Aspetti metodologici e nuovi sviluppi degli studi di associazione genetica di popolazione

Methodological issues and new developments in genetic association studies

Cristina Canova

Dipartimento di medicina ambientale e sanità pubblica, Università degli studi di Padova

***Corrispondenza:** Cristina Canova, Dipartimento di medicina ambientale e sanità pubblica, Sede di igiene, Università di Padova, via Loredan 18, 35131 Padova; e-mail: cristina.canova@unipd.it*

Riassunto

I fattori genetici interessano l'eziologia di molte malattie comuni, anche se è difficile identificare correttamente le variazioni geniche responsabili. L'obiettivo degli studi di associazione genetica è quello di identificare polimorfismi che variano sistematicamente tra gli individui con differenti stati di malattia. Il passaggio dagli studi su pochi geni candidati a studi sull'intero genoma è stato repentino per l'abbattimento dei tempi e dei costi di genotipizzazione, anche se i problemi metodologici sono rimasti i medesimi e ulteriormente aumentati. Uno dei problemi fondamentali è che il genoma è ampio e comprende molte variazioni polimorfiche che possono essere associate casualmente. Inoltre, le malattie complesse richiedono analisi complesse nelle quali molte varianti sono valutate simultaneamente.

Per aiutare a distinguere relazioni causali da quelle spurie, si propone che vengano stabiliti standard restrittivi per la significatività statistica o considerati solo gruppi di polimorfismi che potrebbero a-priori essere associati alla malattia.

(*Epidemiol Prev* 2008; 32(4-5): 254-57)

Abstract

Genetic factors are involved in the etiology of many common diseases, even if it is difficult to correctly identify the responsible genetic variations. The main goal of genetic association studies is to identify polymorphisms that vary systematically between people with different conditions of disease. Recent advances in time and cost for genotyping have rapidly moved from a candidate gene approach to genome-wide studies, but the methodological issues remain the same and furthermore increased. One of the major problem is that the genome is large and includes many polymorphic variants which could be found associated by chance. In addiction, complex diseases will require complex analyses in which many variants are assessed simultaneously.

To help to distinguish causal associations from spurious ones, restrictive standard for statistical significance or considering polymorphisms which could be a-priori causally associated to disease are needed.

(*Epidemiol Prev* 2008; 32(4-5): 254-57)

Introduzione

Questo intervento vuole focalizzare l'attenzione sui nuovi sviluppi degli studi di associazione genetica, in particolare sugli aspetti metodologici legati al passaggio da studi di associazione su geni candidati a studi sull'intero genoma.

Il genoma umano costituisce la completa sequenza del DNA. Circa il 3% del genoma consiste in sequenze codificate, con 30.000-40.000 geni codificatori di proteine.[1] Poiché i geni specificano la codifica delle proteine che costituiscono la struttura elementare e funzionale dell'organismo umano, ogni alterazione nel materiale genetico che porta alla variazione nella struttura e nella funzione della proteina vitale può portare a una malattia.

Una mutazione è definita in termini generali come un cambiamento nel materiale genetico, che può variare dalla sostituzione di una base nel DNA ad alterazioni che coinvolgano ampi segmenti cromosomici. Una delle più importanti classi di variazioni del DNA è rappresentata dai polimorfismi di un singolo nucleotide (SNP). Un polimorfismo è definito come la presenza in una popolazione di due o più fenotipi alternativi geneticamente determinati; in pratica un locus è considerato polimorfico se uno o più alleli rari hanno una frequenza maggiore di 0,01, con il risultato che l'eterozigota di questo allele si presenta con una frequenza maggiore del 2%.[2] A marzo 2007, il numero di SNP conosciuti nel genoma umano superava i 12 milioni (NCBI- National Center for Biotechnology Information).[3]

Sebbene molte malattie siano ereditate come semplici tratti Mendeliani o associate ad anomalie cromosomiche, la maggior parte delle comuni malattie di un uomo adulto non lo sono. Ad ogni modo, è evidente che fattori genetici giochino un ruolo determinante nell'eziologia di molti disturbi. La predisposizione genetica ad ammalarsi riflette l'effetto cumulativo di variazioni genetiche a differenti loci, ciascuno con un relativamente piccolo effetto sul fenotipo. Trovare tali varianti è importante, perché anche se una variazione genica risulta incrementare poco il rischio relativo di una malattia comune, può avere un'importanza rilevante in sanità pubblica in termini di frazione eziologia (la frazione della malattia nella popolazione attribuibile al fattore di rischio studiato).[4]

Gli studi di associazione genetica esaminano la relazione tra l'occorrenza di uno specifico allele in un generico locus e la malattia in una popolazione, utilizzando usualmente un disegno

caso-controllo. L'associazione è generalmente studiata su un locus suscettibile, che incrementa la probabilità di essere affetti dalla malattia, ma che non è «necessario» o «sufficiente» per l'espressione della malattia. Se uno specifico allele o genotipo risulta essere più frequente tra gli individui affetti dalla malattia rispetto agli individui sani, è possibile che essi giochino un ruolo importante nell'eziologia e nella patogenesi della malattia.
Gli studi di associazione genetica di popolazione possono utilizzare un'ipotesi di gene candidato, focalizzandosi su un particolare gene o area del genoma, o coinvolgere l'intero genoma senza alcuna ipotesi a priori (studi di associazione *genome-wide* - GW).
La prima tipologia di studi, esclusivamente e largamente utilizzata fino a un paio di anni fa, non ha portato a risultati consistenti. Una review del 2002 ha evidenziato in studi pubblicati tra il 1986 e il 2000 più di 600 associazioni statisticamente significative tra variazioni geniche e malattie, di cui tra 166 associazioni rilevate in più di tre studi indipendenti, solo 6 erano state replicate.[5] Tuttavia, molti di questi studi pubblicati non presentavano un appropriato disegno in termini di definizione di casi e controlli, selezione di marcatori genetici e in particolare di numerosità campionaria.
Gli studi di associazione con geni candidati hanno comunque identificato molti geni che potrebbero contribuire alla suscettibilità di malattie comuni.[6-8] Tuttavia, la validità di questi studi dipende dall'aver predetto correttamente l'identità del corretto gene o geni, sulla base di ipotesi biologiche o la locazione del gene candidato entro una regione di linkage precedentemente determinata.[9]
Poiché non necessitano assunzioni sulla locazione delle variazioni causali del genoma, gli studi di associazione GW possono sfruttare la forza degli studi di associazione in assenza di un'evidenza convincente riguardante la funzione o locazione del gene causale.[9] L'obiettivo degli studi GW non sembra essere più tanto lontano come quando questo approccio è stato proposto per la prima volta 10 anni fa,[10] grazie alla riduzione dei costi e alle tecniche di genotipizzazione più sofisticate come quelle basate sui *microarray*.
Per identificare alleli associati a una malattia con modesto effetto in studi di associazioni sono necessari, infatti, migliaia di soggetti e campioni. Perciò, l'utilità di sistemi di genotipizzazione basati sui *microarray* in larghi studi di associazione è determinata non solo del numero degli SNP (fino a centinaia di migliaia), ma dal numero dei campioni che deve essere processato in parallelo (migliaia di casi e migliaia di controlli).[11]
Gli studi di associazione GW sono ormai un fatto compiuto, grazie anche alle scoperte su larga scala dello *SNP consortium,*[12] e del progetto *HapMap* (The international Hapmap consortium).[13] La fase II di Hapmap[14] fornisce informazioni sulla locazioni di circa 3 milioni di SNP di quattro popolazioni di origine etnica differente. Per ciascuna popolazione è a disposizione l'informazione sulla struttura del linkage disequilibrium (LD) tra i polimorfismi che può aiutare a selezionare un sottogruppo di SNP (tagSNP) per catturare le informazioni di altri polimorfismi vicini senza genotipizzare queste varianti. Per essere utili, infatti, i marcatori testati nell'associazione dovrebbero essere o l'allele causale, o altamente correlati (in LD) con l'allele causale. La maggior parte del genoma si colloca in segmenti di forte LD, entro i quali le varianti risultano altamente correlate ciascuna con l'altra. Una volta che sono conosciuti pattern di SNP in una determinata regione, possono essere scelti relativamente pochi tagSNP (alcune centinaia di migliaia) per catturare la maggior parte della variazione all'interno di una regione. Il numero preciso di tagSNP necessario deve essere ancora determinato, e dipende dai metodi utilizzati per selezionare gli SNP e l'efficienza con la quale questi SNP possono essere etichettati in regioni con basso LD.[5,9]
Comunque, le limitazioni più ovvie agli studi GW rimangono, a oggi, l'elevato costo e il significativo sforzo richiesto per genotipizzare centinaia di migliaia di polimorfismi per ciascun individuo. A causa degli elevati costi, la numerosità campionaria è spesso limitata, con una conseguente riduzione della potenza. Tuttavia, poiché le variazioni che contribuiscono a tratti complessi hanno molto spesso un effetto modesto, è cruciale ottenere elevate numerosità campionarie.[9]
Una procedura per limitare, ma non risolvere, questi problemi è un approccio a 2 o più stadi, nel quale una proporzione dei campioni è genotipizzata su un elevato numero di marcatori allo stadio 1, utilizzando una bassa soglia di significatività (per esempio un p-value intorno a 0,05 senza correzione per confronti multipli) per considerare come positivi i marcatori valutati al primo stadio.
La soglia determinata al primo stadio è quella che permette di rilevare loci che spieghino anche solo una piccola frazione della variazione fenotipica, tollerando che vi sia un largo ma ragionevole numero di risultati falsi positivi. Tutti i marcatori che passano lo screening iniziale sono testati in un secondo campione di popolazione indipendente, di ampiezza maggiore alla popolazione iniziale, che può essere visto come uno studio di replicazione. Utilizzando metodi differenti di genotipizzazione nei diversi stadi, adeguati a set di polimorfismi di differenti grandezze, si potrebbe inoltre minimizzare la possibilità che vi siano associazioni falso-positive dovute alla tecnica di genotipizzazione.[9]
Comparati con i disegni a uno stadio che genototipizzano tutti i marcatori su tutti i campioni, i disegni su due stadi mantengono la potenza, mentre riducono sostanzialmente i costi e tempi di genotipizzazione.[15]
La potenza degli studi di associazione genetica a 2 o più stadi, per identificare varianti che predispongano alla malattia, dipende da un numero di fattori controllati dall'investigatore: quanti marcatori sono selezionati, quanti campioni sono suddivisi tra la fase 1 e le successive, la proporzione di marcatori testati dopo la prima fase e la strategia utilizzata per verificare le associazioni.[16] E' stato implementato un programma (CaTS)[16] per determinare le numerosità necessarie per un disegno a uno stadio e le percentuali ottimali di campioni da genotipizzare al primo stadio in un disegno multistadio.

Studi genome-wide: il problema dei confronti multipli

I metodi statistici per gli studi di associazione GW sono simili a quelli disponibili per geni candidati,[17] anche se ci sono importanti aspetti riguardanti l'efficienza statistica e computazionale. Poiché vi sono problemi computazionali nell'analisi di larghe basi di dati, i test sui singoli polimorfismi rimangono la procedura statistica principale negli studi GW. Per questi motivi, negli ultimi anni, sono stati implementati dei software *ad hoc* (come PLINK)[18] per queste tipologie di studi.

Il numero di possibili effetti genetici che possono essere testati negli studi di associazione è molto grande, specialmente quando s'iniziano a includere analisi di sottogruppi e analisi di interazione genetiche-ambientali. Sono necessarie delle procedure che distinguano tra associazioni spurie e reali, e che controllino per i risultati falsi positivi dovuti al caso.

Il problema dei test multipli è uno dei più rilevanti nel campo della statistica genetica e non vi è ancora una soluzione universalmente riconosciuta. Quest'aspetto, presente anche negli studi di associazione basati su geni candidati, in particolare quando si vogliono andare ad analizzare sottogruppi, è divenuto assolutamente rilevante negli studi sull'intero genoma.

La questione non è realmente legata al numero di test che vengono considerati: anche se un ricercatore testasse solamente uno SNP per un fenotipo, se molti altri ricercatori facessero lo stesso e venissero riportate le associazioni significative, ci sarebbe un problema di possibili risultati falsi-positivi. Poichè è altamente improbabile, a priori, che ogni data variante (o gruppo di varianti) sia associata causalmente con il fenotipo sotto il modello considerato, è richiesta un'evidenza molto forte per superare il giustificato scetticismo circa un'associazione.[17]

Il paradigma frequentista cerca di controllare l'intero errore di tipo I in modo tale che tutti i test generino nel complesso non più della probabilità α di risultati falso-positivi. In situazioni semplici l'approccio frequentista dice che, se n SNP sono testati e i test sono approssimativamente indipendenti, il livello di significatività più appropriato per ciascuno SNP α' dovrebbe soddisfare $\alpha=(1-\alpha')^n$, che porta alla correzione di Bonferrori $\alpha' \sim \alpha/n$.[19] Per esempio, testare $\alpha= 5\%$ su 1 milione di test indipendenti porterebbe ad $\alpha'=5\times10^{-8}$. Tuttavia, il numero effettivo di test indipendenti è legato a molti fattori, inclusa la numerosità campionaria e il test utilizzato.

Per SNP strettamente correlati, la correzione di Bonferroni è altamente conservativa. Esistono altri metodi frequentisti basati su procedure di permutazione e sulla proporzione di test falsi-positivi tra tutti i positivi (il tasso di scoperta negativa, *false discovery rate*) che permettono in parte di superare i problemi legati alla correzione di Bonferroni.[20-23]

L'approccio alternativo a quello frequentista è l'approccio Bayesiano in cui sono stimate le probabilità a priori dell'associazione tra il polimorfismo e la malattia, prima di procedere all'analisi dei dati.[24] La probabilità di ottenere risultati falsi positivi (*false-positive report probability*) è determinata da tre fattori:

- la dimensione del p-value dell'associazione trovata;
- la potenza statistica dello studio;
- la probabilità a priori che l'associazione tra la variazione genica e la malattia sia vera.

Stime approssimative della probabilità a priori possono includere informazioni da studi epidemiologici precedenti, dati funzionali sul gene e altre malattie con possibile simile eziologia. Le probabilità a priori possono essere successivamente combinate con l'analisi dei dati per ottenere probabilità a posteriori.

Anche la numerosità richiesta negli studi di associazione genome-wide risente del problema del numero di ipotesi che sono testate. Rish e Merikangas[10] proposero che un p-value di 5×10^{-8} (equivalente a un p-value di 0,05 dopo la correzione di Bonferroni per 1 milione di test indipendenti) fosse un livello conservativo opportuno per dichiarare un'associazione come significativa in uno studio sull'intero genoma. Considerando un allele con frequenza di 15% e un odds ratio di 1,25, sarebbero necessari 6.000 casi e 6.000 controlli per provvedere a una potenza dell'80% per rilevare associazioni con un p-value di 5×10^{-8}. Alternativamente, utilizzando un p-value di 0,05, sarebbero comunque necessari 1.200 casi e 1.200 controlli con una potenza dell'80%. Tuttavia, a dispetto dei benefici di permettere una numerosità campionaria più bassa, un p-value rilassato di 0,05 garantirebbe che il 5% di tutti i polimorfismi genotipizzati potrebbe essere associato alla malattia per caso. Considerando 500.000 polimorfismi, questo porterebbe a 25.000 possibili associazioni falso-positive. Perciò una soglia di 0,05 necessita studi di follow-up che distinguano i risultati falso-positivi da associazioni reali.

E' stato messo a disposizione un sito web per i calcoli della potenza statistica negli studi di linkage e di associazione (Genetic power Calculation).[25]

Conclusioni

L'obiettivo degli studi di associazione genetica è quello di identificare polimorfismi che variano sistematicamente tra gli individui con differenti stati di malattia. Uno dei problemi fondamentali è che il genoma è così ampio che i modelli che sono suggeriti da un polimorfismo causale possono essere dovuti al caso. Per aiutare a distinguere relazioni causali da quelle spurie, devono essere stabiliti standard restrittivi per la significatività statistica. Un'altra tattica è di considerare solo gruppi di polimorfismi che potrebbero plausibilmente essere generati da variazioni causali. Inoltre, le malattie complesse richiedono analisi complesse nelle quali molte varianti sono valutate simultaneamente.

Il passaggio dagli studi su pochi geni candidati a studi sull'intero genoma è stato repentino per l'abbattimento dei tempi e dei costi di genotipizzazione. I problemi metodologici, tuttavia, sono rimasti i medesimi, se non aumentati. Gli sforzi fino a ora condotti per incorporare informazioni funzio-

nali attraverso analisi Bayesiane risultano più complessi negli studi GW. Nello stesso tempo, essendoci maggiori probabilità di ottenere risultati falso-positivi, è reso ancor più indispensabile cercare di discriminare le vere associazioni da quelle false attraverso una corretta analisi dei risultati.

Sebbene l'approccio frequentista sia molto conveniente in situazioni semplici, un'aderenza rigorosa a questi metodi potrebbe essere restrittiva per l'analisi di studi di associazione, anche se a oggi la correzione di Bonferroni rimane il metodo più frequentemente utilizzato negli studi di associazione GW.[26] Utilizzando un approccio Bayesiano, viceversa, non vi sono penalità nell'analizzare esaustivamente i dati, poiché la probabilità a priori di un'associazione non dovrebbe risentire di quali e quanti test l'investigatore decida di eseguire.

Il problema dei confronti multipli non dovrebbe scoraggiare i ricercatori dall'eseguire analisi supplementari al di là dei test sui singoli polimorfismi, anche se è necessario che siano considerati correttamente i classici criteri epidemiologici per definire un'associazione, anche se forte, come causale. La mancanza di consistenza di molti studi di associazione riportati in letteratura può essere dovuta a diversi errori dello studio tra cui: errori nella genotipizzazione, problemi nel disegno dello studio in particolare nella selezione dei controlli, insufficiente potenza statistica per analisi a posteriori o di sottogruppi, bias di pubblicazione, comparazioni multiple non giustificate, insufficiente considerazione del modello di ereditabilità di una malattia genetica, probabilità che il gene studiato spieghi una piccola proporzione della variabilità del rischio.

Non si deve inoltre pensare che i classici problemi degli studi di associazione genetica, come quelli statistico/computazionali e la mancanza di replicazioni positive, possano essere risolti attraverso studi GW.

E' ipotizzabile che gli studi di associazione su geni candidati siano destinati a esaurirsi nei prossimi anni. Con l'avanzamento degli sviluppi tecnologici è naturale prevedere, infatti, un forte incremento, sino a pochi anni fa insospettabile, degli studi GW. Il pericolo è che, in particolare piccoli gruppi di ricerca, investano per genotipizzare l'intero genoma a scapito della potenza statistica.

Il principale vantaggio degli studi GW è quello di poter individuare geni che non erano stati precedentemente indiziati a essere associati con la malattia in studio. Per far questo è necessario che i risultati vengano replicati in più studi indipendenti, possibilmente con una maggiore numerosità campionaria, anche all'interno dello stesso studio con disegno multistadio.

E' infine auspicabile che, nei futuri studi di associazione GW, sia distinta la ricerca di nuove ipotesi, dai test di ipotesi specifiche (anche per replicare risultati precedentemente rilevati) in cui è definita a priori la relazione funzionale tra la variazione genica e la malattia indagata.

Conflitti di interesse: nessuno

Bibliografia

1. Carlson CS, Eberle MA, Kruglyak L, Nickerson DA. Mapping complex disease loci in whole-genome association studies. *Nature* 2004; 429(6990): 446-52.
2. Khoury MJ, Beaty TH, Cohen BH. *Fundamentals of Genetic Epidemiology.* 1993, New York, Oxford University Press.
3. http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi
4. Zondervan KT, Cardon LR. Designing candidate gene and genome-wide case–control association studies. *Nat Protoc* 2007; 2(10): 2492-501.
5. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. *Genet Med* 2002; 4(35): 45-61.
6. Cardon LR, Bell JI. Association study designs for complex diseases. *Nat Rev Genet* 2001; 2(2): 91-99.
7. Tabor HK, Risch NJ, Myers RM. Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat Rev Genet* 2002; 3(5): 391-97.
8. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* 2003; 33(2): 177-82.
9. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 2005; 6(2): 95-108.
10. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996; 273(5281): 1516-17.
11. Syvänen AC. Toward genome-wide SNP genotyping. *Nat Genet* 2005; 37(Suppl): S5-10.
12. Miller RD, Phillips MS, Jo I et al.; The SNP Consortium Allele Frequency Project. High-density single-nucleotide polymorphism maps of the human genome. *Genomics* 2005; 86(2): 117-26.
13. International HapMap Consortium. A haplotype map of the human genome. *Nature* 2005; 437(7063): 1299-320.
14. International HapMap Consortium et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007; 449(7164): 851-61.
15. Satagopan JM, Venkatraman ES, Begg CB. Two-stage designs for gene-disease association studies with sample size constraints. *Biometrics* 2004; 60(3): 589-97.
16. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* 2006; 38(2): 209-13.
17. Balding D. J. A tutorial on statistical methods for population association studies. *Nat Rev Genet* 2006; 7(10): 781-91.
18. Purcell S, Neale B, Todd-Brown K et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; 81(3): 559-75.
19. Bland J Martin, Altman Douglas G. Multiple significance tests: the Bonferroni method. *BMJ* 1995; 310(21): 170.
20. Dudbridge F, Koeleman BPC. Efficient Computation of Significance Levels for Multiple Associations in Large Studies of Correlated Data, Including Genomewide Association Studies. *Am J Hum Genet* 2004; 75(3): 424-35.
21. Nyholt Dale R. A Simple Correction for Multiple Testing for Single-Nucleotide Polymorphisms in Linkage Disequilibrium with Each Other. *Am J Hum Genet.* 2004; 74(4): 765-69.
22. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* 2003; 100(16): 9440-45.
23. Dudbridge F, Gusnanto A, Koeleman BP. Detecting multiple associations in genome-wide studies. *Hum Genomics* 2006; 2(5): 310-17.
24. Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Ins* 2004; 96(6): 434-42.
25. Purcell S, Cherny SS, Sham PC. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* 2003; 19(1): 149-50.
26. Pearson TA, Manolio TA. How to interpret a genome-wide association study. *JAMA* 2008; 299(11): 1335-44.