Supplementary Information

TAGster: Efficient Selection of LD Tag SNPs in Single or Multiple Populations
---Evaluation of the algorithms implemented in *TAGster*

Contents

1. Introduction

We implemented 3 algorithms for tag SNP selection in *TAGster*. These algorithms are:

Algorithm 1: A Greedy algorithm for single or multi-population tag SNP;
Algorithm 2: An efficient exhaustive search algorithm for single population tag SNP;
Algorithm 3: A two-stage solution algorithm for multi-population tag SNP.

Please refer to file "*algorithm.pdf*" on *TAGster* website (http://dir.niehs.nih.gov/direb/tag*ster*/) for the detail of these algorithms.

We evaluated these algorithms against algorithms in existing software *ldSelect* (Carlson et al. 2004), *FESTA* (Qin, et al. 2006) and *MultiPop-TagSelect* (Howie, et al. 2006) using SNP genotype data from Environmental Genome Project (EGP) (http://egp.gs.washington.edu/), HapMap ENCODE project (http://hapmap.org/downloads/encode1.html.en).

2. Data

2.1 EGP Panel 2

At the time of this study, 207 genes were resequenced by EGP across 95 DNA samples from 4 populations (27 Africans, 24 Asians, 22 Europeans, and 22 Hispanics). There were a total of 16,153 SNPs with minor allele frequency (MAF) ≥ 0.05 in at least one of the 4 populations.

2.2 HapMap ENCODE

HapMap ENCODE (Encyclopedia of DNA Elements) Project resequenced ten 500 kb genomic regions in 48 individuals and subsequently genotyped all discovered SNPs as well as all SNPs in dbSNP at the time in 270 HapMap DNA samples from 3 populations including 30 CEPH (Utah residents with ancestry from northern and western Europe) trios, 90 Asians (45 unrelated JPT (Japanese in Tokyo, Japan), 45 unrelated CHB (Han Chinese in Beijing, China) and 30 YRI (Yoruba from Ibadan, Nigeria) trios. There were a total of 11,700 SNPs with minor allele frequency (MAF) ≥ 0.05 in at least one of the 3 populations.

3. Single population tag SNP

We applied both the refined greedy algorithm in *TAGster* and the greedy algorithm in ldSelect to select population specific tag SNPs at r^2 threshold of 0.8 from each population specific data set. Table 1 shows that, in EGP data, the modified greedy algorithm selected l42 fewer tag SNPs than the greedy algorithm as implemented in *ldSelect* (Carlson, et al., 2004) in EGP. For 62 genes the modified greedy algorithm selected fewer tags in at least one of the 4 populations, whereas the greedy algorithm had fewer tag SNPs in only 2 genes in one population. Table 2 shows the modified greedy algorithm selected 30 fewer tag SNPs than ldSelect using HapMap ENCODE data.

Table 1. Comparison between the refined greedy algorithm in *TAGster* and the greedy algorithm in *ldSelect* using EGP Panel 2 data.

Population	# SNP	# tag SNPs	
		TAGster	*ldSelect*
African	12807	5444	5509
Asian	7920	2139	2161
European	8538	2449	2477
Hispanics	8619	2605	2634

Table 2. Comparison between the refined greedy algorithm in *TAGster* and the greedy algorithm in *ldSelect* using HapMap ENCODE data.

Population	# SNP	# tag SNPs	
		TAGster	ldSelect
YRI	9005	3074	3085
Asian	6704	1193	1204
CEU	7615	1309	1317

We applied both the exhaustive search algorithms in *TAGster* and the comprehensive search algorithm in *FESTA* (Qin, et al., 2006) to select population specific tag SNPs at r^2 threshold of 0.8 and an exhaustive search step limit specification of 1,000,000 (the default setup of *FESTA*) for both algorithms for each of the 4 populations in EGP Panel 2.

Table 3 shows that the exhaustive search algorithm in *TAGster* greatly improved the computational efficiency in all 4 populations. Moreover, *FESTA* did not find an optimal solution for the number of tag SNPs for 1 gene in Africans and 1 gene in Europeans. *FESTA* exceeded the 1,000,000 step limit and defaulted to use of the greedy algorithm 20 times in order to provide a result while *TAGster* only used greedy algorithm 4 times (Table 4). Evaluation of HapMap ENCODE data to generate table 5 showed a similar pattern of computational efficiency and requirements for defaulting to the greedy algorithm.

Table 3. Comparison between *FESTA* and *TAGster* using EGP Panel 2 data

Pop	*FESTA*			*TAGster*			Time fold difference
	# tags	Time*	Greedy[1]	#tags	Time*	Greedy[1]	
African	5434	9308	6	5433	498	1	18.7
Asian	2132	5446	4	2132	206	1	26.4
European	2448	11427	4	2445	441	1	25.9
Hispanics	2592	9271	6	2592	300	1	30.9

*: Time is in second

[1]: Number of times that Greedy algorithm was used for tag SNP selection

Table 4. Gene list in EGP that greedy algorithm has to be used for selection of tag SNPs

Population	Gene	# SNP	Time (second)		# Tags		Greedy used*	
			FESTA	*TAGster*	*FESTA*	*TAGster*	*FESTA*	*TAGster*
African	tdp1	237	3750	12	45	45	1	0
African	Blm	151	435	20	47	47	1	0
African	App	191	1019	166	118	118	1	1
African	adh4	116	1198	2	18	17	1	0
African	fancd2	172	2222	6	27	27	1	0
African	trpm2	139	435	59	64	64	1	0
Asian	App	161	2436	76	51	51	1	1
Asian	capn3	64	357	1	14	14	1	0
Asian	fancd2	117	2173	2	11	11	1	0
Asian	trpm2	69	341	3	21	21	1	0
CEU	Blm	112	412	2	20	20	1	0
CEU	Tpo	194	5317	290	50	50	1	1
CEU	mlh3	42	928	1	15	12	1	0
CEU	aldh1a2	101	4624	14	22	22	1	0
Hisp	abcc1	187	191	7	60	60	1	0
Hisp	App	164	1484	7	60	60	1	0
Hisp	Tpo	174	785	75	58	58	1	1
Hisp	tp53bp1	68	2421	5	9	9	1	0
Hisp	fancd2	126	3063	2	16	16	1	0
Hisp	rad18	136	1198	88	21	21	1	0

*: Greedy algorithm has to be used for tag SNP selection

Table 5. Comparison between *FESTA* and *TAGster* using HapMap ENCODE data

Pop	*FESTA*			*TAGster*			Time fold difference
	# tags	Time*	Greedy[1]	#tags	Time*	Greedy[1]	
African	3067	2383	2	3067	1739	0	1.37
Asian	1187	11852	6	1184	2093	2	5.66
CEU	1302	3417	2	1302	1212	0	2.82

*: Time is in second

[1]: Number of times that Greedy algorithm was used for tag SNP selection

4. Multiple population tag SNP

We applied the modified greedy algorithm (Algorithm 1) and 2-stage method (Algorithm 3) to select multi-population tag SNP in 207 genes for the 4 populations from EGP Panel 2 and used as a benchmark measure the number of tag SNPs found using *ldSelect* followed by *MultiPop-TagSelect* (Howie, et al., 2006). The generalized modified greedy algorithm (generalized algorithm 1 for multiple populations) reduced tag SNP requirements by 183 SNPs whereas the two-stage method (Algorithm 3) reduced tag SNP requirements by 159 SNPs. If for each gene we selected the minimum of these two methods, it reduced tag SNP requirements by 233 SNPs below that required by *ldSelect* followed by *MultiPop-TagSelect* (Table 4). Evaluation in 3 populations from HapMap ENCODE shows a similar pattern of reduction (Table 6).

Both *TAGster* and *MultiPop-TagSelect* allow an investigator to specify *a priori* a set of SNPs for inclusion as tag SNP. *MultiPop-TagSelect* algorithm selects from population-specific tag SNPs. Thus if an investigator-specified SNP is not one of these population specific tag SNPs, then it can not serve as a proxy for any population specific LD bin. Conversely, in the *TAGster* selection process, every investigator-specified SNP can serve as a proxy for other SNPs unless it is a singleton SNPs.

Table 6: Multi-population tag SNPs for 4 populations from EGP Panel 2

Method		# tag SNPs
MultiPop-TagSelect		7612
TAGster	Greedy	7429
	2-stage	7453
	Hybrid*	7379

*: minimal number of tag SNPs of the greedy and 2-stage method for each gene

Table 7: Multi-population tag SNPs for 3 populations from HapMap ENCODE

Method		# tag SNPs
MultiPop-TagSelect		3917
TAGster	Greedy	3882
	2-stage	3845
	Hybrid*	3843

*: minimal number of tag SNPs of the greedy and 2-stage method for each gene

5. Multiple SNP bin tag SNP

In order to further reduce the number of tag SNPs, investigators may choose to select tag SNPs only for bins that contain multiple SNPs. The minimum bin size can be specify using the parameter *-minimum* in the parameter file *params.txt.* For example setting

-minimum: 2

requires that bins contain at least two SNPs and eliminates singleton bin tag SNPs. Elimination of singleton bin tag SNPs can dramatically cut down the number of tag SNPs, while still capturing the majority of SNPs. It is particularly useful when selecting multiple population tag SNPs. For example, if parameter *–minimum* is set to a value of 2, *TAGster* selected **4094** multiple population multiple SNP bin (MPMS) tag SNPs for the 4 populations in EGP, compared to **7429** SNPs required if singleton bins are tagged. This smaller number of tag SNPs still captures ~95% common SNPs in Asian and CEPH populations, 91% in Hispanic population and 84% in Africans. For HapMap ENCODE data, **2095** MPMS tag SNPs (out of total of **3882** tag SNPs if singleton bin tags are included) can capture ~96% of common SNPs in Asian and CEU and 86% of SNPs in YRI.