Multipoint linkage disequilibrium mapping using multilocus allele frequency data

Toby Johnson

`toby.johnson@ed.ac.uk`

Rothamsted Research & University of Edinburgh

Motivation

- Many diseases have a heritable component; **mapping the underlying gene(s)** has many potential benefits
- **Linkage disequilibrium (LD) mapping** (a.k.a. association mapping) has potential to achieve greater resolution than pedigree studies (more meioses in population history than in a pedigree)
- **Large samples** (individuals × markers) are required when LD is weak, e.g. if there is
  - Ancient origin of disease allele
  - Complex genetic basis underlying the disease
  - Phenocopies (individuals with disease status but without the disease allele)
- A technology called **DNA pools** allows cheap genotyping of many individuals
  - There is at least one “Pooled Genome Scan” dataset of approximately 16,000 bi-allelic markers where phenotypes are complex disease thought to have polygenic basis – **a potentially very informative dataset**

Linkage disequilibrium mapping

- **Looks for association between disease status and allelic state at marker locus or loci**
- Example (Muir *et al.* 2001)

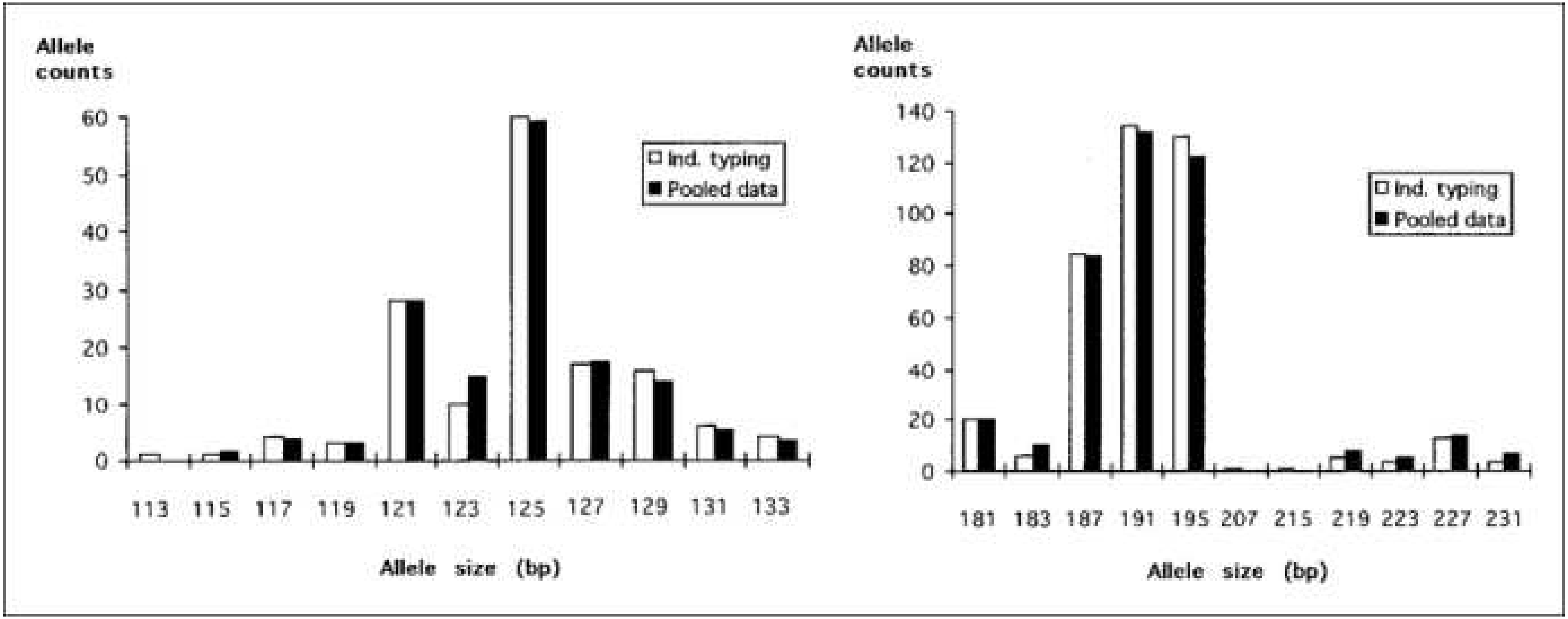
DRD5 microsatellite	Control		Schizophrenia	
allele	count	frequency	count	frequency
134	15	1.72	4	1.27
136	22	2.51	3	0.95
138	78	8.92	29	9.18
140	39	4.46	6	1.90
142	31	3.55	12	3.80
144	35	4.00	18	5.70
146	67	7.67	12	3.80
148	384	43.9	169	53.5
150	110	12.9	32	10.1
152	64	7.32	21	6.65
154	22	2.52	10	3.16
156	7	0.80	0	0.00

- Assume **ancient polymorphism in marker** DRD5 microsatellite
- Assume schizophrenia **predisposing allele arose on unique genetic background**
  (there was complete LD at some time in the past)
- Interpret weak association because of either weak effect, or recombination, or both

DNA Pools

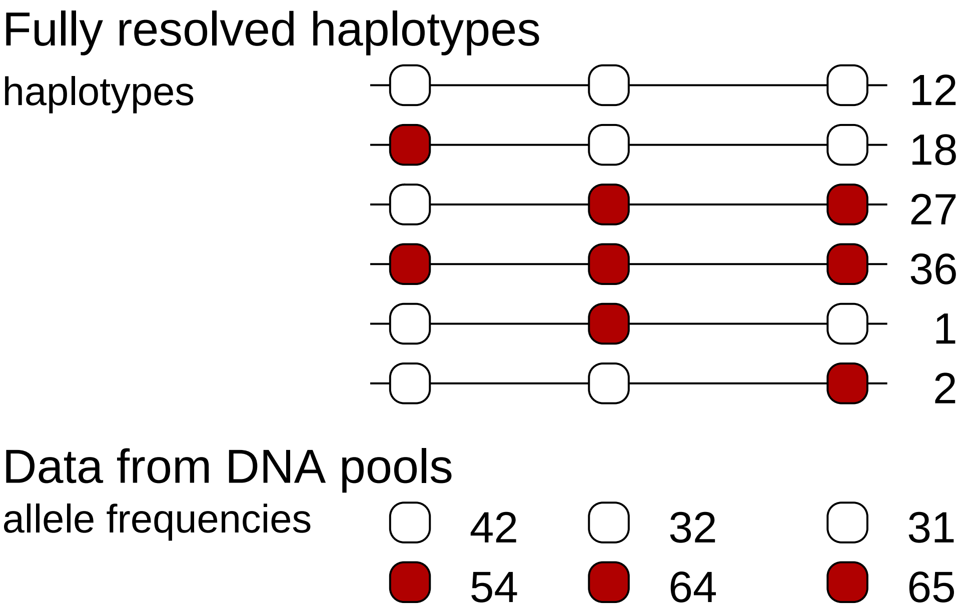
- A pool consisting of exactly **equal quantities of DNA** from many individuals **is mixed together and then typed**. The ratio of peak heights on the chromatograph inform us about the frequencies of the alleles present in the pool
- **Advantage** Effort saved can be used to type more individuals and/or markers
- **Disadvantages**
  - Peak height estimation and differential amplification of alleles lead to imprecise estimates of allele frequencies (but this is a small problem)
  - **No phase / linkage information acquired**
  - **No multipoint analysis available**
    - Multipoint analysis uses data from several markers **simultaneously** to weigh the evidence for the disease locus being at a given position
    - Several multipoint methods are available for analysing haplotypes or phase-unknown diploid genotypes (e.g. `DMLE+` Reeve & Rannala 2002, `BLADE` Liu *et al.* 2001, `COLDMAP` Morris *et al.* 2002, 2003, 2004)

Accurate estimation of allele frequencies



From Barcellos *et al.* 1997 AJHG **61**:734

DNA pools throw away phase information



Fully resolved haplotypes

haplotypes

○	○	○	12
●	○	○	18
○	●	●	27
●	●	●	36
○	●	○	1
○	○	●	2

Data from DNA pools

allele frequencies

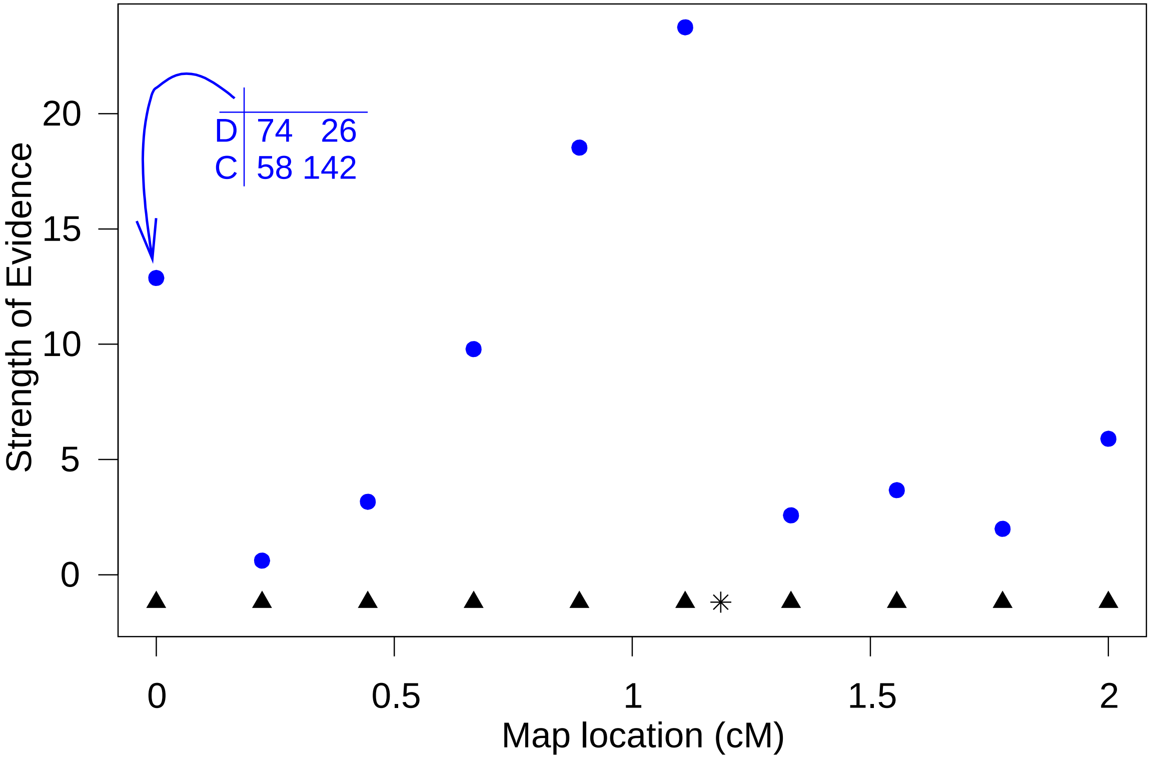
○	42	○	32	○	31
●	54	●	64	●	65

- No information from DNA pools about strong LD between second and third markers
- At present marker loci must be analysed one by one
- Obviously pooling only considered within diseased and control groups

Single point methods are inadequate

- p-values **confound effect size** (strength of LD) **and power** (heterozygosity of marker; number of alleles), leading to “incoherent” conclusions
- Decision to attempt positional cloning should be based on a quantified **region estimate**
- Failure to use all the information leads to **inefficiently large region estimates**

▲ = marker, $*$ = disease locus, ● = $-\log_{10}(p\text{-value})$

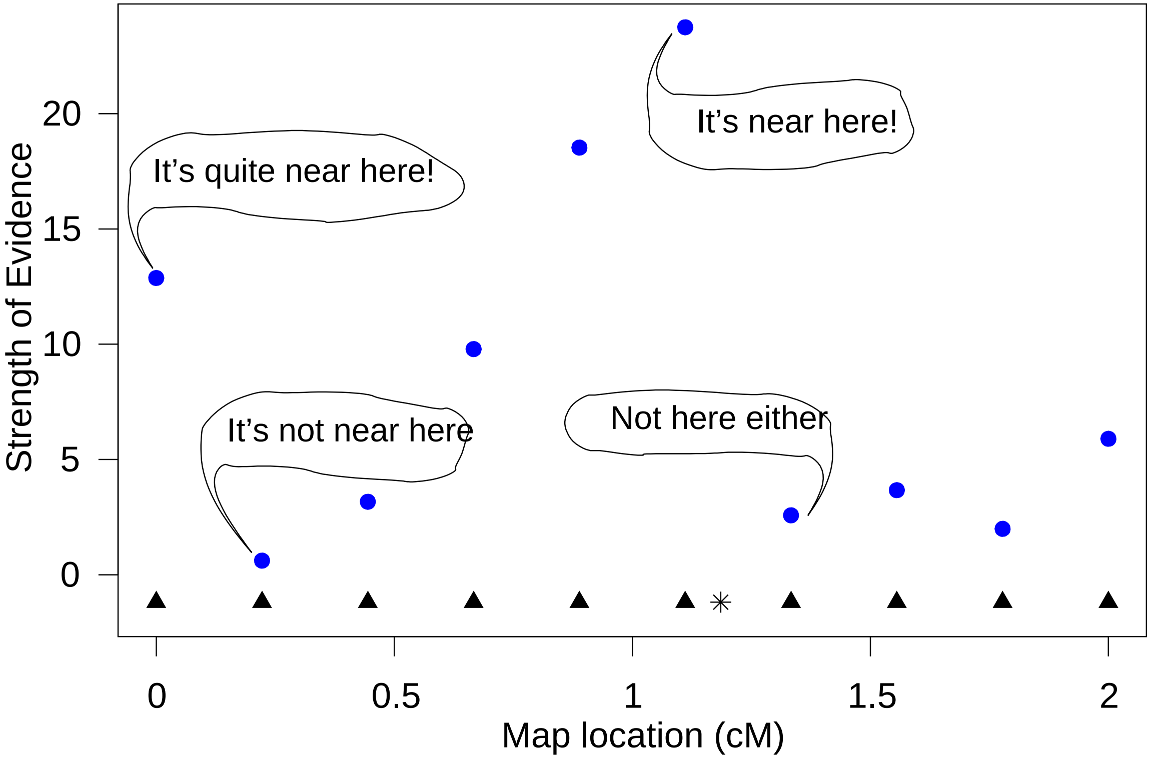


D	74	26
C	58	142

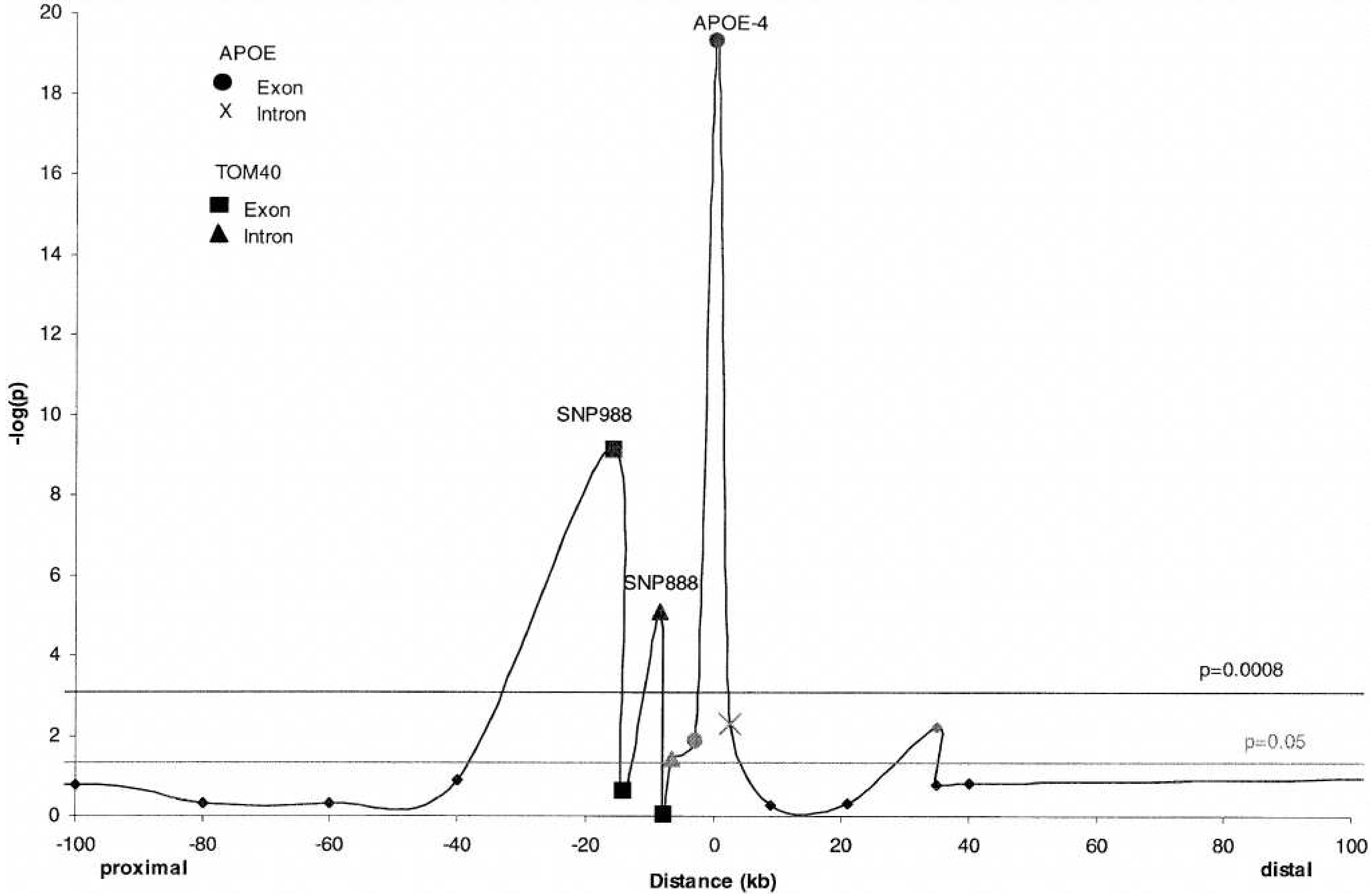
Single point methods are inadequate

- p-values **confound effect size** (strength of LD) **and power** (heterozygosity of marker; number of alleles), leading to “incoherent” conclusions
- Decision to attempt positional cloning should be based on a quantified **region estimate**
- Failure to use all the information leads to **inefficiently large region estimates**

▲ = marker, $*$ = disease locus, ● = $-\log_{10}(p\text{−value})$

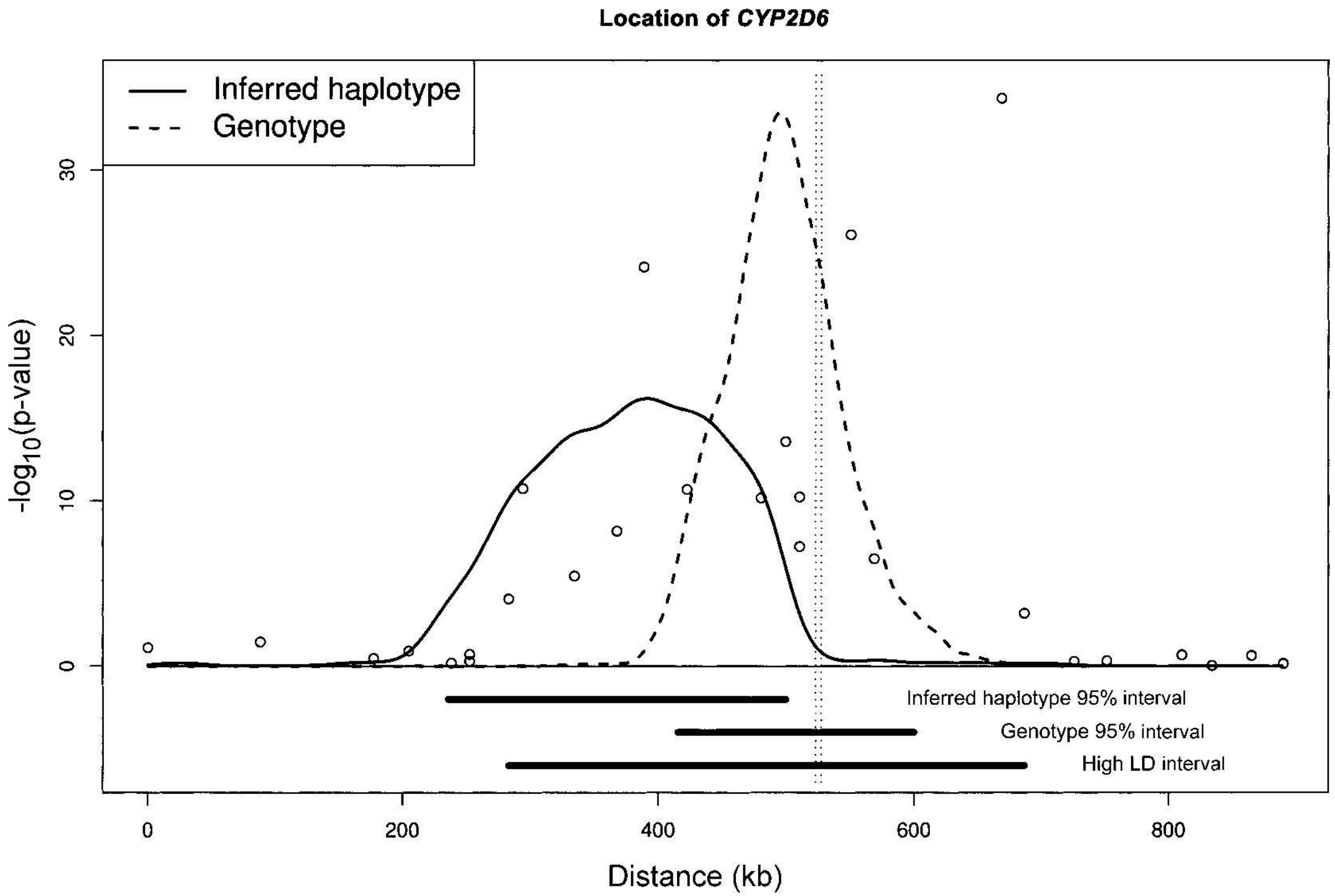


Real Data: APOE and Alzheimers



From Martin *et al.* 2000 AJHG **67**:383

Real Data: Cytochrome p450 Enzyme

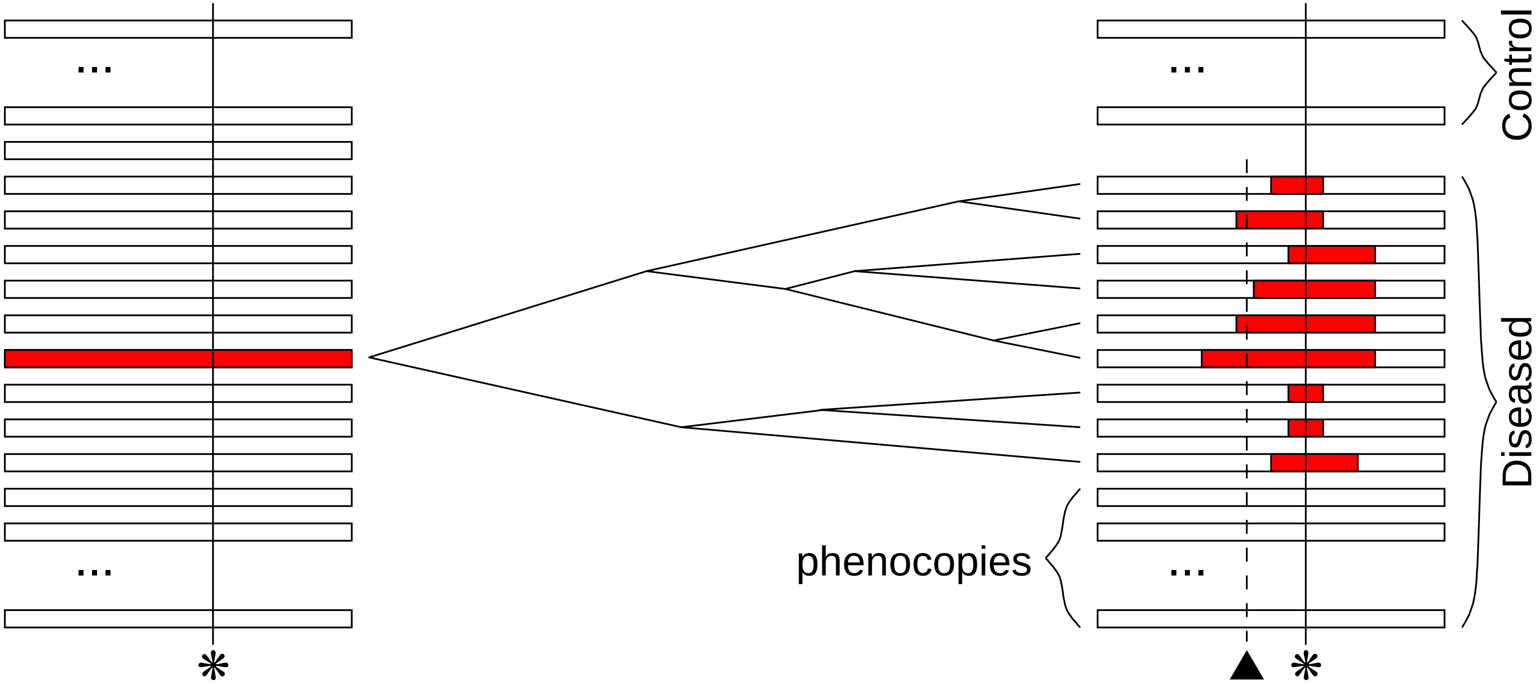


From Morris *et al.* 2004 AJHG **74**:945

A new mapping method: “Poolmap”

- Advantages
  - Uses multilocus allele frequency data, not haplotypes
  - Non-parametric model for genealogy at disease locus
  - No assumption about map distances between markers
  - Robust to (unknown) rate of phenocopies, and to dominance at disease locus
  - Computationally rapid
  - Calculates profile likelihood comparable to posterior density
- Disadvantage: Less precise inferences because
  - Less information (used) from data
  - Non-parametric model
  - Conservative elimination of nuisance parameters

Poolmap method uses a nonparametric model



Disease allele (✻) arises on a unique ancestral chromosome

Arbitrary genealogy with recombination (disease allele stays rare; no two point crossovers)

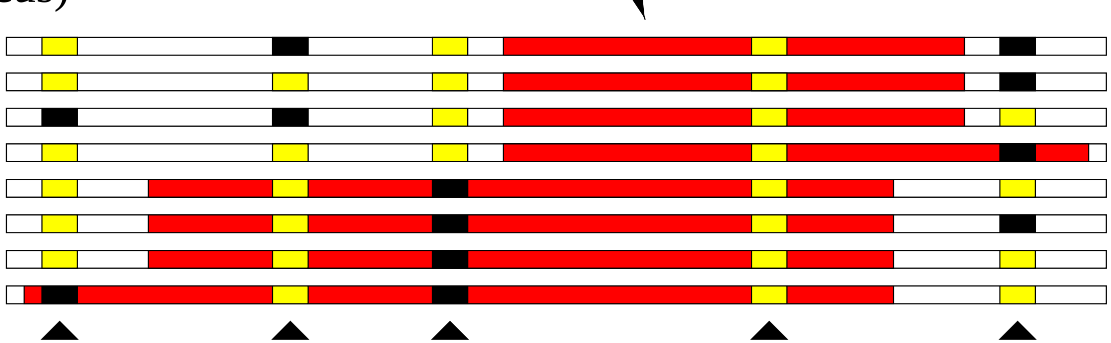
Disease allele surrounded by single blocks of ancestral chromosome

Data, model and parameters ⇝ likelihood function

data from control pool :	🟨	$c_{11}=6$	4	7	5	4
	⬛	$c_{12}=2$	4	1	3	4
data from diseased pool :	🟨	$d_{11}=6$	6	4	8	4
	⬛	$d_{12}=2$	2	4	0	4

parameter of interest :
(position of disease locus)

$z =$



missing data : $a_1 =$ ⬛ etc. for a_2 a_3 ...

nuisance parameters : $x_1 = 1$ $x_2 = 4$ $x_3 = 4$ $x_4 = 8$ $x_5 = 1$

$p_1($⬛$) = 0.2$

$p_1($🟨$) = 0.8$

etc. for p_2 p_3 ...

The likelihood function

$$\mathrm{L}(z, \boldsymbol{x}, P; D, C) \propto \prod_{i=1,\ldots,n} \left(\sum_{a_i} \left(p_{ia_i} \; \mathrm{I}(d^*_{ij} \geq 0) \; \times \; \frac{(\sum_j d^*_{ij})!}{\prod_j d^*_{ij}!} \prod_j p_{ij}^{(d^*_{ij}+c_{ij})} \right) \right)$$

where $d^*_{ij} = d_{ij} - \delta_{ja_i} \, x_i$ are the counts that is not “explained” by $\boldsymbol{a}$ and $\boldsymbol{x}$

and $\mathrm{I}(d^*_{ij} \geq 0) \in \{0, 1\}$ is an indicator function

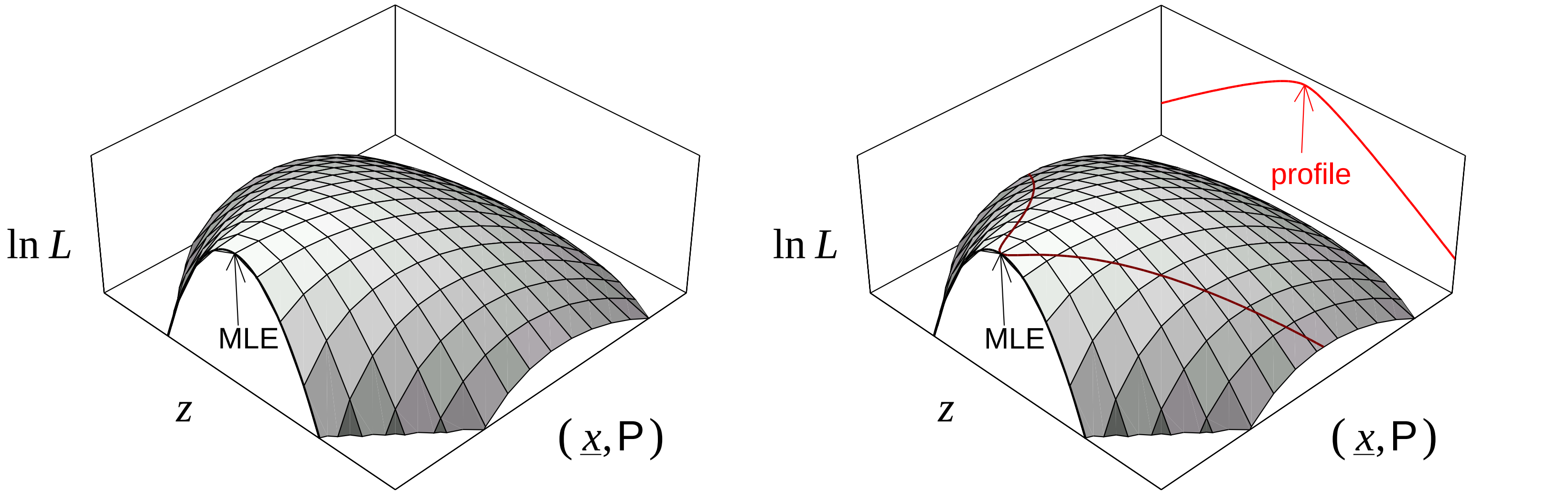
- Applies for arbitrary numbers of alleles at each locus
- Awkward to work with, but efficient numerical exploration possible by using the Pool Adjacent Violators Algorithm (PAVA; Brunk 1955)

Crucial Assumptions Made

- **Rare disease predisposing allele**
- **Linkage Equilibrium and Hardy–Weinberg proportions in blocks of non-ancestral chromosome**

Profile likelihood for reducing dimensionality

$$\mathrm{L}_{\max}(z; D, C) = \max_{\boldsymbol{x}, P} \mathrm{L}\,(z, \boldsymbol{x}, P; D, C)$$



Profile likelihoods “behave” like ordinary likelihoods in several respects:

- **Maximum at same value** of z
- **Equivalence between support regions**

$$\Theta(c) = \{(z, \boldsymbol{x}, P) : \mathrm{L}\,(z, \boldsymbol{x}, P; D, C) > c\} \text{ is a level } c \text{ support region}$$
$$\mathcal{Z}(c) = \{z : \mathrm{L}_{\max}\,(z; D, C) > c\} \text{ is a level } c \text{ profile support region}$$

$$z \in \mathcal{Z}(c) \iff \exists\,(\boldsymbol{x}, P) \text{ s.t. } (z, \boldsymbol{x}, P) \in \Theta(c)$$

a value of z is “in one iff it’s in the other”

Why this might not work

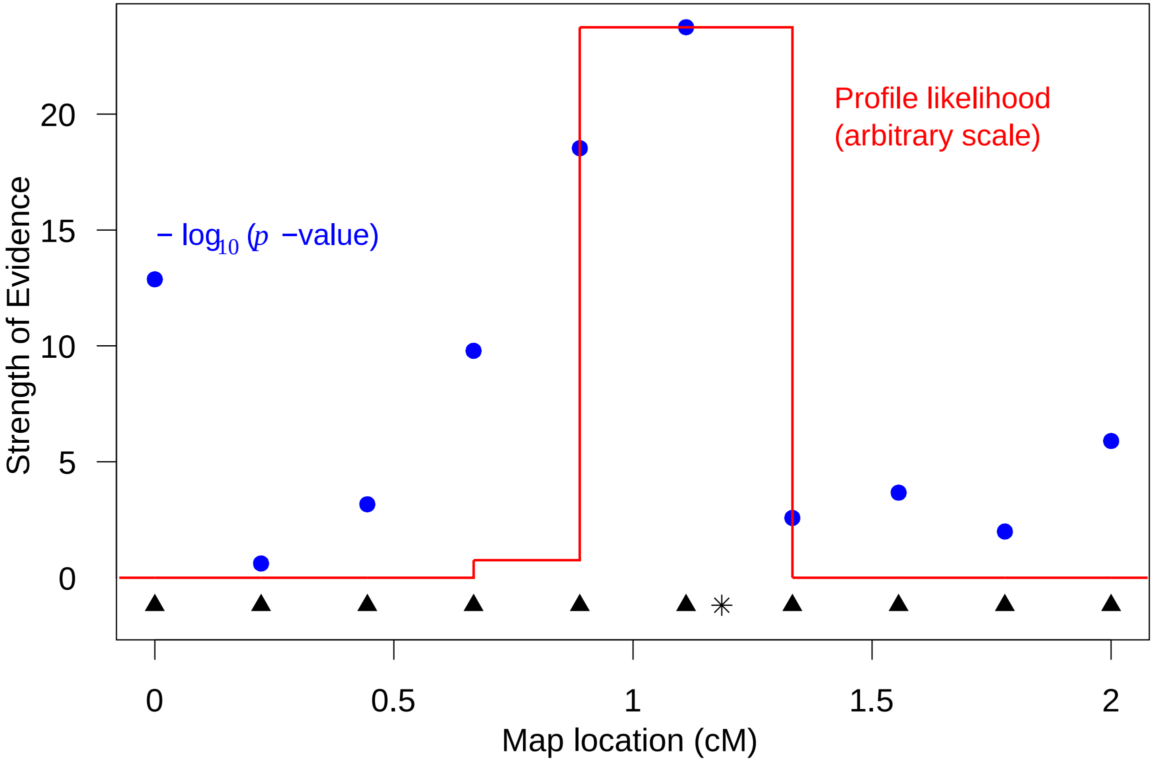
- I’ve deliberately abused the likelihood framework, choosing a “parameter” $\boldsymbol{x}$ so that the likelihood function has a simple form. Ordinarily $\boldsymbol{x}$ would be a random variable with distribution indexed by age of disease allele and other parameters
- Whereas nuisance parameters can be eliminated by maximisation, nuisance random variables must be eliminated by integration
- Treating $\boldsymbol{x}$ as a parameter means that all (isotonic–antitonic) x are equally “plausible” a priori, but e.g. highly asymmetric $\boldsymbol{x}$ should be “less plausible”

R. A. Fisher (Design of Experiments, 1935) on the subject of nonparametric inference:

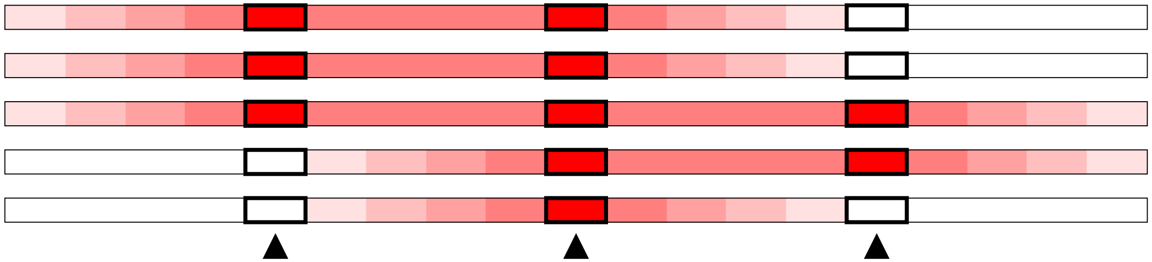
> *an erroneous assumption of ignorance is not innocuous* [*in inductive inference*]*; it often leads to manifest absurdities.* (with apologies to Sprott)

- Nonparametric model has higher dimensional parameter space than sample space $(z, \boldsymbol{x}, P) \in \mathbb{R}\ \mathbb{Z}^n\ \mathbb{R}^n$ and $(D, C) \in \mathbb{Z}^{2n}$ for biallelic loci
- Summary using profile likelihood is both contraversial (may lead to **misinference**) and conservative (may lead to **non-inference** or huge loss of information)

Poolmap generally produces “coherent” conclusions



Maximum precision of inference is 2 inter-marker intervals

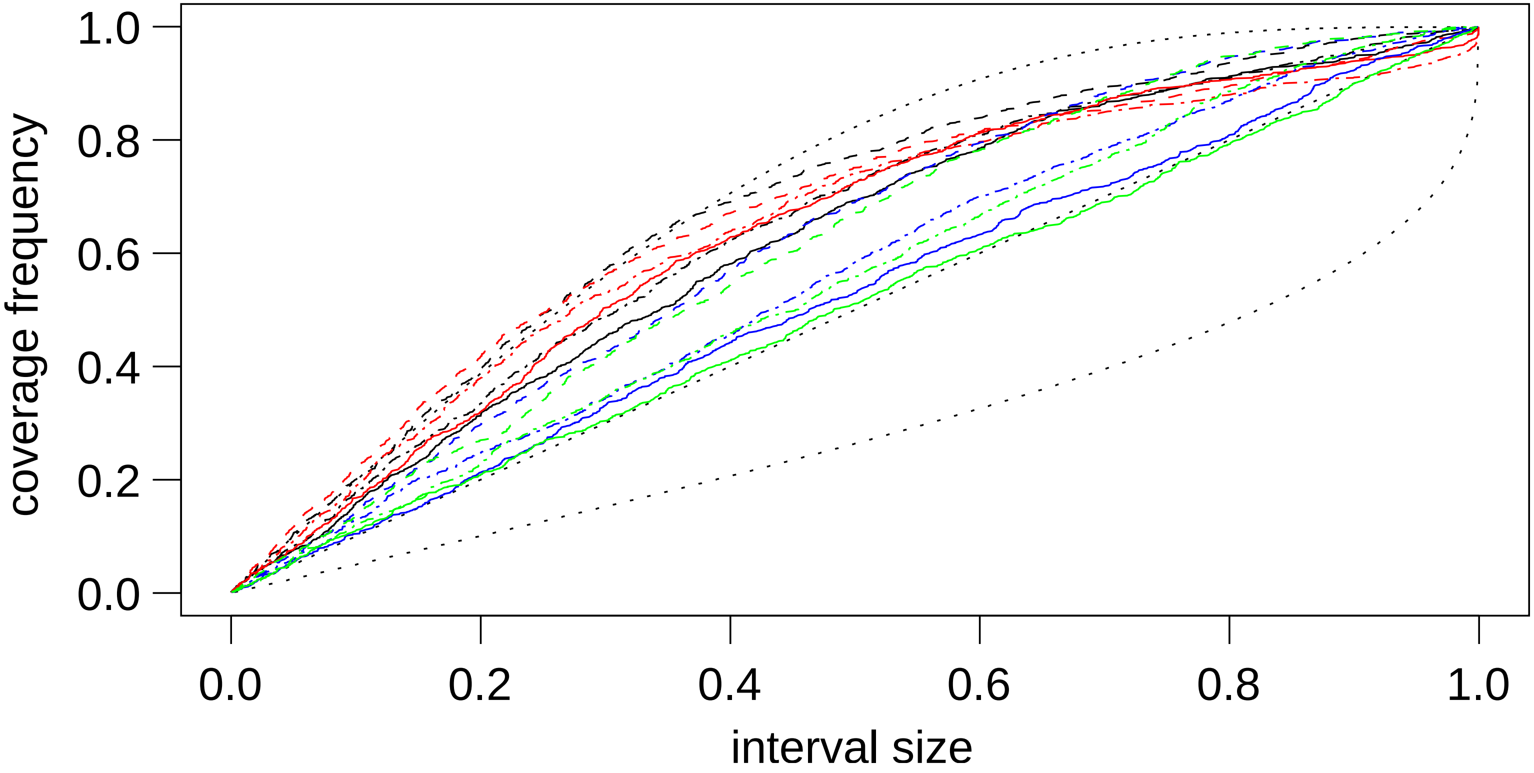


Test on simulated data sets

- Model assumed by DMLE+ Bayesian analysis program of Rannala and Reeve (2001,2002)
- 100 disease haplotypes and 200 control haplotypes
- $n = 10$, 28 or 82 markers at locations m_i uniform on $[0\text{cM}, 2\text{cM}]$ interval
- Marker loci biallelic with allele frequencies uniform on $[0.2, 0.8]$
- Position of disease locus, z^*, uniform on $[(m_1 + m_2)/2, (m_{n-1} + m_n)/2]$
- 1000 replicates for each combination of parameter values
  - **Y**oung: Allele age 100 generations, no phenocopies
  - **A**ncient: Allele age 1500 generations, no phenocopies
  - **P**henocopies: Allele age 100, 50% or 75% phenocopy chromosomes
    - Chromosomes in disease pool carry disease allele with probability 0.25
    - E.g. Disease allele at 0.5%, risk ratio $R_{Dd}/R_{dd} = 100$, 1% phenocopies in population
- Note only age × $(m_n - m_1)$ is identifiable

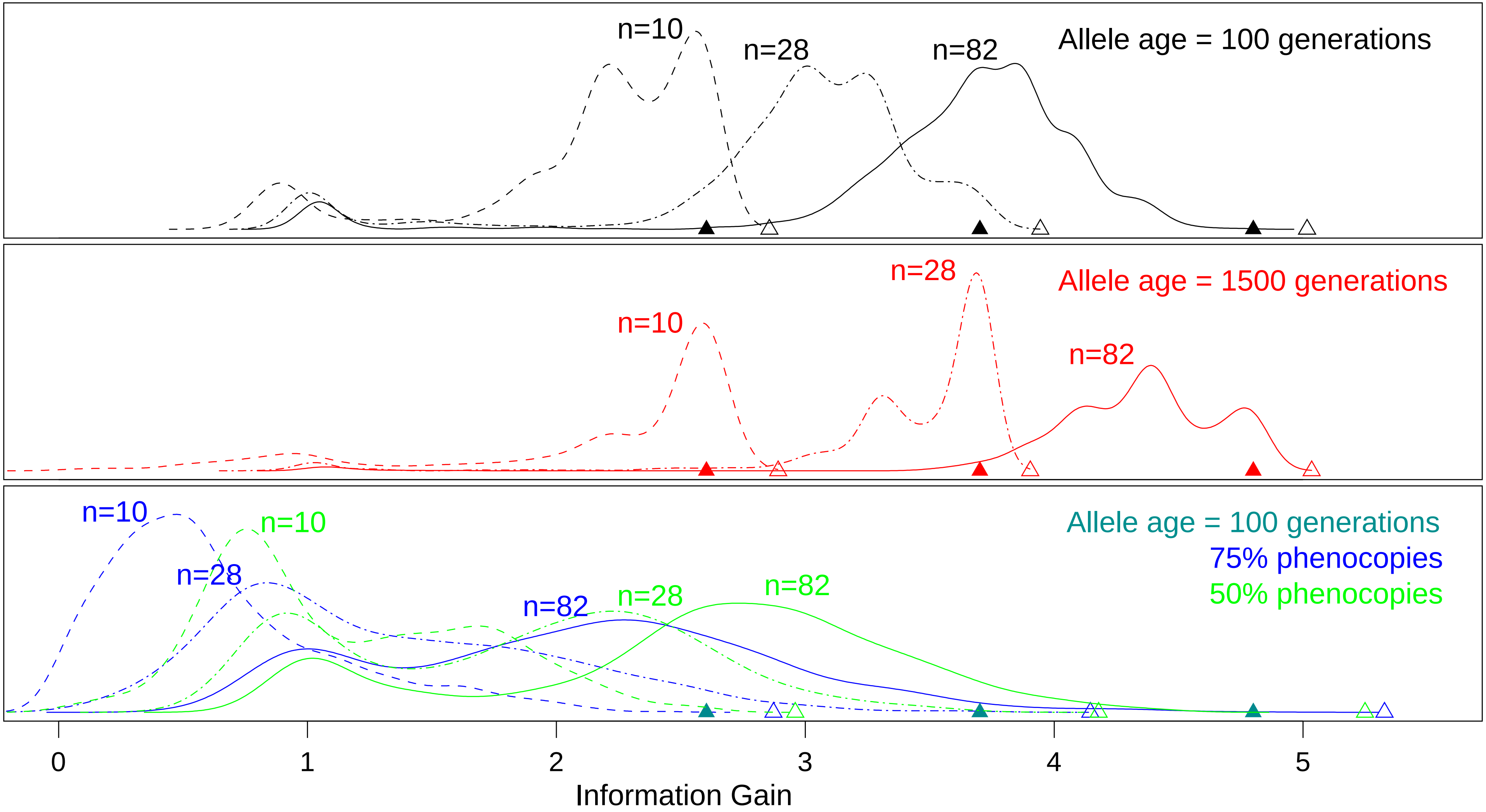
Does $\mathrm{L}_{\max}(\cdot)$ behave like an ordinary likelihood?

Yes, to the extent that confidence intervals based on a “Pretend Bayes” procedure (interpret normalized $\mathrm{L}_{\max}(z)$ as a density $\pi(z)$) have coverage properties (slightly) better than their size would suggest

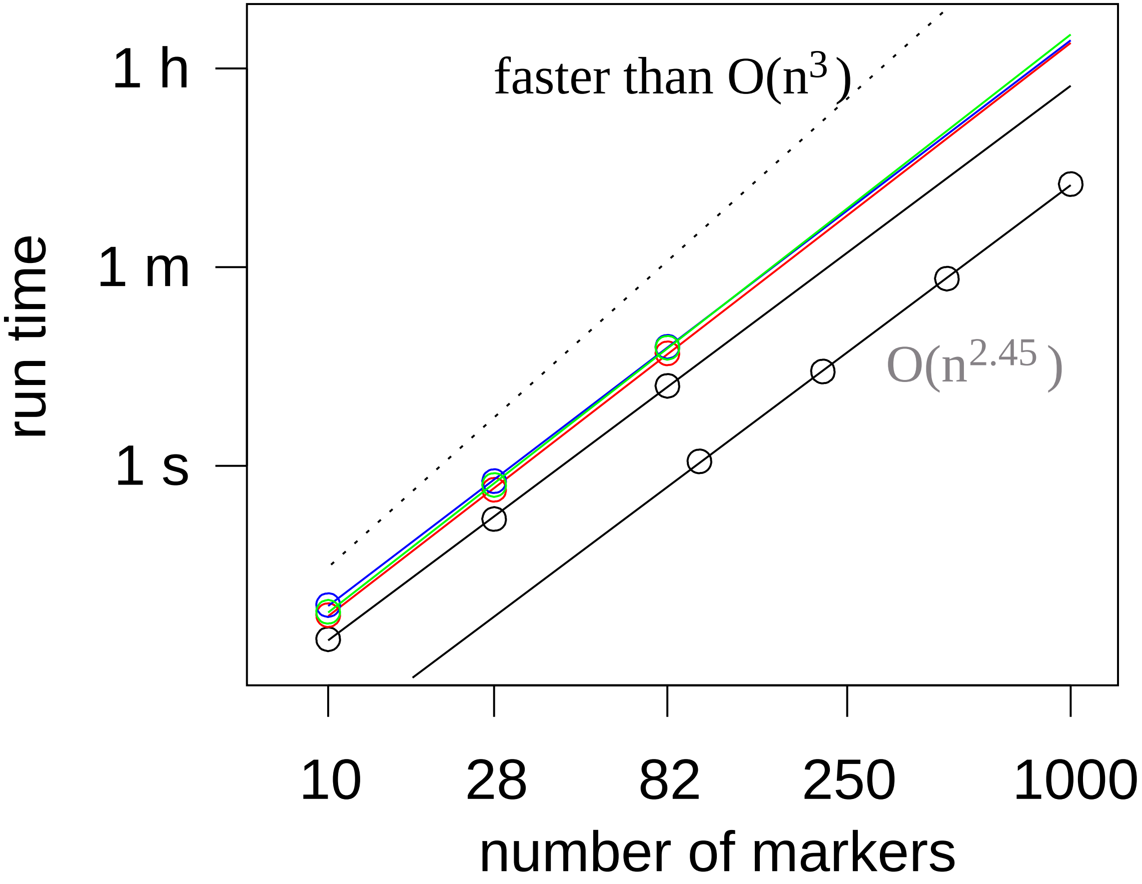


key: **Y**oung **A**ncient 50% **P**henocopies 75% **P**henocopies n =10/28/82 (dash/dotdash/solid)

Information Gain increases with marker density

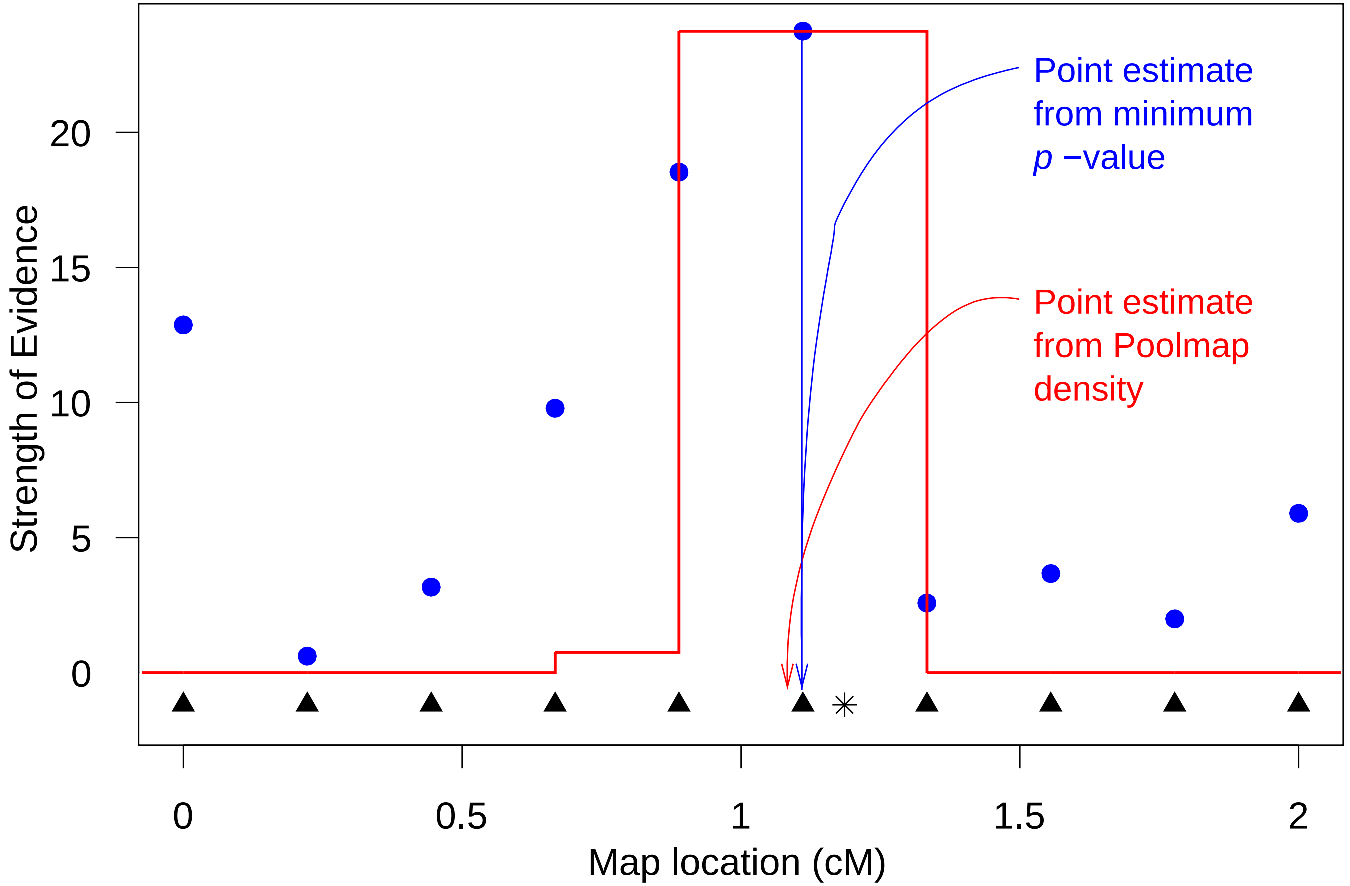


(Very) Fast Runtimes



- MCMC methods typically take **days** for $n \simeq 30$
- Composite likelihood methods are effectively $\mathrm{O}(n^2)$

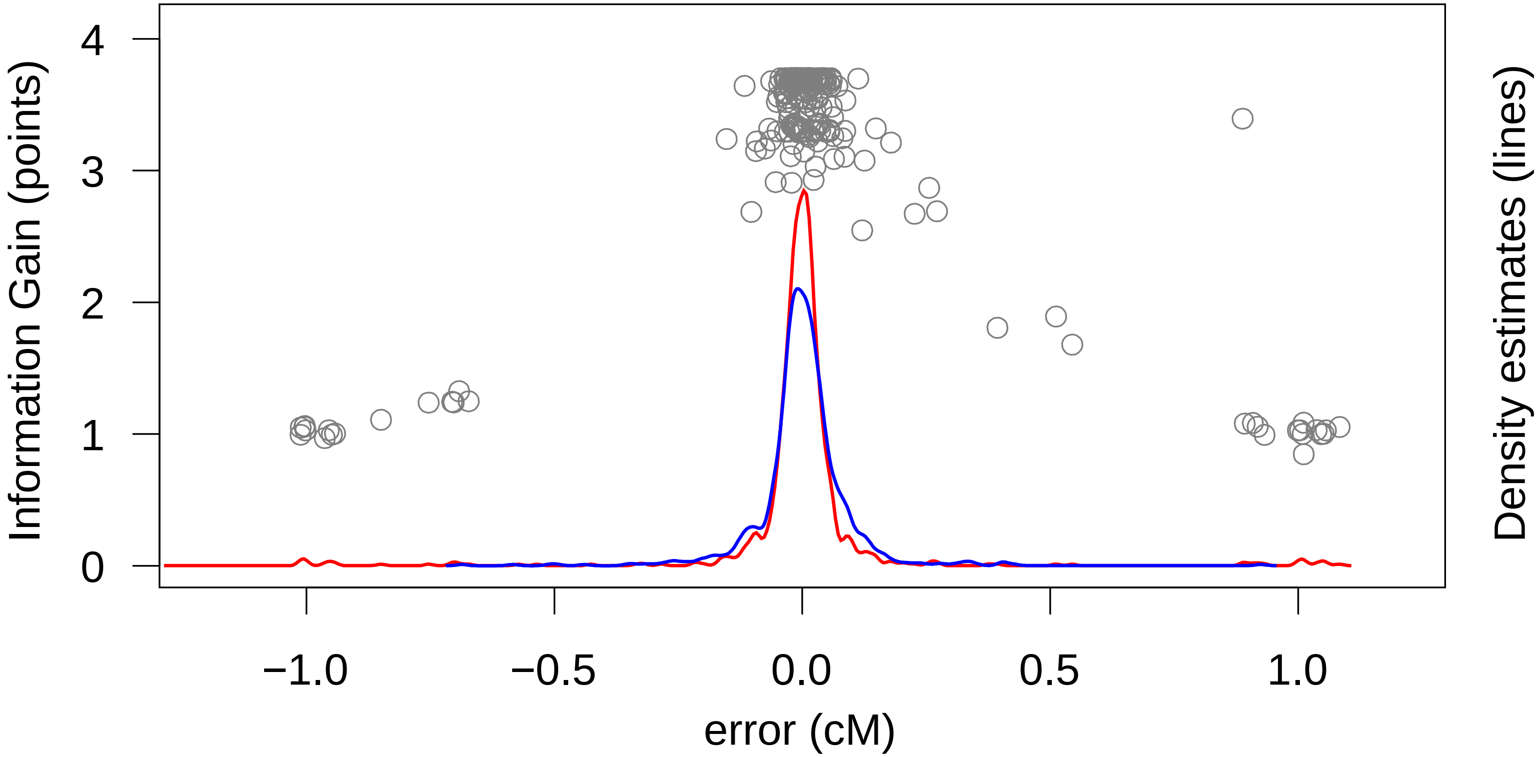
Comparison: Poolmap vs. Minimum p-value



Compare point estimates μ_z (mean of density $\pi(z)$) vs. $z_{\min p}$

Statistical metatheorem: Likelihood method will be as or more efficient (have smaller variance of error distribution) than frequentist method

Comparison: Poolmap vs. Minimum p-value



Poolmap estimator is technically *LESS* efficient than minimum p-value estimator because of rare extremely large errors, but has more density around small errors

Width of **profile likelihood will give a “warning” when a large error occurs**; there is no analogue in the minimum p-value procedure

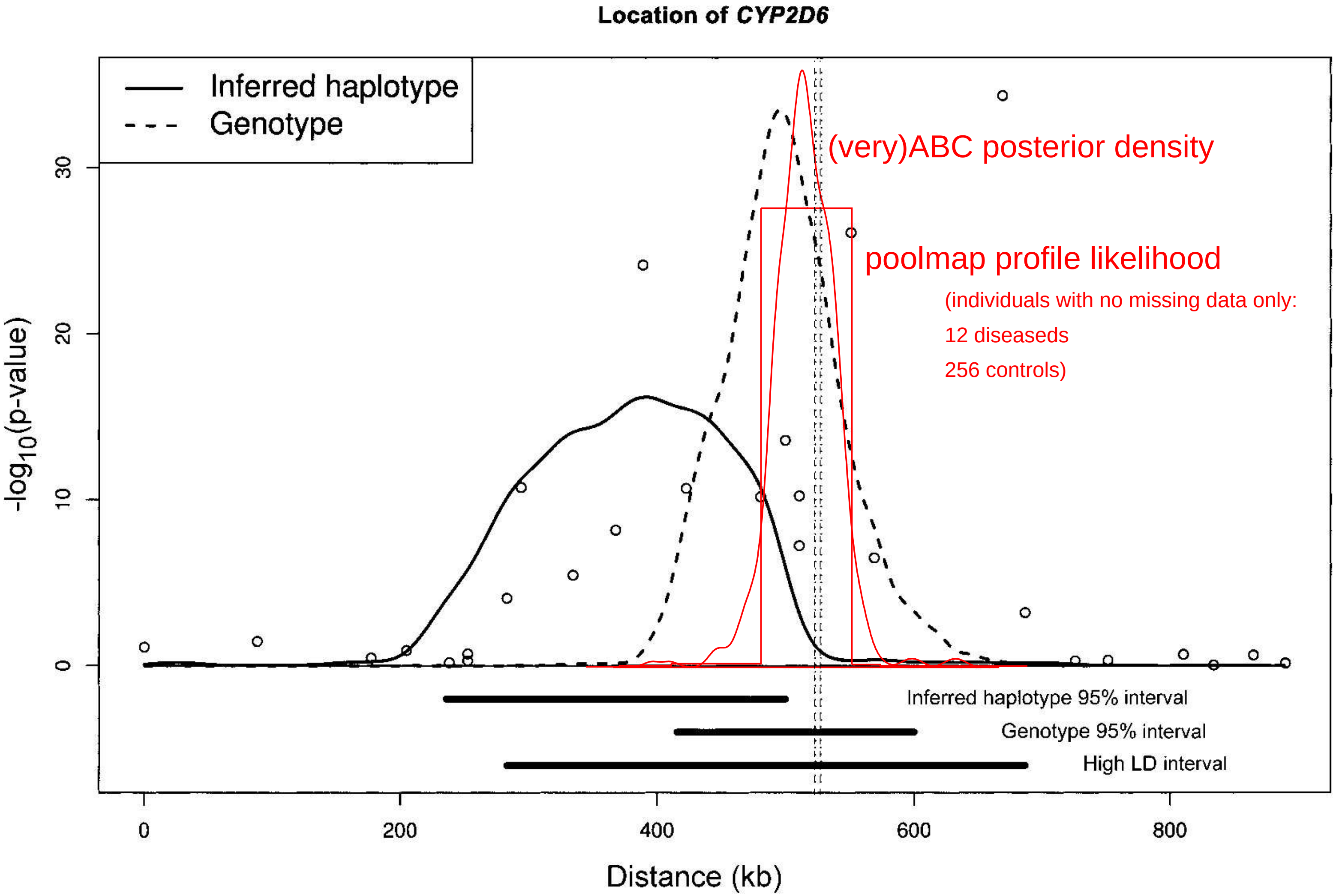
Comparison: Poolmap vs. Minimum p-value

When disease locus position z^* is restricted to $[0.5\text{cM}, 1.5\text{cM}]$ Poolmap generally gives more efficient point estimates that the minimum p-value method

n	Allele age	f_p	Poolmap			Minimum p-value		
			$(\mu_z - z^*)$	$\lvert\mu_z - z^*\rvert$		$(z_{\min p} - z^*)$	$\lvert z_{\min p} - z^*\rvert$	
			s.d.	$Q_{0.5}$	$Q_{0.9}$	s.d.	$Q_{0.5}$	$Q_{0.9}$
			cM	cM	cM	cM	cM	cM
10	100	0	0.159^{-}	0.063^{-}	0.187^{-}	0.296^{+}	0.158^{+}	0.492^{+}
28	100	0	0.058^{-}	0.035^{-}	0.091^{-}	0.202^{+}	0.109^{+}	0.342^{+}
82	100	0	0.037^{-}	0.021^{-}	0.057^{-}	0.156^{+}	0.082^{+}	0.241^{+}
10	1500	0	0.255^{ns}	0.069^{ns}	0.358^{ns}	0.224^{ns}	0.079^{ns}	0.330^{ns}
28	1500	0	0.102^{ns}	0.019^{-}	0.090^{-}	0.102^{ns}	0.033^{+}	0.132^{+}
82	1500	0	0.026^{-}	0.011^{-}	0.033^{-}	0.050^{+}	0.023^{+}	0.078^{+}
10	100	0.75	$0.532^{(+)}$	0.361^{ns}	$0.892^{(+)}$	$0.477^{(-)}$	0.311^{ns}	$0.803^{(-)}$
28	100	0.75	0.453^{ns}	0.225^{ns}	0.794^{ns}	0.423^{ns}	0.229^{ns}	0.716^{ns}
82	100	0.75	0.255^{-}	0.116^{-}	0.384^{-}	0.344^{+}	0.190^{+}	0.600^{+}

$+/-$: $p \leq 0.01$; $(+)/(-)$: $0.01 < p \leq 0.05$; ns: $0.05 < p$ estimated by bootstrapping

Real Data: Cytochrome p450 Enzyme



(very)ABC posterior density

poolmap profile likelihood

(individuals with no missing data only:
12 diseaseds
256 controls)

Modified from Morris *et al.* 2004 AJHG **74**:945

Summary

- In gene *mapping* **region estimates are *REQUIRED*** (and not merely preferable)
- Multipoint analysis of multilocus allele frequency data is possible
- Method described is **robust** to unknown population history, unknown rate of phenocopies, and unknown dominance
- Works "quite well" if modelling assumptions are violated, e.g. allele affecting trait is common, and markers not at linkage equilibrium
- Data sets of up to 1000 markers can be analysed quickly
- Power analysis (for one case; not shown) suggested that
  - Roughly $3\times$ wider region estimates are obtained by Poolmap than by Bayesian analysis of fully resolved haplotypes
  - Roughly $3\times$ marker density can compensate for this
- Bias and efficiency of point estimates should not be sole criteria for judging performance
- Functions of $\mathrm{L}_{\mathrm{max}}(\cdot)$ provide rapidly calculatable summary statistics that can be used for e.g. Approximate Bayesian Computation, or multipoint significance tests

Acknowledgements

- Intellectual
  - Stuart Baird
  - Nick Barton
  - Kevin Dawson
  - Dick Hudson
  - Mark Kirkpatrick
  - Monty Slatkin
  - Jay Taylor
  - Peter Visscher
  - Jeff Reeve for help with DMLE+
  - Psychiatric Genetics Group at Western General Hospital, Edinburgh
- Financial
  - Wellcome Trust
  - BBSRC
  - Mathematical Population Genetics programme at Erwin Schrödinger Institute
  - vn PIII/Linux cluster in the Department of Physics at UBC