UNIVERSITÀ DEGLI STUDI DI VERONA

DIPARTIMENTO MATERNO INFANTILE E DI BIOLOGIA-GENETICA
SEZIONE DI BIOLOGIA E GENETICA

DOTTORATO DI RICERCA IN
BIOTECNOLOGIE APPLICATE ALLE SCIENZE BIOMEDICHE

CICLO XX

TITOLO DELLA TESI DI DOTTORATO

APPROCCIO INFORMATICO NELLO STUDIO DI MALATTIE COMPLESSE PER L'INDIVIDUAZIONE DI POSSIBILI PARENTELE NON NOTE TRAMITE L'ANALISI DEL GENOTIPO IN INDIVIDUI NON RELATI

S.S.D. BIO/13

Coordinatore: Prof. G.F. FUMAGALLI
Firma _____

Tutor: Prof. P.F. PIGNATTI
Firma _____

Dottorando: Dott. LUCIANO XUMERLE
Firma _____

Riassunto

Uno studio per la ricerca di componenti genetiche di malattie complesse richiede, al giorno d'oggi, la caratterizzazione di migliaia di individui per centinaia di migliaia di marcatori. Gli studi di associazione caso-controllo necessitano della presenza di individui non imparentati tra loro per stimare correttamente la frequenza allelica. La presenza di parentele non rilevate potrebbe confondere la stima di tali frequenze in casi e/o nei controlli e portare il ricercatore a descrivere associazioni tra genotipo e fenotipo errate.

Al fine di identificare se nel gruppo studiato sono presenti coppie di individui aventi un possibile grado di parentela, non identificato nella fase di raccolta del campione, abbiamo sviluppato una procedura *in silico* che compara la probabilità di parentela rispetto alla probabilità di non parentela tra due individui condizionata al genotipo. I casi testati sono i più comuni: I grado (padre-figlio e coppia di fratelli) e II grado (zio-nipote e nonno-nipote).

Lo studio mostra come il linkage disequilibrium tra i marcatori diminuisca l'informatività dei marcatori stessi nei test di parentela. Ad esempio, per supportare l'ipotesi di una parentela di II grado con un potere del 80% ed un rate di falsi positivi del 5% necessitano: 100 SNP indipendenti tra loro o 275 SNP organizzati a blocchi di 5 SNP in linkage disequilibrium con un $r^2 \geq 0.4$ e fase sconosciuta. Tuttavia, ricostruire probabilisticamente la fase degli aplotipi aumenta l'informatività dei marcatori a disposizione. Ad esempio, per supportare l'ipotesi suddetta con aplotipi a fase ricostruita necessitano: 20 aplotipi (100 SNP a blocchi di 5 con $r^2 \geq 0.4$) o 40 aplotipi (200 SNP a blocchi di 5 con $r^2 \geq 0.8$).

Indice

Capitolo 1

Introduzione

1.1 Malattie Complesse

Le malattie complesse (o multifattoriali) sono malattie dove sia fattori ambientali che genetici giocano un ruolo nella determinazione del rischio di sviluppare la patologia (ad esempio, asma e osteoporosi) [Lander and Schork, 1994]. Poichè le malattie complesse si manifestano con una elevata prevalenza nella popolazione, sono studiate in campo medico con sempre maggior interesse [GlaxoSmithKline and SIGU, 2002]. L'obiettivo primario di una ricerca genetica su malattie multifattoriali è di identificare fattori genetici che possono portare ad una variazione del rischio malattia. Il modello di trasmissione ed i meccanismi molecolari associati nell'insorgenza e sviluppo di tali patologie sono spesso sconosciuti e la ricerca di fattori genetici coinvolti richiede la disponibilità di molte informazioni cliniche e genetiche per un numero elevato di individui. Inoltre le componenti geniche, ognuna delle quali potrebbe dare un piccolo contributo per la determinazione di un profilo di rischio, devono essere valutate

assieme a quelle ambientali [Stephens and Humphries, 2003].

Il passaggio dallo studio delle malattie monogeniche, dove l'alterazione del gene è la causa della malattia, alle malattie multifattoriali, dove l'alterazione è associata ad un rischio, ha visto un cambiamento nelle metodologie applicate alla ricerca dei fattori genetici che sono associati ad una aumentata suscettibilità alla malattia (Figura 1.1).

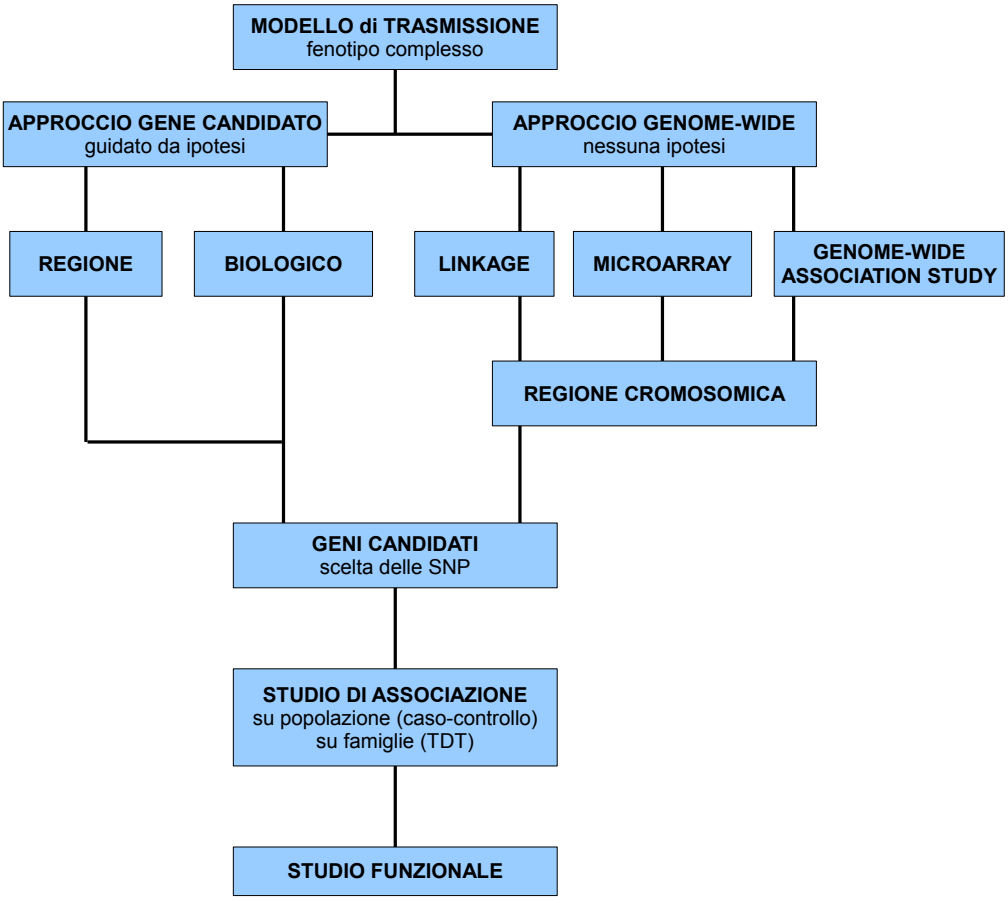


Figura 1.1. Il diagramma di flusso mostra schematicamente i diversi approcci che il ricercatore può utilizzare nell'individuare i fattori genetici coinvolti nella variazione del rischio di malattia.

Esistono diverse strategie per identificare geni associati alle malattie

complesse: attraverso lo studio di geni candidati [Hirschhorn and Daly, 2005] (scelti sulla base delle conoscenze biochimiche e fisiologiche che possono legarsi all'eziopatogenesi della malattia); tramite scansioni genomiche in cui vengono analizzati molti marcatori lungo tutto il genoma [Wills-Karp and Ewart, 2004]; studi di associazione Genome-Wide in cui viene condotta un'analisi di associazione usando centinaia di migliaia di SNP lungo tutto il genoma [Hirschhorn and Daly, 2005]. Dopo aver identificato una regione come associata, è possibile condurre uno studio dettagliato di linkage disequilibrium [Risch, 2000].

Gli studi di associazione si propongono di identificare i fattori che vanno a modificare il rischio malattia attraverso l'utilizzo delle varianti genetiche.

La forma più semplice di studio di associazione è lo studio *caso-controllo*: viene impiegato per valutare il ruolo di uno o più fattori di rischio (ambientali e/o genetici) nell'eziopatogenesi di una malattia. L'impostazione dello studio considera due gruppi di soggetti: i malati che costituiscono i casi ed i controlli ovvero i soggetti con le stesse caratteristiche dei primi ma dai quali differiscono solo per il fatto che non presentano la malattia. L'attendibilità e l'affidabilità dello studio dipendono evidentemente dalla corretta selezione dei casi e dei controlli. Un'associazione tra malattia e fattore di rischio è presente quando una percentuale degli esposti tra i casi è significativamente maggiore di quella degli esposti nel gruppo di controllo. La presenza di un'associazione statistica significativa non comporta però direttamente la conclusione dell'esistenza di una relazione causa-effetto, l'associazione tra fenotipo e genotipo potrebbe essere data dal caso (falso positivo) [Grimes and Schulz, 2002].

I risultati dello studio caso-controllo risentono della presenza di fattori di distorsione (bias) che sono spesso occulti e che possono portare il ricercatore ad una falsa associazione tra i fattori studiati. Il primo bias da considerare è il

bias di selezione, che può verificarsi per una scelta inadeguata dei casi o, più comunemente, dei controlli. Se il gruppo dei controlli non è rappresentativo della popolazione generale, l'eventuale associazione osservata potrebbe dipendere dalle diverse frequenze alleliche proprie dei due gruppi e non da alleli associati con il fenotipo in studio [Hirschhorn and Daly, 2005]. Per evitare i bias di selezione è possibile condurre un'analisi di associazione che utilizzi le famiglie come, ad esempio, il *transmission disequilibrium test (TDT)* in cui il numero osservato di alleli trasmessi da genitori eterozigoti a figli affetti è comparato con quello atteso [Spielman and Ewens, 1996].

Gli studi di associazione sono normalmente condotti su individui non imparentati perciò la presenza di parentele non identificate all'interno del campione porterebbe ad una valutazione errata dell'associazione ed ad una stima errata delle frequenze alleliche e genotipiche del gruppo in studio [Trégouët et al., 1997].

Per escludere la presenza di errori sui dati vengono condotte alcune semplici analisi statistiche utili per produrre una statistica descrittiva del campione analizzato come, ad esempio, il calcolo della frequenza degli alleli, dei genotipi, degli individui non genotipizzati per ogni marcatore. Nel caso di studi familiari viene abitualmente testata la segregazione degli alleli all'interno della famiglia. Per individuare alterazioni delle frequenze nel campione uno dei test abitualmente utilizzati è il test dell'equilibrio di *Hardy-Weinberg* in cui le frequenze dei genotipi osservati vengono comparate rispetto la distribuzione attesa sotto l'ipotesi del libero assortimento degli alleli negli individui di una popolazione [Gomes et al., 1999]. Le cause di una differenza significativa tra le frequenze dei genotipi osservati e le frequenze attese possono essere attribuite ad un qualsiasi fenomeno biologico che impedisce il libero assortimento degli alleli (fenomeni selettivi dove

viene selezionato un particolare genotipo e fenomeni migratori) oppure a cause sperimentali (una non accurata lettura dei genotipi) [Suzuki et al., 1992].

1.2 Test di parentela basato sull'analisi del DNA

Un test di parentela basato sull'analisi del DNA è un esame eseguito per verificare se due o più persone sono parenti biologici. Questo tipo di test include il Test di Paternità, il Test di Zigosità, l'Analisi dei Nonni, utile per verificare l'eventuale relazione biologica tra una persona ed i suoi nonni presunti, così come l'Analisi degli Zii. Il test di Parentela viene di solito eseguito per motivi legali come nel caso di paternità dubbie; eredità controverse; verifica dell'albero genealogico; custodia di minori [Justice, 2000].

Il test viene condotto estraendo il DNA dai campioni biologici (es. sangue o saliva) degli individui coinvolti ed esaminando l'ipotesi di parentela contro l'ipotesi di non parentela su un numero sufficiente di marcatori altamente polimorfici sul DNA. Per aumentare l'efficacia di un test di consanguineità l'analisi può essere estesa al cromosoma *Y (YSTR)* per determinare la linea paterna in caso di soggetti maschi o al *DNA Mitocondriale (mtDNA)* per determinare la linea materna. Il DNA mitocondriale (mtDNA) rappresenta l'informazione genetica contenuta nei mitocondri, presenti nel citoplasma della cellula alla quale forniscono quasi tutta l'energia di cui ha bisogno. Il DNA mitocondriale viene ereditato esclusivamente dalla madre, ogni individuo entro una determinata linea materna avrà lo stesso DNA mitocondriale [Justice, 2000].

Un buon marcatore per studiare la parentela tra individui dovrebbe presentare un'elevata *eterozigosità*. L'eterozigosità (1.2) di un marcatore indica il numero

di individui eterozigoti che sono attesi in una popolazione in equilibrio di Hardy-Weinberg. Similmente l'omozigosità (1.1) indica il numero di individui omozigoti [Weir, 1996].

$$H = \sum_{i=1}^{n} p_i^2 \tag{1.1}$$

$$E = 1 - \sum_{i=1}^{n} p_i^2 = 1 - H \tag{1.2}$$

L'elevata *eterozigosità* consente di stabilire l'origine di un allele: se l'allele portato da un individuo è presente nel padre e non nella madre, ovviamente l'individuo lo avrà ricevuto dal padre. È chiaro quindi che tanto maggiore sarà l'*eterozigosità* di un locus marcatore, tanto maggiore sarà la possibilità che i genitori abbiano genotipo differente e maggiore sarà la probabilità di individuare l'origine dell'allele [Presciuttini et al., 2002]. Una delle ragioni della grande diffusione dei microsatelliti negli studi di linkage e nelle analisi di parentela è dovuto proprio all'elevatissima eterozigosità che essi presentano.

Il grado di informatività, cioè di efficacia, di un marcatore genetico viene misurato con il parametro PIC (Polymorphism Information Content) che rappresenta la proporzione di individui informativi, cioè per i quali è possibile distinguere la provenienza parentale degli alleli [Botstein et al., 1980].

$$PIC = 1 - \sum_{i=1}^{n} p_i^2 - 2\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} p_i^2 * p_j^2 \tag{1.3}$$

La formula (1.3) esprime numericamente la percentuale di individui informativi. Gli individui non informativi (non possiamo attribuire la provenienza

paterna o materna degli alleli) sono gli individui omozigoti e tutti i figli eterozigoti nati da genitori che sono entrambi eterozigoti [Botstein et al., 1980].

1.3 Linkage Disequilibrium

In genetica di popolazione il *linkage disequilibrium* (LD) indica l'associazione non casuale di alleli appartenenti a due o più loci, non necessariamente sullo stesso cromosoma. Il LD descrive una situazione in cui alcune combinazioni di alleli di diversi marcatori si osservano più o meno frequentemente di quanto ci si attenderebbe supponendo la libera formazione degli aplotipi date le frequenze alleliche osservate [Devlin and Risch, 1995].

Supponiamo di avere due marcatori genetici A e B le cui frequenze aplotipiche (1.4) ed alleliche (1.5) siano rispettivamente:

$$\begin{aligned} A_1B_1 &= x_{11} \\ A_1B_2 &= x_{12} \\ A_2B_1 &= x_{21} \\ A_2B_2 &= x_{22} \end{aligned} \tag{1.4}$$

dove x_{11} è la frequenza osservata dell'aplotipo A_1B_1 costituito dal primo allele di A e dal primo allele di B e così via e:

$$\begin{aligned} p_1 &= x_{11} + x_{12} \\ p_2 &= x_{21} + x_{22} \\ q_1 &= x_{11} + x_{21} \\ q_2 &= x_{12} + x_{22} \end{aligned} \tag{1.5}$$

dove con p_1 e p_2 indichiamo la frequenza del primo e secondo allele del marcatore A rispettivamente e con q_1 e q_2 le frequenze del primo e secondo allele di B rispettivamente (le frequenze alleliche si possono ricavare dalle frequenze aplotipiche). Utilizzando le frequenze degli aplotipi e degli alleli è possibile definire il parametro D per la misurazione del LD:

$$D = x_{11} - p_1 q_1 \tag{1.6}$$

in cui $p_1 q_1$ è la frequenza attesa (in caso di libera segregazione degli alleli) dell'aplotipo $A_1 B_1$. Nel caso di presenza di *linkage equilibrium* (assenza di LD) si osserva $D = 0$.

I valori estremi di D sono funzione delle frequenze alleliche nei due loci e così il massimo (minimo) $D = 0.25$ ($D = -0.25$) per un marcatore biallelico è osservabile solo quando tutti e quattro gli alleli hanno frequenza 0.5 [Hedrick and Kumar, 2001].

Per ovviare a questo inconveniente *Lewontin* suggerì una normalizzazione di D definita come:

$$D^1 = \frac{D}{D_{max}} \tag{1.7}$$

Dove D_{max} è il massimo disequilibrio possibile per le frequenze alleliche osservate. Se $D > 0$ allora $D_{max} = min(p_1 q_2, p_1 q_1)$, mentre se $D < 0$ consideriamo il valore assoluto di D_1 e $D_{max} = min(p_1 q_1, p_2 q_2)$.

Sviluppi teorici nello studio del decadimento del LD all'aumentare della distanza tra marcatori hanno portato ad un ulteriore misura di LD (1.8):

$$r^2 = \frac{D^2}{p_1 p_2 q_1 q_2} \tag{1.8}$$

1.4 Scopo del Lavoro

L'analisi di associazione caso-controllo è usualmente condotta su un gruppo di individui non imparentati. La presenza di parentele non note all'interno del dataset potrebbe confondere l'analisi di associazione. Saranno presentati strumenti informatici in grado di calcolare la probabilità condizionata ai genotipi che due individui presi a caso nella popolazione siano imparentati tra di loro. Saranno stimati, attraverso la simulazione di un set di 1000 famiglie, il numero di marcatori necessari per ottenere un rate di falsi positivi inferiore al 5% con un potere dell'80% per parentele di primo e secondo grado. Sarà stimato inoltre il numero di marcatori necessari per ottenere un potere 95% e 99% con un rate di falsi positivi inferiore a 5%. Il potere del test sarà verificato utilizzando marcatori altamente polimorfici ed informativi e Single Nucleotide Polymorphism (SNP). Sarà valutato anche l'effetto del Linkage Disequilibrium sul test di parentela simulando un dataset composto da blocchi di SNP in LD.

Capitolo 2

Materiali e Metodi

2.1 Linguaggi di programmazione

I programmi realizzati per produrre e gestire tutti i dati presentati sono sviluppati in *Java* o *perl*. Entrambi i linguaggi sono liberamente scaricabili e multipiattaforma. Lo stesso codice (Java o perl) può essere eseguito su un qualsiasi sistema operativo (ad esempio, MS Windows, GNU/Linux o Mac OS) installando l'interprete di byte-code *Java* (http://java.sun.com/) e l'interprete *perl* (http://www.perl.org/).

2.1.1 Perl

Il *Perl (Practical Extraction and Report Language)* è un linguaggio interpretato che fornisce una sintassi molto simile al linguaggio C. Permette di trattare file di testo facilmente e con poche righe di codice poiché dispone di costrutti per

il "pattern matching" molto sofisticati e di molte funzioni per il trattamento di stringhe e di liste di stringhe [Wall, 2002]. Questo lo rende uno strumento adatto per il parsing dei dati, la creazione di strumenti per il controllo dei risultati, la riscrittura di file di testo in formati diversi.

2.1.2 Java

Il linguaggio *Java* segue il paradigma di programmazione ad oggetti in cui sono centrali i concetti di classe e metodo: una classe può essere vista come una scatola contenente dei dati (ad esempio, una stringa) che possono essere modificati attraverso procedure (i metodi) fornite dalla stessa classe. L'insieme delle classi e dei metodi definisce una *Application Programming Interface (API)*. Uno dei principali vantaggi che derivano dall'utilizzo di Java (e della sua API) è la disponibilità di molto codice già scritto e testato per cui lo sviluppatore può concentrarsi sul proprio specifico problema anziché sul come preparare gli strumenti per risolverlo.

2.2 Programmi aggiuntivi di supporto

2.2.1 *merlin*

merlin (Multipoint Engine for Rapid Likelihood INference) è stato sviluppato per l'analisi di linkage con mappe genetiche ad alta densità [Abecasis et al., 2002]. *merlin* implementa un metodo efficiente per risolvere i comuni problemi di condivisione allelica e ricostruzione degli aplotipi [Abecasis et al., 2002]. *merlin*

è in grado di calcolare l'esatta likelihood per marcatori singoli o per una mappa di marcatori.

Il programma viene usato per simulare la segregazione degli alleli all'interno di un set di famiglie.

2.2.2 *PHASE*

Il programma *PHASE* implementa metodi per la stima degli aplotipi attraverso i genotipi di una popolazione. Il programma implementa i metodi Bayesiani già utilizzati in altri algoritmi e, in aggiunta, considera il decadimento del Linkage Disequilibrium con l'aumentare della distanza tra i marcatori ed il loro ordine lungo il cromosoma [Stephens and Donnelly, 2003].

2.2.3 *Gevalt* e *Gerbil*

Gevalt (GEnotype Visualization and ALgorithmic Tool) è un software nato per semplificare ed automatizzare il lavoro di analisi dei genotipi [Davidovich et al., 2007]. L'interfaccia grafica di *Gevalt* si serve del motore grafico di *Haploview* [Barrett et al., 2005] al fine di usare programmi per ricostruire probabilisticamente la fase di genotipi (*Gerbil*), eseguire lo SNP tagging (STAMPA) e condurre test di permutazione utilizzati nella verifica della significatività dell'associazione.

2.3 Identificazione di parentele

Tradizionalmente, in medicina forense e nei test di consanguineità, le parentele vengono testate su individui la cui posizione all'interno del pedigree è ben definita.

Lo scenario offerto da uno studio di associazione caso-controllo è del tutto diverso in quanto si vuole verificare se nel set di individui non imparentati sono presenti dei legami ignoti al momento del reclutamento.

		Parent-Child	Full sibs	2.nd degree	Non-relatives
A) Genotype combination					
1	**AA, AA**	p_A^3	$p_A^2(1+p_A)^2/4$	$p_A^3(1+p_A)/2$	p_A^4
2	**AA,AB**	$2p_A^2p_B$	$p_A^2p_B(1+p_A)$	$p_A^2p_B(1+2p_A)$	$4p_A^3p_B$
3	**AA,BB**	0	$p_A^2p_B^2/2$	$p_A^2p_B^2$	$2p_A^2p_B^2$
4	**AB,AB**	$p_Ap_B(p_A+p_B)$	$p_Ap_B(2p_Ap_B+p_A+p_B+1)/2$	$p_Ap_B(4p_Ap_B+p_A+p_B)/2$	$4p_A^2p_B^2$
5	**AA,BC**	0	$p_A^2p_Bp_C$	$2p_A^2p_Bp_C$	$4p_A^2p_Bp_C$
6	**AB,AC**	$2p_Ap_Bp_C$	$p_Ap_Bp_C(2p_A+1)$	$p_Ap_Bp_C(4p_A+1)$	$8p_A^2p_Bp_C$
7	**AB,CD**	0	$2p_Ap_Bp_Cp_D$	$4p_Ap_Bp_Cp_D$	$8p_Ap_Bp_Cp_D$

Figura 2.1. Le probabilità condizionate ai sette possibili accoppiamenti tra i genotipi di un marcatore multi-allelico in funzione della frequenza allelica, indicata con p_i [Presciuttini et al., 2002].

Utilizzando le probabilità di Figura 2.1 è possibile testare l'ipotesi di parentela contro l'ipotesi di non parentela (condizionate al genotipo) attraverso una analisi dei *LOD score*. Il *LOD score* è definito come il logaritmo in base 10 della *likelihood ratio*. Le due ipotesi alternative nel calcolo della *likelihood ratio* sono H_0 - la likelihood sotto l'ipotesi di assenza di parentela (ipotesi nulla) - ed H_1 - la likelihood sotto l'ipotesi di un legame di parentela [Terwilliger and Jurg Ott, 1994].

Supponiamo di avere due individui entrambi con genotipo AA e che p_A sia la frequenza dell'allele A. Dalle probabilità espresse in Figura 2.1 la *likelihood ratio* per una coppia di individui aventi genotipo AA sarà [Kaiser and Seber, 1985]:

$$\Lambda = \frac{H_1}{H_0} = \frac{p_A^2}{p_A^4} \tag{2.1}$$

Da cui si ottiene il *LOD score*:

$$LOD = \log_{10}(\Lambda) = \log_{10}(p_A^2) - \log_{10}(p_A^4) \tag{2.2}$$

Il *LOD score* ottenuto è relativo ad un singolo marcatore. Per ottenere un *LOD score* totale è necessario sommare il contributo di *LOD score* di tutti i marcatori disponibili per una coppia di individui [Terwilliger and Jurg Ott, 1994].

Il risultato del test statistico è soggetto a due tipi di errore [Kaiser and Seber, 1985]:

- *Errore di Tipo I o falso positivo o α error*: il test assegna l'ipotesi di parentela ma i due individui non sono parenti.
- *Errore di Tipo II o falso negativo o β error*: il test assegna l'ipotesi di non parentela ma i due individui in realtà sono parenti.

Il rate di falsi positivi (2.3) e di falsi negativi (2.4) sono definiti come:

$$\alpha = \frac{numero\ di\ falsi\ positivi}{numero\ di\ negativi\ reali} \tag{2.3}$$

$$\beta = \frac{numero\ di\ falsi\ negativi}{numero\ di\ positivi\ reali} \tag{2.4}$$

Il valore $1 - \alpha$ indica la *specificità* del test mentre il valore $1 - \beta$ indica il *potere* del test [Kaiser and Seber, 1985].

In Figura 2.2 sono mostrati schematicamente i passi necessari a stimare il numero di marcatori necessari per condurre un test di parentela con un rate di falsi positivi α ed un potere $(100 - \beta)$ prestabiliti.

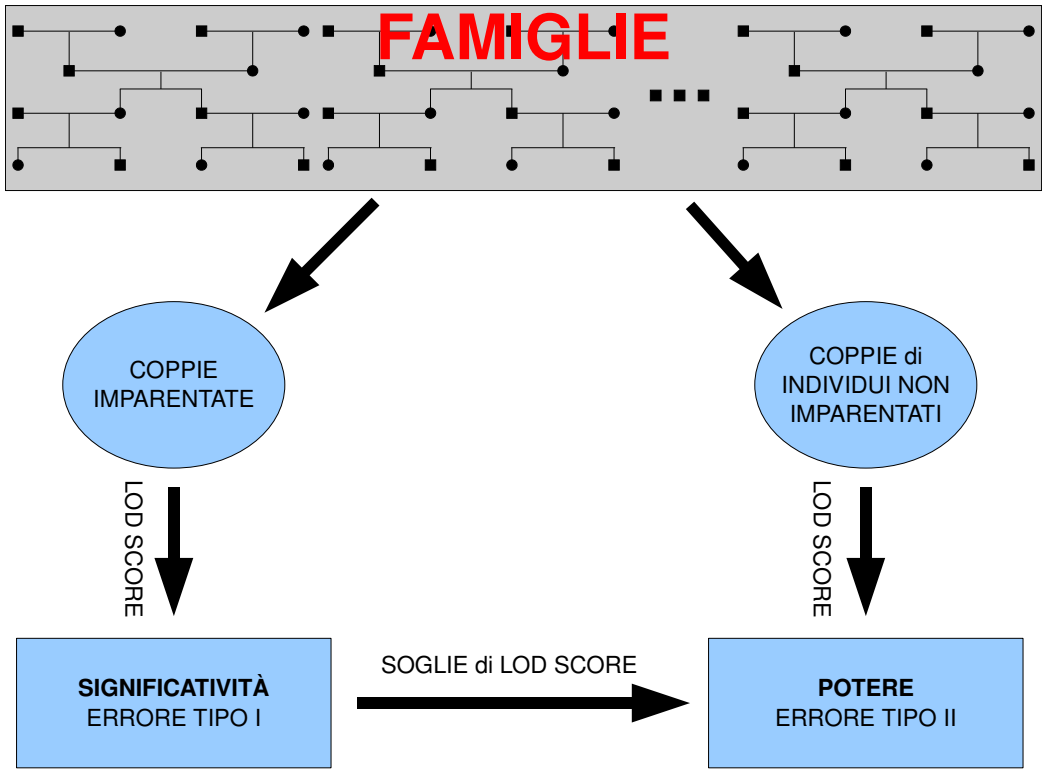


Figura 2.2. Calcolo del potere.

Nel test viene creato un set di famiglie in cui la segregazione degli alleli è simulata *in silico* per procurarsi coppie di individui imparentati e genotipizzati, vengono fissati i valori di α (rate di falsi positivi) e β (rate di falsi negativi) e viene ipotizzato un numero di marcatori N. Dal set di famiglie si selezionano le coppie di individui che sono legate con il grado di parentela d'interesse. Per ogni coppia selezionata si calcola il relativo *LOD score* totale attraverso le coppie di

genotipi degli N marcatori. Dopo aver calcolato il *LOD score* totale per tutte le coppie si possono determinare i valori di: *LOD score* massimo, minimo e medio ed il valore x di *LOD score* che consente di ottenere il rate di falsi negativi β (e potere $1 - \beta$). Dallo stesso gruppo di famiglie viene selezionato un gruppo di individui fondatori (non imparentati tra loro). Per ogni coppia di individui non imparentati si calcola il relativo *LOD score* totale. Se il valore di *LOD score* calcolato è maggiore della soglia x, allora si è in presenza di un falso positivo. Il numero di marcatori N utilizzati per il test è sufficiente quando il rate di falsi positivi è minore della soglia α prefissata.

Capitolo 3

Risultati

3.1 Jenoware: una libreria Java per il trattamento di dati clinici e genetici

Jenoware (http://medgen.univr.it/jenoware) è una libreria sviluppata in *JAVA* per facilitare la realizzazione di programmi in grado di trattare dati clinici e genetici.

Le analisi sui dati spesso tendono a reiterare lo stesso workflow per ogni fattore clinico o genetico disponibile lasciando al ricercatore il solo compito di preparare il set di dati che sarà elaborato dal programma di analisi. Poiché la mole di informazioni è in continua crescita, diventa assai gravoso aggiornare i database e gestire tutte le operazioni che seguono l'immagazzinamento dei dati. Jenoware fornisce una *API* progettata per gestire tutti questi dati. Tramite la *API* di *Jenoware* il ricercatore dispone di procedure (ad esempio, calcolare le frequenze di un marcatore, verificare se un marcatore è o no in equilibrio di Hardy-Weinberg) per sviluppare strumenti in grado di risolvere specifici problemi (ad

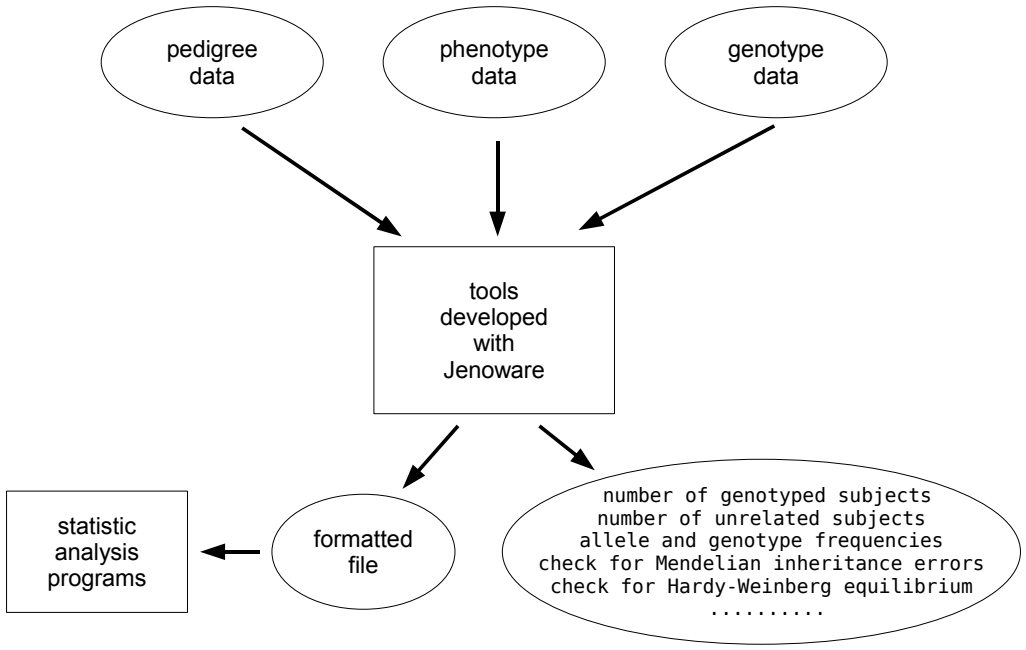


Figura 3.1. Il flusso delle informazioni in Jenoware.

esempio, per mostrare una statistica descrittiva del campione).

In Figura 3.1 è possibile vedere uno dei possibili schemi del flusso di informazioni quando si utilizzano strumenti sviluppati con la API di Jenoware: il ricercatore ha a disposizione un set di individui o famiglie caratterizzati con informazioni cliniche e genetiche che possono essere elaborate al fine di ottenere una statistica descrittiva del campione o formattate per essere poi inoltrate ai programmi di statistica utilizzati in laboratorio.

3.1.1 Simulazione del dataset con *SimulateMerlinFreqFile*

La simulazione di un dataset di famiglie viene condotta in due fasi: in fase 1, utilizzando il programma *SimulateMerlinFreqFile*, generiamo un pedigree file che verrà poi, in fase 2, simulato attraverso il programma *merlin*.

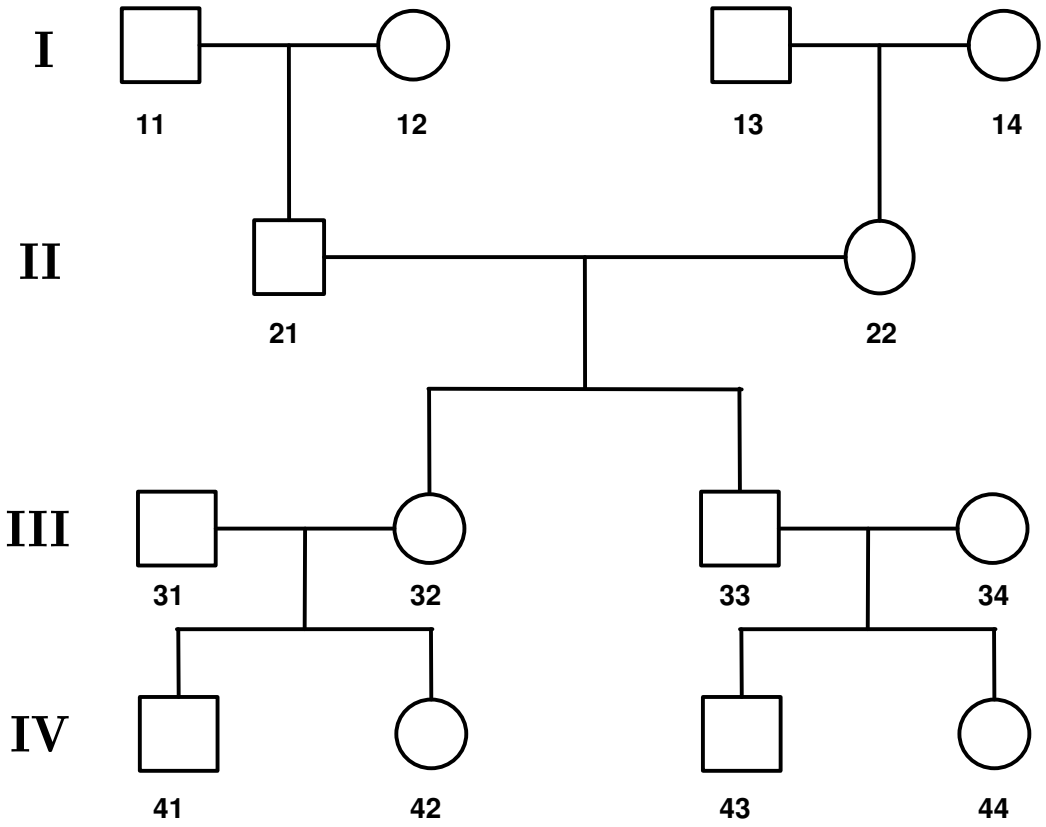


Figura 3.2. Il pedigree utilizzato per la simulazione delle famiglie.

Definizione del pedigree

Il pedigree - Figura 3.2 - utilizzato nelle simulazioni è stato progettato in modo da poter osservare tutti i legami di parentela d'interesse che sono di primo (padre-figlio, coppie di fratelli) e secondo (zii e nonni) grado. Ogni famiglia simulata sarà di 14 individui disposti su 4 generazioni.

In Tabella 3.1 lo stesso pedigree è presentato in un formato matematico nel quale ogni riga descrive un individuo. Questa rappresentazione viene utilizzata all'interno di tutti gli strumenti informatici utilizzati in questa tesi.

family ID	subject ID	father ID	mother ID	sex
1	11	0	0	1
1	12	0	0	2
1	13	0	0	1
1	14	0	0	2
1	21	11	12	1
1	22	13	14	2
1	31	0	0	1
1	32	21	22	2
1	33	21	22	1
1	34	0	0	2
1	41	31	32	1
1	42	31	32	2
1	43	33	34	1
1	44	33	34	2

Tabella 3.1. Rappresentazione matematica del pedigree. Ogni riga corrisponde ad un individuo e contiene le informazioni relative a: codice univoco della famiglia (family ID), codice univoco dell'individuo nella famiglia (subject ID), codice univoco del padre dell'individuo nella famiglia (father ID), codice univoco della madre dell'individuo nella famiglia (mother ID), sesso (1 maschio, 2 femmina). A seconda dei programmi utilizzati la stringa che indica un valore assente è 0 o x. In tabella gli individui fondatori che non hanno genitori hanno i campi father ID e mother ID posti a 0.

SimulateMerlinFreqFile

SimulateMerlinFreqFile fa parte del gruppo di strumenti software sviluppati utilizzando la *API* di *Jenoware*. È lo strumento che permette di creare i file necessari a condurre i diversi test sul potere dei marcatori. Il ricercatore può settare il numero di famiglie da generare (tutte con il pedigree di Figura 3.2), il numero di marcatori genetici per individuo, il numero massimo di alleli per marcatore (il numero di alleli di ogni marcatore avrà un valore random che va da un minimo di due al massimo indicato) e l'eterozigosità minima che si desidera per ogni marcatore.

Il programma produce 4 file:

- un file ped contenente tutte le famiglie. Su ogni riga vengono descritte le informazioni di un singolo individuo che sono quelle riportare in Tabella 3.1 (family ID; subject ID; father ID; mother ID; sex) seguite da un valore per lo stato di malattia (0 dato assente, 1 sano, 2 malato) e da tutti i marcatori (rappresentati come coppie di numeri).
- un file dat con i nomi dei marcatori presenti nel file ped.
- un file $freq$ in cui sono indicati tutti gli alleli con relativa frequenza di tutti i marcatori che appaiono nei file dat e ped.
- un file map che riporta la posizione dei marcatori indicando il cromosoma di appartenenza, il nome del marcatore e la posizione del marcatore sul cromosoma.

I quattro file diventano l'input del programma *merlin*:

```
merlin --simulate --reruns 1 --save -d test.dat
-p test.ped -m test.map -f test.freq
```

L'esecuzione di *merlin* con questi parametri simula (*–simulate*) una sola volta (*–reruns 1*) la segregazione dei marcatori e salva il risultato (*–save*) su 4 nuovi file:

```
merlin-replicate.ped
merlin-replicate.dat
merlin-replicate.freq
merlin-replicate.map
```

SimulateMerlinFreqFile con marcatori in Linkage Disequilibrium

I marcatori presenti nel set di famiglie aventi SNP in linkage disequilibrium formeranno degli aplotipi. Questo permette di simulare la segregazione di blocchi di alleli non indipendenti con il programma *merlin*. Per creare un dataset con aplotipi il programma *SimulateMerlinFreqFile* necessita di ulteriori parametri come il numero di SNP che formano un aplotipo ed il valore di r^2 minimo tra le SNP adiacenti.

Nonostante *SimulateMerlinFreqFile* possa generare blocchi più lunghi, la dimensione massima dei blocchi simulati è fissata a 5 poiché *merlin* può gestire marcatori con massimo 32 alleli. La dimensione 5 perché le combinazioni tra gli alleli di 5 SNP sono $2^5 = 32$.

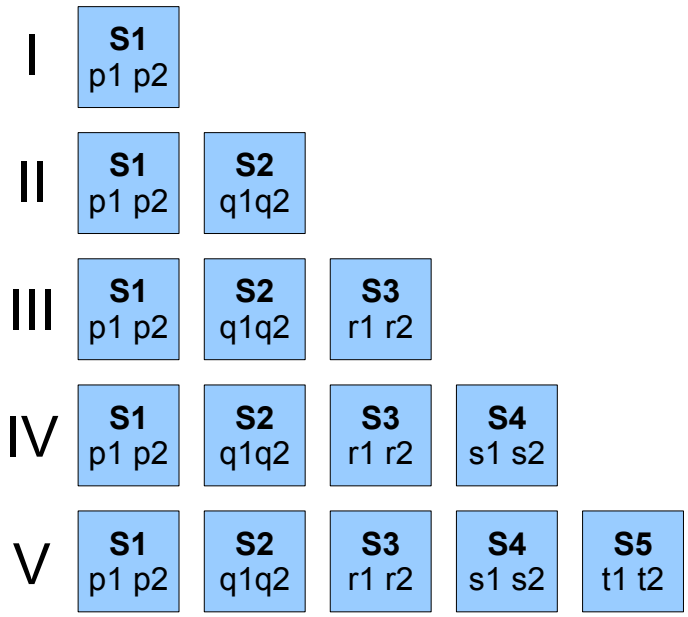


Figura 3.3. Rappresentazione grafica della creazione di un blocco di 5 SNP ($S1$-$S5$) in linkage disequilibrium. I possibili aplotipi $p[12]q[12]r[12]s[12]t[12]$ sono 32.

In Figura 3.3 viene mostrato graficamente l'algoritmo con cui viene formato un blocco di 5 SNP ed in Figura 3.4 viene mostrato il risultato della computazione.

```
$ java -cp jenoware.jar SimulateMerlinFreqFile -f 1 -m 1 -ldh 5 -r2 0.1 -e 0.3

Marker: '0'.
eterozigosit: 0.35420000000000007

p1: 0.23 p2: 0.77 q1: 0.64 q2:0.36 x11: 4/100
D: -0.10719999999999999 D: 0.7282608695652173 R: 0.2816362381579772

p1: 0.64 p2: 0.36 q1: 0.81 q2:0.19 x11: 63/100
D: 0.11159999999999992 D: 0.9177631578947362 R: 0.351242690058479

p1: 0.81 p2: 0.19 q1: 0.56 q2:0.44 x11: 55/100
D: 0.09639999999999999 D: 0.9060150375939847 R: 0.24506130646481517

p1: 0.56 p2: 0.44 q1: 0.59 q2:0.41 x11: 41/100
D: 0.07959999999999995 D: 0.34668989547038304 R: 0.10630398952019443

=== Occorrenze Aplotipi ===
{22221=2, 21221=1, 22222=6, 21121=8, 12111=4, 12112=1,
 12122=1, 21122=10, 21111=29, 12211=1, 21112=12, 12221=2,
 12222=7, 11111=2, 12121=3, 22112=2, 22121=2, 11122=2, 22111=5}
```

Figura 3.4. Un esempio di computazione di *SimulateMerlinFreq*. La prima riga riporta la sintassi del comando eseguito per ottenere l'output mostrato. I parametri dati permettono di creare un dataset in cui esisterà una sola famiglia ($-f$ 1) ed un solo marcatore ($-m$ 1) costituito da 5 SNP in linkage disequilibrium ($-ldh$ 5) con r^2 maggiore di 0,1 ($-r2$ 0,1). La prima SNP deve avere eterozigosità maggiore di 0,3 ($-e$ 0,3). L'output del programma mostra, per ogni coppia di SNP, le frequenze alleliche $p1$, $p2$, $q1$, $q2$, la frequenza del primo aplotipo $x11$ ed i valori di di D, D^1 e r^2. Le ultime righe dell'output riportano le occorrenze di ogni aplotipo generato.

Il primo passo consiste nel generare in modo random le frequenze degli alleli della SNP $S1$ (Figura 3.3) rispettando l'eterozigosità minima richiesta. Le successive SNP $S2$, $S3$, $S4$, $S5$ saranno poi generate in modo da soddisfare il parametro minimo di r^2 richiesto. Le frequenze alleliche $p1$ e $p2$ di $S1$ a questo punto sono note.

Per ottenere le frequenze alleliche $q1$ e $q2$ di $S2$ (Figura 3.3) ed avere un r^2 maggiore di 0.1 vengono generati randomicamente i valori di $q1$ ($q2 = 1 - q1$) e $x11$ (frequenza dell'aplotipo $p1q1$) e quindi viene calcolato r^2. Se r^2 soddisfa la condizione richiesta passiamo a calcolare le frequenze alleliche $r1$ e $r2$ di $S3$

(Figura 3.3) allo stesso modo (avendo $q1$ e $q2$ noti) altrimenti generiamo due nuovi valori per $q1$ e $x11$. Quando le frequenze di $p1$, $p2$, $q1$, $q2$ e $x11$ sono note si hanno tutti i dati per calcolare le frequenze dei tre aplotipi mancanti (con due SNP ci sono 4 possibili aplotipi) $x12$ ($p1q2$), $x21$ ($p2q1$) e $x22$ ($p2q2$) tramite le equazioni 1.5 (Pagina 7).

Una volta generate le frequenze di tutti gli alleli e di tutti gli aplotipi il programma può settare tutti gli aplotipi più relativa occorrenza per le 5 SNP $S_1S_2S_3S_4S_5$ come mostrato in Figura 3.4.

ped2gerbil

ped2gerbil è un programma di utilità scritto in perl che viene utilizzato per generare i file necessari ai successivi test del potere a partire da un set di file simulati (ped, dat, $freq$ e map) con aplotipi.

ped2gerbil crea tutti i file necessari ai programmi *PHASE* e *Gerbil* per ricostruire probabilisticamente la fase degli aplotipi su un set di individui non imparentati che viene selezionato dai file di partenza.

Per verificare l'informatività dei marcatori in presenza di linkage disequilibrium *ped2gerbil* genera un set di file ped, dat, $freq$ e map in cui gli aplotipi sono scomposti. Il risultato è un set di file in cui tutti i marcatori sono SNP.

Per ogni marcatore presente nei file di partenza il programma genera i file di input del programma *Gevalt* per poter così mostrare graficamente il livello di r^2 nei blocchi di SNP in linkage disequilibrium.

3.1.2 Il calcolo dei *LOD score* con *IsRelated*

Il dataset simulato viene processato per rilevare i legami di parentela d'interesse, i genotipi degli individui e selezionare un gruppo di individui non imparentati su cui testare il potere del test.

Per calcolare il *LOD score* di una coppia di genotipi è stata sviluppata una classe *SbjRelationship* nella libreria *Jenoware*. La classe fornisce le funzioni necessarie per calcolare la probabilità di parentela o non parentela condizionate ai genotipi seguendo le formule di Figura 2.1 (Pagina 13).

IsRelated

Per calcolare i contributi di *LOD score* è stato sviluppato in Java - utilizzando la libreria *Jenoware* - il programma *IsRelated*.

Il programma carica i dati degli individui e dei marcatori dai file ped, dat e $freq$ simulati da *merlin*.

Nel caso di simulazioni con linkage disequilibrium il programma deve caricare tutti i file out_pairs dove sono memorizzati gli individui non imparentati per i quali sono stati ricostruiti probabilisticamente gli aplotipi. Il programma riconosce i file prodotti dal programma *PHASE*, mentre per i file di output di *Gerbil* è stato sviluppato in perl un programma - *gerbil2out_pairs* - per convertire i risultati di *Gerbil* in file out_pairs di *PHASE*.

Se il programma non ha in input dei file out_pairs, viene selezionato un individuo fondatore da ogni famiglia per avere il set di individui non imparentati su cui calcolare il rate di falsi positivi.

IsRelated non calcola il *LOD score* totale, ma solo i singoli contributi di *LOD score* di ogni marcatore. Il risultato dell'elaborazione è un file res in cui ogni riga contiene: identificatore del primo individuo, identificatore del secondo individuo, un codice per identificare la parentela testata, un codice per indicare la parentela osservata, la lista di tutti i contributi di *LOD score* (sono tanti quanti i marcatori presenti nel file ped) ed il numero totale di contributi di *LOD score* presenti sulla riga. Questa computazione intermedia consente al ricercatore di condurre lo studio del potere in modo più efficiente poiché non occorre ripetere ad ogni esperimento il calcolo dei contributi di *LOD score*.

3.1.3 Stima del numero di marcatori necessari con *ProcessIsRelated*

ProcessIsRelated è scritto in Java utilizzando la libreria *Jenoware*. Viene utilizzato per sommare tutti i contributi di *LOD score* per ottenere *LOD score* totale e mostrare una statistica descrittiva del test. Il programma ha due possibili parametri: il numero di marcatori N da utilizzare per il calcolo del *LOD score* totale (se omesso sono utilizzati tutti i marcatori disponibili) ed il nome del file res prodotto da *IsRelated* da cui estrarre i contributi di *LOD score*. *ProcessIsRelated* segue lo schema di Figura 2.2 per calcolare il rate di falsi positivi ed il rate di falsi negativi con N marcatori.

Al termine dell'elaborazione il programma salva su file il risultato della computazione (Figura 3.5). Il numero di marcatori utilizzato in ogni test è scritto nel nome del file; in Figura 3.5 il nome del file è nella prima riga (*merlin-replicate-10.pro* per cui il numero di marcatori è 10). Per ogni tipo di parentela sono riportati il numero totale di test effettuati (8000 confronti tra coppie padre

```
$ cat merlin-replicate-10.pro
RISULTATI:

PADRE FIGLIO:
total : 8000
LOD max   : 7.645404315970895
LOD min   : 1.0198489251107201
media LOD : 3.741346276872023
unrelated con LOD > 1.7697853224255735 [1%]    : 418 [0.08351648351648353%]
unrelated con LOD > 2.012636501096001 [2.5%]  : 344 [0.06873126873126872%]
unrelated con LOD > 2.2475502343566296 [5%]    : 267 [0.053346653346653346%]
unrelated con LOD > 2.5357770102447943 [10%]   : 175 [0.03496503496503497%]
unrelated con LOD > 2.726937705344465 [15%]   : 134 [0.02677322677322675%]
unrelated con LOD > 2.8965764604462216 [20%]   : 104 [0.02077922077922078%]
unrelated con LOD > 3.741346276872023 [media] : 10 [0.001998001998001998%]

FRATELLI:
total : 3000
LOD max   : 10.010907132081607
LOD min   : -2.7495171476246294
media LOD : 3.0876499499929015
unrelated con LOD > -0.9551493087369951 [1%]    : 60290 [12.045954045954046%]
unrelated con LOD > -0.3216415104267502 [2.5%]  : 26331 [5.2609390609390605%]
unrelated con LOD > 0.1566001329150064 [5%]    : 12971 [2.591608391608392%]
unrelated con LOD > 0.7929124257990042 [10%]   : 4506 [0.9002997002997002%]
unrelated con LOD > 1.2613956005051674 [15%]   : 1914 [0.38241758241758245%]
unrelated con LOD > 1.564070289931044 [20%]   : 1073 [0.2143856143856144%]
unrelated con LOD > 3.0876499499929015 [media] : 37 [0.007392607392607393%]

SECONDO GRADO:
total : 12000
LOD max   : 4.5379582128093405
LOD min   : -2.3833187738983823
media LOD : 0.8516758518516792
unrelated con LOD > -1.1337949993888121 [1%]    : 333324 [66.5982017982018%]
unrelated con LOD > -0.8684900562537311 [2.5%]  : 266807 [53.3080919080919%]
unrelated con LOD > -0.6054133891020068 [5%]    : 200215 [40.002997002997006%]
unrelated con LOD > -0.2919044451789904 [10%]   : 129951 [25.964235764235767%]
unrelated con LOD > -0.0851853187311275 [15%]   : 92666 [18.514685314685313%]
unrelated con LOD > 0.0777460413217741 [20%]   : 68675 [13.72127872127872%]
unrelated con LOD > 0.8516758518516792 [media] : 11038 [2.205394605394605%]

UNRELATED:
total p-f: 500500
total bro: 500500
total s-d: 500500
LOD max p: 4.909656422439044
LOD max b: 4.939159924700488
LOD max s: 3.555270947143343
```

Figura 3.5. Un esempio di file contenente i risultati prodotti dal programma *ProcessIsRelated*.

figlio, 3000 tra fratelli, 12000 tra parenti di secondo grado e 500500 individui non imparentati), i *LOD score* massimo e minimo osservati, la media dei *LOD score* ed il rate di falsi negativi per ogni soglia di potere testata.

Ad esempio, la riga di Figura 3.5:

```
unrelated con LOD > 1.7697853224255735 [1%]
: 418 [0.08351648351648353%]
```

indica che il numero di individui non imparentati con *LOD score*> 1.7697853224255735 (la soglia che corrisponde ad un rate di falsi negativi $\beta = 1\%$ ovvero un potere del $100 - \beta = 99\%$) è 418. Con 10 marcatori il rate di falsi positivi è $\alpha = 0.08351648351648353\%$.

3.1.4 Automatizzazione dei test con *doSimulate* e *doSimulate-ld*

Per eseguire in modo automatico le simulazioni ed il calcolo del potere sono stati realizzati due programmi perl: *doSimulate* e *doSimulate-ld*. I programmi richiamano in modo automatico tutti gli strumenti sviluppati in modo da eseguire tutti i passi delle simulazioni e dei test.

Il programma *doSimulate* prende in input il numero di famiglie, il numero di marcatori voluto per ogni individuo, il numero di alleli massimo per ogni marcatore e l'eterozigosità minima di ogni marcatore. Viene utilizzato per simulare i set di famiglie con marcatori multiallelici ad alta eterozigosità e per testare l'informatività delle SNP in assenza di linkage disequilibrium.

Il programma *doSimulate-ld* prende in input il numero di famiglie, il numero di marcatori (in questo caso sono aplotipi), il numero di SNP utilizzate su ogni

aplotipo, l'eterozigosità minima delle SNP ed il valore di r^2 minimo tra due SNP adiacenti.

Test con *doSimulate*

In Figura 3.6 sono mostrati tutti i comandi eseguiti per verificare il potere di marcatori altamente polimorfici nel test di parentela. Sono utilizzate 1000 famiglie e 30 marcatori per ogni individuo aventi una eterozigosità di almeno 0.7 ed un numero massimo di alleli pari a 10.

```
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m SimulateMerlinFreqFile -f 1000 -m 30 -a 10 -e 0.70
/home/ciano/geno_exe/merlin --simulate --reruns 1 --save -d test_file.dat -p test_file.ped -m test_file.map -f test_file.freq
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m IsRelated merlin-replicate.ped > /dev/null 2> /dev/null
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 30 merlin-replicate.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 25 merlin-replicate.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 20 merlin-replicate.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 15 merlin-replicate.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 10 merlin-replicate.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 5 merlin-replicate.res
```

Figura 3.6. Comandi eseguiti in automatico dal programma *doSimulate* per verificare il potere di marcatori altamente polimorfici nel test di parentela.

Dopo aver simulato il dataset con *SimulateMerlinFreqFile* e *merlin* vengono calcolati tutti i contributi di *LOD score* con il programma *IsRelated*. Il potere del test viene valutato eseguendo il programma *ProcessIsRelated* ed utilizzando ogni volta un numero di marcatori minore: si parte considerando tutti i 30 marcatori per via via arrivare a 5 marcatori con step di 5.

In Figura 3.8 sono descritti i comandi eseguiti per condurre il test utilizzando 1000 famiglie e 250 SNP per ogni individuo aventi una eterozigosità di almeno 0.15. Il numero massimo di alleli è 2.

I passi della computazione sono del tutto simili a quelli visti in Figura 3.6 ma il potere del test viene valutato partendo dal totale delle 250 SNP ed arrivando a 25 con step di 25.

```
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m SimulateMerlinFreqFile -f 1000 -m 250 -a 2 -e 0.15
/home/ciano/geno_exe/merlin --simulate --reruns 1 --save -d test_file.dat -p test_file.ped -m test_file.map -f test_file.freq
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m IsRelated merlin-replicate.ped > /dev/null 2> /dev/null
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 250 merlin-replicate.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 225 merlin-replicate.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 200 merlin-replicate.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 175 merlin-replicate.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 150 merlin-replicate.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 125 merlin-replicate.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 100 merlin-replicate.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 75 merlin-replicate.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 50 merlin-replicate.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 25 merlin-replicate.res
```

Figura 3.7. Comandi eseguiti in automatico dal programma *doSimulate* per verificare il potere di 250 SNP nel test di parentela.

Test con *doSimulate-ld*

In Figura **??** sono descritti i passi della computazione eseguita per testare l'informatività di SNP in linkage disequilibrium tra loro. Ogni test ha utilizzato 1000 famiglie e gli individui sono caratterizzati con 60 aplotipi di 5 SNP. La prima SNP di ogni aplotipo ha una eterozigosità di almeno 0.15. Per verificare l'informatività delle SNP si sono utilizzate 4 soglie di r^2 minimo nel test: 0.01, 0.1, 0.4, 0.8. I passi della computazione sono gli stessi in tutti i casi a meno del valore minimo di r^2 richiesto.

Per eseguire tutti i test necessari, il dataset - simulato con *SimulateMerlinFreqFile* e *merlin* - viene processato con il programma *ped2gerbil*: vengono così creati i file che saranno passati come input di *PHASE*, *Gerbil* e *Gevalt*. Su ogni gruppo di 5 SNP vengono ricostruiti probabilisticamente gli aplotipi con *PHASE* e quindi vengono calcolati i contributi di *LOD score* con *IsRelated*. Il programma ha come input il file ped e tutti i file out_pairs prodotti da *PHASE*. Il risultato della computazione è un file res che viene valutato con il programma *ProcessIsRelated* usando un numero di aplotipi che parte dal totale di 60 per arrivare a 10 con step di 10.

In modo del tutto simile valutiamo il potere del test ricostruendo la fase degli

```
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m SimulateMerlinFreqFile -f 1000 -m 60 -ldh 5 -r2 0.1 -e 0.15
/home/ciano/geno_exe/merlin --simulate --reruns 1 --save -d test_file.dat -p test_file.ped -m test_file.map -f test_file.freq
/home/ciano/geno_exe/ped2gerbil.pl merlin-replicate.ped > /dev/null
/home/ciano/geno_exe/PHASE merlin-replicate_phase00.inp merlin-replicate_phase00.inp.out
/home/ciano/geno_exe/PHASE merlin-replicate_phase00.inp merlin-replicate_phase01.inp.out
.....
/home/ciano/geno_exe/PHASE merlin-replicate_phase58.inp merlin-replicate_phase58.inp.out
/home/ciano/geno_exe/PHASE merlin-replicate_phase59.inp merlin-replicate_phase59.inp.out
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m IsRelated merlin-replicate.ped test_file0.out_pairs \
        test_file1.out_pairs ...  test_file58.out_pairs test_file59.out_pairs > /dev/null 2> /dev/null
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 60 merlin-replicate.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 50 merlin-replicate.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 40 merlin-replicate.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 30 merlin-replicate.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 20 merlin-replicate.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 10 merlin-replicate.res
/home/ciano/geno_exe/Gevalt/gerbil.exe merlin-replicate_gerbil01.txt
/home/ciano/geno_exe/Gevalt/gerbil.exe merlin-replicate_gerbil02.txt
.....
/home/ciano/geno_exe/Gevalt/gerbil.exe merlin-replicate_gerbil58.txt
/home/ciano/geno_exe/Gevalt/gerbil.exe merlin-replicate_gerbil59.txt
/home/ciano/geno_exe/gerbil2out_pairs.pl
/bin/ln -s merlin-replicate.ped merlin-replicate-gerbil.ped
/bin/ln -s merlin-replicate.dat merlin-replicate-gerbil.dat
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m IsRelated merlin-replicate-gerbil.ped test_file0.out_pairs \
        test_file1.out_pairs ... test_file58.out_pairs test_file59.out_pairs > /dev/null 2> /dev/null
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 60 merlin-replicate-gerbil.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 50 merlin-replicate-gerbil.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 40 merlin-replicate-gerbil.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 30 merlin-replicate-gerbil.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 20 merlin-replicate-gerbil.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 10 merlin-replicate-gerbil.res
/bin/ln -s merlin-replicate.ped merlin-replicate-noinferred.ped
/bin/ln -s merlin-replicate.dat merlin-replicate-noinferred.dat
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m IsRelated merlin-replicate-noinferred.ped > /dev/null 2> /dev/null
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 60 merlin-replicate-noinferred.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 50 merlin-replicate-noinferred.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 40 merlin-replicate-noinferred.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 30 merlin-replicate-noinferred.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 20 merlin-replicate-noinferred.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 10 merlin-replicate-noinferred.res
/home/ciano/geno_exe/pedHaplo2snp.pl merlin-replicate.ped
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m IsRelated merlin-replicate_snp.ped > /dev/null 2> /dev/null
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 300 merlin-replicate_snp.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 275 merlin-replicate_snp.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 250 merlin-replicate_snp.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 225 merlin-replicate_snp.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 200 merlin-replicate_snp.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 175 merlin-replicate_snp.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 150 merlin-replicate_snp.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 125 merlin-replicate_snp.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 100 merlin-replicate_snp.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 75 merlin-replicate_snp.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 50 merlin-replicate_snp.res
/usr/bin/java -cp /home/ciano/geno_exe/jenoware.jar -Xmx1500m ProcessIsRelated -m 25 merlin-replicate_snp.res
/bin/ln -s /home/ciano/geno_exe/Gevalt/gerbil.exe
/usr/bin/java -Xmx1500m -jar /home/ciano/geno_exe/Gevalt/Gevalt.jar -png BOTH -ldcolorscheme RSQ -nogui -pedfile \
        merlin-replicate_separate00.ped -info merlin-replicate_separate00.dat
/usr/bin/java -Xmx1500m -jar /home/ciano/geno_exe/Gevalt/Gevalt.jar -png BOTH -ldcolorscheme RSQ -nogui -pedfile \
        merlin-replicate_separate01.ped -info merlin-replicate_separate01.dat
.....
/usr/bin/java -Xmx1500m -jar /home/ciano/geno_exe/Gevalt/Gevalt.jar -png BOTH -ldcolorscheme RSQ -nogui -pedfile \
        merlin-replicate_separate58.ped -info merlin-replicate_separate58.dat
/usr/bin/java -Xmx1500m -jar /home/ciano/geno_exe/Gevalt/Gevalt.jar -png BOTH -ldcolorscheme RSQ -nogui -pedfile \
        merlin-replicate_separate59.ped -info merlin-replicate_separate59.dat
```

Figura 3.8. Comandi eseguiti in automatico dal programma *doSimulate-ld* per verificare il potere di 60 marcatori composti da 5 SNP nel test di parentela. I puntini indicano l'assenza di righe uguali alle precedenti e successive a meno degli indici numerici dei file.

aplotipi con il programma *Gerbil*. Il risultato di *Gerbil* va convertito con il programma *gerbil2out_pairs* per poter essere processato poi da *IsRelated*. Il file res risultante viene valutato con il programma *ProcessIsRelated* partendo da 60 aplotipi ed arrivando a 10 con step di 10.

Per valutare l'effetto che la ricostruzione probabilistica degli aplotipi ha sui nostri dati viene eseguito il test del potere anche sul dataset di partenza in cui non vengono ricostruiti probabilisticamente gli aplotipi. Il programma *IsRelated* ha come input il solo file ped. Il potere del test viene poi valutato con il programma *ProcessIsRelated* usando un numero di aplotipi che parte dal totale di 60 per arrivare a 10 con step di 10.

L'ultimo test condotto in presenza di linkage disequilibrium è quello in cui tutte le SNP vengono considerate marcatori separati. Il potere del test viene valutato partendo da 300 SNP ed arrivando a 25 con step di 25.

Infine per apprezzare graficamente il livello di r^2 in ognuno dei 60 blocchi di linkage disequilibrium simulati utilizziamo il programma *Gevalt* per generare le mappe di linkage disequilibrium. *Gevalt* viene solitamente utilizzato attraverso la sua interfaccia grafica, ma in questo caso si è sfruttata la possibilità del funzionamento tramite riga di comando per automatizzare la creazione delle immagini.

3.2 Calcolo del Potere

Il programma *ProcessIsRelated* produce, per ogni test effettuato, un file di testo con i risultati della computazione (Figura 3.5).

Il numero totale di file di testo prodotti con *ProcessIsRelated* nei vari test effettuati è 146. Per ogni dataset simulato abbiamo un gruppo di file contenenti risultati diversi. Per ogni gruppo di file verranno mostrati solamente i risultati di interesse per lo scopo del lavoro: il legame di parentela; il potere (80%, 95% o 99%); il numero di marcatori utilizzati; il valore soglia di *LOD score* usato per determinare il rate di falsi negativi e di falsi positivi; il rate di falsi positivi α. Gli individui non imparentati con *LOD score* maggiore del valore soglia indicato nelle tabelle sono falsi positivi mentre gli individui con il grado di parentela aventi *LOD score* minore del valore soglia indicato nelle tabelle sono falsi negativi. Qualora il numero di marcatori non fosse sufficiente a vedere un rate di falsi positivi al di sotto della soglia $\alpha \leq 5\%$, verrà riportato il migliore osservato. Per estrarre dai file dei risultati i valori di interesse è stato creato il programma perl *createTableResults*.

Per tutti i test sono stati utilizzati dataset simulati di 1000 famiglie per un totale di 14000 individui. Per ogni test vengono calcolati i *LOD score* totali di 8000 coppie padre figlio, 3000 fratelli, 12000 parenti di secondo grado e 500500 individui non imparentati.

Di seguito verranno riportate le tabelle contenenti i risultati d'interesse per tutti i test effettuati.

3.2.1 Marcatori altamente polimorfici

Legame	$100-\beta$	Numero marcatori	*LOD score*	α
Padre Figlio	80%	5	1.3274	1.05%
	95%	5	0.9132	2.08%
	99%	5	0.6205	2.66%
Fratelli	80%	5	0.4735	3.74%
	95%	10	0.1566	2.59%
	99%	15	-0.355	2.04%
2^0 grado	80%	20	0.6122	2.75%
	95%	30	-0.080	4.89%
	99%	30	-1.131	18.05%

Tabella 3.2. Sono riportati il legame di parentela, il potere $100-\beta$, il numero di marcatori utilizzati, il *LOD score* ed il rate di falsi positivi α. Parametri della simulazione: 1000 famiglie; 30 marcatori; 10 alleli massimo per marcatore; eterozigosità minima 0.70. Il test è condotto aumentando via via il numero di marcatori da 5 a 30 facendo salti di 5.

3.2.2 Marcatori biallelici indipendenti

Legame	$100-\beta$	Numero marcatori	*LOD score*	α
Padre Figlio	80%	25	0.8924	2.24%
	95%	50	1.4862	0.20%
	99%	50	0.6896	0.50%
Fratelli	80%	25	0.4976	4.09%
	95%	50	0.3414	1.87%
	99%	75	-0.171	1.32%
2^0 grado	80%	100	0.4676	4.60%
	95%	200	0.1219	3.20%
	99%	250	-0.769	6.11%

Tabella 3.3. Sono riportati il legame di parentela, il potere $100-\beta$, il numero di marcatori utilizzati, il *LOD score* ed il rate di falsi positivi α. Parametri della simulazione: 1000 famiglie; 250 SNP; 2 alleli; eterozigosità minima 0.15. Il test è condotto aumentando via via il numero di SNP da 25 a 250 facendo salti di 25.

3.2.3 Marcatori biallelici in Linkage Disequilibrium

Per verificare l'effetto del linkage disequilibrium nel test di parentela sono stati simulati quattro dataset con diverse soglie di r^2 minimo: 0.01, 0.1, 0.4 e 0.8. Gli individui sono caratterizzati genotipicamente con 60 aplotipi di 5 SNP.

Per ogni soglia di r^2 vengono presentate 4 tabelle relative ai 4 esperimenti compiuti per ogni dataset avente SNP in linkage disequilibrium. Ognuna delle 4 tabelle riporta il legame di parentela, il potere $100 - \beta$, il numero di marcatori utilizzati, il *LOD score* ed il rate di falsi positivi α.

Una tabella riporta i risultati ottenuti utilizzando solo il dataset contenente gli aplotipi. Due tabelle riportano i risultati ottenuti attraverso il dataset contenete gli aplotipi ed i file out_pairs ottenuti con *PHASE* o *Gerbil*. L'ultima tabella mostra i risultati ottenuti con un dataset in cui gli aplotipi sono convertiti in singole SNP per un totale di 300 SNP (60×5 SNP).

Dataset simulato con $r^2 > 0.01$

I parametri utilizzati per la simulazione del dataset sono: 1000 famiglie; 60 aplotipi di 5 SNP; valore minimo di $r^2 = 0.01$; eterozigosità minima 0.15.

La Tabella 3.4 riporta i risultati ottenuti utilizzando solo il dataset simulato.

Legame	$100 - \beta$	Numero marcatori	*LOD score*	α
Padre Figlio	80%	10	4.2239	5×10^{-4}%
	95%	10	3.2632	0.003%
	99%	10	2.5464	0.008%
Fratelli	80%	10	2.5528	0.01%
	95%	10	0.7780	0.48%
	99%	10	-0.747	4.87%
2^0 grado	80%	20	1.3422	0.37%
	95%	20	0.0364	3.84%
	99%	30	-0.766	4.94%

Tabella 3.4. Sono riportati il legame di parentela, il potere $100 - \beta$, il numero di marcatori utilizzati, il *LOD score* ed il rate di falsi positivi α. Il test è condotto aumentando via via il numero di aplotipi da 10 a 60 facendo salti di 10.

La Tabella 3.5 riporta i risultati ottenuti utilizzando il dataset simulato e gli individui non imparentati in cui le fasi degli aplotipi sono ricostruite con *PHASE*.

Legame	$100 - \beta$	Numero marcatori	*LOD score*	α
Padre Figlio	80%	10	4.2239	0%
	95%	10	3.2632	0.005%
	99%	10	2.5464	0.02%
Fratelli	80%	10	2.5528	0.02%
	95%	10	0.7780	0.75%
	99%	20	1.6424	0.02%
2^0 grado	80%	20	1.3422	0.55%
	95%	30	0.5579	1.65%
	99%	40	-0.496	4.23%

Tabella 3.5. Sono riportati il legame di parentela, il potere $100 - \beta$, il numero di marcatori utilizzati, il *LOD score* ed il rate di falsi positivi α. Il test è condotto aumentando via via il numero di aplotipi da 10 a 60 facendo salti di 10.

La Tabella 3.6 riporta i risultati ottenuti utilizzando il dataset simulato e gli individui non imparentati in cui le fasi degli aplotipi sono ricostruite con *Gerbil*.

Legame	$100 - \beta$	Numero marcatori	*LOD score*	α
Padre Figlio	80%	10	4.2239	6×10^{-4}%
	95%	10	3.2632	0.005%
	99%	10	2.5464	0.01%
Fratelli	80%	10	2.5528	0.03%
	95%	10	0.7780	0.62%
	99%	10	-0.747	4.92%
2^0 grado	80%	20	1.3422	0.28%
	95%	20	0.0364	2.37%
	99%	30	-0.766	1.79%

Tabella 3.6. Sono riportati il legame di parentela, il potere $100 - \beta$, il numero di marcatori utilizzati, il *LOD score* ed il rate di falsi positivi α. Il test è condotto aumentando via via il numero di aplotipi da 10 a 60 facendo salti di 10.

La Tabella 3.7 riporta i risultati ottenuti utilizzando il dataset in cui gli aplotipi vengono convertiti in singole SNP per un totale di 300 SNP (60×5 SNP).

Legame	$100 - \beta$	Numero marcatori	*LOD score*	α
Padre Figlio	80%	25	0.8665	4.53%
	95%	50	1.1328	1.02%
	99%	50	0.3358	1.80%
Fratelli	80%	50	1.0647	1.74%
	95%	75	-0.158	3.52%
	99%	100	-1.508	4.57%
2^0 grado	80%	150	0.6172	4.96%
	95%	300	-0.155	3.78%
	99%	300	-2.118	17.57%

Tabella 3.7. Sono riportati il legame di parentela, il potere $100 - \beta$, il numero di marcatori utilizzati, il *LOD score* ed il rate di falsi positivi α. Il test è condotto aumentando via via il numero di aplotipi da 25 a 300 facendo salti di 25.

Dataset simulato con $r^2 > 0.1$

I parametri utilizzati per la simulazione del dataset sono: 1000 famiglie; 60 aplotipi di 5 SNP; valore minimo di $r^2 = 0.1$; eterozigosità minima 0.15.

La Tabella 3.8 riporta i risultati ottenuti utilizzando solo il dataset simulato.

Legame	$100 - \beta$	Numero marcatori	*LOD score*	α
Padre Figlio	80%	10	3.2592	0.006%
	95%	10	2.3302	0.03%
	99%	10	1.6064	0.07%
Fratelli	80%	10	1.8802	0.08%
	95%	10	0.2479	1.74%
	99%	20	0.7552	0.11%
2^0 grado	80%	20	0.9137	1.14%
	95%	30	0.1365	2.42%
	99%	40	-0.847	4.42%

Tabella 3.8. Sono riportati il legame di parentela, il potere $100 - \beta$, il numero di marcatori utilizzati, il *LOD score* ed il rate di falsi positivi α. Il test è condotto aumentando via via il numero di aplotipi da 10 a 60 facendo salti di 10.

La Tabella 3.9 riporta i risultati ottenuti utilizzando il dataset simulato e gli individui non imparentati in cui le fasi degli aplotipi sono ricostruite con *PHASE*.

Legame	$100 - \beta$	Numero marcatori	*LOD score*	α
Padre Figlio	80%	10	3.2592	0.005%
	95%	10	2.3302	0.04%
	99%	10	1.6064	0.16%
Fratelli	80%	10	1.8802	0.09%
	95%	10	0.2479	2.23%
	99%	20	0.7552	0.18%
2^0 grado	80%	20	0.9137	1.46%
	95%	30	0.1365	3.76%
	99%	50	-0.532	2.95%

Tabella 3.9. Sono riportati il legame di parentela, il potere $100 - \beta$, il numero di marcatori utilizzati, il *LOD score* ed il rate di falsi positivi α. Il test è condotto aumentando via via il numero di aplotipi da 10 a 60 facendo salti di 10.

La Tabella 3.10 riporta i risultati ottenuti utilizzando il dataset simulato e gli individui non imparentati in cui le fasi degli aplotipi sono ricostruite con *Gerbil*.

Legame	$100 - \beta$	Numero marcatori	*LOD score*	α
Padre Figlio	80%	10	3.2592	0.003%
	95%	10	2.3302	0.03%
	99%	10	1.6064	0.08%
Fratelli	80%	10	1.8802	0.07%
	95%	10	0.2479	1.42%
	99%	20	0.7552	0.07%
2^0 grado	80%	20	0.9137	0.53%
	95%	20	-0.243	3.90%
	99%	30	-1.034	3.52%

Tabella 3.10. Sono riportati il legame di parentela, il potere $100 - \beta$, il numero di marcatori utilizzati, il *LOD score* ed il rate di falsi positivi α. Il test è condotto aumentando via via il numero di aplotipi da 10 a 60 facendo salti di 10.

La Tabella 3.11 riporta i risultati ottenuti utilizzando il dataset in cui gli aplotipi vengono convertiti in singole SNP per un totale di 300 SNP (60×5 SNP).

Legame	$100 - \beta$	Numero marcatori	*LOD score*	α
Padre Figlio	80%	50	1.9195	0.85%
	95%	50	0.6776	2.29%
	99%	50	-0.319	3.46%
Fratelli	80%	50	0.7646	4.04%
	95%	100	0.1667	2.15%
	99%	150	-1.531	2.29%
2^0 grado	80%	200	0.8080	4.15%
	95%	300	-0.565	7.76%
	99%	300	-2.720	28.07%

Tabella 3.11. Sono riportati il legame di parentela, il potere $100 - \beta$, il numero di marcatori utilizzati, il *LOD score* ed il rate di falsi positivi α. Il test è condotto aumentando via via il numero di aplotipi da 25 a 300 facendo salti di 25.

Dataset simulato con $r^2 > 0.4$

I parametri utilizzati per la simulazione del dataset sono: 1000 famiglie; 60 aplotipi di 5 SNP; valore minimo di $r^2 = 0.4$; eterozigosità minima 0.15.

La Tabella 3.12 riporta i risultati ottenuti utilizzando solo il dataset simulato.

Legame	$100 - \beta$	Numero marcatori	*LOD score*	α
Padre Figlio	80%	10	2.3494	0.05%
	95%	10	1.3511	0.26%
	99%	10	0.5654	0.66%
Fratelli	80%	10	1.3751	0.32%
	95%	10	-0.031	3.95%
	99%	20	0.0607	0.66%
2^0 grado	80%	20	0.5859	2.75%
	95%	40	0.1534	2.17%
	99%	60	-0.695	2.41%

Tabella 3.12. Sono riportati il legame di parentela, il potere $100 - \beta$, il numero di marcatori utilizzati, il *LOD score* ed il rate di falsi positivi α. Il test è condotto aumentando via via il numero di aplotipi da 10 a 60 facendo salti di 10.

La Tabella 3.13 riporta i risultati ottenuti utilizzando il dataset simulato e gli individui non imparentati in cui le fasi degli aplotipi sono ricostruite con *PHASE*.

Legame	$100 - \beta$	Numero marcatori	*LOD score*	α
Padre Figlio	80%	10	2.3494	0.05%
	95%	10	1.3511	0.35%
	99%	10	0.5654	1.02%
Fratelli	80%	10	1.3751	0.35%
	95%	10	-0.031	4.49%
	99%	20	0.0607	0.81%
2^0 grado	80%	20	0.5859	3.01%
	95%	40	0.1534	2.64%
	99%	60	-0.695	3.25%

Tabella 3.13. Sono riportati il legame di parentela, il potere $100 - \beta$, il numero di marcatori utilizzati, il *LOD score* ed il rate di falsi positivi α. Il test è condotto aumentando via via il numero di aplotipi da 10 a 60 facendo salti di 10.

La Tabella 3.14 riporta i risultati ottenuti utilizzando il dataset simulato e gli individui non imparentati in cui le fasi degli aplotipi sono ricostruite con *Gerbil*.

Legame	$100 - \beta$	Numero marcatori	*LOD score*	α
Padre Figlio	80%	10	2.3494	0.03%
	95%	10	1.3511	0.25%
	99%	10	0.5654	0.73%
Fratelli	80%	10	1.3751	0.25%
	95%	10	-0.031	3.08%
	99%	20	0.0607	0.37%
2^0 grado	80%	20	0.5859	1.23%
	95%	30	-0.170	1.97%
	99%	40	-1.105	2.32%

Tabella 3.14. Sono riportati il legame di parentela, il potere $100 - \beta$, il numero di marcatori utilizzati, il *LOD score* ed il rate di falsi positivi α. Il test è condotto aumentando via via il numero di aplotipi da 10 a 60 facendo salti di 10.

La Tabella 3.15 riporta i risultati ottenuti utilizzando il dataset in cui gli aplotipi vengono convertiti in singole SNP per un totale di 300 SNP (60×5 SNP).

Legame	$100 - \beta$	Numero marcatori	*LOD score*	α
Padre Figlio	80%	50	1.8224	2.31%
	95%	50	0.5042	4.80%
	99%	75	0.1120	1.67%
Fratelli	80%	75	1.5855	2.44%
	95%	125	0.1139	2.87%
	99%	175	-1.932	3.56%
2^0 grado	80%	275	1.0874	4.79%
	95%	300	-1.360	17.90%
	99%	300	-3.900	45.60%

Tabella 3.15. Sono riportati il legame di parentela, il potere $100 - \beta$, il numero di marcatori utilizzati, il *LOD score* ed il rate di falsi positivi α. Il test è condotto aumentando via via il numero di aplotipi da 25 a 300 facendo salti di 25.

Dataset simulato con $r^2 > 0.8$

I parametri utilizzati per la simulazione del dataset sono: 1000 famiglie; 60 aplotipi di 5 SNP; valore minimo di $r^2 = 0.8$; eterozigosità minima 0.15.

La Tabella 3.16 riporta i risultati ottenuti utilizzando solo il dataset simulato.

Legame	$100 - \beta$	Numero marcatori	*LOD score*	α
Padre Figlio	80%	10	0.7847	2.13%
	95%	20	0.9638	0.48%
	99%	20	0.0635	1.60%
Fratelli	80%	10	0.3883	4.32%
	95%	20	0.1719	2.44%
	99%	30	-0.486	2.49%
2^0 grado	80%	40	0.6422	2.19%
	95%	60	-0.078	4.11%
	99%	60	-1.216	18.48%

Tabella 3.16. Sono riportati il legame di parentela, il potere $100 - \beta$, il numero di marcatori utilizzati, il *LOD score* ed il rate di falsi positivi α. Il test è condotto aumentando via via il numero di aplotipi da 10 a 60 facendo salti di 10.

La Tabella 3.17 riporta i risultati ottenuti utilizzando il dataset simulato e gli individui non imparentati in cui le fasi degli aplotipi sono ricostruite con *PHASE*.

Legame	$100 - \beta$	Numero marcatori	*LOD score*	α
Padre Figlio	80%	10	0.7847	2.17%
	95%	20	0.9638	0.53%
	99%	20	0.0635	1.82%
Fratelli	80%	10	0.3883	4.38%
	95%	20	0.1719	2.54%
	99%	30	-0.486	2.60%
2^0 grado	80%	40	0.6422	2.24%
	95%	60	-0.078	4.28%
	99%	60	-1.216	19.45%

Tabella 3.17. Sono riportati il legame di parentela, il potere $100 - \beta$, il numero di marcatori utilizzati, il *LOD score* ed il rate di falsi positivi α. Il test è condotto aumentando via via il numero di aplotipi da 10 a 60 facendo salti di 10.

La Tabella 3.18 riporta i risultati ottenuti utilizzando il dataset simulato e gli individui non imparentati in cui le fasi degli aplotipi sono ricostruite con *Gerbil*.

Legame	$100 - \beta$	Numero marcatori	*LOD score*	α
Padre Figlio	80%	10	0.7847	1.91%
	95%	20	0.9638	0.43%
	99%	20	0.0635	1.42%
Fratelli	80%	10	0.3883	3.88%
	95%	20	0.1719	2.01%
	99%	30	-0.486	1.90%
2^0 grado	80%	30	0.3236	3.78%
	95%	50	-0.248	4.09%
	99%	60	-1.216	8.76%

Tabella 3.18. Sono riportati il legame di parentela, il potere $100 - \beta$, il numero di marcatori utilizzati, il *LOD score* ed il rate di falsi positivi α. Il test è condotto aumentando via via il numero di aplotipi da 10 a 60 facendo salti di 10.

La Tabella 3.19 riporta i risultati ottenuti utilizzando il dataset in cui gli aplotipi vengono convertiti in singole SNP per un totale di 300 SNP (60×5 SNP).

Legame	$100 - \beta$	Numero marcatori	*LOD score*	α
Padre Figlio	80%	75	2.0745	3.67%
	95%	100	1.0766	3.60%
	99%	125	-0.087	2.71%
Fratelli	80%	100	1.9929	3.52%
	95%	175	0.3669	3.39%
	99%	250	-2.428	3.91%
2^0 grado	80%	300	0.9536	8.53%
	95%	300	-2.266	31.57%
	99%	300	-5.118	61.13%

Tabella 3.19. Sono riportati il legame di parentela, il potere $100 - \beta$, il numero di marcatori utilizzati, il *LOD score* ed il rate di falsi positivi α. Il test è condotto aumentando via via il numero di aplotipi da 25 a 300 facendo salti di 25.

Capitolo 4

Discussione

Vengono riportati i risultati relativi ad uno studio sull'informatività di marcatori genetici nell'analisi di parentele (di primo e secondo grado) non note in studi di associazione caso controllo. La presenza di parentele ignote all'interno del dataset potrebbe alterare le frequenze alleliche del campione e portare il ricercatore ad associazioni false tra fenotipo e genotipo.

Per condurre lo studio sono stati sviluppati strumenti software in grado di elaborare le informazioni cliniche e genetiche su un set di famiglie (*Jenoware*) che sono indispensabili al successivo sviluppo dei programmi adibiti al calcolo dell'informatività di un set di marcatori (*SimulateMerlinFreqFile*, *IsRelated* e *ProcessIsRelated*). A corredo di questi strumenti sviluppati in *Java*, sono stati sviluppati alcuni programmi in *perl* che hanno la funzione di automatizzare tutte le fasi dei test e di offrire gli strumenti di conversione dati necessari per poter usufruire dei programmi di terze parti usati nel lavoro (*merlin*, *PHASE*, *Gerbil* e *Gevalt*). Utilizzando tutti gli strumenti sviluppati il ricercatore può condurre il test lanciando un programma (*doSimulate* o *doSimulate-ld*) che provvederà a sua

volta all'esecuzione di tutti i programmi richiesti.

Per determinare l'informatività dei marcatori genetici nell'analisi di parentele tra coppie di individui si costruisce in silico un set di famiglie in cui gli individui hanno genotipi (definiti dal ricercatore) la cui segregazione dei marcatori è simulata. Per verificare le diversi scenari sono stati creati in silico diversi set di famiglie. Avere un set di famiglie permette di testare su dati simulati di cui si conoscono con esattezza le caratteristiche (ad esempio, la frequenza di tutti gli alleli) permette di stabilire il rate di falsi negativi attraverso il calcolo su coppie realmente imparentate e poi di calcolare il rate di falsi positivi su un set di coppie formato da individui non imparentati all'interno dello stesso set di famiglie.

Storicamente, il mezzo più potente offerto dalla genetica per condurre test di parentela è offerto dai microsatelliti [Jones and Ardren, 2003] [Justice, 2000]. Sono marcatori ad elevata eterozigosità che, grazie alle molte varianti alleliche, permettono di stabilire nella maggioranza dei casi la segregazione dell'allele all'interno di una famiglia. Per testare l'informatività di marcatori ad alta eterozigosità è stato generato un set di 1000 famiglie in cui gli individui sono caratterizzati genotipicamente con 30 marcatori aventi eterozigosità $H > 0.70$. Il risultato del test è descritto in Tabella 3.2. Il risultato minimo voluto è stabilire quanti marcatori di un certo tipo sono necessari per ottenere un potere dell'80% con un rate di falsi positivi inferiore al 5%. Utilizzando 5 marcatori è possibile verificare se coppie di individui sono padre-figlio o fratelli con un potere del 80% ed un rate di falsi positivi di 1.05 e 3.74 rispettivamente. Con 5 marcatori è possibile verificare coppie padre-figlio con un potere del 99% (95) con un rate di falsi positivi 2.66 (2.08). Aumentando il numero di marcatori a 10 è possibile verificare la parentela su coppie di fratelli con un potere del 95% e rate di falsi positivi del 2.59. Per verificare parentele di secondo grado con potere 80% e rate

di falsi positivi 2.75 sono necessari 20 marcatori. I 30 marcatori simulati non sono sufficienti per verificare parentele di secondo grado con potere 99% con un rate inferiore al 5%.

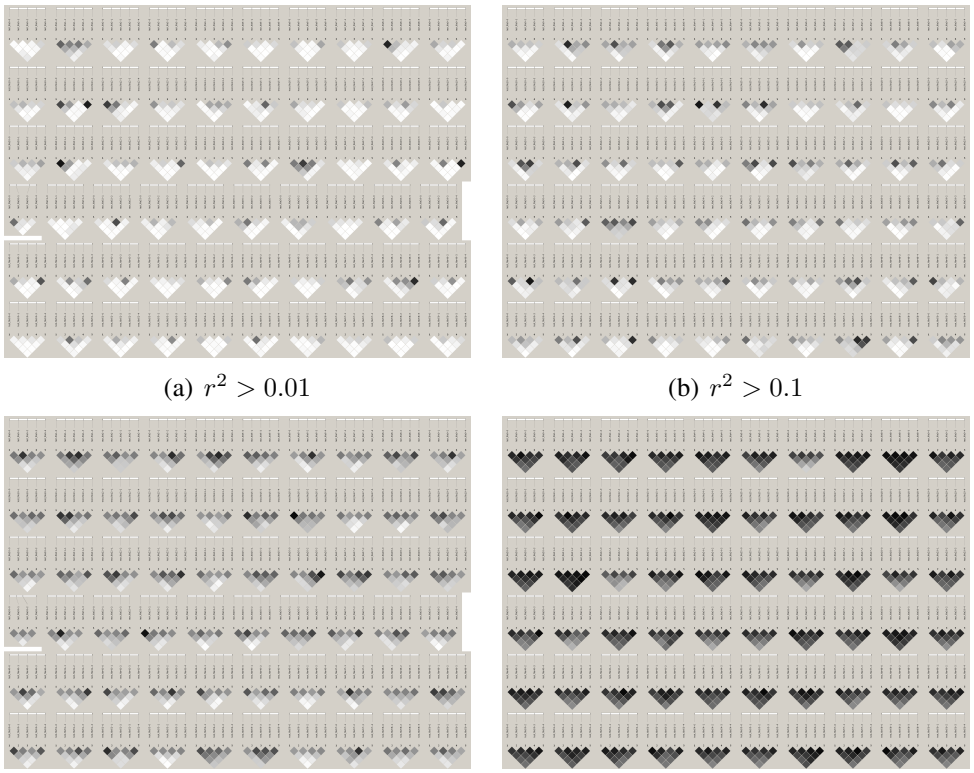
Negli studi di malattie complesse gli SNP sono i marcatori più utilizzati data la loro abbondanza nel genoma ed i bassi costi di genotipizzazione [Anderson and Garza, 2006]. Per testare l'informatività degli SNP è stato generato un set di 1000 famiglie in cui gli individui sono caratterizzati genotipicamente con 250 marcatori aventi eterozigosità $H > 0.15$. Il risultato del test è descritto in Tabella 3.3. Utilizzando 25 SNP è possibile verificare la parentela di primo grado su coppie di individui padre-figlio e di fratelli con un potere del 80% ed un rate di falsi positivi 2.24 e 4.09 rispettivamente. Per verificare parentele di secondo grado con potere 80% e rate di falsi positivi 4.60 sono necessari 100 marcatori. Per verificare parentele di primo grado su coppie padre-figlio e di fratelli con potere 99% e rate di falsi positivi 0.5 e 1.32 sono necessari 50 e 75 marcatori rispettivamente. I 250 marcatori simulati non sono sufficienti per verificare parentele di secondo grado con potere 99% con un rate inferiore al 5%.

Le analisi di parentela appena condotte testano la parentela su un marcatore alla volta per poi combinare tutti i risultati poiché assumono l'indipendenza dei marcatori [Jones and Ardren, 2003]. Nel caso di malattie complesse l'indipendenza di marcatori non è sempre vera. Dalla letteratura è noto come il linkage disequilibrium abbassi l'informatività dei marcatori nelle analisi di parentela [Anderson and Garza, 2006] [Ayres, 2000] [Jones and Ardren, 2003]. Se i loci sono in linkage disequilibrium gli alleli non segregano indipendentemente poiché alleli su loci diversi tendono a segregare assieme con una probabilità maggiore rispetto l'atteso [Devlin et al., 1988]. Per condurre analisi di parentela le SNP dovrebbero essere scelte in modo da essere indipendenti [Anderson and

Garza, 2006] [Ayres, 2000] ma questo non sempre è possibile poiché negli studi di associazione si ricercano sempre marcatori che siano casuali della malattia o abbiano un alto linkage disequilibrium con il marcatore causa malattia. Gran parte del genoma è costituito da regioni in forte linkage disequilibrium nelle quali le varianti alleliche sono molto correlate le une alle altre con il risultato che la regione può essere descritta con poche varianti aplotipiche molto comuni [Hirschhorn and Daly, 2005] [Consortium, 2005]. Tuttavia l'informatività di un set di marcatori in linkage disequilibrium può essere incrementata qualora si conosca la fase degli alleli [Jones and Ardren, 2003] [Devlin et al., 1988].

Per verificare l'effetto del linkage disequilibrium nell'analisi di parentela su individui raccolti per uno studio di associazione sono stati generati 4 set di 1000 famiglie in cui gli individui sono caratterizzati genotipicamente con 60 aplotipi di 5 SNP per un totale di 300 SNP. I programmi *IsRelated* e *ProcessIsRelated* sono in grado di gestire aplotipi con più alleli ma la simulazione del set di famiglie con il programma *merlin* vincola la dimensione massima degli alleli di un marcatore a 32 ovvero il numero di aplotipi possibili con 5 SNP ($2^5 = 32$). Le SNP adiacenti sono simulate in modo da presentare un livello minimo di r^2 che può essere 0.01, 0.1, 0.4 o 0.8. In Figura 4.1 sono mostrate 4 immagini ricavate usando *Gevalt* ognuna delle quali mostra, al variare del livello minimo di r^2, i valori di r^2 tra le SNP di un blocco rappresentandoli con toni di grigio che vanno dal bianco ($r^2 = 0$) fino al nero ($r^2 = 1$). Le 4 figure permettono di apprezzare graficamente l'aumentare di r^2 con le soglie 0.01, 0.1, 0.4 o 0.8.

Per ogni soglia di r^2 testata è stato generato un set di famiglie. Per ogni set sono state condotte 4 analisi per verificare l'efficacia dell'aplotipo a fase nota rispetto alle SNP singole in presenza di linkage disequilibrium e l'effetto della ricostruzione probabilistica della fase nel gruppo di individui non imparentati nel



(a) $r^2 > 0.01$ (b) $r^2 > 0.1$

(c) $r^2 > 0.4$ (d) $r^2 > 0.8$

Figura 4.1. Le immagini (a), (b), (c) e (d) riportano graficamente i valori di r^2 per tutti i 60 aplotipi formati da 5 SNP in linkage disequilibrium con i valori di r^2 minimo indicati.

test di paternità.

In Tabella 4.1 sono riportate - dalle tabelle 3.3 (SNP in assenza di LD), 3.7 (SNP con LD e $r^2 > 0.001$), 3.11 (SNP con LD e $r^2 > 0.01$), 3.15 (SNP con LD e $r^2 > 0.04$) e 3.19 (SNP con LD e $r^2 > 0.08$) - le numerosità di SNP sufficienti ad ottenere un potere del 80% con un rate di falsi positivi minore del 5%.

All'aumentare del linkage disequilibrium il numero di marcatori necessari per verificare un legame di parentela cresce. Per parentele di primo grado (Tabella 4.1) passiamo da 25 marcatori nel caso di SNP indipendenti a 75/100 marcatori necessari per verificare un legame padre-figlio/fratelli. Per verificare parentele di

Legame	$r^2 = 0$	$r^2 > 0.001$	$r^2 > 0.01$	$r^2 > 0.04$	$r^2 > 0.08$
Padre Figlio	25	25	50	50	75
Fratelli	25	50	50	75	100
2^0 grado	100	150	200	275	> 300

Tabella 4.1. Sono riportati il numero di SNP utilizzate per verificare un legame di parentela con potere 80% e rate di falsi positivi minore di 5%. Il valore > 300 indica che sono necessarie più di 300 SNP per verificare il legame di parentela per ottenere un rate di falsi positivi minore del 5%.

secondo (Tabella 4.1) grado bastano 100 SNP indipendenti mentre con un $r^2 > 0.8$ sono necessari più di 300 marcatori per avere un rate di falsi positivi inferiore al 5%.

In Tabella 4.2 sono riportate - dalle tabelle 3.4, 3.5, 3.6, 3.8, 3.9, 3.10, 3.12, 3.13, 3.14, 3.16, 3.17, 3.18 - le numerosità di aplotipi sufficienti ad ottenere un potere del 80% con un rate di falsi positivi minore del 5% (indicato a fianco del numero di SNP tra parentesi quadre).

Legame	Tipo	$r^2 > 0.001$ [α]	$r^2 > 0.01$ [α]	$r^2 > 0.04$ [α]	$r^2 > 0.08$ [α]
Padre	S	10 [$5x10^{-4}$]	10 [0.006]	10 [0.05]	10 [2.13]
Figlio	P	10 [0]	10 [0.005]	10 [0.05]	10 [2.17]
	G	10 [$6x10^{-4}$]	10 [0.003]	10 [0.03]	10 [1.91]
Fratelli	S	10 [0.01]	10 [0.08]	10 [0.32]	10 [4.32]
	P	10 [0.02]	10 [0.09]	10 [0.35]	10 [4.38]
	G	10 [0.03]	10 [0.07]	10 [0.25]	10 [3.88]
2^0 grado	S	20 [0.37]	20 [1.14]	20 [2.75]	40 [2.19]
	P	20 [0.55]	20 [1.46]	20 [3.01]	40 [2.24]
	G	20 [0.28]	20 [0.53]	20 [1.23]	30 [3.78]

Tabella 4.2. Sono riportati il numero di aplotipi utilizzati per verificare un legame di parentela con potere 80% e rate di falsi positivi minore di 5% (riportato tra parentesi quadre). La colonna $Tipo$ indica se si sono utilizzati gli aplotipi simulati (S) o se le fasi degli aplotipi sono state ricostruite probabilisticamente con *PHASE* (P) o con *Gerbil* (G); Ogni aplotipo è composto da 5 SNP per cui 10 aplotipi = 50 SNP e 60 aplotipi = 300 SNP.

Osservando i risultati mostrati in Tabella 4.2 si nota come il numero di aplotipi utilizzati per verificare una parentela sia sempre uguale con ognuno dei tre metodi

utilizzati. L'unica eccezione è la parentela di secondo grado in cui $r^2 > 0.8$ e gli aplotipi sono ricostruiti con il programma *Gerbil*.

Con 40 aplotipi (200 SNP) il test riesce a verificare l'ipotesi di parentela di secondo grado, mentre con 300 SNP prese singolarmente era impossibile. Un risultato interessante è che 10 aplotipi (50 SNP) sono sufficienti a verificare l'ipotesi di parentela di prima grado per qualsiasi livello di linkage disequilibrium tra SNP. L'aumentare del Linkage disequilibrium peggiora il rate di falsi positivi che rimane comunque al di sotto della soglia 5%. Per verificare parentele di secondo grado occorrono almeno 20 aplotipi (100 SNP) o 40 aplotipi (200 SNP) nel caso di forte linkage disequilibrium ($r^2 > 0.8$).

Capitolo 5

Conclusioni

In conclusione, le analisi eseguite in questo lavoro hanno permesso di sviluppare strumenti informatici che consentono di individuare parentele biologiche non note fra gli individui che partecipano agli studi di associazione effettuati per la ricerca della componente genetica di malattie complesse. Questi studi sono particolarmente frequenti nella letteratura scientifica e possono portare a conclusioni errate nel caso in cui siano presenti delle parentele biologiche non identificate fra gli individui reclutati.

In secondo luogo questo lavoro dimostra come l'utilizzo di marcatori in linkage disequilibrium – utilizzati normalmente in malattie complesse nelle analisi basate su geni candidati o su regioni genomiche associate per un mappaggio fine del fattore genetico coinvolto – riduca il numero effettivo dei marcatori informativi e come la conoscenza della fase dei marcatori permetta di aumentarne l'informatività. Il confronto dei risultati ottenuti con il test fatto utilizzando gli aplotipi simulati ed i test fatti utilizzando gli aplotipi ricostruiti probabilisticamente suggerisce che i programmi *PHASE* e *Gerbil* possono essere

utilizzati per condurre test di parentela nel caso di SNP in linkage disequilibrium. I rate di falsi positivi ottenuti con diversi valori di r^2, gli aplotipi simulati e gli aplotipi ricostruiti con *PHASE* e *Gerbil* mostrano come all'aumentare del linkage disequilibrium *Gerbil* riesca ad abbassare il rate di falsi positivi. Successive indagini dovranno verificare l'affidabilità dei risultati ottenuti con *Gerbil* e con *PHASE*.

Bibliografia

Abecasis, G. R., Cherny, S. S., Cookson, W. O., and Cardon, L. R. (2002). Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet*, 30(1):97–101.

Anderson, E. C. and Garza, J. C. (2006). The power of single-nucleotide polymorphisms for large-scale parentage inference. *Genetics*, 172(4):2567–2582.

Ayres, K. L. (2000). Relatedness testing in subdivided populations. *Forensic Sci Int*, 114(2):107–115.

Barrett, J. C., Fry, B., Maller, J., and Daly, M. J. (2005). Haploview: analysis and visualization of ld and haplotype maps. *Bioinformatics*, 21(2):263–265.

Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet*, 32(3):314–331.

Consortium, I. H. (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299–1320.

Davidovich, O., Kimmel, G., and Shamir, R. (2007). Gevalt: an integrated software tool for genotype analysis. *BMC Bioinformatics*, 8:36.

Devlin, B. and Risch, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, 29(2):311–322.

Devlin, B., Roeder, K., and Ellstrand, N. C. (1988). Fractional paternity assignment: theoretical development and comparison to other. *TAG Theoretical and Applied Genetics*, Volume 76(Number 3):369–380.

GlaxoSmithKline and SIGU (2002). *Incontri 3 - Dalla Ricerca Genetica alla pratica clinica*. GlaxoSmithKline e Società Italiana di Genetica Umana.

Gomes, I., Collins, A., Lonjou, C., Thomas, N. S., Wilkinson, J., Watson, M., and Morton, N. (1999). Hardy-weinberg quality control. *Ann Hum Genet*, 63(Pt 6):535–538.

Grimes, D. A. and Schulz, K. F. (2002). An overview of clinical research: the lay of the land. *Lancet*, 359(9300):57–61.

Hedrick, P. and Kumar, S. (2001). Mutation and linkage disequilibrium in human mtdna. *Eur J Hum Genet*, 9(12):969–972.

Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet*, 6(2):95–108.

Jones, A. G. and Ardren, W. R. (2003). Methods of parentage analysis in natural populations. *Mol Ecol*, 12(10):2511–2523.

Justice, U. D. O. (2000). *The Future of Forensic Dna Testing: Predictions of the Research and Development Working Group*. University Press of the Pacific.

Kaiser, L. and Seber, G. A. (1985). Paternity testing. 2: Likelihood ratio tests. *Am J Med Genet*, 20(2):209–219.

Lander, E. S. and Schork, N. J. (1994). Genetic dissection of complex traits. *Science*, 265(5181):2037–2048.

Presciuttini, S., Toni, C., Tempestini, E., Verdiani, S., Casarino, L., Spinetti, I., Stefano, F. D., Domenici, R., and Bailey-Wilson, J. E. (2002). Inferring relationships between pairs of individuals from locus heterozygosities. *BMC Genet*, 3:23.

Risch, N. J. (2000). Searching for genetic determinants in the new millennium. *Nature*, 405(6788):847–856.

Spielman, R. S. and Ewens, W. J. (1996). The tdt and other family-based tests for linkage disequilibrium and association. *Am J Hum Genet*, 59(5):983–989.

Stephens, J. W. and Humphries, S. E. (2003). The molecular genetics of cardiovascular disease: clinical implications. *J Intern Med*, 253(2):120–127.

Stephens, M. and Donnelly, P. (2003). A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet*, 73(5):1162–1169.

Suzuki, D. T., Griffiths, A. J. F., Miller, J. H., and Lewontin, R. C. (1992). *Genetica, principi di analisi formale*. Zanichelli.

Terwilliger, J. D. and Jurg Ott, J. (1994). *Handbook of Human Genetic Linkage*. The Johns Hopkins University Press, 1 edition.

Trégouët, D. A., Ducimetière, P., and Tiret, L. (1997). Testing association between candidate-gene markers and phenotype in related individuals, by use of estimating equations. *Am J Hum Genet*, 61(1):189–199.

Wall, L. (1993-2002). Embperl.

Weir, B. S. (1996). *Genetic Data Analysis II*. Sinauer Associates, 2 edition.

Wills-Karp, M. and Ewart, S. L. (2004). Time to draw breath: asthma-susceptibility genes are identified. *Nat Rev Genet*, 5(5):376–387.