Genetic epidemiology: association and prediction

- Uses of genetic association studies
- Genome-wide association studies
  - Cross-sectional, case-control, cohort
  - Haplotype structure, tag SNPs and imputation
  - Controlling for population stratification
  - Criteria for declaring association
- Why do confirmed SNP associations explain only a small fraction of the genetic variance ("missing heritability")?
- Prediction of clinical outcome from genotypes

Uses of genetic association studies

- Understanding molecular basis of disease
  - But molecular mechanism of a SNP-disease association may not be obvious: variant may alter transcription at distant site
- Prediction of disease risk or response to treatment
  - prediction of complex traits from genotype is difficult
- Inferring causal relationships by exploiting genotype as a randomized "instrument"
  - Requires a genotypic predictor of at least modest effect

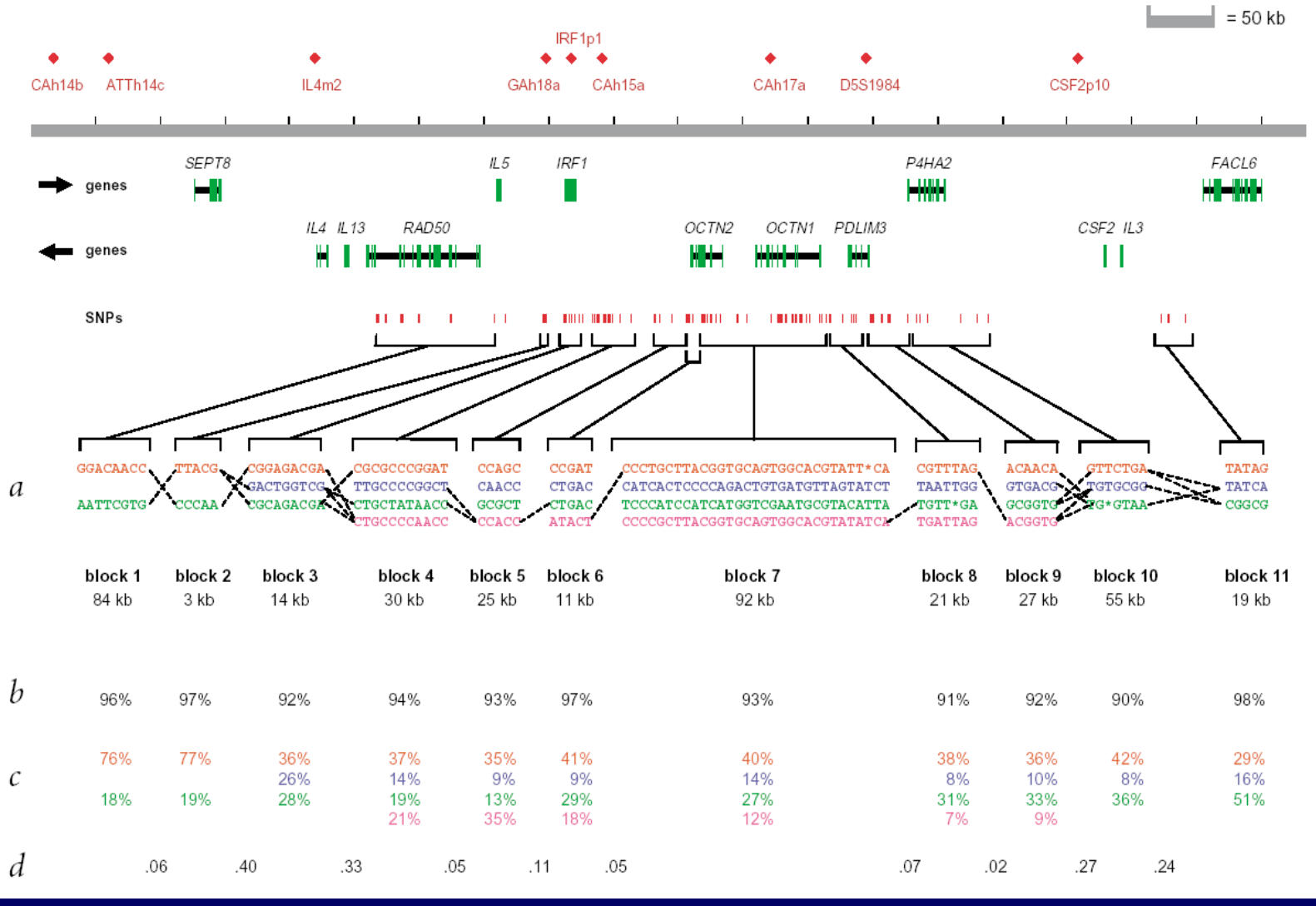
Methodological differences between studies of genetic and environmental risk factors

Genetic factors	*Environmental factors*
Bias in measurement of exposure (genotype) is eliminated in correctly-designed studies	Bias in measurement of exposure (e.g. dietary intake) is often difficult to eliminate
Exposure (genotype) can be measured retrospectively: case-control designs are usually the approach of choice	Measurements of exposure may be affected by disease onset: prospective cohort studies are often necessary
Selection bias less serious: can usually be dealt with by controlling for population stratification	Selection bias can be difficult to eliminate, especially in case-control studies
Only possible confounders of an association with genotype are:- (i) population stratification – which can be controlled in the design or analysis (ii) haplotypes – can exploit this confounding using tag SNPs	Confounding is a serious problem – often impossible to control adequately

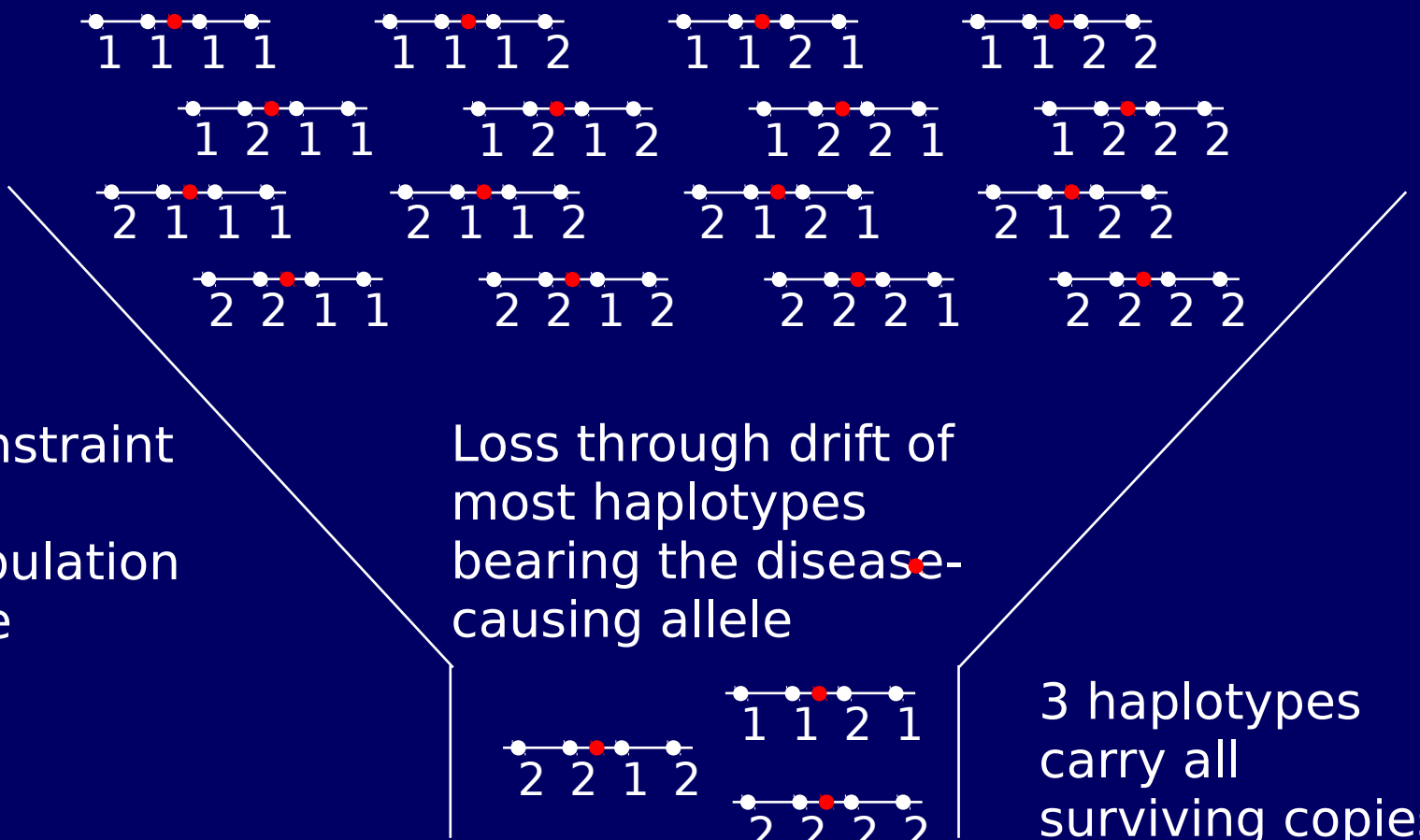
Design of genetic association studies

- Case-control study: most powerful design for studying genotype-disease associations
  - Allows accurate outcome classification: e.g. can ensure that stroke cases have CT scans to classify subtype
- Cross-sectional study: can study biomarkers and other quantitative traits
- Cohort study: can model genotype, biomarkers and disease jointly
  - Not powerful enough for discovery of genotype-disease associations
  - Can model genotypes, biomarkers and outcome jointly

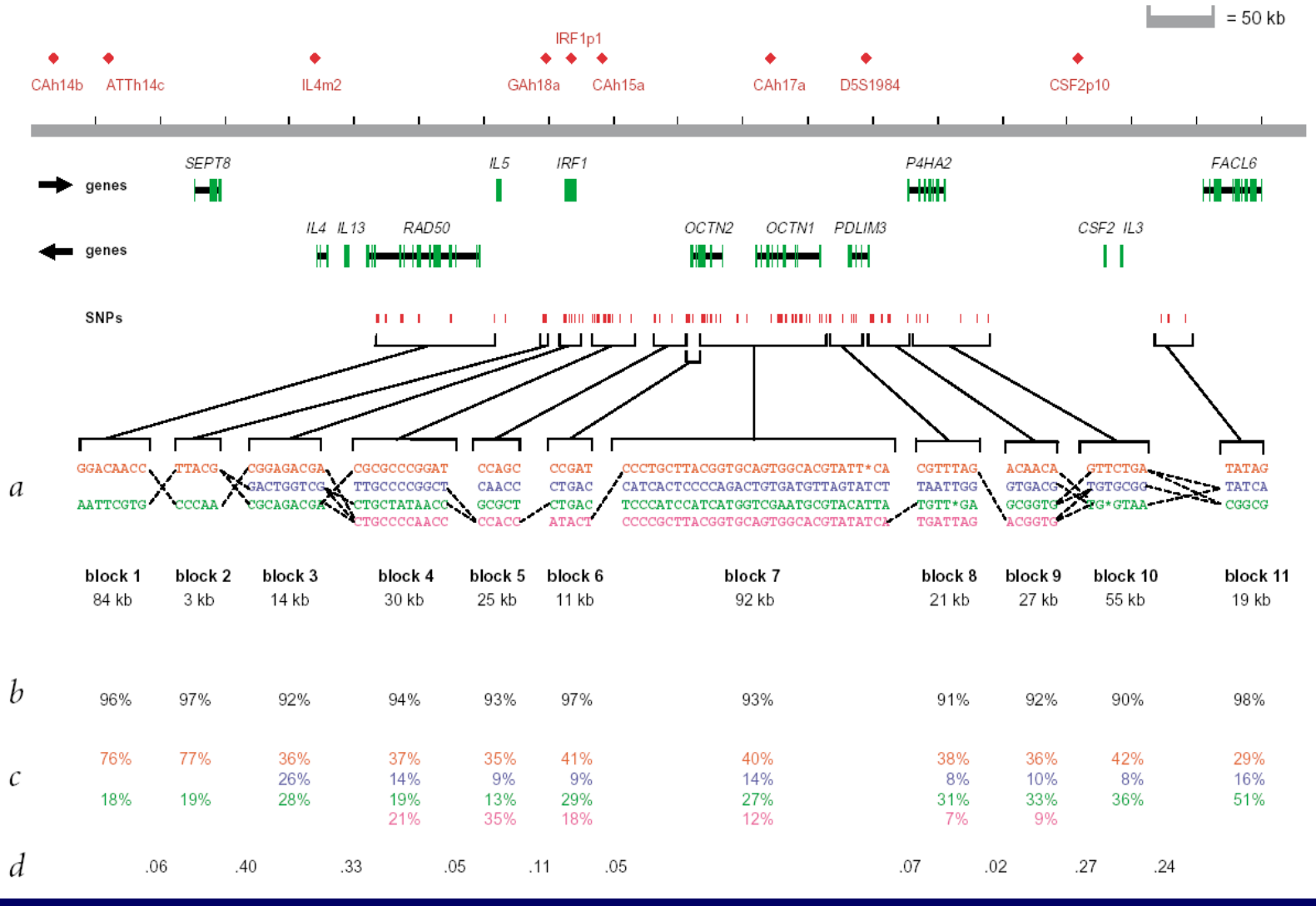
Haplotype structure: example



How constraint of population size leads to association of disease with haplotypes



Haplotype structure: example



Haplotype reference panels

- HapMap: ~3 million common SNPs typed in samples from three continental groups
- 1000 Genomes project: deep sequencing on 2504 individuals from 14 populations across 5 continents
  - Each individual's genome is repeatedly chopped into short (~200 bp) near-random fragments that are sequenced and aligned to reference sequence
  - combines variant discovery and genotyping
  - ~ 30 million common SNPs typed

Imputation of common 1000G SNPs given genotypes of tag SNPs or low-coverage sequencing

- Genotyping arrays include up to 1 million tag SNPs
  - tag SNPs are those that "tag" other common SNPs but are less strongly associated with each other
    - Illumina: 700K SNPs
    - Affymetrix: 1 million SNPs
    - genotyping cost ~ £50 per individual
- Can use chip genotypes with reference haplotypes from HapMap or 1000G to impute genotypes at untyped SNPs

Practical steps in designing a genome-wide association study

- Define the traits to be studied
  - Can use genetic relationship matrix to define a phenotype or sub-phenotype that has high proportion of variance explained by common SNPs
- Define the sampling protocol: case-control, cross-sectional or (rarely) cohort, consent, data sharing agreements
- Type a genome-wide SNP array: usually at least 500K SNPs
  - cost now ~ £50 per chip

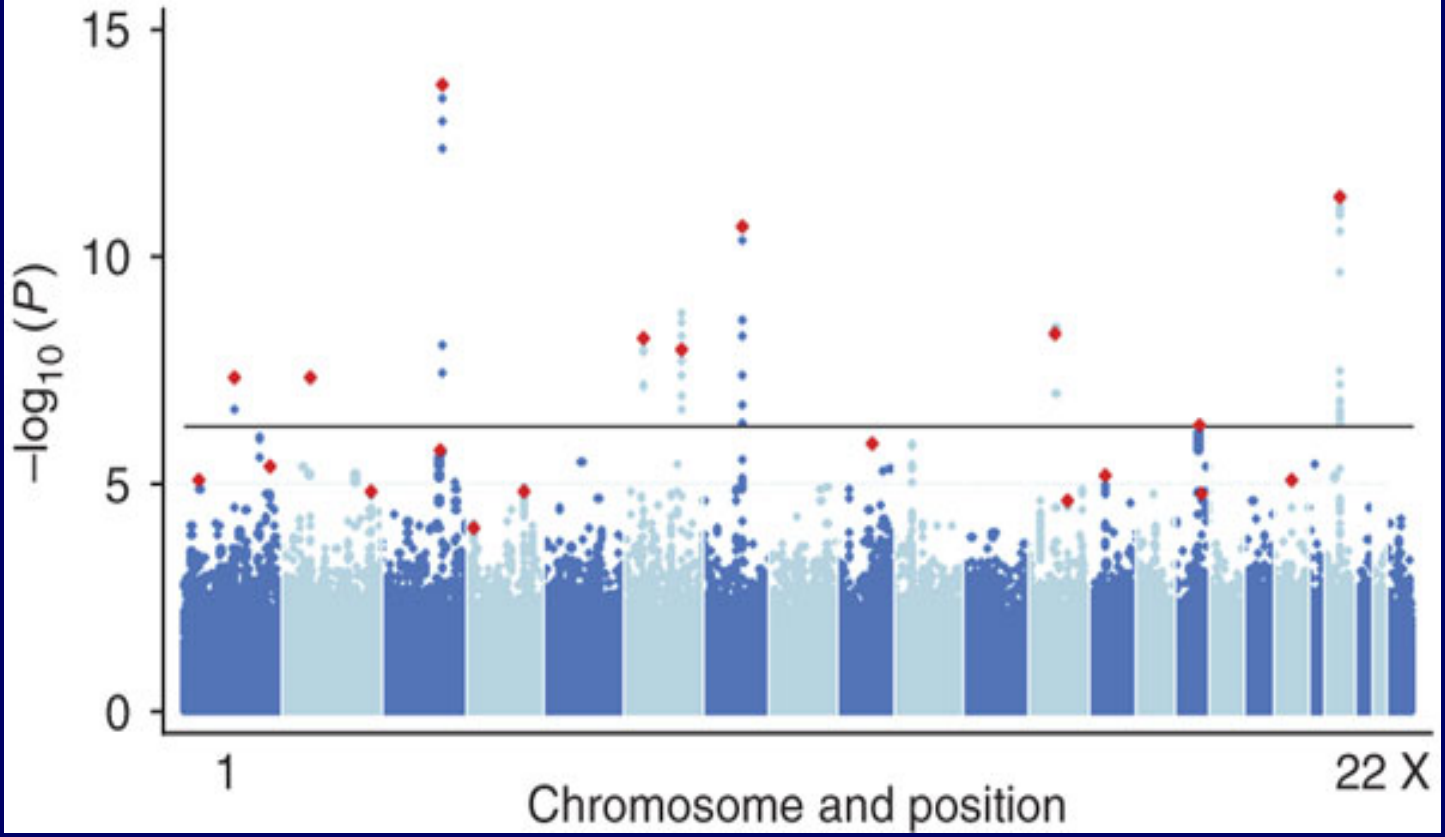
Practical steps in preparing genome-wide association data for analysis

- Data warehousing
  - use a relational database for clinical data, variant call format for genotypes
- Drop SNPs with poor quality scores
  - proportion of missing values, Hardy-Weinberg equilibrium
- Score individuals for genetic background
  - Use principal components analysis with thinned SNPs to derive first few principal components as covariates to control for population stratification

Analysis of genome-wide association studies

- Impute genotypes at untyped loci using 1000 Genomes as reference panel (MACH, IMPUTE)
- Test imputation of SNPs with low minor allele freq using a subset of typed loci set to missing
- Test SNPs for association one at a time, adjusting for population stratification
  - QQ quantile plot to see if variance of test statistical is inflated
  - Manhattan plot is a useful visual summary

Manhattan plot for 402,951 SNPs from meta-analysis of genome-wide association studies of adult height



Alternatives to univariate SNP-based analyses

- Gene-based tests
  - Intragenic SNPs annotated by gene name
  - can test each gene for association/enrichment
- Pathway-based tests
  - Pathway databases (e.g. KEGG) can be used to annotate genes and thus intragenic SNPs
  - Can test each pathway for association/enrichment
- Multivariate tests
  - where you have multivariate outcome measurements

What do you do with a GWAS "hit"?

- Annotate the associated SNPs
  - which genes are nearest?
  - do any of the associated SNP have predicted effect on protein sequence or expression?
  - are any of the SNPs eQTLs?
- Study possible functional effects of nearby genes
  - expression in disease states
  - knockouts / transgenic animals
- Study effect of SNPs that predict disease on intermediate phenotypes / biomarkers

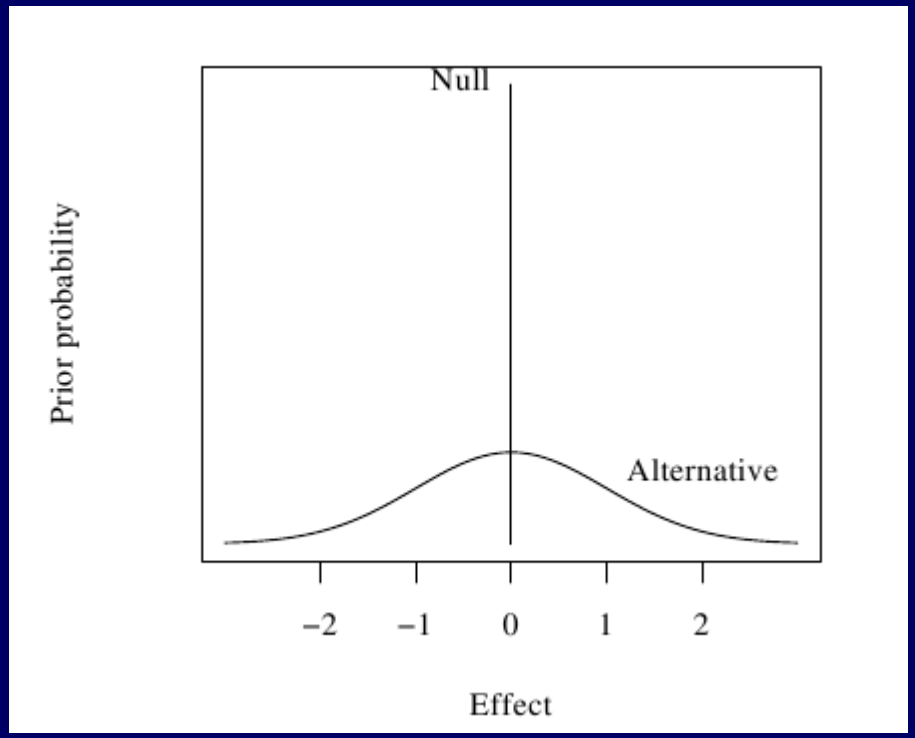
What strength of evidence is required to declare a genetic association?

- Classical approach: correct for multiple tests
  - For 1 million independent tests, reduce threshold *p*-value by factor of 1 million to 5 x 10^{-8} (Bonferroni correction)
- Bayesian argument:
  - Posterior odds = likelihood ratio x prior odds
  - Prior odds that the SNP is associated with disease are very low (~ 10^{-5})
  - For posterior odds of 10 to 1, we should require Bayes factor (likelihood ratio) of 10^{6}

Problems with correction for multiple tests

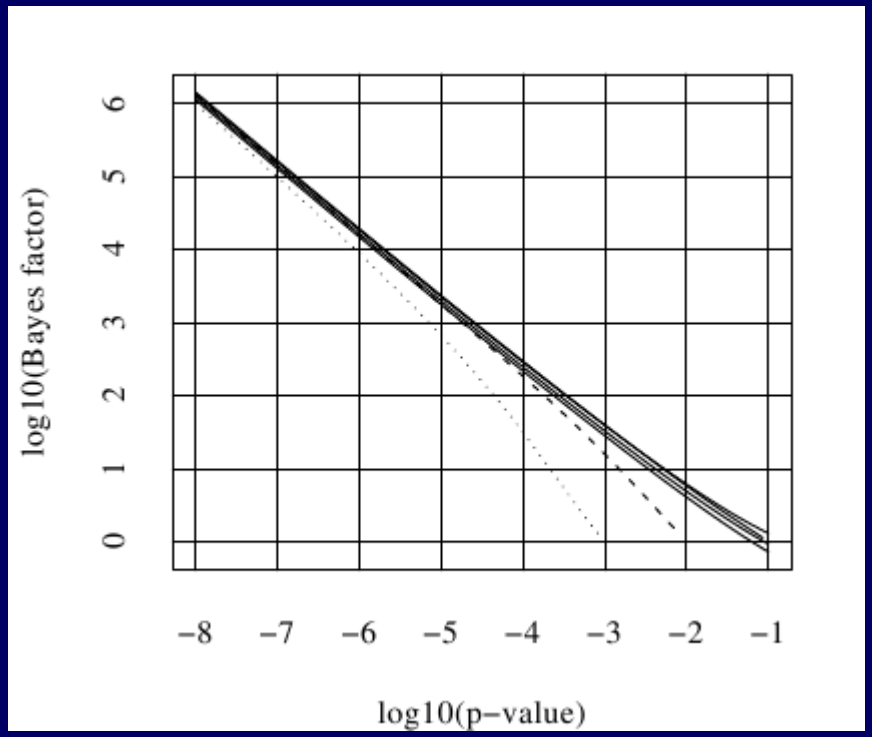
- How many tests were done?
  - number of tests reported in this paper
  - number of tests that could have been done in this study
  - number of tests that could have been done in all studies of this outcome (assuming that only positive tests are reported)
- False discovery rate
  - estimate proportion of true positives assuming a uniform distribution of *p*-values under the null

Bayesian hypothesis testing requires us to specify a prior on effect size



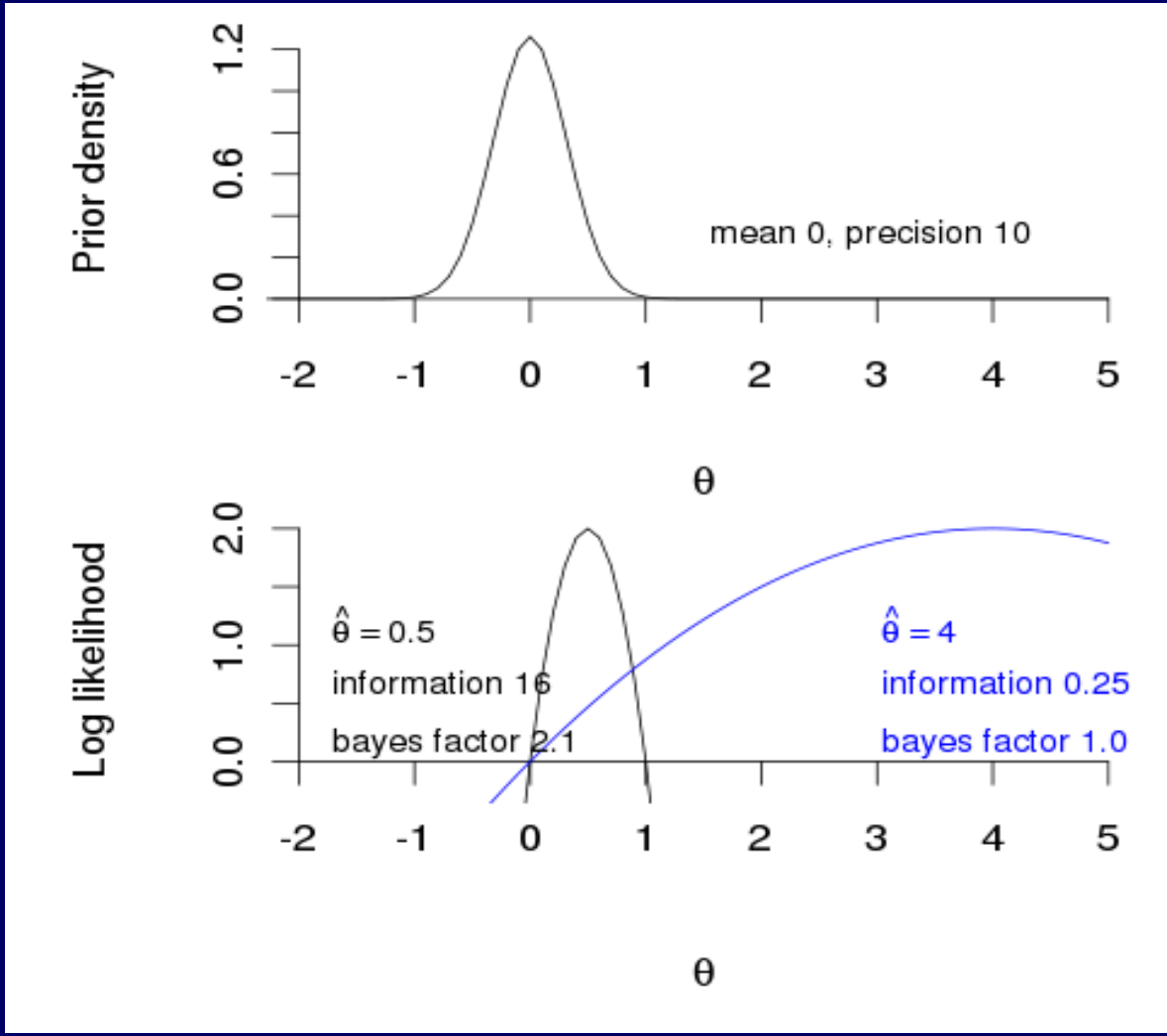
- Null hypothesis H_0: prior on effect size is spike at 0
- Alternative hypothesis H_1: gaussian prior, SD based on belief about plausible effect sizes

Clayton 2003: relationship between p-value and Bayes factor (gaussian prior on effect size under non-null hypothesis H_1)



- Lines are for powers of 0.001, (dots), 0.01 (dashes), 0.1, 0.5, 0.9 and 0.99 (lines) to detect effect of size expected under prior
- Severely underpowered studies require smaller -values to convince

How Bayes factor penalizes implausibly large effects in underpowered studies



Bayesian interpretation of *p*-values

- Given a positive result in a diagnostic test likelihood ratio = sensitivity / (1 – specificity)
- Significance test can be viewed as a diagnostic test:
  - threshold *p*-value = 1 – specificity
  - power to detect effects of plausible size = sensitivity
  - Likelihood ratio = power / threshold p-value
- *p*-values are misleading if study is underpowered to detect effects of plausible size

How much of the genetic effect on complex diseases and traits is explained by known variants?

- Heritability of a continuous trait
  - = genetic variance / total variance of trait
  - Total heritability = sum of locus-specific contributions
- Sibling recurrence risk ratio for a binary (disease) trait
  - = risk to sibling of case / risk in general population
  - Total sibling risk ratio = product of locus-specific contributions

Established Type 2 diabetes susceptibility loci (McCarthy MI 2009)

Index SNP	Chr	Position	Region / gene	Identification	λ_s*
rs10010131	4	6343816	*WFS1*	Candidate gene	1.004
rs1801282	3	12368125	*PPARG*	Candidate gene	1.005
rs757210	17	33170628	*HNF1B (TCF2)*	Candidate gene	1.002
rs5219	11	17366148	*KCNJ11*	Candidate gene	1.005
rs7901695	10	114744078	*TCF7L2*	Linkage peak fine-mapping	1.022
rs10811661	9	22124094	*CDKN2A/B*	GWA	1.003
rs10946398	6	20769013	*CDKAL1*	GWA	1.002
rs13266634	8	118253964	*SLC30A8*	GWA	1.003
rs4402960	3	186994381	*IGF2BP2*	GWA	1.002
rs5015480	10	94455539	*HHEX/IDE*	GWA	1.002
rs8050136	16	52373776	*FTO***	GWA	1.009
rs2237892	11	2796327	*KCNQ1*	GWA	1.031
rs10830963	11	92348358	*MTNR1B****	GWA	1.001
rs10923931	1	120319482	*NOTCH2*	GWA meta-analysis	1.001
rs12779790	10	12368016	*CDC123/CAMK1D*	GWA meta-analysis	1.002
rs4607103	3	64686944	*ADAMTS9*	GWA meta-analysis	1.002
rs7578597	2	43586327	*THADA*	GWA meta-analysis	1.002
rs7961581	12	69949369	*TSPAN8/LGR5*	GWA meta-analysis	1.001
rs864745	7	28147081	*JAZF1*	GWA meta-analysis	1.001

Why is most familial aggregation still unexplained by known variants? height as an example

- In populations where childhood malnutrition is rare, heritability of adult height is 0.8 – 0.9
- GWAS have identified 40 loci that affect height, but their total effect is < 5% of population variance
  - All effects discovered so far are very small

Possible explanations for missing heritability ("dark matter")

- Epistatic effects (gene-gene interactions)
- Copy number variants not detected by standard genotyping assays
- Rare variants with large effects, not detected by tag SNP mapping because of allelic heterogeneity
- Polygenic model
  - many common variants of tiny effect, undetectable at genome-wide significance even in large meta-analyses

How much genetic variance is accounted for by epistasis?

- Contribution of epistasis can be estimated by comparing concordance rates for monozygotic twins and parent-offspring pairs
  - if all genetic effects are additive, MZ twin concordance is twice parent-offspring concordance
  - Only for a few disorders (notably schizophrenia) is MZ twin concordance much higher than predicted from additive model
- Evolution selects only additive genetic effects

How much unexplained genetic variance is accounted for by copy number variants?

- Copy number variation: duplication or deletion of all or part of a gene
  - Not detectable by standard genotyping assays
  - run of SNP genotypes with increased signal of one allele may indicate duplication
  - run of homozygous SNP genotypes with reduced signal may indicate deletion
- Systematic studies of copy number variation don't support hypothesis that this accounts for a high proportion of genetic variance

Example of rare variant effects: hypertriglyceridemia (Johanson 2010)

- 4 top hits in GWAS: *APOA5, GCKR, LPL, APOB*
- These 4 genes resequenced in 438 cases and 327 controls
- Carrier frequency of rare variants: 28% in cases, 15% in controls

Fisher (1918): genetic variance can be estimated from correlation of "genes" between relatives

- Additive polygenic model
  - Genotypic value (X, expected value of trait given genotype) = population mean + sum of many small effects at loci that segregate independently
  - genetic effects (g_{1i}, ..., g_{Ti}) are standardized to zero mean
  - Trait value Y is X plus random environmental effect
- $Y_i = X_i + \varepsilon_i = \mu + g_{1i} + ... + g_{Ti} + \varepsilon_i$

Correlation between relatives (assuming no environmental covariance) is the correlation of their genetic effects

- $Y_i = X_i + \varepsilon_i = \mu + g_{1i} + ... + g_{Ti} + \varepsilon_i$
- $\mathrm{Cov}(Y_i, Y_j) = \langle Y_i\ Y_j \rangle$
- $= \langle g_{1i}\ g_{1j} \rangle + \ ... + \langle g_{Ti}\ g_{Tj} \rangle$
  - equivalent to correlation coefficient if variances of effects are scaled to sum to 1
  - geometric interpretation: dot product of two vectors measures how similar they are (cosine of angle between vectors)

Genetic variance inferred from correlation between relatives

- Fisher showed that
- $Cov(X_1, X_2) = R V_A$ where V_A is additive variance, R is relationship coefficient
- Relationship coefficient is proportion of genome identical by descent
  - 1 for MZ twins, 0.5 for parent-offspring, 0 for unrelated individuals
- Fisher showed good fit to studies of height: correlation between relatives is proportional to degree of relationship

Estimating genetic variance from unrelated individuals

- Additive genetic variance can be estimated from the relationship of trait covariance to the relationship coefficient
- Relationship coefficient can be computed directly from SNP genotypes as the correlation of genotypes between individuals
- These correlations can also be computed for unrelated individuals

Interpretation of genetic variance estimated from SNP relationship matrix

- For related individuals, matrix of relationship coefficients can be calculated either from pedigree or from SNP genotypes
  - additive variance estimate includes effects of common and rare variants
- For unrelated individuals, relationship matrix can still be computed from SNP genotypes but most off-diagonal coefficients will be close to 0
  - additive variance estimate excludes effects of rare variants

How much genetic variance is explained by common SNPs? (Yang 2010, many subsequent papers from same group)

- ~4000 individuals typed with ~300K SNPs
- SNP genotypes used to calculate "relationship matrix" between apparently unrelated individuals
  - SNP relationships are represented in a matrix of correlations between genotypes of persons
  - diagonal elements are 1, off-diagonal elements have mean zero
  - coefficients between unrelated individuals typically vary between -0.02 and +0.02
- Genetic variance estimated from dependence of pairwise concordance on relationship coefficient

How much variance in height is explained by common SNPs? (Yang 2010)

- Additive genetic variance of height is ~ 80% of total variance
- Associations with 50 SNPs that meet genome-wide significance account for ~5% of variance.
- SNP relationship matrix shows that GWAS SNPs account for 45% of variance
- Allowing for incomplete tagging of causal variants, SNPs may explain 54% to 84% of total variance.

How is genetic variance estimated from SNP relationship matrix?

- Crude method: regress squared difference in trait values on pairwise relationship coefficients
  - Genetic variance is minus half the regression slope
- More advanced method: fit a model in which the trait values arise from a multivariate gaussian distribution with covariance matrix proportional to the SNP relationship matrix
  - Software is available (e.g. GCTA)

Can genetic variance explained by common SNPs be partitioned?

- Yang 2011: height, body mass index, vonWillebrand factor, Q-T interval in ~12000 individuals typed with ~570K SNPs
- Variance explained by SNPs on each chromosome scales with length of chromosome
- Intragenic SNPs (50% of panel) account for 3 x variance explained by intergenic SNPs

Implications of SNP genetic variance studies

- GWASs to discover novel associations with SNPs will have diminishing returns
  - "Missing heritability" isn't really missing: undermines case for resequencing studies to discover rare variants of large effect
- Intermediate pathway effects may be sparser than genotypic effects
- Prediction from SNP genotypes is possible in principle, but learning models from genotype-outcome associations alone will require very large sample sizes

How much genetic variance is accounted for by common variants with weak effects?

- Best prediction of outcome in a test dataset is improved by allowing the model to retain many SNPs with small effects that are below the conventional threshold for genome-wide significance
- Using SNP genotypes to estimate genetic variance explained by common SNPs

Prediction from high-dimensional data

- Dimensionality reduction where many variables are correlated
  - Principal components analysis (PCA)
    - Can use supervised or sparse PCA to select relevant variables
- Non-parametric methods
  - learn a function (kernel) that evaluates the similarity between pairs of observations
- Sparse priors (e.g. LASSO regression)
  - encode prior belief is that effects are mostly small or near zero
  - sparsity parameter can be learned from the data

LASSO regression

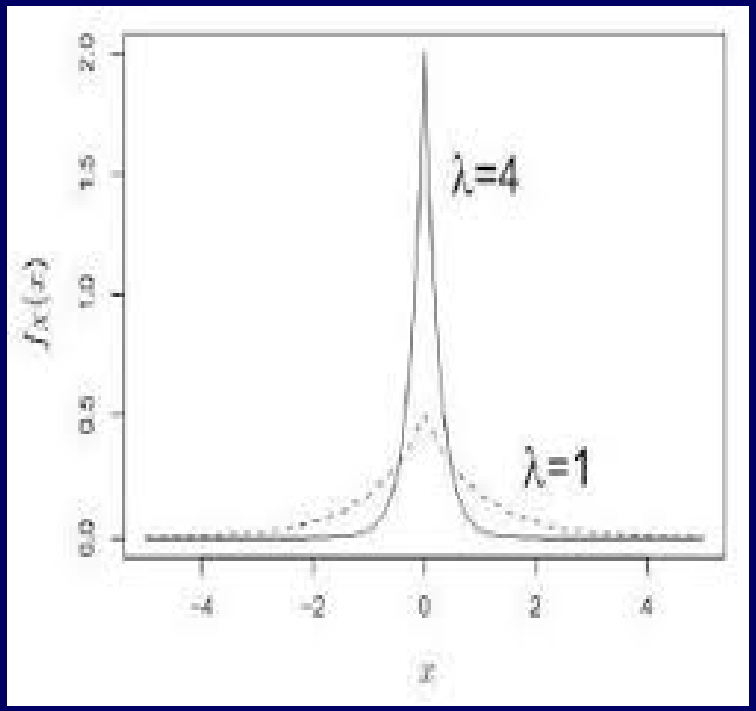
- Least Absolute (value) Shrinkage and Selection Operator
- Standard regression programs maximize log-likelihood (probability of data given model) as a function of regression coefficients β
- LASSO regression maximizes log-likelihood - $\lambda \Sigma |\beta_i|$, where λ is a parameter controlling sparsity
  - best value of λ is learned by cross-validation against withdrawn observations
  - value of λ determines how many variables are retained in the model (non-zero coefficients)

Bayesian interpretation of LASSO regression

- LASSO regression is equivalent to specifying a prior belief that large effects are less probable than small effects, and many effects are close to zero
  - Specifically, the LASSO penalty is equivalent to double exponential priors on the regression coefficients)
  - λ is a scale parameter that controls the strength of the prior: large values force regression coefficients towards zero.

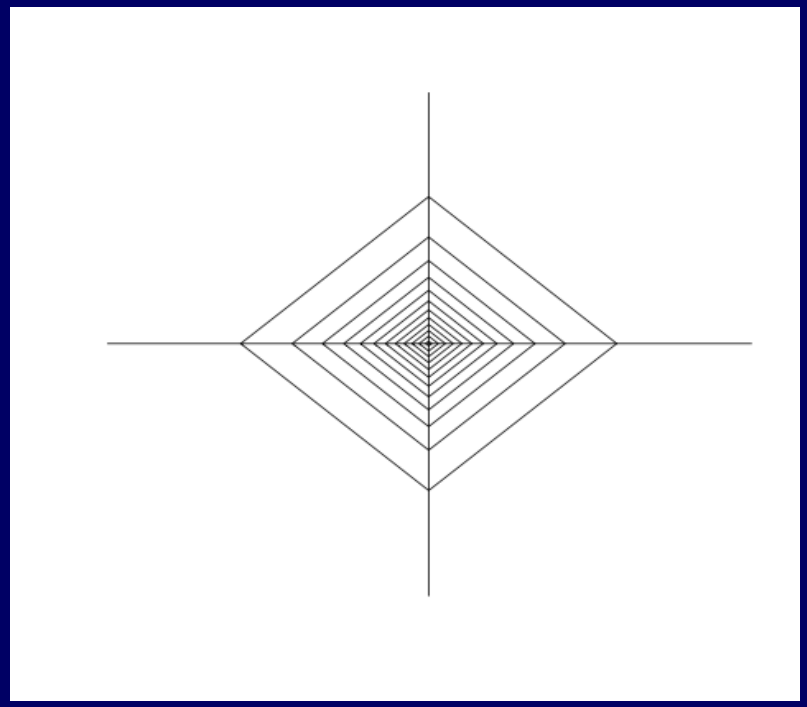
LASSO regression and the double exponential prir

- Parameter λ specifies the strength of the prior (penalty for large effect sizes)
  - learned from data by cross-validation



How double exponential prior encodes sparsity

- 2D probability density looks like a pyramid
  - See contour plot
- Density varies inversely with sum of absolute values of effect parameters



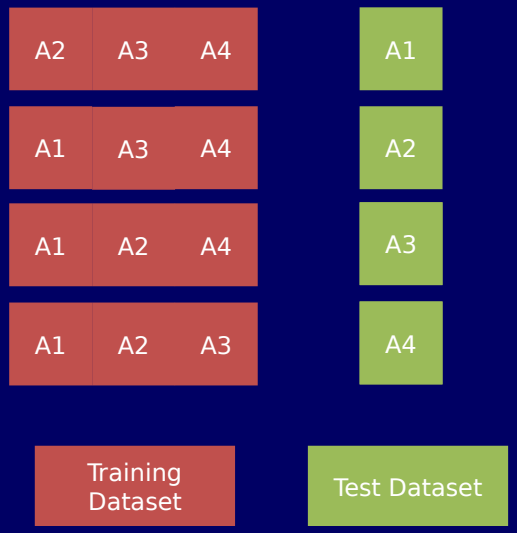
Why do we need to use cross-validation to learn and to evaluate a predictive model?

- To evaluate predictive performance
  - on test data that have never been used to learn the model
  - no need for a separate validation study (but may be hard to convince reviewers / regulatory agencies of this)
- To tune the learning algorithm
  - Optimal number of variables to retain
  - More generally, learn parameters that control how much the model adapts to the data
  - Models that adapt too much will *overfit*

N-fold cross-validation

- Partition dataset into N disjoint *test folds*
  - For each test fold, all other observations are the corresponding *training set*
- For each test/training fold
  - a model is fitted to the training fold and predictions are evaluated on the test fold
  - Predictive performance is evaluated by summing over all test folds
    - For each observation, can compare observed value with value predicted from model fitted to the corresponding training fold
    - Can compute area under ROC curve

Cross-validation compared with a conventional test/training split



With 4-fold cross-validation, each observation appears in one test fold and in 3 training folds

Nested cross-validation

If we are using cross-validation to tune the model and also to evaluate predictive performance, we need *nested* cross-validation.

Inner folds are used to tune the model (e.g. learn the optimal setting of the LASSO penalty parameter

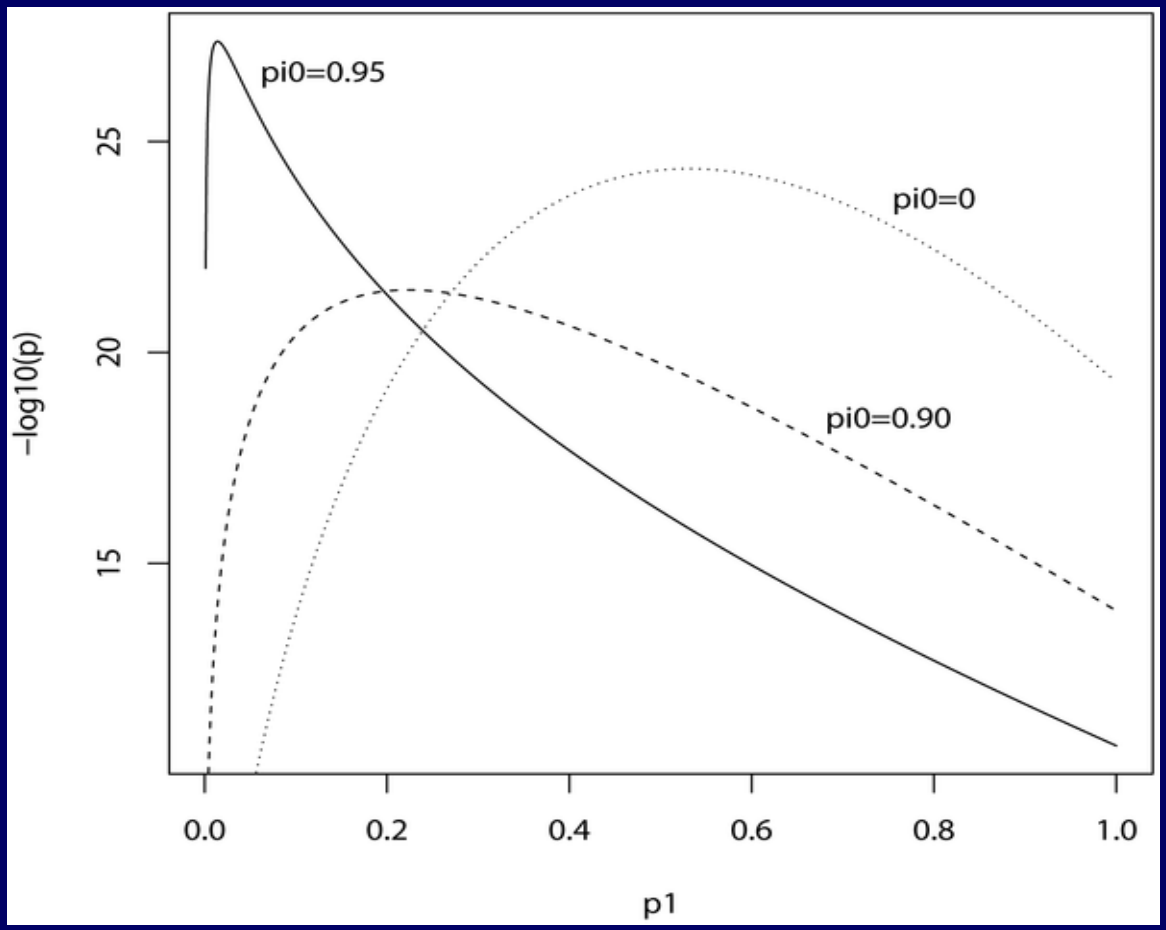
Outer folds are used to evaluate performance of the tuned model

With 10-fold cross-validation, nested cross-validation requires 100 model fitting runs

Using allele scores to predict outcome

- Allele scores can be computed from summary results of a GWAS
- (1) filter SNPs to select those that have p-value below some threshold
  - can be less stringent than the conventional threshold for declaring genome-wide significance
- (2) Calculate individuals'scores as sum of filtered SNP genotypes weighted by the regression coefficients
  - Use this score as a predictor

Performance of allele score for schizophrenia as predictor on test data as function of p-value threshold for filtering SNPs on training data (Dudbridge 2013)



The future of genotypic prediction

- Allele scores can be computed from summary level meta-analyses which are available for very large datasets
- LASSO predictors should outperform allele scores but constructing them requires access to individual-level data
- Genotypic effects on biomarkers are more oligogenic than effects on disease
  - Can learn genotypic predictors of biomarkers from cross-sectional studies, then use them as "features" to construct disease predictors