Polygenic Prediction using GWAS Summary Statistics

SISG – Module 15

Dr Loic Yengo

l.yengo@imb.uq.edu.au

[Slides credit: Dr Jian Zeng (UQ) & Prof. Naomi Wray (UQ)]

Institute for Molecular Bioscience

The University of Queensland

The aim of this lecture is to introduce a few general classes of methods for genomic prediction applicable when individual-level data is unavailable.

Linear Predictor = $\widehat{\beta_1} x_{i1} + \widehat{\beta_2} x_{i2} + \widehat{\beta_3} x_{i3} + \cdots = \sum_{j=1}^{n_{SNP}} \widehat{\beta_j} x_{ij}$

0, 1 or 2 alleles

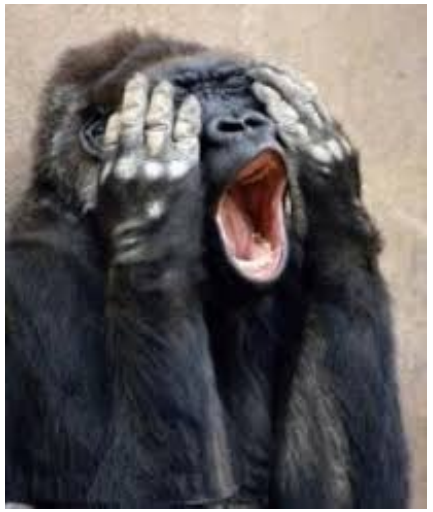
Which SNPs?

What weights?

How should we name those predictors?

Among Human geneticists

- **PRS** - Polygenic risk score
- **PGS** -Polygenic score
- **GRS** - Genetic risk score
- **Genetic score**
- **Genotypic score**
- **Allele score**
- **Profile score**
- …



Among Animal Breeders…

- **GEBV**: Genomic Estimated Breeding Value

In this lecture: I'll use **PGS** or **Linear Predictor** as generic term

Credit: Prof. N. Wray

Two families of methods…

1) Clumping + P-value Thresholding (GWAS-derived)

2) Summary Statistics-based Approximation of Individual-level data methods

CP+T Method (nice and easy*)

Input
GWAS summary statistics
Validation sample (genotypes + phenotypes)

Algorithm
1) Ascertain independent SNPs (**C**lumping) associated with the trait at a certain **P**-value threshold (**T**hresholding)

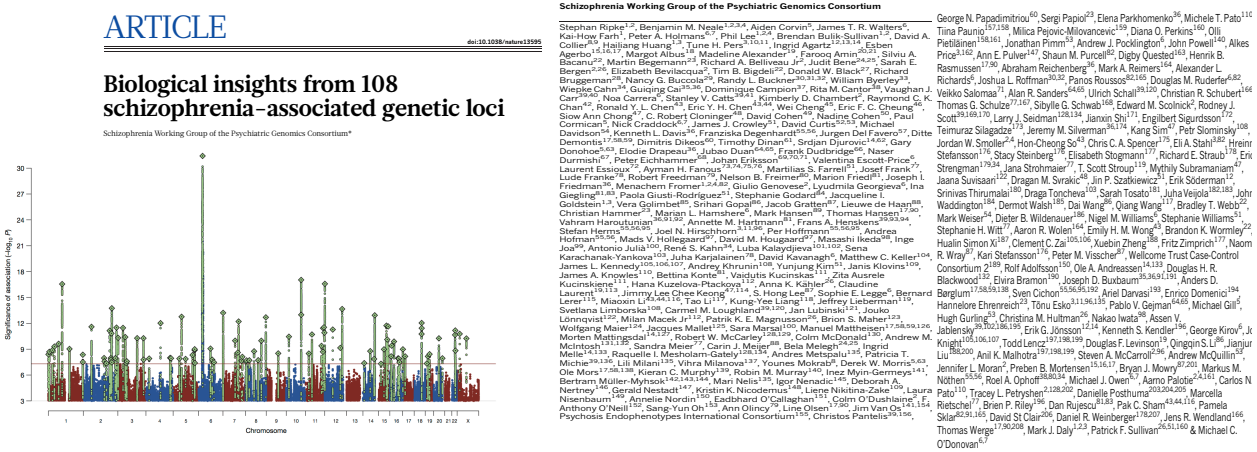
2) Calculate PGS for each set

3) Select the set with largest prediction accuracy in the validation sample

We'll use this method in the practical.

Prediction using whole-genome variation

- Best linear unbiased prediction (GBLUP)
- Bayesian alphabet methods (BayesA, B, C, R etc)
- Require individual-level genotype and phenotype data.
- Computationally demanding when # individuals and SNPs are large.
- Data privacy and ethical considerations.



ARTICLE

doi:10.1038/nature13595

Biological insights from 108 schizophrenia-associated genetic loci

Schizophrenia Working Group of the Psychiatric Genomics Consortium*

Schizophrenia Working Group of the Psychiatric Genomics Consortium

Stephan Ripke[1-3], Benjamin M. Neale[1-3,4], Aiden Corvin[5], James T. R. Walters[6], Kai-How Farh[1], Peter A. Holmans[6,7], Phil Lee[1,3,4], Brendan Bulik-Sullivan[1,3], David A. Collier[8,9], Haijiang Huang[1,3], Tune H. Pers[2,10,11], Ingrid Agartz[12,13,14], Esben Agerbo[15,16,17], Margot Albus[18], Madeline Alexander[19], Farooq Amin[20,21], Silviu A. Bacanu[22], Martin Begemann[23], Richard A. Belliveau Jr.[1], Judit Bene[24,25], Sarah E. Bergen[2,26], Elizabeth Bevilacqua[3], Tim B. Bigdeli[22], Donald W. Black[27], Richard Bruggeman[28], Nancy G. Buccola[29], Randy L. Buckner[30,31,32], William Byerley[33], Wiepke Cahn[34], Guiqing Cai[35,36], Dominique Campion[37], Rita M. Cantor[38], Vaughan J. Carr[39,40], Noa Carrera[6], Stanley V. Catts[39,41], Kimberly D. Chambert[3], Raymond C. K. Chan[42], Ronald Y. L. Chen[43], Eric Y. H. Chen[44,45], Wei Cheng[46], Eric F. C. Cheung[46], Siow Ann Chong[47], C. Robert Cloninger[48], David Cohen[49], Nadine Cohen[50], Paul Cormican[5], Nick Craddock[6,7], James J. Crowley[51], David Curtis[52,53], Michael Davidson[54], Kenneth L. Davis[55], Franziska Degenhardt[56,57], Jurgen Del Favero[57], Ditte Demontis[17,58,59], Dimitris Dikeos[60], Timothy Dinan[61], Srdjan Djurovic[14,62], Gary Donohoe[5,63], Elodie Drapeau[36], Jubao Duan[64,65], Frank Dudbridge[66], Naser Durmishi[67], Peter Eichhammer[68], Johan Eriksson[69,70,71], Valentina Escott-Price[6], Laurent Essioux[72], Ayman H. Fanous[73,74,75,76], Martilias S. Farrell[51], Josef Frank[77], Lude Franke[78], Robert Freedman[79], Nelson B. Freimer[80], Marion Friedl[81], Joseph I. Friedman[36], Menachem Fromer[1,2,4,82], Giulio Genovese[3], Lyudmila Georgieva[6], Ina Giegling[83,84], Paola Giusti-Rodriguez[51], Stephanie Godard[64], Jacqueline I. Goldstein[1,3], Vera Golimbet[85], Srihari Gopal[86], Jacob Gratten[87], Lieuwe de Haan[88], Christian Hammer[89], Marian L. Hamshere[6], Mark Hansen[69], Thomas Hansen[17,90], Vahram Haroutunian[36,91,92], Annette M. Hartmann[83], Frans A. Henskens[93,93,94], Stefan Herms[95,96,97], Joel N. Hirschhorn[3,11,98], Per Hoffmann[95,96,97], Andrea Hofman[95,96], Mads V. Hollegaard[99], David M. Hougaard[99], Masashi Ikeda[98], Inge Joa[99], Antonio Julià[100], René S. Kahn[34], Luba Kalaydjieva[101,102], Sena Karachanak-Yankova[103], Juha Karjalainen[78], David Kavanagh[6], Matthew C. Keller[104], James L. Kennedy[105,106,107], Andrey Khrunin[108], Yunjung Kim[51], Janis Klovins[109], James A. Knowles[110], Bettina Konte[83], Vaidutis Kucinskas[111], Zita Ausrele Kucinskiene[111], Hana Kuzelova-Ptackova[112], Anna K. Kähler[69], Claudine Laurent[19,113], Jimmy Lee Chee Keong[47,114], S. Hong Lee[22], Sophie E. Legge[6], Bernard Lerer[115], Miaoxin Li[42,44,116], Tao Li[117], Kung-Yee Liang[118], Jeffrey Lieberman[119], Svetlana Limborska[108], Carmel M. Loughland[39,120], Jan Lubinski[121], Jouko Lönnqvist[122], Milan Macek Jr[112], Patrik K. E. Magnusson[26], Brion S. Maher[123], Wolfgang Maier[124], Jacques Mallet[125], Sara Marsal[100], Manuel Mattheisen[17,58,59,126], Morten Mattingsdal[14,127], Robert W. McCarley[128,129], Colm McDonald[130], Andrew M. McIntosh[131,132], Sandra Meier[77], Carin J. Meijer[88], Bela Melegh[24,25], Ingrid Melle[14,133], Raquelle I. Mesholam-Gately[128,134], Andres Metspalu[135], Patricia T. Michie[39,136], Lili Milani[135], Vihra Milanova[137], Younes Mokrab[6], Derek W. Morris[5,62], Ole Mors[17,58,138], Kieran C. Murphy[139], Robin M. Murray[140], Inez Myin-Germeys[141], Bertram Müller-Myhsok[142,143,144], Mari Nelis[135], Igor Nenadic[145], Deborah A. Nertney[146], Gerald Nestadt[147], Kristin K. Nicodemus[148], Liene Nikitina-Zake[108], Laura Nisenbaum[149], Annelie Nordin[150], Eadbhard O'Callaghan[151], Colm O'Dushlaine[3], F. Anthony O'Neill[152], Sang-Yun Oh[153], Ann Olincy[79], Line Olsen[17,90], Jim Van Os[141,154], Psychosis Endophenotypes International Consortium[155], Christos Pantelis[39,156], George N. Papadimitriou[60], Sergi Papiol[23], Elena Parkhomenko[36], Michele T. Pato[11C], Tiina Paunio[157,158], Milica Pejovic-Milovancevic[159], Diana O. Perkins[160], Olli Pietiläinen[158,161], Jonathan Pimm[53], Andrew J. Pocklington[6], John Powell[140], Alkes Price[3,162], Ann E. Pulver[147], Shaun M. Purcell[82], Digby Quested[163], Henrik B. Rasmussen[17,90], Abraham Reichenberg[36], Mark A. Reimers[164], Alexander L. Richards[6], Joshua L. Roffman[30,32], Panos Roussos[82,165], Douglas M. Ruderfer[4,82], Veikko Salomaa[71], Alan R. Sanders[64,65], Ulrich Schall[39,120], Christian R. Schubert[166], Thomas G. Schulze[77,167], Sibylle G. Schwab[168], Edward M. Scolnick[2], Rodney J. Scott[39,169,170], Larry J. Seidman[128,134], Jianxin Shi[171], Engilbert Sigurdsson[172], Teimuraz Silagadze[173], Jeremy M. Silverman[36,174], Kang Sim[47], Petr Slominsky[108], Jordan W. Smoller[2,4], Hon-Cheong So[43], Chris C. A. Spencer[175], Eli A. Stahl[3,82], Hreinn Stefansson[176], Stacy Steinberg[176], Elisabeth Stogmann[177], Richard E. Straub[178], Eric Strengman[179,94], Jana Strohmaier[77], T. Scott Stroup[119], Mythily Subramaniam[47], Jaana Suvisaari[122], Dragan M. Svrakic[48], Jin P. Szatkiewicz[51], Erik Söderman[179], Srinivas Thirumalai[180], Draga Toncheva[103], Sarah Tosato[181], Juha Veijola[182,183], John Waddington[184], Dermot Walsh[185], Dai Wang[66], Qiang Wang[117], Bradley T. Webb[22], Mark Weiser[54], Dieter B. Wildenauer[186], Nigel M. Williams[6], Stephanie Williams[51], Stephanie H. Witt[77], Aaron R. Wolen[104], Emily H. M. Wong[22], Brandon K. Wormley[22], Hualin Simon Xi[187], Clement C. Zai[105,106], Xuebin Zheng[188], Fritz Zimprich[177], Naom R. Wray[87], Kari Stefansson[176], Peter M. Visscher[87], Wellcome Trust Case-Control Consortium 2[189], Rolf Adolfsson[150], Ole A. Andreassen[14,133], Douglas H. R. Blackwood[132], Elvira Bramon[190], Joseph D. Buxbaum[35,36,91,191], Anders D. Børglum[17,58,59,138], Sven Cichon[10,56,95,192], Ariel Darvasi[193], Enrico Domenici[194], Hannelore Ehrenreich[23], Tõnu Esko[3,11,96,135], Pablo V. Gejman[64,65], Michael Gill[5], Hugh Gurling[63], Christina M. Hultman[26], Nakao Iwata[98], Assen V. Jablensky[39,120,136,195], Erik G. Jönsson[12,14], Kenneth S. Kendler[196], George Kirov[6], Jo Knight[105,106,107], Todd Lencz[79,136,197], Douglas F. Levinson[73], Qingqin S. Li[86], Jianjun Liu[198,200], Anil K. Malhotra[197,198,199], Steven A. McCarroll[2,96], Andrew McQuillin[132], Jennifer L. Moran[2], Preben B. Mortensen[15,16,17], Bryan J. Mowry[87,201], Markus M. Nöthen[95,96], Roel A. Ophoff[88,80,34], Michael J. Owen[6,7], Aarno Palotie[2,4,161], Carlos N. Pato[110], Tracy L. Petryshen[2,128,202], Danielle Posthuma[203,204,205], Marcella Rietschel[77], Brien P. Riley[196], Dan Rujescu[83,83], Pak C. Sham[43,44,116], Pamela Sklar[82,91,165], David St Clair[206], Daniel R. Weinberger[178,207], Jens R. Wendland[166], Thomas Werge[17,90,208], Mark J. Daly[1,2,3], Patrick F. Sullivan[26,51,160] & Michael C. O'Donovan[6,7]

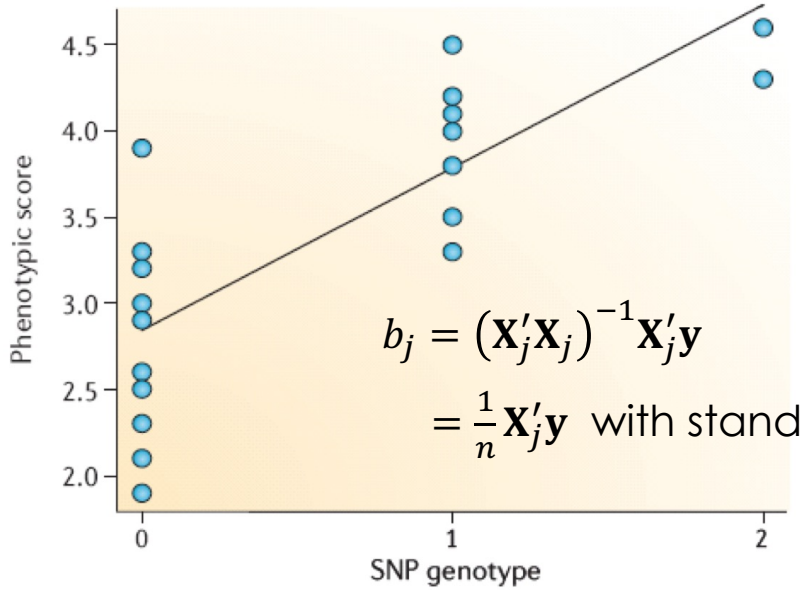
Results generated from a meta-analysis of 52 individual GWAS datasets.

GWAS summary statistics (summary-level data)

- Marginal SNP effects
- Standard errors
- (Per-SNP) sample sizes

- Effect and other alleles (A1 and A2)
- Effect allele frequencies

- But we need more information…



$$b_j = (\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j'\mathbf{y}$$
$$= \frac{1}{n}\mathbf{X}_j'\mathbf{y} \text{ with standardised } \mathbf{X}_j$$

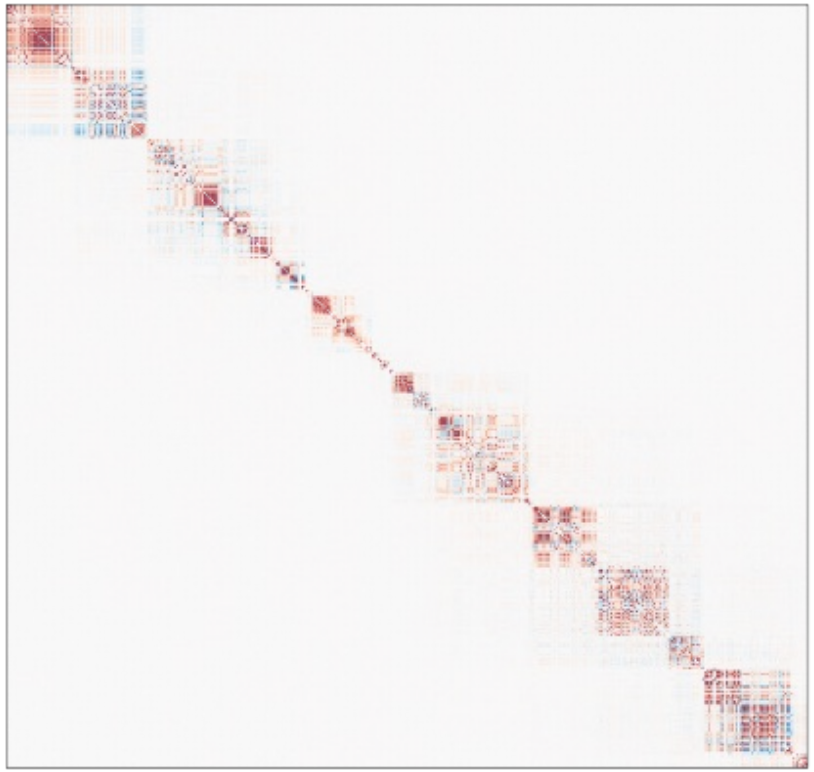
Linkage disequilibrium (LD) data

- Usually obtained from a reference population
- LD correlation matrix

$$\mathbf{R} = \frac{1}{n}\mathbf{X}'\mathbf{X}$$

with standardised $\mathbf{X}$

($E[X_j] = 0$ and $\text{var}[X_j] = 1$)



From individual-level to summary-level model (*ideal conditions*)

Consider an individual-data model with a standardized genotype matrix **X:**

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

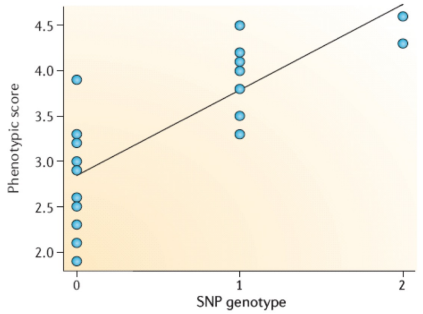
Multiply both sides by $\frac{1}{n}\mathbf{X}'$ gives

$$\frac{1}{n}\mathbf{X}'\mathbf{y} = \frac{1}{n}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \frac{1}{n}\mathbf{X}'\mathbf{e}$$

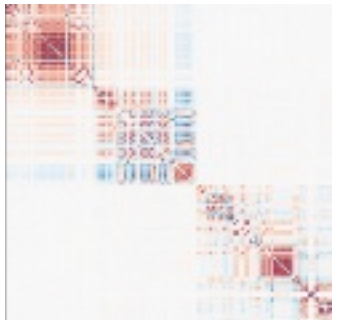
$$\mathbf{b} = \mathbf{R}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

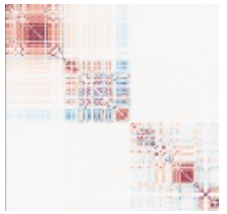
$$Var(\boldsymbol{\epsilon}) = \frac{1}{n}\mathbf{R}\sigma_e^2$$

GWAS marginal SNP effects



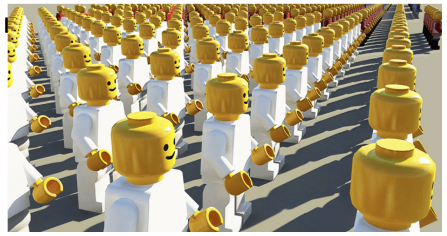
LD correlation matrix

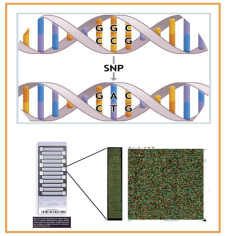




Individual-level analysis

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$





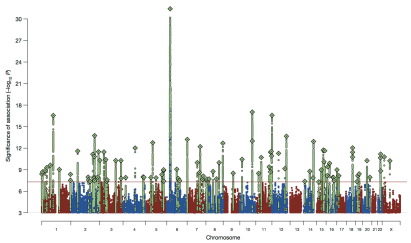
BLUP*

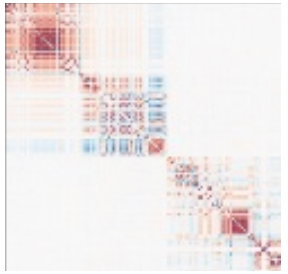
Bayes

→

Summary-level analysis

$$\mathbf{b} = \mathbf{R}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$





SBLUP

SBayes

*Best Linear Unbiased Predictor

SNP-BLUP vs. SBLUP

SNP-BLUP

- $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$
- $Var(\mathbf{e}) = \mathbf{I}\sigma_e^2$
- Mixed model equations (MME):

$$[\mathbf{X}'\mathbf{X} + \mathbf{I}\lambda]\widehat{\boldsymbol{\beta}} = \mathbf{X}'\boldsymbol{y}$$

- $\lambda = \frac{\sigma_e^2}{\sigma_\beta^2}$
- $\widehat{\boldsymbol{\beta}} = [\mathbf{X}'\mathbf{X} + \mathbf{I}\lambda]^{-1}\mathbf{X}'\boldsymbol{y}$

Genotype matrix

Phenotypes

SBLUP

- $\mathbf{b} = \mathbf{R}\boldsymbol{\beta} + \boldsymbol{\epsilon}$
- $Var(\boldsymbol{\epsilon}) = \frac{1}{n}\mathbf{R}\sigma_e^2$
- Mixed model equations (MME):

$$[\mathbf{R}'n\mathbf{R}^{-1}\mathbf{R} + \mathbf{I}\lambda]\widehat{\boldsymbol{\beta}} = n\mathbf{R}'\mathbf{R}^{-1}\mathbf{b}$$

$$[n\mathbf{R} + \mathbf{I}\lambda]\widehat{\boldsymbol{\beta}} = n\mathbf{b}$$

- $\widehat{\boldsymbol{\beta}} = [n\mathbf{R} + \mathbf{I}\lambda]^{-1}n\mathbf{b}$

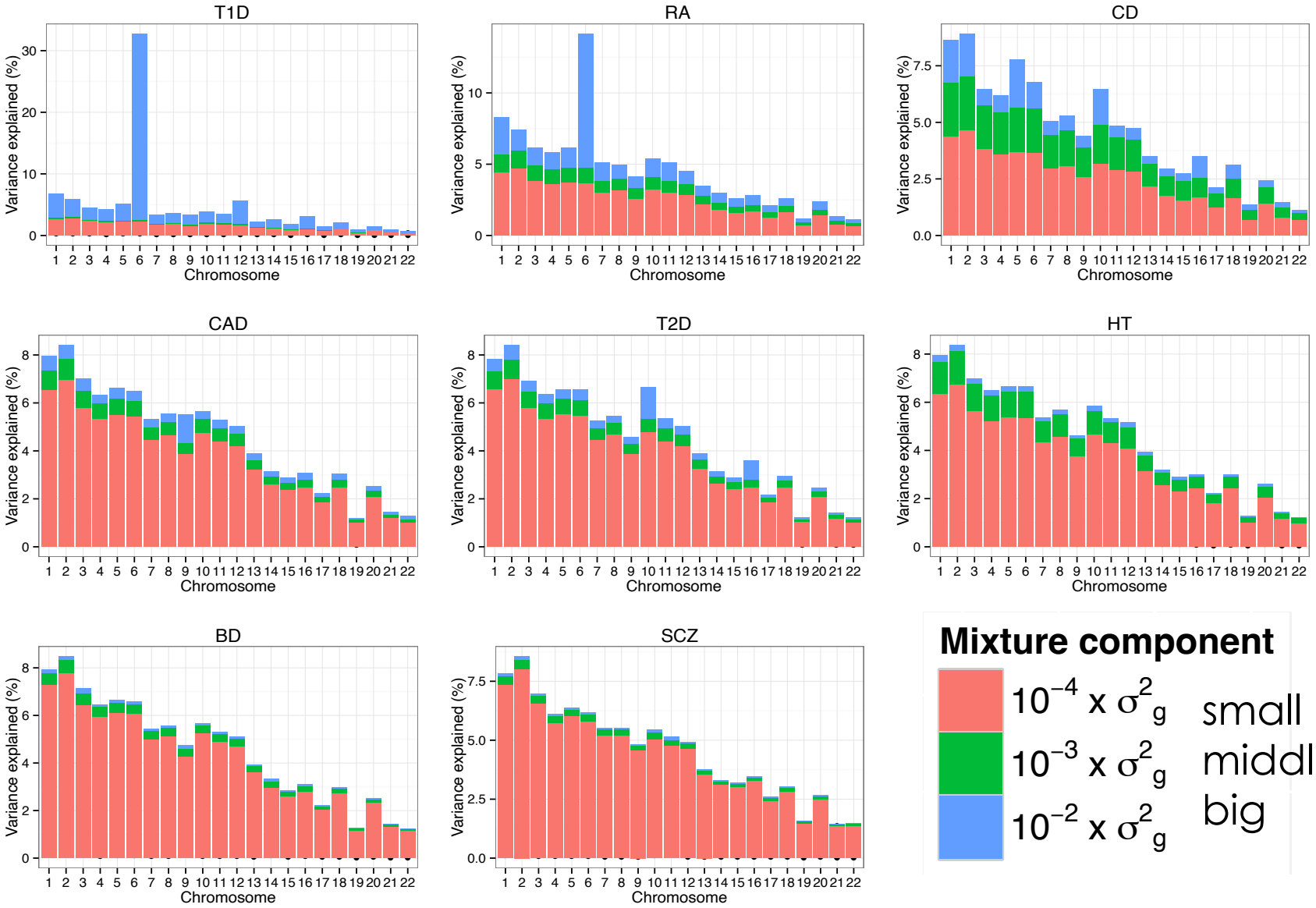
GWAS sample size

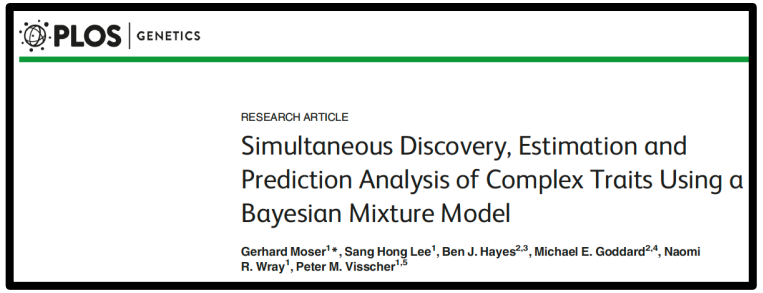
LD correlation matrix

Marginal SNP effects

Individual-level data
Summary-level data

Trait-specific genetic architecture





SBayes

$$\mathbf{b} = \mathbf{R}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

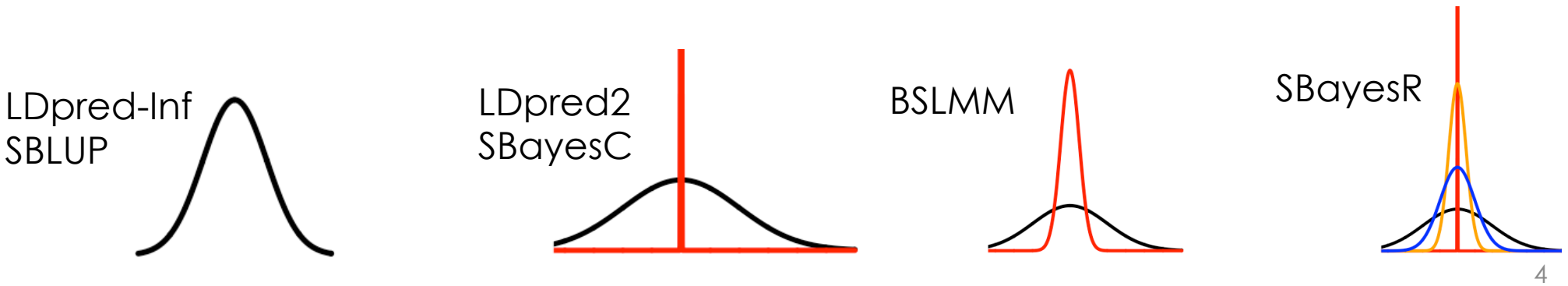
GWAS marginal SNP effects

LD correlation matrix

SNP joint effects

$$Var(\boldsymbol{\epsilon}) = \frac{1}{n}\mathbf{R}\sigma_e^2$$

- Prior distribution for each SNP effect



SBayesR

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-12653-0 OPEN

Improved polygenic prediction by Bayesian multiple regression on summary statistics

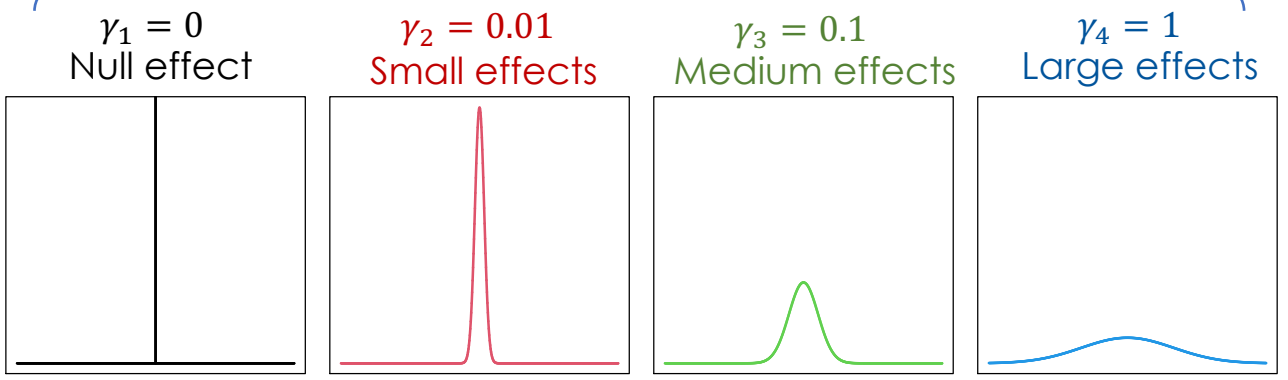
Luke R. Lloyd-Jones[1,9*], Jian Zeng[1,9*], Julia Sidorenko[1,2], Loïc Yengo[1], Gerhard Moser[3,4], Kathryn E. Kemper[1], Huanwei Wang[1], Zhili Zheng[1], Reedik Magi[2], Tõnu Esko[2], Andres Metspalu[2,5], Naomi R. Wray[1,6], Michael E. Goddard[7], Jian Yang[1,8*] & Peter M. Visscher[1*]

- Each SNP effect follows a mixture distribution:

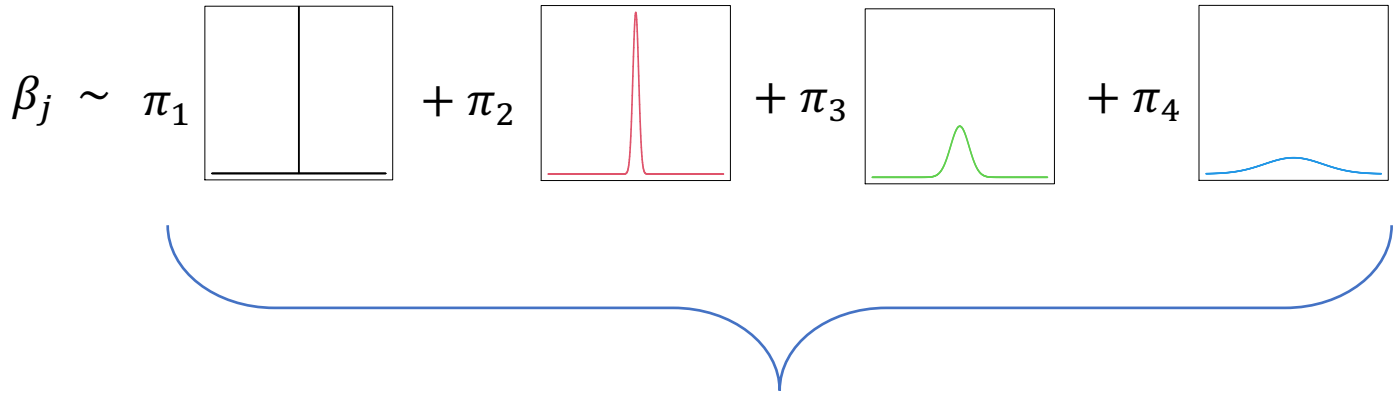
$\boldsymbol{\pi} \sim Dirichlet\ (\mathbf{1})$

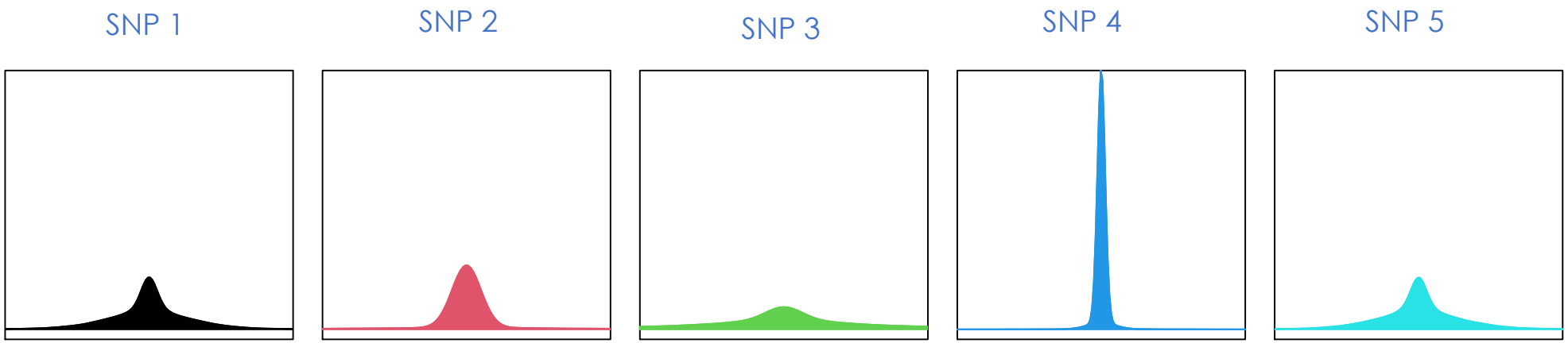
$$\beta_j \sim \sum_{k=1}^{4} \pi_k\, N\left(0, \gamma_k \sigma_\beta^2\right)$$

$\sigma_\beta^2 \sim \chi_v^{-1}(v, \tau^2)$



Account for various SNP effect distributions





The posterior distribution of SNP effects

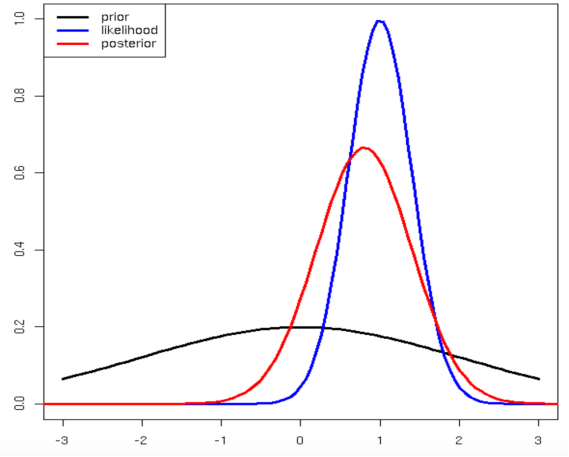
$$\textbf{Posterior} \propto \textbf{Likelihood} \times \textbf{Prior}$$

$$f(\boldsymbol{\beta}|\textbf{Summary data}) \propto f(\textbf{Summary data}|\boldsymbol{\beta}) \times f(\boldsymbol{\beta})$$

$\boldsymbol{\beta}|\mathbf{b} \sim \mathrm{N}(\mathbf{C}^{-1}\mathbf{r}, \mathbf{C}^{-1}\sigma_e^2)$

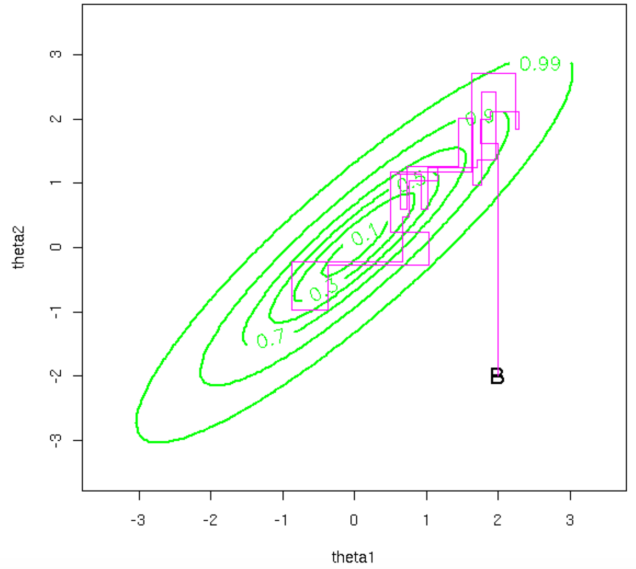
where

- $\mathbf{C} = n\mathbf{R} + \mathbf{G}^{-1}\frac{\sigma_e^2}{\sigma_\beta^2} = \mathbf{X}'\mathbf{X} + \mathbf{G}^{-1}\frac{\sigma_e^2}{\sigma_\beta^2}$
- $\mathbf{r} = n\mathbf{b} = \mathbf{X}'\mathbf{y}$
- $\mathbf{G} = diag\{\gamma_j\}$



Summary-level data
Individual-level data

Single-site Gibbs sampling



Full conditional distribution for β_j

$$f\left(\beta_j \mid \mathbf{b}, else\right) = N\left(\frac{r_j}{C_j}, \frac{\sigma_e^2}{C_j}\right)$$

where

- $r_j = nb_j - \sum_{k \neq j} R_{jk}\beta_k = \mathbf{X}_j'\left(\mathbf{y} - \sum_{k \neq j} \mathbf{X}_k\beta_k\right)$
- $C_j = n + \frac{\sigma_e^2}{\gamma_j \sigma_\beta^2} = \mathbf{X}_j'\mathbf{X}_j + \frac{\sigma_e^2}{\gamma_j \sigma_\beta^2}$

Summary-level data
Individual-level data

Comparison between individual and summary level algorithms

Algorithm 1 – Individual level data algorithm

Initialise parameters and read genotypes and phenotypes in PLINK binary format
Initialise $\mathbf{y}^* = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$
for i :=1 **to** number of iterations **do**
  for i :=1 **to** p **do**
    Calculate $r_j^* = \mathbf{x}_j'\mathbf{y}^*$
    Calculate $r_j = r_j^* + \mathbf{x}_j'\mathbf{x}_j\beta_j^{(i-1)}$
    Calculate $\sigma_c^2 = \sigma_\beta^2\gamma_{\delta_j=c}$ for each of C classes (e.g., BayesR C=4 and $\boldsymbol{\gamma} = (0, 0.0001, 0.001, 0.01)$)
    Calculate the left hand side $l_{jc} = \mathbf{x}_j'\mathbf{x}_j + \frac{\sigma_\varepsilon^2}{\sigma_c^2}$ for each of the C classes
    Calculate the log densities of given $\delta_j = c$ using $\log(\mathcal{L}_c) = -\frac{1}{2}\left[\log\left(\frac{\sigma_c^2 l_{jc}}{\sigma_\varepsilon^2}\right) - \frac{r_j^2}{\sigma_\varepsilon^2 l_{jc}}\right] + \log(\pi_c)$, where π_c is the current
    Calculate the full conditional posterior probability for $\delta_j = c$ for C classes with $\mathbb{P}(\delta_j = c|\boldsymbol{\theta}, \mathbf{y}) = \frac{1}{\sum_{l=1}^{C}\exp[\log(\mathcal{L}_l) - \log(\mathcal{L}_c)]}$
    Using full conditional posterior probabilities sample class membership for $\beta_j^{(i)}$ using categorical random variable sampler
    Given class sample SNP effect $\beta_j^{(i)}$ from $N\left(\frac{r_j}{l_{jc}}, \frac{\sigma_\varepsilon^2}{l_{jc}}\right)$
    Given SNP effect adjust corrected phenotype side $(\mathbf{y}^*)^{(i)} = (\mathbf{y}^*)^{(i-1)} - \mathbf{x}_j\left(\beta_j^{(i)} - \beta_j^{(i-1)}\right)$
  od
  Sample update from full conditional for σ_β^2 from scaled inverse chi-squared distribution $\widetilde{\nu}_\beta = \nu_\beta + q$ and $\widetilde{S^2}_\beta = \frac{\nu_\beta S_\beta^2 + \sum_{j=1}^{q}\frac{\beta_j^2}{\gamma_c}}{\nu_\beta + q}$, where q is the number of non-zero variants
  Sample update from full conditional for σ_ε^2 from scaled inverse chi-squared distribution $\widetilde{\nu}_e = n + \nu_e$ and scale parameter $\widetilde{S}_\varepsilon^2 = \frac{SSE + \nu_\varepsilon S_\varepsilon^2}{n + \nu_\varepsilon}$ and $SSE = \mathbf{y}^{*\prime}\mathbf{y}^*$
  Sample update from full conditional for $\boldsymbol{\pi}$, which is Dirichlet($C, \mathbf{c} + \boldsymbol{\alpha}$), where $\mathbf{c}$ is a vector of length C and contains the counts of the number of variants in each variance class and $\boldsymbol{\alpha} = (1, \ldots, 1)$
  Calculate genetic variance for h_{SNP}^2 calculation using $\sigma_g^2 = \mathrm{Var}(\mathbf{X}\boldsymbol{\beta})$
  Calculate $h_{SNP}^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\varepsilon^2}$
od

Algorithm 2 Summary data algorithm

Initialise parameters and read summary statistics
Reconstruct $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{y}$ from summary statistics and LD reference panel
Calculate $\mathbf{r}^* = \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta}$
for i :=1 **to** number of iterations **do**
  for i :=1 **to** p **do**
    Calculate $\mathbf{r}_j = \mathbf{r}_j^* + \mathbf{x}_j'\mathbf{x}_j\beta_j$
    Calculate $\sigma_c^2 = \sigma_\alpha^2\gamma_{\delta_j=c}$ for each fo C classes (e.g., SBayesR C=4 and $\boldsymbol{\gamma} = (0, 0.01, 0.1, 1)'$)
    Calculate the left hand side $l_{jc} = \mathbf{x}_j'\mathbf{x}_j + \frac{\sigma_\varepsilon^2}{\sigma_c^2}$ for each of the C classes
    Calculate the log densities of given $\delta_j = c$ using $\log(\mathcal{L}_c) = -\frac{1}{2}\left[\log\left(\frac{\sigma_c^2 l_{jc}}{\sigma_\varepsilon^2}\right) - \frac{r_j^2}{\sigma_\varepsilon^2 l_{jc}}\right] + \log(\pi_c)$, where π_c is the current
    Calculate the full conditional posterior probability for $\delta_j = c$ for C classes with $\mathbb{P}(\delta_j = c|\boldsymbol{\theta}, \mathbf{y}) = \frac{1}{\sum_{l=1}^{C}\exp[\log(\mathcal{L}_l) - \log(\mathcal{L}_c)]}$
    Using full conditional posterior probabilities sample class membership for $\beta_j^{(i)}$ using categorical random variable sampler
    Given class sample SNP effect $\beta_j^{(i)}$ from $N\left(\frac{\mathbf{r}_j}{l_{jc}}, \frac{\sigma_\varepsilon^2}{l_{jc}}\right)$
    Given SNP effect adjust corrected right hand side $(\mathbf{r}^*)^{(i+1)} = (\mathbf{r}^*)^{(i)} - \mathbf{X}'\mathbf{x}_j\left(\beta_j^{(i+1)} - \beta_j^{(i)}\right)$. $\mathbf{X}'\mathbf{x}_j$ is the jth column of $\mathbf{X}'\mathbf{X}$.
  od
  Sample update from full conditional for σ_α^2 from scaled inverse chi-squared distribution $\widetilde{\nu}_\alpha = \nu_0 + q$ and $\widetilde{\tau^2}_\alpha = \frac{\nu_0\tau_0^2 + \sum_{j=1}^{q}\frac{\beta_j^2}{\gamma_{\delta_j}}}{\nu_0 + q}$, where q is the number of non-zero variants
  Sample update from full conditional for σ_ε^2 from scaled inverse chi-squared distribution $\widetilde{\nu}_e = n + \nu_e$ and scale parameter $\widetilde{\tau}_\varepsilon^2 = \frac{SSE + \nu_\varepsilon\tau_\varepsilon^2}{n + \nu_\varepsilon}$ and $SSE = \mathbf{y}'\mathbf{y} - \boldsymbol{\beta}'\mathbf{r}^* - \boldsymbol{\beta}'\mathbf{X}'\mathbf{y}$
  Sample update from full conditional for $\boldsymbol{\pi}$, which is Dirichlet($C, \mathbf{c} + \boldsymbol{\alpha}$), where $\mathbf{c}$ is a vector of length C and contains the counts of the number of variants in each variance class.
  Calculate genetic variance for h_{SNP}^2 calculation using $\sigma_g^2 = MSS/n$, where $MSS = \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \widehat{\boldsymbol{\beta}}'r^*$
  Calculate $h_{SNP}^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\varepsilon^2}$
od

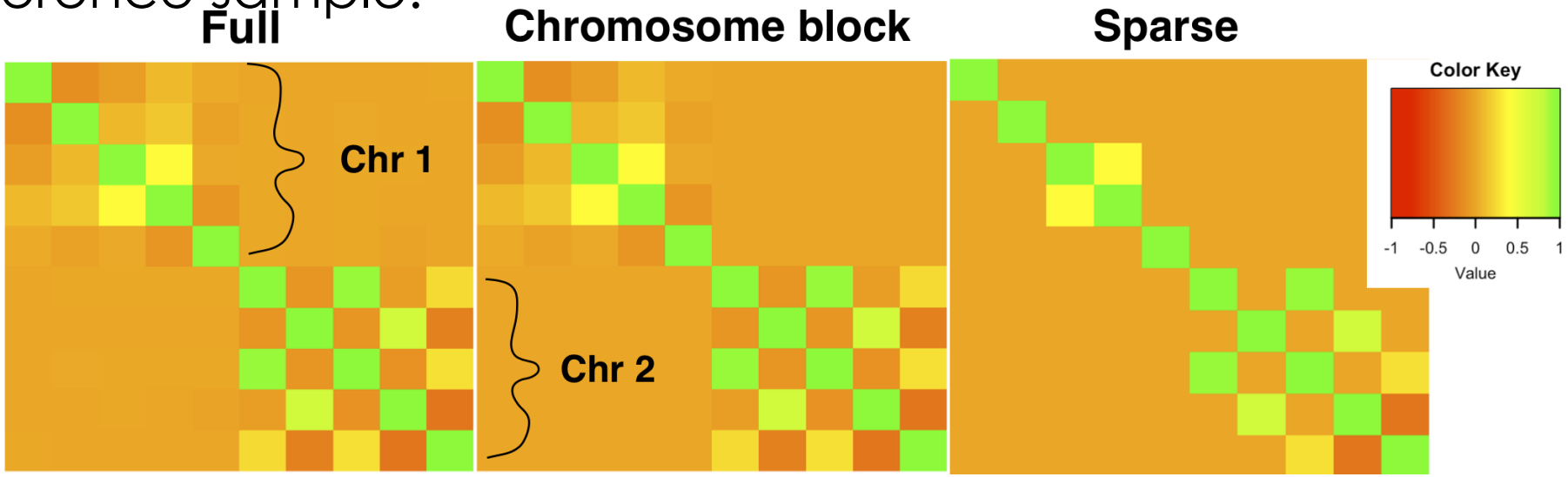
X'y and **X'X** are sufficient statistics in the algorithm!

Scaling the summary statistics

- Derivation is based on standardised genotypes.
- GWAS effects are usually in unit of allele count (per-allele effects).
- Need to rescale GWAS effects by $b_j^* = s_j b_j$ where $s_j = \sqrt{2p_j(1-p_j)}$
- Allele frequency p_j could be reported with errors.
- Use of standard error is more robust:
  - $SE_j = \sqrt{\frac{1}{\mathbf{X}_j'\mathbf{X}_j}\hat{\sigma}_e^2} = \sqrt{\frac{1}{nh_j}\left(\sigma_y^2 - s_j^2 b_j^2\right)} \quad \rightarrow \quad s_j = \sqrt{\frac{\sigma_y^2}{nSE_j^2 + b_j^2}} \approx \sqrt{\frac{\sigma_y^2}{nSE_j^2}}$

LD matrix from a reference

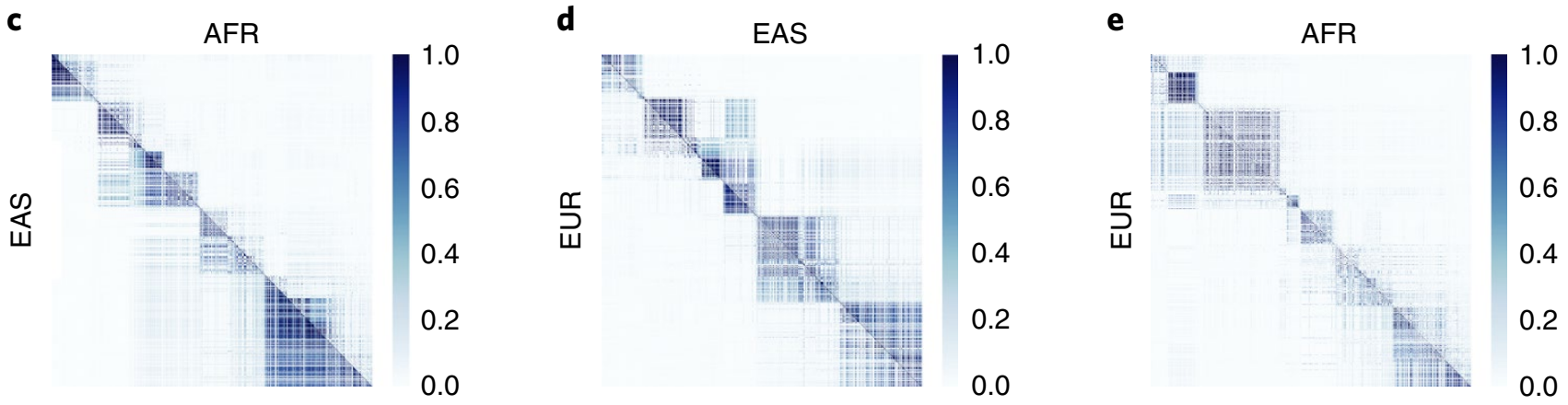
- Often cannot use genome-wide full LD matrix from the GWAS sample.
- Use a reduced (banded, sparse, or shrunk) LD matrix from a reference sample.



Challenging to extend to non-human cases,
where cross-chromosomal LD is NOT negligible

Implicit assumptions

- LD reference population matches with GWAS population in genetics
  - No systematic differences in LD → same ancestry and population structure
  - Minimum sampling variance in LD → LD ref sample size cannot be too small

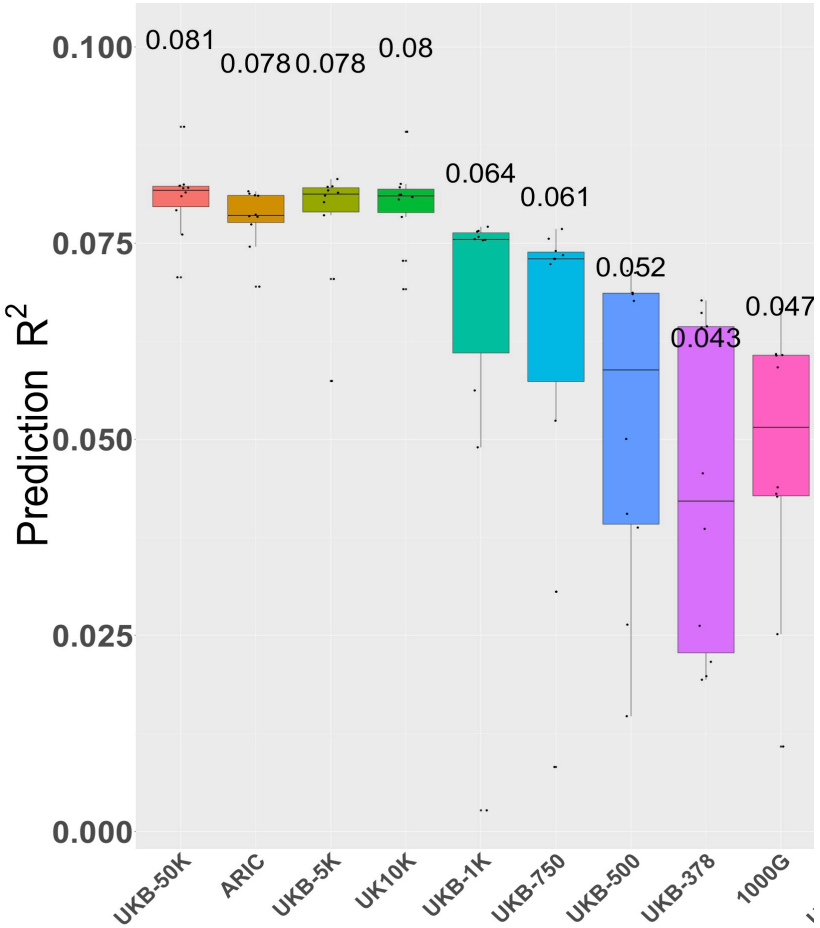


Martin et al. 2019 NG

Implicit assumptions

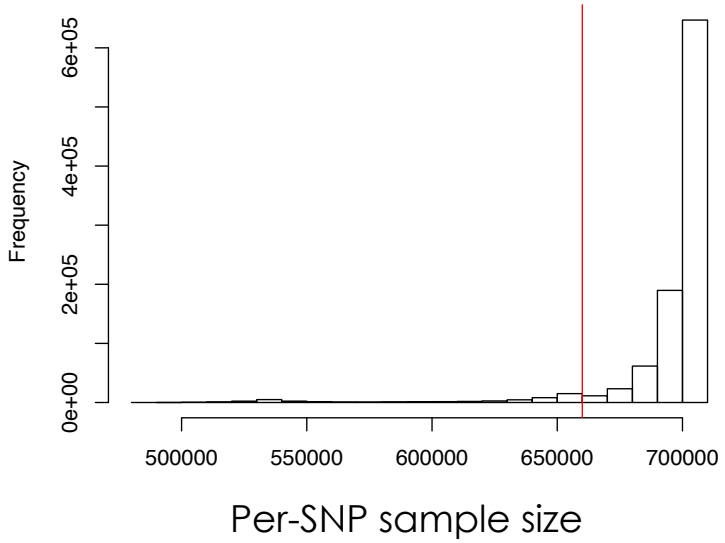
- LD reference population matches with GWAS population in genetics
  - No systematic differences in LD → same ancestry and population structure
  - Minimum sampling variance in LD → LD ref sample size cannot be too small
- GWAS data are collected on the same set of individuals
  - Often an issue in GWAS meta-analysis
  - Consistent genotyping platforms or imputation panels across cohorts
  - Remove SNP outliers in per-SNP sample size
- Violation these assumptions can cause model misspecification, resulting in attenuated prediction accuracy or even failure to reach convergence.

Influence of choice of LD dataset

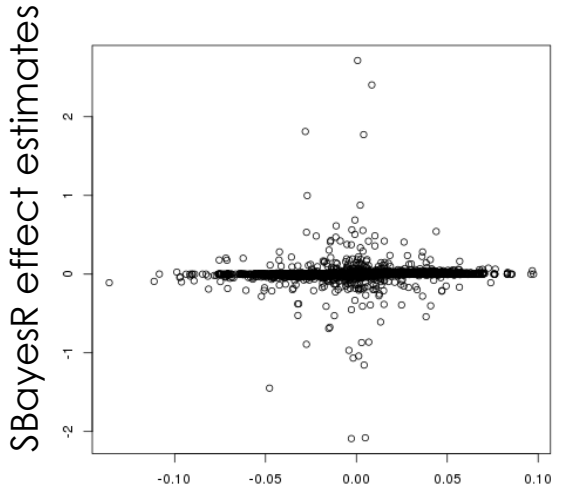


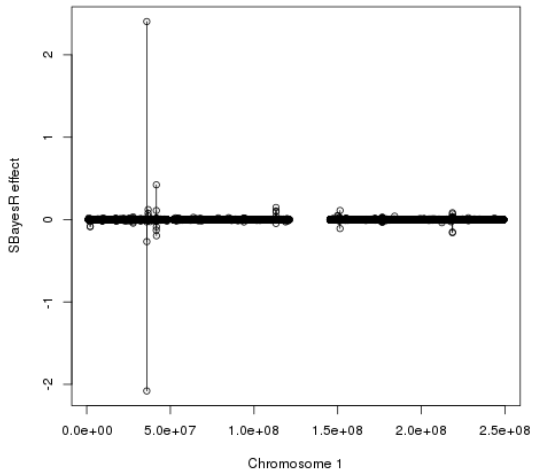
Lloyd-Jones & Zeng et al. 2019 NC

Influence of heterogeneity in per-SNP sample size

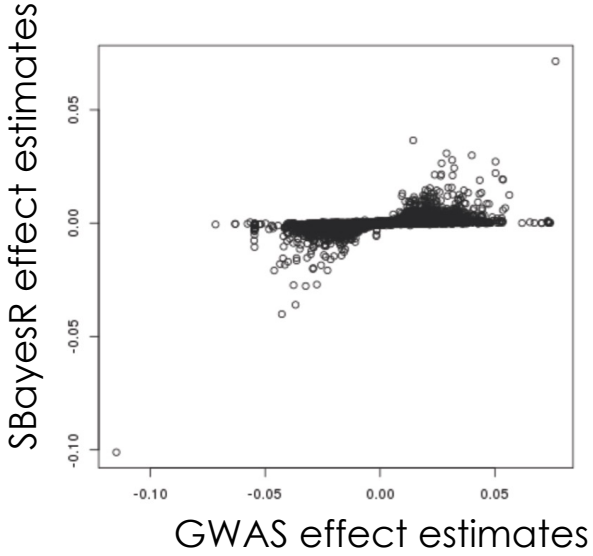


Abnormal



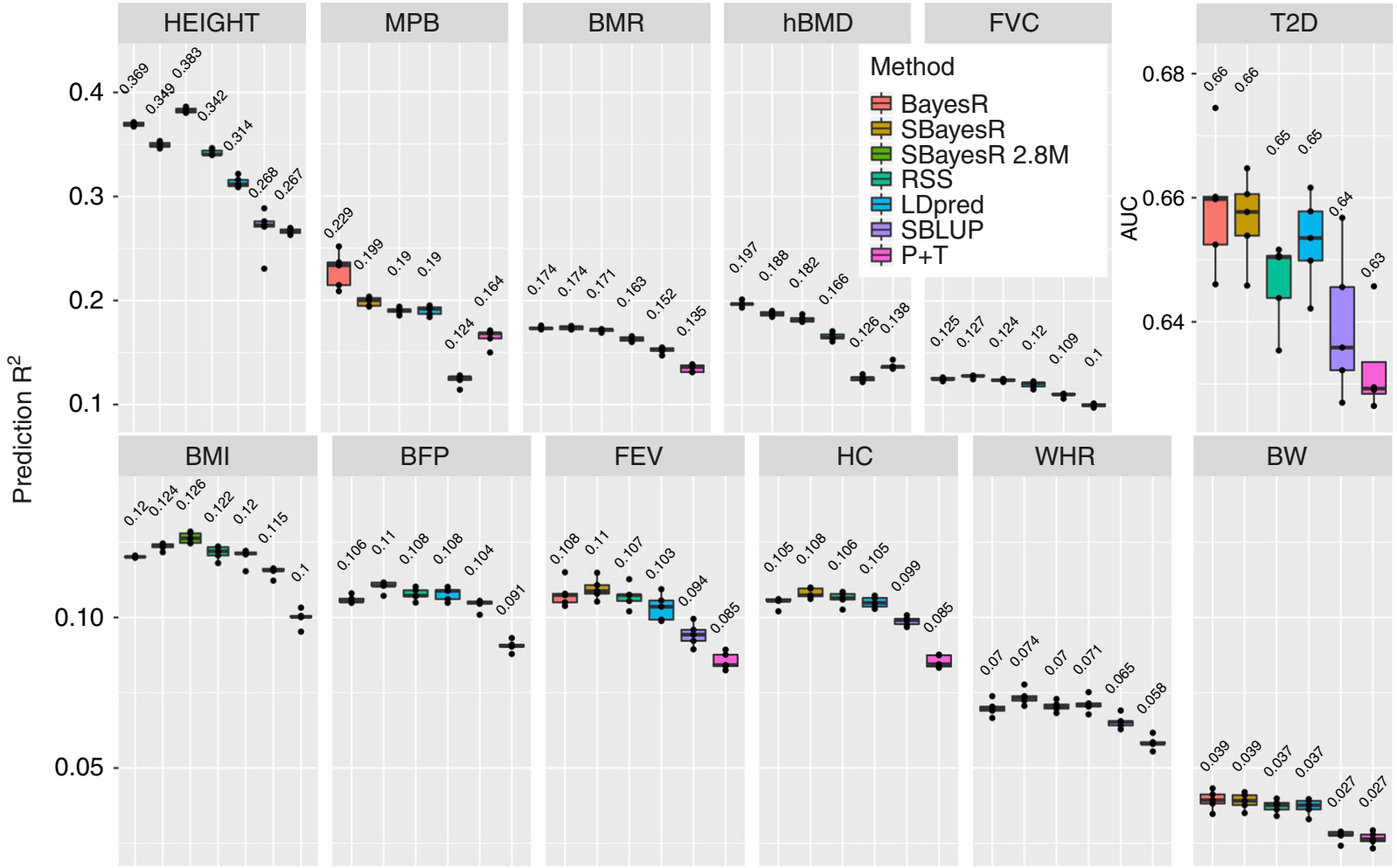


Normal



https://cnsgenomics.com/software/gctb/#FAQ

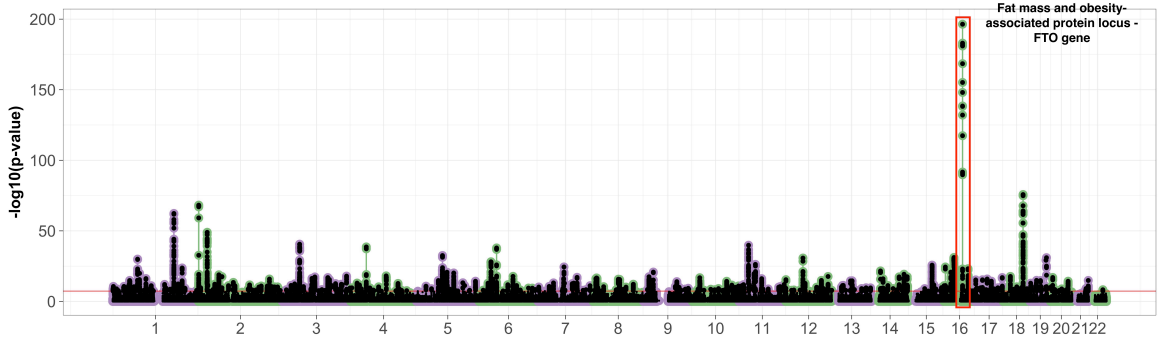
Method comparison



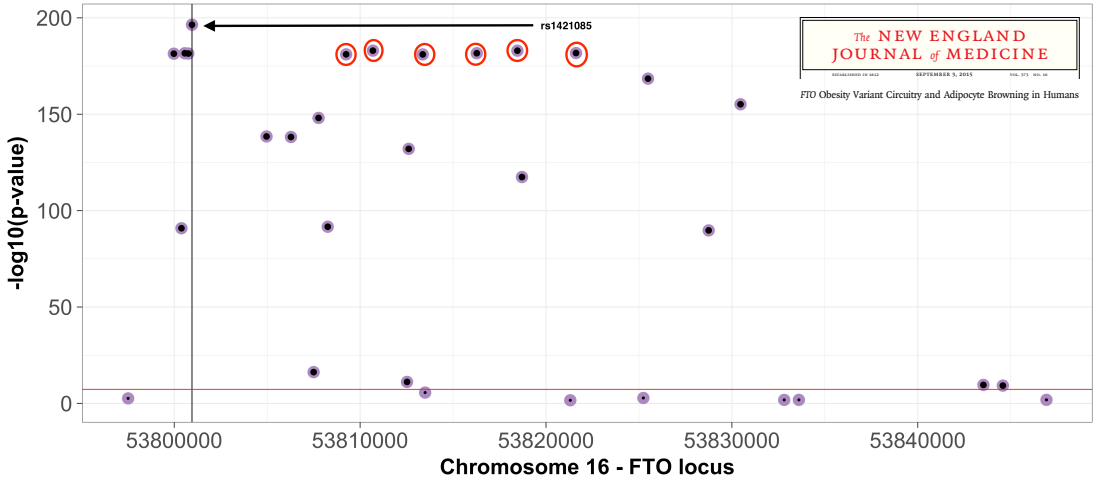
Lloyd-Jones & Zeng et al. 2019 NC

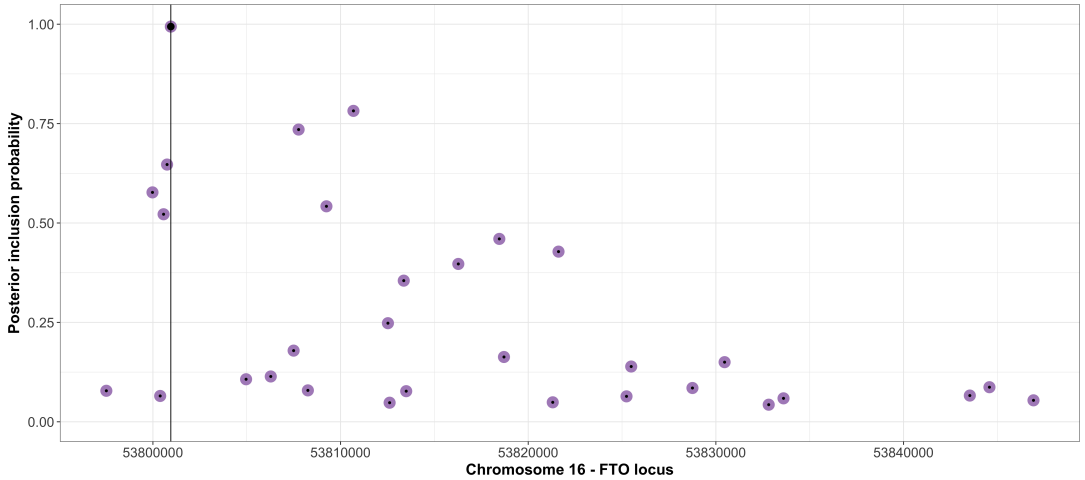
Fine-mapping

Real data - body mass index



Real data - FTO - GWAS

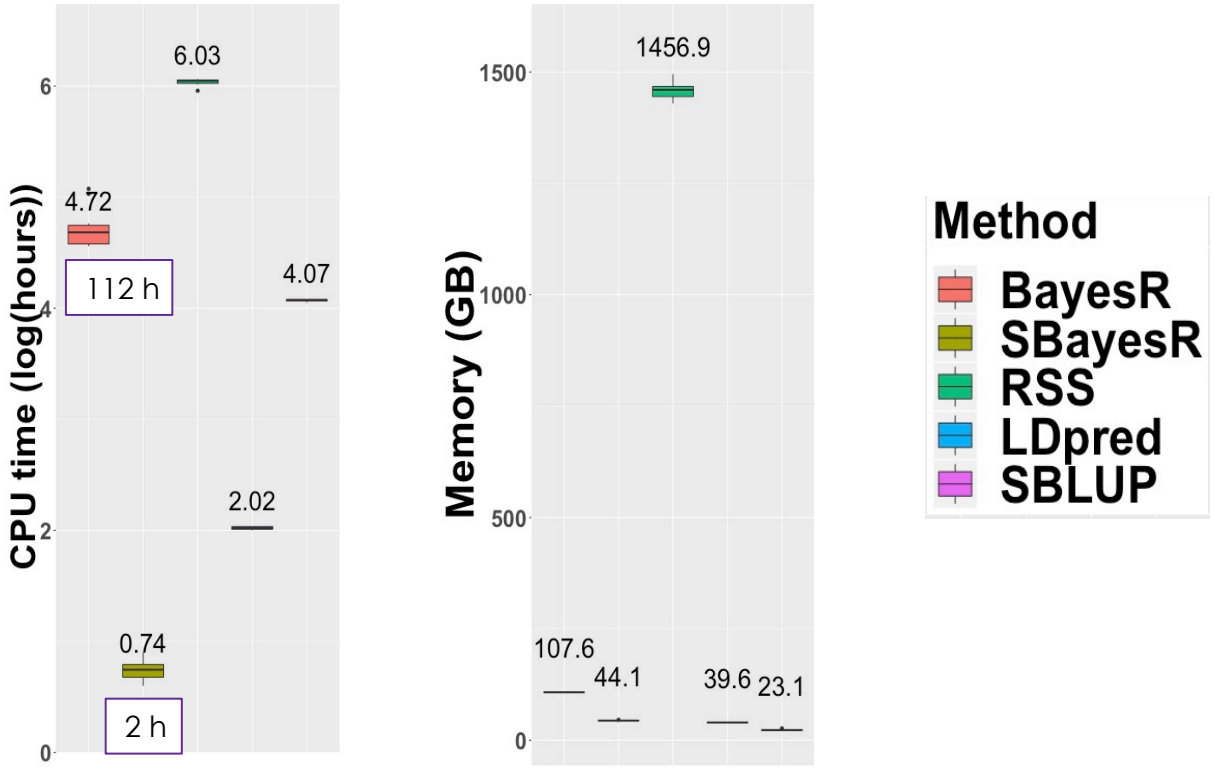




Computational efficiency

- 100k individuals
- 1M SNPs

- Resource consumption for summary-data-based methods is independent of sample size once GWAS summary statistics are obtained.



Summary

- Summary-level methods exploit the full power of GWAS of large sample sizes for polygenic prediction.
- Approximations to the individual-level methods.
- Free from limitation of data accessibility.
- Computationally efficient.
- Flexible to incorporate other information, such as gene expression.

Limits of PGS

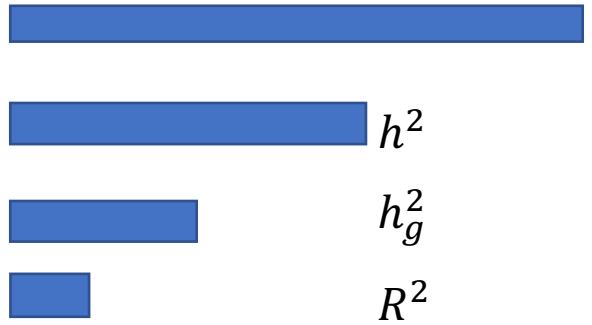
Understanding PGS limits

- PGS have a **theoretical** upper limit dependent on the **heritability of the trait** (how much of the variance of trait values between people is attributed to genetic factors).

- PGS have a **technical** upper limit associated with the proportion of **variance tagged** by the DNA variants measured.

- PGS have a **practical** upper limit dependent on the **sample size of the discovery sample** used to estimate effect sizes of risk alleles, and the **quality** of the discovery sample.

- PGS can be pushed closer to the technical upper limit by the **statistical methodology** used to generate the optimal weighting given to the risk alleles, and new methods integrate new biological data.



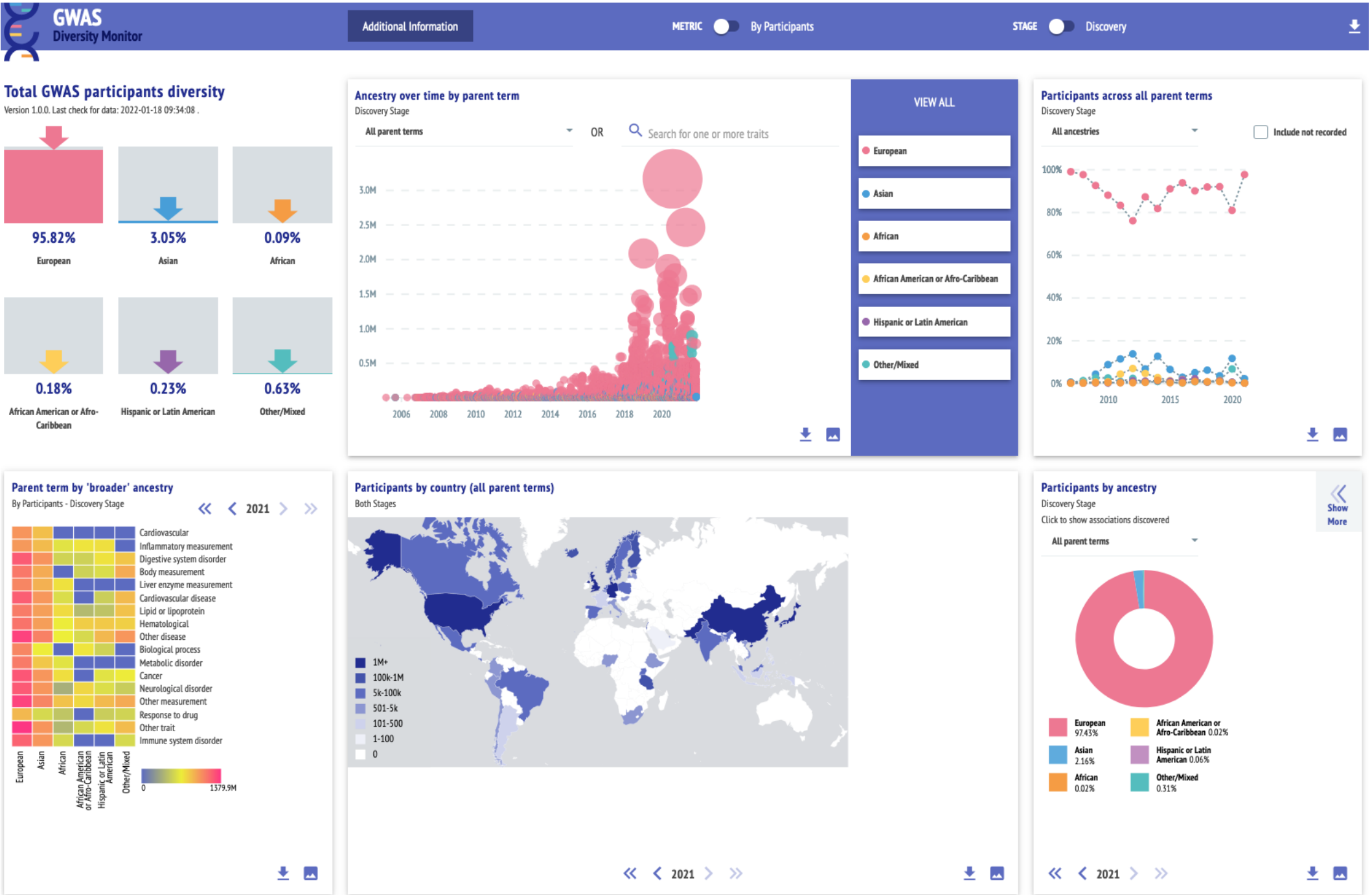
Schizophrenia

Max:
25% Liability
AUC 0.84

Current:
11% Liability
AUC 0.74

Polygenic scores cannot be highly accurate predictors of phenotypes

Loss of prediction accuracy of polygenic predictors across populations



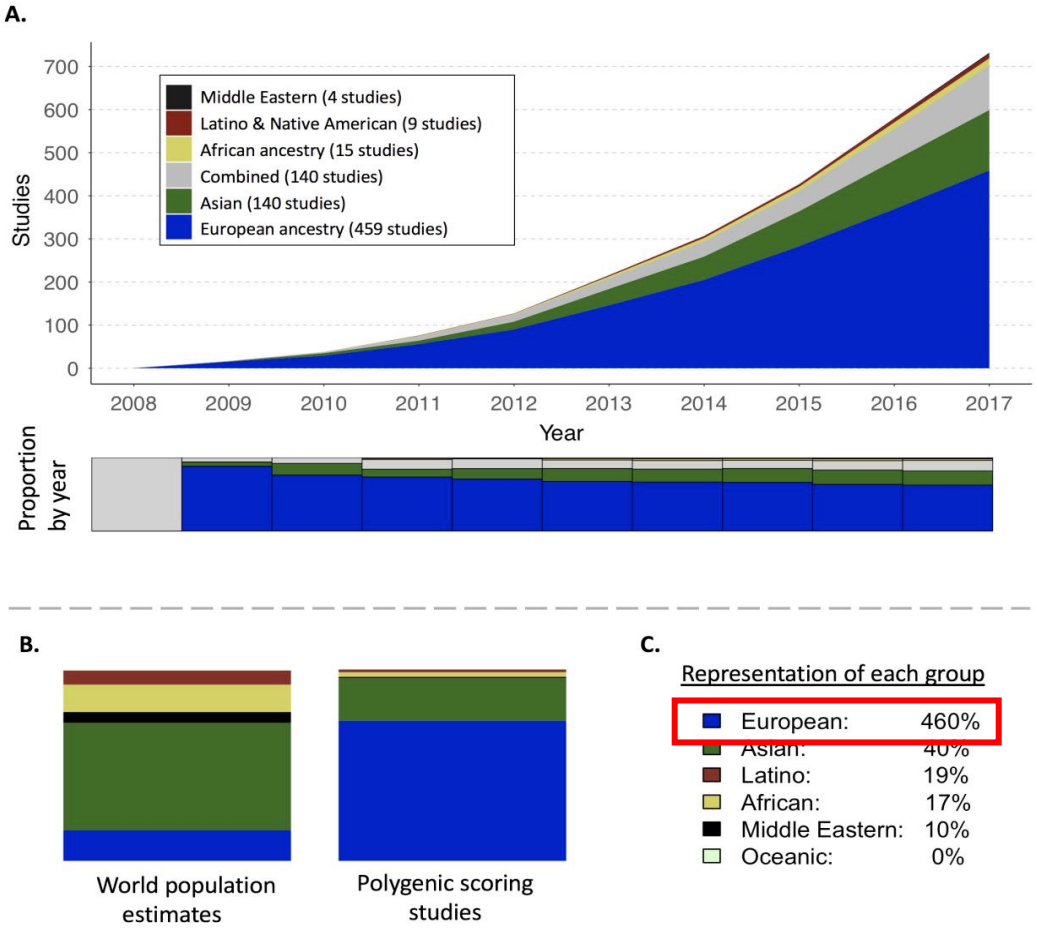
In 2022…

>95% European

<4% Asian

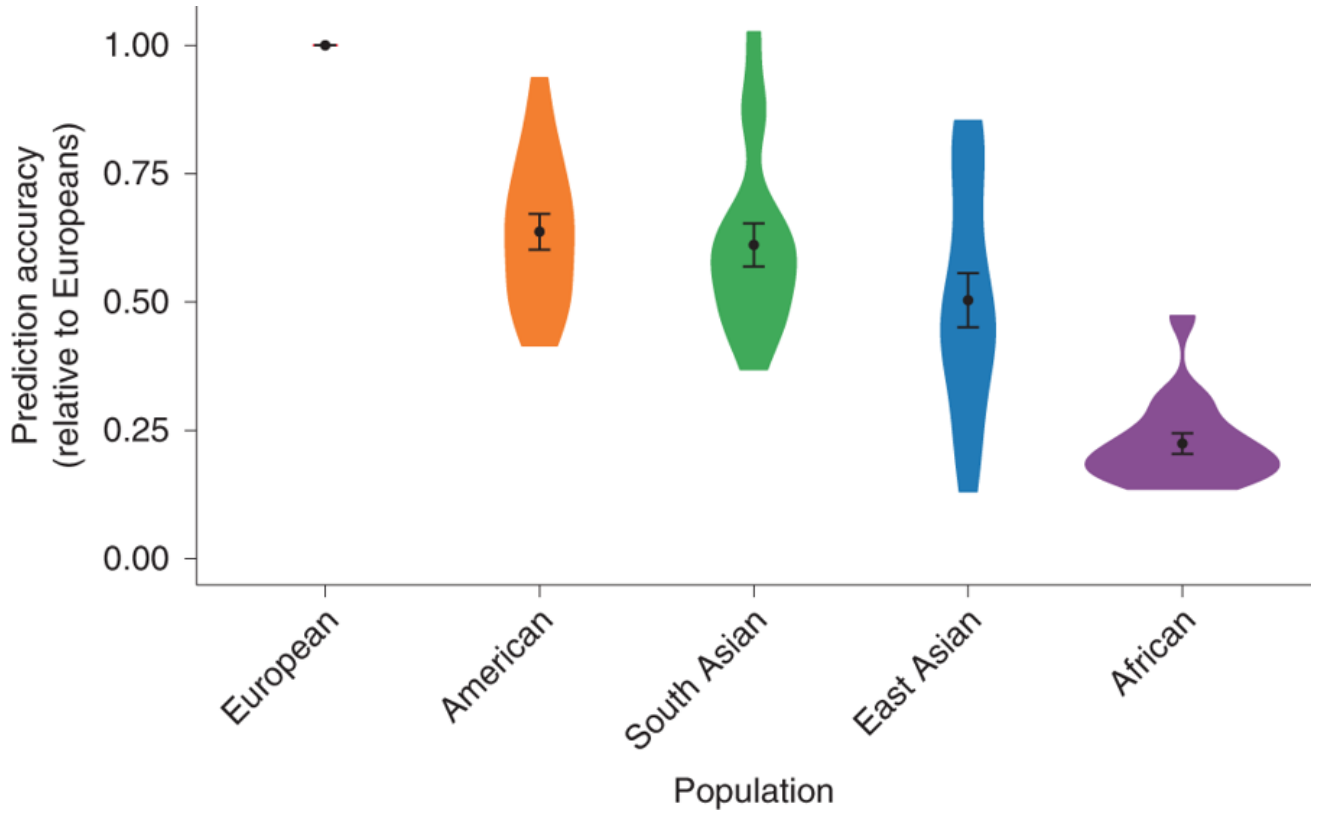
<0.1% African

Euro-centric GWASs bias



Duncan et. al., *Nature Communications*, 2019

Loss of prediction accuracy



Martin et. al., *Nature Genetics*, 2019

We are missing GWAS in non-European ancestries populations

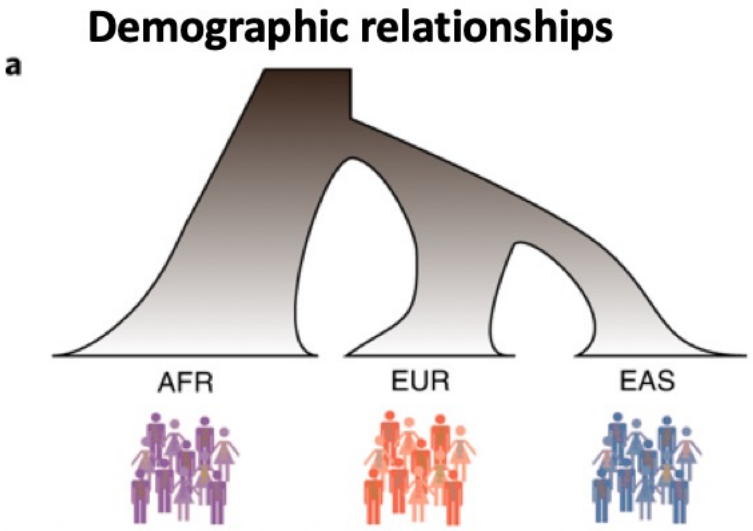
Ancestry-specific causal variants (worse case scenario)

Same causal variants but different effect sizes

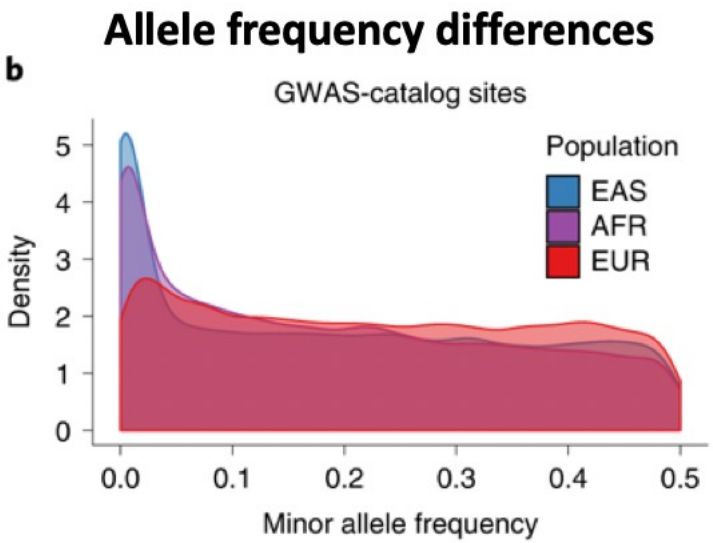
Same causal variants, same effect sizes, same heritability, but different haplotype frequency

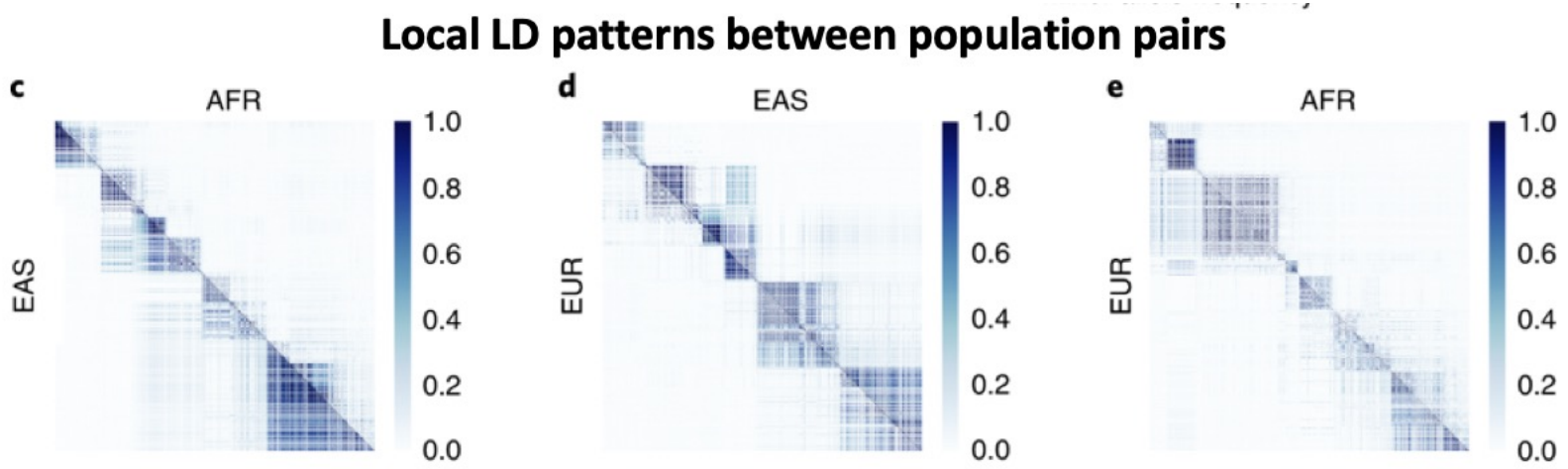
…

Factors affecting PGS accuracy disparity

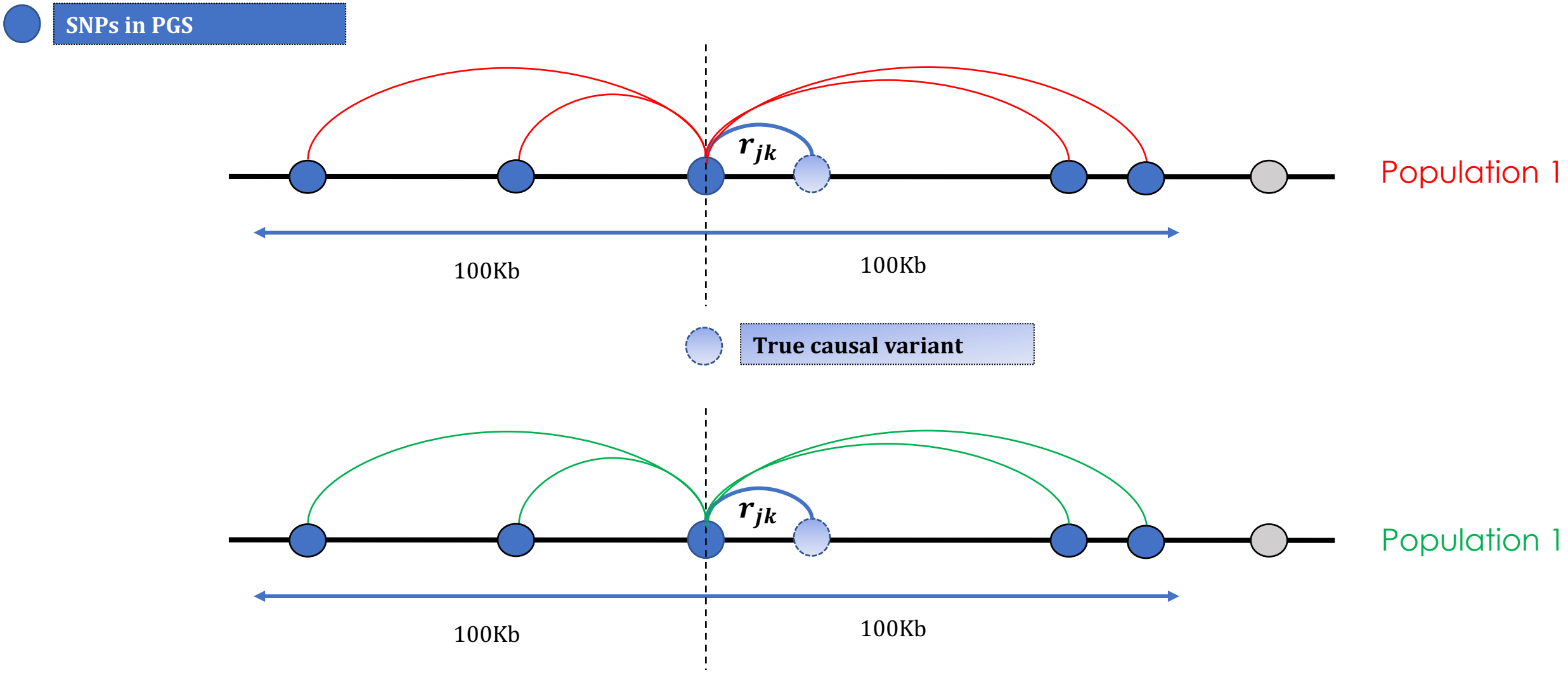


Martin et. al., *Nature Genetics*, 2019





LD difference



Factors affecting PGS accuracy disparity

How to **quantify** the **loss of accuracy** attributable to MAF and LD?

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-17719-y OPEN

Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations

Ying Wang[1], Jing Guo[1], Guiyan Ni[1], Jian Yang[1,2], Peter M. Visscher[1] & Loic Yengo[1]✉

1.1 Theory

Relative Accuracy (RA)

- $\frac{R_2^2}{R_1^2}$ = Function of $r_b^2 h_2^2 / h_1^2$, $p_{k1}^{(t)}, p_{k2}^{(t)}, \frac{\mathrm{var}(\hat{y}_1)}{\mathrm{var}(\hat{y}_2)}$ and $r_{jk,1}, r_{jk,2}$

$\boldsymbol{r_b^2}$: squared genetic correlation

$\boldsymbol{h_l^2}$: trait heritability in Population l

$\boldsymbol{p_{k,l}}$: MAF in Population l

$\boldsymbol{r_{jk,l}}$: LD correlation between j-th causal variant and k-th tagSNP in Population l

$\boldsymbol{var\ (\hat{y}_l)}$: variance of PGS in Population l

Conclusions

(1) Relative accuracy is lower than expected if assuming GWAS SNPs are causal variants (=> wrong conclusion)

(2) LD and MAF differences between populations account for the majority of the loss of prediction accuracy

(3) Empirical data are consistent with causal variants being largely shared across populations.

Trends in
Genetics

CellPress

Feature Review

Genetic prediction of complex traits with polygenic scores: a statistical review

Ying Ma[1] and Xiang Zhou [1,2,*]

More methods…

Methods = What SNPs + What Weights (Linear Models)

- Clumping + Thresholding (Historical – less popular now)

- COJO + Thresholding

- Random Regression (Genome-wide) Methods 
  - Individual-level data: BLUP, Bayesian Regression (e.g., BayesR, BOLT-LMM)
  - Summary-statistics methods

SS Parametric Methods*

Brisbane Teams
Melbourne Teams
Arhus Teams
(Speed + Vihhjalmsson)

- SBayes- family: no tuning but need to specify the model (e.g., S, C, R, etc.)
- SBLUP (tuning parameter $\lambda = M(1-h^2)/(Nh^2)$ is fixed)

Arhus Teams
(Speed + Vihhjalmsson + Prive)

- LDpred infinitesimal ⇔ SBLUP
  (tuning parameter λ is selected on a grid – **need tuning sample**)
- LDpred ⇔ SBayesC parameters π and h^2 either estimated with MCMC or selected on a grid – parametrization slightly differs from SBayesC

Pak Sham Group (2017)

- LASSOSUM (penalizes the L_1 norm): tuning parameter λ and s (shrinkage on the LD matrix) selected on a grid – **need tuning sample** or GWAS summary statistics.

- Deterministic Bayesian Sparse Linear Mixed Model (DBSLMM) – combines C+T, COJO, LDSC and SBLUP

Zhou Group (2020)

*RSS paper by Zhu & Stephens (2017) – formalization

SS Nonparametric Methods

- Continuous Shrinkage Prior (Dirichlet Prior) ⇔ infinite number of components in mixture

DPR method (2017)
Zhou Group [Michigan]

PRS CS (2019, 2022)
Ge Group [Boston]

- Non-parametric Shrinkage [Partitioned GWAS summary stats]

Sunyaev Group (2020)
[Boston]

- Semi-Parametric: MEGA-PRS (combines multiple methods and learn weights from validation sample)

Parametric vs Nonparametric

Bayesian Parametric

- Quest for the "Best" prior -- flexibility
- Flexibility has cost
- Modelling flexibility to incorporate additional information
- MCMC can be expensive

Nonparametric

- No need to specify THE best prior
- Not straightforward to included external information
- Bayesian => MCMC expensive
- Non-Bayesian => need tuning data

Empirical Bayes Parametric (e.g., SBLUP)

- Strong assumptions on effect sizes distribution
- Inference is fast if no MCMC
- May require tuning sample

…The eternal tension between Bias (explanation) and Variance (Prediction)

Notable extensions

- Multi-traits

Improving genetic prediction by leveraging genetic correlations among human diseases and traits

Robert M. Maier, Zhihong Zhu, Sang Hong Lee, Maciej Trzaskowski, Douglas M. Ruderfer, Eli A. Stahl, Stephan Ripke, Naomi R. Wray, Jian Yang, Peter M. Visscher & Matthew R. Robinson

Nature Communications **9**, Article number: 989 (2018) | Cite this article

- Functional annotations

Incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets

Carla Márquez-Luna, Steven Gazal, Po-Ru Loh, Samuel S. Kim, Nicholas Furlotte, Adam Auton, 23andMe Research Team & Alkes L. Price

Nature Communications **12**, Article number: 6052 (2021) | Cite this article

- Tuning using summary statistics (LASSOSUM)

- Combine individual-level and summary data

Article

Leveraging both individual-level genetic data and GWAS summary statistics increases polygenic prediction

Clara Albiñana [1, 2], Jakob Grove [1, 3, 7, 8], John J. McGrath [2, 4, 5], Esben Agerbo [1, 2], Naomi R. Wray [6, 5], Cynthia M. Bulik [9, 10, 11], Merete Nordentoft [1, 12, 13], David M. Hougaard [1, 14], Thomas Werge [1, 15, 13, 16], Anders D. Børglum [1, 3, 7], Preben Bo Mortensen [1, 2], Florian Privé [1, 2, 17], Bjarni J. Vilhjálmsson [1, 2, 8, 17]

stacked clumping and thresholding (SCT) and meta-PRS. We find that, when large individual-level data are available, the linear combination of PRSs (meta-PRS) is both a simple alternative to meta-GWAS and often more accurate.

- A few Deep Learning Applications…

Incorporating functional annotation data

Credit: slides from Dr Jian Zeng and Shouye Liu

Literature

The University Of Queensland AUSTRALIA

nature communications

Explore content | About the journal | Publish with us

nature > nature communications > articles > article

Article | Open Access | Published: 18 October 2021

Incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets

Carla Márquez-Luna, Steven Gazal, Po-Ru Loh, Samuel S. Kim, Nicholas Furlotte, Adam Auton, 23andMe Research Team & Alkes L. Price

LDpredFunct method

PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Leveraging functional annotations in genetic risk prediction for human complex diseases

Yiming Hu, Qiongshi Lu, Ryan Powles, Xinwei Yao, Can Yang, Fang Fang, Xinran Xu, Hongyu Zhao

AnnoPred method

Winner's Curse Correction and Variable Thresholding Improve Performance of Polygenic Risk Modeling Based on Genome-Wide Association Study Summary-Level Data

Jianxin Shi, Ju-Hyun Park, Jubao Duan, Sonja T. Berndt, Winton Moy, Kai Yu, Lei Song, William Wheeler, Xing Hua, Debra Silverman, Montserrat Garcia-Closas, Chao Agnes Hsiung, Jonine D. Figueroa, [···], Nilanjan Chatterjee [view all]

P+T-funct-LASSO method

Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits

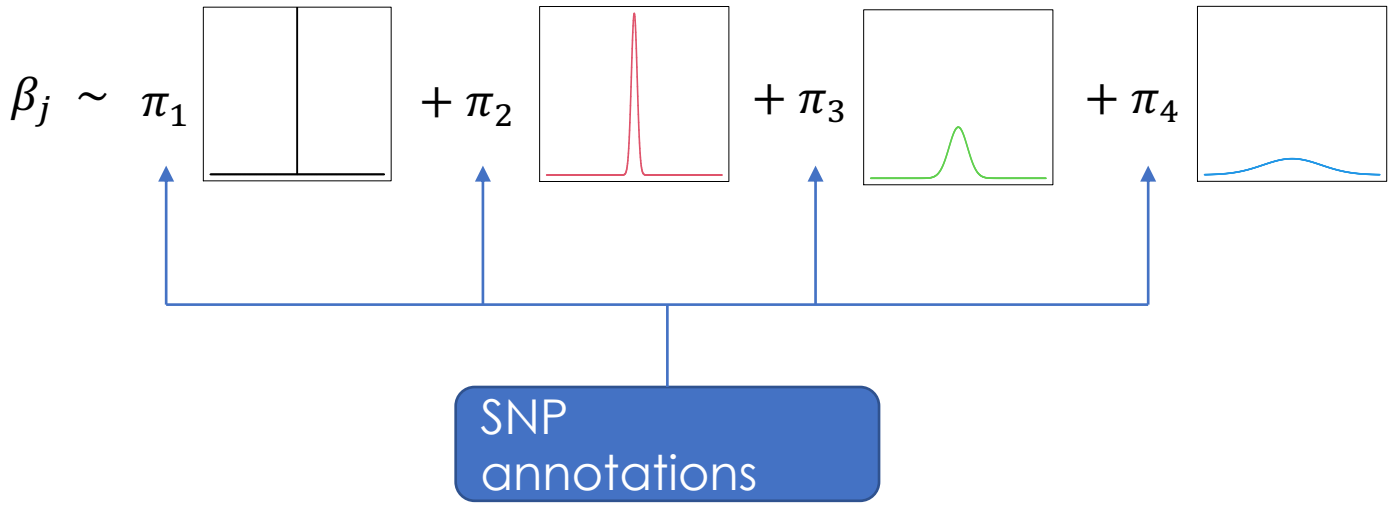
I. M. MacLeod, P. J. Bowman, C. J. Vander Jagt, M. Haile-Mariam, K. E. Kemper, A. J. Chamberlain, C. Schrooten, B. J. Hayes & M. E. Goddard

BMC Genomics **17**, Article number: 144 (2016) | Cite this article

6209 Accesses | **146** Citations | **9** Altmetric | Metrics

BayesRC method

Model annotation effects



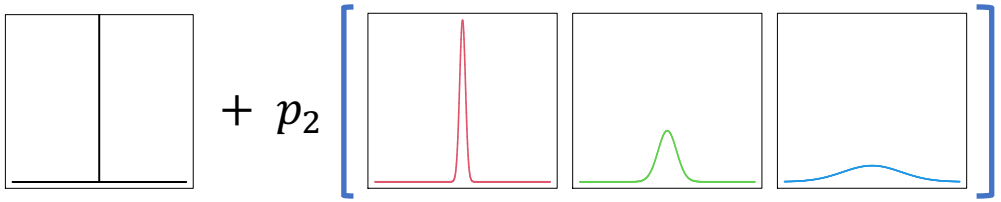
$$f(\pi_{jk}) = \mu_k + \sum \text{SNP annotation} \times \text{annotation effect}$$

Assumption	Pros	Cons
• Annotation effects are additive at the GLM scale.	• Estimation of conditional effects. • Allow annotation overlap. • Interpretation.	• # annotation effect parameters x 4. • $\pi_{j1} + \pi_{j2} + \pi_{j3} + \pi_{j4} = 1$.

Model annotation effects

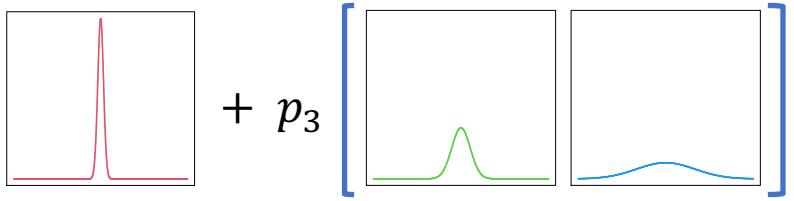
- Three independent 2-component models:
- For all SNPs

$$\beta_j \sim (1 - p_2) + p_2$$



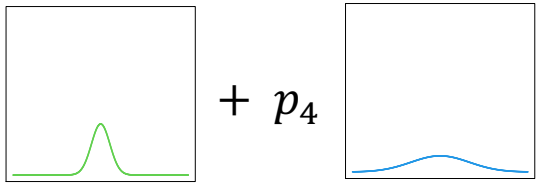
- For SNPs with nonzero effects

$$\beta_j \sim (1 - p_3) + p_3$$



- For SNPs with at least medium effects

$$\beta_j \sim (1 - p_4) + p_4$$



$$f(p_{jk}) = \mu_k + \sum \text{SNP annotation} \times \text{annotation effect}$$

GCTB software

https://cnsgenomics.com/software/gctb

GCTB
A tool for Genome-wide Complex Trait Bayesian analysis

GCTA | SMR | GSMR | OSCA | GCTB | Program in CTG | CTG forum

Download

Executable files

gctb_2.03beta_Linux.zip (*Lastest version*) *In this version, we introduced a more robust parameterisation for SBayesR and SBayesS to address the convergence issue which sometimes occurs when the GWAS summary statistics are from a meta-analysis. With the same command line code, the program will start with the original model but if a convergence issue is detected (indicated by a negative sampled value of residual variance) it will restart the MCMC process with a more robust parameterisation. The robust model can also be directed invoked by --robust. The robust parameterisation for SBayesR and SBayesS is inspired by the parameterisation in LDpred (Vilhjálmsson et al. 2015). Details will soon be available on our website.*

gctb_2.02_Linux.zip

gctb_2.0_Linux.zip

gctb_1.0_Linux.zip

gctb_1.0_Mac.zip

Tutorial data

GCTB tutorial data

LD matrices

The following LD matrices were computed based on 1.1 million common SNPs in a random sample of 50K unrelated individuals of European ancestry in UK Biobank dataset unless otherwise noted.

- Shrunk sparse matrix
- Shrunk sparse LD matrix (2.8 million common SNPs)

In the shrunk sparse matrices, described in Lloyd-Jones et al. (2019), the observed LD correlations computed from a reference sample were shrunk toward the expected values defined by a genetic map, following the algorithm in Wen and Stephens (2010). After shrinkage, LD correlations smaller than a threshold (default 1e-5) were set to be zero to give a sparse format, which is more efficient in storage and computation.

- Sparse matrix (including MHC regions)

The sparse matrices described in Zeng et al. (2021) were computed by setting the likely chance LD to zero based on a chi-squared test (default threshold at chi-squared test statistic of 10).

- Banded matrix (including MHC regions)

While the shrunk sparse matrices were used in our original SBayesR paper, Prive et al. (2021) found that using a banded matrix with a window size of 3 cM per SNP can improve prediction accuracy. Therefore, we have created such a LD matrix in GCTB format for SBayesR analysis.

Source code

GCTB 2.0 standard version

GCTB 2.0 MPI version

GCTB 1.0 standard version

GCTB 1.0 MPI version

The MPI version implements a distributed computing strategy that scales the analysis to very large sample sizes. A significant improvement in computing time is expected for a sample size > 10,000. The MPI version needs to be compiled on user's machine. See README.html in the tarball for instructions of compilation and usage. A testing dataset is also included in each tarball.

Tutorial

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

http://cnsgenomics.com/software/gctb/#Tutorial

GCTB
A tool for Genome-wide Complex Trait Bayesian analysis

GCTA | SMR | GSMR | OSCA | GCTB | Program in CTG | CTG forum

Tutorial

This updated version of the GCTB software (version 2.0) includes summary-data based versions of the individual-data Bayesian linear mixed models previously implemented. These methods require summary statistics from genome-wide association studies, which typically include the estimated univariate effect for each single nucleotide polymorphism (SNP), the standard error of the effect, the sample size used to estimate the effect for each SNP, the allele frequency of an allele, and an estimate of LD among SNPs, all of which are easily accessible from public databases.

This tutorial will outline how to use the summary data based methods and accompanies the manuscript Improved polygenic prediction by Bayesian multiple regression on summary statistics. To recreate the tutorial you will need the PLINK 2 software and the updated version of the GCTB software compiled from source or the binary executable in your path. The data for the tutorial are available at Download. The tutorial is designed so that each part can be reconstructed or picked up at any point with the following directory structure.

```
tutorial
  gctb # Static binary executable for Linux 64-bit systems
   data # 1000 Genomes data in PLINK format
   pheno_generation # R scripts for generating simulated phenotypes
   pheno # Simulated phenotypes
   gwas # PLINK GWAS summary statistics
   ldm # LD correlation matrices and scripts to calculate them
   ma # Summary statistics in GCTB compatible format
   sbayesr # Scripts for running and SBayesR analysis using GCTB and subsequent results
```

Data

The tutorial will run through all aspects of the software using genotype data from chromosome 22 of Phase 3 of the 1000 Genomes Project (1000G). The genotype data have been filtered to exclude variants with minor allele frequency (MAF) < 0.01, which left 15,938 SNPs available for analysis. These data include a subset of 378 individuals of European ancestry from the CEU, TSI, GBR and FIN populations.

N.B. The results generated from this example tutorial using data from the 1000G are designed to be lightweight and should be only used to explore how to run summary data based analyses using GCTB. The results do not make the best use of GCTB capabilities given the small sample size $(N = 378)$ in the example.

Phenotype simulation and GWAS

Using these genotypes, phenotypes were generated under the multiple regression model $y_i = \sum_{j=1}^{p} w_{ij}\beta_j + \epsilon_i$, where $w_{ij} = (x_{ij} - 2q_j)/\sqrt{2q_j(1-q_j)}$ with x_{ij} being the reference allele count for the ith individual at the jth SNP, q_j the allele frequency of the jth SNP and ϵ_i was sampled from a normal distribution with mean 0 and variance $\mathrm{Var}(W\beta)(1/h^2_{SNP} - 1)$ such that $h^2_{SNP} = 0.1$ for each of 20 simulation replicates, which is much larger than the contribution to the genome-wide SNP-based heritability (h^2_{SNP}) estimate for chromosome 22 for most quantitative traits. All phenotypes were generated using the R programming language with scripts used available in the `pheno_generation` subdirectory.

For each scenario replicate, we randomly sampled a new set of 1,500 causal variants. The genetic architecture simulated contained two causal variants of large effect explaining 3% and 2% of the phenotypic variance respectively and a polygenic tail of 1,498 causal variants sampled from a N(0, 0.05/1, 498) distribution such that the expected total genetic variance explained by all variants was 0.1.

For each of the 20 simulation replicates, simple linear regression for each variant was run using the PLINK 2 software to generate summary statistics. Code for running the GWAS and the output is available in the `gwas` subdirectory.

GCTB summary statistics input format

The GCTB summary-based methods have inherited the GCTA-COJO `.ma` format.

```
SNP A1 A2 freq b se p N
rs1001 A G 0.8493 0.0024 0.0055 0.6653 129850
rs1002 C G 0.0306 0.0034 0.0115 0.7659 129799
rs1003 A C 0.5128 0.0045 0.0038 0.2319 129830
```

Columns are SNP identifier, the effect allele, the other allele, frequency of the effect allele, effect size, standard error, p-value and sample size. The headers are not keywords and will be omitted by the program. Important: "A1" needs to be the effect allele with "A2" being the other allele and "freq" should be the frequency of "A1".

The transformation of your summary statistics file to the `.ma` format will depend on the software used to analyse your data. The `ma` subdirectory contains a simple R script for constructing `.ma` files from PLINK 2 `--linear` output. There is no need to filter you summary statistics to match the LD reference as GCTB will perform this data management for you.