Am. J. Hum. Genet. 75:1143–1148, 2004

Report

A Combined Linkage-Physical Map of the Human Genome

X. Kong,[1] K. Murphy,[2] T. Raj,[2] C. He,[1] P. S. White,[2] and T. C. Matise[1]

[1]Department of Genetics, Rutgers University, Piscataway, NJ; and [2]Division of Oncology, Children's Hospital of Philadelphia, Philadelphia

We have constructed de novo a high-resolution genetic map that includes the largest set, to our knowledge, of polymorphic markers ($N = 14,759$) for which genotype data are publicly available; that combines genotype data from both the Centre d'Etude du Polymorphisme Humain (CEPH) and deCODE pedigrees; that incorporates single-nucleotide polymorphisms; and that also incorporates sequence-based positional information. The position of all markers on our map is corroborated by both genomic sequence and recombination-based data. This specific combination of features maximizes marker inclusion, coverage, and resolution, making this map uniquely suitable as a comprehensive resource for determining genetic map information (order and distances) for any large set of polymorphic markers.

Accurate and comprehensive linkage maps are critical for the success of positional cloning projects and for several other types of genetics studies. The physical position—and, hence, the order—of the vast majority of polymorphic markers can now be readily determined from the assembled sequence of the human genome, and several large-scale genomewide linkage maps have been published and are in common usage (Dib et al. 1996; Broman et al. 1998; Kong et al. 2002). However, unless a given set of markers are all present on a single linkage map, identification of recombination-based meiotic map distances for any large set of markers remains difficult. In theory, physical map distances can be used to interpolate and estimate linkage map distances (Kong et al. 2002; Bahlo et al. 2004; Nievergelt et al. 2004). However, the existence of extreme variability in the genomic distribution of recombination (McVean et al. 2004) necessitates a painstaking effort to identify and utilize appropriate region-specific rates of cM/kb, making such large-scale interpolation generally impractical. Accurate estimates of meiotic map distance cannot be obtained by any means other than linkage analysis using genotype data.

To help address these issues, we have constructed de novo a high-resolution genetic map that includes the largest set, to our knowledge, of polymorphic markers ($N = 14,759$) for which genotype data are publicly available, that combines genotype data from both the CEPH and deCODE pedigrees, that incorporates SNPs, and that also incorporates sequence-based positional information. This specific combination of features maximizes marker inclusion, coverage, and resolution, making this map uniquely suitable as a comprehensive resource for determining genetic map information (order and distances) for any large set of polymorphic markers.

We initially identified 13,339 polymorphic markers for which genotype data could be obtained. All of these markers have been genotyped in the CEPH reference pedigrees (Dausset et al. 1990), in the deCODE pedigrees (Kong et al. 2002), or in both. Of these, 13,051 markers with genotypes were already present in our MAP-O-MAT genotype database (MAP-O-MAT Web site) and had been previously obtained primarily from the CEPH (CEPH Genotype Database Web site) and Marshfield Clinic (Center for Medical Genetics Web site) data sets, 239 were newly identified from the latest version of the CEPH database (version 9.0), and an additional 49 markers were identified from the deCODE linkage map. Finally, 2,821 SNPs scored in the CEPH pedigrees from the SNP Consortium linkage map (Matise et al. 2003; SNP Consortium Web site) were added to our set, which already contained 1,209 SNPs obtained from the other sources. This resulted in genotype data for a total of 16,160 markers (with >7.5 million genotypes) that were next subjected to a comprehensive cleaning step.

Received August 9, 2004; accepted for publication September 30, 2004; electronically published October 14, 2004.

Address for correspondence and reprints: Dr. Tara C. Matise, Department of Genetics, Rutgers University, 604 Allison Road, Piscataway, NJ 08840. E-mail: matise@biology.rutgers.edu

0002-9297/2004/7506-0021$15.00

A concerted effort was made to help ensure the uniqueness of the markers in our set. We used a stringent comparison of marker name aliases and primer sequences to identify markers that were represented in our data set more than once. Whenever possible, multiple lines of evidence, including comparison of physical positions, were sought to confirm an identified redundancy, and observation of tight pairwise linkage was always used as a final confirmation of redundancy.

These redundancies included 127 sets of markers that were identified by alias sharing, 61 sets of which included markers that had previously been assigned to more than one chromosome; 30 sets of redundant markers identified through sharing the same primers or the same physical positions; and 251 sets of markers previously identified as redundant by Marshfield. Genotype data for the duplicated sets of markers were merged so that there remained only a single entry for each marker in our data set. A description of the redundant markers is available on the authors' Web site. Removal of these markers and of markers not containing informative meioses resulted in a data set of 15,601 markers that, to the best of our knowledge, represent nonredundant loci within the genome. These markers are comprised of 9,473 STRs, 4,030 SNPs, 1,430 RFLPs/VNTRs and other hybridization-based markers, and 668 whose type could not be identified.

We used the PEDCHECK program (O'Connell and Weeks 1998) to identify and remove genotypes that lead to non-Mendelian transmission and are likely to be erroneous, as well as to search for problematic pedigrees. However, all of these genotype data have been previously cleaned, either by the groups who determined the genotypes or secondarily by other groups who have used these data in mapping; therefore, we detected and removed a negligible number of Mendelian inconsistencies (40 genotypes from six markers) and detected no problematic pedigrees.

Our working set contains 5,083 markers genotyped in both the deCODE and CEPH pedigrees (maximum 2,026 meioses; average 626 informative meioses), 10,469 markers genotyped only in the CEPH pedigrees (maximum 1,207 meioses; average 219 informative meioses), and 49 genotyped only in the deCODE pedigrees (maximum 922 meioses; average 388 informative meioses). The distribution of the number of informative meioses per marker is shown in figure 1.

To improve the accuracy of our map, we attempted to identify the genomic sequence position of all of the STS markers. Current sequence positions for 3,846 SNP markers were readily identified from the dbSNP database (dbSNP Home Page). We then used the me-PCR adaptation of the e-PCR program (Schuler 1997; Murphy et al. 2004) to search for the location of our STS (non-SNP) markers in assembled sequence downloaded

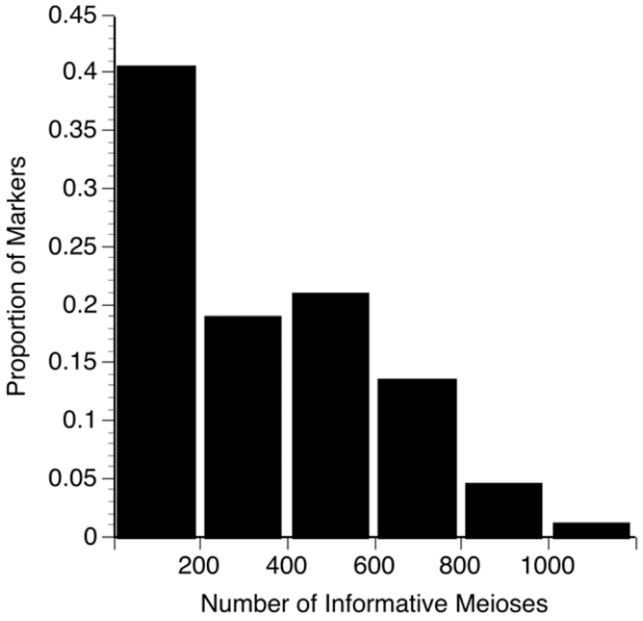

Figure 1 Frequency distribution of the number of informative meioses per marker.

from the National Center for Biotechnology Information (NCBI) ftp site (NCBI Build 34, July 2003). We then searched in the University of California–Santa Cruz STS database (UCSC Genome Bioinformatics Web site) for positions of markers not identified by e-PCR analysis. A marker was given a physical position if our searches yielded either a single hit in the genome ($N = 8,126$), or multiple hits located within a total of 1,000 bases on the same chromosome ($N = 666$). If the search results did not meet these criteria, no physical position was assigned. By this approach, we were able to determine a unique physical location for 8,792 (93%) of the STS markers. When combined with the SNPs, there were a total of 12,638 markers (81% of total set; 94% of PCR-based markers) with physical positions available for mapping.

The initial step toward constructing linkage maps was to test each marker for linkage to its assigned chromosome. Chromosome assignments were determined by physical map data, when available, and otherwise by chromosomal assignment on previously published linkage maps. We used the CRI-MAP computer program (Lander and Green 1987) for all likelihood calculations. We identified 166 markers that did not show linkage to at least one other marker on the same chromosome with a LOD score of ≥ 3.0. These unlinked markers were excluded from further mapping steps and are listed on the authors' Web site.

Our mapping algorithm used sequence-based positions to determine an initial map; this was followed by

Table 1

Description of the Rutgers Combined Linkage-Physical Maps

Chromosome	No. of Markers Mapped to		Physical Length (kb)	Map Length (Kosambi cM)		
	Positions	Intervals[a]		Sex Averaged	Female	Male
1	968	239	2,451.3	286.5	358.0	221.7
2	906	151	2,431.6	263.3	338.6	191.6
3	802	199	1,989.6	225.1	282.3	170.2
4	677	161	1,915.1	212.2	273.2	154.7
5	677	189	1,803.2	208.2	264.9	155.5
6	689	150	1,699.6	192.2	247.6	140.9
7	624	160	1,580.6	189.0	237.8	142.2
8	603	109	1,457.0	173.3	220.2	132.3
9	502	90	1,358.1	168.7	198.4	141.1
10	619	158	1,346.0	173.5	216.5	133.2
11	572	163	1,340.8	163.8	205.3	124.3
12	550	94	1,314.3	174.2	213.8	137.2
13	374	99	944.5	128.9	157.0	102.5
14	395	89	851.3	123.8	146.7	101.8
15	344	87	796.9	130.2	157.1	106.5
16	378	118	895.1	134.2	159.5	110.9
17	438	213	811.8	137.5	164.6	113.0
18	342	47	753.3	124.2	148.3	102.3
19	308	93	627.8	112.2	126.3	99.3
20	322	68	613.8	102.5	125.3	82.1
21	194	35	332.2	68.5	80.5	58.4
22	175	69	331.1	86.1	93.4	79.4
23	411	108	1,522.7	184.8	184.8	11.7
Total	11,870	2,889	29,167.7	3,762.6	4,600.1	2,812.8

[a] These markers could not be localized into a single map position and were instead localized to a larger map interval or bin.

several steps to modify this map on the basis of the assumptions that the initial placement of several of the markers would be incorrect and that there exist undetected genotype errors. The initial map consisted of 12,536 markers for which physical positions were identified and linkage groups were confirmed. This map had a total sex-averaged length of 4,623 cM (Kosambi). We performed linkage analyses using the genotype data to confirm or refute the proposed sequence-based order. For this "remapping" analysis, each marker was subsequently removed from the map; linkage analysis was then used to identify its recombination-based map position on the remaining map. On a well-ordered and statistically well-supported map, each marker would map back to its putative map position, with high statistical support. However, even though the pedigrees in the genotype data set had relatively high numbers of meioses, the extremely high density of markers on the map exceeded the resolving power of the genotype data and caused the linkage analysis for most markers to result in several map intervals that showed equal likelihood for placement. Therefore, a marker was kept in its original sequence-based position as long as the physical position was within any of the most likely recombination-based map intervals (those within a LOD unit of 1.0 from the position with the highest likelihood)—in other words, as long as the physical position *was consistent with* the linkage position. Markers whose linkage result did not encompass, or was inconsistent with, the physical position were removed from this initial map. This step resulted in removal of 1,769 (14%) of the markers.

We next cleaned the genotype data by removal of genotypes that were likely to be erroneous, identified as those that led to close double-recombination events, given the map that resulted from the initial remapping step. This cleaning step resulted in removal of 0.51% of the genotypes. We noted that the relative contribution of SNPs to the apparent double-recombination events was much greater than expected. We found that 50% of the markers leading to the double-recombination events were SNPs, whereas SNPs represented only 30% of the set of 12,536 markers being mapped. This finding is not surprising because, as a result of having only two alleles, genotyping errors were less likely to have been detected among SNPs than STRs during the earlier tests of non-Mendelian inheritance.

Now that the genotype data were more accurate, we performed a second remapping step in which linkage analysis was used to try to add the previously removed

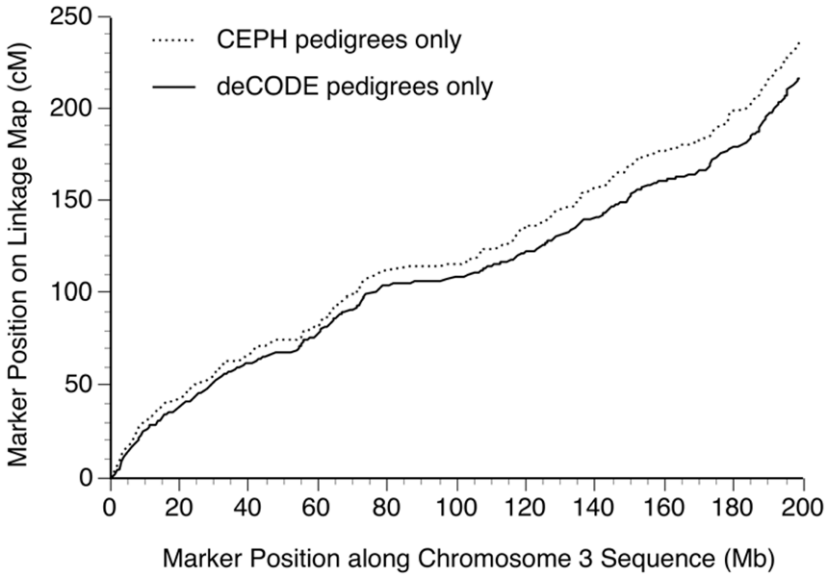

Figure 2 Comparison of map distances on our map of chromosome 3, using only the genotype data from the CEPH pedigrees (*dotted line*) versus using only the genotype data from the deCODE pedigrees (*solid line*). Similar graphs for all chromosomes can be viewed on the authors' Web site.

markers back to the map. By use of the same procedure as described above, markers for which the linkage data were now consistent with their physical position (N = 1,168) were reincorporated into the map at their sequence-based locations. An additional round of genotype cleaning, as described above, was then performed (with 0.25% of genotypes removed), followed by a final remapping test on all of the mapped markers. This map contains 11,870 markers whose physical map position is corroborated by recombination-based mapping data, representing 95% of the markers that were available for our combined linkage-physical mapping.

The list of markers whose physical position did not match the linkage-based position can be found on the authors' Web site. We evaluated this set of inconsistent markers to search for factors contributing to their inconsistency. We did not observe any physical clustering of these inconsistent markers along the chromosomes. We did note that, on average, these markers were considerably less informative than those whose linkage-based position was consistent with their physical position, with the inconsistent markers having an average of 230 informative meioses and heterozygosity of 0.53, whereas the consistent markers averaged 422 informative meioses and heterozygosity of 0.61. It is, therefore, not surprising that we find a slight overrepresentation of SNPs among these inconsistent markers (35%, compared with 30% among the consistent markers). These results suggest that the inconsistent set may not have had adequate power for the statistical process of linkage mapping.

Each marker on this map has a unique physical map position, with an average distance of 246 kb between markers. The average linkage map distance is 0.3 cM (0.7 cM if the 0-cM intervals are excluded). This map resolution far exceeds the resolving power of even the relatively large number of meioses for which these markers were genotyped, such that there are no observed recombination events and, hence, a linkage map distance of zero within many (56%) of the map intervals.

The final mapping step was to localize the remaining markers in our data set onto this map. Since most of these markers did not have a sequence-based position, we used linkage analysis to identify map interval placements. For those markers for which sequence position was known, an interval position was accepted only if the linkage interval encompassed the physical position. By this analysis, we were able to identify statistically well-supported map intervals for 2,889 markers. Thus, in total, our map contains both physical and meiotic position information for 14,759 markers. The total sex-averaged length of our maps is 3,763 cM (Kosambi). The female and male map lengths are 4,600 cM and 2,813 cM, respectively (table 1). Our mapping and cleaning steps resulted in a 19% reduction in map length from the initial map, supporting our supposition that the initial genotype data contained previously undetected genotyping errors and that some of the markers on the initial map were incorrectly localized.

Our map covers an additional 28.7 Mb and 84 cM beyond that covered by the deCODE linkage map. The sex-averaged map lengths of the genomic regions

spanned by both our map and the deCODE map are very consistent (Rutgers: 3,679 cM; deCODE: 3,615 cM). This similarity in overall map lengths is one line of evidence indicating that there are no gross differences in rates of recombination between the CEPH and the deCODE pedigrees. Further support for this conclusion is drawn from a comparison of map lengths on our map determined using only the genotype data for the CEPH pedigrees versus those determined using genotype data from only the deCODE pedigrees. Visual comparison of these map lengths shows good agreement on all chromosomes. An example comparison for chromosome 3 is shown in figure 2, and the others can be found on the authors' Web site.

We compared both marker order and map lengths on our linkage-physical map with those on the deCODE (Kong et al. 2002) and Marshfield (Broman et al. 1998) maps. When compared with the deCODE map, 99.7% of the markers in common showed the same order, and the corresponding sizes of map intervals were also highly correlated ($r = 0.92$). When compared with the Marshfield map, 95.8% of the markers in common showed the same order, but the correlation of interval sizes was much lower ($r = 0.51$). It is not surprising to find a lower map distance correlation with the Marshfield map since, unlike the deCODE map, this map was made without the benefit of corroboration from the assembled sequence and contains a few markers whose Marshfield position is quite different than on our map or the deCODE map.

Additional data describing our maps is available on the authors' Web site. This set of combined genotype data (exclusive of the deCODE data) and maps has been incorporated into the MAP-O-MAT Web-based linkage-mapping server for analysis of human linkage maps (Matise and Gitlin 1999; Kong and Matise 2004; MAP-O-MAT Web site). MAP-O-MAT facilitates the verification of marker order and calculation of map distances for custom mapping sets by running the CRI-MAP program for linkage analysis using the genotype data described in this article. Our maps have also been integrated into the eGenome genomics resource (eGenome Web site), which presents marker and map distances in the context of additional genomic features. We are pursuing the addition of thousands of additional SNP markers to our data set and map, including the forthcoming ABI (Applied Biosystems) SNPlex Human Linkage Mapping Set 4K, the Illumina Linkage III SNP Panel (Murray et al. 2004), and *Eco*RI SNPs genotyped using Affymetrix technology (Kennedy et al. 2003).

Acknowledgments

We are grateful to deCODE Genetics for allowing us to use their genotype data. We thank Suzanne Leal and Linda Brzustowicz for helpful discussions. This work was partially supported by National Institutes of Health grants HG01691 (to T.C.M.) and MH60240 (to P.S.W.) and by March of Dimes grant 12-FY02-108 (to T.C.M.).

Electronic-Database Information

The URLs for data presented herein are as follows:

Authors' Web site, http://compgen.rutgers.edu/maps/ (for Rutgers Linkage-Physical Maps)
Center for Medical Genetics, Marshfield Clinic, http://research.marshfieldclinic.org/genetics/ (for the Marshfield genetic maps)
CEPH Genotype Database, http://www.cephb.fr/cephdb/
dbSNP Home Page, http://www.ncbi.nlm.nih.gov/SNP/
eGenome, http://genome.chop.edu/
MAP-O-MAT, http://compgen.rutgers.edu/mapomat/
NCBI ftp site, http://www.ncbi.nlm.nih.gov/Ftp/
SNP Consortium, http://snp.cshl.org/linkage_maps/ (for the SNP Consortium linkage maps)
UCSC Genome Bioinformatics, http://genome.ucsc.edu/

References

Bahlo M, Xing L, Wilkinson CR (2004) HumanMSD and MouseMSD: generating genetic maps for human and murine microsatellite markers. Bioinformatics. http://bioinformatics.oupjournals.org/cgi/reprint/bth375v1 (electronically published June 24, 2004; accessed October 12, 2004)

Broman KW, Murray JC, Sheffield VC, White RL, Weber JL (1998) Comprehensive human genetic maps: individual and sex-specific variation in recombination. Am J Hum Genet 63:861–869

Dausset J, Cann H, Cohen D, Lathrop M, Lalouel J-M, White R (1990) Centre d'Etude du Polymorphisme Humain (CEPH): collaborative genetic mapping of the human genome. Genomics 6:575–577

Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E, Lathrop M, Gyapay G, Morissette J, Weissenbach J (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature 380:152–154

Kennedy G, Matsuzaki H, Dong S, Liu W, Huang J, Liu G, Su X, Cao M, Chen W, Zhang J, Liu W, Yang G, Di X, Ryder T, He Z, Surti U, Phillips M, Boyce-Jacino M, Fodor S, Jones K (2003) Large-scale genotyping of complex DNA. Nat Biotechnol 21:1233–1237

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K (2002) A high-resolution recombination map of the human genome. Nat Genet 31:241–247

Kong X, Matise T (2004) MAP-O-MAT: internet-based linkage mapping. Bioinformatics. http://bioinformatics.oupjournals.org/cgi/reprint/bti024v1 (electronically published September 16, 2004; accessed October 14, 2004)

Lander ES, Green P (1987) Construction of multi-locus genetic linkage maps in humans. Proc Natl Acad Sci USA 84:2363–2367

Matise T, Gitlin J (1999) MAP-O-MAT: marker-based linkage

mapping on the World Wide Web. Am J Hum Genet Suppl 65:A435

Matise TC, Sachidanandam R, Clark AG, Kruglyak L, Wijsman E, Kakol J, Buyske S, et al (2003) A 3.9-centimorgan-resolution human single-nucleotide polymorphism linkage map and screening set. Am J Hum Genet 73:271–284

McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P (2004) The fine-scale structure of recombination rate variation in the human genome. Science 304:581–584

Murphy K, Raj T, Winters RS, White PS (2004) me-PCR: a refined ultrafast algorithm for identifying sequence-defined genomic elements. Bioinformatics 20:588–590

Murray SS, Oliphant A, Shen R, McBride C, Steeke RJ, Shannon SG, Rubano T, Kermani BG, Fan J-B, Chee MS, Hansen MST (2004) A highly informative SNP linkage panel for human genetic studies. Nat Methods 1:113–117

Nievergelt CM, Smith DW, Kohlenberg JB, Schork NJ (2004) Large-scale integration of human genetic and physical maps. Genome Res 14:1199–1205

O'Connell JR, Weeks DE (1998) PedCheck: a program for identification of genotype incompatibilities in linkage analysis. Am J Hum Genet 63:259–266

Schuler GD (1997) Sequence mapping by electronic PCR. Genome Res 7:541–550