A map of the human genome in linkage disequilibrium units

W. Tapper*, A. Collins, J. Gibson, N. Maniatis, S. Ennis, and N. E. Morton*

Human Genetics Division, University of Southampton, Southampton General Hospital, Southampton SO16 6YD, United Kingdom

Contributed by N. E. Morton, June 23, 2005

Two genetic maps with additive distances contribute information about recombination patterns, recombinogenic sequences, and discovery of genes affecting a particular phenotype. Recombination is measured in morgans (w) over a single generation in a linkage map but may cover thousands of generations in a linkage disequilibrium (LD) map measured in LD units (LDU). We used a subset of single nucleotide polymorphisms from the HapMap Project to create a genome-wide map in LDU. Recombination accounts for 96.8% of the LDU variance in chromosome arms and 92.4% in their deciles. However, deeper analysis shows that LDU/w, an estimate of the effective bottleneck time (t), is significantly variable among chromosome arms because (i) the linkage map is approximated from the Haldane function, then adjusted toward the Kosambi function that is more accurate but still exaggerates w for all chromosomes, especially shorter ones; (ii) the nonpseudoautosomal region of the X chromosome is subject to hemizygous selection; and (iii) at resolution less than ~40,000 markers per w, there are indeterminacies (holes) in the LD map reflecting intervals of very high recombination. Selection and stochastic variation in small regions must have effects, which remain to be investigated by comparisons among populations. These considerations suggest an optimal strategy to eliminate holes quickly, greatly enhance the resolution of sex-specific linkage maps, and maximize the gain in association mapping by using LD maps.

**Materials and Methods**

**Genotypic Data for LD Map Construction.** The HapMap data (www.hapmap.org) were obtained on 60 parental DNA samples from Utah Mormons of northwestern European ancestry collected by the Centre d’Etude du Polymorphisme Humain. These samples are a subset of 270 in the database that includes 3 other populations. A total of 665,335 single nucleotide polymorphism (SNP) genotypes were downloaded from the September 2004 public release of the HapMap data. A little more than one-quarter of the downloaded SNPs (25.8%, 171,927) were removed by a screening procedure that rejected 6,795 SNPs with $\chi^2 > 10$ for the Hardy–Weinberg test (11) or a minor allele frequency <5% (165,132), leaving 493,408 SNPs for LD map construction. These SNPs are a subset of the ~1 million SNPs included in the HapMap release after our analysis was completed. Both samples take their nucleotide position from the July 2003 Golden Path database (http://genom.ucsc.edu). The number of SNPs is expected to reach 3 million next year, with a subsequent increase likely in an updated physical map.

**LD Map Construction.** LD maps were constructed by the methods in the next paragraph for 23 chromosomes, 1–22 and X, covering 98% (2,790 Mb) of the euchromatin. Each SNP has an LD location, and distance between adjacent SNPs in the LD map was constrained to a maximum of 3 LDU. Such intervals are called holes (12), constituting 0.6% of the total map intervals, 17.8% of the total LDU length, and 2.2% of the total physical length. Ignoring stochastic variation and selection in the LD map and errors in estimating the linkage map in morgans (w), the Malecot model predicts that the ratio of corresponding distances in LD and linkage maps estimates $t$, the number of generations over which recombination has accumulated after one or more population bottlenecks (8). On these assumptions, $t$ would be constant between chromosomal arms and their deciles. To test this hypothesis, the initial unit of analysis, neglecting acrocentric short arms, centromeres, and pseudoautosomal regions, was the length of a chromosome arm between the first and last physical locations shared by the two genetic maps (linkage and LD). For the X chromosome, the two pseudoautosomal regions were removed and the residual linkage map length in females was multiplied by 2/3 to allow for the absence of crossing over in males. LDMAP was used for autosomes and the female X chromosome to calculate pairwise association probabilities $\rho$ and information $K_P$ under the null hypothesis that $\rho = 0$, so that $\rho^2K_P = \chi^2$ (13–15). Haplotype frequencies for pairs of loci were directly observed for male X chromosomes, and the association probabilities and information were determined from these counts. Association probabilities were then obtained as the weighted mean of sex-specific estimates $r_m$ and $r_f$, and the information was taken as the sum of the weights $K_{r_m}$ and $K_{r_f}$. Default parameters were used to construct LD maps by estimating the length of an interval between adjacent SNPs (maximal window of adjacent intervals = 100, maximum distance between any SNP pair = 500 kb, segment size = 500 markers with a 25-marker overlap in adjacent segments, and overlap

**Abbreviations:** LD, linkage disequilibrium; LDU, linkage disequilibrium unit; w, morgans; cm, centimorgans.

*To whom correspondence may be addressed. E-mail: wtj@soton.ac.uk or nem@soton.ac.uk.

© 2005 by The National Academy of Sciences of the USA
distance is averaged). These defaults give rapid construction of a good LD map, with the flanking intervals contributing efficiently to each interval estimate.

The predicted value of association $\rho$ between two SNPs at a distance $\Sigma d_i$ kb is $\rho = (1 - L)M e^{-2\xi d_i} + L$, where $d_i$ is the kb length of the $i$th included interval between adjacent SNPs and $\Sigma d_i$ is the corresponding LDU distance (4). On certain assumptions, $\Sigma d_i$ has expectation $wt$ (8). However, $w$ and $t$ are not used when constructing the LD map. The model is fitted from the composite likelihood $\exp[-\Sigma K_i (\hat{\rho} - \rho)^2/2]$, where $\hat{\rho}$ is an estimate with prediction $\rho$ and the summation is over $n$ pairs of SNPs used for LD analysis within a given window containing $r$ SNPs. The asymptote $L$ was predicted for each segment as the weighted mean deviation for a normal distribution (4).

Statistical Analysis of LD Maps. LD maps were analyzed by chromosome arms between the first and last physical location shared by the linkage and LD map, omitting heterochromatic, centromeric, and pseudoautosomal regions. This procedure excludes 3,709 SNPs distal to the linkage map, comprising 48.1 Mb of the LD map span and leaving 489,699 SNPs covering 2,790 Mb to be analyzed. Correlations were examined to suggest models for stepwise linear regression. Inferences from chromosome arms were confirmed by partitioning each physical map into deciles. Assuming that the variance of LDU is proportional to $w$, we calculated effective bottleneck time in generations ($t$) and its variance ($V_t$). Weighting each of $s$ arms or deciles by length on the linkage map in $w$, with length LDU on the LD map, these estimates are

$$t = \frac{\sum w(LDU/w)}{\sum w} = \frac{\sum LDU}{\sum w}$$

$$V_t = \frac{\sum w[(LDU/w)^2 - \left(\frac{\sum w(LDU/w)}{\sum w}\right)^2]}{\sum w(s - 1)/s}$$

For both arms and deciles, the smallest values of $V_t$ were obtained when using the most recent linkage map (1), compared with earlier linkage maps constructed from deCODE data by using a multipoint approach (16) and a composite likelihood method (17). We confirmed this choice by verifying that it gave the smallest reduction in error variance when independent parameters were introduced with LDU/w as a dependent variable that is weighted by $w$. The independent variables for all maps were $h/w$, Mb/w, $m/w$, and their squares, where $h$ and $m$ are counts of holes and markers, respectively. We verified that forward and backward stepwise regression gave the same model. When linear and quadratic terms were both significant for a particular variable, we accepted an exponential function if it gave a better fit.

Results

Recombination Dominates Patterns of LD. Regressing LDU on $w$ through the origin and weighting by $w$ gives the same slope $t = \Sigma LDU/\Sigma w = 1.435$ for the 41 chromosome arms and their 410 deciles. This result is the ratio of their capabilities to resolve causal from predictive markers in association mapping. On simple assumptions, it is also the number of generations over which recombination has accumulated in the LD map during a succession of bottlenecks in population size. Deviations from this regression account for only 3.2% of the variance of LDU/w for arms and 7.6% for deciles, including random errors in both LDU and $w$ and significant effects of other variables (Fig. 1). Assuming a maximum of 25 years per generation, the effective bottleneck time is no more than $25t = 35,875$ years for this population, diminished if generation time were reduced. This finding is less than half the time since the out-of-Africa bottleneck ($\approx 100,000$ years), reflecting subsequent population bottlenecks and justifying the term effective bottleneck time as an analogue of effective population size. Recovery of diversity after a bottleneck can be rapid if determined by migration into a local population but is extremely slow if dominated by mutation within a species (18).

Unexpected Variation in Estimates of Effective Bottleneck Time. Although recombination accounts for agreement with the Malecot prediction of uniform $t$, deviations may be significant. We therefore examine a range of variables, including LDU, $w$, markers ($m$), megabases (Mb), and holes ($h$), for deciles and performed stepwise regression with LDU/w as the dependent variable weighted by $w$. Chromosome arm size (Mb/w), marker density ($m/w$), and hole density ($h/w$) are all significant predictors and were next studied separately.

Chiasma Interference. The first analysis examined the relationship between $t$ as LDU/w and arm length as Mb/w. Computer programs in common use for constructing linkage maps assume no chiasma interference for the computation of multilocus likelihoods and may then convert the resulting map to approximate conformity with the Kosambi function, which is not multipoint feasible. Studies of chiasma interference in humans have shown that it exists at greater levels than the Kosambi function provides and appears to vary among chromosomes (19–21). The estimate of $t$ is significantly lower for smaller autosomes (Fig. 2), supporting evidence that chiasma interference becomes more intense with decreasing chromosome size, as in the mouse (21).

Natural Selection. The X chromosome is exceptional in having unusually low LDU/w (high LD), despite correcting for the absence of recombination in males and is therefore excluded from the weighted exponential fit. This finding is consistent with more rapid selection against deleterious mutations when the X is monosomic in males, and to a lesser extent, under random inactivation (Lyonization) in females (22–23). The smaller effective population size $N_e$ for the X chromosome than for autosomes is not a factor, because $N_e$ affects the M parameter of the Malecot equation but not the LDU length $\Sigma e d_i$ (15).
Incomplete Data. Some values of $e_d$ are indeterminate and are assigned a maximum value of 3 LDU, which may become as little as 2.5 in subsequent iterations. These holes reflect segments with elevated recombination and/or insufficient SNP density in steps where holes occur. This uncertainty will not be resolved until the density of SNPs within holes increases, which would also increase the power of association mapping. Here we take $e_d > 2.5$ as defining a hole. There is a significant relationship between density of holes ($h/w$) and markers ($m/w$). Fig. 3 suggests that the number of holes will decline with progress of the HapMap Project and would decline more rapidly if SNP selection focused on the 3,144 holes in the LD map or if holes were defined on a cosmopolitan map that averages data for two or more populations (7). However, the factors that determine holes are complex, dominated by recombination but also including SNP distributions (7). The number of holes in future databases cannot be estimated reliably but is likely to decrease to a nonzero limit as LD maps evolve.

**High Resolution Sex-Specific Linkage Maps.** We see that both linkage and LDU maps are more complex than their simple models, despite nearly a century of development for the former and nearly three years for the latter. Nevertheless, the low resolution of the sex-specific linkage maps can be greatly increased by interpolation from the LD map. To motivate this increase, we examined LDU/Mb and cM/Mb (cM, centimorgans) in 2-Mb sliding windows for chromosome 19 (Fig. 4) comparing the high-resolution LD map with the low-resolution linkage map. At this resolution, blocks and steps in the LD map are not visible, but major peaks and troughs in recombination rate (cM/Mb) and strength of association (LDU/Mb) are apparent and show high concordance between LD and linkage maps. This agreement remains but to a lesser extent as window size is reduced. Such graphs illustrate variation within chromosomes and the degree of sharing between maps at arbitrary resolution but do not provide either an LD map or linkage map. The q arm was previously examined in a similar way by a coalescent method, with less detail and conspicuous distortion near the centromere (24).

The sex of ancestors in whom recombination took place is unknown for LD maps, but this fact does not imply that LD maps cannot be used to create sex-specific linkage maps at high resolution. All that is required is to conserve framework loci from the appropriate linkage map, interpolating locations from the LD map between adjacent framework loci. Because the same physical map was used for linkage and LD files, all linkage markers not assigned to the same base pair in the LD file were interpolated from the physical map and used as framework loci if the distance to the preceding framework locus on the LD map was zero (i.e., in the same block) or if the distance in the linkage map was nonzero (i.e., recombination had been detected and, therefore, not in the same block). In this way, the sex-specific linkage framework was maintained during interpolation from the high-resolution LD map. Fig. 5 illustrates the profound sex differences in recombination, with males accounting for most crossovers near the telomeres and females responsible for most recombination near the centromere. The ratio cM/Mb is proportional to the Malecot parameter $\kappa = \Sigma e_d / \Sigma d$, the rate of change of the LDU map with respect to the physical map. The LDU length $\Sigma e_d$ is much smoother and increases monotonically (Fig. 6), but it identifies the same peaks. Coalescent methods conceal these differences by interpolating between the pair of most distant markers shared between the sex-averaged linkage map and the coalescent construct, which is neither a linkage map nor an LD map (24). Applied to LD for a particular chromosome in different populations, coalescence adjusts all of them to the sex-averaged linkage map and therefore to the same length determined by misrepresentation of interference, leaving only selected sequences as a memento of LD differences.

---

**Fig. 2.** A graph showing variation among chromosome arms for the ratio of the LD map in LDU/Mb to the physical map in Mb/Mb.

**Fig. 3.** The declining density of holes with marker density among chromosome arms.

**Fig. 4.** A graph of chromosome 19 for LDU/Mb (blue) and centimorgans (cM)/Mb (red) against the physical scale in Mb.
Discussion

Progress in human genomics has been so rapid that terminology
has not kept up. Cloning is no longer necessary to localize a gene,
and so the term positional cloning is currently being replaced by
association mapping. PubMed lists many papers on linkage
disequilibrium mapping, meaning that some aspect of LD is
being used. We argue that an LD map should not refer to the
annotation of genomic sequence with LD blocks and other
features, but most usefully defines a map with additive distances
that describes the pattern of LD. It is not sufficient to construct
a high-resolution recombination map, because association mapping
requires characterization of LD patterns that reflect dura-
tion together with the effects of drift, selection, and mutation.
An LD map provides a tool for localizing genetic effects that
takes the same role for association mapping at high resolution
that the linkage map provides for low-resolution mapping.
However, an LD map is not merely a scaled linkage map, but it
is a logically different entity with its own story to tell of recombina-
tion, selection, population history, and gene expression.

In contrast, the linkage map is uniquely able to identify and
use sex-specific recombination patterns. Although the prop-
eties of linkage and LD maps are different, their relation is,
at present, one-sided. Because of its much greater resolution,
the LD map can be usefully interpolated into nonzero intervals
of sex-specific linkage maps. On the contrary, current linkage
maps have nothing to contribute to LD maps, which have no
chiasma interference because multiple recombination within
small regions takes place in different generations, giving a
composite likelihood that provides a benchmark for associa-
tion mapping (5, 6). The best sex-averaged linkage map should
be most nearly proportional to an LD map, and conversely,
without trying to impose one entity on the other. Evidence on
selection and population history is fully retained, with no
incentive to incorporate information from a linkage map at a
much lower resolution. Linkage and LD are separate but
complementary, with their different evidence as uncon-
founded, as is consistent with the goal of enhancing resolution of
the linkage map without degrading the LD map. It remains
to be determined whether one linkage map will suffice for
different populations and whether a single composite LD map
will be efficient for association mapping in samples with
different Malecot parameters (7).

Mathematical geneticists frequently warn about problems
absent from mutation models but central to coalescent theory
of recombination. “The full likelihood model will be very
difficult to specify without unrealistically stringent assump-
tions about population history” (25). “The lineages we follow
never recombine with each other (the probability of such an
event is vanishingly small). They always recombine with the
(infinitely many) nonancestral chromosomes” to generate a
tree that bifurcates with no junctions (26). We may agree that
such a model is only “slightly nonintuitive” if applied to
speciation over many millions of years, leaving few surviving
haplotypes. However, the assumption cannot be so easily
swallowed that each lineage always traces in a much shorter
time to a single ancestral haplotype within a restricted popu-
lation of a single species that lacks an alternative population
to provide “infinitely many nonancestral chromosomes.” The
assumption that each lineage always traces in estimable time to
a single ancestral haplotype (“most recent common ancestor”) that
can be reliably inferred has other flaws: the rarest allele
for each polymorphism in a haplotype has a finite probability of
being the oldest one, and so the probability that m polymor-
phisms all trace to a single ancestral haplotype tends to zero
as m increases, and the most recent common ancestor and its
associated time vary greatly along a chromosome (27).

Inevitably these theoretical arguments have practical corol-
laries. In the Malecot model, the effective size is the harmonic
mean over generations and affects the M parameter but not
LDU, coalescent time is irrelevant, bottleneck time is estimable,
a founder haplotype is not inferred, the LD map estimate is
direct, degrees of freedom are specified, neighboring intervals
are not smoothed for either linkage or LD, and the population
is not assumed to be at equilibrium. In contrast, a coalescent
model assumes constant effective size N with proportional time,
does not recognize bottlenecks, assumes a unique founder
genotype, cannot directly estimate LD map length, has unspec-
ified degrees of freedom, arbitrarily smooths adjacent intervals,
and assumes equilibrium. Despite their logical differences, the
two coalescent maps currently available can be scaled to fairly
good agreement with LD maps of the same small regions (24,
28). A definitive comparison cannot be made until the coalescent
model is applied to the whole genome. Association mapping
poses a greater problem because coalescence assigns the most
recent common ancestor to a common haplotype composed of
markers with high minor allele frequencies (MAF), excluding
markers with a smaller MAF that may be predictive or causal for
association in an LD map. Experience with the two approaches
will determine the most powerful (8), but coalescent models are
heavily handicapped.

Competition between LDU maps and arbitrarily scaled sub-
stitutes is part of a larger development of analytic genomics that
will enrich the colored diagrams that conventionally represent genomes (29) by location databases (http://cedar.genetics.soton.ac.uk/public.html) in which each point on the physical map is associated with a vector of locations on other maps including, but not restricted to, LDU in multiple populations, sex-specific linkage maps, chromosome bands, and isochores. Only such location databases can provide the composite likelihoods on linkage maps, chromosome bands, and isochores. Only such identification of sequences that are recombinogenic or subject to regional selection.

We thank the Utah participants in the HapMap Project and all of the many research groups cited by the HapMap Consortium that were involved in producing the data and making them publicly available as a community resource project. This work was supported by research grants from the U.K. Biotechnology and Biological Sciences Research Council and the National Institutes of Health. The data used here are a subset of one sample from release 13 of the International HapMap Consortium (3).