J Hum Genet (2008) 53:96–99
DOI 10.1007/s10038-007-0221-7

SHORT COMMUNICATION

HapMap coverage for SNPs in the Japanese population

Fumihiko Takeuchi · Masakuni Serizawa · Norihiro Kato

Received: 2 September 2007 / Accepted: 1 November 2007 / Published online: 28 November 2007

Abstract The coverage of human genomic variations is known to substantially affect the success of genome-wide association studies. We therefore assessed the SNP coverage in the HapMap database for a total of 1,304 subjects from the Japanese population by combining resequencing and high-density genotyping approaches. First, we resequenced 48 Japanese subjects in 86 genes (572 kb in total), and we then genotyped the subset of tag SNPs and also imputed genotypes for all of the detected SNPs in an additional panel of 1,256 subjects. Subsequently, we genotyped 555,352 tag SNPs selected from the HapMap in 72 Japanese subjects (from the panel of 1,256 subjects) and further imputed genotypes for all SNPs currently included in the HapMap. Of 738 common genic SNPs (1.3 per kb) that we detected by resequencing, 58% had already been genotyped in the HapMap, and 31% were not genotyped but had a proxy SNP in the HapMap with a linkage disequilibrium coefficient $r^2 \geq 0.8$, whereas 11% were not represented in the current HapMap database. Thus, the HapMap coverage appears to be high although not thorough for SNPs in the Japanese population as compared to its coverage reported in Caucasians, and this needs to be considered when we interpret association results.

Keywords HapMap · SNP · Japanese population · Linkage disequilibrium · Genome-wide association study

Introduction

The increased coverage of human genomic variations in the HapMap database has been a key factor in the recent success of genome-wide association studies of Caucasian populations (Altshuler and Daly 2007). Here, hundreds of thousands of "tag SNPs" that capture variations of 3.9 million SNPs in the HapMap (Release 22) are genotyped in the study samples. By combining genotypes of characterized SNPs with the catalog of haplotypes in the HapMap database, genotypes of the HapMap SNPs that are not directly characterized in a given study panel can be imputed with high accuracy. As a consequence, most of the HapMap SNPs can be tested for association (Marchini et al. 2007; Pe'er et al. 2006; Servin and Stephens 2007). However, if a particular SNP associated with disease is not included in the HapMap database, the chance of identifying such a disease association relies on the degree of linkage disequilibrium (LD) between a given SNP and the HapMap SNPs tested in the study. Indeed, the sample size required to identify association when a HapMap SNP in LD is tested becomes larger than the case when the given SNP itself is tested, and the increase in size is roughly the inverse of the LD coefficient r^2 between the two SNPs (Pritchard and Przeworski 2001). Thus, the statistical power of genome-wide association studies largely depends on the HapMap coverage of SNPs that are observable in the ethnic group

Electronic supplementary material The online version of this article (doi:10.1007/s10038-007-0221-7) contains supplementary material, which is available to authorized users.

F. Takeuchi
Department of Medical Ecology and Informatics, Research Institute, International Medical Center of Japan, Tokyo, Japan

M. Serizawa · N. Kato (✉)
Department of Gene Diagnostics and Therapeutics, Research Institute, International Medical Center of Japan, 1-21-1 Toyama, Shinjuku-ku, Tokyo 162-8655, Japan
e-mail: nokato@ri.imcj.go.jp

F. Takeuchi
Wellcome Trust Sanger Institute, Cambridge, UK

under investigation; that is, the proportion of all of the SNPs covered by the HapMap.

The overall HapMap coverage is estimated to be high in the Caucasian and Asian populations (The International HapMap Consortium 2007). There are ~1.9 common SNPs, which have a minor allele frequency (MAF) of ≥5%, per kilobase (kb) on average (The International HapMap Consortium 2005), and 1.4 SNPs per kb (not all are common) have been selected and are currently genotyped in the HapMap. As for SNPs in the Asian populations, it has been estimated that 92% of common SNPs have a "proxy SNP" in the HapMap with an LD coefficient $r^2 \geq 0.8$ to a given SNP. On the other hand, it has been reported that only 51% of the SNPs not directly genotyped in the HapMap may have a proxy in the HapMap (Tantoso et al. 2006). These estimations were made by resequencing a relatively modest number of individuals in particular chromosomal regions; namely 16 Asians (48 subjects in total when different ethnic groups were analyzed together) in the HapMap ENCODE regions covering 5 Mb, and 24 Asians (95 in total) in 84 genes investigated by the NIEHS Environmental Genome Project. Because of the limited number of individuals used for resequencing, some rare SNPs and/or ethnically specific SNPs may have been missed. For example, a SNP with a MAF of 5% can be missed with a probability of 0.22 when only 16 subjects are resequenced: the minor allele is not included in 32 chromosomes with a probability of $(1 - 0.05)^{32} = 0.194$, and also a minor allele included in one chromosome is missed with a probability of $32 \times 0.05 \times (1 - 0.05)^{31} \times 0.07 = 0.023$, assuming a 7% experimental miss rate (Stephens et al. 2006). The missing probability decreases to 0.10 for 24 subjects and 0.01 for 48 subjects.

Under these circumstances, we evaluate the HapMap coverage for SNPs in the Japanese population by combining resequencing and high-density genotyping approaches.

Materials and methods

In order to capture genic SNPs comprehensively in 86 genes, we first resequenced 48 Japanese subjects (572 kb in total, see Table S1 of the "Electronic Supplementary Material"). The SNP search was performed as part of our ongoing projects regarding atherosclerosis candidate genes. The SNPs were screened by resequencing genomic DNA in all exons, exon–intron borders, and 5′- and 3′-untranslated regions of each gene; in particular, the entire genomic sequences including introns were screened in genes whose total length was <4 kb, namely *ADM*, *ADRA2B*, *ADRB2*, *APOA1*, *APOA2*, *APOA4*, *APOB48R*, *AVP*, *AVPR2*, *CCL2*, *CMA1*, *CSF2*, *GH1*, *GP1BA*, *INS*, *KCNJ11*, *NPPA*, *NPPB*, *NPPC*, *PNMT*, *TNF*, *UTS2R*. In total, 738 common SNPs were detected.

Next, in addition to the 48 subjects used for the initial SNP search, genotype data of the SNPs detected in a panel of 1,256 Japanese subjects were obtained by characterizing tag SNPs and by imputing genotypes of the remaining non-tag SNPs. We selected 92% of the detected SNPs as tag SNPs ($n = 676$) by (1) grouping SNPs with mutual $r^2 \geq 0.6$ for haplotype construction, (2) inferring haplotypes in each group, and (3) then removing redundant SNPs that are in complete LD with reference to haplotypes showing a frequency of at least 5% (Takeuchi et al. 2005). Genotyping was performed either with TaqMan® SNP Genotyping Assays in ABI 7900HT (Applied Biosystems, Foster City, CA, USA) or with the MassARRAY® Compact system (Sequenom-Bruker, San Diego, CA, USA) according to the manufacturer's protocols. Genotypes of the non-tag SNPs were imputed in the 1,256 subjects by incorporating genotype data from the 48 subjects from the initial search using the Mach software (Li and Abecasis 2006). The accuracy of imputation depends on the selection of tag SNPs and the imputation algorithm. We assessed this accuracy by masking genotypes of the non-tag SNPs in an individual who was selected from the 48 subjects resequenced, performing the imputation in the 1,256 subjects plus one subject (chosen from the 48 subjects from the initial search) based on genotype data from the remaining 47 subjects, and then testing concordance between the imputed genotypes and the masked true genotypes. For each SNP, the genotype error rate—defined as 0, 50 or 100% when zero, one or two alleles differ between the two genotypes—was averaged over 48 trials where genotypes of one subject were masked at a time. In total, 97% of the detected SNPs were either genotyped directly or imputed with a genotype error rate of <5%, which is sufficiently accurate to evaluate the r^2 coefficient for the purposes of this study.

In addition, to see whether the detected SNPs were rather specific to the Japanese, we genotyped the tag SNPs in samples from 100 African Americans and 100 Caucasians, as purchased from the Coriell Cell Repositories (Camden, NJ, USA).

Subsequently, in order to evaluate the proportion of SNPs that were captured by the HapMap SNPs to those detected in the Japanese subjects by resequencing, we obtained genotype data on the HapMap SNPs in 72 Japanese subjects from among the 1,256 subjects mentioned above (these subjects also did not overlap with the 48 subjects used for resequencing). First, we genotyped 555,352 SNPs that were primarily selected as tag SNPs from the HapMap on the Illumina HumanHap 550 K array (San Diego, CA, USA). Then we imputed genotypes of the remaining SNPs that were not included in the 550 K array

but were available in the HapMap by incorporating genotype data for 90 Asian samples (45 Japanese in Tokyo and 45 Han Chinese in Beijing) from the HapMap Release 21a with the Mach software. Again, we assessed the accuracy of imputation by masking genotypes of SNPs that were not on the 550 K array in one of the 90 Asian samples, performing imputation for the 72 subjects plus one Asian subject, and then testing the concordance between the imputed genotypes and the masked true genotypes. Of the HapMap SNPs, 15% had been included in the 550 K array and 83% were imputed with a genotype error rate of <5%; together, these accounted for 97%.

Accordingly, in the 72 Japanese subjects, we obtained genotype data for all SNPs detected by resequencing in 86 genes and for all SNPs in the HapMap. We then analyzed the proportion of the detected SNPs that had been included in the HapMap, and the proportion of the missed SNPs that was in strong LD with the SNPs in the HapMap. For each SNP detected by our resequencing, if it had not been genotyped yet in the HapMap by itself, we sought an alternative HapMap SNP within a distance of 500 kb by achieving the highest LD in r^2 with reference to the genotypic information thus obtained in the 72 Japanese subjects. When r^2 was at least 0.8, the detected SNP was considered to have the HapMap SNP as a proxy, otherwise the detected SNP was not captured by the HapMap. All Japanese subjects gave informed consent for participation and details that might disclose the identities of the subjects under study were omitted. This study was approved by the ethics committee of the International Medical Center of Japan.

Results and discussion

We detected a total of 738 common SNPs (1.3 per kb), of which 430 (58%) were subsequently found to be included in the current HapMap database (Table 1). Looking at the remaining 308 SNPs that were not included in the HapMap, 229 (31% of the total) had a proxy SNP. The HapMap coverage can be evaluated as the total proportion of SNPs that are either genotyped or that have a proxy in the HapMap, which was 89% on average and was almost concordant regardless of the MAF. Among the SNPs not included in the HapMap, the proportion of SNPs with a proxy was 74%, which was higher than the value of 51% reported in a previous study (Tantoso et al. 2006). In the present dataset, the coverage of the HapMap varied according to the location of the SNPs: 91% in coding regions (n = 169), 87% in untranslated regions (n = 67), 91% in introns (n = 442), and 73% in intergenic regions including 5′-upstream regions of the gene (n = 60).

It should be noted that the HapMap coverage was relatively low in a subset of 22 genes (59 kb in size) where SNPs were screened for the entire gene including introns. Of 71 common SNPs detected by resequencing in these 22 genes, 38% (27 SNPs) had been included in the HapMap, whereas among the missing SNPs only 23% (16 SNPs) had a proxy SNP, and 39% (28 SNPs) could not be represented by the HapMap SNPs with high LD ($r^2 \geq 0.8$) to a given SNP. The lower coverage in this subset of genes may be explained in part by the relatively high proportion of intergenic regions to resequenced regions in these small genes (18%) compared to the other genes studied (5%). Since this estimation is based on the results for a limited number of genes that were selected and screened in our experimental setting, further extensive evaluation is warranted.

In order to see whether the SNPs not in the HapMap were missed because of ethnic specificity, we measured the allele frequency in African American and Caucasian samples and compared the distribution of MAF between the SNPs included and those not included in the HapMap. We found no remarkable differences between the two categories of SNPs (supplementary Fig. S1), suggesting that the SNPs were likely to be missed by chance rather than because of ethnic specificity.

Also, we looked for "recombination hotspots" within 10 kb from each SNP using the HapMap database. 65% of the SNPs in weak LD ($r^2 \leq 0.4$) with the HapMap SNPs were located close to a hotspot, while the corresponding proportion was 50% as for the SNPs with a proxy ($r^2 \geq 0.8$) in the HapMap. Thus, SNPs not captured by the HapMap were not exclusively located in the recombination hotspots, where a given SNP would be in weaker LD with nearby SNPs.

We consider a previous estimate of 92% in Asians to be a somewhat overestimated coverage of the HapMap (The International HapMap Consortium 2007) because of the following two reasons. First, since the number of resequenced individuals in the ENCODE regions is limited (16 Asians within a total of 48 subjects), the total number of SNPs may be underestimated. Second, while a *random* subset of common SNPs were expected to be included in the HapMap, there was overestimation in the number of SNPs not included yet having a proxy in the HapMap. A large proportion of the SNPs that are registered in dbSNP/HapMap were originally discovered by resequencing a small number of individuals. Here, a limited number of chromosomes were chosen in a population and SNPs showing a certain level of polymorphism between these chromosomes were subsequently discovered. Such SNPs tend to be mutually in strong LD (e.g., SNPs in complete LD are all discovered together), and accordingly, a subset of SNPs that are selected from the SNP resources thus discovered should have served, in reality, as proxies for a smaller number of SNPs, compared to an equivalent-sized set of SNPs that are literally *randomly* selected from the

Table 1 SNPs detected by resequencing classified and counted by the minor allele frequency (MAF) or genic location and the maximum linkage disequilibrium coefficient r^2 to a SNP in the HapMap

Classification	Maximum r^2 to a SNP in the HapMap					Included in HapMap (%)	Total
	$0 \le r^2 \le 0.2$	$0.2 < r^2 \le 0.4$	$0.4 < r^2 \le 0.6$	$0.6 < r^2 \le 0.8$	$0.8 < r^2 \le 1$[a] (%)		
MAF							
$5\% \le$ MAF $\le 10\%$	1	5	4	6	47 (37)	65 (51)	128
$10\% <$ MAF $\le 20\%$	2	7	7	10	63 (30)	121 (58)	210
$20\% <$ MAF $\le 30\%$	1	1	2	5	34 (25)	91 (68)	134
$30\% <$ MAF $\le 40\%$	1	2	8	8	48 (33)	79 (54)	146
$40\% <$ MAF $\le 50\%$	0	1	0	8	37 (31)	74 (62)	120
Location in the gene							
Coding region	0	2	0	13	32 (19)	122 (72)	169
Untranslated region	0	2	5	2	17 (25)	41 (61)	67
Intron	4	11	10	14	158 (36)	245 (55)	442
Intergenic	1	1	6	8	22 (37)	22 (37)	60
Total	5	16	21	37	229 (31)	430 (58)	738

[a] The detected SNP has a proxy in the HapMap

full SNP resources, which involve the current catalog of SNPs in dbSNP/HapMap plus SNPs that remain to be discovered (or registered). This literally *randomly* selected set of SNPs are less likely to be mutually in strong LD.

In conclusion, we found that 58% of common genic SNPs detected by resequencing had been already genotyped in the HapMap, and 31% were not genotyped but had a proxy SNP in the HapMap; in other words, the remaining 11% did not appear to be represented in the current HapMap database. Taken together, while association studies using the current HapMap-based SNPs seem to be promising in the identification of disease susceptibility, the HapMap coverage is not thorough for SNPs in the Japanese population, and this needs to be considered when we interpret the association results.

All of the SNPs discovered in this project have been submitted to dbSNP, and detailed information is available from the JMDBase website. A list of detected common SNPs and their coverage by the HapMap is provided in Table S2 of the "Electronic supplementary material."

Websites

dbSNP: http://www.ncbi.nlm.nih.gov/projects/SNP/
HapMap: http://www.hapmap.org
JMDBase: http://www.jmdbase.jp
Mach: http://www.sph.umich.edu/csg/abecasis/MACH/index.html
NIEHS Environmental Genome Project: http://egp.gs.washington.edu/

Acknowledgments The authors thank Dr. Kazuyuki Yanai for his valuable advice and support during data preparation, and the anonymous referees for helpful comments. This work was supported by a grant-in-aid from the Ministry of Education, Cultures, Sports, Science and Technology of Japan (#18018046).

References

Altshuler D, Daly M (2007) Guilt beyond a reasonable doubt. Nat Genet 39:813–815
Li Y, Abecasis GR (2006) Mach 1.0: rapid haplotype reconstruction and missing genotype inference. Am J Hum Genet S79:2290
Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39:906–913
Pe'er I, de Bakker PI, Maller J, Yelensky R, Altshuler D, Daly MJ (2006) Evaluating and improving power in whole-genome association studies using fixed marker sets. Nat Genet 38:663–667
Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. Am J Hum Genet 69:1–14
Servin B, Stephens M (2007) Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet 3:e114
Stephens M, Sloan JS, Robertson PD, Scheet P, Nickerson DA (2006) Automating sequence-based detection and genotyping of SNPs from diploid samples. Nat Genet 38:375–381
Takeuchi F, Yanai K, Morii T, Ishinaga Y, Taniguchi-Yanai K, Nagano S, Kato N (2005) Linkage disequilibrium grouping of single nucleotide polymorphisms (SNPs) reflecting haplotype phylogeny for efficient selection of tag SNPs. Genetics 170:291–304
Tantoso E, Yang Y, Li KB (2006) How well do HapMap SNPs capture the untyped SNPs? BMC Genomics 7:238
The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–320
The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861