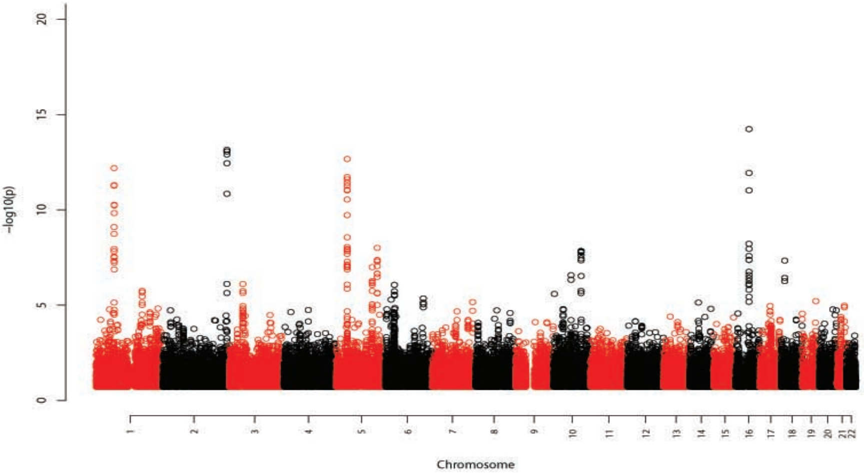
Fine-mapping the results from genome-wide association studies of primary biliary cholangitis using SuSiE and h2-D2

Aida Gjoka and Heather J. Cordell

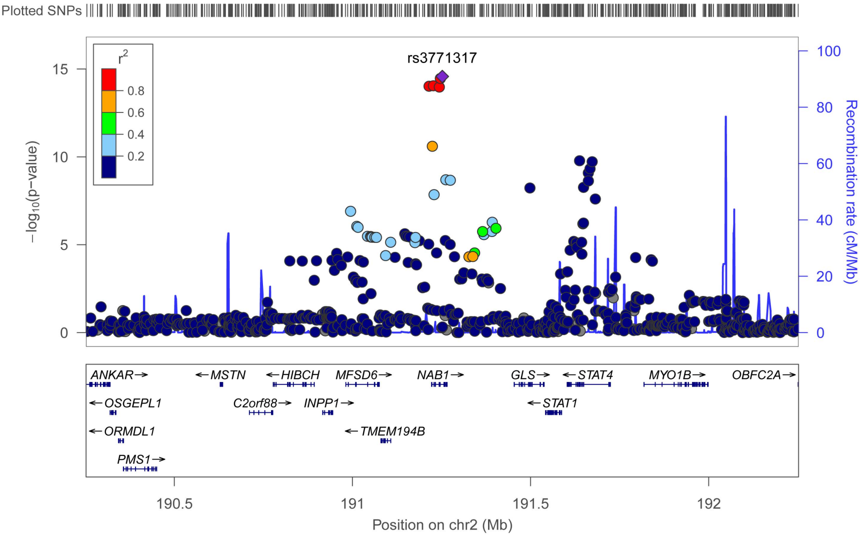
Population Health Sciences Institute
Faculty of Medical Sciences
Newcastle University, UK

heather.cordell@ncl.ac.uk

Genetic fine-mapping



Genetic fine-mapping



Genetic fine-mapping

- Aim is to identify the genetic variants (predictors) that causally affect the trait of interest
    - To learn something about the world
        - Rather than to build a better predictor

Genetic fine-mapping

- Aim is to identify the genetic variants (predictors) that causally affect the trait of interest
  - To learn something about the world
    - Rather than to build a better predictor
- This helps us identify the genes (stretches/sub-sequences of DNA) that causally affect the trait of interest
  - And thus uncover the underlying biological mechanisms

Genetic fine-mapping

- Aim is to identify the genetic variants (predictors) that causally affect the trait of interest
    - To learn something about the world
        - Rather than to build a better predictor
- This helps us identify the genes (stretches/sub-sequences of DNA) that causally affect the trait of interest
    - And thus uncover the underlying biological mechanisms
- If it is not possible to distinguish between several highly-correlated predictors, then we would like our conclusions to reflect this
    - E.g. "either x_1 or x_2 is relevant, and we cannot decide which"
    - Rather than arbitrarily selecting one variable and ignoring the other
        - As done, for example, by many penalized regression approaches

Genetic fine-mapping

- Aim is to identify the genetic variants (predictors) that causally affect the trait of interest
  - To learn something about the world
    - Rather than to build a better predictor
- This helps us identify the genes (stretches/sub-sequences of DNA) that causally affect the trait of interest
  - And thus uncover the underlying biological mechanisms
- If it is not possible to distinguish between several highly-correlated predictors, then we would like our conclusions to reflect this
  - E.g. "either x_1 or x_2 is relevant, and we cannot decide which"
  - Rather than arbitrarily selecting one variable and ignoring the other
    - As done, for example, by many penalized regression approaches
- Most current approaches frame the problem as a variable selection problem
  - Building a regression model where the outcome is the trait of interest
  - And the candidate predictor variables are the genetic variants (SNPs) that have been measured

Methods/programs for genetic fine-mapping

- CAVIAR (Hormozdiari et al. 2014)
- PAINTOR (Kichaev et al. 2014)
- CAVIARBF (Chen et al. 2015)
- FINEMAP (Benner et al. 2016)
- JAM (Newcombe et al. 2016)
- DAP (Wen et al. 2016)
- SuSiE (Wang et al. 2020) and SuSiE-RSS (Zou et al. 2022)

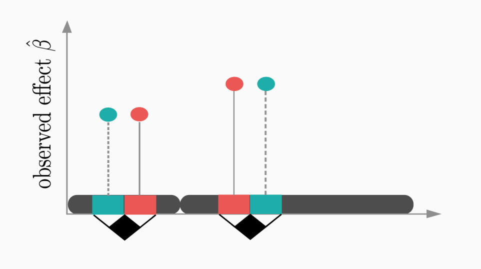
- h2-D2 (Li et al. 2024)
  - Uses a “continuous global-local shrinkage prior’ in contrast to the discrete mixture prior used by previous methods

Toy example from SuSiE authors

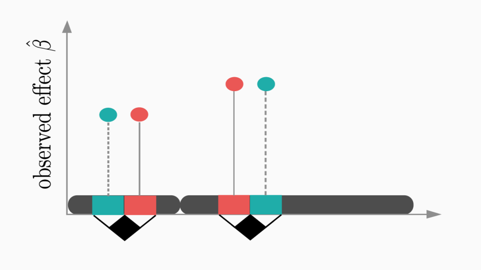
- Suppose we model the relationship between an n-vector $\mathbf{y}$ and an $n \times p$ matrix $\mathbf{X}$ as a multiple regression

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

- Assume there are exactly two effect variables – variables 2 and 3, say – and each is completely correlated with another non-effect variable:
    - $x_1 = x_2$ and $x_3 = x_4$, say.



Toy example

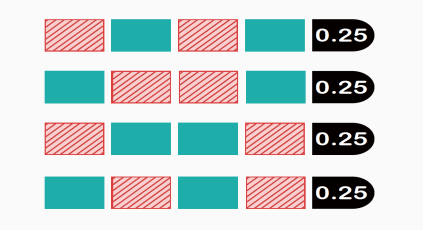


- Because the effect variables are completely correlated with other variables, it is impossible to select the correct variables confidently, even if *n* is very large
- However, given sufficient data, we should be able to conclude that there are (at least) two effect variables, and that

$$(\beta_1 \neq 0 \text{ or } \beta_2 \neq 0) \text{ and } (\beta_3 \neq 0 \text{ or } \beta_4 \neq 0)$$

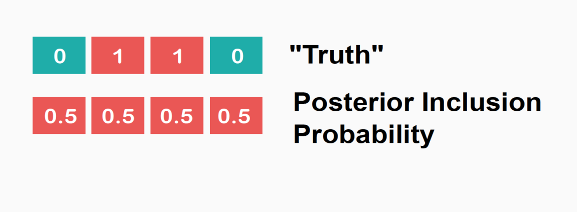
Toy example

- Most sparse/penalized regression approaches do not produce statements like this (nor do they attempt to do so)
- In principle, Bayesian variable selection approaches can produce such statements, as the posterior distributions should put roughly equal mass on the four equivalent combinations $(1, 3)$, $(2, 3)$, $(1, 4)$, $(2, 4)$:



Toy example

- However, in practice, due to the large number of possible combinations of variables, most BVS implementations rather summarize the posterior distribution by the marginal posterior inclusion probability (PIP) of each variable:

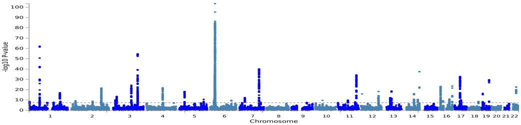


0	1	1	0	**"Truth"**
0.5	0.5	0.5	0.5	**Posterior Inclusion Probability**

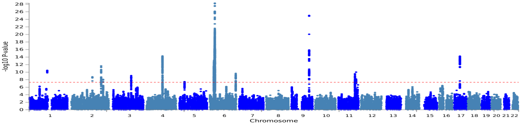
- Obtaining the true posterior distributions (and thus the desired inference) generally involves a lot of manual post-processing of results...
- However this is naturally output by SuSiE
  - Reports one or more "credible sets" of variants (at a user-specified coverage threshold e.g. 95%)

International PBC GWMA (Cordell et al. 2021)

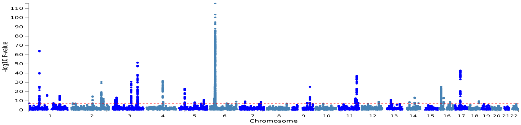
- European (5 cohorts: 8021 cases, 16,489 controls):



- Asian (2 cohorts: 2495 cases, 4283 controls):



- Combined (10,516 cases, 20,722 controls):



Fine-mapping using SuSiE-RSS

- We focussed on fine-mapping the 56 loci (excluding *HLA*) that were significant in the combined analysis
  - Using the European PBC cases and controls
- We re-derived the GWAS summary statistics using logistic regression (with 10 PCs as covariates)
- We used our own samples to estimate the correlation (LD) matrix
  - After initial attempts using a reference sample (European 1000 Genomes data) to estimate the LD matrix failed
- Compared the results to those previously obtained with FINEMAP

Fine-mapping using SuSiE-RSS

- We focussed on fine-mapping the 56 loci (excluding *HLA*) that were significant in the combined analysis
  - Using the European PBC cases and controls
- We re-derived the GWAS summary statistics using logistic regression (with 10 PCs as covariates)
- We used our own samples to estimate the correlation (LD) matrix
  - After initial attempts using a reference sample (European 1000 Genomes data) to estimate the LD matrix failed
- Compared the results to those previously obtained with FINEMAP
- In February 2024, the h2-D2 method was published in AJHG (Li et al. 2024)
  - Uses essentially the same inputs as SuSiE-RSS
  - While producing similar outputs in terms of credible sets

Fine-mapping using h2-D2

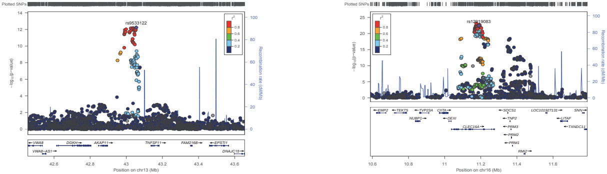
- Put in an HPC software install request on 8th February - was installed by 13th February
  - h2-D2 failed to work on some nodes
  - Gave an error about a deprecated R function
- I contacted the h2-D2 authors on 21st February - they responded on 22nd February
  - Saying we must have downloaded an old version of the software
- Put in another HPC software install request - was installed by 23rd
- On 7th March the h2-D2 authors contacted me to tell me that the h2-D2 package had been updated to version 1.1, with an important update for when an in-sample LD matrix is used
- Put in yet another HPC software install request (!) - was installed by 12th March

PBC fine-mapping results

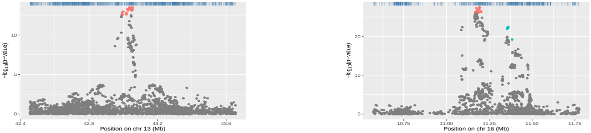
Results obtained from SuSiE-RSS and h2-D2 in comparison with previously-obtained posterior probabilities from FINEMAP. No of CS is the number of credible sets generated. CS sizes are the sizes of the generated credible sets.

Locus		FINEMAP posterior probabilities			SuSiE results		h2-D2 results	
number	Locus	1 variant	2 variants	3 variants	No of CS	CS sizes	No of CS	CS sizes
1	1p36.32	0.95	0.05	0	1	65	1	79
2	1p31.3	0.05	0.45	0.5	2	1,4	1	1
3	1p13.1	0.95	0.05	0	1	29	1	29
4	1q23.1	0.92	0.08	0	1	33	1	48
5	1q31.3	0.8	0.19	0.01	1	6	1	20
6	1q32.1	0.85	0.14	0	1	33	1	37
7	2p25.1	0.88	0.12	0	1	7	1	11
8	2p23.3	0.1	0.8	0.09	2	2, 8	1	8
9	2q21.3	0.88	0.12	0	1	12	5	3, 2, 9, 32, 35
10	2q32.2	0	0	1	4	1, 14, 32, 19	3	1, 18, 33
11	2q33.2	0.61	0.37	0.02	1	39	1	54
12	3p24.3	0.88	0.11	0	1	14	1	25
13	3p24.2	0.91	0.02	0	1	10	1	13
14	3q13.33	0.8	0.17	0.02	1	2	1	3
15	3q25.33	0	0	1	4	1, 4, 11, 39	4	1, 4, 10, 45
16	4q24(1)	0.71	0.27	0.02	1	63	1	76
.	.	.	.	.	.	.	.	.
26	7p14.2	0.91	0.09	0	1	33	1	40
27	7q32.1	0	0.89	0.11	2	20, 6	2	10, 21
.	.	.	.	.	.	.	.	.
35	11q23.1	0.22	0.73	0.04	2	10, 57	2	38, 61
36	11q23.3	0.67	0.28	0.05	1	12	1	9
.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	
.	.	.	.	.	.	.	.	

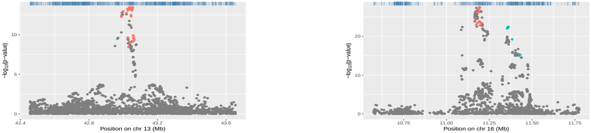
Some concordant loci



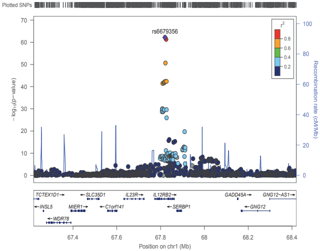
SuSiE



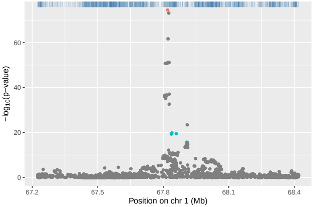
h2-D2



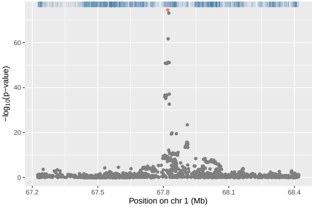
A less concordant locus



SuSiE

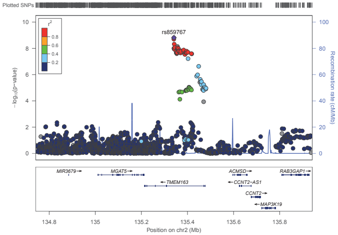


h2-D2

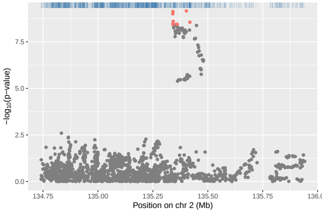


Another less concordant locus

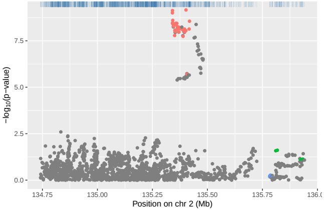
Plot from Cordell et al. (2021)



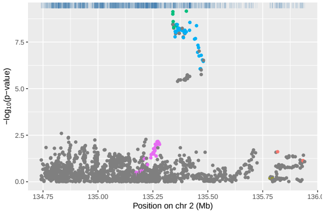
SuSiE results (coverage 0.6)



h2-D2 results (coverage 0.95)



h2-D2 results (coverage 0.6)

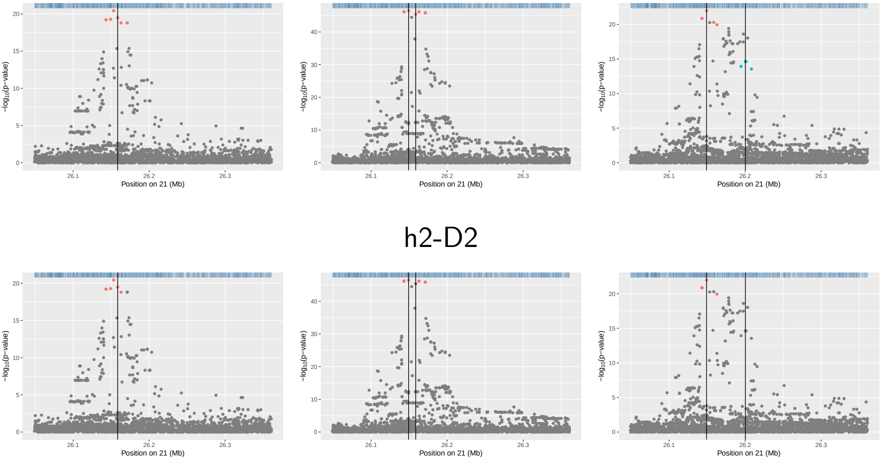


Simulation study

- We used the HAPGEN2 software to simulate data (based on CEU HapMap genotypes in a 0.4 Mb region of chromosome 21) under three different scenarios:
  - A single causal variant
  - Two causal variants, close together in LD
  - Two causal variants, further apart
- Results from 100 replicates:

Analysis Method	Scenario	Power (1st)	Power (2nd)	Power (both)	Mean no of CS	SD no of CS	Mean CS size	SD CS size
SuSiE-RSS	1	0.95	-	-	0.99	0.10	7.18	2.27
SuSiE-RSS	2	0.94	0.98	0.92	1.00	0.00	5.27	1.25
SuSiE-RSS	3	0.85	0.76	0.76	1.77	0.45	7.17	3.15
h2-D2	1	0.84	-	-	0.90	0.30	9.11	4.76
h2-D2	2	0.80	0.88	0.72	0.98	0.14	4.18	1.43
h2-D2	3	0.71	0.23	0.15	1.08	0.46	9.36	5.65

Results from 3 example replicates



SuSiE

h2-D2

Scenario 1

Scenario 2

Scenario 3

Conclusions

- In application to PBC, SuSiE-RSS and (to a lesser extent) h2-D2 produced similar results to those previously obtained with FINEMAP
    - Where there were discrepancies, the results from SuSiE-RSS seemed more plausible

Conclusions

- In application to PBC, SuSiE-RSS and (to a lesser extent) h2-D2 produced similar results to those previously obtained with FINEMAP
    - Where there were discrepancies, the results from SuSiE-RSS seemed more plausible
- In simulations, SuSiE-RSS performed better than h2-D2 at identifying the true causal variant(s)
    - Consistent with simulation results presented by Li et al. (2024)
    - They showed SuSiE to have a slight power advantage over h2-D2 in several scenarios

Conclusions

- In application to PBC, SuSiE-RSS and (to a lesser extent) h2-D2 produced similar results to those previously obtained with FINEMAP
  - Where there were discrepancies, the results from SuSiE-RSS seemed more plausible
- In simulations, SuSiE-RSS performed better than h2-D2 at identifying the true causal variant(s)
  - Consistent with simulation results presented by Li et al. (2024)
  - They showed SuSiE to have a slight power advantage over h2-D2 in several scenarios
- In more extensive simulations, Li et al. (2024) found some cases where h2-D2 outperformed SuSiE
  - We used the default parameter options for h2-D2 in terms of coverage, purity, mcmc iterations, burn-in, stepsize etc. (except when tweaking the coverage in order to match SuSiE or to produce any credible sets)
  - Altering these could potentially result in better performance

Conclusions

- In application to PBC, SuSiE-RSS and (to a lesser extent) h2-D2 produced similar results to those previously obtained with FINEMAP
    - Where there were discrepancies, the results from SuSiE-RSS seemed more plausible
- In simulations, SuSiE-RSS performed better than h2-D2 at identifying the true causal variant(s)
    - Consistent with simulation results presented by Li et al. (2024)
    - They showed SuSiE to have a slight power advantage over h2-D2 in several scenarios
- In more extensive simulations, Li et al. (2024) found some cases where h2-D2 outperformed SuSiE
    - We used the default parameter options for h2-D2 in terms of coverage, purity, mcmc iterations, burn-in, stepsize etc. (except when tweaking the coverage in order to match SuSiE or to produce any credible sets)
    - Altering these could potentially result in better performance
- Applying FUMA to our SuSiE-RSS results (to identify genes and pathways important in PBC) gave results largely consistent with those previously obtained (using FINEMAP results)

Acknowledgements

- Aida Gjoka
- Wellcome Trust

- We are recruiting! (Closing Date: 14 April 2024):

  https://jobs.ncl.ac.uk/job/Newcastle-Research-Assistant-Research-Associate-in-Statistical-Genetics/1044343001/