UNIVERSITÀ DEGLI STUDI DI CAGLIARI
FACOLTÀ DI MEDICINA E CHIRURGIA
Dipartimento di Scienze Biomediche e Biotecnologie

XXIII CICLO DOTTORATO DI RICERCA IN

TERAPIA PEDIATRICA E FARMACOLOGIA DELLO SVILUPPO

Studi di associazione e sequenziamento estesi a tutto il genoma: il Progetto ProgeNIA

Relatore:

Prof. Antonio Cao

Dottorando:

Dott. Fabio Busonero

Coordinatore Scientifico: **Prof. Renzo Galanello**

Anno Accademico 2009-2010

(Settore scientifico disciplinare di afferenza MED/03)

…….sempre e comunque a testa alta

Ai miei genitori che, in questo lungo cammino, mi hanno sempre sostenuto, a volte con una parola, a volte con un semplice sguardo

INDICE

INTRODUZIONE

Negli ultimi decenni l'allungamento della vita media e della sua durata massima, l'elevata prevalenza di soggetti anziani nella popolazione generale, specialmente nei paesi occidentali, e l'incremento della spesa sanitaria e sociale ascrivibile all'assistenza ed alla cura degli anziani, hanno stimolato, sia nei ricercatori dell'area biomedica che in quelli dell'area economico-sociale, un particolare interesse nello studio dei processi dell'invecchiamento.

La definizione di invecchiamento di per sé è aperta a varie interpretazioni, sebbene possa essere condivisa la raffigurazione di tale processo come la somma di tutti i cambiamenti che determinano la riduzione delle abilità psicofisiche e la progressiva involuzione morfologica e strutturale dell'organismo; le fondamentali modificazioni età-dipendenti possono essere attribuite al naturale processo di crescita e riguardano alterazioni morfologiche delle cellule, dei tessuti e degli organi; rallentamento delle funzioni biologiche; alterazione delle funzioni dei vari apparati e sistemi; diminuzione dell'energia di riserva.

Quali siano invece le cause alla base dell'invecchiamento e perché alcuni individui invecchino lentamente, riuscendo a vivere più a lungo (i cosiddetti longevi), mentre altri al contrario manifestino segni di deterioramento in età precoce, non è stato ancora del tutto chiarito. È però evidente che molti meccanismi agiscono simultaneamente, operando a diversi livelli di organizzazione funzionale, scaturendo dall'interazione tra fattori genetici, epigenetici, ambientali e puramente stocastici. Pertanto la visione dell'invecchiamento come processo multifattoriale complesso ha sostituito le precedenti teorie "monofattoriali" che vedevano una singola causa come responsabile di tale fenomeno.

Negli ultimi anni numerosi studi sono stati condotti in tutto il mondo con l'obiettivo di identificare le potenziali basi genetiche ed ambientali associate all'invecchiamento. Nonostante gli avanzamenti nelle conoscenze in questo campo, i tradizionali studi genetici si sono rivelati difficili (Risch *et al*., 2000) in quanto complicati dal fatto che le condizioni associate all'invecchiamento comprendono tratti/patologie complesse o multifattoriali, con un numero elevato di geni causativi che, individualmente, forniscono un modesto contributo al fenotipo finale; inoltre, tali condizioni presentano un'insorgenza tardiva, sebbene l'invecchiamento organico sia un processo fisiologico che ha inizio precocemente nel corso della vita, divenendo maggiormente manifesto nelle persone anziane. A questo si aggiunge il fatto che per ottenere dalle analisi il

potere statistico sufficiente per l'individuazione di loci candidati è necessario concepire degli studi che arruolino un numero elevato, nell'ordine di migliaia, di pazienti e/o volontari, cosa non sempre attuabile.

In questo complesso scenario, hanno trovato larga diffusione negli ultimi 5-10 anni, nel campo della ganetica umana, gli studi di associazione estesi a tutto il genoma (GWAS) focalizzati sull'analisi di tratti quantitativi, ovvero tratti biologici che hanno una variazione continua (come l'altezza o il colesterolo), assunti quali fattori di rischio per patologie complesse nella popolazione generale.

Il recente progresso della genomica ha evidenziato come una parte rilevante della variabilità tra individui sia da attribuirsi ai polimorfismi a singolo nucleotide (SNPs). Gli SNPs hanno, quindi, acquistato particolare rilevanza in campo biomedico dal momento in cui sono stati messi in relazione a patologie/tratti complessi.

Grazie alla disponibilità della sequenza completa del genoma umano, ed al completamento della fase pilota del progetto 1000 Genomi (1000 Genomes Project 2010), con la descrizione di più di 15.000.000 di SNPs, 1.000.000 di inserzioni e delezioni, 20.000 varianti strutturali, la maggior parte delle quali non descritte in precendenza, sono state sviluppate nuove piattaforme di genotipizzazione del DNA ad alta processività e la tecnologia degli SNPs, quale strumento di *gene mapping*, é maturata a tal punto che la scansione dell'intero genoma può essere condotta con GeneChip arrays ad alta densità per studiare simultaneamente fino a circa 1 milione di SNPs in un solo saggio per individuo.

Questo disegno di studio si è rivelato di grande successo per l'identificazione di geni responsabili di malattie comuni; infatti, l'importanza dello studio dei tratti quantitativi risiede proprio in queste connessioni tratto-malattia. Raccogliere un gruppo di migliaia di pazienti affetti da una malattia per uno studio dell'intero genoma può essere difficile o, a volte, impossibile; se invece esiste un tratto quantitativo che la rappresenta, misurabile oggettivamente con appositi strumenti, sarà più semplice studiarne la genetica nella popolazione generale ed, in seguito, caratterizzare i geni così identificati nei pazienti, direttamente a livello molecolare piuttosto che genetico.

Per esempio, lo studio di tratti quantitativi quali i livelli sierici dei lipidi, ed i livelli dell'emoglobina fetale, argomenti che verranno sviluppati di seguito in questa tesi, ha permesso l'identificazione di nuovi geni non solo responsabili della variabilità del tratto ma anche implicati in patologie ad essi correlate, quali le patologie coronariche e la talassemia ed anemia falciforme, rispettivamente.

In questo contesto si colloca il progetto ProgeNIA, la cui forza è costituita, oltre che dalla peculiare struttura genetica della popolazione sarda, dal fatto che un'ampia coorte di individui appartenenti alla popolazione generale sia stata caratterizzata negli anni per oltre 100 tratti quantitativi e genotipizzata con GeneChip arrays a bassa ed alta densità.
ProgeNIA, inoltre, fa parte di numerosi "consorzi genetici", in cui i risultati di associazione di ciascun gruppo (seppure ottenuti con piattaforme differenti quali, ad es., Affymetrix 10K-500K-6.0 GeneChip arrays, Illumina HumanHap 300 BeadChip) vengono congiunti tramite meta-analisi per aumentare la dimensione campionaria e permettere l'identificazione di geni con un effetto ridotto sul fenotipo, con il vantaggio di riuscire ad abbattere i costi e i tempi necessari per la raccolta e la genotipizzazione di un campione più ampio da parte di un singolo gruppo.

Le popolazione fondatrici

Da diversi anni l'attenzione dei genetisti si è focalizzata sulle cosiddette popolazioni fondatrici le quali, già dimostratesi utili nello studio di alcune malattie mendeliane, renderebbero più semplice anche l'analisi delle patologie multifattoriali (Peltonen *et al*., 2000).
Tali rare popolazioni hanno vissuto in un secolare o millenario isolamento geografico (Iceland, Finlandia, Sardegna) o, in alcuni casi, culturale (Amish, tribù arabe beduine), con minimo tasso di immigrazione ed elevata percentuale di matrimoni fra consanguinei.
L'alto grado di parentela che le caratterizza fa si che un numero ridotto di geni di predisposizione/varianti alleliche causino fenotipi complessi, cosicché le moderne analisi non-parametriche estese a tutto il genoma (*genome-wide)* possono essere applicate più facilmente per localizzare tali varianti geniche (Heutink *et al*., 2002).
La Sardegna, 1.659.443 abitanti, è un esempio di tali rare popolazioni; vi sono numerose evidenze che la popolazione sarda, pur collocandosi nell'ambito della variabilità europea, manifesti tutta una serie di caratteristiche di unicità: alcune varianti genetiche sono particolarmente frequenti in Sardegna e talvolta rare o assenti in altre popolazioni; sono varianti antiche ed erano già presenti in quegli individui che diverse migliaia di anni fa hanno popolato l'isola (effetto fondatore).
Queste caratteristiche peculiari si spiegano attraverso un lungo isolamento plurimillenario rispetto ad altre popolazioni.

Tali condizioni di omogeneità si ritrovano ancora più facilmente nelle aree rurali dell'isola ed in particolare nell'Ogliastra, una provincia nella parte orientale dell'isola, circondata da montagne e dal mare.
La Sardegna inoltre ha anche un'altra importante caratteristica che la rende vantaggiosa per gli studi genetici: le condizioni ambientali, quali il clima, le abitudini alimentari, etc sono condivise in tutta l'isola.

Il Progetto ProgeNIA

Sulla base di questi presupposti, nel 2001, a seguito di una collaborazione decennale tra l'Istituto di Neurogenetica e Neurofarmacologia (INN) del Consiglio Nazionale delle Ricerche (CNR) di Monserrato, ed il Laboratorio di Genetica del National Institute of Aging (NIA) dell'NIH di Baltimora, è iniziato il progetto ProgeNIA/SardiNIA, dal titolo "Genetica ed Epidemiologia di tratti associati all'invecchiamento nella popolazione Sarda".
ProgeNIA é uno studio longitudinale che si propone di identificare i determinanti genetici di importanti cambiamenti, patologici e non, associati al processo dell'invecchiamento.
La coorte arruolata nel progetto è costituita dagli abitanti di 4 paesi dell'Ogliastra (Lanusei, Arzana, Ilbono ed Elini) ed ha interessato tutti i volontari con età compresa tra i 14 ed i 102 anni, indipendentemente dal loro stato di salute, che avessero una chiara origine Sarda (nati in Sardegna e con i 4 nonni nati in uno dei 4 comuni in studio); dopo i primi tre anni, sono stati reclutati 6.148 volontari (~62% del campione eligibile).
Poiché la maggior parte della società è interrelazionata, gli individui condividono buona parte della loro informazione genetica, rendendo più semplice l'individuazione degli effetti genetici attraverso le generazioni.
Da notare come, tra le persone fenotipizzate, si possano 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. Complessivamente, 5.610 individui sono organizzati in 711 unità familiari connesse (Pilia *et al*., 2006), ciascuna delle quali comprende fino a 5 generazioni.
Nella coorte sono state raccolte informazioni longitudinali per oltre 100 tratti quantitativi o endofenotipi, cosi come circa 200 tratti dicotomici; in particolare, sono state studiate più di 38 misure ematologiche, più di 5 misure antropometriche, più di 25

valutazioni cardiovascolari, e più di 35 aspetti della personalità, così come numerose valutazioni legate alla fragilità fisica quali la velocità di camminata, la forza muscolare, il livello di attività fisica, le disabilità fisiche, la densità mineraria ossea, la funzionalità tiroidea e renale, etc (Pilia *et al*., 2006).

Inoltre, i volontari sono stati seguiti con visite ripetute ogni 3-4 anni: il primo *follow-up* longitudinale è stato completato a febbraio del 2008 ed è attualmente in corso la 3° fase di raccolta dei dati (2° follow-up), la cui valutazione complessiva permetterà di rilevare "*outcomes"*, quali l'incidenza delle malattie (infarto del miocardio, angina, ictus, diabete mellito, dislipidemie, sindrome metabolica, depressione, etc) e le cause di morte nella coorte. La valutazione di quali volontari contrarranno, nel tempo, particolari malattie sarà utile per testare se i geni, trovati in associazione con i fattori di rischio per tali patologie, hanno valore prognostico.

Inizialmente, il progetto ha preso in esame quei fenotipi strettamente correlati alla salute dell'anziano, quali il fenotipo cardiovascolare ed, in particolare, l'elasticità dei vasi arteriosi, una caratteristica delle arterie che si riduce progressivamente negli anni, ed il fenotipo della personalità, essendo noto che alcuni tipi di personalità predispongono ad ammalarsi di patologie quali la depressione, particolarmente frequenti in età avanzata.

A questi fenotipi, legati allo stato di salute degli anziani, si sono aggiunti tutta una serie di altri parametri di più ampio interesse biomedico, tra cui l'emoglobina fetale (HbF) (Tabella 1 per un elenco completo di tutti i tratti esaminati).

Come già accennato, i GWAS hanno permesso lo studio di diverse malattie e dei tratti correlati, con l'identificazione di centinaia di associazioni indipendenti (descritti nel *Catalog of published genome-wide association studies*, NHGRI).

Per questi motivi, questo approccio è stato scelto per l'analisi genetica del campione ProgeNIA, e la genotipizzazione del DNA è stata condotta con l'ausilio della tecnologia Affymetrix, piattaforma che permette l'analisi su larga scala di polimorfismi a singolo nucleotide (SNPs) su un GeneChip array di oligonucleotidi (leggere il paragrafo "Materiali e Metodologie Sperimentali" per i dettagli relativi al disegno sperimentale ed all'analisi statistica).

OBIETTIVI DELLA TESI

La mia collaborazione con l'Istituto di Neurogenetica e Neurofarmacologia del CNR di Monserrato è iniziata a Marzo del 2005. In particolare, nel contesto del progetto ProgeNIA, ho partecipato alla raccolta e processamento del materiale biologico, mi sono occupato dell'analisi genetica del campione e dell'analisi mutazionale dei geni candidati.

Inoltre, sono stato coinvolto nell'analisi ed interpretazione dei risultati, in particolare per lo studio relativo ai livelli dei lipidi, dell'emoglobina fetale, bilirubina, ferro e transferrina.

Infine, recentemente ho iniziato un percorso di ricerca all'estero, presso "The Center for Statistical Genetics, Department of Biostatistics, University of Michigan" in Ann Arbor (USA), nell'ambito del progetto dal titolo "Genetics of lipid levels: draft sequencing of 1.000 genomes" (advisor Gonçalo R. Abecasis), sempre in collaborazione con l'INN-CNR.

Questo progetto mi vede coinvolto nella preparazione delle librerie di DNA genomico e nel loro sequenziamento, nonchè in una parte dell'analisi delle sequenze prodotte, che comprende il mappaggio delle *reads* al genoma di riferimento, la rimozione dei duplicati e la ricalibrazione delle basi.

Tra i lavori a cui ho preso parte negli ultimi anni, ho deciso di discutere in questa tesi i risultati di due progetti che ritengo di particolare interesse e rilevanza; il primo relativo alla regolazione dei livelli ematici di emoglobina fetale (HbF), di importanza notevole considerate le possibili ricadute in Sardegna sulla talassemia e, a livello internazionale, sull'anemia falciforme; il secondo progetto riguarda l'ereditarietà dei tratti che caratterizzano il profilo lipidico, di notevole rilievo a livello mondiale, in considerazione della ormai stabilita correlazione con la malattia cardiaca e per il notevole impatto che questa ha sui costi della sanità pubblica.

In ultimo, anticiperò alcuni risultati preliminari, anche questi relativi ai livelli ematici dei lipidi, derivati dal progetto di sequenziamento esteso all'intero genoma per il quale ho trascorso parte dell'ultimo anno di attività del dottorato di ricerca all'estero.

MATERIALI E METODOLOGIE SPERIMENTALI

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

Consenso informato

Al fine di poter partecipare a questa ricerca, ciascun volontario ha firmato un consenso informato, precedentemente approvato dal Comitato etico della ASL 4 di Lanusei e dal MedStar Research Institute (Hyattsville, MD), NIA (Baltimore, MD).

Prelievo di sangue ed Analisi ematochimiche

Dopo un digiuno di 12 ore, ciascun volontario è stato sottoposto ad un prelievo di ~25 ml di sangue, successivamente aliquotato in 3 provette: una per la separazione dei linfociti, una per l'estrazione del DNA, e la terza sulla quale è stata eseguita una batteria di esami ematologici.

Relativamente alla sperimentazione oggetto di questa tesi, i livelli ematici di trigliceridi, colesterolo HDL e colesterolo totale sono stati determinati analizzando una aliquota di siero con l'Express Plus, analizzatore che sfrutta il principio della lettura colorimetrica o torbidometrica; i livelli di colesterolo LDL sono stati calcolati usando la formula di Friedewald. Altre valutazioni ematologiche hanno riguardato globuli rossi, leucociti, piastrine, emoglobina, MCV e gli indici derivati MCH e MCHC, analizzati con il Coulter LH 700 (Beckman); HbA2, HbF e HbA1c sono invece stati determinati con il Variant II (Bio-Rad), il quale sfrutta il principio della cromatografia liquida (HPLC).

Separazione ed Immortalizzazione dei linfociti

La trasformazione o blastizzazione dei linfociti con il virus di Epstein-Bar (EBV) ha permesso di ottenere linee cellulari stabilizzate che, adeguatamente crio-conservate, sono state utilizzate come fonte di DNA, RNA e proteine e per effettuare studi funzionali.

Il sangue, prelevato in litio-eparina, è stato diluito nel buffer HBSS (Hank's Balanced Salt Solution) contenente penicillina (10.000 U/ml)-streptomicina (10.000 µg/ml), e stratificato delicatamente sul gradiente Histopaque-1077 (Sigma-Aldrich).

Dopo centrifugazione, l'anello di linfociti è stato recuperato e lavato con HBSS; infine, il pellet di cellule è stato riportato in soluzione mediante agitazione e risospeso in siero bovino fetale, inattivato al calore, integrato con il 10% di DMSO.

Dopo la conta dei linfociti nella camera di Bürker (Knittelglaser 0.100 mm, Tiefe Depth 0.0025 mm^2), la sospensione cellulare è stata pre-incubata a -20°C per circa 3 ore e successivamente trasferita in azoto liquido.

Successivamente, è stato possibile espandere i linfociti separati e crioconservati attraverso la messa in coltura degli stessi nel terreno RPMI 1640 contenente il 14% di siero fetale bovino (FBS) inattivato al calore, L-Glutamina 2 mM, Sodio Piruvato 1 mM, Penicillina (10000 U/ml), Streptomicina (10000 µg/ml) (Biowhittaker Cambrex) ed una aliquota del virus di Epstein-Bar.

Dopo circa 30-45 giorni di crescita in coltura, i linfoblasti (ovvero i linfociti immortalizzati) sono stati contati, la sospensione cellulare è stata suddivisa in 3 aliquote, ciascuna contenente ~15-30 milioni di linfoblasti, le quali sono state crioconservate in azoto liquido.

La fonte di EBV necessaria per le trasformazioni dei linfociti, è garantita dalla crescita in coltura di una linea stabilizzata (B95-8) di linfociti della scimmia Marmoset infettati con il virus. Questi sono stati coltivati in un terreno la cui composizione è simile a quello usato per i linfociti umani, eccetto che per la percentuale di FBS che, in questo caso, è del 10%. Quando si raggiunge un numero sufficiente di cellule (~40-80 milioni), queste vengono raccolte sia come soluzione stock (pellet) di linfoblasti sia come soluzione pronta all'uso per la trasformazione (surnatante contenente il virus), quest'ultima filtrata (0.22 µm), aliquotata in criovials e conservata in azoto liquido o a -80°C.

Estrazione del DNA

Il DNA genomico è stato estratto da sangue intero prelevato in EDTA mediante estrazione salina (*salting-out*). La qualità del DNA è stata verificata mediante corsa su gel di agarosio allo 0.8%, mentre la quantificazione del DNA e la determinazione del rapporto $OD_{260/280}$ è stata effettuata mediante lettura spettrofotometrica. Sono state preparate delle diluizioni alla concentrazione di 50 ng/µl, per un totale di 5 µg, usate per la genotipizzazione, il fine mapping e la replica dei dati di associazione.

Disegno sperimentale del progetto ProgeNIA ed Analisi statistica

Il disegno sperimentale del progetto ProgeNIA è stato concepito per sfruttare le relazioni parentali all'interno del campione (Scuteri *et al*., 2007), al fine di inferire successivamente *in silico* i genotipi in tutta la coorte.

Può essere così riassunto:

1. 1.412 volontari non imparentati tra loro, selezionati per la loro posizione centrale nelle strutture familiari (entrambi i genitori e un figlio nelle famiglie più numerose, solo i genitori in quelle meno numerose), sono stati genotipizzati con il pannello dei GeneChip arrays 500K (contenenti 500.000 SNPs sparsi uniformememte su tutto il genoma);
2. 5.540 volontari, appartenenti a piccoli nuclei familiari, ai nuclei familiari più numerosi e coppie di fratelli (essi sono per la maggior parte figli e fratelli dei 1.412 volontari tipizzati con il pannello dei 500K), sono stati genotipizzati con i GeneChip arrays 10K (contenenti 10.000 SNPs);
3. 436 volontari sono stati genotipizzati con entrambe le piattaforme al fine di valutare con maggior accuratezza lo stato di identità per discendenza (IBD *state*) dei markers presenti solo in una delle due.

Grazie all'alto grado di parentela della coorte, è stato possibile utilizzare una versione modificata dell'algoritmo di Lander-Green (Chen *et al*., 2007), per inferire probabilisticamente (Li Y. *et al.*, 2009) gli aplotipi condivisi (tipicamente >10 Mb) e completare, a partire dalle informazioni genetiche più dettagliate sui 1.412 individui tipizzati con il pannello di 500K, i genotipi mancanti in 2.893 individui selezionati tra i 5.540 tipizzati con il pannello di 10K (Figura 1).

L'approccio funziona in maniera ottimale quando tutti gli individui sono direttamente genotipizzati per almeno alcuni SNPs (tutte le famiglie più numerose nel campione, a tale scopo, sono state genotipizzate con l'array 10K) e quando tutti i fondatori e almeno un figlio per fondatore sono genotipizzati ad alta densità (allo scopo, i volontari sono stati selezionati accuratamente con questo criterio).

In questo modo, è stato possibile aumentare l'effettiva dimensione del campione fino a 4.305 individui selezionati per l'analisi di associazione.

Sono stati presi in esame un totale di 362.129 SNPs (M.A.F.>5%), tra 10K e 500K GeneChip arrays, selezionati in base al superamento dei filtri di qualità (Scuteri *et al*., 2007) quali la completezza dei dati (>90%), la trasmissione mendeliana (<3% di incongruenze) ed il rispetto dell'equilibrio di Hardy-Weinberg ($p > 10^{-6}$).

Attraverso l'inferenza statistica, utilizzando il pannello di HapMap, è stato possibile propagare a tutto il campione le informazioni su oltre 2.500.000 di SNPs.

È stata quindi eseguita un'analisi di associazione *genome-wide* su base familiare al fine di valutare gli effetti additivi genetici e, per la maggior parte dei tratti esaminati, è stato identificato almeno 1 locus con significatività *genome-wide* che contribuisce alla variabilità genetica degli stessi.

Successivamente, al fine di ottimizzare l'analisi dei loci e dei geni candidati, identificati mediante l'analisi di associazione, è stata eseguita sia una replica interna sia un *fine mapping* mediante *targeted genotyping* su 2.496 individui utilizzando un GeneChip array custom 12K (piattaforma ParAllele, Affymetrix).

In totale sono stati selezionati 11.617 SNPs, scelti in base ai dati pubblici disponibili, inclusi i dati del progetto HapMap e di altri studi su larga scala basati su SNPs; in particolare, sono stati selezionati 10 SNPs tag presenti nella regione codificante di ciascuno dei 106 geni, che corrispondono ai candidati per tutti i tratti, mentre i restanti marcatori sono stati scelti considerando i primi 90 SNPs per ogni tratto, la cui lista é stata poi ridotta per escludere SNPs in forte *linkage disequilibrium* o in comune tra piu tratti (per esempio, i primi 15 SNPs sono gli stessi per RBC, MCV, MCH, etc.).

Il campione genotipizzato con questo chip personalizzato è stato cosi suddiviso: 1.862 volontari sono stati scelti come campione di replica (chiamato SardiNIA *stage* 2), in quanto non imparentati con i volontari inclusi nel primo GWA scan (500K); su 638 volontari è stato, invece, eseguito il fine mapping, in quanto strettamente imparentati con i campioni gia genotipizzati (10K e 500K), al fine di ricostruire gli aplotipi tra 10K, 500K e TG e ripetere l'analisi di associazione (in particolare, essi comprendono 100 coppie, marito e moglie, e 434 figli delle famiglie più numerose).

Quando nel 2005 è iniziata l'analisi genetica del campione, erano disponibili sul mercato piattaforme tecnologiche che permettevano di studiare contemporaneamente solo un numero limitato di SNPs (tra queste, i chips Affymetrix 10K, 50K, 100K ed i *beadchip* Illumina 370K e 510K).

Non appena nel 2008 si sono resi disponibili i GeneChips arrays 6.0 (Affymetrix), contenenti 906.000 SNPs e 946.000 sonde per variazioni del numero di copie (CNVs), si è deciso di genotipizzare con questo pannello ulteriori 1.000 volontari ProgeNIA, per ottenere una mappa genetica più dettagliata.

In linea con la precedente strategia, la selezione di questi campioni ha tenuto conto dei vantaggi derivanti dalle strette relazioni di parentela nel campione ProgeNIA, con

conseguente risparmio economico. Anche in questo caso, sono stati inclusi nel GWAS non solo gli individui genotipizzati con gli arrays ad alta densità (500K e 6.0), ma anche i loro parenti stretti, tipizzati con la mappa più rada, i cui genotipi sono stati inferiti utilizzando le informazioni di *IBD* (identity-by-descent)-*sharing.*

È stato quindi ripetuto il GWAS, dopo aver inferito i genotipi su tutto il campione utilizzando il pannello di HapMap e quello del progetto 1000 Genomi, recentemente completato, per un totale di ~5.893.000 SNPs, tra genotipizzati direttamente ed inferiti con qualità di imputazione r^2>0.5.

Genotipizzazione con le piattaforme 10K, 500K e 6.0

La genotipizzazione con le piattaforme 10K, 500K e 6.0 é stata eseguita seguendo il protocollo standard (Matsuzaki *et al*., 2004). In particolare, i protocolli dei GeneChip arrays si basano sulla discriminazione allelica a seguito di ibridazione del DNA al chip contenente oligonucleotidi di 25 basi locus- ed allele-specifici.

Il protocollo prevede una riduzione della complessità del DNA genomico attraverso digestione con endonucleasi di restrizione, appropriate per il numero di SNPs da interrogare, quali XbaI per il *chip* 10K, StyI ed NspI per i *chips* 500K e 6.0; i frammenti ottenuti, compresi nel range di 400-800 bp, sono stati selezionati per la ligazione degli adaptors. Dopo l'amplificazione ed un'ulteriore frammentazione con DNasi I, il DNA è stato marcato, ibridato sul chip, e si è proceduto alla scansione per la discriminazione degli alleli di ciascuno SNPs.

Genotipizzazione con la piattaforma ParAllele (chip personalizzato)

In base alla tecnologia *molecular inversion probe* (Hardenbol *et al*., 2003), le sonde sono state ibridate al DNA genomico, in maniera tale da fiancheggiare lo SNP interrogato e, mediante reazione enzimatica, sono state circolarizzate in maniera allele-specifica; una reazione esonucleasica ha, quindi, eliminato le sonde non legate o che hanno cross-reagito. Le sonde circolarizzate (*pad-locked*) sono state invertite, ovvero linearizzate, e rilasciate dal DNA genomico.

Dopo amplificazione mediante PCR, le sonde sono state catturate nel chip e, quindi, marcate mediante ibridazione con un fluoroforo allele-specifico. I campioni sono stati quindi ibridati sul GeneChip custom Tag Arrays e scansionati 4 volte, una volta per ciascun allele interrogato.

Genotipizzazione di *single* SNPs con il TaqMan

In questo saggio i genotipi degli SNPs sono stati generati mediante discriminazione allelica con 5′-nucleasi (Livak *et al*., 1995), in un saggio implementato nella piattaforma 7900HT Fast Real-Time PCR System (Applied Biosystems).

In questo caso, la regione fiancheggiante lo SNP é stata amplificata in presenza di due sonde fluorescenti allele-specifiche, contenenti un differente *reporter dye* all'estremità 5' (FAM or VIC) ed un *quencher* non fluorescente all'estremità 3' (le sonde non emettono fluorescenza in soluzione a causa della presenza del *quencher*).

Dopo la reazione di PCR, condotta con una *mastermix* standard e con il file suggerito dal produttore, i campioni sono stati letti con la piattaforma 7900HT Fast Real-Time PCR System, mentre i segnali di fluorescenza sono stati analizzati con il software SDS (Perkin-Elmer) per determinare i genotipi degli SNPs.

Sequenziamento tradizionale (Sanger) del gene *BCL11A*

L'intera regione codificante del gene *BCL11A,* la giunzione introne/esone, le regioni non tradotte in 5' ed in 3', e l'ipotetica regione del promoter ad 1kb a monte del sito di inizio della trascrizione, sono state sequenziate dopo che 26 distinti frammenti sono stati amplificati mediante PCR.

I parametri per il disegno dei primers includevano: lunghezza di 18-24 bp, temperatura di *melting* (Tm) di 62-64°C e contenuto medio in GC del 45-60%. Per il disegno del *set* di primers specifici, le regioni contenenti sequenze ripetute sono state mascherate con il software "RepeatMasker".

La reazione di sequenza è stata condotta con il kit *Big Dye Terminator* (Applied Biosystems), seguita da precipitazione con isopropanolo, ed il sequenziamento eseguito con la piattaforma ABI Prism 3130 *xl* Genetic Analyzer (Applied Biosystems).

Re-sequencing* del gene *BCL11A

Per il sequenziamento dell'intera regione genomica di 105 Kb, contenente la sequenza del gene *BCL11A* (esoni, introni, la regione a valle ed a monte del gene), sono stati prodotti 23 distinti ampliconi sovrapposti della lunghezza di 3.8Kb-5.8Kb mediante *long-range* PCR (Sequalprep long-range PCR, Invitrogen).

I 23 ampliconi appartenenti a ciascun campione sono stati frammentati con il Bioruptor al fine di ottenere frammenti di lunghezza inferiore a 800 bp.

Nel processo di preparazione delle librerie genomiche il DNA, a cui sono stati ligati gli adattatori, viene marcato con una sequenza etichetta o index di 6 bp durante la successiva PCR; in questo modo, fino a 12 campioni possono essere caricati come *pool* nella stessa lane della *flow-cell*, per un totale di 96 campioni sequenziati in una singola corsa *single-read* da 36 bp eseguita con il Genome Analyzer *IIx* (Illumina).

Preparazione delle Librerie di DNA e sequenziamento dell'intero genoma mediante la tecnologia Illumina

Il sequenziamento esteso a tutto il genoma é stato eseguito con le piattaforme Genome Analyzer *IIx* ed Hi-Seq (Illumina). Le librerie di DNA genomico sono state generate in accordo con le indicazioni della Illumina, con alcune modifiche nel protocollo (Quail *et al*., 2008).

Brevemente, il DNA genomico é stato frammentato in maniera random, mediante sonicazione (Covaris-S), in frammenti di dimensione inferiore alle 800 bp e successivamente le estremità 5' e 3' dei frammenti sono state riparate e fosforilate.

I frammenti di DNA riparati sono stati adenilati in 3' con una DNA polimerasi *Klenow exo-* (New England BioLabs) e poi sono stati aggiunti degli adattatori (IDT) con l'impiego di DNA ligase. I prodotti di ligazione, di dimensione compresa tra le 300 e le 400 bp, sono stati selezionati su gel di agarosio al 2%, purificati (Gel Extraction Kit, Qiagen) e, successivamente, pre-amplificati mediante PCR, utilizzando dei primers (IDT) compatibili con gli adattatori.

Dopo purificazione degli ampliconi con delle biglie magnetiche (Agencourt Ampure XP, Beckman), la concentrazione e la distribuzione dei frammenti delle librerie sono state determinate mediante corsa su Chip DNA 1000 nel Bioanalyzer 2100 (Agilent Technologies); inoltre, le librerie sono state validate anche mediante PCR quantitativa (qPCR) con il sistema Kapa SYBR fast qPCR (Kapa Biosystems), che permette di confermare la concentrazione delle stesse, nonchè di verificare quale percentuale di frammenti contengono gli adattatori ad entrambe le estremità.

Dopo opportuna titolazione, le librerie sono state ibridate e amplificate sulla superficie di una *flow-cell*, mediante *bridge amplification*, con formazione dei *clusters*, quindi sequenziate con il GA*IIx*, in corse *paired-end* da 240 basi (fino a 320 bp con la chimica V5), o con l'Hi-Seq in corse da 208 basi, ottenendo un *coverage* medio di 2-4X.

I campioni sono stati inoltre tipizzati mediante la piattaforma Sequenom, con un pannello di 40 SNPs, presenti anche nelle piattaforme Illumina 317K e Affymetrix

500K, al fine di poterli unicamente identificare mediante confronto con dati genotipici pre-esistenti per ciascuno di essi.

Analisi delle sequenze prodotte con la piattaforma Illumina

Le piattaforme di sequenziamento di nuova generazione, come i sequenziatori GA*IIx* e Hi-Seq (Illumina), permettono di leggere due *strand* e producono sequenze corte dell'ordine di 100-200 basi, chiamate *reads*.

La qualità delle sequenze é stata valutata con una serie di metriche: percentuale dei *clusters* che superano i filtri di qualità Illumina, numero di *reads* per *lane*, numero di basi che vengono mappate in maniera univoca sul genoma, la dimensione dei frammenti, etc.

Le *reads* sono state mappate sul genoma di riferimento mediante un algoritmo di allineamento per *short reads* (Burrows-Wheeler *Alignment tools*) e, mediante superdeduper, sono state rimosse le eventuali *reads* duplicate.

Ad ogni base è stato assegnato un parametro di qualità (*Phred score* originale), in base ai segnali di intensità luminosa assegnati dal sequenziatore alle stesse; il *Phred score* originale, dopo il mappaggio delle *reads*, è stato ri-calibrato (*Phred score* empirico) in base al confronto delle sequenze delle *reads* con il genoma di riferimento ed al tasso di errore rilevato dopo aver raggruppato le basi secondo il *Phred score* originale.

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

Le *reads* con un adeguato *quality score* ricalibrato sono state analizzate per rilevare varianti genetiche, quali SNPs e, sebbene l'analisi iniziale del nostro progetto *low-coverage* è stata focalizzata alla ricerca di SNPs, le sequenze generate saranno utilizzate nell'immediato futuro per identificare anche altri polimorfismi, quali inserzioni e delezioni o varianti strutturali.

RISULTATI, DISCUSSIONE E CONCLUSIONI

Ereditabilità dei tratti

Il primo lavoro del progetto ProgeNIA, pubblicato nel 2006 nella rivista PLoS Genetics (Pilia *et al*., 2006), rappresenta una pietra miliare nella genetica dei tratti complessi in quanto ha valutato per la prima volta l'impatto dei geni e dell'ambiente su più di 100 tratti quantitativi, descrivendo una significativa componente genetica per ciscuno di essi; ha inoltre dimostrato che uno stesso fattore genetico potrebbe influenzare diversi tratti (agendo per esempio su molteplici caratteristiche delle funzioni cardiovascolare e della personalità).

In particolare, gli effetti genetici spiegano il 40% della varianza dei 38 test ematologici, il 51% della varianza delle 5 misure antropometriche, il 25% per i parametri cardiovascolari e il 19% per i 35 tratti della personalità.

La varianza genetica risulterebbe generalmente più marcata nelle donne e negli individui più giovani.

La successiva analisi genetica, a dimostrazione della validità dell'approccio scelto, ha permesso di confermare il coinvolgimento di loci già noti nella regolazione della variabilità di alcuni tratti studiati, tra cui alcuni parametri ematologici (RBC, MCV, Hb, HbA1c, HbA2, HbF, MCH, MCHC), la bilirubina, l'attività della G6PD e di identificare nuovi loci per la maggior parte di essi.

Alcuni dei lavori in cui sono stato coinvolto hanno appunto descritto, per la prima volta, l'associazione tra nuovi geni e i tratti correlati con il profilo lipidico e le dislipidemie (Willer *et al*., 2008), l'obesità (BMI, circonferenza ai fianchi, peso, Scuteri *et al*., 2007, Speliotes *et al*., 2010), l'uricemia (Li *et al*., 2007), la regolazione dei livelli di emoglobina fetale (Uda *et al*., 2008), i livelli di TSH (Arnaud-Lopez *et al*., 2008), l'iperbilirubinemia (Sanna *et al*., 2009), ed alcune caratteristiche della personalità (Terracciano *et al*., 2008, Terracciano *et al*., 2009).

LO STUDIO DELL'EMOGLOBINA FETALE ED IL GENE *BCL11A*

Presupposti dello studio

La β-talassemia e l'anemia falciforme (*sickle cell disease,* SCD) sono tra le malattie genetiche più diffuse al mondo. In particolare, in Sardegna la frequenza dei portatori di β-talassemia è del 12% di cui la grande maggioranza presenta la stessa mutazione puntiforme ($\beta^0$39).

I portatori sono soggetti più resistenti alla malaria rispetto ai soggetti sani, e questa è stata presumibilmente la potente forza selettiva che ha determinato l'elevata frequenza delle mutazioni causa di tali patologie in zone a pregressa endemia malarica, insieme ovviamente all'effetto fondatore presente in questa popolazione.

Malgrado mostrino una marcata omogeneità genetica, entrambe le condizioni manifestano una notevole eterogeneità fenotipica (Weatherall *et al*., 2001; Cao *et al* 1994; Cao *et al*., 2004), determinata sia da fattori ambientali che genetici.

Questa eterogeneità può essere dovuta all'azione di geni modificatori che, parzialmente, attenuano la severità della malattia; tra questi, la co-ereditarietà dell'α-talassemia, che riduce lo sbilanciamento delle catene globiniche, e la persistenza di emoglobina fetale (HbF, α2γ2) nell'adulto (Thein *et al.*, 1998; Cao *et al*., 1994; Galanello *et al*., 2009) sono tra i fattori meglio conosciuti che conferiscono ai pazienti un quadro clinico più favorevole, rendendoli talvolta meno dipendenti dalla terapia trasfusionale. Per esempio, in alcuni casi è stata osservata una elevata espressione dei geni γ-globinici che determina un aumento dei livelli di emoglobina fetale che supera l'1%, compensando la mancata produzione di HbA.

Il meccanismo molecolare responsabile non è stato ancora chiarito completamente, ma numerose evidenze genetiche, tra cui la presenza di delezioni nei geni δ e β-globinici o mutazioni puntiformi del promotore dei geni γ, come il polimorfimo a -158 del gene Gamma, sostengono la relazione tra diverse mutazioni ed il fenotipo. Molti sforzi sono stati quindi rivolti all'identificazione di altri fattori genetici che determinano la persistenza di HbF con lo scopo di migliorare il quadro clinico della beta talassemia e dell'anemia falciforme.

Obiettivi

Sulla base di queste premesse, tenuto conto dell'alta frequenza di portatori di β-talassemia nell'isola, l'INN-CNR nel contesto del progetto ProgeNIA, ha cercato di approfondire le attuali conoscenze relative alle β-talassemie ed in particolare di

identificare e caratterizzare nuovi fattori genetici associati con la persistenza dell'espressione di emoglobina fetale nell'adulto. Tali fattori, se in grado di indurre la produzione di emoglobina fetale anche in pazienti affetti da β-talassemia o SCD, possono determinare una conseguente attenuazione del quadro clinico di tali patologie e rappresentare dei potenziali bersagli terapeutici per la cura della malattia.

Risultati – GWAS e replica dei dati

A tal fine, tutti i partecipanti allo studio sono stati caratterizzati, dal punto di vista fenotipico, per otto tratti che definiscono la composizione dell'emoglobina e degli indici eritrocitari: RBC, MCV, Hb, HbA1c, HbA2, HbF e gli indici derivati MCH ed MCHC.

Come atteso, l'analisi genetica ha evidenziato associazioni significative (figura 2a, b) nei loci del cromosoma 11 (cluster β-globinico) e del cromosoma 16 (cluster α–globinico), oltre al locus MYB/HBS1L sul cromosoma 6 già coinvolto nella regolazione dei livelli di HbF.

Il dato più interessante è risultato dall'analisi dei livelli di HbF: i risultati dell'analisi *genome-wide* mostrano un'associazione statisticamente significativa sul cromosoma 2 con lo SNP rs11886868 (figura 3), nell'introne 2 del gene *BCL11A* (P=6.74x10^{-35}). Questo dato, insieme a quelli relativi agli altri 7 tratti ematologici, è stato replicato (P=8.5x10^{-10}) in un gruppo indipendente di 521 sardi (chiamato SardiNIA *stage* 2), non imparentati con gli individui compresi nel GWAS, utilizzando il chip *custom* Affymetrix (tabella 2).

Il *BCL11A*, espresso nei precursori eritroidi del sangue e precedentemente noto per essere implicato nella insorgenza di linfomi e leucemie, codifica per un fattore trascrizionale *Kruppel-like zinc finger* con isoforme multiple (XS, S, L ed XL) che condividono un comune dominio N-terminale, ma differiscono nel numero dei *zinc-fingers* C-terminali. Il *BCL11A* lega direttamente i motivi GC-rich delle regioni regolatrici dei suoi geni bersaglio agendo come repressore trascrizionale che interagisce con numerose proteine, tra le quali BCL6, COUP-TFII, SIRT1 (Liu *et al*., 2006).

Osservando la distribuzione dei genotipi risulta evidente la differenza (tabella 3, figura 4) tra gli individui con valori fisiologici di HbF (≤0.8%) e quelli con livelli di HbF elevati (>0.8% dell'emoglobina totale, la soglia per HPFH (persistenza ereditaria eterocellulare di HbF). Le frequenze genotipiche risultavano infatti 67% (T/T), 30% (C/T) e 3% (C/C) nei 1.268 individui della coorte ProgeNIA con livelli normali di HbF (frequenza allele C=0.18), mentre nei 134 individui con HPFH, queste frequenze erano

40%, 47% e 13% (frequenza dell'allele C=0.37), corrispondenti ad un arricchimento di 2 volte dell'allele C, e di 5 volte del genotipo C/C ($P<2.7x10^{-13}$ per le differenze nelle frequenze alleliche).

Inoltre, l'allele C dello SNP rs11886868 è aumentato in frequenza in 66 soggetti sardi con HPFH, indipendentemente valutati mediante screening per β-talassemia nell'intera isola, se confrontati con il campione di 1.412 volontari direttamente tipizzati per l'rs11886868, dato che ha permesso di correlare l'allele C dello SNP rs11886868 agli alti livelli di emoglobina fetale (figura 4).

Risultati – Correlazione del polimorfismo del *BCL11A* con un migliorato fenotipo della β-talassemia omozigote in Sardegna ed effetti sull'anemia falciforme.

Successivamente, è stata valutata l'ipotesi che varianti nel *BCL11A*, influenzando i livelli di emoglobina fetale, potessero modulare il fenotipo clinico delle β-talassemie. Per testare questa ipotesi, lo SNP rs11886868 è stato genotipizzato in 52 pazienti affetti da talassemia intermedia e 74 pazienti con talassemia major (tutti omozigoti per la mutazione $\beta^0$39 e nessuno di essi portatori della mutazione Gγ-158(C>T) (associata con l'attivazione dell'HbF in condizioni di stress eritropoietico), anch'essi reclutati attraverso uno screening di popolazione per β-talassemia nell'intera isola (tabella 3).

L'allele C dello SNP rs11886868, associato ad elevati livelli di HbF, risultava significativamente più frequente nei pazienti con talassemia intermedia ($P_{genotipo}<6,5x10^{-6}$, $P_{allele}<2,9x10^{-6}$), ad indicare che la variante del *BCL11A* portatrice dell'allele C dello SNP rs11886868, incrementando i livelli di HbF, contribuisce allo sviluppo di un fenotipo più lieve. Infatti, mentre i pazienti con talassemia major sono trasfusione-dipendenti, quelli con talassemia intermedia non ricevono, se non sporadicamente, trasfusioni e sono caratterizzati da alti livelli di emoglobina, quasi interamente composta da HbF.

Aumentati livelli di HbF sono associati con una ridotta mortalità oltre che nella β-talassemia, anche nell'anemia falciforme. Per questo motivo, è stata genotipizzata una coorte di 1.242 pazienti affetti da anemia falciforme provenienti dal "*Cooperative Study of Sickle Cell Disease*" (CSSCD) e, anche in questi individui, l'allele C dello SNP rs11886868 è risultato fortemente associato all'aumento dei livelli di HbF ($P<10^{-20}$), giustificando da solo l'8,6% della varianza del tratto.

Questi risultati sono di estrema importanza, in quanto suggeriscono che le varianti del gene *BCL11A*, oltre ad influenzare i livelli di emoglobina fetale negli individui sani, hanno anche un importante effetto nel contesto delle emoglobinopatie.

Risultati - Analisi mutazionale

Allo scopo di identificare mutazioni causative nel gene *BCL11A*, é stata eseguita l'analisi mutazionale del gene su un campione di 20 individui con valori di HbF≥0,8%, costituiti da 10 controlli sani e 10 portatori di β-talassemia suddivisi, a loro volta, in base al genotipo dello SNP rs11886868 (5 omozigoti C/C e 5 omozigoti T/T).

Sono state sequenziate interamente le regioni codificanti (esoni), le giunzioni introne/esone, il 5′- e 3′-UTR, e la presumibile regione del promoter ad 1,3 Kb a monte del sito di inizio della trascrizione; dalla lettura delle sequenze risultava che 4 campioni presentano nel 5′-UTR un numero variabile di repeat (c.-43GCC[9]+[11]), 10 campioni hanno una sostituzione sinonima nell'esone 4 (c.2088T>C, p.Ser696Ser), tutti i campioni presentano 2 delezioni di un nucleotide (c.*222delT, c.*2719delA) nel 3′-UTR ed 1 campione ha una sostituzione nucleotidica nel 3′-UTR (c.*557C>T).

La nostra attenzione è stata inizialmente focalizzata sulla variabilità del repeat, essendo noto che nel 5′-UTR sono contenute importanti sequenze regolatrici che controllano l'espressione genica, sia a livello trascrizionale che post-trascrizionale; inoltre tre dei quattro campioni suddetti mostravano i più alti livelli di HbF tra i 20 individui sequenziati. Lo screening di questa variazione del DNA, eseguito tramite analisi di polimorfismi di conformazione a singolo filamento (SSCP) su un campione di 240 volontari, ha evidenziato che la variabilità del repeat è presente indistintamente nell'intero campione, sia negli individui con alti livelli di HbF che negli individui con livelli normali, a dimostrazione del fatto che si tratta di un polimorfismo e non di una mutazione causativa.

Conclusioni

Il nostro studio ha dimostrato che la variante C dello SNP rs11886868 nel locus del *BCL11A* é più frequente sia in soggetti con HPFH eterocellulare che in soggetti talassemici con un fenotipo lieve, rispetto a quelli con una forma severa, probabilmente grazie alla parziale compensazione della sbilanciata produzione di emoglobina.

Il gene *BCL11A* sarebbe quindi in grado di modificare il fenotipo della β-talassemia attraverso un aumento dei livelli di HbF.

L'aspetto notevole di questa scoperta è costituito dal fatto che le varianti del gene *BCL11A* modulando i livelli di HbF, probabilmente attraverso il legame a regioni regolatorie nel cluster beta-globinico, e quindi modulando lo switch che determina la produzione relativa di catene globiniche fetali e adulte, possono contribuire allo sviluppo di nuovi approcci terapeutici per la β-talassemia.

Inoltre, i dati della coorte di pazienti con anemia falciforme, indicano che il polimorfismo del *BCL11A* é presente con effetti comparabili in altre popolazioni, e che la tipizzazione del locus può avere una utilità pratica non solo nella popolazione sarda. La determinazione di quali alleli siano presenti a tale locus, in giovani pazienti con β-talassemia e anemia falciforme, può servire come indicatore prognostico della severità di tali malattie.

Prospettive future

Dal momento che il sequenziamento "tradizionale" non ha portato all'individuazione di varianti codificanti del gene, suggerendo che varianti regolatorie possano essere responsabili dell'associazione individuata, la ricerca di mutazioni causative è proseguita con il progetto di *re-sequencing* del *BCL11A*, che ha previsto il sequenziamento dell'intera regione genomica (110 kb) contenente il gene *BCL11A* in 96 soggetti, di cui 63 volontari del progetto ProgeNIA, selezionati per avere in parte livelli di HbF≥0.8 e in parte livelli di HbF≤0.8, e 33 pazienti con talassemia intermedia, reclutati attraverso uno screening di popolazione per β-talassemia nell'intera isola.

Questo approccio si è basato sull'utilizzo delle nuove piattaforme di sequenziamento ad alta processività e, in particolare, gli esperimenti sono stati condotti con la tecnologia Illumina, Genome Analyzer *IIx*. Le analisi dei dati sono al momento in corso.

Infine attualmente si stanno conducendo studi di espressione e analisi funzionali per fare luce sui meccanismi attraverso i quali il *BCL11A* regola il livello dell'HbF.

I 4 SNPs maggiormente associati ai livelli di HbF, rs766432, rs4671393, rs11886868 e rs1427407, mappano nel secondo introne del gene *BCL11A.* Una ipotesi è che questi SNPs possano avere una funzione regolatoria sul gene.

Per testare questa ipotesi si stanno perseguendo diverse vie:

1) Una delle ipotesi è che possano fungere da *enhancer* della trascrizione del gene *BCL11A*. Pertanto, sequenze contenenti ciascuno SNPs preso singolarmente ed una sequenza di DNA di circa 2.4Kb contenente tutti e quattro gli SNPs sono state isolate mediante PCR da DNA genomico e clonate all'interno del vettore

d'espressione pGL3, in 5′ al promotore del *BCL11A*, entrambi a monte del gene reporter per la luciferase. Il costrutto contenente il frammento clonato verrà utilizzato in saggi di transattivazione in linee cellulari con fenotipo eritroide (K562 e Mel C88) e non (HeLa), e gli estratti cellulari verranno saggiati per l'attività luciferasica con il *Dual Reporter Assay System*.

2) La sequenza nella quale mappano gli SNPs potrebbe anche contenere un sito target per microRNA endogeni non ancora descritti; pertanto questa verrà inserita nel vettore pmirGLO e transfettata in linee cellulari eritroidi e non; dopo 24 ore dalla transfezione le cellule verranno analizzate per l'attività luciferasica che, nel caso in cui un microRNA si leghi alla sequenza bersaglio, risulterà ridotta.
3) Un altra ipotesi è che la sequenza contenente gli SNPs possa contenere dei siti bersaglio per fattori di trascrizione capaci di legare il DNA. Tali siti potrebbero essere quindi eliminati o creati dalla presenza di uno dei tre alleli.

   Questa ipotesi verrà verificata con esperimenti di ritardo della mobilità elettroforetica (*band shift* e *super shift* con anticorpi specifici); in questi esperimenti, estratti proteici totali e nucleari derivati da linee cellulari non eritroidi (HeLa) ed eritroidi (K562 e Mel C88), indotte a differenziarsi, verrano saggiati con sequenze oligonucleotidiche contenente gli SNPs marcati.

Se da questi esperimenti dovesse risultare che dei fattori legano le sequenze contenenti gli SNPs, verrà eseguita un'analisi di transattivazione sul promotore del *BCL11A*.

Con prove di immunoprecipitazione della cromatina cercheremo di provare che il legame di questi fattori alla sequenza in esame avviene anche in vivo sul gene endogeno e non solo in vitro, quale si evincerebbe dall'analisi di ritardo della mobilità elettroforetica.

In esperimenti complementari tramite infezione con vettori lentivirali di ultima generazione introdurremo invece nelle cellule progenitrici eritroidi shRNA specifici per i fattori di cui vogliamo determinare il ruolo di regolazione sul gene *BCL11A*. L'interferenza con gli RNA messaggeri endogeni codificanti per questi fattori determinerà la degradazione dell'RNA e, in tal modo, potremo analizzare l'effetto di ciascun fattore sull'espressione del gene *BCL11A*.

LO STUDIO DEL PROFILO LIPIDICO E LE META-ANALISI

Presupposti dello studio

Un secondo studio condotto nell'ambito del progetto ProgeNIA (Willer *et al.*, 2008), è stato focalizzato sulla caratterizzazione del profilo lipidico correlato al rischio cardiovascolare.

È noto che la malattia coronarica (*coronary artery disease*, CAD) e l'infarto sono le principali cause di morte e di disabilità nei paesi industrializzati (Mackay *et al*., 2004); alla base di questi patologie vi è l'aterosclerosi, la progressiva deposizione di colesterolo LDL (*low-density lipoprotein cholesterol* o LDL-C) nelle arterie che forniscono sangue al cuore che, nel tempo, determina il ridotto o assente rifornimento di ossigeno e conseguente infarto del miocardio (Kuulasmaa *et al*., 2000).

Numerose evidenze hanno dimostrato la correlazione tra i livelli dei lipidi associati a lipoproteine e l'incidenza di malattia cardiaca (Clarke *et al*., 2007): mentre alte concentrazioni di colesterolo LDL sono associate con aumentato rischio di CAD, elevate concentrazioni di colesterolo HDL sono associate con riduzione dello stesso. In particolare, è stato stimato che la riduzione dell'1% dei livelli ematici di colesterolo LDL riduce il rischio di CAD di circa l'1%, mentre l'incremento dell'1% del colesterolo HDL riduce il rischio di CAD del 2% (Gotto *et al*., 2004).

Una recente meta-analisi dei dati su 150.000 individui, fra i quali 3.000 casi di decesso CAD-correlati, mostra che i due fattori sono indipendentemente associati con il rischio di CAD (Prospective Studies Collaboration 2007). Vi sono anche evidenze che una elevata concentrazione di trigliceridi rappresenti un ulteriore fattore indipendente di rischio per la malattia cardiovascolare, sebbene sia ancora da stabilire se questa associazione sia causale.

Obiettivi

Sebbene numerosi geni (*LDLR, APOB, APOE*) siano stati trovati associati a livello *genome-wide* con le variazioni individuali nella concentrazione dei lipidi, gran parte della variabilità genetica del tratto rimane non identificata; per cui nell'ambito del progetto ProgeNIA si è deciso di approfondire le attuali conoscenze relative alla genetica dei lipidi facendo ricorso alla meta-analisi di studi di associazione *genome-wide* in quanto, come per altri tratti complessi, l'identificazione dei geni che influenzano i livelli dei lipidi è facilitata dal ricorso ad un campione numeroso, trattandosi di geni dal ridotto effetto fenotipico.

Meta-analisi di studi di associazione *genome-wide*

Con questo obiettivo quindi, abbiamo deciso di combinare i dati del GWAS di 4.184 volontari ProgeNIA, con i dati dei GWAS di 1.874 volontari del "Fusion study of type 2 diabetes" e di 2.758 individui del "Diabetes Genetics Initiative" (tabella 4).

Dal momento che i tre studi in oggetto hanno utilizzato differenti sets di marcatori polimorfici, con un overlap di soli 44.998 SNPs, è stato necessario integrare i dati delle tre genotipizzazioni con le informazioni disponibili su HapMap (*Haplotype Mapping Project*), al fine di identificare SNPs tag, effetuare comparazioni del linkage disequilibrium e ricostruire gli aplotipi.

Infatti, con la messa a punto di nuovi metodi di statistica inferenziale, gli aplotipi disponibili su HapMap sono stati utilizzati per inferire "*in silico*" i genotipi mancanti, ottenendo informazioni su un totale di 2.261.000 SNPs tra genotipizzati e inferiti.

Per ciascun marker é stato selezionato un allele di riferimento arbitrario ed é stato calcolato un parametro Z che, incorporando il p-*value* e la direzione dell'effetto, caratterizza l'evidenza di associazione in ciascuno studio. È stato poi calcolato un parametro Z generale, come media pesata dei parametri Z delle tre diverse coorti, e calcolato il p-*value* complessivo.

I pesi sono stati definiti come la radice quadrata del numero di individui esaminati in ciascuna coorte, normalizzati in modo tale che la somma della radice quadrata dei pesi sia uguale ad 1. Inoltre, poiché il campione includeva individui imparentati, che forniscono informazioni ridondanti, per questi è stata utilizzata una differente strategia di scelta del peso.

Risultati - GWAS

L'analisi combinata dei dati relativi alla genotipizzazione condotta in queste 3 popolazioni, per un totale di 8.816 individui, ha permesso l'individuazione di più di 18 varianti genetiche comuni (ciascuna con $p<5x10^{-7}$) indipendentemente associate con la concentrazione plasmatica di colesterolo LDL, HDL o dei trigliceridi (figura 5).

Alcune varianti sono localizzate in loci precedentemente implicati nel metabolismo lipidico, a dimostrazione della validità dell'approccio usato, mentre altre sono state mappate in loci nuovi, non descritti in precedenza (tabella 5).

Tra i loci precedentemente implicati nel metabolismo lipidico, che hanno mostrato maggior evidenza di associazione nel nostro studio, ricordo:

1) associate con i livelli del colesterolo HDL, le regioni vicine a *CETP* (rs3764261, p $<10^{-18}$, i livelli di colesterolo HDL aumentavano di 2.42 mg/dl per copia dell'allele A), *LPL* (rs12678919, p $<10^{-10}$, +2.44 mg/dl per copia dell'allele G), *LIPC* (rs10468017, p $<10^{-10}$, +1.76 mg/dl per l'allele T), *ABCA1* (rs4149274, p$<10^{-8}$, +1.51 mg/dl per copia dell'allele G) e *LIPG* (rs4939883, p$<10^{-7}$, +1.87mg/dl per copia dell'allele C);
2) associati con i livelli del colesterolo LDL, il cluster *APOE-APOC1-APOC4-APOC2* (rs4420638, p $<10^{-20}$, i livelli di LDL aumentavano di 8.02 mg/dl per copia dell'allele G), *APOB* (rs515135, p$<10^{-13}$, +6.08 mg/dl per copia dell'allele C) e *LDL-R* (rs6511720, p$<10^{-9}$, +8.03 mg/dl per copia dell'allele C);
3) associate con i livelli ematici dei trigliceridi, le regioni vicine al cluster *APOA5-APOA4-APOC3-APOA1* (rs964184, p$<10^{-15}$, i livelli di trigliceridi aumentavano di 18.12 mg/dl per copia dell'allele G), *GCKR* (rs1260326, p$<10^{-14}$, +10.25 mg/dl per copia dell'allele T) e *LPL* (rs6993414, p$<10^{-12}$, +14.20 mg/dl per copia dell'allele A).

In diversi di questi loci, gli SNPs trovati in associazione sono il linkage disequilibrium (r^2>0.80) con varianti precedentemente identificate, oppure essi stessi sono stati precedentemente descritti in associazione.

In altri loci, al contrario, in particolare le regioni vicine a *LIPC*, *LIPG*, *LDLR* e *APOB*, gli SNPs che mostrano le associazioni più significative sono solo debolmente in LD (r^2 >0.30) con varianti precedentemente identificate, rappresentando quindi nuovi segnali.

Tra questi, gli SNPs vicini ai geni *GRIN3A* (N-methyl-D-aspartate receptor subtype NR3A), *GALNT2* (Polypeptide N-acetylgalactosaminyltransferase 2), *CELSR2* (Cadherin EGF LAG seven-pass G-type receptor 2 precursor), *PSRC1* (Proline/serine-rich coiled-coil protein 1), *SORT1* (Neurotensin receptor 3), *NCAN-SF4* (Novel protein similar to vertebrate chondroitin sulfate proteoglycan family) e *TRIB1* (G-protein-coupled receptor-induced protein 2) hanno tutti *pvalues*<$5x10^{-7}$ per almeno uno dei tre tratti del profilo lipidico (tabella 5). È stata infatti osservata associazione con set di geni distinti per ciascun tratto, coerentemente con il modesto grado di correlazione tra di essi; infatti, la correlazione tra HDL ed LDL è praticamente zero nel nostro campione, la correlazione tra HDL e trigliceridi è -0.4, e la correlazione tra LDL e trigliceridi è 0.2.

Risultati – *Follow-up* dell'iniziale GWAS

Per la validazione degli SNPs trovati nello studio iniziale, è stato esaminato un sottogruppo di SNPs in sei addizionali coorti di origine europea, per un totale di 11.569 individui (tabella 5). Le analisi di *follow-up* sono state condotte in diversi steps (figura 6): inizialmente gli SNPs inclusi nell'array Affymetrix (usato nel GWAS SardiNIA e DGI), ed inferiti o genotipizzati in FUSION, sono stati selezionati per il *follow-up* sulla base della meta-analisi iniziale. In questo modo, sono stati selezionati circa 90 SNPs da testare nelle coorti ISIS, HAPI e SUVIMAX, e 67 SNPs da testare nella coorte FUSION stage 2. Una volta che l'imputazione degli SNPs presenti in HapMap é stata completata per i campioni SardiNIA e DGI, ed una ulteriore meta-analisi effettuata, sono stati esaminati 9 ulteriori SNPs in loci non scelti nell'iniziale *follow-up* nei campioni FUSION stage 2 e SUVIMAX (figura 6).

In ultimo, é stato genotipizzato un singolo SNP in ciascuno dei 21 loci che nostravano evidenze convincenti per la replica nel campione iniziale stage 2 nei campioni Caerphilly e BWHHS.

La tabella 6 riassume i risultati dello stage 2 e presenta una analisi combinata dei dati dello stage 1 e 2; tutti i loci con un *pvalue*$<5x10^{-7}$ nell'analisi iniziale sono stati confermati, ad eccezione del segnale di associazione vicino al gene *GRIN3A.*

Particolarmente degna di nota l'associazione dei livelli del colosterolo LDL con il locus *CELSR2-PSRC1-SORT1*, in quanto varianti nella regione non sono state precedentemente implicate nel metabolismo lipidico (figura 7c). Non esiste un nesso diretto tra i geni più vicino al segnale di associazione, *CELSR2* e *PSRC1*, ed il metabolismo lipidico ma una possibilità è che l'rs599839, o una variante associata, influenzi l'espressione di *SORT1*, un vicino gene che media l'endocitosi e la degradazione della lipopotein-lipasi.

Nel nostro campione, l'allele A dello SNP rs599839 è associato con un aumento dei livelli di colesterolo LDL di 5.48 mg/dl; da notare, inoltre, che lo stesso allele dello SNP rs599839 é stato recentemente trovato associato, in uno studio indipendente (Samani *et al*., 2007), ad un aumentato rischio di CAD, suggerendo che l'associazione con il rischio di CAD può essere mediato dall'effetto sulla concentrazione di colesterolo LDL.

Un altra serie di loci che ha raggiunto significatività *genome-wide* include SNPs vicini ai geni *ABCA1*, *LIPC*, *LIPG* e *PCSK9* (tabella 6); sebbene questi geni abbiano un ruolo ben noto nel metabolismo dei lipidi, alcuni dei segnali identificati nel nostro studio non

si sovrappongono con associazioni note e indicando, probabilmente, nuovi alleli di rischio. Per esempio, varianti precedentemente associate con i livelli di colesterolo LDL, in *PCSK9*, hanno $r^2<0.10$ con le varianti identificate nel nostro studio. Altri esempi di nuovi alleli di rischio includono *LIPG* (rs2156552), *LIPC* (rs4775041) e *LDLR* (rs6511720).

Questo gruppo comprende anche 6 loci in cui varianti genetiche non erano state precedentemente implicate nel metabolismo dei lipidi; una associazione è stata trovata tra il livelli di colesterolo HDL e SNPs in prossimità dei geni *GALNT2*, e *MVK-MMAB* (figura 7a-b); tra il colesterolo LDL e i trigliceridi e SNPs in una regione estesa vicino a *NCAN* e *CILP2* (figura 7c-d); e tra i trigliceridi e SNPs nelle vicinanze dei geni *TRIB1*, *MLXIPL* e *ANGPTL3* (figura 7e-h). Tra i geni presenti in queste 6 regioni, abbiamo osservato la più chiara connessione con il colesterolo ed il metabolismo delle lipoproteine per il gene *MLXIPL*, il quale codifica una proteina che lega e attiva specifici motifi nel promoter del gene di sintesi dei trigliceridi, e per il gene *ANGPTL3*, il cui omologo proteico è un regolatore principale del metabolismo lipidico nei topi. Varianti rare in un gene correlato, *ANGPTL4*, sono state associate con le concentrazioni di colesterolo HDL e trigliceridi nell'uomo. Una connessione con il metabolismo lipidico è stata inoltre osservata per *MVK* e *MMAB*, due geni vicini che sono regolati da *SREBP2* e che condividono un promoter comune (Murphy *et al*., 2007): *MVK* codifica una mevalonato chinase, la quale catalizza uno step precoce nella biosintesi del colesterolo, mentre *MMAB* codifica una proteina che partecipa ad una via metabolica che degrada il colesterolo. Negli altri 3 loci, non sono stati trovate connessioni note con il metabolismo del colesterolo.

Sono singole le associazioni che coinvolgono i geni *GALNT2* e *TRIB1*. *GALNT2* codifica una glicosiltransferasi ampiamente espressa che potrebbe potenzialmente modificare una lipoproteina o un recettore, mentre il gene *TRIB1* codifica una proteina indotta da recettori accoppiati a proteine G (GPCR) coinvolta nella regolazione della protein-chinase attivata da mitogeni e può regolare il metabolismo lipidico attraverso questo *pathway*. Al contrario, il segnale di associazione in prossimità di *NCAN* si estende per oltre 500 kb e comprende 20 geni. Nei nostri dati combinati, l'rs16996148 (uno SNP presente nel chip Affymetrix vicino a *CILP2*) è stato selezionato per il *follow-up* mostrando associazione significativa sia con il colesterolo LDL ($P\sim2.7\times10^{-9}$) sia coi trigliceridi ($P\sim2.5\times10^{-9}$). L'allele oltre ad essere associato con aumentate concentrazioni di colesterolo LDL, è anche associato con aumentati livelli di trigliceridi, dato coerente

con la modesta correlazione positiva tra i due tratti, ma in contrasto con altri SNPs associati con entrambi i tratti che, nel nostro campione, mostravano associazione con uno solo di essi. Si noti che, con l'analisi combinata dei tre GWAS e degli SNPs imputati con HapMap, uno SNP *coding* non-sinonimo nel gene *NCAN* (rs2228603, Pro92Ser) ha mostrato la più forte evidenza di associazione ($P\sim1.8x10^{-7}$). Questo SNP che non era stato incluso nella iniziale analisi di *follow-up,* la quale considerava solo SNPs sul chip Affymetrix, era in forte linkage disequilibrium con rs16996148 ($r^2=0.89$). *NCAN* é un proteoglicano specifico del sistema nervoso coinvolto nella formazione del pattern neuronale, rimodellamento dei networks neuronali e nella regolazione della plasticità sinaptica (Rauch *et al*., 2001), per il quale non esiste una scontata correlazione con le concentrazioni del colesterolo LDL o trigliceridi.

Risultati – Associazione con la malattia coronarica (CAD)

In considerazione della nota associazione tra i livelli ematici dei lipidi e la malattia delle arterie coronarie (CAD), abbiamo voluto verificare se gli alleli associati con i livelli di lipidi nel presente studio fossero anche associati con la malattia coronarica in un campione di ~2.000 casi di CAD e nel pannello di referenza esteso di ~13.000 britannici (inclusi ~3.000 controlli random e ~2.000 casi per ciascuna di 5 malattie comuni) del WTCCC (The Welcome Trust Case Control Consortium 2007).

Considerati i relativamente modesti cambiamenti nelle concentrazioni del colesterolo LDL associati con gli alleli identificati nel presente studio (variazioni di ~2-9 mg/dl per allele), ci aspettavamo che un subset di SNPs potesse essere anche associato con un piccolo incremento nella suscettibilità a CAD. In particolare, i risultati hanno indicato che tutti gli alleli associati nel nostro campione con un aumento dei livelli di colesterolo LDL sono più comuni tra i casi di CAD rispetto al pannello di referenza esteso (tabella 7).

Tutti e 11 gli alleli indipendentemente associati ($r^2<0.30$ tra alleli vicini) con un aumento della concentrazione di colesterolo LDL nel nostro campione ($P<10^{-6}$), presentavano una aumentata frequenza tra i casi di CAD ($P<0.0005$). L'aumento era significativo ($P<0.05$) per nove SNPs, e vicino alla significatività ($P<0.06$) per altri due. Nonostante le stime del rischio associato siano piccole (aumento del rischio relativo di 1.04-1.29 per allele), é estremamente improbabile ($P<10^{-11}$) che 10 degli 11 SNPs mostrino associazione significativa con CAD con $P<0.06$ per effetto del caso, rendendo la connessione tra LDL, SNP associato e CAD particolarmente degna di nota.

Complessivamente, sebbene sia stata osservata una correlazione tra la forza dell'associazione con CAD e l'impatto di ciascun allele sui livelli di colesterolo LDL (coefficiente di correlazione di Spearman r = 0.71, P=0.015), é stato anche trovato qualche allele che aveva forte associazione con il colesterolo LDL ma associazione non significativa con CAD (per esempio, rs562338 nel locus APOB). Non è stato trovato un simile pattern di associazione per gli alleli associati con gli altri tratti del profilo lipidico, per quanto alleli nelle vicinanze del gene *TRIB1* associati con aumentati livelli di trigliceridi (per esempio, l'rs17321515) risultavano anche associati con aumentato rischio di CAD (P=0.0008).

Sebbene i dati suggeriscano che quasi tutti gli alleli associati con aumentate concentrazioni di colesterolo LDL siano anche associati con aumentato rischio di CAD (considerata la notevole dimensione campionaria), il contrario non è vero; infatti, alleli in un *locus* del cromosoma 9 che mostrano forte associazione con CAD e infarto del miocardio (Helgadottir *et al.*, 2007) non sembrano influenzare la concentrazione dei lipidi nel nostro campione (P=0.31 per l'associazione tra colesterolo LDL e lo SNP associato con la più alta significatività statistica con CAD, rs1333049, nel nostro stadio 1, e P>0.50 per il colesterolo HDL e trigliceridi).

Studi futuri dimostreranno se queste varianti sono anche associate con la longevità (Barzilai *et al*., 2003), l'infarto (Baigent *et al*., 2005) e gli altri *outcomes* della salute associati con elevate concentrazioni di colesterolo LDL.

Conclusioni

I geni individuati nel presente studio influenzano l'intero ciclo di formazione, attività e *turnover* di lipoproteine e trigliceridi; per esempio, essi codificano diverse apolipoproteine (*APOE*, *APOB* e *APOA5*), un fattore di trascrizione attivante la sintesi dei trigliceridi (*MLXIPL*), un enzima coinvolto nella biosintesi del colesterolo (*MVK*), alcuni trasportatori del colesterolo (*ABCA1*) ed esteri del colesterolo (*CETP*), un recettore per la "lipoproteina" (*LDLR*), una potenziale glicosiltransferasi (*B4GALT4*, *B3GALT4* e *GALNT2*), lipasi (*LPL*, *LIPC* e *LIPG*), una proteina coinvolta nella degradazione del colesterolo (*MMAB*), un inibitore della lipasi (*ANGPTL3*) ed un possibile recettore dell'endocitosi per LPL (*SORT1*).

Da rilevare che numerosi altri geni identificati (per es., vicino al gene *TRIB1* e nella regione del gene *NCAN*) non includono buoni candidati per funzione, per cui questi risultati aprono la via a nuovi studi, quali il sequenziamento esteso a tutto il genoma

(attualmente in corso), volti all'identificazione di possibili nuove vie metaboliche riguardanti trigliceridi e colesterolo, che potranno portare a importanti nuove intuizioni sulla regolazione del metabolismo dei lipidi.

Prospettive future

Dai modelli di regressione multipla, che valutano più variabili e/o tratti contemporaneamente, risulta che le varianti identificate nel presente studio complessivamente giustificano solo il 5-8% della variabilità dei livelli dei tre tratti del profilo lipidico nella coorte presa in esame, cosicchè gran parte della ereditabilità di questi tratti rimane non spiegata.

I fattori genetici non identificati sono ipoteticamente rappresentati da una lista molto più lunga di varianti comuni dal ridotto effetto sul fenotipo, varianti rare con un maggior effetto sul fenotipo, non individuate dall'approccio di associazione usato, o dalla interazione tra queste ed altri fattori genetici e ambientali.

Per charire il ruolo dei loci-geni identificati in questo studio, é fondamentale il sequenziamento degli esoni e delle regioni conservate in un grande numero di individui, al fine di identificare e valutare tutte le potenziali varianti all'interno di ciascun gene o cluster di geni.

Il *re-sequencing*, oltre ad identificare le varianti funzionali coinvolte in ciascuna regione, potrà identificare mutazioni *non-sense*, non sinonime o altri cambiamenti che sono associati con la variabilità nella concentrazione dei lipidi, chiarendo così l'identità dei geni coinvolti in regioni con numerosi candidati; inoltre, il *re-sequencing* di geni candidati ha dimostrato, in altri studi, che le suddette varianti rare possono talvolta essere identificate in individui agli estremi della distribuzione della concentrazione dei lipidi (Cohen *et al.*, 2004); cosi, lo studio mirato delle regioni identificate in questo lavoro, in individui con dislipidemie, potrebbe essere particolarmente informativo.

Molti dei loci identificati, per la prima volta, in questo lavoro rappresentano potenziali bersagli farmacologici, e la possibilità di stratificare gli individui sulla base dello specifico profilo genetico può fornire benefici futuri per l'ottimizzazione della terapia, dato che i farmaci che riducono i livelli dei lipidi sono già ampiamente prescritti al fine di gestire meglio il profilo lipidico individuale e ridurre il rischio di eventi cardiovascolari (Law *et al.*, 2003); così, la nostra speranza è che le varianti comuni da noi identificate portino allo sviluppo di nuovi farmaci e determinino un profilo di trattamento ottimale per ciascun individuo.

SEQUENZIAMENTO ESTESO A TUTTO IL GENOMA PER LO STUDIO DEL PROFILO LIPIDICO

Presupposti dello studio

Nonostante i progressi degli ultimi anni nel mappaggio delle varianti comuni (frequenza dell'allele minore o MAF>5%), recenti studi hanno dimostrato che anche le varianti rare hanno un ruolo importante nell'eziologia dei tratti complessi e si suppone che la loro identificazione avrà un impatto notevole sulla valutazione del rischio e prevenzione delle malattie, a causa del loro maggior effetto sul fenotipo (Li *et al*., 2009).

I microchips commerciali utilizzati per gli studi di associazione *genome-wide* sono disegnati per fornire eccellente copertura degli SNPs comuni, mediante genotipizzazione di tag-SNPs che sono *proxies* per varianti causali comuni, ma hanno solo un limitato potere nel catturare le varianti rare (MAF<5%), e non valutano direttamente il contributo di corti polimorfismi *indels* (inserzioni o delezioni); per esempio, nel chip Affymetrix 500K, usato nel GWAS della coorte SardiNIA, sono presenti solo 55.000 SNPs con MAF<0.05 (varianti rare) e 17.000 SNPs con MAF<0.01 (varianti rarissime).

La misura con la quale queste varianti a bassa frequenza e penetranza intermedia contribuiscono alla predisposizione delle malattie, spiegando la grande proporzione del rischio genetico ancora non mappato, rappresenta una delle principali domande della genetica umana ancora senza risposta. Tali varianti infatti non vengono rilevate dalle moderne tecnologie *genome-wide*, in quanto nè abbastanza penetranti per mostrare segregazione mendeliana, ed essere mappate mediante approcci di *linkage* tradizionali, nè abbastanza frequenti per essere rilevate dagli approcci di GWA (McCarthy *et al*., 2008).

Nonostante l'identificazione delle varianti rare rappresenti, quindi, una notevole sfida, i recenti avanzamenti delle tecnologie di sequenziamento *high-throughput*, congiuntamente al completamento della fase pilota del "Progetto 1000 genomi" (1000 Genomes Project Consortium *et al*., 2010), hanno permesso un rapido progresso nel sequenziamento dei geni candidati e, addirittura, di interi genomi consentendo di ottenere dati sulle varianti rare da utilizzare negli studi di associazione delle malattie complesse.

Obiettivi

Sulla base di queste premesse, a seguito di una consolidata collaborazione tra l'INN-CNR ed il Prof. Goncalo R. Abecasis, del Center for Statistical Genetics, Dept. of Biostatistics, University of Michigan (USA), ho avuto la possibilità di svolgere parte dell'ultimo anno di dottorato di ricerca presso il suddetto centro, dove sono stato coinvolto nell'ambizioso progetto di sequenziamento dell'intero genoma di 2.000 volontari del progetto ProgeNIA, "*Genetics of lipid levels: draft sequencing of 1.000 genomes*", impiegando le nuove piattaforme di sequenziamento ad alta processività.

Attraverso il sequenziamento del DNA dei volontari della nostra popolazione fondatrice, già dimostratasi utile negli studi di associazione *genome-wide*, e l'identificazione dei genotipi di SNPs, corti polimorfismi *indels*, polimorfismi del numero di copie (CNP) ed altre varianti strutturali, ci si pone l'obiettivo di studiare l'architettura genetica sia dei tratti che caratterizzano il profilo lipidico, sia di numerosi altri tratti per cui sono disponibili le informazioni fenotipiche longitudinali degli stessi volontari reclutati nello studio.

Disegno Sperimentale - Scelta del tipo di sequenziamento da eseguire

Sebbene il sequenziamento ad alto *coverage* (*depth* ~30X) di singoli genomi sia in grado di rilevare tutte le varianti, comuni e rare, presenti negli individui sequenziati, i costi per la sua attuazione sono talmente elevati che può essere preso in esame solo un numero limitato di individui.

È quindi essenziale considerare strategie alternative che forniscano informazioni sull'intero genoma di migliaia di individui; l'idea alla base del progetto "*Genetics of lipid levels: draft sequencing of 1.000 genomes*", qui descritto, è sequenziare a basso coverage (~2-4X) 1.000-2.000 individui della coorte ProgeNIA e, per fare questo nella maniera più economica, si è deciso di adottare la strategia che combina le tecnologie di sequenziamento *shotgun* con gli stessi strumenti statistici usati per l'imputazione dei genotipi nei GWAS.

Di seguito sono discussi i risultati di una simulazione (Y. Li & G. Abecasis, dati non pubblicati) concepita per predirre cosa é possibile ottenere, supposto un *badget* di spesa fisso, sequenziando 67 individui ad alto coverage (30X) o 1.000 individui a basso coverage (2X), per un totale di 2.000X in entrambi i casi (tabella 8).

Entrambi i metodi hanno un potere eccellente (~100%) nel rilevare varianti con MAF>5%, ed il sequenziamento a basso coverage o *low-pass* mostra un potere maggiore nel rilevare le varianti meno comuni (MAF=0.5-5.0%).

La maggior parte delle varianti con frequenza<0.5% sono risultate non rilevabili con entrambi gli approcci: con il *deep sequencing*, infatti, la maggior parte delle varianti di interesse non sono polimorfiche nei 67 individui selezionati per il sequenziamento; con l'analisi *low coverage* non ci sono abbastanza copie di ciascun aplotipo per poter combinare efficacemente le informazioni tra i campioni.

Per le varianti identificate, l'accuratezza dei genotipi seppur ridotta nell'analisi *low-pass*, era ancora notevole; per esempio, per le varianti con frequenza>1%, l'accuratezza del sequenziamento *low-pass* é sempre maggiore del 99.5% per tutti i siti e dell'89.5% per i siti eterozigoti, i quali sono più difficili da identificare correttamente. Un elevata frequenza di identificazione dei polimorfismi e l'accuratezza dei genotipi sono possibili in quanto il modello combina efficacemente le informazioni tra gli individui con simili aplotipi.

Un altra importante misura dell'accuratezza é la correlazione tra i genotipi identificati con le sequenze ed i genotipi reali (r^2): questa quantità può essere usata per stimare l'effettiva dimensione campionaria ($n \cdot r^2$) per i successivi test di associazione (Pritchard and Przeworski 2001).

In tutti i casi considerati, il sequenziamento *low-pass* di 1.000 individui fornisce maggiori informazioni rispetto al sequenziamento di 67 individui ad alto coverage; per esempio, per varianti con frequenza 0.5-1.0%, 1.0-2.0%, 2.0-5.0% e 5.0% o maggiore, il sequenziamento 2X di 1.000 individui fornisce una effettiva dimensione del campione di 567, 761, 883 e 978 individui (tabella 8), tutte sostanzialmente maggiori dei 67 individui che possono esere esaminati con il deep sequencing.

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

Disegno Sperimentale - Selezione degli individui da sequenziare

La selezione dei campioni per un progetto di sequenziamento deve tenere conto di diverse possibilità: una possibile strategia è prendere in esame gli individui agli estremi della distribuzione dei livelli di lipidi ematici; se da un lato il valore di questo approccio

é riconosciuto, dall'altro non é stato preferito in quanto questo studio prende in esame contemporaneamente 3 tratti (colesterolo HDL, colesterolo LDL e trigliceridi) che mostrano solo modesta correlazione tra di loro; in aggiunta, dato l'alto grado di parentela del nostro campione, questo approccio risulterebbe nel sequenziare molti individui strettamente imparentati, riducendo la capacità di indagare la diversità genetica del campione.

La strategia da noi usata consiste, invece, nel selezionare individui non imparentati (marito e moglie); infatti, come osservato in precedenza, molti individui strettamente imparentati saranno disponibili per ciascun individuo sequenziato e sarà, in seguito, possible propagare probabilisticamente i dati della sequenza a questi individui attraverso l'identificazione degli aplotipi che essi condividono con i campioni sequenziati (o regioni IBD o "*identity-by-descent*"); in questa maniera, oltre ad aumentare l'accuratezza dei genotipi e ricostruire con maggior precisione gli aplotipi (fasi), sarà possibile aumentare l'effettiva dimensione del campione per le successive analisi di associazione.

Per cui è stato deciso di sequenziare i trios (padre, madre e figlio) disponibili nel campione ProgeNIA; laddove assente 1 genitore, le triplette composte dall'altro genitore disponibile e 2 figli oppure, se assenti entrambi i genitori, triplette di fratelli.

Risultati – Dati preliminari sui campioni sequenziati

Ad oggi sono stati selezionati, secondo i criteri descritti nel paragrafo precedente, un totale di 474 individui così suddivisi:

1. 93 trios: padre, madre e figlio
2. 32 triplette 1 genitore + 2 figli
3. 33 triplette di 3 fratelli

Sono state preparate le librerie di DNA genomico per 379 di questi campioni e 349 sono stati sequenziati con il Genome Analyzer *IIx* o, recentemente, con l'Hi-Seq (entrambi Illumina).

Il sequenziamento dei campioni in corse *paired-end* di 240 basi con il GAII_X e 204 basi con l'Hi-Seq ha prodotto, in media, ~9 milioni di basi (9 GB) ad alta qualità (ciascuna con Q20 o maggiore) per campione, corrispondenti a 37.5 milioni di clusters utilizzabili, e ~15 milioni di basi (15 GB) ad alta qualità per campione (ciascuna con Q20 o maggiore), corrispondenti a 74 milioni di clusters utilizzabili, rispettivamente, dopo il

mappaggio delle *reads* al genoma di referenza, la ricalibrazione del *Phred score* delle basi mappate e la rimozione dei duplicati.

Questi valori corrispondo ad un coverage per campione di 3,6X e 5,5X, per corse della durata di 10-12 giorni eseguite con GA*II*$_x$ e Hi-Seq, rispettivamente, *coverage* sufficienti per uno studio *low-coverage* come questo.

L'analisi dei primi 226 campioni sequenziati ha rilevato, dopo l'applicazione di una serie di filtri di qualità, 10.759.492 SNPs, dei quali il 54.3% sono presenti in dbSNP, per un totale di 5.845.709 SNPs (tabella 8).

Conclusioni

L'approccio da noi scelto per il sequenziamento *low-coverage* di un elevato numero di volontari del progetto ProgeNIA, utilizzato anche nella fase pilota del Progetto 1000 genomi, è risultato efficace per l'individuazione dei genotipi e per la successiva propagazione degli stessi ai parenti degli individui sequenziati, al fine di ricostruire gli aplotipi, attraverso il confronto dei dati delle sequenze con dati genotipici già disponibili.

È attualmente in corso l'analisi dei 300 campioni ad oggi sequenziati, che permetterà la scoperta di numerosi altri polimorfismi, con miglioramento della qualità di imputazione.

È inoltre in corso l'analisi di associazione *genome-wide* estesa a tutto il campione ProgeNIA, per tutti i tratti quantitativi analizzati nel primo GWAS.

Prospettive future

Il miglioramento dei software per la chiamata delle basi, reagenti per le reazioni di sequenza sempre più performanti, insieme ad una ottimale titolazione dei campioni e, quindi, dei *clusters* hanno reso possibile il raggiungimentodi un coverage notevole (5.5X) tale che è stato possibile pianificare, e sono attualmente in corso, esperimenti di *multiplexing* al fine di sequenziare un numero più elevato di campioni per *lane*, con risparmio economico e di tempo.

In un futuro prossimo sarà inoltre possibile individuare, oltre agli SNPs, inserzioni e delezioni e varianti strutturali (CNV, CNP), grazie alla ottimizzazione di algoritmi appositamente in esame.

CONCLUSIONI

Il disegno sperimentale applicato al progetto ProgeNIA, descritto nella mia tesi, si è dimostrato di grande successo in quanto ha permesso di studiare la genetica di numerosi tratti quantitativi in un campione numeroso, reclutato dalla popolazione generale.

È stato così possibile utilizzare un singolo *data set* genetico per condurre un GWAS su tutti i tratti esaminati e, negli ultimi 4 anni, il nostro gruppo di ricerca ha pubblicato i risultati ottenuti in più di 50 lavori nelle riviste più prestigiose nel campo della genetica umana.

Il successo di questo disegno sperimentale ha inoltre ribadito l'importanza di studiare i tratti quantitativi correlati a patologie comuni, piuttosto che le patologie stesse, consentendo la scoperta di numerosi geni e quindi la comprensione delle basi genetiche e molecolari sia di patologie monogeniche e, soprattutto, complesse.

GRAFICI E TABELLE

Figura 1 Inferenza statistica dei genotipi in individui imparentati
(*a*) I dati osservati, consistono nei genotipi di una serie di marcatori genetici. In questo caso, alcuni dei marcatori sono stati tipizzati in tutti gli individui (rosso), i rimanenti sono stati tipizzati solo in alcuni individui (nero, negli individui nelle prime 2 generazioni del pedigree). (*b*) Processo di inferenza delle informazioni sullo stato di IBD, mediante valutazione dei marcatori i cui genotipi sono disponibili in tutti gli individui. A ciascun segmento IBD presente in più di un individuo viene assegnato un unico colore; per es., un segmento blu é condiviso tra il primo individuo nella 1° generazione (in alto nel pedigree), il primo individuo nella 2° generazione e negli individui 3 e 4 nella 3° generazione. (*c*) I genotipi osservati e le informazioni di IBD sono stati combinati per completare una serie di genotipi che erano originariamente mancanti nella 3° generazione.

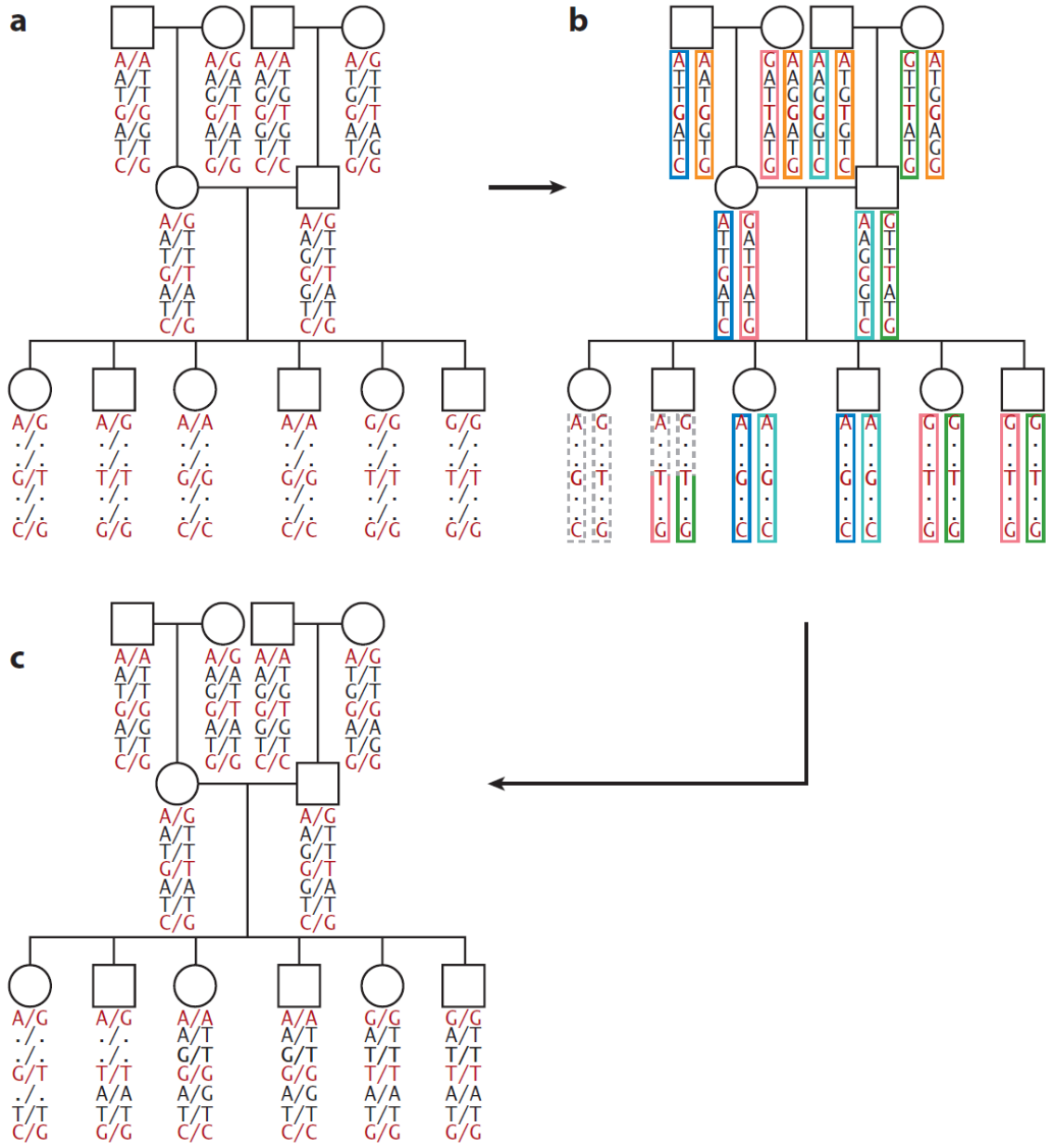


Figura 2a Sommario dei risultati preliminari dello studio di associazione *genome-wide.* Analisi dei 1.412 individui genotipizzati con i chip 500K, prima dell'imputazione dei genotipi, per l'associazione con i globuli rossi (RBC), contenuto emoglobinico corpuscolare medio (MCH), emoglobina (HB) ed emoglobina glicosilata (HBA1C)

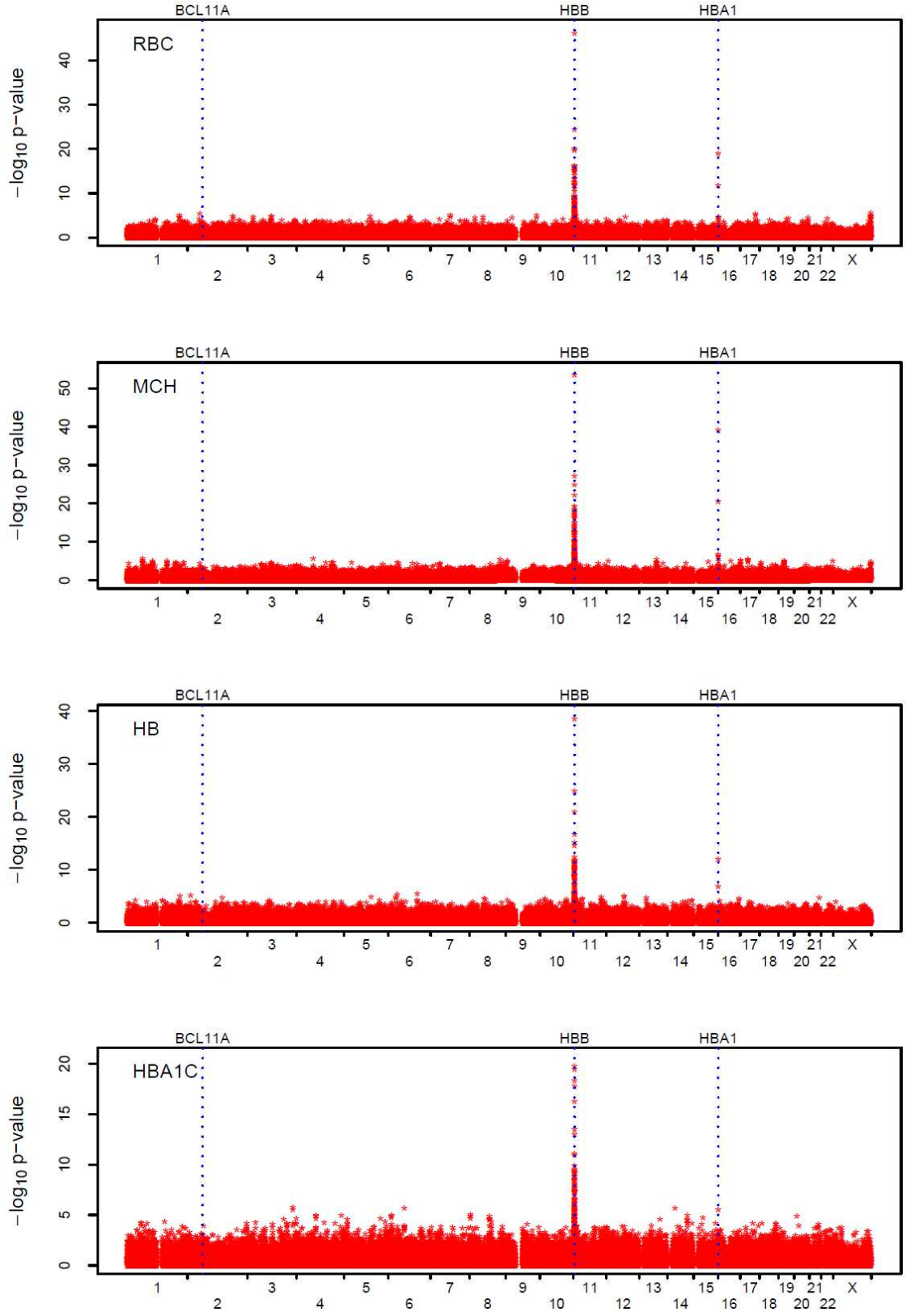


Figura 2b Sommario dei risultati preliminari dello studio di associazione *genome-wide*. Analisi dei 1.412 individui genotipizzati con i chip 500K, prima dell'imputazione dei genotipi, per l'associazione con il volume corpuscolare medio (MCV), contenuto emoglobinico corpuscolare medio (MCHC), HbA2 ed emoglobina fetale (HBF).

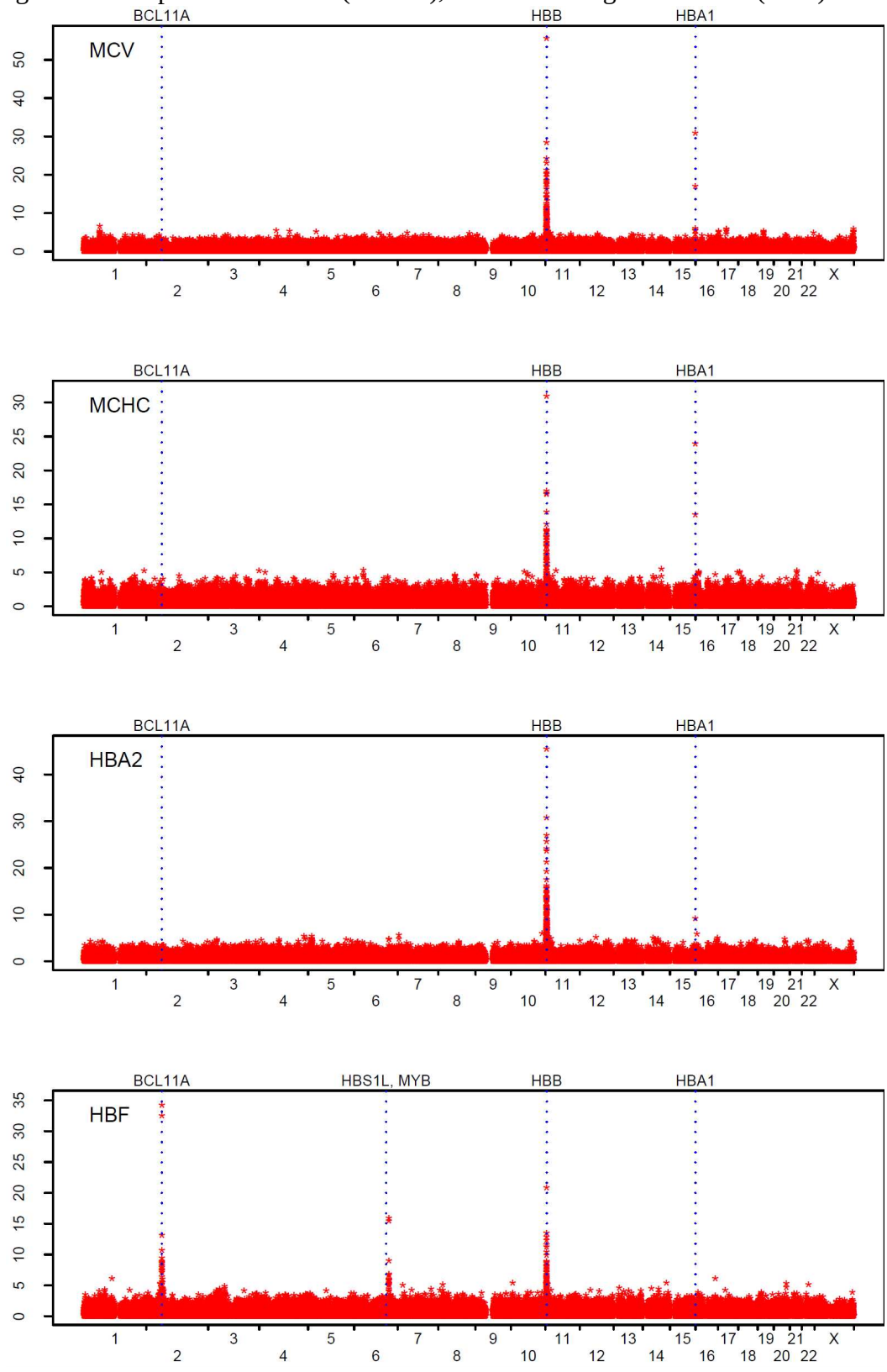


Figura 3

Associazione con i livelli di HbF e pattern di linkage disequilibrium nella regione del gene *BCL11A*. Sommario dell'associazione (A in alto) tra gli SNPs ed i livelli di HbF in ciascun individuo (-$\log_{10}$ del *P* value). In rosso lo SNP che mostra l'associazione più significativa (rs11886868). Gli altri SNPs sono colorati in accordo al loro grado di disequilibrio con lo SNP rs11886868, variando da alto (rosso) a intermedio (verde) a basso (blu). (A in basso) Sono mostrati i trascritti alternativi del *BCL11A*, con una freccia indicante la direzione del trascritto. (B) Sommario del pattern di linkage disequilibrium della regione genomica in Sardegna ed in due delle popolazioni HapMap (CEU e YRI). La legenda dei colori dei valori di r^2 è uguale ad A. La barra grigia indica la regione di associazione e facilita le comparazioni tra immagini.

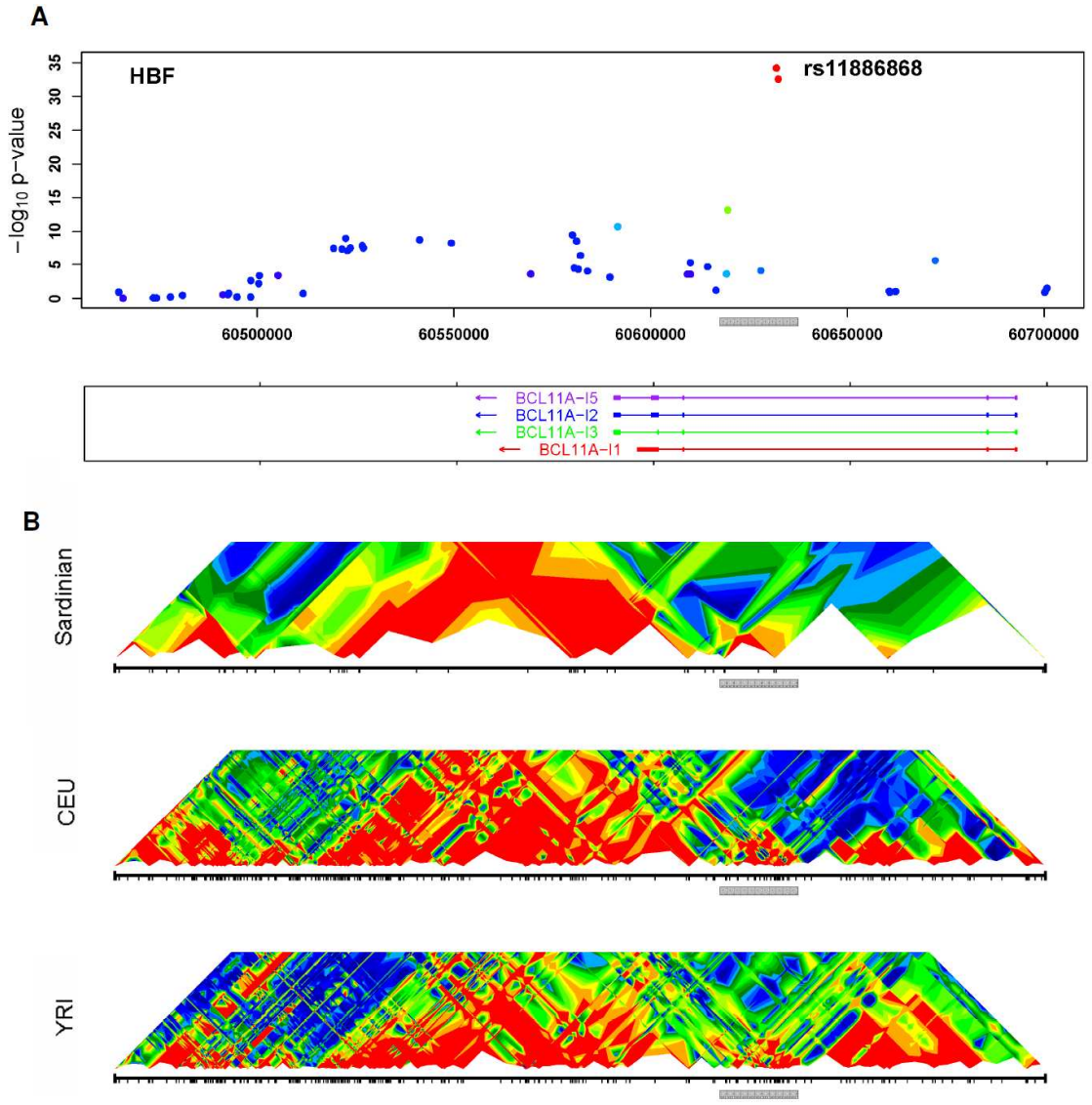


Figura 4
Associazione tra i genotipi dello SNP rs11886868 ed i livelli di HbF.
Il grafico mostra la distribuzione dei livelli di HbF per ciascuna classe di genotipi nel campione di follow-up.

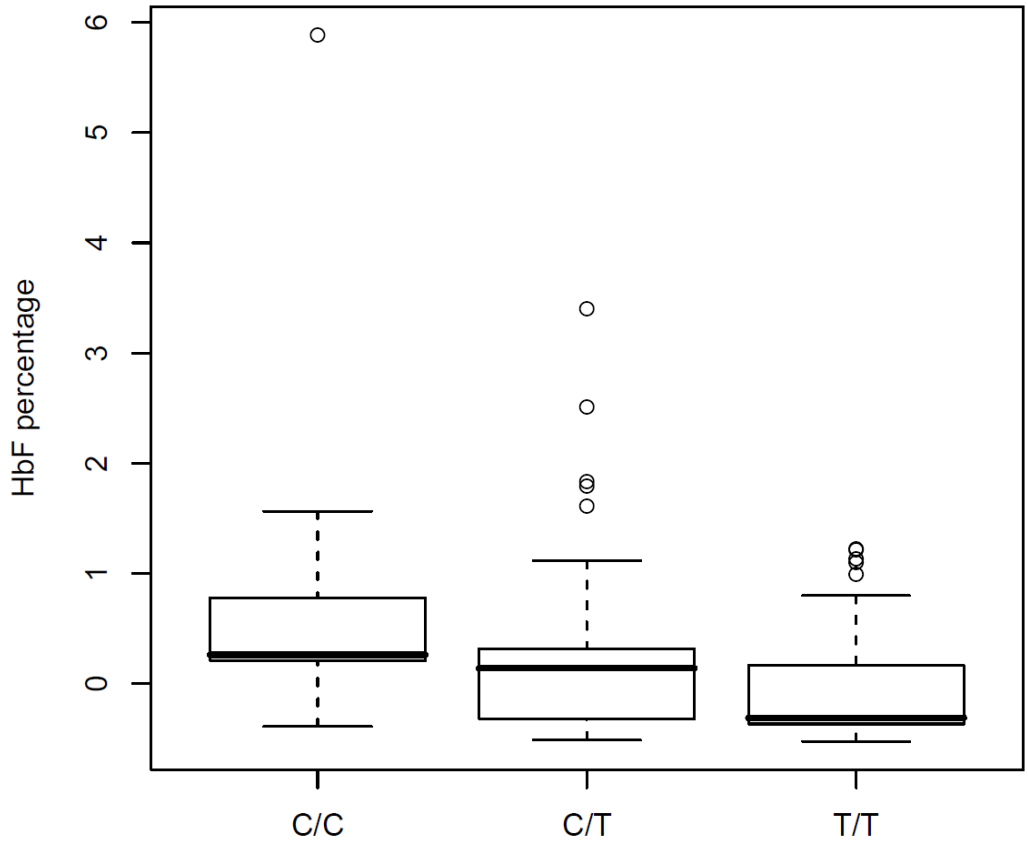


Figura 5
Riassunto dello studio di associazione *genome-wide* per i tratti del profilo lipidico.
I 3 pannelli superiori riassumono i risultati combinati dello studio di associazione *genome-wide* (plottati come -$\log_{10}$ P *value* per il colesterolo HDL, colesterolo LDL ed i trigliceridi). In grigio, i loci per i quali non è stato condotto un *follow-up*; i loci per i quali è stato eseguito il *follow-up* sono in verde (*dataset* combinato con chiare evidenze di associazione, $P<5x10^{-8}$), arancione (*dataset* combinato con promettenti evidenze di associazione, $P<10^{-5}$), o rosso (*dataset* combinato che non indica associazioni, $P>10^{-5}$). I 3 pannelli in basso mostrano i grafici *quantile-quantile* per i test statistici. La linea rossa corrisponde a tutti i test statistici, la linea blu ai risultati dopo aver escluso le statistiche ai loci che hanno replicato (in verde, nel pannello superiore), e l'area grigia corrisponde all'intervallo di confidenza del 90% della distribuzione nulla di P *values* (generata con 100 simulazioni).

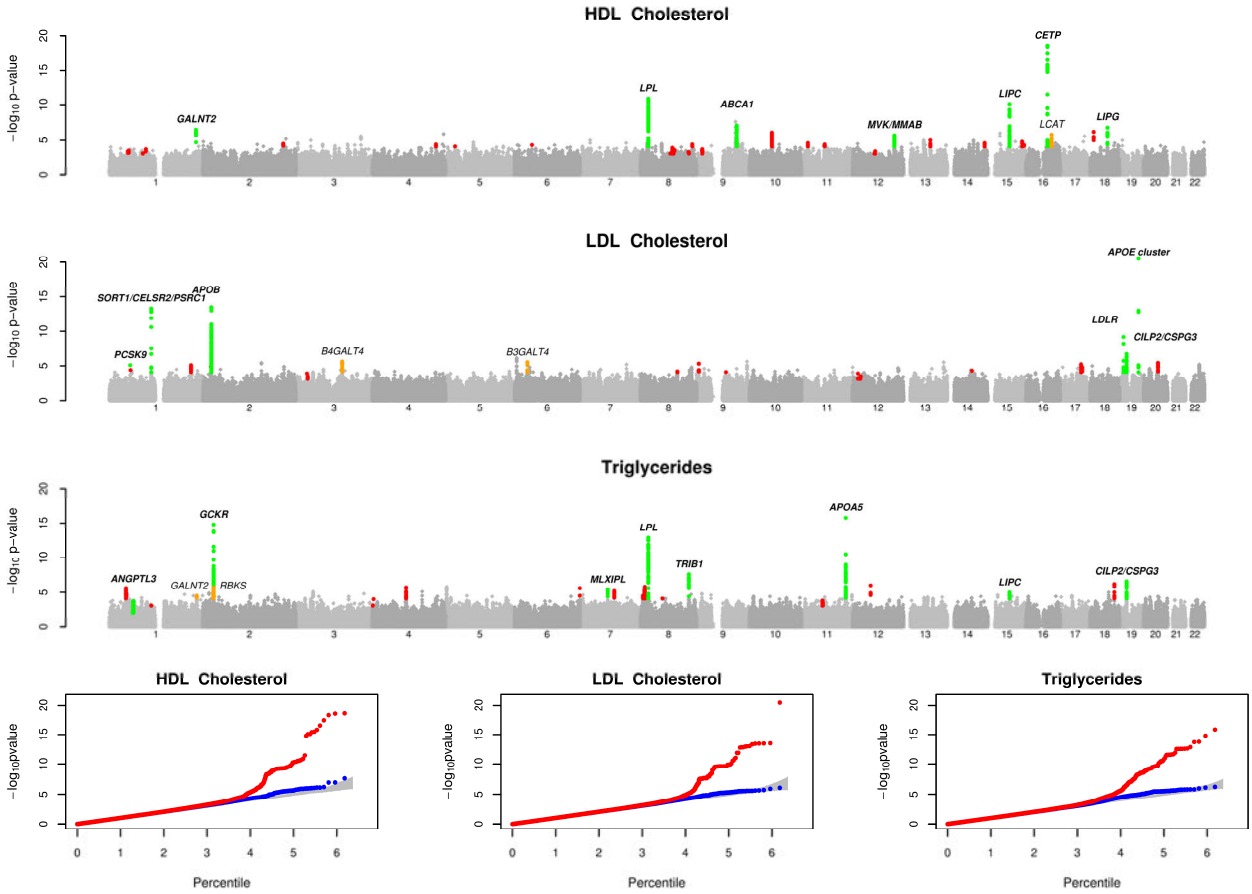
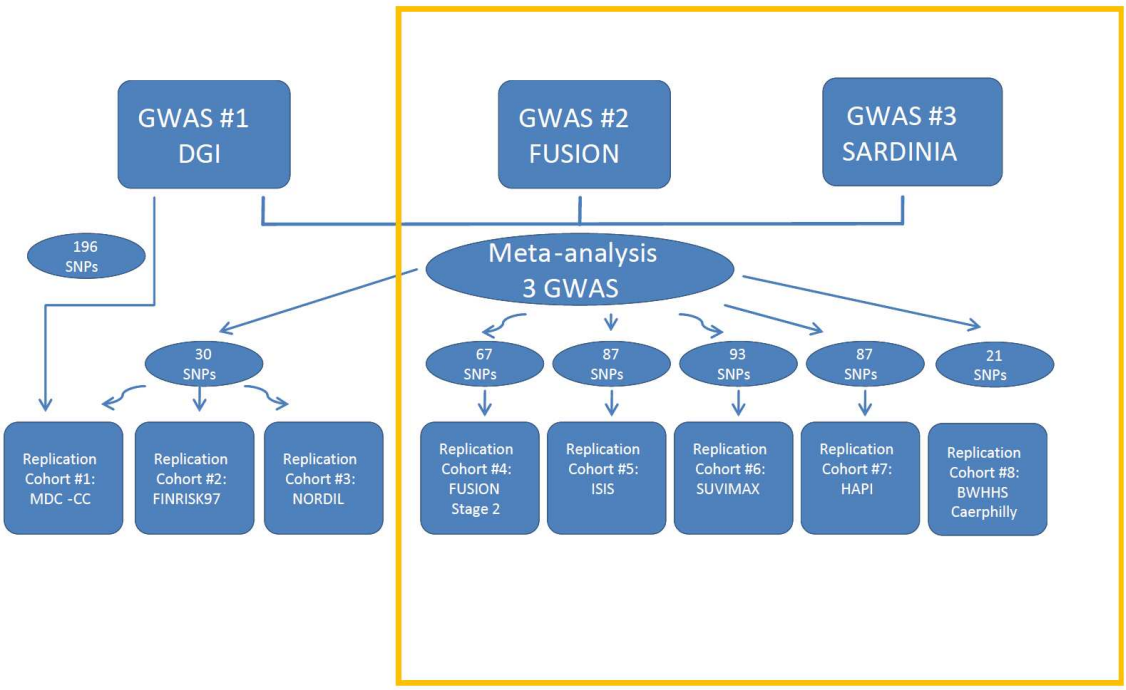


Figura 6
Disegno sperimentale per la meta-analisi dei tre GWAS e *follow-up* dei risultati.



La coorte di FUSION é stata genotipizzata con il BeadChip Illumina HumanHap300 BeadChip. I 4.305 individui dello studio SardiNIA sono stati tipizzati con il chip 10K Affymetrix (N=2.893), con il chip 500K (N=976) o entrambi (N=436) per permettere l'imputazione su base familiare. Un gruppo di 4.184 dei suddetti individui, non sottoposti a terapia per controllare i livelli dei lipidi, è stato usato per le analisi qui descritte. Il campione DGI é stato tipizzato con il chip 500K. Dopo imputazione con il chip 500K nel campione FUSION *stage*1, é stata eseguita una meta-analisi con i risultati di DGI e SardiNIA (N=8.816), e 87-93 degli SNPs associati in maniera più significativa ($P<7x10^{-5}$) sono stati scelti per il *follow-up* nei campioni ISIS, SUVIMAX, e HAPI. 67 SNPs ($p<5x10^{-6}$) sono stati genotipizzati nel campione FUSION *stage*2. Dopo una analisi preliminare dei risultati dello *stage*2 degli studi FUSION, ISIS, SUVIMAX e HAPI, 21 SNPs nei geni più promettenti sono stati selezionati per genotipizzazione nei campioni BWHHS e Caerphilly. Il campione combinato *stage*2 include i genotipi per più di 11.569 individui. Il box arancione indica esperimenti e dati unici di questo manoscritto. Il GWAS dei livelli di lipidi FUSION e SardiNIA sono riportati per la prima volta. La meta-analisi dei 3 GWAS ed i risultati del *follow-up* in 6 coorti (ISIS, HAPI, SUVIMAX, FUSION *Stage*2, BWHHS, Caerphilly) sono ugualmente riportate per la prima volta.

Figura 7 A, B, C, D, E, F, G, H

Riassunto dei nuovi loci trovati in associazione con i tratti del profilo lipidico. Ciascun pannello comprende una regione di 500 kb (eccetto i pannelli D ed F, 800 kb) e mostra evidenze di associazione nei dintorni di uno dei nostri segnali replicati (p*value* $<10^{-8}$). In alto in ciascun pannello, diagrammi a pettine indicano la posizione degli SNPs genotipizzati con successo in FUSION, SardiNIA, e DGI, e degli SNPs inferiti. Al centro di ciscun pannello, é riassunta l'evidenza di associazione per tutti gli SNPs in esame nello *Stage*1. Lo SNP che mostra la più forte evidenza di associazione nella regione è indicato con un quadrato rosso, e gli altri SNPs sono colorati in accordo con il loro grado di linkage disequilibrium con il top SNP. In ciascun locus, uno degli SNPs validati nello *stage*2 é evidenziato insieme ad un p*value* combinatoche prende in considerazione i dati dello *stage*1 e dello *stage*2. Da notare che, dal momento che il *follow-up* è stato principalmente eseguito sugli SNPs presenti nel chip Affymetrix 500K, i più forti segnali di associazione e gli SNP selezionati per il *follow-up* non sempre corrispondono. Il pannello inferiore, riassume la posizione dei geni in ciascuna regione. Per chiarezza visiva, alcuni geni sono stati omessi nei pannelli D ed F. Da sinistra a destra, le etichette per i geni presenti nei pannelli D ed F sono: *SLC25A42, TMEM161A, MEF2B, RFXANK, TRA16, NCAN, HAPLN4, TM6SF2, SF4, KIAA0892, GATAD2A, TSSK6, NDUFA13, FLJ44968, CILP2, PBX4, EDG4, GMIP, ATP13A1, ZNF101, ZNF14, ZNF253, ZNF93, ZNF682, FLJ44894, ZNF626, ZNF85, ZNF430, ZNF714*.

Figura 7A, B Riassunto dei nuovi loci trovati in associazione con i tratti del profilo lipidico.

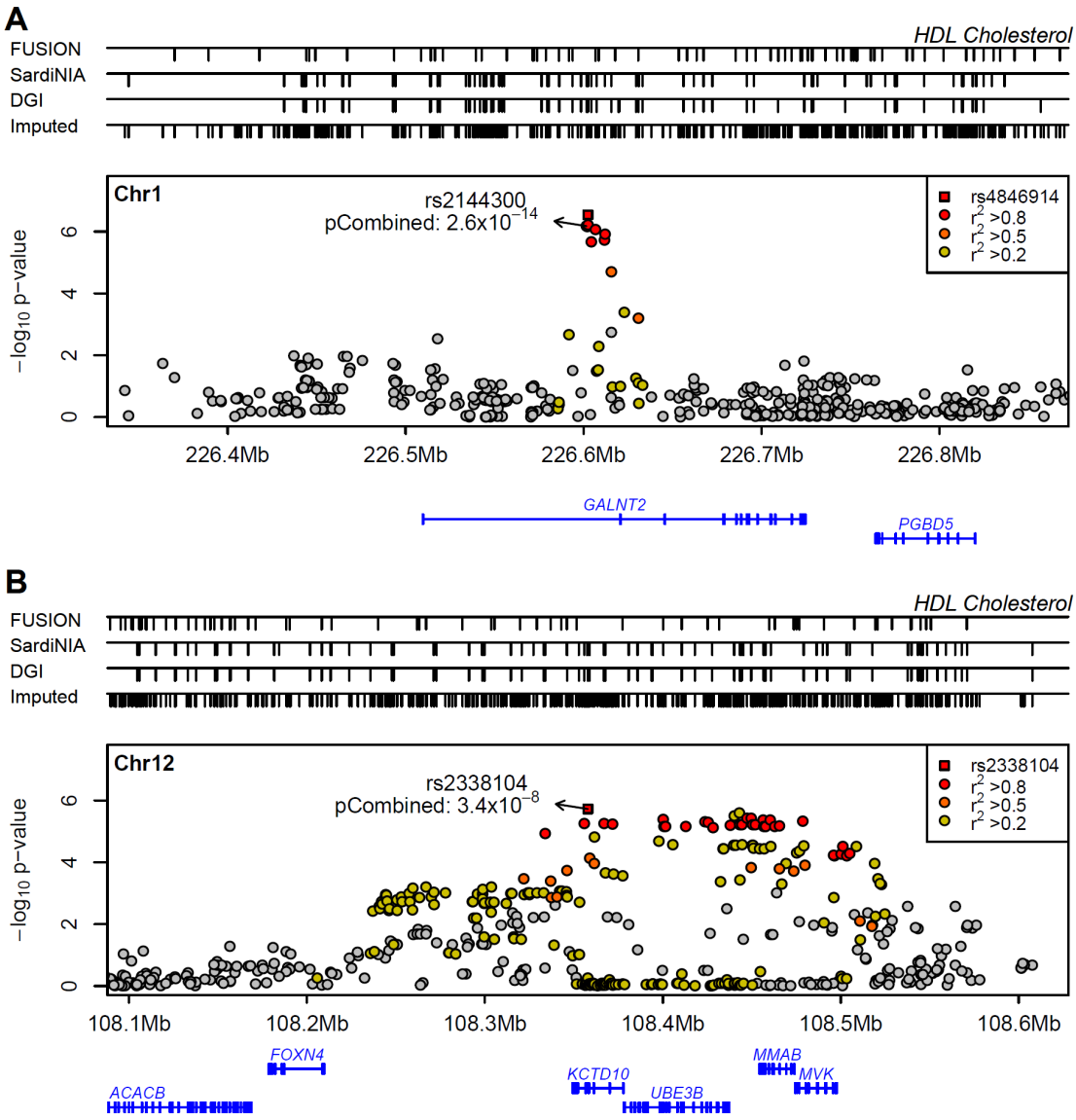


Figura7C, D

Riassunto dei nuovi loci trovati in associazione con i tratti del profilo lipidico

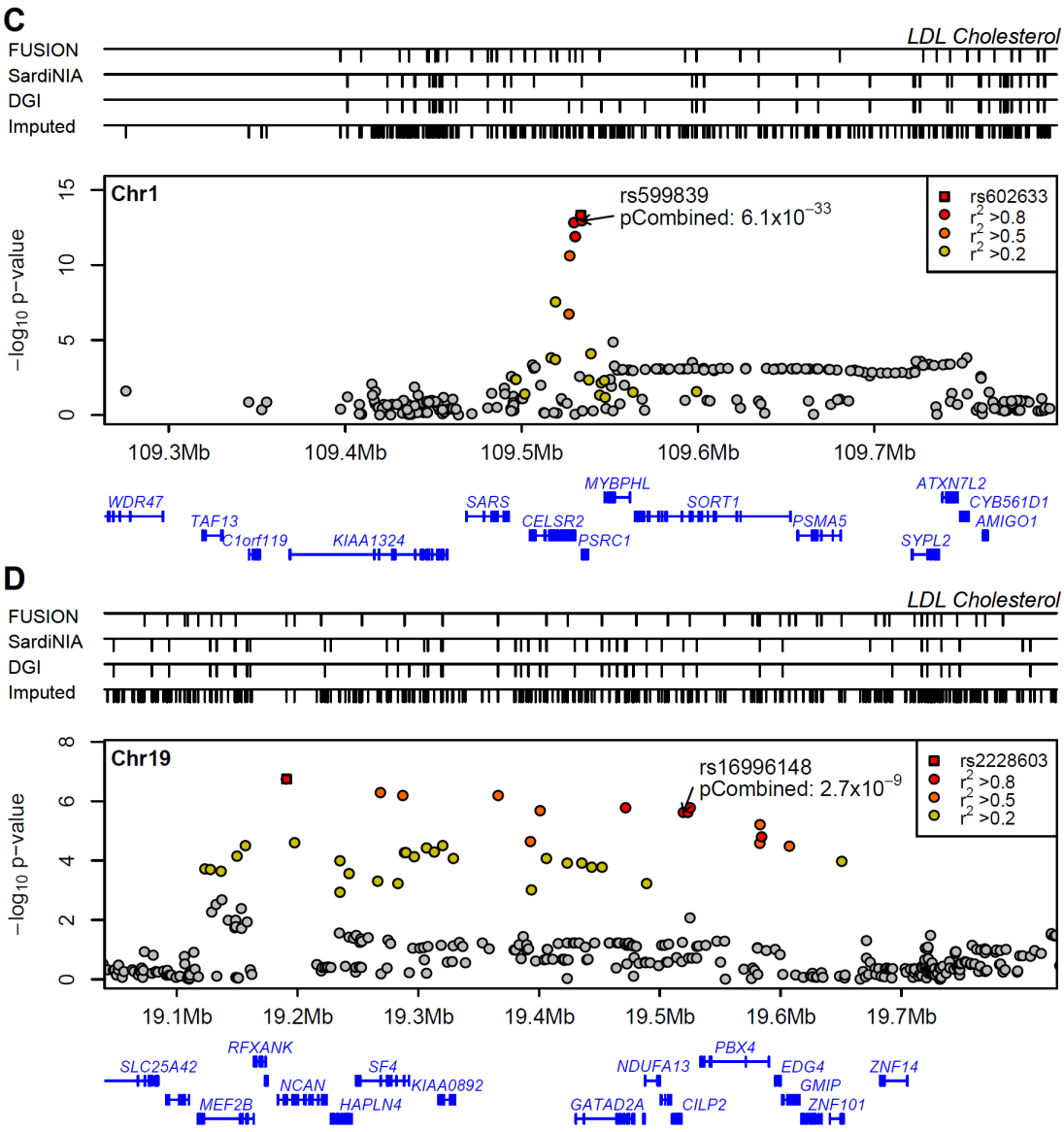


Figura 7E, F

Riassunto dei nuovi loci trovati in associazione con i tratti del profilo lipidico.

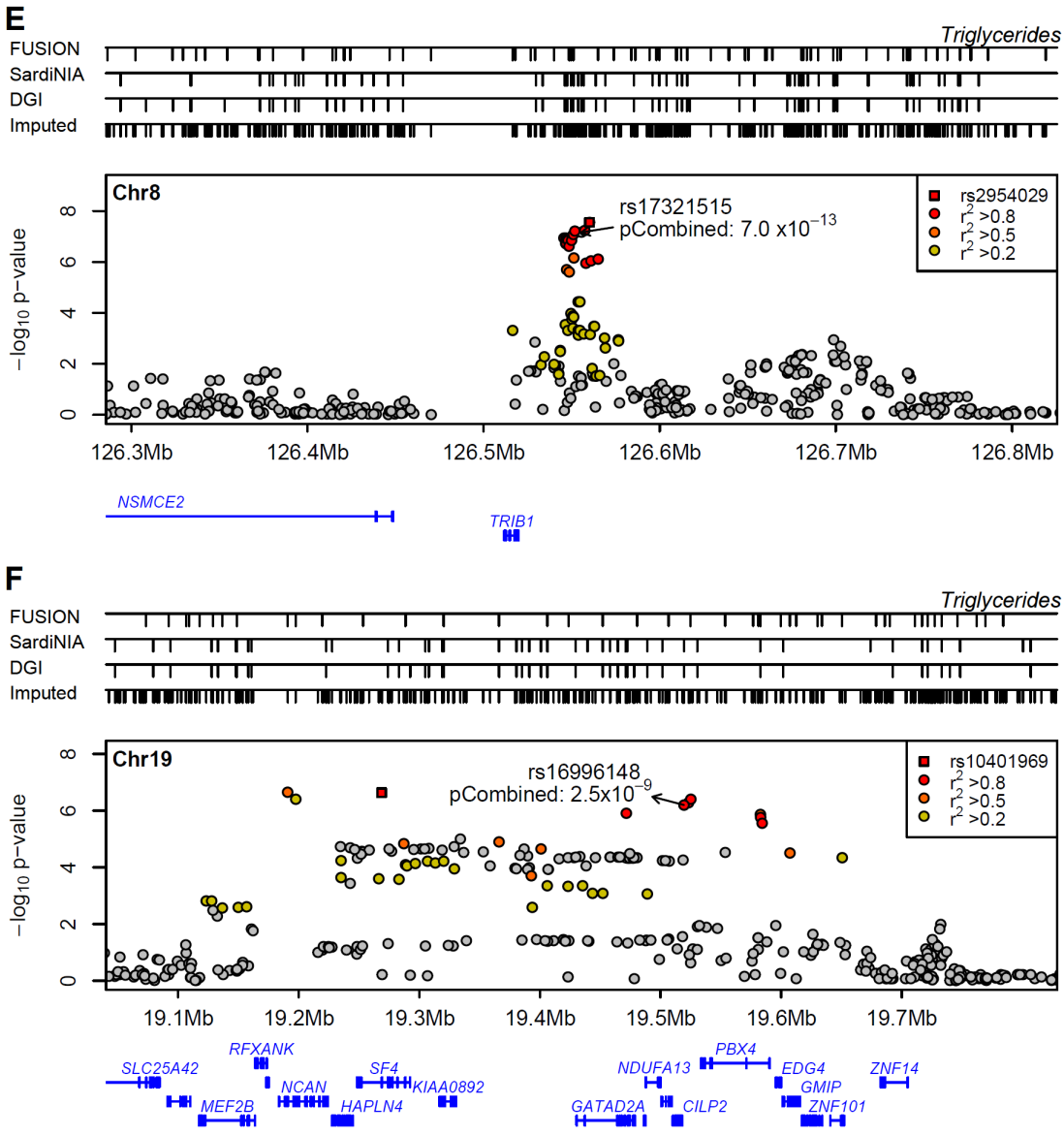


Figura 7G, H

Riassunto dei nuovi loci trovati in associazione con i tratti del profilo lipidico.

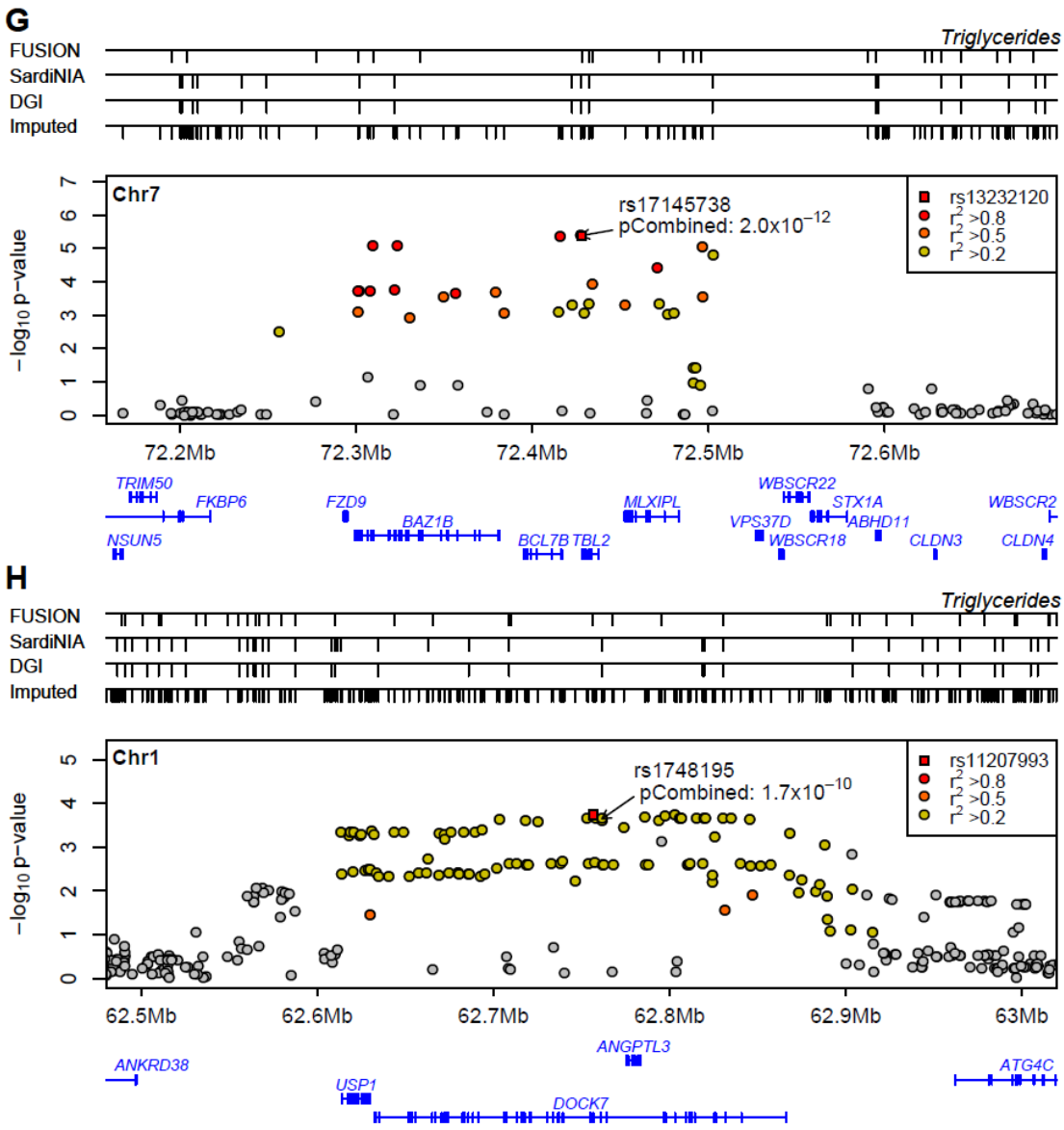


Tabella 1
Tratti quantitativi, con rispettive unità di misura, valutati nella prima fase del progetto ProgeNIA

Parametri ematologici	**unità**	Bilirubina totale	mg/dl
RBC, globuli rossi	10^6/µl	Acido urico	mg/dl
Hb, emoglobina	g/dl	Sodio	mEq/l
MCV (vol. RBC medio)	fl	Potassio	mEq/l
MCH (Hb RBC medio)	Pg	ESR, velocità sedimentazione eritrociti	mm/h
WBC, globuli bianchi	10^3/µl	CRP, proteina C-reattiva	mg/dl
NE, neutrofili	%	TSH, ormone stimolante la tiroide	microU/l
LY, linfociti	%	PSA, antigene prostatico specifico	---
MO, monociti	%	**Misurazioni antropometriche**	**unità**
EO, eosinofili	%	Altezza	Cm
BA, basofili	%	Peso	Kg
PLT, piastrine	10^3/µl	Waist, circonferenza ai fianchi	cm
HbF, emoglobina fetale	%	Rapporto Waist/Hip	---
HbA_2	%	BMI, indice di massa corporea	Kg/m^2
HbA_{1C}	%	**Funzione cardiovascolare**	**unità**
G6PD	UI/dl	Pressione sanguigna sistolica	mmHg
Glucosio	mg/dl	Pressione sanguigna diastolica	mmHg
Insulina	µg/Uml	HR, frequenza cardiaca	Battiti/min
Azotemia	mg/dl	Diametro in sistole	mm
Creatinina	mg/dl	Diametro in diastole	mm
ALT	U/L	IMT, Spessore intimo-mediale	mm
AST	U/L	PWV, velocità dell'onda pulsante	cm/s
γGT	U/L	Variabili derivate (5)	
Fibrinogeno	mg/dl	Variabili elettrocardiografiche (2)	
Colesterolo totale	mg/dl	Variabili ecografiche (6)	
Colesterolo HDL	mg/dl	**Tratti psicologici**	**unità**
Colesterolo LDL	mg/dl	NEO N, Neuroticism (N1-6)	
Trigliceridi	mg/dl	NEO E, Extraversion (E1-6)	
Ferro	µg/dl	NEO O, Openess to exprerience (O1-6)	
Transferrina	mg/dl	NEO A, Agreeableness (A1-6)	
Bilirubina frazionata	mg/dl	NEO C, Conscientiousness (C1-6)	

Tabella 2
Replica dei segnali di associazione più significativi
Con l'ausilio del chip personalizzato Affymetrix, é stato genotipizzato un gruppo indipendente di 521 volontari non imparentati. Lo SNP rs9389269 é stato utilizzato come proxy nel genotyping di *follow-up* (HapMap r^2=1.0), poichè non é stato possible includere l'rs4895441 nel chip stesso. È stato fatto il *follow-up* per l'rs6600143, un altro SNP fortemente associato nella regione.Per lo SNP rs6037828, a causa di problemi tecnici nella preparazione delle sonde per il chip, non abbiamo dati di *follow-up*.

Locus	SNP	Chr	Position, bp	Trait	GWA *P* value	Replication *P* value
HBB	rs4910742	11	5263085	RBC	5.80×10^{-47}	4.60×10^{-17}
				HbA_2	3.00×10^{-46}	1.60×10^{-22}
				MCV	2.50×10^{-56}	3.80×10^{-16}
				MCH*	2.80×10^{-54}	2.20×10^{-15}
				Hb	2.30×10^{-39}	4.10×10^{-13}
				MCHC*	1.20×10^{-31}	4.90×10^{-13}
				HbF	1.80×10^{-21}	3.40×10^{-12}
BCL11A	rs11886868	2	60631897	HbF	6.70×10^{-35}	8.50×10^{-10}
MYB/HBS1L	rs4895441	6	135468266	HbF	1.20×10^{-16}	0.00002†
HBA	rs6600143‡	16	141389	MCH*	4.10×10^{-7}	0.002
				MCV	3.00×10^{-6}	0.002
				Hb	0.0012	0.04
				RBC	0.0003	0.04
				MCHC*	0.0002	0.02
				HbA_2	0.3	0.4
CSNK2A1	rs6037828	20	437009	HbF	1.10×10^{-9}	NA§

Tabella 3
Distribuzione dei genotipi dello SNP rs11886868 in individui HPFH, in pazienti talassemici (intermedi e major) e nella popolazione generale del progetto SardiNIA. A sinistra, sono indicate le frequenze genotipiche per ciascun gruppo. A destra, il P value per i test allelici e genotipici.

		Frequenze			**P *value* test allelico**			
	N individui	C/C	C/T	T/T	HPFH	Talassemia Intermedia	Talassemia Major	SardiNIA
HPFH	66	0.227	0.546	0.227	---	0.987	$3.29x10^{-7}$	$2.15x10^{-16}$
Talassemia intermedia	52	0.192	0.597	0.211	0.847	---	$2.91x10^{-6}$	$1.23x10^{-12}$
Talassemia Major	74	0.040	0.355	0.635	$1.72x10^{-6}$	$6.49x10^{-6}$	---	0.963
SardiNIA	1.412	0.040	0.316	0.644	$8.52x10^{-13}$	$3.16x10^{-12}$	0.969	---

Tabella 4
Caratteristiche dei campioni utilizzati nelle analisi *genome-wide* e di *follow-up*

Samples	Phenotyped individuals[a] (% female)	Demographics			Median trait concentrations (quartile ranges)		
		Geographic origin	Median age (quartile range)	Median BMI (quartile range)	HDL-C (mg/dl)	LDL-C (mg/dl)	Triglycerides (mg/dl)
Genome-wide analyses (n = 8,816)							
FUSION							
Type 2 diabetics	773 (41%)	Finland	63.0 (11.1)	29.8 (6.1)	44.9 (15.9)	135.6 (44.5)	150.6 (106.3)
Controls	1,101 (48%)	Finland	62.0 (14.5)	26.6 (5.0)	54.6 (21.7)	141.1 (44.9)	103.7 (60.2)
SardiNIA	4,184 (57%)	Sardinia (in Italy)	42.4 (28.0)	24.9 (6.4)	62.7 (18.6)	124.6 (47.6)	70.0 (54.0)
DGI	2,758 (51%)	Finland, Sweden	62.8 (15.5)	27.3 (5.4)	46.2 (15.9)	148.3 (51.8)	121.7 (81.9)
Follow-up samples (n = 11,569)							
FUSION							
Type 2 diabetics	970 (41%)	Finland	60.0 (11.0)	30.2 (6.5)	49.1 (17.0)	123.5 (51.6)	139.1 (90.4)
Controls	1,249 (39%)	Finland	59.0 (10.5)	26.4 (4.9)	56.1 (21.3)	138.4 (46.1)	103.2 (55.8)
ISIS							
Myocardial infarction survivors	1,254 (28%)	United Kingdom	52.0 (14.0)	26.0 (6.0)	40.6 (12.4)	144.0 (48.4)	n/a
Controls	1,252 (35%)	United Kingdom	48.0 (14.0)	24.0 (5.0)	49.9 (16.3)	124.2 (41.4)	132.0 (102.8)
HAPI	861 (46%)	United States	43.0 (22.0)	25.9 (5.9)	55.8 (18.0)	139.1 (56.0)	68.5 (38.0)
SUVIMAX	1,551 (62%)	France	50.0 (9.0)	23.3 (4.1)	61.9 (21.9)	135.8 (41.4)	80.0 (41.6)
BWHHS	3,358 (100%)	United Kingdom	69.0 (9.0)	26.9 (6.1)	61.9 (23.2)	158.3 (54.2)	141.8 (90.4)
Caerphilly	1,074 (0%)	United Kingdom	57.0 (8.0)	26.1 (4.1)	51.5 (17.0)	142.3 (54.3)	132.9 (102.8)

Tabella 5 Riassunto dei risultati della meta-analisi dei GWAS nello *stage*1 (sono inclusi tutti i segnali con P<5x10^{-7})

La tabella riassume i segnali di associazione osservati nell'analisi della concentrazione dei lipidi nei tre GWAS. L'attribuzione del cromosoma, la posizione e l'annotazione genica si riferiscono al Genome Build (UCSC), del Marzo 2006. Gli alleli sono ordinati in maniera tale che il primo allele (+) é associato con un aumentato livello dei lipidi.
Per ciascun locus, é indicato lo SNP associato in maniera più significativa, insieme con la sua posizione relativa rispetto ai geni vicini, in maniera tale da poter fare delle correlazioni rispetto a geni precedentemente implicati nel metabolismo lipidico.

Locus			Association signal				Corroborating signals ($P < 10^{-6}$)		Nearby genes
SNP	Chr	Position (Mb)	Allele (+/−)	Freq (+)	Effect (mg/dl)	*P* value	SNPs	LD groups ($r^2 < 0.2$)	(Relative position) (−upstream, +downstream)
HDL cholesterol (n = 8,656)									
rs3764261	16	55.6	A/C	0.29	2.42	$\mathbf{2.8 \times 10^{-19}}$	14	2	*CETP* (–2.4 kb)
rs12678919	8	19.9	G/A	0.12	2.44	$\mathbf{1.3 \times 10^{-11}}$	84	2	*LPL* (+19.5 kb)
rs10468017	15	56.5	T/C	0.32	1.76	$\mathbf{8.6 \times 10^{-11}}$	18	2	*LIPC* (–45.7 kb)
rs1323432	9	101.4	A/G	0.87	1.93	$\mathbf{2.5 \times 10^{-8}}$	4	1	*GRIN3A* (Intron 6); *PPP3R2* (–5.7 kb)
rs4149274	9	104.7	G/A	0.69	1.51	7.4×10^{-8}	20	1	*ABCA1* (Intron 5)
rs4939883	18	45.4	C/T	0.86	1.87	1.4×10^{-7}	2	1	*LIPG* (+47.9 kb)
rs4846914	1	226.6	A/G	0.62	1.15	2.9×10^{-7}	4	1	*GALNT2* (Intron 1)
LDL cholesterol (n = 8,589)									
rs4420638	19	50.1	G/A	0.16	8.02	$\mathbf{3.2 \times 10^{-21}}$	2	1	*APOE/APOC* cluster
rs515135	2	21.2	C/T	0.83	6.08	$\mathbf{3.1 \times 10^{-14}}$	116	3	*APOB* (−19.1kb)
rs602633	1	109.5	G/T	0.80	6.09	$\mathbf{4.8 \times 10^{-14}}$	8	1	*CELSR2* (+3.1kb); *PSRC1* (+668 bp); *SORT1* (−30 kb)
rs6511720	19	11.1	C/A	0.91	8.03	$\mathbf{6.8 \times 10^{-10}}$	1	1	*LDLR* (Intron 1)
rs2228603	19	19.2	C/T	0.93	6.46	1.8×10^{-7}	3	1	*NCAN* (Pro92Ser)
Triglycerides (n = 8,684)									
rs964184	11	116.2	G/C	0.12	18.12	$\mathbf{1.5 \times 10^{-16}}$	29	2	*APOA5* (+11.2 kb)
rs1260326	2	27.6	T/C	0.40	10.25	$\mathbf{1.5 \times 10^{-15}}$	52	2	*GCKR* (Leu446Pro)
rs6993414	8	19.9	A/G	0.46	14.20	$\mathbf{1.4 \times 10^{-13}}$	85	2	*LPL* (+78.1 kb)
rs2954029	8	126.6	A/T	0.56	6.42	$\mathbf{2.8 \times 10^{-8}}$	15	1	*TRIB1* (+40.3 kb)
rs10401969	19	19.3	T/C	0.92	12.28	2.3×10^{-7}	5	1	*NCAN* (+44.7 kb); *SF4* (Intron 8)

Tabella 6 Scoperte più significative degli *stages* 1 e 2

La tabella riassume i segnali di associazione dopo *follow-up* degli SNPs più promettenti nel campione *stage*2. Gli SNPs con un P*value* combinato (stage 1 + 2) $<10^{-5}$ sono stati inclusi, sebbene è indicato in tabella anche *GRIN3A*, per completezza, dal momento che era significativo nello studio iniziale. Le righe corrispondenti agli SNPs con un P*value* combinato di $5x10^{-8}$ sono in grassetto. Gli SNPs in questa tabella possono non corrispondere a quelli nella tabella 5, in quanto quest'ultima mostra solamente i segnali più forti a ciascun locus.

						Association *P* values			Sample sizes		
SNP	Chr	Pos(Mb)	Alleles (+/−)	Freq (+)	Effect (mg/dl)	Stage 1 (two-sided)	Stage 2 (one-sided)	Combined (two-sided)	Stage 1	Stage 2	Nearby genes
SNPs associated with HDL cholesterol											
rs3764261	**16**	**55.6**	**A/C**	**0.69**	**3.47**	**2.8×10^{-19}**	**6.4×10^{-43}**	**2.3×10^{-57}**	**8,656**	**8,072**	***CETP***
rs1864163	**16**	**55.6**	**G/A**	**0.80**	**4.12**	**3.0×10^{-17}**	**4.3×10^{-28}**	**6.9×10^{-39}**	**8,656**	**3,684**	***CETP***
rs9989419	**16**	**55.5**	**G/A**	**0.65**	**1.72**	**8.0×10^{-16}**	**1.8×10^{-17}**	**3.2×10^{-31}**	**8,656**	**6,981**	***CETP***
rs12596776	**16**	**55.5**	**G/C**	**0.13**	**1.26**	**3.7×10^{-5}**	**1.0×10^{-4}**	**2.8×10^{-8}**	**8,656**	**7,030**	***CETP***
rs1566439	**16**	**55.6**	**C/T**	**0.45**	**0.96**	**2.0×10^{-5}**	**2.1×10^{-4}**	**3.3×10^{-8}**	**8,656**	**4,881**	***CETP***
rs4775041	**15**	**56.5**	**C/G**	**0.67**	**1.38**	**2.8×10^{-9}**	**9.6×10^{-13}**	**3.2×10^{-20}**	**8,656**	**11,426**	***LIPC***
rs261332	**15**	**56.5**	**A/G**	**0.19**	**1.41**	**1.7×10^{-9}**	**1.3×10^{-7}**	**2.3×10^{-15}**	**8,656**	**6,956**	***LIPC***
rs10503669	**8**	**19.9**	**A/C**	**0.10**	**2.09**	**3.2×10^{-10}**	**9.4×10^{-11}**	**4.1×10^{-19}**	**8,656**	**11,431**	***LPL***
rs2197089	**8**	**19.9**	**A/G**	**0.42**	**1.38**	**3.4×10^{-8}**	**3.2×10^{-5}**	**1.0×10^{-11}**	**8,656**	**3,644**	***LPL***
rs6586891	**8**	**20**	**A/C**	**0.34**	**1.00**	**3.5×10^{-5}**	**9.7×10^{-6}**	**2.9×10^{-9}**	**8,656**	**7,017**	***LPL***
rs2144300	**1**	**226.6**	**T/C**	**0.40**	**1.11**	**6.6×10^{-7}**	**4.0×10^{-9}**	**2.6×10^{-14}**	**8,656**	**11,406**	***GALNT2***
rs2156552	**18**	**45.4**	**T/A**	**0.84**	**1.20**	**8.4×10^{-7}**	**7.1×10^{-7}**	**6.4×10^{-12}**	**8,656**	**11,437**	***LIPG***
rs4149268	**9**	**104.7**	**C/T**	**0.355**	**0.82**	**3.3×10^{-7}**	**2.2×10^{-5}**	**1.2×10^{-10}**	**8,656**	**11,327**	***ABCA1***
rs2338104	**12**	**108.4**	**G/C**	**0.45**	**0.48**	**1.9×10^{-6}**	**7.6×10^{-4}**	**3.4×10^{-8}**	**8,656**	**11,399**	***MVK/MMAB***
rs255052	16	66.6	A/G	0.17	0.74	1.5×10^{-6}	0.0087	1.2×10^{-7}	8,656	4,534	*LCAT*
rs1323432	9	101.4	A/G	0.88	−0.03	2.5×10^{-8}	0.82	7.7×10^{-4}	8,656	8,176	*GRIN3A*
SNPs associated with LDL cholesterol											
rs4420638	**19**	**50.1**	**G/A**	**0.82**	**6.61**	**3.2×10^{-21}**	**4.9×10^{-24}**	**3.0×10^{-43}**	**8,589**	**10,806**	***APOE/C1/C4***
rs10402271	**19**	**50**	**G/T**	**0.67**	**2.62**	**9.8×10^{-6}**	**1.5×10^{-5}**	**1.2×10^{-9}**	**8,589**	**6,519**	***APOE/C1/C4***
rs599839	**1**	**109.5**	**A/G**	**0.77**	**5.48**	**1.2×10^{-13}**	**2.7×10^{-21}**	**6.1×10^{-33}**	**8,589**	**10,783**	***CELSR2/PSRC1/SORT1***
rs6511720	**19**	**11.1**	**G/T**	**0.90**	**9.17**	**6.8×10^{-10}**	**3.3×10^{-19}**	**4.2×10^{-26}**	**8,589**	**7,442**	***LDLR***
rs562338	**2**	**21.2**	**G/A**	**0.18**	**4.89**	**1.2×10^{-11}**	**3.6×10^{-12}**	**5.6×10^{-22}**	**8,589**	**10,849**	***APOB***
rs754523	**2**	**21.2**	**G/A**	**0.28**	**2.78**	**7.0×10^{-7}**	**1.3×10^{-6}**	**8.3×10^{-12}**	**8,589**	**6,542**	***APOB***
rs693	**2**	**21.1**	**A/G**	**0.42**	**2.44**	**1.2×10^{-7}**	**0.0034**	**3.1×10^{-9}**	**8,589**	**3,222**	***APOB***
rs11206510	**1**	**55.2**	**T/C**	**0.81**	**3.04**	**7.5×10^{-6}**	**5.4×10^{-7}**	**3.5×10^{-11}**	**8,589**	**10,805**	***PCSK9***
rs16996148	**19**	**19.5**	**G/T**	**0.89**	**3.32**	**2.4×10^{-6}**	**8.3×10^{-5}**	**2.7×10^{-9}**	**8,589**	**10,841**	***NCAN/CILP2***
rs2254287	6	33.3	G/C	0.38	1.91	2.9×10^{-6}	0.0015	5.1×10^{-8}	8,589	7,440	*B3GALT4*
rs12695382	3	120.4	A/G	0.90	2.23	4.9×10^{-6}	0.0067	1.0×10^{-6}	8,589	10,802	*B4GALT4*
SNPs associated with triglycerides											
rs780094	**2**	**27.7**	**T/C**	**0.39**	**8.59**	**1.7×10^{-14}**	**2.0×10^{-19}**	**6.1×10^{-32}**	**8,684**	**9,723**	***GCKR***
rs11127129	2	28.0	C/G	0.79	3.77	2.0×10^{-4}	3.2×10^{-4}	4.7×10^{-7}	8,684	9,700	*RBKS/GCKR*
rs12286037	**11**	**116.2**	**T/C**	**0.94**	**25.82**	**1.1×10^{-7}**	**1.6×10^{-22}**	**1.0×10^{-26}**	**8,684**	**9,738**	***APOA5/A4/C3/A1***
rs662799	**11**	**116.2**	**G/A**	**0.05**	**16.88**	**4.3×10^{-8}**	**2.7×10^{-10}**	**2.4×10^{-15}**	**8,684**	**3,248**	***APOA5/A4/C3/A1***
rs2000571	11	116.1	A/G	0.17	6.93	4.7×10^{-5}	8.7×10^{-5}	5.7×10^{-8}	8,684	3,209	*APOA5/A4/C3/A1*
rs486394	11	116.0	C/A	0.28	1.50	1.7×10^{-4}	0.0073	7.4×10^{-6}	8,684	3,597	*APOA5/A4/C3/A1*
rs10503669	**8**	**19.9**	**C/A**	**0.895**	**11.57**	**1.4×10^{-9}**	**1.6×10^{-14}**	**3.9×10^{-22}**	**8,684**	**9,711**	***LPL***
rs2197089	**8**	**19.9**	**G/A**	**0.58**	**3.38**	**3.1×10^{-11}**	**0.0029**	**1.1×10^{-12}**	**8,684**	**3,202**	***LPL***
rs6586891	8	20.0	C/A	0.66	4.60	2.4×10^{-4}	5.0×10^{-4}	1.1×10^{-6}	8,684	3,622	*LPL*
rs17321515	**8**	**126.6**	**A/G**	**0.56**	**6.42**	**6.8×10^{-8}**	**1.0×10^{-6}**	**7.0×10^{-13}**	**8,684**	**5,312**	***TRIB1***
rs17145738	**7**	**72.4**	**C/T**	**0.84**	**8.21**	**4.1×10^{-6}**	**5.0×10^{-8}**	**2.0×10^{-12}**	**8,684**	**9,741**	***MLXIPL***
rs1748195	**1**	**62.8**	**C/G**	**0.70**	**7.12**	**2.3×10^{-4}**	**5.4×10^{-8}**	**1.7×10^{-10}**	**8,684**	**9,559**	***ANGPTL3***
rs16996148	**19**	**19.5**	**G/T**	**0.92**	**6.10**	**6.3×10^{-7}**	**2.4×10^{-4}**	**2.5×10^{-9}**	**8,684**	**9,707**	***NCAN/CILP2***
rs4775041	**15**	**56.5**	**C/G**	**0.67**	**3.62**	**7.3×10^{-5}**	**2.9×10^{-5}**	**1.6×10^{-8}**	**8,684**	**8,462**	***LIPC***
rs2144300	1	226.6	C/T	0.60	4.25	4.9×10^{-4}	2.4×10^{-4}	7.9×10^{-7}	8,684	8,473	*GALNT2*

Tabella 7 Correlazione tra malattia coronarica (CAD) e SNPs associati ai livelli di colesterolo LDL

Associazione tra malattia coronarica e gli alleli associati con la concentrazione del colesterolo LDL nel nostro studio. L'evidenza per l'associazione é stata testata nel pannello del Wellcome Trust Case Control Consortium, senza che venissero fatti aggiustamenti per covariate addizionali. Le righe corrispondenti agli SNPs che mostrano associazione con il colesterolo LDL con $P<5x10^{-8}$ nel nostro campione sono in grassetto.

					Association with coronary artery disease (WTCCC)						
Locus			LDL-C association (current study)		Expanded reference set		CAD cases				
SNP	Chr	Position (Mb)	Alleles (+/−)	*P* value (two-sided)	*n*	Frequency of LDL+ allele	*n*	Frequency of LDL+ allele	*P* value (one sided)	OR (95% CI)	Nearby genes
rs4420638	19	50.1	G/A	$\mathbf{3.0 \times 10^{-43}}$	12,281	0.184	1,926	0.209	1.0×10^{-4}	1.17 (1.08–1.28)	***APOE/C1/C4***
rs10402271	19	50.0	G/T	$\mathbf{1.2 \times 10^{-9}}$	12,256	0.319	1,921	0.339	0.0068	1.10 (1.02–1.18)	***APOE/C1/C4***
rs599839	1	109.5	A/G	$\mathbf{6.1 \times 10^{-33}}$	12,292	0.778	1,923	0.808	1.3×10^{-5}	1.20 (1.10–1.31)	***PSRC1/SORT1***
rs6511720[a]	19	11.1	G/T	$\mathbf{4.2 \times 10^{-26}}$	12,301	0.890	1,926	0.902	6.7×10^{-4}	1.29 (1.10–1.52)	***LDLR***
rs562338	2	21.2	G/A	$\mathbf{5.6 \times 10^{-22}}$	12,288	0.824	1,924	0.830	0.18	1.04 (0.95–1.14)	***APOB***
rs754523	2	21.2	G/A	$\mathbf{8.3 \times 10^{-12}}$	12,292	0.332	1,926	0.353	0.0042	1.10 (1.03–1.18)	***APOB***
rs693	2	21.1	A/G	$\mathbf{3.1 \times 10^{-9}}$	12,292	0.520	1,924	0.536	0.028	1.07 (1.00–1.14)	***APOB***
rs11206510	1	55.2	T/C	$\mathbf{3.5 \times 10^{-11}}$	12,284	0.807	1,925	0.825	0.0042	1.13 (1.03–1.23)	***PCSK9***
rs16996148	19	19.5	G/T	$\mathbf{2.7 \times 10^{-9}}$	12,182	0.915	1,921	0.922	0.055	1.11 (0.98–1.26)	***NCAN/CILP2***
rs2254287[a]	6	33.3	G/C	5.1×10^{-8}	12,301	0.385	1,926	0.399	0.039	1.07 (0.99–1.14)	*B3GALT4*
rs12695382	3	120.4	A/G	1.0×10^{-6}	12,292	0.865	1,924	0.874	0.051	1.09 (0.98–1.20)	*B4GALT4*

Tabella 8 Criteri di scelta del tipo di sequenziamento da eseguire
Sono stati generati *dataset* di 67 e 1.000 individui diploidi e sono stati, quindi, simulati i dati delle sequenze *shotgun*. La tabella fornisce la proporzione dei siti simulati rilevati, al variare della frequenza dell'allele minore (MAF), l'accuratezza delle chiamate genotipiche a ciascun sito, e la correlazione (r^2) tra genotipi simulati e inferiti. L'effettiva dimensione campionaria, per le successive analisi di associazione, può essere calcolata dal prodotto $n \cdot r^2$.

Sequenziamento di 67 individui al coverage 30X

Frequenza dell'allele minore	***0.5-1.0%***	***1.0-2.0%***	***2.0-5.0%***	***>5%***
Proporzione di siti rilevati	59.3%	90.1%	96.9%	100.0%
Accuratezza genotipizzazione	100.0%	100.0%	100.0%	100.0%
…...solo siti eterozigoti	100.0%	100.0%	100.0%	100.0%
Correlazione con il valore vero (r^2)	99.8%	99.9%	99.9%	100.0%
Effettiva dimensione campione ($n \cdot r^2$)	67	67	67	67

Sequenziamento di 1.000 individui al coverage 2X

Frequenza dell'allele minore	***0.5-1.0%***	***1.0- 2.0%***	***2.0-5.0%***	***>5%***
Proporzione di siti rilevati	79.6%	98.8%	100.0%	100.0%
Accuratezza genotipizzazione	99.6%	99.5%	99.5%	99.8%
…...solo siti eterozigoti	78.8%	89.5%	95.9%	99.8%
Correlazione con il valore vero (r^2)	56.7%	76.1%	88.2%	97.8%
Effettiva dimensione campione ($n \cdot r^2$)	567	761	882	978

Tabella 9 Dati preliminari dello studio del profilo lipidico mediante *whole-genome sequencing* di 226 volontari SardiNIA

Dati preliminari sulle sequenze di 226 volontari SardiNIA. La tabella mostra il numero totale di SNPs individuati, quanti di questi risultano validati (PASS) e quanti non superano invece i filtri di qualità (FAIL) quali: FFRQ30 (indica gli SNPs che mappano in regioni contenenti *repeat*, e quindi probabilmente non veri), INDEL10 (indica SNPs che mappano a 10 basi di distanza da indels e per i quali, quindi, la determinazione può essere non accurata) e r0.7p0.9 (indica quegli SNPs scartati in quanto con bassa qualità di imputazione).
È inoltre indicato il numero e la percentuale degli SNPs presenti in dbSNP, la percentuale di SNPs in HapMap trovati nelle nostre sequenze (%HM3) ed il rapporto Ts/Tv, assunto comq fattore di qualità, e che per dati *whole-genome* deve essere circa 2.

FILTER	#SNPs	#dbSNP	%dbSNP	Known	Novel	Overall	%HM3
				Ts/Tv ratio			*sites*
FFRQ30	1.367.210	286.422	20.9	0.78	0.75	0.76	0.10
INDEL10	1.063.008	323.186	30.4	0.86	0.84	0.85	0.49
PASS	10.759.492	5.845.709	54.3	2.16	1.90	2.04	94
r0.7p0.9	1.517.717	372.598	24.5	1.10	0.74	0.82	1.127
PASS	10.759.492	5.845.709	54.3	2.16	1.90	2.04	94.135
FAIL	2.834.332	695.493	24.5	0.97	0.77	0.81	1.652
TOTAL	13.593.824	6.541.202	48.1	1.97	1.43	1.66	95.788

REFERENZE BIBLIOGRAFICHE

1. Risch N.J. **Searching for genetic determinants in the new millenium**. *Nature 2000; 405: 847-856.*
2. Li Y., Willer C., Sanna S. and Abecasis G.R. **Genotype imputation**. *Ann. Rev. Genomics Hum. Genetics 2009; 10: 387-406.*
3. Peltonen L. *et al.* **Use of population isolates for mapping complex traits.** *Nature Genetics 2000; 1(3): 182-190.*
4. P. Heutink and B.A. Oostra. **Gene finding in genetically isolated populations.** *Human Molecular Genetics 2002; 11(20): 2507-2515.*
5. G. Pilia, Wei-Min Chen, A. Scuteri, M. Orrù, G. Albai, M. Dei, S. Lai, G. Usala, M. Lai, P. Loi, C. Mameli, L. Vacca, M. Deiana, N. Olla, M. Masala, A. Cao, S.S. Najjar, A. Terracciano, T. Nedorezov, A. Sharov, A.B. Zonderman, G.R. Abecasis, P. Costa, E. Lakatta, D. Schlessinger. **Heritability of cardiovascular and personality traits in 6,148 Sardinians.** *PLoS Genet. 2006;2(8): 1207-1223.*
6. Scuteri A., S. Sanna, W.-M. Chen, M. Uda, G. Albai, J. Strait, S. Najjar, R. Nagaraja, M. Orrù, G. Usala, M. Dei, S. Lai, A. Maschio, F. Busonero, A. Mulas, G.B. Ehret, A.A. Fink, A.B. Weder, R.S. Cooper, P. Galan, A. Chakravarti, D. Schlessinger, A. Cao, E. Lakatta, G.R. Abecasis. **Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits.** *PLoS Genetics 2007; 7: 1220-1210.*
7. Chen W-M., Abecasis G.R. **Family based association tests for genome-wide association scans.** *Am. J. Human Genetics 2007; 81: 913-926.*
8. Matsuzaki H. *et al.* **Genotyping over 100.000 SNPs on a pair of oligonucleotide arrays.** *Nature Methods 2004 Nov; 1(2): 109-11.*
9. Hardenbol P. *Et al.* **Multiplexed genotyping with sequence-tagged molecular inversion probes.** *Nature Biotechnology 2003; 21(6): 673-8.*
10. Livak K.J. *et al.* **Oligonucleotides with fluorescent dyes at opposite ends provide a quenced probe system useful for detecting PCR product and nucleic acid hybridization.** *PCR Methods Appl. 1995; 4: 357-362.*
11. Quail M.A. *et al.***A large genome center's improvements to the Illumina sequencing system.** *Nature Methods 2008; 5(12): 1005-1010.*

12. C.J. Willer, S. Sanna, A.U. Jackson, A. Scuteri, L.L. Bonnycastle, R. Clarke, S.C. Heath, N.J Timpson, S.S Najjar, H.M. Stringham, J. Strait, W.L. Duren, A. Maschio, F. Busonero, A. Mulas, G. Albai, A.J. Swift, M.A. Morken, N. Narisu, D. Bennett, S. Parish, H. Shen, P. Galan, P. Meneton, S. Hercberg, D. Zelenika, W-M. Chen, Y. Li, L.J. Scott, P.A. Scheet, J. Sundvall, R.M. Watanabe, R. Nagaraja, S. Ebrahim, D.A. Lawlor, Y.B. Shlomo, G.D. Smith, A.R. Shuldiner, R. Collins, R.N. Bergman, M. Uda, J. Tuomilehto, A. Cao, F.S. Collins, E. Lakatta, G.M. Lathrop, M. Boehnke, D. Schlessinger, K.L. Mohlke & G.R. Abecasis. **Newly identified loci that influence lipid concentrations and risk of coronary artery disease.** *Nature Genetics 2008 February; 40 (2): 161-9.*
13. Speliotes E.K. *et al*. **Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index.** *Nature Genetics 2010; 42: 937-948.*
14. S. Li, S. Sanna, A. Maschio, F. Busonero, G. Usala, A. Mulas, S. Lai, M. Dei, M. Orrù, G. Albai, S. Bandinelli, D. Schlessinger, E. Lakatta, A. Scuteri, S.S. Najjar, J. Guralnik, S. Naitza, L. Crisponi, A. Cao, G.R. Abecasis, L. Ferrucci, M. Uda, W.-M. Chen, R. Nagaraja **The *GLUT9* gene is associated with serum uric acid levels in Sardinia and Chianti cohorts.** *PLoS Genetics 2007; 3(11): 1-7.*
15. M. Uda, R. Galanello, S. Sanna, G. Lettre, V.G. Sankaran, W.M. Chen, G. Usala, F. Busonero, A. Maschio, G. Albai, M.G. Piras, N. Sestu, S. Lai, M. Dei, A. Mulas, L. Crisponi, S. Naitza, I. Asunis, M. Deiana, R. Nagaraja, L. Perseu, S. Satta, M.D. Cipollina, C. Sollaino, P. Moi, J.N. Hirschhorn, S.H. Orkin, G.R. Abecasis, D. Schlessinger, A. Cao. **Genome-wide association study shows BCL11A associated with persistent fetal haemoglobin and amelioration of the phenotype of β-thalassemia.** *PNAS 2008 February 1; 5: 1620-1625.*
16. Arnaud-Lopez L., G. Usala, G. Ceresini, B.D. Mitchell, M.G. Pilia, M.G. Piras, N. Sestu, A. Maschio, F. Busonero, G. Albai, M. Dei, S. Lai, A. Mulas, L. Crisponi, T. Tanaka, S. Bandinelli, J.M. Guralnik, A. Loi, L. Balaci, G. Sole, A. Prinzis, Stefano M., A.R. Shuldiner, A. Cao, D. Schlessinger, M. Uda, G.R. Abecasis, R. Nagaraja, S. Sanna, and S. Naitza. **Phosphodiesterase 8B gene variants are associated with serum TSH**

levels and thyroid function. The American Journal of Human Genetics 2008; 82: 1270-1280.

17. Sanna S., F. Busonero, A. Maschio, P.F. McArdle, G. Usala, M. Dei, S. Lai, A. Mulas, M.G. Piras, L. Perseu, M. Masala, M. Marongiu, L. Crisponi, S. Naitza, R. Galanello, G.R. Abecasis, A.R. Shuldiner, D. Schlessinger, A. Cao and M. Uda. **Common variants in the *SLCO1B3* locus are associated with bilirubin levels and unconjugated hyperbilirubinemia.** *Human Molecular Genetics 2009; 18(14): 2711-2718.*
18. A. Terracciano, S. Sanna, M. Uda, B. Deiana, G. Usala, F. Busonero, A. Maschio, M. Scally, N. Patriciu, W-M. Chen, M.A. Distel, E.P. Slagboom, DI Boomsma, S. Villafuerte, E. Sliwerska, M. Burmeister, N. Amin, A.C.J.W. Janssens, C.M. van Duijn, D. Schlessinger, G.R. Abecasis and P.T. Costa Jr. **Genome-wide association scan for five major dimensions of personality.** *Molecular Psychiatry 2008; 15: 647-656.*
19. A. Terracciano, L. Balaci, J. Thayer, M. Scally, S. Kokinos, L. Ferrucci, T. Tanaka, A.B. Zonderman, S. Sanna, N. Olla, M.A. Zuncheddu, S. Naitza, F. Busonero, M. Uda, D. Schlessinger, G.R. Abecasis and P.T. Costa Jr. **Variants[Q1] of the serotonin transporter gene and NEO-PI-R neuroticism: no association in the BLSA and SardiNIA samples.** *Am. J. Med. Genet. 2009; Part B 9999:1-8.*
20. Weatherall D.J., Clegg J.B. **The Thalassemia syndromes**. *Blackwell scientific, Oxford 2001.*
21. Cao A., Galanello R., Rosatelli M.C. **Genotype-phenotype correlations in β-thalassemias**. *Blood Reviews 1994; 8: 1-12.*
22. Cao A., Rosatelli M.C. **Thalassemie.** in Cao A., Dallapiccola B., Notarangelo L.D. Malattie genetiche. Molecole e geni. Diagnosi, prevenzione e terapia. *Padova, Piccin Nuova Libraria 2004; 307-330.*
23. Thein S.L. et al. **Genetics of Hb F/F cell variance in adults and heterocellular hereditary persistence of fetal hemoglobin.** *Hemoglobin 1998; 22: 401-414.*
24. Galanello R., Sanna S., Perseu L., Sollaino M.C., Satta S., Lai M.E., Barella S., Uda M., Usala G., Abecasis G.R., Cao A. **Amelioration of Sardinian β^0 thalassemia by genetic modifiers.** *Blood. 2009; 114:3935-7.*

25. H. Liu et al. **BCL11A-XL splice variant and its interaction with BCL6 in nuclear paraspeckles of germinal center B cells**. *Molecular Cancer 2006, 5:18 doi:10.1186/1476-4598-5-18.*
26. Mackay J. & Mensah G.A. **The Atlas of Heart Disease and Stroke** *(World Health Organization, Geneva, 2004).*
27. Kuulasmaa K. *et al*. **Estimation of contribution of changes in classic risk factors to trends in coronary-event rates across the WHO MONICA Project populations**. *Lancet 2000; 355: 675-687.*
28. Clarke R. *et al.* **Cholesterol fractions and apolipoproteins as risk factors for heart disease mortality in older men**. *Arch. Intern. Med. 2007; 167: 1373-1378*
29. Gotto A.M. Jr. & Brinton E.A. **Assessing low levels of high-density lipoprotein cholesterol as a risk factor in coronary heart disease: a working group report and update**. *J. Am. Coll. Cardiol. 2004; 43: 717-724.*
30. Prospective Studies Collaboration. **Blood cholesterol and vascular mortality by age, sex and blood pressure: a meta-analysis of individual data from 61 prospective studies with 55,000 vascular deaths.** *Lancet 2007; 370: 1829-1839.*
31. Samani N.J. *et al*. **Genome-wide association analysis of coronary artery disease**. *N. Engl. J. Med. 2007; 357: 443-453.*
32. Murphy C. *et al.* **Regulation by SREBP-2 defines a potential link between isoprenoid and adenosylcobalamin metabolism**. *Biochem. Biophys. Res. Commun. 2007; 355: 359-364.*
33. Rauch U. *et al.* **Neurocan: a brain chondroitin sulfate proteoglycan**. *Cell. Mol. Life Sci. 2001; 58: 1842-1856.*
34. The Welcome Trust Case Control Consortium. **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature 2007; 447: 661-678.*
35. Helgadottir A. *et al.* **A common variant on chromosome 9p21 affects the risk of myocardial infarction.** *Science 2007; 316: 1491-1493.*
36. Barzilai N. *et al.* **Unique lipoprotein phenotype and genotype associated with exceptional longevity**. *J. Am. Med. Assoc. 2003; 290: 2030-2040.*

37. Baigent C. *et al.* **Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins.** *Lancet 2005; 366: 1267-1278.*
38. Cohen J.C. *et al.* **Multiple rare alleles contribute to low plasma levels of HDL cholesterol.** *Science 2004; 305: 869–872*
39. Law M.R. *et al.* **Quantifying effect of statins on low density lipoprotein cholesterol, ischaemic heart disease, and stroke: systematic review and meta-analysis.** *Br. Med. J. 2003; 326: 1423.*
40. Li B., Leal S.M. **Discovery of Rare Variants via Sequencing: Implications for the Design of Complex Trait Association Studies**. *PLoS Genet. 2009; 5(5): 1-9.*
41. McCarthy M.I., Hirschhorn J.N. **Genome-wide Association studies: potential next steps on a genetic journey.** *Human Molecular Genetics 2008; Vol 17 Review Issue 2: R156-R165.*
42. Y. Li & G. Abecasis, *unpublished data.*
43. 1000 Genomes Project Consortium, Durbin R.M., Abecasis G.R., Altshuler D.L., Auton A., Brooks L.D., Durbin R.M., Gibbs R.A., Hurles M.E., McVean G.A. **A map of human genome variation from population-scale sequencing.** *Nature 2010 Oct; 467(7319): 1061-73.*
44. Pritchard J.K. and Przeworski M. **Linkage disequilibrium in humans: models and data.** *Am. J. Hum. Genet. 2001; 69(1):1-14.*

RINGRAZIAMENTI

Voglio ricordare il Prof. Giuseppe Pilia che 6 anni fa mi diede la possibilità di iniziare questo percorso e che, grazie alla sua eccezionale intuizione, ha ideato, fondato e diretto il Progetto ProgeNIA fino al 2005, anno della sua prematura scomparsa.

Ringrazio la Dott.ssa Manuela Uda per avermi fornito, in questi anni, supporto logistico e scientifico, indispensabile per la mia crescita, nonchè il gruppo di ricerca dell'INN-CNR di Monserrato e di Lanusei per la produttiva collaborazione.

Ringrazio il Prof. Antonio Cao ed il Prof. Francesco Cucca per il fondamentale supporto e la costante disponibilità.

Voglio ringraziare il Dott. Carlo Sidore e la Dott.ssa Serena Sanna, colleghi ed amici che, guidati da epica pazienza, mi hanno introdotto in quello che ritengo il campo minato delle analisi statistiche.

Ringrazio anche una persona su cui ho potuto contare anche da molto lontano, il mio amico Dott. Andrea Maschio, e sopra ogni cosa i miei genitori, che mi hanno sempre, incondizionatamente, appoggiato nelle scelte che ho fatto.