European Journal of Human Genetics (2008) 16, 506–515

www.nature.com/ejhg

ARTICLE

Singleton SNPs in the human genome and implications for genome-wide association studies

Xiayi Ke*[,1,2,3], Martin S Taylor[1,4] and Lon R Cardon[1]

[1]*Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK;* [2]*Centre for Integrated Genomic Medical Research (CIGMR), School of Medicine, University of Manchester, Manchester, UK;* [3]*Arthritis Research Campaign (arc) Epidemiology Unit, School of Medicine, University of Manchester, Manchester, UK*

The human genome is estimated to contain one single nucleotide polymorphism (SNP) every 300 base pairs. The presence of LD between SNP markers can be used to save genotyping cost via appropriate SNP tagging strategies, whereas absence or low level of LD between markers generally increase genotyping cost. It is quite common that a large proportion of tagging SNPs in a tagging scheme often turn out to be singleton SNPs, that is, SNPs that only tag themselves rather than contribute power to the rest of a region. If genotyping cost is a major concern, which often is the case at the present time for genome-wide association studies, these singleton tagging SNPs would be the primary targets to be removed from genotyping. It is important, however, to understand the characteristics of such SNPs and estimate the impact of removing them in a study. Using the HapMap genotype data and genome wide expression data, we assessed the distribution and functional implications of singleton SNPs in the human genome. Our results demonstrated that SNPs of potentially higher functional importance (eg, nonsynonymous SNPs, SNPs in splicing sites and SNPs in 5′ and 3′ UTR) are associated with a higher tendency to be singleton SNPs than SNPs in intronic and intergenic regions. We further assessed whether singleton SNPs can be tagged using haplotypes of tagSNPs in the three genome wide chips, that is, GeneChip 500k of Affymetrix, HumanHap300 and HumanHap550 of Illumina, and discussed the general implications on genetic association studies.

European Journal of Human Genetics (2008) **16,** 506–515; doi:10.1038/sj.ejhg.5201987; published online 16 January 2008

Keywords: singleton SNPs; functionality; linkage disequilibrium; genetic association

Introduction

Large scale genome-wide association studies using single nucleotide polymorphisms (SNPs) are now becoming the state of the art in disease genetic studies.[1–7] For these studies, it is still unfeasible to obtain genotype information for every SNP in the human genome by complete resequencing or genotyping despite rapid technological advancement. Often tagging SNP markers are selected based on information on linkage disequilibrium (LD) between SNP markers,[8–11] and this has been greatly facilitated by the International HapMap Project.[12,13]

The human genome is thought to contain one SNP every 100–300 bp. It is now known that LD is not only present between SNPs in close physical proximity along the genome, but is also often present between widely spaced markers to form haplotype blocks.[12,14] The extent of LD determines the number of markers needed to cover a region or the whole genome. Absence or low level of LD between markers, generally increase genotyping cost. This

*Correspondence: Dr X Ke, Centre for Integrated Genomic Medical Research (CIGMR), School of Medicine, University of Manchester, Oxford Road, Manchester, M13 9PT, UK.
Tel: +44 161 2751674; Fax: +44 161 2755043;
E-mail: Xiayi.Ke@manchester.ac.uk
[4]Current address: European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK.
Received 9 February 2007; revised 16 November 2007; accepted 22 November 2007; published online 16 January 2008

is often a problem for many tagging schemes since a large proportion of tagging SNPs turns out to be singleton tagSNP markers, which do not contribute power to the rest of a region at all. If genotyping cost is the primary concern, which may often be the case at the present time for genome-wide association studies, these singleton tagging SNPs would be the primary targets for further removal from the final list of SNPs to be genotyped.

An agnostic view about functionality of SNPs is to regard all SNPs, whether singletons or not, as equally important.[15] Whether this is true is not yet clear, and knowledge about this will therefore help understand the practical implications of removing singleton SNPs from genotyping.

In this study, we investigate the distribution and functional implications of singleton SNPs in the human genome using the phase II HapMap data.[13] Since marker density is an important factor in determining whether or not an SNP is independent of other SNPs, we also assessed the HapMap ENCODE data where SNPs were obtained not only from public databases but also from resequencing (http://www.hapmap.org). Our results demonstrated that singleton SNPs distribute across various functional groups and more interestingly a higher proportion of SNPs that have potentially higher importance (such as nonsynonymous SNPs, SNPs in splicing sites, etc) are singletons than of SNPs in introns and intergenic regions. Various factors are also examined on their effect on singleton status and functionality of an SNP. Furthermore, the capability of current commercial genome-wide chips, such as Affymetrix 500 k and Illumina 300 k, in capturing information on singleton SNPs are assessed and discussed.

Materials and methods

SNPs and their functional categorization

A total of 8 876 160 SNPs and their related information were downloaded from the UCSC genome assembly website (http://genome. ucsc.edu). If an SNP has multiple annotations because of the complexity of human genome such as alternative splicing, classification was carried out in the order: nonsynonymous, synonymous, splicing, 5′UTR, 3′UTR, transcribed, intronic, intergenic. Promoter SNPs were defined as those SNPs within 150 nucleotides from the transcription starting site. Conservation scores were also downloaded from the UCSC genome assembly website (http://genome.ucsc.edu).

Genotype data, LD analysis and definition of singleton SNPs

Phase II HapMap genotype data (autosomes only), including all the 10 ENCODE regions, were downloaded from the International HapMap Project website (http://www.hapmap.org). The datasets included about four million SNPs genotyped in 30 CEPH trios, 45 unrelated Han Chinese from Beijing, 45 unrelated Japanese from Tokyo and 30 Yoruban trios from Nigeria.

Pairwise r^2 values were calculated for all polymorphic SNPs in individual populations. For each polymorphic SNP along a chromosome, the number of marker pairs having r^2 value over a threshold (eg, 0.8) within a window of 1 mb was counted. If there were no such marker pairs for an SNP within the window, that SNP was regarded as a singleton SNP.

Fst values and recombination rates

Unbiased Fst values for SNPs that were genotyped in CEU, CHB + JPT and YRI.[16,17] Recombination rates were downloaded from the HapMap Website (Nov 2005).

Genome-wide expression data and association analysis

The Affymetrix genome-wide expression data was reported by Cheung *et al*,[18] and downloaded from NCBI at accession number of GSE2552, which includes 58 CEPH individuals (57 individuals as reported by Cheung *et al*, plus GM12056). The raw data was processed by MAS5.0 program and $\log_2$ transformed. Within- as well as between-subject variances were then calculated for each gene in the array. A total number of 1516 genes were studied, which had between-subject variance/within-subject variance >2 and had missing trait values in fewer than 5% of the individuals who were selected for genome-wide association analysis. Only genes that had 'Presence (P)' calls for all the 100 samples/replicates of the 58 individuals were used in the study. For individuals with two replicates, their averages were obtained.

Phase II genotypes for the 58 CEPH individuals were obtained from the HapMap website as above. Genotypes were coded as 0, 1, 2 according to how many copies of the minor alleles were present for an individual, and genome-wide association analysis was carried out using regression. SNPs that had a P-value <0.05 after bonferroni correction were taken as potential causal variants (gene expression regulators).

Tagging using single markers and multimarker haplotypes

The method of single marker tagging was based on the algorithm developed by Carlson *et al*,[8] whereas multimarker tagging was based on the algorithm developed by de Bakker *et al*.[9] A 200 kb window size (100 kb on each side of a hidden SNP) was used for the testing.

Results

Singleton SNPs in the human genome and implications for SNP tagging

The number of singleton SNPs in the genome obviously depends on a variety of factors, including populations, sample size, the measure of LD, the LD threshold and

marker density and ascertainment. In this study, we based our estimates on the phase II HapMap population samples (with CHB and JPT combined), and used pairwise r^2 as the measure. As r^2 threshold increased from 0.5 to 0.8 and further to 1.0, the proportion of SNPs being singletons in the genome increased in all populations—from 0.06, 0.14 to 0.29 in CEU; 0.06, 0.13 to 0.34 in CHB + JPT and 0.12, 0.29 to 0.49 in YRI. As the marker density increased, the proportion of SNPs being singletons generally decreased as shown in Figure 1a for CEU samples (bars). It should be noted that the ascertainment schemes have a very important impact on the proportion but also on the composition of singleton SNPs in the genome, as demonstrated in Figure 1a where a much higher proportion of singleton SNPs in ENCODE regions (than in phase I and phase II HapMap) are rare SNP variants with minor allele frequency (MAF) $<5\%$. It is known that both phase I and phase II HapMap tended to bias towards common variants whereas the ENCODE regions were accomplished with resequencing efforts.[12,13]

The most important implication of singleton SNPs in the genome may be their genome typing cost in a genome-wide association study. Because singleton SNPs do not have surrogates to represent them, they themselves have to be genotyped or chosen as tagSNPs in a tagging scheme. As shown in Figure 1b, for phase II HapMap samples at a pairwise tagging threshold of r^2 0.80, at least half of the tagSNPs turned out to be singleton tagSNPs in all the HapMap population samples. These singleton tagSNPs do not contribute power at all to the genome and therefore present as a serious cost constraint for a study.

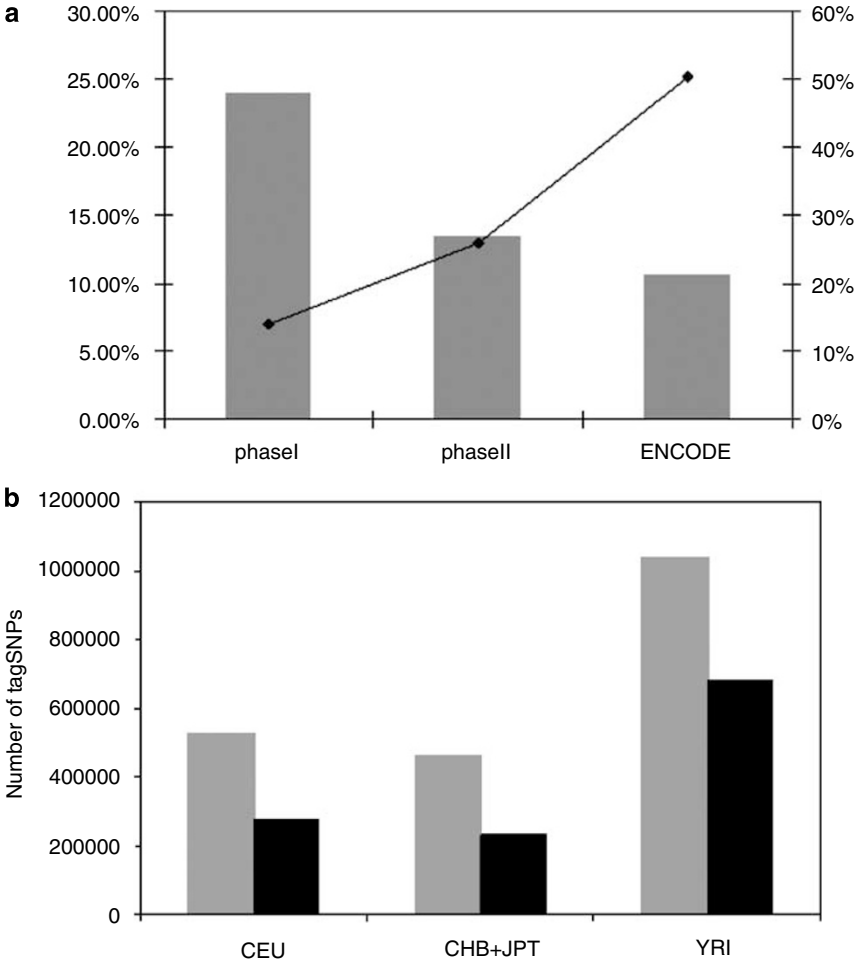

Figure 1 Singleton SNPs in the human autosomal genome. (**a**) Estimate of proportion of SNPs being singleton in the human genome using HapMap phase I, phase II and ENCODE regions respectively. The primary *y* axis and the columns denote the proportion of SNPs being singletons, whereas the secondary *y* axis and line denote the proportion of rare SNPs among the singleton SNPs. (**b**) Singleton tagSNPs in tagSNP selection using phase II HapMap data (pairwise $r^2 \geq 0.80$). Grey bars denote the total number of tagSNPs needed to capture all common variants (MAF $>5\%$). Black bars denote the number of singleton tagSNPs.

Distribution and functional implications of singleton SNPs

It is tempting to exclude singleton SNPs during marker selection to save genotyping cost. Among the 8 876 160 SNPs surveyed in this study, 0.51% of them are nonsynonymous SNPs, 0.46% are synonymous SNPs, 0.04% in splicing sites, 0.2% in 5′ UTR and 0.84% in 3′ UTR, with the majority located either within introns (49.78%) or intergenic regions (46.83%). Nonsynonymous SNPs, SNPs in splicing sites, 5′ UTR and 3′ UTR are presumably more likely to have functional consequences than intronic or intergenic SNPs. Although these groups of SNPs are only a small minority genome wide, a significant higher proportion of them turned out to be singleton SNPs than the intronic and intergenic SNPs (Figure 2a). SNPs were also classified according to whether they locate within a region showing significant cross-species sequence conservation (4.66%), or not (95.34%). Again, it is interesting to observe that there are always a higher proportion of SNPs in conserved regions being singletons than in non-conserved regions (Figure 2b).

A higher proportion of nonsynonymous SNPs, SNPs in splicing sites, SNPs in 5′ UTR and 3′ UTR are singletons than intronic and intergenic SNPs, and this observation is consistent across different minor allele frequency spectrum (Figure 3a). A higher proportion of SNPs in conserved regions were also found to be singletons than of SNPs in non-conserved regions (Figure 3b). It is expected that singleton SNPs are generally located in regions of higher recombination rate than non-singletons, as indeed observed in the recent paper by the International HapMap Consortium,[13] where untaggable SNPs were found strongly

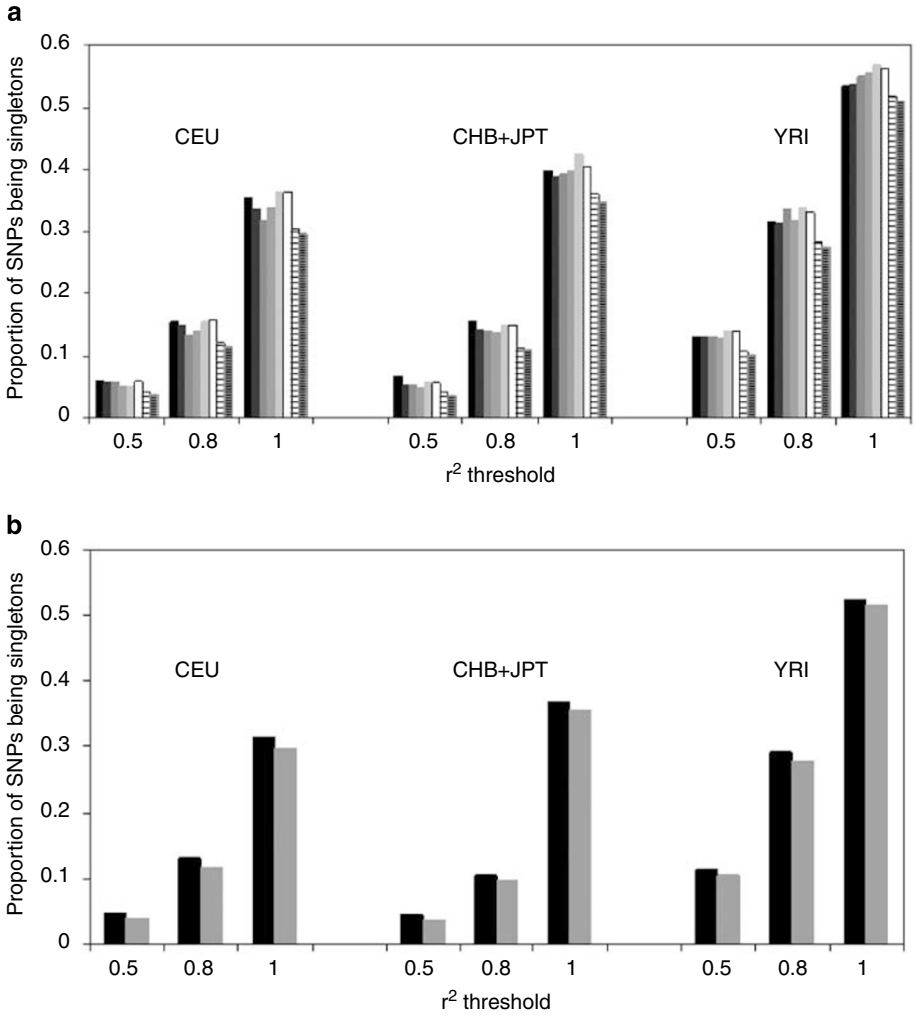

Figure 2 Distribution of singleton SNPs in the genome. HapMap phase II data were used for the analysis. (**a**) Proportion of SNPs being singletons at pairwise r^2 threshold 0.50, 0.80 and 1.0 in different functional groups. Black bars denote nonsynonymous SNPs; grey bars with decreasing intensity denote synonymous SNPs, SNPs in splicing sites, SNPs in transcribed regions, SNPs in 5′ UTR; bars with black border and white fill denotes SNPs in 3′ UTR; bars with black dashes denotes SNPs in introns; bars with grey dashes denote SNPs in intergenic regions. (**b**) Proportion of SNPs being singleton SNPs at pairwise r^2 threshold 0.50, 0.80 and 1.0 in conserved (black bars) and non-conserved (grey bars) regions.

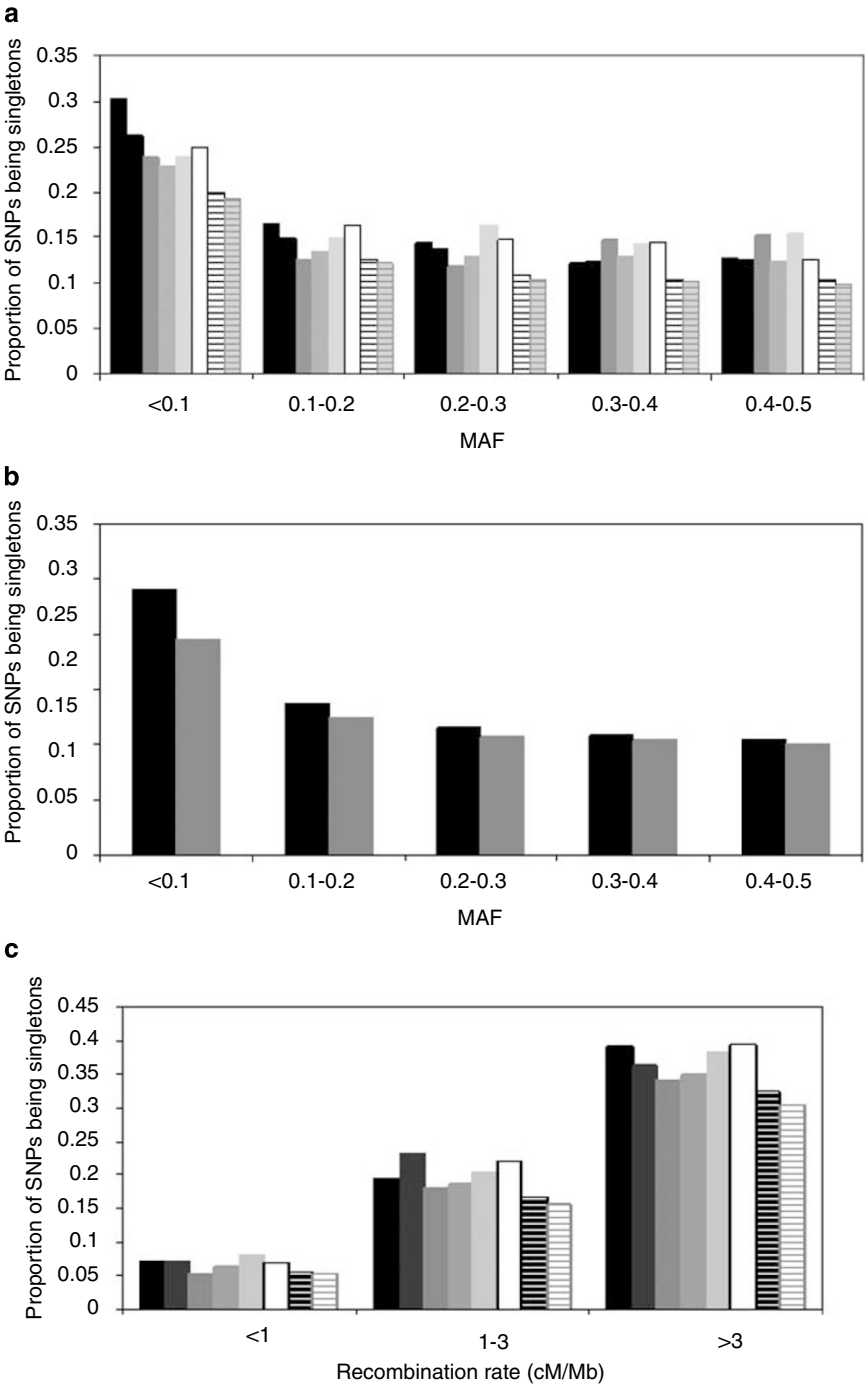

Figure 3 Distribution of singleton SNPs at different MAF and at different recombination rates. Singleton SNPs were defined using phase II HapMap-CEU samples with r^2 threshold at 0.80. (**a**) Distribution of singleton SNPs of different functional groups at different MAF spectrum. (**b**) Distribution of singleton SNPs in conserved *vs* non-conserved regions at different MAF spectrum. (**c**) Distribution of singleton SNPs of different functional groups at regions of high (>3 cM/Mb), intermediate (1–3 cM/Mb) and (<1 cM/Mb) recombination rates. In (**a** and **c**) black bars denote nonsynonymous SNPs; grey bars with decreasing intensity denote synonymous SNPs, SNPs in splicing sites, SNPs in transcribed regions, SNPs in 5′ UTR; bars with black border and white fill denotes SNPs in 3′ UTR; bars with black dashes denotes SNPs in introns; bars with grey dashes denote SNPs in intergenic regions. In (**b**) black bars denote SNPs in conserved regions and grey bars SNPs in non-conserved regions.

enriched in recombination hotspots. By separating SNPs into three different groups according to their recombination rates (<1.0, 1–3, >3 cM/Mb), SNPs distribution was assessed. As shown in Figure 3c, in all situations, the same pattern was observed. SNPs located in promoter regions (150 bp from transcription starting sites) were also examined separately. At r^2 threshold of 0.8, 17% of SNPs located in promoter regions are singletons for CEU, 19% for CHB + JPT and 34% for YRI.

It is well known that African populations generally have lower LD than Caucasians and Asians. So it is not surprising to observe that YRI have a higher proportion of overall singleton SNPs than CEU, CHB + JPT (Figure 2). Among SNPs that are polymorphic in all three populations and singletons in at least one population, a large number is shared among them (Figure 4a). Unsurprisingly, singleton SNPs shared among populations were found to be associated with the highest recombination rates, whereas those population-specific ones tended to locate in regions with lower recombination rates (Figure 4b). Consistent with observations made above and in particular with Figure 3c, SNPs of potentially higher functional importance were associated with a higher proportion of being singletons in all situations, that is, whether they were singletons shared by all three populations, shared by only two populations or observed in only one population (Figure 4c).

Factors affecting functional implications of singleton SNPs

Singleton status was defined by pairwise LD, which is known to be affected by various factors such as recombination, selection and mutation. Effect of these factors, therefore, need to be accounted for in assessing the functional implications of singleton SNPs. For this purpose, we separate SNPs into two distinct groups: non-synonymous SNPs, synonymous SNPs, SNPs in splicing sites, SNPs in 5′utr and 3′utr and SNPs being transcribed were included all in one group which was regarded as functionally more important, whereas the other group was composed of intronic SNPs and SNPs in intergenic regions. MAF, recombination rate and Fst value of a SNP were first examined on their predicting capability on singleton status (Table 1, upper panel). Although MAF had a very strong impact on singleton status with rare SNPs (MAF <0.1), recombination rate and Fst were found to be more consistently associated with singleton status across different MAF spectrum (Table 1, upper panel).

These three factors were then assessed together with singleton status on their predicting capability on functional status. As shown in Table 1 (lower panel), singleton status of a SNP was the most persistent predictor of functional status. Both its strength of association (with functional status) and level of significance were consistent across different MAF spectrum. When only MAF and singleton status were used as the predictors in the regression model, singleton status was always more significantly associated with functional status than MAF, with the latter only showing significant association with less common SNPs (MAF <0.3) (data not shown). These results demonstrated that overall singleton SNPs in the human genome tended to locate in functionally important regions more often than non-singleton SNPs.

Singleton SNPs as potential causal variants of natural variation of gene expression

We carried out allelic association testing for 2418 genes which had expression data (ie, presence calls) in at least 95% of the individuals in GSE2552 dataset[18] and with between-subject-variance/within-subject variance >2.0. A genome-wide bonferroni correction was applied to the association testing. Although SNPs with $P<0.05$ after correction were not necessarily all causal variants which regulate the expression of genes under study, they nevertheless represented a pool of the most likely causal variants. This pool of candidates of gene expression regulators distributed across different regions of the genome, including introns and intergenic regions. Among this pool, more than 10% of SNPs turned out to be singleton SNPs, which was about the genome-wide average (Figure 1). For genes with more than 10 significant SNPs, the proportion of these SNPs being singletons was about 6%, and the figure was about 12% for genes with 2–10 significant SNPs. For genes with only one significant SNP, expectation would be that the majority of the significant SNPs were singletons, but the ratio was about 40%, indicating a very large proportion of the significant SNPs were non-singletons, but their surrogates just failed the corrected *P*-value threshold.

The power of commercial chips on detecting singleton SNPs in the human genome

We have used the genotype data of the ENCODE regions to assess the potential of using haplotypes formed by non-singleton tags to tag singletons. Almost all the singleton SNPs were detectable (haplotype $r^2>0.80$) with 2-marker predictors formed between non-singleton tags within a 200 kb window (data not shown). Various other studies also showed that by using haplotypes formed from multimarker predictors, there was always a big gain in the power of tagSNP sets and tagging efficiency.[3,9] Here we mainly assess the power on singleton SNPs of the current genome-wide chips, that is, 500 k GeneChip Mapping Sets of Affymetrix and the Sentrix HumanHap300 and HumanHap550 BeadChip by Illumina.

GeneChip 500 k is a random SNP set,[19] whereas HumanHap300 is a tagSNP set using LD information of the phase I CEPH HapMap.[20,21] HumanHap550 is currently under development and is using LD information of the phase II HapMap (http://www.illumina.com). These three products contain 504 152, 317 503 and 561 287 SNPs, respectively.

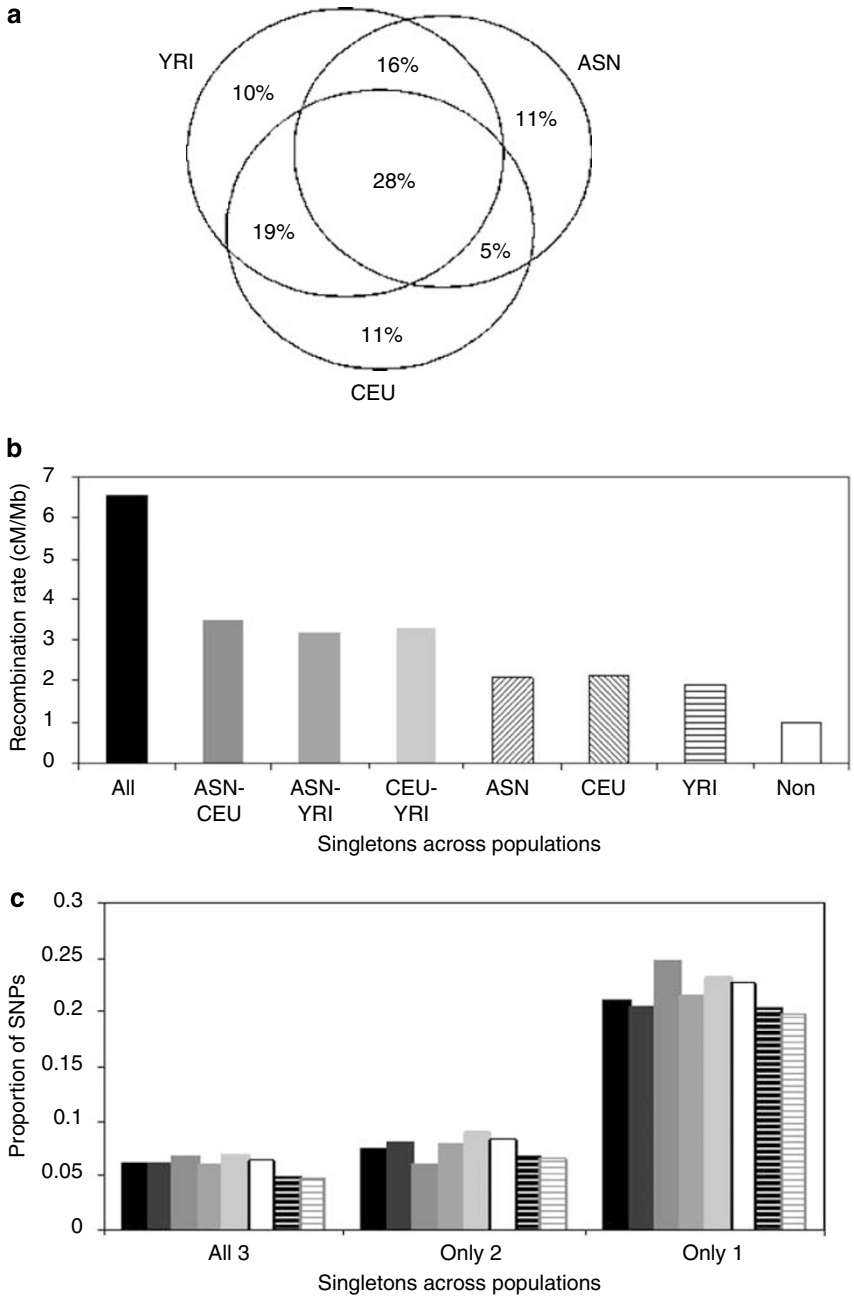

Figure 4 Comparison of singleton SNPs between populations. HapMap phase II data were used and only SNPs, which were polymorphic in all three populations and had MAF > 5% in at least one population were included to define singleton SNPs with r^2 threshold at 0.80. (**a**) Singleton SNPs shared between populations (ASN = CHB + JPT, CEU and YRI). (**b**) Average recombination rates of singleton SNPs shared between populations. Black bar denotes SNPs that were singletons in all the three populations; dark grey bar denotes SNPs being singletons in both CHB + JPT and YRI; medium-light grey bar denotes SNPs being singletons in both CEU and YRI; light grey bar denotes SNPs being singletons in both CEU and CHB + JPT; bar with forward slashes denotes SNP being singletons in CEU only; bar with back slashes denotes SNP being singletons in CHB + JPT only; bar with dashes denotes SNP being singletons in YRI only; white bar denotes SNPs being non-singletons in all three populations. (**c**) Distribution of singleton SNPs of different functional groups across populations. black bars denote nonsynonymous SNPs; grey bars with decreasing intensity denote synonymous SNPs, SNPs in splicing sites, SNPs in transcribed regions, SNPs in 5′ UTR; bars with black border and white fill denotes SNPs in 3′ UTR; bars with black dashes denotes SNPs in introns; bars with grey dashes denote SNPs in intergenic regions. 'All 3' denotes singletons in all three populations; 'Only 2' denotes singletons in only two of the three populations; 'Only 1' denotes singletons in only one of the three populations.

Table 1 Singleton status and functional implications. A total of 1 765 115 SNPs from HapMap–CEU samples were used for this analysis

	Prediction of singleton status					
MAF	<0.1	0.1–0.2	0.2–0.3	0.3–0.4	0.4–0.5	All
MAF (as continuous variable)	OR = 12.7 (8.62–18.6); $P = 2 \times 10^{-38}$	OR = 1.24 (0.85–1.80); NS	OR = 1.16 (0.78–1.72); NS	OR = 1.03 (0.69–1.54); NS	OR = 0.79 (0.53–1.18); NS	OR = 1.74 (1.68–1.80); $P = 1 \times 10^{-212}$
Recombination rate (*cM/Mb*)	OR = 0.89 (0.89–0.89); $P = 0$	OR = 0.87 (0.86–0.87); $P = 0$	OR = 0.86 (0.86–0.86); $P = 0$	OR = 0.86 (0.86–0.86); $P = 0$	OR = 0.86 (0.86–0.86); $P = 0$	OR = 0.87 (0.87–0.87); $P = 0$
Fst	OR = 3.42 (3.13–3.72); $P = 1 \times 10^{-171}$	OR = 1.41 (1.29–1.54); $P = 2 \times 10^{-13}$	OR = 1.29 (1.17–1.43); $P = 4 \times 10^{-7}$	OR = 1.27 (1.15–1.41); $P = 4 \times 10^{-6}$	OR = 1.49 (1.34–1.65); $P = 6 \times 10^{-14}$	OR = 1.69 (1.61–1.76); $P = 2 \times 10^{-127}$
	Prediction of functional status					
MAF	<0.1	0.1–0.2	0.2–0.3	0.3–0.4	0.4–0.5	All
MAF (as continuous variable)	OR = 1.82 (0.93–3.58); NS	OR = 1.721 (0.92–3.22); NS	OR = 2.13 (1.12–4.07); $P = 0.02$	OR = 1.09 (0.57–2.10); NS	OR = 1.06 (0.56–2.03); NS	OR = 1.19 (1.12–1.26); $P = 9 \times 10^{-9}$
Recombination rate (*cM/Mb*)	OR = 1.00 (0.99–1.01); NS	OR = 1.01 (1.00–1.02); $P = 4 \times 10^{-4}$	OR = 1.01 (1.01–1.02); $P = 9 \times 10^{-6}$	OR = 1.01 (1.00–1.02); $P = 5 \times 10^{-4}$	OR = 1.00 (0.99–1.01); NS	OR = 1.01 (1.01–1.01); $P = 1 \times 10^{-9}$
Fst	OR = 1.21 (1.05–1.39); $P = 0.009$	OR = 0.94 (0.81–1.09); NS	OR = 0.97 (0.83–1.14); NS	OR = 0.83 (0.71–0.98); $P = 0.03$	OR = 0.78 (0.66–0.92); $P = 0.003$	OR = 0.95 (0.89–1.02); NS
Singleton status	OR = 1.30 (1.23–1.37); $P = 5 \times 10^{-23}$	OR = 1.28 (1.21–1.35); $P = 4 \times 10^{-17}$	OR = 1.40 (1.32–1.49); $P = 5 \times 10^{-30}$	OR = 1.37 (1.29–1.45); $P = 1 \times 10^{-24}$	OR = 1.29 (1.22–1.37); $P = 9 \times 10^{-18}$	OR = 1.33 (1.29–1.36); $P = 1 \times 10^{-103}$

A threshold of $r^2 = 0.8$ was used to define singleton status. Nonsynonymous SNPs, synonymous SNPs, SNPs in splicing sites, SNPs in 5′-utr and 3′-utr, SNPs being transcribed were regarded as functionally more important and grouped together, as compared to the group composed of SNPs in intergenic regions and intronic SNPs. Minor allele frequency (MAF), recombination rate and Fst were used as continuous variables. Odds ratios (ORs) and their corresponding 95% confidence intervals as well as *P*-values were obtained by multiple logistic regression.
NS – not significant at $P < 0.05$.

For phase II HapMap CEU samples, 13, 26 and 47% of singleton SNPs, defined at r^2 threshold 0.8 with MAF $\geq 5\%$, in the autosomes are included into the three products, respectively, well reflecting the nature of these products (Figure 5a). Given that only about 520 000 SNPs are needed to tag all the autosome common SNPs of the phase II CEU HapMap (Figure 1b), it is somewhat surprising that only 47% of CEU singletons are included in HumanHap550. This is perhaps due to the fact that HumanHap550 is aimed for more diverse markets other than designed solely for Caucasians as in the case of HumanHap300.

For singleton SNPs that are not included as tags, we used multimarker predictors of up to 3 tagSNPs (ie, those SNPs present in the products) to increase the performance of the products. 31, 34 and 31% of these SNPs are captured by GeneChip 500 k, HumanHap300 and HumanHap550, respectively, giving the total power over all singleton SNPs at 44, 60 and 78% (Figure 5b). There was virtually no gain by increasing the number of markers for the multimarker predictors or by increasing the window sizes (data not shown) and this suggest that the remaining recalcitrant singleton SNPs either locate in or close to recombination hotspots and therefore become untaggable as observed by the International HapMap Consortium,[13] or are too distant from their nearest tagSNPs available in the products.

Discussion

Singleton status was defined by LD, and LD is known to be affected by various factors, such as recombination, selection and mutation. It is known that both positive selection for advantageous variants and purifying selection against deleterious variants can increase LD.[22] As functionally more important regions, such as genic regions and conserved regions, are subject to more selection pressure, the level of LD in these regions is expected to be higher than in functionally less important regions. In the context of this study, a lower frequency of singleton SNPs is expected to be observed in the functionally more important regions. The finding of this study was, however, the opposite. Previous studies showed that repeat sequence can inflate LD in a genomic scale,[12,23] and lack of interspersed repeat elements were indeed found to be the reason behind weaker LD in evolutionarily more conserved regions.[24] This was consistent with singleton SNP distribution observed in this study.

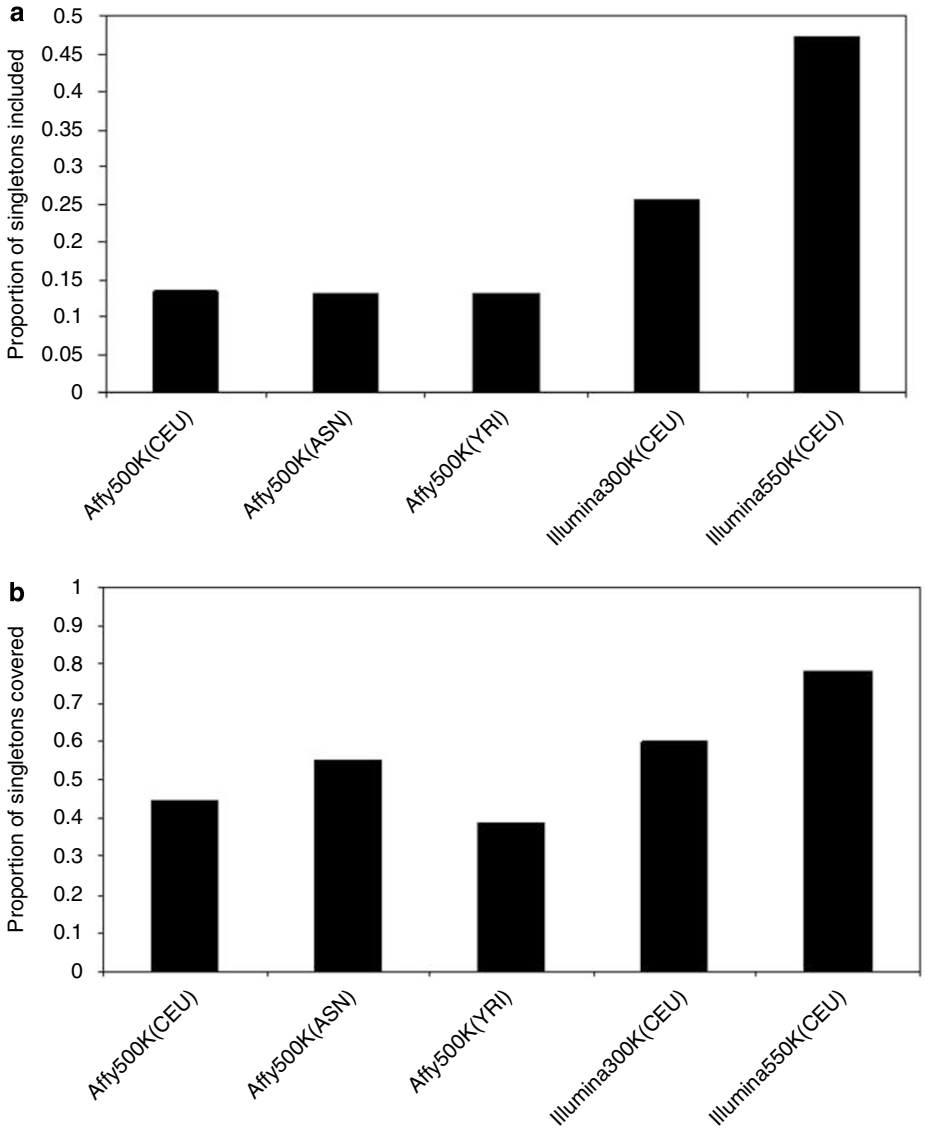

Figure 5 Performance of GeneChip 500 k, HumanHap 300 and HumanHap550 on detecting singleton SNPs. Singleton SNPs were defined at r^2 threshold 0.8 for the phase II HapMaps and only common SNPs (MAF $\geq$5%) were studied. (**a**) Proportion of singleton SNPs being included in the products. (**b**) The proportion of singleton SNPs covered, that is either included in the products or being captured (haplotype $r^2 \geq 0.80$) by multimarker predictors of up to 3 tagSNPs in the products. ASN denotes the combined CHB and JPT HapMap samples.

Purifying selection is expected to produce rarer SNPs in functionally important regions. The HapMap is also known to be biased towards common SNPs and therefore led no or fewer proxies for rare SNPs. As a result, singleton SNPs were expected to have generally low MAF, or at least any specific association of singleton status with SNP functionality was expected to be seriously confounded by its association with rare SNPs. Our genome-wide survey revealed that, after effect of rare SNPs was accounted for, singleton SNPs in the human genome tend to locate in regions of functionally high importance more often than non-singleton SNPs. More specifically, a higher proportion of nonsynonymous SNPs, SNPs in splicing sites, SNPs in 5′ and 3′ UTR and SNPs in promoter regions were identified to be singleton SNPs than of intronic and intergenic SNPs in the genome. It was also observed that recombination rate was a better predictor of singleton status than MAF and that increased proportion of singleton SNPs among the functionally more important groups coincided with increased recombination rate.

It is not clear whether there is any difference between singleton and non-singleton SNPs in the intergenic or intronic regions in terms of their functionality. A comprehensive assessment of this difference is obviously difficult. The list of significant SNPs identified in the genome-wide association analysis of gene expression data represented a pool of possible candidates of gene expression regulators.

Many of these SNPs located in intergenic and intronic regions, but overall no difference was observed between proportion of singleton SNPs among this pool and that of the genome-wide average.

The present study was mainly carried out with the phase II HapMap data. Although the number of SNPs in the 5 Mb ENCODE regions prevented a detailed stratified analysis, similar patterns also emerged. For example, for HapMap-CEU, at r^2 threshold of 0.8 and MAF > 5%, 14% of nonsynonymous SNPs were found to be singletons, whereas this figure dropped to about 6% for SNPs in intronic and intergenic regions. More importantly, the observations made with the HapMap phase II data is of practical importance as the HapMap continues to serve as one of the most important sources of information for study design.

These observations have some important implications for genome-wide association studies, as LD information derived from the phase II HapMap data continues to be used for marker selection in the foreseeable future. More than 320 000 polymorphic SNPs in the phase II CEU and CHB + JPT HapMaps are singleton SNPs (r^2 threshold at 0.8), more than 70% of which are common SNP variants, whereas for YRI HapMap the figure is about 800 000 with more than 80% being common variants. Based on estimates from the ENCODE regions, if all polymorphic SNPs are genotyped, there would possibly be about 700 000 singleton SNPs in CEU samples, half of which are common SNP variants. For YRI samples, the figure is more than doubled to 1 500 000 SNPs with 65% of them being common variants.

For singleton SNPs located in functionally more important regions, such as coding and conserved regions, they can always be explicitly included into a genotyping scheme on top of the set of markers selected based on an agnostic approach.[1,3] For other singleton SNPs, haplotype information may be used to improve the power to detect them, as shown by the performance of GeneChip 500 k of Affimetrix and HumanHap300 and HumanHap550 of Illumina. There are still, however, a large proportion of singleton SNPs not properly covered by the chips, especially GeneChip 500 k and HumanHap300. Even by increasing the genome coverage of these commercial chips, some of these SNPs will remain untaggable unless included directly, as suggested by the International HapMap Consortium.[13] For singletons that can be tagged by multimarker predictors, association testing may be complicated by issues, such as how to choose a predictor from potentially multiple alternative sets and dependence on the successful genotyping of predictor markers. As genotyping cost continues to drop, therefore, it may become more attractive to genotype singleton SNPs directly. Under such a scheme, for less common singleton SNPs (such non-singleton SNPs can still be tagged by more commoner SNPs), however, it may still be desirable to exclude them in a disease association study, as there is generally less power to detect their effects.

References

1 Barrett JC, Cardon LR: Evaluating coverage of genome-wide association studies. *Nat Genet* 2006; **38**: 659–662.
2 Evans DM, Cardon LR: Genome-wide association: a promising start to a long race. *Trends Genet* 2006; **22**: 350–354.
3 Pe'er I, de Bakker PI, Maller J *et al*: Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat Genet* 2006; **38**: 663–667.
4 Smyth DJ, Cooper JD, Bailey R *et al*: A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. *Nat Genet* 2006; **38**: 617–619.
5 Frayling TM, Timpson NJ, Weedon MN *et al*: A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 2007; **316**: 889–894.
6 Zeggini E, Weedon MN, Lindgren CM *et al*: Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* 2007; **316**: 1336–1341.
7 The Wellcome Trust Case Control Consortium: Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature* 2007; **447**: 661–678.
8 Carlson CS, Eberle MA, Rieder MJ *et al*: Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* 2004; **74**: 106–120.
9 de Bakker PI, Yelensky R, Pe'er I *et al*: Efficiency and power in genetic association studies. *Nat Genet* 2005; **37**: 1217–1223.
10 Ke X, Miretti MM, Broxholme J *et al*: A comparison of tagging methods and their tagging space. *Hum Mol Genet* 2005; **14**: 2757–2767.
11 Iles MM: The effect of SNP marker density on the efficacy of haplotype tagging SNPs – a warning. *Ann Hum Genet* 2005; **69**: 209–215.
12 The International HapMap Consortium: A haplotype map of the human genome. *Nature* 2005; **437**: 1299–1320.
13 The International HapMap Consortium: A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007; **449**: 851–861.
14 Gabriel SB, Schaffner SF, Nguyen H *et al*: The structure of haplotype blocks in the human genome. *Science* 2002; **296**: 2225–2229.
15 Carlson CS: Agnosticism and equity in genome-wide association studies. *Nat Genet* 2006; **38**: 605–606.
16 Weir BS, Cockerham CC: Estimating F-statistics for the analysis of population structure. *Evolution* 1984; **38**: 1358–1370.
17 Weir BS: *Genetic Data Analysis II*. Sunderland, MA: Sinauer Associates, 1996, pp 161–173.
18 Cheung VG, Spielman RS, Ewens KG *et al*: Mapping determinants of human gene expression by regional and genome-wide association. *Nature* 2005; **437**: 1365–1369.
19 Matsuzaki H, Dong S, Loi H *et al*: Genotyping over 100 000 SNPs on a pair of oligonucleotide arrays. *Nat Methods* 2004; **1**: 109–111.
20 Gunderson KL, Steemers FJ, Lee G *et al*: A genome-wide scalable SNP genotyping assay using microarray technology. *Nat Genet* 2005; **37**: 549–554.
21 Steemers FJ, Chang W, Lee G *et al*: Whole-genome genotyping with the single-base extension assay. *Nat Methods* 2006; **3**: 31–33.
22 Ardlie KG, Kruglyak L, Seielstad M: Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet* 2002; **3**: 299–309.
23 Smith AV, Thomas DJ, Munro HM *et al*: Sequence features in regions of weak and strong linkage disequilibrium. *Genome Res* 2005; **15**: 1519–1534.
24 Kato M, Sekine A, Ohnishi Y *et al*: Linkage disequilibrium of evolutionarily conserved regions in the human genome. *BMC Genomics* 2006; **7**: 326.