Needles, haystacks and genetic causes of disease

Chris Wallace chr1swallace chr1swallace.github.io

Jenn Asimit

Mary Fortune

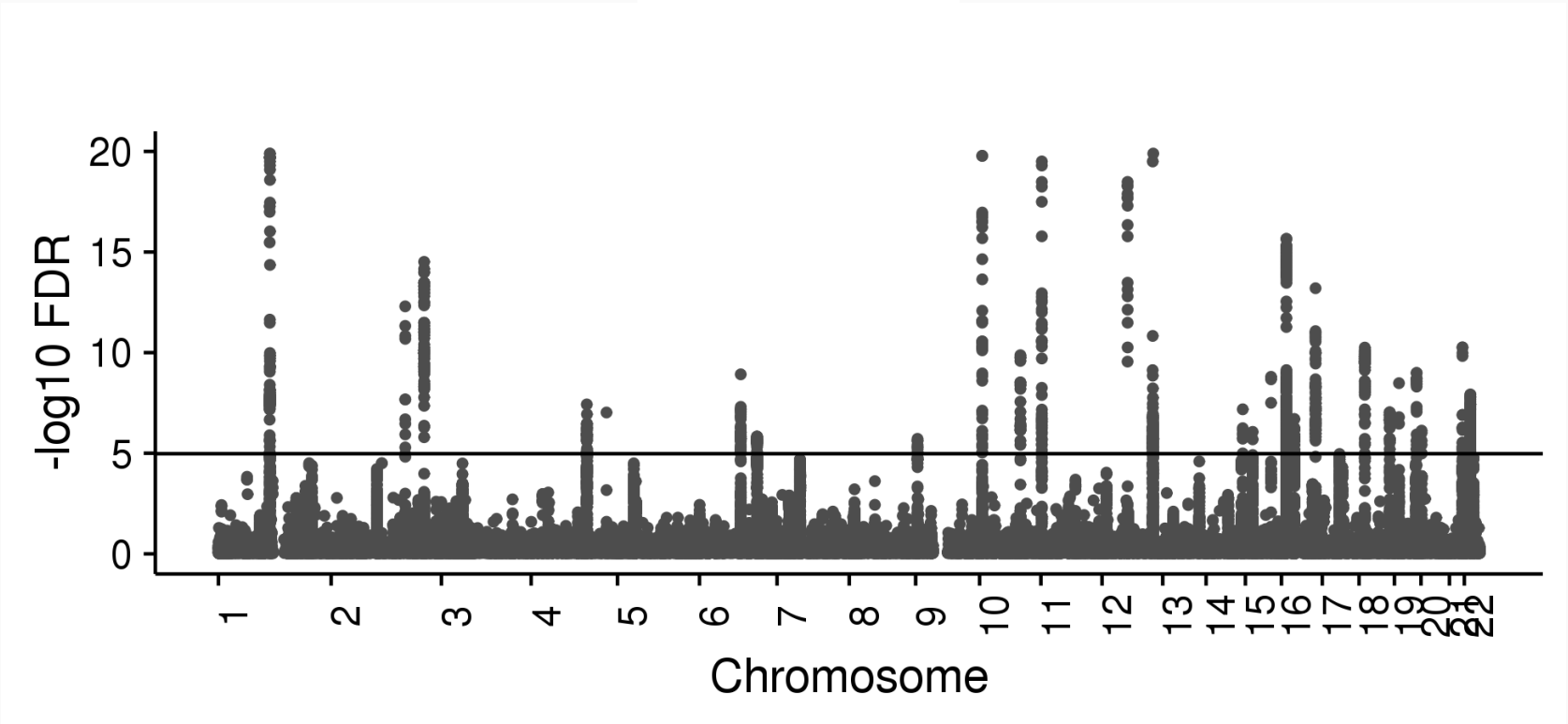
Dan Rainbow

wellcome

UNIVERSITY OF CAMBRIDGE

MRC | Biostatistics Unit

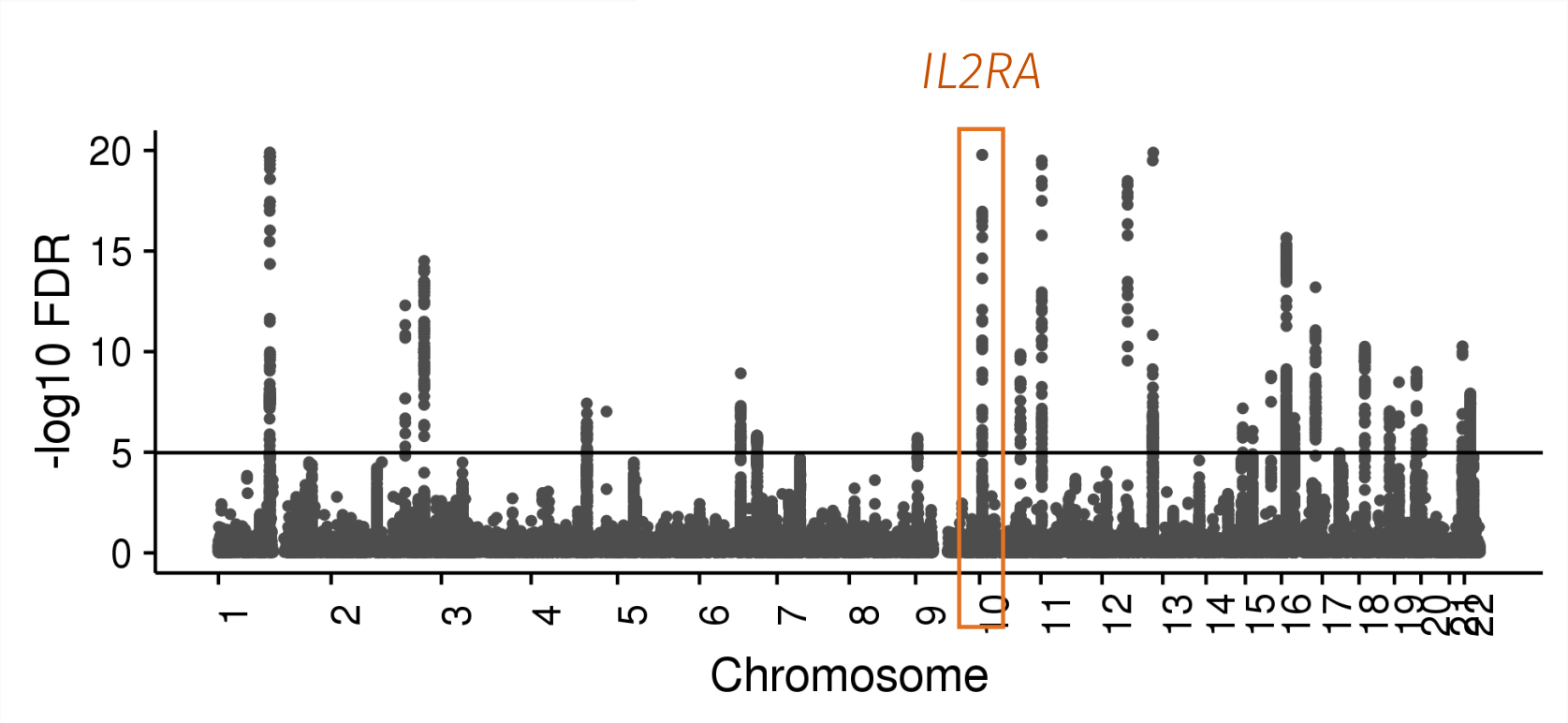
Manhattan plots (haystack plots?)



How do we use GWAS to identify disease genes/proteins and drug targets?

Onengut et al, Nat Genet 2015

Manhattan plots (haystack plots?)



How do we use GWAS to identify disease genes/proteins and drug targets?

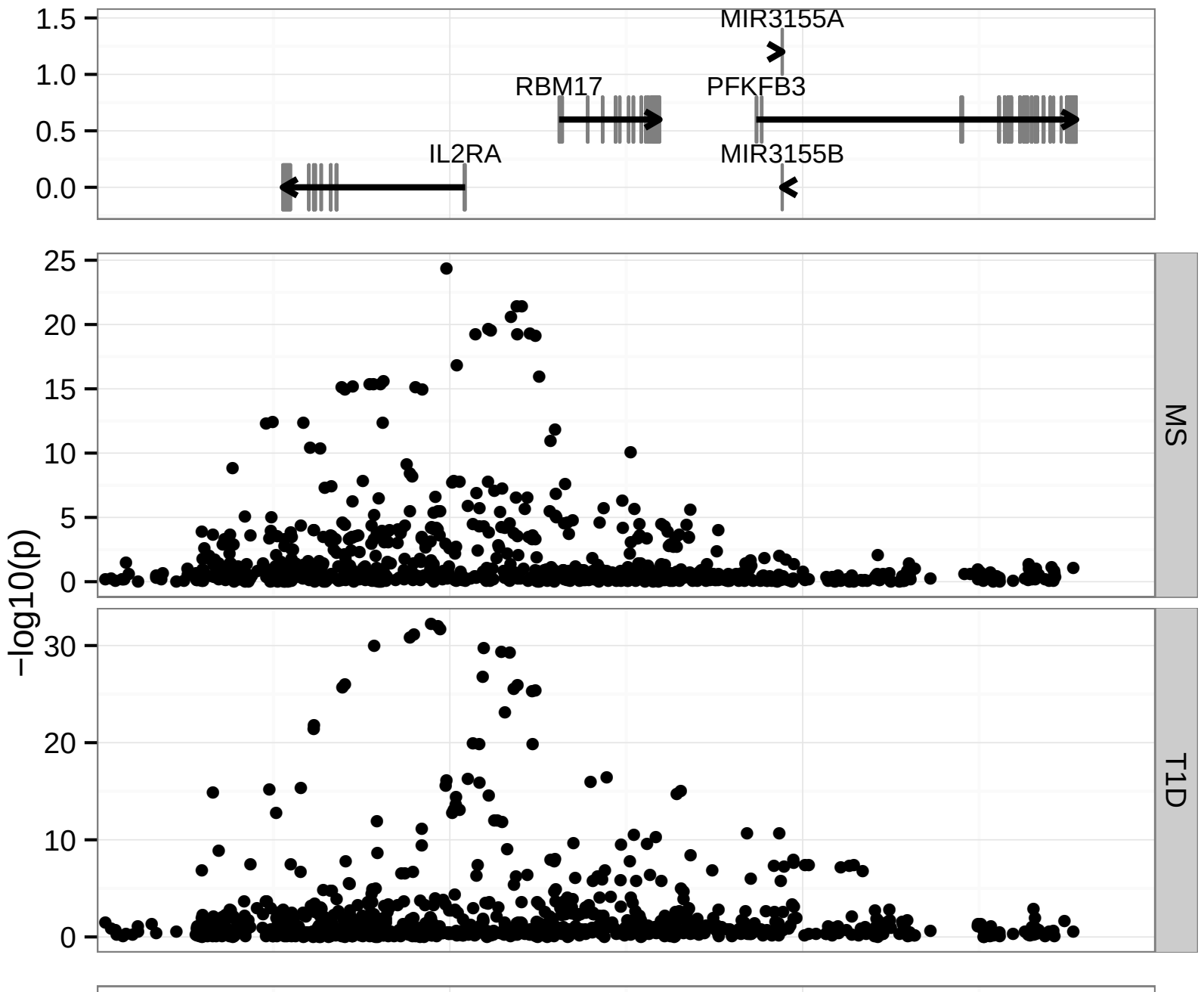
Onengut et al, Nat Genet 2015

Overview

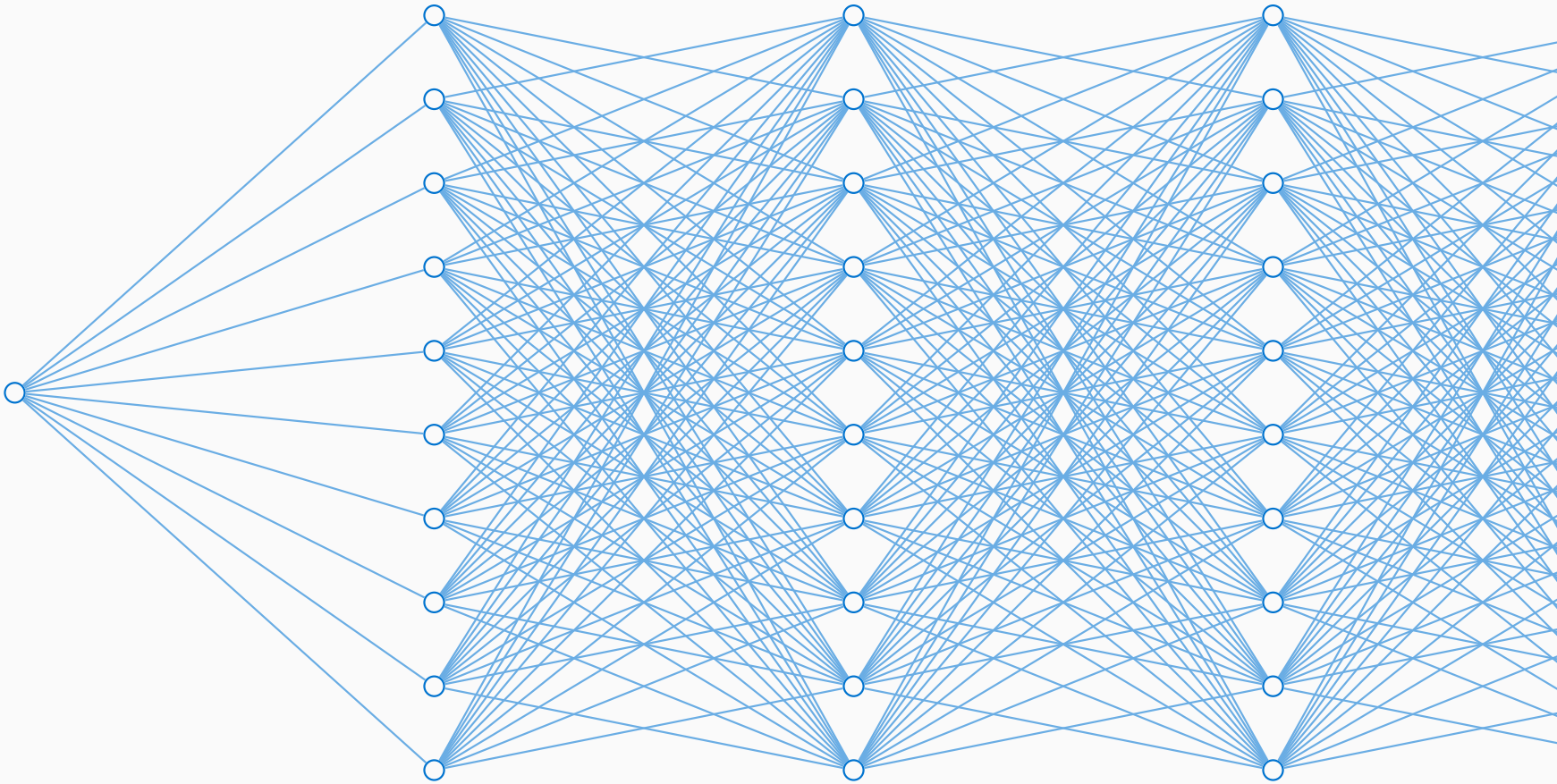
- Fine mapping causal variants
- Joint tagging of causal variants
- Improved fine mapping by exploiting shared aetiology
- Functional validation of causal effects on *IL2RA*

Fine mapping causal variants

Association of MS and T1D in *IL2RA* region

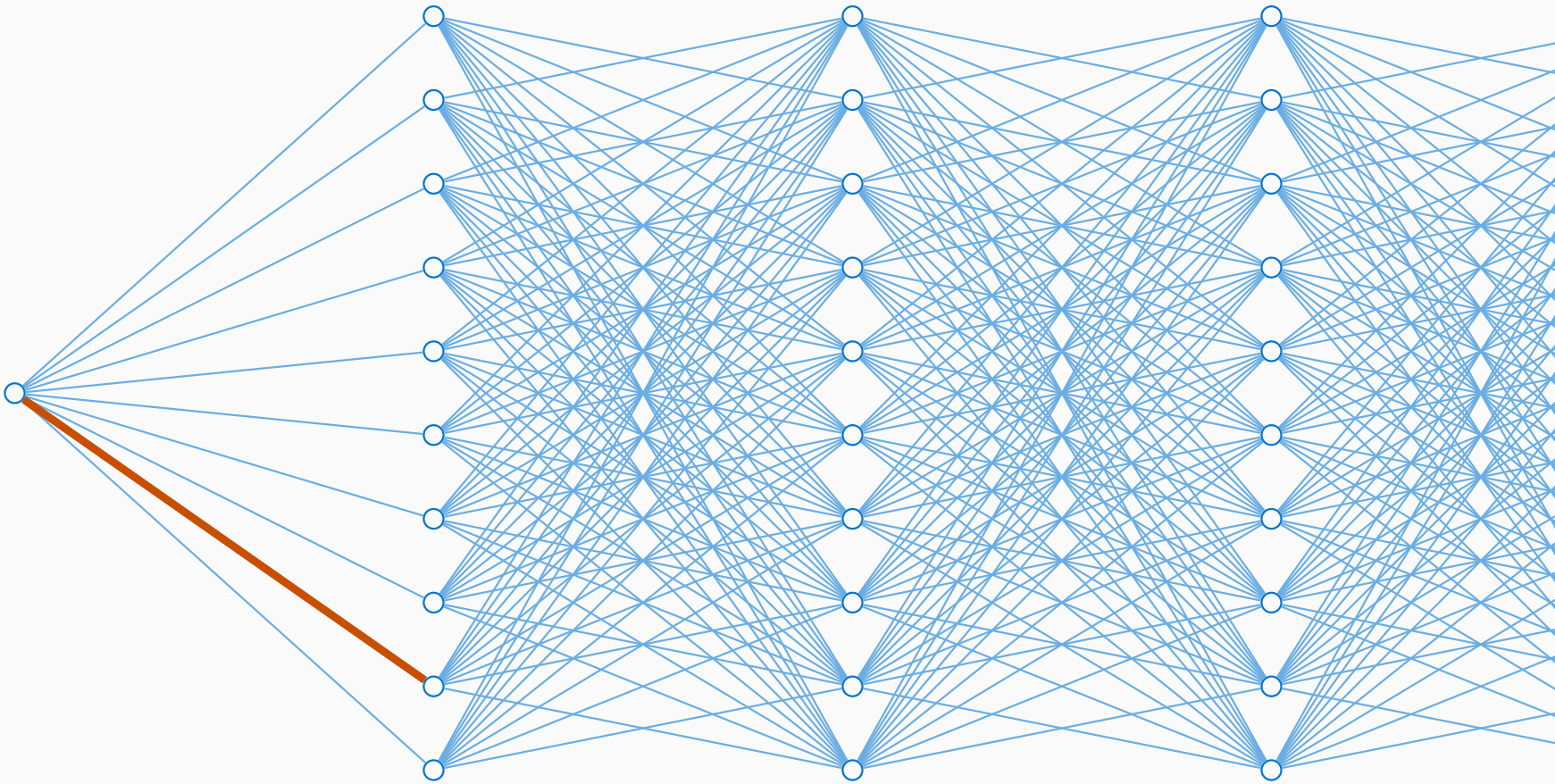


Stepwise search



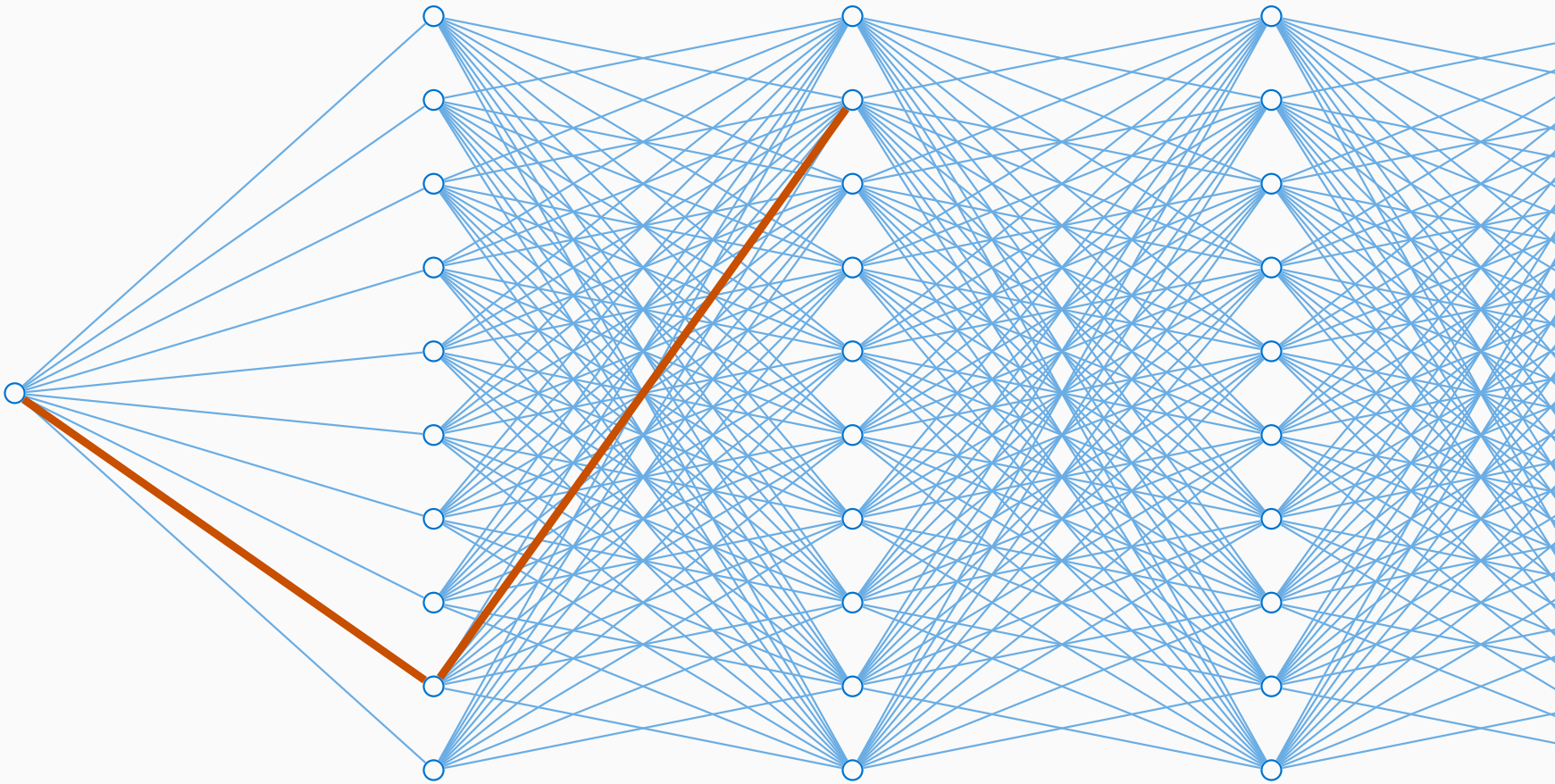
no SNPs 1 SNP 2 SNPs 3 SNPs

Stepwise search



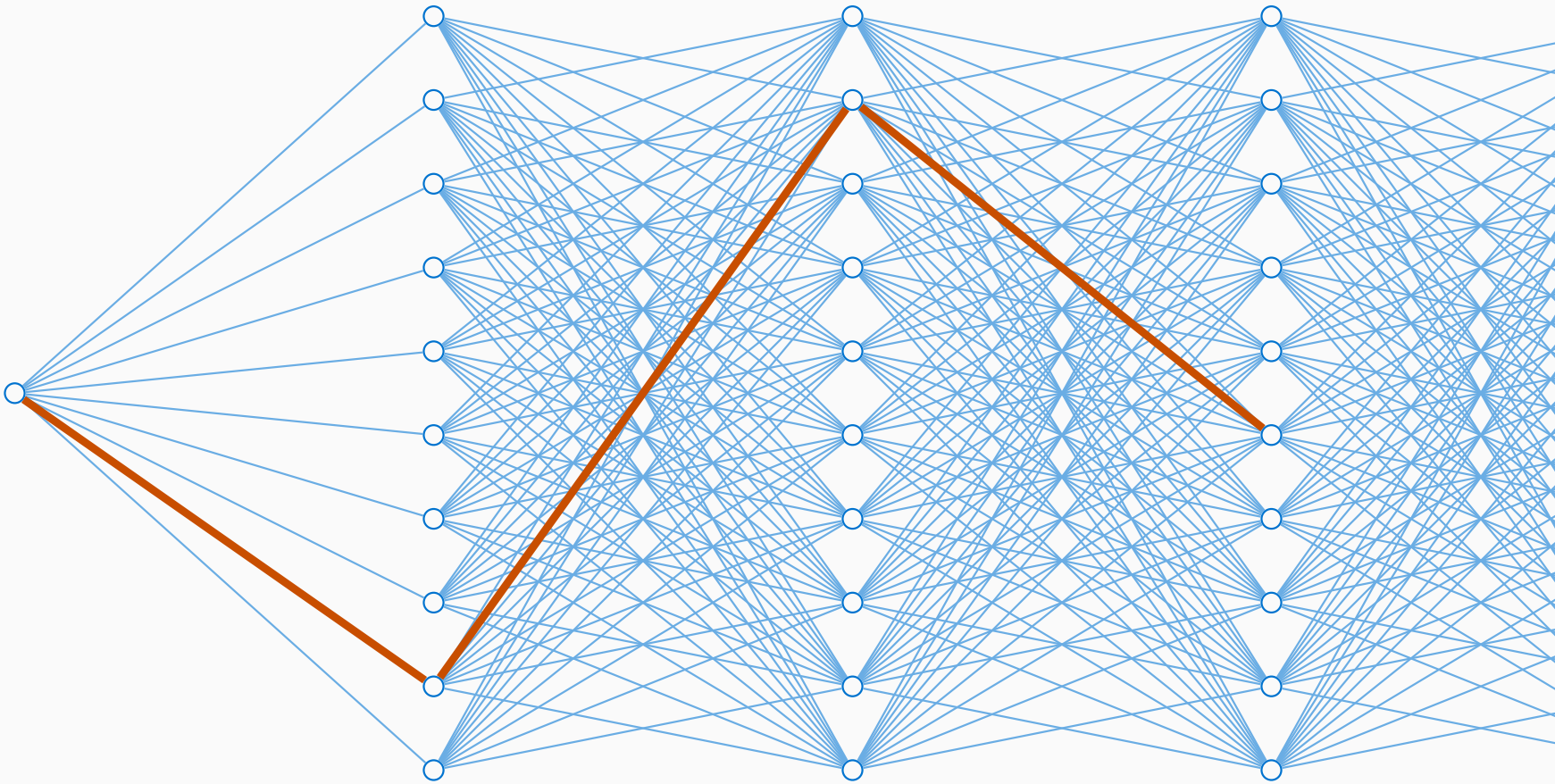
no SNPs 1 SNP 2 SNPs 3 SNPs

Stepwise search



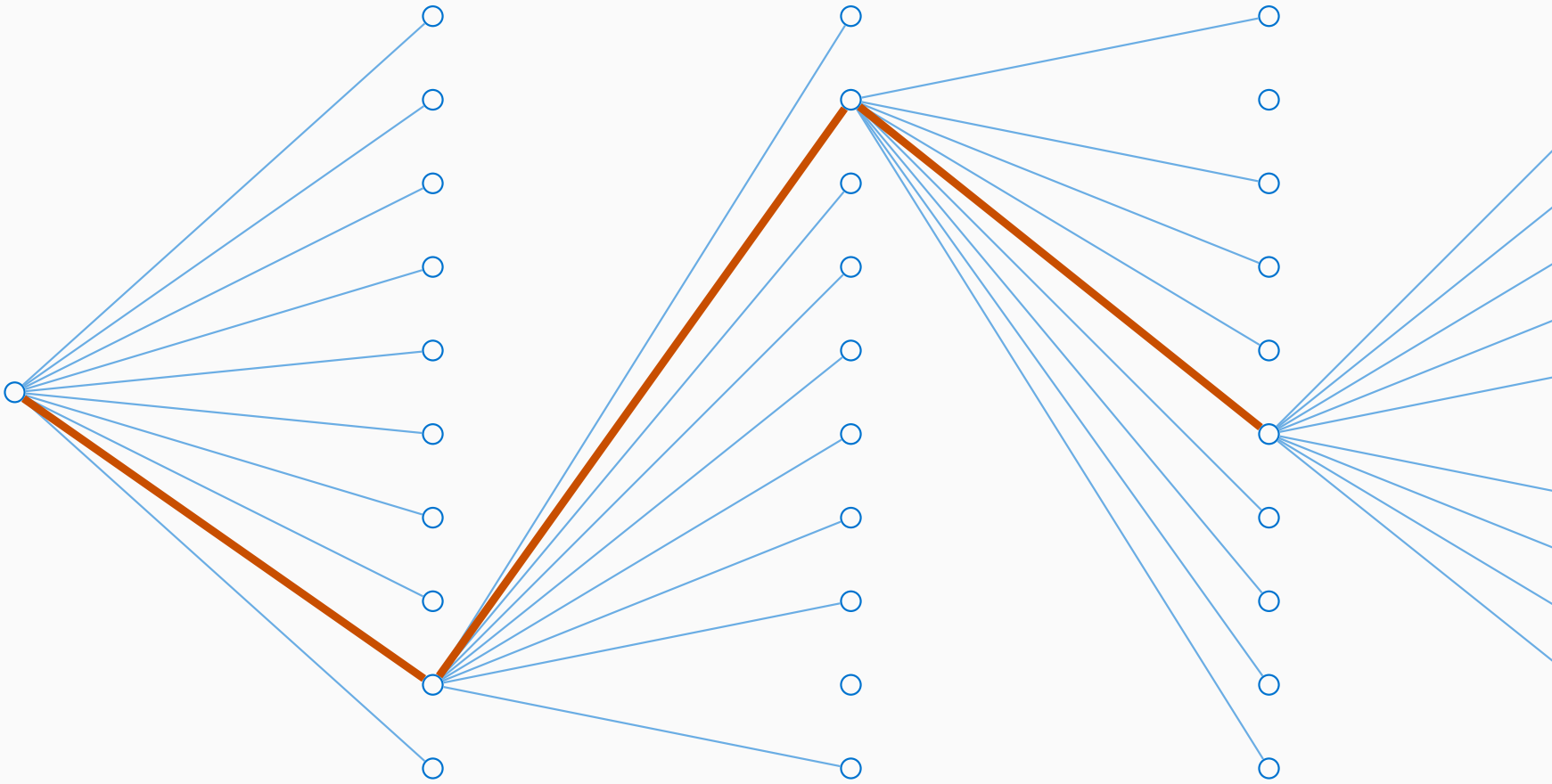
no SNPs 1 SNP 2 SNPs 3 SNPs

Stepwise search



no SNPs 1 SNP 2 SNPs 3 SNPs

Stepwise search



no SNPs 1 SNP 2 SNPs 3 SNPs

Stepwise search of IL2RA region

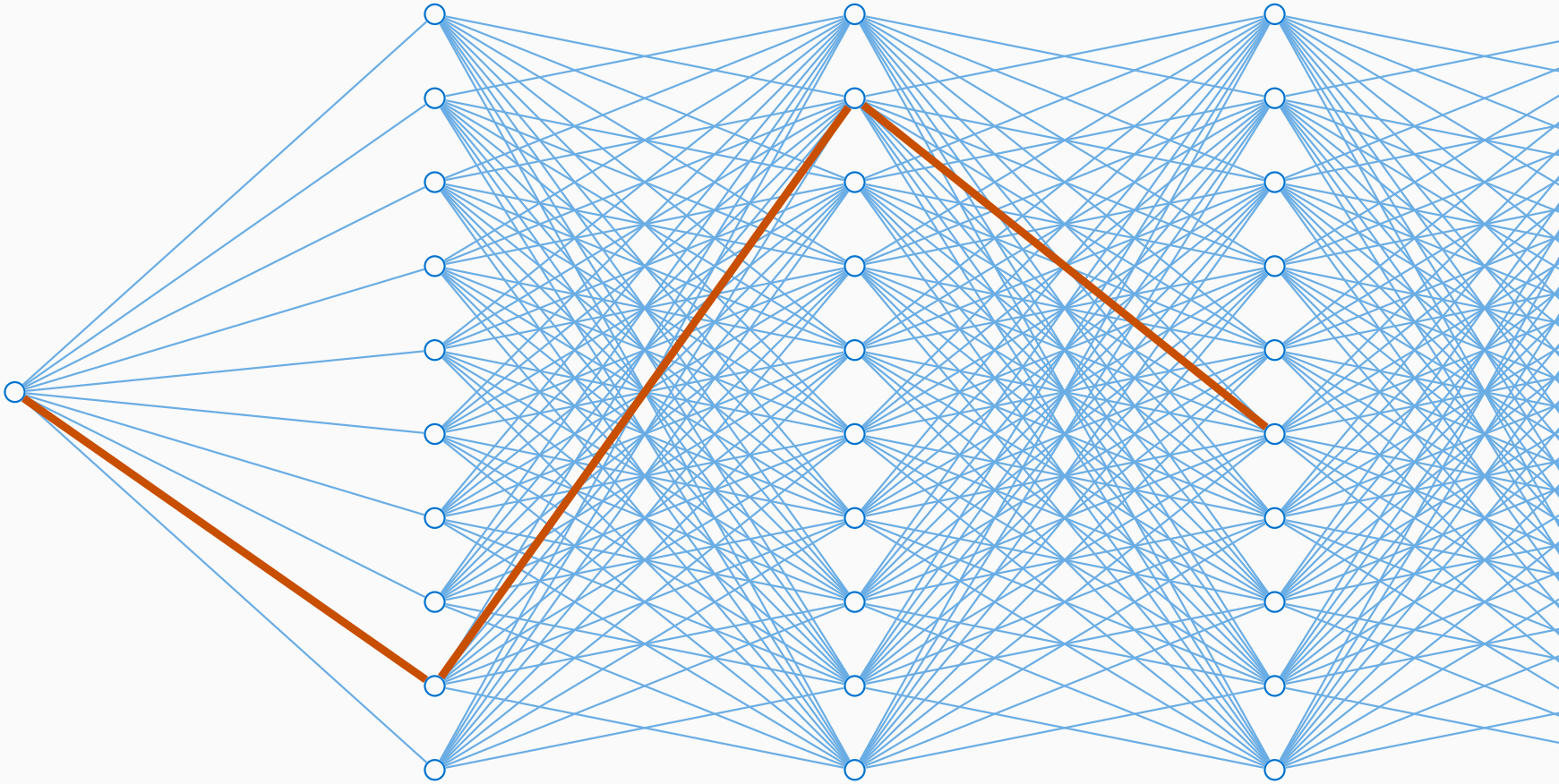
MS

SNP	p
rs2104286	$< 2 \times 10^{-16}$
rs11256593	1×10^{-5}

T1D

SNP	p
rs61839660	$< 2 \times 10^{-16}$
rs11594656	7×10^{-12}
rs12220852	1×10^{-9}
rs41295159	1×10^{-7}

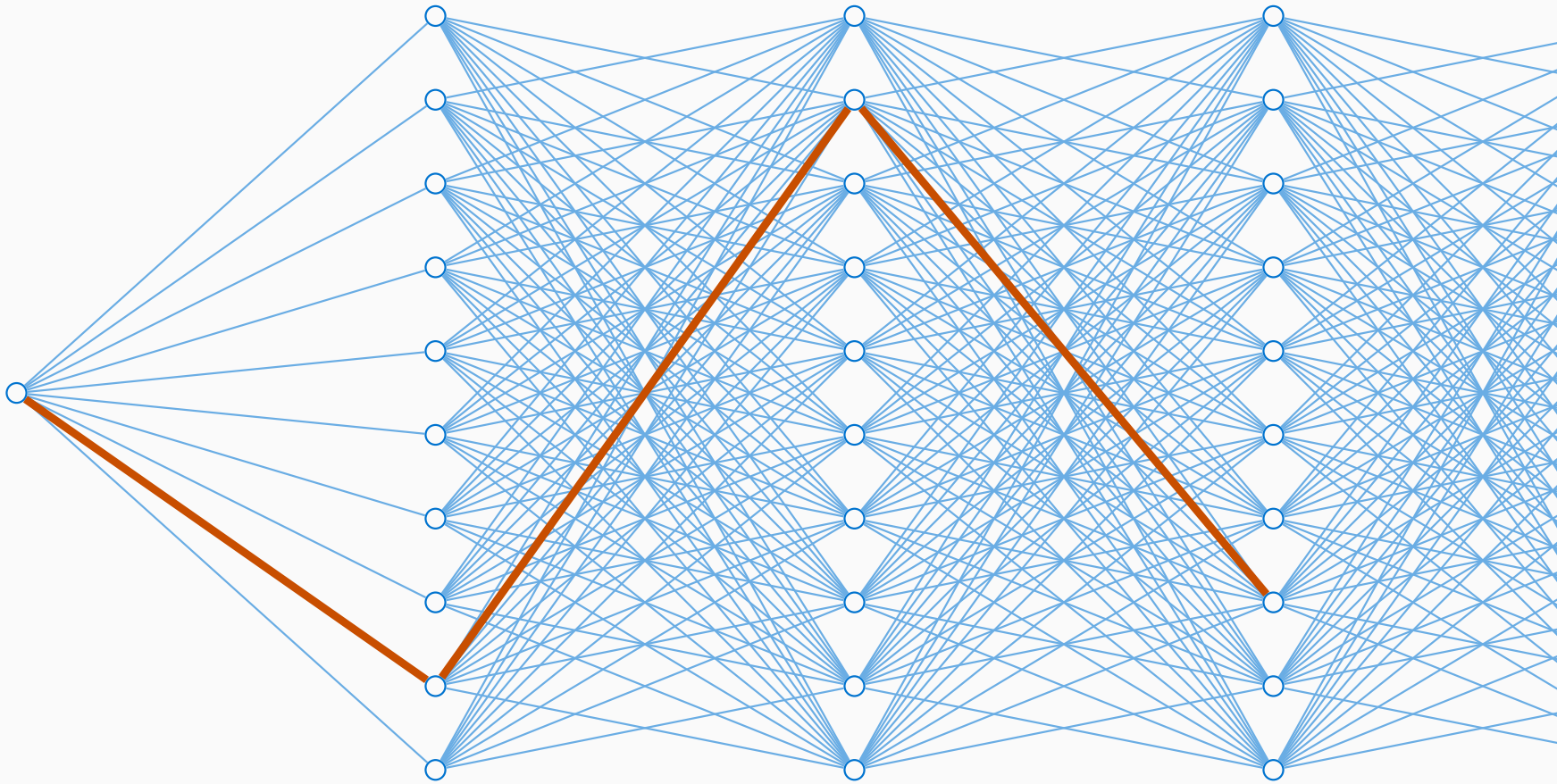
Alternative: evolutionary stochastic search



no SNPs 1 SNP 2 SNPs 3 SNPs

Botolo et al, PLoS Genet 2013; Wallace et al, PLoS Genet 2015

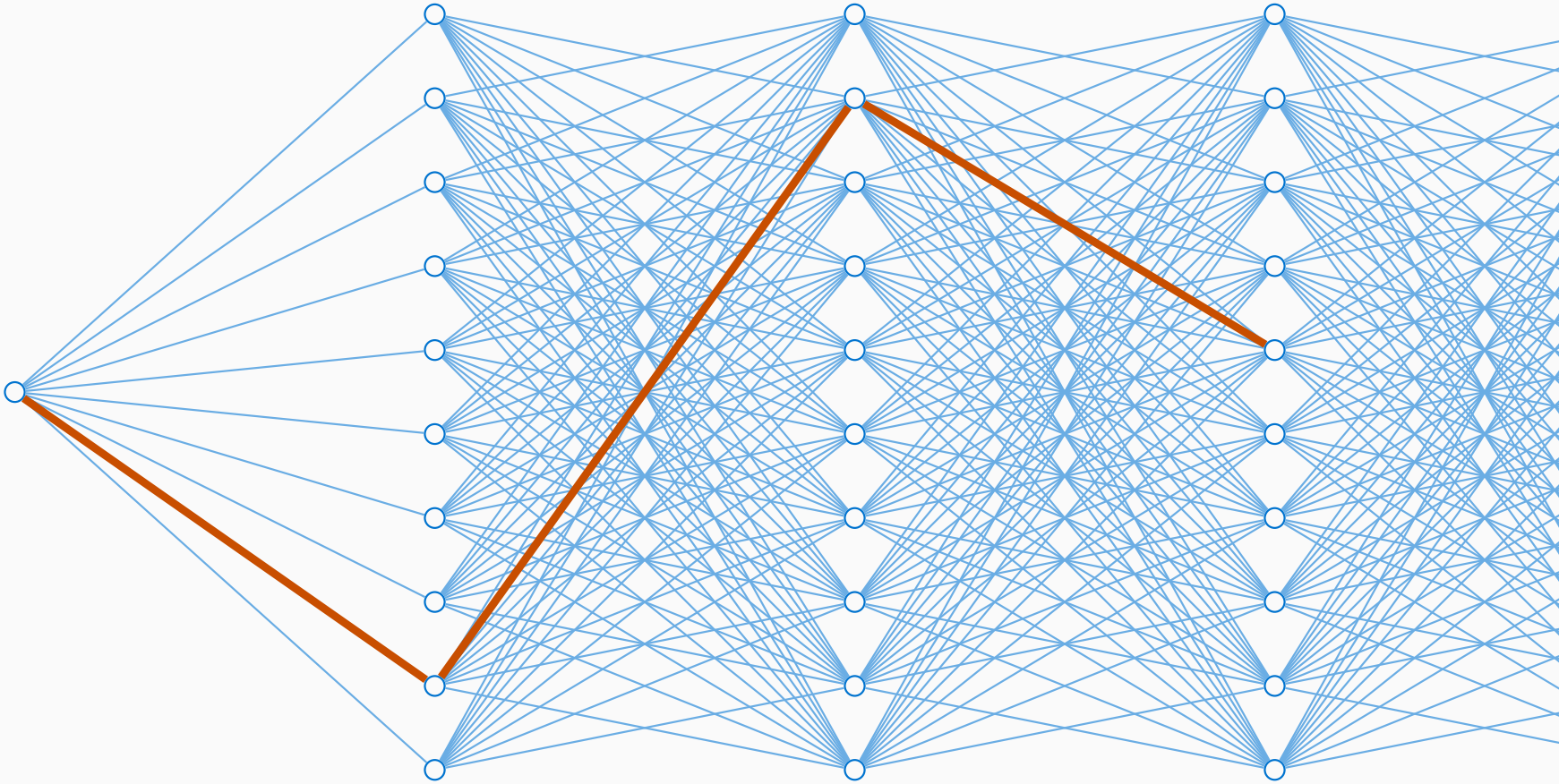
Alternative: evolutionary stochastic search



no SNPs 1 SNP 2 SNPs 3 SNPs

Botolo et al, PLoS Genet 2013; Wallace et al, PLoS Genet 2015

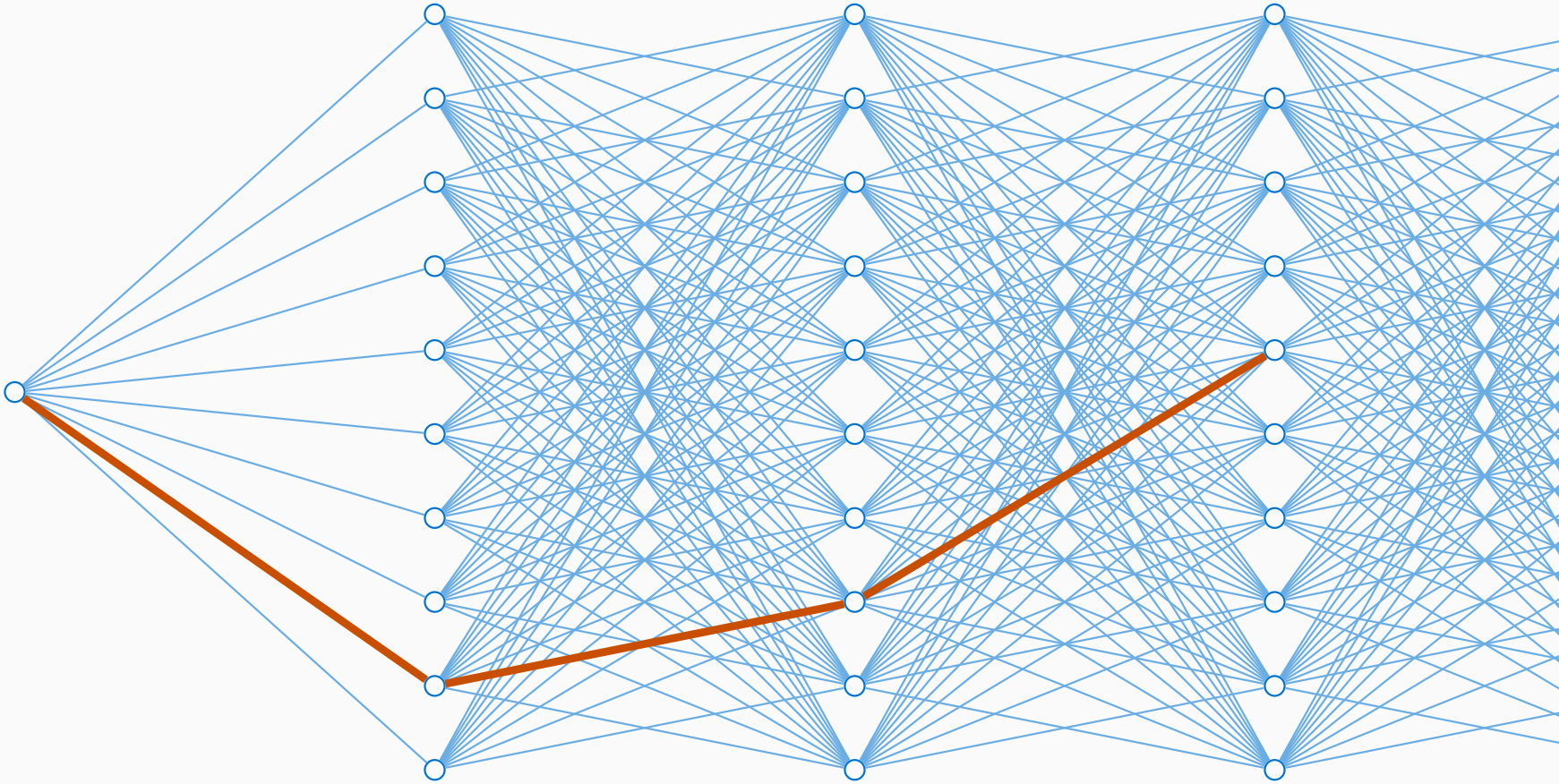
Alternative: evolutionary stochastic search



no SNPs 1 SNP 2 SNPs 3 SNPs

Botolo et al, PLoS Genet 2013; Wallace et al, PLoS Genet 2015

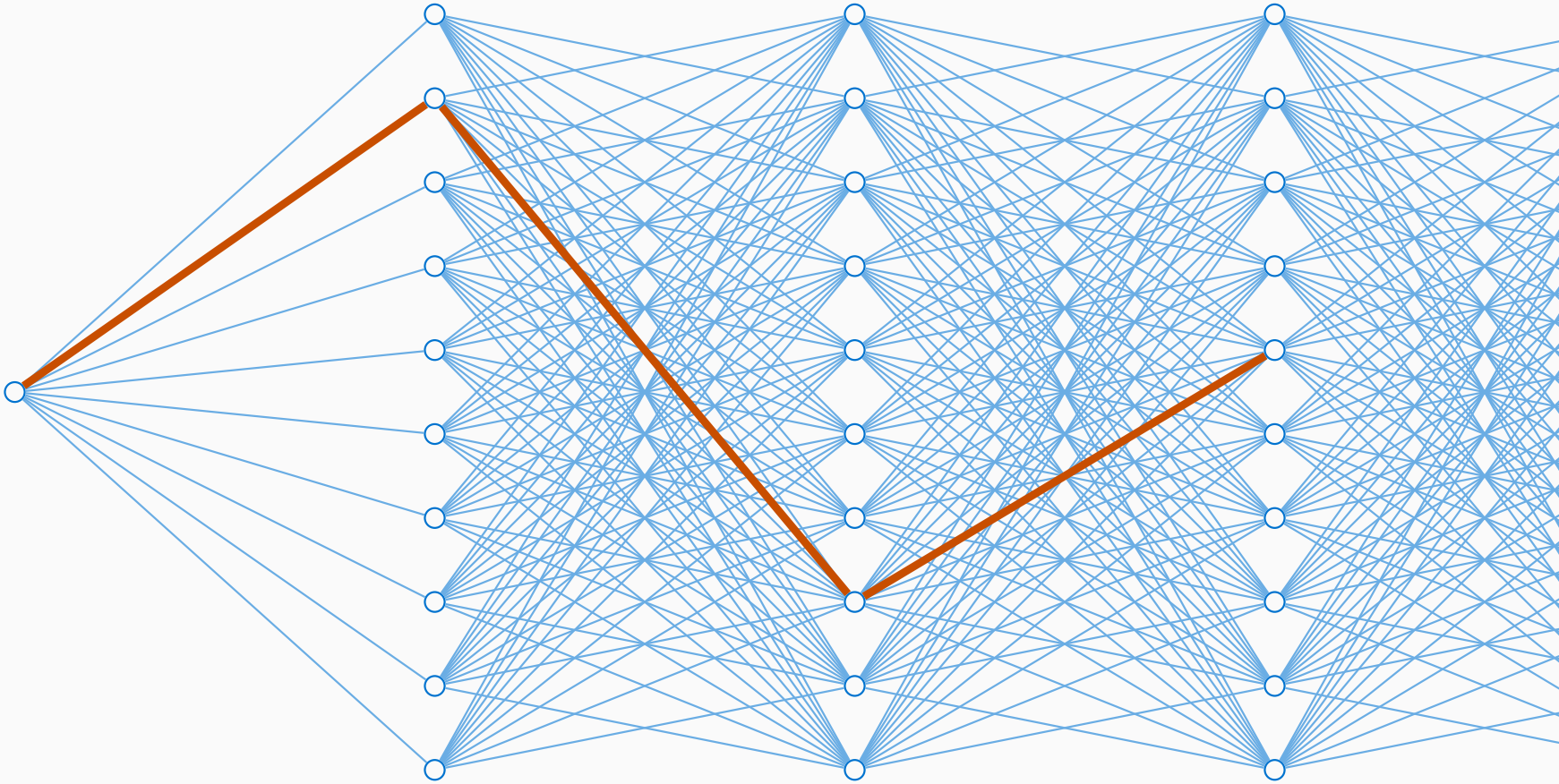
Alternative: evolutionary stochastic search



no SNPs 1 SNP 2 SNPs 3 SNPs

Botolo et al, PLoS Genet 2013; Wallace et al, PLoS Genet 2015

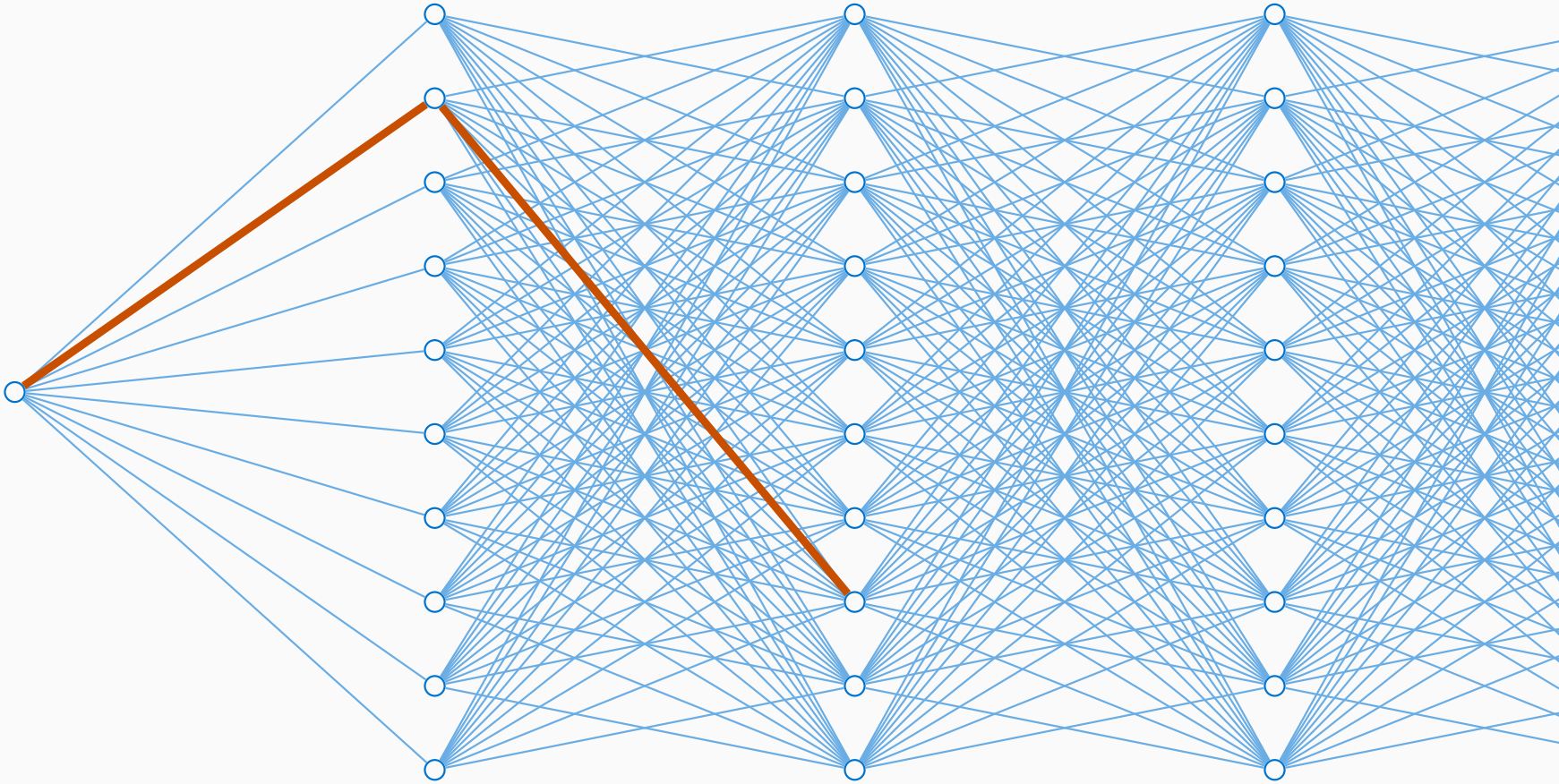
Alternative: evolutionary stochastic search



no SNPs 1 SNP 2 SNPs 3 SNPs

Botolo et al, PLoS Genet 2013; Wallace et al, PLoS Genet 2015

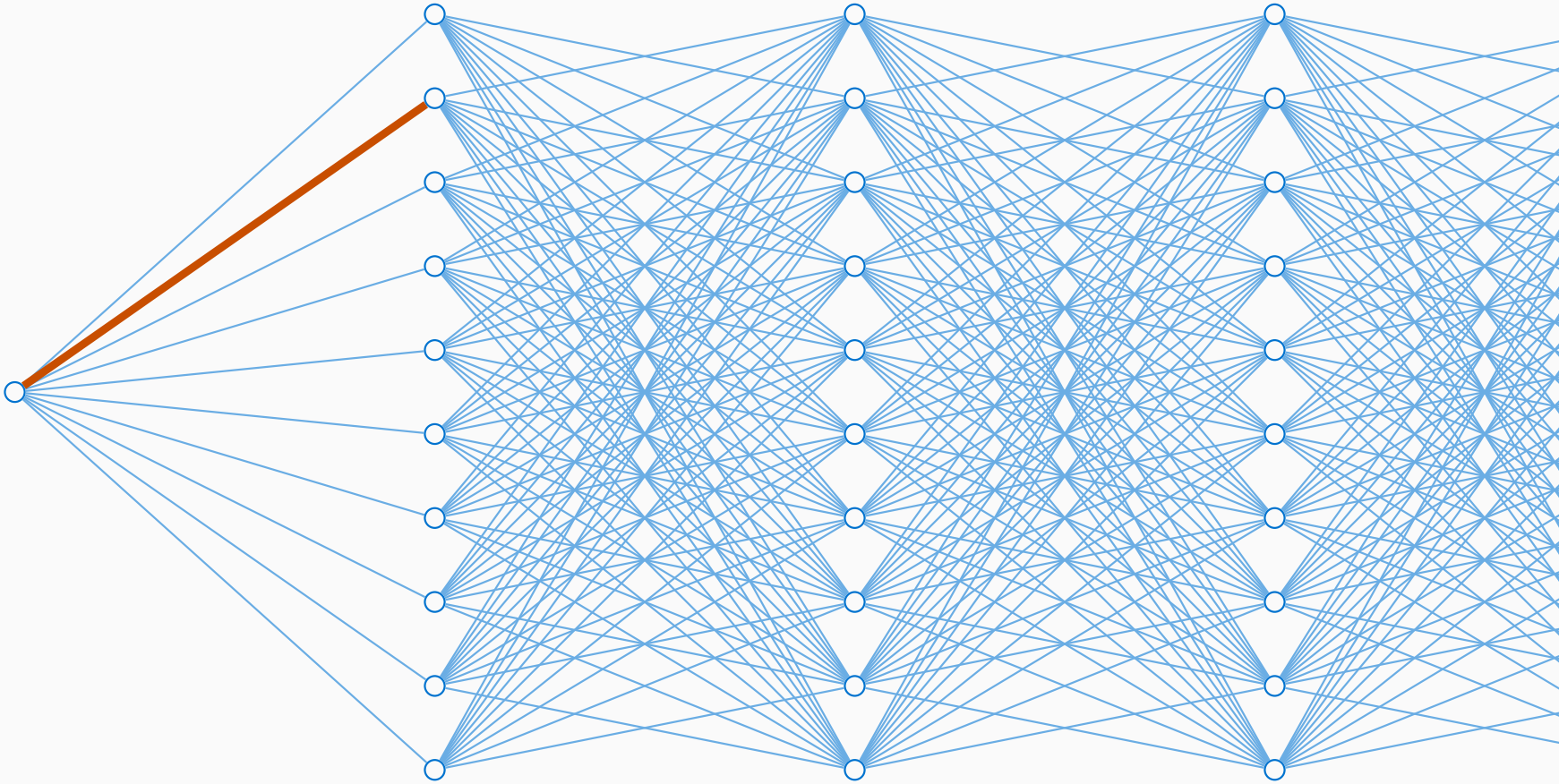
Alternative: evolutionary stochastic search



no SNPs 1 SNP 2 SNPs 3 SNPs

Botolo et al, PLoS Genet 2013; Wallace et al, PLoS Genet 2015

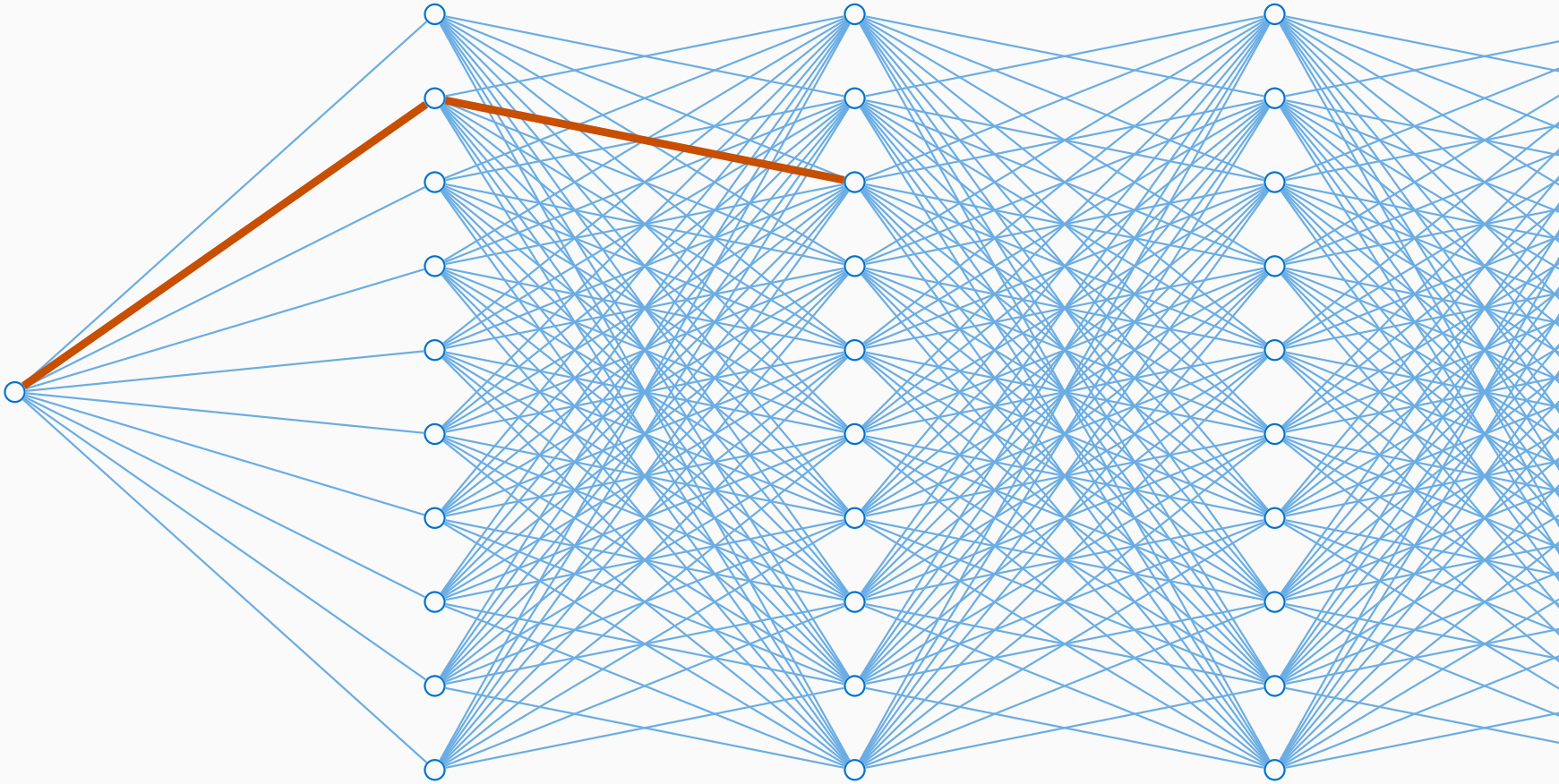
Alternative: evolutionary stochastic search



no SNPs 1 SNP 2 SNPs 3 SNPs

Botolo et al, PLoS Genet 2013; Wallace et al, PLoS Genet 2015

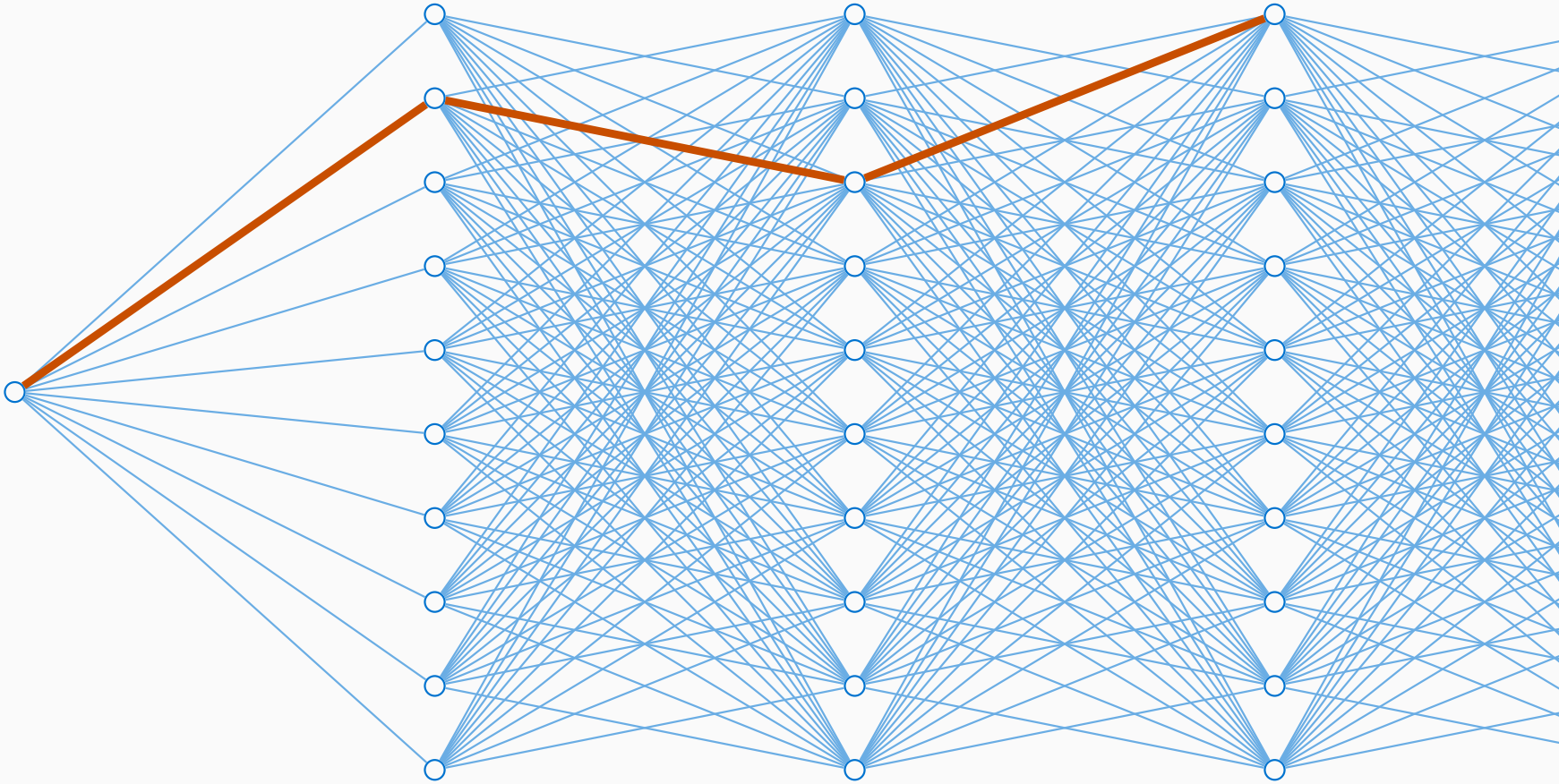
Alternative: evolutionary stochastic search



no SNPs 1 SNP 2 SNPs 3 SNPs

Botolo et al, PLoS Genet 2013; Wallace et al, PLoS Genet 2015

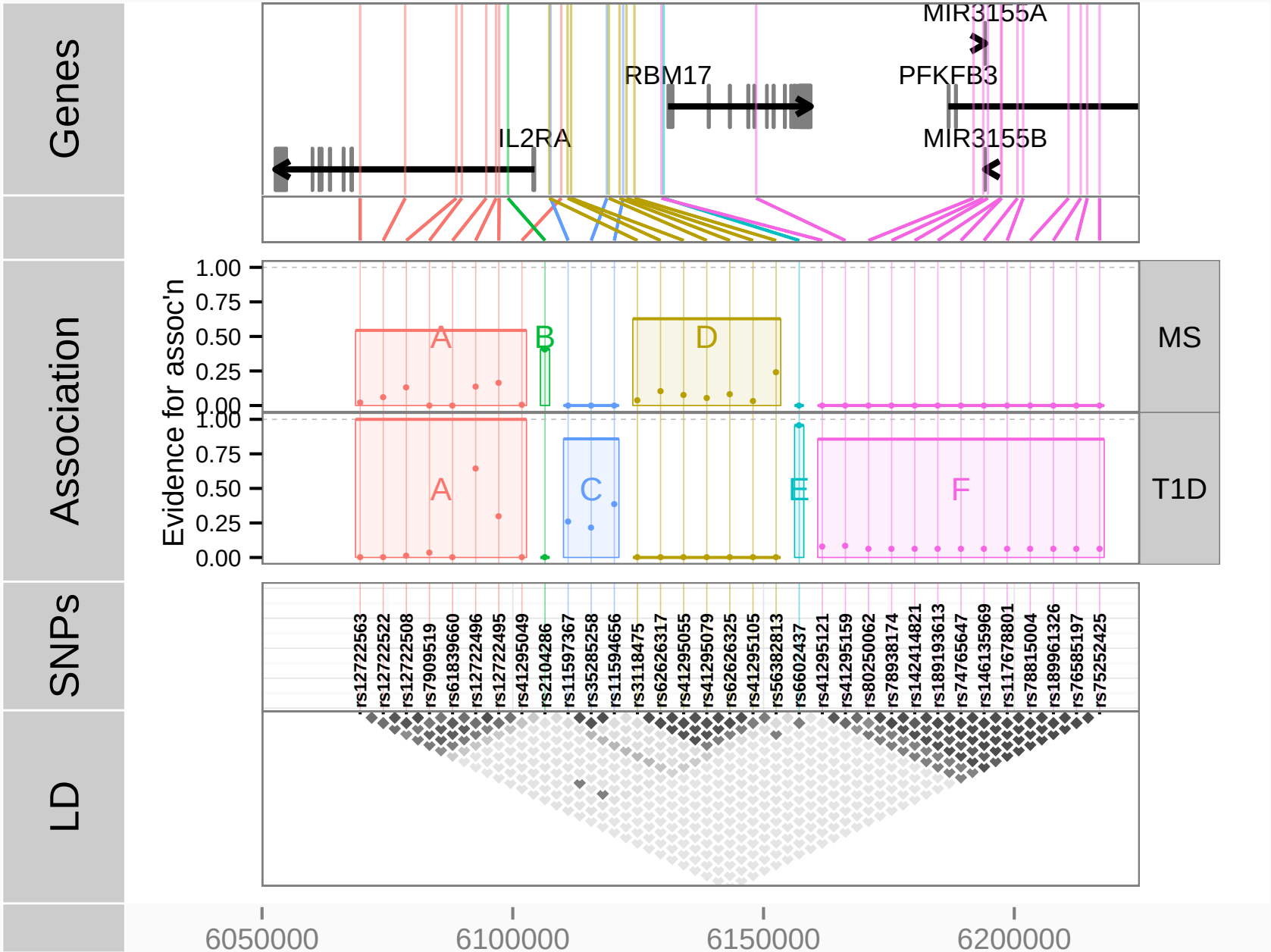
Alternative: evolutionary stochastic search



no SNPs 1 SNP 2 SNPs 3 SNPs

Botolo et al, PLoS Genet 2013; Wallace et al, PLoS Genet 2015

Stochastic search of IL2RA region



(Dis-) agreement between stepwise and stochastic search

MS

Stepwise Stochastic	rs2104286 (B)	rs11256593
M1: rs2104286 (B)	**1.00**	0.33
M2: rs12722496 (A)	0.33	0.08
M2: rs56382813 (D)	0.31	0.29

r^2 between SNPs

T1D

Stepwise Stochastic	rs61839660 (A)	rs11594656 (C)	rs12220852 (E)	rs41295159 (F)
rs12722496 (A)	**0.88**	0.02	0.02	0.00
rs11594656 (C)	0.02	**1.00**	0.26	0.02
rs6602437 (E)	0.05	0.25	**0.62**	0.01
rs41295159 (F)	0.00	0.02	0.00	**1.00**

Wallace et al, PLoS Genet 2015

Systematic comparison: 89 genetic regions, 6 diseases

Group	Number	Group	Number
Autoimm. Thyroid Disease, ATD	2772	Celiac Disease, CEL	12041
Juvenile Idiopathic Arthritis, JIA	1214	Multiple Sclerosis, MS	4461
Rhemuatoid Arthritis, RA	11475	Type 1 Diabetes, T1D	6681
CONTROL	22997		

201 region/disease pairs showing association (min. $p < 10^{-6}$)

Systematic comparison: 89 genetic regions, 6 diseases

Group	Number	Group	Number
Autoimm. Thyroid Disease, ATD	2772	Celiac Disease, CEL	12041
Juvenile Idiopathic Arthritis, JIA	1214	Multiple Sclerosis, MS	4461
Rhemuatoid Arthritis, RA	11475	Type 1 Diabetes, T1D	6681
CONTROL	22997		

201 region/disease pairs showing association (min. $p < 10^{-6}$)

Regions	Region-disease pairs	
62	171	matched
2	2	stochastic null ($p \simeq 1 \times 10^{-6}$)
15	17	stepwise nested in stochastic
5	5	different top SNP (two weak signals)
5	6	non-nested mismatch

Systematic comparison: 89 genetic regions, 6 diseases

Group	Number	Group	Number
Autoimm. Thyroid Disease, ATD	2772	Celiac Disease, CEL	12041
Juvenile Idiopathic Arthritis, JIA	1214	Multiple Sclerosis, MS	4461
Rhemuatoid Arthritis, RA	11475	Type 1 Diabetes, T1D	6681
CONTROL	22997		

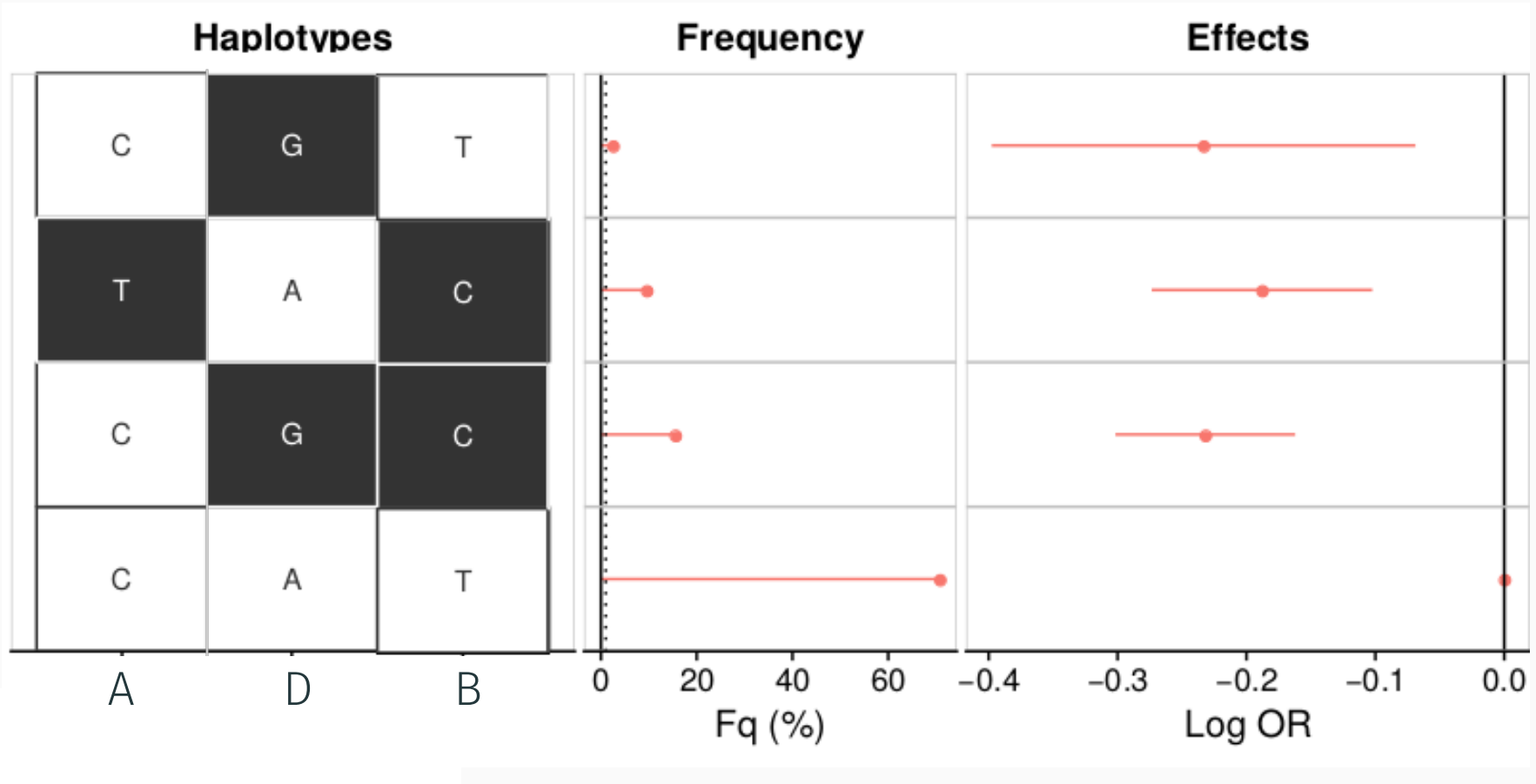
201 region/disease pairs showing association (min. $p < 10^{-6}$)

Regions	Region-disease pairs	
62	171	matched
2	2	stochastic null ($p \simeq 1 \times 10^{-6}$)
15	17	stepwise nested in stochastic
5	5	different top SNP (two weak signals)
5	6	non-nested mismatch

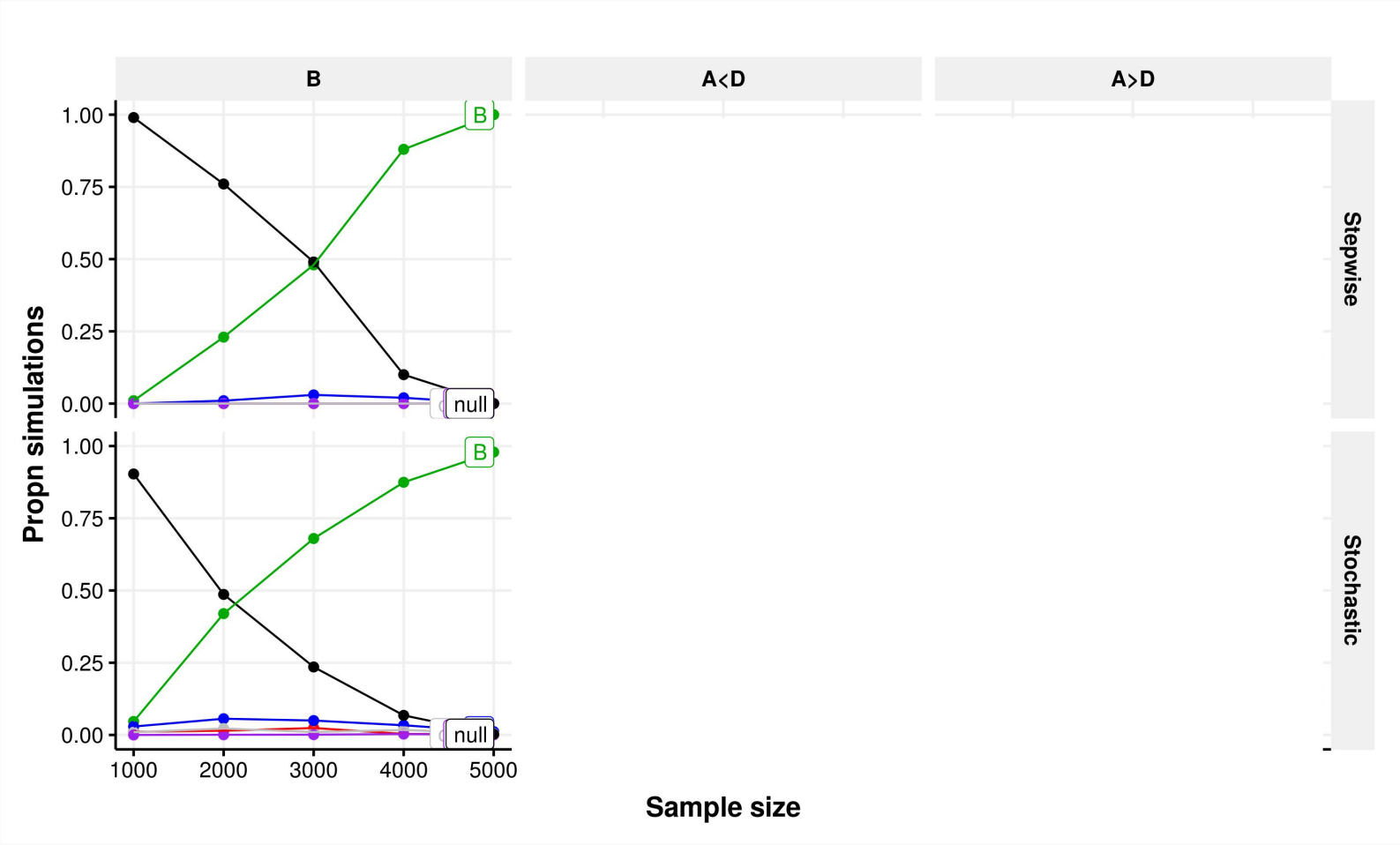
20/30 regions with > 1 associated disease had a shared signal

Joint tagging of causal variants

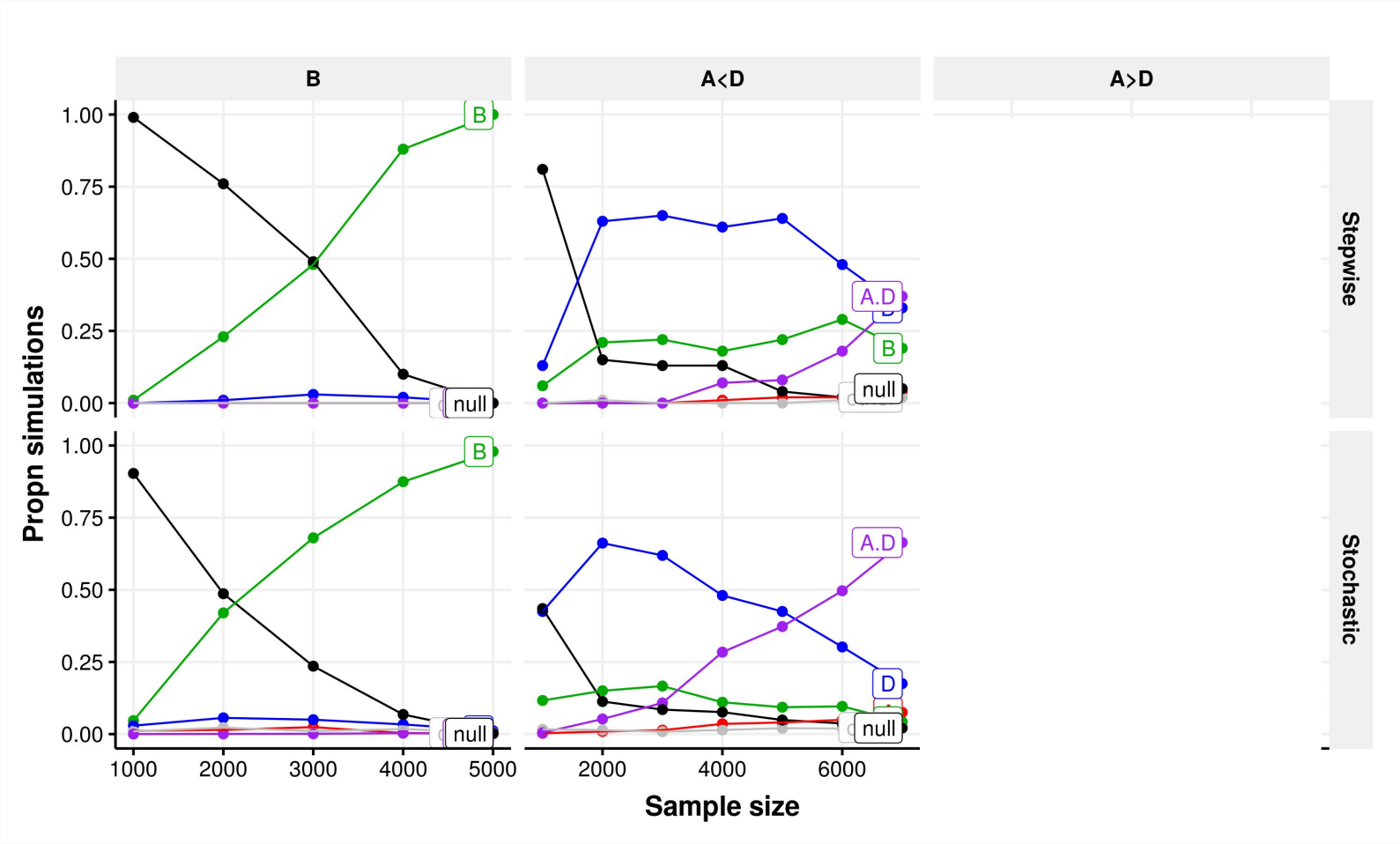
Haplotype analysis of MS



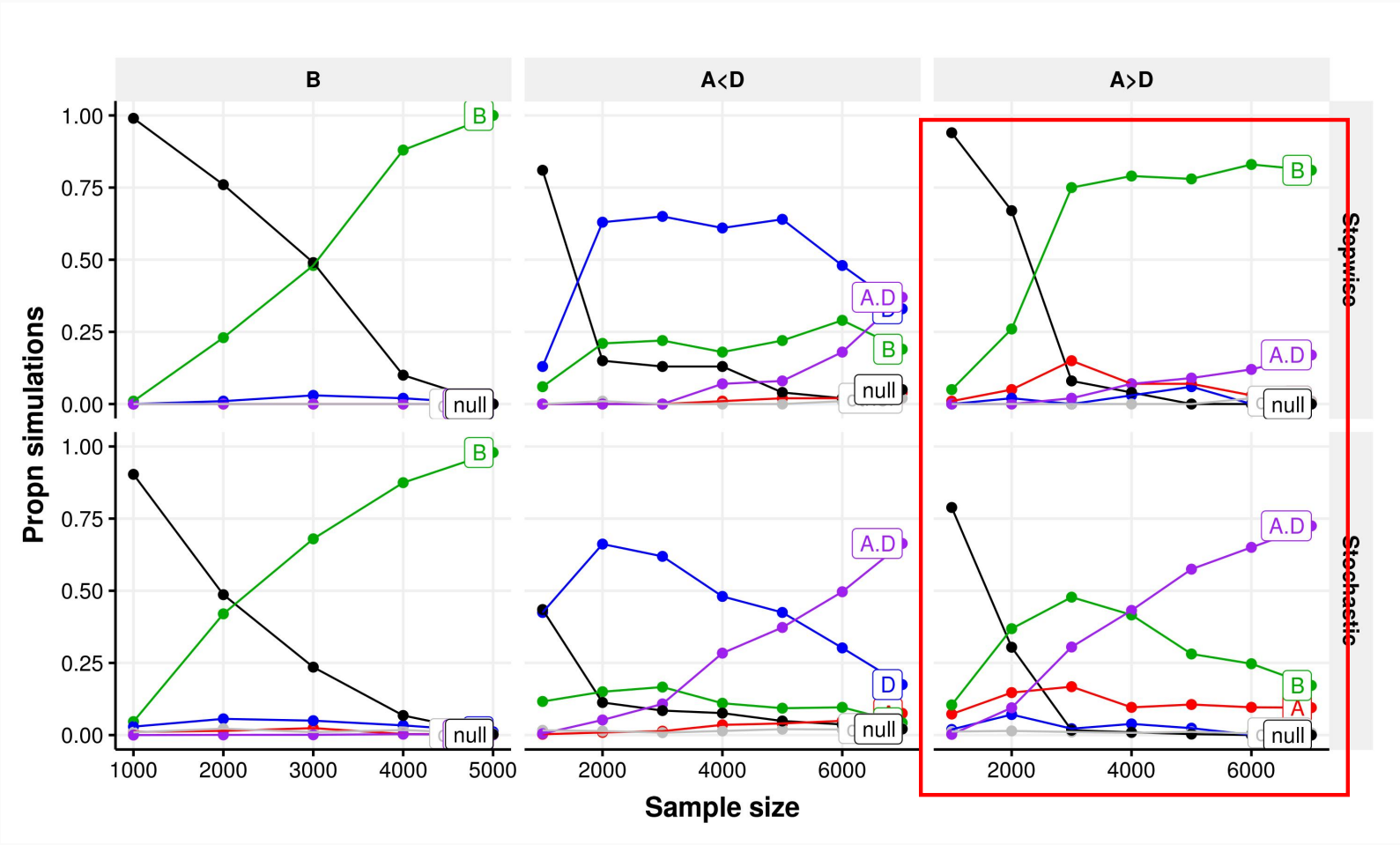
Stochastic search “correctness” sample size dependent



Stochastic search “correctness” sample size dependent



Stochastic search “correctness” sample size dependent



Conditions for joint tagging

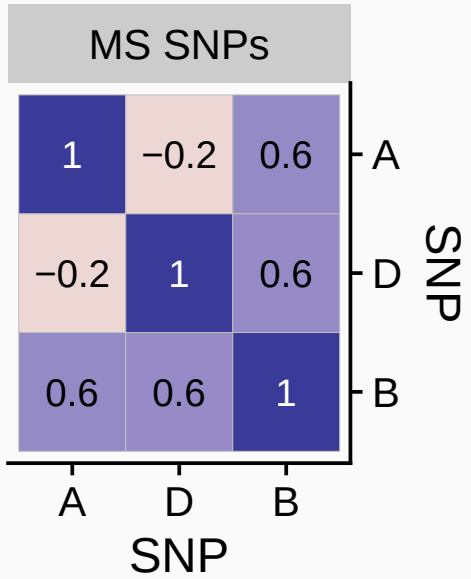
SNPs X_i, $i = 1, .., p$. $X_1, .., X_k$ are causal ($k < p$) correlation matrix is Σ

Expected Z score from a joint model is μ

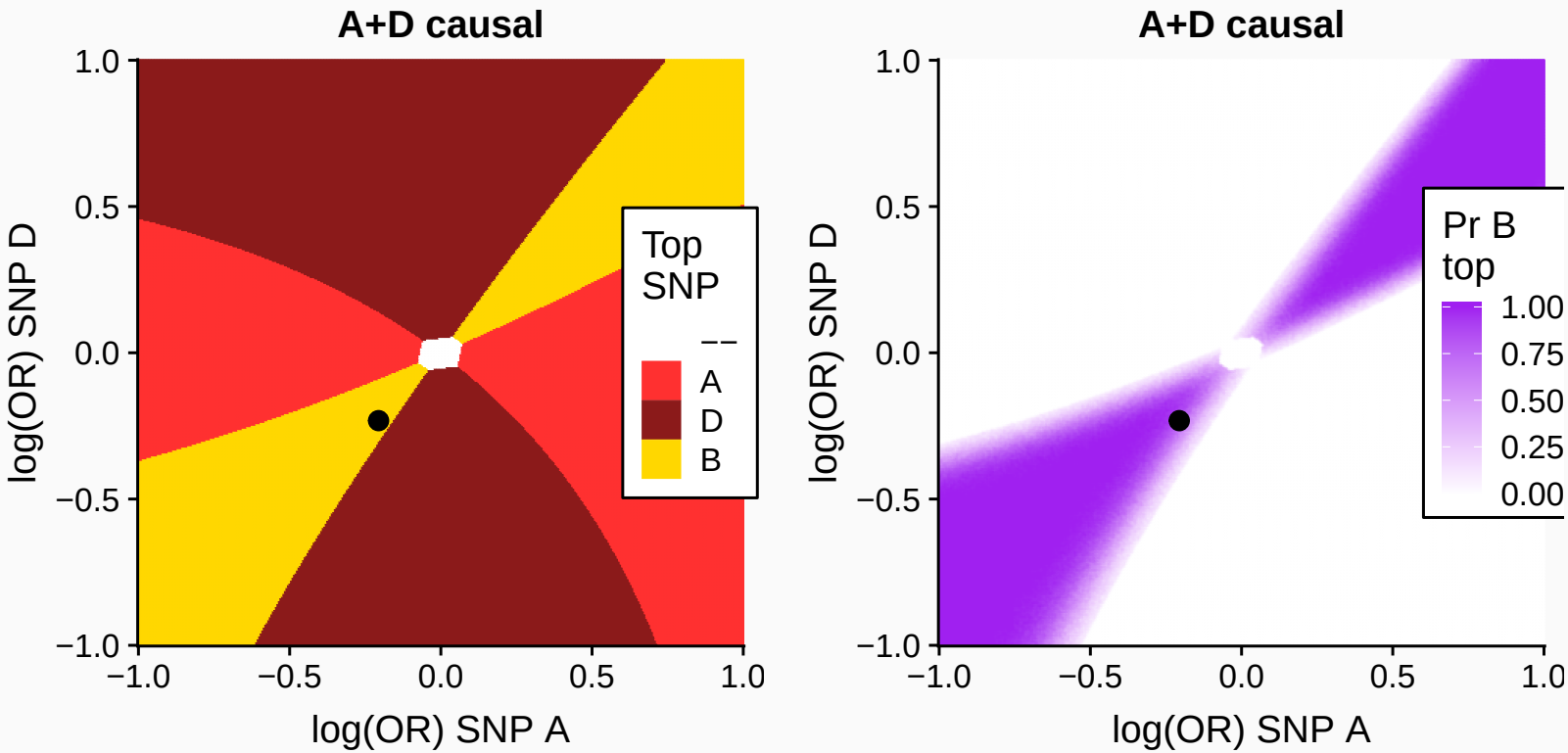
$$\mu_i = \begin{cases} f(\text{MAF, OR, N}) & i \leq k \\ 0 & i > k \end{cases}$$

Marginal Z scores across all SNPs:

$$Z \sim N(\Sigma\mu, \Sigma)$$

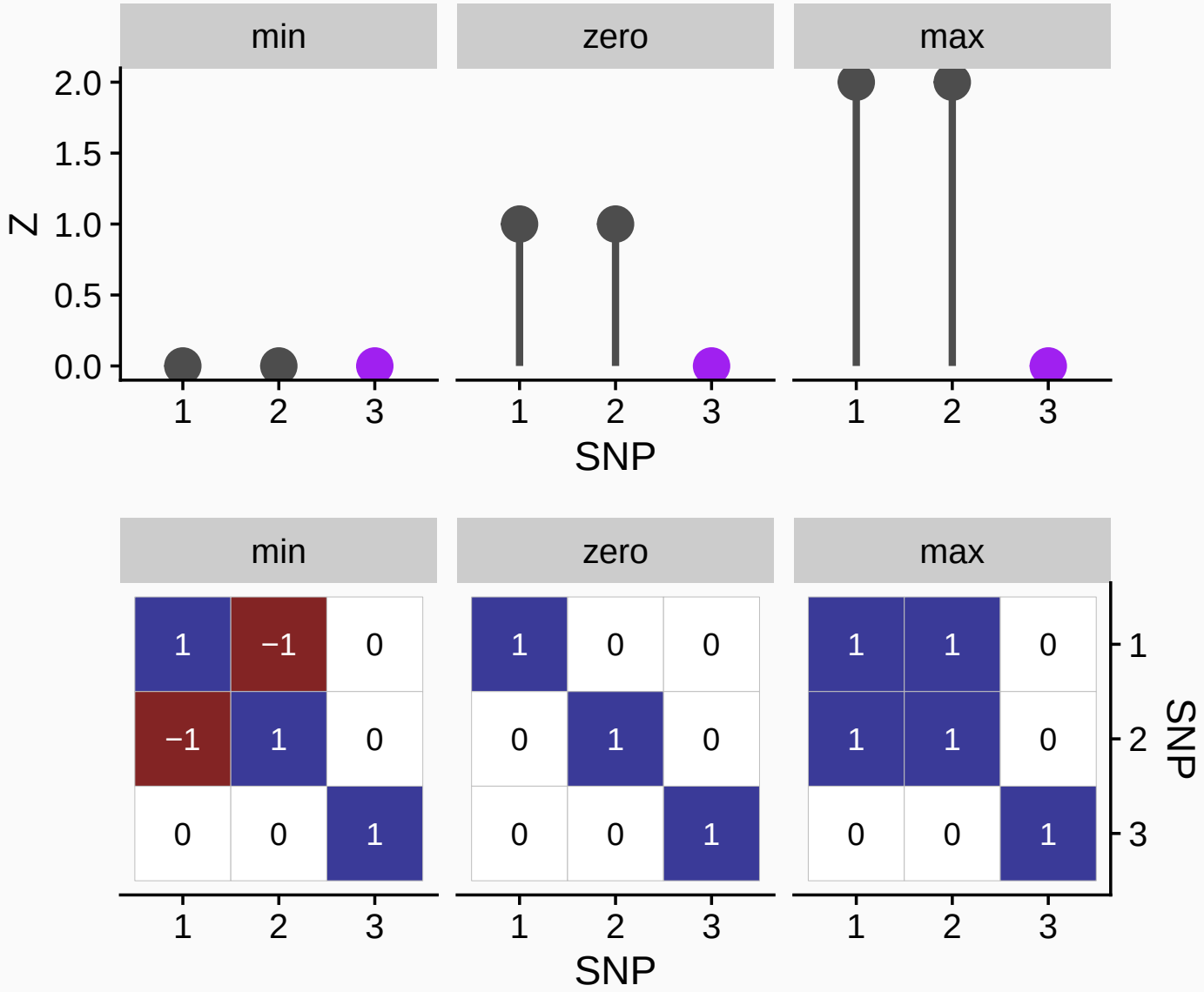


A, D causal, similar effects: expect B to have smallest p value



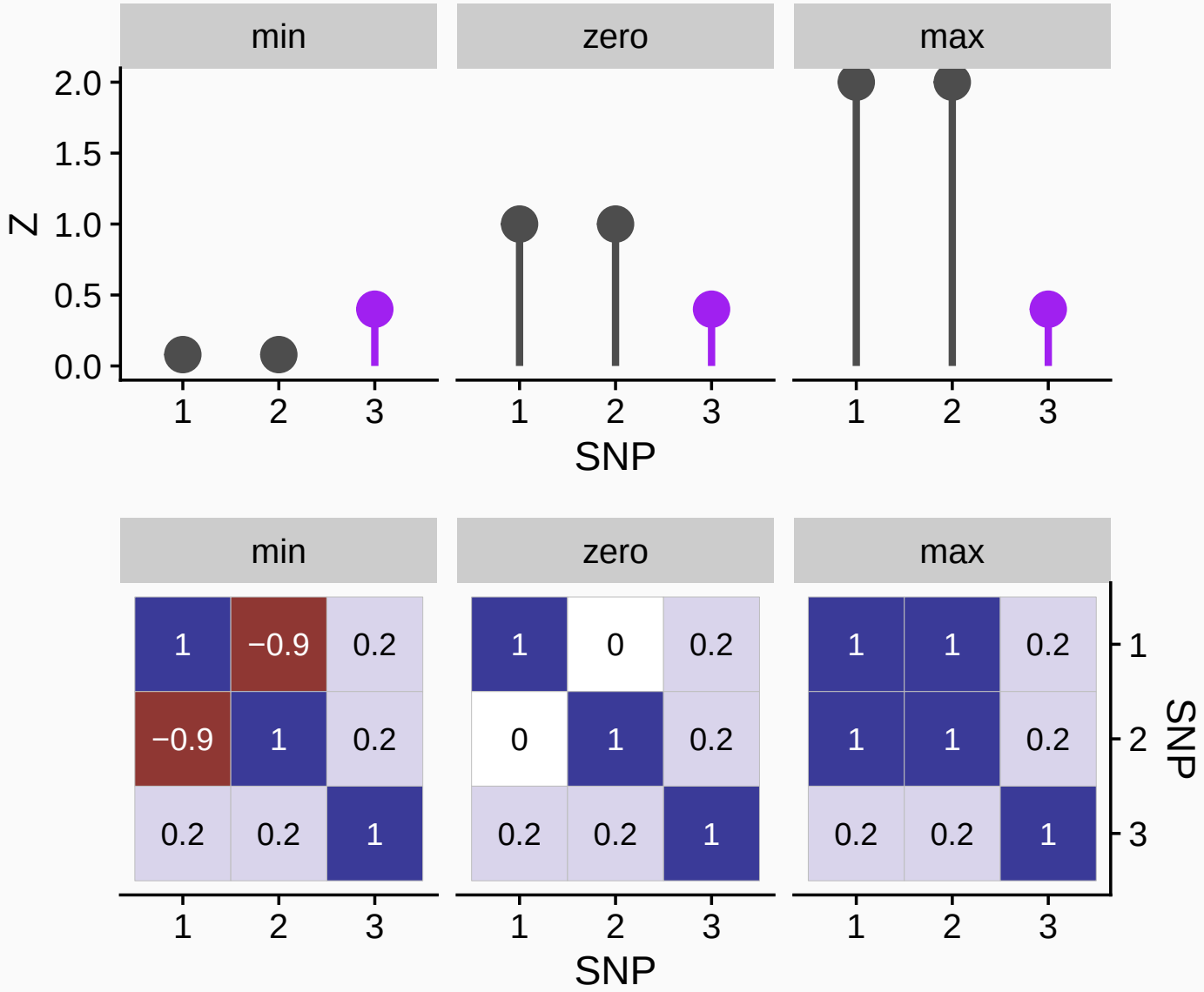
Effect of Σ on potential for joint tagging

SNPs 1, 2 causal. SNP 3 potential tag. $\mu = (1, 1, 0)'$



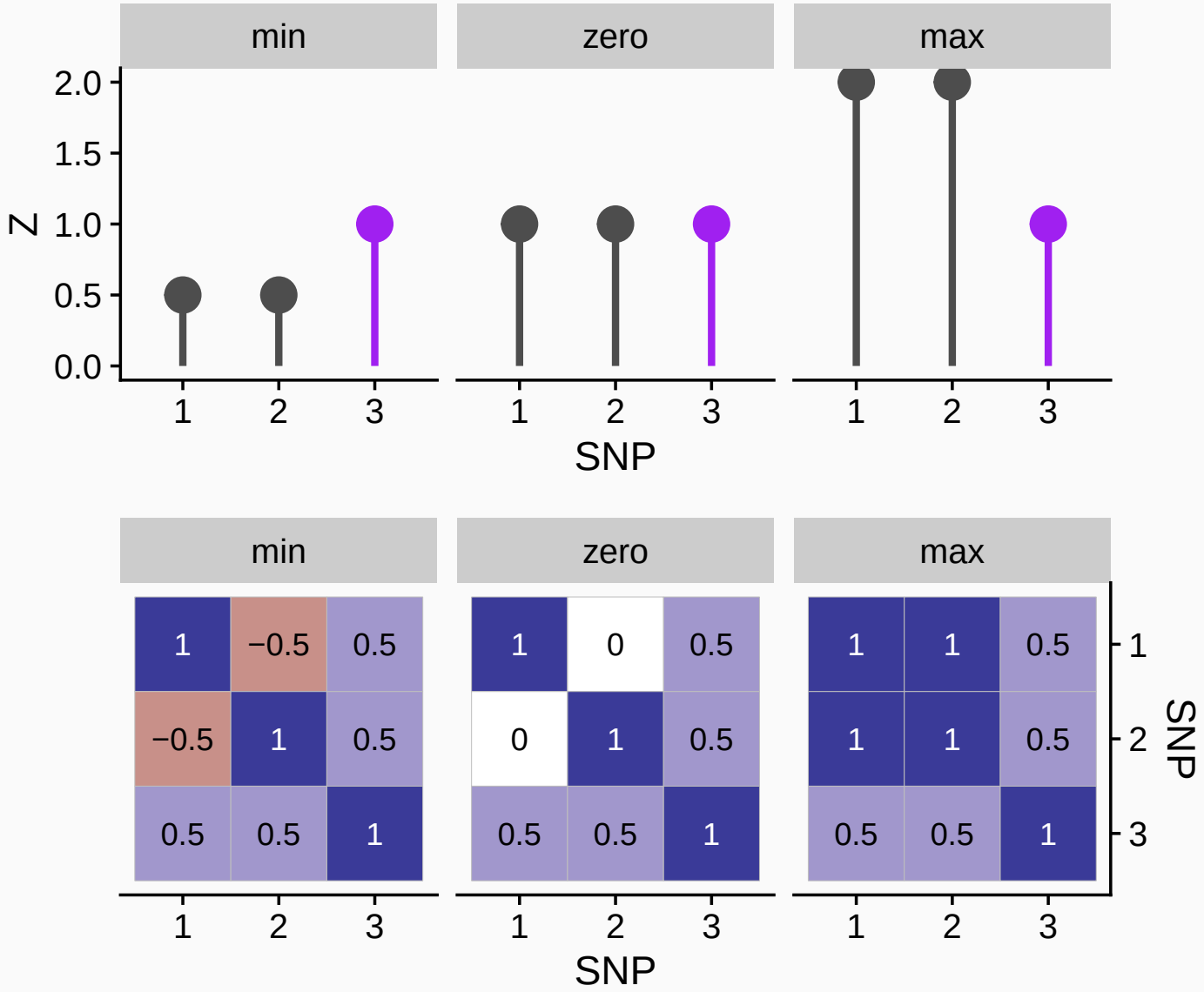
Effect of Σ on potential for joint tagging

SNPs 1, 2 causal. SNP 3 potential tag. $\mu = (1, 1, 0)'$



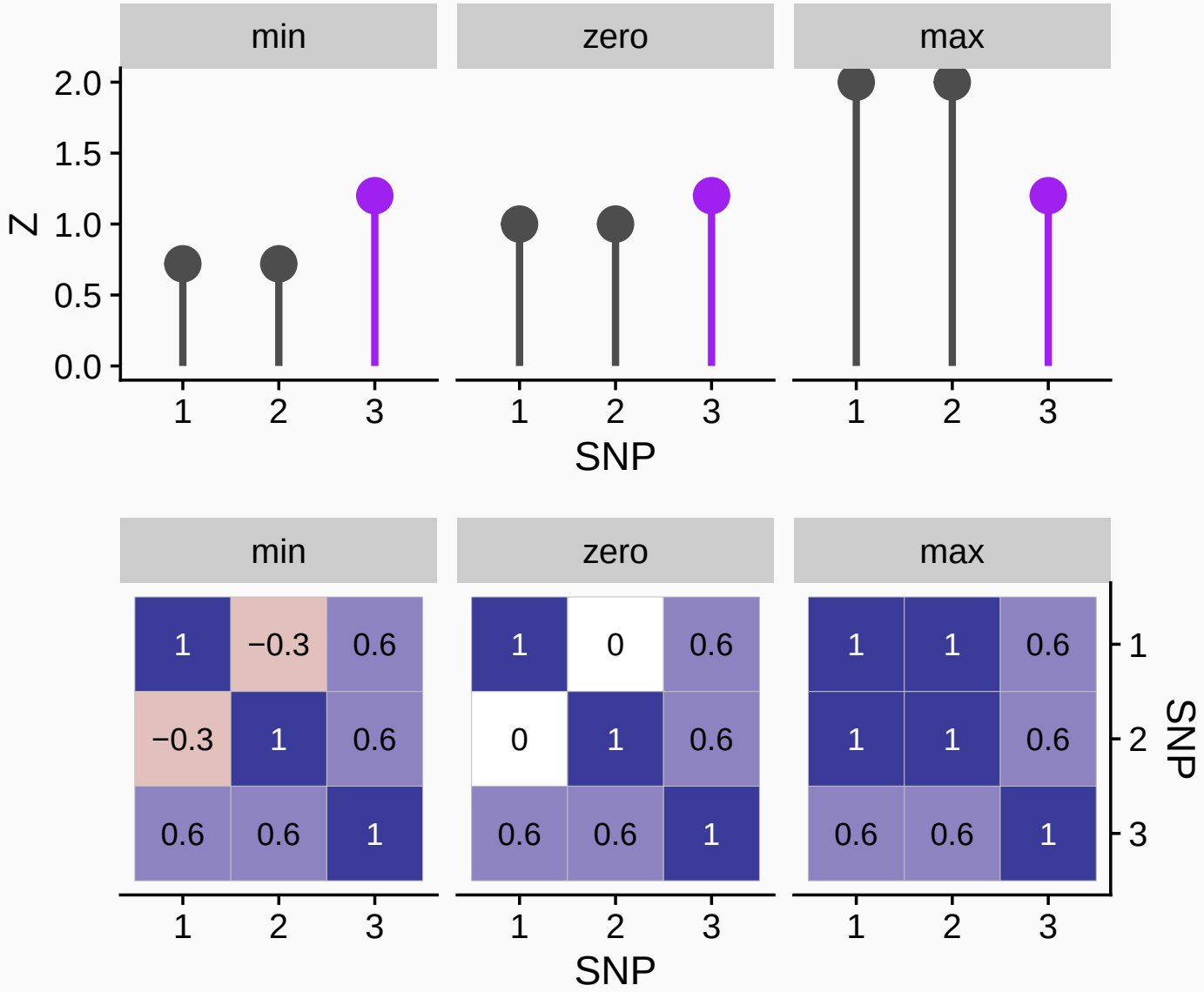
Effect of Σ on potential for joint tagging

SNPs 1, 2 causal. SNP 3 potential tag. $\mu = (1, 1, 0)'$



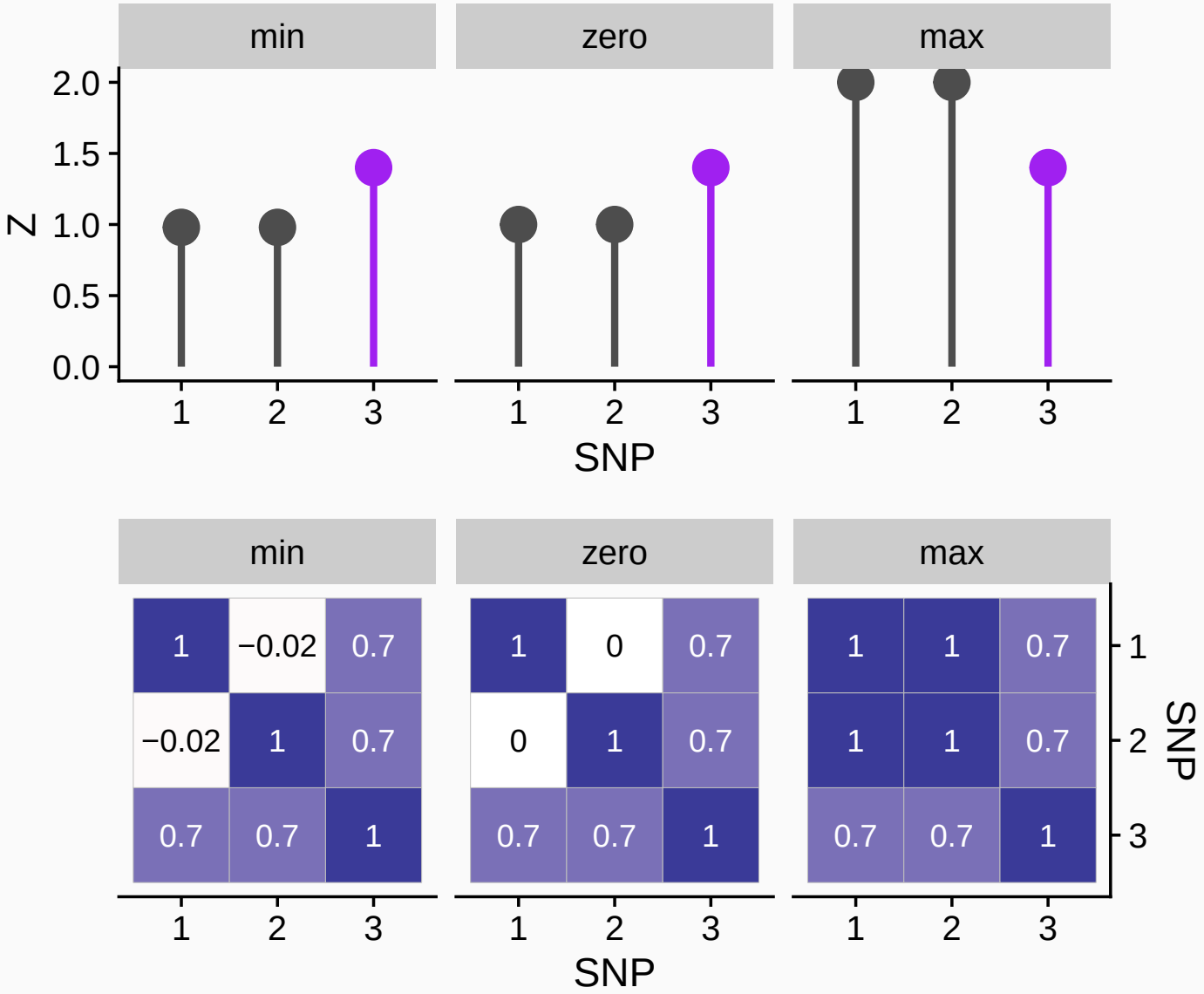
Effect of Σ on potential for joint tagging

SNPs 1, 2 causal. SNP 3 potential tag. $\mu = (1, 1, 0)'$

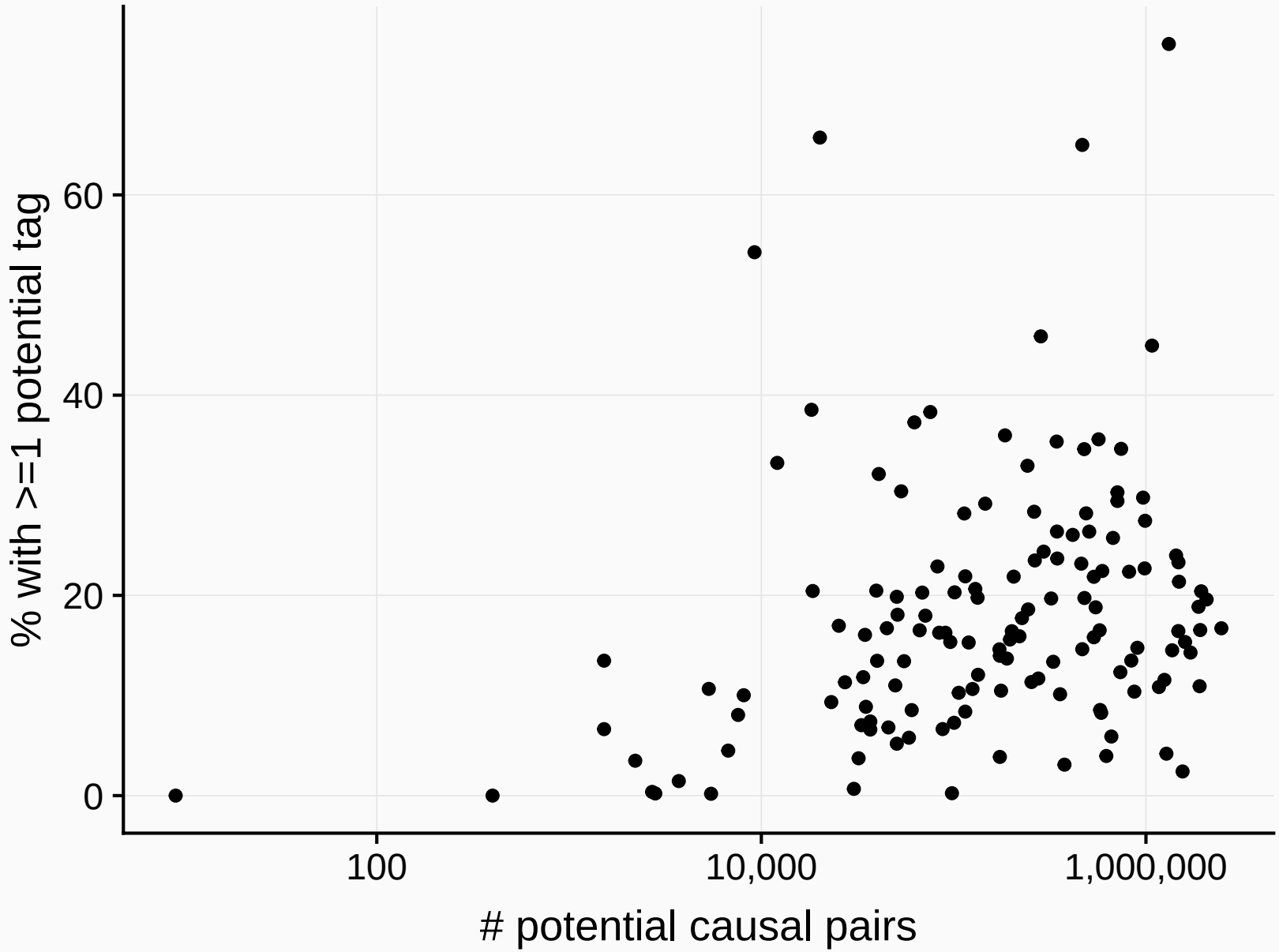


Effect of Σ on potential for joint tagging

SNPs 1, 2 causal. SNP 3 potential tag. $\mu = (1, 1, 0)'$



Potential tags by LD pattern at 20% of variant pairs



Improved fine mapping by exploiting shared aetiology

Bayesian fine mapping

Single disease

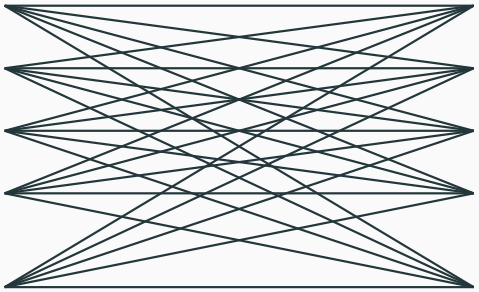
Model	Prior	Data	Posterior
A	π_A	BF_A	$\propto \pi_A BF_A$
B	π_B	BF_B	$\propto \pi_B BF_B$
D	π_D	BF_D	$\propto \pi_D BF_D$
B+D	π_{B+D}	BF_{B+D}	$\propto \pi_{B+D} BF_{B+D}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Bayesian fine mapping

Two diseases

Disease 1

Model	Data
A	BF_A
B	BF_B
D	BF_D
B+D	BF_{B+D}
⋮	⋮



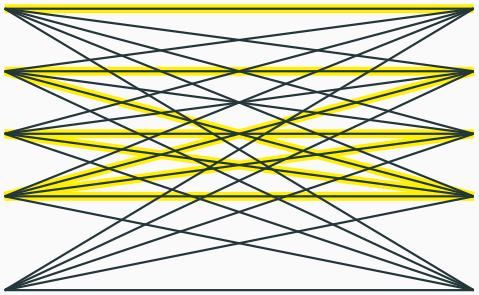
Disease 2

Model	Data
A	BF_A
B	BF_B
D	BF_D
B+D	BF_{B+D}
⋮	⋮

Bayesian fine mapping

Two diseases

Disease 1	
Model	Data
A	BF_A
B	BF_B
D	BF_D
B+D	BF_{B+D}
⋮	⋮



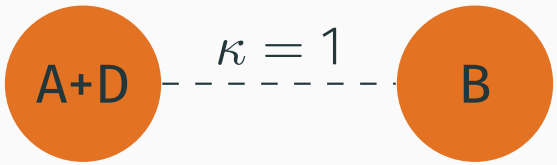
Disease 2	
Model	Data
A	BF_A
B	BF_B
D	BF_D
B+D	BF_{B+D}
⋮	⋮

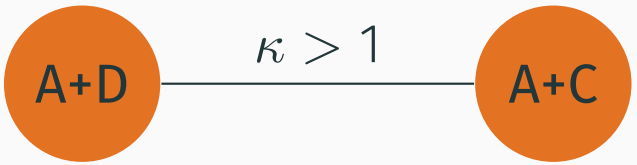
Use prior to borrow information between diseases

Define configurations: sets of models for each disease

$$C_{i,j} = \{M_i \text{ for disease 1}, M_j \text{ for disease 2}\}$$

$$Pr(C_{i,j}) = \begin{cases} Pr(M_i)Pr(M_j) \times \kappa \times \tau_{ij} & M_i \cap M_j \neq \varnothing \\ Pr(M_i)Pr(M_j) \times 1 \times \tau_{ij} & M_i \cap M_j = \varnothing \end{cases}$$





κ: upweighting factor

τ_{ij}: normalisation factor, fixes prior on the number of causal variants

Computational challenges of Bayesian fine mapping

Single disease

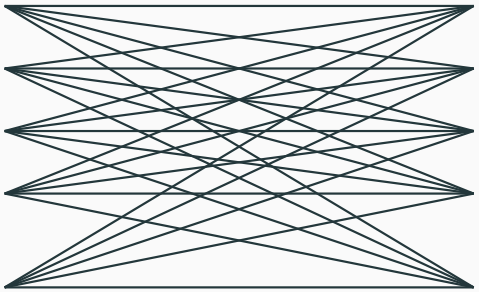
Model	Prior	Data	Posterior
A	π_A	BF_A	$\propto \pi_A BF_A$
B	π_B	BF_B	$\propto \pi_B BF_B$
D	π_D	BF_D	$\propto \pi_D BF_D$
B+D	π_{B+D}	BF_{B+D}	$\propto \pi_{B+D} BF_{B+D}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Model space: exponential in number of causal variants

Computational challenges of Bayesian fine mapping

Two diseases

Disease 1	
Model	Data
A	BF_A
B	BF_B
D	BF_D
B+D	BF_{B+D}
⋮	⋮



Disease 2	
Model	Data
A	BF_A
B	BF_B
D	BF_D
B+D	BF_{B+D}
⋮	⋮

Model space: (exp. causal variants)$^{\text{number of diseases}}$

Challenges: memory, computational time

Fast, memory efficient calculation of marginal posteriors

Speed: Joint Bayes factor approximated by function of single disease Bayes factors

$$BF(\{M_i, M_j\}) \underset{\sim}{\propto} BF(M_i) \times BF(M_j) \times \eta$$

η function of numbers of cases, shared controls and causal variants

Memory: linear (not exponential) in number of diseases, by storing only marginal single disease posteriors

Fast, memory efficient calculation of marginal posteriors

Speed: Joint Bayes factor approximated by function of single disease Bayes factors

$$BF(\{M_i, M_j\}) \propto BF(M_i) \times BF(M_j) \times \eta$$

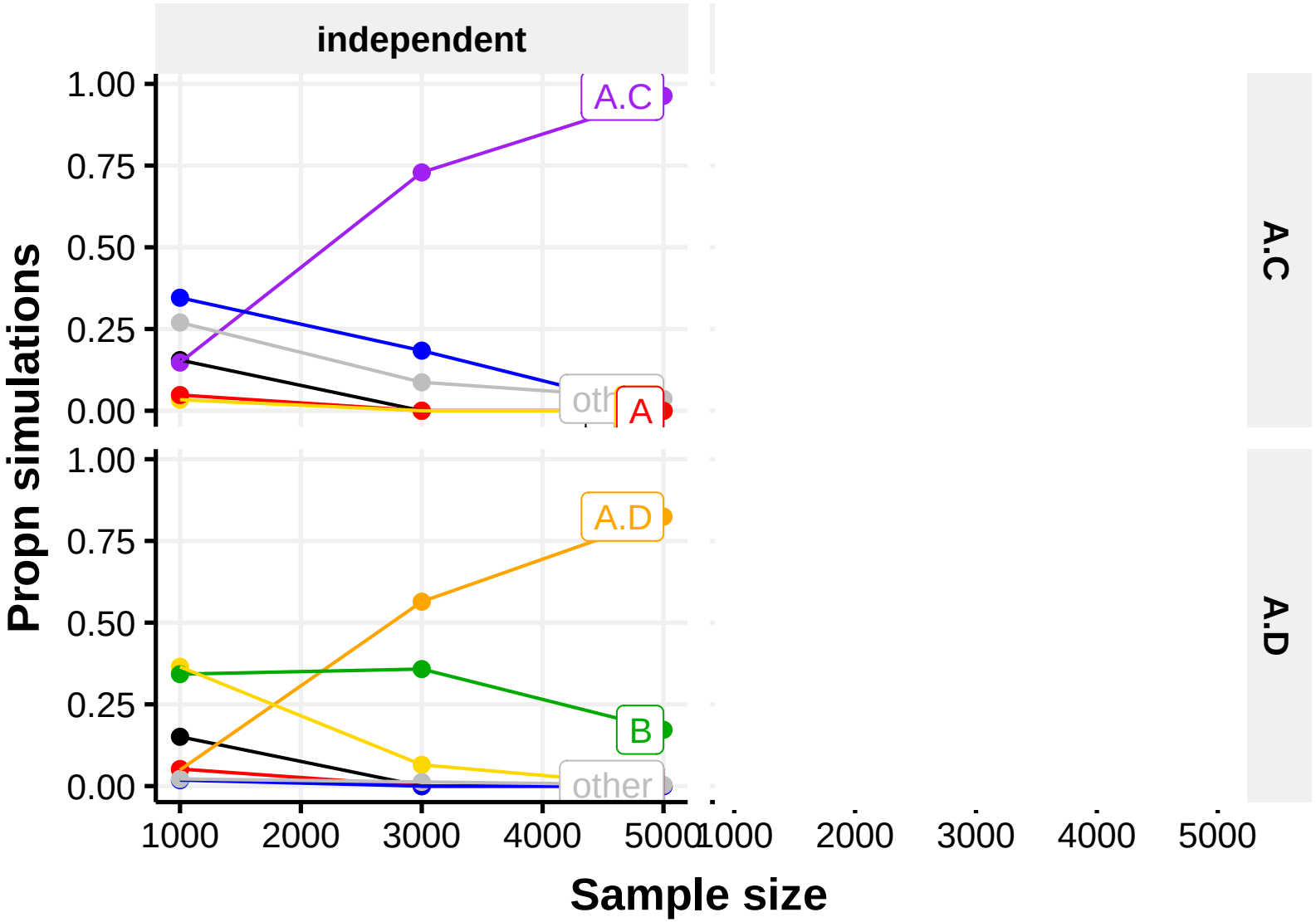
η function of numbers of cases, shared controls and causal variants

Memory: linear (not exponential) in number of diseases, by storing only marginal single disease posteriors

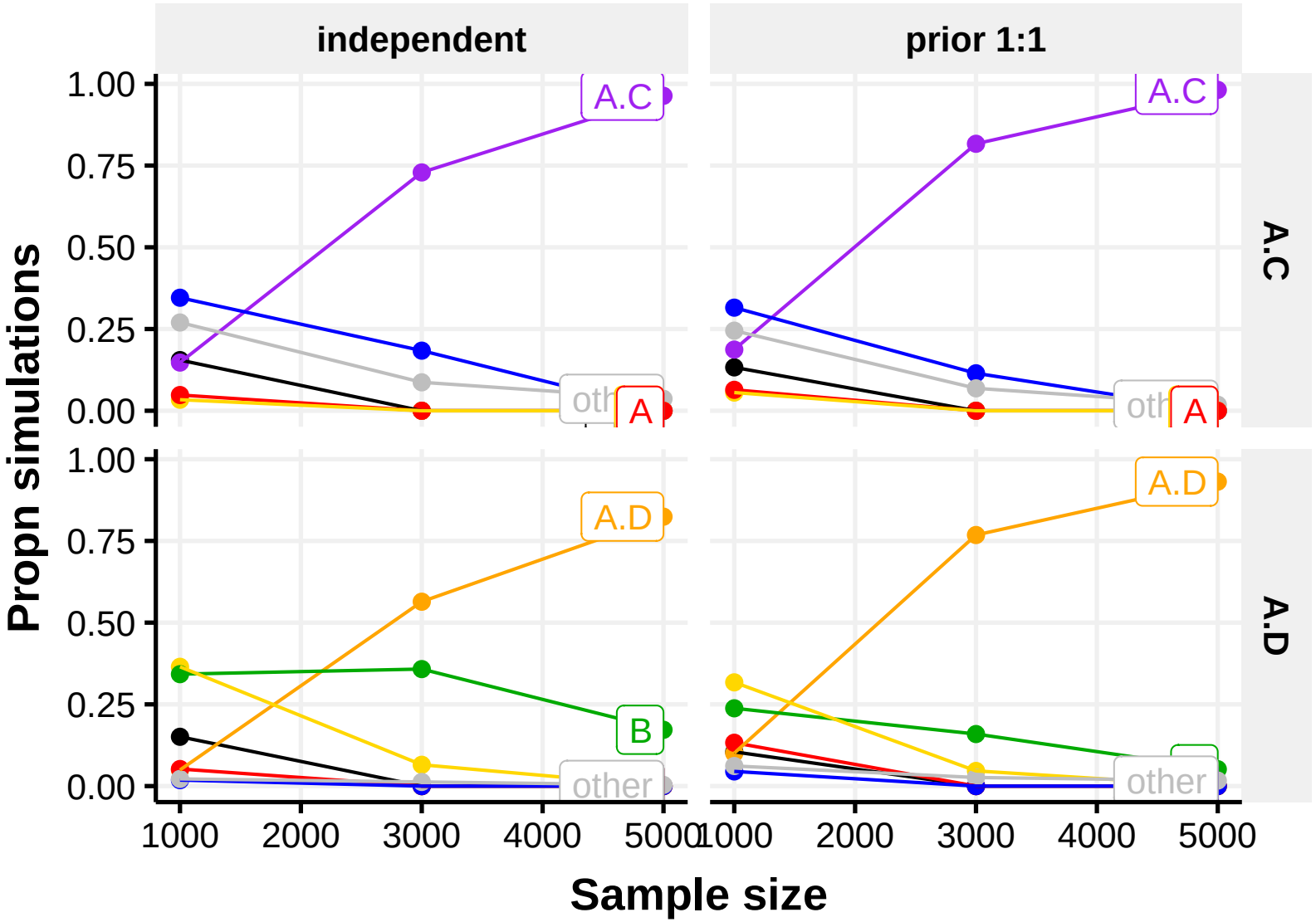
- Running time: 15 seconds (2 diseases) — 83 seconds (6 diseases)

`https://github.com/jennasimit/MTFM`

Joint fine mapping improves accuracy at smaller sample sizes



Joint fine mapping improves accuracy at smaller sample sizes



Joint fine mapping of 30 regions

7/30 regions differed between single and multi-disease analysis

4/4 cases: MFM results matched single disease analysis in larger international dataset

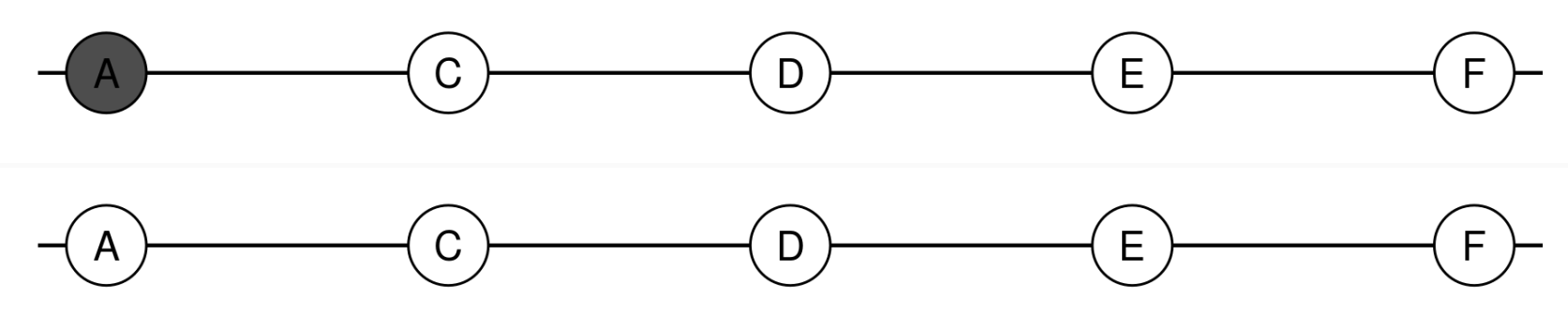
Region	Disease	Stochastic	MFM	Mean r^2
1p	RA	D/rs4648662	C/rs10752749	0.36
	iRA	C/rs141426426	C/rs10797431	
6q	RA	G/rs56258221	C/rs72928038	0.33
BACH2	iRA	C/rs72928038	C/rs72928038	
18p	CEL	F/rs34799913	C/rs12967678	0.4
PTPN2	iCEL	C/rs67878610	C/rs12967678	
10p	MS	B/rs2104286	A/rs12722496 + D/rs7089861	0.2, 0.3
IL2RA	iMS	A/rs12722496 + D/rs56382813	A/rs12722496 + D/rs7089861	

Functional validation of causal effects on *IL2RA*

Allele specific expression

Allele specific expression: quantify relative expression of two chromosomes using targetted PCR and sequencing

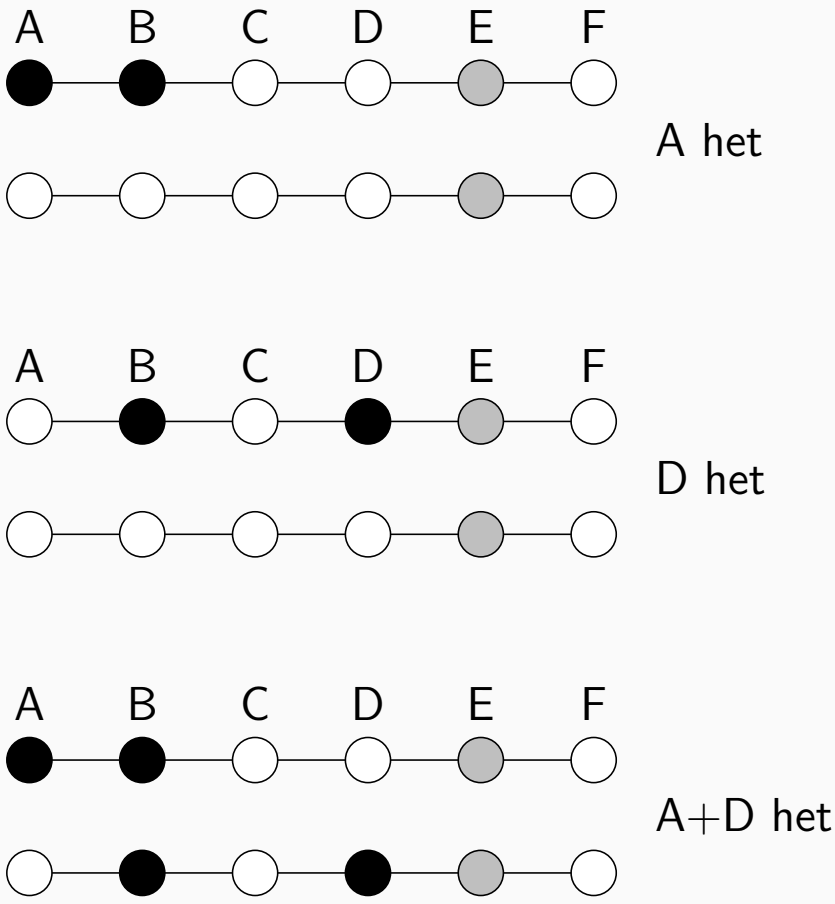
Within-individual: controls for between individual variation in environment, other genetics etc



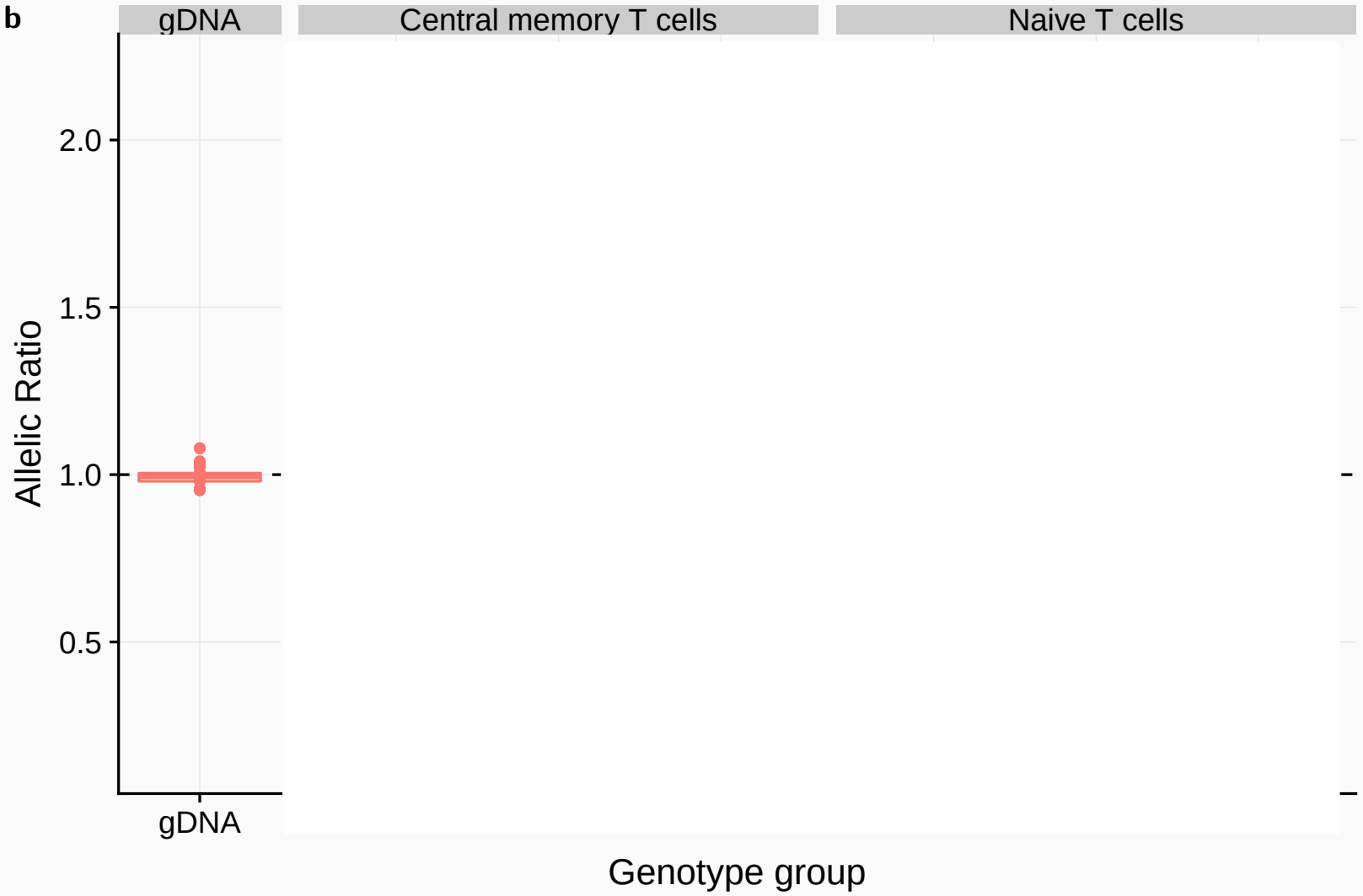
Cambridge NIHR BioResource

Effects of A, D and B

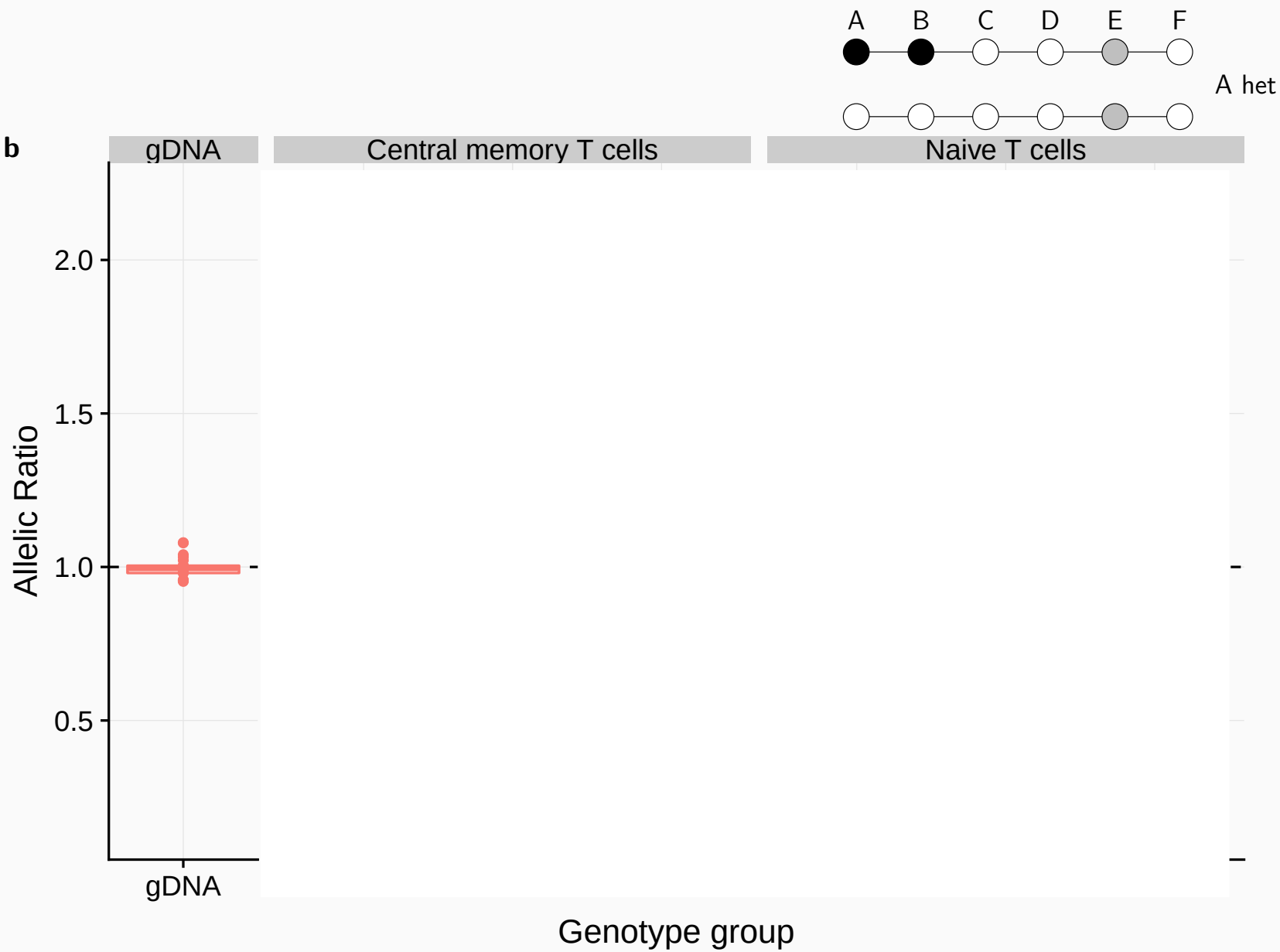
a



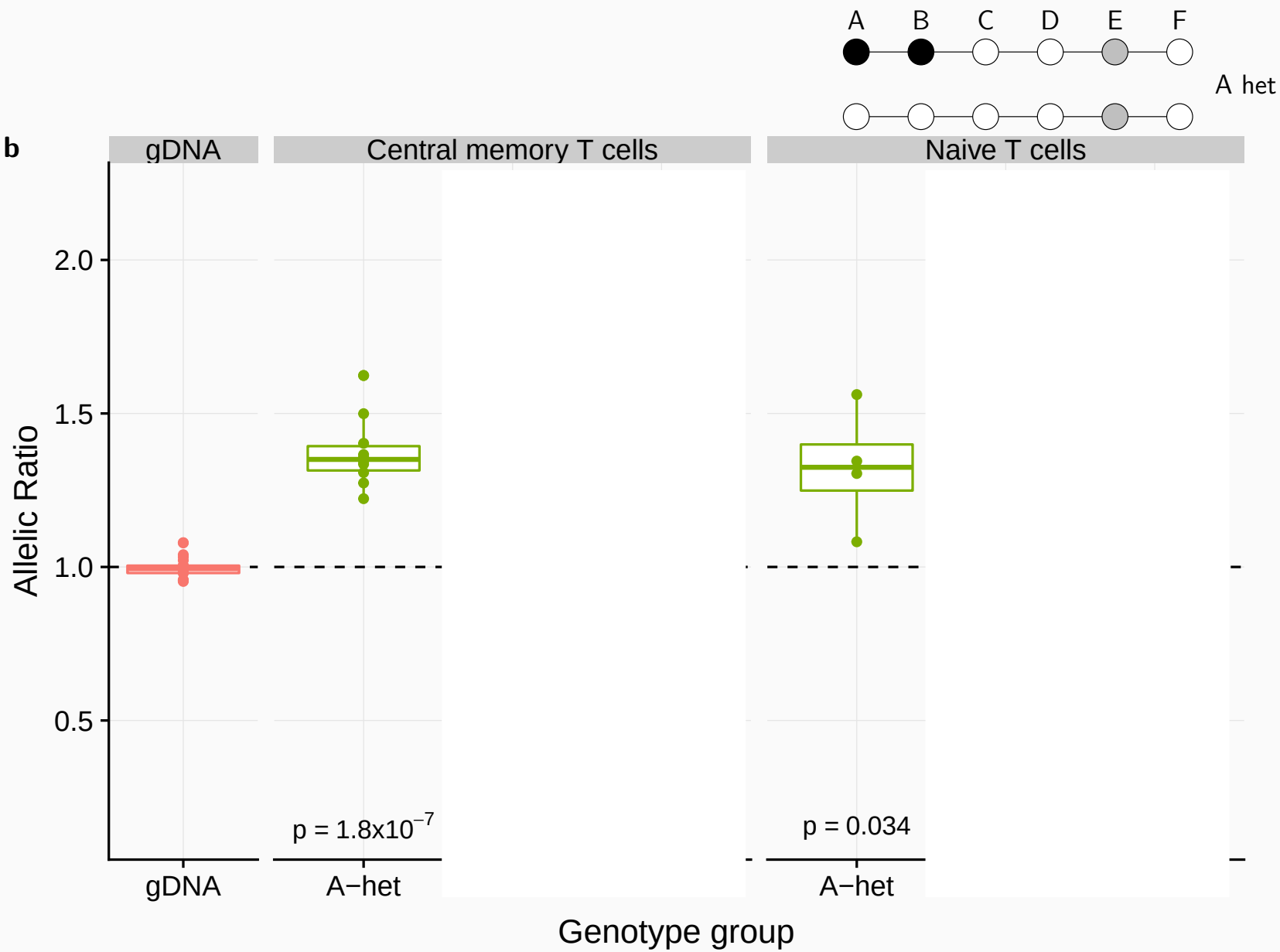
Effects of A, D and B



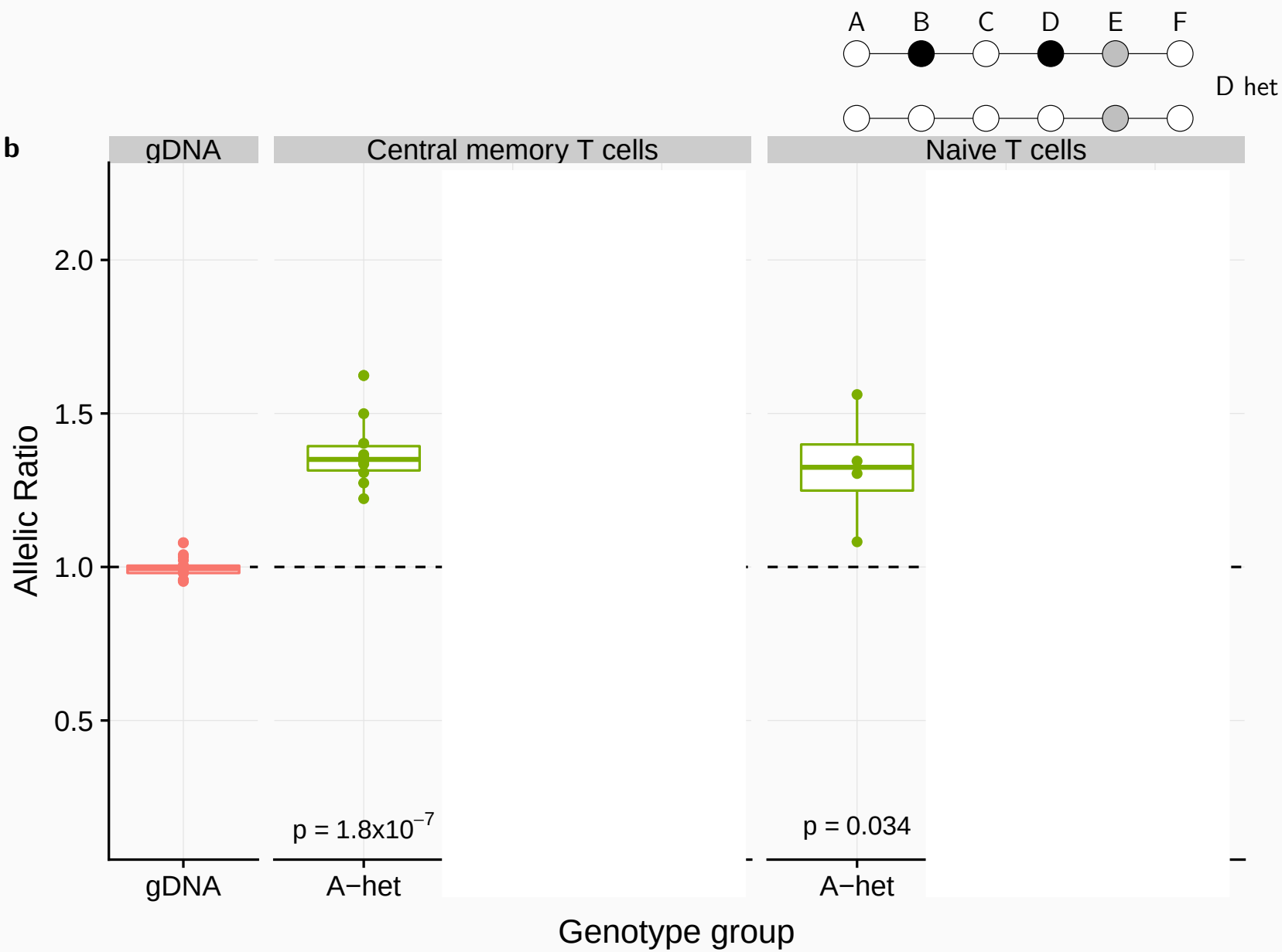
Effects of A, D and B



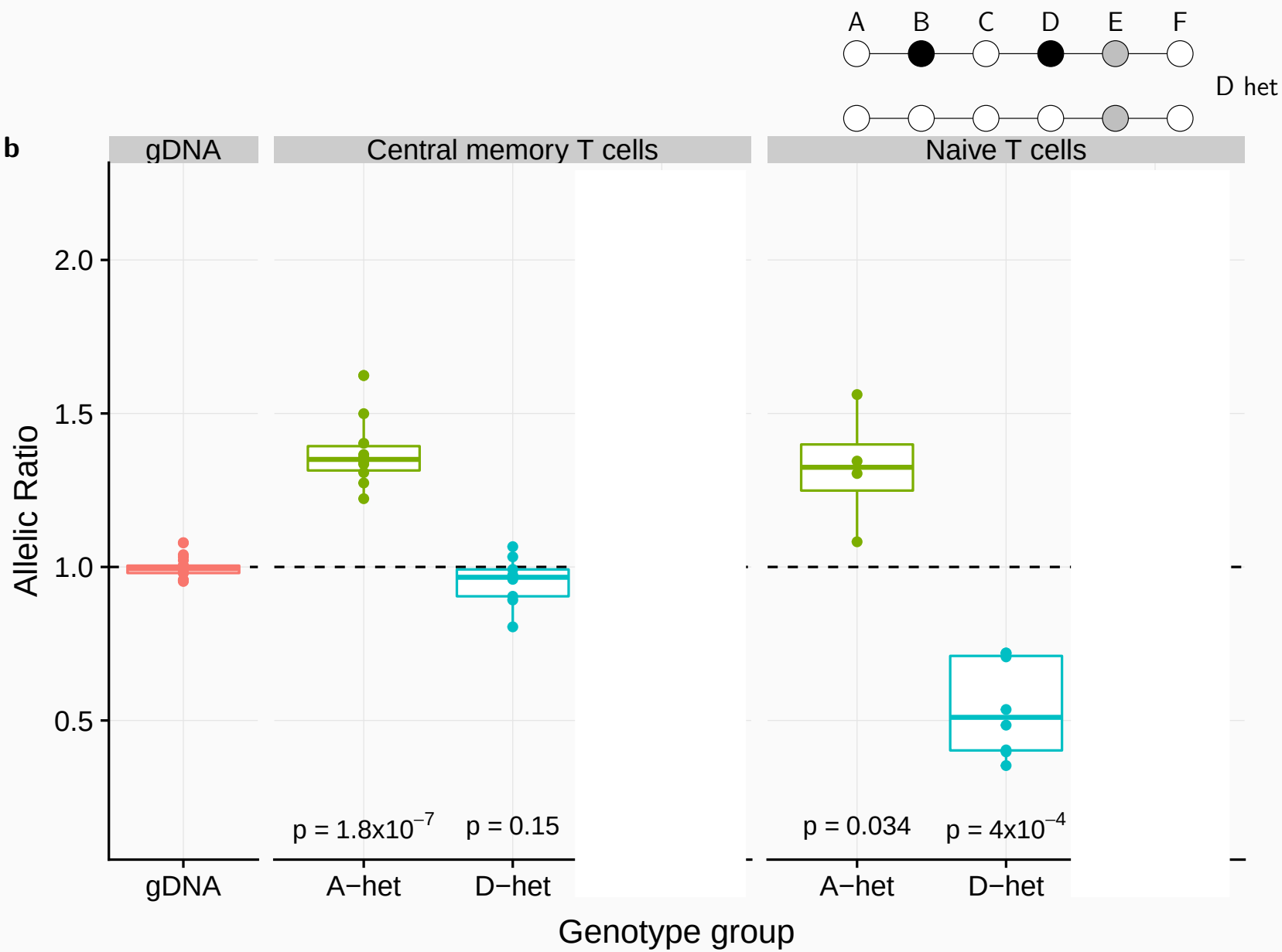
Effects of A, D and B



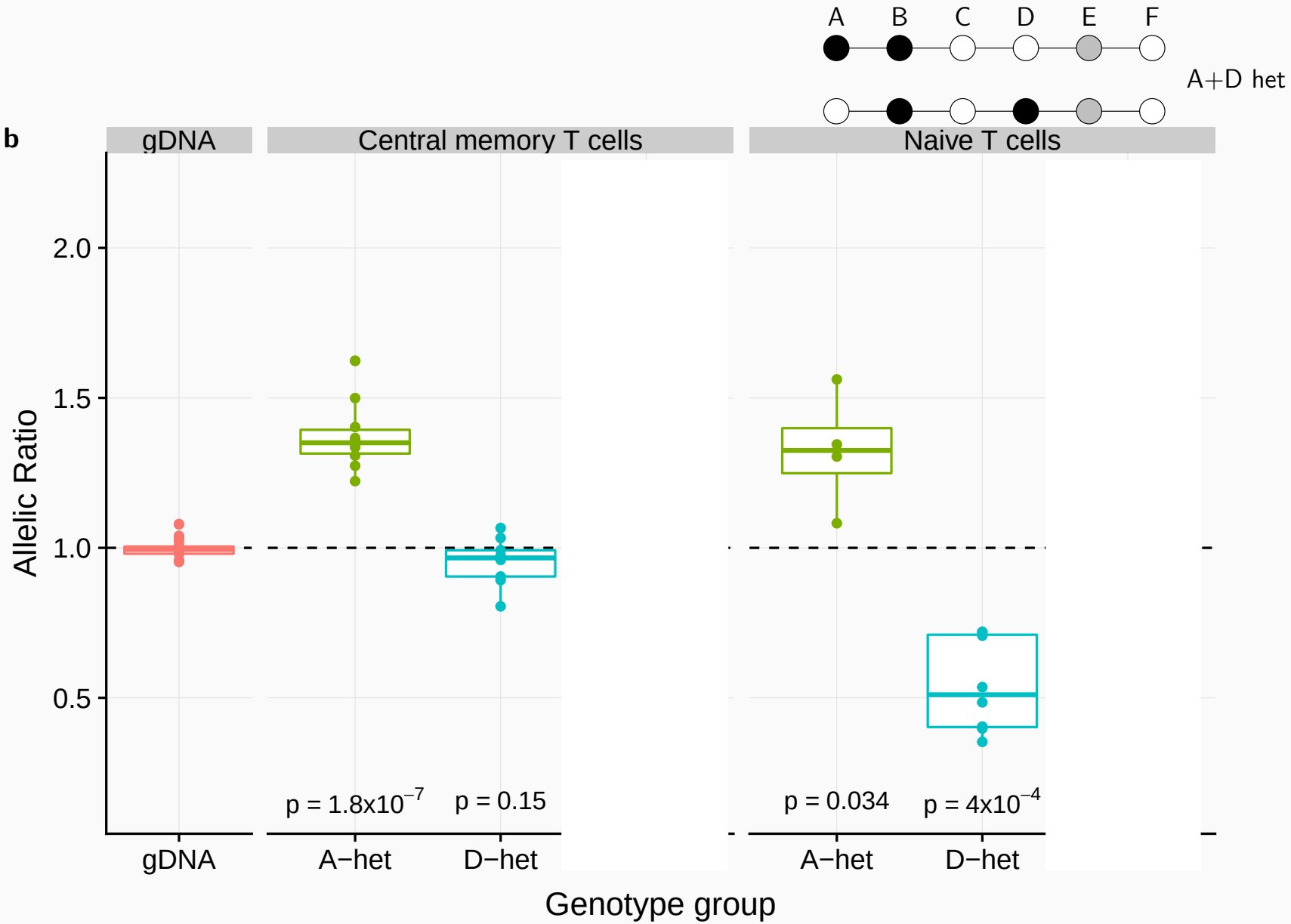
Effects of A, D and B



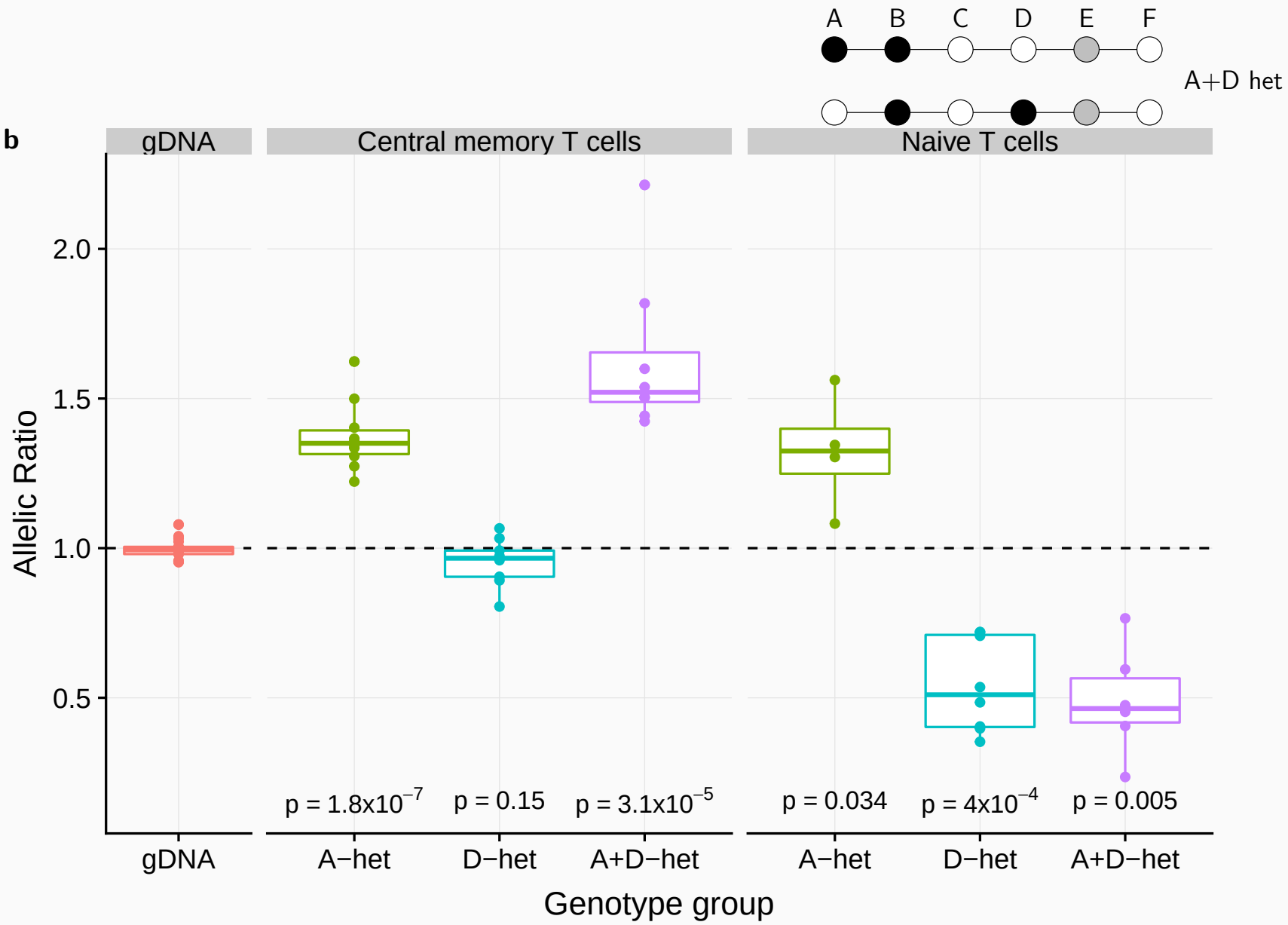
Effects of A, D and B



Effects of A, D and B



Effects of A, D and B



Summary

- Finding associated variants by GWAS variants is “easy”
- Fine mapping is harder
- Knowing causal variants, mechanism can be explored, or explicitly tested (ASE, Hi-C, eQTL, chipseq, CRISPR)

Joint tagging is a thing

- Most regions likely to contain > 1 causal variants
- Joint tagging might affect $\sim 20\%$ of causal variant pairs
- Key assumptions in stepwise search are not about number of / LD between causal variants, but about whether another SNP exists which acts as a lower dimensional summary of disease effect

Reasons for caution

- *IL2RA* “famous”: multiple, complex associations
- Other regions of greatest a-priori interest show strongest associations, learning they are also complex (e.g. *IL2*, *CTLA4*)

What if all regions are “complex”?

Reasons for optimism

- Borrowing information between related diseases can help overcome sample size limitations
- Inference on shared/distinct genetic causality comes “for free”

Thanks to

Jenn Asimit

Mary Fortune

Dan Rainbow

Linda Wicker

Disease investigators Steve Eyre (RA), Steve Rich, John Todd (T1D), Stephen Sawcer, IMSGC (MS), Wendy Thomson (JIA), David van Heel (celiac disease), Stephen Gough (ATD)

Cambridge NIHR BioResource

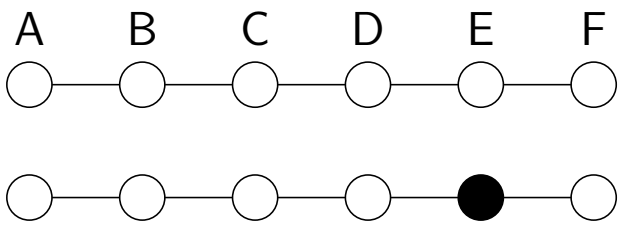
`chr1swallace/GUESSFM` `jennasimit/MFM`

wellcome

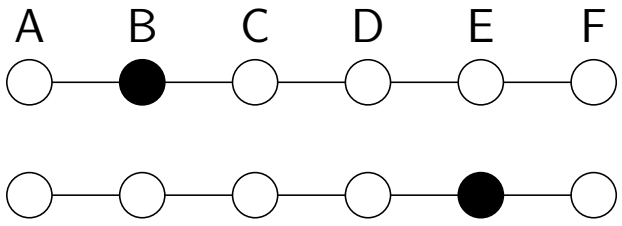
UNIVERSITY OF CAMBRIDGE

MRC Biostatistics Unit

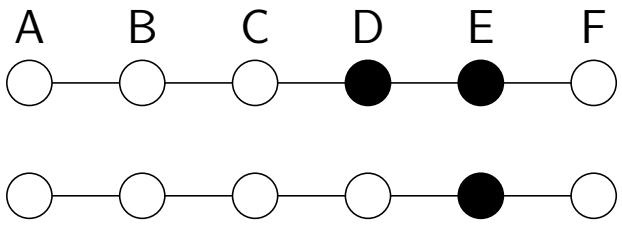
Individuals with rare recombination events



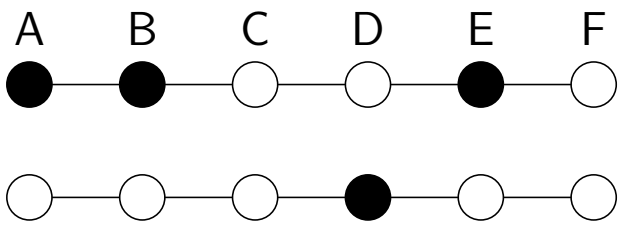
Donor 1



Donor 2



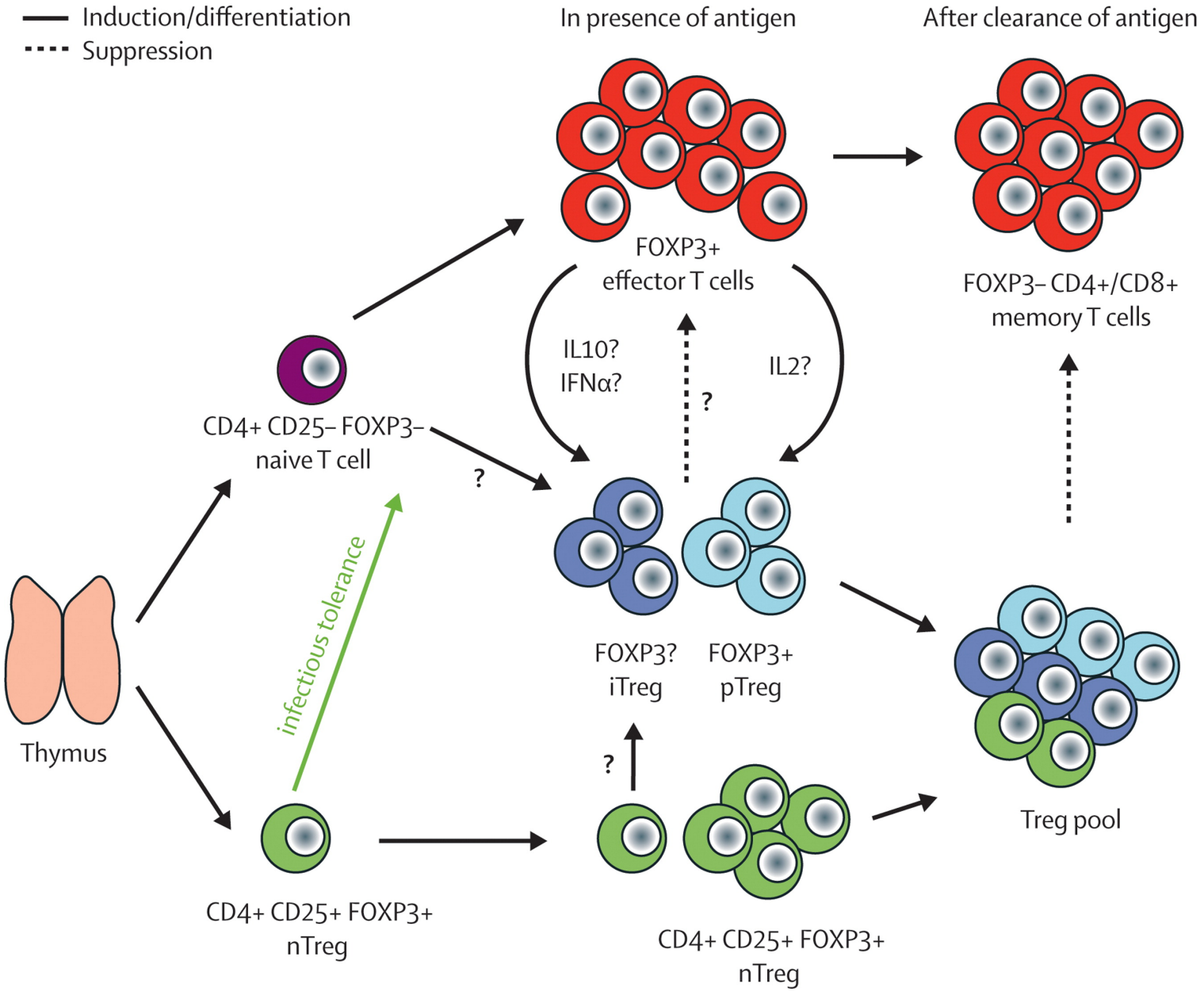
Donor 3



Donor 4

Individuals with rare recombination events

T cell subsets in immune-mediated diseases



T cell subsets in immune-mediated diseases

Disease	Model
T1D	A + C + E + F
MS	A + D
ATD	A + other
Celiac	A+C? / A / C / other