

271

Towards a fine-scale linkage disequilibrium map of human chromosome 20. L.R. Cardon¹, X. Ke¹, F. Lawrence¹, N. Carter², J. Rogers², G. Stavrides², D. Willey², J. Mullikin², S. Hunt², D.R. Bentley², P. Deloukas². 1) Wellcome Trust Ctr Human Genet, Univ Oxford, Oxford, United Kingdom; 2) Wellcome Trust Sanger Institute, Hinxton, United Kingdom.

Genome-wide maps of linkage disequilibrium are being developed to facilitate association studies of complex diseases. At present, it is not clear how many SNPs will be needed to meet this objective, nor how to usefully assess LD to guide marker selection. Various definitions of haplotype blocks have been put forth, as have different metrics based on genetic maps. Until recently, little genomic data were available to fruitfully compare such measures. We have constructed a dense map < 1 SNP/2 kb of chromosome 20 to evaluate LD profiles and statistical measures. Chromosome 20 was flow-sorted from a Caucasian, an African American, an African Pygmy and a Chinese cell line, and individual small-insert libraries were shotgun sequenced to a depth of 2x coverage. Over 130,000 new SNPs were discovered, of which 60,000 assays have been designed and are being genotyped across DNA panels of CEPH families, and unrelated Caucasians, African-Americans and Asians. We have completed a 10 Mb region of 20q12-13.2, comprising 5293 SNPs with minor allele frequency $\geq .04$. To evaluate the effects of map density and allele frequency, we took random subsets of the dense map and applied the most commonly used block definitions and genetic map-based methods, then assessed the robustness of the different approaches against the unselected marker panel. The results indicate the extent to which LD at coarse SNP densities reflects genuine patterns and illustrate the strengths/weaknesses of the different statistical approaches. We show that sequence coverage and boundaries of haplotype blocks are dependent on marker density, such that increasing density yields more blocks of apparently shorter length. Genetic-map based approaches appear more robust to marker density, but can be heavily influenced by the allele frequencies of the markers selected to study. These outcomes have implications for association studies, which we discuss in the context of future expectations and past examples.

273

How useful are the tagging SNPs for identifying complex disease genes? H. Zhao^{1,2}, R. Pfeiffer², M. Gail². 1) Epidemiology & Public Health, Yale Univ Sch Medicine, New Haven, CT; 2) Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD.

Although millions of genetic polymorphisms have been identified in the human genome, due to linkage disequilibrium, a small proportion of these markers may suffice to capture the majority of the diversity due to linkage disequilibrium and to identify complex disease genes. Several methods have been proposed to select these representative markers, commonly called haplotype tagging SNPs. One use of these representative markers is to study aspects of population genetics. For this purpose, selecting markers that capture the diversity or otherwise represent the full set of markers seems reasonable, and selection can be based on a random sample of individuals from the population. A second goal is to select a set of markers to detect an association of disease with haplotype (disease gene mapping). We investigated the usefulness of three methods that have been proposed to select markers that preserve information or diversity in population based samples for identifying disease associations with haplotypes in case-control studies: entropy, a haplotype diversity statistic defined by Clayton, and Strams correlation measure. We examined five genes (ADH, APOE, CCR2/CCR5, COMT, NOD2) with known disease associations. We found that these three selection criteria procedures designed to preserve information or diversity often lead to the selection of markers with poor power to detect the disease association. In contrast to these population-sample based tagging SNP selection criteria, alternative strategies that incorporate disease information in SNP selections, such as a two-stage design that uses data on cases and controls in the initial stage, may offer a more powerful approach to selecting most informative markers for disease gene identification.

272

The first metric linkage disequilibrium map of a human chromosome. W.J. Tappe, N. Maniatis, N.E. Morton, A. Collins. Human Genetics, School of Medicine, General Hospital, University of Southampton, Southampton, SO16 6YD, UK.

Recent descriptions of linkage disequilibrium (LD) patterns have focused on delimiting blocks corresponding to regions of low haplotype diversity. The HapMap project intends to aid positional cloning by determining common haplotype patterns within these blocks so that only a few haplotype tag single-nucleotide polymorphisms (tSNPs) need to be typed to define each haplotype. However, there are difficulties since block definitions are arbitrary and only a proportion of the genome is composed of blocks. Furthermore, the relationship between blocks is important since LD may extend across blocks. In comparison, a metric LD map, with map distances analogous to the centiMorgan (cM) scale of linkage maps provides additional information by characterising inter-block regions which define the relationship between blocks. Such maps avoid arbitrary block definitions and give an additive scale that is useful to determine optimal marker spacing for positional cloning. LD maps also provide insights into the relationship between sequence motifs and recombination since the pattern of LD is closely related to recombination hot-spots and their resolution is higher than existing linkage maps. Using LDMAP to analyse SNP data spanning chromosome 22 (Dawson et al., 2002), we have obtained the first whole-chromosome metric LD map. This map identifies regions of high LD as plateaus and regions of low LD as steps reflecting variable recombination intensity. The intensity of recombination is related to the height of the step and thus intense recombination hot-spots can be distinguished from more randomly distributed historical events. The map identifies holes in which greater marker density is required and defines the optimal SNP spacing for positional cloning which suggests that some multiple of around 50,000 SNPs will be required to efficiently screen Caucasian genomes. Further analyses which investigate selection of informative SNPs and the effect of SNP allele frequency and marker density will refine this estimate. The map is also closely correlated with the most recent high-resolution linkage map and a range of sequence motifs including GT/CA repeats and GC content.

274

Determinants of success in genetic association studies of complex traits. K. Zondervan, A. Morris, L. Cardon. Wellcome Trust Ctr Human Gen, Univ Oxford, Oxford, United Kingdom.

Case-control studies of multifactorial diseases and genetic variants have been notable by their lack of success and replicability, partly due to poor epidemiological practice. However, the apparent effect size - the marker-odds ratio (OR) - is statistically determined by 4 parameters: the OR of the disease variant; disease and marker allele frequencies (DAF & MAF), and linkage disequilibrium (LD) between marker and disease variant. We derived this relationship, showing that under complete LD, and MAFDAF, the relationship between marker and disease OR reduces to a simple expression. We used 5 complex disease associations to illustrate: Deep Vein Thrombosis & FVL (DAF: 0.03); Crohn's disease & NOD2 (0.06); Alzheimer's & APOE (0.15); Bladder cancer & GSTM1 (0.7); and NIDDM & PPAR (0.85). Empirical LD data from chromosome 19 was used to investigate the variability in common MAFs within haplotype blocks.

ORs of the low frequency disease alleles were substantial: 3.3-4.6 (allelic ORs), and 11.7-40.0 (homozygote GRRs). Using a range of marker frequencies in moderate LD with the disease loci ($D > 0.5-0.6$), allelic ORs decreased considerably but remained detectable with 80% power in a sample of 1000 cases and 1000 controls. ORs for the common disease alleles were low (allelic ORs: 1.2-1.3; GRRs: 1.5-2.0). Using samples of 5000 cases and 5000 controls, MAFs had to closely resemble DAFs to allow detection with 80% power but only when $D > 0.7$. Chromosome 19 data showed that ~60% of markers were within 0.1 of the most common marker frequency within blocks.

Associations of complex traits with rare alleles conferring large ORs, and common alleles with modest ORs should be detectable in large case-control studies using common markers (MAF > 0.1) and information on genomic LD patterns. Currently, rare alleles with small ORs will not be detectable in feasible samples, unless rare markers in very high LD with the disease locus are used. The development of the HapMap aimed at providing a genomic map of LD may provide help in study design.