Supplementary information is available on *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We thank J. Finklestein, N. Atigapramoj, E. Dabagyan, S. Huang, A. Ambriz, A. Harxhi and K. Sutton for their contributions to this work, which was supported by NIH and DOE.

Correspondence and requests for materials should be addressed to H.C.R. (e-mail: Riethman@wistar.upenn.edu).

Comparison of human genetic and sequence-based physical maps

Adong Yu*, Chengfeng Zhao*, Ying Fan*, Wonhee Jang†, Andrew J. Mungall‡, Panos Deloukas‡, Anne Olsen§, Norman A. Doggett‖, Nader Ghebranious*, Karl W. Broman¶ & James L. Weber*

** Center for Medical Genetics, Marshfield Medical Research Foundation, Marshfield, Wisconsin 54449, USA*
† National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland 20894, USA
‡ The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK
§ Joint Genome Institute, Walnut Creek, California 94598, USA
‖ Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA
¶ Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland 21205-2179, USA

Recombination is the exchange of information between two homologous chromosomes during meiosis. The rate of recombination per nucleotide, which profoundly affects the evolution of chromosomal segments, is calculated by comparing genetic and physical maps. Human physical maps have been constructed using cytogenetics[1], overlapping DNA clones[2] and radiation hybrids[3]; but the ultimate and by far the most accurate physical map is the actual nucleotide sequence. The completion of the draft human genomic sequence[4] provides us with the best opportunity yet to compare the genetic and physical maps. Here we describe our estimates of female, male and sex-average recombination rates for about 60% of the genome. Recombination rates varied greatly along each chromosome, from 0 to at least 9 centiMorgans per megabase (cM Mb^{-1}). Among several sequence and marker parameters tested, only relative marker position along the metacentric chromosomes in males correlated strongly with recombination rate. We identified several chromosomal regions up to 6 Mb in length with particularly low (deserts) or high (jungles) recombination rates. Linkage disequilibrium was much more common and extended for greater distances in the deserts than in the jungles.

All nucleated human cells contain two homologous copies of each chromosome, except for the sex chromosomes in males. During the formation of the sperm and egg cells, the number of each chromosome is reduced to one so that fertilization restores the normal diploid number in the next generation. The process of chromosome reduction, meiosis, is usually accompanied by exchange or recombination of DNA between the homologous parental chromosomes. Genetic maps, which are based on meiotic recombination, order and estimate distances between DNA sequences that vary between parental homologues (polymorphisms). The primary unit of distance along the genetic maps is the centiMorgan (cM), which is equivalent to 1% recombination.

The genetic maps used in our analysis were based upon the genotyping of 8,031 short tandem repeat polymorphisms (STRPs) from Généthon, the University of Utah and the Cooperative Human Linkage Center in eight reference CEPH families[5]. Excluding the sex chromosomes, the maps cover about 4,250 cM in females and 2,730 cM in males. The genetic maps are relatively marker dense, with an average of 2–3 STRPs per cM, but are also relatively low resolution because only 184 meioses (92 in each sex) were analysed.

The physical maps used were all DNA sequence assemblies. For chromosomes 21 and 22, we used the finished, published sequences[6,7]. For the other 20 autosomes and for the X chromosome, we used the public draft sequence assemblies, 5 September 2000 version (http://genome.cse.ucsc.edu)[4]. As we required relatively long stretches of sequence, we used only sequence assemblies that were over 1.5 Mb long (between terminal STRPs), contained more than three STRPs and had a marker order that agreed with published genetic and radiation hybrid maps. The amount of sequence used from each chromosome is shown in Fig. 1. Some chromosomes had much better coverage than others. We analysed 253 sequence assemblies ranging in length up to 70 Mb and spanning a total of 1,806 Mb (roughly 58% of the portion of the genome that is not highly repetitive). By far the most common reason for rejecting sequence assemblies was insufficient length; only seven assemblies were rejected for incompatible marker order.

Recombination rates varied greatly across the genome, from 0 to 8.8 cM Mb^{-1} (Table 1). Sex-average recombination rates (the average for males and females combined) did not vary as much as the sex-specific rates (for males and females considered separately) because male and female recombination rates at specific sites often differed substantially. We identified 19 recombination deserts up to 5 Mb in length with sex-average recombination rates below 0.3 cM Mb^{-1}, and 12 recombination jungles up to 6 Mb in length with sex-average recombination rates greater than 3.0 cM Mb^{-1} (see Supplementary Information). Wide variation in recombination rates across chromosomes has been observed previously for humans[8–11] and for other eukaryotic species[12–15], and is clearly the rule rather than the exception.

In an effort to identify the basis of differences in recombination rates, we compared the rates to several marker and sequence parameters. These parameters included GC content, STRP informativeness, position of the marker relative to the centromeres and telomeres, density of runs of various short tandem repeats, especially $(A)_n$, $(AC)_n$, $(AGAT)_n$, $(AAN)_n$ and $(AAAN)_n$ sequences, and the density of various interspersed repetitive elements, including

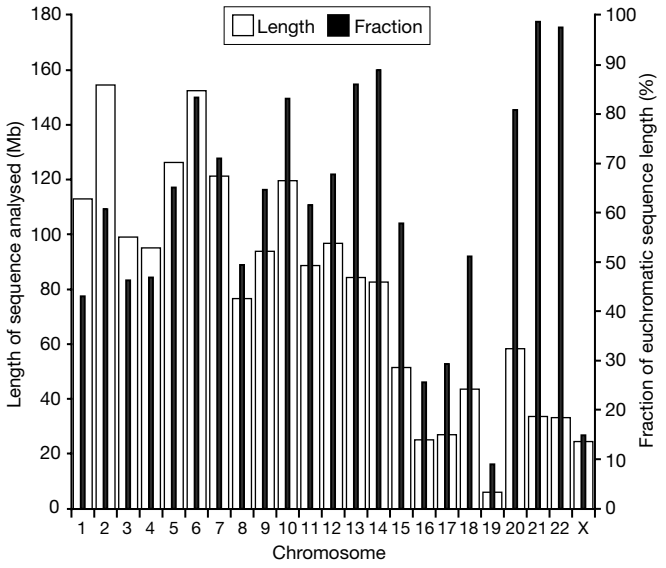

Figure 1 Sequence coverage for comparison of the genetic and physical maps. The total length of sequence used in the analysis (open bars) and the approximate percentage of the euchromatic sequence length (solid bars) are shown for each chromosome.

Table 1 Distribution of recombination rates

	Range		Mean ± s.d.	
Sex average	0.0–6.0		1.30 ± 0.80	
Male	0.0–7.9		0.92 ± 0.96	
Female	0.0–8.8		1.68 ± 1.07	
	0–0.5	0.5–1.5	1.5–3.0	>3.0
Sex average	12%	58%	26%	4%
Male	38%	45%	12%	5%
Female	10%	38%	42%	10%

Recombination rates given in cM Mb^{-1}. These data were derived from individual recombination rates at 4,088 STRPs.

Alu, L1, MIR (mammalian-wide interspersed repeats) and MER (medium reiterated repeats) sequences (Table 2).

With one exception, we found only weak correlations between the parameters and recombination rates. As controls, we found the expected negative correlation between short interspersed nuclear element (SINE) and long interspersed nuclear element (LINE) densities[16], and the positive correlation between SINE density and $(A)_n$ density. In agreement with ref. 17, we found no correlation between STRP heterozygosity and recombination rate, despite reports of positive correlation of nucleotide (sequence) diversity values with recombination rates[18,19]. However, STRP heterozygosities are probably much more dependent upon relatively high mutation rates than selection and are therefore likely to be poor measures of nucleotide diversity. Similarly, we found only weak correlation between $(AC)_n$ density ($n \geq 11$ or 19) and recombination rates despite the report of such a correlation for chromosome 22 (ref. 20). GC content is of interest because the genome appears to be segmented into isochores of varying GC content and because GC content is strongly correlated with gene density[21]. We did confirm a positive relationship between recombination and GC content (ref. 22), but the correlation was weak. By far the strongest relationship detected was for the position of the markers along the metacentric chromosome arms in males. Male (but not female) recombination rates increased markedly near the telomeres.

Some important limitations apply to our comparison of human genetic and physical maps. First, the resolution of the genetic maps is modest, owing to the small number of meioses examined. This places relatively broad confidence intervals on the genetic map distances and similar broad confidence intervals on the recombination rates. Only sex-average recombination rates smaller than about 0.3 cM Mb^{-1} and greater than about 2.5 cM Mb^{-1} are statistically different from uniform recombination at the $P = 0.05$ significance level. Second, the draft sequences used in our analysis were often short, contained many gaps and still had some errors in marker order. When the finished sequences become available, additional recombination deserts and jungles, for example, will undoubtedly be discovered. Third, there is mounting evidence for at least modest individual and possibly population variation in recombination rates[5,23,24]. The genetic maps in our analysis were based on meiosis in only eight mothers and eight fathers, all or nearly all of European ancestry. Examination of a large sample of individuals and/or other populations might give different results. Finally, our analysis is only a long-range (megabase) analysis. We can reach no conclusions about recombination over short (kilobase) ranges. There is growing evidence for recombination hot spots no more than a few kilobases long[13,25,26]. Megabase-sized chromosomal segments may turn out to be comprised of regions with little or no recombination separated by short recombination hot spots. Perhaps the primary difference between recombination deserts and jungles lies in the density and strength of recombination hot spots.

Despite the limitations, there is strong evidence that our results are reliable first estimates of human recombination rates. Genetic maps based on 40 CEPH families show good agreement with the eight family maps (see, for example, ref. 8). Plots of the ratio of female to male recombination from the eight family data show maxima at the centromeres and minima at the telomeres for virtually all metacentric chromosomes[5]. The shapes produced by plotting centiMorgans against megabases obtained from the draft sequence assemblies for chromosomes 6 and 20 match closely those obtained using physical distances from restriction enzyme fingerprinting of overlapping genomic clones. Lengths of the draft sequence assemblies (17 July 2000 version) for chromosomes 21 and 22 matched the lengths of the finished sequences with only 0.1% error. And, probably most importantly, recombination deserts and jungles differ significantly in linkage disequilibrium (when two polymorphic alleles are not in random association).

The decay of linkage disequilibrium is expected to be much slower in recombination-poor than in recombination-rich regions. We tested this hypothesis by comparing linkage disequilibrium among pairs of STRPs within the recombination deserts and jungles. Although the power to detect linkage disequilibrium in genotyping data from only eight families is low, it was still found to be much higher for close pairs of markers in the deserts than in the jungles (Fig. 2). For marker pairs less than 0.5 Mb apart, 32% of pairs in the deserts showed significant linkage disequilibrium, as compared with only 7% in the jungles ($P = 0.001$).

In conclusion, our work shows that recombination rates vary greatly across the human genome, by at least two orders of magnitude. Linkage disequilibrium will generally extend over longer distances in regions with low recombination. Mapping genes responsible for traits and diseases by association studies will be easier and require a lower density of polymorphisms in regions of

Table 2 Correlations of recombination rates with sequence parameters

Parameter 1	Parameter 2	Relationship	R^2	P
SINE density	LINE density	Negative	0.20	<0.001
SINE density	$(A)_n$ density	Positive	0.80	<0.001
Sex average recombination rate	STRP heterozygosity	None	0.00	0.49
Sex average recombination rate	GC content	Positive	0.05	<0.001
Sex average recombination rate	$(AC)_n$ density ($n \geq 11$)	Positive	0.001	0.03
Sex average recombination rate	$(A)_n$ density	Positive	0.02	<0.001
Sex average recombination rate	SINE density	Positive	0.01	<0.001
Sex average recombination rate	LINE density	Negative	0.02	<0.001
Male recombination rate	Chromosome position (metacentrics)	Positive towards telomeres	0.18	<0.001

Correlation was evaluated by linear regression analysis. Relationship indicates the sign of the slope of the best fitting line. R^2 is a measure of the fit of the points to the line. R^2 can vary from 0 to 1.0 with 1.0 indicating a perfect fit. P is the probability of obtaining an R^2 value as large as observed given no correlation. For the original plots from which these parameters were derived, see Supplementary Information.

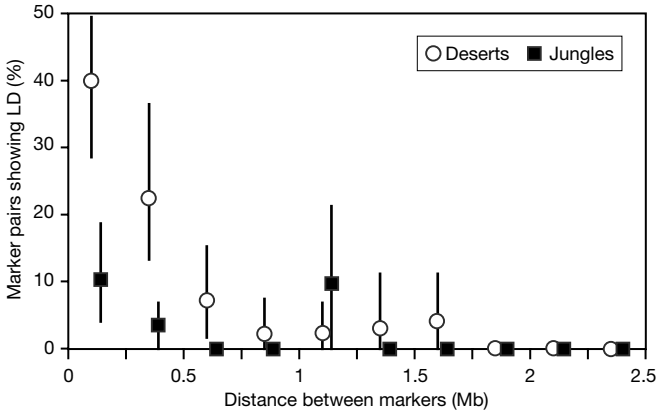

Figure 2 Linkage disequilibrium (LD) among pairs of STRPs within autosomal recombination deserts and jungles. Deserts and jungles are listed in the Supplementary Information. Marker pairs were binned into 0.25-Mb spacing intervals.

low recombination. Nucleotide and haplotype diversity will also probably parallel recombination rates. Although our baseline long-range recombination rates will be useful, they should be recalculated when the human genomic sequences are finished and as higher resolution genetic maps become available. In the more distant future, genotyping greater numbers of reference families at much higher polymorphism densities will lead to short-range maps of recombination hot spots. □

Methods

Connection of genetic and physical maps

We used short, single-pass genomic sequences and/or PCR primer sequences for STRPs to identify draft or finished bacterial artificial chromosome (BAC) or cosmid sequences within GenBank that encompass the STRPs using BLAST[27] and ePCR[28]. Blast criteria were score (bits) > 200, expect (E) value $< e^{-50}$, and ratio of matched bases to marker sequence length $> 85\%$. ePCR criteria were no more than one base mismatch in each primer and size of PCR product within allele size range for the STRP. About 75% of the STRPs were connected to the long genomic sequences. The reasons for failure of the remaining 25% are not fully understood, but include absence of the corresponding sequence in GenBank and poor quality of the STRP sequences. As the genetic maps are marker rich, the absence of 25% was not a serious limitation. Tables of STRPs with GenBank sequence accession numbers for encompassing BACs, genetic map positions and recombination rates are available from the Marshfield web site.

Determination of recombination rates

For each sequence assembly we built new female, male and sex-average genetic maps, using the marker order provided by the assemblies and using the genotyping data from the eight CEPH reference families[5]. We fitted cubic splines to plots of genetic versus physical distance, and from these curves we obtained recombination rates as first derivatives[15]. The statistical significance of the recombination rates was estimated by computer simulation of 1,000 iterations of recombination within each interval between markers, assuming a constant level of recombination across the genome for each sex. The constant levels of recombination were taken as the total genetic lengths of all the assemblies analysed divided by the total physical lengths of these assemblies.

Computation of marker and sequence parameters

We calculated STRP heterozygosities using genotypes of individuals within the eight CEPH families. We obtained STRP positions relative to centromeres and telomeres as the fractional sex-average genetic map distances from the centromeres to the telomeres (value of 0 for a STRP at the centromere and 1.0 for a STRP at the telomere)[5]. GC content and STR densities were obtained from programs written and tested at Marshfield[29]. STR densities were measured as numbers of runs of non-interrupted repeats rather than total numbers of repeats. Minimum values of n for $(A)_n$, $(AC)_n$, $(AGAT)_n$, $(AAN)_n$ and $(AAAN)_n$ sequences were 12, 11 or 19 ($(AC)_n$), 5, 7 and 5, respectively. We obtained interspersed repetitive element densities using the program Repeat Masker (http://ftp.genome.washington.edu/RM/RepeatMasker.html). SINEs and LINEs were defined by Repeat Masker and consist primarily of Alu and L1 elements, respectively. We computed all DNA sequence parameters over 250-kb windows centred about each STRP. For markers ≤ 125 kb from the ends of the sequence assemblies, we defined the window as the 125 kb of proximal sequence plus all available distal sequence. Unknown bases in the sequence assemblies were excluded from analysis. All parameters were corrected for reduced window size owing to unknown bases or proximity to ends.

Measurement of linkage disequilibrium

Recombination deserts and jungles were selected as those chromosomal regions with sex-average recombination rates of <0.3 or >3.0, respectively. We measured linkage disequilibrium for all pairs of STRPs within the deserts (449 pairs) and jungles (467 pairs) using Fisher's exact test[30]. Only disequilibrium results that were significant at $P \leq 0.01$ were plotted in Fig. 2. An overall P-value was obtained by a permutation test treating the regions as units in order to account for the dependence between marker pairs within a region.

Received 27 October; accepted 8 December 2000.

1. The BAC Resource Consortium. Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* **409,** 953–958 (2001).
2. The International Human Genome Mapping Consortium. A physical map of the human genome. *Nature* **409,** 934–941 (2001).
3. Deloukas, P. *et al.* A physical map of 30,000 human genes. *Science* **282,** 744–746 (1998).
4. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
5. Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L. & Weber, J. L. Comprehensive human genetic maps: Individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63,** 861–869 (1998).
6. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402,** 489–495 (1999).
7. Hattori, M. *et al.* The DNA sequence of human chromosome 21. *Nature* **405,** 311–319 (2000).
8. Fain, P. R., Kort, E. N., Yousry, C., James, M. R. & Litt, M. A high resolution CEPH crossover mapping panel and integrated map of chromosome 11. *Hum. Mol. Genet.* **5,** 1631–1636 (1996).
9. Bouffard, G. G. *et al.* A physical map of human chromosome 7: An integrated YAC contig map with average STS spacing of 79 kb. *Genome Res.* **7,** 673–692 (1997).
10. Nagaraja, R. *et al.* X chromosome map at 75-kb STS resolution, revealing extremes of recombination and GC content. *Genome Res.* **7,** 210–222 (1997).
11. Mohrenweiser, H. W., Tsujimoto, S., Gordon, L. & Olsen, A. S. Regions of sex-specific hypo- and hyper-recombination identified through integration of 180 genetic markers into the metric physical map of human chromosome 19. *Genomics* **47,** 153–162 (1998).
12. Nicolas, A. Relationship between transcription and initiation of meiotic recombination: toward chromatin accessibility. *Proc. Natl Acad. Sci. USA* **95,** 87–89 (1998).
13. Wahls, W. P. Meiotic recombination hotspots: shaping the genome and insights into hypervariable minisatellite DNA change. *Curr. Top. Dev. Biol.* **37,** 37–75, (1998).
14. Faris, J. D., Haen, K. M. & Gill, B. S. Saturation mapping of a gene-rich recombination hot spot region in wheat. *Genetics* **154,** 823–835 (2000).
15. Kliman, R. M. & Hey, J. Reduced natural selection associated with low recombination in *Drosophila melanogaster*. *Mol. Biol. Evol.* **10,** 1239–1258 (1993).
16. Chen, T. L. & Manuelidis, L. SINEs and LINEs cluster in distinct DNA fragments of Giemsa band size. *Chromosoma* **98,** 309–316 (1989).
17. Payseur, B. A. & Nachman, M. W. Microsatellite variation and recombination rate in the human genome. *Genetics* **156,** 1285–1298 (2000).
18. Nachman, M. W., Bauer, V. L., Crowell, S. L. & Aquadro, C. F. DNA variability and recombination rates at X-linked loci in humans. *Genetics* **150,** 1133–1141 (1998).
19. Nachman, M. W. & Crowell, S. L. Contrasting evolutionary histories of two introns of the Duchenne muscular dystrophy gene, Dmd, in humans. *Genetics* **155,** 1855–1864 (2000).
20. Majewski, J. & Ott, J. GT repeats are associated with recombination on human chromosome 22. *Genome Res.* **10,** 1108–1114 (2000).
21. Bernardi, G. Isochores and the evolutionary genomics of vertebrates. *Gene* **241,** 3–17 (2000).
22. Eisenbarth, I., Vogel, G., Krone, W., Vogel, W. & Assum, G. An isochore transition in the NF1 gene region coincides with a switch in the extent of linkage disequilibrium. *Am. J. Hum. Genet.* **67,** 873–880 (2000).
23. Yu, J. *et al.* Individual variation in recombination among human males. *Am. J. Hum. Genet.* **59,** 1186–1192 (1996).
24. Lien, S., Szyda, J., Schechinger, B., Rappold, G. & Arnheim, N. Evidence for heterogeneity in recombination in the human pseudoautosomal region: High resolution analysis by sperm typing and radiation-hybrid mapping. *Am. J. Hum. Genet.* **66,** 557–566 (2000).
25. Carrington, M. Recombination within the human MHC. *Immunol Rev.* **167,** 245–256 (1999).
26. Jeffreys, A. J., Ritchie, A. & Neumann, R. High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hot spot. *Hum. Mol. Genet.* **9,** 725–733 (2000).
27. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389 (1997).
28. Schuler, G. D. Sequence mapping by electronic PCR. *Genome Res.* **7,** 541–550 (1997).
29. Zhao, C., Heil, J., & Weber, J. L. A genome wide portrait of short tandem repeats. *Am. J. Hum. Genet.* **65,** (Suppl.) A102 (1999).
30. 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).

Supplementary information is available from *Nature's* World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*, and from the Marshfield Web site (http://research.marshfieldclinic.org/genetics).

Acknowledgements

This work was supported by contracts from the US National Institutes of Health and Department of Energy. Assistance was also provided by the chromosome 6 and 20 project groups at the Sanger Centre, supported by the Wellcome Trust.

Correspondence and requests for materials should be addressed to J.L.W. (e-mail: weberj@cmg.mfldclin.edu).

Integration of cytogenetic landmarks into the draft sequence of the human genome

The BAC Resource Consortium*

* *Authorship of this paper should be cited using the names of authors that appear at the end.*

We have placed 7,600 cytogenetically defined landmarks on the draft sequence of the human genome to help with the characterization of genes altered by gross chromosomal aberrations that cause human disease. The landmarks are large-insert clones mapped to chromosome bands by fluorescence *in situ* hybridization. Each clone contains a sequence tag that is positioned on the genomic sequence. This genome-wide set of sequence-anchored clones allows structural and functional analyses of the genome. This resource represents the first comprehensive integration of cytogenetic, radiation hybrid, linkage and sequence maps of the