Gene-Environment Interaction in the Era of Precision Medicine

Jingjing Li, Xiao Li, Sai Zhang, and Michael Snyder
Cell, Volume 177, Issue 1, P38-44 (March 21, 2019)

Presentation by Mykhaylo M. Malakhov

The Current Paradigm

$$P = G + E$$

Assumptions:

- genetic (G) and environmental (E) factors independently contribute toward a given phenotypic trait (P)
- G is the aggregated contribution from multiple independent individual loci
- E is any variance in P not explained by G

Definition: (narrow-sense) heritability is the fraction of phenotypic variance explained by genetic variance.

Problem: Missing Heritability

For many diseases with a strong genetic underpinning, the classical association framework usually cannot identify significant signals.

Claim

Missing heritability results from model insufficiency. Linear models cannot capture the intrinsic complexity of complex diseases involving hundreds or thousands of at-risk loci.

Problem: Individual Diversity

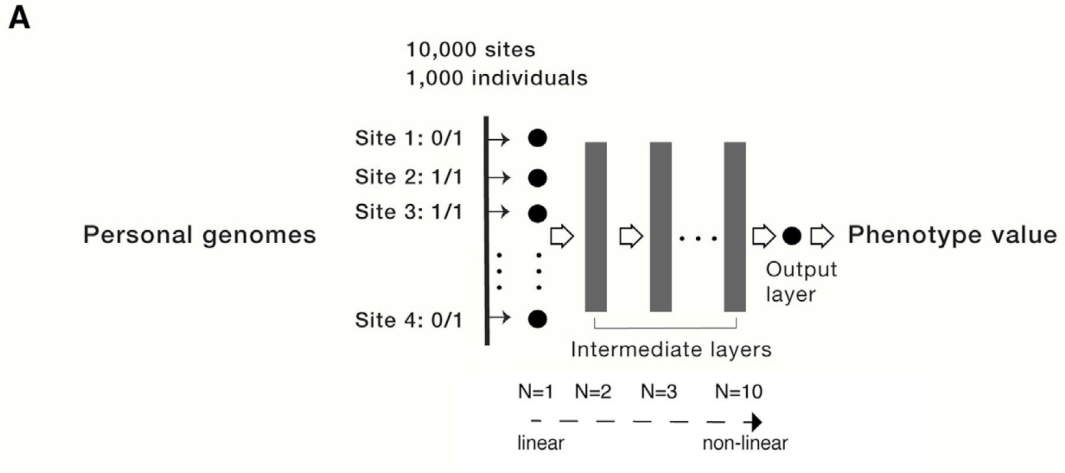
Epidemiological studies estimate risk at a population level, which may not be applicable to personalized health management since individuals have vast differences in lifestyles, behaviors, physiologies, exposomes, and genetic predispositions.

Claim

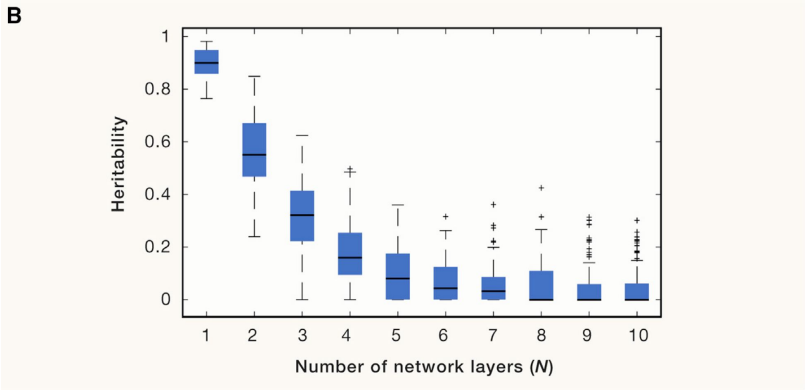
Technological advancement is needed to acquire and assemble non-genetic data at a personal resolution.

Missing Heritability

A Simulation Study



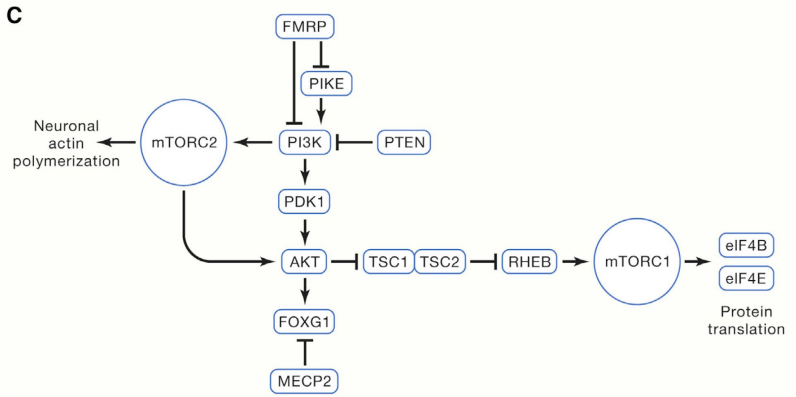
Simulation Results



Note: heritability estimated using GCTA (genome-wide complex trait analysis) package.

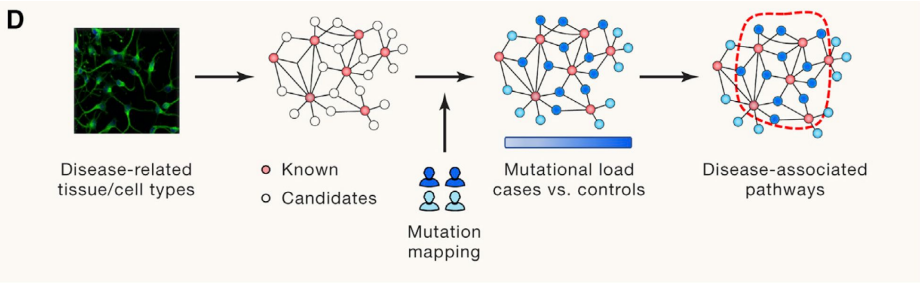
Example: PI3K/AKT/mTOR Pathway

Duplicate genes and epistatic interactions between genes yield nonlinear genotype-phenotype mappings.



Potential Solutions

1. Re-construct a reference cellular map using multiomic approaches
2. Select the minimal subset of genes whose mutation patterns distinguish cases from controls using sparse learning



Individual Diversity

Measuring Environmental Factors

Limitations of epidemiological studies:

- Population-level estimates might not apply to a specific individual
- Data from participant surveys or clinical records is sparse and collected at (relatively few) discrete time points

Benefits of wearable technology:

- Widely available (e.g. smart watches and fitness trackers)
- Continuously track individual-level physiological and environmental parameters in real time
- More accurate than surveys, e.g. galvanic skin response devices instead of survey questions on stress levels

Bayesian View of Heritability

BADGE (Bayesian Aggregation of Disease Genomics and Environment)

Assumption: disease outcome (D), genetic risk (G), and environment contributors (E) can be modeled by a joint probability $Pr(D, G, E)$.

Then (population level) disease prevalence is

$$Pr(D) = \int Pr(D \mid G, E)\, Pr(G, E)\, dGdE \tag{1}$$

(Population level) environmental contribution to disease outcome is

$$Pr(D \mid E) = \int Pr(D \mid G, E)\, Pr(G)\, dG \tag{2}$$

BADGE (Continued)

(Personal) genomic contribution to disease outcome is

$$Pr\left(D \mid G\right) = \int Pr\left(D \mid G, E\right) P\left(E\right) dE \tag{3}$$

Personal genome given personal clinical outcome is

$$Pr\left(G \mid D\right) = \frac{Pr\left(D \mid G\right) Pr\left(G\right)}{\int Pr\left(D \mid G\right) Pr\left(G\right) dG} \tag{4}$$

Genetic Coefficient

Definition: genetic coefficient of a disease:

$$C = JS\left(Pr\left(G \mid D = 1\right) \parallel Pr\left(G \mid D = 0\right)\right) \tag{5}$$

where JS is the Jensen-Shannon divergence between two distributions.

Intuitively, C represents the distinguishability of case genomes ($D = 1$) from control genomes ($D = 0$).