coelom of stage 16/17 embryos by making a slit in the dorsal side of the embryo, between the prospective limb buds and lateral to the somites. The graft was inserted through the slit with the cut edge medial.

Recombinant limbs were prepared as described[14], except that mesenchyme was derived from the distal-most 75–100 µm of the anterior two-thirds of the wing bud and that digestion with collagenase was omitted. Recombinant buds were grafted to the somites, allowed to develop for 7 days, and stained with Victoria blue as described[19].

To rescue limb development in the absence of an AER, heparin acrylic beads were incubated in 1 mg ml^{-1} FGF4 (a gift from V. Rosen) at room temperature for 1 h, then stored on ice until use. After removal of the AER, two heparin beads were attached to the limb bud with platinum staples. One bead was placed at the posterior margin and the other just anterior to the first. After attachment of beads, embryos were kept at room temperature for 30 min before returning them to the incubator. Embryos were collected after 6 days of incubation and fixed in 4% paraformaldehyde. Skeletons were stained with 0.02% alcian blue 8GX in 70% ethanol/30% acetic acid at 37 °C, then cleared in 0.5% KOH and stored in glycerol.

Analysis of cell proliferation and death

After AER removal, embryos were returned to the incubator for 6–8 h. In cases where cell proliferation was examined, 200 µl of 5 mg ml^{-1} BrdU (Sigma) in PBS was injected around the embryo after 7 h, and the egg was returned to the incubator for 60 min. Embryos were collected, fixed in 4% paraformaldehyde/PBS overnight at 4 °C, washed in PBS, dehydrated through an ethanol series and embedded in paraffin. Sections (6 µm) were then floated onto TESPA (3-aminopropylethoxysilane)-treated slides, and apoptotic cells were stained using the method of ref. 25, except that biotin-16-dUTP was replaced by fluorescein-12-dUTP to allow for direct detection of transferred nucleotides. BrdU-labelled cells were detected using an anti-BrdU monoclonal antibody (clone BU-33, Sigma) followed by a Cy-2 conjugated goat anti-mouse secondary antibody (Jackson ImmunoResearch Laboratories). Slides for fluorescent microscopy were stained in 0.5 µg ml^{-1} 4,6-diamidino-2-phenylindole (D-1306; Molecular Probes) in PBS for 1 min before mounting in 80% glycerol.

In situ hybridizations

Both the whole mount and non-radioactive section *in situ* hybridizations were performed using standard protocols for whole-mount *in situ* hybridization. Detailed protocols are available on request.

1. Summerbell, D., Lewis, J. H. & Wolpert, L. Positional information in chick limb morphogenesis. *Nature* **244**, 492–496 (1973).
2. Rubin, L. & Saunders, J. W. Ectodermal-mesoderal interactions in the growth of limb buds in the chick embryo: constancy and temporal limits of ectodermal induction. *Dev. Biol.* **28**, 94–112 (1972).
3. Saunders, J. W. The proximo-distal sequence of origin of the parts of the chick wing and the role of the ectoderm. *J. Exp. Zool.* **108**, 363–403 (1948).
4. Summerbell, D. A quantitative analysis of the effect of excision of the AER from the chick limb-bud. *J. Embryol. Exp. Morphol.* **32**, 651–660 (1974).
5. Rowe, D. A., Cairns, J. M. & Fallon, J. F. Spatial and temporal patterns of cell death in limb bud mesoderm after apical ectodermal ridge removal. *Dev. Biol.* **93**, 83–91 (1982).
6. Niswander, L., Tickle, C., Vogel, A., Booth, I. & Martin, G. R. FGF-4 replaces the apical ectodermal ridge and directs outgrowth and patterning of the limb. *Cell* (1993).
7. Fallon, J. F. *et al.* FGF-2: apical ectodermal ridge growth signal for chick limb development. *Science* **264**, 104–107 (1994).
8. Lewis, J. H. Fate maps and the pattern of cell division: a calculation for the chick wing bud. *J. Embryol. Exp. Morphol.* **33**, 419–434 (1975).
9. Sun, X., Mariani, F. V. & Martin, G. R. Functions of FGF signalling from the apical ectodermal ridge in limb development. *Nature* **418**, 501–508 (2002).
10. Vargesson, N., Clarke, J. D. W., Coles, V. K., Wolpert, L. & Tickle, C. Cell fate in the chick limb bud and relationship to gene expression. *Development* **124**, 1909–1918 (1997).
11. Gibson-Brown, J. J. *et al.* Evidence of a role for T-box genes in the evolution of limb morphogenesis and the specification of forelimb/hindlimb identity. *Mech. Dev.* **56**, 93–101 (1996).
12. Chapman, D. L. *et al.* Expression of the T-box family genes, Tbx1-Tbx5, during early mouse development. *Dev. Dyn.* **206**, 379–390 (1996).
13. Yokouchi, Y. *et al.* Misexpression of *Hoxa-13* induces cartilage homeotic transformation and changes cell adhesiveness in chick limb buds. *Genes Dev.* **9**, 2509–2522 (1995).
14. Ros, M. A., Lyons, G. E., Mackem, S. & Fallon, J. F. Recombinant limbs as a model to study homeobox gene regulation during limb development. *Dev. Biol.* **166**, 59–72 (1994).
15. Wada, N., Ohsugi, K., Yokouchi, Y., Kuroiwa, A. & Ide, H. Cell sorting and chondrogenesis aggregate formation in limb bud recombinants and in culture. *Dev. Growth Diff.* **35**, 421–430 (1998).
16. Hardy, A. *et al.* Gene expression, polarising activity and skeletal patterning in reaggregated hind limb mesenchyme. *Development* **121**, 4329–4337 (1995).
17. Crosby, G. M. & Fallon, J. F. Inhibitory effect on limb morphogenesis by cells of the polarizing zone coaggregated with pre- or postaxial wing bud mesoderm. *Dev. Biol.* **46**, 28–39 (1975).
18. Frederick, J. M. & Fallon, J. F. The proportion and distribution of polarizing zone cells causing morphogenetic inhibition when coaggregated with anterior half wing mesoderm in recombinant limbs. *J. Embryol. Exp. Morphol.* **67**, 13–25 (1982).
19. Piedra, M. E., Rivero, F. B., Fernandez-Teran, M. & Ros, M. A. Pattern formation and regulation of gene expressions in chick recombinant limbs. *Mech. Dev.* **90**, 167–179 (2000).
20. Wolpert, L., Tickle, C. & Sampford, M. The effect of cell killing by X-irradiation on pattern formation in the chick limb. *J. Embryol. Exp. Morphol.* **50**, 175–198 (1979).
21. Hornbruch, A. & Wolpert, L. Cell division in the early growth and morphogenesis of the chick limb. *Nature* **226**, 764–766 (1970).
22. Logan, M. & Tabin, C. Targeted gene misexpression in chick limb buds using avian replication-competent retroviruses. *Methods* **14**, 407–420 (1998).
23. Hamburger, V. & Hamilton, H. L. A series of normal stages in the development of the chick embryo. *J. Morphol.* **88**, 49–92 (1951).
24. Kieny, M. & Pautou, M.-P. Proximo-distal pattern regulation in deficient avian limb buds. *Wilhelm Roux Arch.* **183**, 177–191 (1977).
25. Gavrieli, Y., Sherman, Y. & Ben-Sasson, S. A. Identification of programmed cell death in situ via specific labeling of nuclear DNA fragmentation. *J. Cell Biol.* **119**, 493–501 (1992).

....

A first-generation linkage disequilibrium map of human chromosome 22

Elisabeth Dawson*†, Gonçalo R. Abecasis†‡§, Suzannah Bumpstead*, Yuan Chen*, Sarah Hunt*, David M. Beare*, Jagjit Pabial*, Thomas Dibling*, Emma Tinsley*, Susan Kirby*, David Carter*, Marianna Papaspyridonos*, Simon Livingstone*, Rocky Ganske‖, Elin Lõhmussaar¶#, Jana Zernant#, Neeme Tõnisson#, Maido Remm¶☆, Reedik Mägi#, Tarmo Puurand¶#, Jaak Vilo**, Ants Kurg¶, Kate Rice*, Panos Deloukas*, Richard Mott‡, Andres Metspalu¶☆, David R. Bentley*, Lon R. Cardon‡ & Ian Dunham*

* *The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK*
‡ *Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK*
§ *Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109-2029, USA*
‖ *Third Wave Technologies Inc., Madison, Wisconsin 53719-1256, USA*
¶ *IMCB of the University of Tartu, and ☆ Estonian Biocentre, University of Tartu, 23 Riia Street, Tartu 51010, Estonia*
# *Asper Ltd., 3 Oru Street, Tartu 51014, Estonia*
** *European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK*
† *These authors contributed equally to this work*

....

DNA sequence variants in specific genes or regions of the human genome are responsible for a variety of phenotypes such as disease risk or variable drug response[1]. These variants can be investigated directly, or through their non-random associations with neighbouring markers (called linkage disequilibrium (LD))[2–8]. Here we report measurement of LD along the complete sequence of human chromosome 22. Duplicate genotyping and analysis of 1,504 markers in Centre d'Etude du Polymorphisme Humain (CEPH) reference families at a median spacing of 15 kilobases (kb) reveals a highly variable pattern of LD along the chromosome, in which extensive regions of nearly complete LD up to 804 kb in length are interspersed with regions of little or no detectable LD. The LD patterns are replicated in a panel of unrelated UK Caucasians. There is a strong correlation between high LD and low recombination frequency in the extant genetic

map, suggesting that historical and contemporary recombination rates are similar. This study demonstrates the feasibility of developing genome-wide maps of LD.

Present-day chromosomes are mosaics of ancestral chromosomes that have arisen through multiple recombination events in the past. Each copy of a chromosome within a population can be uniquely characterized on the basis of a specific pattern of sequence variants, which together comprise an individual 'haplotype'. When these haplotypes do not occur in the population at the frequencies expected from the component variants, the variants are said to be associated or in linkage disequilibrium (LD). Characterizing the empirical patterns of LD across the genome will help to reconstruct the genetic history of human populations, enhance our understanding of the biological processes of recombination and natural selection, and facilitate association mapping studies that seek to localize genetic variants influencing complex traits and diseases. Previous studies of LD in humans have shown a high degree of variability, indicating that LD may extend between a few and several hundred kilobases[1–5]. For practical applications, however, the local patterns of haplotype conservation are of primary interest[6–8], as the variability in LD overwhelms the average level. Therefore we characterized 59 independent haplotypes of human chromosome 22, derived from 77 members of three-generation pedigrees of the CEPH reference data set[9]. Using family samples helped to construct long haplotypes and detect genotyping error. All markers were selected from publicly available single-nucleotide polymorphisms (SNPs) and small insertions/deletions (indels)[10–12] at regularly spaced intervals along the chromosome. A total of 951 of the 1,504 (63%) polymorphisms genotyped in the CEPH panels were common (minor allele frequency of ≥ 0.2; see Methods), which is similar to the proportion of common variants expected from comparing two chromosomes drawn from a constant size neutral

population (60%). LD between pairs of markers was calculated using the measure D', following its usage in previous empirical studies, and the r^2 measure, which is preferred by population geneticists[13].

LD decays with increasing distance, but also shows extensive variability (Fig. 1a, b). Maximal D' values extend up to distances over 400 kb, contrasting with occurrences of no detectable LD ($D' < 0.20$) between markers less than 5 kb apart. The distribution of r^2 values shows a similar degree of variability, differing from D' mainly in measurement scale (Spearman rank correlation $\rho(r^2, D') = 0.95$). Similar results were obtained in analysis of two separate populations of unrelated individuals, also of Caucasian origin, one from the UK (90 individuals) and one from Estonia (51 individuals; Fig. 1c, d). In the combined CEPH and UK data sets, average D' declines from 0.70 for adjacent markers to 0.11 for unlinked markers, whereas average r^2 declines from 0.38 to 0.01. Although the two measures differ in scale, their decay profiles are similar.

We assessed the pattern of LD along the chromosome by calculating average D' and r^2 for markers within contiguous 1.7-Mb stretches of DNA (sliding windows). In addition, statistical models of LD decay were fitted to summarize the patterns and to account for marker density. The results highlight areas with very high levels of LD, notably at positions 11–16 Mb and 21–27 Mb of the reference sequence (Fig. 2a, b). The degree to which these exceed background LD levels was confirmed by significance testing using extreme value distribution theory applied to 'runs' of LD (see Methods). Figure 2d shows various regions where LD is slightly above background levels (light bands), as well as shorter runs of extremely high disequilibrium relative to the rest of the chromosome (dark bands). These regional differences are not due to unequal marker spacing, as they align well with the model-fitting
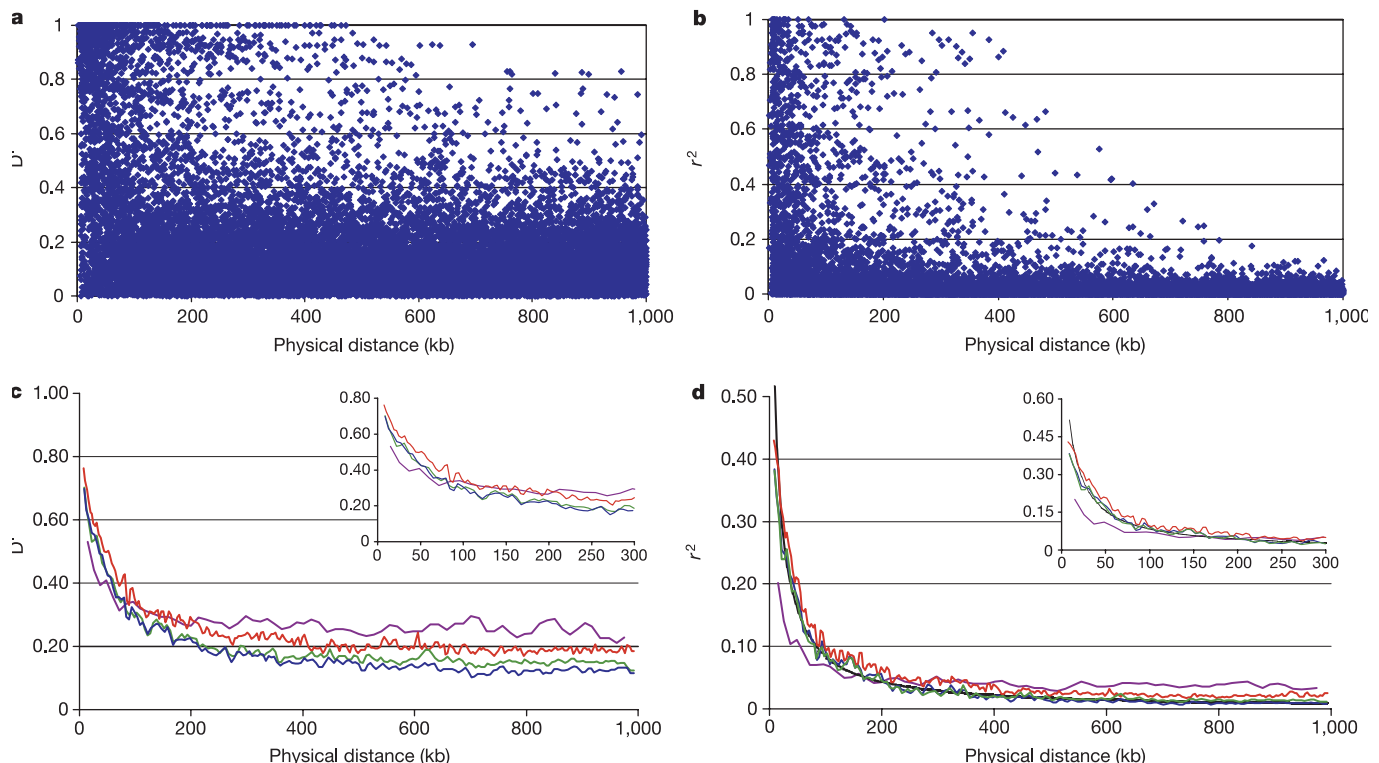

Figure 1 Distribution of linkage disequilibrium on chromosome 22. **a**, **b**, Variability of D' (**a**) and r^2 (**b**) for the CEPH samples, using all pairwise values for markers with minor allele frequency >0.20 separated by ≤ 1 Mb. **c**, **d**, Sliding window results of average D' and r^2

in successive bins of 200 markers (100 marker overlap). Insets provide enhanced views of the observed LD decay from 0–300 kb. CEPH, red; unrelated UK, green; combined CEPH/UK, dark blue; Estonia, violet

results (Fig. 2b) that account for such variability. The estimates were consistent between the CEPH and UK samples (Fig. 2b), indicating the usefulness of both family-based and unrelated samples for initial detection of long LD runs. The general patterns of LD also appear similar in the Estonian samples (Fig. 1a), but the marker density in the smaller Estonian data set (median 34.72 kb) is too coarse for formal delineation of specific regions. In the CEPH and UK samples, average D' levels in the regions of high LD are 2–5 times greater than the background levels, presenting obvious distinctions of high and low LD tracts, whereas in the Estonian data, the highest LD region is less than twice the background level. Extrapolating these results to the genomic scale suggests that a median marker density greater than one marker per 35 kb is required for any first-generation map, and that, in the absence of family data, LD
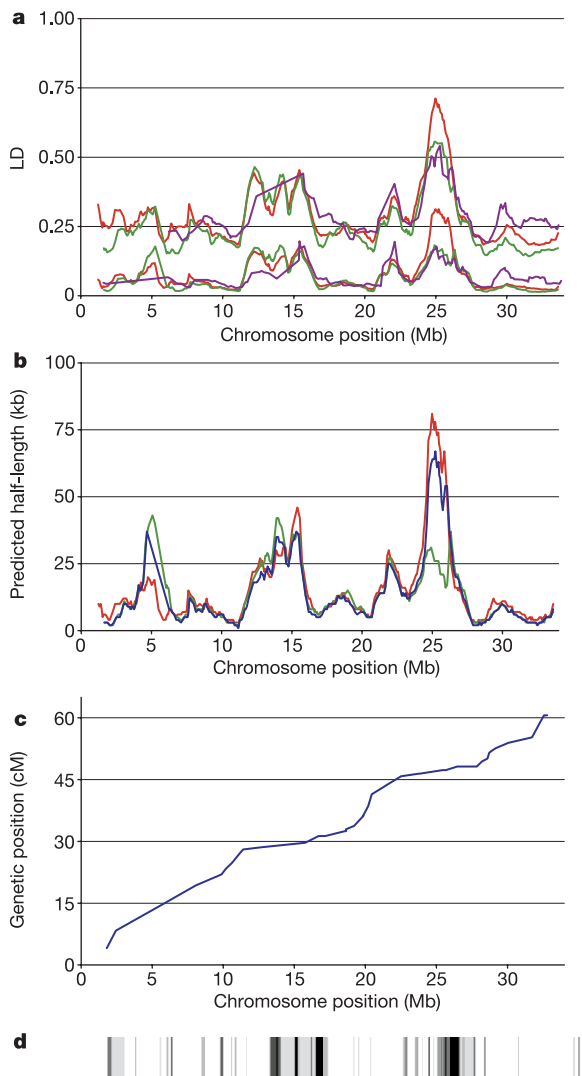
Table 1 Correlations between pairwise LD coefficients and sequence features

Sequence feature	D'	r^2	$\hat{D}'_s$ (resid.)	$\hat{D}'_r$ (resid.)
Physical distance	−0.36	−0.37	0.05*	−0.11
Genetic distance	−0.50	−0.51	−0.30	0.02*
CpG coverage	0.04*	0.04*	0.15	0.00*
Coding coverage	0.12	0.12	0.19	0.05*
Gene coverage	0.13	0.13	0.20	0.06*
Pseudogene coverage	−0.10	−0.09	−0.02*	−0.06*
SINEs	0.20	0.20	0.24	0.04*
LINEs	−0.12	−0.12	−0.09	0.00*
LTRs	−0.27	−0.26	−0.24	−0.15
DNA repeats	−0.13	−0.14	−0.07	−0.04*
Other repeats	−0.01*	−0.01*	0.11	0.05*

Correlations were Spearman rank correlations. CpG coverage, coding coverage, gene coverage and pseudogene coverage refer to the proportion of bases contained in predicted CpG islands, protein coding exons, all genic exons or pseudogenes, respectively, between each common SNP pair separated by ≤150 kb. All sequence characteristics were defined using the annotation in release 2.4 from the Sanger Centre chromosome 22 annotation group (http://www.sanger.ac.uk/HGP/Chr22/cwa_archive/Release_2.4_19-05-2000.shtml). The residual (resid.) columns represent deviates of pairwise linkage disequilibrium (LD) values from their expected values using exponential decay models[4] fitted with sequence-based (bp) separation, $\hat{D}'_s$ (resid.) or genetic distance, $\hat{D}'_r$ (resid.). All correlations are significant at $P < 0.001$ except where noted by an asterisk. *Not significant.

estimates in 100 chromosomes or fewer may be too variable for reproducible patterns at a broad scale.

There is considerable evidence that sites of recombination in humans are not randomly distributed, but are often localized into specific hotspots[7,14]. Current, low-resolution genetic maps can be used to model local recombination rates and provide additional predictors of LD beyond physical distance. Chromosome 22 has an elevated degree of recombination[15–17], averaging 2.46 cM Mb^{-1} in our interpolated sex-averaged genetic map (see Supplementary Information), compared with the genome average of approximately 1.3 cM Mb^{-1}. All components of the high LD tracts at 11–16 Mb and 21–27 Mb are situated in regions of exceptionally low recombination ($<$1 cM Mb^{-1}; Fig. 2c) relative to the chromosome average. Indeed, nearly all of the high LD runs on chromosome 22 are located in regions of low recombination (Fig. 2d), a pattern previously noted in a localized region of this chromosome[18], but which differs from a previous assessment of chromosome 22 microsatellite markers[19]. Collectively, the most exceptionally high-LD/low-recombination tracts cover about 9% of the chromosome, and 40% of the total variation in D' along chromosome 22 can be explained by the (interpolated) genetic distance between markers. These results indicate that the extant genetic maps may be used in practice as guides to genomic regions having high or low LD, and are of immediate use in position-based association studies.

To search for other predictors of LD, we examined correlations between pairwise LD measures and various sequence features (Table 1). LD is positively correlated with gene density and short interspersed nuclear elements (SINEs) such as Alu repeats (which are features of $(G + C)$-rich sequence), and negatively correlated with long interspersed nuclear elements (LINEs), long terminal repeats (LTRs) and other DNA repeats $((G + C)$-poor sequence). There is

Figure 2 Linkage disequilibrium across chromosome 22. **a**, Average D' and r^2 coefficients (top and bottom groups, respectively) plotted in sliding windows containing all common polymorphisms separated by 50 and 500 kb in successive 1.7-Mb segments (1.6-Mb overlap). Sequence position 1 refers to the centromeric q-arm origin (ftp://ftp.sanger.ac.uk/pub/human/chr22/sequences/Chr_22/complete_sequence/Chr_22_19-05-2000.fa). The colour scheme is as in Fig. 1. **b**, Expected half length estimated from application of the model $E(r^2) = 1/(1 + 4Nc)$ to each sliding window. The model could not be fitted to the Estonian data because of the sparser marker density. **c**, Relationship between genetic and physical distance on chromosome 22 (refs 15, 16). **d**, Significant regions of excess LD (see Methods). The longest runs of LD (using $f = 0.50\sigma_{D'}$) are shown in light grey; shorter runs of high LD are shown in dark grey ($f = 1.00\sigma_{D'}$) and black ($f = 1.50\sigma_{D'}$).

Figure 3 Haplotype networks on chromosome 22. The figure is oriented with the leftmost position as the centromeric q-arm origin. The top panel shows the complete set of CEPH founder haplotypes (common alleles in blue; ambiguously phased or ungenotyped regions in white). Marker locations are given in the comb diagram underneath. The middle panels show 1-Mb-wide diagonal sections of colour-coded pairwise disequilibrium matrices[30] for the CEPH families (top), UK unrelated (middle) and combined samples (bottom), respectively. Each sample is followed by a comb diagram for marker location and a pictorial representation of haplotype networks. Networks of three or four markers are in blue; longer networks are in red. The bottom panels show a detailed view of common haplotypes ($>$5%) for the combined data set. Within each block, haplotypes are listed in descending frequency. Gene composition and integrated genetic and physical maps are indicated. Transcribed sequences are grouped according to orientation[15].

also a negative correlation between the (G + C)-rich sequence features and genetic distance. When genetic distance is factored out, by assessing sequence correlations with residual LD measures that are independent of genetic distance by multiple regression or using model-based regression residuals, nearly all of the relationships are reduced or effectively eliminated. The single exception to this collinearity occurs with LTRs, which maintain a significant negative correlation with LD independently of genetic distance ($\rho = -0.15$). It is conceivable that LTRs predict recombination at a relatively fine resolution, whereas the currently available coarse genetic map predicts recombination at a broader resolution.

In addition to pairwise disequilibrium assessments, our data provide an opportunity to identify common conserved haplotypes along the chromosome. We conducted a systematic search for regions of limited haplotype diversity and strong disequilibrium in the combined CEPH and UK samples, in which we detected 97 'haplotype networks' (see Methods), each including three or more markers (Fig. 3), including 329 out of 787 (41.8%) common polymorphisms and covering approximately 9.1 Mb of the long arm of chromosome 22 (22q). Interestingly, the two most common haplotypes are complementary at all sites in 55 of these networks. In some regions the networks overlap, reflecting the difficulty in precisely locating ancestral crossover events and possibly other phenomena such as gene conversion.

All of the common variants in any one of these networks can be surveyed with minimal genotyping effort. For example, in the CEPH families, the longest haplotype network extends 804 kb (at 11.83 Mb, including 16 markers) and a single network in the 21–27 Mb region includes 25 markers and extends 758 kb. These 25 markers could define up to 32 million chromosomes and in the absence of LD every chromosome in our sample would be unique. Instead, five haplotypes account for 76% of the CEPH founder chromosomes. Genotyping three SNP markers is sufficient to distinguish these five common haplotypes and retain 94.7% of haplotype heterozygosity. Similar long conserved haplotypes are present in the UK and combined samples.

Our ability to derive common haplotypes was significantly enhanced by the use of family-based samples (the CEPH panel). This is an important conclusion for a haplotype map project, because the relative usefulness of families will depend on empirical haplotype patterns[20]. Comparison of the haplotype frequency estimates from the unrelated UK compared with partially-phased CEPH chromosomes indicated that, for an average set of five consecutive markers, each CEPH founder contributed 1.91 times more information than an unrelated individual (Methods). This suggests that although genotyping extended pedigrees such as the CEPH families requires twice as much effort as that for unrelated individuals, the information for LD mapping is approximately equal. Given this similarity and the advantages of families for detection of genotyping errors, integration with meiotic maps and quality comparison across laboratories, the use of families for LD mapping is clearly preferable to unrelated individuals.

The primary motivation for construction of any LD map in the human genome is to facilitate identification and characterization of genetic variants for common complex diseases. The present data indicate that considerable information is available for fine mapping of disease loci even in first-generation maps such as our 15-kb chromosome-wide resolution. For example, linkages to schizophrenia and other psychiatric disorders have been reported around 22q12, near marker D22S278 (ref. 21) (at approximately 19.8 Mb); schizophrenia has also been associated with microdeletions in 22q11, near the velo-cardiofacial syndrome locus ($\sim$5.6 Mb). Neither of these regions shows high LD in our data, suggesting that fine mapping may require a high density of markers. Conversely, linkage to type 2 diabetes has been reported around marker D22S423 (ref. 22) (approximately 23.8 Mb), which is on the edge of some of the longest tracts of high LD on the chromosome. Initial

allelic association in this region may be facilitated by this extensive conservation. □

Methods

Selection of markers, DNA samples and genotyping

Markers for genotyping were selected by walking along chromosome 22 in 15-kb steps through all available SNPs and small indels and choosing the nearest variant that was suitable for a unique polymerase chain reaction-based genotyping assay. Markers were genotyped in duplicate on 77 CEPH family DNAs and 90 unrelated UK Caucasian DNAs using the Third Wave Technologies Invader assay[23], and on 51 unrelated Estonian DNAs using allele specific primer extension in microarray format[24].

A final set of 1,504 markers were polymorphic in the CEPH DNA panel and were not rejected because of mendelian segregation errors, Hardy–Weinberg equilibrium deviations or other quality issues (the CEPH SNP set). There are 27 gaps of greater than 100 kb in this set (maximal gap of 293 kb), yielding a mean spacing of 22.95 kb and a median spacing of 15.07 kb. A total of 1,262 markers from the final CEPH set and 23 additional markers were successfully genotyped on the UK sample of unrelated individuals, for 1,286 markers on the UK Caucasians (the UK SNP set, median spacing of 17.53 kb, mean = 26.86 kb). We refer to the overlapping collection of CEPH and UK SNPs as the 'combined' marker set. The final Estonian SNP set had 908 SNPs, 661 of which had minor allele frequencies $\geq$0.20, and a median spacing of 34.72 kb (mean = 61.42 kb). The Estonian SNP set included 594 SNPs in common with the initial CEPH SNP set. The final data used in these analyses is available at http://www.sanger.ac.uk/HGP/Chr22.

Error checking

We tested all markers for Hardy–Weinberg equilibrium, ignoring family structure, and excluded from the analysis those where equilibrium was rejected at the 10^{-4} level. We verified familial relationships within the CEPH samples and checked that presumed unrelated UK individuals were truly unrelated using the GRR program[25]. In addition, we excluded all genotypes that produced mendelian errors or unlikely recombination patterns ($P < 0.001$) in the CEPH sample[26]. The duplicate genotyping resulted in a very low error rate, as indicated by mendelian segregation errors (480 out of 98,095 genotypes = 0.5%) and unlikely double recombinants (315 out of 98,095 genotypes = 0.3%) in the CEPH data set.

Haplotyping

For the CEPH pedigrees, we used MERLIN[26] to list all alternate sets of non-recombinant founder haplotypes including small sets of neighbouring markers. Haplotype frequencies in families were then estimated using an expectation–maximization (E–M) algorithm[27] (software available on request from G.R.A.). For unrelated individuals and the combined data set, we estimated haplotype frequencies using the E–M algorithm.

Pairwise disequilibrium and distance modelling

For pairwise comparisons we calculated D' and r^2 following standard procedures[13]. We also fitted decay models to all pairwise coefficients within successive 1-Mb sliding windows: $E(r_{ij}^2) = 1/(1 + 4Nc_{ij})$, where N is the effective population size and c_{ij} is the recombination fraction between markers i and j estimated from the physical distance between markers using the chromosome 22 average 1 Mb $\approx$ 2 cM. We refer to the half-length of disequilibrium as the distance at which $E(r_{ij}^2) = 0.5$. Estimates from this model are largely independent of the underlying marker density.

Regions of excess disequilibrium

To define boundaries for regions of unusual disequilibrium, we used a method based on the Smith–Waterman algorithm[28]. For the ith ordered marker pair within 500 kb of each other, we define the score $S_i = D'_i - k$ and then identify and compare high scoring segments using the Smith–Waterman accumulation approach and related statistical theory[29]. We used a penalty $k = \bar{D'} \times f\sigma_{D'}$ (with scale $f = 0.5, 1.0, 1.5$) to detect increasingly extreme runs of LD along the chromosome, where $\sigma_{D'}$ refers to the standard deviation of all pairwise LD coefficients.

Haplotype networks

We defined regions of limited haplotype diversity as those in which five haplotypes accounted for 75% or more of all haplotypes observed in the population and in which disequilibrium between each marker and haplotypes of surrounding markers exceed 0.75. We searched for sets of markers (networks) that met these conditions using a single marker as a seed and adding as many neighbouring markers as possible. We used MERLIN[26] to identify all non-recombinant haplotypes in a growing network and the E–M algorithm to estimate haplotype frequencies. We did not require markers in a network to be consecutive, but instead allowed up to six intervening markers to be excluded.

Comparison of families and unrelated individuals

Samples of unrelated individuals include more independent chromosomes, but less phase ambiguity exists in families. To compare the two approaches, we used the combined data set to estimate allele frequencies for each set of five consecutive markers. We then calculated the log-likelihood of each CEPH founder and unrelated individual using equilibrium allele frequencies and the haplotype frequencies estimated by E–M. The change in log-likelihood for each individual provides an indication of the amount of information contributed.

More detailed descriptions of marker selection, genotyping and error-checking

letters to nature

protocols, genotyped marker characteristics and statistical procedures are provided as Supplementary Information.

1. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22,** 139–144 (1999).
2. Taillon-Miller, P. *et al.* Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nature Genet.* **25,** 324–328 (2000).
3. Eaves, I. A. *et al.* The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. *Nature Genet.* **25,** 320–323 (2000).
4. Abecasis, G. R. *et al.* Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.* **68,** 191–197 (2001).
5. Reich, D. E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411,** 199–204 (2001).
6. Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. High-resolution haplotype structure in the human genome. *Nature Genet.* **29,** 229–232 (2001).
7. Jeffreys, A. J., Kauppi, L. & Neumann, R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genet.* **29,** 217–222 (2001).
8. Patil, N. *et al.* Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294,** 1719–1723 (2001).
9. Weissenbach, J. *et al.* A second-generation linkage map of the human genome. *Nature* **359,** 794–801 (1992).
10. Dawson, E. *et al.* A SNP resource for human chromosome 22: extracting dense clusters of SNPs from the genomic sequence. *Genome Res.* **11,** 170–178 (2001).
11. Mullikin, J. C. *et al.* An SNP map of human chromosome 22. *Nature* **407,** 516–520 (2000).
12. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409,** 928–933 (2001).
13. Weir, B. S. *Genetic Data Analysis II* (Sinauer Associates, Sunderland, Massachusetts, 1996).
14. Petes, T. D. Meiotic recombination hot spots and cold spots. *Nature Rev. Genet.* **2,** 360–369 (2001).
15. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402,** 489–495 (1999).
16. Yu, A. *et al.* Comparison of human genetic and sequence-based physical maps. *Nature* **409,** 951–953 (2001).
17. Payseur, B. A. & Nachman, M. W. Microsatellite variation and recombination rate in the human genome. *Genetics* **156,** 1285–1298 (2000).
18. Eisenbarth, I., Striebel, A. M., Moschgath, E., Vogel, W. & Assum, G. Long-range sequence composition mirrors linkage disequilibrium pattern in a 1.13 Mb region of human chromosome 22. *Hum. Mol. Genet.* **10,** 2833–2839 (2001).
19. Huttley, G. A., Smith, M. W., Carrington, M. & O'Brien, S. J. A scan for linkage disequilibrium across the human genome. *Genetics* **152,** 1711–1722 (1999).
20. Thompson, E. A., Deeb, S., Walker, D. & Motulsky, A. G. The detection of linkage disequilibrium between closely linked markers: RFLPs at the AI-CIII apolipoprotein genes. *Am. J. Hum. Genet.* **42,** 113–124 (1988).
21. Pulver, A. E. *et al.* Sequential strategy to identify a susceptibility gene for schizophrenia: report of potential linkage on chromosome 22q12-q13.1: Part 1. *Am. J. Med. Genet.* **54,** 36–43 (1994).
22. Ghosh, S. *et al.* The Finland–United States investigation of non-insulin-dependent diabetes mellitus genetics (FUSION) study. I. An autosomal genome scan for genes that predispose to type 2 diabetes. *Am. J. Hum. Genet.* **67,** 1174–1185 (2000).
23. Mein, C. A. *et al.* Evaluation of single nucleotide polymorphism typing with invader on PCR amplicons and its automation. *Genome Res.* **10,** 330–343 (2000).
24. Kurg, A. *et al.* Arrayed primer extension: solid-phase four-colour DNA resequencing and mutation detection technology. *Genet. Test* **4,** 1–7 (2000).
25. Abecasis, G. R., Cherny, S. S., Cookson, W. O. & Cardon, L. R. GRR: graphical representation of relationship errors. *Bioinformatics* **17,** 742–743 (2001).
26. Abecasis, G. R., Cherny, S. S., Cookson, W. O. & Cardon, L. R. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genet.* **30,** 97–101 (2001).
27. Weir, B. S. & Cockerham, C. C. Estimation of linkage disequilibrium in randomly mating populations. *Heredity* **42,** 105–111 (1979).
28. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147,** 195–197 (1981).
29. Karlin, S. & Altschul, S. F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA* **87,** 2264–2268 (1990).
30. Abecasis, G. R. & Cookson, W. O. GOLD—graphical overview of linkage disequilibrium. *Bioinformatics* **16,** 182–183 (2000).

Competing interests statement

The authors declare competing financial interests: details accompany the paper on *Nature*'s website (http://www.nature.com/nature).

Correspondence and requests for materials should be addressed to I.D. (e-mail: id1@sanger.ac.uk) or L.C. (e-mail: lon.cardon@well.ox.ac.uk).

Golgi biogenesis in *Toxoplasma gondii*

Laurence Pelletier*, **Charlene A. Stern**†, **Marc Pypaert***, **David Sheff***, **Huân M. Ngô**‡, **Nitin Roper***, **Cynthia Y. He***, **Ke Hu**§, **Derek Toomre***, **Isabelle Coppens**‡, **David S. Roos**§, **Keith A. Joiner**‡ & **Graham Warren***

* *Department of Cell Biology, Ludwig Institute for Cancer Research, Yale University School of Medicine, New Haven, Connecticut 06520, USA*
† *Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School and The Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA*
‡ *Section of Infectious Diseases, Department of Internal Medicine, Yale University School of Medicine, New Haven, Connecticut 06520, USA*
§ *Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA*

....

Two models have been put forward to explain the growth of new Golgi during the cell cycle. The first suggests that a new Golgi grows out of the endoplasmic reticulum by *de novo* synthesis[1]. The second suggests that a pre-existing Golgi is needed for the growth of a new one, that is, the Golgi is an autonomously replicating organelle[2]. To resolve this issue, we have exploited the simplicity of the apicomplexan parasite *Toxoplasma gondii*[3], which has only a single Golgi stack[4]. Here we show, by using video fluorescence microscopy and three-dimensional reconstructions of serial thin sections, that the Golgi grows by a process of lateral extension followed by medial fission. Further fission leads to the inheritance by each daughter of a pair of Golgi structures, which then coalesce to re-form a single Golgi. Our results indicate that new Golgi grow by autonomous duplication and raise the possibility that the Golgi is a paired structure that is analogous to centrioles[5].

The doubling of cell mass before division is accompanied by the duplication of cellular organelles. In the case of chromosomes and centrosomes, duplication is manifested as a doubling in number[5,6]. In the case of membrane-bound organelles such as the endoplasmic reticulum (ER) and mitochondria, it is generally manifested as a doubling of mass because the number of these organelles varies, as does their unit size[7]. The Golgi apparatus may be an exception because it has a nearly constant unit size in all eukaryotes, comprising flattened cisternal membranes, typically 0.5–1.0 μm in cross-sectional diameter, that are most often arranged in a stack[8,9]. Golgi duplication therefore results in a doubling of both number and mass.

Two different models have been proposed to explain the growth of new Golgi. The first suggests that new Golgi are assembled *de novo* from components that originate from the ER. The reversible absorption of the Golgi by the ER during mitosis or treatment with Brefeldin A (BFA) has been taken as evidence in favour of this view[10,11]. The fact that cytoplasts lacking the Golgi cannot make a new one from the remaining ER is evidence against it[12]. The second model suggests that new Golgi arise from pre-existing Golgi which either act as templates for assembly[2] or, by analogy with centrosomes[5], instruct the assembly process. Morphological studies have provided images of Golgi, in organisms ranging from earthworms to plants, that are in the apparent act of medial fission (see ref. 7), but static images do not indicate the direction of events. Golgi in the act of separation could just as easily represent two Golgi in the act of fusing.

A significant problem in studying Golgi duplication is the number of Golgi in many cells. Mammalian cells typically contain 100–250 Golgi units, which are most often stitched together in a ribbon-like structure located near the cell nucleus[13,14]. Even fungal cells, which are several orders of magnitude smaller in volume,