Genome-wide tagging for everyone

Anna C Need & David B Goldstein

The recently completed International HapMap Project has provided detailed information about patterns of genetic variation in four different population samples. Two new studies show that the patterns of variation documented in the HapMap can be applied to other human populations, suggesting it is time to establish a standardized platform for all whole-genome association studies.

The motivation for the HapMap project was to develop a tool that would make it economical to assess how common genetic variation influences common diseases. The basic idea is simple. Polymorphisms in human populations often 'travel together' such that once you have identified one genetic variant, you can often predict the form of many others. The HapMap project set out to describe these patterns of association in four population samples in order to identify a minimum set of SNPs ('tagging SNPs') sufficient to represent all common variants in our genome.

Immediately, however, the project drew intense criticism: there was not enough association in the human genome to allow a meaningful reduction in the number of SNPs that would need to be typed. Variants that were not identified in the reference populations would not be represented. And patterns of association would not be consistent enough among human populations. Convinced that such difficulties would prove insurmountable, some geneticists described the effort as a make-work project for the major sequencing centers that brought us the human genome[1]. These concerns have been progressively resolved[2–5], with one nagging exception: how applicable are HapMap tags to other populations? In this issue, new papers by de Bakker *et al.*[6] and Conrad *et al.*[7] provide convincing reassurance.

Anna C. Need and David B. Goldstein are at the Institute for Genome Sciences and Policy, Center for Population Genomics and Pharmacogenetics, Duke University, Durham, North Carolina 27710, USA.
e-mail: d.goldstein@duke.edu

Tagging works

These papers both focus on the simplest tagging strategy, in which each untyped SNP is represented by a tagging SNP, and the simple correlation coefficient r^2 is used to assess how well the tag predicts the allele present at the untyped SNP (when r^2 is close to 1, there is no loss of power in typing only the tag). In their study, de Bakker

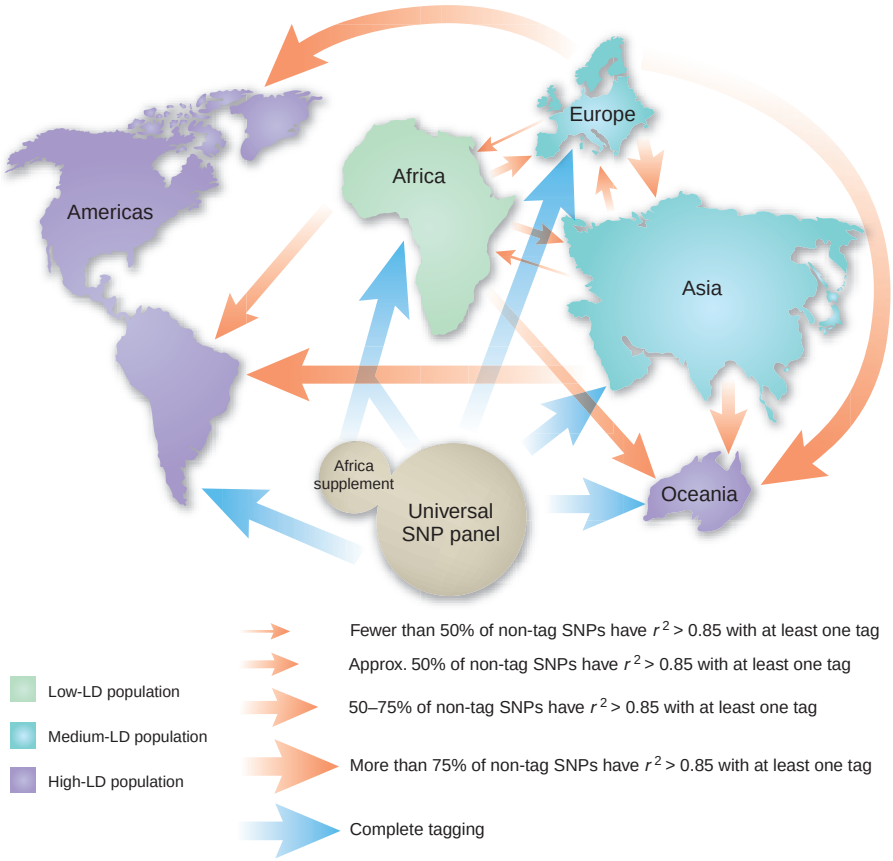

Figure 1 The two determinants of tag transferability are proximity between the reference and target populations and, more importantly, the amount of association between variants (LD) in the target population. In the figure above, arrows originate from the reference population and point to the target population. Because of different patterns of LD among populations, tagging sets could be optimized for specific population groups (for example, the continental groupings reflected in the figure). However, given the now marginal cost of genotyping a small number of extra tagging SNPs, it seems preferable to use a single universal set of SNPs, perhaps including a specific supplement for tagging in populations with significant African ancestry.

et al.[6] initially screened a panel of samples to discover new SNPs and then genotyped these variants in the HapMap samples. Tags were then selected and tested in independent samples from the same population and in samples from different populations. They found that the tags showed only a modest reduction in performance when applied to an independent sample from the same population, although this was more significant in the Asian and Yoruban populations, which could reflect a greater degree of population structure relative to CEPH samples (a narrow sample from Europe).

The authors then studied transferability to other populations with similar ancestry. They tested the Yoruban tags in an African-American population from the Multiethnic Cohort (MEC, which samples from Los Angeles and Hawaii) and in an African-American population from Chicago. The CEPH tags were applied to MEC self-identified 'whites' and individuals from Finland, and the Japanese tags were applied to the MEC Japanese sample. They found that the HapMap tags represented the Japanese and the Caucasian populations very well, but in the African American samples, only 50% of the SNPs had an $r^2 \geq 0.8$, and about 20% had an $r^2 \leq 0.5$.

In addition to the original populations, the authors tested the power of the HapMap tags in Latinos and Native Hawaiians. They found that both of these populations, as well as African Americans, were better represented using a 'cosmopolitan' tag set in which tags are derived from more than one reference population[8].

The authors also used an empirical approach (nominating non-HapMap SNPs in turn as causal variants in simulated case-control panels) to directly test what experimenters are most interested in: how much power is lost when the causal variant is represented by tagging rather than being typed directly. They found very modest power loss in samples of European and Asian ancestory, and even in African Americans, the power was 82% of what it would be if the causal variant were directly ascertained by typing everything. It should be noted that although these analyses do test the applicability of tags derived from the HapMap samples, they cannot be viewed as a direct evaluation of the HapMap resource itself, as the initial SNP set was augmented by SNP discovery in target populations before tag selection occurred.

Conrad *et al.*[7] expand on these findings by evaluating a much broader range of populations. Again using pairwise r^2, they selected three tag sets from the HapMap samples, using only publicly available genotype data from the Phase II release and keeping the number of tags for each set consistent with a genome-wide panel of 400,000 markers. Predictably, these tags represented best populations close to the tag source, but other populations were also tagged well by their closest HapMap population panel (with some exceptions, attributed to admixture).

A key insight emerging from these analyses is that the most important determinant of tag portability is the level of linkage disequilibrium (LD) in the population to be tagged: those with high LD tend to be tagged well, and those with low LD are difficult to tag regardless of their proximity to the tag source population. Consistent with this, the authors found a decline in haplotype diversity with distance from Africa and an increased 'taggability'.

The overall impression from these studies is that tagging works remarkably well and that the HapMap samples provide an appropriate resource for selecting globally useful tags. However, because of their high diversity, African populations will not be represented well unless substantially more tags are used, and even then, many low-LD SNPs will remain unrepresented.

The difficulty of tagging in Africa is clearly illustrated by Illumina's HumanHap650Y, which includes 550,000 tags from the general product and a supplemental 100,000 tags selected to round out tagging in Africa. Analyses of the 'complete' variability data emerging from the ENCODE project show that even with the supplemental SNPs, fewer taggable SNPs are tagged in the Yoruban sample compared with the European one, in which tagging is nearly perfect (S. Dickson and D.B.G., unpublished data). This seems to be an unacceptable asymmetry, and it would be appropriate to develop methods that allow tagging to proceed to a similar degree of efficiency in both European and African population samples. The two possible directions to achieve this are adding more tagging SNPs to the African supplement (**Fig. 1**) or moving beyond simple pairwise approaches in tag selection for African population samples to multimarker approaches. Our own analyses suggest that a combination of these approaches may be the best option. Still, there is also a more fundamental difference that should be noted. Again using the ENCODE data, in the European sample, fewer than 10% of SNPs have no partner SNP with an $r^2 > 0.8$ (*i.e.*, are 'untaggable'), whereas that figure is nearly 20% in the African sample[9]. This means that even if tagging were equally complete in both populations, somewhat less variation would be represented in African samples.

A short game of tag

Thus, tagging works, but can it be used to find anything? Although it is in its early days, there have been encouraging signs. Last year, for instance, Haines *et al.* used tagging to focus on a chromosomal region implicated in age-related macular degeneration[10]. They resolved the original association from 24 million nucleotides to a single coding variant in the complement factor H gene, a result that has since been replicated by a number of researchers.

How, then, should we proceed? One lesson in genetic association studies has been that simplest is sometimes best. For example, despite the reasonable efficiency of pairwise r^2 and how easy it is to implement, the detritus of more sophisticated tagging strategies still fills journal pages, with little impact on real empirical studies. In this regard, the concern raised in both of these papers about optimizing tag selection for individual populations seems out of date. Given that the current SNP per sample cost for the Illumina 550 chip is only $0.001 (and is sure to keep dropping), what does it matter if a few more SNPs than necessary are typed in some populations? Surely the best strategy is to develop a single universal tagging set, perhaps with a supplement for Africa, to be used in all whole-genome association studies (**Fig. 1**). This not only would ensure comparability for replication efforts within complex traits but would also allow direct comparison of the role of the same variants as risk factors for different conditions. Given the possibility that the same gene variants may contribute to the risk of multiple common diseases (as suggested by the comorbidities common in neuropsychiatric and other conditions) it would be a tremendous advantage to have the same sets typed in a broad range of conditions.

Finally, it is well worth bearing in mind that the lifespan of tagging is probably going to be short. Just as large-scale candidate gene studies appeared only recently and are already being replaced in many laboratories by whole-genome tagging, so will tagging be replaced by economical whole-genome sequencing. For this reason, there seems little point in trying to push tagging toward more and more complete representation of variation. Instead it is time to settle on a single platform of variants and type them in a diverse set of cohorts in order to accelerate our rate of discovery. Comprehensive studies of human genetic variation will follow soon enough.

1. Wade, N. *New York Times* 30 October 2002.
2. Hinds, D.A. *et al. Science* **307**, 1072–1079 (2005).
3. de Bakker, P.I. *et al. Nat. Genet.* **37**, 1217–1223 (2005).
4. Montpetit, A. *et al. PLoS Genet.* **2**, e27 (2006).
5. Mueller, J.C. *et al. Am. J. Hum. Genet.* **76**, 387–398 (2005).
6. De Bakker P.I. *et al. Nat. Genet.* **38**, 1298–1303 (2006).
7. Conrad, D.F., *et al. Nat. Genet.* **38**, 1251–1260 (2006).
8. Ahmadi, K.R. *et al. Nat Genet.* **37**, 84–89 (2005).
9. International HapMap Consortium. *Nature* **437**, 1299–1320 (2005).
10. Haines, J.L. *et al. Science* **308**, 419–421 (2005).