UNIVERSITA' DEGLI STUDI DI SASSARI

FACOLTA' DI MEDICINA E CHIRURGIA

DIPARTIMENTO DI SCIENZE BIOMEDICHE

DOTTORATO IN GENETICA MEDICA, MALATTIE METABOLICHE E NUTRIGENOMICA

CICLO XXIV

LA SCLEROSI MULTIPLA; APPROCCI PER IDENTIFICARE LOCI COINVOLTI NELLA PATOGENESI. SEQUENZIAMENTO *LOW-PASS* E STUDIO DI ASSOCIAZIONE SU TUTTO IL GENOMA IN SARDEGNA.

Relatore:
Prof. Francesco Cucca

Dottoranda:
Dott.ssa Maristella Pitzalis

Anno Accademico 2010-2011

(Settore scientifico disciplinare di afferenza MED/03)

La Sclerosi Multipla

"…la memoria è diminuita,
i pensieri si formano
lentamente, e le facoltà
intellettive ed emozionali
sono indebolite nella loro
globalità…"

Charcot, 1877

INDICE

ELENCO FIGURE

ELENCO TABELLE

1 PREFAZIONE

L'argomento trattato, che corrisponde alla principale linea di ricerca seguita in questi anni, è la dissezione genetica di malattie complesse a carattere autoimmune ed a elevata incidenza in Sardegna, quali il Diabete di Tipo 1 (DT1) e la Sclerosi Multipla (SM). Fin dal 2001, ho svolto tale attività presso il laboratorio di Immunogenetica del Prof. Francesco Cucca.

Per diversi anni, a causa dei limiti tecnologici e metodologici è stato difficile ottenere risultati certi ed inequivocabili nell'ambito delle malattie complesse, per qualunque variante localizzata al di fuori della regione HLA e di pochissime altre regioni geniche. Conseguentemente, ben poco si conosceva sulle basi genetiche sia della sclerosi multipla che del diabete di tipo 1 e di altre malattie autoimmuni.

Per anni, la strategia d'elezione è stata lo studio di Linkage su tutto il genoma, la quale prevedeva l'analisi di un numero limitato di marcatori microsatelliti, dislocati su tutto il genoma, in famiglie con almeno due figli affetti. Le analisi di Linkage hanno avuto un buon successo nello svelare loci causali in malattie di tipo Mendeliano, per questo motivo si pensava potessero essere efficaci anche nelle malattie complesse. Solo dopo diversi anni e molti sforzi, questi studi si sono rivelati del tutto inefficaci ed inappropriati nelle malattie complesse.

Negli ultimi anni, che comprendono i miei tre anni di dottorato in Genetica Medica, Malattie Metaboliche e Nutrigenomica, la situazione è notevolmente mutata.

Grazie al progresso delle tecnologie di tipizzazione e all'abbattimento dei costi, è diventata accessibile la tipizzazione di centinaia di migliaia di marcatori polimorfici (fino ad circa 1 milione contemporaneamente), dislocati lungo tutto il genoma, in migliaia di individui, allo scopo di effettuare studi di associazione su tutto il genoma, (Genome Wide Association Study, GWAS).

Allo stato dell'arte, la prima generazione di GWAS, grazie anche all'ampia casistica richiesta in questa tipologia di studi, in soli 5 anni, ha generato 951 pubblicazioni ed identificato oltre 1400 varianti inequivocabilmente associate con 221 tratti e malattie multifattoriali (NHGRI GWA Catalog, www.genome.gov/GWAStudies). In particolare sono state descritte oltre 40 varianti geniche implicate nella patogenesi del T1D e oltre 50 nella SM.

Attraverso un studio di associazione che ha analizzato circa 6.6 milioni di marcatori a singolo nucleotide, (Single Nucleotide Polymorphism, SNP) in 882 pazienti SM e 872 controlli, abbiamo identificato un nuovo gene (l'ubiquitin-protein ligase *CBLB*) implicato nella patogenesi della SM, lavoro pubblicato in Nature Genetics il 9 maggio 2010 (S. Sanna, **M. Pitzalis**, M. Zoledziewska et all. Nature Genetics 2010, "Variants within the immunoregulatory CBLB gene are associated with multiple sclerosis").

Nonostante i successi ottenuti dai GWAS, ancora una larga porzione dell'ereditabilità dei tratti complessi rimane inspiegata. Al tempo stesso, grazie ai progressi della tecnologia Next Generation Sequencing (NGS) e alla diffusione dei sequenziatori di nuova generazione, è diventato possibile sequenziare tutto il genoma umano in tempi ridotti ed a costi accessibili, aprendo una nuova era della genetica che consentirà di ampliare ulteriormente le conoscenze delle basi genetiche delle malattie umane.

Ho trascorso i tre anni di eccellente formazione, in gran parte, presso il laboratorio di Immunogenetica del Dipartimento di Scienze Biomediche, Università di Sassari, dislocato all'interno del Bioparco Sardegna Ricerche a Pula, e guidato dal Prof. Francesco Cucca.

Ho partecipato ai progetti relativi alla dissezione delle basi genetiche della SM e del DT1, ad un progetto di ri-sequenziamento dell'intero genoma in migliaia di individui sardi per l'inerente catalogazione di varianti rare e fondatrici, progetto per il quale ho anche ottenuto un finanziamento dalla Regione Sardegna, grazie al bando Giovani Ricercatori. Nel contesto di tali progetti ho contribuito al disegno sperimentale degli studi, organizzato la raccolta dei campioni, la gestione del *database* dei dati clinici ed anamnestici, la preparazione dei campioni, la tipizzazione e l'interpretazione dei risultati.

Ho scelto di discutere la tesi sulla dissezione genetica nella sclerosi multipla, con lo scopo di esporre i risultati conseguiti fin ora, grazie a studi di associazione su tutto il genoma, integrati da metodi di imputazione statistica che beneficiano di dati di sequenza *low-pass* di tutto il genoma di individui sardi e di origine sarda.

2 INTRODUZIONE

2.1 Sclerosi Multipla; caratteristiche generali della patologia

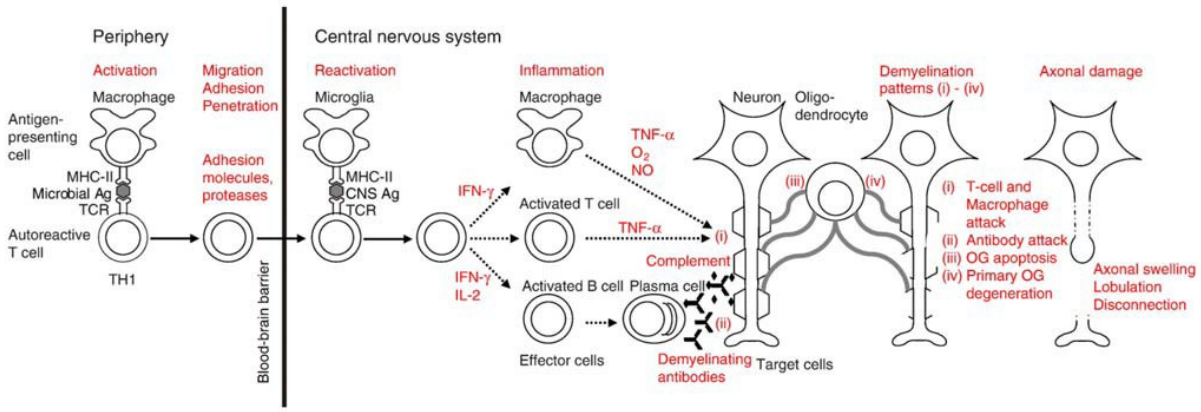
La sclerosi multipla (SM) è una grave malattia infiammatoria cronica demielinizzante, a carattere autoimmune, che colpisce il sistema nervoso centrale (SNC) [1,2]. A seguito dell'attacco autoimmune viene gradualmente distrutta la guaina mielinica (demielinizzazione) che riveste parte del corpo dei neuroni (assoni). La mielina, una sostanza ricca di acidi grassi, permette la trasmissione rapida ed integra degli impulsi nervosi, che dal cervello e dal midollo spinale si dipartono verso le parti periferiche del corpo e viceversa. La distruzione della guaina mielinica, nel sistema nervoso centrale, causa il blocco o il rallentamento della normale conduzione degli impulsi nervosi portando al manifestarsi di un'estrema varietà di sintomi, propri di questa malattia. Le aree in cui la mielina viene danneggiata vengono chiamate "placche" da cui prende anche il nome di "Sclerosi a Placche".

Le placche sono tipiche lesioni infiammatorie causate dall'attacco del sistema autoimmunitario, da parte dei linfociti T autoreattivi attivati (in particolare CD8+) nei confronti del rivestimento mielinico. Il linfociti T sono capaci di guidare l'evento infiammatorio, con produzione di citochine pro-infiammatorie (Interferone gamma (INFγ), Tumor necrosis factor alfa (TNFα) e Interleuchina 2 (IL-2) e richiamo di ulteriori cellule mononucleate che superano la barriera ematoencefalica quali; linfociti B, macrofagi, che fagocitano i frammenti di mielina, e i polimorfonucleati, che liberano sostanze citotossiche e citolitiche.

Il modello patogenetico della SM attualmente proposto, rappresentato schematicamente in figura 1, vede l'attivazione dei linfociti T pro-infiammatori nella periferia. L'attivazione è causata dal riconoscimento, da parte del recettore delle cellule T (T cell receptor, TCR), di antigeni presentati sul complesso maggiore di istocompatibilità di classe II (Major histocompatibility complex, MHC-II), dalle cellule presentanti l'antigene (Antigen Presentig Cell, APC). Tali linfociti T, migrano, aderiscono e penetrano la barriera ematoencefalica mediante meccanismi di adesione molecolare, e con l'intervento di proteasi e citochine.

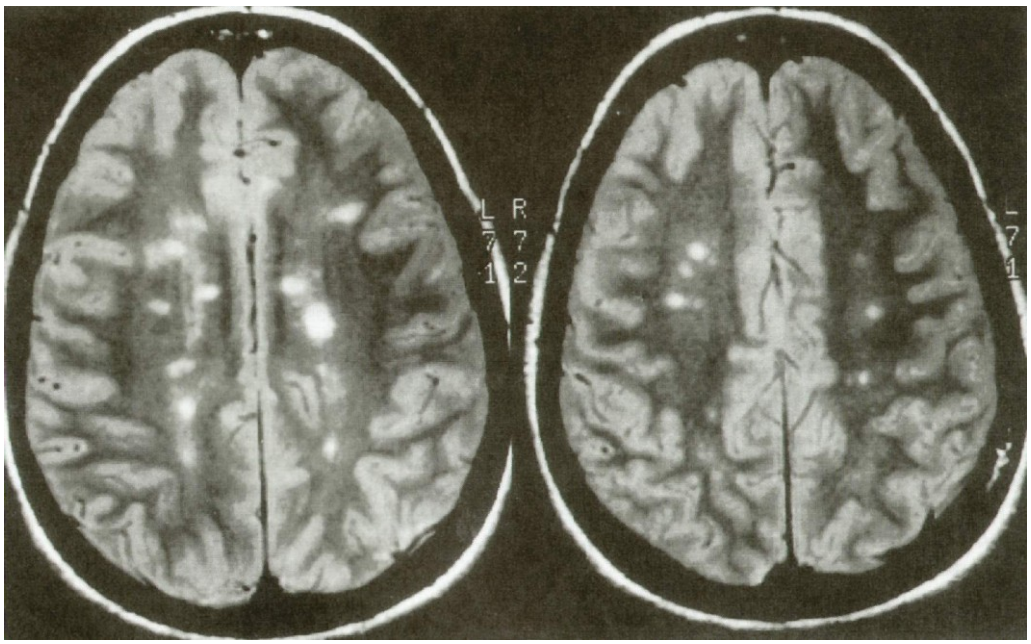
All'interno del sistema nervoso centrale, i linfociti T vengono riattivati dal MHC-II sulle APC e iniziano a produrre citochine pro-infiammatorie, che promuovono lo stato infiammatorio nel SNC con conseguente attivazione di molecole effettrici come macrofagi, linfociti B e altri linfociti T. I macrofagi ed i linfociti T attaccano la guaina mielinica attraverso dei mediatori citotossici, soprattutto il TNF-α, le specie radicaliche dell'ossigeno (O2) e l'ossido nitrico (NO). I linfociti B si differenziano in plasmacellule che secernono anticorpi demielinizzanti. Quest'ultimi attivano altri i macrofagi e la cascata del complemento che causano il danno mielinico [3].

Figura 1. Meccanismo patogenetico della SM



Le placche, generatesi da tale processo autoimmune, vengono definite multifocali sia in senso spaziale, in quanto possono comparire in diverse aree del sistema nervoso centrale, che in senso temporale, in quanto alcune placche regrediscono completamente, ma in generale il numero di lesioni aumenta nel tempo. Queste regioni d'infiammazione possono essere rilevate con tecniche di *neuroimaging*, quali la risonanza magnetica. La figura 2 rappresenta un esempio d'immagine estrapolata da una risonanza magnetica di un paziente SM.

Figura 2. Risonanza Magnetica nella SM



In figura 2 sono rappresentate due sezioni assiali al di sopra del livello dei ventricoli. Le lesioni caratteristiche della sclerosi multipla, identificate dalle aree con maggior contrasto (macchie bianche) variano per localizzazione, dimensione e intensità.

Le manifestazioni della malattia possono variare a seconda delle aree colpite (encefalo e midollo spinale) ed i sintomi possono interessare diverse funzioni dell'organismo, regolate dal sistema nervoso centrale; quali il movimento e la coordinazione con un generale senso di affaticamento, la sensibilità, la vista, l'equilibrio, la parola, le funzioni sfinteriche e talvolta anche le funzioni cognitive.

La SM è caratterizzata da un decorso clinico variabile e vengono riconosciute diverse forme, riconducibili ad un diverso andamento della patologia, cosi come rappresentato in figura 3 [4].

Si possono distinguere quattro forme; la recidivante-remittente, la secondariamente progressiva, la primariamente progressiva e la progressiva con ricadute. A queste si aggiunge una quinta forma detta **SM benigna**, la quale ha la peculiarità di esordire con uno o due episodi acuti, seguiti da un recupero completo che non lascia tracce di

disabilità e non peggiora con il passare del tempo. Questa forma viene individuata anche quando è presente solo una minima disabilità, per almeno 15 anni dalla data di esordio. In generale la SM benigna tende a essere associata a sintomi sensitivi (parestesie) o visivi (neurite ottica).

La forma clinica più frequente (circa l'85%) è rappresentata dalla SM a **decorso recidivante-remittente (SM-RR)**, nella quale si presentano episodi acuti di malattia(detti 'poussè' o 'ricadute', che regrediscono del tutto o in parte) alternati a periodi di benessere (definiti 'remissioni'). Le recidive si verificano circa una volta all'anno e queste ricadute inducono la rapida insorgenza di difetti neurologici, differenti in base alle regioni, del cervello o del midollo spinale, coinvolte. Queste recidive sono di solito seguite da un certo recupero delle funzioni neurologiche perse, chiamata fase di remissione.

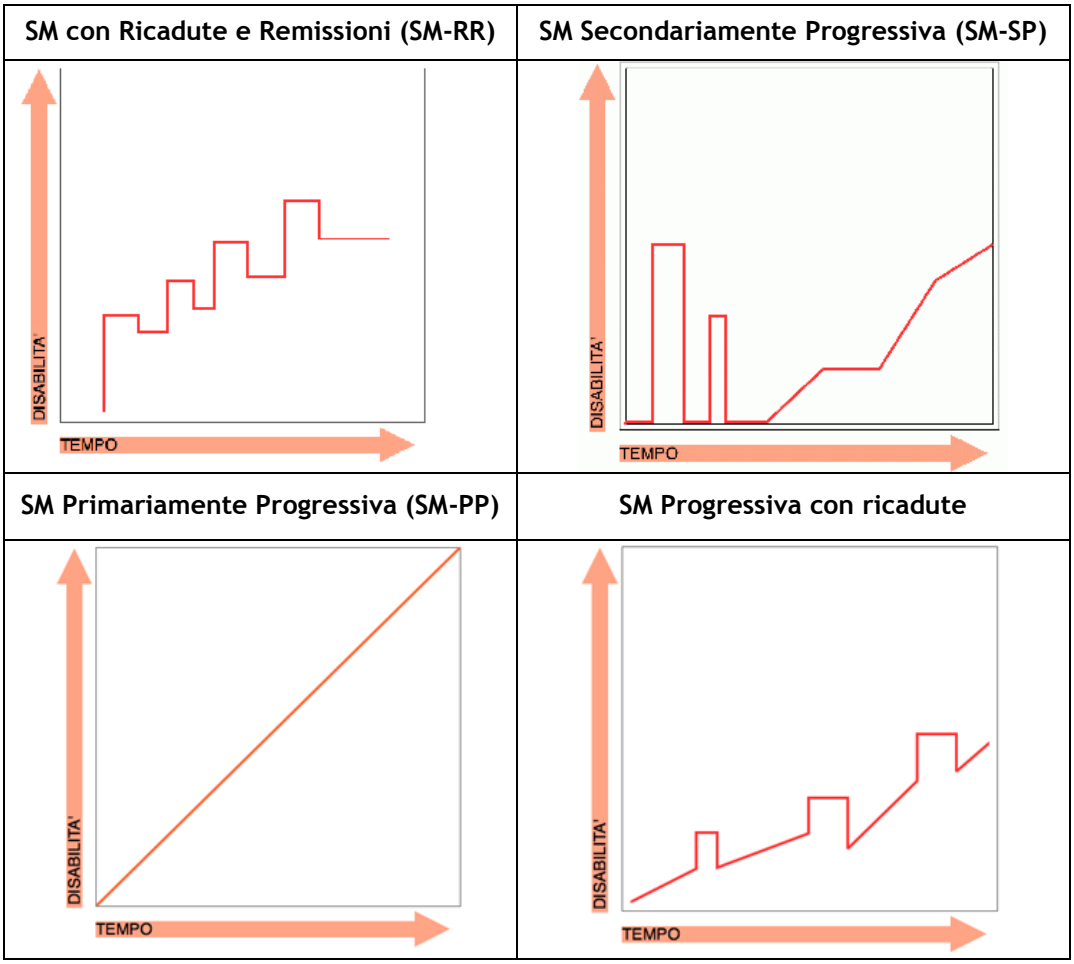
La **SM secondariamente progressiva (SM-SP),** si sviluppa come evoluzione della forma recidivante-remittente ed è caratterizzata da una disabilità persistente che progredisce gradualmente nel tempo. Circa il 30-50% delle persone con SM, che inizialmente hanno una forma recidivante-remittente, sviluppano entro 10 anni circa, una forma secondariamente progressiva.

La **SM primariamente progressiva (SM-PP),** che colpisce circa il 10% degli affetti SM**,** è caratterizzata dall'assenza di vere e proprie ricadute; all'esordio i sintomi iniziano in modo graduale e tendono a progredire lentamente nel tempo.

Infine nel 5% dei casi, oltre al presentarsi di un andamento progressivo dall'esordio, si manifestano anche episodi acuti di malattia, con scarso recupero dopo l'episodio (**decorso progressivo con ricadute**).

I modelli animali di SM, noti come encefalomielite autoimmune sperimentale (EAE), possiedono sia la forma progressiva che la recidivante-remittente della malattia, riflettendo così le due principali forme di più SM [5].

Figura 3. Rappresentazione grafica delle varianti cliniche della SM

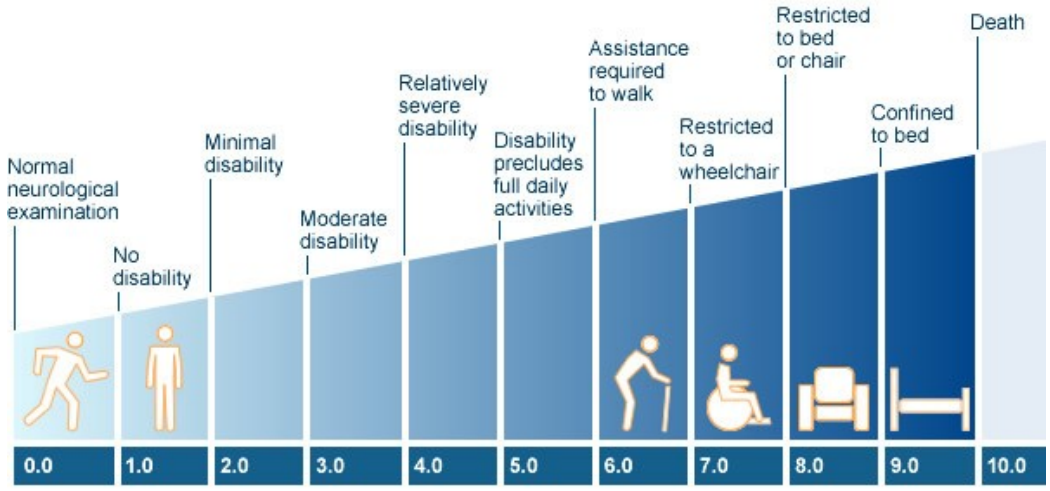


Secondo la classificazione dell'Organizzazione Mondiale della Sanità (OMS), il peso della malattia sulla qualità di vita del paziente SM può essere descritto in termini di;

- "*impairment*"(insieme di deficit neurologici)
- "*handicap*" (limitazioni nelle attività sociali e lavorative)
- "*disability*" (limitazioni nelle attività di vita quotidiana)

Il grado di severità della malattia viene valuta attraverso un punteggio da 0 a 10, definito dalla scala clinica proposta dal neurologo americano Kurtzke nel 1983; Expanded Disability Status Scale, EDSS [6] e riproposta in figura 4.

Figura 4. Valutazione del grado di disabilità; EDSS



Sino ad ora, non esistono trattamenti curativi a disposizione per la SM, sebbene siano disponibili diversi farmaci che svolgono un ruolo nel modificare e rallentare il decorso della malattia, ovvero, riducono il numero di attacchi della forma più comune recidivante-remittente della malattia.

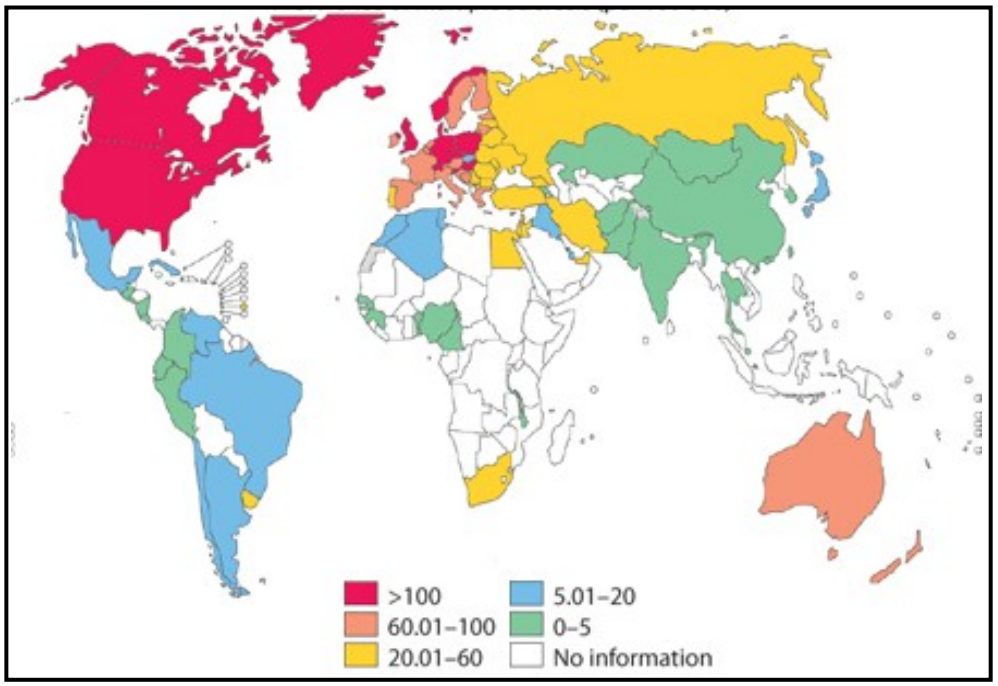
La SM è una patologia importante, altamente disabilitante che colpisce prevalentemente i giovani adulti, e per questo rappresenta un notevole impegno del capitolo di spesa della salute pubblica, rappresentando un forte impatto socio-economico.

Trovare i fattori genetici di suscettibilità alla malattia è fondamentale per svelare i meccanismi ed i *pathways* coinvolti nell'insorgenza dei danni al SNC, nelle persone affette da SM, al fine di comprenderne anche i fondamenti per la protezione dello stesso. Tali risultati si rendono necessari per evidenziare nuovi bersagli terapeutici e sviluppare nuove ed efficaci terapie.

2.2 Epidemiologia

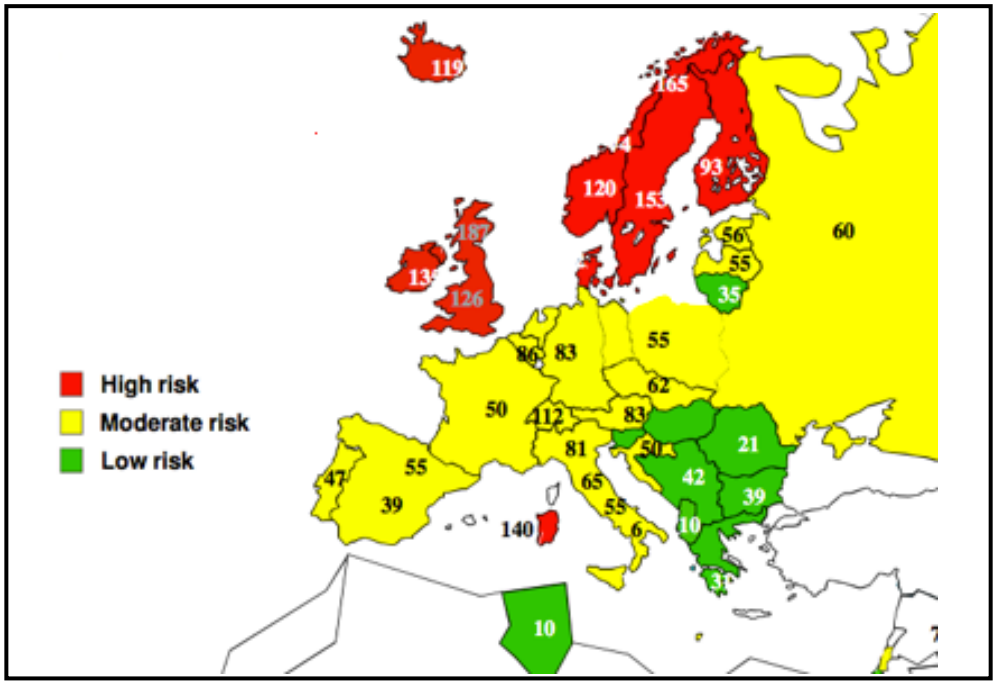
La SM ha un esordio variabile, tra i 15 e i 50 anni, anche se si manifesta soprattutto tra i giovani adulti, tra i 30 e i 40 anni, e prevalentemente nel sesso femminile, in un rapporto di uno a due, rispetto agli uomini. L' OMS, stima che nel mondo ci siano oltre 2,5 milioni di persone affette. Negli Stati Uniti, la patologia colpisce circa 400.000 persone ed in Italia si stima ci siano circa 50.000 individui affetti. La patologia è molto frequente tra le popolazioni caucasiche (soprattutto tra quelle residenti nel nord-ovest europeo), nel nord America, nel sud-est dell'Australia e in Nuova Zelanda, sud-Africa e America meridionale, mentre si riscontra una bassa incidenza in Asia e nelle regioni caraibiche. La figura 5 mostra, il differente *range* di prevalenza nei paesi di tutto il mondo .

Figura 5. Prevalenza della sclerosi multipla nel mondo (per 100,000)



In Europa l'incidenza sembra seguire un gradiente nord-sud, con una più alta prevalenza nei paesi del nord, soprattutto in Scandinavia, e bassa nei paesi del sud, ad eccezione della Sardegna che mostra una prevalenza due volte superiore rispetto al resto della popolazione italiana e alla maggior parte delle popolazioni caucasiche (140 casi per 100,000 abitanti) equiparabile alla prevalenza dei paesi nord Europei, come indicato dalla figura 6 [7,8].

Figura 6. Prevalenza della sclerosi multipla in Europa (per 100,000)



2.3 Basi eziopatogenetiche

L'eziologia della malattia non è ancora del tutto nota. In questi ultimi cinque anni, grazie agli studi di associazione su tutto il genoma, è stata notevolmente ampliata la conoscenza delle basi genetiche della SM, tuttavia ancora una larga fetta dell'ereditabilità genetica (circa l'80%) della malattia rimane ancora da spiegare .

La SM è un malattia complessa, con una forte componente genetica, definita dall'azione congiunta di più varianti alleliche a diversi loci, dislocati lungo il genoma, ancora in larga parte sconosciuti. Numerosi studi effettuati sull'ereditarietà della patologia, condotti sui gemelli e sui fratelli di individui affetti, hanno messo in evidenza l'importanza del ruolo dei fattori genetici nella predisposizione alla malattia. I geni giocano un ruolo importante nello sviluppo della SM, ma non possono spiegarne completamente l'eziologia. Infatti il rischio d'ammalare di SM aumenta con l'aumentare delle relazioni parentali con la persona affetta [9], ma i gemelli monozigoti sono concordanti per la malattia solo per circa il 30% [10]. Ciò evidenzia chiaramente una penetranza incompleta e l'importanza della genetica nello sviluppo della SM, ma sottolinea la necessità, per la piena estrinsecazione del rischio, di fattori ambientali permissivi [11], nonché della modulazione di fattori post-trascrizionali ed epigenetici (processi che determinano modificazioni ereditarie del funzionamento dei geni senza variazioni nella sequenza del DNA), altrettanto importanti per il rischio globale di sviluppare la SM.

Attualmente, non sono ancora noti i fattori ambientali coinvolti nell'eziologia della sclerosi multipla, sebbene alcuni elementi, come particolari infezioni nei primi anni di vita, esposizione all'Epstein Barr virus (EBV) e all'Human Herpes Virus 6 (HHV6), carenza di vitamina D ed esposizione al fumo, siano stati indicati più volte [12-14],

L'identificazione dei fattori ambientali è complicata da numerose variabili. La difficoltà principale deriva dal fatto che la manifestazione clinica della malattia, conseguenza della progressiva demielizzazione degli assoni, si riscontra solo dopo molti anni dall'inizio del processo patogenetico. Sarebbero quindi necessari lunghi e faticosi studi prospettici che seguano gli individui, reclutati nello studio, per diversi decenni fino all'esordio della malattia.

Il campo della genetica e l'identificazione dei geni di malattia, potrebbero altresì, fornire un aiuto importante nell'indicare i fattori ambientali che svolgono un ruolo nel rischio di sviluppare la patologia. Infatti i geni di malattia, le proteine correlate ed i meccanismi d'azione, potrebbero suggerire diversi fattori ambientali implicati nell'eziopatogenesi della SM, difficilmente evidenziabili con studi di tipo prospettico.

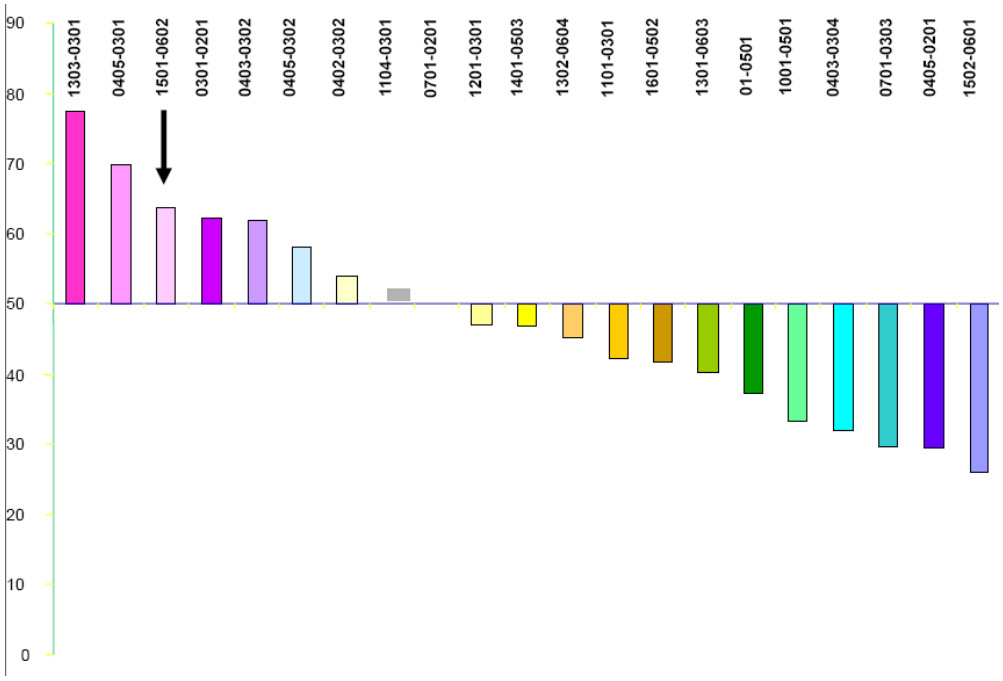
Tuttavia, anche la ricerca dei fattori genetici, per molti anni ha rappresentato, ed ancora oggi rappresenta, un ardua sfida.

Nei primi anni '70 è stata illustrata l'associazione di alcuni aplotipi della regione dell'Istocompatibilità maggiore (MHC)/ Antigene Leucocitario Umano (HLA), in 6q31, con la SM [15], ma ancor oggi rimangono da chiarire gli effetti primari delle varianti genetiche che contribuiscono al rischio ereditato. La regione HLA/MHC rappresenta ad oggi il locus che esercita il maggiore effetto di rischio genetico della SM.

E' nota la presenza di un esteso aplotipo predisponente, DRB1*1501-DQB1*0602 (anche definito sierologicamente, aplotipo DR2, DQw1), costantemente associato in diversi gruppi etnici [16-20].

Nella popolazione sarda, la SM è stata associata a diversi aplotipi dei loci DRB1 - DQB1; quali il DRB1*0301-DQB1*0201 e DRB1*0405-DQB1*0301 [21-22]. In particolare sono stati descritti cinque aplotipi DRB1- DQB1 associati positivamente con la SM; DRB1*1303-DQB1*0301, DRB1*0405-DQB1*0301, DRB1*0301-DQB1*0201, DRB1*1501-DQB1*0602, DRB1*0405-DQB1*0302 e nessun aplotipo negativamente associato alla malattia in maniera significativa [23], come evidenziato dalla figura 7.

Figura 7.Aplotipi dei loci *DRB1-DQB1* predisponenti e protettivi nella SM in Sardegna



L'esatto meccanismo con cui il locus *DRB1* influenza la suscettibilità della SM rimane ancora da chiarire. Tali meccanismi, tuttavia, evidentemente solo legati alla funzione fisiologica delle molecole HLA nei vari processi immunologici, come il legame dell'antigene, la presentazione, e la determinazione del repertorio delle cellule T.

Il recente progresso della genomica ha dimostrato come una parte rilevante della variabilità tra individui sia da attribuirsi ai polimorfismi con variazione di un singolo nucleotide, i quali hanno acquistato particolare rilevanza in campo biomedico, in quanto possono essere utilizzati come marcatori degli studi di mappaggio genico ma possono essere loro stessi causa di patologie o tratti complessi.
Secondo alcune stime, nella popolazione mondiale, si trovano circa dieci milioni di SNPs per i quali entrambi gli alleli sono presenti con una frequenza superiore all'1% [24].
Tali SNPs (*common SNPs*) costituiscono il 90% della variabilità nella popolazione mondiale, mentre il restante 10% è costituito da un insieme diversificato di altre varianti rare.
Grazie alla disponibilità della sequenza completa del genoma umano, al progetto HapMap, ed al completamento della fase I del progetto 1000 Genomes, con la descrizione di più di 15.000.000 di SNPs, 1.000.000 di inserzioni e delezioni, 20.000 varianti strutturali, sono state sviluppate nuove piattaforme di caratterizzazione genotipica degli SNPs che consentono lo studio contemporaneo fino ad un milione di marcatori in migliaia di individui in tempi ridotti e con una notevole riduzione dei costi.
Questi progressi della tecnologia hanno incentivato la diffusione degli studi di associazione su tutto il genoma che hanno avuto un notevole impatto nell'identificazione dei geni responsabili di malattie e tratti comuni.

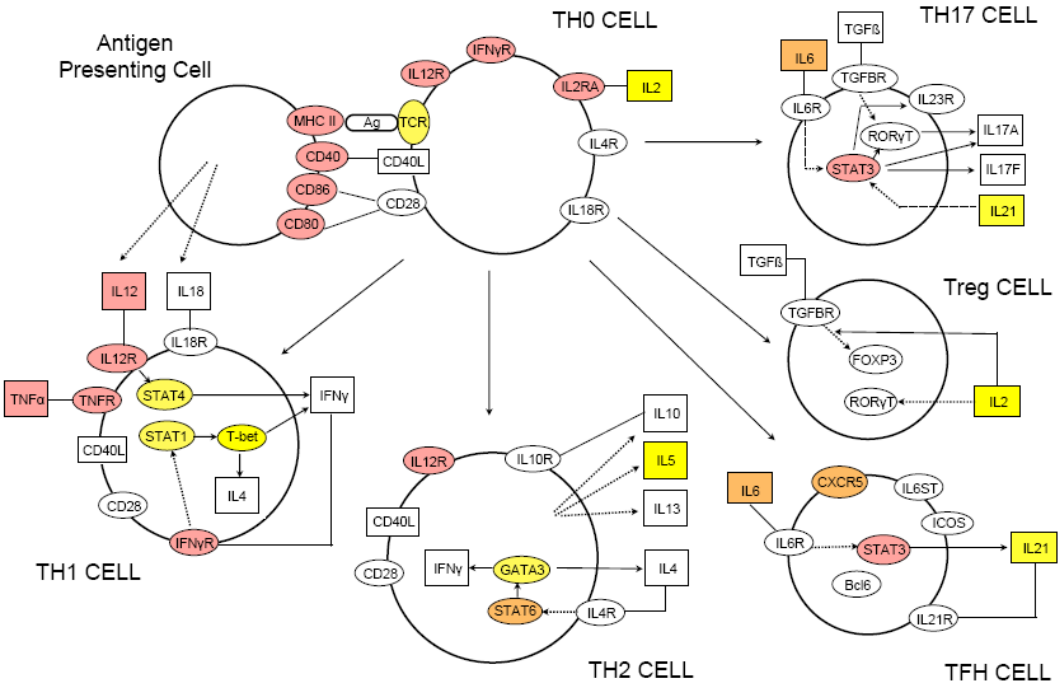
A partire dal 2006, sono stati pubblicati nove studi GWAS che hanno identificato 57 loci che inequivocabilmente contribuiscono alla biologia della SM, riassunti in tabella 1, dimostrando che nel complesso le varianti che esercitano individualmente un modesto effetto genetico, giocano un ruolo chiave nella malattia suscettibilità alla malattia.

Lo studio più recente risale ad Agosto 2011, in cui il Consorzio Internazionale della Sclerosi Multipla (IMSC) ha confermato 23 dei 26 loci già noti, 29 nuovi loci significativi a livello genome wide (*P-value* < $5x10^{-8}$) e 5 nuovi loci con forte evidenza di associazione (*P-value* < $5x10^{-7}$) [25].

In particolare, allo stato dell'arte sono stati descritti diversi geni che codificano per molecole che possiedono una particolare rilevanza immunologica quali; citochine (*CXCR5*, *IL2RA*, *IL7R*, *IL7*, *IL12RB1*, *IL22RA2*, *IL12A*, *IL12B*, *IRF8*, *TNFRSF1A*, *TNFRSF14*, *TNFSF14*), molecole co-stimolatorie (*CD37*, *CD40*, *CD58*, *CD80*, *CD86*, *CLECL1*) e molecole implicate nella trasduzione del segnale (*CBLB*, *GPR65*, *MALT1*, *RGS1*, *STAT3*, *TAGAP*, *TYK2*). Sono stati, altresì, descritte molecole correlate al metabolismo della Vitamina D (*CYP27B1*, *CYP24A1*) e alle terapie per la SM, quali il Natalizumab (*VCAM1*) e il Daclizumab (*IL2RA*) e solo due geni con un ruolo nella neurodegenerazione assonale (*GALC*, *KIF21B*).

Analisi Gene Ontology dimostrano che il 30% delle varianti di suscettibilità alla malattia sono localizzate in geni del sistema immunitario, soprattutto geni coinvolti nel percorso di differenziazione delle cellule T helper, confermando il carattere autoimmune della sclerosi multipla, come indicato dalla figura 8.

Figura 8. Rappresentazione del pathway di differenziazione delle cellule T helper



In figura 8, le etichette alfanumeriche rappresentano i geni coinvolti nel pathway di differenziazione delle cellule T. Sono stati colorati tutti i geni per i quali sono state riportate associazioni con SNPs all'interno dello stesso gene o in prossimità di esso. I diversi colori rappresentano il diverso grado di associazione, comunque superiore a P value < 1x 10 $^{-3}$.

Ciascuno di questi geni contribuisce solo in minima parte al rischio totale di sviluppare la malattia (Odds Ratio, OR ~ 1.2) ed una larga fetta di ereditabilità della SM (circa l'80%) rimane ancora inspiegata. Ciò significa che probabilmente molte altre varianti alleliche, verosimilmente definite da varianti a bassa frequenza e varianti rare (MAF < 5%), stanno contribuendo in maniera importante all'eziologia della SM. Si tratta probabilmente di ulteriori 100-200 geni, ciascuno dei quali potrebbe contribuire in minima parte al rischio di malattia (OR = 1.1-1.3).

Tabella 1. Riepilogo dei GWAS pubblicati sulla SM

Studio	**Disegno sperimentale**	**Popolazione**	**Numero di campioni***	**Numero di SNPs**	***Loci o geni descritti***
Wellcome Trust Case–Control Consortium (2007)	Caso-controllo	UK	1000 casi e 1500 controlli	14,436	*IL7R*
International Multiple Sclerosis Genetics Consortium (2007)	Famiglie	US e UK	931 famiglie trios	334,923	*HLA, IL2R, IL7R, CLEC16, CD58, EVI5, TYK2*
Comabella et al. (2008)	Caso-controllo pooled	Spagna	242 casi e 242 controlli	500,000	*HLA, 13q31.3*
Gene Associations in Multiple Sclerosis Consortium (2009)	Caso-controllo	US, Paesi Baasi e Svizzera	978 casi e 883 controlli	551,642	*HLA, GPC5, PARK2, PDZRN4, CSMD1*
Australia and New Zealand Genetics in Multiple Sclerosis Consortium (2009)	Caso-controllo	Australia e Nuova Zelanda	1618 casi e 3413 controlli	303,431	*HLA, METTL1, CD40*
De Jager et al.(2009)	Meta analisi e caso-controllo	US, UK, Paesi Bassi, Svizzera	2624 casi e 7220 controlli	2,557,248 (imputati)	*TNFRSF1A, IRF8, CD6, RGS1*
Jakkula et al. (2010)	Caso-controllo	Finlandia	68 casi e 136 controlli	297,343	*STAT3*

Sanna et al.(2010)	Caso-controllo	Sardegna	882 casi e 872 controlli	6,600,000 (genotipizzati e imputati)	*HLA, CBLB*
International Multiple Sclerosis Genetics Consortium and the Wellcome Trust Case Control Consortium 2(2011)	Caso-controllo	Discendenti Europei di 23 paesi	9772 casi e 17376 controlli	441,547	*VCAM1, rs12466022, PLEK, MERTK, SP140, EOMES, rs669607, CD86, IL12B,BACH2, THEMIS, MYB, IL22RA2, TAGAP, ZNF767, MYC, PVT1, HHEX, CLECL1, ZFP36L1, BATF, GALC, MALT1, TNFSF14, MPV17L2, DKKL1, CYP24A, MAPK1, SCO2*

Tutti gli studi includono una fase di replicazione. Numeri della fase iniziale.*

Anche se, ad oggi, sono stati compiuti notevoli progressi nella comprensione delle basi ezio-patogenetiche della SM, le risposte fornite da questi studi non sono sufficienti a risolvere gran parte della complessità delle patologia.

2.3 Difficoltà nella dissezione della basi genetiche

La dissezione dei fattori genetici che contribuiscono al rischio di SM è complicata da diversi fattori. La principale difficoltà è rappresentata dall'estrema eterogeneità della patologia sclerosi multipla, descritta nei paragrafi precedenti, e dalla possibile presenza di fenocopie (casi non genetici) che complica estremamente l'analisi. La strategia d'elezione richiede un'accurata selezione dei casi, concentrandosi solo su pazienti fenotipicamente ben caratterizzati e selezionando sottogruppi maggiormente rappresentati, quali la forma recidivante-remittente.
Tuttavia, la maggior parte delle difficoltà si riferiscono invece, al quadro più ampio della dissezione genetica dei tratti complessi in generale.

Negli studi di associazione di tutto il genoma, il primo problema riguarda la dimensione dell'effetto genetico, che è determinato dalla differenza di frequenza delle varianti causali nei pazienti e nei controlli. La maggior parte delle varianti di suscettibilità contribuisce alle malattie complesse con piccoli effetti individuali, quindi é necessario disporre di una casistica caso-controllo di grandi dimensioni, capace di fornire un'adeguata potenza statistica allo studio.
Analisi di potere statistico dei GWAS indicano che per raggiungere l'80% del potere di svelare una variante con OR 1.5 sono necessari circa 1,000 casi e 1,000 controlli, ma per identificare varianti con effetti genetici di minori dimensioni (OR < 1.5) si rendono necessari oltre 5,000 casi e 5,000 controlli e, per essere identificabile, la variante non deve essere né troppo frequente, né troppo rara (frequenza ottimale > 5%). Varianti genetiche moderatamente rare possono essere identificate se contribuiscono alla malattia con un grande effetto (rischio relativo > 2.0)[26].

Un altro limite dei GWAS nello studio delle basi genetiche della SM, riguarda fondamentalmente limitazioni tecniche, ovvero la capacità di estrapolare la massima quantità d'informazioni genetiche dai campioni analizzati, attraverso l'utilizzo di mappe di marcatori prestabilite dai chip commerciali.
Infatti, i chip commerciali finora in uso, interrogano prevalentemente varianti comuni (MAF > 5%) tendenzialmente ubiquitarie e non considerano varianti rare e varianti fondatrici, ovvero varianti frequenti in alcune popolazioni ma rare o assenti in altre. Le mappe di marcatori sono state sviluppate prevalentemente dall'integrazione delle informazioni del progetto Genoma Umano e del progetto HapMap, includendo anche importanti informazioni sull'architettura del Genoma Umano, in termini di correlazione allelica (*linkage disequilibrium*, LD).
L'LD, rappresentato dal parametro D' o r^2, è definito come correlazione allelica non casuale a due o più loci.
Esso deriva dal fatto che alleli vicini vengono più facilmente coereditati in blocchi di aplotipi risultando, quindi, associati all'interno di una popolazione. Poiché la probabilità di ricombinazione fra due SNPs aumenta con l'aumentare della distanza fisica fra i due, di norma il grado di associazione fra SNPs diminuisce progressivamente con la distanza [27].
Il genoma umano è costituito da regioni cromosomiche, dette blocchi di aplotipi, all'interno delle quali gli SNPs hanno un alto grado di associazione. Questo significa

che, all'interno di ciascun blocco, la diversità di una popolazione umana è rappresentata solo da pochi aplotipi [28].

La conoscenza della struttura del LD e della struttura in blocchi del genoma, rappresentata come una forza che tiene unite varianti alleliche a differenti loci, permette la selezione di TagSNPs, ovvero SNPs rappresentativi di interi blocchi di aplotipi.

I chip commerciali sono stati disegnati per fornire un'eccellente copertura degli SNPs comuni, mediante caratterizzazione genotipica di TagSNPs, *proxies* per varianti causali comuni, che non catturano, se non in minima parte, le varianti rare (MAF <5%), e non valutano direttamente il contributo di corti polimorfismi quali inserzioni o delezioni (*indels*). Infatti, sono stati ottimizzati per aumentare la possibilità di trovare varianti comuni di malattia, con piccoli rischi relativi, e varianti rare con grandi rischi relativi, ma non esiste la possibilità pratica di individuare varianti rare, con modesti effetti genetici.

Attualmente, sono in corso notevoli sforzi per risolvere i punti di criticità di queste mappe, e la prossima generazione di chip comprenderà varianti a bassa frequenza e rare, svelate con il risequenziamento *low-pass* (sequenziamento a bassa copertura) di migliaia di individui attraverso sequenziatori di nuova generazione (Next Generation Sequencing, NGS) nell'ambito del progetto internazione 1000 Genomes, 10,000 Genomes UK e del progetto di risequenziamento del genoma Sardo, oggetto di questa tesi.

Un ulteriore complessità dell'analisi genetica della SM, deriva da un possibile incompleto LD e da differenze di frequenza, fra i marcatori genotipizzati e gli alleli della variante di malattia, che è causa dell'indebolimento del segnale rilevato. Anche per superare questo livello di difficoltà è richiesto un campione molto ampio che possa mitigare tale problematica.

La necessità di analizzare casistiche di grandi dimensioni nei GWAS è anche dettata dal numero elevato di test che vanno eseguiti e per i quali è stato proposto un livello di significatività corrispondente ad un *P-value* = $1x10^{-8}$ per ridurre il tasso di falsi positivi, anche se osservazioni più recenti indicano accettabili soglie meno stringenti (*P-value* $5x10^{-7}$).

3 SCOPO DELLA RICERCA E PRESUPPOSTI DELLO STUDIO

L'arduo compito d'identificare la componente genetica delle malattie non mendeliane è, indubbiamente, una meta fondamentale della genetica umana.
Il progetto di ricerca, oggetto di questa tesi, è volto alla dissezione delle basi genetiche della sclerosi multipla in Sardegna, per la quale è stata raccolta un ampia casistica che fornisce un adeguato potere statistico negli studi di associazione su tutto il genoma.
La condizione d'interesse è stata esaminata attraverso un'attenta caratterizzazione fenotipica ed anamnestica, e generazione di dati ad alta risoluzione, che includono caratterizzazione genotipica attraverso chip commerciali (Affymetrix ed Illumina), e ri-sequenziamento *low-pass* dell'intero genoma (a bassa copertura, 3-4 x), tramite sequenziatori di nuova generazione, in un sub-gruppo di individui sardi e di origine sarda.
Il progetto è stato articolato in due principali fasi; una fase preliminare in cui abbiamo testato per associazione circa 6.6 milioni di SNPs (direttamente genotipizzati con chip Affymetrix 6.0 o imputati con pannello di referenza HAPMAP CEU, TSI e 1000 Genomes) in 882 casi di SM e 872 controlli sardi e di origine sarda ed una seconda fase, maggiormente complicata, chiamata di seguito GWAS#2, in cui abbiamo testato per associazione circa 15 milioni di marcatori, direttamente genotipizzati con Affymetrix 6.0 e Illumina 1 M Duo, e imputati con il pannello di referenza 1,000 Genomes e con il pannello di referenza sardo, in 2.280 casi e 1.922 controlli.
Quest'ultimo pannello è stato costituito all'interno del medesimo progetto a partire dal ri-sequenziamento di 505 individui sardi.
Il progetto ha consentito di descrivere l'associazione del gene *CBLB* (Ubiquitin-protein ligase *CBL-B*) con l'aumentato rischio di SM [29].

I presupposti del nostro studio si basano sulla considerazione che, come già descritto nei capitoli precedenti, la prima generazione di studi GWAS ha apportato un notevole avanzamento nelle conoscenze delle basi genetiche della SM, ma nel contesto ha spiegato solo il 20% dell'ereditarietà della malattia.

Questi studi hanno usufruito dei chip di genotipizzazione commerciali (Affymetrix ed Illumina), che hanno utilizzato mappe di marcatori adeguate per raggiungere un eccellente copertura degli SNPs comuni.
Tali studi hanno, infatti, identificato prevalentemente varianti comuni con bassa e media penetranza e con basso o medio rischio genetico OR tra 1.15 a 1.3.
La maggior parte dell'ereditabilità mancante può essere spiegata da varianti rare con un maggiore effetto, e da varianti strutturali, entrambe scarsamente rappresentate negli array attualmente disponibili [30]. Le varianti rare sono qui definite come polimorfismi a bassa frequenza dell'allele minore (MAF) di 0,5% - 5%.
Recenti studi hanno dimostrato che le varianti rare hanno un ruolo importante nell'eziologia dei tratti complessi e si suppone che la loro identificazione avrà un impatto notevole sulla valutazione del rischio, prevenzione, diagnosi e trattamento delle malattie, a causa del loro maggior effetto sul fenotipo [31]. In letteratura si trovano già esempi paradigmatici del coinvolgimento e dell'importanza delle varianti rare nell'eziologia delle malattie complesse, nonché della promettente linea di studio che prevede l'integrazione degli studi di associazione di tutto il genoma con dati di risequenziamento [32-33].

3.1 Caratteristiche della popolazione sarda.

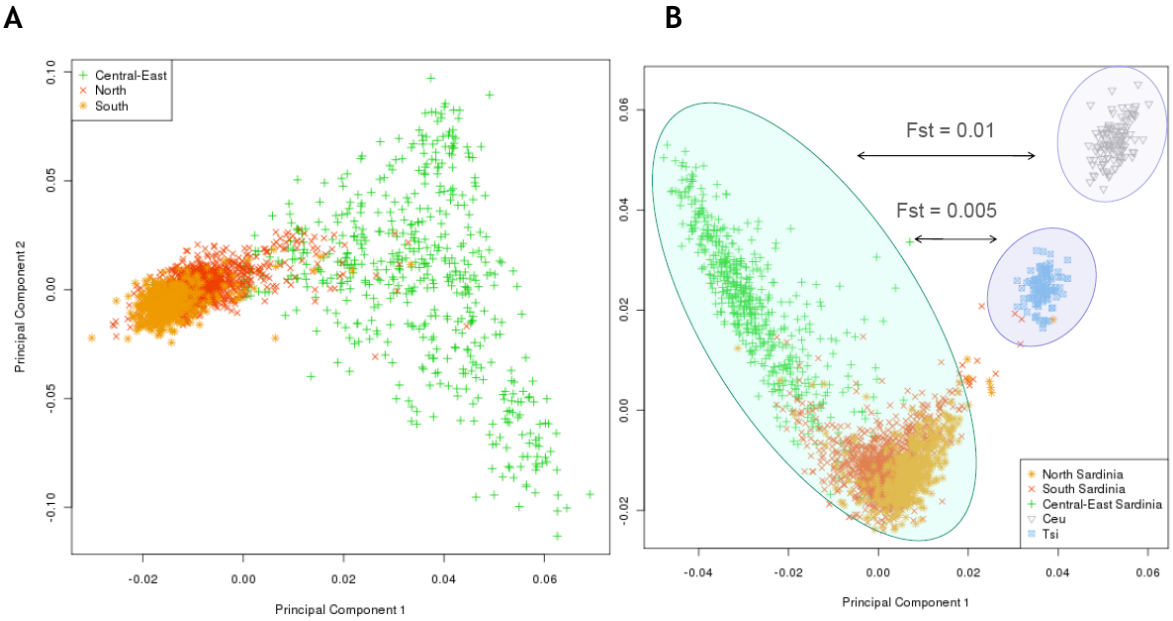
Come abbiamo visto, uno dei limiti dei GWAS è rappresentato dall'assenza, all'interno delle mappe utilizzate, di varianti a bassa frequenza, varianti rare e fondatrici. E' verosimile che la misura, con la quale tali varianti contribuiscono alla predisposizione alle malattie, spieghi la grande proporzione di rischio genetico non ancora mappato.
Ciò assume particolare rilievo quando la popolazione in esame è quella sarda, nella quale il numero di varianti fondatrici è atteso essere più alto che in altre popolazioni.

La Sardegna con i suoi circa 1,600,000 abitanti, è infatti, definita come un macro isolato genetico, con un'elevata frequenza di patologie autoimmuni tra le quali la sclerosi multipla.

La sua attuale popolazione mostra l'intervento di diversi effetti fondatori che hanno determinato una variabilità genetica omogenea nelle differenti macro-aeree dell'isola.

L'assenza di stratificazione e l'omogeneità genetica è confermata da analisi delle componenti principali che utilizzano sia dati aplotipici del cromosoma Y e del locus HLA DRB1-DQA1-DQB1 [34] che da dati, più recenti, ad alta risoluzione di caratterizzazione genotipica, attraverso chip Affymetrix 6.0 in un ampia casistica di 2,615 individui non imparentati, reclutati da diverse aree della Sardegna (figura 9).

Figura 9. Analisi delle componenti principali



La figura 9 mostra l'analisi delle compenti principali, effettuata utilizzando 661,238 SNPs genotipizzati in 2,615 individui non imparentati reclutati da diverse sub-regioni dell'isola. A) mostra l'analisi tra le diverse sub-regioni della Sardegna. B) mostra il confronto della popolazione sarda con popolazioni HapMap di origine europea (CEU) e italiana (TSI).

Gli effetti fondatori hanno determinato un aumento nelle frequenze alleliche e aplotipiche di varianti rare, le cosiddette varianti fondatrici, ovvero varianti molto frequenti in Sardegna e molto rare in altre popolazioni.

Quindi, seppur in un contesto genetico europeo, vi sono alleli e aplotipi piuttosto rari o assenti altrove, come DRB1*1501-DQB1*0602, DRB1*1601-DQB1*0502 e l'aplotipo esteso

DRB1*0301-DQA1*0501-DQB1*0201-B18-A30, molto comuni fuori dalla Sardegna e molto rari nell'isola [35].

Altri esempi sono rappresentati sia dalla frequenza elevata delle mutazioni eziologiche per la Beta Talassemia e la Malattia di Wilson ma anche dalla frequenza di marcatori genetici come l'M26 (marcatore del cromosoma Y) che si ritrova nel 36% dei maschi sardi mentre e raro assente altrove. La popolazione mostra altresì, un'elevata variabilità interindividuale, con la concomitante presenza di varianti genetiche presenti sia nell'est che nell'ovest europeo.

Tutte queste caratteristiche, la rendono quindi una popolazione ideale sia per estensivi studi di catalogazione della variabilità umana ma anche per la dissezione genetica della SM che preveda l'impiego di studi di associazione su tutto il genoma.
Infatti, alcune delle varianti fondatrici sarde potrebbero svolgere un ruolo rilevante come fattori di rischio sia per la SM che per altre malattie comuni in Sardegna.

Dunque, questo lavoro riveste una duplice importanza; creare un catalogo della variabilità sarda che, per le caratteristiche intrinseche di questa popolazione, può rappresentare uno strumento utile anche per studi di associazione in altre popolazioni europee ed aumentare lo spettro delle varianti genetiche che influenzano il rischio di ammalare di sclerosi multipla.
L'identificazione delle varianti di rischio rappresenta un primo importante traguardo nella battaglia contro la SM. Conoscere e comprendere i meccanismi genetici alla base della patologia potrebbe svelare nuove potenziali vie terapeutiche. In generale, è quindi probabile che una migliore comprensione dei fattori e dei meccanismi che svolgono un ruolo nell'attacco del nostro sistema autoimmune verso il sistema nervoso centrale avrà un impatto importante verso trattamenti più efficaci ed anche per la prevenzione della malattia.

4 DISEGNO SPERIMENTALE DELLO STUDIO

Il disegno sperimentale utilizzato è di tipo caso-controllo, e prevede uno studio di associazione su tutto il genoma di migliaia di marcatori SNPs, caratterizzati attraverso piattaforma di genotipizzazione estensiva, Affymetrix e Illumina.
Per incrementare lo spettro delle varianti testate è stato, inoltre, utilizzato un approccio basato sull'inferenza statistica (o imputazione) che ha permesso di ricostruire i genotipi di varianti non direttamente genotipizzate. Tale approccio è basato su algoritmi capaci di ricostruire i genotipi delle varianti mancanti in tutti i campioni dello studio, sulla scorta di pannelli aplotipici di individui caratterizzati per molti più loci, individuando i marcatori in comune e ricercando gli aplotipi simili.

4.1 Studio caso-controllo

Lo studio caso-controllo valuta differenze di frequenza delle varianti, tra la popolazione dei casi di malattia ed una popolazione, omogenea per etnia, di individui sani non imparentati tra loro, né con i casi. Quando varianti alleliche mostrano una frequenza statisticamente maggiore nei casi rispetto ai controlli sono definite varianti associate alla malattia.

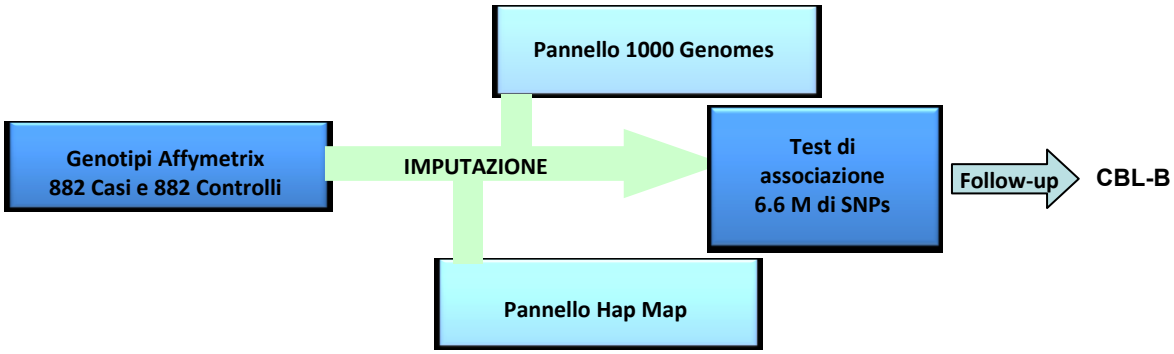
4.1.1 Studio preliminare

Abbiamo condotto uno studio preliminare analizzando 1,754 individui sardi e di origine sarda, definiti da 882 casi di SM e 872 controlli (campioni che hanno superato i filtri di qualità). Tutti i campioni sono stati caratterizzati con chip Affymetrix v 6.0, che include 1.800.000 varianti; più di 906.600 sonde per SNPs e oltre 946.000 sonde per l'identificazione di variazioni nel numero di copie (*CNV - Copy Number Variations*).
I genotipi sono stati attribuiti attraverso l'utilizzo dell'algoritmo Birdseed, considerando i casi e i controlli in un unico cluster. I dati grezzi sono stati filtrati sulla

base dei rigorosi controlli di qualità effettuati sia sugli SNPs che sugli individui, descritti in dettaglio nei materiali e metodi.
Grazie all'impiego del software MACH, sono state enormemente incrementate il numero delle varianti testate per associazione con la SM. A tale scopo sono stati utilizzati i pannelli di referenza HapMap fase II di origine europea (CEU), HapMap fase III CEU e toscani (TSI) e 1000 Genomes che hanno permesso di testare per associazione, un totale di 6,607,266 direttamente genotipizzate e imputate.

Nel *follow-up*, eseguito con metodica Taqman, abbiamo replicato l'associazione della variante rs9657904, in 1,775 casi e 2,005 controlli sardi e 1,441 casi e 1,465 controlli Italiani [36]. L'analisi combinata ha mostrato un *P-value* = 3.3×10^{-13} (figura 10).

Figura 10. Schema studio preliminare



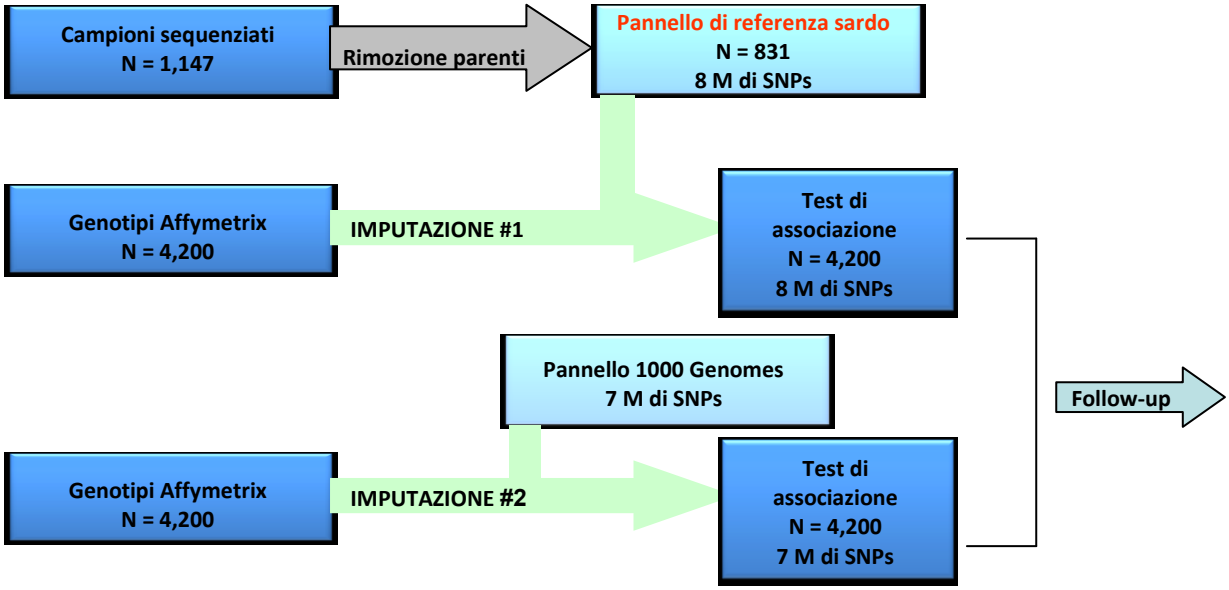
4.1.2 Seconda fase di studio (GWAS#2)

Per incrementare il potere statistico dell'analisi, è stata ampliata la casistica in esame. Sono stati tipizzati con chip Affymetrix v 6.0 un totale di 4,200 individui sardi e di origine sarda, definiti da 2,280 casi SM e 1,922 controlli che hanno superato i filtri di qualità. Per incrementare il numero di varianti testate e analizzare anche varianti rare (frequenza allelica < 5%), un sub-gruppo di 837 individui (279 famiglie trios) sono stati caratterizzati anche con chip 1M Duo Illumina, il quale ha permesso di analizzare quasi 1.200.000 di polimorfismi che includevano un notevole numero di varianti non presenti

nel chip Affymetrix, studiati attraverso la definizione di una super mappa Affymetrix-Illumina.

Al fine di studiare varianti rare e fondatrici della popolazione sarda, ovvero varianti frequenti nella popolazione sarda ma rare o assenti altrove, che potrebbero svolgere un ruolo importante nella patogenesi della SM, sono stati altresì, inclusi dati di sequenza di individui sardi (Pannello di referenza sardo), generati con piattaforme di sequenziamento NGS (figura 11).

Figura 11. Schema disegno sperimentale GWAS#2



4.2 Pannello di referenza sardo

La generazione del pannello aplotipico di referenza sardo prevede l'utilizzo di dati di sequenza *low-pass* di tutto il genoma (a bassa coverage, 3-4x) generati da 3,000 individui sardi e di origine sarda, appartenenti a famiglie trios (genitori e figlio)

Il disegno dello studio del sequenziamento è stato deciso sulla scorta delle analisi e simulazioni eseguite dal gruppo di Abecasis e da nostri collaboratori.

Il sequenziamento *depth* coverage (alta copertura ~30X) è in grado di rilevare tutte le varianti, comuni e rare, presenti negli individui sequenziati, a costi ancora molto elevati che limita, per questo, l'analisi a pochi individui. Tuttavia è possibile combinare i dati di sequenza *low-pass* al fine di generare un catalogo accurato di varianti e costituire un pannello di riferimento che consenta di guidare l'imputazione genotipica nei campioni supplementari, per aumentare la potenza dello studio di associazione.
Il sequenziamento *low-pass* di circa 3,000 individui è stato definita una strategia basilare negli studi di genetica dei tratti complessi, in quanto, consente di svelare putative varianti causali, con frequenza > 0,2%. Un simile potere è ottenibile solo con il sequenziamento *depth coverage* di oltre 2,000 individui, ma richiede circa l'80% dello sforzo in più, in termine di costi e di risorse [37].
La nostra strategia ha previsto la selezione di individui appartenenti a famiglie trios (genitori e figlio) in quanto, nostri dati preliminari indicano una migliore efficienza dell'algoritmo di imputazione nella chiamata genotipica delle varianti ed una migliore ricostruzione delle fasi aplotipiche, nonché un miglioramento delle prestazioni dell'inferenza statistica nel propagare probabilisticamente i dati della sequenza a supplementari individui non sequenziati.

4.3 Mappaggio fine del gene *CBLB*

Al fine d'identificare, nel gene *CBLB*, le varianti primariamente associate all'aumentato rischio di SM, è stato condotto uno studio di mappaggio fine della regione genomica contenete il gene.
Sono stati utilizzati due pannelli di referenza per l'imputazione di varianti non tipizzate nella regione *CBLB* (chr3:105,377,110-105,587,887 ± 2Mb). Un primo pannello generato da dati di sequenza di 280 individui di origine Europea del progetto 1,000 Genomes, ed il pannello sequenze sarde, generato da dati di sequenza di 154 individui sardi, contenenti 21,255 e 16,451 varianti rispettivamente.
Inoltre, sono stati sequenziati 93 casi SM, attraverso metodica Sanger, per tutti gli esoni, le regioni introne-esone e la regione promotrice del gene *CBLB*. Sono state identificate 41 varianti di cui 8 codificanti e 13 non descritte. I 93 pazienti sequenziati fanno parte del sub-gruppo caratterizzato con chip Affymetrix v 6.0 e 1M Duo Illumina.

Tutte le varianti sono state imputate e testate per associazione nell'intera casistica a disposizione.

5 MATERIALI E METODI

In questa sezione verranno descritte tutte le metodologie sperimentali utilizzate nella prima e nella seconda fase del progetto GWAS, incluse le metodiche e gli approcci utilizzati per la catalogazione della variabilità genetica della popolazione sarda e del mappaggio fine del gene *CBLB*.

5.1 Descrizione della casistica

Per questo studio abbiamo reclutato un totale di 4,200 individui, suddivisi in 2,280 pazienti SM e 1,920 controlli sani. Tutta la casistica selezionata ha compilato una scheda con informazioni anagrafiche, anamnestiche e biometriche ed ha, inoltre, firmato il consenso informato, approvato nelle rispettive ASL di provenienza.

6.1.1 Selezione del campione caso-controllo per GWAS

Nella coorte di studio GWAS sono stati arruolati 4,200 individui sardi e di origine sarda, ovvero con almeno tre linee parentali native della Sardegna. Sono stati selezionati esclusivamente individui che non mostravano tra loro relazione, entro il primo grado, di parentela. Quest'analisi è stata primariamente effettuata attraverso una comparazione dei dati anagrafici ed informazioni personali ed a posteriori con una più accurata analisi dei dati genotipici, effettuata attraverso il software Relative Finder (http://www.isogg.org/wiki/Relative_Finder).
La casistica è composta da 2,280 pazienti MS di cui 51 affetti da diabete di tipo 1, altra patologia autoimmune ad elevata prevalenza in Sardegna, e da 1,920 controlli sani.
La raccolta dei campioni di pazienti SM è iniziata oltre 20 anni fa. I campioni provengono da diversi centri clinici dell'isola; Clinica Universitaria di Sassari, Centro Sclerosi Multipla ASL8 di Cagliari ed Azienda Ospedaliera "G. Brotzu". La diagnosi di tutti pazienti è stata effettuata nel rispetto dei criteri Mc Donald ed il rapporto maschi-femmine osservato è pari a 1:2,2.

Tutti campioni sono stati caratterizzati con chip Affymetrix 6.0 e solo un sub-gruppo di 837 campioni (279 trios completi) è stato ulteriormente genotipizzato con chip Illumina 1 M Duo.
La raccolta degli individui sani (controlli) è stata effettuata presso diversi centri trasfusionali dell'isola; Centro Trasfusionale dell'Azienda Ospedaliera "G. Brotzu", Centro trasfusionale di Sassari e Centro Trasfusionale ASL6 di San Gavino Monreale. Il 75% dei controlli utilizzati nello studio sono maschi, che riflette la percentuale degli uomini rispetto alle donne, dei donatori sardi.

Tabella 2. Caratteristiche demografiche dei pazienti e controlli

	Pazienti SM	%	Pazienti SM-T1D	%	Controlli sani	%
Maschi	681	31	13	0.25	1,441	75
Femmine	1,548	69	38	0.75	479	25
Totale	2,229		51		1,920	

5.1.2 Selezione del campione caso-controllo per sequenziamento low-pass

Nella coorte del progetto di sequenziamento a bassa copertura,"*low-sequencing*" sono stati reclutati 1,700 campioni sardi e di origine sarda provenienti da tutte le provincie della Sardegna, che consistono in un sub-gruppo di individui caratterizzati geneticamente con chip Affymetrix v 6.0. Sono stati selezionati trios d'individui appartenenti a famiglie (genitori e figlio) in quanto analisi preliminari hanno mostrato che gli algoritmi utilizzati, di seguito descritti in maniera specifica, funzionano meglio utilizzando famiglie rispetto al sequenziamento di individui non correlati.

5.2 Estrazione del DNA

I campioni di DNA utilizzati sono stati ottenuti da sangue periferico, trattato con l'anticoagulante acido etilendiamminotetracetico (EDTA), attraverso la classica metodica dell'estrazione salina (*salting-out*).

Questa metodica si basa sul processo di solubilizzazione delle proteine, la quale è dipendente da caratteristiche fisico-chimiche, dalla temperatura, dal pH e dalla concentrazione salina della soluzione. Infatti, ad alte concentrazioni di sali, la solubilità delle proteine diminuisce, causandone la loro precipitazione.
Il metodo prevede l'isolamento della cellule nucleate, presenti nel campione di sangue, dopo aver rimosso per lisi i globuli rossi. I leucociti isolati, vengono trattati con un tampone di lisi (Buffer A, SDS e proteinasi K) allo scopo di estrarre gli acidi nucleici e degradare le proteine presenti, che vengono allontanate mediante precipitazione con i sali (NaCl soprasaturo). Infine, mediante trattamento con isopropanolo si ottiene la precipitazione del DNA, il quale viene recuperato sotto forma di nubecola e risospeso in una soluzione di contenete 1mM Tris-HCl, 0.1mM EDTA 1M (TE 1:0.1).

6.3 Controllo di qualità del DNA

Il DNA genomico estratto è stato sottoposto a controlli di qualità che prevedevano una valutazione visiva dopo corsa elettroforetica su gel d'agarosio al 1%, stima della concentrazione e valutazione della purezza (rapporto OD260/280), con analisi spettrofotometrica al Nanodrop 1000.
Sono state preparate aliquote alla concentrazione di 50 ng/µl per un totale di 5 µM di tutti i campioni selezionati per la genotipizzazione con piattaforma Affymetrix, Illumina e sequenziamento. Per la genotipizzazione nel *follow-up,* con metodica TaqMan, sono invece, state preparate ed utilizzate aliquote alla concentrazione di 5 ng/µl.

5.4 Genotipizzazione Affymetrix

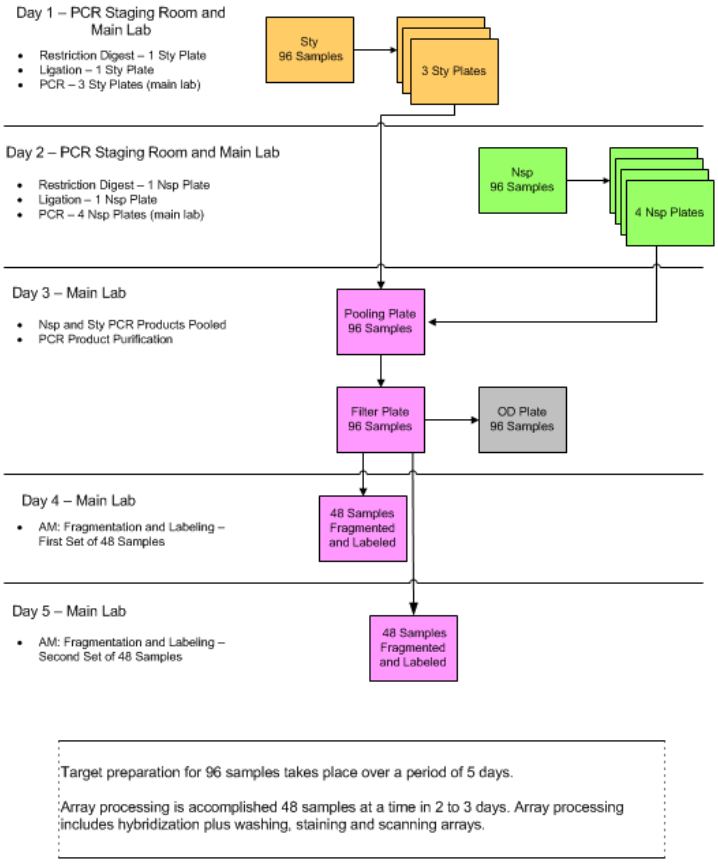
La genotipizzazione mediante il chip Affymetrix v 6.0 include 1.800.000 varianti; 906.600 sonde per SNPs e 946.000 sonde per l'identificazione di variazioni CNV. La genotpizzazione mediante piattaforma Affymetrix è stata eseguita secondo protocollo della ditta produttrice, di seguito brevemente descritto.
La quantità di 500 ng di DNA genomico è stata digerita, in maniera casuale su tutto genoma, utilizzando separatamente gli enzimi di restrizione Nsp I e Sty I. Alle estremità dei frammenti prodotti, è stata aggiunta, tramite una ligasi, una breve sequenza

nucleotidica, a sequenza nota, detti adattatori, che riconoscono le 4 bp coesive del primer. Un primer generico complementare alla sequenza dell'adattatore viene, di seguito, utilizzato per l'amplificazione.

E' stato utilizzato il seguente programma di amplificazione, come da protocollo: 94° C 3 min di denaturazione, 94° C 30 sec, 60° C 45 sec, 68° C 15 sec per 30 volte, 68° C per 7 min.

Le condizioni di PCR sono state ottimizzate per amplificare preferenzialmente frammenti da 200 a 1.100 bp. I prodotti di PCR corrispondenti a ciascuno dei due enzimi di restrizione sono stati uniti, purificati con biglie di polistirene (Invitrogen) e quantificati con spettrofotometro (Spectramax, Molecular Devices). Il DNA amplificato è stato in seguito frammentato, marcato ed ibridato al chip. Dopo una serie di lavaggi, eseguiti nella stazione fluidica, è stata effettuata la lettura mediante lo *scanner* Gene Chip (figura 12).

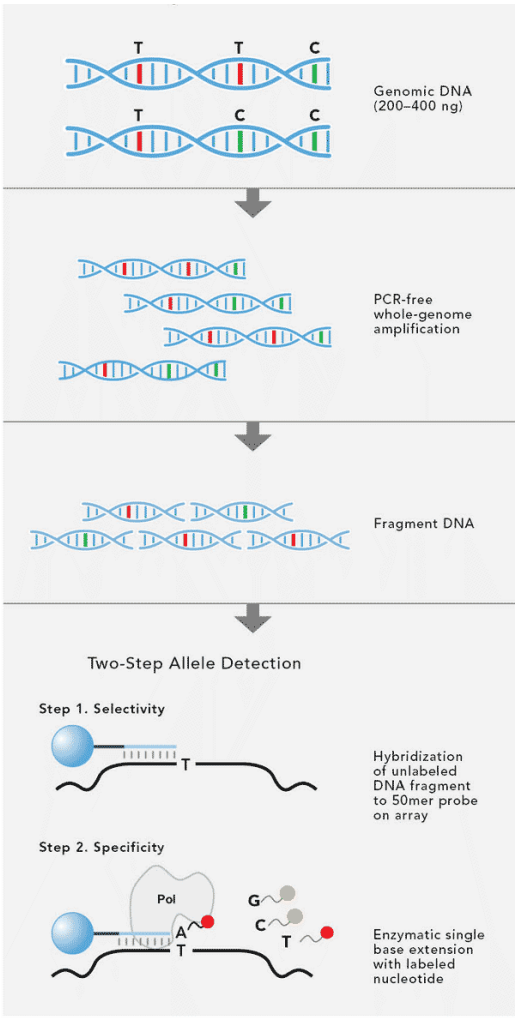
Figura 12. Workflow protocollo Affymetrix v 6.0



5.5 Genotipizzazione Illumina

La genotipizzazione mediante il chip 1M Duo include 1.200.000 varianti; più di 900.000 sonde per SNPs e più di 200.000 sonde per l'identificazione di CNV.
In breve, 400 ng di DNA è stato amplificato, tramite amplificazione isotermica di tutto genoma (*Whole Genome Amplification*, WGA), e in seguito frammentato. I frammenti di DNA si ibridizzano a oligonucleotidi di 50pb, specifici ad ogni locus, che si attaccano covalentemente a una delle 1.100.000 biglie immobilizzate sulla superficie dello chip. Il DNA in eccesso è stato eliminato durante il lavaggio. Il passaggio successivo è stata l'amplificazione isoterica con l'estensione di una singola base per ogni oligomero e quindi per ogni variante. Il beadchip è stato scannerizzato con Illumina iScan che utilizza un laser per eccitare la base estesa che essendo marcata con un fluoroforo emette fluorescenza e lo scanner raccogliere l'immagine della luce prodotta.
Nella discriminazione allelica vengono utilizzati due oligonucleotidi, di 50 pb, per ciascun SNP, che sono specifici per ogni allele di ciascun sito (Allele Specific Oligo, ASO). La fase dell'ibridazione delle sonde, rende questa metodica altamente specifica. A seguito dell'ibridazione, una reazione enzimatica di PCR estende di una singola la base il primers specifico, annilato alla sequenza target. I nucleotidi nella soluzione di PCR sono marcati con differenti fluorocromi che permettono la discriminazione in maniera altamente sensibile della variante wild-type dalla variante mutata di ciascun SNP (figura 13).

Figura 13. Workflow genotipizzazione 1M Duo Illumina



5.6 Sequenziamento Next Generation di DNA genomico

Le sequenze sono state prodotte attraverso sequenziatori NGS; Genome Analyzer IIx e Hi-Seq 2000 (Illumina). Sono state dapprima preparate le librerie di DNA, in accordo con il protocollo Illumina [38].

Brevemente, il DNA genomico é stato frammentato in maniera casuale, mediante un processo di nebulizzazione o sonicazione (Covaris S, Applied Biosystems), in frammenti

sotto 800 paia di basi. Successivamente le estremità dei frammenti, in 3'e 5', sono state riparate e fosforilate. I frammenti di DNA riparati sono stati adenilati in 3´ con una DNA polimerasi *Klenow exo-* (NEB) e poi sono stati aggiunti degli adattatori (IDT) con l'impiego di DNA ligase. I prodotti di ligazione, di dimensione compresa tra le 300 e le 400 paia di basi, sono stati selezionati su gel di agarosio 2%, purificati con il Kit Qiagen Gel Extraction e, successivamente, preamplificati mediante PCR, utilizzando dei primers (IDT) compatibili con gli adattatori.

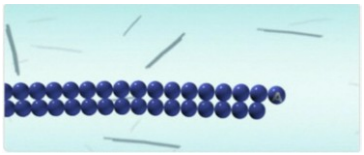
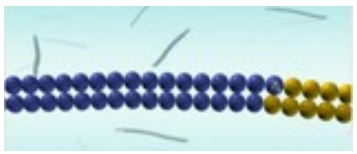
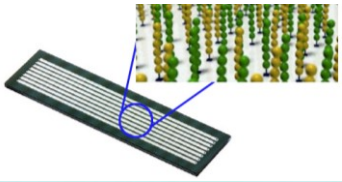
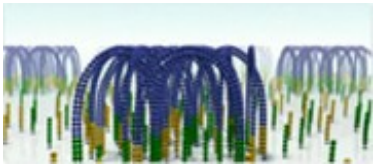
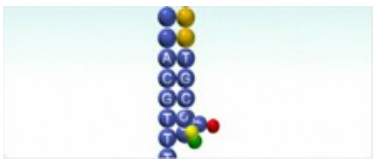
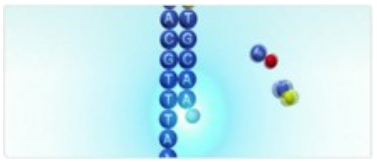
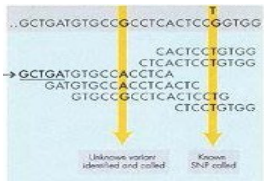
Gli ampliconi ottenuti sono stati purificati con Kit Qiagen Gel Extraction e successivamente è stata valutata la concentrazione e la distribuzione dei frammenti delle librerie mediante corsa su Chip DNA 1000 nel Bioanalyzer 2100 (Agilent Technologies).

Le librerie sono state ibridizzate e amplificate sulla superficie di uno specifico vetrino, detto *flow-cell,* mediante un processo di amplificazione a ponte "*bridge amplification*", formando i *clusters*, quindi sequenziate con il GA*IIx*, in corse *paired-end* da 240 basi (con Paired End Cluster Generation Kit versioni 3,4 e SBS Cycle Sequencing Kit versioni 2, 4, 5 di Illumina), o con l'Hi-Seq 2000 in corse da 202 basi, ottenendo un copertura media di 3-4X.

Il sequenziamento, nella piattaforma Illumina impiega un metodo ciclico di incorporazione, fluorescenza, immagine e lavaggio, fruttando la chimica di particolari dinucleotidi modificati, sviluppati sulla base dei dideossinucleotidi utilizzati nella tecnologia Sanger. Tali nucleotidi sono rappresentati dalla molecola 3'_O_Ozydomethyl, che riproducono dei terminatori reversibili. Durante il sequenziamento è presente un miscela delle quattro basi dinuclotidiche modificate e fluorescinate con quattro diversi fluorocromi che competono tra di loro. A seguito dell'incorporazione della base complementare alla prima base del filamento, il laser eccita la molecola ed il segnale luminoso viene raccolto come immagine per identificare la base appena aggiunta. Nel passaggio successivo viene ripristinato il 3'OH del dinucleotide modificato, appena incorporato, che sarà quindi in grado di accogliere la seconda base. Questo processo si verifica contemporaneamente per tutti i filamenti all'interno della *flow-cell* e le immagini vengono raccolte indipendentemente per ciascuno degli otto canali della *flow-cell* (tabella 3).

L'Hi-Seq 2000, in otto giorni di lavoro, è in grado di produrre circa 200 miliardi di basi. Le risorse computazionali, necessarie per ogni corsa, sono imponenti (30 terabasi) come del resto l'analisi bioinformatica che è prevista, descritta nel capitolo 6.

Tabella 3. Workflow del sequenziamento con piattaforma NGS Illumina

Preparazione della libreria		Il DNA viene frammentato attraverso nebulizzazione o sonicazione . Le estremità 5' e 3' vengono riparate ed adenilate.
		Ai frammenti vengono aggiunti degli adattatori attraverso una DNA ligasi I. Vengono selezionati frammenti di dimensione compresa tra 300 e 400 pb , purificati e pre-amplificati.
Cluster Station		I campioni vengono inseriti nella *flow cell*, vetrino con otto canali che possiede nella sua superficie una densa quantità densa oligonucleotidi.
		I frammenti, attraverso gli adattatori, si legano agli oligonucleotidi presenti nella superficie della *flow-cell*.
		Ogni frammento si ripiega creando un legame covalente con gli oligo della superficie della *flow-cell*. Ogni frammento è amplificato attraverso un'amplificazione isotermica "*bridge amplification*" con la formazione dei clusters.
Sequenziatore		La reazione di sequenza utilizza quattro nucleotidi modificati e marcati con 4 diversi fluorocromi (terminatori reversibili). I nucleotidi competono tra di loro e verrà incorporata la base complementare alla prima base del filamento.
		Dopo ogni ciclo il laser eccita la molecola fluorescinata ed identifica la base appena aggiunta. Il nucleotide viene ripristinato e sarà in grado di accogliere la seconda base. Il ciclo si ripete fino determinare la sequenza del filamento
Analisi		Allineamento delle corte sequenze rispetto ad un genoma di referenza e identificazioni delle differenze e chiamata degli SNPs.

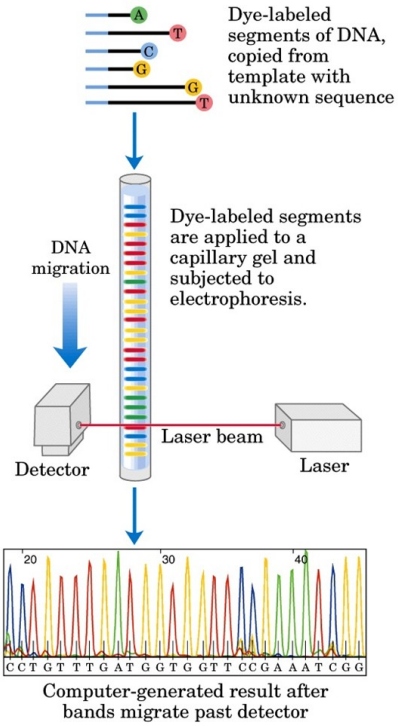
5.7 Sequenziamento Sanger automatizzato

Nel sequenziamento Sanger la sequenza dell'amplicone in esame avviene mediante generazione di frammenti di varie dimensioni grazie ad una interruzione controllata della replicazione enzimatica.

Il metodo Sanger automatizzato si basa sulla sintesi di una reazione di sequenza a partire da un templato di DNA, preventivamente amplificato e purificato. Per questo scopo si utilizza una DNA polimerasi I, che ha funzione di copiare una particolare sequenza di DNA a singolo filamento a partire da inneschi (primers) complementari alla sequenza del DNA target. Alla miscela di reazione vengono aggiunti quattro deossiribonucleotidi trifosfato (dATP, dCTP, dGTP, dTTP), che permettono l'allungamento del filamento; e analoghi 2',3'-dideossi di ogni base (ddATP, ddCTP, ddGTP, ddTTP) marcati con una diversa fluorescenza. L'incorporazione dei dideossi interrompe a diversi livelli la crescita della nuova catena, perché privi del terminale ossidrilico in 3', necessario per formare il successivo legame fosfodiesterico.

In questo modo si producono casualmente frammenti di lunghezza diversa, tutti con un dideossiribonucleotide marcato all'estremità 3'. Il prodotto caricato nel sequenziatore automatico (Applied Biosystem, AbiPrism 3100) viene separato, all'interno di capillari, per via elettroforetica, grazie alla differenza di potenziale applicata agli estremi dei capillari, dallo strumento. Al passaggio delle molecole davanti alla lampada laser, i fluorocromi vengono eccitati con conseguente emissione di fluorescenza che verrà captata e rielaborata dal software con la produzione del classico elettroferogramma (figura 14).

Figura 14. Sequenziamento Sanger Automatizzato



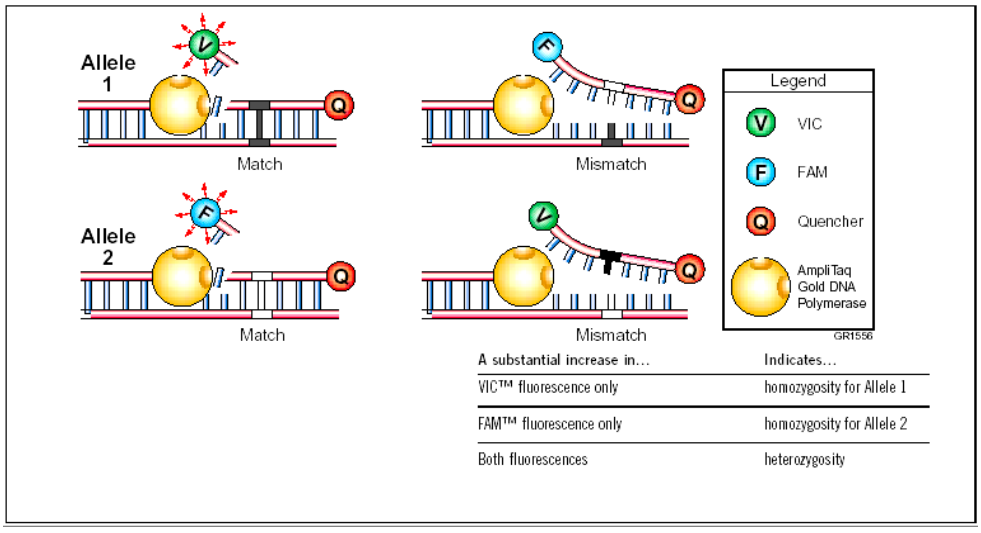
5.8 Genotipizzazione TaqMan

Nel follow-up, la genotipizzazione è stata effettuata tramite saggi di discriminazione allelica con 5' nucleasi [39], attraverso la piattaforma TaqMan 7900HT Real-Time PCR System (Applied Biosystem) con blocco da 384 pozzetti. Questa metodica si propone per efficienza, sensibilità e specificità come una tra i principali metodi per la genotipizzazione di un numero medio-grande di SNPs.

Il metodo impiega degli oligonucleotidi allele-specifici (Allele specific oligonucleotide, ASO) come sonde per l'ibridazione. Il saggio si avvale di primers, forward e reverse, che consentono l'amplificazione della regione contenete il polimorfismo e di due oligonucleotidi (sonde MGB), complementari alla sequenza bersaglio che riconoscono lo SNP, una nella sua variante *wild-type* ed una nella variante mutata. La sonda è caratterizzata dalla presenza in 5' di una molecola "accettore" marcata in VIC o in FAM e in 3' da una molecola "donatore" La sonda complementare all'allele 1 (wild type) sarà marcata con il fluorocromo VIC e quella complementare all'allele 2 (mutato) sarà

marcata con il fluorocromo FAM, al fine di generare un segnale sequenza specifico, consentendo quindi la discriminazione della variazioni nucleotidiche del polimorfismo.
La FRET (Fluorescence Resonance Energy Transfer) è il metodo di rilevazione utilizzato dallo strumento. Il trasferimento di energia avviene solamente nel caso in cui siano soddisfatte due condizioni: lo spettro di emissione della molecola fluorescente "donatore" (*reporter*) deve sovrapporsi con la lunghezza d'onda d'eccitamento della molecola "accettore" (*quencher*) e in secondo luogo, le due molecole devono trovarsi molto vicine fra loro altrimenti il trasferimento di energia cade rapidamente all'aumentare della distanza.
Durante l'amplificazione la Dna polimerasi scalza la molecola "accettore", se e solo se la sonda ha riconosciuto correttamente l'allele complementare dello SNP, che trovandosi libera in soluzione potrà emettere fluorescenza. La rilevazione della sola fluorescenza VIC indicherà l'omozigosi del campione per l'allele 1, della solo fluorescenza FAM, l'omozigosi dell'allele 2 e in caso di rilevazione di entrambe, l'eterozigosi (figura 15).

Figura 15. Rappresentazione del principio di genotipizzazione Taqman



6 ANALISI STATISTICA

6.1 Controlli di qualità sui dati genotipici

Nella piattaforma Illumina i genotipi sono stati assegnati attraverso il "*GenomeStudio Genotyping Module*". Questo processo include la normalizzazione dei dati grezzi, la formazione degli *clusters*, la chiamata dei genotipi e il controllo di qualità di genotipizzazione. L'algoritmo assegna un genotipo ad ogni individuo, effettuando una classificazione a 3 classi mediante l'utilizzo di conoscenza a priori.
Nella piattaforma Affymetrix i genotipi sono stati assegnati considerando i casi e i controlli in un unico cluster, attraverso il "*Birdseed genotyping algorithm*" implementazione degli Affymetrix Power Tools. L'algoritmo assegna un genotipo ad ogni individuo attraverso una classificazione a 3 classi e l'utilizzo di conoscenza a priori. Tale conoscenza a priori è costituita su parametri di un modello gaussiano misto, calcolati sulla base dei genotipi di 270 individui tipizzati nell'ambito del progetto HapMap. L'algoritmo di *Expectation-Maximization* viene applicato ricorsivamente fino alla convergenza tra il modello e i dati osservati. Per le sue caratteristiche, che lo rendono dipendente dai dati che gli si forniscono in ingresso, l'algoritmo deve essere applicato sull'intero set di dati che si intende includere nello studio.
I dati grezzi sono stati di seguito sottoposti ad una serie di stringenti filtri di qualità in base a severi controlli riguardanti sia gli individui che gli SNPs.
Sono stati eliminati tutti gli individui con *Contrast Quality Control* (cQC) < di 0.4 e controllato la media del cQC per piastra, ripetendo la genotipizzazione di tutti quei campioni locati in piastre in cui la media del cQC sia risultata inferiore a 1.7. Il cQC è un indice di qualità del genotipo attribuito. Sono stati eliminati gli individui con tasso di chiamata dei genotipi (*call-rate*) minore del 90%, gli individui in duplicato, con relazione di parentela (software Relative Finder) e con sesso discordante (sesso reale vs sesso assegnato da Affymetrix Power tools, http://www.affymetrix.com/partners_programs/programs/developer/tools/powertool.affx).
Sono stati eliminati gli SNPs che presentavano: frequenza dell'allele minore (Minor Allelic Frequency, MAF) < 1%, call rate < 98% nei soli casi o nei soli controlli, differenze di *call-rate* tra casi e controlli superiore al 5%, deviazione eccessiva dall'equilibrio di

Hardy-Weinberg nei controlli (*P-value* >0.000001), con tasso di discordanza < 0.009 su dupliche. Per il cromosoma X abbiamo anche eliminato gli SNPs per cui il numero di maschi eterozigoti era > del 2% e per cui c'era una differenza eccessiva tra il *P-value* calcolato per i soli maschi e quello calcolato per le sole femmine.

6.2 Analisi delle sequenze NGS

Le analisi dei dati di sequenza prodotte con sequenziatori NGS rappresentano un forte impegno per i ricercatori sia in termini di tempo che di risorse computazionali che devono essere dedicate. Infatti, i dati grezzi delle sequenze sono costituiti da brevi sequenze di 100-200 paia di basi, chiamate reads, che devono essere allineate con un genoma umano di referenza.
Nel nostro studio le corte sequenze sono state allineate attraverso il software Burrows-Wheeler Aligner (BWA), http://bio-bwa.sourceforge.net.
Dai dati totali sono state filtrate le reads duplicate, attraverso il programma Picard, http://picard.sourceforge.net. I duplicati si generano, verosimilmente, durante la fase d'amplificazione (Polymerase Chain Reaction, PCR).
L'introduzione di errori di PCR potrebbero alterare le frequenze alleliche delle varianti, e quindi, diminuire la sensibilità e specificità di rilevamento delle stesse [40].
La qualità delle basi, assegnata dal sequenziatore, è stata ricalibrata considerando diversi fattori confondenti, quali la presenza di regioni ricche in GC , la posizione della reads nella *flowcell*, e verificando la presenza nei database pubblici la presenza di un polimorfismo della stessa base.
La qualità delle basi è stata riassegnata in *Phred Score*: -log10(p), dove p è la probabilità che la base sia errata. Sono state scartate tutte le basi con qualità < 20. Questo processo è stato eseguito dapprima indipendentemente per ciascun campione e successivamente analizzando insieme tutti i campioni sequenziati. Attraverso il software SAMTOOL, http://samtools.sourceforge.net, è stata creata una lista di basi nucleotidiche che rappresentano potenziali polimorfismi e quindi definita la probabilità dei possibili genotipi, salvati nel formato GLF (Genotype Likelihood File) (http://genome.sph.umich.edu/wiki/GLF) [41].

Le basi che hanno mostrato un eccesso di *coverage* rispetto alla distribuzione generale sono state filtrate, poiché verosimilmente si trattava di regioni ripetute ed infine sono stati caratterizzati i genotipi. Quest'ultimo processo è stato eseguito con un programma scritto dai nostri collaboratori, che tiene conto delle *reads* lette in individui con aplotipi simili e delle relazioni familiari degli individui sequenziati. Poiché quest'algoritmo ricostruisce la trasmissione dei cromosomi all'interno delle famiglie, i genotipi risultanti sono sottoforma di aplotipi. La lista degli aplotipi del pannello di referenza sardo è stato salvato in formato standard vcf (http://www.1000genomes.org/wiki/Analysis/vcf4.0).

Prima di procedere all'imputazione statistica, è stato effettuato un ulteriore controllo tra i dati di sequenza generati e i genotipi attribuiti dalla genotipizzazione Affymetrix. Ovvero per ciascun individuo sequenziato, sono stati confrontati tutti gli SNPs, caratterizzati con piattaforma Affymetrix, localizzati nel cromosoma 20 (scelto per la sua ridotta dimensione che consente tempi di analisi più rapidi) con i genotipi estrapolati dai dati di sequenza per i medesimi SNPs. Tale controllo ha permesso di escludere errori e artefatti nel processo di imputazione, dovuti a *cross match* di individui.

Il processo è stato eseguito utilizzando il software VerifyBamID con verifica delle reads mappate nel Bam file (Binary Alignment Map) versus un file di genotipi, inputGenotypes, definito da dati genotipici in strand positivo, formato binario PLINK. La linea di comando utilizzata è la seguente: verifyBamID --reference [reference.fa] --in [inputReads.bam] --bfile [inputGenotypes] –out [outPrefix] –verbose. Per ogni coppia di campioni è stato generato un score di probabilità chiamato P ibd, valore tra 0 e 1. Il valore 1 definisce la coincidenza tra il campione genotipizzato e quello sequenziato, mentre il valore 0 indica l'indipendenza dei campioni. Valori di P ibd uguali a 0.5 indicano un rapporto di parentela tra i campioni analizzati mentre un valore di 0.95 è indice di una probabile contaminazione del campione.

6.3 Imputazione statistica

Imputazione statistica è il termine utilizzato per descrivere il processo di predizione dei genotipi di varianti non direttamente genotipizzate, in un campione definito di

individui. Il termine si riferisce alla situazione generale in cui un pannello di riferimento, costituito da aplotipi di una densa mappa di SNPs, viene usato per ricostruire i genotipi in un campione di individui che sono stati caratterizzati solo per un sottoinsieme di questi SNPs.
Nel nostro studio, per incrementare lo spettro delle varianti testate per associazione, abbiamo utilizzato metodi d'inferenza statistica, usufruendo di pannelli aplotipici di referenza generati da dati di sequenza su tutto il genoma a basso *coverage* .
Tali metodi, consentono di analizzare, in studi di associazione su tutto il genoma, varianti non direttamente genotipizzate, in quanto assenti nelle mappe dei chip commerciali. I metodi di imputazione, per questo motivo, aumentano il potere degli studi di associazione su tutto il genoma di svelare loci di suscettibilità nei tratti in esame.
L'imputazione genotipica combina due o più insiemi di dati e si basa sul confronto degli individui dello studio, genotipizzati per un numero relativamente alto di marcatori genetici, e del pannello aplotipico di referenza utilizzato, che include informazioni genotipiche dettagliate, di un numero molto maggiore di marcatori. L'algoritmo identifica i marcatori comuni tra i due gruppi e gli aplotipi condivisi affinché i genotipi mancanti possano essere compilati in ciascun campione dello studio, copiando gli alleli osservati nell'aplotipo di riferimento corrispondente [42].
Nel nostro studio abbiamo utilizzato il software MACH (http://www.sph.umich.edu/csg/abecasis/MACH/download/).

Sono stati utilizzati due differenti pannelli di referenza di aplotipi sardi, generati da 347 e 831 aplotipi di individui sardi e di origine sarda ed un pannello di referenza generato nel contesto del progetto 1000 Genomes, costituito da aplotipi di 280 individui di origine Europea.

Dapprima è stata convertita la mappa dei dati genetici dalla Build 36 (dbSNP129) alla build 37(dbSNP131), di seguito è stata ricostruita la fase aplotipica dei marcatori genotipizzati con chip commerciali, attraverso l'utilizzo del software MACH. Quindi attraverso l'impiego MiniMAC, un'estensione di MACH, che utilizza gli aplotipi in fase come punti iniziali della catena di Markov (http://genome.sph.umich.edu/wiki/Minimac), sono stati integrati i dati di sequenza dei campioni sardi, in tutta l'intera casistica genotipizzata, considerando le varianti in

comune tra i due insiemi di dati ed utilizzando come principio la ricerca di aplotipi più simili (figura 16). L'imputazione dei dati di sequenza della popolazione sarda sulla super mappa Affymetrix-Illumina è stata condotto con il software Impute2 http://mathgen.stats.ox.ac.uk/impute/impute_v2.html. Impute2 è un software che permette di mettere in fase i genotipi osservati e d'imputare i genotipi mancanti, inoltre può essere utilizzato in analisi d'imputazione che necessitano di combinare due pannelli di referenza contenenti differenti sets di SNPs, come nel nostro caso.

Dopo ciascuna fase d'imputazione è stata valutata la qualità dei genotipi inferiti. Questa fase è necessaria al fine di valutare la qualità dei genotipi inferiti, in particolare per ciascun SNPs per i quali vi è assenza di veri genotipi da mettere a confronto. Il metodo prevede che tutti gli SNPs con bassa qualità d'imputazione vengono filtrati.

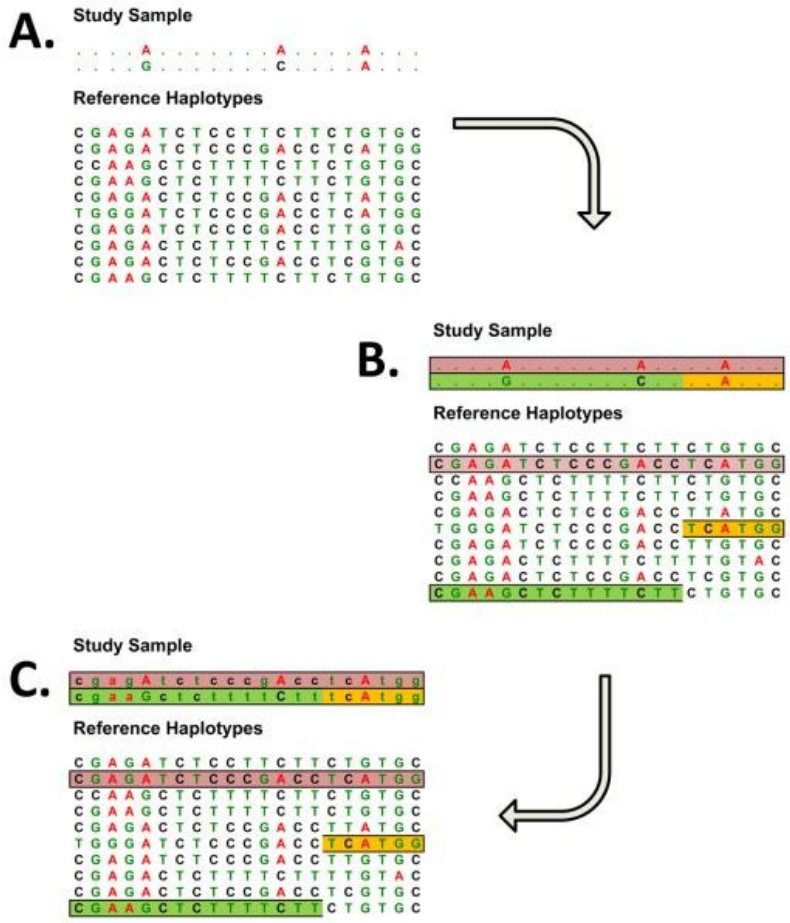
Ci sono diverse metriche con la quale viene valutata la qualità e l'accuratezza dell'imputazione; il parametro RSQR (Russell Square Quality Representatives), che rappresenta il rapporto della varianza media osservata rispetto all'atteso di ciascun SNP ed il parametro r^2 che indica il rapporto tra la varianza osservata del dosaggio allelico e la varianza attesa, in equilibrio di Hardy-Weinberg.

Entrambe le metriche sono dei valori che variano tra 0 e 1, nella quale il valore 1 indica la presunta certezza dei genotipi assegnati ed i valore 0, la totale incertezza circa i genotipi imputati.

E' consigliata l'applicazione della soglia di 0.3, la quale è statisticamente capace di rimuove il 90% degli SNPs inferiti in maniera errata. In genere, si guardano con sospetto gli SNPs con RSQR tra 0.3 e 0.5, con prudenza quelli con RSQR tra 0.5 e 0.8, e con maggiore confidenza quelli con valori al di sopra di 0.8.

In questo studio sono stati esclusi dall'analisi tutti gli SNPs con un RSQR minore di 0.3.

Figura 16. Imputazione di genotipi utilizzando aplotipi di individui non imparentati



La figura 16 illustra le fasi dell'imputazione dei genotipi di varianti non tipizzate. A) Illustra due campioni genotipizzati (Study Sample) ad un modesto numero di varianti ed un pannello di referenza costituito da aplotipi di individui genotipizzati ad un numero superiore di marcatoti. B) Illustra il processo di individuazione di aplotipi simili tra i campioni dello studio e gli individui del pannello di referenza. C) Illustra il risultato dell'imputazione dei genotipi nei campioni di studio.

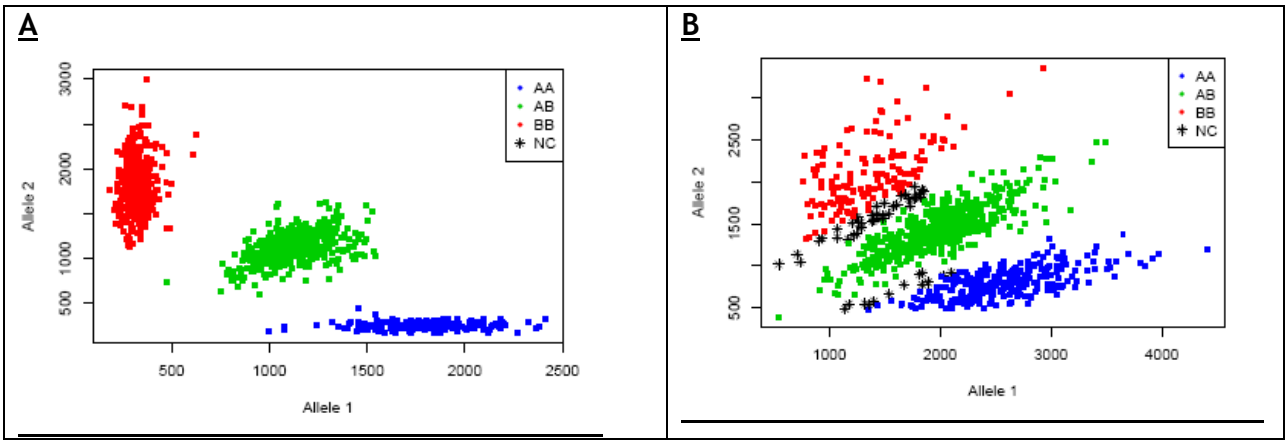
6.4 Test di associazione

Nell'analisi preliminare d'associazione sono stati testati 6,606,267 milioni di SNPs, direttamente genotipizzati con Affymetrix 6.0 ed imputati, in 882 casi e 872 controlli.
Mentre lo studio GWAS#2 ha previsto l'analisi di SNPs direttamente genotipizzati con Affymetrix ed Illumina, ed imputati dal pannello di sequenze sarde e 1,000 Genomes, in 2,280 casi e 1,922 controlli;
In particolare, nello studio GWAS#2 sono stati previsti due *round* di analisi;

1) associazione effettuata per 14,820,141 SNPs in 2,280 affetti MS e 1,922 controlli, di cui 471.724 SNPs direttamente genotipizzati con chip Affymetrix, che hanno superato i controlli di qualità, 7,669,642 e 6,678,775 imputati con il pannello di referenza sardo ed il pannello di referenza 1,000 Genomes, rispettivamente.
2) associazione di 15,563,116 SNPs in 2,280 affetti MS e 1,922 controlli, di cui 1,214,699 SNPs della super mappa Affymetrix-Illumina, che hanno superato i controlli di qualità (923,929 e 291,307 SNPs genotipizzati con chip Illumina ed Affymetrix rispettivamente) e 7,669,642 e 6,678,775 imputati con il pannello di referenza sardo ed il pannello di referenza 1,000 Genomes, rispettivamente.

Abbiamo applicato il test d'associazione (basato sulle tipizzazioni di pazienti versus controlli) sia sui marcatori autosomici che localizzati sul cromosoma X e calcolato il chi quadro con correzione per sottostruttura utilizzando il software Eigenstrat [http://genepath.med.harvard.edu/~reich/Software.htm].
Abbiamo, quindi, valutato visivamente i plot della discriminazione genotipica di ciascun SNP che mostrava un *P-value* $<10^{-5}$, eliminando tutti i marcatori che presentavano un'alterata attribuzione dei genotipi. In figura 17 sono riportati due esempi di plot di discriminazione allelica esaminati, i quali includono un esempio di SNP che ha superarto il controllo di qualità ed un esempio di SNP che è stato eliminato.

Figura 17. Esempi di plot di discriminazione allelica



La figura 17-A. mostra un plot con adeguata discriminazione allelica, mentre la 17-B mostra un esempio di SNPs eliminato dalle analisi per cattiva attribuzione dei genotipi. Per entrambe le figure In ordinata sono espressi i valori di fluorescenza dell'allele 1 ed in ascissa per l'allele 2. I puntini in blu rappresentano gli individui con genotipo omozigote AA, i verdi rappresentano gli individui eterozigoti AB ed i rossi gli omozigoti BB. Nella figura 17-A, i genotipi identici formano dei cluster ben definiti, mentre in 17-B è completamente assente la distinzione dei cluster, che riflette una maggiore o totale imprecisione del dato.

7 RISULTATI E DISCUSSIONE

Il primo risultato raggiunto in questo progetto è stata la capacità di collaborazione dimostrata da parte di diversi enti di ricerca (l'Istituto di Ricerca Genetica e Biomedica - CNR, Università di Sassari ed il Centro di Ricerca, Sviluppo e Studi Superiori in Sardegna) e di diversi centri clinici dell'isola (i centri Sclerosi Multipla dell' Università ed ASL di Cagliari, di Ozieri, ed il centro dell'Università di Sassari), che hanno costituito un consorzio sardo della SM, al fine di perseguire e raggiungere l'obiettivo comune di aumentare le conoscenze sulla biologia della sclerosi multipla, di trovare una spiegazione all'alta prevalenza di questa patologia in Sardegna e di mettere in luce i meccanismi ed i pathways coinvolti nella patologia, che possano aiutare tutta la comunità scientifica a trovare nuove ed efficaci cure nel trattamento e nella prevenzione della SM.

7.1 Risultati del GWAS preliminare e gene *CBLB*

Nell'analisi preliminare di associazione su tutto il genoma sono stati analizzati circa 6.6 milioni di varianti direttamente genotipizzate e imputate con pannelli di referenza HapMap II e III, e 1000 Genomes, in 882 casi e 872 controlli che hanno superato i controlli di qualità descritti nei materiali e metodi.
In dettaglio, sono stati analizzati 575,678 SNPs caratterizzati con chip Affymetrix, 2,533,753 e 1,277,673 SNPs imputati rispettivamente con HapMap II e III e 6,338,706 SNPs imputati con il pannello di referenza generato da un primo rilascio di dati dal progetto 1000 Genomes.
Sono stati testati per associazione 6,607,266 SNPs non sovrapposti nelle diverse mappe (tabella 4). Per il follow-up sono state selezionate nove varianti ma è stata replicata l'associazione solo per la variante rs9657904 del gene Cas-Br-M (murine) ecotropic retroviral transforming sequence b (*CBLB*) che mappa nel cromosoma 3q13.11 (105,377,110-105,587,887).

Tabella 4. Dati di qualità dell'analisi di associazione preliminare

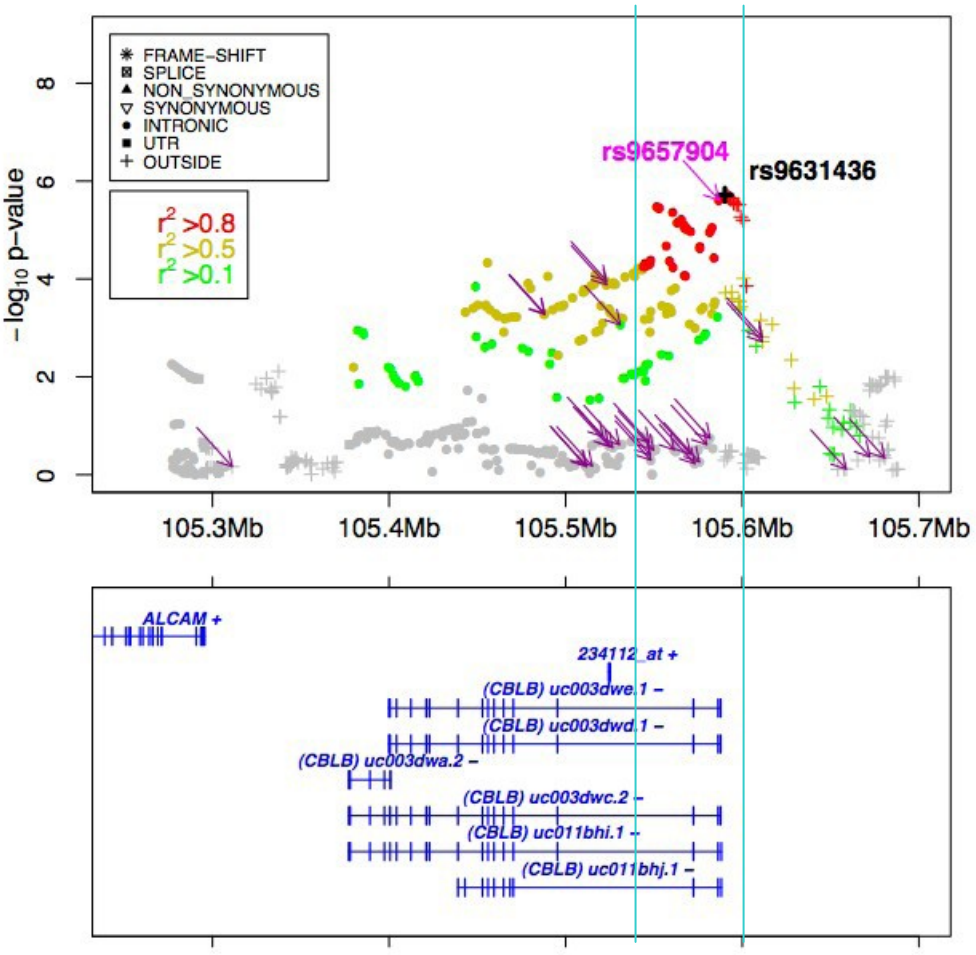
Samples QC	**# Samples** (cases / controls)
All	915/ 909
call rate < 90%	12 / 4
gender discordance	7 / 13
Presence of a relative[1]	14 / 20
Total QCed samples	882 / 872
Genotyped Markers QC	**# Markers**
All	934,968
Call rate <= 90%	115,444
abs(Cases call rate – controls call rate) > 5%	21,190
MAF < 0.05	241,002
HWE pvalue[1] < 10^{-6}	5,404
Excess of Mendelian Error [2]	389
Excess of discordances in duplicates[3]	4,733
Low quality genotype discrimination plot	85
SNPs present twice in Affymetrix chip	5
Total QCed autosomal markers	*555,334*
Total QCed chrX markers	*20,344*
Total QCed markers	*575,678*
Imputed Markers QC(HapMap II / HapMap III/ 1000G)	**# Markers**
All autosomal markers (genotyped and imputed)	2,543,887 / 1,321,001 / 8,317,360
Imputed	2,046,155 /745,323 / 7,775,000
rsqr ≤ 0.3 (HapMap II / HapMap III/ 1000G[6])	88,075 / 43,279 / 2,011,401
Monomorphic imputed	0 / 49 / 9
Total QCed markers (imputed only)	1,958,080 / 701,995 / 5,763,567
Total QCed markers (imputed and genotyped)[5]	2,533,753 / 1,277,673 / 6,338,706
Total QCed non overlapping markers	6,607,266

L'associazione della variante rs9657904, localizzata nell'introne 1 di CBLB, è stata replicata nella popolazione sarda in 1,775 casi e 2,005 controlli (analisi congiunta $P = 1.60 \times 10^{-10}$, OR = 1.40) e nella popolazione Italiana in 1,441 casi e 1,465 controlli ($P = 1.2 \times 10^{-4}$) [36].

Analisi combinata in 4,098 casi e 4,342 controlli *P-value* = 3.3×10^{-13}.

L'analisi di mappaggio fine, ha previsto l'imputazione di oltre 21 mila varianti nella regione del cromosoma 3, paia di basi da 105,377,110 a 105,587,887 ± 2Mb, derivanti da dati di sequenza NGS del progetto 1,000 Genomes e di 154 individui sardi e da dati di sequenze Sanger di 93 individui. Tale analisi ha confermato la forte associazione della variante primariamente descritta, rs9657904, e di altre varianti in forte LD (r^2 = 0.9), tra le quali rs9631436 (variante intronica) ed in generale di tutta la regione contenete *CBLB,* che si estende dal promoter fino al secondo introne del gene (figura 18). In tale regione è presente un forte LD che tiene unite le varianti più fortemente associate e che quindi rende difficoltosa l'individuazione della variante primariamente associata al rischio di SM. Possiamo concludere che ad una simile risoluzione, a causa di un forte LD non si è in grado di differenziare le varianti primariamente associate.

Figura 18. Mappaggio fine della regione *CBLB*

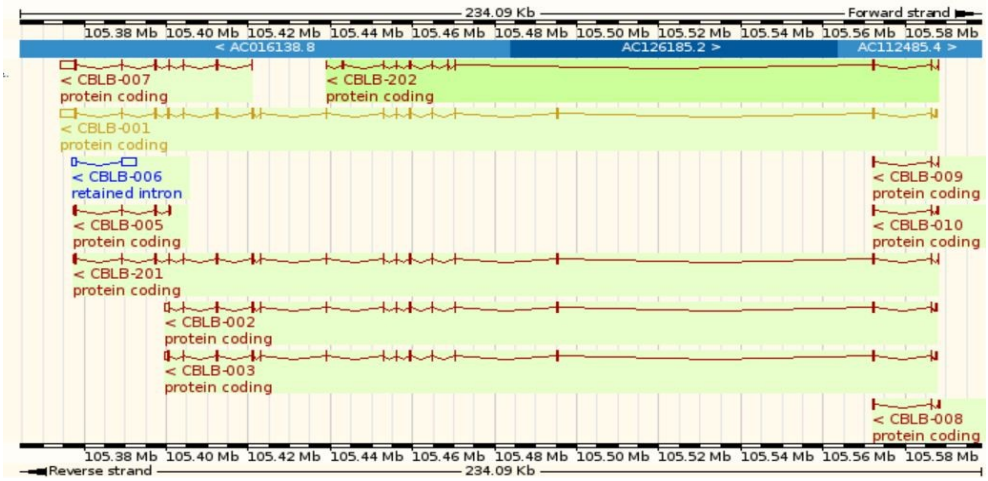


La figura 19, nel primo riquadro in alto illustra la regione CBLB, in ascissa le paia di basi e in ordinata il -log10 del P-value. I puntini, le freccette e le crocette rappresentano gli SNPs imputati dai diversi pannelli di referenza, mentre i colori rappresentano il grado di correlazione delle varianti, espresso come r^2. I puntini rossi (r^2> 0.8) rappresentano gli SNPs del pannello di referenza sardo in forte LD tra loro. Nel secondo riguardo, la schematizzazione del gene CBLB. La zona compresa tra due linee azzurre identifica la regione di maggiore associazione, rappresentata da SNPs in r^2> 0.8 .

Il gene *CBLB* fa parte della famiglia dei geni *CBL*. Il gene è costituito da 21 esoni e sono stati descritti diversi splicing alternativi che danno origine a 11 isoforme di cui 10 espresse. L'isoforma maggiore codifica per una proteina di 982-aminoacidi (figura 19) .

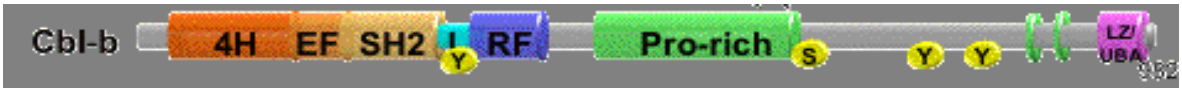
Figura 19. Isoforme gene *CBLB* tratte da Ensembl

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
CBLB-001	ENST00000264122	6780	ENSP00000264122	982	Protein coding	CCDS2948
CBLB-201	ENST00000394027	3430	ENSP00000377595	960	Protein coding	-
CBLB-007	ENST00000394030	4475	ENSP00000377598	321	Protein coding	-
CBLB-005	ENST00000407712	1493	ENSP00000384170	197	Protein coding	-
CBLB-002	ENST00000403724	3350	ENSP00000384816	770	Protein coding	-
CBLB-003	ENST00000405772	3237	ENSP00000384938	810	Protein coding	-
CBLB-008	ENST00000443752	561	ENSP00000393906	139	Protein coding	-
CBLB-010	ENST00000447441	944	ENSP00000400949	139	Protein coding	-
CBLB-009	ENST00000438603	519	ENSP00000409750	161	Protein coding	-
CBLB-202	ENST00000545639	1632	ENSP00000446116	227	Protein coding	-
CBLB-006	ENST00000476370	4722	No protein product	-	Retained intron	-



La proteina *CBLB* è caratterizzata come tutte le proteine della famiglia *CBL* da quattro principali domini, tra i quali il **dominio PTB** (N-terminal phosphotyrosine-binding) costituito da 3 differenti sottodomini: il four-helix bundle (4H), calcium-binding EF hand e il dominio SH2, il dominio **RING-type zinc finger** che possiede attività di ligasi E3 ubiquitin-protein , il **dominio UBA** che interagisce con le proteine poli-ubiquitinate e **domini ricchi di prolina (figura 18).** Le isoforme minori sono caratterizzate dall'assenza di uno o diversi motivi.

Figura 20. Domini della proteina *CBLB* della isoforma 001



L'espressione di *CBLB* è stata osservata nel polmone, rene, milza, e testicoli, cervello, fegato e cellule ematopoietiche.

La molecola codificata è implicata in una varietà di funzioni. Il gene codifica per una proteina E3 ubiquitin - ligasi, la quale accetta l'ubiquitina da specifici enzimi coniuganti E2 ubiquitina e la trasferisce a diversi substrati, promuovendo la loro degradazione nel proteasoma.

Tuttavia, la funzione di maggiore importanza è il ruolo che essa svolge nel sistema immune. *CBLB* rappresenta, infatti, un regolatore negativo del segnale di trasduzione, mediato dal recettore sia delle cellule T che delle cellule B [43].

Nelle T cellule naive, inibisce VAV1 attivando il TCR e richiedendo la stimolazione da parte di CD28 per la produzione e la proliferazione dell' Interleuchina 2, rappresenta quindi un regolatore negativo della risposta infiammatoria.

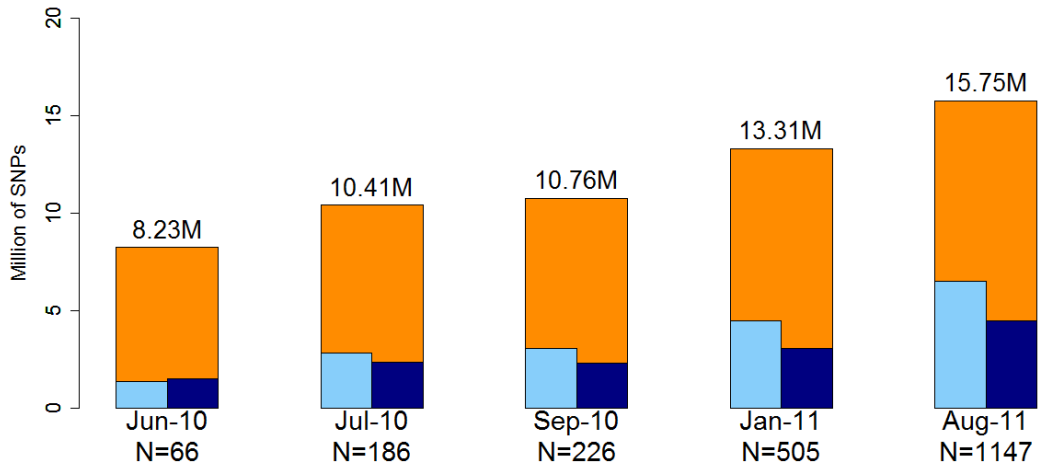
Inoltre topi deficienti dell'ortologo gene Cbl-b sono maggiormente suscettibili di sviluppare l'encefalomielite autoimmune, forma sperimentale della SM, suggerendo che le alterazioni del segnale di trasduzione, modulate da *CBLB*, possono contribuire allo sviluppo di patologie autoimmune umane, come la sclerosi multipla [44].

Ulteriori studi, in parte già in corso, sono necessari per capire come le varianti associate modifichino il rischio di SM nella popolazione umana.

7.2 Risultati delle analisi di sequenze nella popolazione sarda

Attualmente sono stati sequenziati 1,700 individui sardi e di origine sarda a bassa copertura (3-4 x) ma finora è stato analizzato solo un sub-gruppo di 1,147 individui.
Il sequenziamento dei campioni in corse *paired-end* di 240 basi con il GAIIX e 204 basi con l'Hi-Seq 2000 ha prodotto, in media, ~9 e ~15 milioni di basi per campione, corrispondenti a 37.5 e 74 milioni di clusters, per campione, rispettivamente.
Le enormi moli di dati che vengono costantemente generate dai sequenziatori HiSeq 2000 e GAIIx, il tempo e lo spazio disco richiesto per l'analisi ci hanno imposto di generare periodicamente dei *data freeze,* che raccolgano di volta in volta, tutte le informazioni ottenute alla data stabilita.
l risultati del confronto del numero di varianti identificate, rispetto al numero degli individui sequenziati nei diversi *data freeze*, mostra che tanto maggiore è il numero degli individui sequenziati e tanto maggiore sono il numero delle varianti identificate (figura 21).
La figura 21 mostra che a Giugno del 2010 erano stati analizzati 66 individui e solo 8.23 milioni di SNPs erano stati identificati. In quattordici mesi sono stati sequenziati ed analizzati ulteriori 1,081 campioni e sono stati quasi raddoppiati il numero degli SNPs identificati (15.57 milioni di SNPs). Il numero degli individui sequenziati ha subito un discreto incremento nel tempo, grazie al miglioramento dei software per la chiamata delle basi ed ai reagenti di reazione di sequenza, che hanno aumentato le performance della piattaforma, consentendo di raggiungere un maggior numero di *clusters* ed un aumentata *coverage* per individuo.
Sebbene, questi continui *update* abbiano creato la necessità di riadattare ogni volta i protocolli previsti dalla casa madre, hanno permesso, infine, di mettere a punto esperimenti di *multiplexing* che hanno consentito di sequenziare più individui in una stessa *lane,* riducendo lo sforzo economico e migliorando le tempistiche.

Figura 21. Numero delle varianti descritte vs il numero dei campioni sequenziati



La figura 21 illustra in ordinata le date dei diversi data freeze ed il numero di campioni analizzati. In ascissa sono rappresentati, nell'ordine di milioni, gli SNPs identificati. In arancio il numero totale degli SNPs identificati, in azzurro il numero di SNPs non presenti in dbSNP132 e in blu il numero di SNPs non presenti nel catalogo 1000 Genomes.

Il pannello di referenza, generato nel data *freeze* di Gennaio 2011, è stato utilizzato allo scopo di inferire e testare per associazione, le varianti in esso contenute nel GWAS#2. Tale pannello è costituito da dati di sequenza di 505 individui (347 aplotipi dopo deplezione degli individui imparentati) e contiene 13.313.964 SNPs, di cui il 60.8% contenute in dbSNP129 e il 67.2% in dbSNP131.

7.3 Risultati dell'analisi GWAS#2

Lo studio di associazione su tutto il genoma è stato condotto dopo la fase d'imputazione dei genotipi delle varianti non direttamente genotipizzate in 4,200 individui sardi e di origine sarda, costituiti da 2,280 casi SM e 1,922 controlli ed utilizzando metodi di imputazione che sfruttano sia dati di sequenza di 280 individui di origine Europea del progetto 1000 Genomes che dati di sequenza della popolazione

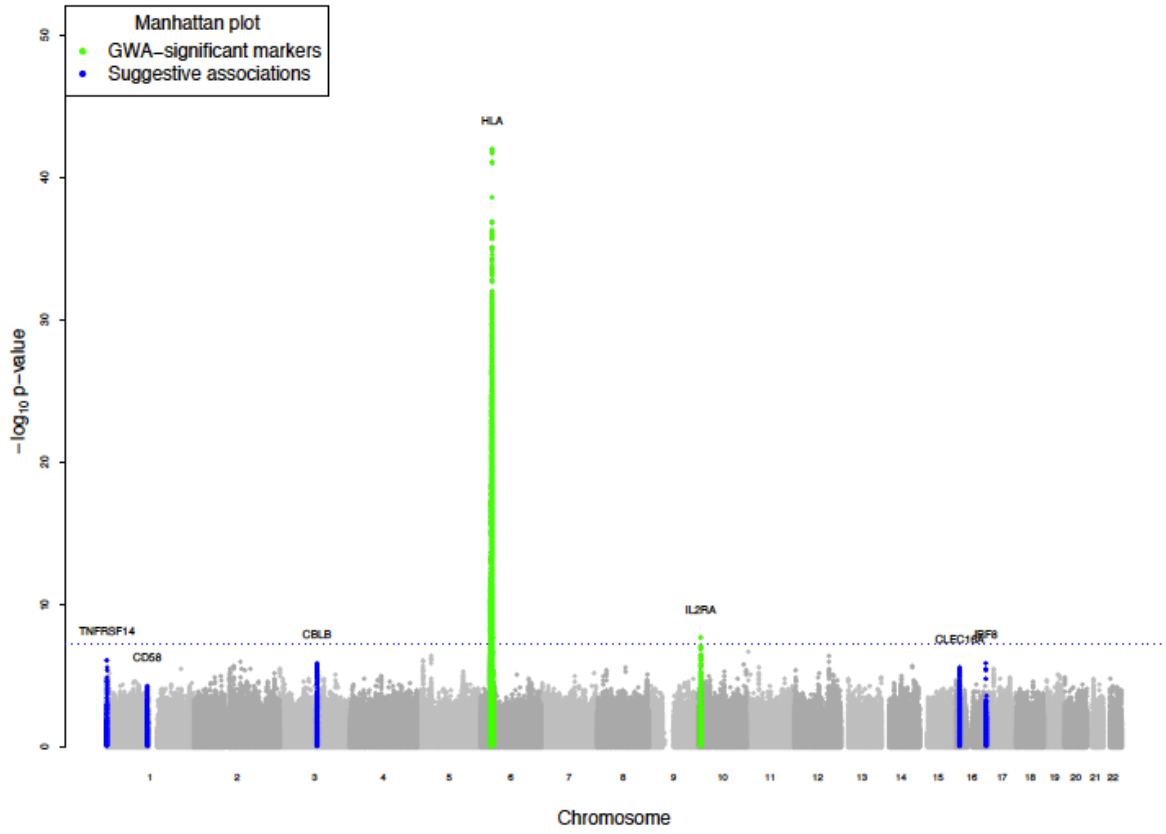
sarda (505 individui, pannello di referenza con 347 aplotipi). I risultati dello studio di associazione su tutto i genoma non hanno rivelato nessuna nuova associazione. Tuttavia, i risultati, illustrati dal Manhattan plot in figura 22, hanno confermano l'associazione a livello *Genome-Wide* (*P-value* > 10^{-8}) del locus di predisposizione maggiore HLA (*P-value*= 1.06 x 10^{-42}) e del recettore dell'IL2 (*IL2RA*).

E' stata, altresì, confermata l'associazione, con significatività statistica suggestiva ai loci noti; *TNFRSF14* che mappa nel cromosoma 1p36.32, *CD58* in 1p13.1 , *CBLB* in 3q13.11, *CLEC16A* sul cromosoma 16p13.13 ed infine di *IRF8* in 16q24.1. Infine, è stata confermata l'associazione, con suggestiva significatività statistica, a 25 nuovi loci descritti recentemente dal consorzio internazionale SM (tabella 5).

Tabella 5. Associazione nella popolazione sarda dei nuovi loci descritti dal IMSC

Gene	*SNP*	*RSQR*	*P-value*
PLEK	chr2:68580836	1	4.98E-05
MERTK	chr2:112642370	0.8	1.65E-04
SP140	chr2:231077725	1	1.42E-05
EOMES	chr3:27840071	1	1.15E-04
CD86	chr3:121769522	0.9	6.29E-04
NFKB1(MANBA)	chr4:103552068	0.4	6.58E-06
IL12B	chr5:158649872	0.9	3.08E-03
BACH2	chr6:90607855	0.6	8.05E-04
THEMIS	chr6:127985785	1	7.39E-05
MYB(AHI1)	chr6:135891265	0.9	3.35E-05
IL22RA2	chr6:137540335	0.9	5.28E-04
TAGAP	chr6:159474624	0.8	3.60E-05
ZNF767	chr7:149414709	1	3.55E-03
MYC	chr8:129033872	1	1.61E-03
PVT1	chr8:129033872	1	1.61E-03
HHEX	chr10:94479107	1	1.28E-04
CXCR5	chr11:118743286	1	6.57E-05
ZFP36L1	chr14:69332968	0.5	8.35E-04
BATF	chr14:75931070	0.6	1.59E-03
GALC(GPR65)	chr14:88556085	0.8	8.26E-05
SOX8	chr16:1008647	0.7	2.17E-06
RPS6KB1	chr17:57932314	0.9	1.32E-04
MALT1	chr18:56258036	0.7	1.83E-03
TNFSF14	chr19:6668972	0.5	1.89E-04
MPV17L2 (IL12RB1)	chr19:18241857	1	2.58E-05
CYP24A1	chr20:52783652	0.7	5.64E-04
TNFRSF6B	chr20:62260845	0.4	5.83E-04
MAPK1	chr22:22245556	0.9	5.69E-03
SCO2	chr22:50952989	0.4	5.46E-03

Figura 22. Manhattan plot dell'analisi GWAS#2



La figura 22, rappresenta i risultati di associazione GWAS#2. In ascissa sono rappresentati i cromosomi, in ordinata è rappresentato il -log10 del P-value. Ogni puntino rappresenta uno SNP. In verde i loci replicati con associazione Genome Wide (P-value < 10^{-8}) mentre in blu i geni replicati con suggestiva associazione.

Il pannello di referenza sardo ha mostrato una migliore efficienza rispetto al pannello di referenza 1,000 Genomes. Le migliori prestazioni si sono manifestate non solo nella qualità dell'imputazione, valutata dal parametro RSQR, con un deciso miglioramento dei genotipi delle varianti non tipizzate, ma anche con un incremento della significatività statistica ai loci replicati (tabella 6).

Tabella 6. Accuratezza dell'imputazione del pannello sardo vs 1,000 Genomes attraverso l'esempio di alcuni geni noti

Geni	Pannello di referenza sardo			Pannello di referenza 1000 Genomes			stesso SNP?	r2
	RSQR	*P-value*	I/G	RSQR	*P-value*	I/G		
TNFRSF14	0.82	8.41×10^{-7}	I	0.66	9.86×10^{-7}	G	NO	0.41
CD58	0.89	4.88×10^{-5}	I	0.85	1.80×10^{-4}	I	NO	0.94
CBLB	0.99	1.36×10^{-6}	G	0.99	2.29×10^{-6}	I	NO	1
HLA	0.89	1.06×10^{-42}	I	0.67	1.06×10^{-39}	I	NO	_
IL2RA	0.94	1.80×10^{-8}	I	0.97	2.27×10^{-9}	I	SI	1
CLEC16A	0.99	2.36×10^{-6}	I	0.96	3.86×10^{-4}	I	NO	0.35
IRF8	0.98	7.07×10^{-4}	I	0.94	1.30×10^{-4}	I	NO	0.9

I = Inferito, G = Genotipizzato

Un numero maggiore di SNPs ha superato il filtro della qualità d'imputazione (RSQR > di 0.3). Sono stati, infatti, testati per associazione un totale di 7,220,405 SNPs con RSQR > di 0.3 (media RSQR = 0.81) imputati attraverso il pannello di referenza sardo, rispetto a 6,234,041 SNPs con RSQR > di 0.3 (media RSQR = 0.76) imputati con il pannello 1000 Genomes.

La qualità dell'imputazione è migliorata soprattutto per le varianti a bassa frequenza e per le varianti rare (MAF 1-3%), media RSQR 0.73 vs 0.66, rispettivamente per il pannello di referenza sardo e per il pannello 1000 Genomes (tabella 7).

Tabella 7. Varianti imputate; pannello sardo vs pannello 1000 Genomes

MAF	Varianti Imputate Referenza Sarda	Media RSQR	Varianti Imputate Referenza 1000 G	Media RSQR
1-3%	1,275,990	0.73	884,402	0.66
3-5%	779,669	0.81	752,123	0.78
>5%	5,164,746	0.89	4,597,516	0.86

Inoltre, i risultati dimostrano che il processo d'imputazione, valutato anche dal parametro r^2 di MACH che valuta l'accuratezza dell'imputazione, definito dal rapporto tra la varianza osservata del dosaggio allelico e la varianza attesa, in equilibrio di Hardy-Weinberg. L'imputazione, anche in questo caso, ha dimostrato una performance migliore con l'utilizzo del pannello referenza sardo, generato da dati di sequenza di un maggior numero di individui, soprattutto per le varianti a bassa frequenza e varianti rare (tabella 8).

Tabella 8. Accuratezza dell'imputazione (r^2)

MAF	Pannello di referenza 1000 Genomes (563 ID non imparentati)	Pannello di referenza sardo (347 ID non imparentati)	Pannello di referenza sardo (831 ID non imparentati)
1-3%	0.75	0.90	0.94
3-5%	0.88	0.95	0.97
>5%	0.94	0.97	0.98

7.4 Problematiche incontrate nello studio

Durante la fase di analisi sono state incontrate numerose difficoltà e sfide, alle quali, tuttavia, di volta in volta sono state trovare rapide ed efficaci soluzioni.
La prima sfida riscontrata in questo studio è stata l'attribuzione dei genotipi con il software Birdseed-v2. Infatti, per le caratteristiche del software è necessaria l'analisi in un unico cluster di tutti casi e tutti i controlli, al fine di evitare distorsioni nell'assegnazione dei genotipi, dovute al processamento dei campioni in diversi laboratori e in diversi tempi. Il cluster unico richiede ingenti risorse hardware e tempi lunghi. Per l'attribuzione dei genotipi, dopo diverse prove, è stata scelta la strategia più efficiente e capace di rispettare il criterio del cluster unico nel minor tempo d'analisi. Tale metodo ha previsto la distribuzione di gruppi di SNPs per tutti gli

individui. Ciascun gruppo conteneva 50,000 SNPs ed i gruppi si sovrapponevano per un totale di 5,000 SNPs.

Il software è stato eseguito in maniera distribuita su macchine dedicate, facenti parte del cluster di calcolo del CRS4. Il centro di calcolo del CRS4 è in grado di fornire fino a 47 Teraflops di potenza di calcolo, 1,2 Pbyte di spazio disco e 800 Terabyte circa di spazio per backup. Alle macchine utilizzate per i *runs* è stato collegato un sistema di *storage* appositamente configurato perché fosse molto robusto e potesse ospitare e gestire gli accessi alla grossa mole di dati analizzata.

L'altra maggiore sfida è stata l'assenza di un sistema centralizzato o (*database*) di raccolta di dati anagrafici ed anamenstici che ha reso più difficoltosa la collezione delle informazioni necessarie al fine dell'analisi. Per agevolare queste fasi, si sta attualmente costituendo, attraverso una collaborazione con il *CRS4* ed il sistema sanitario regionale, un *database* clinico (realizzato sulla base del software client-server OMERO) con lo scopo di unificare tutte le informazioni finora reperite ed ottenute dai dati genetici, ed interfacciare le stesse con i dati disponibili dalla rete sanitaria.

Un'altra difficoltà incontrata, ha riguardato l'allineamento dei genotipi rispetto allo *strand* del genoma umano di riferimento e delle coordinate di posizione per i set di dati utilizzati. Infatti i genotipi prodotti dalle piattaforme di genotipizzazione e gli algoritmi utilizzati per la chiamata delle basi possono essere espressi nello strand positivo (+) o negativo (-) rispetto al genoma di riferimento. Quindi è stato fondamentale, prima di procedere all'imputazione, che i genotipi di ciascun SNP fossero espressi nello stesso *strand* in tutto il set di dati utilizzato, ma anche che il sistema di riferimento della posizione di ciascun SNPs in paia di basi fosse il medesimo per ciascun set di dati (Build 37, dbSNP129).

8 CONCLUSIONI E SVILUPPI FUTURI

Gli studi di associazione sull'intero genoma hanno, in pochi anni, consentito l'identificazione di numerose varianti geniche, oltre 1,400, legate a numerosi tratti e fenotipi complessi.
Oltre 57 varianti sono state descritte associate la sclerosi multipla, con soglie genome wide ($p>5 \times 10^{-8}$); il nostro gruppo di ricerca ha contribuito, attraverso la descrizione dell'associazione del gene *CBLB*, uno dei geni più importanti nella suscettibilità alla SM. L'associazione del gene *CBLB* è stata replicata anche dal recente lavoro del consorzio internazionale IMSGC [25]. Abbiamo inoltre confermato associazioni descritte da altri gruppi, in particolare dal consorzio internazionale IMSGC, e grazie al nostro progetto di sequenziamento esteso all'intero genoma ridotto tali associazioni ai loro elementi primari, il cosidetto " mappaggio fine".
Nel loro insieme i risultati ottenuti dal nostro e da altri gruppi definiscono in maniera inequivocabile la natura immunologica, autoimmune della sclerosi multipla, in quanto una larga quota dei prodotti proteici dei geni associati con la malattia sono coinvolti nella regolazione delle risposte immuni. I risultati ottenuti definiscono anche il modello genetico della componente ereditabile della malattia, con una regione genica, la regione HLA, con effetti più rilevanti sul rischio di sviluppare la malattia e altre varianti coinvolte, al di fuori della regione HLA, con effetti genetici più modesti e con una larga fetta di ereditabilità, molto verosimilmente legata a varianti rare e molto rare, ancora da spiegare.

I risultati del nostro studio confermano l'importanza di esaminare ampie casistiche, necessarie allo scopo di aumentare il potere statistico dello studio. Tale constatazione impone, sempre più fortemente, linee strategiche che prevedano collaborazioni tra clinici e ricercatori e tra diversi gruppi di ricerca per perseguire e centrare gli obiettivi comuni.

I risultati documentano inoltre l'utilità di studiare la variabilità genetica della popolazione sarda, attraverso metodi di sequenziamento *low-pass* con NGS. In particolare, dimostrano che l'utilizzo di dati di sequenza generati nella medesima popolazione oggetto di studio, favoriscono un miglioramento delle performance dei

metodi di imputazione statistica, grazie all'analisi di aplotipi più simili tra gli individui del pannello di referenza e gli individui studio tipizzati con i chip array e non sequenziati. Il nostro lavoro pioneristico nell'integrazione attraverso imputazione dei dati ottenuti con i chip array e quelli generati in un sub gruppo di individui attraverso sequenziamento dell'intero genoma a bassa copertura, rappresenta un modello degli studi GWAS che saranno effettuati in futuro anche dal resto della comunità scientifica internazionale.

La migliore conoscenza della variabilità genetica sarda consentirà inoltre la realizzazione di nuovi e più performanti chip array di tipizzazione molecolare (come ad esempio l'Exome chip, prossimo chip in commercio), che potranno essere utili non solo per studi nella popolazione sarda, ma in maniera più estesa anche per studi in altre popolazioni e soprattutto in quella Europea, infatti varianti fondatrici, in particolare quelle con elevata frequenza nell'isola (frequenza allelica > 20%) tendono ad essere presenti, seppur molto rare, anche fuori dalla Sardegna. Tali chip array renderanno inoltre ancora più accurata l'imputazione statistica, in futuri studi.

L'obiettivo finale di tutte le ricerche genetiche nelle malattie complesse, cosi come nella SM è quello di risolvere il problema dell'eziologia della malattia e trovare modi efficaci per la diagnosi, trattamento e prevenzione della malattia. L'effettuazione di studi funzionali sui prodotti proteici delle varianti fin qui identificate e la ricerca dell'ereditabilità restante rimangono gli obiettivi primari degli studi attuali e futuri.

L'analisi di casistiche ancora più numerose e l'inclusione delle varianti rare e fondatrici nel disegno sperimentale svolgeranno un'importante ruolo in questo senso, Grazie alla crescente disponibilità dei dati di sequenziamento (1,000 Genome e 10,000 Genomes UK) ed in particolare dei dati di sequenziamento nella popolazione sarda sarà possibile migliorare ulteriormente l'accuratezza dell'inferenza statistica nell'analisi delle varianti rare.
In particolare, nel nostro progetto, i prossimi pannelli di referenza, generati da dati di sequenza di un numero maggiore di individui, fino a 3,000, consentiranno una migliore estrazione dell'informazione genetica e precisione del pannello di referenza sardo, ed includeranno variazioni genetiche che possono contribuire alla suscettibilità della SM, cosi come alla sua prevenzione, diagnosi e trattamento appropriato della patologia.

Le varianti rare saranno, altresì, indagate utilizzando dati di sequenza dell'RNA (RNA-seq) di 1,000 individui sardi, progetto attualmente in corso.
L'analisi del trascrittoma consentirà di esaminare, inoltre, da un punto di vista qualitativo e quantitativo i profili di espressione e correlarli con i profili genetici (eQTL) e con i tratti fenotipici esaminati, fornendo ulteriori informazioni sulle conseguenze funzionali delle associazioni genetiche.
Tale metodo consentirà anche una migliore estrazione delle varianti rare, e permetterà, altresì, di valutare variazioni del numero di copie, come inserzioni, delezioni ma anche inversioni e traslocazioni, per comprenderne il loro eventuale ruolo nell'eziologia della SM e di altre patologie autoimmuni comuni in Sardegna.
La caratterizzazione genotipica di varianti incluse nell'ImmunoChip Illumina, chip commerciale con numerose di varianti in geni coinvolti nell'immunità e nell'infiammazione e varianti già descritte in associazione a malattie autoimmuni, in tutta la casistica finora raccolta, di 3,100 pazienti e 4,500 controlli (progetto in corso), aumenterà il potere statistico di identificare le varianti coinvolte nella SM .
Inoltre, il nostro studio parallelo dei tratti immuno-fenotipici quantitativi, alcuni dei quali potenzialmente correlati alla SM e l'identificazione delle associazioni coincidenti con la malattia e con specifiche variabili immuno-fenotipiche, fornirà informazioni chiave per la comprensione delle basi funzionali dell'associazione genetica, in quanto consentirà di correlare in maniera diretta, ogni sostituzione nucleotidica del DNA, associata alla SM, con un particolare tipo cellulare responsabile della malattia stessa.

Tutte le evidenze saranno infine sviluppate anche con successivi studi funzionali, al fine di confermare il ruolo di ciascuna specifica variante genetica associata.

Nel suo insieme il lavoro in corso deluciderà i meccanismi cellulari nei quali i prodotti proteici dei geni intervengono, e consentirà una migliore comprensione di come una singola variazione nel DNA può modificare i meccanismi in senso predisponente nei confronti della SM e di altre malattie autoimmuni, il delicato equilibrio attraverso il quale le risposte autoimmuni sono regolate. Infine, ciò permetterà di evidenziare potenziali bersagli terapeutici e aiuterà i ricercatori impegnati nella ricerca farmacologica pre-clinica, che potranno, attraverso un approccio basato sulla conoscenza, selezionare composti utili per la generazione di nuovi e più efficaci farmaci.

BIBLIOGRAFIA

1. McFarlin, D.E., and H.F. McFarland. Multiple sclerosis (first of two parts). N Engl J Med 307:1183-1188 (1982).

2. McFarlin, D.E., and H.F. McFarland. Multiple sclerosis (second of two parts). N Engl J Med 307:1246-1251(1982).

3. Henry F McFarland and Roland Martin. Multiple sclerosis: a complicated picture of autoimmunity. Nature Immunology 8, 913 - 919 (2007).

4. EL Waubant, et all. Pathophysiology of multiple sclerosis lesions- Science & Medicine, (1997).

5. Steinman L. et al. Multiple sclerosis: a two stage disease. Nature Immunol. 2, 762–765 (2001).

6. Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology 33 (11): 1444–52(1983).

7. Granieri, E.,et al. The increasing incidence and prevalence of MS in a Sardinian province. Neurology 55:842-848 (2000).

8. Pugliatti, M. et al. The epidemiology of multiple sclerosis in Europe. Eur J Neurol 13:700-722 (2006).

9. Fisher R.A.. The Correletion between relatives on the supposition of mendelian inheritance. Transactions of the Royal Society of Edinburgh 52:399-433 (1918).

10. Ebers, G.C. et al. A genetic basis for familial aggregation in multiple sclerosis. Canadian Collaborative Study Group. Nature 377:150-151 (1995).

11. Sospedra, M., and R. Martin. Immunology of multiple sclerosis. Annu Rev Immunol 23:683-747 (2005).

12. Christensen T, et al. Gene-environment interactions in multiple sclerosis: innate and adaptive immune responses to human endogenous retrovirus and herpesvirus antigens and the lectin complement activation pathway. J Neuroimmunol. Feb;183(1-2):175-88 (2007).

13. Blanco-Kelly F, et al. Members 6B and 14 of the TNF receptor superfamily in multiple sclerosis predisposition. Genes Immun. Mar;12(2):145-8 (2011).

14. Smolders J. et al. Vitamin D as an immune modulator in multiple sclerosis, a review. J Neuroimmunol. Feb;194(1-2):7-17 (2008).

15. Jersild, C. et al. HL-A antigens and multiple sclerosis. Lancet 1:1240-1241 (1972).

16. Hillert, J. and Olerup, O. Multiple sclerosis is associated with genes within or close to the HLA–DR–DQ subregion on a normal DR15, DQ6, Dw2 haplotype. Neurology, 43, 163–168(1993).
17. Hauser, S.L. et al. Extended major histocompatibility complex haplotypes in patients with multiple sclerosis. Neurology, 39, 275–277(1989).
18. Kwon, O.J. et al. HLA class II susceptibility to multiple sclerosis among Ashkenazi and non-Ashkenazi Jews. Arch. Neurol., 56, 555–560 (1999).
19. Saruhan-Direskeneli, G. et al. HLA–DR and –DQ associations with multiple sclerosis in Turkey. Hum. Immunol., 55, 59–65(1997).
20. Alvarado-de la Barrera, C. et al. HLA class II genotypes in Mexican Mestizos with familial and nonfamilial multiple sclerosis. Neurology, 55, 1897–1900 (2000).
21. Marrosu, M.G., et al. Multiple sclerosis in Sardinia is associated and in linkage disequilibrium with HLA–DR3 and –DR4 alleles. Am. J. Hum. Genet., 61, 454–457(1997).
22. Marrosu, M.G., et al. DRB1–DQA1–DQB1 loci and multiple sclerosis predisposition in the Sardinian population. Hum. Mol. Genet., 7, 1235–1237 (1998).
23. Marrosu, M.G. et al. Dissection of the HLA association with multiple sclerosis in the founder isolated population of Sardinia. Hum. Mol. Genet. 10 (25): 2907-2916 (2001).
24. Bhasi K, et al. Information-theoretic identification of predictive SNPs and supervised visualization of genome –wide association studies. 34, 14:e101 (2006).
25. The International Multiple Sclerosis Genetics Consortium & The Wellcome Trust Case Control Consortium 2. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 476, 214–219 (2011).
26. Zondervan K.T. and Cardon L.R. Designing candidate gene and genome-wide case-control association studies. Protocol; doi:10.1038 (2007).
27. Reich E., et al. Linkage disequilibrium in the human genome. Nature 411, 199-204 (2001).
28. Patil N. et al. Blocks of Limited Haplotype Diversity Revealed by High-Resolution Scanning of Human Chromosome 21. Science, 294 no. 5547, 1719-1723 (2001).
29. Sanna et all.Variants within the immunoregulatory *CBLB* gene are associated with multiple sclerosis Nat Genet. 42, 495–497 (2010).
30. Manolio T. A et al. Finding the missing heritability of complex diseases. Nature 461(7265): 747–753 (2009)

31. Li B et al. Discovery of Rare Variants via Sequencing: Implications for the Design of Complex Trait Association Studies. PLoS Genet, 5(5), 1-9 (2009).
32. Hilma Holm H. et al. A rare variant in MYH6 is associated with high risk of sick sinus syndrome. Nature Genetics 43, 316–320 (2011).
33. Nejentsev S. et al. Rare Variants of IFIH1, a Gene Implicated in Antiviral Responses, Protect Against Type 1 Diabetes. Science, 324 no. 5925, 387-389 (2009).
34. Lampis, R. et al. The inter-regional distribution of HLA class II haplotypes indicates the suitability of the Sardinian population for case-control association studies in complex diseases. Hum. Mol. Genet. 9, 2959-65. (2000).
35. Cucca, F. et al. The distribution of DR4 haplotypes in Sardinia suggests a primary association of insulin dependent diabetes mellitus with DRB1 and DQB1 loci. Hum. Immunol. 43, 301-308 (1995).
36. Corrado L. et al. Association of the CBLB gene with multiple sclerosis: new evidence from a replication study in an Italian population. J Med Genet;48:210-211 (2011).
37. Yun Li et al. Low-coverage sequencing: Implications for design of complex trait association studies. Genome Res. 21: 940-951 (2011).
38. Quail M.A. et al.A large genome center's improvements to the Illumina sequencing system. Nature Methods; 5(12), 1005-1010 (2008).
39. Livak K.J. et al. Oligonucleotides with fluorescent dyes at opposite ends provide a quenced probe system useful for detecting PCR product and nucleic acid hybridization. PCR Methods Appl, 4: 357-362 (1995).
40. Kozarewa I. et al. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (GþC)-biased genomes. Nat Methods, 6(4):291–5 (2009).
41. Li H, et al. The sequence alignment/map format and SAMtools. Bioinformatics, 25(16):2078–9 (2009).
42. Li Y. et al. Genotype imputation. Ann. Rev. Genomics Hum. Genetics; 10: 387-406 (2009).
43. Bachmaier, K. et al. Negative regulation of lymphocyte activation and autoimmunity by the molecular adaptor Cbl-b. Nature 403, 211–216 (2000).
44. Chiang, Y.J. et al. Cbl-b regulates the CD28 dependence of T-cell activation. Nature 403, 216–220 (2000).

RINGRAZIAMENTI

Ringrazio il responsabile nonché fondatore del progetto, Prof. Francesco Cucca e tutto il team interdisciplinare di lavoro costituito da ricercatori IRGB, CRS4, UNISS ed del laboratorio del Centro Sclerosi Multipla di Cagliari, che ha consentito lo svolgersi della ricerca.
Ringrazio tutti i clinici che hanno collaborato ed infine tutti i pazienti ed i volontari che hanno partecipato alla ricerca.