Transferability of Tag SNPs to Capture Common Genetic Variation in DNA Repair Genes Across Multiple Populations

Paul I.W. De Bakker, Robert R. Graham, David Altshuler, Brian E. Henderson, and Christopher A. Haiman

Pacific Symposium on Biocomputing 11:478-486(2006)

TRANSFERABILITY OF TAG SNPS TO CAPTURE COMMON GENETIC VARIATION IN DNA REPAIR GENES ACROSS MULTIPLE POPULATIONS

PAUL I.W. DE BAKKER, ROBERT R. GRAHAM, DAVID ALTSHULER
Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA

BRIAN E. HENDERSON, CHRISTOPHER A. HAIMAN
Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA

Genetic association studies can be made more cost-effective by exploiting linkage disequilibrium patterns between nearby single-nucleotide polymorphisms (SNPs). The International HapMap Project now offers a dense SNP map across the human genome in four population samples. One question is how well tag SNPs chosen from a resource like HapMap can capture common variation in independent disease samples. To address the issue of tag SNP transferability, we genotyped 2,783 SNPs across 61 genes (with a total span of 6 Mb) involved in DNA repair in 466 individuals from multiple populations. We picked tag SNPs in samples with European ancestry from the Centre d'Etude du Polymorphisme Humain, and evaluated coverage of common variation in the other samples. Our comparative analysis shows that common variation in non-African samples can be captured robustly with only marginal loss in terms of the maximum r^2. We also evaluated the transferability of specified multi-marker haplotypes as predictors for untyped SNPs, and demonstrate that they provide equivalent coverage compared to single-marker tests (pairwise tags) while requiring fewer SNPs for genotyping. The efficacy of a tagging-based approach in studying genotype-phenotype correlations in complex traits is strongly supported by our empirical results.

1. Introduction

A significant fraction of the risk of developing common diseases such as cancer is due to genetic variation. Knowledge of the genetic basis of these complex traits may lead to new insights into disease pathogenesis, the identification of novel drug targets, and ultimately contribute to human health.

Family-based linkage analysis has been very successful in localizing causal variants for monogenic, Mendelian diseases. However, success has been rather limited for common diseases, where multiple loci are likely to act in concert and contribute only probabilistically [1]. Testing genetic variants for association between cases and appropriate controls offers a more powerful approach to detect putative causal variants, but require large sample sizes to achieve adequate power [2].

Complete ascertainment of genetic variation by resequencing is the only comprehensive approach to test all variants (both common and rare) directly for association. For the foreseeable future, routine resequencing in thousands of individuals will not be practical. But high-throughput technology to type large numbers of SNPs in thousands of people is rapidly improving, making it possible to probe the vast majority of human heterozygosity due to common variations. In addition, public databases of SNP variation have swelled to 10 million variants. The International HapMap Project provides genome-wide data in 269 individuals from four different population groups [3], and supports the selection of informative markers ("tag SNPs") by exploiting redundancies among nearby polymorphisms due to linkage disequilibrium (LD) [4]. Tagging approaches may substantially improve the cost-effectiveness of association studies by delivering greater power and better genotyping efficiency through the selection of tag SNPs and definition of statistical tests based on the empirical LD patterns in HapMap (and similar resources such as [5]).

An important outstanding question is whether tag SNPs picked from HapMap will be transferable across independent disease samples, and how this varies for different testing strategies (especially methods based on haplotypes). The precise LD patterns are likely to differ between population groups, given the many forces that determine the patterns of genetic variation. Thus, empirical evidence is required to study the efficiency and coverage of tag SNPs across different populations, and to validate LD-based tagging approaches in general. The work described here builds upon early studies that have begun to address this issue [6-8].

We report a large data set of genes involved in DNA repair within which we have performed dense genotyping in multiple population samples. By picking tag SNPs in one population sample, we can perform a blind assessment as to how well these tag SNPs—and allelic tests for association based on them—capture the variation in any of the other population samples.

2. Methods and Materials

2.1. *DNA samples*

We have collected genotype data from seven population samples. The CEPH (Centre d'Etude du Polymorphisme Humain) samples are a subset (20 trios, 60 individuals in total) of the 30 trios used in HapMap (designated as CEU samples) [3]. The African American (AA, $n = 70$), Native Hawaiian (NH, $n = 67$), Latino (LA, $n = 70$), Japanese (JA, $n = 70$) and White (WH, $n = 70$) samples were selected from the Multiethnic Cohort (MEC) conducted in Hawaii

and California (mainly Los Angeles) [9]. The Chinese samples (CH, n = 59) were selected from an ongoing study in Shanghai and from the Singapore Chinese Health Study [10].

2.2. *Genotyped SNPs*

SNPs for genotyping were selected from dbSNP in 61 DNA repair genes (Table 1) with a total span of 5.7 Mb. This resulted in a working set of 2,783 successfully genotyped SNPs from all samples with an average marker density of 1 SNP every ~2 kb. Criteria for successful conversion are: Hardy-Weinberg $P > 0.01$ (for five of the six ethnic groups), genotyping percentage >75%, no more than one discordant blinded replicate (9 total) or Mendel inconsistency in parent-offspring trios (CEPH only).

As expected, we observe more common SNPs in AA than in the other samples, reflecting greater genetic diversity (heterozygosity) in African-derived populations.

The data sets were phased using the program EMPHASE (written by Nick Patterson) to give 140 unrelated chromosomes (haplotypes) for AA, LA, JA, WH; 134 for NH; 118 for CH and 80 for CEPH. EMPHASE is based on the expectation-maximization algorithm [11].

Table 1. List of selected DNA repair genes and number of successfully genotyped SNPs in all population samples.

Locus	# SNPs	Locus	# SNPs	Locus	# SNPs
APE1	30	Ku80	58	POLE	74
ATM	57	LIG1	55	POLI	34
ATR	33	LIG3	25	POLK	32
Artemis	45	LIG4	28	RAD50	42
BLM	71	MGMT	114	RAD51	20
BRCA1	32	MLH1	41	RAD52	45
BRCA2	59	MLH3	26	RPA1	68
CHEK1	44	MRE11	47	RPA2	33
CHEK2	39	MSH2	38	RPA3	73
CSA	52	MSH3	133	TP53	19
CSB	69	MSH6	25	XPA	40
DNA-PK	43	NEIL1	13	XPB	41
ERCC1	26	NEIL2	46	XPC	55
FANCA	64	OGG1	39	XPD	28
FANCC	50	PARP1	66	XPF	55
FANCD2	38	PCNA	27	XPG	61
FANCE	34	PMS1	49	XRCC1	42
FANCF	17	PMS2	29	XRCC2	38
FANCG	22	POLB	31	XRCC3	39
FEN1	19	POLD	43	XRCC4	83
Ku70	22				
				Total	**2,783**

2.3. *Selection of tag SNPs*

Many different methods have been proposed for selecting tag SNPs [12-16]. Pairwise methods offer straightforward analysis, but fail to exploit long-range haplotype structure. We have developed a tagging approach—called Tagger—that combines the simplicity of pairwise methods with the potential efficiency gains of multi-marker approaches [17].

In this study, we focus specifically on *common* variants with a frequency of ≥ 5%, given the limited ascertainment of less common SNPs in this data set. We picked tags from the CEPH samples as the reference panel so that all observed common variants are captured with $r^2 \geq 0.8$. Use of this threshold has become common practice in the field [15].

Tagging was performed in two modes: (*a*) by a greedy pairwise approach, in which every common allele is captured by a single tag at the prescribed r^2 threshold [15], and (*b*) by aggressively searching for specific multi-marker (haplotype) tests to improve tagging efficiency. We achieve the latter by first picking pairwise tags, and then iteratively dropping tags, one by one, and replacing them with a specific multi-marker predictor (using any of the remaining tag SNPs). That predictor is accepted only if it can capture the alleles originally captured by the discarded tag at the required r^2; otherwise, that provisionally dropped tag is considered indispensable and kept. This multi-marker approach essentially finds an identical set of 1 d.f. tests of association, only now using certain specific haplotypes as effective surrogates for single tag SNPs, thereby requiring fewer tag SNPs for genotyping. To minimize risk of overfitting, tag SNPs within a specified multi-marker test are forced to be in strong LD (defined as LOD > 3) with one another and with the predicted allele.

Tagger thus outputs (1) a list of tag SNPs, and (2) a list of allelic tests, both central for the evaluation of tag SNP transferability.

Tagger is available in the stand-alone application Haploview [18] and as a web server at http://www.broad.mit.edu/mpg/tagger/.

2.4. *Evaluation of tag SNPs*

Given the lists of tag SNPs, we evaluated the coverage of the common variants in the population samples (other than CEPH) by computing the maximum r^2 between the common variants observed in those samples and the specified allelic tests. For pairwise tagging, these tests simply correspond to the genotypes of every tag SNP (as single-marker tests). For multi-marker tagging, tests were specified during tag SNPs selection from the reference panel. (Importantly, in the evaluation of tag SNPs, we do not allow ourselves to derive better allelic tests by looking at LD patterns in the population sample under evaluation.)

3. Results

3.1. *Selection of tag SNPs*

To mimic how investigators will be using the HapMap resource, we used the CEPH samples as the reference panel for picking tag SNPs. For all 61 loci, we required all common variants (≥5%) observed in the reference panel to be captured at $r^2 \geq 0.8$.

We picked a total of 718 tag SNPs by pairwise tagging, and 631 tag SNPs when we allowed Tagger to form multi-marker predictors in place of single-marker tests (Table 2). For both tagging approaches, the mean r^2 for all common alleles (in the reference panel) was 0.97, and the minimum r^2 was 0.86 (these are averages over all 61 loci).

Table 2. Tag SNPs picked from CEPH as the reference panel. The mean and minimal r^2 are averages over all 61 loci studied.

Method	Number of tag SNPs	Mean r^2	Minimum r^2
Pairwise tagging	718	0.97	0.86
Multi-marker tagging (specified haplotype tests)	631	0.97	0.86

This suggests that a nontrivial boost in genotyping efficiency can be achieved by multi-marker tagging, exploiting the underlying haplotype structure, in contrast to pairwise tagging which relies solely on single-marker relationships between SNPs.

We note that the efficiency gain between pairwise and multi-marker tagging observed here (~12%) is significantly lower than that typically obtained in broader genomic regions such as the data from the HapMap-ENCODE project [17]. It is not uncommon for distant (> 100 kb) markers to be in strong LD and to form haplotypes that proxy for other SNPs. Since this study was performed on multiple genes (with an average span of 94 kb), overall efficiency was reduced compared to tagging in large contiguous regions of the genome.

3.2. *Evaluation of tag SNPs*

Having picked tag SNPs and defined statistical tests from the CEPH reference panel, we evaluated the performance of pairwise tagging in terms of the r^2 at which common variants are captured in each of the other six population samples (AA, HA, LA, JA, CH and WH). For every locus, we computed the percentage

of common SNPs captured at $r^2 \geq 0.2$, 0.5 and 0.8 as well as the mean r^2 and minimum r^2. We present these metrics as averages over all 61 loci (Table 3).

Table 3. Coverage of common (≥5%) SNPs in six population samples by pairwise tag SNPs picked in CEPH as the reference panel. Values are averages over all 61 loci studied.

Population sample	Number of common (≥5%) SNPs	Percentage of common SNPs captured at $r^2 \geq$			Mean r^2	Minimum r^2
		0.2	0.5	0.8		
AA	2347	88.8%	69.3%	50.4%	0.68	0.06
HA	2196	97.2%	92.7%	85.3%	0.90	0.45
LA	2273	97.9%	93.9%	80.5%	0.88	0.40
JA	2028	95.9%	92.4%	82.3%	0.88	0.33
CH	2030	97.0%	91.7%	79.2%	0.87	0.37
WH	2191	98.6%	95.8%	87.3%	0.92	0.51

Most importantly, coverage of common alleles in the HA, LA, JA, CH and WH samples appears to be robust. In the WH samples (which is most "similar" from a population-genetic standpoint), we observe a marginal drop in mean r^2 from 0.96 (in the CEPH reference panel) to 0.92. Between 80% and 87% of common variants are captured at $r^2 \geq 0.8$ in the non-African samples, and the overwhelming majority (> 92%) are captured at $r^2 \geq 0.5$. Of course, not all alleles are captured equally well: a small fraction (3-4%) of the common alleles in the non-African samples are not captured at all ($r^2 < 0.2$) by any of the allelic tests.

Not surprisingly, fewer common variants are captured in the AA samples: only 50% of the common alleles are captured with $r^2 \geq 0.8$; and the mean r^2 dropped down to 0.68. This can be attributed to the significantly lower extent of LD in African populations [19]. We emphasize that in practice, however, investigators will likely pick tag SNPs from a reference panel that is more representative (such as the HapMap samples of Yoruba from Ibadan, Nigeria). Due to greater genetic diversity and less LD, more tag SNPs will be required for capturing common variation in African-derived samples.

We next evaluated the performance of the multi-marker predictors on the basis of the 631 tag SNPs picked by Tagger. Again, we computed the percentage of common alleles captured at $r^2 \geq 0.2$, 0.5 and 0.8 as well as the mean r^2 and minimum r^2 (Table 4). The coverage with our haplotype-based approach is roughly equivalent to that of pairwise tagging but require fewer tag SNPs. Thus, the multi-marker approach in Tagger is not only more efficient than a pairwise tagging method, but the specified haplotype predictors capture

common variation in the other (non-African) population samples almost as well as the single-marker tests (Table 3).

Table 4. Coverage of common (≥5%) SNPs in six population samples by tag SNPs picked and specified multi-marker tests defined in CEPH as the reference panel. Values are averages over all 61 loci studied.

Population sample	Number of common (≥5%) SNPs	Percentage of common SNPs captured at r^2 ≥			Mean r^2	Minimum r^2
		0.2	0.5	0.8		
AA	2347	88.8%	66.6%	46.7%	0.66	0.06
HA	2196	97.3%	92.3%	83.7%	0.89	0.45
LA	2273	97.8%	92.8%	78.8%	0.86	0.38
JA	2028	95.3%	90.9%	79.1%	0.86	0.32
CH	2030	96.9%	90.5%	76.7%	0.85	0.35
WH	2191	98.6%	95.6%	87.0%	0.91	0.51

4. Discussion

Using empirical genotype data in genes of medical relevance, we find that (*a*) tag SNPs picked in the CEPH samples provide good coverage of common variants in the non-African population samples studied here; and (*b*) specified haplotype tests can improve overall tagging efficiency with minimal loss of coverage.

Even though the fine details of LD patterns are known to differ between population samples, these results demonstrate that tag SNPs chosen from the CEPH reference panel (used in HapMap) are able to effectively capture the majority of common alleles in other (non-African) samples in a cost-effective manner.

Although this work focuses only on a limited set of parameters, we believe that the results presented here are fairly representative of the practical decisions that investigators face in the design of tag SNP sets.

Our tagging approach, like that of others, is explicitly not based on haplotype "blocks," hotspots of recombination, or other features of empirical data. We agree with commentators who have noted that while blocks may be a convenient descriptor of genotype data, a block-by-block approach ignores the sometimes substantial correlations between blocks, and as not all SNPs are contained within blocks, block-based selection of tag SNPs is likely to give inadequate coverage [20].

While many different approaches exist for selecting tag SNPs from a reference panel and for performing tests, these concepts are sufficiently

intertwined and should be considered as a unit. Tag SNPs may perform well under the particular analytical strategy for which they were designed, but not under another. We do not address in this study the tradeoff between the amount of required genotyping and statistical power to detect an association in an actual disease study. We have addressed these issues elsewhere [17].

Acknowledgments

We would like to thank Melissa A. Frasco for laboratory assistance and Xin Sheng, John T. Casagrande and David Van Den Berg for technical support. We would also like to acknowledge Laurence N. Kolonel, Loïc Le Marchand, Ronald K. Ross, Mimi C. Yu and Juan-Min Yuan for providing the samples for this study.

References

1. Altmuller, J., et al., *Genomewide scans of complex human diseases: true linkage is hard to find.* Am J Hum Genet, 2001. **69**(5): p. 936-50.
2. Lohmueller, K.E., et al., *Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease.* Nat Genet, 2003. **33**(2): p. 177-82.
3. The International HapMap Consortium, *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-96.
4. Johnson, G.C., et al., *Haplotype tagging for the identification of common disease genes.* Nat Genet, 2001. **29**(2): p. 233-7.
5. Hinds, D.A., et al., *Whole-genome patterns of common DNA variation in three human populations.* Science, 2005. **307**(5712): p. 1072-9.
6. Nejentsev, S., et al., *Comparative high-resolution analysis of linkage disequilibrium and tag single nucleotide polymorphisms between populations in the vitamin D receptor gene.* Hum Mol Genet, 2004. **13**(15): p. 1633-9.
7. Ahmadi, K.R., et al., *A single-nucleotide polymorphism tagging set for human drug metabolism and transport.* Nat Genet, 2005. **37**(1): p. 84-9.
8. Mueller, J.C., et al., *Linkage disequilibrium patterns and tagSNP transferability among European populations.* Am J Hum Genet, 2005. **76**(3): p. 387-98.
9. Kolonel, L.N., et al., *A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics.* Am J Epidemiol, 2000. **151**(4): p. 346-57.
10. Hankin, J.H., et al., *Singapore Chinese Health Study: development, validation, and calibration of the quantitative food frequency questionnaire.* Nutr Cancer, 2001. **39**(2): p. 187-95.

11. Excoffier, L. and M. Slatkin, *Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population.* Mol Biol Evol, 1995. **12**(5): p. 921-7.
12. Stram, D.O., 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**(1): p. 27-36.
13. Weale, M.E., 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**(3): p. 551-65.
14. Meng, Z., et al., *Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes.* Am J Hum Genet, 2003. **73**(1): p. 115-30.
15. Carlson, C.S., et al., *Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium.* Am J Hum Genet, 2004. **74**(1): p. 106-20.
16. Lin, Z. and R.B. Altman, *Finding haplotype tagging SNPs by use of principal components analysis.* Am J Hum Genet, 2004. **75**(5): p. 850-61.
17. de Bakker, P.I.W., et al., *Efficiency and power in genetic association studies.* Nat Genet, 2005. **In the press.**
18. Barrett, J.C., et al., *Haploview: analysis and visualization of LD and haplotype maps.* Bioinformatics, 2005. **21**(2): p. 263-5.
19. Reich, D.E., et al., *Linkage disequilibrium in the human genome.* Nature, 2001. **411**(6834): p. 199-204.
20. Wall, J.D. and J.K. Pritchard, *Haplotype blocks and linkage disequilibrium in the human genome.* Nat Rev Genet, 2003. **4**(8): p. 587-97.