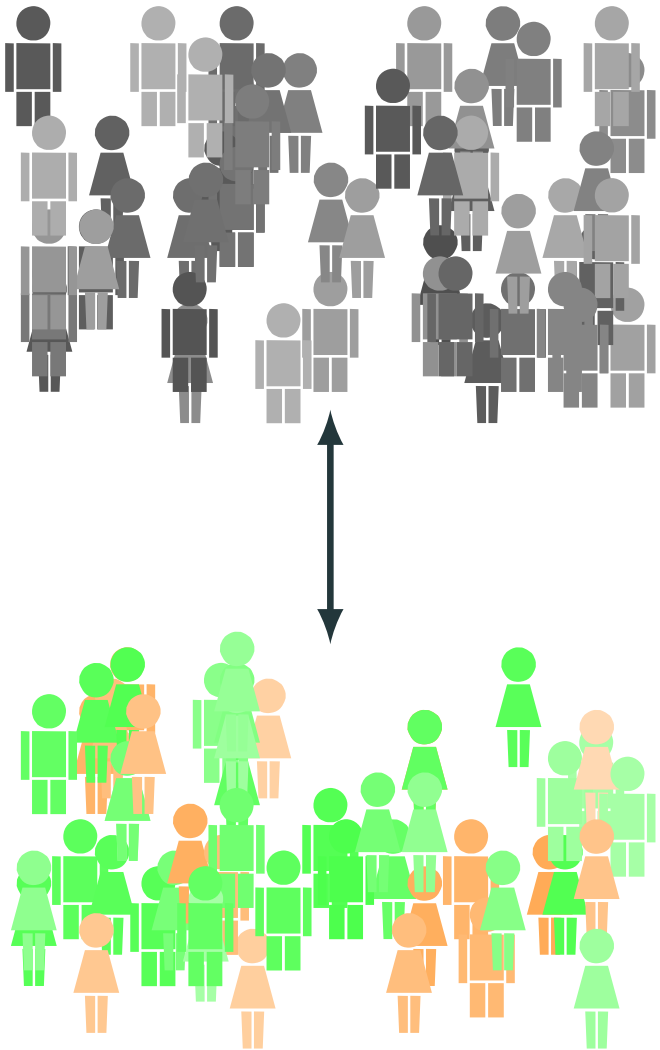
A genetic test for differential causative pathology in disease subgroups

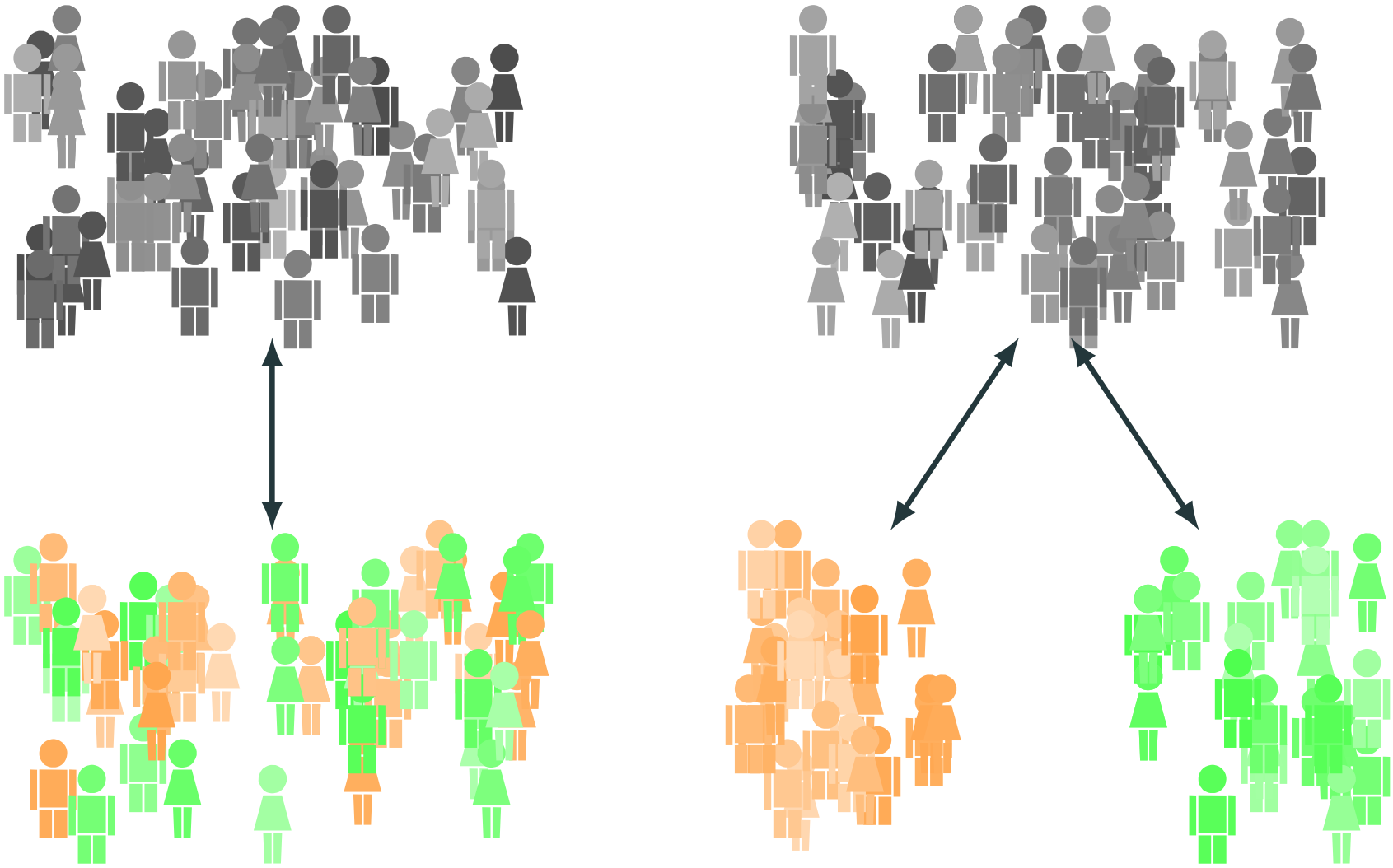
James Liley, John Todd and Chris Wallace

chr1swallace chr1swallace.github.io cew54@cam.ac.uk

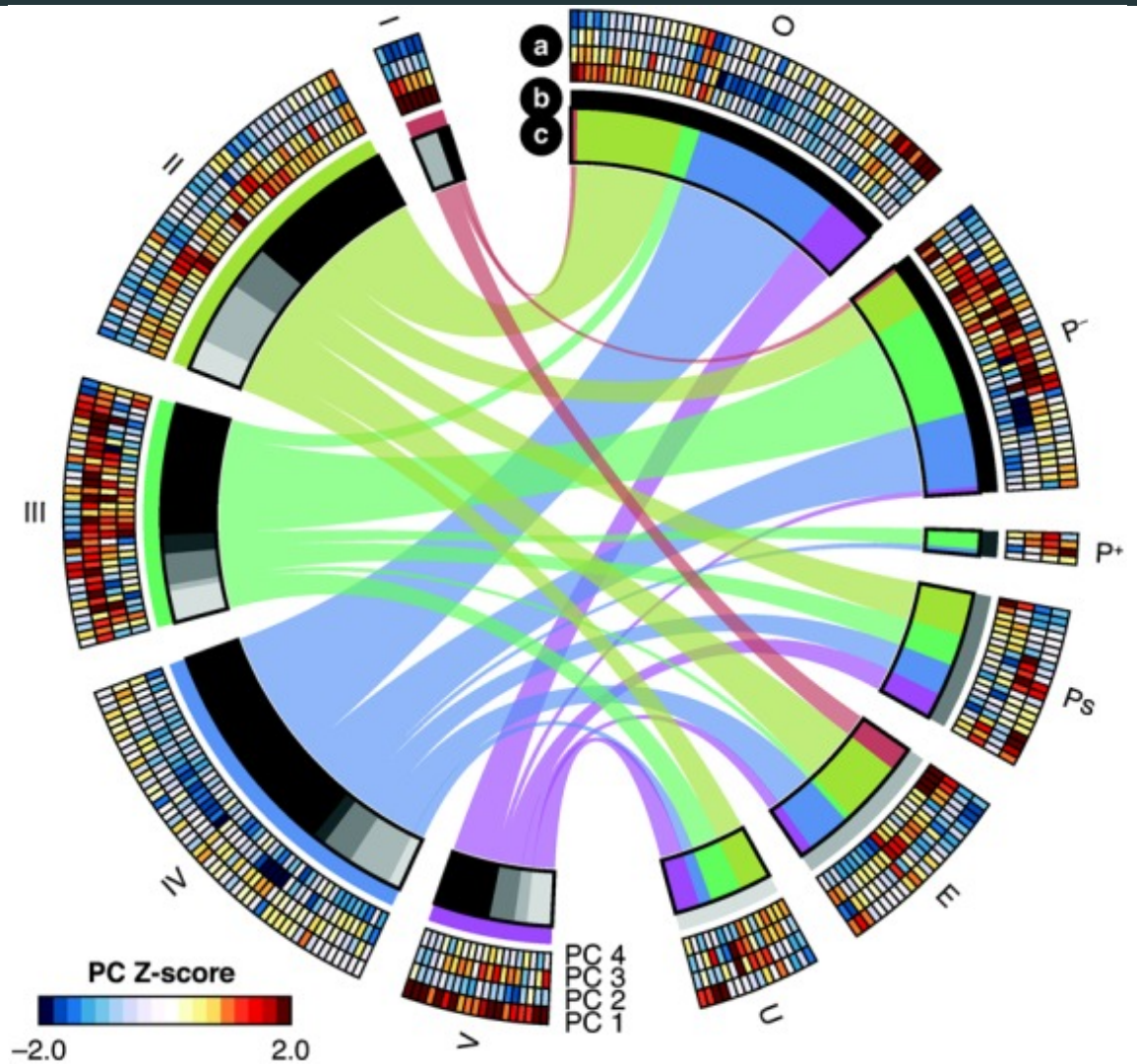
Clinical heterogeneity in disease often ignored in GWAS



Clinical heterogeneity in disease often ignored in GWAS

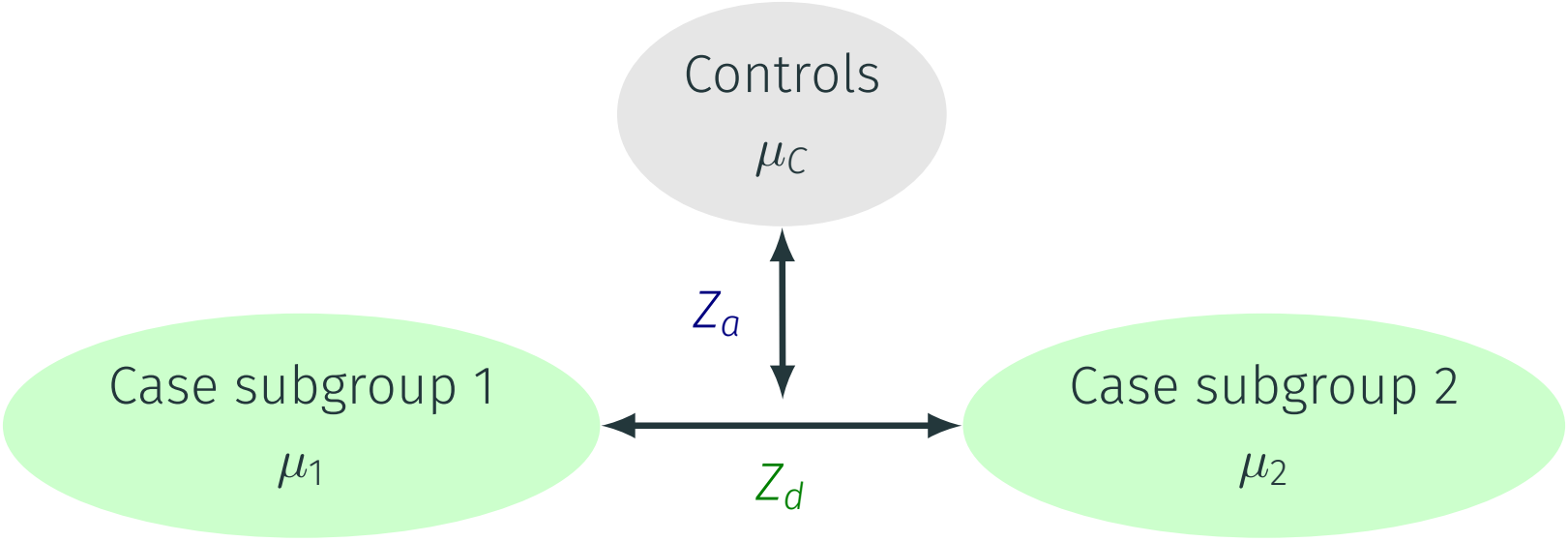


Juvenile Idiopathic Arthritis is a heterogeneous family of diseases



Eng et al., 2014

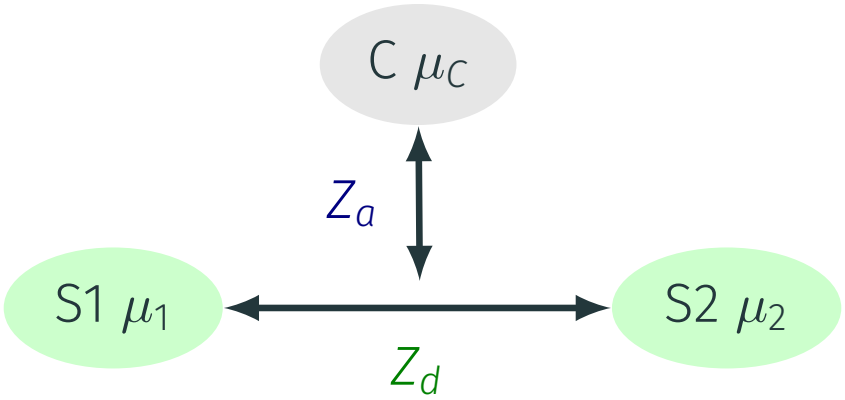
Two dimensional GWAS model



μ represents population allele frequency at a given genetic variant (SNP) in each group

Test hypotheses of the form $\mu_1 = \mu_2$ to derive a Z score at each SNP

Two dimensional GWAS model



Joint mixture Gaussian model of (Z_a, Z_d). SNPs may fall into one of three groups:

Group 1 SNPs not associated with the disease and with the same frequency in subgroups ($\mu_1 = \mu_C = \mu_2$)

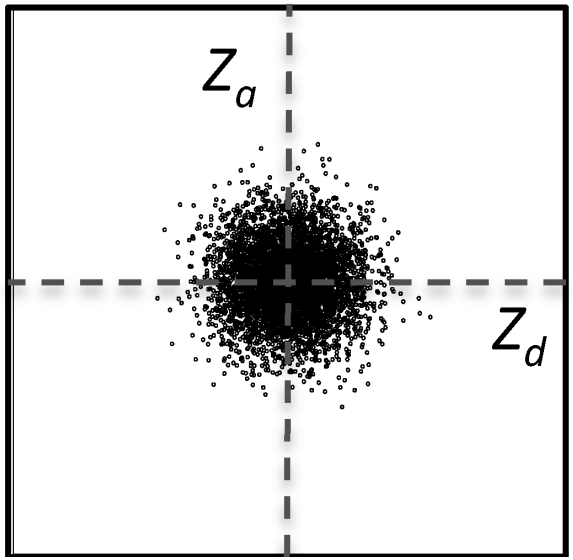
Group 2 SNPs associated with the disease, but with the same frequency in subgroups ($\mu_1 = \mu_2 \neq \mu_C$)

Group 3 SNPs with different frequencies in subgroups ($\mu_1 \neq \mu_2$)

Group 1: $\mu_C = \mu_1 = \mu_2$

Z_a, Z_d both $\sim N(0,1)$, and are independent.

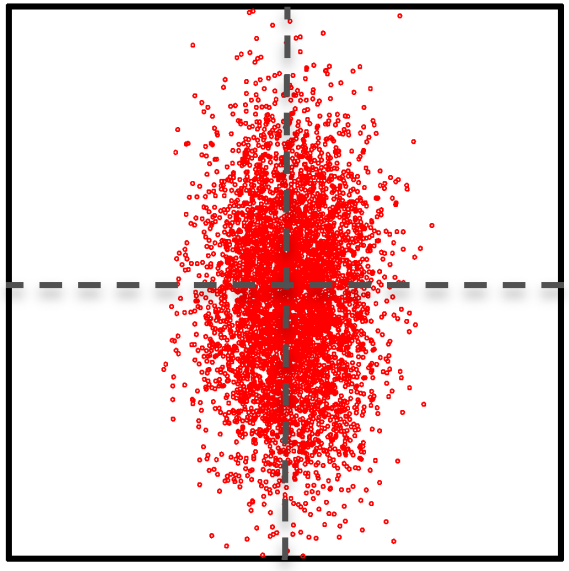
$$\begin{pmatrix} Z_d \\ Z_a \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right)$$



Group 2: $\mu_C \neq \mu_1 = \mu_2$

Assume that underlying case-control effect sizes $logOR(\mu_C, \mu_{12})$ are normally distributed with mean 0

$$\begin{pmatrix} Z_d \\ Z_a \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & \sigma_1^2 \end{pmatrix}\right)$$

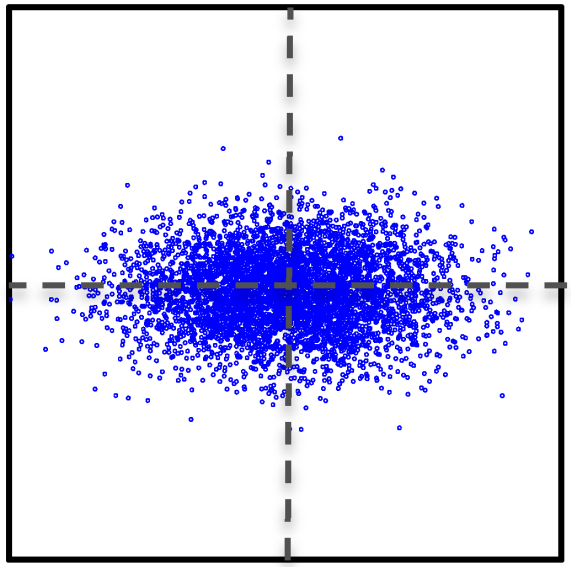


Group 3 $\mu_1 \neq \mu_2$ – Null hypothesis

Assume that underlying between subgroup effect sizes $logOR(\mu_1, \mu_2)$ are normally distributed with mean 0

The overall allele frequency should be the same in cases and controls, so $Z_a \sim N(0, 1)$

$$\begin{pmatrix} Z_d \\ Z_a \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau^2 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

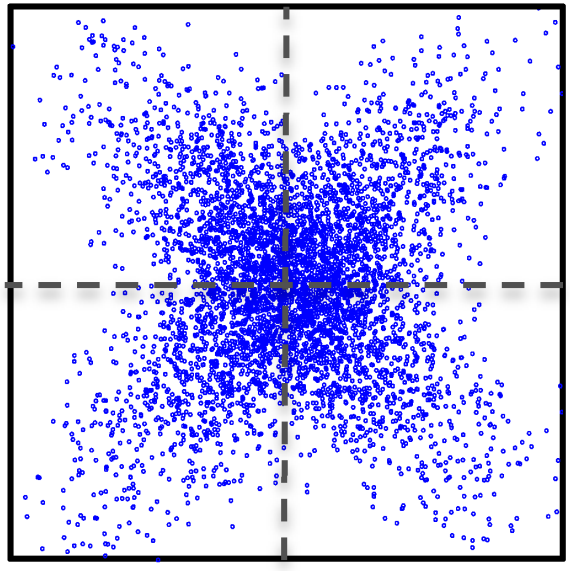


Group 3 $\mu_1 \neq \mu_2$, $\mu_{12} \neq \mu_C$ – Alternative hypothesis

If SNPs have different effect sizes between subgroups, and are associated with the phenotype as a whole, then we expect both SD(Za) > 1 and SD(Zd) > 1.

They may also be correlated.

$$\begin{pmatrix} Z_d \\ Z_a \end{pmatrix} \sim \begin{cases} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau^2 & \rho \\ \rho & \sigma_2^2 \end{pmatrix}\right) \\ N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau^2 & -\rho \\ -\rho & \sigma_2^2 \end{pmatrix}\right) \end{cases}$$



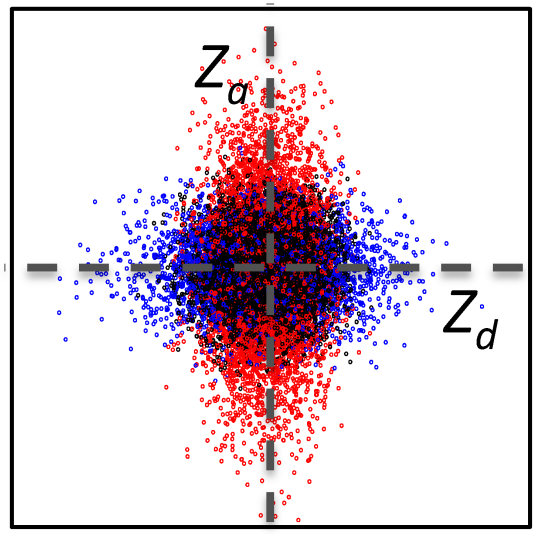
Three-Gaussian model

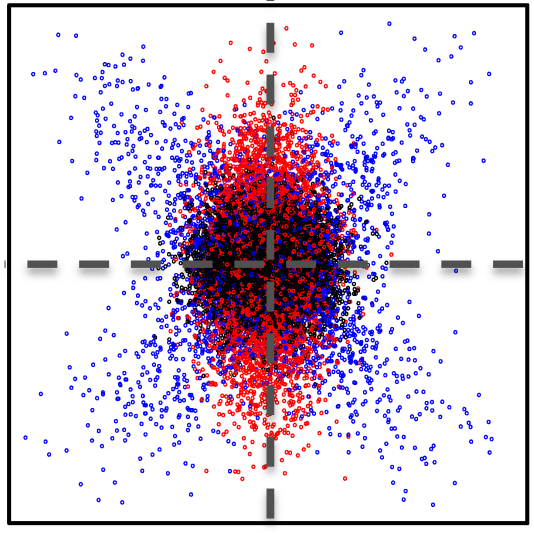
Assume proportion of SNPs in each group is π_0, π_1, π_2.

Find MLE of $\Theta_1 = (\pi_0, \pi_1, \pi_2, \sigma_1^2, \sigma_2^2, \rho)$ using E-M algorithm

Find MLE of $\Theta_0 = (\pi_0, \pi_1, \pi_2, \sigma_1^2 | \sigma_2^2 = 1, \rho = 0)$ (null model)

Compare likelihood under Θ_1 and Θ_0





Pseudo likelihood ratio test - challenges

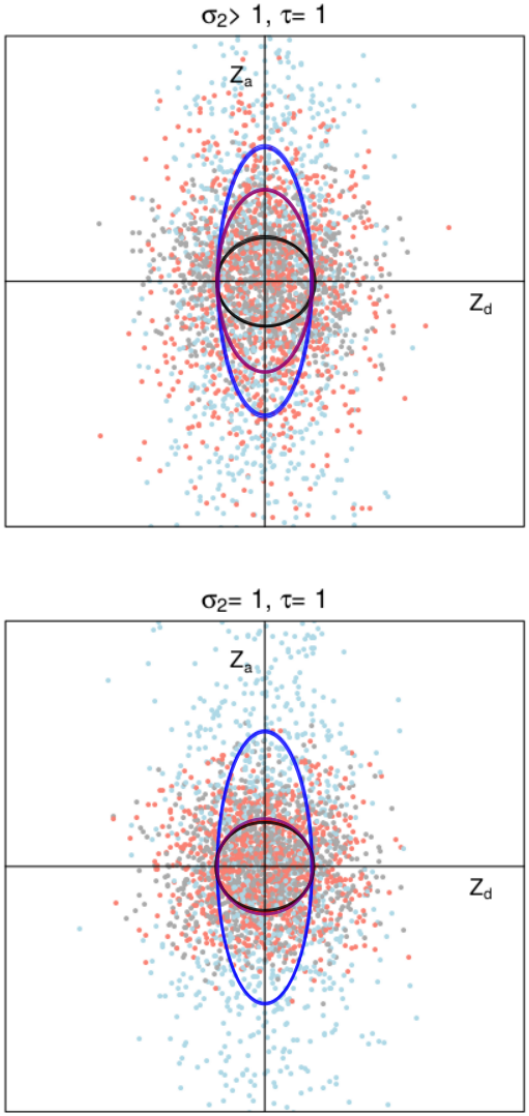
1. Observations are dependent due to linkage disequilibrium between SNPs. We weight individual contributions from individual SNPs using LDAK[1], but some residual correlation remains.

[1] *Speed et al., AJHG 2012*

Pseudo likelihood ratio test - challenges

1. Observations are dependent due to linkage disequilibrium between SNPs. We weight individual contributions from individual SNPs using LDAK[1], but some residual correlation remains.
2. If there are no SNPs in group 3 and log OR not exactly normal, then H_1 will always fit better.
   We condition on Z_a

$$PLR = \frac{\prod_i w_i \times PDF\left(Z_d^{(i)} | Z_a^{(i)}; \Theta_1\right)}{\prod_i w_i \times PDF\left(Z_d^{(i)} | Z_a^{(i)}; \Theta_0\right)}$$



[1] *Speed et al., AJHG 2012*

Null distribution of PLR

Null parameter values are on a boundary, so PLR will have a mixture χ^2 distribution

Non-independence between SNPs results in scaling of mixture χ^2 distribution

Null distribution of PLR is a scaled and transposed χ^2 distribution:

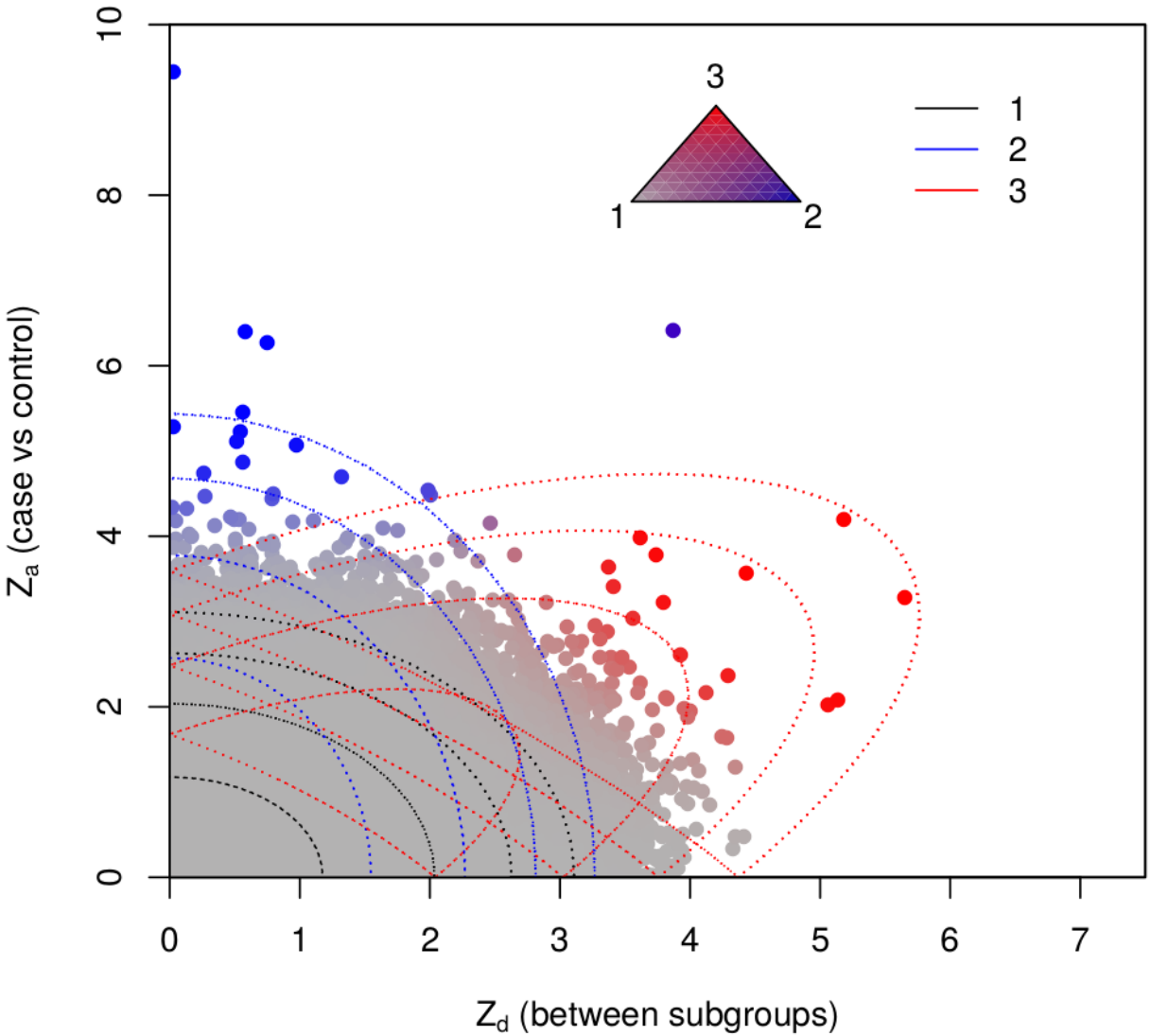
$$PLR \sim \begin{cases} \gamma\chi^2_1 & p = \kappa \\ \gamma\chi^2_2 & p = 1 - \kappa \end{cases}$$

γ depends on the covariance matrix (LD) between Z scores through the weights $\{w_i\}$

κ depends on probability $\rho = 0$ - approximately 0.5.

These parameters can be estimated by resampling.

Results for T1D/RA as subgroups of “autoimmune disease”



$p = 3 \times 10^{12}$

data from WTCCC

Post-hoc single SNP analysis

Several options.

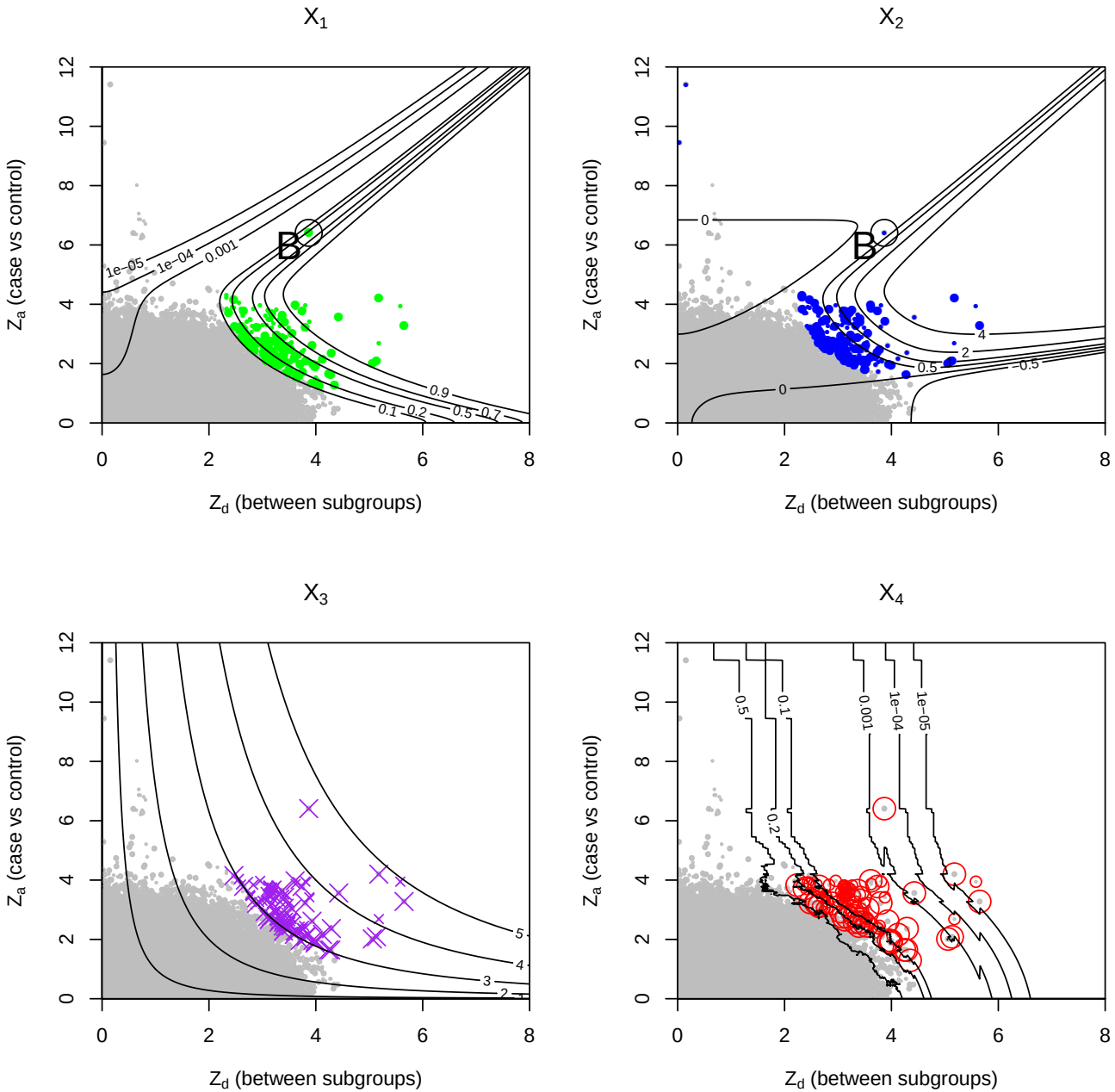
- Posterior probability of group 3 membership - can be large when $|Z_a|$ large but $|Z_d|$ small
- $\log P\left(Z_a, Z_d \mid \Theta_1\right) - \log P\left(Z_a, Z_d \mid \Theta_0\right)$ - sensitive to fitted Θ_1
- Conditional false discovery rate for related null hypothesis $H_0' : \mu_1 = \mu_2$

$$P\left(H_0' \mid \widetilde{Z}_a \geq z_a, \widetilde{Z}_d \geq z_d\right) = \frac{P\left(\widetilde{Z}_a \geq z_a, \widetilde{Z}_d \geq z_d \mid \mu_1 = \mu_2\right) P\left(H_0'\right)}{P\left(\widetilde{Z}_a \geq z_a, \widetilde{Z}_d \geq z_d\right)}$$

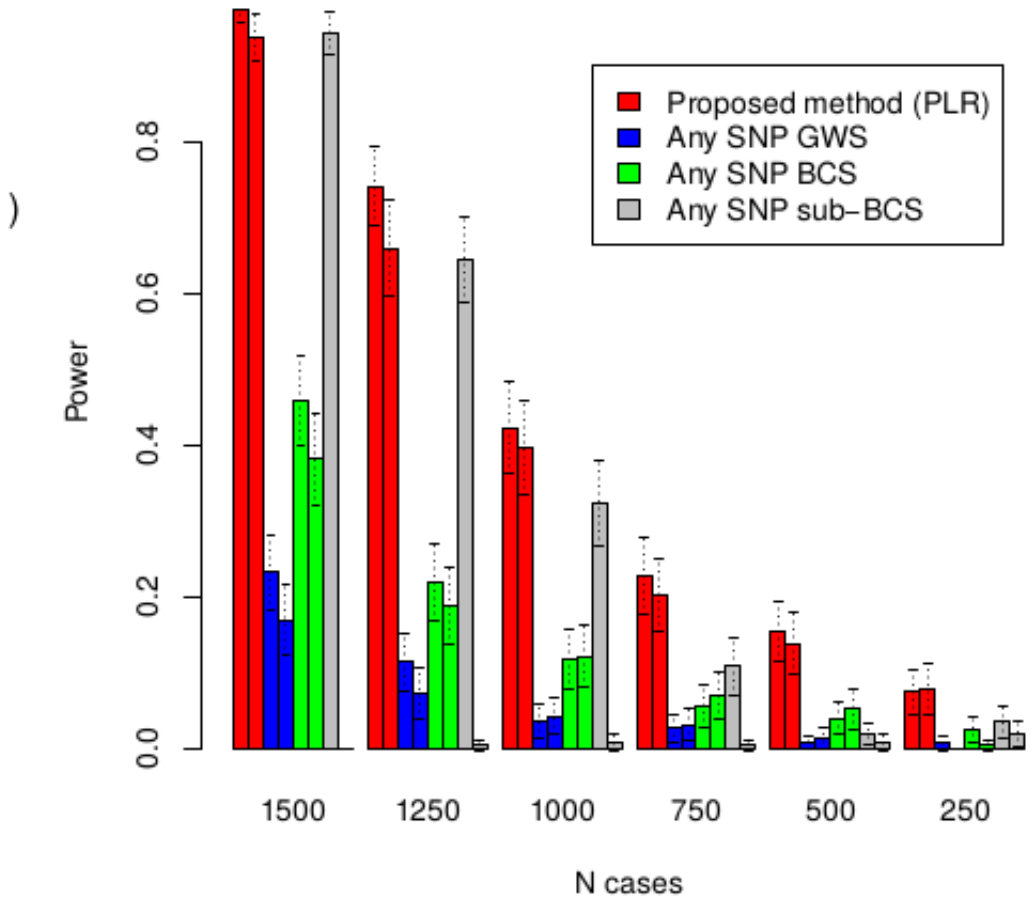
$$\leq \frac{P\left(\widetilde{Z}_d > z_d \mid \widetilde{Z}_d \sim N(0,1)\right) \times P\left(\widetilde{Z}_a \geq z_a\right) \times 1}{P\left(\widetilde{Z}_a \geq z_a, \widetilde{Z}_d \geq z_d\right)}$$

where $\widetilde{Z} = |Z|$

Post-hoc single SNP analysis



Power of PLR vs single SNP significance



Auto-antibody specific type 1 diabetes subtyping

	Model	π_0	π_1	π_2	σ_1	σ_2	τ	ρ	p-value
TPO-Ab	Full	0.511	0.487	2.407×10^{-3}	0.994	6.545	1.552	0.991	$< 1 \times 10^{-20}$
	Null	0.987	2.333×10^{-3}	0.011	6.634	-	1.308	-	
TPO-Ab	Full	0.997	2.898×10^{-4}	3.031×10^{-3}	4.698	2.291	1.497	0.338	1.5×10^{-4}
no MHC	Null	0.989	1.882×10^{-3}	9.087×10^{-3}	3.11	-	1.318	-	
GAD-Ab	Full	0.995	3.557×10^{-3}	1.057×10^{-3}	2.832	8.866	2.295	5.484	$< 1 \times 10^{-20}$
	Null	0.997	2.328×10^{-3}	3.002×10^{-4}	6.639	-	2.153	-	
GAD-Ab	Full	0.997	2.9×10^{-3}	3.434×10^{-4}	2.279	4.531	1.055	3.424	0.002
no MHC	Null	0.792	1.883×10^{-3}	0.206	3.111	-	0.997	-	
IA2-Ab	Full	0.995	3.275×10^{-3}	1.244×10^{-3}	2.804	8.291	3.027	1.575	$< 1 \times 10^{-20}$
	Null	0.997	2.287×10^{-3}	3.805×10^{-4}	6.674	-	3.852	-	
IA2-Ab	Full	0.998	1.362×10^{-3}	7.904×10^{-4}	3.318	2.212	2.145	0	0.008
no MHC	Null	0.998	1.88×10^{-3}	2.073×10^{-4}	3.112	-	2.889	-	
PCA-Ab	Full	0.997	2.336×10^{-3}	3.413×10^{-4}	6.631	0.37	2.097	0.422	> 0.5
	Null	0.998	2.335×10^{-3}	1.276×10^{-4}	6.632	-	2.54	-	
PCA-Ab	Full	0.997	2.759×10^{-3}	1.303×10^{-4}	2.508	5.58	2.256	0	> 0.5
no MHC	Null	0.998	1.884×10^{-3}	1.384×10^{-4}	3.111	-	2.5	-	

Relationship to genetic correlation

Genetic heritability, σ_g^2, can be estimated by partitioning the covariance matrix for a single trait X measured in n individuals with kinship matrix Φ

$$\Omega = 2\Phi\sigma_g^2 + I_n\sigma_e^2$$

where $\Omega_{i,j} = cov(X_i, X_j)$.

Similarly, genetic correlation $r_g = \frac{\sigma^2_{g_{XY}}}{\sigma_{g_X}\sigma_{g_Y}}$ between two traits can be estimated from by partitioning the bivariate correlation matrix

$$\Omega_{XY} = 2\Phi\sigma^2_{g_{XY}} + I_n\sigma_e^2$$

where $\Omega_{XY_{i,j}} = cov(X_i, Y_j)$.

Relationship to genetic correlation

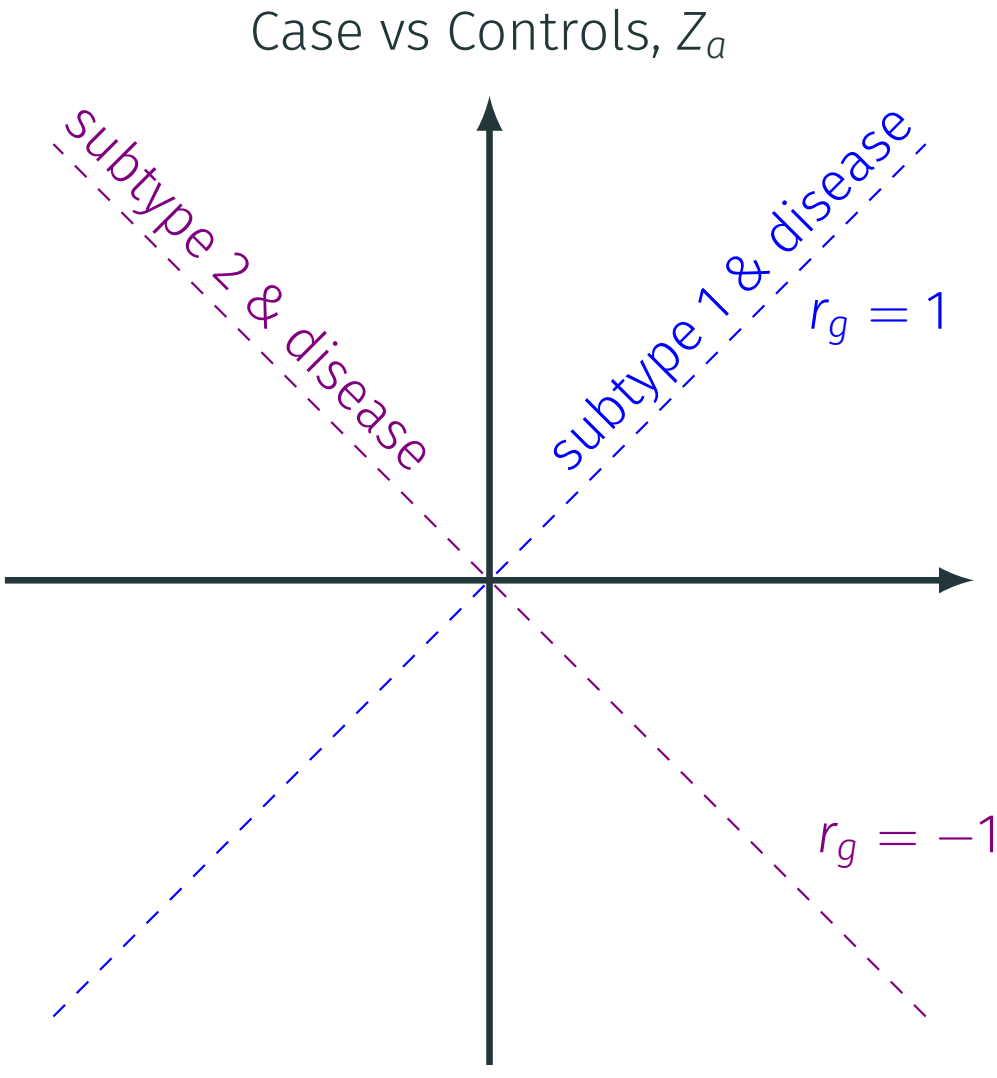
Genetic correlation can also be estimated directly from GWAS data for two traits from distinct datasets[1].

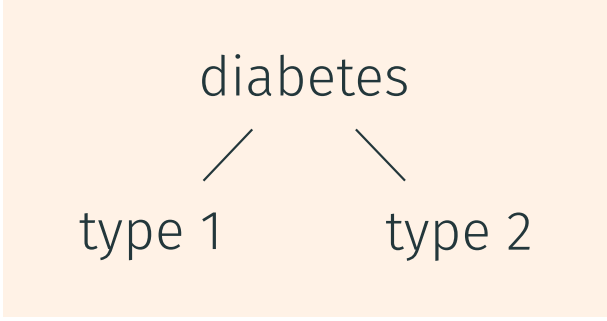
Can subtypes be detected by testing r_g for specific GWAS comparisons?

$$r_g(Z_a, Z_d) > 0 \qquad \text{or} \qquad r_g(S1 \textit{ vs } C, S2 \textit{ vs } C) < 1$$

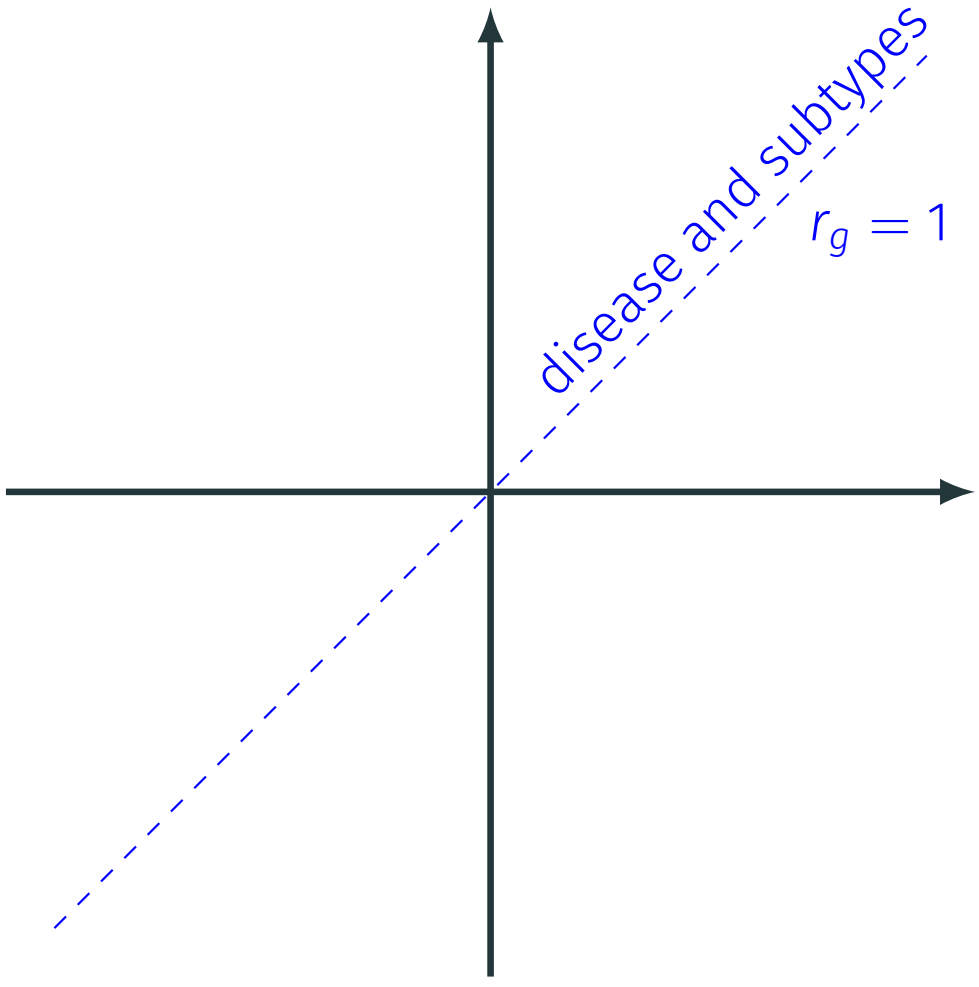
[1] *Bulik-Sullivan et al., Nature 2015*

$r_g(Z_a, Z_d) > 0$
tests correlation of signed rather than absolute Z scores

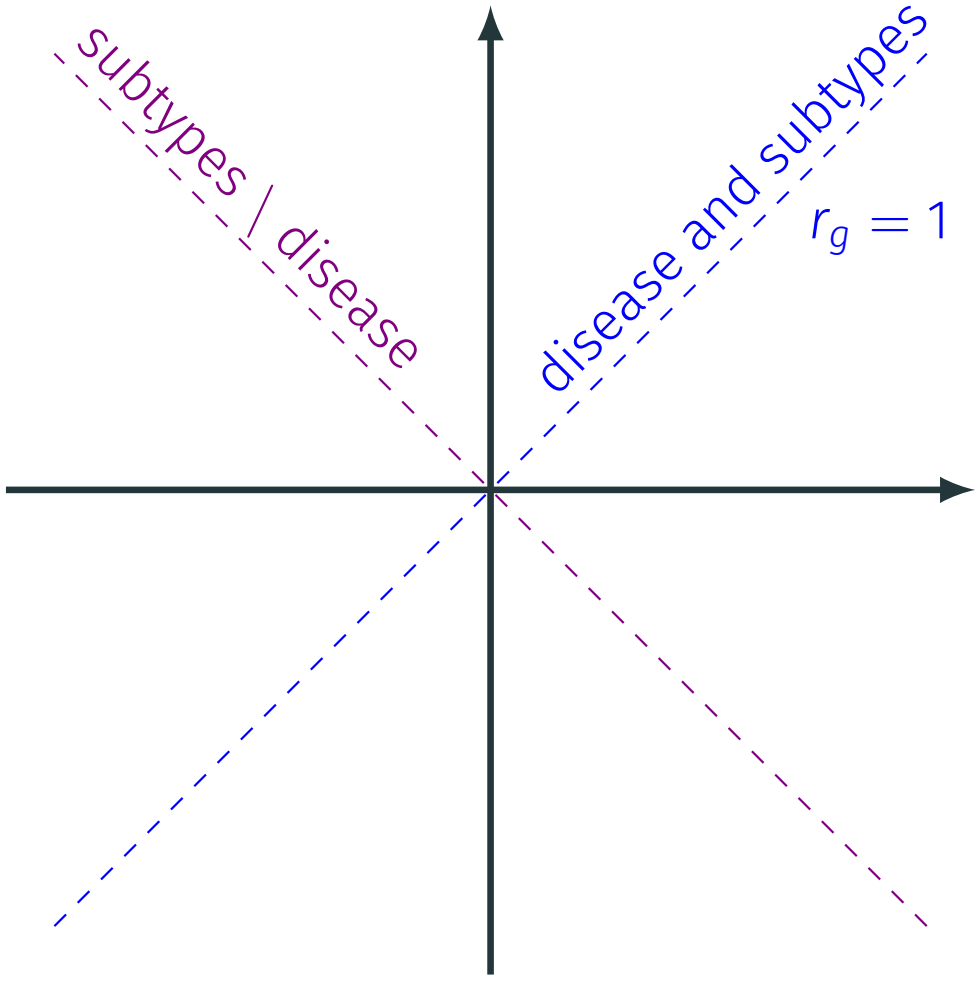




$r_g(S1 \text{ vs } C, S2 \text{ vs } C) < 1$
assumes no disease-independent variants distinguish subtypes

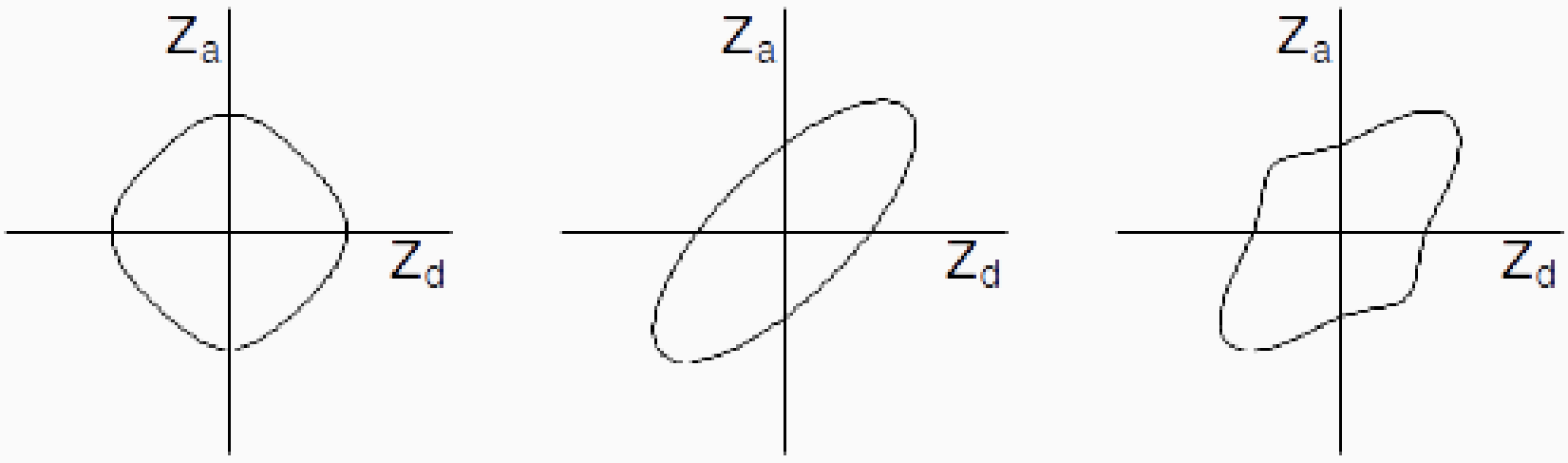


$r_g(S1 \text{ vs } C, S2 \text{ vs } C) < 1$
assumes no disease-independent variants distinguish subtypes



Future Directions

Further inference of causes of heterogeneity



Applications to other diseases

JIA, vasculitis

Acknowledgements

UNIVERSITY OF CAMBRIDGE

wellcome trust

MRC | Biostatistics Unit

NHS
National Institute for Health Research

JDRF IMPROVING LIVES. CURING TYPE 1 DIABETES.

James Liley

Software: https://github.com/jamesliley/subtest

Preprint: http://biorxiv.org/content/early/2016/08/02/037713