European Journal of Human Genetics (2006) 14, 1209–1214

www.nature.com/ejhg

ARTICLE

The value of gene-based selection of tag SNPs in genome-wide association studies

Steven Wiltshire[1], Paul IW de Bakker[2,3,4,5] and Mark J Daly*[,2,5,6]

[1]*Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK;* [2]*Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA, USA;* [3]*Department of Molecular Biology, Massachusetts General Hospital, Boston, MA, USA;* [4]*Department of Genetics, Harvard Medical School, Boston, MA, USA;* [5]*Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA;* [6]*Department of Medicine, Harvard Medical School, Boston, MA, USA*

Genome-wide association scans are rapidly becoming reality, but there is no present consensus regarding genotyping strategies to optimise the discovery of true genetic risk factors. For a given investment in genotyping, should tag SNPs be selected in a gene-centric manner, or instead, should coverage be optimised based on linkage disequilibrium alone? We explored this question using empirical data from the HapMap-ENCODE project, and we found that tags designed specifically to capture common variation in exonic and evolutionarily conserved regions provide good coverage for 15–30% of the total common variation (depending on the population sample studied), and yield genotype savings compared with an anonymous tagging approach that captures all common variation. However, the same number of tags based on linkage disequilibrium alone captures substantially more (30–46%) of the total common variation. Therefore, the best strategy depends crucially on the unknown degree to which functional variation resides in recognisable exons and evolutionarily conserved sequence. A hypothetical but reasonable scenario might be one in which trait-causing variation is equally distributed between exons plus conserved sequence, and the rest of the genome. In this scenario, our analysis suggests that a tagging approach that captures variation in exons and conserved sequence provides only modestly better coverage of putatively causal variation than does anonymous tagging. In HapMap CEU samples (with northern and western European ancestry), we observed roughly equivalent coverage for equal investment for both tagging strategies.
European Journal of Human Genetics (2006) **14,** 1209–1214. doi:10.1038/sj.ejhg.5201678; published online 28 June 2006

Keywords: tag SNP selection; association studies; linkage disequilibrium; evolutionarily conserved regions

Introduction

Significant efforts have been made to characterise common genetic variation throughout the human genome, such as the International HapMap Project,[1,2] and to examine gene-based variation at greater depth through resequencing.[3] Yet, our understanding of the genetic basis of complex traits and common disease remains far from complete. Considerable advances in SNP genotyping technology have led to genome-wide association studies of complex traits becoming a realistic prospect.[4] There is considerable divergence of opinion, however, regarding the optimal approach to selecting markers to capture the genetic variation underlying complex traits. These views range from screens of 'anonymous' SNPs across the genome chosen solely on the basis of regional patterns of linkage disequilibrium (LD) to those focusing explicitly on SNPs in protein coding or evolutionarily conserved regions.[5] The

*Correspondence: Dr MJ Daly, Center for Human Genetic Research, Massachusetts General Hospital, 185 Cambridge Street, CPZN-6818, Boston, MA 02114-2790, USA.
Tel: +1 617 643 3290; Fax: +1 617 643 3293;
E-mail: mjdaly@chgr.mgh.harvard.edu
Received 7 February 2006; revised 2 May 2006; accepted 2 May 2006; published online 28 June 2006

paucity of confirmed complex-trait susceptibility genes thus far identified precludes a definitive conclusion as to what may be considered the best approach. Intuitively, the best approach will largely depend on the assumed distribution of causal variation between coding and conserved regions, and the rest of the genome.

Here, we address the more technical aspect of the genotyping effort involved in these two alternative approaches using empirical data from the HapMap-ENCODE project.[2] This resource contains 17 944 SNPs (one SNP per 279 bp) genotyped in the HapMap DNA samples across ten 500 kb regions, and can be considered near-complete with respect to common variation (SNPs with $\geq 5\%$ frequency).[2] Collectively, these 10 regions are representative of the genome in terms of gene density and nonexonic conservation.[2] We have determined the number of tag SNPs needed to capture the common genetic variation for several gene-based tagging approaches, evaluated the amount of total common variation captured by these tags, and finally, evaluated one of these gene-based tagging strategies in the realistic scenario where genotyping resources (ie, number of tag SNPs) are considered fixed.

Materials and methods

Data sets

We used phased genotype data generated as part of the HapMap-ENCODE project (release 16c.1; http://www.hapmap.org/downloads/encode1.html.en) for 10 genomic regions (each spanning 500 kb) on 2p16.3 (ENr112), 2q37.1 (ENr131), 4q26 (ENr113), 7p15.2 (ENm010), 7q21.13 (ENm013), 7q31.33 (ENm014), 8q24.11 (ENr321), 9q34.11 (ENr232), 12q12 (ENr123) and 18q12.1 (ENr213), genotyped in 269 HapMap samples. These are 30 parent–offspring trios from the Yoruba people in Ibadan, Nigeria (YRI); 30 parent–offspring trios from Utah, with northern and western European ancestry (from the Centre d'Etude du Polymorphisme Humain; CEU); 45 unrelated Han Chinese from Beijing, China (CHB); and 44 unrelated Japanese from Tokyo, Japan (JPT). For the purposes of the present study, we combined data for CHB and JPT to give three analysis panels: YRI, CEU and CHB + JPT. We focused exclusively on common SNPs with minor allele frequency $\geq 5\%$. The limited ascertainment of the ENCODE data prevents an unbiased assessment of less common (or rare) variants.

Tagging strategies

Using GENCODE (http://genome.imim.es/gencode/) sequence annotations from the UCSC browser (http://genome.ucsc.edu/ENCODE/), we specified four different sets of putatively causal alleles that are to be captured with a tagging strategy. The first set includes all common SNPs that fall within a gene footprint based on the complete transcription unit of known and validated gene transcripts identified by the human and vertebrate analysis and annotation protocol (HAVANA; http://www.sanger.ac.uk/HGP/havana/). We termed this set 'transcription SNPs'. The second set includes all common SNPs that fall solely within exons of known and validated genes; we termed this set 'exon SNPs'. The third set includes all common SNPs found both in exons and in regions of strong evolutionary conservation, defined as the intersect of elements detected by three conservation algorithms (PhastCons,[6] BinCons and GERP[7]) applied to multiple sequence alignments of 23 vertebrate genomes generated by TBA[8] and by M-LAGAN.[9] (These regions correspond to the 'intersect consensus elements' from the 'ENCODE Comparative Genomics' track at the UCSC browser.) We termed this set 'excon SNPs'. These three SNP sets reflect specific gene-based tagging strategies. The 'excon' approach captures the spirit of the 'exon SNP' strategy, but makes the rather uncontroversial extension that conserved sequence points to as yet uncharacterised genes or other regions of potentially functional importance.[10,11] The fourth tagging strategy was simply to capture all observed common SNPs across all 10 ENCODE regions, regardless of gene annotation; we termed these 'anonymous SNPs'. The characteristics of these four sets of putatively causal alleles are shown in Table 1.

Tag SNP selection

For a given set of putatively causal alleles (transcription, exon, excon, anonymous SNPs), we used the program Tagger[12] (http://www.broad.mit.edu/mpg/tagger/) to derive a set of tag SNPs such that each common SNP ($\geq 5\%$) in that set was captured with $r^2 \geq 0.8$ either by a single marker[13] or by a specified haplotype.[12] This multimarker approach essentially maintains an identical set of 1 d.f.

Table 1 Characteristics of the four SNP sets as putatively causal alleles in the ENCODE data

SNP set	*% DNA sequence coverage*	*Number of common SNPs (MAF≥5%)*		
		YRI	*CEU*	*CHB+JPT*
Anonymous (all)	100.0	9043 (100%)	7627 (100%)	6711 (100%)
Transcription	34.4	2783 (31%)	2340 (31%)	2050 (31%)
Exon	3.3	199 (2%)	184 (2%)	162 (2%)
Excon	5.0	302 (3%)	270 (4%)	241 (4%)

Table 2 Performance of the selected tag SNPs (pairwise and multimarker) for the four tagging strategies

	YRI			CEU			CHB+JPT		
Tagging strategy	*Number of tags*	*% SNPs $r^2 \geq 0.8$*	*Mean maximum r^2*	*Number of tags*	*% SNPs $r^2 \geq 0.8$*	*Mean maximum r^2*	*Number of tags*	*% SNPs $r^2 \geq 0.8$*	*Mean maximum r^2*
				Pairwise tagging					
Anonymous	3140	100	0.96	1360	100	0.96	1361	100	0.96
Transcription	961	38	0.44	464	43	0.48	479	41	0.46
Exon	152	10	0.18	99	18	0.25	98	15	0.23
Excon	240	17	0.30	157	28	0.41	152	25	0.39
				Multimarker tagging					
Anonymous	1878	100	0.96	843	100	0.96	871	100	0.96
Transcription	610	39	0.45	296	45	0.49	319	42	0.47
Exon	132	12	0.20	88	19	0.27	85	17	0.25
Excon	203	18	0.32	139	31	0.44	133	28	0.42

The tags are evaluated with respect to their ability to capture all common ($\geq$5%) variation in the ENCODE data.

tests (compared with pairwise tagging) by performing an aggressive search for haplotype tests that serve as effective surrogates for single tag SNPs. This reduces the total number of tag SNPs required for genotyping.

Comparative evaluation of tagging strategies

For the four tagging strategies – transcription, exon, excon and anonymous – we evaluated the selected tags by their ability to capture the total common variation across all 10 ENCODE regions, in terms of the proportion of common SNPs captured with $r^2 \geq 0.8$, and the mean maximum r^2 with which each common SNP is captured.

We also characterised the relative cost-effectiveness of excon and anonymous tags given finite genotyping resources. First, we compared the performance of excon tags with that of the same number of randomly chosen tags. These 'random *N* tags' were a set of SNPs, equal in number to the excon tags, but selected at random from all the common SNPs in the ENCODE regions, and as such did not exploit the observed LD relationships between the SNPs. Second, we compared the excon tags with the same number (*N*) of best-performing anonymous tags (ie those selected solely on the basis of LD relationships). These 'best *N*' tags were the subset of anonymous tags with the most proxies (ie SNPs captured at $r^2 \geq 0.8$).[12] Thirdly, we evaluated the performance of these best *N* tags at capturing the common SNPs that reside in exons and conserved sequence.

It is likely, however, that the common variation in exons and conserved sequence will represent only a fraction *f* of the total putatively causal variation in the genome. The proportion of the total trait-causing variation (C_{causal}) that is captured by a set of tags can be approximated by:

$$C_{\text{causal}} \approx f \cdot C_{\text{excon}} + (1 - f) \cdot C_{\text{all}}$$

where C_{excon} is the proportion of excon variation captured, and C_{all} is the proportion of total variation captured. For each analysis panel (YRI, CEU and CHB+JPT), we estimated C_{causal} for *f* ranging from 0.05 to 1.0, comparing the excon tags to the same number of anonymous tags (based on LD alone).

Results

We have examined the performance of three gene-based tagging approaches using common SNPs (frequency $\geq$5%) from 10 ENCODE regions together with sequence annotations from the GENCODE project, and compared them with an anonymous tagging approach (in which tags are selected solely on the basis of LD structure, irrespective of gene annotation).

Using Tagger[12] we picked tags such that each SNP in a given set (listed in Table 1) is captured by a single marker with pairwise $r^2 \geq 0.8$. We found that 3140 anonymous tags were needed to capture all the common variation in the YRI samples, 1360 in CEU and 1361 in CHB+JPT, which correspond to genotype savings of three- to six-fold relative to the total number of common SNPs in the data (Table 2). As expected, these savings track with the extent of LD in the respective population samples.

When we considered the transcription tagging strategy, we found genotype savings of about three-fold compared with anonymous tagging. Between 464 (CEU) and 961 (YRI) tag SNPs captured roughly 40% of the total common variation with $r^2 \geq 0.8$ (with a mean maximum r^2 of 0.46). As gene footprints are large contiguous segments that make up 34% of this data set, this result is not surprising as tagging a subset of chromosomes would require an effort roughly proportional to the fraction of the genome those chromosomes cover.

The greatest genotyping savings can be achieved by the exon tagging strategy (Table 2). Beyond the exons themselves, however, the exon tags in general perform poorly, capturing not more than 18% of the total common variation with $r^2 \geq 0.8$ in CEU, and as little as 10% in YRI. However, the focused tagging of excon SNPs (those SNPs in exons and regions of convincing evolutionary conserva-

tion) yields genotype savings of between eight-fold (CEU and CHB + JPT) and 13-fold (YRI) compared with anonymous tagging, and provides tags that capture approximately a quarter of the total common variation. For example, 157 tags in CEU captured 28% of the total common variation and all excon SNPs at $r^2 \geq 0.8$ (with a mean maximum r^2 of 0.41) in the complete set of 7627 common SNPs (Table 2). The tagging performance in YRI was not as good: 240 tags captured only 17% of all common SNPs at $r^2 \geq 0.8$ (with a mean maximum r^2 of 0.30).

As the excon tagging strategy is commonly proposed for reasons noted earlier, and appears to offer good coverage given genotyping investment, we sought to characterise further how well these excon tags performed in terms of coverage of the total common variation. First, we examined whether the excon tags provided better, worse or equivalent coverage than a randomly selected set of markers of equivalent density. We randomly picked common SNPs as tags from the complete ENCODE data (without consideration of LD structure), equal in number (*N*) to the excon tags, to generate a set of 'random *N*' tags. Strikingly, excon tags were significantly worse than these 'random *N*' tags at capturing the total common variation. In 100 random trials, the fraction of common SNPs captured with $r^2 \geq 0.8$ was higher for the 'random *N*' tags than for the excon tags 93 times for YRI, 96 times for CEU and all 100 times for CHB + JPT samples. In terms of the mean maximum r^2, the 'random *N*' tags were consistently better than the excon tags (Table 3).

We next evaluated the best set of anonymous (LD-based) tags, again equal in number (*N*) to the excon tags. We did this by preferentially picking those SNPs as tags that have the most proxies.[12] Not surprisingly, coverage of this 'best *N*' tag set was much better with respect to the total common variation than random tags: >40% of the SNPs were captured with $r^2 \geq 0.8$ (with a mean maximum $r^2 > 0.50$) for CEU and CHB + JPT samples, and >30% for YRI samples (with a mean maximum $r^2 > 0.40$) (Table 3). (The 'best *N*' tag set also captured between 40% (YRI) and 59% (CEU) of the common SNPs (with $r^2 \geq 0.8$) found solely in exons and conserved sequence.) Therefore, focusing exclusively on excon tags does enable great efficiency, but it comes at a considerable penalty for the detection of causal variants that reside in the remaining 95% of the genome.

As multimarker tagging approaches are becoming more popular,[14,15] we decided to repeat some of these analyses using the haplotype-based approach that we described recently.[12] For all tagging strategies, the multimarker approach improved genotyping efficiency significantly, although the *relative* efficiency savings made by focused tagging of transcription SNPs were much the same as for pairwise tagging. Multimarker excon tagging – in which between 133 (CHB + JPT) and 203 (YRI) tags were needed – yields genotype savings of between six-fold (for CEU and CHB + JPT samples) and nine-fold (for YRI) (Table 2). These efficiency savings are not as impressive as the corresponding values in pairwise tagging. The reduced effectiveness of the multimarker approach is likely to be the result of the small size of the excon regions, limiting efficiency by not taking advantage of long-range LD.[16] Again, we observed that excon tags performed worse at capturing the total common variation than equivalent numbers of 'random *N*' tags and 'best *N*' tags. The 'best *N*' tags themselves, however, captured with $r^2 \geq 0.8$ between 49% (YRI) and 68% (CEU) of common variation in excon region (Table 3).

Known exons and evolutionarily conserved regions are likely to contain only a fraction of the total putatively trait-causing variation in the genome. We have estimated the impact of the relative distribution of putatively causal variation between excons and the rest of the genome on the performance of the excon and 'best *N*' anonymous tagging strategies (Figure 1). Both excon and anonymous

Table 3 Comparative evaluation of excon, random and LD-based tags

		YRI		*CEU*		*CHB+JPT*	
Tag SNPs	*Evaluated SNP set*	*% SNPs $r^2 \geq 0.8$*	*Mean maximum r^2*	*% SNPs $r^2 \geq 0.8$*	*Mean maximum r^2*	*% SNPs $r^2 \geq 0.8$*	*Mean maximum r^2*
			Pairwise tagging				
Excon	Anonymous	17	0.30	28	0.41	25	0.39
Random *N*[a]	Anonymous	18	0.36	32	0.50	30	0.48
Best *N*[a]	Anonymous	32	0.44	46	0.56	44	0.54
Best *N*[a]	Excon	40	0.54	59	0.68	52	0.66
			Multimarker tagging				
Excon	Anonymous	18	0.32	31	0.44	28	0.42
Random *N*[a]	Anonymous	18	0.36	34	0.51	33	0.50
Best *N*[a]	Anonymous	36	0.45	53	0.58	50	0.55
Best *N*[a]	Excon	49	0.60	68	0.72	64	0.70

[a]'Random *N*' and 'best *N*' refer to tags picked at random and based on LD, respectively, equal in number to the set of excon tags; 'Anonymous' refers to all common (≥5%) SNPs in the ENCODE data.

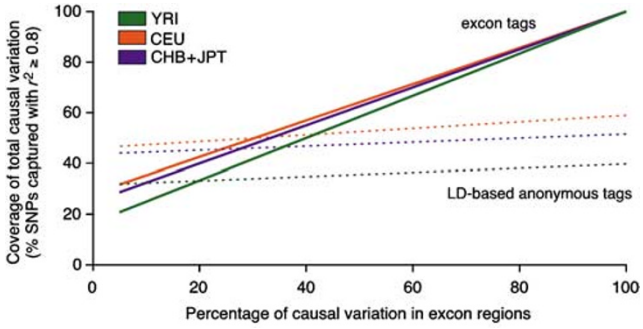

Figure 1 Impact of the relative distribution of causal variation on the performance of excon and anonymous tagging strategies. The coverage of the total causal variation captured by excon tags (solid lines) and the same number (best *N*) of LD-based anonymous tags (broken lines) is plotted as a function of the proportion of the causal variation residing in excon regions. The coverage (*y*-axis) is given in terms of % common SNPs captured with pairwise $r^2 \geq 0.8$ for the YRI (green), CEU (orange) and CHB + JPT (magenta) analysis panels.

tagging demonstrate equal coverage for equal genotyping investment when a minority of the putatively causal variation – approximately 20–26% (YRI), 30–41% (CEU), 28–37% (CHB + JPT) – lies within recognisable excon regions (Figure 1). At these proportions where the cost-effectiveness is equal for excon and anonymous tagging strategies, we observe that 33% of all causal variation is captured with pairwise $r^2 \geq 0.8$ in YRI, 50% in CEU and 46% in CHB + JPT. As the distribution of causal variation shifts more towards excon regions, the excon tagging strategy captures more of the total functional variation and consequently will become significantly more cost-effective. Clearly, the converse is true when causal variation is found overwhelmingly in regions of the genome other than exons and conserved sequence. The optimal tagging approach will therefore depend crucially on the assumed genetic architecture of the trait under investigation.

Discussion

Using the most extensive resequencing and annotation data sets available at present, we have examined a number of seemingly distinct tagging strategies to capture common genetic variation. The respective performance of gene-based and anonymous tagging approaches to capture putatively causal variation in a genome-wide context will obviously depend on the relative distribution of this variation between exons and conserved sequence, and the rest of the genome. If *all* trait-causing variation were to lie in the 5% of DNA found in excons, then there are substantial gains (eight- to 13-fold, in our study) in terms of genotyping effort to be made by adopting the excon-tagging approach. If, at the opposite end of the genetic spectrum, trait-causing variation is uniformly distributed throughout the genome, in such a way as not to be over-represented in exonic or conserved regions, then an excon tagging approach comes at a cost of missing more than half of the total trait-causing variation. Genotyping the same number of anonymous tags based on LD provides significantly better genome-wide coverage despite the risk of missing functionally important variants in regions of low LD.[17] The true state of nature, of course, lies between these two extremes.

Our study suggests that for a plausible scenario of an equal distribution of causal variation between excon regions and the rest of the genome, genotyping excon tags provides somewhat better coverage than with genotyping the same number of anonymous tags. This improvement is quite small (~7%) in the case of multimarker tagging of CEU samples, but more sizeable (~16%) for YRI samples (Figure 1). However, we note that these estimates may be biased given that the ENCODE data set covers only a tiny fraction of the genome. We conclude that these apparent differences may amount to little practical significance, and that we see, perhaps surprisingly, roughly equal coverage for equal investment.

As genome-wide genotyping products are becoming available, each investigator will have to decide how causal variation is likely to be distributed across the genome for the phenotype of interest, and how well such products capture these putatively causal variants. This is relevant because, in reality, investigators are not likely to have the resources to customise an array with SNPs of their choice (that optimally capture the presumed set of putatively causal alleles). A recent analysis demonstrates that the Affymetrix GeneChip Mapping 500K and Illumina Sentrix HumanHap300 BeadChip arrays achieve comparable coverage of common variation across the genome despite differences in the design of these products.[18]

Lastly, we note that our analysis did not include less common SNPs or rare sequence variants. It is clear, however, that rare variation can be an important component of the genetic architecture of complex diseases. Although indirect haplotype-based methods have been proposed for testing such variants, complete ascertainment by resequencing will be the only comprehensive approach to expose the full spectrum of causal variants that contribute to trait heritability. This is currently feasible for selected genomic regions (eg, for follow-up of initial findings). It is also possible to design genome-wide panels supplemented with less common (rare) variants of biological importance (for instance, coding variants and splice site mutations) and therefore likely to have a much higher prior probability of playing a role in disease.

Acknowledgements

PIWDB thanks Daryl Thomas from the UCSC team for help with the ENCODE annotation tracks. SW is supported by a Career Development Fellowship from the Wellcome Trust.

References

1 The International HapMap Consortium: The International HapMap Project. *Nature* 2003; **426**: 789–796.
2 The International HapMap Consortium: A haplotype map of the human genome. *Nature* 2005; **437**: 1299–1320.
3 The ENCODE Project Consortium: The ENCODE (ENCyclopedia Of DNA elements) project. *Science* 2005; **306**: 636–640.
4 Hirschhorn JN, Daly MJ: Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 2005; **6**: 95–108.
5 Botstein D, Risch N: Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* 2003; **33** (Suppl): 228–237.
6 Siepel A, Bejerano G, Pedersen JS *et al*: Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 2003; **15**: 1034–1050.
7 Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A: Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* 2005; **15**: 901–913.
8 Blanchette M, Kent WJ, Riemer C *et al*: Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* 2004; **14**: 708–715.
9 Brudno M, Do CB, Cooper GM *et al*: LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res* 2003; **13**: 721–731.
10 Dermitzakis ET, Reymond A, Antonarakis SE: Conserved nongenic sequences – an unexpected feature of mammalian genomes. *Nat Rev Genet* 2005; **6**: 151–157.
11 Drake JA, Bird C, Nemesh J *et al*: Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nat Genet* 2006; **38**: 223–227.
12 de Bakker PIW, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D: Efficiency and power in genetic association studies. *Nat Genet* 2005; **37**: 1217–1223.
13 Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA: Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* 2004; **74**: 106–120.
14 Stram DO, Haiman CA, Hirschhorn JN *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**: 27–36.
15 Weale ME, Depondt C, Macdonald SJ *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**: 551–565.
16 Pe'er I, Chretien YR, de Bakker PIW *et al*: Biases and reconciliation in estimates of linkage disequilibrium in the human genome. *Am J Hum Genet* 2006; **78**: 588–603.
17 Smith AV, Thomas DJ, Munro HM, Abecasis GA: Sequence features in regions of weak and strong linkage disequilibrium. *Genome Res* 2003; **15**: 1519–1534.
18 Pe'er I, de Bakker PIW, Maller J *et al*: Evaluating and improving power in whole genome association studies using fixed marker sets. *Nat Genet* 2006; **38**: 663–667.