

# MaGIC: a program to generate targeted marker sets for genome-wide association studies

Claire L. Simpson<sup>1</sup>, Valerie K. Hansen<sup>1</sup>, Pak C. Sham<sup>1,2</sup>, Andrew Collins<sup>3</sup>, John F. Powell<sup>1</sup>, and Ammar Al-Chalabi<sup>1</sup>

<sup>1</sup>Institute of Psychiatry, Kings College London, London, UK, <sup>2</sup>University of Hong Kong, Hong Kong, and <sup>3</sup>University of Southampton, Southampton, UK

*BioTechniques* 37:996-999 (December 2004)

*High-throughput genotyping technologies such as DNA pooling and DNA microarrays mean that whole-genome screens are now practical for complex disease gene discovery using association studies. Because it is currently impractical to use all available markers, a subset is typically selected on the basis of required saturation density. Restricting markers to those within annotated genomic features of interest (e.g., genes or exons) or within feature-rich regions, reduces workload and cost while retaining much information. We have designed a program (MaGIC) that exploits genome assembly data to create lists of markers correlated with other genomic features. Marker lists are generated at a user-defined spacing and can target features with a user-defined density. Maps are in base pairs or linkage disequilibrium units (LDUs) as derived from the International HapMap data, which is useful for association studies and fine-mapping. Markers may be selected on the basis of heterozygosity and source database, and single nucleotide polymorphism (SNP) markers may additionally be selected on the basis of validation status. The import function means the method can be used for any genomic features such as housekeeping genes, long interspersed elements (LINES), or Alu repeats in humans, and is also functional for other species with equivalent data. The program and source code is freely available at <http://cogent.iop.kcl.ac.uk/MaGIC.cogx>.*

is only now underway. We have started to use these data by developing maps with distances in LD units (LDUs) for a subset of the HapMap data used in the ENCODE project. Our approach uses genome assembly data to identify areas rich in particular annotated features of interest, such as genes or exons (9) and matches these with markers at a user-defined marker spacing. Except for the ENCODE region we have already converted, marker spacing is defined currently in base pairs, but the remainder will be converted to LDUs when whole genome LD information becomes available from the HapMap Project.

## MATERIALS AND METHODS

### Computer Program

A database was constructed from the July 2003 construction of the human genome [National Center for Biotechnology Information (NCBI) build 34; <http://genome.ucsc.edu/>] containing both known and predicted genes, their exons, National Institutes of Health (NIH) reference SNPs from dbSNP, and polymorphic microsatellites with unambiguous genetic and physical map data from the Marshfield (4) and ABCC databases (5). A program was designed (Marker Gene Interspacing and Correlation; MaGIC) to use the database to create feature-targeted marker lists. MaGIC is written in Visual FoxPro and compiled for Microsoft® Windows® 95, 98, ME, NT, 2000, and XP platforms.

### Algorithm for Targeting Genes

Each gene with start position  $s$  base pairs is placed into one of many sequential bins, each  $w$  base pairs wide, the bin number  $n_b$  calculated by:

$$n_b = \left( \frac{s - \lfloor \frac{s}{w} \rfloor}{w} \right)$$

The rank order of bins by number of genes per bin is used to define gene-rich regions by counting the number of genes in each bin and then ranking by gene count in descending order. This allows gene-dense regions to be specifically targeted at the expense of gene-poor regions. A user-defined cutoff of

## INTRODUCTION

High-throughput genotyping technologies such as DNA pooling (1,2) and DNA microarrays (3) mean that whole genome screens are now practical for complex disease gene discovery using association studies. Databases such as the Marshfield genetic map (4), the Advanced Biomedical Computing Center (ABCC) database (5), or the single nucleotide polymorphism (SNP) database dbSNP (6) contain large numbers of polymorphic markers. It is currently impractical to use all available markers, and a subset is typically selected on the basis of required saturation density. Evenly spacing markers is not efficient, as genes are not evenly spaced along chromosomes (7), and linkage disequilibrium (LD) is variable between and within chromosomes (8). Thus, for disease locus discovery, a restricted subset of markers could have the same information as the full set. Restricting markers to those

within annotated genomic features of interest, or within feature-rich regions, reduces workload and cost while retaining much information. For instance, markers could be restricted to those within exons, which occupy only a small proportion of the genome. Moreover, the identity of functional elements is continually being determined, and this information is being added to the genome annotation (e.g., as part of the ENCODE project; <http://www.genome.gov/10005107>). Another endeavour, the International HapMap project (<http://www.hapmap.org/>), is designed to delineate the LD structure of the genome and generate a highly informative but restricted set of markers that tag most common genetic variation. Combining this LD information with the targeting of genomic features may ultimately yield the most efficient choice of markers, but the HapMap data are available in preliminary form, and the conversion of the genotype data to a genome-wide representation of LD

genes per bin is used to define valid bins and one marker selected per bin. In this algorithm, marker spacing corresponds to bin width, so marker density is a function of bin size. A nested algorithm allows multiple markers to be selected for each bin if required.

### Determination of LDUs

Both gene position and bin width can be measured in LDUs instead of base pair units. LD maps (10) are constructed from the Malecot equation,

$$\rho = (1-L)Me^{-\varepsilon d} + L$$

modeling the decline in association  $\rho$  as a function of distance  $d$ . The parameters of the model include  $M$ , which is the maximum association at zero distance, reflecting association at the last major bottleneck.  $L$  is the residual association at large distance, and  $\varepsilon$  is the exponential decline of  $\rho$ , with distance in kilobases. The Malecot parameters  $\varepsilon$  and  $M$  are estimated by fitting multiple pairwise association probabilities,  $\rho$ , and corresponding information,  $K\rho$ , using composite likelihood. Construction of LD maps requires the estimation of the  $\varepsilon$  parameter within each map interval, and a distance in LDUs is defined as  $\varepsilon d$ , where  $d$  is the interval width in kilobases.

### Statistical Power Estimation

The targeting process was modeled for the physical map. Model assumptions were of equal sized bins, each containing an informative marker in the center. We assumed average LD ( $R^2$ ) decay of  $\frac{1}{4N_e\theta}$ , where  $\theta$  is derived according to the Haldane map function of  $\theta = \frac{1}{2}(1 - e^{-2\omega})$ ,  $N_e$  is the effective population size (taken to be 10,000 for the human population), and  $\omega$  is genetic distance in cM (11). Since the distribution of LD is not yet known in detail, we used an average of 0.88 Mb/cM (12). A disease-associated or phenotype-modifying locus was regarded as being equally likely to be found at any position within the bin and modeled as a single point based on the start position for each gene. The expectation for detection of a locus using a completely informative marker was defined as the mean LD for the bin. The noncentrality parameter ( $\lambda$ ) of the chi-squared distribution for a test of a locus contributing to a quantitative phenotype was taken as  $\lambda = n\rho \frac{V_a}{(1-V_a)}$ , where  $n$  is the number of samples,  $\rho$  is the mean LD within a bin, and  $V_a$  is the proportion of phenotypic variance accounted for by the locus (modified from Reference 11). Power can then be estimated using the Genetic Power Calculator (13). A

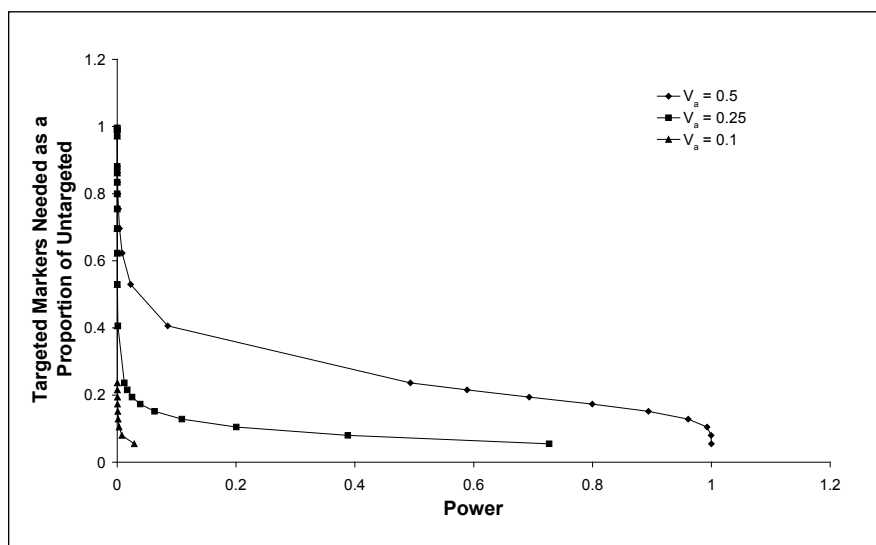
locus contributing 50% to phenotypic variance was modeled assuming a sample size of 1000. Chromosome 1 was used as an example, and the results were extrapolated genome-wide. To examine the effect of this approach on power, a comparison was made between a classic evenly spaced marker set and the targeted marker set generated by the program.

### RESULTS AND DISCUSSION

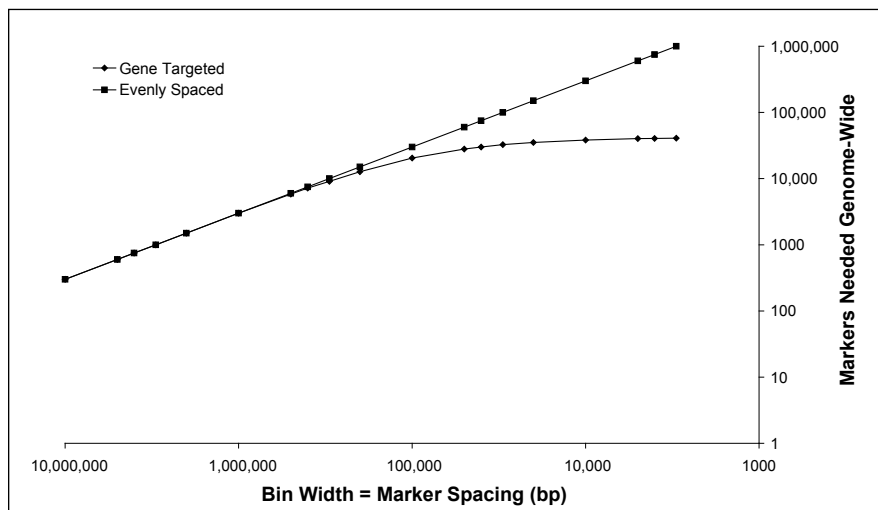
MaGIC is a graphical user interface-based program designed for generating genome-wide feature targeted marker lists. The current version of the underlying database contains marker data for 449,740 microsatellites and 9,178,414 SNPs. The program allows markers of any spacing and targeting genomic features at a user-defined density, to be selected on the basis of heterozygosity and source database. SNP markers may additionally be selected on the basis of validation status and their use by the HapMap project. Markers can be evenly spaced in LDUs, or fractions of LDUs, which is a very efficient selection method for association studies. The import function means that the method can also be used for other genomic features such as housekeeping genes, long interspersed elements (LINES) or Alu repeats in human and other species with equivalent data.

In MaGIC, the genome is modeled as a series of bins, with bin width equivalent to marker spacing. Markers and annotated features are sorted into the bins based on physical or LDU map position. As an example, we demonstrate the method by targeting genes in physical units, but any annotated feature (e.g., exons, promoters, introns, etc.) can be targeted with this program.

The statistical model showed that power increases as a function of marker density, but marker number required increases as a function of the density of the feature being targeted. In the case of gene targeting, because of the nonrandom distribution of genes, while marker density increased exponentially, the required marker number increased much more slowly for 100-kb resolutions or finer. As a result, >95% power could be achieved with 40,000 markers



**Figure 1. Targeted versus untargeted marker sets.** Proportion of gene-targeted markers needed to achieve the same power as untargeted evenly spaced markers for a given phenotypic variance at a locus ( $V_a$ ) under the modeled assumptions. For example, a locus contributing 50% of phenotypic variance can be detected with 80% power with about one-fifth of the markers needed for an evenly spaced screen of the same power assuming genes act as a point and linkage disequilibrium (LD) decays according to the Haldane function. Note that for smaller effect sizes, the proportion drops for equivalent power because the number of evenly spaced markers needed rises exponentially.



**Figure 2. Number of markers needed.** Graph showing the number of markers in a gene-targeted or an evenly spaced marker set for different bin widths, on a logarithmic scale. In the untargeted case, this is simply an evenly spaced marker set with the required resolution. For the targeted set, only bins containing a gene were targeted with a marker.

under our modeled assumptions (targeting a single marker per gene, with each gene acting as a point), corre-

sponding to 600,000 evenly spaced markers and a 5-kb marker spacing (bin width) (Figures 1 and 2). Because the

curve is so flat at 5-kb marker spacing, this marker set is almost equivalent to a million evenly spaced markers and 3-kb marker spacing. In fact, for any screen denser than 300 kb, gene targeting is more efficient than an evenly spaced marker set, but the saving is not large until screens of 100 kb or finer are contemplated, which required 70% or less of the evenly spaced marker set. In reality, one would target each exon with an SNP or a haplotype block with an informative set of SNPs, but for the simplicity of modeling here, we used a simple gene-targeted approach. Nevertheless, this demonstrates that a targeted approach is efficient in time and workload and therefore in cost.

Generating a targeted marker list has several immediate benefits. First, for those wishing to fine-map a region of interest, this method allows the automatic selection of targeted markers at any particular density. Second, for existing commercial or

in-house marker sets, it is possible to check to what extent the markers target genomic features of interest, such as coding or promoter regions, given that it is unlikely that the set will have been designed for such a purpose. The program also allows for the import of custom genome data, so in-house annotations or data from other species can be used. This of course also allows users to update the program with the latest data sets as needed. Finally, for those wishing to design a marker set targeting features such as promoters, genes, or exons, this presents an automated method that can generate a highly efficient genome-wide marker set (8). The genome-wide set can be targeted at features with any set threshold of density, so that for example, only the most gene-rich regions that contain 80% of all genes at a particular resolution could be targeted. This allows a trade-off between workload, cost, and power.

Existing online databases also allow selection of markers with a powerful set of selection tools. There are a number of differences with MaGIC, particularly the ability to evenly space in LDUs rather than physical map position. MaGIC also allows marker selection of user-defined spacing with the density of features as a selection parameter. The ability to target any annotated feature and to import custom marker or feature lists also allows great flexibility. This is particularly true for fine-mapping an area that may be a few megabases in size, but in which the first pass requires searching available SNPs prioritized in some way. Importing also allows the search to be restricted, for example, to markers already available to the laboratory. Custom data from other species can also be imported.

LDU maps have been shown to have additive distances given sufficient marker density (14). In practice, this can be achieved stepwise by constructing interim maps and adding additional markers within "holes," where the number of LDUs in an interval has reached a threshold set at 3 (15). Morton et al. (16) defined a "swept radius," the extent of "useful" LD, as  $1/\epsilon$ , and LDU maps have the property that one LDU corresponds to one swept radius. Equal spacing of markers on the LDU scale,

rather than the kilobase scale is optimal for LD mapping by association, although the relationship between marker density within each LDU and power for gene mapping has not yet been established. The LD map of an ENCODE region used in this version of the program spans 8.4 LDUs in approximately 500 kb, implying approximately 50,000 LDUs in the genome. This is consistent with an earlier estimate (15).

We have designed a method to provide both a rapid automated check of the distribution of available marker sets and targeted marker selection, either genome-wide in feature-rich regions or for fine-mapping studies in the final stages of gene hunting projects. The MaGIC Visual FoxPro source code and executable is available freely from the authors at <http://cogent.iop.kcl.ac.uk/MaGIC.cogx>.

#### ACKNOWLEDGMENTS

*This work is funded by grants from the Medical Research Council G9901258, Wellcome Trust 055379, National Eye Institute EY12562, and Motor Neurone Disease Association. C.S. is funded by Guy's, King's, & St. Thomas' (GKT) Medical School Ph.D. Studentship. At the time of this work, A.A.C. was a Medical Research Council Clinician Scientist Fellow.*

#### COMPETING INTERESTS STATEMENT

*The authors declare no conflicts of interest.*

#### REFERENCES

1. Fisher, P.J., D. Turic, N.M. Williams, P. McGuffin, P. Asherson, D. Ball, I. Craig, T. Eley, et al. 1999. DNA pooling identifies QTLs on chromosome 4 for general cognitive ability in children. *Hum. Mol. Genet.* 8:915-922.
2. Daniels, J., P. Holmans, N. Williams, D. Turic, P. McGuffin, R. Plomin, and M.J. Owen. 1998. A simple method for analyzing microsatellite allele image patterns generated from DNA pools and its application to allelic association studies. *Am. J. Hum. Genet.* 62:1189-1197.
3. Meltzer, P.S. 2001. Spotting the target: microarrays for disease gene discovery. *Curr. Opin. Genet. Dev.* 11:258-263.

4. Broman, K.W., J.C. Murray, V.C. Sheffield, R.L. White, and J.L. Weber. 1998. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* 63:861-869.
5. Collins, J.R., R.M. Stephens, B. Gold, B. Long, M. Dean, and S.K. Burta. 2003. An exhaustive DNA micro-satellite map of the human genome using high performance computing. *Genomics* 82:10-19.
6. Sherry, S., M. Ward, and K. Sirotkin. 1999. dbSNP—database for single nucleotide polymorphisms and other classes of minor genetics variation. *Genome Res.* 9:677-679.
7. Oliver, J.L., P. Carpena, R. Roman-Roldan, T. Mata-Balaguer, A. Mejias-Romero, M. Hackenberg, and P. Bernaola-Galvan. 2002. Isochore chromosome maps of the human genome. *Gene* 300:117-127.
8. Goldstein, D.B. 2001. Islands of linkage disequilibrium. *Nat. Genet.* 29:109-111.
9. Inglehearn, C.F. 1997. Intelligent linkage analysis using gene density estimates. *Nat. Genet.* 16:15.
10. Maniatis, N., A. Collins, C-F. Xu, L.C. McCarthy, D.R. Hewett, W. Tapper, S. Ennis, X. Ke, and N.E. Morton. 2002. The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. *Proc. Natl. Acad. Sci. USA* 99:2228-2233.
11. Sham, P.C., S.S. Cherny, S. Purcell, and J.K. Hewitt. 2000. Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *Am. J. Hum. Genet.* 66:1616-1630.
12. Strachan, T. and A.P. Read. 1999. Genetic mapping of mendelian characters: genetic markers. *In* T. Strachan and A.P. Read (Eds.), *Human Molecular Genetics 2*. Bios Scientific Publishers Ltd, Oxford.
13. Purcell, S., S.S. Cherny, and P.C. Sham. 2003. Genetic power calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* 19:149-150.
14. Ke, X., S. Hunt, W. Tapper, R. Lawrence, G. Stavrides, J. Ghori, P. Whittaker, A. Collins, et al. 2004. The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Hum. Mol. Genet.* 13:577-588.
15. Tapper, W.J., N. Maniatis, N.E. Morton, and A. Collins. 2003. A metric linkage disequilibrium map of a human chromosome. *Ann. Hum. Genet.* 67:487-94.
16. Morton, N.E., W. Zhang, P. Taillon-Miller, S. Ennis, P.-Y. Kwok, and A. Collins. 2001. The optimal measure of allelic association. *Proc. Natl. Acad. Sci. USA* 98:5217-5222.

Received 4 June 2004; accepted 9 August 2004.

*Address correspondence to Ammar Al-Chalabi, Department of Neurology, Institute of Psychiatry, Kings College London, De Crespigny Park, London SE5 8AF, UK. e-mail: ammar@iop.kcl.ac.uk*