UNIVERSITÀ DEGLI STUDI DI SASSARI

FACOLTÀ DI MEDICINA E CHIRURGIA

DIPARTIMENTO DI SCIENZE BIOMEDICHE, CENTRO DI GENETICA CLINICA
Direttore: Chiar.mo Prof. A. Montella

DOTTORATO DI RICERCA IN SCIENZE BIOMEDICHE

INDIRIZZO IN GENETICA MEDICA,

MALATTIE METABOLICHE E NUTRIGENOMICA

Ciclo XXIV

Direttore: Chiar.mo Prof. E.Tolu

Sequenziamento e studi di associazione su tutto il genoma per svelare i geni di suscettibilità al diabete di tipo 1.

Relatore:
Prof. Francesco Cucca

Dottoranda
Dott.ssa Francesca Deidda

Anno Accademico 2010-2011

Sommario

ABSTRACT

During my third doctoral year, I worked in the Genome-Wide Association (GWA) and sequencing study of type 1 diabetes (T1D), directed by Prof. Francesco Cucca. The project aims to find type 1 diabetes risk loci using a large collection of Sardinian patients and controls. T1D is a multifactorial autoimmune disease, common in Sardinia, so the population is an appropriate cohort for the study. The GWA study design was a scan of 1,377 patients and 1,917 healthy unrelated controls genotyped with the Affymetrix 6.0 chip that contains over 900K single nucleotide polymorphisms (SNPs) across the genome. These data are then imputed with whole genome sequencing data from reference panels: our 508 sequenced Sardinian individuals, or 280 Europeans (of the 1000 genomes project), respectively. This let us test more than 13 million variants per person. Our analysis confirmed ($P=1x10^{-5}$) several known associations outside the HLA region (*PTPN22, CTLA4, IL2RA* and *INS* gene). Imputation quality improved with the Sardinian reference panel. For instance, the INS gene association p-value increased from $7.3*10^{-8}$ (1000 Genomes) to $5.5*10^{-13}$ (Sardinian), suggesting that DNA sequencing data from a specific population improves accuracy of imputation in that population. To increase the chance to find new T1D susceptibility loci, we are expanding the dataset (to > 2,000 cases) and sequencing 1,200 Sardinian individuals for the reference panel.

PREMESSA

Lo studio delle patologie complesse negli ultimi anni è passato dall'analisi specifica di un singolo locus all'analisi simultanea di più loci situati su cromosomi differenti.
Nel corso degli ultimi anni è aumentata la conoscenza della variabilità genetica, grazie alla disponibilità della sequenza completa del genoma umano, del progetto HapMap, e del completamento della fase I del progetto 1000 Genomi, con la descrizione di più di 15.000.000 di SNPs, 1.000.000 inserzioni e delezioni, 20.000 varianti strutturali. Tali avanzamenti abbinati ai progressi tecnologici nelle metodologie di miniaturizzazione molecolare hanno consentito la generazione di micro-arrays commerciali che hanno consentito una rapida evoluzione nei metodi di genotipizzazione. Tali metodi, implementati in piattaforme di genotipizzazioni delle maggiori compagnie sul mercato, Affymetrix ed Illumina, sono in grado di testare contemporaneamente centinaia di migliaia di varianti geniche (sino ad un milione di polimorfismi a singolo nucleotide, SNP) in migliaia di individui a costi accessibili, attraverso l'utilizzo di Chip-array.

Questi progressi hanno reso possibile, a partire dal 2006, l'attuarsi degli studi di associazione su tutto il genoma (Genome-Wide Association Study, GWAs). Gli studi di associazione si basano sul confronto delle frequenze alleliche o genotipiche in un campione di pazienti, definiti casi, rispetto ad un campione di individui sani, definiti controlli, non imparentanti ne tra loro ne con i casi. Il test valuta differenze statistiche nella frequenza di specifici alleli nei casi rispetto ai controlli.
L'ipotesi alla base di uno studio di associazione è che la presenza di polimorfismi genetici sia correlata all'aumento o alla diminuzione del rischio di sviluppare patologie complesse; esistono varianti alleliche con ruolo di predisposizione alle malattie e varianti alleliche con un ruolo protettivo più frequenti negli individui sani.

Il primo studio GWAs è stato pubblicato su Nature nel Febbraio del 2007 per la ricerca delle varianti di suscettibilità del diabete di tipo II. In seguito sono stati condotti numerosi studi GWAs sulle più importanti patologie complesse (artrite reumatoide,

disordini bipolari, ipertensione, patologie coronariche, psoriasi) . I GWAs, a partire dal 2006 hanno rilevato 1449 associazioni significative per 237 tratti comuni e malattie complesse (http://www.genome.gov/gwastudies/), contribuendo a definirne le basi ezio-patogenetiche che fino ad allora rimanevano in gran parte sconosciute.
In particolare in 5 anni i GWAs hanno identificato 41 nuovi loci associati con il diabete di tipo (DT1), patologia trattata in questa tesi .

Il metodo ha rivoluzionato il modo di fare ricerca, in quanto ha permesso l'analisi dell'intero genoma contemporaneamente a livelli di risoluzione mai raggiunti in precedenza. Sono stati individuati geni che non si pensava potessero avere un ruolo nelle patologie analizzate. I chip-array utilizzati nella prima generazione dei GWAs sono stati ottimizzati per aumentare la possibilità di trovare varianti comuni di malattia, con piccoli rischi relativi, e varianti rare con grandi rischi relativi.
Infatti, 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 indels (inserzioni o delezioni).
Solo una piccola frazione e rappresentata da SNPs non sinonimi. Questi SNPs hanno un ruolo più specifico nel determinare la variazione dell'espressione genica. L'analisi simultanea degli SNPs e dell'espressione genica permette, inoltre, di mappare i fattori genetici alla base delle differenze individuali dei livelli quantitativi di espressione (eQTLs).

Sebbene i GWAs abbiano determinato un notevole progresso nella conoscenza del rischio genetico di diversi tratti e patologie, queste in nessun caso spiegano completamente la loro ereditabilità.
E' stato stimato che per il DT1, il locus di predisposizione maggiore HLA, il gene *INS* insieme ai nuovi 41 loci descritti spiegano solo circa il 60% dell'ereditabilità stimata.
La dissezione dei fattori genetici che contribuiscono al rischio di DT1 è un compito arduo ed è complicata da diversi fattori:
1 La dimensione dell'effetto genetico, Odds Ratio (OR), definito da piccole differenze di

frequenza nel campione dei casi rispetto ai controlli delle varianti associate al DT1. La maggior parte delle varianti di suscettibilità delle malattie complesse, cosi come del DT1, ha piccoli effetti individuali. Per mitigare questa problematica sono necessarie ampie casistiche di casi e di controlli capaci di fornire un'adeguata potenza statistica allo studio.

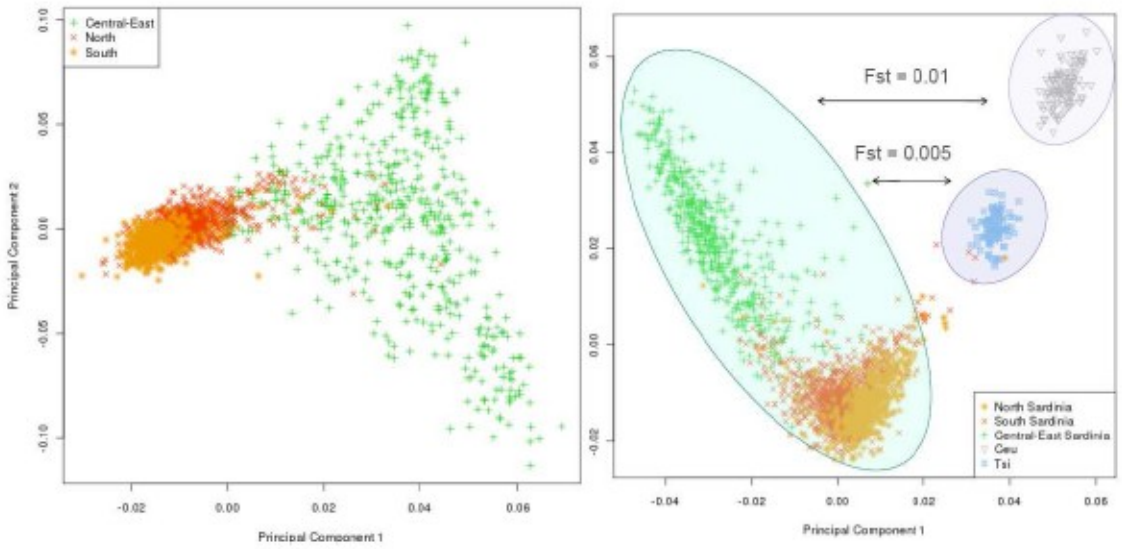
2 Associazione indiretta, definita dal un possibile incompleto linkage disequilibrium (LD) tra il marcatore esaminato e la variante primariamente associata che riduce il segnale d'associazione della variante testata. Anche in questo caso è richiesta una casistica molto ampia che consenta di moderare le conseguenze di questa problematica.

La necessità di analizzare grandi casistiche è anche dettata anche 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.

3 Problematiche tecniche definite dalle mappe di marcatori utilizzate negli attuali chip di genotipizzazione. Infatti i chip commerciali attualmente disponibili sono costruiti con sonde che provengono da un numero limitato di popolazioni, esse hanno solo un limitato potere nel catturare le varianti rare (MAF<5%), e non valutano direttamente il contributo di corti polimorfismi (inserzioni o delezioni). Pertanto i chip commerciali in uso interrogano prevalentemente varianti comuni e non considerano varianti rare e varianti fondatrici frequenti in alcune popolazioni ma rare o assenti in altre. Tali varianti potrebbero svolgere un ruolo importante nella predisposizione alle malattie e spiegare la proporzione di rischio genetico non ancora mappato. Ciò ovviamente assume un importanza significativa 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 è un macro isola genetico, la cui la popolazione ha vissuto un lungo isolamento geografico, con minimo tasso di immigrazione ed elevata percentuale di matrimoni fra consanguinei. L'alto grado di parentela che la caratterizza fa si che un numero ridotto di geni di predisposizione/varianti alleliche causino fenotipi complessi. La popolazione sarda, è una popolazione omogenea e anche nostri dati mostrano una

sostanziale assenza di sottostruttura (figura 1). Essa, pur collocandosi nell'ambito della variabilità europea, manifesta tutta una serie di caratteristiche di unicità: alcune varianti genetiche sono particolarmente frequenti in Sardegna e talvolta rare o assenti in altre popolazioni. Si tratta di varianti antiche che erano già presenti in quegli individui che diverse migliaia di anni fa hanno popolato l'isola, chiamato effetto fondatore.



La figura 1 mostra a sinistra l'analisi tra le diverse sub-regioni della Sardegna e a destra il confronto della popolazione sarda con popolazioni HapMap di origine europea (CEU) e italiana (TSI).

Lo scopo del lavoro descritto in questa tesi e' aumentare le conoscenze delle basi genetiche del DT1, utilizzando metodologie e tecniche innovative, e sfruttando le peculiarità della popolazione sarda.

INTRODUZIONE

Il diabete di Tipo 1

Il DT1 è una malattie multifattoriali a carattere autoimmune, causata dall'interazione di fattori genetici predisponenti e di fattori ambientali sconosciuti.
Il DT1 rappresenta la patologia con base genetica, attualmente maggiormente studiata dal punto di vista genetico, presumibilmente in quanto è accessibile all'analisi genetica, per la sua diagnosi clinica inequivocabile e per il suo esordio giovanile che consente la raccolta dei DNA parentali.
Il contributo relativo della componente genetica e della componente ambientale nella predisposizione al DT1 è indicato dal diverso rischio di contrarre la malattia in diverse categorie di individui.

Il diabete è una malattia che impedisce all'organismo di utilizzare correttamente l'energia derivante dagli alimenti. Si manifesta quando il pancreas non produce insulina o quando, anche se prodotta, non viene assimilata dall'organismo che non è in grado di utilizzarla (insulino-resistenza). Essa è una malattia cronica dovuta ad una alterazione della produzione o della funzione di un ormone pancreatico, l'insulina, che ha la funzione di ridurre la quantità di glucosio presente nel sangue.

In generale il diabete mellito, si origina o per la ridotta produzione d'insulina, ormone chiave nella regolazione del metabolismo glucidico, o per una difettosa utilizzazione dell'insulina stessa legata ad alterazioni recettoriali e post-recettoriali a livello degli organi bersaglio dell'insulina. Esempi paradigmatici di queste due forme di diabete sono rappresentati rispettivamente dal DT1, caratterizzato da una severa deficienza nella produzione d'insulina e dal diabete di tipo 2 (DT2), in cui l'insulina, pur essendo presente e in certe fasi della malattia perfino aumentata, non agisce per una cosiddetta insulino-resistenza a livello degli organi bersaglio.
Nel diabete mellito si verifica una grave turba del metabolismo glucidico con conseguente incapacità dell'organismo a mantenere il livello di glucosio del sangue al di sotto di un certo valore. L'Organizzazione Mondiale della Sanità ha stabilito tale valore

come una glicemia dopo pasto di 200 mg/dl o come una glicemia a digiuno di 126 mg/dl, riscontrati in almeno due rilevazioni indipendenti. Esistono diverse forme di diabete mellito che presentano profonde differenze per patogenesi, quadro clinico all'esordio, età media d'insorgenza, severità del decorso e necessità di misure terapeutiche specifiche.

Il diabete mellito di tipo 1 è causato dalla distruzione irreversibile delle cellule β pancreatiche da parte di linfociti T autoreattivi, con conseguente deficit della sintesi dell'ormone insulina e permanente dipendenza dalla somministrazione dell'insulina esogena.

La malattia rappresenta la forma di diabete mellito clinicamente più grave e ad insorgenza più precoce. Il processo autoimmune insorge in individui geneticamente suscettibili in presenza di fattori ambientali permissivi.

Il DT1 è più comunemente causato dall'autodistruzione, operata da linfociti T autoaggressivi, delle cellule del pancreas, le uniche dell'organismo capaci di produrre efficientemente l'insulina. La malattia è in realtà il risultato finale di un processo infiammatorio cronico e si manifesta clinicamente quando circa il 90% delle cellule β pancreatiche è andato distrutto e quindi solo il 10% è ancora funzionante.

A questo punto del processo autodistruttivo la produzione d'insulina da parte delle β cellule residue non è più sufficiente per regolare in maniera ottimale l'ingresso e l'utilizzazione del glucosio all'interno delle cellule dell'organismo; compare quindi il quadro sintomatologico tipico della malattia, legato all'iperglicemia plasmatica e al concomitante squilibrio metabolico.

Nel soggetto diabetico si manifesta minzione frequente, poliuria (cioè aumento della quantità di urina emessa), enuresi o emissione involontaria di urina dopo aver raggiunto il controllo degli sfinteri, polidipsia (una sensazione di sete intensa che porta ad un frequente bisogno di bere), diminuzione della massa muscolare, stanchezza e dimagramento.

La causa di questa sintomatologia è una cheto-acidosi metabolica, legata allo scompenso metabolico di base, associata ad una disidratazione iperosmolare, in quanto la concentrazione osmotica del plasma aumenta,per un incremento della concentrazione del glucosio extra-cellulare. Ciò determina richiamo di acqua dalle cellule al plasma con conseguente grave disidratazione intra-cellulare e marcata sofferenza di vari organi, fra i

quali il sistema nervoso centrale, che in situazioni estreme determina il coma diabetico.
Fino a pochi anni fa si riteneva che il DT1 fosse una patologia essenzialmente pediatrica e che fosse inevitabilmente associata ad una totale dipendenza dalla somministrazione di insulina esogena fin dall'esordio clinico della malattia. In realtà si è visto che questa forma autoimmune di DT1 pur essendo più frequente nei bambini può insorgere in qualunque epoca della vita. In generale la malattia tende a scompensarsi molto più velocemente nel bambino e richiede obbligatoriamente una terapia insulinica sostitutiva per sopravvivere.
L'esordio clinico della malattia è preceduto da una fase di latenza di durata variabile ma spesso di diversi anni, durante la quale si ha una progressiva distruzione delle cellule β pancreatiche producenti l'ormone insulina. In questa fase è possibile attuare una previsione di suscettibilità, analizzando le varianti di predisposizione genetica ad oggi note e rilevando la presenza dei marcatori auto-anticorpali che precedono la manifestazione del diabete.
L'attuale cura del DT1è rappresentata da una sostituzione ormonale a partire dall'esordio - che avviene tipicamente anni dopo l'insorgenza del processo patogeno - e per tutto il resto della vita. Quando comincia la terapia, purtroppo una proporzione rilevante di cellule negli organi bersaglio possono essere state irreversibilmente distrutte.

Incidenza della malattia

Un ulteriore motivo d'interesse della popolazione sarda è rappresentato dall'elevata incidenza, di malattie multifattoriali autoimmuni quali il DT1 e la sclerosi multipla. La Sardegna, in particolare, condivide con la Finlandia la più alta incidenza di DT1al mondo.
Ciò rappresenta una rimarchevole eccezione al gradiente nord-sud mostrato dall'incidenza di tale patologia.
A parte le similitudini sul piano epidemiologico, il DT1e la sclerosi multipla mostrano in Sardegna una tendenza alla co-morbidità negli stessi individui e alla co-occorrenza nelle

stesse famiglie .

Ciò suggerisce che in Sardegna siano frequenti alleli di suscettibilità nei confronti di entrambe le patologie.

Il DT1 mostra una differente incidenza in Europa: in particolare in Scandinavia, e bassa nell'area mediterranea, con un'eccezione: la Sardegna. La malattia è inoltre molto rara in Asia e relativamente rara nelle popolazioni africane fin qui studiate.

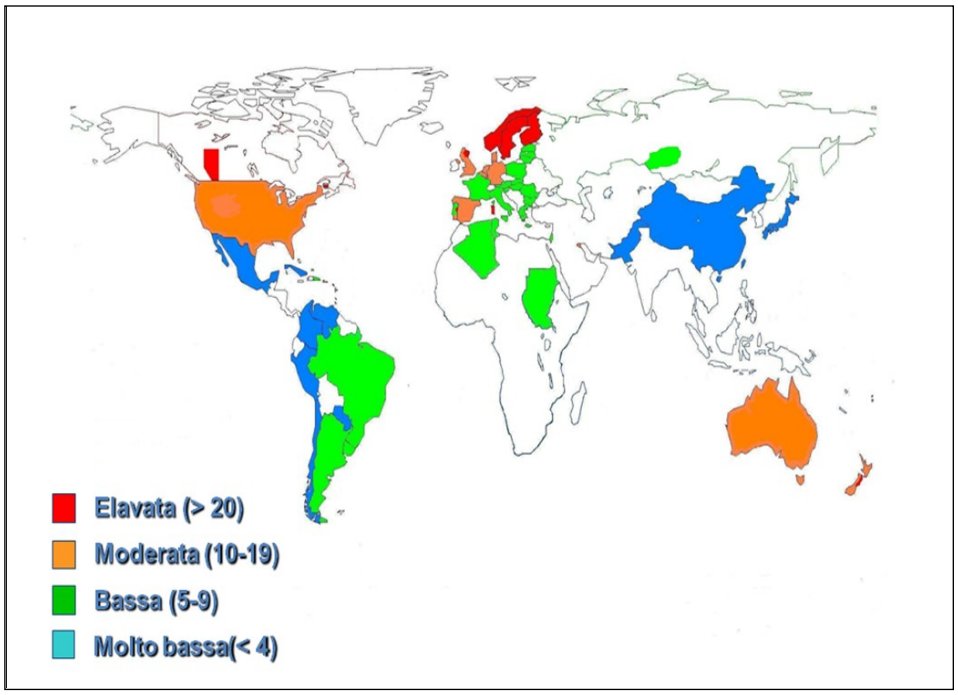


Figura 2 Mappa dell'incidenza del DT1 in alcuni paesi del mondo. In rosso i paesi ad elevata incidenza > di 20 casi annui per 100.000 nati vivi, il colore arancio regioni a incidenza intermedia, in verde le regioni a bassa incidenza, mentre in celeste sono rappresentate le regioni a bassissima.

Il rischio medio di DT1 per un individuo scelto a caso nella popolazione Europea è pari a circa lo 0.4%.

Questo rischio incrementa al 6 % quando l'individuo in questione è un fratello di un

paziente diabetico.

Se questo ipotetico fratello condivide con il malato una totale identità genotipica a livello della regione HLA/MHC il suo rischio di contrarre la malattia incrementa al 12-15 %. Infine il rischio di un gemello monozigote (MZ) di un paziente diabetico è pari al 70%. Questi dati epidemiologici indicano che la malattia mostra una marcata tendenza a segregare nelle famiglie in quanto il rischio di malattia è 15 volte più alto in un fratello di un malato (6%) rispetto a quello della popolazione generale (0.4%).

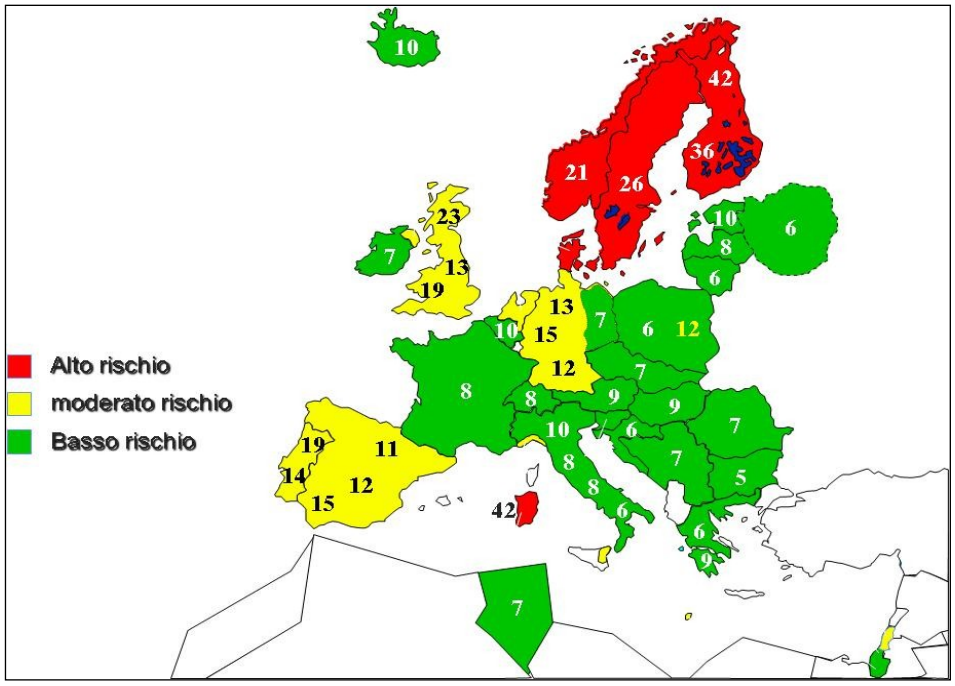


Figura 3 Mappa dell'incidenza del DT1in Europa e in alcuni paesi del mediterraneo I numeri riportati in corrispondenza delle diverse regioni geografiche si rifericono ai nuovi casi annui x 100.000 nati vivi nella fascia d'età 0-14 anni (Green 2001.)

I dati indicano inoltre che geni contenuti nella regione HLA sono importanti in quanto il rischio di un fratello di un diabetico è più alto se condivide con il probando un'identità

genetica a livello della regione HLA (12-15%). Anche geni al di fuori della regione HLA sono importanti in quanto il rischio dei gemelli MZ, che condividono una completa identità non solo a livello della regione HLA ma a livello di tutto il genoma, è ancora più alto (70%). Infine, queste osservazioni suggeriscono che anche fattori ambientali o epigenetici sono implicati, e quindi la penetranza del DT1 è incompleta, in quanto il rischio empirico per gemelli MZ di pazienti DT1 è inferiore al 100% (70%).

La quantificazione del rischio nei gemelli MZ è particolarmente importante, in quanto fornisce una stima diretta della penetranza per l'intero corredo genetico di suscettibilità. Occorre sottolineare che un ambiente permissivo appare necessario per la piena estrinsecazione del rischio genetico. La principale evidenza epidemiologica a supporto della componente ambientale è rappresentata dal fatto che nei paesi cosiddetti "industrializzati" il numero di pazienti con questa patologia è in rapida crescita. Questo fenomeno non può essere spiegato dai fattori genetici, perchè l'aumento d' incidenza è avvenuto in un arco temporale troppo ristretto per essere accompagnato da modifiche sostanziali dell'assetto genetico di quelle popolazioni.

Il rapido incremento d'incidenza della malattia negli ultimi 40 anni nei paesi occidentali potrebbe essere legato sia alla comparsa di fattori ambientali predisponenti e/o alla scomparsa di fattori protettivi. Sfortunatamente i cambiamenti ambientali avvenuti in questo arco di tempo sono stati così tanti da rendere problematica l'identificazione di quelli effettivamente coinvolti in tale aumento d'incidenza del DT1. La ricerca dei fattori ambientali implicati nella predisposizione al diabete è ulteriormente complicata dalla loro interazione con i geni, in particolare con le proteine codificate dai geni. Indipendentemente dal peso del suo contributo, la componente ambientale sembra operare a diversi livelli nel processo patologico, influenzando il decorso della malattia, sia nel senso di una predisposizione che nel senso di una protezione.

Alcuni fattori ambientali sembrano coinvolti in stadi precocissimi, per esempio la rosolia congenita incrementa il rischio di DT1 di circa 15 volte e il rischio nei gemelli dizigotici (13%) è più alto che nei fratelli (6%) ; altri sembrano operare nell'accelerare o precipitare l'esordio della malattia. Tutto lascia pensare, comunque, che gli ipotetici fattori diabetogenici ambientali, almeno nei paesi occidentali siano ubiquitari e per questo di difficilissima identificazione.

LOCI NOTI DI SUSCETTIBILITÀ AL DIABETE DI TIPO 1

Gli studi che indicano una componente genetica nel DT1 sono iniziati nel 1970 ed hanno rivelato il contributo fondamentale del complesso maggiore di istocompatibilità (MHC) alla suscettibilità a questa malattia. Analisi genetiche sistematiche, hanno rivelato che le varianti ai loci DRB1 e DQB1 da sole non spiegano l'intera predisposizione genetica alla malattia conferita dalla regione HLA considerata nel suo complesso. Altre varianti localizzate al di fuori della regione HLA sono inoltre necessarie per spiegare le complesse basi genetiche della malattia. Tuttavia, attraverso gli studi di associazione genetica sono stati rivelati ulteriori loci che sono associati con DT1. A partire dal 2007, attraverso questi studi sono stati aumentati il numero dei loci che sono noti per essere associati con DT1 ad oltre 40.(Pociot, Akolkar et al. 2010). In totale, le varianti di suscettibilità alla malattia rappresentano circa il 48% del rischio genetico, in cui il 41% di questo è da attribuirsi alla sola regione HLA.

Si stima che il contributo ambientale alla malattia rappresenti circa un 20% per questo un terzo del rischio globale rimane non ancora spiegato.

Questo residuo è dovuto a numerosi loci di suscettibilità non ancora identificati per i quali ci si attende piccoli effetti genetici dovuti a varianti rare. Vediamo ora nel dettaglio i vari loci di suscettibilità per il DT1.

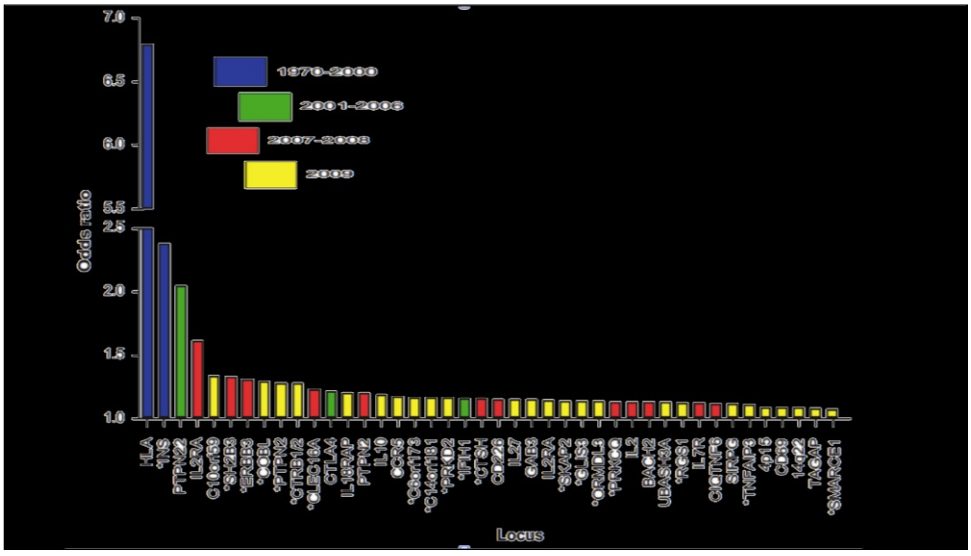


Figura 4 Andamento delle scoperte dei geni nel diabete tipo 1 attraverso gli studi di associazione.

Organizzazione genica, struttura e funzione della regione HLA.

Prima di esaminare in dettaglio l'associazione del sistema HLA con il DT1, è opportuno discutere brevemente gli aspetti salienti inerenti la composizione genica e la funzione dei principali geni contenuti in questa regione.

La regione HLA rappresenta la versione umana del complesso maggiore di istocompatibilità (MHC) e contiene oltre 200 geni localizzati nel contesto di una sequenza lineare di circa 3.500.000 paia di basi di DNA sul cromosoma 6 (6p21.31).

Le proteine codificate da molti di questi geni svolgono importanti funzioni nella regolazione delle risposte immuni e pertanto sono state e sono tutt'ora sottoposte a forti pressioni selettive.

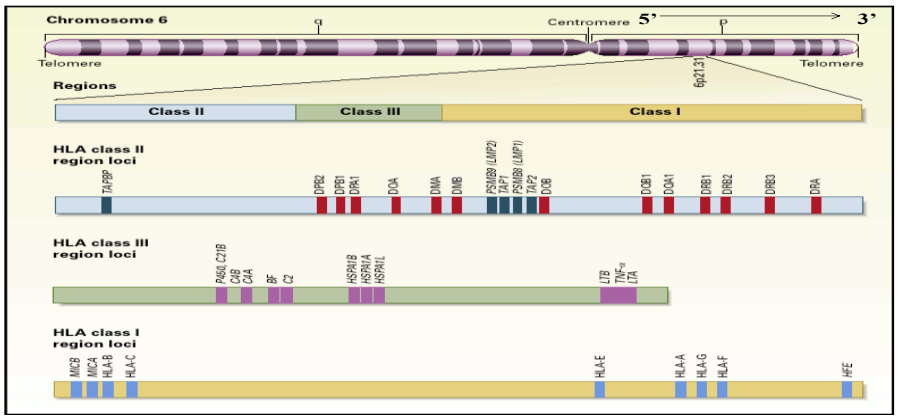


Figura 5 Localizzazione dei "clusters" genetici nella regione HLA.

Questo è un sistema sofisticato e integrato di regolazione dell'espressione genica. Più specificamente dal centromero al telomero del braccio corto del cromosoma 6 si trovano rispettivamente i geni cosiddetti di classe II, di classe III e di classe I.Nella classe II e nella classe I sono contenuti una serie di geni che codificano per proteine altamente polimorfiche denominate molecole HLA classiche, in quanto in grado di presentare peptidi ai linfociti T.

In particolare, le molecole HLA classiche sono codificate nella sub-regione di classe I dai loci A, B, C. Questi loci codificano per una catena pesante altamente polimorfica che

forma un dimero con una molecola non polimorfica, la b2 microglobulina, che viene codificata sul braccio lungo del cromosoma 15.

Queste molecole di classe I sono espresse in tutte le cellule. Le molecole hanno la funzione di presentare peptidi virali ai linfociti citotossici (o CTL) denominati anche CD8+ per la presenza di tale molecola co-accessoria sulla superficie cellulare.

Nella subregione di classe II sono presenti geni, tra i quali DP, DQ e DR, che presentano due loci, che vengono indicati con l'aggiunta delle lettere A e B, per esempio DPB e DPA, che codificano per un dimero funzionale formato rispettivamente dalla catena a e della catena b.

Queste molecole di classe II sono espresse in cellule specializzate, denominate APC o "antigen presenting cells" che sono rappresentate da macrofagi, cellule dendritiche, monociti e linfociti B. I prodotti di tali geni di classe II sono deputati alla presentazione di peptidi di origine batterica e protozoaria ai linfociti T helper (Th) o CD4+. Nell'ambito della sub-regione di classe II esistono anche altri geni i cui prodotti proteici costituiscono molecole HLA non classiche, in quanto non hanno la capacità di presentare direttamente peptidi ai linfociti T. In particolare, come vedremo in dettaglio più avanti, i prodotti dei loci denominati DM e DO sono coinvolti nel trasporto e nel corretto posizionamento di peptidi di origine batterica e protozoaria nel contesto delle molecole di classe II.

Fra la classe II e la classe I, nella cosiddetta classe III, esiste inoltre un terzo gruppo di geni, che codificano per citochine e altre molecole coinvolte nei processi infiammatori. In particolare questa regione include numerosi loci "tumor necrosis factor" (TNF), denominati a, b, g, d, infine i loci i cui prodotti proteici costituiscono alcune frazioni del complemento.

È importante rilevare che a livello della regione HLA esistono anche geni senza una relazione evidente con le risposte immunitarie, o geni con funzione ignota. Altri geni ancora, denominati pseudogeni, non sono espressi e rappresentano vestigia evoluzionistiche di loci che hanno accumulato mutazioni che gli hanno resi non funzionali. Riguardo alla struttura delle molecole HLA classiche, le proteine codificate dai geni di classe I e di classe II presentano una struttura abbastanza simile. Ciò non sorprende in quanto si tratta di recettori specializzati nella presentazione di peptidi ai linfociti T e in particolare al loro recettore denominato TCR (o T cell receptor).

Gli antigeni di classe I non sono erano il fattore eziologico che determina suscettibilità alla malattia, piuttosto erano solo marcatori in linkage disequilibrium (LD) con le reali varianti eziologiche. Il LD rappresenta un'associazione non casuale di alleli localizzati su uno stesso cromosoma (Zavattari, Deidda et al. 2000) e nel caso della regione HLA tende a mantenere insieme specifiche combinazioni di alleli di classe I, III e II. L'elevato LD ha semplificato l'osservazione iniziale ma rese difficoltosa l'attribuzione della suscettibilità ad un singolo locus piuttosto che all'aplotipo esteso (Cucca and Todd 1996),(Gyllensten and Erlich 1993),(Klitz, Stephens et al. 1995),(Jorde, Watkins et al. 1994). Successivamente fu evidente che l'associazione con la malattia era maggiore con gli alleli di classe II HLA-DR3 e -DR4, positivamente associati con la malattia, i quali si trovano in LD rispettivamente con il B8 e il B15 (Wolf, Spencer et al. 1983).

Studi successivi evidenziarono una forte associazione anche con alcuni alleli al locus HLA-DQB1 (in particolare gli alleli DQB1*0302 e DQB1*0201) (Owerbach, Lernmark et al. 1983), (Todd, Bell et al. 1987). L'associazione deriva dalla presenza nella catena β di un amminoacido diverso dall'acido aspartico in posizione 57. Di fatto l'acido aspartico in posizione 57 è importante per la configurazione della tasca, sito di legame con l'antigene (Todd, Bell et al. 1987), (Ettinger, Liu et al. 2000). La tecnologia e le strategie di mappaggio attuali consentono di giungere alle stesse conclusioni con esperimenti eseguibili in tempi brevi (Zavattari, Lampis et al. 2000) (Herr, Dudbridge et al. 2000).

In particolare è stato dimostrato (Zavattari, Lampis et al. 2000)con chiarezza che nella popolazione sarda l'associazione della regione HLA con il diabete è dominata dalle molecole *DQB1* e *DRB1*. Simili risultati sono stati ottenuti anche in altre popolazioni, confermando che l'associazione con questi loci è primaria (Herr, Dudbridge et al. 2000). Di fatto anche studi sul modello murino di DT1 (il topo NOD) hanno dimostrato in modo diretto il ruolo primario dei prodotti IA (ortologo murino del DQ) e IE (ortologo murino del DR) nella suscettibilità e protezione nei confronti della malattia (Kwok, Nepon et al. 1995), (Lund, O'Reilly et al. 1990), (Miyazaki, Uno et al. 1990),(Slattery, Kjer-Nielsen et al. 1990; Singer, Tisch et al. 1998), (Wicker, Todd et al. 1995). Lo studio dell'associazione genetica fra il DT1 e le molecole HLA di classe II nella popolazione sarda hanno esaustivamente spiegato la complessa natura dell'associazione della regione HLA (Cucca, Lampis et al. 1995), (Cucca, Muntoni et al. 1993),(Cucca, Dudbridge et al.

2001), (Zavattari, Lampis et al. 2001).

Inoltre questa popolazione per la peculiare distribuzione dei vari aplotipi HLA, ha consentito di risolvere gli effetti confondenti dovuti al linkage disequilibrium, chiarendo così il contributo individuale dei vari loci presenti nella regione.

Lo studio dell'associazione degli aplotipi DR-DQ in un' ampia casistica di pazienti sardi, inglesi e americani (Cucca, Lampis et al. 2001).

ha consentito di definire in modo ancora più esaustivo l'associazione dei vari aplotipi DR-DQ con il DT1. È possibile suddividere l'associazione della malattia con i vari aplotipi DR-DQ in quattro gruppi principali.

Aplotipi positivamente associati o ad alto rischio:

(DR4) DRB1*0405/*0401/*0402/*0404-DQA1*0301-DQB1*0302

(DR3) DRB1*0301-DQA1*0501-DQB1*0201;

Aplotipi neutrali:

(DR8) DRB1*08-DQA1*0401-DQB1*0402,

(DR9) DRB1*0901-DQA1*0301-DQB1*0303,

(DR6) DRB1*1302-DQA1*0102-DQB1*0604,

(DR2) DRB1*1601-DQA1*0102-DQB1*0502;

Aplotipi negativamente associati o a basso rischio:

(DR6) DRB1*1301-DQA1*0103-DQB1*0603,

(DR7) DRB1*0701-DQA1*0201-DQB1*0201,

(DR5) DRB1*11/12-DQA1*0501-DQB1*0301;

Aplotipi con gradi estremi di associazione negativa o fortemente protettivi:

(DR5) DRB1*11/12-DQA1*0501-DQB1*0301,

(DR2) DRB1*1501-DQA1*0102-DQB1*0602,

(DR7) DRB1*0701-DQA1*0201-DQB1*0303,

(DR6) DRB1*1401-DQA1*01-DQB1*0503.

Esistono anche altri loci HLA diversi dal DR e DQ che sono in grado di influenzare ulteriormente il rischio di sviluppare la malattia. In particolare, alcune varianti alleliche al locus DPB1, che codifica per la terza molecola di classe II potrebbero avere un ruolo primario nell'associazione con il DT1 .

Il gene dell'insulina.

L'associazione del DT1 con la sequenza polimorfica contenuta in 5' del gene dell'insulina è nota fin dal 1981 . Il gene dell'insulina umana (INS) è localizzato sul braccio corto del cromosoma 11 a livello della banda 15.5 (IDDM2), tra i geni per la tirosina-idrossilasi (TH) e il fattore di crescita II insulino-simile (IGF2) (Bell, Pictet et al. 1980; Ullrich, Dull et al. 1980; Harper, Ullrich et al. 1981).

A livello del promoter del gene dell'insulina, 365 bp dal sito d'inizio della trascrizione del gene, è presente una regione polimorfa composta di ripetizioni in tandem (VNTR) di sequenze di 14-15 bp, collegate alla ripetizione più comune A(C/T)AGGGGT(G/C)C(T)GGGG (Bell, Karam et al. 1981).

Questo polimorfismo, dato da un'inserzione-delezione di sequenze di DNA, pur non trovandosi all'interno del gene ha effetti nell'espressione di INS (Bell, Karam et al. 1981).Gli alleli VNTR sono suddivisi in tre classi in base al numero d'unità ripetute: la classe I è la regione polimorfica più corta, è costituita da 26-63 ripetizioni per una lunghezza media di 590 pb. La prevalenza degli alleli di classe I nei Caucasici è circa il 70%. La classe II ha una lunghezza intermedia, è costituita da 64-139 ripetizioni per una lunghezza di circa 1200 pb ed molto rara nella popolazione Europea (1976). Infine, la classe III rappresenta la variante più lunga. E' costituita da 140-200 ripetizioni, con una taglia media di 2200 pb e la prevalenza nei Caucasici è stata stimata intorno al 30% . La prima descrizione di un'associazione positiva tra genotipo omozigote per alleli di classe I e diabete di tipo1, su un ampio data-set degli USA, si deve a Bell e collaboratori (Bell, Horita et al. 1984)

L'associazione è stata poi successivamente confermata in numerose casistiche

indipendenti, anche attraverso test di associazione intra-familiari (Julier, Hyer et al. 1991). I risultati positivi di queste analisi di associazione non ricevettero inizialmente conferma da analisi di linkage, per cui fu da più parti suggerito che i risultati ottenuti negli studi di associazione fossero in realtà legati ad una stratificazione genetica, legata ad un'incompleta omogenità dei pazienti e dei controlli per origine etnica. Queste difficoltà sono state superate con la messa a punto di test di associazione intrafamiliare: la selezione di famiglie con almeno un genitore eterozigote al locus INS, utilizzata nel TDT, ha confermato che l'associazione fra la VNTR 5' al gene dell'insulina e il DT1 non è spuria ma è dovuta a linkage con la malattia stessa). Inizialmente il ruolo funzionale della VNTR nell'influenzare l'espressione del gene dell'insulina non fu compreso. Ritenendo improbabile che fosse primariamente e direttamente coinvolta nella suscettibilità al DT1, fu considerata un marcatore in LD con un polimorfismo sconosciuto in grado di influenzare direttamente la predisposizione nei confronti della malattia (Spielman, McGinnis et al. 1993),(Field 1991),

Un ulteriore contributo è apportato da varianti polimorfiche del promoter del gene dell'insulina, che mappa sul braccio corto del cromosoma 11 (IDDM2), esso rappresenta un effetto genetico più modesto. L' analisi di associazione del IDDM2 ha messo in evidenza che varianti alleliche di un minisatellite (VNTR) presente nella zona del promoter del gene del insulina (INS) risultano essere eziologiche (Bennett, Wilson et al. 1996). Le differenze di lunghezza del VNTR determinano la suddivisione in tre classi di cui una predisponente (I /I) e due (I / III e III / III) protettive .

In base al grado d'associazione o linkage disequilibrium fra i vari polimorfismi e il DT1, fu delineata la regione minima d'associazione (i.e l'intervallo cromosomico più breve associato con la malattia) nei 4,1 Kb intorno al gene dell'insulina, comprendente il locus VNTR in 5' e altri 9 polimorfismi (Julier, Hyer et al. 1991; Lucassen, Julier et al. 1993). La maggiore associazione con il DT1 era data dal VNTR stesso (Bennett and Todd 1996) e dai loci –23/HphI e + 1140 A/C.

Tuttavia, per lo stretto linkage disequilibrium tra i vari polimorfismi era difficile determinare con certezza quale fra questi fosse la mutazione eziologica primaria e quali fossero solo polimorfismi secondariamente associati. Utilizzando un set di dati familiari combinati di famiglie diabetiche e mettendo a confronto i diversi aplotipi, risultò che 6

dei 10 polimorfismi normalmente associati ad un aumentato rischio di sviluppare il DT1 erano presenti sia su aplotipi di suscettibilità e che di protezione.

Assumendo un modello con un solo polimorfismo casuale, era quindi molto improbabile che tali polimorfismi rappresentassero le componenti eziologiche presenti nella regione. A questo punto i candidati più plausibili per un effetto eziologico erano:-2733A/C, la VNTR, -23/HphI, e +1140A/C presenti solo su aplotipi predisponenti (Julier, Lucassen et al. 1994).

Uno studio su un data-set della popolazione finnica confermò questa ricerca: si paragonarono le frequenze genotipiche di 7 dei 10 polimorfismi (esclusi –2733 A/C, +1140 A/C e +1355 T/C) in pazienti finnici e in soggetti di controllo e si trovò che solo i loci VNTR e –23/HphI erano significativamente associati al DT1 (Undlien, Bennett et al. 1995). Gli esperimenti successivi riguardarono la subtipizzazione degli alleli VNTR per arrivare all'esclusione dei polimorfismi –2733A/C, –23/HphI e +1140A/C.

La subtipizzazione degli alleli VNTR di classe III portò alla suddivisione di questi in 15 sottoclassi di lunghezza definibile (da 301 a 315) (Bennett, Lucassen et al. 1995).

Queste furono raggruppate in base all'allele al locus HUMTH01 con cui erano in forte LD. Questo è un microsatellite tetranucleotidico, localizzato 9kb a monte dell'INS nel primo introne del gene TH, con 5 alleli comuni, Z, Z-4, Z-8, Z-12 e Z-16, ognuno diverso per una singola unità ripetuta di 4bp.

HUMTH01 di per se stesso non era un candidato per l'effetto IDDM2 perché esterno alla regione d'associazione più forte (Lucassen, Julier et al. 1993). I suoi alleli, però, sono in forte linkage disequilibrium con gli alleli VNTR di classe III, soprattutto l'allele Z di HUMTH01 e le sottoclassi 306-310 da un lato e l'allele Z-8 e le sottoclassi 304-306 dall'altro (Bennett, Lucassen et al. 1995). Furono considerati soltanto gli aplotipi VNTR che contenevano l'allele negativamente associato con la malattia ai tre loci polimorfici: -2733A/C, -23/HphI e +1140A/C.

Questi aplotipi, quando associati con l'allele Z-8, erano significativamente più protettivi, rispetto agli aplotipi associati con l'allele Z. Gli alleli -2733A/C, -23/HphI e +1140A/C erano identici in questi aplotipi esibenti diversi gradi di protezione nei confronti della malattia, mentre l'unica differenza consisteva negli specifici alleli di classe III. Questo suggeriva che IDDM2 corrispondesse ad una variazione allelica nel locus VNTR e non

nei loci -2733A/C, –23/HphI e +1140A/C.

L'analisi degli alleli VNTR di classe I diede un'ulteriore, indiretta conferma di questo risultato. Nelle popolazioni caucasiche furono definiti almeno 21 alleli di classe I che differivano fra loro di una singola unità di ripetizione di 14 o 15 bp .

Anche se gli alleli VNTR di classe I erano globalmente trasmessi da genitori eterozigoti significativamente più spesso alla progenie diabetica rispetto a quelli di classe III, fu osservato che non tutti gli alleli di classe I erano trasmessi con uguale frequenza ai pazienti. In altri termini si rilevò un'eterogeneità nel grado d'associazione positiva con il DT1 da parte degli alleli di classe I. Anche in questo caso la differenza nella trasmissione di alleli specifici di classe I poteva essere spiegata solo dalla variazione allelica nello stesso VNTR (Bennett, Lucassen et al. 1995)

Gli specifici alleli VNTR di classe I, cosi come quelli di classe III, non differivano solo in base alla lunghezza della sequenza (i.e in base al numero di unità ripetute), ma esistevano anche differenze nella struttura dell'unità ripetuta ACAGGGGTGTGGGG, determinate dalla presenza d'inserzioni e mutazioni di basi.

Nel loro insieme questi risultati indicano che differenze vere e significative nella sequenza allelica della VNTR sono correlate a differenze nei livelli di associazione con il DT1. La suscettibilità o la protezione nei confronti della malattia, codificata dal minisatellite VNTR-INS, può essere attribuita tanto alla lunghezza dell'unità ripetuta, quanto alla struttura della sequenza allelica VNTR, o a una combinazione di queste che vanno a formare una particolare configurazione VNTR.

Nel 2010 in uno studio condotto da Durinovic-Bellò caratterizzarono in soggetti affetti da DT1 e in controlli la suscettibilità malattia rispetto minisatellite VNTR-INS, . Soggetti con diabete di tipo I e controlli sani con elevati livelli di proinsulina specifiche per le cellule T sono stati caratterizzati con il polimorfismo del gene VNTR-INS.

Al contrario, i soggetti con un polimorfismo 'protettiva' nel INS-VNTR del gene avevano livelli quasi non rilevabili di proinsulina nelle cellule T. Negli esseri umani, i livelli di espressione di insulina sono regolate,in parte dall' influenza della trascrizione dalla regione del gene dell'insulina (INS-VNTR) associato con la regione promotore del gene proinsulina.

Negli individui con classe 3 avranno Livelli di espressione più alta di proinsulina, a causa

della presenza VNTR, e quindi la probabilità di sviluppare la malattia rispetto alla classe è 3-4 volte minore. Negli individui con classe 1 si ha una bassa espressione di proinsulina a livello delle cellule midollari del timo che reagiscono contro l'insulina perchè è poco espressa e questo porta ad un rischio di ammalarsi (Durinovic-Bello, Wu et al. 2010).

Gene PTPN22

Negli ultimi anni, sono stati trovati numerosi geni e varianti non-HLA e non-INS sono associati con il DT1 (Bottini, Musumeci et al. 2004). Nel 2004, Mediante l'approccio del gene candidato, Bottini e colleghi hanno identificato un'associazione tra il DT1 ed un SNP del gene PTPN22, che mappa in1p13.3-p13.1. PTPN22 codifica per una tirosina fosfatasi specifica delle cellule linfoidi (LYP: Lymphoid Phosphatase). La proteina LYP fa parte della grande famiglia delle tirosin fosfatasi, importante per la regolazione negativa dell'attivazione T-cellulare (Bottini, Musumeci et al. 2004). LYP è espresso nei linfociti e si lega attraverso un motivo ricco di proline al dominio SH3 della proteina CsK (C-terminal Src tyrosine kinase), la quale è un importante soppressore delle kinasi che mediano l'attivazione delle cellule T.

Lo SNPs associato è un polimorfismo +1858C>T localizzato nell'esone 14 del gene PTPN22 (1p13.3), mostra un consistente effetto sulla suscettibilità del DT1 (OR costantemente >1,5 in tutte le popolazioni dove è presente con frequenze apprezzabili) (Smyth, Cooper et al. 2004).

Anche se le ragioni per la diversa associazione con distinte malattie autoimmuni non sono ancora chiare, i dati disponibili puntano alla variante +1858C>T come uno dei pochi polimorfismi non-HLA con effetto rilevante nel rischio ereditario per il DT1. Questa variante causa una sostituzione dell'amminoacido arginina (R) in triptofano (W) nel codone 620 della proteina linfoide tirosin-fosfatasi (Lyp) (Vang, Congia et al. 2005).

Che la variante +1858C> sia primariamente associata con il DT1, e quindi che il suo prodotto proteico Lyp R620W sia direttamente coinvolto nella patogenesi della malattia, è chiaramente indicato da un estensivo risequenziamento del gene e successiva analisi di

associazione dei polimorfismi identificati in un'ampia casistica proveniente dalla popolazione sarda (Zoledziewska, Perra et al. 2008). Questo polimorfismo e quindi la sostituzione aminoacidica R620W nella proteina LYP darebbe luogo a una mutazione con acquisizione di funzione, che la rende un regolatore negativo dell'attivazione dei T linfociti più potente (Vang, Congia et al. 2005).

In particolare, la variante di suscettibilità 620W sarebbe più efficiente da un punto di vista cinetico e si assocerebbe ad una ridotta produzione di interleuchina 2 e ad una ridotta proliferazione T cellulare. Questo studio ha evidenziato che la variante +1858C>T, sebbene rara nella popolazione sarda (con frequenza allelica 0,014) era positivamente associata con il DT1 (P = 3,7 * 10-3) mentre l'aplotipo nel quale questa mutazione è insorta è comune (frequenza aplotipica 0,117) e non è associato con la malattia.

Questa evidenza sperimentale è la prova principe da un punto di vista genetico dell'effetto primario di tale polimorfismo nell'associazione con la malattia. Lo stesso studio non ha confermato l'associazione della malattia riscontrata in altre popolazioni per varianti non +1858C>T (rs2488457, rs1310182, e rs3811021), anche se avevano frequenze significative in Sardegna.

Quindi lo studio di Zoledziewska e colleghi non solo dimostra che la variante +1858C>T è primariamente associata con la malattia ma anche che a livello di tale gene rappresenta l'unica variante associata, escludendo un modello di eterogenità allelica precedentemente ipotizzato in altri studi (Kawasaki, Awata et al. 2006).

Il ruolo primario della sostitituzione la variante Lyp-W620 è anche indicato da studi funzionali, che hanno evidenziato un ruolo critico di tale variante suggerendo che essa causi una acquisizione di funzione fisiologica della fosfatasi wild-type; ovvero la W620 ha un'attività regolatoria negativa più marcata rispetto al più comune allele wild-type R620 (Bottini, Vang et al. 2006).

Gene CTLA4

Dall'analisi di diversi GWAs sono state identificate tre loci IDDM nel cromosoma 2q31-35 (IDDM7, IDDM12, IDDM13) (Copeman, Cucca et al. 1995),(Nistico, Buzzetti et al. 1996). Il locus IDDM12 in 2q33, contiene diversi geni candidati per il DT1 e per altre patologie autoimmuni; il gene CTLA-4 (Cytotoxic T Lymphocyte-associated Antigen 4), ICOS (Inducible T-cell Costimulator) and CD28 (Todd and Wicker 2001) .

Il prodotto del gene CTLA4 fa parte della famiglia delle immunoglobuline ed è espresso nella superfice dei linfociti T CD4 e CD8 attivati . CTLA4 è un inibitore della proliferazione dei linfociti T, promuove l'apoptosi delle cellule T, attraverso l'incremento dell'espressione dell'interleuchina 2 (IL2). Iniziali studi hanno descritto l'associazione del DT1 con uno SNP (A/G) in posizione 49, responsabile della sostituzione amminoacidica Tre17Ala , ma questa associazione non venne replicata in tutte le popolazioni studiate (Nistico, Buzzetti et al. 1996). Ulteriori marcatori rappresentativi del locus IDDM12 hanno mostrato associazione con la malattia, non fornendo un indicazione precisa riguardo all'associazione per il gene CTLA4. Era infatti possibile che essi fossero rappresentativi dell'associazione con altri geni del locus IDDM12, presumibilmente CD28 o ICOS. Nel 2003 il gruppo di Cambridge conferma il coinvolgimento del gene CTLA-4 nella suscettibilità del DT1 e di altre malattie autoimmuni (Ueda, Howson et al. 2003) sebbene l'effetto genetico di questo locus nel determinare il rischio di sviluppare la malattia appaia estremamente modesto (OR=1.1).

Il gene IL2RA/CD25

IL2RA/CD25 è un ottimo gene candidato che codifica per la subunità del recettore dell' interleuchina 2 e del CD25. Rare mutazioni del gene IL2RA/CD25 causano severe forme di patologie autoimmuni (Sharfe, Dadi et al. 1997). Nel 2005 è stata descritta l'associazione di IL2RA/CD25 con il DT1 (Vella, Cooper et al. 2005). In seguito è stato dimostrato che polimorfismi del gene IL2RA/CD25 contribuivano alla patogenesi di altre patologie autoimmuni come la SM (Matesanz, Caro-Maldonado et al. 2007).

Gene CLEC16A

Nel 2007 Todd e colleghi descrissero l'associazione per varianti nucleotidiche nell'introne 19 del gene CLEC16A (KIAA0350), attraverso uno studio di GWAs in popolazioni nord europee, (Todd, Walker et al. 2007). Il gene CLEC16A è localizzato sul cromosoma 16p13, e la funzione non è ancora ben compresa. CLEC16A appartiene alla famiglia delle lectine tipo C, ed è espressa in cellule immunitarie. In particolare, le lectine tipo C hanno una funzione anti-infiammatorio; inibiscono l'attivazione di NFkB, e regolano negativamente la produzione di citochine. L'associazione di questo gene è stata confermata nella popolazione sarda (Zoledziewska, Costa et al. 2009). Due varianti (rs725613 e rs12708716) sono localizzate nell'introne 19 di questo gene e sono in perfetto LD (r2=1). L' allele A di rs725613 è positivamente associato sia con non DT1 (odds ratio = 1.15, one-tail=5.1 × 10-3), ma anche con la SM (odds ratio = 1,21 , one-tail 6.7 × 10-5). Nel loro insieme questi dati forniscono evidente associazione con entrambe le patologie ed un probabile percorso di malattia comune.

IFIH1

IFIH1 (Interferon Induced with Helicase C domain) mappa sul cromosoma 2q24.3 e il codifica per una proteina capace di attivare risposte immunitarie verso cellule infettate da virus (Meylan, Tschopp et al. 2006), (Yoneyama, Kikuchi et al. 2005). Uno studio di GWAs in popolazioni nord europee, che prevedeva la tipizzazione di SNPs non sinonimi, ha evidenziato associazione per lo SNP rs1990760 (A946T) con il DT1 (Smyth, Cooper et al. 2006). Dopo risequenziamento del gene IFIH1 sono state descritte quattro varianti rare (rs35667974, rs35337543, rs35732034 e rs35744605) che possono causare cambiamenti nell' espressione del gene. e non escludono la possibilità di ulteriori varianti con deboli effetti.

DISEGNO SPERIMENTALE DELLO STUDIO

Alla luce dei risultati della ricerca scientifica attuale, si può constatare che il rischio genetico può essere scoperto solo da studi con campioni di dimensioni sempre più grandi. Tuttavia, una percentuale dei segnali genetici ancora sconosciuti ma potenzialmente significativi possono essere deboli a causa di una bassissima frequenza di un allele che altera fortemente la funzione di un gene importante . Queste considerazioni rendono importante identificare varianti altamente penetranti e a bassa frequenza, questi loci potrebbero essere di importanza cruciale in almeno un terzo dei casi (Polychronakos and Li 2011). Il progetto di questa tesi di dottorato si colloca pienamente in questo scenario. Nei 3 anni di dottorato il nostro gruppo di ricerca si è concentrato sull' identificazione dei determinanti genetici del DT1.

L'obiettivo principale di questo progetto è stato quello di identificare le varianti di suscettibilità al diabete di tipo 1 (DT1) in una casistica di soggetti sardi. Partendo dall'obiettivo generale della ricerca proposta, ovvero definire le basi genetiche del DT1 e colmare il deficit di conoscenza sulla variabilità genetica in Sardegna, è stata considerata e genotipizzata con chip array Affymetrix una casistica di 1377 casi di DT1, di cui 51 individui affetti anche da sclerosi multipla, e 1917 controlli sani.

Oltretutto un sub-gruppo di, 279 famiglie trios costituiti da madre, padre e figlio affetto, per un totale di 837 individui sono stati genotipizzati anche con chip Illumina 1M-Duo. Tale chip che ha permesso di studiare circa 1.200.000 polimorfismi tra cui un notevole numero di varianti non presenti nel chip Affymetrix 6.0, consentendo di estendere l'analisi anche ad un sub-set di varianti rare (frequenza allelica < 5%) che non potevano essere caratterizzate in maniera robusta con i chip Affymetrix 6.0. L'integrazione delle due mappe, grazie all'approccio di inferenza statistica, ha consentito l'analisi di una supermappa Affymetrix/Illumina sull'intero data set esaminato.

Con lo scopo di incrementare ulteriormente il numero di varianti identificate, includendo varianti fondatrici sarde non presenti nei chip commerciali, i dati di genotipizzazione sono stati integrati con i dati del sequenziamento a bassa copertura "*low coverage* " dell'intero genoma in 505 individui. In particolare, è stato condotto il sequenziamento di

famiglie o trios ciascuna delle quali risulta composta da un probando (paziente affetto dalla patologia) e dai suoi genitori; inoltre sono stati sequenziati anche 32 trios di individui sani chiamati trios di controlli. Tutti i campioni che sono stati sequenziati sono stati caratterizzati anche con la piattaforma Affymetrix che quella Illumina 1M-Duo.

Durante il terzo anno del dottorato è stato proseguito il sequenziamento, ad oggi sono sequenziati circa 1700 individui di cui 505 anche analizzati. Le sequenze con un'adeguata qualità sono state analizzate per rilevare varianti genetiche, quali SNPs e, sebbene l'analisi iniziale del nostro progetto low *coverage* è stata focalizzata alla ricerca di SNPs, le sequenze generate saranno utilizzate nell'immediato futuro per identificare anche altri polimorfismi, quali inserzioni e delezioni o varianti strutturali.

MATERIALI E METODI

Per gli esperimenti descritti in questa tesi sono state utilizzate procedure standard di biologia molecolare e cellulare. Si è inoltre fatto uso delle più innovative piattaforme di genotipizzazione e sequenziamento, nonché dei più moderni strumenti statistici per l'analisi dei dati e l'inferenza degli aplotipi.

Descrizione dei campioni.

La coorte arruolata nel progetto è costituita da volontari con età compresa tra i 18 e 60 anni chiamati di seguito controlli non imparentati fra loro né con i pazienti e da individui con diagnosi di DT1 di seguito chiamati casi. Sia i casi che i controlli dovevano avere una chiara origine sarda ossia nati in Sardegna e con almeno 3 nonni sardi. Al fine di partecipare a questa ricerca, a ciascun volontario, sia esso un caso o un controllo, è stata richiesta la compilazione e sottoscrizione del consenso informato, del modulo anamnestico e genealogico sulla provenienza dei nonni e sulla familiarità entro il primo grado, per malattie autoimmuni.

I consensi informati sia dei casi che dei controlli sono stati approvati dal comitato etico.

In tutti i pazienti la diagnosi clinica di DT1 era inequivocabile, con una condizione di dipendenza assoluta dalla somministrazione della terapia sostitutiva con insulina esogena fin dall'esordio clinico della malattia.

Sulla base di questi presupposti ed a seguito di una collaborazione decennale tra il Laboratorio di Immungenetica coordinato dal Prof. Francesco Cucca e i maggiori Centri Trasfusionali dell' isola, parallellelamente alla collaborazione con i servizi di diabetologia della Sardegna principalmente la diabetologia pediatrica dell'ospedale Brotzu coordinato dal Dott. Paolo Pusceddu, la Clinica Medica di Sassari sezione di diabetologia coordinata dal Prof. Mario Maioli e in minor misura ma altrettanto assidua collaborano i centri di diabetologia dell'Ospedale di Ozieri, Oristano, Lanusei e Olbia, è stato possibile raccogliere migliaia di campioni DNA (operazione assai difficile o a volte

impossibile) e cosi dar inizio al progetto dal titolo "Sequenziamento e studi di associazione su tutto il genoma per svelare i geni di suscettibilità al DT1". (Figura 6)

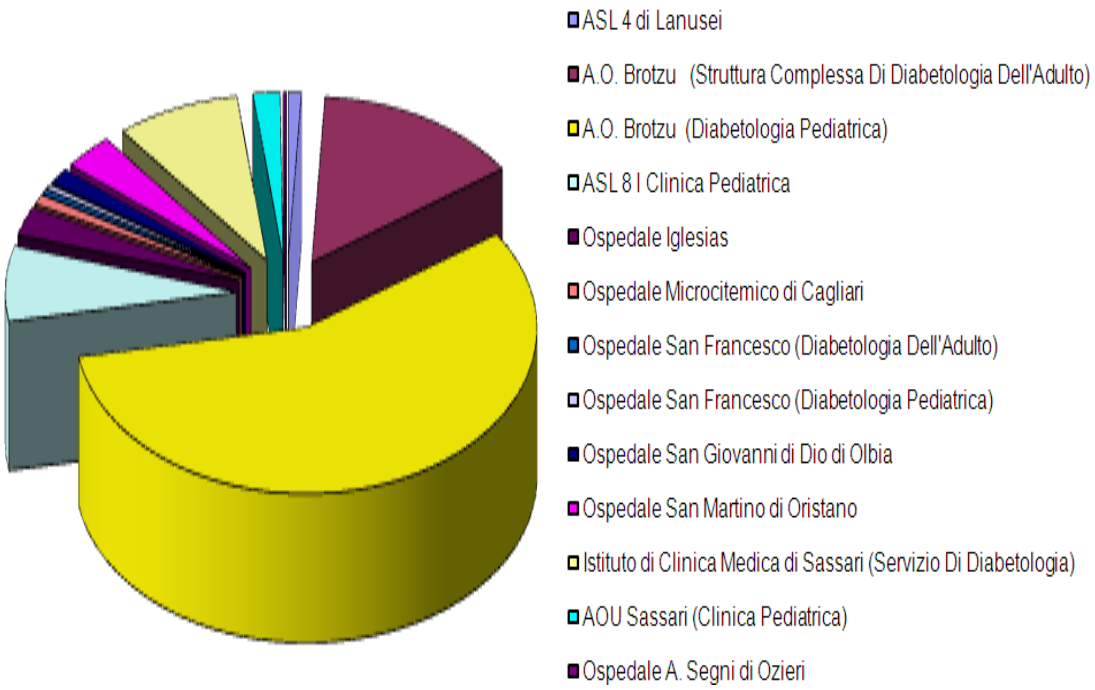


Figura 6. Grafico che rappresenta il contributo nella raccolta dei pazienti di ciascun gruppo clinico.

Estrazione del DNA

Il DNA è stato estratto da 10 ml circa di sangue periferico raccolto in apposite provette contenenti una soluzione anticoagulante (Potassio-EDTA 0.78 M) e conservato a 4ºC. Il DNA è stato estratto, dai campioni di sangue pervenuti mediante il classico metodo salino. La metodica prevede l'aggiunta a 10-12 ml di sangue al Lysis Buffer, che ha funzione di lisare le cellule, successivamente si effettuano due lavaggi con Fisio Buffer (NaCl ed EDTA) per rimuovere i globuli rossi ricchi di emoglobina e per ridurre il più possibile eventuali contaminazioni da ferro, che potrebbero avvenire durante l'estrazione. Dopo che il pellet è ben pulito si risospende in Buffer A (Tris, HCl ed EDTA) e si

procede alla lisi dei globuli bianchi tramite l'aggiunta di SDS (sodio dodecilfosfato) e Proteinasi K, lasciando i campioni in incubazione a 57°C o/n oppure 2 ore a 65 °C.
Il giorno successivo si effettua la precipitazione delle proteine con NaCl soprasaturo, si trasferisce il surnatante in una nuova provetta da 15ml e si ricentrifuga per eliminare eventuali depositi. Il DNA viene ottenuto per precipitazione con isopropanolo assoluto, attraverso un movimento oscillante della provetta; esso si addenserà ed apparirà sotto forma di un flocculo biancastro. Dopo l'asciugatura all'aria del flocculo , si risospende in 300-500 µl di buffer TE (1M Tris-HCl; 0.5M EDTA; pH 8.0) a seconda della grandezza.

Verifica in agarosio del DNA estratto.

Per controllare la riuscita del protocollo di estrazione , un'aliquota di 2 µl del DNA estratto è stata sottoposta ad elettroforesi orizzontale in gel di agarosio al 1% p/v, al fine di verificare la presenza dell'integrità della banda di DNA.

Il gel viene preparato mescolando l'agarosio in polvere, al TAE 0,5X (Tris-Acetato 0.04mM, EDTA 10mM pH=8) riscaldando i reagenti fino ad ebollizione, per sciogliere la soluzione. Ottenuta una soluzione limpida, si aggiunge, sotto cappa chimica, 20 µL di Syber Green 10 mg/mL, e la si fa colare lentamente nell'apparato elettroforetico, dove sono stati inseriti i pettini per la formazione dei pozzetti per il caricamento dei campioni. Durante questa operazione si deve evitare di fare bolle, in quanto potrebbero essere di intralcio per la corsa dei campioni. Si lascia solidificare il gel per circa 30-45 min. I campioni da caricare su gel sono stati allestiti in un volume finale di 10 µL, si aliquotano 2 µl del DNA con 2 µL di colorante , quindi, si porta a volume con acqua sterile . Come marcatore di peso molecolare è utilizzato 1 Kb DNA Ladder (10000, 8000, 6000, 5000, 4000, 3500, 3000, 2500, 1500, 1000, 750, 500, 250 bp della MBI-Fermentas). La corsa elettroforetica è eseguita a 120 Volts in tampone di TAE 0,5 X per 1 ora. Le molecole che migrano nel gel si separeranno in base alla loro carica e al loro peso molecolare e durante la corsa, il colorante indica la posizione dei campioni. Terminata la migrazione

elettroforetica, il DNA è visualizzato mediante esposizione a luce UV e fotografato con l'apposito apparecchio fotografico : l'acido nucleico è visibile grazie all'emissione di luce del Syber Green, una molecola fluorescente non specifica che si lega al solco minore del DNA, intercalatosi tra le basi azotate del DNA quando è eccitato da luce ultravioletta. Dalla foto ottenuta è possibile ricavare informazioni riguardanti lo stato del materiale, ossia è possibile vedere se i campioni sono stati degradati o meno durante il processo di estrazione.

Quantificazione del DNA allo spettrofotometro.

Le basi puriniche e pirimidiniche degli acidi nucleici assorbono radiazioni UV con un picco di intensità massima a 260 nm, è dunque possibile determinare la concentrazione di DNA nella soluzione madre mediante lettura della densità ottica (OD) alla lunghezza d'onda di 260 nm. Il DNA estratto è sottoposto ad analisi allo spettrofotometro (Nanodrop 1000 Thermofischer). Lo strumento viene tarato con una soluzione di TE pH 7.5 (Tris 10 mM-EDTA 1mM). L'unità di lettura dello strumento è l'OD (optical density) e la concentrazione si ricava dalla formula: µg/mL=ODx50xDIL, dove OD è il valore letto dallo strumento, 50 è il coefficiente di correzione per la lettura del DNA allo spettrofotometro (secondo la legge di Lambert-Beer) e DIL è il coefficiente di diluizione del DNA nella cuvetta. Per verificare il grado di purezza del DNA analizzato, bisogna valutare il rapporto tra le assorbanze a 260 nm e a 280 nm e, poiché 1 OD corrisponde a 50 µg/mL, è possibile determinare la concentrazione dei campioni di DNA. Inoltre è stato valutato il rapporto delle densità ottiche a 260/280 nm: se questo rapporto è compreso tra 1.6 e 1.8 la lettura spettrofotometrica corrisponde con buona probabilità alla concentrazione di acidi nucleici. Un rapporto inferiore indica che nella soluzione sono presenti ancora molte proteine e il DNA deve essere ri-precipitato, un rapporto superiore indica che il campione potrebbe essere contaminato. Una volta misurate le concentrazioni di DNA nelle soluzioni madre, per ciascun campione sono state allestite due diluizioni a 50 e a 5 ng/µl da utilizzare nelle analisi molecolari.

METODICHE UTILIZZATE

Genotipizzazione con la piattaforma Affymetrix

I microarray sono costituiti da una serie di oligonucleotidi o sequenze di DNA legati chimicamente ad una superficie solida tramite un processo di stampa o di sintesi in situ. Le sequenze specifiche marcate con fluorocromi fluorescenti vengono ibridate ad una sequenza nota sui chip di DNA ed individuate con uno scanner elettronico che rileva l'eccitazione del marcatore fluorescente.

I supporti solidi usati nei dosaggi su array includono vetrini da microscopia, matrici di agarosio e microsfere.

La specificità dei dosaggi di ibridazione su array è stata ulteriormente potenziata con l'applicazione di un campo elettrico o tramite passaggi di processazione enzimatica come l'estensione o la ligation di primer.

Negli array di ibridazione è possibile individuare il risultato sulla base dell'ibridazione a siti specifici dell'array stesso, caratteristica che consente l'analisi contemporanea di più marcatori SNP in un singolo campione.

I microarray ad alta densità della Affymetrix vengono creati legando centinaia di migliaia di oligonucleotidi ad una superficie solida di silicio secondo una disposizione ordinata.

La genotipizzazione con la piattaforma 6.0 é stata eseguita seguendo il protocollo di Affymetrix (Matsuzaki et al., 2004).

In particolare, il protocollo dei GeneChip arrays si basa sulla discriminazione allelica a seguito di ibridazione del DNA al chip contenente oligonucleotidi di 25 basi locus- ed allele-specifici.

Le sequenze perfettamente corrispondenti si ibridano con maggiore efficienza ai loro oligomeri nell'array e forniscono quindi segnali di fluorescenza più forti rispetto alle combinazioni sonda-bersaglio parzialmente o solo minimamente simili.

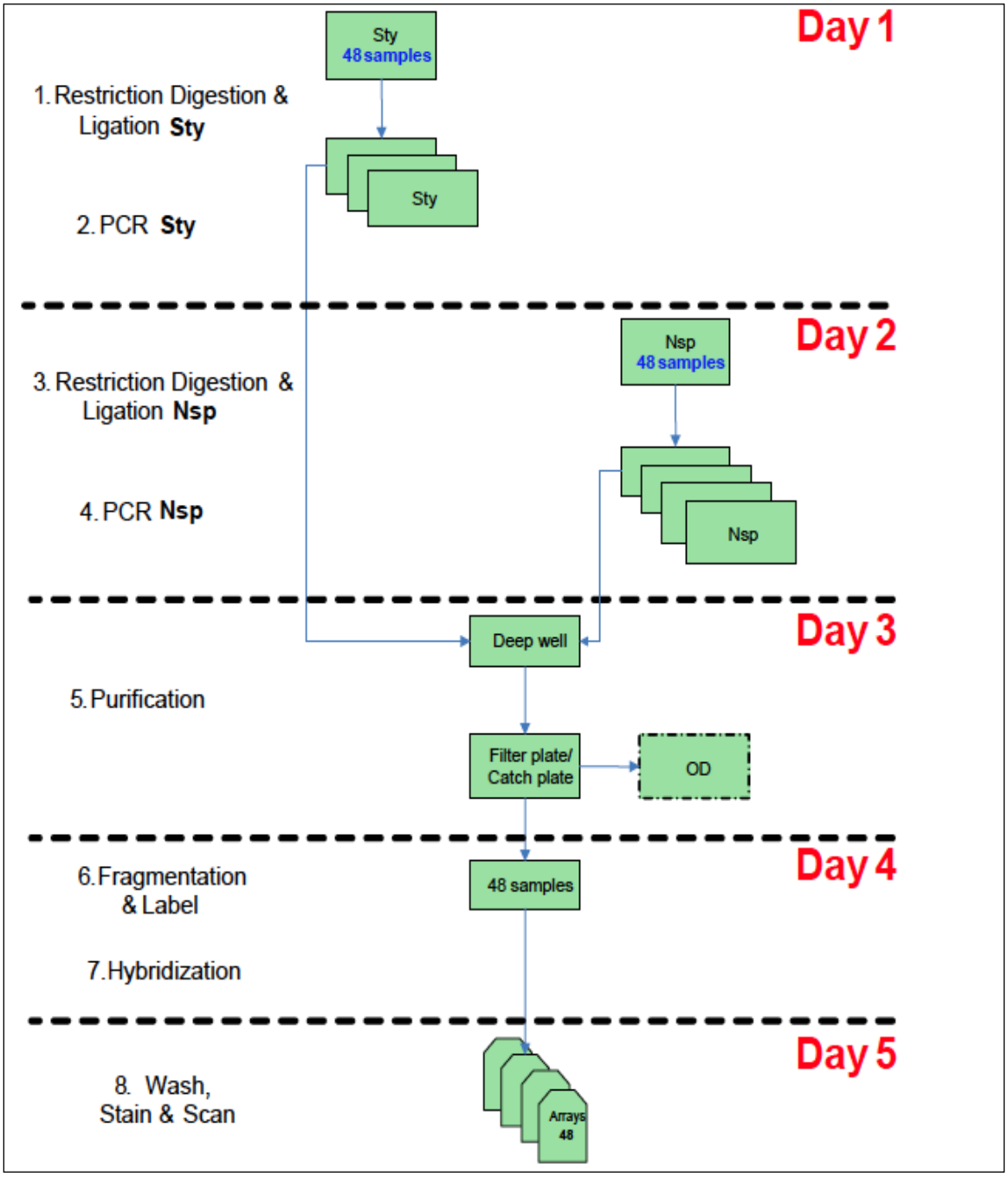


Figura 7 Protocollo di genotipizzazione Affymetrix 6.0.

I segnali di ibridazione vengono quantificati tramite scansione in fluorescenza ad alta risoluzione ed analizzati con un software dedicato.

Brevemente, il protocollo della piattaforma 6.0 prevede una riduzione della complessità del DNA genomico attraverso digestione con endonucleasi di restrizione, appropriate per il numero di SNPs da interrogare, StyI ed NspI per i chips; i frammenti ottenuti,compresi tra 400 e 800 bp, sono stati selezionati per la ligazione degli adattatori. Dopo l'amplificazione ed un'ulteriore frammentazione con DNasi I, il DNA è stato marcato, ibridato sul chip, e si è proceduto alla scansione per la discriminazionedegli alleli di ciascuno SNPs.

Genotipizzazione con la tecnologia Illumina 1M-DUO

La genotipizzazione mediante il chip 1M Duo include 1.200.000 varianti di cui più di 900.000 sonde per SNPs e circa di 200.000 sonde per l'identificazione di CNV. Il protocollo l'amplificazione di 400 ng di DNA, tramite amplificazione di tutto genoma (Whole Genome Amplification, WGA), e in seguito la frammentazione.

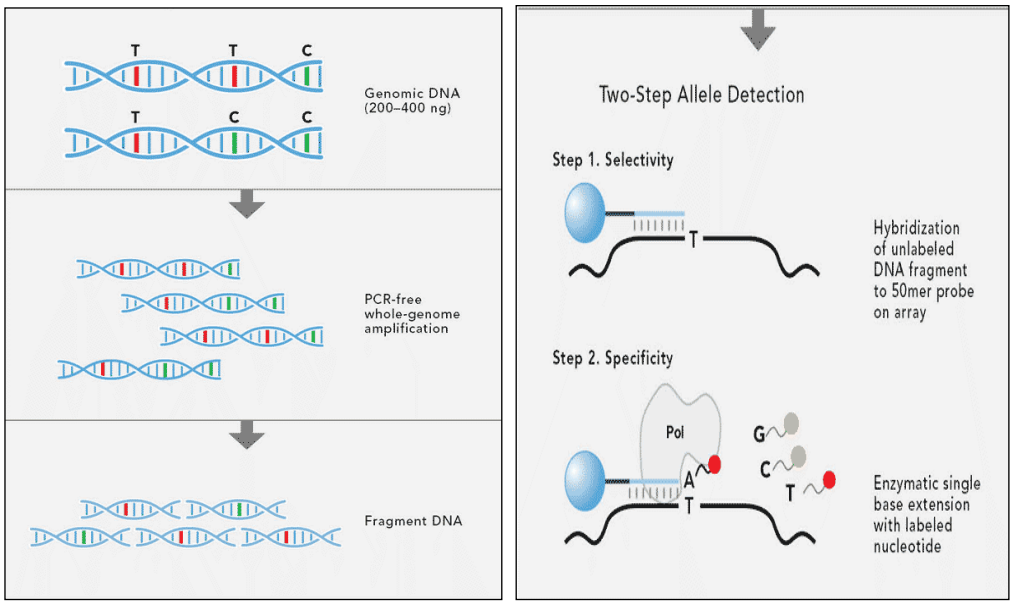


Figura 8 Protocollo di genotipizzazione Illumina 1M-DUO

I frammenti di DNA si ibridizzano a 50pb specifiche ad ogni locus e si attaccano covalentemente a una delle 1.100.000 biglie immobilizzate sulla superficie dello chip. Tutto il DNA in eccesso viene eliminato con un lavaggio, e l'amplificato viene incorporato in un chip. Nella discriminazione allelica vengono utilizzati oligonucleotidi, di 50 pb ciascuno, per identificare gli SNP che sono specifici per ogni allele di ciascun sito (Allele specific Oligo o 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. Il beadchip viene scannerizzato con Illumina iScan che utilizza un laser per eccitare la base estesa che essendo marcata con un fluoroforo emette fluorescenza.

Genotipizzazione con la tecnologia TaqMan

L'utilizzo di oligonucleotidi allele-specifici come sonde per l'ibridazione è alla base dei moderni metodi automatizzati di genotipizzazione degli SNPs .

La metodica discriminazione allelica usando il metodo della 5' nucleasi si propone per efficienza, sensibilità e specificità come uno tra i metodi principali per la genotipizzazione degli SNPs.

Tale strategia (che implica l'ibridazione con oligonucleotidi allele-specifici e la determinazione attraverso il trasferimento di energia di risonanza fluorescente, detta FRET) si basa sull'utilizzo di un oligonucleotide innovativo, il quale risulta essere complementare alla sequenza bersaglio da amplificare nella regione del polimorfismo d'interesse ed è in grado di appaiarsi tra i due primers forward e reverse. La sonda è in grado di riconoscere specifiche sequenze nel DNA per la complementarietà delle basi, e di generare un segnale sequenza specifico, che viene rivelato durante la PCR, permettendo quindi l'identificazione di una variazione puntiforme. Con la

discriminazione allelica è, infatti, possibile identificare forme differenti dello stesso gene che differiscono per un solo SNP. 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. In primo luogo, lo spettro di emissione della molecola fluorescente "donatore" (nel caso della metodica in esame, viene chiamata Reporter) deve sovrapporsi con la lunghezza d'onda d'eccitamento della molecola "accettore" (Quencher); in secondo luogo, le due molecole devono trovarsi molto vicine fra loro altrimenti il trasferimento di energia cade rapidamente all'aumentare della distanza.Tale metodica fa parte dei metodi definiti omogenei, in quanto non presenta fasi di separazione e l'andamento della tipizzazione dei campioni può essere monitorato in tempo reale durante la PCR. In essa sono contemporanee due fasi che nella maggior parte delle altre strategie di indagine risultano distinte: la PCR e i processi di analisi post-PCR. Questo consente una notevole riduzione dei tempi di esecuzione e di analisi. La discriminazione allelica attraverso il TaqMan è completamente automatizzabile e presenta un'altissima sensibilità e affidabilità.

Sequenziamento Hiseq

Il sequenziamento esteso a tutto il genoma é stato eseguito con le piattaforme Genome Analyzer IIx ed Hi-Seq (Illumina). Le librerie di DNA genomico sono state generate in accordo con le indicazioni della Illumina, con alcune modifiche nel protocollo . Il DNA genomico é stato frammentato in maniera random, mediante nebulizzazione o sonicazione (Covaris S, Applied Biosystems), in frammenti sotto 800 bp e successivamente le estremità 3'e 5' dei frammenti sono state riparate e fosforilate. I frammenti di DNA riparati sono stati adenilati in 3´ con una DNA polimerasi Klenow exo- (New England BioLabs) 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 bp, sono stati caricati su gel di agarosio al 2%, successivamente vengono purificati (Qiagen Gel Extraction Kit) e successivam preamplificati mediante PCR, utilizzando dei

primers (IDT) compatibili con gli adattatori.

Dopo purificazione degli ampliconi (Qiagen Gel Extraction Kit), la concentrazione e la distribuzione dei frammenti delle librerie sono state determinate mediante corsa su Chip DNA 1000 nel Bioanalyzer 2100 (Agilent Technologies).

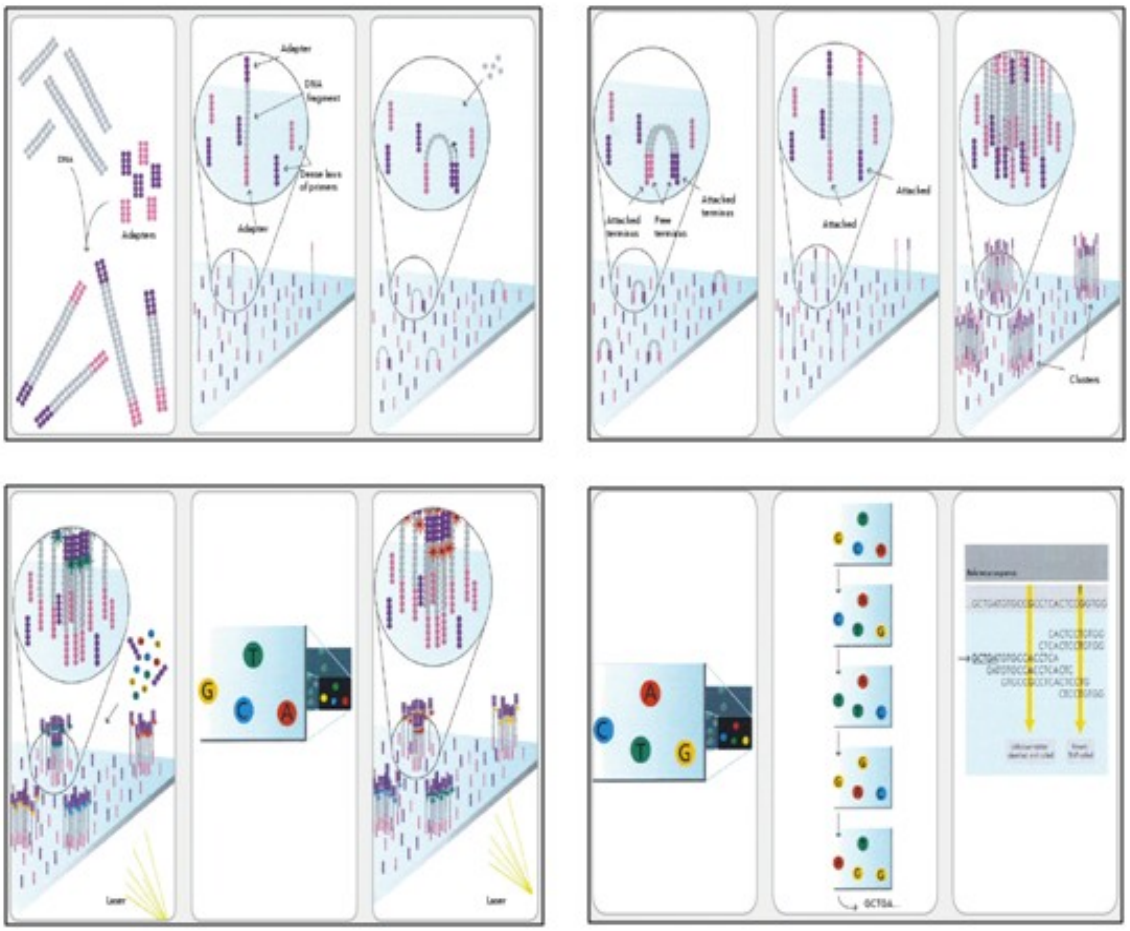


Figura 9 Protocollo di sequenziamento con Hi-Seq (Illumina)

Le librerie sono state ibridizzate e amplificate sulla superficie della flow-cell mediante bridge amplification, formando i clusters, quindi sequenziate con il GAIIx, 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 in corse da 202 basi, ottenendo un *coverage* medio di 3-4X.

SCELTA DEL TIPO DI SEQUENZIAMENTO DA ESEGUIRE

Sebbene il sequenziamento ad alto *coverage* (~30X) di singoli genomi sia in grado di rilevare tutte le varianti, comuni e rare, presenti negli individui sequenziati, i costi per la sua attuazione sono talmente elevati che può essere preso in esame solo un numero limitato di individui. Per ovviare a questo problema è stato preso in considerazione altre strategie alternative che forniscano informazioni sull'intero genoma di migliaia di individui; si è attuato il sequenziamento a basso *coverage* (~2-4X) 1.500 individui e, per fare questo nella maniera più economica, si è deciso di adottare la strategia che combina le tecnologie di sequenziamento shotgun con gli stessi strumenti statistici usati per l'imputazione dei genotipi negli studi di associazione. Tale scelta è nata anche a seguita della pubblicazione di Li et al dove riportano una simulazione di sequenziamento supponendo un badget di spesa fisso, sequenziando 67 individui ad alto *coverage* (30X) e 1.000 individui a basso *coverage* (2X). Entrambi i metodi hanno dimostrato un potere eccellente (~100%) nel rilevare varianti con MAF>5%, ma il sequenziamento a basso *coverage* o *low-pass* mostra un potere maggiore nel rilevare le varianti meno comuni (MAF=0.5-5.0%). La maggior parte delle varianti con frequenza<0.5% sono risultate non rilevabili con entrambi gli approcci: con il deep sequencing, infatti, la maggior parte delle varianti di interesse non sono polimorfiche nei 67 individui selezionati per il sequenziamento; con l'analisi *low coverage* non ci sono abbastanza copie di ciascun aplotipo per poter assemblare efficacemente le informazioni tra i campioni Per le varianti identificate, l'accuratezza dei genotipi seppur ridotta nell'analisi lowpass, era ancora notevole; per esempio, per le varianti con frequenza>1%, l'accuratezza del sequenziamento low-pass é sempre maggiore del 99.5% per tutti i siti e dell'89.5% per i siti eterozigoti, i quali sono più difficili da identificare in maniera corretta. In tutti i casi considerati, il sequenziamento low-pass di 1.000 individui fornisce maggiori informazioni rispetto al sequenziamento di 67 individui ad alto coverage; per esempio, per varianti con frequenza 0.5-1.0%, 1.0-2.0%, 2.0-5.0% e 5.0% o maggiore, il sequenziamento 2X di 1.000 individui fornisce una effettiva dimensione del campione di 567, 761, 883 e 978 individui, tutte sostanzialmente maggiori dei 67 individui che possono essere esaminati con il deep sequencing.

SEQUENZIAMENTO DI 67 INDIVIDUI AL *COVERAGE* 30X				
Frequenza dell'allele minore	0.5-1.0%	1.0-2.0%	2.0-5%	>5%
Proporzione di siti rilevati	59.3%	90.1%	96.9%	100.0%
Accuratezza genotipizzazione	100.0%	100.0%	100.0%	100.0%
solo siti eterozigoti	100.0%	100.0%	100.0%	100.0%
Correlazione con il valore vero (r2)	99,8%	99.9%	99.9%	100.0%
Effettiva dimensione campione (n*r2)	67	67	67	67
SEQUENZIAMENTO DI 1000 INDIVIDUI AL *COVERAGE* 2X				
Frequenza dell'allele minore	0.5-1.0%	1.0-2.0%	2.0-5%	>5%
Proporzione di siti rilevati	79.6%	98.8%	100.0%	100.0%
Accuratezza genotipizzazione	99.6%	99.5%	99.5%	99.8%
solo siti eterozigoti	78.8%	89.5%	95.9%	99.8%
Correlazione con il valore vero (r2)	56.7%	76.1%	88.2%	97.8%
Effettiva dimensione campione (n*r2)	567	761	882	978

Tabella 1: Tabella riassuntiva con i criteri di scelta del tipo di sequenziamento da eseguire.

In conclusione, il low pass shotgun sequencing sembra una promettente alternativa al deep sequencing di un piccolo numero di individui ad alto *coverage* poichè ci permetterà di rilevare più varianti (con MAF>0.5%) e fornirà una più numerosa dimensione campionaria per i test di associazione che utilizzano le varianti rilevate.

ANALISI STATISTICA

Analisi dei genotipi

I microchips commerciali utilizzati per gli studi di associazione genome-wide sono disegnati per fornire eccellente copertura degli SNPs comuni, mediante genotipizzazione di tag-SNPs che sono proxies per varianti causali comuni, ma hanno solo un limitato potere nel catturare le varianti rare (MAF<5%), e non valutano direttamente il contributo di corti polimorfismi (inserzioni o delezioni).

Tutta la casistica di 1377 casi e 1917 controlli è stata caratterizzata attraverso la piattaforma Affymetrix [http://www.affymetrix.com/partners_programs/programs/developer/tools/powertools.affx] .

I genotipi sono stati attribuiti mediante l'algoritmo Birdseed .

Tutti i campioni sono stati tipizzati con Affymetrix e solo un sub gruppo di 837 individui è stato caratterizzato anche con chip Illumina 1M Duo.

I genotipi sono stati assegnati attraverso il "GenomeStudio Genotyping Module".

Tali algoritmi assegnano i genotipi ad ogni individuo effettuando una classificazione a 3 classi mediante l'utilizzo di conoscenza a priori.

Prima di applicare il test di associazione sono stati utilizzati stringenti e rigorosi controlli di qualità sia dei campioni che degli SNPs.

Sono stati eliminati gli individui con:

1. Contrast quality control (cQC) < di 0.4 e controllato la media del cQC per piastra, genotipizzando campioni in piastre in cui la media del cQC sia < 1.7;
2. Call rate < del 90%;
3. Individui in duplicato e con relazione di parentela - software Relative Finder [http://genome.sph.umich.edu/wiki/RelativeFinder];
4. Individui con sesso discordante (sesso reale versus sesso assegnato dagli Affymetrix

Power tools) .

Inoltre sono stati applicati controlli di qualità anche sui marcatori, eliminando tutti gli SNPs con:

1. Minor frequency allelic (MAF) < 5%;
2. Call rate < 90% nei soli casi o nei soli controlli;
3. Differenza di call rate tra casi e controlli superiore al 5%;
4. Deviazione eccessiva dall'equilibrio di Hardy-Weinberg nei controlli;
5. Tasso di errore superiore al 0.008 negli individui in duplicato.

Per incrementare il numero di varianti testate è stato utilizzato l'approccio d'inferenza aplotipica, descritto sopra in dettaglio, per la ricostruzione di marcatori non direttamente tipizzati.

I genotipi sono stati ricostruiti probabilisticamente utilizzando sia dati disponibili nei database pubblici, quali il progetto 1000 Genomi che dati di sequenza di 505 individui sardi che ci hanno consentito di creare un pannello di referenza sardo.

Analisi delle sequenze

Ci siamo concentrati sulle varianti fondatrici sarde e varianti rare, fin ora non indagate in quanto non presenti nei chip commerciali.

Una limitazione degli attuali GWAs è che, sebbene interroghino milioni di SNPs, direttamente genotipizzati o imputati, sparsi nell'intero genoma, questi sono prevalentemente varianti ubiquitarie e non contengono varianti rare o fondatrici che sono frequenti solo in alcune popolazioni ma non in altre.

I progetti Hap Map e 1000 Genomi prendono in esame le stesse popolazioni riducendo così la rappresentatività delle varianti geniche identificabili.

Per colmare questa lacuna, grazie all'utilizzo di sequenziatori di nuova generazione (Illumina) a nostra disposizione, sono stati ad oggi risequenziati sull'intero genoma a bassa risoluzione (circa 3-4 x) circa 1700 individui che includono anche un numero

elevato di affetti DT1 appartenenti a famiglie trios interamente genotipizzate sia con chip Affymetrix 6.0 che con chip Illumina 1M-Duo.

L'analisi delle sequenze è stata condotta dal gruppo "Advanced Genomics Computing Technology" del Centro di Ricerca, Sviluppo e Studi Superiori in Sardegna (CRS4), coordinato dal dott. Christopher M. S. Jones, ad oggi il più grande centro di calcolo in Italia. Il CRS4 è localizzato al Parco tecnologico di Pula ed è strettamente connesso con le attività della piattaforma di sequenziamento e genotipizzazione; insieme costituiscono, al momento, il centro di sequenziamento di maggior capacità produttiva in Italia.
La dotazione tecnologica, solo per il sequenziamento, è costituita da 3 piattaforme Illumina HiSeq 2000 e 2 Illumina GAIIx, e dalla piattaforma di genotipizzazione Affymetrix.
L'ampia infrastruttura di calcolo possiede un supercomputer di 47 teraflops e un disco di 1,5 petabyes di memoria disco indispensabile per la registrazione e l'analisi dei dati. Il sequenziamento dei campioni in corse *paired-end* di 240 basi con il GAIIX e 204 basi con l'Hi-Seq ha prodotto, in media, ~9 milioni di basi (9 GB) ad alta qualità (ciascuna con Q20 o maggiore) per campione, corrispondenti a 37.5 milioni di *clusters* utilizzabili, e ~15 milioni di basi (15 GB) ad alta qualità per campione (ciascuna con Q20 o maggiore), corrispondenti a 74 milioni di *clusters* utilizzabili, rispettivamente, dopo il mappaggio delle *reads* al genoma di referenza, la ricalibrazione del *Phred score* delle basi mappate e la rimozione dei duplicati.
Questi valori corrispondono ad un *coverage* per campione di 3-4 X, per corse della durata di 10-12 giorni eseguite con GAIIx e Hi-Seq, rispettivamente, *coverage* sufficienti per uno studio *low-coverage* come questo.
Le piattaforme di sequenziamento di nuova generazione, come i sequenziatori GAIIx e Hi-Seq (Illumina), permettono di leggere due *strand* e producono sequenze corte dell'ordine di 100-200 basi, chiamate *reads*.
La qualità delle sequenze é stata valutata con una serie di metriche: percentuale dei *clusters* che superano i filtri di qualità Illumina, numero di *reads* per lane, numero di basi che vengono mappate in maniera univoca sul genoma, la dimensione dei frammenti, etc.
Le *reads* sono state mappate sul genoma di riferimento mediante un algoritmo di

allineamento per *short reads* (Burrows-Wheeler Alignment tools) e, mediante un programma sono state rimosse le eventuali *reads* duplicate. Ad ogni base è stato assegnato un parametro di qualità (*Phred score* originale), in base ai segnali di intensità luminosa assegnati dal sequenziatore alle stesse; il *Phred score* originale, dopo il mappaggio delle *reads*, è stato ri-calibrato (*Phred score* empirico) in base al confronto delle sequenze delle *reads* con il genoma di riferimento ed al tasso di errore rilevato dopo aver raggruppato le basi secondo il *Phred score* originale.

Il *Phred score* descrive la probabilità di errato assegnamento della base. Per esempio, se su 100 basi con un *quality score* originale di 40 (1 errore stimato ogni 10.000 basi sequenziate) é osservata 1 lettura errata e 99 letture corrette, a queste basi verrà riassegnato un *quality score* empirico di 20, essendo il *Phred score* il logaritmo negativo della probabilità che una base venga letta in maniera errata: -10 log10[1/(1+99)]=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).

Quando le basi hanno mostrato un eccesso di *coverage* in relazione alla distribuzione generale sono state eliminate e successivamente sono stati caratterizzati i genotipi.

Questo metodo è un algoritmo stilato dal gruppo "Advanced Genomics Computing Technology", esso mette in relazione le *reads* lette in individui con aplotipi simili e le relazioni familiari degli individui sequenziati, ricostruendo la trasmissione dei cromosomi all'interno delle famiglie. La lista degli aplotipi del pannello di referenza sardo è stato salvato in formato standard vcf (http://www.1000genomes.org/wiki/Analysis/vcf4.0).

Inferenza statistica

Per poter identificare e caratterizzare la variabilità genetica del DT1 il gruppo degli statistici coordinato dalla dott.ssa Serena Sanna, ricercatrice del CNR, ha utilizzato la strategia dell'inferenza genetica per l'analisi dei dati derivanti dai GWAs e dalle nuove piattaforme di sequenziamento ad alta processività.

La base di questa strategia è stata descritta da circa due anni dal gruppo di Abecasis nella pubblicazione dal titolato "Genotype Imputation" (Li, Willer et al. 2009).

In questo articolo vengono descritte le strategie statistiche partendo da informazioni genetiche di tipo "incompleto" o meglio indicano i modi con cui gestire una grossa mole di dati partendo da un gruppo di individui e fornendo informazioni utili su molte altre varianti genetiche inosservato negli stessi individui.

L'imputazione è ormai uno strumento essenziale per l'analisi del genoma, questa tecnica permette di avere la tipizzazione a marcatori genetici che non sono direttamente genotipizzati; essa è particolarmente utile per combinare i risultati tra gli studi che si basano su piattaforme di genotipizzazione diverse, perchè hanno la peculiarità di aumentare la potenza delle genotipizzazioni di ciascun individuo. la qualità dell'imputazione è definita dal parametro RSQR (Russell Square Quality Representatives).

Tutti gli SNPs con RSQR < 0.3 sono stati esclusi.

Questo valore varia tra 0 e 1. La soglia del 0.3 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 più "confidenza" quelli con valori al di sopra di 0.8.

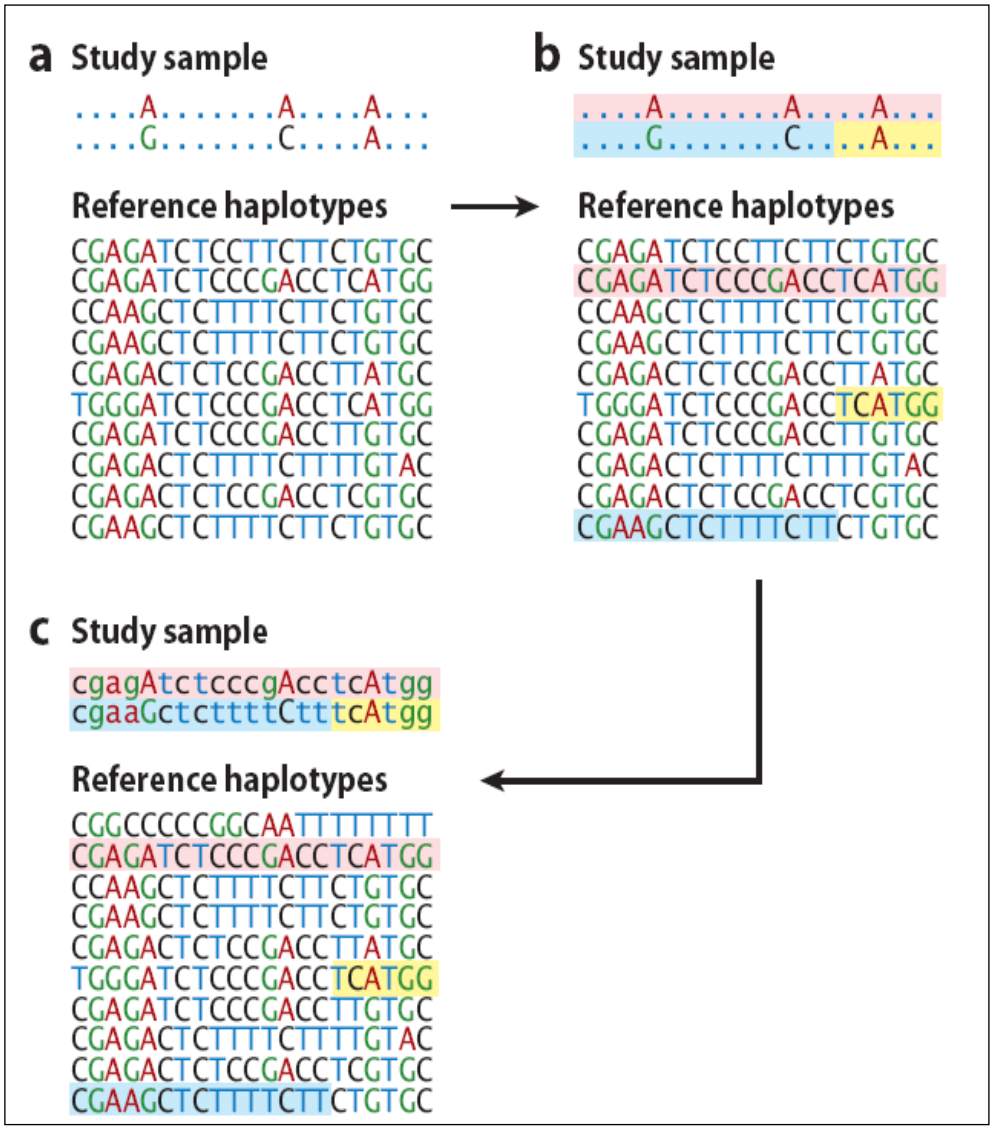


Figura 10 Imputazione dei genotipi in un campione di individui apparentemente non correlati.(a) I dati osservati consistono in genotipi ad un modesto numero di marcatori genetici in ogni campione oggetto di studio e, inoltre, di informazioni dettagliate su genotipi (o aplotipi) per un campione di riferimento. (b)Il processo di inferenza è capace di identificare le regioni del cromosoma condivisa tra un campione di studio e individui del panello di riferimento. Quando un campione viene confrontato con aplotipi nel pannello di riferimento HapMap, si possono identificare i più grandi segmenti condivisi. (c)I genotipi osservati e aplotipo condivisione delle informazioni sono stati combinati per compilare una serie di genotipi inosservato nel campione di studio.

Imputazione: analisi congiunta dei dati di genotipizzazione e sequenziamento.

L'obiettivo è stato quello di integrare i dati di genotipizzazione derivanti dai chip commerciali Affymetrix e Illumina con i dati generati dal sequenziamento di 505 individui sardi e di estendere quest'ultimi in tutta la casistica esaminata, di 1377 casi DT1 e 1917 controlli.

L'imputazione dei dati di sequenza della popolazione sarda sulla super mappa Affymetrix-Illumina è stata condotta con il software Impute2 [http://mathgen.stats.ox.ac.uk/impute/impute_v2.html].

Durante la fase d'imputazione sono stati utilizzati differenti sets di marcatori polimorfici, per i quali è presente un overlap di SNPs, è stato quindi necessario integrare i dati dei tre sub-sets di dati (genotipizzazione Affymetrix, Illumina e dati di sequenziamento) al fine di identificare SNPs indipendenti, e ricostruire gli aplotipi..

I genotipi mancanti sono stati ottenuti attraverso il software Impute2 mettendo in fase i genotipi osservati. Il software ha quindi permesso d'imputare dati che necessitano di combinare più pannelli di referenza contenenti differenti sets di SNPs.

L'analisi di inferenza utilizzando i marcatori in comune tra il pannello di referenza ed i campioni dello studio identifica brevi tratti di un aplotipo, o di aplotipi incompleti ma simili nei due set di dati, allo scopo di estendere probabilisticamente ma con notevole precisione i genotipi a varianti che non sono state direttamente genotipizzati in tutti i campioni dello studio.

L'approccio da noi scelto per il sequenziamento *low-coverage* di un elevato numero individui, è risultato efficace per l'individuazione dei genotipi e per la successiva propagazione degli stessi ai parenti degli individui sequenziati, al fine di ricostruire gli aplotipi, attraverso il confronto dei dati delle sequenze con dati genotipici già disponibili.

Test di associazione

Per entrambi i set di dati, ottenuti con l'imputazione HapMap e 1000 Genomi, è stato eseguito il test di associazione e calcolato il chi quadro con correzione per sottostruttura utilizzando il software Eigenstrat.

Per il cromosoma X sono stati eliminati gli SNPs per cui il numero di maschi eterozigoti era superiore del 3% e per i quali vi era una differenza eccessiva tra il Pvalue calcolato per i soli maschi e quello calcolato per le sole femmine. Infine è stato utilizzato il software incluso nel pacchetto Eigensoft - SmartPCA per l'analisi delle componenti principali sugli SNPs autosomici per identificare una eventuale sottostruttura di popolazione ed eliminare gli individui outliers [http://genepath.med.harvard.edu/~reich/Software.htm].

Sono stati infine testati per associazione 1,214,669 SNPs direttamente genotipizzati, derivanti dall'integrazione della mappa Affymetrix- Illumina (923,299 SNPs Illumina e 291,370 SNPs Affymetrix) e 14,386,887 imputati, in particolare 7,690,836 con il pannello di referenza 1000 genomi e 6.696,051 con il pannello di referenza sardo.

Sono stati, quindi, valutati visivamente i plot di discriminazione allelica di ciascun SNP che mostrasse un Pvalue <10-5, eliminando tutti i marcatori con un attribuzione errata dei genotipi.

RISULTATI E DISCUSSIONE

I nostri dati indicano che l'imputazione ha funzionato correttamente e che il panello di referenza sardo ha migliorato l'estrapolazione dell'informazione genetica nella casistica esaminata.

L'analisi di associazione di circa 15 milioni di marcatori direttamente genotipizzati con affymetrix ed Illumina ed imputati con il pannello 1000 Genomi e il pannello di referenza sardo, costituito da dati di sequenza di 505 individui , in 1425 casi e 2487 controlli ha evidenziato un discreto potere statistico dello studio.

L'analisi di associazione condotta attraverso l'imputazione con il pannello di referenza sardo è stata più efficiente rispetto a quella condotta con il pannello 1000 genomi.

Oltre il locus HLA, sono stati confermati anche ulteriori geni noti, quali *PTPN22, CTLA4, IL2RA* e *CLEC16A* e *INS*, come mostrato dal Manhattan plot in figura 11.

E stato ottenuto il valore del *genomic control* inferiore a 1, e il parametro Lambda di 1.074, come evidenziato nel Q-Q plot (riquadro in alto della figura 11).

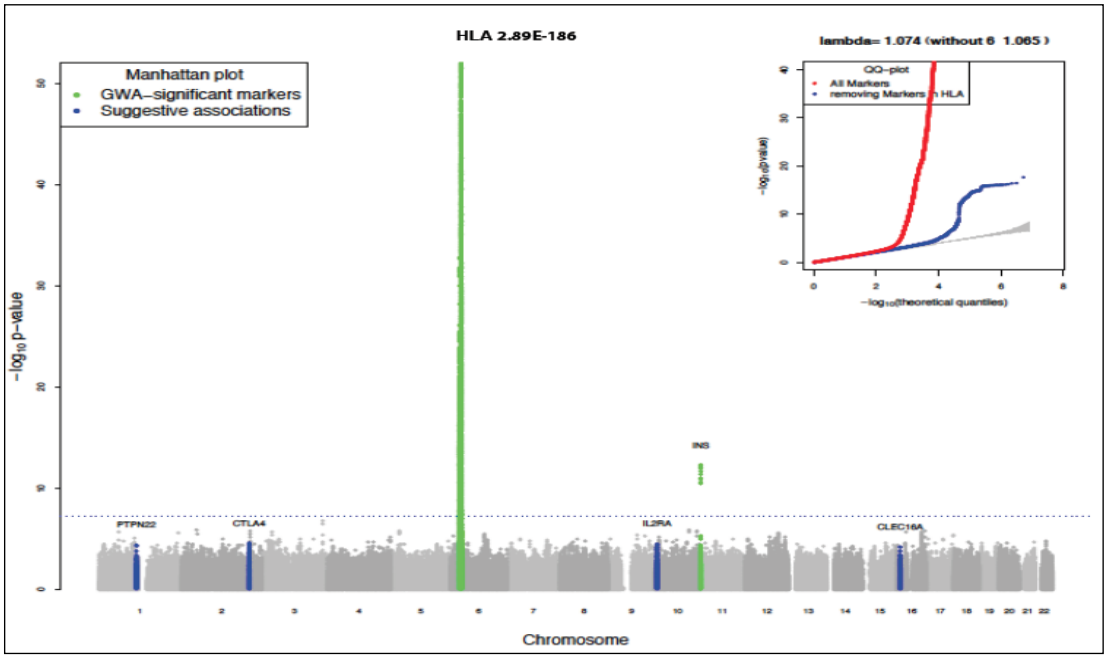


Figura 11 Test di associazione utilizzando il pannello di referenza sardo.

Nella tabella successiva è rappresentata una sintesi dei geni confermati.

	Pannello di refenza sardo		***Pannello dei 1000 Genomi***	
Geni noti	**RSQR**	***p-value***	**RSQR**	***p-value***
PTPN22	0,95	5,27 E -05	0,77	5,74 E -05
CTLA4	0,98	3,96 E -05	0,99	3,72 E -05
HLA	0,93	2,89 E -186	0,45	1,59 E -178
IL2RA	0,99	3,45 E -05	0,99	4,03 E -05
INS	0,55	5,53 E -13	0,34	7,32 E -08
CLEC16A	0.85	6,65 E -05	0,89	1,10 E -03

Tabella 2: Qualità e significatività dell'imputazione con l'uso del pannello di referenza sardo e il pannello di referenza dei 1000 Genomi.

La tabella 2 riassume gli indici di qualità dell'imputazione e la significatività a ciascun locus, con il pannello di referenza 1000 Genomi versus il pannello di referenza generato da dati di sequenziamento nella popolazione sarda. Dai risultati di quest'analisi si evince un netto miglioramento della qualità dell'imputazione effettuata utilizzando i dati di sequenza generati nella medesima popolazione, definita dal parametro RSQR.

Inoltre ha aumentato il potere statistico dello studio, definito dall'incremento della significatività statistica ai geni replicati. Risultato particolarmente evidente al locus *INS* (gene dell'insulina).

Come riportato in tabella 2, considerando il locus *INS*, il parametro RSQR aumenta da 0,34 a 0,55, con imputazione 1000 Genomi e con imputazione del pannello di referenza sardo, rispettivamente.

In generale 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 7 232 174 SNPs con RSQR > di 0.3 (media RSQR = 0.81) imputati attraverso il pannello di referenza sardo, rispetto a 6 242 682 SNPs con RSQR > di 0.3 (media RSQR = 0.76) imputati con

il pannello 1000 Genomi.

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 Genomi (tabella 3).

	Pannello referenza sardo		***Pannello di referenza dei 1000 Genomi***	
MAF	***QC delle varianti imputate***	***RSQR***	***QC delle varianti imputate***	***RSQR***
1% - 3%	1.286. 649	0.74	888 455	0.66
3% - 5%	784. 118	0.81	758 727	0.78
> 5%	5.161 407	0.89	4.595.500	0.86

Tabella 3 Qualità dell'imputazione e varianti imutante con l'uso del pannello di referenza sardo e dei 1000 Genomi in relazione alla MAF.

I risultati ottenuti indicano che il metodo definito dall'integrazione di dati di genotipici integrati da dati di sequenza generati nella medesima popolazione dei campioni dello studio, è capace di aumentare il potere statistico dello studio.

Inoltre i nostri dati indicano che il sequenziamento di un numero sempre più elevato di campioni permette di aumentare il numero di varianti testate.

Infatti con il sequenziamento di 66 individui sardi erano stati identificati solo 8.23 milioni di SNPs. L'analisi di 1147 campioni ha consentito l'identificazione di 15.57 milioni di SNPs, aumentando lo spettro delle varianti testate per associazione con la malattie con l'inclusione di varianti rare.

Lo scopo ultimo è avere creare un pannello di riferimento costituito da 3000 individui utilizzabile per l'inferenza statistica che potrà fornire una continua e nuova risorsa per studi di associazione che saranno condotti sia nella popolazione sarda che piu in generale

in popolazioni europee.

Appare anche evidente che la casistica fin qui raccolta e analizzata non sia ancora sufficiente per identificare varianti nuove.

In tal senso si rileva che la collaborazione di alcuni centri clinici diabetologici è stata ampiamente inferiore rispetto alle attese e alle possibilità. Questo spiega come per una malattia relativamente più rara, quale la sclerosi multipla, sia stata raccolta una casistica più che doppia rispetto a quella raccolta nel caso del DT1. E' quindi necessario aumentare il potere statistico dello studio, attraverso collaborazioni con tutte le strutture sanitarie che nella regione Sardegna gestiscono i pazienti affetti da DT1, cercando di superare le difficoltà, logistiche o di altra natura, del passato. Parallelamente, uno studio caso controllo in cui si analizzino un maniera congiunta i campioni con DT1 e quelli sclerosi multipla (4900 individui finora raccolti) comparati con 4500 controlli aumenterà notevolmente il potere statistico di identificare varianti condivise in entrambe le malattie. Tali varianti sono attese in base ai dati epidemiologici e genetici, indicanti che in Sardegna la frequenza di DT1 nei pazienti con sclerosi multipla e 5 volte quella della popolazione generale e che la regione HLA spiega solo in parte la co-morbidità di queste due patologie. Ci attendiamo quindi che almeno una variante con un ruolo più generale nell'autoimmunità (e in particolare nei confronti sia del DT1 che della Sclerosi Multipla) si trovi a frequenze uniche in Sardegna e abbia un effetti genetici non banali che potrebbero essere identificati dall'analisi congiunta della nostra casistica.

Le informazioni derivanti da tali analisi potrebbero essere rilevanti in senso terapeutico per una progettazione di farmaci razionale, basata sulla delucidazione delle cause e sulla comprensione della patogenesi delle malattie, Una volta che siano state identificate le proteine codificate dai geni responsabili di una malattia e sia stata compresa la loro funzione anormale, i farmaci possono essere progettati in modo da stimolare, inibire, o sostituire tale funzione. L'identificazione delle variazioni del genoma umano potrà infine consentire ai clinici di sotto- classificare le malattie e di adattare le terapie in modo che siano più appropriate per ogni singolo paziente. Le possibilità di una medicina basata sulla genetica sono infinite e si può predire che questi rapidi progressi modificheranno grandemente gli approcci clinici alle malattie.

CONCLUSIONI E PROSPETTIVE FUTURE

I nostri risultati hanno dimostrato che il metodo d'integrazione di dati di caratterizzazione genotipica con dati di sequenza generati nella medesima popolazione oggetto dello studio ha un aumento del potere statistico dello studio.

Inoltre, i nostri risultati confermano l'importanza di esaminare casistiche ancora più ampie di quelle considerate nel presente studio. Tali difficoltà potranno essere superate attraverso un impegno congiunto dei principali centri di diabetologia sardi. Sarà anche effettuata un'analisi congiunta di pazienti con differenti patologie autoimmuni, in particolare, DT1 e sclerosi multipla, raccolte dal nostro gruppo che aumenteranno considerevolmente il potere statistico di identificare varianti condivise predisponenti per l'autoimmunità in generale. Infine sono previste meta-analisi dei dati raccolti e analizzati dal nostro gruppo e di quelli raccolti da altri gruppi, in particolare dal Prof.John Todd a Cambridge, UK.

A prescindere dall'identificazione dei geni di suscettibilità per il DT1, il presente studio ci ha consentito di migliorare considerevolmente le informazioni della variabilità genetica della popolazione sarda, costituendo un nuovo pannello di referenza definito da aplotipi di individui sardi.

Attualmente abbiamo confermato loci noti non HLA di suscettibilità al DT1 quale i geni *PTPN22, CTLA4, IL2RA e CLEC16A* ed *INS*.

La precisione del metodo di inferenza utilizzato per l'analisi dei dati è stato confermato dall'applicazione con successo nella sclerosi multipla. Il metodo ha permesso al nostro gruppo di evidenziare un nuovo gene implicato nella suscettibilità di questa malattia (Sanna, Pitzalis et al. 2010).

L'approccio di inferenza utilizzato per l'analisi dei dati ha funzionato in maniera ottimale, soprattutto con l'uso del pannello di referenza sardo ha aumentato la precisione e l'accuratezza dell'imputazione, soprattutto per le varianti rare (MAF 1-3%).

Il rilevamento di varianti altamente penetranti e a bassa frequenza presenti nella nostra popolazione richiederà l'ampliamento del progetto di questa tesi di dottorato.

Attualmente sono in corso le analisi dei dati di sequenza di 1700 individui sardi che aumenterà il potere di descrivere numerosi altri polimorfismi, con miglioramento della

qualità di imputazione.

In primo luogo prevediamo l'ampliamento della casistica dei campioni dello studio.

Attualmente stiamo caratterizzando tutta la casitica a disposizione circa 1800 pazienti e 4,500 controlli con l'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.

Inoltre stiamo iniziando studi di analisi del mappaggio fine di regioni che hanno mostrato una maggiore significatività, esso ci permetterà di effettuare uno zoom oltre che dei geni noti anche di nuove varianti sardo-specifiche. Anche il miglioramento dei software per la chiamata delle basi, reagenti per le reazioni di sequenza sempre più performanti, insieme ad una ottimale titolazione dei campioni e, quindi, dei *clusters* consentirà il raggiungimento di un *coverage* notevole.

Le varianti rare saranno indagate utilizzando anche dati di sequenza dell'RNA (RNA-seq) di 1,000 individui sardi. Pertanto l'analisi del trascrittoma consentirà di esaminare qualitativamente e quantitativamente i profili di espressione e correlarli con i profili genetici (eQTL) e con il fenotipo DT1.

A breve sarà quindi inoltre possibile estendere lo studio, oltre che agli SNPs, a altre variazioni, quali inserzioni e delezioni e varianti strutturali (CNV, CNP), grazie alla ottimizzazione di appositi algoritmi.

L'integrazione di tutte queste informazioni delineranno i nostri esperimenti futuri allo scopo di aumentare le conoscenze sulle basi genetiche del DT1 che porranno le basi per la prevenzione della malattia man mano che altre informazioni saranno delineate riguardo allo stile di vita e ad altri fattori di rischio di ammalarsi.

Bibliografia

(1976). "American Academy of Pediatrics, The Task Force on Genetic Screening: The pediatrician and genetic screening (every pediatrician a geneticist)." Pediatrics 58(5): 757-764.

Bell, G. I., S. Horita, et al. (1984). "A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus." Diabetes 33(176-83).

Bell, G. I., J. H. Karam, et al. (1981). "Polymorphic DNA region adjacent to the 5' end of the human insulin gene." Proc Natl Acad Sci U S A 78(9): 5759-5763.

Bell, G. I., R. L. Pictet, et al. (1980). "Sequence of the human insulin gene." Nature 284(5751): 26-32.

Bennett, S. and J. Todd (1996). "Human type 1 diabetes and the insulin gene:principles of mapping polygenes." Annu. Rev. Genet. 30: 343-370.

Bennett, S., A. Wilson, et al. (1996). "IDDM2-VNTR-encoded susceptibility to type 1 diabetes: dominant protection and parental transmission of alleles of the insulin gene-linked minisatellite locus." Journal of Autoimmunity .

Bennett, S. T., A. M. Lucassen, et al. (1995). "Susceptibility to human type 1 diabetes at *IDDM2* is determined by tandem repeat variation at the insulin gene minisatellite locus." Nat Genet 9: 284-292.

Bottini, N., L. Musumeci, et al. (2004). "A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes." Nat Genet 36(4): 337-338.

Bottini, N., T. Vang, et al. (2006). "Role of PTPN22 in type 1 diabetes and other autoimmune diseases." Semin Immunol 18(4): 207-213.

Copeman, J. B., F. Cucca, et al. (1995). "Linkage disequilibrium mapping of a type 1 diabetes susceptibility gene (IDDM7) to chromosome 2q31-q33." Nat Genet 9(1): 80-85.

Cucca, F., F. Dudbridge, et al. (2001). "The HLA-DPB1-associated component of the IDDM1 and its relationship to the major loci HLA-DQB1, -DQA1, and -DRB1." Diabetes 50(5): 1200-1205.

Cucca, F., R. Lampis, et al. (2001). "A correlation between the relative predisposition of MHC class II alleles to type 1 diabetes and the structure of their proteins." Hum. Mol. Genet. 10(19): 2025-2037.

Cucca, F., R. Lampis, et al. (1995). "The distribution of DR4 haplotypes in Sardinia suggests a primary association of insulin dependent diabetes mellitus with DRB1 and DQB1 loci." Hum. Immunol. 43(4): 301-308.

Cucca, F., F. Muntoni, et al. (1993). "Combinations of specific DRB1, DQA1, DQB1 haplotypes are associated with insulin-dependent diabetes mellitus in Sardinia." Hum. Immunol. 37(2): 85-94.

Cucca, F. and J. Todd (1996). HLA/MHC: genes, molecules and function. Oxford, BIOS Scientific Publishers.

Durinovic-Bello, I., R. P. Wu, et al. (2010). "Insulin gene VNTR genotype associates with frequency and phenotype of the autoimmune response to proinsulin." Genes Immun 11(2): 188-193.

Ettinger, R. A., A. W. Liu, et al. (2000). "Beta 57-Asp plays an essential role in the unique SDS stability of HLA-DQA1*0102/DQB1*0602 alpha beta protein dimer, the class II MHC allele associated with protection from insulin-dependent diabetes mellitus." J Immunol 165(6): 3232-3238.

Field, L. L. (1991). "Non-HLA region genes in insulin dependent diabetes mellitus." Baillieres Clin Endocrinol Metab 5(3): 413-438.

Green, A. (2001). "The EURODIAB studies on childhood diabetes 1988-1999. Europe and Diabetes." Diabetologia 44 Suppl 3: B1-2.

Gyllensten, U. B. and H. A. Erlich (1993). "MHC class II haplotypes and linkage disequilibrium in primates." Hum Immunol 36(1): 1-10.

Harper, M. E., A. Ullrich, et al. (1981). "Localization of the human insulin gene to the distal end of the short arm of chromosome 11." Proc Natl Acad Sci U S A 78(7): 4458-4460.

Herr, M., F. Dudbridge, et al. (2000). "Evaluation of fine mapping strategies for a multifactorial disease locus: systematic linkage and association analysis of IDDM1 in the HLA region on chromosome 6p21." Hum Mol Genet 9(9): 1291-1301.

Jorde, L. B., W. S. Watkins, et al. (1994). "Linkage Disequilibrium Predicts Physical Distance in the Adenomatous Polyposis Coli Region." Am J Hum Genet 54: 884-898.

Julier, C., R. N. Hyer, et al. (1991). "Insulin-IGF2 region on chromosome 11p encodes a gene implicated in HLA-DR4-dependent diabetes susceptibility." Nature 354: 155-159.
Julier, C., A. Lucassen, et al. (1994). "Multiple DNA variant association analysis: application to the insulin gene region in type 1 diabetes." Am J Hum Genet 55: 1247-1254.
Kawasaki, E., T. Awata, et al. (2006). "Systematic search for single nucleotide polymorphisms in a lymphoid tyrosine phosphatase gene (PTPN22): association between a promoter polymorphism and type 1 diabetes in Asian populations." Am J Med Genet A 140(6): 586-593.
Klitz, W., J. C. Stephens, et al. (1995). "Discordant patterns of linkage disequilrium of the peptide transporter loci within the HLA class region." Am J Hum Genet 57: 1436-1444.
Kwok, W. W., G. T. Nepon, et al. (1995). "HLA-DQ polymorphisms are highly selective for peptide binding interactions." J. Immunol. 155: 2468-2476.
Li, Y., C. Willer, et al. (2009). "Genotype imputation." Annu Rev Genomics Hum Genet 10: 387-406.
Lucassen, A., C. Julier, et al. (1993). "Susceptibility to insulin dependent diabetes mellitus maps to a 4.1 kb segment of DNA spanning the insulin gene and associated VNTR." Nature Genet 4: 305-310.
Lund, T., L. O'Reilly, et al. (1990). "Prevention of insulin-dependent diabetes mellitus in non-obese diabetic mice by transgenes encoding modified I-A beta-chain or normal I-E alpha- chain." Nature 345(6277): 727-729.
Matesanz, F., A. Caro-Maldonado, et al. (2007). "IL2RA/CD25 polymorphisms contribute to multiple sclerosis susceptibility." J Neurol 254(5): 682-684.
Meylan, E., J. Tschopp, et al. (2006). "Intracellular pattern recognition receptors in the host response." Nature 442(7098): 39-44.
Miyazaki, T., M. Uno, et al. (1990). "Direct evidence for the contribution of the unique I-ANOD to the development of insulitis in non-obese diabetic mice." Nature 345(6277): 722-724.
Nistico, L., R. Buzzetti, et al. (1996). "The CTLA-4 gene region of chromosome 2q33 is linked to, and associated with, type 1 diabetes. Belgian Diabetes Registry." Hum Mol Genet 5(7): 1075-1080.
Owerbach, D., A. Lernmark, et al. (1983). "HLA-D region β-chain DNA endonuclease fragments differ between HLA-DR identical healthy and insulin-dependent diabetic individuals." Nature 303: 815-817.
Pociot, F., B. Akolkar, et al. (2010). "Genetics of type 1 diabetes: what's next?" Diabetes 59(7): 1561-1571.
Sanna, S., M. Pitzalis, et al. (2010). "Variants within the immunoregulatory CBLB gene are associated with multiple sclerosis." Nat Genet 42(6): 495-497.
Sharfe, N., H. K. Dadi, et al. (1997). "Human immune disorder arising from mutation of the alpha chain of the interleukin-2 receptor." Proc Natl Acad Sci U S A 94(7): 3168-3171.
Singer, S. M., R. Tisch, et al. (1998). "Prevention of diabetes in NOD mice by a mutated I-Ab transgene." Diabetes 47(10): 1570-1577.
Slattery, R. M., L. Kjer-Nielsen, et al. (1990). "Prevention of diabetes in non-obese diabetic I-Ak transgenic mice." Nature 345(6277): 724-726.
Smyth, D., J. D. Cooper, et al. (2004). "Replication of an Association Between the Lymphoid Tyrosine Phosphatase Locus (LYP/PTPN22) With Type 1 Diabetes, and Evidence for Its Role as a General Autoimmunity Locus." Diabetes 53(11): 3020-3023.
Smyth, D. J., J. D. Cooper, et al. (2006). "A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region." Nat Genet 38(6): 617-619.
Spielman, R. S., R. E. McGinnis, et al. (1993). "Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM)." Am J Hum Genet 52(3): 506-516.
Todd, J. A., J. I. Bell, et al. (1987). "HLA-DQ beta gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus." Nature 329(6140): 599-604.
Todd, J. A., N. M. Walker, et al. (2007). "Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes." Nat Genet 39(7): 857-864.
Todd, J. A. and L. S. Wicker (2001). "Genetic protection from the inflammatory disease type 1 diabetes in humans and animal models." Immunity 15(3): 387-395.
Ueda, H., J. M. Howson, et al. (2003). "Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease." Nature 423(6939): 506-511.
Ullrich, A., T. J. Dull, et al. (1980). "Genetic variation in the human insulin gene." Science 209(4456): 612-615.

Undlien, D. E., S. T. Bennett, et al. (1995). "Insulin gene region-encoded susceptibility to IDDM maps upstream of the insulin gene." Diabetes 44: 620-625.
Vang, T., M. Congia, et al. (2005). "Autoimmune-associated lymphoid tyrosine phosphatase is a gain-of-function variant." Nat Genet 37(12): 1317-1319.
Vella, A., J. D. Cooper, et al. (2005). "Localization of a type 1 diabetes locus in the IL2RA/CD25 region by use of tag single-nucleotide polymorphisms." Am J Hum Genet 76(5): 773-779.
Wicker, L. S., J. A. Todd, et al. (1995). "Genetic control of autoimmune diabetes in the NOD mouse." Ann. Rev. Immunol. 13: 179-200.
Wolf, E., K. M. Spencer, et al. (1983). "The genetic susceptibility to type 1 (insulin-dependent) diabetes: analysis of the HLA-DR association." Diabetologia 24(4): 224-230.
Yoneyama, M., M. Kikuchi, et al. (2005). "Shared and unique functions of the DExD/H-box helicases RIG-I, MDA5, and LGP2 in antiviral innate immunity." J Immunol 175(5): 2851-2858.
Zavattari, P., E. Deidda, et al. (2000). "Major factors influencing linkage disequilibrium by analysis of different chromosome regions in distinct populations: demography, chromosome recombination frequency and selection." Hum. Mol. Genet. 9(20): 2947-2957.
Zavattari, P., R. Lampis, et al. (2001). "Conditional linkage disequilibrium analysis of a complex disease superlocus, IDDM1 in the HLA region, reveals the presence of independent modifying gene effects influencing the type 1 diabetes risk encoded by the major HLA-DQB1, -DRB1 disease loci." Hum. Mol. Genet. 10(8): 881-889.
Zavattari, P., R. Lampis, et al. (2000). "Confirmation of the DRB1-DQB1 loci as the major component of IDDM1 in the isolated founder population of Sardinia." Hum. Mol. Genet. 9(20): 2967-2972.
Zoledziewska, M., G. Costa, et al. (2009). "Variation within the CLEC16A gene shows consistent disease association with both multiple sclerosis and type 1 diabetes in Sardinia." Genes Immun 10(1): 15-17.
Zoledziewska, M., C. Perra, et al. (2008). "Further evidence of a primary, causal association of the PTPN22 620W variant with type 1 diabetes." Diabetes 57(1): 229-234.

Ringraziamenti

Quando ho spedito la domanda di tesi per conseguire il titolo di dottorato ho pensato: "E' fatta tra qualche mese è tutto finito, avrò il titolo nel cassetto."
In un attimo tre anni di studio, di lavoro, di scelte hanno percorso i miei pensieri.

Ringrazio quanti con il loro tributo hanno reso possibile questo progetto, superando ogni avversità con determinazione e .costanza.

Far parte di questo progetto mi ha permesso di collaborare con eccelse professionalità, in primis il prof.Cucca che mi ha dato l'opportunità di "Salire su questo treno", coinvolgendomi e interessandomi a questo obiettivo, con entusiasmo e passione;la ricerca scientifica ha bisogno di persone innamorate sinceramente della loro professione come il prof. Cucca al quale va un ringraziamento speciale.

Ringrazio le mie colleghe, dott.ssa Magdalena Zoledziewska e la dott.ssa Maristella Pitzalis per la particolare attenzione dedicatami con continua disponibilità e pazienza durante l'intero periodo formativo,

Un pensiero speciale a Michael per le sue correzioni in madre lingua e ad Elena per il supporto tecnico – logistico

Ringrazio i miei genitori, Cinzia e Roberto che hanno sempre creduto in me, permettendomi di raggiungere questa nuova meta con la loro presenza e incoraggiamento.

Ringrazio Ilenia per il suo supporto tecnico nei lavori al computer.

Rivolgo un pensiero speciale a mia nonna venuta a mancare in questo percorso di studi.

Un bacio ad Andrea, che mi ha affiancato con una pazienza degna di Giobbe, consolandomi e incoraggiandomi in tutti i momenti, di fatica e apprensione.

E infine, non certo perché il meno importante ma perché a Lui tutto confluisce, ringrazio il Signore, perché mi ha donato questa possibilità e ha voluto circondarmi di tante persone che mi amano e mi hanno accompagnato in questa fase bellissima della vita.
Grazie, Signore, per la forza che mi hai dato nei momenti più bui e difficili, grazie perché hai permesso che tutto venisse superato nel bene. A Te affido il nostro futuro.