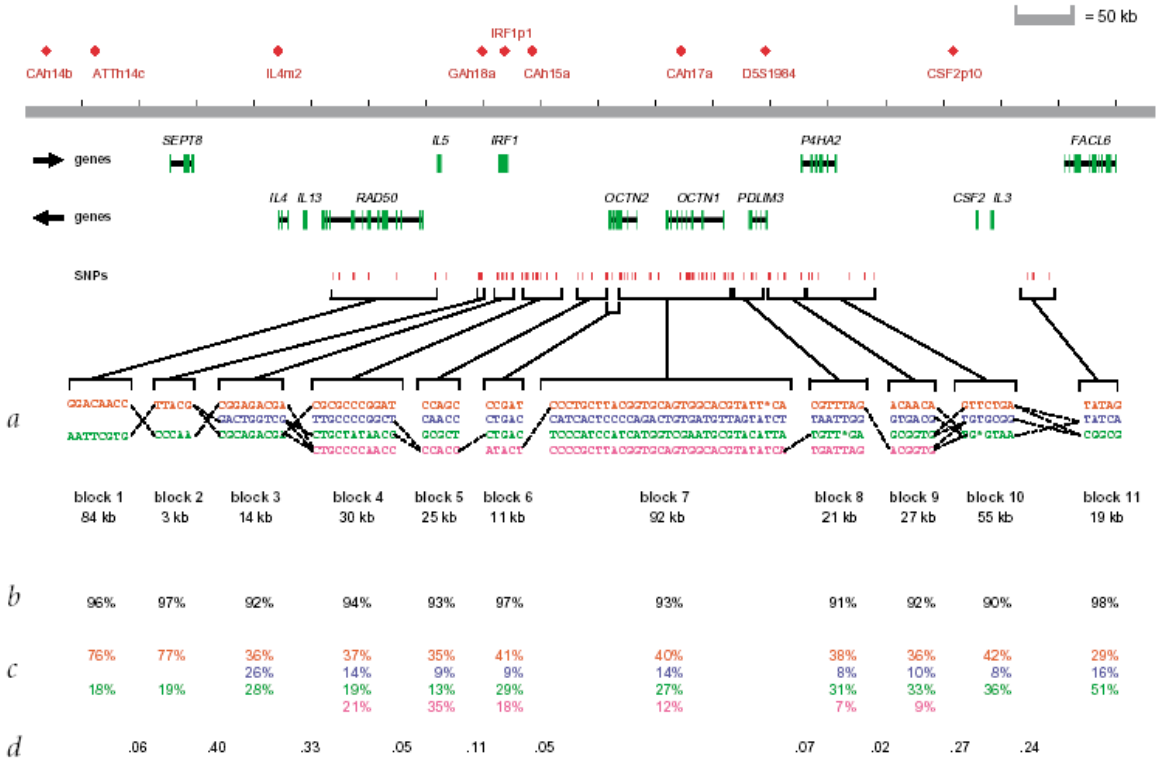
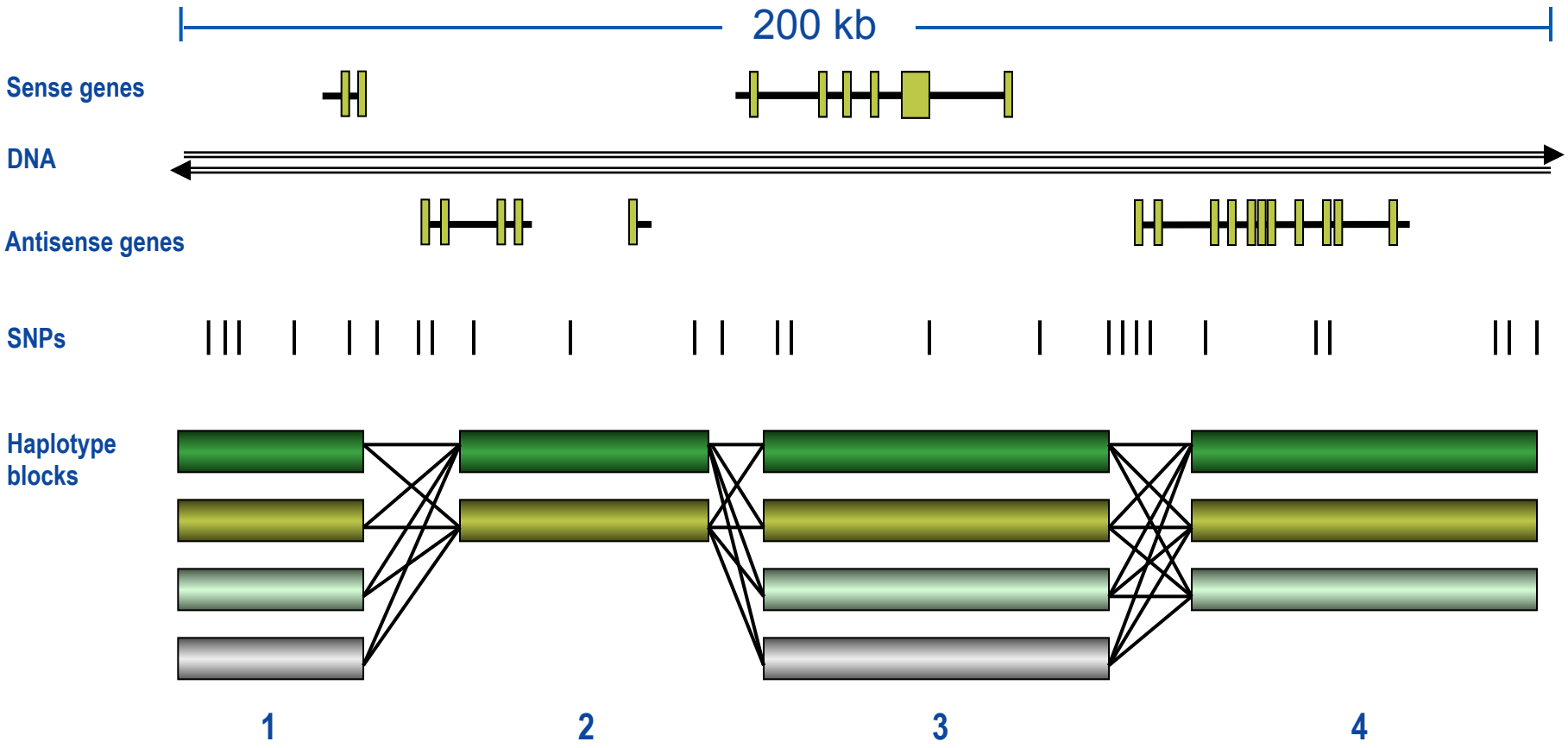
Compresssion based on Haplotype Blocks

Hypothesis – Haplotype Blocks?



- The genome consists largely of blocks of common SNPs with relatively little recombination within the blocks
  - Patil et al., Science, 2001;
  - Jeffreys et al., Nature Genetics, 2001;
  - Daly et al., Nature Genetics, 2001

Haplotype Block Structure
LD-Blocks, and 4-Gamete Test Blocks



Four Gamete Block Test

- Hudson and Kaplan 1985

A segment of SNPs is a **block** if between every pair of SNPs at most 3 out of the 4 gametes (00, 01,10,11) are observed.

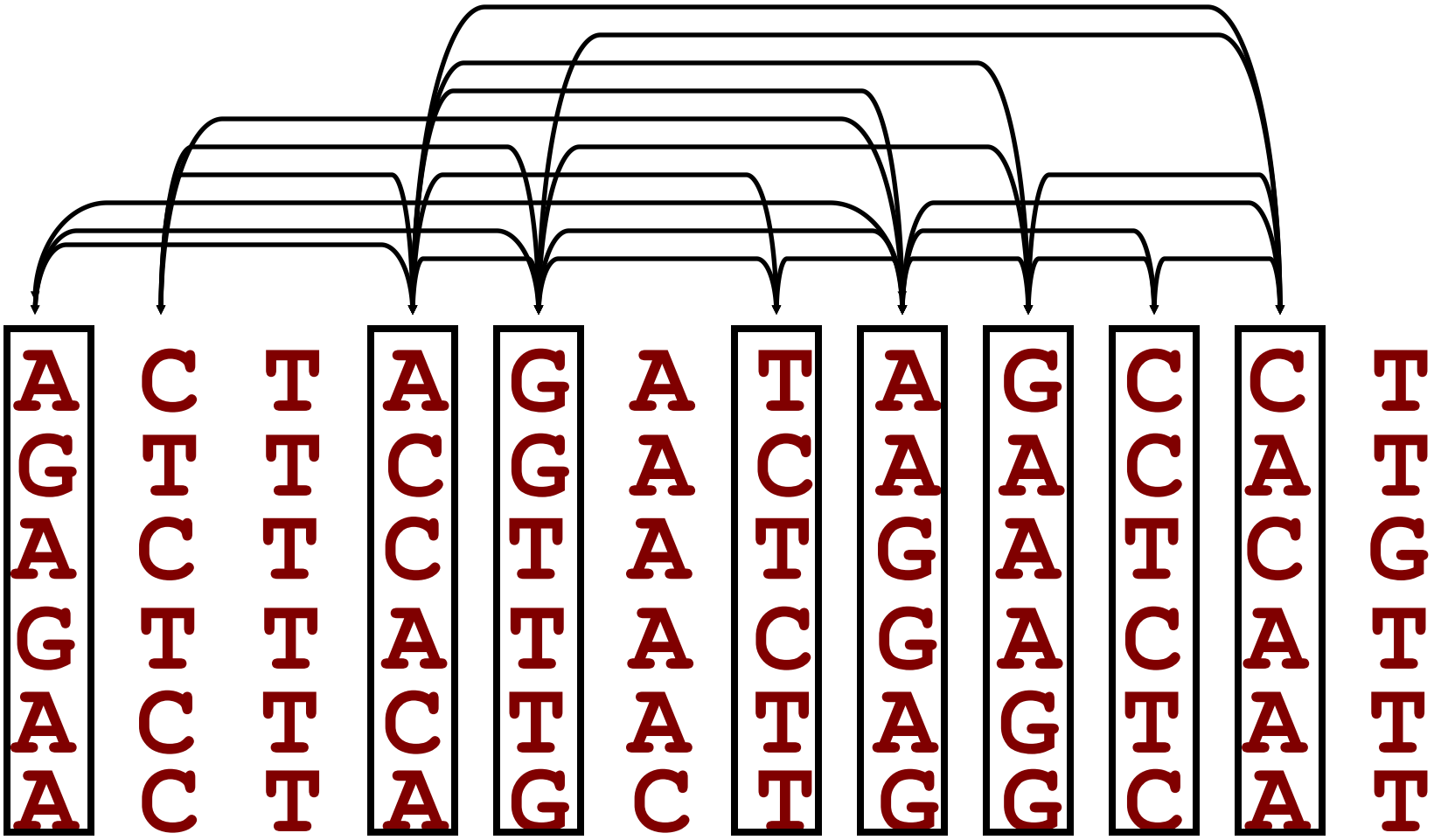
0	0	1
0	1	1
1	1	0
1	1	1

BLOCK

0	0	1
0	1	1
1	1	0
1	0	1

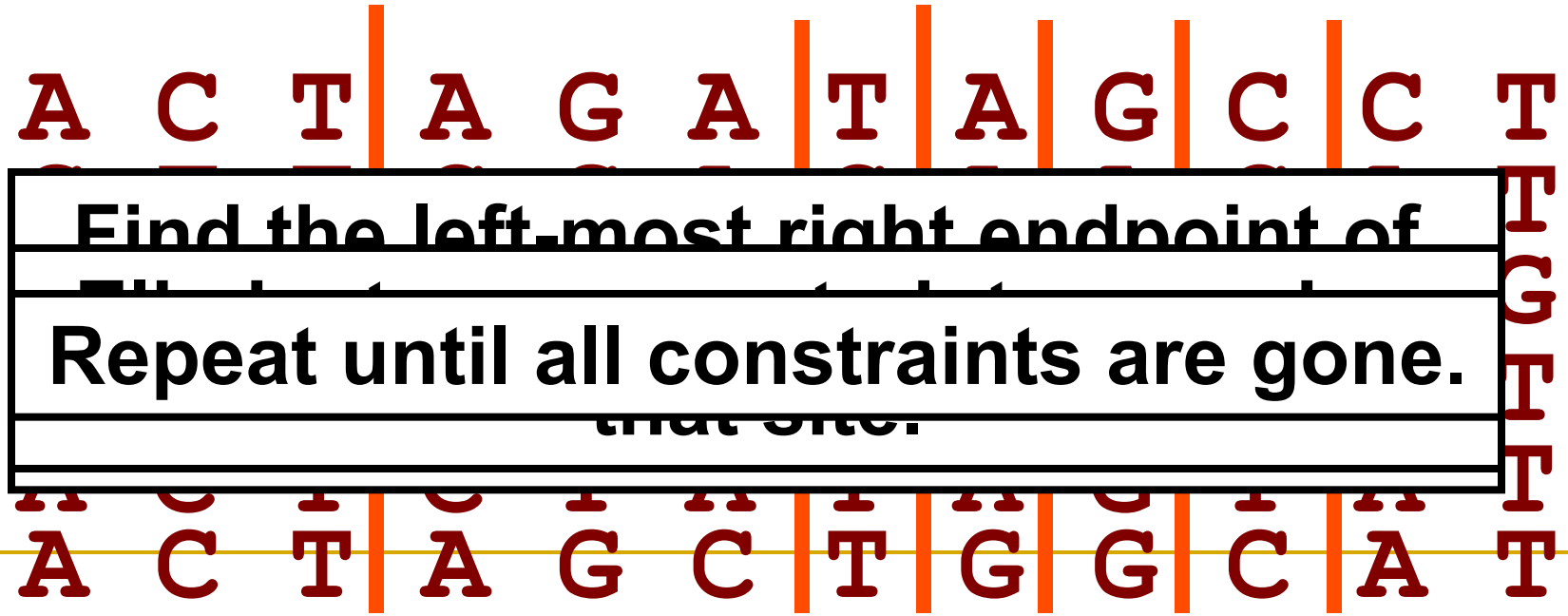
VIOLATES THE BLOCK DEFINITION

Finding Recombination Hotspots:
Many Possible Partitions into Blocks



All four gametes are present:

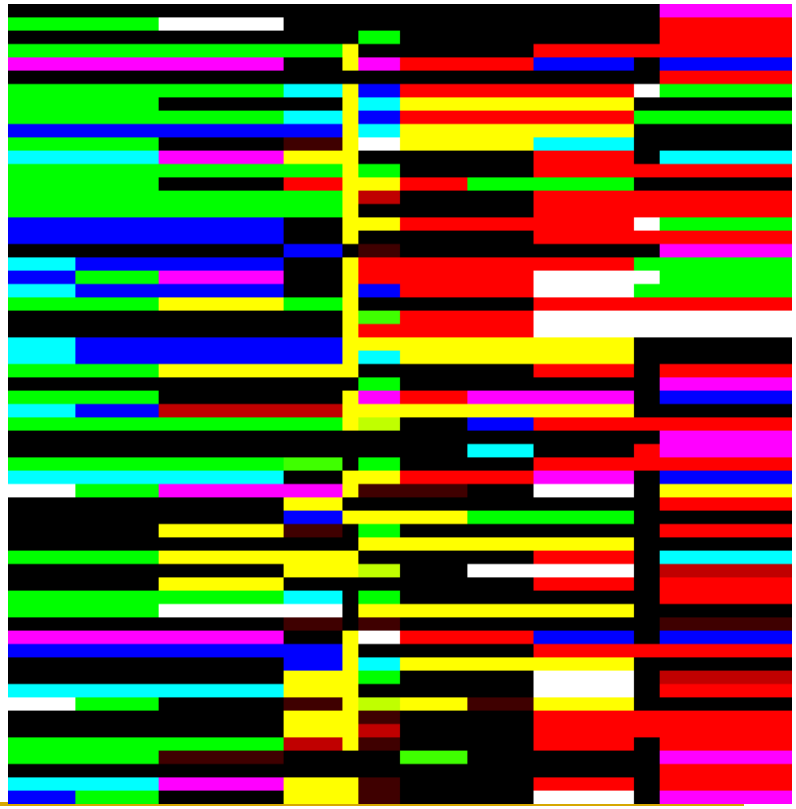
The final result is a minimum-size set of sites crossing all constraints.



Dynamic programming framework

Partitioning a chromosome into blocks

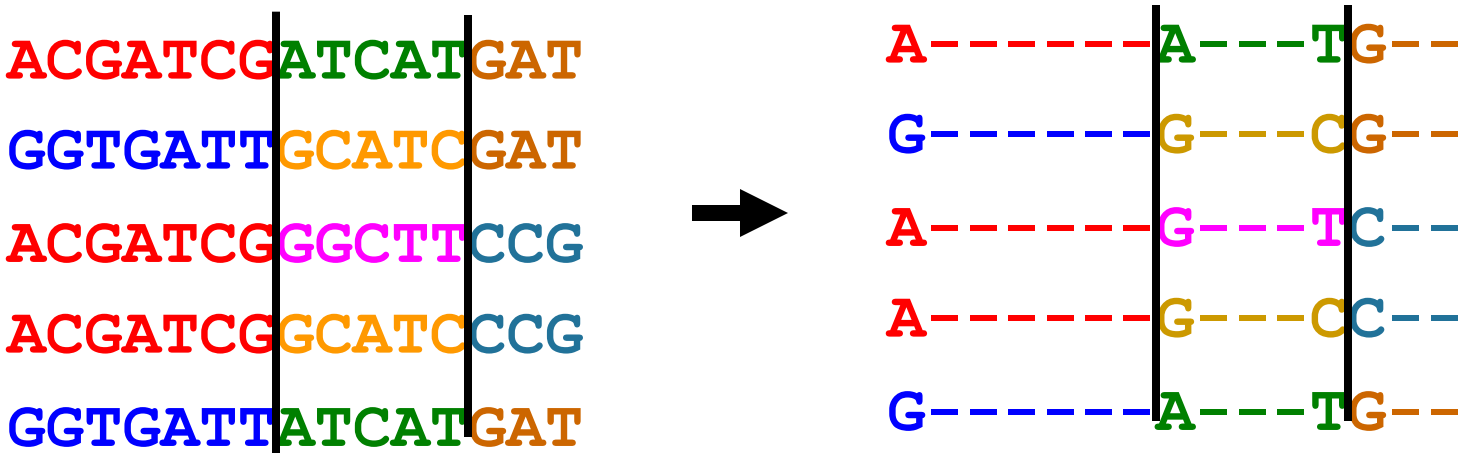
- Zhang et al. (PNAS, 2002).
- Zhang et al. RECOMB, 2003
- H. I. Avi-Itzhak et al. PSB, 2003
- Sebastiani et al. PNAS 2003
- Patil et al., PNAS 2002.



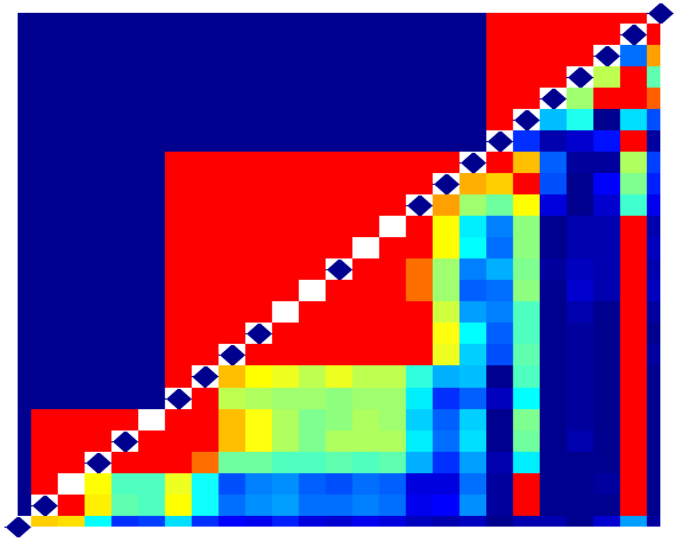
Parametric in block test

- Solve a dynamic program
  - Optimal block partition requires the minimal number of blocks.
  - Within blocks one can select the SNPs that maximize entropy, diversity or r2 correlation

Data Compression



Selecting Tagging SNPs in blocks



Haplotype Blocks based on LD
(Method of Gabriel et al.2002)

SNP-Selection Axioms: LD consistency and Block-freeness

The highly concordant results of the block detection methods make the interior of LD blocks adequate for sparse SNP selection. However, block boundaries defined by these methods are not sharp, with no single "true" block partition. SNP selection should avoid dependence of particular definitions of " haplotype block."

A New Measure

Informativeness

A New SNP Selection Measure:

Informativeness

It satisfies the following six **Axioms**:

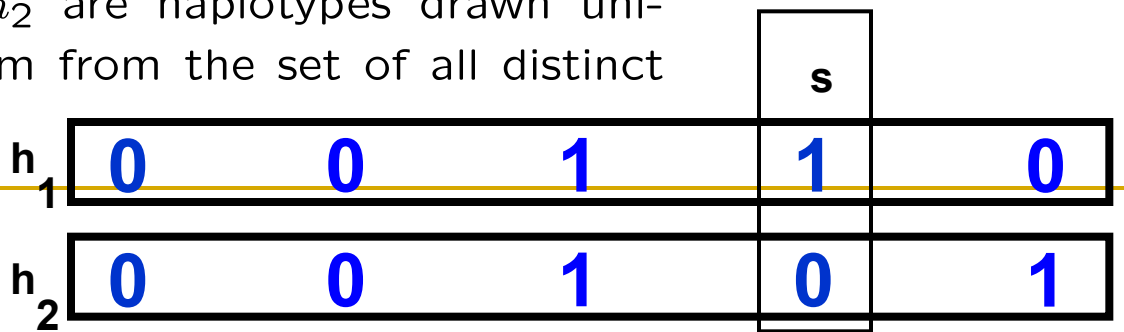
1. **Multi-allelic measure**
2. **LD consistency**: compares well with measures of LD
3. **Block-freeness**: independence on any particular block definition
4. **Hypothesis-free associations**: optimization achieves maximum haplotype resolution
5. **Algorithmically sound**: practical for genome-wide computations
6. **Statistically sound**: passes overfitting and imputation tests

Informativeness

Informativeness of a SNP s with respect to another SNP t quantifies the confidence with which we can predict t from s.

Let s be a SNP and h_1, h_2 two haplotypes. Let $D^s(h_1, h_2)$ be the event that h_1 and h_2 have different alleles at s.

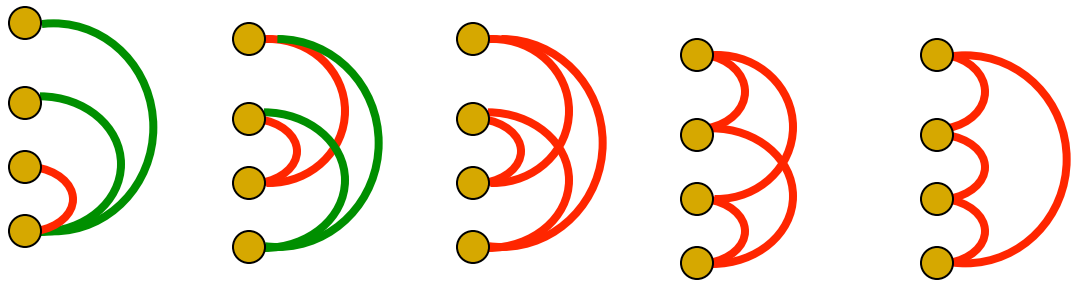
The **informativeness** of s with respect to t is given by $I(s,t) = Pr[D^s(h_1,h_2) \mid D^t(h_1,h_2)]$, where h_1 and h_2 are haplotypes drawn uniformly at random from the set of all distinct haplotype pairs.



Informativeness

For each SNP define a bipartite graph, having the set of haplotypes as nodes and an edge exists between two haplotypes when the two alleles at s are different. Let $E(s)$ be the set of edges. We estimate I from the sample as follows.

- $I(S,t) = \frac{|E(s) \cap E(t)|}{E(t)}$
- $I(S,t) = \frac{|(\cup_{s \in S} E(s)) \cap E(t)|}{E(t)}$
- $I(S,T) = \sum_{t \in T} I(S,t)$



0	0	1	1	0
0	0	1	0	1
0	1	0	0	0
1	1	0	1	1

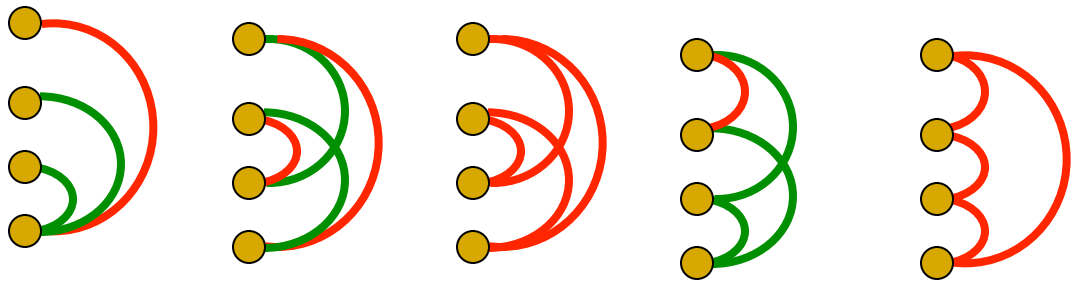
$I(s_1,s_2) = 2/4 = 1/2$

s_1 s_2 s_3 s_4 s_5

Informativeness

For each SNP define a bipartite graph, having the set of haplotypes as nodes and an edge exists between two haplotypes when the two alleles at s are different. Let $E(s)$ be the set of edges. We estimate I from the sample as follows.

- $I(S,t) = \frac{|E(s) \cap E(t)|}{E(t)}$
- $I(S,t) = \frac{|(\cup_{s \in S} E(s)) \cap E(t)|}{E(t)}$
- $I(S,T) = \sum_{t \in T} I(S,t)$

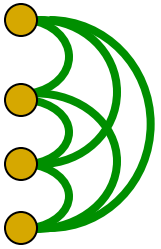


0	0	1	1	0
0	0	1	0	1
0	1	0	0	0
1	1	0	1	1

$I(\{s_1, s_2\}, s_4) = 3/4$

s_1 s_2 s_3 s_4 s_5

Informativeness



0	0	1	1	0
0	0	1	0	1
0	1	0	0	0
1	1	0	1	1

$I(\{s_3,s_4\},\{s_1,s_2,s_5\}) = 3$

$S=\{s_3,s_4\}$ is a
Minimal Informative Subset

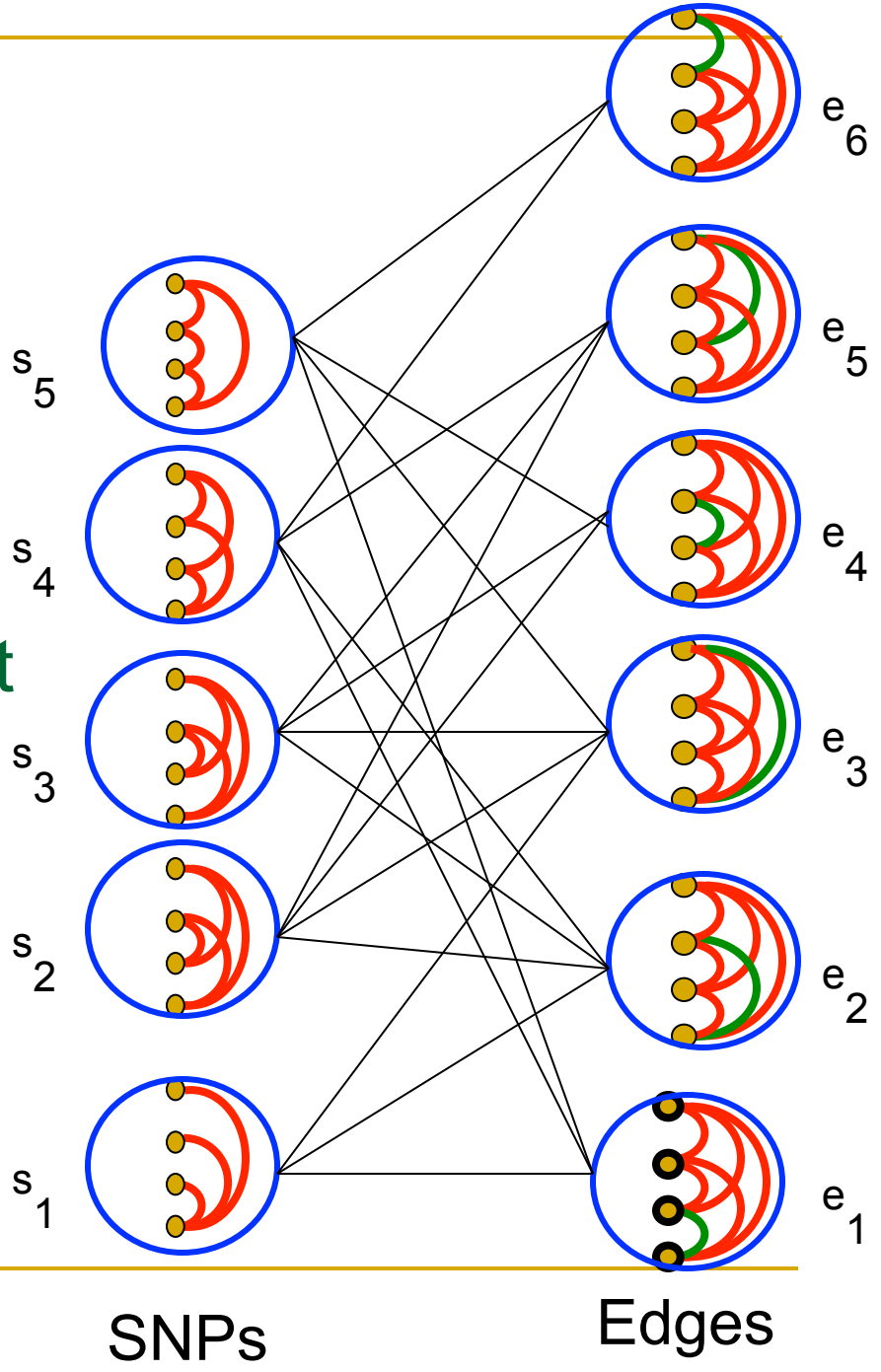
s_1 s_2 s_3 s_4 s_5

Informativeness

Graph theory insight

Minimum Set Cover
=
Minimum Informative Subset

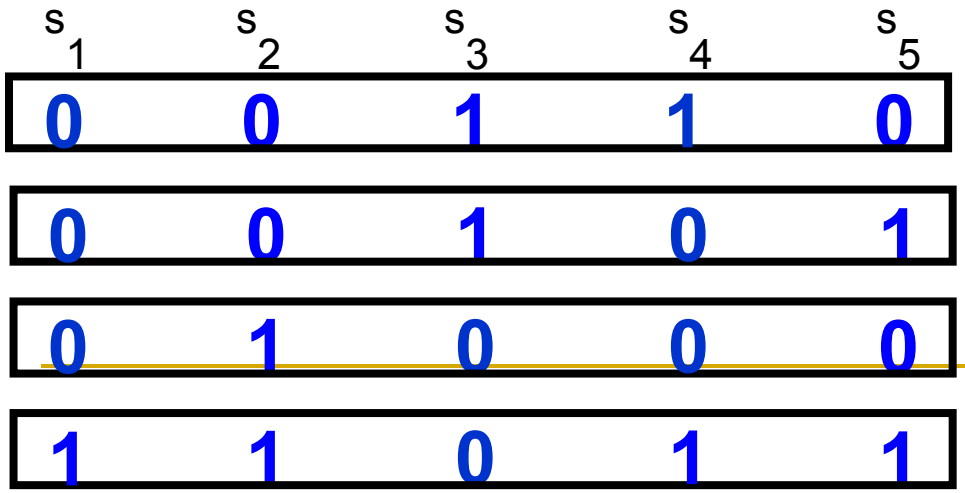
s_1	s_2	s_3	s_4	s_5
0	0	1	1	0
0	0	1	0	1
0	1	0	0	0
1	1	0	1	1

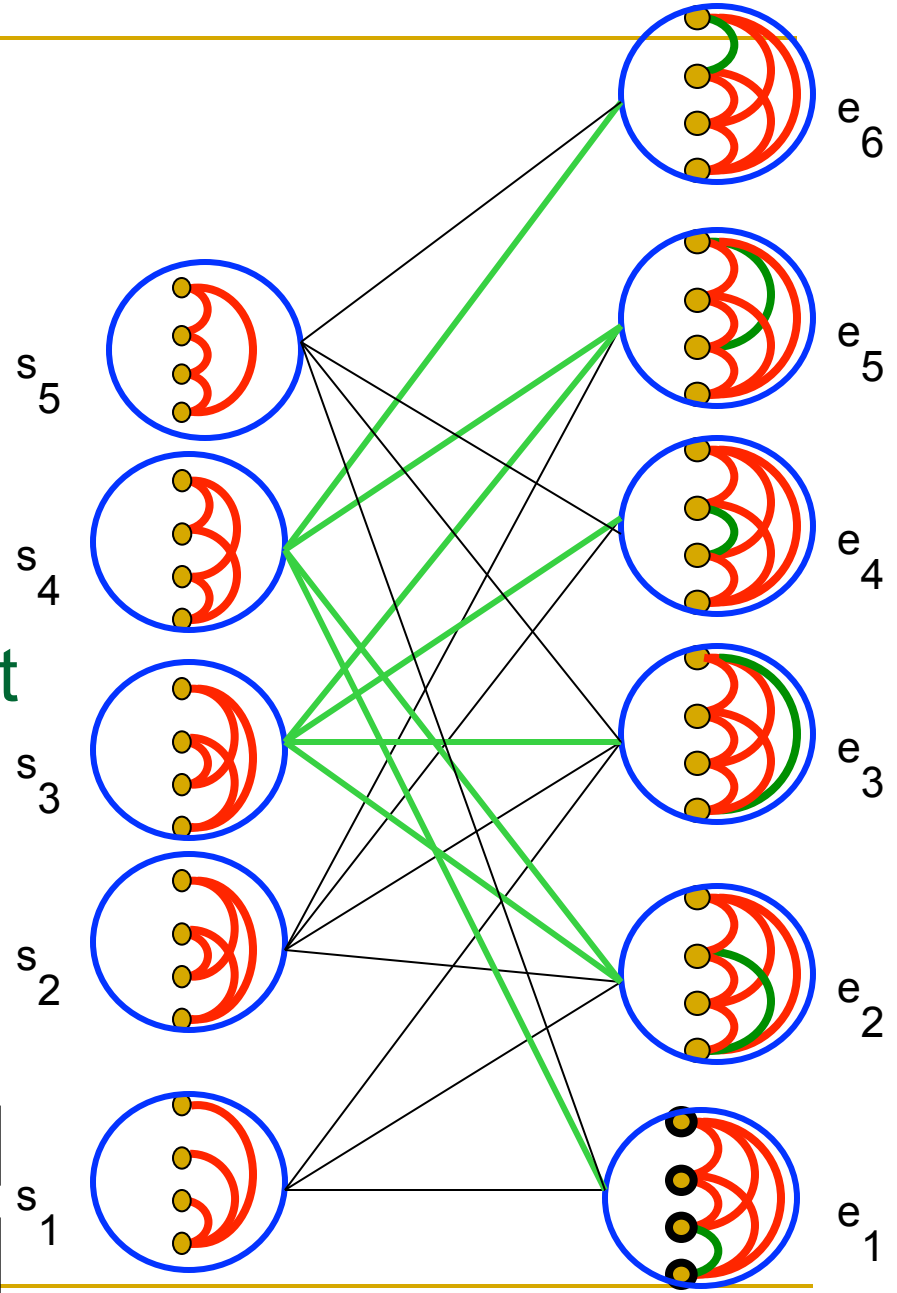


Informativeness

Graph theory insight

Minimum Set Cover $\{s_3, s_4\}$
=
Minimum Informative Subset





Connecting Informativeness with Measures of LD

The most basic population genetics measures for linkage at two loci are $| D' |$ and r^2. And a recent important addition is the measure d^2 of Devlin and Risch 1995, Kruglyak 1999.

- $I(s,t) = 1$ when $| D' | = 1$ and $r^2 = 1$ and $I(s,t) = 0$ when $| D' | = 0$ and $r^2 = 0$.

- $I(s,t) = 1$ when $d^2 = 1$ and $I(s,t) = 0$ when $d^2 = 0$

All these pairwise LD measures present analytical difficulties when they are extended to sets of SNPs.

The Minimum Informative SNPs in a Block of Complete LD

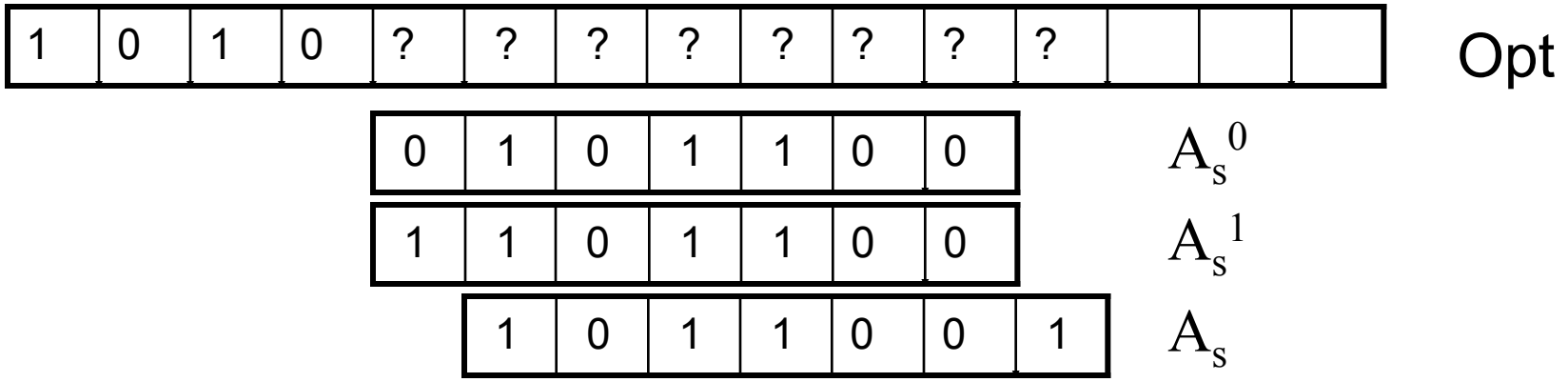
Input: A set m haplotypes, a set of n SNPs S that define a complete LD block, a subset $T \subset S$, and $0 < k \leq n$

Output: Doeas there exist a subset $S' \subset S - T$ such that $I(S', T) = \mid T \mid$.

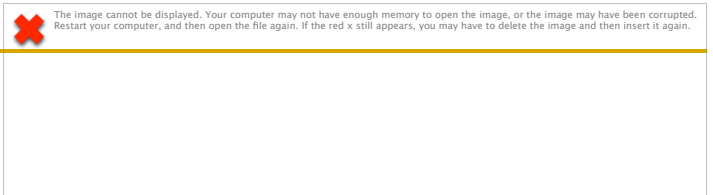
The Minimum Informative SNPs in a Block of Complete LD Problem can be solved exactly in time $O(mn)$ when $\mid T \mid = 1$.

(k,w)-MIS Problem

(k,w)-MIS: O(nk2^w) solution



For s **from** 1 to n
For l **from** 1 to k
Forall A_s
$A_s^0 = 0A_s[1..w-1]$
$A_s^1 = 1A_s[1..w-1]$
$I(s,l,A_s) = I(S(A_s),s) +$
$max(I(s-1,l-A_s[w],A_s^0), I(s-1,l-A_s[w],A_s^1))$



Validation
Tests on Publicly-Accessible Data

- We performed tests using two publicly available datasets:

  LPL dataset of Nickerson et al. (2000):

  142 chromosomes typed at 88 SNPs

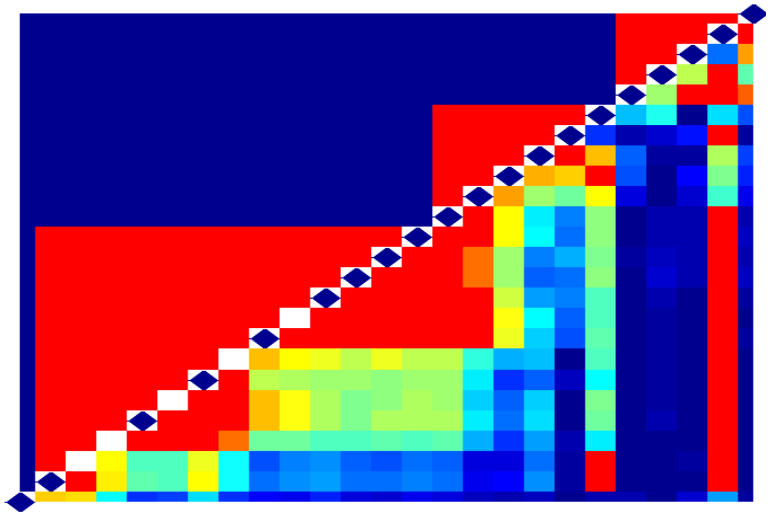
  Chromosome 21 dataset of Patil et al. (2001):

  20 chromosomes typed at 24,047 SNPs

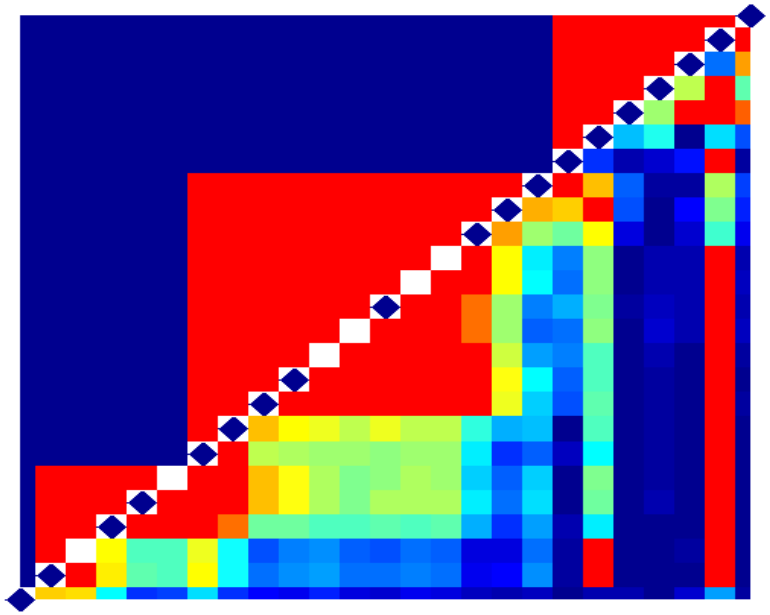
- We also performed tests on an AB dataset

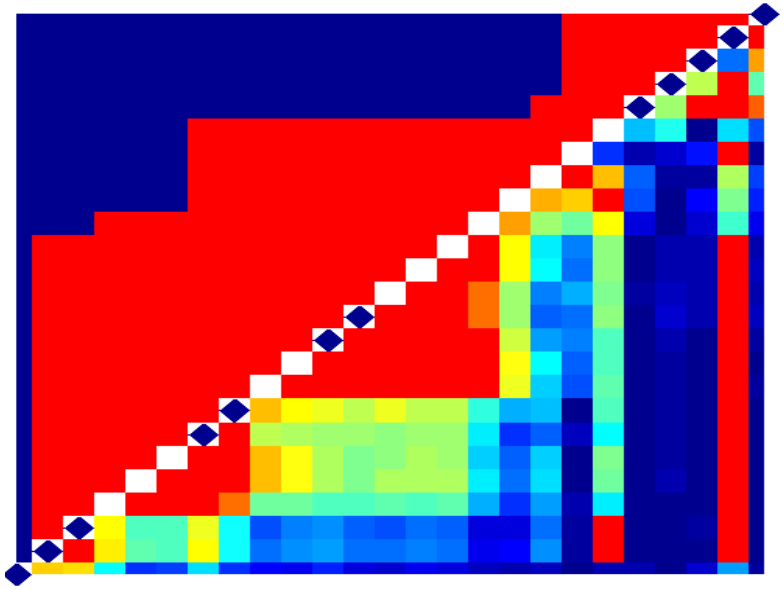
  Most of Chromosome 22

  45 chromosomes typed at 4102 SNPs



A region of Chr. 22
45 Caucasian samples



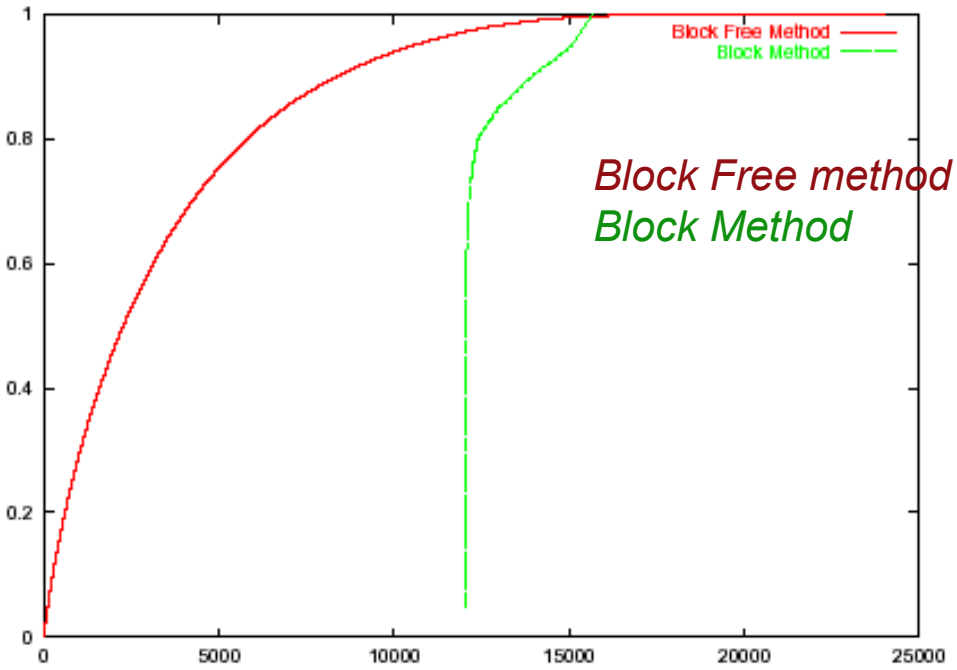


Two different runs of the Gabriel el al Block Detection method + Zhang et al SNP selection algorithm

Our block-free algorithm

Block free tagging
Minimum informative SNPs

Informativeness

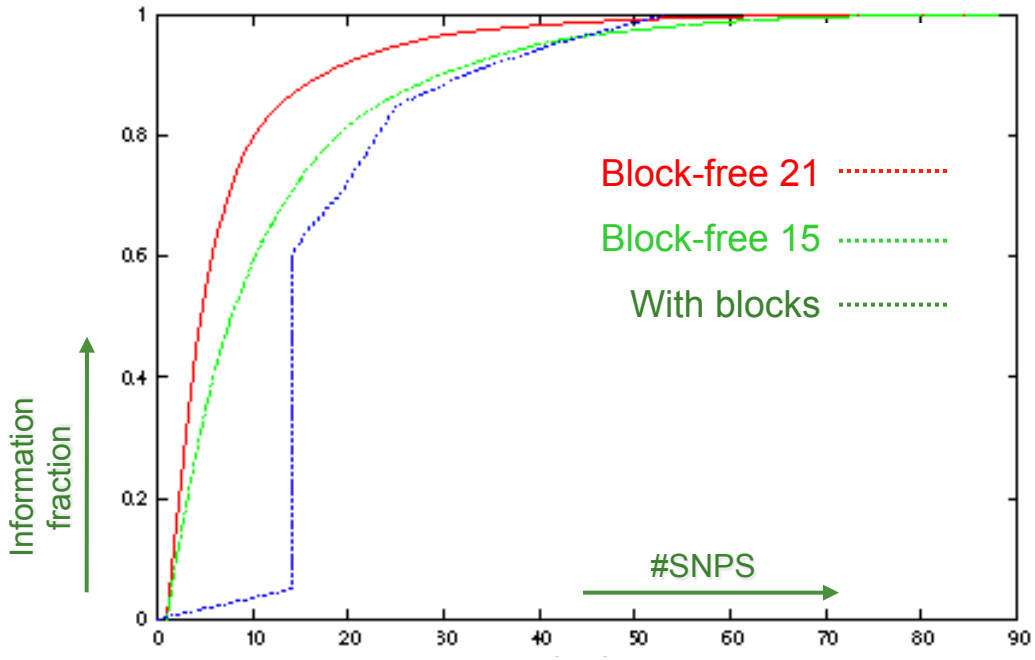


Block Free method
Block Method

Number of SNPs

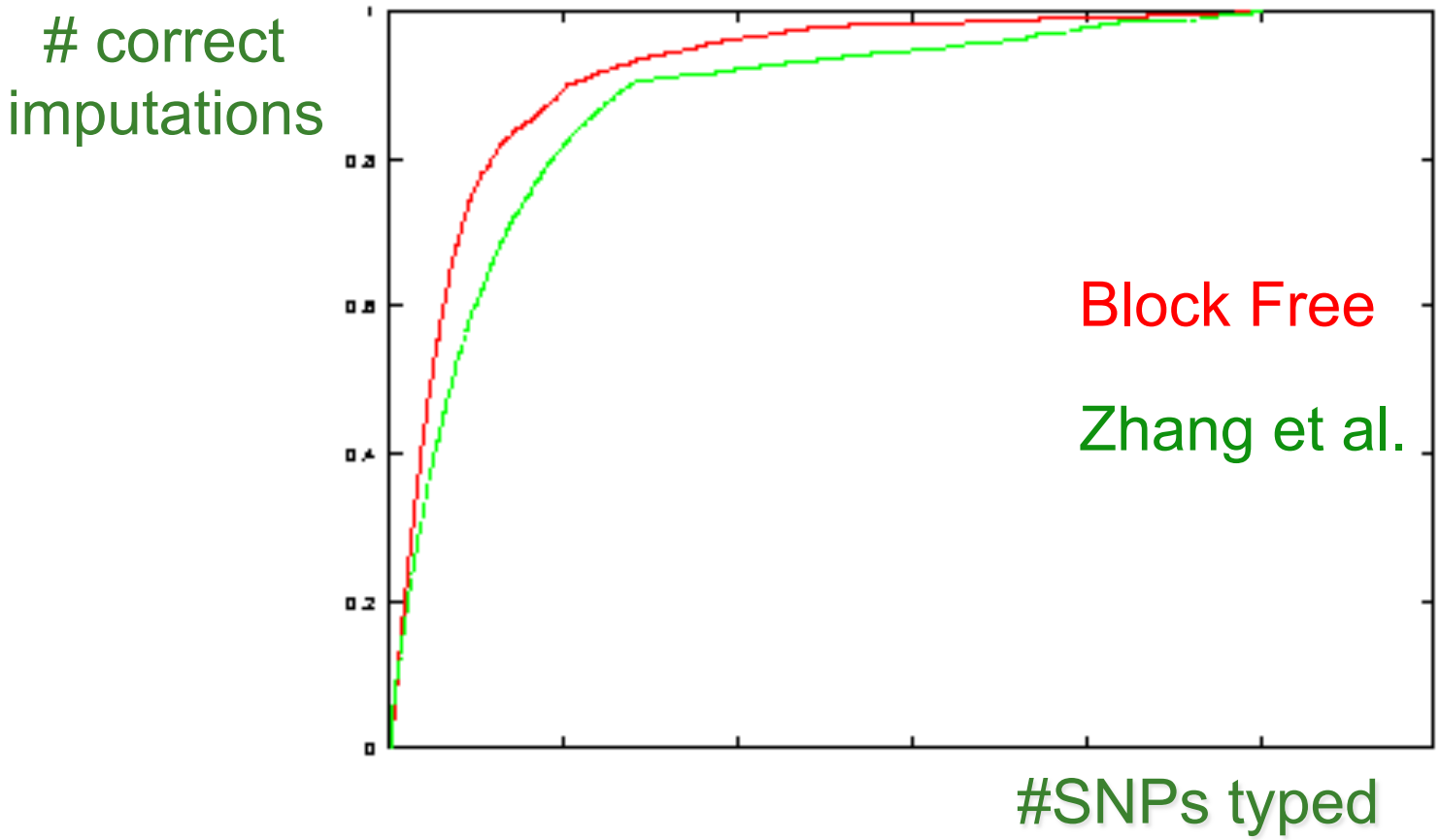
Perlegen Data Set Chromosome 21:
20 individuals, 24047 SNPs

Block free tagging
Minimum informative SNPs



Lipoprotein Lipase Gene, 71 individuals, 88 SNPs

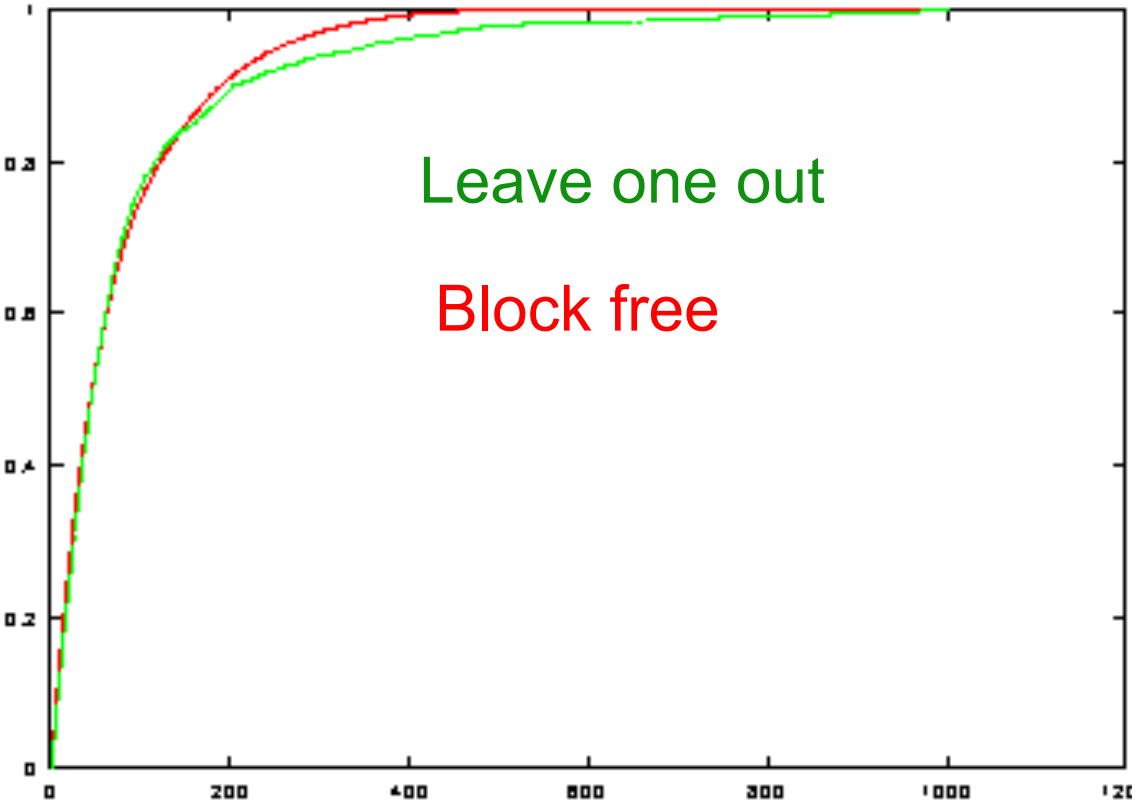
Correct imputation block vs. block free



Perlegen dataset

Correlations of informativeness with imputation in leave one out studies

Informativeness



#SNPs

Perlege dataset

Conclusions

Conclusions

- Existing LD based measures are not adequate for SNP subset selection, and do not extend easily to multiple SNPs
- The **Informativeness** measure for SNPs is Block-free, and extends easily to multiple SNPs.
- Practically feasible algorithms for genome-wide studies to compute minimum informative SNP subsets
- We are able to show that by typing only 20-30% of the SNPs, we are able to retain 90% of the informativeness.