The Whole Genome TagSNP Selection and Transferability Among HapMap Populations

Reedik Magi, Lauris Kaplinski, and Maido Remm

Pacific Symposium on Biocomputing 11:535-543(2006)

THE WHOLE GENOME TAGSNP SELECTION AND TRANSFERABILITY AMONG HAPMAP POPULATIONS

REEDIK MÄGI, LAURIS KAPLINSKI, MAIDO REMM

Department of Bioinformatics, University of Tartu, Riia str. 23
Tartu, 51010, Estonia

One of the crucial issues of association studies is the selection of markers. One possible approach would be to select tagging SNPs (tSNPs) according the HapMap information. In this study we present the number of tSNPs required for the association analysis of the entire human genome for all available HapMap population samples: CEPH, Nigerian, Chinese and Japanese. For future association studies, it is also important to know how well the tSNP set of one population sample can describe the markers of another population. Therefore, we have calculated the proportion of markers adequately described by tSNPs and how many additional tSNPs we need to describe all markers of another population.

1. Introduction

The selection of markers is a very important step for a successful association study. The amount of markers and thereby the cost of the study can be significantly reduced by selecting markers according to any popular tagging SNP (tSNP) selection method[1-3]. Studies have shown that it takes approximately 500,000 – 1,000,000 to cover the whole human genome[4-6]. Lately, haploblock independent methods have gathered more popularity among researchers, particularly the r^2-bin based method, which was introduced by Carlson et al.[7].

The HapMap Project started in 2001 with the goal of determining the common patterns in the human genome and to make that information freely available in the public domain[8]. The HapMap information gives us an opportunity to calculate the approximate number of tSNPs necessary to cover the whole human genome. This approach can be used for marker selection in national genome projects or in other large association studies.

An important aspect of marker selection in standardized populations is the transferability of chosen tSNP sets to other populations[2,9]. As the allele frequencies of many markers differ among populations, the selected markers of one population sample may be insufficient to describe the whole heterogeneity of another population. Ahmadi *et al.* found, in their study of European and Japanese

population samples, that it is possible to identify tSNPs that work adequately in multiple population groups[10]. Also, it is known that despite of the role of a populations' demographic history, the linkage disequilibrium (LD) pattern of all four HapMap populations is remarkably similar[11]. However, the extension of LD and haplotype allele frequencies of areas with high LD, may be very different[12].

Two important goals were set for this work:

- To find the number of tSNPs necessary to describe the whole human genome, according to HapMap population samples;
- To evaluate the transferability of HapMap tSNPs among populations – the proportion of markers adequately described by tSNP sets and the number of additional tSNPs that have to be chosen from a population sample to describe all of its markers.

2. Methods

2.1. *Population samples*

In this study we used public data from HapMap release #16c.1.* Genotype information of the following four population samples has been used for tSNP selection:

- CEPH (Utah residents with ancestry from northern and western Europe) (abbreviation: CEU) 1,104,996 markers genotyped at 60 founders
- Japanese in Tokyo, Japan (abbreviation: JPT) 1,087,297 markers genotyped at 45 population samples
- Han Chinese in Beijing, China (abbreviation: CHB) 1,087,297 markers genotyped at 45 population samples
- Yoruba in Ibadan, Nigeria (abbreviation: YRI) 1,076,381 markers genotyped on 60 founders

SNP data of all chromosomes, except for Y-chromosome, were used in this study.

2.2. *r^2-bin based tSNP selection*

We created a software program, REAPER, which is optimized for fast tSNPs selection. The program implements the r^2-bin algorithm[7] for tSNP selection and is specifically designed for full genome scale analysis. It is written in C++ and is available for 32bit Linux and Windows operating systems.

---

* Publicly available final data freeze of Phase I (21. June 2005) from http://www.hapmap.org/. Redundant-filtered data was used.

The algorithm works as follows: REAPER uses the greedy approach, always starting from the biggest possible bin. In the first step, all r^2 values between markers less than N positions apart are calculated for the full genome. Lowering the distance threshold N increases the calculation speed and lowers memory consumption, but increases the risk of leaving some relevant SNPs out of bins. In the current analysis, a default distance threshold of 1024 positions was used, giving us reasonable calculation times with the assumption that no two markers more than 1024 positions apart show significant r^2. Only boolean values are stored for each analyzed marker pair, based on whether the calculated r^2 was above or equal to the linkage threshold or not. The default linkage threshold is 0.8. Lowering the threshold results in larger r^2-bins and smaller number of tSNPs. However, with a lower linkage threshold, the average prediction power of tSNPs decreases, thus requiring more individuals in further association studies. The program allows the use of any r^2 threshold between 0 and 1. For each bin, the marker with the largest number of "good" r^2 values (above or equal to the threshold) from the set of unbinned markers, is selected as the r^2-bin seed marker. All other markers, whose r^2 with the seed, is above or equal to the threshold are put into the bin, together with the seed marker. The markers in the bin are sorted by the number of "good" r^2 values with other bin members. All markers in the bin having, "good" r^2 values with all other bin members, are marked as alternative tSNPs. The candidate tSNP with the highest average r^2 will be reported for the given bin. The same algorithm is repeated, choosing a new seed marker from the remaining, ungrouped markers of the genome, until no SNP has a r^2 value with other ungrouped markers above or equal to the linkage threshold. The remaining SNPs are then added to the tSNP list as single-marker bins. REAPER can also be forced to use a pre-selected list of tSNPs from a file. In this case, the tSNP selection procedure is identical, except that in the first step bins are constructed by picking seeds from the tSNP list. Remaining markers are then distributed to bins using the abovementioned greedy algorithm.

The calculations described in this paper (ca 25 sets of whole-genome tSNPs) were made by parallel computing on 14 Pentium 4 2.200MHz computers with 1GB memory each. This calculation process took about 1 week to complete. REAPER is freely available for academic users at http://bioinfo.ebc.ee/download/.

2.3. *tSNP transferability calculation*

The HapMap data is likely to be used for the selection of tSNPs in the future. However, the cases for the association study may be collected from different populations. In order to evaluate the suitability of the HapMap based tSNPs are on other populations, we calculated the transferability of tSNP sets between HapMap populations. For each population, the tSNP set was calculated with the r^2-bin method ($r^2 \geq 0.8$). These tSNP sets were used in other populations' tSNP calculations

as pre-selected seeds. For each population, the number of additional tSNPs to describe all markers in the dataset was calculated.

2.4. *Strategies for tSNP selection from other populations*

If the tSNP set is calculated from a different population, it might be possible to genotype some additional individuals from the observed population to increase the efficiency of the tSNP set. The tSNP sets that were defined in the CEPH trios were tested on all other populations. We added an additional, randomly selected, 10, 20 and 30 persons from the other datasets and tested how efficiently the tSNP sets of these mixed populations can describe Chinese, Japanese and Nigerian population samples. The efficiency of a given tSNP set in the other population samples is evaluated by the following criteria: an average r^2 among all typed SNPs and 25% lower quartile of r^2 among all typed SNPs.

3. Results

3.1. *The number of tSNPs required for the whole human genome*

The idea of tSNP based association analysis is currently very popular. Using this analysis method, it is therefore interesting to understand how many tSNPs does it take to cover the whole human genome. To answer this question, we have calculated the tSNP sets for all four HapMap populations using the r^2-bin method. The number of tSNPs needed to describe the population samples varied from 579,978 in the CEPH sample set to up to 716,617 in the Nigerian set (Table 1).

We have also analyzed the distribution of r^2-bins by size (Fig.1). The proportion of markers in single-marker-bins is varying from 75.9% in the CEPH samples to 79.6% in the Nigerian sets. The r2-bins over 30 markers were rarely found in any population. R^2-bins with more than 10 markers are most common in the CEPH population sample (1.9%) and are quite rare in the Nigerian samples (0.6%). Largest r^2-bins are in chromosome 12 of the CEPH population (257 markers) and in chromosome X of the Japanese population (255 markers).

Table 1. The number of tSNPs necessary to describe all markers in each HapMap population sample.[†] The numbers on the diagonal show how many tSNPs are necessary to describe each sample. Each populations sample tSNP sets were also tested on other samples to test the transferability of tSNPs. Additional tSNPs have been found for the markers that are not described by a tSNP set (added after '+' signs) of each corresponding population sample. tSNP sets were calculated according to the r^2-bin method with $r^2 \geq 0.8$ in bins.

		population sample of tSNP selection			
		CEU	JPT	CHB	YRI
studied population sample	CEU	579,978	+91,495 (15.5%)	+63,088 (9.9%)	+67,555 (9.4%)
	JPT	+92,840 (16.0%)	590,979	+50,993 (8.0%)	+82,165 (11.5%)
	CHB	+88,018 (15.2%)	+54,470 (9.2%)	639,459	+78,163 (10.9%)
	YRI	+200,358 (34.6%)	+209,398 (35.4%)	+122,019 (19.0%)	716,617

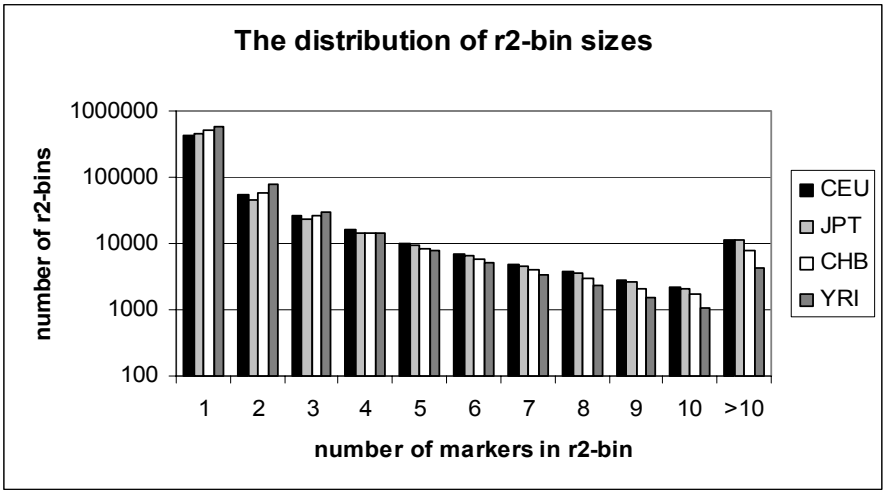

Fig 1. The distribution of r^2-bin sizes. The counts of r2-bins with all marker numbers have been determined for all four populations.

[†] Chromosome Y not included.

3.2. *tSNP transferability between populations*

To find the union and the intersection of different tSNP sets of population samples, we selected the tSNP list of one population sample and used this as a pre-selected seed list for REAPER, while calculating r^2-bins of another population. The percentage of additional tSNPs needed is indirectly reflecting the union between the tSNP sets of the two populations. We have also determined the proportion of tSNPs that is not used in another population sample by comparing a population's own tSNP count with the number of tSNPs required from another population (Table 1, outside of diagonals). The number of additional tSNPs required to cover the other populations is the smallest for the Nigerian population sample tSNP set (additional 9.4% - 11.5%) and the largest in the Japanese populations sample tSNPs (additional 35.4% of tSNPs to cover the Nigerian samples). The number of redundant tSNPs can be found by conducting a comparison of the number of tSNPs found from the population sample with the number of transferred tSNPs + additional tSNPs. For example, if we use the Japanese tSNP set on the Chinese population sample, we need 590,979 markers + 54,470 additional tSNPs. It takes 639,459 markers to describe the Chinese population sample with its own tSNPs. Therefore, we have genotyped 639,459 – 645,449 = 5990 tSNPs in excess.

3.3. *Strategies for tSNP selection from other populations*

In the future, the association studies will be performed on various populations, many of which are not covered by the current HapMap. If the standard HapMap population tSNPs were used for studying other population samples, then the performance of tSNP might be improved by genotyping a small amount of people from the same population before tSNP selection (local HapMap approach). We have used the CEPH tSNP set on the other HapMap populations in order to test how many individuals should be added to increase the performance of tSNPs. To understand how many extras were required, local individuals were mixed with the CEPH individuals' samples. As the tSNP set should describe all genotyped markers in the dataset, we determined the best r^2 score between each marker and any tSNP. For the whole population's dataset, we found the average of these r^2 scores and the lesser quartile of 25%. As expected, additional genotyped persons increase the r^2 values between tSNPs and other markers most efficiently by adding the Nigerian population samples (Fig. 2). Also note that the performance of some poor markers (indicated by lower quartile) increases significantly by using local population samples for the selection of tSNPs.

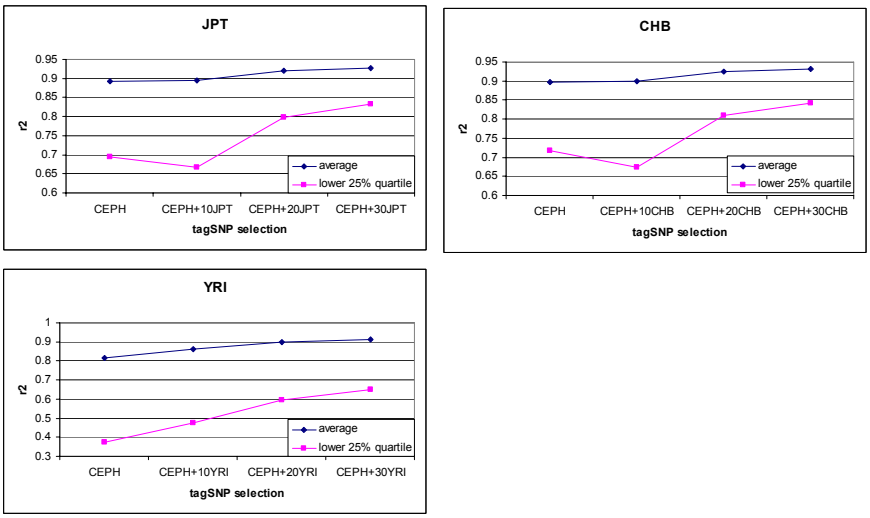

Fig 2A-C. Mean and 25% lower quartile of r^2 value between tSNPs and all the markers in the observed population sample. tSNPs were calculated according to the CEPH population sample; CEPH + 10 random persons from the observed population sample; CEPH + 20 random persons from the observed population sample; CEPH + 30 random persons from the observed population sample.

4. Discussion

The results of the association analysis are strongly determined by marker selection. Thereby the selection of tSNPs is one of the most important aspects while designing a new study. The HapMap Consortium has contributed greatly to this research by making four population samples, with more than million markers genotyped in each, publicly available. That data gives researchers an opportunity to use fine scale LD data for designing and using different marker selection algorithms.

The approximate number of tSNPs to cover the whole human genome has been predicted to be between 500,000 and 1,000,000 – one marker per 3 - 5kb[4,5]. Our results re-confirm the order of magnitude for HapMap populations. The large number of singleton markers indicates that the current HapMap marker density is still too sparse for adequate coverage of the whole human genome. Therefore, the number of tSNPs may grow, when new data is added to HapMap in Phase II.

As the idea of tSNP based association analysis is currently very popular, it is important to find out how well the tSNP set of HapMap population will work on other population samples. Our results indicate that tSNP sets from different populations can describe other populations reasonably well. This result is similar to

the findings of Ahmadi *et al.*[10]. In their study of Japanese and European (CEPH) populations they found that the number of tSNPs that adequately cover both populations is only 19% higher than the one required for only one population. Our findings show that the percentage is somewhat lower – 15.5%-16% in the same pair of populations. However, we also found that the tSNP sets of non-African populations describe Nigerian data poorly. And *vice versa*, if we use Nigerian tSNPs on non-African populations, we found that we genotyped a large number of unnecessary markers.

The current study required numerous calculations of the whole-genome tSNP sets, which was possible thanks to the speed and low memory demand of REAPER.

Acknowledgments

This work is supported by the Estonian Ministry of Education and Research grant no. 0182649s04 and by the applied research grant EU19730 from Enterprise Estonia. We thank Ulvi Gerst-Talas, Jody Novakoski and Katre Palm for valuable help with English grammar.

References

1. Halldorsson, B. V., Istrail, S. & De La Vega, F. M. Optimal selection of SNP markers for disease association studies. *Hum Hered* **58**, 190-202 (2004).
2. Mueller, J. C. et al. Linkage disequilibrium patterns and tagSNP transferability among European populations. *Am J Hum Genet* **76**, 387-98 (2005).
3. Gabriel, S. B. et al. The structure of haplotype blocks in the human genome. *Science* **296**, 2225-9 (2002).
4. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet* **22**, 139-44 (1999).
5. Dunning, A. M. et al. The extent of linkage disequilibrium in four populations with distinct demographic histories. *Am J Hum Genet* **67**, 1544-54 (2000).
6. Judson, R., Salisbury, B., Schneider, J., Windemuth, A. & Stephens, J. C. How many SNPs does a genome-wide haplotype map require? *Pharmacogenomics* **3**, 379-91 (2002).
7. Carlson, C. S. et al. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* **74**, 106-20 (2004).
8. The International HapMap Project. *Nature* **426**, 789-96 (2003).

9. Liu, N. et al. Haplotype block structures show significant variation among populations. *Genet Epidemiol* **27**, 385-400 (2004).
10. Ahmadi, K. R. et al. A single-nucleotide polymorphism tagging set for human drug metabolism and transport. *Nat Genet* **37**, 84-9 (2005).
11. De La Vega, F. M. et al. The linkage disequilibrium maps of three human chromosomes across four populations reflect their demographic history and a common underlying recombination pattern. *Genome Res* **15**, 454-62 (2005).
12. Sawyer, S. L. et al. Linkage disequilibrium patterns vary substantially among populations. *Eur J Hum Genet* **13**, 677-86 (2005).