Causal inference in drug discovery

NOVAMATH Thematic Weeks 2024

Short Course Lecture 2

Tom Michoel

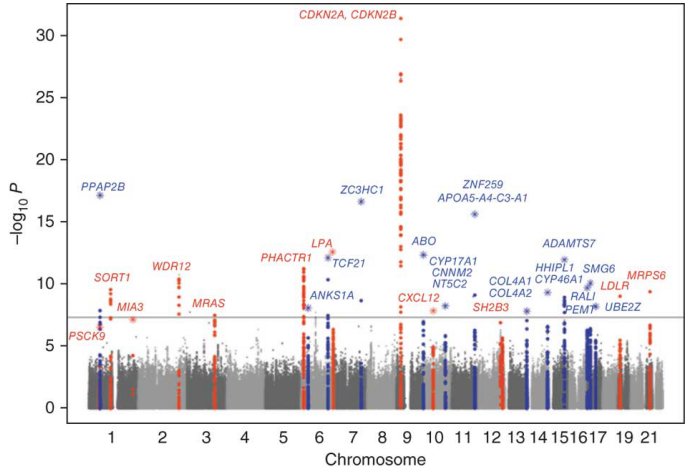
19 June 2024

Part I

A crash course in genetics & molecular biology

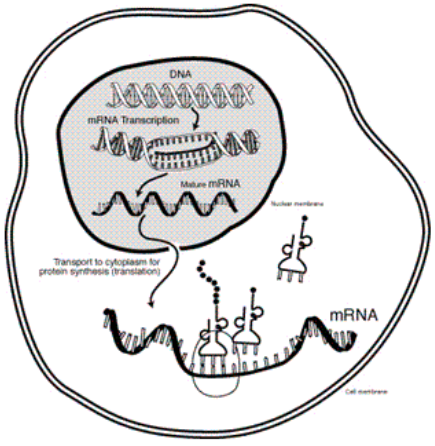
Genome-wide association studies (GWAS) identify genetic variants associated with complex traits and diseases

Genetic variation *causes* variation in complex traits.

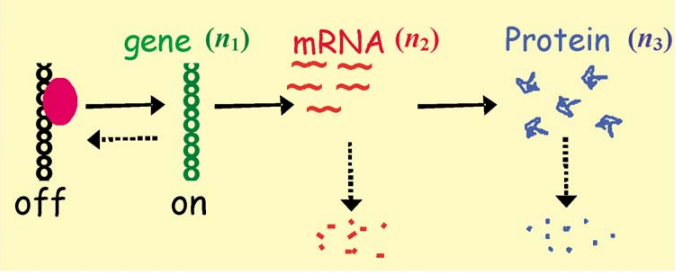


- Why *"causes"*?
  - Correlation ≠ causation.
  - Treatment randomly assigned vs. self-selected.
  - Genotype is fixed from conception and randomly distributed in population w.r.t. lifestyle and environment.
- How does genetic variation cause trait variation?

Gene expression determines cellular states



[Wikipedia]



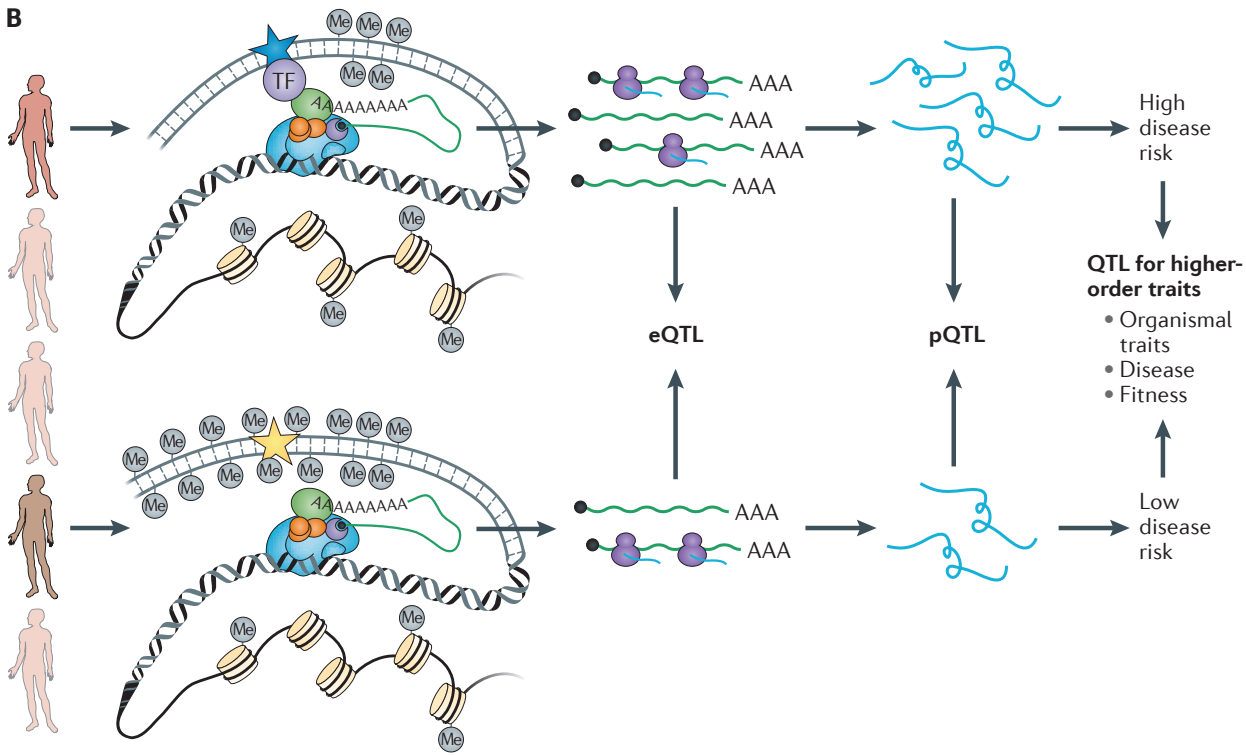
- Genes are transcribed into mRNA and translated into protein.
- Every step in this process can be regulated by genetic or environmental factors.
- The repertoire and relative levels of proteins expressed determines cellular identity, fate, cell-to-cell communication, etc.
- Variation in cellular properties causes variation in phenotypes.

Genetic variation causes trait variation by affecting gene expression.

How does genetic variation affect mRNA expression?

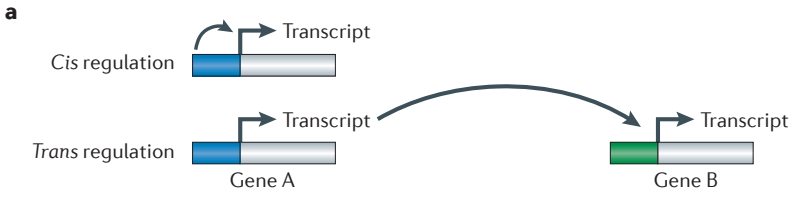
Genetic variation causes variation in transcription factor binding

- TFs bind to accessible regulatory DNA elements to control gene transcription.
- Environmental perturbations affect TF concentrations in nucleus.
- Genetic perturbations affect ability of TFs to bind to DNA.

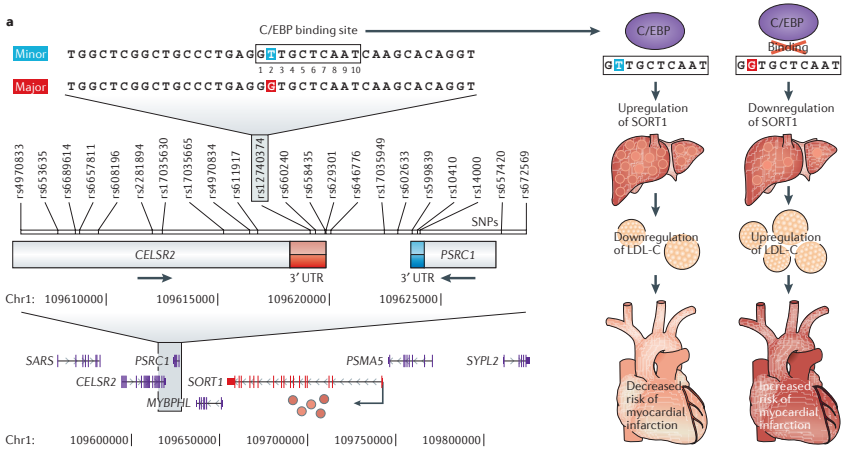


[Albert and Kruglyak, Nat Rev Genet (2016)]

Genes are organized in hierarchical, multi-tissue causal networks



[Civelek and Lusis, Nat Rev Genet (2014)]

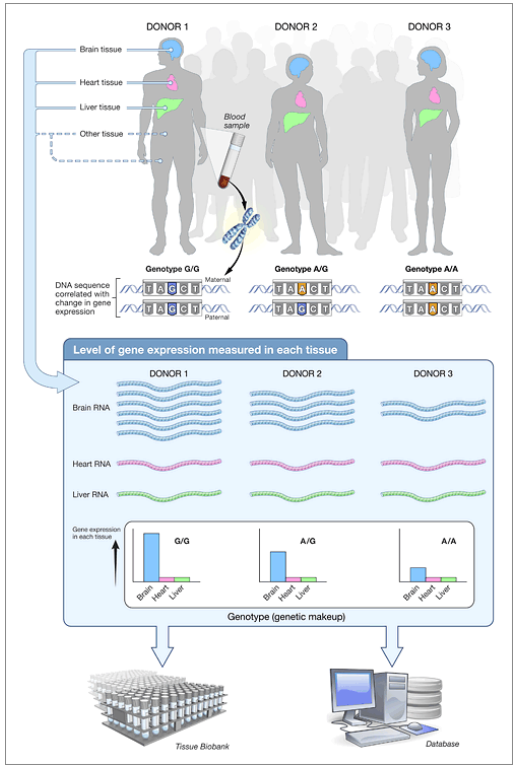


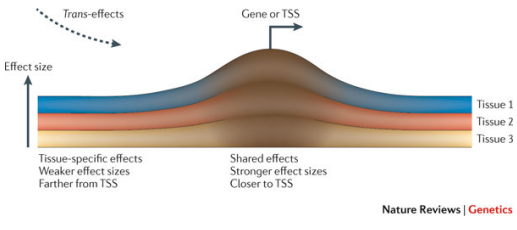
[Albert and Kruglyak, Nat Rev Genet (2016)]

- Variation in expression of one gene has downstream consequences on expression of other genes.
- Example: Introduction of just 4 TFs ("Yamanaka factors") converts adult cells into stem cells.
- Hundreds to thousands of genes are differentially expressed between cellular states (e.g. healthy vs. disease).
- Gene expression in one tissue can affect gene expression in other tissues.
- Phenotype variation also causes gene expression variation ("environmental" perturbation).

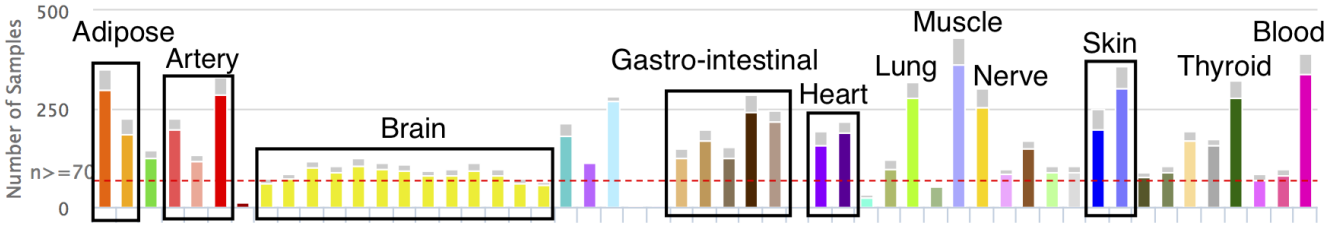
Inference (reconstruction) of causal gene networks is essential to understand how the genotype determines the phenotype.

Genome-wide transcriptome variation studies map the genetic architecture of gene expression





- eQTL = expression QTL = genetic variant associated with gene expression level.
- Strongest effects in gene proximity (regulatory region).
- RNA-sequencing measures expression levels of ~40k transcripts in a single sample.
- Gene expression is highly tissue-specific, but only few tissues are easily accessible (blood, immune cells).
- Genotype-Tissue Expression project (GTEx): 48 tissues, 620 donors (deceased), 10,294 samples → 341,316 eGenes (31,403 unique)



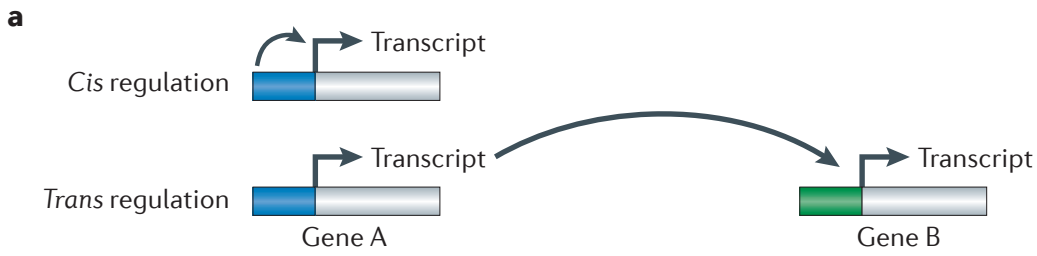
[GTEx website (2017)]

Part II

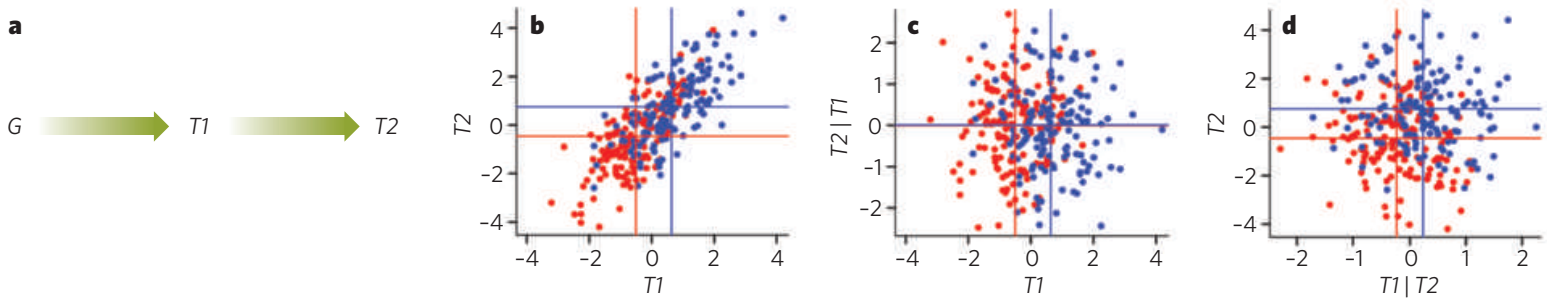
Causal inference in systems genetics

Genome-transcriptome variation studies inform on causal interactions

- Genetic variation between individuals precedes gene expression variation ⇒ eQTLs act as instrumental variables to infer causal direction between correlated expression traits.
- Molecular version of a random trial.



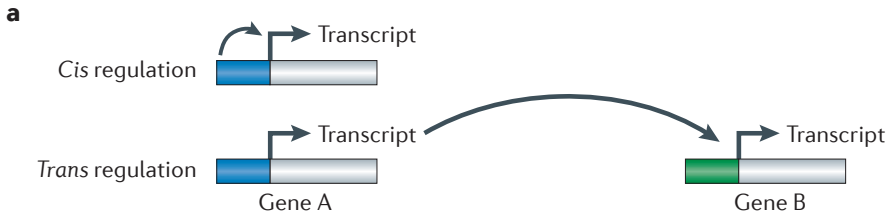
[Civelek and Lusis, Nat Rev Genet (2015)]



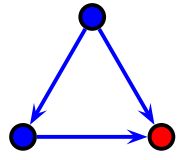
[Rockman, Nature (2008)]

Genome-transcriptome variation studies inform on causal interactions

- *Cis*-eQTLs act as genetic instruments to infer the causal direction between coexpressed genes



- Main differences with "standard" causal inference/Mendelian randomization:
  - Thousands of traits are analyzed simultaneously:
    - Computationally challenging ☹
    - Bayesian inference of real vs. null distributions ☺
  - eQTL effect sizes are often large ☺
  - Pleiotropy is rare ☺
  - Confounding by shared upstream regulators is common: FFL is most abundant motif in gene regulatory networks ☹



Traditional causal inference in genome-wide studies is based on a conditional independence test

Given a pair of gene expression traits A, B, and a *cis*-eQTL E of gene A, **Chen, Emmert-Streib and Storey (Genome Biol 2007)** estimate:

$$P(E \to A \to B) = P(E \to A) \times P(E \to B) \times P(B \perp E \mid A)$$

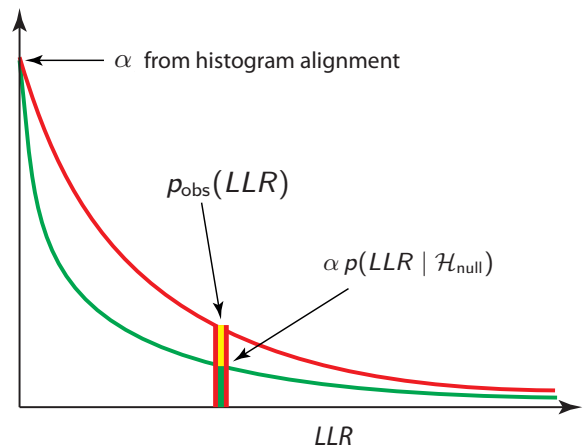
Causal interaction from A to B, using E as instrument	Association between E and A	Association between E and B	B independent of E given A

Approach: For each factor:

1. Likelihood ratio between null and alternative model.
2. Probability that alternative hypothesis is true, given LLR test statistics from real and random distribution.

$$p_{\text{obs}}(LLR) = \alpha\, p(LLR \mid \mathcal{H}_{\text{null}}) + (1-\alpha)\, p(LLR \mid \mathcal{H}_{\text{alt}})$$

$$P(\mathcal{H}_{\text{alt}} \mid LLR) = 1 - \alpha\, \frac{p(LLR \mid \mathcal{H}_{\text{null}})}{p_{\text{obs}}(LLR)},\ p(LLR \to 0 \mid \mathcal{H}_{\text{alt}}) = 0$$



$p_{\text{obs}}(LLR)$ from testing 1000s of B's for each A.
$p(LLR \mid \mathcal{H}_{\text{null}})$ from random permutations.

LLRs are computed using linear models and ML parameter estimates

Association between E and A

Alternative model: $A \mid E \sim \mathcal{N}(\mu_E, \sigma_A^2)$

Null model: $A \sim \mathcal{N}(\mu, \sigma^2) \perp E$

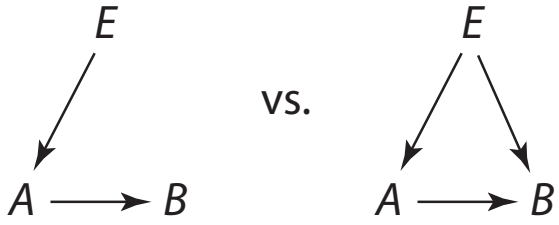
$$LLR = \log \frac{\prod_i p(a_i \mid e_i, \hat{\mu}_{e_i}, \hat{\sigma}_A^2)}{\prod_i p(a_i \mid \hat{\mu}, \hat{\sigma}^2)}$$

B independent of E given A

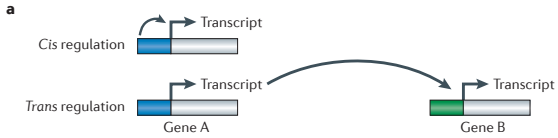
Alternative model: $A \mid E \sim \mathcal{N}(\mu_E^A, \sigma_A^2)$

$B \mid A \sim \mathcal{N}(a x_A + b, \sigma_{BA}^2)$

Null model: $A, B \mid E \sim \mathcal{N}\left(\begin{bmatrix} \mu_E^A \\ \mu_E^B \end{bmatrix}, \begin{bmatrix} \sigma_A^2 & \rho_{AB}\sigma_A\sigma_B \\ \rho_{AB}\sigma_A\sigma_B & \sigma_B^2 \end{bmatrix}\right)$



Mediation allows causal model selection only in the absence of hidden confounders



M_1 $E_A \longrightarrow A \longrightarrow B$ *vs.* M_2 $E_A \longrightarrow A \longleftarrow B$ or M_3 $E_A \rightarrow A$, $E_A \rightarrow B$ or M_4 $E_A \rightarrow A \leftarrow B$, $E_A \rightarrow B$

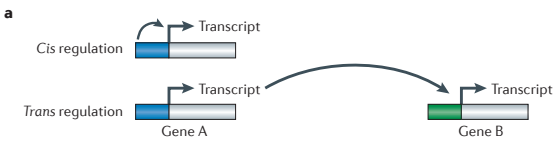
1. *Trans-association:* Statistical dependence between E_A and B excludes model M_2.
2. *Mediation:* Statistical independence between E_A and B conditioned on A excludes model M_3 and M_4.

BUT

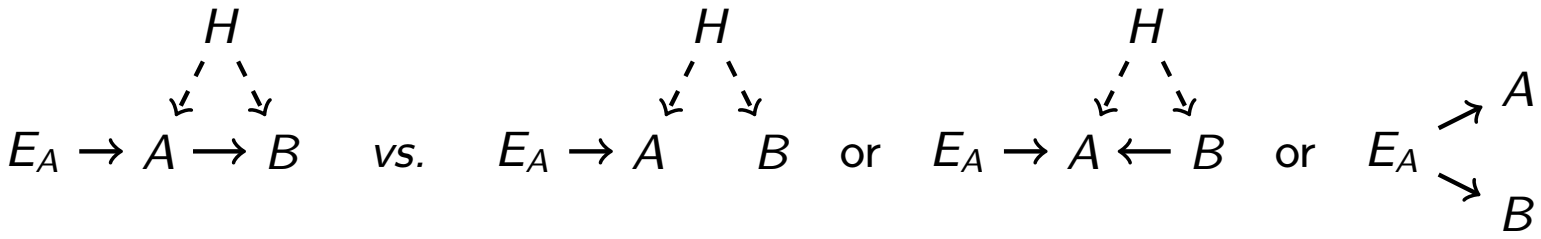
Mediation fails in the presence of hidden confounders due to collider effect.

$E_A \rightarrow (A) \rightarrow B$, $H \rightarrow A$, $H \rightarrow B$ $\Rightarrow$ E_A, $H \rightarrow B$

Instrumental variables allows causal model selection with hidden confounders



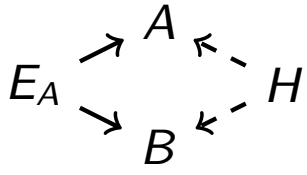
M_1 M_2 M_3 M_4



1. *Trans-association:* Statistical dependence between E_A and B excludes model M_2 and M_3.
2. *Instrumental variables:* Statistical dependence between A and B conditioned on E_A excludes model M_4.

BUT

Instrumental variables fails if there is a hidden confounder in model M_4.



*Because eQTL-gene interactions are **local**, a direct effect of E_A on B can only occur if A and B are co-located on the genome.*

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Efficient and accurate causal inference with hidden confounders from genome-transcriptome variation data

Lingfei Wang, Tom Michoel*

- Findr software implements 6 likelihood-ratio tests and outputs probabilities (1-FDR) of hypotheses being true.
- Analytic results to avoid random permutations, resulting in massive speed-up.
- Combining LLR tests gives probabilities of causal effects.
  - **Mediation**: $P = P_1\,P_2\,P_3$
  - **Instrumental variables**:

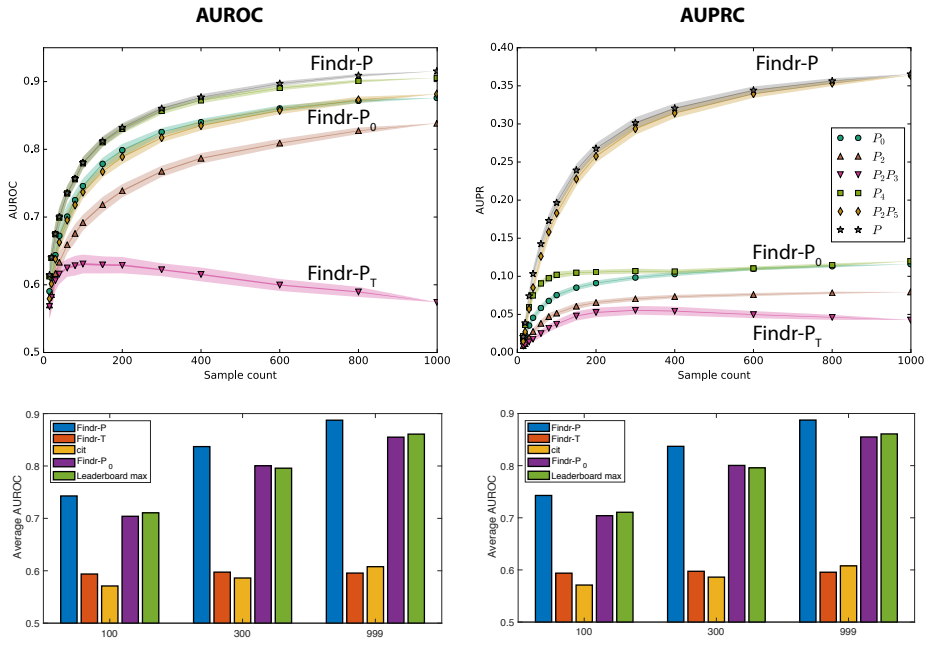
    $P = P_1\,P_2\,P_5$ or $P_1\,(P_2\,P_5 + P_4)$

Test ID	Test name	Null (hypothesis)	Alternative (hypothesis)	Selected hypothesis
0	Correlation	A B	A — B	Alternative
1	Primary (Linkage)	E A	E → A	Alternative
2	Secondary (Linkage)	E B	E → B	Alternative
3	(Conditional) Independence	E, A → B	E, A → B	Null
4	Relevance	E, A B	E, A → B	Alternative
5	Controlled	E, A B	E, A → B	Alternative

https://github.com/lingfeiwang/findr

Mediation-based causal inference fails in the presence of hidden confounders and weak regulations

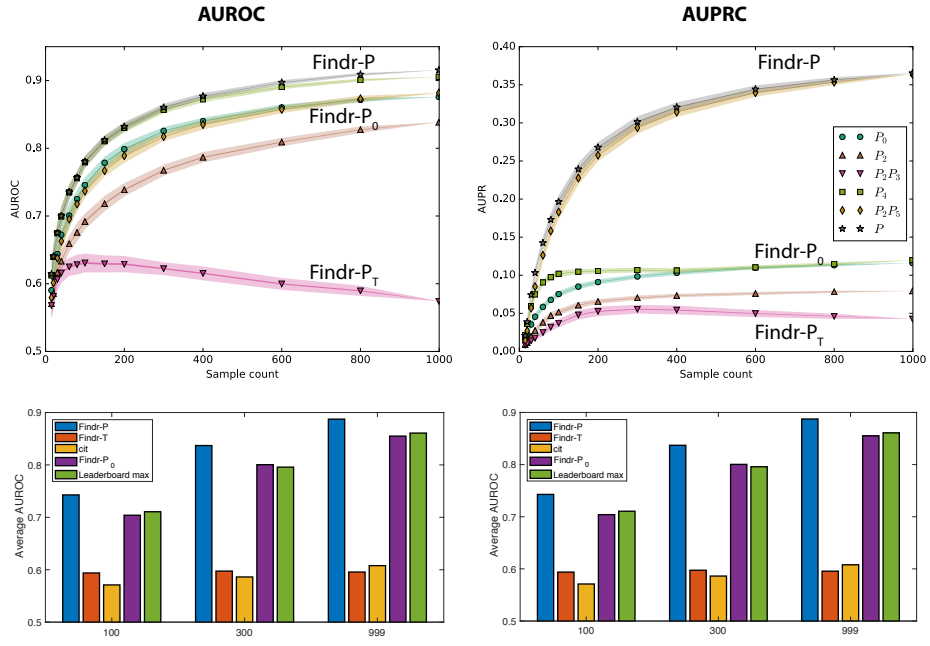
DREAM5 Systems Genetics Challenge
Synthetic networks (1000 genes), simulated data
(RILs and ODE model with *cis* and *trans* effects)



- Traditional test performs worse than correlation test.
- Traditional test performs worse with increasing sample size.
- Inclusion of conditional independence worse than secondary linkage alone.

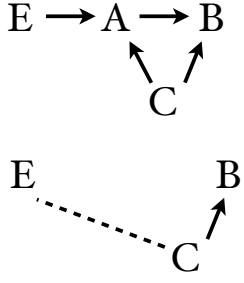
Mediation-based causal inference fails in the presence of hidden confounders and weak regulations

DREAM5 Systems Genetics Challenge
Synthetic networks (1000 genes), simulated data
(RILs and ODE model with *cis* and *trans* effects)



Hypothesis: elevated false negative rate (FNR) due to:

- Weak secondary linkage $E \to B$ in true $E \to A \to B$ relations.
- Conditional independence fails in the presence of confounders (common regulators) due to collider effect.

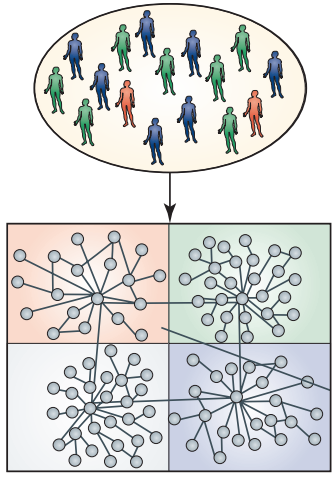


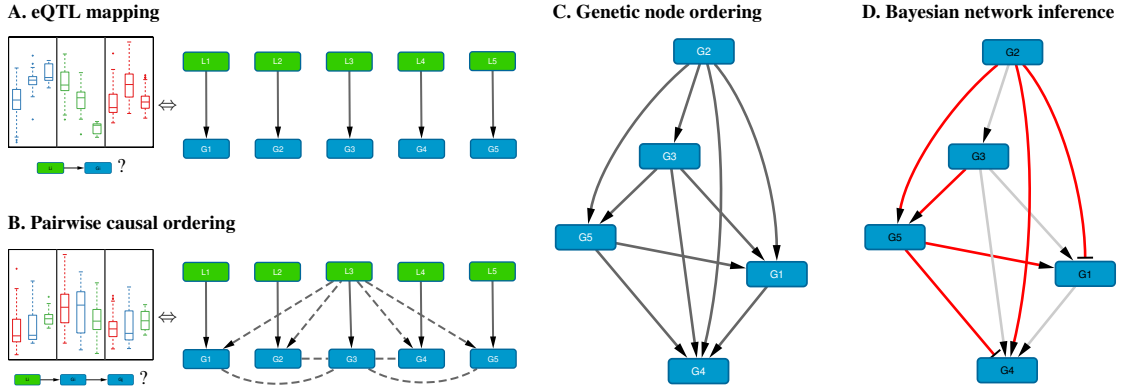
Reconstructing global causal Bayesian networks from pairwise causal inferences

Aim: Develop $O(n^2)$ / $O(n^2 \log n)$ algorithms for reconstructing Bayesian networks from genome-transcriptome variation data.

Steps:

- Calculate $P_{ij} = P(L_i \to X_i \to X_j)$ from causal inference test.
- Rank edges by P_{ij} and create DAG by adding ranked edges and incremental cycle detection ($O(m^{3/2})$ for m edges; $O(n^2 \log n)$ for all edges).
- Sparse Bayesian network from DAG via n independent variable selection problems.





[Lingfei Wang, Pieter Audenaert, TM. bioRxiv]

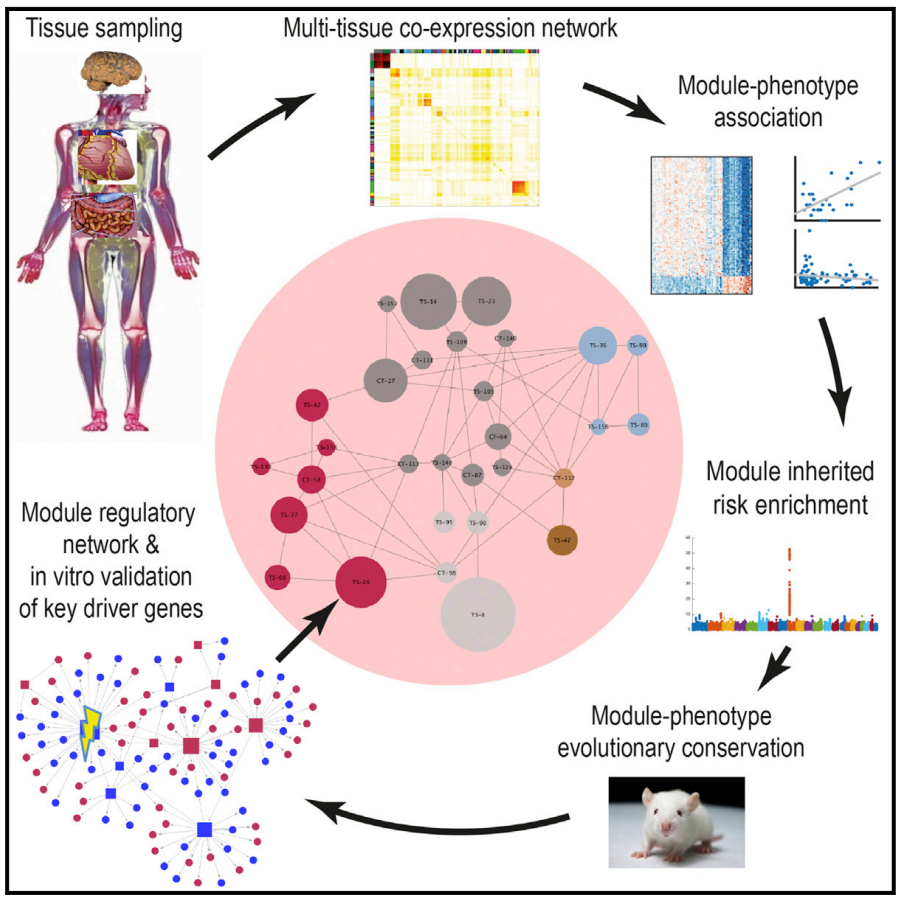
Part III

Using gene networks for candidate drug target discovery and drug repurposing

Cell Systems

Cross-Tissue Regulatory Gene Networks in Coronary Artery Disease

Graphical Abstract



Authors

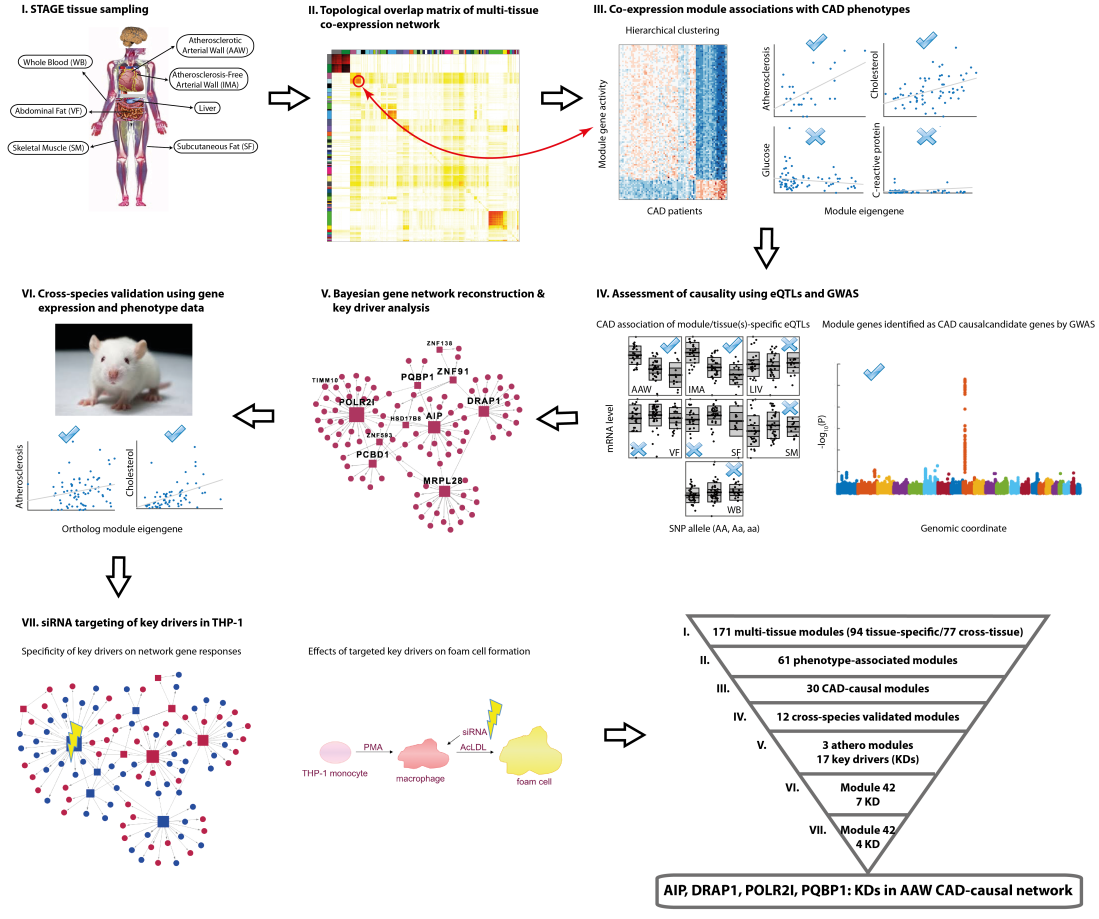
Husain A. Talukdar,
Hassan Foroughi Asl, Rajeev K. Jain, ...,
Josefin Skogsberg, Tom Michoel,
Johan L.M. Björkegren

Highlights

- We reconstruct regulatory gene networks across seven vascular and metabolic tissues
- Integrative analysis using GWASs reveals 30 networks causally related to CAD
- 12 CAD-causal networks are indicated to be evolutionarily conserved from mouse
- An atherosclerotic arterial wall RNA-processing network affects foam cell formation

Talukdar et al., 2016, Cell Systems *2*, 196–208

- Multi-tissue clustering identified 171 co-expression clusters (94 tissue-specific/77 cross-tissue).
- 61 clusters associated to key CAD phenotypes (athero, cholesterol, glucose, CRP).
- 30 CAD-causal clusters (CAD risk enriched eSNPs/GWAS genes).
- 12 clusters conserved in mouse with same phenotype association in same tissue.
- Key drivers of athero-causal Bayesian gene networks validated by siRNA targeting in THP-1 foam cells.



[Talukdar *et al*, Cell Systems (2016)]

SCIENTIFIC REPORTS

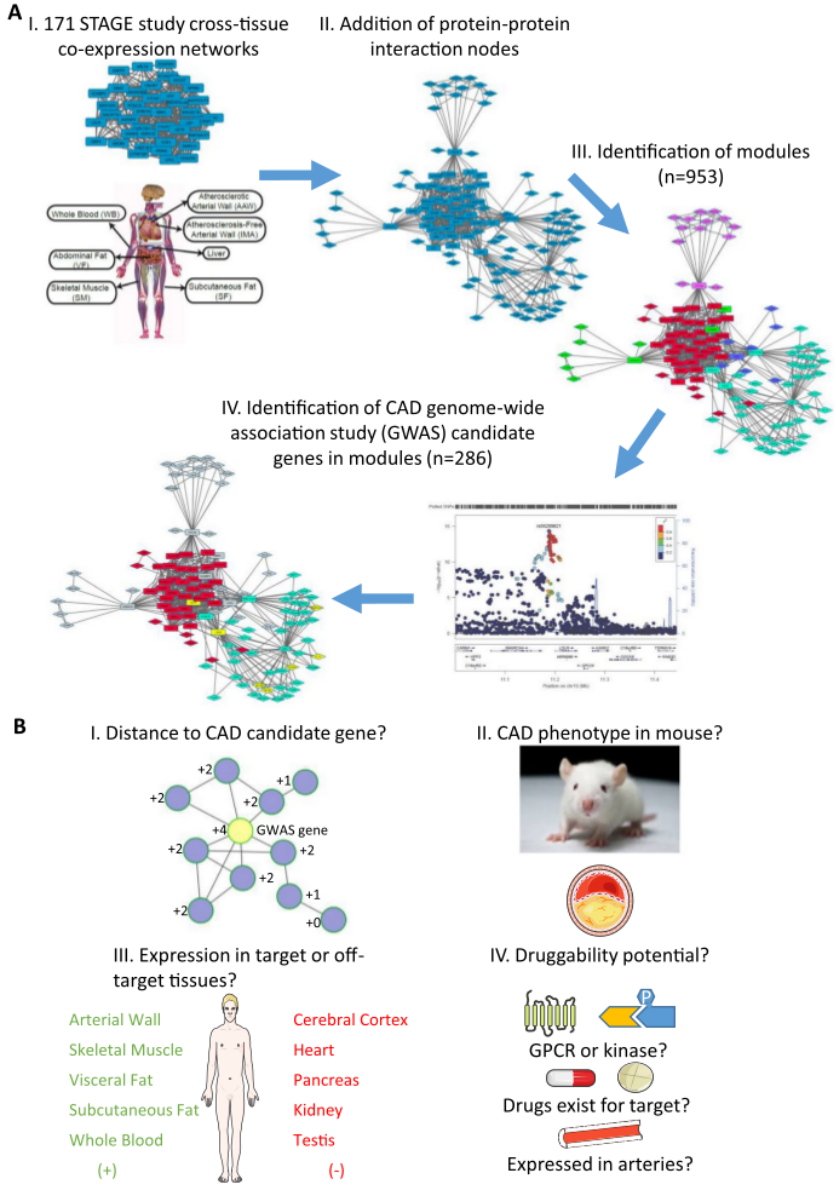
OPEN

Received: 8 August 2017
Accepted: 6 December 2017
Published online: 21 February 2018

Network analysis of coronary artery disease risk genes elucidates disease mechanisms and druggable targets

Harri Lempiäinen[1], Ingrid Brænne[2], Tom Michoel[3,4], Vinicius Tragante[5], Baiba Vilne[6,7], Tom R. Webb[8], Theodosios Kyriakou[9], Johannes Eichner[1], Lingyao Zeng[6], Christina Willenborg[2], Oscar Franzen[10], Arno Ruusalepp[4], Anuj Goel[9], Sander W. van der Laan[11], Claudia Biegert[1], Stephen Hamby[8], Husain A. Talukdar[12], Hassan Foroughi Asl[12], CVgenes@target consortium*, Gerard Pasterkamp[11,13], Hugh Watkins[9], Nilesh J. Samani[8], Timo Wittenberger[1], Jeanette Erdmann[2], Heribert Schunkert[6,7], Folkert W. Asselbergs[5,14] & Johan L. M. Björkegren[4,10,12]

ATC group	ATC group code	Significant modules	
		n	ID
Antineoplastic and immunomodulating agents	L	13	18_4, 130_2, 134_1, 82_4, 171_2, 108_3, 137_2, 126_2, 143_5, 91_4, 116_1, 130_3, 124_1
Cardiovascular system	C	4	36_4, 17_3, 84_3, 72_6
Musculoskeletal system	M	4	17_3, 82_4, 163_1, 69_5
Blood and blood-forming organs	B	2	84_3, 10_3
Nervous system	N	1	124_1
Genitourinary system and sex hormones	G	1	130_3
Sensory organs	S	1	163_1
Alimentary tract and metabolism	A	0	
Dermatologicals	D	0	
Systemic hormonal preparations, excluding sex hormones and insulins	H	0	
Anti-infectives for systemic use	J	0	
Antiparasitic products, insecticides, and repellents	P	0	
Respiratory system	R	0	
Various	V	0	



Circulation: Genomic and Precision Medicine

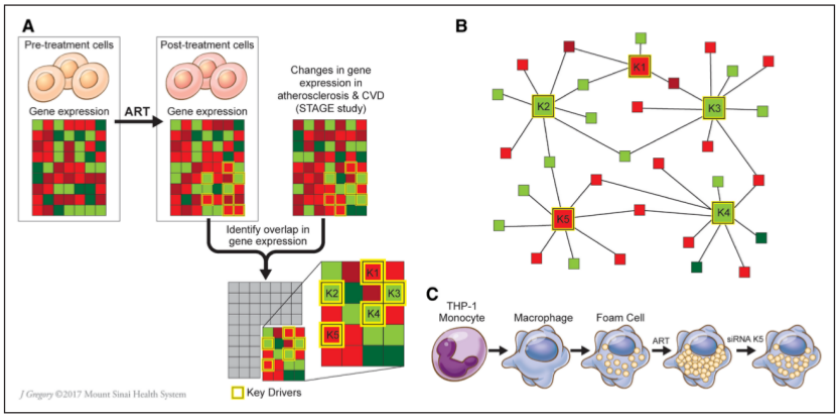
ORIGINAL ARTICLE

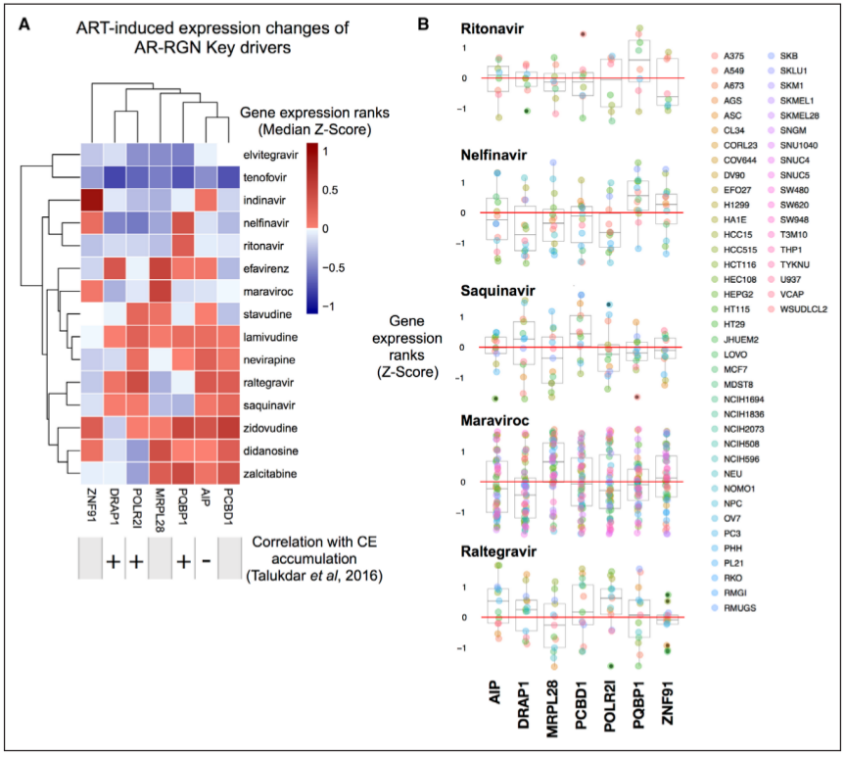
Systems Pharmacology Identifies an Arterial Wall Regulatory Gene Network Mediating Coronary Artery Disease Side Effects of Antiretroviral Therapy

BACKGROUND: Antiretroviral therapy (ART) for HIV infection increases risk for coronary artery disease (CAD), presumably by causing dyslipidemia and increased atherosclerosis. We applied systems pharmacology to identify and validate specific regulatory gene networks through which ART drugs may promote CAD.

METHODS: Transcriptional responses of human cell lines to 15 ART drugs retrieved from the Library of Integrated Cellular Signatures (overall 1127 experiments) were used to establish consensus ART gene/transcriptional signatures. Next, enrichments of differentially expressed genes and gene-gene connectivity within these ART-consensus signatures were sought in 30 regulatory gene networks associated with CAD and CAD-related phenotypes in the Stockholm Atherosclerosis Gene Expression study.

Itziar Frades, PhD*
Ben Readhead, MBBS*
Letizia Amadori, PhD*
Simon Koplev, MScEng
Husain A. Talukdar, PhD
Heidi M. Crane, MD, MPH
Paul K. Crane, MD, MPH
Jason C. Kovacic, MD, PhD
Joel T. Dudley, PhD
Chiara Giannarelli, MD, PhD
Johan L.M. Björkegren, MD, PhD†
Inga Peter, PhD†





Relevant databases and resources

Druggability

- Protein kinase coding genes: Uniprot
- G-protein-coupled receptors: Guide to pharmacology
- Protein-protein interactions: ConsensusPathDB, and many others
- Drug-gene interactions: DGIdb, Guide to pharmacology
- ...

Drug repurposing

- Connectivity map
- SigCom LINCS

Gene set enrichment

- Enrichr
- ...