WHOLE GENOME TAG SNP SELECTION STRATEGY

Tai-Chun Wang , twan5724@it.usyd.edu.au

Supervisor: Prof. Albert Y Zomaya
School of Information Technologies

FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

1. SINGLE NUCLEOTIDE POLYMORPHISM AND HAPLOTYPE

Single Nucleotide Polymorphisms (SNPs)

- Single nucleotide polymorphism (SNP), a type of DNA mutation.
- SNPs exist with low mutation rate in DNAs, they are usually the best material for allele-related diseases studies and/or plant breeding purposes [1].
- The association between pair-wise SNPs is measured by the linkage disequilibrium (LD), and the common statistical correlation measurements are r^2 and D' values.

Haplotype

- A set of alleles at multiple loci on one chromosome that are transmitted together is called “haplotype”.

2. WHY USE TAG SNPs?

- Tag SNP is a informative SNPs subset. This subset can represents the haplotype.
- Genotyping all samples’ SNP is time consuming and costly. Tag SNP helps researchers to overcome these issues.
- Tag SNP has been used in type 1 diabetes related studies [2].

3. PROBLEMS WHEN SELECTING TAG SNPs

Models of Tag SNP

- Tag SNP selection can be classified into LD-based, block-free and block-based models.
- In LD-based model, one of the most common methods is to find a set of LD clusters which their pair-wise SNPs in each cluster are in high LD (e.g. $r^2 \geq 0.8$).
- In black-free model regard tag SNPs as a subset of all SNPs. Current genome-wide tag SNP selection methods is based on this model, but all of them are both time-consuming and memory-consuming.
- For block-based model, there are different definitions for a block in different methods, they may partition haplotypes into completely different blocks. These methods then try to find a minimum subset of SNPs which is capable of distinguishing all common haplotypes in each block.

Disadvantages

- For LD-based model, it cannot represent the complete haplotype correctly if a pair of SNPs’ distance is too long.
- For block-free model, most of these methods are restricted by a small-bounded location or fix number of tag SNPs that allowed to be selected.
- For block-based model, none of them consider about the LD between a pair of tag SNPs. The result may not be able to represent the original haplotype in some case. For instance, if a pair-wise tag SNPs has low LD, and if we treat major allele as “1” and minor allele as “0”, there would be less chance to represent “11” situation [3].

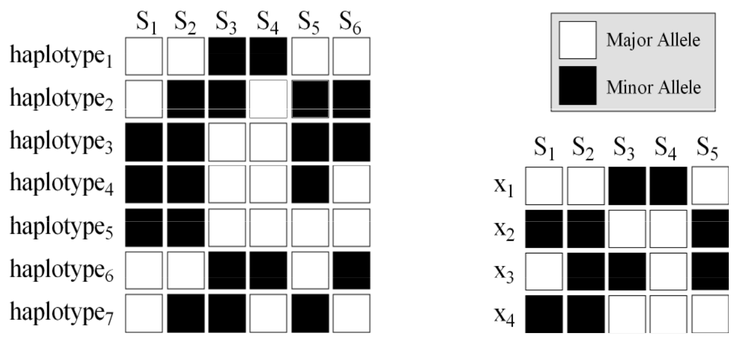
4. THE COMBINATIVE STRATEGY – tagSNPicker

Aim:

- In order to overcome the drawbacks of current methods, we combined the advantages of LD-based and block-based models to design a new method named tagSNPicker.
- The result generated by tagSNPicker would has the ability to handle and represent long haplotypes correctly.
- tagSNPicker is going to refer to the correlation between each candidate when selecting tag SNPs.

Methods:

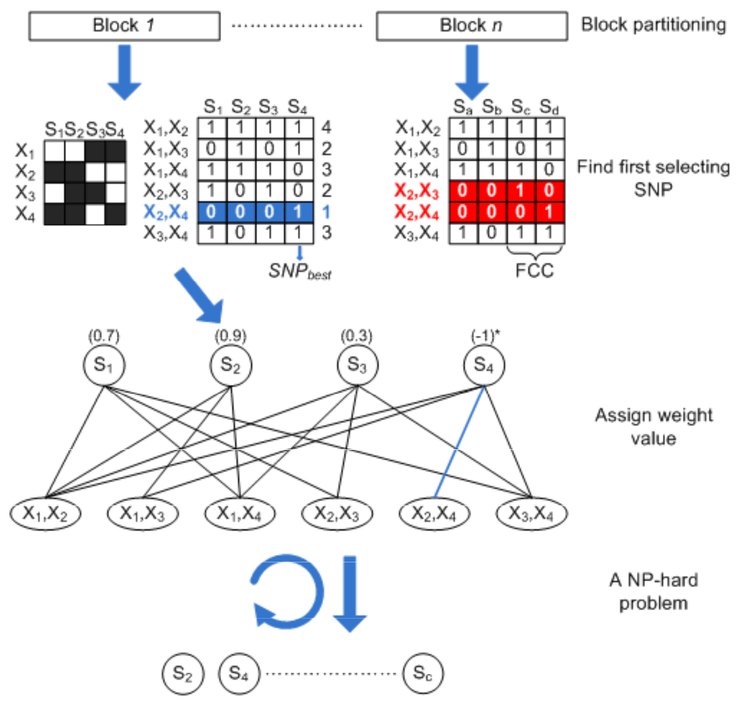
- Block partition method is not the main concern in this research. Hence, we use the *gene diversity* (*GD*) definition.



- We first transferred minimum tag SNP selection problem into weighted bipartite matching problem.

  Given a graph $G = \{(V_1 \cup V_2, E) \mid V_1$: SNPs, V_2: pair-wise haplotype patterns$\}$ with vertex weights ω: $V_1 \rightarrow (0, \infty)$.

  The weighted bipartite matching can be defined as finding the maximum match subgraph $G' = \{u \cup v, E' \mid u \subseteq V_1, v \subseteq V_2, E' \subseteq E\}$ with maximum weight, $\omega(u)$.



- Because weighted bipartite matching is a NP-hard problem, we translated it again into a minimum cost set covering problem; and the problem can be formulated as a zero-one integer linear problem as follows:

$$min \sum_{j=1}^{n} c_j x_j$$

subject to

$$\sum_{j=1}^{n} a_{ij} x_j \leq b_i, \quad (i = 1, 2, \ldots, m)$$

$$x_j = 0 \text{ or } 1, \quad (j = 1, 2, \ldots, n)$$

where $x_j = 1$ if j is in the cover, otherwise, it equals to zero. $a_{ij} = 1$ if $i \in S_j$, otherwise, it equals to zero.

- We deploy the additive algorithm developed by Balas *et al.* [4] to solve the associated integer programming of set cover problem. The algorithm simplify to the operation of addition and subtraction. As a result, it can eventually obtain an optimal or feasible solution faster.

5. EXPERIMENTAL RESULTS

Experimental result summary

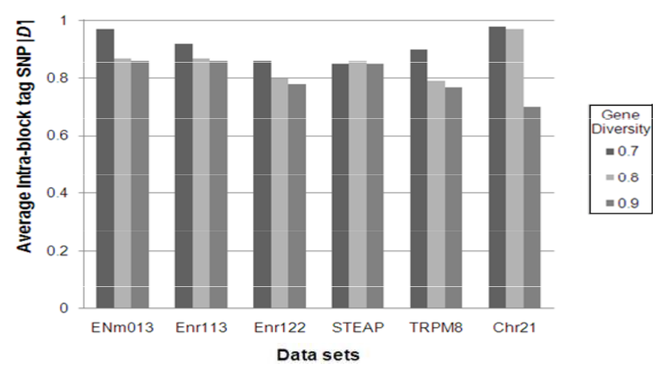
		Data Set					
		ENm013	Enr113	Enr122	STEAP	TRPM8	Chr21
Total SNPs		1023	785	560	31	166	11867
Block number		34	13	11	2	7	912
tag SNP number		252	106	75	17	43	6362
Avg. intra-block LD	(all SNPs)[1]	0.90	0.93	0.87	0.83	0.88	0.80
	(tag SNPs)[2]	0.85	0.89	0.83	0.90	0.80	0.78
Avg. inter-block LD	(all SNPs)[3]	0.57	0.56	0.55	0.79	0.65	0.50
	(tag SNPs)[4]	0.56	0.56	0.56	0.86	0.62	0.49

[1] The average D' value of all subset SNPs within each block.
[2] The average D' value of tag subset SNPs within each block.
[3] The average D' value of pair-wise SNPs.
[4] The average D' value of pair-wise tag SNPs cross block.

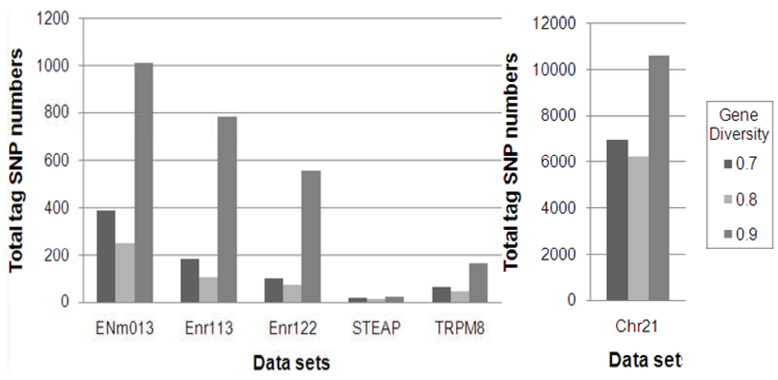
The comparison between tagSNPicker and GPT algorithm

		Data Set					
	Methods	ENm013	Enr113	Enr122	STEAP	TRPM8	Chr21
tag SNPs	(tagSNPicker)	252	106	75	17	43	6362
	(GPT)	250	103	73	14	43	6214
Avg. intra-block LD	(tagSNPicker)	0.85	0.89	0.83	0.90	0.80	0.78
	(GPT)	0.79	0.8	0.7	0.75	0.54	0.37

Average intra-block tag SNPs |D’| values under different gene diversity values



Total tag SNP numbers under different gene diversity values



The performance of tagSNPicker when using Patil’s[5] data set

	GD Value			
	0.7	0.8	0.9	GPT[1]
Block number	4030	2657	1454	4030
tag SNPs	14100	12842	10260	13867

[1] The GD value of GPT was equal to 0.8.

6. REFERENCES

[1] He, J. and Zelikovsky, A. (2007). Informative snp selection methods based on snp prediction. *IEEE Trans Nanobioscience*, 6(1), pp 60–67.

[2] Chapman, J., Cooper, J., Todd, J., and Clayton,D.(2003). Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Human Hered*, 56(1-3), pp 18–31.

[3] Wang, T.-C., Taheri, J. and Zomaya, A.Y. (2010). A Combinative Strategy for Higher Reliable tag SNPs Selection. *In Proceeding of The 11th International Conference on Bioinformatics and Computational Biology (BIOCOMP’10)*.

[4] Balas, E., Glover, F., and Zionts, S. (1965). An additive algorithm for solving linear programs with zero-one variables. *Operations Research*, 13(4), pp 517–549.

[5] Patil, N., Berno, A., Hinds, D., Barrett, W., Doshi, J., Hacker, C., Kautzer, C., Lee, D., Marjoribanks, C., McDonough, D., *et al*. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Sciences*, 294(5547), pp 1719.