European Journal of Human Genetics (2004) 12, 469–477

www.nature.com/ejhg

ARTICLE

Haplotype diversity and SNP frequency dependence in the description of genetic variation

Michael PH Stumpf*,[1]

[1]*Department of Biological Sciences, Biochemistry Building, Imperial College London, London SW7 2AY, UK*

Haplotype diversity is controlled by a variety of processes, including mutation, recombination, marker ascertainment and demography. Understanding the extent to which genetic variation at physically linked loci is co-inherited is crucial for the design of the HapMap project and the correct interpretation of the resulting data. In the absence of an analytical theory extensive coalescent simulations are used to disentangle the influence of all of these factors on haplotype diversity. In addition to these qualitative insights, this study also demonstrates (i) that marker spacing and frequency profoundly influence observed levels of haplotype diversity; (ii) that the spectrum of haplotypes contains information about how exhaustively genetic variation in a region is described by a given marker set; and (iii) that so-called haplotype blocks can be generated due by the stochasticity inherent in the recombination process without having to assume variation in the recombination rate.
European Journal of Human Genetics (2004) **12**, 469–477. doi:10.1038/sj.ejhg.5201179
Published online 17 March 2004

Keywords: population genetics; HapMap project; haplotype tagging; haplotype blocks

Introduction

An increasing number of empirical[1–3] studies investigate whether the inheritance of genetic variants occurs in a block-like manner and the potential implications of this for association studies.[4] Reported haplotype diversities along extended stretches of DNA appear surprisingly simple with most chromosomes belonging to one of roughly a handful of different haplotypes.[5] The levels of linkage disequilibrium (LD) are also reported to be consistently high between markers that are in the same block although LD can also extend beyond block-boundaries.[6,7] If such a picture were to prevail it would have obvious consequences for the design of association studies.[8]

In an important paper Jeffreys *et al*[1] showed that at least sometimes blocks may be delimited by recombination hotspots; the recombination rate in fairly localized regions can exceed the background or block recombination rate by up to four orders of magnitude and LD does not extend beyond the block boundaries. In many cases, however, there is as yet no conclusive evidence for block boundaries to coincide with recombination hotspots.[9,10] If this were generally the case then we could hope that block-boundaries and possibly knowledge of haplotypes in one population would allow us to make predictions of inferences for other populations. Unfortunately, however, many reports of blocks fail to show evidence for such a connection with hotspots[11] and the methods by which blocks are ascertained[3,8] may at least be partly to blame for this.

There are three main objectives of this study of haplotype diversity. On a fundamental level we gain insight (in the absence of an analytical theory) into how physical proximity between markers, the marker frequencies and the intensity of recombination interact to determine the complexity of the haplotype spectrum. Second, recent theoretical work by Wiuf *et al*[12] is followed up. These authors have shown that the number T of haplotype tagging SNPs (htSNPs)[2] necessary to describe a given set of M haplotypes defined by N SNPs is bounded by $\log_2 M$

*Correspondence: Dr MPH Stumpf, Department of Biological Sciences, Biochemistry Building, Imperial College London, London SW7 2AZ, UK.
Tel: +44 20 7594 5114; Fax: +44 20 7594 5789;
E-mail: m.stumpf@imperial.ac.uk
Revised 12 December 2003; accepted 21 January 2004

$< T <$ min(N,M-1). Here we determine, for different scenarios, the relationship between M and N. Third, a simple block definition is used to evaluate properties of blocks and how inferred blocks (which do not correspond to recombination hotspots) depend on a marker characteristics and recombination rate.

All three aspects of this work have implications in the current run-up to the HapMap project.[8] The study provides guidance into when and how the resulting SNP data are best summarized in terms of haplotypes. Moreover, as will become clear, haplotype diversity and the combinatorial structure of haplotypes also hold information about how exhaustively genetic variation in a region has been sampled.

Methods

Simulation procedures

In the following discussion we simulate the ancestral recombination graph[13] assuming uniform recombination and mutations rates ρ and μ, respectively. Assuming constant ρ allows the study of the behaviour in blocks of low recombination rates, or the expected behaviour of haplotype diversity under a Null model; both aspects will be considered here. Throughout we assume a single panmictic population with an effective population size of $N_e = 10\,000$ diploid individuals. Throughout we use a sample size of $n = 500$ chromosomes and consider a stretch of 50 kb length; the sample size is large compared to most studies of LD performed today,[3,9] but smaller than the population samples predicted for future case–control studies.[14] The mutation rates are assumed to be 10^{-8}/(nucleotide $\times$ generation) and 10^{-9}/(nucleotide $\times$ generation), whence the population mutation rates along the whole stretch are $\mu = 50$ and 5; the mutation model used here is the infinite sites model. We consider two recombination rates which correspond to 1 and 0.1 cM/Mb in addition to the case of no recombination; the corresponding population recombination rates are thus: 50, 5 and 0, respectively.

Human genetic diversity for a stretch of 50 kb corresponds approximately to $\mu = 50$ and $\rho = 50$.[15] The other values therefore correspond loosely to cases where the recombination and/or average marker density (via the mutation rate) is decreased. The case of $\mu = 5$ and $\rho = 5$, however, can also be interpreted as the correct description of a 5 kb stretch. We also use the $\mu = 5$ case, which gives rise to a sparser marker set, as a qualitative example for SNP ascertainment.[16] Results for the lower mutation rate to model as representing the case of a sparser set of markers.

In addition to the constant population size we also investigate the effects of population growth on the resulting haplotype diversity but refrain from a more detailed study of the effects of demography. For each scenario, 2000 independent runs of the ancestral recombination graph were performed. Frequency cutoffs for the minor marker allele (and not always the derived allele) are enforced by counting the copies of each allele in the sample. Cutoff frequencies considered are 1, 5, 10 and 20%.

Haplotype analysis and tagging approach

The minimum number of necessary tagging SNPs to tag a given set of haplotypes is evaluated using a brute-force implementation of the algorithm described in Wiuf *et al*[12]. Starting from the $k = M$, where M is the number of haplotypes, we evaluate each possible combination of k SNPs to see if it could be used as a basis for the set of haplotypes. If one of the *N!/(N-k)/k!* possible SNP combinations forms a valid basis then k is decreased by 1 until the first time a basis cannot found. For large N and M the number of SNP combinations can become enormous but in smaller simulations it was observed that the distribution of the minimum number of tags required to tag a given number of haplotypes is relatively flat: many different combinations can be used to tag haplotypes. Thus, for large values of N and M it is possible to proceed heuristically[12] and investigate, for exmple, a maximum of 100 Million combinations of candidate tags and an inferred minimal basis will be close to optimal. Similarly, we have also implemented a strategy where we start from $k = \min(N,M\text{-}1)$ and increment k until a basis has been found. Using either approach (which of course yield identical results) when a set of k SNPs is found the procedure stops and the number of necessary and sufficient tagging SNPs is set to $T = k$. At most min(N,M-1) tagging SNPs are required to describe all observed haplotypes in the sample. As the algorithm is in the NP-complete class we only evaluate the number of tagging SNPs for the case of SNP ascertainment outlined above. Our heuristic approach can also be implemented more formally in a Markov Chain Monte Carlo setting.

Results

Here we discuss how the number of haplotypes depends on the number of SNPs, the recombination rate and the cutoff frequency for the fraction of chromosomes that should be included. Analytic results are only available for the case of no recombination and free recombination, respectively, and we therefore use coalescent simulations as outlined above.

Determinants of haplotype diversity

In Figure 1 we show how the number of haplotypes depends on the number of SNPs, their frequency and the recombination rate. The relationship between SNP number (for each frequency cutoff) and the total number of haplotypes in a sample already carries information about the recombination rate and how exhaustively a given set of SNPs represents or resembles underlying genetic variation. Large SNP sets (with a low cutoff frequency) will contain correlations among SNPs but if marker sets are sparse, recombination

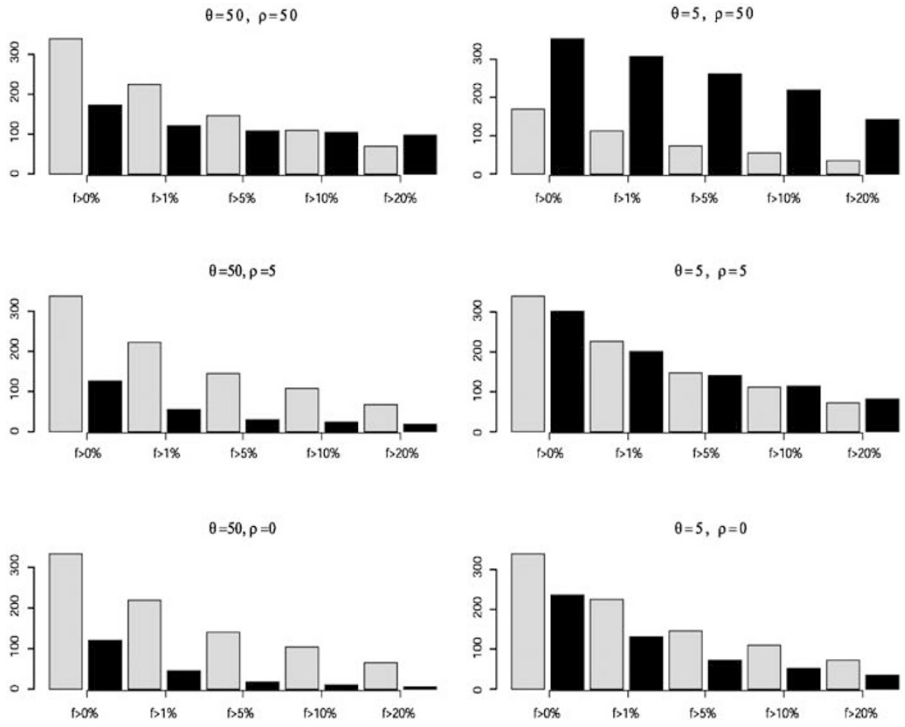

Figure 1 Average SNP (grey) and haplotype numbers (black) *versus* minor allele frequency cutoff for $\theta = 5$ and 50 and $\rho = 50$, 5 and 0, respectively.

will be more effective at breaking up associations between markers; we therefore expect a lower value for the ratio for $\theta = 5$ than for $\theta = 50$, irrespective of cutoff frequency and recombination rate. The number of tags required to adequately describe variation in a region will therefore be a function of both marker frequency and marker density.

For $\theta = 50$ we find that a 10-fold decrease in the recombination rate from $\rho = 50$ to 5 already brings the observed number of haplotypes very close to the $\rho = 0$ results. For lower values of ρ the average number of haplotypes is virtually indistinguishable from the $\rho = 0$ case. Note that for the decay of LD measured by the same decrease in ρ from $\rho = 50$ to 5 does not yield a behaviour anywhere near the $\rho = 0$ case (not shown). Haplotype diversity and LD, although related, show somewhat different dependence on the population recombination rate ρ. This is also observed for growing and bottleneck populations (data not shown).

The dependence of haplotype diversity on the minor SNP allele frequency is further exemplified in Figure 2. Here we show the number of haplotypes needed to describe 90, 95 and 99%, and all of the 500 chromosomes in the sample. These numbers are displayed for five different marker frequency cutoffs, three recombination and two mutation rates. Such a table can either be used to assess the genotyping cost necessary to capture a given amount of variation or in case all the available genetic variation has been characterized, to obtain an indication of the average recombination/mutation rate ratio. While for high marker density or mutation rates rare ($f < 1\%$) alleles give rise to a large number of haplotypes we find that for $f \geq 5\%$ there is no big reduction in genotyping effort as the cutoff frequency is further increased. Also for $f \geq 5\%$ the frequency distribution of haplotypes holds some information about the recombination rate: a higher recombination rate will lead to more rare haplotypes even at moderate to high frequency cutoffs, as is also intuitively obvious.

The frequency distribution of haplotypes is displayed in Figure 3. At the reported genome wide average of the recombination rate a stretch of 50 kb is not expected to have any haplotypes at a frequency greater than 10%, irrespective of the cutoff frequency. If the marker spacing is decreased, however, some haplotypes will gain in frequency and at $\theta = 5$ and $\rho = 50$ we therefore observe some haplotypes at moderate frequencies, especially at high cutoffs. Low values of θ result in a shift of weight to higher haplotype frequencies. For $\rho < 1$ (results not shown) the resulting haplotype distributions are very similar to the special $\rho = 0$ case apart from the origin. At low recombination rates and for cutoffs $f \geq 5$ the haplotype distribution obtains a mode at the cutoff frequency f.

The shift of the mode to the cutoff frequency is simply a result of the fact that in the absence of excessive

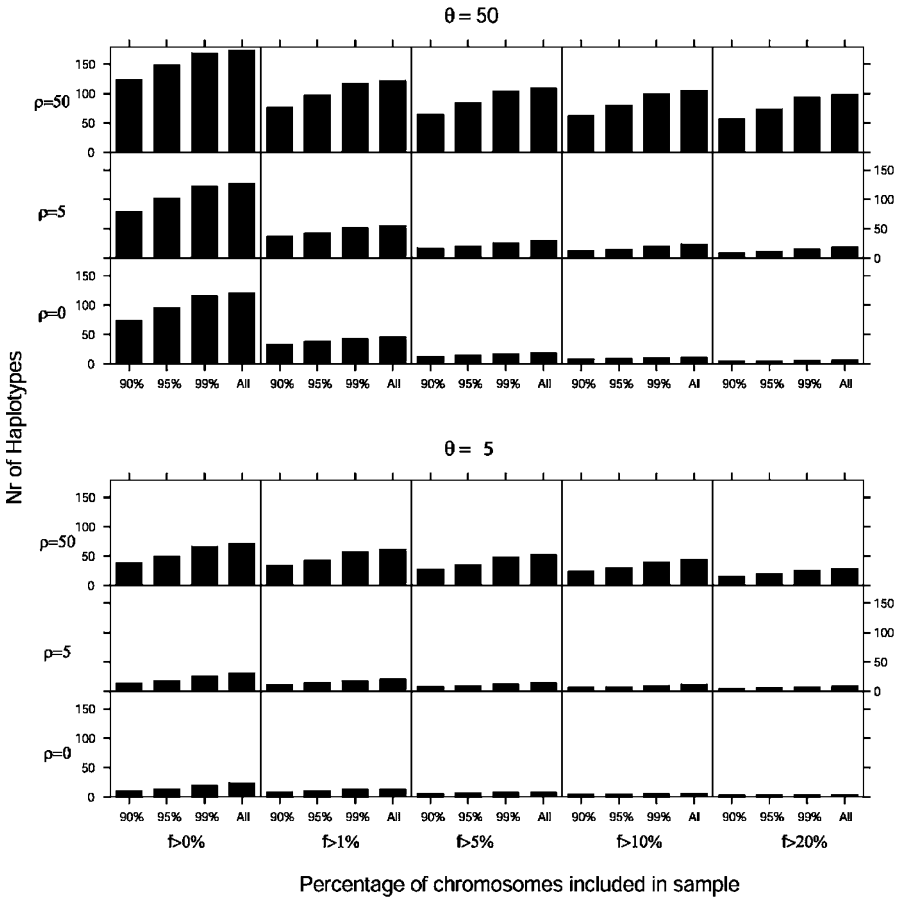

Figure 2 Average number of haplotypes needed to explain 90, 95, 99 and 100% of observed chromosomes in a sample for frequency cutoffs of 1, 5, 10 and 20%, respectively, for $\theta = 5$ and 50 and $\rho = 50$ and 5.

recombination, an SNP with a minor allele frequency of x will define a haplogroup of frequency x if x is very close to the cutoff frequency; thus an excess of haplotypes with frequency x will be observed. If the recombination rate is high then haplotypes defined by the youngest SNP can be broken up by recombination and here $\rho = 50$ appears to yield results that are very close to the case of free recombination. As a result the mode of the haplotype frequency distribution shifts back to the origin.

Haplotype tagging

Only one tagging strategy is investigated here[12] and at the moment it is by no means clear what tagging strategy is best suited for association studies.[8] Rather than focusing on tagging haplotypes, it may for example be better to define tags that capture the patterns of LD and/or association between SNPs. Simulation-based power analysis along the lines taken here will help to assess such questions in further detail. The tagging approach used here is quite likely not optimal for association studies, but its easy interpretation in terms of a geometric basis for the space spanned by the SNP defined haplotypes nicely highlights the combinatorial nature of haplotypes and the complexity introduced by recombination. Other haplotype tagging frameworks, however, are likely to behave qualitatively similarly to the approach taken here.

We only consider an allele frequency cutoff of 5%. In Table 1 we show mean values of the ratios T/N_5 (where N is the number of SNPs with a minor allele frequency of 5%) and the corresponding 5 and 95 percentiles. We find an obvious dependence on ρ and for $\rho = 5$ the results are already quite similar to $\rho = 0$. The results for $\rho = 50$ are discouraging: on average over 90% of SNPs need to be typed in order to reliably distinguish between haplotypes. This suggests that reports of low haplotype diversity indicate regions of low recombination rate. We note, however, that the majority of currently published studies has marker density that is at least a factor of 5 lower than the one obtained here.[9] Moreover our results concern true, not inferred haplotypes. Haplotype inference may systematically bias tagging approaches.

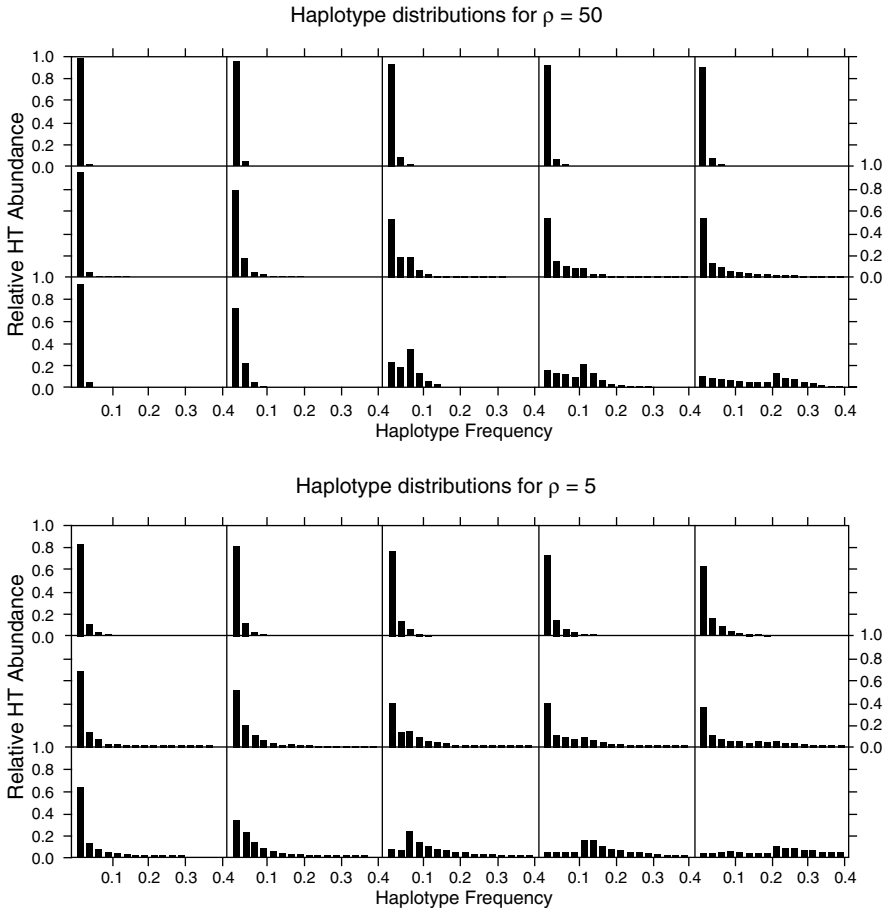

Figure 3 Frequency distribution of haplotypes and their dependence on θ, ρ and minor allele frequency cutoff.

Table 1 Average fraction of SNPs (in percent) needed to capture *x*% of chromosomes in the sample ($x = 99$, 95 and 90%, respectively) for three different values of the recombination rate together with their 5 and 95 percentiles (in parentheses)

Population recombination rate ρ	*Average number of haplotypes*	*Percentage of SNPs needed to explain fraction x of haplotypes*		
		$x = 99\%$	$X = 95\%$	$x = 90\%$
0	7.3	49.7 (22.0–92.5)	45.0 (17.9–83.3)	37.5 (14.3–66.7)
5	14.2	70.6 (41.2–100)	58.9 (32.0–90.0)	49.4 (25.0–78.6)
50	52.3	94.7 (83.3–100)	93.1 (80.0–100)	91.7 (76.2–100)

The average number of segregating sites is 33.9 of which 14.67 had a frequency $\geq 5\%$; the corresponding 5 and 95 percentiles are 21 and 50 and 5 and 28, respectively.

Dynamics of haplotype blocks

Notions and possible uses of extended haplotype blocks that are characterized by high levels of pairwise LD between SNPs within the same block (and accordingly low haplotype diversity compared to the extreme case of free recombination) have attracted considerable interest.[4,6,17–19] Here we follow Wang *et al*[17] and use probably the simplest definition of a block: all SNP pairs that are within the same block must fail the four-gamete test, that is, at most three out of the possible four two-locus haplotypes are observed for each pair of bi-allelic markers. This definition has some shortcomings but is (i) easily implemented, and (ii) we expect it to give at least some insight into how SNP frequencies and ascertainment affect the behaviour of blocks. Insights gained for this simple model will be transferable to other, more involved, block-ascertainment methods.

In Figure 4 we show how the average number and average size of blocks, as well as the proportion of DNA and SNPs that are found within blocks depend on minor allele frequency cutoff and recombination rate ρ. We only consider $\theta = 50$ but in each case we show both the results

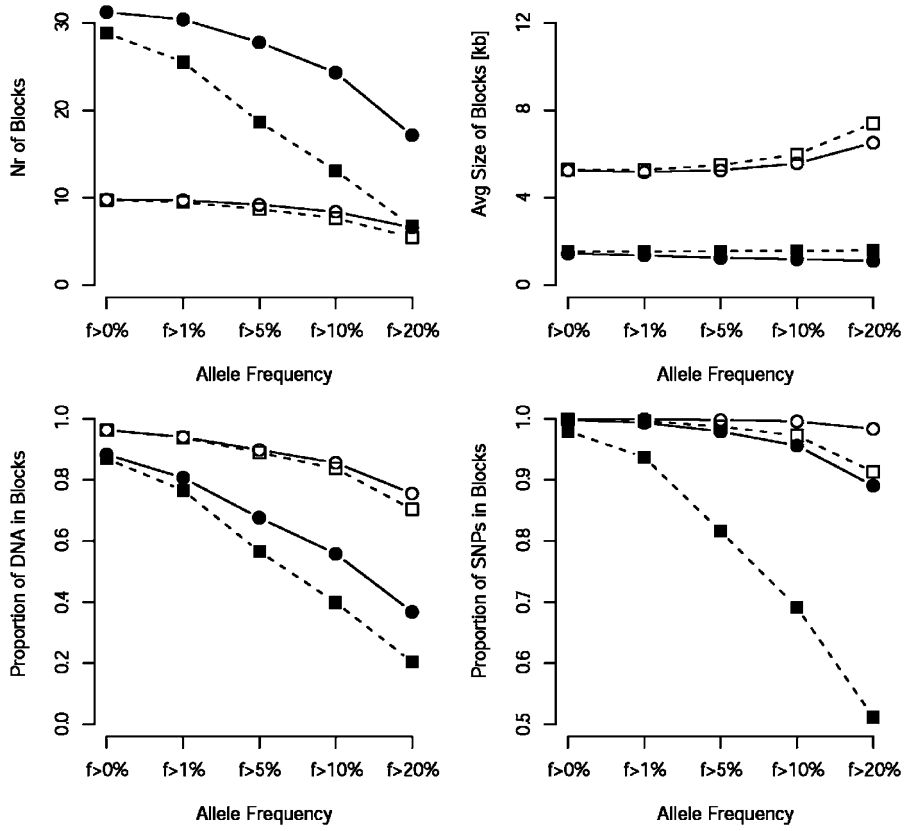

Figure 4 Average no. of blocks, average block-size, average of the total proportion of DNA in blocks and average of the total number of SNPs in blocks calculated for a sample of 500 chromosomes drawn from a constant size population with $\theta = 50$ *versus* frequency cutoff. Solid symbols represent the case $\rho = 50$, empty symbols $\rho = 5$. Circles (solid lines) represent results obtained for all blocks, boxes (dashed lines) represent results for blocks containing at least four SNPs.

for all blocks that adhere to our definition and of 'long' blocks. 'Long' blocks are blocks that contain at least four SNPs while other blocks may also contain pairs of SNPs that fulfil our four-gamete test criterion. Full symbols denote results for $\rho = 50$, empty symbols $\rho = 5$; circles (full lines) are for all blocks while boxes (dashed lines) refer only to the long blocks.

We observe that for low frequency cutoffs there are many more but shorter blocks for $\rho = 50$ than for 5 where the two curves are in very close agreement. At $\rho = 50$ the average block-size is determined largely by the long blocks but for all other measures displayed in Figure 4 we observe significant differences between long and short blocks. The number of long blocks, the proportion of DNA in long blocks, and perhaps most severely, the proportion of SNPs that are found in long blocks decreases more dramatically with minor allele frequency cutoff than the same measures do for all blocks. At a minor allele frequency of 20% only approximately 20% of DNA and 50% of SNPs are found in long blocks. For all blocks these values increase to 40 and 90%, respectively. It is obvious that small blocks, containing only two or three SNPs, will offer little or no reduction in genotyping effort. Long blocks, on the other hand, account for only a small part of the total sequence.

The average block-size remains approximately constant for all allele frequency cutoffs. This result can be explained by considering those pairs of SNPs that are the most likely to give rise to four observed two-locus haplotypes. These SNPs have to be old enough to have undergone at least one recombination event and therefore will have reasonably large minor allele frequencies. Pairs of younger markers, which by and large will have a smaller minor allele frequency, are less likely to give rise to four haplotypes and therefore we expect SNPs with moderate to high minor allele frequencies to determine block-size. Undersampling of diversity (eg restricting the analysis to already known SNPs such as those in dbSNP) could therefore systematically overestimates average block-lengths. This result is in agreement with the study of Phillips *et al*[3] who find that block-length increases with marker spacing; it is likely to hold for other definitions as suggested by recent studies of the effects of SNP ascertainment.[16] Thus, interpretation of haplotype diversity (like LD and block

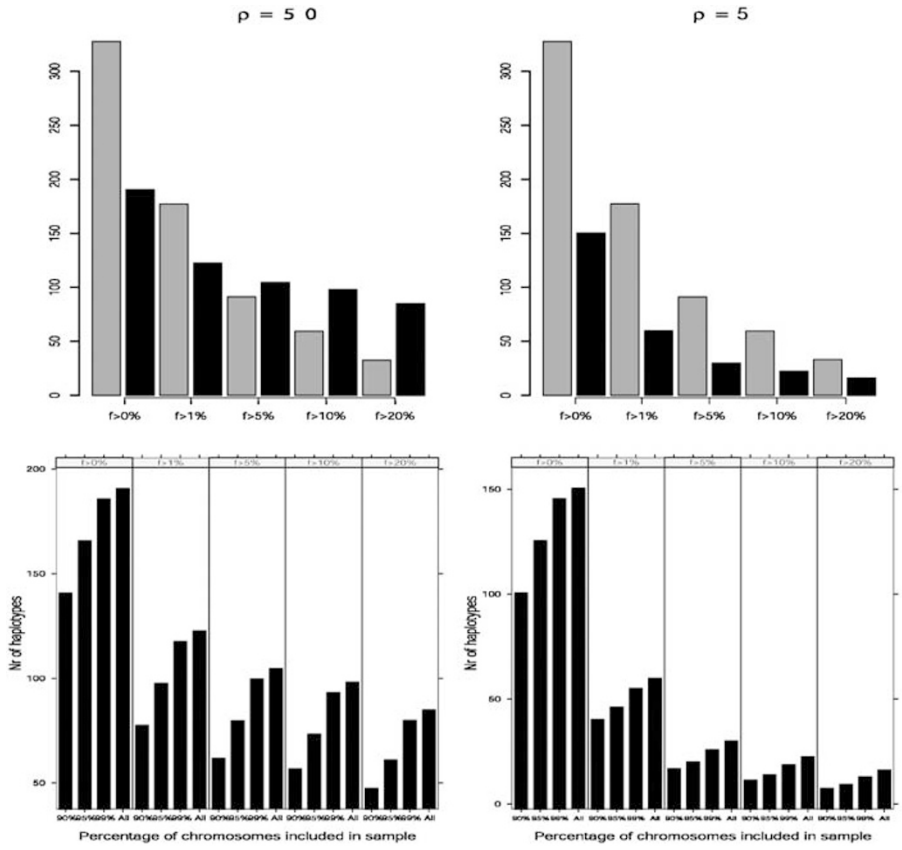

Figure 5 Top row: average numbers of SNPs (grey) and haplotypes (black) resulting for $\theta \approx 65$ and $\rho = 50$ and 5, respectively. Bottom row: number of haplotypes that need to be considered in order to cover 90, 95 and 99%, and all of the chromosomes in the sample. In each case the demographic model outlined in the text was used in the coalescent simulations.

boundaries) is problematic if not supported by extensive simulations.[3,7]

Demography and haplotype diversity

Demography and population structure are known to have profound effects on the frequency spectrum of segregating sites, LD and thus also on haplotype diversity.[3,4,9] Simulations of population-growth scenarios suggest that the effect of minor-allele frequency still persists. We only show results for one particular demographic scenario where the population has grown from 1% of its present size to its present size over a time $\tau = 1$ (in coalescent units); before the onset of growth the population size is assumed to be constant at 1% of the present size. Other cases are easily assessed using coalescent simulations. Owing to the problems associated with diversity discussed by Pritchard and Przeworski[15] the mutation rate was adjusted such that the number of segregating sites in the sample is the same in the population growth scenario as in the constant population scenario discussed above.

Comparing Figure 1 with the top row of Figure 5 shows only quantitative differences that are easily explained by the different SNP allele frequency distribution resulting from a population growth scenario. We find at the higher recombination rate that haplotype numbers exceed SNP numbers already for lower frequency cutoffs (ie $f > 5\%$ instead of $f > 20\%$). At the same cutoff frequency the ratio of [haplotype number]/[SNP number] is less for the growth demography considered here than for the constant size population. Comparison of Figure 2 with the bottom row of Figure 5 shows only a minor vertical shift: the average number of haplotypes needed to describe $x\%$ ($x = 90$, 95, 99, 100) of the chromosomes in the sample is higher for population growth than for constant population size. Again this is easily understood because population growth results in a relative excess of rare alleles compared to the case of constant population size. These results suggest that the basic patterns of haplotype dependence (on allele frequency cutoff, marker spacing and recombination rate) elucidated above may remain valid for a range of demographic scenarios.

Conclusions

In the search for the genetic components of complex diseases or drug response phenotypes haplotype-based

approaches have recently been heralded as particularly promising. A host of early studies suggested that relatively few (eg 2–6) haplotypes may suffice to describe the genetic variation along extended stretches of DNA.[3,5,9,10] The aim of this study was to (i) gain some understanding of the factors influencing observed haplotype diversities, (ii) evaluate the behaviour of haplotypes expected for simple population genetic models, and (iii) see to what extent haplotype blocks can appear without underlying local variation in the recombination rate.

Before discussing the application of the results presented here to real world data, it is important to acknowledge the limitations of the approach taken here. The population model is of course incorrect and at best over-simplified. While a quantitative interpretation of the results is thus impossible they seem to reflect qualitative trends. For example, for many if not all population models (including the unknown true model), haplotype diversity will increase with increased recombination rate and decrease dramatically with increased SNP frequency cutoff. This is a general result confirmed by simulations of a wide range of demographic models (data not shown) and intuitively obvious in the light of what is known about the ancestral recombination graph.

The reported haplotype frequencies and diversities are not easily reconciled with the standard neutral constant size model of evolution although the generally small sample sizes will result in overestimation of LD and of haplotype frequencies. For the sample size considered here, $n = 500$, which is by no means large compared to what will be required for genetic association studies,[14] the number of segregating sites is very large for a region of 50 kb, $S \approx 330$. Even a moderate reduction of the recombination rate brings haplotype diversities and the number of required tSNPs into the range observed for $\rho = 0$. This suggests that at least some of the reported blocks may occur in regions where the recombination rate ρ is less than the reported genome wide average $\rho = 1$ cM/Mb. The simulations also show that haplotype diversity and block behaviour depend on both allele frequency and marker spacing. A number of reports of long-range disequilibrium and/or low haplotype diversity, based on incomplete sampling of the genetic SNP diversity, need to be reassessed in the light of this. A detailed assessment of local recombination rate variation becomes important and should provide crucial information about the usefulness of blocks. Similarly, predictions about the success/efficiency gains to be gained from the HapMap project that are based on present studies may systematically underestimate the number of tagging SNPs required to describe human genetic diversity.

Generally, we find that for complete ascertainment of segregating sites/SNPs haplotype diversity along a 50-kb stretch is almost unmanageably large if all markers or those with a minor allele frequency of $f \leq 1\%$ are to be typed. From a cutoff of '5%' and above no big efficiency gains are obtained and if the common variant/common disease should turn out to be correct than 5% may be a reasonable cutoff frequency. The genotyping effort, even if tagging approaches are used, may be considerably more than had been hoped.[2,9,10]

There are considerable problems in interpreting current experimental data sets and the simulation study presented here gives some clues as to what factors may compromise inferences drawn from summaries of the data such as LD and/or haplotype diversity. Many of these problems could be directly addressed if the underlying recombination rate variation were known. In addition to approaches using sperm-typing,[1,20] a number of inferential procedures has recently developed that allow direct estimation of the recombination rate.[21–25] These use mainly information from informative sites with high minor allele frequency and their inferences should be robust against the problems associated with low marker density and bias in allele frequencies. Knowledge of local recombination rate variation along the human genome will provide crucial guidance in the setup of genetic epidemiology studies.

Acknowledgements

I thank Carsten Wiuf and Gil McVean for many discussions on this topic and Monty Slatkin for his helpful comments on a earlier version of this manuscript. This work was funded through a Wellcome Trust Career Development Fellowship and a Royal Society Project Grant.

References

1 Jeffreys AJ, Kauppi L, Neumann R: Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet* 2001; **29**: 217–222.
2 Johnson GC, Esposito L, Barrat BJ *et al*: Haplotype tagging for the identification of common disease genes. *Nat Genet* 2001; **29**: 233–237.
3 Phillips MS, Lawrence R, Schidanandam R *et al*: Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nat Genet* 2003; **33**: 382–387.
4 Stumpf MP, Goldstein DB: Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. *Curr Biol: Cb* 2003; **13**: 1–8.
5 Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES: High-resolution haplotype structure in the human genome. *Nat Genet* 2001; **29**: 229–232.
6 Wall JD, Pritchard JK: Assessing the performance of haplotype block models of linkage disequilibrium. *Am J Hum Genet* 2003; **73**: 2003.
7 Wall JD, Pritchard JK: Haplotype blocks and linkage disequilibrium in the human genome. *Nat Rev Genet* 2003; **4**: 587–597.
8 Cardon LR, Abecasis GR: Using haplotype blocks to map human complex trait locl. *Trends Genet* 2003; **19**: 135–140.
9 Gabriel SB, Schaffner S, Nguyen H *et al*: The structure of haplotype blocks in the human genome. *Science* 2002; 1069424.
10 Patil N, Berno AJ, Hinds DA *et al*: Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 2001; **294**: 1719–1723.
11 Anderson EC, Slatkin M: Population-genetic basis of haplotype blocks in the 5q31 region. *Am J Hum Genet* 2004; **74**: 40–49.
12 Wiuf C, Laidlaw Z, Stumpf MPH: Some notes on the combinatorial properties of haplotype tagging. *Math Biosci* 2003; **185**: 205–216.

13 Griffiths RC, Marjoram P: Ancestral inference from samples of DNA sequences with recombination. *J Comput Biol* 1996; **3**: 479–502.
14 Weiss KM, Clark AG: Linkage disequilibrium and the mapping of complex human traits. *Trends Genet* 2002; **18**: 19–24.
15 Pritchard JK, Przeworski M: Linkage disequilibrium in humans: models and data. *Am J Hum Genet* 2001; **69**: 1–14.
16 Akey JM, Zhang K, Xiong MM, Jin L: The effect of single nucleotide polymorphism identification strategies on estimates of linkage disequilibrium. *Mol Biol Evol* 2003; **20**: 232–242.
17 Wang N, Akey JM, Zhang K, Chakraborty R, Jin L: Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am J Hum Genet* 2002; **71**: 1227–1234.
18 Zhang K, Deng M, Chen T, Waterman MS, Sun F: A dynamic programming algorithm for haplotype block partitioning. *PNAS* 2002; **99**: 7335–7339.
19 Anderson EC, Novembre J: Finding haplotype block boundaries by using the minimum-description length principle. *Am J Hum Genet* 2003; **73**: 336–354.
20 Arnheim N, Calabrese P, Nordborg M: Hot and cold spots of recombination in the human genome: the reason we should find them and how this can be achieved. *Am J Hum Genet* 2003; **73**: 5–16.
21 Fearnhead P, Donnelly P: Estimating recombination rates from population genetic data. *Genetics* 2001; **159**: 1299–1318.
22 McVean G, Awadalla P, Fearnhead P: A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* 2002; **160**: 1231–1241.
23 Hudson RR: Two-locus sampling distributions and their application. *Genetics* 2001; **159**: 1805–1817.
24 Li N, Stephens M: Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 2003; **165**: 2213–2293.
25 Stumpf MPH, McVean GAT: Estimating recombination rates from population-genetic data. *Nat Rev Genet* 2003; **4**: 959–968.