UNIVERSITÀ DEGLI STUDI DI SASSARI

SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE BIOMEDICHE

Direttore della Scuola: Prof. Andrea Fausto Piana

INDIRIZZO IN GENETICA MEDICA, MALATTIE METABOLICHE E NUTRIGENOMICA

XXVIII CICLO

Caratterizzazione molecolare di tratti quantitativi di potenziale interesse biomedico nella popolazione Sarda
Studio SardiNIA

Direttore:

Prof. Andrea Fausto Piana

Tesi di dottorato di:

Dott.ssa Mulas Antonella

Tutor: *Prof. Francesco Cucca*

Co-Tutor: *Dott.ssa Magdalena Zoledziewska*

Anno Accademico 2014 - 2015

INDICE

PREFAZIONE:

Nei mie tredici anni di lavoro nel progetto SardiNIA ho avuto la grande opportunità di studiare la genetica di una delle più antiche e interessanti popolazioni d'Europa – la popolazione Sarda. Tali anni di ricerca finalizza, hanno condotto all'elaborazione di questa tesi di dottorato che descrive la caratterizzazione molecolare di tratti quantitativi di potenziale interesse biomedico nella popolazione sarda, nel contesto del progetto SardiNIA.

Nell'ultimo decennio i grandi progressi delle tecnologie *genome-wide* hanno permesso di studiare una notevole parte del genoma umano. I *microarrays* utilizzati per gli studi GWAS (*Genome Wide Association Scan*) sebben notevolmente migliorati in termini di copertura genomica, catturano solo una frazione della variabilità genotipica presente, e spiegano solo una parte della componente genetica dei tratti genetici lasciando poco rappresentata la sfera delle varianti rare, a bassa frequenza e varianti fondatrici. L'approccio *population-sequencing* che usa le tecnologie di sequenziamento di ultima generazione combinato con l'imputazione statistica come esemplificato in questa tesi, permette di superare questa barriera e testare centinaia di migliaia di varianti non ancora testate.

Perciò la mia tesi descriverà l'attività di ricerca svolta nei tre anni della scuola di dottorato in cui ho partecipato allo studio GWAS condotto nell'ambito del progetto SardiNIA con lo scopo di .caratterizzare dal punto di vista molecolare i tratti quantitativi di potenziale interesse biomedico

Alla base del nostro studio vi è stata la creazione di una mappa integrata nella quale abbiamo unito i dati ottenuti dalla genotipizzazione e dal sequenziamento dell'intera coorte SardiNIA per creare un pannello interno sardo-specifico usato come riferimento per l'inferenza statistica dei genotipi ottenuti su tutti gli individui della coorte. Cardine dei nostri studi, sono stati i processi di estensiva genotipizzazione, sequenziamento e imputazione statistica grazie ai quali in tutti e tre i lavori abbiamo potuto dimostrare come l'uso del *Whole Genome Sequencing-based* GWAS possa essere un utile strumento nell'identificazione di varianti rarefondatrici prendendo come esempio alcuni tratti di particolare interesse biomedico quali i tratti cardivascolari, antropometrici ed ematologici.

Ho scelto di discutere la tesi sui GWAS condotti nella coorte SardiNIA con lo scopo di consentire di effettuare ricerche nel campo della biomedicina. Il mio studio è stato in parte guidato dalla grande curiosità di scoprire le basi genetiche della mia popolazione di origine quella Ogliastrina ed in parte guidato dall'onore di far parte del gruppo di lavoro guidato dal Prof. Francesco Cucca che ha condotto alla finalizzazione di tre recenti pubblicazioni nell' importante rivista scientifica internazionale Nature Genetics, con editoriale e copertina dedicato alla Sardegna.

In questi anni di lavoro ho contribuito alla pubblicazioni di diversi lavori che hanno come scopo la dissezione genetica di tratti quantitativi di potenziale interesse biomedico:

Scuteri, A. et al, Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity- related traits. PLoS Genet.3(7):e115: 1200-1210. (2007)

Li, S.et al, The GLUT9 Gene is Associated with Serum Uric Acid Levels in Sardinia and Chianti Cohorts. PLoS Genetics, 3 (11) e194: 1-7. (2007)

Uda, M. et al. Genome wide association scan shows BCL11A associated with persistent HbF and amelioration of the phenotype of α-thalassemia. PNAS,105(5):1620-5. (2008)

Lopez, L. A. et al. Phosphodiesterase 8B Gene Variants Are Associated with Serum TSH Levels and Thyroid Function American Journal Human Genetics.82(6):1270-80. (2008)

Willer,CJ. et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet.40(2):161-9. (2008)

Willer,CJ. et al. Six new loci associated with body mass index highlight a neuronal influence on body weight Regulation. Nature Genetics. 41(1):25-34. (2009)

Lettre,G. et al. Identfication of ten loci associated with height highlights new biological pathway in human growth Nature Genetics. 40(5):489-90. (2008)

Tanaka, T et al. Genome-wide Association Study of Vitamin B6, Vitamin B12, Folate and Homocysteine Blood Concentrations. American Journal Human Genetics. (2009)

Sanna, S. et al. Common variants in the SLCO1B3 locus are associated with bilirubin levels and unconjugated hyperbilirubinemia. Hum Mol Genet.18(14):2711-8. (2009)

Lango,AH. et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 467(7317):832-8. (2010)

Terracciano, A. et al Genetics of serum BDNF: Meta-analysis of the Val66Met and genome-wide association study.World J Biol Psychiatry. (2011)

Sanna, S. et al. Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. PLoS Genet. 7(7):e1002198. (2011)

Naitza, S. et al. A genome-wide association scan on the levels of markers of inflammation in Sardinians reveals associations that underpin its complex regulation. PLoS Genet8(1):e1002480. (2012)

Scott,RA. et al.Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. Nat Genet. doi: 10.1038/ng.2385. (2012).

Voight,BF. et al. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS Genet. 8(8):e1002793. (2012)

Orrù,V. et al. Genetic variants regulating immune cell levels in health and disease. Cell. 155(1):242-56. doi: 10.1016/j.c. (2013)

Medici,M. et al. Identification of novel genetic Loci associated with thyroid peroxidase antibodies and clinical thyroid disease. PLoS Genet. 10(2):e1004123. doi: 10.1371/journal.pgen.1004123. (2014)

Pistis,G. et al. Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. Eur J Hum Genet. doi: 10.1038/ejhg.2014.216. (2014)

Arking, DE. et al Genetic association study of QT interval highlights role for calcium signaling pathways in myocardial repolarization. Nat Genet. 46(8):826-36. doi: 10.1038/ng.3014. (2014)

Sidore, C. et al. Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. Nat. Genet. doi:10.1038/ng.3368.(2015)

Danjou, F. et al. Genome-wide association analyses based on whole-genome sequencing in Sardinia provide insights into regulation of hemoglobin levels. Nat. Genet. doi:10.1038/ng.3307. (2015)

Zoledziewska, M. *et al.* Height-reducing variants and selection for short stature in Sardinia. *Nat. Genet.* doi:10.1038/ng.3403 (2015)

CAPITOLO 1

1.1 La definizione di tratti quantitativi

L'efficace combinazione tra le tecniche di sequenziamento di ultima generazione e le moderne piattaforme di genotipizzazione ha permesso l'individuazione di centinaia di migliaia di varianti genetiche che hanno consentito di catalogare una discreta parte di variabilità genetica e di correlarla con le differenze fenotipiche osservate. Ma da un attenta analisi dei risultati ottenuti in questi anni dai diversi gruppi di ricerca, ci rendiamo conto che la maggior parte delle varianti genetiche individuate hanno un modesto effetto sul fenotipo esaminato spiegando quindi sono una piccola parte della variabilità osservata. E' abbastanza chiaro il fatto che l'ampia variabilità fenotipica tra gli individui vada ulteriormente indagata e soprattutto, dobbiamo tener presente che questa sia di natura quantitativa. Lo studio della genetica quantitativa è fortemente complicato proprio dalla natura dei tratti quantitativi che per definizione sono soggetti ad una variabilità genetica e fenotipica dovuta alla segregazione simultanea di molti geni, con effetti fenotipici piccoli e difficilmente distinguibili tra loro, il tutto accompagnato dall'influenza ambientale. Da questo particolare assetto di interazioni si evince come l'architettura genetica dei tratti quantitativi sia il risultato delle azioni cumulative di molti geni e di questi con l'ambiante, da cui deriva la ormai comune, definizione di "tratti complessi". Il locus genici occupati da queste varianti corrispondono a loci per caratteri quantitativi e vengono identificati come QTL, Quantitative Trait Loci. All'aumentare del numero di geni coinvolti nella determinazione di un tratto quantitativo, aumenta anche il numero di varianti che concorrono alla variabilità del tratto, varianti molto spesso rare e/o popolazioni specifiche. Un ruolo importante è svolto dal contesto genetico ed ambientale in cui gli QTL si trovano.

Fenomeni quali epistassi, pleiotropia e background genetici non uniformi possono ulteriormente complicare il quadro di analisi. I passi in avanti che la comunità scientifica sta compiendo per la dissezione genetica dei tratti complessi vanno verso un particolare assetto di analisi che conduce ad aumentare il potere di risoluzione in termini di mappaggio fine delle regioni genomiche ed all'aumento del numero degli individui oggetto di studio. Questo approccio analitico fortemente aiutato dai progressi compiuti dalle nuove tecnologie può essere ulteriormente affiancato da altre strategie di analisi utili ad aumentare il potere di risoluzione. A tal scopo nei capitoli successivi mostreremo la strategia di analisi scelta dal nostro gruppo di ricerca che ha condotto ad importanti risultati.

1.2 I caratteri multifattoriali e le distribuzioni di frequenza

I tratti quantitativi mostrano un ampia variabilità a livello di popolazione. Questa variabilità è la base per l'evoluzione e la selezione naturale, fenomeni che agiscono per favorire l'adattamento di un individuo all'ambiente. La distribuzione dei tratti quantitativi varia entro un range di classi ben definite, può essere misurata e descritta in termini statistici al fine di individuare le modalità tramite le quali si distribuiscono nella popolazione, individuando la variazione dei valori all'interno dei campioni stessi e di quanto si discostino dalla media. Generalmente l'analisi della distribuzione di uno o più caratteri quantitativi viene fatta a livello di popolazione e può prevedere un tipo di indagine che si protrae nel tempo al fine di identificare dei cambiamenti del tratto di interesse nell'arco di un determinato decorso temporale. Studi longitudinali di questo tipo implicano misurazioni ripetute nel tempo. Un esempio di studi di questo tipo è rappresentato dallo studio SardiNIA oggetto di questo tesi. Un aspetto importante da considerare è come questi caratteri possano essere intesi in termini di predisposizione o in termini di suscettibilità. Da qui la definizione di carattere quantitativo inteso come carattere soglia, carattere che si manifesterà in seguito ad effetti cumulativi di alleli di suscettibilità o di rischio che superano una certa soglia.
La soglia può essere definita come il punto di equilibrio della bilancia, superato l'equilibrio si avrà uno sbilanciamento che conduce ad un aumento del numero di alleli predisponenti con il conseguente manifestarsi della patologia. Il livello di aggregazione familiare è naturalmente in funzione del grado di parentela quindi della percentuale di geni condivisi tra gli individui ed il probando.

1.3 L'ereditabilità dei caratteri multifattoriali

Il concetto di ereditabilità riveste un ruolo fondamentale nella comprensione di quanto della variabilità fenotipica osservata in un tratto quantitativo possa essere attribuibile alla genetica e di quanto la conoscenza di questa parte di variabilità possa essere utile nella prevenzione, diagnosi e trattamento di alcune patologie. Uno sguardo ai risultati dei vari studi di associazione pubblicati ci fa capire, però che una grossa parte di ereditabilità rimane inspiegata e non attribuita.
Diverse sono le ipotesi formulate sul motivo che conduce a questa perdita di ereditabilità. La prima ipotesi riguarda sicuramente la natura poligenica di questi caratteri che da luogo ad uno scenario in cui abbiamo diverse varianti che conferiscono ognuna un piccolo effetto sul tratto.

La seconda ipotesi può essere rappresentata dalla variati rare spesso popolazione specifiche, che hanno un effetto importante sul tratto ma non sono rappresentate nei normali array di genotipizzazione che risultano essere implementati con varianti a frequenza superiore all'5%. Un altro limite è rappresentato dal tipo di popolazione scelta per lo studio, la maggior parte degli studi GWAS si focalizzano su popolazioni di origine Europea, che per definizione mostrano background genetici ed ambientali non uniformi mentre le popolazioni isolate, come discusso in seguito rappresentano un utile strumento per l'individuazione di varianti rare spesso responsabili di una grossa parte di variabilità.Va precisato che l'ereditabilità è in relazione alla popolazione in esame in quanto la variabilità genica, additiva e non additiva come pure la varianza ambientale, sono popolazione specifici. La variabilità genetica è strettamente legata alla segregazione a livello di popolazione degli alleli che influenzano il tratto, dalla loro frequenza e dal loro effetto. Queste variabili, come pure la variabilità ambientale, possono essere diverse tra le popolazioni, di conseguenza l'ereditabilità di un determinato tratto potrebbe non essere la stessa se consideriamo popolazioni differenti. L'ereditabilità di un tratto potrebbe variare in base al genere ed anche in base alla stadio di vita considerato.

In conclusione diciamo che l'ereditabilità è un importante parametro da considerare nell'ambito della genetica quantitativa, permette la dissezione tra i fattori genetici e non genetici, ma molto ancora deve essere fatto per poter elucidare tutti gli aspetti che concorrono alla determinazione del fenomeno. Il lavoro esposto in questa tesi è sicuramente un esempio di quali strategie possono essere attuate per detestare varianti genetiche ad ampio affetto che concorrono alla determinazione di importanti caratteri quantitativi.

CAPITOLO 2

In questo capitolo forniamo la descrizione della caratteristiche peculiari della popolazione Sarda che ci hanno condotto a sceglierla come scenario dei nostri studi permettendoci di ottenere i risultati conseguiti.

2.1 Uso delle popolazioni isolate per lo studio dei caratteri multifattoriali e vantaggi presenti negli isolati genetici

Avendo come punto di partenza la dissezione dei tratti quantitativi di potenziale interesse biomedico, ci siamo subito dovuti misurare con una serie di problemi tecnici e pratici, tipici degli studi che implicano l'analisi dell'architettura genetica dei tratti complessi. Come precedentemente detto il principale ostacolo è rappresentato dalla natura poligenica di questi tratti, dai limiti degli array commerciali ma nello stesso tempo, dall'importanza e dal conseguente fatto che fosse necessario aumentare il numero di varianti da poter testare nei nostri GWAS. Questo avrebbe comportato sicuramente un notevole dispendio in termini economici ma soprattutto non sarebbe stato sufficiente. Per poter ottenere risultati soddisfacenti è necessario agire sotto più fronti, aumentare la possibilità di individuare varianti rare, utilizzare individui con un background genetico uniforme e per livellare l'effetto dell'ambiente utilizzare individui che condividono lo stesso ambiente ed in parte lo stesso stile di vita. Chi incarna tutti questi aspetti? gli isolati genetici.

Dalle analisi del DNA cromosomale e mitocondriale emerge che le popolazioni isolate mostrano una diversità genetica minore rispetto alle altre popolazioni, derivano da un piccolo gruppo di individui fondatori in cui la nuova popolazione formatasi possiede solo una parte della variabilità genetica della popolazione originale. Fenomeni evolutivi quali l'effetto fondatore e la deriva genica hanno avuto un effetto molto importante sulla composizione del pool genico dei Sardi conducendo alla fissazione di alleli rari i con l'effetto di far permanere tali alleli nella nuova popolazione formatasi. Ne consegue che alcuni alleli sono stati eliminati dal pool genico mentre altri vengono "spinti" a frequenze molto più elevate. In conclusione l'utilità della popolazione Sarda risiede nel fatto che l'eterogeneità genetica tipica delle altre popolazione Europee, viene superata da un background genetico uniforme e dalla condivisione dello stesso ambiente e stile di vita che riducono ulteriormente la complessità della varianza fenotipica dovuta all'ambiente.

2.2 Caratteristica della popolazione Sarda

La particolare posizione geografica della Sardegna, inserita al centro del Mediterraneo occidentale e la montuosità del suo territorio hanno fatto sì che tra la popolazione sarda, per via dell'isolamento e dell'azione di particolari processi evolutivi, quali la deriva genetica e la selezione naturale, si siano venute a creare particolari caratteristiche geniche che l'hanno resa uno scenario di particolare interesse per lo studio dei tratti e patologie multifattoriali, quali ad esempio le malattie autoimmuni come il diabete di tipo I e la sclerosi multipla che mostrano un alta incidenza in Sardegna.

I vari ritrovamenti archeologici documentano la presenza umana sull'Isola sin dall'epoca prenuragica. I dati genetici e le espansioni demografiche sono coerenti con i dati archeologici classici, che indicano che la Sardegna ha raggiunto un notevole sviluppo in termini di numero di individui sin dai tempi della preistoria con la popolazione stimata durante il periodo nuragico (circa 2500-3700 anni fa) maggiore di 300.000 abitanti [6] [7].
Le pressioni Cartaginesi e Romaniche spinsero i nuragici a spostarsi verso la regione centrale collinare dell'isola, questo fenomeno rese minimi i contatti culturali esterni rendendo questa parte della Sardegna la regione più conservatrice dell'isola sia dal punto di vista genetico che linguistico.
Tale caratteristica di omogeneità della popolazione Sarda non è però unanimemente accettata. Alcuni studi evidenziano un'elevata variabilità tra sub-popolazioni geografiche o linguistiche [6] [7] ed altri dimostrano come l'intera popolazione sarda presenterebbe livelli di eterogeneità paragonabili a quelli delle popolazioni continentali [10]. Altri studi supportano la tesi dell'omogeneità genetica. L'analisi della distribuzione degli aplotipi relativi a HLA DRB1-DQA1-DQB1condotta a livello di sette sub-regioni della Sardegna: Cagliari, Carbonia, Oristano, Lanusei, Tempio, Sassari e Sorgono, [9], dimostra invece che la distribuzione degli aplotipi più frequenti risulta essere uniforme tra le diverse sub-regioni. Quindi la mancata individuazione di eterogeneità genetica tra le regioni costiere, maggiormente soggette ad invasioni da parte di altre popolazioni, rispetto alle regioni interne, suggerisce un basso flusso genico tra gli invasori e le popolazioni del luogo. La spiegazione di ciò risiede nel fatto che la Sardegna risultava essere densamente popolata come precedentemente descritto, con circa 300.000 abitanti già durante il periodo Nuragico, prima dell'eventuale invasione da parte di altre popolazioni.
L'assenza di eterogeneità genetica tra le diverse aree della Sardegna non implica che non ci sia un elevato grado di variabilità inter-individuale. Un'altra caratteristica della popolazione Sarda è la presenza di un esteso linkage disequilibrium [8]. Come detto precedentemente tutte le popolazioni isolate si originano da un piccolo gruppo di fondatori, quindi un utile strumento di analisi potrebbe

essere quello della ricerca di un aplotipo comune mostrato nei pazienti. Ricordiamo la definizione di aplotipo, partendo da quella del *linkage disequilibrium*: associazione non-random di alleli, corrispondenti a marcatori (loci) in stretta vicinanza fisica sullo stesso cromosoma,(G.Pilia,Prospettive in Pediatria). Questa associazione genera un assortimento di alleli di diversi marcatori su un singolo cromosoma, ovvero un aplotipo. Naturalmente con il passaggio da una generazione all'altra si verifica " l'erosione" dell'aplotipo ancestrale a seguito degli eventi di ricombinazione. La parziale conservazione dell'aplotipo in un paziente suggerisce che il locus malattia possa risiedere nella regione conservata dell'aplotipo e fenomeni quali la deriva genica e la selezione naturale possono ad esempio tenere in fase determinate combinazioni di alleli. Questo comporta che diverse varianti genetiche, rare nelle altre popolazioni europee possano invece avere una frequenza elevata in Sardegna.

2.3 La descrizione dello studio SardiNIA

Lo studio SardiNIA, conosciuto in Italia come progetto ProgeNIA, nasce nel 2001dalla collaborazione tra l'allora Istituto di Neurogenetica e Neurofarmacologia, ora IRGB, ed il National Institute of Aging (NIA) degli Stati Uniti. Ideatore del progetto è stato il Professor Giuseppe Pilia, prematuramente scomparso nel 2005. Il progetto prese avvio grazie alla profonda intesa tra il Professor Pilia ed il Dottor David Schlessinger responsabile del laboratorio di genetica del NIA, e grazie all'indispensabile supporto del Professor Antonio Cao. Il progetto SardiNIA/ProgeNIA è uno studio longitudinale che si propone di identificare i determinanti genetici di importanti cambiamenti, patologici e non, associati al processo dell'invecchiamento, attraverso l'analisi di oltre 500 tratti quantitativi. Lo scenario dello studio è rappresentato dall'Ogliastra ed in particolare da quattro paesi: Lanusei, Ilbono, Elini ed Arzana. E' proprio a Lanusei, per forte volere del Professor Pilia, che nascono i laboratori del progetto ProgeNIA. Lo studio viene attualmente diretto dal Professor Francesco Cucca, direttore dell'istituto di ricerca genetica e biomedica del CNR di Cagliari nonché Professore ordinario di Genetica Medica presso la facoltà di Medicina dell'Università degli studi di Sassari. Il progetto si avvale di numerosi collaboratori internazionali.
Cardine dello studio sono i volontari che giornalmente si presentano al centro ProgeNIA per dare il loro prezioso e unico contributo alla ricerca.

I volontari vengono sottoposti ad una serie di misurazioni che complessivamente durano circa due ore che comprendono: prelievo di sangue usato per emocromo, analisi biochimiche, estrazione DNA/RNA ed analisi citofluorimetriche. Anamnesi personale e familiare, misurazione dell'altezza e del peso, esame cardiovascolare comprendente misurazioni quali la pressione arteriosa, valutazione della PWV, valutazione dell' IMT, ecocardiogramma ed elettrocardiogramma. Rilevazione di alcuni parametri tiroidei, epatici e renali mediante ecografia. Densitometria ossea. Retinoscopia ed esame audiometrico.

In totale vengono raccolti dati per oltre 500 tratti quantitativi. Oggetto dello studio riguardante la mia tesi di dottorato sono stati i tratti relativi all'altezza umana, ai livelli di HDL,LDL e colesterolo totale nonché i livelli delle emoglobina HbA1,HbA2 ed HbF.

2.4 Whole Genome Sequenced-based Genome Wide Association Scan

Gli studi di associazione su tutto il genoma prevedono l'utilizzo di un ampio set di varianti genetiche, che catturano una sostanziale parte di variabilità genetica, in una serie di campioni di DNA che sono informativi per un tratto o una patologia di interesse. L'obiettivo è quello di evidenziare eventuali effetti di suscettibilità attraverso l'individuazione di associazioni tra un dato genotipo e la patologia od il tratto in esame. Attualmente è uno dei metodi di analisi largamente impiegato per la dissezione dei caratteri multifattoriali, in quanto ha permesso notevoli progressi nella comprensione delle basi genetiche che concorrono alla determinazione della variabilità fenotipica osservata in diversi tratti o patologie di particolare interesse biomedico. Questo approccio di studio può avere differenti campi applicativi. Generalmente la metodica si applica a studi di tipo caso-controllo oppure a studi che prevedono l'uso di popolazioni generali. La base di partenza per uno studio di tipo caso-controllo è il confronto tra un gruppo di individui, casi, che manifestano la patologia in esame e che si presume abbiano un'elevata prevalenza di alleli di suscettibilità per quel tratto, e un secondo gruppo di individui, i controlli, che non manifestano la patologia in esame e che si presume abbiano una minore prevalenza di alleli di suscettibilità. Sempre più spesso però si ricorre all'uso di popolazioni nelle quali sono disponibili gruppi di individui definiti coorti, ben caratterizzate dal punto di vista fenotipico, spesso facenti parte di studi longitudinali che implicano la misurazione e quindi la disponibilità di diversi tratti quantitativi utili per lo studio.

La base di partenza, sia che lo studio sia condotto a livello di popolazione generale o sia di tipo caso controllo, è la caratterizzazione genotipica degli individui in esame. Maggiore sarà il numero di varianti da poter utilizzare come base per lo studio di associazione tanto più efficienti potranno essere i risultati ottenuti. Gli approcci seguiti per creare il set di varianti sono due: quello di genotipizzare un ampio set di individui mediante l'uso di arrays o procedere con il sequenziamento diretto mediante tecniche di sequenziamento di ultima generazione. Brevemente diciamo che il requisito fondamentale che un array di genotipizzazione deve possedere è la capacità di garantire una buona copertura genomica al fine di catturare la maggior parte della variabilità genotipica osservata. Il metodo di interrogazione del locus in esame può essere diretto, includendo varianti genetiche responsabili delle differenze fenotipiche osservate o indiretta includendo SNP che sono in linkage disequilibrium con la varianti che si vuole interrogare. Attualmente i costi relativi ai processi di genotipizzazione sono notevolmente diminuiti, questo consente l'utilizzo di questa metodica su un campione di individui sempre più ampio. Naturalmente la copertura genomica offerta del sequenziamento è notevolmente maggiore, ma i costi sono ancora abbastanza elevati. Il nostro gruppo di ricerca per poter sopperire a questa lacuna è ricorso all'utilizzo di una particolare variante degli studi di associazione classici chiamata Whole Genome Sequencing-based GWAS. In questo approccio di si usa una popolazione isolata, come per esempio quella Ogliastrina, per la dissezione dei tratti genetici complessi in esame. Fulcro della metodica è la creazione del pannello di riferimento Sardo.

Materiale e Metodi

Capitolo 3

3.1 Descrizione della casistica oggetto di studio SardiNIA

Il progetto SardiNIA è costituito da 6.805 individui, di cui 6.142 arruolati negli anni tra il 2001 ed il 2009, ed un ulteriore gruppo di 665 individui arruolati a partire dal 2010 al fine arricchire le fasce di età oltre i 75 anni che, come vediamo dalla figura 4.1 sono le fasce d'età meno rappresentate nello studio.

I 6.807 individui della coorte sono originari di quattro paesi Ogliastrini ubicati nella collina di Lanusei: Lanusei, Ilbono Elini ed Arzana. L'Età dei volontari SardiNIA è compresa tra i 14 ed i 102 anni.

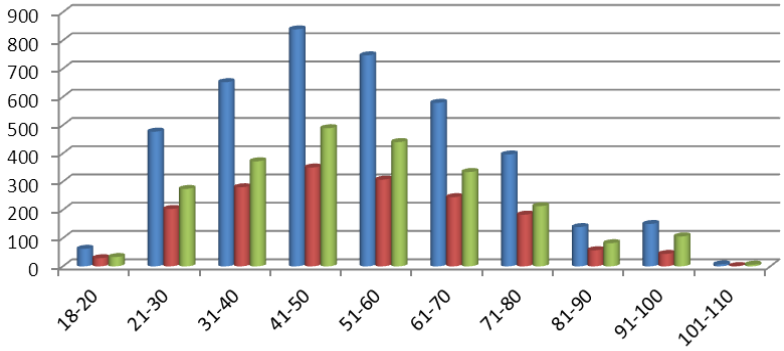


Figura 3.1 Distribuzione per fasce di età nella coorte SardiNIA

In base all'ultimo censimento effettuato nel 2013 la popolazione eleggibile nel territorio è costituita da 10,510 individui, di cui risultano essere stati arruolati nello studio circa il 64,7% (Figura 4.2) della popolazione eleggibile comprendente 3,862 femmine e 2,943 maschi. Complessivamente, i 6.807 individui sono organizzati in 711 unità familiari ciascuna delle quali comprende fino a 5 generazioni. Nella popolazione campione (Figura 4.3) si possono individuare 34,469 coppie di parenti che includono:

- 4.256 coppie bigenerazionali (genitore-figlio)
- 675 coppie trigenerazionali (nonno-nipote)
- 4.933 coppie di fratelli-sorelle
- 4.014 coppie di cugini di primo grado
- 6.459 coppie avuncolari (zio-nipote)
- 180 coppie di fratellastri-sorellastre
- 11 gemelli monozigoti

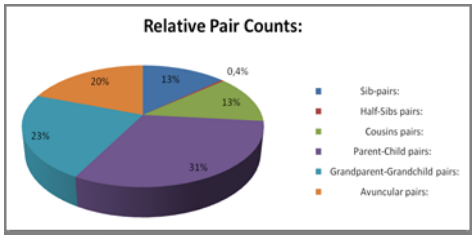


Figura 3.3 Distribuzione familiare coorte SardiNIA

Attualmente lo studio prevede la caratterizzazione di oltre 500 tratti quantitativi o endofenotipi, tra i quali troviamo: misure antropometriche (peso, altezza, indice di massa corporea, circonferenza vita), esami ematochimici ed ematologici (LDL, HDL, TG, insulina, RBC, MCH, MCV, bilirubina, hsCRP, MCP-1, IL-6), tratti cardiovascolari (HR, SBP, DBP, PP, PWV, IMT, QT), tratti della personalità (nevroticismo, estroversione), ed oltre 1000 tratti immunologici che caratterizzano le principali sottoclassi leucocitarie: monociti, granulociti, cellule dendritiche circolanti, linfociti e loro sottoclassi tra cui B,T e NK. Nel contesto dello studio sono stati misurati 200 tratti dicotomici che comprendono patologie di maggior incidenza in Sardegna e relativi fattori di rischio.

3.2 Estrazione del DNA e controlli di qualità per successive applicazioni

Il materiale biologico di partenza per i nostri studi è rappresentato dalle cellule nucleate del sangue - leucociti ematici considerati la fonte maggiore di DNA per studi di routine in biologia molecolare. A partire da prelievi di circa 7 ml di sangue venoso periferico si possono ottenere in media 400 ng di DNA.

Il protocollo seguito per estrarre il DNA leucocitario prevede l'impiego di sangue trattato con anticoagulante. Preferibilmente viene utilizzato l'acido etilendiaminotetracetico (EDTA): $(HOOC\text{-}NH_2)_2N\text{-}CH_2\text{-}CH_2\text{-}N(CH_2\text{-}COOH)_2$, vista al sua capacità chelante. In seguito alla reazione di chelazione si forma un complesso, composto dal legante esadentato, l'EDTA, e uno ione accettore di elettroni, ad esempio lo ione calcio, che costituisce uno dei fattori della coagulazione. Il legame tra l'EDTA e lo ione calcio inibisce l'attività dell'enzima trombochinasi il quale non può trasformare la protrombina. Il compito dell'enzima proteolitico trombina è quello di far precipitare il fibrinogeno presente nel sangue e di trasformarlo in fibrina, le cui molecole formano un reticolo in cui restano intrappolati gli elementi figurate del sangue, mentre si separa la parte liquida, detta siero.

Con l'uso di anticoagulanti si prevengono tutti i fenomeni della coagulazione sanguigna che intrappola le cellule nucleate nel coagulo e interferisce negativamente con il processo di estrazione del DNA, determinando una resa quantitativa nettamente inferiore alle aspettative. L'estrazione può avvenire sia da sangue fresco, quindi appena prelevato, che da sangue conservato a -20ºC per diversi mesi. In questo ultimo caso è necessario che le provette in cui il sangue è contenuto siano costruite in polietilene (PET). Questo materiale risulta essere resistente alle basse temperature, (sino a -80ºC), non poroso e mantenente il sottovuoto.

Il tappo delle provette è in polibutilene. Dopo il prelievo venoso effettuato dall'equipe infermieristica si è proceduto all' estrazione del DNA genomico mediante la classica tecnica del salting-out che utilizza alte concentrazioni saline per rimuovere le proteine (NaCl 6M) (Sambrook et al.,1989). Vediamo in dettaglio il protocollo utilizzato:

3.2.1 Protocollo di Estrazione DNA da sangue periferico tecnica del salting-out

- Trasferire 7ml di sangue intero (10 mM di EDTA pH8) in una falcon da 50 ml,
- Aggiungere 43 ml di Lysis Buffer a 4°C.
- Agitare delicatamente e lasciare in ghiaccio (o in frigorifero a 2-8°C, se non si dispone di ghiaccio) 10-15 minuti.
- Centrifugare 20 minuti 350 rpm a 4°C,
- Eliminare il surnatante tramite pompa a vuoto o per svuotamento,
- Aggiungere 10ml di Fisio Buffer.
- Centrifugare 20 minuti 3500 rpm a 4°C,
- Eliminare il surnatante tramite pompa a vuoto o per svuotamento
- Aggiungere 10ml di Fisio Buffer
- Centrifugare 20 minuti 3500 rpm 4°C,
- Eliminare il surnatante tramite pompa a vuoto o per svuotamento
- Sospendere nuovamente il pellet in 4,2 ml Buffer A
- Vortexare per 10-15 secondi (comunque finché il pellet non si dissolve finemente),
- Aggiungere 140µl di SDS al 10% e 16.8µl (25mg/ml) di proteinase K
- Incubare a 65°C per 1h nel bagnetto termostatico,
- Aggiungere 0,7ml di una soluzione NaCl satura (approssimativamente 6M)
- Agitare al vortex per 15 secondi.
- Centrifugare per 15 minuti a 3000 rpm a 4°C,
- Trasferire il surnatante in un'altra falcon e aggiungere 6 ml di Isopropanolo a T.a.

- Agitare delicatamente finché non si osserva il DNA precipitato,
- Centrifugare per 25minuti 3500 rpm a 4°C
- Eliminare il surnatante tramite pompa a vuoto o per svuotamento,
- Aggiungere al DNA precipitato 10ml di etanolo al 70% a -20°C
- Centrifugare 20minuti 3500 rpm a 4°C
- Eliminare il surnatante tramite pompa a vuoto o per svuotamento
- Eliminare l'etanolo lasciando la falcon capovolta per almeno 1h.
- Sospendere il DNA in TE
- Agitare delicatamente la falcon
- Riporre la falcon nella scarabattola appoggiando il tappo senza avvitare, per consenti re l'evaporazione dei residui di etanolo.
- Avvolgere la scarabattola con stagnola
- Lasciare a temperatura ambiente per tutta la notte
- Eseguire il controllo di qualità su gel di agarosio 1% e lettura spettrofotometrica.

3.2.2 Controlli di qualità: elettroforesi su gel di agarosio 1%

I principali controlli di qualità che ho effettuato dopo il processo di estrazione del DNA genomico sono la verifica dell'integrità del DNA stesso mediante corsa elettroforetica attraverso gel di agarosio all'1%; determinazione della purezza per verificare l'assenza di contaminanti quali ad esempio le proteine, che potrebbero interferire con le successive applicazioni e determinazione della resa attraverso valutazione della concentrazione del DNA tramite lettura spettrofotometrica con Nanodrop 1000 (Thermo Scientific).L'elettroforesi su gel di agarosio è una tecnica che consente alle molecole dotate di carica di essere separate in base al loro peso molecolare attraverso migrazione su un gel in presenza di un campo elettrico. La concentrazione dell'agarosio può essere variata per ottimizzare la risoluzione. I gel di uso più comune sono allo 0,8-1%. Questa concentrazione riesce a separare un range di dimensioni da circa 500 bp a circa 20,000 bp.

L'agarosio è un polimero lineare e neutro estratto dalle alghe, formato da unità di D-galattosio e di 3,6-anidro-L-galattosio legate alternativamente con legami glicosidici. E' costituito da una rete tridimensionale attraverso le cui maglie migrano le molecole sotto l'azione di un campo elettrico generato da un apparecchio detto alimentatore.

Le molecole di DNA cariche negativamente per la presenza di gruppi fosfato, migrano dal polo negativo, anodo, verso il polo positivo, catodo. Per un certo intervallo di pesi molecolari, la velocità di migrazione è funzione del loro peso molecolare: tanto più grande è la molecola di DNA, tanto minore è la velocità di migrazione, tanto più piccola è la molecola di DNA, tanto più velocemente migra. Le molecole di DNA di diversa lunghezza vengono pertanto separate in base alla diversa velocità di migrazione. La conduzione della corrente elettrica ed il controllo del pH durante l'elettroforesi avviene per mezzo di una soluzione salina chiamata tampone di elettroforesi, solitamente TBE 1X. Per poter determinare la lunghezza delle molecole di DNA in esame separate mediante elettroforesi, viene "caricato" sul gel anche il cosiddetto marcatore di peso molecolare, ossia una miscela di frammenti di DNA di cui è noto il peso molecolare. Confrontando la posizione dei frammenti a peso molecolare noto con quella dei frammenti di DNA in esame, è possibile calcolarne il peso molecolare, ossia la lunghezza.

Dato che il peso molecolare di un frammento di DNA è proporzionale al numero di coppie di nucleotidi (basi) che lo costituiscono, di solito esso viene espresso in paia di basi (bp). La preparazione del campione per la corsa

elettroforetica include il tampone di caricamento costituito da un addensante più colorante il gel loading solution che si aggiunge ai campioni di DNA prima dell'elettroforesi, ed usato per "appesantire" i campioni, permettendo la loro permanenza nei pozzetti, e a fornire un marker visivo del progredire della corsa elettroforetica.

La separazione elettroforetica dura circa 30 minuti. Il DNA delle diverse classi di peso molecolare è visibile sotto forma di bande distinte. Per poter visualizzare il DNA dopo corsa elettroforetica, durante la preparazione del gel, all'agarosio si aggiunge bromuro di etidio, una sostanza che ha la proprietà di intercalarsi al DNA e di emettere fluorescenza se esposto alla luce UV tale proprietà consente di rendere visibili le bande di DNA.

Alla fine della corsa, le bande si visualizzano esponendo il gel alla luce ultravioletta.

Preparazione del gel di agarosio 1%:

- Predisporre il supporto per il gel e il pettine per la formazione dei pozzetti;
- Pesare la quantità di polvere per preparare un gel di agarosio all'1%
- Aggiungere il tampone TBE 1X per sciogliere l'agarosio
- Sciogliere il tutto in forno a microonde portando ad ebollizione e agitando ripetutamente sino a che il liquido non diviene trasparente;
- Sotto cappa chimica aggiungere 10 ul di bromuro di etidio ogni 100 ml di soluzione
- Versare il gel sul supporto in modo continuo e deciso evitando di creare delle bolle
- Lasciare raffreddare sino a che il gel polimerizza
- Rimuovere delicatamente il pettine e immergere il gel nella camera elettroforetica, accertandosi che rimanga completamente coperto dal tampone di corsa;
- Caricare i campioni nei rispettivi pozzetti del gel polimerizzato preparati nel seguente modo: 1 ul di DNA 50ng/ul, 5ul di acqua bi-distillata sterile e 2 ul di loading solution.
- Aggiungere il marker di peso molecolare nel primo pozzetto.
- Avviare la corsa elettroforetica accertandosi che la polarità sia dal polo negativo verso quello positivo con un voltaggio pari a 100V;
- Dopo circa 30 minuti visualizzare il gel al transilluminatore.

3.2.3 Controlli di qualità: Lettura spettrofotometrica con Nanodrop 1000

Il metodo spettrofotometrico sfrutta la capacità degli acidi nucleici di assorbire la luce UV con un massimo assorbimento alla lunghezza d'onda di 260 nm ed è questa la lunghezza d'onda che viene utilizzata per misurarne la concentrazione allo spettrofotometro in quanto si è misurato che un assorbimento pari a 1 OD corrisponde a una concentrazione di DNA equivalente a 50 microgrammi/millilitro. Applicando la formula O.D X 50 otteniamo la concentrazione del campione. Gli spettrofotometri attuali quale ad esempio il Nanodrop 1000 forniscono già la stima della concentrazione calcolata come nanogrammi/ microlitri a partire da una quantità di caricamento pari ad 1 microlitro di DNA genomico.

Per valutare il grado di purezza del campione di DNA in esame sfruttiamo il rapporto di assorbanza a 260/280 nm. Un rapporto di circa 1.8 è generalmente indice di purezza mentre un rapporto inferiore indica la presenza di proteine, fenolo o altri contaminanti che assorbono intorno ai 280 nm. Un secondo indice di purezza è rappresentato dal rapporto 260/230. I valori di 260/230 sono spesso superiori ai rispettivi valori 260/280. Valori normali sono previsti nell'intorno di 2,0-2,2. Un rapporto inferiore al previsto, può indicare la presenza di contaminanti che assorbono a 230 nm come EDTA, carboidrati e fenolo.

3.3 Selezione del campione ProgeNIA per GWAS e disegno sperimentale

I 6,805 individui appartenenti alla coorte SardiNIA sono stati genotipizzati a partire dal 2009 a diversi livelli di risoluzione utilizzando la nota piattaforma di genotipizzazione Illumina attraverso l'iScan system.

La fase di genotipizzazione ha previsto l'utilizzo di quattro beadchip Illumina a diversi gradi di densità. Tre beadchip custom: Human Exome BeadChip, Human CardioMetabochip e lo Human Immunochip, contenenti circa 200,000 marcatori ognuno ed un quarto beadchip ad alta densità, lo HumanOmniExpress BeadChip contenente circa 750,000 marcatori.

La strategia scelta è stata quella di genotipizzare con tutti e quattro i beadchip l'intera coorte afferente al progetto SardiNIA comprendente quindi tutti i 6.805 individui al fine di poter utilizzare il set di SNPs ottenuto dalla genotipizzazione come base di partenza per poter imputare il set di varianti provenienti dal sequenziamento dei 2,120 individui sardi su tutti i 6,805 individui della coorte sfruttando gli streach aplotipici condivisi da tutti gli individui della coorte e creando una mappa integrata. Procedimento che descrivo nel paragrafo 4.9. Per gli studi focalizzati su specifiche regioni genomiche per le quali non esistono prodotti standard per la genotipizzazione, il nostro gruppo ha avuto la possibilità di far parte di consorzi, in cui i vari gruppi di ricerca afferenti, hanno la possibilità di creare un pannello di marcatori condiviso relativo a determinate regioni genomiche, in genere con l'intento di seguire endofenotipi comuni di particolare interesse con l'obiettivo di sviluppare un ampio catalogo di varianti che forniscono una migliore copertura genomica nella popolazione. A tale scopo l'Illumina ha commercializzato diversi array tra cui lo Human CardioMetabochip, lo Human Immunochip ed infine l'ultimo usato lo Human Exome, contenenti circa duecento mila marcatori. Il nostro gruppo ha partecipato nel disegno di questi arrays.

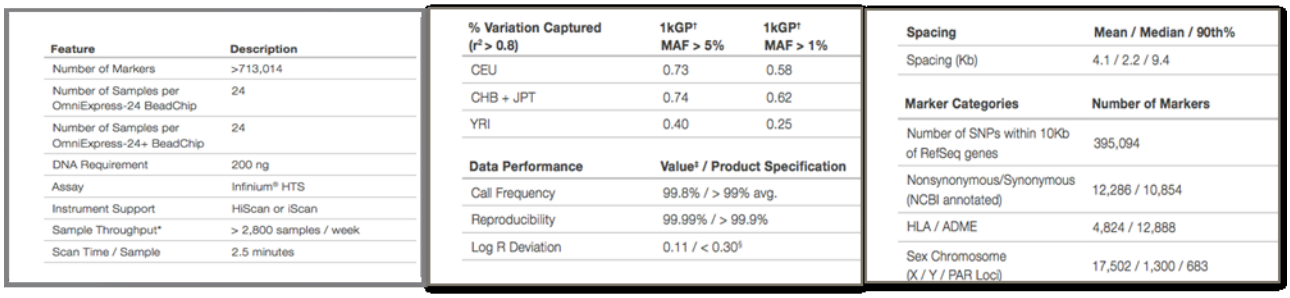
In dettaglio, l'Infinium Human CardioMetabochip contiene circa 200.000 marcatori identificati attraverso studi di metanalisi condotti a livello genome-wide per tratti/malattie metaboliche e cardiovascolari. Questo array è stato disegnato sulla base dei risultati dei GWAS condotti dai principali rappresentanti dei seguenti consorzi: CARDIoGRAM (coronary artery disease), DIAGRAM (type 2 diabetes), GIANT (height and weight), MAGIC (glycemic traits), Lipids (lipids), ICBP- GWAS (blood pressure), and QT- IGC (QT interval).

Invece l'Infinium Human Immunochip contiene circa 195,806 marcatori con 718 piccole inserzioni-delezioni provenienti da regioni genomiche che mostrano evidenza di associazione, almeno nominale, per 12 tratti o patologie legate alle malattie autoimmuni come: la cirrosi biliare primaria, la malattia infiammatoria intestinale, il diabete di tipo 1 e la sclerosi multipla. Lo scopo dell'Immunochip è quello di permettere un fine mapping ed uno studio di replicazione dei loci precedentemente trovati in associazione attraverso la loro genotipizzazione in una coorte molto più ampia rispetto a quella dello studio iniziale. L'array è stato disegnato nel 2009 da un folto gruppo di ricercatori afferenti a ben 11 differenti consorzi riguardanti patologie infiammatorie ed a carattere autoimmune.L'Infinium HumanExome BeadChip è stato sviluppato da un team di ricercatori internazionali di cui ha fatto parte il nostro gruppo. L'array è costituito da circa 250,000 (Figura 3.4) varianti funzionali esoniche identificate da oltre 12,000 singole sequenze derivanti dal sequenziamento del solo esoma così come dall'intero genoma con l'obiettivo di sviluppare un ampio catalogo di varianti esoniche. Le varianti incluse nell'array sono rappresentative di diverse popolazioni tra cui quella europea, africana, cinese ed ispanica, riguardanti una serie di condizioni comuni, quali il diabete di tipo 2, malattie metaboliche e disturbi psichiatrici.

Marker Categories	Number of Markers
Total markers	> 240,000
Number of unique RefSeq entries covered by at least 1 probe	> 20,000
Nonsynonymous SNPs (NCBI)	219,621
SNPs in splice sites	10,675
Stop variants	5,637
SNPs in promoter regions	7,012
SNPs in extended MHC region	5,158
GWAS tag markers[a]	4,761
HLA tags	2,061
Ancestry informative markers	3,468
Identity by descent markers	3,369
X / Y / mitochondrial markers	470 / 101 / 177
Indels	180
Variation Captured (r^2 > 0.8)[b]	**Fraction**
MAF > 5.0%	0.10
MAF > 2.5%	0.096
MAF > 1.0%	0.088

Figura 3.4 Rappresentazione del contenuto in termini di SNPs HumanExomechip (www.illumina.com)

L'Infinium HumanOmniExpress BeadChip è un potente strumento di analisi per gli studi di associazione genome-wide che contiene circa 750,000 marcatori ottimizzati nel contenuto di tag SNP da tutte e tre le fasi HapMap. Grazie al suo contenuto di marcatori è in grado di catturare una grossa parte di variabilità comune e guidare la scoperta di nuovi associazioni con tratti e malattie (Figura 3.5).



Feature	Description
Number of Markers	>713,014
Number of Samples per OmniExpress-24 BeadChip	24
Number of Samples per OmniExpress-24+ BeadChip	24
DNA Requirement	200 ng
Assay	Infinium® HTS
Instrument Support	HiScan or iScan
Sample Throughput*	> 2,800 samples / week
Scan Time / Sample	2.5 minutes

% Variation Captured (r^2 > 0.8)	1kGP† MAF > 5%	1kGP† MAF > 1%
CEU	0.73	0.58
CHB + JPT	0.74	0.62
YRI	0.40	0.25

Data Performance	Value‡ / Product Specification
Call Frequency	99.8% / > 99% avg.
Reproducibility	99.99% / > 99.9%
Log R Deviation	0.11 / < 0.30§

Spacing	Mean / Median / 90th%
Spacing (Kb)	4.1 / 2.2 / 9.4

Marker Categories	Number of Markers
Number of SNPs within 10Kb of RefSeq genes	395,094
Nonsynonymous/Synonymous (NCBI annotated)	12,286 / 10,854
HLA / ADME	4,824 / 12,888
Sex Chromosome (X / Y / PAR Loci)	17,502 / 1,300 / 683

Figura 3.5 Rappresentazione del contenuto in termini di SNPs HumanOmniExpress, (www.illumina.com)

3.4 Selezione del campione ProgeNIA per NGS e disegno sperimentale

Dei 6,805 individui appartenenti alla coorte SardiNIA, sono stati selezionati 2,120 campioni di DNA genomico per il pair-end whole genome sequencing. I campioni di DNA genomico alla opportuna concentrazione sono stati successivamente trasferiti a temperatura controllata presso i due laboratori di riferimento per la preparazione delle librerie: Università del Michigan presso il Medical School Core Sequencing Lab ed il CRS4, Centro di Ricerca, Sviluppo e Studi Superiori in Sardegna. I 2,120 campioni di DNA genomico sono costituiti da individui appartenenti a trios completi (genitori più figlio), un genitore più due figli e triplette di fratelli. L'uso delle famiglie aumenta il potere dello studio perché permette di definire con maggiore accuratezza la ricostruzione degli aplotipi condivisi tra gli individui (phasing) che a sua volta determina una qualità di imputazione maggiore.Tutti i campioni sono stati sequenziati ad una bassa copertura di 3-4 X, che significa che tutto il genoma è stato letto 3-4 volte. Eccezione fatta per due campioni sequenziati rispettivamente ad una coverage di 13X e 24X. Lo scopo dello studio era di analizzare l'individui nell'insieme ed usare l'informazione relativa al linkage disequilibrium per chiamare le varianti mediante inferenza statistica di genotipi mancanti in tutti i 6,805 individui della coorte, sfruttando come base di partenza lo scaffold di SNPs proveniente dalla genotipizzazione dell'intera coorte per imputare.

3.5 Genotipizzazione mediante piattaforma Illumina

Il protocollo utilizzato per effettuare la genotipizzazione mediante piattaforma Illumina è stato l'Infinium HD Assay che prevede l'uso di beadchip che permettono il caricamento simultaneo di 12 o 24 campioni sfruttando la tecnologia del BeadArray e dell'iScan system (Figura 3.6).

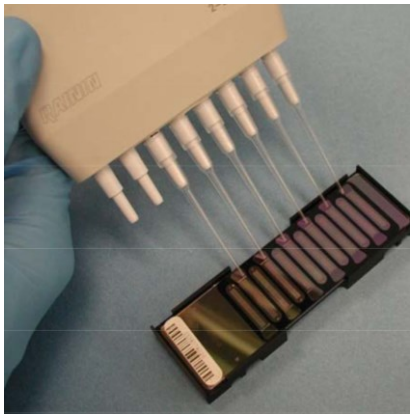


Figura 3.6 Rappresentazione schematica di caricamento di campioni sul Beadchip 12 posti, (www.illumina.com)

Questa tecnologia è basata sull'utilizzo di biglie in silice della grandezza di 3 micron assemblate in micro pozzetti ad una distanza di circa 5.7 micron.
Ad ogni biglia sono unite covalentemente centinaia di migliaia di copie di una sonda che è costituita da un oligonucleotide specifico che si andrà a legare al DNA target (Figura 3.7). La tecnologia Infinium usa due tipi di sonde I e II. Nella tecnologia Infinium II le sonde altamente specifiche, hanno una lunghezza di 50 basi, alle quali si andranno ad ibridare selettivamente frammenti di DNA con il locus di interesse fermandosi esattamente una base prima rispetto al marcatore interrogato. Invece le sonde Infinum I hanno l'estremità 3'terminale che si sovrappone con la base del marcatore interrogato e si verificherà l'estensione della base solo se vi è stato un match perfetto con il focus esaminato. La specificità di ibridazione viene ulteriormente conferita dall'estensione enzimatica di una singola base che prevede l'incorporazione di un nucleotide marcato che inseguito all'immissione di fluorescenza permetterà al sistema di visualizzazione dell'iScan di identificare sia il colore che l'intensità del segnale permettendo una corretta discriminazione allelica. I vari passaggi del protocollo sono descritti nel seguenti paragrafi :

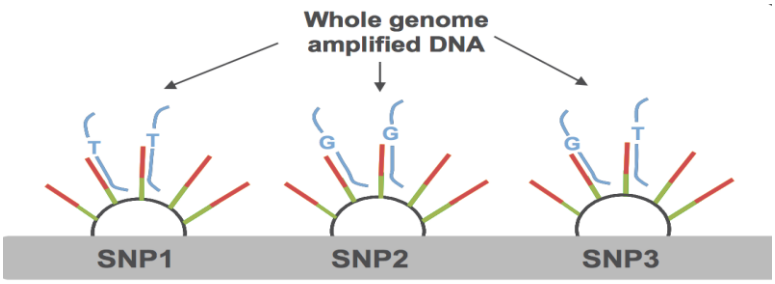


Figura 3.7 Rappresentazione schematica oligonicleotidi legati alle biglie.
(C.DalFiume,www.Illumina.com)

- Whole genome amplification WGA

WGA parte da 200 ng di DNA che vengono denaturati e poi amplificati mediante una reazione isotermica. Dopo questo passaggio il DNA viene amplificato uniformemente fino a mille volte.

• Frammentazione

Lo scopo di questa fase è quello di scindere il DNA in frammenti di lunghezza ottimale per l'ibridazione el campione sull'Illumina BeadChip. Si tratta di una frammentazione enzimatica che produrrà frammenti di DNA della lunghezza compresa tra 300 a 600 paia di basi. Il processo utilizza frammentazione di tipo endpoint, non sensibile al tempo evitando quindi un eccesso di frammentazione del campione di DNA.

• Precipitazione e risospensione

Il campione subisce una precipitazione alcolica con isopropanolo e successiva risospensione in un Buffer di ibridazione che permette il mantenersi delle condizioni ottimali affinché l'ibridazione del campione nel beadarray avvenga con successo.

• Ibridazione del campione sulla Beadarray

Durante l'ibridazione (Figura 3.8), il campione di DNA viene iniettato sul BeadChip dove avverrà l'ibridazione con le sonde specifiche per ogni locus. Le sonde sono complementari alle basi adiacenti allo SNP di interesse ma non lo includono. L'elevata specificità di questa ibridazione è assicurata dalla lunghezza delle sonde, dall'elevata stringendo del buffer di ibridazione utilizzato e dalle temperature elevate.

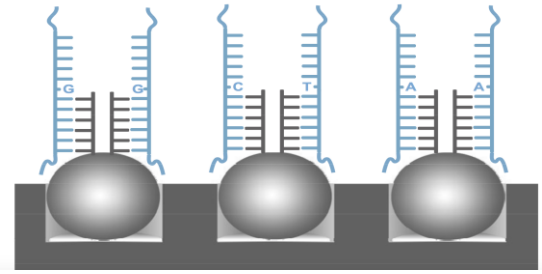


Figura 3.8 Schema di ibridazione (C.DalFiume,www.Illumina.com)

 Estensione

Lo scopo di questa fase è quello di permettere la discriminazione tra genotipi. La reazione di estensione aggiunge (Figura 3.9), una singola base marcata per ogni sonda. L'estensione di una singola base utilizza dideossinucleotidi terminatori di catena. I nucleotidi A e T sono marcati con dinitrofenile, DNP, ed i nucleotidi G e C sono marcati con biotina, quindi il tipo di marcatura identifica il genotipo dello SNP in questione.

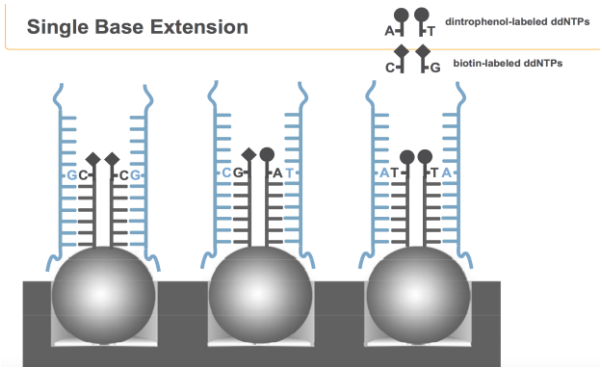


Figura 3.9 Schema di estensione a singola base. (C.DalFiume,www.Illumina.com)

 Colorazione

Lo scopo di questa fase è quello di applicare un segnale fluorescente specifico per ogni sonda marcata, figura 3.10.Il primo round di colorazione utilizza la streptavidina che emette sul verde che si legherà alla biotina ed il fluoroforo rosso contro il DNP. Il segnale viene amplificato attraverso i successivi cicli di incorporazione

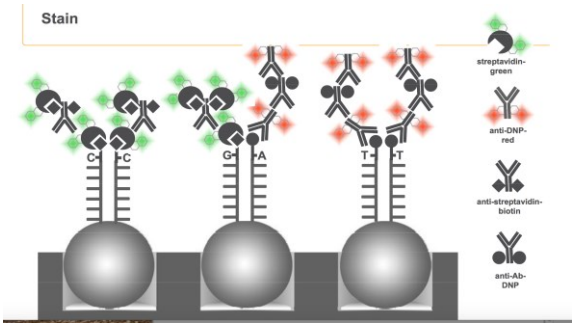


Figura 3.10 Rappresentazione schematica procedura di colorazione
(C.DalFiume,www.Illumina.com)

• Visualizzazione

Dopo la colorazione, il BeadChip è pronto per essere letto sul'iScan. Lo scopo di questa fase è quello di generare dati di intensità di fluorescenza che possono essere analizzati per effettuare le chiamate relative al genotipo di ogni SNP. Lo scanner iScan di Illumina utilizza il laser rosso e verde per eccitare i fluorofori e misurare il relativo segnale di intensità di fluorescenza. Il genotipo omozigote TT essendo marcati con DNP a cui si lega il fluoroforo rosso anti DNP produrrà un segnale principalmente rosso, mentre l' omozigote CC marcato con biotina a cui si lega la streptavidina produce principalmente segnale verde. L'eterozigote AG produrrà un segnale giallo dovuto all'emissione di fluorescenza rossa e verde nella stessa posizione della sonda.
L'immagine attraverso lo scanner produrrà dei file che racchiudono l'intensità di fluorescenza relative ad ogni singolo locus interrogato dalle sonde che vengono poi utilizzati dall'algoritmo GeneCall del software GenomeStudio per determinare la chiamata genotipica effettiva necessaria per gli step successivi dello studio Figura 3.11.

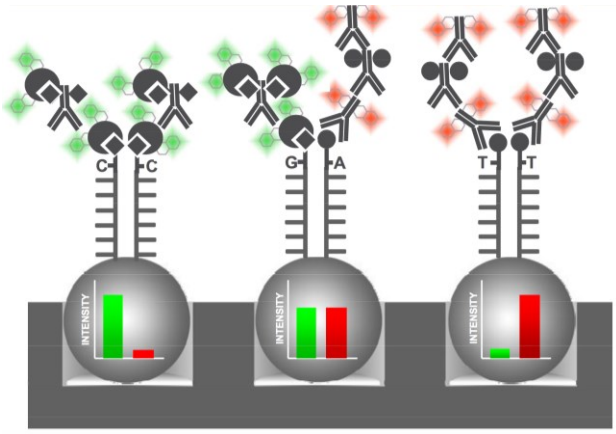


Figura 3.11 Rappresentazione schematica procedura di visualizzazione.
(C.DalFiume,www.Illumina.com)

Creazione di genotipi con il software GenomeStudio

Il software GenomeStudio analizza i dati generati con protocolli Illumina ed è composto da diversi moduli di analisi specifici ed integrati in un'unica piattaforma. Il software fornisce una serie di funzionalità mirate alla visualizzazione dei dati e dei risultati delle analisi generati dai singoli moduli. Il Framework di GenomeStudio visualizza i risultati in tre modalità differenti: global visualization, graphs e Tables (Figura 3.12) per consentire di esaminare in modo più gerarchico i dati ottenuti.

TABLES
Samples Table
Sequence or Lane Tables
SNP Table
Alleles Table
Probe Table
Gene, Exon, or Junction Tables
Plus other assay results tables

GLOBAL VISUALIZATION
Illumina Genome Viewer (IGV)
Illumina Chromosome Browser (ICB)
Illumina Sequencing Viewer (ISV)
GRAPHS
Dendrograms and Clustering Analysis
Heat Maps
Scatter Plots
Histograms
Line Graphs
Box Plots
Frequency Plots
Pie Charts

Figura 3.12 Schema di visualizzazione dati GenomeStudio (www.Illumina.com)

I dati possono essere esportati in diversi formati compatibili con i più comuni software di analisi. I dati relativi alla genotipizzazione vengono generati utilizzando il sistema iScan ed analizzati con il modulo di genotipizzazione del GenomeStudio. L'applicazione GenCall del GenomeStudio, incorpora un algoritmo di clustering, il GenTrain ed un algoritmo di chiamata.

In prima istanza si verifica la normalizzazione dei dati grezzi, il clustering degli SNPs ed infine la chiamata genotipica. L'algoritmo di clustering, GenTrain, prevede che i dati grezzi provenienti da ogni array vengano auto normalizzati utilizzando le informazioni contenute nella matrice dello stesso array mediante un algoritmo di normalizzazione che regola le variazioni nominali di intensità osservate nei due canali di colore, quali le differenze di sfondo tra i canali e le possibili interferenze tra i coloranti. Per ogni locus viene riportato un saggio a due colori, un colore per ognuno degli alleli e viene rilevata l'intensità di fluorescenza relativa ad ognuno dei due canali di colore A e B. Il comportamento di ciascun locus viene poi regolato utilizzando le varie intensità di fluorescenza i cui valori vengono rappresentati graficamente mediante assi cartesiani. I cluster corrispondenti alle intensità di fluorescenza rilevate sono rappresentate dai relativi genotipi AA, AB e BB, figura 3.13. Per semplificare il processo di raggruppamento, Gentrain trasforma le intensità A e B in due nuovi valori, chiamati θ e R (Figura 2) . θ quantifica la quantità relativa del segnale misurato dalle intensità A e B , definita dall'equazione : 2π - 1 arctan (AB e -1). R è una misura dell'intensità totale osservata dai segnali A e B, definito come: R = A + B.

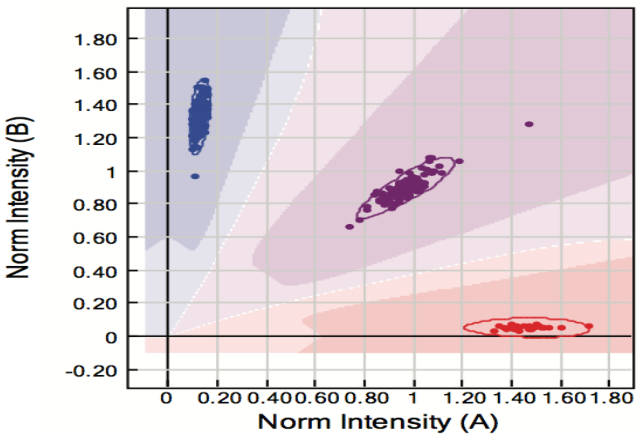


Figura 3.13 Rappresentazione grafica dei clusters (www.illumina.com)

In seguito all'attribuzione dei valori di fluorescenza viene stimato un punteggio statistico chiamato GenTrain score che prende in considerazione parametri quali la bassa intensità e la mancata corrispondenza tra i cluster esistenti e quelli previsti. Questo score viene incorporato nel modello di analisi utilizzato dall'algoritmo di chiamata al fine di permettere la corretta individuazione dei genotipi per ogni locus in esame definendo anche i limiti entro i quali la chiamata può essere ritenuta valida: il GenCall score. Punteggi superiori a 0,7 indicano genotipi utilizzabili per le analisi successive. Si definisce Callrate il rapporto tra il numero di genotipi superiori al valore soglia diviso il numero totale di genotipi.Il valore soglia di callrate corrisponde allo 0.98.

3.6 Sequenziamento Next Generation (NGS) di DNA genomico

Le sequenze relative ai 2.120 individui usate per la costruzione del pannello di referenza sardo-specifico sono state prodotte mediante tecnologia Illumina attraverso due sequenziatori di ultima generazione il Genome Analyzer IIx e l' Hi-Seq 2000.

La tecnologia Illumina sfrutta la chimica del sequenziamento per sintesi SBS utilizzando 4 nucleotidi fluorescenti per sequenziare le decine di milioni di cluster presenti sulla superficie della flow-cell. Durante ogni ciclo di sequenziamento, un singolo dNTP marcato viene aggiunto alla catena di acido nucleico. Il nucleotide marcato funge da terminatore reversibile di catena per l'attività polimerasica: dopo l'incorporazione del dNTP, il colorante fluorescente viene identificato tramite eccitazione laser e enzimaticamente staccato per consentire il prossimo ciclo di incorporazione. Poiché sono presenti tutti i 4 terminatori reversibili di catena A,C,T,G, la concorrenza naturale si riduce al minimo non pregiudicando il processo d'incorporazione. La chiamata delle basi viene effettuata ad ogni ciclo direttamente attraverso il rilevamento della fluorescenza riducendo notevolmente gli errori di chiamata. Il protocollo di sequenziamento consiste in quattro fasi principali: preparazione delle librerie, generazione dei clusters, sequenziamento ed analisi dei dati, Figura 3.14.

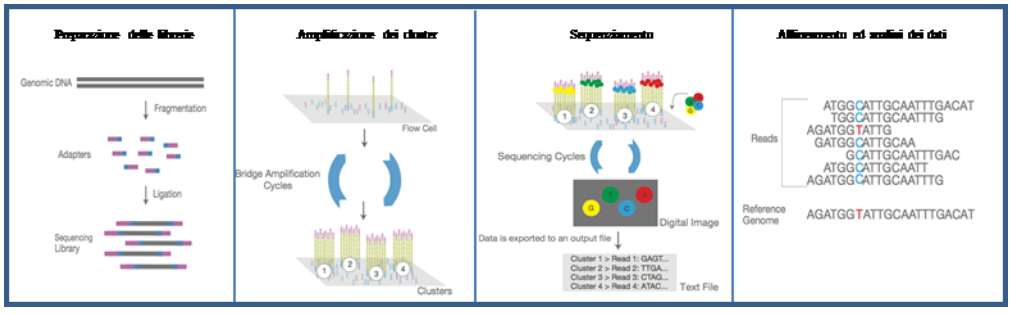


Figura 3.14 Schema protocollo di sequenziamento (www.illumina.com)

La preparazione delle librerie, consiste nella frammentazione meccanica del DNA genomico mediante nebulizzazione o sonicazione in frammenti non eccedenti le 800 paia di basi alle cui estremità 3'e 5' una volta essere state riparate e fosforilate sono stati aggiunti degli adattatori con l'impiego di DNA ligase. Gli adattatori sono complementari agli oligonucleotidi spotati sulla superficie della flowcell.
I prodotti di ligazione, di dimensione compresa tra le 300 e le 400 paia di basi vengono amplificati e purificati su gel di agarosio 2%. La concentrazione e la distribuzione dei frammenti delle librerie viene valutata mediante corsa su Bioanalyzer 2100 Agilent Technologies. Durante la preparazione delle librerie avviene un importante processo rappresentato dall'inserimento degli index che rappresentano una componente di adattatori o primer PCR costituiti generalmente da 8-12 paia di basi che permettono di raggruppare in maniera univoca le diverse librerie caricate in una delle lane della flowcell. I vari index di cui è nota la sequenza vengono poi individuati e ordinati mediante software permettendo di discriminare le diverse librerie. Il processo descritto è noto con il termine inglese " multiplexing ".

Una volta costituite, le librerie vengono ibridate sulla flowcell e caricate su uno strumento a flusso chiamato cBotTM. La flowcell è un vetrino con 1, 2 o 8 corsie (lane), separate fisicamente. Ogni lane è rivestita da 2 tipi di oligonucleotidi distinti denominati " p5 " e " p7 spotati a tappeto su tutta la superficie della flowcell che risultano essere complementari agli adattatori usati durante la fase di preparazione delle librerie. Sulla flowcell si verifica la formazione dei clusters attraverso il processo di amplificazione a ponte chiamato "bridge amplification" nel quale le due estremità di ogni inserto si legano covalentemente ai vari oligo complementari presenti sulla superficie della flowcell con la polimeri che determina dando luogo alla formazione di una struttura a ponte per diversi cicli di amplificazione ripetuti, il tutto si traduce in una amplificazione isotermica localizzata dei frammenti che da luogo alla formazione dei cluster. Ogni cluster è costituito da circa mille copie clonali dello stesso frammento in cui i primer di ogni frammento sono bloccati per evitare legami non specifici. Il passaggio successivo è rappresentato dal sequenziamento con il GAIIx, in corse paired- end da 240 basi o con l'Hi-Seq 2000 in corse da 202 basi, ottenendo un copertura media di 3-4X che significa che ogni base del genoma viene letta mediamente tre quattro volte. Il processo di sequenziamento utilizza quattro nucleotidi marcati con quattro diversi fluorocromi. Durante ogni ciclo di sequenziamento, un singolo dNTP viene aggiunto alla catena di acido nucleico ed il nucleotide marcato viene usato come terminatore di catena per il processo di polimerizzazione guidato dalla polimerasi. Dopo ogni ciclo di incorporazione di dNTP, il colorante fluorescente eccitato dal laser viene

utilizzato per identificare la base incorporata in base all'intensità di fluorescenza rilevata ed enzimaticamente clivato per consentire l'incorporazione del successivo nucleotide solo dopo aver ripristinato il gruppo ossidrilico OH sull'estremità 3' della base appena aggiunta in modo che possa accogliere la base successiva. Questo schema di incorporazione si ripete ad ogni ciclo e si svolge simultaneamente per ognuno dei frammenti di cui è costituito ciascun cluster. Terminata la corsa di sequenziamento si procede all'allineamento delle sequenze prodotte ed all'analisi dei dati. Le sequenze prodotte sono lunghe 100-200 basi, vengono chiamate reads e sono relative ad entrambi gli strand, da cui deriva la denominazione sequenziamento pair-end. L'allineamento delle reads contro il genoma di riferimento avviene mediante l'algoritmo di allineamento Burrows-Wheeler Alignment tools. In base ai segnali di intensità luminosa registrati dal sequenziatore avviene l'assegnazione di un primo parametro di qualità chiamato Phred score originale che viene successivamente rielaborato in seguito al confronto delle reads con il genoma di riferimento ed al tasso di errore rilevato dopo aver raggruppato le basi secondo il Phred score originale. In termini pratici il Phred score rappresenta la probabilità di errore nell'identificazione delle basi in termini matematici abbiamo: -10 log10[1/(1+99)]=20. Le reads con Phred score ricalibrato maggiore di 40 vengono ritenute valide al fine dell'identificazione di varianti quali SNPs, inserzioni, delezioni o varianti strutturali.

3.7 Sequenziamento Sanger automatizzato mediante elettroforesi capillare

Il sequenziamento mediante piattaforma ABI 3130xl è stato condotto per la validazione delle varianti imputate mediante pannello di referenza sardo specifico risultate essere associate ai livelli di HbA1 per chr12:123681790 ed ai livelli di HbF per rs183437571.

Il sequenziamento Sanger è stato sviluppato da Fred Sanger negli anni settanta. Viene comunemente denominata sequenziamento tradizionale in contrapposizione alle metodiche di sequenziamento di ultima generazione NGS. Il metodo Sanger è un metodo definito enzimatico poiché richiede l'utilizzo di un enzima, la DNA polimerasi per effettuare la reazione. Il metodo sfrutta la chimica dei dideossinucleotidi terminatori di catena i ddNTPs. Si tratta di nucleotidi modificati corrispondenti ai nucleotidi naturali, ma si differenziano per l'assenza del gruppo idrossilico sul carbonio 2' e 3' della molecola I dideossinucleotidi, a causa della loro conformazione, impediscono che un altro nucleotide si leghi ad essi, in quanto non si possono formare legami fosfodiesterici provocando quindi la terminazione della polimerizzazione ad opera della DNA polimerasi.Il protocollo classico

richiede un templato di DNA denaturato, i primers per iniziare la reazione di polimerizzazione, una DNA polimerasi, deossinucleotidi e dideossinucleotidi per terminare la reazione di polimerizzazione. Una volta effettuata l'amplificazione dei prodotti di estensione questi vengono quindi iniettati in un capillare per la successiva separazione elettroforetica. Durante l'elettroforesi capillare, i prodotti della reazione vengono iniettati nei capillari riempiti di polimero. L'alta tensione è applicata in modo che i frammenti di DNA carichi negativamente si muovano attraverso il polimero nei capillari verso l'elettrodo positivo. L'elettroforesi capillare è in grado di risolvere molecole di DNA che differiscono nel peso molecolare anche di un solo nucleotide.

Poco prima di raggiungere l'elettrodo positivo, i frammenti di DNA marcati e separati per dimensioni, si muovono attraverso il percorso di un raggio laser. Il raggio laser eccita i fluorofori con conseguente emissione di fluorescenza che viene catturata da un dispositivo di rilevamento ottico presente sul sequenziatore. Il software di raccolta dati converte tale segnale di fluorescenza in dato digitale registrando i dati in un file .ab1 *. Poiché ciascun colorante emette luce ad una lunghezza d'onda diversa quando eccitato dal laser, tutti i quattro colori, e pertanto tutte le quattro basi, possono essere rilevate e distinte in una iniezione capillare. I dati vengono visualizzati sotto forma di elettroferogramma (Figura 3.15) in cui ad ogni picco è associato un colore a cui corrisponde la rispettiva base nucleotidica.

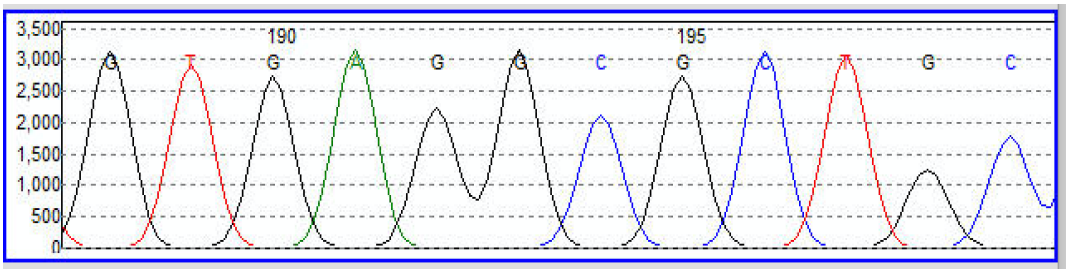


Figura 3.15 Rappresentazione elettroferogramma tipico (www.appliedbiosystem.com)

3.8 Genotipizzazione con la metodica TaqMan

La metodica è stata usata per la validazione delle varianti imputate usando il pannello di riferimento Sardo specifico. Le varianti validate con questa tecnologia sono state rs112233623 e rs7936823 per quanto riguarda i livelli di HbA2, le variati rs121909358 su GHR e rs2075870 come proxy del nostro segnale di associazione rs150199504 nel KCNQ1.

Per la replicazione dei locus in esame abbiamo utilizzato il protocollo relativo alla discriminazione allelica che si basa sulla chimica chiamata TaqMan probe-based chemistry e ABI Prism 3700 HT. Nel nostro saggio abbiamo utilizzato le probe definite TaqMan MGB (minor groove binder) che di-

spongono di un quencher (NFQ) non fluorescente. Brevemente, la sonda fluorogenica oligonucleotidica risulta essere marcata con un fluoroforo legato covalentemente all'estremità 5' terminale ed un quencher legato all' estremità 3',Figura 3.16. I fluorofori utilizzati per la genotipizzazione sono VIC e FAM per ogni di due alleli.

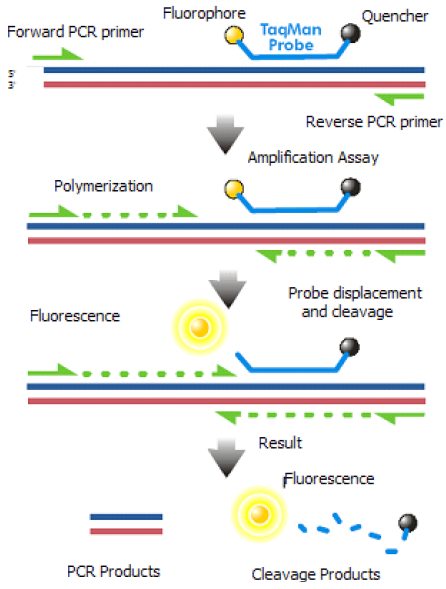


Figura 3.16 Schema TaqMan probe-based chemistry (www.appliedbiosystem.com)

Quando la sonda è intatta, quindi reporter e quencher sono presenti, la vicinanza del quencher riduce fortemente la fluorescenza emessa dal fluorocromo reporter mediante il trasferimento spaziale dell'energia di risonanza di fluorescenza (FRET; Förster resonance, Förster, V. T. 1948). Durante i cicli di estensione, in seguito alla presenza dell' allele target, la sonda esegue un annealing nella zona tra i primer e viene distrutta dall'attività 5′ nucleasica della Taq DNA polimerasi durante l'estensione. In questo modo il fluorocromo reporter viene separato dal quencher, permettendo alla emissione del segnale del fluorocromo reporter.

Il software SDS elabora i segnali di fluorescenza rilevati dallo strumento e li trasforma in chiamata genotipica effettuando uno discriminazione allelica in base al tipo di fluorescenza rilevata. Ognuno dei due alleli del polimorfismo in esame viene indagato da un fluoroforo specifico (Figura 3.17). Gli omozigoti per l'allele 1 ad esempio avranno un tipo di fluorescenza riconducibile esclusivamente al fluoroforo VIC mentre gli omozigoti per l'allele 2 avranno un tipo di fluorescenza riconducibile esclusivamente al fluoroforo FAM.

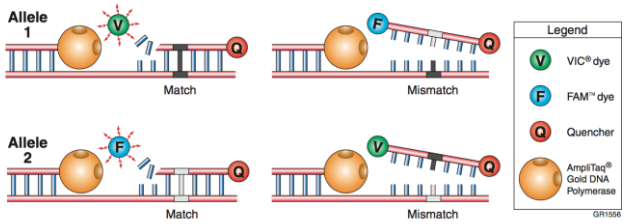


Figura 3.17 Rappresentazione annealing (www.appliedbiosystem.com)

3.9 Analisi statistica: costruzione del pannello di imputazione ed imputazione statistica

L'imputazione statistica si basa sul principio dell'inferire probabilisticamente varianti non tipizzate direttamente combinando gli aplotipi parziali trovati in un individuo, con gli aplotipi meglio caratterizzati in termini di copertura genomica, derivanti da pannelli di riferimento come quelli risultanti dal sequenziamento dell'intero genoma al fine di propagare i genotipi mancanti a tutti gli individui dello studio che non sono stati direttamente tipizzati per un dato set di marcatori.
Nello specifico il nostro metodo utilizza le informazioni relative alla vicinanza fisica dei marcatori fiancheggianti lo polimorfismi puntiformi (SNP) di interesse nonché il grado di relazione tra familiari per stimare l'IBD identificando porzioni di aplotipo condivise con parenti stretti che sono stati caratterizzati per un elevato numero di marcatori ed inferendo probabilisticamente i genotipi mancanti. La base di partenza nel processo di imputazione è rappresentata dal subset di varianti da utilizzare come base di genotipi per l'imputazione.

A tale scopo i 6,805 individui della coorte SardiNIA, come precedentemente descritto, sono stati genotipizzati a diversi gradi di risoluzione mediante quattro arrays. I genotipi ottenuti sono stati sottoposti a rigorosi controlli di qualità. Sono stati esclusi i campioni con call rate inferiore al 95% per lo HumanOmniExpress e con call rate inferiore al 98% per i tre arrays custom. I quattro arrays sono stati analizzati indipendentemente rimuovendo i marcatori con call rate inferiore al 98%, quelli che si discostavano dall'equilibrio di Hardy Weinberg definito come $p < 1 \times 10\text{-}6$, i marcatori monomorfici o con una frequenza dell'allele minore (MAF) inferiore all'1% per quanto riguarda l'HumanOmniexpress, e quelli che presentavano un eccesso di errori mendeliani. Sono stati rimossi gli SNPs in comune tra i quattro arrays che mostravano un elevato livello di discordanza o che generavano una non concordanza maggiore dell'1% quando analizzati all'interno di 13 coppie di ge-

melli. I genotipi degli SNPs in comune tra i quattro arrays sono stati paragonati tra loro ed i genotipi discordanti sono stati esclusi, il tasso di concordanza per il set di SNPs finale è superiore al 99.99%. Dopo aver effettuato i sopracitati controlli di qualità abbiamo ottenuto una base di 890,542 SNPs da poter usare nel processo di imputazione. Il passo successivo è stato quello di inferire probabilisticamente i genotipi mancanti su tutti i 6,805 individui della coorte utilizzando le 17 milioni varianti presenti nel pannello di riferimento creato in seguito al sequenziamento a bassa copertura, circa 3-4X, relativo ai 2,120 individui sardi. Creando come riportato nella Figura 3.18 il panello di referenza per l'imputazione - la mappa integrata dei genomi Sardi.

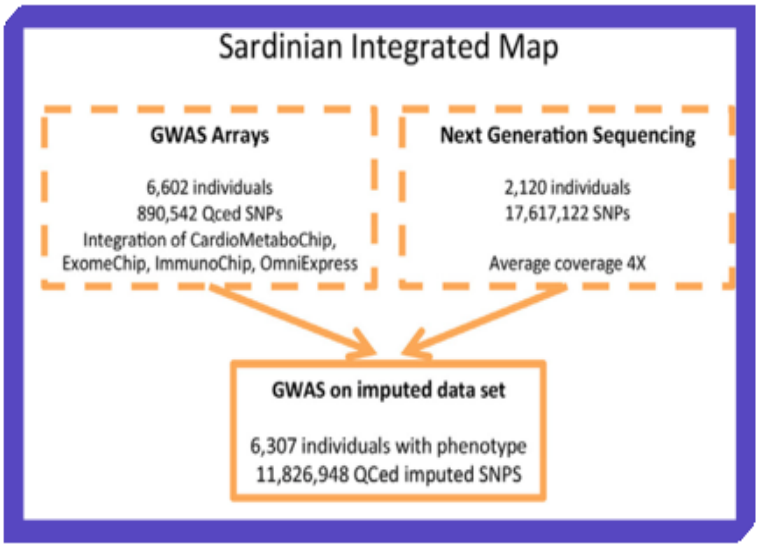


Figura 3.18 Schematica rappresentazione della costruzione del panello di referenza.

RISULTATI

Capitolo 4

In questo capitolo espongo tre esempi di studi di associazione dove il GWAS è stato basato sui dati mirati del sequenziamento dell'intero genoma, estensiva genotipizzazione e l'imputazione. Come esempio ho scelto il GWAS sui livelli di emoglobine, lipidi ed altezza umana.

4.1 Studio di associazione su tutto il genoma riguardanti i livelli di emoglobina

Il GWAS relativo ai tratti ematologici comprendenti i livelli di emoglobina HbA1,HbA2 ed HbF è stato condotto su 6.305 individui per i quali erano disponibili i dati di genotipizzazione e le misurazioni relativi ai livelli delle tre forme di emoglobina in esame. La tabella 4.0 riporta le statistiche in dettaglio con la popolazione in esame divisa tra individui portatori e non portatori della mutazione beta zero 39, nota per essere la principale causa della forma di talassemia beta in Sardegna.

Emoglobina	Portatori β039 (N=643) (5th percentile - 95th percentile)	Non portatori β039 (N=5,662) (5th percentile - 95th percentile)
Total Hb (g/dl)	12.000 (10.300 - 14.000)	14.000 (11.900 - 16.200)
HbA1 (g/dl)	11.168 (9.557 - 13.057)	13.538 (11.519 - 15.746)
HbA2 (g/dl)	0.699 (0.556 - 0.871)	0.382 (0.277 - 0.483)
HbA2 (%)	5.800 (5.000 - 6.700)	2.700 (2.200 - 3.200)
HbF (g/dl)	0.107 (0 - 0.275)	0.036 (0 - 0.126)
HbF (%)	0.900 (0 - 2.400)	0.300 (0 - 0.900)

Tabella 4.0 Descrizione fenotipica di tratti ematologici della coorte SardiNIA

I risultati del GWAS condotto sui livelli di emoglobina HbA1, HbA2 ed HbF sono presenti nella Tabella 4.1. In totale sono stati individuati 23 segnali in 10 locus, di cui 5 erano nuovi.

					SNP with highest CADD score and r>0.9 with lead			
	Lead SNP (chr:position)	rsID from dbsnp142	CADD score		SNP	score	r^2 with lead	Annotation
HbA1 (g/dl)	12:123681790	–	0,153		12:123465483	12.05	1	TFBS region
HbA2	16:88601281	rs141006889	5.19		–	–		–
	6:41952511	rs113267280	2.99		rs112233623	6.34	0.99	TFBS region
	20:44547672	rs59329875	0.19		rs1057208	12.85	0.97	TFBS region
HbF (g/dl)	19:13121899	rs183437571	12,90		–	–	–	fattore importante per l'eritropoiesi

Tabella 4.1 Annotazione funzionale dei 5 nuovi segnali trovati nel GWAS per tratti ematologici mediante CADD score

Per quanto riguarda i livelli di HbA1 il segnale sul cromosoma 12: 123681790 è localizzato nell'introne del gene *MPHOSPH9*. Questo locus comprende diversi SNPs associati con il tratto in forte linkage disequilibrium con il nostro top che ricadono in una regione ad alto contenuto genico come mostrato sulla Figura 4.2.

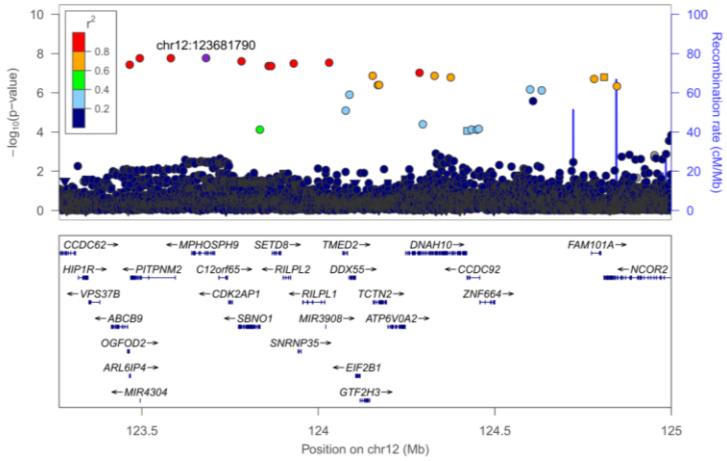


Figura 4.2 Plot regionale per il locus associato con i livelli HbA1 sul chr12

Tra i vari segnali individuati in questa regione genomica menziono il 12:123465483, localizzato in una regione altamente conservata e ricca di siti di legame per fattori di trascrizione che possiede un punteggio di CADD score pari a 12.5. Questo segnale è poco sotto il valore empirico di significati-

vità di $p=1.4x10^{-8}$ ma va notato che risulta essere associato anche ad una forma di emoglobina l'HbA2,con significatività pari a $p=5.9x10^{-5}$.

La Figura 4.3 presenta 3 nuovi segnali individuati nel GWAS per i livelli di HbA2. Il primo segnale si trova sul chr:88601281 (rs141006889) nel gene *ZFPM1*, chiamato anche *FOG1*, che codifica un importante cofattore per i fattori di trascrizione eritropoietica GATA1 e GATA2. Mutazioni a carico di questo complesso che ne alterano la sua stabilità sono responsabili di forme familiari di anemia diseritropoietica e trombocitopenia.

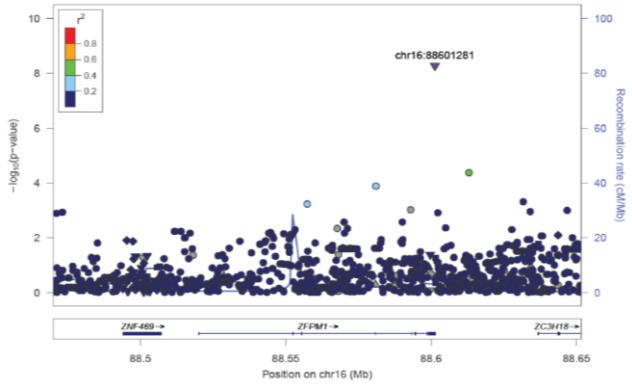


Figura 4.3 Plot regionale con la variante chr16:88601281 più associata a i livelli di HbA

Il secondo segnale ricade in chr6:41952511. Come si vede dal plot regionale presentato nella figura 4.4, il nostro top rappresentato dal cerchio viola, risulta essere in alto linkage disequilibrium con un altra variante: rs112233623.

Le due varianti sono localizzate nel gene *CCND3*, il cui prodotto, la ciclina D3 svolge un ruolo critico nel processo eritropoietico. La variante top chr6:41952511 è rappresentata dal cerchio viola.

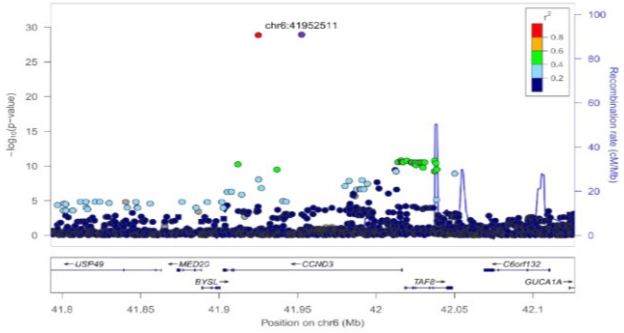


Figura 4.4 Plot regionale del locus localizzato sul chr6 associato con i livelli HbA2.

Il terzo segnale associato ai livelli di HbA2 è la variante chr20:44547672, (rs59329875), rappresentato nella Figura 4.5. La variante ricade tra il gene *PLTP* e *PCIF1*.

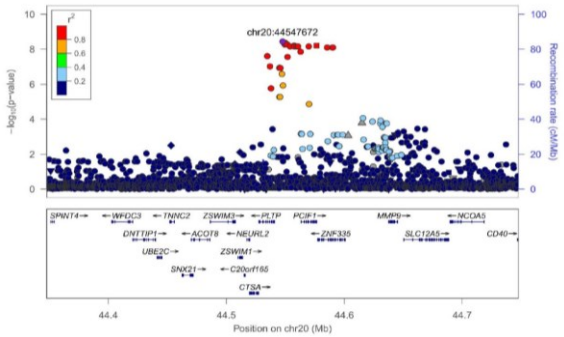


Figura 4.5 Plot regionale dei livelli di HbA2 con la variante top - chr20: 44547672

Il gene PLTP risulta essere associati con i livelli di trigliceridi e diverse lipoproteine plasmatiche, mentre PLTP è regolatore dell'espressione genica attraverso RNA polimerasi di tipo II. In stretto LD con il nostro top c'è un'altra variante - rs1057208, localizzata nel gene *PCIF1* che possiede un CADD score pari a 12.85.

La Figura 4.6 mostra il quinto nuovo segnale trovato associato ai livelli di emoglobina fetale - la variante chr19:13121899 (rs183437571). Questa variante è ,localizzata nell'introne del gene *NFIX* che codifica per un fattore di trascrizione che lega il motivo CCAAT. La variante in esame risulta essere di poco sotto il livello di significatività $p=1.4x10^{-8}$, ma la sua importanza biologica è giustificata dal fatto che ricade in una regione ricca di siti GC che risulta essere mutilata in maniera differenziale nelle cellule progenitrici della linea eritroide, nell'età fetale e adulta. In aggiunta, nel topo il gene *NFIX* è stato identificato come un fattore che regola la sopravvivenza delle cellule staminali e progenitrici durante il processo eritropoietico.

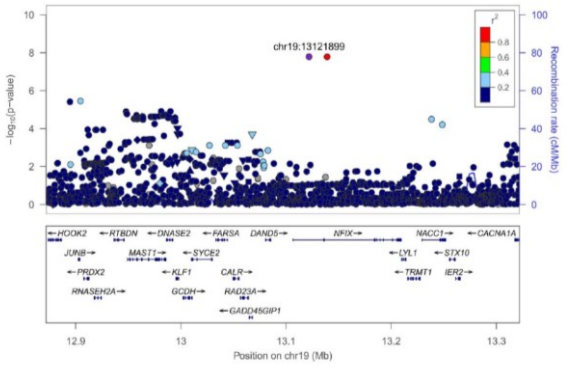


Figura 4.6 Plot regionale di livelli HbF con la variante top - chr19:13121899

Dei 5 nuovi segnali di associazione individuati nel GWAS per i livelli di emoglobine, i tre segnali riportati nella tabella 4.2 sono stati imputati mediante l'uso del pannello di riferimento Sardo. L'imputazione statistica porta un tasso di errore. E' stato quindi necessario procedere alla validazione delle varianti. La tabella 4.2 mostra per ogni locus i risultati della validazione, la piattaforma utilizzata ed il livello di significatività ottenuto nell'analisi primaria ed in quella ottenuta dopo avere sostituito i genotipi inferiti con quelli genotipizzati tramite metodi diretti Sanger o Taqman. Le metodiche di validazione utilizzate sono state descritte nel capitolo 3 materiali e metodi.

SNP	Metodo validazione	p-value studio primario	p-value replicazione
chr12:123681790	Sanger	$1,68x10^{-08}$	$2,10x10^{-08}$
rs112233623	TaqMan	$1,11x10^{-29}$	$3,28x10^{-30}$
rs183437571	Sanger	$1,61x10^{-08}$	$1,79x10^{-08}$

Tabella 4.2 Validazione segnali di associazione provenienti dal GWAS sui livelli di emoglobine

La replicazione dei segnali di associazione individuati nel nostro studio è stata condotta usando la coorte TwinsUK del Regno Unito, costituita da 4,131 individui genotipizzati a diversi gradi di risoluzione mediante gli array Illumina: HumanHap300, HumanHap610Q, 1M-Duo and 1.2 MDuo 1M. Dei 5 nuovi segnali di associazione individuati sono stati replicati con successo i segnali chr20: 44547672 (rs59329875) ed il chr6: 41952511 (rs113267280), entrambi riguardanti i livelli di HbA2. La tabella 4.3 riporta i risultati della replicazione dei loci. Per ogni SNP viene riportato l'allele, l'effetto dell'allele testato ed il livello di significativa.

Locus#	Trait	SNP	Allele	Effect (SE)	P-value	Note
1	HbA1 (g/dl)	chr12:123681790		-	-	Non inclusa nel pannello 1000 Genomi
2	HbA2 (%)	rs113267280	G/T	0.442 (0.118)	$1.73x10^{-4}$	
3	HbA2 (%)	rs59329875	C/T	0.132 (0.029)	$6.98x10^{-6}$	
4	HbA2 (%)	rs141006889			-	Non inclusa nel pannello 1000 Genomi
5	HbF (%)	rs183437571	T/C	-	-	Monomorfico nella coorte TwinsUK

Tabella 4.3 Risultati replicazione emoglobine nella coorte TwinsUK

Per quanto riguarda i locus 1, 4, 5 rispettivamente associati con i livelli di HbA1 ed HbA2 la replicazione non è stata possibile in seguito alla mancanza di queste varianti nei pannelli pubblici di riferimento analizzati come riportato nella Tabella 4.4.

Locus#	Trait	SNP	Allele	Effect (SE)	P-value	Note
1	HbA1 (g/dl)	chr12:123681790		-	-	Non inclusa nel pannello 1000 Genomi
2	HbA2 (%)	rs113267280	G/T	0.442 (0.118)	$1.73x10^{-4}$	
3	HbA2 (%)	rs59329875	C/T	0.132 (0.029)	$6.98x10^{-6}$	
4	HbA2 (%)	rs141006889			-	Non inclusa nel pannello 1000 Genomi
5	HbF (%)	rs183437571	T/C	-	-	Monomorfico nella coorte TwinsUK

Tabella 4.4 Risultati di replicazione 5 nuovi segnali per HbA1, HbA2 ed HbF nella coorte TwinsUK

Tre dei nostri cinque nuovi segnali sono stati identificati esclusivamente mediante l'uso del pannello di riferimento sardo specifico. In dettaglio vediamo che la variante chr12:123681790 per i livelli di HbA1 non è presente nel pannello 1000 genomi, mentre la variante rs141006889 per i livelli di HbA2 non è presente nel pannello 1000 genomi ma è stato incluso nel disegno di Exomechip in seguito ai risultati del nostro sequenziamento. La variante rs183437571 più associata con i livelli di HbF mostra invece una scarsa qualità di imputazione con l'uso del pannello 1000 genomi con un segnale di associazione che non raggiunge la soglia di significativa di $p=1.4x10^{-8}$. L'elevata efficienza del nostro pannello di riferimento interno è mostrata nella figura 4.7, notiamo che il potere di imputazione quando si usa il pannello di riferimento sardo specifico raggiunge l'80% prendendo come riferimento cinque diverse classi di frequenza allelica e differenti numeri di marcatori testati.

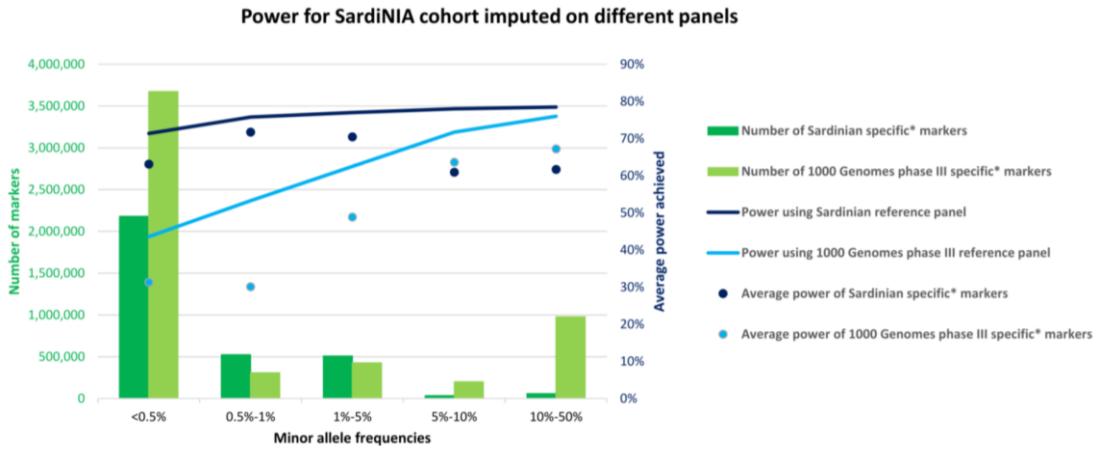


Figura 4.7 Efficienza di imputazione tramite pannello Sardo specifico e pannello 1000 genomi

L'ereditabilità stimata per i livelli di HbA1,HbA2 ed HbF è pari a 0.520 g/dl, 0.728% e 0.633% rispettivamente. La proporzione di varianza fenotipica del tratto spiegata dalla varianti trovate associate è pari a 0.240 g/dl per HbA1, 0.492% per HbA2 e 0.383% per HbF. Questo dimostra il caratteristico assetto poligenico dei tratti quantitativi che prevede che la varianza totale sia il risultato di tante varianti con effetto modesto.Tre dei cinque nuovi segnali individuati mostrano associazione con un secondo tipo di emoglobina. La tabella 4.5 mostra come la direzione degli effetti sia la stessa in entrambe le forme di emoglobina modulate dalla variante.

Tratto	cromosoma: posizione	Allele	Effetto (SE)	Pvalue	Direzione dell'effetto		
					HbA1	HbA2	HbF
HbA1	12:123681790	A/C	- 0.3606 (0.064)	1.68×10^{-8}	—	—	
HbA2	16:88601281	G/A	- 0.5074 (0.087)	5.33×10^{-9}		—	
	6:41952511	G/T	0.2923 (0.026)	1.11×10^{-29}		+	+
	20:44547672	C/T	- 0.1399 (0.024)	3.64×10^{-9}		—	
HbF	19:13121899	T/C	0.4607 (0.081)	1.61×10^{-8}			+

Tabella 4.5 Effetti pleiotropici delle tre forme di emoglobina.

4.2 Studio di associazione riguardante due varianti con ampio effetto sulla statura umana

I risultati del GWAS sui tratti antropometrici ed in particolare all'altezza umana sono mostra ti nella Figura 4.8.

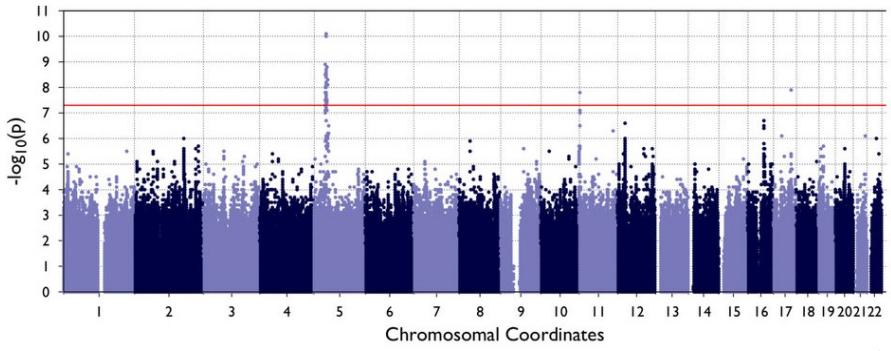


Figura 4.8 Manhattan plot del GWAS sull'altezza umana

Il primo dei due segnali di associazione è quello relativo al locus rs121909358 con $P = 1.07 \times 10^{-10}$ come riportato nella Figura 4.9.

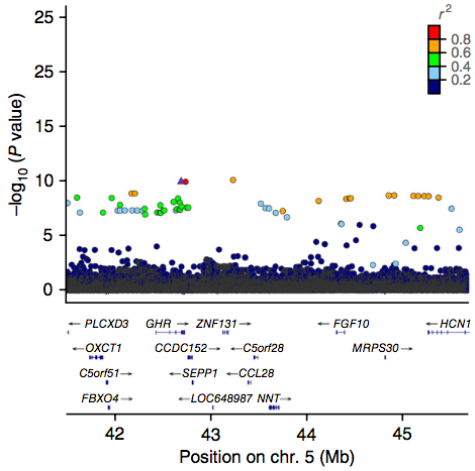
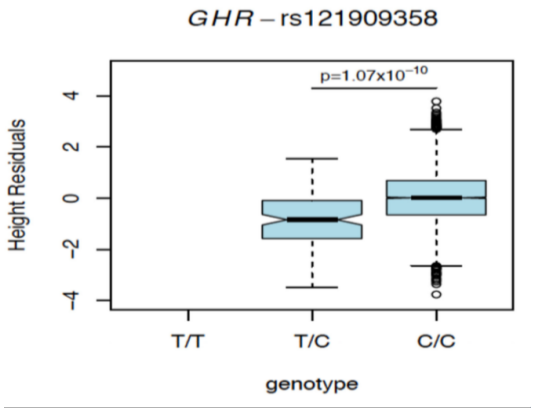


Figura 4.9 Plot regionale del locus *GHR* con la variante top - rs121909358

La variante è localizzata sul gene *GHR* che codifica per il recettore dell'ormone della crescita. Il legame dell'ormone della crescita al recettore porta alla dimerizzazione del recettore stesso ed all'attivazione di un segnale di trasduzione intra e intercellulare.



La variante determina una perdita di funzione del codone di terminazione che codifica per la sostituzione R61X. In particolare l'allele T conferisce una diminuzione della statura pari a - 4,2 cm, in portatori della variante come riportato nella Figura 4.10

Figura 4.10 Distribuzione dei residui rispetto ai genotipi della variante rs121909358

La variante è stata individuata grazie all'uso del pannello di referenza sardo specifico, con una qualità di imputazione pari a 0.94 e risulta essere molto rara o assente nelle altre popolazioni con una frequenza inferiore a 1/60.000 individui.

L'associazione è stata replicata nella una coorte indipendente di 5,314 individui sardi appartenenti allo studio Ogliastra Genetic Park (OGP). La frequenza e l'effetto di rs121909358 stimato nella coorte di replicazione sono inferiori rispetto ai valori identificati nello studio primario (MAF = 0.46% in 857 individui non imparentati; *one-tailed P* = 0.015; effetto = −0.31 s.d. corrispondente ad una riduzione pari a 1.89 cm). Interessante notare che gli omozigoti T/T per questa variante sono affetti da una sindrome autosomica recessiva congenita, rara, chiamata Sindrome di Laron caratterizzata da una ridotta statura.La frequenza mondiale dei portatori della sindrome di Laron è in base ai dati Orphanet (http://www.orpha.net), sotto lo 0.01%. Interessante notare che tra il sottogruppo di 1,481 individui della coorte SardiNIA non imparentati tra loro, calcoliamo una frequenza pari allo 0.8% e consistentemente con ciò, individuiamo un solo individuo affetto dalla Sindrome di Laron sul bacino di 10,721 individui che popolano i quattro paesi oggetto del studio SardiNIA. Nella coorte abbiamo individuato altre due mutazioni note per essere associate alla Sindrome di Laron come riportato nella tabella 4.6. Entrambe le mutazioni però raggiungono una frequenza troppo bassa per poter determinare effetti fenotipici negli eterozigoti.

SNP	Cambio amminoacidico	Conta allelica	MAF %
rs121909358	p.R61X	129	0,87
rs121909362	p.R179C	1	0,033
rs143814221	p.Y240H	6	0,2

Tabella 4.6 Frequenza delle mutazioni della Sindrome di Laron trovate nei 1.481 individui della coorte SardiNIA

66 individui eterozigoti per la mutazione R61X e 49 individui non portatori della mutazione appartenenti alla coorte SardiNIA sono stati sottoposti alle misurazioni antropometriche, tabella 4.7 L'analisi tramite t-test tra i due gruppi non ha evidenziato disproporzioni corporee significative (p=0.23). Il 30% degli individui portatori esaminati evidenziavano una ridotta estensione del gomito tipica degli individui affetti.

Portatori uomini Sindrome di Laron 54, Wild type 2620		Media	SD	Correlazione con BMI	Correlazione con altezza	T-test p value
Età	Portatori	45.7	14.04	0.05	-0.27	0.79
	Non portatori	45.2	16.08	0.44	-0.53	
Altezza	Portatori	160.6	7.34	0.19	x	$1.55x10^{-7}$
	Non portatori	167	7.17	-0,260	x	
BMI	Portatori	26.7	3.05	x	-0.19	0.79
	Non portatori	26.4	4.03	x	-0,26	
Peso	Portatori	69	10.09	0.77	0.45	$5.7x10^{-4}$
	Non portatori	73.7	11.54	0.84	0.28	
Portatrici donne Sindrome di Laron 75, Wild type 3558		Media	SD	Correlazione con BMI	Correlazione con altezza	T-test p value
Età	Portatori	42.2	17.01	0.55	-0.45	0.87
	Non portatori	44.56	16.07	0.53	-0.52	
Altezza	Portatori	150.9	6.18	-0.34	x	$9.21x10^{-8}$
	Non portatori	155.1	6.56	-0.32	x	
BMI	Portatori	24.6	4.09	x	-0.34	0.98
	Non portatori	24.8	4.09	x	-0.32	
Peso	Portatori	55.9	10.06	0.9	0.06	0.0126
	Non portatori	59.4	11.26	0.9	0.09	
Età menarca	Portatori	13.03	1.03	0.25	-0.15	0.22
	Non portatori	13.01	1.74	0.06	-0.13	

Tabella 4.7 Caratteristiche antropometriche dei portatori e non portatori della mutazione R61X nel gene GHR

Il secondo segnale individuato nel GWAS riguardante l'altezza umana si trova nel locus rs150199504 sul gene *KCNQ1*, MAF = 7.7% in Sardegna ed inferiore all'1% nelle altre popolazioni Europee. Questo locus risulta essere sotto imprinting per cui è stato necessario valutare il *parent of origin effect* per stabilire se l'allele con effetto sul tratto fosse stato ereditato dal padre o dalla madre. La Figura 4.11 e la tabella 4.8 mostrano i risultati di questa analisi.

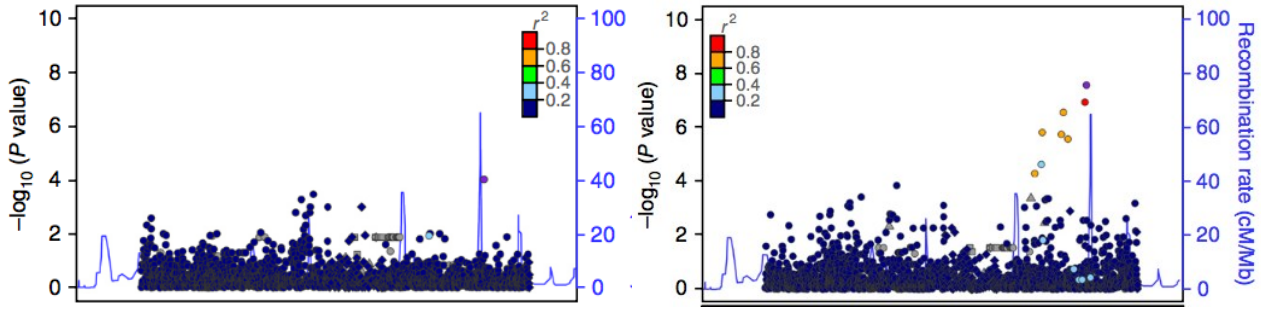


Figura 4.11 Plot regionale dell'effetto paterno (A) e materno (B) nel locus *KCNQ1*

B

rs ID	Alleli	MAF	Effetto Materno	Effetto Paterno
rs150199504	C/G	0.083	5.56×10^{-9}	0.9488
rs143840904	T/C	0.094	3.92×10^{-8}	0.9653
rs2075870	A/G	0.094	6.97×10^{-8}	0.793
rs149658560	A/G	0.076	2.93×10^{-7}	0.8183
rs12790610	G/A	0.095	4.73×10^{-7}	0.3531
rs67004488	G/A	0.104	5.21×10^{-7}	0.3875

Tabella 4.8 Risultato dell'analisi del *parent of origin* effect nella regione del gene *KCNQ1*

Non abbiamo riscontrato nessuna significatività per quanto riguarda l'effetto paterno $P = 0.95$, mentre l'effetto materno ha raggiunto una significatività pari a $P = 5.6 \times 10^{-9}$ con un effetto di −0.315 s.d. corrispondente ad una riduzione dell'altezza pari a 1.83 cm.

In particolare la riduzione dell'altezza è conferita dall'alleale G quando ereditato dalla madre come mostrato nella figura 4.12.

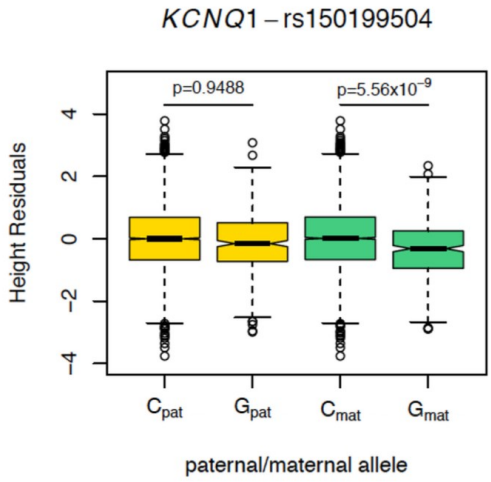


Figura 4.12 Distribuzione dei residui rispetto ai genotipi rs150199504

Come mostrato nella figura 4.11 e nella tabella 4.8 nella regione del gene *KCNQ1* sono presenti diversi segnali di associazione che mostrano un moderato LD con il nostro top. Per chiarire meglio il contributo di queste varianti abbiamo effettuato la replicazione dei risultati in un gruppo di 19,053 individui appartenenti a 6 coorti europee: TwinsUK, ALSPAC, TEENAGE, HA, HP, UKHLS. La tabella 4.9 mostra i risultati della replicazione nelle 6 coorti.

rs ID	Alleli	MAF %	pvalue
rs150199504	G/C	0.9	$2.82x10^{-4}$
rs143840904	T/C	2.0	$1.23x10^{-3}$
rs2075870	A/G	3.3	$9.19x10^{-2}$
rs149658560	A/G	3.0	$6.52x10^{-2}$
rs67004488	G/A	4.7	$2.34x10^{-1}$

Tabella 4.9 Replicazione dei segnali di associazione nel locus KCNQ1

La variante rs12790610 è stata rimossa dal pannello in quanto non ha passato i controlli di qualità, mentre il nostro top continua a mantenere una significatività ed un effetto maggiore pur avendo la MAF più bassa. Anche l'analisi condizionale del locus ha mostrato la variante rs150199504 come la principale responsabile del segnale di associazione nella regione.

I risultati sino ad ora mostrati relativi al GWAS sull'altezza hanno individuato due varianti rare nelle altre popolazioni ma relativamente più frequenti in Sardegna e per di più con un importante effetto sulla statura.

Per stimare l'effetto sulla statura di tutti varianti descritti finora nella popolazione Sarda e altre popolazioni Europee abbiamo applicato il *Polygenic height scores*. I risultati di questa analisi sono presenti nella figura 4.13

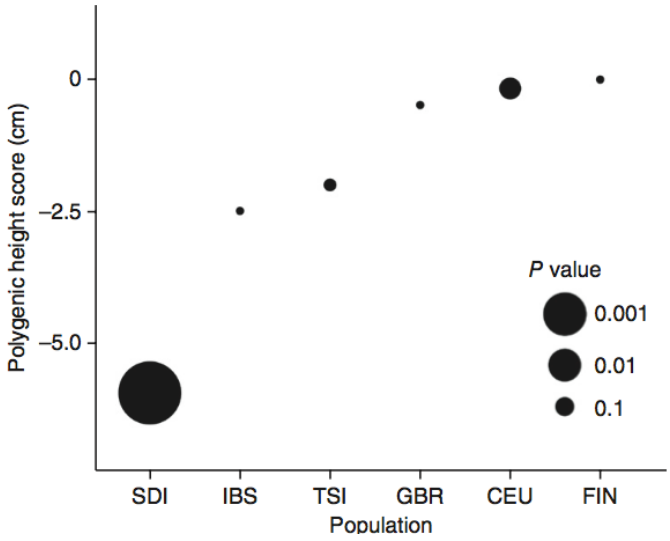


Figura 4.13 Stima del *Polygenic height scores* per la popolazione Sarda e diverse popolazioni Europee.

Nell'analisi abbiamo considerato la misura della frequenza totale degli alleli legati all'altezza ponderando ogni allele in base al suo *effect size* ed alla frequenza. Nel modello sono state incluse due nuove varianti: rs121909358 sul *GHR*, la su rs150199504 *KCNQ1* ed i 691 locus precedentemente descritti come associati all'altezza nello studio GIANT.

Le popolazioni prese in esame sono quella Sarda (SDI), Iberici (IBS), toscani (TSI), Britannici (GBR), Finlandesi (FIN) ed Utha (CEU). I risultati indicano un basso score poligenico per gli individui Sardi, il che comporta un arricchimento degli alleli che conferiscono una riduzione dell'altezza nella popolazione sarda, indice questo di un probabile effetto di selezione sul tratto.

4.3 Studio di associazione relativo ai livelli di LDL, colesterolo totale e trigliceridi

Per quanto riguarda i tratti cardiovascolari ed in particolare i livelli di LDL, Colesterolo Totale e Trigliceridi i risultati del nostro GWAS indicano 14 varianti indipendenti distribuite su 11 locus di cui 2 nuovi locus, tabella 4.10.

Tratto	Gene	Cromosoma: posizione	Alleli	Frequenza	Effetto (SD)	pvalue
LDL						
	PCSK9	1:55505647	T/G	0.038	−0.406 (0.053)	1.73×10^{-14}
	SORT1	1:109821307	G/T	0.180	0.156 (0.027)	1.87×10^{-8}
	HBB	11:5248004	A/G	0.048	−0.473 (0.051)	1.17×10^{-20}
	CILP2	19:19456917	T/C	0.074	−0.232 (0.042)	2.58×10^{-8}
	APOE	19:45412079	T/C	0.036	−0.645 (0.053)	2.47×10^{-33}
	APOE	19:45411941	C/T	0.074	0.264 (0.039)	1.21×10^{-11}
COLESTEROLO TOTALE						
	PCSK9	1:55505647	T/G	0.038	−0.390 (0.053)	1.69×10^{-13}
	TMEM33	4:41980435	G/A	0.013	−0.520 (0.091)	6.94×10^{-9}
	HBB	11:5248004	A/G	0.048	−0.490 (0.05)	6.88×10^{-22}
	CILP2	19:19456917	T/C	0.074	−0.260 (0.041)	2.15×10^{-10}
	APOE	19:45412079	T/C	0.036	−0.544 (0.053)	2.06×10^{-24}
	APOE	19:45411941	C/T	0.074	−0.210 (0.038)	2.18×10^{-8}
HDL						
	LPL	8:19815256	T/A	0.125	0.257 (0.046)	2.70×10^{-8}
	LIPC	15:58687603	T/C	0.467	0.136 (0.021)	7.96×10^{-11}
	CETP	16:56989590	T/C	0.268	0.190 (0.023)	2.37×10^{-16}
	TGIF1	18:3412386	T/C	0.026	−0.448 (0.082)	4.49×10^{-8}
TRIGLICERIDI						
	LPL	8:19845376	T/C	0.209	−0.160 (0.026)	8.36×10^{-10}
	APOA5	11:116661101	T/G	0.025	−0.450 (0.064)	1.24×10^{-12}
	APOA5	11:116664040	G/A	0.172	0.160 (0.027)	4.64×10^{-9}
	CILP2	19:19456917	T/C	0.074	−0.260 (0.039)	2.14×10^{-11}

Tabella 4.10 Risultati del GWAS sui Lipidi utilizzando il pannello di imputazione sardo specifico

Il primo segnale - chr11:5248004 ricade sul gene *HBB*. La variante conosciuta come **β ^{0}39,** causa la principale forma in Sardegna di talassemia β. Questa variante è associata ad una riduzione dei livelli di LDL pari a 13.9 mg/dl, ed ad una riduzione dei livelli di colesterolo totale pari a 16.9 mg/dl nel

nostro GWAS. Questi effetti sono concordi con quanto precedentemente descritto in letteratura da Maioli et al.,(1989)

Nei plot regionali della figura 4.14 vediamo i risultati del processo di imputazione relativi all'uso del pannello di riferimento sardo specifico rispetto al pannello 1000 genomi.
Utilizzando il pannello 1000 genomi non individuiamo la variante chr11:5248004, ma un segnale intergenico relativo a rs76053862 (tabella 4.10) distante 122 Kb da $\beta^0$39che rappresenta il secondo segnale usando il pannello di referenza Sardo con una significatività $P = 1.4 \times 10^{-13}$

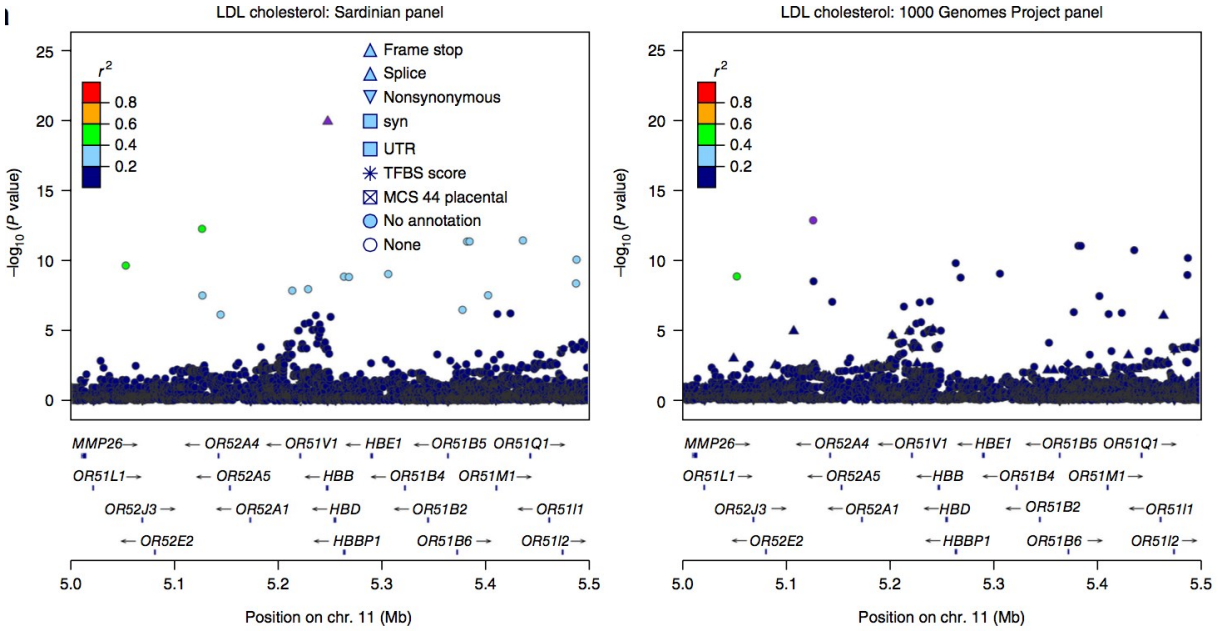


Figura 4.14 Il confronto di efficienza dei due pannelli di referenza usati per l'imputazione – il pannello di referenza Sardo e il panello di referenza del progetto di 1000 Genomes

Il secondo segnale è rappresentato dalla variante chr11:116661106 localizzata sul gene *APOA5* ed è associata con un livello di significatività pari a $1.2x10^{-12}$ad una diminuzione dei livelli di trigliceridi pari a 20.7 mg/dl con una frequenza del 3% nella popolazione Sarda. La variante spiega circa l'1% della variabilità fenotipica osservata per il tratto ed è indipendente dal noto segnale chr11:116664040-rs10750097 ,(MAF = 17%; effetto = 11.9 mg/dl; $P = 4.6x10^{-9}$).

Questa variante è stata inclusa nel ExomeChip in seguito al sequenziamento condotto dal nostro gruppo di ricerca che ha partecipo al disegno sperimentale dell'array. Queste due varianti dimostrano come nello stesso locus possano trovarsi due varianti indipendenti tra loro, una Sardo specifica, l'altra molto frequente nelle altre popolazioni europee. Oltre i due locus su HBB e APOA5, abbiamo individuato altri due segnali intergenici (tabella 4.10), uno tra TMEM33 e DCAF4L1-

chr4:41980435 associato con i livelli di colesterolo totale e l'altro vicino a TGIF1-chr18:3412386 associato ai livelli di LDL e relativo all'analisi sesso specifica, solo nelle femmine. Entrambi i segnali sono al di poco sotto la nostra soglia di significatività di $6.9x10^{-09}$ ed il loro significato biologico non è chiaro. Come mostrato dalla tabella 4.11 entrambi i segnali non sono identificabili se usiamo pannello di 1000 genomi.

Tratto	Gene	Cromosoma: posizione	Alleli	Frequenza	Effetto (SD)	pvalue	R2
LDL	PCSK9	1:55505647	T/G	0.038	-0.406(0.053)	$1.51x10^{-14}$	1.0
	SORT1	1:109821307	G/T	0.180	-0.156(0.027)	$2.06x10^{-08}$	0.5
	OR52E, OR52A5	11:5125982	T/C	0.049	-0.403(0.054)	$1.44x10^{-13}$	0.9
	CILP2	19:19456917	T/C	0.074	-0.233(0.042)	$2.59x10^{-08}$	0.5
	APOE	19:45412079	T/C	0.037	-0.645(0.053)	$2.47x10^{-33}$	2.4
	APOE	19:45411941	C/T	0.074	0.264(0.039)	$1.21x10^{-11}$	0.8
COLESTEROLO TOTALE							
	PCSK9	1:55505647	T/G	0.038	-0.38(0.053)	$9.79x10^{-14}$	1.0
	OR52E, OR52A5	11:5125982	T/C	0.048	-0.407(0.054)	$5.95x10^{-14}$	1.0
	CILP2	19:19456917	T/C	0.074	-0.264(0.041)	$2.15x10^{-10}$	0.6
	APOE	19:45412079	T/C	0.036	-0.544(0.053)	$2.26x10^{-24}$	1.7
	APOE	19:45411941	C/T	0.074	-0.215(0.038)	$3.09x10^{-08}$	0.5
HDL							
	LPL	8:19815256	T/A	0.125	0.257(0.046)	$2.70x10^{-08}$	1.2
	LIPC	15:58687603	T/C	0.467	0.137(0.021)	$1.19x10^{-10}$	0.7
	CETP	16:56987015	C/T	0.268	0.190(0.023)	$1.96x10^{-16}$	1.1
TRIGLICERIDI							
	LPL	8:19938902	CAAA T/C	0.2078	-0.169(0.027)	$3.14x10^{-10}$	0.7
	APOA5	11:116661101	T/G	0.025	-0.450(0.064)	$1.24x10^{-12}$	0.9
	APOA5	11:116664040	G/A	0.171	0.174(0.027)	$1.13x10^{-10}$	0.7
	CILP2	19:19456917	T/C	0.074	-0.260(0.039)	$2.14x10^{-11}$	0.8

Tabella 4.11 Risultati del GWAS sui Lipidi utilizzando il pannello di imputazione 1000

4.4 Discussione dei risultati

In questa tesi di dottorato ho presentato i risultati di tre GWAS sui livelli di emoglobine, LDL, trigliceridi, colesterolo totale ed altezza umana.

Il punto di forza dei risultati ottenuti è stato l'uso del *Whole Genome Sequenced-Based GWAS* basato su una popolazione isolata, quale quella Sarda che ci ha permesso di chiarire l'architettura genetica dei Sardi i quali grazie a fenomeni quali la deriva genica e la selezione naturale mostrano un arricchimento di varianti che risultano essere rare nelle altre popolazioni. Questa particolare differenziazione genetica ci ha permesso di condurre il nostro studio ed ottenere grazie all'applicazione del disegno sperimentale scelto dei risultati estremamente soddisfacenti. In più l'uso di un disegno di studio di tipo *population sequencing* quindi sequenziamento dell' intero genoma a bassa copertura in molti individui con carattere di*"discovery"* permette di trovare più variabilità nella popolazione in confronto a quando si usano pochi individui e si sequenzia ad alta copertura.

Il primo aspetto che voglio mettere in evidenza è l'elevata efficienza in termini di *discovery* del pannello di referenza Sardo specifico che ci ha permesso di individuare alcuni locus genici contenenti varianti che forniscono importanti informazioni biologiche, come la variante chr11:5248004 la cui associazione non veniva evidenziata con l'imputazione mediante il pannello di referenza 1000 genomi. Abbiamo mostrato che questa variante fondatrice ha molti effetti pleiotropici. Nel GWAS sui livelli di lipidi, abbiamo visto come questa variante, nota in Sardegna per essere la causa della principale forma di talassemia di tipo beta è risultata essere associata ad una diminuzione dei livelli di LDL e colesterolo totale, spiegandone una grossa parte di variabilità. Per la prima volta è stata identificata la variante associata a questa riduzione già mostrata a livello plasmatico nel lavoro Maioli et al. in 1989. Gli eterozigoti per questa variante mostrano una riduzione dei livelli di LDL e colesterolo totale ed una conta assoluta di globuli rossi pari al 23% più alta rispetto ai non portatori della mutazione. Questo è perfettamente in linea con il fatto che la diminuzione plasmatica dei lipidi negli individui portatori della mutazione è la conseguenza dell'aumentato assorbimento di colesterolo totale ed LDL da parte dei macrofagi e istiociti del sistema reticolo-endoteliale.

Il GWAS sui livelli di emoglobine ha messo in evidenza altri tre segnali di associazione - chr12:123681790 per HbA1, rs141006889 per HbA2 ed rs183437571 per HbF, identificati esclusivamente mediante il pannello di referenza sardo specifico, segnali molto importanti dal punto di vista funzionale in quanto localizzati in regioni genomiche ricche di siti di legame per i fattori di trascrizione ed il segnale su HbF indica invece un importante fattore per il processo eritropoietico. Le varianti descritte nel GWAS per i livelli di emoglobine sono di particolare interesse biomedico per-

ché possono avere la potenziale funzione di modificatori, che possono modulare o predire lo stato dei pazienti con la talassemia.

Un altro importante risultato ottenuto grazie all'uso della popolazione Sarda, è stato quello di dimostrare per la prima volta, come una grossa parte di ereditabilità ancora non spiegata per quanto riguarda l'altezza umana sia dovuta a varianti rare responsabili di patologie monogeniche come il caso della variante rs121909358 sul *GHR*, come pure a varianti comuni nella popolazione Sarda ma rare nelle altre popolazioni Europee come la variante rs150199504 sul gene *KCNQ1*.

Lo studio sulla statura ha portato tante domande di cui per una parte stiamo ancora cercando le risposte. L'analisi tramite *Polygenic Height Score* ci ha mostrato come la popolazione Sarda presentasse un arricchimento degli alleli che conferiscono una riduzione dell'altezza umana. Una delle ipotesi che sono state fatte era che la selezione del tratto sembra sia antecedente al popolamento della Sardegna da parte della linea di discendenza che conduce ai primi contadini europei che si ritiene abbiano colonizzato l'isola. Di conseguenza è possibile che la bassa statura sia stata selezionata negli antichi contadini per risparmiare le risorse energetiche? Per quale motivo i Sardi hanno mantenuto o acquisito questo carattere? La risposta a questa domanda ha portato all'analisi di altre popolazioni che discendono dalle popolazioni neolitiche Europee, come Toscani e Spagnoli. La statura dei sardi rimane sempre più bassa. Sulla base di queste analisi lo studio conclude che l'abbassamento della statura è proseguito nella popolazione Sarda dopo il popolamento della Sardegna come risposta alle condizioni di ristrette risorse nutrizionali nell'ambiente dell' isola. Esempi di selezione per la ridotta statura sono stati descritti in letteratura per i grandi mammiferi.

Molto curioso il fatto che uno dei loci associati con la bassa statura che è stato selezionato in Sardegna, è localizzato nella regione sotto imprinting dove l'espressione dell'allele materno della variante nel gene *KCNQ1* porta l'abbassamento della statura dei figli. Rimane intricante e da svelare se la madre ha un particolare beneficio dalla selezione dell' espressione del suo allelenei figli nel contesto della riproduzione e dello sviluppo.

Sfortunatamente il fatto che in Sardegna alcune varianti siano "spinte" ad un alta frequenza rimanendo rare o assenti nelle altre popolazioni pone il problema della difficoltà di replicare questi segnali di associazione come dimostrano i risultati dei tre GWAS. In questi casi replicare l'associazione in altre popolazioni richiede grandi numeri di centinaia di migliaia di persone per ottenere l'associazione a livello nominale come abbiamo mostrato per tutti e tre i tratti oggetto dello studio. D'altra parte la frequenza più alta di certe varianti in Sardegna permette di valutare la loro causalità più facilmente.

Un altro problema degli studi GWAS è individuare la variante causativa. Molto spesso il picco di associazione mostra molte varianti con comparabile valore p. In questo caso è possibile arrivare alla variante causativa usando le popolazioni con diverso *linkage disequilibrium* nella regione. Nel caso in cui la variante sia rara, ma con frequenze diverse, tra le due popolazioni e con sufficiente potere statistico, ma nella seconda popolazione non la troviamo associata con il tratto si può concludere che la variante non è causativa. Nello specifico abbiamo condotto il *fine-mapping* e individuato la variante causativa utilizzando per il lavoro dell'altezza come coorte di replicazione, gli individui appartenenti alla popolazione inglese e individui Sardi appartenenti al progetto SardiNIA.

Lo studio SardiNIA costituisce una risorsa preziosa di dati genetici e fenotipici. Concludendo ci sono ancora tante domande a cui rispondere ma si può affermare che il *Sequencing based GWAS* basato sulla popolazione Sarda ed in particolare su quella Ogliastrina, ci abbia permesso di ottenere dei risultati estremamente incoraggianti grazie alla validità del disegno sperimentale scelto.

Conclusioni

L'oggetto della mia tesi di dottorato è stata la dissezione dei tratti quantitativi di potenziale interesse biomedico nell'ambito del progetto SardiNIA utilizzando come metodica di analisi il Whole Genome Sequenced-based GWAS. Partecipare ad uno studio di questo tipo implicava la capacità di muoversi nell'ambito della genetica quantitativa, un mondo ricco di complessità vista la natura multifattoriale dei caratteri in esame. Il nostro gruppo di ricerca è impegnato da anni in questo campo come dimostrano le oltre 100 pubblicazioni che ci vedono coinvolti a vario grado. Nonostante tutto, l'impegno è stato massimo e sempre volto alla massima onestà intellettuale. Questo percorso è iniziato con l'elaborazione del progetto sperimentale e la sua successiva attuazione. Le scelte effettuate e la strategia utilizzata hanno comportato la collaborazione e l'impegno di diverse figure professionali: medici, biologi e biostatistici, con l'impiego di non indifferenti risorse economiche e logistiche. Il tutto è iniziato con la scelta della piattaforma di genotipizzazione da utilizzare e nel contempo la scelta dell'array, che nello specifico sono stati quattro. Scelta non facile ma decisiva per lo sviluppo dello studio. Quasi contemporaneamente alla genotipizzazione è iniziato il processo di sequenziamento, anche qui innumerevoli ragionamenti per la scelta del gruppo di individui che sarebbe stato oggetto del sequenziamento. Dalla scelta del gruppo di individui dipendeva il tipo di informatività genetica ottenuta. Terminati questi due processi abbiamo unito i dati provenienti dalla genotipizzazione e dal sequenziamento per poter creare la cosiddetta mappa integrata tramite la quale abbiamo potuto costruire un pannello di referenza sardo specifico ed inferire i genotipi mancanti a tutti gli individui della coorte, aumentando così il potere statistico delle nostre analisi. Questo ci ha permesso di ottenere circa undici milioni di varianti da poter usare nei GWAS relativi a diversi tratti quantitativi in esame: LDL, colesterolo totale e trigliceridi per i tratti cardiovascolari, altezza umana per i tratti antropometrici ed i livelli di HbA1, HbA2 ed HbF per i tratti ematologici. I risultati sono stati estremamente soddisfacenti e si sono conclusi con la pubblicazione nella prestigiosa rivista Nature Genetics che a prova dell'importanza dei nostri studi ha dedicato la copertina e l'editoriale alla Sardegna. La nostra isola ha avuto un ruolo fondamentale nella riuscita dei nostri studi che ci hanno permesso di raggiungere un grado di risoluzione genetica molto elevato concedendoci di chiarire ulteriormente l'architettura genetica dei Sardi.

Abbiamo evidenziato un ampio set di varianti rare nelle altre popolazioni ma con un alta frequenza in Sardegna, varianti molto importanti dal punto di vista biologico. Un esempio è la variante responsabile della principale forma di talassemia di tipo beta in Sardegna, una mutazione stop-gain che nel nostro studio è stata trovata associata ai livelli di LDL e colesterolo totale di cui spiega una

grossa parte di variabilità. Qui è entrato in gioco il potere del nostro disegno sperimentale nel creare ed usare un pannello di riferimento sardo specifico. La variante in questione non veniva identificata usando il pannello di riferimento relativo al progetto 1000 genomi a causa della sua bassa frequenza nelle altre popolazioni e la frequenza inferiore allo 0.1%. Lo stesso scenario vale per altre tre varianti, la rs121909358 sul gene *GHR* associata con la statura umana e le varianti chr12:123681790, rs141006889 rispettivamente implicate nella regolazione dei livelli di HbA1 ed HbA2. Questo dimostra come varianti spinte ad alta frequenza in Sardegna ed assenti o addirittura rare, nelle altre popolazioni possano avere un importante significato biologico e possano condurre verso studi funzionali mirati. Questo scenario evidenzia però anche la difficoltà di poter replicare varianti popolazione specifiche rare o assenti nelle restanti popolazioni, nello specifico abbiamo potuto superare questa difficoltà usando una coorte Sarda indipendente dalla nostra, afferente al progetto Ogliastra Genetic Park (OGP). Le nostre scoperte dimostrano per la prima volta come parte dell'ereditabilità non spiegata possa essere dovuta a varianti rare con ampio effetto sul tratto, ed implicate in patologie monogeniche come il caso della variante sul gene *GHR*. In particolare i risultati del GWAS sull'altezza umana ci hanno premesso di individuare due varianti con ampio fatto non individuate in studi che implicavano il coinvolgimento di centinaia di migliaia di individui, come il caso del consorzio GIANT che riporta circa 691 varianti implicate nella riduzione dell'altezza umana che nel loro insieme spiegano circa il 16% della variabilità del tratto con un effetto modesto che non va oltre lo 0.3 cm contro l'effetto del nostro segnale su *GHR* e *KCNQ1* che conduce ad una riduzione dell'altezza umana rispettivamente pari a 4.2 ed 1.83 cm in uno studio che coinvolge 6,182 individui contro i 183,727 dello studio GIANT.

In conclusione si può dire che la differenziazione genetica dei sardi, dovuta a fenomeni quali la deriva genica e la selezione naturale, ha fatto in modo che un elevato numero di varianti rare in Europa potessero avere invece un alta frequenza in Sardegna. Dal nostro studio si è potuto stimare che 76,286 varianti ritenute rare o assenti con frequenza minore dello 0,5%, nel pannello fase 3 del progetto 1000 Genomi, sono risultate avere una frequenza maggiore del 5% nella nostra coorte di riferimento.

Abbiamo sfruttato questo particolare assetto genetico dei Sardi per applicare il Whole genome SGWAS creando un pannello di riferimento Sardo specifico che ci ha permesso in prima istanza di imputare variante rare estremamente bene e nello stesso tempo di studiare un totale di 6,805 individui, allo stesso costo usando metodi di analisi tradizionali quali il sequenziamento ad alta copertura, avremo potuto sequenziare 2,120 genomi.

I risultati ottenuti ci hanno consentito di dimostrare come il Whole genome Sequenced-based GWAS inserito nel contesto di uno studio longitudinale che coinvolge un popolazione isolata quale quella Sarda, possa essere un utile strumento per poter spiegare parte dell'ereditabilità fino ad ora non spiegata, ponendo le basi per effettuare ricerche nel campo della medicina personalizzata e sull'origine e la storia della popolazione Sarda.

Referenze

1. Sidore, C. et al. Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. Nat. Genet. doi:10.1038/ng.3368. (2015)

2. Danjou, F. et al. Genome-wide association analyses based on whole-genome sequencing in Sardinia provide insights into regulation of hemoglobin levels. Nat. Genet. doi:10.1038/ng.3307. (2015)

3. Zoledziewska, M. *et al.* Height-reducing variants and selection for short stature in Sardinia. *Nat. Genet.* doi:10.1038/ng.3403. (2015).

4. Pilia, G. et al. Heritability of cardiovascular and personality traits in 6,148 Sardinians. PLoS Genet. 2, e132. (2006)

5. Peltonen, L., Palotie, A. & Lange, K. Use of population isolates for mapping complex traits. Nat. Rev. Genet. 1, 182–190. (2000).

6. Francalacci, P. et al. Peopling of three Mediterranean islands (Corsica, Sardinia, and Sicily) inferred by Y-chromosome biallelic variability. Am. J. Phys. Anthropol. 121, 270–279 (2003).

7. Francalacci, P. et al. Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. Science 341, 565–569. (2013).

8. Zavattari, P.et al. Major factors influencing linkage disequilibrium by analysis of different chromosome regions in distinct populations: demography, chromosome recombination frequency and selection. Hum. Mol. Genet. 9, 2947–2957. (2000)

9. Lampis, R. et al. The inter-regional distribution of HLA class II haplotypes indicates the suitability of the Sardinian population for case-control association studies in complex diseases. Hum Mol Genet. 9(20):2959-65. (2000).

10. Eaves, IA. Et al. The genetically populaions of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. Nat Genet.25(3):320-3 (2000).

11. Bansal, V. et al. Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet. 11(11):773-85. doi: 10.1038/nrg2867.(2010).

12. Manolio, TA. et al. Finding the missing heritability of complex diseases. Nature 461(7265): 747–53. (2009).

13. Visscher, PM. et al. Heritability in the genomics era, concepts and misconceptions. Nat Rev Genet. 9(4):255-66. doi: 10.1038/nrg2322. (2008).

14. Civelek, M. et al. Systems genetics approaches to understand complex traits. Nat Rev Genet. 15(1):34-48. doi: 10.1038/nrg3575.(2015).

15. Li, B. et al. A likelihood-based framework for variant calling and de novo mutation detection in families. PLoS Genet.;8(10):e1002944. doi: 10.1371/journal.pgen. (2012).

16. Pistis, G. et al Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. Eur J Hum Genet. 23(7):975-83. doi: 10.1038/ejhg.2014.216. (2015)

17. Chen, W. Et al. Genotype calling and haplotyping in parent-offspring trios. Genome Res. 23(1):142-51. doi: 10.1101/gr.142455.112. (2013)

18. Cirulli ET, Goldstein DB Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 11(6): 415–25. (2010).

19. Cappello, N. et al. Data on 12 polymorphisms in 21 linguistic domains. Ann Hum Genet. 60(Pt 2):125-41.(1996)

20. Piras, I. et al. Genome-wide scan with nearly 700 000 SNPs in two Sardinian sub-populations suggests some regions as candidate targets for positive selection. *European Journal of Human Genetics* **20,** 1155–1161; doi:10.1038/ejhg.2012.65. (2012).

21. Lango Allen, H. et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467, 832–838 (2010).

22. Laron, Z. Laron Syndrome—From Man to Mouse (Springer-Verlag, 2011).
Turchin, M.C. et al. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. Nat. Genet. 44, 1015–1019 (2012).

23. Soranzo, N. et al. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. Nat. Genet. 41, 1182–1190 (2009).

24. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816–834 (2010).

25. Xu, C. *et al.* Estimating genome-wide significance for whole-genome sequencing studies. Genet. Epidemiol. **38**, 281–290 (2014).

26. Millien, V. Morphological evolution is accelerated among island mammals. PLoS Biol. 4, e321 (2006).

27. Mathieson, I. et al. Eight thousand years of natural selection in Europe. bioRxiv doi:10.1101/016477 (2015).

28. Trecartin, R.F. et al. Beta zero thalassemia in Sardinia is caused by a nonsense mutation. J. Clin. Invest. 68, 1012–1017 (1981).

29. Zuk, O. et al. Searching for missing heritability: designing rare variant association studies. Proc. Natl. Acad. Sci. USA 111, E455–E464 (2014).

30. Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. Nat. Genet. 45, 1274–1283 (2013).

31. Jun, G. et al. An efficient and scalable analysis framework for variant extraction and refinement from population scale DNA sequence data. Genome Res. 25, 918–925 (2015).

32. Li, Y. et al. Low-coverage sequencing: implications for design of complex trait association studies. Genome Res. 21, 940–951 (2011).

33. Sherry, S.T. et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 29, 308–311 (2001).

34. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65 (2012).

35. Gravel, S. et al. Demographic history and rare allele sharing among human populations. Proc. Natl. Acad. Sci. USA 108, 11983–11988 (2011).

36. Maioli, M. et al. Plasma lipoprotein composition, apolipoprotein(a) concentration and isoforms in β-thalassemia. Atherosclerosis 131, 127–133 (1997).

37. Maioli, M. et al. Plasma lipids in β-thalassemia minor. Atherosclerosis 75, 245–248 (1989).

38. Yang, J. et al. Genome-wide complex trait analysis (GCTA): methods, data analyses, and interpretations. Methods Mol. Biol. 1019, 215–236 (2013).

39. Pruim, R.J. et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 26, 2336–2337 (2010).

40. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. 46, 310–315 (2014).

41. Sankaran, V.G. et al. Cyclin D3 coordinates the cell cycle during differentiation to regulate erythrocyte size and number. Genes Dev. 26, 2075–2087 (2012).

42. Kathiresan, S. et al. Common variants at 30 loci contribute to polygenic dyslipidemia. Nat. Genet. 41, 56–65 (2009).

43. Karolchik, D. et al. The UCSC Genome Browser database: 2014 update. Nucleic Acids Res. 42, D764–D770 (2014).

Ringraziamenti

In conclusione di questo percorso è doveroso fare dei ringraziamenti.

Al Professor Francesco Cucca, direttore dell'istituto di ricerca genetica e biomedica del CNR di Cagliari, nonché professore ordinario di genetica medica presso l'Università degli studi di Sassari, per avermi concesso la possibilità di intraprendere questo percorso di studio e per la fiducia dimostratami nel corso degli anni.

Alla mia co-Tutor, la dottoressa Magdalena Zoledziewska, esempio di inconfutabile valore scientifico, per la disponibilità e la costante supervisione avuta nei miei confronti.

A tutti gli autori e co-autori che hanno contribuito a vario grado alla stesura dei tre lavori che contengono i dati presentati in questa tesi di dottorato. In particolar modo al Dottor David Schlessinger, Professor Gonçalo Abecasis, Dottor John Novembre, Dottor Fabrice Danjou, Carlo Sidore, Charleston W.K. Chiang e Serena Sanna.

A tutti i colleghi del progetto SardiNIA/ProgeNIA ed in particolare ai 6,805 Ogliastrini afferenti allo studio ProgeNIA, che hanno reso possibile questa ricerca.