Laboratorio di Bioinformatica - Presentazione NAR

Relazione associata

Chiara Solito e Aurelia Timis

Corso di Laurea in Bioinformatica
Università degli studi di Verona
A.A. 2021/22

La presente è una relazione riguardante la presentazione nell'ambito dei database trattati nell'issue di. Per il corso di **Laboratorio di Bioinformatica** del CdS in Bioinformatica (Università degli Studi di Verona). Per la stesura di questa dispensa si è fatta fede al materiale didattico fornito direttamente dal professore nell'Anno Accademico 2021/2022. Eventuali variazioni al programma successive al suddetto anno non saranno quindi incluse.
Insieme a questo documento in formato PDF viene fornito anche il codice LaTeXcon cui è stato generato.

Contents

1 Introduzione

1.1 Lessico e nozioni di base

Varianti Causali Nell'ambito degli studi di associazione, le varianti genetiche responsabili del segnale di associazione in un locus sono indicate nella letteratura come VARIANTI CAUSALI. Esse hanno un effetto biologico sul fenotipo.

Trait-Associated Loci *Loci tratto associati*
Un locus a cui è associao un particolare tratto fenotipico.

QTL *Quantitative trait loci*
Un locus dei caratteri quantitativi (ovvero tratti che possono essere studiati e indagati mediante parametri numerici) è un locus che si correla con la variazione di un tratto quantitativo nel fenotipo di una popolazione di organismi.

GWAS *Genome Wide Association Study*
Un approccio della ricerca genetica per associare, a specifiche variazioni genetiche, particolari malattie.

Lead Variant La variante col miglior p-value per combinazioni gene/fenotipo significative.

P-Value La probabilità, per un'ipotesi supposta vera (ipotesi nulla), di ottenere risultati ugualmetnte o meno compatibili, di quelli osservati durante i test, con la suddetta ipotesi.

Tag Variants Varianti rappresentative in una regione del genoma con un alto linkage disequilibrium.

Linkage Disequilibrium

Fine Mapping Dal momento che i risultati dei GWAS (che otteniamo) non sempre ci danno una sintesi completa delle statistiche, e avendo a disposizione solo le Variant Lead, dobbiamo apliare alle Variants Tag, in modo da avere un insieme più completo. Il fine mapping è uno dei metodi: è una tecnica dei GWAS per identificare la varianti genetiche che possono influenzare causalmente il tratto esaminato, in particolare cerca di determinare la variante genetica responsabile di ????

Single Evidence Score

Proxy Un proxy è una misura indiretta del risultato desiderato che è esso stesso fortemente correlato a quel risultato. È comunemente usato quando le misure dirette del risultato non sono osservabili e/o non disponibili.

2 Open Targets Genetics

2.1 Cos'è Open Target Genetics?

Open Target Genetics è l'ultima *release* della piattaforma Open Targets: una *partnership* tra pubblico e privato che utilizza i dati genetici e genomici umani per l'identificazione sistematica e la prioritizzazione dei bersagli farmacologici.
Il portale offre tre caratteristiche al fine di mettere in luce le associazioni tra **geni, varianti e tratti**:

- Sfogliare e classificare le associazioni di geni e varianti identificate dalla pipeline di punteggio **Locus-to-Gene (L2G)**
- Scoprire set credibili per associazioni di varianti e tratti basati sulla pipeline di analisi di *fine mapping*.
- Esplorare e confrontare gli studi della UK BioBank, di FinnGen e del catalogo GWAS utilizzando lo strumento di confronto multi-tratto

La novità di OTG

La maggior parte delle varianti, individuate attraverso i GWAS, si trova nella parte non codificante del genoma: ciò suggerisce che tali varianti vadano ad intaccare tratti complessi, alterando l'espressione dei geni vicini, attraverso meccanismi di regolazione, e influenzando in maniera significativa le malattie studiate dai GWAS. Identificare un gene causale è difficile poiché bisogna integrare dati dai GWAS con dati di trascrittomica, proteomica ed epigenomica prendendo in considerazione un'ampia tipologia cellulare o tissutale.In assenza di un portale già esistente che consenta di rispondere sistematicamente a un'ampia gamma di domande biologiche, è stato costruito OTG sulla base della tecnologia più recente per consentire di aggiungere e sfogliare facilmente i dati.

2.2 L'Obiettivo

Identificare bersagli farmacologici per lo sviluppo di medicinali sicuri ed efficaci è una priorità per l'industria farmaceutica; lo sviluppo di farmaci porta spesso a perdite di tempo e risultati fallimentari. I **farmaci con targets** che hanno evidenziato prove genetiche per associazioni a malattie, hanno dimostrato di essere vincenti nello sviluppo clinico. Ecco che, una sistematica valutazione di associazioni genetiche a particolari malattie o tratti può aiutare nella scoperta di targets (genes) per lo sviluppo di farmaci:
l'obiettivo di Open Targets Genetics è quindi di aggregare gli evidenti collegamenti tra VARIANTI e MALATTIE, e VARIANTI e GENI, così che, per una specifica malattia, potenziali bersagli farmaclogici possano essere prioritizzati basandosi su informazione genetica robusta, traducendo i segnali da GWAS e Biobank data in geni target, attraverso centinaia di tratti genome-wide.

Obiettivo di Open Targets Genetics

aggregare le prove che collegano

1. Varianti alla malattia
2. Varianti ai geni
3. Geni alle malattie

in modo che per una specifica malattia i potenziali bersagli farmacologici (drug targets) possano essere prioritizzati sulla base di solide informazioni genetiche.

2.3 Il Metodo

Aggregazione e fusione di:

- associazioni genetiche curate da letteratura e BioBank (UK)
- dati di genomica funzionale (sempre da UK BioBank)
  - conformazione della cromatina
  - interazione della cromatina
- loci dei tratti quantitativi
  - eQTL
  - pQTL

Viene applicata la "fine-mapping" (mappatura) statistica su migliaia di loci associati ai tratti per risolvere i segnali di associazione e colelgare ogni variante ai suoi geni bersaglio, prossimali e distali, usando uno score "single evidence".

3 Come funziona?

S = Study, Desease Association Information

Informazioni associate alla malattia, sono ottenute dai GWAS (Genome Wide Association Study) che collega lo "status" della malattia alla comune variazione genetica.

V_L = Lead Variant

Dato come sono riportati i GWAS è spesso l'unica variante che si conosce per ogni locus associato. Non può perl essere assunto che la lead variant causi l'associazione.

V_T = Tag Variants

Si espande la lead variant ad includere tutte le tag variants, che crea un set più completo di potenziali varianti causali.
Metodi:

1. fine mapping / credible set analisys
2. linkage disequilibrium

G = Genes

Dato il set di tag variants, si prosegue assegnandole ai geni, usando la V2G pipeline.

L'informazione sulle malattie e sui tratti associati (**S = study**) è ottenuta dai GWA Study. In base a come i risultati ottenuti dai GWAS sono riportati, spesso conosciamo solo la **VL = Variant lead**, a ciascun locus associato. In particolare, mentre alcuni studi offrivano una completa sintesi statistica, altri ne riportavano solo le variant lead. Tuttavia, non si può assumere che la VL stia causando l'associazione → si espande la VL per includere tutte le **VT = Variant Tag**, che costituiscono un insieme più completo di varianti potenzialmente causali. L'espansione viene fatta in due modi nei due modi sopra riportati. Questa fase prevede l'utilizzo della pipeline **V2D**.

4 Pipeline

Open Targets Genetics utilizza tre pipeline diverse.

4.1 Assigning Variants to Genes (V2G)

La pipeline V2G collega Varianti e geni, combinando dati provenienti da 4 fonti.

- Esperimenti sui loci dei tratti quantitativi del fenotipo molecolare (eQTL pQTL)
- Esperimenti di interazioni con cromatina
- In silico predizioni funzionali
- Distanza dal sito di inizio della trascrizione canonica

Per ciascuna variante, la pipeline prima assegna una prova funzionale della coppia V-G su tutte le fonti, poi applica un algoritmo di punteggio per produrre punteggi V2G aggregati.

4.2 Assigning Variants to Disease (V2D)

Open Targets Genetics prende in considerazione le associazioni variante-fenotipo riportate nel catalogo GWAS con p $\leq$ 1e-5. (Attualmente i dati sono derivanti da campioni di origine prevalentemente europea).

4.3 Prioritising causal genes at GWAS loci (L2G)

Ovvero la locus to gene, per cui si ha la prioritizzazione dei geni causali ai loci GWAS. Nonostante possa sembrare simile, è diversa da V2G, infatti usa un modello Machine Learning (si tratta di un addestramento di classificatore).
Al fine di consentire un apprendimento automatico supervisionato dei geni causali sono stati curati manualmente una serie di geni, utilizzando una repository di geni gold standard aperta al contributo della comunità (di questa repository si ha un'alta confidenza della funzionalità del gene implicato).
[i GSP sono i geni gold standard positivi, mentre gli altri sono definiti come GSN, ovvero geni gold negativi]

5 Un esempio di utilizzo

6 Possibili domande