Testing epistatic effect in complex trait using related intermediate trait

Yanyu Liang

University of Chicago

yanyul@uchicago.edu

March 13, 2018

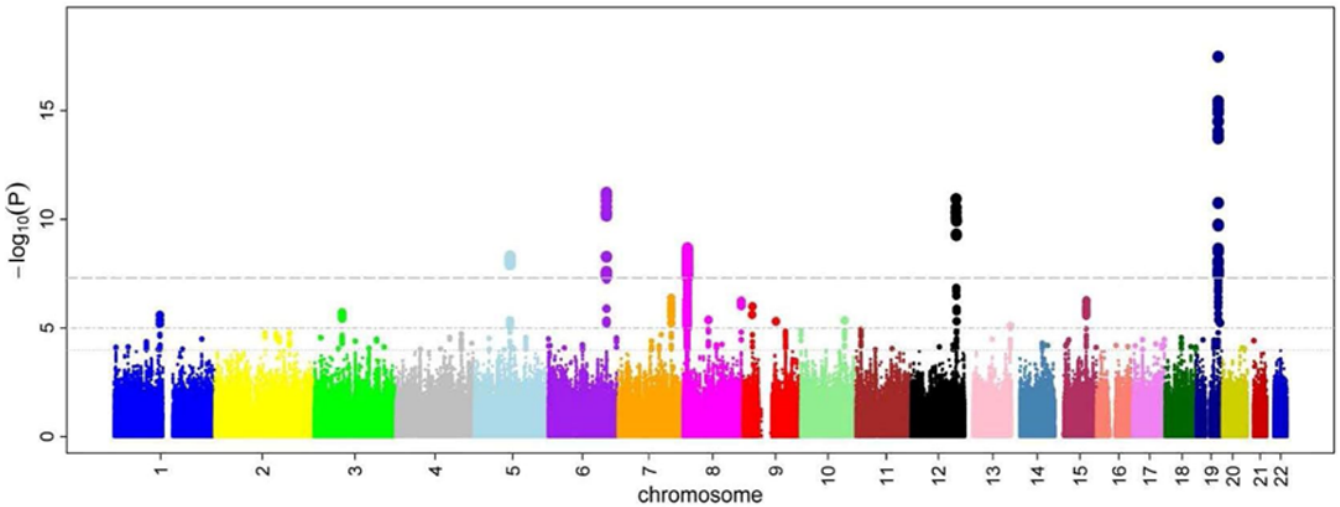
Overview

Mendelian disease and complex trait

- Caused by single variant on a gene (or regulatory region), *i.e.* Phenylketonuria is caused by abnormal *PAH* gene which encodes phenylalanine hydroxylase
- Highly heritable and is inherited in recessive or dominant way, *i.e.* Mendel's laws
- They are in general rare

- Caused by many variants along with environmental factors, *i.e.* high blood pressure, type II diabetes
- Also heritable but no clear pattern of recessive or dominant
- They are more common in population, *i.e.* more than 30% people (older than 20) have high blood pressure in U.S.

GWAS and complex trait

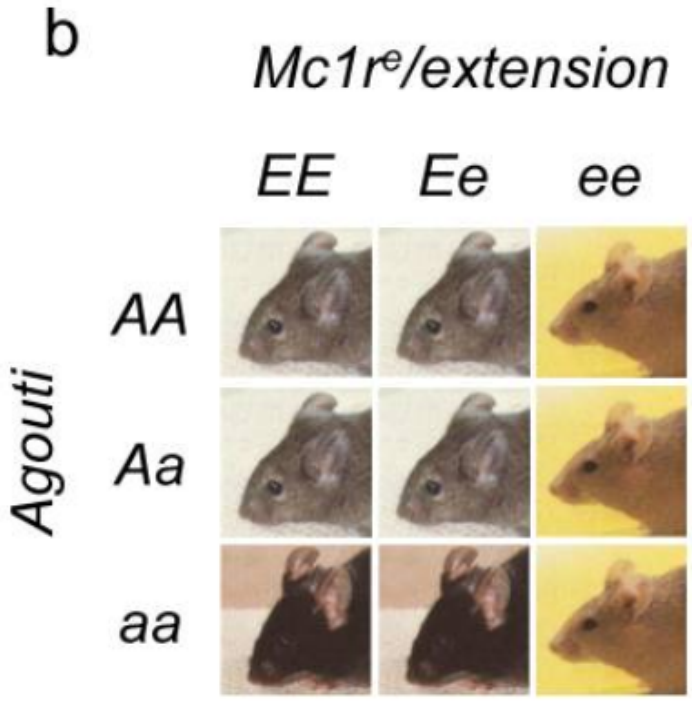


- Genome-wide association study (GWAS)
  - Case-control study
  - Look for association between disease status and genotype
  - Perform test each locus at a time
- GWAS has been extensively used to identify causal genes for complex trait

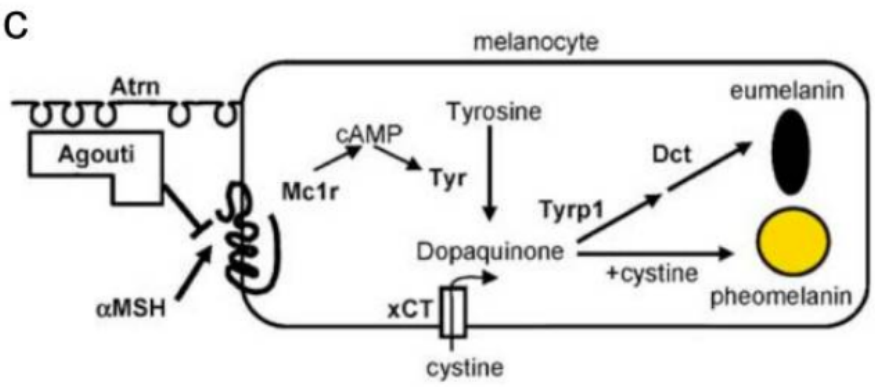
Genetic background and disease

- Carrying pathogenic variant may not lead to disease, *incomplete penetrance* (ExAC [4])
- In complex trait, risk loci only explain a small fraction of disease risk
- Explanations
  - Rare variant also contribute
  - Weak effect loci also contribute additively (polygenic assumption)
  - The effect of risk loci depends on other loci non-additively (epistatic effect of genetic background)

Epistatic effect - an example of color coating [6]



- The effect of *Agouti* is masked by $Mc1r^e$



- In the pathway *Agouti* is upstream of $Mc1r^e$ and the phenotypic effect of *Agouti* acts through $Mc1r^e$. So, when $Mc1r^e$ loses function, *Agouti* 'loses' its effect

Epistatic effect

- Epistasis occurs when the effect of one gene depends on other genes, modifiers (*Agouti* depends on $Mc1r^e$ to have normal function)
- More generally, the effect of loci as a whole is different from their individual effect

Question

1. In complex trait, is there any epistatic effect?
2. If so, how to identify them from current data?

Ideas

1. Beside the epistatic signals that have already be found by performing SNP-SNP interaction test, epistatic effect makes biological sense.
   - Molecules in the same pathway may act depending on each others
   - The outcome of one pathway may modify the effect of the gene in another pathway
2. Currently, SNP-SNP interaction is computationally tractable but with huge statistically burden. It is better to shrink SNP searching space and aggregate signals in some way (hypothesis-based)

Hypothesis

For some complex trait, the related pathway (let say they are **intermediate traits**) is known. And these intermediate traits potentially modify the effect of risk gene. For instances,

- Low density lipo-protein (LDL) level may modify coronary heart disease (CHD) risk
- The sensitivity of immune system (baseline activity) may modify Crohn's disease (CD) risk
- Baseline glucose level may modify type II diabetes (T2D) risk

Method overview

- Polygenic risk score (PRS) is computed from genotype data
- Obtain a set of loci which are potentially modified by the intermediate trait
- Test interaction between $\hat{I}$ and X_j on the disease risk Y

- LDpred [8] is used to obtain posterior mean effect $\bar{\beta}_i$ and $\hat{I} = \sum_i \bar{\beta}_i X_i$
- i) GWAS significant SNPs;
  ii) SNPs that are likely to act through the intermediate trait
- $\text{logit}\Pr(Y = 1|\hat{I}, X_j) = \alpha + \beta_I \hat{I} + \beta_j X_j + \gamma \hat{I} X_j$ using `plink`

Data overview

- Genotype of a case-control study, WTCCC (2009) [2]
  - Three diseases along with two shared controls: CHD, CD, T2D (sample size 4000-5000 each)
- Summary statistic of intermediate traits
  1. LDL, HDL, triglycerides (TG) [3]
  2. Insulin-like growth factor 1 (IGF1) [7]
  3. White blood cell count (WBC) [1]
  4. Fast glucose, fast insulin [5]

Intermediate trait PRS is correlated with disease status

We first test whether intermediate trait PRS $\hat{I}$ is correlated with the disease of interest Y. The following pairs are tested

- CHD vs. LDL, HDL, TG
- CD vs. IGF1, WBC
- T2D vs. FastGlu, FastInsulin

In short, we find the following significant association (p-value $<$ 0.05) under $\text{logit}(\Pr(Y = 1|\hat{I})) = \alpha + \beta\hat{I}$

- CHD $\leftarrow^{-}$ LDL, CHD $\leftarrow^{+}$ TG, HDL (n.s.)
- CD $\leftarrow^{+}$ WBC, IGF1 (n.s.)
- T2D $\leftarrow^{+}$ FastGlu, T2D $\leftarrow^{+}$ FastInsulin

Test interaction on GWAS significant hits: CHD - LDL

CHR	SNP	BP	A1	TEST	NMISS	OR	STAT	P	GWAS.P
19	rs7250581	0	A	ADDxPRS	4864	1.237	2.395	0.01664	8.756e-06
12	rs2398486	0	T	ADDxPRS	4864	1.31	2.232	0.02561	3.841e-05
9	rs564398	0	C	ADDxPRS	4864	1.161	2.147	0.03183	9.123e-09
9	rs7865618	0	G	ADDxPRS	4864	1.151	2.041	0.04126	1.476e-09
9	rs7049105	0	G	ADDxPRS	4864	0.8724	-2	0.04552	1.426e-09
12	rs2167512	0	A	ADDxPRS	4864	1.285	1.987	0.04688	2.798e-05
9	rs10965215	0	A	ADDxPRS	4864	0.8768	-1.926	0.05413	6.17e-10
21	rs2838756	0	C	ADDxPRS	4864	1.158	1.924	0.05438	2.743e-05
9	rs523096	0	G	ADDxPRS	4864	1.134	1.818	0.06899	1.382e-06
9	rs10965219	0	G	ADDxPRS	4864	0.8938	-1.656	0.09781	8.495e-11

- Most of them locate near CDKN2B antisense RNA 1 (*CDKN2B-AS1*) and the rest is at *TMEM132D* region, with little effect on LDL
- The effect of the locus is masked by high LDL level (namely the sign of β_j is opposite to γ)

Test interaction on GWAS significant hits: CD - WBC

CHR	SNP	BP	A1	TEST	NMISS	OR	STAT	P	GWAS.P
17	rs3816769	0	C	ADDxPRS	4686	0.9517	-2.578	0.00994	2.27e-05
17	rs1026916	0	A	ADDxPRS	4686	0.9523	-2.539	0.01112	1.649e-05
17	rs7211777	0	G	ADDxPRS	4686	0.9525	-2.532	0.01136	2.435e-05
19	rs8111071	0	G	ADDxPRS	4686	1.071	2.091	0.03656	9.812e-06
16	rs2076756	0	G	ADDxPRS	4686	0.9609	-2.029	0.04247	1.263e-14
6	rs9469615	0	C	ADDxPRS	4686	1.058	1.973	0.04848	1.635e-05
5	rs16869934	0	T	ADDxPRS	4686	1.042	1.925	0.05418	5.284e-11
5	rs10512734	0	G	ADDxPRS	4686	1.036	1.689	0.09127	1.889e-10
16	rs2066843	0	T	ADDxPRS	4686	0.9693	-1.637	0.1015	6.319e-13

- Some loci locate near *NOD2* (IBD gene), *STAT3* (related o immune response), with little effect on WBC
- The effect of the *NOD2* locus is masked by high WBC level
- The effect of the *STAT3* locus is enhanced by high WBC level (namely the sign of β_j is the same as γ)

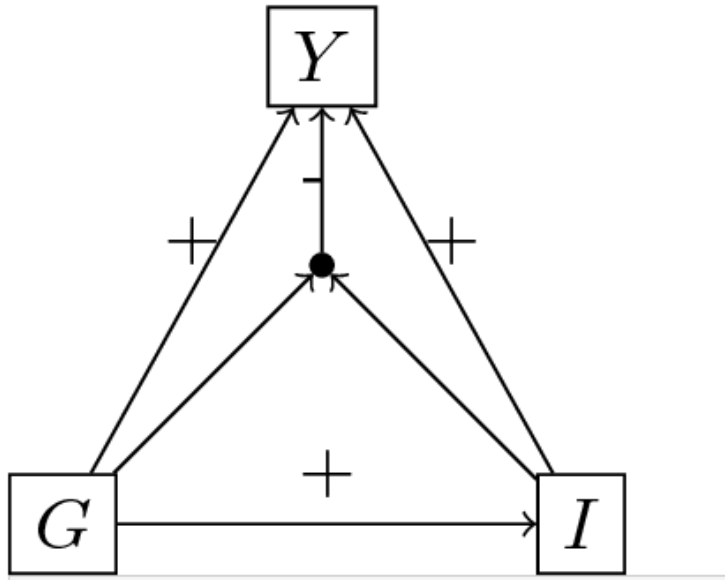
Test interaction on GWAS significant hits: T2D - FastInsulin

CHR	SNP	BP	A1	TEST	NMISS	OR	STAT	P	GWAS.P
14	rs8012854	0	G	ADDxPRS	4862	1.702	2.838	0.004541	4.036e-05
3	rs440646	0	G	ADDxPRS	4862	1.359	2.146	0.03187	3.317e-05
16	rs11075123	0	A	ADDxPRS	4862	0.7449	-1.851	0.06422	2.377e-05
12	rs7961581	0	C	ADDxPRS	4862	1.312	1.782	0.07468	5.76e-06

- *rs440646* locates in the intronic region of *HRH1* which acts as a stimulator of insulin-induced adipogenesis
- It has little effect to fast insulin level and the effect on T2D is enhanced by high fast insulin level

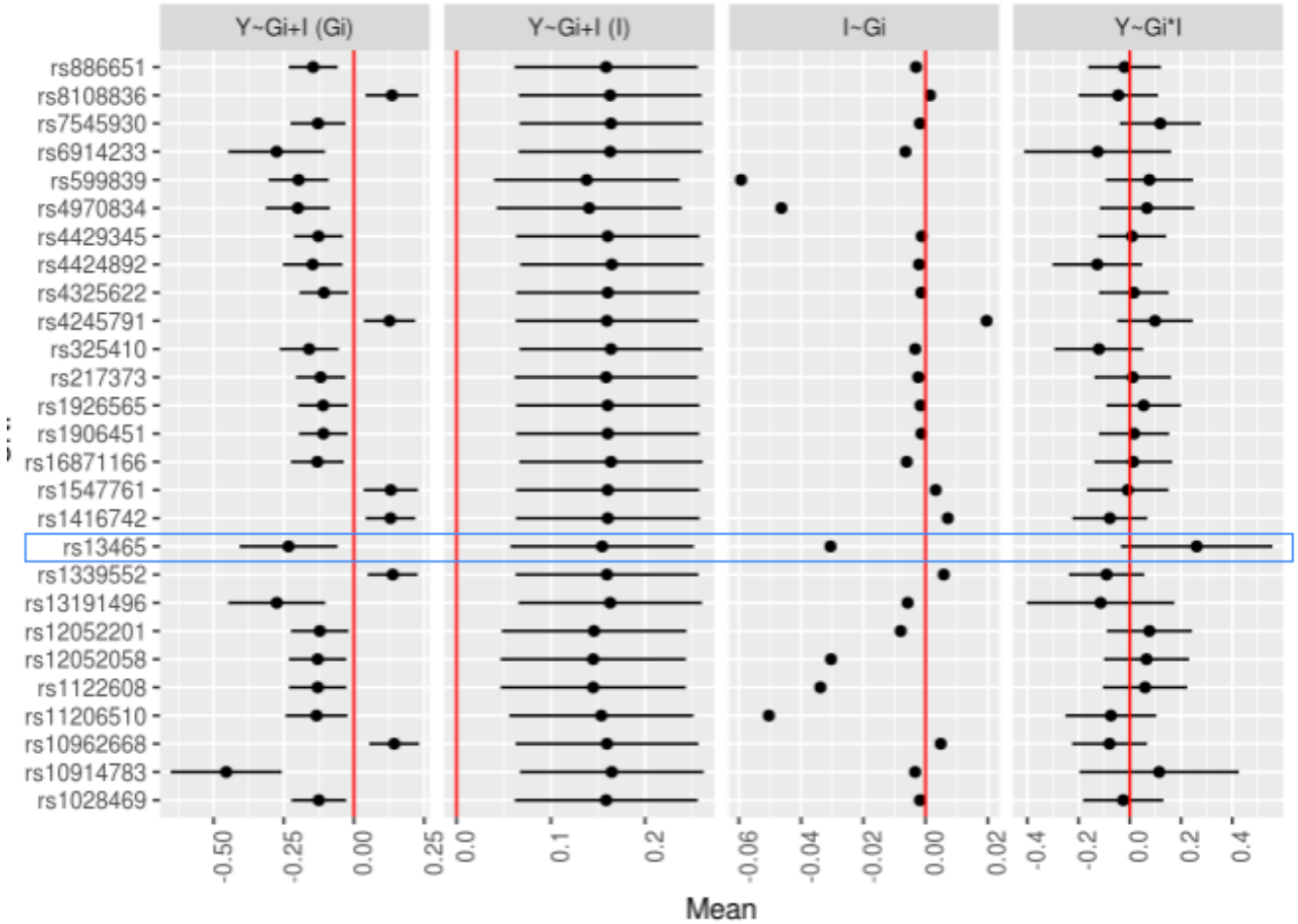
Test interaction on SNPs that potentially act through the intermediate trait

	Y ~ I	Y ~ Gi	I ~ Gi
case1	+	+	+
case2	+	-	-
case3	-	+	-
case4	-	-	+



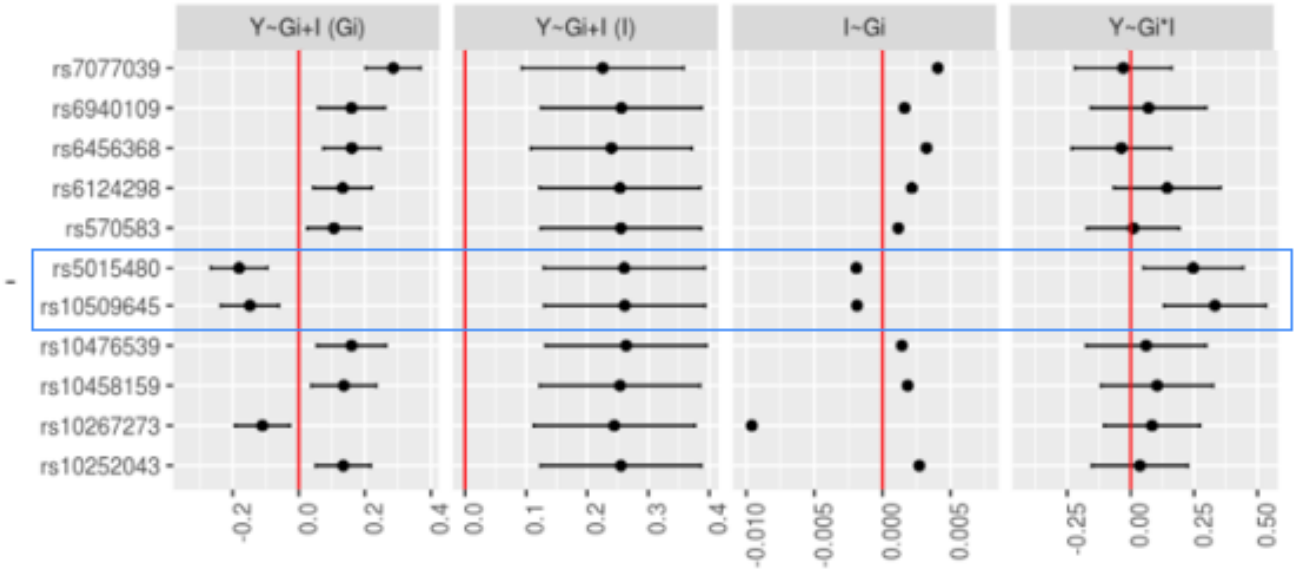
- If SNP acts through the pathway, it is likely to be modified by the pathway
- The marginal effect of SNP on disease and SNP on intermediate trait should be consistent with the effect of intermediate trait on disease
- We collect SNPs that act consistently as a **consistent** set

Test interaction on consistent SNPs: CHD - LDL



- *rs13465* locates in the intronic region of *ILF3*. Its protective effect is masked by high LDL level
- Interestingly, it has been reported that *ILF3* affects myocardial infarction risk only when LDL level is low [9].

Test interaction on consistent SNPs: T2D - FastGlu



- *rs10509645* and *rs5015480* locate in the intronic region of *IDE* and *HHEX* respectively (the genes are both T2D candidate genes)
- Their protective effect is masked by high fast glucose level

Summary

- We test the modifier effect of several intermediate trait on a subset of SNPs (GWAS hits or consistent SNPs)
- Overall, there are some interaction signals which indicates that the selected intermediate does as a whole modify the effect of some risk genes
- In particular,
  - For GWAS hits, the epistatic effect either masks or enhances the effect of the locus when the baseline pathway level is high
  - For consistent loci, the epistatic effect tend to mask the effect

Issues and future directions

1. Sample size is small and the interaction signal is weak
2. Validate the signal
3. Hard to interpret the result: what biological insight we can get out of such interaction

1. Use larger data sets (UK Biobank?)
2. Seek some well-established epistasis in complex trait? Simulation analysis?
3. Perform gene-based analysis. *I.e.* to test the modifier effect of intermediate trait on how gene expression affect disease risk

The End

References I

William J Astle, Heather Elding, Tao Jiang, Dave Allen, Dace Ruklisa, Alice L Mann, Daniel Mead, Heleen Bouman, Fernando Riveros-Mckay, Myrto A Kostadima, et al.

The allelic landscape of human blood cell trait variation and links to common complex disease.

Cell, 167(5):1415–1429, 2016.

Wellcome Trust Case Control Consortium et al.

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.

Nature, 447(7145):661, 2007.

Johannes Kettunen, Ayşe Demirkan, Peter Würtz, Harmen HM Draisma, Toomas Haller, Rajesh Rawal, Anika Vaarhorst, Antti J Kangas, Leo-Pekka Lyytikäinen, Matti Pirinen, et al.

Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of lpa.

Nature communications, 7:11122, 2016.

References II

Monkol Lek, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H O'Donnell-Luria, James S Ware, Andrew J Hill, Beryl B Cummings, et al.

Analysis of protein-coding genetic variation in 60,706 humans.

Nature, 536(7616):285, 2016.

Alisa K Manning, Marie-France Hivert, Robert A Scott, Jonna L Grimsby, Nabila Bouatia-Naji, Han Chen, Denis Rybin, Ching-Ti Liu, Lawrence F Bielak, Inga Prokopenko, et al.

A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance.

Nature genetics, 44(6):659, 2012.

Patrick C Phillips.

Epistasis-the essential role of gene interactions in the structure and evolution of genetic systems.

Nature Reviews Genetics, 9(11):855, 2008.

References III

Bram P Prins, Karoline B Kuchenbaecker, Yanchun Bao, Melissa Smart, Delilah Zabaneh, Ghazaleh Fatemifar, Jian'an Luan, Nick J Wareham, Robert A Scott, John RB Perry, et al.

Genome-wide analysis of health-related biomarkers in the uk household longitudinal study reveals novel associations.

Scientific reports, 7(1):11008, 2017.

Bjarni J Vilhjálmsson, Jian Yang, Hilary K Finucane, Alexander Gusev, Sara Lindström, Stephan Ripke, Giulio Genovese, Po-Ru Loh, Gaurav Bhatia, Ron Do, et al.

Modeling linkage disequilibrium increases accuracy of polygenic risk scores.

The American Journal of Human Genetics, 97(4):576–592, 2015.

Tetsuro Yoshida, Kimihiko Kato, Mitsutoshi Oguri, Hideki Horibe, Toshiki Kawamiya, Kiyoshi Yokoi, Tetsuo Fujimaki, Sachiro Watanabe, Kei Satoh, Yukitoshi Aoyagi, et al.

Association of polymorphisms of btn2a1 and ilf3 with myocardial infarction in japanese individuals with different lipid profiles.

Molecular medicine reports, 4(3):511–518, 2011.