The Signatures of Autozygosity among Patients with Colorectal Cancer

Manny D. Bacolod,1 Gunter S. Schemmann,7 Shuang Wang,2 Richard Shatlock,1 Sarah F. Giardina,1 Zhaoshi Zeng, Jiru Shia, Robert F. Stengel,1 Norman Gerry,7 Josephine Hoh,10 Tomasz Kirchhoff,5 Bert Gold,1 Michael F. Christman,10 Kenneth Offit,5 William L. Gerald,1 Daniel A. Netterman,4 Jurg Ott,6 Philip B. Paty,3 and Francis Barany1

1Department of Microbiology, Weill Medical College of Cornell University; 2Department of Biostatistics, Mailman School of Public Health, Columbia University; Departments of 3Surgery, 4Pathology, and 5Medicine, Memorial Sloan-Kettering Cancer Center; 6Laboratory of Statistical Genetics, Rockefeller University, New York, New York; 7School of Engineering and Applied Science and 8Department of Molecular Biology, Princeton University, Princeton, New Jersey; 9Laboratory of Genomic Diversity, National Cancer Institute at Frederick, Maryland; 10Department of Genetics and Genomics, Boston University, Boston, Massachusetts; and 11School of Public Health, Yale University, New Haven, Connecticut

Abstract

Previous studies have shown that among populations with a high rate of consanguinity, there is a significant increase in the prevalence of cancer. Single nucleotide polymorphism (SNP) array data (Affymetrix, 50K Xba1) analysis revealed long regions of homozygosity in genomic DNAs taken from tumor and matched normal tissues of colorectal cancer (CRC) patients. The presence of these regions in the genome may indicate levels of consanguinity in the individual’s family lineage. We refer to these autozygous regions as identity-by-descent (IBD) segments. In this study, we compared IBD segments in 74 mostly Caucasian CRC patients (mean age of 66 years) to two control data sets: (a) 146 Caucasian individuals (mean age of 80 years) who participated in an age-related macular degeneration (AMD) study and (b) 118 cancer-free Caucasian individuals from the Framingham Heart Study (mean age of 67 years). Our results show that the percentage of CRC patients with IBD segments (2.4 Mb length and 50 SNPs probed) in the genome is at least twice as high as the AMD or Framingham control groups. Also, the average length of these IBD regions in the CRC patients is more than twice the length of the two control data sets. Compared with control groups, IBD segments are found to be more common among individuals of Jewish background. We believe that these IBD segments within CRC patients are likely to harbor important CRC-related genes with low-penetrance SNPs and/or mutations, and, indeed, two recently identified CRC predisposition SNPs in the 8q24 region were confirmed to be homozygous in one particular patient carrying an IBD segment covering the region. [Cancer Res 2008;68(8):2610–21]

Introduction

Colorectal cancer (CRC) is one of the four most prevalent cancers in the United States. In 2007, there will be 153,760 new cases of CRC in the United States, resulting in 52,180 deaths (1). According to a recent worldwide statistical compilation, over a million people suffered from the disease in 2002, with the majority of cases in industrialized countries (2). Genetics aside, the incidence of CRC correlates with diets rich in fat and calories, and low in vegetables, fruits, and fibers as well as alcohol consumption and smoking (3). Traditionally, CRC cases are divided into two categories: sporadic and familial (or hereditary; ref. 4). Approximately 70% of the cases are classified as sporadic, afflicting people with apparently no family history of the disease. Of the familial cases, the two most commonly occurring are familial adenomatous polyposis (FAP) and hereditary nonpolyposis CRC (HNPCC). FAP, characterized by formation of polyps within the gastrointestinal tracts of affected individuals, is caused by highly penetrant, autosomal dominant germ line mutations in the adenomatous polyposis coli (APC) gene, and can account for ~1% of all CRCs (5). HNPCC (Lynch syndrome) cases, seen in as many as 2.5% of all CRCs, are caused by highly penetrant mutations in DNA mismatch repair genes (primarily MLH1 and MSH2; ref. 6). Much less common genetically linked CRCs are those arising from hamartomatous polyp syndromes such as juvenile polyposis, Peutz-Jeghers, and Cowden’s, which are caused by mutations in SMAD4 (7), STK11 (8), and PTEN (9), respectively. However, the exact genetic causes of a great percentage of familial CRCs remain undiscovered and likely due to low penetrating alleles. Moreover, the distinction between spontaneous and familial CRCs may be understated. Some cancers classified as sporadic cases may in fact have underlying genetic components (4, 10). Several statistical analyses of huge cancer databases have attempted to quantify the heritable components of cancers. The cohort studies from Sweden (11) and Utah (12) showed that the CRC family risk ratios, which is a direct measure of heritability (13), are 4.41 (considered high) and 2.54 (considered moderate), respectively. In addition, the Scandinavian twin study (involving a little less than 45,000 pairs of twins) showed that hereditary factors affect colon cancer 35% of the time (14). Therefore, a significant proportion of heritable CRC remains unaccounted for.

Our research group aims to characterize CRCs using a variety of molecular techniques, including expression profiling (15, 16), methylation profiling (17), mutational scanning (18, 19), and single nucleotide polymorphism (SNP) array-based chromosomal analysis (15). The latter technique (Affymetrix Human Mapping array) readily reveals cancer tissue chromosomal aberrations such as amplifications and loss of heterozygosity (LOH). Indeed, we initially set out to identify regions of varying amplification and to determine if any correlation existed between these chromosomal...
aberrations and expression data generated from the same samples. When we began to examine the results of individual patients comparing their copy number and LOH between the tumor and the matched normal, we often saw chromosomal gains and losses in the tumor but not in the normal tissue (Fig. 1A). Surprisingly, we often identified samples where the regions of high homozygosity [identity-by-descent (IBD) segments] were found in both the tumor and the corresponding normal mucosa (Fig. 1B). These homozygous segments are most probably indicators of an individual's autozygosity—an indication that parents share a common ancestor.

**Figure 1.** A, SNP array (Affymetrix Xba 240 50K) whole-genome analysis of the colon cancer tissue C0114A and its matching normal mucosa C0114H. The charts (copy number and LOH) were generated by Affymetrix CNAT version 3 (26). The aberrations in C0114A include losses in chromosomes 4, 22, 8p, and parts of 10 and 20p, as well as gains in chromosomes 13 and 20q. The copy number chart indicates deviations from the normal copy number of 2 (baseline of the chart). High LOH values (for the charts, the LOH value is capped at 20), indicated by tall blue bars represent segments in the chromosome of contiguous homozygous SNPs. In the CRC sample, regions of copy loss usually correspond to regions of high LOH. The matching normal (C0114H) indicates neutral copy number (equal populations of red and blue bars only represent noise) throughout the genome. B, SNP array (Affymetrix Xba 240 50K) whole-genome analysis of the colon cancer tissue C0111A and its matching normal mucosa C0111H. Unlike in C0114A (A), the regions of high homozygosity in C0111A can also be found in its corresponding normal mucosa (C0111H). These homozygous segments may in fact be indicator of genomic autozygosity in the patient C0111.
The question was then raised whether these signatures of autozygosity occur more frequently among CRC patients. Several studies have shown that cancer occurs more frequently among groups with higher degrees of consanguinity, that is, groups that share a common ancestor. Among these studies is the comparison of incidence of cancer and other late-onset complex diseases between individuals from genetically isolated islands in middle Dalmatia, Croatia, and a control population (21). The investigators concluded that inbreeding can be a positive predictor for a number of late-onset diseases such as heart disease, stroke, and cancer. The same observations were noted in a Pakistani study where cancer patients, on average, have higher coefficient of inbreeding compared with the general population (22). In another study involving descendants of an Italian immigrant group in Wisconsin, 94% of the subjects with reported adenocarcinomas (mostly colorectal) were products of consanguineous parentage (23). The detrimental effects of inbreeding on cancer is likely more complex than a simple Mendelian genetics, with many more genetic components involved. Nonetheless, studying these genetically isolated populations may eventually lead to discovery of other genes that contribute to cancer predisposition. It is the same argument backed by a growing number of researchers who believe that studying the genetics of purebred dogs known to have high incidence of cancer may eventually help in the discovery of cancer-related genes in humans (24).

Figure 2. The locations of the identified IBD segments (black horizontal bars) among the genomes of the 74 colon cancer patients (A), and the AMD (B) and Framingham (C) control data sets. The threshold limit was set to a minimum of 4-Mb length encompassing at least 50 consecutive homozygous SNPs (but allowing at most 2% errors).

There are a number of ways to measure an individual’s degree of consanguinity (25). Rudan and coworkers (21) used Wright’s path method in measuring average inbreeding coefficients for both the case and control populations. Using short tandem repeat polymorphisms (STRP) as markers, Broman and Weber described the presence of homozygous segments in some individuals from reference families genotyped by Centre d’Etude du Polymorphisme Humain (CEPH). With the advent of high-density human SNP arrays (also known as genotyping arrays), the process of identifying the homozygous segments in the genome has become easier (26). In this study, we show that signatures of autozygosity correlate to CRC incidence and that these IBD regions may be the locations of genes that contribute to CRC heritability.

Materials and Methods

Tissue Acquisition, Sample Selection, and DNA Extraction

Tissue acquisitions followed the protocols of the institutional review boards of Memorial Sloan Kettering Cancer Center (MSKCC) and Cornell University Weill Medical College (institutional review board nos. 0201005297 and 9807003424). Our initial objective was to study the chromosomal aberrations (copy number changes, LOH) in CRCs using high-density SNP mapping arrays (Affymetrix). Based on the pathologist reports (MSKCC), 74 CRC samples showing ≥70% pure tumor cells were chosen for SNP array analysis. Most of these samples were from Caucasian patients (average age...
SNP Array Procedure

The procedure for the Affymetrix GeneChip Human Mapping 50K SNP array was carried out according to the manufacturer’s guidelines. Briefly, 0.25 μg of genomic DNA was digested with XhoI. The digests were then ligated to oligonucleotide adapters, PCR-amplified (such that the amplifications were in the range of 250–2000 bp), fragmented, biotin-labeled, and hybridized to the array for 16 h. Following hybridization, the array chips were washed and then stained with streptavidin-phycoerythrin and a biotinylated anti-streptavidin antibody in the Affymetrix Fluidics Station 450. The arrays were scanned in GeneChip Scanner 3000 to generate the image (DAT) and cell intensity (CEL) files. The CEL files were imported to GeneChip Genotyping Analysis Software 4.0 (GT4.0, Affymetrix) to generate the SNP calls using the dynamic model mapping algorithm (27). It should be noted that the analyses of the current study were undertaken before the release of GTYPE 4.1 and its new Bayesian robust linear model with Mahalanobis distance classifier algorithm.

Use of Other Control Data Sets

To determine the frequency of IBD on the general population, we used the following controls: (a) The age-related macular degeneration (AMD) data set representing the 146 non-Hispanic Caucasian individuals who participated in AMD study (28). These include the 96 cases (mean age 79 ± 5.2 years old) and the 50 controls (mean age 82 ± 2.2 years old). (b) The Framingham data set: 118 Caucasian individuals who are a subset of the National Heart, Lung, and Blood Institute Framingham Heart Study (ages 61–81 years; ref. 29). Clinical data indicate that the individuals in the Framingham data set had no known cancer at the time they participated in the study. Other control data sets used were SNP array data from (a) 30 Ashkenazi Jewish group afflicted with breast cancer (AJBC), (b) 133 Ashkenazi Jewish group with no incidence of cancer (AJNC), and (c) the subgroup of 48 Caucasian individuals in the Affymetrix reference data set. The last control group did not have any available clinical information. Detailed analyses of the Ashkenazi Jewish data sets are to be described elsewhere. We also examined the possibility of population stratification using the EIGENSTRAT method (30).

Verification of SNPs by Direct DNA Sequencing

Tumors containing IBD segments covering SNPs that are recently associated with colon cancer and Crohn’s disease were subjected to dideoxy-sequencing to genotype the associated SNP and to verify homozygosity at that region. DNA sequencing was performed using the Applied Biosystems Automated 3730 DNA Analyzer, along with Big Dye Terminator chemistry and AmpliTaq-FS DNA Polymerase (Applied Biosystems). Universal primers (forward: 5’-CGTACGCCAGACGAA-3’ and reverse: 5’-CTGACGACGAAAC-3’) were used for the sequencing and the following DNA-specific primers were used to amplify the DNA segment covering the SNP in question: rs9469220 (forward: 5’-AATGTCAGCCTGACTGCACTGGA-3’ and reverse: 5’-CTGGGACACAAGGGATTT-GACTGTC-3’), rs11805303 (forward: 5’-AGTGCACTCGTGTAATCAATA-3’ and reverse: 5’-AGTTGTTCCAGGTCGTTATAC-3’), rs11888365 (forward: 5’-AGTGCACTCGTGTAATCAATA-3’ and reverse: 5’-AGTTGTTCCAGGTCGTTATAC-3’), rs11888365 (forward: 5’-AGTGCACTCGTGTAATCAATA-3’ and reverse: 5’-AGTTGTTCCAGGTCGTTATAC-3’), rs11888365 (forward: 5’-AGTGCACTCGTGTAATCAATA-3’ and reverse: 5’-AGTTGTTCCAGGTCGTTATAC-3’).

Method 2: logit model of the odds calculation. Another statistical method we used in identifying the IBD segments is an extension of the Broman and Weber approach (20), in which the autozygosity logit of the odds (LOD, base 10) score for a 5-Mb segment (ranging from SNP position j to SNP position k) in the genome was calculated. As defined in that article,

\[
LOD(j, k) = \sum_{i=0}^{k-j} \log R_i
\]

where \( P(g_i|\text{autozygous at } i) \) refers to the probability of the observed genotype \( g_i \) at the \( i \)th position in the genome, given that the \( i \)th position is autozygous, whereas \( P(g_i|\text{not autozygous at } i) \) refers to the probability of the observed genotype \( g_i \) at the \( i \)th position in the genome, given that the \( i \)th position is not autozygous. If the SNP call (genotype) at position \( i \) is AA or BB, then

\[
R_i = (1 - e^{-z}) / P_A + e^{-z}
\]

and if the genotype is AB, then

\[
R_i = e^{-z}
\]

where \( z \) denotes the combined rate of genotyping error and mutations (maximum of 2%). \( P_A \) and \( P_B \) are the frequencies of alleles A and B, respectively, in the study group (i.e., CRC or control population separately). Described in detail in a separate manuscript, the algorithm used a sliding window method (5 Mb from one end of a chromosome to the other with 0.5 Mb step size) to form segments along the genome.
The PCR reaction (25 μL) contained 20 mmol/L Tricine (pH 8.7), 16 mmol/L (NH₄)₂SO₄, 2.5 mmol/L MgCl₂, 0.2 mmol/L deoxynucleotide triphosphate, 0.2 μmol/L of each gene-specific primer, 2.5 units of AmpliTaq Gold DNA polymerase, and 100 ng of genomic DNA. Thermocycling conditions were as follows: 95°C for 10 min to activate AmpliTaq Gold polymerase; followed by 25 cycles of 94°C for 30 s, 60°C for 1 min, 72°C for 1 min; followed by a final extension step at 72°C for 30 min.

Table 1. Partial list of the IBD segments in the CRC patients

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Patient ID</th>
<th>IBD segments (all)</th>
<th>Homozygosity analysis (method 1; Mb-Mb)</th>
<th>Same region as identified by autozygosity analysis (method 2; Mb-Mb)</th>
<th>Length* (Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>C0111</td>
<td>14.837–41.8932</td>
<td>15.4744–41.9744</td>
<td>27.0562</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>C0111</td>
<td>51.3465–84.5851</td>
<td>51.9744–84.9744</td>
<td>33.246</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>C0253</td>
<td>68.1947–76.2018</td>
<td>66.9744–76.4744</td>
<td>8.0069</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>C0111</td>
<td>142.4068–157.5006</td>
<td>144.9744–159.9744</td>
<td>15.0945</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>02308</td>
<td>97.1053–118.0407</td>
<td>98.6008–119.1008</td>
<td>20.9355</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>C0221</td>
<td>191.3785–205.8568</td>
<td>191.1008–207.1008</td>
<td>14.4783</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>C0221</td>
<td>218.8516–228.8536</td>
<td>218.1008–229.6008</td>
<td>10.0021</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>C0181</td>
<td>94.2878–108.8576</td>
<td>93.2605–110.2605</td>
<td>14.5697</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>C0111</td>
<td>12.3928</td>
<td>12.3928</td>
<td>12.3928</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>C0111</td>
<td>14.8368–148.2447</td>
<td>131.7605–149.2605</td>
<td>16.8061</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>C0181</td>
<td>148.1866–158.304</td>
<td>148.2605–158.7605</td>
<td>10.174</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>C0221</td>
<td>70.9658–80.6667</td>
<td>73.1512–81.6512</td>
<td>7.9099</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>C0153</td>
<td>120.9134–136.3389</td>
<td>120.6512–138.1512</td>
<td>15.6255</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>C0111</td>
<td>2.7994–17.945</td>
<td>2.7994–17.945</td>
<td>15.1456</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>C0111</td>
<td>7.46258–91.0269</td>
<td>7.3608–90.6806</td>
<td>16.4012</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>C0153</td>
<td>126.8113–136.872</td>
<td>125.1806–137.6806</td>
<td>10.0607</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>C0111</td>
<td>2.5937–12.4867</td>
<td>2.5937–12.4867</td>
<td>12.3928</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>C0111</td>
<td>121.2258–130.8439</td>
<td>121.2394–130.7394</td>
<td>9.6181</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>C0111</td>
<td>0.0939–12.4867</td>
<td>0.0939–12.4867</td>
<td>12.3928</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>C0111</td>
<td>42.3244–59.9483</td>
<td>42.5937–60.0937</td>
<td>17.625</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>C0111</td>
<td>71.2942–92.3212</td>
<td>71.5937–92.3212</td>
<td>21.1821</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>C0111</td>
<td>33.6001–42.4575</td>
<td>33.8211–43.3211</td>
<td>8.9475</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>02308</td>
<td>40.2645–57.9614</td>
<td>42.4512–57.9512</td>
<td>17.6968</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>C0153</td>
<td>45.2834–55.8778</td>
<td>44.1999–56.1499</td>
<td>10.5944</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>C0192</td>
<td>50.5348–68.1576</td>
<td>50.6499–68.4999</td>
<td>17.6228</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>C0111</td>
<td>58.4454–73.4044</td>
<td>58.1499–74.1499</td>
<td>14.959</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>C0111</td>
<td>29.31–54.4835</td>
<td>27.5937–53.5937</td>
<td>25.1735</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>C0161</td>
<td>20.7118–31.7663</td>
<td>20.0748–32.0748</td>
<td>11.0544</td>
<td></td>
</tr>
</tbody>
</table>

NOTE: Shown are IBD segments of at least 8 Mb in length. The complete list is found on Supplement S-B7. The segments are identified through both homozygosity (method 1) and LOD calculation (method 2) analyses (described in Materials and Methods).

*The length of the IBD segment identified through homozygosity analysis.
†The average LOD of the segments determined by autozygosity analysis.
‡The number of SNPs covered by the IBD region identified through homozygosity analysis.
§Portion of IBD segment (homozygosity analysis) of at least 4 Mb in length and not overlapping any IBD segment found in the two control data sets (AMD and Framingham).
¶Portion of IBD segment (method 1) of at least 4 Mb in length and not overlapping any IBD segment found in only the AMD control data set.
*Portion of IBD segment found to be specific to CRC patients when compared with the AMD controls using LOD calculations (method 2).
Results

**IBD segments as extended runs of homozygous SNPs.** The first approach to score regions of autozygosity (method 1) was to isolate the long stretches of homozygosity in the genomic DNAs taken from mostly noncancerous tissues (explained in Materials and Methods). We set the threshold limit to a minimum of 4-Mb length encompassing at least 50 consecutive homozygous SNPs, but allowing at most 2% heterozygous SNPs. These identified IBD regions (on chromosomes 1 to 22) are indicated as bars in Fig. 2 (A, CRC patients; B, AMD controls; C, Framingham controls). A partial list (at least 8 Mb in length) of these IBD segments are shown in Table 1. The longest IBD segment is the chromosome 1 region from 51,3405 to 84,5851 Mb found in patient C0111, who has a total of 271.6 Mb of homozygous segments distributed in 19 IBD segments (see Supplements S-B3, S-B4, and S-B5) for the total IBD segment lengths of every CRC patient, as well as all the AMD, and Framingham control individuals. As shown in Table 1, the same IBD regions in chromosome 1 of patient C0111 is also identified by the autozygosity (method 2) analysis, having an average LOD score of 23.5. Within this segment is a region (51,34045–63,01534 Mb) not overlapping with any of the IBD regions in both control data sets. In all, the homozygosity analysis identified a total of 117 IBD segments of at least 4 Mb in length (Supplement S-B7). In another approach (Method 2), a LOD score was calculated to compare the strength of autozygosity versus nonautozygosity for a defined genomic region. This calculation was able to identify the 5 Mb regions (many regions were overlapping, and further inspection identified these autozygous regions as contiguous) in the CRC patients' genome with LOD values of at least 5 (see Supplement S-B8). Of the 34 IBD segments identified by homozygosity analysis

![Table 1. Partial list of the IBD segments in the CRC patients (Cont'd)](image)
The Kolmogorov-Smirnov test showed significant differences between the CRC and AMD (along with AJBC and AJNC patients). Statistical comparison (Kolmogorov-Smirnov test) also showed a clear difference between the CRC Jewish and non-Jewish patients compared with either the AMD (P = 0.91). The use of LOD calculations (method 2) also identified most of the IBD segments detected by the homozygosity (method 1) analysis. On average, CRC patients have LOD of 2.76, which is significantly higher than either the AMD (0.67) or Framingham (1.64) controls (Table 2). Eighty-eight percent of the CRC patients had LOD score of at least 5, whereas it is 16% and 57% for the AMD and Framingham controls, respectively.

**CRC patients of Jewish ancestry have higher percentage of IBD regions compared with the rest of the cohort, and the control groups.** If the CRC patients are divided into Jewish and non-Jewish groups, 94% of the former and 35% of latter have IBD regions. There is also a disparity in IBD segment size—8.3 Mb for the Jewish and 5.1 Mb among non-Jewish patients (calculated based on information and data listed in Supplements S-A2 and S-B3). Statistical comparison (Kolmogorov-Smirnov analysis) also showed a clear difference between the CRC Jewish and similar regions.

The graph is presented in such a way that each data point represents the cumulative fraction (y axis) of the samples with the corresponding minimum cumulative IBD segment length (x axis). In other words, $Y = f(X > x)$. For example, the graph tells us that ~35% of the CRC have total IBD length of at least 10 Mb, whereas it is only 10% for both controls. The clear difference between the CRC patients and the control data sets can be seen even up to a cumulative frequency of 20 Mb IBD segment/sample. The Kolmogorov-Smirnov test (31) showed significant differences between the CRC and AMD (P = 1.28 × 10^-5) and CRC and Framingham (P = 1.13 × 10^-5) distributions. On the other hand, there was no significant difference between the distributions of AMD and Framingham data sets (P = 0.91). The use of LOD calculations (method 2) also identified most of the IBD segments detected by the homozygosity (method 1) analysis. On average, CRC patients have LOD of 2.76, which is significantly higher than either the AMD (0.67) or Framingham (1.64) controls (Table 2). Eighty-eight percent of the CRC patients had LOD score of at least 5, whereas it is 16% and 57% for the AMD and Framingham controls, respectively.

**Table 2. Summary of the results of IBD segment analysis using the homozygosity mapping and autozygosity (LOD) approach**

<table>
<thead>
<tr>
<th>Data set</th>
<th>Homozygosity analysis (method 1)</th>
<th>LOD approach (Method 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average IBD ≥ threshold (Mb)</td>
<td>Average LOD*</td>
</tr>
<tr>
<td></td>
<td>Average IBD without max IBD sample</td>
<td>% Samples with IBD ≥ 4 Mb (threshold)</td>
</tr>
<tr>
<td>All CRC patients</td>
<td>12.7</td>
<td>62.2</td>
</tr>
<tr>
<td>AMD control</td>
<td>5.3</td>
<td>35.6</td>
</tr>
<tr>
<td>Framingham control</td>
<td>5.5</td>
<td>28.8</td>
</tr>
</tbody>
</table>

*Average of positive LOD.
†Samples that have at least one segment with LOD ≥ 5 are eligible.

Figure 3. A, the cumulative distributions of the lengths of IBD segments for the CRC patients, as well as AMD and Framingham control individuals. The graph is presented in such a way that each data point represents the cumulative fraction (y axis) of the samples with the corresponding minimum cumulative IBD segment length (x axis). In other words, $Y = f(X > x)$. The clear difference between the CRC patients and the control data sets can be seen even up to a cumulative frequency of 20 Mb IBD segment/sample. The Kolmogorov-Smirnov test showed significant differences between the CRC and AMD (P = 1.28 × 10^-5), as well as between CRC and Framingham (P = 1.13 × 10^-5) distributions. On the other hand, there was no significant difference between the distributions of AMD and Framingham data sets (P = 0.91). B, the cumulative distributions of the lengths of IBD segments for Jewish and non-Jewish subgroups of the CRC patients, the AMD, and Framingham controls, along with AJBC and AJNC patients. Statistical comparison (Kolmogorov-Smirnov test) also showed a clear difference between the CRC Jewish and non-Jewish distributions (P = 0.0170). Nonetheless, both the percentages of samples with IBD segments and the average IBD segment size are significantly higher for non-Jewish patients compared with either the AMD (P = 4.30 × 10^-10) or Framingham controls (P = 1.08 × 10^-13). B. We then compared the IBD segment distributions in the Ashkenazi Jewish (AJBC and AJNC) data sets with those of our CRC control (AMD and Framingham) data sets. The IBD segment distributions of AJBC and AJNC are indistinguishable from each other (P = 0.922). However, it is very clear that the fraction of samples with at least 5 Mb total IBD length is higher in both Ashkenazi Jewish data sets than in the CRC non-Jewish, as well as AMD and Framingham data sets. Statistical comparisons show that AJBC versus AMD, AJNC versus AMD, AJBC versus Framingham, and AJNC versus Framingham have P values of 1.31 × 10^-10, 9.48 × 10^-11, 2.09 × 10^-12, and 2.54 × 10^-17, respectively. The data from AJBC and AJNC groups were generated using the more dense Affymetrix 500K SNP array. Before the comparing the IBD segments identified from the 500K and the 50K Xba data set, we identified the SNPs whose genomic positions are closely matched in the two sets (maximum separation of 10,000 bp, although 9,360 SNPs are identical, in the two array sets; see Supplement S-D2). Thus, the IBD regions identified and plotted for B were from the analyses of 39,097 SNPs.
non-Jewish distributions ($P = 0.0170$). Nonetheless, both the percentages of samples with IBD segments and the average IBD segment size are significantly higher for non-Jewish patients compared with either the AMD ($P = 4.30 \times 10^{-6}$) or Framingham controls ($P = 1.08 \times 10^{-4}$; Fig. 3B). This observation also led us to examine additional data sets generated specifically for a genome-wide association study at MSKCC: 30 AJBC, along with 133 AJNC. We then compared the IBD segment distributions in the Ashkenazi Jewish (AJBC and AJNC) data sets with those of our CRC and control (AMD and Framingham) data sets. The IBD segment distributions of AJBC and AJNC are virtually indistinguishable from each other ($P = 0.922$). It is very clear that the fraction of samples with at least 5 Mb total IBD length is higher in both Ashkenazi Jewish data sets than in the CRC non-Jewish, as well as AMD and Framingham data sets. Statistical comparisons show that AJBC versus AMD, AJNC versus AMD, AJBC versus Framingham, AJNC versus Framingham have $P$ values of $1.31 \times 10^{-6}$, $9.48 \times 10^{-17}$, $2.09 \times 10^{-7}$, and $2.54 \times 10^{-11}$, respectively.

Autozygosity increases CRC risk. From the data plotted in Fig. 3A, it is possible to calculate the extent to which autozygosity adds to CRC risk by using Bayes’ rule, a formula of conditional probabilities: $P(B|A) = P(A|B) \times P(B)/P(A)$. If we assume that $A$ refers to IBD $\geq x$, where $x$ is the IBD length, and $B$ refers to CRC incidence, then:

$$P(CRC|IBD \geq x) = P(IBD \geq x|CRC) \times P(CRC)/P(IBD \geq x)$$  \hspace{1cm} (B)

From Fig. 3A, we can see that $P(IBD \geq 10 \, \text{Mb}|CRC) = 0.3$ and $P(IBD \geq 20 \, \text{Mb}|CRC) = 0.18$. Furthermore, the data from control data sets (which represents 95% of the population) suggest that $P(IBD \geq 10 \, \text{Mb}) = 0.1$; $P(IBD \geq 20 \, \text{Mb}) = 0.05$. Therefore

$$P(CRC|IBD \geq 10 \, \text{Mb}) = 3 \times P(CRC); P(CRC|IBD \geq 20 \, \text{Mb}) = 3.6 \times P(CRC)$$  \hspace{1cm} (C)

Equation B shows that having total IBD of at least 10 Mb increases CRC risk 3-fold, whereas having a total IBD of at least 20 Mb increases the risk almost 4-fold.

Discussion and Conclusion

The most plausible explanation for the presence of long stretches of homozygous regions in an individual’s genome is that his or her parents can trace their lineage to a common ancestor. UPD (an instance when an offspring inherits both copies or segments of chromosomes from a single parent), although possible, is highly unlikely. In cancer tissues, the appearance of a UPD may be attributed to autozygosity. In one particular family, all the progeny showed 4 to 12 autozygous segments with an average length of 19 cM per segment. The fact that both parents did not show any significant homozygosity suggests that the parents can trace their ancestry to a common individual. Using the publicly available SNP genotype data for 209 individuals from the International Hapmap Project (34), Gibson and coworkers identified 1,393 homozygous segments (with at least 1-Mb length and minimum SNP density of 1 SNP per 5 kb; ref. 35). The longest identified homozygous segment (17.91 Mb) is that of a Japanese individual whom the authors consider to be a progeny of related parents. Yorubas from Ibadan, Nigeria, have the fewest long tracts of homozygosity when compared with Han Chinese from Beijing, Japanese from Tokyo, and CEPH Utah individuals of Northern and Western Europe ancestry. This observation is consistent with the belief that the African race has been established earlier (thus higher incidence of recombination subdividing the haplotype regions) than the Asiatic and Caucasian races. Another important conclusion from their study is that these homozygous segments are more prevalent in regions of high linkage disequilibrium (and thus, of low recombination). Based on the analysis of Li and coworkers, the genomes of 34 of 515 (6.6%) unrelated Han Chinese individuals also contained these homozygous segments (which they referred to as long contiguous stretches of homozygosity). The segment size ranged from 294 to 26,27 Mbp (36). Using the publicly available Affymetrix data sets, they also found that 26.2% of Caucasians and 4.76% of African Americans also have these IBD segments in their genomes. When they analyzed the genomes of siblings of a consanguineous marriage, they found out that the genomes of all the siblings exhibited multiple long contiguous stretches of homozygosity ranging from 3.06 to 52.17 Mb. This served as clear proof that genomic IBD regions result from inbreeding. Most recently, the International Hapmap Project (phase II; ref. 37) was able to identify these extended runs of homozygosity among 51 of 270 individuals (19%). Although they used more dense SNP arrays, and set different specifications (minimum of 3 Mb), the percentages of individuals with long homozygous segments were comparable with what we found in the AMD and Framingham controls. The authors also contended that these were most probably due to recent co-ancestry in the individuals’ parents. We then examined the possibility that the IBD segments among our CRC subjects may actually be haploblocks or groups of alleles (or SNPs) that are usually in linkage disequilibrium. However, of the 117 IBD segments identified by the homozygosity analysis, only 11 (9%) have at least 30% overlap (see Supplement S-B9) with the long haplotype regions identified by the International Hapmap project (phase I; ref. 34).

There are clear correlations between the incidence of cancer and degrees of inbreedings on a number of population-based studies (21, 22). The results of our own study clearly show the difference in degrees (both the percentage and lengths) of autozygous segments between the MSKCC CRC patients and the control data sets. However, it is important to note that of all the 74 CRC patients in our study, 16 (22%) indicated Judaism as their religious affiliation. However, it is important to note that of all the 74 CRC patients in our study, 16 (22%) indicated Judaism as their religious affiliation. This is greatly due to the location of MSKCC (New York City). According to a 2002 survey, there are 1.4 million individuals of Jewish ancestry (constituting 15% of all the households) living in the five New York City boroughs plus three surrounding counties (38). Unfortunately, we do not have any information on the religious affiliations of the subjects making up the AMD and Framingham controls data sets. It is very likely that the incidence of autozygosity among people of Jewish ancestry are more prevalent compared with the average Caucasian population. Historically, Jewish communities have maintained high degree of endogamy (marrying within its own group) for cultural and religious reasons, thus
increasing the chances of autozygotic signatures in their genomes. The patient C0111 who has the most IBD segments of all the MSKCC patients is of Jewish descent. We can only speculate whether the incidence of autozygosity is a contributing factor to the fact that Ashkenazi Jews have the highest incidence of colon cancer of any ethnic group in the world (39). Aside from dietary factors, genetics can also play a major role. The APC variant I1307K, almost unique to Ashkenazi Jews, has been identified as a CRC susceptibility factor among this group (40). The results of our principal components analysis (EIGENSTRAT method) did not identify population stratification between those CRC patients of Jewish ancestry and the rest of the CRC cohort (Supplement S-C4).

Table 3. The recently identified predisposition SNPs for colon cancer and Crohn’s disease whose locations are covered by the IBD regions in some of the CRC patients in the study

<table>
<thead>
<tr>
<th>Identified IBD region</th>
<th>Chromosomal region</th>
<th>SNP</th>
<th>Associated disease</th>
<th>Reference</th>
<th>Actual SNP genotype</th>
<th>Allele associated to added risk (Y/N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRC patient</td>
<td>Position (Mb)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C0153</td>
<td>5q23</td>
<td>rs10505477</td>
<td>Colon cancer</td>
<td>(46)</td>
<td>TT</td>
<td>Y</td>
</tr>
<tr>
<td>C0153</td>
<td>5q23</td>
<td>rs6983267</td>
<td>Colon cancer (and prostate cancer)</td>
<td>(47, 48)</td>
<td>GG</td>
<td>Y</td>
</tr>
<tr>
<td>C0111</td>
<td>5q23</td>
<td>rs17234657</td>
<td>Colon cancer (strong association)</td>
<td>(49)</td>
<td>CC</td>
<td>N</td>
</tr>
<tr>
<td>C0253</td>
<td>5q23</td>
<td>rs6596075</td>
<td>Crohn's disease (strong association)</td>
<td>(49)</td>
<td>CC</td>
<td>Y</td>
</tr>
<tr>
<td>10216</td>
<td>5p13</td>
<td>rs6596075</td>
<td>Crohn's disease (strong association)</td>
<td>(49)</td>
<td>CC</td>
<td>Y</td>
</tr>
<tr>
<td>C0111</td>
<td>5q23</td>
<td>rs11805303</td>
<td>Crohn's disease (moderate association)</td>
<td>(49)</td>
<td>CC</td>
<td>Y</td>
</tr>
<tr>
<td>00485</td>
<td>5q23</td>
<td>rs6596075</td>
<td>Crohn's disease (moderate association)</td>
<td>(49)</td>
<td>CC</td>
<td>Y</td>
</tr>
<tr>
<td>C0181</td>
<td>5q33</td>
<td>rs1000113</td>
<td>Crohn's disease (strong association)</td>
<td>(49)</td>
<td>CC</td>
<td>N</td>
</tr>
<tr>
<td>C0170</td>
<td>6p21</td>
<td>rs9469220</td>
<td>Crohn's disease (moderate association)</td>
<td>(49)</td>
<td>GG</td>
<td>N</td>
</tr>
<tr>
<td>C0159</td>
<td>6p21</td>
<td>rs9469220</td>
<td>Crohn's disease (moderate association)</td>
<td>(49)</td>
<td>GG</td>
<td>N</td>
</tr>
<tr>
<td>C0111</td>
<td>6p22</td>
<td>rs9086425</td>
<td>Crohn's disease (moderate association)</td>
<td>(49)</td>
<td>CC</td>
<td>Y</td>
</tr>
<tr>
<td>A7223</td>
<td>10q24</td>
<td>rs10883365</td>
<td>Crohn's disease (strong association)</td>
<td>(49)</td>
<td>AA</td>
<td>N</td>
</tr>
<tr>
<td>07061</td>
<td>10q24</td>
<td>rs10883365</td>
<td>Crohn's disease (strong association)</td>
<td>(49)</td>
<td>GG</td>
<td>Y</td>
</tr>
</tbody>
</table>

NOTE: The actual genotypes of the SNPs were verified by direct DNA sequencing.
state. Moreover, longer IBD segments may also cover multiple low-penetrance, dose-dependent genes that have additive effects, which is now believed to occur in both sporadic and familial types of CRC (4). We can only presume that functionally, such low-penetrance, dose-dependent alleles do not necessarily have to be associated with tumor suppressors. For instance, it is possible that a mutation in the regulatory region of a proto-oncogene may result in protein overexpression or the dysregulation during stress.

This is an exploratory study on the possibility that autozygosity increases the risk of cancer, and there are limitations in our study. First, there is the lack of information on the cancer status of the AMD control subjects. The average age of AMD subjects is 80 years. According to the statistics provided by National Cancer Institute Surveillance Epidemiology and End Results, which is accessible through the Web site, the incidence of CRC in the United States (for all races between 1975 and 2003) is 0.322%, 0.377%, and 0.416% for age groups 75 to 79, 80 to 84, and 85+, respectively (45). Thus, there is only a small chance that an AMD study participant has also been afflicted with CRC. All of the subjects in the other control data set (Framingham) did not have any cancer at the time of their participation according to clinical records. Second, our cohort data was enriched for patients of Jewish ancestry, and the work would have benefited from availability of another sizable set of control individuals of Jewish ancestry who had not been diagnosed of CRC at a late age (75 years or older). Third, we chose to evaluate copy number and alleles only among matched normal and tumor samples, where the tumor samples had -30% stromal infiltrates. Such samples may have had more homogenous tumors, which, in turn, may have had a higher incidence of underlying genetic factors.

None of the widely recognized CRC predisposing genes (APC, MLH1, MSH2) fall within our identified IBD regions. Most recently, several laboratories have performed large-scale, genome-wide association studies and identified several loci associated with increased risk to colon cancer (46–48). All of the three newly identified colon cancer–associated SNPs of highest risk (46, 47) are located within the identified IBD regions among the CRC patients, three (among four patients) were found to be homozygous for the CRC-predisposing alleles. However, the 9p24 SNP was found to be homozygous for the non-CRC predisposition SNP. According to our clinical records, both patients C0111 and C0153 did not have any family history of CRC. We also examined the genotypes of IBD region SNPs that have been associated to Crohn’s disease (49), a possible precursor of colon cancer (50). The Crohn’s disease–associated SNPs were interrogated despite the fact that none of our CRC patients had any clinical documentation for the disease (Supplement S-A2). Of the seven Crohn’s disease–associated SNPs located within the identified IBD regions among the CRC patients, three (among four patients) were found to be homozygous for the Crohn’s disease–predisposing SNPs: rs6596075 (within IBD regions of C0111H and 00485K), rs6908425 (within IBD region of C0111H), and rs10883365 (within IBD region of 07061). However, SNPs rs11805303 (C0253K), rs17234657 (10216H), rs1000113 (C0181H), rs9469220 (C0170H and C0159H), and rs10883365 (A7223H) were all genotyped to be homozygous for the non–disease-associated alleles.

We have shown that there is a higher frequency and a longer length of IBD segments within our CRC patients compared with a number of control groups. Whether these IBD segments result in cancer or lead to the progression of cancer has yet to be determined. There is clearly a need to expand this study to include the sampling of a wider cohort and just as importantly to examine the identified IBD regions for potential cancer-causing genes.

**Disclosure of Potential Conflicts of Interest**

No potential conflicts of interest were disclosed.

**Acknowledgments**

Received 9/6/2007; revised 1/15/2008; accepted 1/30/2008.

**Grant support:** National Cancer Institute grant P01-CA59390, the Gilber Family Foundation, and generous funding from the Ludwig Institute for Cancer Research/Conrad N. Hilton Foundation joint Hilton-Ludwig Cancer Metastasis Initiative.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

We thank Dr. Jenny Xiang and her team at the Microarray Core Facilities, Department of Microbiology, Cornell University Weill Medical College, New York, NY, for the help and services and Shoshana Rosenberg of MSKCC for all the valuable assistance.

**References**


The Signatures of Autozygosity among Patients with Colorectal Cancer

Manny D. Bacolod, Gunter S. Schemmann, Shuang Wang, et al.


Updated version
Access the most recent version of this article at:
doi:10.1158/0008-5472.CAN-07-5250

Supplementary Material
Access the most recent supplemental material at:
http://cancerres.aacrjournals.org/content/suppl/2008/05/15/0008-5472.CAN-07-5250.DC1

Cited articles
This article cites 43 articles, 9 of which you can access for free at:
http://cancerres.aacrjournals.org/content/68/8/2610.full#ref-list-1

Citing articles
This article has been cited by 5 HighWire-hosted articles. Access the articles at:
http://cancerres.aacrjournals.org/content/68/8/2610.full#related-urls

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, use this link
http://cancerres.aacrjournals.org/content/68/8/2610.
Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site.