Human Molecular Genetics, 2014, Vol. 23, No. 8 **2198–2209**
doi:10.1093/hmg/ddt584
Advance Access published on November 20, 2013

Identification and characterization of functional risk variants for colorectal cancer mapping to chromosome 11q23.1

Michela Biancolella[1,†], Barbara K. Fortini[1,†], Stephanie Tring[1], Sarah J. Plummer[1], Gustavo A. Mendoza-Fandino[1], Jaana Hartiala[1], Michael J. Hitchler[3,‡], Chunli Yan[2], Fredrick R. Schumacher[1], David V. Conti[1], Christopher K. Edlund[1], Houtan Noushmehr[1,2,¶], Simon G. Coetzee[2,¶], Robert S. Bresalier[5], Dennis J. Ahnen[6], Elizabeth L. Barry[7], Benjamin P. Berman[1], Judd C. Rice[4], Gerhard A. Coetzee[1,2] and Graham Casey[1,*]

[1]Department of Preventive Medicine, [2]Department of Urology, [3]Eli Broad Center for Regenerative Medicine and [4]Department of Biochemistry and Molecular Biology, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA, [5]Department of Gastroenterology, Hepatology and Nutrition, University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA, [6]Department of Medicine, Department of Veterans Administration, Eastern Colorado Health Care System and University of Colorado School of Medicine, Denver, CO 80220, USA and [7]Department of Community and Family Medicine, Geisel School of Medicine at Dartmouth, Lebanon, NH 03766, USA

Received July 25, 2013; Revised October 14, 2013; Accepted November 13, 2013

Genome-wide association studies of colorectal cancer (CRC) have identified a number of common variants associated with modest risk, including rs3802842 at chromosome 11q23.1. Several genes map to this region but rs3802842 does not map to any known transcribed or regulatory sequences. We reasoned, therefore, that rs3802842 is not the functional single-nucleotide polymorphism (SNP), but is in linkage disequilibrium (LD) with a functional SNP(s). We performed ChIP-seq for histone modifications in SW480 and HCT-116 CRC cells, and incorporated ChIP-seq and DNase I hypersensitivity data available through ENCODE within a 137-kb genomic region containing rs3802842 on 11q23.1. We identified SNP rs10891246 in LD with rs3802842 that mapped within a bidirectional promoter region of genes *C11orf92* and *C11orf93*. Following mutagenesis to the risk allele, the promoter demonstrated lower levels of reporter gene expression. A second SNP rs7130173 was identified in LD with rs3802842 that mapped to a candidate enhancer region, which showed strong unidirectional activity in both HCT-116 and SW480 CRC cells. The risk allele of rs7130173 demonstrated reduced enhancer activity compared with the common allele, and reduced nuclear protein binding affinity in electromobility shift assays compared with the common allele suggesting differential transcription factor (TF) binding. SNPs rs10891246 and rs7130173 are on the same haplotype, and expression quantitative trait loci (eQTL) analyses of neighboring genes implicate *C11orf53*, *C11orf92* and *C11orf93* as candidate target genes. These data imply that rs10891246 and rs7130173 are functional SNPs mapping to 11q23.1 and that *C11orf53*, *C11orf92* and *C11orf93* represent novel candidate target genes involved in CRC etiology.

*To whom correspondence should be addressed. Tel: +1 3234427865; Fax: +1 3234427886; Email: gcasey@usc.edu
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint first authors.
‡Present address: Radiation Oncology Department, Kaiser Permanente, Los Angeles, CA, USA.
¶Present address: Department of Genetics, Faculty of Medicine at Ribeirão Preto, University of São Paulo, São Paulo 14049-900, Brazil.

INTRODUCTION

Genome-wide association studies (GWAS) of colorectal cancer (CRC) have led to the identification of risk variants mapping to 20 chromosomal regions including 1q41, 3q26.2, 6p21, 8q23.3, 8q24.21, 10p14, 11q13.4, 11q23.1, 12q13.12, 14q22.2, 15q13.3, 16q22.1, 18q21.1, 19q13.11, 20p12.3, 20q13.33 and Xp22.2 in Caucasian populations (P-value $\leq 5 \times 10^{-8}$) (1–12). Most of the variants map to non-coding regions of the genome and the majority of them are not expected to be functional due to the fact that the commercial SNP arrays used in most GWAS use tagSNPs to represent large areas of the genome. Only limited effort has been directed towards identifying and characterizing the functional variants related to GWAS implicated regions. Identifying functional variants remains a major challenge as fine mapping approaches often provide only limited information due to the extent of linkage disequilibrium (LD) across these regions.

One example of a risk variant that maps to a non-coding region is the rs6983267 CRC risk variant mapping to 8q24.21. Identified through several GWAS, rs6983267 maps to an enhancer region that is believed to interact with *MYC* (13–17). However, few CRC-associated tagSNPs are expected to be functional, and there is growing evidence that risk associated SNPs are likely to be in LD with functional SNPs that map within nearby genes, regulatory regions such as enhancers and promoters or long non-coding RNAs (lncRNAs) (16,18). This implies that the identification of functional SNPs could be facilitated through incorporation of chromatin status such as histone modification and/or DNase I hypersensitivity sites into post-GWAS analyses, and that this may represent a powerful approach to identify tissue-specific regulatory elements (19–21).

The common SNP rs3802842 mapping to chromosome 11q23.1 is associated with CRC risk in populations of European descent (CEU) (4). Within the LD block ($r^2 > 0.2$) defined by rs3802842 lie *POU2AF1*, three putative genes, *C11orf53*, *C11orf92* and *C11orf93*, and one microRNA *MIR4491*. There are no known lncRNAs or other transcription products within this region. TagSNP rs3802842 does not map to a coding region or a putative regulatory region, and none of the SNPs in LD ($r^2 > 0.2$) with rs3802842 map to exons. We hypothesized, therefore, that rs3802842 is in LD with a nearby functional SNP(s) within a genetic regulatory element. To test this hypothesis, we generated chromatin immunoprecipitation-sequencing (ChIP-seq) data of histone modifications for SW480 and HCT-116 CRC cells and interrogated ENCODE ChIP-seq and DNase I hypersensitivity data from CRC cell lines to identify putative promoters and enhancers within this region (22).

Using this approach, we identified two potentially functional SNPs, rs10891246 and rs7130173 that are in LD with the CRC-associated SNP rs3802842. In CRC cell lines, rs10891246 mapped to a H3K4me3 peak suggesting it fell within an active promoter of genes *C11orf92* and *C11orf93*, whereas rs7130173 mapped to both H3K4me1 and DNase I hypersensitivity peaks suggesting the presence of an enhancer. We found that a genomic fragment containing the SNP rs10891246 demonstrated bidirectional promoter activity in luciferase-based cellular assays, and the risk allele (A) resulted in lower reporter gene expression activity than the major allele (G). Furthermore, a 1 kb genomic fragment containing SNP rs7130173 showed enhancer activity, and the risk (A) allele led to reduced enhancer activity compared with the common (C) allele. Additionally, an oligonucleotide containing the rs7130173 risk (A) allele showed reduced protein binding affinity compared with the common (C) allele in electromobility shift assays suggesting differential TF binding. We also showed that the rs7130173 risk (A) variant correlated with reduced expression of three predicted open-reading frame genes, *C11orf53*, *C11orf92* and *C11orf93*, in normal human colon tissues. These data implicate rs7130173 and rs10891246 as functional variants mapping to the 11q23.1 region and *C11orf53*, *C11orf92* and *C11orf93* as novel candidate genes involved in CRC etiology.

RESULTS

Region analysis

To investigate the functional basis of the chromosome 11q23.1 association with CRC etiology, we began by identifying potential functional variants using the LD between tagSNP rs3802842 and the other known SNPs within 1 Mb (Fig. 1). There are 89 SNPs in LD with rs3802842 with an $r^2 \geq 0.2$ defining a 137-kb region of chromosome 11 from coordinates 111,119,694 to 111,256,668 (1000 Genomes Project release June 2011; CEU population, hg19) (23,24). Of those, 20 SNPs are in very high LD with rs3802842 with an $r^2 \geq 0.8$. None of these variants mapped to coding exons. Additionally, there are no known lncRNAs in this region (25–27). When we applied FunciSNP (21), an R/Bioconductor tool which systematically identifies correlated SNPs coinciding with chromatin features, to the ENCODE data available for the region and ChIP-seq data we generated (discussed below), we found a short list of SNPs, including rs7130173 and rs10891246 discussed below, residing in potential biofeatures. Thus, our attention turned to variants that may affect genetic regulatory elements.

One potentially active promoter region and several candidate enhancer regions were identified using high-resolution genome wide ChIP-seq profiles for histone modifications including H3K4me1, H3K4me3 and H3K9/14ac (Figure 1A and Supplementary Material, Fig. S1), which we generated for the SW480 and HCT-116 CRC cell lines. DNase I hypersensitive (DNase I)-seq data were also available in the UCSC genome browser for HCT-116 and Caco-2 CRC cells through the ENCODE consortium (22,28–30) and a single DNase I hypersensitivity site in HCT-116 mapped within a candidate enhancer fragment (Fig. 1A). These data were overlaid with SNPs in strong LD with the tagSNP to determine which elements contained potential functional variants. Putative regulatory elements lacking SNPs in LD ($r^2 \geq 0.2$) with the tagSNP were not tested for activity.

Promoter activity determination

The only promoter-correlated histone peak (H3K4me3) in CRC cell line SW480 corresponded to the *C11orf92* and *C11orf93* bidirectional promoter region near their transcriptional start sites (Fig. 1A). The area encompassing the peak (Ch11: 111,169,359–111,171,279, hg19) was cloned into a luciferase assay vector and activity was tested in both SW480 and HCT-116 cells. This region showed bidirectional promoter

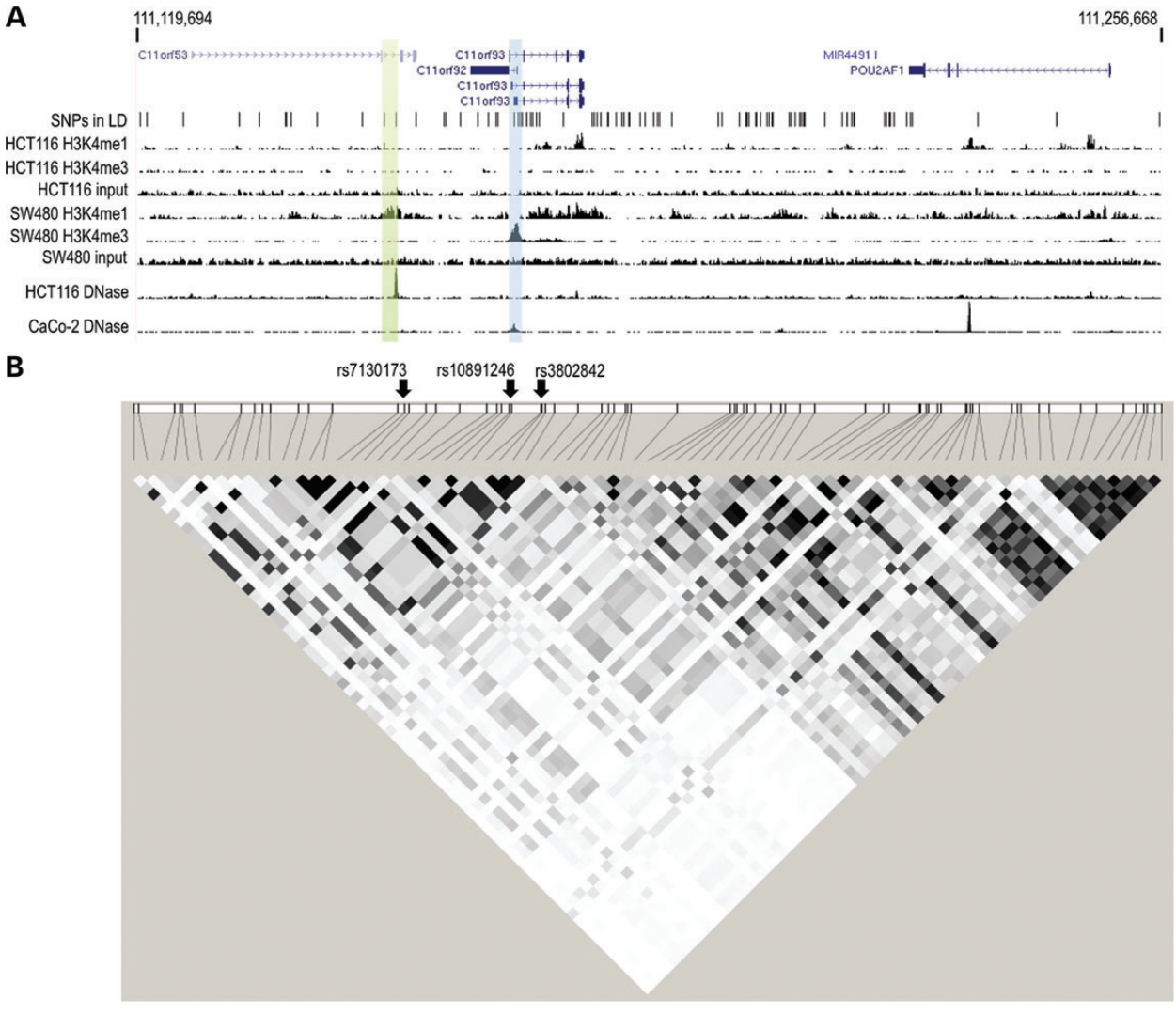

Figure 1. Chromatin features and LD structure for the 11q23.1 CRC GWAS locus. (**A**) Coordinates for the LD block ($r^2 \geq 0.2$, CEU population) surrounding tagSNP rs3802842 are noted above coding transcripts from Genome Browser. SNPs in LD $r^2 \geq 0.2$ with rs3802842 are noted. ChIP-seq tracks for H3K4me1 and H3K4me3 in HCT-116 and SW480 are presented along with DNAse hypersensitivity tracks from UW ENCODE dataset for HCT-116 and the CaCo-2 cell lines (29,30). The blue stripe highlights the promoter fragment described in Figure 2. The green stripe highlights the enhancer fragment described in Figure 3. (**B**) In line with the above panel, a linkage disequilibrium plot for SNPs in the region is shown. Arrows denote the tagSNP rs3802842, and putative functional SNPs rs7130173 and rs10891246. The LD plot was created using 1000 Genomes Project data and created with Haploview. $r^2 = 0$—white, $0 < r^2 < 1$—shades of gray, $r^2 = 1$—black.

activity in both cell lines (Fig. 2A). The cloned H3K4me3 peak contained two SNPs in LD with the tagSNP: rs10891246 ($r^2 = 0.967$, $D' = 1$) and rs7105857 ($r^2 = 0.826$, $D' = 1$). Two constructs were mutated to the minor allele using site-directed mutagenesis and promoter activities were measured (Fig. 2B). Only the rs10891246 SNP demonstrated a statistically significant reduction in bidirectional promoter activity when changed from the major allele G to minor allele A. SNPs rs7122375 ($r^2 = 1$, $D' = 1$ with the tagSNP) and rs11213823 ($r^2 = 0.878$, $D' = 1$ with the tagSNP), included within the cloned fragment but outside of the promoter peak, had no allele-specific effects on promoter activity (Supplementary Material, Fig. S2A).

Enhancer activity determination

The presence of overlapping intronic H3K4me1 and DNase I hypersensitivity peaks suggested the presence of an enhancer, which contained two SNPs in LD with the tagSNP (Fig. 1A). To investigate potential enhancer activity of candidate enhancer regions, several DNA fragments centered over H3K4me1 chromatin marks were PCR amplified from DNA isolated from a normal human lymphoblastoid cell line and cloned in both directions upstream of a luciferase reporter gene driven by the thymidine kinase (TK) minimal promoter (13). Enhancer activities of these fragments were determined following transient transfection in HCT-116 and SW480 CRC cells (Supplementary Material, Fig. S2B). Only one region showed enhancer activity, chr11:111,152,290–111,154,290, and this was seen in both CRC cell lines (Fig. 3A). The enhancer activity was unidirectional and was only observed in the forward orientation. This enhancer region contained two SNPs (rs1987128, $r^2 = 0.77$ and rs7130173, $r^2 = 0.96$) correlated with the risk variant (Fig. 3A). Notably, the rs7130173 SNP mapped within the DNase I hypersensitive site seen in HCT-116 CRC cells (Fig. 1).

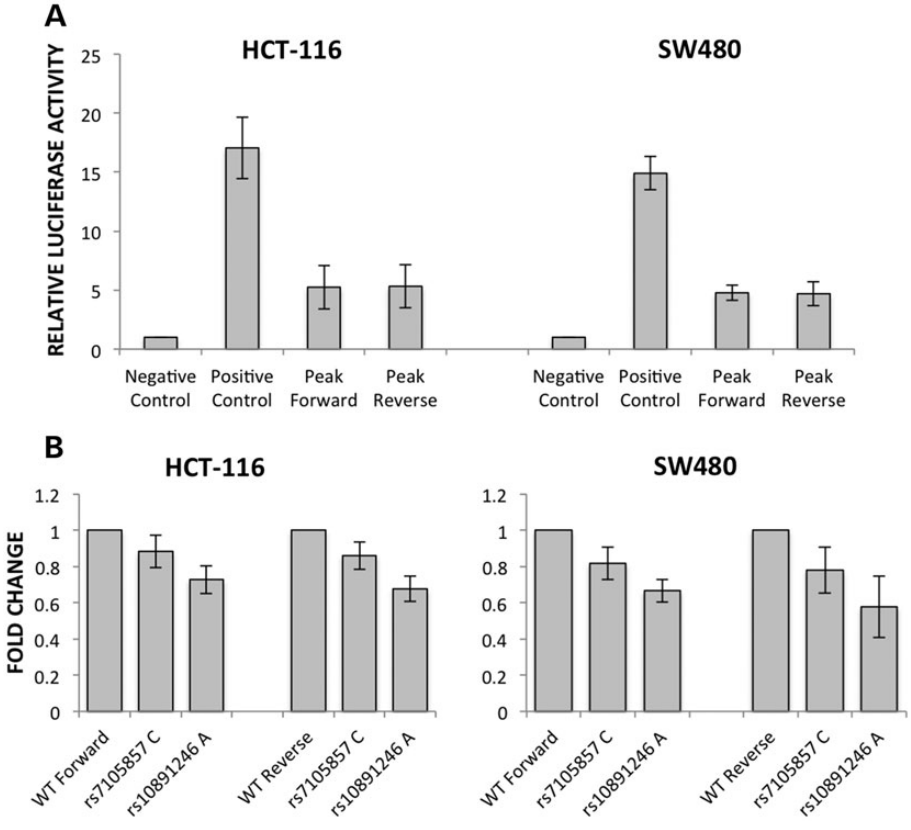

Figure 2. Bidirectional promoter activity for DNA fragment surrounding rs10891246. (**A**) A 2 kb DNA fragment centered over the H3K4me3 peak on 11q23.1 (forward and reverse) along with a negative control sequence from chromosome 10 were tested for promoter activity in HCT-116 and SW480 CRC cells lines in transient assays. Bidirectional promoter activity is seen in both cell lines. (**B**) Two SNPs within the fragment highly correlated with rs3802842 (rs7105857, $r^2 = 0.83$ and rs10891246, $r^2 = 0.96$) were mutated from their common alleles, T and G, to the risk alleles, C and A, respectively. Fold change in activity values are presented as mean $\pm$ SD of at least four-independent transfections. A statistically significant reduction in promoter activity was observed bidirectionally for rs10891246 in both cell lines.

To refine the region responsible for the observed enhancer activity, the 2 kb fragment was cloned as two smaller overlapping fragments $\sim$1 kb each. One of the fragments contained rs1987128, the other contained rs7130173. Both fragments were evaluated for enhancer activity as described above. Enhancer activity (once again unidirectional) was only seen with the fragment containing the rs7130173 variant (Fig. 3A). No enhancer activity was seen in either direction with the fragment carrying the rs1987128 SNP, excluding that SNP as a candidate functional SNP. Given that rs7130173 mapped within a DNase I hypersensitive site, we tested a smaller 300 bp region encompassing the DNase 1 peak for enhancer activity and unidirectional enhancer activity was again observed. This enhancer element displays some tissue specificity as neither the 1 kb-B nor the 300 bp fragment exhibited enhancer activity in the prostate cancer cell line PC-3 (Supplementary Material, Fig. S3).

To determine whether the rs7130173 allele modulated enhancer activity, the rs7130173 SNP was changed from the common (C) to the risk (A) allele by site-directed mutagenesis and enhancer activity was re-assessed (Fig. 3B). Following mutagenesis, the 1 kb fragment containing the risk (A) allele exhibited an enhancer activity statistically lower than that seen for the common (C) allele (Fig. 3B). We obtained the same result when the 2 kb fragment was also mutagenized (data not shown). Surprisingly, when the risk allele was tested in the context of the 300 bp fragment, enhancer activity increased compared with the major allele. This result was seen in both SW480 and HCT-116 cells. This result is not entirely without precedent, as a regulatory element, the enhancer region and its bound TFs by definition interact with other factors bound to additional regulatory elements corresponding to the target genes. The effect of the 700 bp of neighboring sequence upon the action of the smaller enhancer element containing rs7130173 in an allele-specific manner may be key to understanding the CRC mechanism at 11q23.1. As the 2 and 1 kb fragments represent a larger portion of the true genomic context, we postulate that they are more representative of the effect of the SNP *in situ*, and therefore it would be expected that the minor allele of rs7130173 would correlate with a reduction in gene expression of the target gene(s). To test this hypothesis, we looked for eQTLs with rs7130173 among nearby genes.

eQTL analysis

Four genes map within 100 kb of rs7130173 (and within an $r^2 \geq 0.2$ with the tagSNP rs3802842): *POU2AF1* and three putative open reading frame genes, *C11orf53*, *C11orf92* and *C11orf93* (Fig. 4). To determine the relationship between these genes and the candidate functional variant rs7130173, we examined their gene expression in 308 pathologically normal human

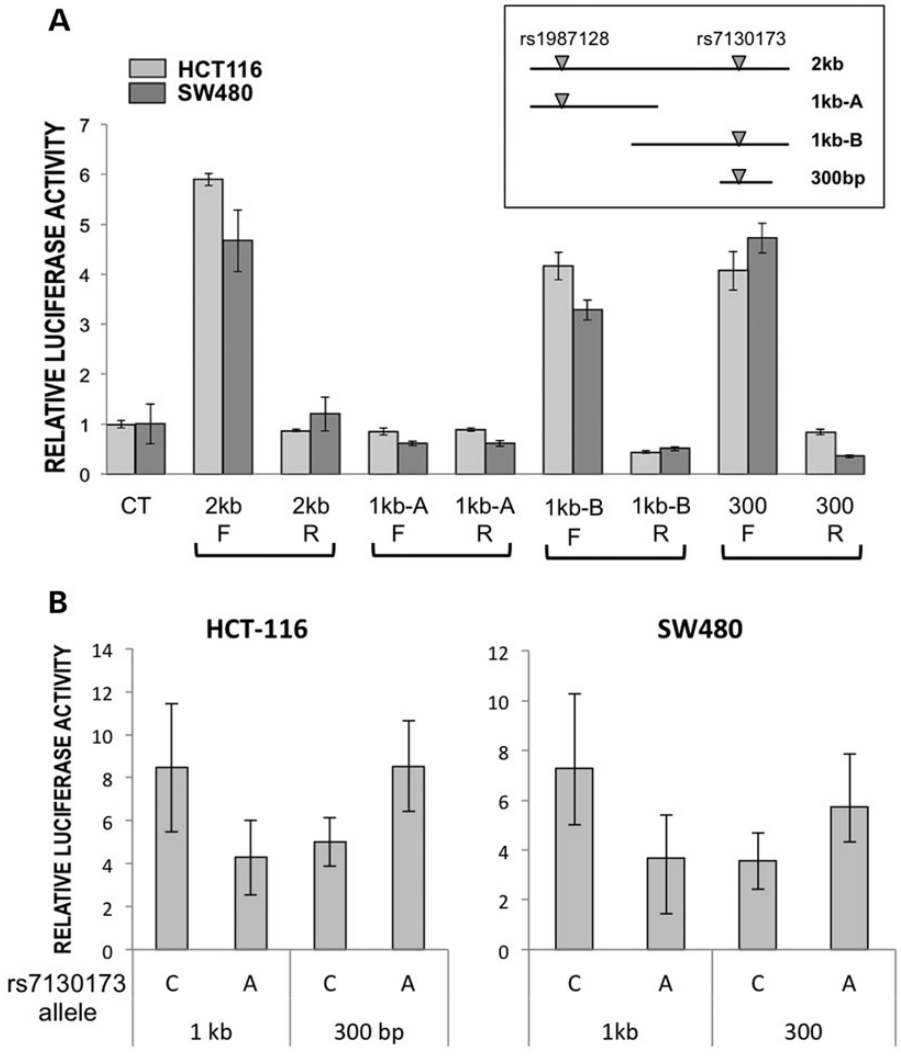

Figure 3. Unidirectional enhancer activity for DNA fragment surrounding rs7130173. (**A**) 2 kb DNA fragment centered over H3K4me1 peak on 11q23.1, chr11:111,152,290–111,154,290, (forward, F, and reverse, R) along with a negative control sequence (CT) from the chromosome 8q24 region (13) were tested for enhancer activity in HCT-116 and SW480 CRC cells lines in transient assays along with pRL-TK Renilla luciferase plasmid. Relative luciferase activity is presented as mean ± SD of at least four independent transfections and reveals unidirectional enhancer activity in both HCT-116 and SW480 CRC cells. The locations of two SNPs within the fragment highly correlated with rs3802842 (rs1987128, $r^2 = 0.77$ and rs7130173, $r^2 = 0.96$) are depicted in the inset. The 2 kb fragment was sub-cloned into two 1 kb fragments, '1 kb-A' and '1 kb-B', and a 300-bp DNA fragment encompassing only the DNase I hypersensitive site, '300 bp'. Enhancer activity was seen only in those fragments containing the common SNP rs7130173 in the forward orientation. (**B**) The common (C) and risk (A) alleles of rs7130173 were tested in enhancer activity assays following site directed mutagenesis of fragments '1 kb-B' and '300 bp' in both HCT-116 and SW480. The risk allele A reduces enhancer activity for the 1 kb fragment, while it increases enhancer activity for the 300 bp fragment.

colon tissue samples obtained by colonoscopy through the Aspirin/Folate Polyp Prevention Study (31–33). We show a statistically significant correlation between reduced expression of *C11orf53* ($P = 1.8 \times 10^{-7}$), *C11orf92* ($P = 1.2 \times 10^{-4}$) and *C11orf93* ($P = 2.6 \times 10^{-9}$) and the risk (A) allele, whereas expression of *POU2AF1* was not statistically different ($P = 0.52$). The two other genes expressed in colon tissues within 400 kb of rs7130173, *LAYN* and *SIK2* were also investigated but expression of these genes did not show any significant relationship with either allele of rs7130173. The promoter SNP rs10891246 and enhancer SNP rs7130173 are in complete LD with one another ($r^2 = 1$, $D' = 1$), and thus the eQTL results are essentially equivalent for both SNPs. When the eQTL analysis was performed for the same genes and the tagSNP rs3802842, we found a similar result to rs7130173 with slightly higher (i.e., less significant) *P*-values [*C11orf53* $P = 2.4 \times 10^{-7}$, *C11orf92* $P = 1.2 \times 10^{-3}$, *C11orf93* $P = 1.4 \times 10^{-8}$, *POU2AF1* $P = 0.90$].

Allele-specific electromobility shift

Considering our observed allele-specific luciferase assay activity (Fig. 3) and the eQTL results in normal colon tissues (Fig. 4), we predicted that rs7130173 was more likely to be a functional SNP than the tagSNP rs3802842. Further, we hypothesized that differential enhancer activity with the risk allele at rs7130173 is due to changes in binding of TFs to the sequence containing and surrounding the SNP. To test this

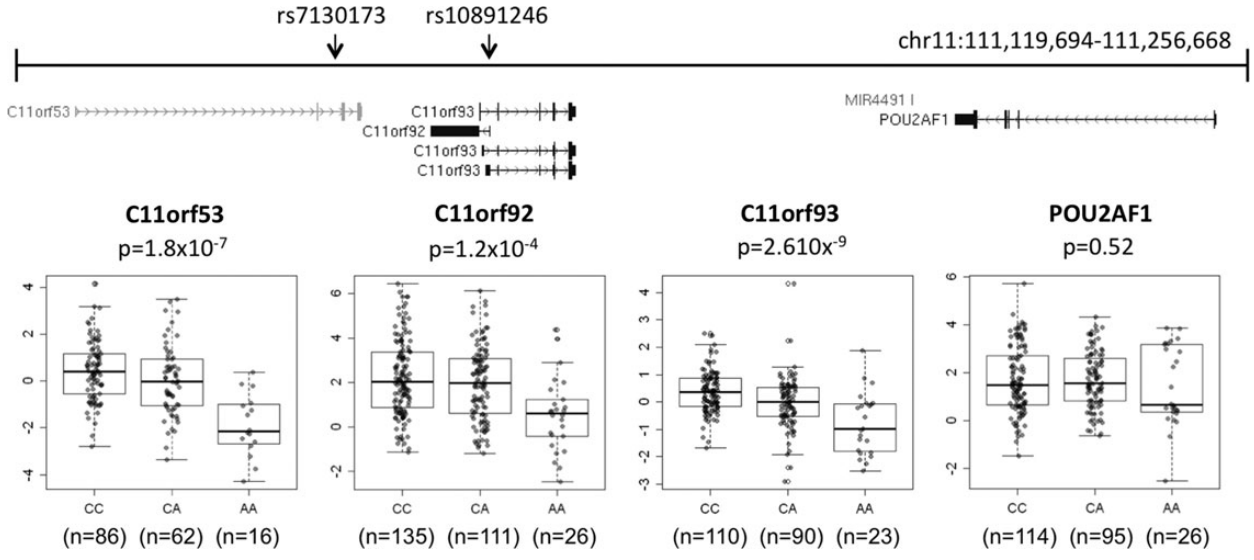

Figure 4. Minor allele of the rs7130173 variant was associated with reduced gene expression levels of three putative genes (*C11orf53*, *C11orf92* and *C11orf93*) mapping to 11q23.1. Relative gene expression quantities (RQs) in *C11orf53* (Hs00736614_m1), *C11orf92* (Hs04186919_m1), *C11orf93* (Hs00416978_m1) and *POU2AF1* (Hs01573371_m1) were measured in normal colon epithelial tissue samples from surveillance colonoscopies. Linear regression was used to estimate the effect of each extra minor allele on log-RQs (additive model). Two-sided *P*-values were obtained from a likelihood ratio test. Levels of log-RQs were plotted as a function of genotypes using a box plot—dot plot overlay. The positions of SNPs rs7130173 and rs10891246 are shown relative to the genes tested.

hypothesis, we used electromobility shift assay (EMSA) to compare the migration patterns of 41mers labeled with near-infrared fluorescent dyes representing both alleles following incubation with nuclear protein extracts from colon cancer cell lines SW480 and HCT-116. The ability to label each allele's probe with a different color allows the direct comparison of binding products to the individual alleles in the same binding reaction and gel lane.

Figure 5A shows the merged color image with the C allele probe in red and the A (risk) allele probe in green; areas with equal amounts of each probe appear yellow. Figure 5B and C shows the individual probe emission channels (700 or 800 nm) of the same gel as black and white images (for reproduction clarity). Lanes 1 and 2 contain the probes alone in the absence of nuclear proteins. In Lanes 3 and 5, we see several distinct protein/DNA bands containing the C allele probe. These bands are faint or absent with the A allele probe (Lanes 4 and 5). To determine if these complexes are specific to the sequence surrounding rs7130173, unlabeled competitor DNA oligonucleotides were added to the reactions in Lanes 6 and 7. The bands marked with arrows are lost upon incubation with unlabeled competitor DNA, and are thus likely specific to the probe sequence. These bands may represent several proteins that bind to the sequence independently, the components of a multiprotein complex, or degradation products of a single protein or protein complex that binds the probe. Both alleles, in 200-fold excess, effectively compete with the C allele for binding factors. Experiments with varying ratios of labeled C and A allele probes show that the binding affinity of the C allele is roughly four times greater than the A allele (data not shown).

Lanes 3 through 7 show the binding of the rs7130173 allele probes to nuclear proteins from colon cancer cell line SW480. Recall that this cell line exhibited histone marks for an active enhancer element containing rs7130173 in our ChIP-seq experiments (Fig. 1). Lanes 8 through 12 show the same set of binding experiments using nuclear extracts prepared from the HCT-116 cell line which did not exhibit strong histone marks for an active enhancer. It is notable that several of the protein/DNA bands appear different when comparing the two cell line extracts. The prominent, specific band in SW480 is less distinct in HCT-116, while there are additional bands in HCT-116 near the top of the lane absent in SW480. The physiological relevance of these cell line differences remains to be elucidated.

Candidate Transcription Factors

Determining the TF responsible for the differential enhancer activity is of great interest to understanding the cancer risk mechanism at chromosome 11q23. Using the Biobase Match tool built upon the TRANSFAC matrix of TF motifs, a list of candidate TFs was generated (Table 1) (34,35). This table catalogues all TFs predicted to have affinity for the region surrounding rs7130173 with the C allele present which show a reduction in binding score if the A allele is present instead. TFs with no change or increase in binding score with the A allele were not included on this list. Future experimental efforts will be needed to determine which of these candidates is responsible for the effects observed in our studies, but several: p53, HIC1, PPARα/γ, ZF5, Spz1 and AP-2rep, are intriguing possibilities due to their known links to carcinogenesis.

DISCUSSION

Identification of functional variants mapping to GWAS regions has proven to be a significant challenge as several hypotheses must be tested for every GWAS locus, requiring a number of complex molecular, genetic and bioinformatics approaches (36). Until now, no candidate functional SNP had been identified at 11q23.1. No non-synonymous SNPs in high LD with the

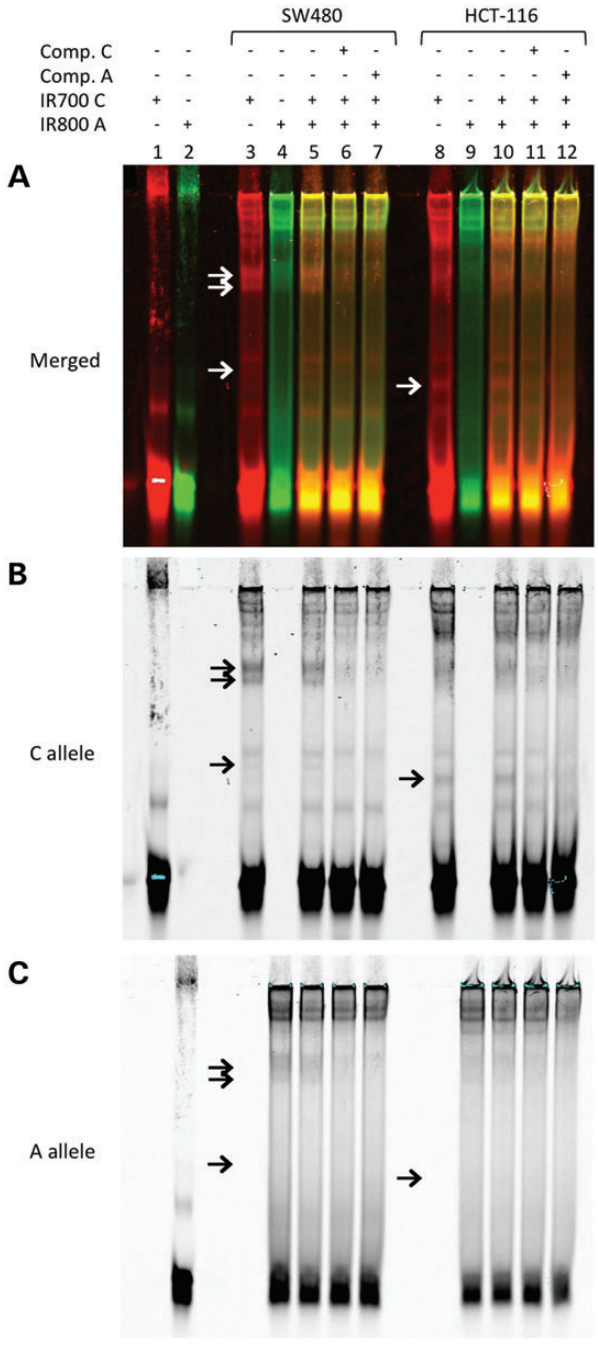

Figure 5. DNA fragments containing the A (risk) allele of rs7130173 show reduced DNA-protein binding compared with fragments containing the C (common) allele. EMSA were performed with IRDye-labeled oligos: C (major) allele (700/red), A (risk) allele (800/green). (**A**)–(**C**) The same gel and image with (A) showing the dual color scan, (**B**) showing the 700 nm channel (red above) in black and white and (C) showing the 800 nm channel (green above) in black and white for clarity of reproduction. Lanes 1 and 2 are probes incubated without nuclear extract. Lanes 3–7 show the probe shift following incubation with 5 μg SW480 nuclear extract. Lanes 8–12 show the probe shift following incubation with 5 μg HCT-116 nuclear extract. Lanes 6 and 11 and Lanes 7 and 12 show the effect of 200-fold excess competition with unlabeled C and A oligos, respectively. Bands that show protein/DNA complexes specific to rs7130173 alleles are noted with arrows.

tagSNP have been found in neighboring genes (3,4). In this comprehensive study, we describe the identification and preliminary characterization of two candidate functional SNPs, rs10891246 and rs7130173, mapping to the 11q23.1 CRC GWAS region. These two SNPs modulate gene expression in cell-based assays in an allele specific manner, with rs10891246 mapping within a promoter region and rs7130173 in an enhancer region. This is manifest in normal human colon tissues as eQTL for three genes, *C11orf53*, *C11orf92* and *C11orf93*. Using bicolor EMSA, we further showed that protein binding is markedly different for the sequences containing the two alleles of rs7130173, with binding affinity of the C allele roughly four times greater than the A allele. This level of reduction in affinity for the risk allele is consistent with the moderate effect the risk allele has on enhancer activity as seen in the luciferase assays. However, a 4-fold change in TF binding affinity may have significant consequences for gene expression in cells and colon tissues. In fact, this was demonstrated for the 8q24 *MYC* enhancer containing rs6983267, the SNP associated with more human cancers, including CRC and prostate cancers, than any other GWAS identified SNP. In *Myc-335*$^{-/-}$ mice lacking the enhancer element, expression of *MYC* in colon crypts was only modestly reduced, but when this is mutation is made in a *APCmin* mouse, there were profound effects on the number of polyps formed (37,38).

There are four candidate genes that map to the risk region *C11orf53*, *C11orf92*, *C11orf93* and *POU2AF1*, with rs7130173 mapping within intron 2 of *C11orf53*, and rs10891246 mapping within the bidirectional promoter between *C11orf92* and *C11orf93*. *POU2AF1*, also known as *BOB.1* or *OBF1*, was an attractive candidate for the CRC risk mechanism as it is a TF known to regulate *TCF12*, a TF linked to colon cancer metastasis and invasion (39,40). However, we did not see a relationship between our SNP alleles and the level of *POU2AF1* in normal colon tissue samples. The three open reading frame genes represent better candidate targets as the eQTL analysis revealed a statistically significant correlation between the rs7130173 risk allele and reduced expression of *C11orf53*, *C11orf92* and *C11orf93*.

Data available through the Cancer Genome Atlas reveal a low frequency (<5% of cases) of overexpression for genes *C11orf53*, *C11orf92* and *POU2AF1* in the ~193 CRC tumors for which RNA-seq and other data are available. In addition, a very low frequency (<1%) of CRC tumors showed somatic mutations in *C11orf53*, *C11orf92*, *C11orf93* or *POU2AF1*. There have been no published reports implicating any of the genes in cancer other than a single study that found *C11orf92* up-regulated in ovarian cancers, and the clinical relevance of these uncharacterized gene products remains unclear. Immunohistochemical staining for *C11orf92* shows the protein in extracellular granules in normal mucosa and at the periphery of colon cancer cells, with the level of protein staining and genotype correlation consistent with our eQTL result (41). Additional functional analyses will be required to determine the role of these candidate genes in CRC development.

While none of the candidate genes have been implicated strongly in CRC, chromosome 11q23–q24 has been reported frequently deleted in CRCs and other tumor types (42). Two published studies report however that allelic imbalance across this region does not correlate with the rs3802842 risk allele (4,43) suggesting any allelic imbalance seen in tumors is independent of the risk effect of this locus.

The 11q23.1 tagSNP rs3802842 (along with the risk variant from 8q23.3) has been implicated in increased risk and earlier

Table 1. Biobase match analysis of differential binding candidates

Sequence	Transcription factor(s)	C allele		Difference with A allele	
		Core	Matrix	Core	Matrix
gagccaccgagccTGCCCcaccaagg	AIRE	0.891	0.842	−0.419	−0.185
gagccaccgagccTGCCCcacca	PPARγ:RXRα, PPARγ	0.853	0.487	−0.408	0.039
agccaccgagcCTGCCcca	PPARα:RXRα	0.628	0.704	−0.299	−0.137
gccaccgaGCCTGccc	AhR:Arnt	0.787	0.761		−0.008
ccaccgagccTGCCCca	LXR, PXR, CAR, COUP, RAR	0.821	0.716	−0.147	−0.064
ccaccgagccTGCCC	VDR	0.843	0.772	−0.271	−0.330
ccgagcCTGCCccacc	MAF	0.800	0.579	−0.200	−0.112
cgagcCTGCCccac	GC box	0.874	0.826	−0.317	−0.222
cgagccTGCCCcaccaag	HIC1	1.000	0.907	−0.025	−0.020
gaGCCTGccccaccaaggga	p53	0.664	0.579		−0.063
gagcctgccccacCAAGGga	p53	0.645	0.567		−0.063
gaGCCTGccccac	ZF5	0.738	0.762		−0.014
agcctgCCCCAccaagggaaa	Muscle initiator sequence-20	1.000	0.805	−0.004	−0.002
gCCTGCccc	CAC-binding protein	0.914	0.883		−0.093
gcctgccCCACCaag	Spz1	0.989	0.891		−0.045
gccTGCCCc	Zic3	0.854	0.803	−0.230	−0.191
gcctGCCCC	LRF	0.943	0.886	−0.311	−0.281
CCTGCccca	TFII-I	0.919	0.868		−0.022
cctGCCCCa	Zic3	0.883	0.756	−0.235	−0.196
cctGCCCC	CACD	0.710	0.740	−0.151	−0.103
cctgCCCCA	LRF	0.923	0.904	−0.310	−0.281
cctgcCCCACca	SREBP	0.678	0.700	−0.259	−0.173
ctgccCCACCaagg	GC box	0.954	0.835		−0.045
ctgCCCCAccaag	MAZR	0.930	0.873	−0.236	−0.159
gCCCCA	Churchill	0.986	0.980	−0.326	−0.323
gccCCACCa	Zic3	0.731	0.710		−0.030
GCCCCac	MOVO-B	0.780	0.789	−0.261	−0.250
cCCCACcaagggaa	RREB-1	0.901	0.639	−0.235	−0.131
cCCCACcaagggaa	Sp3	0.925	0.709	−0.235	−0.090
cccCACCA	CACD	0.768	0.795		−0.092
ccCACCA	AP-2rep	0.700	0.733		−0.030
CCCACca	MOVO-B	0.938	0.901	−0.220	−0.210

Search parameters used for Biobase match program were Matrix library: TRANSFAC MATRIX TABLE, Release 2013.1, Profile: vertebrate_non_redundant.prf, only high-quality matrices, minimize false negatives. Candidate TF list was determined using the C allele ± 20 bp. SNP position is underlined in binding sequence. TFs with no change or increase in core and matrix score upon search using the A allele were ignored. Reduction in core and matrix scores with the A allele are listed in column Group 2.

onset of CRC in Lynch syndrome females in a dose-dependent manner (44–46). Lynch syndrome is the most common inherited form of CRC and is caused by an inherited mutation in one of the DNA mismatch repair (MMR) genes. The individual risk within an MMR-deficient CRC family varies widely among family members and this variability may in part be accounted for by the risk variant on 11q23.1. Interestingly, the HCT-116 CRC cell line, which is MMR deficient, does not show histone marks for an active promoter around rs10891246 nor an active enhancer encompassing rs7130173 in our ChIP-seq data. The histone biofeature marks that sparked our interest in the rs1089126/rs7130173 haplotype in CRC risk were found using the MMR proficient CRC cell line SW480. SW480 carries a truncation in the APC gene, the gene commonly associated with autosomal dominant familial adenomatous polyposis (47). It also appeared that the proteins that bound the major allele of rs7130173 in our EMSA assay were in higher abundance in SW480. Further studies will be needed to determine the relationship between these variants, the MMR pathway, and the expression of *C11orf53*, *C11orf92* and *C11orf93*, not only in CRC cells lines but also in colon stem cells, during colon development or in normal colon crypts. Our eQTL data imply that one or more of the genes *C11orf53*, *C11orf92* and *C11orf93* functions as tumor suppressors since lower levels of expression correlated with the CRC risk alleles.

Characterizing the enhancer at rs7130173 led to the interesting observation that the SNP alleles correlate with differential luciferase activity depending on the length of the fragment assayed in CRC cell lines. Our eQTL result strongly implies that the more physiologically relevant construct is the 1 kb fragment with 700 bp of additional sequence upstream of the enhancer element containing rs7130173. Some of the functional significance of the allele specific activity appears to lie in an interaction between proteins bound to additional upstream regulatory elements in the 700 bp region. The chromatin landscape of the entire region is of interest and further studies with chromatin capture technologies are needed to untangle the effects of each DNA–protein interaction. It is important to note that downstream of rs7130173 (within 100 bp) is a CCCTC-binding factor, CTCF and Rad21/Cohesin binding site in HCT-116 and other cell lines (48–52). Once regarded as only an insulator element, it is now known that CTCF and cohesion can also drive gene expression by linking enhancer elements to their gene targets (53,54). It is therefore possible that the enhancer element encompassing rs7130173 also regulates another gene(s) in *trans* at quite a linear distance.

In this study, ChIP-seq and DNase I hypersensitivity peak chromatin biofeatures were used to identify an active promoter whose activity could be modulated by SNP rs10891246 and

several candidate enhancer regions including one whose activity could be modulated by the SNP rs7130173. Instead of identifying a single causal SNP, we found a haplotype with two SNPs having complementary, or perhaps even synergistic, effects on gene expression. These two SNPs are part of a haplotype including four additional SNPs (rs3087967, rs4477469, rs10789822 and rs7103178, $r^2 = 1$, $D' = 1$, MAF = 0.22, CEU population) in strong LD with the tagSNP rs3802842. Our data suggest that two of these SNPs are functional and contribute to the risk for this region. We did not identify any functional effects for the remaining four SNPs within this haplotype. SNPs rs3087967, rs4477469 and rs10789822 map to intergenic regions lacking histone modifications in our ChIP-seq studies, but rs7103178 lies within the 3′ UTR of *C11orf92*.

The use of biofeature information such as ChIP-seq data to identify candidate enhancers is similar to the approach used previously by members of this team to identify enhancers within the 8q24 region (13) and represents a powerful tool to identify candidate functional SNPs that map to regulatory regions. We have performed a comprehensive analysis of the 11q23.1 region associated with CRC: this region is not highly complex as it has relatively few ChIP-seq peaks corresponding to histone modifications or open chromatin and lacks putative long non-coding RNAs. Applying the same strategy to other CRC GWAS regions represents a powerful systematic and comprehensive approach to understanding CRC risk. This comprehensive evaluation may be quite challenging, however, for those regions with larger LD structure, multiple regions of open and active chromatin, large numbers of genes and lncRNAs. Use of bioinformatics tools, such as FunciSNP, will help facilitate identification of high priority candidates, but it should be noted that comprehensive experimental strategies are still required to elucidate biochemical and genetic mechanisms unique to each region in order to develop the predictive diagnostics and potential therapies that will demonstrate the importance of these studies. This case also highlights the importance of agnostic GWAS approaches in cancer research as this region of chromosome 11q23.1 was formerly overlooked and largely uncharacterized, but is now clearly relevant to understanding colorectal cancer in large populations.

MATERIALS AND METHODS

Cell culture

HCT-116 and SW480 CRC cell lines and the PC-3 prostate cancer cell line were obtained from the American Type Culture Collection (Manassas, VA, USA). HCT-116 and SW480 cells were grown in McCoy's 5A (Mediatech) supplemented with 10% fetal bovine serum (Omega Scientific, Inc.), and 1% penicillin/streptomycin, and incubated at 37°C and 5% CO_2. PC-3 cells were grown in DMEM (Mediatech) supplemented with 10% fetal bovine serum (Omega Scientific, Inc.), and 1% penicillin/streptomycin, and incubated at 37°C and 5% CO_2.

Chromatin-immunoprecipitation

ChIP was performed as previously described (55) using SW480 and HCT-116 CRC cells. Quantitative real-time PCR was performed using SYBR green (Bio-Rad) and Platinum Taq DNA Polymerase (Life Technologies) using an iQ5 Real Time System (Bio-Rad) and the following cycling conditions: 95°C for 5 min, followed by 40 cycles of 95°C for 30 s and 60°C for 30 s. To calculate the enrichment of each immunoprecipitation at a target PCR amplicon a signal threshold was selected and a cycle threshold (Ct) determined. Fold Enrichment for signal was then determined using the formula: fold enrichment = $2^{((\mathrm{Ct\ Ip}) - (\mathrm{Ct\ Input}))}$. The antibodies used were: H3K4me1 (#ab8895, Abcam), H3K4me3 (#04-745, Millipore), Acetylated H3 (#06-599, Millipore), Rabbit IgG control (#ab46540, Abcam) and Histone H3 (#ab1791 Abcam).

Sequencing

ChIP DNA along with ChIP input DNA were prepared as described above, and high-throughput sequencing of the fragments was performed using the Illumina GAII platform. Enrichment for known target sequences was verified by SBYR green quantitative real-time PCR before ChIP and input DNA were sequenced. Libraries for ChIP-seq were prepared, and single-end sequencing of 40 cycles was performed following protocols recommended by Illumina. For SW480, four libraries were sequenced to generate from 18–65 million tags each after standard Illumina quality (PF) filtering. Sequence tags were aligned to UCSC genome assembly hg19 using MAQ (PMID 18714091), and those that were unambiguously mapped to a single position in the genome (Mapping Quality score greater than or equal to 30) were retained for downstream analysis. All sequencing lanes for each library were merged using SAMTOOLS (PMID 19505943), and potential PCR duplicates (tags mapping to identical genomic positions) were removed. On average, 70% of all sequencing tags were unambiguously mappable and passed duplicate filtering. In order to create genomic coverage (wiggle) tracks, each tag was extended by half the median fragment size (200 bp) relative to the mapped strand, and this estimate of the fragment midpoint was used to calculate total tag coverage along the genome.

Plasmids and luciferase assays

DNA fragments corresponding to candidate promoter and enhancer regions were PCR amplified using genomic DNA from a normal lymphoblastoid cell line. Fragments were amplified and cloned using CloneAmp HiFi PCR Premix and the In-Fusion HD cloning kit (Clontech). For the promoter assay, the candidate region Chr11: 111,169,359-111,171,279 was cloned into plasmid pGL4.10 (Promega) at the Kpn1 site. A region of Ch10 with no open chromatin peaks or proximal genes, Chr10: 8,690,709–8,690,917 served as the negative control, while vector gGL4.13 served as the positive control. For the enhancer assays, PCR fragments (2 kb: chr11: 111,152,290–111,154,290; 1 kb-A: chr11: 111,152,290–111,153,217; 1 kb-B: chr11: 111,153,185–111,154,290; 300 bp: chr11: 111,153,991–111,154,290) were subcloned into the Sac II restriction enzyme site upstream of a TK minimal promoter-firefly-luciferase vector in both directions. PCR fragments were sequenced in both directions using Sanger sequencing to confirm the presence of the candidate variants and the absence of PCR amplification-induced mutations (Genewiz). More than three independent constructs in both directions were evaluated. For both promoter and enhancer assays, HCT-116 and

SW480 cells (1.25×10^4 cells/well) were seeded into 96-well plates. Cells were co-transfected with reporter plasmids and constitutively active pRL-TK Renilla luciferase plasmid (Promega) using Lipofectamine 2000 Reagent (Life Technologies) according to the manufacturer's instructions. After 24 h, cells were harvested and extracts were assayed for luciferase activity using the Dual-Luciferase Reporter Assay System (Promega) according to the manufacturer's instructions, and measured using a Dynex MLX Microtiter Plate Luminometer or a Tecan Infinite F200Pro Microplate Reader. The ratio of luminescence from the experimental sample to the control reporter was calculated for each sample, and defined as the relative luciferase activity. The data are presented as mean $\pm$ SD of at least three independent transfection experiments each conducted in triplicate. To assess allele-specific effects, specific SNP alleles were generated by mutagenesis using the QuikChange site-directed mutagenesis kit (Agilent Technologies). Plasmids were sequenced and transfected into the cells as above. At least six-independent clones of each construct were generated, confirmed by Sanger sequencing in both directions and tested for activity as above. The mutagenesis data are presented as mean fold change $\pm$ SD of at least three independent transfection experiments each conducted in triplicate. Two-sided *P*-values between alleles were calculated using the Student's *t*-test.

DNA and RNA isolation

RNA and DNA were extracted from 320 fresh frozen tissue biopsies of normal colorectal mucosa that were obtained from surveillance colonoscopy as part of the Aspirin/Folate Polyp Prevention Study (31–33). The tissue was homogenized using the Precellys Minilys bead mill (Bertin Technologies) and total RNA was extracted using the mirVana kit (Life Technologies). Genomic DNA was extracted using the MELT kit (Life Technologies). The Aspirin/Folate Polyp Prevention Study was approved by ethics committees at the participating institutions and all subjects provided written informed consent.

Genotyping

The genotypes of rs7130173, rs3802842, rs7105857 and rs10891246 were determined using Taqman SNP genotyping assays (Life Technologies) and the Type-it Fast SNP Probe PCR Kit (Qiagen). Assays were read using an Applied Biosystems 7900HT Real Time Instrument and analyzed with the manufacturer's software, SDS2.3 (Life Technologies).

eQTL analysis

Following quantitation, 308 samples had sufficient RNA to proceed to cDNA synthesis using 250 ng of total RNA with the High Capacity RNA-to-cDNA kit (Life Technologies). cDNA samples were then preamplified with TaqMan Preamp Master Mix and 96 TaqMan Gene Expression Assays (Life Technologies), and loaded on a microfluidics chamber for real-time PCR analysis (96.96 Dynamic Array and BioMark HD system, Fluidigm). Relative quantity (RQ) of expression of *C11orf53* (Hs00736614_m1), *C11orf92* (FLJ45803) (Hs04186919_m1), *C11orf93* (Hs00416978_m1) and *POU2AF1* (Hs01573371_m1) was measured using the comparative C_T method ($\Delta\Delta C_T$). Expression was normalized using beta-glucuronidase.

The effect of each extra minor allele (0, 1 or 2) on RQs was evaluated utilizing linear regression and assuming an additive genetic model. RQs were log-transformed prior to analysis to ensure the normal distribution. Test for significance was obtained from a likelihood ratio test and a two-sided *P*-value of < 0.05 was considered statistically significant. Our findings remained statistically significant even after applying overly conservative Bonferroni correction for multiple testing ($P < 4.2 \times 10^{-3}$). Levels of log-RQs were plotted as a function of genotypes using a box plot—dot plot overlay. All data were analyzed with R 12.13.1 using SNPassoc package (56).

Electrophoretic mobility shift assay

Near-infrared dye (Li-Cor Bioscience)-labeled EMSA oligonucleotide probes spanning the SNP rs7130173 were synthesized by Integrated DNA Technologies and annealed in 1× TE (5′-IRDye 700-GTGTGAGCCACCGAGCCTGCCCCACCAA GGGAAACTTTATG-3′ to 5′-IRDye 700-CATAAAGTTTCC CTTGGTGGGGCAGGCTCGGTGGCTCACAC-3′, and 5′-IRDye 800-GTGTGAGCCACCGAGCCTGCACCACCAAGGGAAA CTTTATG-3′ to 5′-IRDye 800-CATAAAGTTTCCCTTGG TGGTGCAGGCTCGGTGGCTCACAC-3) Nuclear extracts from HCT116 and SW480 CRC cell lines were prepared using the NE-PER nuclear and cytoplasmic extraction kit (Pierce, Thermo Scientific). Binding reactions containing 10× binding buffer (100 mM Tris, 500 mM KCl, 10 mM DTT, pH 7.5), 1 μg poly(dI•dC), 2.5 mM DTT/0.25% Tween 20 and 5 μg nuclear extract were preincubated with 20 pmol unlabeled competitor DNA at room temperature for 10 min (5′-GAGCCACCGA GCCTGCCCCACCAAGGG-3′, 5′-GAGCCACCGAGCCT GCACCACCAAGGG-3′, Integrated DNA Technologies). IRDye-labeled oligos were added (50 fmol per oligo) and reactions were incubated in darkness for 20 min at room temperature prior to addition of 10× Orange loading dye (Li-Cor Biosciences). Complexes were resolved on a 6% native polyacrylamide gel run at 200 V for 90 min at 4°C in 0.5× TBE and imaged in the glass plates using a Li-Cor Odyssey Imager.

Analysis of TF binding sites

Forty-one base pair DNA sequences centered on rs7130173, C allele, were scanned for TF binding motifs using the BioBase match tool (BioBase Biological Databases). TRANSFAC MATRIX TABLE, Release 2013.1, profile: vertebrate_non_redundant.prf, only high-quality matrices, minimize false negatives. This analysis was repeated for the rs7130173 A allele. Core and matrix scores were compared between alleles and binding factors with no change in binding scores or increases with the A allele were removed from the candidate TF list.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

We thank Dr John Baron, collaborators and participants of the Aspirin/Folate Polyp Prevention Study for the normal colon tissues used in this study.

Conflict of Interest statement. None declared.

FUNDING

This work was supported by the National Institutes of Health (R01 CA143237 to G.C.; U19 CA148107 to G.C. and G.A.C.; R01 CA059005 to J.B.). The scientific development and funding of this project was in part supported by the USC Norris Comprehensive Cancer Center Support Grant (NCI P30 CA014089) and by the CORECT consortium on behalf of the Genetic Associations and Mechanisms in Oncology (GAME-ON) network. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health.

REFERENCES

1. Houlston, R.S., Cheadle, J., Dobbins, S.E., Tenesa, A., Jones, A.M., Howarth, K., Spain, S.L., Broderick, P., Domingo, E., Farrington, S. *et al.* (2010) Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nat. Genet.*, **42**, 973–977.
2. Houlston, R.S., Webb, E., Broderick, P., Pittman, A.M., Di Bernardo, M.C., Lubbe, S., Chandler, I., Vijayakrishnan, J., Sullivan, K., Penegar, S. *et al.* (2008) Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat. Genet.*, **40**, 1426–1435.
3. Tomlinson, I.P., Webb, E., Carvajal-Carmona, L., Broderick, P., Howarth, K., Pittman, A.M., Spain, S., Lubbe, S., Walther, A., Sullivan, K. *et al.* (2008) A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat. Genet.*, **40**, 623–630.
4. Tenesa, A., Farrington, S.M., Prendergast, J.G., Porteous, M.E., Walker, M., Haq, N., Barnetson, R.A., Theodoratou, E., Cetnarskyj, R., Cartwright, N. *et al.* (2008) Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat. Genet.*, **40**, 631–637.
5. Tomlinson, I., Webb, E., Carvajal-Carmona, L., Broderick, P., Kemp, Z., Spain, S., Penegar, S., Chandler, I., Gorman, M., Wood, W. *et al.* (2007) A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat. Genet.*, **39**, 984–988.
6. Broderick, P., Carvajal-Carmona, L., Pittman, A.M., Webb, E., Howarth, K., Rowan, A., Lubbe, S., Spain, S., Sullivan, K., Fielding, S. *et al.* (2007) A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nat. Genet.*, **39**, 1315–1317.
7. Zanke, B.W., Greenwood, C.M., Rangrej, J., Kustra, R., Tenesa, A., Farrington, S.M., Prendergast, J., Olschwang, S., Chiang, T., Crowdy, E. *et al.* (2007) Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat. Genet.*, **39**, 989–994.
8. Dunlop, M.G., Dobbins, S.E., Farrington, S.M., Jones, A.M., Palles, C., Whiffin, N., Tenesa, A., Spain, S., Broderick, P., Ooi, L.Y. *et al.* (2012) Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. *Nat. Genet.*, **44**, 770–776.
9. Peters, U., Hutter, C.M., Hsu, L., Schumacher, F.R., Conti, D.V., Carlson, C.S., Edlund, C.K., Haile, R.W., Gallinger, S., Zanke, B.W. *et al.* (2012) Meta-analysis of new genome-wide association studies of colorectal cancer risk. *Hum. Genet.*, **131**, 217–234.
10. Peters, U., Jiao, S., Schumacher, F.R., Hutter, C.M., Aragaki, A.K., Baron, J.A., Berndt, S.I., Bezieau, S., Brenner, H., Butterbach, K. *et al.* (2012) Identification of genetic susceptibility loci for colorectal tumors in a genome-wide meta-analysis. *Gastroenterology*, **144**, 799–807.
11. Jiao, S., Hsu, L., Berndt, S., Bezieau, S., Brenner, H., Buchanan, D., Caan, B.J., Campbell, P.T., Carlson, C.S., Casey, G. *et al.* (2012) Genome-wide search for gene-gene interactions in colorectal cancer. *PLoS ONE*, **7**, e52535.
12. Fernandez-Rozadilla, C., Cazier, J.B., Tomlinson, I.P., Carvajal-Carmona, L.G., Palles, C., Lamas, M.J., Baiget, M., Lopez-Fernandez, L.A., Brea-Fernandez, A., Abuli, A. *et al.* (2013) A colorectal cancer genome-wide association study in a Spanish cohort identifies two variants associated with colorectal cancer risk at 1p33 and 8p12. *BMC Genomics*, **14**, 55.
13. Jia, L., Landan, G., Pomerantz, M., Jaschek, R., Herman, P., Reich, D., Yan, C., Khalid, O., Kantoff, P., Oh, W. *et al.* (2009) Functional enhancers at the gene-poor 8q24 cancer-linked locus. *PLoS Genet.*, **5**, e1000597.
14. Pomerantz, M.M., Ahmadiyeh, N., Jia, L., Herman, P., Verzi, M.P., Doddapaneni, H., Beckwith, C.A., Chan, J.A., Hills, A., Davis, M. *et al.* (2009) The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat. Genet.*, **41**, 882–884.
15. Pomerantz, M.M., Beckwith, C.A., Regan, M.M., Wyman, S.K., Petrovics, G., Chen, Y., Hawksworth, D.J., Schumacher, F.R., Mucci, L., Penney, K.L. *et al.* (2009) Evaluation of the 8q24 prostate cancer risk locus and MYC expression. *Cancer Res.*, **69**, 5568–5574.
16. Meyer, K.B., Maia, A.T., O'Reilly, M., Ghoussaini, M., Prathalingam, R., Porter-Gill, P., Ambs, S., Prokunina-Olsson, L., Carroll, J. and Ponder, B.A. (2011) A functional variant at a prostate cancer predisposition locus at 8q24 is associated with PVT1 expression. *PLoS Genet.*, **7**, e1002165.
17. Tuupanen, S., Yan, J., Turunen, M., Gylfe, A.E., Kaasinen, E., Li, L., Eng, C., Culver, D.A., Kalady, M.F., Pennison, M.J. *et al.* (2012) Characterization of the colorectal cancer-associated enhancer MYC-335 at 8q24: the role of rs67491583. *Cancer Genet.*, **205**, 25–33.
18. Visser, M., Kayser, M. and Palstra, R.J. (2012) HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. *Genome Res.*, **22**, 446–455.
19. Visel, A., Blow, M.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F. *et al.* (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, **457**, 854–858.
20. Rahimov, F., Marazita, M.L., Visel, A., Cooper, M.E., Hitchler, M.J., Rubini, M., Domann, F.E., Govil, M., Christensen, K., Bille, C. *et al.* (2008) Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip. *Nat. Genet.*, **40**, 1341–1347.
21. Coetzee, S.G., Rhie, S.K., Berman, B.P., Coetzee, G.A. and Noushmehr, H. (2012) FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs. *Nucleic Acids Res.*, **40**, e139.
22. ENCODE (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636–640.
23. Barrett, J.C. (2009) *Haploview: Visualization and Analysis of SNP Genotype Data*. Cold Spring Harbor Protocols, CSHL Press **2009**, pdb ip71.
24. Barrett, J.C., Fry, B., Maller, J. and Daly, M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.
25. Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P. and Burge, C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
26. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.
27. Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A. and Rinn, J.L. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.*, **25**, 1915–1927.
28. ENCODE (2009) Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature*, **457**, 1028–1032.
29. Sabo, P.J., Hawrylycz, M., Wallace, J.C., Humbert, R., Yu, M., Shafer, A., Kawamoto, J., Hall, R., Mack, J., Dorschner, M.O. *et al.* (2004) Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proc. Natl. Acad Sci. USA*, **101**, 16837–16842.
30. Sabo, P.J., Kuehn, M.S., Thurman, R., Johnson, B.E., Johnson, E.M., Cao, H., Yu, M., Rosenzweig, E., Goldy, J., Haydock, A. *et al.* (2006) Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat. Methods*, **3**, 511–518.
31. Baron, J.A., Cole, B.F., Sandler, R.S., Haile, R.W., Ahnen, D., Bresalier, R., McKeown-Eyssen, G., Summers, R.W., Rothstein, R., Burke, C.A. *et al.* (2003) A randomized trial of aspirin to prevent colorectal adenomas. *New England J. Med.*, **348**, 891–899.

32. Figueiredo, J.C., Grau, M.V., Wallace, K., Levine, A.J., Shen, L., Hamdan, R., Chen, X., Bresalier, R.S., McKeown-Eyssen, G., Haile, R.W. *et al.* (2009) Global DNA hypomethylation (LINE-1) in the normal colon and lifestyle characteristics and dietary and genetic factors. *Cancer Epidemiol. Biomarkers Prevent.*, **18**, 1041–1049.
33. Wallace, K., Grau, M.V., Levine, A.J., Shen, L., Hamdan, R., Chen, X., Gui, J., Haile, R.W., Barry, E.L., Ahnen, D. *et al.* (2010) Association between folate levels and CpG Island hypermethylation in normal colorectal mucosa. *Cancer Prevent. Res.*, **3**, 1552–1564.
34. Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.
35. Kel, A.E., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V. and Wingender, E. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, **31**, 3576–3579.
36. Freedman, M.L., Monteiro, A.N., Gayther, S.A., Coetzee, G.A., Risch, A., Plass, C., Casey, G., De Biasi, M., Carlson, C., Duggan, D. *et al.* (2011) Principles for the post-GWAS functional characterization of cancer risk loci. *Nat. Genet.*, **43**, 513–518.
37. Sur, I., Tuupanen, S., Whitington, T., Aaltonen, L.A. and Taipale, J. (2013) Lessons from functional analysis of genome-wide association studies. *Cancer Res.*, **73**, 4180–4184.
38. Sur, I.K., Hallikas, O., Vaharautio, A., Yan, J., Turunen, M., Enge, M., Taipale, M., Karhu, A., Aaltonen, L.A. and Taipale, J. (2012) Mice lacking a Myc enhancer that includes human SNP rs6983267 are resistant to intestinal tumors. *Science*, **338**, 1360–1363.
39. Bordon, A., Bosco, N., Du Roure, C., Bartholdy, B., Kohler, H., Matthias, G., Rolink, A.G. and Matthias, P. (2008) Enforced expression of the transcriptional coactivator OBF1 impairs B cell differentiation at the earliest stage of development. *PloS ONE*, **3**, e4007.
40. Lee, C.C., Chen, W.S., Chen, C.C., Chen, L.L., Lin, Y.S., Fan, C.S. and Huang, T.S. (2012) TCF12 protein functions as transcriptional repressor of E-cadherin, and its overexpression is correlated with metastasis of colorectal cancer. *J. Biol. Chem.*, **287**, 2798–2809.
41. Peltekova, V.D., Lemire, M., Qazi, A.M., Zaidi, S.H., Trinh, Q.M., Bielecki, R., Rogers, M., Hodgson, L., Wang, M., D'Souza, D.J. *et al.* (2013) Identification of genes expressed by immune cells of the colon that are regulated by colorectal cancer-associated variants. *Int. J. Cancer*, doi: 10.1002/ijc.28557. [Epub ahead of print].
42. Ong, D.C., Ho, Y.M., Rudduck, C., Chin, K., Kuo, W.L., Lie, D.K., Chua, C.L., Tan, P.H., Eu, K.W., Seow-Choen, F. *et al.* (2009) LARG at chromosome 11q23 has functional characteristics of a tumor suppressor in human breast and colorectal cancer. *Oncogene*, **28**, 4189–4200.
43. Niittymaki, I., Tuupanen, S., Li, Y., Jarvinen, H., Mecklin, J.P., Tomlinson, I.P., Houlston, R.S., Karhu, A. and Aaltonen, L.A. (2011) Systematic search for enhancer elements and somatic allelic imbalance at seven low-penetrance colorectal cancer predisposition loci. *BMC Med. Genet.*, **12**, 23.
44. Wijnen, J.T., Brohet, R.M., van Eijk, R., Jagmohan-Changur, S., Middeldorp, A., Tops, C.M., van Puijenbroek, M., Ausems, M.G., Gomez Garcia, E., Hes, F.J. *et al.* (2009) Chromosome 8q23.3 and 11q23.1 variants modify colorectal cancer risk in Lynch syndrome. *Gastroenterology*, **136**, 131–137.
45. Talseth-Palmer, B.A., Brenne, I.S., Ashton, K.A., Evans, T.J., McPhillips, M., Groombridge, C., Suchy, J., Kurzawski, G., Spigelman, A., Lubinski, J. *et al.* (2012) Colorectal cancer susceptibility loci on chromosome 8q23.3 and 11q23.1 as modifiers for disease expression in Lynch syndrome. *J. Med. Genet.*, **48**, 279–284.
46. Giraldez, M.D., Lopez-Doriga, A., Bujanda, L., Abuli, A., Bessa, X., Fernandez-Rozadilla, C., Munoz, J., Cuatrecasas, M., Jover, R., Xicola, R.M. *et al.* (2012) Susceptibility genetic variants associated with early-onset colorectal cancer. *Carcinogenesis*, **33**, 613–619.
47. Yang, J., Zhang, W., Evans, P.M., Chen, X., He, X. and Liu, C. (2006) Adenomatous polyposis coli (APC) differentially regulates beta-catenin phosphorylation and ubiquitination in colon cancer cells. *J. Biol. Chem.*, **281**, 17751–17757.
48. Fields, S. (2007) Molecular biology. Site-seeing by sequencing. *Science*, **316**, 1441–1442.
49. Johnson, D.S., Mortazavi, A., Myers, R.M. and Wold, B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
50. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
51. Li, Q.B., Brown, J.B., Huang, H. and Bickel, P.J. (2011) Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.*, **5**, 1752–1779.
52. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
53. Holwerda, S.J.B. and de Laat, W. (2013) CTCF: the protein, the binding partners, the binding sites and their chromatin loops. *Phil. Trans. R. Soc. B Biol. Sci.*, 368, 20120369.
54. Rubio, E.D., Reiss, D.J., Welcsh, P.L., Disteche, C.M., Filippova, G.N., Baliga, N.S., Aebersold, R., Ranish, J.A. and Krumm, A. (2008) CTCF physically links cohesin to chromatin. *Proc. Natl. Acad. Sci. USA*, **105**, 8309–8314.
55. Hitchler, M.J. and Rice, J.C. (2011) Genome-wide epigenetic analysis of human pluripotent stem cells by ChIP and ChIP-Seq. *Methods Mol. Biol.*, **767**, 253–267.
56. Gonzalez, J.R., Armengol, L., Sole, X., Guino, E., Mercader, J.M., Estivill, X. and Moreno, V. (2007) SNPassoc: an R package to perform whole genome association studies. *Bioinformatics*, **23**, 644–645.