A high-resolution recombination map of the human genome

Augustine Kong, Daniel F. Gudbjartsson, Jesus Sainz, Gudrun M. Jonsdottir, Sigurjon A. Gudjonsson, Bjorgvin Richardsson, Sigrun Sigurdardottir, John Barnard, Bjorn Hallbeck, Gisli Masson, Adam Shlien, Stefan T. Palsson, Michael L. Frigge, Thorgeir E. Thorgeirsson, Jeffrey R. Gulcher & Kari Stefansson

Published online: 10 June 2002, doi:10.1038/ng917

Determination of recombination rates across the human genome has been constrained by the limited resolution and accuracy of existing genetic maps and the draft genome sequence. We have genotyped 5,136 microsatellite markers for 146 families, with a total of 1,257 meiotic events, to build a high-resolution genetic map meant to: (i) improve the genetic order of polymorphic markers; (ii) improve the precision of estimates of genetic distances; (iii) correct portions of the sequence assembly and SNP map of the human genome; and (iv) build a map of recombination rates. Recombination rates are significantly correlated with both cytogenetic structures (staining intensity of G bands) and sequence (GC content, CpG motifs and poly(A)/poly(T) stretches). Maternal and paternal chromosomes show many differences in locations of recombination maxima. We detected systematic differences in recombination rates between mothers and between gametes from the same mother, suggesting that there is some underlying component determined by both genetic and environmental factors that affects maternal recombination rates.

Introduction

The draft sequence of the human genome[1] has markedly advanced the understanding of human genetics. Because the available sequence is that of a reference genome, however, it does not provide insight into the genomic variability that is responsible for much of human diversity. Along with mutation, a major mechanism generating variability in the eukaryotic genome is intergenerational mixing of DNA through meiotic recombination of homologous chromosomes. The standard approach to studying rates of recombination across the genome is to build a genetic map by genotyping, with a high density of markers, a large number of individuals in families and then match this to the corresponding physical map.

Existing genetic maps[2–4] have been used extensively in linkage analysis in the mapping of disease genes and the assembly of human DNA sequences. One limitation of present genetic maps is the low resolution inherent to modest sample size. The Marshfield map[4], considered the current standard by most scientists, is based on only 188 meioses. This affects the accuracy of the estimates of recombination probabilities, and for markers separated by no more 3 cM, makes even the marker order somewhat unreliable. We collected a substantially larger set of genetic mapping data that provides information on 1,257 meioses. The draft sequence of the human genome facilitates the construction of a high-resolution genetic map by clarifying the order of the markers where the genetic data lack resolution; the genetic data, in turn, can be used to check and improve the sequence assembly. In addition, a higher-resolution genetic map and an accurate physical map together provide better estimates of recombination rates with respect to physical distances, which are essential to understanding the intergenerational variability of the genome. Thus, our results should facilitate the formulation and testing of hypotheses about the relationships between sequence content and recombination rate and between recombination rate and the degree of linkage disequilibrium.

Results

Data collection

We genotyped 869 individuals in 146 Icelandic families, consisting of 149 sibships and providing information on 628 male/paternal and 629 female/maternal meioses, with 5,136 microsatellite markers[2,4–8]. Both parents of 95 sibships were genotyped and for 52, a single parent was genotyped (see Web Tables A and B online for more details of the families and the microsatellite markers used). As compared to the eight large sibships on which the Marshfield map is based, most with grandparents genotyped, the information on recombinations in our data set is slightly less complete. With more than six times the number of meioses, the average resolution of our map is probably about five times that of the Marshfield map.

Of the 5,136 markers, 4,690 (91.3%) were placed in sequence contigs of the August 2001 freeze (released in October 2001) of the Human Genome Project Working Draft at the University of California, Santa Cruz[1]. We placed another 82 (1.6%) markers through our own *in silico* analysis of the public sequence. The remaining 364 markers, or 7.1%, could not be located in the current public sequence.

deCODE genetics, Sturlugotu 8, IS-101 Reykjavik, Iceland. Correspondence should be addressed to A.K. (e-mail: kong@decode.is) or K.S. (e-mail: kstefans@decode.is).

The entire set of genotype data, coded for anonymity to protect privacy, is available to investigators with a valid research plan.

Determination of marker order

The Marshfield map and all previous genetic maps were constructed without the benefit of the draft sequence as a reference. Because of the low resolution of the data, simply determining the order of the markers was a substantial undertaking. Our higher resolution, resulting from the large sample size, and the availability of the draft sequence made our task easier. The correct ordering of the markers was still not straightforward, however, as there are discrepancies between the draft sequence and other genetic and physical mapping data[9]. We ordered the markers using our genetic data, and used the draft sequence as a default when our data lacked resolution. We used our genetic data to resolve cases where there was more than one hit for a particular marker in the draft sequence from BLAT or ePCR (Web Note A online). Also, where our genetic data provide strong support for a marker order different from that of the draft sequence, we modified the physical locations of the markers along with the corresponding sequence. We made some additional changes to the draft sequence using other physical mapping data (see Methods and Web Note B online). In total, we made 104 modifications to the August 2001 UCSC sequence assembly, amounting to about 3.4% of the genome, affecting 84 of 543 contigs (15% of contigs) and representing 40% of the genome sequence (Web Tables C and D online). For ordering markers, the average resolution of our marker map is about 0.5 cM. (Web Table E online gives the physical positions of our markers and incorporates these modifications.)

Genetic distances

We used an extended version of our multipoint linkage program, Allegro[10], to estimate the genetic distances between consecutive markers on our corrected marker map in males and females by applying the method of maximum likelihood and the expectation-maximization (EM) algorithm[11]. (Web Table E online contains the resulting sex-averaged and sex-specific genetic maps.)

Aside from sampling errors, genotyping errors not causing inheritance incompatibilities can inflate the genetic distances substantially, as one genotyping error can lead to one or more false double recombinations. We examined each genotype using the extended version of Allegro and identified 2,123 problematic genotypes. The removal of each one led to a reduction of two or more obligate crossovers. We eliminated these 2,123 genotypes from the final estimation of genetic distances and thereby reduced the estimated genetic length of the genome by about 11%. Although most of these genotypes probably reflected genotyping errors, some may represent mutations. Also, apparent multiple recombinations could result from gene conversions or DNA rearrangements; a common inversion of an approximately 3 Mb region of 8p was recently identified from the CEPH genetic data and later confirmed by FISH[12]. To ensure that data that could lead to similar interesting discoveries remain available, we have included these 2,123 problematic genotypes (flagged as such) in the data distribution.

Our estimate of the total genetic length of the genome (the 22 autosomal chromosomes and the X chromosome) spanned by our markers is 3,615 cM—not significantly different from the estimate of 3,567 cM indicated by the Marshfield map (Table 1). Notably, however, the length of chromosome 1 (the longest chromosome) indicated by our map is 13.8 cM (4.9%) less than that indicated by the Marshfield map. For the two shortest chromosomes (chromosomes 21 and 22), however, our lengths are

Table 1 • Physical and genetic lengths of individual chromosomes

Chromosome	Physical length (Mb)	Marshfield sex-averaged Genetic length (cM)	Genetic length according to this study (cM)			Recombination rate (cM Mb^{-1})		Number of markers
			Sex averaged	Male	Female	Sex averaged	Excluding centromere	
1	282.61	284.07	270.27	195.12	345.41	0.96	1.08	468
2	252.48	261.61	257.48	189.55	325.41	1.02	1.05	407
3	224.54	219.34	218.17	160.71	275.64	0.97	0.99	369
4	205.35	206.59	202.80	146.54	259.06	0.99	1.00	302
5	199.24	197.54	205.69	151.20	260.19	1.03	1.06	334
6	190.87	189.00	189.60	137.62	241.59	0.99	1.03	293
7	168.50	178.84	179.34	128.35	230.33	1.06	1.09	246
8	158.14	164.25	158.94	107.94	209.94	1.01	1.04	247
9	150.21	159.61	157.73	117.25	198.20	1.05	1.25	193
10	145.63	168.81	176.01	133.89	218.13	1.21	1.25	256
11	152.96	145.66	152.45	109.36	195.53	1.00	1.03	260
12	153.39	168.79	171.09	135.54	206.64	1.12	1.17	239
13	100.44	114.98	128.60	101.31	155.88	1.28	1.28	175
14	87.09	127.84	118.49	94.62	142.36	1.36	1.36	161
15	87.25	117.36	128.76	102.57	154.96	1.48	1.48	125
16	106.45	129.33	128.86	108.10	149.62	1.21	1.47	151
17	89.45	125.83	135.04	108.56	161.53	1.51	1.56	181
18	89.37	125.12	120.59	98.62	142.57	1.35	1.41	158
19	69.44	100.61	109.73	92.64	126.82	1.58	1.75	120
20	59.37	95.70	98.35	74.72	121.97	1.66	1.84	141
21	29.97	50.06	61.86	47.31	76.40	2.06	2.06	67
22	31.19	56.55	65.86	48.96	82.76	2.11	2.11	66
X	156.83	179.95	179.00		179.00	1.14	1.19	177
Total	3,190.77	3,567.44	3,614.71	2,590.48	4,459.94	1.13	1.19	5,136

The lengths, including those from the Marshfield map, correspond to the chromosome regions spanned by our markers and will in general be shorter than the actual total lengths. The recombination rate for a chromosome excluding the centromere is calculated by deleting the genetic length and physical length of the two markers flanking the centromere.

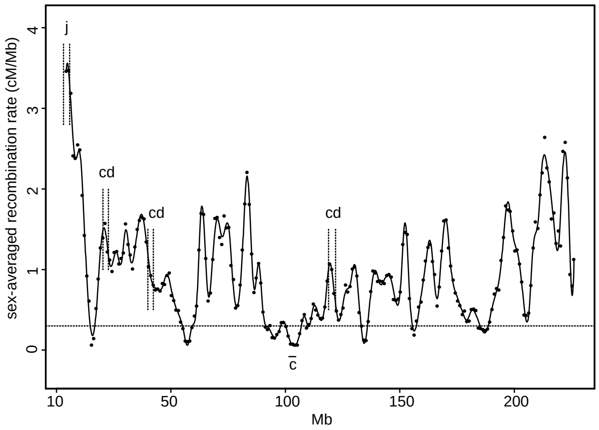

Fig. 1 Sex-averaged recombination rate for chromosome 3. Points correspond to sex-averaged crossover rates, calculated using moving windows 3 Mb in width; the shift from the center of one bin to the next is 1 Mb. The sex-averaged genetic distance for each 3-Mb window was calculated on the basis of our genetic map, and assumes a constant crossover rate between two adjacent markers. The solid curve was fitted to the points using smoothing splines[26]. c represents the centromere; cd represents the three recombination deserts and j the recombination jungle identified by Yu *et al.*[13] using data obtained from CEPH families.

11.8 cM (23.6%) and 9.3 cM (16.5%) greater than the lengths indicated by the Marshfield map.

Because of differences in recombination rates between the sexes, the estimated genetic length of the female autosomal genome (4,281 cM) differs from that the male genome (2,590 cM) by a ratio of 1.65.

High-resolution map shows fine structure of recombinations

We compared the high-resolution genetic map with our corrected sequence to derive recombination rates in centimorgans per megabase across the genome (Web Table E online gives estimated sex-averaged and sex-specific recombination rates at the marker locations). The shorter chromosomes usually have higher recombination rates than the longer ones, and the relationship between the average recombination rate and the physical length of a chromosome can be fitted well by a smooth curve (see Web Fig. A online). The average recombination rates of chromosomes 21 and 22 are twice as high as those of chromosomes 1 and 2. Recombination rates also vary across individual chromosomes, as illustrated by the sex-averaged crossover rates for chromosome 3 (Fig. 1; Web Fig. B online contains the corresponding plots for the other chromosomes). The crossover rate varied from over 3 cM Mb^{-1} at the telomere of the short arm to less than 0.1 cM Mb^{-1} at the centromere and its immediate surroundings on the short arm. Most interesting is the large number of local recombination peaks and valleys throughout each chromosome. Using the same 8 CEPH families from which the Marshfield map was constructed, Yu *et al.*[13] identified 19 recombination 'deserts', defined as regions with crossover rates less than 0.3 cM Mb^{-1}, and 12 recombination 'jungles', defined as regions with crossover rates greater than 3 cM Mb^{-1}. Three of the deserts and one of the jungles identified by Yu *et al.*[13] are on chromosome 3 (locations indicated in Fig. 1). We identified the same jungle, but none of their three deserts; however, we identified other potential desert regions. With respect to the whole genome, we detected 8 of the 19 deserts identified by Yu *et al.*[13] (5 with recombination rates between 0.3 and 0.5) but found better agreement with the 12 jungle regions, all located at the telomeres. It is likely that the discrepancy with regard to recombination deserts is due to the small sample size of the original study[13].

We calculated the crossover rates in males and females across chromosomes 1 and 7 (Fig. 2) much as we did the sex-averaged crossover rates (Fig. 1), but using a larger bin size (6 Mb as compared to 1 Mb). This provides poorer resolution but is necessary to ensure that the estimates have similar precision. We confirmed that crossover rates in females are much higher around the centromeres, whereas those in males tend to be higher towards the telomeres[4]. Our data show a more complicated pattern, however (for comparison, see ref. 4). Notably, although locations of local peaks and valleys for the two sexes tend to coincide, there are some instances of phase shifts, such that a peak for males corresponds to a valley for females and vice versa. Two such regions lie near the centromere on the p arm of chromosome 1 and around 25 Mb on chromosome 7. In addition, the ratio of sex-specific recombination rates fluctuates greatly across the chromosomes (Web Fig. C online). Over the whole genome, the correlation between male and female crossover rate is 0.57, high enough to lend support to the notion that underlying variables, such as sequence content and physical location, may explain a large fraction of the variation in sex-averaged crossover rates.

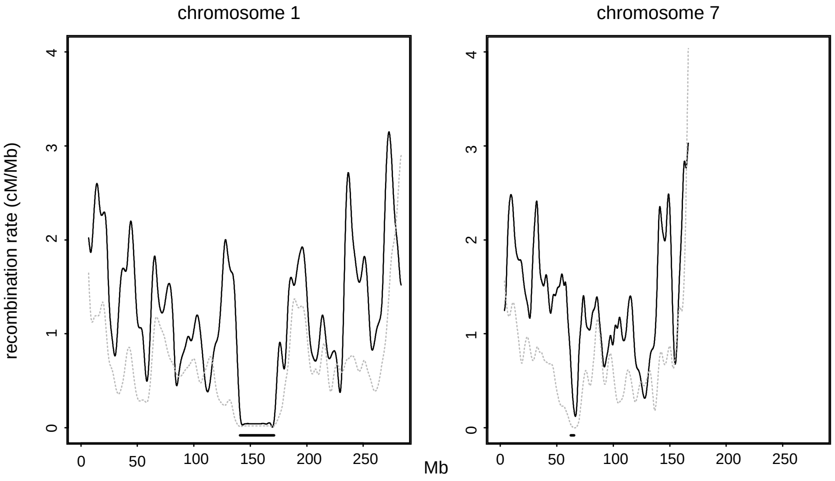

Fig. 2 Sex-specific recombination rates for chromosomes 1 and 7. Solid line, female; dashed line, male.

Correlation of recombination with sequence parameters

Many statistically significant correlations between recombination rates and sequence content have been identified using genetic maps such as the Marshfield map. These correlations are usually small, however, and parameters explaining a substantial percentage of the variance of recombination rates have not been identified. For example, among parameters relating to sequence content, the highest correlation seen in the study that identified recombination deserts and jungles[13] was with GC content, and this explained only 5% of the variation in sex-averaged recombination rates ($R^2 = 0.05$). In contrast, we saw much stronger correlation with GC content (correlation = 0.39, $R^2 = 0.15$) and other sequence parameters (Table 2). When we used the parameters

Table 2 • Results of simple and multiple regressions with sex-averaged recombination rate as the response

	Predicting sex-averaged recombination rates							
	Simple regression				Multiple regression (all predictors)		Multiple regression (best three predictors)	
Predictor	Coef.	Std. err.	R^2	*P*-value	Coef.	Std. error	Coef.	Std. error
Poly(A)/poly(T) fraction	–0.44	0.03	0.19	0	–2.23	0.13	–1.96	0.13
CpG fraction	0.40	0.03	0.16	0	1.42	0.10	1.27	0.10
GC content fraction	0.39	0.03	0.15	0	–3.27	0.22	–2.70	0.20
RefSeq gene count	0.14	0.03	0.02	0	–0.30	0.05		
PPY/PPU fraction	0.30	0.03	0.09	0	0.20	0.04		
UniGene cluster count	0.22	0.03	0.05	0	0.29	0.05		
					$R^2 = 0.37$		$R^2 = 0.32$	

We used 957 non-overlapping 3 Mb bins for individual data points. Fractions of sequence contents are all adjusted for the number of 'N' bases in the draft sequence, and only bins with less than 50% N bases are used. Results presented are based on standardized values (linearly transformed to have mean 0 and variance 1) of the response variable (recombination rate) and the six predictors. This does not affect R^2 or the *P* values, but makes the fitted coefficients more readily interpretable.

simultaneously to predict sex-averaged recombination rates using multiple regression, six parameters together explained about 37% of the variance and just three—CpG motif fraction, GC content and poly(A)/poly(T) ($(A)_{n\geq4}$ and $(T)_{n\geq4}$) tract fraction—explained about 32% of the variance. Although GC content was positively correlated with recombination rate when assessed separately, in the multiple regression fit it was negatively correlated with recombination rate. Close inspection showed that the three best predictors are all highly correlated pairwise: the correlation between CpG fraction and GC content is 0.94, the correlation between CpG fraction and poly(A)/poly(T) fraction is –0.85 and the correlation between GC content and poly(A)/poly(T) fraction is –0.96. This might suggest that these parameters capture essentially the same predictive information and that using two or three of them together would not substantially improve the prediction. But that is not the case: in particular, GC content is negatively correlated with recombination rates after adjustment for poly(A)/poly(T) fraction and CpG fraction. Thus, regions with the highest recombination rates tend to be those with high CpG fraction but low GC content and poly(A)/poly(T) fraction. Three other parameters—reference sequence genes, UniGene cluster and polypyrimidine/polypurine ratio (PPY/PPU)—are weakly, but statistically significantly, predictive of recombination rates.

Table 3 • Pearson sample correlation coefficients between the number of maternal recombinations on individual chromosomes and the number of maternal recombinations in the corresponding genome complement

Correlations of maternal gametic recombination rates across the genome (after adjusting for the mother effect)		
Genomic region	Correlation with the rest of the genome	*P*-value
Chr. 1	0.14	0.0224
Chr. 2	0.19	0.0020
Chr. 3	0.32	<0.0001
Chr. 4	0.19	0.0014
Chr. 5	0.24	0.0001
Chr. 6	0.21	0.0004
Chr. 7	0.16	0.0074
Chr. 8	0.15	0.0120
Chr. 9	0.17	0.0051
Chr. 10	0.17	0.0040
Chr. 11	0.13	0.0288
Chr. 12	0.18	0.0040
Chr. 13	0.18	0.0040
Chr. 14	0.16	0.0107
Chr. 15	0.01	0.8686 (NS)
Chr. 16	0.17	0.0045
Chr. 17	0.07	0.2333 (NS)
Chr. 18	0.18	0.0028
Chr. 19	0.26	<0.0001
Chr. 20	0.14	0.0240
Chr. 21	0.13	0.0378
Chr. 22	0.05	0.3820 (NS)
Chr. X	0.23	0.0001
Chr. 1–8	0.40	$<10^{-10}$

Also computed is the correlation between the first eight chromosomes with their complement. The correlations are calculated adjusted for the mother effect; for each chromosome, we compute the average number of maternal recombinations in a family, and this is subtracted from the number of maternal recombinations of each child in that family. NS, nonsignificant.

The substantially greater power to predict recombination rates, as compared with previous studies, that we obtained by using sequence parameters is probably a consequence of the substantially higher resolution of our genetic map, the availability of the draft sequences and our use of multiple regression.

We also observed a significant correlation between sex-averaged recombination rates and cytogenetic bands as defined by FISH mapping[14] ($R^2 = 0.06$, $P < 0.00001$). Specifically, among G bands, staining intensity (G25, G50, G75, G100) is inversely correlated to recombination rate. The G-negative bands have recombination rates somewhere between those of the G50 and G75 bands. This correlates well with GC content: in G bands, staining intensity decreases with GC content, but G-negative bands have a GC content somewhere between those of G50 and G75 bands.

Individual differences in recombination rates

On the basis of the 62 sibships with four or more sibs and both parents genotyped, comprising 269 male and 269 female meioses, we confirmed previous findings[4] of a systematic difference in recombination rates between mothers ($P = 0.002$) but not fathers. Notably, even after we adjusted for this 'mother effect', the number of recombinations was still positively correlated among chromosomes within the same maternal gamete. Thus, after adjustment, the correlation between the number of maternal recombinations on chromosome 3 and the sum of the maternal recombinations on the other 22 chromosomes was 0.32 ($P < 0.0001$). The correlation with the corresponding complement of the genome was positive for each of the 23 chromosomes (Table 3) and statistically significant in 20 of 23 cases. We artificially divided the genome into two halves of about equal genetic lengths, chromosomes 1–8 and chromosomes 9–22 plus X, and observed a correlation between the number of maternal recombinations in the two halves of 0.40 ($P < 1 \times 10^{-10}$). We did not detect similar correlation for the paternal recombinations.

The systematic mother effect and the maternal gamete effect that exists even after adjustment for the mother effect suggest that there is some yet unidentified factor—which may be partly genetic and partly environmental and varies within and between mothers—that has a global influence on maternal recombination rates affecting most, if not all, chromosomes simultaneously.

Comparison of the high-resolution and Marshfield maps

We saw a few large discrepancies and many small ones between our map and the Marshfield map. In the Marshfield map and to a lesser extent in ours, often more than one marker has been assigned the same position, reflecting a lack of resolution. The 5,012 markers shared by the two maps are assigned to 2,866 distinct positions in the Marshfield map and to 3,690 positions in our map. Even when we considered only pairs of markers that were apparently resolved on the Marshfield map, our marker order often did not agree with theirs. For example, among pairs of markers separated by 0.05–3.0 cM in the Marshfield map (not limited to adjacent pairs on the map), the two markers were ordered in reverse on our map in 6.7% of cases (5.5% where we had apparent resolution in our genetic data and 1.2% where our ordering of markers was based entirely on the draft sequence). Even when there was agreement as to marker order, the differences in estimated genetic distances were sometimes substantial (see Web Table F online for more details).

An accurate genetic map is crucial for linkage analysis, in which the locations of disease-susceptibility genes relative to a set of markers are estimated—and particularly for multipoint analysis, in which information from multiple markers is processed simultaneously[15,16]. In theory it is better to use sex-specific maps for linkage analysis[17], but in practice, nearly all published linkage scans are based on a sex-averaged map. Our map, based on over 600 meioses per sex, may make it possible to realize the theoretical gain obtainable by using sex-specific maps.

Corrections to the human sequence

In the process of collecting and analyzing our genetic data, we compared them with three Golden Path assemblies of successive freezes of the draft sequence, those of December 2000, April 2001 and August 2001. Apart from combining the information from both sources to obtain a best estimate of the marker order, this process also serves as a monitor of the changes and progress made in the sequence assembly. Using our genetic data, we ordered the markers by minimizing the number of obligate crossovers[18]. When the relative order of two or more markers have no effect on the number of obligate crossovers, the genetic data are considered to have no resolution. There were many instances where our genetic data had resolution, but our preferred order differed from that of either the April 2001 or August 2001 freeze, as illustrated for chromosome 3 (Fig. 3). The most illuminating case is that of region 5, covering approximately 22 Mb. The December 2000 freeze (not illustrated in the figure) and the August 2001 freeze both inverted the black and red segments together, a change involving about 8 Mb of the sequence. Compared with the December 2000 freeze, the April 2001 freeze further inverted the blue and black segments together, expanding the problematic region to about 22 Mb. This second error was corrected in the August 2001 freeze, but the 8 Mb inversion remained.

For the genome as a whole, although many changes occurred between the December 2000 freeze and the April 2001 freeze, there was no real improvement at the level of macro-assembly—some errors were corrected, but they were replaced by a similar number of new errors. But the August 2001 freeze appears to be a real improvement over the April 2001 freeze: most errors involving large segments of DNA were corrected and the total number of errors was reduced.

A genetic map for SNPs

Given a reliable assembly of the human sequence, markers for which we have no direct genetic mapping data can be assigned positions on the genetic map through linear interpolation between the sequence/physical map and our genetic map. Indeed, we have assigned genetic locations to about 2 million SNPs in the public databases[19] in a way that can be used by scientists in selecting and using SNPs for genetic mapping analysis (see Web Table G online).

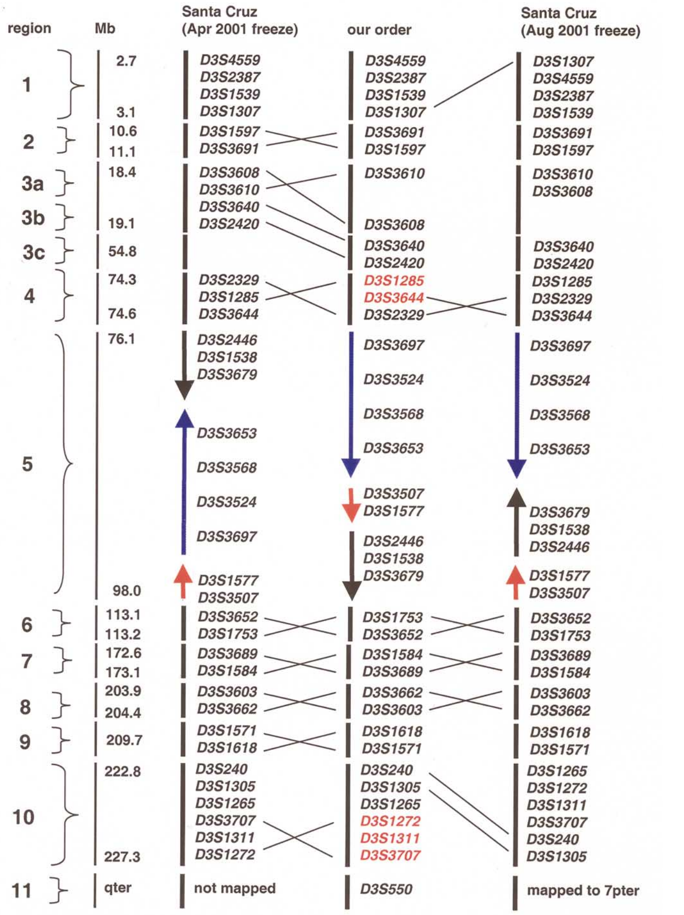

Fig. 3 Comparisons of our order and the Santa Cruz orders for the April and August 2001 freezes of the draft sequence. All discrepancies between our order and these freezes for chromosome 3 are shown. Physical positions of the genomic segments in the figure are indicated either by giving the starting and ending position of the segment, or by giving a single position when the segment is less than 100 kb in length. Physical positions shown are with respect to our modified sequence. Region 5 contains over 40 markers, of which only a subset are shown. Adjacent markers in red indicate that our genetic data lack sufficient resolution to determine the order. The misplacement of the marker *D3S1307* in region 1 appears to be an error in the annotations instead of an actual problem with the assembly, as our *in silico* analysis using the August freeze sequence put the marker at the right place. With regions 7 and 8, our order agreed with the December 2000 freeze. Regions 4 and 10 were apparently difficult regions for the assemblies, and although our genetic data revealed some inconsistencies, some uncertainty as to the order remained.

Discussion

The recombination map of the human genome described here reveals marked regional differences in recombination rates. Because meiotic recombination probably contributes to evolutionary change in humans, the regional differences in recombination rate raise the possibility that DNA changes contributing to evolution may not be entirely random, but rather may be more concentrated within specific regions. The regional differences in the recombination rate have prompted the speculation that recombination may be driven by sequence features such as the density of genes, the nature of genes and the presence of sequence repeats, among others. But differences in recombination rates between men and women demonstrate that there is more to recombination than just sequence. First, the frequency of recombination in the autosomes of females is 1.65 times that in the autosomes of males, although the autosomes are not known to contain any sex-specific sequence differences. If recombination events drive evolution, women may contribute more, in this regard, than do men. Second, there are regions in the genome where the recombination rate is particularly high in women and particularly low in men, and vice versa (data presented here and ref. 20). This indicates that forces outside the sequence contribute substantially to the determination of recombination rate.

Our observation of interfamily variation in maternal recombination rates is in agreement with previous reports of variable rates of chiasma formation[21] and crossover[4] in humans, and suggests that genetic factors may directly influence maternal recombination rates. This is in accordance with the finding in maize of a gene that controls recombination rates[22].

We saw significant differences in recombination rate among maternal gametes even after accounting for interfamilial differences. This suggests that stochastic factors operating during development or gametogenesis, or environmental factors acting over the many years of prophase of meiosis I in females, may affect the recombination rates of particular gametes.

We have achieved our original goal of constructing a more accurate genetic map of over 5,000 polymorphic microsatellite markers. But the intrinsic value of our primary data goes beyond that of the map from which it was constructed. For example, there have always been numerous disagreements between various human physical and genetic maps, which have not disappeared with the availability of the draft sequence. Theoretically, it is preferable to obtain a consensus order by combining and evaluating the original sources of data rather than by combining the resulting maps. In addition, although most discrepancies arise from limitations of the data, some may result from polymorphisms of macro-rearrangements[12]. Indeed, rearrangement polymorphisms, together with differences in individual maternal recombination rates, may account for some of the discrepancies in marker order and distance between the Marshfield map and our genetic map. Our data, by themselves or in conjunction with other data, can help to identify such rearrangement polymorphisms. These may be more frequent than expected and may contribute substantially to human phenotypic variation and, hence, natural selection. Recent studies[23,24] of linkage disequilibrium at a few locations suggest that local recombination hot spots tend to occur every 50–100 kb. When such data become available for the whole genome, it will be possible to determine whether the regions of high recombination rate that we have identified are driven by higher densities or higher intensities of recombination hot spots.

Methods

Data collection and genotyping. We obtained all biological samples used in this study according to protocols approved by the Data Protection Commission of Iceland (DPC) and the National Bioethics Committee of Iceland. We obtained informed consent from all patients and their relatives whose DNA samples were used in linkage studies. We encrypted all personal identifiers using an algorithm whose key was held by the DPC[25]. Details concerning genotyping, allele-calling, and genotype quality control are in Web Note C online.

Genotype data. Investigators interested in obtaining a copy of the genotype data should submit a completed agreement form (see Web Form A online) by fax to 354-570-1903 or by mail to Statistics Map, deCODE Genetics ehf, Sturlugata 8, IS-101 Reykjavik, Iceland. Data will be distributed on a CD-ROM, in a manner consistent with the protection of privacy. In addition to the removal of personal identifiers, the genotype data provided is also coded for anonymity. Specifically, alleles for each marker are randomly coded, but the coding is consistent across families. As a consequence, all results reported here can be reproduced independently with this data.

Ordering markers. With the genetic data, we evaluated an order of the markers based on the corresponding number of obligate recombinations[18]. This is a robust method based on the simple idea that if there is a crossover between two markers and if the order of the two markers is reversed, the single crossover will appear to be three consecutive crossovers, one in front of, one between, and one after the two markers. We performed computations by modifying our program, Allegro[10], and used a simulated annealing approach to search efficiently for orders that minimize the number of obligate crossovers. When the relative order of two or more markers had no effect on the number of obligate crossovers, we considered the genetic data to lack resolution. When our genetic mapping data had resolution and our preferred order was in disagreement with the sequence assembly, we considered modifying the sequence assembly. When the data was informative, we took a single recombination between the two markers as enough to determine the order of the two markers, as the wrong order would require two more recombinations than the right order, and this led to a likelihood ratio $>2 \times 10^3$ for distance smaller than 2 cM. We made 86 modifications to the assembly of the August 2001 sequence freeze with support from our genetic data: 53 supported by a reduction of four or more obligate recombinations (likelihood ratio $>4 \times 10^6$) and 33 supported by a reduction of two obligate recombinations (likelihood ratio $>2 \times 10^3$). We made an additional 18 modifications in cases where our genetic data lacked resolution, but there was strong support from alternative sources of physical mapping data (see Web Note D online for details on how sequence modifications were carried out).

Genetic distances. We estimated recombination probabilities between adjacent markers and then converted these to genetic distances using the Kosambi map so that they were directly comparable with the Marshfield map. We first calculated sex-specific distances and then averaged these to obtain the sex-averaged distances.

Correlation with cytogenetic bands. We determined statistical significance by one-way analysis of variance where recombination rate was the response and band-type was treated as a factor with five unordered categories: the G-negative bands and the G bands of four different staining intensities.

Differences in recombination rates. Because the data were not fully informative, there is some, though relatively little, uncertainty regarding the actual number of recombinations. To minimize the impact of the uncertainty in the data without unnecessarily complicating the presentation here, we used only sibships with four or more children and for which both parents were genotyped (62), accounting for a total of 269 meioses, to study maternal and paternal recombinations. We used Allegro to simulate 100 replicates of recombination patterns conditional on the genotype data and the estimated male and female maps. For each gamete, we used the number of maternal and paternal recombinations averaged over the 100 replicates for subsequent calculations. We used one-way analysis of variance, treating identity as mother or father as a factor, to obtain P-values when testing for mother and father effects. For the mother effect, the between-mother mean square and within-mother (residuals) mean square were 97.9 and 56.5, respectively, giving a F statistic of 1.73 (97.9/56.5) with 61 and 207 degrees of freedom ($P = 0.002$). Information on the individual families is in Web Table H online.

We obtained the *P*-values in Table 3 for the correlations of individual chromosomes with their genome complement on the basis of a permutation test; we performed 10,000 random permutations of the 269 mother-adjusted recombination counts. The *P*-value for the correlation between the first eight chromosomes with their genome complement was supported by asymptotic approximations corresponding to tests based on either the Pearson product moment, Spearman's rho or Kendall's tau. Also, the largest correlation coefficient obtained based on 500,000 permutations of the 269 recombination counts was only 0.28, substantially smaller than the observed value of 0.40.

Note: Supplementary information is available on the Nature Genetics website.

Acknowledgments
The authors wish to extend their gratitude to the 869 individuals providing DNA to this study, the Human Genome Project for providing the draft human genome sequence, and J.L. Weber and his colleagues for constructing the Marshfield genetic map.

Competing interests statement
The authors declare competing financial interests: see the Nature Genetics website (http://genetics.nature.com) for details.

Received 28 February; accpeted 21 May 2002.

1. International Human Genome Sequence Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Dib, C. *et al.* A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**, 152–154 (1996).
3. Murray, J.C. *et al.* A comprehensive human linkage map with centimorgan density. *Science* **265**, 2049–2054 (1994).
4. Broman, K.W., Murray, J.C., Sheffield, V.C., White, R.L. & Weber, J.L. Comprehensive human genetic map: individual and sex-specific variation in recombination. *Am. J. Hum. Genet*. **63**, 861–869 (1998).
5. Sheffield, V.C. *et al.* A collection of tri- and tetranucleotide repeat markers used to generate high quality, high resolution human genome-wide linkage maps. *Hum. Mol. Genet.* **4**, 1837–1844 (1995).
6. Sunden, S.L. *et al.* Chromosomal assignment of 2900 tri- and tetranucleotide repeat markers using NIGMS somatic cell hybrid panel 2. *Genomics* **32**, 15–20 (1996).
7. Utah Marker Development Group. A collection of ordered tetranucleotide-repeat markers from the human genome. *Am. J. Hum. Genet*. **57**, 619–628 (1995).
8. Rosenberg, M. *et al.* Characterization of short tandem repeats from thirty-one human telomeres. *Genome Res.* **7**, 917–923 (1996).
9. DeWan, A.T., Parrado, A.R., Matise, T.C. & Leal, S.M. The map problem: a comparison of genetic and sequence-based physical maps. *Am. J. Hum. Genet.* **70**, 101–107 (2002).
10. Gudbjartsson, D.F., Jonasson, K., Frigge, M.L. & Kong, A. Allegro, a new computer program for multipoint linkage analysis. *Nature Genet.* **25**, 12–14 (2000).
11. Dempster, A.P., Laird, N.M. & Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B* **39**, 1–38 (1997).
12. Giglio, S. *et al*. Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. *Am. J. Hum. Genet.* **68**, 874–883 (2001).
13. Yu, A. *et al.* Comparison of human genetic and sequence-based physical maps. *Nature* **409**, 951–953 (2001).
14. The BAC Resource Consortium. Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* **409**, 953–958 (2001).
15. Halpern, J. & Whittemore, A.S. Multipoint linkage analysis. A cautionary note. *Hum. Hered.* **49**, 194–196 (1999).
16. Gretarsdottir, S. *et al*. Localization of a susceptibility gene for common forms of stroke to chromosome 5q12. *Am. J. Hum. Genet.* **70**, 593–603 (2002).
17. Daw, E.W., Thompson, E.A. & Wijsman, E.M. Bias in multipoint linkage analysis arising from map misspecification. *Genet. Epidemiol.* **19**, 366–380 (2000).
18. Thompson, E.A. Crossover counts and likelihood in multipoint linkage analysis. *IMA J. Math. Appl. Med. Biol*. **4**, 93–108 (1987).
19. The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
20. Mohrenweiser, H.W., Tsujimoto, S., Gordon, L. & Olsen, A. Regions of sex-specific hypo- and hyper-recombination identified through integration of 180 genetic markers into the metric physical map of the human chromosome 19. *Genomics* **47**, 153–162 (1998).
21. Laurie, D.A. & Hulten, M.A. Further studies on bivalent chiasma frequency in human males with normal karyotypes. *Ann. Hum. Genet.* **49**, 189–201 (1985).
22. Ji, Y., Stelly, D.M., DeDonato, M., Goodman, M.M. & Williams, C.G. A candidate recombination modifier gene for *Zea mays* L. *Genetics* **151**, 821–830 (1999).
23. Jeffreys, A.J., Kauppi, L. & Neuman, R. Intensely punctuate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genet*. **29**, 217–222 (2001).
24. 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).
25. Gulcher, J.R., Kristjansson, K., Gudbjartsson, H. & Stefansson, K. Protection of privacy by third-party encryption in genetic research in Iceland. *Eur. J. Hum. Genet.* **8**, 739–742 (2000).
26. Venables, W.N. & and Ripley, B.D. *Modern Applied Statistics with S-plus* (Springer, New York, 1994).