European Journal of Human Genetics (2008) 16, 718–723

www.nature.com/ejhg

ARTICLE

To what extent do scans of non-synonymous SNPs complement denser genome-wide association studies?

David M Evans*[,1,2], Jeffrey C Barrett[1] and Lon R Cardon[1]

[1]*Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK;* [2]*MRC Centre for Causal Analyses in Translational Epidemiology, Department of Social Medicine, University of Bristol, Bristol, UK*

Several studies involving genome-wide scans of non-synonymous SNPs (nsSNPs) have successfully identified loci contributing to common complex diseases. We were interested in the extent to which these small scans involving a few thousand non-synonymous markers might complement the results from denser genome-wide association studies. We assessed the degree to which three commercially available genome-wide marker panels tagged nsSNPs on the Illumina HumanNS-12 BeadChip, a product specifically designed to capture non-synonymous variation. We demonstrate that commercially available genome-wide panels already tag the majority of common non-synonymous variants on the NS-12 BeadChip, indicating that with respect to capturing common non-synonymous variation, information from the NS-12 BeadChip is largely redundant. In contrast, genome-wide panels fail to capture most of the rare SNPs present on the NS-12 BeadChip. Power calculations reveal that non-synonymous scans involving sample sizes typical of the current wave of genome-wide association studies are unlikely to identify rare variants of small effect, but could conceivably identify rare variants of intermediate penetrance. We conclude that non-synonymous scans may facilitate the identification of rare variants of intermediate penetrance that would not otherwise be detectable using dense genome-wide panels, but are unlikely to uniquely identify common variants contributing to complex disease variation.
European Journal of Human Genetics (2008) **16,** 718–723; doi:10.1038/sj.ejhg.5202011; published online 16 January 2008

Keywords: non-synonymous SNPs; genome-wide association; coverage

Introduction

With the advent of genome-wide association (GWA) analysis, genetic mapping has now entered an exciting new phase, where for the first time it has become possible to robustly identify many of the genetic variants underlying complex traits and diseases. Three recent developments have made this a possibility. First, the availability of genotyping chips containing hundreds of thousands of markers, which provide good coverage of much of the common genetic variation within the genome, has meant that GWA studies are now financially and technically feasible.[1] Second, the publication of the International Haplotype Map, which documents the pattern of linkage disequilibrium (LD) across the genome, has facilitated the design and analysis of GWA studies.[2] Finally, the existence of large patient cohorts has been a necessary prerequisite to obtain the power necessary to detect common loci of small-to-moderate effect.[1] These developments have led to a flood of GWA studies that have successfully identified genes conferring risk to a variety of diseases including (but not limited to) breast cancer,[3] coronary heart disease,[4] inflammatory bowel disease,[5] and types I and II diabetes.[6,7]

The GWA approach implicitly assumes that common genetic variants contribute to complex disease aetiology and that it is possible to identify these polymorphisms by indirectly tagging common variation at hundreds of

*Correspondence: Dr DM Evans, Department of Social Medicine, University of Bristol, Canynge Hall, Whiteladies Road, Bristol BS8 2PR, UK. Tel: +44 (0)117 9287200; Fax: +44 (0)117 9287292;
E-mail: dave.evans@bristol.ac.uk
Received 13 September 2007; revised 10 December 2007; accepted 20 December 2007; published online 16 January 2008

thousands of sites across the genome.[2] The fact that this strategy has worked so well across such a wide spectrum of diseases attests to the notion that common genetic variants contribute to the risk of many complex diseases. In addition to these genome-wide studies, there have been a smaller number of less costly scans involving a few thousand non-synonymous SNP (nsSNP) markers scattered across the genome.[8–10] The rationale underpinning this approach is that non-synonymous polymorphisms, which produce amino-acid substitutions and often adverse conformational changes in protein structure, are likely to be overrepresented in disease aetiology. This has certainly been true for Mendelian disorders,[11] and recent evidence demonstrates that non-synonymous variation is also an important contributor to complex disease risk as well.[8–10,12–15] Importantly, many nsSNPs are rare, and consequently may not be tagged well by dense genome-wide panels, which have primarily been designed to indirectly capture common genetic variation.[2] In the situation where a rare nsSNP confers disease risk, it may be necessary to genotype the variant directly to detect disease association.

To date there have been three genome-wide scans that have specifically examined the association between non-synonymous variation and disease risk.[8–10] All three of these studies have yielded successes, resulting in the identification of novel loci underlying Crohn's disease (rs2241880, minor allele frequency (MAF) = 0.47),[8] type I diabetes (rsrs1990760, MAF = 0.40),[9] and ankylosing spondylitis (rs27044, MAF = 0.34; rs11209026, MAF = 0.04).[10] These studies validate the non-synonymous approach and conclusively demonstrate that such scans can identify loci involved in complex disease aetiology. Following on from these successes, the Illumina biotechnology company has recently released a product, the HumanNS-12 SNP chip, which specifically focuses on non-synonymous variation throughout the genome. We were particularly interested in the extent to which these panels might complement the results from GWA studies given that (a) existing dense genome-wide panels tag a large fraction of common variation within the genome (and hence presumably common nsSNPs as well) and (b) there may not be much to be gained by directly genotyping the remaining uncommon nsSNPs, which are not tagged by dense genome-wide panels, as there is low power to detect rare variants by genetic association anyway.[16] The aim of our study therefore was to assess the amount of information a genome-wide scan of non-synonymous polymorphisms might provide over and above that afforded by a denser GWA study.

Materials and methods

The content of the Illumina HumanNS-12 Genotyping BeadChip is largely based upon a collection of SNPs selected by the Wellcome Trust Case Control Consortium that included all known nsSNPs with >1% MAF in European populations at the time of study design.[10] The complete list of 13 917 SNPs present on the HumanNS-12 chip is displayed in Supplementary Table 1. Of these, the present analyses concern only the 11 649 SNPs on the panel that were non-synonymous (ie, a dense set of MHC tagging SNPs across the MHC region was excluded as were several other SNPs). In addition, a further 15 nsSNPs whose existence or genomic position was uncertain, as well as 174 nsSNPs on the X chromosome were excluded from analyses, leaving a total of 11 460 nsSNPs.

We were interested in the extent to which SNPs on commercially available, dense GWA panels tagged nsSNPs on the Illumina HumanNS-12 Genotyping BeadChip. Therefore, for each nsSNP on the NS-12 panel, the highest r^2 between it and a marker on each of the three dense genome-wide panels was recorded. The dense genome-wide panels were the Affymetrix 500K SNP chip (500 568 SNPs), the Illumina HumanHap300 Genotyping BeadChip (317 503 SNPs), and the Illumina 550K BeadChip (555 352 SNPs). MAF and LD calculations were based upon cleaned genotype data uploaded onto the International HapMap website on 2 March 2007, which represented the most recent data collected at the time of manuscript submission. Calculations were performed using all four HapMap population samples, specifically the 60 founder individuals from the Centre d'Etude Polymorphism Humain collection (CEU), 60 founder individuals from the Yoruba in Ibadan, Nigeria (YRI), 45 Han Chinese in Beijing, China (CHB), and 44 Japanese individuals from Tokyo (JPT). MAF and LD estimates were so similar in these last two population samples, that the groups were combined. SNPs on the Illumina NS-12 SNP chip that were not present in the latest cleaned release of the HapMap data were excluded from LD and MAF calculations (ie, even if the nsSNPs were present on one or more of the GWA panels). Thus, a further 788, 775, and 743 SNPs in the CEU, YRI, and combined CHB and JPT samples, respectively, were excluded from analyses. SNPs were included in MAF and LD calculations even if they were monomorphic within one or more HapMap samples. Calculation of r^2 was by a custom written programme and only involved SNPs 500 kB either side of the SNP of interest on the NS-12 BeadChip.

Finally, we examined the power to detect association in a genome-wide scan of non-synonymous variants using a web-based Genetic Power Calculator tool.[17] In all cases, we assumed an underlying multiplicative disease model.

Results

Figure 1 displays the distribution of MAFs for nsSNPs on the Illumina HumanNS-12 Genotyping BeadChip for each of the four population groups in the HapMap. In sharp contrast to the roughly uniform distribution of MAFs, which is often observed using dense genome-wide panels

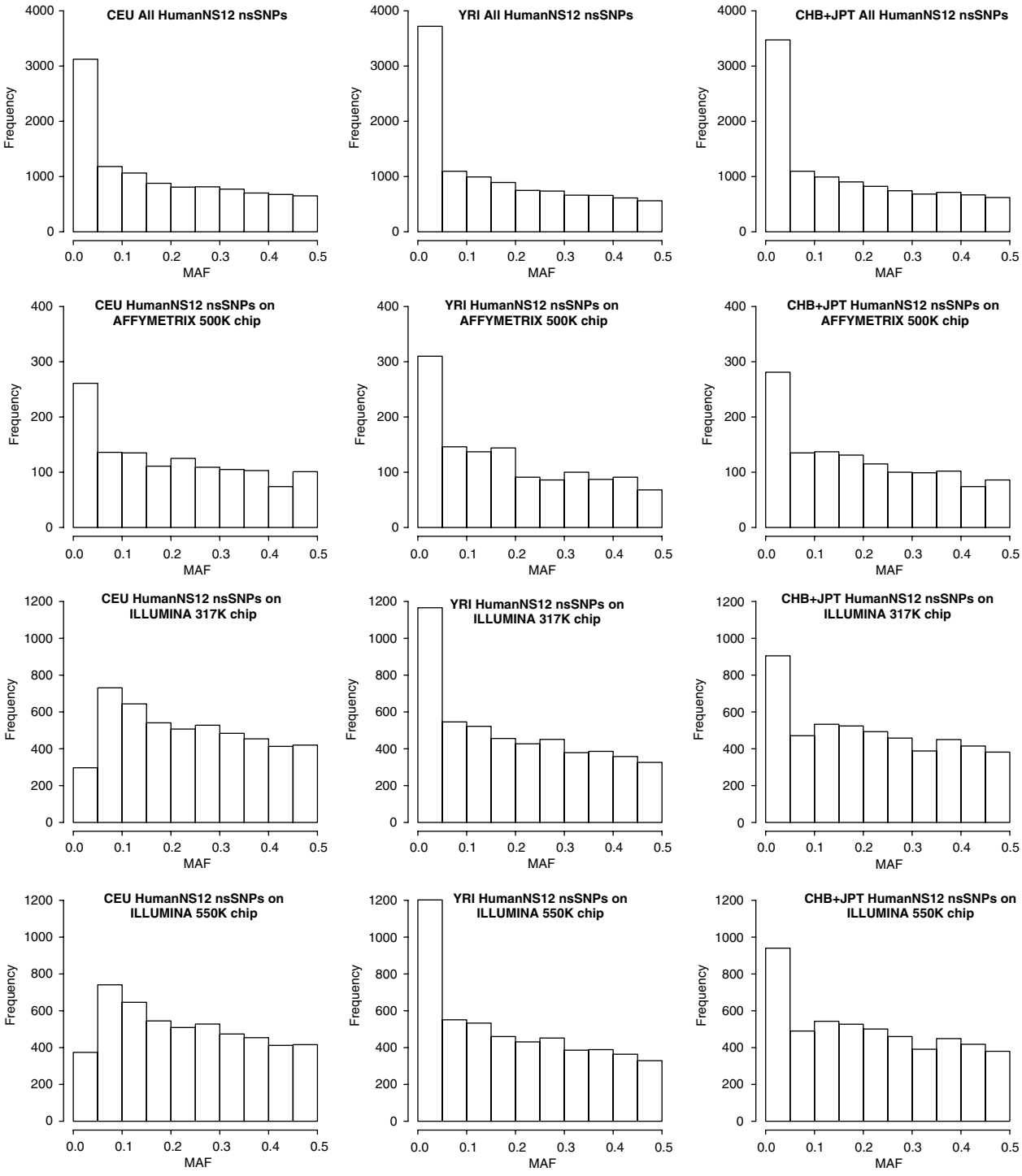

Figure 1 Histograms displaying MAFs of nsSNPs on the Illumina HumanNS-12 SNP chip. The first row displays MAFs for all nsSNPs on the Illumina HumanNS-12 SNP chip. The lower three rows display MAFs of nsSNPs on the Illumina HumanNS-12 SNP chip that are also present on the Affymetrix 500K (second row), Illumina 317K (third row), or Illumina 550K (fourth row) chips, respectively. Each column refers to a different HapMap population sample. In sharp contrast to the roughly uniform distribution of MAFs, characteristic of dense genome-wide panels, nsSNPs on the HumanNS-12 SNP chip exhibit a clear skew towards the rare end of the frequency spectrum. This figure suggests that rare SNPs in the CEU population sample were deliberately not included on the Illumina 317 and 550K chips.

(a consequence of SNP discovery and ascertainment), the non-synonymous panel exhibits a clear skew towards the rare end of the frequency spectrum. Since many nsSNPs on both the Illumina 317 and 550K SNP chips were deliberately included as part of the product design, the MAF calculations were repeated using only nsSNPs that were physically present on the Affymetrix 500K (1260 nsSNPs), Illumina 317K (5030 nsSNPs), or Illumina 550K (5110 nsSNPs) panels. Figure 1 suggests that SNPs that are rare in the CEU population sample were deliberately not included on the Illumina 317 and 550K chips.

Table 1 displays the proportion of nsSNPs on the Illumina HumanNS-12 BeadChip that are tagged by markers on the Affymetrix 500K, Illumina 317 and 550K SNP chips. The columns are divided according to whether the SNP on the NS-12 panel is common (ie, MAF$\geq$5%, corresponding to 7732, 7113, and 7259 SNPs in the CEU, YRI, and JPT+CHB samples, respectively) or rare (ie, MAF$<$5%, corresponding to 2940, 3572, and 3457 SNPs in the CEU, YRI, and JPT+CHB samples, respectively) in the relevant population sample. The Affymetrix 500K chip tags 65 and 67% of common nsSNPs ($\geq$5% MAF) on the NS-12 BeadChip at an $r^2\geq0.8$ in the CEU, and combined CHB and JPT samples, respectively. These figures are in good agreement with previous estimates of the coverage among all common SNPs in the genome.[16] If the r^2 threshold is lowered to $\geq$0.5, coverage increases to $>$80% in all of these populations. This percentage roughly translates to about 1540 common nsSNPs that are not tagged well by the Affymetrix 500K chip (in the CEU sample), indicating that the vast majority of common variants on the NS-12 panel are tagged adequately. While coverage of common non-synonymous variation is similar in these populations, it is substantially lower in the YRI sample consistent with the observation that LD tends to be lower in African populations on average. Finally, we note that coverage of low-frequency nsSNPs is much poorer in all samples ($r^2\leq0.30$), indicating that the Affymetrix 500K chip does not capture rare non-synonymous variation well.

Both the Illumina 317K and Illumina 550K chips tag over 90% of common nsSNPs in the CEU sample at a threshold of $r^2\geq0.8$. However, this high level of coverage is partly a consequence of overfitting, since SNPs on both panels were specifically selected for their ability to tag common variation in the CEU HapMap data set. Therefore, the ability of these chips to capture non-synonymous variation is more accurately estimated by the CHB, JPT, and YRI population groups, which do not suffer from the same overfitting problem. LD estimates from the combined CHB and JPT samples indicate that coverage is still very high (ie, $>$80% at $r^2\geq0.8$). Coverage in the YRI group is significantly lower, but still impressive at the same threshold (69%). Table 1 also indicates that the Illumina panels do not tag rare nsSNPs well, presumably due to their focus on common genetic variation.

Another reason why the Illumina 317 and 550K chips provide such high coverage of non-synonymous variation is that many nsSNPs (in particular, common nsSNPs; see Figure 1) were deliberately selected as part of the product design. If the same LD analyses are performed on NS-12 SNPs, which are not physically present on the dense genome-wide chips (Table 2), it is possible to estimate the

Table 1 Percentage of nsSNPs on the Illumina HumanNS-12 BeadChip that are tagged by markers on the Affymetrix 500K, and Illumina 317 and 500K SNP chips

	Affymetrix 500K			*Illumina 317K*			*Illumina 550K*		
	All	*<5%*	*≥5%*	*All*	*<5%*	*≥5%*	*All*	*<5%*	*≥5%*
CEU									
$r^2\geq0.5$	65	27	80	72	10	95	77	25	97
$r^2\geq0.8$	53	24	65	67	8	90	73	21	92
$r^2=1.0$	39	23	45	59	7	78	64	20	81
YRI									
$r^2\geq0.5$	50	21	66	64	37	78	71	39	86
$r^2\geq0.8$	37	18	47	58	35	69	64	37	77
$r^2=1.0$	27	17	33	54	35	63	57	37	67
JPT+CHB									
$r^2\geq0.5$	64	27	81	72	35	90	77	39	95
$r^2\geq0.8$	53	25	67	66	34	82	73	37	89
$r^2=1.0$	37	23	44	58	33	70	63	36	76

CEU, Centre d'Etude Polymorphism Humain collection; CHB, Han Chinese in Beijing, China; JPT, Japanese individuals from Tokyo; MAF, minor allele frequency; nsSNP, non-synonymous SNP; YRI, Yoruba in Ibadan, Nigeria.
Calculations are presented for all nsSNPs, rare nsSNPs (<5% MAF), and common nsSNPs (≥5% MAF).

Table 2 Proportion of nsSNPs on the Illumina HumanNS-12 Beadchip that are tagged by markers on the Affymetrix 500K, Illumina 317K and 500K SNP chips (Only markers not physically present on the dense genome-wide chips are included in the analyses)

	Affymetrix 500K			*Illumina 317K*			*Illumina 550K*		
	All	*<5%*	*≥5%*	*All*	*<5%*	*≥5%*	*All*	*<5%*	*≥5%*
CEU									
$r^2\geq0.5$	61	21	77	47	5	87	56	18	91
$r^2\geq0.8$	47	17	59	38	2	72	47	14	79
$r^2=1.0$	31	17	37	22	1	42	32	13	49
YRI									
$r^2\geq0.5$	44	14	60	33	8	51	44	11	69
$r^2\geq0.8$	29	10	39	20	6	31	30	8	47
$r^2=1.0$	18	10	22	12	6	17	19	8	27
JPT+CHB									
$r^2\geq0.5$	59	20	78	48	13	76	56	16	88
$r^2\geq0.8$	47	18	61	36	10	58	48	14	75
$r^2=1.0$	29	17	35	21	10	31	29	13	43

CEU, Centre d'Etude Polymorphism Humain collection; CHB, Han Chinese in Beijing, China; JPT, Japanese individuals from Tokyo; MAF, minor allele frequency; nsSNP, non-synonymous SNP; YRI, Yoruba in Ibadan, Nigeria.

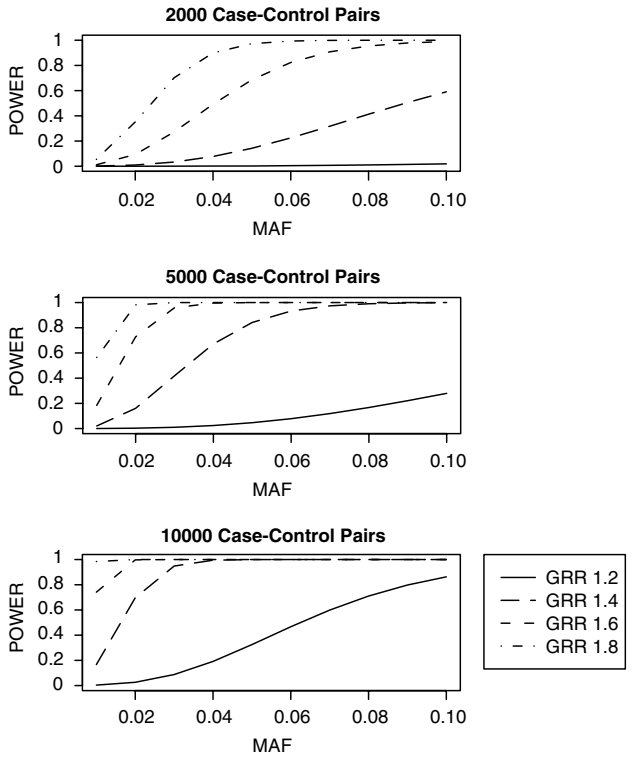

Figure 2 Relationship between MAF, heterozygote GRR, and power to detect association assuming a multiplicative disease model. Results are shown for 2000, 5000, and 10 000 case–control pairs assuming a disease prevalence of 1% and a type I error rate of $\alpha = 3.6 \times 10^{-6}$. The figure illustrates that it is possible to detect rare variants of intermediate penetrance using current sample sizes of 2000 case–control pairs. To detect rare alleles of smaller effect, far larger sample sizes will need to be employed.

ability of the Illumina 317 and 550K chips to indirectly tag non-synonymous variation. These analyses demonstrate that indirect coverage of common variation is still quite high in the CEU, and combined CHB and JPT population groups, but that rare variants are covered extremely poorly.

It appears then that existing dense genome-wide products capture common non-synonymous variation well, particularly in non-African populations. However, none of the genome-wide panels adequately capture rare non-synonymous variation. What are the consequences of this for genetic association studies? Figure 2 displays the power to detect a rare nsSNP contributing to disease risk using 2000, 5000, and 10 000 case–control pairs, across a variety of different MAFs and genetic effect sizes. Two thousand case–control pairs reflect the size of GWA studies, which are currently appearing in the literature, whereas 5000 and 10 000 pairs represent what could be achieved by a realistic doubling/quintupling of these sample sizes.

Figure 2 confirms that sample sizes typical of the current wave of non-synonymous scans are unlikely to identify rare variants (<5%) of small effect (ie, genotypic relative risk (GRR) < 1.4). However, a study of 2000 case–control pairs is large enough to detect rare alleles of 'intermediate' penetrance (ie, GRR > 1.6–1.8). This is important because as Tables 1 and 2 demonstrate, dense genome-wide products do not tag the majority of rare non-synonymous variation in the genome. Increasing the sample size further to 5000 case–control pairs is adequate to identify some low-risk rare alleles (MAF > 2%), but is still not large enough to detect rarer variants. In fact, even with 10 000 case–control pairs, there is only low-to-moderate power to detect a variant of 1% MAF responsible for an allelic risk ratio of 1.4.

Discussion

We examined the degree to which a genome-wide scan of nsSNPs complemented the more costly and comprehensive whole-genome approaches involving 300 000–500 000 markers. We found that whole-genome panels tagged the majority of common nsSNPs on the Illumina HumanNS-12 BeadChip, but only captured a small proportion of the rare non-synonymous variants. Our power calculations revealed that current sample sizes of 1000–2000 case–control pairs are insufficient to identify rare variants of small effect and that vastly larger studies in the order of 10 000+ case–control pairs will be required to detect low-frequency, low-risk variants. However, one to two thousand pairs are sufficient to detect rare variants of intermediate penetrance that may not be identified via dense whole-genome approaches.

A major conclusion from our study has been that existing genome-wide panels capture the majority of common non-synonymous variation in the genome. The corollary is that there is substantial redundancy on the HumanNS-12 BeadChip when compared with these larger marker panels. This is particularly true of the Illumina 317K and Illumina 550K SNP chips, where many common nsSNPs were deliberately included as part of the product design. A particularly vivid illustration of this redundancy is provided by the three successful non-synonymous scans to date.[8–10] In each case, the non-synonymous variant associated with the disease would also have presumably been flagged had subjects been typed using any of the dense genome-wide panels, since either the SNP itself or a marker in high LD with it (ie, $r^2 > 0.7$) is present on the genome-wide chips. Given that genome-wide panels have the added advantage of tagging other genomic regions besides nsSNPs, from the perspective of identifying common disease-associated variants, apart from cost, there appears little to recommend on the use of non-synonymous scans. A critical exception is in the case of African populations, where the level of LD is lower and even common non-synonymous variants are not tagged well by existing whole-genome panels.

For current sample sizes, the degree to which scans of nsSNPs will complement the results from GWA studies will largely depend upon the extent to which rare alleles of

intermediate penetrance contribute to common disease pathogenesis. Although several studies have implicated low-frequency alleles of moderate effect in the aetiology of complex disease,[18] it is too early to say how 'common' these instances will be, particularly in the case of non-synonymous variation. In addition, even if one is willing to accept the premise that rare non-synonymous variants are important in complex disease aetiology, the HumanNS-12 BeadChip directly types only a very small fraction of the estimated ~60 000 nsSNPs listed in the dbSNP database. Since SNP discovery is biased towards common variants by definition, we might expect the total number of rare nsSNPs in existence to be far greater than this figure. In other words, the NS-12 BeadChip does a mediocre job at capturing the total amount of rare non-synonymous variation in the genome.

Finally, we note that our study quantifies the power to detect a single disease locus as a function of effect size and MAF. In reality, the risk of developing complex disease will almost certainly be a result of many disease-predisposing variants, several of which may be non-synonymous mutations. Assuming that disease loci are independent, the expected number of true variants identified using a non-synonymous scan will equal the sum of the power to detect each individual locus. In other words, the number of nsSNPs correctly identified using such a scan will depend upon (a) the number of nsSNPs that contribute to disease risk, (b) whether these nsSNPs are present on the chip of interest, and (c) the MAF and relative risk conferred by each variant.

In conclusion, genome-wide scans of non-synonymous markers represent a cheap method of screening for disease-associated variants, which have yielded a number of recent successes. However, upon close scrutiny of the Illumina NS-12 Human BeadChip, we have found that the majority of common nsSNPs on the panel are already tagged by existing whole-genome products, indicating that non-synonymous scans are unlikely to add significantly to the results from GWA studies in terms of identifying common variants. In contrast, non-synonymous scans may enable the identification of rare non-synonymous variants of intermediate penetrance that would not otherwise be detectable via current whole-genome marker panels. Thus, the degree to which non-synonymous scans might complement the results from dense whole-genome scans will depend on (a) the degree to which rare variants of intermediate penetrance contribute to common complex disease and (b) the degree to which the scan captures the total amount of rare non-synonymous variation in the genome.

Acknowledgements

This work was supported by the Wellcome Trust and the National Institutes of Health (EY-126562 to LRC).

Web resources

The URL for data presented herein is as follows: the International HapMap website (http://www.hapmap.org/genotypes/latest/fwd_strand/non-redundant/).

References

1 The Wellcome Trust Case Control Consortium: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007; **447**: 661–678.
2 The International HapMap Consortium: A haplotype map of the human genome. *Nature* 2005; **437**: 1299–1320.
3 Easton DF, Pooley KA, Dunning AM *et al*: Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* 2007; **447**: 1087–1093.
4 Samani NJ, Erdmann J, Hall AS *et al*: Genomewide association analysis of coronary artery disease. *N Engl J Med* 2007; **357**: 443–453.
5 Parkes M, Barrett JC, Prescott NJ *et al*: Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nat Genet* 2007; **39**: 830–832.
6 Todd JA, Walker NM, Cooper JD *et al*: Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat Genet* 2007; **39**: 857–864.
7 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.
8 Hampe J, Franke A, Rosenstiel P *et al*: A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. *Nat Genet* 2007; **39**: 207–211.
9 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.
10 The Wellcome Trust Case Control Consortium: Association scan of 14,500 nonsynonymous SNPs in four disease identifies autoimmunity variants. *Nat Genet* 2007; **39**: 1329–1337.
11 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.
12 Altshuler D, Daly M, Kruglyak L: The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat Genet* 2000; **26**: 76–80.
13 Bertina RM, Koeleman BP, Koster T *et al*: Mutation in blood coagulation factor V associated with resistance to activated protein C. *Nature* 1994; **369**: 64–67.
14 Cox A, Dunning AM, Garcia-Closas M *et al*: A common coding variant in CASP8 is associated with breast cancer risk. *Nat Genet* 2007; **39**: 352–358.
15 Duerr RH, Taylor KD, Brant SR *et al*: A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* 2006; **314**: 1461–1463.
16 Barrett JC, Cardon LR: Evaluating coverage of genome-wide association studies. *Nat Genet* 2006; **38**: 659–662.
17 Purcell S, Cherny SS, Sham PC: Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* 2003; **19**: 149–150.
18 Rahman N, Seal S, Thompson D *et al*: PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat Genet* 2007; **39**: 165–167.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)