12. H. Yoshikawa, N. Sueoka, *Proc. Natl. Acad. Sci. U.S.A.* **49**, 559 (1963).
13. K. P. Lemon, A. D. Grossman, *Science* **282**, 1516 (1998).
14. V. Bidnenko, S. D. Ehrlich, L. Jannière, *Mol. Microbiol.* **28**, 1005 (1998).
15. C. Bruand, S. D. Ehrlich, L. Jannière, *EMBO J.* **14**, 2642 (1995)
16. C. Bruand, M. Farache, S. McGovern, S. D. Ehrlich, P. Polard, *Mol. Microbiol.* **42**, 245 (2001).
17. C. Bruand, P. Polard, unpublished data.
18. C. Bruand, S. D. Ehrlich, L. Jannière, *EMBO J.* **10**, 2171 (1991).
19. L. Boe, M. F. Gros, H. te Riele, S. D. Ehrlich, A. Gruss, *J. Bacteriol.* **171**, 3366 (1989).
20. A. E. Pritchard, H. G. Dallmann, B. P. Glover, C. S. McHenry, *EMBO J.* **19**, 6536 (2000).
21. S. Moriya, Y. Imai, A. K. Hassan, N. Ogasawara, *Plasmid* **41**, 17 (1999).
22. M. E. Sharpe, P. M. Hauser, R. G. Sharpe, J. Errington, *J. Bacteriol.* **180**, 547 (1998).
23. H. Araki, P. A. Ropp, A. L. Johnson, L. H. Johnston, A. Morrison, A. Sugino, *EMBO J.* **11**, 733 (1992).
24. R. Karthikeyan, E. J. Vonarx, A. F. Straffon, M. Simon, G. Faye, B. A. Kunz, *J. Mol. Biol.* **299**, 405 (2000).
25. R. Dua, D. L. Levy, J. Campbell, *J. Biol. Chem.* **274**, 22283 (1999).
26. B. P. Glover, C. S. McHenry, *Cell* **105**, 925 (2001).
27. R. E. Yasbin, D. Cheo, D. Bol, in B. subtilis *and Other Gram-Positive Bacteria*, A. Sonenshein, J. A. Hoch, R. Losick, Eds., (American Society for Microbiology, Washington DC, 1993), pp. 529–537.
28. M. F. Goodman, *Trends Biol. Sci.* **25**, 189 (2000).
29. P. J. Lewis, S. Thaker, J. Errington, *EMBO J.* **19**, 710 (2000).
30. E. LeChatelier, S. D. Ehrlich, L. Jannière, *Mol. Microbiol.* **20**, 1099 (1996).

31. We thank N. Givernaud for the construction of the IPTG-controlled *dnaE* mutant and S. Séror for the choice of probes to measure the *ori:ter* ratio. Supported in part by grant BIO4 CT95-0278 from the European Commission (S.D.E.) and by the UK Biotechnology and Biological Sciences Research Council (J.E.).

18 September 2001; accepted 11 October 2001

Blocks of Limited Haplotype Diversity Revealed by High-Resolution Scanning of Human Chromosome 21

Nila Patil, Anthony J. Berno, David A. Hinds, Wade A. Barrett, Jigna M. Doshi, Coleen R. Hacker, Curtis R. Kautzer, Danny H. Lee, Claire Marjoribanks, David P. McDonough, Bich T. N. Nguyen, Michael C. Norris, John B. Sheehan, Naiping Shen, David Stern, Renee P. Stokowski, Daryl J. Thomas, Mark O. Trulson, Kanan R. Vyas, Kelly A. Frazer, Stephen P. A. Fodor, David R. Cox*

Global patterns of human DNA sequence variation (haplotypes) defined by common single nucleotide polymorphisms (SNPs) have important implications for identifying disease associations and human traits. We have used high-density oligonucleotide arrays, in combination with somatic cell genetics, to identify a large fraction of all common human chromosome 21 SNPs and to directly observe the haplotype structure defined by these SNPs. This structure reveals blocks of limited haplotype diversity in which more than 80% of a global human sample can typically be characterized by only three common haplotypes.

Human DNA sequence variation accounts for a large fraction of observed differences between individuals, including susceptibility to disease. The majority of human sequence variation is due to substitutions that occurred once in the history of mankind at individual base pairs, called single nucleotide polymorphisms (SNPs) (*1–3*). Although most of these biallelic SNPs are rare, it has been estimated that 5.3 million common SNPs, each with a frequency of 10 to 50%, account for the bulk of the DNA sequence difference between humans. Such SNPs are present in the human genome once every 600 base pairs (*4*). Alleles making up blocks of such SNPs in close physical proximity are often correlated, resulting in reduced genetic variability and defining a limited number of "SNP haplotypes," each of which reflects descent from a single, ancient ancestral chromosome (*5*). Although a block of N independent biallelic SNPs could in theory generate 2^N different haplotypes, in the absence of recurrent mutation and/or recombination the number of observed haplotypes should be no greater than (N + 1). The complexity of local haplotype structure in the human genome and the distance over which individual haplotypes extend is poorly defined. Empirical studies investigating different segments of the human genome in different populations have revealed tremendous variability in local haplotype structure. These studies indicate that the relative contributions of mutation, recombination, selection, population history, and stochastic events to haplotype structure vary in an unpredictable manner, resulting in some haplotypes that extend for only a few kilobases (kb) and others that extend for greater than 100 kb (*6–8*). These findings suggest that any comprehensive description of the haplotype structure of the human genome, defined by common SNPs, will require empirical analysis of a dense set of SNPs in many independent copies of the human genome. As a first step toward achieving this goal, we have used high-density oligonucleotide arrays, in combination with somatic cell genetics, to identify a large fraction of all common human chromosome 21 SNPs and to directly observe the haplotype structure they define.

SNPs were discovered by using a publicly available panel of 24 ethnically diverse individuals (*9*). We physically separated the two copies of chromosome 21 from each individual using a rodent-human somatic cell hybrid technique (*10*). Twenty independent copies of chromosome 21, representing African, Asian, and Caucasian chromosomes, were analyzed for SNP discovery and haplotype structure. Finished human chromosome 21 genomic DNA sequence consisting of 32,397,439 bases was masked for repetitive sequences and the resulting 21,676,868 bases (67%) of unique sequence were assayed for variation with high-density oligonucleotide arrays (*11*). In total, we synthesized 3.4×10^9 oligonucleotides on 160 wafers to scan 20 independent copies of human chromosome 21 for DNA sequence variation. Each unique chromosome 21 was amplified from a rodent-human hybrid cell line by using long range–polymerase chain reaction (LR-PCR) (*12*). We designed unique oligonucleotides to generate 3253 minimally overlapping LR-PCR products of 10-kb average length spanning 32.4 Mb of contiguous chromosome 21 DNA. LR-PCR products corresponding to the bases present on a single wafer were pooled and hybridized to the wafer as a single reaction (*13*). SNPs were detected as altered hybridization by using a pattern recognition algorithm (*14*). In total, we identified 35,989 SNPs in our sample of 20 chromosomes. The position and sequence of these human polymorphisms have been deposited in the National Center for Biotechnology Information (NCBI) dbSNP database, accession numbers ss#3995623 through ss#4020948. We used dideoxy sequencing to assess a random sample of 227 of these SNPs in the original DNA samples and confirmed 220 (97%) of the SNPs assayed. In order to achieve this low rate of 3% false-positive SNPs, we

Perlegen Sciences, Inc., 2021 Stierlin Court, Mountain View, CA 94043, USA.

*To whom correspondence should be addressed. E-mail: david_cox@perlegen.com

required stringent thresholds for SNP detection on wafers that resulted in a high false-negative rate. Approximately 65% of all bases present on the wafers yielded data of high enough quality for use in SNP detection. (*15*).

The allele frequency distribution of the SNPs is presented in Fig. 1A. Genetic variation, normalized for the number of chromosomes in our sample, was estimated with two measures of nucleotide diversity: π, the average heterozygosity per site, and θ, the population mutation parameter (*16*). The estimates of average nucleotide diversity for the total data set ($\pi = 0.000723$ and $\theta = 0.000798$) as well as the distribution of nucleotide diversity, measured in contiguous 200,000 base pair bins of chromosome 21 (Fig. 1B), are within the range of values previously described (*2*, *3*, *17*). The estimate of θ was observed to be greater than the estimate of π for 129 of the 162 200-kb bins of contiguous DNA sequence analyzed. This difference is consistent with a recent expansion of the human population and is similar to the finding of a recent study of nucleotide diversity in human genes (*18*). We found that 11,603 of the SNPs (32%) had a minor allele observed a single time in our sample (singletons), as compared to the neutral model expectation of 43% singletons given the observed amount of nucleotide diversity (*19*). The difference between the observed and expected values is largely attributable to our reduced power to identify rare as compared to common SNPs (*15*).

We identified 47% of the 53,000 common SNPs with an allele frequency of 10% or greater estimated to be present in 32.4 Mb of the human genome (*4*). This compares with an estimate of 18 to 20% of all such common SNPs present in the collection generated by the International SNP Map Working Group and the SNP Consortium (*17*). The difference in coverage is explained by the fact that we used larger numbers of chromosomes for SNP discovery. To assess the replicability of our findings, we performed SNP discovery for one wafer design with 19 additional copies of chromosome 21 derived from the same diversity panel as the original set of samples. We identified a total of 7188 SNPs using the two sets of samples. On average, 66% of all SNPs found in one set of samples were discovered in the second set, consistent with previous findings (*20*, *21*). As expected, failure of a SNP to replicate in a second set of samples is strongly dependent on allele frequency. We found that 80% of SNPs with a minor allele present two or more times in a set of samples were also found in a second set of samples, whereas only 32% of SNPs with a minor allele present a single time were found in a second set of samples. These findings suggest that the 24,047 SNPs in our collection with a minor allele represented more than once are highly replicable in different global samples and that this set of SNPs is useful for defining common global haplotypes (*22*). In addition to the replicability of SNPs in different samples, the distance between consecutive SNPs in a collection of SNPs is critical for defining meaningful haplotype structure. Haplotype blocks, which can be as short as several kilobases, may go unrecognized if the distance between consecutive SNPs in a collection is large relative to the size of the actual haplotype blocks. Our collection of SNPs is very evenly distributed across the chromosome, even though we did not include repeat sequences in our SNP discovery process (Fig. 1C). The average distance between consecutive SNPs was 900 bases when all SNPs are considered, and 1300 bases when one considers only the 24,047 common SNPs. For this set of common SNPs, 93% of the intervals between consecutive SNPs in genomic DNA, including repeated DNA, were 4000 bases or fewer (Fig. 1C).

The construction of haplotypes from diploid data is complicated by the fact that the relation between alleles for any two heterozygous SNPs is not directly observable. Consider an individual with two copies of chromosome 21 and two alleles, A and G, at one chromosome 21 SNP, as well as two alleles, A and G, at a second chromosome 21 SNP. In such a case, it is unclear whether one copy of chromosome 21 contains allele A at the first SNP and allele A at the second SNP but the other copy of chromosome 21 containins allele G at the first SNP and allele G at the second SNP, or whether one copy of chromosome 21 contains allele A at the first SNP and allele G at the second SNP but the other copy of chromosome 21 contains allele G at the first SNP and allele A at the second SNP. Current methods used to circumvent this problem include statistical estimation of haplotype frequencies, direct inference from family data, and allele-specific PCR amplification over short segments (*23*, *24*). To avoid the uncertainty and missing information inherent in all of these methods (*25*), we characterized SNPs on haploid copies of chromosome 21 isolated in rodent-human somatic cell hybrids, a process that allowed us to directly determine the full haplotypes of these chromosomes. We have used the set of 24,047 SNPs with a minor allele represented more than once in our data set to define the haplotype structure (Fig. 2). Although no two chromosomes shared an identical haplotype pattern for these 147 SNPs, there are numerous regions in which multiple chromosomes shared a common pattern. One such region, defined by 26 SNPs spanning 19 kb, is expanded for more detailed analysis (Fig. 2). This block defines seven unique haplotype patterns in 20 chromosomes. Despite the fact that some data is missing because it did not pass the threshold for data quality, in all cases a given chromosome can be assigned unambiguously to one of the seven haplotypes. The four most frequent haplotypes, each of which is represented by three or more chromosomes, account for 80% of all chromosomes in the sample. Only 2 SNPs out of the total of 26 are required to distinguish the four most frequent haplotypes from one another (Fig. 2). In this example, four chromosomes with

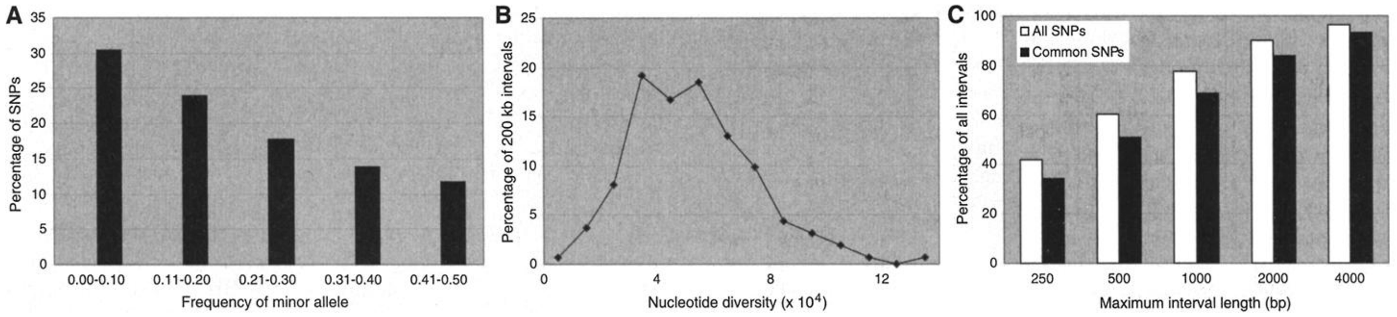

Fig. 1. (**A**) The distribution of minor allele frequencies of all 35,989 SNPs discovered in our sample of globally diverse chromosomes. (**B**) The distribution of nucleotide diversity. The 32,397,439 bases of finished genomic chromosome 21 DNA were divided into 200,000 base pair segments, and the high-quality base pairs used for SNP discovery in each segment were examined. The observed heterozygosity of these bases was used to calculate an average nucleotide diversity (π) for each segment. (**C**) The distribution of SNP coverage across 32,397,439 bases of finished chromosome 21 DNA sequence. An interval is the distance between consecutive SNPs. There are a total of 35,988 intervals for the entire SNP set and a total of 24,046 intervals for the common SNP set (i.e., SNPs with a minor allele present more than once in the sample).

infrequent haplotypes would be misclassified as common haplotypes by using information from only these two SNPs. Nevertheless, it is remarkable that 80% of the haplotype structure of the entire global sample is defined by less than 10% of the total SNPs in the block. Several different possibilities exist in which three SNPs can be chosen so that each of the four common haplotypes is defined uniquely by a single SNP (Fig. 2). One of these "three SNP" choices would be preferred over the two SNP combination in an experiment that involves genotyping of pooled samples, because the two SNP combination would not permit determination of frequencies of the four common haplotypes in such a situation. In summary, although the particular application may dictate the selection of SNPs to capture haplotype information, it is clear that the majority of the haplotype information in the sample is contained in a very small subset of all the SNPs. It is also clear that random selection of two or three SNPs from this block of SNPs will often not provide enough information to assign a chromosome to one of the four common haplotypes.

An unresolved issue is how to define a set of contiguous blocks of SNPs spanning the entire 32.4 Mb of chromosome 21 while minimizing the total number of SNPs required to define the haplotype structure. We use greedy optimization algorithm to address this problem. We begin by considering all possible blocks of physically consecutive SNPs of size one SNP or larger. We exclude all blocks in which less than 80% of the chromosomes in the sample that provide data are defined by haplotypes represented more than once in the block (i.e., 80% coverage). Ambiguous haplotypes are treated as missing data and are not included when calculating percent coverage. Considering the remaining overlapping blocks simultaneously, we select the one with the maximum ratio of total SNPs in the block to the minimal number of SNPs required to uniquely discriminate haplotypes represented more than once in the block. Any of the remaining blocks that physically overlap with the selected block are discarded, and the process is repeated until we have selected a set of contiguous, nonoverlapping blocks that cover the 32.4 Mb of chromosome 21 with no gaps and with every SNP assigned to a block. Given our sample size of 20 chromosomes, the algorithm produces a maximum of ten common haplotypes per block, each represented by two independent chromosomes. Applying this algorithm to our data set of 24,047 common SNPs, we defined 4135 blocks of SNPs spanning chromosome 21 (Table 1) (*26*). A total of 589 blocks, which is 14% of the total number of blocks, contain greater than ten SNPs per block and compose 44% of the total 32.4 Mb. In contrast, 2138 blocks, which is 52% of all blocks, contain less than three SNPs per block and make up only 20% of the physical length of the chromosome. The largest block contains 114 common SNPs and spans 115 kb of genomic DNA. Over all the average physical size of a block is 7.8 kb. The size of a block is not correlated with its order on the chromosome, and large blocks are interspersed with small blocks along the length of the chromosome.

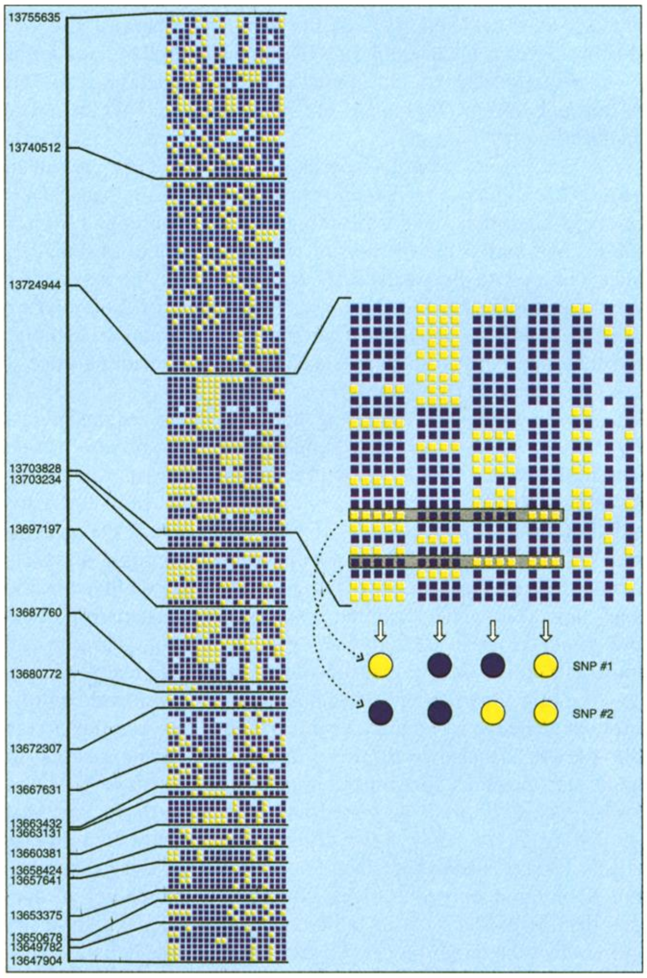

Fig. 2. The haplotype patterns for 20 independent globally diverse chromosomes defined by 147 common human chromosome 21 SNPs. The 147 SNPs span 106 kb of genomic DNA sequence. Each row of colored boxes represents a single SNP. The blue boxes in each row represent the major allele for that SNP, and the yellow boxes represent the minor allele. Absence of a box at any position in a row indicates missing data. Each column of colored boxes represents a single chromosome, and the SNPs are arranged in their physical order on the chromosome. Invariant bases between consecutive SNPs are not represented in the figure. The 147 SNPs are divided into 18 blocks, defined by black horizontal lines. The position of the base in chromosome 21 genomic DNA sequence defining the beginning of one block and the end of the adjacent block is indicated by each number to the left of the vertical black line. The expanded boxes on the right of the figure represent a SNP block defined by 26 common SNPs spanning 19 kb of genomic DNA. Of the seven different haplotype patterns represented in the sample, the four most common patterns include 16 of the 20 chromosomes sampled (i.e., 80% of the sample). The blue and yellow circles indicate the allele patterns of two SNPs, which unambiguously distinguish the four common haplotypes in this block.

Table 1. Properties of SNP blocks. Common SNPs are the 24,047 SNPs with a minor allele present more than once in the sample of 20 chromosomes used to define SNP blocks. Common haplotypes are the haplotypes in each block present more than once in the sample of 20 chromosomes. All bases are the 32,397,439 bases of finished chromosome 21 genomic DNA sequence which were the basis of this study.

Common SNPs/block	No. of blocks	Avg. π (×10^4)	Avg. size/block (kb)	Avg. no common haplotypes/block	All blocks (%)	Common SNPs (%)	% of all bases	% of all exonic bases
>10	589	8.27	23.90	3.75	14.2	56.8	43.5	33.80
3 to 10	1408	6.48	8.52	2.92	34.1	30.7	37.0	45.20
<3	2138	6.26	2.96	2.30	51.7	12.4	19.5	20.90
Total	4135	7.23	7.83	2.72	100.0	100.0	100.0	100.0

On average, there are 2.7 common haplotypes per block, defined as haplotypes that are observed on multiple chromosomes. The most frequent haplotype in a block is represented by 9.6 chromosomes of the 20 in our sample, the second most frequent haplotype is represented by 4.2 chromosomes, and the third most frequent haplotype, if present, is represented by 2.1 chromosomes. It is remarkable that such a large fraction of globally diverse chromosomes are represented by such limited haplotype diversity. Our findings are consistent with the observation that when haplotype frequency is considered, 82% of the haplotypes observed in a collection of 313 human genes are observed in all ethnic groups, whereas only 8% of haplotypes are population-specific (*18*).

We performed several experiments to measure the influence of parameters of the haplotype algorithm on the resulting block patterns. We varied the fraction of chromosomes required to be covered by common haplotypes, from our original 80%, to 70 and 90%. As would be expected, changing the algorithm to require more complete coverage results in somewhat larger numbers of shorter blocks [Web table 2 (*26*)]. Using only the 16,503 SNPs with a minor allele frequency of at least 20% in our sample resulted in somewhat longer blocks, but the numbers of SNPs per block did not change significantly [Web table 3 (*26*)]. For one region of about 3 Mb, we analyzed a larger sample of 38 chromosomes for SNPs and common haplotype blocks with at least 10% frequency to be comparable to our 20-chromosome analysis. The resulting distribution of block sizes closely matched our original results [Web table 4 (*26*)]. We also performed a randomization test in which the nonambiguous alleles at each SNP were permuted and then used for haplotype block discovery. In this analysis, 94% of blocks contained fewer than three SNPs, and only one block contained more than five SNPs (Web table 2). This confirms that the larger blocks we see in our original data cannot be produced by chance associations nor can they be artifacts of our block selection algorithm.

In an effort to determine whether genes were proportionately represented in both large and small blocks, we determined the number of exonic bases in blocks containing >10 SNPs, 3 to 10 SNPs, and <3 SNPs (Table 1). Exonic bases are somewhat overrepresented as compared to total bases in blocks containing 3 to 10 SNPs ($P < 0.05$, as determined by a permutation test).

On the basis of known haplotype structure within blocks, we can select subsets of the 24,047 common SNPs to capture any desired fraction of the common haplotype information. We define common haplotype information as complete information for haplotypes that are present more than once and include more than 80% of the sample across the entire 32.4 Mb (Fig. 3). For example, a minimum of 4563 SNPs are required to capture all the common haplotype information, but only 2793 SNPs are required to capture the common haplotype information in blocks containing three or more SNPs, which cover 81% of the 32.4 Mb. A total of 1794 SNPs are required to capture all the common haplotype information in genic DNA, representing approximately 220 distinct genes.

Our results have particular relevance for whole-genome association studies mapping common disease genes. This approach relies on the hypothesis that common genetic variants are responsible for susceptibility to common diseases (*27*, *28*). By comparing the frequency of genetic variants in unrelated cases and controls, genetic association studies can identify specific haplotypes in the human genome that play important roles in disease. Although this approach has been used to successfully associate candidate genes with disease (*29*), the recent availability of the human DNA sequence offers the possibility of surveying the entire genome, dramatically increasing the power of genetic association analysis (*30*). A major limitation to the implementation of this method has been lack of knowledge of the haplotype structure of the human genome, which is required in order to select the appropriate genetic variants for analysis. The unpredictable nature of the haplotype structure in any particular genomic region demands a comprehensive, empirical approach. Our results demonstrate that high-density oligonucleotide arrays in combination with somatic cell genetic sample preparation provide a high-resolution approach to empirically define the common haplotype structure of the human genome. Although the length of genomic regions with a simple haplotype structure is extremely variable, a dense set of common SNPs enables our systematic approach to define blocks of the human genome in which 80% of the global human population is described by only three common haplotypes. In general, when applying our particular algorithm, the most common haplotype in any block is found in 50% of individuals, the second most common in 25% of individuals, and the third most common in 12.5% of individuals. It is important to note that blocks are defined based on their genetic information content and not on knowledge of how this information originated or why it exists. As such, blocks do not have absolute boundaries and may be defined in different ways, depending on the specific application. Our algorithm provides only one of many possible approaches. Our results indicate that a very dense set of SNPs is required to capture all the common haplotype information. Once in hand, however, this information can be used to identify much smaller subsets of SNPs useful for comprehensive whole-genome association studies.

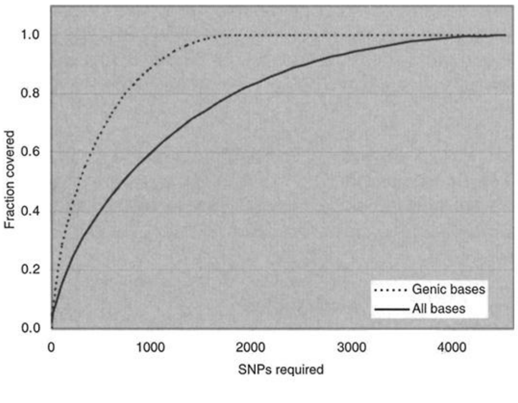

Fig. 3. The number of SNPs required to capture the common haplotype information for 32.4 Mb of chromosome 21. For each SNP block, we determine the minimum number of SNPs required to unambiguously distinguish haplotypes in that block which are present more than once (i.e., common haplotype information). These SNPs provide common haplotype information for the fraction of the total physical distance defined by that block. Beginning with the SNPs that provide common haplotype information for the greatest physical distance, the cumulative increase in physical coverage (i.e., fraction covered) is plotted relative to the number of SNPs added (i.e., SNPs required). Genic DNA includes all genomic DNA beginning 10 kb 5′ to the first exon of each known chromosome 21 gene and extending 10 kb 3′ to the last exon of that gene.

References and Notes

1. D. G. Wang *et al.*, *Science* **280**, 1077 (1998).
2. M. Cargill *et al.*, *Nature Genet.* **22**, 231 (1999).
3. M. K. Halushka *et al.*, *Nature Genet.* **22**, 239 (1999).
4. L. Kruglyak, D. A. Nickerson, *Nature Genet.* **27**, 235 (2001).
5. S. M. Fullerton *et al.*, *Am. J. Hum. Genet.* **67**, 881 (2000).
6. A. G. Clark *et al.*, *Am. J. Hum. Genet.* **63**, 595 (1998).
7. G. R. Abecasis *et al.*, *Am. J. Hum. Genet.* **68**, 191 (2001).
8. D. E. Reich *et al.*, *Nature* **411**, 199 (2001).
9. F. S. Collins, L. D. Brooks, A. Chakravarti, *Genome Res.* **8**, 1229 (1998).
10. J. A. Douglas *et al.*, *Nature Genet.* **28**, 361 (2001).
11. Eight unique oligonucleotides, each 25 bases in length, were used to interrogate each of the unique chromosome 21 bases, for a total of 1.7 × 10^8 different oligonucleotides. These oligonucleotides were distributed over a total of eight different wafer designs using a previously described tiling strategy [M. Chee *et al.*, *Science* **274**, 610 (1996)]. Light-directed chemical synthesis of oligonucleotides was carried out on 5 inch by 5 inch glass wafers by Affymetrix, Inc. (Santa Clara, CA).
12. LR-PCR assays were designed using Oligo 6.23 primer design software with high to moderate stringency parameters. The resulting primers were typically 30 nucleotides in length with the melting temperature of >65°C. The range of amplicon size was from 3 to 14 kb. A primer database for the entire chromosome

was generated and custom software (pPicker; Perlegen Sciences, Inc., Mountain View, CA) was designed to choose a minimal set of nonredundant primers that yield maxium coverage of chromosome 21 sequence with a minimal overlap between adjacent amplicons. LR-PCR reactions were performed using the Expand Long Template PCR Kit (Roche Biosciences, Palo Alto, CA) with minor modifications.

13. LR-PCR targets were prepared as previously described with some modifications (*1*). For each wafer hybridization, corresponding LR-PCR products were pooled and purified using Qiagen tip 500 (Qiagen, Valencia, CA). A total of 280 μg of purified DNA was fragmented using 37 μl of 10× One-Phor-All buffer PLUS (Promega, Madison, WI) and 1 unit of DNAase (Life Technolgies/Invitrogen, Carlsbad, CA) in 370 μl total volume at 37°C for 10 min, which was then heat-inactivated at 99°C for 10 min. The fragmented products were end labeled using 500 units of Tdt (Boehringer Manheim) and 20 nmoles of biotin-N6-ddATP (DuPont NEN, Boston, MA) at 37°C for 90 min and heat inactivated at 95°C for 10 min. The labeled samples were hybridized to the wafers in 10 mM tris-HCL (pH 8), 3 M tetramethylammonium chloride, 0.01% Tx-100, 10 μg/ml denatured herring sperm DNA in a total volume of 14 ml per wafer at 50°C for 14 to 16 hours. The wafers were rinsed briefly in 4× SSPE, washed three times in 6× SSPE for 10 min each, and stained with streptavidin R-phycoerythrin (SAPE; 5 ng/ml) at room temperature for 10 min. The signal was amplified by staining with an antibody against streptavidin (1.25 ng/ml) and by repeating the staining step with SAPE. The wafers were scanned using a custom-built confocal scanner.
14. A combination of previously described algorithms (*1*), was used to detect SNPs based on altered hybridization patterns.
15. Consistent failure of LR-PCR in all samples analyzed accounts for 15% of the 35% false negative rate. The remaining 20% false negatives are distributed between bases that never yield high-quality data (10%) and bases that yield high-quality data in only a fraction of the 20 chromosomes analyzed (10%). In general, it is the sequence context of a base that dictates whether or not it will yield high-quality data. Our finding that approximately 20% of all bases give consistently poor data is very similar to the finding that approximately 30% of bases in single dideoxy sequencing reads of 500 bases have quality scores too low for reliable SNP detection [D. Altschuler *et al.*, *Nature* **407**, 513 (2000)]. The power to discover rare SNPs as compared to more frequent SNPs is disporportionately reduced in cases where only a limited number of the samples analyzed yield high-quality data for a given base. As a result, our SNP discovery is biased in favor of common SNPs.
16. D. L. Hartl, A. G. Clark, *Principles of Population Genetics* (Sinauer, Sunderland, MA, 1997), pp. 57–60.
17. The International SNP Map Working Group, *Nature* **409**, 928 (2001). We compared the overlap of 15,549 chromosome 21 SNPs discovered by The SNP Consortium (TSC) with the SNPs found in this study. Of the TSC SNPs, 5087 were found to be in repeated DNA and were not tiled on our wafers. Of the remaining 10462 TSC SNPs, we identified 4705 (45%).
18. J. C. Stephens *et al.*, *Science* **293**, 489 (2001)
19. Y. Fu, W.-H. Li, *Genetics* **133**, 693 (1993).
20. G. Marth *et al.*, *Nature Genet.* **27**, 371 (2001).
21. Z. Yang *et al.*, *Nature Genet.* **26**, 13 (2000).
22. In the course of SNP discovery, we identified 339 SNPs that appeared to have more than two alleles. These SNPs were not included in any analyses.
23. L. Excoffier, M. Slatkin, *Mol. Biol. Evol.* **12**, 921 (1995).
24. S. Michalatos-Beloin *et al.*, *Nucleic Acids Res.* **24**, 4841 (1996).
25. S. E. Hodge, M. Boehnke, M. A. Spence, *Nature Genet.* **21**, 360 (1999).
26. Supplementary data delineating the precise boundaries of the SNP blocks described in this paper as well as the haplotypes identified for each block in the 20 chromosomes sampled are available at *Science* Online at www.sciencemag.org/cgi/content/full/294/5547/1719/DC1 and www.perlegen.com/haplotype.
27. N. Risch, K. Merikangas, *Science* **273**, 1516 (1996).
28. E. Lander, *Science* **274**, 536 (1996).
29. D. Altschuler *et al.*, *Nature* Genet. **26**, 76 (2000).
30. L. Kruglyak, *Nature Genet.* **22**, 139 (1999).
31. We thank B. Margus, E. Rubin, A. Chakravarti, and E. Lander for helpful discussions and an anonymous reviewer for suggestions that significantly improved the manuscript.

20 August 2001; accepted 1 October 2001

A Genomic-Systems Biology Map for Cardiovascular Function

Monika Stoll,[1] Allen W. Cowley Jr.,[1] Peter J. Tonellato,[1,2] Andrew S. Greene,[1] Mary L. Kaldunski,[1] Richard J. Roman,[1] Pierre Dumas,[1,3] Nicholas J. Schork,[4,5,6] Zhitao Wang,[1,2] Howard J. Jacob[1,3*]

With the draft sequence of the human genome available, there is a need to better define gene function in the context of systems biology. We studied 239 cardiovascular and renal phenotypes in 113 male rats derived from an F_2 intercross and mapped 81 of these traits onto the genome. Aggregates of traits were identified on chromosomes 1, 2, 7, and 18. Systems biology was assessed by examining patterns of correlations ("physiological profiles") that can be used for gene hunting, mechanism-based physiological studies, and, with comparative genomics, translating these data to the human genome.

Genetic studies of multifactorial disorders in human populations remain challenging due to the modest nature of gene effects and the heterogeneity of patient populations. The difficulties investigators face in identifying QTLs in multifactorial diseases have become apparent from the results obtained from recent total genome scans for asthma (*1*), hypertension (*2, 3*), NIDDM (*4*), and IDDM (*5*) in diverse human populations. Hypertension is one such multifactorial disorder that develops as a consequence of an "error" in the complex and redundant biological systems that determine blood pressure. The present manuscript describes the results of studies in which 239 phenotypes in each animal have been analyzed (i) to develop a model of the systems biology of the rat for renal, vascular, and neurohumoral function; (ii) to develop a correlational physiological model of the relationships among these phenotypes; and (iii) to develop bioinformatic tools to link the genetic model and the physiological model. The output for the linkage map and use of the physiological profiling tool can be found at (*6*).

A comprehensive genetic linkage map of 239 "likely determinant phenotypes" of blood pressure was first developed (*7*) using a total genome scan with a ~10-cM interval between markers to produce a detailed system biology map (*6*). Many of the quantitative trait loci (QTLs) for blood pressure aggregate (six or more QTLs with overlapping 95% confidence intervals) in discrete regions on rat chromosomes 1, 2, 7, and 18 (Fig. 1) (*6*). In four of these five aggregates, the phenotypes were independent, indicating that the cluster of traits is likely to be the result of separate genes rather than pleiotropy. In the fifth set, on chromosome 18, significant correlations were found among the phenotypes that could be divided into three functional groups, i.e., vascular reactivity, plasma lipid concentrations, and renal function, suggesting a functional genes cassette, as has been observed for QTLs in agriculture (*8*) and biomedical research (*9–11*).

To date, the majority of genome wide scans searching for the genetic basis of hypertension in rats have focused on a limited number of phenotypes, typically blood pressure and heart rate. The results of these studies have identified QTLs on almost every rat chromosome, with loci confirmed on chromosomes 1, 2, 3, 5, 10, and 13 (*12*). Unfortunately none of these have been translated into genes. The need for improved tools, including better phenotypes, to identify the genes responsible for these QTLs has been well articulated by Nadeau and Frankel (*11*). The present comprehensive linkage study in

[1]Department of Physiology, [2]Bioinformatics Research Center, and [3]Human and Molecular Genetics Center, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, WI 53226–0509, USA. [4]Case Western Reserve University, Cleveland, OH 44106, USA. [5]The Jackson Laboratory, Bar Harbor, ME 04609, USA. [6]Harvard School of Public Health, Boston, MA 02115, USA.

*To whom correspondence should be addressed. E-mail: jacob@mcw.edu