Tag SNP selection in genetic association studies

Yuan Lin
Post-doc Research Associate
@ Dr. Kirk Wilhelmsen’s Lab
Department of Genetics, School of Medicine
UNC at Chapel Hill

Power & efficiency of association studies

- Statistical power of association studies increases with the number of individuals and the density of SNPs being genotyped.
- Genotyping cost (efficiency) is affected by the overall number of genotyped SNPs.
- Select a minimal subset of markers (tag SNPs) that predict remaining SNPs (target SNPs) with high accuracy.

“Predict a SNP”

Hap1	A	G	T	A
Hap2	A	C	A	C
SNP #	1	2	3	4

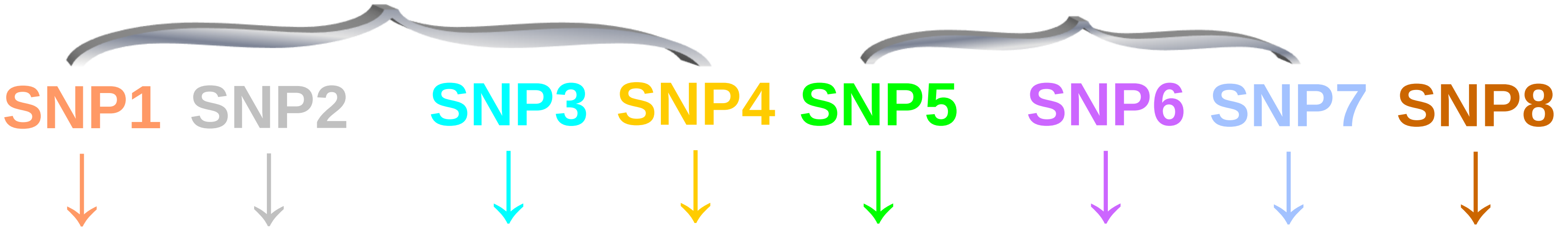
SNP 2 can predict SNP 3

SNP 3 can predict SNP 2

SNP 3 can predict SNP 4

Hap1	G	T	A	G
Hap2	C	T	A	T
Hap3	G	G	T	T
SNP #	1	2	3	4

SNPs 1 and 3 together predict SNP 4



CAGATCGCTGGATGAATCGCATCTGTAAGCAT

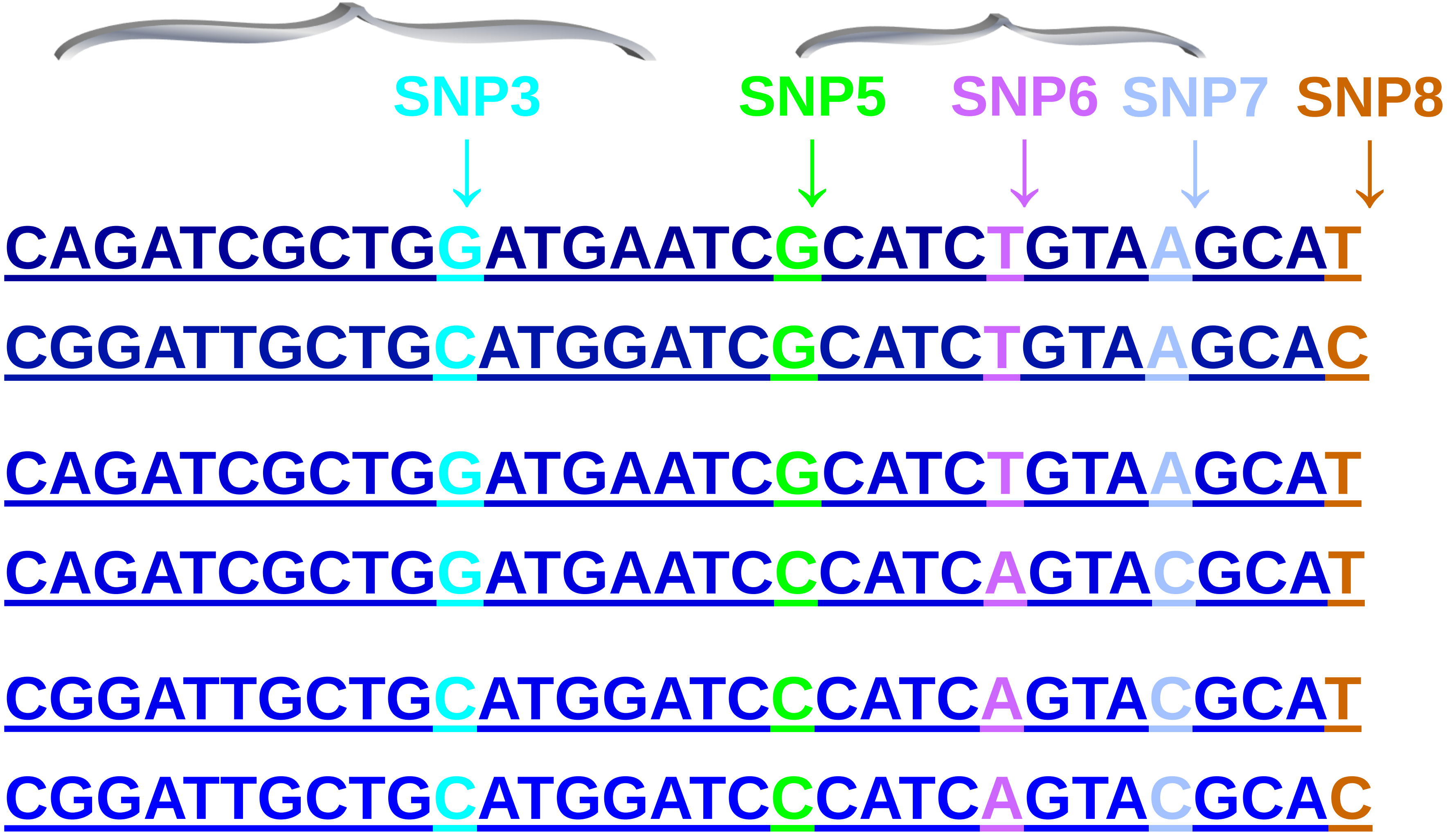
CGGATTGCTGCATGGATCGCATCTGTAAGCAC

CAGATCGCTGGATGAATCGCATCTGTAAGCAT

CAGATCGCTGGATGAATCCCATCAGTACGCAT

CGGATTGCTGCATGGATCCCATCAGTACGCAT

CGGATTGCTGCATGGATCCCATCAGTACGCAC



SNP3 SNP5 SNP6 SNP7 SNP8

CAGATCGCTGGATGAATCGCATCTGTAAGCAT

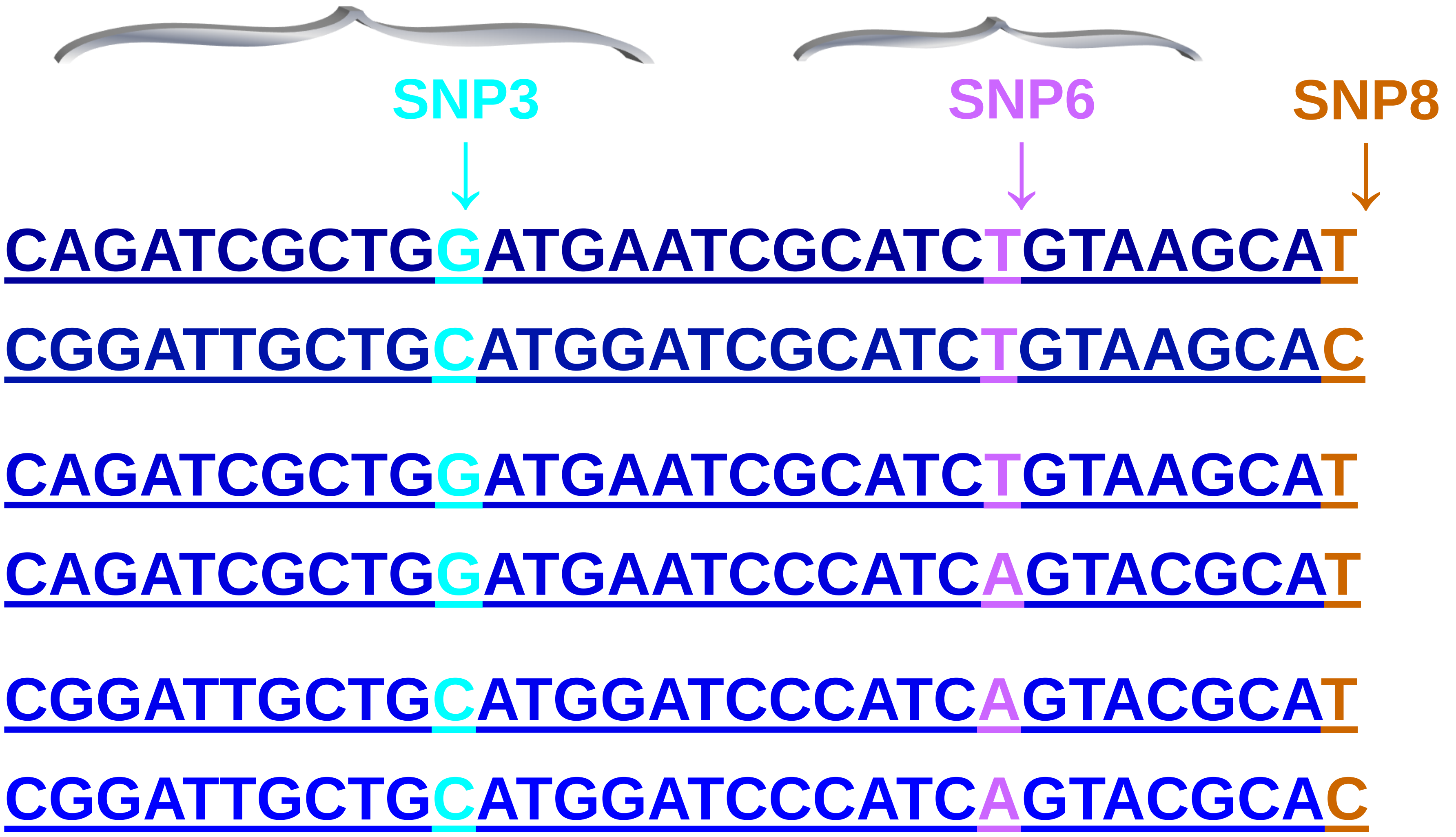
CGGATTGCTGCATGGATCGCATCTGTAAGCAC

CAGATCGCTGGATGAATCGCATCTGTAAGCAT

CAGATCGCTGGATGAATCCCATCAGTACGCAT

CGGATTGCTGCATGGATCCCATCAGTACGCAT

CGGATTGCTGCATGGATCCCATCAGTACGCAC



SNP3 SNP6 SNP8

CAGATCGCTGGATGAATCGCATCTGTAAGCAT

CGGATTGCTGCATGGATCGCATCTGTAAGCAC

CAGATCGCTGGATGAATCGCATCTGTAAGCAT

CAGATCGCTGGATGAATCCCATCAGTACGCAT

CGGATTGCTGCATGGATCCCATCAGTACGCAT

CGGATTGCTGCATGGATCCCATCAGTACGCAC

GTT	35%
CTC	30%
GTT	10%
GAT	8%
CAT	7%
CAC	6%
other haplotypes	4%

Three SNPs predict 96% different haplotypes

The Tagging problem

- **Given** a sample *S* of genotypes from a population *P*; each sample has *m* SNPs
- **Find** positions of *k* (*k* < *m*) tag SNPs
- **Such that** one can reconstruct genotype *g* on all *m* SNPs in *P* from its restriction *g'* on *k* tag SNPs with certain accuracy

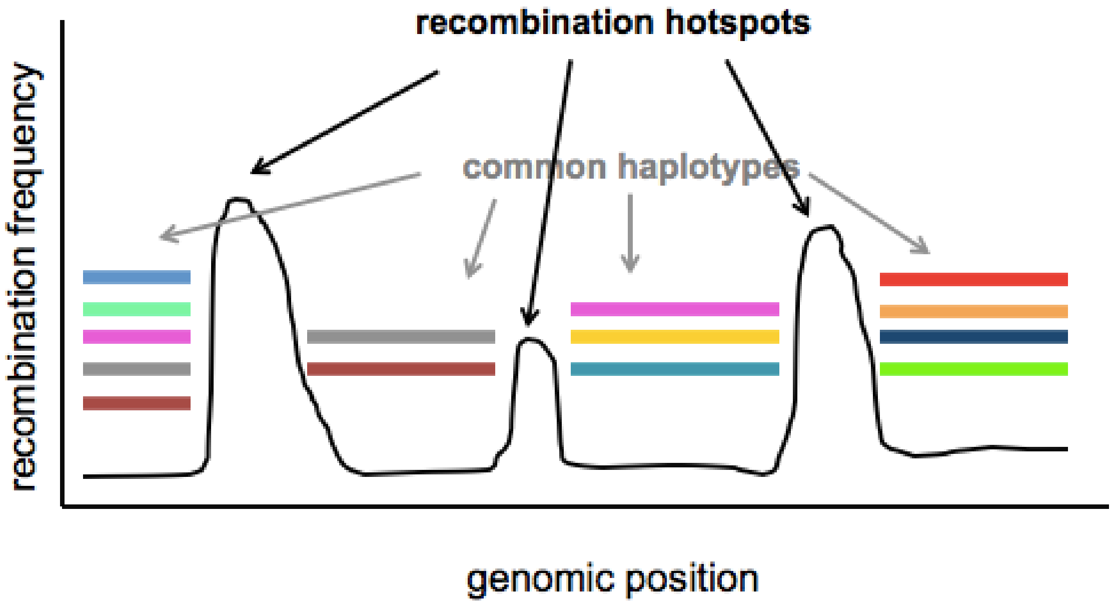
General framework of a tagging method

(Halldorsson et al., 2004)

1. Define a genomic region to search for tag SNPs.
2. Define a quality metric that quantifies how well a set of tag SNPs capture all the variance in the full data set.
3. Design an algorithm that selects a minimal number of tag SNPs that meet a desired quality threshold or optimizes the quality metric (as an objective function).

Define a search region

- Haplotype-block-based vs block-free methods
- Human genome consists of haplotype blocks (Daly et al., 2001; Dawson et al., 2002; Gabriel et al., 2002; Patil et al., 2001; Wall & Pritchard 2003).



Block-based tagging

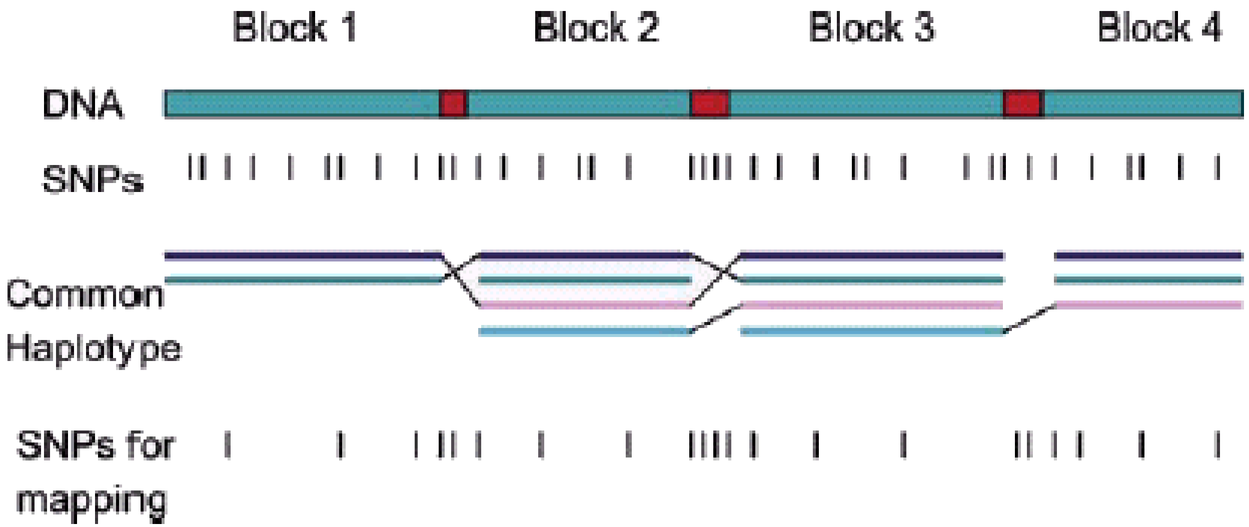
- Find a small set of SNPs in each block that captures the majority of SNP variation and identity common haplotypes in that block.
- But what exactly is a haplotype block?
  - High LD inside
  - Low haplotype diversity
  - Little recombination

  …

No consensus on a practical definition

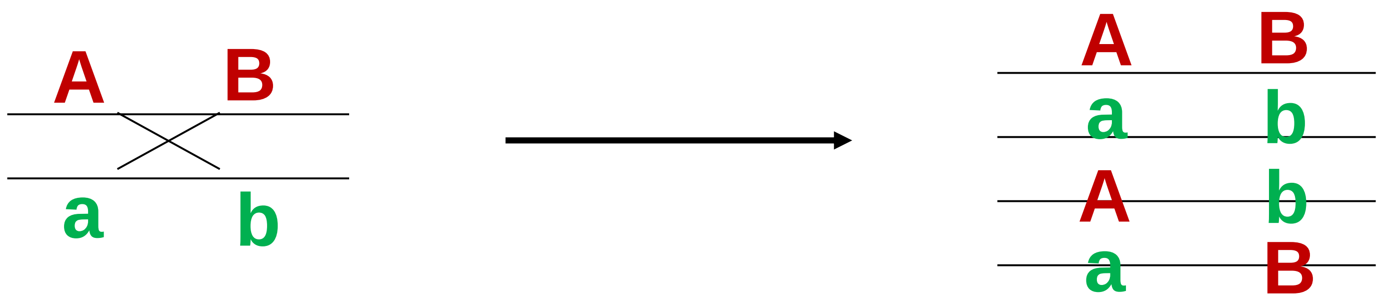
Block: Low haplotype diversity

- Patil et al., 2001
  - – In each block, at least a certain proportion of observed or inferred haplotypes should be common haplotypes.



Block: no historical recombination

- Wang et al. 2002
  - A set of consecutive SNPs form a block if there is no historical recombination events (based on the four-gamete test)



< 4 haplotypes, D´=1 → block

4 haplotypes, D´<1 → boundary

Block: strong pairwise LD inside

- Gabriel et al. 2002
  - Blocks are partitioned based on whether the upper and lower confidence bounds on pairwise *D*’ meet certain thresholds.
  - Specifically, the proportion of SNP pairs with strong LD (upper confidence bound > .98 and lower bound > .7) must account for at least 95% of all SNP pairs

Problems

- Block boundaries are ambiguous: they are sensitive to block definition and marker density (Boundary SNPs are often SNPs within recombination hotspots; until today they are not well tagged. Fine mapping is often needed.)
- Haplotype blocks are assumed to be independent, but adjacent blocks can still have substantial correlation.
- Not all genomic regions fit the haplotype-block model (Wall and Pritchard 2003).

Block-free Tagging

- Search for tag SNPs in a predefined neighborhood of each target SNP
- It is non-trivial to define the neighborhood (a sliding window).
  - There is usually an upper bound on the distance between a tag SNP and a target SNP (i.e., the maximal size of the window)
  - A small fixed window size (Meng et al., 2003)
  - A dynamically adjusted window size based on local LD extent (Halldorsson et al., 2004)

Define a quality metric

- Pairwise vs multivariate metrics
- LD measures (e.g., D', r^2)
  - – Select tags until a r^2 threshold (often > 0.8) is exceeded for every pair of target and tag SNPs (Carlson et al., 2004; Zhang and Jin, 2003)
  - – Select the "best N" tags by the number of target SNPs they can surrogate at a given r^2 (de Bakker et al., 2005)
  - – The power to directly detect a causal SNP in Nr^2 samples is equivalent to the power to detect it indirectly (via markers) in N samples (Pritchard & Przeworski 2001).

Define a quality metric (cont.)

- Haplotype R^2 (Stram et al., 2003; Weale et al., 2003)
  - Extension of r^2 to Haplotypes
  - R_h^2 stands for the correlation between the frequency of haplotype *h* inferred from tag SNPs and all SNPs
- Statistical power (Genin 2001; Hu et al., 2004)
  - Assume, one at a time, that every SNP could be the disease mutation, which is unknown, and calculate pairwise power between the putative causal SNP and other SNPs
- Classic multivariate statistics used in PCA (Meng et al., 2003; Lin & Altman 2004), clustering (Ao 2005), or regression (He 2006)

Define a quality metric (cont.)

- Haplotype diversity
  - Coverage of common haplotypes (Patil et al, 2001; Zhang et al., 2002)
  - Coverage of overall haplotype diversity (Johnson et al., 2001)
  - “Informativeness” (Halldorsson et al., 2004)
  - Entropy (Hampe et al., 2003; Zhang et al., 2005)
    - If there are *n* haplotypes and the frequency of haplotype *i* is denoted by p_i , then the entropy of these haplotypes is defined as $S = -\sum_{i=1}^{n} p_i \log p_i$

Problems

- Not all the metrics have clear implications on the power-efficiency trade-off of association studies.
- Using pairwise metrics tend to overestimate the required number of tag SNPs
- Using multivariate metrics must deal with the fact that haplotypes are often unknown and need to be inferred.
- These metrics are based on one SNP or one block. The values need to be appropriately combined for genome-wide SNP selection.

Design an optimizing algorithm

- Computing the optimal solution to selecting the most informative SNPs is generally NP-hard (Bafna et al, 2003).
- Existing tagging methods use greedy (Carlson et al., 2004) or brand-and-bound (Avi-Itzhak et al., 2003).
- Dynamic programming is also applied (Zhang et al., 2002, 2003, 2004; Halldorsson et al., 2004).

Comparison of tagging methods

- Pairwise vs multivariate metrics
  - Multi-marker tagging tends to have fewer tags but more missed signals
- There is a lack of consistency across SNP sets selected by different methods, whether or not LD was present (Ding & Kullo, 2007; Goode et al., 2007).
- Quality metrics may not be as important to performance as optimizing algorithms.

Problems

- SNPs that are rare or have low r^2 with others are poorly tagged.
- Tagging loses its cost-saving advantage in regions of low LD.
- Tagging can be inaccurate when there is population stratification and allele frequencies are significantly different in subpopulations.
- Controversy exists over the extent to which tag SNPs (and GWAS) can help explore untyped structural polymorphism.
- Are these problems caused by tagging methods' dependency on LD? What other information can we to find out the correlation of SNPs? What about genealogy? Can we find a set of tag SNPs such that a coalescent model can be as well simulated by these SNPs alone as by all SNPs?

References

- N. Patil et al., (2001), Blocks of Limited Haplotype Diversity Revealed by High-Resolution Scanning of Human Chromosome 21, Science, vol. 294, pp. 1719-1723
- N. Wang et al., (2002), Distribution of Recombination Crossovers and the Origin of Haplotype Blocks: The Interplay of Population History, Recombination and Mutation, Am. J. Hum. Genet., vol. 71, pp. 1227-1234.
- G. C. Johnson, L. Esposito, B. J. Barratt, et al. Haplotype tagging for the identification of common disease genes. *Nat Genet.*, 29(2):233 – 7, Oct 2001.
- Lin Z., Altman R. B. Finding haplotype tagging SNPs by use of principal components analysis. Am J Hum Genet. 2004 Nov;75(5):850-61.
- Hampe J., Schreiber S., Krawczak M. Entropy-based SNP selection for genetic association studies. (2003) Hum Genet 114:36-43.
- Gabriel SB, Schaffner SF, Nguyen H et al: The structure of haplotype blocks in the human genome. Science 2002; 296: 2225– 2229.
- Zhang K, Deng M, Chen T, Waterman MS, Sun F: A dynamic programming algorithm for haplotype block partitioning. PNAS 2002; 99: 7335–7339.
- Zhang K, Qin ZS, Liu JS, Chen T, Waterman MS, Sun F: Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies. Genome Res 2004; 14: 908– 916.
- Bafna V, Halldorsson BV, Schwartz R, Clark AG, Istrail S: Haplotypes and informative SNP selection algorithms: don’t block out information. Proceedings of the 7th Annual International Conference on Research in Computational Molecular Biology 2003
- Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA: Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 2004; 74: 106–120.. New York, USA: ACM Press, pp 19–27.
- Meng Z, Zaykin DV, Xu CF, Wagner M, Ehm MG: Selection of genetic markers for association analyses, using linkage disequilibrium and hap
- Weale ME, Depondt C, Macdonald SJ et al: Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: implications for linkage-disequilibrium gene mapping. Am J Hum Genet 2003; 73: 551– 565.lotypes. Am J Hum Genet 2003; 73: 115– 130.
- Stram DO, Haiman CA, Hirschhorn JN et al: Choosing haplotype- tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. Hum Hered 2003; 55: 27– 36.
- Halldorsson BV, Istrail S, De La Vega FM: Optimal selection of SNP markers for disease association studies. Hum Hered 2004; 58: 190–202.
- **Halldórsson, Bjarni V., Sorin Istrail, and Francisco M. De La Vega. "Optimal selection of SNP markers for disease association studies." Human heredity 58.3-4 (2004): 1**
- Ding, Keyue, and Iftikhar J. Kullo. "Methods for the selection of tagging SNPs: a comparison of tagging efficiency and performance." European Journal of Human Genetics 15.2 (2007): 228-236.90-202.
- Goode, Ellen L., et al. "Comparison of tagging single-nucleotide polymorphism methods in association analyses." *BMC proceedings*. Vol. 1. No. Suppl 1. BioMed Central Ltd, 2007.

Thank you