*Carcinogenesis*, 2016, Vol. 37, No. 6, 547–556

**doi:10.1093/carcin/bgw046**
Advance Access publication April 18, 2016
Original Manuscript

ORIGINAL MANUSCRIPT

# Genome-wide association study of colorectal cancer in Hispanics

Stephanie L.Schmit[1,2,3], Fredrick R.Schumacher[1,2], Christopher K.Edlund[1,2], David V.Conti[1,2], Ugonna Ihenacho[1,2], Peggy Wan[1,2], David Van Den Berg[1], Graham Casey[1,2], Barbara K.Fortini[4], Heinz-Josef Lenz[1,2], Teresa Tusié-Luna[5,6], Carlos A.Aguilar-Salinas[5], Hortensia Moreno-Macías[7], Alicia Huerta-Chagoya[5,6], María Luisa Ordóñez-Sánchez[7], Rosario Rodríguez-Guillén[7], Ivette Cruz-Bautista[7], Maribel Rodríguez-Torres[7], Linda Liliana Muñóz-Hernández[7], Olimpia Arellano-Campos[5], Donají Gómez[7], Ulices Alvirde[7], Clicerio González-Villalpando[8,9], María Elena González-Villalpando[9], Loic Le Marchand[10], Christopher A.Haiman[1,2] and Jane C.Figueiredo[1,2,*]

[1]Department of Preventive Medicine, [2]University of Southern California Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, CA 90089, USA, [3]Department of Cancer Epidemiology, Moffitt Cancer Center, Tampa, FL 33612, USA, [4]Department of Biology, Claremont McKenna College, Claremont, CA 91711, USA, [5]Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Sección XVI, Tlalpan, 14000 México City, México, [6]Instituto de Investigaciones Biomédicas, UNAM. Unidad de Biología Molecular y Medicina Genómica, UNAM/INCMNSZ, Coyoacán, 04510 México City, México, [7]Universidad Autónoma Metropolitana, Tlalpan 14387, México City, México, [8]Unidad de Investigación en Diabetes, Instituto Nacional de Salud Pública, México City, México, [9]Centro de Estudios en Diabetes, 01120 México City, México and [10]Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI 96813, USA

*To whom correspondence should be addressed. Jane C. Figueiredo, Harlyne Norris Cancer Research Tower, 1450 Biggy Street 1509J, Los Angeles CA 90089, USA. Tel: 323 442 7752; Fax: 323 442 7954; Email: janefigu@med.usc.edu

## Abstract

Genome-wide association studies (GWAS) have identified 58 susceptibility alleles across 37 regions associated with the risk of colorectal cancer (CRC) with $P < 5 \times 10^{-8}$. Most studies have been conducted in non-Hispanic whites and East Asians; however, the generalizability of these findings and the potential for ethnic-specific risk variation in Hispanic and Latino (HL) individuals have been largely understudied. We describe the first GWAS of common genetic variation contributing to CRC risk in HL (1611 CRC cases and 4330 controls). We also examine known susceptibility alleles and implement imputation-based fine-mapping to identify potential ethnicity-specific association signals in known risk regions. We discovered 17 variants across 4 independent regions that merit further investigation due to suggestive CRC associations ($P < 1 \times 10^{-6}$) at 1p34.3 (rs7528276; Odds Ratio (OR) = 1.86 [95% confidence interval (CI): 1.47–2.36); $P = 2.5 \times 10^{-7}$], 2q23.3 (rs1367374; OR = 1.37 (95% CI: 1.21–1.55); $P = 4.0 \times 10^{-7}$), 14q24.2 (rs143046984; OR = 1.65 (95% CI: 1.36–2.01); $P = 4.1 \times 10^{-7}$) and 16q12.2 [rs142319636; OR = 1.69 (95% CI: 1.37–2.08); $P=7.8 \times 10^{-7}$]. Among the 57 previously published CRC susceptibility alleles with minor allele frequency ≥1%, 76.5% of SNPs had a consistent direction of effect and 19 (33.3%) were nominally statistically significant ($P < 0.05$). Further, rs185423955 and rs60892987 were identified as novel secondary susceptibility variants at 3q26.2 ($P = 5.3 \times 10^{-5}$) and 11q12.2 ($P = 6.8 \times 10^{-5}$), respectively. Our findings demonstrate the importance of fine mapping in HL. These results are informative for variant prioritization in functional studies and future risk prediction modeling in minority populations.

**Received:** December 7, 2015; **Revised:** April 11, 2016; **Accepted:** April 13, 2016

**Abbreviations**

| | |
|---|---|
| AMR | Ad Mixed American |
| CI | confidence interval |
| CRC | colorectal cancer |
| GWAS | genome-wide association study |
| HCCS | Hispanic Colorectal Cancer Study |
| HL | Hispanic and/or Latino |
| Indel | short insertion or deletion |
| LD | linkage disequilibrium |
| MAF | minor allele frequency |
| MEC | multiethnic Cohort |
| OR | odds ratio |
| PC | principal component |
| SIGMA | Slim Initiative in Genomic Medicine for the Americas Type 2 Diabetes Consortium |
| SNP | single nucleotide polymorphism |

## Introduction

Colorectal cancer (CRC) is the third most common cancer and the fourth leading cause of cancer deaths worldwide (1). The Hispanic/Latino (HL) population is the fastest growing ethnic group in the United States, with its size expected to reach 26.5% of the total population by 2050 (2,3). CRC remains the second most common cancer and third most common cause of cancer-related death in the USA among HL (4). Further, disparities in disease presentation and outcomes are evident in this ethnic group. Several studies have observed an increasing trend of early-onset disease (<50 years) and a greater likelihood of late-stage tumors or metastatic disease, especially in the last few decades (5–9).

In addition to well-characterized environmental influences, family history is among the strongest risk factors for CRC, with genetic factors accounting for an estimated 12–35% of the variation in risk of developing the disease among Europeans (10,11). Genome-wide association studies (GWAS) of CRC have been instrumental in identifying common (MAF ≥ 5%) low penetrance susceptibility variants; such efforts have identified 58 susceptibility alleles across 37 regions associated with $P < 5 \times 10^{-8}$ (12–32). To date, the majority of CRC GWAS have been limited to non-Hispanic whites and East Asians, and the generalizability of resultant findings to other ethnic groups where CRC-specific incidence and mortality disparities exist have yet to be comprehensively explored. Specifically, HL individuals have been largely understudied in terms of genetic susceptibility to CRC. In addition, novel CRC-associated variation specific to other populations may exist due to relevant alleles being more common or to different distributions of important environmental factors. Recent examples of ethnic-specific variation are evident in other complex diseases including Latino-specific susceptibility alleles associated with breast cancer and type 2 diabetes (29,33–35). The possibility of ethnic-specific variation has not been widely studied in diverse populations in relation to risk of CRC beyond East Asians, and to a lesser extent, African Americans (26–29,36). To our knowledge, only one small study in Colombians has examined the association of genetic ancestry with colorectal adenomas and adenocarcinomas and described a positive association between African ancestry and CRC (37).

Fine-mapping of genetic association signals can reduce the number of candidate single nucleotide polymorphisms (SNPs) or insertion/deletions (indels) considered for time-consuming and expensive functional follow-up of GWAS-identified risk regions (31,38–40). Fine-mapping studies in multiethnic and admixed populations have been suggested to be more powerful than studies in a single or genetically homogenous ethnic group for localizing functional variants, as shorter linkage disequilibrium (LD) blocks can help to decrease the set of SNPs correlated with a functional allele (41–44). Indeed, the limited set of prior CRC studies focusing on racial/ethnic minorities have proven informative in fine-mapping known risk regions as well as in identifying novel risk loci undetected by GWAS in non-Hispanic white populations. For example, 15 CRC risk SNPs have been discovered in East Asian and African-American populations, two from fine-mapping efforts (26–29,31,36). Individuals of Ad Mixed American (AMR) descent, including HL with diverse backgrounds of European, Native American and African ancestries, present an additional opportunity for fine-mapping because of the group's shorter shared haplotypes around variants of all frequencies as compared to European-only populations (45). In combination, the unique LD structure and allele frequency spectrum of HL populations may assist in localizing association signals (46,47).

The goals of this study were to identify novel variants conferring genetic susceptibility to CRC in the rapidly growing HL ethnic group (48) and to leverage this population's unique LD structure for the fine-mapping of known risk regions identified by GWAS in other ethnic groups.

## Materials and methods

### Study participants

This investigation of genetic contributions to risk of CRC in HL includes cases and controls from three main studies. Epidemiologic and clinical characteristics of the studies are summarized in Supplementary Table 1, available at *Carcinogenesis* Online, and described briefly below.

### Hispanic colorectal cancer study

The Hispanic Colorectal Cancer Study (HCCS) is a population-based study of individuals self-identified as Hispanic with a diagnosis of CRC. Cases are identified from the California Cancer Registry or directly from local hospitals in the Los Angeles region [LAC + USC County Hospital and University of Southern California (USC) Norris Comprehensive Cancer Center]. All men and women over 21 years of age with a first time diagnosis of CRC (ICD-O-3 codes: C18–C21) after January 1, 2008 were eligible for participation. Risk factor/dietary questionnaires, pathology reports and saliva samples (for genotyping) were collected using methodologies developed in the Colon Cancer Family Registry (70) and the Multiethnic Cohort (MEC) (71). The present study includes 950 cases recruited into the HCCS who were born in Mexico (42.3%), the USA (31.4%), Central/South America (16.6%), Cuba or the Caribbean Islands (1.8%) or Europe (0.4%). This study was approved by the University of Southern California Institutional Review Board and the California Committee for the Protection of Human Subjects, and all participants provided written informed consent.

### Multiethnic cohort study

The Multiethnic cohort study (MEC) is a large prospective cohort study that includes subjects from various ethnic groups, including HL primarily from California and mainly, Los Angeles (71). Between 1993 and 1996, participants returned a self-administered baseline questionnaire that obtained general demographic, medical and risk factor information. The MEC used state driver's license files as the primary source to identify study subjects in California. Surnames were used to identify HL individuals because race/ethnicity was not available in driver's license files.

In the cohort, incident cancer cases are identified annually through cohort linkage to population-based Surveillance, Epidemiology and End Results cancer registries in Los Angeles County as well as to the California State cancer registry in the same manner as in the HCCS. All men and women over age 21 with a first time diagnosis of CRC (ICD-O-3 codes: C18–C21) were included as eligible cases. The current study used questionnaire

data and DNA samples derived from whole blood or buccal cells for 661 HL prevalent or incident CRC cases born in the USA (57.8%), Mexico (27.7%), Central/South America (9.7%), Cuba or the Caribbean Islands (4.4%), or Europe (0.2%). Individuals without a diagnosis of CRC were used as controls (n=2,106). All MEC controls self-reported being born in the USA (52.2%), Mexico (34.3%), or Central/South America (13.2%).

This study was approved by the University of Southern California and the University of Hawaii Institutional Review Boards, and all participants provided informed consent.

## Slim initiative in genomic medicine for the Americas

Additional controls for this CRC GWAS consisted of participants from a GWAS of type 2 diabetes conducted by the SIGMA Type 2 Diabetes Consortium. The primary goal of this consortium was to characterize the genetic basis of type 2 diabetes in four component studies: (i) Diabetes in Mexico Study (DMS, $n$ = 472), (ii) Mexico City Diabetes Study (MCDS; $n$ = 614), (iii) Universidad Nacional Autónoma de México (UNAM)/Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán Diabetes Study (UIDS; $n$ = 1138) and (iv) the MEC ($n$ = 2,106; described above) (35). Whole blood-derived DNA samples in the present analysis included SIGMA participants without a diagnosis of diabetes. All participants in DMS, MCDS and UIDS were identified as Mexican.

## Genotyping and imputation

The HCCS and MEC CRC cases were genotyped using the Illumina HumanOmni2.5Exome-8v1.0 and HumanOmni2.5Exome-8v1.1 BeadChip arrays in the USC Norris Comprehensive Cancer Center Molecular Genomics Core (Los Angeles, CA, USA). MEC controls and controls derived from the Mexican SIGMA studies were genotyped using the Illumina HumanOmni2.5-4v1 SNP array at the Broad Institute Genetic Analysis Platform (Cambridge, MA, USA). Samples not passing SIGMA's standard QC procedures for raw data, as described previously, were removed prior to downstream QC steps on the combined set of cases and controls (35). Controls from the MEC-SIGMA study ($n$ = 93) were also genotyped on the HumanOmni2.5Exomev1 array ($n$ = 62) and HumanOmni2.5Exome-8v1.1 array at the University of Southern California to allow for cross-platform validation.

Genotype data were cleaned based on QC metrics at the individual subject and SNP levels. In brief, samples with <95% call rate, unintended replicates, sex mismatches between self-reported and genotypic predicted sex, and identity-by-descent with another sample were removed. Monomorphic SNPs, SNPs with <95% call rate and SNPs with mismatching alleles across platforms were eliminated. We also removed: SNPs with low concordance between intentional cross-platform replicates; SNPs not compared due to low call rate; SNPs discordant between platforms (using HapMap samples genotyped by Illumina); SNPs discordant in within-platform duplicates; and SNPs not present in all datasets. Eleven CRC cases from the MEC study were identified since selection for SIGMA participation, so these individuals were genotyped with controls but treated as cases for analytic purposes.

All SNPs overlapping 1000 Genomes Project genotypes were matched to the forward strand. Imputation of genotypes was performed for both autosomal and X chromosome markers, and all samples were imputed together. The target panel was pre-phased using SHAPE-IT v2.r790 (72), and IMPUTE2 v.2.3.2 (73) was used to impute missing genotypes based on the multiethnic panel of reference haplotypes from Phase 3 of the 1000 Genomes Project (October 12, 2014 release) for autosomal markers and from Phase 1 of the 1000 Genomes Project (March 2012 release) for chromosome X markers (45,74). Genetic markers resulting from the imputation were required to pass stringent imputation quality and accuracy filters prior to entering the analysis phase (info ≥ 0.7, certainty ≥ 0.9, concordance ≥ 0.9) between directly measured and imputed genotypes after masking input genotypes (for genotyped markers only).

## Statistical analysis

### *Ancestry analysis*

Percent ancestry from major population subgroups was estimated for each participant using fastSTRUCTURE software with $k$ = 4 and including HapMap3 samples (European = CEU, TSI; Asian = CHB, JPT, CHD; African = LWK, MKK, YRI) (75). Principal components analysis was conducted using EIGENSOFT v6.0.1 on a panel of ancestry informative markers derived from the literature, the Illumina Infinium HumanExome BeadChip and the Affymetrix Axiom® Exome Array (76–79). Principal components analysis was run twice, once on study samples in combination with HapMap3 samples (2254 ancestry informative markers) to identify ethnic outliers, and subsequently, with study samples only (2616 ancestry informative markers) to generate PCs for global ancestry adjustment in association analyses.

### *Discovery*

A genome-wide association analysis with risk of CRC was conducted using 9 875 636 directly genotyped and high-quality imputed SNPs and indels with MAF ≥ 1% in our full study sample. The association between the allelic dosage of each variant, assuming a log-additive genetic model and the risk of CRC was evaluated using PLINK v1.07 (80). Per-allele odds ratios and 95% confidence intervals (CI) were estimated using unconditional logistic regression adjusted for age, sex and the first ten PCs for global ancestry. The $P$ value threshold for statistical significance in the discovery GWAS was set at the traditional genome-wide value of $5 \times 10^{-8}$.

### *Investigation of previously reported susceptibility regions*

*Replication of index variants*: In addition to the search for novel susceptibility alleles, we examined the association between 57 previously reported susceptibility alleles (i.e. index SNPs or index variants) with MAF ≥ 1% in our study and risk of CRC. Again, association testing was conducted using unconditional logistic regression assuming a log-additive genetic model, with adjustment for age, sex and 10 PCs. Criteria for replication included (i) a consistent direction of effect with the previously published risk allele and (ii) a nominally statistically significant $P$ value (< 0.05). Next, we characterized the broader region surrounding each index SNP (± 500 kilobase, kb) to examine generalizability in HL. We summarized the strongest association signals in our HL study along with the $r^2$ values corresponding to their respective index variants in the original GWAS population.

*Identification of secondary susceptibility alleles*: Finally, we characterized independent secondary signals (i.e. novel markers) in each known susceptibility region. To accomplish this, we conducted fine-mapping in ±500 kb windows surrounding each of the 57 index SNPs that had MAF ≥ 1% in our dataset. To screen for regions of interest, we calculated empirical $P$ value thresholds for statistical significance that accounted for the number of correlated SNPs in each region. These region-specific thresholds were based on a Bonferroni correction for the number of markers needed to tag all SNPs with MAF ≥ 5% in high LD ($r^2 \geq 0.8$) with the index based on the 1000 Genomes Phase I AMR population. The region-specific $P$ values are detailed in Supplementary Table 3, available at *Carcinogenesis* Online. For further analysis, we selected regions in which the most highly associated SNP (i) correlated weakly with the index SNP ($r^2 < 0.2$ in the original discovery population) and (ii) exceeded our region-specific $P$ value threshold.

For these regions, we conducted an association analysis with logistic regression conditional on the index SNP's dosage in R version 3.2.2. If a variant in moderate LD ($r^2 \geq 0.2$) with the index demonstrated a more statistically significant association with CRC in our unconditional examination of known susceptibility regions, then we instead conditioned on that variant. A secondary signal was defined as a variant that remained associated with risk of CRC with a $P$-value below the region-specific threshold upon conditional analysis. LocusZoom plots with LD shading based on $r^2$ in 1000 Genomes AMR samples were generated to visualize the unconditional and conditional association results in each 1 Mb region surrounding the index or the lead variant which was in at least moderate LD ($r^2 \geq 0.2$) with the index (81). We also conducted a sensitivity analysis that excluded CRC cases with diabetes.

# Results

## Characteristics of study sample

Demographic and clinical characteristics of the 1611 CRC cases and 4330 controls included in this study are summarized in Table 1. Supplementary Table 1, available at *Carcinogenesis* Online, provides detailed descriptive statistics for participants in each component study. Case and control groups were statistically

**Table 1.** Characteristics of Hispanic/Latino (HL) colorectal cancer cases and controls in a genome-wide association study of 5941 participants

| | | Cases[a] | Controls[b] | *P* |
|---|---|---|---|---|
| | | N = 1611 | N = 4330 | |
| Age [mean (SD)] | | 61.2 (12.3) | 62.4 (10.2) | <0.01 |
| BMI [mean (SD)] | | 29.2 (6.1) | 27.5 (4.2) | <0.01 |
| Sex (%) | Male | 910 (56.5) | 1845 (42.6) | <0.01 |
| | Female | 701 (43.5) | 2485 (57.4) | |
| Place of birth | United States | 680 (42.2) | 1100 (25.4) | <0.01 |
| | Mexico | 585 (36.3) | 2947 (68.1)[c] | |
| | Central or South America[d] | 222 (13.8) | 277 (6.4) | |
| | Europe | 5 (0.3) | 5 (0.1) | |
| | Cuba or Caribbean Islands | 46 (2.9) | 0 (0.0) | |
| Ancestry estimates [mean (SD)][e] | | | | |
| | European | 0.50 (0.24) | 0.39 (0.27) | <0.01 |
| | East Asian | 0.05 (0.10) | 0. 03 (0.07) | <0.01 |
| | African | 0.02 (0.06) | 0.01 (0.02) | <0.01 |
| | Amerindian | 0.43 (0.25) | 0.57 (0.28) | <0.01 |
| Diabetes | No | 1151 (74.2) | 4330 (100.0) | <0.01 |
| | Yes | 400 (25.8) | 0 (0.0) | |
| Family history of CRC (first degree relative) | No[f] | 1274 (79.1) | 1675 (38.7) | <0.01 |
| | Yes | 146 (9.1) | 114 (2.6) | |
| Cancer site | Colon | 987 (61.3) | — | |
| | Rectum | 347 (21.5) | — | |
| | Other | 14 (0.9) | — | |
| Stage at diagnosis[g] | 0 | 7 (0.4) | — | |
| | 1 | 436 (27.1) | — | |
| | 2 | 259 (16.1) | — | |
| | 3 | 347 (21.5) | — | |
| | 4 | 122 (7.6) | — | |

[a]From the Hispanic Colorectal Cancer Study and the Multiethnic Cohort (California).
[b]From the Slim Initiative in Genomic Medicine for the Americas (California and Mexico).
[c]All non-MEC SIGMA participants were assumed to have been born in Mexico.
[d]Argentina, Belize, Bolivia, Brazil, Chile, Colombia, Costa Rica, Ecuador, El Salvador, Guatemala, Honduras, Nicaragua, Panama, Peru or Uruguay.
[e]% European, Asian and African ancestries were estimated using fastSTRUCTURE with HapMap3 European, Asian and African samples ($k = 4$).
[f]2224 controls were missing family history information.
[g]440 cases were missing stage information.

significantly different with respect to age, sex and body mass index. However, the absolute differences for age and body mass index were minimal, and as in standard GWAS practice, we adjusted for age and sex in our models. Differences in estimated European, Asian, African and Amerindian ancestries and place of birth were accounted for by adjustment for the first 10 principal components (PCs) for global ancestry. The differences in diabetes status and family history between cases and controls were driven by inclusion criteria and missing data for the Slim Initiative in Genomic Medicine for the Americas (SIGMA) controls.

When combined with HapMap3 samples, there were no outliers (>5 standard deviations from the mean) on PCs 1–3. Therefore, all samples were retained for subsequent analysis. Examination of PCs 1–10 showed that only PCs 1 and 2 were statistically significantly different between cases and controls. Supplementary Figure 1, available at *Carcinogenesis* Online, shows pairwise plots of the first three PCs for global ancestry in our total study sample and indicates that cases and controls still had overlapping distributions of PCs 1 and 2, supporting our ability to appropriately adjust for these as covariates.

## Discovery

In total, our GWAS scan included 9 875 636 genetic variants with MAF ≥ 1% that passed stringent quality control (QC) procedures, as depicted in the Manhattan plot in Supplementary Figure 2, available at *Carcinogenesis* Online. A genomic control inflation factor ($\lambda$) of 1.09 indicated adequate control for population stratification (Supplementary Figure 2, available at *Carcinogenesis* Online). At the standard genome-wide significance level of $P < 5 \times 10^{-8}$, we did not observe any genetic markers that were statistically significantly associated with risk of CRC. However, 17 variants across 4 regions with highly suggestive CRC associations ($P < 1 \times 10^{-6}$) were identified on chromosomes 1p34.3 [rs7528276; OR = 1.86 (95% CI: 1.47–2.36); $P = 2.5 \times 10^{-7}$], 2q23.3 [rs1367374; OR = 1.37 (95% CI: 1.21–1.55); $P = 4.0 \times 10^{-7}$], 14q24.2 [rs143046984; OR = 1.65 (95% CI: 1.36–2.01); $P = 4.1 \times 10^{-7}$] and 16q12.2 [rs142319636; OR = 1.69 (95% CI: 1.37–2.08); $P = 7.8 \times 10^{-7}$] (Supplementary Table 2, available at *Carcinogenesis* Online).

## Replication

Among the 57 previously published CRC susceptibility alleles with MAF ≥ 1% in our study, 19 (33.3%) were associated with risk of CRC in HL at a nominal level of statistical significance ($P < 0.05$) (Supplementary Table 3, available at *Carcinogenesis* Online). The known susceptibility alleles that replicated most strongly in HL included rs10505477, rs6983267 and rs7014346 at 8q24.21 (rs6983267: $P = 2.8 \times 10^{-5}$), rs3217810 at 12p13.32 ($P = 2.2 \times 10^{-4}$) and rs4939827 at 18q21.1 ($P = 1.2 \times 10^{-4}$). In HL, the 8q24.21 and 12p13.32 association signal regions were led by previously identified ('index') variants (rs6983267 and rs3217810, respectively), but the most strongly associated

SNP at the 18q21.1 locus (rs4939827 ± 500 kb) was rs11874392 ($6.4 \times 10^{-8}$; $r^2_{EUR}$ = 0.93). A comparison of effect sizes from the current study and the initial published report indicated that 76.5% of SNPs had a consistent direction of effect (Figure 1). Only two potential outliers with respect to discordance were identified: rs35509282 on 4q32.2 and rs73208120 on 12q24.22 (Figure 1). With respect to broader generalization of risk regions (index ± 500 kb), the SNPs most statistically significantly associated with risk of CRC in HL are summarized in Supplementary Table 4, available at *Carcinogenesis* Online. In each of three regions where the top marker was not the index (12q24.22, 16q22.1 and 18q21.1), the lead variant was correlated with the index at $r^2 \geq 0.2$ in the original GWAS population, suggesting that a different variant (or variant set) in HL as compared to other populations may better tag the same underlying functional element.

### Identification of secondary susceptibility alleles

Using our fine-mapping strategy outlined in the statistical methods, we identified two risk regions on chromosomes 3q26 and 11q12.2 in which the most strongly associated variant was weakly correlated with its corresponding index SNP in the original discovery population ($r^2 < 0.2$) and in which the association *P* value was smaller than our pre-specified region-specific threshold (Table 2). The unconditional and conditional analysis results for the proposed independent SNPs are summarized in Table 2. At 3q26.2, the most statistically significantly associated SNP from the unconditional analysis was rs116626941 (OR = 1.35 (95% CI: 1.17–1.56); $P = 4.0 \times 10^{-5}$, Figure 2A). However, after conditioning on the region's most strongly associated index-correlated variant, rs56012908, rs116626941 was no longer associated with a *P* value less than the pre-specified region-specific threshold of $1.3 \times 10^{-4}$ (Figure 2D). Nonetheless, a second SNP, rs185423955, was associated with risk of CRC below our region-specific *P* value threshold in both unconditional and conditional analyses (Figure 2D). At 11q12.2, we identified rs60892987 as a novel secondary signal which exceeded our region-specific *P* value threshold after conditioning on the region's most strongly associated variant that was in high LD with the index (OR = 1.32, $P = 6.6 \times 10^{-5}$), rs28456 (LD with the index SNP rs1535: $r^2_{ASN}$ = 0.93) (Figure 2B and E).

We identified a third region at 2q32.3 with suggestive evidence of a secondary signal, but where the most strongly associated marker did not meet the *P*-value threshold for that region upon conditional analysis ($P = 1.1 \times 10^{-4}$, Table 2). Rs7604359 was in low LD with the index SNP (rs11903757) in the original GWAS population ($r^2_{EUR}$ = 0.002). This SNP was statistically significantly associated with risk of CRC in our unconditional analysis ($P = 1.2 \times 10^{-4}$) but was borderline significant with respect to our pre-specified threshold in an analysis conditional on the region's lead index-correlated variant, rs12474044 ($P = 2.6 \times 10^{-4}$). Rs7604359, approximately 250 kb away from the index and downstream of the myosin IB (*MYO1B*) gene, represents a candidate that tags a potential independent signal in this known region (Figure 2C and F). Notably, a sensitivity analysis that excluded CRC cases with diabetes showed no appreciable differences in results for rs185423955, rs60892987 and rs7604359 (data not shown).

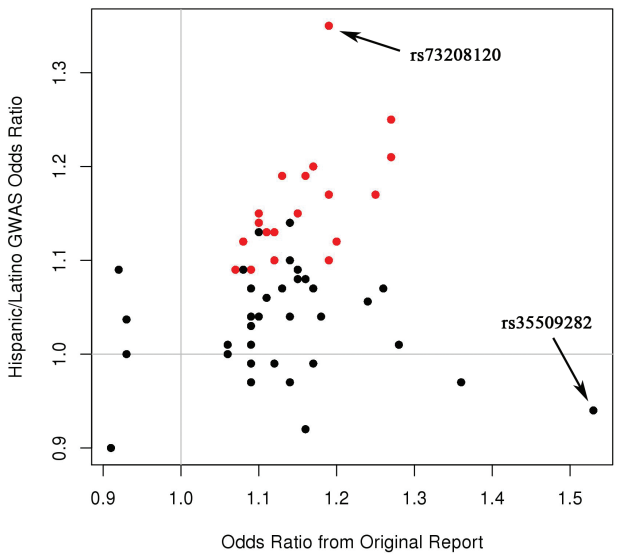


**Figure 1.** Comparison of association effect sizes for previously published CRC risk SNPs (*n* = 57) in the original GWAS population and in Hispanic/Latinos. Red shading denotes $P < 0.05$ in the HL study. GWAS = genome-wide association study. OR = odds ratio. Pearson correlation coefficient = 0.13.

## Discussion

Evaluating the genetic susceptibility to CRC in diverse racial/ethnic groups is important for understanding the generalizability of previous findings, localizing functional variants that underlie known risk regions, and identifying population-specific variation. Our study represents the first large-scale GWAS of CRC in the HL population, an ethnic group experiencing increasing incidence of early-onset and late-stage disease (5–9). Although we identified novel susceptibility regions with only suggestive levels of statistical evidence, our study characterized the generalizability of previous findings from studies mainly in non-Hispanic whites and East Asians to HL. Importantly, our fine-mapping analyses identified two known risk loci (rs185423955/3q26.2 and rs60892987/11q12.2) with a novel secondary association signal within 500 kb of the index SNP(s), which may help guide future risk modeling in HL.

Germline genetic studies of racial/ethnic minorities present a unique opportunity to better understand factors contributing to disparities in CRC incidence and to narrow down the list of candidate variants in known risk regions for functional follow-up and risk modeling. Here, we sought to characterize novel genome-wide genetic variation as well as to better understand the association of known susceptibility SNPs in relation to the risk of developing CRC in HL. Although our study was not powered to identify low-penetrance risk variants at the conventional genome-wide significant level of $5 \times 10^{-8}$, we did identify 17 variants across 4 regions with suggestive CRC associations ($P < 1 \times 10^{-6}$). The variants with the most statistically significant associations with risk of CRC in all four of these regions (rs7528276, rs1367374, rs143046984 and rs142319636) warrant special attention during replication efforts, as examples of HL-specific variation have been demonstrated in relation to risk of CRC, cancers at other organ sites and other complex diseases. For example, two concurrent studies of CRC with gene discovery conducted in Japanese and African American, and East Asian subjects, respectively, identified a novel genome-wide significant risk locus at 10q25.2 (28,36). Other illustrative examples of disease-associated variants that are common in HL but rare in other populations include a breast cancer susceptibility SNP at 6q25 (5′ of the Estrogen Receptor 1 gene) (34) and a type II diabetes risk haplotype spanning *SLC16A11* from SIGMA study of genetic risk factors for type II diabetes (35). Of particular interest

**Table 2.** Known colorectal cancer susceptibility regions harboring a variant that is statistically significantly associated with CRC with $P <$ region-specific threshold in HL ($N = 5941$)

| | | | | | | | | Unconditional | | | | Conditional | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Region | rsID | Chr | Position (hg19) | Eff | Alt | Frq Eff | Info | OR | SE | 95% LCL | 95% UCL | P | OR | SE | 95% LCL | 95% UCL | P |
| 3q26.2 | rs185423955 | 3 | 169950156 | C | T | 0.04 | 0.96 | 1.61 | 0.11 | 1.29 | 2.01 | 3.2E-05 | 1.57 | 0.11 | 1.26 | 1.97 | 7.5E−05 |
| 11q12.2 | rs60892987[a] | 11 | 61982418 | A | G | 0.10 | 1.00 | 1.34 | 0.07 | 1.17 | 1.53 | 2.7E-05 | 1.32 | 0.07 | 1.15 | 1.51 | 6.6E−05 |
| Borderline significant variants | | | | | | | | | | | | | | | | | |
| 2q32.3 | rs7604359[a] | 2 | 192335294 | C | A | 0.05 | 1.00 | 1.43 | 0.09 | 1.19 | 1.71 | 1.2E-04 | 1.40 | 0.09 | 1.17 | 1.68 | 2.6E−04 |

Results are derived from unconditional and conditional logistic regression adjusted for age, sex and 10 PCs for global ancestry. The index or lead variant adjusted for in conditional analyses had $r^2 \geq 0.2$ with the index in the original discovery population.

[a]Directly genotyped.

in the present study is the suggestive risk region on chromosome 1p34.3, which lies in an intron of the microtubule-actin crosslinking factor 1 (*MACF1*) gene. *MACF1* (formerly *ACF7*) knockdown experiments in a mouse embryonic carcinoma cell line resulted in the inhibition of Wnt signaling, a critical signal transduction pathway in the colon, while knockdown in colonic mucosa led to altered mucosal epithelial arrangement and proliferation due to changes in cytoskeleton dynamics (49,50). Interestingly, MACF1 is found in complexes with APC and is regulated by GSK3 in skin and breast carcinoma cells (51,52).

Next, our study investigated the ability to confirm known susceptibility alleles, and more broadly, characterized variants in the surrounding regions in HL. We replicated approximately 33% of the 57 previously identified CRC risk SNPs available for analysis (MAF ≥ 1%) with $P < 0.05$. This is comparable to 31% of the first 29 risk SNPs replicated in a similarly sized study of African-Americans (31). Further, 29 risk regions (±500 kb surrounding the index) included a SNP in at least moderate LD with the index ($r^2 \geq 0.2$ in the original discovery ethnic group) that was associated with risk of CRC at $P < 0.05$. Potential explanations for the lack of replication of some susceptibility alleles in HL include: (i) limited power due to modest sample size and differences in allele frequencies across populations; (ii) differential tagging of the underlying functional/causal variant across racial/ethnic groups with different LD structures and (iii) true biologic heterogeneity, potentially driven by differences in the distribution of important environmental and lifestyle factors across populations.

An additional goal of this study was to fine-map known risk regions in an effort to identify novel secondary signals potentially specific to the HL population. Our analysis revealed two known risk regions (3q26.2 and 11q12.2) with at least one variant weakly correlated with the index that exceeded a region-specific threshold for statistical significance. For each region, this study identified a putative secondary signal following analysis conditional on the index or a more strongly-associated proxy ($r^2 \geq 0.2$ with the index). For the region on 2q32.3, a borderline secondary signal was identified. Further, we observed that known risk loci at 8q24.21 and 18q21.1 harbored SNPs exceeding the respective region's *P* value threshold.

The risk locus at 3q26.2 was among the earliest CRC susceptibility regions identified using a GWAS meta-analysis approach. The index SNP, rs10936599, lies in a coding region of the myoneurin gene (*MYNN*), which encodes a zinc finger protein with largely unknown function (53). However, it also lies upstream of the telomerase RNA component (*TERC*) locus, and the SNP has been associated with longer telomere length (54,55). It has also been associated with increased risk of other cancers including multiple myeloma, bladder cancer, and chronic lymphocytic leukemia (56–58). Our study did not replicate the index SNP but found an additional variant within 100 kb, rs185423955, that was statistically significantly associated with risk of CRC in HL in both unconditional and conditional analysis. This SNP leads a putative secondary association signal. This SNP is an intronic variant in the protein kinase C, iota form (*PRKCI*) gene that encodes a protein implicated in Ras-mediated transformation of colon tissues and CRC progression (59–62). Our findings provide additional evidence in support of this susceptibility locus and highlight the complicated nature of this gene-rich region in relation to risk of CRC.

At 11q12.2, four highly correlated risk SNPs (forming two common haplotypes) have been identified through prior work in East Asians (28). These SNPs were hypothesized to affect the development of CRC as expression quantitative trait loci

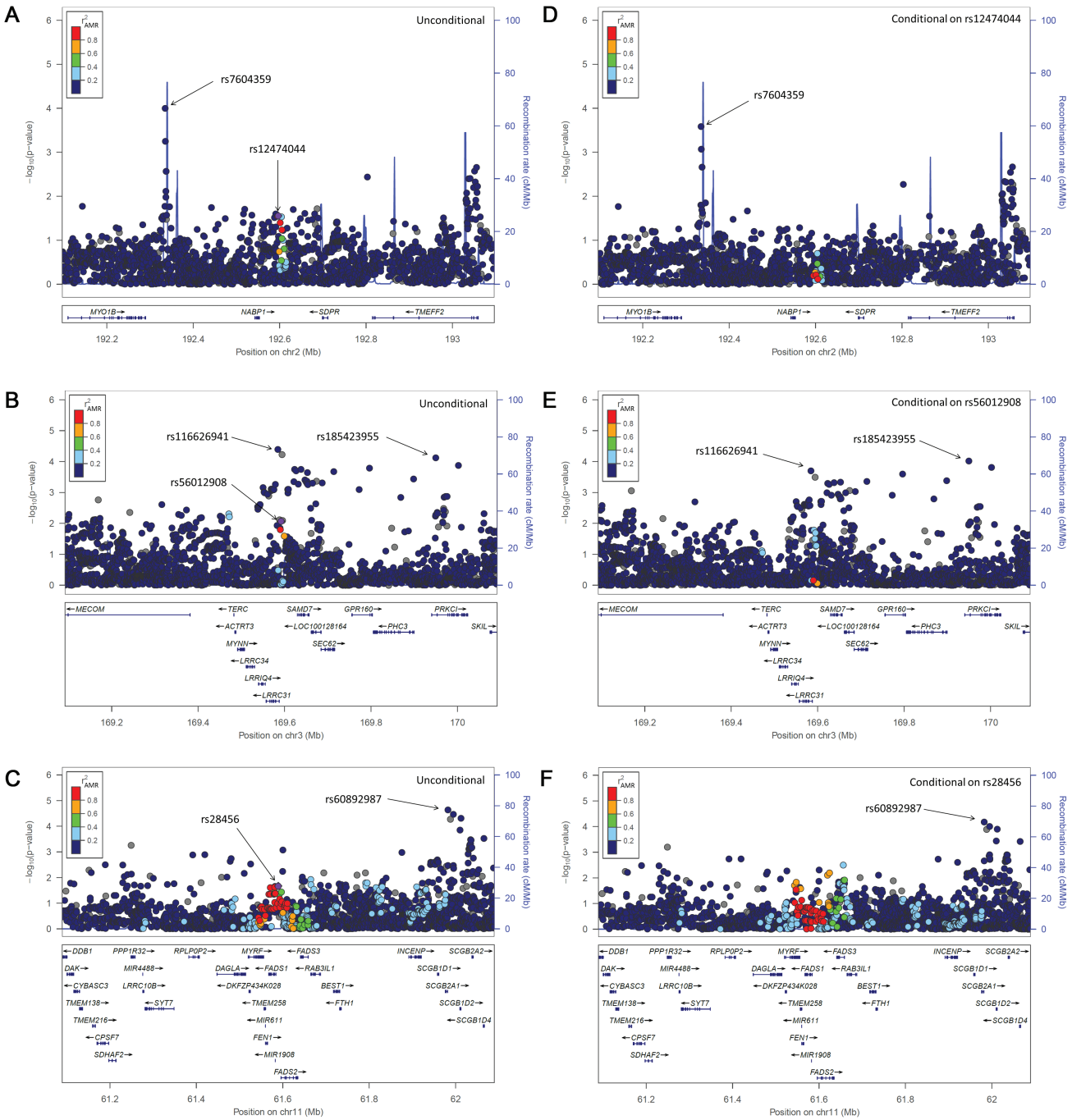


**Figure 2.** LocusZoom regional plots (±500 kb from the index SNP and/or the region's most strongly associated variant in LD ($r^2 \geq 0.2$) with the index) for 2q32.3, 3q26.2 and 11q12.2 based on analyses using best genotype calls. **A–C** represent association results from logistic regression adjusted for age, sex and global ancestry. **D–F** represent association results from logistic regression adjusted for age, sex, global ancestry and allelic dosage of the known region's lead variant. Linkage disequilibrium shading is based on 1000 Genomes Project Phase 3 AMR samples. Diamond-shaped points in purple represent these regions' lead variants.

for *FEN1*, *FADS1* and/or *FADS2* (28,63). Our examination of the surrounding ~1 megabase (Mb) region identified a secondary signal through conditional analysis that is tagged by the intergenic SNP rs60892987. This SNP lies about 384 kb upstream of the most strongly associated index SNP (rs1535) and about 393kb from the region's lead variant correlated with that index (rs28456). Rs60892987 and rs28456 are in low LD with each other ($r^2_{AMR}$ = 0.04; $r^2_{ASN}$ = 0.01), suggesting the independent nature of these signals. However, it is possible that both SNPs are in moderate LD with a shared functional variant(s) yet to be discovered. This region was recently identified and has not yet been a major focus of fine-mapping or functional characterization efforts aside from quantitative trait loci analysis. The potential biological significance of this gene-rich region remains largely unknown, and our identification of a novel secondary association signal supports the complexity of this locus.

An intergenic SNP between the nucleic acid binding protein 1 (*NABP1*) and serum deprivation response

phosphatidylserine-binding protein (*SDPR*) genes on 2q32.3, rs11903757, was first identified as a CRC susceptibility allele in a multiethnic sample of Europeans (discovery and replication) and East Asians (replication) (12). Subsequently, this SNP was replicated in one recent meta-analysis of Europeans and East Asians but not in a second (20,32). Although the most proximal gene is not necessarily the most functionally relevant, it is worth noting that single-stranded DNA binding protein NABP1's potential link to cancer development has been suggested in relation to the maintenance of genomic instability as a component of the sensor of ssDNA (SOSS) complex; also, NAPB1 was found in a proteomic screen of differentially expressed proteins in CRCs (64,65). Here, we did not replicate this index SNP, but we did observe that rs12474044, a SNP in LD with the index that is approximately 10 kb away ($r^2_{AMR}$ = 0.12; $r^2_{EUR}$ = 0.16), was associated with risk of CRC at a nominal level of statistical significance. Further, we identified a suggestive novel secondary signal led by rs7604359, centromeric to the previously published tag SNP. This marker is in low LD with the index SNP ($r^2_{AMR}$ = 0.000; $r^2_{EUR}$ = 0.002), which is not surprising given that it is adjacent to a recombination hotspot. This SNP lies downstream of *MYO1B*, a gene that encodes myosin, a molecular motor protein. In general, this risk region has not been a focus of prior fine-mapping or functional characterization efforts, so additional evidence is needed to replicate this potential independent signal in other populations.

Finally, we examined in detail the well-characterized CRC susceptibility regions at 8q24.21 and 18q21.1, both of which had SNPs that reached our region-specific thresholds despite being in high LD with their index variants. The low penetrance risk locus at 8q24.21 was the first to be identified in association with CRC, and three index SNPs in the region have been described previously (16–19,27,66–68). In this region, our investigation suggested that the index SNP, rs6983267, is also the best marker of CRC risk in HL. The CRC risk region at 18q21.1 was the second genome-wide significant locus to be identified in European ancestry populations (23). The index SNP, rs4939827, lies within intron 4 of *SMAD7*, a gene that encodes a critical regulator of TGF-β signaling. A study in East Asians recently identified another genome-wide significant SNP, rs7229639, which is uncorrelated with rs4939827 in Asians ($r^2_{ASN}$ = 0.02) and weakly correlated in Europeans ($r^2_{EUR}$ = 0.10) (29). We replicate this finding at a nominal level of statistical significance for rs7229639 [OR (95% CI) = 1.12 (1.01, 1.25); $P = 3.9 \times 10^{-2}$] but find that the variant in this region most strongly associated with CRC risk in our HL study is rs11874392 (OR (95% CI) = 1.27 (1.17, 1.39); $P = 6.9 \times 10^{-8}$). This region has been a recent focus of in-depth functional exploration, and it has been proposed that the increased CRC risk is driven by four variants (rs6507874, rs6507875, rs8085824 and rs58920878) that have allele-specific enhancer effects on *SMAD7* expression (69). In our study, all four variants replicated at $P < 0.005$ levels of statistical significance (data not shown).

These results should be interpreted in the context of the study's limitations. Primarily, this investigation with a modest sample size was underpowered for detecting novel low-penetrance susceptibility alleles (MAF < 5%) at the standard genome-wide significant level of statistical significance. However, we did identify potential novel regions and secondary-signals in known regions that merit follow-up in future studies. Thus, the lack of novel risk regions at the $P < 5 \times 10^{-8}$ level should not be interpreted as definitive evidence regarding the absence of population-specific variation influencing CRC risk among HL. With regard to fine-mapping, the incorporation of imperfect genotype imputation can be limiting. However, imputation based on a multiethnic reference panel performed exceptionally well for common variation in this population, as evidenced by high info scores in Supplementary Table 2, available at *Carcinogenesis* Online. Further, conditional analysis is not necessarily the most robust approach for identifying independent signals in a region, and SNPs with the smallest *P* values for association often are not generally the most likely functional candidates. Finally, there are few CRC studies in HL populations with high-throughput genotype data available and therefore, replication of our fine-mapping findings is difficult at this time.

In summary, this study demonstrates the utility of conducting genetic studies in racial/ethnic minorities to better understand the complicated genetic architecture of known risk regions. Future work is needed to replicate and evaluate the biological significance of the identified top variants in known regions and secondary signals, to conduct admixture mapping and to examine the ability of these newly identified SNPs to improve HL-specific risk prediction modeling.

## Supplementary material

Supplementary Tables 1–4 and Figures 1 and 2 can be found at http://carcin.oxfordjournals.org/

## Funding

This work was supported by the National Institutes of Health [R01CA155101 to J.C.F., U01HG004726 to C.A.H., R01CA140561 to D.V.C. and F.R.S., T32ES013678 to S.L.S, U19CA148107 and P30CA014089].

## Acknowledgements

We are indebted to the individuals who participated in this study. Without their assistance, we could not have conducted any of our research. We would like to thank Nathalie Nguyen, Julissa Ramirez, Yaquelin Perez, Daniel Collin, Alicia Rivera, Lauren Gerstmann and the student intern staff for their assistance in logistical support and management, interviewing patients and data entry. Finally, we would like to especially acknowledge Dr. Brian E. Henderson, who passed away before this article was submitted. Without his mentorship and tremendous efforts in co-founding the Multiethnic Cohort, this work would not have been possible.
*Conflict of Interest Statement:* None declared.

## References

1. Ferlay, J. et al. (2015) Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int. J. Cancer, 136, E359–E386.
2. Bureau, U.S.C. (2015) 2014 National Population Projections: Summary Tables. http://www.census.gov/population/projections/data/national/2014/summarytables.html.
3. SR, E. et al. (2011) The Hispanic Population: 2010. 2010 Census Briefs. U.S Department of Commerce, Economics and Statistics Administration, United States Census Bureau.
4. American Cancer Society. (2015) Cancer facts & figures 2015. American Cancer Society, Atlanta, GA.
5. Stefanidis, D. et al. (2006) Colorectal cancer in Hispanics: a population at risk for earlier onset, advanced disease, and decreased survival. Am. J. Clin. Oncol., 29, 123–126.
6. Jafri, N.S. et al. (2013) Incidence and survival of colorectal cancer among Hispanics in the United States: a population-based study. Dig. Dis. Sci., 58, 2052–2060.
7. Soto-Salgado, M. et al. (2009) Incidence and mortality rates for colorectal cancer in Puerto Rico and among Hispanics, non-Hispanic whites, and non-Hispanic blacks in the United States, 1998-2002. Cancer, 115, 3016–3023.

8. Cruz-Correa, M. (2013) Increasing colorectal cancer burden among young US Hispanics: is it time to change current screening guidelines? Dig. Dis. Sci., 58, 1816–1818.
9. Lathroum, L. et al. (2012) Ethnic and sex disparities in colorectal neoplasia among Hispanic patients undergoing screening colonoscopy. Clin. Gastroenterol. Hepatol., 10, 997–1001.
10. Lichtenstein, P. et al. (2000) Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from Sweden, Denmark, and Finland. N. Engl. J. Med., 343, 78–85.
11. Burt, R. (2007) Inheritance of Colorectal Cancer. Drug Discov. Today. Dis. Mech., 4, 293–300.
12. Peters, U. et al. (2013) Identification of genetic susceptibility loci for colorectal tumors in a genome-wide meta-analysis. Gastroenterology, 144, 799–807.
13. Houlston, R.S. et al.; COGENT Consortium; CORGI Consortium; COIN Collaborative Group; COINB Collaborative Group. (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.
14. Dunlop, M.G. et al.; Colorectal Tumour Gene Identification (CORGI) Consortium; Swedish Low-Risk Colorectal Cancer Study Group; COIN Collaborative Group. (2012) Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. Nat. Genet., 44, 770–776.
15. Tomlinson, I.P. et al.; CORGI Consortium; EPICOLON Consortium. (2008) A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. Nat. Genet., 40, 623–630.
16. Tomlinson, I. et al.; CORGI Consortium. (2007) A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nat. Genet., 39, 984–988.
17. Tenesa, A. 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.
18. Zanke, B.W. et al. (2007) Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. Nat. Genet., 39, 989–994.
19. Houlston, R.S. et al. (2008) Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat. Genet., 40, 1426–1435.
20. Whiffin, N. et al. (2014) Identification of susceptibility loci for colorectal cancer in a genome-wide meta-analysis. Hum. Mol. Genet., 23, 4729–4737.
21. Tomlinson, I.P. 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.
22. Jaeger, E. et al. (2008) Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. Nat. Genet., 40, 26–28.
23. Broderick, P. et al. (2007) A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. Nat. Genet., 39, 1315–1317.
24. Peters, U. et al. (2012) Meta-analysis of new genome-wide association studies of colorectal cancer risk. Hum. Genet., 131, 217–234.
25. Schmit, S.L. et al. (2014) A novel colorectal cancer risk locus at 4q32.2 identified from an international genome-wide association study. Carcinogenesis, 35, 2512–2519.
26. Jia, W.H. et al. (2013) Genome-wide association analyses in East Asians identify new susceptibility loci for colorectal cancer. Nat. Genet., 45, 191–196.
27. Cui, R. et al. (2011) Common variant in 6q26-q27 is associated with distal colon cancer in an Asian population. Gut, 60, 799–805.
28. Zhang, B. et al. (2014) Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk. Nat. Genet., 46, 533–542.
29. Zhang, B. et al. (2014) Genome-wide association study identifies a new SMAD7 risk variant associated with colorectal cancer risk in East Asians. Int. J. Cancer, 135, 948–955.
30. Peters, U. et al. (2015) Genetic architecture of colorectal cancer. Gut, 64, 1623–1636.
31. Wang, H. et al. (2013) Fine-mapping of genome-wide association study-identified risk loci for colorectal cancer in African Americans. Hum. Mol. Genet., 22, 5048–5055.
32. Schumacher, F.R. et al. (2015) Genome-wide association study of colorectal cancer identifies six new susceptibility loci. Nat. Commun., 6, 7138.
33. Wang, H. et al. (2014) Trans-ethnic genome-wide association study of colorectal cancer identifies a new susceptibility locus in VTI1A. Nat. Commun., 5, 4613.
34. Fejerman, L. et al. (2014) Genome-wide association study of breast cancer in Latinas identifies novel protective variants on 6q25. Nat. Commun., 5, 5260.
35. Williams, A.L. et al. (2014) Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. Nature, 506, 97-101.
36. Wang, H. et al. (2014) Trans-ethnic genome-wide association study of colorectal cancer identifies a new susceptibility locus in VTI1A. Nat. Commun., 5, 4613.
37. Hernandez-Suarez, G. et al. (2014) Genetic ancestry is associated with colorectal adenomas and adenocarcinomas in Latino populations. Eur. J. Hum. Genet., 22, 1208–1216.
38. Spain, S.L. et al. (2012) Refinement of the associations between risk of colorectal cancer and polymorphisms on chromosomes 1q41 and 12q13.13. Hum. Mol. Genet., 21, 934–946.
39. Carvajal-Carmona, L.G. et al. (2011) Fine-mapping of colorectal cancer susceptibility loci at 8q23.3, 16q22.1 and 19q13.11: refinement of association signals and use of in silico analysis to suggest functional variation and unexpected candidate target genes. Hum. Mol. Genet., 20, 2879–2888.
40. Tomlinson, I.P. 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.
41. Liu, C.T. et al. (2014) Multi-ethnic fine-mapping of 14 central adiposity loci. Hum. Mol. Genet., 23, 4738–4744.
42. Franceschini, N. et al. (2012) Discovery and fine mapping of serum protein loci through transethnic meta-analysis. Am. J. Hum. Genet., 91, 744–753.
43. Wu, Y. et al. (2013) Trans-ethnic fine-mapping of lipid loci identifies population-specific signals and allelic heterogeneity that increases the trait variance explained. PLoS Genet., 9, e1003379.
44. Replication, D.I.G. et al. (2014) Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat. Genet., 46, 234–244.
45. Abecasis, G.R. et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature, 491, 56–65.
46. Ko, A. et al. (2014) Amerindian-specific regions under positive selection harbour new lipid variants in Latinos. Nat. Commun., 5, 3983.
47. Weissglas-Volkov, D. et al. (2013) Genomic study in Mexicans identifies a new locus for triglycerides and refines European lipid loci. J. Med. Genet., 50, 298–308.
48. U.S. Census Bureau. http://www.census.gov.
49. Chen, H.J. et al. (2006) The role of microtubule actin cross-linking factor 1 (MACF1) in the Wnt signaling pathway. Genes Dev., 20, 1933–1945.
50. Liang, Y. et al. (2013) ACF7 regulates colonic permeability. Int. J. Mol. Med., 31, 861–866.
51. Wu, X. et al. (2011) Skin stem cells orchestrate directional migration by regulating microtubule-ACF7 connections through GSK3β. Cell, 144, 341–352.
52. Zaoui, K. et al. (2010) ErbB2 receptor controls microtubule capture by recruiting ACF7 to the plasma membrane of migrating cells. Proc. Natl. Acad. Sci. USA, 107, 18517–18522.
53. Alliel, P.M. et al. (2000) Myoneurin, a novel member of the BTB/POZ-zinc finger family highly expressed in human muscle. Biochem. Biophys. Res. Commun., 273, 385–391.
54. Jones, A.M. et al. (2012) TERC polymorphisms are associated both with susceptibility to colorectal cancer and with longer telomeres. Gut, 61, 248–254.
55. Codd, V. et al. (2013) Identification of seven loci affecting mean telomere length and their association with disease. Nat. Genet., 45, 422–427.
56. Chubb, D. et al. (2013) Common variation at 3q26.2, 6p21.33, 17p11.2 and 22q13.1 influences multiple myeloma risk. Nat. Genet., 45, 1221–1225.
57. Figueroa, J.D. et al. (2014) Genome-wide association study identifies multiple loci associated with bladder cancer risk. Hum. Mol. Genet., 23, 1387–1398.

58. Speedy, H.E. et al. (2014) A genome-wide association study identifies multiple susceptibility loci for chronic lymphocytic leukemia. Nat. Genet., 46, 56–60.
59. Hashimoto, N. et al. (2005) PKClambda regulates glucose-induced insulin secretion through modulation of gene expression in pancreatic beta cells. J. Clin. Invest., 115, 138–145.
60. Atwood, S.X. et al. (2013) GLI activation by atypical protein kinase C iota/lambda regulates the growth of basal cell carcinomas. Nature, 494, 484-8.
61. Murray, N.R. et al. (2004) Protein kinase Ciota is required for Ras transformation and colon carcinogenesis *in vivo*. J. Cell Biol., 164, 797–802.
62. Murray, N.R. et al. (2009) Protein kinase C betaII and PKCiota/lambda: collaborating partners in colon cancer promotion and progression. Cancer Res., 69, 656–662.
63. Liu, L. et al. (2012) Functional FEN1 genetic variants contribute to risk of hepatocellular carcinoma, esophageal cancer, gastric cancer and colorectal cancer. Carcinogenesis, 33, 119–123.
64. Broderick, S. et al. (2010) Eukaryotic single-stranded DNA binding proteins: central factors in genome stability. Subcell. Biochem., 50, 143–163.
65. Lim, S.R. et al. (2011) Analysis of differentially expressed proteins in colorectal cancer using hydroxyapatite column and SDS-PAGE. Appl. Biochem. Biotechnol., 165, 1211–1224.
66. Gruber, S.B. et al. (2007) Genetic variation in 8q24 associated with risk of colorectal cancer. Cancer Biol. Ther., 6, 1143–1147.
67. Haiman, C.A. et al. (2007) A common genetic risk factor for colorectal and prostate cancer. Nat. Genet., 39, 954–956.
68. Hutter, C.M. et al. (2010) Characterization of the association between 8q24 and colon cancer: gene-environment exploration and meta-analysis. BMC Cancer, 10, 670.
69. Fortini, B.K. et al. (2014) Multiple functional risk variants in a SMAD7 enhancer implicate a colorectal cancer risk haplotype. PLoS One, 9, e111914.
70. Newcomb, P.A. et al. (2007) Colon Cancer Family Registry: an international resource for studies of the genetic epidemiology of colon cancer. Cancer Epidemiol Biomarkers Prev., 16, 2331–2343.
71. Kolonel, L.N. et al. (2000) A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. Am. J. Epidemiol., 151, 346–357.
72. Delaneau, O. et al. (2013) Haplotype estimation using sequencing reads. Am. J. Hum. Genet., 93, 687–696.
73. Howie, B.N. et al. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet., 5, e1000529.
74. Abecasis, G.R., et al. (2010) A map of human genome variation from population-scale sequencing. Nature, 467, 1061–1073.
75. Raj, A. et al. (2014) fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics, 197, 573–589.
76. Price, A.L. et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet., 38, 904–909.
77. Liu, J. et al. (2013) Confounding and heterogeneity in genetic association studies with admixed populations. Am. J. Epidemiol., 177, 351–360.
78. Smith, M.W. et al. (2004) A high-density admixture map for disease gene discovery in african americans. Am. J. Hum. Genet., 74, 1001–1013.
79. Seldin, M.F. et al. (2006) European population substructure: clustering of northern and southern populations. PLoS Genet., 2, e143.
80. Purcell, S. et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet., 81, 559–575.
81. Pruim, R.J. et al. (2010) LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics, 26, 2336–2337.