A New Model of Multi-Marker Correlation for Genome-Wide Tag SNP Selection

Wei-Bung Wang
Tao Jiang

Outline

- Introduction
- Problem
- Related Work
- Our Approach
- Result

Single Nucleotide Polymorphism

- Single Nucleotide Polymorphism (SNP)
  - A genetic variation

94% → C T T A G **C** T T

6% → C T T A G **T** T T

SNP

Modified from slide by Yao-Ting Huang,
National Taiwan University
Department of Computer Science and Information Engineering

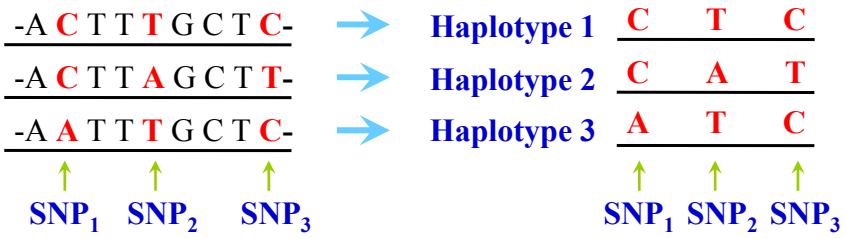
SNPs

- SNPs are usually bi-allelic
  - Major allele 94% → C T T A G **C** T T
  - Minor allele 6% → C T T A G **T** T T

SNP

- Minor allele frequency > 1% (or 5%)
- Tri-allelic: very rare

Haplotype



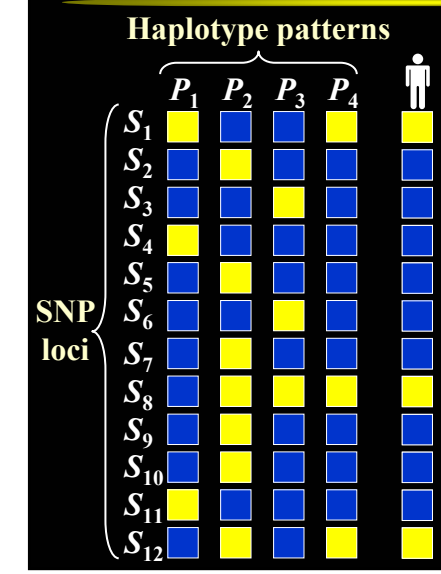
Modified from slide by Yao-Ting Huang,
National Taiwan University
Department of Computer Science and Information Engineering

Tag SNP

- What is a tag SNP?
- Here I use some slides by Yao-Ting Huang and Kun-Mao Chao

National Taiwan University
Department of Computer Science and Information Engineering

Examples of Tag SNPs



← An unknown haplotype sample

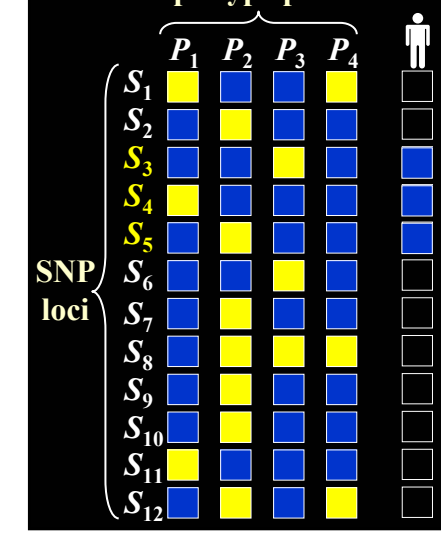
- Suppose we wish to distinguish an unknown haplotype sample.
- We can genotype all SNPs to identify the haplotype sample.

■ : Major allele

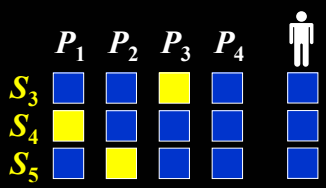
■ : Minor allele

National Taiwan University
Department of Computer Science and Information Engineering

Examples of Tag SNPs

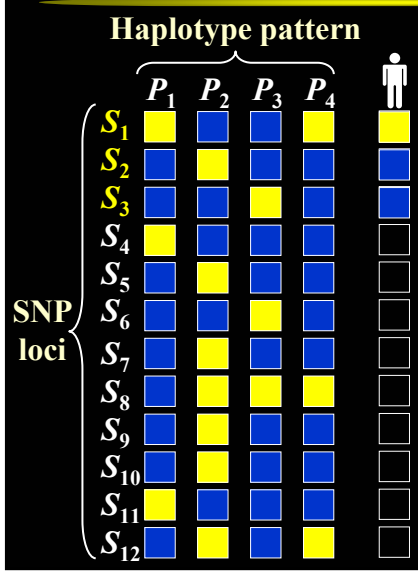


- In fact, it is not necessary to genotype all SNPs.
- SNPs S_3, S_4, and S_5 can form a set of tag SNPs.

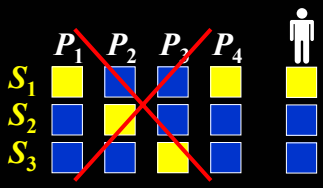


National Taiwan University
Department of Computer Science and Information Engineering

Examples of Wrong Tag SNPs

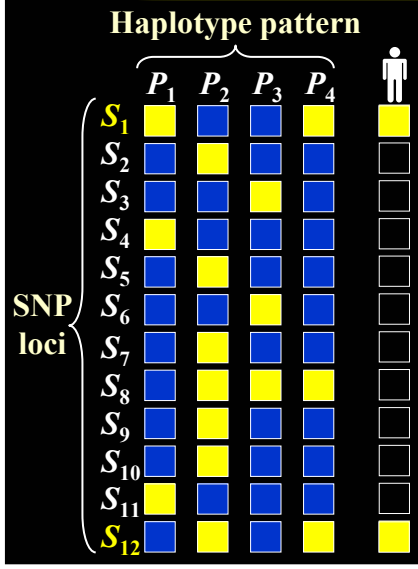


- SNPs S_1, S_2, and S_3 can not form a set of tag SNPs because P_1 and P_4 will be ambiguous.

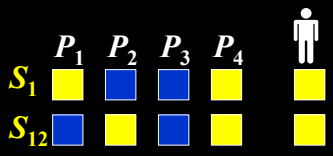


National Taiwan University
Department of Computer Science and Information Engineering

Examples of Tag SNPs



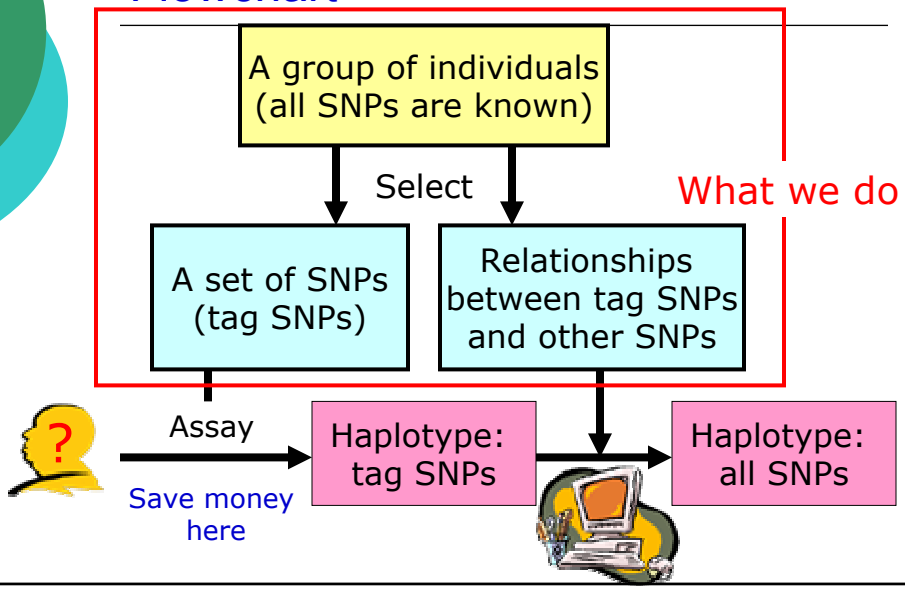
- SNPs S_1 and S_{12} can form a set of tag SNPs.
- This set of SNPs is the minimum solution in this example.



Problem

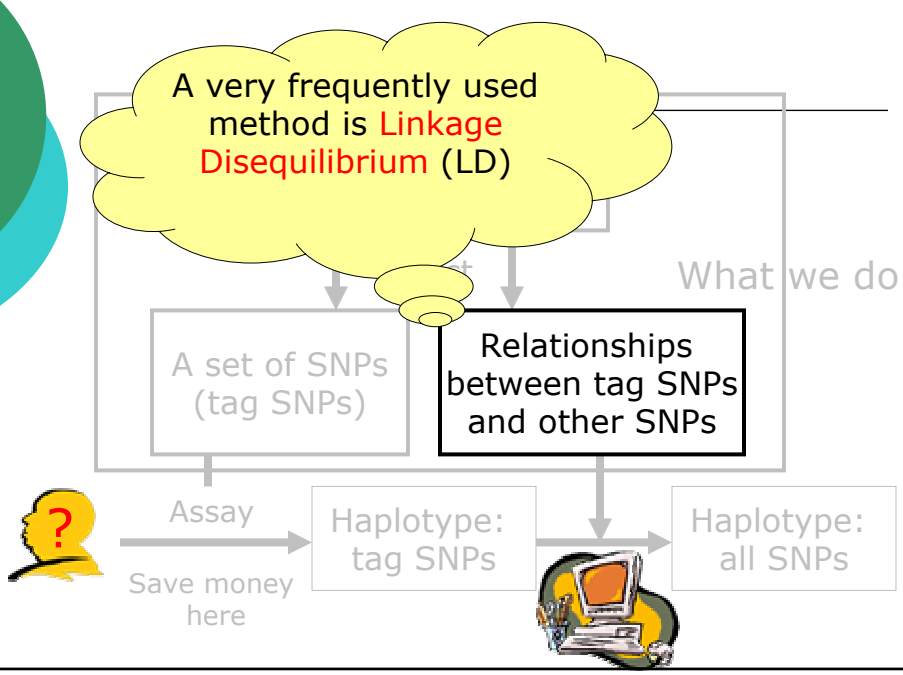
- Tag SNP selection
  - How to select representatives?
  - Many different ways

Flowchart



Problem

- Perfect world
  - Minimum set of tag SNPs
  - Save most money
  - NP-hard
- Real life
  - Relatively small set
  - Sufficient accuracy/confidence



Linkage Disequilibrium (LD)

- Non-random association of alleles at two or more loci
- Correlated coefficient: estimation of dependency
- LD = correlated coefficient = r^2

Linkage Disequilibrium

$$r^2 = \frac{(P_{AB} - P_A P_B)^2}{P_A P_a P_B P_b}$$

- r^2 2 [0; 1]
- $r^2 = 1$: perfect correlation (A , B ; a , b)

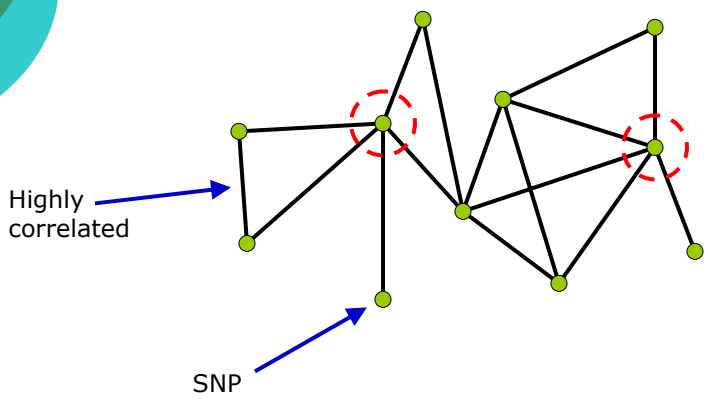
  $r^2 = 1$, $P_{AB} = P_A = P_B$

- $r^2 = 0.9$: strong correlation (0.95, etc.)
- $r^2 = 0$: no correlation

An Example

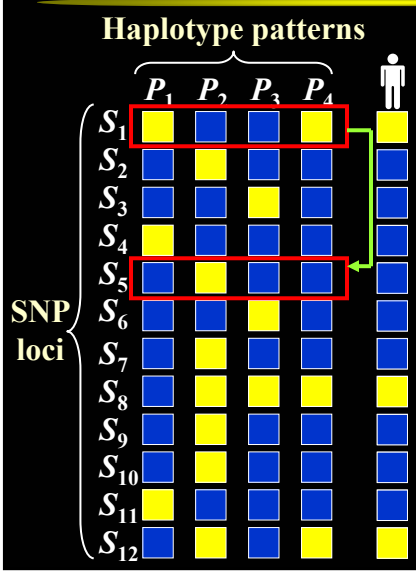
Individual	SNP_1	SNP_2	SNP_3	SNP_4	SNP_5	SNP_6
1	A	G	A	C	G	T
2	T	G	C	C	G	C
3	A	A	A	T	A	T
4	T	G	C	T	A	C
5	T	A	C	C	G	C
	A T		A C			T C

Minimum Dominating Set Problem



National Taiwan University
Department of Computer Science and Information Engineering

Examples of Tag SNPs



← An unknown haplotype sample

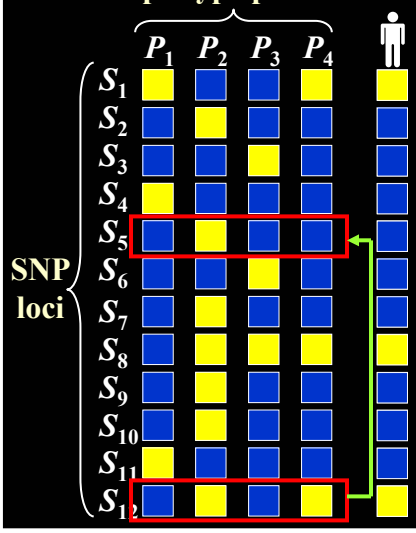
- Suppose we wish to distinguish an unknown haplotype sample.

: Major allele

: Minor allele

National Taiwan University
Department of Computer Science and Information Engineering

Examples of Tag SNPs



← An unknown haplotype sample

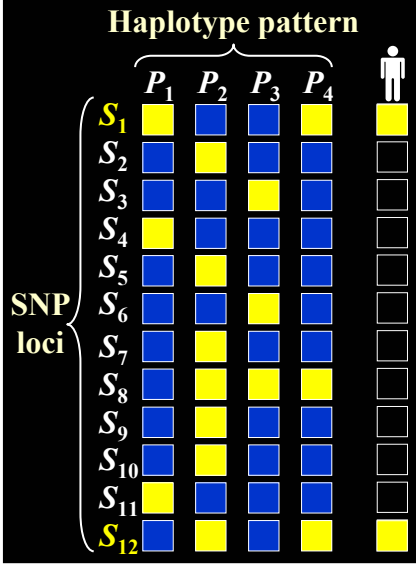
- Suppose we wish to distinguish an unknown haplotype sample.

: Major allele

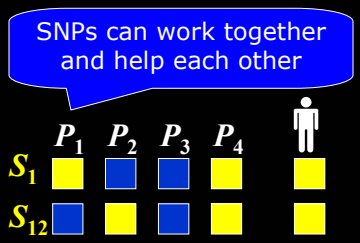
: Minor allele

National Taiwan University
Department of Computer Science and Information Engineering

Examples of Tag SNPs



- SNPs S_1 and S_{12} can form a set of tag SNPs.
- This set of SNPs is the minimum solution in this example.



Our Approach

Our Approach

	snp_1	snp_2	snp_3
$haplotype_1$	A	C	G
$haplotype_2$	A	T	T
$haplotype_3$	A	C	G
$haplotype_4$	A	T	T
$haplotype_5$	C	T	T
$haplotype_6$	A	T	T
$haplotype_7$	C	T	G
$haplotype_8$	C	C	T
$haplotype_9$	C	C	T
$haplotype_{10}$	C	T	T
	A else	C	G T

- We introduce new allele AC and ¬AC

	: AC	AC	
T	0.7	0	0.7
G	0.1	0.2	0.3
	0.8	0.2	

- Only one mistake

Our Approach

	snp_1	snp_2	snp_3	snp_4
$haplotype_1$	A	C	G	A
$haplotype_2$	A	T	T	C
$haplotype_3$	A	C	G	A
$haplotype_4$	A	T	T	C
$haplotype_5$	C	T	T	A
$haplotype_6$	A	T	T	C
$haplotype_7$	C	T	G	A
$haplotype_8$	C	C	T	C
$haplotype_9$	C	C	T	C
$haplotype_{10}$	C	T	T	A

- (snp_1, snp_2) vs. snp_3

  AC , G

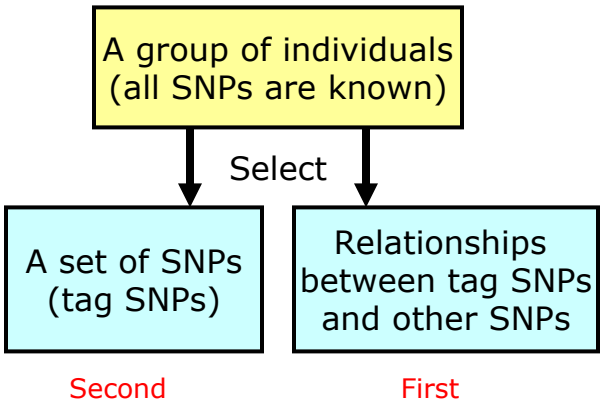
  : AC , T

- (snp_1, snp_2) vs. snp_4

  (AC _ CT) , A

  : (AC _ CT) , C

In the Right Order



Our Approach

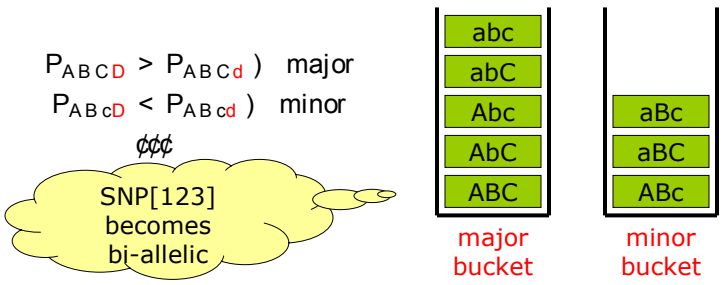
- Generate relationships

If SNP 1, 4, 10 are tag SNPs Predict SNP 17 with patterns … Accuracy / LD: 0.97
If SNP 5, 8, 13 are tag SNPs Predict SNP 11 with patterns … Accuracy / LD: 0.62

.....

How to Predict / Determine the Alleles?

- LD: (tag) SNP 1, 2, 3 vs. SNP 4
- Allele A/a, B/b, C/c, D/d



(ABC _ AbC _ Abc _ abC _ abc) = M) D
(ABc _ aBC _ aBc) = m) d

Similar Work

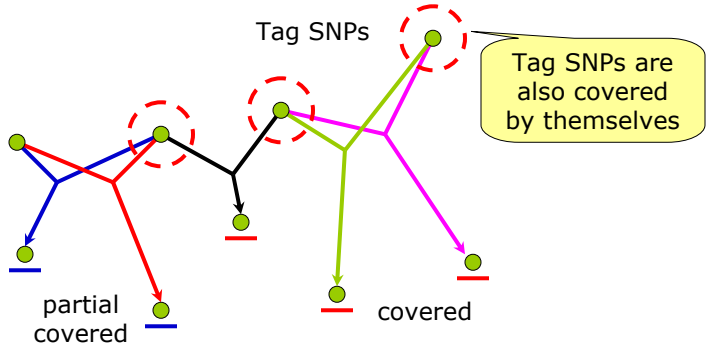
- Ke Hao also did a similar work
  - The same LD model
  - Different way to determine alleles for composite SNPs
    - Less flexibility
    - A special case of our model
  - Related paper: "Genome-wide selection of tag SNPs using multiple-marker correlation," Bioinformatics, 2007

Sketch

- Get r^2 value for all possible combinations
- Find a small subset of SNPs according to LD

Sketch

- Find a small subset of SNPs according to LD

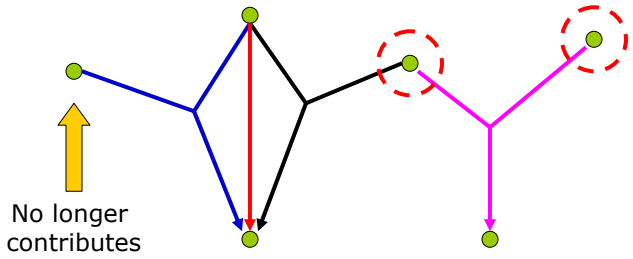


Our Approach

Sketch

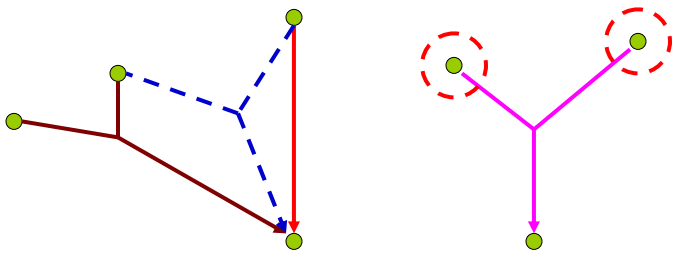
- Simple greedy algorithm (Ke Hao)
  - Cover more SNPs in each iteration
- Modified greedy algorithm (my work)
  - A SNP that can't be covered by others
    - High priority
  - A SNP that is not picked but covered
    - OK
  - Break tie: partial cover

Supersede



Our Approach

Supersede



Our Approach

My Program: MMTagger

Algorithm 1 Two-Marker MMTagger

Require: set of triplets

```
1:  while there are SNPs uncovered do
2:      if there is a SNP s with no incoming edge
3:          s¤ Ã s
4:      else
5:          s¤ Ã the SNP that covers the most uncovered SNPs
6:      for each triplet t of form (s_i ; s_j B s¤) do
7:          remove t and its corresponding edges
8:      Put s¤ into tag SNP set                    /* s¤ is \picked"   */
9:      for each triplet t of form (s¤; s_i B s_j) or (s_i ; s¤ B s_j) do
10:         if s_i is picked then
11:             put s_j into covered SNP set
12:             remove t and its corresponding edg
13:         else
14:             remove all triplets of form (s_i ; s'B s_j) or (s'; s_i B s_j)
```

Pick a SNP

Data structure

Our Approach

Complexity

- Computing r^2 value
  - $O(n^{k+1})$ for k-marker
- Picking tag SNPs
  - - (T) where T is the number of relationships
  - $O(T \log T)$ time algorithm

Result

- Our program: MMTagger
- Vs. Single-marker approach (LRTag)
  - A state-of-the-art program
  - Single-marker
- Vs. Hao's program (MultiTag)
  - Multi-marker

Vs. Single-Marker Approach

Region	ENm010	ENm013	ENm014	ENr112	ENr113
# SNP	459	731	874	868	1035
$r^2 \geq 0.8$					
LRTag	119	88	134	148	133
2-marker MultiTag	75	57	80	87	75
2-marker MMTagger	72	52	78	85	73
3-marker MultiTag	68	53	75	78	64
3-marker MMTagger	62	48	75	68	59
$r^2 \geq 0.9$					
LRTag	148	121	172	204	190
2-marker MultiTag	100	76	111	118	122
2-marker MMTagger	92	73	100	109	115
3-marker MultiTag	91	66	102	101	100
3-marker MMTagger	79	58	85	81	81
$r^2 \geq 0.95$					
LRTag	192	148	196	268	247
2-marker MultiTag	127	96	131	157	156
2-marker MMTagger	117	92	122	141	149
3-marker MultiTag	120	83	119	138	145
3-marker MMTagger	97	66	102	107	112

Result

MMTagger Vs. MultiTag

Table 2. MMTagger vs. MultiTag

Chromosome	# SNP	mode	r^2	program	# SNPs Selected	Time (hours)	Memory (M bytes)
JPT+CHB chr19	28931	2-marker	0.9	MultiTag	9600	26hrs	30–35
				MMTagger	9145	2mins	125
		3-marker	0.95	MultiTag	N/A	>700hrs	30–35
				MMTagger	10032	<1hr	657
JPT+CHB chr21	28914	2-marker	0.9	MultiTag	7115	42hrs	30–35
				MMTagger	6766	2mins	187
		3-marker	0.95	MultiTag	N/A	>700hrs	30–35
				MMTagger	7404	<1hr	1210
JPT+CHB chr22	26595	2-marker	0.9	MultiTag	7557	93hrs	30–35
				MMTagger	7221	2mins	183
		3-marker	0.95	MultiTag	N/A	>700hrs	30–35
				MMTagger	7788	3hrs	1216

Note: Both programs were run on a desktop PC with dual AMD Athlon(tm) processors of 2.1 GHz.

Conclusion

- We provide a new multi-marker model
  - Size of tag SNP set
    - 2- vs. 1-marker: apparently better
    - 3- vs. 2-marker: slightly better
    - 4-marker or more: slow, unacceptable
- Performance
  - Our program outperforms the only other program with similar model

Thank you!