Supplementary Information

Genome-wide association analyses in East Asians identify new susceptibility loci for colorectal cancer

Wei-Hua Jia, Ben Zhang, Keitaro Matsuo, Aesun Shin, Yong-Bing Xiang, Sun Ha Jee, Dong-Hyun Kim, Zefang Ren, Qiuyin Cai, Jirong Long, Jiajun Shi, Wanqing Wen, Gong Yang, Ryan J. Delahanty, GECCO and CCFR, Bu-Tian Ji, Zhi-Zhong Pan, Fumihiko Matsuda, Yu-Tang Gao, Jae Hwan Oh, Yoon-Ok Ahn, Eun Jung Park, Hong-Lan Li, Ji Won Park, Jaeseong Jo, Jin-Young Jeong, Satoyo Hosono, Graham Casey, Ulrike Peters, Xiao-Ou Shu, Yi-Xin Zeng, & Wei Zheng

Corresponding author contact information:

Wei Zheng, M.D., Ph.D.
Vanderbilt Epidemiology Center and Vanderbilt-Ingram Cancer Center
Vanderbilt University School of Medicine
2525 West End Avenue, 8th Floor, Nashville, TN 37203-1738, USA
Phone: (615) 936-0682 Fax: (615) 936-8241
E-mail: wei.zheng@vanderbilt.edu

Contents

Supplementary Note

1. Description of Study Participants

Shanghai-Vanderbilt Colorectal Cancer Genetic Project (SVCRCGP): Colorectal cancer (CRC) cases for the project were derived from the Shanghai Women's Health Study (SWHS) and the Shanghai Men's Health Study (SMHS), both population-based cohort studies which are being conducted in urban Shanghai, China.

The SWHS includes 75,049 Chinese women who were between the ages of 40 and 70 years at enrollment during 1997 to 2000 and lived in urban Shanghai [1]. In-person interviews were conducted to collect exposure information, and anthropometrics were measured. The response rate is 92% for the baseline interview. Approximately 88% of study participants provided biological samples, either a blood sample (n = 56,833) or exfoliated buccal cell sample (n = 8,921).

Using similar study protocols, the SMHS enrolled 61,582 men between the ages of 40 and 74 years in urban Shanghai between 2001 and 2006 with an overall response rate of 74% [2]. Approximately 90% of study participants provided either a blood sample (76%) or a buccal cell sample (14%).

Ongoing follow-up for cancer incidence and cause-specific mortality is conducted in both the SMHS and SWHS via a combination of periodic in-person surveys and annual linkage with data routinely collected by the population-based Shanghai Cancer Registry and Vital Statistics Unit (for death certificates). A total of 777 CRC cases with DNA available were identified in participants of the SWHS and SMHS and included in this study.

A total of 758 cancer-free controls were derived from the SWHS/SMHS and frequency-matched to CRC cases by age and sex. To increase statistical power in Stage 1, we also included 2,131 community controls which were scanned using Affymetrix Genome-Wide Human SNP Array 6.0 (Affymetrix 6.0) as part of an ongoing GWAS of breast cancer [3]. These controls were recruited primarily from the Shanghai Breast Cancer Study (SBCS), a population-based case-control study conducted in urban Shanghai during 1996 through 2004 [4]. Similar to the SWHS/SMHS, all study participants were interviewed to obtain exposure data, and anthropometrics were measured. The vast majority of study participants provided a blood sample or exfoliated buccal cell sample. Because SBCS and SWHS/SMHS participants all reside in the Shanghai urban area, genetic background of the controls from these studies is comparable.

For clarity, we named the 481 cases and 2,632 controls genotyped using Affymetrix 6.0 Shanghai-1. The 296 cases and 257 controls genotyped using Illumina HumanOmniExpress BeadChip (Illumina OmniExpress) were named Shanghai-2.

Guangzhou Study: The Guangzhou study contributed 694 cases and 1,025 controls (Guangzhou-1) to Stage 1. Two additional sets of samples from this study were utilized in Stage 2, including 1,371 cases and 1,521 controls (Guangzhou-2), and 1,355 cases and 1,283 controls (Guangzhou-3). Histopathologically diagnosed CRC cases were recruited from Sun Yat-Sen University Cancer Center between January 2002 and January 2012. All cases and controls were self-reported Han Chinese who lived in Guangdong Province at the time of recruitment. Healthy controls were recruited from physical examination centers of several large general hospitals in Guangdong province communities. At enrollment, controls reported no history of any cancer. Blood samples from all cases and controls were obtained as the source of genomic DNA for the study. Informed consent was obtained.

Guangzhou-1 controls in Stage 1 were genotyped with Illumina Human610-Quad BeadChips (Illumina Inc., San Diego, CA, USA). Methods for genotyping and data preparation have been described previously [5]. We excluded samples with call rate < 96%, heterozygosity more than three standard deviation from the sample mean percent, or confused population stratification. We also excluded SNPs with call rate < 95%, minor allele frequency (MAF) < 3%, or significant deviation from Hardy-Weinberg Equilibrium in controls ($P < 10^{-6}$). A total of 972 controls and 435,925 autosomal SNPs met the quality control criteria and were included in the study.

Aichi Study: This study is part of the Hospital-based Epidemiologic Research Program at Aichi Cancer Center (HERPACC-II), Japan [6]. All first-visit outpatients 20-79 years of age at Aichi Cancer Center during December 2000 to November 2005 were asked to participate in HERPACC-II. Of 29,736 eligible patients approached, 28,766 participated in the study, with a response rate of 96.7%. All participants completed self-administered questionnaires about their lifestyle and demographic characteristics and provided blood samples. Case status was confirmed via the HERPACC-II database and the hospital-based cancer registry database. A total of 589 CRC cases were identified in this cohort and 497 were included in the GWAS. A total of 942 controls without any cancer at recruitment were randomly selected and frequency-matched to cases by age and sex. DNA samples from these controls were scanned using Illumina HumanHap610 BeadChips (Illumina Inc., San Diego, CA, USA) [7]. These subjects served as the control group for CRC cases in Stage 1 (Aichi-1). Stage 2 included 391 cases and 1,116 controls recruited from the same hospital (Aichi-2). Study protocol was approved by the institutional review board at Aichi Cancer Center (Nagoya, Japan).

Korean Cancer Prevention Study-II (KCPS-II): The KCPS-II included 266,258 individuals, 20-77 years of age, who visited 16 health promotion centers nationwide from April 2004 to December 2008 in South Korea. Subjects were interviewed at baseline to obtain exposure data. Cancer diagnoses were identified through 2008 using data from the national cancer registry and hospitalization records. Mortality outcomes were ascertained through 2009 by reviewing death certificates. A computerized search of death-certificate data from the National Statistical Office in Korea was performed using the unique identification number assigned at birth. For the study, we selected 325 CRC patients who provided a blood sample. Cancer-free cohort members (N= 977) were randomly selected as controls.

DNA samples were isolated and genotyped using Affymetrix Genome-Wide Human SNP Array 5.0 (Affymetrix Inc., Santa Clara, CA, USA) at DNA Link Inc. (South Korea). Methods for genotyping and data processing have been described previously [8]. SNPs with call rate < 95%, MAF < 0.01, or significant deviation from HWE ($P < 1.0\times10^{-5}$) were excluded. The final dataset of acceptable markers included 312,869 autosomal SNPs. Samples with call rate < 95 %, or sex-mismatch or confused population structure were removed. A total of 1,301 participants (325 cases, 976 controls) were included in this study.

Seoul Study (Korea-Seoul): This is a multicenter case-control study of CRC conducted in South Korea [9]. Cases consisted of patients with histologically confirmed CRC ages 30-79 years, who were admitted to two university hospitals and one general cancer hospital in the Seoul Metropolitan Area between 1995 and 2004. Controls were selected from the same hospitals during the same period from a wide spectrum of inpatients with non-neoplastic conditions ages 30-79 years. Trained nurse interviewers collected information using a structured questionnaire covering smoking habits, alcohol intake, diet, and other lifestyle factors. Venous blood was collected at the time of interview with written informed consent for genetic studies from subjects

enrolled after 1998. Included in the current study are 849 cases and 673 controls who provided a blood sample. DNA was extracted from the buffy coat by using a commercial kit (Qiagen GmbH). Methods for genotyping and data preparation have been described previously [9].

Korean-NCC Study (Korea-NCC): The Korean National Cancer Center (NCC) Study is a hospital-based case-control study conducted in South Korea. Cases were histologically confirmed patients with CRC who received surgery between 2000 and 2004. A total of 1,392 patients were eligible for this study and provided blood samples to the Tumor Bank of the NCC. Controls (n = 1,329), frequency-matched to cases by age and sex, were selected from participants of the Cancer Screening Cohort of the NCC recruited between August 2002 and December 2004. Genomic DNA was isolated from the buffy coat using a MagAttract DNA Blood Midi M48 Kit (Qiagen, Inc., Valencia, CA, USA) on a Qiagen BioRobot M48 workstation.

2. Genotyping and QC Methods at Vanderbilt and Shanghai

Methods for Stage 1 genotyping and data processing are described in the Study Participants section for controls included in the Guangzhou-1 and Aichi-1 studies, and for cases and controls in KCPS-II. Described here are methods for genotyping and data processing conducted at Fudan University (Shanghai, China) for Guangzhou-3, and at Vanderbilt University (Nashville, TN, USA) for all other samples.

2.1. Stage 1 genotyping and quality control

Genome-wide scanning for 481 CRC cases and 2,632 controls included in Shanghai-1was conducted using Affymetrix Genome-Wide Human SNP Array 6.0 (Affy 6.0). Genotyping for all these samples was performed at the Vanderbilt Microarray Shared Resources (VMSR) following manufacturer's protocol as described previously [1]. Briefly, total genomic DNA (250 ng) at 50 ng/μl is digested with a restriction enzyme (Nsp I or Sty I) and ligated to adaptors which recognize the cohesive 4 bp overhangs. Following ligation, a generic primer which recognizes the adaptor sequence is used to amplify adaptor ligated DNA fragments. The PCR material is purified, quantitated, and analyzed on a 2% agarose gel. Successful PCR products are fragmented, and the fragmented products are again analyzed on a 4% agarose gel. The prepared samples are hybridized to Affy 6.0, then stained and washed, and subsequently scanned. Birdseed v2 algorithm (www.broad.mit.edu/mpg/birdsuite/) was used to call genotypes.

A series of stringent methods was implemented. Before genotyping, quantity of genomic DNA was assessed using both Fluorometer and Nanodrop. All samples were arranged on the 96-well plate at a concentration of 50 ng/μl for a volume of 15 μl. During the genotyping process, we included one negative control (water) and 1 to 4 positive QC samples (NA10851, NA15510, NA18505) in each of the 96-well plates genotyped using Affy 6.0. QC samples were purchased from Coriell Cell Repositories (http://www.coriell.org/). These QC samples were successfully genotyped for 311 times in the assay using Affymetrix 6.0 array, and these data were used to assess the across-batch validation. The average concordance percentage was 99.7% with a median of 100% [3, 10, 11].

Stage 1 genotyping for the 296 cases and 257 controls in Shanghai-2 was performed using Illumina HumanOmniExpress BeadChip (Illumina OmniExpress) (Illumina Inc., San Diego, CA, USA) at the HudsonAlpha Institute for Biotechnology (Huntsville, AL, USA). The same method was used to genotype cases from Aichi-1 (N = 497) and Guangzhou-1 (N = 694) in Stage 1. Samples were processed according to manufacturer protocol. Briefly, 200 ng of DNA at 50ng/μl was amplified in an overnight multistrand displacement amplification assay. Following

amplification, the DNA was enzymatically fragmented, using end-point fragmentation to avoid over-fragmentation. DNA samples were then purified by precipitation using isopropanol. The DNA was resuspended in Illumina buffer RA1. After denaturing the resuspended DNA for 20 minutes at 95°C and incubating for 30 minutes at room temperature, the DNA samples were loaded onto the arrays (12 samples per array) and incubated at 48°C for 16 hours. After hybridization, the arrays were washed to remove unhybridized and nonspecifically hybridized DNA. Labeled nucleotides were added during the XStain step to extend the primers hybridized to the DNA. The extended primers were then stained and coated for protection prior to staining. The coated arrays were then scanned per manufacturer protocol using Illumina iScan. The resulting data were processed using GenomeStudio with genotype calls exported for further analysis.

For the samples genotyped using Illumina OmniExpress, we included one negative control (water) and one positive QC sample (either NA15220 or NA18505) in each of the 96-well plates. These positive QC samples were successfully genotyped for 20 times with an average concordance rate of 99.41% and median value of 99.97%.

Sample exclusion criteria: The sex of study subjects was consistent with the sex referred from X-chromosome genotype data. To detect additional first-degree cryptic relationships in this population, an identity-by-descent (IBD) analysis was performed using PLINK v1.07 (http://pngu.mgh.harvard.edu/purcell/plink/). Of the initial 3,113 samples genotyped using Affy 6.0, 11 samples were excluded due to a call rate < 95%. Of 1,744 samples genotyped using Illumina OmniExpress, 214 were excluded due to a call rate < 95% (N = 211) or sample duplication or contamination (N = 3). After these exclusions, a final data set of 4,632 individuals genotyped at Vanderbilt remained in the association analyses.

Marker exclusion criteria: The following quality control criteria were applied: 1) genotype call rate < 95%; 2) MAF < 5%; 3) concordance rate across genotyping batches < 95% based on QC samples; 4) *P*-value of test for HWE < 1×10^{-5} in controls; and 5) SNPs not in the 22 autosomes. We filtered SNPs for each participating study. After exclusion, the numbers of SNPs remaining for final analysis were 502,145 for 3,102 samples in Shanghai-1, 522,096 for 641 cases in Guangzhou-1, 478,246 for 404 cases in Aichi-1, and 515,701 for 485 samples in Shanghai-2.

2.2. Stage 2 genotyping

We used the Sequenom method for Stage 2 genotyping. Polymerase chain reaction (PCR) and extension primers were designed using MassARRAY Assay Design 4.0 software (Sequenom, Inc., San Diego, CA, USA). PCR and extension reactions were performed according to the manufacturer's instructions, and extension product sizes were determined by mass spectrometry using the Sequenom MassARRAY Analyzer 4.

In Stage 2, we genotyped 64 SNPs for all samples in Guangzhou-2, Korea-NCC, Korea-Seoul, and 799 samples in Japan Aichi-2. We then genotyped 8 SNPs (rs647161, rs2695220, rs4146184, rs10774214, rs1432881, rs1580743, rs1665650, and rs2423279) for 708 samples in Japan Aichi-2 and all samples in Guangzhou-3. All genotyping experiments were performed in the Vanderbilt Molecular Epidemiology Laboratory except those conducted for some samples from the Guangzhou study, which were analyzed in Fudan University (Shanghai, China) using the protocol developed and validated at the Vanderbilt Molecular Epidemiology Laboratory.

We included four negative controls (water) and eight positive QC samples in each 384-well assay plate analyzed at Vanderbilt. We included six negative controls (water) and six blinded

duplicate samples in each 384-well plate analyzed at Fudan University (Shanghai, China). The average concordance rate of data generated from these QC samples was greater than 99% with median value 100% for each of the five studies included in the replication stage. SNPs were excluded in analysis if: 1) call rate < 95%; 2) genotyping concordance rate < 95% in QC samples; 3) unclear genotyping cluster; or 4) *P*-value for Hardy-Weinberg equilibrium < 0.00078 (0.05/64). The numbers of SNPs remaining for analysis in Stage 2 was 61, 61, 59, 8, and 8 for Guangzhou-2, Korea-NCC, Korea-Seoul, Aichi-2, and Guangzhou-3, respectively.

3. Statistical Analyses

Imputation: We used the program MACH 1.0 [12] to impute genotypes for autosomal SNPs which were present in HapMap Phase II release 22 separately for each of the five studies included in Stage 1 (Shanghai-1, Shanghai-2, Guangzhou-1, Aichi-1, and KCPS-II). Genotype data from the 90 Asian HapMap subjects were used as reference. For Guangzhou-1 and Aichi-1, cases and controls were genotyped using different platforms. To improve imputation quality [13], we identified SNPs shared between cases and controls (250,612 SNPs in Guangzhou-1 and 232,426 SNPs in Aichi-1) and used them to impute genotyping data. A total of 1,636,380 imputed SNPs with high imputation quality in all the five studies were tested for association with CRC**.**

Evaluation of population structure: Principal component analysis was performed for each of the five studies using the software EIGENSTRAT, version 2.5 [14] to evaluate the population structure and identify potential genetic outliers. We selected a set of ~6,000 SNPs with 1) a neighboring distance >200kb, 2) MAF>0.2, 3) r^2<0.1, and 4) call rate>99%), which were shared between clean samples in Stage 1 and the International HapMap Project. Individual genotyping data from the five Stage 1 studies were pooled with HapMap r23a data to generate the first ten principal components. The first two principal components for each sample were plotted using R, version 2.13.0 (see **URLs**). We identified and excluded one participant of the KCPS-II who was more than 6 σ away from the means of PC1 and PC2 (**Figure S1**, **e**). The remaining 7,847 samples with clear East Asian origin were included in the final genome-wide association analysis (**Figure S1**). Cases and controls in each of the five studies were in the same cluster as compared with HapMap Asian samples. The quantile-quantile (Q-Q) plot for the meta-analysis was generated using R, and the inflation factor (λ) among all 7,847 participants included in Stage 1 was 1.01 (**Figure S2**). This suggests that population substructure, if present, should not have any substantial effect on our results.

Stage 1 association analysis: Dosage data of genotyped and imputed SNPs generated from MACH software were analyzed separately for each of the five studies in Stage 1 using the mach2dat program [12]. Associations between genotype dosage and CRC risk were assessed based on the log-additive model after adjusting for age, sex, and the first ten principal components. ORs associated with each SNP and 95% CIs were estimated using logistic regression models. PLINK, version 1.07 (http://pngu.mgh.harvard.edu/~purcell/plink/) was also used to analyze genotyping data and yielded results virtually identical to those derived from analyzing dosage data using mach2dat in Stage 1 [15]. A meta-analysis was conducted to combine results from the five Stage 1 studies using the program METAL based on the inverse-variance method for fixed effects [16]. SNPs were selected for replication based on the following criteria: 1) data available in each of the five stage 1 studies; 2) MAF > 5% in each of the five Stage 1 studies; 3) no heterogeneity across five studies included in Stage 1 (*P* for heterogeneity > 0.05 and I^2 < 25); 4) not in LD (r^2 < 0.2) with any of the known risk variants reported from previous GWAS; 5) not in

LD ($r^2 < 0.2$) with each other; 6) high imputation quality in each of the five studies (RSQ > 0.5); and 7) $P < 0.01$ in the combined analysis of all Stage 1 studies. When multiple SNPs showed strong LD ($r^2 \geq 0.8$), only the SNP with the lowest *P*-value was selected. A total of 67 SNPs were selected, and 64 were successfully designed and genotyped. Pooled analyses of data from all studies using SAS version 9.2 (SAS Institute) yielded results virtually identical to those from the meta-analyses described above.

Statistical analysis for replication studies: Logistic regression models were used to estimate ORs and 95% CIs associated with each SNP included in this study, assuming a log-additive model with adjustment for age and sex. A joint analysis was performed for combined data from Stages 1 and 2 of all studies, and additional adjustment for study site was performed. All analyses were performed using SAS version 9.2 (SAS Institute).

4. The Genetics and Epidemiology of Colorectal Cancer (GECCO) Consortium and the Colon Cancer Family Registry (CCFR)

The GECCO and CCFR GWAS meta-analysis included 11,870 CRC cases and 14,190 controls of European ancestry from 14 studies conducted in the USA, Europe, Canada, and Australia [17-19]. Each study is described in detail, below. CRC cases were defined as colorectal adenocarcinoma and confirmed by medical records, pathologic reports, or death certificates. All participants gave written informed consent, and studies were approved by their respective Institutional Review Boards. The GECCO GWAS consisted of participants within the French Association Study Evaluating RISK for sporadic colorectal cancer (ASTERISK); Hawaiian Colorectal Cancer Studies 2 and 3 (Colo2&3); Darmkrebs: Chancen der Verhütung durch Screening (DACHS); Diet, Activity, and Lifestyle Study (DALS); Health Professionals Follow-up Study (HPFS); Multiethnic Cohort (MEC); Nurses' Health Study (NHS); Ontario Familial Colorectal Cancer Registry (OFCCR); Physician's Health Study (PHS); Postmenopausal Hormone study (PMH); Prostate, Lung, Colorectal Cancer, and Ovarian Cancer Screening Trial (PLCO); VITamins And Lifestyle (VITAL); and the Women's Health Initiative (WHI). Phase one genotyping was conducted on a total of 1,709 colon cancer cases and 4,214 controls from PLCO, WHI, and DALS (PLCO Set 1, WHI Set 1, and DALS Set 1) using Illumina HumanHap 550K, 610K, or combined Illumina 300K and 240K, and has been described previously [18]. A total of 650 colorectal cancer cases and 522 controls from OFCCR were included in GECCO from previous genotyping using Affymetrix platforms [19]. A total of 5,540 colorectal cancer cases and 5,425 controls from ASTERISK, Colo2&3, DACHS Set 1, DALS Set 2, MEC, PMH, PLCO Set 2, VITAL, and WHI Set 2 were successfully genotyped using Illumina HumanCytoSNP. A total of 2,004 colorectal cancer cases and 2,244 controls from HPFS (2 sets), NHS (2 sets), PHS (2 sets), and DACHS set 2 were successfully genotyped using Illumina HumanOmniExpress. The two GWAS from CCFR included a population-based case-control set (CCFR Set 1; 1,171 cases and 983 controls) genotyped using Illumina Human1M or Human1M-Duo, and a sibling-pair set (CCFR Set 2; 796 cases and 802 controls) genotyped using Illumina Omni1.

Samples were excluded based on call rate, heterozygosity, unexpected duplicates, gender discrepancy, and unexpectedly high identity-by-descent or unexpected concordance (> 65%) with another individual. All analyses were restricted to samples clustering with the Utah residents with Northern and Western European ancestry from the CEPH collection (CEU) in principal component analysis, including the HapMap II populations as a reference. SNPs were excluded if they were triallelic, not assigned an rs number, or were reported as not performing consistently across platforms. Additionally, genotyped SNPs were excluded based on call rate (<

98%), lack of Hardy Weinberg Equilibrium in controls (HWE, $P < 1 \times 10^{-4}$), and minor allele frequency (MAF < 5% in Set 1 for PLCO, WHI, DALS, and OFCCR; minor allele count < 10 for remaining studies). All studies were imputed to HapMap II release 24, with the exception of OFCCR, which was imputed to HapMap II release 22. CCFR Set 1 and Set 2 were imputed using IMPUTE [20]. OFCCR was imputed using BEAGLE [21], and all other studies were imputed using MACH [12]. Imputed data were merged with genotype data such that genotype data were preferentially selected if a SNP had both types of data, unless there was a difference in terms of reference allele frequency (> 0.1) or position (> 100 base pairs), in which case imputed data were used. As a measurement of imputation accuracy we calculated R^2. Analyses of imputed data had different QC cutoffs than those for directly genotyped SNPS, as discussed above, and were restricted to SNPs with either MAF ≥ 1% or $R^2 > 0.3$, with the exception of CCFR Set 2, which was restricted to SNPs with both MAF ≥ 1% and $R^2 \geq 0.3$. After imputation and QC, a total of 2,684,119 SNPs were used in the meta-analysis of GECCO and CCFR studies.

For each study, we estimated the association between each SNP and risk for colorectal cancer by calculating betas, odds ratios (ORs), standard errors, 95% confidence intervals (CIs), and *P*-values using logistic regression models with log-additive genetic effects. Each directly genotyped SNP was coded as 0, 1, or 2 copies of the minor allele, and for imputed SNPs, we used the expected number of copies of the minor allele. Using the expected number of copies of the minor allele (dosage) has been shown to give unbiased estimates in meta-analyses [22]. We adjusted for age, sex (when appropriate), center (when appropriate), smoking status (PHS only), batch effects (ASTERISK only), and the first three principal components from EIGENSTRAT to account for population substructure. As CCFR Set 2 is a family-based study, we used conditional logistic regression stratified by family ID while adjusting for age and sex only. We conducted inverse-variance weighted, fixed-effects meta-analysis to combine beta estimates and standard errors across individual studies. We used PLINK [15] and R (see **URLs**) to conduct the statistical analysis.

5. GECCO and CCFR Study Descriptions

Ontario Familial Colorectal Cancer Registry (OFCCR): In GECCO, a subset of the Assessment of Risk in Colorectal Tumors in Canada (ARCTIC) from the Ontario Registry for Studies of Familial Colorectal Cancer (OFCCR) was used. Both the case-control study [23] and the OFCCR [24] have been described in detail previously, as have GWAS results [19].

French Association Study Evaluating RISK for sporadic colorectal cancer (ASTERISK): Participants were recruited from the Pays de la Loire region in France between December 2002 and March 2006. Details of this study have been previously described [25].

Colon Cancer Family Registry (CCFR): The CCFR is an NCI-supported consortium consisting of six centers dedicated to the establishment of a comprehensive collaborative infrastructure for interdisciplinary studies in the genetic epidemiology of colorectal cancer [26]. The Set 1 scan, which has been described previously [17], includes population-based cases and age-matched controls from three population-based centers: Seattle, Toronto and Australia. The Set 2 scan includes population-based cases and matched controls from all six Colon CFR centers including Mayo Clinic, Hawaii, University of Southern California, Seattle, Toronto, and Australia. As with Set 1, cases were genetically enriched by over-sampling those with a young age at onset or positive family history. Controls were same-generation family controls.

Darmkrebs: Chancen der Verhütung durch Screening (DACHS): This German study was initiated as a large population-based case-control study in 2003 in the Rhine-Neckar-

Odenwald region to assess the potential of endoscopic screening for reduction of colorectal cancer risk and to investigate etiologic determinants of disease. Details of this study have been described previously [27, 28]. The Set 1 scan consisted of a subset of participants recruited through 2007, and samples were frequency-matched on age and gender. The Set 2 scan consisted of additional subjects that were recruited through 2010 as part of this ongoing study.

Diet, Activity, and Lifestyle Study (DALS): DALS is a population-based case-control study of colon cancer. Participants were recruited between 1991 and 1994 from three locations: the Kaiser Permanente Medical Care Program (KPMCP) of Northern California, an eight-county area in Utah, and the metropolitan Twin Cities area of Minnesota. Details of this study have been described previously [29]. The Set I scan consisted of a subset of the study designed above, from Utah, Minnesota, and KPMCP, and was restricted to subjects who self-reported as White non-Hispanic. The Set 2 scan consisted of subjects from Utah and Minnesota that were not genotyped in Set 1. Set 2 was restricted to subjects who self-reported as White non-Hispanic and those that had appropriate consent to post data to dbGaP.

Hawai'i Colorectal Cancer Studies 2 & 3 (Colo2&3): Colo2&3 are Hawaiian population-based case-control studies of colorectal cancer that enrolled participants in the Oahu area between 1994 and 1998. Details of these studies have been described previously [30].

Health Professionals Follow-up Study (HPFS): The HPFS is a parallel prospective study to the Nurses' Health Study (NHS) that was conducted with the purpose of evaluating nutritional factors in relation to incidence of serious illnesses. The cohort includes 51,529 men in various health professions who responded to a mailed questionnaire in 1986. Details of the study have been described elsewhere [31]. Two case-control sets were constructed from which DNA was isolated from either buffy coat or buccal cells for genotyping: 1) a case-control set with cases of colorectal cancer matched to randomly selected controls who provided a blood sample and were free of colorectal cancer at the same time the colorectal cancer was diagnosed in the cases; 2) a case-control set with cases of colorectal cancer matched to randomly selected controls who provided a buccal sample and were free of colorectal cancer at the same time the colorectal cancer was diagnosed in the case. For both case-control sets, matching criteria included year of birth (within 1 year) and month/year of blood or buccal cell sampling (within six months). Cases were pair matched 1:1, 1:2, or 1:3 with a control participant(s).

Multiethnic Cohort Study (MEC): MEC was initiated in 1993 to investigate the impact of dietary and environmental factors on major chronic diseases, particularly cancer, in ethnically diverse populations in Hawai'i and California. The study recruited 96,810 men and 118,441 women ages 45 to 75 years between 1993 and 1996. Details of this study have been described previously [32].

Nurses' Health Study (NHS): The NHS cohort began in 1976 when 121,700 married female registered nurses ages 30 to 55 years returned the initial questionnaire that ascertained a variety of important health-related exposures. Details of this study have been described previously [33], and study resources are available online (http://www.channing.harvard.edu/nhs/). We constructed two case-control sets from which DNA was isolated from either buffy coat or buccal cells for genotyping: 1) a case-control set with cases of colorectal cancer matched to randomly selected controls who provided a blood sample and were free of colorectal cancer at the same time the colorectal cancer was diagnosed in the case; 2) a case-control set with cases of colorectal cancer matched to randomly selected controls who provided a buccal sample and were free of colorectal cancer at the same time the colorectal cancer was diagnosed in the cases. For both case-control sets, matching criteria included year of birth (within one year) and month/year

of blood or buccal cell sampling (within six months). Cases were pair matched 1:1, 1:2, or 1:3 with a control participant(s).

Physician's Health Study (PHS): The PHS was established as a randomized, double-blind, placebo-controlled trial of aspirin and ß-carotene among 22,071 healthy U.S. male physicians, between 40 and 84 years of age in 1982. Details have been described elsewhere [34, 35]. Among those who provided baseline blood samples, colorectal cases were ascertained through March 31, 2008, and controls were matched on age (within one year for younger participants, up to five years for older participants) and smoking status (never, past, current). Cases were "pair" matched 1:1, 1:2 or 1:3 with a control participant(s). Due to DNA availability, samples were genotyped in two batches on the same platform at the same genotyping center at different time points.

Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO): PLCO enrolled 154,934 participants (men and women between 55 and 74 years of age) at ten centers into a large, randomized, two-arm trial to determine the effectiveness of screening to reduce cancer mortality. Details of this study have been previously described [36, 37] and are available online (http://dcp.cancer.gov/plco). The Set 1 **s**can included a subset of 577 colon cancer cases self-reported as non-Hispanic White with available DNA samples, questionnaire data, and appropriate consent for ancillary epidemiologic studies. Controls come from the Cancer Genetic Markers of Susceptibility (CGEMS) prostate cancer scan [38] (all male) and the GWAS of Lung Cancer and Smoking [39] (enriched for smokers) along with an additional 92 non-Hispanic White female controls. The Set 2 scan included CRC cases from both arms of the trial, which were not already included in Set 1. Controls were frequency matched 1:1 to cases without replacement, and cases were not eligible to be controls. Matching criteria were age at enrollment (two year blocks), enrollment date (two year blocks), sex, race/ethnicity, trial arm, and study year of diagnosis (i.e. controls must be cancer free into the case's year of diagnosis).

Postmenopausal Hormones Supplementary Study to the Colon Cancer Family Registry (PMH-CCFR): PMH is a population-based case-control study evaluating the effect of postmenopausal hormone use on CRC. Female participants from 13 counties in Washington State were enrolled from 1998-2002. Details of this study have been previously described [40]. Only participants that were not part of the CCFR Seattle site were included in the sample set.

VITamins And Lifestyle (VITAL): The VITamins And Lifestyle (VITAL) cohort is comprised of 77,721 Washington State men and women ages 50 to 76 years, recruited from 2000 to 2002 to investigate the association of supplement use and lifestyle factors with cancer risk. Details of this study have been described elsewhere [41]. In GECCO, a nested case-control set was genotyped. Samples included CRC cases with DNA. Controls were matched on age at enrollment (within one year), enrollment date (within one year), sex, and race/ethnicity. One control was randomly selected per case among all controls that matched on the four factors above and where the control follow-up time was greater than follow-up time of the case until diagnosis.

Women's Health Initiative (WHI): WHI is a long-term health study of 161,808 post-menopausal women ages 50 to 79 years recruited from 40 clinical centers throughout the U.S. WHI is comprised of a Clinical Trial (CT) arm, an Observational Study (OS) arm, and several extension studies. Details of WHI have been previously described [42, 43] and are available online (https://cleo.whi.org/SitePages/Home.aspx). Set 1 cases were selected from a 2005 database and were comprised of centrally adjudicated colon cancer cases from the OS who self-reported as White. Controls were first selected among controls previously genotyped as part of a Hip Fracture GWAS conducted within the WHI OS and matched to cases on age (within three years)

enrollment date (within 365 days), hysterectomy status, and prevalent conditions at baseline. Set 2 scan cases were selected from the August 2009 database, and were comprised of centrally adjudicated colon and CRC cases from the OS and CT who were not genotyped in Set 1. Matching criteria included age (within years), race/ethnicity, WHI date (within three years), WHI Calcium and Vitamin D study date (within three years), and randomization arms (OS flag, hormone therapy assignments, dietary modification assignments, calcium/vitamin D assignments). In addition, they were matched on the four regions of randomization centers. Each case was matched with one control (1:1) that met the exact matching criteria. Control selection was done in a time-forward manner, selecting one control for each case first from the risk set at the time of the case's event.

References

1. Zheng,W. *et al.* The Shanghai Women's Health Study: rationale, study design, and baseline characteristics. *Am. J. Epidemiol.* **162**, 1123-1131 (2005).
2. Cai,H. *et al.* Dietary patterns and their correlates among middle-aged and elderly Chinese men: a report from the Shanghai Men's Health Study. *Br. J. Nutr.* **98**, 1006-1013 (2007).
3. Zheng,W. *et al.* Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat. Genet.* **41**, 324-328 (2009).
4. Zheng,W. *et al.* Genetic and clinical predictors for breast cancer risk assessment and stratification among Chinese women. *J. Natl. Cancer Inst.* **102**, 972-981 (2010).
5. Bei,J.X. *et al.* A genome-wide association study of nasopharyngeal carcinoma identifies three new susceptibility loci. *Nat. Genet.* **42**, 599-603 (2010).
6. Matsuo,K. *et al.* Association between an 8q24 locus and the risk of colorectal cancer in Japanese. *BMC. Cancer* **9**, 379 (2009).
7. Nakata,I. *et al.* Association between the SERPING1 gene and age-related macular degeneration and polypoidal choroidal vasculopathy in Japanese. *PLoS. One.* **6**, e19108 (2011).
8. Jee,S.H. *et al.* Adiponectin concentrations: a genome-wide association study. *Am. J. Hum. Genet.* **87**, 545-552 (2010).
9. Kim,J. *et al.* Dietary intake of folate and alcohol, MTHFR C677T polymorphism, and colorectal cancer risk in Korea. *Am. J. Clin. Nutr.* **95**, 405-412 (2012).
10. Long,J. *et al.* Genome-wide association study in East asians identifies novel susceptibility Loci for breast cancer. *PLoS. Genet.* **8**, e1002532 (2012).
11. Shu,X.O. *et al.* Identification of new genetic risk variants for type 2 diabetes. *PLoS. Genet.* **6**, (2010).
12. Li,Y., Willer,C.J., Ding,J., Scheet,P., & Abecasis,G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816-834 (2010).
13. Sinnott,J.A. & Kraft,P. Artifact due to differential error when cases and controls are imputed from different platforms. *Hum. Genet.* **131**, 111-119 (2012).
14. Price,A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904-909 (2006).
15. Purcell,S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559-575 (2007).
16. Willer,C.J., Li,Y., & Abecasis,G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics.* **26**, 2190-2191 (2010).
17. Figueiredo,J.C. *et al.* Genotype-environment interactions in microsatellite stable/microsatellite instability-low colorectal cancer: results from a genome-wide association study. *Cancer Epidemiol. Biomarkers Prev.* **20**, 758-766 (2011).
18. Peters,U. *et al.* Meta-analysis of new genome-wide association studies of colorectal cancer risk. *Hum. Genet.* **131**, 217-234 (2012).
19. Zanke,B.W. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat. Genet.* **39**, 989-994 (2007).
20. Marchini,J., Howie,B., Myers,S., McVean,G., & Donnelly,P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906-913 (2007).

21. Browning,S.R. & Browning,B.L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum. Genet* **81**, 1084-1097 (2007).
22. Jiao,S., Hsu,L., Hutter,C.M., & Peters,U. The use of imputed values in the meta-analysis of genome-wide association studies. *Genet. Epidemiol.* **35**, 597-605 (2011).
23. Cotterchio,M., Manno,M., Klar,N., McLaughlin,J., & Gallinger,S. Colorectal screening is associated with reduced colorectal cancer risk: a case-control study within the population-based Ontario Familial Colorectal Cancer Registry. *Cancer Causes Control* **16**, 865-875 (2005).
24. Cotterchio,M. *et al.* Ontario familial colon cancer registry: methods and first-year response rates. *Chronic. Dis. Can.* **21**, 81-86 (2000).
25. Kury,S. *et al.* The thorough screening of the MUTYH gene in a large French cohort of sporadic colorectal cancers. *Genet. Test.* **11**, 373-379 (2007).
26. Newcomb,P.A. *et al.* Colon Cancer Family Registry: an international resource for studies of the genetic epidemiology of colon cancer. *Cancer Epidemiol. Biomarkers Prev.* **16**, 2331-2343 (2007).
27. Brenner,H., Chang-Claude,J., Seiler,C.M., Rickert,A., & Hoffmeister,M. Protection from colorectal cancer after colonoscopy: a population-based, case-control study. *Ann. Intern. Med.* **154**, 22-30 (2011).
28. Lilla,C. *et al.* Effect of NAT1 and NAT2 genetic polymorphisms on colorectal cancer risk associated with exposure to tobacco smoke and meat consumption. *Cancer Epidemiol. Biomarkers Prev.* **15**, 99-107 (2006).
29. Slattery,M.L. *et al.* Energy balance and colon cancer--beyond physical activity. *Cancer Res.* **57**, 75-80 (1997).
30. Le,M.L. *et al.* Combined effects of well-done red meat, smoking, and rapid N-acetyltransferase 2 and CYP1A2 phenotypes in increasing colorectal cancer risk. *Cancer Epidemiol. Biomarkers Prev.* **10**, 1259-1266 (2001).
31. Rimm,E.B. *et al.* Validity of self-reported waist and hip circumferences in men and women. *Epidemiology* **1**, 466-473 (1990).
32. Kolonel,L.N. *et al.* A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am J Epidemiol.* **151**, 346-357 (2000).
33. Belanger,C.F., Hennekens,C.H., Rosner,B., & Speizer,F.E. The nurses' health study. *Am J Nurs.* **78**, 1039-1040 (1978).
34. Hennekens,C.H. & Eberlein,K. A randomized trial of aspirin and beta-carotene among U.S. physicians. *Prev. Med.* **14**, 165-168 (1985).
35. Christen,W.G., Gaziano,J.M., & Hennekens,C.H. Design of Physicians' Health Study II--a randomized trial of beta-carotene, vitamins E and C, and multivitamins, in prevention of cancer, cardiovascular disease, and eye disease, and review of results of completed trials. *Ann. Epidemiol.* **10**, 125-134 (2000).
36. Gohagan,J.K., Prorok,P.C., Hayes,R.B., & Kramer,B.S. The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: history, organization, and status. *Control Clin. Trials* **21**, 251S-272S (2000).
37. Prorok,P.C. *et al.* Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. *Control Clin. Trials* **21**, 273S-309S (2000).
38. Yeager,M. *et al.* Identification of a new prostate cancer susceptibility locus on chromosome 8q24. *Nat. Genet.* **41**, 1055-1057 (2009).

39. Landi,M.T. *et al.* A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am J Hum. Genet.* **85**, 679-691 (2009).
40. Newcomb,P.A. *et al.* Estrogen plus progestin use, microsatellite instability, and the risk of colorectal cancer in women. *Cancer Res.* **67**, 7534-7539 (2007).
41. White,E. *et al.* VITamins And Lifestyle cohort study: study design and characteristics of supplement users. *Am J Epidemiol.* **159**, 83-93 (2004).
42. Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. *Control Clin. Trials* **19**, 61-109 (1998).
43. Hays,J. *et al.* The Women's Health Initiative recruitment methods and results. *Ann. Epidemiol.* **13**, S18-S77 (2003).

Supplementary Table 1. Sample size and genotyping methods used in Stage 1

	Genotyped		After quality control		Genotyping Platform		Number	Inflation
Study	Cases	Controls	Cases	Controls	Cases	Controls	of SNPs[a]	factor (λ)[b]
Shanghai-1	481	2,632	474	2,628	Affymetrix 6.0	Affymetrix 6.0	502,145	1.03
Shanghai-2	296	257	254	231	Illumina OmniExpress	Illumina OmniExpress	515,701	1.03
Guangzhou-1	694	972	641	972	Illumina OmniExpress	Illumina Human610-Quad	250,612	1.02
Aichi-1	497	942	404	942	Illumina OmniExpress	Illumina HumanHap610	232,426	1.04
KCPS-II	325	977	325	976	Affymetrix 5.0	Affymetrix 5.0	312,869	1.02
Overall	2,293	5,780	2,098	5,749				1.01

[a] Number of SNPs in autosome used for imputation in GWAS.

[b] Genomic inflation factor (lambda) derived from 1,636,780 imputed SNPs with minor allele frequency > 0.05 and high imputation quality (RSQ > 0.50), adjusted with age, sex and the first ten principal components.

Supplementary Table 2. Summary for characteristics of participants in GWAS and replication studies in East Asians

Study	Population ethnicity	Sample size[a]		Mean age (years)[b]		Female (%)[b]	
		Cases	Controls	Cases	Controls	Cases	Controls
GWAS		2,098	5,749	57.13	49.07	45.85	65.39
Shanghai-1	Chinese	474	2,628	60.02	51.99	73.84	94.82
Shanghai-2	Chinese	254	231	61.16	60.75	54.72	56.71
Guangzhou-1	Chinese	641	972	54.86	47.40	36.51	27.06
Aichi-1	Japanese	404	942	59.43	47.88	37.38	47.77
KCPS-II	Korean	325	976	51.38	41.27	27.08	43.34
Replication		5,358	5,922	58.08	52.99	37.95	51.74
Guangzhou-2	Chinese	1,371	1,521	58.22	54.64	37.96	39.25
Korea-NCC	Korean	1,392	1,329	58.19	55.59	37.64	38.60
Korea-Seoul	Korean	849	673	59.05	57.19	40.99	47.85
Aichi-2	Japanese	391	1,116	59.87	54.90	36.06	72.58
Guangzhou-3	Chinese	1,355	1,283	56.71	44.45	36.90	64.09
Overall		7,456	11,671	57.81	51.06	40.17	58.46

[a] Final sample size used in statistical analysis.

[b] Samples with missing data were not included.

Supplementary Table 3. Associations of previously reported SNPs with colorectal cancer risk in East Asians

SNP[a]	Chromosome/Gene	Position (bp)[b]	Alleles[c]	Our study[d]			Published GWAS[e]			$P_{heterogeneity}$[f]
				MAF	OR (95% CI)	*P* value	MAF	OR (95% CI)	*P* value	
rs6691170	1q41/*DUSP10*	220,112,069	**T**/G	0.00	NA	NA	0.36	1.06 (1.03-1.09)	9.55×10^{-10}	NA
rs6687758	1q41/*DUSP10*	220,231,571	**G**/A	0.22	1.09 (1.03-1.16)	**0.005**	0.20	1.09 (1.06-1.12)	2.27×10^{-9}	0.950
rs10936599	3q26.2/*MYNN*	170,974,795	**T**/C	0.58	0.94 (0.89-0.99)	**0.030**	0.25	0.93 (0.91-0.96)	3.39×10^{-8}	0.640
rs7758229	6q26-q27/*SLC22A3*	160,760,242	**T**/G	0.23	1.02 (0.96-1.08)	0.580	0.22	1.28 (1.18-1.39)	7.92×10^{-9}	NA
rs16892766	8q23/*EIF3H*	117,699,864	**C**/A	0.00	NA	NA	0.07	1.25 (1.19-1.32)	3.30×10^{-18}	NA
rs10505477	8q24.21/*Unknown*	128,476,625	**A**/G	0.41	1.12 (1.03-1.21)	**0.005**	0.51	1.17 (1.12-1.23)	3.16×10^{-11}	0.350
rs6983267	8q24.21/*Unknown*	128,482,487	**G**/T	0.41	1.11 (1.05-1.17)	$\mathbf{8.53\times10^{-5}}$	0.52	1.21 (1.15-1.27)	1.27×10^{-14}	**0.021**
rs7014346	8q24.21/*Unknown*	128,493,974	**A**/G	0.30	1.11 (1.02-1.21)	**0.016**	0.37	1.19 (1.14-1.24)	8.60×10^{-26}	0.147
rs10795668	10p14/*FLJ3802842*	8,741,225	**A**/G	0.38	0.85 (0.80-0.90)	$\mathbf{3.84\times10^{-9}}$	0.33	0.89 (0.86-0.91)	2.50×10^{-13}	0.137
rs3802842	11q23/*Unknown*	110,676,919	**C**/A	0.41	1.06 (1.01-1.12)	**0.021**	0.29	1.11 (1.08-1.15)	5.82×10^{-10}	0.140
rs7136702	12q13.13/*LARP4*,*DIP2*	49,166,483	**T**/C	0.49	1.01 (0.95-1.06)	0.816	0.35	1.06 (1.04-1.08)	4.02×10^{-8}	0.087
rs11169552	12q13.3/*DIP2B*,*ATF1*	49,441,930	**T**/C	0.38	0.96 (0.91-1.02)	0.171	0.28	0.92 (0.90-0.95)	1.89×10^{-10}	0.138
rs4444235	14q22/*BMP4*	53,480,669	**C**/T	0.51	1.06 (1.01-1.12)	**0.025**	0.46	1.11 (1.08-1.15)	8.10×10^{-10}	0.159
rs4779584	15q13/*GREM1*	30,782,048	**T**/C	0.82	1.12 (1.04-1.20)	**0.002**	0.18	1.26 (1.19-1.34)	4.44×10^{-14}	**0.011**
rs9929218	16q22/*CDH1*	67,378,447	**A**/G	0.19	0.91 (0.85-0.98)	**0.010**	0.29	0.91 (0.89-0.94)	1.20×10^{-8}	0.920
rs4939827	18q21/*SMAD7*	44,707,461	**C**/T	0.74	0.89 (0.84-0.95)	$\mathbf{2.86\times10^{-4}}$	0.48	0.85 (0.81-0.89)	1.00×10^{-12}	0.191
rs10411210	19q13/*RHPN2*	38,224,140	**T**/C	0.19	0.87 (0.81-0.93)	$\mathbf{6.61\times10^{-5}}$	0.10	0.87 (0.83-0.91)	4.60×10^{-9}	0.925
rs961253	20p12/*BMP2*	6,352,281	**A**/C	0.09	1.11 (1.01-1.21)	**0.030**	0.36	1.12 (1.08-1.16)	2.00×10^{-10}	0.794
rs4925386	20q13.33/*LAMA5*	60,354,439	**T**/C	0.24	0.98 (0.92-1.05)	0.636	0.32	0.93 (0.91-0.95)	1.89×10^{-10}	0.090

[a] Two SNPs (rs6691170 and rs16892766) were monomorphic in Asians and not available in this study. SNP rs7758229 was originally identified in Japanese for distal colon cancer (heterogeneity was not tested). Two SNPs (rs10505477 and rs7014346) in LD with rs6983267 were not genotyped in Stage 2.

[b] The chromosome physical position is based on the National Center for Biotechnology Information (NCBI) database, build 36.3.

[c] Minor allele/major allele (in Europeans) is based on forward allele coding in NCBI, build 36.3.

[d] Results based on all samples (2,098 cases and 5,749 controls) in Stage 1 and 3,154 cases and 3,322 controls in Stage 2. NA, not available.

[e] Results (including MAF, ORs, 95% CIs and *P*-value) from the original studies.

[f] *P* for heterogeneity between East Asians and Europeans was calculated using a Cochran's *Q* test. NA, not available.

Supplementary Table 4. Results for 61 additional SNPs evaluated in Stage 2

					Stage 1		Stage 2	
SNP	Chr.	Position (bp)[a]	Alleles	MAF[b]	OR (95% CI)[c]	P_{trend}	OR (95% CI)[d]	P_{trend}
rs3795263	1	2,928,849	A/G	0.16	0.83 (0.74-0.93)	0.002	0.94 (0.86-1.04)	0.22
rs396617	1	6,748,016	C/T	0.11	1.19 (1.06-1.35)	0.005	0.85 (0.76-0.94)	0.002
rs484107	1	70,726,848	A/G	0.15	1.20 (1.08-1.34)	8.19×10^{-4}	1.07 (0.97-1.17)	0.16
rs17380127	1	158,661,561	C/A	0.09	1.28 (1.12-1.47)	3.75×10^{-4}	0.98 (0.88-1.10)	0.76
rs7525615	1	227,902,036	T/C	0.15	0.79 (0.70-0.89)	7.28×10^{-5}	0.99 (0.90-1.08)	0.80
rs11096507	2	18,067,272	T/G	0.11	0.78 (0.68-0.90)	5.95×10^{-4}	1.03 (0.92-1.15)	0.60
rs4666200	2	29,391,915	G/A	0.22	0.89 (0.81-0.98)	0.02	0.98 (0.91-1.07)	0.67
rs1521935	2	53,608,235	C/T	0.30	0.85 (0.78-0.93)	2.76×10^{-4}	1.05 (0.98-1.13)	0.20
rs11681872	2	53,634,294	G/A	0.49	0.87 (0.79-0.95)	0.001	1.02 (0.95-1.09)	0.62
rs6724312	2	123,272,644	T/C	0.20	1.20 (1.09-1.32)	3.33×10^{-4}	0.99 (0.91-1.08)	0.86
rs4666880	2	181,378,359	A/G	0.16	1.25 (1.10-1.42)	6.39×10^{-4}	0.96 (0.88-1.05)	0.40
rs11688453	2	220,873,163	G/A	0.18	1.18 (1.07-1.30)	0.001	0.97 (0.90-1.06)	0.50
rs16862908	3	189,236,583	T/G	0.27	1.15 (1.04-1.26)	0.004	0.97 (0.91-1.04)	0.41
rs11943776	4	36,320,611	T/C	0.31	0.86 (0.79-0.94)	5.69×10^{-4}	0.98 (0.92-1.05)	0.63
rs17329945	4	94,204,948	G/A	0.44	0.89 (0.82-0.97)	0.007	0.98 (0.92-1.05)	0.55
rs7685402	4	99,879,787	T/G	0.34	1.15 (1.06-1.25)	8.89×10^{-4}	1.03 (0.96-1.10)	0.37
rs2695220	4	102,303,638	G/A	0.20	0.79 (0.71-0.88)	2.99×10^{-5}	0.94 (0.87-1.00)	0.07
rs2850966	4	102,398,272	T/C	0.15	0.79 (0.70-0.89)	8.08×10^{-5}	0.88 (0.80-0.97)	0.01
rs12646525	4	120,721,909	T/C	0.11	1.22 (1.08-1.38)	0.001	0.98 (0.88-1.08)	0.67
rs1580743	4	121,806,719	C/T	0.10	1.24 (1.09-1.42)	0.001	1.13 (1.02-1.25)	0.02
rs10428477	4	128,509,713	G/A	0.11	0.74 (0.64-0.85)	2.35×10^{-5}	1.04 (0.94-1.16)	0.39
rs4146184	5	62,032,400	A/G	0.34	1.17 (1.08-1.28)	2.44×10^{-4}	1.03 (0.97-1.09)	0.35
rs1432881	5	166,865,098	C/T	0.41	1.21 (1.12-1.32)	4.78×10^{-6}	1.03 (0.98-1.09)	0.26
rs17732485	5	168,027,033	C/T	0.09	1.23 (1.07-1.43)	0.005	0.93 (0.83-1.04)	0.20
rs729342	5	172,208,934	G/A	0.39	1.17 (1.08-1.27)	9.02×10^{-5}	0.98 (0.92-1.05)	0.64
rs13197045	6	6,078,433	C/T	0.13	1.21 (1.08-1.35)	0.001	1.01 (0.92-1.12)	0.77
rs9385124	6	120,647,098	A/G	0.41	0.83 (0.76-0.91)	7.54×10^{-5}	0.98 (0.92-1.05)	0.53
rs2092601	6	131,759,702	A/C	0.25	1.19 (1.09-1.30)	1.32×10^{-4}	0.93 (0.86-0.99)	0.04
rs13228166	7	47,860,851	G/A	0.33	0.86 (0.79-0.94)	7.60×10^{-4}	1.00 (0.93-1.07)	0.99
rs6980283	7	146,381,187	T/G	0.26	0.86 (0.79-0.94)	0.001	1.01 (0.94-1.09)	0.79
rs4875441	8	4,988,177	A/G	0.13	0.81 (0.71-0.91)	3.66×10^{-4}	1.22 (1.10-1.35)	1.61×10^{-4}
rs4503064	8	10,616,802	G/A	0.48	1.17 (1.07-1.28)	4.17×10^{-4}	1.08 (1.01-1.16)	0.02
rs17485069	8	27,921,826	T/C	0.17	1.25 (1.13-1.39)	1.49×10^{-5}	1.08 (0.99-1.18)	0.07
rs2448170	8	98,619,180	G/A	0.15	0.79 (0.71-0.89)	9.38×10^{-5}	1.03 (0.94-1.13)	0.51
rs3002335	9	91,335,225	A/T	0.43	0.84 (0.77-0.91)	1.94×10^{-5}	0.99 (0.93-1.06)	0.78
rs6477730	9	111,700,596	T/C	0.17	0.83 (0.75-0.93)	8.43×10^{-4}	0.98 (0.90-1.07)	0.62
rs1665650	10	118,477,090	T/C	0.32	1.20 (1.10-1.31)	3.88×10^{-5}	1.10 (1.04-1.17)	0.002
rs9423343	10	125,150,957	A/G	0.22	1.18 (1.08-1.30)	4.37×10^{-4}	1.03 (0.95-1.11)	0.54
rs3987740	11	2,852,376	C/T	0.35	1.18 (1.09-1.28)	9.13×10^{-5}	1.06 (0.99-1.13)	0.11
rs2351044	11	15,535,033	A/G	0.50	0.85 (0.79-0.92)	5.88×10^{-5}	1.00 (0.94-1.07)	0.96
rs7110302	11	66,447,030	T/C	0.41	0.86 (0.79-0.93)	2.97×10^{-4}	1.02 (0.95-1.09)	0.57
rs12282262	11	74,071,586	T/C	0.27	0.80 (0.73-0.88)	3.90×10^{-6}	0.95 (0.88-1.02)	0.18
rs4384353	11	94,953,939	T/C	0.22	0.82 (0.74-0.91)	1.53×10^{-4}	0.94 (0.87-1.02)	0.11
rs1259158	11	133,529,637	A/G	0.29	1.18 (1.09-1.29)	1.33×10^{-4}	0.98 (0.91-1.05)	0.61
rs11063569	12	5,213,272	C/A	0.36	0.85 (0.78-0.93)	3.74×10^{-4}	1.04 (0.97-1.11)	0.24
rs10878048	12	62,135,182	A/G	0.27	1.18 (1.07-1.30)	6.87×10^{-4}	1.07 (0.99-1.15)	0.08
rs4913279	12	66,920,530	T/C	0.14	1.22 (1.07-1.40)	0.003	1.02 (0.93-1.12)	0.62
rs17125469	14	51,935,567	A/G	0.07	1.32 (1.13-1.53)	2.85×10^{-4}	1.03 (0.91-1.17)	0.63
rs7162303	15	27,055,170	T/G	0.14	1.21 (1.08-1.36)	0.001	1.02 (0.93-1.12)	0.71
rs8042401	15	31,845,882	A/C	0.38	0.86 (0.80-0.94)	4.67×10^{-4}	0.98 (0.92-1.05)	0.63
rs12922569	16	82,060,491	A/G	0.11	0.84 (0.73-0.96)	0.009	1.14 (1.02-1.27)	0.02
rs464274	16	88,184,132	A/G	0.32	1.15 (1.05-1.25)	0.002	0.98 (0.91-1.05)	0.52
rs2291676	17	59,558,361	T/A	0.12	1.29 (1.12-1.48)	4.47×10^{-4}	0.88 (0.80-0.97)	0.01
rs7219156	17	65,060,075	G/A	0.30	1.18 (1.09-1.29)	9.73×10^{-5}	1.00 (0.93-1.08)	0.94
rs12937299	17	66,466,419	A/G	0.37	0.86 (0.79-0.93)	3.26×10^{-4}	0.97 (0.91-1.04)	0.37
rs11664910	18	57,179,043	G/A	0.14	0.81 (0.70-0.93)	0.002	0.98 (0.89-1.07)	0.63
rs4468878	20	59,361,632	C/T	0.47	1.16 (1.07-1.26)	5.55×10^{-4}	0.93 (0.87-0.99)	0.02
rs1056930	21	15,258,675	T/C	0.06	1.26 (1.07-1.48)	0.005	0.95 (0.83-1.09)	0.49
rs2822993	21	15,280,253	C/A	0.45	0.87 (0.80-0.94)	4.38×10^{-4}	0.99 (0.92-1.05)	0.66
rs2831699	21	28,594,828	G/A	0.32	1.14 (1.05-1.24)	0.002	1.01 (0.94-1.08)	0.82
rs9615799	22	46,988,149	C/T	0.11	1.30 (1.15-1.46)	1.79×10^{-5}	0.99 (0.89-1.11)	0.89

Abbreviations: Chr., Chromosome; MAF, minor allele frequency; OR, odds ratio; CI, confidence interval.

[a] Location based on NCBI Human Genome Build 36.3.

[b] MAF in controls from Stage 1 and 2.

[c] Adjusted for age, sex, the first ten principal components and study site.

[d] Adjusted for age, sex, and study site.

Supplementary Table 5. Association of the top three SNPs with colorectal cancer risk by study populations

SNP (alleles)[a]	Population	MAF[b]	Cases	Controls	Per-allele association		$P_{heterogeneity}$[d]
					OR (95% CI)[c]	P_{trend}	
rs10774214	Chinese	0.317	3,969	6,571	1.13 (1.05-1.21)	6.45×10^{-4}	
(T/C)	Korean	0.396	2,531	2,936	1.19 (1.09-1.29)	3.56×10^{-5}	0.212
	Japanese	0.401	795	2,039	1.30 (1.12-1.49)	3.58×10^{-4}	
rs647161	Chinese	0.316	3,983	6,578	1.12 (1.05-1.20)	3.84×10^{-4}	
(A/C)	Korean	0.294	2,537	2,947	1.19 (1.09-1.29)	8.34×10^{-5}	0.153
	Japanese	0.333	795	2,039	1.29 (1.14-1.46)	8.33×10^{-5}	
rs2423279	Chinese	0.313	3,993	6,577	1.14 (1.07-1.22)	8.78×10^{-5}	
(C/T)	Korean	0.256	2,537	2,945	1.19 (1.09-1.30)	1.53×10^{-4}	0.180
	Japanese	0.334	795	2,038	1.02 (0.90-1.17)	0.71	

Abbreviations: MAF, minor allele frequency; OR, odds ratio; CI, confidence interval.

[a] Minor/major allele.

[b] MAF in controls.

[c] Adjusted for age, sex, and study site.

[d] *P* for heterogeneity between different populations was calculated using a Cochran's *Q* test.

Supplementary Table 6. Association of the top three SNPs with colorectal cancer risk by sex

SNP (alleles)[a]	Sex	MAF[b]	Cases	Controls	Per-allele association		$P_{heterogeneity}$[d]
					OR (95% CI)[c]	P_{trend}	
rs10774214	Male	0.354	4,359	4,798	1.16 (1.08-1.24)	2.95×10^{-5}	0.667
(T/C)	Female	0.350	2,935	6,746	1.18 (1.10-1.27)	8.56×10^{-6}	
rs647161	Male	0.318	4,374	4,804	1.16 (1.09-1.24)	4.69×10^{-6}	0.814
(A/C)	Female	0.310	2,940	6,758	1.16 (1.08-1.24)	6.50×10^{-5}	
rs2423279	Male	0.293	4,379	4,803	1.15 (1.07-1.23)	5.69×10^{-5}	0.696
(C/T)	Female	0.308	2,945	6,755	1.12 (1.05-1.21)	0.0016	

Abbreviations: MAF, minor allele frequency; OR, odds ratio; CI, confidence interval.

[a] Minor/major allele.

[b] MAF in controls.

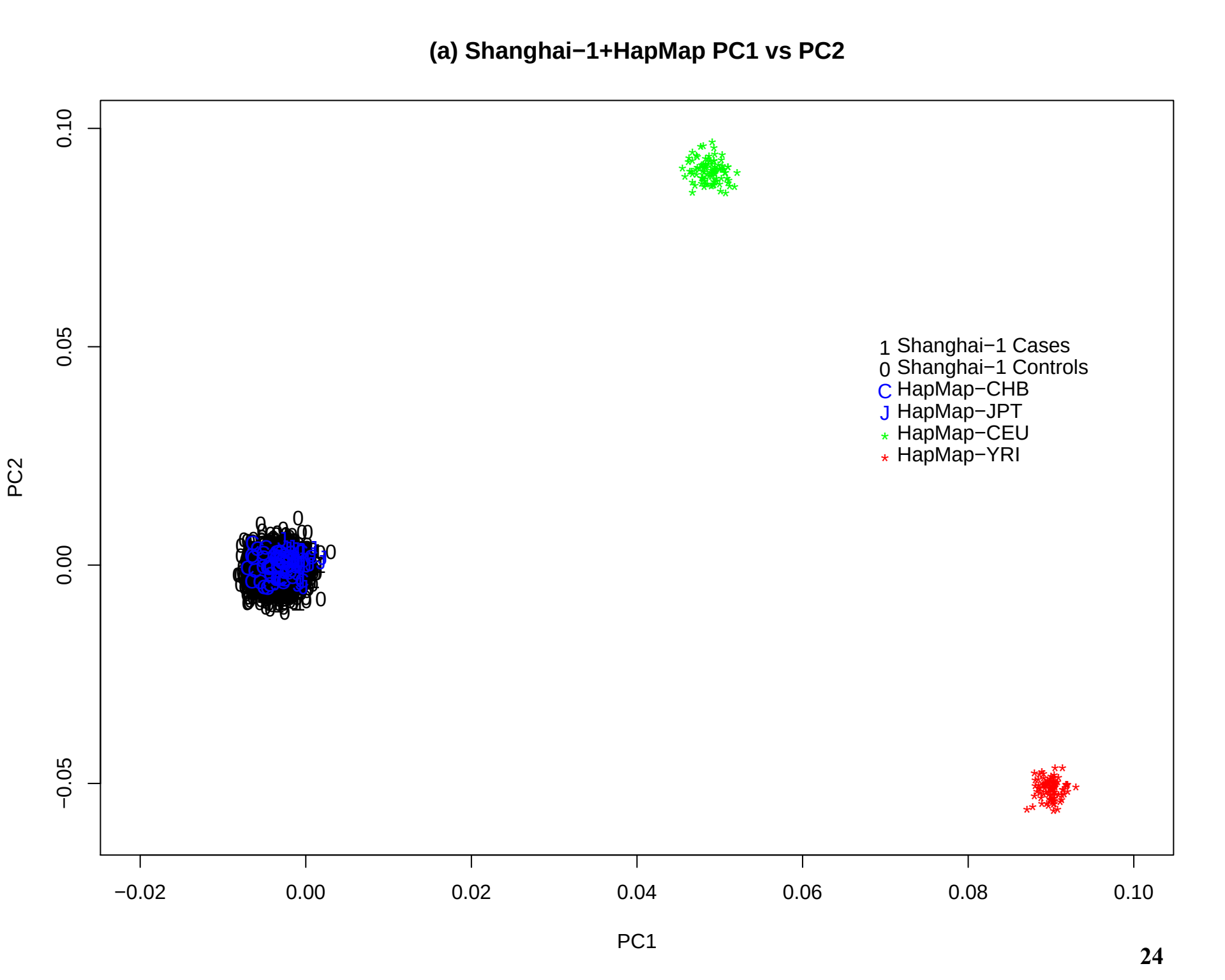
[c] Adjusted for age and study site.

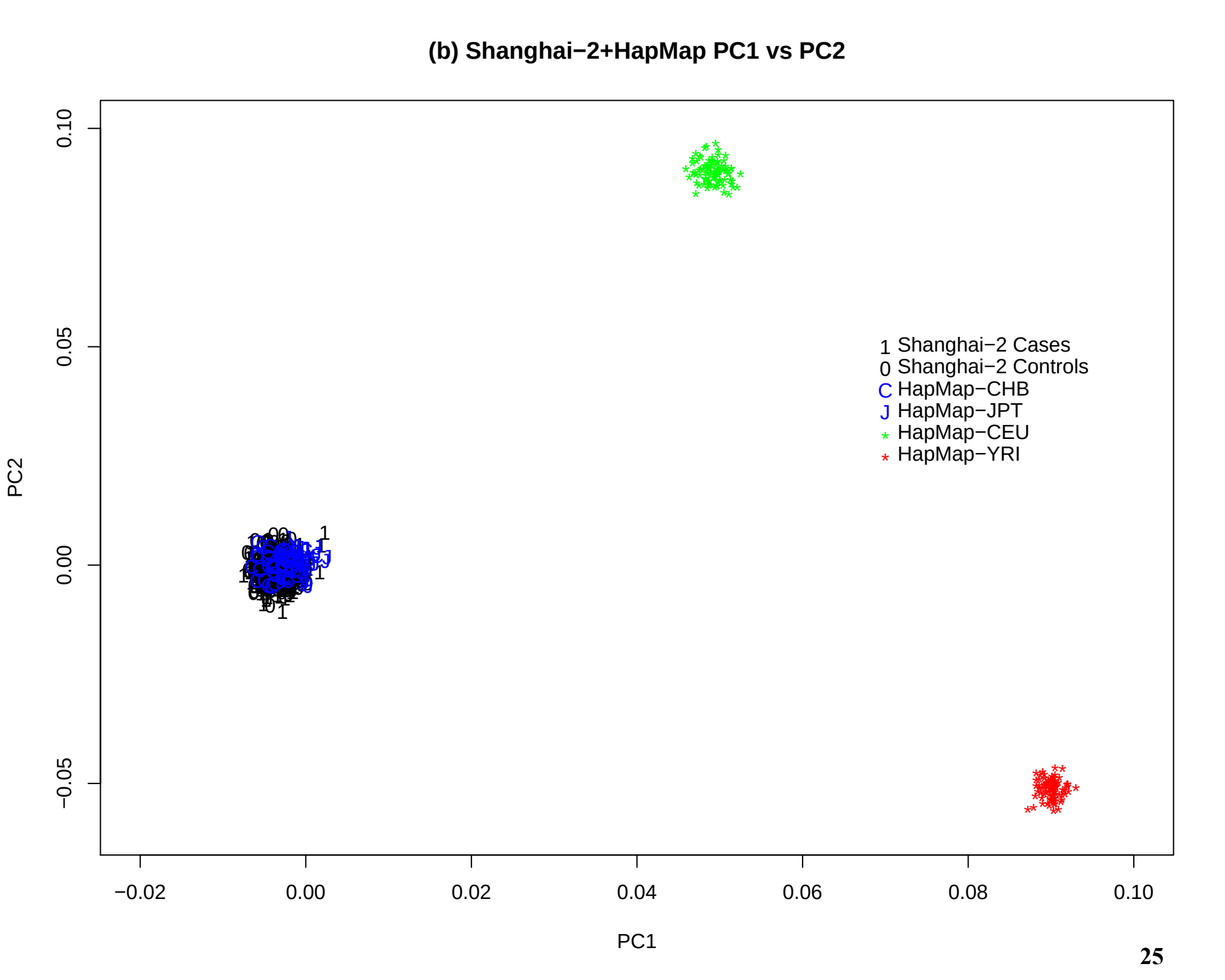
[d] *P* for heterogeneity between male and female was calculated using a Cochran's *Q* test.

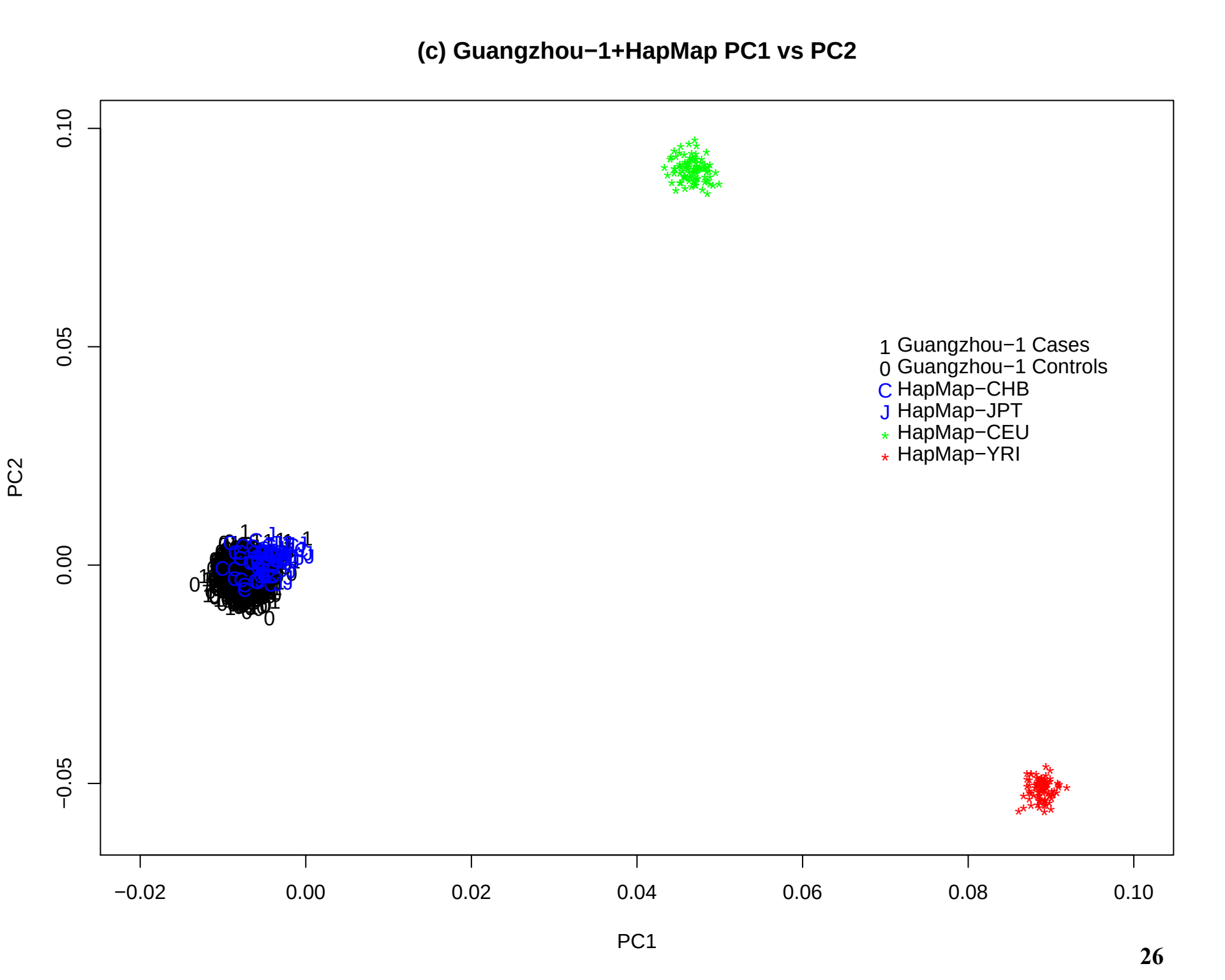
Supplementary Table 7. Concordant call between imputed and genotyped data

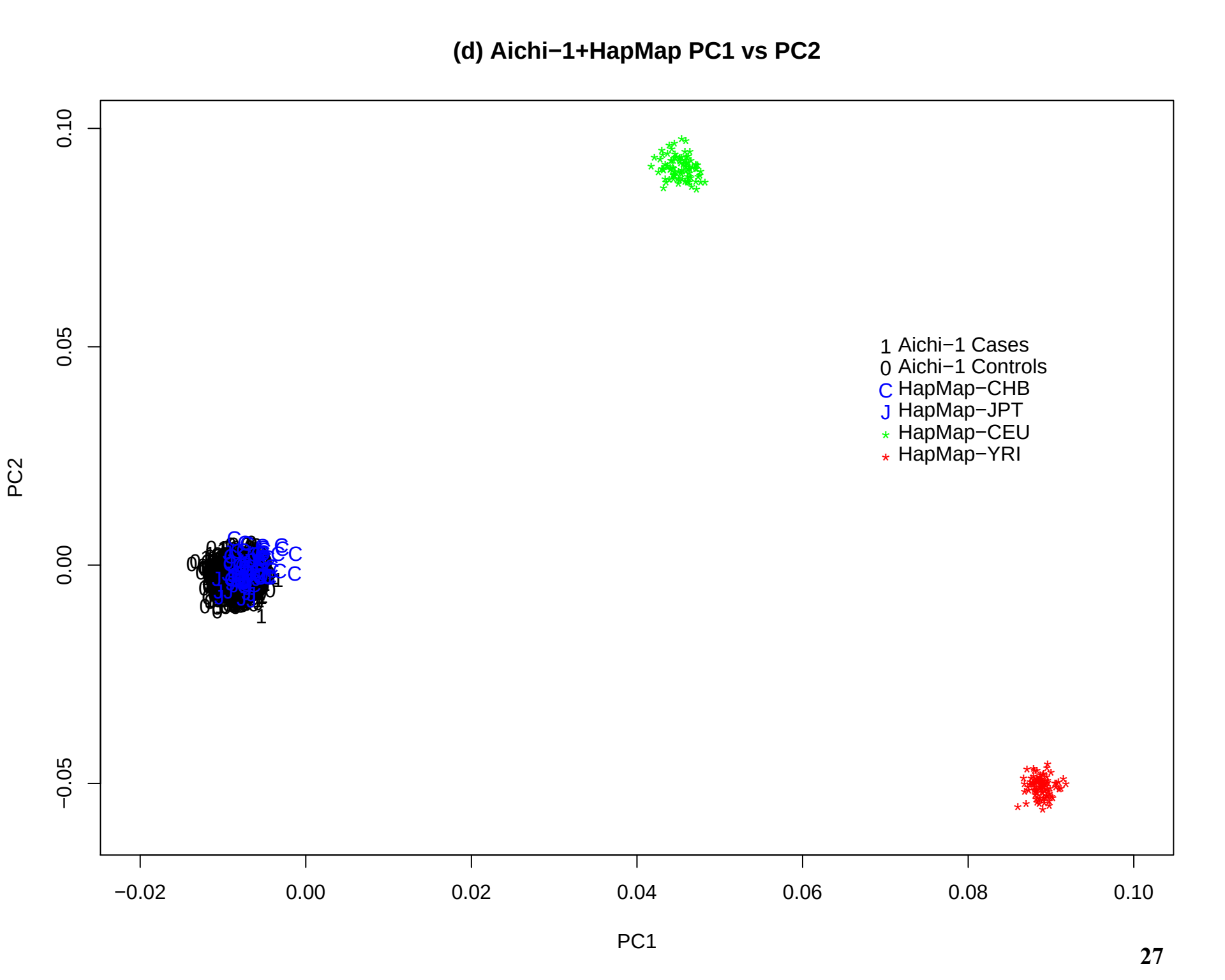
	SNP			
Study	**rs647161**	**rs10774214**	**rs2423279**	**rs1665650**
Shanghai-1	0.9874	0.9937	0.9990	0.9979
Shanghai-2	0.9868	0.9370	1.0000	0.9667
Guangzhou-1	0.9730	0.9228	0.9969	0.9656
Aichi-1	0.9667	0.9219	0.9975	0.9672
Overall	0.9805	0.9561	0.9984	0.9790

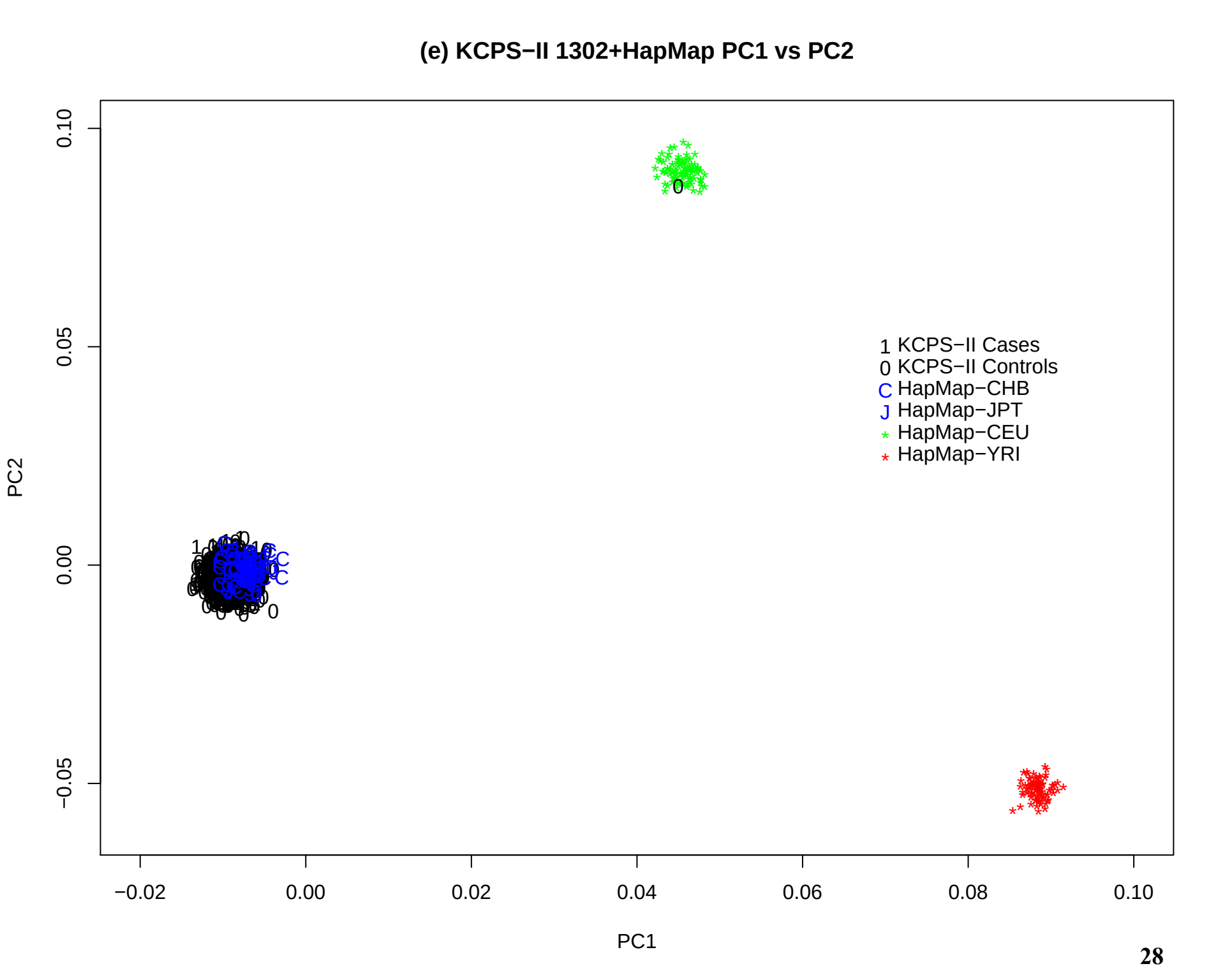
Supplementary Figure 1. Plots of the first two principal components in five studies.

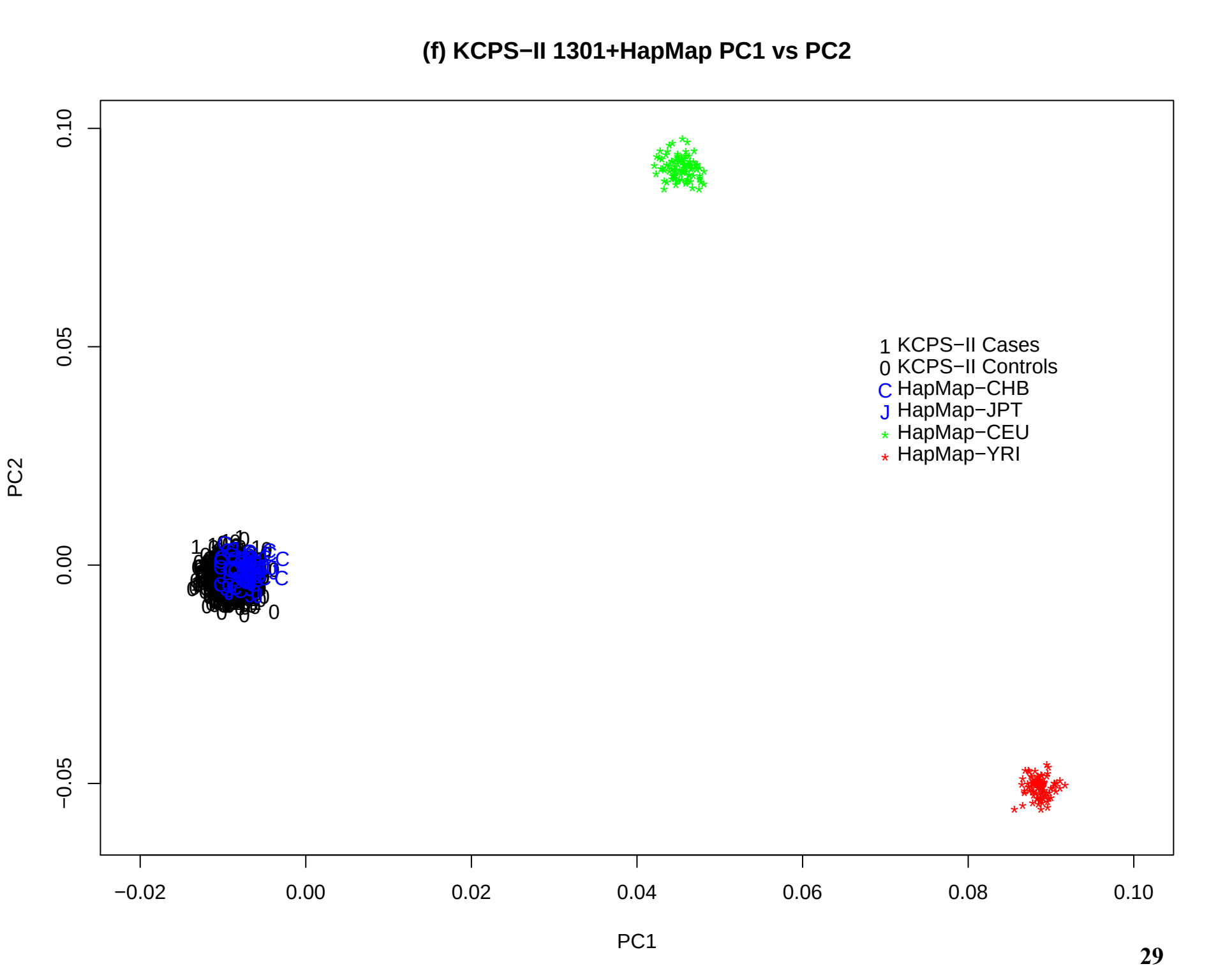

(a) Shanghai−1+HapMap PC1 vs PC2
PC2
PC1
0.10
0.05
0.00
−0.05
−0.02
0.00
0.02
0.04
0.06
0.08
0.10
1 Shanghai−1 Cases
0 Shanghai−1 Controls
C HapMap−CHB
J HapMap−JPT
HapMap−CEU
HapMap−YRI


(b) Shanghai−2+HapMap PC1 vs PC2
PC2
PC1
0.10
0.05
0.00
−0.05
−0.02
0.00
0.02
0.04
0.06
0.08
0.10
1 Shanghai−2 Cases
0 Shanghai−2 Controls
C HapMap−CHB
J HapMap−JPT
* HapMap−CEU
* HapMap−YRI

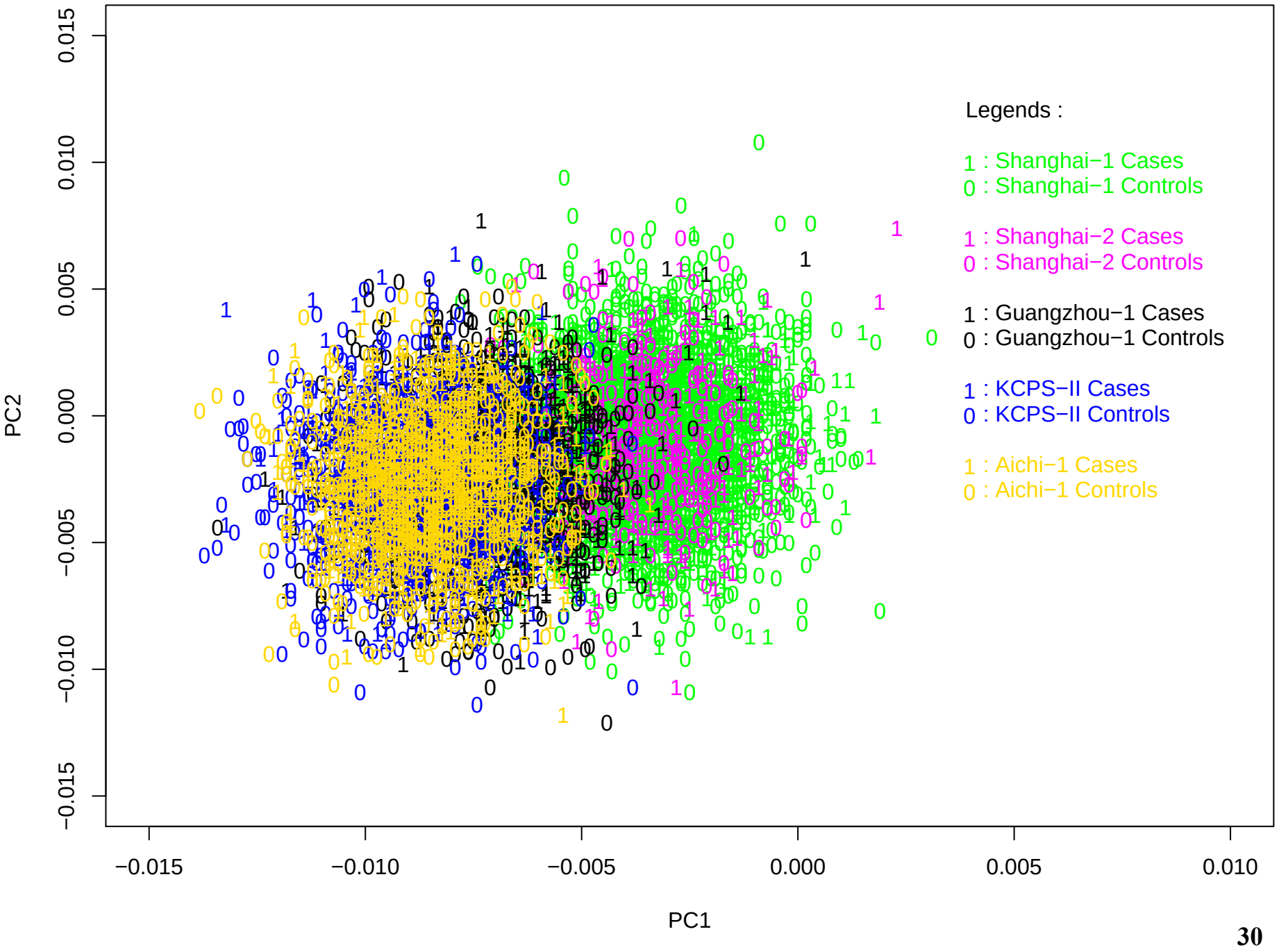

(c) Guangzhou−1+HapMap PC1 vs PC2
PC2
PC1
0.10
0.05
0.00
−0.05
−0.02
0.00
0.02
0.04
0.06
0.08
0.10
1 Guangzhou−1 Cases
0 Guangzhou−1 Controls
C HapMap−CHB
J HapMap−JPT
* HapMap−CEU
* HapMap−YRI


(d) Aichi−1+HapMap PC1 vs PC2
1 Aichi−1 Cases
0 Aichi−1 Controls
C HapMap−CHB
J HapMap−JPT
* HapMap−CEU
* HapMap−YRI
PC1
PC2
−0.02
0.00
0.02
0.04
0.06
0.08
0.10
−0.05
0.00
0.05
0.10


(e) KCPS−II 1302+HapMap PC1 vs PC2
PC2
PC1
0.10
0.05
0.00
−0.05
−0.02
0.00
0.02
0.04
0.06
0.08
0.10
1 KCPS−II Cases
0 KCPS−II Controls
C HapMap−CHB
J HapMap−JPT
* HapMap−CEU
* HapMap−YRI

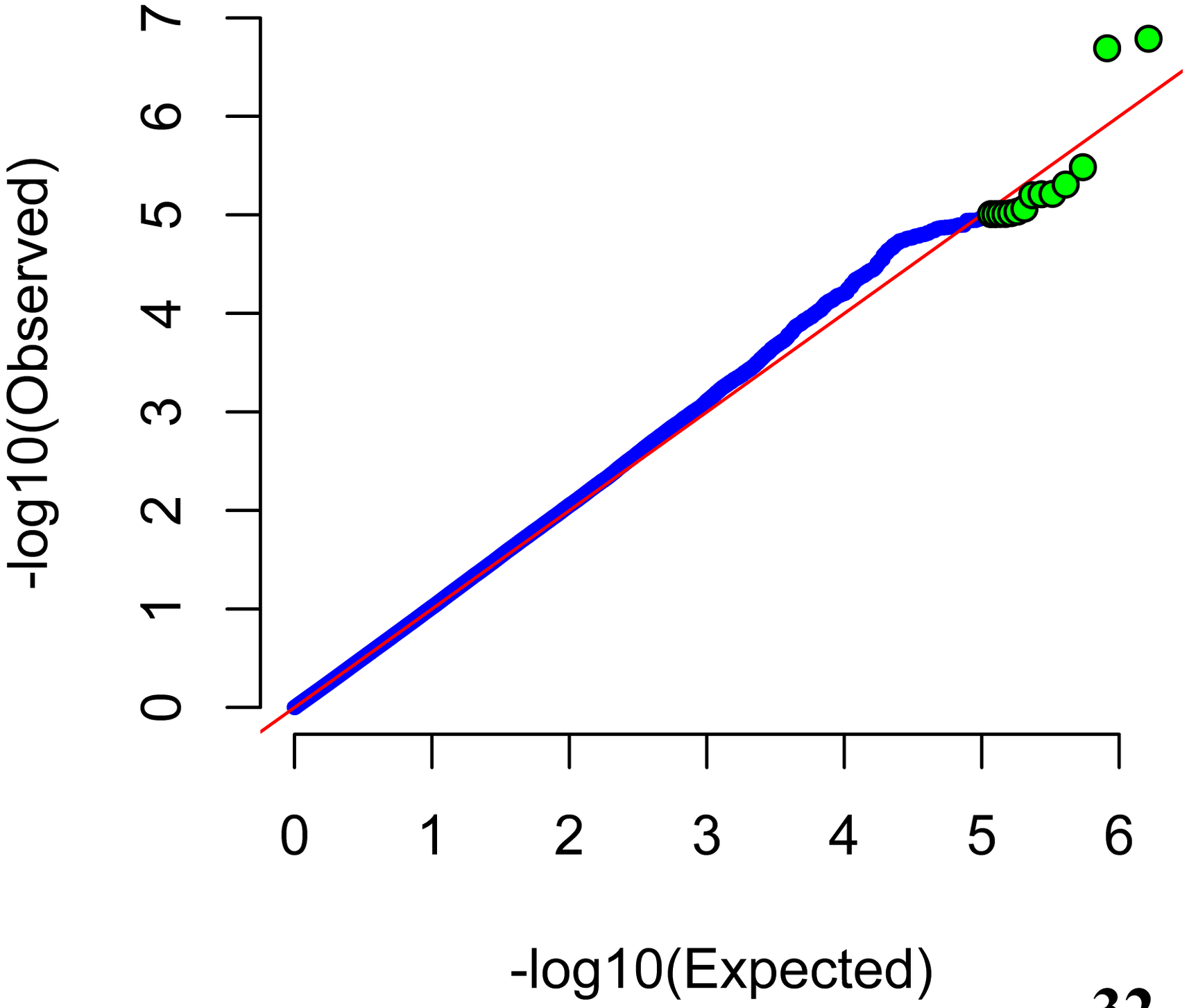

(f) KCPS−II 1301+HapMap PC1 vs PC2
1 KCPS−II Cases
0 KCPS−II Controls
C HapMap−CHB
J HapMap−JPT
* HapMap−CEU
* HapMap−YRI
PC1
PC2
−0.02
0.00
0.02
0.04
0.06
0.08
0.10
−0.05
0.00
0.05
0.10

(g) All Stage 1 samples PC1 vs PC2

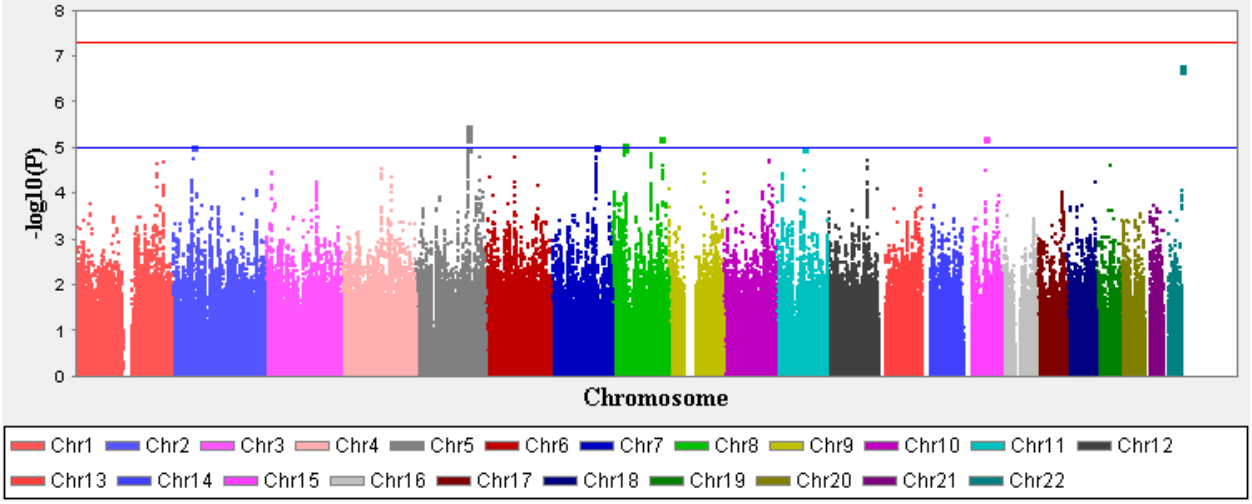


Supplementary Figure 2. Genome-wide quantile-quantile (Q-Q) plot of *P*-values for the association of SNPs with CRC risk in combined analysis of all Stage 1 data ($\lambda = 1.01$).

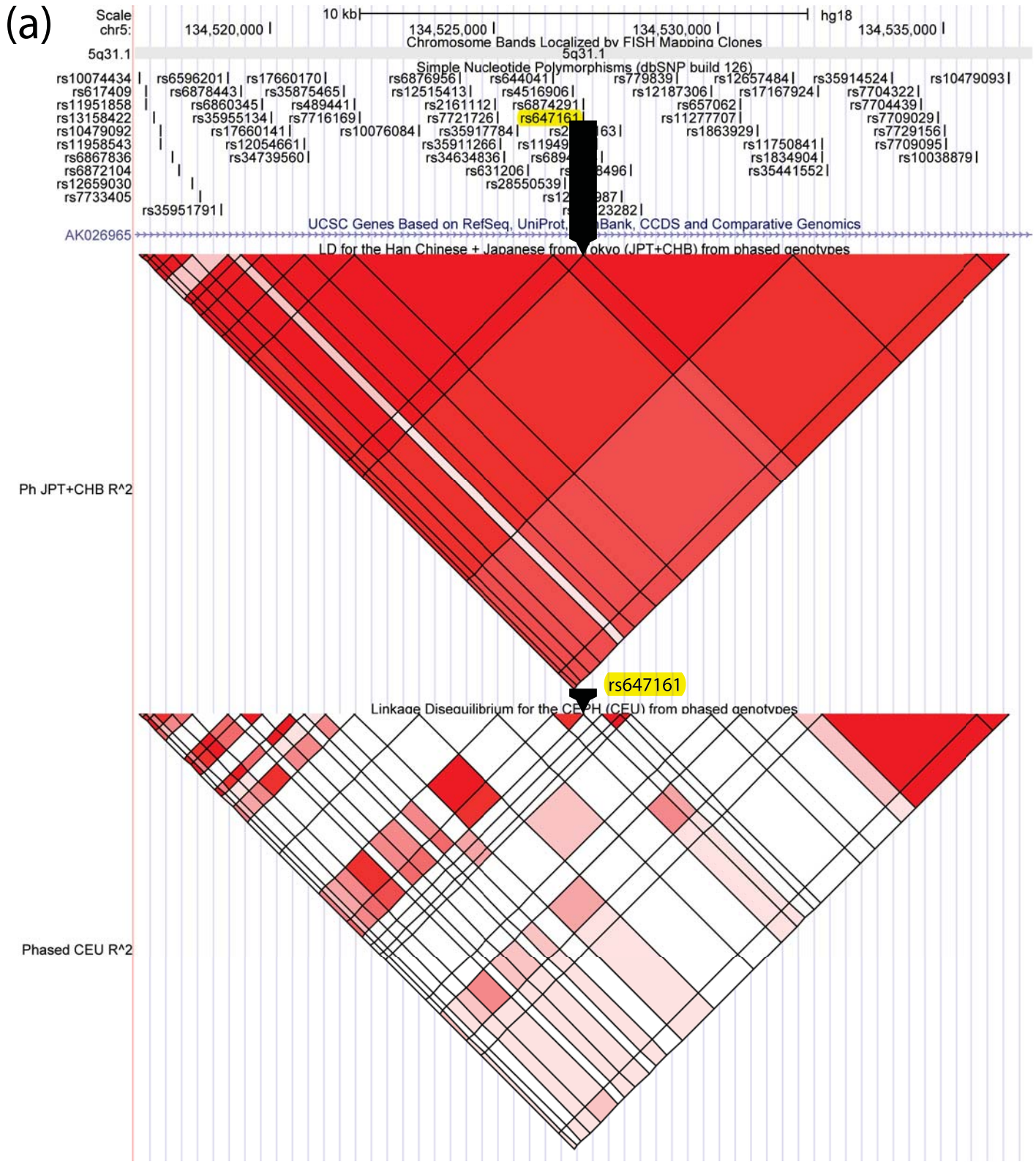
Supplementary Figure 2. Q-Q plot for Stage 1 meta-analysis (lambda=1.01)

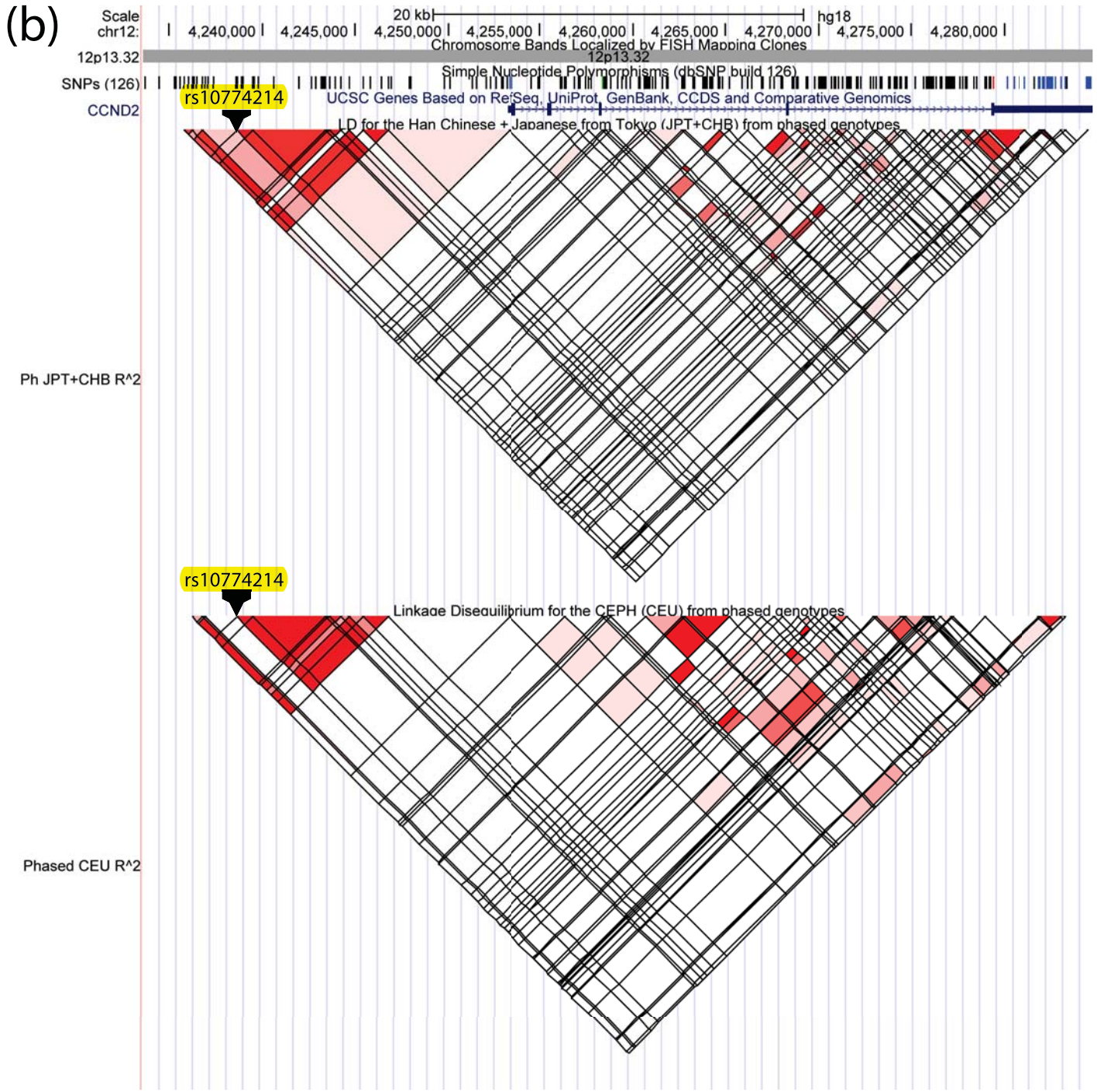


Supplementary Figure 3. Genome-wide Manhattan plot for results derived from Stage 1 meta-analysis. Shown are the log10 (*P*) using the trend test for SNPs in the 22 autosomes, adjusted for age, sex, and the first ten principal components. A total of 1,636,780 SNPs with MAF > 0.05 and high imputation quality (RSQ > 0.50) were used to generate the plot. The blue horizontal line indicates $P = 1\times10^{-5}$, and the red line indicates $P = 5\times10^{-8}$.

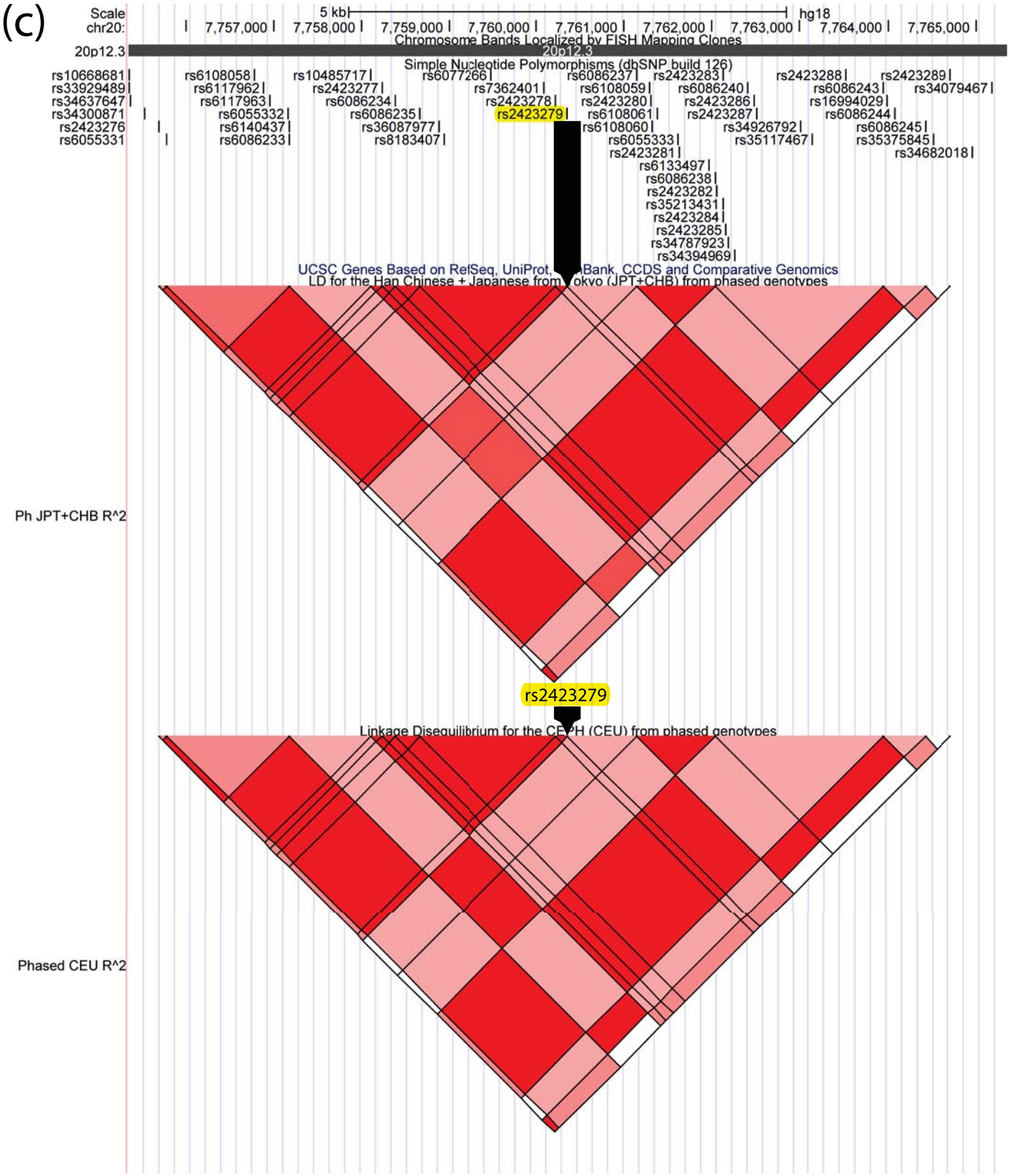


Supplementary Figure 4. LD structures of 5q31.1 (a), 12p13.32 (b), and 20p12.3 (c) loci in East Asians and Europeans.


(a)
Scale
chr5:
10 kb
hg18
134,520,000
134,525,000
134,530,000
134,535,000
Chromosome Bands Localized by FISH Mapping Clones
5q31.1
Simple Nucleotide Polymorphisms (dbSNP build 126)
rs10074434
rs6596201
rs17660170
rs6876956
rs644041
rs779839
rs12657484
rs35914524
rs10479093
rs617409
rs6878443
rs35875465
rs12515413
rs4516906
rs12187306
rs17167924
rs7704322
rs11951858
rs6860345
rs489441
rs2161112
rs6874291
rs657062
rs7704439
rs13158422
rs35955134
rs7716169
rs7721726
rs647161
rs11277707
rs7709029
rs10479092
rs17660141
rs10076084
rs35917784
rs1863929
rs7729156
rs11958543
rs12054661
rs35911266
rs7709095
rs6867836
rs34739560
rs34634836
rs1834904
rs10038879
rs6872104
rs631206
rs35441552
rs12659030
rs28550539
rs7733405
rs35951791
AK026965
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
LD for the Han Chinese + Japanese from Tokyo (JPT+CHB) from phased genotypes
Ph JPT+CHB R^2
rs647161
Linkage Disequilibrium for the CEPH (CEU) from phased genotypes
Phased CEU R^2


(b)
Scale
chr12:
20 kb
hg18
4,240,000
4,245,000
4,250,000
4,255,000
4,260,000
4,265,000
4,270,000
4,275,000
4,280,000
Chromosome Bands Localized by FISH Mapping Clones
12p13.32
12p13.32
Simple Nucleotide Polymorphisms (dbSNP build 126)
SNPs (126)
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
CCND2
rs10774214
LD for the Han Chinese + Japanese from Tokyo (JPT+CHB) from phased genotypes
Ph JPT+CHB R^2
rs10774214
Linkage Disequilibrium for the CEPH (CEU) from phased genotypes
Phased CEU R^2

(c)


Scale
chr20:
5 kb
hg18
7,757,000
7,758,000
7,759,000
7,760,000
7,761,000
7,762,000
7,763,000
7,764,000
7,765,000
Chromosome Bands Localized by FISH Mapping Clones
20p12.3
Simple Nucleotide Polymorphisms (dbSNP build 126)
rs10668681
rs33929489
rs34637647
rs34300871
rs2423276
rs6055331
rs6108058
rs6117962
rs6117963
rs6055332
rs6140437
rs6086233
rs10485717
rs2423277
rs6086234
rs6086235
rs36087977
rs8183407
rs6077266
rs7362401
rs2423278
rs2423279
rs6086237
rs2423283
rs6108059
rs6086240
rs2423280
rs6108061
rs6108060
rs6055333
rs2423281
rs6133497
rs6086238
rs2423282
rs35213431
rs2423284
rs2423285
rs34787923
rs34394969
rs2423288
rs2423289
rs6086243
rs34079467
rs2423286
rs16994029
rs2423287
rs6086244
rs34926792
rs6086245
rs35117467
rs35375845
rs34682018
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
LD for the Han Chinese + Japanese from Tokyo (JPT+CHB) from phased genotypes
Ph JPT+CHB R^2
rs2423279
Linkage Disequilibrium for the CEPH (CEU) from phased genotypes
Phased CEU R^2