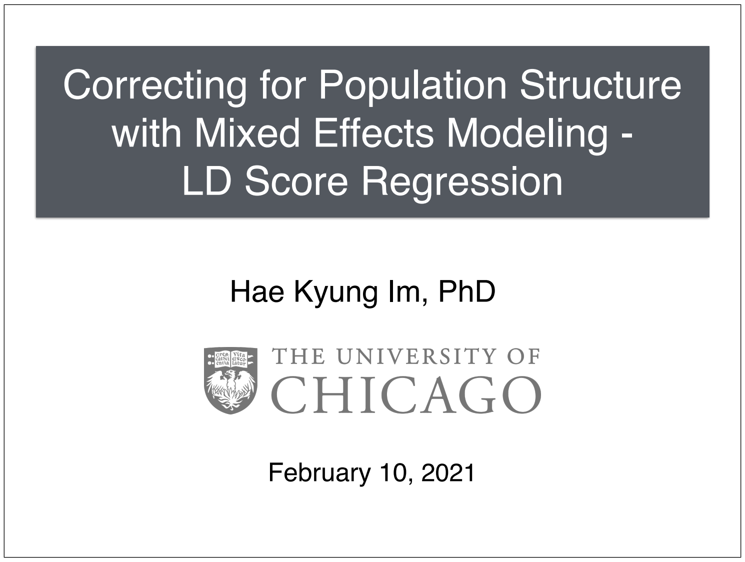


Correcting for Population Structure with Mixed Effects Modeling - LD Score Regression

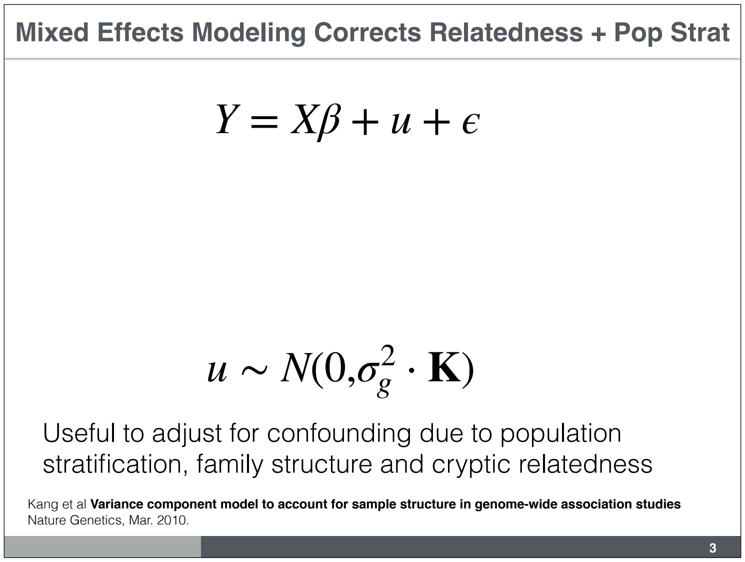
Hae Kyung Im, PhD

THE UNIVERSITY OF CHICAGO

February 10, 2021

Mixed Effects Modeling

In our previous lecture on population structure, we saw that genomic control and genetic principal components are useful tools to correct for population structure. Today we will see two additional approaches. One is based on mixed effects modeling and the other is using LD score regression.



Mixed Effects Modeling Corrects Relatedness + Pop Strat

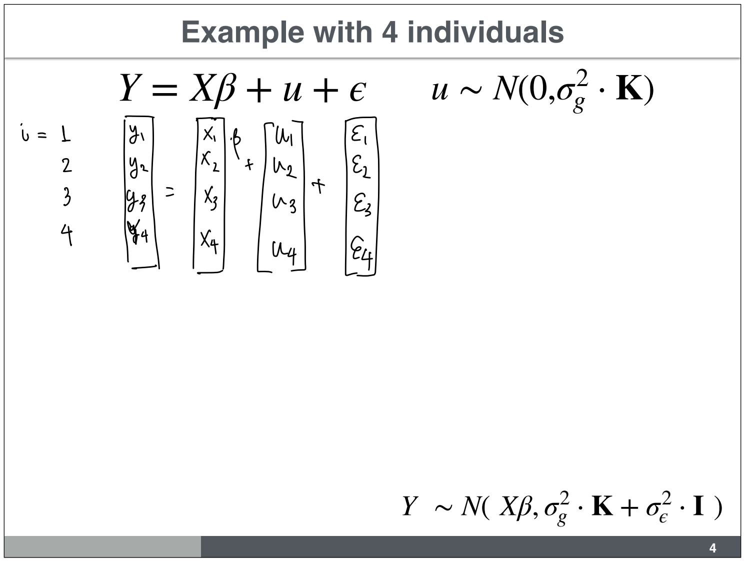
$$Y = X\beta + u + \epsilon$$

$$u \sim N(0, \sigma_g^2 \cdot \mathbf{K})$$

Useful to adjust for confounding due to population stratification, family structure and cryptic relatedness

Kang et al **Variance component model to account for sample structure in genome-wide association studies** Nature Genetics, Mar. 2010.

Mixed effects models are just regression models, for simplicity think linear regression, where in addition to the traditional "fixed" effects of the covariates, there are "random" effects. In this case, β is a fixed effect, whereas u is a random effect. Random effects are specified by their distribution, in this case u is normally distributed with mean 0 and variance σ$_g$·K. When K is the genetic relatedness matrix (genetic correlation between individuals), then this model is able to account for population structure, family structure, and cryptic relatedness.



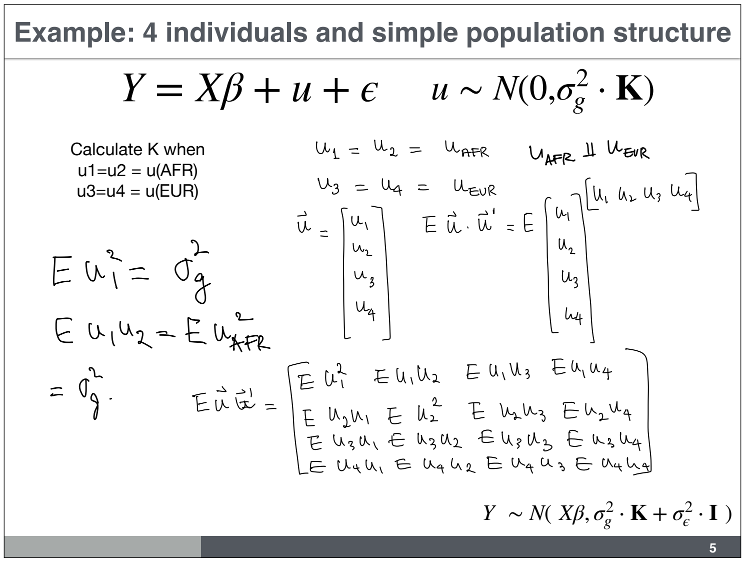
Example with 4 individuals

$$Y = X\beta + u + \epsilon \qquad u \sim N(0,\sigma_g^2 \cdot \mathbf{K})$$

$$\begin{array}{l} i = 1 \\ \;\;\;\;\; 2 \\ \;\;\;\;\; 3 \\ \;\;\;\;\; 4 \end{array} \quad \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix} \cdot \beta + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}$$

$$Y \sim N(\, X\beta, \sigma_g^2 \cdot \mathbf{K} + \sigma_\epsilon^2 \cdot \mathbf{I}\,)$$

Let's look at a simple example with 4 individuals, two from EUR ancestry and two from AFR ancestry. No family structure or cryptic relatedness. Here u represents the effect of population status on the phenotype $\boldsymbol{Y}$. The matrix $\mathbf{K}$ represents the population pattern in the data and σ$_g$ is the scale of the effect, a measure of the effect of the population membership on the phenotype.



Example: 4 individuals and simple population structure

$$Y = X\beta + u + \epsilon \qquad u \sim N(0,\sigma_g^2 \cdot \mathbf{K})$$

Calculate K when
u1=u2 = u(AFR)
u3=u4 = u(EUR)

$$u_1 = u_2 = u_{AFR} \qquad u_{AFR} \perp\!\!\!\perp u_{EUR}$$

$$u_3 = u_4 = u_{EUR}$$

$$\vec{u} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} \qquad E\,\vec{u} \cdot \vec{u}' = E \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} \begin{bmatrix} u_1 & u_2 & u_3 & u_4 \end{bmatrix}$$

$$E\,u_1^2 = \sigma_g^2$$

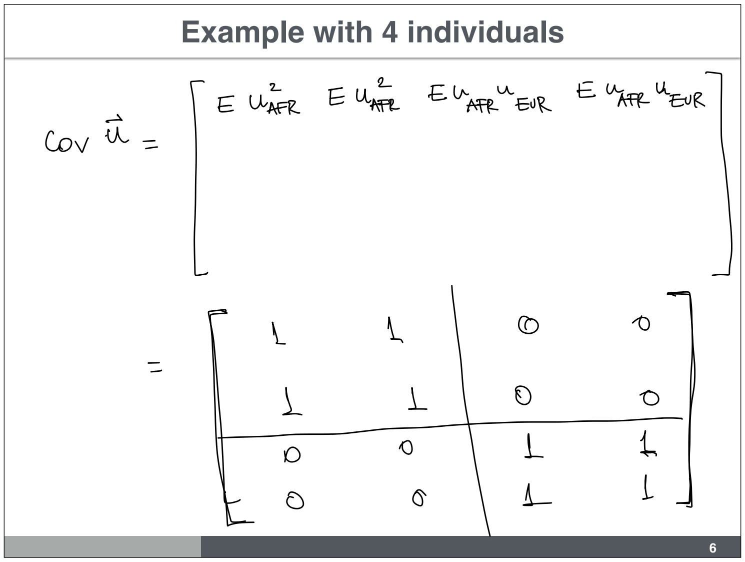
$$E\,u_1 u_2 = E\,u_{AFR}^2$$

$$= \sigma_g^2.$$

$$E\,\vec{u}\,\vec{u}' = \begin{bmatrix} E\,u_1^2 & E\,u_1u_2 & E\,u_1u_3 & E\,u_1u_4 \\ E\,u_2u_1 & E\,u_2^2 & E\,u_2u_3 & E\,u_2u_4 \\ E\,u_3u_1 & E\,u_3u_2 & E\,u_3u_3 & E\,u_3u_4 \\ E\,u_4u_1 & E\,u_4u_2 & E\,u_4u_3 & E\,u_4u_4 \end{bmatrix}$$

$$Y \sim N(\, X\beta, \sigma_g^2 \cdot \mathbf{K} + \sigma_\epsilon^2 \cdot \mathbf{I}\,)$$

Assuming individuals 1 and 2 are of African descent and 3 and 4 are of European descent. $u_1 = u_2 = u_{\text{AFR}}$ and $u_3 = u_4 = u_{\text{EUR}}$. We also assume that the the population effects are independent across populations, i.e. that u_{EUR} is orthogonal to u_{AFR}. Since we are assuming that the mean is 0, the variance of the u's are given by E u^2. With these assumptions, we can calculate the covariance matrix of u, which is by definition = σ$_g$·K. Recall also that you would calculate the covariance matrix of a vector as E u·u'.



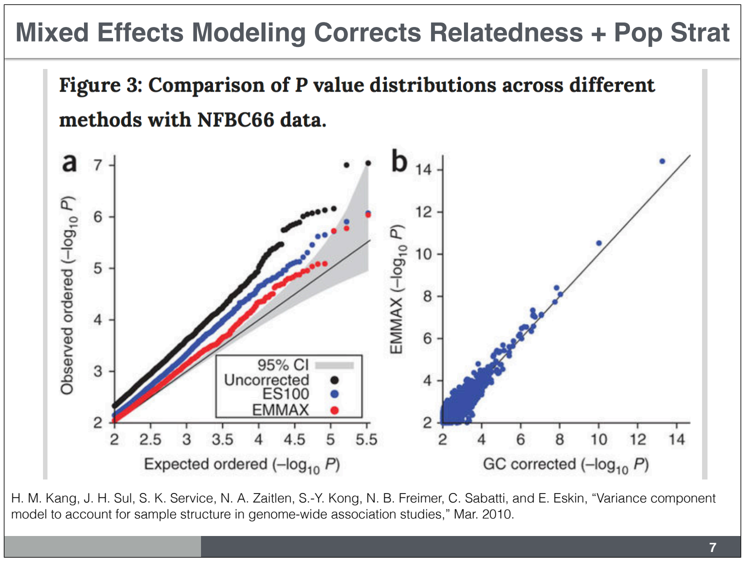
Example with 4 individuals

$$\text{cov } \vec{u} = \begin{bmatrix} E\,u_{AFR}^2 & E\,u_{AFR}^2 & E\,u_{AFR}u_{EUR} & E\,u_{AFR}u_{EUR} \\ & & & \\ & & & \\ & & & \end{bmatrix}$$

$$= \left[\begin{array}{cc|cc} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \hline 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{array}\right]$$

Using all these facts, we can calculate K in this simple two ancestry population. K is formed with block matrices of ones, on the diagonal and 0's off the diagonal, which seems consistent with our intuition of what these should look like.

We seen last week, in these simple structure cases, we can either add a fixed effects that represents the population effect or perform separate analysis of each population and combine the results via meta-analysis. Mixed effect modeling gives us an alternative, which can be extended to more general cases where population may not be clearly separated, with a gradient for example. It also allows more general structure that includes relatedness.



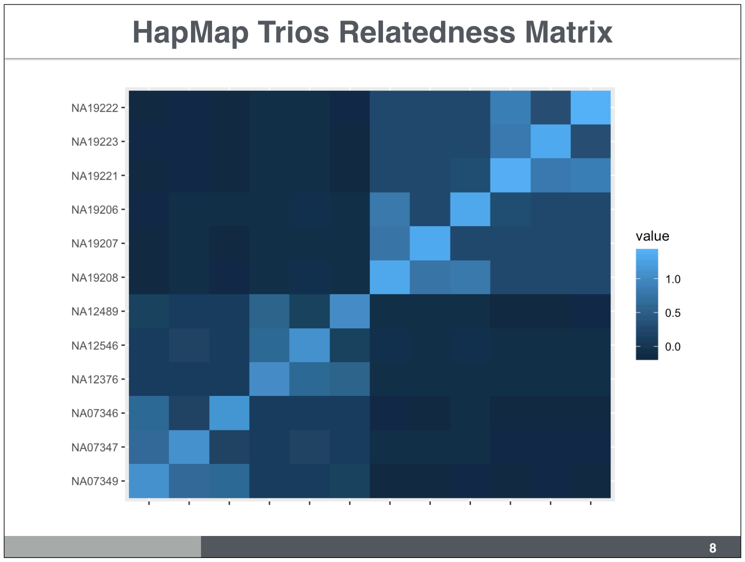
Mixed Effects Modeling Corrects Relatedness + Pop Strat

Figure 3: Comparison of P value distributions across different methods with NFBC66 data.

H. M. Kang, J. H. Sul, S. K. Service, N. A. Zaitlen, S.-Y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin, "Variance component model to account for sample structure in genome-wide association studies," Mar. 2010.

In this figure, Kang et all show compare the uncorrected p-values, which look clearly inflated, with the p-values after correcting for 100 principal components (calculated with eigensoft) and the EMMAX-corrected (mixed effects approach) p-values which look much less inflated. Panel b shows the comparison of EMMAX to the simple genomic control approach (divide the chi2 statistic by the inflation factor λ) showing similar correction of inflation (because the points are located around the identify line).

Notice that the authors use of genomic control as a measure of goodness of fit.



HapMap Trios Relatedness Matrix

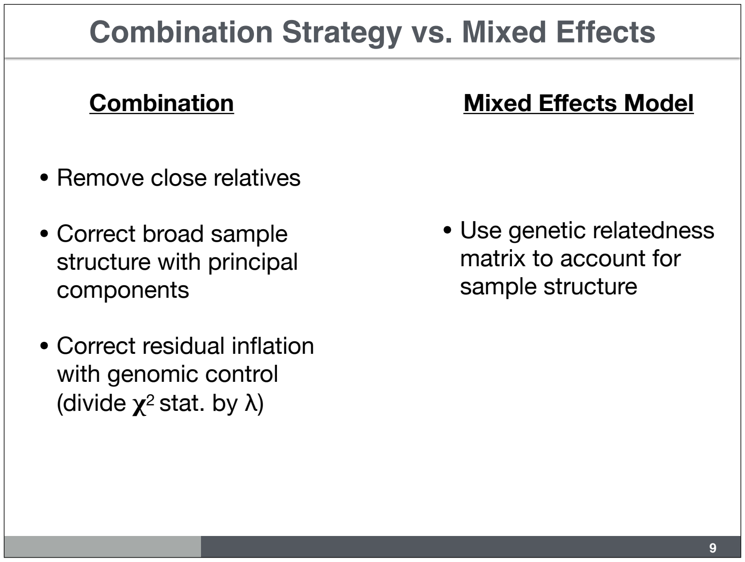
Here is the genetic relatedness matrix of 2 European and 2 African trios (mother, father, and child) from HapMap (YRI, EUR) calculated using plink's make-grm-gz command.

The population structure is apparent in the 6 by 6 distinct blocks, and the family structure (trios) manifests as smaller blocks of 3 by 3.

Random mating seems to be a reasonable assumption. Explain why.

Can you recognize which individual is the child within each trio?

```
## 1345    NA07349 NA07347 NA07346 1     0     CEU
## 1353    NA12376 NA12546 NA12489 2     0     CEU
## Y051    NA19208 NA19207 NA19206 1     0     YRI
## Y058    NA19221 NA19223 NA19222 2     0     YRI
```
https://hakyimlab.github.io/hgen471/L9-GRM.html



Combination Strategy vs. Mixed Effects

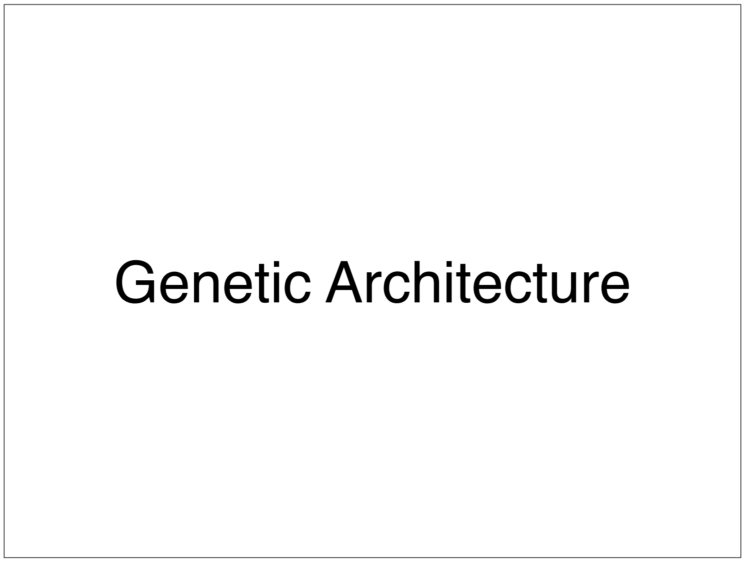
Combination

- Remove close relatives
- Correct broad sample structure with principal components
- Correct residual inflation with genomic control (divide χ^2 stat. by λ)

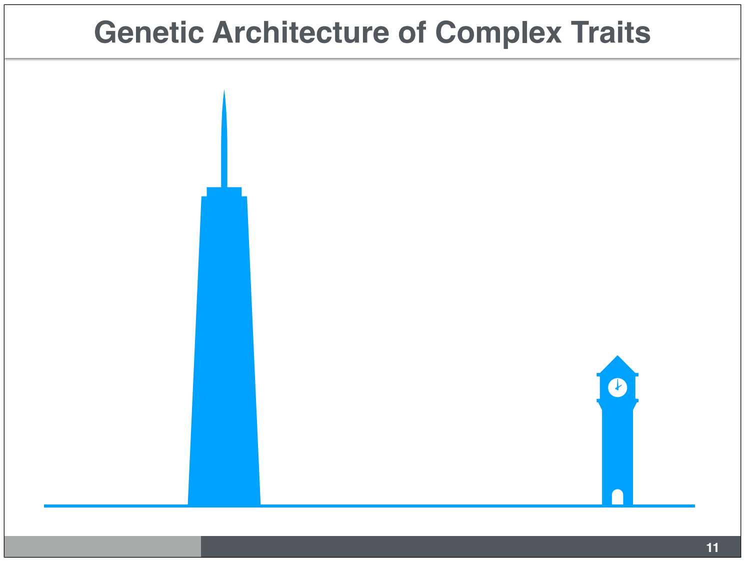
Mixed Effects Model

- Use genetic relatedness matrix to account for sample structure

Mixed effects models are great tools to model data with complex underlying structure but can be computationally expensive and has been shown that sometimes it can over correct causing deflation of association statistics. Still today, the more naive approach of simply removing close relatives is quite common (you can find tons of paper in the UKB where over 100K individuals were excluded to avoid dealing with relatedness).



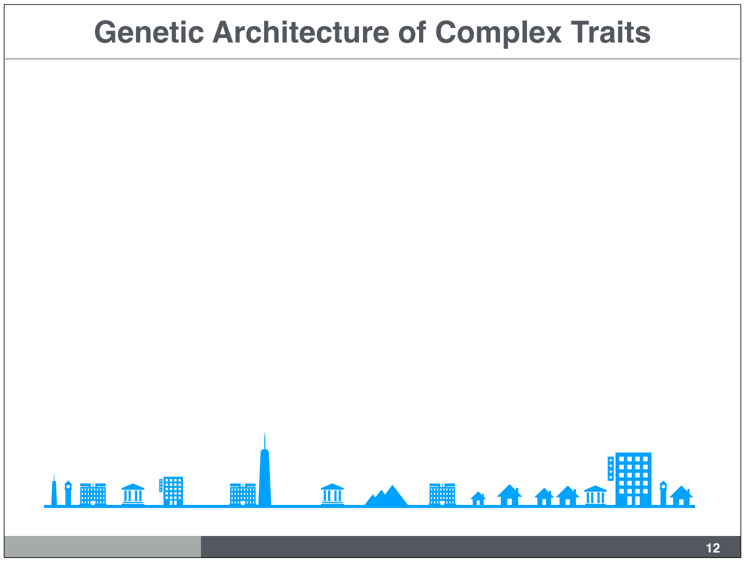
Genetic Architecture



Genetic Architecture of Complex Traits

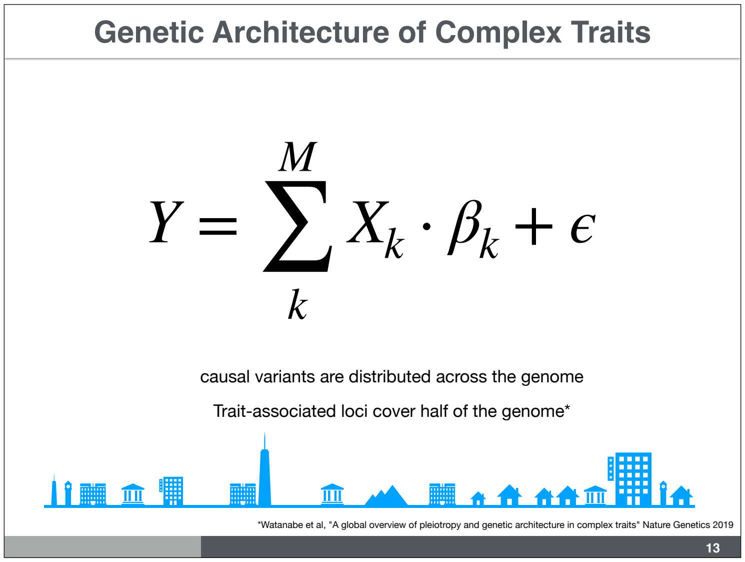
What is the genetic architecture of complex traits? By genetic architecture, we mean the distribution of effect sizes (for example, a few variants of large effects or many variants of small effects) or the dependence of the effect sizes as a function of minor allele frequencies.

When the first GWAS were performed in common disease, it was thought that we would find a few genes which cause the disease (large effect sizes = high penetrance) since the diseases with known genetics at the time were mostly monogenic.



Genetic Architecture of Complex Traits

After many attempts to find large effect variants, the GWAS community came to terms with the fact that the genetic architecture of common diseases and traits is highly polygenic with likely many causal variants sort of uniformly distributed across the genome, each with a modest effect size. The alternative explanation that there are multiple genes with high penetrance (large effect sizes) and that because each person carries a different causal gene, the estimated effects in GWAS ended up being diluted and look like they are very small. This model of diseases is known as the "Anna Karenina" model. Anna Karenina is a novel by Tolstoi which starts with the paragraph "All happy families are alike; each unhappy family is unhappy in its own way." Large family studies (which should be enriched with the same causal gene) have failed to identify effect genes providing overwhelming evidence that most diseases do not follow this "unhappy in its own way" pattern.



Genetic Architecture of Complex Traits

$$Y = \sum_{k}^{M} X_k \cdot \beta_k + \epsilon$$

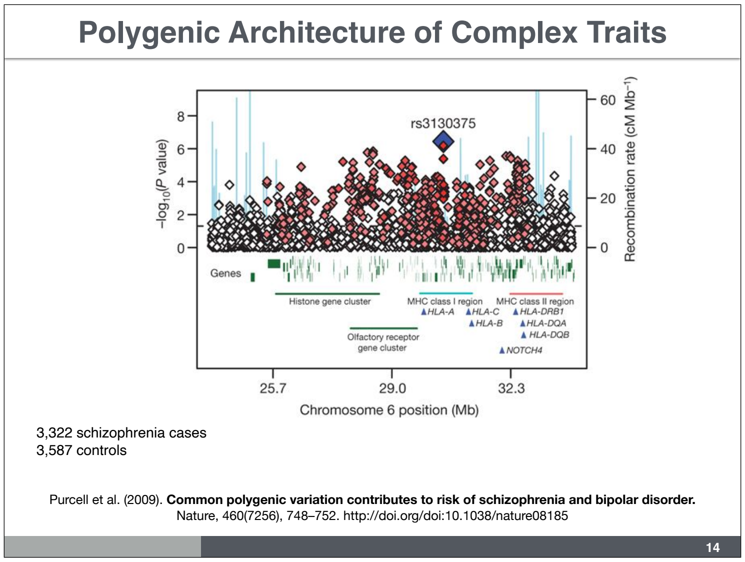
causal variants are distributed across the genome

Trait-associated loci cover half of the genome*

*Watanabe et al, "A global overview of pleiotropy and genetic architecture in complex traits" Nature Genetics 2019

We will be using this polygenic additive model as our default model for complex diseases and traits.

Watanabe et al analyzed 4155 GWAS and found that trait-associated loci cover half of the genome.

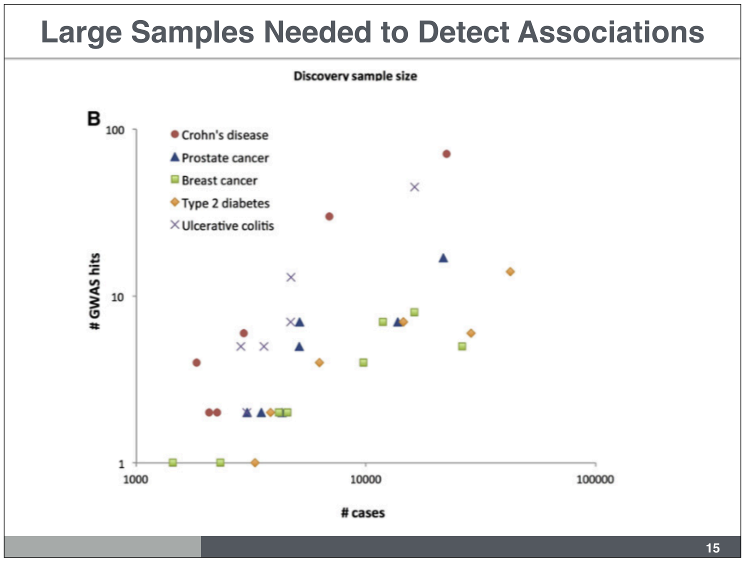


Polygenic Architecture of Complex Traits

3,322 schizophrenia cases
3,587 controls

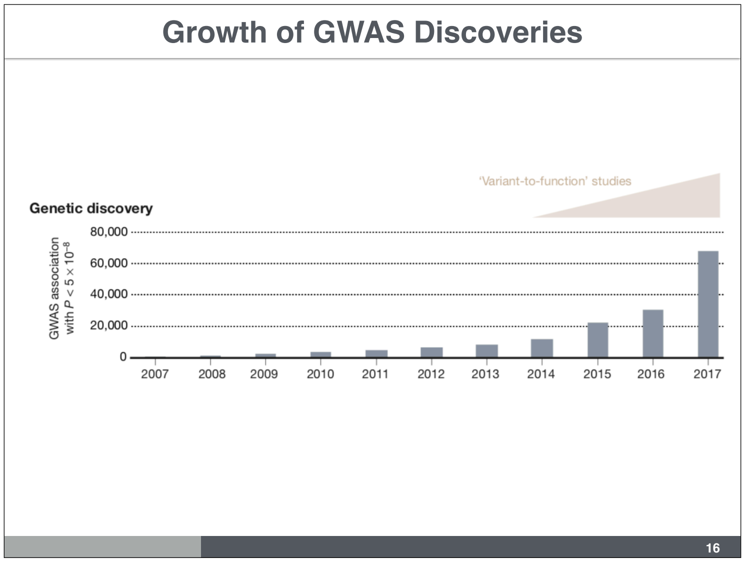
Purcell et al. (2009). **Common polygenic variation contributes to risk of schizophrenia and bipolar disorder.** Nature, 460(7256), 748–752. http://doi.org/doi:10.1038/nature08185

In 2009, the lack of GWAS significance results forced the investigators to look beyond single variant approaches and discovered that aggregating many variants with a loose cut of threshold for p-value could predict schizophrenia disease status. Their conclusion was that common polygenic variation contributes to risk of schizophrenia and bipolar disorder. This was the first major influential publication that used polygenic risk scores to predict complex traits.



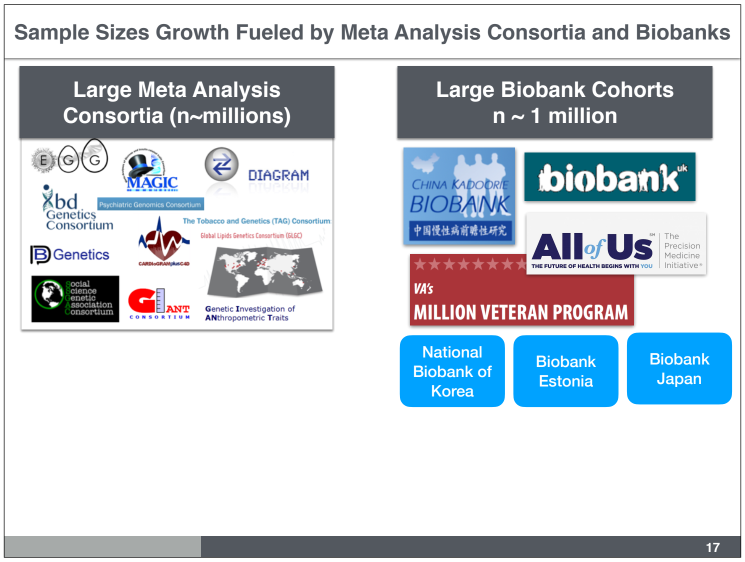
Large Samples Needed to Detect Associations

Number of GWAS loci discovered goes up with sample size. As the example with schizophrenia showed, some minimum sample size is needed to start identifying the associated loci.



Growth of GWAS Discoveries

Over time, the GWAS community has continued adding more individuals and additional phenotypes leading to a rapid increase in the number of discoveries. We still don't understand the mechanism behind most of these variants.



Sample Sizes Growth Fueled by Meta Analysis Consortia and Biobanks

Large Meta Analysis Consortia (n~millions)

Large Biobank Cohorts n ~ 1 million

National Biobank of Korea

Biobank Estonia

Biobank Japan

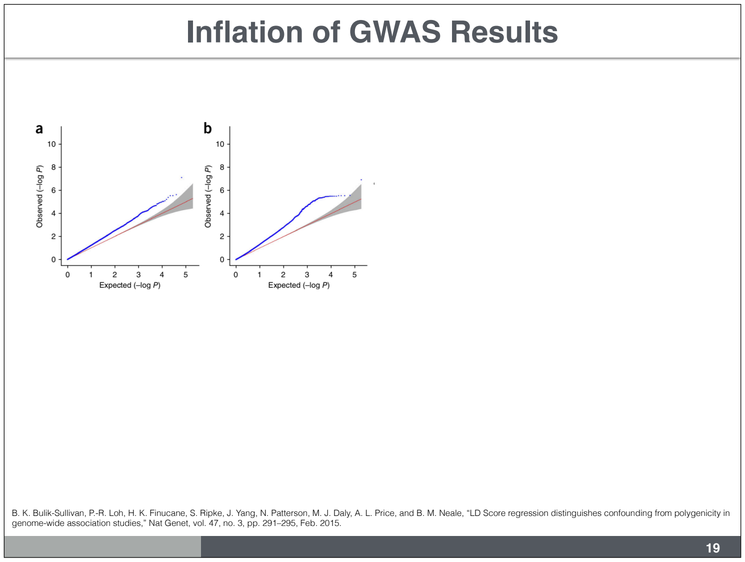
EGG
MAGIC
DIAGRAM
ibd Genetics Consortium
Psychiatric Genomics Consortium
The Tobacco and Genetics (TAG) Consortium
Global Lipids Genetics Consortium (GLGC)
CARDIoGRAMplusC4D
Genetics
Social Science Genetic Association Consortium
GIANT CONSORTIUM
Genetic Investigation of ANthropometric Traits
CHINA KADOORIE BIOBANK
biobank uk
All of Us
THE FUTURE OF HEALTH BEGINS WITH YOU
The Precision Medicine Initiative
VA's MILLION VETERAN PROGRAM

Meta analysis consortia have been very successful because different investigators had to share only the summary results. Individual level data sharing is much more difficult because of regulatory, consent issues. Also computational cost of handling massive sample sizes limits the size. Methods that integrate summary level data are in high demand.

Biobanks are other ways in which the field has increased sample sizes. But this has created the need for statistical methods that can handle the massive datasets.

Both data types have created the need to develop novel methods. We will see some of these methods in the coming lectures.

LD Score Regression



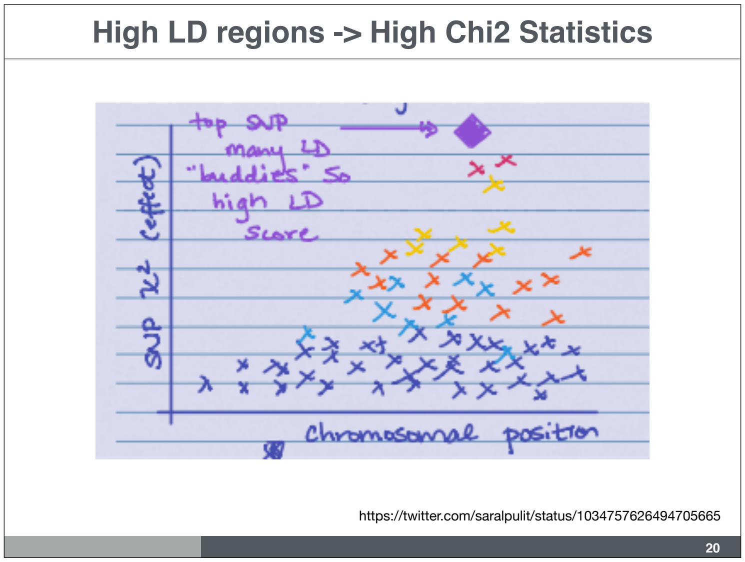
Inflation of GWAS Results

B. K. Bulik-Sullivan, P.-R. Loh, H. K. Finucane, S. Ripke, J. Yang, N. Patterson, M. J. Daly, A. L. Price, and B. M. Neale, “LD Score regression distinguishes confounding from polygenicity in genome-wide association studies,” Nat Genet, vol. 47, no. 3, pp. 291–295, Feb. 2015.

Inflation of summary statistics (-log10 p here) can be due to two effects:

- confounders such as population structure, relatedness, etc
- true polygenicity (most variants have a causal effect or are in LD with causal variants)

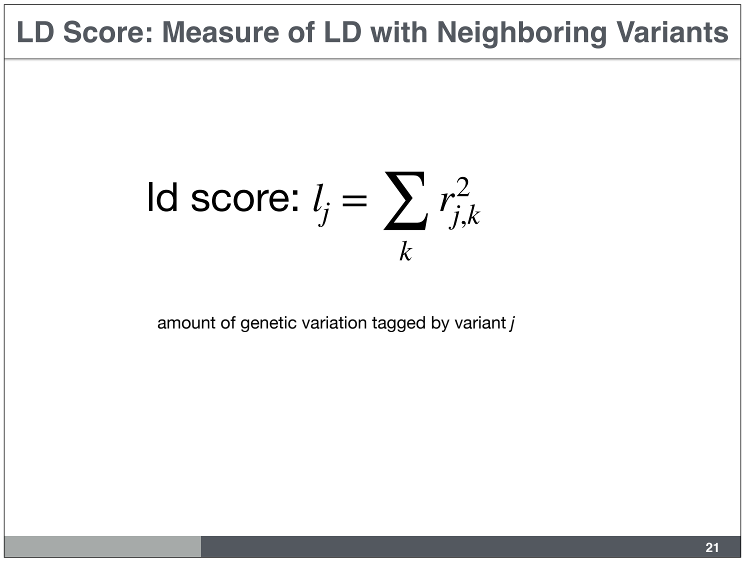
Genomic control method does not distinguish between true polygenicity and inflation due to population stratification.



High LD regions -> High Chi2 Statistics

https://twitter.com/saralpulit/status/1034757626494705665

This represents a locus zoom-like plot with Chi2 statistics instead of p-values. The basic principle of LD score regression relies on the fact that if we assume a polygenic model, then variants that are in high LD regions, will have higher association statistic (chi2 here). More informally, SNPs with many LD-friends will be lifted up in the chi2 chart because they are more likely to tag causal variants that SNPs without LD-friend.

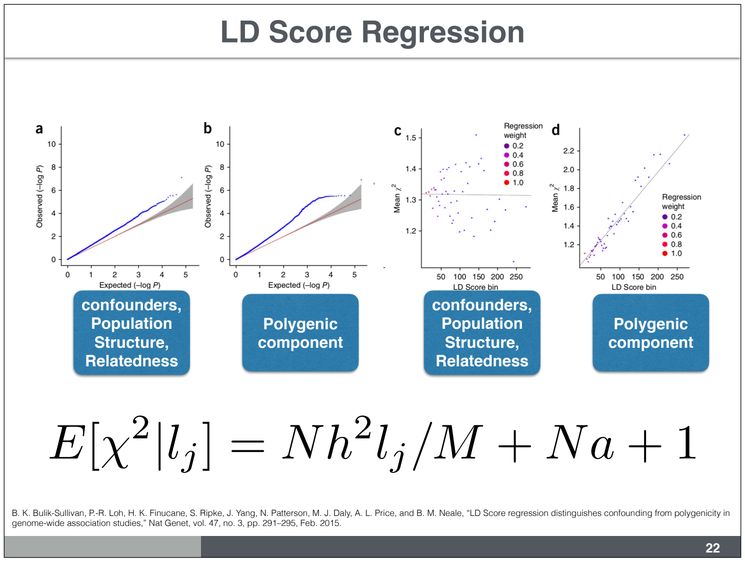


LD Score: Measure of LD with Neighboring Variants

$$\text{ld score: } l_j = \sum_k r^2_{j,k}$$

amount of genetic variation tagged by variant *j*

ld-score is a measure the number of "LD-friends", and it's calculated as the sum of LD. Each genetic variant will have one such score.



LD Score Regression

confounders, Population Structure, Relatedness

Polygenic component

confounders, Population Structure, Relatedness

Polygenic component

$$E[\chi^2|l_j] = Nh^2 l_j/M + Na + 1$$

B. K. Bulik-Sullivan, P.-R. Loh, H. K. Finucane, S. Ripke, J. Yang, N. Patterson, M. J. Daly, A. L. Price, and B. M. Neale, "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies," Nat Genet, vol. 47, no. 3, pp. 291–295, Feb. 2015.

Bulik-Sullivan et al show that we can distinguish inflation due to confounding (a) and polygenicity (b) by regressing the chi2 statistic against ld-score. The intercept should be 1 if there were no inflation. Any number above 1 can be interpreted as inflation due to population or relatedness confounding. The slope of the regression allows us to calculate the heritability.

N is the sample size, M is the number of variants, h2 is the heritability explained by the M variants in aggregate, and a represents a measure of confounding. Notice that the effects of both confounding and polygenicity increases with the sample size.
Note: regression weights are used to account for heteroskedasticity, i.e. the fact that errors are larger for higher LD score.

LD Score Regression Distinguishes Confounding from Polygenicity

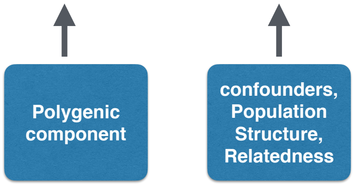
- Variants in LD with a causal variant show inflation in test statistics proportional to their LD with the causal variant.
- The more genetic variation an index variant tags, the higher the probability that this index variant will tag a causal variant
- Inflation from cryptic relatedness or population stratification purely from genetic drift will not correlate with LD
- **Assumptions**: polygenic model, effect sizes for variants drawn independently from distributions with variance proportional to 1/(p(1 – p))

$$E[\chi^2|l_j] = Nh^2 l_j/M + Na + 1$$

LD Score regression distinguishes confounding from polygenicity https://rdcu.be/b07sl

LD Score Regression

$$E[\chi^2|l_j] = Nh^2 l_j/M + Na + 1$$


Polygenic component

confounders, Population Structure, Relatedness

$N =$ sample size

$M =$ number of SNPs

$h^2/M =$ variance explained per SNP

$a =$ confounding biases, cryptic relatedness and population stratification

$l_j = \sum_k r_{jk}^2$ amount of genetic variation tagged by SNP j

LD Score Regression - Schizophrenia

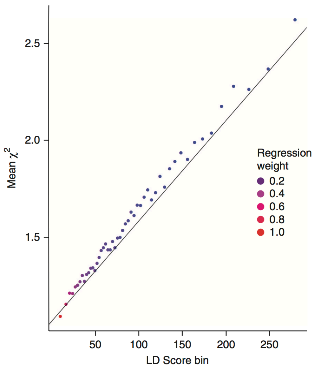


Figure 2 LD Score regression plot for the most recent schizophrenia meta-analysis. Each point represents an LD Score quantile, where the *x*

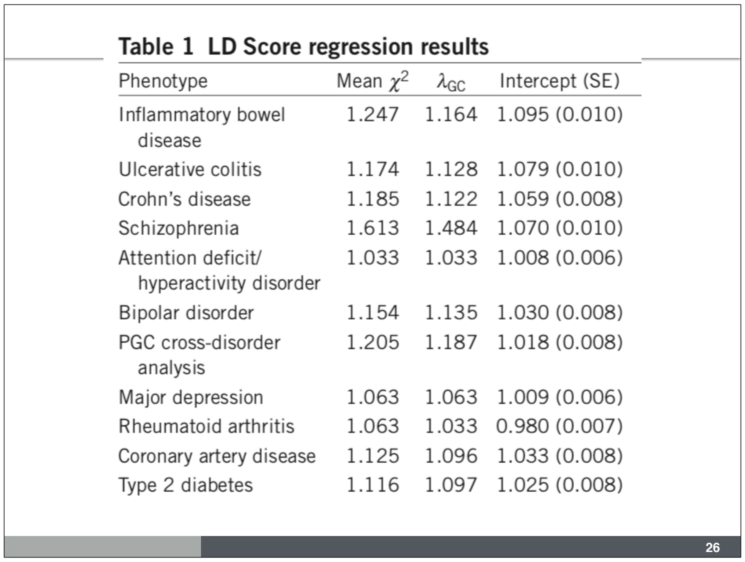
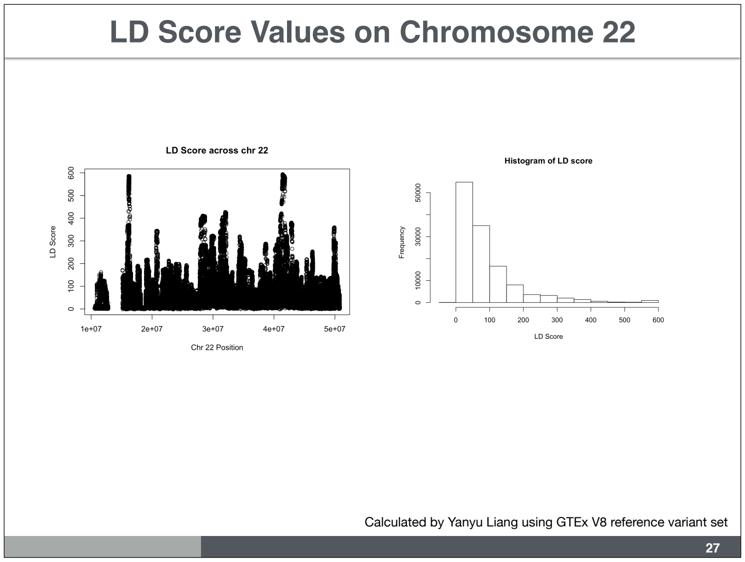


Table 1 LD Score regression results

Phenotype	Mean χ^2	λ_{GC}	Intercept (SE)
Inflammatory bowel disease	1.247	1.164	1.095 (0.010)
Ulcerative colitis	1.174	1.128	1.079 (0.010)
Crohn's disease	1.185	1.122	1.059 (0.008)
Schizophrenia	1.613	1.484	1.070 (0.010)
Attention deficit/ hyperactivity disorder	1.033	1.033	1.008 (0.006)
Bipolar disorder	1.154	1.135	1.030 (0.008)
PGC cross-disorder analysis	1.205	1.187	1.018 (0.008)
Major depression	1.063	1.063	1.009 (0.006)
Rheumatoid arthritis	1.063	1.033	0.980 (0.007)
Coronary artery disease	1.125	1.096	1.033 (0.008)
Type 2 diabetes	1.116	1.097	1.025 (0.008)

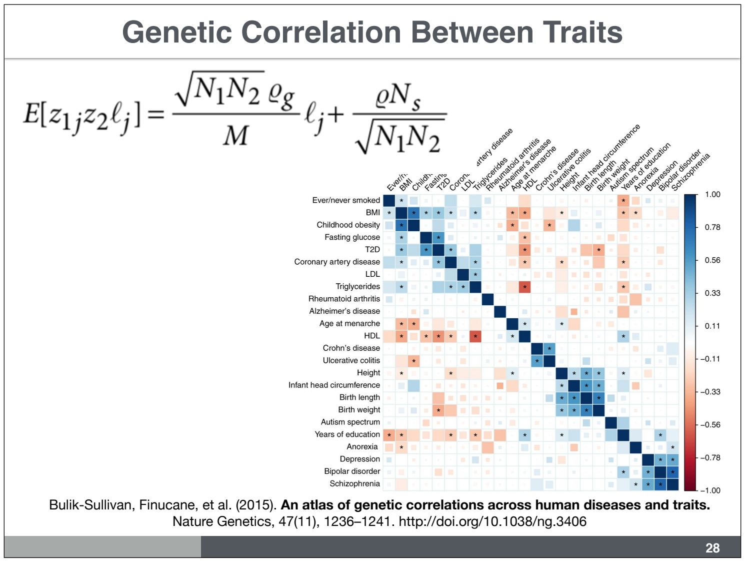
Notice that the intercept, a new measure of confounder driven inflation, is systematically smaller than the one calculated by the genomic control λ_{GC}.



LD Score Values on Chromosome 22

Calculated by Yanyu Liang using GTEx V8 reference variant set

To get a sense of the range of values LD score can take, here is a plot of the LD score values calculated on chromosome 22.



Genetic Correlation Between Traits

$$E[z_{1j}z_{2}\ell_j] = \frac{\sqrt{N_1N_2}\,\varrho_g}{M}\ell_j + \frac{\varrho N_s}{\sqrt{N_1N_2}}$$

Bulik-Sullivan, Finucane, et al. (2015). **An atlas of genetic correlations across human diseases and traits.** Nature Genetics, 47(11), 1236–1241. http://doi.org/10.1038/ng.3406

With the same assumptions as used for the derivation of the LD score, one can calculate the genetic correlation between traits using the same techniques for the ld-score regression. This genetic correlation can provide additional insight into some of the epidemiological/observed correlation between these traits. These avoid some of the confounders in observational studies providing additional insights and orthogonal sources of evidence for the associations.

References

- B. Devlin and Kathryn Roeder (1999) "Genomic Control for Association Studies", Biometrics, Vol. 55, No. 4, 997-1004.
- Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) "Association mapping in structure populations" Am J Hum Genet 67: 170-181.
- H. M. Kang, J. H. Sul, S. K. Service, N. A. Zaitlen, S.-Y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin, “Variance component model to account for sample structure in genome-wide association studies,” Mar. 2010.
- A. L. Price, N. A. Zaitlen, D. Reich, and N. Patterson, “New approaches to population stratification in genome-wide association studies,” Nat Rev Genet, vol. 11, no. 7, pp. 459–463, Jun. 2010.
- B. K. Bulik-Sullivan, P.-R. Loh, H. K. Finucane, S. Ripke, J. Yang, N. Patterson, M. J. Daly, A. L. Price, and B. M. Neale, “LD Score regression distinguishes confounding from polygenicity in genome-wide association studies,” Nat Genet, vol. 47, no. 3, pp. 291–295, Feb. 2015.