturati nel tag-set. Non esiste un unico insieme di tagSNP che soddisfa le esigenze del probelma. Saranno i ricercatori a selezionare quali SNPs lavorano meglio con un particolare sistema di genotipizzazione, e a effettuare il compromesso tra costi di sequenziamento di uno studio di popolazione e la forza del livello di associazione che possono identificare.

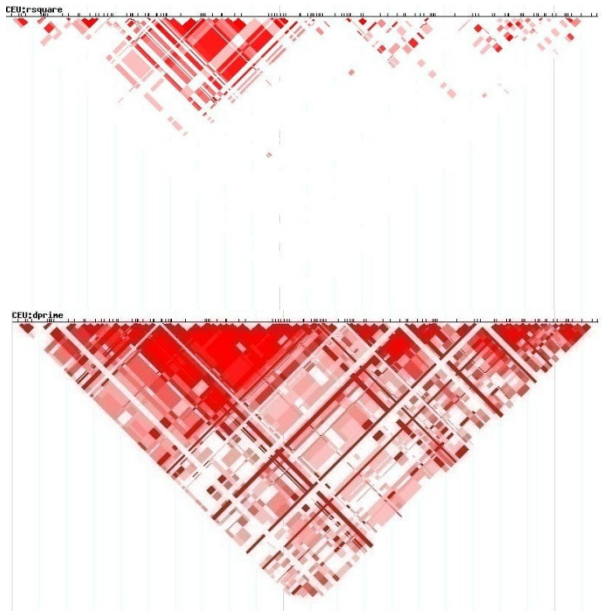


Figura 4.7: visualizzazione dei plot di LD utilizzando i soli dati relativi al gruppo CEU. Il riquadro superiore è ottenuto plottando i valori di LD calcolati con r^2, mentre il riquadro inferiore è ottenuto plottando i valori di LD misurati tramite D'. Il plot a triangolo è costruito congiungendo ogni coppia di SNPs su una linea orientata di 45° rispetto l'orizzontale. Il colore più intenso indica un grado maggiore di LD, mentre le zone colorate in grigio indicano dati mancanti.

Per questo motivo, il sito HapMap non offre un set di tagSNPs preselezionati, ma invece offre ai ricercatori un mezzo per selezionare interattivamente i tag basandosi su criteri scelti dall'utente. La lista di tagSNPs è generata da algoritmi supportati dal programma Tagger. (http://www.broad.mit.edu/mpg/tagger/, de Bakker et al. 2005). Nella sezione *Scaricamento, Ricerca e altre operazioni* si sceglie *Annota TagSNP Picker* per identificare i TagSNP. Se si seleziona Configura, è possibile settare le diverse opzioni che includono: la scelta di una popolazione e di un algoritmo (Tagger Pairwise o Tagger Multimaker), l'inclusione o esclusione di una lista di SNPs dai tag da selezionare, inclusione di una lista di "punteggi" che pesano diversamente i maker, selezione di soglie di cutoff sui valori di LD accettabili e le frequenze alleliche per gli SNPs da includere. Il metodo Tagger Pairwise, sviluppato da Carlson et al, seleziona uno SNP come tag se presenta alta correlazione con un altro. Il metodo Tagger Multimaker invece utilizza un approccio basato non sul confronto a coppie come per il caso precedente, ma su predittori multi-marker e per questo risulta avere una maggiore efficienza. Oltre al tipo di algoritmo, è possibile settare anche il valore minimo del coefficiente r^2 con il quale selezionare lo SNP (se impostato a 1 si ottiene un set di SNP

non ridondante) e il valore della minor frequency allele (MAF). Il risultato viene mostrato graficamente nel riquadro inferiore di Figura x.8, dove nella sezione *tagSNP Picker*, vengono riportati i tagSNPs del fattore di trascrizione TFC7L2.

Nel riquadro superiore di Figura 4.8, è presente invece il riquadro relativo alla **OMIM disease associations**. Posizionando il cursore sullo SNP evidenziato e cliccando su esso, si accede alla pagina web OMIM (Online Mendelian Inheritance in Man) relativa al fattore di trascrizione TFC7L2 dove si può leggere non solo quanto riguarda il fattore stesso, ma come quel particolare SNP è implicato nella predisposizione genica verso determinate malattie. Nel caso in esame l'articolo riporta brevemente i risultati più importanti conseguiti dai gruppi di ricerca con i relativi riferimenti bibliografici.

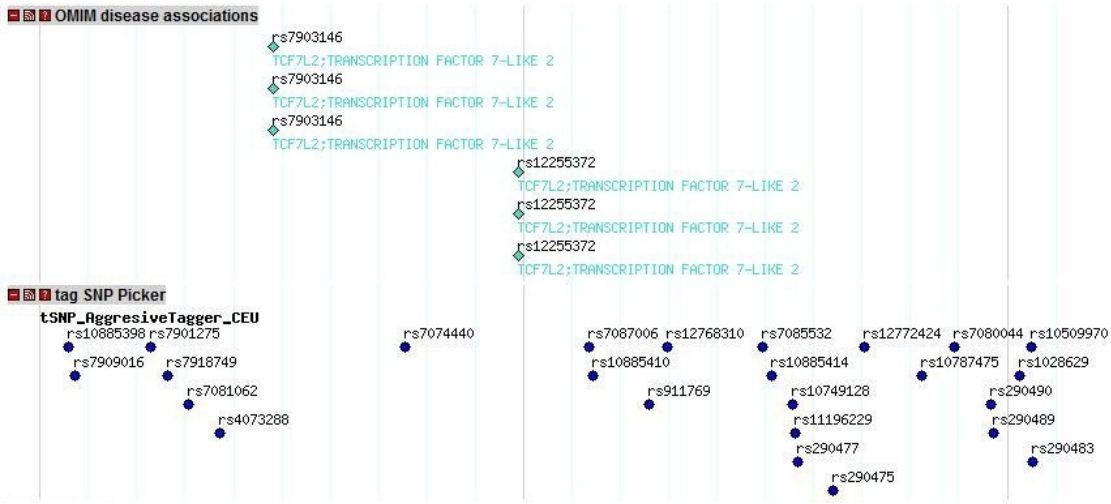


Figura 4.8: visualizzazione di OMIM disease association e i TagSNP

Per la visualizzazione degli aplotipi, il sito HapMap si appoggia al programma PHASE versione 2.1 (Stephens e Donnely, 2003). In Figura 4.9 sono riportati gli **aplotipi** relativi al fattore di trascrizione ricostruiti statisticamente utilizzando il software PHASE. Nel plot vengono mostrati tutti i 120 cromosomi con gli alleli colorati in giallo o in blu.

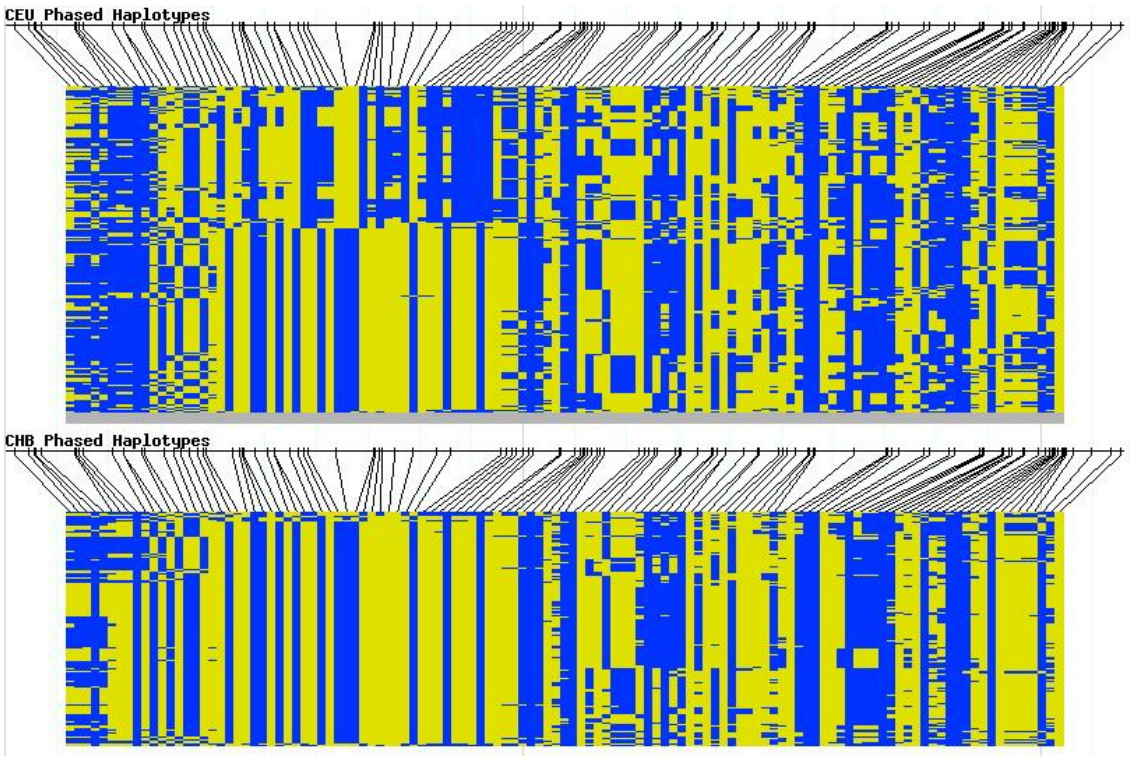


Figura 4.9: visualizzazione degli aplotipi di due popolazioni. Nel riquadro superiore è rappresentato l'aplotipo del gruppo CEU, mentre quello inferiore è relativo al gruppo CHB.

Durante il *phasing*, ogni allele in un genotipo è assegnato a uno o all'altro cromosoma parentale usando un algoritmo di *maximum likelihood* che usa l'informazione sui nuclei familiari di tre persone (genitori e un figlio) nei gruppi della popolazione HapMap. Se questa informazione non è disponibile (non è presente cioè informazione sui gruppi familiari, ma solo su individui singoli), si fittano i dati in un modello che minimizza il numero di crossover nella popolazione. Gli aplotipi phased sono visualizzabili graficamente mediante un codice a due colori. Ogni cromosoma è rappresentato come una linea di altezza 1 pixel e ogni allele dello SNP è arbitrariamente colorato in blu o giallo. Una regione ad alto LD apparirà come una regione nella quale ci sono lunghe file di SNPs in cui gli alleli hanno il medesimo colore, indicando una bassa ricombinazione. Una regione a basso LD apparirà come un'area dove invece i segmenti sono più corti e più frammentati. L'ordine dei cromosomi è determinato da un algoritmo di clustering gerarchico che raggruppa i cromosomi che condividono che condividono aplotipi simili. In Figura x.9 vengono messi a confronto gli aplotipi dei gruppi CEU e CHB.

4.4 HapMart

L'analisi condotta finora ha consentito di visualizzare i dati prevalentemente da un punto di vista grafico. Per avere listati gli SNPs di una determinata regione di interesse e/o di un determinato gruppo etnico, con le relative caratteristiche (frequenza allelica, genotipica..), si accede ad *HapMart*. HapMart è una versione modificata di BioMart, un sistema di data management orientato alle query. BioMart è stato sviluppato in modo che i ricercatori potessero eseguire query anche complesse, consultando i maggiori database di sequenze biomolecolari, pathway, e di annotazione, come Ensembl, Uniprot, Reactome HGNC, Wormbase e PRIDE. HapMart utilizza la stessa interfaccia grafica di BioMart e la stessa modalità di esecuzione della query, ma quest'ultima è limitata al database di HapMap, pertanto la ricerca risulta circoscritta ai soli SNPs del progetto.

Nella sezione di HapMart è possibile selezionare in dettaglio i criteri con cui effettuare la query. In particolare :

- Tipo di database (anche se di fatto la scelta è limitata a un'unica risorsa);
- popolazione (posso considerare tutte le popolazioni o un gruppo soltanto);
- il valore della MAF (%);
- supporto per il sequenziamento dei campioni: Perlegen amplicon-based platform (che sequenzia gli SPNs da frammenti di DNA amplificati con PCR), Affimetrix GeneChip Mapping Array, Illumina HumanHap100, MIP,..;
- la regione genica di interesse (numero cromosoma e/o intervallo di basi);
- inclusione o meno di determinati SNP.

Si selezionano infine i dettagli, nella sezione *Attributes*, che si vogliono inclusi nel report finale:

- ID, cromosoma di appartenenza, posizione, alleli, allele di riferimento,..;
- codice della popolazione, genotipo;
- frequenza allelica e genotipica dello SNP.

Nel risultato finale vengono riportati per default i primi 10 risultati; è possibile scaricarne una versione anche in formato Excel. La stessa ricerca può essere effettuata con BioMart, selezionando il database relativo ad HapMap.

4.5 Ricerca per malattia

Se nel campo ricerca del browser si inserisce il nome della malattia, ad esempio *diabetes*, il database provvede a dare come risultato tutte le regioni su ogni cromosoma che risultano coinvolte con questa malattia. In particolare, per ogni singolo risultato sono indicati il numero del cromosoma e l'intervallo di basi che interessano la regione. Se disponibile viene riportata una breve descrizione che spiega cosa codifica quella sequenza, e come è coinvolta nella patologia. Nel caso in esame del diabete, la query produce 401 risultati. Andando ad indagare per ognuno di essi, si trovano collegamenti ad altri database (Reactome, BioXRT,..) che permettono di studiare come e in quali processi sono coinvolte le sequenze. Il database permette quindi di selezionare ogni singolo risultato e ottenere un'analisi uguale a quella condotta in precedenza per il fattore di trascrizione TCF7L2.

Capitolo 5

PLINK

Gli studi di associazione sull'intero genoma (WGAS) hanno comportato una nuova sfida computazionale e analitica per i ricercatori. Molti supporti già esistenti per l'analisi genetica non sono stati progettati per maneggiare un così ampio data set in modo pratico ed efficiente e non riescono a spiegare la complessità di indagine che deriva dall'utilizzo dei dati dell'intero genoma. Uno degli strumenti più diffusi per analizzare questo tipo di dati è PLINK (di Shaun Purcell, http://pngu.mgh.harvard.edu/~purcell/plink/), uno strumento open-source, che permette di condurre le analisi di routine in modo computazionalmente più efficiente, e offre la possibilità di introdurre nuovi metodi che sfruttano al meglio le potenzialità di data set così grandi.

Le principali funzionalità di PLINK WGAS permettono di :

- provvedere a un modo semplice per gestire grandi set di WGAS;
- stimare gli errori dovuti al problema della stratificazione della popolazione [ref] e ai genotipi errati [ref];
- operare una varietà di test di associazione standard in modo efficiente su grandi data set (su una popolazione, su famiglie, con o senza covariate, test su aplotipi,…).

5.1 Formato dei dati PLINK

I dati utilizzati per l'analisi con PLINK sono tipicamente di due tipi: *mydata.ped* e *mydata.map* .

5.1.1 PED files

Un file .ped (Tabella 5.1) è caratterizzato dalla presenza di 6 colonne obbligatorie, dove ogni individuo è rappresentato da una riga, caratterizzate dai seguenti campi:

1. Family ID
2. Individual ID
3. Paternal ID
4. Maternal ID
5. Sex (1 = maschio, 2 = femmina, other = non specificato)
6. Phenotype.

Gli ID sono alfanumerici; la combinazione degli ID dei campi relativi all'appartenenza a un gruppo famigliare e all'individuo (rispettivamente prima e seconda colonna) deve identificare univocamente una persona. La sesta colonna, Phenotype, deve essere caratterizzata dalla presenza di un solo codice relativo a un preciso fenotipo che può assumere uno dei seguenti valori: 1 corrisponde alla presenza del fenotipo, 2 indica l'assenza del fenotipo, 0, -9 o *missing genotype* identificano, in forma alternativa, la mancanza di informazione sulla presenza o assenza del fenotipo. Quest'ultimo può essere una caratteristica, uno stato o una malattia: PLINK individua automaticamente lo status dell'individuo basandosi sul valore osservato in corrispondenza di questo campo. Nel caso in cui il sesso dell'individuo non sia noto, allora viene identificato con un carattere diverso da 1 e 2. Se nel data set non sono presenti informazioni sul fenotipo o sul sesso dell'individuo, questo sarà automaticamente escluso dalle analisi che fanno uso di queste stesse informazioni. Ad esempio, verrà dato un messaggio di errore nel caso in cui si sta eseguendo un'indagine nella quale si richiede di ricostruire il nucleo famigliare e uno degli individui, di cui manca il sesso, deve essere identificato come padre o madre. In caso in cui sia disponibile anche l'informazione sul genotipo, questa viene riportata dalla settima colonna in poi, e può essere denotata con un codice di caratteri qualsiasi (ad esempio 1,2,3,4 o A,C,T,G) ad eccezione dello 0 che indica il genotipo mancante. I caratteri devono essere separati da uno spazio bianco. Tutti i markers devono essere biallelici e, per tutti gli SNPs, entrambi gli alleli devono essere specificati.

1°	2°	3°	4°	5°	6°	7°
1	1	0	0	1	1	A A G T
2	1	0	0	1	1	A C T G
3	1	0	0	1	1	C C G G
4	1	0	0	1	2	A C T T
5	1	0	0	1	2	C C T T
6	1	0	0	1	2	C C T T

Tabella 5.1: prime righe di esempio di un file PED di cui vengono specificati i campi per colonna. Il file è composto da 6 individui, tutti di sesso maschile (il campo in colonna 5 è sempre pari a 1).3 di essi hanno fenotipo pari a 1, quindi casi. I rimanenti sono sani (controlli).

5.1.2 MAP files

Per default, un file .map (Tabella 5.2) descrive, in ogni riga, un singolo marker e deve contenere esattamente 4 colonne caratterizzate dai seguenti campi:

1. Numero cromosoma
2. rs (o identificatore dello SNP)
3. distanza genetica o cromosomica (cM)
4. posizione della coppia di basi (bp).

La distanza cromosomica deve essere espressa in centimorgan (unità di misura della distanza genetica tra 2 loci; 2 loci che presentano frequenza di ricombinazione dell1% , sono definiti distanti 1 cM) con il comando *--cm*. Alternativamente, si può usare un file MAP con la distanza genetica esclusa, allora si dovrà specificare *--map3*; in questo caso sono richieste solo 3 colonne. La maggior parte delle analisi non richiede la distanza genetica, quest'ultima risulta utile solo nel caso in cui l'analisi si occupi di individuare eventuali segmenti che possono essere condivisi da più individui. Per i test di associazione di base, la colonna della distanza genetica può essere settata a 0. La posizione della coppia di basi è un numero positivo con un range di valori che copre tipicamente le dimensioni dei cromosomi umani. L'identificatore dello SNP può contenere ogni tipo di carattere eccetto spazi vuoti; dovrebbe essere evitato anche il carattere *. Per escludere uno SNP dall'analisi, si pone nella quarta colonna un simbolo – davanti al valore corrispondente. Il file MAP deve quindi contenere tanti marker quanti sono quelli presenti nel file PED.

1°	2°	3°	4°
1	snp1	0	5000650
2	snp2	0	5000830

Tabella 5.2: prime righe di esempio di un file MAP. I due SNPs si trovano rispettivamente sul cromosoma 1 e 2 come specifica la prima colonna.

5.1.3 Fileset trasposti

Un altro possibile formato di file è chiamato trasposto. Vengono presi i formati dei file MAP/PED, e invertite tutte le informazioni dei genotipi tra i file, mediante un'operazione analoga alla trasposizione di matrice. Il risultato è caratterizzato da due file di testo: uno (TPED) contenente per ogni riga uno SNP e informazione del genotipo, l'altro (TFAM) contenente informazioni sull'individuo e sulla famiglia, dove ogni riga è un individuo. Le prime 4 colonne del file TPED sono le stesse 4 colonne standard di un file MAP. Dalla quinta colonna in poi, tutti i genotipi sono listati per tutti gli individui per ogni particolare SNP. Il TFAM contiene invece solo le prime 6 colonne di un file PED standard. In Tabella 5.3 e 5.4 vengono proposti due esempi rispettivamente di TPED e TFAM.

1°	2°	3°	4°	5°
1	snp1	0	5000650	A A A C C C A C C C C C
2	snp2	0	5000830	G T G T G G T T G T T T

Tabella 5.3: prime righe di esempio di un file TPED. Le prime 4 colonne corrispondono esattamente alle 4 colonne del file MAP in Tabella 5.2. La quinta colonna contiene il genotipo.

1°	2°	3°	4°	5°	6°
1	1	0	0	1	1
2	1	0	0	1	1
3	1	0	0	1	1
4	1	0	0	1	2
5	1	0	0	1	2
6	1	0	0	1	2

Tabella 5.4: prime righe di esempio di un file TFAM.

Questo tipo di formato è conveniente da utilizzare nel caso in cui siano presenti molti più SNP rispetto ai soggetti (ad esempio nei dati WGAS). Il programma permette di generare dataset trasposti grazie al comando *–transpose*.

5.1.4 File binari

Per un risparmio di spazio e tempo computazionale, è possibile creare un file PED binario (.bed). Questo permette di salvare le informazioni di pedigree e fenotipo in file separati (.fam) e crea un file MAP esteso (.bim). Nel file .bim vengono memorizzate le informazioni relative agli alleli, che altrimenti verrebbero perse nella creazione del BED. Riassumendo si ha:

- *mydata.bed* : file binario con le informazioni del genotipo;
- *mydata.fam*: prime sei colonne del file mydata.PED;
- *mydata.bim*: file MAP esteso con 2 colonne extra per i nomi degli alleli

5.1.5 Codifica cromosomi e alleli

Gli autosomi devono essere indicati con un numero che va da 0 a 22. I codici di seguito elencati servono per specificare cromosomi di altro tipo:

- X cromosoma X (23)
- Y cromosoma Y (24)
- XY regione pseudoautosomale di X (25)
- MT cromosoma mitocondriale (26)

I numeri riportati tra parentesi rappresentano il codice interno utilizzato da PLINK per questi cromosomi; questo codice numerico apparirà in tutti gli output del programma. Per cromosomi aploidi, i genotipi dovrebbero essere specificati come omozigoti. Il campo NM indica il numero di alleli non mancanti (non missing) per ogni SNP, questo perché i genotipi non validi sono automaticamente considerati come mancanti. I genotipi, etichettati come mancanti, vengono comunque preservati in un file a parte, e possono comunque essere riconsiderati nelle analisi successive.

Per quanto riguarda gli alleli, per default, l'allele minore è codificato con A1, mentre il maggiore con A2.

5.2 Data management

PLINK provvede a riordinare, ricodificare e filtrare le informazioni del genotipo. Vengono brevemente elencate le principali operazioni possibili:

- riordino e ricodifica del file,
- trasposizione dei data set,
- elenco in base al conteggio dell'allele minore, o genotipo o SNPs,
- aggiornamento dei dati sugli SNPs, sugli alleli e sui singoli individui,
- creazione di file per le covariate,
- unione di due o più file set,
- estrazione di un sottogruppo di SNPs (in base al singolo cromosoma,o al range di SNPs, in riferimento a un unico SNP coinvolgendo il gruppo di SNPs limitrofi, o a più SNPs considerando per ognuno la finestra di SNPs vicini,..),
- rimozione di un sottoinsieme di SNPs,
- rimozione di uno specifico sottoinsieme di genotipi,
- estrazione e/o rimozione e/o filtraggio di un sottoinsieme di individui.

È possibile unire due o più dataset contenenti dati che possono sovrapporsi parzialmente, in termini sia di individui che di marker, e produrre reports per l'individuazione di eventuali discrepanze tra i dataset.

5.3 Summary statistic

PLINK genera una serie di misurazioni statistiche che risultano utili per il controllo qualità dei dati a disposizione (quantità di genotipi mancanti, MAF, errori nell'equilibrio Hardy-Weinberg, tasso di ereditarietà non Mendeliana,..). I risultati possono poi essere utilizzati come soglia per le analisi successive. Tutte le statistiche per gli SNPs sono state condotte dopo aver rimosso individui con un alto tasso di informazioni genotipiche mancanti.

5.3.1 Missingness

Questa operazione crea due file: *myData.imiss* e *myData.lmiss* che descrivono dettagliatamente per individuo e per SNP le eventuali mancanze. Per testare le eventuali differenze nei dati mancanti tra casi e controlli si sceglie l'opzione - - *test-missing*. Questa crea un file .missing; l'operazione è ovviamente effettuabile nel caso si abbia un dataset con casi e controlli. Per gli individui, l'output presenta i seguenti campi:

1. FID Family ID
2. IID Individual ID

3. MISS_PHENO fenotipo mancante? (Y/N)
4. N_MISS numero di SNPs mancanti
5. N_GENO numero di genotipi mancanti
6. F_MISS proporzione di SNP mancanti

Per ogni SNP, l'output ha i seguenti campi:

1. SNP identificativo dello SNP
2. CHR numero del cromosoma
3. N_MISS numero di invidui a cui manca quello SNP
4. N_GENO numero di genotipi mancanti
5. F_MISS proporzione di SNP mancanti

Nel caso in cui si scelga di effettuare un controllo separato per casi e controlli, l'output dell'analisi provvederà a dare la percentuale di dati mancanti separatamente per casi e controlli e un p_value calcolato applicando il test esatto di Fisher. PLINK inoltre testa la possibilità di predire la mancanza di un dato dall'aplotipo locale circostante, per determinare il mancato sequenziamento rispettando il genotipo. Per ogni SNP, ci chiediamo se gli aplotipi formati da due (o più) SNPs adiacenti possono determinare quali individui sono mancanti rispetto allo SNP di riferimento. Questo test assume una informazione sul genotipo degli SNP densa, tali che gli SNPs limitrofi siano in LD tra di loro. Ne consegue che un risultato negativo in questo test può semplicemente riflettere il fatto che c'è un basso LD in quella regione.

5.3.2 Equilibrio Hardy-Weinberg

Per creare una lista dei conteggi del genotipo e del test statistico HW per ogni SNP si utilizza l'opzione *–hardy* che crea un file *.hwe* . Il file di output ha i seguenti campi:

1. SNP identificativo dello SNP
2. TEST codice identificativo: AFF(solo casi), UNAFF(solo controlli), ALL
3. A1 codice allele minore
4. A2 codice allele maggiore
5. GENO conteggio genotipico (11/12/22)
6. O(HET) eterozigosità osservata
7. E(HET) eterozigosità stimata

8. P p_value.

PLINK utilizza, per default., un test esatto per il calcolo dell'HWE, descritto e implementato da Wiggington et al. (*A Note on Exact Test of Hardy-Weinberg Equilibrium*, Wiggington JE, Cutler DJ, Abecasis GR, *Am J Hum Genet*, 2005).

5.3.3 Frequenza allelica

Per generare una lista di MAF per ogni SNP, si crea un file *.frq*. il file contiene 5 colonne:

1. CHR Cromosoma
2. SNP identificativo dello SNP
3. A1 codice allele minore
4. A2 codice allele maggiore
5. MAF Minor Allele Frequency
6. NCHROBS Non-missing allele count

5.4 Analisi di stratificazione di popolazione

PLINK propone un approccio statisticamente efficace al problema di stratificazione, in grado di tenere conto di tutti i dati di SNPs dell'intero genoma. Sulla base della proporzione media di alleli condivisi *identical by state* (IBS) tra due individui qualsiasi, PLINK offre uno strumento per clusterizzare gli individui in sottoinsiemi omogenei, eseguire uno scaling multidimensionale (MSD) dei dati per visualizzare le sottostrutture, ricavare indici quantitativi della variazione genetica di popolazione e identificare gli individui. PLINK utilizza il metodo di clustering gerarchico agglomerativo basandosi su una misura della distanza che viene aggiornata con il metodo complete-linkage, per il quale vengono considerati tutti i dati degli SNPs del genoma. Questa procedura agglomerativa parte considerando ogni individuo come un cluster separato di dimensione 1, continua poi unendo i due clusters più vicini. Il complete-linkage clustering specifica che i clusters sono confrontati sulla base dei due componenti più dissimili; l'algoritmo termina quando tutti gli individui appartengono a un unico cluster o quando vengono soddisfatte determinate condizioni predefinite. L'obiettivo è assicurarsi che tutti i membri di ogni cluster appartengano alla stessa sottopopolazione. I metodi proposti per l'identificazione e la correzione della

stratificazione sono necessari per determinare e quantificare gli errori di tipo I e II che derivano dalle analisi di associazione.

5.4.1 Definizione di similarità e distanza tra individui

Per default, l'algoritmo di clustering è basato sul calcolo della matrice delle similarità IBS. Questa matrice è quadrata e simmetrica; se N è il numero di individui considerati nel problema, sarà una matrice NxN, e in ogni posizione saranno riportati i valori della distanza tra tutte le coppie di individui. La misura di similarità tra l'individuo *j* e l'individuo *k* è calcolata con la seguente formula:

$$d_{jk} = 1 - \frac{\sum_{i=1}^{N} |g_{ij} - g_{ik}|}{2M}$$

dove g_{ij}=0,1,2 e *M* è il numero totale di marker (quindi SNPs). Questi valori variano in un range che va da 0 a 1. In pratica non ci si aspetta mai un valore pari a 0, che corrisponde a una coppia di individui totalmente scorrelati, in quanto si presume che anche tra questi ci sia una porzione di genoma condivisa. Un valore pari a 1 indica invece una coppia di gemelli monozigoti, o un casi di duplicazione del campione.

Per generare la matrice delle distanze è necessario specificare un altro comando (distance matrix) il quale darà in output la matrice 1-IBS. Con questa nuova definizione, i valori della matrice prossimi a 1 indicheranno individui molto simili, viceversa valori prossimi allo 0 indicheranno individui scorrelati.

5.4.2 Vincoli sul clustering

Un primo vincolo che PLINK applica sulla procedura di clustering è il PPC test (Pairwise Population Concordance): è un semplice test per verificare che due individui appartengano alla stessa popolazione. Il test si basa sulla proporzione osservata di coppie di SNP IBS 2 {Aa,Aa} e IBS 0 {AA,aa}. Per una data coppia di individui, se appartengono alla stessa popolazione, il conteggio di IBS 2 e IBS 0 dovrebbe essere in rapporto 2:1. Se la coppia è composta da individui che provengono da due diverse sottopopolazioni, allora è atteso un conteggio più alto di SNPs IBS 0. Un test per valutare lo scostamento dal rapporto 2:1 è dato dall'approssimazione normale di una binomiale: per una particolare coppia, se L è il numero totale di SNP indipendenti e informativi, e L_2 è il conteggio di IBS 2,

$$Z = \frac{\frac{L_2}{L} - \frac{2}{3}}{\sqrt{\frac{2}{3} x \frac{1}{3} x \frac{1}{L}}}$$

È possibile scegliere di unire i clusters solo se non ci sono coppie con un risultato del PPC test statisticamente significativo rispetto una data soglia di significatività.

In aggiunta al PPC test, si possono valutare altri vincoli. Un'altra possibilità è raggruppare solo individui che hanno profili simili di dati mancanti, per i quali specifichiamo una soglia per la proporzione massima ammissibile di siti per i quali i due individui sono discordanti nel loro status genotipico. Per campioni casi/controlli, un altro possibile vincolo è imporre che ogni cluster di due o più individui abbia almeno un caso e un controllo. Alternativamente, si può fissare la dimensione massima o il numero massimo di cluster. È inoltre possibile combinare i vincoli su fenotipo e dimensione del cluster specificando che ogni cluster contenga, ad esempio, non più di un caso e tre controlli. In ultimo è possibile tener conto di altre specifiche come età, sesso, variabili ambientali o misure QC come la genotype call rate per ogni individuo.

5.4.3 Algoritmo di clustering

L'algoritmo completo è indicato di seguito: la distanza IBS tra l'individuo *k* appartenente al cluster *i* e l'individuo *l* appartenente al cluster *j* è denotata come d_{ijkl} (definita nel paragrafo *Definizione di distanza tra individui*); la distanza tra cluster è indicata con D_{ij}.

1. START: trova una coppia i,j che soddisfi $\min_{ij}(D_{ij})$, dove $D_{ij}=\max_{kl}(d_{ijkl})$;
2. Test (opzionale) sui vincoli per il nuovo potenziale cluster: il nuovo cluster contiene entrambi casi e i controlli? Il cluster ottenuto unendo *i*+*j* è più piccolo del vincolo imposto sulla dimensione massima? Si eccede il numero massimo di casi o di controlli?
3. Per ogni coppia *i* e *j*, si testano i seguenti vincoli (opzionali): PPC test significativo? superamento della soglia imposta sui dati mancanti? Individuo già selezionato dal gruppo?
4. Vengono soddisfatti i vincoli? →Si uniscono i cluster.
5. Non ci sono altri cluster che possono essere confrontati? →STOP.
6. Ritorno a START.

5.4.4 Scaling multidimensionale (MDS)

PLINK inoltre provvede a proporre un metodo alternativo per l'analisi della stratificazione di popolazione: piuttosto che clusterizzare in gruppi discreti, si può utilizzare la tecnica dello scaling multidimensionale per ottenere una rappresentazione in *k* dimensioni di ogni "sottostruttura". Tuttavia l'utilizzo principale di questo approccio è la visualizzazione grafica; i valori di ognuna delle *k* dimensioni, invece che essere utilizzati per formare clusters dicreti, possono essere usati come covariate nelle analisi di associazione successive per controllare la stratificazione. La tecnica di MDS classica si basa sulla distanza metrica euclidea. L'output del programma darà un file .mds con i seguenti campi:

- FID Family ID
- IID Individual ID
- SOL codice assegnato nella custerizzazione
- C1 posizione della prima dimensione
- C2 posizione della seconda dimensione
- Ck posizione della k-esima dimensione

Plottando i valori di C1 vs i valori di C2, per esempio, otteniamo un plot a due dimensioni nel quale ogni punto corrisponde a un individuo; i due assi corrispondono a una rappresentazione ridotta dei dati in due dimensioni. Il grafico aiuta visivamente a distinguere i cluster che rappresentano i vari sottogruppi. Si prenda in considerazione, ad esempio, un dataset composto da 89 individui, di cui 45 cinesi e 44 giapponesi. È possibile visualizzare le due sottopopolazioni graficamente creando una matrice delle distanze IBS, e integrando il risultato in R, per generare lo scaling multidimensionale. Il plot in uscita è, nel caso in esempio, quello riportato in Figura 5.1

Un grafico di questo tipo suggerisce che nel dataset in considerazione esistono almeno due distinti sottogruppi. Basandosi su tale risultato, si è facilitati nell'analisi di stratificazione di popolazione.

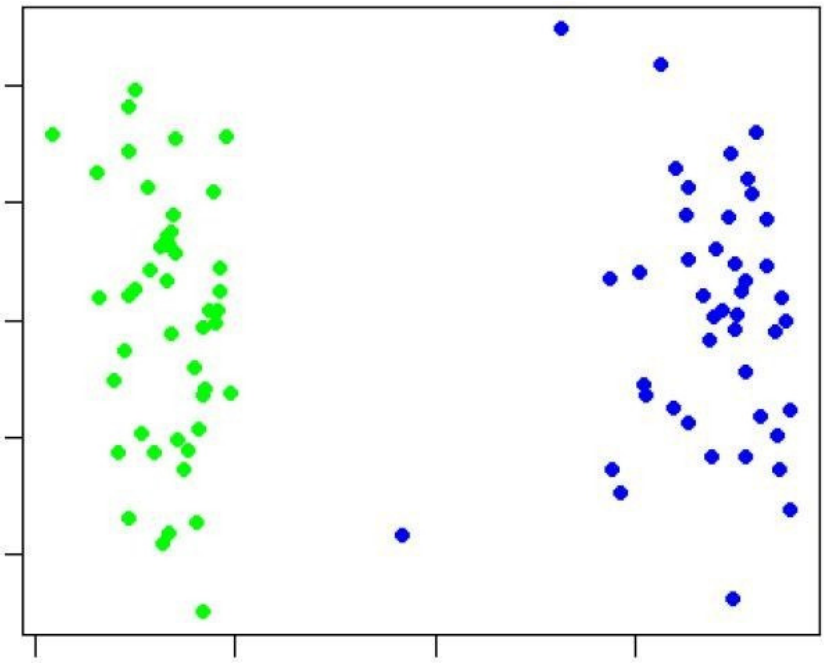


Figura 5.1: visualizzazione grafica delle due sottopopolazioni; il verde rappresenta gli individui cinesi, mentre il blu gli individui giapponesi.

5.4.5 Individuazione di individui outlier

Talvolta, può risultare utile individuare un gruppetto di individui che non appartengono in modo omogeneo a nessun cluster. È possibile utilizzare dei metodi di misura che quantificano quanto si discosta un individuo da un sottogruppo di campioni considerato, basandosi sempre sulle misure riportate nella matrice delle distanze IBS già calcolata per le analisi precedenti. Per ogni individuo, andiamo a classificare tutti gli altri individui sulla base della loro distanza (in termini di IBS) dall'individuo considerato. Dalla distribuzione dei punteggi assegnati ai "vicini più vicini", uno per ogni individuo, si calcolano una media e una varianza campionaria e si trasforma questa misura in uno Z score. Se questo individuo avrà uno Z score estremamente basso, ad esempio inferiore di 4 unità di standard deviation, allora si può concludere che rappresenta un outlier rispetto al resto dei campioni. Così come si esegue il test con il vicino più vicino, è possibile considerare anche la distribuzione del secondo vicino più vicino per ogni individuo, del terzo più vicino, etc. . Può essere talvolta più informativo andare a guardare proprio le misure del secondo e del terzo vicino più vicino, per identificare, ad esempio, una coppia di individui che sono molto simili tra loro, ma molto distanti dal resto del campione.

5.5 Analisi di associazione

I test di associazione sono basati sul confronto delle frequenze alleliche tra casi e controlli.

PLINK offre i seguenti test per eseguire l'analisi di associazione:

- Test standard sul confronto allelico tra casi/controlli,
- Test Cochran-Armitage,
- Test esatto di Fisher,
- Test sul genotipo (con implementazione dei vari modelli dominante, recessivo,..),
- Test Cochran-Armitage-Haenszel (che permette di eseguire l'analisi di associazione condizionato per ogni cluster o altre categorizzazioni dei campioni),
- Test Berslow-Day,
- Test omogeneità della odd-ratio.

Gli ultimi due test elencati utilizzano la regressione lineare standard del fenotipo. I test si trovano già implementati nel programma per permettere una velocizzazione dell'analisi di associazione. In alternativa, è possibile condurre l'analisi con un approccio più generale, mediante l'uso di modelli di regressione lineare o logistica che permettono l'utilizzo di covariate che esprimono sia gli effetti principali, che le varie interazioni. È inoltre possibile testare l'interazione tra geni o tra gene e ambiente. Per eseguire i test di associazione, è possibile specificare il tipo di modello o lasciare che sia il programma stesso a scegliere il miglior modello da applicare (quest'ultima opzione è settata per default). Per le analisi sui nuclei famigliari è implementato il transmission/disequilibrium test (TDT).

Un esempio di output che si può ottenere eseguendo un test standard di confronto tra casi e controlli contiene i seguenti campi:

1. CHR cromosoma
2. SNP SNP ID
3. BP posizione fisica (coppia di basi)
4. A1 nome dell'allele minore
5. F_A frequenza di questo allele nei casi

6. F_U frequenza di questo allele nei controlli
7. A2 nome dell'allele maggiore
8. CHISQ test chi-quadrato (1 grado di libertà)
9. P p-value asintotico
10. OR odd-ratio attesa.

5.6 Stima IBD (identical by descent)

In campioni omogenei, PLINK prevede un'opzione utile a stimare i coefficienti di condivisione IBD sull'intero genoma tra individui apparentemente scorrelati. Questa misurazione può risultare particolarmente utile per il QC, per diagnosi di errori nella ricostruzione dell'ascendenza, e per l'individuazione di scambio di campioni e eventi di contaminazione. PLINK adotta una procedura semplice per trovare tratti di genoma condivisi nell'intero dataset (regioni che comprendono più di un certo numero di SNPs e/o kilobasi) che ricorrono in modo relativamente frequente, e propone un approccio efficace nel mappare i geni. Attraverso una permutazione, può essere calcolato un p-value per ogni SNP sulla base di un test per verificare la presenza significativa di segmenti omozigoti in una data posizione confrontando casi e controlli. PLINK calcola inoltre il coefficiente di incrocio (imbreeding coefficient) per ogni individuo.

È stato implementato un nuovo metodo per identificare condivisioni IBD su segmenti di cromosomi tra coppie di individui lontanamente imparentati utilizzando un modello di Markov (Hidden Markov model, HMM), nel quale lo stati IBD "sottostante" è stimato una volta osservata la condivisione IBS e il livello di relazione tra le coppie. È disponibile anche un test per la correlazione tra la condivisione di segmenti di cromosomi e la condivisione del fenotipo. Questo test, un'analisi di linkage basato sulla popolazione, potenzialmente offre un approccio complementare ai dati dell'intero genoma che non assume l'ipotesi che una variazione genetica sia correlata alla presenza di una malattia. La stima della condivisione IBD può essere usata per QC e indica errori sui campioni e sui genotipi, inclusi eventuali scambi, duplicazioni ed eventi di contaminazione, così come relazioni famigliari errate o non specificate. Se, ad esempio, il DNA proveniente da uno o più individui contamina altri campioni, può portare a un pattern distintivo di campioni contaminati che mostrano alto IBD con tutti gli altri individui. Questo è dovuto al fatto che la contaminazione induce false chiamate eterozigote (ad esempio, AA messo insieme con CC potrebbe essere identificato come

AC). Per di più, i campioni contaminati mostreranno un forte e negativo coefficiente di incrocio, indicativo di un numero maggiore di eterozigoti rispetto a quanto previsto.

Capitolo 6

Obiettivi della tesi

6.1 Motivazione

L'idea del metodo proposto in questo elaborato nasce da un'attenta osservazione dello stato dell'arte degli studi di associazione genome wide, di cui è stata offerta una panoramica generale nei capitoli precedenti. Si è rivolta in particolare l'attenzione su come i ricercatori affrontano le fasi di preprocessing dei dati e feature selection. Sono questi infatti i passaggi più critici di tutta l'analisi in quanto possono influire in modo rilevante sulla significatività del risultato.

Si nota innanzitutto che, per esigenze computazionali, si è costretti a suddividere la ricerca sull'intero genoma in cromosomi, da analizzare separatamente. I polimorfismi a singolo nucleotide hanno infatti la caratteristica di essere estremamente densi nel genoma umano, pertanto i dati ricavati dal loro sequenziamento occupano molto spazio nella memoria ram di un normale calcolatore (Si può arrivare fino a 15 Mb per ogni cromosoma). Inoltre, l'elevato numero di variabili in gioco, dell'ordine di 10^8 SNPs, rende necessaria una fase di feature selection prima dell'analisi multivariata. Se da una parte un dataset estremamente ricco come quello derivante dal sequenziamento degli SNPs può costituire un vantaggio ai fini della quantità dell'informazione, dall'altra costituisce un limite per l'applicazione di un qualsiasi algoritmo di classificazione. La letteratura in merito a questo problema propone due approcci alternativi: il primo prevede di applicare in successione un'analisi univariata e quindi multivariata; il secondo, quest'ultimo molto frequente in letteratura, prevede l'estrazione di un sottoinsieme di tagSNP (capitolo 2.4) scelti sulla base di una soglia di correlazione (maggiore dello 0.8) calcolata considerando tutti i soggetti. Il limite evidente che consegue l'applicazione di questi approcci è la pesante riduzione dell'informazione contenuta nel dataset di partenza. Considerare infatti un sottoinsieme di tagSNPs

piuttosto che l'insieme totale di tutti i marcatori non permette di tenere in considerazione la pluralità di fattori che intervengono a regolare l'espressione e la regolazione genica. È infatti sempre più confermato che le malattie a base genica non siano a carico di un unico gene o regione cromosomica, ma piuttosto siano dovute alla compresenza di numerosi fattori. Nel momento però in cui considero solo un sottoinsieme delle variabili di partenza, questo tipo di regolazione viene a mancare. L'applicazione di un'analisi multivariata, a questo punto, risulta comunque "impoverita", e il suo risultato sarà comunque poco rappresentativo della situazione reale.

Alla luce di quanto evidenziato dallo stato dell'arte, si propone in questa tesi un nuovo approccio che permette di rendere il dataset fruibile per l'applicazione di un qualsiasi metodo di classificazione attraverso una riduzione del numero di variabili attraverso un criterio che permetta di mantenere complessivamente invariata la quantità dell'informazione iniziale.

6.2 Strategie

La metodologia propone innanzitutto di abbandonare l'analisi mirata ai singoli cromosomi per andare a isolare i singoli pathway e i geni che vi appartengono. Infatti essendo particolarmente interessati ad individuare eventuali interazioni tra geni diversi, si ritiene che sia più sensato individuare interazioni tra geni che sono localizzati su cromosomi diversi ma appartengono allo stesso pathway, piuttosto che individuare alti valori di correlazione tra geni che appartengono allo stesso cromosoma, ma a pathway diversi.

Si introducono quindi le definizioni di entropia e mutua informazione come criterio per la costruzione di metavariabili. La metavariabile è costituita da quegli SNPs altamente correlati tra di loro mediante alti valori di mutua informazione. I valori di MI vengono calcolati sfruttando tutte le possibili coppie tra i marcatori e la scelta dei valori più significativi viene fatta eseguendo un test di significatività. Generalmente, dato che la distribuzione in ipotesi nulla dei valori di MI non è nota a priori, l'ipotesi nulla viene ricostruita calcolando la MI di coppie di SNP permutando ripetutamente e indipendentemente, per ogni SNP, i soggetti del dataset. In questo modo, si opera una

riduzione del numero di variabili iniziali, ma la quantità di informazione iniziale rimane sostanzialmente invariata.

Una volta ottenute le metavariabili e, calcolate per ognuna di esse i possibili stati che possono assumere, si procede con la classificazione.

L'idea generale è quella di analizzare più pathway, isolarne i geni e gli SNPs corrispondenti, ricostruire le metavariabili e creare un classificatore per ognuno. Una volta scelti un determinato numero di pathway, si procede ricostruendo un classificatore aggregato di tutti i classificatori costruiti sui diversi pathway biologici.

Capitolo 7

Dati e Preprocessing

7.1 Descrizione dei dati

I dati sono stati rilasciati dal Wellcome Trust Case-Control Consortium (WTCCC, [24]) e sono dati sensibili.

I dati sono stati ricavati utilizzando chip Affymetrix 500K e si compongono come segue:

- 1504 campioni (controlli) dalla 1958 British Birth Cohort (**1958 BC**, 752 maschi, 752 femmine; i partecipanti sono originari di Inghilterra, Scozia e Galles);
- 1500 campioni (controlli) dal gruppo del National Blood Service (**NBS**, 720 maschi, 780 femmine);
- 2000 campioni di individui (casi) affetti da diabete di tipo 1 (**T1D**, 1015 maschi, 985 femmine);
- 2000 campioni di individui (casi) affetti da diabete di tipo 2 (**T2D**, 1162 maschi, 837 femmine).

I dati a disposizione per l'analisi condotta in questa tesi utilizzano l'algoritmo Chiamo[4] per l'assegnazione delle calling dei genotipi. I primi due gruppi costituiscono i controlli, i due gruppi rimanenti costituiscono i casi, considerati singolarmente nei due studi. I dati sono memorizzati in formato PLINK e in fileset trasposti (si veda Capitolo 5, in particolare il paragrafo 5.1.3). Ogni gruppo è caratterizzato da un unico file TFAM che contiene le informazioni relative al campione e agli individui, e da 23 file TPED, uno

[4] Chiamo è un algoritmo utilizzato per sequenziale i genomi in presenza di più coorti. È stato inizialmente sviluppato per analizzare i chip Affymetrix 500K, ma può essere applicato anche a dati provenienti da altre tecnologie. Implementa un approccio gerarchico che riconosce correlazioni tra i parametri di ogni coorte.

per ogni cromosoma (compreso il cromosoma X). Gli SNPs nel file TPED sono nominati con identificativi Affymetrix (ad esempio: SNP_A-224303). Gli SNPs, la cui posizione sul genoma è incerta, sono riportati con numero di cromosoma e posizione nulli. Le informazioni sulla colonna "phenotype" sono settate a 0.

Si è scelto di lavorare su un numero ristretto di SNPs, limitato a quelli appartenenti ai geni del pathway dell'insulina. Questa scelta rappresenta già una prima novità rispetto agli approcci che generalmente vengono adottati dai ricercatori che intendono affrontare un test di associazione con un dataset analogo. Infatti, normalmente, si procede nell'analisi considerando cromosoma per cromosoma in maniera distinta, mentre qui si sono considerati tutti i geni, e quindi i cromosomi, apparteneti a uno specifico pathway. Maggiori dettagli verranno illustrati in Capitolo 7. La lista dei geni è stata ricavata dal database HUGO e comprende inizialmente 293 geni. Successivamente si è ulteriormente limitata la ricerca degli SNPs a quelli compresi all'interno del gene, escludendo quelli localizzati nelle regioni intrageniche e quelli sul cromosoma X. Si è voluto infatti evitare qualsiasi ambiguità di attribuzione degli SNPs a uno dei due geni tra i quali sono compresi. Questo passaggio ha permesso di ridurre ulteriormente il numero di geni e marcatori coinvolti nell'analisi a 1071 SNPs per ogni dataset.

7.2 Preprocessing dei dati

Prima di procedere con l'analisi sul dataset, è necessario stimare la bontà dei dati a disposizione, in modo da rimuovere quelli meno informativi, di bassa qualità, e ridurre il tasso di errori di tipo I.

I sistemi che includono informazione di tipo biologico sono sempre affetti da errori dovuti principalmente a:

- preparazione e qualità dei campioni di DNA (stato di degradazione delle molecole, efficienza del processo di ibridazione,…);
- condizioni sperimentali e operatore-dipendenti specifiche (il tipo di protocollo adottato per l'estrazione dei campioni può introdurre un bias nell'analisi);
- detection del segnale non corretta e assegnazione del campione al genotipo sbagliato.

Se questi errori avessero distribuzione random su tutto il dataset, non influirebbero più di tanto sul risultato statistico. Ma poiché si è sperimentalmente osservato una distribuzione fortemente non casuale, è necessario procedere a una correzione dell'errore per evitare un numero di falsi positivi. In studi GWAS in cui l'obiettivo è identificare differenze anche minime delle frequenze tra casi e controlli, anche la presenza di un piccolo errore sperimentale può falsare pesantemente il risultato.

In particolare, con l'ausilio di PLINK, si sono andati a verificare, per ogni SNP:

- Tasso di *missingness* (percentuale di dati mancanti);
- Hardy-Weinberg Equilibrium (HWE);
- Minor Allele Frequency (MAF).

Per quanto riguarda il primo punto, fissata una soglia del 10%, sono stati mantenuti tutti gli SNPs, in quanto nessun marker sembra evidenziare una percentuale di dati mancanti maggiore della soglia imposta.

La verifica dell' equilibrio HW permette di identificare errori di imbreeding, genotipizzazione, e eterogeneità del campione. La presenza di uno di questi problemi viene rilevata quando esiste una differenza significativa tra le frequenze alleliche misurate nel campione e quelle attese dall'equilibrio (paragrafo 1.3). La verifica si effettua solo sui controlli, quindi, nel caso in esame, sui dataset 1958BC e NBS unificati. In generale, infatti, si presume che solo i controlli rispettino sicuramente questo equilibrio, a differenza dei casi che, presentando differenze di frequenza allelica dovute alla malattia, possono o meno rispettare l'HWE. Con PLINK, viene fissata una soglia sul *p-value* calcolato (paragrago 5.3.2) pari a 10^{-4}; questo permette di escudere 19 SNPs dal dataset 1958BC, 17 SNPs dal dataset NBS, 22 SNPs dal dataset T1D e 45 dal dataset T2D.

Si procede quindi con la verifica dell'ultimo punto che prende in considerazione la Minor Allele Frequency su tutti i dataset (casi e controlli). Per i motivi già citati in paragrafo 1.1.1, è necessario eliminare dall'analisi quei marcatori che hanno una MAF minore di una certa soglia, in quanto rappresentano varianti troppo rare, pertanto non conducono a nessun risultato significativo nell'analisi di associazione. La soglia imposta è pari all'1%, e per ogni dataset vengono scartati: 141 SNPs dal 1958 BC, 141 SNPs dal NBS, 138 SNPs dal T1D, 135 SNPs dal T2D.

A questo punto si sono uniti i dataset 1958BC, NBS e T1D in modo da formare un unico dataset contenente tutti i controlli e i casi del diabete di tipo 1, mantenendo in maniera opportuna solo quei marcatori che hanno superato tutte le tre fasi di preprocessing sopra descritte. Il dataset contiene 918 SNPs. Allo stesso modo si è fatto per i controllli e i casi del diabete di tipo 2, e il dataset finale ha anch'esso gli stessi 918 SNPs. I file .map e .ped dei due dataset sono quindi stati ricodificati in binario (sempre mediante PLINK) nei file .bim, .bed e .fam (paragrafo 5.1.4). Successivamente sono stati nuovamente ricodificati con il programma *bed2num* (Francesco Sambo) che permette di convertire i file .bed e .fam appena creati in un tab delimited text file (con estensione .out) che codifica ogni SNP nei tre stati:

- 1 se il genotipo è *aa*;
- 2 se il genotipo è *aA*;
- 3 se il genotipo è *AA*.

Capitolo 8

Definizione di metavariabili basata sulla MI

In Capitolo 7 si è già fatto riferimento alla decisione di lavorare su geni appartenenti al pathway dell'insulina, a differenza dei normali approcci che prevedono invece una divisione dell'intero genoma, fatta considerando i cromosomi separatamente. Si ritiene che l'approccio adottato in questa tesi possa portare dei vantaggi rispetto a quello più "tradizionale" (dello stato dell'arte) , in quanto permette di considerare tutti i cromosomi contemporaneamente e permette inoltre di individuare, se esiste, una rete di regolazione tra i geni, anche posizionati su cromosomi diversi, che appartengono allo stesso pathway e sono quindi coinvolti nella regolazione di uno stesso prodotto (l'insulina).

A questo punto, per poter procedere con l'analisi e la classificazione dei dati, è necessario ridurre la mole di dati a disposizione, quindi il numero di variabili. Lo stato dell'arte prevede che, una volta effettuato il preprocessing sul dataset, si proceda alternativamente mediante due approcci: una primo approccio percorribile è quello di eseguire nell'ordine una analisi univariata e successivamente multivariata; un secondo approccio è quello di sfruttare le informazioni e le misure di LD sulle regioni geniche e estrarre i tagSNP (paragrafo 2.4). Quest'ultimo approccio prevede quindi di selezionare da un sottoinsieme di SNPs, un unico SNP "rappresentativo" (il tagSNP) a cui tutti gli altri sono strettamente correlati. Questi metodi permettono quindi di selezionare un numero limitato di variabili rispetto all'insieme di partenza, e quindi rendono fattibile il processo di classificazione, a scapito però di una importante riduzione di informazione. L'obiettivo della tesi è invece quello di ridurre il numero di variabili su cui lavorare, mantenendo però intatta la quantità di informazione iniziale. In particolare, si fa ricorso

alle definizioni di entropia e mutua informazione come criterio per la costruzione di metavariabili categoriali.

8.1 Calcolo di entropia, mutua informazione e definizione delle metavariabili

I file .out creati (paragrafo 6.2) presentano i dati in forma matriciale con i soggetti sulle colonne e gli SNPs sulle righe. I valori su ogni riga sono gli stati (variabili categoriali) e variano tra 1 e 3 come descritto precedentemente, e vanno a rappresentare gli stati che lo SNP può assumere. Per ogni SNP, quindi per ogni riga, viene calcolata l'entropia H come:

$$H(SNP_i) = -\sum_{k=1}^{3} P(SNP_i = k) \log_2 P(SNP_i = k)$$

successivamente, per ogni coppia di SNP (420903 coppie in totale), viene calcolata l'entropia congiunta come:

$$H\big(SNP_i, SNP_j\big) = -\sum_{k=1}^{3} \sum_{s=1}^{3} P(SNP_i = k, SNP_j = s) \log_2 P(SNP_i = k, SNP_j = s)$$

e mutua informazione (MI):

$$MI\big(SNP_i, SNP_j\big) = H(SNP_i) + H\big(SNP_j\big) - H(SNP_i, SNP_j)$$

Con metavariabile si intende una variabile in grado di riassumere più dati (in questo caso più SNPs) contemporaneamente e che può assumere più stati. Il criterio con cui vengono costruite le metavariabili è la selezione, mediante opportuni test statistici di significatività, dei valori di mutua informazione più significativi, in modo tale che siano costituite da gruppi di SNPs altamente correlati l'uno con l'altro da valori di alta mutua informazione.

È quindi necessario quindi andare a selezionare i valori di MI più significativa: in questo modo le metavariabili che si andranno a ricostruire saranno caratterizzate da SNPs legati tra di loro da valori di alta mutua informazione. Per la selezione dei valori di mutua informazione più significativa si procede con il metodo delle permutazioni, in modo da costruire una distribuzione in ipotesi nulla dei valori di mutua informazione

ottenuti dai dati permutati, verso i quali confrontare i valori di MI originali rispetto una soglia predefinita. Prima di procedere con le permutazioni, si vanno a normalizzare i valori della MI calcolata dai dati originali come:

$$MI(SNP_i, SNP_j)_{NORM} = \frac{MI(SNP_i, SNP_j)}{\max\{H(SNP_i, SNP_j)\}}$$

I valori di MI normalizzata vengono poi ordinati in ordine decrescente. Si permutano quindi in modo indipendente le righe della matrice originale per un totale di 100 permutazioni (quindi 100 matrici permutate), e per ogni matrice si calcolano i valori di MI normalizzata dai dati appena permutati. Applicando una correzione di Bonferroni, vengono selezionate per il dataset dei controlli e casi T1D 7977 MI, mentre per il dataset dei controlli e casi T2D vengono selezionate 7842 MI. Si ottengono quindi rispettivamente 7977 e 7842 coppie di SNPs per ogni dataset.

Ottenuti i valori di MI (normalizzata) significativa, si vanno a ricostruire le coppie degli SNPs coinvolti in ogni MI e si ricostruisce la rete (si veda Capitolo 8). La rete viene plottata con l'ausilio di Cytoscape. Dalla visualizzazione dell'organizzazione della rete, vengono ricostruite le metavariabili e per ognuna di esse vengono calcolati gli stati possibili.

8.2 Classificazione con Naive Bayes

Il classificatore Naive Bayes (NB) è uno degli algoritmi di classificazione più efficienti per machine learning e data mining. È ampiamente utilizzato per scopi di classificazione soprattutto in ambito biomedico, e più recentemente nell'ambito dei test di associazione Genome Wide (GWAS). Le ragioni per cui viene ampiamente impiegato sono essenzialmente le buone performance di classificazione e l'elevata efficienza computazionale. Il classificatore NB, come dice il nome stesso, è basato sull'assunzione semplificativa che tutti gli attributi che descrivono una certa istanza sono tra loro condizionalmente indipendenti data la categoria a cui appartiene l'istanza. Questa affermazione viene detta *assunzione del Naive Bayes*. Quando questa ipotesi è verificata, il NB esegue una classificazione di tipo MAP. Nonostante questa assunzione sia violata nella maggior parte dei problemi reali, il NB si comporta molto bene e risulta essere molto efficiente. L'assunzione di indipendenza permette di apprendere separatamente i parametri di ogni attributo, semplificando molto l'apprendimento

specialmente in quelle situazioni in cui il numero di attributi è molto elevato e in cui i dati a disposizione non sono molto numerosi. Il dominio di applicazione del Naive Bayes riguarda la classificazione di istanze che possono essere descritte mediante un insieme di attributi di cardinalità anche molto elevata.

Mediante il software *Orange Canvas*, si esegue la classificazione NB inizialmente sui dataset di partenza, non ancora divisi in metavariabili, considerando quindi ogni SNP come una variabile categoriale. Si procede quindi con l'applicazione dell'algoritmo, e, per conoscere le capacità predittive del modello, si procede con la validazione mediante il metodo leave one out ,il calcolo della matrice di confusione e il coefficiente di correlazione di Matthews il quale misura la bontà della classificazione. Questo coefficiente viene calcolato direttamente dalla matrice di confusione come:

$$MCC = \frac{TPxTN - FPxFN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Dove TP indica il numero di veri positivi (true positive), TN indica il numero di veri negativi (true negative), FP il numero di falsi positivi (false positive) e FN il numero di falsi negativi. La misura di correlazione MCC fornisce un raffronto tra le prestazioni dell'algoritmo di classificazione e quelle di un classificatore casuale. Il valore di questo coefficiente varia tra -1 e 1: più precisamente, vale uno in corrispondenza di una classificazione perfetta, in cui tutti i dati sono assegnati correttamente alla classe di appartenenza, assume valore -1 nel caso opposto; un valore pari a zero, invece, indica che il classificatore ha una prestazione paragonabile a quella ottenuta tramite scelta casuale della classe di appartenenza.

Successivamente si esegue nuovamente la classificazione sui dataset considerando la suddivisione in metavariabili.

Capitolo 9

Risultati

Vengono di seguito riportate le reti ricostruite per i due dataset, costituiti dai controlli e casi del diabete di tipo 1 e tipo 2 rispettivamente, ricostruite con Cytoscape. Ogni illustrazione riporta:

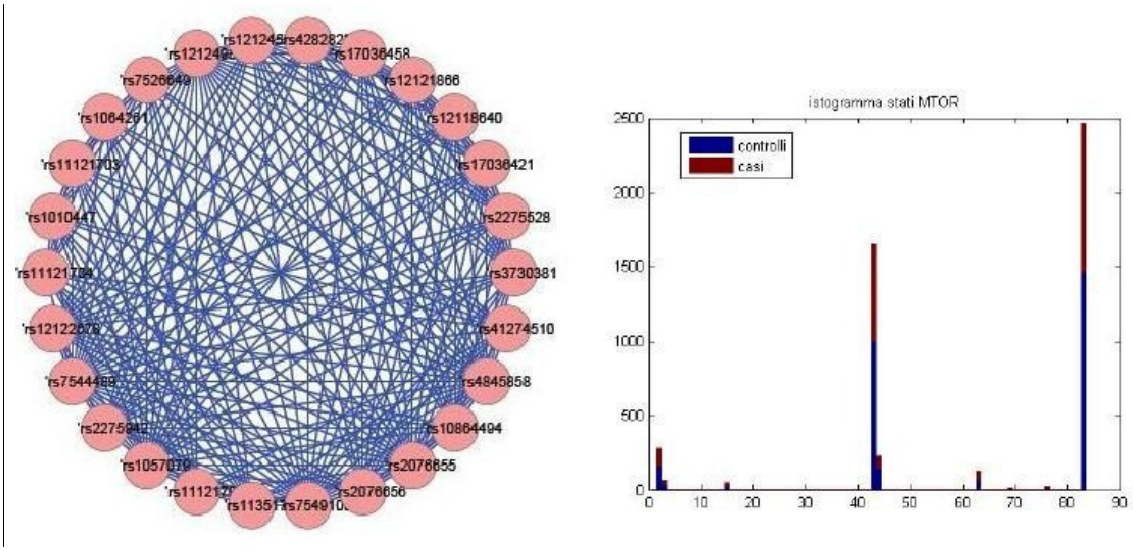
- un *gene*, rappresentato mediante gli SNPs che ricavo dalle coppie di MI selezionata (come descritto in Capitolo 7), e le connessioni tra essi. Gli SNPs sono raffigurati come nodi di un grafo non orientato, in quanto nel calcolo della MI normalizzata si perde la direzionalità della connessione tra i due marcatori. Vengono evidenziate, se sono più di una, le metavariabili costruite mediante tratteggio.
- un *istogramma* degli stati che le metavariabili presenti (una o più di una) assumono. L'istogramma è plottato in modo tale da evidenziare, mediante due colori, il contributo di casi e controlli per ogni stato della metavariabile. È da sottolineare che un gene può costituire una o più meta variabili, a seconda dell'organizzazione e disposizione degli SNPs.

Gli SNPs dei geni non inclusi nelle coppie di mutua informazione selezionate non sono stati riportati nell'elenco che segue, e devono essere trattati come variabili a sé stanti caratterizzate dalla presenza dei tre stati 1,2 e 3.

Per il primo dataset, vengono costruite 149 metavariabili; per il secondo ne vengono costruite 116.

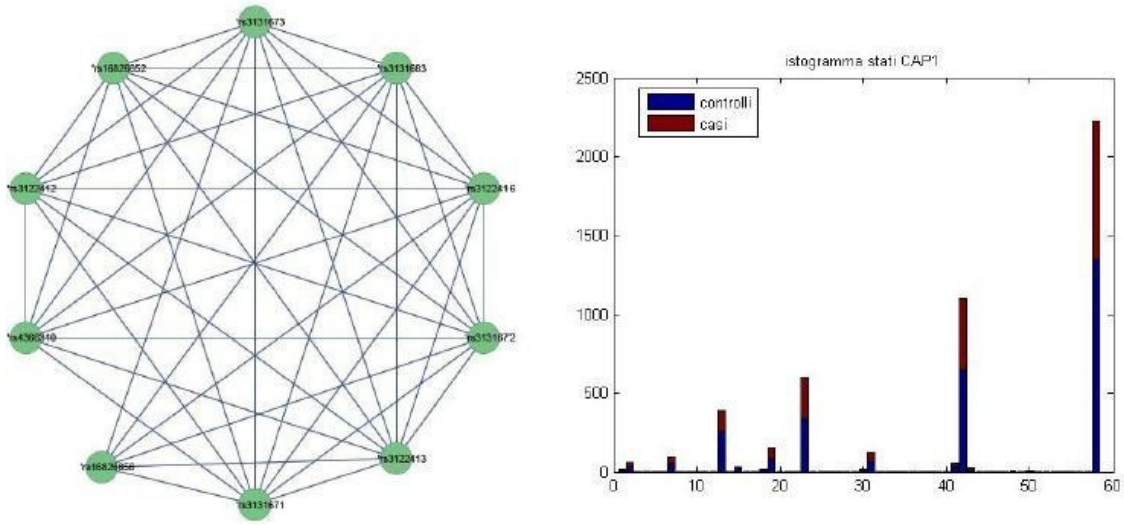
9.1 Rete ricostruita dai dati di controlli e casi del diabete di tipo 1

MTOR



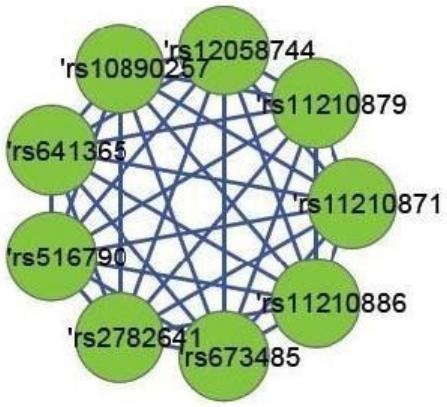
MTOR (cromosoma 1) è costituito da 26 SNPs e costituisce un'unica metavariabile, con 83 stati.

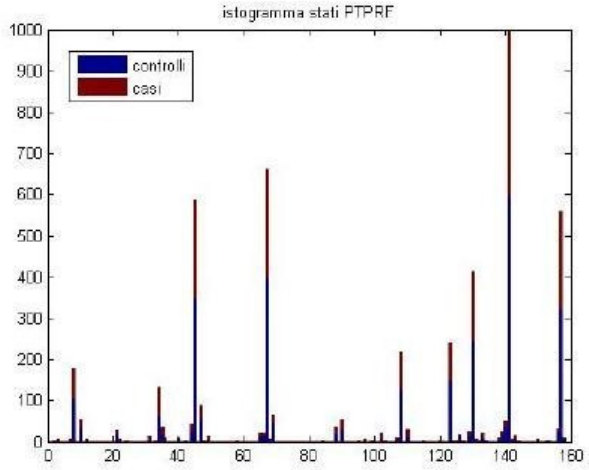
CAP1



CAP1 (cromosoma 1) è costituito da 10 SNPs, costituisce un'unica metavariabile e ha 58 stati.

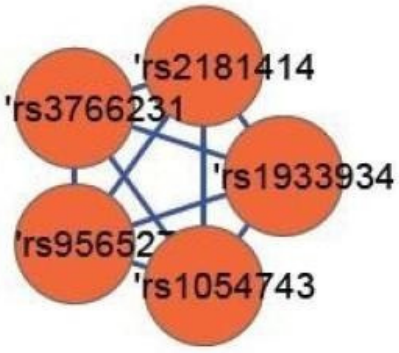
PTPRF

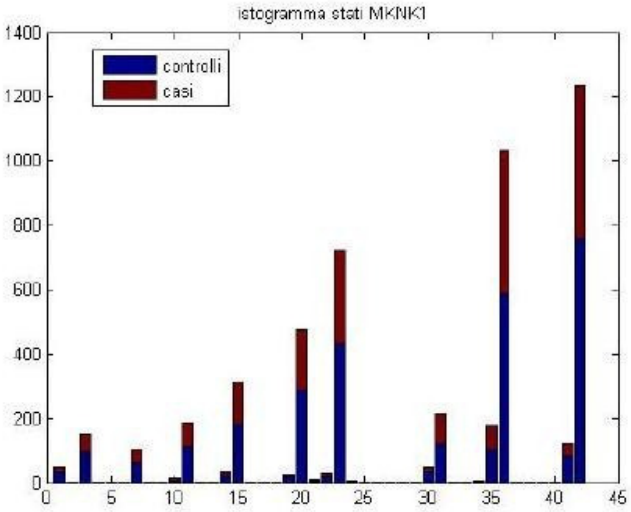




PTPRF (cromosoma 1) ha 9 SNPs, costituisce un'unica metavariabile e 158 stati.

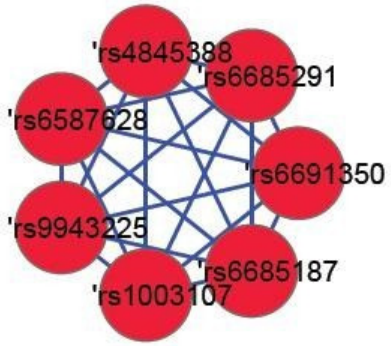
MKNK1

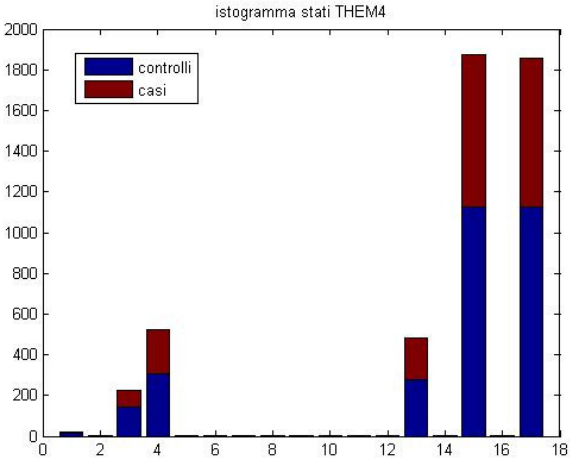




MKNK1 (cromosoma 1) ha 5 SNPs, costituisce un'unica metavariabile e ha 42 stati.

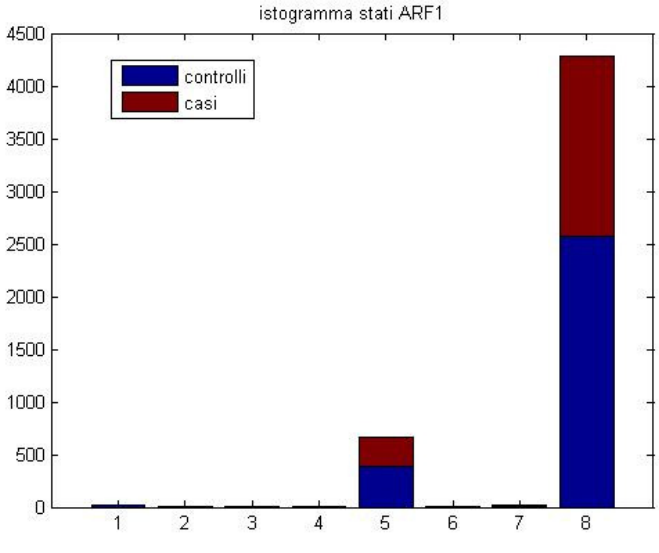
THEM4





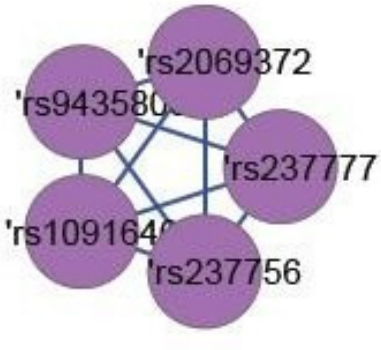
THEM4 (cromosoma 1) ha 7 SNPs, costituisce un'unica metavariabile e ha 7 stati.

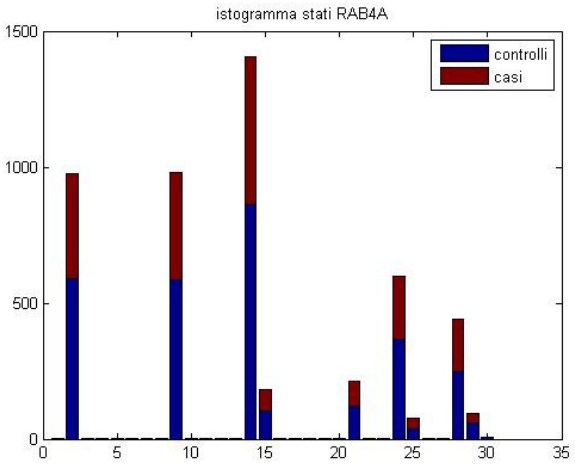
ARF1



ARF1 (cromosoma 1) ha 2 SNPs, costituisce un'unica metavariabile e ha 8 stati.

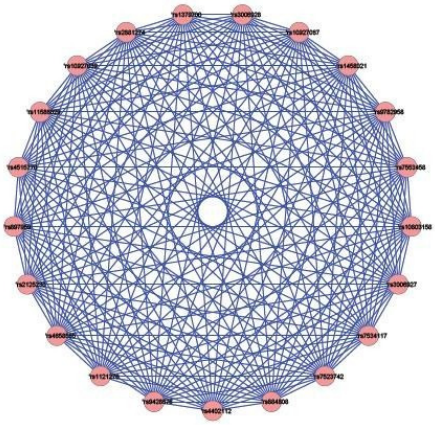
RAB4A

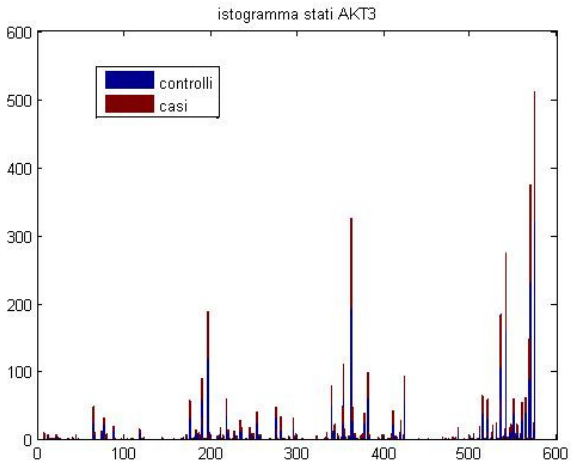




RAB4A (cromosoma 1) ha 5 SNPs, costituisce un'unica metavariabile e ha 30 stati.

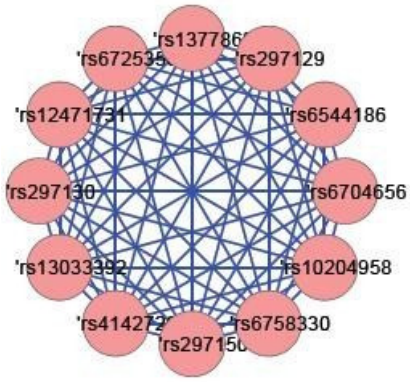
AKT3

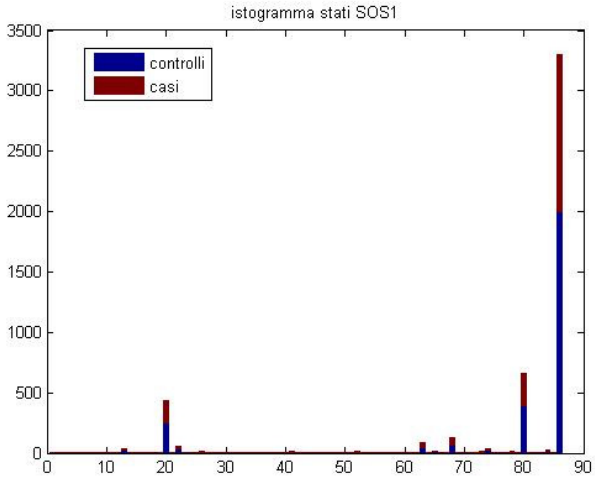




AKT3 (cromosoma 1) ha 21 SNPs, ma solo 20 selezionati nel grafico, costituisce un'unica metavariabile e ha 573 stati.

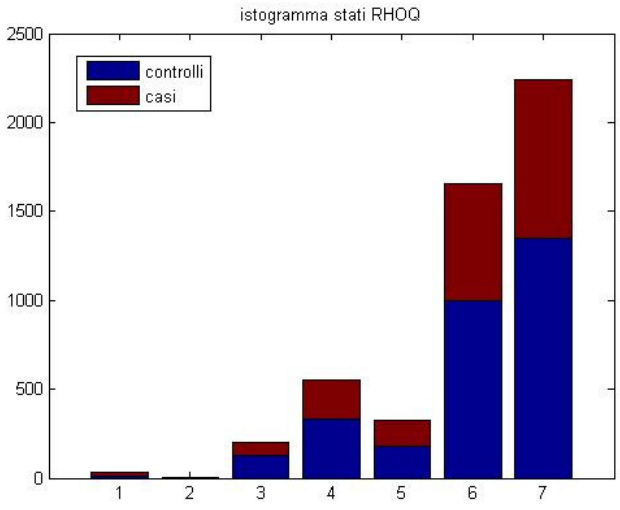
SOS1





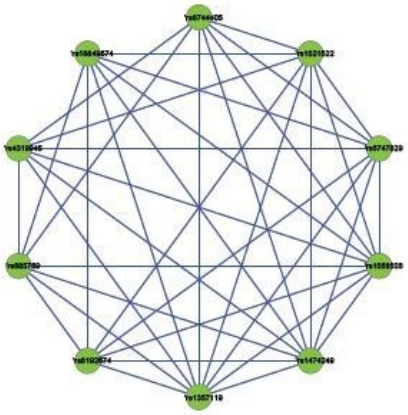
SOS1 (cromosoma 2) ha 13 SNPs, ma solo 12 selezionati nel grafico, costituisce un'unica metavariabile e ha 86 stati.

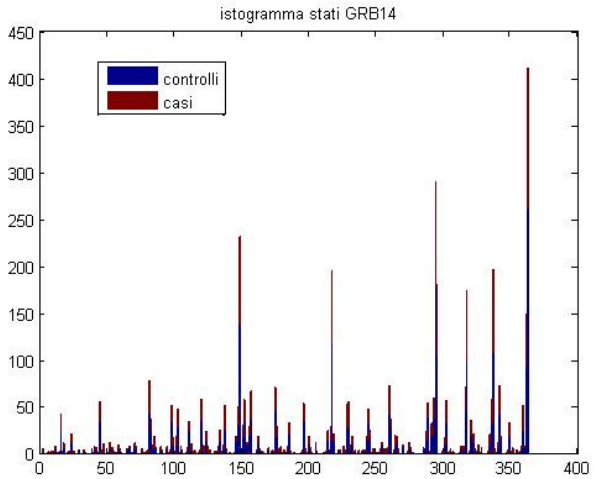
RHOQ



RHOQ (cromosoma 2), ha 2 SNPs, costituisce un'unica metavariabile e ha 7 stati.

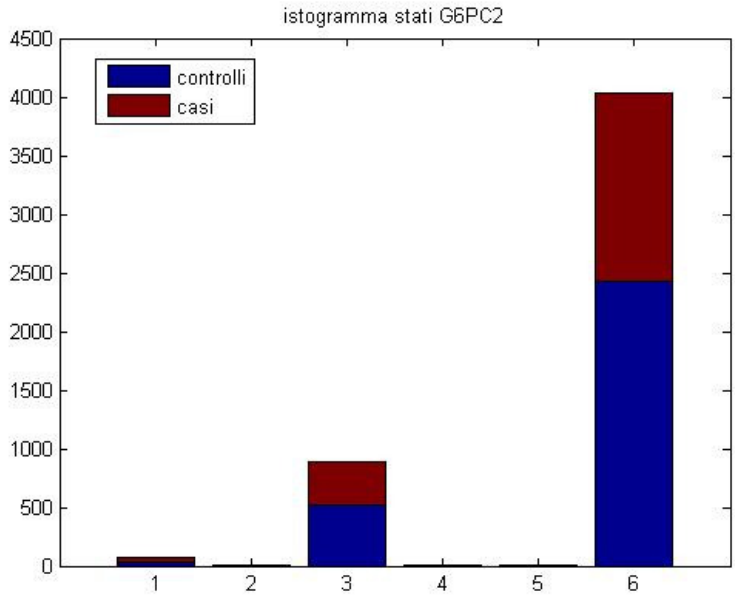
GRB14





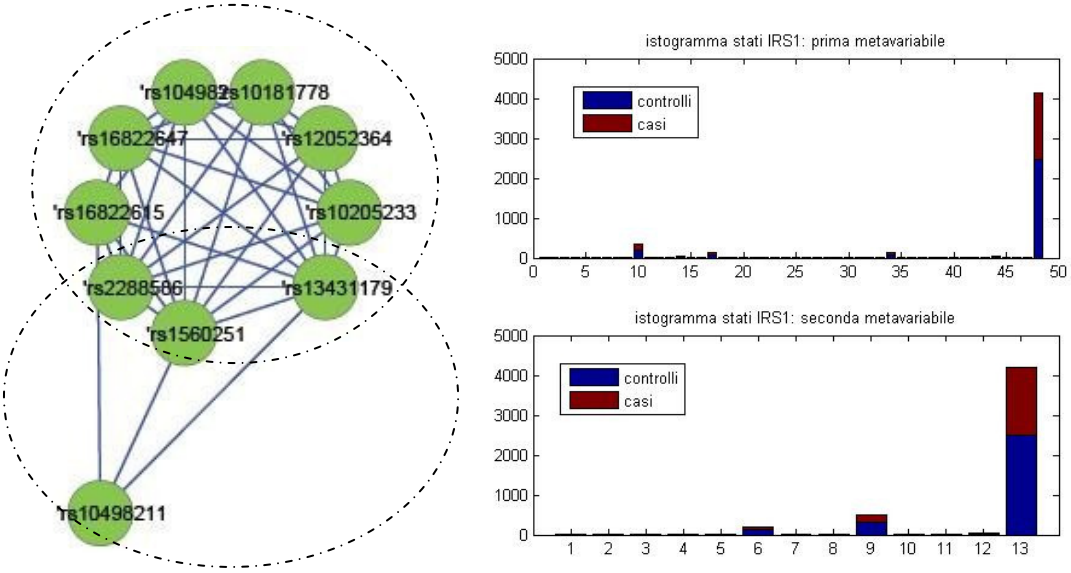
GRB14 (cromosoma 2) ha 11 SNPs, ma solo 10 riportati nel grafico, costituisce un'unica metavariabile e ha 362 stati.

G6PC



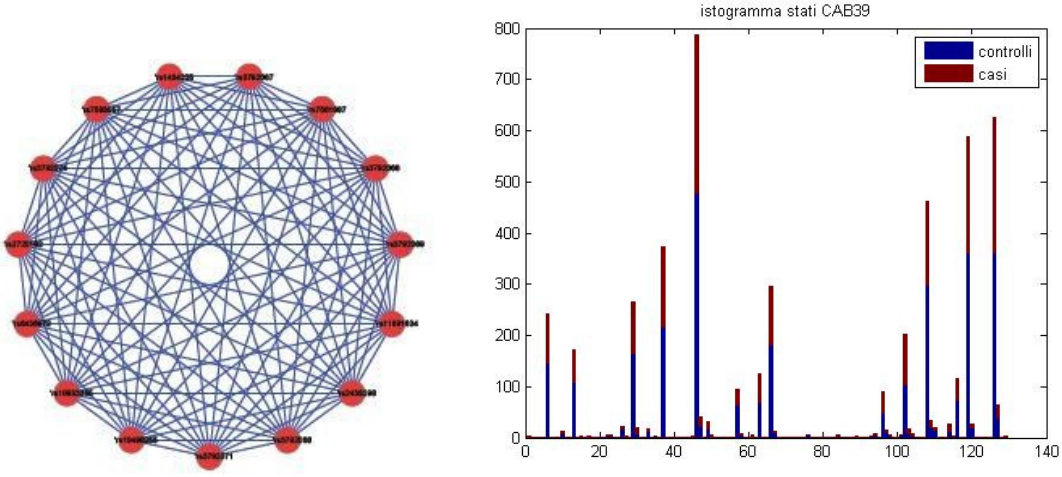
G6PC (cromosoma 2) ha 2 SNPs, costituisce un'unica metavariabile e ha 6 stati.

IRS1



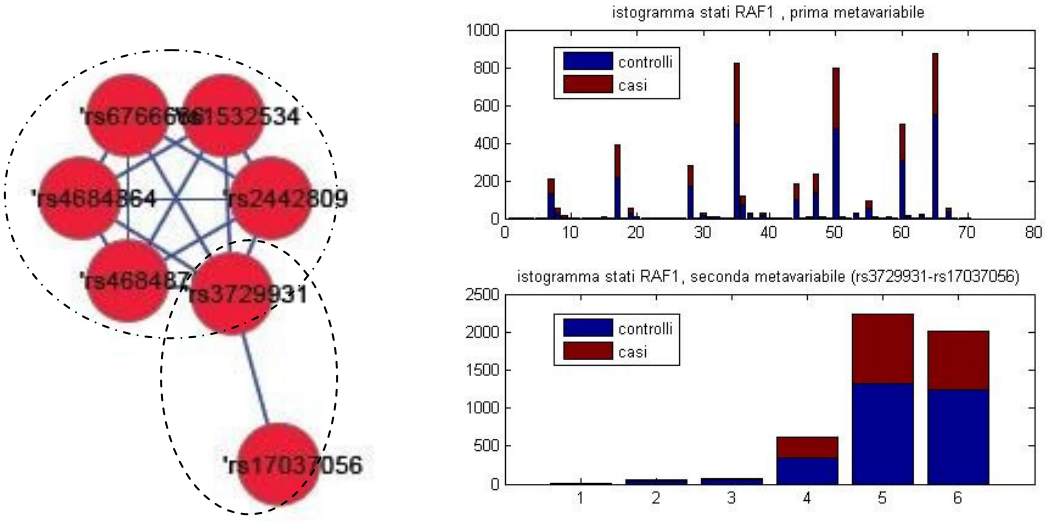
IRS1 (cromosoma 2) ha 10 SNPs e costituisce due meta variabili, la prima (9 SNPs) è rappresentata dall'istogramma e dagli SNPs nel riquadro superiori, con 48 stati. La seconda metavariabile (4 SNPs) è rappresentata dagli SNPs e dall'istogramma inferiore con 13 stati.

CAB39



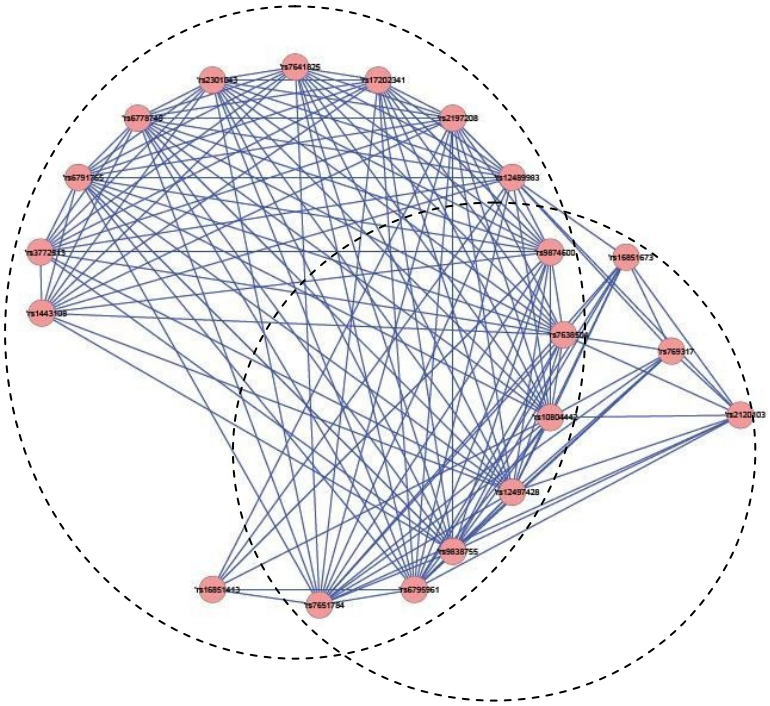
CAB39 (cromosoma 2) ha 15 SNPs e costituisce un'unica metavariabile, con 129 stati.

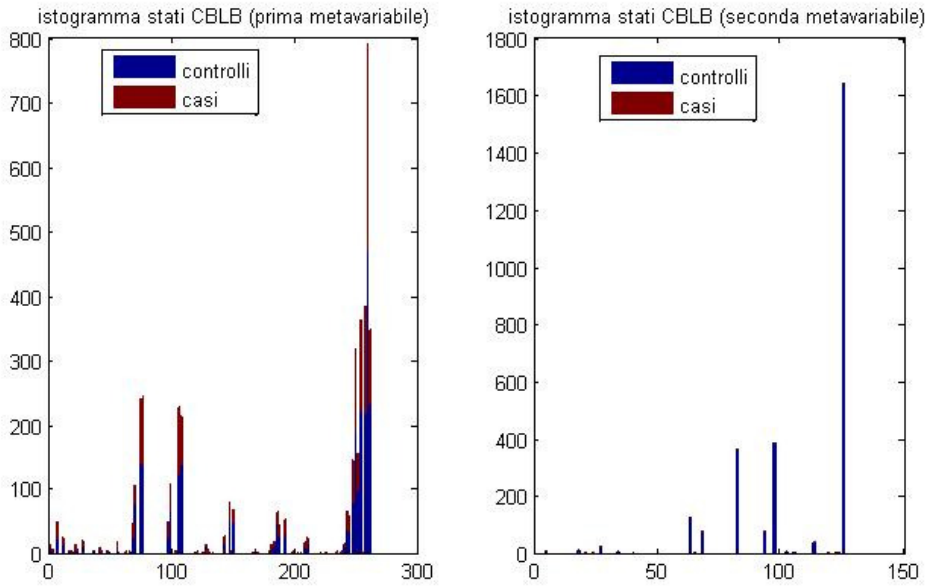
RAF1



RAF1 (cromosoma 2) ha 7 SNPs e costituisce due metavariabili, la metavariabile rappresentata superiormente ha 6 SNPs e ha 70 stati, mentre la seconda metavariabile con 2 SNPs ha 6 stati.

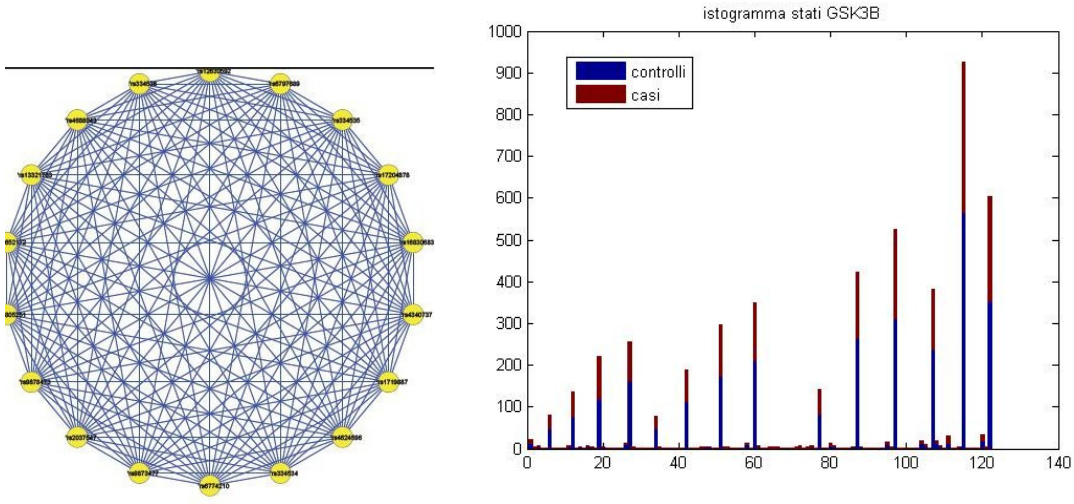
CBLB





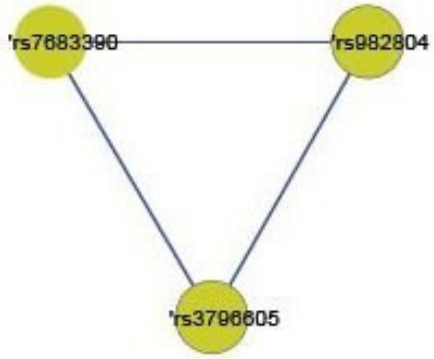
CBLB (cromosoma 3) ha 20 SNPs e si suddivide in due metavariabili, la prima ha 261 stati, mentre la seconda metavariabile ha 125 stati.

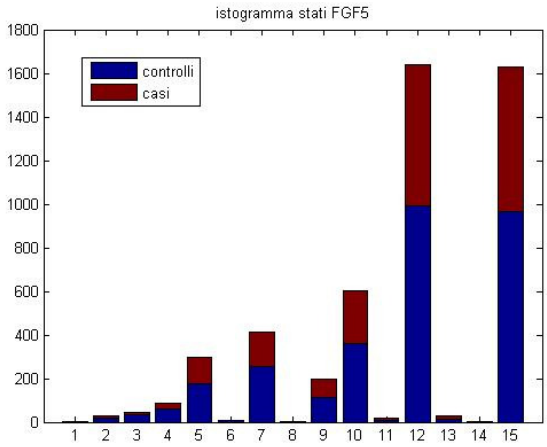
GSK3B



GSK3B (cromosoma 3) ha 18 SNPs e costituisce un'unica variabile, ha 122 stati.

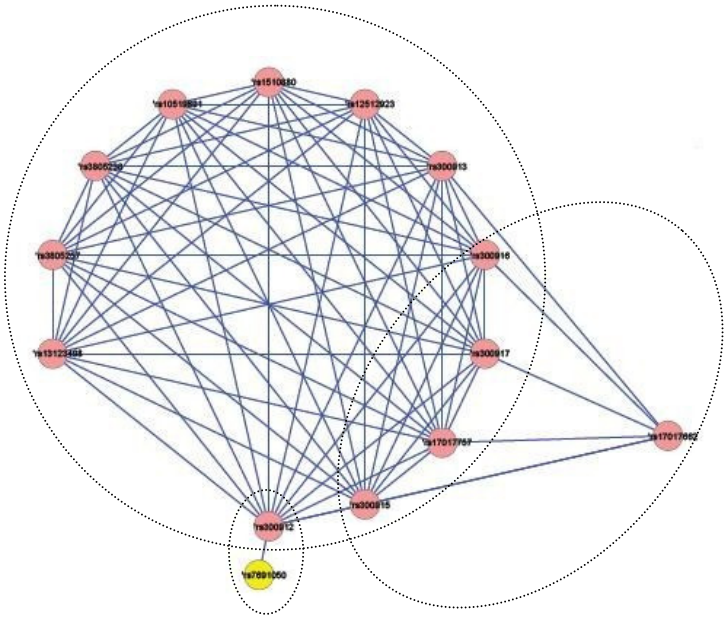
FGF5

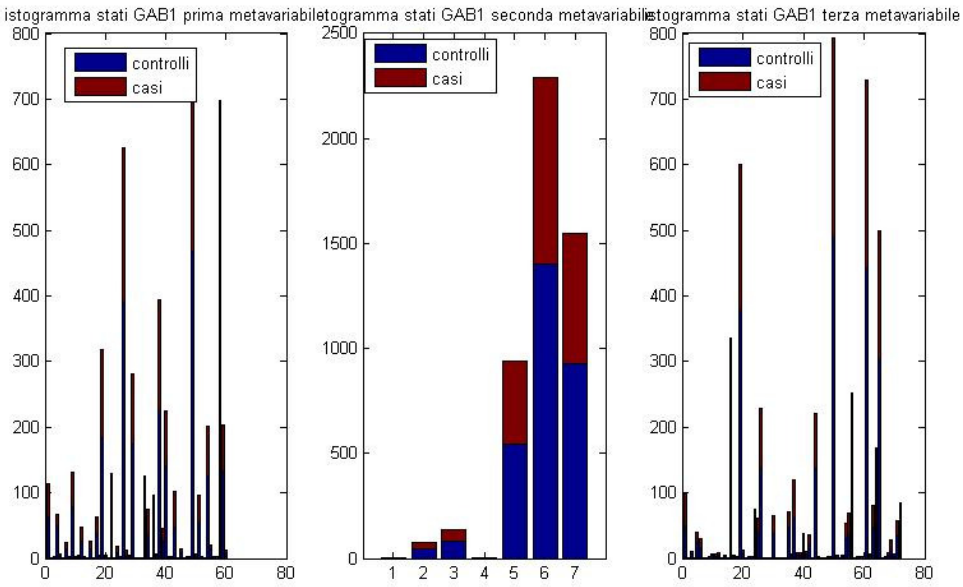




FGF5 (cromosoma 4) ha 4 SNPs, ma solo 3 sono inclusi nel grafico, costituisce un'unica metavariabile e ha 15 stati.

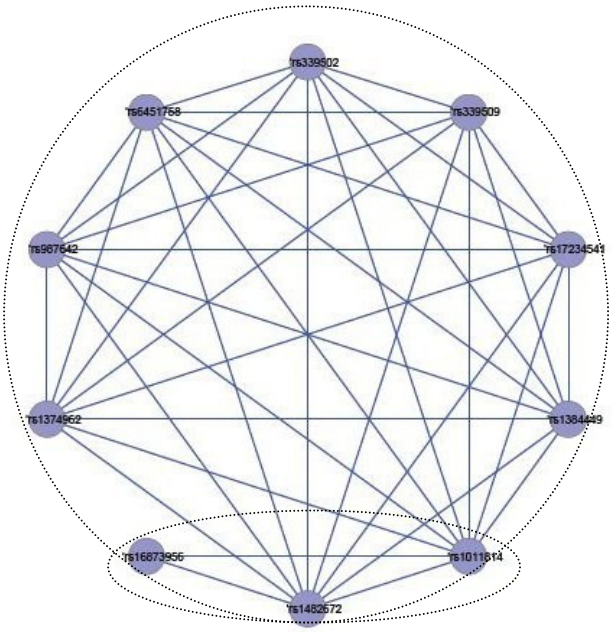
GAB1

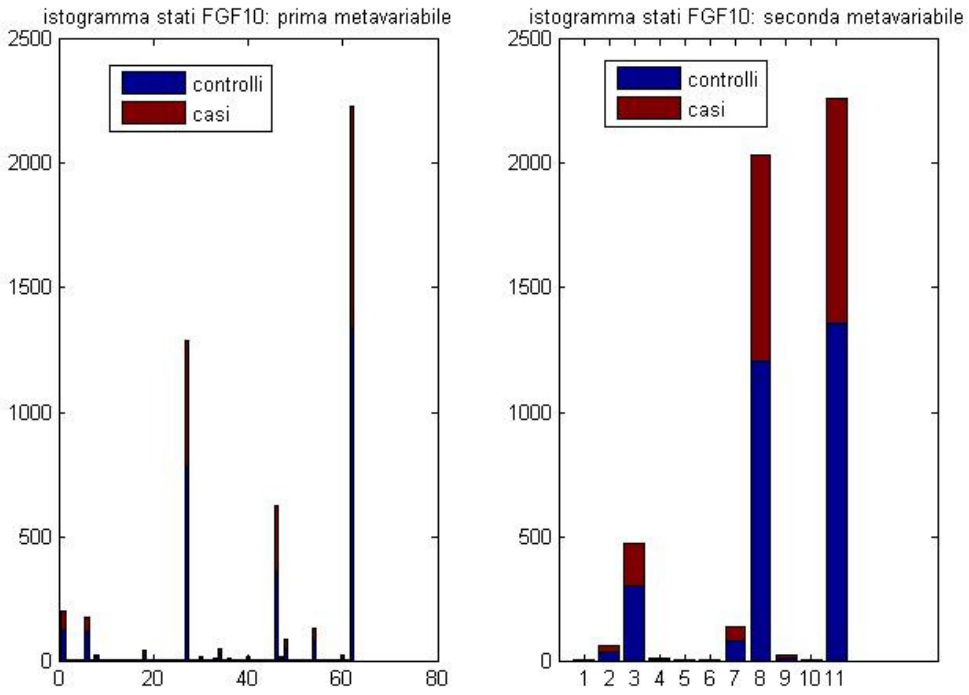




GAB1 (cromosoma 4) ha 14 SNPs e costituisce 3 metavariabili caratterizzate da 60 , 7 e 72 stati rispettivamente.

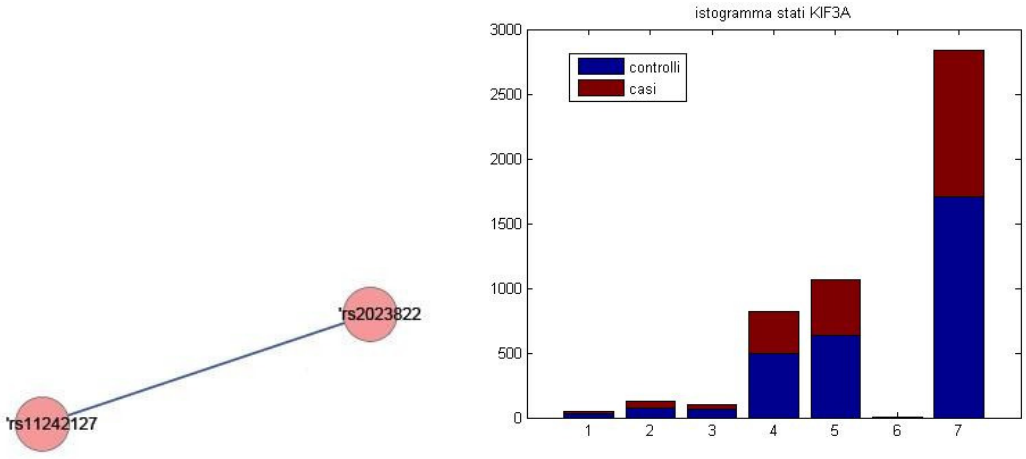
FGF10





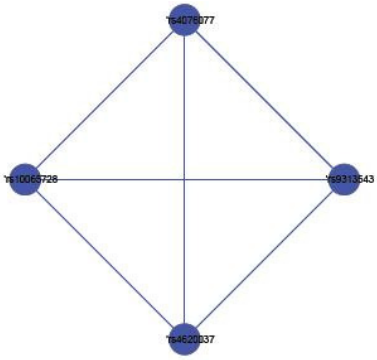
FGF10 (cromosoma 5) ha 10 SNPs e costituisce due meta variabili, con 62 e 11 stati rispettivamente.

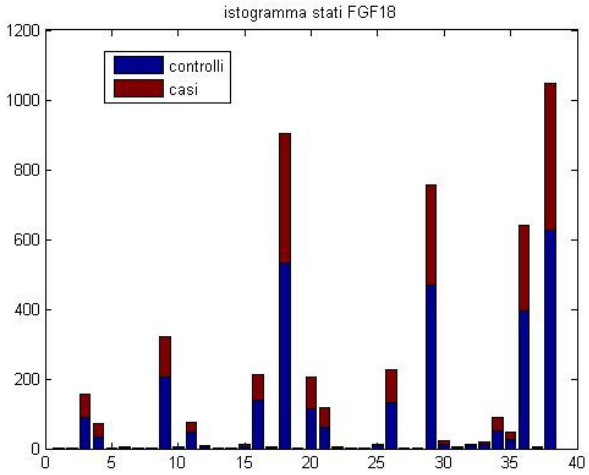
KIF3A



KIF3A (cromosoma 5) ha 2 SNPs, costituisce un'unica metavariabile e ha 7 stati.

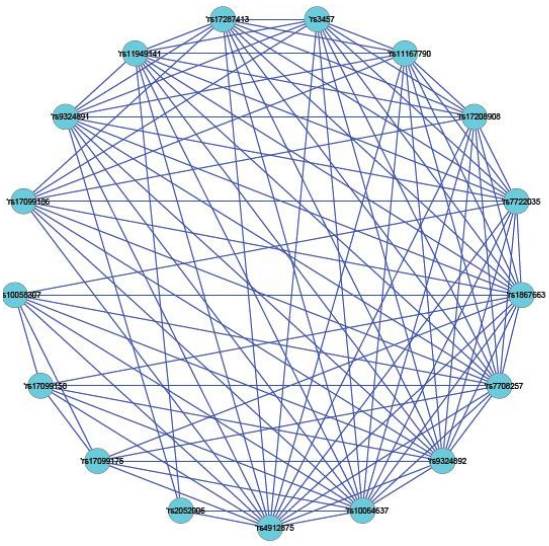
FGF18

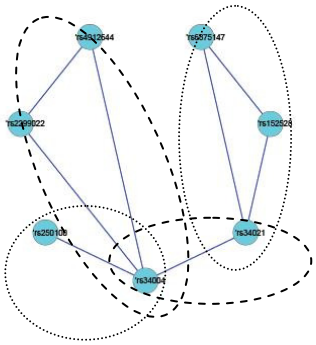


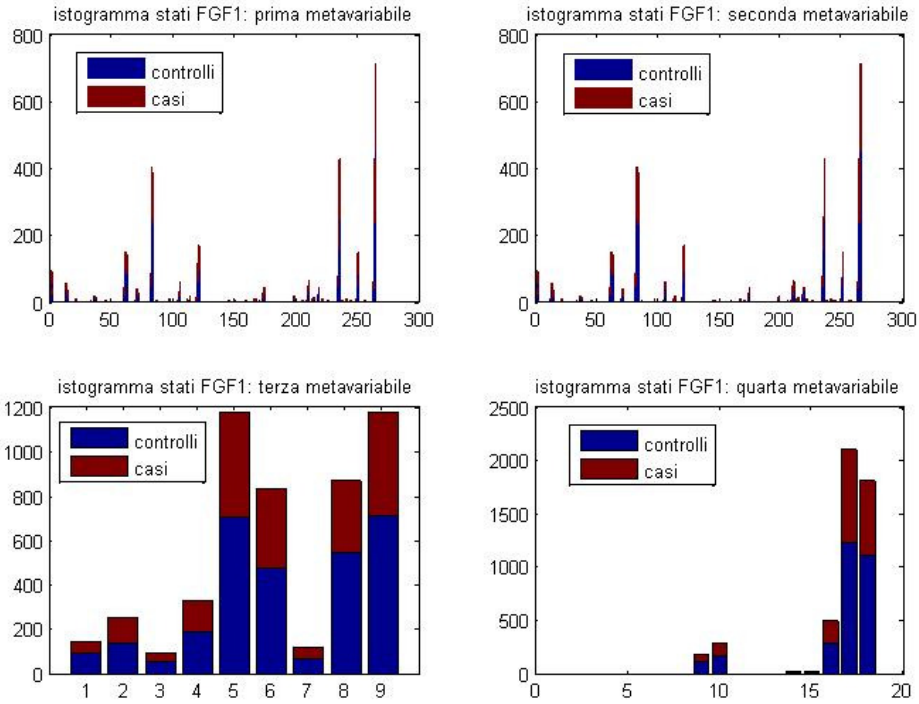


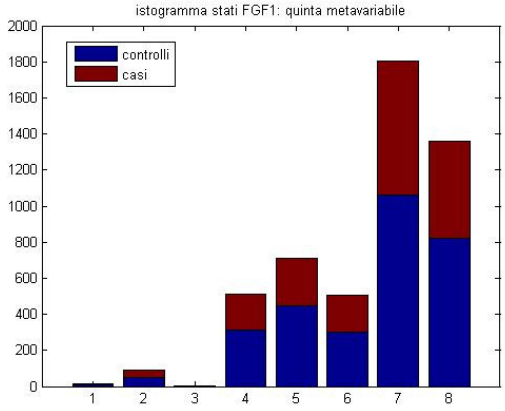
FGF18 (cromosoma 5) ha 4 SNPs e costituisce un'unica metavariabile con 38 stati.

FGF1



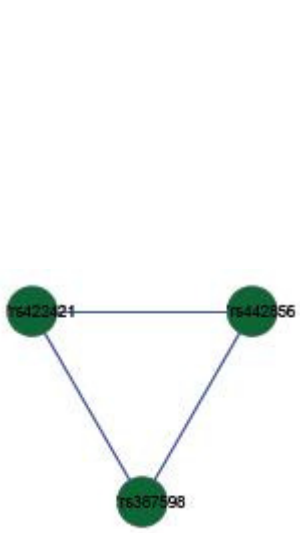


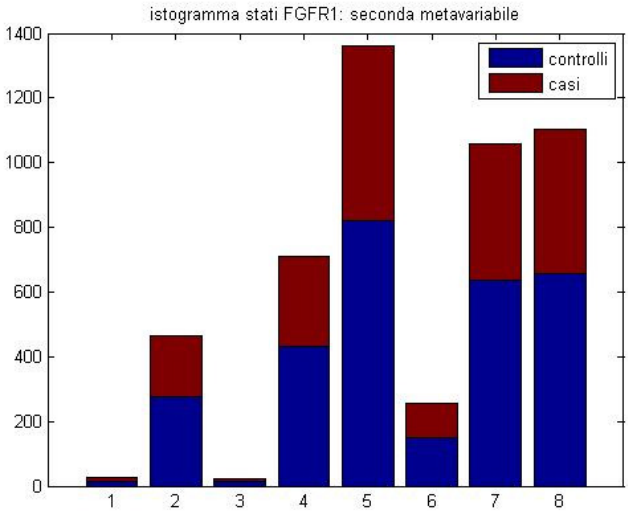




FGF1 (cromosoma 5) ha 26 SNPs in totale divisi in due gruppi separati (situazione dovuta presumibilmente alla presenza di due aplotipi separati all'interno del gene e di un punto di ricombinazione tra essi). Costituisce 5 metavariabili con numero di stati pari a 264, 22, 9, 18, e 8 stati.

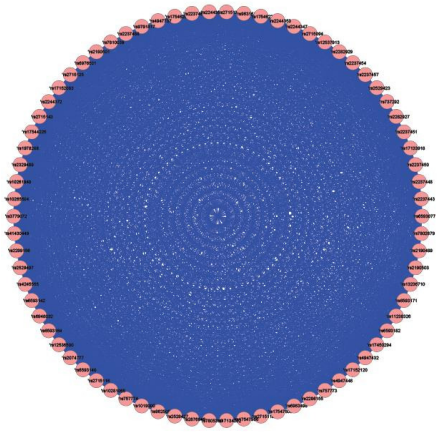
FGFR4

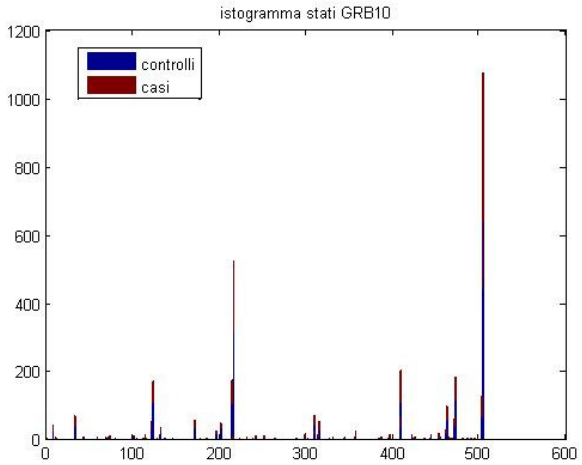




FGFR4 (cromosoma 5) ha 3 SNPs e costituisce un'unica metavariabile, con 8 stati.

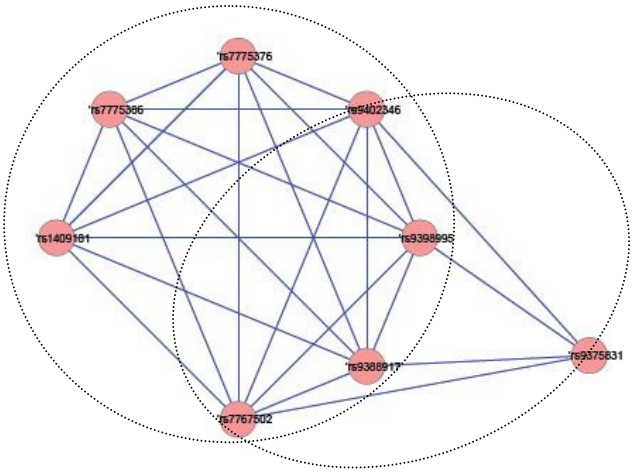
GRB10

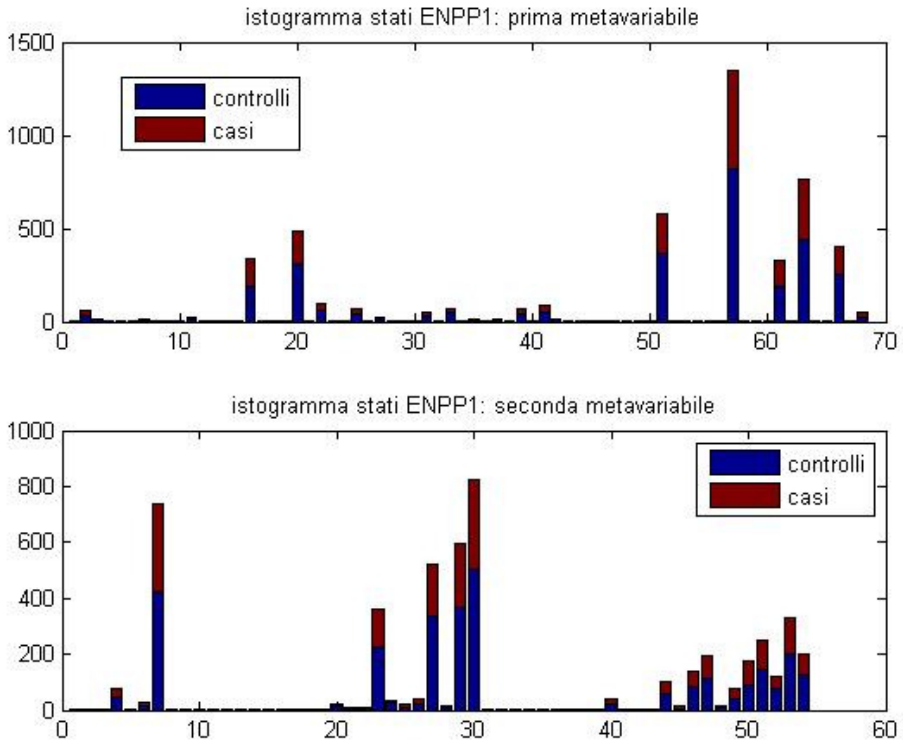




GRB10 (cromosoma 7) ha 74 SNPs ma solo 72 vengono selezionati nella rete. E' un gene della rete fortemente connesso e costituisce un'unica metavariabile con 504 stati, anche se, dall'istogramma, si può notare come solo alcuni di questi stati siano presenti in maniera preponderante rispetto agli altri.

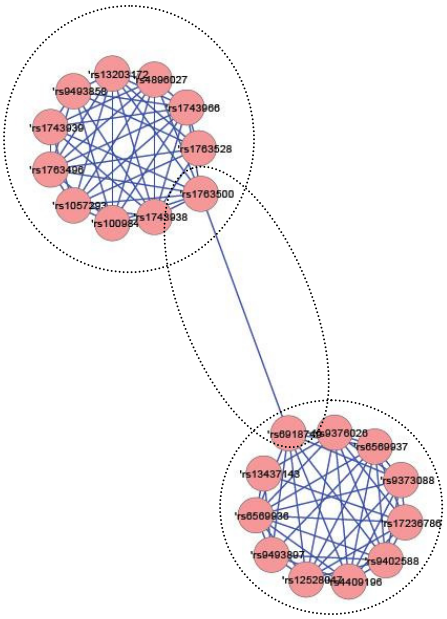
ENPP1

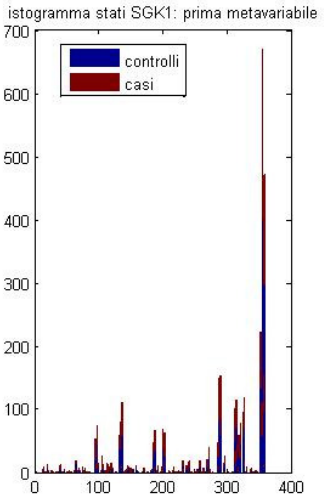


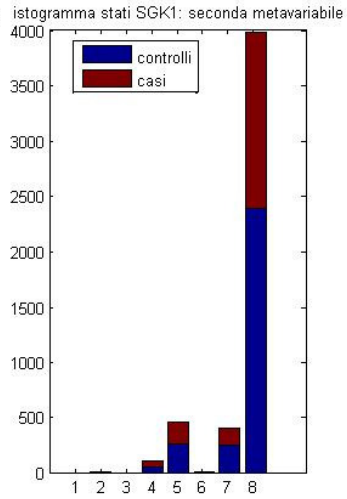


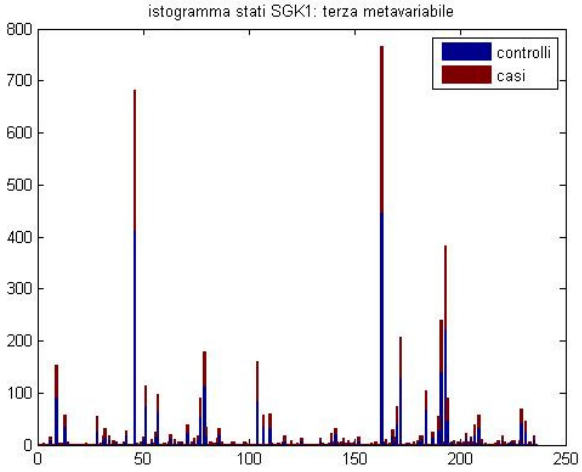
ENPP1 (cromosoma 6) ha 8 SNPs, costituisce due metavariabili che hanno 68 e 54 stati rispettivamente.

SGK1


rs1320372
rs4896027
rs9493858
rs1743966
rs1743930
rs1763528
rs1763498
rs1763500
rs1057293
rs1743938
rs10098
rs9376028
rs6569937
rs13437143
rs9373088
rs6569936
rs17236786
rs9493897
rs9402588
rs12528047
rs4409196

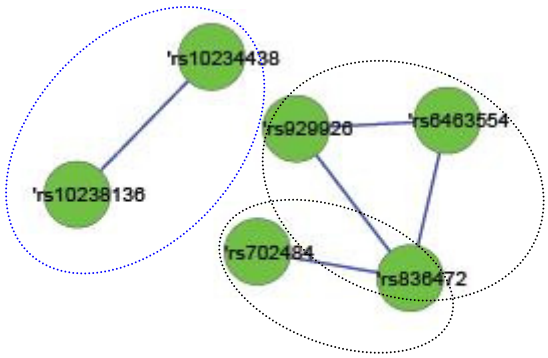

istogramma stati SGK1: prima metavariabile
controlli
casi
0
100
200
300
400
500
600
700

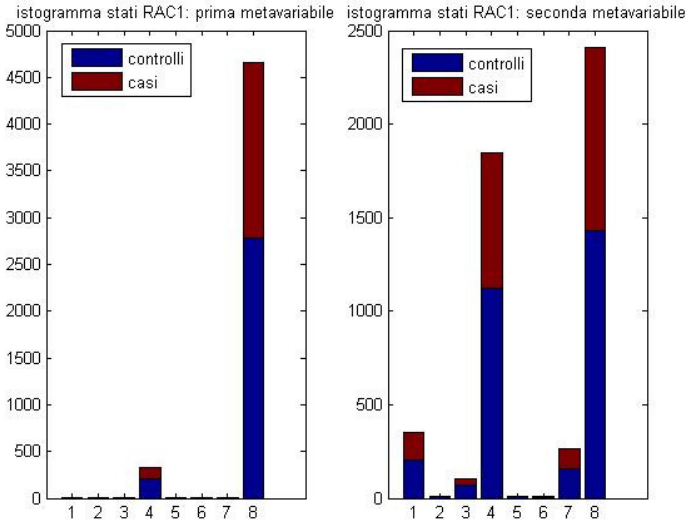

istogramma stati SGK1: seconda metavariabile
controlli
casi
0
500
1000
1500
2000
2500
3000
3500
4000
1
2
3
4
5
6
7
8

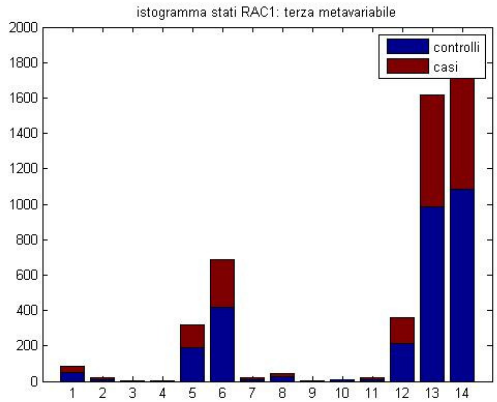

istogramma stati SGK1: terza metavariabile
controlli
casi
0
100
200
300
400
500
600
700
800
0
50
100
150
200
250

SGK1 (cromosoma 6) ha 22 SNPs che costituiscono tre metavariabili, si 357, 8 e 236 stati .

RAC1

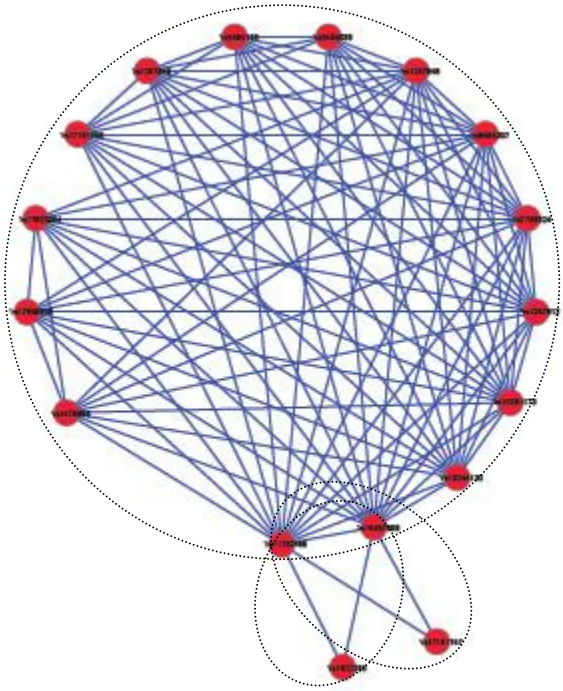


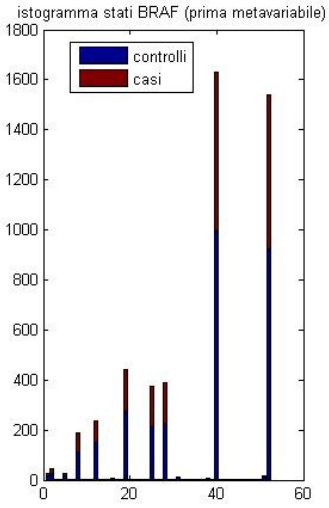


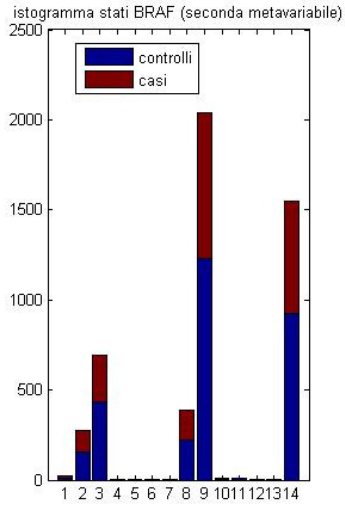


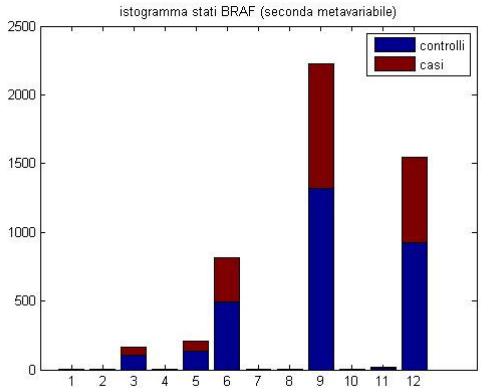
RAC1 (cromosoma 7) ha 6 SNPs e costituisce tre metavariabili di 8 e 14 stati rispettivamente.

BRAF



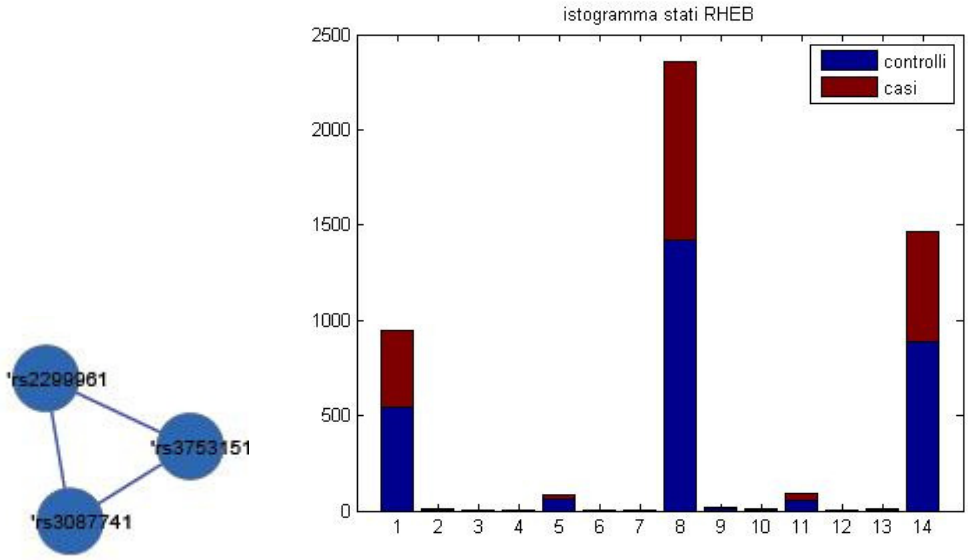

istogramma stati BRAF (prima metavariabile)
controlli
casi
0
200
400
600
800
1000
1200
1400
1600
1800
0
20
40
60


istogramma stati BRAF (seconda metavariabile)
controlli
casi
0
500
1000
1500
2000
2500
1 2 3 4 5 6 7 8 9 10 11 12 13 14


istogramma stati BRAF (seconda metavariabile)
controlli
casi
0
500
1000
1500
2000
2500
1 2 3 4 5 6 7 8 9 10 11 12

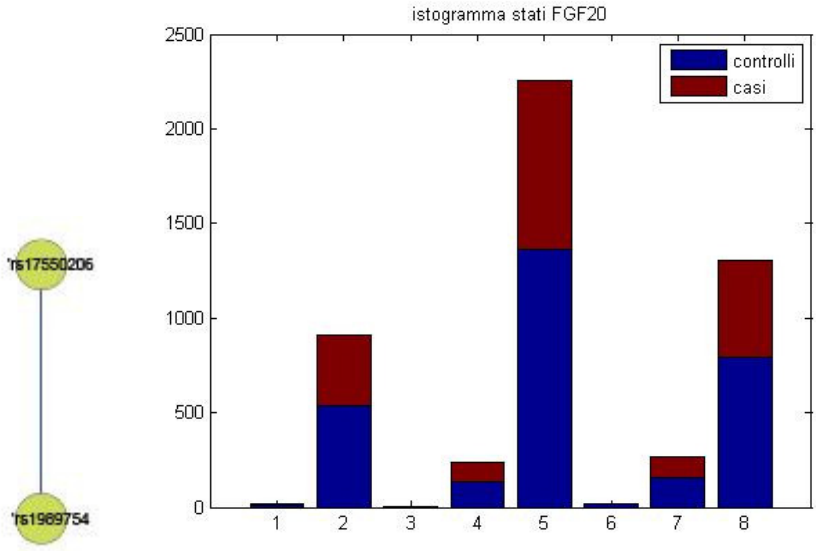
BRAF (cromosoma 7) ha 17 SNPs organizzati in tre metavariabili di 52, 14 e 12 stati.

RHEB



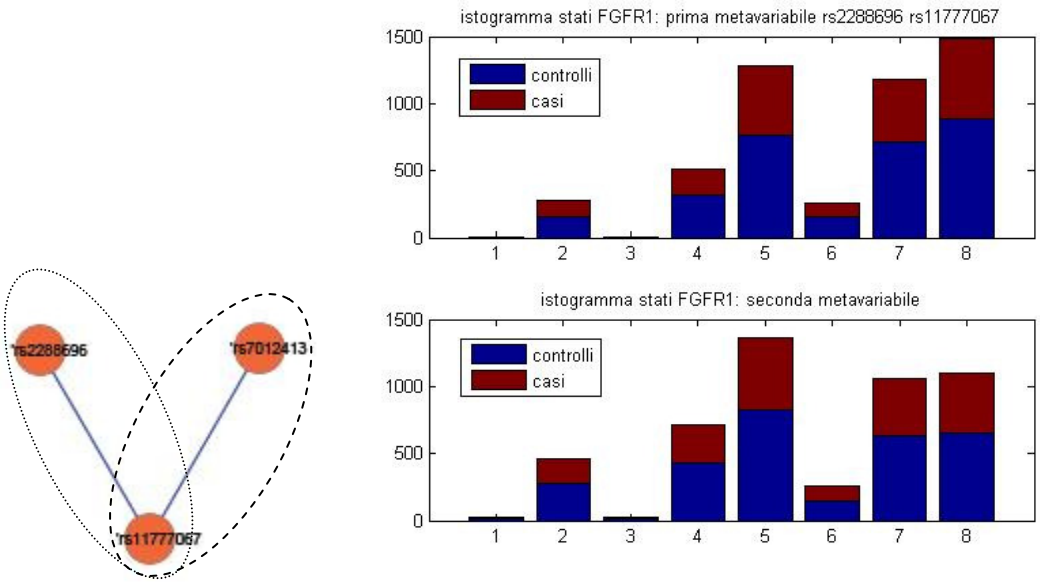
RHEB (cromosoma 7) ha 7 SNPs, costituisce un'unica metavariabile con 14 stati.

FGF20



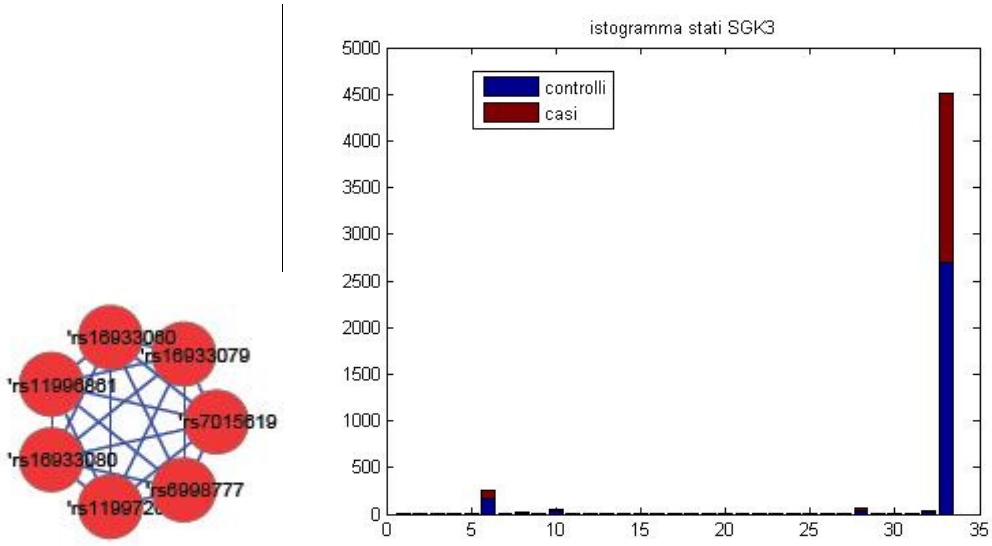
FGF20 (cromosoma 8) ha 2 SNPs e costituisce un'unica metavariabile con 8 stati.

FGFR1



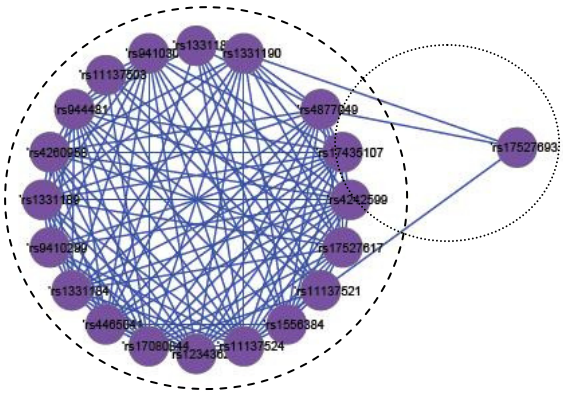
FGFR1 (cromosoma 8) ha 3 SNPs e costituisce due metavariabili di 8 stati ciascuna.

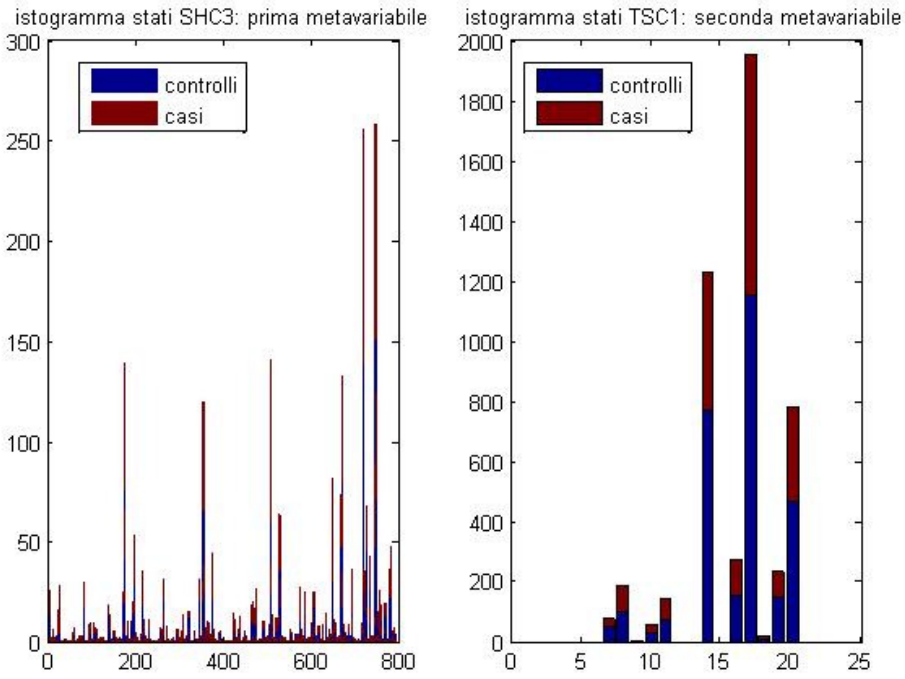
SGK3



SGK3 (cromosoma 8) ha 7 SNPs, costituisce un'unca metavariabile con 33 stati (dall'istogramma si può notare come uno stato sia nettamente prevalente sugli altri).

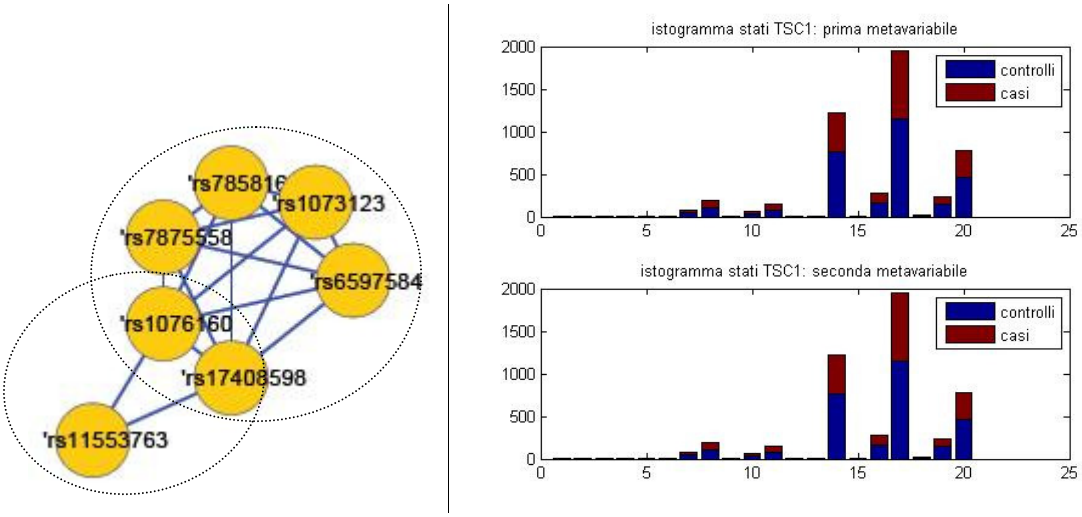
SHC3





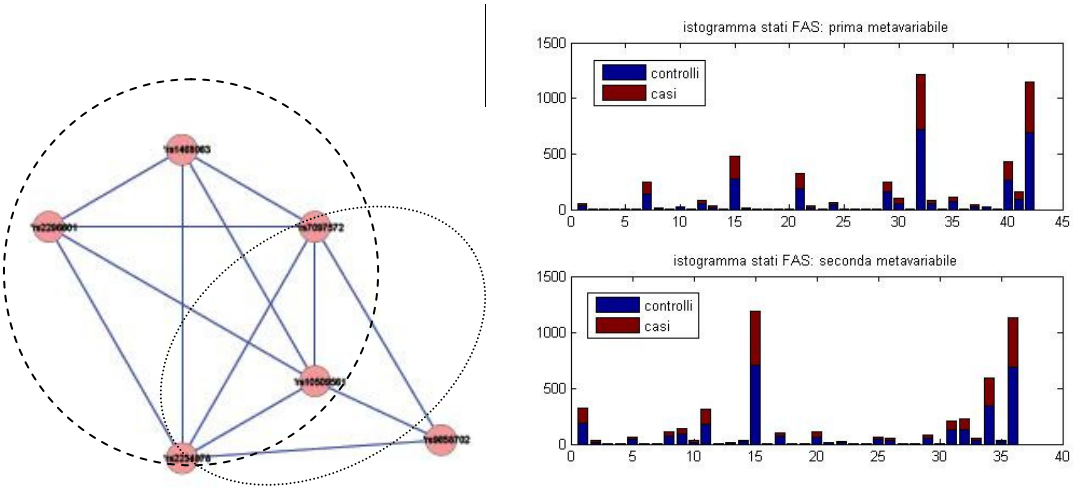
SCH3 (cromosoma 9) ha 21 SNPs, ma uno è escluso dalla rete, costituisce due metavariabili di 792 e 21 stati rispettivamente.

TSC1



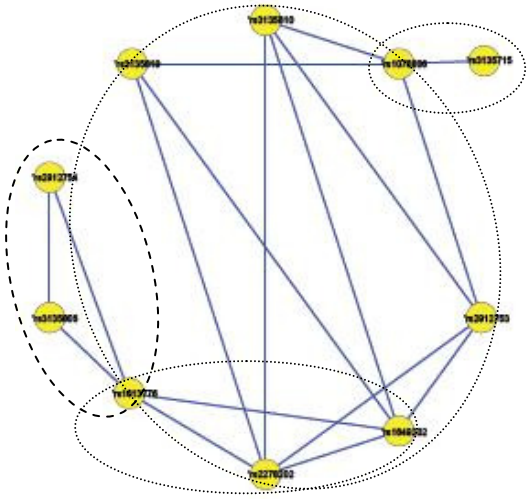
TSC1 (cromosoma 9) ha 7 SNPs, e costituisce due metavariabili di 20 stati ciascuna.

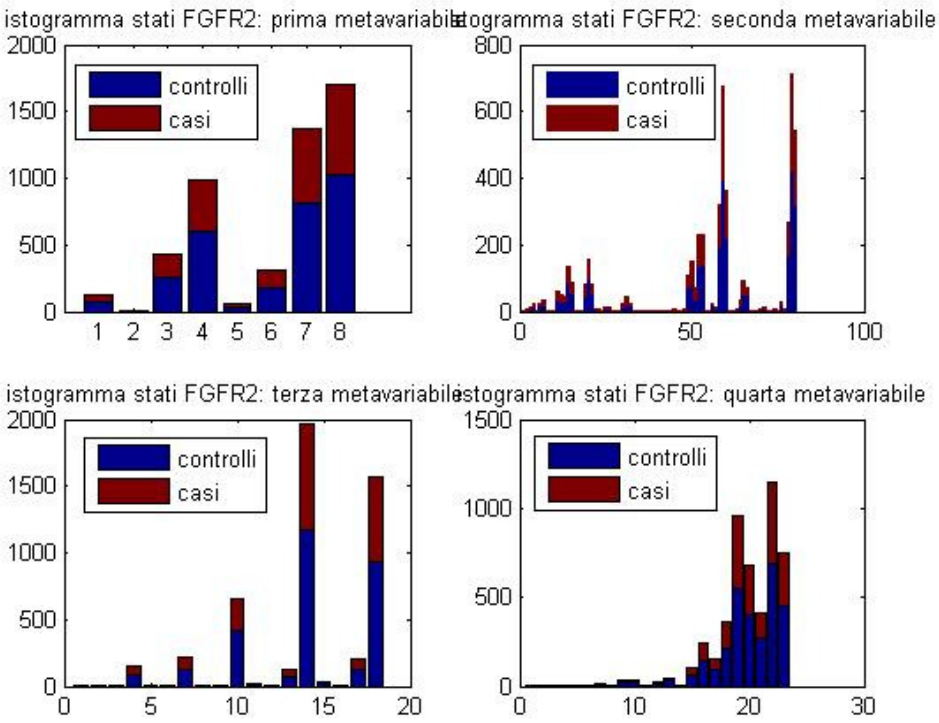
FAS



FAS (cromosoma 10) ha 6 SNPs e costituisce due metavariabili di 42 e 36 stati rispettivamente.

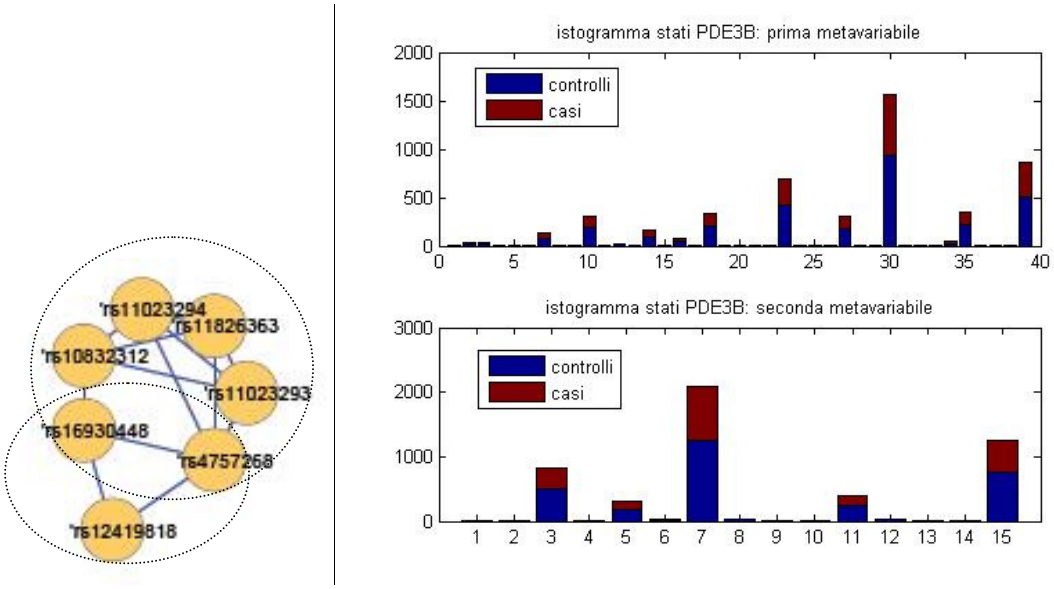
FGFR2





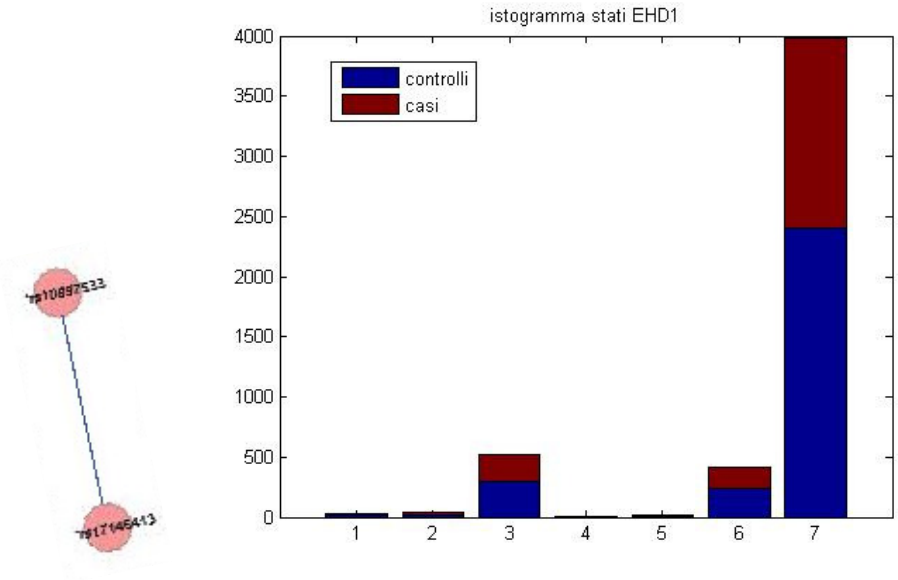
FGFR2 (cromosoma 10) ha 10 SNPs organizzati in quattro metavariabili di 8, 80, 18 e 23 stati ciascuna.

PDE3B



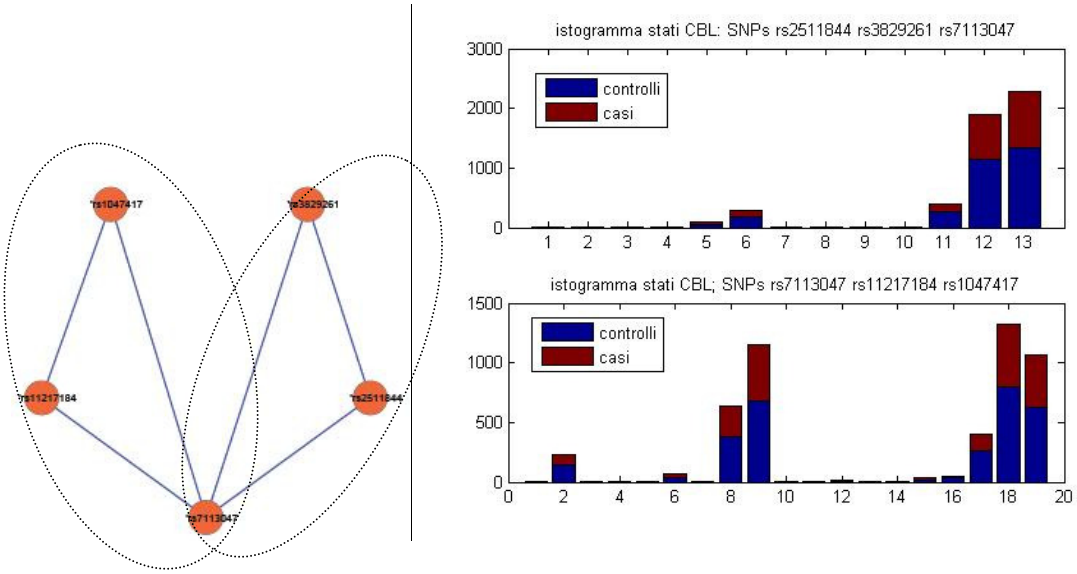
PDE3B (cromosoma 11) ha 7 SNPs e costituisce due metavariabili di 39 e 15 stati rispettivamente.

EHD1



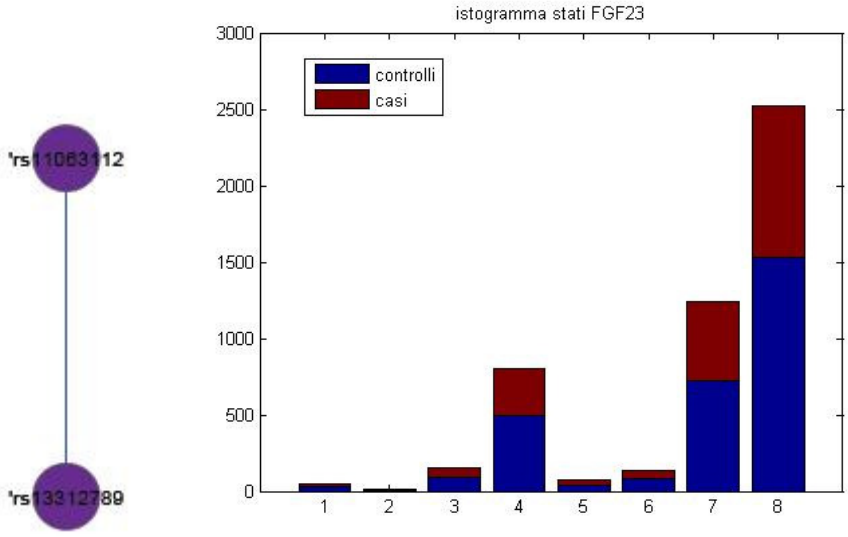
EHD1 (cromosoma 11) ha 2 SNPs e costituisce un'unica metavariabile con 7 stati.

CBL



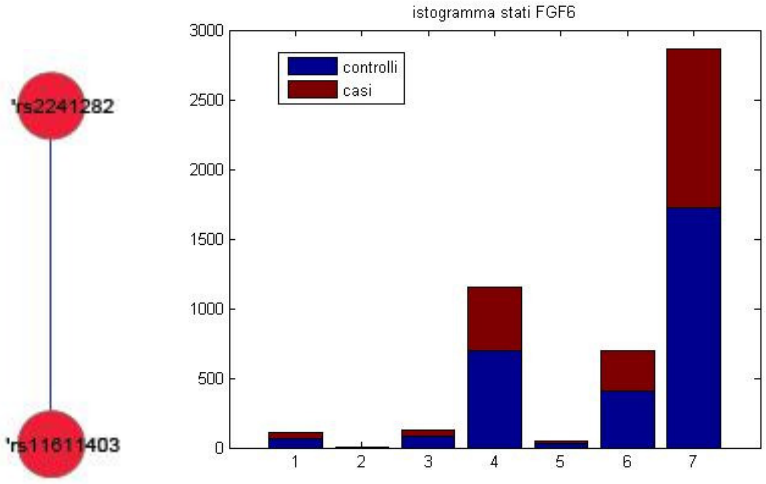
CBL (cromosoma 11) ha 5 SNPs e costituisce due metavariabili con 13 e 19 stati rispettivamente.

FGF23



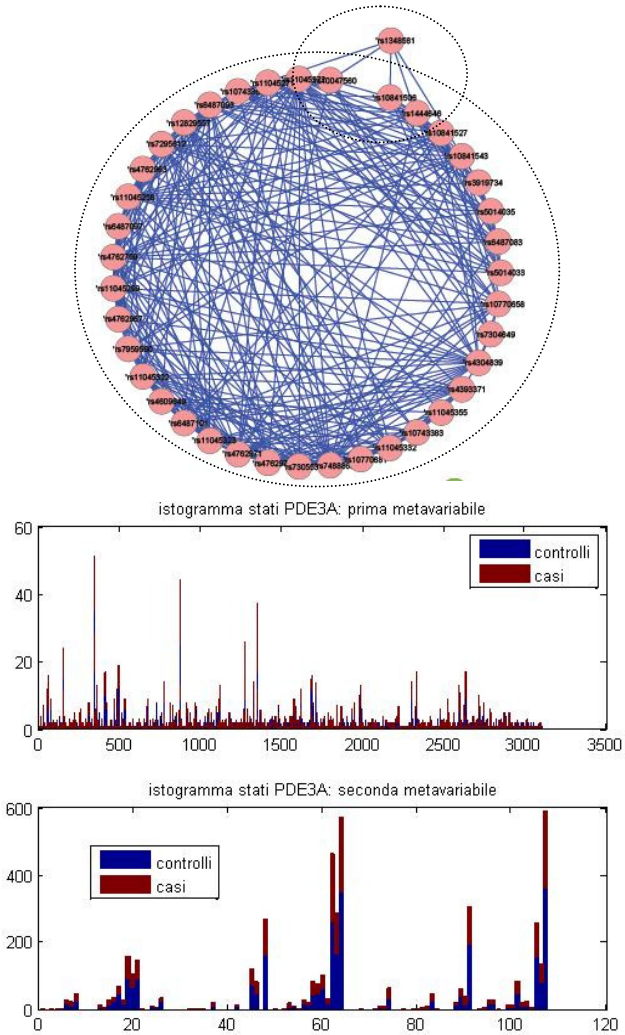
FGF23 (cromosoma 12) ha 2 SNPs, costituisce un'unica metavariabile e ha 8 stati.

FGF6



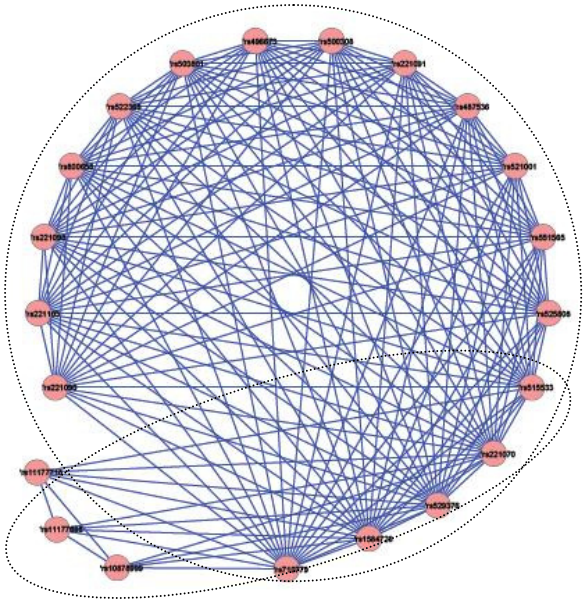
FGF6 (cromosoma 12) ha due SNPs, costituisce un'unica metavariabile con 7 stati.

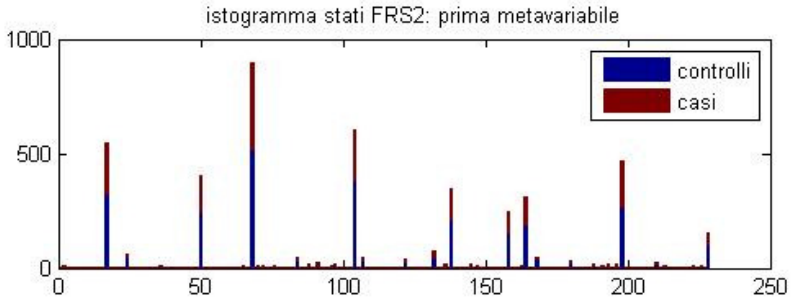
PDE3A

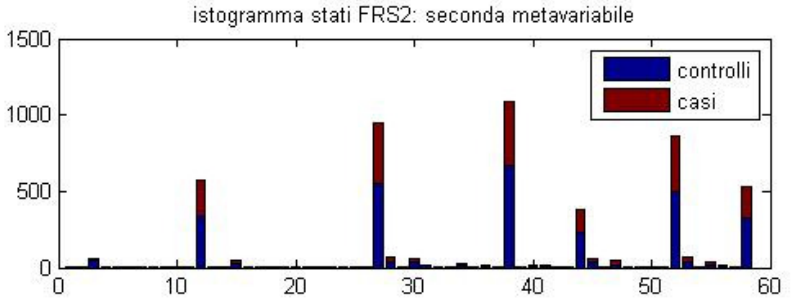


PDE3A (cromosoma 12) ha 40 SNPs e costituisce due meta variabili con 3095 e 170 stati rispettivamente.

FRS2

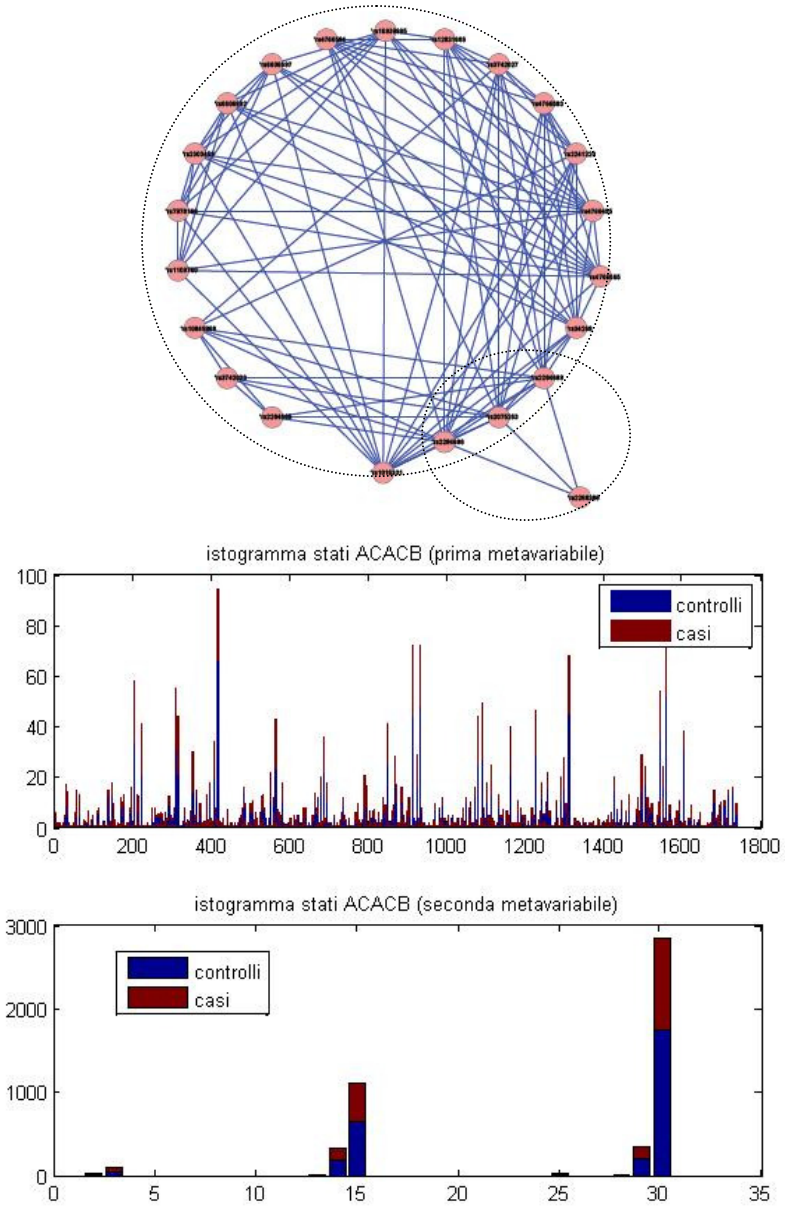






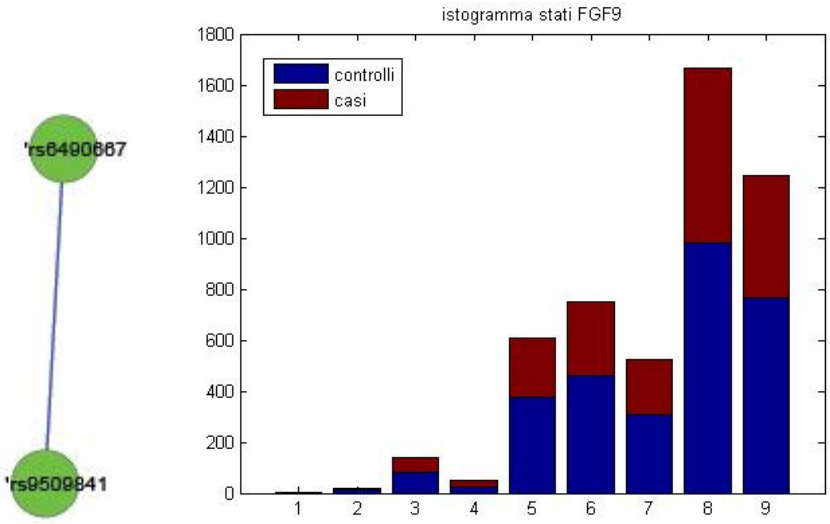
FRS2 (cromosoma 12) ha 21 SNPs, costituisce due metavariabili con 228 e 58 stati rispettivamente.

ACACB



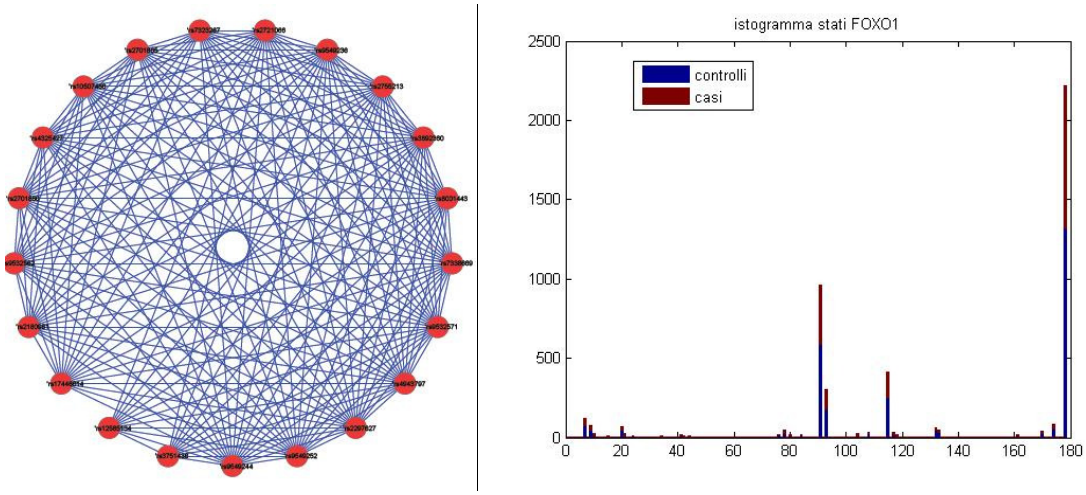
ACACB (cromosoma 12) ha 22 SNPs e costituisce due metavariabili, di 1732 e 30 stati rispettivamente.

FGF9



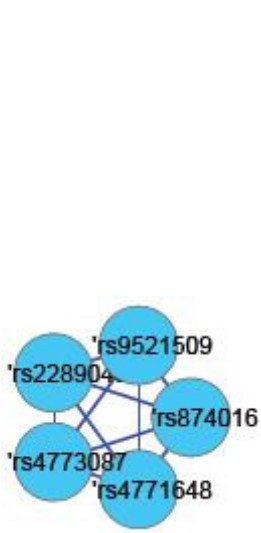
FGF9 (cromosoma 13) ha due SNPs, e costituisce un'unica metavariabile con 9 stati.

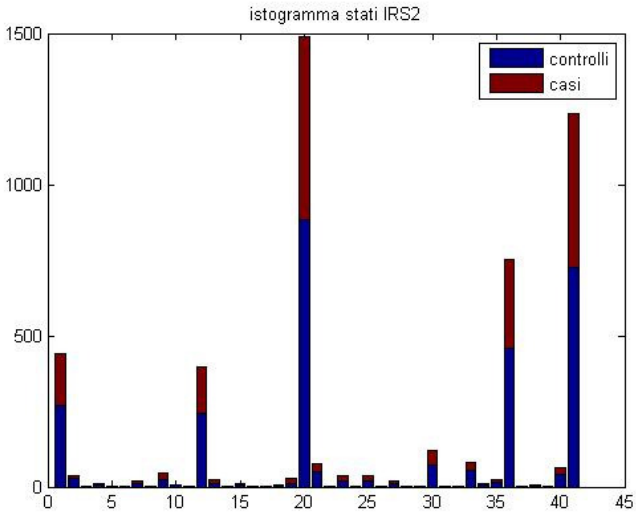
FOXO1



FOXO1 (cromosoma 13) ha 21 SNPs e costituisce un'unica metavariabile con 179 stati.

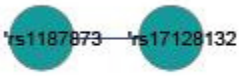
IRS2

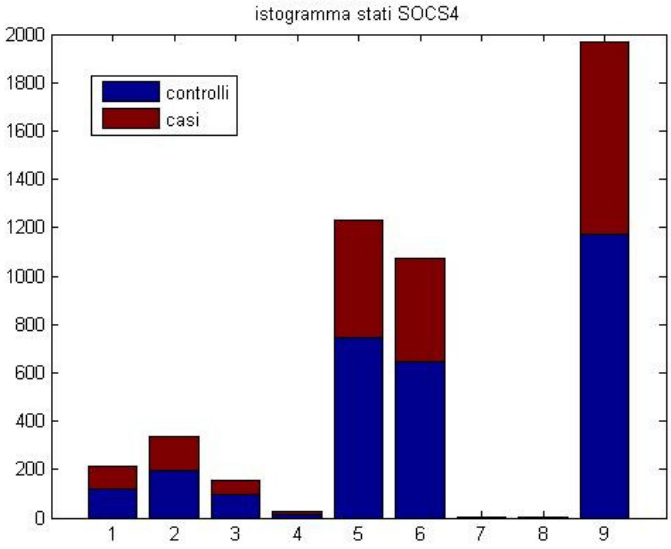




IRS2 (cromosoma 13) ha 5 SNPs, costituisce un'unica metavariabile con 41 stati.

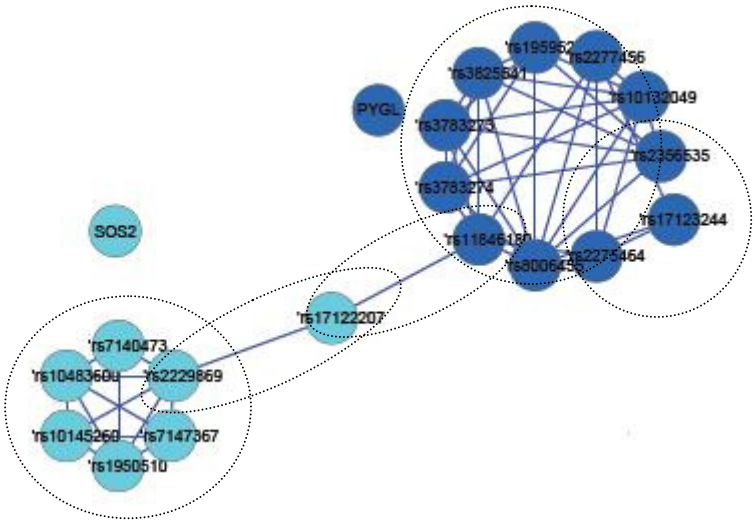
SOCS4

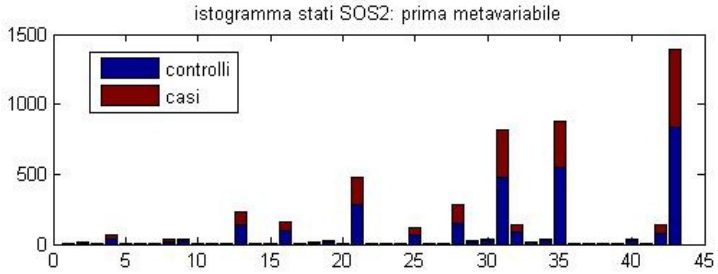


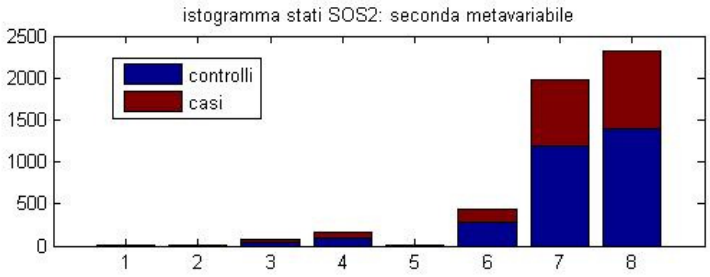


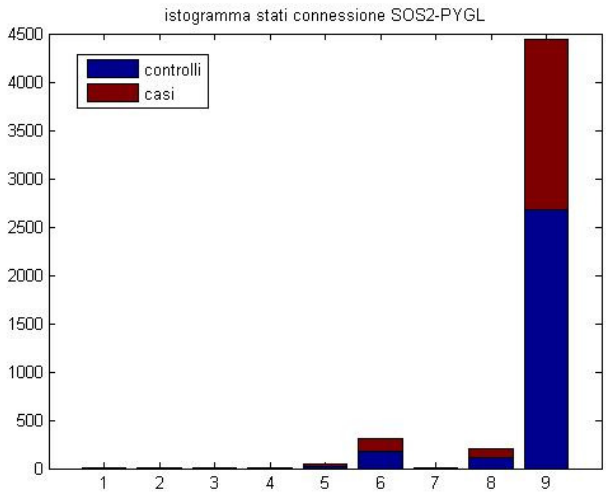
SOCS4 (cromosoma 14) ha 2 SNPs e costituisce un'unica metavariabile con 9 stati.

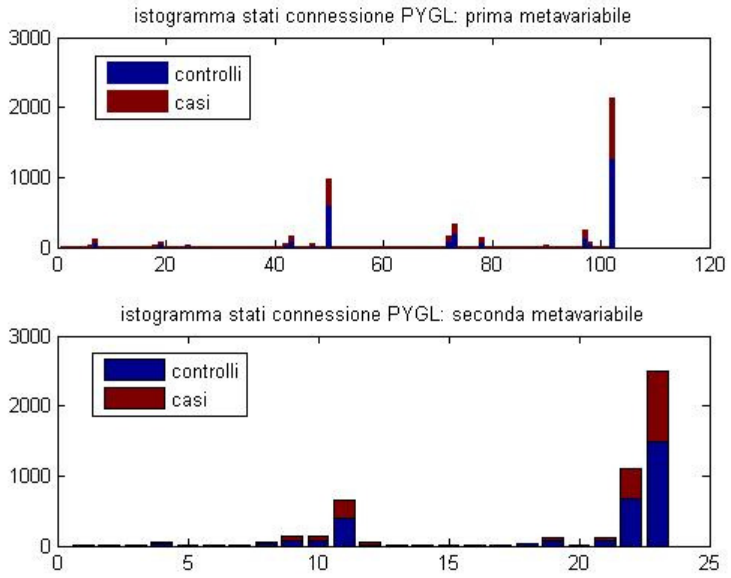
SOS2 - PYGL


SOS2
PYGL
rs17122207
rs7140473
rs2229869
rs7147367
rs1950510
rs17123244


istogramma stati SOS2: prima metavariabile
controlli
casi
0
500
1000
1500
0
5
10
15
20
25
30
35
40
45

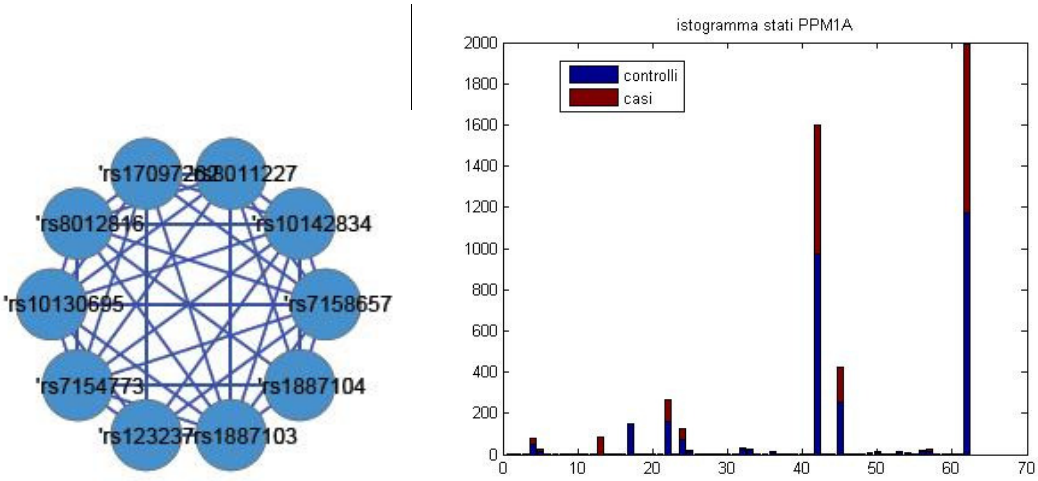

istogramma stati SOS2: seconda metavariabile
controlli
casi
0
500
1000
1500
2000
2500
1
2
3
4
5
6
7
8


istogramma stati connessione SOS2-PYGL
controlli
casi
0
500
1000
1500
2000
2500
3000
3500
4000
4500
1
2
3
4
5
6
7
8
9



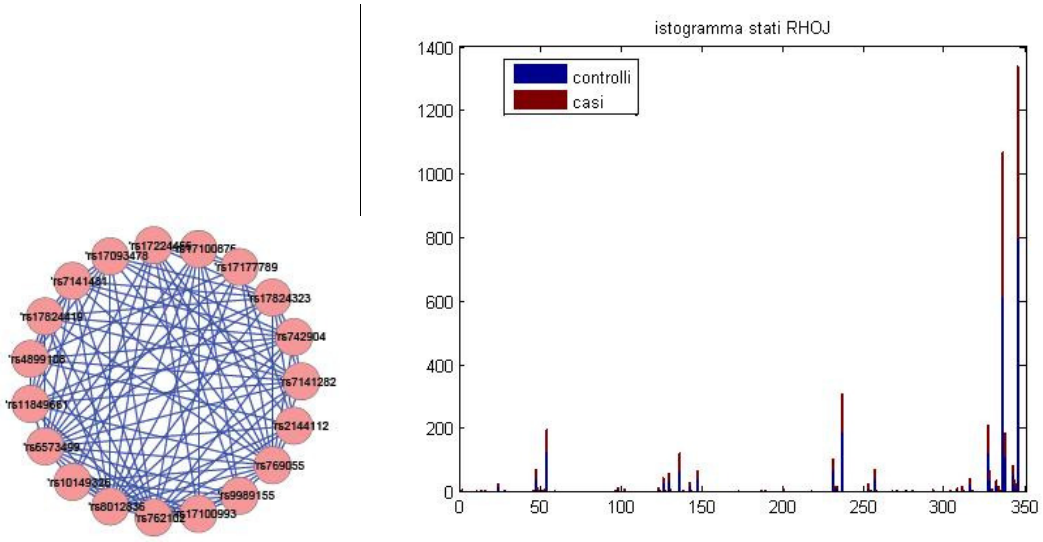
Qui vengono rappresentati due geni, SOS2 e PYGL, che risultano legati tra di loro mediante una connessione tra due SNPs. SOS2 si trova sul cromosoma 14, menre PYGL si trova sul cromosoma 11. SOS2 è costituito da due metavariabili (di 43 e 8 stati), e così pure PYGL (di 102e 23 stati). Anche gli SNPs di collegamento tra i due geni costituiscono una metavariabile di 9 stati.

PPM1A



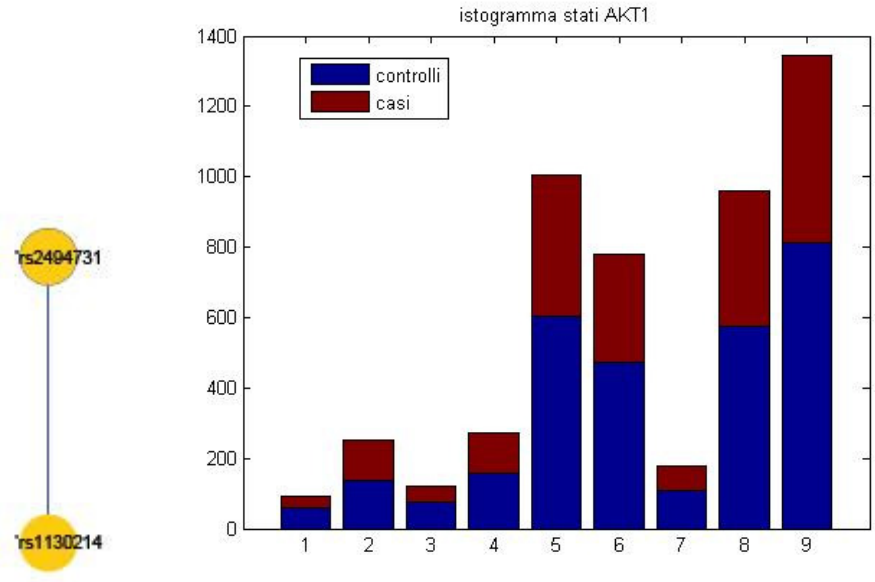
PPM1A (cromosoma 14) ha 10 SNPs e costituisce un'unica metavariabile con 62 stati.

RHOJ



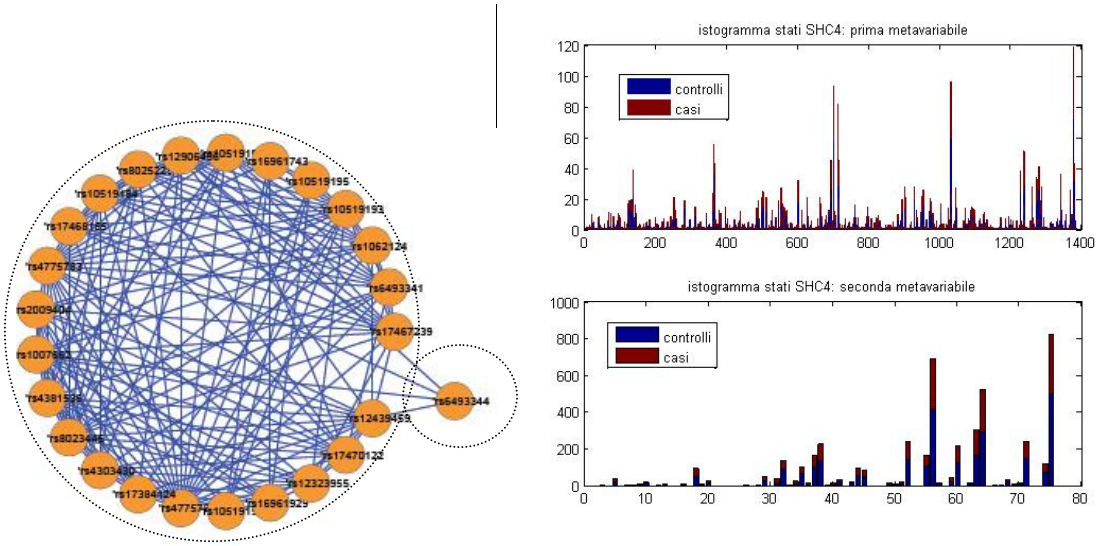
RHOJ (cromosoma 14) ha 19 SNPs e costituisce un'unica metavariabile con 344 stati.

AKT1



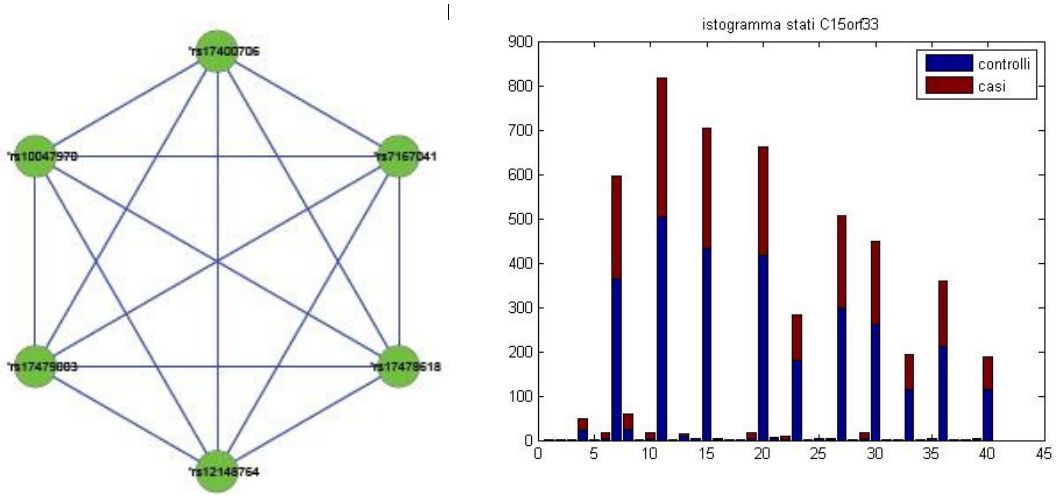
AKT1 (cromosoma 14) ha 2 SNPs e costituisce un'unica metavariabile, con 9 stati.

SCH4



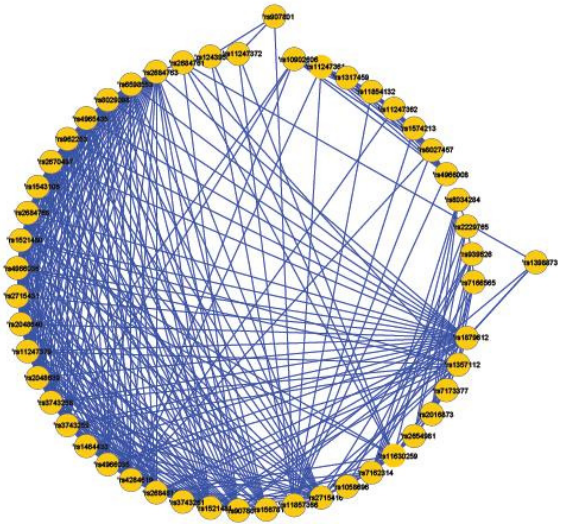
SHC4 (cromosoma 15) ha 25 SNPs, costituisce due metavariabili con 1382 e 75 stati rispettivamente.

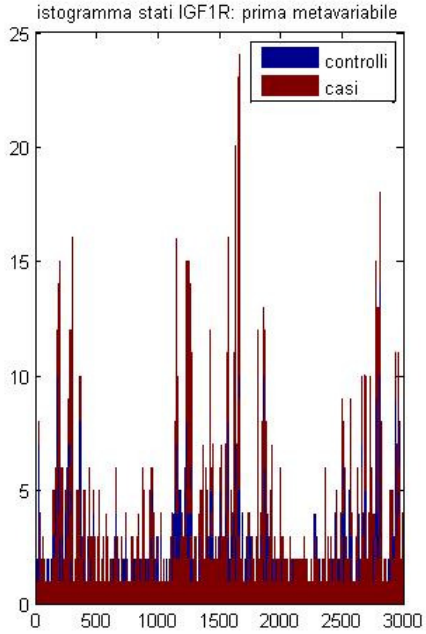
C15orf33

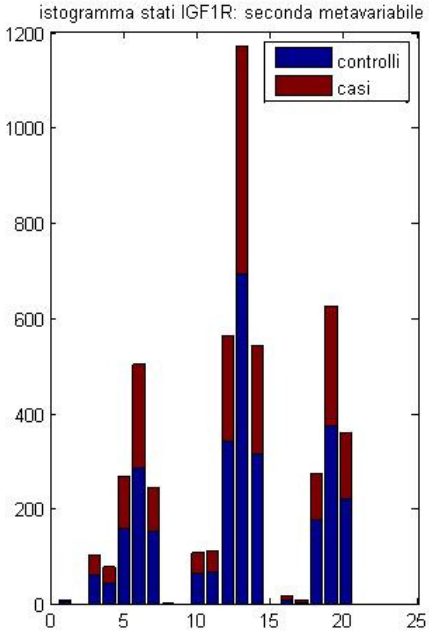


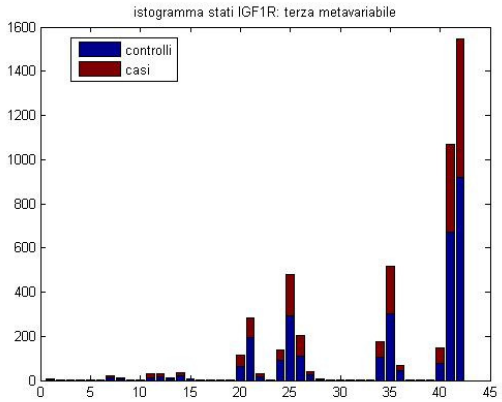
C15orf33 (cromosoma 15) ha 6 SNPs, costituisce un'unica metavariabile con 40 stati.

IGF1R



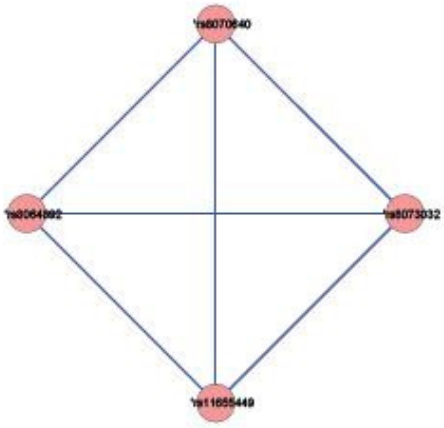

istogramma stati IGF1R: prima metavariabile
controlli
casi
0
5
10
15
20
25
0
500
1000
1500
2000
2500
3000

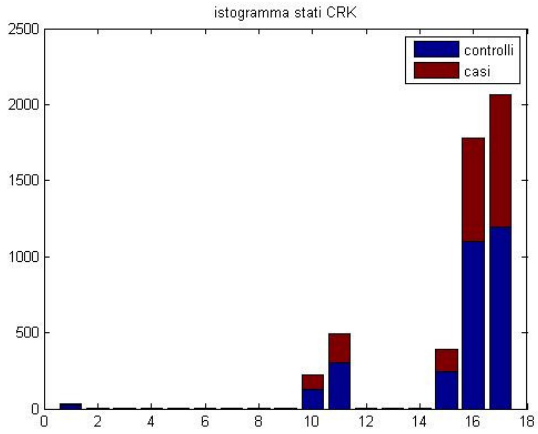

istogramma stati IGF1R: seconda metavariabile
controlli
casi
0
200
400
600
800
1000
1200
0
5
10
15
20
25


istogramma stati IGF1R: terza metavariabile
controlli
casi
0
200
400
600
800
1000
1200
1400
1600
0
5
10
15
20
25
30
35
40
45

IGF1R (cromosoma 15) ha 52 SNPs, ma solo 51 rappresentati nel grafico, costituisce tre metavariabili con 2997, 20 e 42 stati.

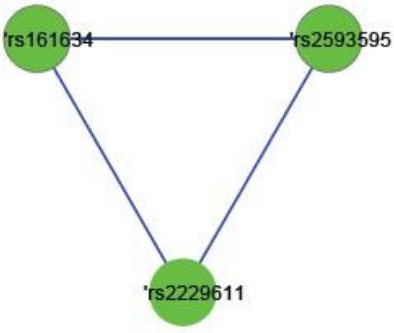
CRK

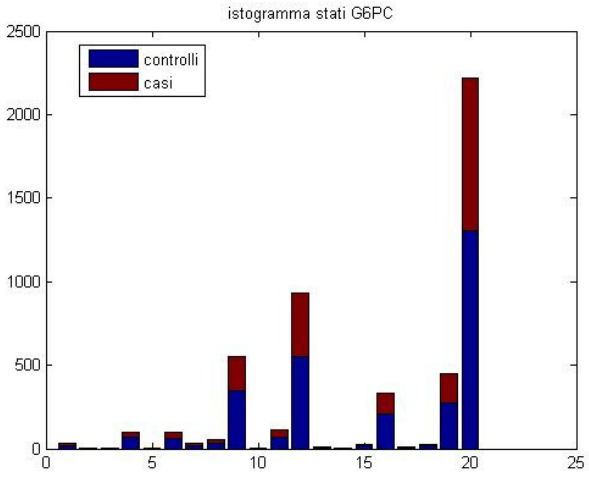




CRK (cromosoma 17) ha 4 SNPs e costituisce un'unica metavariabile con 17 stati.

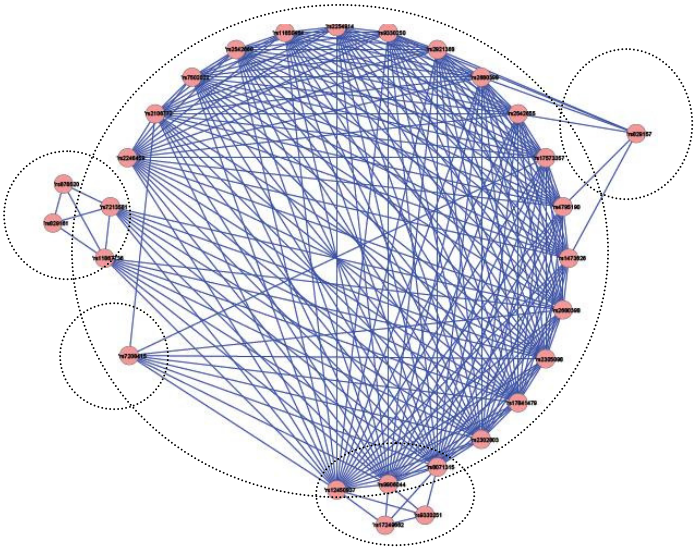
G6PC

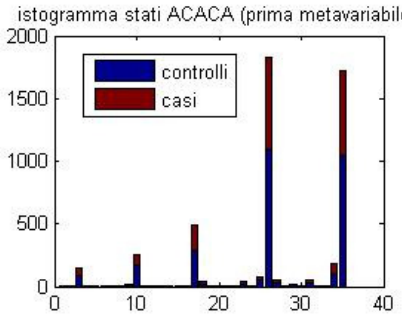


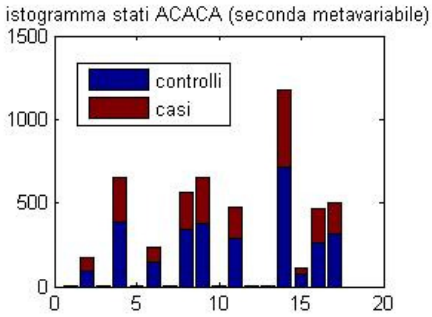


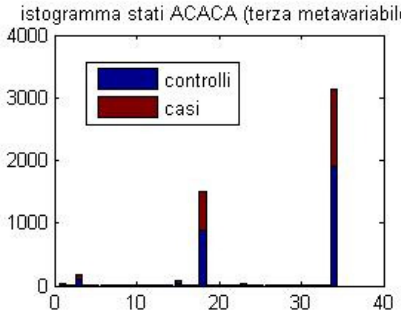
G6PC (cromosoma 17) ha 3 SNPs e costituisce un'unica metavariabile con 20 stati.

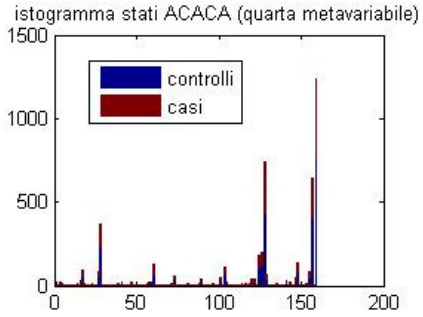
ACACA

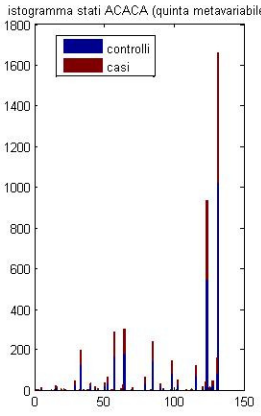


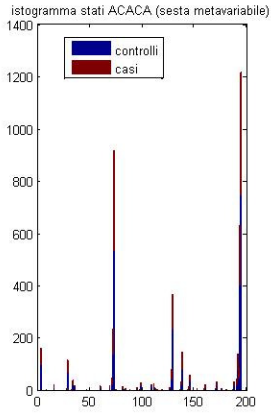

istogramma stati ACACA (prima metavariabile)
controlli
casi
0
500
1000
1500
2000
0
10
20
30
40


istogramma stati ACACA (seconda metavariabile)
controlli
casi
0
500
1000
1500
0
5
10
15
20


istogramma stati ACACA (terza metavariabile)
controlli
casi
0
1000
2000
3000
4000
0
10
20
30
40

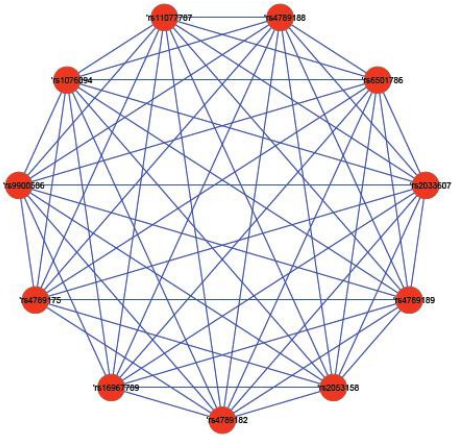

istogramma stati ACACA (quarta metavariabile)
controlli
casi
0
500
1000
1500
0
50
100
150
200

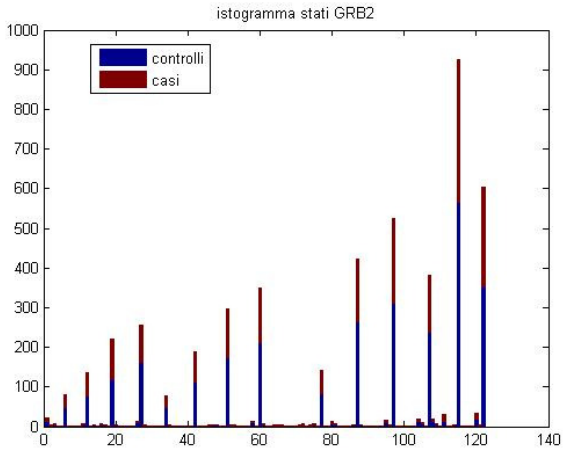

istogramma stati ACACA (quinta metavariabile)
controlli
casi
0
200
400
600
800
1000
1200
1400
1600
1800
0
50
100
150


istogramma stati ACACA (sesta metavariabile)
controlli
casi
0
200
400
600
800
1000
1200
1400
0
50
100
150
200

ACACA (cromosoma 17) ha 28 SNPs e costituisce 5 metavariabili, di 35, 17, 34, 159 e 131 stati ciascuno.

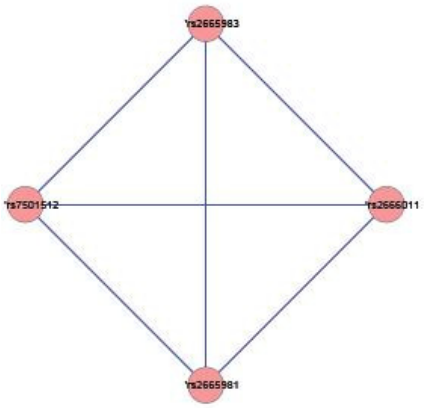
GRB2

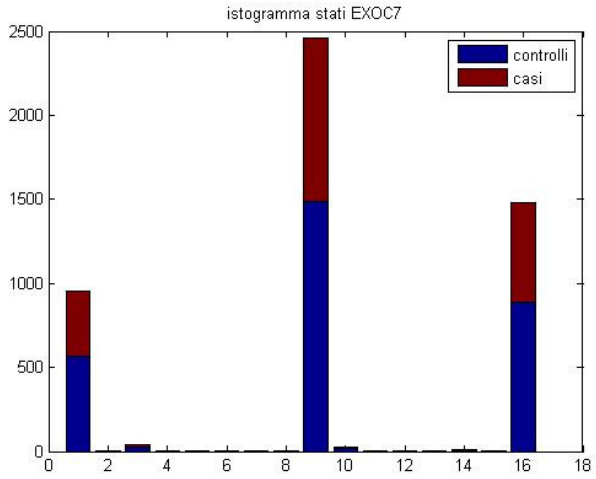




GRB2 (cromosoma 17) ha 11 SNPs, costituisce un'unica metavariabile e ha 96 stati.

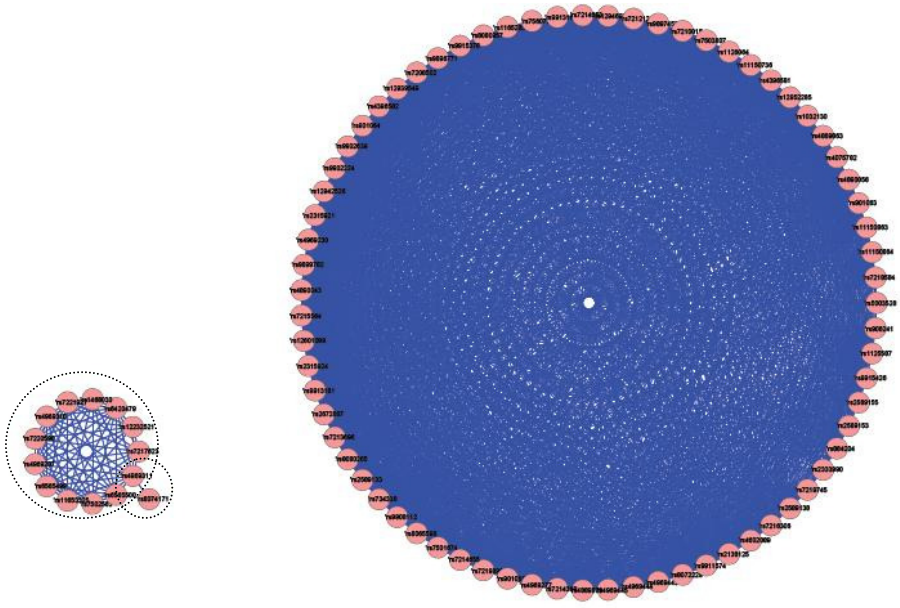
EXOC7

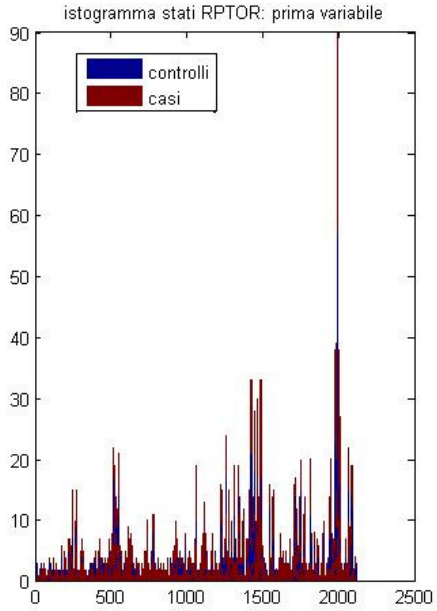


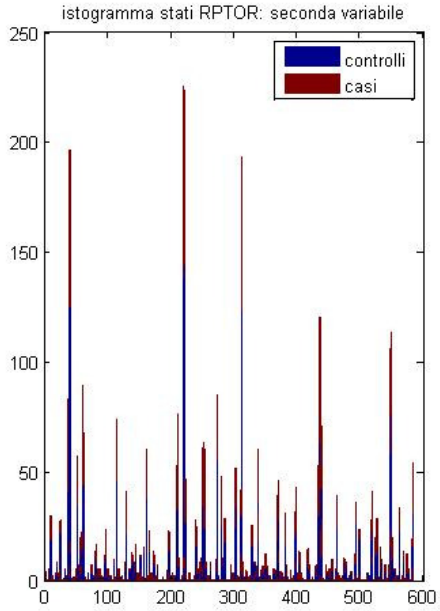


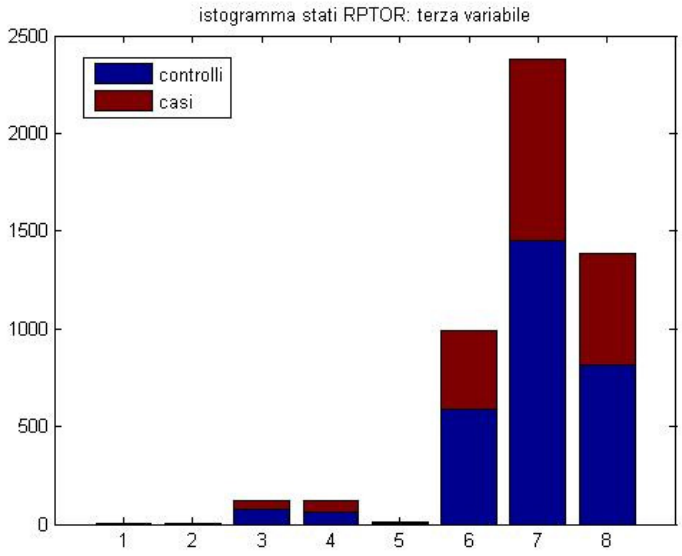
EXOC7 (cromosoma 17) ha 4 SNPs e costituisce un'unica metavariabile di 16 stati.

RPTOR



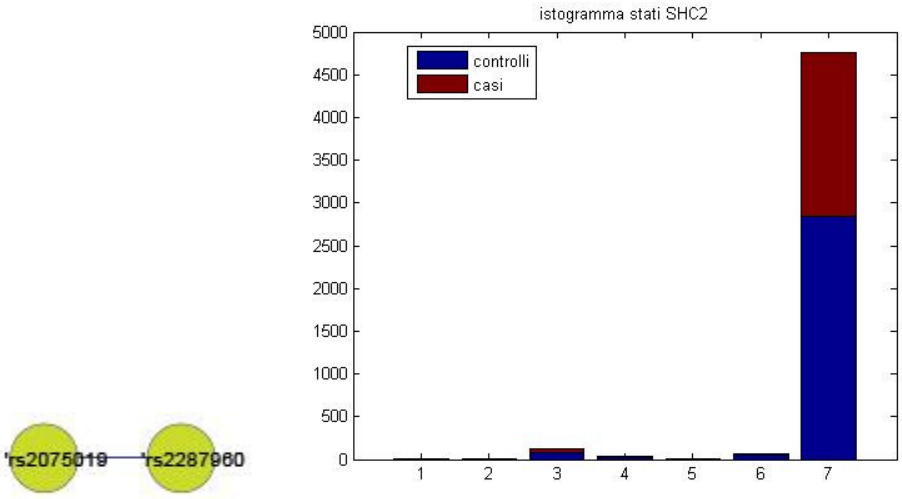

istogramma stati RPTOR: prima variabile
controlli
casi
0
10
20
30
40
50
60
70
80
90
0
500
1000
1500
2000
2500


istogramma stati RPTOR: seconda variabile
controlli
casi
0
50
100
150
200
250
0
100
200
300
400
500
600



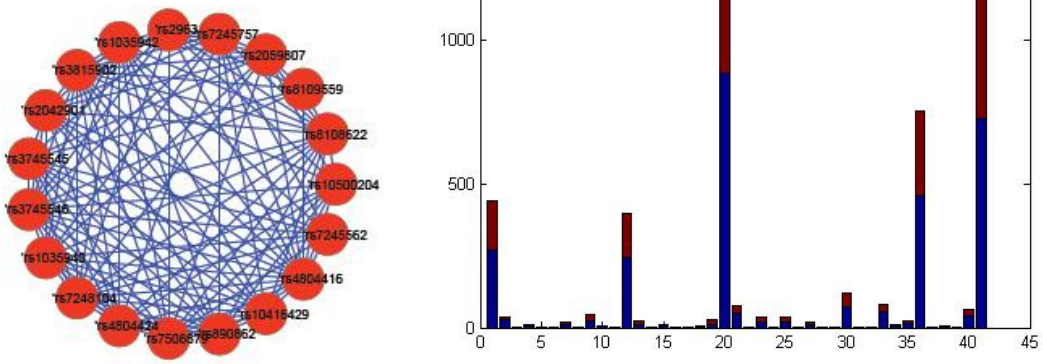
RPTOR rappresenta un gene fortemente connesso della rete , ha 85 SNPs e è diviso in tre meta variabili di 2107, 582 e 8 stati ciascuna.

SHC2



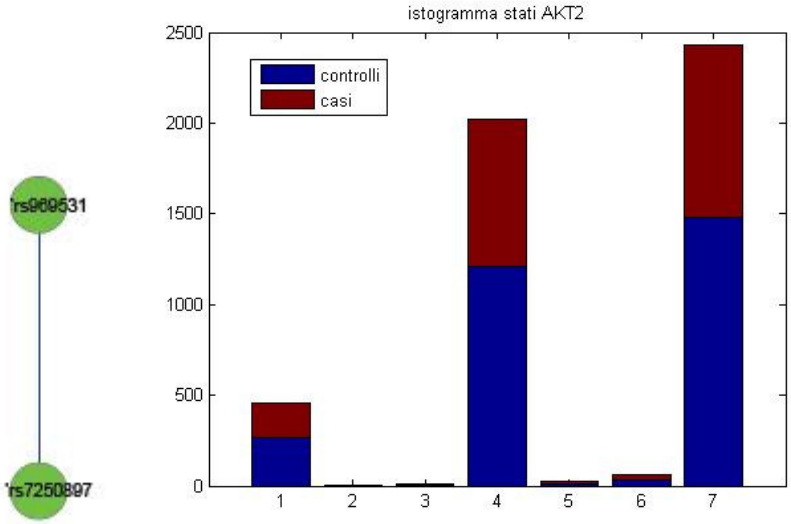
SHC2 (cromosoma 19) ha 2 SNPs e costituisce un'unica metavariabile con 7 stati.

INSR



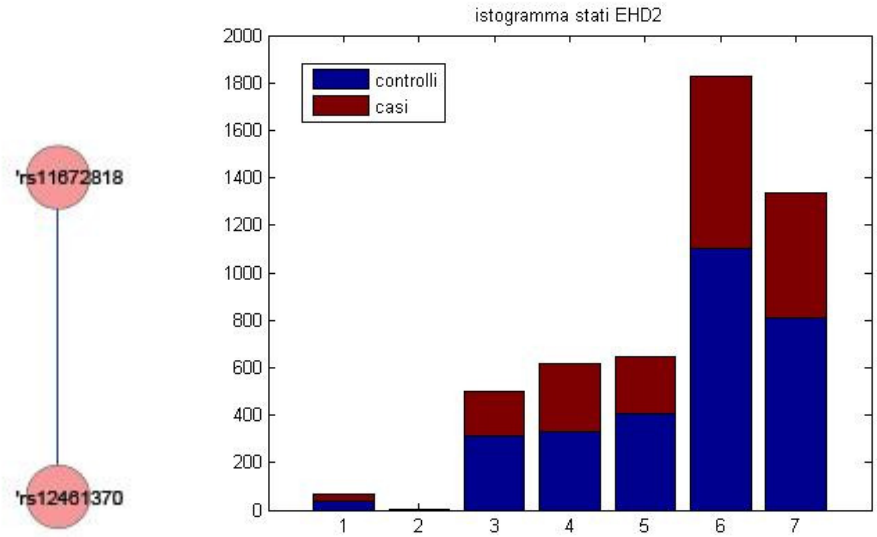
INSR (cromosoma 19) ha 23 SNPs, ma solo 19 compresi nel grafico, costituisce un'unica metavariabile con 41 stati.

AKT2



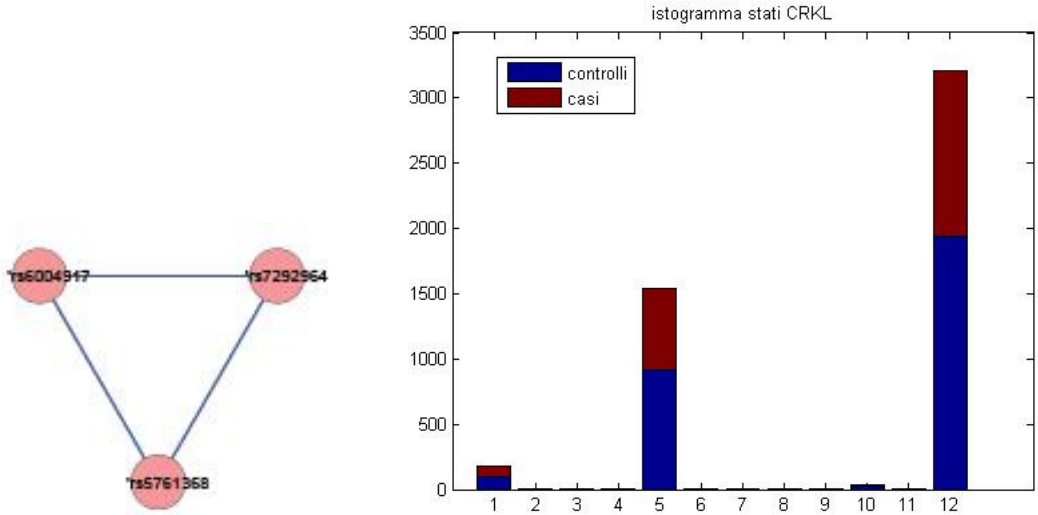
AKT2 (cromosoma 19) ha 2 SNPs, costituisce un'unica metavariabile con 7 stati.

EHD2



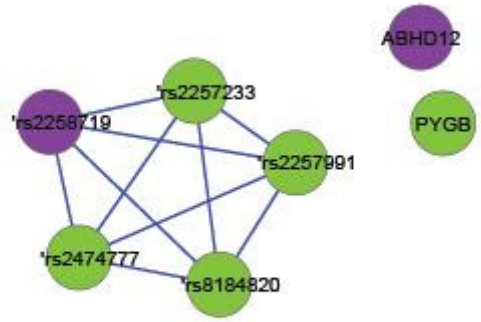
EHD2 (cromosoma 19) ha 2 SNPs, costituisce un'unica metavariabile con 7 stati.

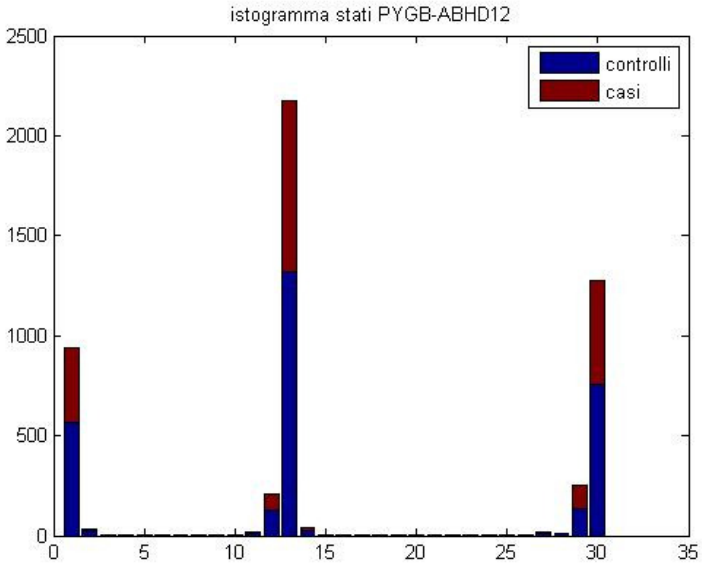
CRKL



CRKL (cromosoma 22) ha 3 SNPs, costituisce un'unica metavariabile con 12 stati.

PYGB – ABHD12

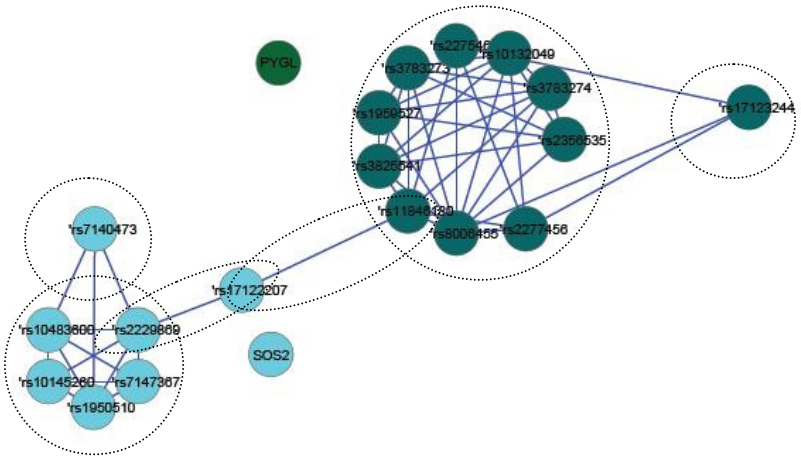


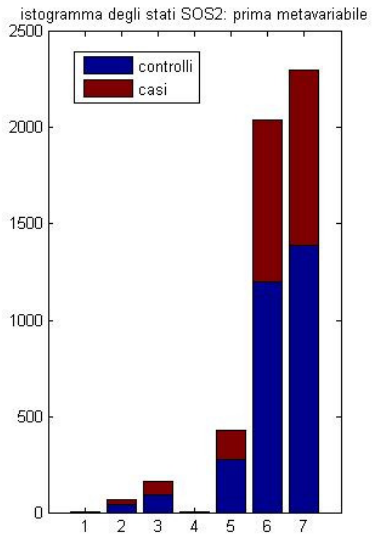


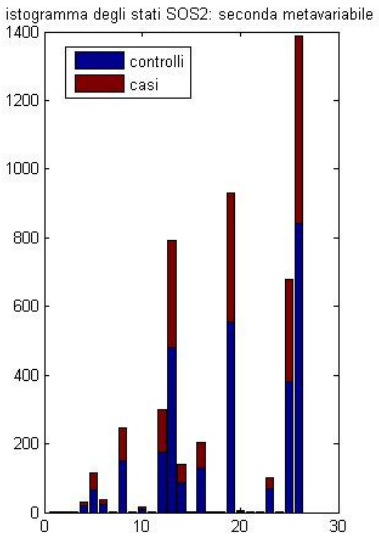
Sono rappresentati due geni ABHD12 e PYGB (entrambi sul cromosoma 20) che risultano connessi tra di loro e costituiscono un'unica metavariabile con 30 stati.

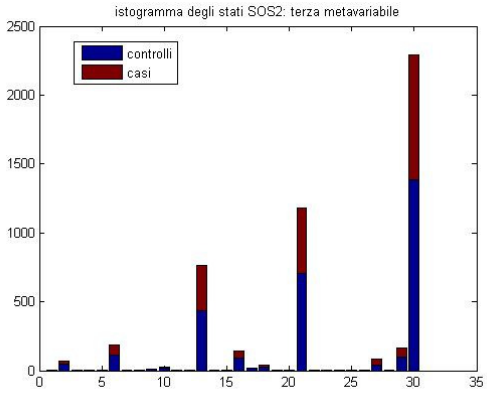
9.2 Rete ricostruita dai dati di controlli e casi del diabete di tipo 2

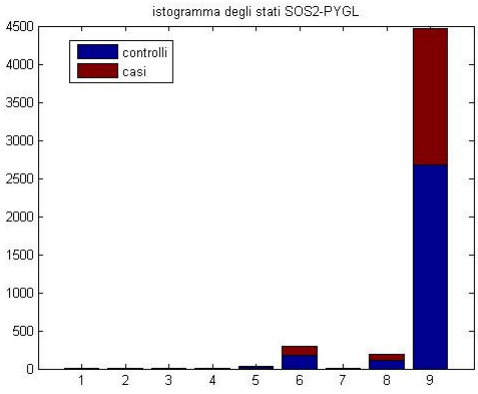
PYGL – SOS2

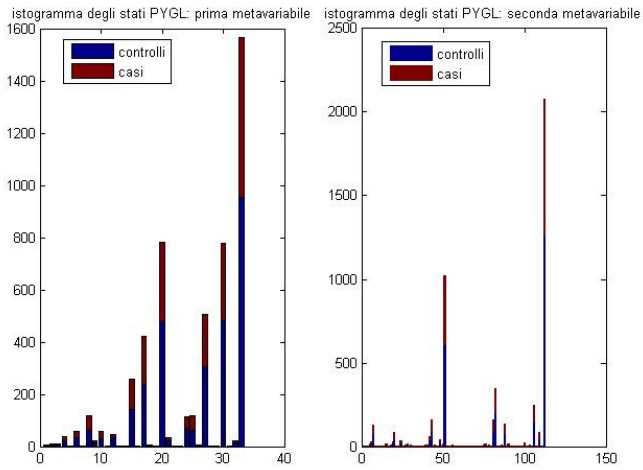






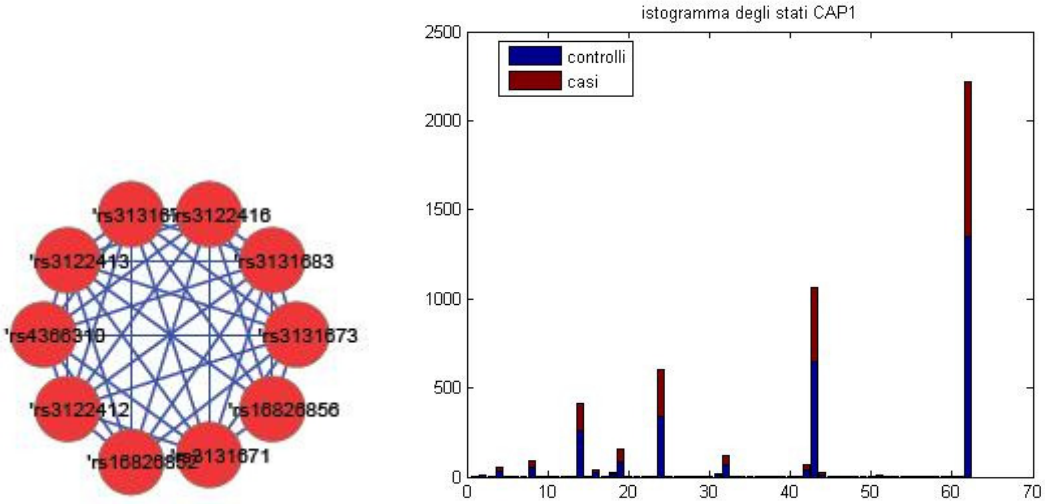






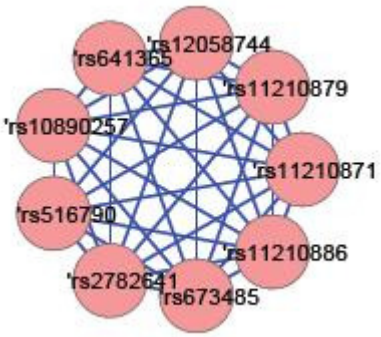
Viene rappresentata a connessione tra i geni PYGL e SOS2 (già trovato nella rete precedente) con 6 metavariabili: 3 metavariabili per SOS2, una metavariabile per gli SNPs di collegamento e 2 metavariabili per PYGL.

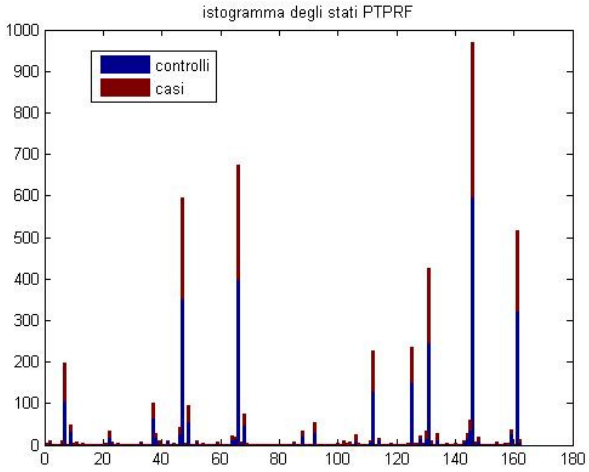
CAP1



CAP1 (cromosoma 1) ha 10 SNPs e costituisce un'unica metavariabile con 62 stati.

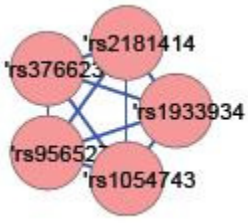
PTPRF

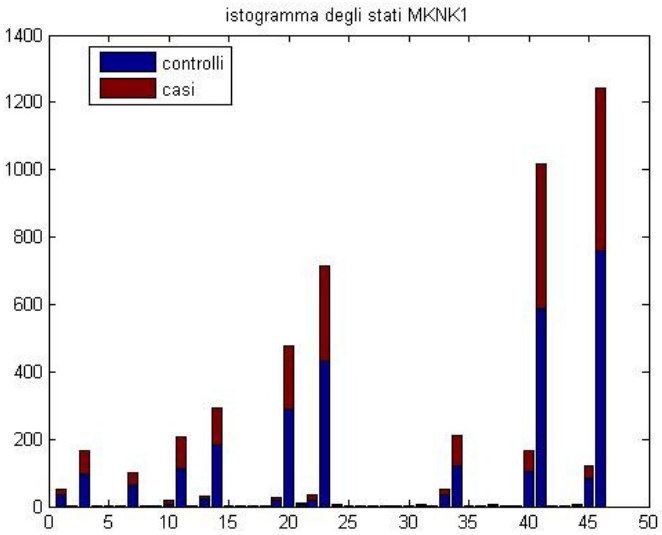




PTPRF (cromosoma 1) ha 9 SNPs e costituisce un'unica metavariabile con 162 stati.

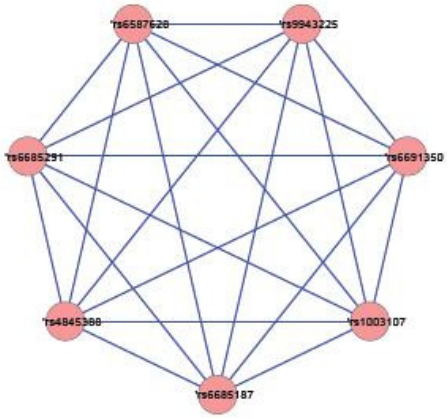
MKNK1

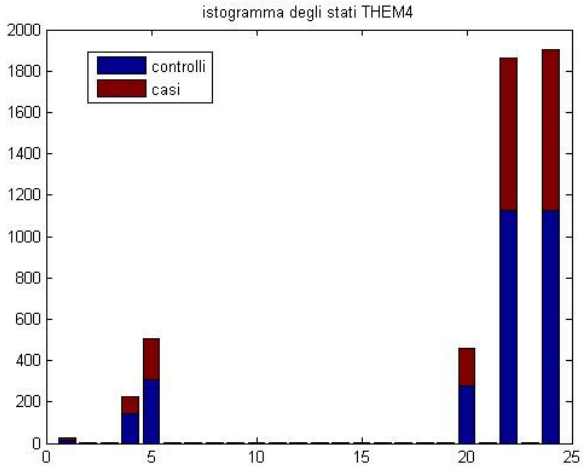




MKNK1 (cromosoma 1) ha 5 SNPs, costituisce un'unica metavariabile con 46 stati.

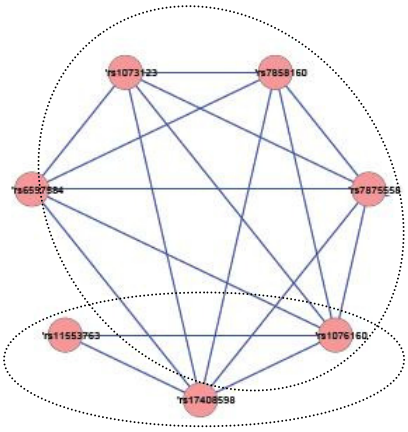
THEM4

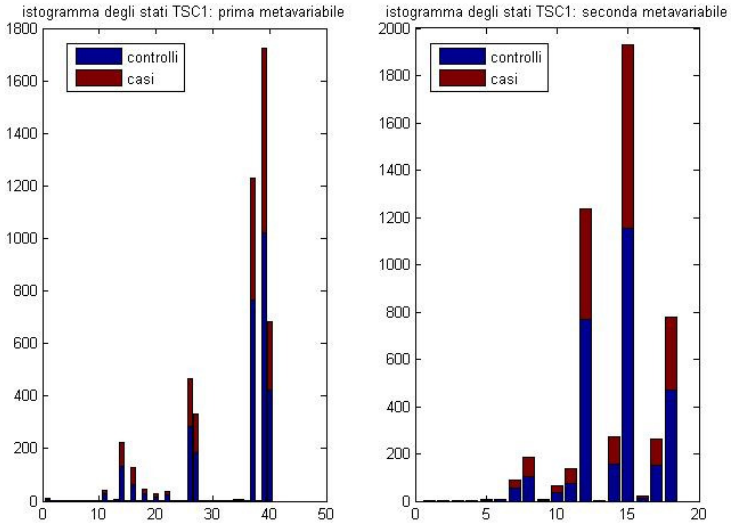




THEM4 (cromosoma 1) ha 7 SNPs, costituisce un'unica metavariabile con 24 stati.

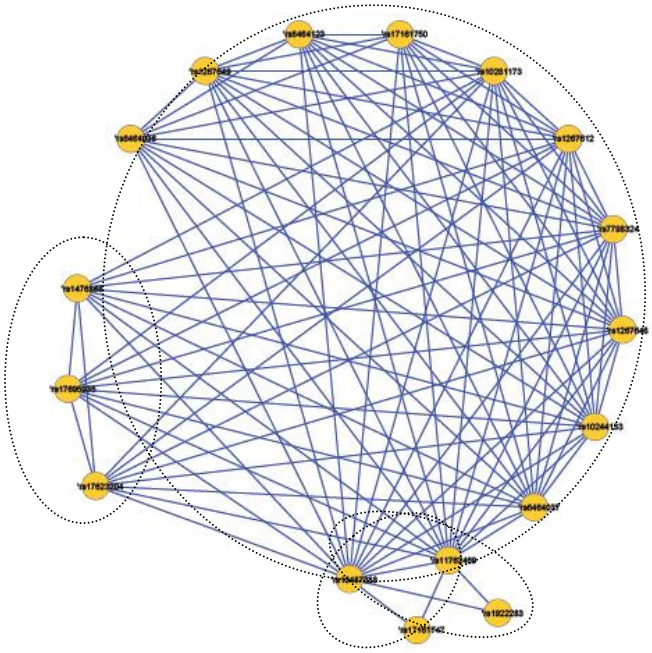
TSC1

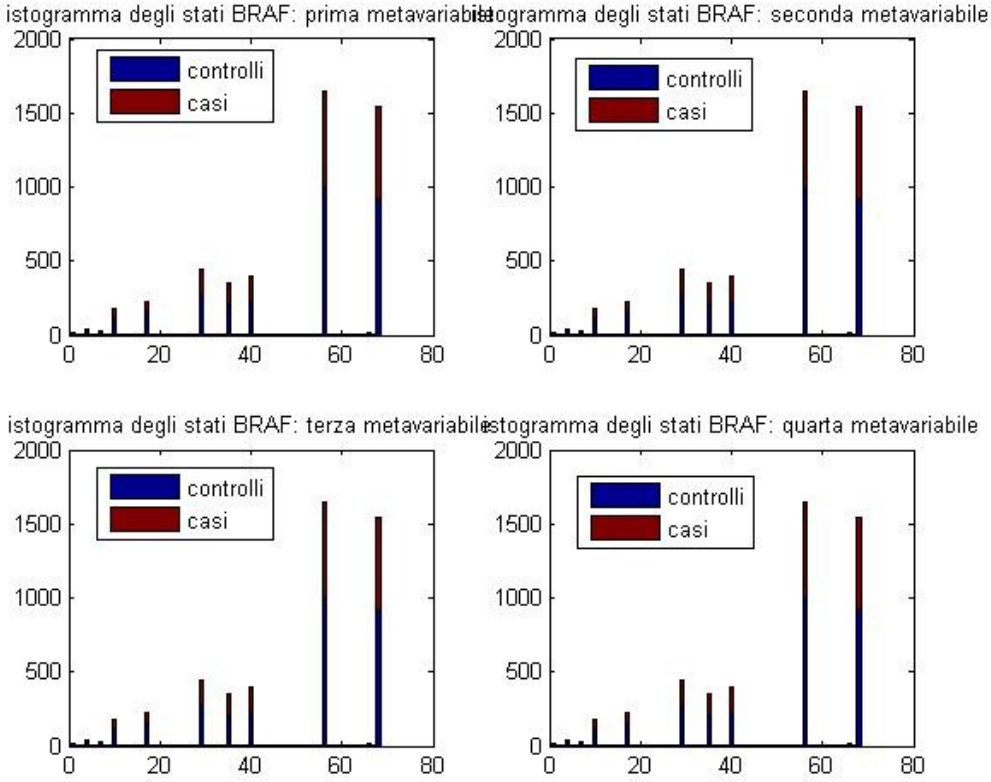




TSC1 (cromosoma 9) ha 7 SNPs e costituisce due metavariabili di 40 e 18 stati ciascuna.

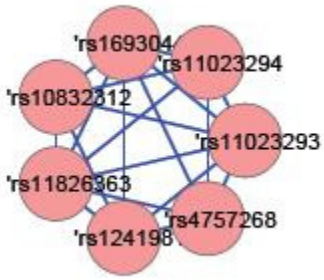
BRAF

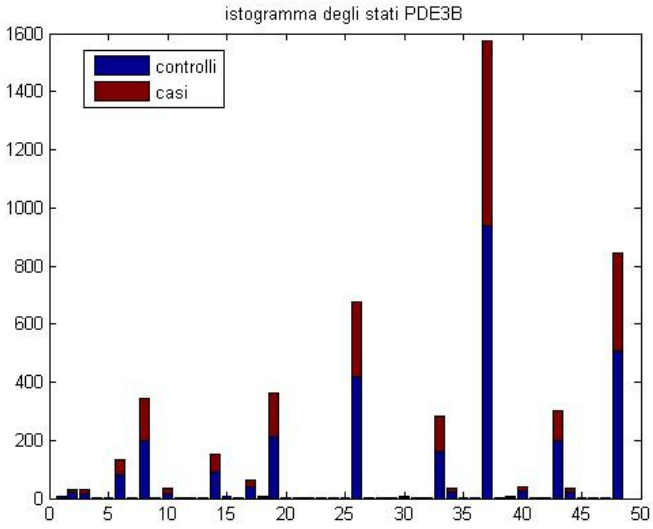




BRAF (cromosoma 7) è organizzato in maniera analoga a quanto visto per il dataset precedente, ma ha una metavariabile in più.

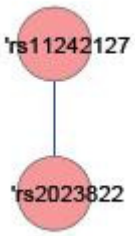
PDE3B

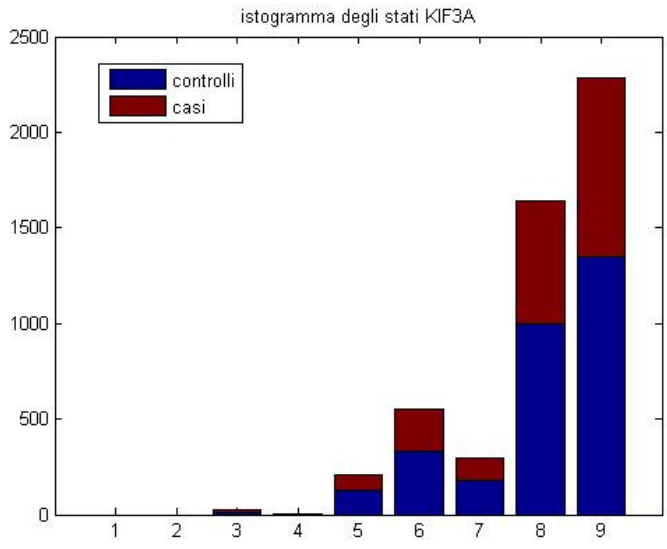




PDE3B (cromosoma 11) ha 7 SNPs e costituisce un'unica metavariabile con 47 stati.

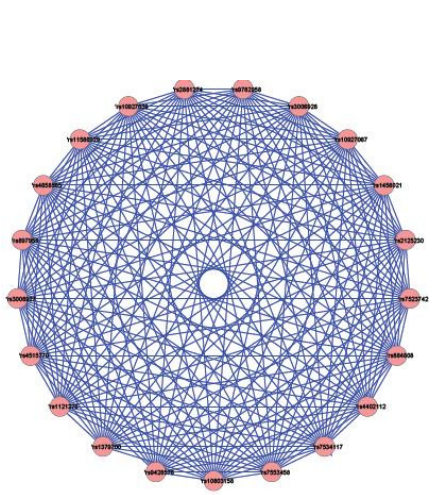
KIF3A

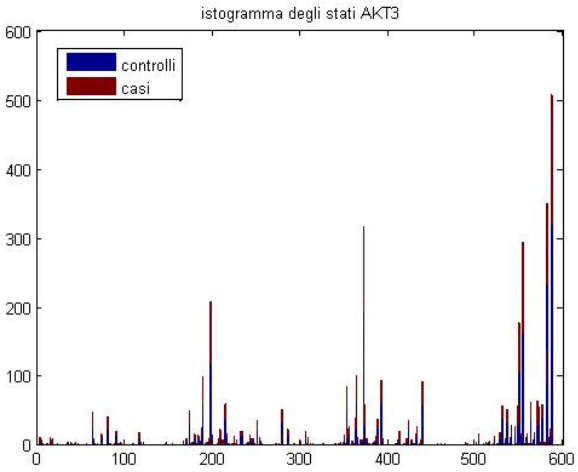




KIF3A (cromosoma 5) ha 2 SNPs e costituisce un'unica metavariabile con 9 stati.

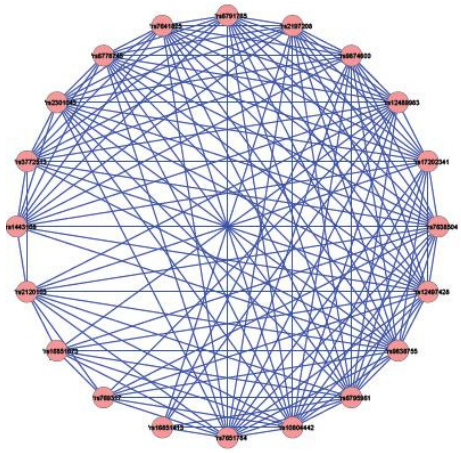
AKT3

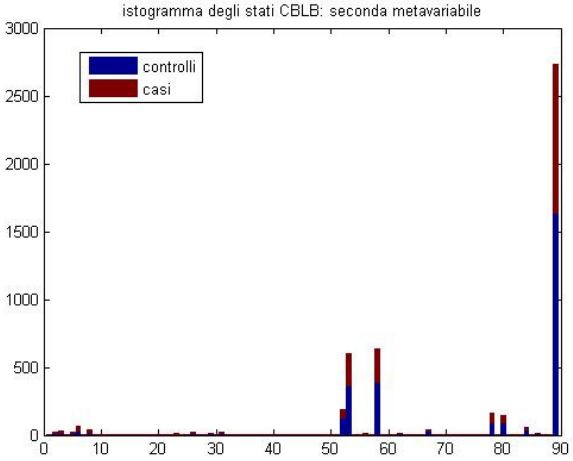




AKT3 (cromosoma 1) ha 21 SNPs e costituisce un'unica metavariabile con 586 stati.

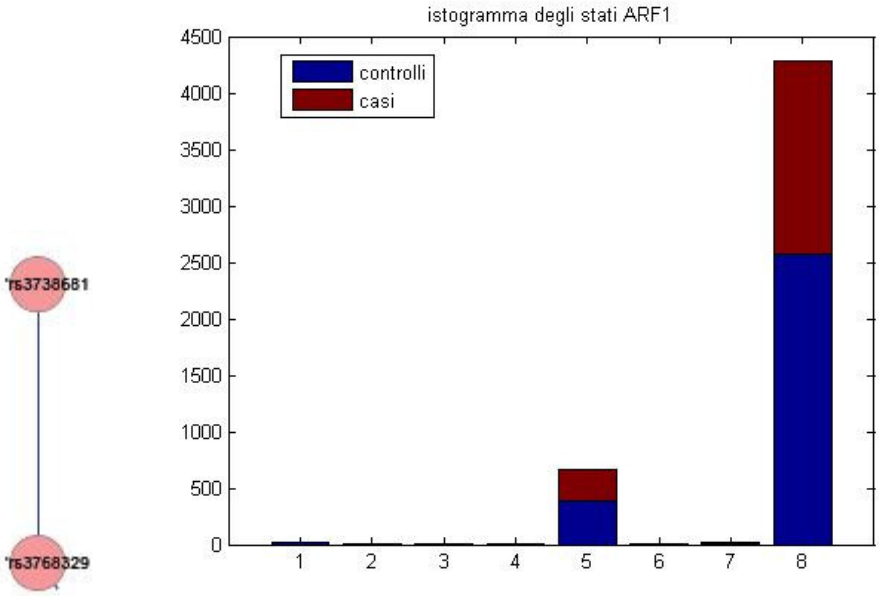
CBLB





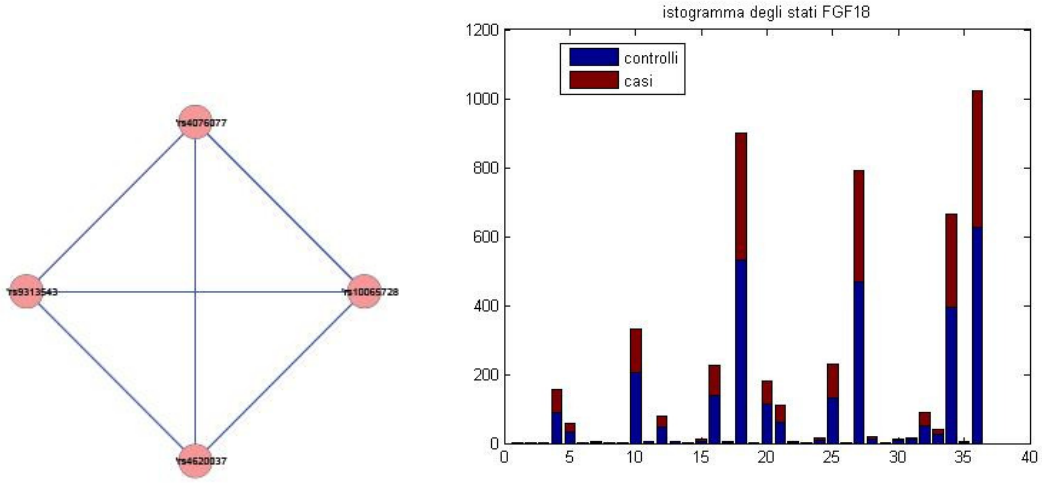
CBLB (cromosoma 3) è costituito da 20 SNPs che possono essere suddivisi in due metavariabili. Viene riportato l'istogramma di una sola delle due metavariabili, in quanto l'altra presenta un unico stato, pertanto non viene considerata nel procedimento di classificazione. Infatti, assumendo sempre lo stesso stato sia per i casi che per i controlli, non dà alcuna informazione aggiuntiva e non contribuisce in alcun modo alla classificazione.

ARF1



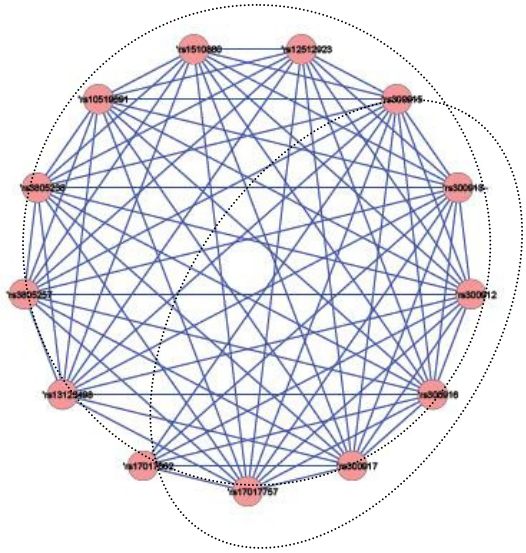
ARF1 (cromosoma 1) ha 2 SNPs e costituisce un'unica metavariabile con 8 stati (due stati nettamente prevalenti rispetto agli altri).

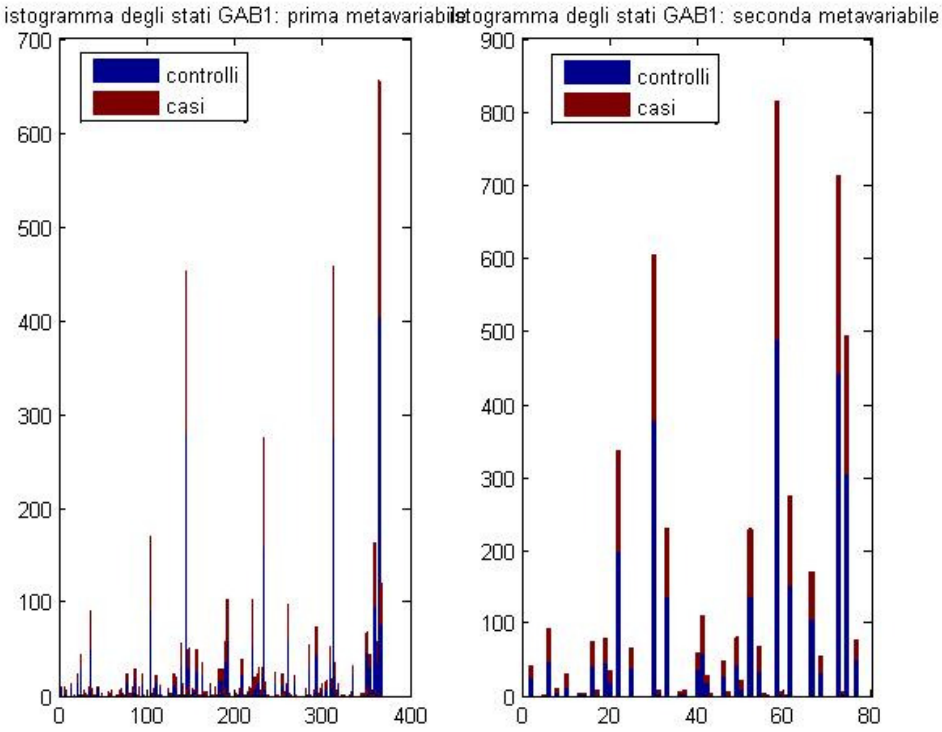
FGF18



FGF18 (cromosoma 5) ha 4 SNPs e costituisce un'unica metavariabile, con 36 stati.

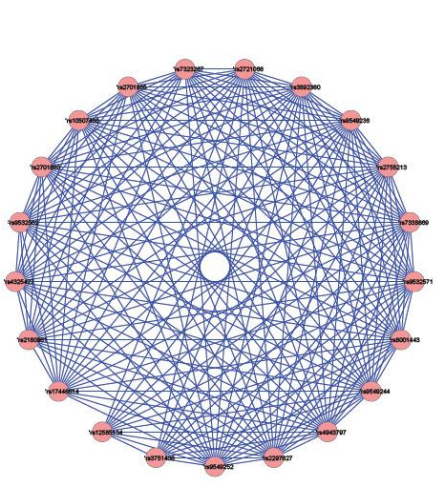
GAB1

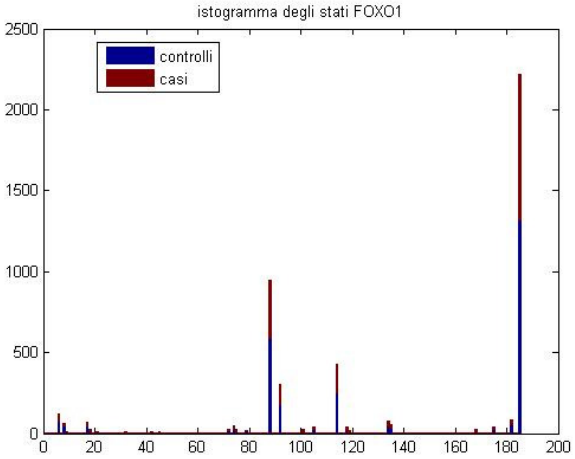




GAB1 (cromosoma 4) ha 14 stati, e costituisce due metavariabili con 366 e 66 stati rispettivamente.

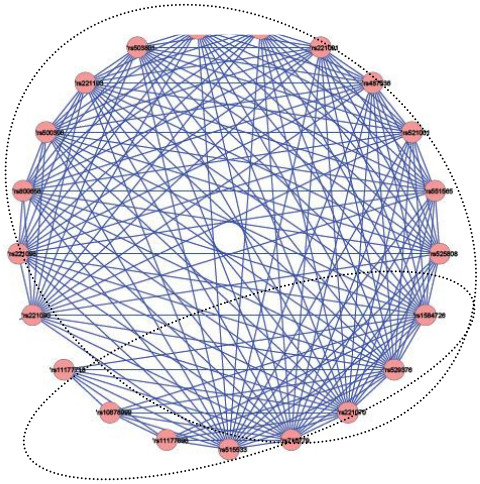
FOXO1

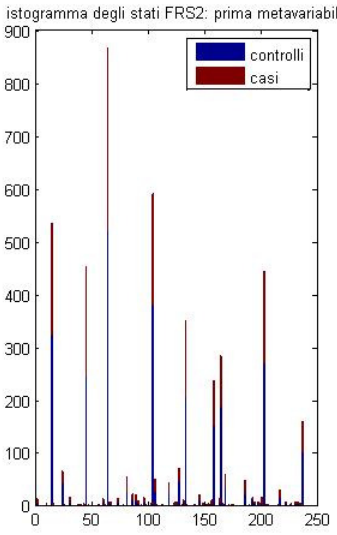


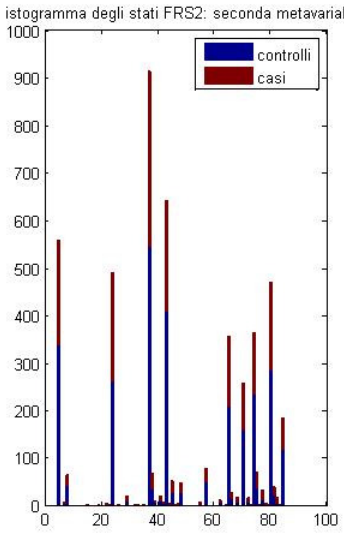


FOXO1 (cromosoma 13) ha 21 SNPs e costituisce un'unica metavariabile con 185 stati.

FRS2

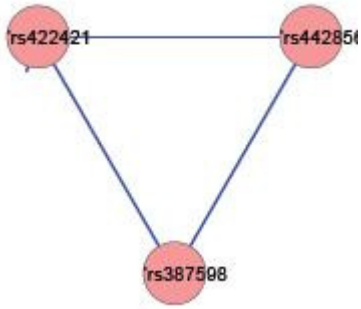


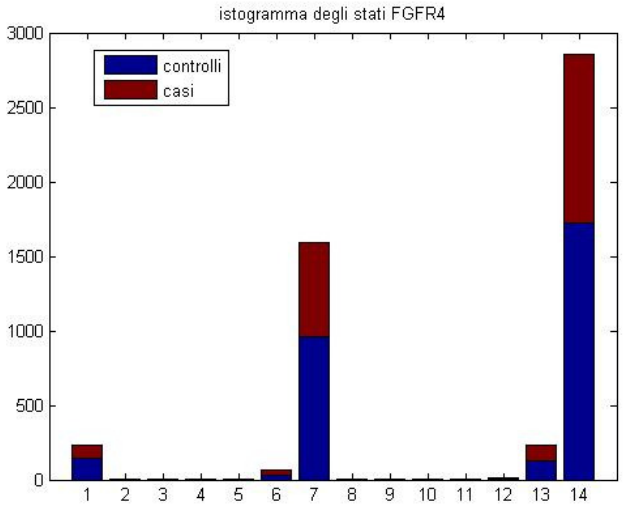




FRS2 (cromosoma 12) ha 21 SNPs organizzati in metavariabili con 236 e 84 stati rispettivamente.

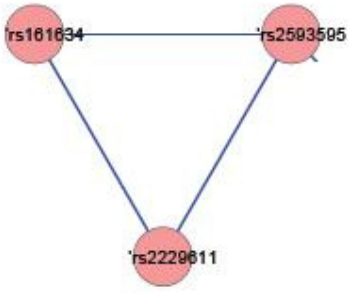
FGFR4

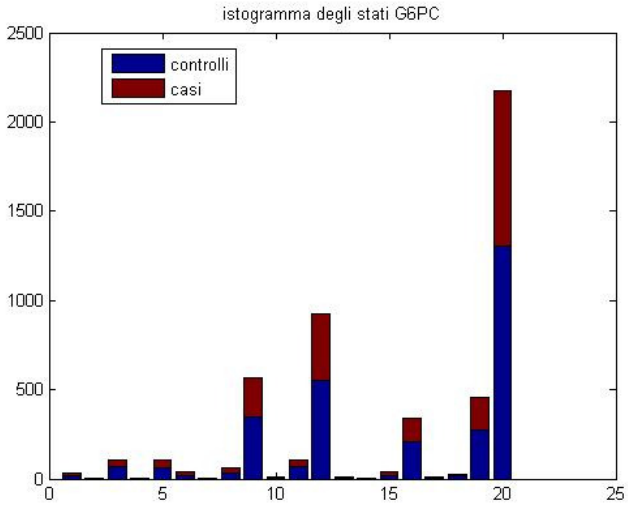




FGFR4 (cromosoma 5) ha 3 SNPs e costituisce un'unica metavariabile con 14 stati.

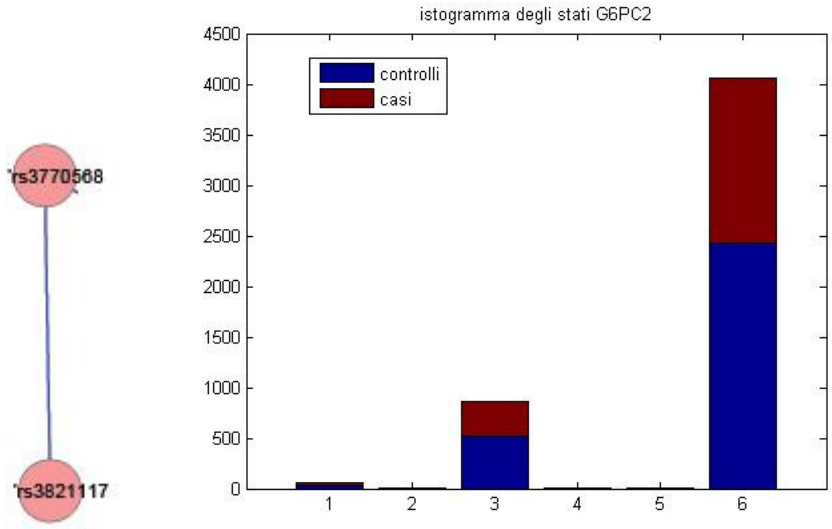
G6PC





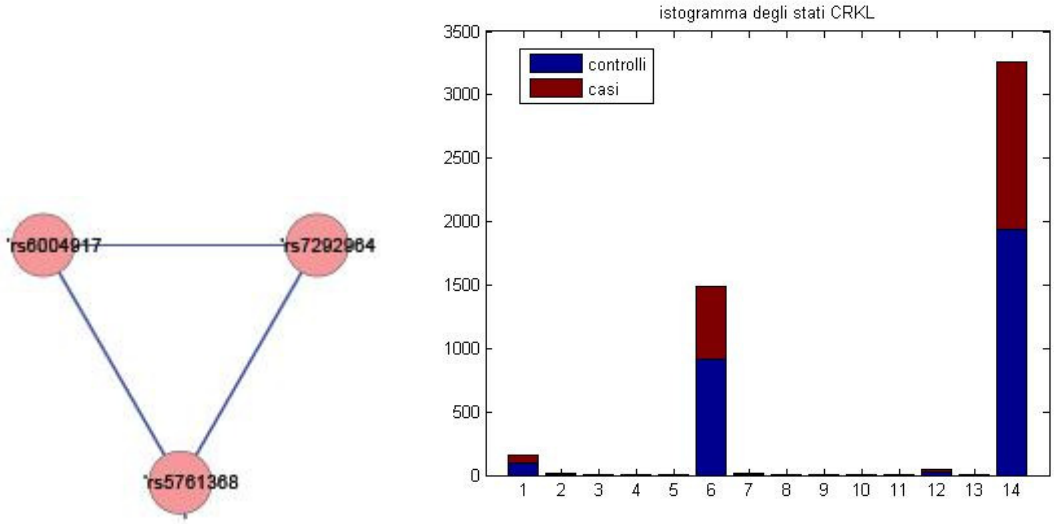
G6PC (cromosoma 17) ha 3 SNPs e costituisce un'unica metavariabile con 20 stati.

G6PC2



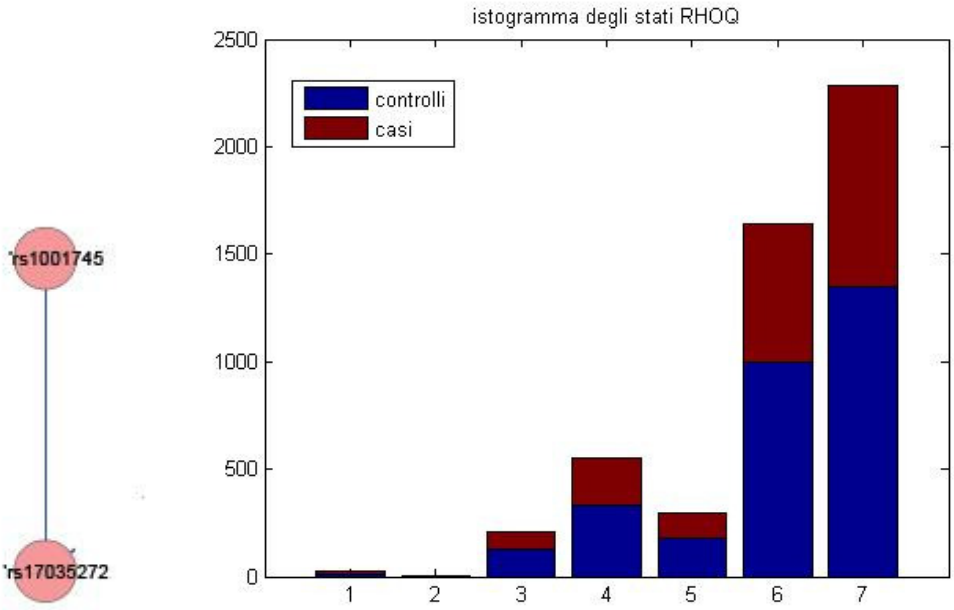
G6PC2 (cromosoma 2) ha 2 SNPs e costituisce un'unica metavariabile con 6 stati.

CRKL



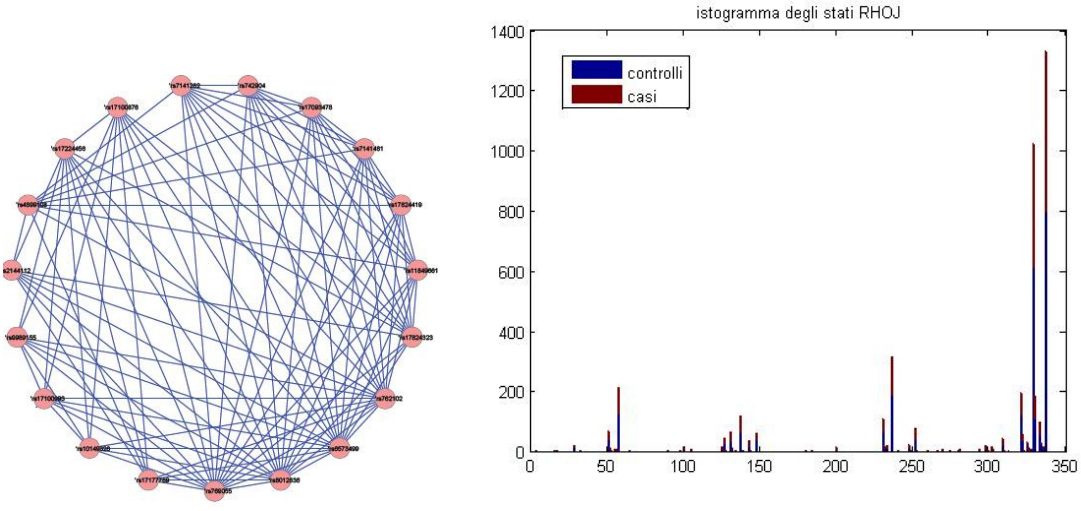
CRKL (cromosoma 22) ha 3 SNPs e costituisce un'unica metavariabile con 14 stati.

RHOQ



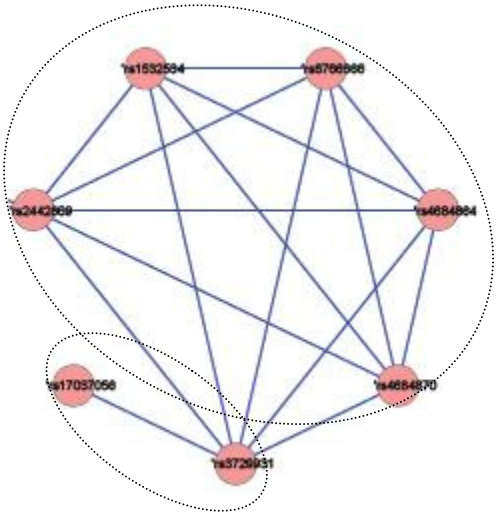
RHOQ (cromosoma 2) ha 2 SNPs e costituisce un'unica metavariabile con 7 stati.

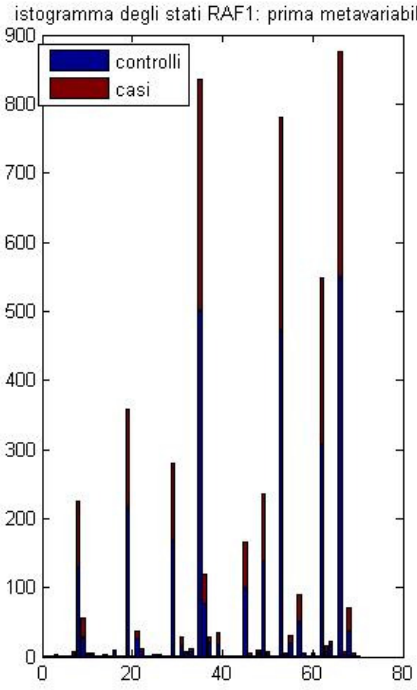
RHOJ

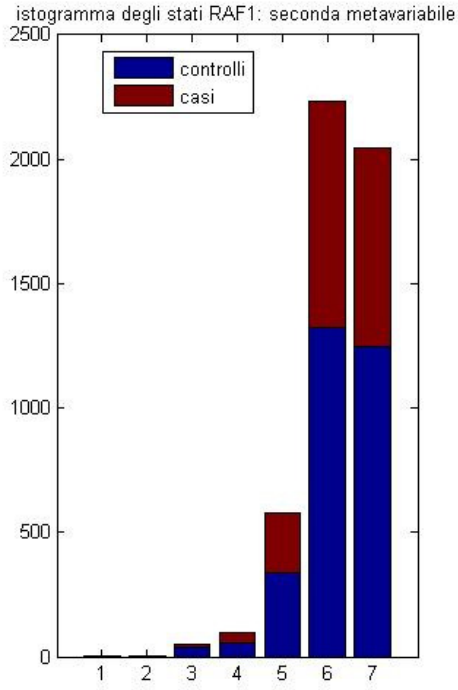


RHOJ (cromosoma 14) ha 19 SNPs e costituisce un'unica metavariabile con 336 stati.

RAF1

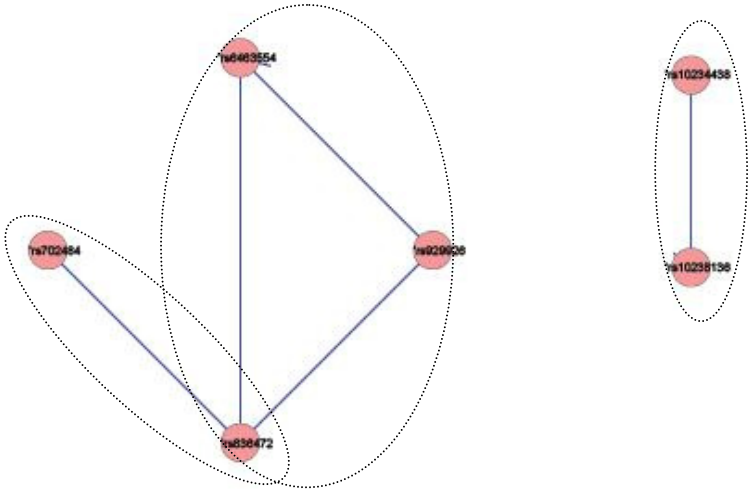


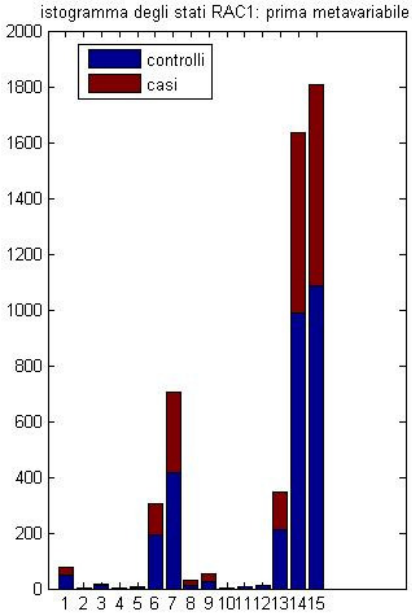


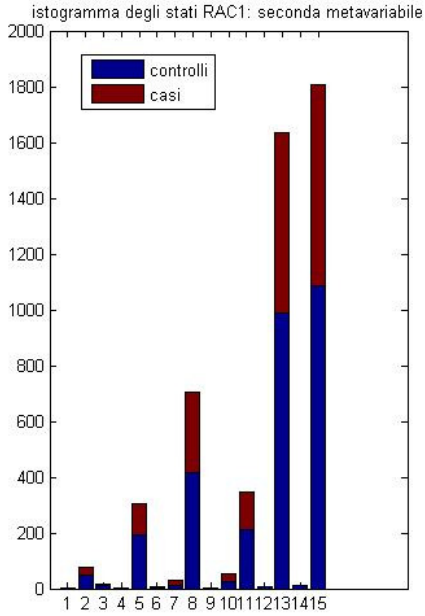


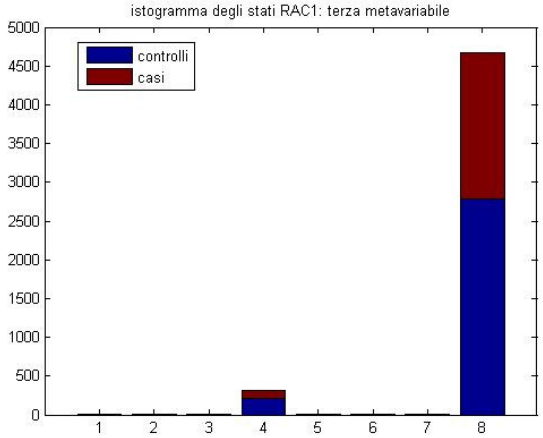
RAF1 (cromosoma 3) ha 7 SNPs e costituiscono due metavariabili con 70 e 7 stati rispettivamente.

RAC1


rs6483554
rs10234438
rs702484
rs10238136
rs836472

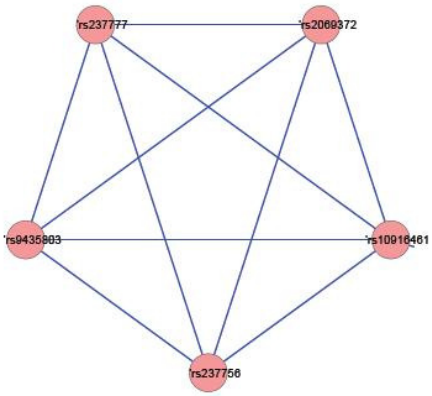

istogramma degli stati RAC1: prima metavariabile
controlli
casi
2000
1800
1600
1400
1200
1000
800
600
400
200
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

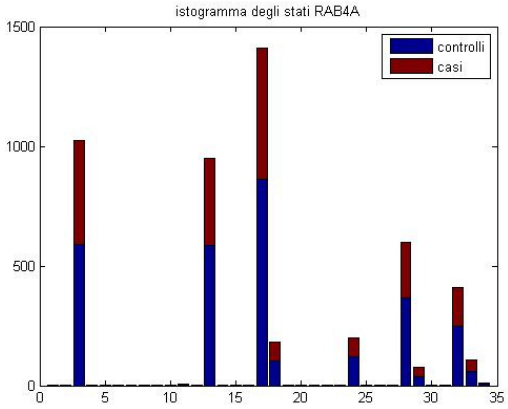

istogramma degli stati RAC1: seconda metavariabile
controlli
casi
2000
1800
1600
1400
1200
1000
800
600
400
200
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15


istogramma degli stati RAC1: terza metavariabile
controlli
casi
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
1 2 3 4 5 6 7 8

RAC1 (cromosoma 7) ha 6 SNPs organizzati in tre metavariabili: le prime due di 15 stati, mentre la restante da 8 stati.

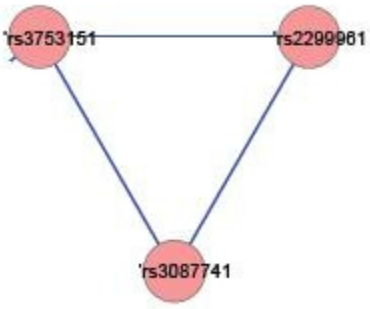
RAB4A

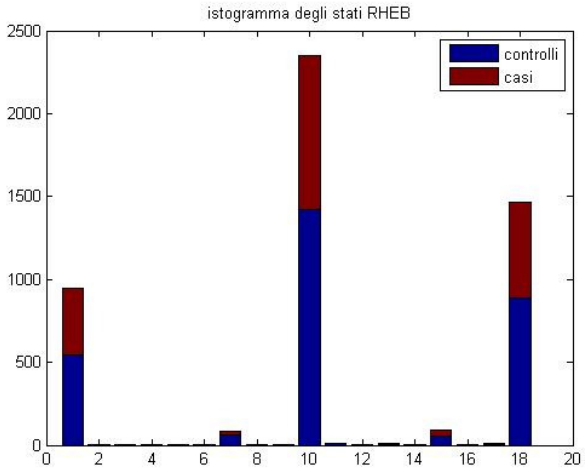




RAB4A (cromosoma 1) ha 5 SNPs e costituisce un'unica metavariabile con 34 stati.

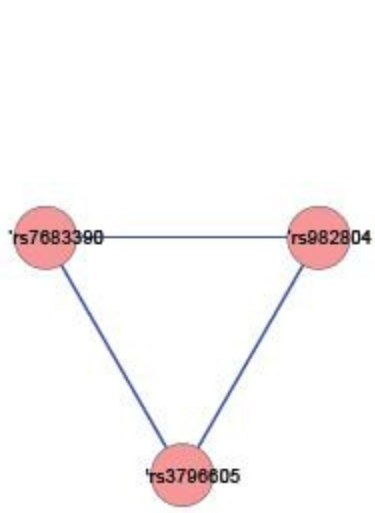
RHEB

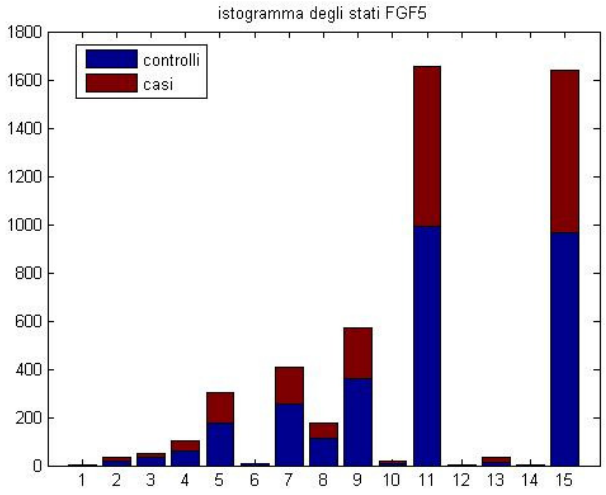




RHEB (cromosoma 7) ha 3 SNPs e costituisce un'unica metavariabile con 18 stati.

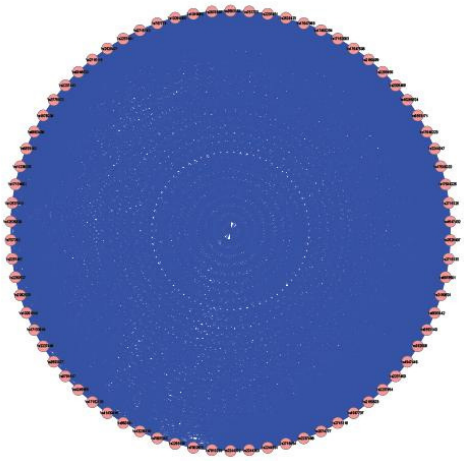
FGF5

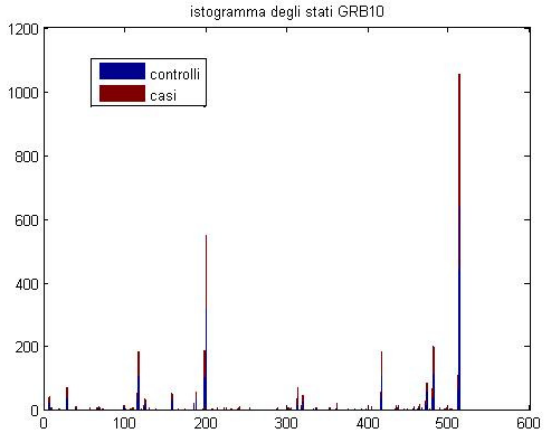




FGF5 (cromosoma 4) ha 3 SNPs e costituisce un'unica metavariabile con 15 stati.

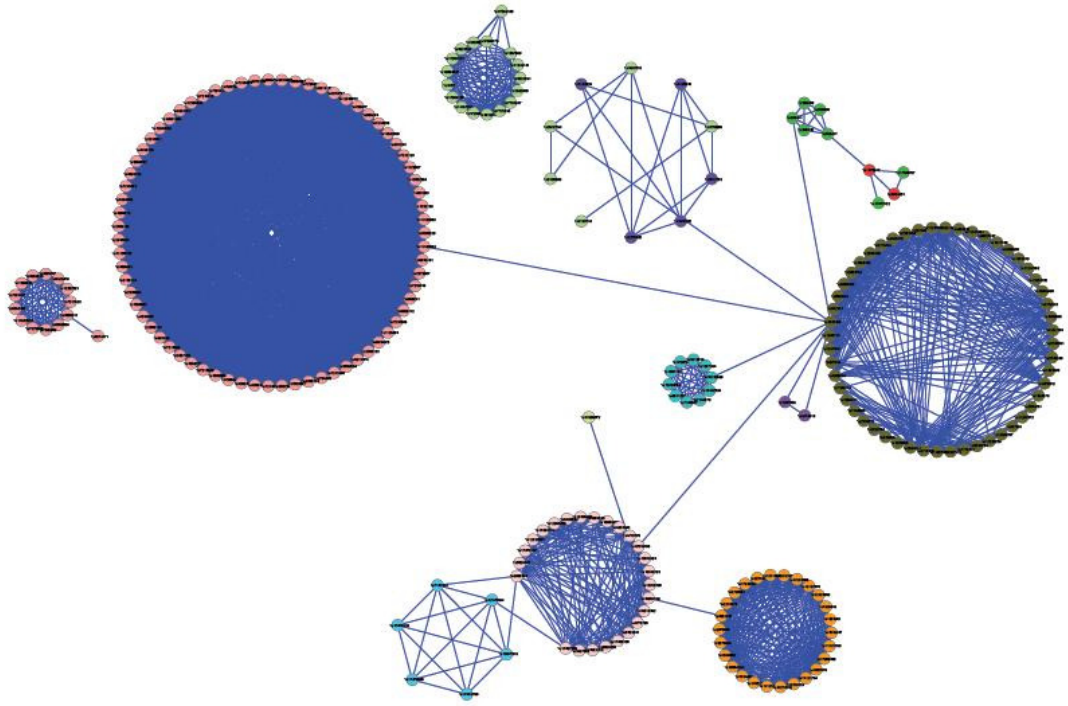
GRB10



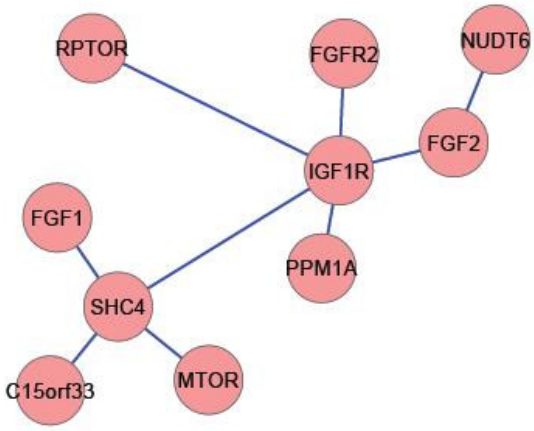


GRB10 (cromosoma 7) ha una rappresentazione molto simile a quanto già visto per la rete precedente (paragrafo 7.1): risulta ancora organizzato come un grafo fortemente connesso e vengono riportati tutti gli SNPs del gene (74 in totale). È organizzato come un'unica metavariabile e ha 512 stati.

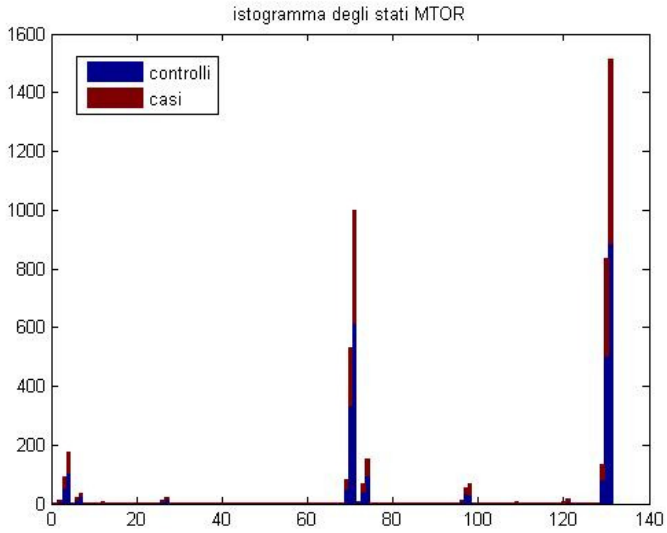
FGF1 – SHC4 – MTOR – IGF1R – C15orf33 – FGFR1 – PPM1A – NUDT6 – FGF2

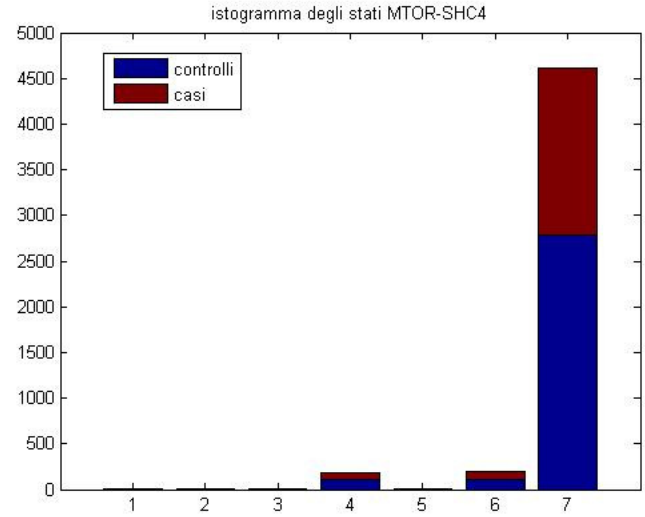


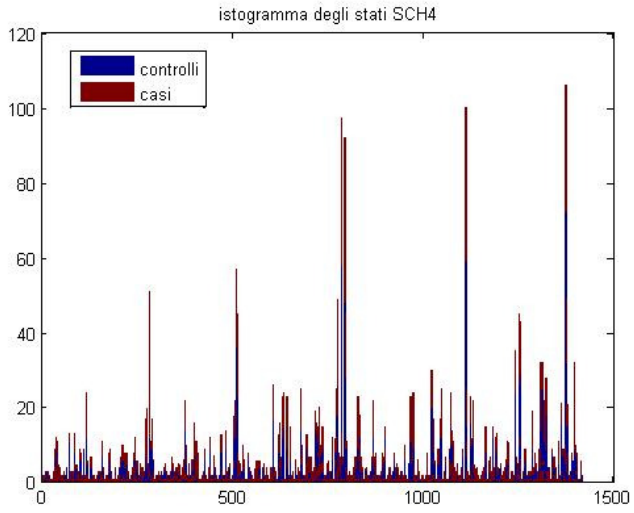
Questo porzione di rete riporta i geni FGF1, SHC4, MTOR, IGF1R, C15orf33, FGFR1, PPM1A, NUDT6 e FGF2 che risultano connessi tra di loro mediante uno o due collegamenti. Di seguito è riportata un'organizzazione semplificata di questa rete.

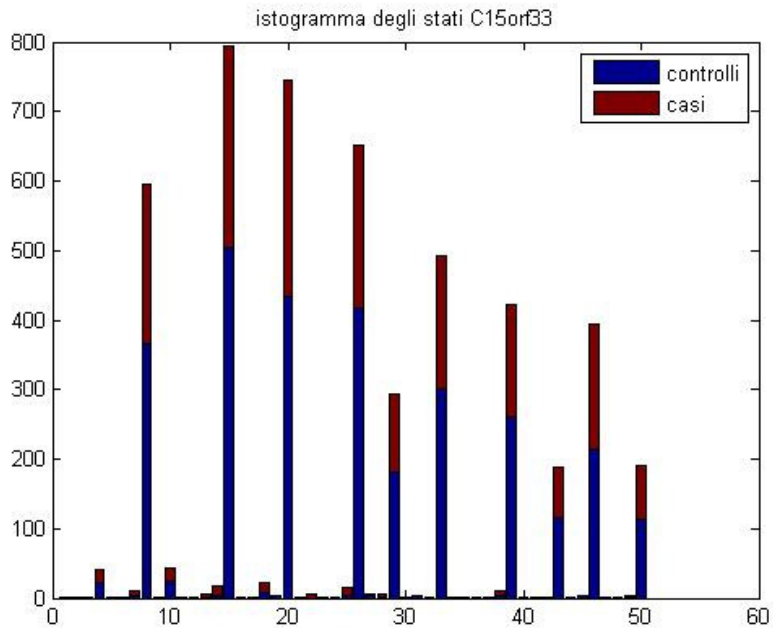


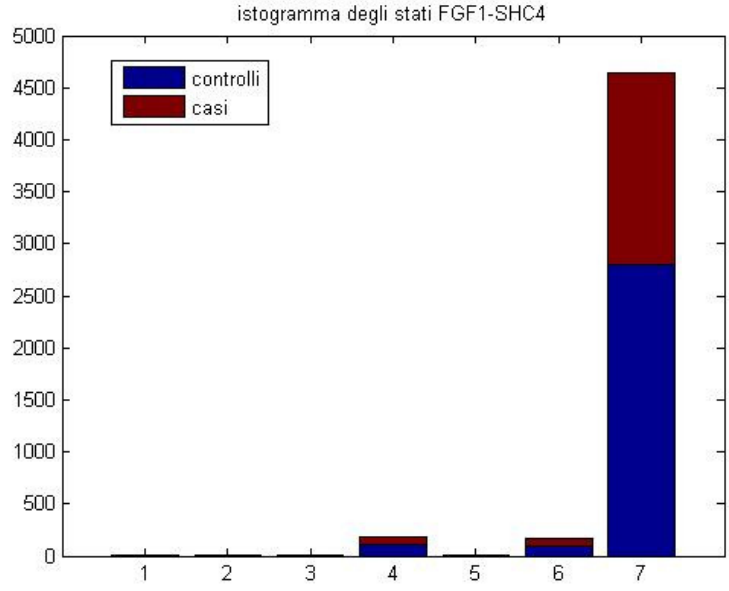
Di seguito sono riportati gli istogrammi di ciascuna metavariabile del grafico

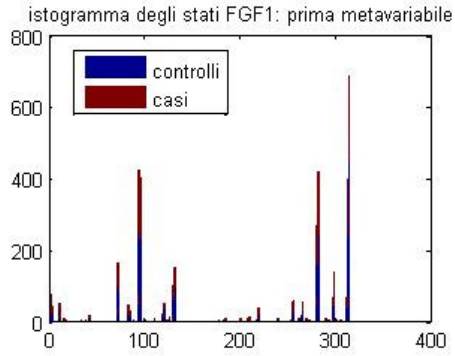

istogramma degli stati MTOR
controlli
casi
1600
1400
1200
1000
800
600
400
200
0
0
20
40
60
80
100
120
140

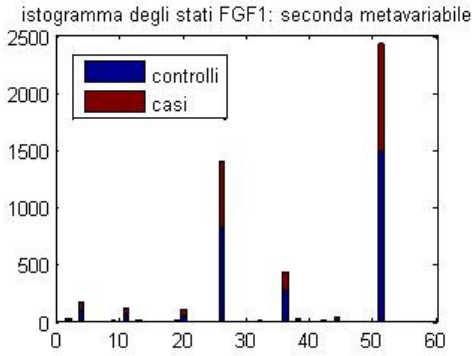

istogramma degli stati MTOR-SHC4
controlli
casi
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
1
2
3
4
5
6
7

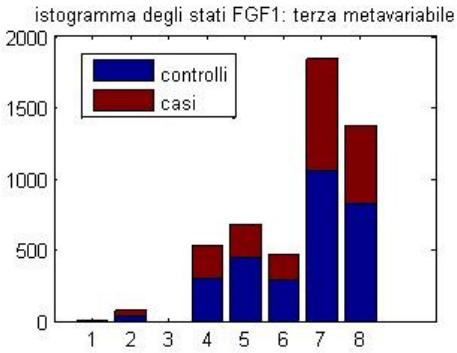

istogramma degli stati SCH4
controlli
casi
120
100
80
60
40
20
0
0
500
1000
1500

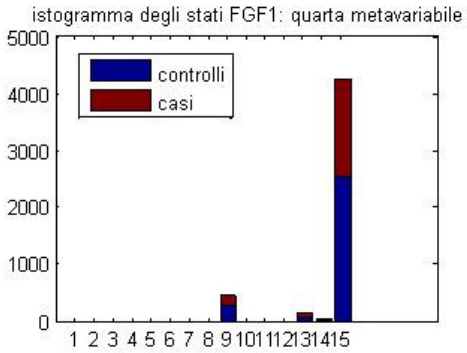

istogramma degli stati C15orf33
controlli
casi
800
700
600
500
400
300
200
100
0
0
10
20
30
40
50
60

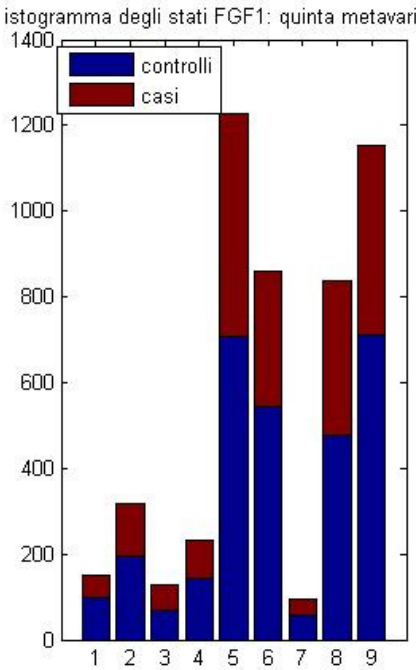

istogramma degli stati FGF1-SHC4
controlli
casi
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
1
2
3
4
5
6
7

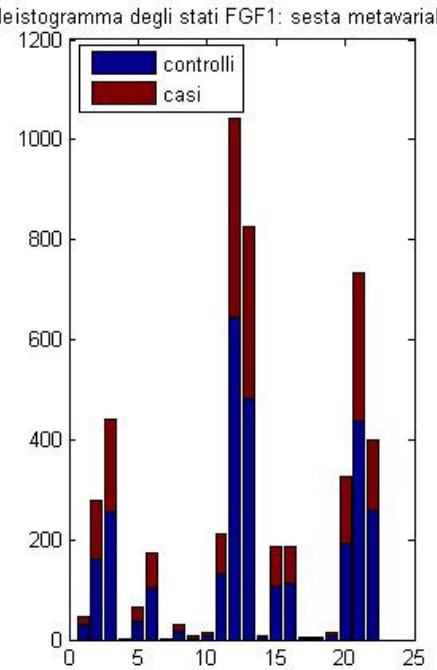

istogramma degli stati FGF1: prima metavariabile
controlli
casi
800
600
400
200
0
0
100
200
300
400

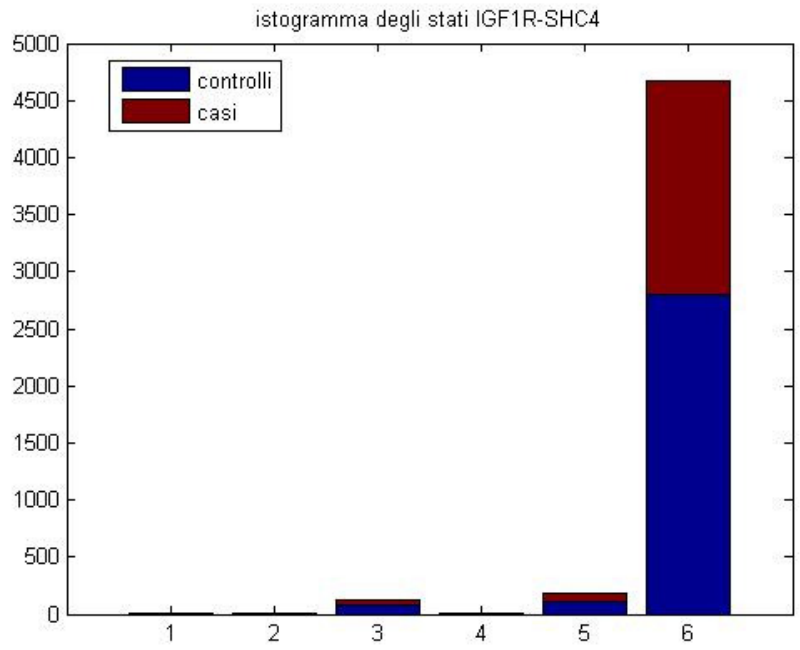

istogramma degli stati FGF1: seconda metavariabile
controlli
casi
2500
2000
1500
1000
500
0
0
10
20
30
40
50
60

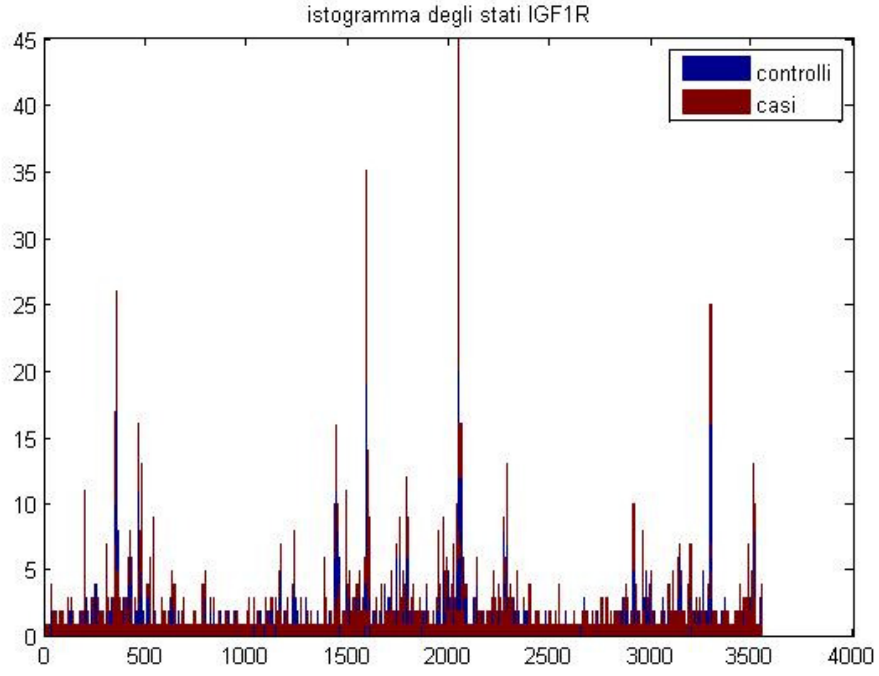

istogramma degli stati FGF1: terza metavariabile
controlli
casi
2000
1500
1000
500
0
1
2
3
4
5
6
7
8

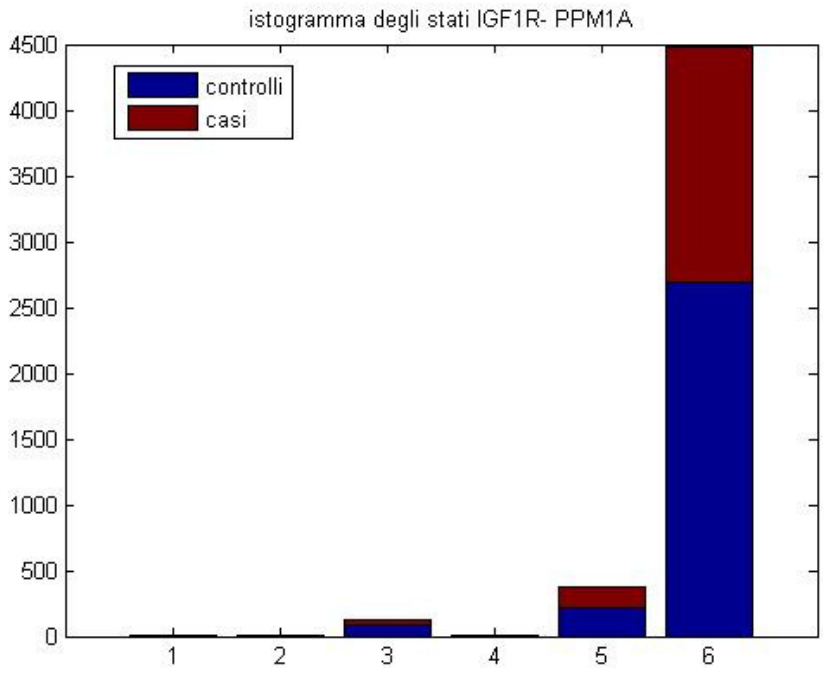

istogramma degli stati FGF1: quarta metavariabile
controlli
casi
5000
4000
3000
2000
1000
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

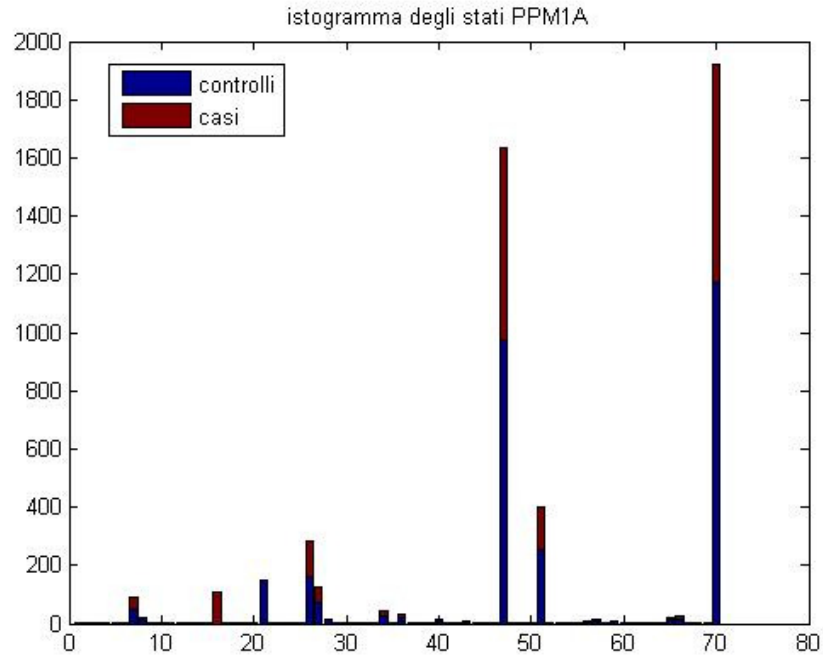

istogramma degli stati FGF1: quinta metavariabile
controlli
casi
1400
1200
1000
800
600
400
200
0
1
2
3
4
5
6
7
8
9

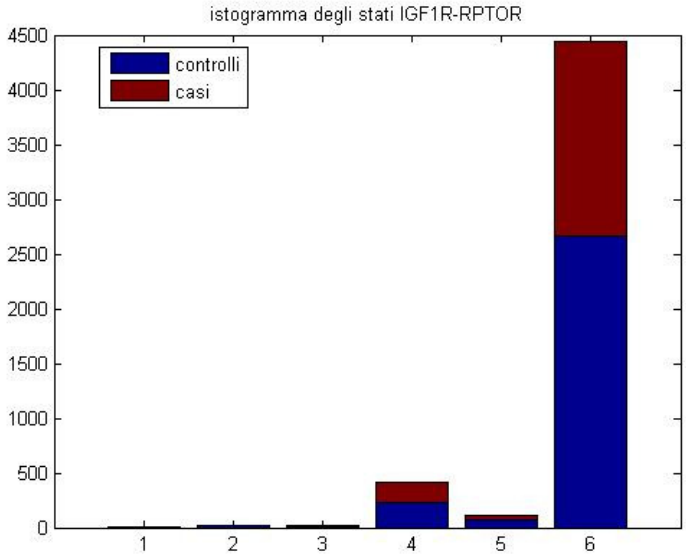

istogramma degli stati FGF1: sesta metavariabile
controlli
casi
1200
1000
800
600
400
200
0
0
5
10
15
20
25

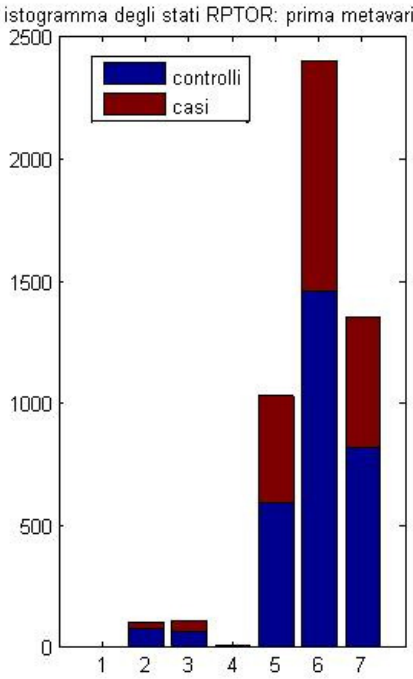

istogramma degli stati IGF1R-SHC4
controlli
casi
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
1
2
3
4
5
6

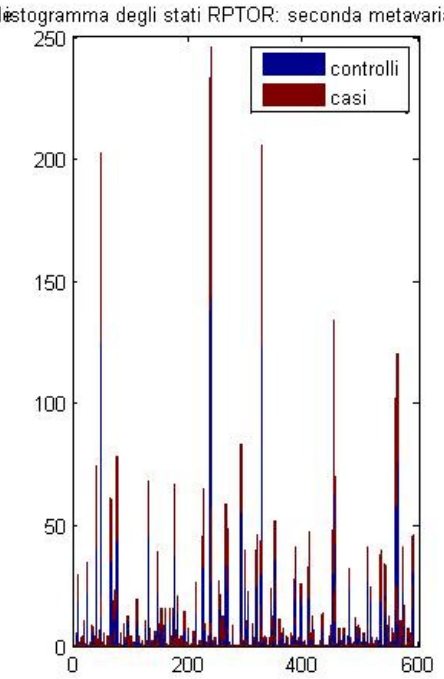

istogramma degli stati IGF1R
controlli
casi
45
40
35
30
25
20
15
10
5
0
0
500
1000
1500
2000
2500
3000
3500
4000

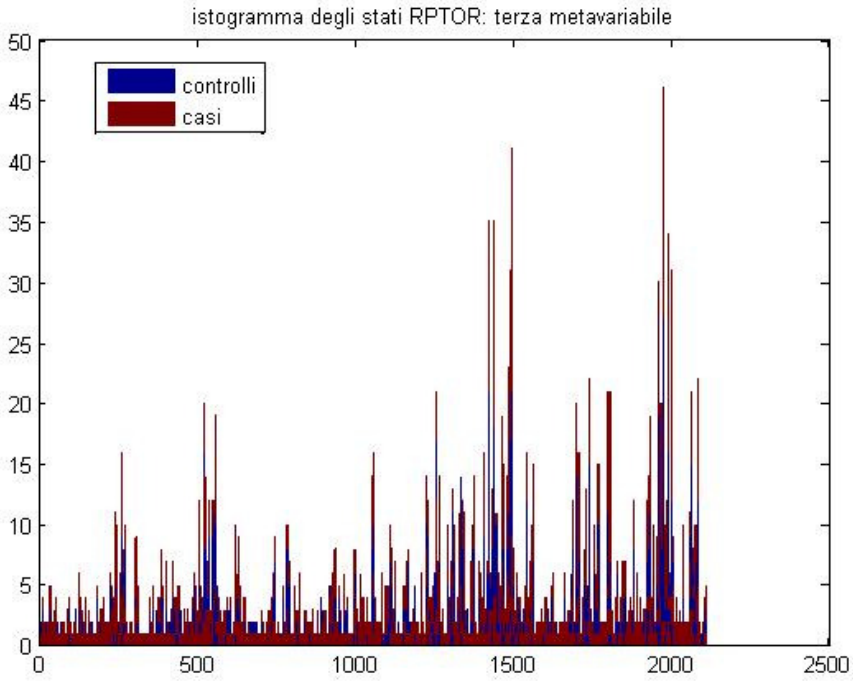

istogramma degli stati IGF1R- PPM1A
controlli
casi
4500
4000
3500
3000
2500
2000
1500
1000
500
0
1
2
3
4
5
6

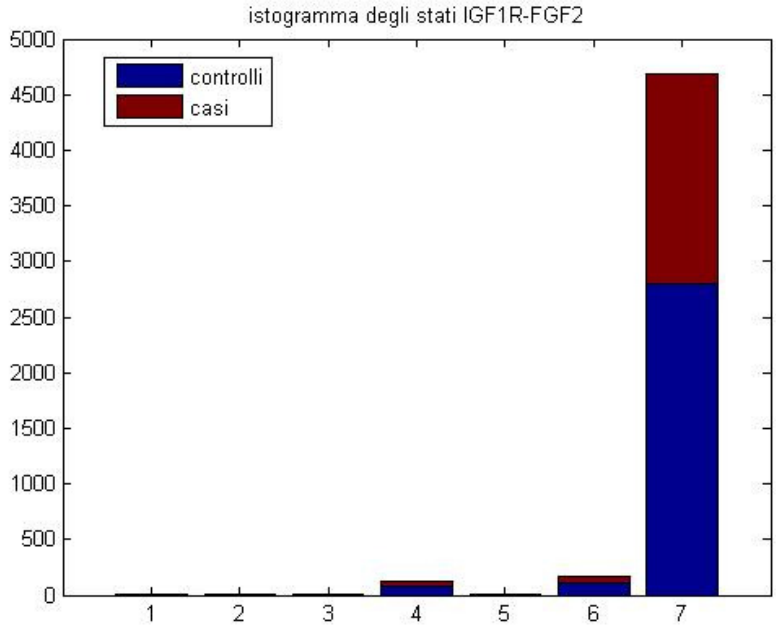

istogramma degli stati PPM1A
controlli
casi
2000
1800
1600
1400
1200
1000
800
600
400
200
0
0
10
20
30
40
50
60
70
80

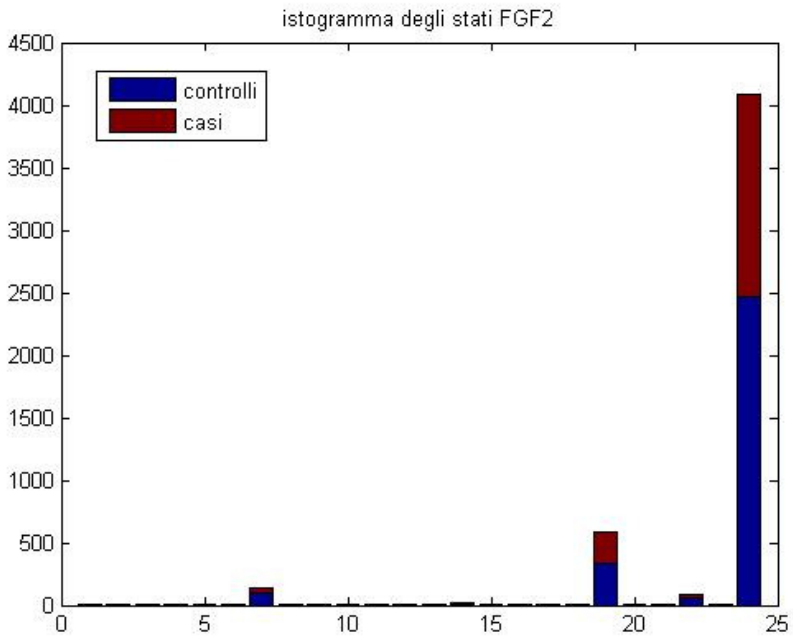

istogramma degli stati IGF1R-RPTOR
controlli
casi
4500
4000
3500
3000
2500
2000
1500
1000
500
0
1
2
3
4
5
6

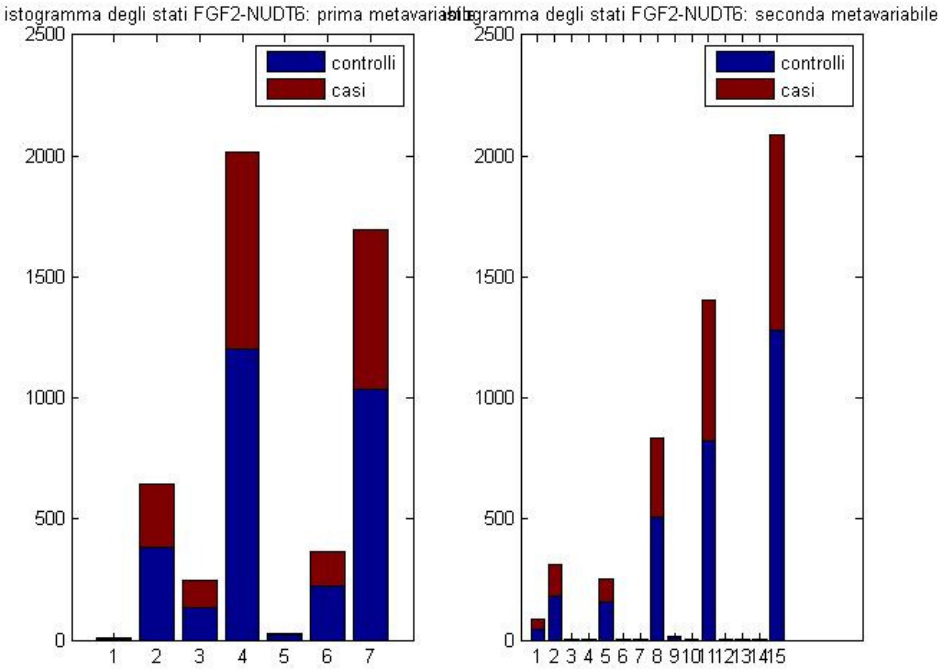

istogramma degli stati RPTOR: prima metavariabile
controlli
casi
2500
2000
1500
1000
500
0
1
2
3
4
5
6
7

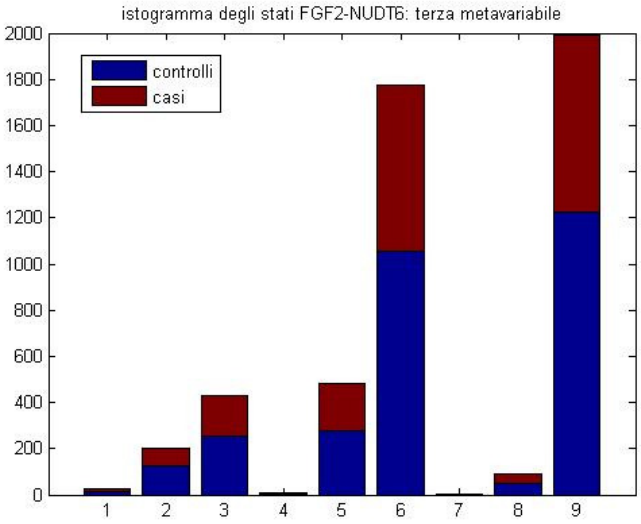

istogramma degli stati RPTOR: seconda metavariabile
controlli
casi
250
200
150
100
50
0
0
200
400
600

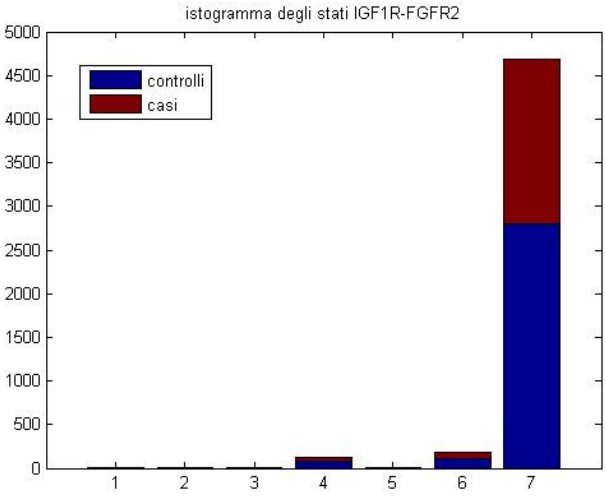

istogramma degli stati RPTOR: terza metavariabile
controlli
casi
0
5
10
15
20
25
30
35
40
45
50
0
500
1000
1500
2000
2500

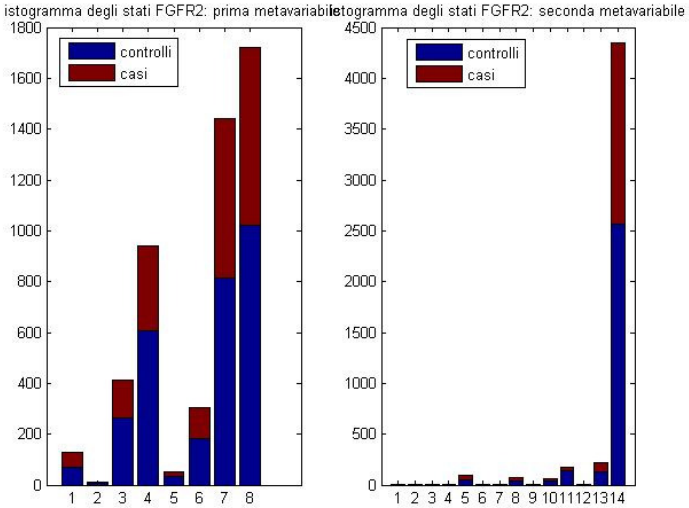

istogramma degli stati IGF1R-FGF2
controlli
casi
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
1
2
3
4
5
6
7

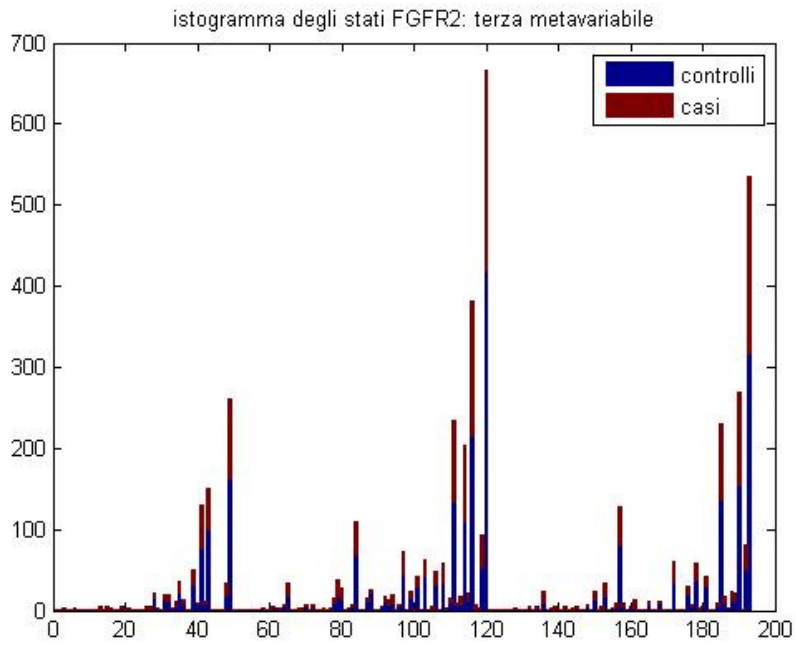

istogramma degli stati FGF2
controlli
casi
4500
4000
3500
3000
2500
2000
1500
1000
500
0
0
5
10
15
20
25


istogramma degli stati FGF2-NUDT6: prima metavariabile
controlli
casi
2500
2000
1500
1000
500
0
1
2
3
4
5
6
7
istogramma degli stati FGF2-NUDT6: seconda metavariabile
controlli
casi
2500
2000
1500
1000
500
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15


istogramma degli stati FGF2-NUDT6: terza metavariabile
controlli
casi
2000
1800
1600
1400
1200
1000
800
600
400
200
0
1 2 3 4 5 6 7 8 9

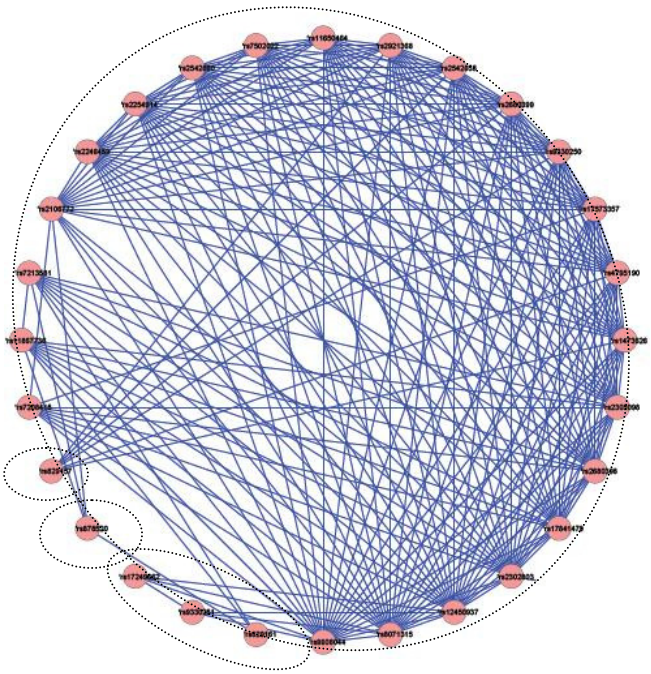

istogramma degli stati IGF1R-FGFR2
controlli
casi
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
1 2 3 4 5 6 7

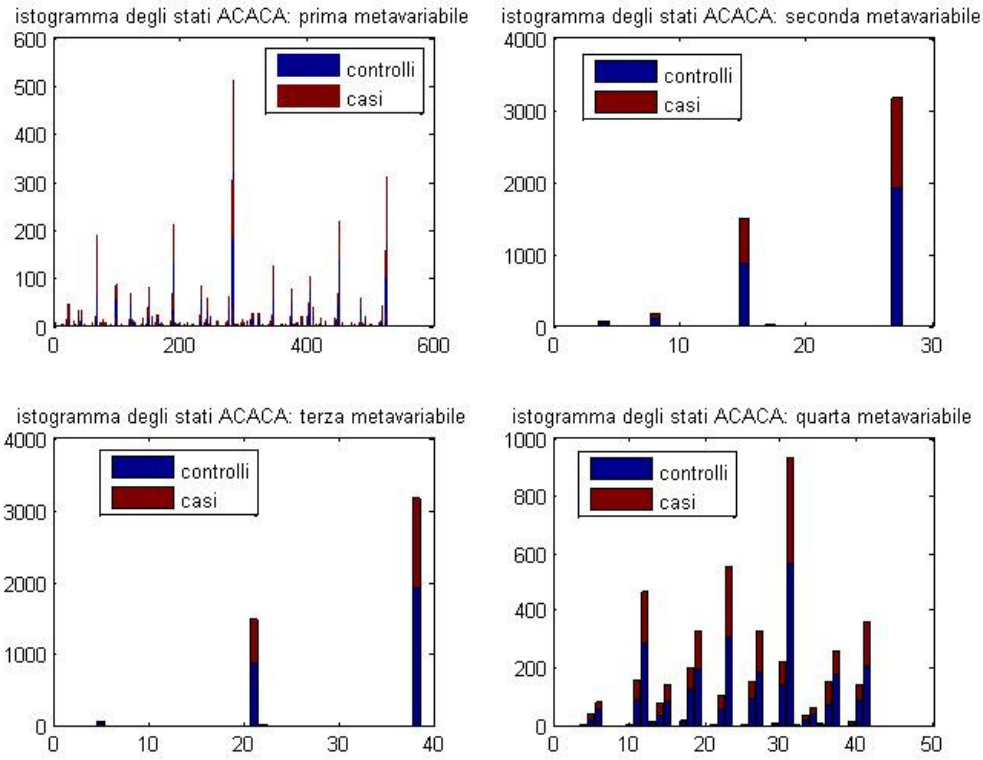

istogramma degli stati FGFR2: prima metavariabile
controlli
casi
1800
1600
1400
1200
1000
800
600
400
200
0
1 2 3 4 5 6 7 8
istogramma degli stati FGFR2: seconda metavariabile
controlli
casi
4500
4000
3500
3000
2500
2000
1500
1000
500
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14



È interessante notare come, a differenza della rete ricostruita nel caso del dataset precedente, qui le interazioni individuate sono per lo più tra geni appartenenti a cromosomi diversi. Viene inoltre riconfermata una connessione trovata anche nella rete precedente tra FGF2 e NUDT6.

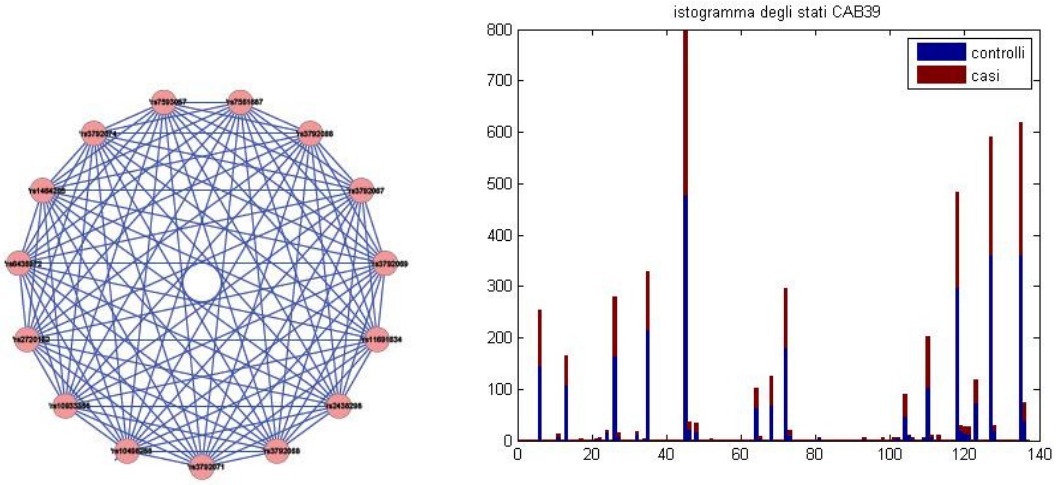
ACACA





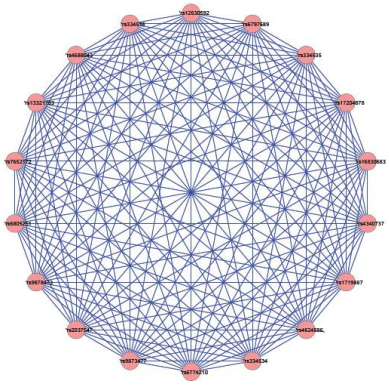
ACACA (cromosoma 17) ha 28 SNPs organizzati in 4 metavariabili di 524, 27, 38 e 41 stati ciascuna.

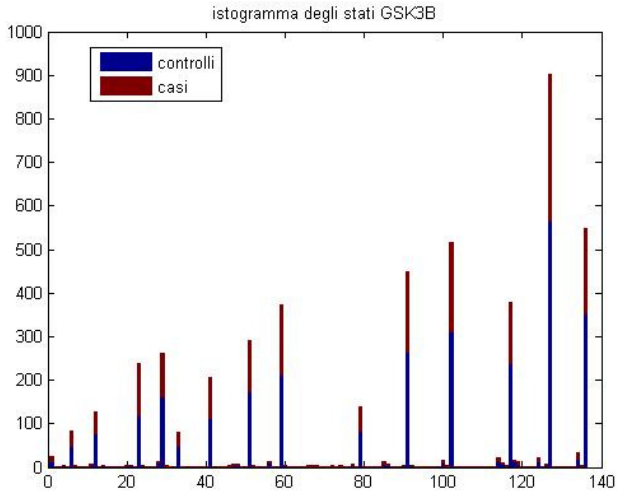
CAB39



CAB39 (cromosoma 2) costituisce un'unica metavariabile con 138 SNPs.

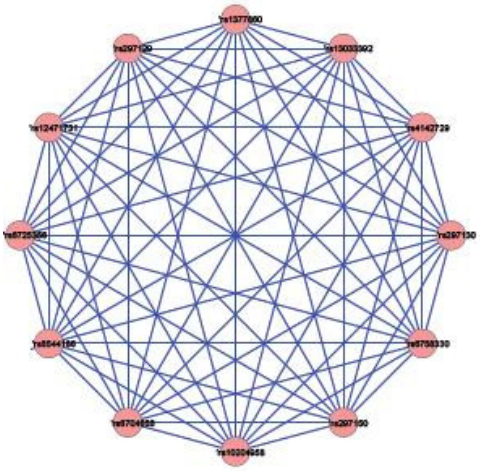
GSK3B

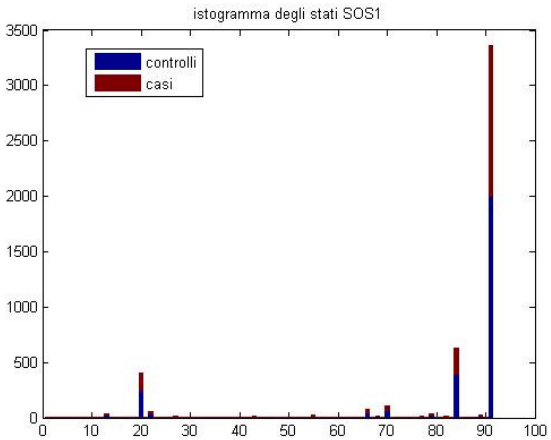




GSK3B (cromosoma 3) costituisce un'unica metavariabile con 138 stati.

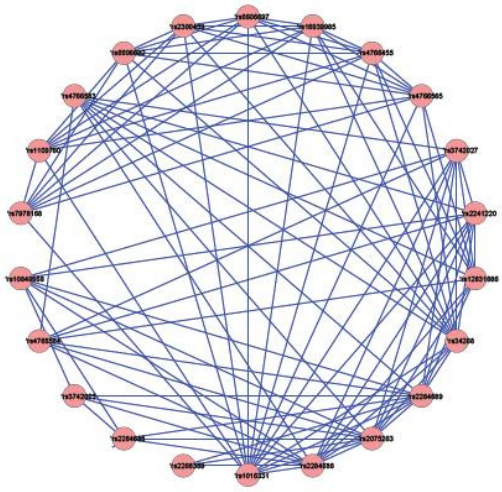
SOS1

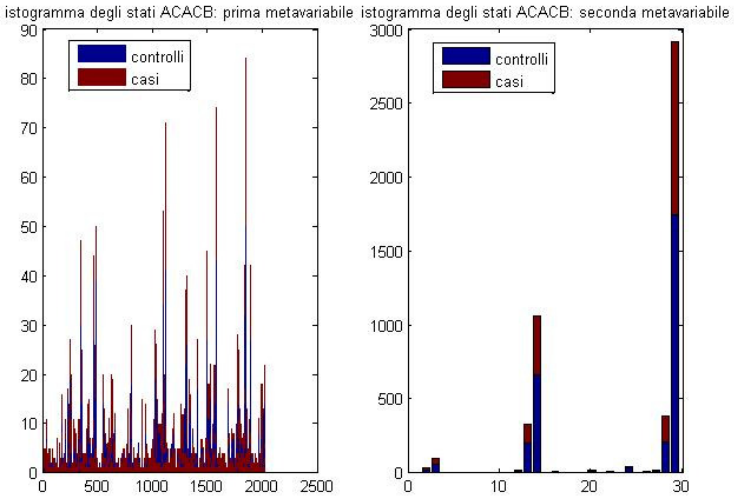




SOS1 (cromosoma 2) costituisce un'unica metavariabile con 52 stati.

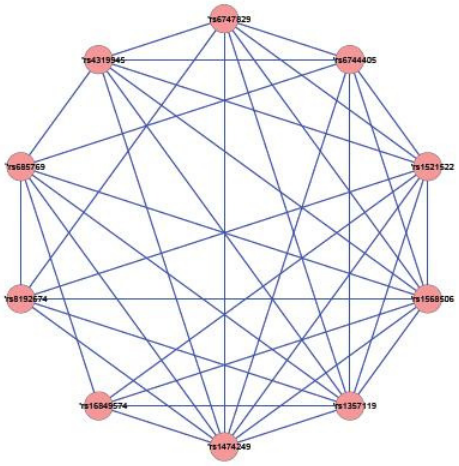
ACACB

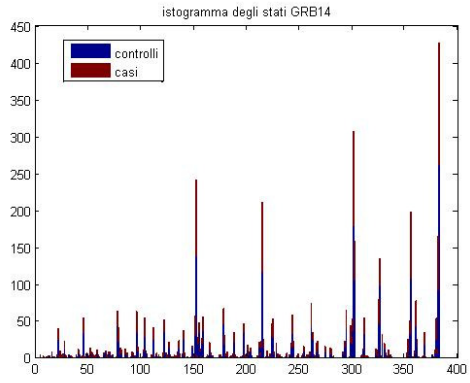




ACACB (cromosoma 12) costituisce due metavariabili di 2017 e 29 stati ciascuna.

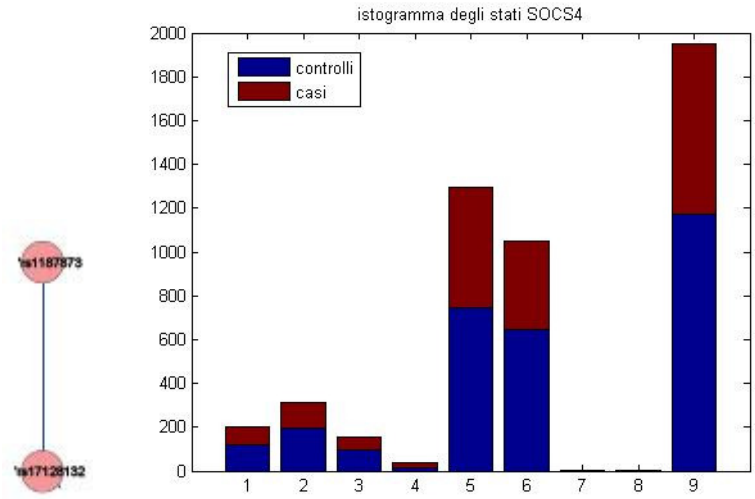
GRB14





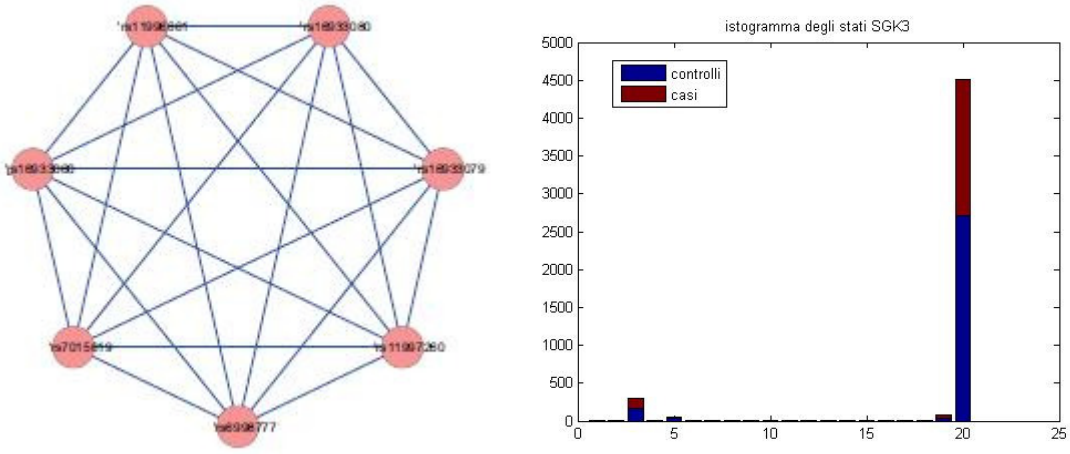
GRB14 (cromosoma 2) ha 10 SNPs e costituisce un'unica metavariabile con 382 stati.

SOCS4



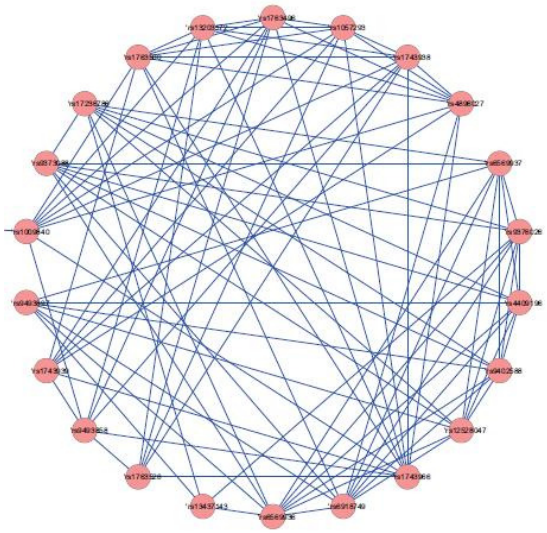
SOCS4 (cromosoma 14) ha due SNPs e costituisce un'unica metavariabile con 9 stati.

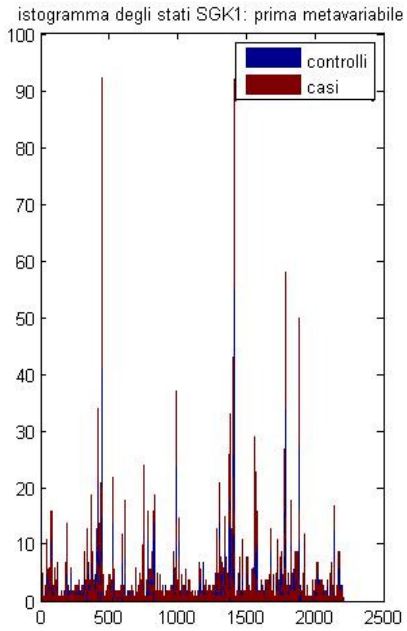
SGK3

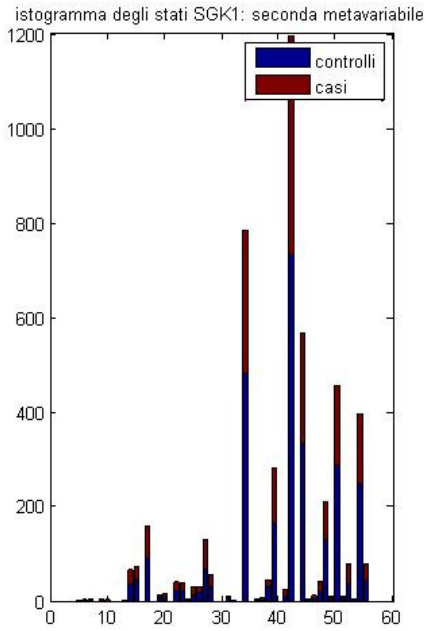


SGK3 (cromosoma 8) costituisce un'unica metavariabile con 20 stati (dall'istogramma risulta comunque evidente la netta prevalenza di due stati rispetto a tutti gli altri).

SGK1

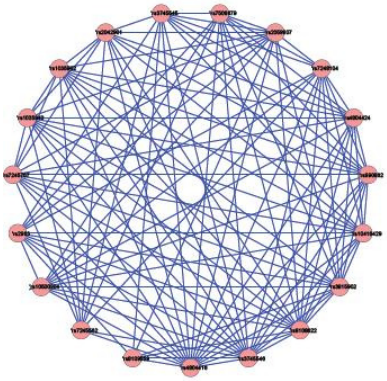


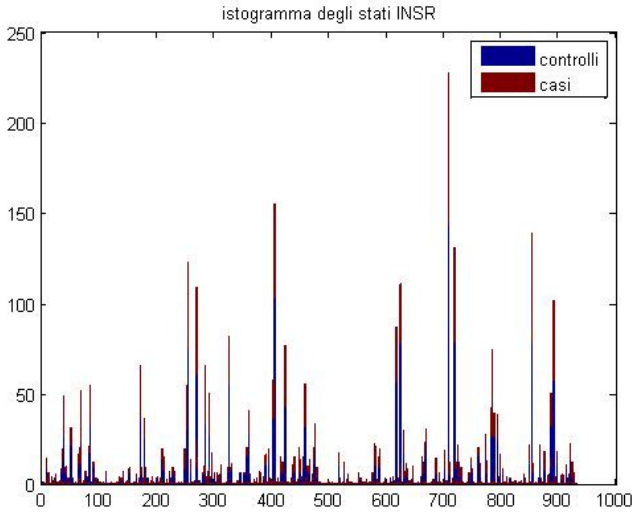




SGK1 (cromosoma 19) è costituito da due metavariabili con 2196 e 55 stati ciascuna.

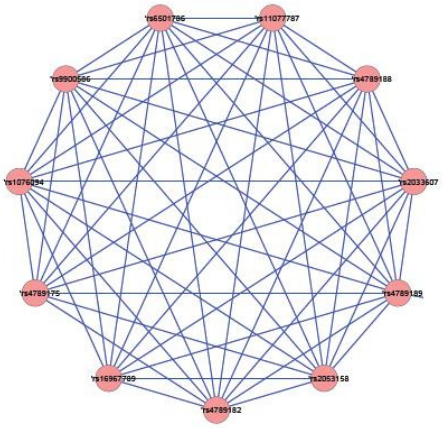
INSR

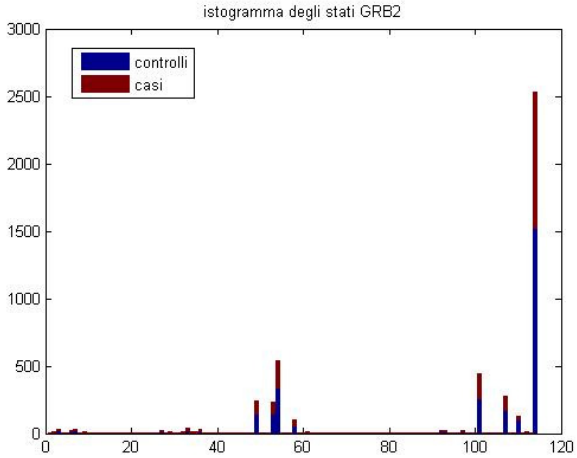




INSR (cromosoma 19) è costituito da un'unica metavariabile con 930 stati.

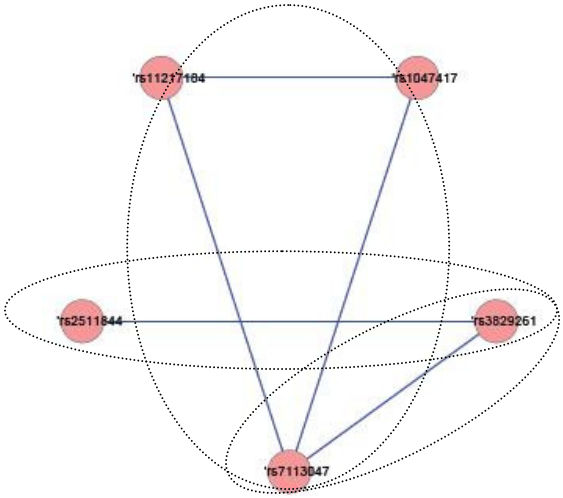
GRB2

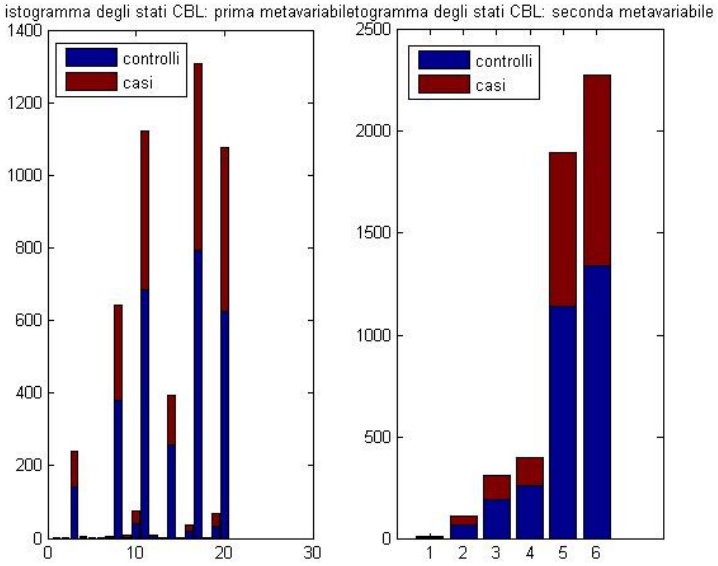


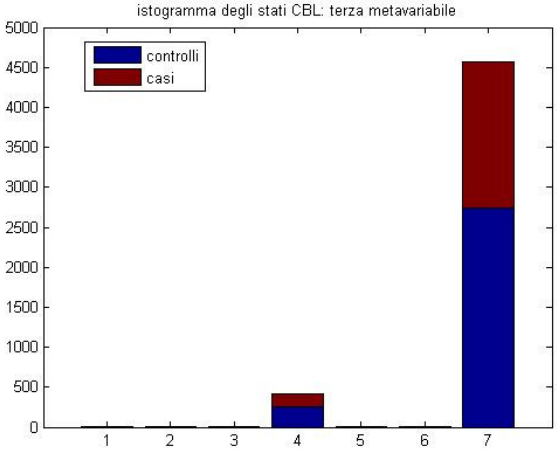


GRB2 (cromosoma 17) è costituito da 11 SNPs e costituisce un'unica variabile con 114 stati.

CBL


'rs1123?7184
'rs1047417
'rs2511844
'rs3829261
'rs7113047


istogramma degli stati CBL: prima metavariabile
controlli
casi
0
200
400
600
800
1000
1200
1400
0
10
20
30
istogramma degli stati CBL: seconda metavariabile
controlli
casi
0
500
1000
1500
2000
2500
1
2
3
4
5
6


istogramma degli stati CBL: terza metavariabile
controlli
casi
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
1
2
3
4
5
6
7

CBL (cromosoma 11) ha 5 SNPs ed è costituito da 3 metavariabili di 21, 6 e 7 stati ciascuna.

Possiamo concludere che le metavariabili ricostruite per il primo dataset sono sostanzialmente riconfermate nel secondo dataset. L'unica differenza che si può osservare è il numero di connessioni individuate tra geni diversi: per il primo dataset il numero di collegamenti tra geni è limitato a pochi casi e sostanzialmente tra geni appartenenti allo stesso cromosoma, a differenza di quanto si può osservare nel secondo dataset dove il numero di collegamenti tra geni diversi è più alto e le connessioni individuate riguardano per lo più geni localizzati su cromosomi diversi.

9.3 Risultati della classificazione naive Bayes

Mediante il software Orange si procede con l'applicazione dell'algoritmo di classificazione naive Bayes su entrambi i dataset.

Cominciando l'analisi con il dataset di partenza (con le 918 variabili) di controlli e casi di diabete di tipo 1, ed eseguendo la classificazione e il leave one out, dalla matrice di confusione si ottengono i risultati in Tabella 9.1.

	casi	Controlli	
casi	923	1077	2000
controlli	1209	1795	3004
	2132	2872	

Tabella 9.1: matrice di confusione del dataset di casi e controlli del diabete di tipo 1.

Assumendo che la matrice di organizzi come:

Veri positivi (true positive, TP)	Falsi positivi (false positive, FP)
Falsi negativi (false negative, FN)	Veri negativi (true negative, TN)

allora si può calcolare il coefficiente MCC andando a sostituire i valori contenuti in Tabella 8.1 nell'espressione riportata in paragrafo 7.2. L'MCC è pari a: 0.0585.

Per quanto riguarda l'applicazione dell'algoritmo di classificazione sul dataset iniziale con 918 marcatori di controlli e casi del diabete di tipo 2 si hanno risultati riportati in Tabella 9.2.

	casi	controlli	
casi	920	1079	2000
controlli	1226	1778	3004
	2132	2872	

Tabella 9.2: matrice di confusione del dataset di casi e controlli del diabete di tipo 2.

Assumendo che la tabella sia organizzata come nel caso precedente, l'MCC calcolato è pari a 0.0513. Dai risultati ottenuti, si può concludere che il pathway dell'insulina non è molto informativo ai fini della classificazione.

Ripetendo la stessa procedura utilizzando le metavariabili, mi aspetto una prestazione analoga. Con le 149 metavariabili ottenute dall'analisi sul primo dataset ottengo i risultati in Tabella 9.3. L'MCC calcolato è pari a 0.048.

	casi	controlli	
casi	2045	959	2000
controlli	1268	732	3004
	3313	1691	

Tabella 9.3: matrice di confusione delle metavariabili del dataset di casi e controlli del diabete di tipo 1.

Con le 116 variabili ricavate dall'analisi del secondo dataset si ottengono i risultati in Tabella 9.4. L'MCC calcolato è pari a 0.09587.

	casi	controlli	
casi	2043	961	2000
controlli	1172	827	3004
	3215	1788	

Tabella 9.4: matrice di confusione delle metavariabili del dataset di casi e controlli del diabete di tipo 2

Conclusioni

I test di associazione Genome Wide si pongono come obiettivo quello di individuare quegli SNPs, su tutto il genoma, che consentono di classificare i soggetti caso/controllo costituenti il dataset. È necessario operare, a monte della classificazione, una selezione del numero delle variabili da considerare. Lo stato dell'arte propone due approcci alternativi che implicano entrambi una riduzione dell'informazione di partenza.

In questa tesi si è proposto un approccio innovativo per la gestione del numero di variabili da analizzare. Rispetto ai metodi che si possono trovare allo stato dell'arte, il ricorso alle definizioni di entropia e mutua informazione ha reso possibile la costruzione di un numero ridotto di metavariabili mantenendo comunque intatta la quantità di informazione iniziale. La decisione di lavorare sul determinato pathway, piuttosto che su ogni singolo cromosoma, ha permesso di identificare alcune relazioni tra geni: soprattutto per quanto riguarda il dataset di controlli e casi di diabete di tipo 2 si trovano numerosi collegamenti tra geni posizionati anche su cromosomi diversi. I risultati della classificazione non sono molto buoni, ma la prestazione è più che altro influenzata dal fatto che il solo pathway analizzato in questa tesi, quello dell'insulina contiene dei dati non estremamente informativi per la classificazione dei soggetti del problema in esame.

Questo tipo di approccio è ovviamente adattabile a qualsiasi tipo di dataset nell'ambito degli studi di associazione caso/controllo ed è possibile estendere la ricerca a più pathway nell'intento di individuare i meccanismi di regolazione alla base della *suscettibilità* alle malattie a base genica. È sempre più accertato infatti che la causa della comparsa di una determinata patologia a base genica non sia imputabile all'azione di un unico gene, ma all'effetto combinato di più fattori. Un approccio di questo tipo, mirato allo studio di un pathway piuttosto che dei singoli cromosomi, rende possibile studiare il problema da un punto di vista più globale. Il metodo che si vuole proporre è quello di analizzare il dataset considerando ogni volta un pathway diverso e costruendo per ognuno un classificatore. In ultimo passo, aggregando tutti i classificatori assieme, si è in grado di avere una visione d'insieme del problema avendo raccolta tutta l'informazione del dataset iniziale.

Bibliografia

1. Smith AV, Thomas DJ, Munro HM, Abecasis GR: Sequence features in regions of weak and strong linkage disequilibrium, Genome Res 2005;
2. Devlin B, Rish N:A comparison of linkage disequilibrium measures for fine scale mapping, Genomics, 1995;
3. Klug WS, Spencer CA: Concetti di Genetica, Prentice Hall, 8va Edizione, 2007
4. Nothnagel M., Furts R., Rohde K: Entropy as a Measure for Linkage Disequilibtium over Multilocus Haplotype Blocks, Hum Hered 2002,54:186-198;
5. Balding DJ: A Tutorial on statistical methods for population association studies, Nature Reviews, Genetics, 2006;
6. Gabriel SB, Schaffner SF: The structure of haplotype blocks in the human genome, Science, vol 296, 2002.
7. Bonizzoni P., Della Vedova G., Dondi R., Li J., The Haplotyping problem: an overview of computational models and solutions,
8. Marchini J, Cardon LR, Philips M, Donnelly P, The effects of human population structure on large genetic association studies, Nature Genetics, 2004.
9. Price AL., Zaitlen NA., Reich D, Patterson N, New approaches to population stratification in genome-wide association studies, Nature Rev Genetics, 2010.
10. Freedman ML, et al, Assessing the impact of population stratification on genetic association studies, Nat Genetics, 2004
11. Cardon LR, Palmer LJ, Population stratification and spurious allelic association, The Lancet, 2003;
12. The International HapMap Consortium, The international HapMap project, Nature Publishing Group, 2003;
13. A second generation human aplotype map of over 3.1 million SNPs, The international HapMap project, Nature, 2007
14. Thorisson GA, Smith A., The international HapMap project Web site, Genome Research, 2005.
15. Stram D.O., Tag SNP selection for Association Studies, Genetic Epidemiology 27 (365-374), 2004

16. Montpetit A., Nelis M., An evaluation of the performance of tagSNPS derived from HapMap in a Caucasian Population, Genetics, 2003
17. Chen J., Chatterjee N., Exploiting Hardy-Weinberg Equilibrium for Efficient Screening of Single SNP Associations for case-control studies., Human Heredity 2006
18. Willer C., Scott L., Tag SNP selection for Finnish individuals based on the CEPH Utah HapMap database, Gen Epidemiology., 2006
19. Purcell S., Neale B., PLINK: a tool set for Whole-genome association and population-based linkage analyses., The Am J of Hum Gen., 2007
20. Zeggini E., Scott L., Meta-analysis of genome wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes., Nature Genetics, 2008
21. Nothnagel N., Furts R., Entropy as a Measure for linkage disequilibrium over multilocus Haplotype Blocks., Hum Hered., 2002.
22. Balding DJ., A tutorial on statistical methods for population association studies, Nature review, Genetics, 2006
23. Lewis C., Genetic association studies: design, analysis and interpretation, 2002.
24. Genome Wide association analysis identifies loci for type 2 diabetes and triglyceride levels., Science, 2007.