Multiple common susceptibility variants near BMP pathway loci GREM1, BMP4, and BMP2 explain part of the missing heritability of colorectal cancer

Citation for published version:

Digital Object Identifier (DOI):
10.1371/journal.pgen.1002105

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
PLoS Genetics

Publisher Rights Statement:
This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 15. May. 2019
Multiple Common Susceptibility Variants near BMP Pathway Loci GREM1, BMP4, and BMP2 Explain Part of the Missing Heritability of Colorectal Cancer

Ian P. M. Tomlinson1,3, Luis G. Carvajal-Carmona1,3, Sara E. Dobbins2, Albert Tenesa3, Angela M. Jones1, Kimberley Howarth1, Claire Palles1, Peter Broderick2, Emma E. M. Jaeger1, Susan Farrington3, Annabelle Lewis1, James G. D. Prendergast3, Alan M. Pittman2, Evropi Theodoratou4, Bianca Oliver5, Marion Walker3, Steven Penегar2, Ella Barclay1, Nicola Whiffin2, Lynn Martin1, Stephane Ballereau3, Amy Lloyd2, Maggie Gorman1, Steven Lubbe2, The COGENT Consortium6, The CORGI Collaborators5, The EPICOLON Consortium7, Bryan Howie5, Jonathan Marchini5, Clara Ruiz-Ponte6, Ceres Fernandez-Rozadilla6, Antoni Castells7, Angel Carracedo6, Sergi Castellvi-Bel7, David Duggan8, David Conti9, Jean-Baptiste Cazier1, Harry Campbell10, Oliver Sieber11, Lara Lipton11, Peter Gibbs11, Nicholas G. Martin12, Grant W. Montgomery12, Joanne Young13, Paul N. Baird14, Steven Gallinger15, Polly Newcomb16, John Hopper17, Mark A. Jenkins17, Lauri A. Aaltonen18, David J. Kerr19, Jeremy Cheadle20, Paul Pharoah21, Graham Casey9, Richard S. Houlston22, Malcolm G. Dunlop3

1 Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom, 2 Section of Cancer Genetics, Institute of Cancer Research, Sutton, United Kingdom, 3 Colon Cancer Genetics Group, Institute of Genetics and Molecular Medicine, University of Edinburgh and Medical Research Council Human Genetics Unit, Edinburgh, United Kingdom, 4 The University of Edinburgh Medical School, Edinburgh, United Kingdom, 5 Department of Statistics, University of Oxford, Oxford, United Kingdom, 6 Galician Public Foundation of Genomic Medicine (FPGMX), Centro de Investigacion Biomedica en Red de Enfermedades Raras (CIBERER), Genomics Medicine Group, Hospital Clinico, Santiago de Compostela, University of Santiago de Compostela, Galicia, Spain, 7 Department of Gastroenterology, Hospital Clinic, CIBERehd, IDIBAPS, University of Barcelona, Barcelona, Catalonia, Spain, 8 Translational Genomics Research Institute, Phoenix, Arizona, United States of America, 9 Department of Preventive Medicine, University of Southern California, Los Angeles, California, United States of America, 10 Public Health Sciences, University of Edinburgh, Edinburgh, United Kingdom, 11 Ludwig Colon Cancer Initiative Laboratory, Ludwig Institute for Cancer Research, Royal Melbourne Hospital, Parkville, Australia, 12 Genetic and Molecular Epidemiology Laboratories, Queensland Institute of Medical Research, Brisbane, Australia, 13 Familial Cancer Laboratory, Queensland Institute of Medical Research, Brisbane, Australia, 14 Centre for Eye Research Australia, University of Melbourne, Melbourne, Australia, 15 Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Canada, 16 Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America, 17 Centre for Molecular, Environmental, Genetic, and Analytic Epidemiology, The University of Melbourne, Australia, 18 Department of Medical Genetics, Genome-Scale Biology Research Program, Biomedicum Helsinki, University of Helsinki, Helsinki, Finland, 19 Department of Clinical Pharmacology, University of Oxford, Oxford, United Kingdom, 20 Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, United Kingdom, 21 Cancer Research UK Laboratories, Strangeways Research Laboratory, Department of Oncology, University of Cambridge, Cambridge, United Kingdom

Abstract

Genome-wide association studies (GWAS) have identified 14 tagging single nucleotide polymorphisms (tagSNPs) that are associated with the risk of colorectal cancer (CRC), and several of these tagSNPs are near bone morphogenetic protein (BMP) pathway loci. The penalty of multiple testing implicit in GWAS increases the attraction of complementary approaches for disease gene discovery, including candidate gene- or pathway-based analyses. The strongest candidate loci for additional pathway loci. The penalty of multiple testing implicit in GWAS increases the attraction of complementary approaches for disease gene discovery, including candidate gene- or pathway-based analyses. The strongest candidate loci for additional predisposition SNPs are arguably those already known both to have functional relevance and to be involved in disease risk. To investigate this proposition, we searched for novel CRC susceptibility variants close to the BMP pathway genes GREM1 (15q13.3), BMP4 (14q22.2), and BMP2 (20p12.3) using sample sets totalling 24,910 CRC cases and 26,275 controls. We identified new, independent CRC predisposition SNPs close to BMP4 (rs1957636, \(P = 3.93 \times 10^{-10}\)) and BMP2 (rs4813802, \(P = 4.65 \times 10^{-11}\)). Near GREM1, we found using fine-mapping that the previously-identified association between tagSNPs rs4779584 and CRC actually resulted from two independent signals represented by rs16969681 (\(P = 5.33 \times 10^{-10}\)) and rs11632715 (\(P = 2.30 \times 10^{-10}\)). As low-penetrance predisposition variants become harder to identify—owing to small effect sizes and/or low risk allele frequencies—approaches based on informed candidate gene selection may become increasingly attractive. Our data emphasise that genetic fine-mapping studies can deconvolute associations that have arisen owing to independent correlation of a tagSNP with more than one functional SNP, thus explaining some of the apparently missing heritability of common diseases.
Introduction

Genome-wide association (GWA) studies of colorectal cancer (CRC) have so far identified 14 common, low-risk susceptibility variants [1]. Of these 14 variants, 3 are close to loci that are secreted members of the bone morphogenetic protein (BMP) signalling pathway: GREM1 (rs4779584; BMP4 (rs4444235); and BMP2 (rs961253). In the colon, GREM1 is one of several BMP antagonists produced by sub-epithelial myofibroblasts (ISEMFs). GREM1 binds to and inactivates the ligands BMP2 and BMP4 that are primarily produced by inter-cryptal stromal cells.

Our GWA studies have utilised a primary phase of genome-wide typing of tagging single nucleotide polymorphisms (tagSNPs), followed by larger validation phases of those SNPs with the strongest signals of association. We have previously used relatively stringent statistical thresholds to take SNPs forward into the final validation phases [1]. Whilst such a design has been cost-effective, the use of a lower threshold may have led to the discovery of more CRC SNPs, albeit at the cost of a relatively high type I error rate.

Results

The rs4779584 CRC signal results from two independent SNPs close to GREM1

In order to refine the location of CRC-associated functional variation close to GREM1, BMP4 and BMP2 loci, we genotyped 442 SNPs close to rs4779584, rs4444235 and rs961253 in 4,878 CRC cases and 4,914 controls from the UK2 and Scotland2 sample sets and imputed other SNPs within these regions. No significant localisation of a functional variant was achieved for rs4444235 or rs961253 (Figure S1), but at GREM1, rs16969681 (chr15:30,780,403 bases) had a notably stronger signal of association than rs4779584 (pairwise LD: r² = 0.18, D’ = 0.70) (Figure 1 and Figure S2). We genotyped rs16969681 in additional independent CRC case-control series (UK1, UK4, VQ58, Helsinki, Cambridge and EPICOLON; see Methods). After combined analysis, a significant association between the minor allele at rs16969681 and CRC risk was seen (P = 5.33 × 10⁻⁶; Table 1). Unconditional logistic regression analysis, incorporating sample series as a co-variate, showed that rs16969681 was more strongly associated with CRC than rs4779584, but that the signals were non-independent (for rs16969681, OR = 1.16, 95% CI 1.07–1.25, P = 1.91 × 10⁻⁴; for rs4779584, OR = 1.08, 95% CI 1.02–1.14, P = 5.27 × 10⁻⁵). Akaike information criteria metrics for rs16969681 and rs4779584 respectively were 25608 and 25922, consistent with a superior fit of the risk model incorporating the former SNP. Intriguingly, we found that rs16969681 maps to a site of open chromatin in GREM1-expressing CRC cell lines, raising the possibility that it may be directly functional (Figure S3).

Haplotype risk analysis (Table S2) provided evidence that rs16969681 alleles do not capture all the CRC risk associated with rs4779584. In brief, data from UK2 and Scotland2 showed that the risk alleles at rs16969681 and rs4779584 were defined by a TGTTG haplotype at rs16969681-rs16969862-rs12594722-rs4779584-rs9888701. The TT rs16969681-rs4779584 haplotype was at a frequency of 0.065 in cases and 0.052 in controls (P = 6.29 × 10⁻⁵). However, there appeared to be a residual effect of the T allele at rs4779584, since there was also an elevated risk associated with the CT rs16969681-rs4779584 haplotype (P = 0.026).

We therefore tested the hypothesis that rs4779584 tags two independent risk SNPs at GREM1. We used reverse stepwise logistic regression to search the set of GREM1 SNPs genotyped in the UK2 and Scotland2 samples (Table S1) for associations that were independent of rs16969681 genotype and that captured the residual rs4779584 signal. This analysis led to elimination of rs4779584 from the regression model and identification of a model in which only rs16969681 (P = 1.04 × 10⁻⁴) and another SNP, rs11632715 (P = 1.00 × 10⁻⁴), produced independent signals.
rs11632715 (chr15:30,791,539) is in low LD with rs16969681 ($r^2 = 0.009$, $D^2 = 0.31$) and modest LD with rs4779584 ($r^2 = 0.18$, $D^2 = 0.90$; Figure S2). Through genotyping of additional case-control series, we showed that rs11632715 was significantly associated with CRC risk ($P = 2.80 \times 10^{-10}$; Table 1). Unconditional logistic regression in the 21,139 samples typed for both rs11632715 and rs16969681 provided confirmatory evidence of the independence of the signals (for rs16969681, $P = 1.84 \times 10^{-6}$ and for rs11632715, $P = 6.36 \times 10^{-7}$); these associations were of very similar magnitude to those obtained when each SNP was analysed individually in those sample sets (Figure S2). Incorporation of rs4779584 into the logistic regression model showed that this SNP had a weaker effect than that of either rs16969681 or rs11632715 and did not significantly improve the model fit (Table S3). Inspection of the region containing rs4779584, rs16969681 and rs11632715 (Figure S2) showed that rs4779584 lay within a recombination hotspot. This finding was consistent with our discovery that rs4779584 tags two independent functional variants that are, in turn, tagged by rs16969681 and rs11632715.

Identification of additional, independent CRC susceptibility SNPs near BMP4 and BMP2

The regions analysed for fine mapping encompassed only a minority of the transcribed and flanking regions of GREM1, BMP4 and BMP2. We therefore tested for further independent CRC-associated SNPs around these loci (Table S4) by undertaking a pooled analysis of data from 5 CRC GWA studies (UK1, Scotland1, VQ58, CCFR, Australia) and from the UK2 and Scotland2 samples that had been genotyped at 55,000 SNPs with the strongest evidence of association from meta-analysis of UK1 and Scotland1 (Figure S4) [1]. Since each of the 7 sample sets had been genotyped using different, but overlapping, SNP panels, we performed the combined analysis irrespective of the number of studies in which any SNP had been typed. Figure 2 shows the resulting signals of association from single SNP analysis in this discovery phase.

We prioritised SNPs for further assessment in the replication data sets if they passed two thresholds. First, we required SNPs to show association with CRC at $P<1 \times 10^{-7}$ under the allelic or Cochran-Armitage tests; this was a less stringent threshold than that used in our previously-reported hypothesis-free GWA studies [1,2,3], reflecting the fact that GREM1, BMP4 and BMP2 were strong candidate susceptibility genes. Four SNPs at BMP4, 3 at BMP2 and 9 at GREM1 fulfilled this criterion (Figure 2). Second, since our aim was to test for novel, independent disease variants rather than to refine existing signals of association, we required that SNP genotypes were not correlated with each other or with previously identified risk SNPs ($r^2 < 0.05$, $D^2 < 0.10$). After applying these criteria, one SNP at BMP4 (rs1957636) and one at BMP2 (rs4813802) were retained for subsequent analyses.

rs1957636 and rs4813802 were then genotyped in the validation sample sets (Figure S4), comprising 15,075 CRC cases and 13,296 controls from six independent European case-control series (COIN/NBS, UK3, UK4, Scotland3, Cambridge, Helsinki). After combined analysis, significant associations (Table 2) were shown for both rs1957636, $P = 1.36 \times 10^{-9}$ (OR = 1.08, 95% CI: 1.06–1.11, $P_{het} = 0.009$, $I^2 = 54\%$) and rs4813802, $P = 7.52 \times 10^{-11}$ (OR = 1.09, 95% CI: 1.06–1.12, $P_{het} = 0.42$, $I^2 = 3\%$). In case-only analysis, neither SNP showed any evidence of association with age or sex ($P>0.05$, details not shown).

We used unconditional logistic regression, adjusting for sample size, to test the independence of the two pairs of SNPs at BMP4 and at BMP2. In both cases, each signal remained independent, reflecting the existence of recombination hotspots between the pairs of SNPs at each locus (Figure S5 and Figure S6). For rs4444235 and rs1957636, association $P$-values were respectively $2.09 \times 10^{-7}$ ($F = 47.7\%$) and $3.93 \times 10^{-10}$ ($I^2 = 0\%$). For rs961253 and rs4813802, $P$-values were $1.89 \times 10^{-15}$ ($I^2 = 5\%$) and $4.65 \times 10^{-11}$ ($I^2 = 5\%$). Thus, all 4 SNPs represented independent signals of association with CRC. Further imputation around BMP4 and BMP2 provided no evidence for the alternative possibility that a single variant was tagged by the two SNPs in each region (details not shown).

Annotation of the regions around rs1957636 and rs4813802

rs1957636 (chr14: 53,629,768) is 136 kb upstream of the transcriptional start site of BMP4, 150 kb telomeric to the previously-identified CRC susceptibility SNP, rs4444235 (chr14: 53,480,609), which is downstream of BMP4. There is a recombination hotspot at chr14:33,351,000 (Figure S5) and LD between rs1957636 and rs4444235 is very weak ($r^2 = 0.004$, $D^2 = 0.73$) with rs1957636 in Europeans. Of those SNPs, several mapped to sites of potential functional importance in BMP4 transcription (H3K4Me1, H3K4Me3, DNase1 hypersensitivity, transcription factor ChIP-Seq), as evidenced by the ENCODE regulation tracks ([http://genome.ucsc.edu/cgi-bin/hgTrackU?g heid=171775907&c=chr14&hg=wgEncodeReg]) of the UCSC Genome Browser. For example, rs12432287 ($r^2 = 0.60$, $D^2 = 1.00$ with rs1957636) and rs728425 ($r^2 = 0.69$, $D^2 = 1.00$) lie within a region of apparently high transcription regulatory activity at chr14:33,642,340–53,652,937. Another SNP, rs8011813 ($r^2 = 0.822$, $D^2 = 0.811$), maps within a similar region at chr14:53,728,957–33,731,647. Although none of the SNPs in the region around rs1957636 is the location of a reported eQTL ([http://eqtl.uchicago.edu/cgi-bin/ gbrowse/eqtl/]), no studies relating transcription to SNP genotype have yet been undertaken in the colorectum.
rs4813802 maps to chr20:6,647,595, about 49 kb upstream of BMP2 and 295 kb telomeric of the previously-identified BMP2 CRC susceptibility SNP, rs961253 (chr20:6,352,281). There is very little LD between these two SNPs ($r^2 = 0.000$, $D^2 = 0.017$ from UK1) owing to a recombination hotspot at chr20:6,587,000 (Figure S5). rs4813802 lies within a region of LD flanked by rs727689 (chr20:6,636,405) and rs6117401 (chr20:6,664,097). This region contains 3 spliced ESTs, BX107852, BG822004 and DB094697; none of these has any known functional role or homology to other human or non-human transcripts or genes. The nearest gene to rs4813802 apart from BMP4 is FERMT1 (transcriptional start site, chr20:6,052,191). From HapMap3 release 2 and 1000 Genomes Pilot 1, 29 SNPs were found to be in moderate or greater LD ($r^2 > 0.20$) with rs4813802 in Europeans. Of those SNPs, several in the region chr20:6,636,405–6,647,595 mapped to sites of potential functional importance in BMP2 transcription. None of the SNPs in the area around rs4813802 is the location of a reported eQTL.

**Gene-gene interactions and other SNPs near BMP pathway loci**

Using a case-control logistic regression design, we searched for pairwise gene-gene interactions between 5 SNPs associated with CRC risk (rs4444235, rs1957636, rs961253, rs4813802 and rs4779584). Risks were additive and no evidence of epistasis was detected ($P > 0.2$ for all SNP pairs).

---

**Figure 1. Fine mapping around the known CRC risk SNP close to GREM1 (15q13.3).** Results for meta-analysis of UK2 and Scotland2 are shown. Both significance of association ($-\log_{10}(P)$) and effect size ($\beta$) are presented. The original CRC-associated tagSNPs are shown in blue. The SNP with the clearly strongest association signal is the genotyped SNP rs16969681. doi:10.1371/journal.pgen.1001250.s001
We also searched for evidence of CRC susceptibility alleles at tagSNPs close to other BMP pathway genes. Using the transcribed regions of flanking genes as boundaries, we identified 4,361 tagSNPs mapping to 37 BMP agonist, antagonist and receptor loci (Table S5). However, we found no statistically significant evidence of associations with disease (P > 10^-3 in all cases).

Discussion

We have identified two new CRC predisposition tagSNPs close to BMP4 (rs1957636) and BMP2 (rs4813802). To date, few other loci have been shown at stringent levels of significance to harbour more than one, independent cancer susceptibility variant. One notable exception is the locus proximal to MYC on chromosome 8q24.21 that contains multiple regions independently associated with the risk of prostate and other cancers [4]. Low-penetrance cancer predisposition loci are becoming increasingly hard to identify, owing to small effect sizes and/or low risk allele frequencies – and a return to candidate gene-based approaches may become increasingly attractive. It is true that in the past, candidate gene approaches have generally been unsuccessful at identifying cancer risk loci, but it is now possible to make use of information, such as expression quantitative trait locus identification, that increasingly permits a more considered approach.

Table 1. Genotype counts and statistics of association at rs16969681, rs11632715, and rs4779584.

<table>
<thead>
<tr>
<th>Sample series</th>
<th>Case genotypes</th>
<th>Control genotypes</th>
<th>Odds ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TT</td>
<td>TC</td>
<td>CC</td>
</tr>
<tr>
<td>rs16969681</td>
<td>1</td>
<td>UK2</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Scotland2</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>UK1</td>
<td>12</td>
</tr>
<tr>
<td>P_{overall}</td>
<td>= 5.33 × 10^-4</td>
<td>4</td>
<td>VQNB5</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>EPICOLON</td>
<td>13</td>
</tr>
<tr>
<td>P_{replication}</td>
<td>= 0.013, I_2 = 60.8%</td>
<td>6</td>
<td>Helsinki</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>UK4</td>
<td>6</td>
</tr>
<tr>
<td>P_{replication}</td>
<td>= 2.73 × 10^-4</td>
<td>8</td>
<td>Cambridge</td>
</tr>
</tbody>
</table>

All data sets in which rs16969681 and/or rs11632715 were genotyped are shown. The sample sets genotyped for the SNPs near GREM1 are overlapping, but non-identical, largely because rs11632715 and rs4779584 (but not rs16969681) are present on the proprietary Illumina genome-wide arrays, and also because the Cambridge data set was additionally genotyped for rs16969681. In addition to the overall association test statistics, the P value for the replication phase (excluding UK2 and Scotland2) is shown for rs16969681 and rs11632715. Although there is considerable overlap, the sample sets genotyped here differ somewhat from those typed for the BMP2 and BMP4 SNPs. These differences result entirely from sample and data availability and practical issues of genotyping, including the following: GWAS data but not samples were available from some data sets, so that SNPs such as rs16969681 could not be genotyped in those sample sets; the 1958 Birth Cohort samples were not available at the time of genotyping rs16969681; and for some sample sets, DNA quantity was limiting.

doi:10.1371/journal.pgen.1002105.t001

We also searched for evidence of CRC susceptibility alleles at tagSNPs close to other BMP pathway genes. Using the transcribed regions of flanking genes as boundaries, we identified 4,361 tagSNPs mapping to 37 BMP agonist, antagonist and receptor loci (Table S5). However, we found no statistically significant evidence of associations with disease (P > 10^-3 in all cases).
Multiple BMP SNPs Associate with Colorectal Cancer

**GREM1 region**
- rs4779584
- rs10318

**BMP4 region**
- rs4444235
- rs1957636

**BMP2 region**
- rs961253
- rs4813802
Figure 2. Search for additional colorectal cancer susceptibility SNPs near GREM1, BMP4, and BMP2. Association signals from discovery phase around GREM1, BMP4 and BMP2 are shown. For GREM1, the labelled SNPs are highly correlated tagSNPs originally reported as associated with CRC; these signals are non-independent. For BMP4 and BMP2, the labelled SNPs are the original tagSNPs and the subsequently proven new signals at rs1957636 and rs4813802 respectively. doi:10.1371/journal.pgen.1002105.g002

We have also found good evidence that the original CRC-associated SNP near GREM1, rs4779584 [5], tags two independent functional SNPs, represented by association signals at rs16969681 and rs11632715. This finding emphasises that genetic fine-mapping studies are valuable not only for detecting stronger association signals, but also for deconvoluting tagSNP associations that have arisen owing to independent correlation of the tagSNP with more than one functional SNP. The original rs4779584 tagSNP signal could be described as an example of “synthetic association”, a term that has been used to describe a situation in which multiple, sometimes rare, variants underlie a tagSNP signal [6,7]. Synthetic association can explain some of the apparently missing heritability of complex diseases. Here, we estimate that the 6 SNPs close to the BMP pathway genes contribute approximately 2% of the heritability of CRC, about double that estimated before this study.

Finally, our data provide evidence that GREM1, BMP4 and BMP2 are the targets of the functional variation in each region. Multiple, independently-acting variants close to these loci contribute to CRC risk. Perhaps unexpectedly, there are no detectable genetic interactions among these variants. If the downstream SMAD effectors that function within both the BMP and TGF-beta pathways are included, the components of BMP signalling involved in CRC risk might comprise up to 3 high-penetrance predisposition genes (SMAD4, BMPRIA, GREM1) and 8 low-penetrance variants at GREM1, BMP4, BMP2, SMAD7 and LAMA5 (tagged respectively by rs16969681 and rs11632715, rs444235 and rs1957636, rs961253 and rs4813802, rs4939827, and rs4925386) [1,2,3,5,8,9,10,11]. Collectively these data emphasise the potential importance of genetic variants in the BMP pathway for CRC predisposition.

**Methods**

**Ethics statement**

Collection of blood samples and clinico-pathological information from patients and controls was undertaken with informed consent and ethical review board approval in accordance with the tenets of the Declaration of Helsinki.

**Overall strategy**

The study had two main components: (i) refinement of existing GWAS signals at the GREM1, BMP4 and BMP2 loci using a dense genotyping and imputation approach in several thousand cases and controls previously used for GWAS validation; and (ii) a search for new, independent CRC tagSNPs at the same three loci using a less stringent threshold for validation than used previously, combined with multiple validation sample sets.

**Discovery screen data sets**

UK1 (CorGI) [1] comprised 922 cases with colorectal neoplasia (47% male) ascertained through the Colorectal Tumour Gene Identification (CoRGI) consortium. All had at least one first-degree relative affected by CRC and one or more of the following phenotypes: CRC at age 75 or less; any colorectal adenoma (CRAd) at age 45 or less; ≥3 colorectal adenomas at age 75 or less; or a large (>1 cm diameter) or aggressive (villous and/or severely dysplastic) adenoma at age 75 or less. The 929 controls (45% males) were spouses or partners unaffected by cancer and without a personal family history (to 2nd degree relative level) of colorectal neoplasia. Known dominant polyposis syndromes, HNPCC/ Lynch syndrome or bi-allelic Mlh1 mutation carriers were excluded. All cases and controls were of white UK ethnic origin.

Scotland1 (COGS) [1] included 980 CRC cases (51% male; mean age at diagnosis 49.6 years, SD±6.1) and 1,002 cancer-free population controls (51% male; mean age 51.0 years; SD±5.9). Cases were for early age at onset (age ≤55 years). Known dominant polyposis syndromes, HNPCC/Lynch syndrome or bi-allelic Mlh1 mutation carriers were excluded. Control subjects were sampled from the Scottish population NHS registers, matched by age (±5 years), gender and area of residence within Scotland.

VQ58 comprised 1,832 CRC cases (1,099 males, mean age of diagnosis 62.5 years; SD±10.9) from the VICTOR [12] and QUASAR2 (www.octo.oxford.org.uk/alltrials/trials/q2.html) trials. There were 2,720 population control genotypes (1,391 males) from the Wellcome Trust Case-Control Consortium 2 (WTCC2) 1958 birth cohort (also known as the National Child Development Study), which included all births in England, Wales and Scotland during a single week in 1958 [13].

The Colon Cancer Family Registry (CCFR) data set [14] comprised 1,332 familial CRC cases and 1,084 controls Colon Cancer Family Registry (Colon-CFR) (http://epi.grants.cancer.gov/CFR/about_colon.html). The cases were recently diagnosed CRC cases reported to population complete cancer registries in the USA (Puget Sound, Washington State) who were recruited by the Seattle Familial Colorectal Cancer Registry; in Canada (Ontario) who were recruited by the Ontario Familial Cancer Registry; and in Australia (Melbourne, Victoria) who were recruited by the Australasian Colorectal Cancer Family Study. Controls were population-based and for this analysis were restricted to those without a family history of colorectal cancer.

The Australian study [15] comprised 591 patients treated for CRC at the Royal Melbourne, Western and St Francis Xavier Cabrini Hospitals in Melbourne from 1999 to 2009. The 2,535 controls were derived from Queensland or Melbourne: for the former, the controls came from the Brisbane Twin Nevus Study [16]; for the latter, individuals were participants in the Genes in Myopia study [17]. There was no overlap between the CFR and Australian data sets. Owing to potential residual ethnic heterogeneity within the Melbourne population, for the Australian cohort only we performed an additional screen to minimise heterogeneity after performing principal components analysis (PCA) to remove individuals who clustered with non-CEU individuals (see below). We achieved this by performing PCA on the Australian cases and controls without reference samples of known ancestry. We then paired each case with a control in a 1:1 ratio based on a maximum separation of 0.050 using the first and second eigenvectors. All unpaired samples were excluded, leaving 441 cases and 441 controls in the study. The genomic inflation factor, λGC, was 1.02 after this filtering.

UK2 (NSCCG) [1] consisted of 2,854 CRC cases (58% male; mean age at diagnosis 59.3 years; SD±8.7) ascertained through two ongoing initiatives at the Institute of Cancer Research/Royal Marsden Hospital NHS Trust (RMHNHST) from 1999 onwards - The National Study of Colorectal Cancer Genetics (NSCCG) and
<table>
<thead>
<tr>
<th>Sample series</th>
<th>Case genotypes</th>
<th>Control genotypes</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CC</td>
<td>CT</td>
<td>TT</td>
</tr>
<tr>
<td>rs4444235</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 UK1</td>
<td>247</td>
<td>441</td>
<td>233</td>
</tr>
<tr>
<td>2 Scotland1</td>
<td>220</td>
<td>500</td>
<td>256</td>
</tr>
<tr>
<td>3 UK2</td>
<td>684</td>
<td>1407</td>
<td>761</td>
</tr>
<tr>
<td>4 Scotland2</td>
<td>449</td>
<td>1017</td>
<td>540</td>
</tr>
<tr>
<td>5 VQ58</td>
<td>410</td>
<td>886</td>
<td>503</td>
</tr>
<tr>
<td>6 CCFR</td>
<td>298</td>
<td>595</td>
<td>290</td>
</tr>
<tr>
<td>Australia</td>
<td>108</td>
<td>208</td>
<td>124</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1957636</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 UK1</td>
<td>171</td>
<td>440</td>
<td>310</td>
</tr>
<tr>
<td>2 Scotland1</td>
<td>172</td>
<td>484</td>
<td>320</td>
</tr>
<tr>
<td>3 UK2</td>
<td>475</td>
<td>1383</td>
<td>1002</td>
</tr>
<tr>
<td>4 Scotland2</td>
<td>365</td>
<td>964</td>
<td>677</td>
</tr>
<tr>
<td>5 VQ58</td>
<td>302</td>
<td>891</td>
<td>606</td>
</tr>
<tr>
<td>6 CCFR</td>
<td>206</td>
<td>594</td>
<td>386</td>
</tr>
<tr>
<td>Australia</td>
<td>68</td>
<td>220</td>
<td>152</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs961253</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 UK1</td>
<td>160</td>
<td>418</td>
<td>343</td>
</tr>
<tr>
<td>2 Scotland1</td>
<td>151</td>
<td>460</td>
<td>366</td>
</tr>
<tr>
<td>3 UK2</td>
<td>418</td>
<td>1333</td>
<td>1099</td>
</tr>
<tr>
<td>4 Scotland2</td>
<td>300</td>
<td>939</td>
<td>767</td>
</tr>
<tr>
<td>5 VQ58</td>
<td>219</td>
<td>771</td>
<td>620</td>
</tr>
<tr>
<td>6 CCFR</td>
<td>159</td>
<td>558</td>
<td>469</td>
</tr>
<tr>
<td>Australia</td>
<td>63</td>
<td>218</td>
<td>159</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs4813802</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 UK1</td>
<td>141</td>
<td>425</td>
<td>350</td>
</tr>
<tr>
<td>2 Scotland1</td>
<td>147</td>
<td>452</td>
<td>373</td>
</tr>
<tr>
<td>3 UK2</td>
<td>374</td>
<td>1299</td>
<td>1026</td>
</tr>
<tr>
<td>4 Scotland2</td>
<td>359</td>
<td>918</td>
<td>724</td>
</tr>
</tbody>
</table>
the Royal Marsden Hospital Trust/Institute of Cancer Research Family History and DNA Registry. The 2,822 controls (41% males; mean age 59.8 years; SD±10.8) were the spouses or unrelated friends of patients with malignancies. None had a personal history of malignancy at time of ascertainment. All cases and controls had self-reported European ancestry, and there were no obvious differences in the demography of cases and controls in terms of place of residence within the UK.

Scotland1 (SOCCS) [1] comprised 2,024 CRC cases (61% male; mean age at diagnosis 56.8 years, SD±8.4) and 2,092 population controls (60% males; mean age 67.9 years, SD±9.0) ascertained in Scotland. Cases were taken from an independent, prospective, incident CRC case series and aged <80 years at diagnosis. Control subjects were population controls matched by age (±5 years), gender and area of residence within Scotland.

Replication data sets
UK3 (NSCCG) [1] comprised 7,912 CRC cases (65% male; mean age at diagnosis 59 years, SD±10.2) and 4,398 controls (40% male; mean age 62 years, SD±11.5) ascertained through NSCCG post-2005.

Scotland2 (SOCCS) [1] comprised 1,145 CRC cases (50% male; mean age at diagnosis 53.2 years, SD±15.4) and 2,203 cancer-free population controls (47% male; mean age 51.8 years, SD±11.5). Controls were recruited as part of the Generation Scotland study.

UK4 (CORGI2/BCD) [1] consisted of 621 CRC cases (46% male; mean age at diagnosis 58.3 years, SD±14.1) and 1,121 cancer-free population or spouse controls (45% male; mean age 45.1 years, SD±15.9).

Cambridge/SEARCH consisted of 2,248 CRC cases (56% male; mean age at diagnosis 59.2 years, SD±8.1) and 2,209 controls (42% males; mean age 57.6 years; SD±15.1). Samples were ascertained through the SEARCH (Studies of Epidemiology and Risk Factors in Cancer Heredity, http://www.cancerhelp.org.uk/trials/a-study-looking-at-genetic-causes-of-cancer) study based in Cambridge, UK. Recruitment started in 2000; initial patient contact was through the general practitioner. Control samples were collected post-2003. Eligible individuals were sex- and frequency-matched in five-year age bands to cases.

The COIN samples [18] were 2,151 cases derived from the COIN and COIN-B clinical trials of metastatic CRC. Median age was 63 years. COIN cases were compared against genotypes from 2,501 population controls (1,237 males,) from the WTCCC2 National Blood Service (NBS) cohort (50% male; mean age at diagnosis 53.2 years, SD±15.4).

The Helsinki (FCCPS) study (http://research.med.helsinki.fi/gsb/aaltonen/) comprised 988 cases from a population-based collection centred on south-eastern Finland and 864 population controls from the same collection.

EPICOLON [19] included 1,410 cases matched with the same number of controls collected in a prospective fashion from centres in Spain. Exclusion criteria were Mendelian CRC syndromes and a personal history of inflammatory bowel disease.

In all cases CRC was defined according to the ninth revision of the International Classification of Diseases (ICD) by codes 153–154 and all cases had pathologically proven adenocarcinomas.

Sample preparation and genotyping
DNA was extracted from samples using conventional methods and quantified using PicoGreen (Invitrogen). The VQ, UK1, Scotland1 and Australia GWA cohorts were genotyped using Illumina Hap300, Hap300, or Hap550 arrays. 1958BC and NBS genotyping was performed as part of the WTCCC2 study on Hap1M arrays. The COIN samples were genotyped using Illumina Hap1M or Hap1M-Duo arrays. In UK2 and Scotland2, genotyping was conducted using custom Illumina Infinium arrays according to the manufacturer’s protocols. Some COIN SNPs were typed on custom Illumina Goldengate arrays. To ensure quality of genotyping, a series of duplicate samples was genotyped, resulting in 99.9% concordant calls in all cases.

Other genotyping was conducted using competitive allele-specific PCR KASPar chemistry (KBiosciences Ltd, Hertfordshire, UK), Taqman (Life Sciences, Carlsbad, California) or MassARRAY (Sequenom Inc, San Diego, USA). All primers, probes and conditions used are available on request. Genotyping quality control was tested using duplicate DNA samples within studies and SNP assays, together with direct sequencing of subsets of samples to confirm genotyping accuracy. For all SNPs, >99% concordant results were obtained.

Quality control and sample exclusion
We excluded SNPs from analysis if they failed one or more of the following thresholds: GenCall scores <0.25; overall call rates <95%; MAF<0.01; departure from Hardy-Weinberg equilibrium (HWE) in controls at P<10^{-4} or in cases at P<10^{-8}; outlying in terms of signal intensity or X-Y ratio; discordance between duplicate samples; and, for SNPs with evidence of association, poor clustering on inspection of X-Y plots.

We excluded individuals from analysis if they failed one or more of the following thresholds: duplication or cryptic relatedness to

<table>
<thead>
<tr>
<th>Table 2. Cont.</th>
<th>GG</th>
<th>TG</th>
<th>TT</th>
<th>GG</th>
<th>TG</th>
<th>TT</th>
<th>freqG(cases)</th>
<th>freqG(controls)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P_{\text{hets}} = 0.416, 12 = 3.1%</td>
<td>5</td>
<td>VO58</td>
<td>268</td>
<td>832</td>
<td>694</td>
<td>351</td>
<td>1208</td>
<td>1131</td>
</tr>
<tr>
<td>6</td>
<td>CCFR</td>
<td>176</td>
<td>546</td>
<td>460</td>
<td>112</td>
<td>445</td>
<td>436</td>
<td>0.380</td>
</tr>
<tr>
<td>7</td>
<td>Australia</td>
<td>53</td>
<td>208</td>
<td>169</td>
<td>53</td>
<td>212</td>
<td>173</td>
<td>0.365</td>
</tr>
<tr>
<td>$P_{\text{discovery}} = 2.75 \times 10^{-7}$</td>
<td>8</td>
<td>Helsinki</td>
<td>131</td>
<td>400</td>
<td>395</td>
<td>88</td>
<td>366</td>
<td>375</td>
</tr>
<tr>
<td>9</td>
<td>Cambridge</td>
<td>280</td>
<td>992</td>
<td>707</td>
<td>284</td>
<td>1039</td>
<td>832</td>
<td>0.392</td>
</tr>
<tr>
<td>$P_{\text{replication}} = 3.47 \times 10^{-5}$</td>
<td>10</td>
<td>COIN/NBS</td>
<td>313</td>
<td>982</td>
<td>818</td>
<td>327</td>
<td>1162</td>
<td>1012</td>
</tr>
<tr>
<td>11</td>
<td>UK3</td>
<td>1037</td>
<td>3582</td>
<td>2969</td>
<td>544</td>
<td>2025</td>
<td>1751</td>
<td>0.373</td>
</tr>
<tr>
<td>12</td>
<td>Scotland3</td>
<td>205</td>
<td>523</td>
<td>451</td>
<td>318</td>
<td>1043</td>
<td>878</td>
<td>0.396</td>
</tr>
<tr>
<td>13</td>
<td>UK4</td>
<td>92</td>
<td>265</td>
<td>221</td>
<td>102</td>
<td>285</td>
<td>301</td>
<td>0.368</td>
</tr>
</tbody>
</table>

Results of allelic test of association in all sample sets are shown.
doi:10.1371/journal.pgen.1002105.t002
estimated identity by descent (IBD) >6.25%; overall successfully genotyped SNPs <95%; mismatch between predicted and reported gender; outliers in a plot of heterozygosity versus missingness; and evidence of non-white European ancestry by PCA-based analysis in comparison with HapMap samples (http://hapmap.ncbi.nlm.nih.gov/). We excluded 6 duplicate samples using PCA (see below) within the UK samples that had undergone analysis of over 200 SNPs (UK1, Scotland1, UK2, Scotland2, VQ, 1958BG, NBS, COIN). We excluded duplicates from other UK cohorts on the basis of names (or initials where release of names was not possible) and dates of birth. No duplicates were found from the CCFR or Australian sample sets.

To identify individuals who might have non-northern European ancestry, we merged our case and control data from all sample sets with the 60 European (CEU), 60 Nigerian (YRI), and 90 Japanese (JPT) and 90 Han Chinese (CHB) individuals from the International HapMap Project. For each pair of individuals, we calculated genome-wide identity-by-state distances based on markers shared between HapMap2 and our SNP panel, and used these as dissimilarity measures upon which to perform principal components analysis. Principal components analysis was performed using EigenStrat/SmartPCA using CEU, YRI and HCB HapMap samples as reference. The first two principal components for each individual were plotted and any individual not present in the main CEU cluster (that is, >5% of the PC distance from HapMap CEU cluster centroid) was excluded from subsequent analyses.

We had previously shown the adequacy of the case-control matching and possibility of differential genotyping of cases and controls using Q-Q plots of test statistics in STATA. The inflation factor ω was calculated by dividing the mean of the lower 90% of the test statistics by the mean of the lower 90% of the expected frequencies in the controls from those expected under HWE was calculated using the formula ω = [E(x) / M(x)]^0.5, where E(x) is the population frequency of the minor allele, M(x) is the proportion of the total variation due to heterogeneity and p is the overall familial relative risk.

 SNP selection and genotyping for fine mapping

Regions selected for fine mapping were: chr15:30,733,560–30,802,752; chr14:53,430,973n 53,530,761; and chr20:6,292,730–6,402,661. These corresponded to the haplotype blocks and immediately flanking regions harbouring rs4779584, rs4444235, and rs6161253. To define these haplotype blocks and the recombinant hotspots harbouring these CRC-associated SNPs, we used Haploview and SequenceLDHot. From dbSNP (build 128), we selected all SNPs between the recombinant hotspots flanking the haplotype block. All these SNPs were submitted to Illumina for assay design and those with a design score >0.3 were genotyped on custom arrays in the UK2 and Scotland2 case-control series. In total, we genotyped 81, 42 and 60 SNPs in the 15q13.3, 14q22.2 and 20p12.3 regions respectively. A list of these SNPs is shown in Table S1. Association statistics, using an additive model, were obtained with SNPTST v2 (www.stats.ox.ac.uk/~marchini/software/gwas/snptest.html). We used genotype data from the 1000 Genomes CEPH (http://www.1000genomes.org/) and HapMap3 CEPH and TSI samples (www.hapmap.org/) and the IMPUTE v2 software (https://mathgen.stats.ox.ac.uk/impute/impute_v2.html) to generate imputed genotypes at additional SNPs in all three regions. This imputation resulted in the addition of 74, 113 and 255 markers in the chromosome 15q13.3, 14q22.2 and 20p12.3 regions respectively (for details on imputed and genotyped markers see Table S1). Association meta-analyses only included markers with proper_info scores >0.5, imputed call rates per SNP >0.9 and minor allele frequencies (MAFs) >0.01. Meta-analyses of the two sample sets were carried out with Meta (http://www.stats.ox.ac.uk/~jliu/meta.html) using the genotype probabilities from IMPUTE v2, where a SNP was not directly typed. To test for the presence of additional independent risk alleles in each region, we carried out logistic regression analysis within each region, both pairwise with the original tagSNP and then in a backwards analysis that included all SNPs with evidence of association in the meta-analysis at P<3×10^-5.

Statistical analysis

Association between SNP genotype and disease status was primarily assessed in STATA v10 (http://www.stata.com/) and PLINK v1.07 (http://pngu.mgh.harvard.edu/~purcell/plink/) using allelic and Cochran-Armitage tests (both with 1df) respectively, or by Fisher’s exact test where an expected cell count was <5. Genotypic (2df), dominant (1df) and recessive (1df) tests were also performed. The risks associated with each SNP were estimated by allelic, heterozygous and homozygous odds ratios (ORs) using unconditional logistic regression, and associated 95% confidence intervals (CIs) were calculated.

Joint analysis of data generated from multiple phases was conducted using standard methods for combining raw data based on the Mantel-Haenszel method in STATA and PLINK. The reported meta-analysis statistics were derived from analysis of allele frequencies, and joint ORs and 95% CIs were calculated assuming fixed- and random-effects models. Tests of the significance of the pooled effect sizes were calculated using a standard normal distribution. Cochran’s Q statistic to test for heterogeneity [20] and the I² statistic [21] to quantify the proportion of the total variation due to heterogeneity were calculated. Large heterogeneity is typically defined as I²≥75%. Where significant heterogeneity was identified, results from the random effects model were reported. Alongside, we also performed meta-analysis based on allele dosage (0, 1, 2) and incorporated age and sex as co-variates. Although age and sex are associated with colorectal cancer risk, they were not associated with SNP genotype and did not materially affect the significance of any of the 6 reported associations (details not shown).

We used Haploview software v4.2 (http://www.broadinstitute.org/haploview) to infer the LD structure of the genome in the regions around GREM1, BMP2 and BMP4. The combined effects of pairs of loci identified as associated with CRC risk were investigated by multiple logistic regression analysis in PLINK to test for independent effects of each SNP and stratifying by sample series. Evidence for interactive effects between SNPs (epistasis) was assessed by likelihood ratio test assuming an allelic model in PLINK.

The sibling relative risk attributable to a given SNP was calculated using the formula

\[ \lambda_s = \frac{p(p r_2 + q r_1)^2 + q(p r_1 + q)^2}{(p^2 r_2 + 2 pq r_1 + q^2)^2} \]

where \( p \) is the population frequency of the minor allele, \( q = 1 - p \), and \( r_1 \) and \( r_2 \) are the relative risks (estimated as OR) for heterozygotes and rare homozygotes, relative to common homozygotes [22]. Assuming a multiplicative interaction, the proportion of the familial risk attributable to a SNP was calculated as \( \log(\lambda_s) / \log(\lambda_0) \), where \( \lambda_0 \) is the overall familial relative risk estimated from epidemiological studies of CRC, assumed to be 2.2 [23]. UK2/NSCCG2 samples were used for this estimation. The Akaike information criterion was calculated using the swaic command in STATA.

Genome co-ordinates were taken from the NCBI build 36/hg18 (dbSNP b126).
Multiple BMP SNPs Associate with Colorectal Cancer

Supporting Information

Figure S1 Fine mapping around the known CRC risk SNPs close to (a) BMP4 (14q22) and (b) BMP2 (20p12).

Figure S2 Pairwise linkage disequilibrium between rs4779584, rs16969681, and rs11632715 near GREMI (upper) and position of recombination hotspot (lower).

Figure S3 Histone methylation and acetylation marks upstream of GREMI.

Figure S4 Study design for discovery and validation phases.

Figure S5 Large-scale LD structure in regions around BMP4 and BMP2.

Figure S6 Locations of recombination hotspots in regions around BMP4 and BMP2.

Table S1 SNPs genotyped directly or predicted by imputation in locations of recombination hotspots in regions

Table S2 Haplotype risk analysis at rs16969681 and rs4779584.

Table S3 Logistic regression model analysis of CRC risk and genotypes at rs4779584, rs16969681, and rs11632715.

Table S4 TagSNPs around GREMI, BMP4, and BMP2 analysed for new associations.

Table S5 Additional BMP pathway genes around which tagSNP associations with CRC were analysed.

Text S1 Consortium co-authors.

Acknowledgments

This study made use of genotyping data on the 1958 Birth Cohort and NBS samples, kindly made available by the Investigators of those studies and the Wellcome Trust Case-Control Consortium 2; a full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk/. Finally, we would like to thank all individuals who participated in the study.

Author Contributions


References