Mathematical models of the genetic architecture in complex human disorders

Oleksandr Frei

November 22, 2019

Table of contents

Home UiO > Faculty of Medicine > NORMENT: Norwegian Centre for Mental Disorders Research

For employees Norwegian website Search

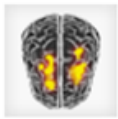
UiO : NORMENT: Norwegian Centre for Mental Disorders Research
Faculty of Medicine

Home Research About the centre People

Research topics

Genetics

Identify rare genetic variants or expression variation to reveal "missing heritability".

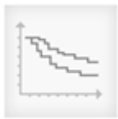


Brain Imaging

Determine new brain imaging phenotypes linking genes and core clinical phenotypes.

Antipsychotic Medication

Define new targets to optimize the ratio of beneficial vs. adverse effects of antipsychotics.



Outcome Predictors

Using genetic and environmental factors to estimate illness course and outcome.

NORMENT

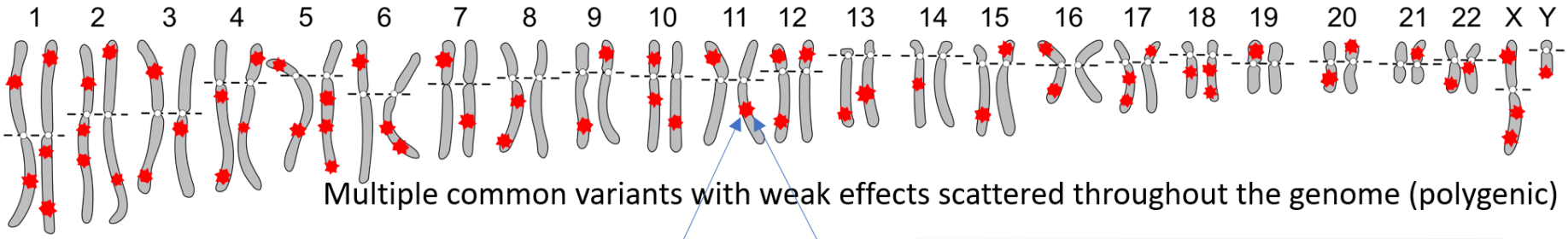
NORMENT is a Centre of Excellence (CoE) funded by the Research Council of Norway.

Our main goal is to find answers to why some people develop severe mental illness.

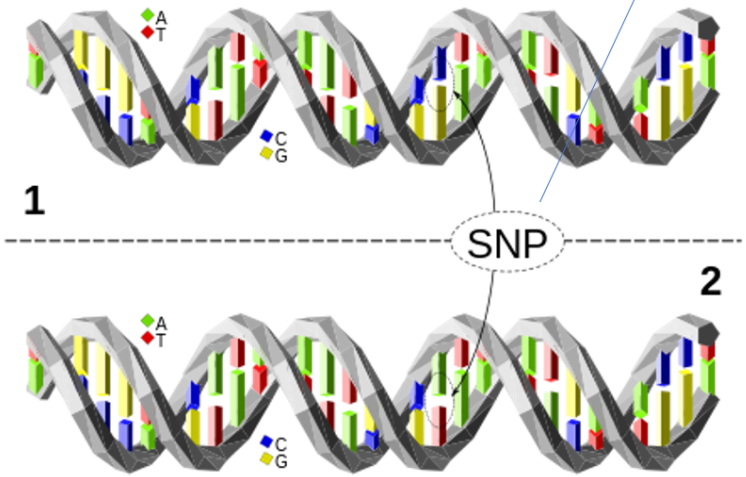
→ Read more about NORMENT

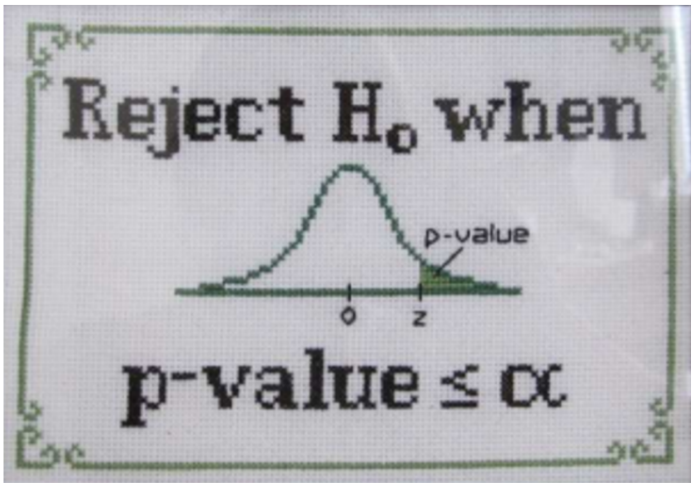
NORMENT in social media

Genetics of complex human traits

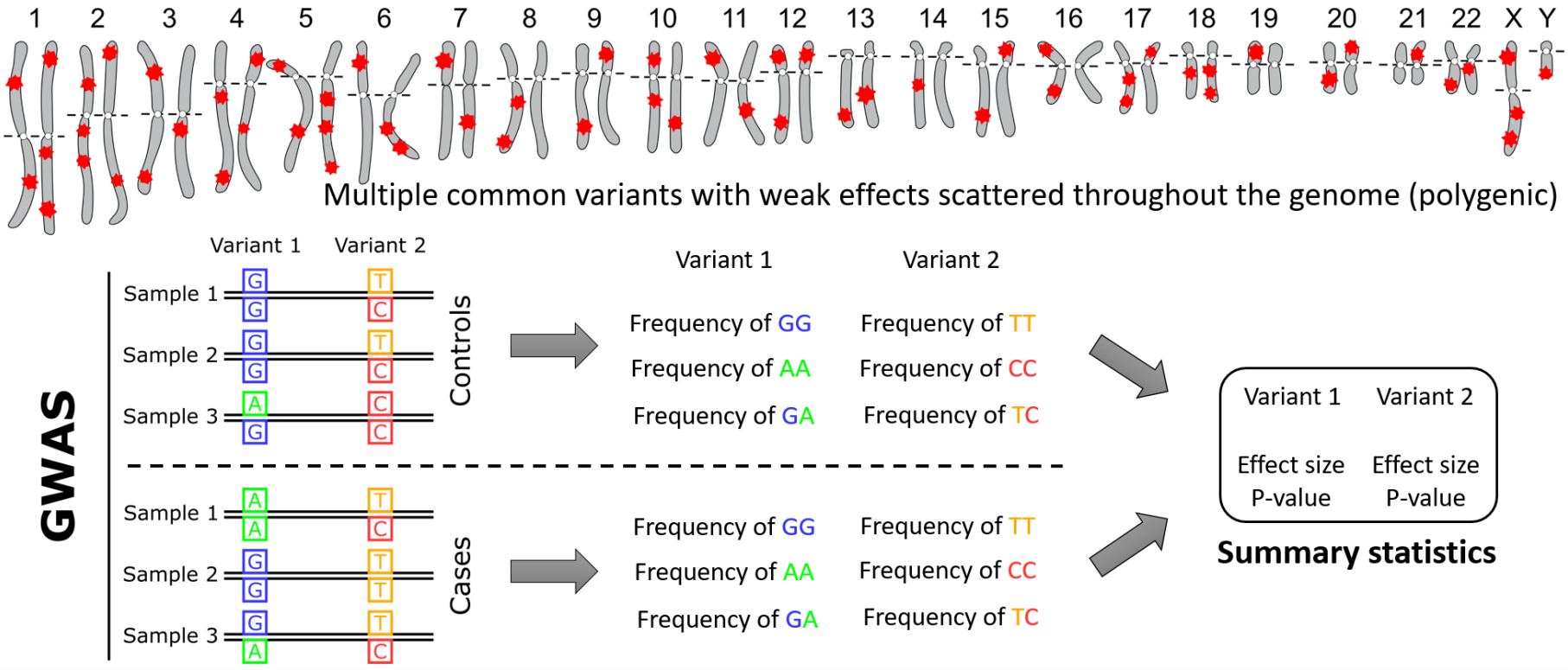


Multiple common variants with weak effects scattered throughout the genome (polygenic)





Genetics of complex human traits



Think of the naïve bayes classifier
- genetic variants = features
- human trait = output

Simple additive genetic model

$$y_k = \sum_{i=1}^{M} g_{ki}\beta_i + e \quad \leftrightarrow \quad \mathbf{y} = G\beta + e$$

where

- N - the number of individuals in the dataset
- M - the number of genetic variants
- $\mathbf{y}$ - N-vector, "phenotype" (e.g. human height)
- G - NxM-matrix
- β - M-vector, genetic effects, random variables
- e - non-genetic effects, random variable
- $\mathbf{y}$, G - known; β, e - unknown

Estimate $\hat{\beta} = G^{+}y$?

$$y_k = \sum_{i=1}^{M} g_{ki}\beta_i + e \quad \leftrightarrow \quad \mathbf{y} = G\beta + e$$

$$\hat{\beta}_i = \frac{\mathbf{y}^T\mathbf{v}_i}{\mathbf{v}_i^T\mathbf{v}_i} \propto corr(\mathbf{y}, \mathbf{v}_i), \text{ where } \mathbf{v}_i = (g_{1i}, g_{2i}, \cdots, g_{Ni});$$

$$z_i = \frac{\hat{\beta}_i}{se(\beta_i)} = r_i\sqrt{N-2}\sqrt{1-r_i^2}, \quad r_i = corr(\mathbf{y}, \mathbf{v}_i)$$

Correlation among genetic features

Simple Additive Genetic Model

$$y_k = \sum_{i=1}^{M} g_{ki}\beta_i + e \quad \leftrightarrow \quad \mathbf{y} = G\beta + e$$

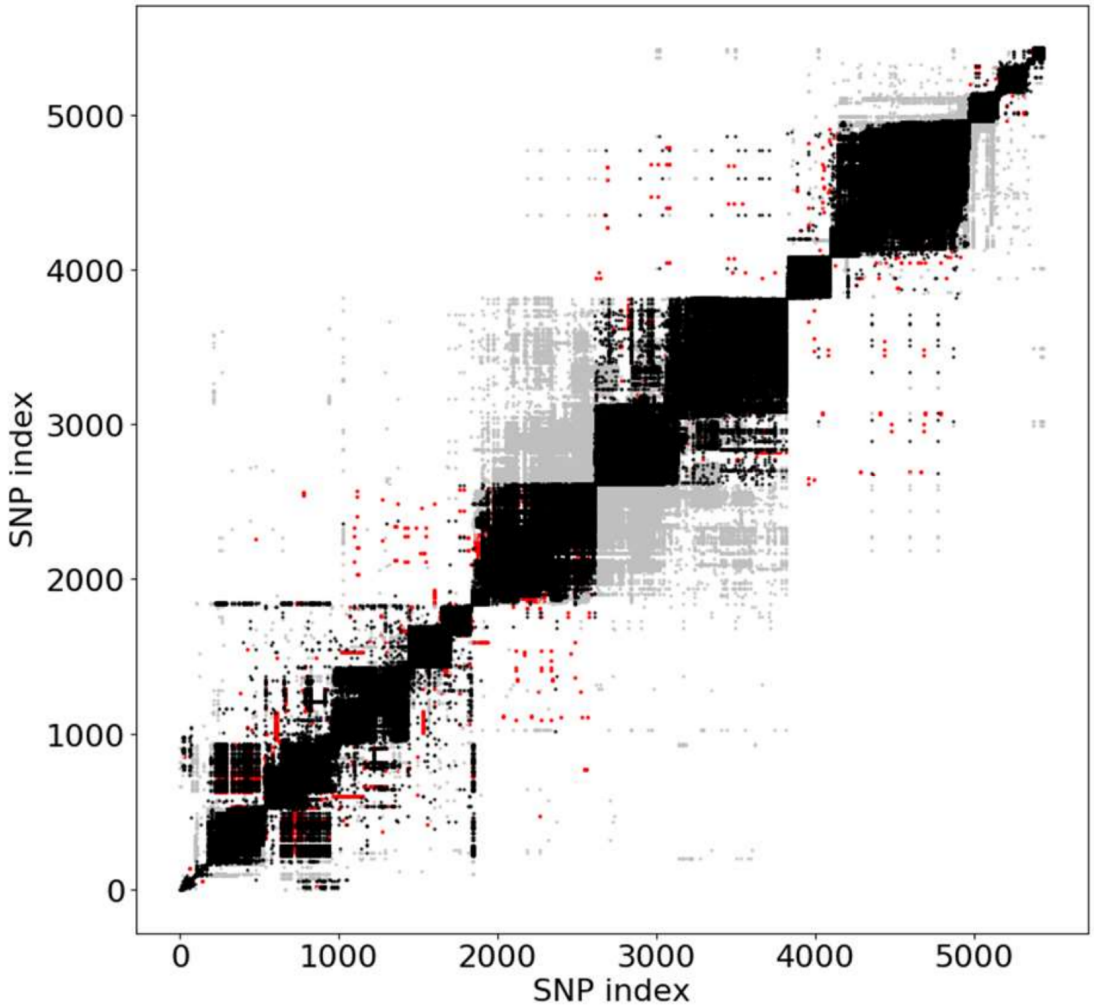
Theorem:

$$z_j = \sum_{i=1}^{M} a_{ij}\beta_i + \epsilon \quad \leftrightarrow \quad \mathbf{z} = A\beta + \epsilon$$

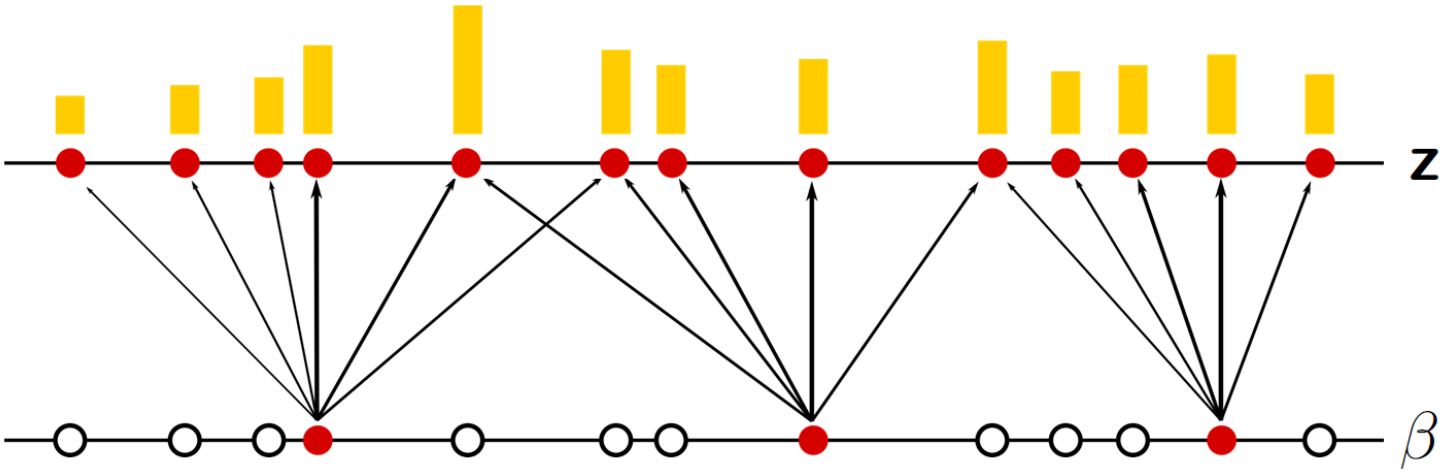
where

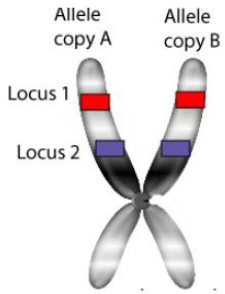
- $\mathbf{z}$ - M-vector, derived from $\mathbf{y}$ and G
- A - MxM matrix, derived from G, sparse banded matrix ($a_{ij} = \sigma_0\sqrt{N_j H_i} r_{ij}$, where $r_{ij} = corr(\mathbf{v}_i, \mathbf{v}_j)$)
- β - as before
- $\epsilon \sim N(0, \sigma_0^2)$

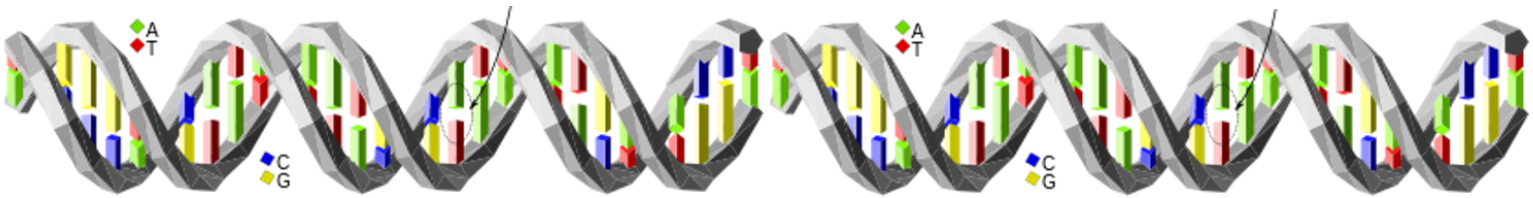
Correlation matrix A ("Linkage Disequilibrium")



Correlation among genetic features: β vs z







MiXeR prior distribution on β

$$\mathbf{y} = G\beta + e, \text{ or}$$
$$\mathbf{z} = A\beta + \epsilon$$

MiXeR:

$$\beta_i \sim (1 - \pi_1)N(0, 0) + \pi_1 N(0, \sigma^2_\beta)$$

where

- π_1 - weight in the mixture
- σ^2_β - variance
- $N(0, 0)$ - probability mass at zero

How to compute the likelihood function?

$$z_j = \sum_{i=1}^{M} a_{ij}\beta_i + \epsilon \quad \leftrightarrow \mathbf{z} = A\beta + \epsilon$$

where

- $\mathbf{z}$ - M-vector, derived from $\mathbf{y}$ and G
- A - MxM matrix, derived from G. Sparse banded matrix
- β - as before, $\beta_i \sim (1-\pi_1)N(0,0) + \pi_1 N(0,\sigma_\beta^2)$
- $\epsilon \sim N(0,\sigma_0^2)$

How to compute $p(\mathbf{z}_j|\theta)$? $(\theta = (\pi_1, \sigma_\beta^2, \sigma_0^2))$
Max likelihood: $\sum_j w_j \; log \; p(z_j|\theta) \rightarrow min_\theta$

Sampling, Convolution, Gaussian approximation

$$z_j = \sum_{i=1}^{M} a_{ij}\beta_i + \epsilon$$

$$\beta_i \sim (1-\pi_1)N(0,0) + \pi_1 N(0,\sigma_\beta^2)$$

$$\epsilon \sim N(0,\sigma_0^2)$$

$$p(\mathbf{z}_j|\pi_1,\sigma_\beta^2,\sigma_0^2) = \text{ ????}$$

Easy if $\pi_1 = 1$:

$$z_j \sim N(0,(\sum_i a_{ij}^2)\sigma_\beta^2 + \sigma_0^2)$$

Sampling: generate T subsets with prior π_1, $C_t \subset \{1,...,M\}$

$$z_j \sim \frac{1}{T}\sum_{t=1}^{T} N(0,(\sum_{i\in C_t} a_{ij}^2)\sigma_\beta^2 + \sigma_0^2)$$

Sampling, **Convolution**, Gaussian approximation

$$pdf_z(z_0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itz_0}\phi_z(t)dt, \quad \phi_z(t) - \text{characteristic function of z}$$

$$pdf_z(z_0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \cos(tz_0)\phi_z(t)dt - \frac{i}{2\pi}\int_{-\infty}^{\infty} \sin(tz_0)\phi_z(t)dt$$

$$pdf_z(z_0) = \frac{1}{\pi}\int_{0}^{\infty} \cos(tz_0)\phi_z(t)dt - \text{for an even function } \phi_z(t)$$

$$z_j = \epsilon + \sum_i \xi_i, \quad \xi_i = \begin{cases} 0, & 1-\pi_1 \\ \mathcal{N}(0,\sigma_{ij}^2), & \pi_1 \end{cases}, \quad \sigma_{ij}^2 = a_{ij}^2\sigma_\beta^2, \quad \epsilon \sim \mathcal{N}(0,\sigma_0^2)$$

$$\phi_z(t) = \phi_\epsilon(t)\prod_i \phi_{\xi_i}(t) - \text{convolution theorem}$$

$$\phi_{\xi_i}(t) = (1-\pi_1) + \pi_1 e^{-\frac{t^2\sigma_{ij}^2}{2}}$$

Sampling, Convolution, **Gaussian approximation**

Lemma. $E\beta^2 = \pi_1\sigma_\beta^2$ and $E\beta^4 = 3\pi_1\sigma_\beta^4$.

Lemma. Let $\delta_j = \sum_{i=1}^{M} a_{ij}\beta_i$, where β_i are independent zero-mean random variables. Then

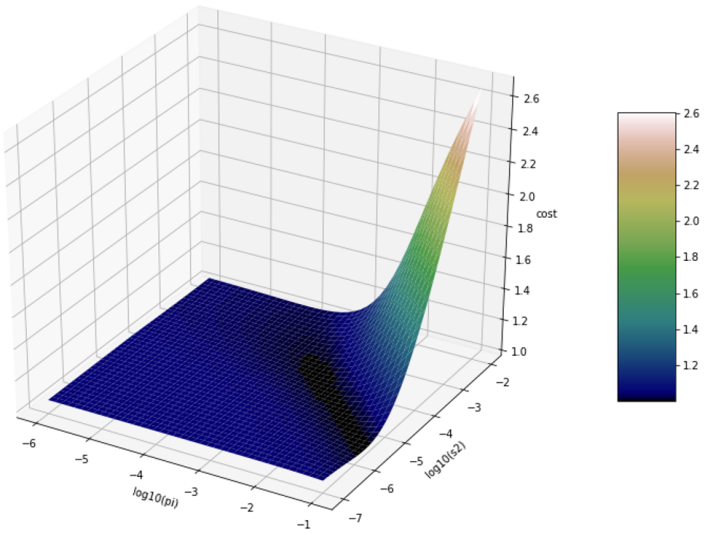
$$\begin{aligned} E\delta_j^2 &= \sum_i a_{ij}^2 E\beta_i^2 \\ E\delta_j^4 &= \sum_i a_{ij}^4 \left(E\beta_i^4 - 3(E\beta_i^2)^2 \right) + 3(E\delta_j^2)^2 \end{aligned} \tag{1}$$

Lemma. Let $A > 0$ and $B > 0$ be two numeric values. Let $\tilde{\delta}_j = p_0 N(0,0) + p_1 N(0, s^2)$ where

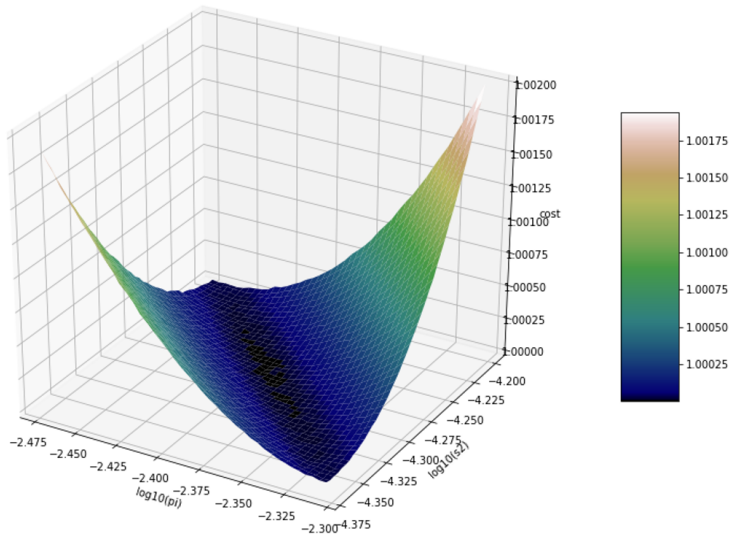
$$p_0 = \frac{B}{B+3A^2}, \quad p_1 = \frac{3A^2}{B+3A^2}, \quad s^2 = \frac{B+3A^2}{3A}. \tag{2}$$

Then $E\tilde{\delta}_j^2 = A$, and $E\tilde{\delta}_j^4 - 3(E\tilde{\delta}_j^2)^2 = B$.

Energy landscape (log-likelihood)

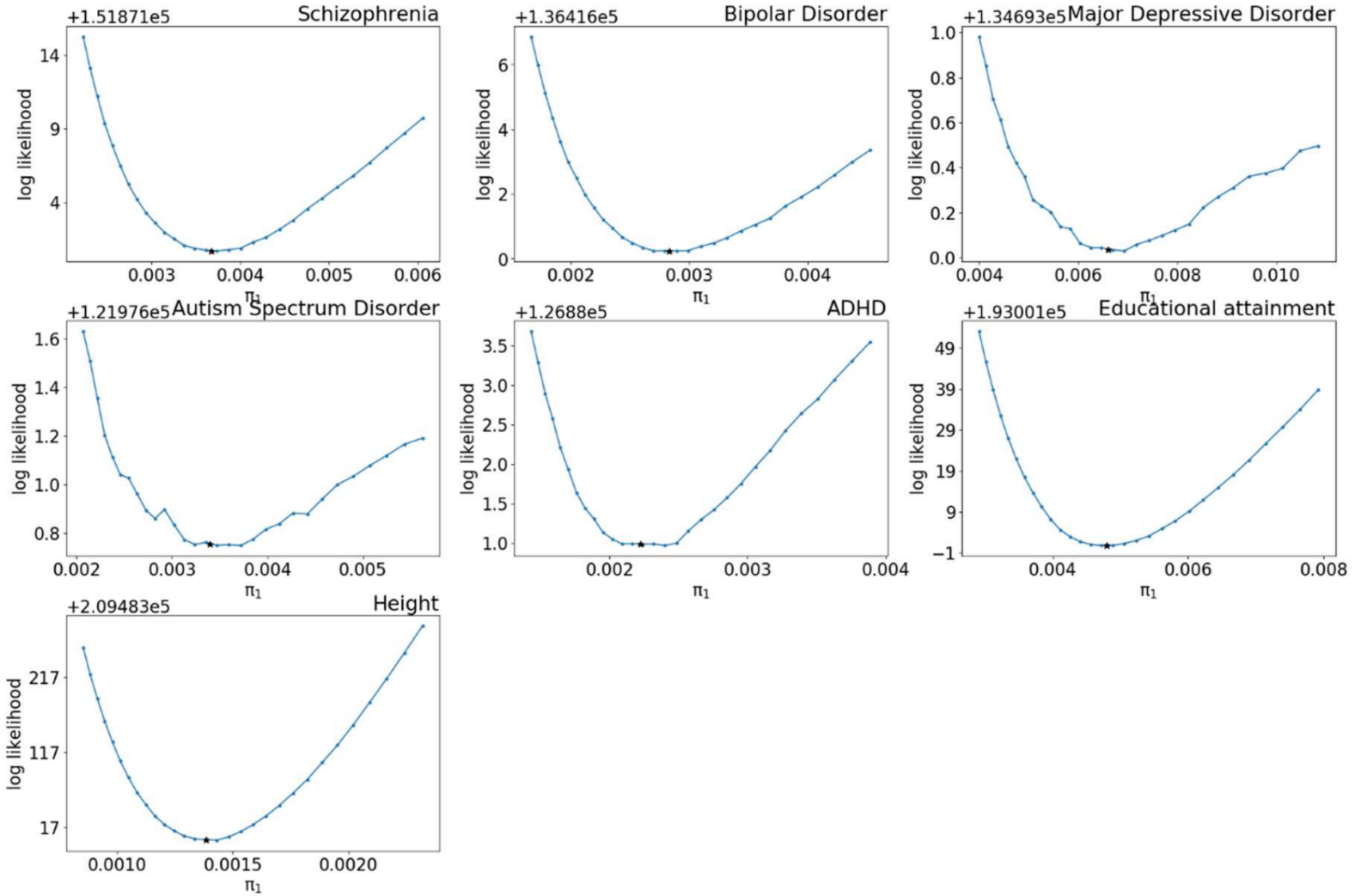


bird's-eye view



zoom into the region around observed minima

Energy landscape (log-likelihood)



Posterior integrals

$$z_j = \delta_j + \epsilon$$

$$C_j^q(z) = \int \delta^q P_j(z|\delta) P_j(\delta) d\delta \quad \propto \quad E(\delta^q | z_j).$$

$$S = \frac{\sum_j \int_{z:\ |z| \geq z_t} C_j^2(z) dz}{\sum_j \int_z C_j^2(z) dz}$$

Let $\delta_j \sim \frac{1}{K} \sum_k N(0, S_{kj}^2)$, $\epsilon \sim N(0, \sigma_0^2)$. Then

$$\int_{z:\ |z| \geq z_t} C_j^2(z) dz = \frac{1}{K} \sum_{k=1}^{K} \Bigg[\frac{\sqrt{\frac{2}{\pi}} S_{kj}^4 z_t}{(\sigma_0^2 + S_{kj}^2)^{\frac{3}{2}}} e^{-\frac{z_t^2}{2(\sigma_0^2 + S_{kj}^2)}} + S_{kj}^2 \mathrm{erfc}\left(\frac{z_t}{\sqrt{2(\sigma_0^2 + S_{kj}^2)}} \right) \Bigg],$$

where $\mathrm{erfc}(z) = \frac{2}{\sqrt{\pi}} \int_z^\infty e^{-t^2} dt$ is the complementary error fun.

nature COMMUNICATIONS

Bivariate causal mixture model quantifies polygenic overlap between complex traits beyond genetic correlation

Oleksandr Frei ✉, Dominic Holland, Olav B. Smeland, Alexey A. Shadrin, Chun Chieh Fan, Steffen Maeland, Kevin S. O'Connell, Yunpeng Wang, Srdjan Djurovic, Wesley K. Thompson, Ole A. Andreassen & Anders M. Dale ✉

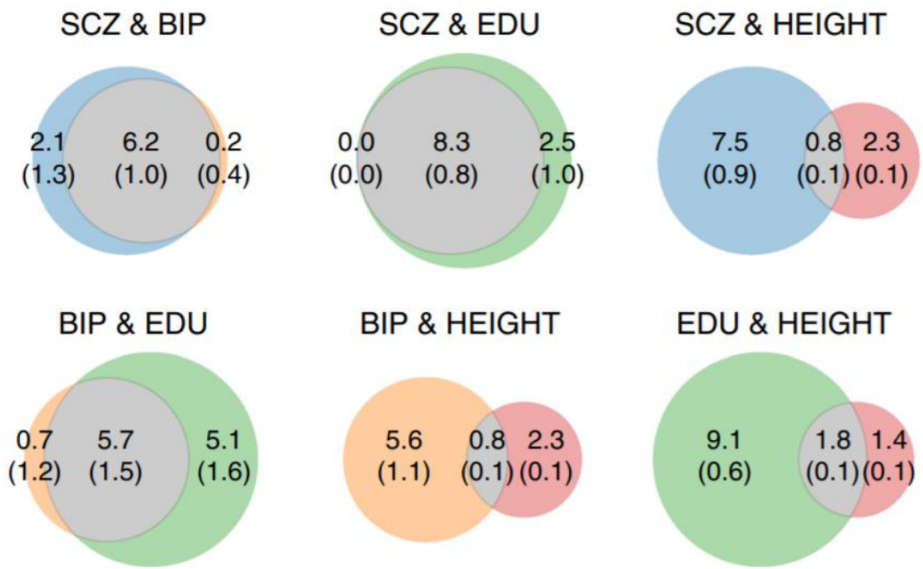
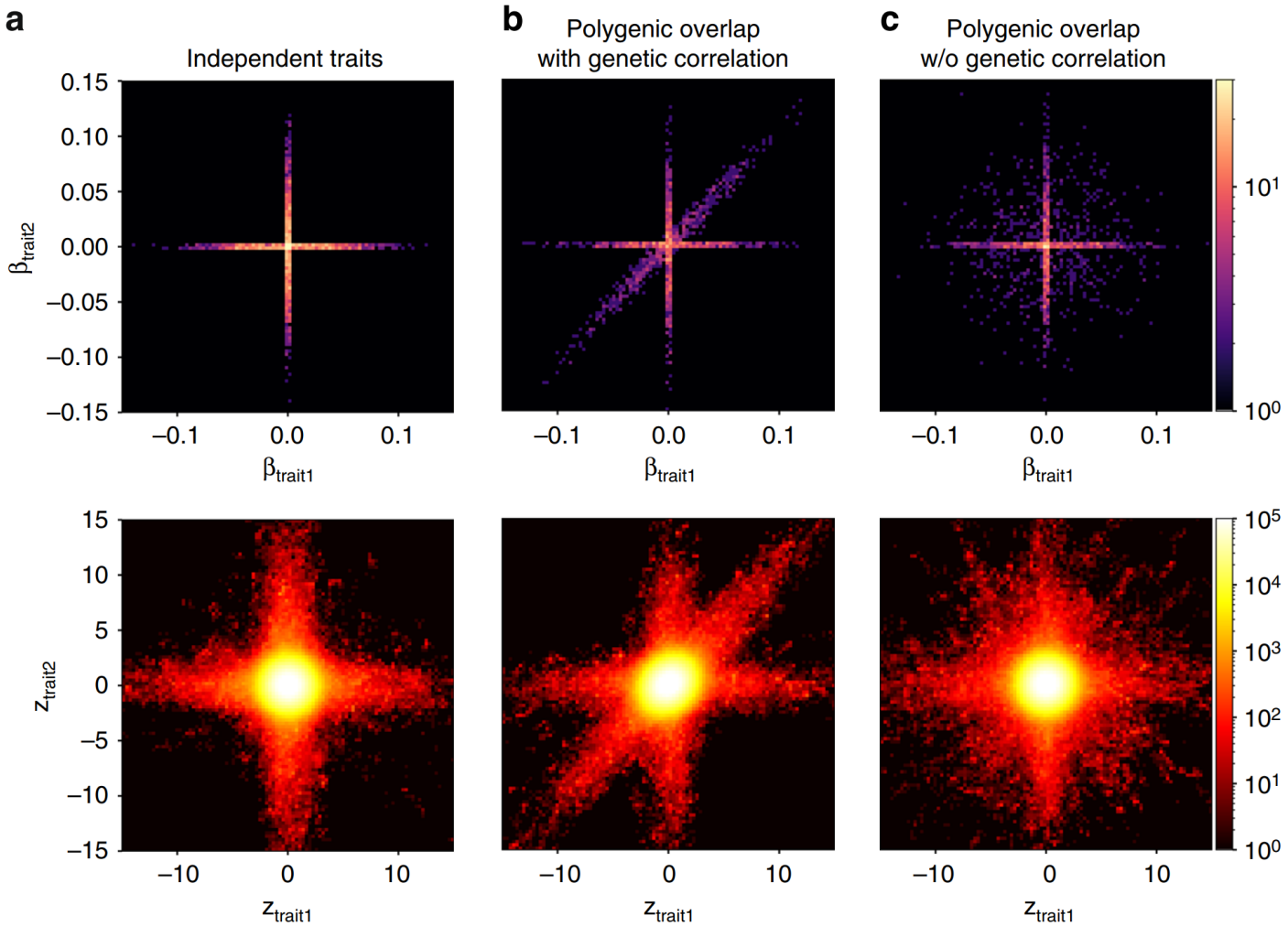
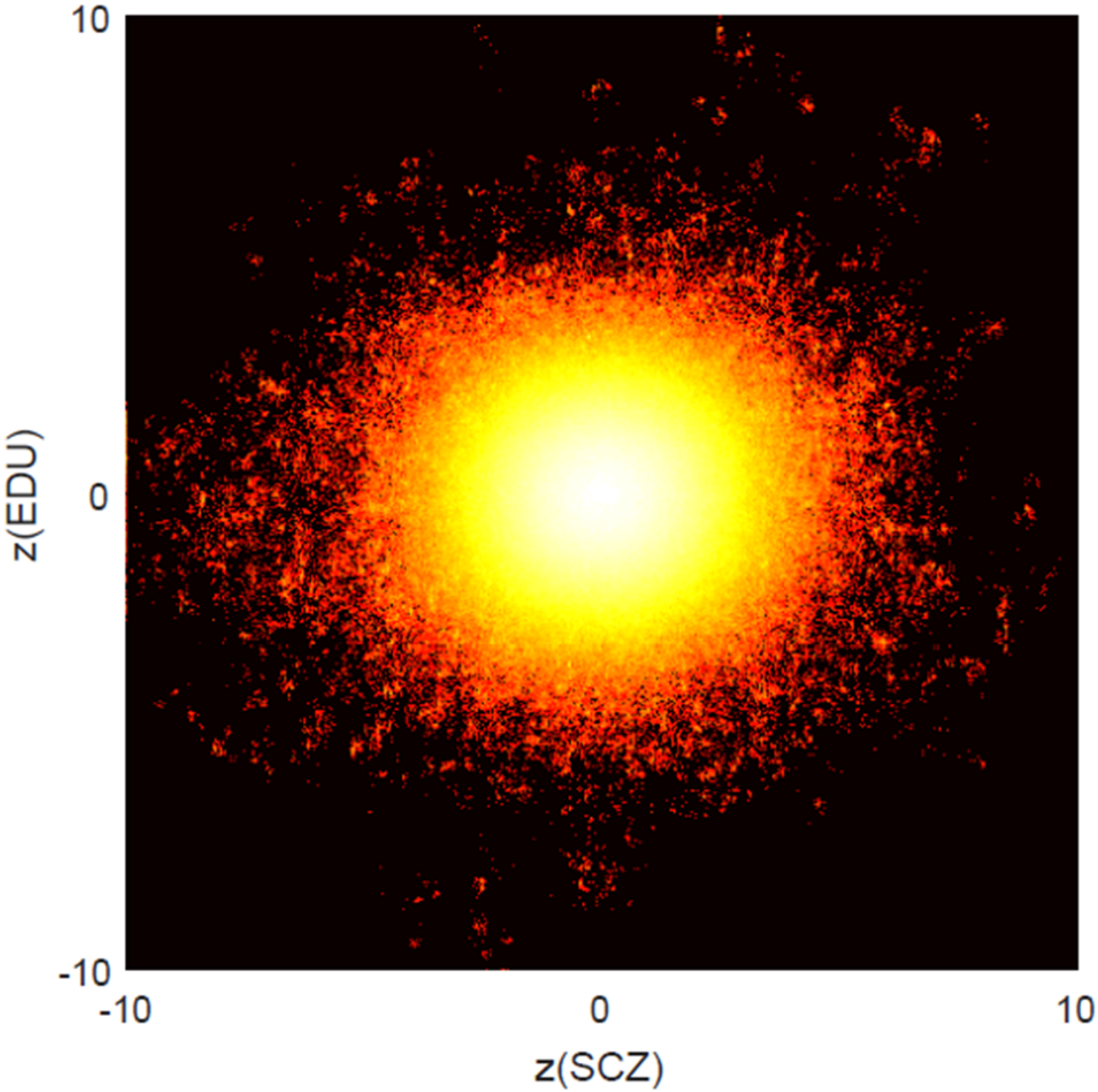


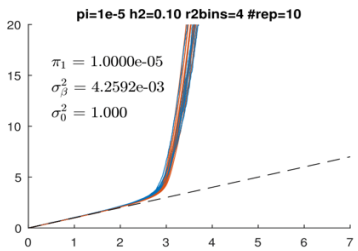
Fig. 3 Venn diagrams of unique and shared polygenic components

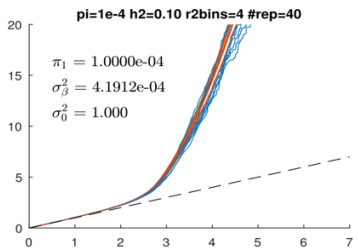
β vs z

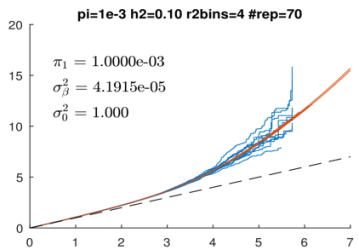

a
Independent traits
b
Polygenic overlap
with genetic correlation
c
Polygenic overlap
w/o genetic correlation
β trait2
β trait1
z trait2
z trait1
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.1
0.0
0.1
15
10
5
0
−5
−10
−15
−10
0
10
10¹
10⁰
10⁵
10⁴
10³
10²
10¹
10⁰

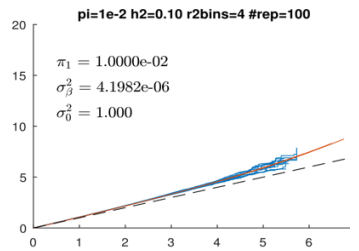

10
0
-10
z(EDU)
-10
0
10
z(SCZ)

Quantile-quantile plots

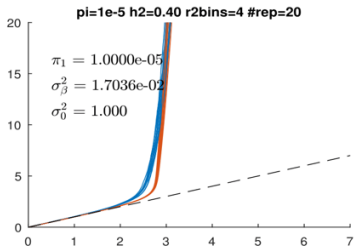


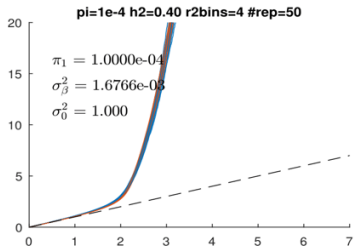


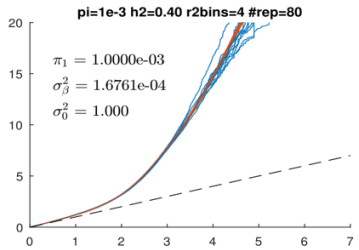


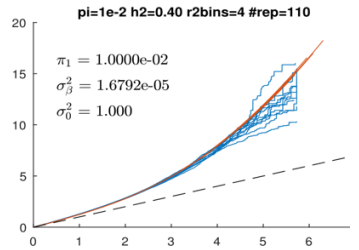


h^2=10%

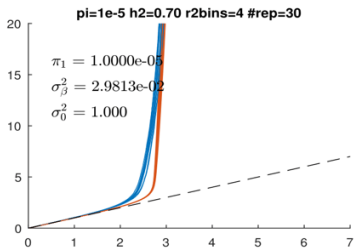


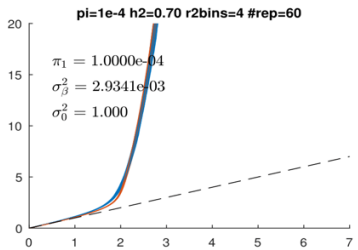


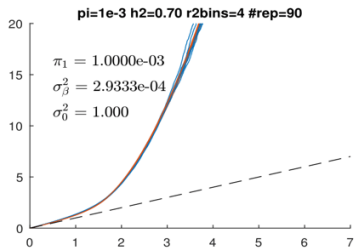


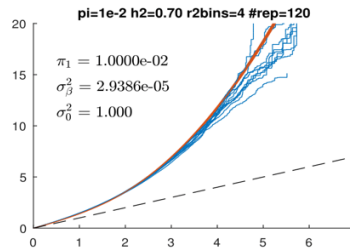


h^2=40%









h^2=70%

1 out of 100000 | 1 out of 10000 | 1 out of 1000 | 1 out of 100

Discussion

- Is there a better prior? (heavy tails, closed-form linear combinations)
- Can we model dependencies between β_{i1} and β_{i2}?
- Can we do posterior $E(\beta_i|\mathbf{z})$? Involve A^+ (pseudoinverse)
- Mixed-effect models, AI (average information), ...

$$y = G\beta + W\mathbf{x} + e$$
$$z = A\beta + B\mathbf{b} + \epsilon$$
$$\beta - \text{ random}, \mathbf{x} \text{ and } \mathbf{b} - \text{ fixed}$$

- Optimization strategy (differential evolution, non-zero OLS, Nedler-Mead)
- Computational issues: C++/Python, OpenMP, numeirc integration, sparse CRS, small integers compression, AVX2
- Better prediction? ($\hat{y} = G\hat{\beta}$)

MiXeR prior distribution (extended)

PLSA-MiXeR:

$$\beta_i \sim \pi_0 N(0, \sigma^2_{inf}\sigma^2_i) + \pi_1 N(0, \sigma^2_\beta \sigma^2_i) + \pi_2 N(0, \sigma^2_\gamma \sigma^2_i)$$

$$\sigma^2_i = \Big(\sum_{t=1\cdots T} [i \in C_t]\sigma^2_{a,t} \Big) H_i^S L_i^\ell,$$

where $t = 1 \cdots T$ runs across annotations

Cross-trait MiXeR:

$$(\beta_{1i}, \beta_{2i}) \sim \pi_{0i}\mathcal{N}(0,0) + \pi_{1i}\mathcal{N}(0, \mathbf{\Sigma}_{1i}) + \pi_{2i}\mathcal{N}(0, \mathbf{\Sigma}_{2i}) + \pi_{12i}\mathcal{N}(0, \mathbf{\Sigma}_{12i}),$$

$$\mathbf{\Sigma}_{1i} = \begin{bmatrix} \sigma^2_{1i} & 0 \\ 0 & 0 \end{bmatrix}, \mathbf{\Sigma}_{2i} = \begin{bmatrix} 0 & 0 \\ 0 & \sigma^2_{2i} \end{bmatrix}, \text{ and } \mathbf{\Sigma}_{12i} = \begin{bmatrix} \sigma^2_{1i} & \rho_{12i}\sigma_{1i}\sigma_{2i} \\ \rho_{12i}\sigma_{1i}\sigma_{2i} & \sigma^2_{2i} \end{bmatrix}$$

Acknowledgements