Human Molecular Genetics, 2013, Vol. 22, No. 24 **5048–5055**
doi:10.1093/hmg/ddt337
Advance Access published on July 12, 2013

Fine-mapping of genome-wide association study-identified risk loci for colorectal cancer in African Americans

Hansong Wang[1,*], Christopher A. Haiman[2], Terrilea Burnett[1], Barbara K. Fortini[2], Laurence N. Kolonel[1], Brian E. Henderson[2], Lisa B. Signorello[3,4], William J. Blot[3,4], Temitope O. Keku[5], Sonja I. Berndt[6], Polly A. Newcomb[7], Mala Pande[8], Christopher I. Amos[9], Dee W. West[10], Graham Casey[2], Robert S. Sandler[5], Robert Haile[11], Daniel O. Stram[2] and Loïc Le Marchand[1]

[1]Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI, USA, [2]Department of Preventive Medicine, Keck School of Medicine and Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA, USA, [3]International Epidemiology Institute, Rockville, MD, USA, [4]Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center and Vanderbilt-Ingram Cancer Center, Vanderbilt University, Nashville, TN, USA, [5]Center for Gastrointestinal Biology and Disease, University of North Carolina, Chapel Hill, NC, USA, [6]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA, [7]Fred Hutchinson Cancer Research Center, Seattle, WA, USA, [8]Department of Epidemiology, The University of Texas M.D. Anderson Cancer Center, Houston, TX, USA, [9]Department of Community and Family Medicine, Geisel School of Medicine, Dartmouth College, Lebanon, NH, USA, [10]Cancer Prevention Institute of California, Fremont, CA, USA [11]Stanford Cancer Institute, Stanford, CA, USA

Received April 17, 2013; Revised June 29, 2013; Accepted July 9, 2013

Genome-wide association studies of colorectal cancer (CRC) in Europeans and Asians have identified 21 risk susceptibility regions [29 index single-nucleotide polymorphisms (SNPs)]. Characterizing these risk regions in diverse racial groups with different linkage disequilibrium (LD) structure can help localize causal variants. We examined associations between CRC and all 29 index SNPs in 6597 African Americans (1894 cases and 4703 controls). Nine SNPs in eight regions (5q31.1, 6q26-q27, 8q23.3, 8q24.21, 11q13.4, 15q13.3, 18q21.1 and 20p12.3) formally replicated in our data with one-sided *P*-values <0.05 and the same risk directions as reported previously. We performed fine-mapping of the 21 risk regions (including 250 kb on both sides of the index SNPs) using genotyped and imputed markers at the density of the 1000 Genomes Project to search for additional or more predictive risk markers. Among the SNPs correlated with the index variants, two markers, rs12759486 (or rs7547751, a putative functional variant in perfect LD with it) in 1q41 and rs7252505 in 19q13.1, were more strongly and statistically significantly associated with CRC ($P < 0.0006$). The average per allele risk was improved using the replicated index variants and the two new markers (odds ratio = 1.14, $P = 6.5 \times 10^{-16}$) in African Americans, compared with using all index SNPs (odds ratio = 1.07, $P = 3.4 \times 10^{-10}$). The contribution of the two new risk SNPs to CRC heritability was estimated to be 1.5% in African Americans. This study highlights the importance of fine-mapping in diverse populations.

INTRODUCTION

Genome-wide association studies (GWAS) in European and East Asian populations have identified 21 regions [characterized by 29 index single-nucleotide polymorphisms (SNPs)] associated with the risk of colorectal cancer (CRC) with *P*-values $<5 \times 10^{-8}$ (1–12). Like for other diseases/traits, many variants identified from CRC GWAS fall outside of coding regions with

*To whom correspondence should be addressed at: University of Hawaii Cancer Center, 701 Ilalo St, Honolulu, HI 96813, USA. Tel: +1 808564 5846; Email: hansongw@hawaii.edu

no known biological function and they together only account for a small proportion of CRC heritability (9,13). Findings from the first wave of GWAS are expected to open up new windows into disease biology; however, at present, the identities of the causal variants remain undetermined for most risk loci. Studies in populations of African ancestry are likely to provide additional insights in that: (i) it is possible to narrow the large chromosomal regions of association and pinpoint the best risk-defining variants, a benefit of the shorter average linkage disequilibrium (LD) blocks in this population (14); and (ii) the greater genetic diversity in African-descent populations may help to identify additional susceptibility variants.

In this study, we tested the 29 GWAS-identified CRC risk SNPs (21 regions) in a large sample set of African Americans (1894 cases and 4703 controls) and searched for additional or more predictive risk variants in the surrounding regions ($\pm$250 kb) using genotyped and imputed markers at the density of the 1000 Genomes Project (1KGP).

RESULTS

The African American CRC cases and controls in this analysis are from seven studies/centers in the USA (Supplementary Material, Methods and Table S1). Samples were genotyped using the Illumina 1M-Duo bead array (except 170 subjects on Omni 2.5 M). After quality control (QC) exclusions (see Methods), the analysis was performed on 1894 cases and 4703 controls; the number of right colon, left colon and rectal cases (mutually exclusive) was 778, 500 and 399, respectively. On average, cases were older than controls (mean 68 and 62 years, respectively) and the proportion of females was higher in cases (49.6%, compared to 35.2% in controls). Cases and controls were similar in global ancestry distribution based on principal components (PCs) (Methods, Supplementary Material, Fig. S1). We adjusted for age, sex and the first four most important PCs in logistic regression and little inflation of test statistics due to population stratification was observed (genomic control $\lambda = 1.04$).

Among the 29 CRC risk variants (Table 1, Supplementary Material, Table S2), 27 were directly genotyped and 2 were imputed with $R^2 \geq 0.90$. All variants are common [minor allele frequencies (MAFs) > 0.05] in African Americans. For 10 SNPs, minor alleles in Europeans or East Asians are major alleles in African Americans. The average difference in allele frequency was 0.18 (all $\leq$ 0.41) (Fig. 1).

Table 1. Associations of the 29 index CRC risk variants in African Americans, with adjustment of age, gender and the first four principal components

SNP	Locus	BP (HG19)	A1	A2	In published GWAS		In African Americans			
					FRQ[a]	OR[a]	FRQ[a]	OR[a] (95% CI)	*P* (two-sided)	OR[a,b]
rs6691170	1q41	222 045 446	T	G	0.36	1.06	0.33	1.02 (0.93, 1.11)	0.72	1.02
rs6687758	1q41	222 164 948	G	A	0.20	1.09	0.19	0.99 (0.90, 1.10)	0.86	0.99
rs10936599	3q26.2	169 492 101	T	C	0.22	0.93	0.08	0.96 (0.82, 1.12)	0.62	0.98
rs647161[c,d]	5q31.1	134 499 092	A	C	0.31	1.17	0.55	1.14 (1.05, 1.24)	0.002	1.14
rs1321311	6p21	36 622 900	A	C	0.23	1.10	0.39	0.96 (0.89, 1.04)	0.34	0.96
rs7758229[d,e]	6q26-q27	160 840 252	T	G	0.23	1.28	0.11	1.08 (0.95, 1.23)	0.056	1.21
rs16892766	8q23.3	117 630 683	C	A	0.07	1.25	0.13	1.17 (1.05, 1.32)	0.0058	1.18
rs10505477	8q24.21	128 407 443	G	A	0.54	0.85	0.19	0.91 (0.82, 1.01)	0.09	0.92
rs6983267	8q24.21	128 413 305	T	G	0.51	0.83	0.13	0.87 (0.76, 0.99)	0.029	0.87
rs7014346	8q24.21	128 424 792	A	G	0.37	1.19	0.39	1.05 (0.97, 1.13)	0.27	1.05
rs10795668	10p14	8 701 219	A	G	0.33	0.89	0.07	0.98 (0.83, 1.15)	0.77	1.00
rs3824999	11q13.4	74 345 550	G	T	0.50	1.08	0.20	1.15 (1.03, 1.27)	0.009	1.12
rs3802842	11q23.1	111 171 709	C	A	0.29	1.11	0.35	1.04 (0.95, 1.12)	0.40	1.03
rs10774214[d]	12p13.32	4 368 352	C	T	0.65	0.85	0.38	0.95 (0.87, 1.03)	0.19	0.95
rs7136702	12q13.13	50 880 216	C	T	0.65	0.94	0.41	0.97 (0.89, 1.05)	0.41	0.96
rs11169552	12q13.13	51 155 663	T	C	0.28	0.92	0.10	1.04 (0.91, 1.19)	0.58	1.07
rs4444235	14q22.2	54 410 919	C	T	0.46	1.11	0.34	1.03 (0.95, 1.12)	0.45	1.03
rs1957636	14q22.2	54 560 018	C	T	0.60	0.93	0.26	0.97 (0.89, 1.06)	0.54	0.97
rs16969681[c]	15q13.3	32 993 111	T	C	0.10	1.18	0.13	1.16 (1.04, 1.31)	0.010	1.16
rs4779584	15q13.3	32 994 756	C	T	0.81	0.74	0.46	0.98 (0.90, 1.06)	0.62	0.99
rs11632715	15q13.3	33 004 247	A	G	0.47	1.12	0.38	1.04 (0.96, 1.13)	0.34	1.04
rs9929218	16q22.1	68 820 946	A	G	0.29	0.92	0.29	0.93 (0.85, 1.02)	0.12	0.94
rs4939827	18q21.1	46 453 463	T	C	0.53	1.18	0.32	1.07 (0.99, 1.17)	0.096	1.08
rs10411210	19q13.1	33 532 300	T	C	0.10	0.87	0.41	0.94 (0.86, 1.02)	0.11	0.97
rs961253	20p12.3	6 404 281	A	C	0.36	1.12	0.36	1.08 (1.00, 1.18)	0.054	1.09
rs4813802	20p12.3	6 699 595	G	T	0.36	1.09	0.14	1.10 (0.98, 1.23)	0.11	1.07
rs2423279[d]	20p12.3_Asi	7 812 350	C	T	0.30	1.14	0.34	0.94 (0.86, 1.03)	0.19	0.95
rs4925386	20q13.33	60 921 044	C	T	0.69	1.08	0.28	1.05 (0.96, 1.15)	0.27	1.04
rs5934683	23p22.2	9 751 474	C	T	0.62	0.93	0.28	1.00 (0.90, 1.11)	0.93	1.03

[a]Frequency (FRQ) and odds ratio (OR) for the A1 allele.
[b]OR further adjusted for local ancestry.
[c]Imputed with $R^2 \geq 0.90$.
[d]Published associations were in East Asians. Other published results were in Europeans.
[e]Association previously reported only with distal (left) colon cancer. Results in our study were also for distal colon cancer. No association was seen when all cancer cases were pooled (OR = 1.08, $P = 0.23$).

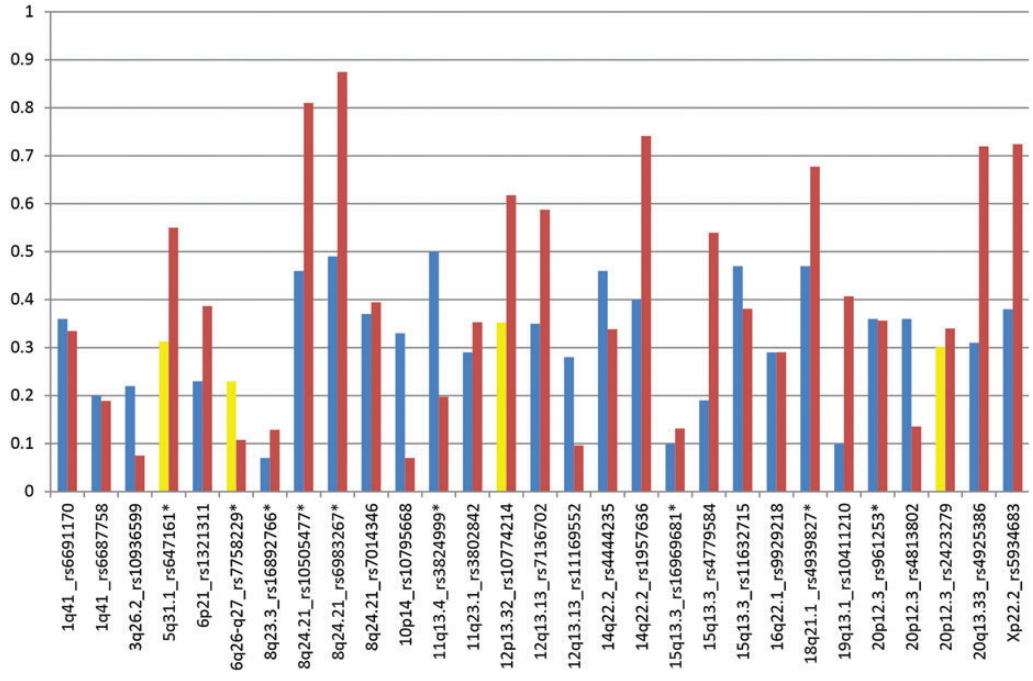

Figure 1. Allele frequencies in published reports (blue for Europeans and yellow for East Asians) and in African Americans (red). SNPs replicated in African Americans are marked with an asterisk.

Given the strong prior evidence for the directions of these genetic effects, we considered as statistically significant replication if the odds ratio (OR) estimates in African Americans were of the same direction as reported in the original GWAS with a one-sided P-value < 0.05 (or equivalently with a two-sided P-value < 0.10). Nine variants were replicated at this level. These included rs7758229 on 6q26-27 that was previously reported to be associated only with distal (left) colon cancer in East Asians [C/A, OR = 1.28 in (2)]—in our data, the OR of the risk allele C was 1.23 (95% CI: 1.00–1.51, two-sided $P = 0.056$ with 500 left colon cases and all controls). These 9 and additional 15 variants showed the same direction of allelic risk as reported before (24 of 29, 83%; Table 1). Estimated locus-specific ancestry (see Methods) (including 250 kb on both sides of the index variants) was not associated with CRC after adjustment for age, gender and the first four PCs—only two strong associations were observed for 19q13.1 ($P = 0.003$) and Xp22.2 ($P = 0.04$). Results for the index SNPs were similar with or without adjustment for local ancestry (Table 1).

Significant heterogeneity of effects across anatomic subsites (left colon, right colon and rectal cancer) was detected for rs10411210 at 19q13 and rs1321311 at 6p21 ($P_{\text{heterogeneity}} = 0.03$ and 0.04, respectively). Stratified analyses by site showed that the rs10411210T allele was only associated with left colon cancer (OR = 0.83, 95% CI = 0.72–0.95, $P = 0.007$) and the rs1321311 A allele was only associated with rectal cancer (OR = 0.85, 95% CI = 0.73–0.99, $P = 0.03$). These site-specific effects were of the same direction as reported in the initial GWAS.

Even with the large sample size, our power to detect the previously reported effect sizes was >80% for only 15 of the 29 index variants at a one-sided α level 0.05 (Supplementary Material, Table S2); among these, 7 were successfully replicated. One possible reason of non-replication could be that the biologically relevant variants are in LD (correlated) with the index variants in the original GWAS population, yet not so in African Americans. Therefore, a complete search among the correlated variants for stronger associations with CRC is necessary. We performed fine-mapping in the 21 risk regions/loci ($\pm$ 250 kb of index variants) using genotyped and imputed data (see Methods). We catalogued SNPs that were correlated ($r^2 > 0.2$) with the index signals from 1KGP [Europeans (EUR) or Asians (ASN) depending upon the initial GWAS population] and tested their associations with CRC using locus-specific α levels, calculated as 0.05 divided by the 'effective' number of independent markers in the 1KGP Africans (AFR) population among the correlated markers in each region (see Methods and Table S3 for locus-specific α levels). This strategy was to balance between the need to correct for multiple comparisons and the prior knowledge that these regions are known to be important for CRC risk.

Among the correlated SNPs, two variants in the 1q41 and 19q13.1 loci passed the significance threshold (Table 2). Specifically, at 1q41, the original risk variants rs6691170 and rs6687758 were not replicated in African Americans (OR = 1.02 and 0.99, $P > 0.7$, power $\leq$ 56%); rs12759486 (T/C) located between the two was much more strongly associated with CRC (OR = 0.86, 95% CI = 0.79–0.93, $P = 1.5 \times 10^{-4}$) in our study (Fig. 2). Rs6691170 and rs6687758 were in low LD with each other ($r^2 = 0.13$, $D' = 0.59$) in EUR and the original GWAS in Europeans (4) suggested that they may represent independent signals of association. From LD structure based on 1KGP data, rs6691170 and rs6687758 are in separate (but not completely distinct) LD blocks in both EUR and AFR (Supplementary Material, Fig. S2). Rs12759486 is located between the two and is in moderate LD with both of them in EUR ($r^2 = 0.37$ and 0.22, respectively) but not so in AFR

Table 2. Associations with CRC at the two regions where a statistically significant better SNP was discovered in African Americans

Region	Name (BP)	A1/A2/A1 Frequency	OR (95% CI)[a]	*P*-value	Index SNP from initial GWAS (BP)	r^2 with index SNP in EUR/AFR
1q41	rs12759486 (222 066 536)	T/C/0.58	0.86 (0.79, 0.93) 0.85 (0.79, 0.92)[c]	1.5×10^{-4} 9.4×10^{-5}	rs6691170 (222 045 446)[b]	0.37/0.004
19q13.1	rs7252505 (33 575 064)	A/G/0.62	0.85 (0.78, 0.93) 0.84 (0.76, 0.93)[c]	1.8×10^{-4} 6.0×10^{-4}	rs10411210 (33 532 300)	0.77/0.05

Analysis was adjusted for age, sex and the first four PCs. Both SNPs were imputed with $R^2 > 0.97$. EUR and AFR are European and African samples from the 1000 Genomes Project.
[a]ORs are for A1 allele.
[b]Selected from the two index SNPs in this region based on a smaller *P*-value.
[c]Analysis conditioning on the index SNPs from the initial GWAS study.

($r^2 < 0.004$, $D' < 0.01$). Rs12759486 maps within an intron of the lincRNA RP11-815M8.1. Although the function of this RNA is unknown, it is expressed in colon tissue (Human BodyMap 2.0 data from Illumina). Interestingly, a nearby SNP, rs7547751, 827 bp upstream and in perfect LD with rs12759486 in both EUR and AFR ($r^2 = 1.0$), was equally associated with CRC ($P = 1.7 \times 10^{-4}$, OR = 0.85 for major allele C). Rs7547751 maps to H3K4me1 and H3K27Ac enhancer peaks with a nearby DNAse I peak and CTCF and cohesion/Rad21 binding sites determined by ChIP-seq (15). Based on these data, it is possible that this enhancer may modulate lincRNA RP11-815M8.1 or a protein-coding gene *in trans*.

At 19q13.1 in African Americans, the published hit, rs10411210, had a *P*-value of 0.11 using all cases combined (OR = 0.94, 95% CI = 0.86–1.02, power = 0.97). Rs7252505, 43 kb downstream from rs10411210, was more strongly associated with CRC (OR = 0.85, 95% CI = 0.78–0.93, $P = 1.8 \times 10^{-4}$) in African Americans (Fig. 3). Rs7252505 and rs10411210 were correlated in EUR ($r^2 = 0.77$, $D' = 0.88$) but not so in AFR in 1KGP ($r^2 = 0.05$, $D' = 0.44$) (see Supplementary Material, Fig. S3 for LD structure in 1KGP).

These results raise the possibility that the original GWAS signals in 1q41 and 19q13 may not have captured the best predictive risk markers. Results for the two best markers in African Americans did not change much (<3% change in ORs) after conditioning on the index variants or after adjusting for local ancestry.

We also looked for new independent associations among variants that were uncorrelated ($r^2 < 0.2$) with the index SNPs in the initial GWAS populations, using a significance level of 2.8×10^{-6}, which is 0.05/the total number of 'effective' markers across the 21 risk regions in the 1KGP YRI population (Methods and Table S3). We did not find any statistically significant associations among SNPs in this category.

To model the cumulative effects of CRC risk variants, we constructed a risk score by summing all independent risk alleles (OR >1) and estimated the average per allele risk (see Methods). We compared the results from summing all GWAS-identified risk variants ($n = 21$ independent) and from summing the variants that were replicated ($n = 7$) or newly discovered ($n = 2$) in African Americans (Methods and Table 3), excluding the one SNP on 6q26-q27 that was only associated with distal colon cancer in the original GWAS (2). From 21 GWAS index variants, the per allele OR was 1.07 (95% CI: 1.05–1.09, $P = 3.4 \times 10^{-10}$); the OR for the highest versus the lowest quartile of the risk score was 1.72 (95% CI: 1.44–2.06, $P = 1.9 \times 10^{-9}$). As expected, there was a slight increase in relative risk when summing risk variants replicated or newly discovered in African Americans: the per allele OR was 1.14 (95% CI: 1.10–1.18, $P = 6.5 \times 10^{-16}$) and the OR for the highest versus lowest quartile of the score was 1.83 (95% CI: 1.55–2.16, $P = 1.1 \times 10^{-12}$). The risk score effect did not differ across left colon, right colon or rectal cancer ($P_{heterogeneity} = 0.54$).

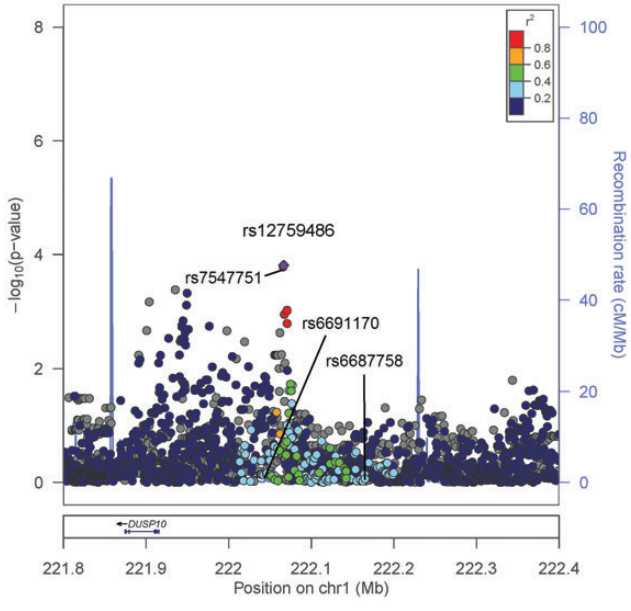

Figure 2. *P*-value plot for the 1q41 risk locus. The SNP with the smallest *P*-value in African Americans, rs12759486, is shown as a purple diamond. r^2 is in relation to this SNP in EUR from the 1000 Genomes Project. Grey circles are SNPs with no r^2 estimation due to low MAF or because the SNP is not in older versions of the 1000 Genomes data. The plot was generated using LocusZoom (16).

The two novel variants rs12759486 and rs7252505 contributed approximately 1.5% of the heritability of CRC (see Methods), estimated in African Americans.

DISCUSSION

It is important to assess the strength and robustness of GWAS-identified association signals in diverse ethnic

populations. This study, to our knowledge, is the first to replicate and characterize risk variants identified from CRC GWAS (all in European and Asian populations) in African Americans with a relatively large sample size. Nine associations were statistically significantly replicated with the same direction of effects as previously reported (one-sided $P < 0.05$). We also identified variants in 1q41 (rs12759486 or rs7547751, two SNPs in perfect LD) and 19q13.1 (rs7252505) that were more strongly related to CRC in African Americans than the index variants (three orders of magnitude change in *P*-value). These two new risk variants explained roughly 1.5% of overall CRC heritability in African Americans and improved on cumulative risk modeling. Failure to replicate other previous associations may be due to low study power (<80% for half of the variants) or to differences in LD patterns between the causal variant and the identified marker between populations—the latter is a known cause for directionally inconsistent associations across studies (a.k.a. the flip-flop phenomenon) (17). No statistically significant reverse association was noted, however.

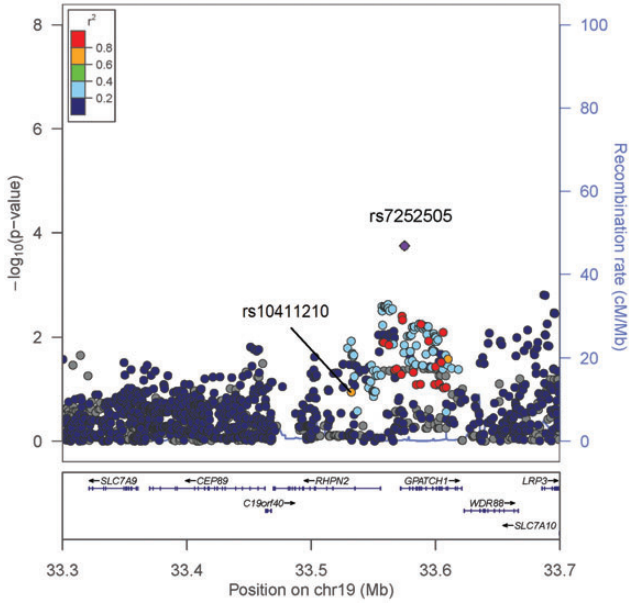

Figure 3. *P*-value plot for the 19q13 risk locus. The SNP with the smallest *P*-value in African Americans, rs7252505, is shown as a purple diamond. r^2 is in relation to this SNP in EUR from the 1000 Genomes Project. Grey circles are SNPs with no r^2 estimation due to low MAF or because the SNP is not in older versions of the 1000 Genomes data. The plot was generated using LocusZoom.

We detected significant heterogeneity of effects across left colon, right colon and rectal cancer for rs10411210 in *RHPN2* at 19q13 and rs1321311 near *CDKN1A* at 6p21 ($P_{\text{heterogeneity}} < 0.05$). Although they were not statistically significantly replicated when all cases were pooled, rs10411210 was associated with left colon cancer and rs1321311 with rectal cancer (*P*'s < 0.05) in African Americans. In the GWAS where rs10411210 was first reported for CRC association, there was no mention of effect heterogeneity across subsites (although the between-study heterogeneity seemed high) (5). Several replication studies conducted in Northern Chinese (18), Southern Chinese (19), Hong Kong subjects (20) and in a Swedish-based cohort (21) have been published for rs10411210; however, none of these replicated the association or reported on differences in effect across subsites. A reason for failure to replicate may be due to effect heterogeneity. In a study (~1000 cases) on clinical and morphological characters of CRC tumors and genetic variation, the TT genotype of rs10411210 was associated with desmoplastic reaction, which is generally considered favorable for survival (22). For rs1321311, we did not find studies including specific results other than the original GWAS, which only reported that this SNP was not associated with tumor location (colon and rectum). Risk heterogeneity by CRC subsites has been previously reported for GWAS index variants. For example, the risk for rs4939827 was greater for rectal cancer than for colon cancer in a GWAS (statistically non-significant heterogeneity) (7); yet, in a replication study in Europeans, this SNP was associated with distal colon cancer but not proximal or rectal cancer ($P_{\text{heterogeneity}} = 0.03$) (23). In our data, there did not seem to be effect heterogeneity for rs4939827 by subsites ($P = 0.23$). However, for the SNP rs7758229 at 6q26 that was reported to be only related to left colon cancer in a GWAS in East Asians (2), no heterogeneity across sites was found in our data ($P_{\text{heterogeneity}} = 0.35$), even though we replicated its association with left colon cancer (one-sided $P < 0.05$) and not with right-sided colon or rectal cancers (*P*'s > 0.41). Sample size could be a reason for these inconsistencies and additional large replication studies are needed.

At 19q13 (index SNP rs10411210), rs7252505 was more strongly associated with CRC in African Americans, with no evidence of heterogeneity across anatomical subsites ($P_{\text{heterogeneity}} = 0.78$). While rs10411210 is located in *RHPN2*, rs7252505 is in an intron of the neighboring gene *GPATCH1*. Although *GPATCH*1 is expressed in the colon, little is known about its function other than the fact that it contains a G-patch domain, a domain typically associated with RNA processing. Rs7252505 or SNPs in high LD ($r^2 > 0.8$) with it in 1KGP AFR do not map to any ChIP-seq peaks marking

Table 3. Effects of summary risk score on CRC risk in African Americans

Quartiles	Using index GWAS SNPs ($n = 21$)			Using risk SNPs in African Americans ($n = 9$)		
	Ca/Co[a]	OR (95% CI)	*P*-value	Ca/Co[a]	OR (95% CI)	*P*-value
Q1	271/862	1.00 (ref.)	—	380/1270	1.00 (ref.)	—
Q2	428/1133	1.29 (1.07, 1.56)	0.008	469/1180	1.63 (1.38, 1.93)	0.0001
Q3	516/1298	1.36 (1.13, 1.63)	0.001	504/1147	1.40 (1.18, 1.65)	6.8×10^{-9}
Q4	679/1410	1.72 (1.44, 2.06)	1.9×10^{-9}	541/1106	1.83 (1.55, 2.16)	1.1×10^{-12}

[a]Ca/Co, numbers of cases/controls.

enhancers or other regulatory elements in colon cancer cell lines. The functional basis for the rs7252505 signal is, thus, not apparent at this stage. Only one CRC study in Spanish subjects reported on rs7252505; however, the association with CRC was non-significant ($P = 0.36$) and no allelic risk estimate was presented (24).

Our approach of separating SNPs that are in LD from those not in LD with the index SNPs and testing the two groups with different significance thresholds was a reasonable balance between prior knowledge and the need for multiple comparison adjustment. However, the two groups may not have been correctly defined even with the detailed coverage featured by currently available public data. For example, we replicated the index SNP rs16892766 at 8q23 with a *P*-value 0.006 (minor/major = C/A, MAF = 0.13, OR = 1.17). Another SNP, rs16892769, with a much smaller *P*-value (minor/major = G/T, MAF = 0.12, OR = 1.31, $P = 1.5 \times 10^{-5}$, imputation $R^2 = 0.91$) is located just 131 bp from rs16892766 (see Supplementary Material, Fig. S4 for *P*-value plot and Supplementary Material, Fig. S5 for LD structure in 1KGP EUR and AFR). After conditioning on each other, the *P*-values became even smaller (10^{-6} for rs16892769 and 0.0004 for rs16892766) and did not change much with additional adjustment for local ancestry. Generally, SNPs within such short distance are likely inherited together because physical chromosomal structure typically prevents recombination events. In 1KGP AFR, the two SNPs are indeed in high LD ($D' = 1$, $r^2 = 0.05$, $P <$ 0.0001 from Pearson's correlation test). However, because rs16892769 was monomorphic in Europeans in both HapMap and 1KGP, no LD estimation was available; hence, rs16892769 was defined as 'uncorrelated' with the index variant and did not pass the corresponding more stringent significance threshold. Given its rarity or non-existence in Europeans, rs16892769 is unlikely a causal variant in this population. In African Americans, haplotype analysis revealed three common haplotypes AT, AG and CT, with frequency 0.75, 0.12 and 0.13, respectively; compared to AT, the haplotype AG had an OR of 1.30 ($P = 3.2 \times 10^{-5}$), suggesting some independent effect for the G allele of rs16892769. There has been no report on the association between rs16892769 and CRC. However, this SNP may be worth investigating in future studies in light of our results. On the other hand, the threshold ($r^2 > 0.2$) for defining markers correlated with the index variants may be too low. If a higher threshold is used (such as $r^2 > 0.4$), our best marker at 1q41, rs12759486, would be classified as 'non-correlated' and would not pass the more stringent significance level for this category. Hence, some caution should be exercised in interpreting our results. On a related note, the effect sizes of our two best SNPs (rs12759486 and rs7252505) may be over-estimated due to the 'winner's curse' (25–27).

In summary, we confirmed in African Americans the risk directions for 83% of the 29 CRC variants identified in previous GWAS and successfully replicated an association for nine of them. No population-specific (except for the possibility of an association with rs16892769) or novel risk loci were discovered. However, at two known loci, we identified two markers that may better define CRC risk than the previous index SNPs. Replication in additional studies is required to confirm these effects.

MATERIALS AND METHODS

Subjects and genotyping and QC

DNA samples were available for 7339 African Americans (2066 CRC cases, 5273 CRC-free controls) from the following studies/centers: the Multiethnic Cohort study (MEC, 442 cases and 4620 controls), Colorectal Cancer Family Registry (CCFR, 999 cases and 290 controls), the Southern Community Cohort Study (SCCS, 164 and 160 controls), MD Anderson Cancer Center (189 cases), the University of North Carolina CanCORS study (UNC-CanCORS, 84 African American cases), the North Carolina Rectal Cancer Study (UNC-Rectal, 112 cases and 108 controls), the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO, 76 cases and 95 controls). See Supplementary Material for details on sample recruitment and data collection.

Except for the PLCO subjects (see below), 7168 samples were genotyped using the Illumina 1M-duo bead array at the USC Epigenome Center. Samples were excluded in the following situations: (i) call rates < 95% ($n = 167$), (ii) no age information ($n = 2$), (iii) gender mismatch ($n = 39$), i.e. when the reported gender is different from that estimated based on X chromosome inbreeding coefficient *F* (calculated by PLINK), (iv) ancestry outliers ($n = 98$) based on PCs (discussed below), and (v) closer than second-degree relatives ($n = 435$), where relationships were derived from estimated probabilities of sharing 0, 1 or 2 allele based on genomic data (calculated by PLINK). Relatives were removed in the following order or priority: (i) subjects with many relatives, (ii) controls, (iii) samples with lower call rates. Sample replicates (2%) were included and the average concordance rates were >99.9%. Starting from 1 192 666 markers, we excluded markers of poor clustering property or with call rates <95%, MAFs <0.005, more than one discordant pair among sample replicates, and *P*-values $< 10^{-10}$ from the Hardy-Weinberg equilibrium test in MEC controls and of poor clustering quality. These resulted in 6427 subjects (4609 controls and1818 cases) on 1 049 327 markers.

PLCO subjects ($n = 171$) were genotyped by the NCI genotyping center on the Illumina Omni 2.5 M array and were pre-filtered by the PLCO data coordinating center using similar criteria as described above. We removed one subject to get rid of closer than second-degree relatives. For this analysis, 76 cases and 94 controls on 527 383 markers that overlapped with other studies were retained. Allele frequencies matched well between PLCO and MEC controls (only 77 differences were >0.1 and all were ≤0.13).

Statistical analysis

Association testing

Logistic regression of allelic dosage with adjustment for age at blood draw, sex and the first four PCs was performed to estimate the OR and 95% CI of per increase in allele count, where age was grouped as <55 years, 5-year intervals from 55 to 80, and ≥80 years. Heterogeneity of genetic effects across left colon, right colon and rectal cancer (mutually exclusive) was assessed with multinomial logistic regression with *SAS* 9.2. PCs were calculated using similar methods as in (28) with our own *R* program, based on about 22 000 SNPs with inter-marker distance >100 kb. Unrelated HapMap CEU, YRI and JPT were

included as population controls to associate PCs with continental ancestries—PC1 was an indicator for European ancestry. Ethnicity outliers were identified on PC plots by visual inspection. The distribution of PCs was similar among all cases and controls after the outliers were removed (Supplementary Material, Fig. S1). The first six PCs sequentially accounted for 2.35, 0.52, 0.10, 0.10, 0.09 and 0.09% of the total variation, which, together with pair-wise PC plots, suggested that the first four PCs were most informative for global ancestry. All *P*-values presented are two-sided, unless otherwise noted.

Local ancestry estimation

The percentage of African ancestry (0, 50 or 100%, i.e. half of the estimated number of African chromosomes) was inferred for each participant at each marker location with the LAMP program v2.4 (29). For the one region on Chromosome X, local ancestry was separately estimated for males and females. To summarize local ancestry at a CRC risk region, for each individual we averaged across all local ancestry estimates that are within the start and end points of the region (Supplementary Material, Table S3). We evaluated whether ORs changed with and without adjustment for local ancestry in logistic regression. Since they did not change materially, we presented results without adjustment for local ancestries, unless otherwise noted.

Imputation

Prediction of un-genotyped SNPs was performed with BEAGLE 3.3 (30) based on the 1000 Genomes EUR and AFR reference panels (phase 1, release 3). Markers with MAFs <0.005 in both EUR and AFR reference panels were excluded from imputation. 10 050 748 markers with imputation accuracy $R^2 > 0.8$ were kept for association analysis.

Fine mapping

We conducted fine mapping in regions within 250 kb of previously identified risk variants. If the regions overlapped for multiple risk variants (i.e. when they are <500 kb apart), one larger region was formed from 250 kb upstream of the first to 250 kb downstream of the last variant (see Supplementary Material, Table S3 for risk regions). For each region, we catalogued SNPs that were correlated ($r^2 > 0.2$) with the index signal in the initial GWAS population (EUR or ASN) in 1KGP (phase 1, release 3). Locus-specific significance levels, 0.05/the 'effective' number of independent markers in the 1KGP AFR population among the correlated markers in each region, were used to declare a better marker than the index SNPs in a region. For variants that were uncorrelated with the index signal in the initial GWAS populations ($r^2 < 0.2$), we estimated the total number of effective markers across all risk regions in 1KGP YRI and used a significance level of 2.8×10^{-6} (0.05/the total number of independent SNPs) in claiming novel association signals. The number of markers correlated and uncorrelated with previous hits and the 'effective' number of SNPs in each region are shown in Supplementary Material, Table S3. The number of independent SNPs was estimated with Keffective (31).

We performed stepwise logistic regression to select among SNPs (dosages) in each region separately for correlated and uncorrelated SNPs, using the significance levels described above. To preserve sample size, sporadic missing values for genotyped markers were replaced with mean allele dosages in their corresponding sex group. One SNP from locus 6q26-27 was previously reported to be associated only with distal (left) colon cancer in East Asians (2) so only association with distal colon cancer was assessed for this variant, unless otherwise noted.

Risk modeling

A risk score was calculated summing the risk (OR > 1) alleles from independent markers. The OR of per one allele increase in the risk score was estimated as an average cumulative risk of all contributing markers. Here, we excluded the one SNP on 6q26-q27 associated with only distal colon cancer in the original GWAS. This sum was calculated first based on previous GWAS hits ($n = 21$). There are six regions containing multiple risk SNPs; these SNPs are in LD ($r^2 > 0.2$) in the initial GWAS population except the two in 20p12.3 (defined in Table 1). Both SNPs in 20p12.3 were included; for the remaining five regions, the SNP with the smallest *P*-value was included since likelihood ratio tests showed that the contribution of other variants in the same region was not important (χ^2's < 0.69, *P*'s > 0.41). Independence of SNPs on the same chromosome was confirmed by comparing regression coefficients from single variant analysis and from the model including all markers on the same chromosome. A sum was similarly calculated with the variants replicated ($n = 7$ independent) and newly identified ($n = 2$) in African Americans. Missing values for ungenotyped markers were replaced with mean allele dosages in their corresponding sex group.

Heritability explained by the newly discovered variants in African Americans

The sibling relative risk attributable to a given SNP was calculated using the formula as in (9,32), $\lambda^* = (p(pr_2 + qr_1)^2 + q(pr_1 + q)^2)/(p^2r_2 + 2pqr_1 + q^2)^2$, where p is the population frequency of the risk allele (OR > 1), $q = 1 - p$, and r_1 and r_2 are the relative risks (estimated as OR) for heterozygous and homozygous genotypes containing the risk allele. For imputed markers, we used the probabilities of heterozygotes and homozygotes in regression models to obtain r_1 and r_2. Assuming a multiplicative effect, the proportion of the familial risk attributable to a SNP was calculated as $\log(\lambda^*)/\log(\lambda_0)$, where λ_0 is the overall familial relative risk estimated from epidemiological studies of CRC, assumed to be 2.2 (32,33).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

We thank Dr David Van Den Berg and Xin Sheng from the USC Genomics Core and Lucy Shen from the University of Hawaii Cancer Center for their technical assistance. CCFR was supported by the National Cancer Institute, National Institutes of Health under RFA # CA-95-011. The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the CFRs, nor does mention of trade names, commercial products or organizations imply endorsement by the US Government or the CCFR.

Conflict of Interest statement. None declared.

FUNDING

This work was supported by National Cancer Institute (NCI) grants (R01CA126895, R01CA126895-S1 and R01CA104132) and California Breast Cancer Research Program grant (15UB-8402). The MEC study is funded by NCI grants (UM1CA164973, R37CA54281, P01CA33619 and R01CA 63464). The CCFR studies included were funded by NCI grants (U01CA074799 to Familial Colorectal Neoplasia Collaborative Group at University of Southern California, U01CA074794 to Seattle Colorectal Cancer Family Registry and U01CA074806 to University of Hawaii Colorectal Cancer Family Registry). The SCCS was funded by NCI grant (R01CA092447). UNC-CanCORS was funded by NCI grant (U01CA93326). The North Carolina Rectal Cancer study was funded by NCI grant (R01CA66635). The MD Anderson data collection was supported in part by the MD Anderson University Cancer Fund, the MD Anderson Cancer Center Duncan Family Institute for Cancer Prevention and Risk Assessment, the Center for Clinical and Translational Sciences of the University of Texas Health Science Center at Houston, NCI Cancer Center Support Grant (CA16672) and NCI grant (K07CA160753).

REFERENCES

1. 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.
2. Cui, R., Okada, Y., Jang, S.G., Ku, J.L., Park, J.G., Kamatani, Y., Hosono, N., Tsunoda, T., Kumar, V., Tanikawa, C. *et al.* (2011) Common variant in 6q26-q27 is associated with distal colon cancer in an Asian population. *Gut*, **60**, 799–805.
3. 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.
4. 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.
5. 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.
6. Jaeger, E., Webb, E., Howarth, K., Carvajal-Carmona, L., Rowan, A., Broderick, P., Walther, A., Spain, S., Pittman, A., Kemp, Z. *et al.* (2008) Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. *Nat. Genet.*, **40**, 26–28.
7. 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.
8. 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.
9. Tomlinson, I.P., Carvajal-Carmona, L.G., Dobbins, S.E., Tenesa, A., Jones, A.M., Howarth, K., Palles, C., Broderick, P., Jaeger, E.E., Farrington, S. *et al.* (2011) Multiple common susceptibility variants near BMP pathway loci GREM1, BMP4, and BMP2 explain part of the missing heritability of colorectal cancer. *PLoS Genet.*, **7**, e1002105.
10. 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.
11. 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.
12. Jia, W.H., Zhang, B., Matsuo, K., Shin, A., Xiang, Y.B., Jee, S.H., Kim, D.H., Ren, Z., Cai, Q., Long, J. *et al.* (2013) Genome-wide association analyses in east Asians identify new susceptibility loci for colorectal cancer. *Nat. Genet.*, **45**, 191–196.
13. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.
14. Zaitlen, N., Pasaniuc, B., Gur, T., Ziv, E. and Halperin, E. (2010) Leveraging genetic variability across populations for the identification of causal variants. *Am. J. Hum. Genet.*, **86**, 23–33.
15. The ENCODE Project Consortium. (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.*, **9**, e1001046.
16. Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R. and Willer, C.J. (2010) Locuszoom: regional visualization of genome-wide association scan results. *Bioinformatics*, **26**, 2336–2337.
17. Lin, P.I., Vance, J.M., Pericak-Vance, M.A. and Martin, E.R. (2007) No gene is an island: the flip-flop phenomenon. *Am. J. Hum. Genet.*, **80**, 531–538.
18. Xiong, F., Wu, C., Bi, X., Yu, D., Huang, L., Xu, J., Zhang, T., Zhai, K., Chang, J., Tan, W. *et al.* (2010) Risk of genome-wide association study-identified genetic variants for colorectal cancer in a Chinese population. *Cancer Epidemiol. Biomarkers Prev.*, **19**, 1855–1861.
19. Li, F., Yang, X., Hu, N., Du, H., Ma, Q. and Li, M. (2012) Single-nucleotide polymorphism associations for colorectal cancer in southern Chinese population. *Chin. J. Cancer Res.*, **24**, 29–35.
20. Ho, J.W., Choi, S.c., Lee, Y.f., Hui, T.C., Cherny, S.S., Garcia-Barcelo, M.M., Carvajal-Carmona, L., Liu, R., To, S.h., Yau, T.k. *et al.* (2011) Replication study of SNP associations for colorectal cancer in Hong Kong Chinese. *Br. J. Cancer*, **104**, 369–375.
21. von Holst, S., Picelli, S., Edler, D., Lenander, C., Dalen, J., Hjern, F., Lundqvist, N., Lindforss, U., Pahlman, L., Smedh, K. *et al.* (2010) Association studies on 11 published colorectal cancer risk loci. *Br. J. Cancer*, **103**, 575–580.
22. Ghazi, S., von Holst, S., Picelli, S., Lindforss, U., Tenesa, A., Farrington, S.M., Campbell, H., Dunlop, M.G., Papadogiannakis, N. and Lindblom, A. (2010) Colorectal cancer susceptibility loci in a population-based study: associations with morphological parameters. *Am. J. Pathol.*, **177**, 2688–2693.
23. Curtin, K., Lin, W.-Y., George, R., Katory, M., Shorto, J., Cannon-Albright, L.A., Bishop, D.T., Cox, A., Camp, N.J. and Colorectal Cancer Study, G. (2009) Meta association of colorectal cancer confirms risk alleles at 8q24 and 18q21. *Cancer Epidemiol. Biomarkers Prev.*, **18**, 616–621.
24. Fernandez-Rozadilla, C., Cazier, J.-B., Tomlinson, I., Carvajal-Carmona, L., Palles, C., Lamas, M., Baiget, M., Lopez-Fernandez, L., 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.
25. Lohmueller, K.E., Pearce, C.L., Pike, M., Lander, E.S. and Hirschhorn, J.N. (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat. Genet.*, **33**, 177–182.
26. Xiao, R. and Boehnke, M. (2009) Quantifying and correcting for the winner's curse in genetic association studies. *Genet. Epidemiol.*, **33**, 453–462.
27. Zöllner, S. and Pritchard, J.K. (2007) Overcoming the winner's curse: estimating penetrance parameters from case-control data. *Am. J. Hum. Genet.*, **80**, 605–615.
28. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.
29. Sankararaman, S., Sridhar, S., Kimmel, G. and Halperin, E. (2008) Estimating local ancestry in admixed populations. *Am. J. Hum. Genet.*, **9**, 290–303.
30. Browning, S.R. and Browning, B.L. (2007) Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. *Am. J. Hum. Genet.*, **81**, 1084–1097.
31. Moskvina, V. and Schmidt, K.M. (2008) On multiple-testing correction in genome-wide association studies. *Genet. Epidemiol.*, **32**, 567–573.
32. Houlston, R.S. and Ford, D. (1996) Genetics of coeliac disease. *Q. J. Med.*, **89**, 737–743.
33. Johns, L.E. and Houlston, R.S. (2001) A systematic review and meta-analysis of familial colorectal cancer risk. *Am. J. Gastroenterol.*, **96**, 2992–3003.