Int J Hum Genet, 6(4): 297-301 (2006)
PRINT: ISSN 0972-3757 ONLINE: 2456-6360
DOI: 10.31901/24566330.2006/06.04.04

Large Scale HapMap Genotyping and the Possibility of Genome Center-Specific Effects

Daniel P K Ng[1,2], David Koh[1,2] and Chia Kee-Seng[1,2]

1. Department of Community, Occupational and Family Medicine, National University of Singapore
2. Centre for Molecular Epidemiology, National University of Singapore

KEYWORDS Single nucleotide polymorphisms; genotyping protocol; HapMap; heterozygosity; Hardy Weinberg equilibrium

ABSTRACT HapMap samples are currently being genotyped using different high throughput protocols at various international genome centres. To determine if there are any differences between SNP genotypes that may be related to these protocols, we analysed an initial set A consisting of 2,200 SNPs (100 SNPs from each autosome) typed in 90 HapMap CEU samples. Although SNP composition in terms of percentage of transitions and transversions was similar across protocols, one (termed "PI") yielded a high prevalence (39.9%) of mono-morphic SNPs (i.e. those with heterozygosity = 0) which was generally double that observed for the other protocols (corrected P-value (P_c) < 0.01). To examine this issue further, we enlarged the dataset to include a total of 22,000 SNPs (1000 SNPs per autosome). While results in this larger dataset B remained similar for all other protocols, the prevalence of mono-morphic SNPs genotyped using "PI" declined by nearly half from 39.9% to 19.1% ($P_c < 1x10^{-7}$). Stratifying both the initial and larger datasets by genome centres, it was observed that the prevalence of polymorphic SNPs (defined as those with heterozygosity >0 and <= 0.5) genotyped using "PI" increased from 42% to more than 70% at two locations while staying relatively consistent at the remaining centres. Although our analysis does not allow us to pinpoint the precise cause for this discrepancy, our findings clearly advocate greater caution when using high throughput technologies in order to ensure consistent genotype calls.

INTRODUCTION

Following successes in pinpointing mutations for rare monogenic diseases, focus has now shifted towards identifying the susceptibility genes for common diseases such as cancer, cardiovascular disease, and diabetes among others (Farrall and Morris 2005). This research endeavour entails the genotyping of a large number of human DNA samples for many single nucleotide polymorphisms (SNPs) to determine if there is a significant disease association (Tanaka et al. 2003). In response to this technical challenge, various technologies for large scale genotyping have emerged implementing the use of beadarrays, microarrays, mass spectrometry, enzyme cleavase methods among others. Large scale genotyping represents a significant departure from more conventional genotyping protocols in which SNP assays (such as PCR-RFLP) are typically designed, executed and analysed singly, allowing individual genotypes to be manually inspected and scored. In contrast, large scale genotyping necessarily employs computer algorithms to design assays and to automatically perform genotype calls while using computational measures in an attempt to ensure quality control, without the need for manual inspection of each genotype call. These quality control measures do not routinely include comparison of genotype calls against other methods including that of DNA sequencing, which is arguably the gold standard. Unlike conventional genotyping, large scale genotyping reactions are typically carried out in a multiplex manner with the simultaneous examination of hundreds to thousands of SNPs in a single reaction.

The International HapMap Project is the premiere example of a large scale genotyping project (The International HapMap Consortium 2003). Since the mammoth task of genotyping millions of SNPs throughout the genome is currently being carried out in various international centres, institutions and organisations and the project employs a diverse range of high throughput genotyping platforms, it may be postulated that these and possibly other factors could impact genotyping results, a hypothesis which we have presently tested using publicly available data from the project.

Corresponding Author: Daniel P K Ng, PhD
Department of Community, Occupational and Family Medicine (MD3) Yong Loo Lin School of Medicine
National University of Singapore 16, Medical Drive, Singapore 117597
Telephone: 65 6516 4967, Fax: 65 6779 1489
E-mail: cofnpkd@nus.edu.sg

MATERIALS AND METHODS

HapMap data

HapMap data release #16c.1 (June 2005) was downloaded from the International HapMap Project website (http://www.hapmap.org/). Genotypes analysed in this study are those of the 90 Caucasian CEU samples. Based on this release, we formed two datasets for our analyses. Set A, used in preliminary analyses, consisted of genotypes for the first 100 SNPs from each of the 22 autosomes, giving rise to a total of 2,200 SNPs. Set B, the larger replication dataset with 19,800 SNPs, was made up of the next 900 SNPs from each of the 22 autosomes. Set A and set B together comprised 22,000 SNPs. Both datasets were characterised with regard to SNP nature (transition/transversions), genotype conformity to Hardy-Weinberg equilibrium (HWE) as well as SNP heterozygosity (HET). Information on factors such as type of protocol and genotyping centre was also included. From inspection of both datasets together with information from the HapMap website as well as a recent report by the International HapMap Consortium (The International HapMap Consortium 2005), we identified six genotyping protocols including Illumina BeadArray (denoted PI), Third Wave (PT), Acycloprime (PF), MassExtend (PM), Affymetrix microarray (PA) and ParAllele (PP). Similarly, genotyping centers/organisations were also identified including Baylor College of Medicine with ParAllele BioScience (denoted CBC), Broad Institute and MIT (CBR), Chinese HapMap Consortium consisting of Beijing Genomics Institute, the Chinese National Human Genome Centers at Beijing and Shanghai, the University of Hong Kong, The Hong Kong University of Science and Technology and the Chinese University of Hong Kong (CCH), Illumina (CIL), RIKEN (CRK), McGill University-Genome Quebec Innovation Center (CMG), Wellcome Trust Sanger Institute (CSG) and University of California, San Francisco with Washington University in St Louis (CCW).

Data Analysis

Only non-redundant data indicated as having passed HapMap quality controls were analysed in this study. SNPs were considered to have conformed to expectations under HWE if P values associated with goodness-of-fit χ^2 test were <=0.05. HET based on actual genotypes was calculated as number of heterozygotes / number of successful genotypes for each SNP. Expected HET was calculated as the product of *2pq* where *p* and *q* are the major and minor allele frequencies respectively. SNPs were grouped as non-polymorphic (i.e. HET = 0), polymorphic with acceptable HET values (0 < HET <=0.5), and finally, polymorphic but with HET values exceeding 0.5 which is the theoretical maximum for bi-allelic SNPs with a maximum minor allele frequency of 0.5. Categorical data was analysed using χ^2 tests. As multiple comparisons were performed, we corrected the native P-values using the Bonferroni procedure. Corrected P-values (indicated as P_c) < 0.05 were considered statistically significant. The CEU samples are formed from 30 trios including two parents and a child. As exclusion of genotype data for the child in each trio did not significantly affect the conclusions in this study, findings presented in this paper are for all 90 CEU samples.

RESULTS

Initial data analysis was performed based on the 2,200 SNPs in Set A (Table 1). Approximately half (54.6%) of these SNPs were genotyped using protocol "PI", with a further 28.5% by protocol "PT". The remaining 16.9% of SNPs were analysed using four other methods. The majority (68.3%) of SNPs were transitions and this was not significantly different between genotyping protocols (P_c = NS). Among SNPs with acceptable HET values (0 < HET <= 0.5), the vast majority were in HWE and this was comparable across genotyping protocols (P_c = NS) (Table 1). Intriguingly, the overall distribution of SNPs when grouped according to HET was significantly different between genotyping protocols (P_c < 0.01). Particularly, a substantial percentage of SNPs (39.9%) genotyped by protocol "PI" was non-polymorphic. In comparison, the prevalence of non-polymorphic SNPs was much lower at 22.3% for the next most common protocol "PT" (Table 1). For the less common protocols, this percentage ranged from 5.1 to 29.6%, this variability likely reflecting the lower number of SNPs genotyped using these methods.

We next repeated our analyses using the

Table 1: SNP distribution in Set A and B categorized according to genotyping protocol.

Protocol	*PI*		*PT*		*PF*		*PM*		*PA*		*PP*	
Dateset	Set A	Set B	Set A	Set B	Set A	Set B	Set A	Set B	Set A	Set B	Set A	Set B
Distribution of	1201	10589	627	5686	39	402	106	1067	119	1171	108	885
sampled SNPs	(54.6)	(53.4)	(28.5)	(28.7)	(1.8)	(2)	(4.8)	(5.4)	(5.4)	(5.9)	(4.9)	(4.5)
Nature of SNP												
Transitions	806	7209	451	3974	25	267	80	693	77	774	63	616
	(67.1)	(68.1)	(71.9)	(69.9)	(64.1)	(66.4)	(75.5)	(64.9)	(64.7)	(66.1)	(58.3)	(69.6)
Transversions	395	3380	176	1712	14	135	26	374	42	397	45	269
	(32.9)	(31.9)	(28.1)	(30.1)	(35.9)	(33.6)	(24.5)	(35.1)	(35.3)	(33.9)	(41.7)	(30.4)
No of SNPs with:												
HET = 0	479*	2021	140	1205	2	65	16	161	9	111	32	175
	(39.9)	(19.1)	(22.3)	(21.2)	(5.1)	(16.2)	(15.1)	(15.1)	(7.6)	(9.5)	(29.6)	(19.8)
HET > 0, <= 0.5	629	7513	425	3967	36	265	82	806	99	969	64	620
	(52.7)	(71)	(67.8)	(69.8)	(92.3)	(65.9)	(77.4)	(75.5)	(83.2)	(82.7)	(59.3)	(70.1)
HET > 0.5	93	1055	62	514	1	72	8	100	11	91	12	90
	(7.7)	(10)	(9.9)	(9)	(2.6)	(17.9)	(7.6)	(9.4)	(9.2)	(7.8)	(11.1)	(10.2)
Total	1201	10589	627	5686	39	402	106	1067	119	1171	108	885
	(100)	(100)	(100	(100)	(100)	(100)	(100)	(100)	(100)	(100)	(100)	(100)
No of SNPs:												
In HWE	623	7473	415	3885	34	264	81	786	97	946	63	616
	(99.1	(99.5)	(97.7)	(97.9)	(94.4)	(99.6)	(98.8)	(97.5)	(98)	(97.6)	(98.4)	(99.4)
Not in HWE	6	40	10	82	2	1	1	20	2	23	1	4
(goodness-of-fit	(1)	(0.5)	(2.4)	(2.1)	(5.6)	(0.4)	(1.2)	(2.5)	(2)	(2.4)	(1.6)	(0.6)
P < 0.05)	629	7513	425	3967	36	265	82	806	99	969	64	620
	(100)	(100)	(100)	(100)	(100)	(100)	(100)	(100)	(100)	(100)	(100)	(100)

Genotypes were based on HapMap data determined for the 90 Hapmap CEU Caucasian samples.
Set A consisted of 2,200 SNPs comprising 100 each per autosome.
Set B consisted of 19,800 SNPs with 900 per autosome.
*$P_c < 1 \times 10^{-7}$ between set A and B for protocol "PI"
#excluding SNPs with HET = 0 or > 0.5.

larger replication dataset B which contained 19,800 SNPs (Table 1). SNP distribution according to SNP nature (transitions/transversions) and conformity to HWE remained comparable to Set A. However, the percentage of SNPs typed as non-polymorphic using protocol "PI" declined by more than half from 39.9% (Set A) to 19.1% (Set B) ($P_c < 1x10^{-7}$). For the remaining five protocols, the distribution of SNPs according to HET categories did not differ significantly between the two datasets (P_c = NS) (Table 1).

Protocol "PI" and "PM" were employed by more than one genotyping centre. As such, SNPs with acceptable HET values were cross-tabulated by both protocol and centre. In two of five centres using 'PI", SNP distribution was comparable between Set A and B (P_c = NS) (Table 2). For centres "CCH" and "CIL", the percentage of SNPs with acceptable HET nearly doubled from ~ 42% (Set A) to more than 70% (Set B) (Table 2); these increments were highly significant even after adjustment for multiple comparisons ($P_c < 1x10^{-6}$). For centre "CSG", a significant increase was also observed but the difference was more modest (Table 2). Protocol "PM" did not show any significant genotyping centre-specific differences between Set A and B (P_c value = NS). There were no statistically significant differences between the two datasets for the remaining protocols, each of which was employed in only one centre (P_c = NS).

DISCUSSION

In this study, we used genotype data obtained for 22,000 SNPs from the International HapMap Project to study whether various factors impact aggregate genotype calls. Our analyses were initially carried out on 2,200 SNPs which provided equal representation from all 22 autosomes (Set A). We then sought to investigate any outstanding findings from this preliminary analysis in a substantially larger replication sample of 19,800 SNPs (Set B).

Table 2:. Percentage of SNPs with HET values >0 and <=0.5 according to genotyping protocol and centre.

Protocol	*PI*		*PT*		*PF*		*PM*		*PA*		*PP*	
Dataset	*Set A*	*Set B*	*Set A*	*Set B*	*Set A*	*Set B*	*Set A*	*Set B*	*Set A*	*Set B*	*Set A*	*Set B*
SNPs (N)	1201	10589	627	5686	39	402	106	1067	119	1171	108	885
Genotyping center												
CBC	-	-	-	-	-	-	-	-	-	-	64/108 (59.3)	620/885 (70.1)
CBR	75/116 (64.7)	509/808 (63)	-	-	-	-	25/29 (86.2)	236/307 (76.9)	99/119 (83.2)	969/1171 (82.7)	-	-
CCH	76/180* (42.2)	1088/1541 (70.6)	-	-	-	-	57/77 (74)	570/760 (75)	-	-	-	-
CIL	101/241* (41.9)	1830/2357 (77.6)	-	-	-	-	-	-	-	-	-	-
CRK	-	-	425/627 (67.8)	3967/5686 (69.8)	-	-	-	-	-	-	-	-
CMG	128/185 (69.2)	1188/1661 (71.5)	-	-	-	-	-	-	-	-	-	-
CSG	249/479* (52)	2898/4222 (68.6)	-	-	-	-	-	-	-	-	-	-
CCW	-	-	-	-	36/39 (92.3)	265/402 (65.9)	-	-	-	-	-	-

Data is presented as number of SNPs with acceptable HET / total number of SNPs genotyped with a specific protocol-centre combination.
*P_c < 1 x 10^{-6} between set A and B for protocol “PI”

Our most salient finding from the initial analysis of Set A was the high prevalence of non-polymorphic SNPs as determined using protocol "PI". This prevalence was nearly twice as high compared to other protocols. With analysis of the larger dataset, this high prevalence of 39.9% was halved to 19.1%, a dramatic decrease that was statistically highly significant. Probing further, it appeared that the reduction in the prevalence of non-polymorphic SNPs was predominantly attributed to two genotyping centres, each of which saw a near doubling in the prevalence of SNPs with acceptable HET values from Set A (~42%) to Set B (~70%). These changes occurred despite the lack of any obvious differences between Set A and B in terms of the nature of the SNPs and the conformity of the genotype distributions to that expected under HWE. A possible interpretation of our finding is that there may be "pockets" (or "groups") of SNPs that have been inaccurately genotyped in the HapMap database, despite the application of various quality control filters (The International HapMap Consortium 2005).

Clearly, our finding is disconcerting and various reasons may be contemplated including the possibility of some form of bias in terms of SNP selection for data analyses. Arguing against this however is that the distribution of SNPs in both Set A and B according to genotyping protocol is consistent with general information available from HapMap website suggesting that the sampling does seem to give a fair representation of SNPs in the HapMap database. For instance, the percentage of SNPs genotyped using the Third Wave method is expected to be 24.3% (according to online information at the HapMap website) which is very similar to what was observed in our study (28.5% in Set A and 28.7% in Set B); similar comparisons were also made for other protocols. Moreover, we did not observe any significant protocol-specific differences for SNP distribution according to SNP nature and inter-marker distance. It is also reasonable to propose that any existence of bias is more likely to exist in smaller datasets. In this respect, our finding related primarily to protocol "PI" which forms the bulk (>50%) of the genotyping data that was analysed. Taken together, we found no overt evidence to suggest that our findings were spurious due to uncontrolled bias or confounding.

The underlying cause for our observations is not clear and cannot be directly answered without extensive re-genotyping of numerous markers, coupled possibly with the need for direct DNA sequencing. Nevertheless, one may entertain several scenarios. One such possibility is that SNPs in the two datasets differed in terms of the DNA sequence context surrounding each polymorphism. This different sequence context may then impact genotyping, leading to a high prevalence of non-polymorphic calls in Set A but not Set B. However, it is notable that this problem was largely confined to one protocol ("PI"). Furthermore, the effect of genotyping centres on aggregate genotypes was only observed for this protocol but not "PM" which was also used in multiple genotyping centres. For the remaining genotyping protocols that were each utilised in one genotyping centre, there was no significant variation between the two datasets. As a further consideration, one may expect that the effect of DNA sequence context should presumably be moderated by the implementation of computer programs to evaluate beforehand if a particular SNP is suitable for genotyping using each particular method. Additional possibilities that may be considered include alterations to the genotyping protocols in specific centres. These changes may potentially affect the experimental genotyping itself or the criterion for declaring a particular genotype call. In conclusion, our findings have highlighted potential genome centre-specific effects on large scale genotyping in the HapMap project. Oversight of these effects can impact haplotype tagging and the detection of gene-disease associations.

REFERENCES

Farrall M, Morris AP 2005. Gearing up for genome-wide gene-association studies. *Human Molecular Genetics*, **14:** 157-162

Tanaka N, Babazono T, Saito S, Sekine A, Tsunoda T, Haneda M, Tanaka Y, Fujioka T, Kaku K, Kawamori R, Kikkawa R, Iwamoto Y, Nakamura Y, Maeda S 2003. Association of solute carrier family 12 (sodium/chloride) member 3 with diabetic nephropathy, identified by genome-wide analyses of single nucleotide polymorphisms. *Diabetes*, **52:** 2848-2853

The International HapMap Consortium 2003. The International HapMap Project. *Nature*, **426:** 789-796

The International HapMap Consortium 2005. A haplotype map of the human genome. *Nature*, **437:** 1299-1320