

Bayesian analysis of genetic association studies

John Whittaker

London School of Hygiene and Tropical Medicine/GlaxoSmithKline

June 10, 2009

Outline

- 1 Analysis of single studies
 - Full Bayes approach
 - Shrinkage estimation

- 2 Meta-analysis
 - Binary traits

Outline

- 1 Analysis of single studies
 - Full Bayes approach
 - Shrinkage estimation

- 2 Meta-analysis
 - Binary traits

Axiom

- A substantial proportion of variation in most interesting phenotypes is due to genetic variation
- Evidence: familial aggregation, twin studies etc
- For most phenotypes the genetic variants contributing to phenotypic variation are unknown
- Measure genetic variants and phenotypes of interest and look for associations

Motivating example: C-reactive protein

- CRP is an acute-phase protein involved in inflammatory response and associated with a number of diseases
- This has motivated interest in using CRP as a predictive biomarker for cardiovascular disease
- However, the existing epidemiological evidence is potentially subject to both reverse causation and confounding
- Here analyse associations between 14 SNPs and log CRP levels in a cohort of 1000 UK caucasians.
- Also include covariates for eg smoking, bmi etc

Analysis methods

- Assume that the functional variant is unobserved and treat as a missing data problem by averaging over possible genotypes at this variant
 - Numerous approaches varying in complexity/computational demands. Eg COLDMAP (Morris, Whittaker, Balding, 2002, 2004), Chapman et al (2003), Scheet & Stephens (2006), IMPUTE (Marchini et al, 2007)
- Look for patterns that indicate multiple SNPs are tagging an unobserved causal variant (ie multi-SNP test)
- Fit a regression model with SNPs as variables
- Do a sequence of single SNP tests

Linear model based analysis

- Treat each SNP as a variable in a regression model: if the functional SNP is typed, this is the 'correct' model
- Advantages:
 - Easy!
 - much existing statistical machinery/software
 - For indirect studies, can apply to imputed genotypes
- We wish to use Bayesian variable selection

Bayesian inference

- Assume, *all sources of uncertainty are regarded as randomness and expressed in terms of probability*
- Inference is done by:
 - writing down our prior beliefs about the parameters, $P(\theta)$
 - writing down a model $P(\mathcal{D}|\theta)$
 - applying Bayes theorem to update the prior:

$$\begin{aligned}P(\theta|\mathcal{D}) &= \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})} \\ &\propto P(\mathcal{D}|\theta)P(\theta)\end{aligned}$$

Variable selection model

- Phenotypes y_i measured on $i = 1, \dots, n$ individuals; for now $y_i \sim N(\mu_i, \sigma^2)$
- Genotypes at Q SNPs, $g_{ij} \in \{0, 1, 2\}$ $i = 1, \dots, n$
 $j = 1, \dots, Q$; code to give variables x_{ij}
- Variable selection model: $\mu_i = W_i\beta$, $i = 1, \dots, n$, where

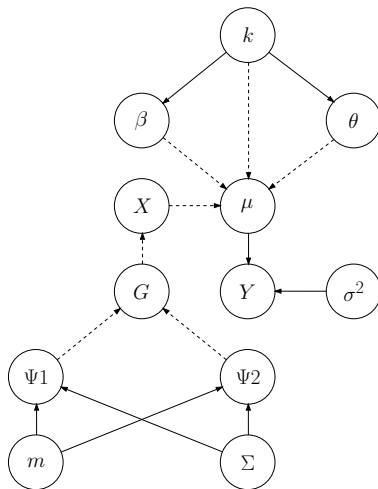
$$W_i = (1, x_{i\theta_1}, x_{i\theta_2}, \dots, x_{i\theta_k})$$

- Here:
 - k is the number of (currently) selected predictor variables;
 - $\theta = (\theta_1, \dots, \theta_k)^T$ pick out the selected variables;
 - β represents the $k + 1$ regression coefficients

Model fitting

- We fit using reversible jump Markov-chain Monte Carlo (RJMCMC) (Green, 1995)
- Incorporation in WinBUGS allows a (partially) generic implementation.
- Also facilitates modelling of missing data, either inferring via a multivariate probit model or by re-weighting eg PHASE output

Graphical model



CRP: Model probabilities (posterior > 0.02 + null)

Model	Posterior probability
{CRP2, CRP1(A)}	0.139
{CRP2, 9, CRP1(A)}	0.139
{CRP1(A)}	0.122
{APOE2, CRP1(A)}	0.0944
{CRP3059, CRP2}	0.0604
{CRP3059, CRP2, APOE2}	0.0581
{}	0.0120

CRP: Marginal probabilities

Marker	Marginal probability
CRP3059	0.290
CRP2	0.545
APOE2	0.468
CRP1(A)	0.667

Motivation

- WinBUGs implementation: ~ 100 SNPs
- Implementation of Bayesian logistic regression using RJMCMC implemented as R package using C code: ~ 10000 SNPs
- What about larger scale studies, for instance genome wide?
- Could prefilter by single locus test, but more satisfactory to do a single analysis.

Bayesian shrinkage

- RJMCMC doesn't scale easily to very large numbers of SNPs
- Instead, fit regression model with **all** SNPs, but with a shrinkage prior
- This shrinks the posteriors for regression coefficients so that many have modes at zero: hence does model selection
- Here use double exponential (DE) and normal exponential gamma (NEG) priors, tuned to control type I error
- Restrict to finding posterior modes rather than exploring full posterior
- Works with $> 500K$ SNPs

Priors

- DE: one parameter distribution commonly used as a shrinkage prior (Genkin, Lewis & Madigan, 2007)

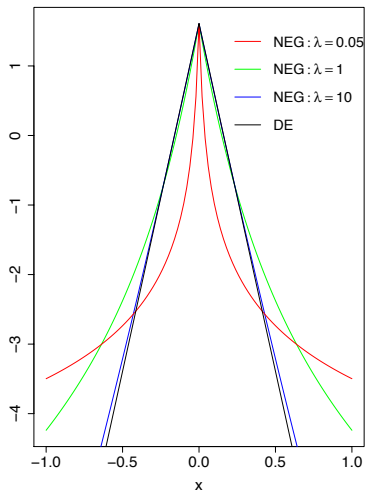
$$\text{DE}(\beta|\xi) = \int_0^\infty \text{N}(\beta|0, \sigma^2) \text{Ga}(\sigma^2|1, \xi^2/2) d\sigma^2 = \frac{\xi}{2} \exp(-\xi|\beta|)$$

- NEG (Griffin & Brown, 2007): generalisation of DE generated by sampling from a DE with parameter drawn from a gamma

$$\begin{aligned} \text{NEG}(\beta|\lambda, \gamma) &= \int_0^\infty \int_0^\infty \text{N}(\beta|0, \sigma^2) \text{Ga}(\sigma^2|1, \phi) \\ &\quad \times \text{Ga}(\phi|\lambda, \gamma^2) d\sigma^2 d\phi \end{aligned}$$

- As λ and γ increase such that $\xi = \sqrt{2\lambda}/\gamma$ is constant, NEG converges to DE with parameter ξ

Priors: log densities



Optimisation

- We wish to maximise the posterior $P(\beta|x, y)$ over β . By Bayes theorem,

$$\log P(\beta|x, y) = L(\beta) - f(\beta) + \text{const}$$

where $L(\beta)$ is log-likelihood and $f(\beta)$ is minus log prior

- Can approximate the type I error rate α for given prior, taking rejection as non-zero mode:

$$f'(\beta = 0) = \sqrt{\frac{n_0 n_1}{n_0 + n_1}} \Phi^{-1}(1 - \alpha/2)$$

where n_0 and n_1 are the numbers of cases and controls

- NB: Posteriors are multi-modal: we do multiple runs and report best mode.

Simulation studies

- 500 sets of 1000 cases / controls, 20 chromosomes of 20 Mb, 5 causal variants, 80K SNPs per data set
- single simulation of 480K SNPs, 10 causal variants in one 20MB region
- Resequencing data: 10 sets of 1000 cases / controls, 20Mb sequence, 5 causal variants

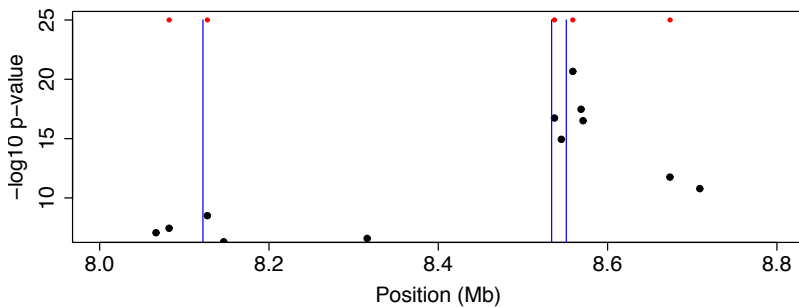
In each case compare with Armitage test for trend with $\alpha = 5 \times 10^{-7}$.

Results: simulation study, 500 replicates

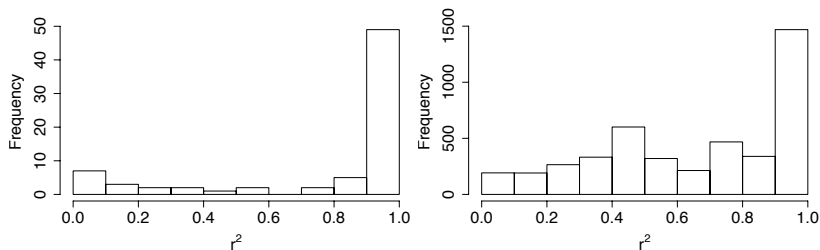
- total of 3000 causal SNPs;
- 'true' positive if it has $r^2 > 0.05$ with any causal SNP
- false positive if $> 20, 40, 200$ Kb from any other already recorded positive

Method	SNPs selected	Causal SNPs tagged	False positives			
			minimum separation (Kb)			
			0	20	40	100
NEG	2097	1576	368	368	368	366
DE	2622	1501	297	277	276	271
ATT	6810	1554	696	536	486	441

Results: simulation study, 480K SNPs



Results: simulation study, resequence data



- Both methods found 55/60 causal variants at $r^2 = 0.01$
- NEG selected 64 SNPs; ATT 599 SNPs

Type 2 diabetes

- Sladek et al (2007) analysed 694 type 2 diabetes cases and 654 controls using Human Hap300 BeadArrays
 - 42 significant SNPs (permutation p-value $< 10^{-5}$), tagging 32 distinct loci
 - Replication: 8 SNPs tagging 5 loci
- We reanalysed using the NEG:
 - 26 SNPs, tagging 25 distinct loci including the five previously-replicated loci
 - One SNP from each of these 5 loci, suggesting only one causal variant per locus.

Outline

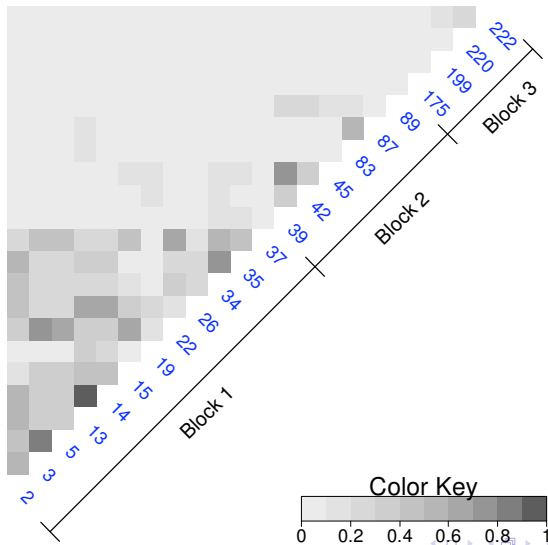
- 1 Analysis of single studies
 - Full Bayes approach
 - Shrinkage estimation

- 2 Meta-analysis
 - Binary traits

Example: PDE4D/stroke

- Candidate gene for stroke—much attention (first report has 275 citations) but 'replication' studies have been contradictory
- Despite this, much interest in PDE4D inhibitors
- Contrast to CRP: interest is in whether there is an association at all
- Recent single snps based meta-analysis looked at one part of the gene (Bevan et al, 2008)
- We consider 14 studies with a total of 12,929 subjects (5,994 cases, 6,935 controls) and 25 SNPs

LD



Model

- Remember we only have **summary** data for each study: ie allele/genotype counts for a subset of SNPs by case/control status
- We model the observed marginal minor allele counts \mathbf{q}_d^s (s for study, d for case/control) in terms of the haplotype probabilities π_d^s
- Assume (unobserved) haplotype counts are multinomial given π_d^s
- Approximate $P(\log(\mathbf{q}_d^s) | \pi_d^s, n_d^s)$ by multivariate normal with covariance matrix calculated from this multinomial via the multivariate delta method
- Note easy to get likelihood if \mathbf{q}_d^s is partially observed, because marginal of a MVN is also MVN.

Model

- Then assume case haplotype probabilities differ from control via SNP log-OR β
- Gives a likelihood in terms of π_0^s and β
- Add study random effects etc
- Fit by RJMCMC

Priors

- Informative priors on log-OR β so that most mass for OR in $(0.5, 2)$
- Informative prior for π_0^s based on HapMap
- Prior on model space: Poisson on dimension k giving a prior probability of 95% on null model

Results

- Univariate frequentist analysis: SNPs 5, 175 and 222 significant at 0.1; 95% CI include OR of 0.7 to 1.4

Results

- Univariate frequentist analysis: SNPs 5, 175 and 222 significant at 0.1; 95% CI include OR of 0.7 to 1.4
- Bayes multivariate analysis: all posterior probabilities of effect less than 5%; overall posterior probability of null around 0.9; Bayes factor of 2

Results

- Univariate frequentist analysis: SNPs 5, 175 and 222 significant at 0.1; 95% CI include OR of 0.7 to 1.4
- Bayes multivariate analysis: all posterior probabilities of effect less than 5%; overall posterior probability of null around 0.9; Bayes factor of 2
- Effects at this gene unlikely/small

Caveats

- Assumes both LD structure and any gene-disease associations are similar across the individual studies:
 - no evidence against this (eg MAF in controls very consistent)
 - however, impossible to exclude heterogeneity
- May be causal variants not well tagged by these SNPs:
 - 1,542 SNPs in HapMap in Gretarsdottir et al region
 - only 10% of the SNPs well tagged by SNPs in current analysis
 - Hence are only tagging SNPs flagged by Gretarsdottir et al
 - Of 260 SNPs in Gretarsdottir et al, only about 130 identified in HapMap: these tag about 30% of HapMap SNPs

Summary

- Linear models provide a flexible and powerful set of tools for the analysis of association tools
- Bayesian framework has advantages: full posteriors or modes?
- Computation is challenging, but can be dealt with
- Many possible extensions

Collaborators

- WinBUGS: Dave Lunn, Nicky Best (ICL)
- Bayesian shrinkage: Clive Hoggart, Maria De Iorio, David Balding (ICL)
- Meta-analysis: Paul Newcombe, Claudio Verzilli, Juliet Chapman, Liam Smeeth, Juan Pablo-Casas (LSHTM), Tina Shah, Aroon Hingorani (UCL)