Table 2 Distribution of mother's age at birth for all offspring

Mother's age at birth	Count	Percent
15	317	2.2
20	3,340	23.6
25	4,662	33.0
30	3,386	23.9
35	1,770	12.5
40	595	4.2
45–50	70	0.5
Total	14,140	100.0

born between 1925 and 1955. Table 2 gives the distribution of maternal age at birth.

Genotype data do not provide complete information on recombination counts, which complicates the analysis. To handle this missing-data problem[13], we applied two different statistical methods. The first method, called 'mean imputation', imputes the recombination counts using the best guesses. The way we implemented this method makes it robust, meaning that the calculated P values are insensitive to model mis-specifications or potential artifacts in the data. There is some loss of efficiency, however, and effects tend to be underestimated. The second method is likelihood-based, is fully efficient but computationally intensive, and can be sensitive to model mis-specification.

Using the robust method, we estimated the effect of maternal age on recombination rate to be 0.043 recombinations per year (s.e. = 0.011; $P = 0.00016$). Because we used family-adjusted recombination counts and the ages of mothers at birth, the age trend that we detected existed 'within family' (*i.e.*, a child born to a mother later in life tends to have more maternal recombinations than a child born earlier in her life) and was not simply a consequence of the possibility that some mothers tend to have children later in life and also happen to have higher recombination rates. The likelihood-based method gives an estimate of 0.082 recombinations per year (s.e. = 0.012; $P < 1 \times 10^{-8}$).

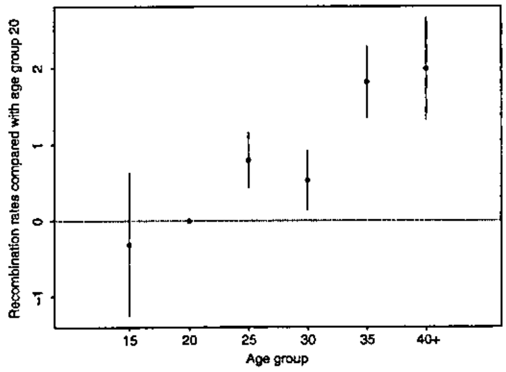

Figure 1 Recombination rate and maternal age. Using the age group 20 as the reference, the estimates and 95% confidence intervals for the differences in recombination rates between the other age groups and age group 20 are shown. Maternal ages of 40 and more were grouped into a single bin. There is a trend towards increase, but the data deviate from linearity with a slight drop in the estimate from age 25 to 30 followed by a big jump from age 30 to 35. Although the drop from age 25 to 30 is not statistically significant and probably not real, the data do support a relatively big incremental increase from age 30 to 35. An exact linear relationship between recombination rate and maternal age is not consistent with the data and is rejected by a goodness of fit test ($P < 0.005$).

Although the effect is significant even with the conservative method, the higher estimate based on the likelihood method is probably better. To determine whether the age effect is well fitted by a linear relationship, we fitted a model treating maternal age as a categorical variable using the likelihood method (Fig. 1; the distributions of maternal recombination counts of individual offspring are shown in Supplementary Fig. 1 online). The age effect is already apparent for relatively young women, and there is a marked increase in recombination rate from age 30 to 35. Notably, the rate of increase of maternal nondisjunction accelerates during this time frame.

The maternal age effect translates into only an additional two recombinations, or ~4% of the average maternal recombination rate, over a period of 25 years. But the relevance and importance of the observed effect depends on the underlying causes. There are at least two possible explanations for the results: first, recombination rate among the eggs of a woman increases as she ages; and second, the recombination rate of eggs does not increase, but there is a selection effect that increases the chance of an egg with more recombinations to produce a successful live birth. This selection effect probably exists even early on, but becomes stronger as the woman ages.

The first explanation is unlikely to be true, because recombinations take place prenatally and a 'production line' hypothesis would have to be invoked as outlined above. Moreover, this increase contradicts the observed decrease in chiasma frequency reported for mice. The second explanation, related to selection forces, is more plausible. A higher number of recombinations along a chromosome might reduce the chance of maternal age–related nondisjunction, the leading cause of pregnancy loss due to aneuploidy in the fetus. Maternal nondisjunction is associated with maternal age and reduced levels or altered placement of recombination[14]. Altered recombination has been identified for all examined cases of trisomy arising at the first stage of maternal meiosis. Consequently, increasing amounts of meiotic recombination may be protective for certain forms of nondisjunction, depending on the location of the additional exchanges. There is evidence, at least for chromosome 15, that multiple recombinants may be more resistant to nondisjunction because of increased stability of the bivalent over time[15]. Age-related abnormalities in spindle morphology and chromosome alignment at the meiotic plate have been reported[16], suggesting that some components of the meiotic apparatus are susceptible to the effects of aging. It has also been proposed that the sister chromatid cohesion complex may suffer an age-related breakdown[17]. If this is true, then meiotic tetrads from older oocytes may retain their integrity on the basis of their chaismata alone. Greater numbers of recombination would then provide additional protection from age-related meiotic breakdown.

Under the selection hypothesis, women with higher recombination rates would have more children. To examine this possibility, we regressed the total number of children of the mother on (i) the (estimated) recombination rate of the mother; (ii) the number of genotyped children of the mother; (iii) the mother's mean age at the

Table 3 Estimated effects for four predictors of family size

Predictor	Estimated effect	Standard error (s.e.)	P value
Recombination rate of mother	0.0109	0.0041	0.0076
Number of genotyped children of the mother	0.6815	0.0212	0.0000
Mother's average age at birth of her children	0.0002	0.0045	0.9637
Birth date of mother	–0.0469	0.0022	0.0000

Table 4 Estimated effect of mother's age on recombination rate for different parts of the genome

	Centromeric half	Telomere minus 6 cM	Telomeric 6 cM
Female genetic length (cM)	23.904	20.483	1.152
Recombination increase per year (s.e.)	0.038 (0.0072)	0.041 (0.0065)	0.009 (0.0014)
Percent increase per year (s.e.)	0.158 (0.0299)	0.201 (0.0316)	0.761 (0.1231)

The centromeric half consists of the part of each chromosome arm that is next to the centromere and accounts for roughly one-half of that chromosome arm in female genetic distance. The telomeric 6-cM region is defined as in **Supplementary Table 1** online.

times of birth of the genotyped children; and (iv) the mother's birth date (**Table 3**). As expected, the number of genotyped children of the mother is correlated with the total number of children of a mother, but the correlation is not perfect ($R^2 = 0.19$). Its inclusion in the regression ensures that any correlation observed between family size and recombination rate of a mother is not spurious: a higher number of recombinations is not estimated or detected simply because more children are genotyped. After accounting for the generational trend, recombination rate has a positive and statistically significant effect ($P = 0.0076$) on family size. With the mother's mean age at the times of birth of the genotyped children, which happens to be nonsignificant, also included in the regression, mothers who have a larger number of children have a higher recombination rate not simply because they have more children at a later age. Although it is significant, the effect of recombination rate on family size is modest. This is not surprising, as many factors affect family size.

We investigated whether the maternal-age effect is specific to certain genomic regions, as data from nondisjoined chromosomes indicate that there is selection against specific chiasmatic configurations[18]. The age effect is roughly the same for long and short chromosomes. Dividing each chromosome arm into two roughly equal parts on the basis of female genetic distance, the telomeric halves have a slightly higher percentage increase per year than the centromeric halves, but the difference is not significant. Focusing on marker intervals within 6 cM of our most telomeric marker (**Supplementary Table 1** online), we determined that the percentage increase per year for these telomeric regions is roughly four times higher than that of the rest of the genome ($P < 0.0001$; **Table 4**). Because these regions account for only ~2.5% of the genome in genetic length, however, ~90% of the yearly increase of recombinations observed is accounted for by the other parts of the genome.

We observed no association of recombination rates with paternal age (**Supplementary Table 2** and **Supplementary Fig. 2** online), nor did we identify a systematic difference in recombination rates between or within men. A previous report using an immunofluorescence method to examine exchanges in human spermatocytes described significant variation in recombination rates within and among men, but no age effect[19]. The observed variation identified among spermatocytes, but not live births, suggests that selection occurs at the level of spermatocytes. Presumably, the checkpoints for such meiotic disruptions are more stringent in spermatocytes than in oocytes[17].

The proposed selection hypothesis explains the maternal-age effect and the correlation of maternal recombination rate with family size. But there could be alternative explanations. A recent paper[20] challenged the doctrine that all the oocytes of a woman are produced when the woman is still at her fetal stage and suggested that follicular renewal may occur in the postnatal mammalian ovary. If true, this would provide a natural time ordering of the oocytes that corresponds to the dates of birth of the children and an alternative explanation for the age effect we observed. But this theory does not explain why mothers who have higher recombination rates have more children. Our observations and hypotheses do not contradict this new theory; there could be both follicle renewal and selection associated with recombination counts.

Among the 5,463 families studied, 1,090 mothers make up 545 independent sister pairs. Based on the correlation of estimated recombination rates of these sisters, the heritability[21] of recombination rate is estimated to be 30.4% (s.e. = 8.5%; $P = 0.0004$), which supports the idea that there is a large genetic component to recombination rate. Together with the hypothesis that reproductive success of eggs is dependent on the number of recombinations they have across the genome, these data imply that not only do recombinations have a role in evolution by yielding diversity of combinations of gene variants for natural selection, but they are also under selective forces acting at the level of chromosome segregation and reduced survival of mis-segregated oocytes.

METHODS

Data collection and genotyping. We obtained all the biological samples used in this study according to protocols approved by the Data Protection Commission of Iceland and the National Bioethics Committee of Iceland. We obtained informed consent from all participants. All personal identifiers were encrypted using a code that is held by the Data Protection Commission of Iceland[12]. Details concerning genotyping, allele-calling and genotype quality control can be found in the supplemental material of our previous study[1].

Statistical methods. The first method we applied to study the age effect is called mean imputations[13] and is similar to the method we used previously[1]. With all the family data, we first fitted a male and a female genetic map using maximum likelihood and the EM algorithm[22]. We then calculated the expected paternal and maternal recombination counts of each child conditional on the observed genotype data and the fitted maps. We calculated the conditional expected recombination counts using the simulation option of our linkage program Allegro[23]; we carried out 100 simulations and computed the averages. We then treated these estimated recombination counts as though they were the actual recombination counts in the subsequent analysis. To study the age effect, we regressed the family-adjusted recombination counts on the family-adjusted age of mother at birth of the child (the family-adjusted value is the difference between the value of a child and the value averaging over all the children in the same family). Using the family-adjusted values not only ensures that any potential artifacts are eliminated, because any bias would have the same effect on all children in the family, but it also implies that any age effect detected will not be confounded by the differences between mothers. We obtained the P values through a randomization test. The children in a family were permuted and the analysis repeated. We did this 25,000 times, and the two-sided P value reported is twice as large as the fraction of times that the permutations produced an estimated effect bigger than or equal to the observed effect of 0.043. This method is robust and completely insensitive to model misspecifications. There is, however, loss of efficiency and the effect tends to be underestimated. This is because the mean imputations are done under the model of no age effect; hence, the estimated effect has a tendency to shrink towards the null hypothesis. The amount of shrinkage is proportional to the amount of missing information, which is expected to be quite small for the data set in our previous study[1] as over 5,000 markers were used there; but, as only ~1,000 markers were used here, the shrinkage is expected to be greater. Also, when there are only two siblings genotyped and the genotypes of grandparents are not available, the data can provide a good estimate of the total maternal and paternal recombination counts in the family, but the data are completely uninformative regarding whether a crossover occurs with one child or the other, and, as a consequence, the mean imputations would be the same for both siblings. Hence, when using mean imputations in conjunction with a family-adjusted analysis, families with two siblings genotyped are completely uninformative. As including these families would not add information but would further shrink the estimated effect, we used only the 2,177 families with three or more siblings genotyped when applying this method. But we used all 5,463 families for the likelihood approach described below, and also when the mean imputations were used to study the relationship between family size and recombination rate of a mother.

The second method we used was a full-likelihood approach, which is maximally efficient but may not be as robust as our first method. It is based on a model that assumes a multiplicative maternal effect which is constant across the genome, that is, any effect on recombination rate is assumed to affect all genomic regions equally, unless otherwise specified. The multiplicative effect is modeled as a function of the mother, the gamete and the age of the mother at birth. The mother and gamete effects are modeled as random effects and the age effect is modeled as a fixed effect. We employed the following model for the mean number of maternal crossovers per gamete per chromosome: $\mu_{mgc} = \exp(\alpha_c + \beta \times \mathrm{age}_{mg}) \times \exp(U_m + U_{mg})$, where m indexes mother, g indexes gamete, c indexes chromosome and U_m and U_{mg} are assumed to be independent and normally distributed. The full model can be viewed as a generalized linear mixed model[24] with a Poisson random conditional component, log link and normally distributed random effects. To simplify the analysis, we assumed the absence of any crossover interference throughout the model. This can create some biases as interference exists[25], but the effect is likely to be modest for the parameters of interest in this study. To transform from the multiplicative scale to the additive scale, we took $(\exp(\beta) - 1)$ times the total estimated length of the maternal genome as the additive effect of maternal age on recombination rate. Because the recombination counts are not directly observed, even with the assumption of no interference, maximizing the likelihood under both the null hypothesis (no age effect) and the alternative hypothesis is challenging, going beyond the difficulties found in standard generalized linear mixed models. To meet these computational challenges we applied various computational techniques that include the Monte Carlo Newton-Raphson algorithm, Monte Carlo EM-algorithm and importance-reweighting of the samples simulated from Allegro under the null hypothesis[24,26,27]. Standard errors for the model parameters are determined on the basis of the observed Fisher information, and P values are obtained on the basis of likelihood ratio tests. When studying potential differences between genomic regions, a separate α_c and β are assigned to each genomic group.

In **Figure 1** and **Supplementary Figure 2** online, the confidence intervals have incorporated the uncertainties of both the recombination rate of the particular age bin and the recombination rate of age bin 20. In **Supplementary Figure 1** online, the box plots are constructed using a central box, indicating the range of the middle 50% of recombination (from the first to the third quartiles) with the median value indicated by a horizontal bar within the box, and whiskers (the dashed vertical lines and the hinged horizontal lines at their edge) that extend to the furthest data point that is not more than one-and-a-half times the width of the interquartile range beyond the central box. All other recombination values further away are indicated individually by horizontal lines.

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

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests (see the *Nature Genetics* website for details).

Received 10 May; accepted 30 August 2004
Published online at http://www.nature.com/naturegenetics/

1. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nat. Genet.* **31**, 241–247 (2002).
2. 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).
3. Henderson, S.A. & Edwards, R.G. Chiasma frequency and maternal age in mammals. *Nature* **218**, 22–28 (1968).
4. Speed, R.M. & Chandley, A.C. Meiosis in the foetal mouse ovary. II. Oocyte development and age-related aneuploidy. Does a production line exist? *Chromosoma* **88**, 184–189 (1983).
5. Polani, P.E. & Jagiello, G.M. Chiasmata, meiotic univalents and age in relationship to aneuploid imbalance in mice. *Cytogenet. Cell. Genet.* **16**, 505–529 (1976).
6. Polani, P.E. & Crolla, J.A. A test of the production line hypothesis of mammalian oogenesis. *Hum. Genet.* **88**, 64–70 (1991).
7. Tanzi, R.E. *et al.* A genetic linkage map of human chromosome 21: analysis of recombination as a function of sex and age. *Am. J. Hum. Genet.* **50**, 551–558 (1992).
8. Haines, J.L. *et al.* A genetic linkage map of chromosome 21: a look at meiotic phenomena. *Prog. Clin. Biol. Res.* **384**, 51–61 (1993).
9. Elston, R.C., Lange, K. & Namboodiri, K.K. Age trends in human chiasma frequencies and recombination fractions. II method for analyzing recombination fractions and applications to the ABO:nail-patella Linkage. *Am. J. Hum. Genet.* **28**, 69–76 (1976).
10. Renwick, J.H. & Schulze, J. Male and female recombination fractions, for the nailpatella:ABO linkage in man. *Ann. Hum. Genet.* **28**, 379–392 (1965).
11. Weitkamp, L.R. *et al.* The relation of parental sex and age to recombination in the HLA system. *Hum. Hered.* **23**, 197–205 (1973).
12. 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).
13. Little, R.J.A. & Rubin, D.B. *Statistical Analysis with Missing Data* (John Wiley & Sons, New York, 1987).
14. Hassold, T.J., Sherman, S. & Hunt, P. Counting cross-overs: characterizing meiotic recombination in mammals. *Hum. Mol. Genet.* **9**, 2409–2419 (2000).
15. Robinson, W. *et al.* Maternal meiosis I nondisjunction of chromosome 15: dependence of the maternal age effect on the level of recombination. *Hum. Mol. Genet.* **7**, 1011–1109 (1998).
16. Hodges, C.A. *et al.* Experimental evidence that changes in oocyte growth influence meiotic chromosome segregation. *Hum. Reproduction.* **17**, 1171–1180 (2002).
17. Hunt, P.A. & Hassold, T.J. Sex matters in meiosis. *Science* **296**, 2181–2183 (2002).
18. Lamb, N.E. *et al.* Characterization of susceptible chiasma configurations that increase the risk for maternal nondisjuction of chromosome 21. *Hum. Mol. Genet.* **6**, 1391–1399 (1997).
19. Lynn, A. *et al.* Covariation of synaptonemal complex length and mammalian meiotic exchange rates. *Science* **296**, 2222–2225 (2002).
20. Johnson, J., Canning, J., Kaneko, T., Pru, J.K. & Tilly, T.L. Germline stem cells and follicular renewal in the postnatal mammalian ovary. *Nature* **428**, 145–150 (2004).
21. Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits* (Sinauer Associates, Massachusetts, 1998).
22. Dempster, A.P., Laird, N.M. & Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Methodol.* **39**, 1–38 (1977).
23. Gudbjartsson, D.F., Jonasson, K., Frigge, M.L. & Kong, A. Allegro, a new computer program for multipoint linkage analysis. *Nat. Genet.* **25**, 12–14 (2000).
24. McCulloch, C.E. Maximum likelihood algorithms for generalized linear mixed models. *J. Am. Stat. Assoc.* **92**, 162–170 (1997).
25. Broman, K.W. & Weber, J.L. Characterization of human crossover interference. *Am. J. Hum. Genet.* **66**, 1911–1926 (2000).
26. Irwin, M., Cox, N.J. & Kong, A. Sequential imputation for multilocus linkage analysis. *Proc. Natl. Acad. Sci. USA* **91**, 11684–11688 (1994).
27. Quintana, F.A., Liu, J.S. & del Pino, G.E. Monte carlo EM with importance reweighting and its applications in random effects model. *Comput. Stat. Data Anal.* **29**, 429–444 (1999).

To order reprints, please contact:

In the Americas: Tel 212 726 9631; Fax 212 679 0843; reprints@natureny.com
Europe/UK/ROW: Tel + 44 (0)20 7843 4937; Fax + 44 (0)20 7843 4749; reprints@nature.com
Japan & Korea: Tel +81 3 3267 8751; Fax +81 3 3267 8746; reprints@naturejpn.com

Printed by The Sheridan Press

Cancer as a Complex Phenotype: Pattern of Cancer Distribution within and beyond the Nuclear Family

Laufey T. Amundadottir[1*], Sverrir Thorvaldsson[1], Daniel F. Gudbjartsson[1], Patrick Sulem[1], Kristleifur Kristjansson[1], Sigurdur Arnason[1¤1], Jeffrey R. Gulcher[1], Johannes Bjornsson[2], Augustine Kong[1], Unnur Thorsteinsdottir[1], Kari Stefansson[1*]

[1] deCODE Genetics, Reykjavik, Iceland, [2] National University Hospital, Reykjavik, Iceland

ABSTRACT

Background

The contribution of low-penetrant susceptibility variants to cancer is not clear. With the aim of searching for genetic factors that contribute to cancer at one or more sites in the body, we have analyzed familial aggregation of cancer in extended families based on all cancer cases diagnosed in Iceland over almost half a century.

Methods and Findings

We have estimated risk ratios (RRs) of cancer for first- and up to fifth-degree relatives both within and between all types of cancers diagnosed in Iceland from 1955 to 2002 by linking patient information from the Icelandic Cancer Registry to an extensive genealogical database, containing all living Icelanders and most of their ancestors since the settlement of Iceland.

We evaluated the significance of the familial clustering for each relationship separately, all relationships combined (first- to fifth-degree relatives) and for close (first- and second-degree) and distant (third- to fifth-degree) relatives. Most cancer sites demonstrate a significantly increased RR for the same cancer, beyond the nuclear family. Significantly increased familial clustering between different cancer sites is also documented in both close and distant relatives. Some of these associations have been suggested previously but others not.

Conclusion

We conclude that genetic factors are involved in the etiology of many cancers and that these factors are in some cases shared by different cancer sites. However, a significantly increased RR conferred upon mates of patients with cancer at some sites indicates that shared environment or nonrandom mating for certain risk factors also play a role in the familial clustering of cancer. Our results indicate that cancer is a complex, often non-site-specific disease for which increased risk extends beyond the nuclear family.

Competing Interests: JB and SA have declared that no competing interests exist. LTA, ST, DFG, PS, KK, JRG, AK, UT, KS have stock in deCODE Genetics as well as equity interests.

Author Contributions: LTA, ST, DFG, PS, KK, SA, JRG, JB, AK, UT, and KS designed the study. LTA, ST, DFG, PS, AK, and UT analyzed the data. LTA, ST, DFG, PS, KK, SA, JRG, JB, AK, UT, and KS contributed to writing the paper.

Academic Editor: Cathryn Lewis, Guy's King's and St. Thomas' School of Medicine, United Kingdom

Citation: Amundadottir LT, Thorvaldsson S, Gudbjartsson DF, Sulem P, Kristjansson K, et al. (2004) Cancer as a complex phenotype: Pattern of cancer distribution within and beyond the nuclear family. PLoS Med 1(3): e65.

Received: August 27, 2004
Accepted: October 27, 2004
Published: December 28, 2004

DOI: 10.1371/journal.pmed.0010065

Abbreviations: ICR, Icelandic Cancer registry; KC, kinship coefficient; RR, risk ratio

*To whom correspondence should be addressed. E-mail: laufey.amundadottir@decode.is (LTA), kari.stefansson@decode.is (KS)

[¤1]Current address: Keflavik County Hospital, Keflavik, Iceland

Introduction

Highly penetrant susceptibility variants explain only a small fraction of the genetics of all cancer cases. As an example, mutations in the *BRCA1* and *BRCA2* genes account for around 2%–3% of all breast cancer cases [1,2], although more prevalent founder mutations in these genes can explain up to about 10% of the disease in some populations [3,4,5,6,7]. However, the role of genetics in the remaining breast cancer cases and the majority of other cancers is not clear.

Family studies have given insight into the contribution of genetic and environmental factors to the etiology of cancer. Case-control, registry- and population-based studies have evaluated familial clustering using either risk ratio (RR) estimations for relatives of cancer patients, or kinship coefficient (KC) estimations for cancer patients. The largest of these studies, utilizing either the Utah Population and Cancer Registry Database or the Swedish Family-Cancer Database, have demonstrated excess familial clustering at practically all cancer sites in the body [8,9,10,11,12]. Most of these studies have been able to evaluate familial clustering only within the nuclear family, thus making it more difficult to separate the roles of shared environmental and genetic factors in the familial aggregation of cancers. However, in one of these studies [12], in which familial clustering was evaluated for more distant relatives, significant clustering outside the nuclear family was demonstrated for a number of cancer sites. Extended familial clustering has also been reported in studies of individual cancers [13,14,15,16,17,18,19,20,21,22].

Twin studies have also evaluated the role of genes versus environment in cancer susceptibility. The largest study involved close to 45,000 twins from Denmark, Sweden, and Finland where the RR of same type of cancer was calculated for individuals with affected twins and compared to those without an affected twin [23]. The authors concluded that for the majority of cancer sites only a limited part of the risk could be explained by heritable factors. Exceptions to this were cancers of the prostate, colon and breast.

In addition to well documented familial clustering for the majority of individual cancers, aggregation of different types of cancers in families has also been observed. Reports have been published on the results of systematic analysis of the aggregation of different cancers using the Utah Population and Cancer Registry Database [24,25]. In addition to demonstrating excess familial clustering for most cancer sites, these studies also indicate that an excess is also shared by different cancer sites. In these studies, cancer clustering was evaluated either by calculating the RR for first-degree relatives or KC between different cancer sites. While distant relationships contributed to the overall calculation of KC, their contributions were not evaluated separately in the studies between cancer sites, hence making it more difficult to separate the effects of genetic and environmental factors in these studies.

We have studied a registry of all cancer cases diagnosed in Iceland from 1 January 1955 to 31 December 2002, with the aim of searching for evidence of genetic factors both at individual cancer sites and those shared by different sites. By cross-referencing cancer prevalence in relatives of cases with the aid of a comprehensive nationwide genealogy database, we have estimated RR separately for first- to fifth-degree relatives of all cancer patients diagnosed in Iceland over 48 y. We demonstrate here an increased cancer risk in relatives outside the nuclear family (third- to fifth-degree relatives) for many cancer sites. These relatives share significant genetic makeup but are less likely to share environmental factors beyond those shared by the general population, indicating that genetic factors may be involved. By applying the analysis across different cancer sites we also demonstrate shared familiality between certain cancer sites both in close and distant relatives. These results suggest that cancer can be considered a broad phenotype with shared genetic factors crossing different cancer sites. That is, the difference between cancers at various sites may in part be the consequence of variable expressivity of the same cancer-predisposing genes.

Methods

This study was approved by the National Bioethics Committee of Iceland, the Data Protection Authority of Iceland, and the Icelandic Cancer Society. All names of patients listed in the Icelandic Cancer Registry (ICR) and the genealogic database were encrypted through a process approved by the National Bioethics Committee and the Data Protection Authority before being analyzed [26].

Cancer Registry

The ICR of the Icelandic Cancer Society is a carefully constructed database containing practically complete records of all cancer cases diagnosed in Iceland after 1 January 1955 [27]. Records are received at the ICR from all hospitals in the country that treat cancer patients, and the very few not listed are individuals who are diagnosed while living abroad. Furthermore, the records are verified by a continuous interaction between the ICR and Icelandic hospitals and clinicians. Approximately 95% of cases are histologically verified [28]. In the present study we used International Classification of Disease version 10 codes as the basis for defining phenotypes. A total of 81 unique phenotypes were analyzed. In this paper we present data from 27 sites with more than 200 cases each (Table 1). For the 48 years (1 January 1955 to 31 December 2002) a total of 32,534 individuals were found in our genealogy database. Cancer incidence in Iceland is comparable to the Nordic countries of Europe and is detailed in [27].

Genealogic Database

deCODE Genetics has built a computerized genealogy database of more than 687,500 individuals [29,30]. The names of all 288,000 Icelanders currently alive and a large proportion of all Icelanders who have ever lived in the country are in the database. The genealogy of the entered individuals is recorded from multiple sources including church records and censuses from previous centuries and, more recently, from published genealogy books. The genealogy database is quite complete from the 18th century on, thus allowing quite distant relationships to be traced accurately.

Mates are individuals of the opposite sex who have one or more children in common, regardless of marital status.

Calculations of RRs

The RR for relatives is a measure of the risk of disease for a relative of an affected person compared to the risk in the population as a whole. More precisely, for a given relation-

Table 1. RR Estimates of Cancer at the Same Site for Relatives and Mates for Cancer Sites with 200 or More Cases

Cancer Site	ICD10	Number Affected	RR [90% Confidence Interval]						Combined p Value for Relatives		
			1° Relatives	2° Relatives	3° Relatives	4° Relatives	5° Relatives	Mates	1°-5° Relatives	1° and 2° Relatives	3°-5° Relatives
Breast	C50	3,812	**2.02 [1.88,2.15]**	**1.36 [1.27,1.43]**	**1.21 [1.15,1.25]**	**1.13 [1.08,1.16]**	**1.05 [1.01,1.06]**	2.02 [0.83,5.68]	**<0.00001**[a]	**<0.00001**	**<0.00001**
Prostate	C61	3,380	**1.89 [1.75,2.01]**	**1.36 [1.26,1.45]**	**1.19 [1.13,1.24]**	**1.10 [1.05,1.13]**	**1.10 [1.07,1.12]**	na[b]	**<0.00001**[a]	**<0.00001**	**<0.00001**
Lung	C34	2,904	**2.00 [1.83,2.16]**	**1.39 [1.26,1.50]**	**1.10 [1.03,1.16]**	1.02 [0.97,1.07]	**1.04 [1.01,1.08]**	**1.68 [1.35,2.06]**	**<0.00001**[a]	**<0.00001**	**0.00135**
Stomach	C16	2,890	**1.90 [1.74,2.05]**	**1.38 [1.25,1.48]**	**1.23 [1.16,1.29]**	**1.15 [1.09,1.19]**	**1.09 [1.04,1.11]**	**1.72 [1.33,2.18]**	**<0.00001**[a]	**<0.00001**	**<0.00001**
Colon	C18	2,224	**1.92 [1.71,2.14]**	**1.26 [1.12,1.40]**	**1.16 [1.07,1.24]**	1.05 [0.98,1.10]	**1.06 [1.01,1.09]**	**1.46 [1.03,2.08]**	**<0.00001**[a]	**<0.00001**	**0.00029**
Bladder	C67	1,384	**1.68 [1.39,2.05]**	1.19 [0.98,1.43]	**1.26 [1.13,1.41]**	1.08 [0.99,1.19]	1.03 [0.97,1.10]	0.41 [0.09,1.23]	**<0.00001**[a]	**0.00002**	**0.00033**
Kidney	C64	1,227	**2.30 [1.89,2.80]**	**1.31 [1.06,1.57]**	**1.32 [1.15,1.48]**	**1.15 [1.03,1.26]**	1.03 [0.95,1.09]	1.20 [0.52,2.30]	**<0.00001**[a]	**<0.00001**	**0.00034**
Thyroid	C73	957	**3.02 [2.33,3.85]**	**1.64 [1.29,2.02]**	**1.30 [1.07,1.51]**	1.08 [0.93,1.21]	1.11 [1.00,1.19]	1.12 [0.23,3.93]	**<0.00001**[a]	**<0.00001**	**0.00506**
Pancreas	C25	930	**2.33 [1.83,2.96]**	1.28 [0.97,1.66]	1.09 [0.90,1.29]	1.06 [0.93,1.21]	0.99 [0.91,1.08]	1.29 [0.53,4.08]	**0.00001**[a]	**<0.00001**	0.20677
Ovary	C56	906	**2.01 [1.48,2.70]**	**1.62 [1.27,2.05]**	**1.24 [1.03,1.47]**	**1.18 [1.03,1.33]**	0.92 [0.83,1.01]	na[b]	**<0.00001**[a]	**<0.00001**	**0.04804**
Non-melanoma skin	C44	781	1.46 [0.97,2.07]	0.96 [0.64,1.35]	**1.44 [1.18,1.71]**	1.04 [0.87,1.20]	1.12 [1.00,1.23]	2.16 [0.85,5.73]	**0.00727**	0.17809	**0.00233**
Rectum	C20	767	**1.68 [1.17,2.42]**	**1.63 [1.22,2.15]**	**1.31 [1.08,1.61]**	0.90 [0.76,1.06]	1.00 [0.90,1.12]	**1.63 [0.52,4.83]**	**0.00051**[a]	**0.00034**	0.15784
Endometrium	C54	753	**1.86 [1.31,2.62]**	1.26 [0.90,1.74]	**1.57 [1.30,1.89]**	1.05 [0.89,1.24]	1.12 [1.00,1.24]	na[b]	**<0.00001**[a]	**0.00298**	**0.00018**
Cervix uteri	C53	724	**1.74 [1.12,2.73]**	**1.71 [1.24,2.26]**	1.10 [0.86,1.40]	1.16 [0.98,1.36]	0.99 [0.88,1.12]	na[b]	**0.00053**[a]	**0.00079**	0.11505
Brain	C71	663	1.41 [0.74,2.40]	1.10 [0.69,1.62]	0.83 [0.59,1.10]	**1.31 [1.09,1.54]**	0.99 [0.85,1.11]	1.22 [0.20,4.45]	0.16029	0.18892	0.30278
Melanoma skin	C43	618	**1.86 [1.06,3.35]**	**1.61 [1.08,2.31]**	1.23 [0.90,1.58]	1.00 [0.78,1.19]	0.92 [0.77,1.04]	1.52 [0.19,4.90]	**0.02082**	**0.00606**	0.52545
Esophagus	C15	535	**2.09 [1.30,3.31]**	**1.62 [1.07,2.38]**	1.04 [0.74,1.41]	1.14 [0.91,1.40]	1.07 [0.92,1.23]	2.40 [0.66,5.94]	**0.0015**[a]	**0.00134**	0.16738
Diffuse NHL[c]	C83	422	1.46 [0.65,4.20]	1.19 [0.61,2.11]	1.42 [0.98,2.02]	0.96 [0.70,1.30]	0.92 [0.74,1.13]	2.07 [0.13,5.70]	0.12743	0.17125	0.24982
Multiple myeloma	C90	391	**2.66 [1.54,6.10]**	1.32 [0.69,2.62]	0.57 [0.29,0.99]	1.15 [0.85,1.55]	1.10 [0.89,1.34]	3.65 [0.80,7.84]	**0.03144**	**0.00743**	0.58679
Lymphoid leukemia	C91	368	**4.56 [2.52,8.86]**	1.53 [0.74,2.80]	**1.57 [1.01,2.31]**	1.25 [0.89,1.67]	1.06 [0.82,1.29]	0/281[d]	**0.00004**[a]	**0.00024**	**0.03474**
Myeloid leukemia	C92	342	0.99 [0.23,3.37]	0.98 [0.38,2.01]	1.02 [0.57,1.70]	0.94 [0.62,1.35]	0.99 [0.76,1.26]	0/317[d]	0.54272	0.48934	0.56875
Meninges	C70	291	**3.15 [1.50,6.84]**	1.36 [0.54,4.03]	1.60 [0.94,2.70]	1.47 [1.00,2.06]	1.20 [0.89,1.53]	0/289[d]	**0.00141**[a]	**0.01853**	**0.01051**
Liver	C22	257	2.00 [0.71,5.08]	1.28 [0.41,3.89]	0.64 [0.24,1.35]	1.14 [0.70,1.74]	1.02 [0.73,1.36]	0/263[d]	0.23887	0.13249	0.60908
Lip	C00	244	**5.04 [2.75,9.52]**	0.49 [0.03,2.37]	0.91 [0.40,1.73]	0.63 [0.30,1.12]	**1.41 [1.06,1.80]**	0/220[d]	**0.02346**	**0.0091**	0.48947
Hodgkin's disease	C81	239	**3.27 [1.19,7.01]**	1.85 [0.61,4.84]	0.79 [0.26,1.74]	0.94 [0.50,1.57]	0.79 [0.50,1.16]	0/191[d]	0.11431	**0.03064**	0.81722
Testis	C62	222	**3.52 [1.18,7.37]**	0.99 [0.09,3.37]	1.81 [0.89,4.78]	**1.86 [1.16,3.04]**	1.21 [0.83,1.73]	na[b]	**0.01076**	0.09278	**0.00647**
Larynx	C32	208	**3.02 [1.06,6.65]**	0.65 [0.05,2.71]	0.27 [0.02,0.90]	1.30 [0.75,2.33]	0.83 [0.53,1.28]	0/244[d]	0.3358	0.14907	0.76038

Shown are the estimated RRs with 90% confidence interval for first- to fifth-degree (1°-5°) relatives and mates of the 27 cancer sites with ≥200 cases, in bold when the 90% CI does not include 1.00, which corresponds to one-sided $p < 0.05$. Also shown are combined p values to evaluate the significance of the increased RR for all relatives (first- to fifth-degree) and for close (first- and second-degree) and distant relatives (third- to fifth-degree). p values nominally significant at the 0.05 level are shown in bold.

[a] Nominal p values that remained significant after Bonferroni correction for the 27 individual tests ($p < 0.00185$).

[b] na, not applicable (sex-specific cancers).

[c] NHL, non-Hodgkin's lymphoma.

[d] Number of mates with cancer/total number of mates for each cancer site.

ICD10, International Classification of Disease version 10.

DOI: 10.1371/journal.pmed.0010065.t001

ship the RR for disease B in the relatives of probands with disease A is defined as

$$RR = \frac{P(R_B|P_A)}{P(R_B)}, \tag{1}$$

where P_A denotes the event that the proband is affected with disease A, and R_B denotes the event that the relative is affected with disease B. Note that disease A and disease B can be the same in this definition. Using Bayes' rule it can be shown that for symmetric relationships, RR is the same if the roles of A and B are switched, i.e., the RR for disease A in the relatives of probands with disease B is the same as the described above. In this study we always chose the less common phenotype as the proband when estimating RR.

A basic underlying assumption in our estimation of RR is that of conditional independence of ascertainment, or censoring, (O_{RB} and O_{PA} are the events that the relative and proband are observed with diseases A and B, respectively):

$$P(O_{RB}, O_{PA}|P_A, P_B) = P(O_{RB}|R_B)P(O_{PA}|P_A). \tag{2}$$

Some form of this assumption is used by most methods estimating RR [31].

Obtaining valid estimates of the RR is not always straightforward, since the method of ascertainment of affected cases critically affects the estimation, and inappropriate estimators can lead to bias or inflated estimates [32]. The use of a nationwide registry of patients covering close to five decades decreases much of the potential sampling bias. However, the ascertainment of the ICR depends on the year of birth of individuals. This dependence needs to be addressed when estimating the RR.

The approach chosen here is to estimate the RR for a number of subpopulations, where prevalence is reasonably constant, and combine them into a single estimate of RR for the full population. Let r be the number of relatives of probands, counting multiple times individuals who are relatives of multiple probands [33], let a be the number of relatives of probands that are affected (again possibly counting the same individual more than once), let n be the size of the population, and finally let x be the number of affected individuals in the population. If $P(R_B)$ and $P(R_B \mid P_A)$ can reasonably be assumed to be constant in the population, then x/n and a/r, respectively, are estimates of these probabilities. Given these estimates, RR is consistently estimated by

$$\frac{a/r}{x/n}. \tag{3}$$

Assuming the population can be split into N subpopulations, such that within each subpopulation $P(R_B)$ and $P(R_B \mid P_A)$ can be assumed to be constant, although they may vary between subpopulations, and assuming furthermore that RR is the same in all the subpopulations, then the RR is consistently estimated by a convex combination of the estimates for the subpopulations. We selected weights for the combination such that the efficiency of the estimator was at maximum for RR equal to one. Making the simplifying assumption that the relatives are independent (while this assumption is not entirely correct, it affects only efficiency, not validity), the optimal weight for group j is

$$w_j = \frac{x_j r_j}{n_j - x_j} \tag{4}$$

(this is the inverse of the variance of the estimate for RR in subpopulation j), where a, r, x, and n are defined as above, restricted to the subpopulation j. Note that probands are not restricted to the subpopulation. Given these weights, our estimate of RR is

$$\frac{\sum_{j=1}^{N} w_j \frac{a_j/r_j}{x_j/n_j}}{\sum_{j=1}^{N} w_j} = \frac{\sum_{j=1}^{N} \frac{a_j n_j}{n_j - x_j}}{\sum_{j=1}^{N} \frac{r_j x_j}{n_j - x_j}}. \tag{5}$$

In this study, the most relevant variations in $P(R_B)$ and $P(R_B \mid P_A)$ stem from time-dependent censoring of affected status and sex-specific differences. Hence, we have stratified the population so that j runs over groups of people of the same sex and born in the same 5-y periods. For a fixed year-of-birth stratum, there is censoring of affected status (missing data) based on year of onset because of the fact that records cover only the period 1955–2002. Our approach is designed to address this type of missing data. As an example of the stratification, the breast cancer patients in our analysis were born in the years 1865 to 1970 (5-y strata), yielding 35 subpopulations, 22 for female patients, but only 13 for male, as this cancer is rare for males.

To assess the significance of the RR obtained for a given group of patients, we compared their observed values with the RR computed for up to 100,000 independently drawn and matched groups of control individuals. Each patient was matched to a single control individual in each control group. The control individuals were drawn at random from the genealogic database with the conditions that they had the same year of birth, the same sex, and the same number of ancestors recorded in the database at five generations back as the matched patients. Empirical p values can be calculated using the control groups; thus, a p value of 0.05 for the RR would indicate that 5% of the matched control groups had values as large as or larger than that for the patient's relatives or spouses. The number of control groups required to obtain a fixed accuracy of the empirical p values is inversely proportional to the p value. We therefore selected the number of control groups generated adaptively up to a maximum of 100,000. When none of the values computed for the maximum number of control groups were larger than the observed value for the patient's relatives and mates, we report the p value as being less than 0.00001. Using a variance-stabilizing square-root transform, an approximate confidence interval may be constructed based on the distribution of RR for control groups [33].

As another test for significance of RR between cancer sites, we used combined estimators for risk in relatives of degree 1 and 2 together, degrees 3, 4, and 5 together, and degrees 1 through 5 together. If RR_d is the RR for relatives of degree d, then $RR_d - 1$ is known to decrease proportional to 2^{-d} as d increases for a monogenetic single variant or additive disease models, and faster for more complex disease models [34]. Denote the estimate of $\hat{RR}_d$, then choose a test statistic of the form

$$\sum_d w_d \frac{(\hat{RR}_d - 1)}{2^d}, \tag{6}$$

with d summed over the relevant degrees. For RR_d close to one, the variance of the estimate $\widehat{RR}_d$ is inversely proportional to the number of relatives of degree d for the proband. Based on the Icelandic genealogy for the cancers being studied here, the number of relatives is proportional to γ^d, where the value of γ quantifies how the number of relatives grows with each degree of relatedness to the proband. This factor γ varies only slightly between cancers and is on average 2.46. Minimizing the variance of the test statistic in equation 6 with respect to the weights yields the statistic

$$\sum_d 1.23^d(\widehat{RR}_d - 1). \qquad (7)$$

As above, the choice of weights and the form of the statistic affects only power, not validity. To assess significance, the observed value of the statistic was compared to its value for multiple matched control groups as described above.

Although our evaluations of familial clustering, for both close and distant relatives, are based on RR, an alternative approach based on comparing KCs among patients and among controls exists [12,24,25]. The two approaches are closely related, and our choice was made in part because relative risk is a less technical concept and its application to genetic counseling more direct. Also, the relationship between relative risk and the power to map disease genes by linkage analysis has been thoroughly investigated [34,35].

Results

We have studied the familial clustering of cancer by estimating RR for first- and up to fifth-degree relatives both within and between all cancer sites. Here we present results for 27 sites that contain 200 or more cancer cases each, based on International Classification of Disease version 10 codes. These 27 sites represent 89% of all cancer cases in the ICR.

Risk Estimations for Cancer at Same Site

A significantly increased RR to first-degree relatives of patients with cancer was seen for 22 of the 27 cancer sites (Table 1). Among the statistically significant RRs, the highest estimates were for lymphoid leukemia, Hodgkin's disease, and cancer of the thyroid, meninges, lip, testis, and larynx (RR above three). These cancers, except for thyroid cancer, were among the least prevalent sites (200–400 cases), as reflected in the large standard deviation of the RR estimates (Table 1). First-degree relatives of individuals with breast,, lung, kidney, pancreatic, ovarian, and esophageal cancer and multiple myeloma, had between 2- and 3-fold increased risk of developing the same cancer.

The medians of the distribution of the estimated RR values in first- to fifth-degree relatives were 2.00, 1.32, 1.21, 1.10, and 1.04, respectively, for the 27 sites.

When calculating a combined p value summarizing the significance of the increased risk for first- to fifth-degree relatives, 21 sites were significant at a nominal p value level of 0.05. Sixteen of those sites remained significant after Bonferroni adjustment for the 27 individual tests (p value < 0.00185) (Table 1). To discriminate between familial clustering in close and distant relatives, combined p values were also calculated for first- and second-degree relatives on one hand and for third- to fifth-degree relatives on the other hand (Table 1). Fourteen sites were nominally significant for the distant relationships (third- to fifth-degree relatives) of which eight were significant after Bonferroni adjustment. These eight sites were all within the group of 16 sites demonstrating significant familial clustering in all relationships.

The RR for developing cancer at the same site was also estimated for mates of cancer patients at 22 out of the 27 individual sites. The remaining five sites are sex-specific and calculations thus not applicable. For seven rare cancer sites, affected mates were not observed, corresponding to a RR of zero. Only lung, stomach, and colon cancer were characterized by significantly increased RR values in mates (Table 1).

Risk Estimations between Cancer Sites

We calculated RR between all cancer sites for first- and up to fifth-degree relatives and mates (results for the 27 largest sites are shown in Table S1). As done for the individual cancer sites, p values were calculated for all (first- to fifth-degree), close (first- and second-degree), and distant (third- to fifth-degree) relationships. Figure 1 shows a diagram representing 20 pairs of cancer sites that associate with a combined p value, significant at a level of 1×10^{-4}, for first- to fifth-degree relationships. This level was significant at the 0.05 level after Bonferroni adjustment for the 351 tests (number of unique pairs of cancers). The strength of the distant familiality (i.e., the p value for third- to fifth-degree relatives) between these pairs of cancers is represented by the thickness of the lines joining sites in Figure 1.

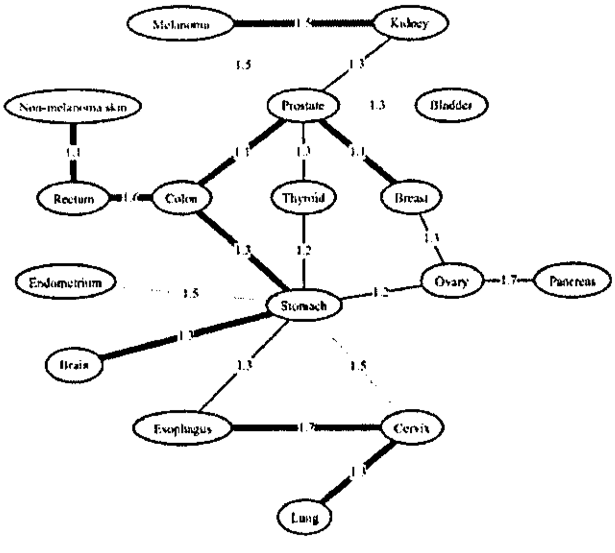

Figure 1. A Schematic Representation of Cancer Pairs Demonstrating Significant Familial Aggregation

Cancer pairs that demonstrate significant familial co-clustering (first- to fifth-degree relatives) at the 0.05 level after adjustment for multiple testing (nominal p value $< 1 \times 10^{-4}$) are joined by lines. The thickness of the lines joining the pairs are based on nominal p values corresponding to the significance of the familiality in distant relatives (third to fifth degree): bold, $p \leq 0.001$; solid, $p \leq 0.01$; and dashed, $p \leq 0.05$. The number on the lines joining each pair indicates the cross-cancer RR in first-degree relatives. Shaded ovals correspond to individual cancer sites that were significant for the combined group of first- to fifth-degree relatives at the 0.05 level after Bonferroni adjustment (see Table 1).

DOI: 10.1371/journal.pmed.0010065.g001

In total, 17 cancer sites were involved in 20 significant pairs of sites (Figure 1). Stomach and prostate cancer were involved in most pairs, seven and six pairs, respectively, followed by colon, ovarian, and cervical cancer, each involved in three pairs. The estimated RRs for the 20 pairs are between 1.1 and 1.7 for first-degree relatives and between 1.1 and 1.5 for second-degree relatives (Figure 1; Table S1). The highest RRs in first-degree relatives between cancer sites were seen for esophagus–cervix, with a RR of 1.74, pancreas–ovary, with a RR of 1.66, and colon–rectum, with a RR of 1.64.

All of the 20 pairs shown in Figure 1 were nominally significant (p value < 0.05) for distant relationships, of which nine were significant at the 0.001 level. In the latter group, prostate, rectum, stomach, and cervical cancers each appeared in two pairs, and colon cancer in three.

Discussion

In this study we have comprehensively analyzed familial aggregation of cancer cases in a whole nation, both within and between pairs of cancer sites. The completeness of our genealogy database allows us to accurately trace distant relationships, which we believe is unique to this study. Linking the ICR to our nationwide genealogy database thus has made it possible to uncover distant familial connections between cancer cases, and reach beyond shared environmental factors to identify individual and combined cancer sites with the strongest genetic influences. Furthermore, even though the genetic effect decreases with more distant relationships, the sample sizes used to estimate familiality are dramatically larger for the distant relationships than for the closer ones. This compensates to some extent for the lower effect and adds considerable statistical power to the study.

In this paper we restrict the presentation and discussion to the most significant findings. However, we provide results for all pairs of 27 cancer sites in Table S1, as a resource for other researchers interested in the familiality of specific cancers.

The largest population-based studies reported to date, evaluating familial clustering within the same cancer site, are from Utah and Sweden [8,9,10]. These studies report RR values for first-degree relatives [36] that are comparable to those presented here for first-degree relatives. For example, the median RRs for the occurrence of the same cancer in first-degree relatives were 2.15, 1.86, and 2.00 for the Utah, the Sweden, and our study, respectively. Also, RR values in first-degree relatives ranged between 1.5 and 3.0 for the majority of sites, i.e., 69%, 82%, and 60%, in Utah, Sweden, and this study, respectively.

As seen in Utah and Sweden, high RR values were found in this study for multiple myeloma, lymphoid leukemia, and thyroid, testicular, and laryngeal cancer. The RR for thyroid cancer in first-degree relatives was much higher in Utah and Sweden (8.48 and 9.51) than in Iceland (3.02). One possible explanation of the lower RR may be the high incidence of thyroid cancer in Iceland, due to an excess of the papillary subtype [18,37], which is not a part of the multiple endocrine neoplasia syndromes.

The cancer sites showing the highest RR for first-degree relatives tend to be among the rarer sites. There are two potential reasons why rare tumors tend to show higher RRs than common cancers. Being common, the baseline frequency is not low and that creates a bound on how large the RR can be. Also, common cancers are expected to be genetically complex, whereas it is more likely for a rare tumor to be closer to a Mendelian trait, caused by rare alleles with high penetrances.

Most individual cancer sites, or 16 out of the 27 studied here, showed familiality as evidenced by significant p values (after adjustment for multiple testing) for the combined group of first- to fifth-degree relatives. Furthermore, eight of these 16 sites remained significant even after exclusion of the first- and second-degree relatives (after adjustment for multiple testing). The majority of the 16 significant cancer sites are among the sites of the most prevalent cancers, indicating that we may lack power to detect extended familiality for the less prevalent cancer sites. Indeed the median number of cases per cancer site was 943 for the 16 significant sites compared to 342 for the non-significant sites. Nevertheless, significant familial clustering (first- to fifth-degree relatives) is seen for some of the less prevalent sites, i.e., lymphoid leukemia and esophagus and meningeal cancer.

The largest cancer twin study reported to date [23] documented significant heritability of prostate (42%), colorectal (35%), and breast cancer (27%) and provided suggestive evidence for limited heritability of leukemia and stomach, lung, pancreas, ovarian, and bladder cancer. All of these cancer sites showed significant familial clustering in our study. However, when the analysis was restricted to distant relatives, lymphoid leukemia, pancreatic, and ovarian cancer were no longer significant. Although close to 45,000 pairs of twins were included in the study (of which 10,803 had been diagnosed with cancer), the study clearly lacked statistical power to detect the effects of heritable factors for the less prevalent cancer sites.

A significantly increased risk of the same cancer was seen in mates only for individuals diagnosed with stomach, lung, or colon cancer. These results are in accordance with previous reports, including Swedish population-based studies, except for colon cancer [38,39,40,41]. Environmental factors in adult life (including lifestyle and infections) or nonrandom mating could explain the higher risk of these cancer types in mates. The RR was not significant or not observed in mates for other sites.

We also assessed the significance of familial clustering between cancer sites by calculating combined p values corresponding to the increased risk for first- to fifth-degree relationships. With this method, we detected 17 cancers that linked into 20 pairs of sites that were significant after adjustment for multiple testing. Stomach and prostate cancer appeared more frequently in the pairs than other cancer types, followed by colon, ovarian, and cervical cancer. We emphasize again, as with the same-cancer calculations, that we might lack power to connect rare cancers to other cancer sites. This possibility is highlighted by the fact that the 17 cancers in the significant pairs are the most prevalent cancer sites in Iceland.

Some connections seen here between cancer sites may be partly explained by known high-risk genes involved in heritable syndromes. Thus, mutations in genes associated with hereditary nonpolyposis colorectal cancers could explain a part of the risk shared between stomach, colon, rectal, and endometrial cancer, and possibly brain and ovarian cancer [42,43]. In a similar manner, mutations in *BRCA1* and

BRCA2 may explain in part the cluster seen between prostate, breast, ovarian, and possibly pancreatic cancer [20,44,45,46]. Other known but even rarer cancer syndromes are likely to explain only a handful of cases.

Undiscovered genetic factors could contribute to some connections seen here to a much greater extent than the known susceptibility factors. Although these could include unknown high-risk susceptibility genes, they are more likely multiple genetic variants, each conferring small to moderate risk.

Familial clusters were identified between cancer sites, both in close and distant relatives, that do not correspond to known cancer syndromes. These include lung, esophageal, cervical, and stomach cancer, which, interestingly, have been associated with environmental rather than genetic factors. One explanation for this excess familiality between these cancer sites is an interaction of genetic susceptibility factors with environmental carcinogens (e.g., tobacco and diet) or infectious agents. Thus, the same environmental factor could interact with the same genetic susceptibility factor or factors to induce different cancers (i.e., smoking in lung and cervical cancer). Alternatively, different environmental factors could interact with the same genetic susceptibility factor or factors to increase the risk for different cancers (i.e., smoking in lung cancer and human papilloma virus in cervical cancer).

Hormone-related cancers form another risk cluster. Thus, shared genetic susceptibility factors could directly influence the hormonal metabolism to induce breast, prostate, thyroid, or ovarian cancer in carriers. Alternatively, shared genetic factors could interact with dietary factors to induce aggregation of cancers at these sites in related individuals. A significantly increased risk of breast, prostate, cervical, and non-melanoma skin cancer was recently reported in first-degree relatives of early-onset breast cancer patients from Sweden that tested negative for *BRCA1* and *BRCA2* mutations [47]. Our data support the notion that unknown susceptibility variants that increase the risk of breast and prostate cancer and melanoma remain to be characterized.

Two more groups of cancers with shared risk were identified that each include sites that share the same developmental progenitors: the prostate, kidney, and bladder are sites derived from the nephrogenic ridge while colon, rectum, and stomach are derived from the primitive gut tube. Therefore, the sites in each group may share risk alleles that regulate embryonic development, which can later play a role in oncogenesis.

Interestingly, three cancer sites/types, non-melanoma skin, brain, and melanoma, that do not have significant same-cancer familial clustering demonstrate significant cross-cancer familial clustering with more prevalent cancer sites, i.e., rectum, stomach, and kidney cancers, respectively.

Previous reports systematically evaluating the significance of co-clustering of cancer pairs in families have utilized the Utah Population Database. In these studies lip and prostate cancers appear to associate most frequently with other cancer sites. The same is true for prostate cancer in our study, whereas lip cancer does not significantly associate with any other cancer sites. This can at least in part be explained by the difference in age-standardized incidence rates for lip cancer in Iceland and Utah (Iceland 1.1 and Utah 2.4) [48]. In contrast, stomach cancer associates with seven other cancer sites out of the 20 significant pairs in our study, but only three other sites in the Utah study. Of the 20 cancer pairs that significantly associate in our study, eight concur with the Utah studies.

Because the increased cross-site RR extends beyond the nuclear family, shared genetic factors may contribute to the risk of more than one cancer type. This suggests that cancer could be considered a broad phenotype with shared genetic factors across cancer sites. Therefore cancer should in certain cases be studied in a broader context than previously done. Combining multiple cancers that show increased cross-site RR may serve to increase the power of linkage and case-control studies. Our results also have implications for genetic counseling and imply that the focus of attention should broaden to the history of multiple cancer types in relatives within and outside the nuclear family. These results also suggest the utility of comparing expression profiles and in vitro biological processes across the cancers that we have identified as sharing genetic risk. The isolation of cancer predisposition genes with broad effects may define new rate-limiting pathways that can be used to search for drug targets for a more focused treatment with fewer side effects but with utility across multiple cancers.

Supporting Information

Table S1. Cross-Site RR Estimates for Relatives and Mates of Patients Diagnosed with Cancers at 27 Sites with 200 Cases or More

Found at DOI: 10.1371/journal.pmed.0010065.st001 (1.9 MB DOC).

Accession Numbers

The LocusLink (http://www.ncbi.nlm.nih.gov/projects/LocusLink/) accession numbers for the genes discussed in this paper are *BRCA1*(LocusLink ID 672) and *BRCA2* (LocusLink ID 675).

Acknowledgments

We thank the ICR for providing lists of cancer patients. We also thank Hersir Sigurgeirsson and Snæbjorn Gunnsteinsson for assistance with data processing, and Jon Thor Bergthorsson, Simon N. Stacey, and Julius Gudmundsson for their critical reading of the manuscript and helpful comments. The study was funded by deCODE Genetics. ■

References

1. Anglian Breast Cancer Study Group (2000) Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases. Anglian Breast Cancer Study Group. Br J Cancer 83: 1301–1308.
2. Syrjakoski K, Vahteristo P, Eerola H, Tamminen A, Kivinummi K, et al. (2000) Population-based study of BRCA1 and BRCA2 mutations in 1035 unselected Finnish breast cancer patients. J Natl Cancer Inst 92: 1529–1531.
3. King MC, Marks JH, Mandell JB (2003) Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. Science 302: 643–646.
4. Johannesdottir G, Gudmundsson J, Bergthorsson JT, Arason A, Agnarsson BA, et al. (1996) High prevalence of the 999del5 mutation in Icelandic breast and ovarian cancer patients. Cancer Res 56: 3663–3665.
5. Thorlacius S, Sigurdsson S, Bjarnadottir H, Olafsdottir G, Jonasson JG, et al. (1997) Study of a single BRCA2 mutation with high carrier frequency in a small population. Am J Hum Genet 60: 1079–1084.
6. Hartge P, Struewing JP, Wacholder S, Brody LC, Tucker MA (1999) The prevalence of common BRCA1 and BRCA2 mutations among Ashkenazi Jews. Am J Hum Genet 64: 963–970.
7. Thorlacius S, Struewing JP, Hartge P, Olafsdottir GH, Sigvaldason H, et al. (1998) Population-based study of risk of breast cancer in carriers of BRCA2 mutation. Lancet 352: 1337–1339.
8. Dong C, Hemminki K (2001) Modification of cancer risks in offspring by sibling and parental cancers from 2,112,616 nuclear families. Int J Cancer 92: 144–150.
9. Vaittinen P, Hemminki K (1999) Familial cancer risks in offspring from discordant parental cancers. Int J Cancer 81: 12–19.
10. Goldgar DE, Easton DF, Cannon-Albright LA, Skolnick MH (1994) Systematic population-based assessment of cancer risk in first-degree relatives of cancer probands. J Natl Cancer Inst 86: 1600–1608.
11. Czene K, Lichtenstein P, Hemminki K (2002) Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. Int J Cancer 99: 260–266.

12. Cannon-Albright LA, Thomas A, Goldgar DE, Gholami K, Rowe K, et al. (1994) Familiality of cancer in Utah. Cancer Res 54: 2378–2385.
13. Pharoah PD, Day NE, Duffy S, Easton DF, Ponder BA (1997) Family history and the risk of breast cancer: A systematic review and meta-analysis. Int J Cancer 71: 800–809.
14. Ziogas A, Gildea M, Cohen P, Bringman D, Taylor TH, et al. (2000) Cancer risk estimates for family members of a population-based family registry for breast and ovarian cancer. Cancer Epidemiol Biomarkers Prev 9: 103–111.
15. Tulinius H, Egilsson V, Olafsdottir GH, Sigvaldason H (1992) Risk of prostate, ovarian, and endometrial cancer among relatives of women with breast cancer. BMJ 305: 855–857.
16. Tulinius H, Olafsdottir GH, Sigvaldason H, Tryggvadottir L, Bjarnadottir K (1994) Neoplastic diseases in families of breast cancer patients. J Med Genet 31: 618–621.
17. Tulinius H, Sigvaldason H, Olafsdottir G, Tryggvadottir L, Bjarnadottir K (1999) Breast cancer incidence and familiality in Iceland during 75 years from 1921 to 1995. J Med Genet 36: 103–107.
18. Hrafnkelsson J, Tulinius H, Jonasson JG, Sigvaldason H (2001) Familial non-medullary thyroid cancer in Iceland. J Med Genet 38: 189–191.
19. Gudbjartsson T, Jonasdottir TJ, Thoroddsen A, Einarsson GV, Jonsdottir GM, et al. (2002) A population-based familial aggregation analysis indicates genetic contribution in a majority of renal cell carcinomas. Int J Cancer 100: 476–479.
20. Tulinius H, Olafsdottir GH, Sigvaldason H, Arason A, Barkardottir RB, et al. (2002) The effect of a single BRCA2 mutation on cancer in Iceland. J Med Genet 39: 457–462.
21. Imsland AK, Eldon BJ, Arinbjarnarson S, Egilsson V, Tulinius H, et al. (2002) Genetic epidemiological aspects of gastric cancer in Iceland. J Am Coll Surg 195: 181–186.
22. Eldon BJ, Jonsson E, Tomasson J, Tryggvadottir L, Tulinius H (2003) Familial risk of prostate cancer in Iceland. BJU Int 92: 915–919.
23. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, et al. (2000) Environmental and heritable factors in the causation of cancer—Analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med 343: 78–85.
24. Thomas A, Cannon-Albright L, Bansal A, Skolnick MH (1999) Familial associations between cancer sites. Comput Biomed Res 32: 517–529.
25. Gholami K, Thomas A (1994) A linear time algorithm for calculation of multiple pairwise kinship coefficients and the genetic index of familiality. Comput Biomed Res 27: 342–350.
26. Gulcher JR, Kristjansson K, Gudbjartsson H, Stefansson K (2000) Protection of privacy by third-party encryption in genetic research in Iceland. Eur J Hum Genet 8: 739–742.
27. Moller B, Fekjaer H, Hakulinen T, Tryggvadottir L, Storm HH, et al. (2002) Prediction of cancer incidence in the Nordic countries up to the year 2020. Eur J Cancer Prev 11: S1–S96.
28. Jonasson JG, Tryggvadottir LT, Bjarnadottir K, Olafsdottir GH, Olafsdottir EJ, et al. (2002) Iceland. In: Parkin DM, SL Whelan SL, Ferlay J, Teppo L, Thomas DB, editors. Cancer incidence in five continents, Volume VIII (IARC Scientific Publications No. 143). Lyon: International Agency for Research on Cancer. pp. 354–355.
29. Gulcher J, Stefansson K (1998) Population genomics: Laying the groundwork for genetic disease modeling and targeting. Clin Chem Lab Med 36: 523–527.
30. Gulcher J, Kong A, Stefansson K (2001) The genealogic approach to human genetics of disease. Cancer J 7: 61–68.
31. Wallace C, Clayton D (2003) Estimating the relative recurrence risk ratio using a global cross-ratio model. Genet Epidemiol 25: 293–302.
32. Guo SW (1998) Inflation of sibling recurrence-risk ratio, due to ascertainment bias and/or overreporting. Am J Hum Genet 63: 252–258.
33. Sveinbjornsdottir S, Hicks AA, Jonsson T, Petursson H, Gugmundsson G, et al. (2000) Familial aggregation of Parkinson's disease in Iceland. N Engl J Med 343: 1765–1770.
34. Risch N (1990) Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46: 222–228.
35. Risch N (1990) Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 46: 229–241.
36. Risch N (2001) The genetic epidemiology of cancer: Interpreting family and twin studies and their implications for molecular genetic approaches. Cancer Epidemiol Biomarkers Prev 10: 733–741.
37. Jonasson JG, Hrafnkelsson J, Bjornsson J (1989) Tumours in Iceland. 11. Malignant tumours of the thyroid gland. A histological classification and epidemiological considerations. Apmis 97: 625–630.
38. Hemminki K, Dong C, Vaittinen P (2001) Cancer risks to spouses and offspring in the Family-Cancer Database. Genet Epidemiol 20: 247–257.
39. Hemminki K, Jiang Y (2002) Cancer risks among long-standing spouses. Br J Cancer 86: 1737–1740.
40. Hemminki K, Chen B (2004) Familial risk for colorectal cancers are mainly due to heritable causes. Cancer Epidemiol Biomarkers Prev 13: 1253–1256.
41. Mellemgaard A, Jensen OM, Lynge E (1989) Cancer incidence among spouses of patients with colorectal cancer. Int J Cancer 44: 225–228.
42. Umar A, Boland CR, Terdiman JP, Syngal S, de la Chapelle A, et al. (2004) Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. J Natl Cancer Inst 96: 261–268.
43. Oliveira Ferreira F, Napoli Ferreira CC, Rossi BM, Toshihiko Nakagawa W, Aguilar S Jr, et al. (2004) Frequency of extra-colonic tumors in hereditary nonpolyposis colorectal cancer (HNPCC) and familial colorectal cancer (FCC) Brazilian families: An analysis by a Brazilian hereditary colorectal cancer institutional registry. Fam Cancer 3: 41–47.
44. Ford D, Easton DF, Bishop DT, Narod SA, Goldgar DE (1994) Risks of cancer in BRCA1-mutation carriers. Breast Cancer Linkage Consortium. Lancet 343: 692–695.
45. Breast Cancer Linkage Consortium (1999) Cancer risks in BRCA2 mutation carriers. The Breast Cancer Linkage Consortium. J Natl Cancer Inst 91: 1310–1316.
46. Baffoe-Bonnie AB, Kiemeney LA, Beaty TH, Bailey-Wilson JE, Schnell AH, et al. (2002) Segregation analysis of 389 Icelandic pedigrees with breast and prostate cancer. Genet Epidemiol 23: 349–363.
47. Loman N, Bladstrom A, Johannsson O, Borg A, Olsson H (2003) Cancer incidence in relatives of a population-based set of cases of early-onset breast cancer with a known BRCA1 and BRCA2 mutation status. Breast Cancer Res 5: R175–R186.
48. Parkin DM, Whelan SL, Ferlay J, Raymond L, Young J, editors (1997) Cancer incidence in five continents, Volume VII (IARC Scientific Publications No. 143). Lyon: International Agency for Research on Cancer. 1,240 p.

Patient Summary

Background Although a few cancers have a fairly simple genetic cause, most, especially the most common cancers, do not, and what makes one person rather than another develop cancer is not clear. One way of trying to work out how much genes rather than environment contribute to disease is to study large populations. One such population is the entire Icelandic nation, for which not only are genetic data available, but there is also very good information on family relationships and health, especially cancers.

What Did the Study Find? Researchers examined all cancer records dating back to 1955 and then analyzed the chances of relatives and mates of these patients having cancer. They found that some cancers, especially rare ones, had a higher than baseline chance of occurring in relatives, but so did many common cancers, and for some cancers, the higher chances extended to quite distant relatives. In addition, the risk sometimes involved different cancer types.

What Does the Study Mean for Patients? Even for the highest risk cancers, the absolute increased risk for relatives remains very small. In addition, despite the large numbers of patients studied, the numbers of cancer cases are still not large enough to be completely certain of the results, apart from very common cancers, which had the lowest chance of occurring in relatives. So these results will not help doctors much at the present time in telling an individual patient what their risk is of getting cancer if a relative has it—but they will be useful for other researchers in knowing how to plan future studies to look at the underlying causes of cancer.

Where Can I Get More Information? Icelandic Cancer Society: http://www.krabb.is/cancer/
The United States National Cancer Institute's Cancer Information Service: http://cis.nci.nih.gov/
CancerHelp UK, a free information service about cancer and cancer care: http://www.cancerhelp.org.uk/
deCODE Genetics:http://www.decode.com

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 0.5cM. (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.

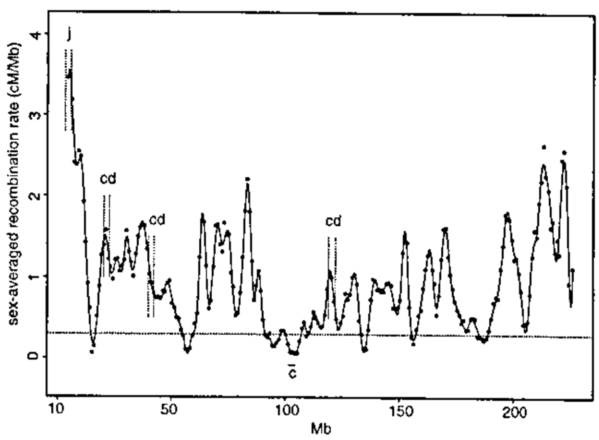

Fig. 1 Sex-averaged recombination rate for chromosome 3. Points correspond to sex-averaged crossover rates, calculated using moving windows 3 Mb in width; the shift from the center of one bin to the next is 1 Mb. The sex-averaged genetic distance for each 3-Mb window was calculated on the basis of our 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.

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

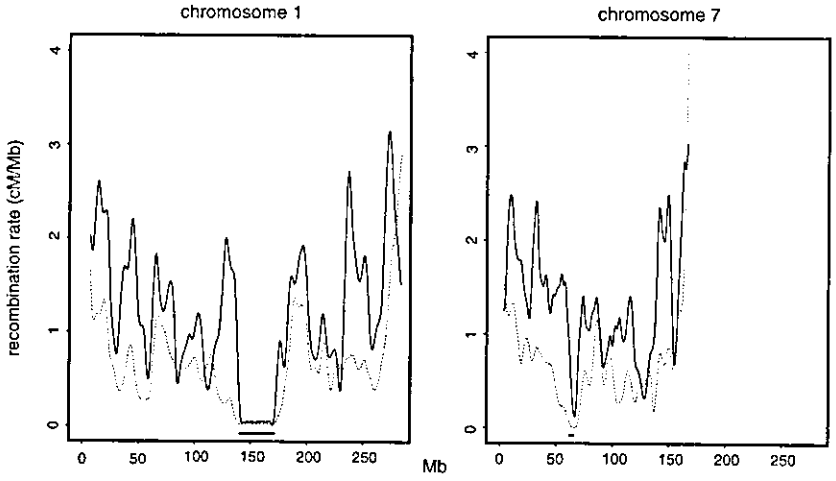

Fig. 2 Sex-specific recombination rates for chromosomes 1 and 7. Solid line, female; dashed line, male.

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).

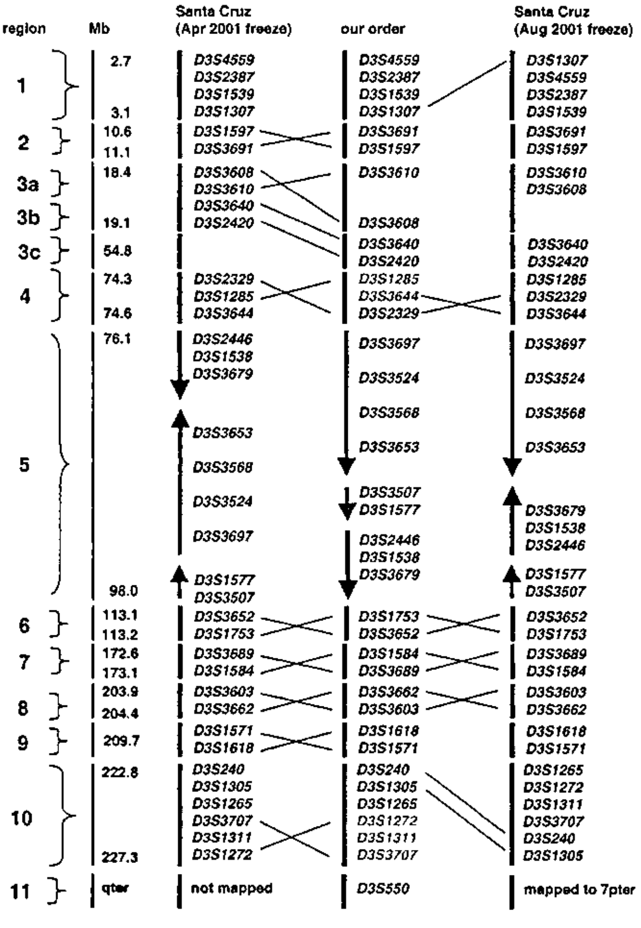

Fig. 3 Comparisons of our order and the Santa Cruz orders for the April and August 2001 freezes of the draft sequence. All discrepancies between our order and these freezes for chromosome 3 are shown. Physical positions of the genomic segments in the figure are indicated either by giving the starting 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).

nature genetics

A common inversion under selection in Europeans

Hreinn Stefansson[1,3], Agnar Helgason[1,3], Gudmar Thorleifsson[1], Valgerdur Steinthorsdottir[1], Gisli Masson[1], John Barnard[2], Adam Baker[1], Aslaug Jonasdottir[1], Andres Ingason[1], Vala G Gudnadottir[1], Natasa Desnica[1], Andrew Hicks[1], Arnaldur Gylfason[1], Daniel F Gudbjartsson[1], Gudrun M Jonsdottir[1], Jesus Sainz[1], Kari Agnarsson[1], Birgitta Birgisdottir[1], Shyamali Ghosh[1], Adalheidur Olafsdottir[1], Jean-Baptiste Cazier[1], Kristleifur Kristjansson[1], Michael L Frigge[1], Thorgeir E Thorgeirsson[1], Jeffrey R Gulcher[1], Augustine Kong[1,3] & Kari Stefansson[1,3]

A refined physical map of chromosome 17q21.31 uncovered a 900-kb inversion polymorphism. Chromosomes with the inverted segment in different orientations represent two distinct lineages, H1 and H2, that have diverged for as much as 3 million years and show no evidence of having recombined. The H2 lineage is rare in Africans, almost absent in East Asians but found at a frequency of 20% in Europeans, in whom the haplotype structure is indicative of a history of positive selection. Here we show that the H2 lineage is undergoing positive selection in the Icelandic population, such that carrier females have more children and have higher recombination rates than noncarriers.

Though important for evolution, large chromosomal rearrangements such as deletions, duplications and inversions are generally thought to be deleterious. These large-scale polymorphisms contribute substantially to genomic variation among humans and account for much of the genomic difference between humans and other primates[1–3]. The architecture of large inversion polymorphisms suggests they may occur through nonallelic homologous recombination assisted by low-copy repeats positioned in the genome in an inverted orientation[4,5].

Genotype analysis may show whether a segment is duplicated or deleted, by means of an apparent gain or loss of heterozygosity, respectively. In contrast, inversions are difficult to detect, particularly those of moderate size. Genotypes of markers inside inverted regions are consistent among relatives and, unless inversions are several megabases in size, they are not easily detected with standard cytogenetic assays. A few large and common inversion polymorphisms have been detected in the human genome, the most notable being a large inversion on chromosome 8p (ref. 5). These may be only the tip of the iceberg.

Here we describe, for the first time to our knowledge, a 900-kb inversion polymorphism at 17q21.31, a region that contains several genes, including those encoding corticotropin releasing hormone receptor 1 (*CRHR1*) and microtubule-associated protein tau (*MAPT*). Previous studies have characterized two highly divergent *MAPT* haplotypes, H1 and H2, and noted the existence of strong linkage disequilibrium (LD) across a 1.6-Mb region containing the gene[6–12]. We provide a detailed description of the unusual haplotype structure in this inverted region, evaluate the impact of natural selection in the past and present, and discuss the implications for our understanding of human evolutionary history.

RESULTS

Discovery of a 900-kb inversion polymorphism

The Build 34 assembly of chromosome 17q21.31 is chimeric, constructed to a large extent from clones representing different *MAPT* haplotypes of type H1. We used a set of chromosome-specific BAC contigs to show that there is a 900-kb inversion polymorphism in this region (Fig. 1). We generated the chromosome-specific assembly from RP11 BAC clones (Roswell Park Cancer Institute Human BAC Library) originating in a DNA sample from one individual. The two RP11 chromosomes represent *MAPT* haplotypes of type H1 and H2 based on the characteristic alleles for a dinucleotide marker in intron nine[6] (DG17S142) and a characteristic 238-bp deletion in the same intron on the H2 background[6]. By genotyping RP11 clones from 17q21.31 for 60 microsatellite markers and assembling the clones into chromosome-specific contigs, we found that the H2 haplotype was structurally different from the Build 34 assembly. The segment from 44.1 to 45.0 Mb is inverted on the H2 background compared with the H1 background and Build 34. Furthermore, a 127-kb tandem duplication containing exons 1–13 of the N-ethylmaleimide-sensitive factor gene (*NSF*) is located upstream of a full-length copy of *NSF* in the H1 variant in Build 34. The same *NSF* exons are also duplicated on the H2 chromosome, but the H2 duplication is larger (280 kb), spanning the H1 duplication and extending to the 5′ end of the gene *LOC284058*. The two *NSF* copies (the partial copy with exons 1–13 and the full-length copy) are separated by only 100 kb on the H1 chromosome in the RP11 library, whereas on the H2 chromosome in the RP11 library, the partial copy of *NSF* is inverted and located 1 Mb upstream of the full-length copy of *NSF*.

[1]deCODE Genetics, Sturlugata 8, 101 Reykjavik, Iceland. [2]Department of Biostatistics and Epidemiology, Cleveland Clinic Foundation, Cleveland, Ohio, USA. [3]These authors contributed equally to this work. Correspondence should be addressed to A.K. (kong@decode.is) or K.S. (kari.stefansson@decode.is).

Published online 16 January 2005; doi:10.1038/ng1508