OPINION

A gene-centric approach to genome-wide association studies

Eric Jorgenson and John S. Witte

Abstract | Genic variants are more likely to alter gene function and affect disease risk than those that occur outside genes. Variants in genes, however, might not be sufficiently covered by the existing approaches to genome-wide association studies. Our analysis of the HapMap ENCODE data indicates that this concern is valid, and that an alternative approach that focuses on genic variants provides a more complete coverage of functionally important regions and a greater genotyping efficiency. We therefore argue that resources should be developed to make gene-centric genome-wide association studies feasible.

Interest in genome-wide association (GWA) studies was instigated in 1996, when Risch and Merikangas noted that association studies have considerably greater power than linkage analysis to detect genetic variants with small or moderate phenotypic effects, even when testing large numbers of variants across the genome[1]. Their power estimates related to a 'direct' or 'sequence-based' study design, using variants that are located within genes. In this design, 500,000 variants (on average, 5 variants for each gene in the then-anticipated 100,000 human genes) would be genotyped and tested for association with the phenotype of interest. This approach has the advantage of a candidate-gene study — in which variants with known biological functions can be examined — but avoids the disadvantage of only testing a limited number of genes.

Risch and Merikangas also noted that the number of genotyped variants could be reduced using an approach that takes advantage of linkage disequilibrium (LD) between variants. This was later termed the 'indirect' or 'map-based' approach to GWA. Because variants in strong LD are likely to be inherited together, one can use a subset of 'tagging' markers as proxies for the entire set. There has been much interest in such an approach, as genotyping an exhaustive set of markers would be prohibitively expensive. The indirect approach has rapidly expanded to include variation in the whole genome rather than only in genes[2,3]. The rationale for this expansion was that potentially important functional polymorphisms also exist outside genes, particularly in *cis*-regulatory regions, which can be located tens of thousands of base pairs away from the genes that they regulate[4].

At present, there are two general strategies for indirect GWAs (FIG. 1). The first uses quasi-random or anonymous SNPs that are spread across the genome, such as the Affymetrix 500K array set. The second uses sets of LD-based tag SNPs that are specifically chosen to saturate the genome, effectively capturing most of the other unmeasured common SNPs at a pre-specified LD threshold. To make the second approach feasible, the International HapMap Project was established, with the initial goal of creating a set of 600,000 LD tagging SNPs[5]. The second phase of this project was recently completed, resulting in a publicly available catalogue of more than 3.9 million validated SNPs, as well as information about the LD between them, from 269 individuals from multiple populations[6]. Along with these SNPs, ten HapMap ENCODE regions across the human genome have been resequenced in 16 subjects from each of the three HapMap populations (Caucasian, African and Asian). The additional SNPs that were detected in these regions have been genotyped in all the HapMap subjects, providing a more complete set of SNPs that can be used to evaluate the performance of SNP genotyping sets.

In addition to these steps towards facilitating GWAs in general, various gene-based SNP discovery projects (for example, the SeattleSNPs Program for Genomic Applications) have identified SNPs in genes that can be used for gene-centric approaches to GWA studies. Such studies can use an indirect approach that focuses on markers that capture variants in genes, or can study putative causal variants directly.

Based on the results of the first few GWA studies[7–9], it is clear that both indirect map-based and direct gene-centric approaches can be successful. It is not clear, however, whether either approach currently provides a comprehensive survey of the genome. For example, although a study of age-related macular degeneration succeeded in identifying a polymorphism in the complement factor *H* gene, the study missed an additional major locus that has since been identified through candidate-gene studies[10,11]. This raises the crucial issue of coverage in GWA studies — the portion of all genetic variants for which information can be captured with a given SNP set. Coverage, in turn, affects the overall power of a study to detect causal variants. A causal variant that is in high LD with SNPs in the set that are genotyped can be detected with a minimal loss of power compared with testing the variant directly, under a number of assumptions[12]. By contrast, causal variants that are located in regions of the genome that are poorly covered by the SNP set can only be detected with an enormous sample size. Therefore, a SNP set with comprehensive coverage can help to both limit false-negative results and reduce the number of subjects needed — and in turn, the overall genotyping burden — important aspects of any GWA in light of their considerable expense.

Here we argue for a gene-centric approach to GWA studies that focuses on variation in genes for two reasons. First, variants in genes have a high probability of being functionally important, so comprehensive

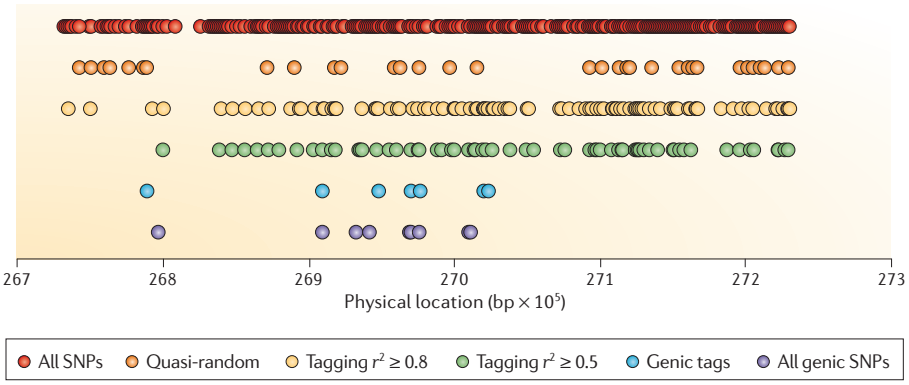


Figure 1 | **Approaches to constructing SNP sets for genome-wide association studies.** All common SNPs (those with minor allele frequencies of ≥5%) are shown for one of the HapMap ENCODE regions (ENm010 on chromosome 7) for the Caucasian population. All genic SNPs from this set are also shown. Four approaches to capturing information about the SNPs in this region in genome-wide association studies are illustrated, each using a different type of genotyping SNP set. Quasi-random SNPs from the Affymetrix 500K array set are shown, as are two sets of tagging SNPs, which are chosen on the basis that they are in linkage disequilibrium with other SNPs that are not genotyped. These tagging SNPs are selected to capture all SNPs at or above pre-specified thresholds of r^2, a measure of linkage disequilibrium; see BOX 1. Finally, the genic tagging set consists of the SNPs that are required to cover SNPs that lie within genes at or above an r^2 threshold of 0.8.

coverage of these variants is essential for GWAs[13–15]. Second, because variants in many genes seem to be in lower LD than variants that lie outside genes[16], they might be more difficult to capture using an indirect approach. To support the validity of this second concern, we describe a quantitative comparison of indirect and gene-centric approaches to GWAs using data from the HapMap ENCODE project. With regard to the feasibility of capturing all casual variants in 'genic' regions, we also determine the number of SNPs that would be required for a comprehensive coverage of genes. Ultimately, the development of such resources to make gene-centric studies possible will improve the efficiency and completeness of any GWA study.

Indirect approaches

It is currently unknown whether the existing SNP sets for indirect GWA approaches of either type — using quasi-random or specifically selected markers — are complete with regard to coverage of SNPs in genes. Also unclear is whether these approaches are fully efficient for detecting causal variants, owing to the large number of SNPs that must be genotyped.

With regard to coverage, previous work indicates that, across the majority of the genome, recombination rates are low within genes[17]. Specifically, SNPs in exons[18], non-conservative coding changes[19] and SNPs in exon–intron boundaries[20] show higher levels of LD than SNPs that lie outside genes but are the same physical distance apart. However, recombination rates in gene-rich regions overall are the highest in the genome. This phenomenon agrees with the Hill–Robertson effect: a high level of recombination around loci that are under selective pressure is beneficial because it allows for selection to act independently on the loci[21]. This suggests that although SNPs that lie within genes might serve as good proxies for each other, SNPs that occur outside genes might do a poor job of capturing variation within genes.

Empirical investigation of coverage. We quantitatively evaluated the coverage (BOX 1) of both genic and non-genic SNPs that is likely to be achieved by following both the quasi-random and the tag SNP-based approaches to indirect GWA studies. For this, we used data on all common SNPs (those with minor allele frequencies (MAFs) of ≥5%) from the HapMap ENCODE regions.

Two recent studies have used the HapMap Phase II data to compare the coverage of the Affymetrix 500K quasi-random SNP set and an LD-based platform, the Illumina HumanHap300 set[22,23], although they did not specifically investigate coverage of genic variants. Here we also evaluate coverage for the Affymetrix 500K set. However, as noted by Pe'er *et al.*[22], the HumanHap300 set was constructed using information from the HapMap Phase I data, which includes data from the ENCODE regions. Therefore, evaluating the HumanHap300 set in the HapMap ENCODE data would lead to an upward bias in the estimate of coverage in our analysis, as this platform is based on a less-complete SNP set across the rest of the genome. For this reason, we do not include the actual Illumina SNP set here, but rather an LD-based tag set. Note that future SNP sets that incorporate HapMap Phase II data should provide even more complete coverage of the genome.

LD tagging SNPs for genotyping can be chosen to ensure that all SNPs, genic or otherwise, are captured at or above a specific value of r^2 (the squared correlation coefficient, a measure of LD between variants; BOX 1). Although SNPs can still be captured with various maximum r^2 values, threshold tagging can eliminate the problem of SNPs being captured at low levels of LD, thereby limiting the effects of variation in coverage on statistical power. The selection of a particular threshold involves a trade-off between more complete coverage and a lower genotyping burden. For these reasons, we chose to evaluate tagging SNP sets that we selected to LD-tag all HapMap ENCODE SNPs (MAF ≥ 5%) at two LD thresholds: $r^2 = 0.8$ and $r^2 = 0.5$. The first ensures that all SNPs are captured at a high level of LD at the cost of genotyping a large number of SNPs; the second provides a 39–40% reduction in the number of SNPs that must be genotyped, although this threshold provides a lower level of coverage for many SNPs[24]. Tagging was implemented using the Tagger server to choose multimarker tags with as many as six markers.

We defined 'genic' SNPs as those SNPs that were annotated as follows in Ensembl: synonymous and non-synonymous coding SNPs, and SNPs in 5′ and 3′ untranslated regions. Although other SNPs that lie near coding regions are also more likely to cause a functional change than SNPs that are further away from genes (owing to effects on RNA processing or transcription regulation[13]), SNPs that lie within introns, particularly those that are not located near intron–exon boundaries, are relatively less likely to have functional significance than those in coding regions or UTRs[14,15].

To measure coverage, we determined the maximum r^2 value between each HapMap ENCODE SNP and the SNPs in the genotyping sets. The cumulative distribution of the maximum r^2 values between SNPs in the quasi-random genotyping set and all common genic and non-genic HapMap ENCODE SNPs (MAF ≥ 5%) is shown in FIG. 2. On the basis of these r^2 values we see that, when using the quasi-random set of

markers, genic SNPs are not as well covered as non-genic SNPs for the Caucasian (CEU) and combined Asian (JPT and CHB) populations. This difference is largely driven by a statistically significant excess of genic SNPs in low LD ($r^2 \leq 0.2$) compared with non-genic SNPs ($P = 0.0012$ in CEU; $P = 0.0025$ in JPT and CHB) (Supplementary information S1 (table)). In the Yoruban African group, there is little difference in the quasi-random SNP set coverage of genic and non-genic SNPs; in fact, there is a slightly higher coverage of genic SNPs for larger r^2 values (Supplementary information S1 (table)). This might, in part, reflect the lower levels of LD that exist overall among the Yoruban African group compared with the Caucasian and Asian groups.

For the tagging SNP sets, coverage of genic SNPs was slightly lower than that of non-genic SNPs in the Caucasian group, indicating that the maximum r^2 values for genic SNPs are more likely to fall closer to the LD threshold (that is, they are more likely to be found at the low end of the LD range) than those of non-genic SNPs (not shown). On the other hand, coverage of genic SNPs in the Asian and Yoruban African groups was similar or slightly higher than that of non-genic SNPs. These results indicate that tagging sets can be selected to give sufficient coverage of genic SNPs, although not all tagging sets can provide equivalent coverage for both genic and non-genic SNPs.

It is also possible to capture many genic SNPs using tagging SNPs from nearby introns and non-coding regions. To determine how well these non-genic SNPs can capture genic SNPs, we constructed a SNP set that included all common (MAF ≥ 5%) intronic SNPs and SNPs that lie within 10 kb of genes, and examined the maximum r^2 values for the common (MAF ≥ 5%) genic SNPs (FIG. 3). The majority of genic SNPs (74–90%) were captured at an r^2 value ≥0.8 in the three populations. However, 9% of genic SNPs in the Caucasian and combined Asian populations and 13% in the Yoruban African population were captured with a maximum $r^2 < 0.5$. This indicates that a tagging set that includes only SNPs that have been selected from introns and nearby regions will not provide comprehensive coverage of high-priority genic SNPs.

Power of indirect studies. When SNP genotyping sets have lower levels of coverage of genic SNPs, as observed above, GWA studies will have a decreased power to detect the causal variants that occur in genes. To quantify the potential difference in power to detect genic and non-genic SNPs, we simulated power on the basis of the empirical distributions of allele frequencies and maximum r^2 values of these types of SNP in the HapMap ENCODE regions, using the cumulative r^2 adjustment for power[25]. We determined power on the basis of a study of 2,000 cases and 2,000 controls, for multiplicative genetic effects on the phenotype that have odds ratios ranging from 1.2 to 2.0, and a genome-wide significance criterion of $\alpha = 10^{-6}$.

Using the quasi-random SNP set, power is decreased by as much as 10% for detecting associations between genic SNPs and the phenotype — compared with non-genic SNP associations — in the Caucasian population (FIG. 4). The greatest difference occurs when the odds ratios and power are high; this is driven by the number of SNPs in low LD, which are difficult to capture. Similarly, in the Asian population, there is a 6% lower power for genic SNPs. In the African population, power is up to 13% higher for genic SNPs when power is low, and 1% greater when the power is high. The power to detect both genic and non-genic SNPs in the African group is lower than the corresponding power in the Caucasian and Asian groups.

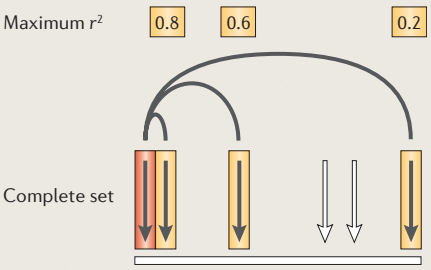
When looking at the tagging SNPs, for both LD thresholds (0.5 and 0.8), power is slightly lower for the genic SNPs among Caucasians (up to 6% lower). Here the largest difference occurs in the middle of the range of odds ratios. For the Asian group, power is 8–10% higher for genic SNPs when using the two tagging sets, again with the largest difference occurring in the middle of the range of odds ratios. For the African population, power was up to 14% higher for genic SNPs when using the two tagging sets, also with the largest difference occurring in the middle of the range of odds ratios. All three groups had similar power to detect both genic and non-genic SNPs at the high and low ends of the range of odds ratios examined.

Our empirical observations that quasi-random SNP sets can provide worse coverage and lower power for genic than non-genic SNPs for the Caucasian and Asian groups, and that tagging SNP sets have lower coverage and power for the Caucasian group, has important implications for the design of indirect GWA studies. In particular, quasi-random and LD tagging SNP sets might require additional SNPs for the complete coverage of genes in some populations.

Gene-centric approaches

Approaches that focus efforts directly on genes can provide two advantages over the indirect approaches that attempt to capture all variants. First, a gene-centric approach could decrease the genotyping burden, an important concern because of the considerable expense

Box 1 | **Coverage and power**



The goal of measuring coverage is to determine how well the SNPs that are part of a genotyping set capture all known variants. Studies of coverage typically use the linkage disequilibrium (LD) measure r^2 (the squared correlation coefficient). For each variant, one can calculate the r^2 between that variant and each SNP in the genotyping set. The highest of these values is the maximum r^2 value, *m*. In the figure, arrows represent SNPs, and the SNPs that are shown on a yellow background represent a subset that have been selected for genotyping. Values are shown for the coverage by these SNPs of the SNP shown in red (which is not genotyped). By determining the maximum r^2 for all SNPs in the complete set, we can estimate the coverage that a particular genotyping set provides.

The maximum r^2 measure can be used to translate coverage to calculate the sample size that is required for an indirect association study. In the figure, the maximum r^2 value is 0.8, corresponding to the value between the SNP of interest (red) and the genotyped SNP that is shown immediately to the right. For a particular variant, the effective sample size of an indirect association study is simply the product of the actual sample size (*n*) and the maximum r^2 value for that variant (*m*); so, as coverage decreases, a larger sample size will be needed to obtain the same power. The overall power of a genome-wide association study can be estimated using the effective sample size for each variant. In the analyses we describe here, we use a metric that is based on the cumulative distribution of maximum r^2 values, or the cumulative r^2 adjusted power, to determine power[25]. Note that caution should be exercised when using summary measures of coverage that are often presented in the literature, such as the average of all maximum r^2 values. Sample size and power do not characteristically vary in a linear way, so using summary coverage measures can overestimate power.

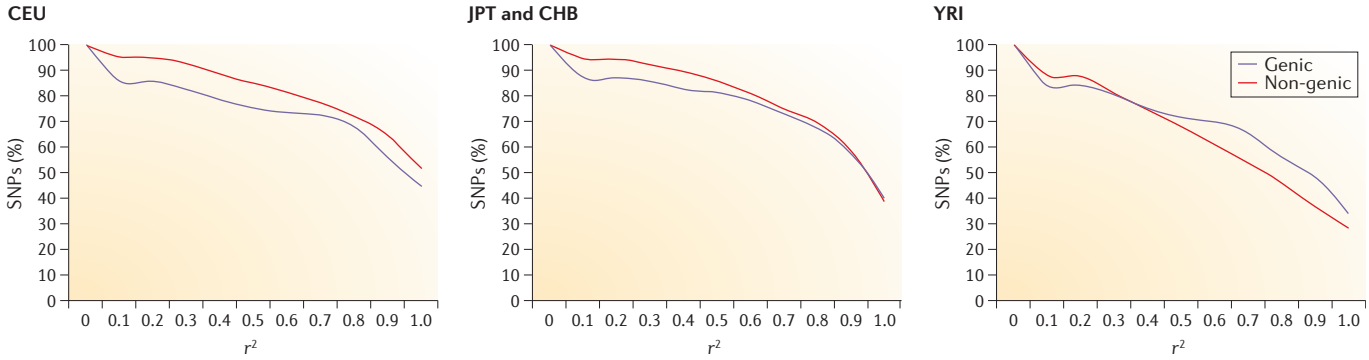


Figure 2 | **Coverage of genic and non-genic SNPs by a quasi-random SNP set.** Coverage, as measured by the cumulative distribution of the maximum r^2 values for each SNP, is lower for genic SNPs in both the Caucasian (CEU) and Asian (JPT and CHB) groups. This is largely owing to an excess of genic variants in low linkage disequilibrium with the SNPs in the genotyping set. This is not the case for the Yoruban African (YRI) group, which has lower overall coverage compared with the Caucasian and Asian groups.

of conducting GWA studies. Second, a gene-centric approach should be more complete with regard to the coverage of genes, which is crucial to detecting causal variants.

Power of the genic SNP approach. Although a gene-centric approach to GWAs requires much less genotyping than an indirect approach, it would be expected to provide little coverage of SNPs that lie outside genes. To investigate this, we used the program Tagger to generate a set of SNPs that tag all common genic SNPs in the HapMap ENCODE regions (MAF ≥ 5%) with an $r^2 \geq 0.8$. We then determined how much power this set of genic-tag SNPs had to detect associations with other genic and non-genic SNPs in these regions. As above, power was calculated for a study of 2,000 cases and 2,000 controls, with odds ratios ranging from 1.2 to 2.0 and a genome-wide significance criteria of $\alpha = 10^{-6}$.

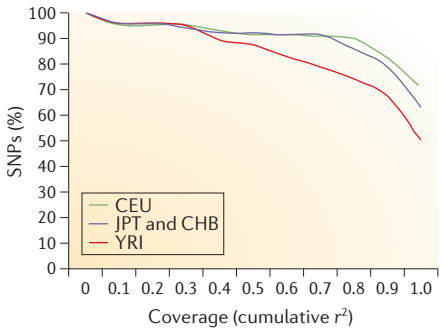


Figure 3 | **Coverage of genic SNPs by a non-genic SNP set.** Coverage of genic SNPs in the HapMap populations using a non-genic SNP set that included all common (minor allele frequencies of ≥5%) intronic SNPs, and SNPs within 10 kb of genes. Coverage is measured by the cumulative distribution of the maximum r^2 values for each SNP.

As expected by design, the genic-tag SNP set had sufficient power to detect causal variants within genes. For example, for a variant with an odds ratio of 1.5 and MAF ≥ 5%, our genic-tag SNP sets have 80% power in Caucasians, 94% in Asians and 88% in Yoruban Africans. However, the power of a genic set to detect non-genic SNPs is considerably lower, equal to 17% in Caucasians, 23% in Asians and 11% in Yoruban Africans. This is not surprising, as many putative causal variants that lie outside genes are in relatively low LD with the genic-tag SNPs. Conversely, although many genic SNPs can be captured using non-genic tags, about 9–13% of genic SNPs have low maximum r^2 values (<0.5), and 5% of genic SNPs cannot be picked up at all by non-genic tags.

Note that the same genome-wide significance level was applied for both the genic and indirect approaches ($\alpha = 10^{-6}$), even though the first approach has substantially fewer tests. Because the genic SNP approach is focused on identifying causal variants in a subset of the genome, a more liberal significance level can be used to reflect the smaller number of tests, resulting in an increase in power and efficiency for the genic approach. For example, using a significance level of 2×10^{-5}, we see an increase in power of 10%, 3% and 6% to detect genic SNPs and 17%, 10% and 21% to detect non-genic SNPs in the Caucasian, Asian and African groups, respectively.

Relative efficiency of GWA approaches

Although there is a loss of power to detect non-genic causal variants when using a genic SNP approach to GWA, this approach has a considerably lower genotyping burden. As shown in TABLE 1, the genic approach requires that far fewer SNPs be genotyped than does the indirect approach. As a result, the genic SNP approach might be more 'efficient' in terms of the effort that is required to detect an association, depending on the proportion of causal variants that reside within genes.

To evaluate this, we determined the efficiency of GWA approaches — here defined as power divided by the genotyping burden — for conditions in which different proportions of causal variants are genic, ranging from 0 to 100%. Again, this was carried out using data from the HapMap ENCODE regions. Our analyses show that, in the Caucasian group, the genic SNP approach is 1.8- to 2.7-fold more efficient than whole-genome approaches, even when all causal variants occur outside genes (FIG. 5). When half of all causal variants lie outside genes, the gains in efficiency for the genic approach are 5.0- to 16.0-fold, compared with whole-genome approaches. For the Asian group, the gains are 2.5- to 3.6-fold when all causal variants occur outside the genes, and 6.0- to 9.3-fold when half of all causal variants lie outside genes. For the African group, the gains are 1.2- to 3.7-fold and 4.9- to 16.0-fold, respectively.

Which GWA approach should be used?

Determining which GWA approach to use depends on how an individual group of researchers wishes to balance completeness, efficiency and *a priori* hypotheses about where causal variants reside for a particular phenotype. For example, rare adverse drug events might occur in only a few hundred subjects each year. Collecting a large enough sample to adequately power an indirect GWA study of such events could take many years. Focusing on a smaller set of genic SNPs and taking advantage of the reduced multiple-testing burden can provide an efficient initial genome-wide association scan that can

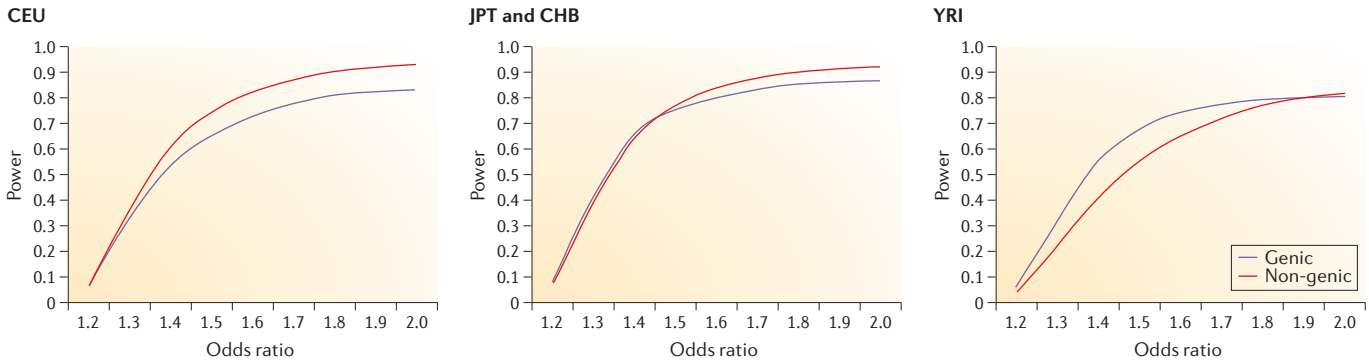


Figure 4 | **Predicted power of genome-wide association studies using a quasi-random SNP set.** The cumulative r^2 adjusted power to detect causal variants that are either genic or non-genic is calculated on the basis of the empirical distributions of coverage and allele frequencies. Power varies depending on the odds ratios of the causal genetic variants (CEU, Caucasian; JPT and CHB, Asian; YRI, Yoruban African).

identify biologically plausible associations while limiting false-positive findings. A second example in which prioritizing variants can lead to greater efficiency is the testing of gene–gene interactions. Testing for all possible interactions requires a number of comparisons that equals the square of the number of variants being tested. By focusing on a subset of high-priority variants, one can dramatically reduce the multiple-testing burden.

Instead of choosing one GWA approach over the other, the most attractive option might be to complement the indirect approach with genic SNPs. For example, indirect GWA studies that use quasi-random SNP sets could add gene-based SNPs to provide more complete coverage of high-priority regions. Alternatively, indirect GWA studies that use an LD-tagging based genotyping set could choose to 'over-capture' genic SNPs by using a more stringent LD threshold in genic regions. This would allow for more complete coverage of genic SNPs, while maintaining a reasonable genotyping burden (that is, in comparison with using more stringent LD criteria for all SNPs).

The genic SNP approach could also be combined with genotyping SNPs in evolutionarily conserved regions to help identify *cis* enhancers, thereby covering all variants with a high prior probability of functional importance. Approximately 5% of the human genome sequence is conserved with the mouse, and about 30% of this is coding sequence. A minimal set of tagging SNPs ($r^2 \geq 0.8$) for genic and conserved non-coding variants in the HapMap ENCODE regions would require 80% fewer SNPs than a complete tagging set ($r^2 \geq 0.8$) in the Caucasian group, 81% fewer in the Asian group and 87% fewer in the African group (TABLE 1). This approach could be particularly attractive for studying African populations, because an LD-tagging SNP set that can provide comprehensive coverage might need to incorporate more than twice as many SNPs as similar sets for Caucasian and Asian populations (TABLE 1). Of course, the potential reductions in the genotyping burden and the ensuing costs from using genic SNPs, either alone or in conjunction with indirect SNP sets, will depend on the expense that is involved with incorporating additional genic SNPs into various genotyping platforms. Given that both Affymetrix and Illumina have platforms that can incorporate specific SNPs, it might be possible to efficiently include additional genic SNPs into future SNP sets.

Developing gene-centric resources

Although the HapMap project focused on providing a set of SNPs that can be used as part of an indirect approach to whole-genome studies that aim to capture all variants, a genome-wide gene-centric SNP set has received less attention. Given the clear functional importance of non-synonymous coding SNPs, genic resequencing efforts in numerous subjects could provide a complete set of high-priority SNPs for gene-centric studies.

The total number of genes in the human genome is probably less than 25,000 (REF. 26). The most recent build of the human genome has identified 22,218 genes (including 1,947 psuedogenes), and another 1,000 to 2,000 protein-coding loci are expected to be identified in the next 5 years (E. Birney, personal communication). Therefore, we now know the location and sequence of more than 95% of human genes, making it possible to identify the vast majority of SNPs that lie within genes. To this end, the Wellcome Trust Exon Resequencing Project has begun to resequence all the known exons in 48

Table 1 | **Multimarker tags for common SNPs in the HapMap ENCODE regions**

	Population		
	CEU	JPT and CHB	YRI
Genotyping sets			
Quasi-random	707	659	712
Tagging $r^2 \geq 0.5$	637	558	1,367
Tagging $r^2 \geq 0.8$	1,036	933	2,246
Genic	76	72	84
Genic and conserved non-coding	212	181	281
Number of common SNPs in the HapMap ENCODE regions			
Total	7,692	6,618	8,481
Genic	140	110	113
Genic and conserved non-coding	400	307	377

Common is defined as a minor allele frequency (MAF) of ≥5%. CEU, Caucasian; JPT and CHB, Asian populations; YRI, Yoruban African; r^2, the squared correlation coefficient.

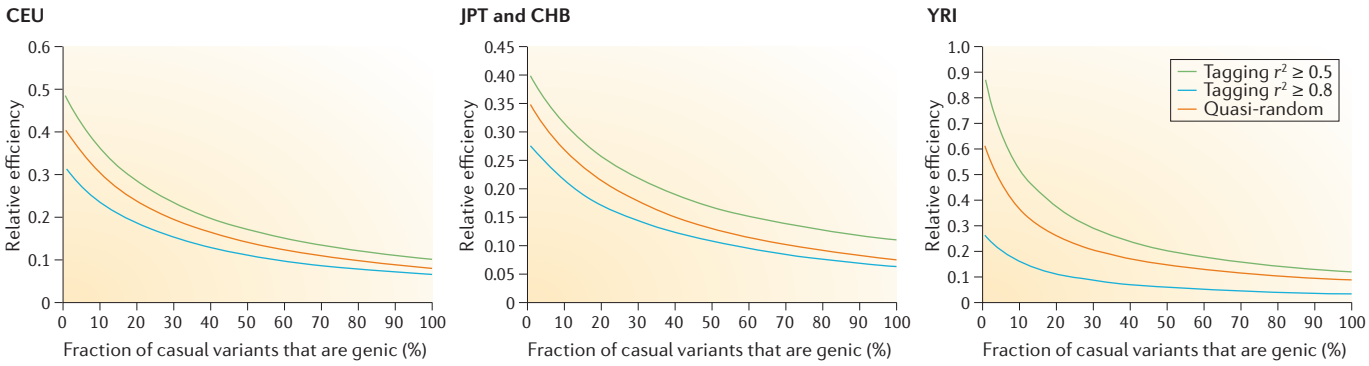


Figure 5 | **Relative efficiencies of approaches to genome-wide association studies.** A comparison of the efficiencies of different approaches based on quasi-random or tagging SNP sets, relative to the efficiency of a genic approach. Note the different scales for efficiency in the three graphs. Efficiency is the power divided by the genotyping burden. Relative efficiency is the efficiency for one approach divided by the efficiency for the genic approach. Values less than one indicate that the genic SNP approach is more efficient than the other approach (CEU, Caucasian; JPT and CHB, Asian; YRI, Yoruban African).

Caucasian subjects. The project has a 99% probability of detecting an exonic SNP with a true population MAF of 5% or greater.

There are several resources that are currently available for gene-based SNPs. For example, EntrezSNP contains 89,931 human synonymous, non-synonymous and UTR SNPs with genotypes and a heterozygosity between 10% and 50% (equivalent to a MAF of ≥5%), including 9,506 non-synonymous SNPs. Furthermore, the HapMap Phase II release has validated a large number of gene-based SNPs, including 9,625 common (MAF ≥ 5%) non-synonymous SNPs in the Caucasian group, 9,418 in the combined Asian group and 10,497 in the Yoruban African group. The more complete HapMap ENCODE regions contain 47, 38, and 41 common non-synonymous coding SNPs in the Caucasian, Asian and Yoruban African groups, respectively. Extrapolating from these regions, which total 5 million bp, to the entire 3,200 million bp of the human genome, the estimated total number of common non-synonymous coding SNPs is 30,080 in the Caucasian group, 24,320 in the Asian group and 26,240 in the Yoruban African group. Similar numbers arise from resequencing projects of candidate genes, which estimate that there are 0.8–1.1 common non-synonymous coding SNPs for each gene, amounting to a total of 20,000–27,500 common non-synonymous SNPs (0.8–1.1 × 25,000) in the human genome[20,27].

On the basis of these estimated numbers, the HapMap project might currently have validated approximately 38–47% of all common non-synonymous coding SNPs. We might then expect the Wellcome Trust Exon Resequencing Project to identify another 11,000 common non-synonymous coding SNPs that have not already been validated by the HapMap project. A study of subjects of African descent might identify an additional 17,000 SNPs. Taken together, these projects will provide valuable resources for studying genes in GWAs.

Of course, creating a comprehensive catalogue of gene-based SNPs requires extensive resequencing in a large number of subjects. Consider the effort that is required to identify all SNPs that lie in exons, of which there are currently known to be 245,231, comprising approximately 60 Mb of DNA. Taking the current cost of sequencing a 500-bp amplicon in one individual to be US$2.00 (P. Kwok, personal communication), and applying this to each exon, the cost of sequencing all exons in one person would be $490,462. We note that multiple subjects could be pooled in each sequencing run for the purpose of discovering SNPs, and then individual subjects could be genotyped later to provide information on SNP frequency and LD between SNPs. Such an approach could cut the number of sequencing reactions, and therefore also the cost, by 50% or more.

The HapMap subjects provide an excellent resource for an exon resequencing project, with a large number of subjects from multiple ethnic groups. The CEU Caucasian and YRI African groups contain 60 unrelated individuals, or 120 independent chromosomes. Resequencing these subjects could identify at least one copy of all common (MAF ≥ 5%) exonic SNPs with a probability of 99.8%, and all SNPs with a

Glossary

Genome-wide significance criterion
The level of significance that an association must reach to reject the null hypothesis of no association, taking into account the large number of tests being conducted.

Linkage analysis
A method for localizing genes that is based on the co-inheritance of genetic markers and phenotypes in families over several generations.

Linkage disequilibrium
The non-random association of alleles of different linked polymorphisms in a population.

Minor allele frequency
The frequency of the less-common allele at a polymorphic locus. It has a value that lies between 0 and 0.5, and can vary between populations.

Multiple-hypothesis testing
The practice of testing more than one hypothesis within an experiment. As a result, the probability of an unusual result from within the entire experiment occurring by chance is higher than the individual probability for one test alone.

Odds ratio
A measurement of association that is commonly used in case–control studies. It is defined as the odds of exposure to the susceptible genetic variant in cases compared with that in controls. If the odds ratio is statistically significantly greater or less than one, then the genetic variant is associated with the disease.

Power
The probability of rejecting the null hypothesis when it is false. In genome-wide association studies, the null hypothesis is that there is no association between a variant and the phenotype of interest.

HapMap
A catalogue of common genetic variation in the human genome that was developed by the International HapMap Project.

MAF of ≥1% with a probability of 70.0%. The combined Asian group contains 90 unrelated subjects for a total of 180 independent chromosomes. Resequencing in this group would have a probability of 99.99% of detecting all common (MAF ≥ 5%) exonic SNPs, and a probability of 83.6% for all SNPs with a MAF of ≥1%. The HapMap Phase II data currently contain 12,027 polymorphic non-synonymous coding SNPs (of any frequency) in the Caucasian group, 12,085 in the Asian group and 13,264 in the Yoruban African group. Resequencing projects that are based on candidate genes, including 24 subjects from each population, have found that there are 1.4–2.0 polymorphic non-synonymous coding SNPs for each gene, or 35,000–50,000 genome wide[27,28], indicating that 27–35% of these are currently available in the HapMap. We note that even this number is likely to be an underestimate of the number of rare (MAF < 5%) SNPs, given the number of chromosomes that must be screened to identify these SNPs[29]. Because rare SNPs are more likely to be specific to one population[20,28], a full catalogue of rare non-synonymous SNPs will require sequencing more subjects from these groups, and collecting subjects from other populations for which rare variants might not be present in the current HapMap samples.

The development of a gene-centric SNP set in a population such as those that were used in the HapMap project would have several benefits. First, such a SNP set would provide a resource for performing the potentially more efficient gene-centric GWA studies. Second, the SNPs that are identified will allow for more complete coverage of high-priority regions that should be included in any comprehensive GWA study. Third, information on linkage disequilibrium among these SNPs can be used to further characterize the effects of individual variants after an initial positive association result. Fourth, because much of the sequence in genes is likely to be evolutionarily conserved, the identification of SNPs within genes will provide a resource for studying the effects of variation in conserved regions. Finally, a full register of common variation in human genes will provide a useful resource for the functional studies that will be needed to assess causality of SNPs that are associated with disease in GWA studies.

Conclusions

In summary, using empirical data from the HapMap ENCODE region, we have shown that quasi-random and tagging SNP sets for indirect approaches to GWA studies can provide lower coverage of genic SNPs than non-genic SNPs, especially among Caucasians. Although indirect GWA studies can have higher overall power than genic SNP studies, we have shown that a genic SNP approach to GWA studies can be more efficient for detecting causal variants than the existing indirect approaches, which attempt to capture information on all variants. Given the greater genotyping efficiency of a genic SNP approach, one might want to combine this approach with the currently available quasi-random or LD-tagging sets of SNPs to maximize coverage of regions with a high prior probability of functional importance, at a minimal cost. We estimate that more than 50% of high-priority non-synonymous coding SNPs have yet to be identified and validated. The identification of additional genic SNPs will make more complete gene-centric studies possible, facilitating the efficient detection of causal variants within genes.

Eric Jorgenson and John S. Witte are at the Department of Epidemiology and Biostatistics, and Center for Human Genetics, University of California, San Francisco, California 94143-0794, USA.
e-mail: Eric.Jorgenson@ucsf.edu
doi:10.1038/nrg1962

1. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
2. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22**, 139–144 (1999).
3. Collins, F. S., Brooks, L. D. & Chakravarti, A. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* **8**, 1229–1231 (1998).
4. Olds, L. C. & Sibley, E. Lactase persistence DNA variant enhances lactase promoter activity in vitro: functional role as a *cis* regulatory element. *Hum. Mol. Genet.* **12**, 2333–2340 (2003).
5. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
6. Altshuler, D. *et al.* A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
7. Ozaki, K. *et al.* Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nature Genet.* **32**, 650–654 (2002).
8. Klein, R. J. *et al.* Complement factor *H* polymorphism in age-related macular degeneration. *Science* **308**, 385–389 (2005).
9. Maraganore, D. M. *et al.* High-resolution whole-genome association study of Parkinson disease. *Am. J. Hum. Genet.* **77**, 685–693 (2005).
10. Conley, Y. P. *et al.* Candidate gene analysis suggests a role for fatty acid biosynthesis and regulation of the complement system in the etiology of age-related maculopathy. *Hum. Mol. Genet.* **14**, 1991–2002 (2005).
11. Rivera, A. *et al.* Hypothetical LOC387715 is a second major susceptibility gene for age-related macular degeneration, contributing independently of complement factor H to disease risk. *Hum. Mol. Genet.* **14**, 3227–3236 (2005).
12. Terwilliger, J. D. & Hiekkalinna, T. An utter refutation of the "Fundamental Theorem of the HapMap". *Eur. J. Hum. Genet.* **14**, 426–437 (2006).
13. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genet.* **33**, 228–237 (2003).
14. Palmer L. J., Cardon, L. R. Shaking the tree: Mapping complex disease genes using linkage disequilibrium. *Lancet* **336**, 1223–1234 (2005).
15. Tabor, H. K., Risch, N. J. & Myers, R. M. Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nature Rev. Genet.* **3**, 391–397 (2002).
16. Smith, A. V., Thomas, D. J., Munro, H. M. & Abecasis, G. R. Sequence features in regions of weak and strong linkage disequilibrium. *Genome Res.* **15**, 1519–1534 (2005).
17. McVean, G. A. *et al.* The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584 (2004).
18. Tsunoda, T. *et al.* Variation of gene-based SNPs and linkage disequilibrium patterns in the human genome. *Hum. Mol. Genet.* **13**, 1623–1632 (2004).
19. Goddard, K. A., Hopkins, P. J., Hall, J. M. & Witte, J. S. Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. *Am. J. Hum. Genet.* **66**, 216–234 (2000).
20. Stephens, J. C. *et al.* Haplotype variation and linkage disequilibrium in 313 human genes. *Science* **293**, 489–493 (2001).
21. Hill, W. G. & Robertson, A. The effect of linkage on limits to artificial selection. *Genet. Res.* **8**, 269–294 (1966).
22. Pe'er, I. *et al.* Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nature Genet.* **38**, 663–667 (2006).
23. Barrett, J. C. & Cardon, L. R. Evaluating coverage of genome-wide association studies. *Nature Genet.* **38**, 659–662 (2006).
24. de Bakker, P. I. *et al.* Efficiency and power in genetic association studies. *Nature Genet.* **37**, 1217–1223 (2005).
25. Jorgenson, E. & Witte, J. S. Coverage and power in genome-wide association studies. *Am. J. Hum. Genet.* **78**, 884–889 (2006).
26. Pennisi, E. Human genome. A low number wins the GeneSweep Pool. *Science* **300**, 1484 (2003).
27. Livingston, R. J. *et al.* Pattern of sequence variation across 213 environmental response genes. *Genome Res.* **14**, 1821–1831 (2004).
28. Crawford, D. C. *et al.* Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. *Am. J. Hum. Genet.* **74**, 610–622 (2004).
29. Pe'er, I. *et al.* Biases and reconciliation in estimates of linkage disequilibrium in the human genome. *Am. J. Hum. Genet.* **78**, 588–603 (2006).

Acknowledgements
We thank N. Risch, D. Thomas, X. Liu and I. Cheng for their helpful comments, as well as L. Edblad and L. Woldin for assistance in the preparation of the manuscript. We also thank anonymous reviewers for their helpful suggestions. This work was supported by grants from the US National Institutes of Health.

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION

Affymetrix 500k Array Set: http://www.affymetrix.com/products/arrays/specific/500k.affx
EntrezSNP: http://www.ncbi.nlm.nih.gov/entrez
Ensembl: http://www.ensembl.org/homo_sapiens/index.html
HapMap ENCODE: http://www.hapmap.org/downloads/encode1.html.en
HapMart: http://hapmart.hapmap.org/BioMart/martview
International HapMap Project: http://www.hapmap.org
National Institute of Environmental Health Science Environmental Genome Project: http://egp.gs.washington.edu/summary_data.html
SeattleSNPs Program for Genomic Applications: http://pga.gs.washington.edu/summary_data.html
Tagger: http://www.broad.harvard.edu/mpg/tagger
Wellcome Trust Exon Resequencing Project: http://www.sanger.ac.uk/Teams/Team76
Witte Laboratory Homepage: http://www.epibiostat.ucsf.edu/witte_lab

SUPPLEMENTARY INFORMATION

See online article: S1 (table)
Access to this links box is available online.