Genome-wide Association Studies of Substance Use and Use Disorder

Where to find them, and what to do with them

Presented By Dr. Alexander S. Hatoum

Twitter handle: @AlexanderHatoum

Funded by NIDA Neuroscience T32 DA007261 and Consulted for *All of Us;*

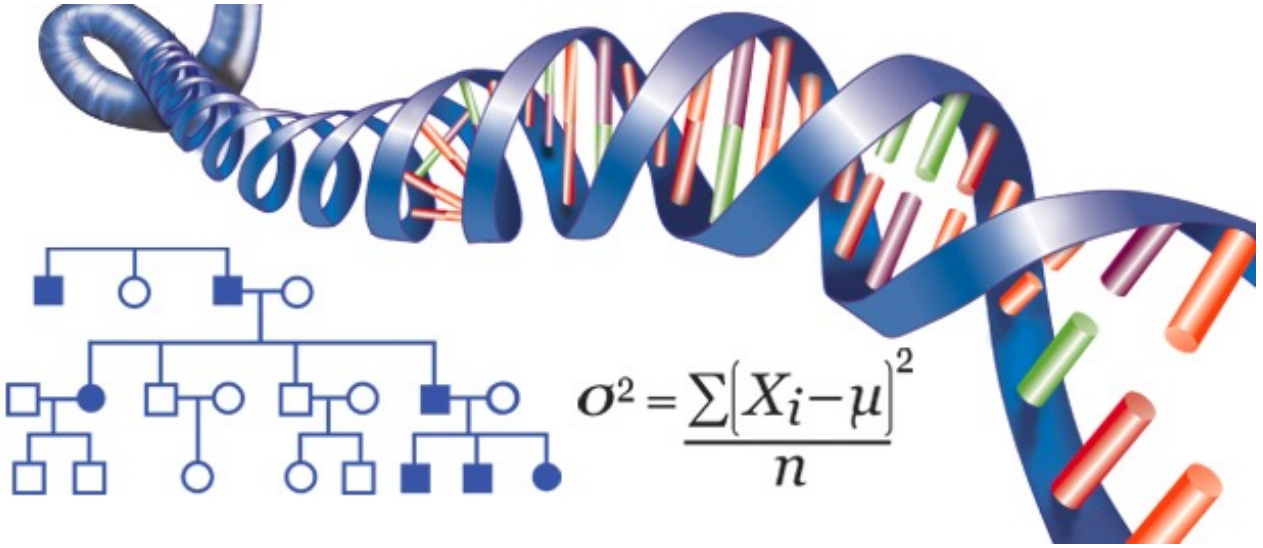
Other funding acknowledgements: K02DA032573; DA54869 (Agrawal)

Roadmap

- Part 1: GWAS data for Substance Use/Use Disorder
  - GWAS background
  - Accessing summary data
- Part 2: Bioinformatics with GWAS summary data
  - Levels of Analysis
  - Online resources
- Part 3: Whole-Genome Analyses
  - Pathway Analysis
  - Genetic Correlations
  - Latent Causal Variable Analysis

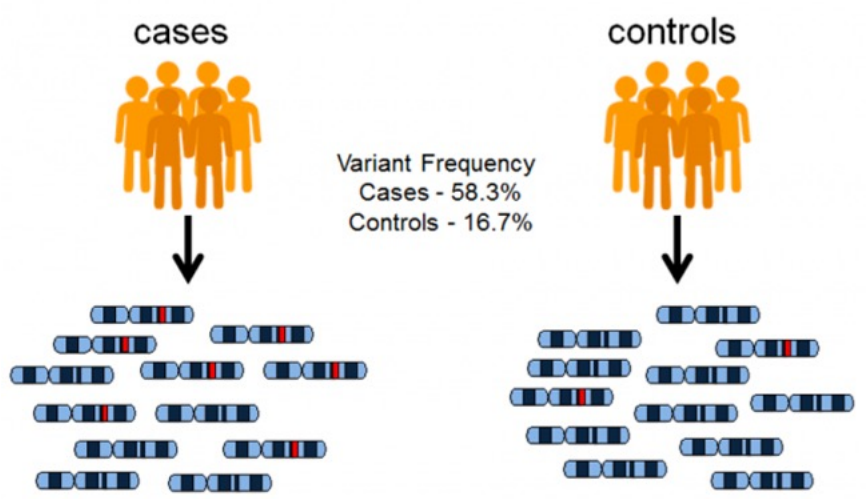
Genetic Epidemiology

- The process of tracking disease via inheritance of measured genotype
- Broadly refers to methods used in human genetics, where samples must be taken from the population and therefore benefit from epidemiological designs
- Includes twin/family studies, genome-wide association studies, and sequencing.
- Suite of tools to integrate with large-scale bioinformatics to make sense of results.



Genome-Wide Association Studies (GWAS)

- Genetic searches, where a linear model is run at genotyped (array) SNPs across the human genome to detect signals/peaks of association;
- Readily integrates with bioinformatics to ask questions about biology, and epidemiology to asks questions about disease (i.e., causality and risk)

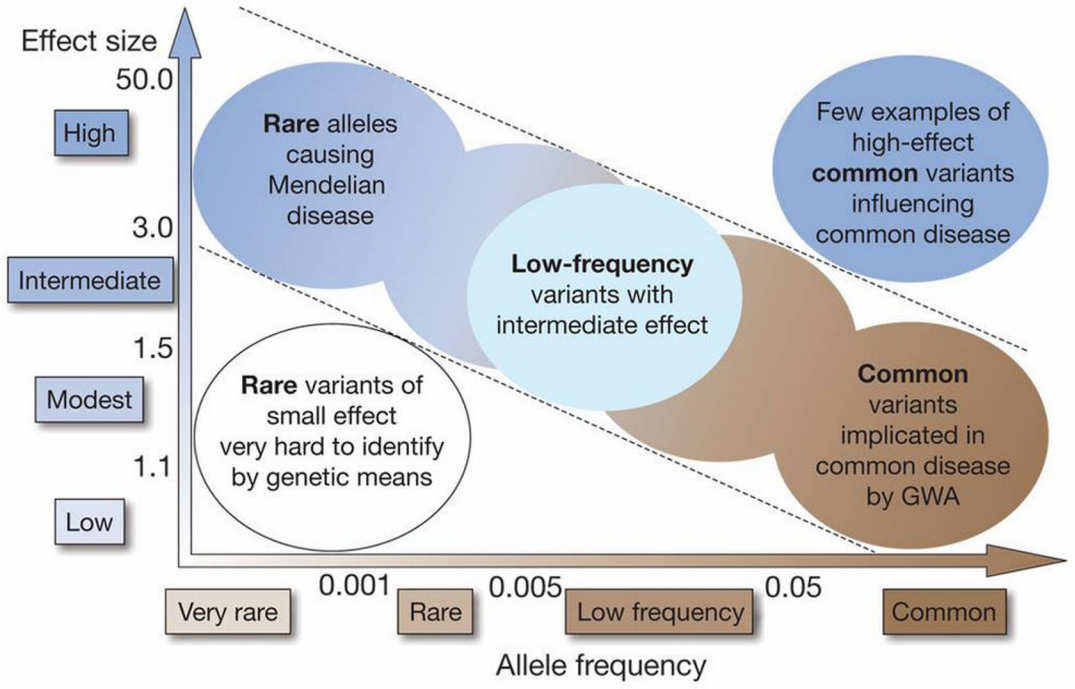


$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$$

Y : Phenotype
β_0 : Intercept
β_i : Covariates (Ancestry)
X = Minor Allele Frequency

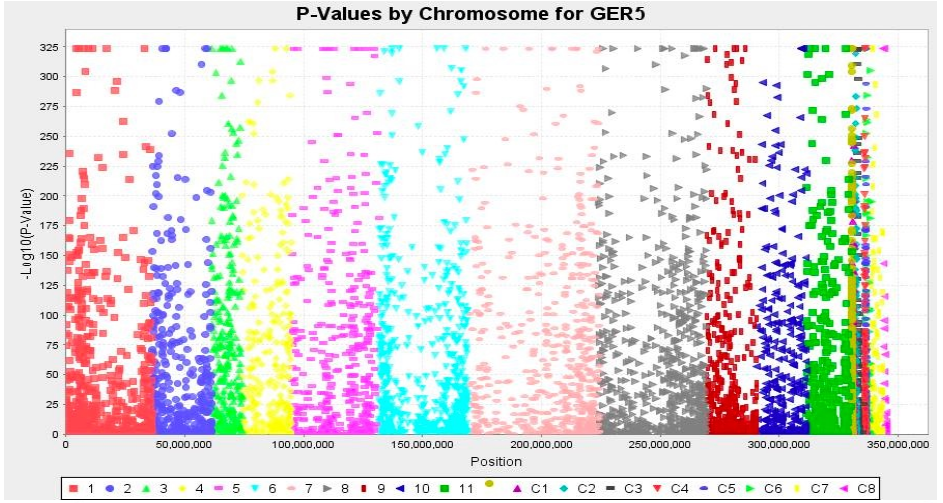
GWAS: Quick Reminder

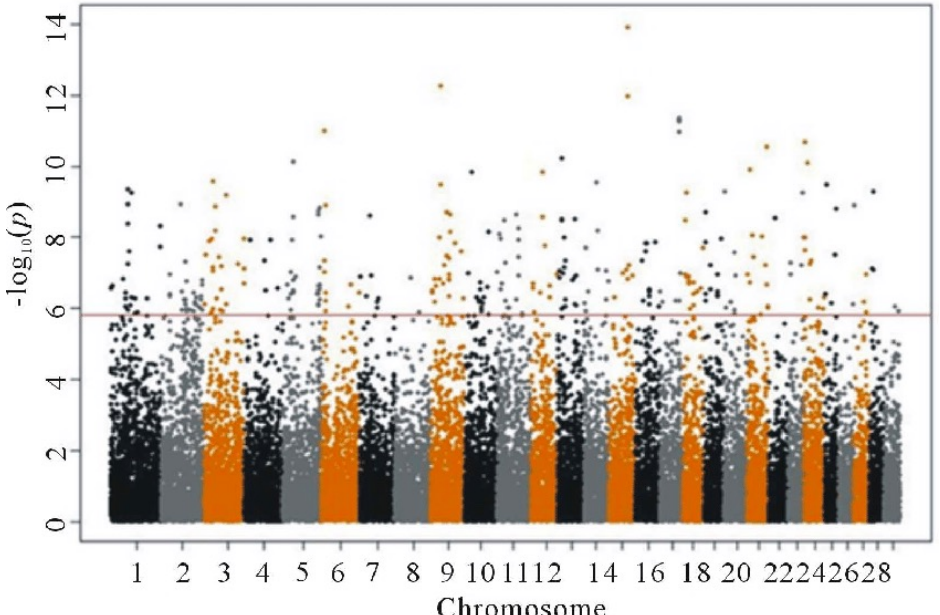
- Most use GWAS arrays
  - About 7 million common variants
  - Many are imputed
- Common variants have relatively small effects (i.e., polygenic)
  - Variants are largely intronic
    - Some exceptions with substance use disorders/substance use;
- Sample size: Want >100K; min >10K.



Standards for trustworthy GWAS

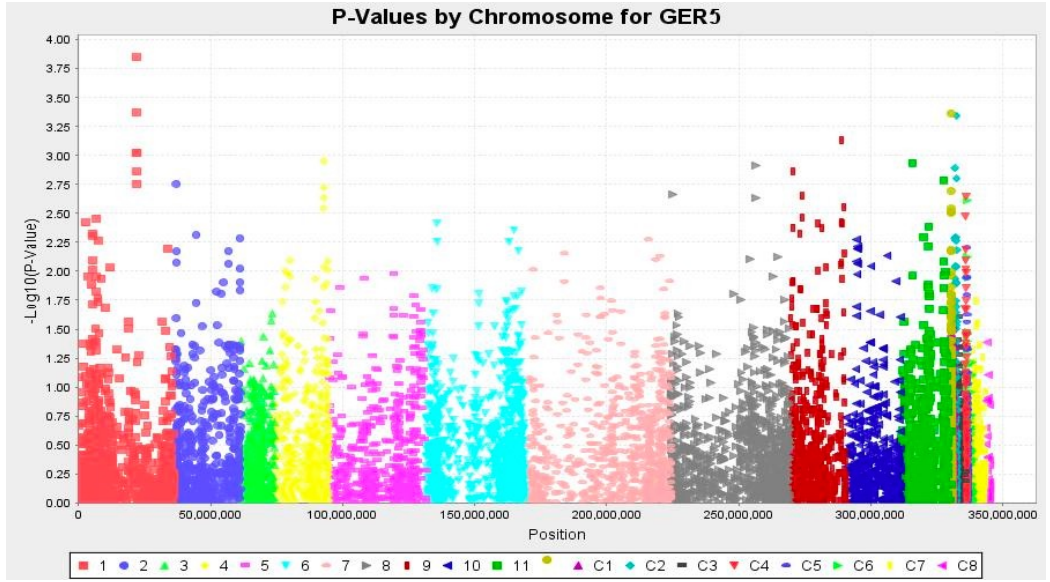
- Control for ancestry
  - Split into separate major geographic populations (African, European
    - Principal components within each group
- Control for confounds/batch effects
- Use ancestry “tracts”
  - Only Tractor does this, and these sumstats don’t work with most bioinformatic pipelines (yet)

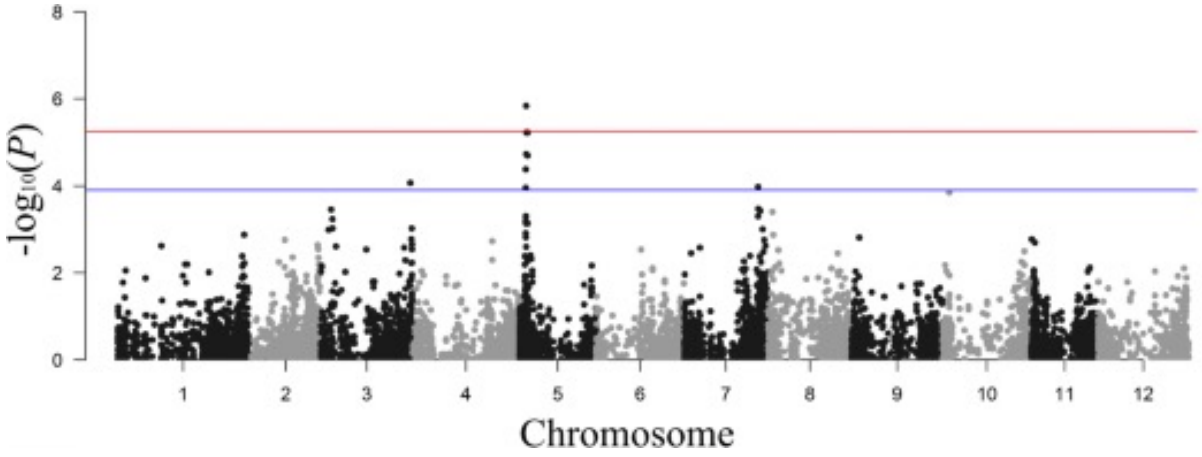




Standards for trustworthy GWAS

- Control for ancestry
  - Split into separate major geographic populations (African, European
    - Principal components within each group
- Control for confounds/batch effects
- Use ancestry “tracts”
  - Only Tractor does this, and these sumstats don’t work with most bioinformatic pipelines (yet)





Getting GWAS results

What to look for and where to find them

What you need depends on your question:

- Most analyses are possible with results files;
  - Gene-based, pathway, network models
  - Prioritizing SNPs based on eQTL, epigenome data
  - Causal analyses
  - Multivariate analyses (e.g., common genetic factor underlying multiple SUDs)
  - Conditional analyses (e.g., loci associated with alcohol when controlling for tobacco genetics)
  - Polygenic risk score weights
- Few require raw/individual level data (e.g., coding polygenic risk scores in a specific dataset; re-doing the GWAS with other covariates);

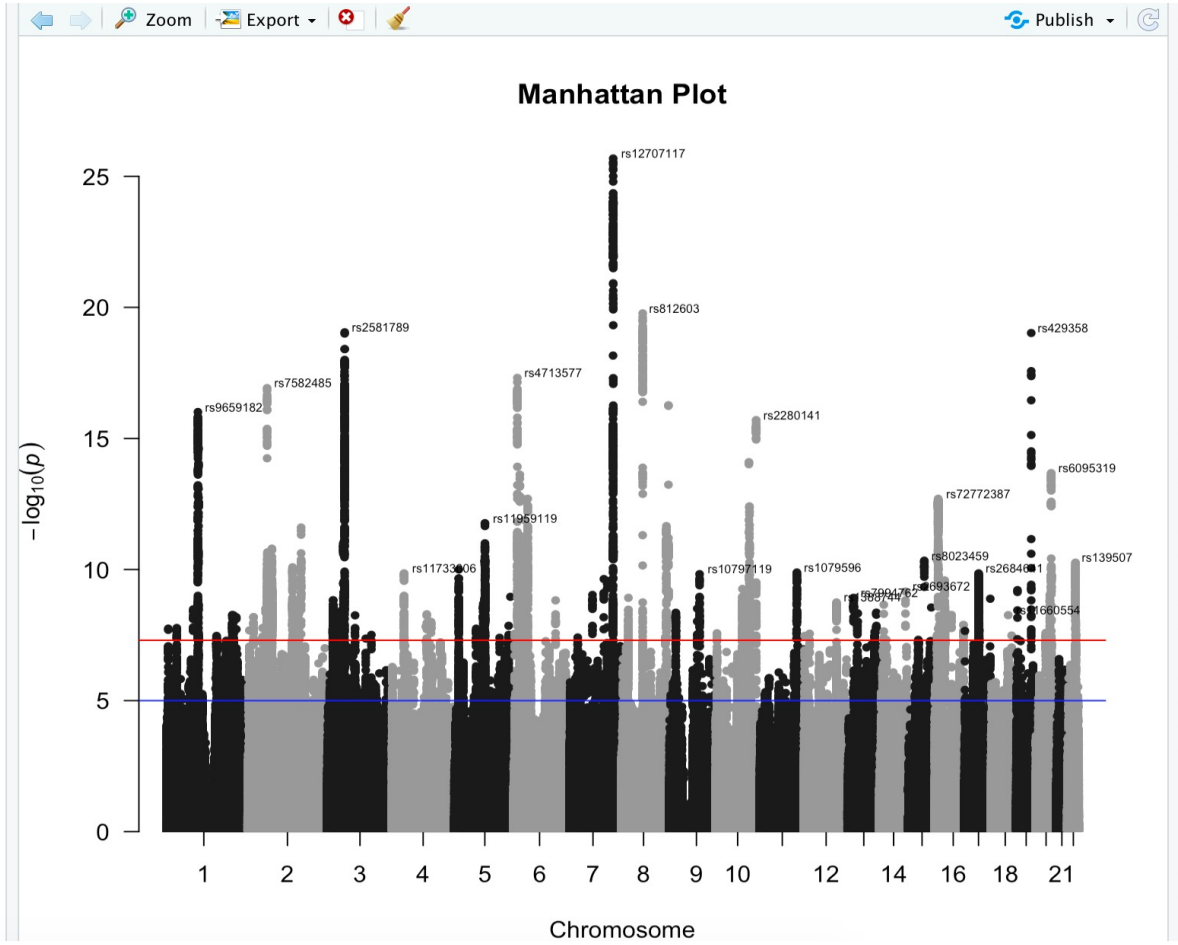
GWAS results sources

- GWAS researchers are (typically) very open to sharing summary statistics from their studies:
  - Many GWAS are available via the GWAS catalogue: https://www.ebi.ac.uk/gwas/
  - Neale lab has run ~1500 traits from UKBiobank: http://www.nealelab.is/uk-biobank

- Best source in substance use/use disorders is from **consortia webpages**. All results can be downloaded immediately (typically post publication) and contain helpful README's;

What you will get (and it is all you need!)

	SNP	Effect	Other_allele	EAF	Beta	SE
1	rs28544273	T	A	0.877114	8.57115e-05	0.00173069
2	rs28527770	T	C	0.877007	3.77144e-05	0.00172833
3	rs3115860	C	A	0.129130	-9.93430e-04	0.00168405
4	rs3131970	T	C	0.124491	-4.54247e-05	0.00171218
5	rs2073813	G	A	0.871339	9.04270e-04	0.00168808
6	rs3131969	A	G	0.129550	-9.84466e-04	0.00168215
7	rs3131968	A	G	0.129453	-1.06321e-03	0.00168283
8	rs3131967	T	C	0.129551	-9.91024e-04	0.00168212
9	rs3115858	A	T	0.129471	-9.93259e-04	0.00167916
10	rs61768170	G	C	0.873968	8.83057e-04	0.00171311
11	rs3131962	A	G	0.129900	-1.09204e-03	0.00167511
12	rs3115853	G	A	0.130620	-1.06686e-03	0.00167357
13	rs4951929	C	T	0.129757	-1.14408e-03	0.00167647
14	rs4951862	C	A	0.129752	-1.14191e-03	0.00167659
15	rs3131956	A	G	0.129744	-1.14928e-03	0.00167663
16	rs3131954	C	T	0.129300	-9.97044e-04	0.00168099
17	rs3115851	T	A	0.125364	-7.64427e-05	0.00170628
18	rs2286139	C	T	0.136005	-1.25929e-03	0.00167285
19	rs1057213	C	T	0.130610	-9.50365e-04	0.00168934
20	rs3115849	G	A	0.133629	-9.70695e-04	0.00168971
21	rs3115848	G	C	0.128421	-6.14764e-04	0.00170931
22	rs3131950	C	G	0.128421	-6.14729e-04	0.00170931
23	rs3131949	T	C	0.128423	-6.10564e-04	0.00170934
24	rs3131948	T	A	0.128053	-5.58076e-04	0.00171028
25	rs7515915	T	G	0.874414	3.58030e-04	0.00171125
26	rs61768174	A	C	0.894481	9.60609e-04	0.00186206
27	rs2977608	A	C	0.236358	-2.24961e-03	0.00132648
28	rs12562034	G	A	0.894405	2.65774e-03	0.00182884
29	rs60320384	C	G	0.871352	1.23661e-03	0.00168558
30	rs7518545	G	A	0.895126	2.53517e-03	0.00184281
31	rs371458725	G	A	0.896041	2.47603e-03	0.00185709
32	rs2977605	T	C	0.129930	-1.25050e-03	0.00167797
33	rs59066358	G	A	0.871294	1.21080e-03	0.00168477
34	rs2905039	A	C	0.129916	-1.25721e-03	0.00167802
35	rs28810152	A	C	0.130001	-1.21070e-03	0.00169128



Major Sources: Psychiatric Genomics Consortium

- Genome-wide meta-analysis consortium that focuses on substance use disorders (and many other psychiatric disorders)
  - med.unc.edu/pgc/download-results/
- Alcohol Use Disorders (not the largest; PMID: 30482948)
  - Walters et al., N case = 14,904 N control = 37,944
- Cannabis Use Disorder (PMID: 33096046)
  - Johnson et al. 2020, N case= 20,916 N control = 363,116
- Opioid Use Disorder (not the largest; PMID: 32099098)
  - Exposed and unexposed controls,
    - Polimanti et al. 2020, N case = 4,503 N exposed control = 4,173 N unexposed control = 32,500

Major Sources: GSCAN consortium

- Genome-wide traits in substance use
  - https://conservancy.umn.edu/handle/11299/201564
- Phenotypes from Liu et al. 2019 PMID: 30643251
  - Cigarettes per day, N = 337,334
  - Drinks per week, N = 914,280
  - Ever smoking, N = 1,232,091
  - Smoking cessation, N = 547,219
  - Age of Smoking Initiation, N = 341,427

Major Sources: International Cannabis Consortium

- Two Cannabis initiation behaviors available here:
  - https://www.ru.nl/bsi/research/group-pages/substance-use-addiction-food-saf/vm-saf/genetics/international-cannabis-consortium-icc/
  - Not very many hits yet
- Cannabis Ever Use (PMID: 30150663)
  - Pasman et al. 2018, N = 184,765
- Age of First Cannabis Use (PMID: 25987507)
  - Minica et al. 2018, N = 24,953

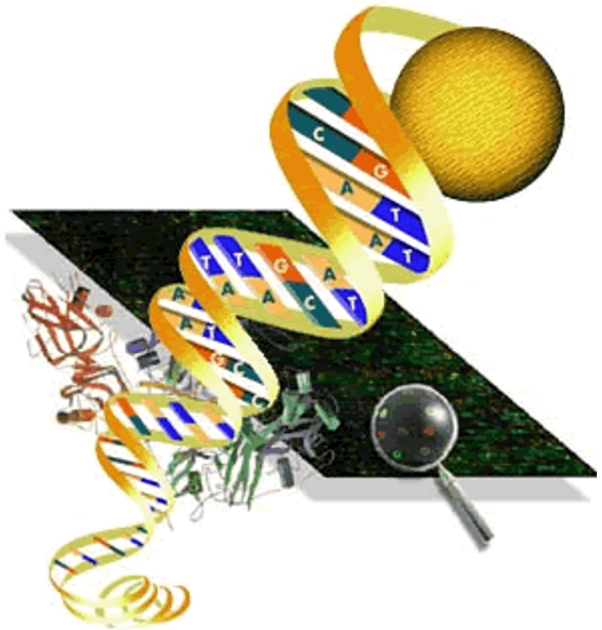
Major Sources: dbGaP Million Veterans Program

- Some of the largest GWAS of substance use disorders;
- https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001672.v3.p1

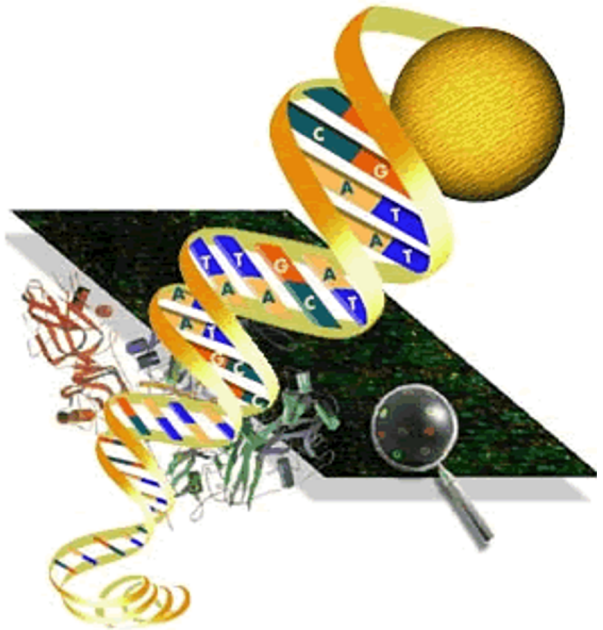
- Must be registered as a “PI” in your institutional eRA account;
- Apply for data (scope of project etc); annual review; no sharing beyond application;
- Tobacco trajectories
  - Xu (PMID: 33082346)
- Alcohol Use Disorder
  - Kranzler (PMID: 30940813) & Zhou (PMID: 32451486)
- Opioid Use Disorder
  - Zhou (PMID: 32492095)

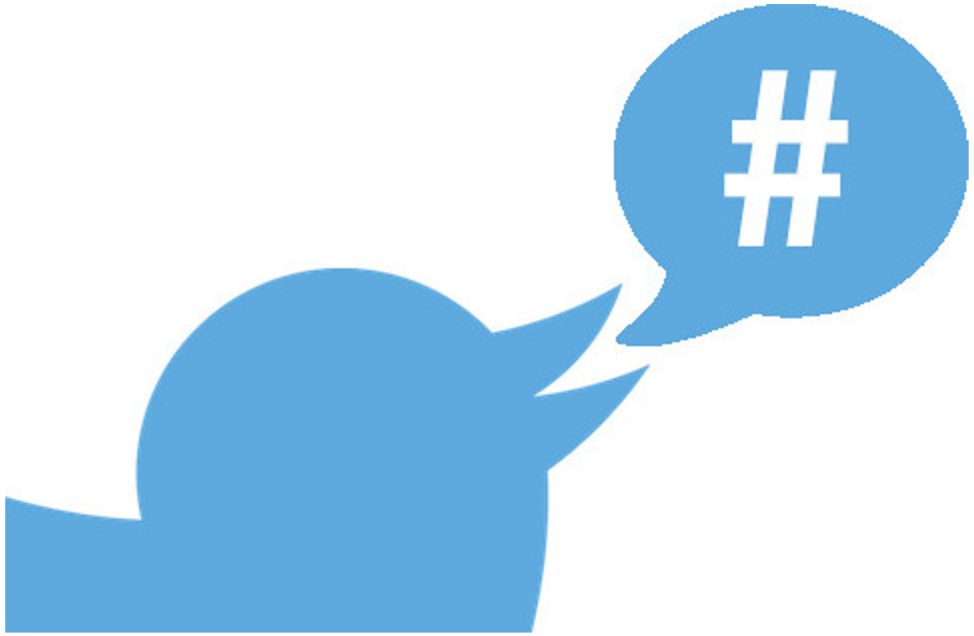
Part 2: Making sense of GWAS summary statistics

Aligning summary stats with other data
Annotation



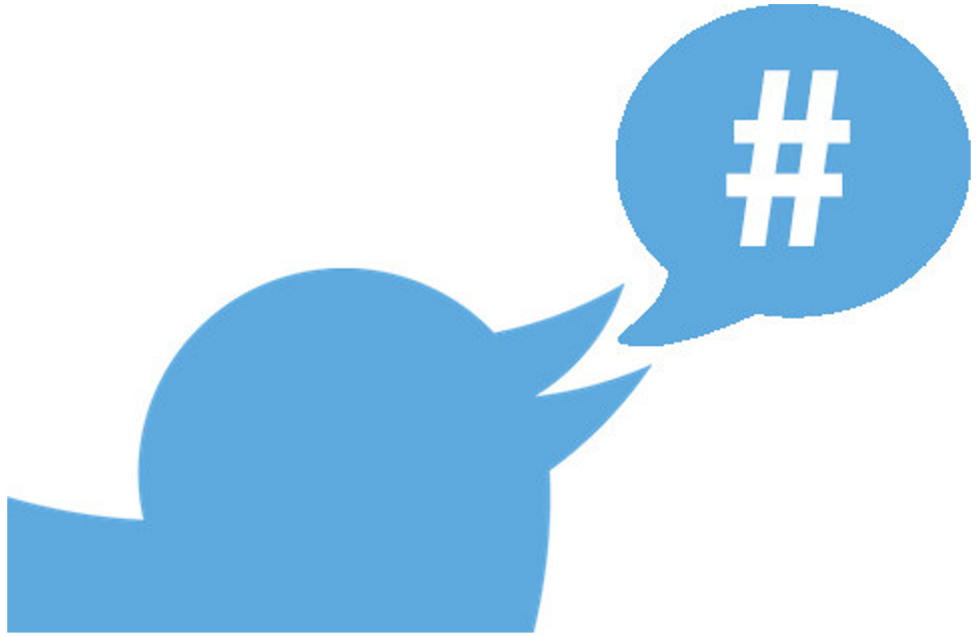
Making sense from results





Making sense from results: Annotation

- Nearest gene;
- Functional
  - Intronic, Exonic, etc.
- enhancer promoter, eQTL, and phenotypic associations



FUMA and MASSIVE contain most summary statistics

- Functional Mapping and Annotation of Genome-Wide Association Studies PMID: 29184056 ;
  - https://fuma.ctglab.nl/

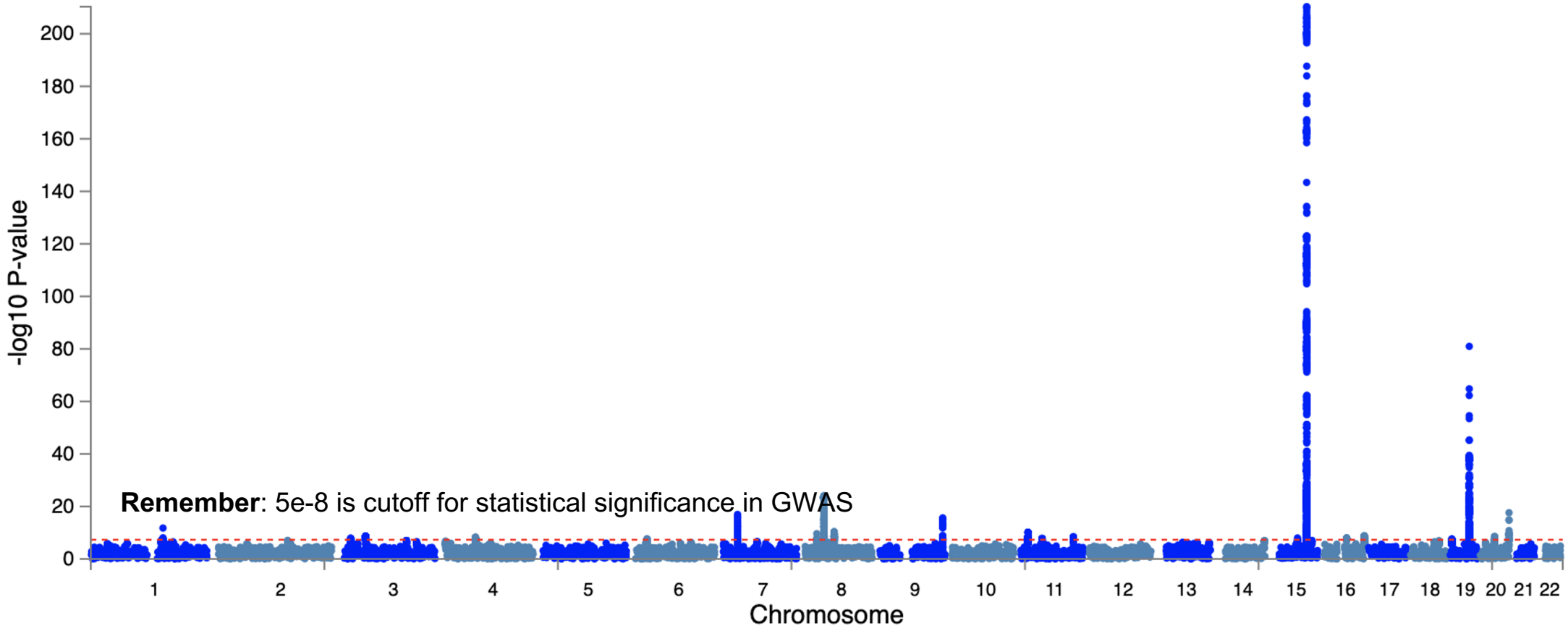
- MASSIVE
  - Genoma.io

Illustration for talk: Cigarettes Per Day

Cigarettes per day (Liu et al. 2019; PMID: 30643251)

N=263,954

Good-looking Manhattan plot for CPD



Levels of Evidence and FUMA

All you can do with FUMA!

- Functional Mapping and Annotation of Genome-Wide Association Studies
- SNP positional annotation
  - Identifying independent and lead loci
  - Mapping loci to function
  - Mapping Loci to genes via functional annotation
  - Mapping loci to eQTLs
  - Filtering, such as CADD and regulome, Hi-C
- Pathway enrichment
  - Gene-set enrichment analysis (GSEA via MAGMA)
    - Pathways and single-cell enrichment databases
  - Gene enrichment analysis (drug bank for example, but 100’s of others)

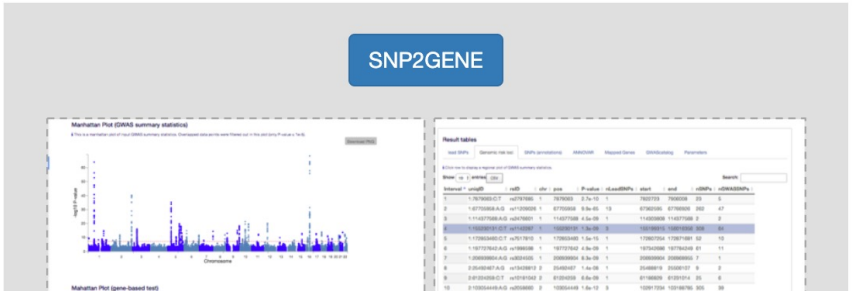
FUMA GWAS

Functional Mapping and Annotation of Genome-Wide Association Studies

FUMA is a platform that can be used to annotate, prioritize, visualize and interpret GWAS results.
The SNP2GENE function takes GWAS summary statistics as an input, and provides extensive functional annotation for all SNPs in genomic areas identified by lead SNPs.
The GENE2FUNC function takes a list of gene IDs (as identified by SNP2GENE or as provided manually) and annotates genes in biological context
To submit your own GWAS, login is required for security reason. If you have't registered yet, you can do from here.
You can browse public results of FUMA (including example jobs) from Browse Public Results without registration or login.

Please post any questions, suggestions and bug reports on Google Forum: FUMA GWAS users.
If you would like to be in the mailing list, please send an email to k.watanabe@vu.nl. Only major updates will be announced through email (low traffic).

Citation:
When using FUMA, please cite the following.
K. Watanabe, E. Taskesen, A. van Bochoven and D. Posthuma. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**:1826. (2017).
https://www.nature.com/articles/s41467-017-01261-5
When using cell type analysis, please cite the following.
K. Watanabe, M. Umicevic Mirkov, C. de Leeuw, M. van den Heuvel and D. Posthuma. Genetic mapping of cell type specificity for complex traits. *Nat. Commun.* **10**:3222. (2019).
https://www.nature.com/articles/s41467-019-11181-1
Depending on which results you are going to report, please also cite the original study of data sources/tools used in FUMA (references are available at links or tutorial for the cell type specificity analysis for scRNA-seq data).



Result tables

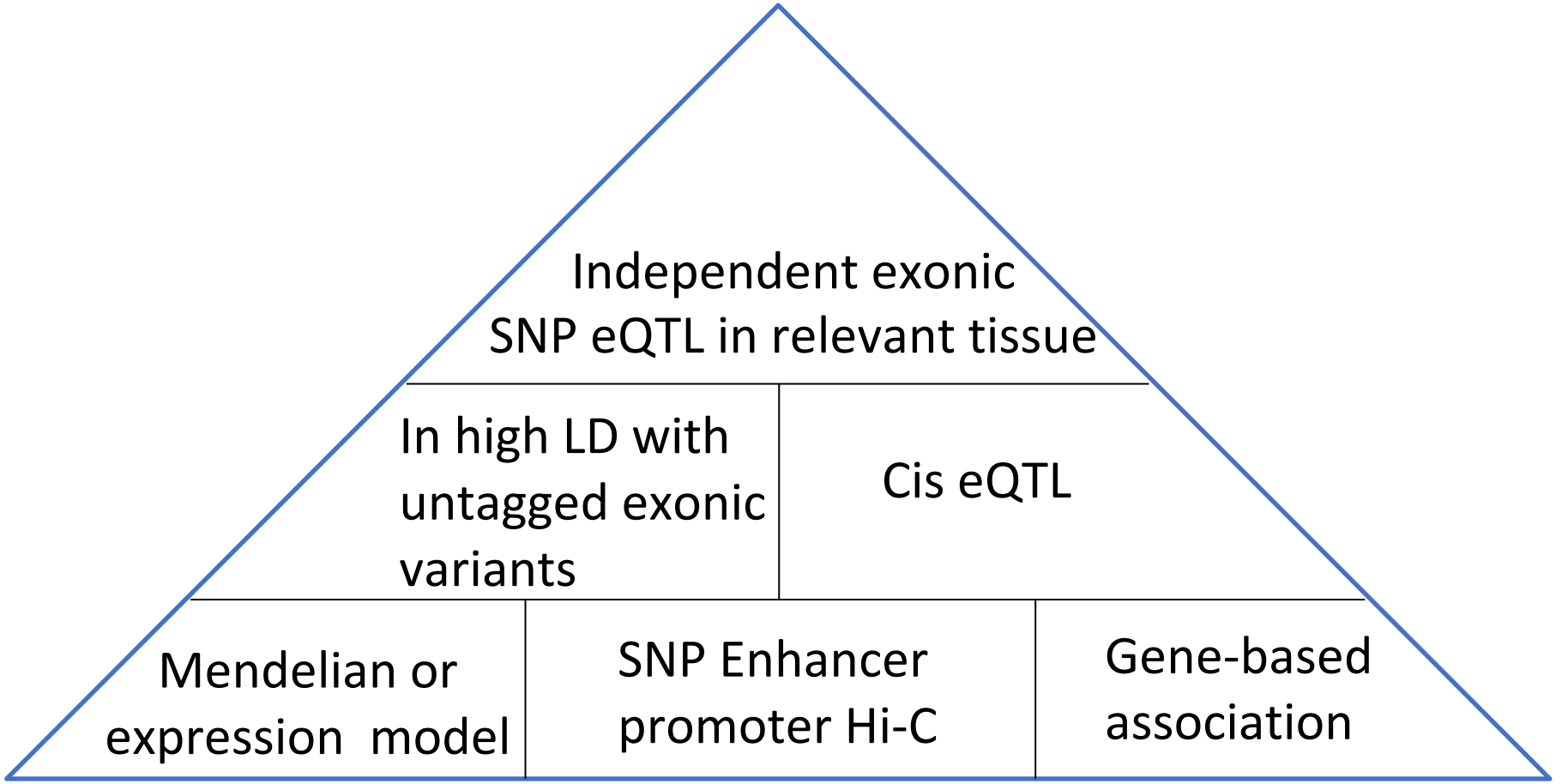
Genomic risk loci | lead SNPs | Ind. Sig. SNPs | SNPs (annotations) | ANNOVAR | Mapped Genes | eQTL | Chromatin interactions | GWAScatalog | Parameters

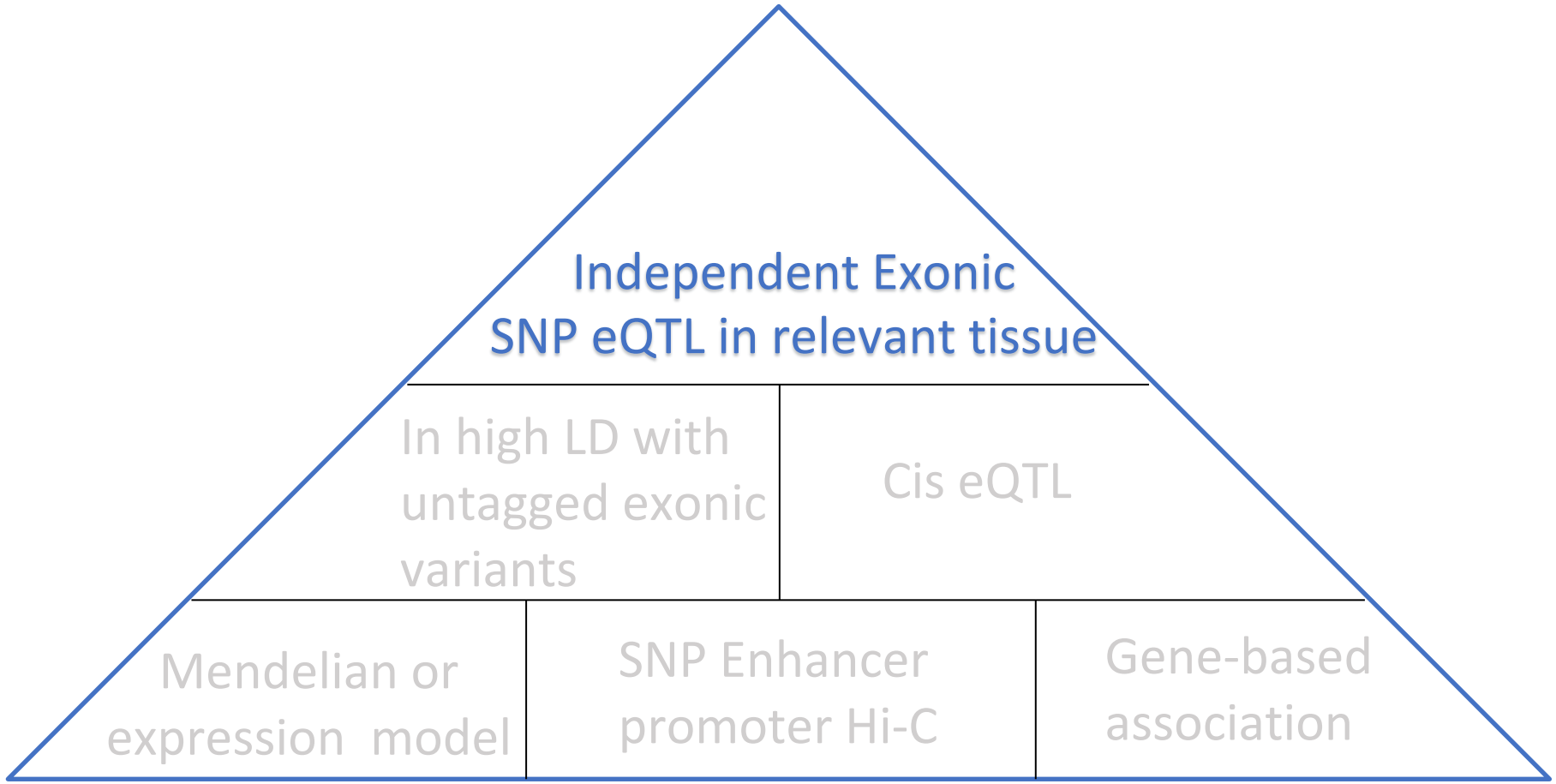
ℹ Click row to display a regional plot of GWAS summary statistics.

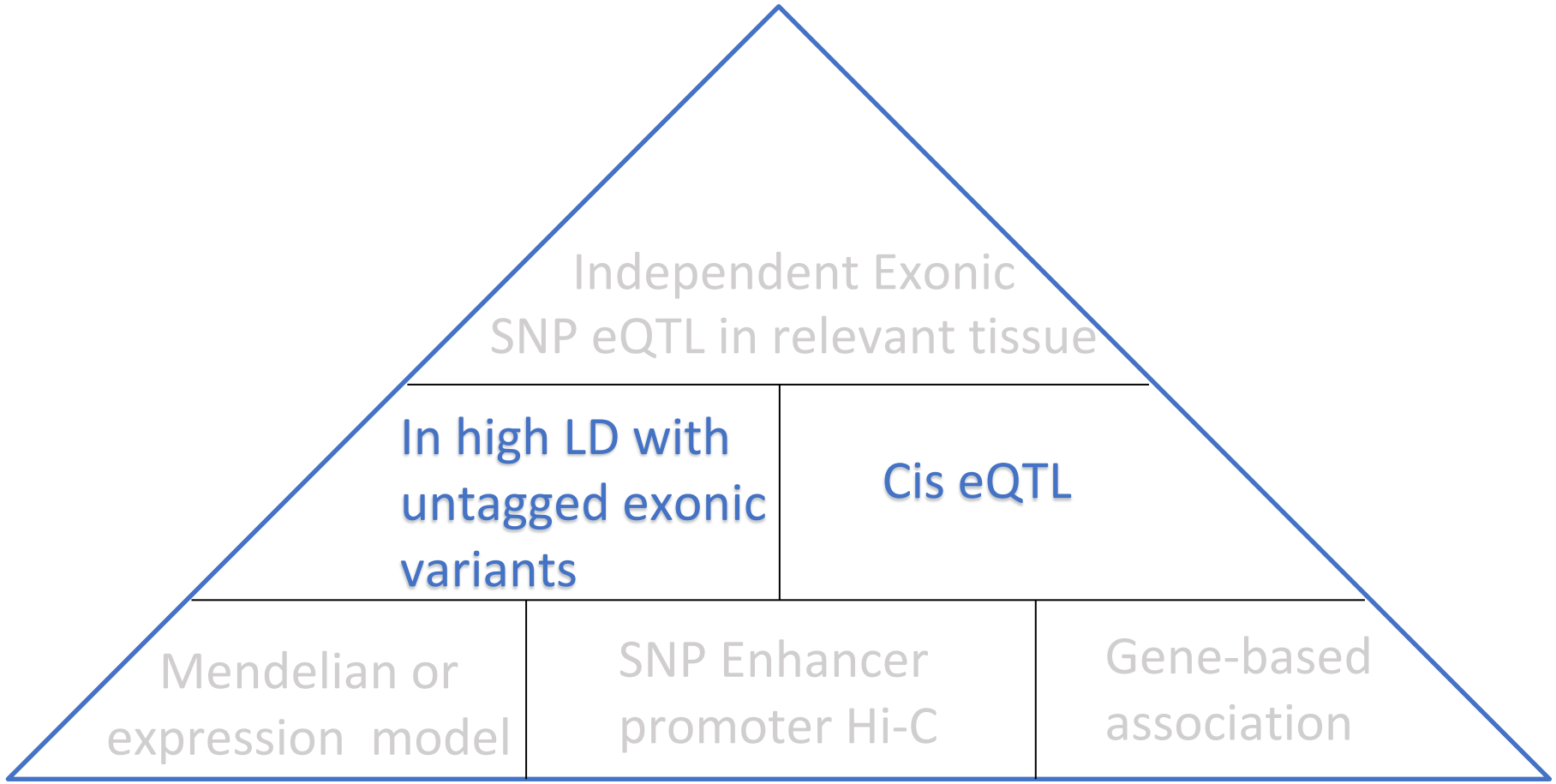
Show 10 entries Search:

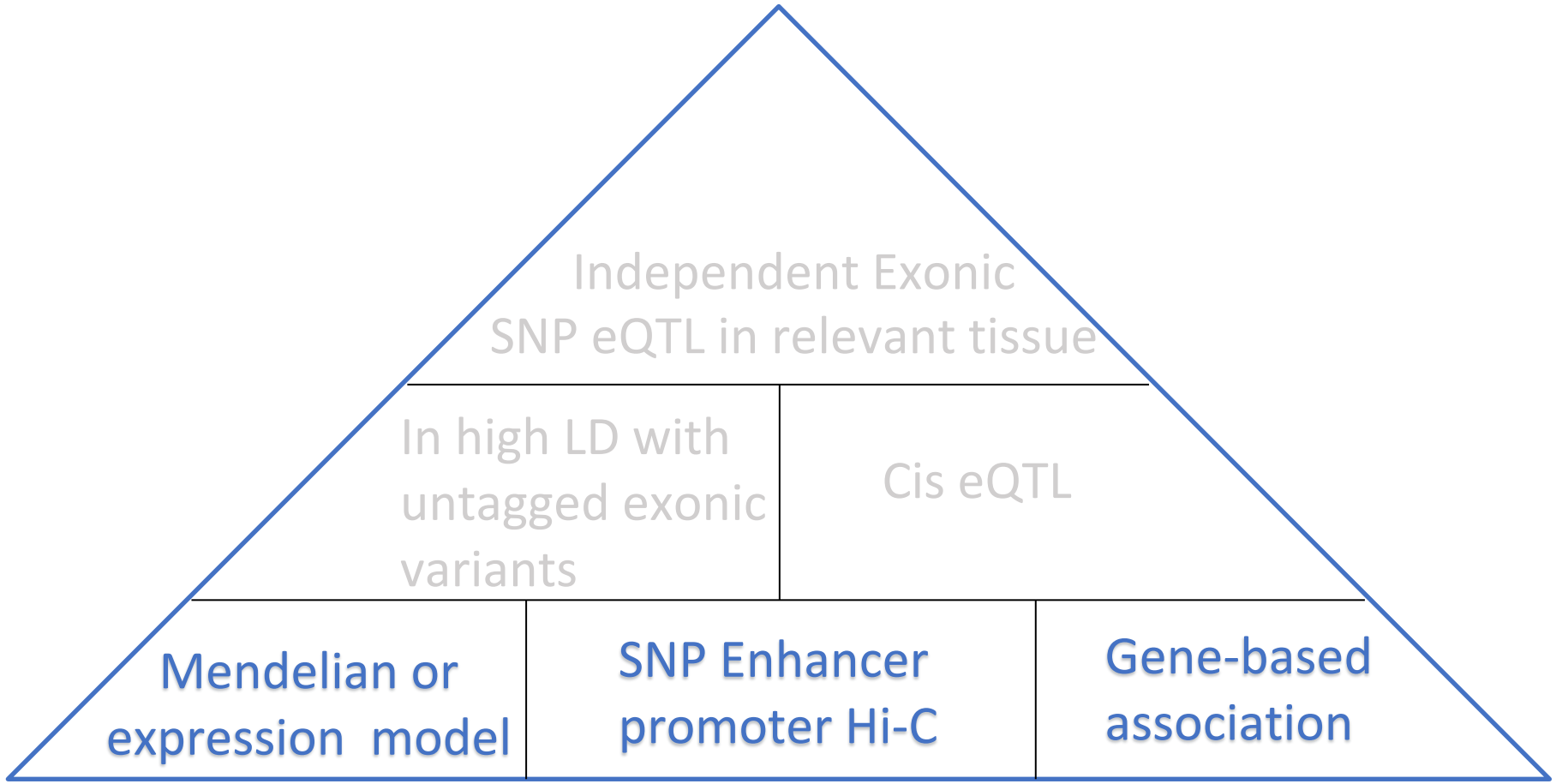
Genomic Locus	uniqID	rsID	chr	pos	P-value	start	end	nSNPs	nGWASSNPs	nIndSigSNPs	IndSigSNPs
1	1:154548521:C:G	rs2072659	1	154548521	1.71e-12	154548521	154618623	13	12	2	rs2072659;rs4
2	3:16872929:C:T	rs2084533	3	16872929	1.22e-08	16846967	16879840	46	45	1	rs2084533
3	3:48935583:A:G	rs7431710	3	48935583	1.82e-09	48719638	49575913	197	143	1	rs7431710
4	4:67053769:C:T	rs11725618	4	67053769	4.67e-09	67053769	67108122	33	27	1	rs11725618
5	4:67891641:C:T	rs13141210	4	67891641	2.77e-08	67792065	67906728	123	105	1	rs13141210
6	6:26214473:C:T	rs806798	6	26214473	2.48e-08	26164824	26302573	53	45	2	rs806798;rs77
7	7:32333642:A:G	rs215600	7	32333642	1.1e-17	32255608	32450072	236	195	7	rs215600;rs28
8	8:27442127:A:C	rs73229090	8	27442127	2.44e-10	27406353	27453579	24	17	3	rs2741351;rs7

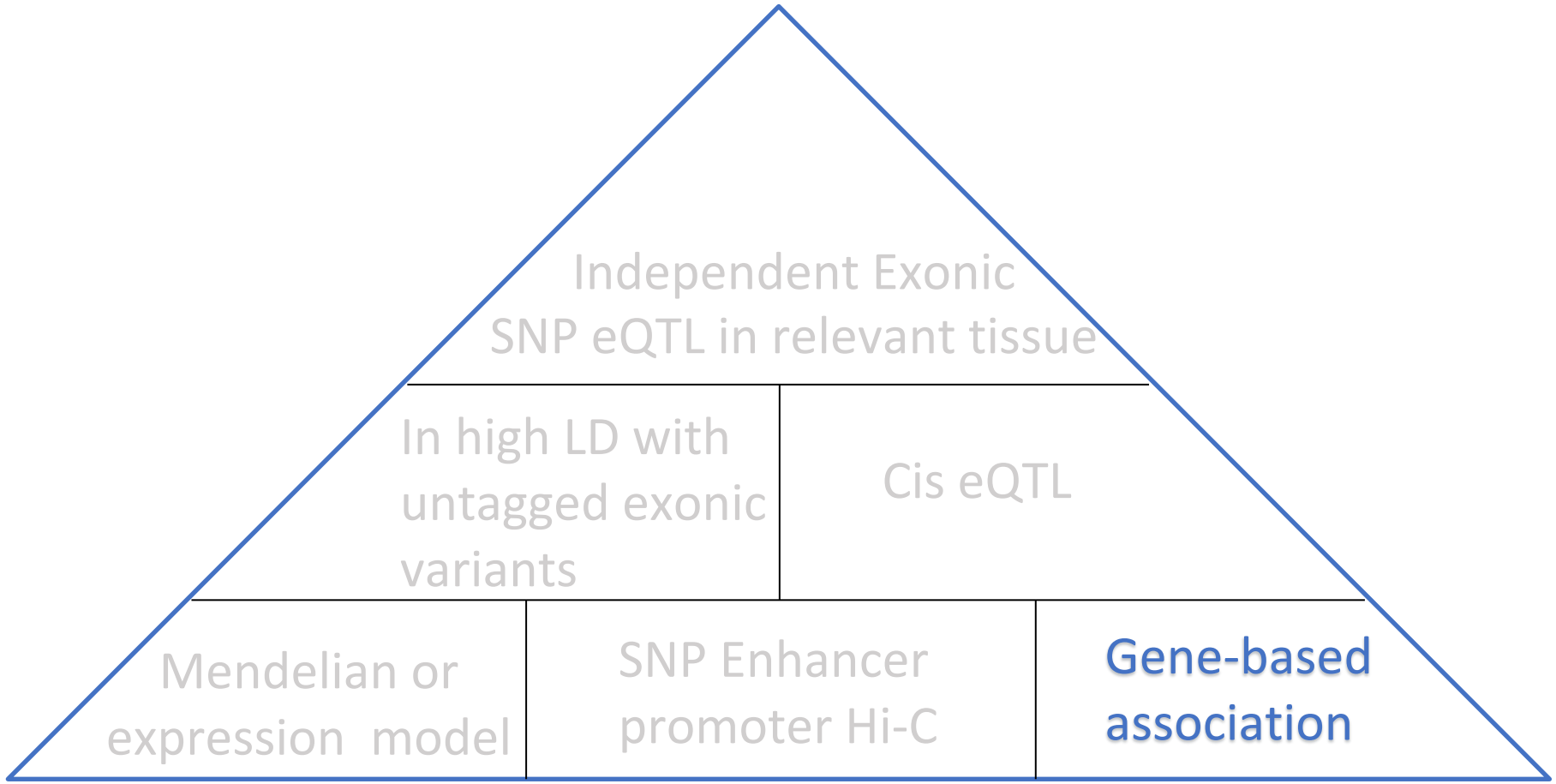
The Pyramid Gene-reporting Certainty




Independent Exonic
SNP eQTL in relevant tissue
In high LD with untagged exonic variants
Cis eQTL
Mendelian or expression model
SNP Enhancer promoter Hi-C
Gene-based association

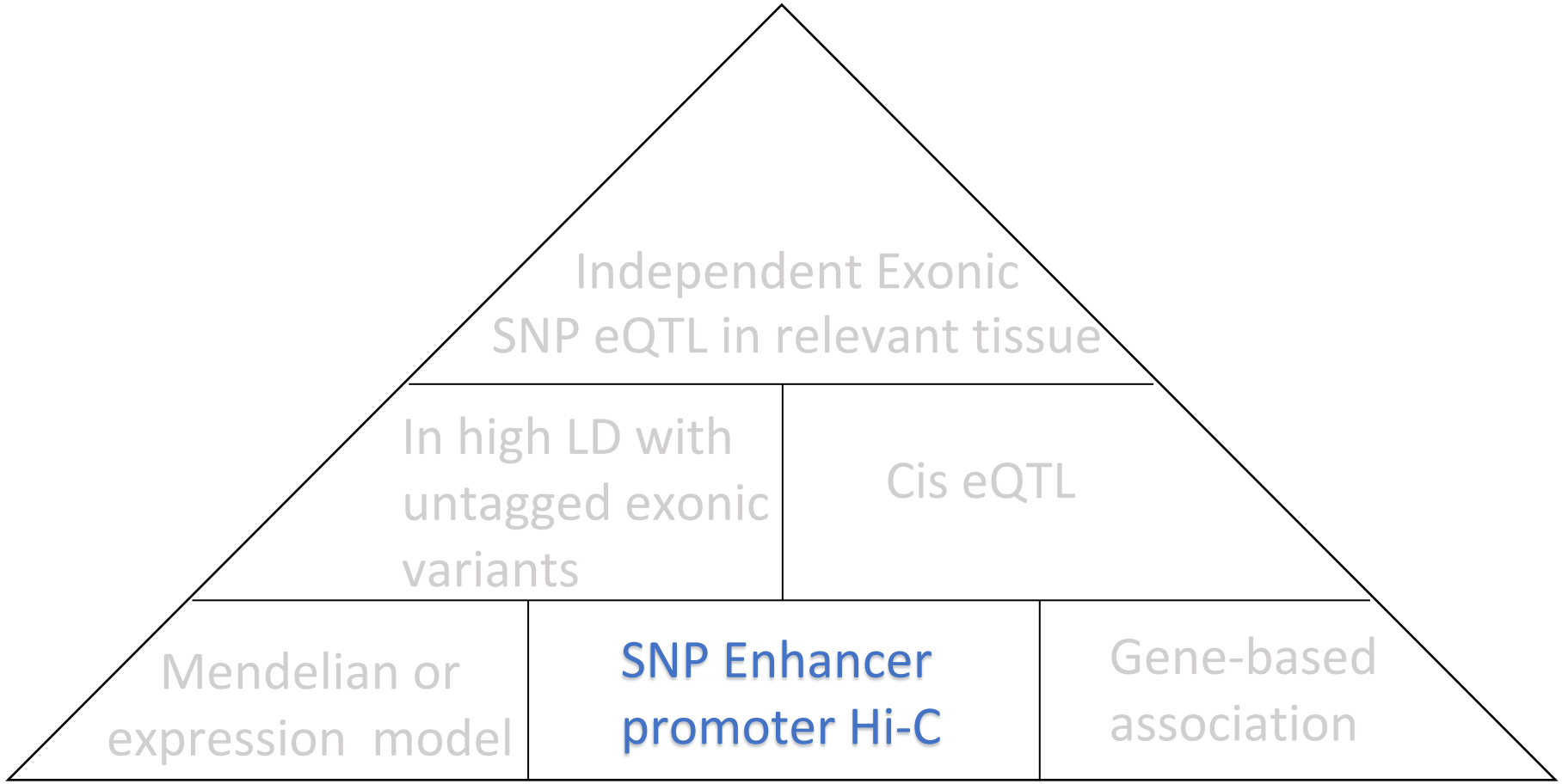

Independent Exonic
SNP eQTL in relevant tissue
In high LD with untagged exonic variants
Cis eQTL
Mendelian or expression model
SNP Enhancer promoter Hi-C
Gene-based association

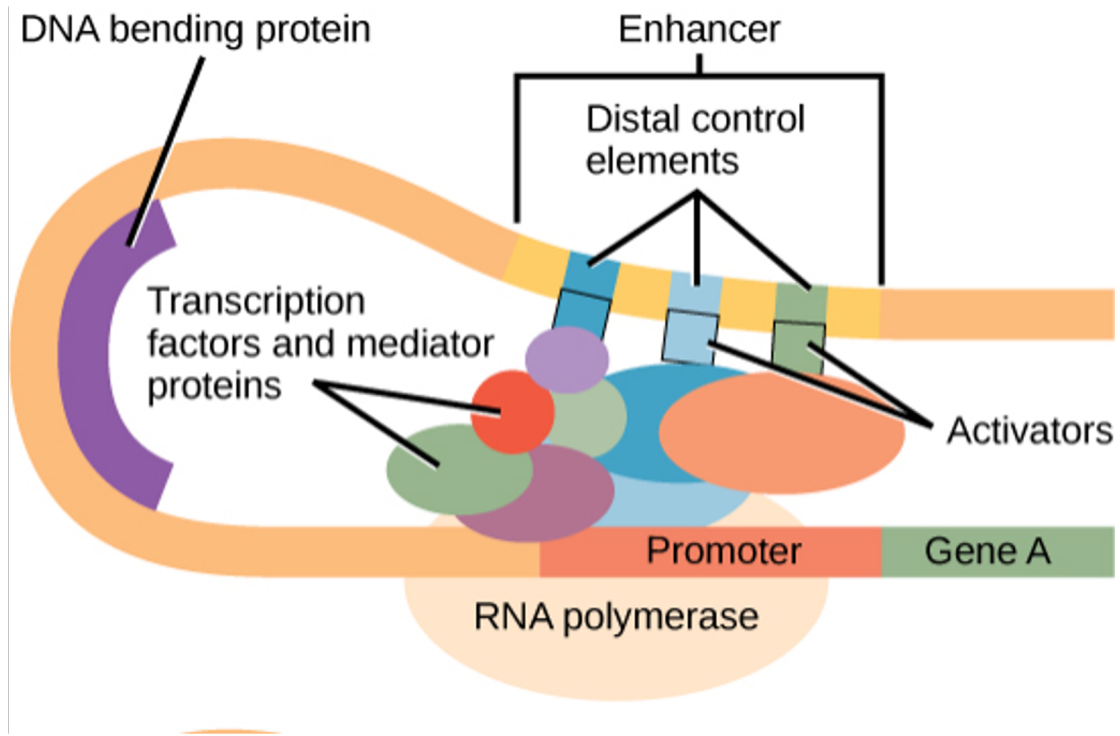

Independent Exonic SNP eQTL in relevant tissue
In high LD with untagged exonic variants
Cis eQTL
Mendelian or expression model
SNP Enhancer promoter Hi-C
Gene-based association


Independent Exonic SNP eQTL in relevant tissue
In high LD with untagged exonic variants
Cis eQTL
Mendelian or expression model
SNP Enhancer promoter Hi-C
Gene-based association

MAGMA Gene-based method

- Test of all SNPs positionally mapped to a gene (typically within 10 kb of a gene);
- Model is a “competitive test”, where we compare a random set of SNPs to those positionally mapped to the gene.
- Model controls for gene length and ancestral principal components (PCs);
  - In FUMA, the default is 1000 Genomes European sample (but you can/should change it if results are from another ancestral population)
- Input is a matrix of SNPs and their LD, and summary statistics from a GWAS.
- Model: $Y = \alpha_{0g}\vec{1} + X_g^*\alpha_g + W\beta_g + \varepsilon_g$


Independent Exonic SNP eQTL in relevant tissue
In high LD with untagged exonic variants
Cis eQTL
Mendelian or expression model
SNP Enhancer promoter Hi-C
Gene-based association

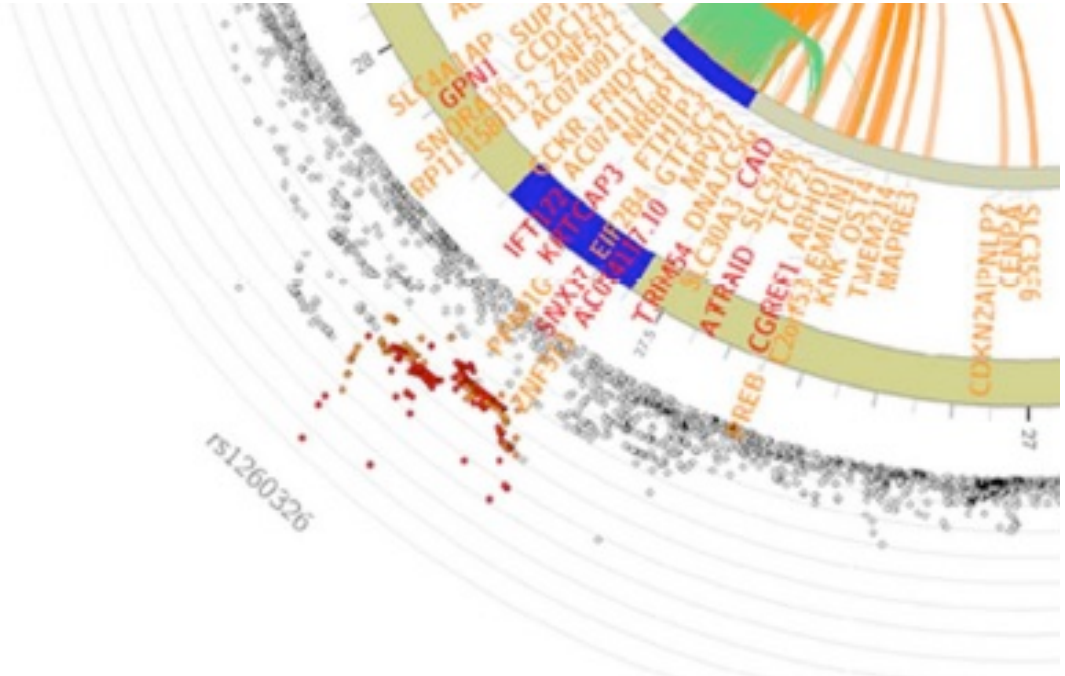

DNA bending protein
Enhancer
Distal control elements
Transcription factors and mediator proteins
Activators
Promoter
Gene A
RNA polymerase

Hi-C Practicalities

- Search for overlap between the significant SNPs from your study and known promoter/enhancer elements;
- Requires a pre-existing dataset of promoter/enhancer conformations
  - Hi-C data included in FUMA for relevant tissues, including hippocampus, dorsolateral prefrontal cortex, hippocampus, ventricle and neural progenitor cells;
  - FUMA has two builds for adult and fetal cortex;
  - PsychEncode generated sets are the densest to date;

- H-MAGMA (Sey et al., PMID: 32152537)
  - Adds enhancer promotor associations to the combined SNP (i.e., gene-based test)
  - Not on FUMA

Circos Plot of Chromosome 15



Let’s look at those two on FUMA!

- CPD FUMA: https://fuma.ctglab.nl/snp2gene/112975

Doing more with MASSIVE

CTG-VL (0.4-beta)

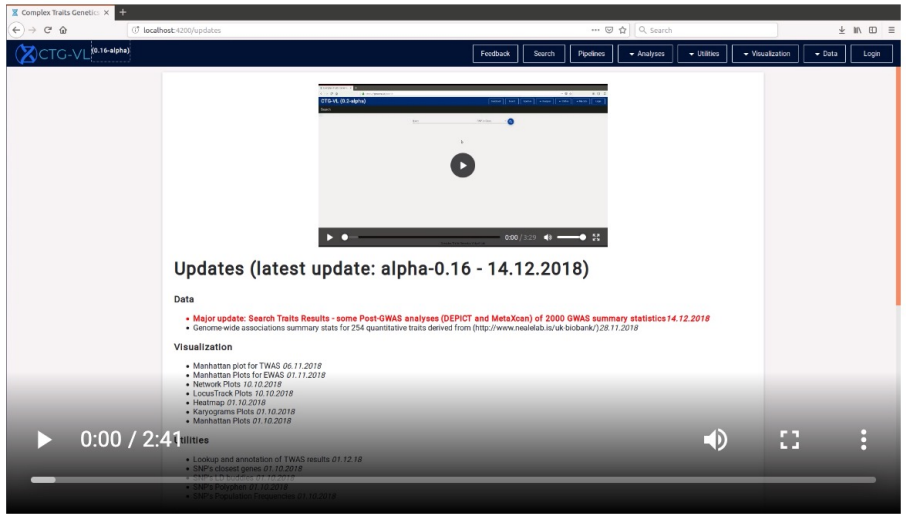
CTG-VIEW Feedback Analyses Pipelines Visualization Data ashatoum@wustl.edu

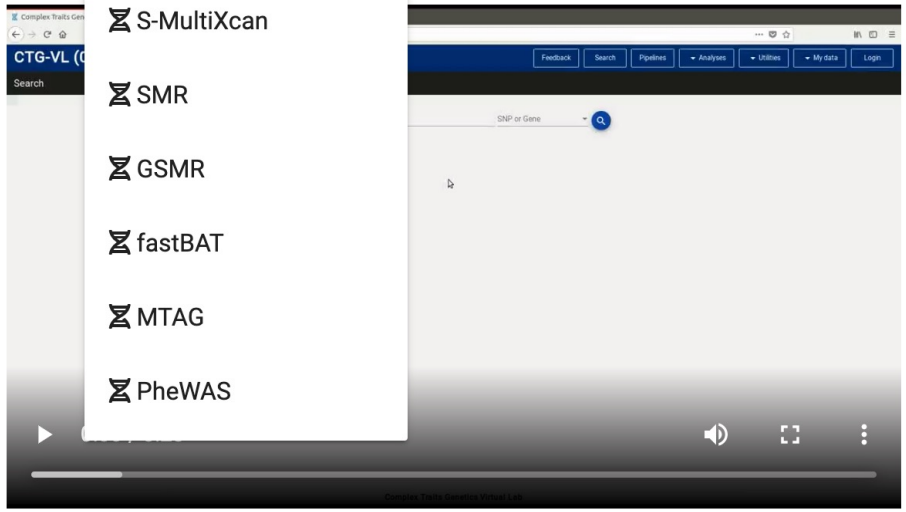
- Latent Causal Variable
- LD score
- DEPICT
- MetaXcan
- S-MultiXcan
- SMR
- GSMR
- fastBAT
- MTAG
- PheWAS

Updates

The platform is in beta version. The only way to build a great tool is with feedback in touch if you notice any errors or woud like to contribute

Preprint describing the platform: Complex-Traits Genetics Virtual Lab: A commun ost-GWAS analyses. bioRxiv, 2019.





Updates (latest update: beta-0.4 - 09.12.2020)

Analysis

- MASSIVE: Massive downstream Analysis of Summary Statistics added *28.10.2019*
- Phenome-wide LCV *01.06.2019*
- PheWAS capability added*20.09.2019*

https://vl.genoma.io/analyses/lcv

CTG-VL (0.4-beta)

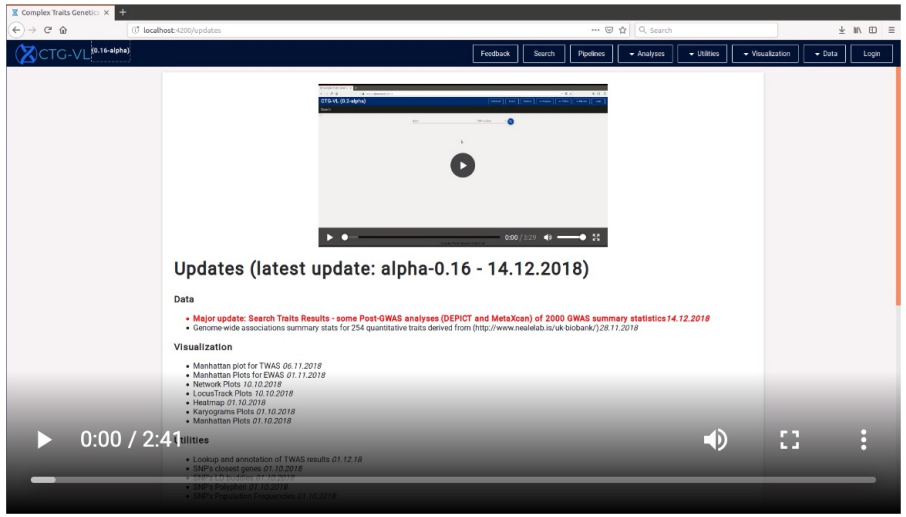
CTG-VIEW Feedback Analyses Pipelines Visualization Data ashatoum@wustl.edu

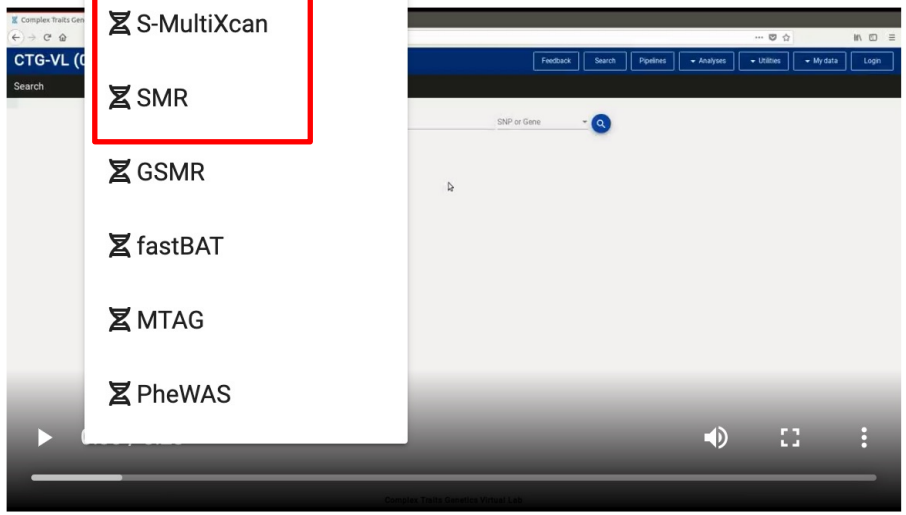
- Latent Causal Variable
- LD score
- DEPICT
- MetaXcan
- S-MultiXcan
- SMR
- GSMR
- fastBAT
- MTAG
- PheWAS

Updates

The platform is in beta version. The only way to build a great tool is with feedback in touch if you notice any errors or woud like to contribute

Preprint describing the platform: Complex-Traits Genetics Virtual Lab: A commun ost-GWAS analyses. bioRxiv, 2019.



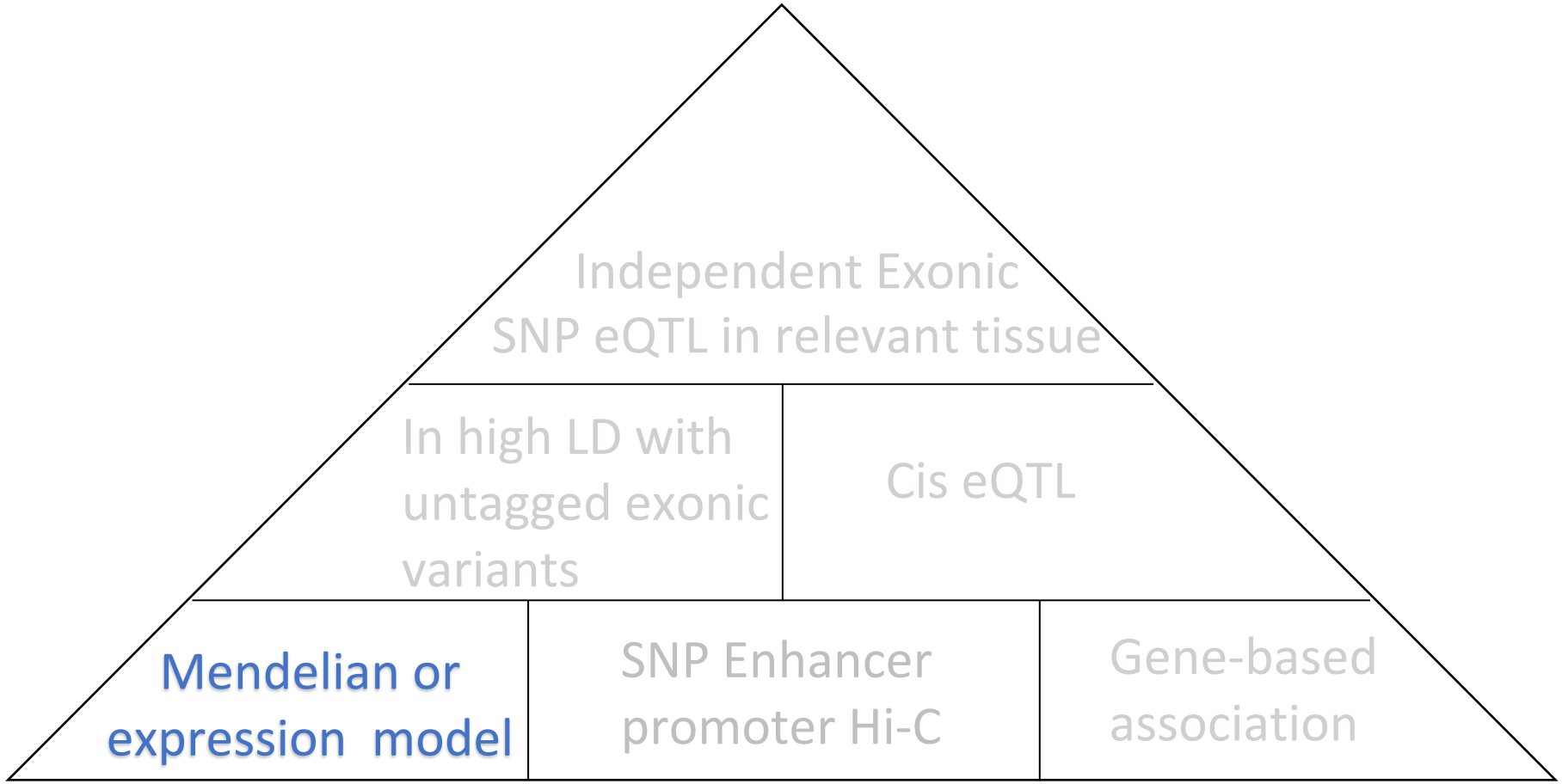


Updates (latest update: beta-0.4 - 09.12.2020)

Analysis

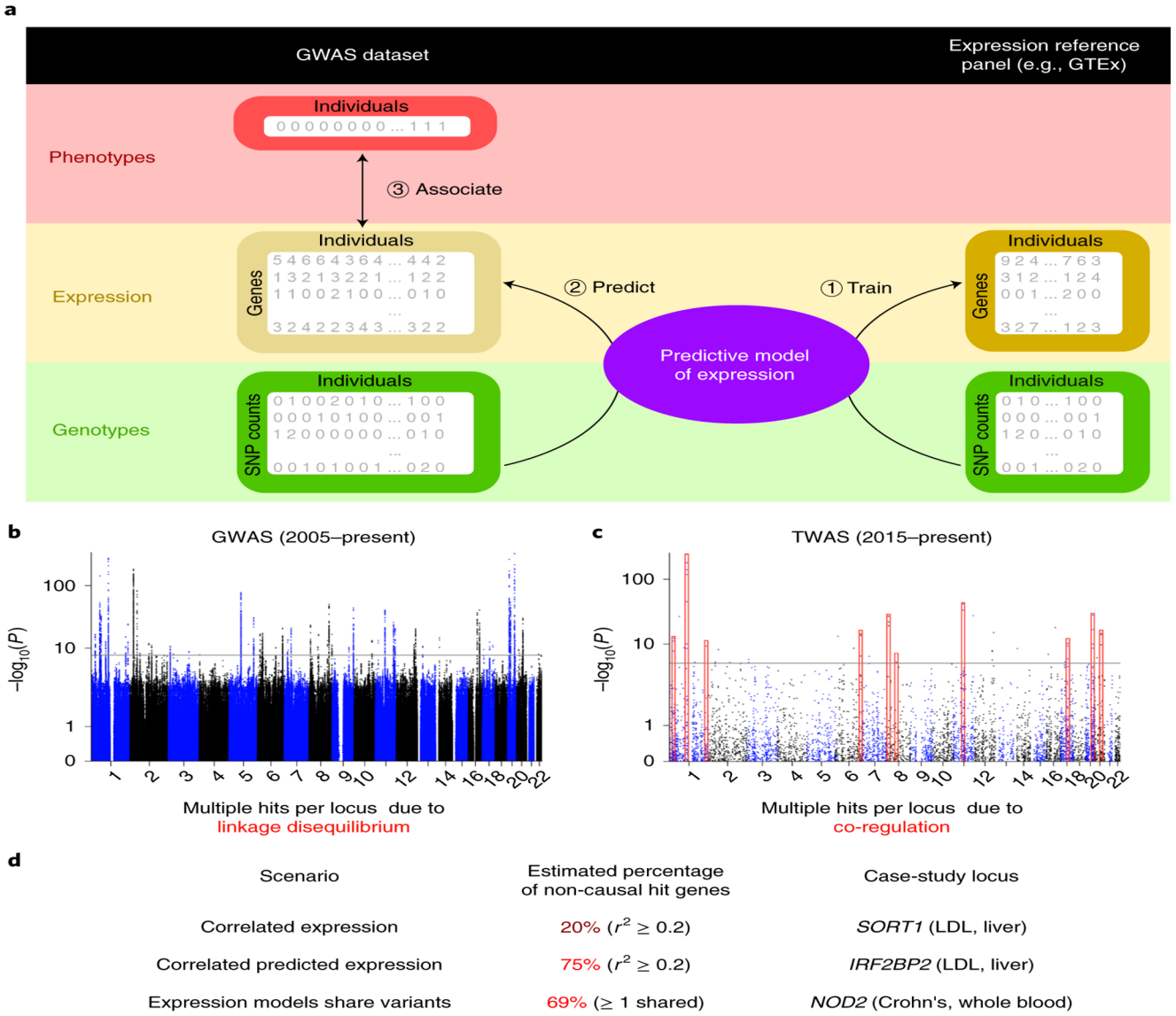
- MASSIVE: Massive downstream Analysis of Summary Statistics added *28.10.2019*
- Phenome-wide LCV *01.06.2019*
- PheWAS capability added*20.09.2019*

https://vl.genoma.io/analyses/lcv

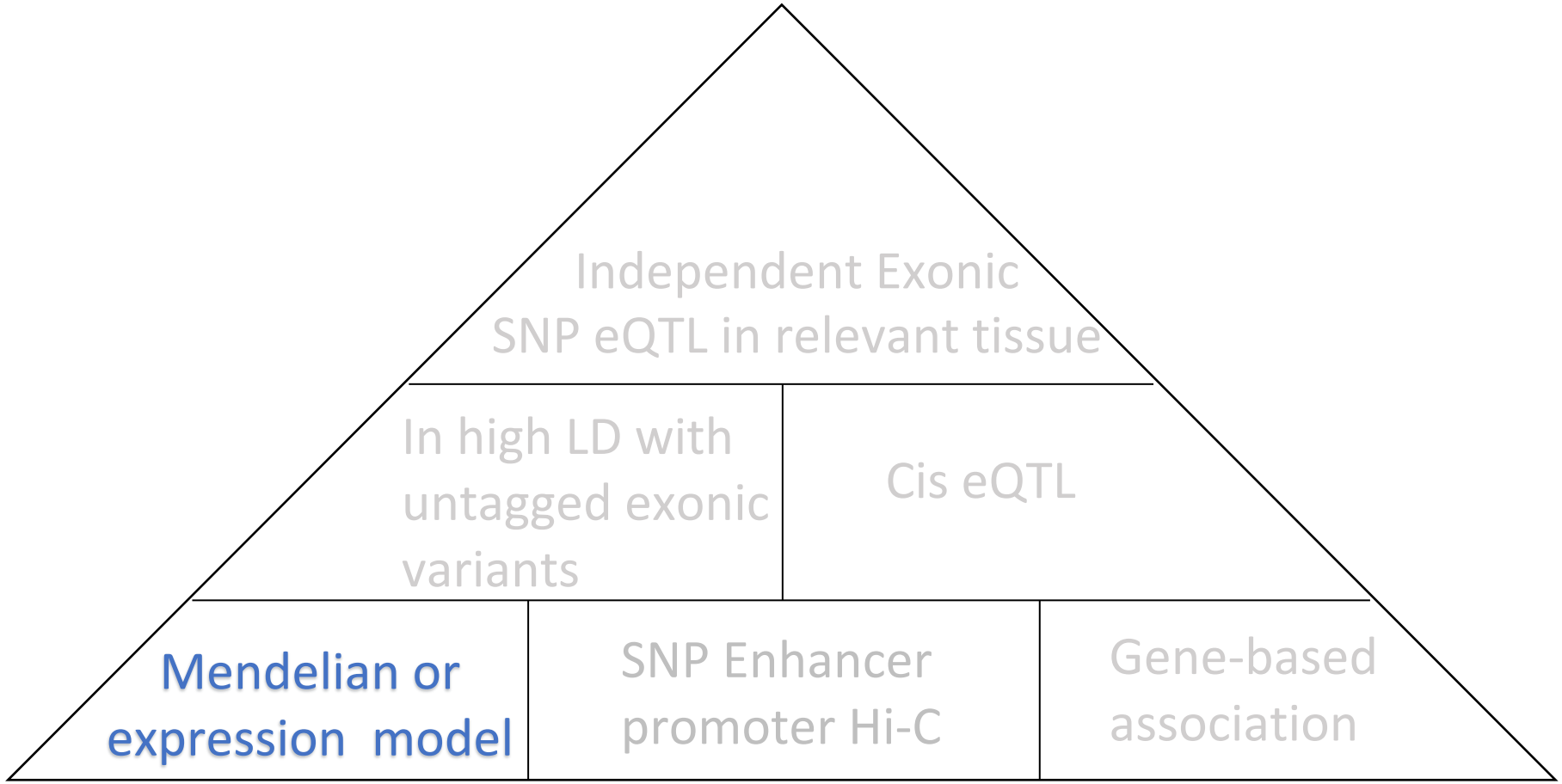

Independent Exonic SNP eQTL in relevant tissue
In high LD with untagged exonic variants
Cis eQTL
Mendelian or expression model
SNP Enhancer promoter Hi-C
Gene-based association

Inferred transcriptional models

- Models built from SNP and expression (eQTL) databases
  - Models are weighted by SNP associations with eQTLs and traits expected to have similar SNP patterns “predict/impute” expression.

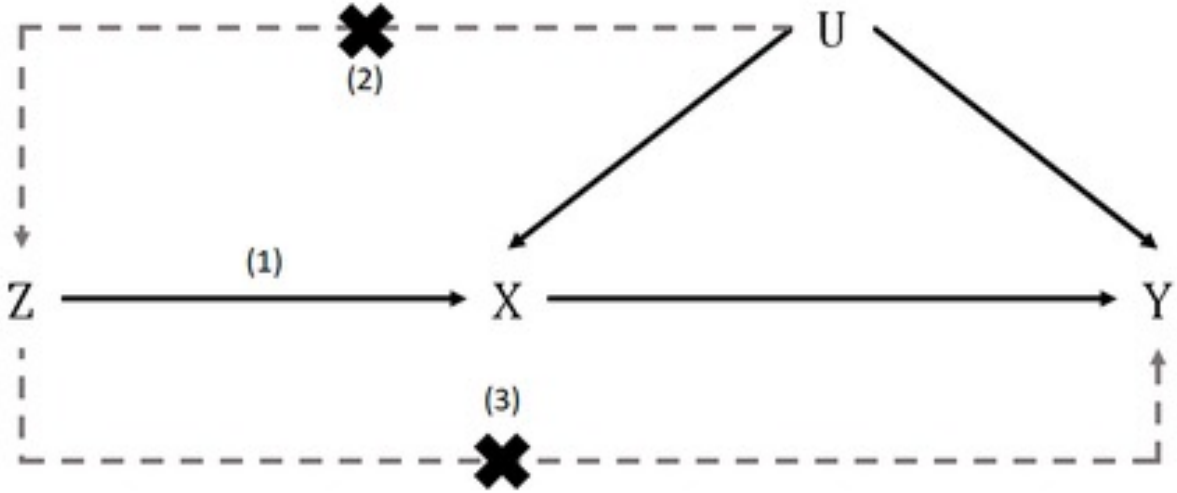


Wainberg et al. 2019 (PMID: 30926968)
https://doi.org/10.1038/s41588-019-0385-z


Independent Exonic SNP eQTL in relevant tissue
In high LD with untagged exonic variants
Cis eQTL
Mendelian or expression model
SNP Enhancer promoter Hi-C
Gene-based association

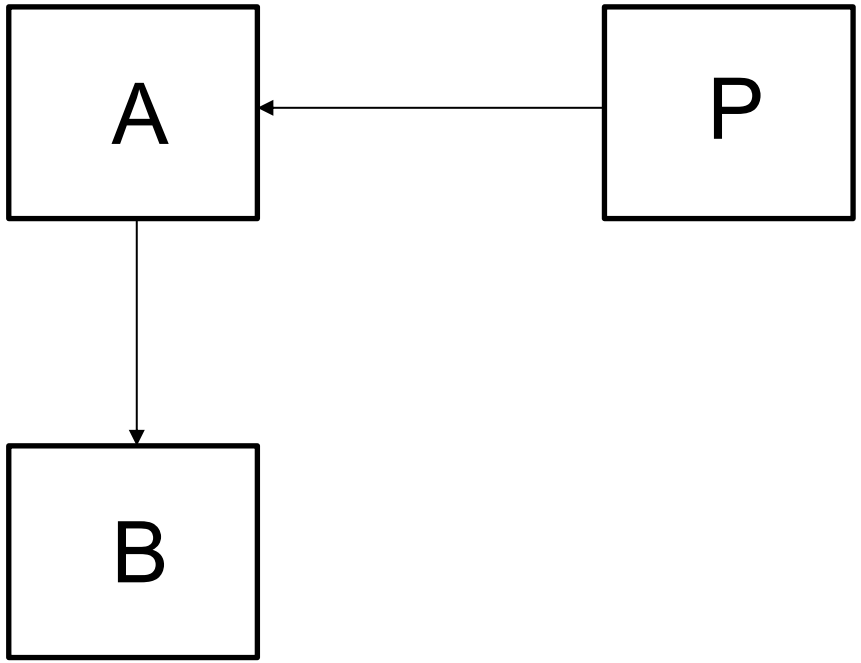
Mendelian Randomization

- Fundamental methodology for genotyped SNPs
  - Extension of classic instrumental variable design
  - Uses genotyped SNPs to “randomize” individuals to groups
- Z is genotyped SNP, X is exposure and Y is outcome
  - U is an unknown confounder
- From an epidemiology perspective: Test whether association between Z and Y is mediated (completely) by X

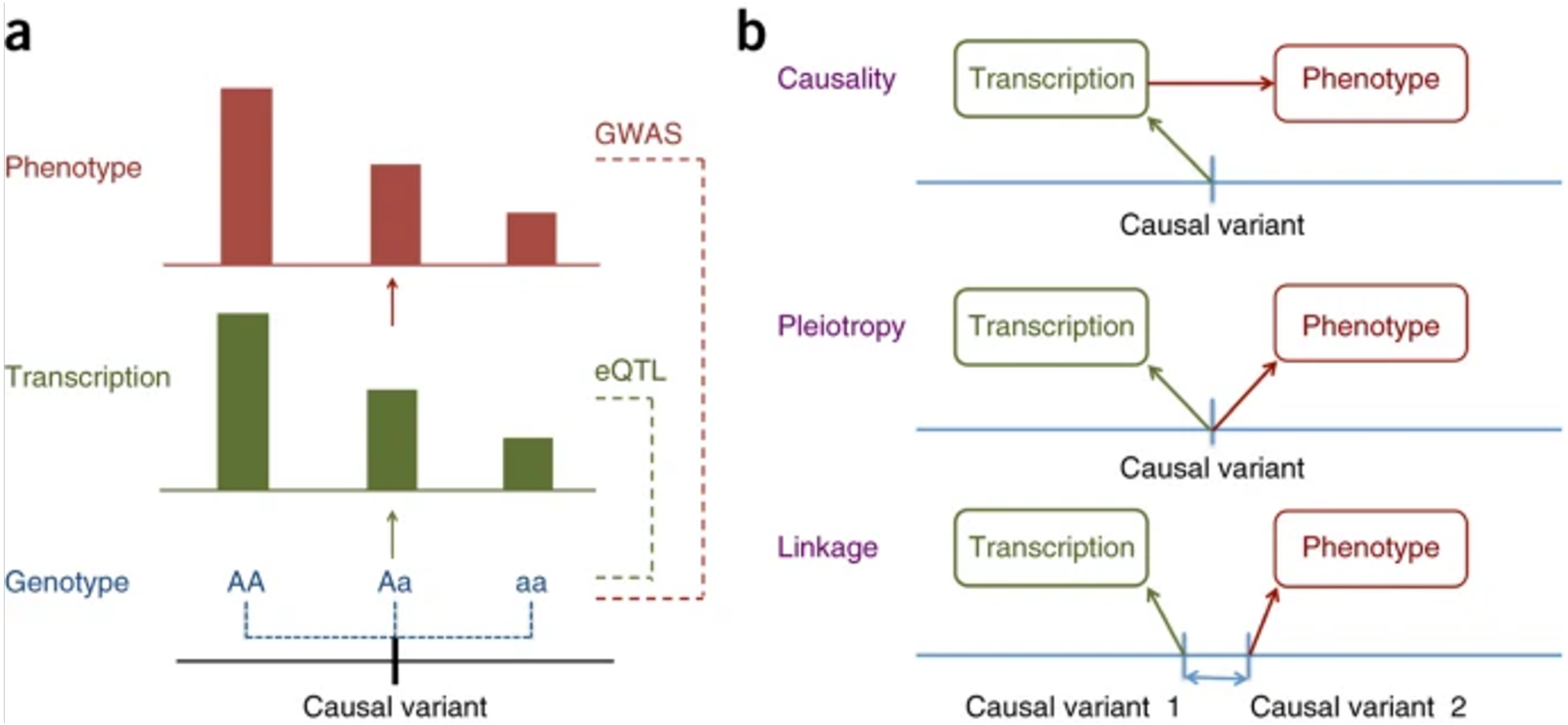


Mendelian Randomization logic

- ”Genetic causality” means that we can establish some process in which the manifestation of phenotype B is due the effects of A
  - We establish the system that is necessary (but not sufficient) for causality.
- If P causes A, and A causes B, then P must also cause B
- If we have causal SNPs of large effects, it’s easy!
  - We only have a couple of those
    - ADH1B/C, CHRNA5

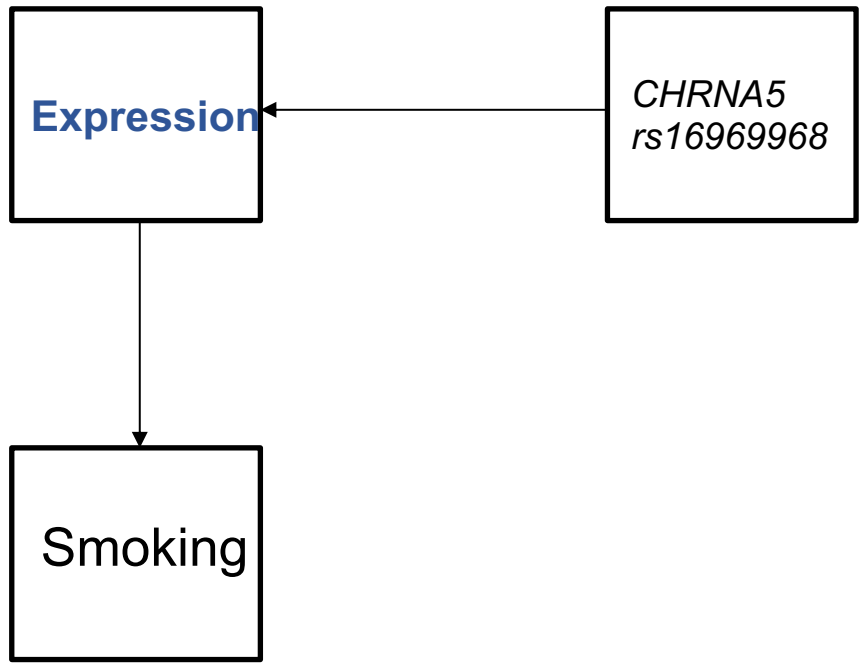


Mendelian Randomization with Expression



Mendelian Randomization logic

- If P causes A, and A causes B, then P must also cause B



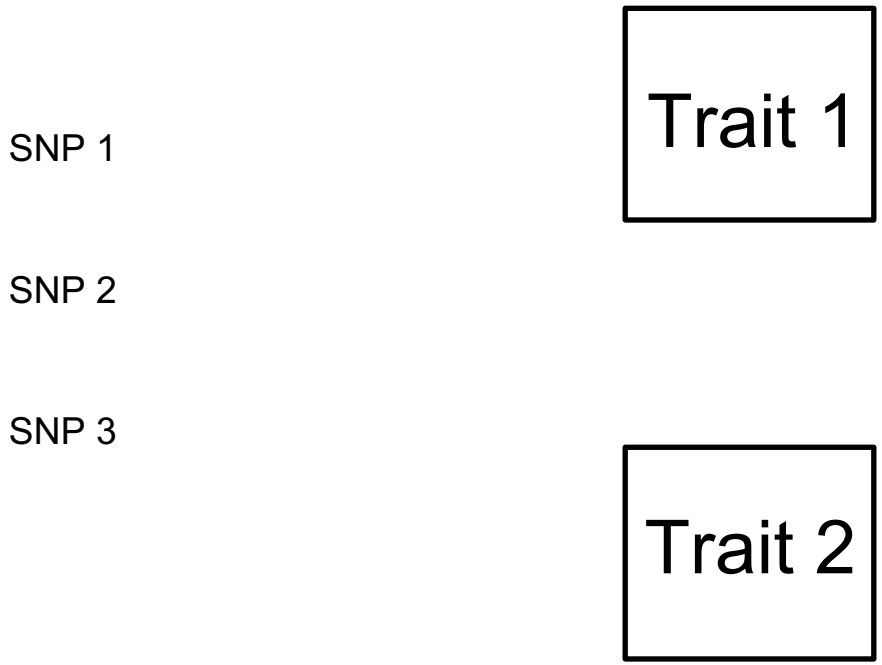
How do we run multi-SNP Mendelian randomization?

- Estimate effects sizes assuming a no effect (non-causality), departure from this is a causal model
- E.g., could constrain SNPs to have the same effects that are mediated by the presence of an exposure
- With summary statistics, we will often estimate a marginal effect size and compare to the expectation (from O'Connor & Price 2018; PMID: 30374074): "If trait 1 is causal for trait 2, Marginal effect sizes on trait one show influence trait 2, but marginal effect sizes on trait 2 shouldn't influence trait 1

What about an example?

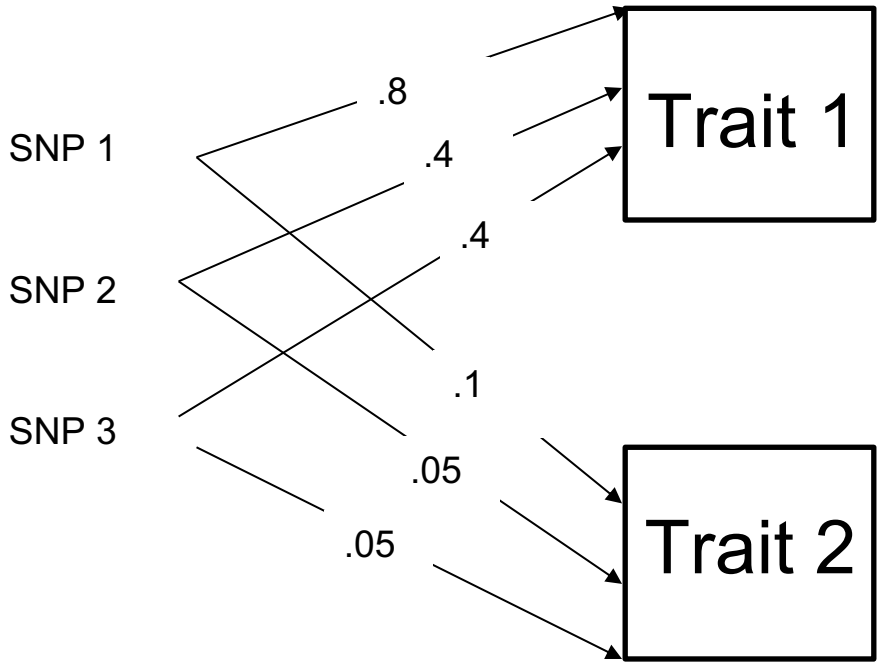
Three SNP example: Causal

Three SNPs that are related to two traits



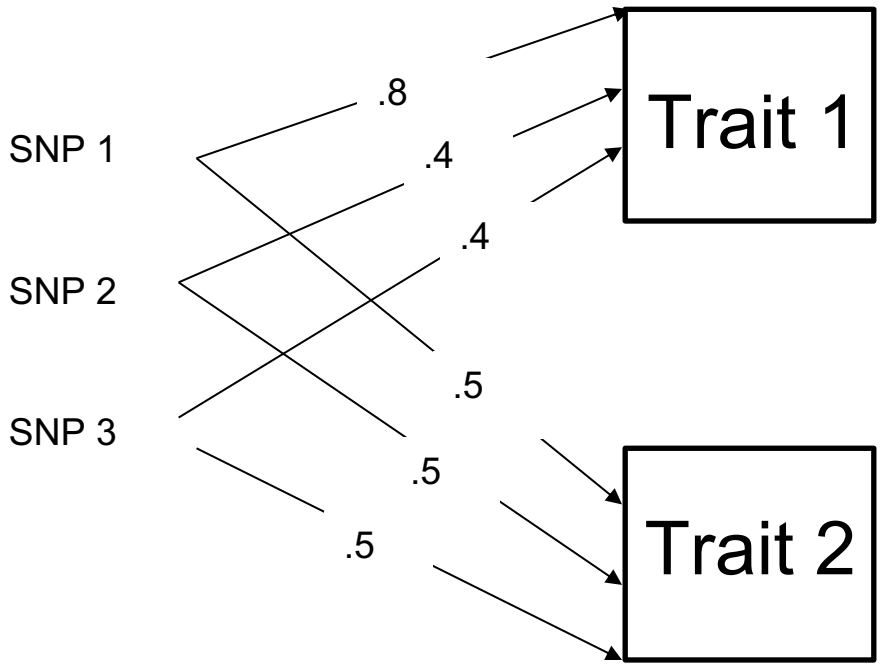
Three SNP example: Causal

Let's assume: all three SNPs are additive and completely penetrant, If Causal, we would expect proportionally similar effects



Three SNP example: Pleiotropy

Let's assume: all three SNPs are additive and completely penetrant, If Causal we would expect proportionally similar effects

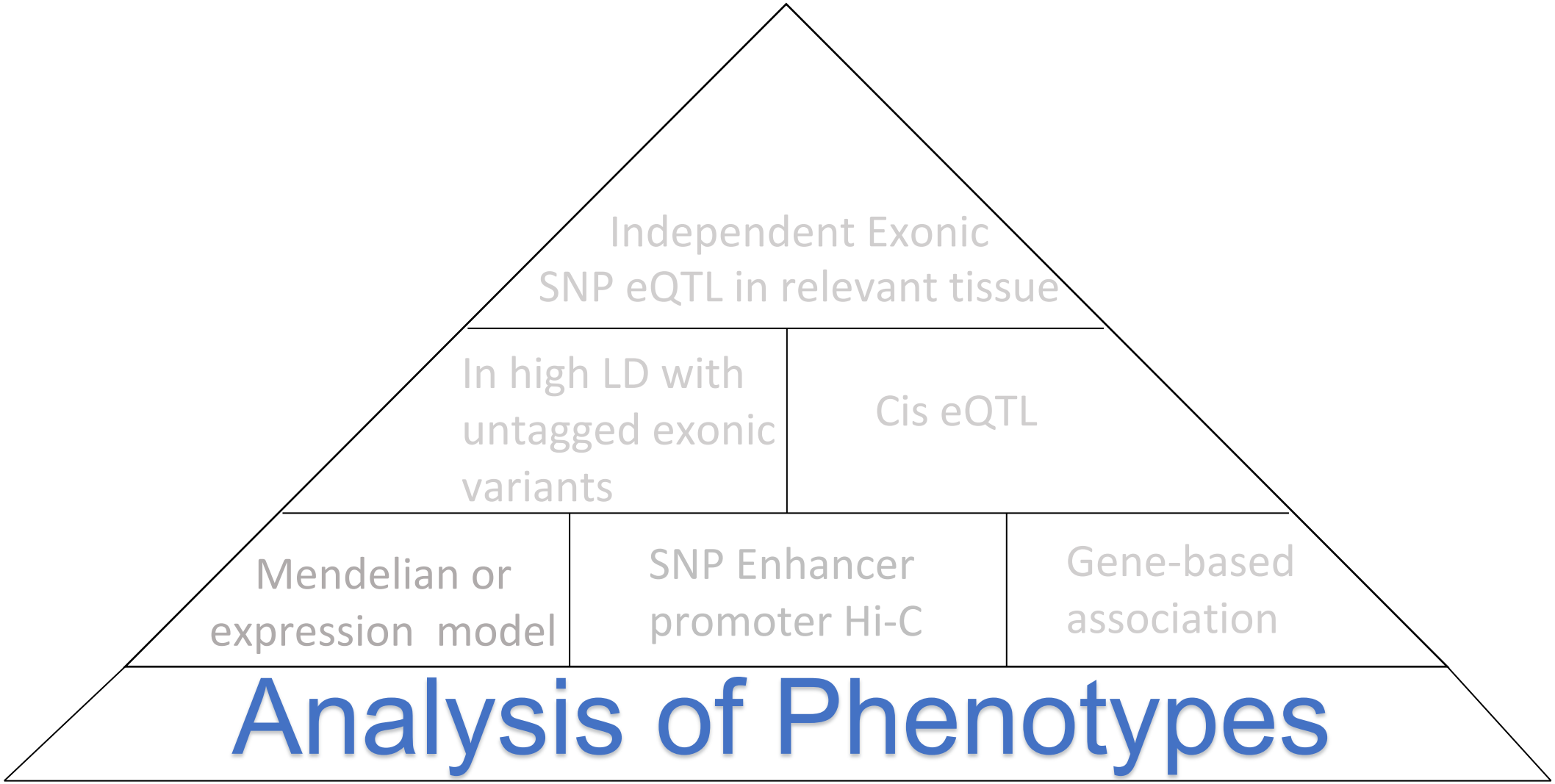


Examples with MASSIVE/genoma.io

- https://view.genoma.io/gwas/QrCBPZiDZVVnfw8DNARoUnEoZur2/I6zZtYeTnPA2IYXA0QRj

Part 3: Moving to the Whole-Genome Level

What We Can Do With Results Across the Genome


Independent Exonic SNP eQTL in relevant tissue
In high LD with untagged exonic variants
Cis eQTL
Mendelian or expression model
SNP Enhancer promoter Hi-C
Gene-based association
Analysis of Phenotypes

Answering Epidemiological Questions About Phenotypes

- Traits sharing pleiotropic effects with our trait: **Genetic Correlations**

- Genetic Causality revisited at the whole genome level: **Latent Causal Variable Analysis**

Genetic Correlation

Genetic Correlations

- In twin studies: rG is estimated using cross-trait cross-twin covariance;
- In GWAS, SNP-rG is similarity in genome-wide effect sizes for two traits:
  - For example, Cannabis Use Disorder and Opioid Use Disorder have a SNP-rG =.7, suggesting similarity in effects across genome
  - Can look at traits we cannot observe correlations between phenotypically, like cannabis ever smoke and cannabis use disorder
- Many methods (LD Score, GCTA);

LD score regression

- LD score regression is a general method for estimating SNP heritability, heritability enrichment, and co-inheritance
  - LD score regression is mostly used for co-inheritance, and we can estimate genetic correlation (SNP-rG) from coinheritance
- Regression of the LD scores (degree of LD in a block) on the degree of association
  - Can include a second line of regression to get estimate of SNP-rG;
- Works well with summary statistics within major geographic ancestry groups;
- Only works well with common, well imputed variants
  - Needs at least 600,000 variants to run effectively
- Separates other sources of similarity (e.g., sample overlap, ancestry) from estimates of SNP-h2 and SNP-rG.
- Run via MASSIVE and many online resources have libraries of trait associations

Genetic Causality Between Phenotypes

Why we need more than MR and Latent causal variable Analysis

Assumptions of MR

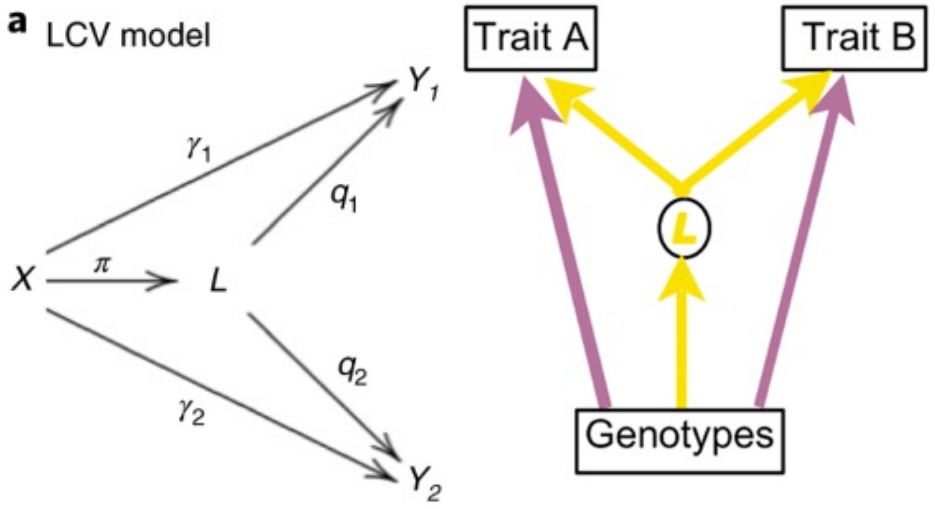
- Depending on the MR methods, there are several assumptions of the models
  - Many MR innovations just relax one of these assumptions.
- No sample overlap;
- Genetic instruments are powerful enough to segregate individuals (weak instrument assumption);
- Genes are not pleiotropic
  - Even when relaxed, often assumes no mixture of pleiotropic and causal SNPs;
- SNPs included do not overlap with each other (must have low LD between those SNPs)

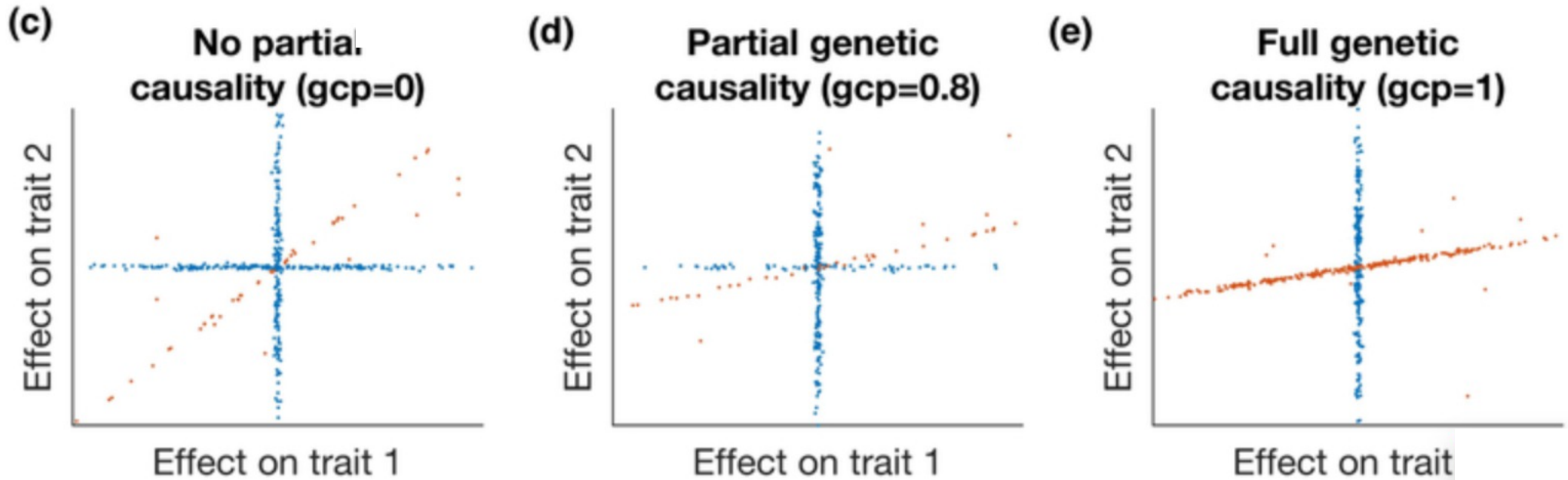
~~Assumptions of MR~~ Assumptions violated by substance use data

- Depending on the MR methods, there are several assumptions of the models
  - Many MR innovations just relax one of these assumptions.
- No sample overlap
- Genetic instruments are powerful enough to segregate individuals (weak instrument assumption)
- Genes are not pleiotropic
  - Even when relaxed, often assumes no mixture of pleiotropic and causal SNPs
- SNPs included do not overlap with each other (must have low LD between those SNPs);

Latent Causal Variable (LCV) Analysis

- Formulated by O'Connor & Price in 2018 (PMID: 30374074)
- Allows SNP-rG between two traits to be mediated by a latent variable with a causal effect on each trait.
- Causality is implied when trait one is strongly correlated with the causal variable in the model compared to the second trait.
- Think of this as a variance partitioning of the LD score correlation (SNP-rG) into causal or pleiotropic effects.




(c)
No partia.
causality (gcp=0)
Effect on trait 2
Effect on trait 1
(d)
Partial genetic
causality (gcp=0.8)
Effect on trait 2
Effect on trait 1
(e)
Full genetic
causality (gcp=1)
Effect on trait 2
Effect on trait

LCV vs. MR

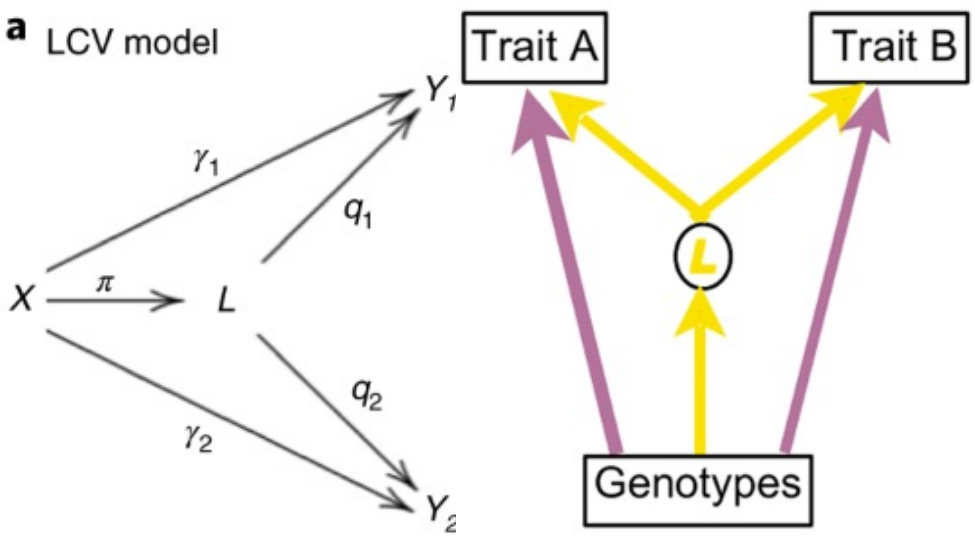
- LCV similar to MR (in that it utilizes a genetic instrument) but has several advantages;
  - Sample overlap is accounted for by LD scores while most MR methods cannot;
  - The gcp is robust to pleiotropy, in contrast to MR;
    - Output looks like: “GCP = .75, P=.001”,
    - The causality can be partial, and the model produces a genetic causality proportion, that is the proportion of genes to genetic causality

- Advantages of MR:
  - Advanced MR can test bidirectional relationships;
  - Some MR methods can be conducted within smaller samples, instead of relying on summary statistics (as long as genetic instruments still have relatively large effects)
  - LCV has low power when SNP-rG is low and trait is not as polygenic;

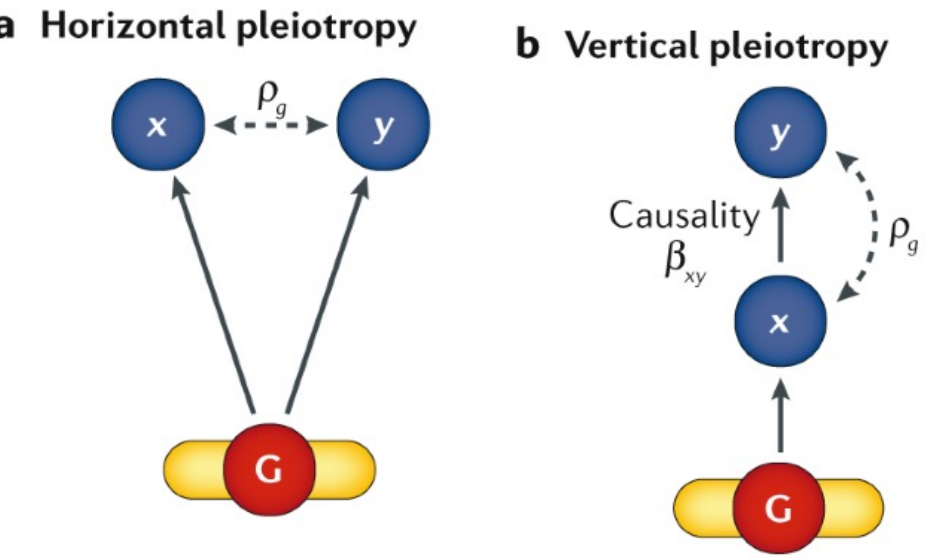
Examples with MASSIVE

- https://fuma.ctglab.nl/snp2gene/112975

Questions


a LCV model
Trait A
Trait B
Y_1
γ_1
q_1
X
π
L
q_2
γ_2
Y_2
L
Genotypes

-
-
-
-
-
-
-
-
-


a Horizontal pleiotropy
x
ρ_g
y
G
b Vertical pleiotropy
y
Causality
β_{xy}
ρ_g
x
G

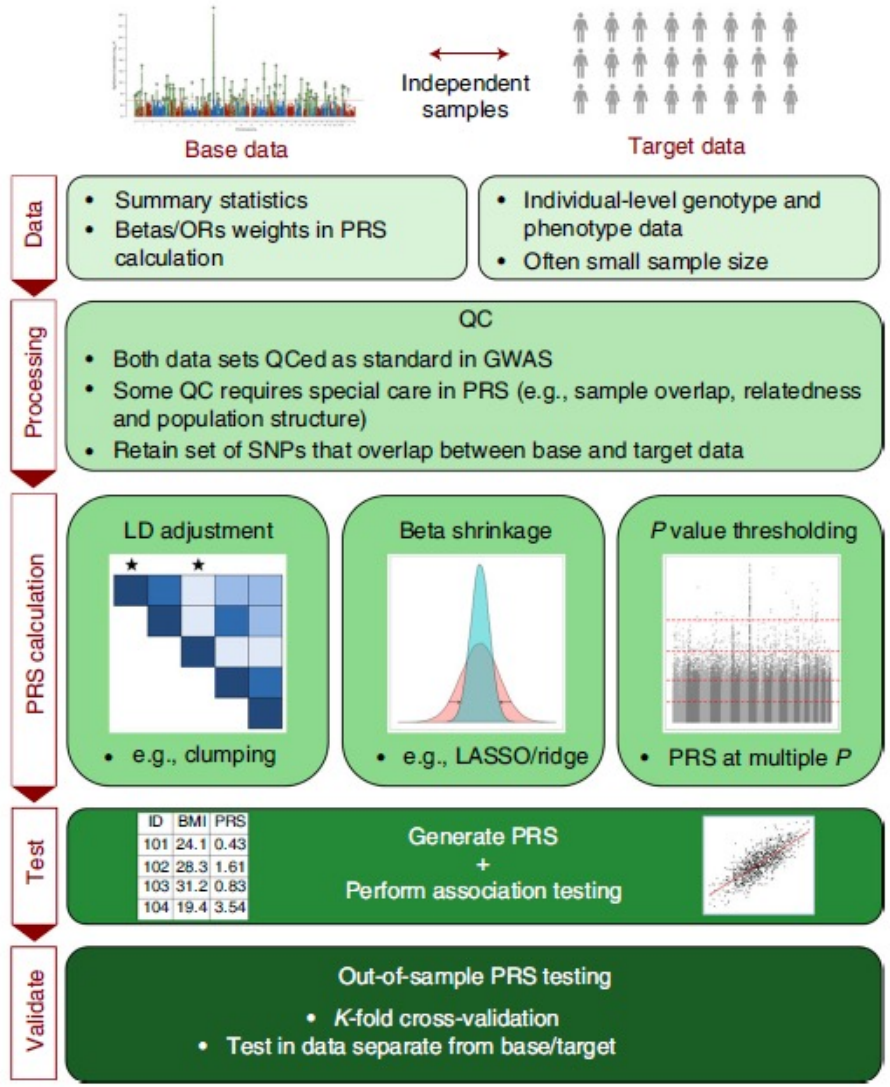


Fig. 1 | The PRS analysis process. PRS analyses can be characterized by their use of base and target data sets. QC of both data sets is described in 'QC of base and target data', while the different approaches to calculating PRSs (e.g., LD adjustment via clumping, beta shrinkage using LASSO regression or *P* value thresholding) are summarized in 'Calculation of PRSs'. Issues relating to utilizing PRSs for association analyses to test hypotheses, including interpretation of results and avoidance of overfitting to the target data, are detailed in 'Interpretation and presentation of results'.