

# 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 del database Open Targets Genetics. È stata scritta nell'ambito dei database trattati in *Nucleic Acids Research, 2021, Vol. 49, Database issue D1-D9*, per il corso di **Laboratorio di Bioinformatica** del CdS in Bioinformatica (Università degli Studi di Verona), A.A. 2021/2022.

## Contents

<b>1</b>	<b>Introduzione</b>	<b>2</b>
1.1	Lessico e nozioni di base . . . . .	2
<b>2</b>	<b>Open Targets Genetics</b>	<b>3</b>
2.1	Cos'è Open Targets Genetics? . . . . .	3
2.2	L'Obiettivo . . . . .	3
2.3	Il Metodo . . . . .	4
<b>3</b>	<b>Come funziona?</b>	<b>4</b>
<b>4</b>	<b>Pipeline</b>	<b>5</b>
4.1	Assigning Variants to Disease (V2D) . . . . .	5
4.2	Assigning Variants to Genes (V2G) . . . . .	6
4.3	Prioritising causal genes at GWAS loci (L2G) . . . . .	7
4.4	Colocalisation Analysis . . . . .	7
<b>5</b>	<b>Un esempio di utilizzo</b>	<b>8</b>
5.1	Study . . . . .	8
5.1.1	Compare Studies . . . . .	9
5.2	Variant . . . . .	10
5.3	Gene . . . . .	12
5.3.1	Gene Priorisation . . . . .	13
5.4	Locus Plot . . . . .	16
<b>6</b>	<b>Possibili domande</b>	<b>16</b>

# 1 Introduzione

## 1.1 Lessico e nozioni di base

### Nozioni di Statistica

**Single Evidence Score** Tra i livelli di Evidenza, anche definiti come gerarchia dell'evidenza, assegnati agli studi, basandosi sulla qualità metodologica della loro progettazione, validità e applicabilità alla cura del paziente: è il sesto livello di evidenza - Evidenza da un singolo studio descrittivo o qualitativo.

**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.

### **Credible Sets** *Set Credibili*

sono l'insieme più piccolo di varianti, selezionate in modo tale che la loro stima di copertura rettificata soddisfi la copertura del target. Ad esempio: in letteratura, gli autori in genere riferiscono di aver trovato un insieme credibile al 90% di cui sono fiduciosi almeno al 90% contenga la vera variante causale.

### **PP - Posterior Probability** *Probabilità a posteriori*

In statistica bayesiana, la probabilità a posteriori di un evento aleatorio o di una proposizione incerta, è la probabilità condizionata che è assegnata dopo che si è tenuto conto dell'informazione rilevante o degli antefatti relativi a tale evento aleatorio o a tale proposizione incerta. Similmente, la distribuzione di probabilità a posteriori è la distribuzione di una quantità incognita, trattata come una variabile casuale, condizionata sull'informazione posta in evidenza da un esperimento o da un processo di raccolta di informazione rilevanti (es. un'ispezione, un'indagine conoscitiva, ecc.).

**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.

---

## Nozioni sui GWAS

### **GWAS** *Genome Wide Association Study*

Un approccio della ricerca genetica per associare, a specifiche variazioni genetiche, particolari malattie.

**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.

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

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

### **Trait-Associated Loci** *Loci tratto associati*

Un locus a cui è associato 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.

**Linkage Disequilibrium** Linkage disequilibrium (LD) è l'associazione non casuale di alleli di diversi loci. Non esiste una singola statistica migliore che quantifica l'entità di LD. Sono state proposte diverse statistiche utili per scopi diversi.

**Fine Mapping** Dal momento che i risultati dei GWAS non sempre ci danno una sintesi completa delle statistiche, e avendo a disposizione solo le Variant Lead, dobbiamo ampliare le informazioni con le Variants Tag, in modo da avere un insieme più completo.

Il fine mapping è uno dei metodi con quei viene eseguita questa operazione: 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 della malattia o del fenotipo analizzato.

**Linkage Disequilibrium Espansione** Sappiamo che il genoma viene ereditato a blocchi e ogni blocco è definito da un aplotipo, che a sua volta è definito da un insieme di SNPs. OTG prende in considerazione l'LD perchè pazienti che hanno ereditato lo stesso segmento cromosomico, definito dal medesimo aplotipo, possono aver ereditato anche la stessa mutazione. L'LD ha permesso nelle analisi di associazioni, di studiare solo gli SNPs necessari a identificare il blocco di LD in disequilibrium.

## 2 Open Targets Genetics

### 2.1 Cos'è Open Targets 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 farmacologici 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 collegare ogni variante ai suoi geni bersaglio, prossimali e distali, usando uno score “single evidence”.

**3 Come funziona?****S = Study, Disease 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 per 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 analysis
2. linkage disequilibrium

## G = Genes

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

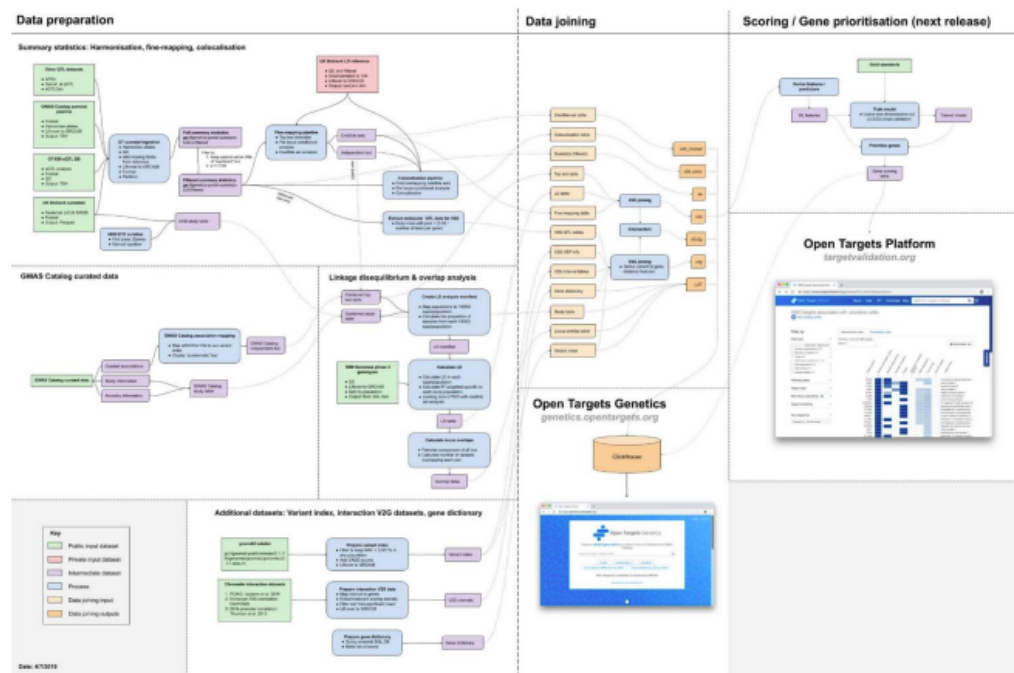
L'informazione sulle malattie e sui tratti associati (**S = study**) è ottenuta dai GWA Study. In base a come i risultati ottenuti dai GWAS sono riportati, spesso conosciamo solo la **VL = Variant lead**, a ciascun locus associato. In particolare, mentre alcuni studi offrivano una completa sintesi statistica, 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

OTG utilizza le sue pipeline interne al fine di aggregare le prove che colleghino variante-malattia, variantegene e gene-malattia, che rispecchiano lo scopo stesso del data base.

Tenendo presente che Open Targets Genetics è solo una risorsa dell'ecosistema Open Targets, che completa l'esistente piattaforma Open Targets Platform, vediamo dalla seguente immagine l'importanza delle pipeline esistenti in OTG che permettono il raggiungimento di evidenze genetiche, le quali comportano una successiva prioritizzazione e identificazione di bersagli farmacologici. I dati di prioritizzazione dei geni che usano il punteggio del modello machine learning L2G di OTG, alimentano le prove di associazioni genetiche per supportare le associazioni target-disease nella piattaforma Open Targets.

Questa pagina fornisce una panoramica di alto livello degli input e degli output e dei passaggi necessari per preparare, elaborare e unire vari set di dati per alimentare il portale OT-Genetics



### 4.1 Assigning Variants to Disease (V2D)

Le associazioni variante-fenotipo che troviamo in letteratura sono state identificate tramite il catalogo NHGRI-EBI GWAS: data base curato manualmente di varianti pubblicate che soddisfano determinati criteri di inclusione dettati dal catalogo stesso. In particolare, il catalogo GWAS estrae i dettagli sulle associazioni variante-fenotipo a un livello di significatività di  $p \leq 1e-5 (0.00001)$ , tenendo presente che il valore p aiuta a capire se la differenza tra il risultato osservato e quello ipotizzato è dovuta alla casualità introdotta dal campionamento, oppure

se tale differenza è statisticamente significativa, cioè difficilmente spiegabile mediante la casualità dovuta al campionamento.

***Nelle analisi di associazione genetica di tratti complessi, determinare la corretta soglia del valore di  $P$  per la significatività statistica è fondamentale per controllare il numero di associazioni di falsi positivi.***

Il termine p-value sta ad indicare la probabilità che quanto stiamo sostenendo sia corretto con un piccolo margine di errore, oppure esso spiega che la probabilità di sbagliarsi è troppo alta per sostenere la veridicità di quello che ipotizziamo. In OTG si includono le associazioni curate dal Catalogo GWAS con  $p \leq 5e^{-8}$  (0.00000005). I dati del repository delle statistiche di riepilogo del catalogo GWAS sono pertanto stati inclusi nel portale.

Oltre ai dati di derivazione del catalogo GWAS, si hanno a disposizione anche ampi archivi di statistiche riassuntive dai dati disponibili nella biobanca britannica. **Questi dati sono sfruttati per scoprire i locus della malattia.**

Vengono usati due metodi per espandere le varianti associate alla malattia in un insieme più completo di varianti possibilmente causali: [ricordando la differenza tra variant lead e variant tag: la VL è la variante].

1. Espansione del LD utilizzando una popolazione di riferimento: questa viene applicata a tutti gli studi in OTG

Le informazioni sul LD vengono calcolate utilizzando il pannello dell'aplotipo 1000 Genomes Phase 3 come riferimento

2. Espansione mediante mappatura fine (analisi del set credibile), utilizzata dove sono disponibili statistiche riassuntive complete (attualmente i tratti della biobanca britannica e quelli inclusi nel riepilogo del catalogo GWAS archivio di statistiche).

## 4.2 Assigning Variants to Genes (V2G)

Tutte le varianti nell'indice delle varianti vengono annotate utilizzando la pipeline di OTG V2G. la pipeline integra le prove V2G che rientrano in diversi tipi di dati principali:

- 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

All'interno di ogni tipo di dato ci sono molteplici fonti di informazioni prodotte da diversi metodi sperimentali. Alcune di questi fonti possono essere ulteriormente suddivise tenendo conto di tessuti/cellule.

STEP DA SEGUIRE:

### 1. Pre-elaborazione

a. **Filtraggio:** i set di dati grezzi vengono elaborati per conformarsi a un formato standard: si produce un file V2G.

b. **Trasformazione del punteggio:** diverse fonti utilizzano metriche diverse per misurare l'associazione V2G: viene estratta quindi una metrica specifica dello studio pertinente seguita dalla trasformazione dei quantili utilizzando una distribuzione uniforme. I punteggi trasformati vengono arrotondati al decile più vicino.

### 2. Annotazione variante-gene: ogni coppia (V, G) è annotata con tutte le prove funzionali disponibili.

### 3. Punteggio: è stato sviluppato un sistema di punteggio in modo tale che per una data variante V possiamo ottenere un elenco di geni (G1, G2..) classificato in base a:

- i. Il punteggio V2G complessivo
- ii. Un punteggio V2G per sorgente

a. AGGREGARE TRA LE CARATTERISTICHE: Siccome alcune fonti di dati forniscono associazioni misurate in pi tessuti o linee cellulari, laddove esistono più funzionalità, le aggregiamo prendendo

il punteggio massimo su tutte le funzionalità per ciascun (V, G) coppia. Questa aggregazione fornisce un punteggio V2G per sorgente per ciascuno (V, G) coppia.

b. AGGREGARE TRA LE ORIGINI: La fase dopo combina le informazioni per produrre un punteggio complessivo V2G. attenzione: poiché alcune fonti possono essere più attendibili di altre, prima dell'aggregazione (con una conoscenza precedente) si ordinano le fonti in base all'attendibilità.

### 4.3 Prioritising causal genes at GWAS loci (L2G)

**Caratteristiche predittive** L2G è un modello che dà la priorità a probabili geni causali in ciascun locus GWAS in base alle caratteristiche genomiche genetiche e funzionali. Le principali categorie di caratteristiche predittive sono (tenendo presente che abbiamo 51 caratteristiche predittive dalle 4 categorie qui sotto elencate):

- Distanza (dalle varianti del set credibile al gene)
- Colocalizzazione molecolare dei QTL
- Interazione della cromatina
- Patogenicità varianti.

Ogni set di dati di genomica funzionale fornisce punteggi da variante a gene → viene convertito il punteggio V-G in punteggi locus to gene, aggregando su varianti credibili identificate attraverso una mappatura fine.

**Per gli studi GWAS con statistiche riassuntive disponibili sono utilizzati insieme credibili dalla mappatura fine approssimata dal fattore di Bayes, altrimenti sono utilizzati insieme credibili basati su LD.**

Il modello L2G è distinto dalla pipeline V2G in quanto:

- Utilizza un modello di apprendimento automatico per apprendere il peso di ciascuna fonte di prove basato su un gold standard di geni causali precedentemente identificati. Per consentire l'apprendimento automatico sono stati curati manualmente una serie di geni.
- Si basa su dati di mappatura fine e colocalizzazione.

Tener presente che alcune caratteristiche sono migliore nel predire il gene: vi è disponibile anche un grafico che indica l'importanza delle caratteristiche.

**Interpretare il punteggio L2G** Il modello produce un punteggio tra 0 e 1: questo può essere interpretato per dire che i geni con un punteggio L2G di 0.5 hanno una probabilità del 50% di essere il gene causale nel locus.

**Attenzione:** è probabile che l'insieme di geni attualmente identificati come "causali" sia sbilanciato verso quelli che sono più vicini al picco GWAS → utile considerare non solo il punteggio, ma anche gli altri parametri disponibili.

Sono disponibili anche dei punteggi parziali: in particolare questi si trovano nella pagina GENE PRIORISATION per un locus. Qui oltre al punteggio complessivo troviamo anche quello parziale: si tratta di una sottoanalisi in cui si prende in considerazione solo un insieme parziale di caratteristiche. Anche in questo caso il punteggio è compreso tra 0 e 1.

I punteggi L2G parziali sono utili per capire quanto sia forte l'evidenza per un dato gene basato su un singolo tipo di perduttore, tenendo presente che questi modelli utilizzano meno info rispetto al modello completo.

### 4.4 Colocalisation Analysis

Utilizzata per verificare se due segnali di associazione indipendenti in un locus sono coerenti con l'avere una variante causale condivisa. Se due tratti condividono una variante causale (= sono colocalizzati) ciò aumenta anche l'evidenza che condividano anche un meccanismo causale.

Si fa una analisi di colocalizzazione che per due tratti, integra l'evidenza su tutte le varianti al locus per valutare le seguenti ipotesi:



- H0: Nessuna associazione con nessuno dei due tratti
- H1: Associazione con il tratto 1, non con il tratto 2
- H2: Associazione con il tratto 2, non con il tratto 1
- H3: Associazione con il tratto 1 e il tratto 2, due SNP indipendenti
- H4: Associazione con il tratto 1 e il tratto 2, un SNP condiviso

In questo quadro, l'evidenza per H4 è considerata evidenza per la colocalizzazione tra due tratti.

## 5 Un esempio di utilizzo

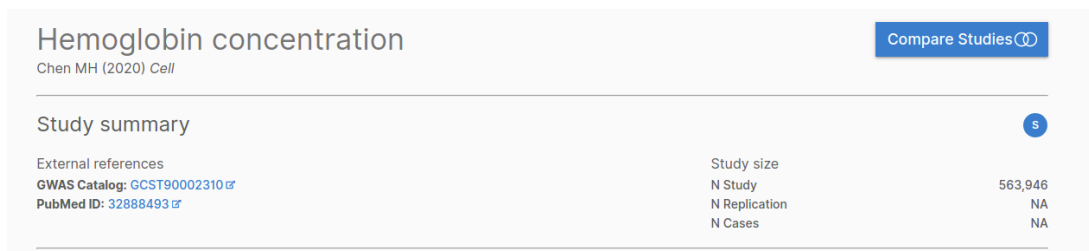
### 5.1 Study

#### Ricerca per Studio

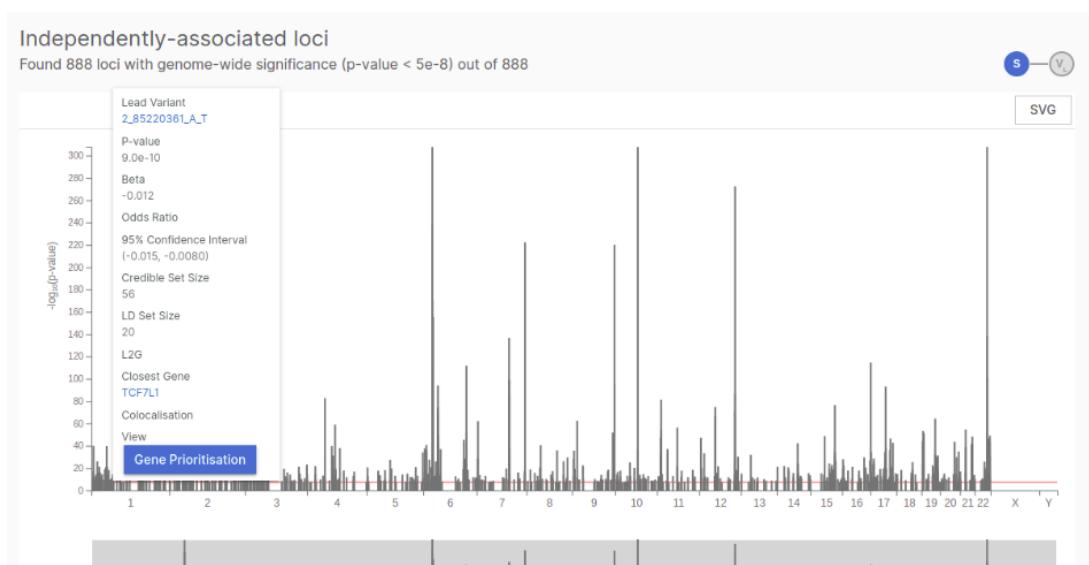
Iniziamo la ricerca a partire dallo studio per:

- Visualizzare i loci associati a un tratto nello studio selezionato
- Identificare i geni prioritari implicati funzionalmente da ciascun locus
- Visualizzare il 95% di set credibili (se disponibili) e proxy in ogni locus

La prima cosa che visualizziamo sono le informazioni generali relative allo studio (il *summary*), come l'ID di PUBMED, il numero di casi studiati, la grandezza dello studio ecc.



Subito dopo troviamo il primo plot: i loci associati indipendentemente sono riportati in un Manhattan plot semplificato. Sull'asse delle x vengono riportati i cromosomi, sull'asse delle y invece il p-value (in logaritmo).



I dettagli completi di ogni locus sono riassunti nella tabella sotto il Manhattan plot. Il gene con il punteggio più alto è definito come il gene con il maggior peso di prove funzionali tra tutte le fonti e i tipi di cellule che lo collegano al locus specificato direttamente o tramite una Variant Tag.

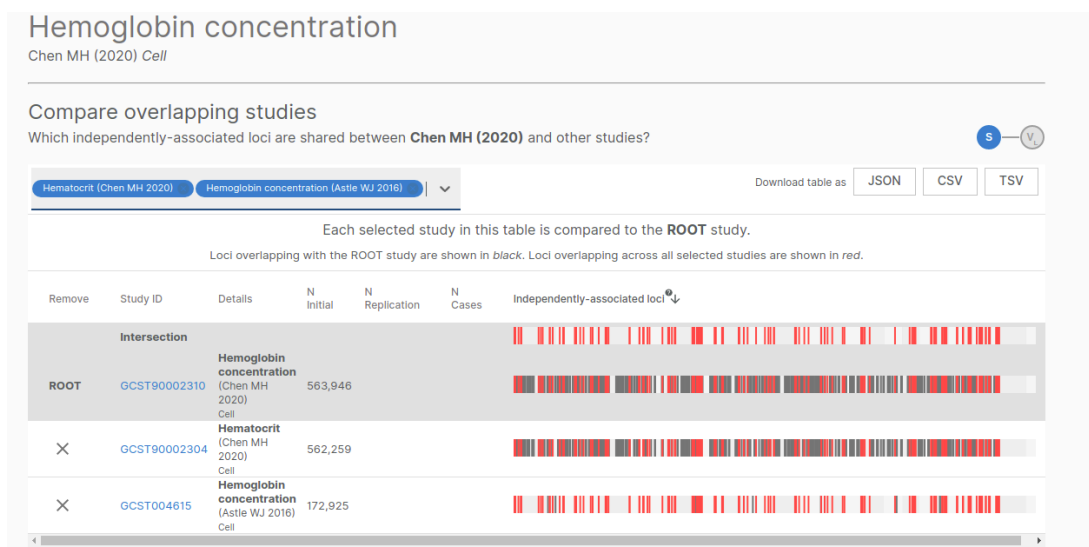
Association Information							Prioritised Genes results from our gene prioritisation pipelines			
Lead Variant	P-value	Beta	Odds Ratio	95% Confidence Interval	Credible Set Size	LD Set Size	L2G	Closest Gene	Colocalisation	View
<a href="#">6_26092913_G_A</a>	2.2e-308	0.19		(0.18, 0.19)	2	13	HFE	HFE		Gene Prioritisation
<a href="#">10_69334748_T_C</a>	2.2e-308	0.14		(0.14, 0.15)	7	9	HK1	HK1	HK1	Gene Prioritisation
<a href="#">22_37066896_A_G</a>	2.2e-308	0.10		(0.097, 0.10)	5	7	TMPSR56	KCTD17		Gene Prioritisation
<a href="#">2_46133768_G_C</a>	2.5e-303	-0.072		(-0.076, -0.069)	1	15		EPAS1		Gene Prioritisation
<a href="#">12_111494996_C_T</a>	3.7e-273	-0.065		(-0.069, -0.061)	3	19	SH2B3	SH2B3	HVCN1	Gene Prioritisation
<a href="#">6_25978174_T_C</a>	4.5e-238	0.091		(0.086, 0.097)	2	53		TRIM38		Gene Prioritisation
<a href="#">7_151716108_T_C</a>	5.1e-223	-0.066		(-0.070, -0.062)	3	31	PRKAG2	PRKAG2		Gene Prioritisation
<a href="#">9_133274295_A_T</a>	8.9e-221	-0.079		(-0.084, -0.074)	1	16	ABO	ABO		Gene Prioritisation
<a href="#">2_46128709_C_A</a>	2.4e-209	-0.10		(-0.11, -0.098)	2	8		EPAS1		Gene Prioritisation
<a href="#">6_26500335_G_T</a>	3.8e-171	0.074		(0.069, 0.079)		5		BTN1A1		Gene Prioritisation

1-10 of 888 | < > >>

### 5.1.1 Compare Studies

Dalla pagina Studio, è possibile confrontare rapidamente più studi per identificare segnali sovrapposti: **Compare Overlapping Studies**.

Il primo studio verrà caricato nella vista di confronto come root, con i loci riportati a significatività dell'intero genoma, plottati in base alla posizione. Solo gli studi con almeno un locus sovrapposto verranno visualizzati come opzione di confronto. Gli studi nel menù a tendina sono ordinati in maniera decrescente in base al numero di sovrapposizioni con la radice caricata.



Quando viene caricato più di uno studio (come in questo caso), loci intersecanti di tutti gli studi caricati vengono visualizzati in rosso sia sulla barra di intersezione nella parte superiore della vista, sia all'interno di ogni studio. I loci all'interno di ogni studio che si sovrappongono allo studio radice vengono visualizzati in nero. I loci non sovrapposti sono tracciati in grigio.

Sotto la visualizzazione del grafico, ogni locus sovrapposto in tutti gli studi caricati è riassunto in una tabella.

Download table as [JSON](#) [CSV](#) [TSV](#)

Loci in this table are shared across all selected studies.

Variant	rsID	Cytoband	Top Ranked Genes
<a href="#">1_3402205_C_G</a>	rs10909942	1p36	ACTRT2 PRDM16 ARHGEF16 MEGF6 TPRG1L WRAP73 TP73 PLCH2 SMIM1 CCDC27 LRRC47 DFFB CEP104
<a href="#">1_3774964_A_G</a>	rs1175550	1p36	CEP104 SMIM1 C1orf174 DFFB LRRC47 CCDC27 MEGF6 WRAP73 TP73 TPRG1L ARHGEF16 NPHP4 KCNAB2
<a href="#">1_10297748_G_C</a>	rs112682076	1p36	LZIC KIF1B NMNAT1 CTNNBIP1 RBP7 CLSTN1 PGD UBE4B CENPS-CORT CORT CENPS DFFA PEX14 PIK3CD TMEM201 SLC25A33
<a href="#">1_10423110_G_A</a>	rs12119893	1p36	KIF1B PGD CENPS DFFA CENPS-CORT PEX14 CASZ1 UBE4B CORT RBP7 NMNAT1 LZIC CTNNBIP1 CLSTN1
<a href="#">1_29159616_C_T</a>	rs113292219	1p35	EPB41 SRSF4 PTPRU MECR TMEM200B SDC3 MATN1 OPRD1 YTHDF2 LAPTM5 GMEB1 TAF12 RAB42 TRNAUT1AP RCC1 SERINC2
<a href="#">1_43295980_C_T</a>	rs4660253	1p34	TIE1 HYI MED8 SZT2 CDC20 MPL ELOVL1 C1orf210 EBNA1BP2 CFAP57 CCDC24 TMEM125 ZNF691 PTPRF SVBP PPCS ST3GAL3 ZMYND12 YBX1 FAM183A ATP6V0B PPIH ERMAP SLC2A1 KDM4A P3H1 B4GALT2 IPO13 DPH2 ARTN
<a href="#">1_45460823_C_G</a>	rs2055145	1p34	MUTYH TESK2 TIE1 MMACHC IPP PRDX1 MED8 HYI CCDC163 UROD MPL NASP ELOVL1 AKR1A1 ZSWIM5 TOE1 GPBP1L1 CDC20 CCDC17 HECTD3 MAST2 HPDL NSUN4 FAAH EIF2B3 BTBD19 SZT2 DYNLT4 PLK3 BEST4 PTPRF PTCH2 RPS8 PIK3R3 ARMH1 TMEM69 KIF2C TMEM53 C1orf210 TMEM125 EBNA1BP2 KDM4A ST3GAL3 CFAP57 FAM183A TSPAN1 RNF220 ARTN POMGNT1 MAST2 PIK3R3 NASP TESK2 GPBP1L1 AKR1A1 IPP MUTYH CCDC163 CCDC17 LURAP1 TMEM69 UROD NSUN4 PRDX1 POMGNT1 TSPAN1 FAAH P3R3URF TOE1 RPS8 MMACHC MOB3C RAD54L P3R3URF-PIK3R3 BTBD19 DYNLT4 PLK3 HECTD3 MKNK1 EIF2B3 UQCRH KIF2C LRRC41 HPDL DMAP1 ZSWIM5 PTCH2 DMBX1 TMEM275 KNCN CMPK1 BEST4
<a href="#">1_147815880_A_T</a>	rs1086605	1q21	GJA5 BCL9 GPR89B ACP6 GJA8 PDE4DIP NBPFI1 PRKAB2
<a href="#">1_161545536_C_T</a>	rs55971447	1q23	FCGR2A FCGR2B FCGR3B FCGR3A HSPA6 ATF6 FCRLB DUSP12 FCRLA MPZ FCE1G1G TOMM40L NIT1 SDHC CFAP126 ARHGAP30 USF1 PPOX APOA2 PCP4L1 NRII3 NDUFS2 ADAMTS4 B4GALT3 DEDD USP21 OLFML2B PFDN2 UFC1 KLHDC9 NECTIN4 TSTD1 NOS1AP F11R

1-10 of 181 | < > >|

I geni con il punteggio più alto visualizzati sono i geni principali implicati direttamente dalla Lead Variant mostrata e non tengono conto dei geni assegnati a nessuna Tag Variant della Lead. Potrebbe quindi esserci un elemento di mancata corrispondenza tra il gene mostrato qui e il gene funzionale previsto nel locus.

## 5.2 Variant



### Ricerca per Variante

Iniziamo la ricerca a partire dalla variante per:

- Identificare un elenco classificato di geni funzionalmente implicati dalla variante
- Visualizzare e analizzare i dati funzionali mediante i quali i geni sono assegnati a questa variante
- Visualizzare i risultati PheWAS per la variante nella biobanca britannica
- Visualizzare la struttura del collegamento intorno alla variante

Anche in questo caso per prima cosa vedremo le informazioni generali riguardo alla variante, che essa sia Lead o Tag le informazioni variano molto poco.

**1\_3402205\_C\_G** [Locus Plot](#)

**Variant summary**

Location  
**GRCh38:** 1:3,402,205  
**GRCh37:** 1:3,318,769  
**Reference allele:** C  
**Alternative allele (effect allele):** G

External references  
**Ensembl:** [rs10909942](#)

Neighbourhood  
**Nearest gene** (52,460 bp to canonical TSS): [ARHGEF16](#)  
**Nearest coding gene** (52,460 bp to canonical TSS): [ARHGEF16](#)

Variant Effect Predictor ([VEP](#))  
**Most severe consequence:** intron variant

Combined Annotation Dependent Depletion ([CADD](#))  
**raw:** -0.508 **scaled:** 0.0820

Population allele frequencies ([gnomAD](#))

African/African-American	0.377
Latino/Admixed American	0.680
Ashkenazi Jewish	0.659
East Asian	0.627
Finnish	0.664
Non-Finnish European	0.643
Non-Finnish European Estonian	0.637
Non-Finnish European North-Western	0.646
European	
Non-Finnish European Southern European	0.642
Other (population not assigned)	0.644

Subito sotto troviamo la tabella **Assigned Genes**, che riassume l'entità delle prove con cui la variante interrogata implica vari geni.

**Assigned genes**  
Which genes are functionally implicated by this variant?

Summary Distance (Canonical TSS) pQTL (Sun, 2018) eQTL Enhancer-TSS corr (FANTOM5) PCHI-C (Javierre, 2016) DHS- >

Download table as JSON CSV TSV

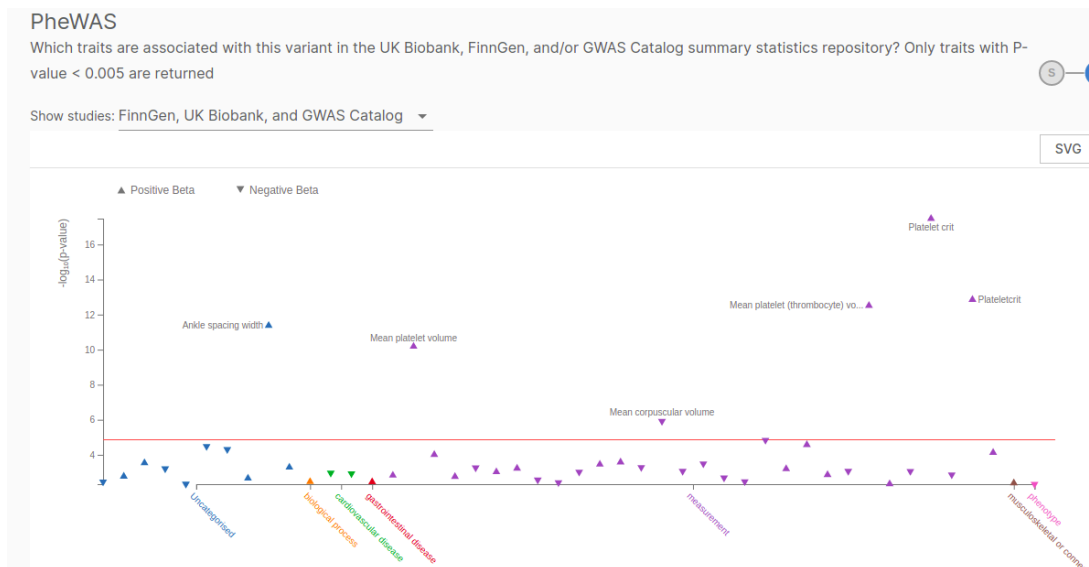
Evidence summary linking this variant to different genes.

Gene	Overall V2G ↓	Distance (Canonical TSS)	pQTL (Sun, 2018)	eQTL	Enhancer-TSS corr (FANTOM5)	PCHI-C (Javierre, 2016)	DHS-promoter corr (Thurman, 2012)	PCHI-C (Jung, 2019)	VEP (Ensembl)
PRDM16	●	333,037				●			intron_variant
ACTRT2	●	380,738				●			
ARHGEF16	●	52,460							
TPRG1L	●	222,810							
MEGF6	●	209,303							
WRAP73	●	250,556							
TP73	●	250,311							
PLCH2	●					●			
CCDC27	●	344,255							
SMIM1	●	370,544							

1-10 of 14 | < > >|

La visualizzazione predefinita riassume le prove combinate per ciascun gene per ciascuna origine dati funzionale, comprese tra i tipi di cellule all'interno dell'origine dati. Il G2V complessivo è una rappresentazione di questa ponderazione combinata delle prove per ciascun gene. La presenza di un **pallino** nella tabella indica che vi sono prove dalla data fonte di dati che collegano il gene alla variante interrogata, in almeno un tipo di cellula. Per visualizzare le prove specifiche del tessuto all'interno di un'origine dati, selezionare l'origine dati dalle schede visibili nella parte superiore del widget della tabella. Verrà aperta una vista equivalente segregata per tipo di cella anziché per origine dati, come sopra per le prove eQTL. Se l'evidenza dell'origine dati in esame può essere interpretata in modo direzionale, i proiettili verranno colorati in base alla direzione dell'effetto.

Subito sotto troviamo la sezione che presenta gli studi che collegano, con la variante in esame, determinati **PheWAS**.



I risultati di PheWAS per la variante selezionata in tutti i fenotipi della BioBank del Regno Unito, vengono visualizzati come un grafico PheWAS segregato per raggruppamento di fenotipi di alto livello e dettagliato in una tabella sottostante.

Download table as [JSON](#) [CSV](#) [TSV](#)

Study ID	Trait	Trait Category	P-value ↑	Beta	Odds Ratio	PMID	Author (Year)	N Cases	N Overall
	None	None							
<a href="#">NEALE2_30090_raw</a>	Platelet crit	measurement	3.3e-18	0.00131		<a href="#">UK Biobank</a>	UKB Neale v2 (2018)		350,471
<a href="#">GCST004607</a>	Plateletcrit	measurement	1.4e-13	0.0367		<a href="#">PMID:27863252</a>	Astle WJ (2016)		164,339
<a href="#">NEALE2_30100_raw</a>	Mean platelet (thrombocyte) volume	measurement	3.1e-13	0.0255		<a href="#">UK Biobank</a>	UKB Neale v2 (2018)		350,470
<a href="#">NEALE2_3143_raw</a>	Ankle spacing width	Uncategorised	4.2e-12	0.115		<a href="#">UK Biobank</a>	UKB Neale v2 (2018)		206,589
<a href="#">GCST004599</a>	Mean platelet volume	measurement	6.6e-11	0.0320		<a href="#">PMID:27863252</a>	Astle WJ (2016)		164,454
<a href="#">NEALE2_30040_raw</a>	Mean corpuscular volume	measurement	0.0000012	-0.0685		<a href="#">UK Biobank</a>	UKB Neale v2 (2018)		350,473
<a href="#">NEALE2_30050_raw</a>	Mean corpuscular haemoglobin	measurement	0.000015	-0.0255		<a href="#">UK Biobank</a>	UKB Neale v2 (2018)		350,472
<a href="#">NEALE2_30080_raw</a>	Platelet count	measurement	0.000027	0.790		<a href="#">UK Biobank</a>	UKB Neale v2 (2018)		350,474
<a href="#">NEALE2_30260_raw</a>	Mean reticulocyte volume	Uncategorised	0.000033	-0.105		<a href="#">UK Biobank</a>	UKB Neale v2 (2018)		344,728
<a href="#">NEALE2_30270_raw</a>	Mean spheroid cell volume	Uncategorised	0.000049	-0.0695		<a href="#">UK Biobank</a>	UKB Neale v2 (2018)		344,729

1-10 of 46 | < > >>

La direzione della freccia corrisponde alla direzione dell'effetto beta e i punti sono colorati in base al loro fenotipo ampio. Infine, vengono visualizzate due tabelle dedicate all'architettura genetica del locus a cui appartiene la variante di interesse - 'GWAS Lead Variants' e 'Tag Variants'. Il primo mostra tutte le varianti di lead GWAS (dal catalogo GWAS o dalla biobanca britannica) a cui la variante interrogata è stata assegnata come proxy (tag) in base a LD o fine-mapping. Se la variante interrogata è essa stessa una variante lead GWAS, la seconda tabella mostra tutte le varianti che le sono state assegnate come proxy. Questa tabella non verrà visualizzata se la variante interrogata non è una variante lead.

### 5.3 Gene



#### Ricerca per Gene

Iniziamo la ricerca a partire dal gene per:

- Identificare i loci che implicano funzionalmente un gene
- Collegarti a informazioni dettagliate sul gene e sui farmaci che lo prendono di mira
- Identificare in quali tratti questo gene può svolgere un ruolo, in base alle varianti a cui è assegnato

La ricerca porta alle principali informazioni sul gene e collega a vari link sulla piattaforma Open Targets Platform:

**PCK1**  
20:57,561,080-57,568,121

Locus Plot

Information about PCK1 from the Open Targets Platform

Target profile overview

Is there known drug data?

Is there mouse phenotype data?

Is there pathway data?

Is there expression data?

Other links

[Ensembl](#)
[GeneCards](#)
[GTEx](#)

[HGNC](#)
[UniProt](#)
[gnomAD](#)

accanto ci sono altri link. E sopra il link per collegarsi al Locus Plot, che vedremo in seguito. Subito sotto troviamo la sezione **Associated studies**, tratta dalla *locus-to-gene pipeline*.

Associated studies: locus-to-gene pipeline

Which studies are associated with PCK1?

Download table as

JSON

CSV

TSV

Study Information				Association Information						
Study ID	Trait	Publication	N Initial	Lead Variant	P-value	Beta	Odds Ratio	95% Confidence Interval	L2G pipeline score	View
	None	None								
GCST90002310	Hemoglobin concentration	Chen MH (2020)	563,946	20_57543981_A_G	1.2e-35	0.023		(0.020, 0.027)	0.84	Gene prioritisation
GCST90002308	Hematocrit	Chen MH (2020)	737,823	20_57561898_G_C	2.0e-28				0.84	Gene prioritisation
GCST90002403	Red blood cell count	Vuckovic D (2020)	408,112	20_57562742_G_A	4.0e-17	0.018		(0.014, 0.022)	0.83	Gene prioritisation
GCST90002384	Hemoglobin	Vuckovic D (2020)	408,112	20_57562778_A_G	2.7e-27	0.023		(0.019, 0.027)	0.83	Gene prioritisation
GCST90002363	Red blood cell count	Chen MH (2020)	545,203	20_57560878_C_G	2.4e-19	0.017		(0.013, 0.020)	0.83	Gene prioritisation
GCST90002383	Hematocrit	Vuckovic D (2020)	408,112	20_57560013_C_T	4.4e-24	0.022		(0.017, 0.026)	0.83	Gene prioritisation
GCST90002304	Hematocrit	Chen MH (2020)	562,259	20_57561898_G_C	1.6e-30	0.021		(0.018, 0.025)	0.83	Gene prioritisation
GCST008734	Waist-hip ratio	Lotta LA (2018)	663,598	20_57560143_C_T	5.0e-8				0.83	Gene prioritisation

La tabella associa gli studi e i fenotipi, da UKB, che sono associati con il gene di cui stiamo visualizzando la pagina. Un gene è connesso a un tratto nei casi in cui il gene è stato funzionalmente assegnato a un locus associato a questo tratto, tramite la Variant Lead o tramite una proxy Variant Tag assegnato. Accanto ci riporta al link **Gene Priorisation**, che vedremo in seguito.

Subito dopo abbiamo altri Studi associati, ma questa volta tramite la *colocalisation analysis*, che risponde alla domanda: quali studi hanno evidenze di colocalizzazione con qtl molecolari per il gene in questione?

Associated studies: Colocalisation analysis

Which studies have evidence of colocalisation with molecular QTLs for PCK1?

Download table as [JSON](#) [CSV](#) [TSV](#)

Study	Trait reported	Author	Lead variant	Phenotype	Tissue	Source	H3	H4	log2(H4/H3)	View
	None									
NEALE2_5098_raw	6mm weak meridian (right)	UKB Neale v2	20_57245990_G_T	ENSG000000124253	small intestine	GTEx-eQTL	0.20	0.061	-1.7	Gene Prioritisation
NEALE2_5134_raw	6mm strong meridian (left)	UKB Neale v2	20_57245990_G_T	ENSG000000124253	small intestine	GTEx-eQTL	0.20	0.061	-1.7	Gene Prioritisation
NEALE2_5133_raw	6mm strong meridian (right)	UKB Neale v2	20_57245990_G_T	ENSG000000124253	small intestine	GTEx-eQTL	0.20	0.060	-1.7	Gene Prioritisation
NEALE2_5099_raw	3mm weak meridian (right)	UKB Neale v2	20_57245990_G_T	ENSG000000124253	small intestine	GTEx-eQTL	0.20	0.060	-1.7	Gene Prioritisation
NEALE2_5132_raw	3mm strong meridian (right)	UKB Neale v2	20_57245990_G_T	ENSG000000124253	small intestine	GTEx-eQTL	0.20	0.060	-1.7	Gene Prioritisation
NEALE2_5135_raw	3mm strong meridian (left)	UKB Neale v2	20_57245990_G_T	ENSG000000124253	small intestine	GTEx-eQTL	0.20	0.060	-1.7	Gene Prioritisation
NEALE2_5097_raw	6mm weak meridian (left)	UKB Neale v2	20_57245990_G_T	ENSG000000124253	small intestine	GTEx-eQTL	0.20	0.059	-1.8	Gene Prioritisation
NEALE2_5096_raw	3mm weak meridian (left)	UKB Neale v2	20_57245990_G_T	ENSG000000124253	small intestine	GTEx-eQTL	0.20	0.059	-1.8	Gene Prioritisation
GCST006661	Male-pattern baldness	Hagenaars SP	20_56815683_C_CA	ENSG000000124253	liver	GTEx-eQTL	0.23	0.058	-2.0	Gene Prioritisation
GCST90002406	Reticulocyte fraction of red cells	Vuckovic D	20_57405659_T_C	ENSG000000124253	testis	GTEx-eQTL	0.27	0.066	-2.0	Gene Prioritisation

1-10 of 49 | < > >>

### 5.3.1 Gene Priorisation

In questa pagina stiamo studiando la malattia selezionata, nel locus suscettibile, possiamo ad esempio capire se l'allele alternativo è protettivo oppure no. Questa pagina è chiamata "Study-Locus" e vi si può accedere da Gene e da Studio. Qui si investiga una specifica Lead Variant e un GWAS study.

Dapprima abbiamo la informazioni di base: ad esempio tramite l'odds ratio capiamo se l'allele alternativo è associato ad un alto o basso rischio della malattia su cui si concentra lo studio. Tutte le dimensioni degli effetti sono in relazione all'allele alternativo, che è l'allele riportato ultimo nell>ID della variante.

**High cholesterol | non-cancer illness code, self-reported** Locus Plot

UKB Neale v2 (2018) NEALE2\_20002\_1473

Locus around **1\_55055640\_G\_T** (rs472495)

---

**Association summary**

P-value: 10e-15  
 Odds ratio: 1.1  
 Odds ratio Confidence Interval: (1.0, 1.1)

Subito sotto una tabella riporta i geni prioritizzati tramite la **Locus To Gene Pipeline**.

**Gene prioritisation using locus-to-gene pipeline**  
 Which genes were prioritised by L2G pipeline at this locus?

Download table as JSON CSV TSV

Gene	Overall L2G score <sup>⬇</sup>	Variant Pathogenicity <sup>⬆</sup>	Distance <sup>⬆</sup>	QTL Coloc <sup>⬆</sup>	Chromatin Interaction <sup>⬆</sup>	Distance to locus (bp)	Evidence of colocalisation
<a href="#">PCSK9</a>	0.85	0.70	0.82	0.45	0.26	16,193	No
<a href="#">BSND</a>	0.030	0.062	0.032	0.017	0.019	56,707	No
<a href="#">USP24</a>	0.028	0.066	0.046	0.038	0.12	159,724	No
<a href="#">TMEM61</a>	0.025	0.062	0.026	0.017	0.019	75,012	No
<a href="#">DHCR24</a>	0.023	0.062	0.025	0.021	0.021	168,445	No
<a href="#">LEXM</a>	0.0094	0.062	0.013	0.017	0.019	249,577	No
<a href="#">ACOT11</a>	0.0082	0.067	0.012	0.031	0.15		No
<a href="#">MROH7</a>	0.0078	0.068	0.012	0.043	0.30	413,886	No
<a href="#">TTC22</a>	0.0075	0.062	0.011	0.017	0.019	254,317	No
<a href="#">TTC4</a>	0.0070	0.062	0.011	0.019	0.020	339,779	No

1-10 of 13 |< < > >|

Questi sono ordinati tramite lo score assegnato appunto dalla pipeline, riportando anche tutte le features usate per fare il training del modello che ha effettuato la prioritizzazione (quindi distanza, colocalizzazione dei QTL, ecc.). Alla luce dei dati troviamo anche la risposta alla domanda: **“L’assegnamento del gene è supportato da evidenze di colocalizzazione?”** Subito dopo visualizziamo sempre informazioni riguardanti la prioritizzazione dei geni, ma ottenute tramite la colocalisation Analysis Pipeline: la colocalizzazione dei geni principali, ottenuti mediante la L2G pipeline, in diversi tessuti è mostrata in una heatmap e in una tabella.

**Gene prioritisation using colocalisation analysis**  
 Which molecular traits colocalise with **High cholesterol | non-cancer illness code, self-reported (UKB Neale v2, 2018)** at this locus?

Download table as JSON CSV TSV

Heatmap Table

Gene ↑	Molecular trait	Source	brain (substantia nigra)
<a href="#">PCSK9</a>	ENSG00000169174	Braineac2	●

1-1 of 1 |< < > >|

Nella tabella la QTL beta è per la variante che stiamo visualizzando, più che per tutte quelle riportate. Questo permette di comparare la direzione degli effetti nei diversi tessuti o geni nei tessuti, così da poter rimanere consistenti attraverso tutti gli studi.

Sotto, la tabella **GWAS Study Colocalisation**, mostra la colocalizzazione tratto-incrociata. Lo "Study beta" è sempre per la variante all’inizio della pagina.

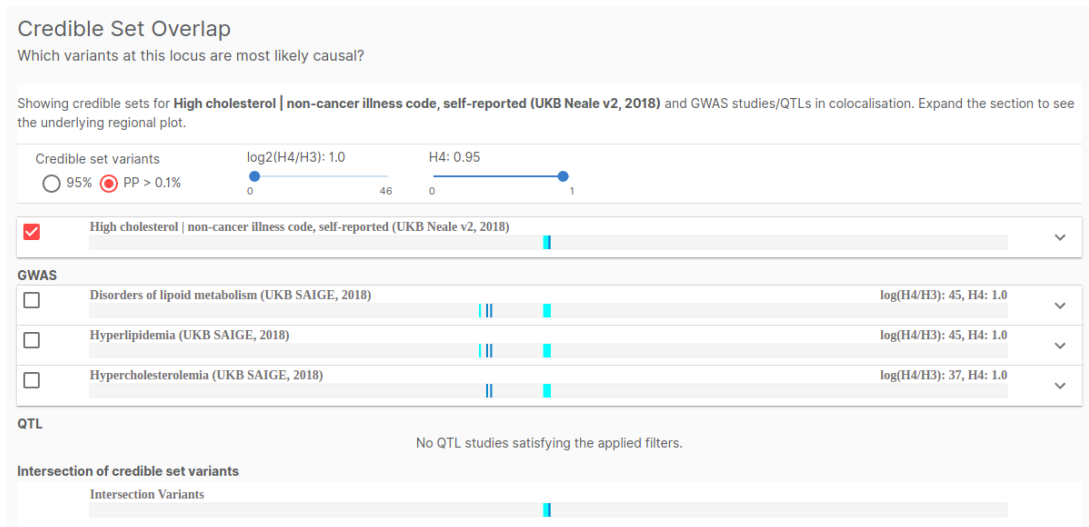
**GWAS Study Colocalisation**  
Which GWAS studies colocalise with **High cholesterol | non-cancer illness code, self-reported (UKB Neale v2, 2018)** at this locus?

Download table as [JSON](#) [CSV](#) [TSV](#)

Study	Trait reported	Author	Lead variant	Study beta <sup>Ⓢ</sup>	H3 <sup>Ⓢ</sup>	H4 <sup>Ⓢ</sup>	log2(H4/H3) <sup>Ⓢ</sup>	View
<a href="#">SAIGE_272</a>	Disorders of lipid metabolism	UKB SAIGE	<a href="#">1_55023869_C_T</a>	0.043	2.7e-14	1.0	45	<a href="#">Gene Prioritisation</a>
<a href="#">SAIGE_272_1</a>	Hyperlipidemia	UKB SAIGE	<a href="#">1_55023869_C_T</a>	0.043	3.6e-14	1.0	45	<a href="#">Gene Prioritisation</a>
<a href="#">SAIGE_272_11</a>	Hypercholesterolemia	UKB SAIGE	<a href="#">1_55023869_C_T</a>	0.044	8.3e-12	1.0	37	<a href="#">Gene Prioritisation</a>

1-3 of 3 |< < > >|

La sezione finale fornisce l'insieme delle probabili varianti causali (set credibili) a questo locus: **Credible Set Overlap** mostra il fine mapping dei set credibili attraverso più studi. Il primo è lo studio analizzato nella pagina:



Sono evidenziati le varianti che rientrano nel set credibile al 95% o con una probabilità a posteriori maggiore dello 0.1%.

Le freccette in basso accanto agli studi consentono agli utenti di visualizzare un grafico regionale di base per le consultare le statistiche di riepilogo. Vengono visualizzate anche le tracce per gli studi che colocalizzano: se due o più studi colocalizzano, è più probabile che condividano una variante causale, quindi confrontando i set credibili per gli studi di colocalizzazione, potremmo essere in grado di perfezionarli. Selezionando più tracce, nella tabella sotto l'elenco degli studi, si possono visualizzare le intersezioni di varianti nei set credibili.

Download table as [JSON](#) [CSV](#) [TSV](#)

Variant	Position	Maximum Posterior Probability <sup>Ⓢ</sup>	Product of Posterior Probabilities (across selected studies) <sup>Ⓢ</sup>
<a href="#">1_55055640_G_T</a>	55055640	0.50	0.50
<a href="#">1_55055436_G_A</a>	55055436	0.35	0.35
<a href="#">1_55053342_G_T</a>	55053342	0.045	0.045
<a href="#">1_55054539_G_A</a>	55054539	0.034	0.034
<a href="#">1_55052794_A_G</a>	55052794	0.023	0.023
<a href="#">1_55055569_T_G</a>	55055569	0.019	0.019
<a href="#">1_55054735_G_C</a>	55054735	0.018	0.018
<a href="#">1_55052855_C_A</a>	55052855	0.0068	0.0068

1-8 of 8 |< < > >|

**Genes**

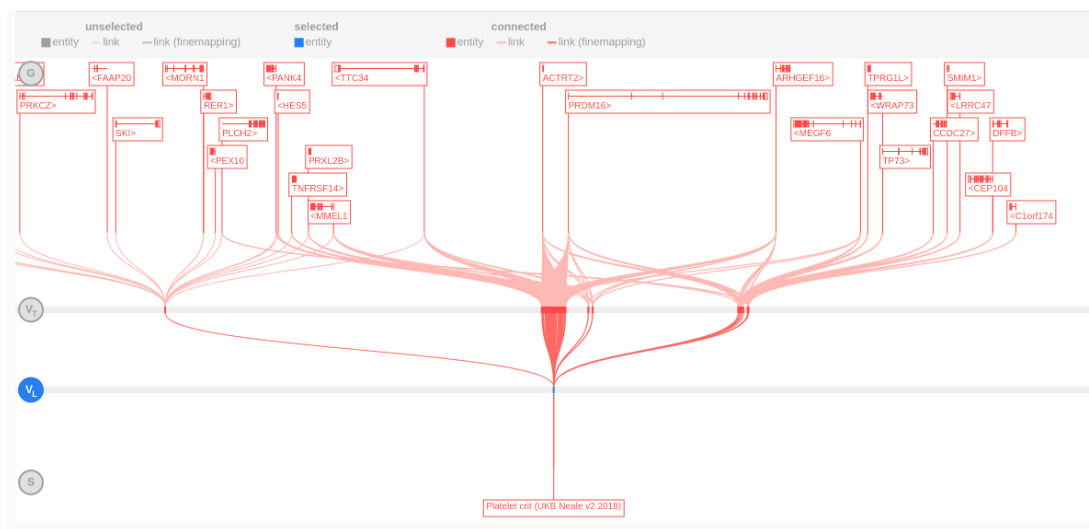


## 5.4 Locus Plot

L'ultima sezione è una delle più interessanti, si tratta del Locus Plot: accessibile da tutte e 4 le query (Study, Variant Lead, Variant Tag e Gene) e riporta, se si accede da una variante, se è in visualizzazione Lead o Tag.



È progettato per riassumere efficacemente la complessità dei collegamenti tra numerosi geni (bersagli), tratti e lead variants. Il Locus Plot utilizza un browser track-style per visualizzare un locus che può essere centrato su un gene, una variante o un tratto di interesse (ognuno noto come entità). Tratti di collegamento, varianti e geni sono stringhe di prova che indicano un'associazione o un collegamento funzionale tra le entità.



## 6 Possibili domande

**Che differenza c'è tra Varianti Lead e Tag?** Inanzitutto Open Targets è un portale di genetica concentrato sulle varianti. Le varianti lead si trovano dall'annotazione basata su dati curati manualmente del catalogo GWAS e dati provenienti dalla BioBanca britannica. La variant lead è la variante leader che guida nella ricerca, ma non si può dare per scontato che questa sia la variante causale, ovvero quella responsabile di un effetto biologico sul fenotipo. Per questo motivo si cerca di prendere in considerazione un più ampio set di variabili potenzialmente causali, che chiamiamo tag variants e che otteniamo da due tipologie di espansione:

- LD expansion (se la VL è presente nel pannello di riferimento) : sappiamo che il genoma viene ereditato a blocchi e ogni blocco è definito da un aplotipo, che a sua volta è definito da un insieme di SNPs. OTG prende in considerazione il LD perchè pazienti che hanno ereditato lo stesso segmento cromosomico, definito dal medesimo aplotipo, possono aver ereditato anche la stessa mutazione. Il LD ha permesso nelle analisi di associazioni, di studiare solo gli SNPs necessari a identificare il blocco di DNA in disequilibrio.
- Espansione per fine-mapping : identifica quali varianti hanno maggiore probabilità di essere causali (= responsabili dell'associazione)

**Cos'è la Colocalizzazione?** L'analisi di colocalizzazione viene utilizzata per verificare se due segnali di associazione indipendenti in un locus hanno una variante causale condivisa. Se due tratti condividono una variante causale (sono colocalizzati), ciò aumenta l'evidenza che condividono anche un meccanismo causale.

Esempio: se il segnale di una malattia e il livello proteico (pQTL) si colocalizzano, ciò può fornire la prova del ruolo della proteina nel causare la malattia.

**Cos'è la Prioritizzazione?** La Prioritizzazione di geni è il processo di assegnazione della probabilità del coinvolgimento del gene nella generazione di un fenotipo di malattia. Questo approccio restringe e organizza in ordine di probabilità nel coinvolgimento della malattia, l'insieme di geni da testare sperimentalmente. La prioritizzazione in OTG permette di dare supporto al portale OTP.

**Qual è la differenza tra GWAS e PheWAS?** Sono due tipologie di studio differente:

- Un PheWAS inizia con una variante genetica di interesse e analizza sistematicamente molti fenotipi (cioè, "a livello di fenotipo") per l'associazione al genotipo. È l'abbreviazione di *studio di associazione sull'intero fenomeno*: uno studio in cui l'associazione tra polimorfismi a singolo nucleotide o altri tipi di varianti del DNA viene testata su un gran numero di fenotipi diversi.
- Come spiegato in precedenza invece uno *studio di associazione sull'intero genoma* (GWAS) è un approccio utilizzato nella ricerca genetica per associare variazioni genetiche specifiche a particolari malattie. Il metodo prevede la scansione dei genomi di molte persone diverse e la ricerca di marcatori genetici che possono essere utilizzati per prevedere la presenza di una malattia.

**Come sono definiti gli Overlap nella pagina del Compare Studies?** Per definire la sovrapposizione per una data Lead Variant  $x$ , le Tag Variants, definite da LD-expansion di  $x$  sono incrociate con le Tag Variants di tutte le Lead Variants entro 5 mB da  $x$ . In ogni caso, quando una Tag Variants di  $x$  è condivisa con un'altra Lead Variants, tale Lead è considerata parte dello stesso segnale di  $x$ . Ogni locus condiviso, quindi, può essere considerato come un insieme di segnali che occupano un aptotipo comune.

**Come si legge il Manhattan Plot?** La linea rossa è la linea di significatività genome-wide.

Si può passare il mouse su un locus per visualizzarne i dettagli, incluso il gene più classificato per il locus secondo la pipeline di Open Targets Genetics. Selezionando un cromosoma, dal menù a discesa o facendo clic sul numero del cromosoma sull'asse x, la visualizzazione dei loci situati su quel cromosoma verrà ampliata e ingrandita e verranno limitati i dettagli dei loci visualizzati nella tabella allegata, di seguito. Vengono fornite le opzioni per scaricare il plot come vettore.

**Come si legge il Plot dei PheWAS?** La linea rossa denota il livello di significatività dopo la correzione di Bonferroni per il numero di fenotipi testati, considerandoli prudentemente indipendenti. Ogni fenotipo è rappresentato da un triangolo nel plot: la direzione del triangolo corrisponde alla direzione dell'effetto beta ed è colorato in base al suo "fenotipo ampio". I dettagli dell'associazione di ciascun fenotipo con il tratto di interesse sono visualizzati in una tabella sotto il grafico, che può essere ordinata, filtrata su colonna e scaricata. Per ciascun fenotipo viene fornito anche un collegamento alla vista locus diretta, che caricherà la vista locus con la variante e il tratto UK Biobank preselezionati.

**Quali sono le "tecnologie recenti" menzionate nell'approccio innovativo di OTG?** Partendo dai GWAS e dai più recenti dati di trascrittomica e proteomica, che appartengono alla BioBank, e quindi dai database più recenti e aggiornati (su cui si appoggia OTG rimanendo costantemente aggiornato), fino ad arrivare alle pipeline usate (di cui parleremo brevemente in seguito perché molto tecniche ed esulano dal portale in sé per sé su cui ci siamo concentrate)

**Come viene creata l'informazione contenuta nel database?** Tutte le pipeline sono interne ad OTG, non sono esterne, ma proprie del database.

**Di che tipo sono i campioni contenuti nel database?** OTG si basa principalmente su campioni europei, basandosi sulla UK BioBank. Non è stata trovata al momento una fonte di campioni differenti, che mantenga la stessa qualità, ma la piattaforma sta lavorando per ampliare i campioni e cercarne una adeguata.

**Cosa intendete quando parli di dati senza statistiche riassuntive dal catalogo GWAS, curati manualmente? Quali sono le statistiche riassuntive?** Le statistiche riassuntive sono i valori di p aggregati e i dati di associazione per ogni variante analizzata in uno studio di associazione sull'intero genoma. Non sempre i GWAS forniscono questi dati, è per questo che ottenendo dati senza statistiche aggiuntive è necessario integrarle alle Tag Variants.

**Cosa intendi con Linkage Disequilibrium espansione?** Sappiamo che il genoma viene ereditato a blocchi e ogni blocco è definito da un aplotipo, che a sua volta è definito da un insieme di SNPs. OTG prende in considerazione l'LD perchè pazienti che hanno ereditato lo stesso segmento cromosomico, definito dal medesimo aplotipo, possono aver ereditato anche la stessa mutazione. L'LD ha permesso nelle analisi di associazioni, di studiare solo gli SNPs necessari a identificare il blocco di LD in disequilibrium.

**Qual è la differenza tra 1\_55039974\_G\_T e rs11591147?** Entrambe le nomenclature fanno riferimento a una variante, mentre la prima ha una forma cromosoma\_posizione\_riferimento\_alternativo, la seconda rappresenta il nome.