Haplotype tagging for the identification of common disease genes

Johnson et al (2001)

Rationale for choosing this paper

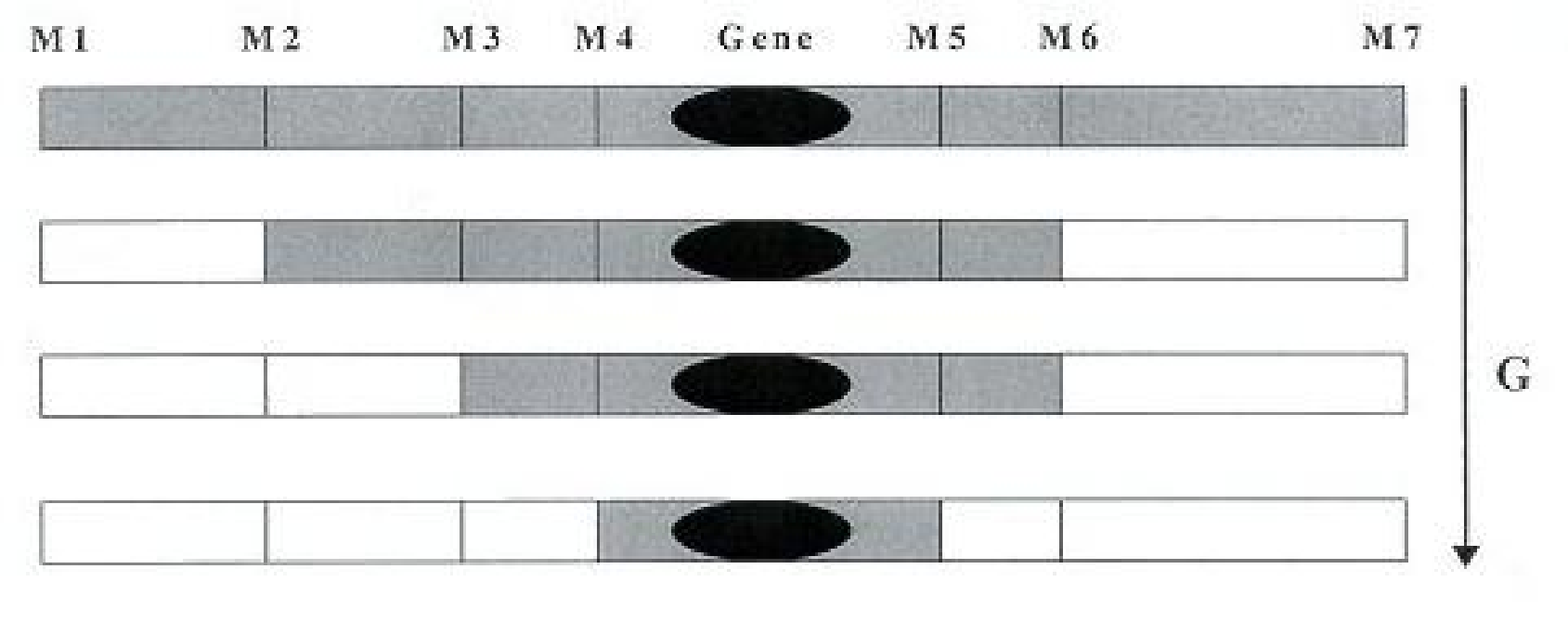
- First paper to suggest use of tagging SNPs
- We use tagging, should understand how/why it works

Context (paper published 2001)

- Genotyping considerably more expensive $\longrightarrow$ even greater incentive to reduce genotyping burder
- SNP coverage much less dense and more variable (pre-HapMap)

BACKGROUND: What is linkage disequilibrium?

Recombination is rare in small distances



Markers today tend to be surrounded by genetic material from ancestral chromosome

Linkage Disequilibrium exists when

- There has been little mutation between polymorhisms that
- Arose on the same haplotype

BACKGROUND: How to measure LD

Measures of LD - D' or d^2?

- Variety of measures
- Pick your favourite!
- Understand what they mean:
  - Both between 0 and 1
  - D' direct measure of deviation from 'equilibrium'
  - r^2 information one snp gives you about the other
- d^2 in paper closely related to r^2 - I will use interchangably

Example

Haplotype	freq	
11	20	
12	50	$D' = 1$
21	0	$r^2 = 0.11$
22	30	

Haplotype	freq	
11	20	
12	0	$D' = 1$
21	0	$r^2 = 1$
22	80	

Paper claims

To demonstrate how knowledge of LD in a gene can

- explain patterns of LD between adjacent markers
- reduce genotyping significantly (in this case from 122 to 34 SNPs)
- scan the common variation of a gene sensitively and comprehensively
- provide key fine-mapping data within regions of strong LD

Limited utility of dbSNP

- Possibly less true now?
- Would probably rely on HapMap
- If using HapMap, be aware of how SNPs are chosen

Methods

Polymorphism detection

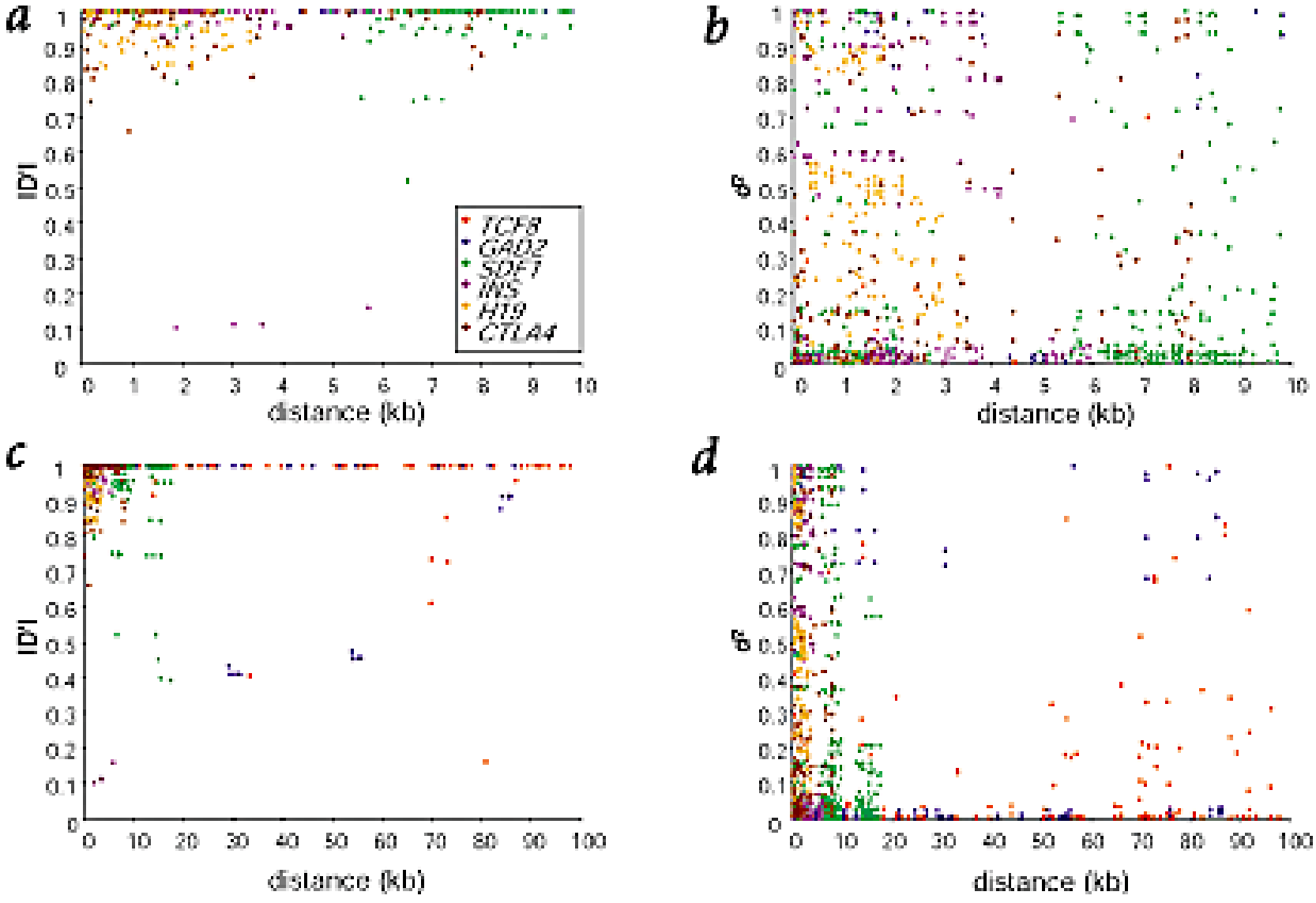
- scanned 135kb of DNA
- in 9 genes: TCF8, GAD2, SDF1, INS, H19, CTLA4, CASP8, CASP10, CFLAR
- within 5 kb of the 5' UTR and 3 kb of the 3' UTR
- using WAVE or resequencing
- screening panel of unrelated individuals: $20 \leq n \leq 46$

Methods

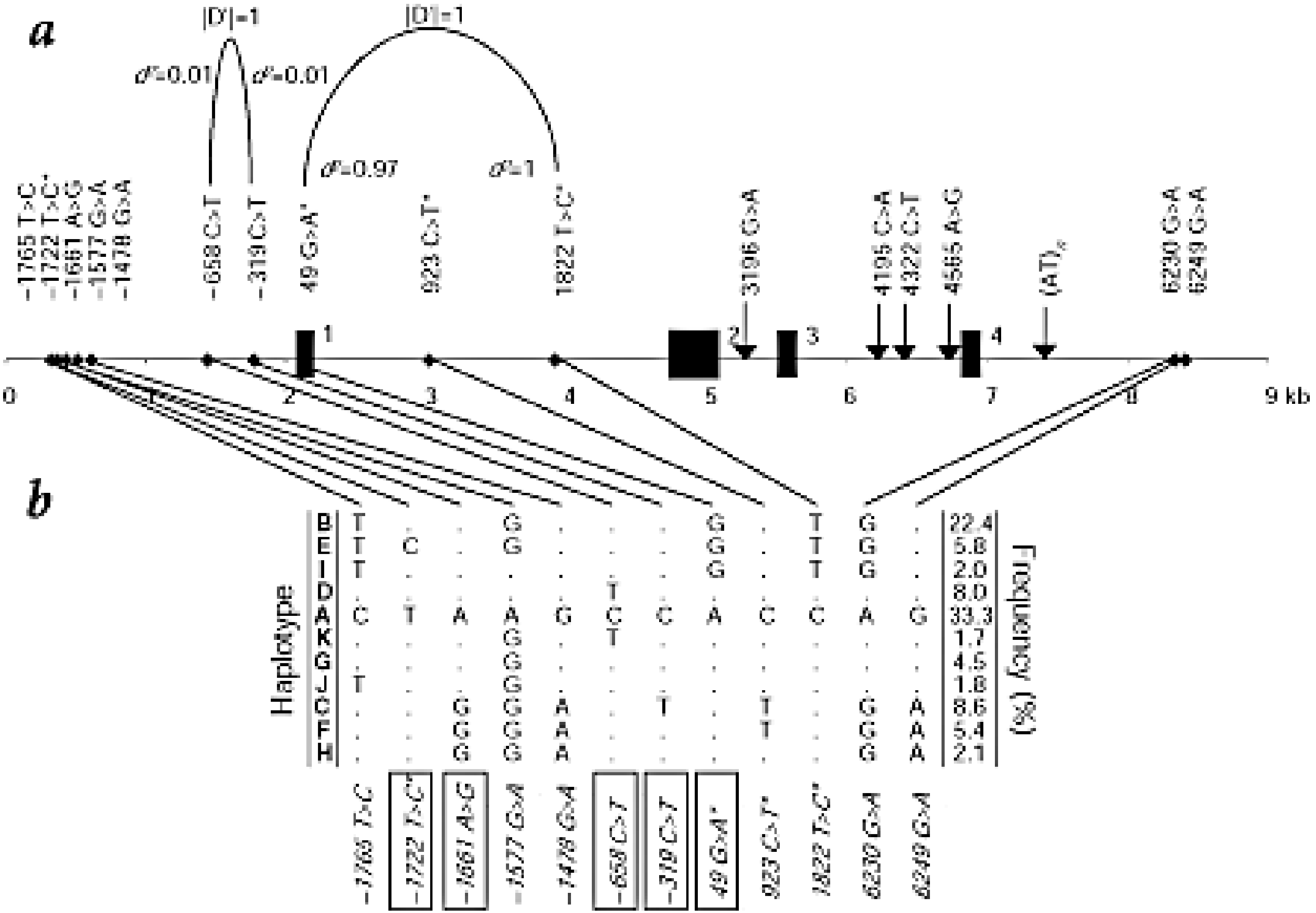
Haplotype reconstruction

- genotyped SNPs in UK or Finnish families or unrelateds (CTLA4)
- family data: constructed 4 unrelated unambigously resolved haplotypes
- unrelateds: estimated haplotype frequencies in CTLA4 (EM algorithm)
- excluded rare haplotypes (observed only once, or $< 1\%$)
- quantified LD - D', d^2

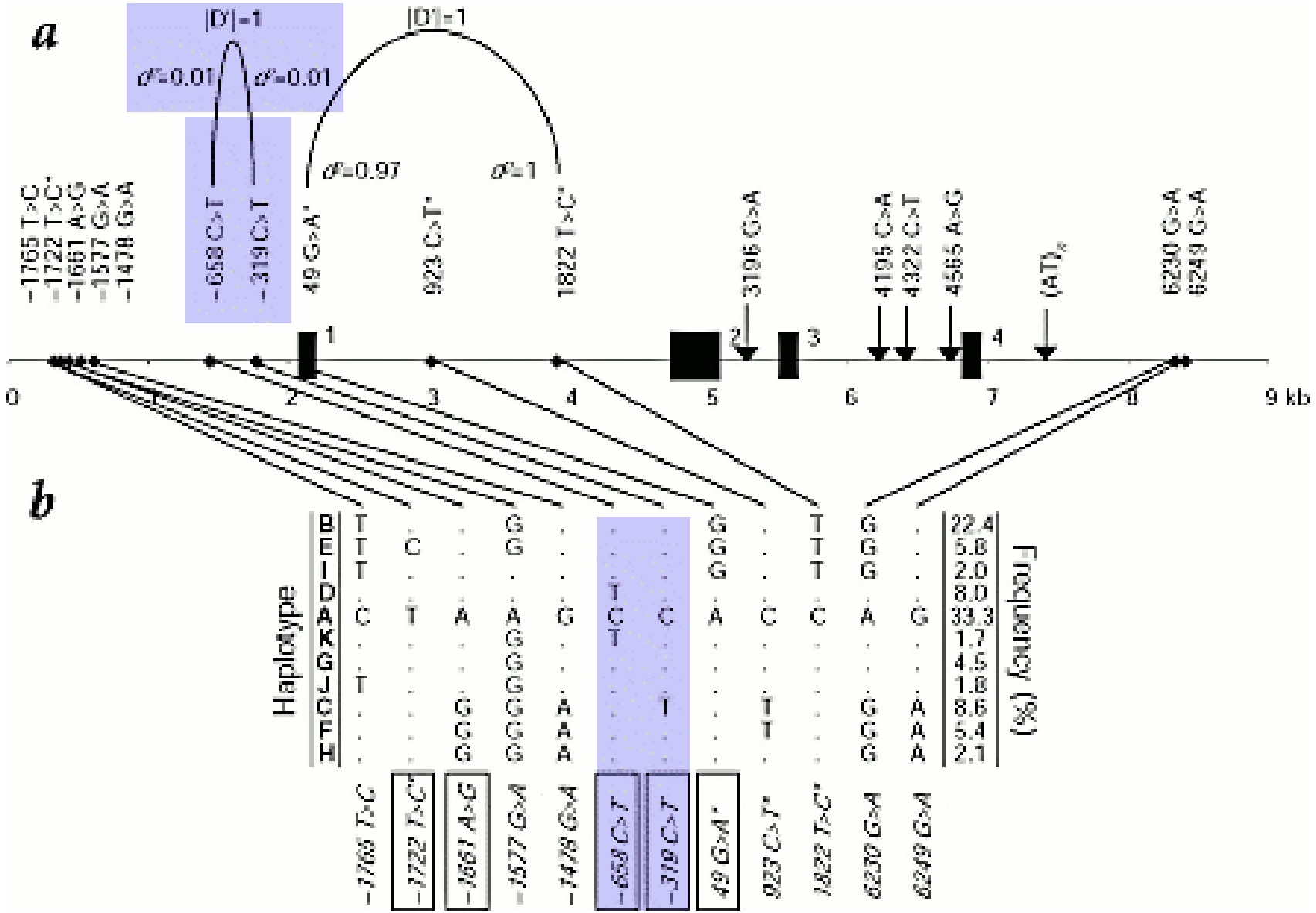
Relationship of LD to physical distance



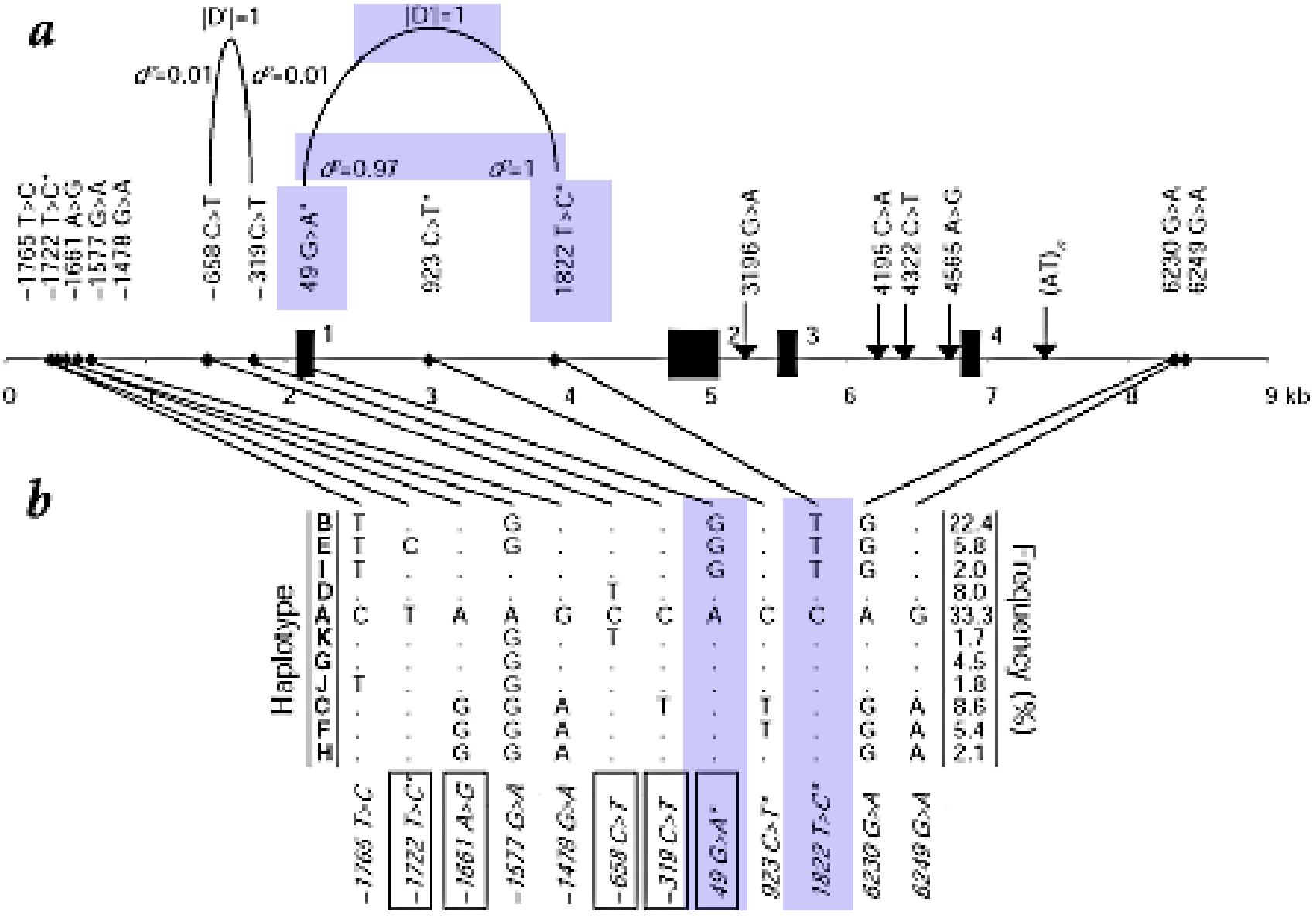
Greater resolution afforded by studying haplotypes



Greater resolution afforded by studying haplotypes



Greater resolution afforded by studying haplotypes



Tagging SNPs

Tag SNP choice

- By eye
- Searching all possible subsets for optimal set of given size

Define 'optimal' in terms of 'residual haplotype diversity within groups of haplo- types identified by the htSNPs'

- maximise overall percent diversity explained
- maximise worst of the locus-specific values

What does haplotype diversity mean?

"Haplotype diversity is measured by the total number of differences recorded in all possible pairwise comparisons of the haplotypes..."

What does haplotype diversity mean?

> *"Haplotype diversity is measured by the total number of differences recorded in all possible pairwise comparisons of the haplotypes..."*

Example

- Genotype 2 people
- Observe haplotypes 111 111 112 222
- Differences:

What does haplotype diversity mean?

> *"Haplotype diversity is measured by the total number of differences recorded in all possible pairwise comparisons of the haplotypes..."*

Example

- Genotype 2 people
- Observe haplotypes **111** **111** 112 222
- Differences: 0

What does haplotype diversity mean?

> *"Haplotype diversity is measured by the total number of differences recorded in all possible pairwise comparisons of the haplotypes..."*

Example

- Genotype 2 people
- Observe haplotypes **111** 111 **112** 222
- Differences: 0 + 1

What does haplotype diversity mean?

> *"Haplotype diversity is measured by the total number of differences recorded in all possible pairwise comparisons of the haplotypes..."*

Example

- Genotype 2 people
- Observe haplotypes **111** 111 112 **222**
- Differences: $0 + 1 + 3$

What does haplotype diversity mean?

> *"Haplotype diversity is measured by the total number of differences recorded in all possible pairwise comparisons of the haplotypes..."*

Example

- Genotype 2 people
- Observe haplotypes $111 \quad \mathbf{111} \quad \mathbf{112} \quad 222$
- Differences: $0 + 1 + 3 + 1$

What does haplotype diversity mean?

> *"Haplotype diversity is measured by the total number of differences recorded in all possible pairwise comparisons of the haplotypes..."*

Example

- Genotype 2 people
- Observe haplotypes $111 \quad \mathbf{111} \quad 112 \quad \mathbf{222}$
- Differences: $0 + 1 + 3 + 1 + 3$

What does haplotype diversity mean?

> *"Haplotype diversity is measured by the total number of differences recorded in all possible pairwise comparisons of the haplotypes..."*

Example

- Genotype 2 people
- Observe haplotypes $111 \quad 111 \quad \mathbf{112} \quad \mathbf{222}$
- Differences: $0 + 1 + 3 + 1 + 3 + 2$

What does haplotype diversity mean?

> *"Haplotype diversity is measured by the total number of differences recorded in all possible pairwise comparisons of the haplotypes..."*

Example

- Genotype 2 people
- Observe haplotypes $111 \quad 111 \quad 112 \quad 222$
- Differences: $0 + 1 + 3 + 1 + 3 + 2 = 10$

What does haplotype diversity mean?

> *"Haplotype diversity is measured by the total number of differences recorded in all possible pairwise comparisons of the haplotypes..."*

Example

- Genotype 2 people
- Observe haplotypes 111 111 112 222
- Differences: $0 + 1 + 3 + 1 + 3 + 2 = 10$

Question

What tag set would **you** pick?

Using haplotype diversity to evaluate a tagging set

> *"...and it is simple to calculate the proportion of diversity 'explained' by a given choice of htSNPs, both overall and locus by locus."*

Using haplotype diversity to evaluate a tagging set

> *"...and it is simple to calculate the proportion of diversity 'explained' by a given choice of htSNPs, both overall and locus by locus."*

Example - tag snps set

- Genotype only snps 1 and 2
- Observe haplotypes 111 111 112 222
- Differences:

Using haplotype diversity to evaluate a tagging set

> *"...and it is simple to calculate the proportion of diversity 'explained' by a given choice of htSNPs, both overall and locus by locus."*

Example - tag snps set

- Genotype only snps 1 and 2
- Observe haplotypes **11**1 **11**1 11**2** 22**2**
- Differences: 0

Using haplotype diversity to evaluate a tagging set

> *"...and it is simple to calculate the proportion of diversity 'explained' by a given choice of htSNPs, both overall and locus by locus."*

Example - tag snps set

- Genotype only snps 1 and 2
- Observe haplotypes **11**1 11**1** **11**2 22**2**
- Differences: 0 + 0

Using haplotype diversity to evaluate a tagging set

> *"...and it is simple to calculate the proportion of diversity 'explained' by a given choice of htSNPs, both overall and locus by locus."*

Example - tag snps set

- Genotype only snps 1 and 2
- Observe haplotypes **11**1 11**1** 11**2** **22**2
- Differences: 0 + 0 + 0

Using haplotype diversity to evaluate a tagging set

> *"...and it is simple to calculate the proportion of diversity 'explained' by a given choice of htSNPs, both overall and locus by locus."*

Example - tag snps set

- Genotype only snps 1 and 2
- Observe haplotypes $111 \quad \mathbf{111} \quad \mathbf{112} \quad 222$
- Differences: $0 + 0 + 0 + 0$

Using haplotype diversity to evaluate a tagging set

> *"...and it is simple to calculate the proportion of diversity 'explained' by a given choice of htSNPs, both overall and locus by locus."*

Example - tag snps set

- Genotype only snps 1 and 2
- Observe haplotypes 111 $\mathbf{111}$ 112 $\mathbf{222}$
- Differences: $0 + 0 + 0 + 0 + 0$

Using haplotype diversity to evaluate a tagging set

> *"...and it is simple to calculate the proportion of diversity 'explained' by a given choice of htSNPs, both overall and locus by locus."*

Example - tag snps set

- Genotype only snps 1 and 2
- Observe haplotypes $111 \quad 111 \quad \mathbf{11}2 \quad \mathbf{22}2$
- Differences: $0 + 0 + 0 + 0 + 0 + 0$

Using haplotype diversity to evaluate a tagging set

> *"...and it is simple to calculate the proportion of diversity 'explained' by a given choice of htSNPs, both overall and locus by locus."*

Example - tag snps set

- Genotype only snps 1 and 2
- Observe haplotypes $111 \quad 111 \quad 112 \quad 222$
- Differences: $0 + 0 + 0 + 0 + 0 + 0 = 0$
- Proportion of diversity explained $= 0/10 = 0\%$

Using haplotype diversity to evaluate a tagging set

> *"...and it is simple to calculate the proportion of diversity 'explained' by a given choice of htSNPs, both overall and locus by locus."*

Using haplotype diversity to evaluate a tagging set

> *"...and it is simple to calculate the proportion of diversity 'explained' by a given choice of htSNPs, both overall and locus by locus."*

Example - tag snps set

- Genotype only snps 1 and 3
- Observe haplotypes 111 111 112 222
- Differences:

Using haplotype diversity to evaluate a tagging set

> *"...and it is simple to calculate the proportion of diversity 'explained' by a given choice of htSNPs, both overall and locus by locus."*

Example - tag snps set

- Genotype only snps 1 and 3
- Observe haplotypes **111** **111** 112 222
- Differences: 0

Using haplotype diversity to evaluate a tagging set

> *"...and it is simple to calculate the proportion of diversity 'explained' by a given choice of htSNPs, both overall and locus by locus."*

Example - tag snps set

- Genotype only snps 1 and 3
- Observe haplotypes **11**1 11**2** 111 222
- Differences: 0 + 1

Using haplotype diversity to evaluate a tagging set

> *"...and it is simple to calculate the proportion of diversity 'explained' by a given choice of htSNPs, both overall and locus by locus."*

Example - tag snps set

- Genotype only snps 1 and 3
- Observe haplotypes **111** 111 112 **222**
- Differences: $0 + 1 + 3$

Using haplotype diversity to evaluate a tagging set

> *"...and it is simple to calculate the proportion of diversity 'explained' by a given choice of htSNPs, both overall and locus by locus."*

Example - tag snps set

- Genotype only snps 1 and 3
- Observe haplotypes $111 \quad \mathbf{111} \quad \mathbf{112} \quad 222$
- Differences: $0 + 1 + 3 + 1$

Using haplotype diversity to evaluate a tagging set

> *"...and it is simple to calculate the proportion of diversity 'explained' by a given choice of htSNPs, both overall and locus by locus."*

Example - tag snps set

- Genotype only snps 1 and 3
- Observe haplotypes $111 \quad \mathbf{111} \quad 112 \quad \mathbf{222}$
- Differences: $0 + 1 + 3 + 1 + 3$

Using haplotype diversity to evaluate a tagging set

> *"...and it is simple to calculate the proportion of diversity 'explained' by a given choice of htSNPs, both overall and locus by locus."*

Example - tag snps set

- Genotype only snps 1 and 3
- Observe haplotypes $111 \quad 111 \quad \mathbf{112} \quad \mathbf{222}$
- Differences: $0 + 1 + 3 + 1 + 3 + 2$

Using haplotype diversity to evaluate a tagging set

> *"...and it is simple to calculate the proportion of diversity 'explained' by a given choice of htSNPs, both overall and locus by locus."*

Example - tag snps set

- Genotype only snps 1 and 3
- Observe haplotypes $111 \quad 111 \quad 112 \quad 222$
- Differences: $0 + 1 + 3 + 1 + 3 + 2 = 10$
- Proportion of diversity explained $= 10/10 = 100\%$

Using haplotype diversity to evaluate a tagging set

> *"...and it is simple to calculate the proportion of diversity 'explained' by a given choice of htSNPs, both overall and locus by locus."*

Example - tag snps set

- Genotype only snps 1 and 3
- Observe haplotypes 111 111 112 222
- Differences: $0 + 1 + 3 + 1 + 3 + 2 = 10$
- Proportion of diversity explained $= 10/10 = 100\%$
- NB - tag snp sets 1,3 or 2,3 equivalent
- Often more than one possible tag snp set

Number of tag SNPs selected

Gene	No. SNPs	No. tag SNPS
CFLAR	6	2
CASP10	11	4
GAD2	13	2
H19	13	4
SDF1	22	5
INS	14	3
TCF8	14	5
CASP8	13	3
CTLA4	12	5

Tag snps capture common variation

> *"... 25 htSNPs can be used to define the six or fewer common haplotypes (greater than 5% population frequency) [which] account for at least 80% of all haplotypes that we observed"*

Tag snps capture common variation

"For common haplotypes ($>$ 5% frequency), 5,893 cases and controls required to have 80% power to detect a disease variant with an odds ratio of 1.5 at $p = 10^{-5}$"

Tag snps capture common variation

"If suitably powered studies that evaluate common variation fail to produce convincing associations, the multiple-rare-variants model of common disease will then need to be considered."

Were claims met?

To demonstrate how knowledge of LD in a gene can

- explain patterns of LD between adjacent markers ?

Were claims met?

To demonstrate how knowledge of LD in a gene can

- explain patterns of LD between adjacent markers ?
- reduce genotyping significantly (in this case from 122 to 34 SNPs) ✓

Were claims met?

To demonstrate how knowledge of LD in a gene can

- explain patterns of LD between adjacent markers ?
- reduce genotyping significantly (in this case from 122 to 34 SNPs) ✓
- scan the common variation of a gene sensitively and comprehensively ✓

Were claims met?

To demonstrate how knowledge of LD in a gene can

- explain patterns of LD between adjacent markers ?
- reduce genotyping significantly (in this case from 122 to 34 SNPs) ✓
- scan the common variation of a gene sensitively and comprehensively ✓
- provide key fine-mapping data within regions of strong LD ✓

Were claims met?

To demonstrate how knowledge of LD in a gene can

- explain patterns of LD between adjacent markers ?
- reduce genotyping significantly (in this case from 122 to 34 SNPs) ✓
- scan the common variation of a gene sensitively and comprehensively ✓
- provide key fine-mapping data within regions of strong LD ✓

Further claims for advantages of haplotype study

- cis interactions ✓(only if we study haplotypes of tag SNPs)

Were claims met?

To demonstrate how knowledge of LD in a gene can

- explain patterns of LD between adjacent markers ?
- reduce genotyping significantly (in this case from 122 to 34 SNPs) ✓
- scan the common variation of a gene sensitively and comprehensively ✓
- provide key fine-mapping data within regions of strong LD ✓

Further claims for advantages of haplotype study

- cis interactions ✓(only if we study haplotypes of tag SNPs)
- exploiting differences in haplotype diversity and frequency between populations to pinpoint which variants are most likely to be the primary etiological determinants of common diseases ?

Other methods for tag SNP selection

All subset searches

As before, but define 'optimal' in terms of

- maximize the minimum r^2 at untyped loci
- maximize the *mean* r^2 at untyped loci

Faster search algorithms

- When all subset searches not possible
- Not guaranteed to find optimal solution

e.g. Carlson - clusters SNPs into groups with similar r^2, picks one from each group

e.g. Principle components - identify set of independent SNPs, dropping redundant

Other methods for tag SNP selection

Use haplotypes of tag SNPs

- The haplotype of a pair of tag SNPs should predict other SNPs more efficiently than the pair of SNPs
- Extra information comes from knowing phase
- WARNING! Must infer phase of tag snps before analysis
- WARNING! Must use all possible phase assignments in analysis

e.g. Paul de Bakker's Tagger (`http://www.broad.mit.edu/mpg/tagger`)

- Implemented, with 'a number of differences', in haploview

So what about rare variants?

Using standard statistical methods

- Rare variants are rarely taggable
- Must be typed directly
- Difficult to detect association without *huge* samples (low power)

So what about rare variants?

Using standard statistical methods

- Rare variants are rarely taggable
- Must be typed directly
- Difficult to detect association without *huge* samples (low power)

Alternative statistical methods under development

- Exploit idea: Affected individuals will be more closely related than unaffecteds
- Therefore should sharing longer haplotypes, on average
- So perhaps tagging will work after all...