Table 1 Haplotypes spanning *CYP2C19* and *CYP2C9* in Europeans and Han Chinese

Gene	SNP	Allele	Haplotype frequencies in Europeans (%)									Haplotype frequencies in Han Chinese (%)										
			21.7	16.6	10.0	9.1	8.2	6.7	5.0	3.3	2.5	24.6	17.7	6.0	4.8	4.8	4.7	3.8	2.7	2.7	2.5	2.4	2.4
CYP2C19	Rs4986893	*3										1	1	1	1	1	1	1	1	1	1	2
	Rs4244285	*2A, *2B	1	1	1	2	1	1	2	1	1	1	2	1	1	1	1	2	1	1	1	1
	Rs3758581	*1B, *1C, *2A	1	1	1	1	1	1	1	2	1	1	1	1	2	1	1	1	1	1	1	1
CYP2C9	Rs1057910	*2	1	1	1	1	1	1	1	2	1	1	1	1	2	1	1	1	1	1	1	1

in **Table 1**. Roughly 50% of Europeans share common haplotypes leading to the extensive metabolism phenotype for both *CYP2C19* and *CYP2C9*. The data suggest that these metabolic phenotypes will occur together in individuals more often than expected by chance and that, except for rare recombination events, metabolic status for the two cytochromes will be inherited together in families.

We suggest that a minimum SNP set taking into account the underlying asymmetric haplotype block structure of the *CYP2C* genomic region will be more likely to capture genetic variation in the cytochromes and will therefore be a more useful tool for research and for potential clinical applications.

Robert Walton[1,2], Martin Kimber[3,4], Kirk Rockett[3], Clare Trafford[3], Dominic Kwiatkowski[3] & Giorgio Sirugo[1]

[1]Medical Research Council Laboratories, PO Box 273, Fajara, The Gambia. [2]Department of Clinical Pharmacology, University of Oxford, Radcliffe Infirmary, Woodstock Road, Oxford OX2 6HE, UK. [3]Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Headington, Oxford OX3 7BN, UK. [4]Tessella Support Services plc, Abingdon, Oxfordshire, UK. Correspondence should be addressed to R.W. (rwalton@mrc.gm).

1. Ahmadi, K.R. *et al.* A single-nucleotide polymorphism tagging set for human drug metabolism and transport. *Nat. Genet.* **37**, 84–89 (2005).

In reply:

Walton *et al.* discuss two points regarding the selection of tagging SNPs (tSNPs) for genetic association studies. The first relates to the genomic boundaries within which tSNPs are selected. The computational constraints on multiple-marker methods of selecting tSNPs make it impractical to consider contiguous stretches of sequence beyond a certain size; beyond this, it is necessary to follow some scheme to subdivide the region. Walton *et al.* note that the division selected for one of our regions, the polygenic *CYP2C* region, will influence the tSNP efficiency (*i.e.*, minimizing the number of tSNPs needed to achieve a desired level of power against untyped causal variants). Second, Walton *et al.* note that there may be long-range haplotypes extending across genes in the *CYP2C* cluster that may form functional units. As general principles, we agree with both these points. In fact, we were among the first to show that combining smaller regions into single large ones, where possible, led to considerable increases in SNP tagging efficiency[1]. We also think that the occurrence of two or more genes in strong linkage disequilibrium (LD) could make it useful to consider intergene haplotypes as integrated units in association studies.

Both these areas are worthy of further evaluation. In our view, there has not yet been any thorough investigation of optimal approaches for subdividing regions in such a way as to maximize the performance of tSNPs. It is a difficult problem because of the inherent trade-off between tagging efficiency and the size of the region being tagged. The larger the region is, the more one can capitalize on LD[1] but the more difficult it becomes to infer haplotypes accurately. There is also a danger of overfitting with haplotype-based models in which the degrees of freedom of the model can grow rapidly with the number of tSNPs required. There comes a point at which regions become too large to tag using aggressive multiple-marker methods. It is, therefore, uncertain how best to set break points between regions in such situations. For example, Walton *et al.* base their assessment on using point estimates of D′. In contrast, the method of Gabriel *et al.*[2], in which the degree of confidence in a given D′ value is used in the assessment of an LD block, does not define an extended LD block covering both *CYP2C19* and *CYP2C9* (see Fig. 2 in ref. 3). There are, however, no *a priori* grounds for assuming either of these approaches will result in an optimal subdivision of the region in terms of optimizing tagging performance (and almost certainly they do not). In the particular case of the *CYP2C* region, however, we note that using the subdivision suggested by Walton *et al.* results in slightly fewer tSNPs than does using the arbitrary subdivision that we used. Finally, we note that our study[3] used just one of many possible tSNP selection methods (based on haplotype r^2), and the relative power of different methods for different genomic regions is a further unresolved issue.

In summary, some method of subdividing regions in order to maximize tSNP efficiency would be useful, but the best method for doing this is currently not known. Similarly, consideration of longer-range functional haplotypes spanning more than one gene has not been given much attention in the tSNP literature. Both these areas are worthy of further consideration, and Walton *et al.* are right to call attention to them.

Kourosh R Ahmadi[1], Michael E Weale[2], Allen D Roses[3], Ann M Saunders[3] & David B Goldstein[4]

[1]Department of Biology (Galton Lab) and [2]Department of Medicine (Institute for Human Genetics and Health), University College London, The Darwin Building, Gower Street, London WC1E 6BT, UK. [3]Genetics Research, GlaxoSmithKline, PO Box 13398, Five Moore Drive, Research Triangle Park, North Carolina 27709, USA. [4]Institute for Genome Sciences and Policy, Duke University, CIEMAS, 101 Science Drive, Box 3382, Durham, North Carolina 27708, USA. Correspondence should be addressed to K.R.A. (k.ahmadi@ucl.ac.uk) or D.B.G. (d.goldstein@duke.edu).

1. Goldstein, D.B., Ahmadi, K.R., Weale, M.E. & Wood, N.W. Genome scans and candidate gene approaches in the study of common diseases and variable drug responses. *Trends Genet.* **19**, 615–622 (2003).
2. Gabriel, S.B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
3. Ahmadi, K.R. *et al.* A single-nucleotide polymorphism tagging set for human drug metabolism and transport. *Nat. Genet.* **37**, 84–89 (2005).