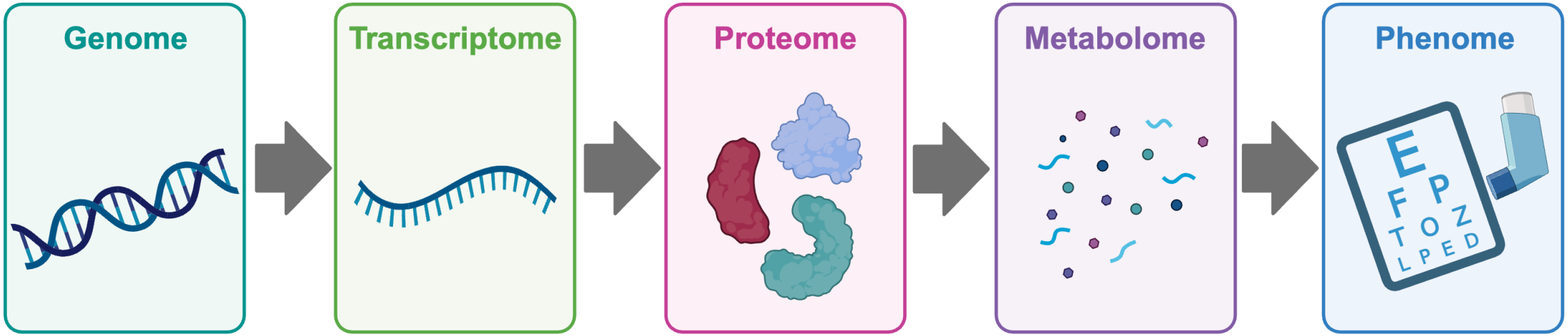
Protein MR in the UK Biobank

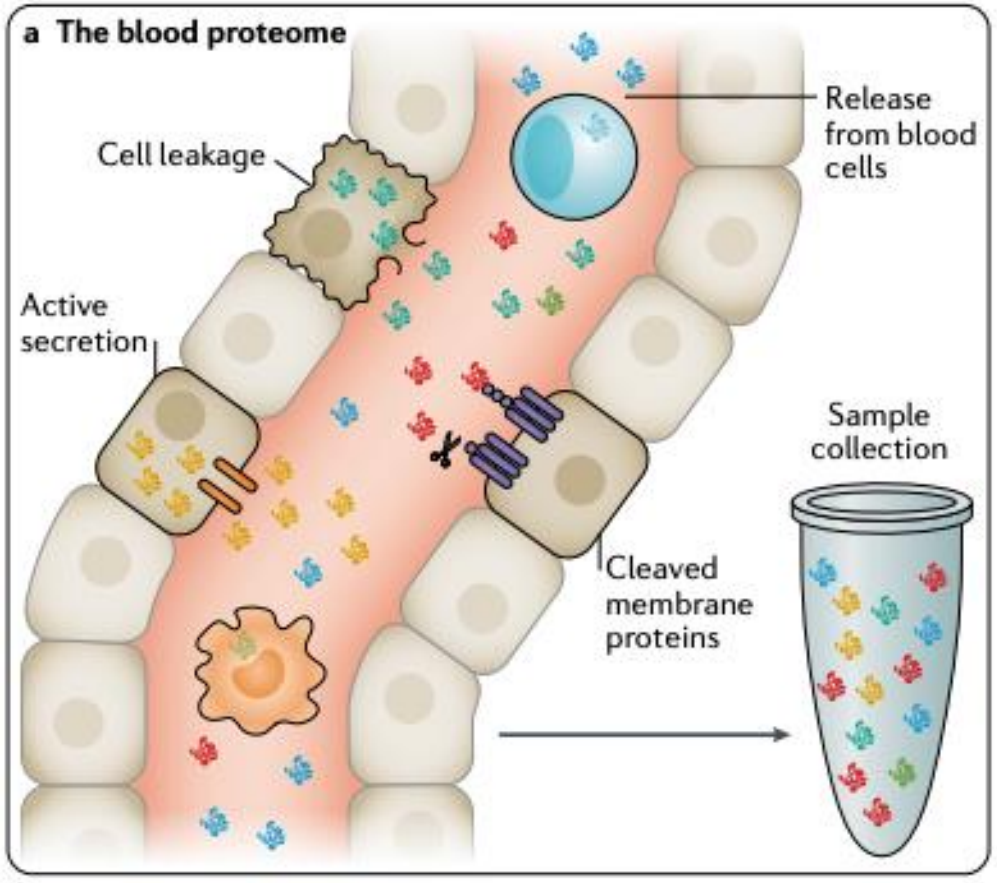
Emily Daubney

Proteins in context



- Proteins = important biological functional units
- Intermediate phenotype meaning it is a combination of our underlying genetics and our environment
- The further away from the genome the more potential to interact with environment and the more dynamic
- Large cross play between these different levels (it's not actually a simple follow-on like the figure implies)

The blood proteome

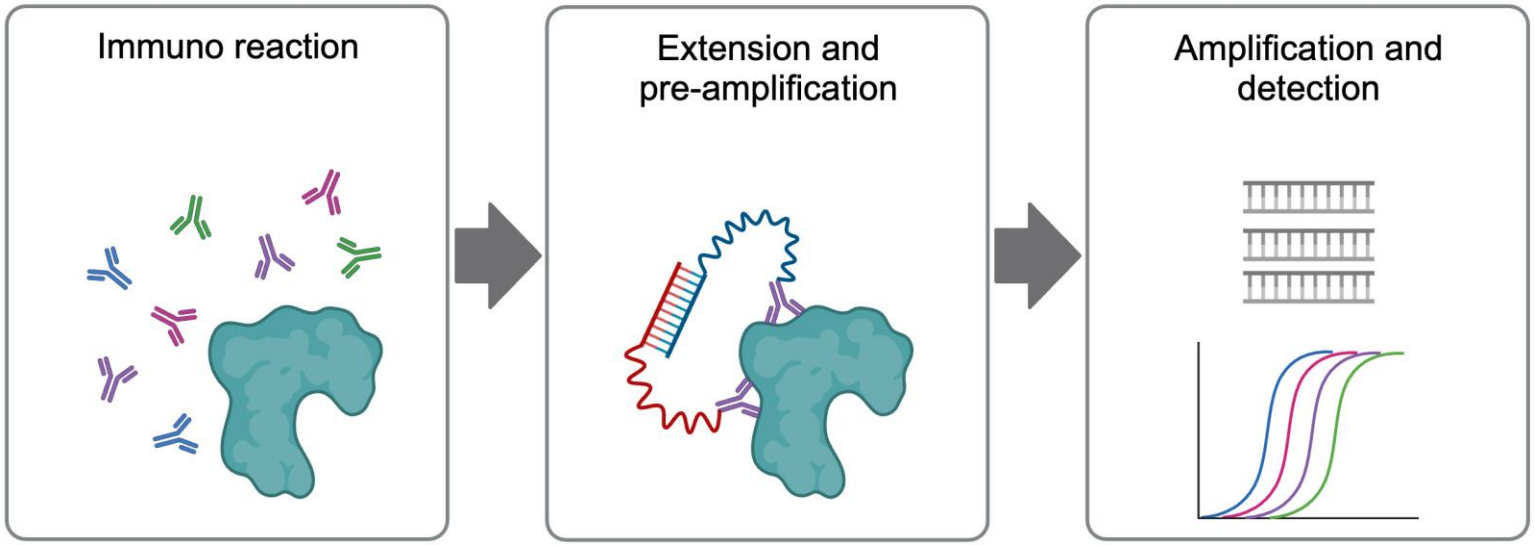


Suhre *et al.* (2021)

- **Proteome GWAS mainly consider the blood proteome**
  - UK Biobank = plasma samples measured with Olink
- **Blood proteome includes proteins from different cells and tissues**
  - Secretion for signalling or enzymatic actions
  - Leakage from damaged cells
  - Most common are those produced and secreted by the liver
- **Most proteome GWAS utilise scalable affinity-based proteomic techniques**
  - Olink = anti-body based affinity proteomics
  - SomaScan = aptamer-based affinity proteomics

Olink overview

- **Antibody-based affinity proteomics**
- **Binding of two separate, complimentary antibodies**
- **Protein-specific binding with a readout that relies on DNA concentrations (through qPCR)**



Olink overview

Advantages	Disadvantages
• High sample throughput • Decent proteome coverage (~3000 protein < SomaScan) • Improved specificity compared to other high throughput technologies	• Relative quantification • Possibility of epitope effects • Does not address the issue of different proteoforms

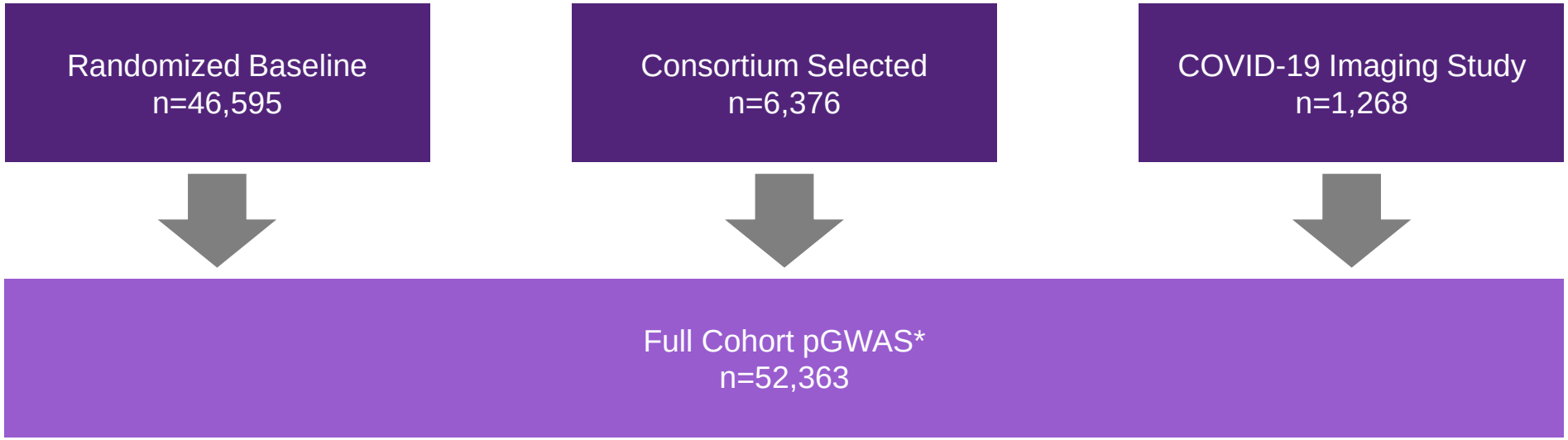
What is NPX?

- **NPX = Normalized Protein eXpression**
- **Measure relative changes in protein expression (i.e. this is not a concentration)**
  - Can only compare NPX for same proteins within the same study
- **Increased NPX corresponds to increased concentration**

	IL6	PCSK9
Sample 1	9.87	10.93
Sample 2	10.64	9.28
Sample 3	6.37	9.68

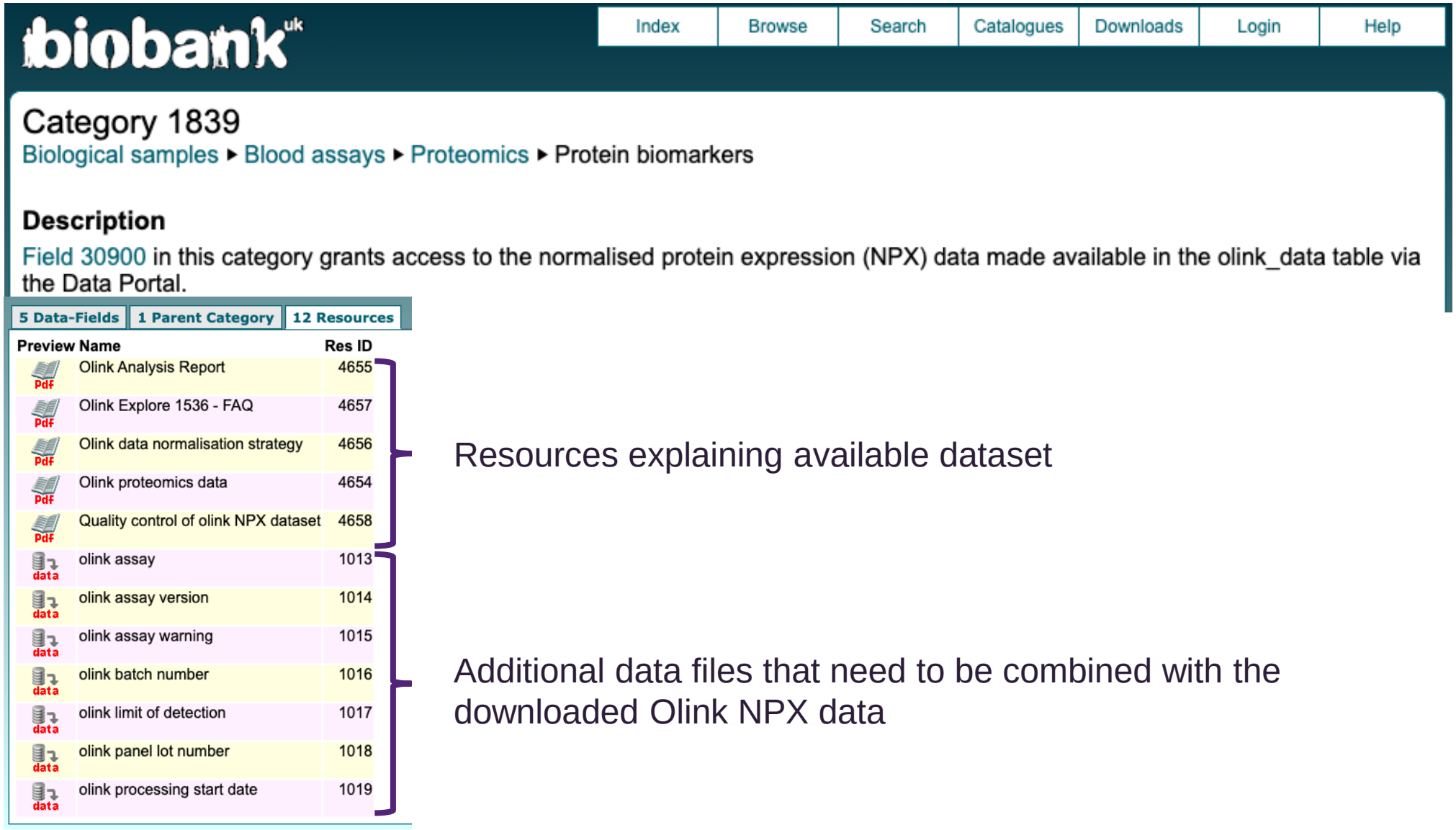
Proteins and UK Biobank

- **UKB-PPP = UK Biobank Pharma Proteomics Project**
- **Proteomic profiling on blood plasma samples collected from 54,219 UKB participants**
- **Measured 2923 unique proteins using Olink Explore 3072 PEA**



*publicly available summary statistics from Sun *et al.* (2023) https://metabolomips.org/ukbbpgwas/

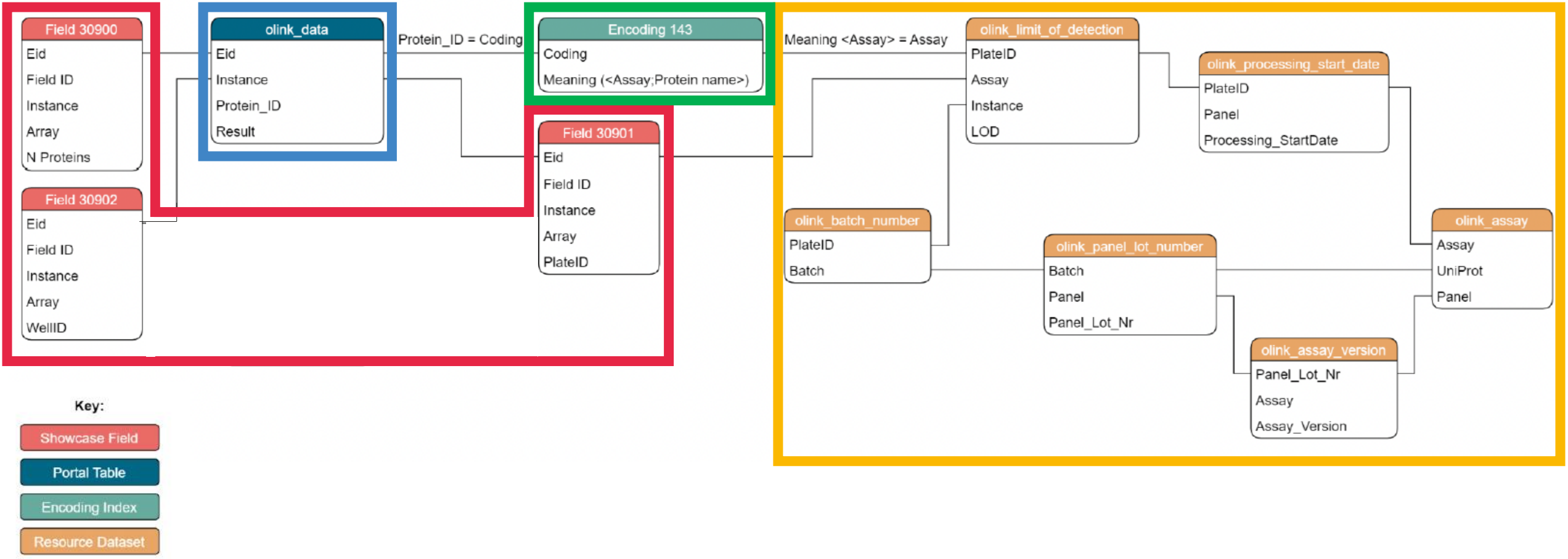
How is the data stored?



Resources explaining available dataset

Additional data files that need to be combined with the downloaded Olink NPX data

How is the data stored?

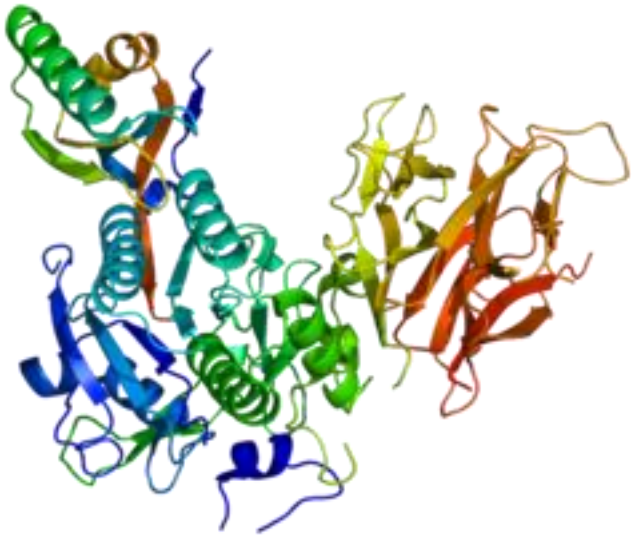


Protein QC options

- **Assay warning flag**
  - Samples that do not pass QC are indicated with “WARN”
  - Failed samples are indicated with “FAIL”
  - The flag is a reflection of sample QC and assay QC
- **Individuals:**
  - Low number of proteins measured
  - Consistently below limit of detection (LOD)
- **Proteins:**
  - Proteins with few individuals
  - Consistently below LOD for majority of individuals
  - GLIPR1 has failed a number of QC steps and should be excluded

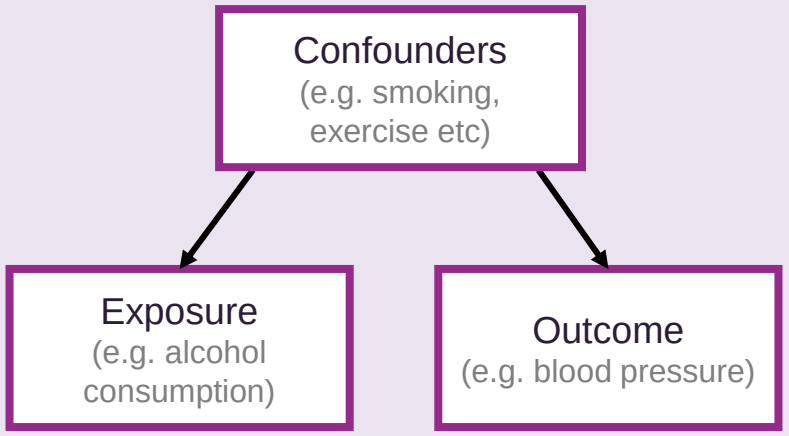
Our example: PCSK9

- **PCSK9 = proprotein convertase subtilisin/kexin type 9**
- **PCSK9 binds LDLR and marks it for destruction**
- **LDLR is responsible for the clearance of LDL-cholesterol**
- **By inhibiting PCSK9, LDLR is able to reused → increasing LDL clearance**
- **Analyses we'll perform:**
  - Observational analysis
  - Instrument selection (*cis*-pQTLs)
  - Mendelian randomisation (MR)
  - Protein specific sensitivity analysis
    - HEterogeneity In Dependent Instruments (HEIDI)
    - Colocalisation (coloc)

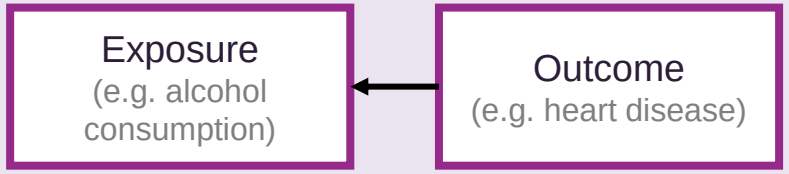


Observational analysis

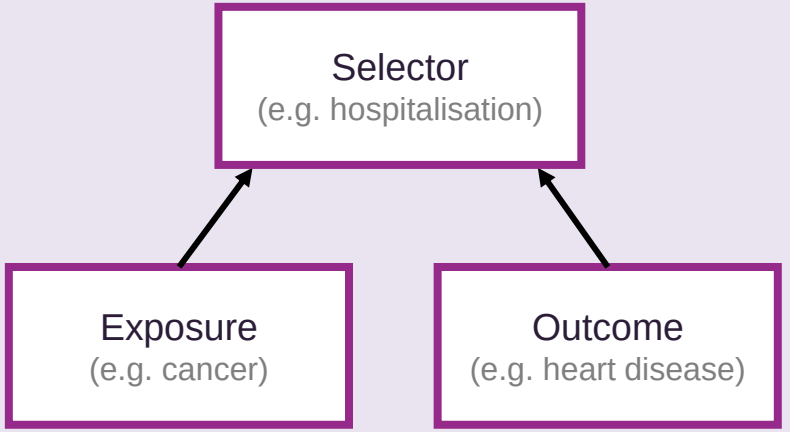
Confounding



Reverse Causation



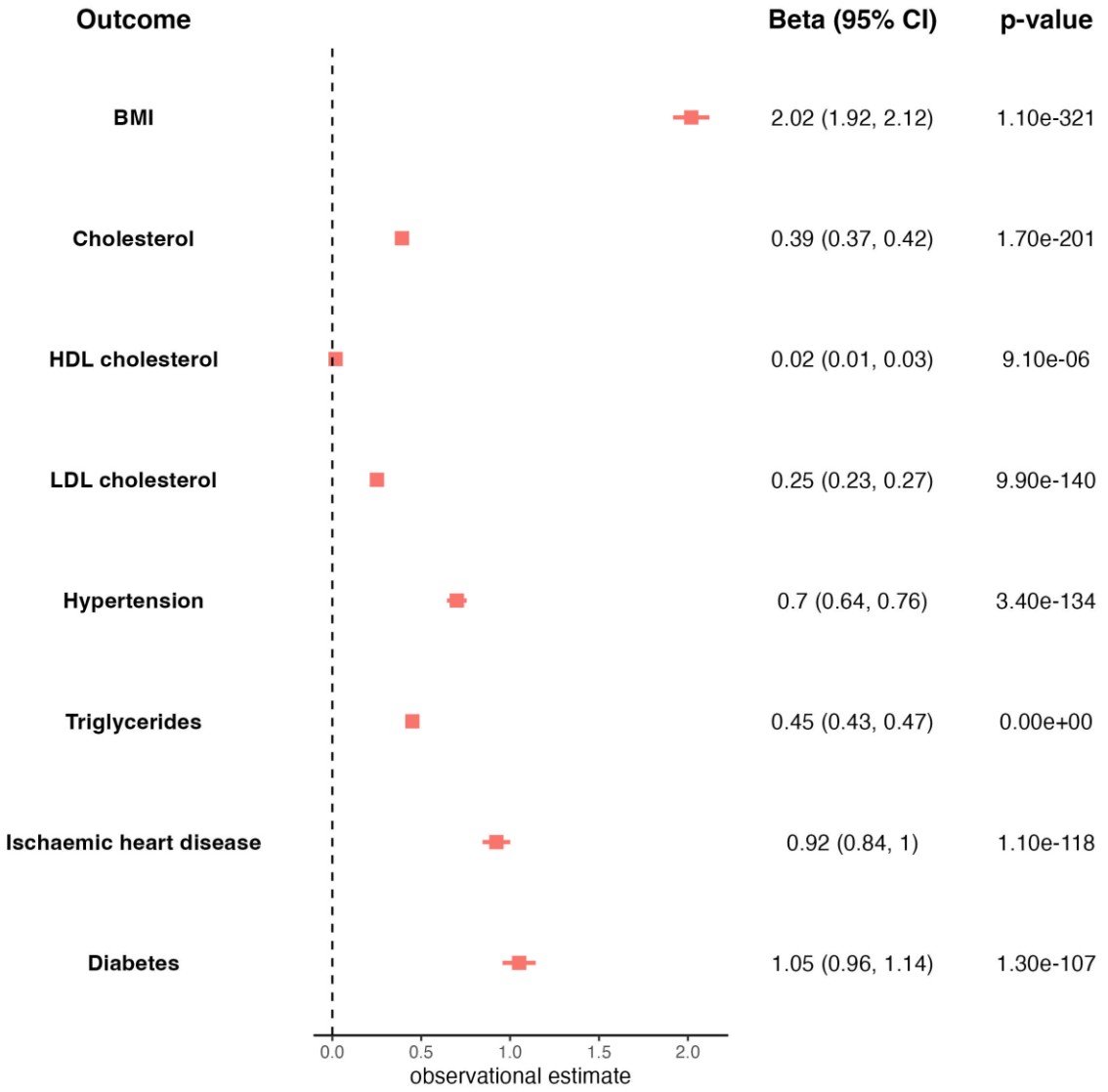
Bias



Observational analysis

- **Is ancestry important?**
  - Ancestry has the possibility to act as a confounder (especially for some traits/diseases)
- **Exclude related individuals or include relatedness in model**
- **Common covariates**
  - Age at sample collection
  - Sex
  - Sample storage duration
  - Protein batch
  - Genetic PCs
- **Remember this is a single timepoint and also a blood protein measure**
  - Proteins are dynamic (this will be more of a factor for some traits than others)

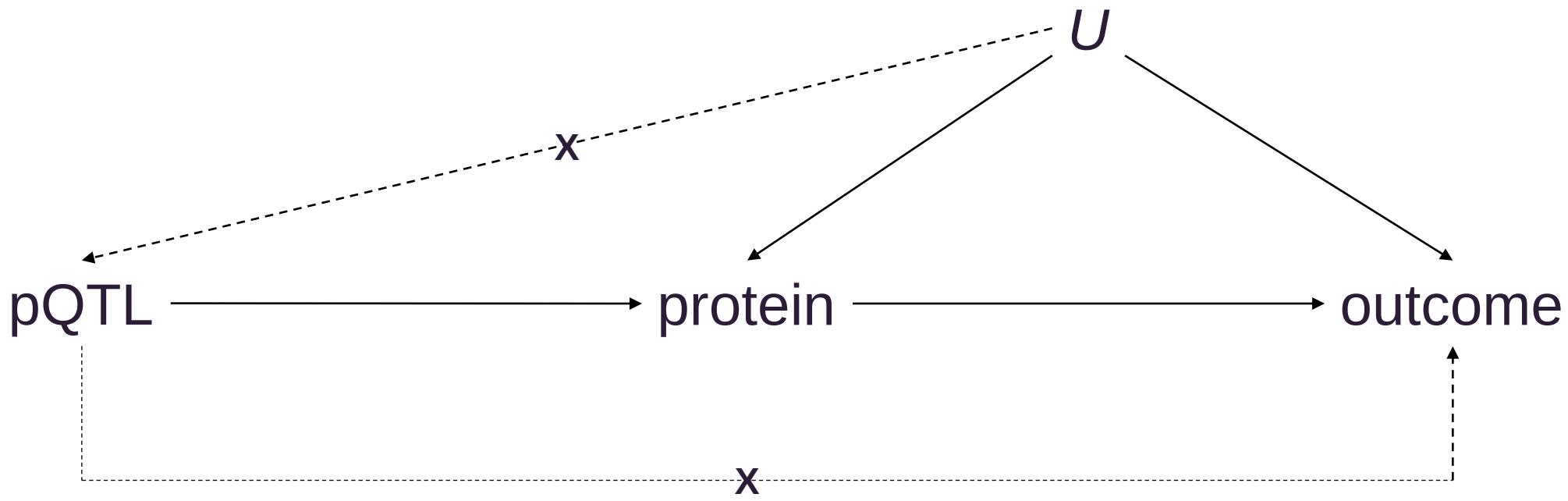
Observational example: PCSK9



outcome ~ PCSK9 + age collection + sex + protein batch + storage time + PC1-5

- Continuous = linear regression
- Binary = logistic regression

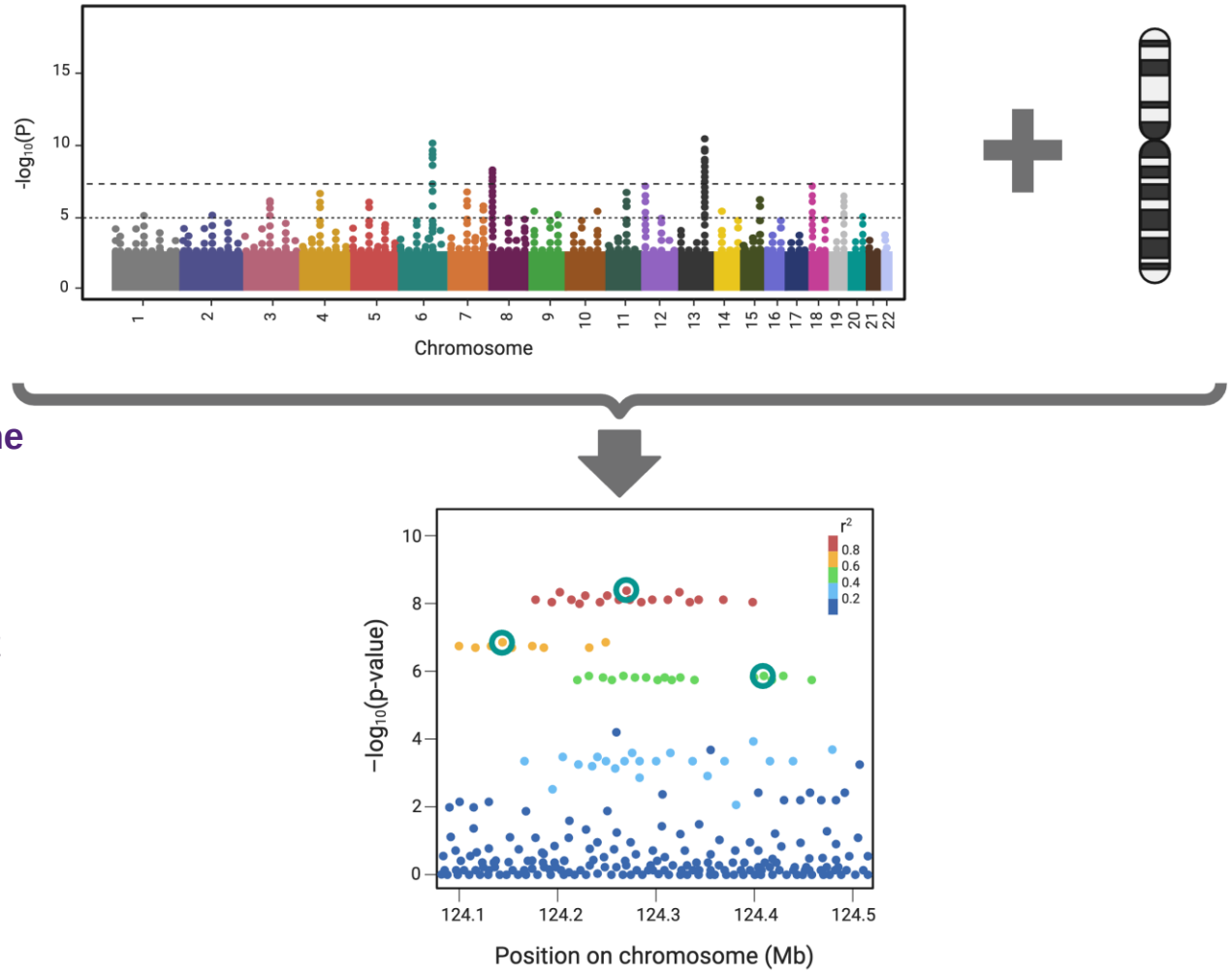
Mendelian Randomization with proteins



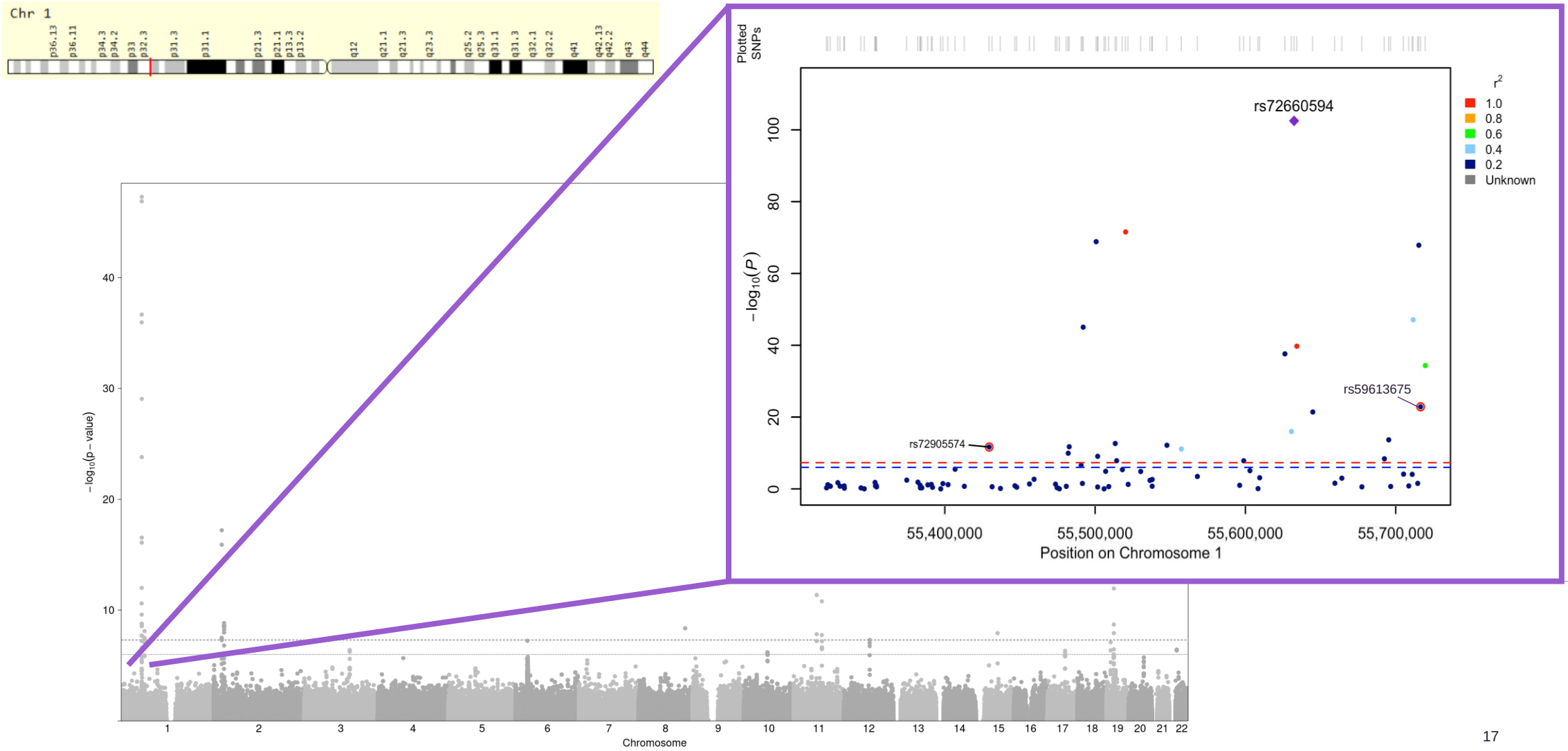
- **General preference for use of *cis*-pQTLs as instruments (excluding *trans*-pQTLs)**
- **Lack of tissue specificity**
- **Potential for reverse causality**

Instrument selection

- **Perform GWAS and select pQTLs**
  - Clumping for independent instruments
  - *cis*-pQTLs are generally ±500kb of the protein gene coding region
- ***cis*-pQTLs = within or close proximity to protein-coding gene**
  - Suggests direct influence of pQTL on protein expression or turnover
  - Considered to be more likely reflect direct or biological effect
  - More likely to have epitope effects
- ***trans*-pQTLs = not within or close to protein-coding gene**
  - Indicates indirect link between pQTL and protein expression or turnover
  - *trans*-pQTLs are likely to be pleiotropic



Instrument selection example: PCSK9



Protein MR in the UK Biobank

- **1-sample MR**
  - For binary traits/diseases check prevalence in the UKB-PPP to ensure decently powered
  - For a disease outcome exclude the “cases” from the GWAS for instrument selection
  - Perform 2 Stage Least Squares (2SLS) with *cis*-pQTLs or an allelic score
- **2-sample MR**
  - Sample overlap is likely to be a problem with most publicly available summary statistics
  - Check the F-statistic for instrument strength → most *cis*-pQTLs will be strong instruments and less likely to be impacted by this problem
  - Perform Inverse Weighted Variance (IVW) or Wald ratio depending on number of available instruments
- **Power calculator available: https://shiny.cnsgenomics.com/mRnd/**

Performing MR

1-sample MR

1. extract allele dosage for *cis*-pQTLs (e.g. plink)

2. estimate $\hat{\beta}_{pQTL-PCSK9}$ (e.g. ivreg)

3. predict values of PCSK9 using $\hat{\beta}_{pQTL-PCSK9}$ (e.g. ivreg)

4. estimate causal effect using predicted values (e.g. ivreg or glm)

2-sample MR

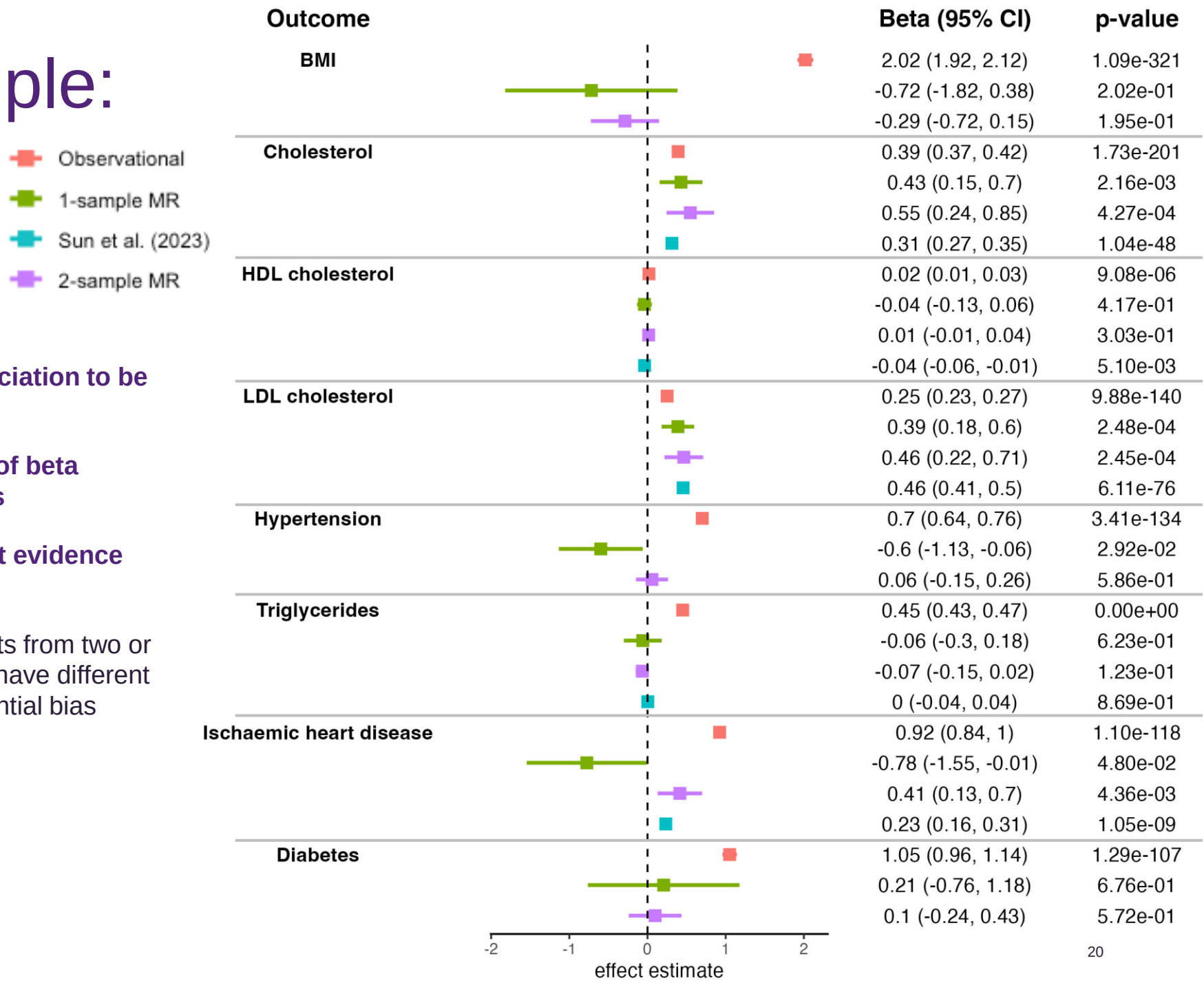
1. extract *cis*-pQTLs in PCSK9 dataset and outcome dataset

2. harmonise alleles between datasets (e.g. TwoSampleMR)

3. estimate causal effect (e.g. TwoSampleMR, GSMR)

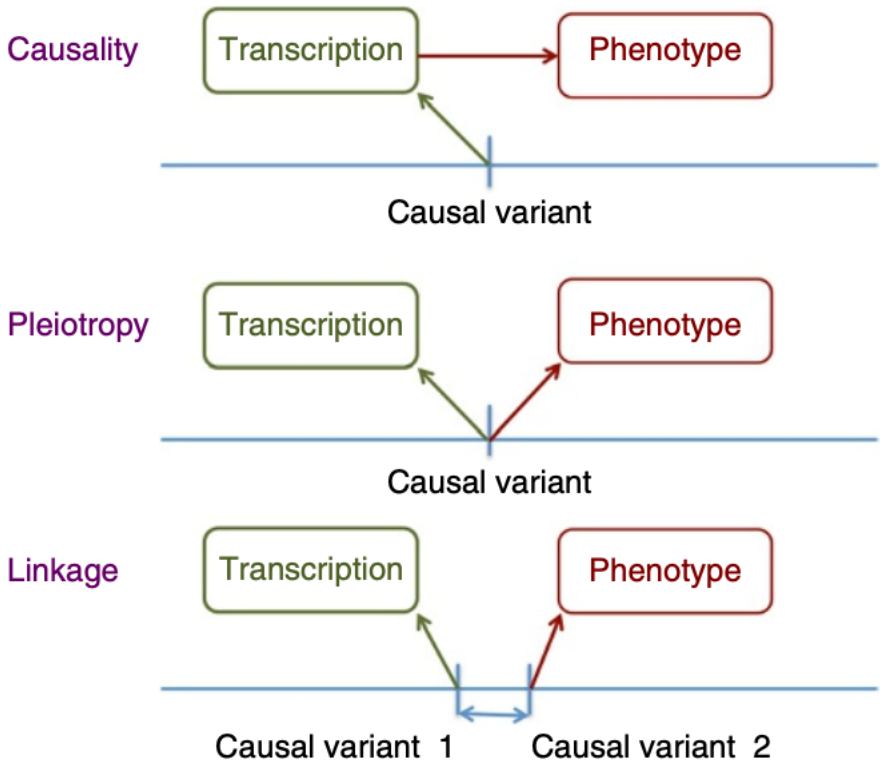
4. sensitivity analysis as appropriate (e.g. TwoSampleMR, HEIDI, coloc)

UKB MR example: PCSK9



- **We expect our observational association to be the most significant association**
- **Looking for consistent alignment of beta direction across different methods**
- **MR findings by themselves are not evidence enough of a causal association**
  - Triangulation = integration of results from two or more different study designs, that have different and unrelated key sources of potential bias

HEIDI (HEterogeneity In Dependent Instruments)



Zhu *et al.* (2015)

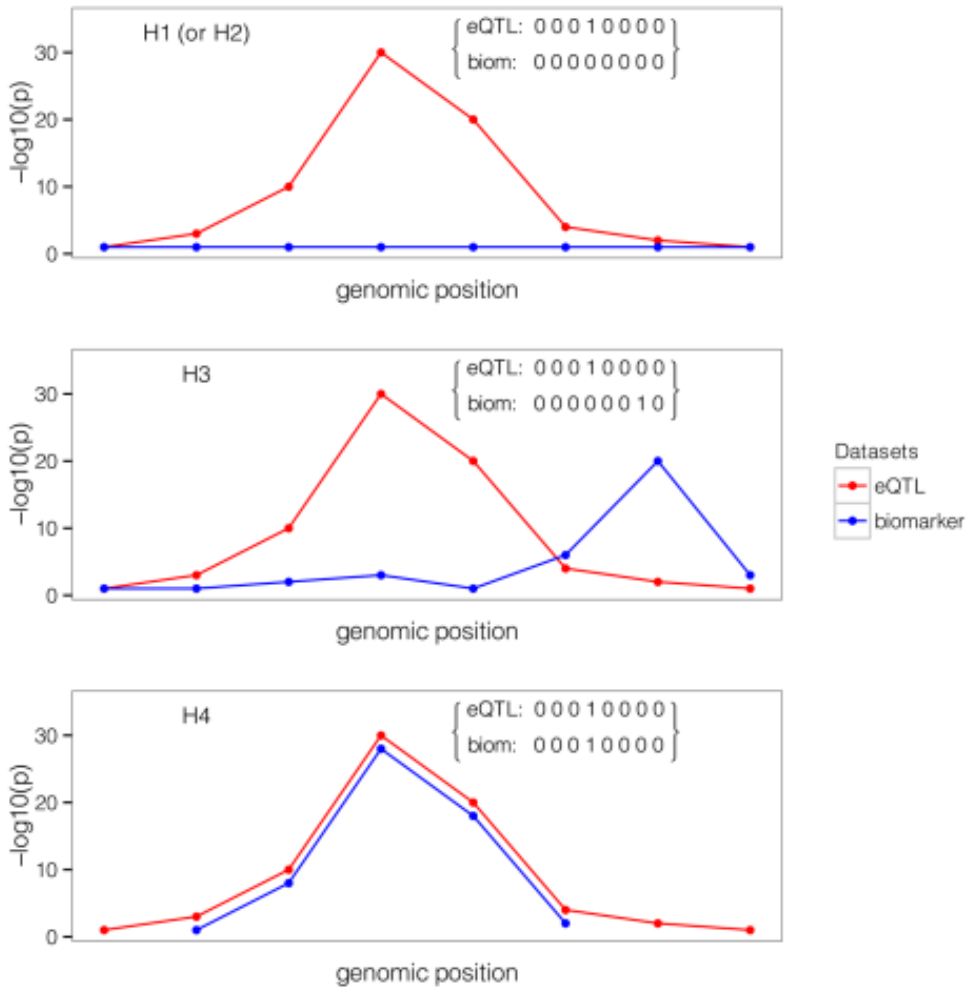
- **Heterogeneity test of valid SNPs within a region of interest**
- **Confounding due to linkage disequilibrium (LD) arises from using tagged SNPs/non-causal SNPs as instruments**
- **Interpreting results: for HEIDI a significant results ($p < 0.05$) indicates different causal variants responsible for exposure and outcome**

Example: PCSK9 and LDL-cholesterol

HEIDI

SNP	$\hat{b}_{zx}$	$\hat{b}_{zy}$	p_{HEIDI}
rs72905574	0.0360	0.0197	0.8269

Colocalisation



Giambartolomei *et al.* (2013)

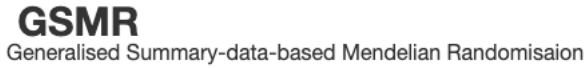
- **Testing for a shared causal variant between our exposure (protein) and outcome of interest**
  - H_0: neither trait has a genetic association in the region
  - H_1: only trait 1 has a genetic association in the region
  - H_2: only trait 2 has a genetic association in the region
  - H_3: both traits are associated, but with different causal variants
  - H_4: both traits are associated and share a single causal variant
- **coloc assumes a single causal variant, to allow for multiple causal variants preform fine mapping with SusieR**

Example: PCSK9 and LDL-cholesterol

PP.H0	PP.H1	PP.H2	PP.H3	PP.H4
0.00	$1.52e^{-279}$	0.00	0.00	1.00

Software applications

- **TwoSampleMR:** https://mrcieu.github.io/TwoSampleMR/
- **MRbase:** https://www.mrbase.org/
- **GSMR and HEIDI:** https://yanglab.westlake.edu.cn/software/gsmr/
- **AER:** https://cran.r-project.org/web/packages/AER/index.html
- **coloc:** https://chr1swallace.github.io/coloc/
- **SusieR:** https://stephenslab.github.io/susieR/reference/susie.html
- **biomaRt:** https://bioconductor.org/packages/release/bioc/html/biomaRt.html
- **plink:** https://www.cog-genomics.org/plink/



Recent publications – UK Biobank proteins

Article

Large-scale plasma proteomics comparisons through genetics and disease associations

https://doi.org/10.1038/s41586-023-06563-x
Received: 4 August 2022
Accepted: 22 August 2023
Published online: 4 October 2023
Open access
Check for updates

Grimur Hjorleifsson Eldjarn[1,7], Egil Ferkingstad[1,7], Sigrun H. Lund[1,2], Hannes Helgason[1,2], Olafur Th. Magnusson[1], Kristbjorg Gunnarsdottir[1], Thorunn A. Olafsdottir[1], Bjarni V. Halldorsson[1,3], Pall I. Olason[1], Florian Zink[1], Sigurjon A. Gudjonsson[1], Gardar Sveinbjornsson[1], Magnus I. Magnusson[1], Agnar Helgason[1,4], Asmundur Oddsson[1], Gisli H. Halldorsson[1], Magnus K. Magnusson[1,5], Saedis Saevarsdottir[1,5], Thjodbjorg Eiriksdottir[1], Gisli Masson[1], Hreinn Stefansson[1], Ingileif Jonsdottir[1,6], Hilma Holm[1], Thorunn Rafnar[1], Pall Melsted[1,2], Jona Saemundsdottir[1], Gudmundur L. Norddahl[1], Gudmar Thorleifsson[1], Magnus O. Ulfarsson[1,6], Daniel F. Gudbjartsson[1,2], Unnur Thorsteinsdottir[1,5], Patrick Sulem[1] & Kari Stefansson[1,5]

Article

Rare variant associations with plasma protein levels in the UK Biobank

https://doi.org/10.1038/s41586-023-06547-x
Received: 5 October 2022
Accepted: 15 August 2023
Published online: 4 October 2023
Open access

Ryan S. Dhindsa[1,30], Oliver S. Burren[2,30], Benjamin B. Sun[3,30], Bram P. Prins[2], Dorota Matelska[2], Eleanor Wheeler[2], Jonathan Mitchell[2], Erin Oerton[2], Ventzislava A. Hristova[1], Katherine R. Smith[2], Keren Carss[2], Sebastian Wasilewski[2], Andrew R. Harper[4], Dirk S. Paul[2], Margarete A. Fabre[2], Heiko Runz[3], Coralie Viollet[2], Benjamin Challis[5], Adam Platt[6], AstraZeneca Genomics Initiative*, Dimitrios Vitsios[2], Euan A. Ashley[7], Christopher D. Whelan[3], Menelas N. Pangalos[8], Quanli Wang[1] & Slavé Petrovski[2,9]

Cell Reports Medicine

Article

Plasma proteins and onset of type 2 diabetes and diabetic complications: Proteome-wide Mendelian randomization and colocalization analyses

Authors
Shuai Yuan, Fengzhe Xu, Xue Li, Jie Chen, Jie Zheng, Christos S. Mantzoros, Susanna C. Larsson

Plasma proteomic associations with genetics and health in the UK Biobank

https://doi.org/10.1038/s41586-023-06592-6
Received: 17 June 2022
Accepted: 31 August 2023
Published online: 4 October 2023
Open access
Check for updates

Benjamin B. Sun[1], Joshua Chiou[2,26], Matthew Traylor[3,26], Christian Benner[4,26], Yi-Hsiang Hsu[5,26], Tom G. Richardson[3,6,26], Praveen Surendran[6,26], Anubha Mahajan[4,26], Chloe Robins[7,26], Steven G. Vasquez-Grinnell[8,26], Liping Hou[9,26], Erika M. Kvikstad[8,26], Oliver S. Burren[10], Jonathan Davitte[7], Kyle L. Ferber[11], Christopher E. Gillies[12], Åsa K. Hedman[13], Sile Hu[3], Tinchi Lin[14], Rajesh Mikkilineni[15], Rion K. Pendergrass[4], Corran Pickering[16], Bram Prins[10], Denis Baird[1], Chia-Yen Chen[1], Lucas D. Ward[17], Aimee M. Deaton[17], Samantha Welsh[16], Carissa M. Willis[17], Nick Lehner[18], Matthias Arnold[18,19], Maria A. Wörheide[18], Karsten Suhre[20], Gabi Kastenmüller[18], Anurag Sethi[21], Madeleine Cule[21], Anil Raj[21], Alnylam Human Genetics*, AstraZeneca Genomics Initiative*, Biogen Biobank Team*, Bristol Myers Squibb*, Genentech Human Genetics*, GlaxoSmithKline Genomic Sciences*, Pfizer Integrative Biology*, Population Analytics of Janssen Data Sciences*, Regeneron Genetics Center*, Lucy Burkitt-Gray[16], Eugene Melamud[21], Mary Helen Black[9], Eric B. Fauman[2], Joanna M. M. Howson[3], Hyun Min Kang[12], Mark I. McCarthy[4], Paul Nioi[17], Slavé Petrovski[10,22], Robert A. Scott[6], Erin N. Smith[23], Sándor Szalma[23], Dawn M. Waterworth[24], Lyndon J. Mitnaul[12], Joseph D. Szustakowski[8,27], Bradford W. Gibson[5,27], Melissa R. Miller[2,27] & Christopher D. Whelan[1,25,27]

medRxiv
THE PREPRINT SERVER FOR HEALTH SCIENCES
Cold Spring Harbor Laboratory · BMJ · Yale

View current version of this article

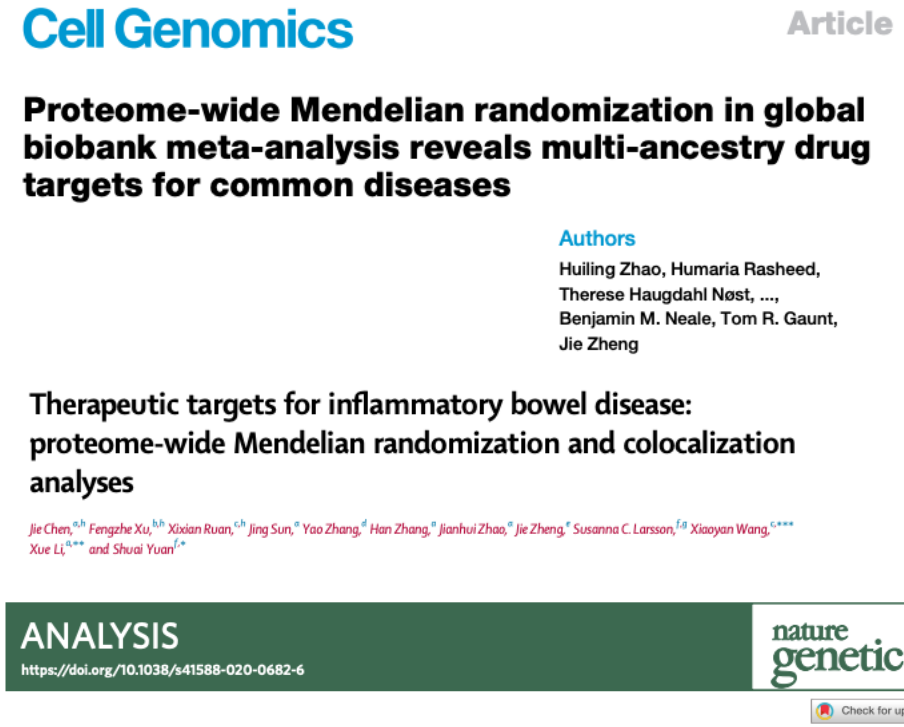
Follow this preprint

Blood protein levels predict leading incident diseases and mortality in UK Biobank

Danni A. Gadd, Robert F. Hillary, Zhana Kuncheva, Tasos Mangelis, Romi Admanit, Jake Gagnon, Tinchi Lin, Kyle Ferber, Heiko Runz, Biogen Biobank Team, Riccardo E. Marioni, Christopher N. Foley, Benjamin B. Sun

doi: https://doi.org/10.1101/2023.05.01.23288879

Recent publications – protein MR



Cell Genomics

Article

Proteome-wide Mendelian randomization in global biobank meta-analysis reveals multi-ancestry drug targets for common diseases

Authors

Huiling Zhao, Humaria Rasheed, Therese Haugdahl Nøst, ..., Benjamin M. Neale, Tom R. Gaunt, Jie Zheng

Therapeutic targets for inflammatory bowel disease: proteome-wide Mendelian randomization and colocalization analyses

Jie Chen, Fengzhe Xu, Xixian Ruan, Jing Sun, Yao Zhang, Han Zhang, Jianhui Zhao, Jie Zheng, Susanna C. Larsson, Xiaoyan Wang, Xue Li and Shuai Yuan

ANALYSIS

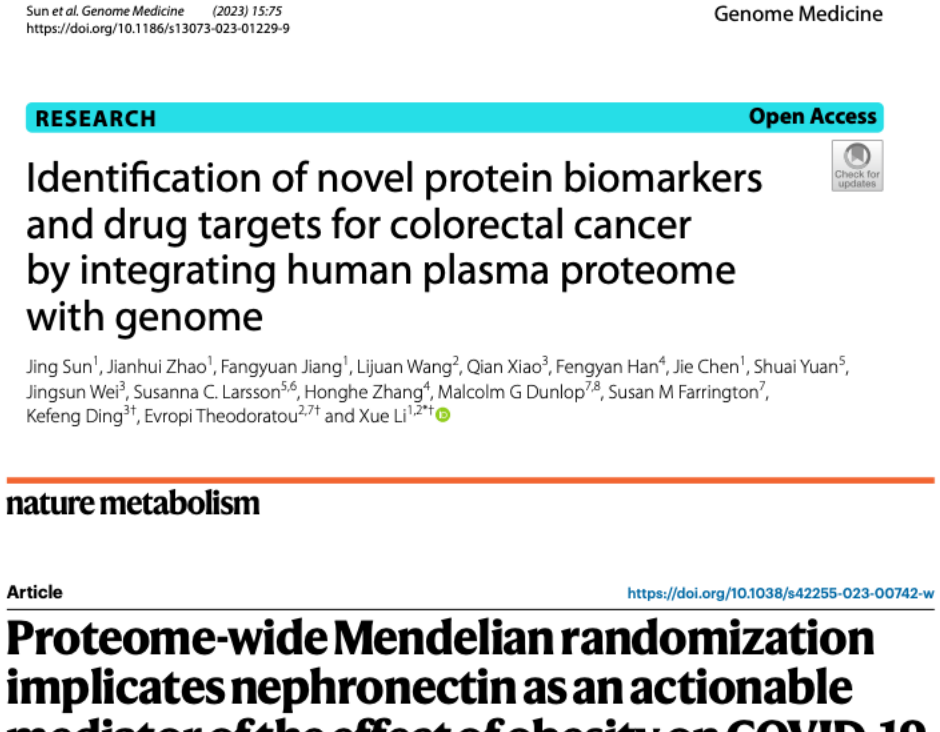
https://doi.org/10.1038/s41588-020-0682-6

nature genetics

Check for updates

Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases

Jie Zheng, Valeriia Haberland, Denis Baird, Venexia Walker, Philip C. Haycock, Mark R. Hurle, Alex Gutteridge, Pau Erola, Yi Liu, Shan Luo, Jamie Robinson, Tom G. Richardson, James R. Staley, Benjamin Elsworth, Stephen Burgess, Benjamin B. Sun, John Danesh, Heiko Runz, Joseph C. Maranville, Hannah M. Martin, James Yarmolinsky, Charles Laurin, Michael V. Holmes, Jimmy Z. Liu, Karol Estrada, Rita Santos, Linda McCarthy, Dawn Waterworth, Matthew R. Nelson, George Davey Smith, Adam S. Butterworth, Gibran Hemani, Robert A. Scott and Tom R. Gaunt



Sun et al. Genome Medicine (2023) 15:75
https://doi.org/10.1186/s13073-023-01229-9

Genome Medicine

RESEARCH Open Access

Identification of novel protein biomarkers and drug targets for colorectal cancer by integrating human plasma proteome with genome

Jing Sun, Jianhui Zhao, Fangyuan Jiang, Lijuan Wang, Qian Xiao, Fengyan Han, Jie Chen, Shuai Yuan, Jingsun Wei, Susanna C. Larsson, Honghe Zhang, Malcolm G Dunlop, Susan M Farrington, Kefeng Ding, Evropi Theodoratou and Xue Li

nature metabolism

Article

https://doi.org/10.1038/s42255-023-00742-w

Proteome-wide Mendelian randomization implicates nephronectin as an actionable mediator of the effect of obesity on COVID-19 severity

Received: 17 June 2022

Accepted: 10 January 2023

Published online: 20 February 2023

Check for updates

Satoshi Yoshiji, Guillaume Butler-Laporte, Tianyuan Lu, Julian Daniel Sunday Willett, Chen-Yang Su, Tomoko Nakanishi, David R. Morrison, Yiheng Chen, Kevin Liang, Michael Hultström, Yann Ilboudo, Zaman Afrasiabi, Shanshan Lan, Naomi Duggan, Chantal DeLuca, Mitra Vaezi, Chris Tselios, Xiaoqing Xue, Meriem Bouab, Fangyi Shi, Laetitia Laurent, Hans Markus Münter, Marc Afilalo, Jonathan Afilalo, Vincent Mooser, Nicholas J. Timpson, Hugo Zeberg, Sirui Zhou, Vincenzo Forgetta, Yossi Farjoun & J. Brent Richards

Acknowledgements

Examples performed using data provided by UK Biobank Application 53641

Helpful papers

- **Proteome GWAS:** Suhre K, McCarthy MI, Schwenk JM. Genetics meets proteomics: perspectives for large population-based studies. Nat Rev Genet. 2021 Jan;22(1):19-37. doi: 10.1038/s41576-020-0268-2. Epub 2020 Aug 28. PMID: 32860016.
- **Mendelian randomisation:** Lawlor DA, Harbord RM, Sterne JA, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Stat Med. 2008 Apr 15;27(8):1133-63. doi: 10.1002/sim.3034. PMID: 17886233.
- **MR Egger:** Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int J Epidemiol. 2015 Apr;44(2):512-25. doi: 10.1093/ije/dyv080. Epub 2015 Jun 6. PMID: 26050253; PMCID: PMC4469799.
- **Weighted Median:** Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. Genet Epidemiol. 2016 May;40(4):304-14. doi: 10.1002/gepi.21965. Epub 2016 Apr 7. PMID: 27061298; PMCID: PMC4849733.
- **Modal Estimator:** Fernando Pires Hartwig, George Davey Smith, Jack Bowden, Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption, *International Journal of Epidemiology*, Volume 46, Issue 6, December 2017, Pages 1985–1998, https://doi.org/10.1093/ije/dyx102
- **Colocalisation:** Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, Plagnol V. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 2014 May 15;10(5):e1004383. doi: 10.1371/journal.pgen.1004383. PMID: 24830394; PMCID: PMC4022491.
- **Fine mapping:** Gao Wang, Abhishek Sarkar, Peter Carbonetto, Matthew Stephens, A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, Volume 82, Issue 5, December 2020, Pages 1273–1300, http://doi.org/10.1111/rssb.12388
- **SMR & HEIDI:** Zhu, Z., Zhang, F., Hu, H. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48**, 481–487 (2016). https://doi.org/10.1038/ng.3538

Useful sites

- **MR Dictionary:** https://mr-dictionary.mrcieu.ac.uk/
- **PCG Video Textbook:** https://pgcanalytics.github.io/pgcvideotextbook/welcome.html
- **Boulder 2021:** https://www.colorado.edu/ibg/international-workshop/2021-syllabus

Pseudocode: 1-sample MR

```
# load libraries
library(AER)

# run first stage
first_stage = iv_reg(exposure ~ instruments, data)
fitted_values = predict(first_stage, new_data)

# run second stage
# for continuous outcome
second_stage = ivreg(outcome ~ fitted_values, new_data)

# for binary outcome
second_stage = glm(outcome ~ fitted_values, family = “binomial”, new_data)
```

Pseudocode: 2-sample MR

```
# load libraries

library(TwoSampleMR)


# format data using TwoSampleMR

exp_dat = format_data(exp_input, snp_col, beta_col, se_col, eaf_col, effect_allele_col, other_allele_col, pval_col,
                      chr_col, pos_col)

out_dat = format_data(out_input, type="outcome", snp_col, beta_col, se_col, eaf_col, effect_allele_col, other_allele_col,
                      pval_col, chr_col, pos_col)


# harmonise exposure and outcome data sets

har_dat = harmonise_data(exp_dat, out_dat)


# perform MR

mr_res = mr(har_dat)
```