Using GWAS summary statistics to construct polygenic scores for testing and prediction

Paul McKeigue
Usher Institute of Population Health Sciences

With help of Athina Spiliopolou and Marco Colombo

Current status of GWAS studies

- Most effects of modest size on easily measurable traits have now been discovered
  - Most journal editors and funding agencies are not interested in publishing new GWAS studies:
- Early optimism about genotypic prediction of clinical outcomes has been tempered by experience
- Challenge is now to make better use of the information available
  - new methods for secondary analysis

Uses of GWAS summary stats

- Construct polygenic scores for prediction
- Construct polygenic scores for hypothesis testing
  - mendelian randomization: test for effect on outcome of scores for intermediate trait
- Interpret novel and established GWAS hits by examining correlations with scores for other traits in same region
- Estimate genetic correlations between traits (cross-trait LD score regression)
- Impute effect size estimates at untyped SNPs
- Evaluate association of genome annotations with SNP effects

Dudbridge (2013): power and predictive accuracy of polygenic risk scores

- Assumes
  - Mixture distribution for effect sizes: spike (SNPs with no effect) and slab (SNPs with nonzero effect)
  - Can select a panel of nearly independent SNPs that explain the genetic variance
  - Threshold model for case-control status: genetic model for underlying “liability”
- Conclusion: sample sizes need to be an order of magnitude larger than current meta-analyses

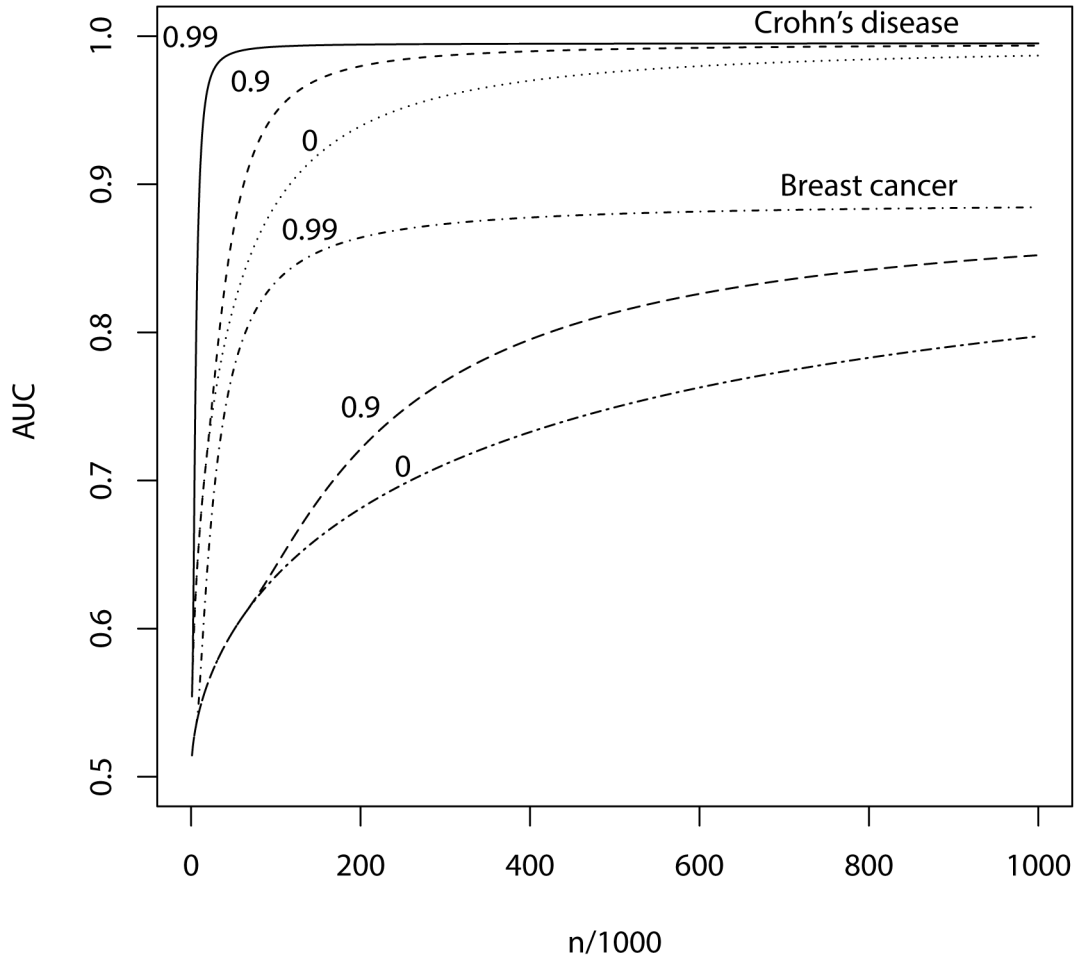
Dudbridge (2013): Relation of predictive performance (AUC) to sample size n

Line labels are proportions of SNPs in the “spike” mixture component (no effect)

Assumptions:

(1) all genetic variation tagged by 1 million independent SNPs

(2) Heritability of liability 0.44 for breast cancer, 0.76 for Crohn’s



Why use genotypic scores rather than SNPs to construct predictors?

- Sample size required to learn a predictor of a clinical outcome directly from SNP associations is unrealistically large for many problems: e.g. prediction of drug response
- Using weighted combinations of SNPs that predict related traits, we obtain a smaller number of features that have higher probability of a nonzero effect.
  - Reduces dimensionality from ~1 million to ~ 20 000
  - exploits meta-analyses based on very large sample sizes

  Region-specific scores facilitate biological interpretation of effects on complex traits
- Genotypic scores can be constructed separately for each region showing association

Constructing genotypic scores from GWAS summary stats

- Can filter SNPs by p-value:
  - optimal threshold is typically much less stringent than that for declaring genomewide significance
  - reduces computation, approximates a heavy tailed prior (e.g. spike and slab mixture)
- Allele score: genotypes weighted by effect sizes scored as +/-1)
- Effect size score: genotypes weighted by effect size estimates
- Multivariate score: allows for covariance between SNPs
- Penalized multivariate regression:
  - ridge regression equivalent to gaussian prior
  - LASSO regression equivalent to double exponential prior

Multivariate regression using univariate effect sizes from GWAS

- Multivariate linear regression coeffs β are estimated as $[\mathbf{X'X}]^{-1}\ [\mathbf{X'Y}]$
- if genotypes and outcome are standardized to zero mean and unit variance, [**X'X**] is the matrix of sample correlations between SNPs and [**X'Y**] is the vector of univariate regression coefficients
- Correlations between SNPs, estimated from a reference panel, can be used to correct the univariate coeffs
  - Adding a positive number to the diagonal of the correlation matrix is equivalent to a penalized regression (ridge regression)
  - Implemented in LDPred
- Almost any multivariate analysis of GWAS data can be approximated using univariate summary GWAS results.
- To exploit this we need is a platform for sharing summary GWAS results

Risk to privacy of study participants from summary-level GWAS data

- Homer et al 2008: tiny correlation of GWAS sample allele freqs with individual's genotypes can be used by an attacker who has access to the individual's genotypes to detect whether that individual was in the sample
- Clayton 2010: derivations based on a gaussian approximation
  - Compared two hypotheses:

    individual is in the case sample (H1) versus individual not in study (H0)
- For P SNPs typed in a case sample of size N, expected *weight of evidence* (log Bayes factor) favouring H1 over H0 is P / 2N natural log units
  - P is usually large compared with N
- Attack requires accurate estimation of population allele frequencies from a reference population matched to the case sample
- In practice, where N > 10 000, uncertainty in the estimates of population allele frequencies limits the information leaked

Availability of summary GWAS data since 2008

NIH and Wellcome withdrew all summary stats from public access

- Managed data access via dbGaP and EGA is permitted, but data cannot be made available to anyone but the approved investigator

- Lumley 2010: recommended “as an interim measure” that no more than 500 SNPs should be shared from a given study.
- Johnson 2011: proportion of GWAS studies providing access to summary stats on > 500 SNPs fell from 20% to 15%
- Most databases of summary-level GWAS data withhold effect sizes
  - GWAS Central, GRASP, GWASdb, Accelerating Medicines Partnership portal
- GWAS summary stats for some privacy-sensitive phenotypes (Psychiatric Genetic Consortium) are freely available
- PloS Genetics instructs authors that GWAS summary statistics for all SNPs “should be available without restriction” but this policy is not enforced

Why address the privacy problem?

- Should not expose patients even to a theoretical risk unless this risk was explained when consent was obtained.
- Even if we think an attack using summary GWAS stats with the individual's genotypes as side information is implausible, we need to be able to reassure research governance bodies that the problem has been addressed
- Existing approaches have limitations:-
  - Adding noise to summary stats protects privacy but reduces predictive performance
  - Platform for generating scores while blocking user access to raw summary stats (? MR-base) does not protect privacy because score itself will leak information if individual was in the case sample
- The same model can be used to calculate both the information leaked to an attacker and the expected predictive performance of a genotypic score

Information leaked by case-control effect size estimates from a GWAS

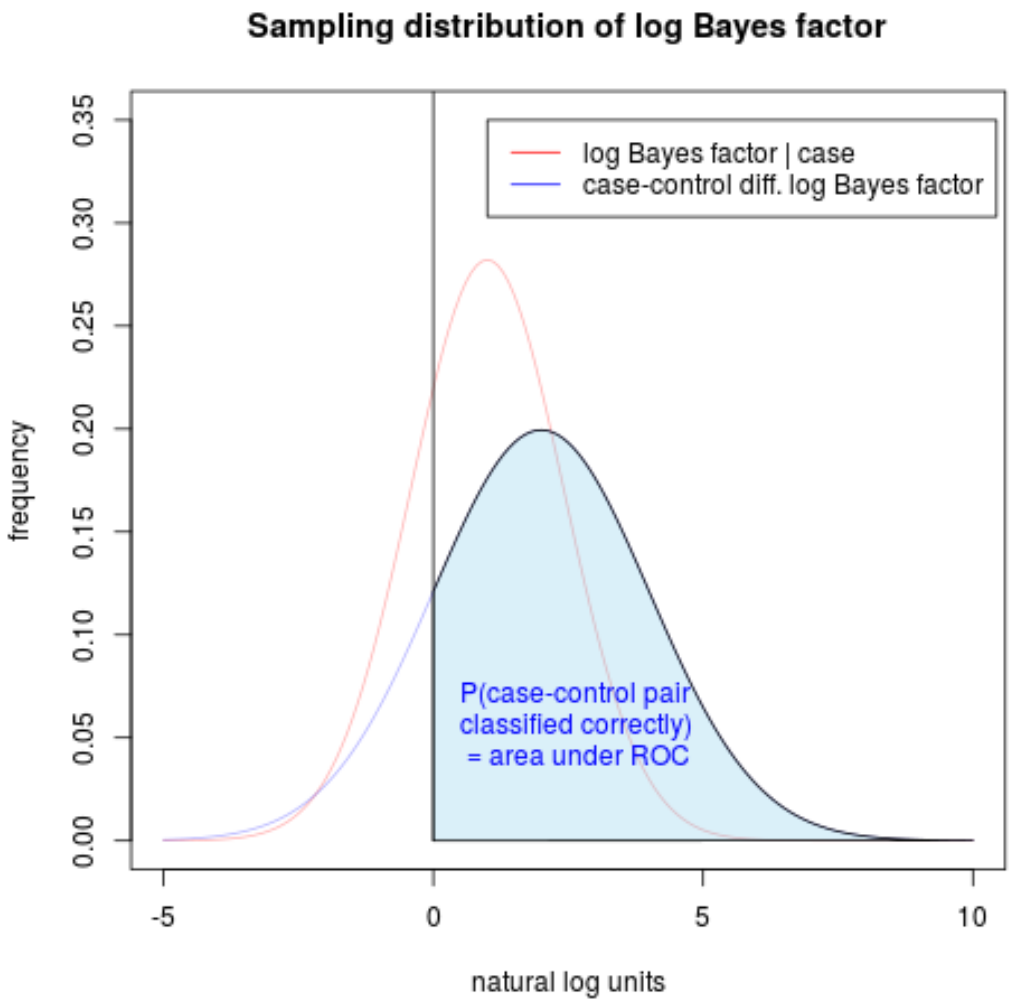
- Prediction of disease status can be calculated from any individual's genotypes if GWAS effect size estimates are available.
- If the individual was in the case sample, this prediction will be more accurate than if the individual was not in the study
  - "training sample" overlaps with test sample
- The *information leak* is the extra information that the attacker gains if the individual agrees to participate in the study

Differential privacy criterion

- Results of any study may allow an attacker to learn something about an individual, given side information
  - for instance unusual facial feature may be identified as a marker of disease
- Differential privacy:
  - information about an individual that attacker can gain from summary stats should not be appreciably more if the individual agrees to participate in the study than it would be if that individual declined to participate.
- To evaluate the extra information gained by the attacker if individual participates, we have to calculate the performance of the genotypic predictor on individuals not seen before.

Information gain and predictive performance

- Weight of evidence favouring H1 over H0 is the log Bayes factor W
- Sampling properties (Turing)
  - gaussian
  - mean = 2 x variance
  - mean $[W(H1/H2)_{H1}]$ = mean $[W(H2/H1)_{H2}]$



Relation of information leak to mutual information between genotypes and summary stats

- 3 hypotheses: H0 individual is not a case and not in study, H1 is a case and in case sample, H2 individual is case but not in study
- Information leak can be calculated as expectation of
- W(H1/H0) given H1 – W(H2/H0) given H2
- = $I(x, b)_{H1}$ - $I(x, b)_{H2}$
- where I(x, b) is the *mutual information* between genotypes x and summary stats b

Information leaked by effect size estimates: gaussian approximation

- If genotypes x and effect size estimates b are multivariate gaussian, the mutual information I(x, b) is half the sum of squared correlations between x_i and b_i
- Mutual information does not depend upon the correlations between SNPs, as long as the correlation matrix is of full rank (no SNPs are redundant)
- Imputed SNPs do not leak any information beyond that contained in the typed SNPs
- Attacker does not need accurate estimates of population allele frequencies
- With a gaussian prior on effect sizes, we can integrate out the effect sizes to calculate the mutual information
- SNPs with extreme p-values leak more information than randomly-chosen SNPs

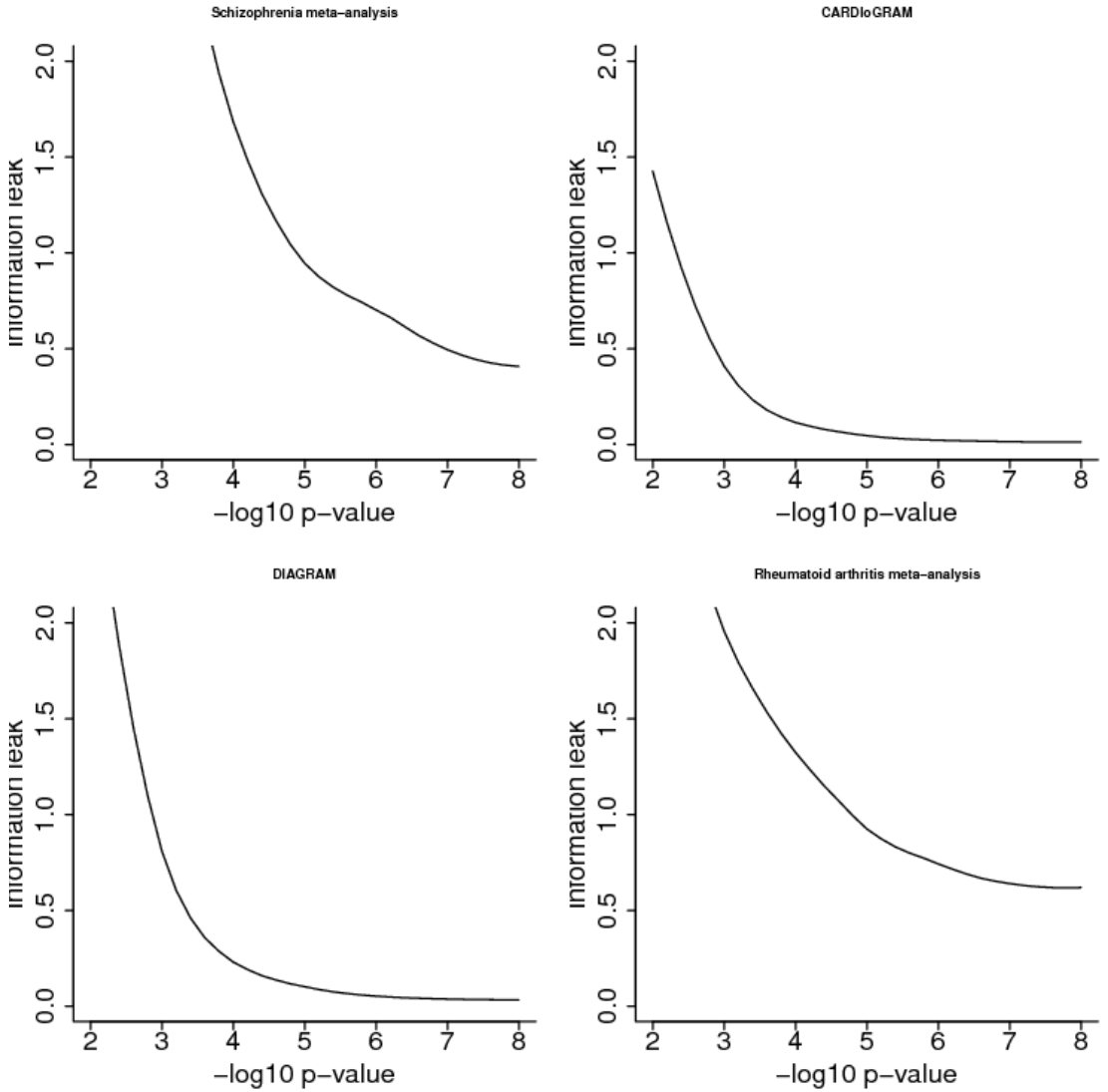
Filtering SNPs to minimize information leakage when sharing GWAS results

- For each phenotype, share only the largest possible meta-analysis
- Calculate information leaked by each typed SNP
- Set a p-value filter that will limit the total information leak, summed over all typed SNPs in or near the filtered regions, to a specified level
  - 1 nat log unit suggested
- Typically with large meta-analyses the threshold p-value determined by this procedure retains >10 000 SNPs and most of the predictive information

Example: filtering four large meta-analyses to keep information leak down to 1 nat log unit

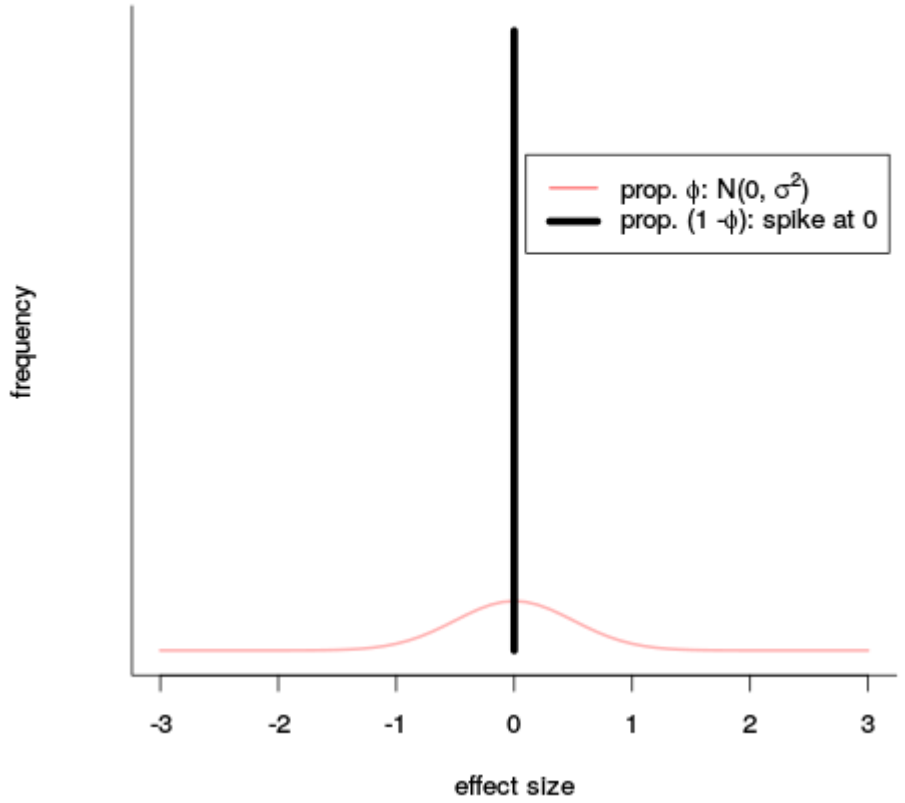
Phenotype	Typed SNPs	Cases /controls	p-value threshold for keeping down info leak	Retained SNPs (including imputed)
Schizophrenia	784K	36989 / 113075	1E-5	34628
Coronary disease	656K	18961 / 57962	0.004	20205
Type 2 diabetes	679K	9739 / 53810	0.001	6522
Rheumatoid arthritis	144K	29880 / 73758	2E-5	31438

Relation of information leak to threshold p-value: four large meta-analyses



Estimating predictive performance of a genotypic from GWAS results

- Realistic prior for effect sizes is “spike and slab” mixture



Estimating predictive performance from a spike and slab model fitted to summary GWAS results

- Spike and slab mixture is specified by two parameters: mixture proportion, variance of gaussian component
- Set these parameters by equating calculated moments (variance and leptokurtosis) of the distribution of effect sizes to observed moments
  - Restrict to typed SNPs
- For given case/control sample sizes of GWAS, can calculate the expected log bayes factor for classifying case-control status of a new individual, based on gaussian approximation to likelihood
- Contrast with Dudbridge's approach: doesn't assume independent SNPs, doesn't assume liability-threshold model but relies on gaussian approximation

Example: colorectal cancer

- Meta-analysis: 445813 typed SNPs in 8323 cases, 9547 controls
- Spike and slab model fitted to effect size estimates: gaussian mixture proportion 0.02, inverse variance of effect sizes 126
- Expected log Bayes factor 0.64 (equivalent to AuROC 0.65) for classifying a new individual using a discriminant learned with this sample size
- Polygenic score derived from meta-analysis calculated in new case-control dataset (SOCCS)
- AuROC 0.56 for prediction of colorectal cancer

Limitations of estimating predictive performance from distribution of effect size estimates

- Simplifying assumptions: no linear dependencies between typed SNPs, correlations of genotypic effects is scalar multiple of correlations of SNPs.
- gaussian approximation is not accurate for mixture distribution of SNP effect sizes
  - Especially for autoimmune disease such as rheumatoid arthritis with very large effects in HLA gene region
- Fitting spike-and-slab mixture parameters depends critically on well-calibrated p-values. Variance of test statistic may be
    - **inflated** by inadequate control of population stratification, relatedness, batch effects in genotyping
    - **deflated** by poorly calibrated (over-confident) probabilities for imputed genotypes

Some implications of the model for predictive performance

- All predictive information is contained in the typed SNPs – imputation may detect interesting effects of imputed SNPs but cannot add to prediction
- For a given total size of genetic effect, learning predictive models is easiest for outcomes where the proportion of SNPs with nonzero effects is small.
- For most complex traits, learning predictive models directly from SNP associations will require much larger sample sizes than current meta-analyses have

GENOSCORES platform

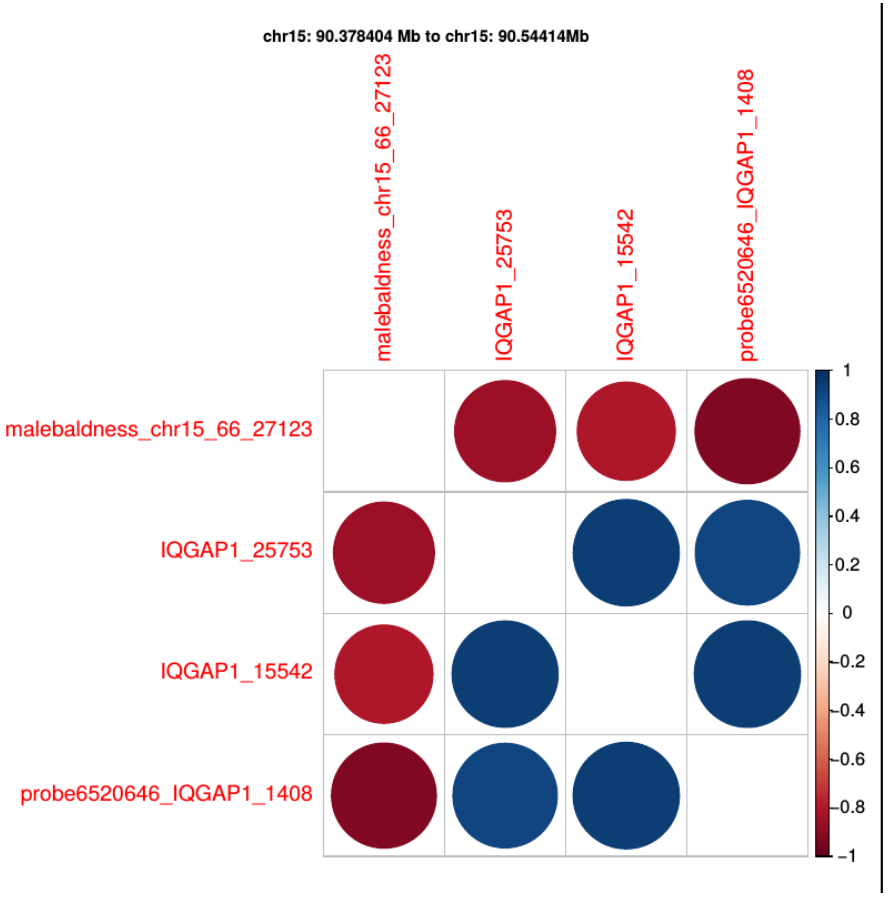
https://pm2.phs.ed.ac.uk/genoscores/

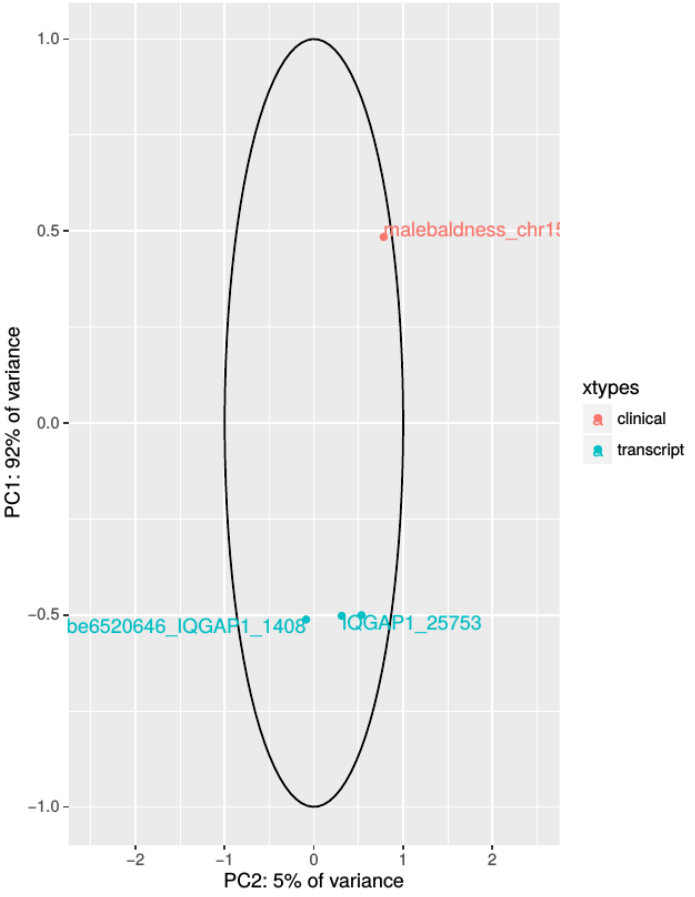
- Database holds effect size estimates for 22003 phenotypes from 57 studies, filtered at p < 10-5
- Trait types: transcript levels (20646), micro-RNA, metabolites, immune cell traits, glycans, proteins, clinical
- R script running on the user's server queries the database and computes genotypic scores on the user's genotype dataset
  - genome build, SNP IDs and allele coding are cleaned up automatically
  - user can specify calculation of global or regional-specific scores, with or without LD adjustment
- Other scripts fit cross-validated predictive models and plot score correlations

Similar platforms

- Bristol: http://www.mrbase.org/ MR-base for 2-sample mendelian randomization
  - currently under development
  - will use LD Pred to calculate scores corrected for LD
  - cannot use summary stats filtered by p-value
- Zhu Z et al 2015: summary data–based Mendelian randomization (SMR)
  - results summarized in database SMRdb
  - Integrates complex trait GWAS summary results with eQTL summary stats
  - Uses multiple SNPs in each cis-eQTL region to distinguish shared haplotype from causal effect of transcript on trait

Using GENOSCORES to identify transcript scores that correlate with a complex trait GWAS hit

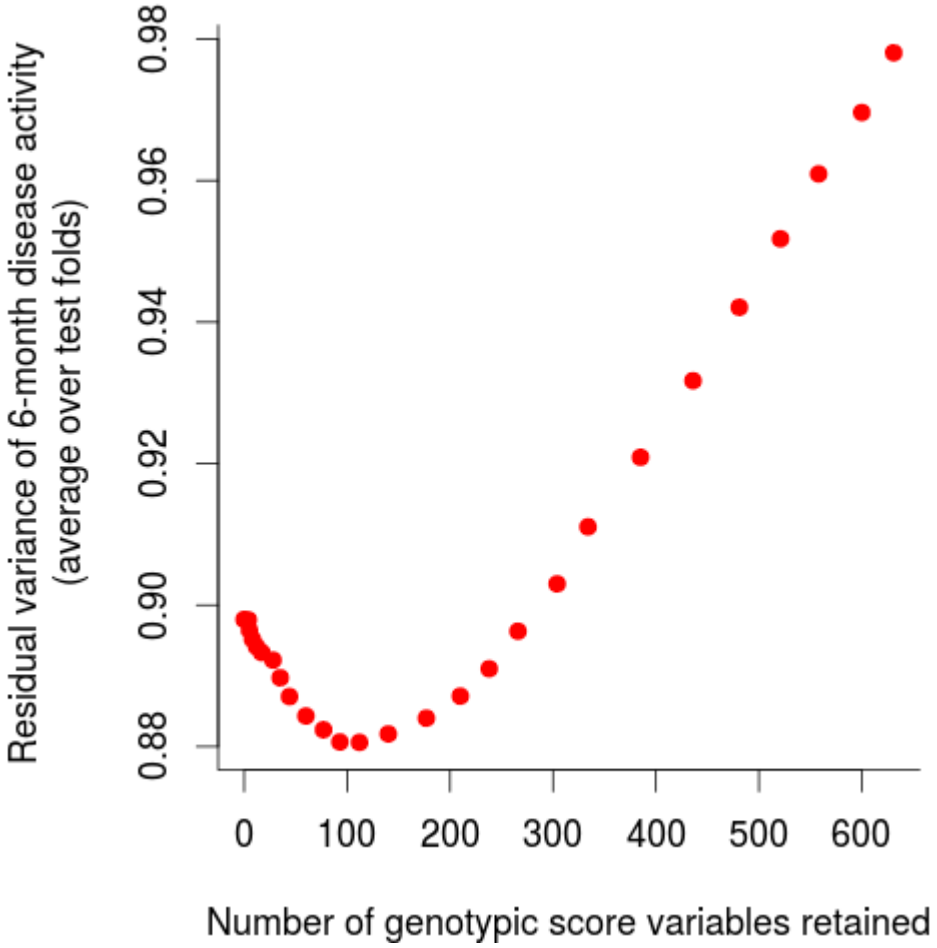




Relation of anti-TNF response in rheumatoid arthritis to genotypic scores: BRAGGSS study

- 1566 individuals with rheumatoid arthritis treated with anti-TNF agents for first time
- Outcome measure: change in disease activity score (swollen joint count + ESR), adjusted for baseline covariates
- LASSO regression model fitted by 50-fold cross-validation
  - Optimal penalty retains 112 genotypic score variables, explaining 1.9% of residual variance in test folds
  - Gain in test log-likelihood 13.6 nat log units, equivalent to $p = 2 \times 10^{-7}$)

Prediction of change in 2-component disease activity score from 12752 genotypic scores



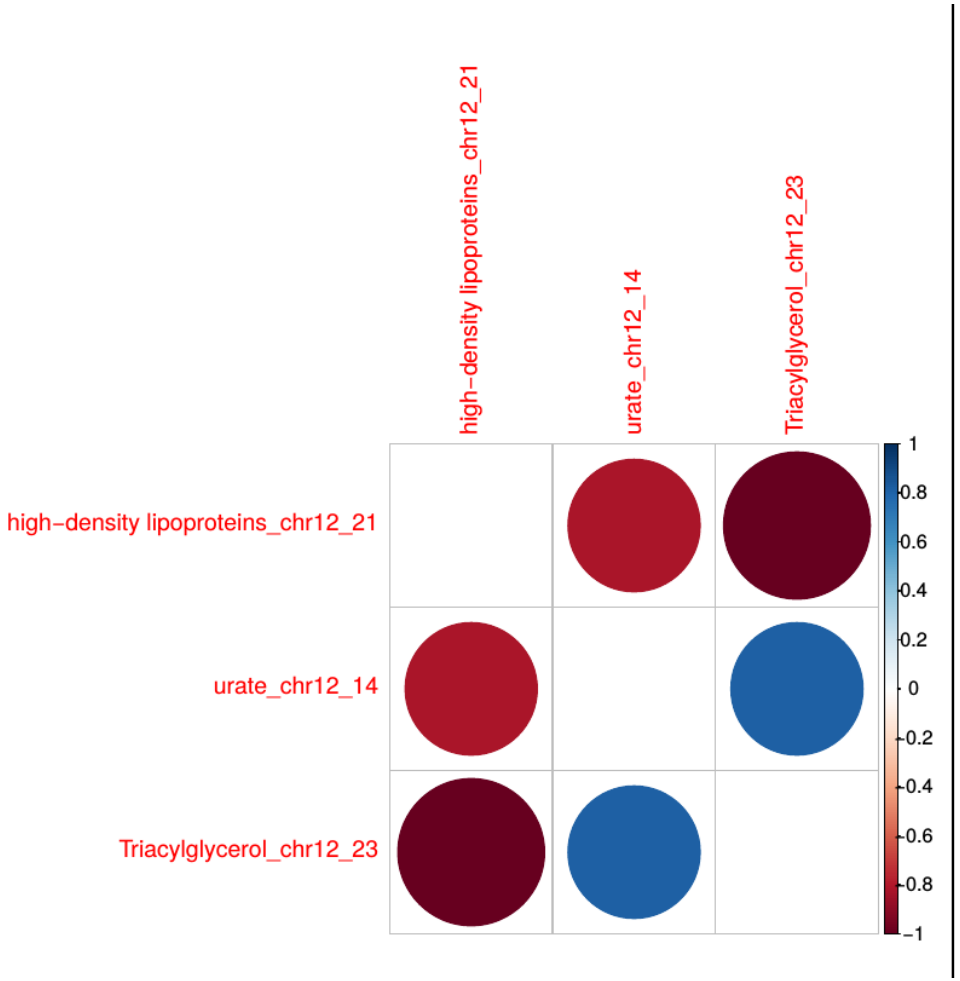
Top four genotypic scores in univariate analyses

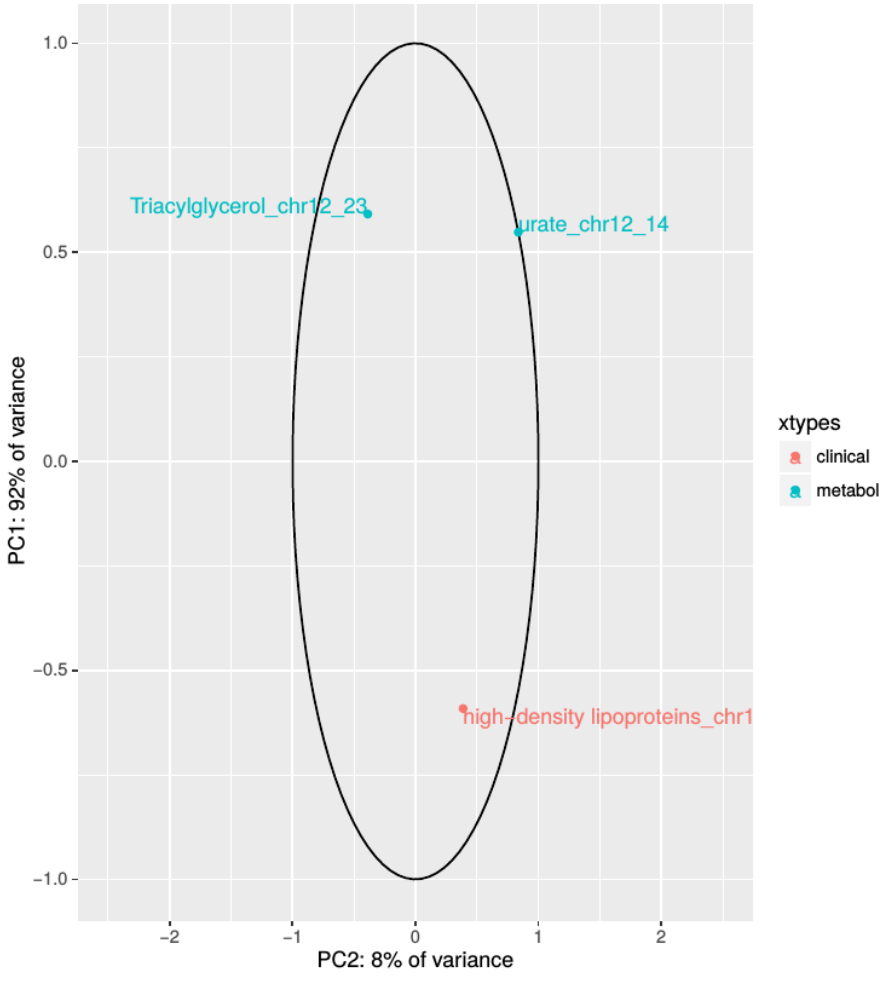
Trait	Gene Name	Location	Function
MR1 transcript	Major histocompatibility complex, class I-related	1q25.3	Antigen-presenting molecule specialized in presenting microbial vitamin B metabolites
DAP transcript	Death-associated protein 1	5p12.2	mediator of programmed cell death that is induced by interferon-gamma
CCDC163P transcript		1p34.1	
Scores for RA, psoriasis, Crohn's in *CDC37* region	Hsp90 co-chaperone Cdc37	19p13.2 (*ICAM3* in same region)	Toll-like receptor mediated macrophage activation

Using genotypic scores to test for common pathway between two or more traits

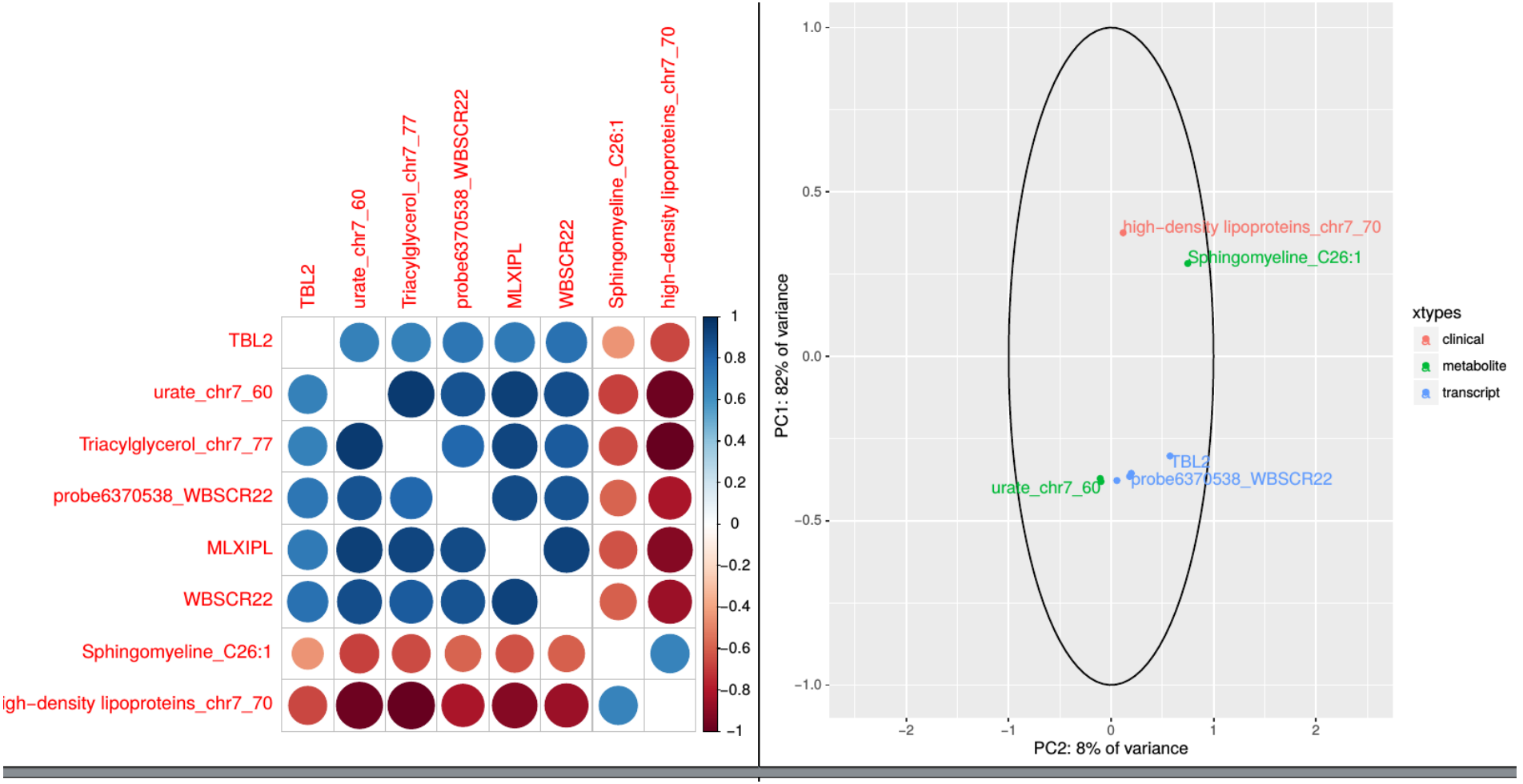
- Region-specific genotypic scores for 2 or more traits may be correlated because
  - Either (1) SNPs influencing them are on same haplotype even though their effects are mediated by different transcripts
  - Or (2) SNPs act through common pathway (same transcripts)
- Can try to distinguish these by modelling joint associations of SNPs (Zhu et al 2015) but correlations may be too strong for effects on different traits to be distinguished.
- If region-specific scores for trait A are associated with region-specific scores for trait B in *multiple regions* containing different gene clusters, this is strong evidence for a common pathway

Correlation of regional scores: urate with lipids on chr 12

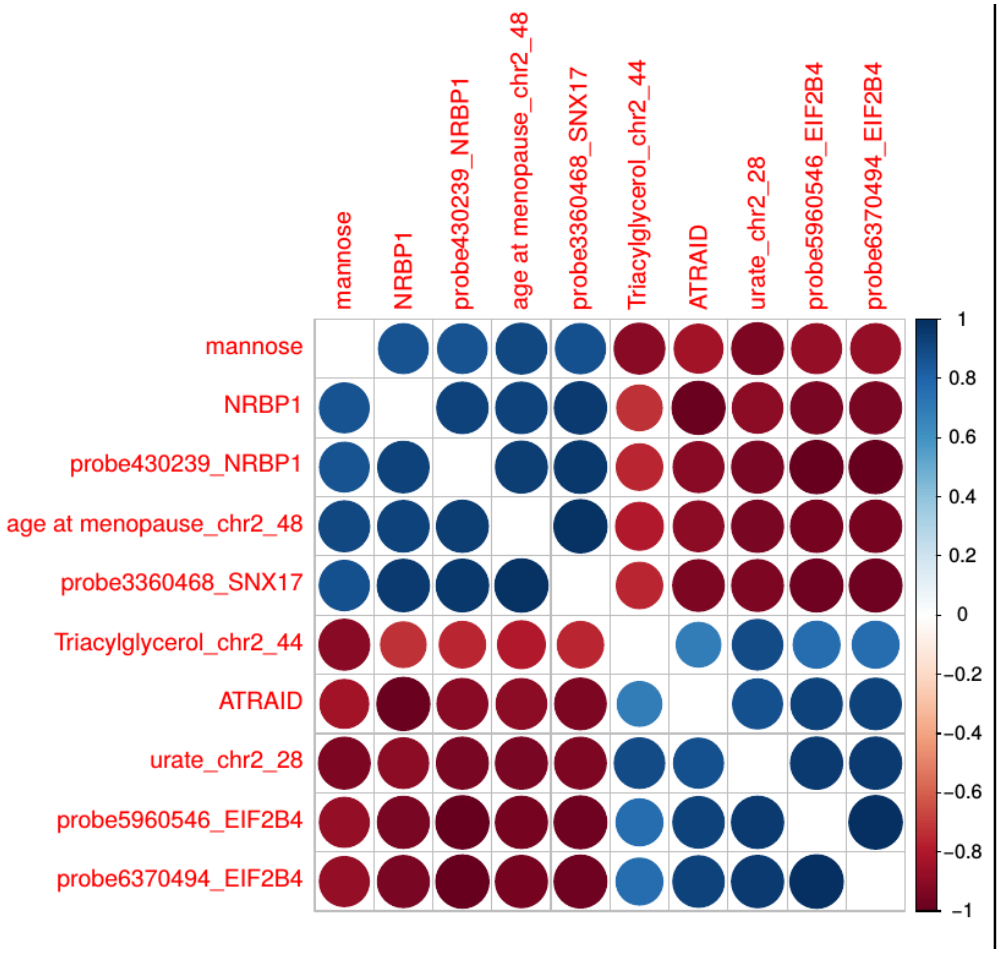


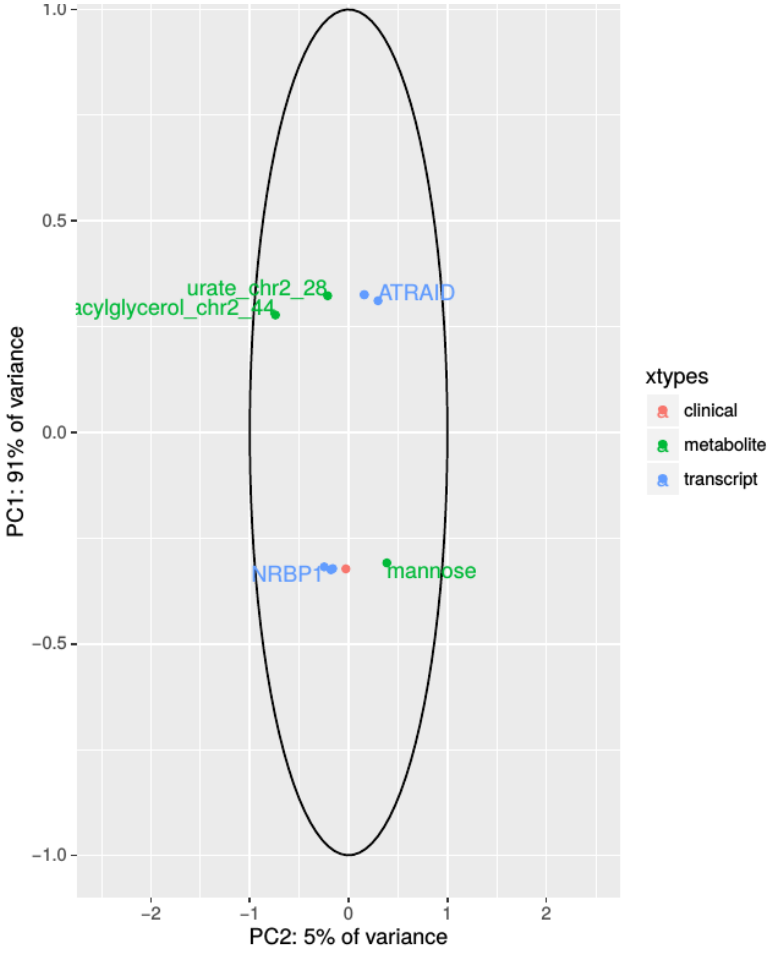


Correlation of regional scores: urate with lipids on chr 7



Correlations of regional scores: urate with triglyceride on chr 2





Future development of GENOSCORES platform

- Increase GWAS coverage:
  - Currently lacks GWAS effect sizes for cancer, serum proteins, DNA methylation
- Scale up platform to support more users with more efficient computation
- Develop methods to
  - Identify and visualize novel correlations
  - Identify and visualize evidence for novel common pathways / causal relations