Finemap-MiXeR: A variational Bayesian approach for genetic finemapping

Bayram Akdeniz, Oleksandr Frei

March 3, 2023

Table of contents

1. Introduction (Alex)
   - Genome-wide association studies (GWAS)
   - Simple additive genetic model
   - MiXeR prior (spike-and-slab)
2. Finemap-MiXeR model (Bayram)
   - What is "finemapping"? Existing tools?
   - Finemap-MiXeR uses Adam to optimize ELBO
   - Results in simulations and with real data (height)
3. Beyond finemapping: challenges in statistical genetics (Alex)
   - Other techniques solving MiXeR prior
   - GSA-MiXeR: gene set heritability enrichment analysis
   - Discussion, conclusions and useful links

GWAS from Machine Learning Perspective

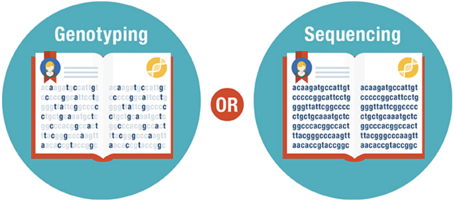
Input: Genotype matrix G (N subjects, M genetic "SNPs");
Input: Binary or continuous target variable (y):

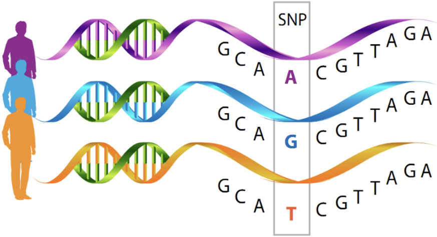
Matrix G	SNP_1	SNP_2	...	SNP_M	Class y
$Subject_1$	1	1	...	0	1
$Subject_2$	0	2	...	1	0
$Subject_3$	1	0	...	2	1
...	...	...	...	...	...
$Subject_N$	2	1	...	1	0

Output: SNPs associated with disease

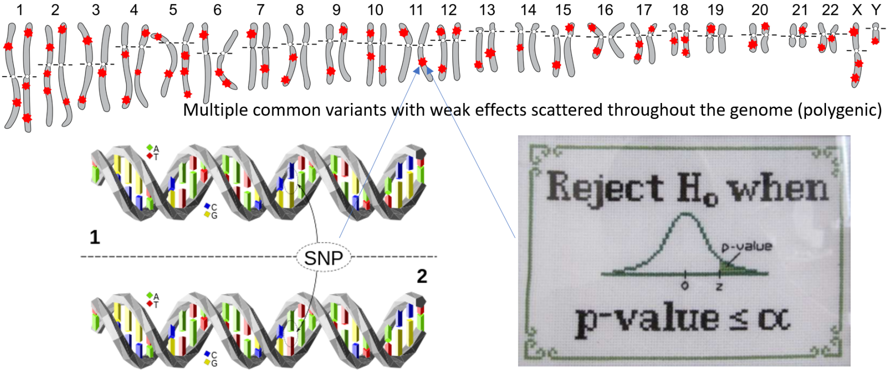
- 1-order: $\{SNP_1\}, \{SNP_2\}, \{SNP_3\}, \cdots$
- ~~2-order: $\{SNP_1, SNP_2\}, \ldots$~~
- ~~3-order: $\{SNP_1, SNP_2, SNP_3\}, \ldots$~~

Genotyping vs sequencing technologies



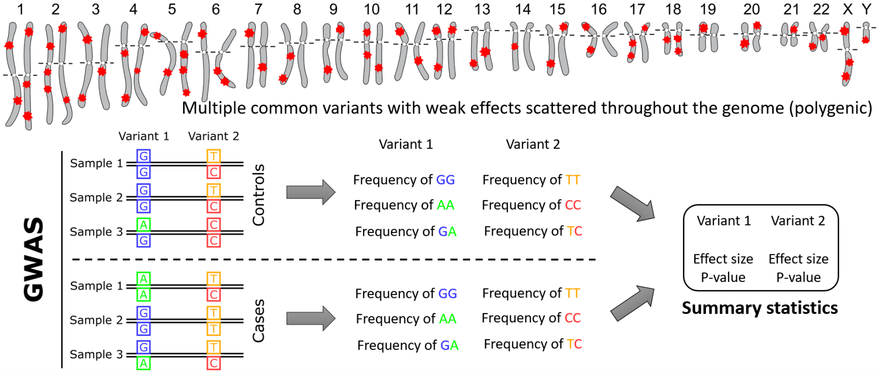


Genetics of complex human traits



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

Case/control genome-wide association study



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

GWAS

Controls

	Variant 1	Variant 2
Sample 1	G G	T C
Sample 2	G G	T C
Sample 3	A G	C C

Variant 1	Variant 2
Frequency of GG	Frequency of TT
Frequency of AA	Frequency of CC
Frequency of GA	Frequency of TC

Cases

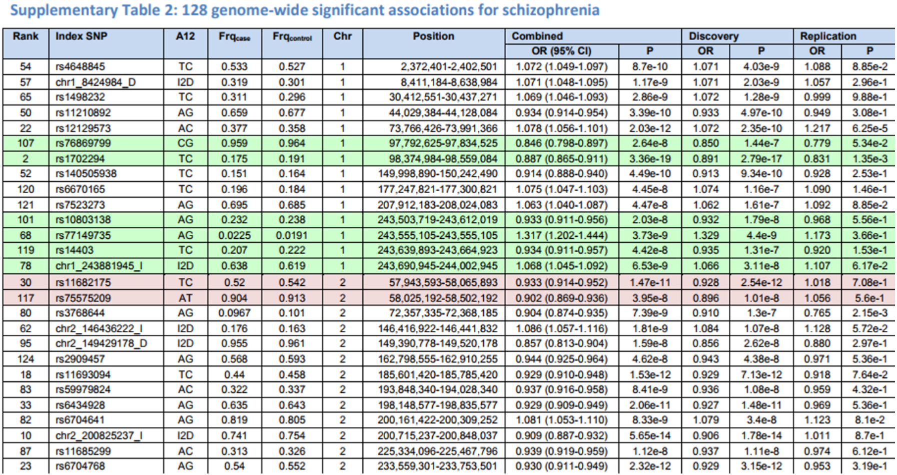
	Variant 1	Variant 2
Sample 1	A A	T C
Sample 2	G G	T T
Sample 3	G A	T C

Variant 1	Variant 2
Frequency of GG	Frequency of TT
Frequency of AA	Frequency of CC
Frequency of GA	Frequency of TC

Variant 1	Variant 2
Effect size P-value	Effect size P-value

Summary statistics

Psychiatric Genomic Consortium GWAS on schizophrenia



Supplementary Table 2: 128 genome-wide significant associations for schizophrenia

Rank	Index SNP	A12	Frq_{case}	$Frq_{control}$	Chr	Position	Combined		Discovery		Replication	
							OR (95% CI)	P	OR	P	OR	P
54	rs4648845	TC	0.533	0.527	1	2,372,401-2,402,501	1.072 (1.049-1.097)	8.7e-10	1.071	4.03e-9	1.088	8.85e-2
57	chr1_8424984_D	I2D	0.319	0.301	1	8,411,184-8,638,984	1.071 (1.048-1.095)	1.17e-9	1.071	2.03e-9	1.057	2.96e-1
65	rs1498232	TC	0.311	0.296	1	30,412,551-30,437,271	1.069 (1.046-1.093)	2.86e-9	1.072	1.28e-9	0.999	9.88e-1
50	rs11210892	AG	0.659	0.677	1	44,029,384-44,128,084	0.934 (0.914-0.954)	3.39e-10	0.933	4.97e-10	0.949	3.08e-1
22	rs12129573	AC	0.377	0.358	1	73,766,426-73,991,366	1.078 (1.056-1.101)	2.03e-12	1.072	2.35e-10	1.217	6.25e-5
107	rs76869799	CG	0.959	0.964	1	97,792,625-97,834,525	0.846 (0.798-0.897)	2.64e-8	0.850	1.44e-7	0.779	5.34e-2
2	rs1702294	TC	0.175	0.191	1	98,374,984-98,559,084	0.887 (0.865-0.911)	3.36e-19	0.891	2.79e-17	0.831	1.35e-3
52	rs140505938	TC	0.151	0.164	1	149,998,890-150,242,490	0.914 (0.888-0.940)	4.49e-10	0.913	9.34e-10	0.928	2.53e-1
120	rs6670165	TC	0.196	0.184	1	177,247,821-177,300,821	1.075 (1.047-1.103)	4.45e-8	1.074	1.16e-7	1.090	1.46e-1
121	rs7523273	AG	0.695	0.685	1	207,912,183-208,024,083	1.063 (1.040-1.087)	4.47e-8	1.062	1.61e-7	1.092	8.85e-2
101	rs10803138	AG	0.232	0.238	1	243,503,719-243,612,019	0.933 (0.911-0.956)	2.03e-8	0.932	1.79e-8	0.968	5.56e-1
68	rs77149735	AG	0.0225	0.0191	1	243,555,105-243,555,105	1.317 (1.202-1.444)	3.73e-9	1.329	4.4e-9	1.173	3.66e-1
119	rs14403	TC	0.207	0.222	1	243,639,893-243,664,923	0.934 (0.911-0.957)	4.42e-8	0.935	1.31e-7	0.920	1.53e-1
78	chr1_243881945_I	I2D	0.638	0.619	1	243,690,945-244,002,945	1.068 (1.045-1.092)	6.53e-9	1.066	3.11e-8	1.107	6.17e-2
30	rs11682175	TC	0.52	0.542	2	57,943,593-58,065,893	0.933 (0.914-0.952)	1.47e-11	0.928	2.54e-12	1.018	7.08e-1
117	rs75575209	AT	0.904	0.913	2	58,025,192-58,502,192	0.902 (0.869-0.936)	3.95e-8	0.896	1.01e-8	1.056	5.6e-1
80	rs3768644	AG	0.0967	0.101	2	72,357,335-72,368,185	0.904 (0.874-0.935)	7.39e-9	0.910	1.3e-7	0.765	2.15e-3
62	chr2_146436222_I	I2D	0.176	0.163	2	146,416,922-146,441,832	1.086 (1.057-1.116)	1.81e-9	1.084	1.07e-8	1.128	5.72e-2
95	chr2_149429178_D	I2D	0.955	0.961	2	149,390,778-149,520,178	0.857 (0.813-0.904)	1.59e-8	0.856	2.62e-8	0.880	2.97e-1
124	rs2909457	AG	0.568	0.593	2	162,798,555-162,910,255	0.944 (0.925-0.964)	4.62e-8	0.943	4.38e-8	0.971	5.36e-1
18	rs11693094	TC	0.44	0.458	2	185,601,420-185,785,420	0.929 (0.910-0.948)	1.53e-12	0.929	7.13e-12	0.918	7.64e-2
83	rs59979824	AC	0.322	0.337	2	193,848,340-194,028,340	0.937 (0.916-0.958)	8.41e-9	0.936	1.08e-8	0.959	4.32e-1
33	rs6434928	AG	0.635	0.643	2	198,148,577-198,835,577	0.929 (0.909-0.949)	2.06e-11	0.927	1.48e-11	0.969	5.36e-1
82	rs6704641	AG	0.819	0.805	2	200,161,422-200,309,252	1.081 (1.053-1.110)	8.33e-9	1.079	3.4e-8	1.123	8.1e-2
10	chr2_200825237_I	I2D	0.741	0.754	2	200,715,237-200,848,037	0.909 (0.887-0.932)	5.65e-14	0.906	1.78e-14	1.011	8.7e-1
87	rs11685299	AC	0.313	0.326	2	225,334,096-225,467,796	0.939 (0.919-0.959)	1.12e-8	0.937	1.11e-8	0.974	6.12e-1
23	rs6704768	AG	0.54	0.552	2	233,559,301-233,753,501	0.930 (0.911-0.949)	2.32e-12	0.929	3.15e-12	0.953	3.19e-1

https://pgc.unc.edu/for-researchers/download-results/

Psychiatric Genomic Consortium GWAS on schizophrenia

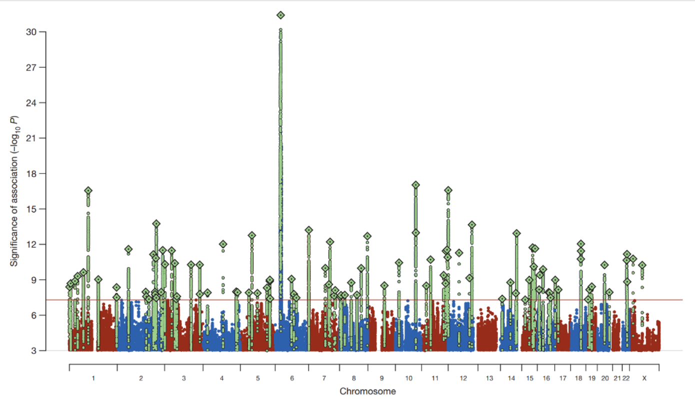
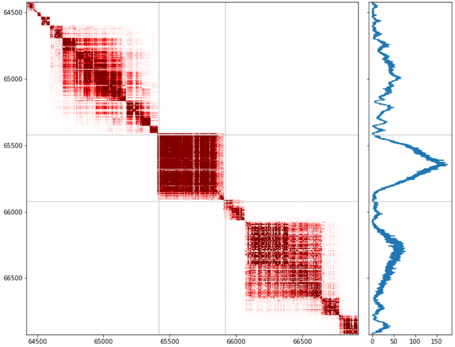


Figure 1 | Manhattan plot showing schizophrenia associations. Manhattan plot of the discovery genome-wide association meta-analysis of 49 case control samples (34,241 cases and 45,604 controls) and 3 family based association studies (1,235 parent affected-offspring trios). The *x* axis is chromosomal position and the *y* axis is the significance ($-\log_{10} P$; 2-tailed) of association derived by logistic regression. The red line shows the genome-wide significance level (5×10^{-8}). SNPs in green are in linkage disequilibrium with the index SNPs (diamonds) which represent independent genome-wide significant associations.

https://pgc.unc.edu/for-researchers/download-results/

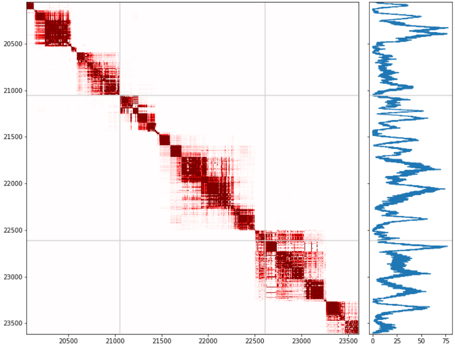
Linkage disequilibrium - SNP correlation matrix

Region around **PIGP** gene
Color: r_{ij}^2 - squared correlation between i-th and j-th SNP genotype
Histogram: total LD score $\ell_i = \sum_j r_{ij}^2$

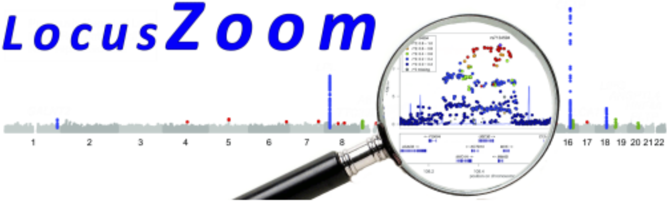


Linkage disequilibrium - SNP correlation matrix

Region around **NCAM2** gene
Color: r_{ij}^2 - squared correlation between i-th and j-th SNP genotype
Histogram: total LD score $\ell_i = \sum_j r_{ij}^2$



Finemapping: zooming into a locus



https://genome.sph.umich.edu/wiki/LocusZoom

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
- e - non-genetic effects
- $\mathbf{y}$, G - known; β, e - unknown

GWAS vs OLS

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

$$\hat{\beta}_{OLS} = (G'G)^{-1}G'\mathbf{y} \quad \textit{- naive implementation works too badly}$$

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

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

Regression from GWAS summary statistics

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 $(a_{ij} = \sigma_0\sqrt{N_j Var(g_i)}r_{ij}$, where $r_{ij} = corr(\mathbf{v}_i, \mathbf{v}_j))$
- β - as before
- $\epsilon \sim N(0, \sigma_0^2)$

MiXeR prior distribution on β

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

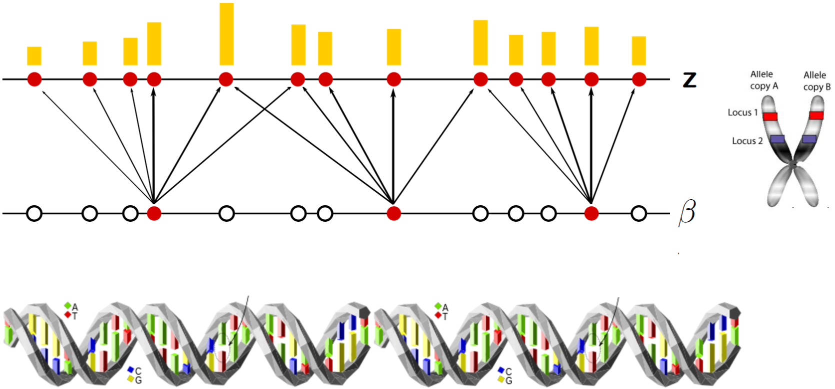
MiXeR:

$$\beta_i \sim (1 - \pi_1)\delta_0 + \pi_1 N(0, \sigma_\beta^2)$$

where

- π_1 - weight in the mixture
- σ_β^2 - variance
- δ_0 - probability mass at zero

Inferences about β using z as input



What is “finemapping”? Existing tools?

Finemap-MiXeR uses Adam to optimize ELBO

Results in simulations and with real data (height)

Sampling from $p(\vec{u}|\theta)$, analytical $p(z_j|\vec{\beta},\theta) \cdot p(\vec{\beta}|\vec{u},\theta)$

Let $\theta = \{\pi_1, \sigma_\beta^2, \sigma_0^2\}$ be the vector of parameters of MiXeR model. Let $u_i \in \{0,1\}$ be latent variable with $p(u_i) = Bern(u_i|\pi_1)$, then

$$p(z_j, \vec{\beta}, \vec{u}|\theta) = p(z_j|\vec{\beta},\theta) \cdot p(\vec{\beta}|\vec{u},\theta) \cdot p(\vec{u}|\theta)),$$

$$p(z_j|\beta_1,\dots,\beta_M,\theta) = N\left(z_j\middle|\sum_{i=1}^{M} a_{ij}\beta_i, \sigma_0^2\right),$$

$$p(\beta_i|u_i=0,\theta) = N(\beta_i|0,0), \quad p(\beta_i|u_i=1,\theta) = N(\beta_i|0,\sigma_i^2),$$

$$p(u_i|\theta) = Bern(u_i|\pi_1)$$

After observing $\vec{z} = (z_1,\dots,z_M)^T$, infer θ by max. likelihood:

$$p(\vec{z}|\theta) = \prod_j \int_u \int_\beta p(z_j,\vec{\beta},\vec{u},\theta)\,du\,d\beta \to \max_\theta$$

Sampling from prior distribution: let U_{jk} be the set of variants with $u_i = 1$. For each k the distribution over $p(z_j|U_{jk},\theta)$ is a normal zero-mean distribution:

$$p(z_j|U_{jk},\theta) = N(z_j|0,\Sigma_{jk}^2), \text{ where } \Sigma_{jk}^2 = \sigma_0^2 + \sum_{i\in U_{jk}} a_{ij}\sigma_i^2.$$

SuSiE and RSS: https://stephenslab.uchicago.edu/

- https://github.com/stephenslab/susieR/
  - sum of single effects model
- https://github.com/stephenslab/rss
  - regression with summary statistic, implements both MCMC and Variational Bayes approaches
- Peter Carbonetto, Matthew Stephens. "Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies." Bayesian Anal. 7 (1) 73 - 108, March 2012.
  https://doi.org/10.1214/12-BA703
- Xiang Zhu and Matthew Stephens (2017). Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. Annals of Applied Statistics 11(3): 1561-1592.
  https://doi.org/10.1214/17-aoas1046
- Wang G, Sarkar A, Carbonetto P, Stephens M. A simple new approach to variable selection in regression, with application to genetic fine mapping. Journal of the Royal Statistical Society, Series B. 2020; 82(5):12731300.
  https://doi.org/10.1111/rssb.12388
- Fine-mapping from summary data with the Sum of Single Effects model PLOS Genetics, July 19, 2022 Yuxin Zou,Peter Carbonetto,Gao Wang,Matthew Stephens
  https://doi.org/10.1371/journal.pgen.1010299

SuSiE - sum of single effects model: $(Bern \rightarrow Mult)$

$$\mathbf{y} = G\beta + \mathbf{e}$$

SuSiE model: $\beta = \sum_{\ell}^{L} \beta_l$, where each vector β_l is a single-effect vector, i.e. a vector with exactly one non-zero element:

$$\begin{aligned} \beta_l &= b_l \gamma_l, \\ \gamma_l &\sim \mathrm{Mult}(1, \pi), \\ b_l &\sim N(0, \sigma_\beta^2), \\ \pi &= (1/p, \ldots, 1/p) \end{aligned}$$

Posterior under single-effect regression model:

$$\gamma_l | G, y, \sigma_\beta^2 \sim Mult(1, \alpha_l)$$
$$b_l | G, y, \sigma_\beta^2 \sim N(\mu_l, \sigma_l^2)$$

OLS with regularization

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

$$\hat{\beta}_{OLS} = (G'G)^{-1}G'\mathbf{y} \quad \textit{- naive implementation works too badly}$$

- Stacked block ridge regression (regenie) `https://www.nature.com/articles/s41588-021-00870-7`
- Change axis to first principal components of the LD matrix $A = G'G$ as in `https://github.com/josefin-werme/LAVA`

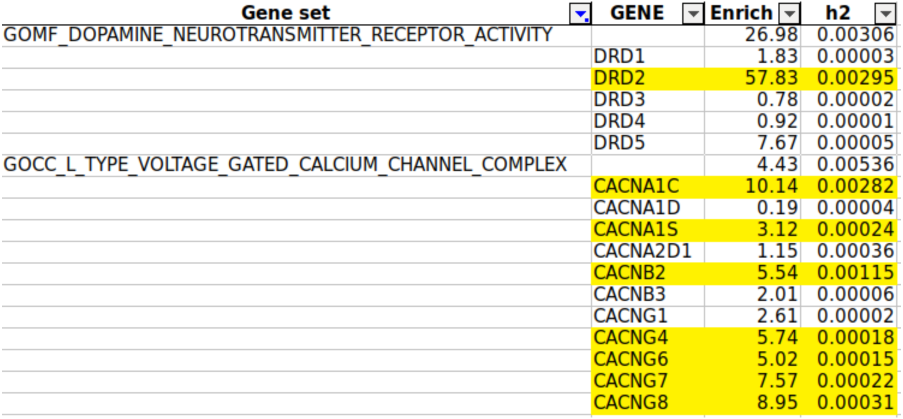
GSA-MiXeR: gene set heritability enrichment analysis

"Fine-map" at the level of genes rather than SNPs:

$$\beta_i \sim \pi_0 \delta_0 + \pi_1 N(0, \sigma^2_{g(i)}),$$

where $g(i)$ indicates the gene that i-th SNP belongs to; $\sigma^2_{g(i)}$ indicates its effect size variance. `https://www.medrxiv.org/content/10.1101/2022.12.08.22283159v1` - applies analysis to 45.000 genes, using Adam to optimize variance parameters.

GSA-MiXeR: gene set heritability enrichment analysis



Gene set	GENE	Enrich	h2
GOMF_DOPAMINE_NEUROTRANSMITTER_RECEPTOR_ACTIVITY		26.98	0.00306
	DRD1	1.83	0.00003
	DRD2	57.83	0.00295
	DRD3	0.78	0.00002
	DRD4	0.92	0.00001
	DRD5	7.67	0.00005
GOCC_L_TYPE_VOLTAGE_GATED_CALCIUM_CHANNEL_COMPLEX		4.43	0.00536
	CACNA1C	10.14	0.00282
	CACNA1D	0.19	0.00004
	CACNA1S	3.12	0.00024
	CACNA2D1	1.15	0.00036
	CACNB2	5.54	0.00115
	CACNB3	2.01	0.00006
	CACNG1	2.61	0.00002
	CACNG4	5.74	0.00018
	CACNG6	5.02	0.00015
	CACNG7	7.57	0.00022
	CACNG8	8.95	0.00031

Discussion (status update from Nov 2019)

- Is there a better alternative to $(1-\pi_1)\delta_0 + \pi_1 N(0, \sigma_\beta^2)$ prior, with closed-form linear combinations, but still heavy tails?
- How to model dependencies between β_{i1} and β_{i2} - partly covered with SuSiE model
- Can we do posterior $p(\beta_i|\mathbf{z})$? Yes, Finemap-MiXeR!
- Optimization strategy (differential evolution, non-zero OLS, Nedler-Mead). Now also Adam!
- Better prediction? $(\hat{y} = G\hat{\beta})$ - TBD

Conclusions

- Statistical genetics - large and active research area with many promising application of Bayesian inference
- Key challenges: extremely high number of genetic features each having a tiny effect on the outcome, and a high correlation among those features
- We presented Finemap-MiXeR - new technique for fine-mapping causal variants from GWAS summary statistics, using direct ELBO optimization with Adam
- We also found other applications to Adam (e.g. GSA-MiXeR)
- In the future we hope such models will improve our understanding of complex psychiatric disorders, including schizophrenia, and lead to better treatment alternatives

Useful links

- Our group runs a GWAS cources at UiO, March 2023, organized by my colleague Alexey Shadrin. Registration link: `https://nettskjema.no/a/325001`
- Previous year program: `https://www.med.uio.no/norment/forskning/aktuelt/arrangementer/andre/2022/genome-wide-association-studies-why-how-and-then-w.html`
- Some practical "hands on" exercises: `https://github.com/ofrei/gwas101`
- AI@MIPT: Using big data for mathematical models of the human genome implications for psychiatric genetics (Kevin O'Connell, Oleksandr Frei) - `https://vk.com/aimipt?z=video-932_456239307`