Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies

David J. Balding

Centre for Biostatistics

Imperial College London

d.balding@ic.ac.uk

from 1/10/09 moving to: Institute of Genetics, University College London

Simultaneous analysis of all SNPs

Why try to do it?

- additional power to detect true positives:
  - multiple true predictors in model ⇒
    - less residual variation
    - better prediction of phenotype
  - capture all main effects for epistatic interactions
  - look for G×E joint/interaction effects with ≈ all the G
- signal at a potential false +ve weakened by true +ves
  - better localisation and interpretability.

Counter-arguments:

- massive optimization problem
  - unlikely to find global maximum
  - so may lose some of the above advantages.

HyperLASSO algorithm

- **Data:** cases and controls GW genotypes (+ covariates)
- **Model:** logistic regression

  $$\text{logit}(y_i) = \beta_0 + \sum_j \beta_j x_{ij}$$

  $i \equiv$ individuals; $j \equiv$ SNPs; $x_{ij} \equiv$ genotype $\in \{0, 1, 2\}$.
- **Problem:** too may predictors - overfitting.
- **Solution:** prior/penalty strongly rewards $\beta_j = 0$ for each j:

Normal-Exponential-Gamma (NEG)

NEG is a generalisation of the DE (Laplace):

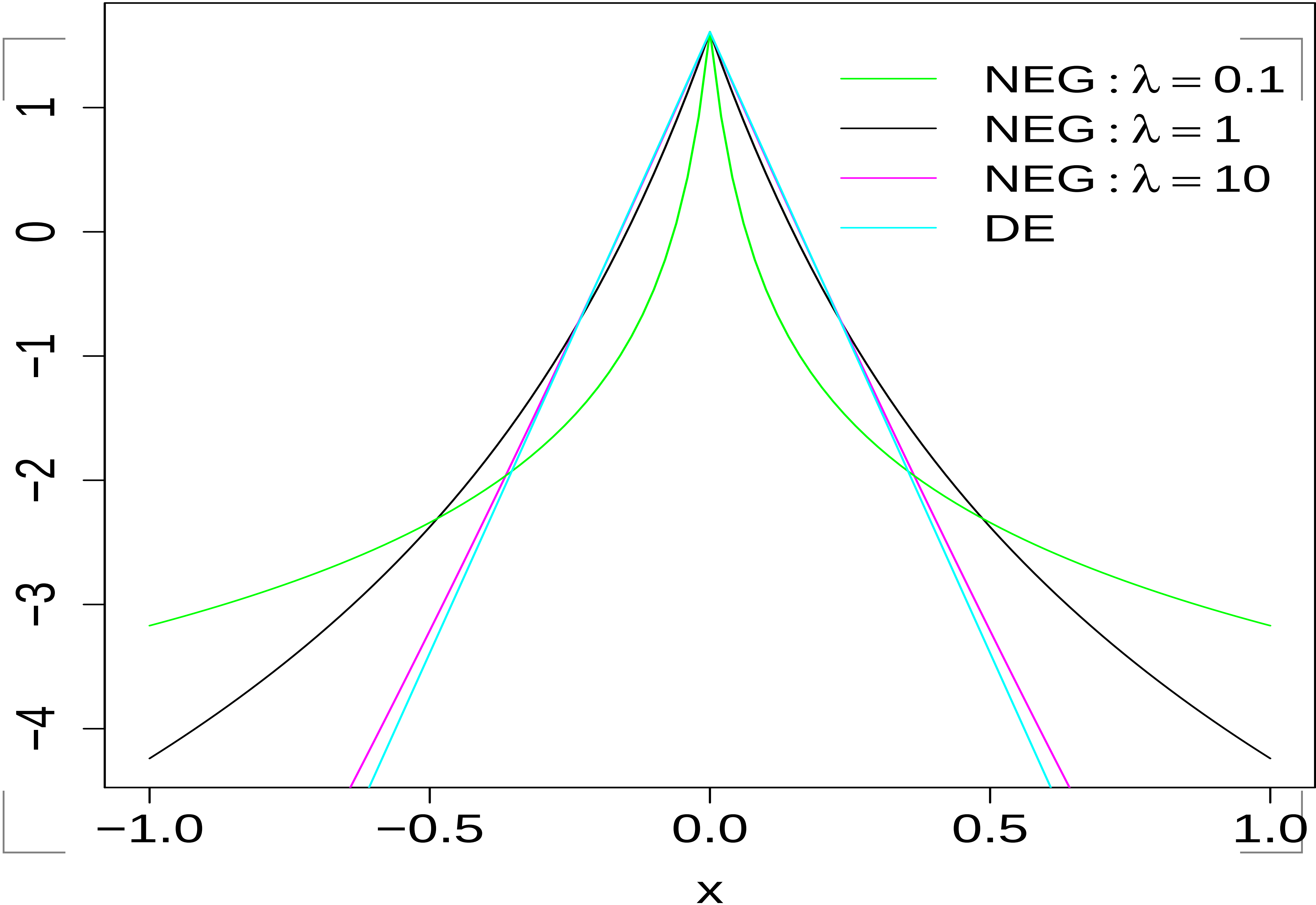
$$\mathsf{DE}(\beta|\xi) = \int_0^\infty \mathsf{N}(\beta|0,\sigma^2)\mathsf{G}(\sigma^2|1,\xi^2/2)\ d\sigma^2 = \frac{\xi}{2}\exp\{-\xi|\beta|\}$$

$$\begin{aligned} \mathsf{NEG}(\beta|\lambda,\gamma) &= \int_0^\infty\int_0^\infty \mathsf{N}(\beta|0,\sigma^2)\mathsf{G}(\sigma^2|1,\psi)\mathsf{G}(\psi|\lambda,\gamma^2)\ d\sigma^2 d\psi \\ &= \kappa\exp\left\{\frac{\beta^2}{4\gamma^2}\right\} D_{-2\lambda-1}\left(\frac{|\beta|}{\gamma}\right), \end{aligned}$$

If $\lambda \uparrow \infty$ and $\gamma \uparrow \infty$ with $\xi = \sqrt{2\lambda}/\gamma$ constant, NEG $\rightarrow$ DE.

- prior interpretation: most effects are $\sim$ zero, agnostic about size of non-zero effects;
- flatter tails means
  - less shrinkage so sparser models
  - initial over-estimate of effect size penalised lightly.

log density of NEG prior



Choice of prior parameters

Some prior quantiles for three choices of λ and γ:

λ	1.0	1.4	1.8
γ	0.0012	0.006	0.015
$\mathsf{P}(x > 0.05)$	5.8×10^{-4}	3.5×10^{-3}	1.7×10^{-2}
$\mathsf{P}(x > 0.1)$	1.4×10^{-4}	5.3×10^{-4}	2.1×10^{-3}
$\mathsf{P}(x > 0.2)$	3.6×10^{-5}	7.8×10^{-5}	2.0×10^{-4}
$\mathsf{P}(x > 0.4)$	9.0×10^{-6}	1.1×10^{-5}	1.7×10^{-5}
$\mathsf{P}(x > 1)$	1.4×10^{-6}	8.5×10^{-7}	6.2×10^{-7}

Larger shape parameter λ gives more weight to small, non-zero effects. Below, we choose λ to be very small (usually 0.05) to have very little shrinkage. Choose scale parameter γ to give desired type 1 error for given λ (see below).

The optimisation algorithm

Cyclic Co-ordinate Descent algorithm:

- start with all $\beta_j = 0$.
- update order is allocated randomly but fixed in each run.
- Newton-Raphson update step:

$$\beta_j^{new} = \beta_j - \frac{L'(\boldsymbol{\beta}) - f'(\beta_j)}{L''(\boldsymbol{\beta}) - f''(\beta_j)}$$

  where each $'$ denotes derivative wrt β_j.
- Key shortcut: use computationally-fast bounds to avoid expensive computation of L' for all but a few SNPs.
- Seek local optimum in 100 runs; choose best solution:
  - may not be global optimum but very similar model.

SNP selection

Given shrinkage prior:

- SNPs with non-zero posterior mode
  - $|(\text{log-lik})'| > |(\text{log-prior})'|$
  - not a Bayesian procedure: assess using type-1 error
  - use Bayesian language for penalised likelihood (“shrinkage regression”)
- rescaling genotype scores alters prior; e.g. if genotypes are standardised to mean zero and unit variance then
  - type-1 error invariant with MAF
  - for 1 SNP, asymptotically $\sim$ Armitage trend test (ATT).
  - supports larger effect sizes for lower MAF.

Type 1 error; Multiple genetic models

- explicit approximation for type 1 error in terms of derivative of log-prior
  - asymptotically correct for 1 SNP, otherwise conservative
  - avoids need for permutation
  - choose desired λ then assign γ to control type-1 error
- Can also consider dominant, recessive and heterozygous (1 df over-dominant) models, in addition to codominant (additive).
  - only one regression coefficient per SNP
  - no type-1 error approximation, but empirically the extra terms approximately double type-1 error.

Main simulation study

- 1K cases, 1K controls;
- 80K SNPs on 20 × 20 Mb chromosomes;
- 6 Causal SNPs all on one chromosome;
- 500 replicates, so 3K causal SNPs total;
- $\alpha = 10^{-5}$; additive-only model.

		Causal	False positives		
	SNPs	SNPs	min. separation (Kb)		
Method	selected	tagged	0	40	100
HyperLASSO	2097	1576	368	368	366
ATT	6810	1554	696	486	441

A causal variant is “tagged" if ≥ 1 selected SNP has $r^2 > 0{\cdot}05$ with it.

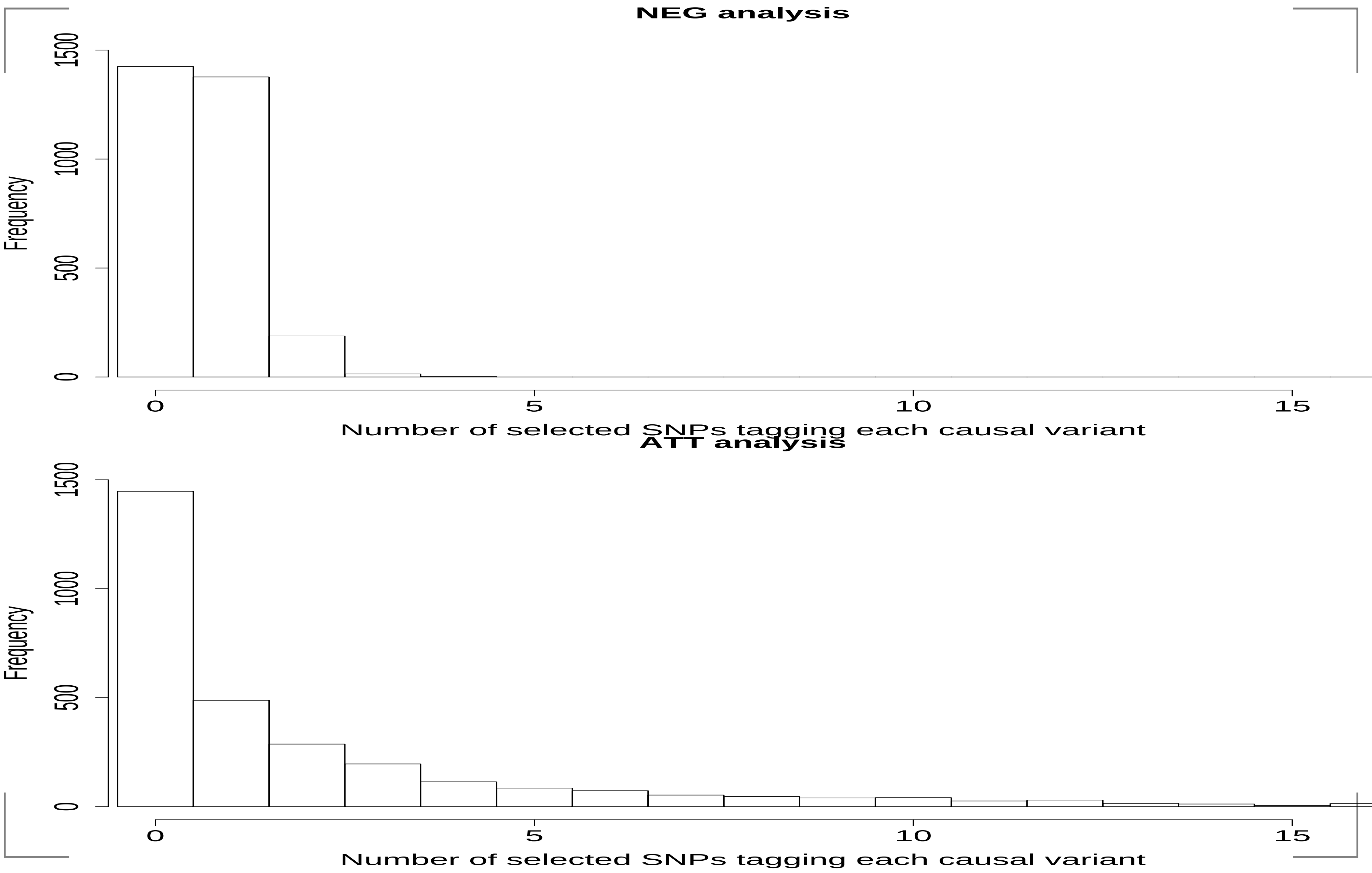
Main simulation study

# causal SNPs tagged (/500) by MAF and risk ratio

	MAF and allelic risk ratio					
	15%		5%		2%	
Method	1.4	1.5	1.8	2.2	2.5	3.0
HyperLASSO	252	360	209	370	146	239
ATT	244	353	209	370	143	235

- 54 SNPs tagged by HyperLASSO and not ATT;
- 32 SNPs tagged by ATT and not HyperLASSO;
- $p = 0{\cdot}11$

Tagging SNPs per causal SNP



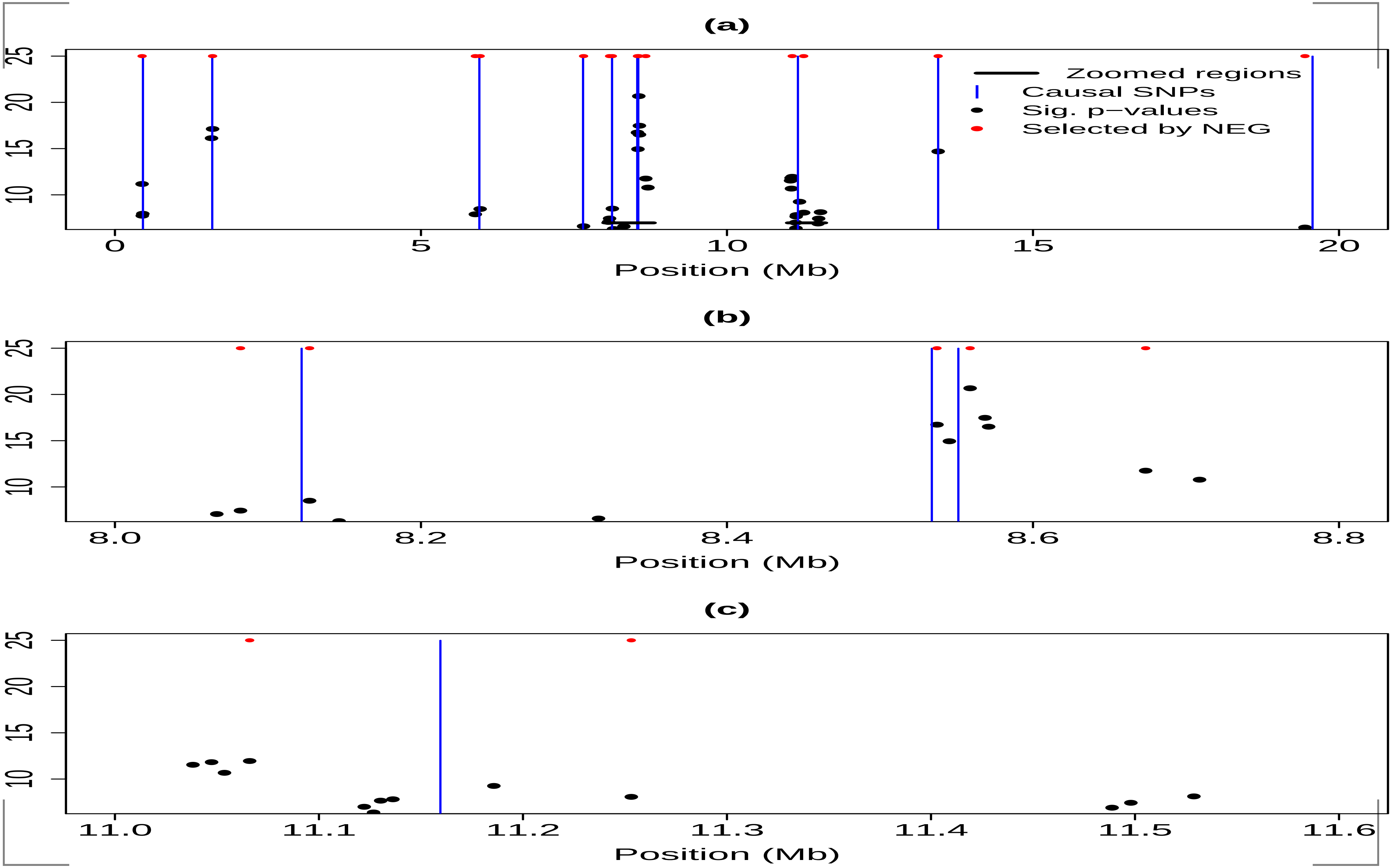
Genome-wide simulation study

- 1K cases, 1K controls;
- 120 × 20 Mb chromosome; 480K SNPs;
- 10 causal variants on one chromosome, each MAF $= 0{\cdot}15$ and risk ratio $= 2$;
- $\alpha = 5 \times 10^{-7}$; additive-only model;
- compute time $\approx$ 1 hour per mode.

Results:

- Both HyperLASSO and ATT tag all 10 causal SNPs;
- HyperLASSO selects 14 SNPs;
- ATT selects 35 SNPs.

Genome-wide simulation



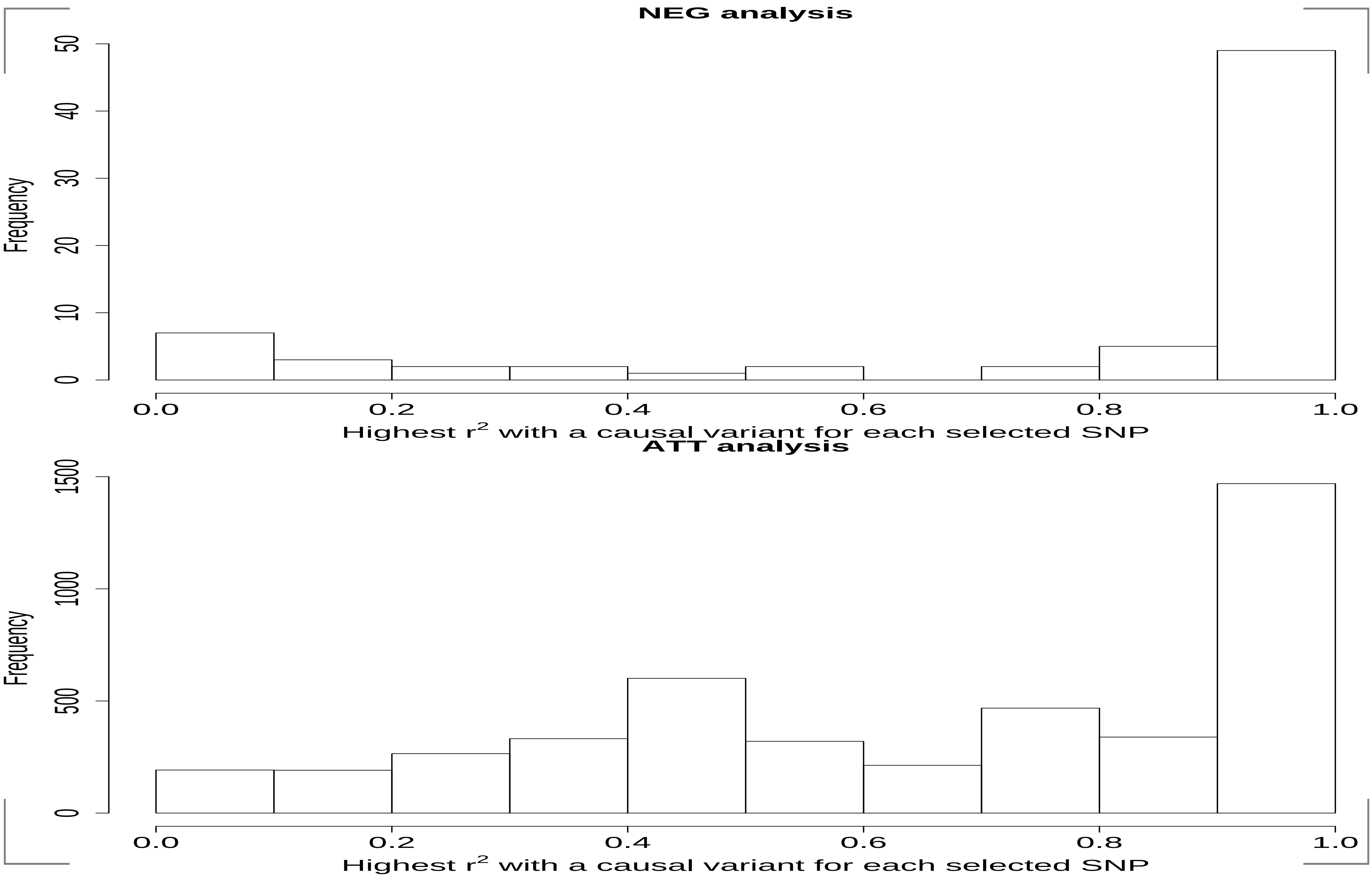
Resequencing simulation study

- 1K cases, 1K controls;
- all 192K polymorphic sites on one 20 Mb chromosome;
- 6 Causal SNPs on the chromosome
  - same disease model as main study;
- 10 replicates;
- $\alpha = 10^{-5}$; additive-only model;

Results:

- Both ATT and HyperLASSO tagged 54 of 60 causals;
- HyperLASSO selected 64 SNPs
- ATT selected 599.

Selected SNPs: best r^2 with causal



T2D GWAS Sladek et al. 2007

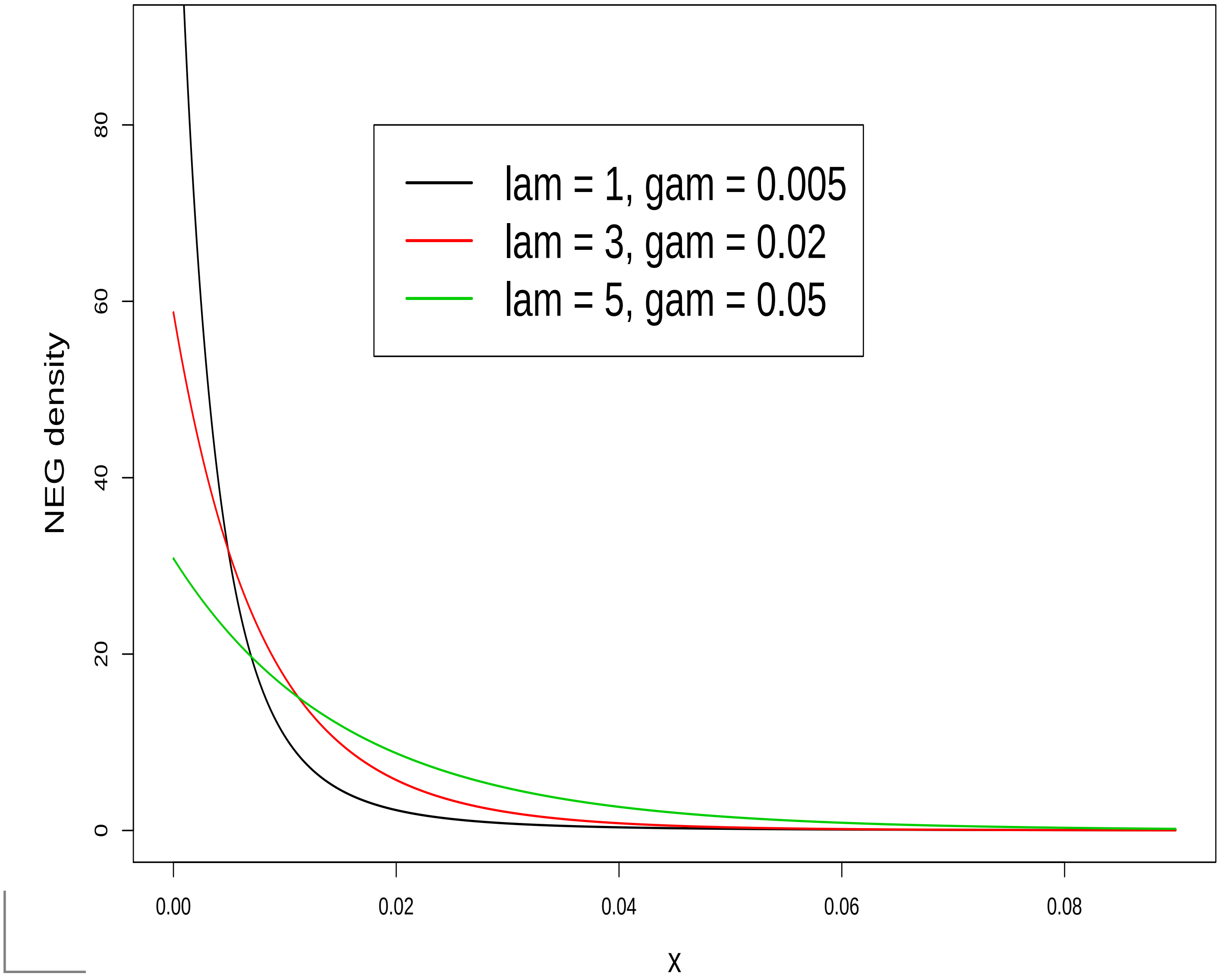
- 694 cases, 654 controls;
- 300K Illumina Hap300 genotyping platform;
- 42 SNPs tagging 32 loci significant at 5×10^{-5} and were progressed to stage 2 (plus 15 from Hap100 chip – not re-analysed here).
- NEG re-analysis using additive, dominant and recessive models:
  - 26 SNPs tagging 25 loci;
  - tagged all 5 loci confirmed in stage 2 – with same model (3 dominant, 2 additive);
  - looking at sub-optimal modes generated 3 new SNPs, each in high LD with a selected SNP.

Prediction of case/control status

- much interest in prediction of phenotype, but widespread view that prospects are poor, e.g.
  - NEJM Nov 08: 18 confirmed SNPs add little to prediction of T2D from known risk factors
  - Nat Genet 08: 20 confirmed SNPs for human height explain 3% of variation

  BUT these only include SNPs significant at very stringent levels. Many more true causal SNPs exist.
- relax penalty for prediction
  - larger models
  - greater shrinkage of effect sizes

NEG prior for prediction



Larger values of shape parameter λ gives more curvature and greater density away from origin, so more shrinkage and more non-zero modes.

Acknowledgments

- funding from UK Medical Research Council
- HyperLASSO software by Clive Hoggart, with help from Maria De Iorio and John Whittaker
- http://www.ebi.ac.uk/projects/BARGEN/
- Hoggart *et al.*, PLoS Genetics 2008