Comprehensive assessment of variation at the transforming growth factor β type 1 receptor locus and colorectal cancer predisposition

Luis G. Carvajal-Carmona1, Mike Churchman2, Carolina Bonilla2,c, Axel Walther3, Jeremy H. Lefèvre2,c, David Kerr5, Malcolm Dunlop4, Richard Houlston2, Walter F. Bodmer2,c,1, and Ian Tomlinson6

1Molecular and Population Genetics Laboratory, Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, United Kingdom; 2Cancer and Immunogenetics Laboratory, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford OX3 9DZ, United Kingdom; 3Department of Clinical Pharmacology, University of Oxford, Oxford OX3 7DQ, United Kingdom; 4Colon Cancer Genetics Group, Medical Research Council, Human Genetics Unit, Western General Hospital, Edinburgh EH4 2XU, United Kingdom; and 5Section of Cancer Genetics, Institute of Cancer Research, Sutton SM2 5NG, United Kingdom

Contributed by Walter F. Bodmer, March 15, 2010 (sent for review February 1, 2010)

The role of transforming growth factor β receptor type 1 (TGFBR1) polymorphisms, particularly a coding CGC insertion (rs11466445, TGFBR1*6A/9A) in exon 1, has been extensively investigated in regard to colorectal cancer (CRC) risk. These investigations have generated conflicting results. More recently, allele-specific expression (ASE) of TGFBR1 mRNA has been suggested as predisposing to CRC, with a relative risk of nearly 10-fold and a population attributable risk of ~10%. Owing to the potential importance of TGFBR1 variants in CRC, we performed a comprehensive examination of tagging SNPs at and around the gene in 3,101 CRC cases and 3,334 controls of northern European ancestry. To test whether rare or subpolymorphic TGFBR1 variants were associated with CRC risk, we sequenced the gene’s exons in a subset of patients. We also evaluated TGFBR1 ASE in a panel of CRC cases and controls. Overall, we found no association between TGFBR1 polymorphisms and CRC risk. The rare variant screen did not identify any changes of potentially pathogenic effects. No evidence of greater ASE in cases than controls was detected, and no haplotype around TGFBR1 could account for the ASE reported in other studies. We conclude that neither genetic variation nor ASE at TGFBR1 is likely to be a major CRC risk factor.

The effects on cancer risk of other common polymorphisms and rare variants at TGFBR1 have not specifically been assessed. A second line of evidence for the importance of TGFBR1 in predisposition to CRC has recently emerged. Valle et al. (7) reported that CRC cases show imbalanced expression of TGFBR1 alleles compared with controls. This so-called “germline” allele-specific expression (ASE) was detected using the SNPshot (Applied Biosystems) method to screen TGFBR1 mRNA from peripheral blood in a panel of 138 cases and 105 controls. Although no causal variation was found, this tendency was said to be present in about 20% of CRC cases, to be dominantly inherited, and to confer a nearly 10-fold increased risk for CRC. A further report from Guda et al. (8) extended the study of Valle et al. (7) using pyrosequencing. Guda et al. (8) found ASE in only about 5% of familial CRC cases and in no sporadic CRC patients or controls.

We wished to undertake a thorough assessment of TGFBR1 in relation to CRC risk in a large set of cases and controls. Using a total of over 3,000 CRC cases and over 3,300 controls, we genotyped haplotype-tagging SNPs from throughout the TGFBR1 region, imputed other SNPs in the region, genotyped the TGFBR1*6A/9A polymorphism, and screened the coding region of the gene for rare variants with potential effects on disease risk. In parallel, we undertook ASE analysis in a subset of patients using two complementary techniques.

Results

Association Between TGFBR1*6A/9A and CRC Risk. We typed TGFBR1*6A/9A in 828 familial colorectal tumor cases and 913 controls from the COloRectal Gene Identification (CORGI) study. There was no association with CRC susceptibility (Pper allele = 0.71, β = 0.04, SE = 0.13; Table 1). To investigate the association between this allele and CRC in two other case–control cohorts, we used IMPUTEv2 (9) to generate in silico genotypes, with the CORGI samples as a diploid reference (https://mathgen.stats.ox.ac.uk/impute/impute.html). The imputation quality was very good, with proper info scores >80% (Table 1) in all cases and >98% correspondence between imputed and typed genotypes in the CORGI study. In the additional cohorts, we found no evidence for an association between TGFBR1*6A/9A and CRC risk (Pper allele in the Scotland series = 0.07, β = 0.21; Pper allele in the VQ58 series = 0.58, β = 0.05; 1To whom correspondence may be addressed. E-mail: luis@well.ox.ac.uk or walter.bodmer@hertford.ox.ac.uk.

This article contains supporting information online at www.pnas.org/cgi/content/full/1002816107/DCSupplemental.

Author contributions: L.G.C.-C., D.K., M.D., R.H., W.F.B., and I.T. designed research; L.G.C.-C., W.F.B., and I.T. wrote the paper. The authors declare no conflict of interest.
Table 1. Genotypes at rs11466445 in each of the three CRC case–control series

<table>
<thead>
<tr>
<th>Series</th>
<th>Cases</th>
<th>Controls</th>
<th>Additive</th>
<th>Dominant</th>
<th>Recessive</th>
<th>Genotypic</th>
<th>P value and associated regression coefficient $\beta$ (SE)$^T$</th>
<th>Proper_info score$^T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CORGI</td>
<td>746/159/8</td>
<td>673/145/10</td>
<td>0.70, 0.04 (0.13)</td>
<td>0.81, 0.02 (0.12)</td>
<td>0.49, 0.32 (0.47)</td>
<td>0.47, 0.01 (0.12)</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>Scotland</td>
<td>772/152/9</td>
<td>843/140/7</td>
<td>0.07, 0.21 (0.15)</td>
<td>0.07, 0.23 (0.13)</td>
<td>0.54, 0.31 (0.50)</td>
<td>0.50, 0.22 (0.13)</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>VQ58</td>
<td>938/201/13</td>
<td>1119/200/14</td>
<td>0.58, 0.05 (0.09)</td>
<td>0.55, 0.06 (0.10)</td>
<td>0.96, 0.01 (0.38)</td>
<td>0.38, 0.01 (0.38)</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td></td>
<td></td>
<td>0.13, 0.09 (0.06)</td>
<td>0.13, 0.10 (0.07)</td>
<td>0.17, 0.18 (0.25)</td>
<td>0.17, 0.09 (0.07)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Tests of association (additive, dominant, recessive, and genotypic) are shown.

*Genotypes in the CORGI study were obtained by direct genotyping. Genotypes in the Scotland and VQ58 series were imputed using IMPUTE2 software. Counts in the Scotland and VQ58 series were derived from individual genotypes with a probability of 90% or higher.

$^T$P values were derived from the additive, dominant, recessive, and genotypic disease models as implemented using logistic regression in the program SNPTEST. For each series, we present the P value, followed by the regression coefficient and the regression coefficient's SE (in parentheses).

$^T$Proper_info scores relate to the quality of the imputation. These scores contain a measure of the relative statistical information about the parameters of interest, $\beta$ in this case. This measure lies in the range [0,1], with 1 indicating perfect information and 0 indicating no information. For more details on this score, see the article by Marchini et al. (12).

A meta-analysis of the three studies found no association between TGFBR1*6A/6A alleles and CRC (P per allele = 0.12; Fig. 1).

We then investigated whether particular haplotypes around TGFBR1*6A/6A could mediate CRC risk. The 6A allele was principally present on two haplotypes with frequencies of $\sim$1% and $\sim$8%. However, neither haplotype was associated with CRC risk (SI Appendix).

Although TGFBR1*6A/6A was not a major detectable CRC risk factor on its own, it might increase risk by interacting with polymorphisms in other genes. Prime candidates are four CRC SNPs that map close to the TGFB/BMP pathway genes GREM1, BMP2, BMP4, and SMAD7 (5). We carried out simple multiplicative genotypic interaction tests in cases only between TGFBR1*6A/6A and the SNPs rs4779584, rs961253, rs4444235, and rs4599827, corresponding to each of the four risk loci above. No interactions were found ($P = 0.05, P = 0.49, P = 0.45$, and $P = 0.56$, respectively; SI Appendix).

Association Between Other SNPs Near TGFBR1 and CRC. We genotyped up to 15 tagging SNPs spanning the $\sim$75-kb haplotype block containing the TGFBR1 locus (chr9: 100,897,875–100,973,680) in 3,125 CRC cases and 3,372 controls, all of northern European ancestry. To generate additional genotypes in this block, we used the genotype data from the Centre d’Étude du Polymorphisme Humain (CEPH) samples ($n = 63$) obtained by the 1000genomes project (http://browser.1000genomes.org/index.html). This dataset contained nearly five times more SNPs and lower frequency variants than those present in release 22 of the Hapmap project (260 SNPs vs. 56 SNPs), including a much larger fraction of variants with MAF $< 0.01$ (97 vs. 1), with 0.01 $< $ MAF $< 0.05$ (67 vs. 1), and with MAF $> 0.05$ (96 vs. 54). Large proportions of these variants (91%, 74%, and 27%, respectively) were not reported in database of single nucleotide polymorphisms (dbSNP, http://www.ncbi.nlm.nih.gov/projects/SNP/). The use of this reference population for imputation permitted a comprehensive evaluation of common variants and some rarer variants close to TGFBR1 for CRC susceptibility. Genotype imputation was successful for 89–97 variants depending on the study cohort (differences in SNP array genotyping are discussed in Materials and Methods); most variants had MAF $> 0.05$ (SI Appendix), bringing the total variants evaluated in this haplotype block to 102.

We found no evidence of association between SNPs in the TGFBR1 haplotype block and CRC risk in the VQ58 and Scottish cohorts. Two SNPs had nominally significant associations in the CORGI study ($9{-}100973331, P = 0.018; 9{-}100973680, P = 0.019$; SI Appendix), but none of these remained significant after correction for multiple testing. An association meta-analysis of the three cohorts found no evidence that any SNP was significantly associated with CRC (SI Appendix). Therefore, common variation at the TGFBR1 locus is unlikely to be associated with CRC risk in populations of northern European ancestry.

To investigate whether long-range regulation could mediate the association between TGFBR1 and CRC risk and explain mechanisms such as ASE, we extended our analysis to $\sim$500 kb on each side of the TGFBR1 haplotype block (chr9: 100,385,599–101,513,263). We found very little evidence for an association between SNPs in this region and CRC risk after multiple testing correction (SI Appendix). However, one marker (rs410180), located $\sim$100 kb downstream of TGFBR1, had a suggestive association ($P_{\text{per allele}} = 0.00008, \beta = -0.24555, SE = 0.06$) in all three sample sets combined. We were intrigued by this possible association between rs410180 and CRC risk and decided to obtain genotypes from two additional cohorts that included $\sim$3,000 sporadic cases and $\sim$3,000 controls from England and $\sim$2,000 sporadic cases and $\sim$2,000 sporadic controls from the Scotland series. We found no support for the association between rs410180 and disease in these latter cohorts ($P_{\text{per allele}} = 0.15$ and $P_{\text{per allele}} = 0.59$, respectively; meta-analysis for all five cohorts: $P = 0.003$, odds ratio = 1.13, 95% confidence interval: 1.04–1.22), suggesting that rs410180 is unlikely to be a CRC susceptibility variant.

A further possibility of long-range regulation of TGFBR1 could involve multimarker haplotypes rather than alleles. To examine this possibility, we used the Haploview program (10) to investigate associations between haplotypes in this region and disease risk. The distal and proximal flanking regions contained 19 and 12 haplotype blocks defining 82 and 45 haplotypes, respectively. We found no haplotypes with significant associations after correction for multiple testing (lowest nominal $P$ value = 0.007). Therefore, neither alleles nor haplotypes at or around TGFBR1 are likely to be associated with CRC risk.

Association Between Other TGFB Pathway Genes and CRC Risk. In previous studies, we have shown that alleles at a number of TGFB pathway genes are associated with CRC risk. These genes include SMAD7, GREM1, BMP2, and BMP4 (5, 11). It is thus possible that other genes in this pathway interact with TGFBR1 to increase risk for cancer. To investigate this possibility, we examined the association between CRC and alleles at and around TGFBR1, TGFBR2, TGFBR3, and TGFBR3 in our three cohorts (SI Appendix). Overall, we found very little evidence for an association between these genes and CRC, with the lowest $P$ values being at two TGFBR2 intragenic SNPs: 3-30632242 ($P = 0.00083, \beta = 0.035$) and 3-30644255 ($P = 0.00099, \beta = 0.43$).

Although no SNPs in TGFBR pathway genes examined in the present study were individually associated with CRC, we tested epistasis between genotyped SNPs in the TGFBR1-containing...
haplotype block and SNPs genotyped close to TGFB1, TGFB2, TGFB3, TGFBR2, and TGFBR3. We failed to detect any significant interaction at $P = 0.001$, suggesting that epistasis between TGFBR1 and these five closely-related genes is unlikely to mediate CRC risk.

**Coding Variation at TGFBR1.** We screened the entire coding sequence of TGFBR1 in 96 CRC cases from the CORGI cohort. All these patients had at least one first-degree relative with CRC and did not have mutations in known highly penetrant CRC genes. We found only one change in the TGFBR1 coding region, a synonymous serine-to-serine change at residue 39 (exon 2). We failed to identify the Tyr401Asn change reported by Valle et al. (7). An examination of the 1000genomes CEPH genotype data did not reveal the existence of this change or any other TGFBR1 non-synonymous variant. Thus, we found no evidence of uncommon disease-associated variants in the coding region of TGFBR1, suggesting that these do not represent a major risk factor for CRC.

**Germline ASE.** We examined ASE in lymphoblastoid cell line-derived cDNA at TGFBR1*6A/9A (located in the 3' UTR of the gene) and in a 5'-UTR marker (rs1590) previously examined by Valle et al. (7). Twenty-five individuals (8 patients and 17 controls) were informative (heterozygous) at TGFBR1*6A/9A, and 44 individuals (16 patients and 28 controls) were informative at rs1590. We initially searched for any evidence of ASE as a rare trait by looking for outliers in terms of their relative allelic expression of TGFBR1 alleles (Fig. 2). One control showed highly biased expression at TGFBR1*6A/9A, and a single case showed similarly biased expression at rs1590. However, apparently biased allelic dosages were also seen at rs1590 in genomic DNA (gDNA) from one control and two cases. In the absence of constitutional copy number variation at TGFBR1 (http://genome.ucsc.edu/ and http://projects.tcag.ca/variation/), it seemed most likely that the outlying individuals resulted from inherent occasional problems with quantitative genotyping.

Despite our lack of clear evidence for a discrete category of ASE, we assessed whether ASE, in the sense of the degree of
biased allelic expression, was more prevalent in our cases than in controls. There was a nominally significant difference between the cases and controls \((P = 0.028, \text{Wilcoxon test; } t = -2.09 \text{ and } P = 0.040, \text{t test})\), but the trend to ASE was actually more pronounced in the controls (mean = +0.027 deviation from mean, SD = 0.12) than in the cases (mean = −0.041, SD = 0.13).

In the above analysis, we reasoned that ASE should be assessed relative to the hypothetical 1:1 ratio of allelic dosage rather than by comparing cDNA and gDNA dosage from a single sample, essentially so that experimental noise was minimized. However, we also replicated the tests performed by Valle et al. (7) and Guda et al. (8) by calculating the dosage ratio, \((A_{\text{cDNA}}/B_{\text{cDNA}})/(A_{\text{gDNA}}/B_{\text{gDNA}})\), for alleles A and B in each individual. Nineteen individuals showed ASE according to the criteria of Valle et al. (7), that is, ratios >1.5, or <0.67. Of these, 7 were patients and 12 were controls, demonstrating a nonsignificant difference \((\chi^2 = 0.049, P = 0.83)\). Quantitative analyses also showed no significant difference in ratios between cases and controls \((P = 0.09, \text{Wilcoxon test; } P = 0.13, \text{t test})\).

**Discussion**

Despite its excellent candidate gene status and previous reports, we found no evidence to show that common or rare genetic variants at or near \(TGFBR1\) are associated with the risk for CRC. There was, moreover, no evidence of a disease-associated haplotype in this region or that variants near \(TGFBR1\) had any effect on risk in association with other variants in the TGFB pathway. Although we did not specifically test for common copy number variants, there is very little evidence of such variation in public databases. Our study was empowered to detect common alleles with moderate or greater effects (e.g., >80% power to find an allele with a frequency of 0.3 with an additive 1.1-fold effect on disease risk at \(P = 0.1\)).

We failed to confirm the reported finding of higher levels of ASE at \(TGFBR1\) in CRC cases than controls, and we did not identify any disease-associated haplotype that could account for the previously reported ASE at this site. The question remains as to whether ASE occurs at all at \(TGFBR1\). In our data, we observed no greater variation in relative allelic dosage in cDNA than in gDNA, suggesting that it is difficult to find rare ASE events using the techniques employed.

To date, hypothesis-free genome-wide association (GWA) studies in colorectal and other cancers have identified tens of variants associated with differential risk for disease. The genes involved act in multiple different pathways, although there is accumulating evidence that the TGFB/BMP pathway plays a central role in CRC predisposition. The prior focus of research groups on \(TGFBR1\) as a candidate gene has therefore been vindicated by GWA approaches. The absence of detectable CRC-associated genetic variation at \(TGFBR1\) might result from

![Fig. 2. Box-and-whiskers plots showing log (allelic dosage) at rs11466445 and rs1590 relative to the geometrical mean for that polymorphism in all samples. Dosage is shown for cases and controls at TGFBR1*6A/9A in cDNA (A), rs1590 in cDNA (B), TGFBR1*6A/9A in gDNA (C), and rs1590 in gDNA (D). Note the outlying samples shown by diamonds.](image-url)
variants of very weak effects or low frequency but may also result from chance or may reflect relatively strong natural selective constraints owing to the central importance of the type 1 TGFβ receptor in development and tissue maintenance.

Materials and Methods

Study Samples. We used three CRC case–control series based on samples of northern European ancestry. The first series was ascertained in England through the CORGI consortium and included 920 familial CRC or significant adenoma cases and 929 cancer-free controls. Familial patients in the CORGI cohort were individuals with two first-degree relatives with colorectal tumors and from whom known Mendelian syndromes had been excluded. The second case–control sample comprised 1,003 early-onset Scottish CRC patients (<55 years of age) who had no mutations in the known highly penetrant genes and from whom known Mendelian syndromes had been excluded. The third case–control sample comprised 1,216 samples from the post-treatment stage of a Phase III, randomised, double blind, placebo-controlled study of rofecoxib (VIOXX) in colorectal cancer patients following potentially curative therapy (VICTOR) (n = 920; http://www.octo-oxford.org.uk/alltrials/infolowup/vic.html) and from the multicentre international study of casecallable ± bevacizumab as adjuvant treatment of colorectal cancer (QUASAR2) (n = 356; http://www.octo-oxford.org.uk/alltrials/q2.html). In the latter study, we used publicly available control genotype data from 1,437 individuals belonging to the UK 1958 Birth Cohort. We refer to these three series as the CORGI, Scotland, and VQ58 series (5).

Nudeic Acid Isolation and Genotyping. DNA samples were isolated from peripheral blood using standard methods and quantified with picogreen. Tagging SNPs around TGFBR1 were typed with Illumina Hap300 (VQ58) or Hap550 (CORGI and Scotland1 series) SNP arrays; these contain 9 and 15 tagging SNPs, respectively, in the haplotype block that comprises the TGFBR1 coding region (chr9: 100,897,000–100,973,999). Duplicate samples were used to check genotyping quality. Genotyping in quality control samples was re-verified as previously described, and all SNPs and samples described herein passed the required thresholds (5). The TGFBR1*6A/9A SNP, rs11466445, was genotyped using standard PCR conditions and the following primers: 5′-GAGGTGGTCTGGGTTGGA and 5′-AGAGGAGCGGAGCGAG. PCR products were run on an ABI3730XL sequencer (Applied Biosystems), and genotypes were read using GeneMapper (Applied Biosystems). For a subset of CRC cases and controls, lymphoblastoid cell lines were made. cDNA was extracted from these samples using standard methods.

Exon Sequencing. We used conventional PCR sequencing to screen all protein coding exons of TGFBR1 for mutations in 96 CRC patients. Primers used to amplify and sequence these exons are listed in SI Appendix. Sequences were visualized using the 4Peaks program. Sequence changes were confirmed by independent amplification and sequence reactions.

Statistical Analyses. Genotype frequencies at each SNP were tested for deviations from Hardy–Weinberg equilibrium (HWE) and rejected at P < 10⁻⁶. Logistic regression was used to test additive, genotypic, dominant, recessive, and genotypic models of association between genetic variants and disease. The program SNPTEST (www.stats.ox.ac.uk/~marchini/software/gwas/snptest.html) was used to obtain association P values and to estimate regression coefficients (β) in these models along with their standard errors (SEs). Epistasis and HWE test statistics were calculated using the PLINK package (12), STAATA software (Stata Corporation), and SNPTEST. We used genotype data from the CEPH 1000-genome samples (ftp://ftp.1000genomes.ebi.ac.uk and IMPUTE1 software (13) to generate in silico genotypes at additional common polymorphisms in and around (within 50 kb of) TGFBR1 and selected other loci in the TGFβ8 pathway. In silico genotypes at rs11466445 were generated in the Scotland and VQ58 series using reference genotype data from the CORGI cohort and the program IMPUTE2 (8). Meta-analysis of association data was carried out with the program Meta (http://www.stats.ox.ac.uk/~marchini/software/gwas/gwas.html).

Linkage disequilibrium analyses, including estimation of haplotype frequencies and haplotype association χ² tests, were carried out with Haploview (9). Haplotype blocks were defined with the solid spine method incorporated into Haploview. The program VCFTools (kindly provided by Adam Auton) was used to summarize genotypes and linkage disequilibrium (LD) patterns at the TGFBR1 locus in the 1000genome CEPH samples.

ASE. TGFBR1 ASE was examined at rs1590 and rs11466445 in 43 CRC cases and 55 controls in at least two replicates of each sample in all cases. ASE at rs1590 was examined in DNAse-treated lymphoblastoid mRNA from cases and controls using the SNaPshot protocol, as also employed by Valle et al. (7). Owing to some inherent noise in the genotyping signal using this method, we initially compared dosage of the minor allele (C) with that of the major allele (A) in cDNA from informative (heterozygote) samples, based on an expectation that the underlying allelic dosage was 1:1, using non-parametric analysis. We then normalized each cDNA dosage to the geometrical mean of all the dosages derived from the full set of cDNAs. Finally, we normalized each cDNA ratio to the gDNA ratio on a per sample basis. For rs11466445, relative allelic dosages were examined using a fluorescent PCR and Genescan/Genennmapper (Applied Biosystems) analysis of peak areas based on allelic separation by size. Other analyses were performed as for rs1590.

ACKNOWLEDGMENTS. We thank Dr. Boris Pasche for sharing the experimental conditions used in the ASE experiments and Sarah Louise West-Spain for sharing her expertise with IMPUTE. All research was approved by UK Ethics Committees and performed according to the Declaration of Helsinki. Funding was provided by Cancer Research UK, the Wellcome Trust, the Oxford Biomedical Research Centre, and the European Union FP7 Genetic study of Common Hereditary Bowel Cancers in Hispanics and the Americas (CHIBCHA) project.