On Genes and Environments

rGE, $G \times E$, and how to study them

Rafael Ahlskog

Uppsala Universitet

2023-11-10

Prelude...

- I will spend about a third of this lecture on gene-environment correlation, and the rest on gene-environment interaction.
- Focus will be on a) a conceptual understanding of GxE effects, and b) methodological tools/designs to investigate them in a more/less robust manner using PGIs.
- Intention is to provide concrete examples (mostly from my own research), and to show how you would go about applying the GxE "paradigm" to your own research questions.
- Will be information dense, but I will make sure to start from square one.
- Slides will contain a fair amount of text – this is intentional; will make them more useful if you choose to go back to them later.

Gene-environment correlation, or *rGE*

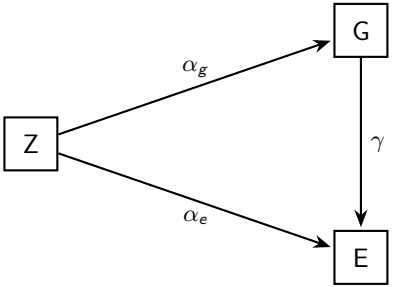
- Definition: gene-environment correlation is whenever genotypes are not randomly distributed across environmental conditions.
- Causes two problems:
  - Makes it harder to identify environmental effects, since they can be genetically confounded
  - Makes it harder to identify genetic effects, since they can be environmentally confounded
- We have seen many examples of this already during this week, but here I will:
  - Introduce more detailed terminology about different sources of *rGE*
  - Talk about which of these sources are problematic and which are not (necessarily)

A typology of *rGE*

As Martin said yesterday, gene-environment correlation is often said to come in three different flavors:

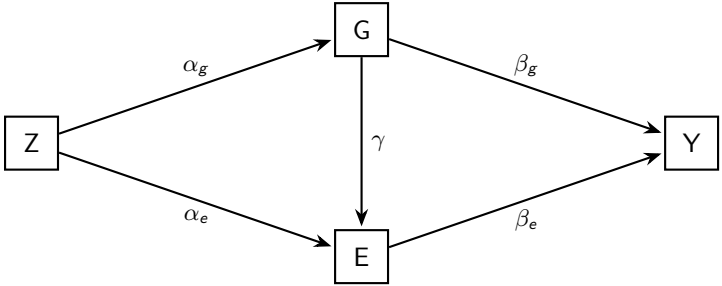
- Passive *rGE* – your environment and genotype are correlated because you inherited both. In a broader sense: any correlation between genotype and environment that is non-causal (i.e. not because your genotype causes your environment). This is the part that is definitively problematic for identification of genetic effects!
- Active *rGE* – your genotype makes you construct, seek out, or select into certain environmental exposures. These are causal effects of genetics, and makes certain environmental variables a part of the causal chain between (i.e. a mechanism for) genetic effects on some other downstream outcome.
- Evocative (or reactive *rGE*): some genetically influenced trait makes people treat you differently, i.e. it "evokes" a response from the environment. These are also causal genetic effects mediated by an environmental variable (other people's treatment of you).

A simple causal framework I



- Passive rGE (non-causal): $\alpha_g \alpha_e$
- Active + evocative rGE (causal): γ
- Complete rGE: $\alpha_g \alpha_e + \gamma$

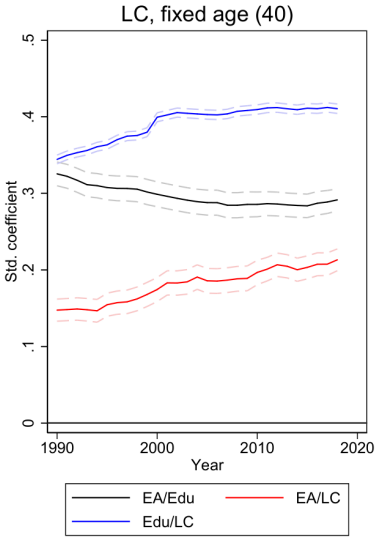
A simple causal framework II

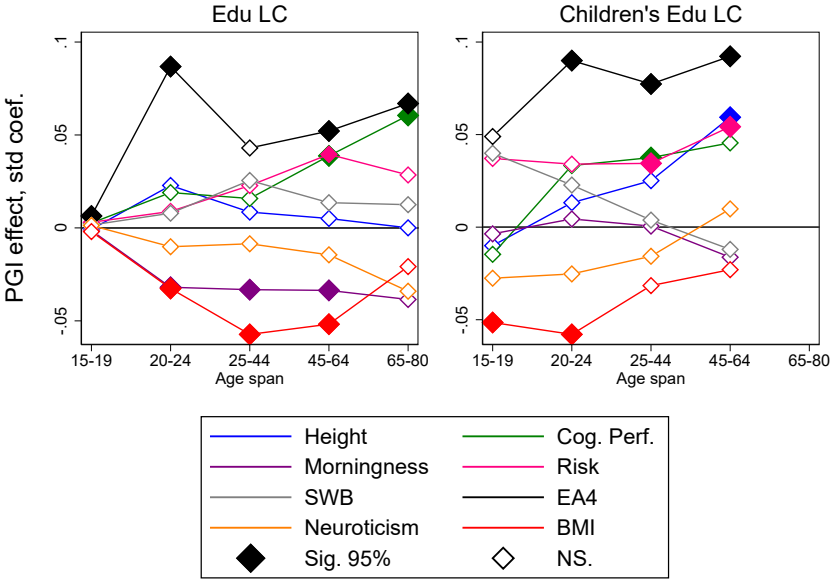


- Causal effect of G on Y: $\beta_g + \gamma\beta_e$
- Estimated bivariate correlation between G and Y: $\beta_g + (\alpha_g\alpha_e + \gamma)\beta_e$
- Simply conditioning on E (i.e. controlling for environment) is often not a good idea: removes a part of the causal pathway ($\gamma\beta_e$)
- Instead, conditioning on parents does solve this: removes passive *rGE*, but not active or evocative.

Educational local contexts in Sweden

- Plenty of evidence on rGE (see e.g. Abdel's work) – here is evidence from Sweden.
- Both an EA PGI and actual education correlates with educational environment.
- Increasing over time.
- Clear evidence of *rGE*!




Edu LC
Children's Edu LC
PGI effect, std coef.
.1
.05
0
-.05
15-19
20-24
25-44
45-64
65-80
Age span
Height
Morningness
SWB
Neuroticism
Sig. 95%
Cog. Perf.
Risk
EA4
BMI
NS.

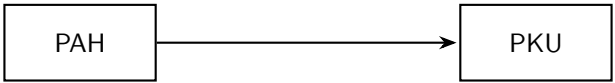
Gene-environment interaction

- $G \times E$: when the effects of genes and environments are not independent.
- More concretely: when the effect of a genetic factor depends on what the environmental conditions are, or when the effect of an environmental factor depends on the genotype of the individual that the environment is acting on.
- This section will outline some concepts and typologies of *GxE* effects, and explore practical tools, methods, and common pitfalls when investigating these types of effects.
- The reason for focusing so much on more "helicopter perspective" theoretical concepts and typologies here is that I agree with Hans' assessment before lunch that this field is highly undertheorized and needs a lot more careful thought about mechanisms and higher-level causal models.

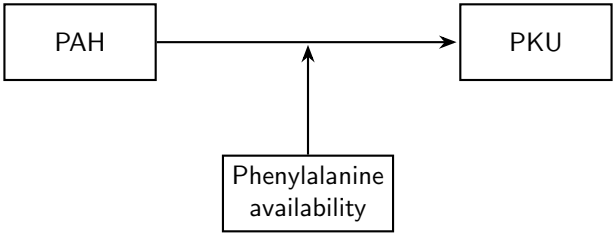
Why are $G \times E$ effects important?

- In a trivial sense, foundational for all of biology: a gene cannot be expressed if there are no amino acids available to build proteins with. All gene expression is dependent on the cellular and extra-cellular environment.

Basic example: Phenylketonuria (PKU)



Basic example: Phenylketonuria (PKU)



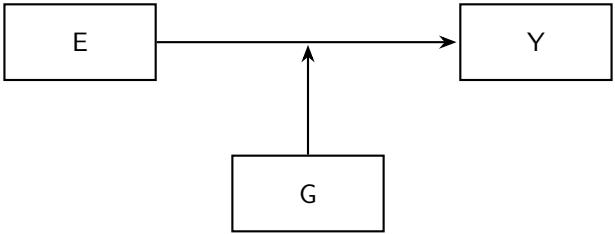
Why are $G \times E$ effects important?

- In a trivial sense, foundational for all of biology: a gene cannot be expressed if there are no amino acids available to build proteins with. All gene expression is dependent on the cellular and extra-cellular environment.
- We (social science genetics!) are typically interested in less "trivial" GxE effects, for example with environmental factors that happen outside of an individual's physical body:
  - Are depression genes expressed more or less depending on traumatic life experiences?
  - Are genetic effects on cognitive performance stronger or weaker if you grew up rich vs. poor?
- Complex (i.e. polygenic) traits should not be expected to have GxE mechanisms of theoretical interest as close to the biology as e.g. PKU, but rather at a somewhat higher level of "aggregation"
- The moderating effect of an environment will usually happen further downstream the causal chain, e.g. between some endophenotype and the outcome

Typologies of $G \times E$

- Genes as moderators of environmental effects
- Environments as moderators of genetic effects
- Relative vs. absolute $G \times E$ effects
- Moderation before/after endophenotype
- (others – e.g. Shanahan & Hofer covered yesterday)

Genes as moderators

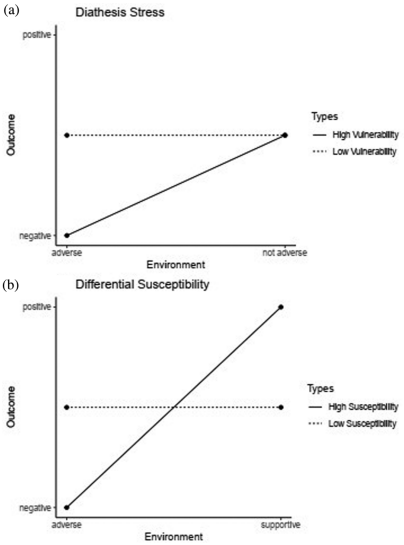


Genes as moderators

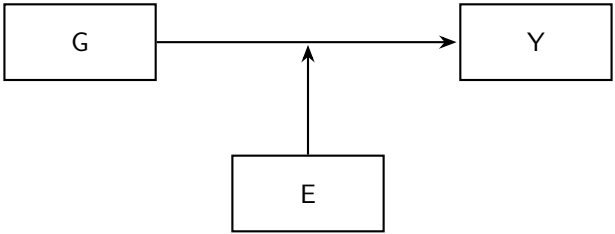
Environmental effects, with some genetic factor as a source of latent heterogeneity:

- Diathesis-stress: an environmental stressor affects only (or affects more) those with a particular genetic sensitivity.
- Differential susceptibility: individuals vary in how susceptible they are to environment influences in general, both negative and positive.

(see e.g. Zhang and Belsky 2020)



Environments as moderators



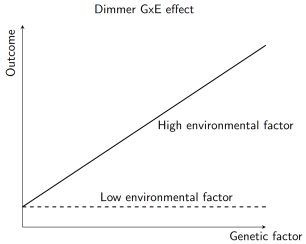
Environments as moderators

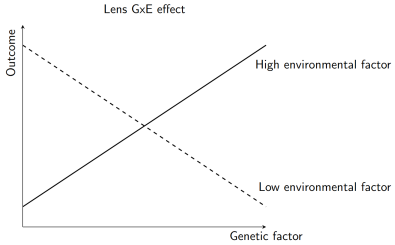
Same functional (but different theoretical) distinction here:

- Dimmer effects: the size of a genetic effect is turned up or down by environmental influences.
- Lens effects: the *sign* of a genetic effect is *flipped* by environmental influences.

(see e.g. Domingue et al. 2020).

For applications using PGI's, this distinction can be crucial, as we shall see later.





Lens...

Relative vs. absolute effects

- Heritability is a relative estimate (0-100%).
- ...means heritability can increase (decrease) if the variability in the environment decreases (increases), without the absolute effect of genetics being affected at all.
- Therefore important to conceptually separate GxE as different heritabilities (i.e. Scarr-Rowe hypothesis), vs. GxE as different absolute effects.
- Going forward, I will focus on the latter.

Endophenotypes?

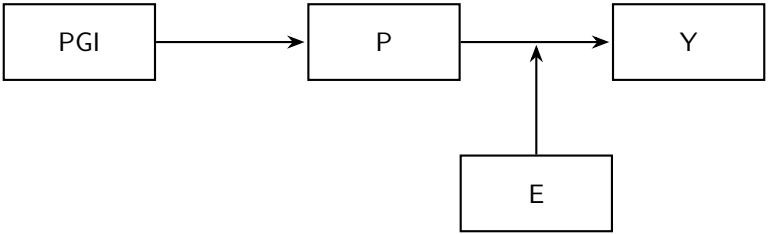
An important consideration – especially with environments as moderators – is often: where in the causal chain does the moderation of the effect happen?

This is important because:

- GWAS is constructed to find main, linear effects.
- Identified causal SNPs are going to be ones *least* likely to be environmentally moderated.
- In particular, ALL lens GxE effects are going to be washed out.

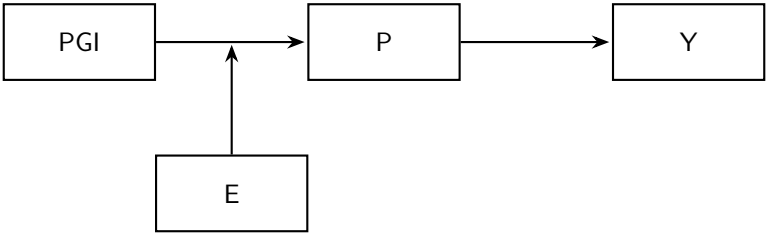
Therefore – helpful to think about PGI target traits as *endophenotypes*

Moderation after endophenotype



This assumes a nice, linear relationship between your PGI and your endophenotype – all good!

Moderation before endophenotype



This may be more problematic, in particular for lens-type GxE effects.

Methodological options

- Traditional variance decomposition models (MZ/DZ twin models) modeling varying heritability (i.e. relative GxE effects, as mentioned before)
- Candidate GxE studies. Not a good idea, as we've seen.

REPORTS

Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene

AVSHALOM CASPI, KAREN SUGDEN, TERRIE E. MOFFITT, ALAN TAYLOR, IAN W. CRAIG, HONALEE HARRINGTON, JOSEPH MCCLAY, JONATHAN MILL, JUDY MARTIN, [...], AND RICHIE POULTON +1 authors Authors Info & Affiliations

SCIENCE · 18 Jul 2003 · Vol 301, Issue 5631 · pp. 386-389 · DOI: 10.1126/science.1083968

Abstract

In a prospective-longitudinal study of a representative birth cohort, we tested why stressful experiences lead to depression in some people but not in others. A functional polymorphism in the promoter region of the serotonin transporter (5-HT T) gene was found to moderate the influence of stressful life events on depression. Individuals with one or two copies of the short allele of the 5-HT T promoter polymorphism exhibited more depressive symptoms, diagnosable depression, and suicidality in relation to stressful life events than individuals homozygous for the long allele. This epidemiological study thus provides evidence of a gene-by-environment interaction, in which an individual's response to environmental insults is moderated by his or her genetic makeup.

nature › molecular psychiatry › original article › article

Published: 04 April 2017

Collaborative meta-analysis finds no evidence of a strong interaction between stress and 5-HTTLPR genotype contributing to the development of depression

R C Culverhouse, N L Saccone, A C Horton, Y Ma, K J Anstey, T Banaschewski, M Burmeister, S Cohen-Woods, B Etain, H L Fisher, N Goldman, S Guillaume, J Horwood, G Juhasz, K J Lester, L Mandelli, C M Middeldorp, E Olié, S Villafuerte, T M Air, R Araya, L Bowes, R Burns, E M Byrne, ... L J Bierut

+ Show authors

Molecular Psychiatry **23**, 133–142 (2018) | Cite this article

Methodological options

- Traditional variance decomposition models (MZ/DZ twin models) modeling varying heritability (i.e. relative GxE effects, as mentioned before).
- Candidate GxE studies. Not a good idea, as we've seen.
- GWASxE studies – add a pre-hypothesized environmental interaction at the discovery stage.
- Other novel methods, like PIGEON (Miao et al. 2022).
- ...or – just use PGIs! Wellpowered even in smaller samples, large range of existing phenotypes and cohorts (thanks Aysu!)

Using PGIs to identify GxE effects

Estimating GxE is no different from attempting to estimate any other interaction effect – meaning that many of the following guidelines and suggestions are not different from what you're used to if you've worked with multiplicative interaction models before. To cleanly identify GxE effects, we want:

- Some source of exogenous in variation in E.
- ...ditto for G – meaning, in practice, within-family differences.

These things don't often co-occur, and when they do, the sample sizes where they do often become very small (with a few rare exceptions – i.e. with sex as environmental factor). Therefore we will often have to think of less ideal options.

No exogenous variation – control based causality

The basic model is:

$$y_i = a + b_g G_i + b_e E_i + b_{ge} G_i E_i + e_i \tag{1}$$

We add controls:

$$y_i = a + b_g G_i + b_e E_i + b_{ge} G_i E_i + \mathbf{b_x X_i} + \mathbf{b_{xe} X_i E_i} + \mathbf{b_{xg} X_i G_i} + e_i \tag{2}$$

These last parts (add interactions between the controls and both G and E) are crucial and are often forgotten (Keller 2014)!

No exogenous variation – control based causality

What should we control for?

- By convention: the famed principal components. But important to remember that these probably won't capture all population stratification, especially for social science outcomes.
- Sex – because it decreases noise and increases power.
- Age at measurement and/or birth year – same reason!
- Perhaps other characteristics that are causally prior to both G and E but that could be associated with both.
- Ideally factors that confound E, i.e. affects both E and Y

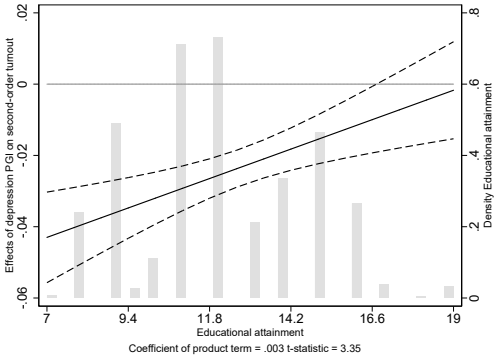
What should we *not* control for?

- Things that are causally somewhere between G and Y – this will absorb possible mechanisms and bias estimates toward zero. This usually conflicts with the above. Try both!
- Other bad controls (e.g. colliders)

Still, important to remember that results will be correlational.

No exogenous variation – control based causality

- Voter turnout in Sweden
- PGI of depression
- Interacted with educational attainment
- Controls for first 10 PCs and sex



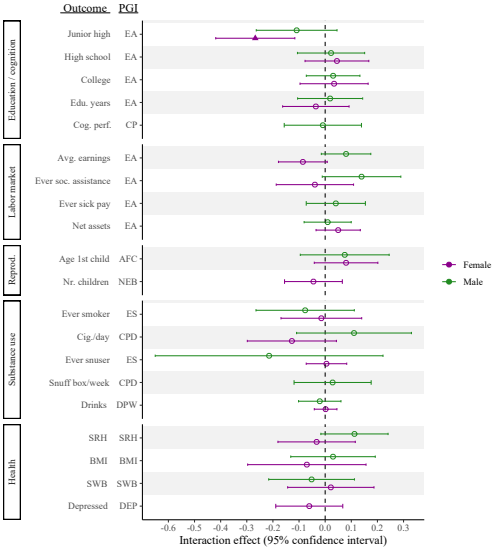
Exogenous variation in E

If we have some plausibly random variation in the environmental exposure, we can use this to estimate GxE effects where we at least can identify the causal effect of the environment at different levels of the PGI.

If so, we solve at least one major problem above: how to control for factors influencing both E and Y without controlling for factors that are in the causal path between G and Y.

Effects of an education reform in Sweden

- Pseudo-randomly implemented reform increasing the mandatory schooling from 7 to 9 years.
- Reform effect on various life outcomes interacted with different PGIs.
- Not much going on – but some!



Exogenous variation in E

We learned about other versions of this before lunch: RD designs, instruments, RCTs etc.

Another way of getting variation in E that at least is not genetically confounded is to use differences between identical twins, so called *discordant twin models*. Differences between identical twins in E are then interacted with between-family differences in G to predict within-family differences in Y.(The use of discordant twin models has a bunch of other problems, as Hans mentioned, and needs a lecture of its own)

Important to remember that all the standard caveats about between-family results for PGIs still apply here: remaining population stratification, genetic nurture etc.

In other words, our identified $G_1 \times E_1$ may still (at least partially) be an $E_1 \times E_2$ interaction effect, or an $G_2 \times E_1$ effect.

Exogenous variation in G

For example, use sibling fixed effects, do sibling differences, or control for parental PGIs. Here we'll do the most straight forward version: sibling differences in Y and G. Depending on what environmental variable we are interested in, can do sibling differences also for E.

$$\Delta Y_j = Y_{j1} - Y_{j2}, \qquad \Delta G_j = G_{j1} - G_{j2},$$

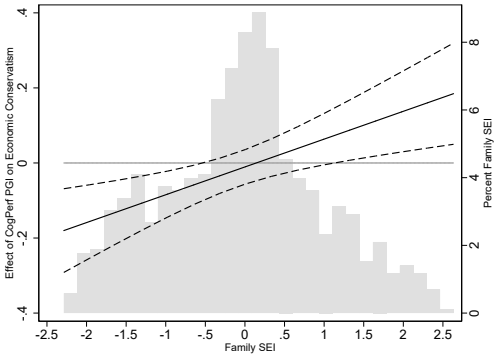
$$\Delta y_j = a + b_g \Delta G_j + b_e E_{ji} + b_{g\times e} \Delta G_j E_{ji} + \mathbf{b_x X_{ji}} + \mathbf{b_{xe} X_{ji} E_{ji}} + \mathbf{b_{xg} X_{ji} \Delta G_j} + e_{ji} \tag{3}$$

or

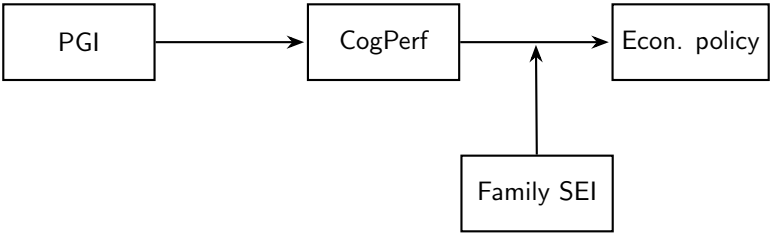
$$\Delta y_j = a + b_g \Delta G_j + b_e \Delta E_j + b_{g\times e} \Delta G_j \Delta E_j + \mathbf{b_x \Delta X_j} + \mathbf{b_{xe} \Delta X_j \Delta E_j} + \mathbf{b_{xg} \Delta X_j \Delta G_j} + e_{ji} \tag{4}$$

The coolest result of my entire research career

- Within-family differences in economic ideology and a cognitive performance PGI
- Between-family differences in parental SEI
- Zero average effects of the PGI on economic ideology
- Clear lens-GxE effects – important implications for gene discovery for political phenotypes



Moderation after endophenotype



Exogenous variation in G

Strength is credible identification of the genetic effect, but the moderating effect of the environmental variable relies on robust controls.

Identified $G_1 \times E_1$ could instead still be (at least partially) a $G_1 \times G_2$ interaction effect, or $G_1 \times E_2$.

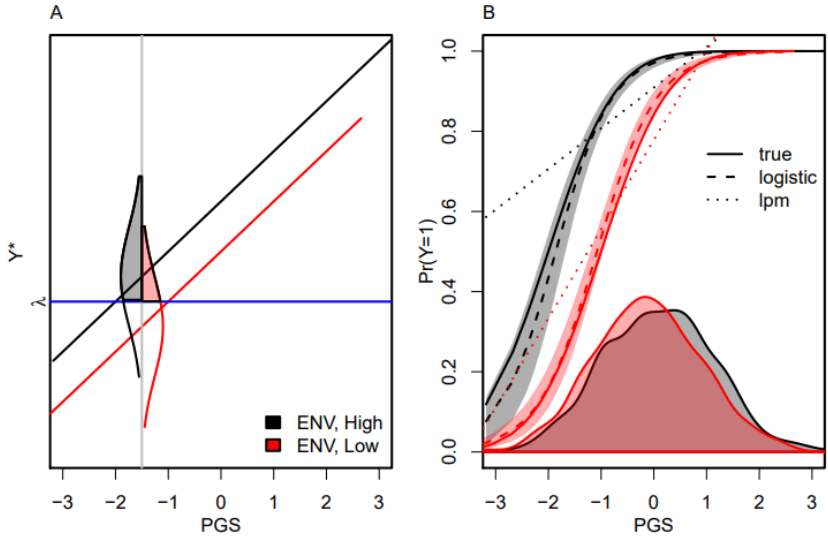
Exogenous variation in both!

There are situations where we can find exogenous variation in both G and E! Use within-family variation in G (i.e. sib differences, family fixed effects, or control for parental genotype), and interact with:

- Conduct an actual RCT in a genotyped sample which contains siblings
- RD for differences in birth month either between families (for DZ twins) or within families (for siblings) – could address voting age, school start, drinking age etc (refer to Hans' slides!)
- Etc...

Other issues to consider

- Beware dichotomized outcomes.
- Power for interactions is tricky – almost always lower than you think.
- Think carefully about non-linearities!


A
Y*
λ
ENV, High
ENV, Low
−3 −2 −1 0 1 2 3
PGS
B
Pr(Y=1)
0.0 0.2 0.4 0.6 0.8 1.0
true
logistic
lpm
−3 −2 −1 0 1 2 3
PGS

Other issues to consider

- Beware dichotomized outcomes.
- Power for interactions is tricky – almost always lower than you think.
- Think carefully about non-linearities!

Questions

(I don't have a picture of a cat, so here is a tiny picture of a chipmunk from google images)