The use of a genetic map of biallelic markers in linkage studies

Leonid Kruglyak

Improvements in genetic mapping techniques have driven recent progress in human genetics. The use of single nucleotide polymorphisms (SNPs) as biallelic genetic markers offers the promise of rapid, highly automated genotyping. As maps of SNPs and the techniques for genotyping them are being developed, it is important to consider what properties such maps must have in order for them to be useful for linkage studies. I examine how polymorphic and densely spaced biallelic markers need to be for extraction of most of the inheritance information from human pedigrees, and compare maps of biallelics with today's genome-scanning sets of microsatellite markers. I conclude that a map of 700–900 moderately polymorphic biallelic markers is equivalent—and a map of 1,500–3,000 superior—to the current 300–400 microsatellite marker sets.

The revolution in human genetics that has unfolded over the past decade and a half has been driven largely by the development of genetic maps. The original concept was proposed by Botstein *et al.*, with restriction fragment length polymorphisms (RFLPs) as markers[1]. The first human RFLP was quickly identified[2], and Huntington's disease soon became the first autosomal disorder linked to an anonymous DNA marker[3]. The first RFLP map of the human genome followed shortly[4]. RFLPs were based on a variety of polymorphisms at the sequence level (single nucleotide changes, insertions and deletions, repeat length polymorphisms) and were assayed by Southern hybridization. Although a great advance, RFLPs were often not very polymorphic, and they were costly and time-consuming to develop and assay in large numbers. Nevertheless, these markers made human molecular genetics a reality and led to the mapping of a number of important mendelian diseases.

The next major advance came with the discovery and development of microsatellites (STRs or SSLPs) as markers[5]. These loci are abundant, have fairly high polymorphism rates and can be assayed by PCR, leading to lower cost and a greater degree of automation. Dense maps of microsatellites are now available[6,7], allowing simple mendelian diseases to be mapped with relative ease and enabling first searches for genetic causes of complex diseases by genome scan. However, the requirements to assay the loci on gels and to distinguish several length-based alleles make it hard to fully automate the genotyping process, and typing large numbers of individuals for markers covering the genome remains beyond the resources of all but a few labs. There is thus a need to move beyond this current technology.

Recent attention has focused on the use of single nucleotide polymorphisms (SNPs) as genetic markers. At first glance, this may appear to represent a step back to the days of low polymorphism rates characteristic of RFLPs. However, modern technology should allow efficient assays of SNPs in numbers sufficiently large to offset their lower polymorphism rates, as discussed below. SNPs offer a number of important advantages over microsatellites. They are highly abundant, with classic estimates of more than 1 per 1,000 base pairs, or more than 3 million in the genome[8,9]. To date, more than 1,000 PCR-amplifiable SNP markers have been discovered and mapped (D. Wang, pers. comm.). Because SNPs have only two (common) alleles (hence the term 'biallelics'), genotyping them requires only a plus/minus assay rather than a length measurement, permitting easier automation. Several non–gel-based assays have been proposed[10–14], with high-density oligonucleotide arrays currently showing great promise for typing large numbers of biallelic markers in parallel[15,16].

Here I consider the feasibility of carrying out linkage studies with a genetic map based on biallelic markers. The key questions are: What level of polymorphism is required? and How many markers adequately cover the genome? These questions are addressed below.

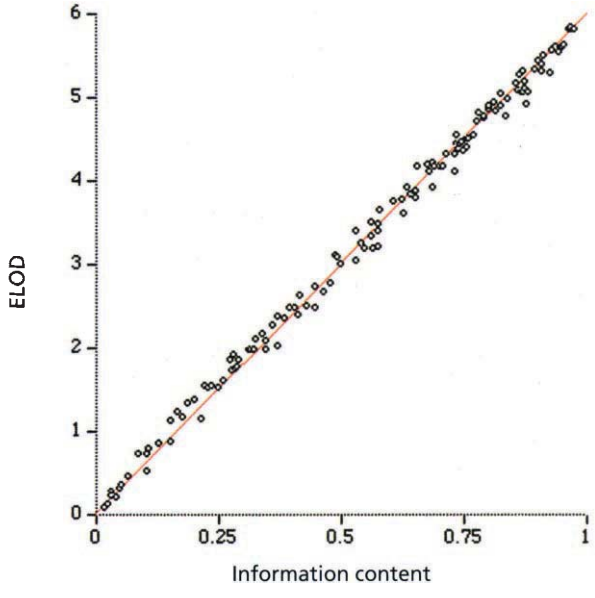

Fig. 1 Expected lod score (ELOD) for a dominant locus is plotted against information content. Each circle represents the results of a simulation for one of 130 maps, as described in Methods. The solid line shows the expected linear correlation if information content of 0 corresponds to an ELOD of 0 and information content of 1 corresponds to the maximum achievable ELOD of 6.02 in these pedigrees.

Assumptions

The effects of marker density and polymorphism were examined by simulating pedigree genotype data and measuring the information content[17,18] for a broad range of map densities and polymorphism levels (see Methods for simulation details). Information content measures the fraction of inheritance information extracted by the map relative to that which

Whitehead Institute for Biomedical Research, 1 Kendall Square, Bldg. 300, Cambridge, Massachusetts 02139, USA. email: leonid@genome.wi.mit.edu

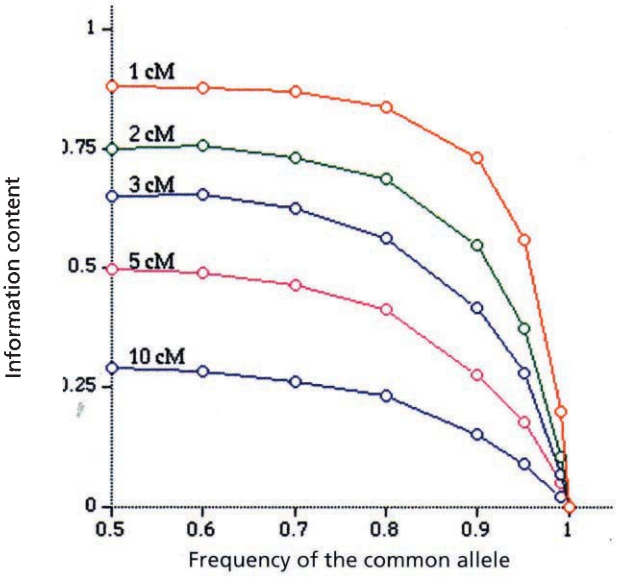

Fig. 2 Information content for five map densities is plotted against the frequency of the more common of the two alleles of a biallelic marker. The circles show actual simulation data points.

would be extracted by an infinitely dense polymorphic map. Thus, an information content of 1 reflects complete information, whereas an information content of 0 reflects no information. Information content incorporates both marker density and polymorphism in a single general measure of map quality that is independent of assumptions about a particular disease locus. It also closely predicts the power of a map to detect linkage—for example, as measured by the expected lod score (ELOD; Fig. 1).

The markers were assumed to be evenly spaced, and information content was measured at a location halfway between two markers, where it is expected to be lowest. For clarity, a single pedigree structure is used throughout: first-cousin pairs with parents but not grandparents available for genotyping. Extensive simulations show that although the absolute numbers differ somewhat for other pedigree structures, all the main conclusions about the relative importance of marker polymorphism and density continue to hold.

How polymorphic do biallelic markers need to be?

Biallelic markers vary in their rates of polymorphism: the more common allele can range in frequency from 50% to nearly 100%. In considering a map of biallelic markers, it is important to ask whether only near-perfect (50–50) biallelics are useful or whether less polymorphic markers can provide comparable amounts of information. To answer this question, I measured information content in simulations of maps of biallelic markers with varying degrees of polymorphism.

Table 1 • Information content for biallelics

	allele distribution				
spacing (cM)	50-50	60-40	70-30	80-20	90-10
1	0.88	0.88	0.87	0.84	0.73
2	0.75	0.76	0.73	0.69	0.55
3	0.65	0.65	0.63	0.56	0.42
4	0.58	0.56	0.53	0.48	0.34
5	0.50	0.49	0.46	0.41	0.27
6	0.45	0.43	0.41	0.36	0.24
7	0.39	0.39	0.37	0.32	0.22
8	0.35	0.35	0.33	0.28	0.19
9	0.32	0.31	0.29	0.25	0.17
10	0.29	0.28	0.26	0.23	0.15

The results (Fig. 2, Table 1) clearly indicate that at higher map densities, allele frequency has only a small effect on information content in the range of frequency distributions from 50-50 to 80-20. Specifically, a 1-cM map of 60-40 biallelics provides an information content of 0.88, essentially the same as perfect 50-50 biallelics at this density, while 70-30 biallelics provide an information content of 0.87, and 80-20 biallelics provide an information content of 0.84. The information content drops to 0.73 for 90-10 biallelics. Thus, the use of biallelic markers with frequency distribution as skewed as 80-20 leads to little reduction in the information content of a dense map. For sparser maps of 5–10 cM, a similar conclusion holds for marker allele frequency distributions as skewed as 70-30.

How dense does a map of biallelic markers need to be?

Although there is a limit on how polymorphic a biallelic marker can be (a 50-50 distribution of the two alleles), there is essentially no theoretical limit on map density (or marker number), as reasonably polymorphic SNPs can be found roughly every 1 kb, or about 3 million times in the human genome (see above). Thus, one answer to how many markers are needed is that more is always better[1]. For common linkage study designs, however, the addition of markers provides diminishing returns once most of the inheritance information has been extracted. As shown above, a 1-cM

Table 2 • Information content for microsatellites

	allele number			
spacing (cM)	3	4	5	10
1	0.93	0.95	0.96	0.97
2	0.87	0.90	0.91	0.94
3	0.80	0.84	0.87	0.90
4	0.74	0.78	0.81	0.87
5	0.68	0.75	0.78	0.82
6	0.64	0.70	0.73	0.80
7	0.58	0.64	0.69	0.76
8	0.53	0.61	0.66	0.72
9	0.49	0.58	0.62	0.69
10	0.45	0.54	0.58	0.68

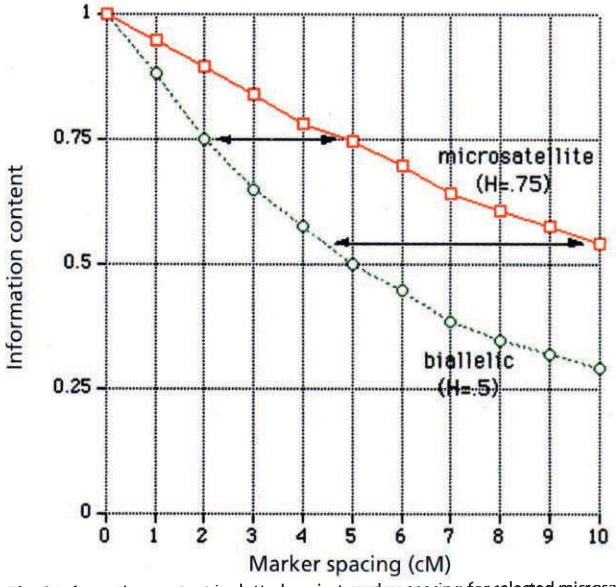

Fig. 3 Information content is plotted against marker spacing for selected microsatellite (heterozygosity H=0.75) and biallelic (heterozygosity H=0.5) markers. Arrows connect the points on the two curves where information content reaches 0.75 (top) and 0.54 (bottom), the values for 5-cM and 10-cM microsatellite maps, respectively.

map of 50–50 biallelic markers extracts 88% of the available information, and it is unlikely that higher information content is needed in an initial screen for linkage. What is the informational cost of decreasing the density of the map? Simulation results (Fig. 2) show that map density plays a more critical role than marker polymorphism. A 2-cM map provides information content of 0.75, a 3-cM map 0.65 and a 5-cM map 0.50. Together with the results of the previous section, these numbers lead to the conclusion that for initial linkage studies it is desirable to screen a dense (1–2-cM) map of moderately polymorphic (50–50 to 80–20) biallelic markers. Interesting regions can then be followed up with all available (biallelic and microsatellite) markers.

It is worth noting that there are two separate issues regarding map density: how many markers exist and how many markers can be genotyped rapidly and cost-effectively. Although current microsatellite maps cover the genome at an average spacing of less than 1 cM (with more than 5,000 markers in the final Généthon map alone[7]), genotyping more than a few hundred markers in a large collection of families remains beyond the power of today's technology and research budgets. Thus, the practical limit on the number of biallelic markers will depend on the techniques for marker development and genotyping. Nonetheless, it is interesting to compare such maps with current maps of microsatellite markers. Such a comparison is carried out in the next section.

Comparison of maps based on biallelics and microsatellites

Current genome scans typically employ a 10-cM map of microsatellite markers for the initial screen[19,20], followed by denser coverage of regions that yield interesting results. (Although one could employ a 'staged search' strategy of starting with a sparser 20–40-cM map and then increasing the density in all moderately positive regions[21,22], economies of scale in large genotyping labs usually argue for a one-stage initial scan: using a single optimized set of markers for all projects is more efficient than 'filling in' different regions for each.) Microsatellite markers typically vary between 0.65 and 0.8 in heterozygosity (for instance, an average of 0.7 in the final Généthon map[7]), and for simplicity I will use microsatellites with four equally frequent alleles (heterozygosity of 0.75) as representative in the following comparisons with biallelics with two equally frequent alleles (heterozygosity of 0.5); results for other values are given in Tables 1 and 2. Intuitively, one would expect two closely linked biallelics to provide the same information as one microsatellite, and simulations largely confirm this intuition. A 10-cM map of microsatellites achieves information content of 0.54 (Fig. 3). The same information content is provided by a 4.5-cM map of biallelic markers. A denser 5-cM microsatellite map achieves an information content of 0.75, as does a 2-cM map of biallelics. In general, maps of biallelic markers at about 2.25–2.5 times the density of microsatellites provide a comparable information content. A 10-cM map of 300 microsatellite markers can therefore be replaced by a 4-cM map of 750 biallelic markers. These conclusions are in rough agreement with the results of an earlier study of the trade-off between marker spacing and polymorphism[23].

As technology improves, it is likely that screening a much denser map of biallelic markers will be cheaper and easier than carrying out today's genome scans employing microsatellites[15,16]. There are reasons to employ such denser maps. As shown above, current scan densities lead to considerable loss of information. This problem is more serious for data-sets consisting of more distantly related affecteds or of progeny of consanguineous marriages used in homozygosity mapping[24]. It is therefore worth noting that a 1-cM map of biallelics (about 3,000 markers) yields much higher information content than a 10-cM map of microsatellites (0.88 vs. 0.54), and is superior to a 5-cM microsatellite map (0.88 vs. 0.75).

Practical linkage analysis using biallelic markers

Because of the lower polymorphism rates of biallelic markers, it is critical to consider many linked markers simultaneously; indeed, all the above results assume complete multipoint analysis of all markers on a chromosome. Such multipoint analysis is even more important for biallelics than for microsatellites. Fortunately, recently developed algorithms and software allow multipoint analysis with an essentially unlimited number of linked markers to be carried out for sib pairs[17] as well as for general pedigrees of moderate size[18]. These methods can also be used for automatic haplotype reconstruction, avoiding the tedious prospect of haplotyping many biallelics by hand. The one remaining challenge is extending multipoint analysis with many markers to large multi-generational families, although even here the situation is improving[25].

Discussion

The results presented here clearly demonstrate that the use of a genetic map of biallelic markers for linkage studies is feasible on theoretical grounds. It is not necessary to find only 'perfect' 50-50 biallelics: markers with allele frequency distributions as skewed as 70-30 or even 80-20 are almost as useful in a dense map. This result should allay the concern that markers discovered in one population may not be sufficiently informative in other populations with different allele frequencies. A 1–2-cM map of moderately polymorphic biallelic markers is superior to today's microsatellite screening sets for extracting inheritance information and should provide a more efficient tool for initial genome scans.

Even denser maps should enable novel study designs for dissecting genetically complex phenotypes. In particular, genome scans for linkage disequilibrium (LD) and association may become practical[26–28]. Because LD mapping relies on detecting recombinationally conserved regions around an ancestral mutation, the required map density will vary with the age and history of the study population, with very dense maps (spacing of 10 kb or less) likely to be needed for LD scans in a mixed general population. A more promising approach may be to screen in parallel functional (coding) biallelic polymorphisms in many genes for direct association (rather than LD) with disease[26–28].

Maps of biallelic markers and the technology to genotype them should be forthcoming[15,16], and the resulting progress in human genetics will be exciting to watch.

Methods

Simulations. Segregation of chromosomes of 100-cM length with evenly spaced markers was simulated. For biallelics, the frequencies of the common allele were 0.5, 0.6, 0.7, 0.8, 0.9, 0.95 and 0.99. For microsatellites, equally frequent alleles were assumed, with allele numbers of 3, 4, 5, 10, 20 and 100. Marker spacings of 1, 2, . . . , 10 cM were examined. Each simulation consisted of 100 replicates of 10 cousin pairs each. Information content was computed with GENEHUNTER[18]. Information content was measured halfway between the two markers closest to the middle of the chromosome. For ELOD computation, a dominant disease locus with full penetrance, no phenocopies and allele frequency of 0.001 was assumed to lie halfway between two markers, and chromosomes were simulated assuming that both cousins were affected. GENEHUNTER was used to compute multipoint lod scores. The relationship between information content and ELOD is preserved for other assumptions about the disease locus (data not shown). Simulation software used to generate the data is available from the author and can be used to explore additional map properties and pedigree structures.

Acknowledgements

I thank M. Daly, E. Lander and D. Wang for helpful discussions and comments on the manuscript. This work was supported in part by a Special Emphasis Research Career Award from NHGRI (HG00017).

1. Botstein, D., White, D.L., Skolnick, M. & Davis, R.W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**, 314–331 (1980).
2. Wyman, A.R. & White, R.W. A highly polymorphic locus in human DNA. *Proc. Natl. Acad. Sci. USA* **77**, 6754–6758 (1980).
3. Gusella, J.F., *et al.* A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**, 234–238 (1983).
4. Donis-Keller, H. *et al.* A genetic linkage map of the human genome. *Cell* **51**, 319–337 (1987).
5. Weber, J.L. & May, P.E. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.* **44**, 388–396 (1989).
6. Cooperative Human Linkage Center. A comprehensive human linkage map with centimorgan density. *Science* **265**, 2049–2054 (1994).
7. Dib, C. *et al.* A comprehensive genetic map of the human genome based on 5264 microsatellites. *Nature* **380**, 152–154 (1996).
8. Hofker, M.H. *et al.* The X chromosome shows less genetic variation at restriction sites than the autosomes. *Am. J. Hum. Genet.* **39**, 438–451 (1986).
9. Cooper, D.N., Smith, B.A., Cooke, H.J., Niemann, S. & Schmidtke, J. An estimate of unique DNA sequence heterozygosity in the human genome. *Hum. Genet.* **69**, 201–205 (1985).
10. Nickerson, D.A. *et al.* Automated DNA diagnostics using an ELISA-based oligonucleotide ligation assay. *Proc. Natl. Acad. Sci. USA* **87**, 8923–8927 (1990).
11. Livak, K.J., Marmaro, J. & Todd, J.A. Towards fully automated genome-wide polymorphism screening. *Nature Genet.* **9**, 341–342 (1995).
12. Saiki, R.K., Walsh, P.S., Levenson, C.H. & Erlich, H.A. Genetic analysis of amplified DNA with immobilized sequence-specific oligonucleotide probes. *Proc. Natl. Acad. Sci. USA* **86**, 6230–6234 (1989).
13. Syvanen, A.-C., Aalto-Setala, K., Harju, L., Kontula, K. & Soderlund, H. A primer-guided nucleotide incorporation assay in the genotyping of apolipoprotein E. *Genomics* **8**, 684–692 (1990).
14. Wu, D.Y., Ugozzoli, L., Pal, B.K. & Wallace, R.B. Allele-specific enzymatic amplification of β-globin genomic DNA for diagnosis of sickle cell anemia. *Proc. Natl. Acad. Sci. USA* **86**, 2757–2760 (1989).
15. Wang, D. *et al.* Toward a third generation genetic map of the human genome based on biallelic polymorphisms. *Am. J. Hum. Genet.* **59**, A3 (1996).
16. Chee, M. *et al.* Accessing genetic information with high-density DNA arrays. *Science* **274**, 610–614 (1996).
17. Kruglyak, L. & Lander, E.S. Complete multipoint sib-pair analysis of qualitative and quantitative traits. *Am. J. Hum. Genet.* **57**, 439-454 (1995).
18. Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. & Lander, E.S. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.* **58**, 1347–1363 (1996).
19. Reed, P.W. *et al.* Chromosome-specific microsatellite sets for fluorescence-based, semi-automated genome mapping. *Nature Genet.* **7**, 390–395 (1994).
20. Dubovsky, J., Sheffield, V.C., Duyk, G.M. & Weber, J.L. Sets of short tandem repeat polymorphisms for efficient linkage screening of the human genome. *Hum. Mol. Genet.* **4**, 449–452 (1995).
21. Elston, R.C. Designs for the global search of the human genome by linkage analysis. in *Proceedings of the 16th International Biometrics Conference* 39–51 (Hamilton, New Zealand, 1992).
22. Brown, D.L., Gorin, M.B. & Weeks, D.E. Efficient strategies for genomic searching using the affected-pedigree-member method of linkage analysis. *Am. J. Hum. Genet.* **54**, 544–552 (1994).
23. Terwilliger, J.D., Ding, Y. & Ott, J. On the relative importance of marker heterozygosity and intermarker distance in gene mapping. *Genomics* **13**, 951–956 (1992).
24. Lander, E.S. & Botstein, D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* **236**, 1567–1570 (1987).
25. O'Connell, J.R. & Weeks, D.E. The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. *Nature Genet.* **11**, 402–408 (1995).
26. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
27. Lander, E.S. The new genomics: global views of biology. *Science* **274**, 536–539 (1996).
28. Collins, F.S. Positional cloning moves from perditional to traditional. *Nature Genet.* **9**, 347–350 (1995).