Karolinska Institutet

http://openarchive.ki.se

This is a Peer Reviewed Accepted version of the following article, accepted for publication in Nature Genetics.

2013-02-07

Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk

Dunlop, Malcolm G; Dobbins, Sara E; Farrington, Susan M; Jones, Angela M; Palles, Claire; Whiffin, Nicola; Tenesa, Albert; Spain, Sarah; Broderick, Peter; Ooi, Li-Yin; Domingo, Enric; Smillie, Claire; Henrion, Marc; Frampton, Matthew; Martin, Lynn; Grimes, Graeme; Gorman, Maggie; Semple, Colin; Ma, Yusanne P; Barclay, Ella; Prendergast, James; Cazier, Jean-Baptiste; Olver, Bianca; Penegar, Steven; Lubbe, Steven; Chandler, Ian; Carvajal-Carmona, Luis G; Ballereau, Stephane; Lloyd, Amy; Vijayakrishnan, Jayaram; Zgaga, Lina; Rudan, Igor; Theodoratou, Evropi; Colorectal Tumour Gene Identification (CORGI) Consortium; Starr, John M; Deary, Ian; Kirac, Iva; Kovacevic, Dujo; Aaltonen, Lauri A; Renkonen-Sinisalo, Laura; Mecklin, Jukka-Pekka; Matsuda, Koichi; Nakamura, Yusuke; Okada, Yukinori; Gallinger, Steven; Duggan, David J; Conti, David; Newcomb, Polly A; Hopper, John L; Jenkins, Mark A; Schumacher, Fredrick; Casey, Graham; Easton, Douglas; Shah, Mitul; Pharoah, Paul; Lindblom, Annika; Liu, Tao; Swedish Low-Risk Colorectal Cancer Study Group; Smith, Christopher G; West, Hannah; Cheadle, Jeremy P; COIN Collaborative Group; Midgley, Rachel; Kerr, David J; Campbell, Harry; Tomlinson, Ian; Houlston, Richard S

Nat Genet. 2012 May 27;44(7):770-6.
Springer Nature
http://doi.org/10.1038/ng.2293
http://hdl.handle.net/10616/41405

If not otherwise stated by the Publisher's Terms and conditions, the manuscript is deposited

Copyright © 2012 Springer Nature. This manuscript is subject to Springer Nature's Terms of Use. Under no circumstances may the manuscript be shared or distributed under a Creative Commons, or other form of open access license, nor may it be reformatted or enhanced.

Common variation at 6p21.2 (*CDKN1A*), 11q13.4 (*POLD3*) and Xp22.2 influences colorectal cancer risk

Author list

Correspondence to Malcolm Dunlop, Ian Tomlinson or Richard Houlston

We performed a meta-analysis of multiple genome-wide association studies to enhance power to identify common variants modestly influencing colorectal cancer (CRC) risk. GWAS datasets comprised 9,498 cases and 10,456 controls, with replication in 9 case-control series, totalling 41,485 subjects. We identified three novel CRC risk loci at 6p21.2 (rs1321312; near *CDKN1A*; $P=2.32\times10^{-10}$), 11q13.4 (rs3824999, intronic to *POLD3*; $P=1.29\times10^{-10}$), and Xp22.2 (rs5934683; $P=3.16\times10^{-9}$) near SHROOM2 This brings to 20 the number of independent loci associated with CRC risk, and provides further insight into the genetic architecture of inherited susceptibility to CRC.

Many colorectal cancers (CRC) develop in genetically susceptible individuals; most of whom are not carriers of germ-line mismatch repair or *APC* mutations[1-3]. Genome-wide association studies (GWASs) have validated the hypothesis that part of the heritable risk of CRC is attributable to common, low-risk variants identifying CRC susceptibility loci at 1q41, 3q26.2, 8q23.3, 8q24.21, 10p14, 11q23.1, 12q13.13, 14q22.2 (x2), 15q13.3, 16q22.1, 18q21.1, 19q13.11, 20p12.3 (x2) and 20q13.33[4-10].

The modest effect sizes of individual variants thus far identified and the need for stringent thresholds for establishing statistical significance, coupled with financial constraints on numbers of variants that can be followed up, have constrained the statistical power of individual GWASs. Meta-analysis of existing GWAS data offers the opportunity to discover additional loci based on current projections for the number of independent regions harbouring common variants associated with CRC risk. In this study, we conducted a meta-analysis of GWAS data, followed by validation in multiple

independent case-control series, enabling us to identify three novel susceptibility loci for CRC.

Two of the UK GWASs were conducted by centres in London and Edinburgh, and were based on a two-phase strategy, using samples from UK populations (Supplementary Table 1). The London phase 1 (UK1) was based on genotyping 940 cases with familial colorectal neoplasia and 965 controls ascertained through the Colorectal Tumour Gene Identification (CoRGI) consortium using Illumina HumanHap550 BeadChip Arrays. Phase 1 in the Edinburgh study (Scotland1) consisted of genotyping 1,012 early-onset (aged ≤55 years) Scottish CRC cases and 1,012 controls using the Illumina HumanHap300 and HumanHap240S arrays (COGS Study). London phase 2 (UK2) samples comprised 2,873 CRC cases and 2,871 controls ascertained through the National Study of Colorectal Cancer Genetics (NSCCG). The Edinburgh phase 2 (Scotland2) was based on 2,057 cases and 2,111 controls (SOCCS Study). For phase 2, the London and Edinburgh samples were genotyped for a common set of SNPs: the 14,982 SNPs most strongly associated with colorectal neoplasia from London phase 1; the 14,972 most strongly associated SNPs from Edinburgh phase 1 (432 of these SNPs were common to both the London and Edinburgh lists of most strongly associated SNPs); and 13,186 SNPs showing the strongest association with CRC risk from a joint analysis of all CRC cases and controls from both phase 1 data sets (that were not already included in any of the preceding categories). Therefore, phase 2 was based on genotyping 43,140 SNPs in total.

The third UK GWAS (VQ58) comprised 1,800 CRC cases from the UK-based VICTOR and QUASAR2 adjuvant chemotherapy clinical trials. The CRC cases from the VQ58 study were genotyped in-house using the Illumina Hap300 and Hap370 arrays. The 2,697 controls, typed on the Illumina Human 1.2M-Duo Custom_v1 Array BeadChips, were from the UK population-based 1958 Birth Cohort, for which genotype data are publicly available from the Wellcome Trust Case-Control Consortium 2.

Prior to undertaking the meta-analysis of all GWAS datasets, we searched for potential errors and biases in data from each case-control series (Supplementary Figure 1). Comparison of the observed and expected distributions showed little evidence for an inflation of the test statistics in any of the data sets (Supplementary Figure 2), thereby excluding the possibility of significant hidden population substructure, cryptic relatedness among subjects or differential genotype calling. Principal component

analysis showed that the cases and controls were genetically well matched (Supplementary Figure 3). Any outliers or individuals identified as related were excluded (Supplementary Methods; Supplementary Figure 1).

We also made use of data on 260 SNPs on 2,151 cases and 2,501 controls which had been genotyped as part of the COINNBS series and which had been selected on the basis of a previous meta-analysis (Supplementary Table 1; Supplementary Methods).

Using data on all CRC cases and controls from these six series we derived joint odds ratios (ORs) and confidence intervals (CIs) under a fixed-effects model for each SNP, and associated *P*-values. Through these analyses we identified two SNPs, rs1321311 and rs3824999 which showed good evidence of association ($P<5.0x10^{-5}$) and mapped to distinct loci that had not previously been associated with CRC risk. This threshold for follow-up did not exclude the possibility that other SNPs represented genuine association signals, but was simply a pragmatic strategy for prioritizing replication.

To validate our findings, we conducted a replication study of rs1321311 and rs3824999 based on eight additional case-control series: UK NSCCG replication (UK2/3), Edinburgh replication (Scotland3), UK CORGI replication (UK4), Cambridge replication (Cambridge), Croatian replication (Croatia), Finnish Colorectal Cancer Predisposition Study (Helsinki), Swedish replication (Sweden), Colon Cancer Family Registry (CCFR1) and the Japanese replication (Japan) totalling 47,278 subjects (Table 1, Supplementary Table 1). In the combined analysis, both rs1321311 ($P=2.32x10^{-10}$; $P_{het}=0.99$, $I^2=0\%$) and rs3824999 ($P=1.29x\ 10^{-10}$; $P_{het}=0.99$, $I^2=0\%$) showed evidence for an association with CRC at genome-wide significance (*i.e.*, $P<5.0x10^{-8}$) (Table 1).

rs3824999 maps to 11q13.4 at 74,345,550bps, within intron 9 of the *POLD3* gene (polymerase DNA-directed delta 3; MIM 611415; Figure 1). POLD3 is a component of the DNA polymerase-δ complex which comprises proliferating cell nuclear antigen (PCNA), the multisubunit replication factor C and the 4-subunit polymerase complex: POLD1, POLD2, POLD3 and POLD4. As well as being involved in suppression of homologous recombination, the DNA polymerase-δ complex participates in DNA mismatch repair and base excision repair, key repair processes previously shown to be defective in germline CRC susceptibility disorders[11].

rs1321312 maps to 6p21.2 at 36,622,874bps within a region of linkage disequilibrium (LD) encompassing the *CDKN1A* gene (cyclin-dependent kinase inhibitor 1A; MIM 116899; Figure 1). Intriguingly, rs13211311 has been previously associated with electrocardiographic QRS duration[12]. *CDKN1A* encodes $p21^{WAF1/Cip1}$ which mediates p53-dependent G1 growth arrest[13]. Moreover, p21 acts as a master effector of multiple tumour suppressor pathways which are independent of classical p53 tumour suppression. In addition, by binding to PCNA, p21 interferes with PCNA-dependent DNA polymerase activity, thereby inhibiting DNA replication and modulating PCNA-dependent DNA repair[13]. Through binding to PCNA, p21 also competes for PCNA binding with DNA polymerase-δ and several other proteins involved in DNA synthesis, thus directly inhibiting DNA synthesis[13]. Similarly, p21 represses MYC-dependent transcription and in turn, MYC disrupts the PCNA–p21 interaction, thus alleviating p21-dependent inhibition of PCNA and DNA synthesis[13]. Decreased p21 expression has been reported to be a feature of dysplastic aberrant crypt foci in colonic mucosa and adenomas, and lymph node involvement and liver metastasis from CRC. The finding that p21 down-regulation inversely correlates with MSI status in CRC, irrespective of p53 status, again invokes a relationship with defective DNA repair and genomic instability.

To date all of the risk SNPs for CRC which have been identified map to autosomal regions of the genome and analysis of the X and Y chromosomes has been limited to the pseudo-autosomal regions. The risk of sporadic CRC is higher for males in both economically developed and less-developed countries. Furthermore, males are at greater overall CRC risk and earlier age at onset in Lynch Syndrome[14-16]. It is possible that some of these differences in risk may be attributable to sex chromosome genetic variation. To explore this hypothesis, we studied the relationship between SNPs mapping to the sex-specific region of the X-chromosome and CRC risk. Due to limited coverage of the X chromosome by UK2 and Scotland2, we made use of data provided by CCFR1 (Supplementary Table 1) in this meta-analysis. X-chromosome genotypes were analysed using an extension to the standard Cochran-Armitage test for trend as proposed by Clayton [17] (Supplementary Methods).

The SNPs showing the strongest association in meta-analysis of UK1, UK2, Scotland1, Scotland2, VQ58 and CCFR1 with support in each of these studies (Combined $P<5.0\text{x}10^{-5}$) was genotyped in UK2/3, Scotland3, UK4, Cambridge, Croatia, Helsinki, Sweden and COINNBS, totalling xxx subjects. In the combined analysis rs5934683

showed evidence for an association with CRC at genome-wide significance (P=3.16x10^{-9}, P_{het}=0.98, I^2=0%; Table 1; Supplementary Table 1).

rs5934683 maps to Xp22.2 within a 43Kb region of LD (Figure 1). Two genes map to this region, *GPR143* (G protein-coupled receptor 143; MIM300808) which is expressed by melanocytes and retinal pigment epithelium and *SHROOM2* (shroom family member 2; MIM 300103) a human homolog of the *Xenopus laevis APX* gene. While there is currently no evidence for a role of *SHROOM2* in CRC, its expression is implicated in the regulation of cellular contractility controlling endothelial morphogenesis[18] thereby representing a candidate gene for determining metastatic potential of cancers *a priori*. Like *GPR143*, SHROOM2 regulates melanosome biogenesis and localisation in the retinal pigment epithelium[19]. Intriguingly, abnormal retinal pigmentation, similar to the congenital hypertrophy of retinal pigment epithelium (CHRPE) lesions that are a component of the familial adenomatous polyposis syndrome, has been previously been shown to be an extra-colonic feature of non-FAP CRC[20,21]. To our knowledge, the relationship between Xp22.2 and CRC risk represents the first evidence for the role of X-chromosome variation in predisposition to a non-sex-specific cancer.

We assessed associations between clinico-pathological variables and genotype through case-only logistic regression. The association of rs5934683 and CRC was significantly stronger in cases with colonic disease compared to rectal disease (P=7.49x10^{-5}; based on 16,284 cases from seven data sets; Supplementary Table 2). Adjusting for multiple testing we did not find any other significant associations between SNP genotype and clinic-pathological data (specifically, sex, age at diagnosis, family history of CRC or microsatellite instability [MSI]; Supplementary Table 2).

To comprehensively analyse associations at 6p21.2, 11q13.4 and Xp22.2, we imputed unobserved genotypes in GWAS and controls using HapMap3 and 1000genomes data for the autosomal regions and HapMap release21 for Xp22.2 (Supplementary Methods; Figure 1). We did not find substantive evidence of stronger associations at the 6p21.2 and Xp22.2 risk loci. However, at the 11q13.4 locus, rs72977282, mapping 3,188bps 5' to *POLD3*, was more strongly associated with CRC than the original tagSNP rs3824999 (Figure 1; Supplementary Table 3). No non-synonymous SNPs showing strong LD (*i.e.* r^2>0.4/D'>0.8) with rs1321311, rs3824999 or rs5934683 at 6p21.2, 11q13.4 and Xp22.2 loci were identified. These data make it

likely that the associations we have identified between 6p21.2, 11q13.4 and Xp22.2 and CRC risk are mediated through changes that influence gene expression rather than impacting on protein sequence.

To examine if any directly typed or imputed SNPs lie within or very close to a putative transcription factor binding/enhancer element, we conducted a bioinformatic search of the region of association using Transfac Matrix Database[22], Encode Chipseq and DNAase I hypersensitivity data. These analyses did not provide evidence that rs4355419**,** rs3824999 and rs5934683 or closely correlated SNP maps with a known or predicted transcription regulatory region (Supplementary Table 3).

To explore whether the rs4355419**,** rs3824999 and rs5934683 associations (or SNP proxies) reflect *cis*-acting regulatory effects on *POLD3*, *CDKN1A*, *GPR143* or *SHROOM2*, we conducted expression studies using Illumina HT-12 arrays using RNA extracted from 42 samples of normal colonic epithelium (Supplementary Table 4). We also analyzed publicly-available mRNA expression data from fibroblasts, lymphoblastoid cell lines (LCL), T-cells, adipose tissue and CRC [23,24] (Supplementary Table 4). In silico analysis revealed a statistically significant relationship between rs1321311 genotype and expression of one of the transcripts of *CDKN1A.* However, this was observed only in LCLs and no effect was observed in data from normal large bowel epithelium (Supplementary Table 4). There was no relationship between rs3824999 and *POLD3* expression from the *in silico* analysis or colonic epithelium expression studies. These exploratory analyses can only detect >5% difference in expression by genotype with 80% power and levels of mRNA at a single time point hence may not adequately capture the total impact of differential expression on CRC. There was, however, a striking relationship between *SHROOM2* expression in normal colonic epithelium and rs5934693 genotype, and this was supported by *in silico* analysis CRC expression data (Supplementary Table 4). The risk allele at rs5934693 was associated with rs5934693 risk genotype in both normal colonic epithelium and CRC tissue. The relationship between *SHROOM2* expression in normal colonic epithelium and rs5934693 genotype is very strong (P=2.7x10^{-6}) and was significant even accounting for all genes tested. Indeed, rs5934693 genotype accounted for 48% of the variation in SHROOM2 expression. Exploring the relationship between *SHROOM2* expression, rs5934693 risk genotype and CRC causation will be of considerable interest, not least because of the observations of the association between excess pigmented lesions in the retinal pigment epithelium previously and

CRC[20,21]. Favored skewed X-inactivation producing a normal phenotype has been documented in X-linked dominant disease[25] and skewed X-inactivation has been implicated as a risk factor for breast cancer[26]. However, the dose-dependent relationship between rs5934693 genotype and SHROOM2 expression argues against X-inactivation as the basis for the Xp22.2 association.

By pooling GWAS data and conducting extensive replication analyses, we have identified three previously unreported loci influencing CRC susceptibility in addition to the 17 loci we have previously shown to be associated with CRC risk. The new loci identified are of modest effect size, which is unsurprising given that those with a larger impact on CRC were discovered in previous reports. While additional studies are required to determine the functional consequences that lead to CRC, our findings highlight the importance of variation in genes encoding components of the $p21^{WAF1/Cip1}$ signalling pathway in CRC. Moreover, this pathway, elucidated through the extended interaction network of *CDKN1A*, incorporates *POLD3* and *MYC* and other genes (including *SMADs* and other *TGF*-β pathway genes) that we have previously identified as risk factors for CRC.

Note: Supplementary information is available on the Nature Genetics website

URLs

The R suite can be found at http://www.r-project.org/
Detailed information on the tag SNP panel can be found at http://www.illumina.com/
dbSNP: http://www.ncbi.nlm.nih.gov/projects/SNP/
HapMap: http://www.hapmap.org/
1000Genomes: http://www.1000genomes.org/
SNAP http://www.broadinstitute.org/mpg/snap/
IMPUTE: https://mathgen.stats.ox.ac.uk/impute/impute.html
SNPTEST: http://www.stats.ox.ac.uk/~marchini/software/gwas/snptest.html
Transfac Matrix Database: http://www.biobase-international.com/pages/index.php?id=transfac
JASPAR2 database: http://jaspar.cgb.ki.se/
Wellcome Trust Case Control Consortium: www.wtccc.org.uk
Mendelian Inheritance In Man: http://www.ncbi.nlm.nih.gov/omim
SIFT: http://sift.jcvi.org/
PolyPhen: http://genetics.bwh.harvard.edu/pph/

Globocan: http://globocan.iarc.fr/

Cancer Genome Atlas project: http://cancergenome.nih.gov

The ENCODE Project: ENCyclopedia Of DNA Elements: http://www.genome.gov

Genevar (GENe Expression VARiation: http://www.sanger.ac.uk/resources

ACKNOWLEDGEMENTS

Cancer Research UK provided principal funding for this study individually to R.S.H. (C1298/A8362 - Bobby Moore Fund for Cancer Research UK), I.P.M.T., and M.G.D. At the Institute of Cancer Research additional funding was provided a Centre grant from CORE as part of the Digestive Cancer Campaign, the National Cancer Research Network and the NHS via the Biological Research Centre of the National Institute for Health Research at the Royal Marsden Hospital NHS Trust. S,L., was in receipt of a PhD studentship from Cancer Research UK, I.C., a Clinical Research Training Fellowship from St. George's Hospital Medical School and N.W., is in receipt of a PhD Studentship from the Institute of Cancer Research. M.H., was in receipt of a Post-Doctoral Training post from Leukaemia Lymphoma Research Fund.

In Oxford additional funding was provided by the Oxford Comprehensive Biomedical Research Centre (to E. Domingo and I.P.M.T.) and the EU FP7 CHIBCHA grant (to L.G.C.-C. and I.P.M.T.). Core infrastructure support to the Wellcome Trust Centre for Human Genetics, Oxford was provided by grant 075491/Z/04.

We are grateful to many colleagues within UK Clinical Genetics Departments (for CORGI) and to many collaborators who participated in the VICTOR and QUASAR2 trials. We also thank colleagues from the UK National Cancer Research Network (for NSCCG).

In Edinburgh funding was provided by a Cancer Research UK Programme Grant (C348/A12076) and a Centre Grant from the CORE Charity. E.T. was funded by a Cancer Research UK Fellowship (C31250/A10107). LYO is supported by A Cancer Research UK Researh Training Fellowship to the Edinburgh Cancer Centre. CS is supported by an MRC Research Studentship to the MRC HGU. We gratefully acknowledge the work of the COGS and SOCCS administrative teams; R. Cetnarskyj and the research nurse teams, who all recruited subjects to the studies; the Wellcome Trust Clinical Research Facility for sample preparation; and to all clinicians and pathologists throughout Scotland at the collaborating centres.

Lothian Birth Cohort Illumina genotyping was supported by the UK's Biotechnology and Biological Sciences Research Council (BBSRC). Phenotype collection in the Lothian Birth Cohort 1921 was supported by the BBSRC, The Royal Society and The Chief Scientist Office of the Scottish Government. Phenotype collection in the Lothian Birth Cohort 1936 was supported by Research into Ageing (continues as part of Age UK's The Disconnected Mind project). The work on the Lothian Birth Cohorts was undertaken in the University of Edinburgh Centre for Cognitive Ageing and Cognitive Epidemiology, part of the cross council Lifelong Health and Wellbeing Initiative (G0700704/84698). Funding from the BBSRC, EPSRC, ESRC and MRC is gratefully acknowledged.

COIN and COINB were funded by the UK Medical Research Council. COIN sample analysis (J. Cheadle) was also funded by Cancer Research Wales, Tenovus and Wales Gene Park.

For the Helsinki study, the work was supported by grants from Academy of Finland (Finnish Centre of Excellence Program 2006-2011), the Finnish Cancer Society and the Sigrid Juselius Foundation.

For the Cambridge study, we thank the SEARCH study team and all the participants in the study. P.P. is a Cancer Research UK Senior Clinical Research Fellow. This study made use of genotyping data from the 1958 Birth Cohort and NBS samples, kindly made available by the Wellcome Trust Case Control Consortium 2. A full list of the investigators who contributed to the generation of the data is available at http://www.wtccc.org.uk/. Finally, we would like to thank all individuals who participated in the study.

AUTHOR CONTRIBUTIONS

The study was designed and financial support was obtained by R.S.H., I.P.M.T., and M.G.D. The manuscript was drafted by R.S.H., I.P.M.T., and M.G.D. Statistical and bioinformatic analyses were conducted by S.E.D., N.W, with contributions from Y.M, M.H., CS, GG, R.S.H, and P.B.

Institute of Cancer Research and local collaborators: subject recruitment and sample acquisition to NSCCG were undertaken by S.P. The coordination of sample preparation and genotyping was performed by P.B. Sample preparation and genotyping were performed by A.L. B.O. and N.W. Tumour pathology analyses were performed by I.C. and S.L.

Oxford and local collaborators: subject recruitment and sample acquisition were done by E.B., M.G., L.M., A.M.L., D.G.R.E., E.R.M., H.J.W.T. and members of the CORGI Consortium, and by R. Mager, R. Midgley, E.J. and D.J.K. Sample preparation was performed by K.H., S.L.S. and E.E.M.J. Genotyping was performed and coordinated by L.G.C.-C., K.H., A.M.J., M.C., E.E.M.J., A.W. and E. Domingo.

Colon Cancer Genetics Group, Edinburgh and local collaborators: subject recruitment and sample acquisition were performed by S.F. SH, ID, HC and MGD and members of the SOCCS and COGS study teams. Sample preparation was coordinated by S.F. Genotyping was performed and coordinated by S.M.F., and M.G.D. Data curation and analysis in Edinburgh was conducted by ET, LZ, JP, AT and SB. Recruitment sample preps, wet lab expression analysis and genotyping was performed by LYO and CS. GG and CS performed the bioinformatic analyses.

The following authors from collaborating groups conceived the local or national study, undertook assembly of case-control series in their respective regions, collected data and samples and variously undertook genotyping and analysis: C.G.S., J. Colley, S.I., T.M. and J. Cheadle in Cardiff; I.N., S.T. and L.A.A. in Finland; and P.P. in Cambridge. All other authors undertook sample collection and phenotype data collection and collation in the respective centres.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

METHODS

Ethics statement

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

Subjects

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

Discovery screen data sets

UK1 (CORGI)[8] comprised 922 cases with colorectal neoplasia (47% male) ascertained through the Colorectal Tumour Gene Identification (CoRGI) consortium. All had at least one first-degree relative affected by CRC and one or more of the following phenotypes: CRC at age 75 or less; any colorectal adenoma (CRAd) at age 45 or less; ≥3 colorectal adenomas at age 75 or less; or a large (>1 cm diameter) or aggressive (villous and/or severely dysplastic) adenoma at age 75 or less. The 929 controls (45% males) were spouses or partners unaffected by cancer and without a personal family history (to 2nd degree relative level) of colorectal neoplasia. Known dominant polyposis syndromes, HNPCC/Lynch syndrome or bi-allelic *MYH* mutation carriers were excluded. All cases and controls were of white UK ethnic origin.

Scotland1 (COGS)[8] included 980 CRC cases (51% male; mean age at diagnosis 49.6 years, SD±6.1) and 1,002 cancer-free population controls (51% male; mean age 51.0 years; SD±5.9). Cases were for early age at onset (age ≤55 years). Known dominant polyposis syndromes, HNPCC/Lynch syndrome or bi-allelic *MYH* mutation carriers were excluded. Control subjects were sampled from the Scottish population NHS registers, matched by age (±5 years), gender and area of residence within Scotland.

VQ58 comprised 1,832 CRC cases (1,099 males, mean age of diagnosis 62.5 years; SD±10.9) from the VICTOR[27] and QUASAR2 (www.octo-oxford.org.uk/alltrials/trials/q2.html) trials. There were 2,720 population control genotypes (1,391 males,) from the Wellcome Trust Case-Control Consortium 2 (WTCCC2) 1958 birth cohort (also known as the National Child Development Study),

which included all births in England, Wales and Scotland during a single week in 1958[28].

UK2 (NSCCG)[8] consisted of 2,854 CRC cases (58% male, mean age at diagnosis 59.3 years; SD±8.7) ascertained through two ongoing initiatives at the Institute of Cancer Research/Royal Marsden Hospital NHS Trust (RMHNHST) from 1999 onwards - The National Study of Colorectal Cancer Genetics (NSCCG)[29] and the Royal Marsden Hospital Trust/Institute of Cancer Research Family History and DNA Registry. The 2,822 controls (41% males; mean age 59.8 years; SD±10.8) were the spouses or unrelated friends of patients with malignancies. None had a personal history of malignancy at time of ascertainment. All cases and controls had self-reported European ancestry, and there were no obvious differences in the demography of cases and controls in terms of place of residence within the UK.

Scotland2 (SOCCS)[8] comprised 2,024 CRC cases (61% male; mean age at diagnosis 65.8 years, SD±8.4) and 2,092 population controls (60% males; mean age 67.9 years, SD±9.0) ascertained in Scotland. Cases were taken from an independent, prospective, incident CRC case series and aged <80 years at diagnosis. Control subjects were population controls matched by age (±5 years), gender and area of residence within Scotland.

The Colon Cancer Family Registry (CCFR) data set comprised 1,332 familial CRC cases and 1,084 controls Colon Cancer Family Registry (Colon-CFR) (http://epi.grants.cancer.gov/CFR/about_colon.html)[30]. The cases were recently diagnosed CRC cases reported to population complete cancer registries in the USA (Puget Sound, Washington State) who were recruited by the Seattle Familial Colorectal Cancer Registry; in Canada (Ontario) who were recruited by the Ontario Familial Cancer Registry; and in Australia (Melbourne, Victoria) who were recruited by the Australasian Colorectal Cancer Family Study. Controls were population-based and for this analysis were restricted to those without a family history of colorectal cancer.

The COIN samples were 2,151 cases derived from the COIN and COIN-B clinical trials of metastatic CRC[31]. Median age was 63 years. COIN cases were compared against genotypes from 2,501 population controls (1,237 males,) from the WTCCC2 National Blood Service (NBS) cohort (50% male; mean age at diagnosis 53.2 years, SD±15.4).

Replication data sets

UK3 (NSCCG)[8] comprised 7,912 CRC cases (65% male; mean age at diagnosis 59 years, SD±8.2) and 4,398 controls (40% male; mean age 62 years, SD±11.5) ascertained through NSCCG post-2005[29].

Scotland3 (SOCCS)[8] comprised 1,145 CRC cases (50% male; mean age at diagnosis 53.2 years, SD±15.4) and 2,203 cancer-free population controls (47% male; mean age 51.8 years, SD±11.5). Controls comprised cancer-free participants in the Lothian Birth Cohort 1921 and Lothian Birth Cohort 1936.

UK4 (CORGI2BCD)[8] consisted of 621 CRC cases (46% male; mean age at diagnosis 58.3 years; SD±14.1) and 1,121 cancer-free population or spouse controls (45% male; mean age 45.1 years, SD±15.9).

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

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

The Swedish study comprised CRC patients were recruited within a Swedish national study conducted by the Swedish Low-Risk Colorectal Cancer Study Group. Samples were obtained during 2004-2009 from 14 different surgical clinics in central Sweden. All CRC patients during the study period were eligible for recruitment and were invited to participate. Only those too ill or too frail to consent were excluded. Controls comprised blood donors from Stockholm and Uppsala. Fully informed consent was obtained in accordance with the Swedish law concerning ethical approval of research on human subjects (2002:489,2003:198,2010:1213-31/4).

The Croatian study subjects were recruited from surgical hospitals in Zagreb. Controls were healthy volunteers from a similar urban population (Split).

The Japanese study comprised 1583 colorectal cancer cases and 1897 control subjects as describe previously[32]. All cases and controls were obtained from Biobank Japan. These samples were genotyped using the Illumina Human610-QuadBeadChip in cases and the Illumina HumanHap550v3 BeadChip in controls. Exclusion criteria: Samples with a call rate of < 0.98, SNP quality call rates <0.95, Hardy-Weinberg $P<1.0\times10^{-7}$ in controls.

Sample preparation and genotyping

DNA was extracted from samples using conventional methods and quantified using PicoGreen (Invitrogen). The VQ, UK1, and Scotland1 GWA cohorts were genotyped using Illumina Hap300, Hap370, or Hap550 arrays. 1958BC and NBS genotyping was performed as part of the WTCCC2 study on Hap1.2M-Duo Custom arrays. The CCFR samples were genotyped using Illumina Hap1M or Hap1M-Duo arrays. In UK2 and Scotland2, genotyping was conducted using custom Illumina Infinium arrays according to the manufacturer's protocols. Some COIN SNPs were typed on custom Illumina Goldengate arrays. To ensure quality of genotyping, a series of duplicate samples was genotyped, resulting in 99.9% concordant calls in all cases. Other genotyping was conducted using competitive allele-specific PCR KASPar chemistry (KBiosciences Ltd, Hertfordshire, UK), Taqman (Life Sciences, Carlsbad, California) or MassARRAY (Sequenom Inc., San Diego, USA). All primers, probes and conditions used are available on request. Genotyping quality control was tested using duplicate DNA samples within studies and SNP assays, together with direct sequencing of subsets of samples to confirm genotyping accuracy. For all SNPs, >99% concordant results were obtained.

Quality control and sample exclusion

We excluded SNPs from analysis if they failed one or more of the following thresholds: GenCall scores <0.25; overall call rates <95%; MAF<0.01; departure from Hardy-Weinberg equilibrium (HWE) in controls at $P<10^{-4}$ or in cases at $P<10^{-6}$; outlying in terms of signal intensity or X:Y ratio; discordance between duplicate samples; and, for SNPs with evidence of association, poor clustering on inspection of X:Y plots. We excluded individuals from analysis if they failed one or more of the following thresholds: duplication or cryptic relatedness to estimated identity by descent (IBD)

>6.25%; overall successfully genotyped SNPs<95%; mismatch between predicted and reported gender; outliers in a plot of heterozygosity *versus* missingness; and evidence of non-white European ancestry by PCA-based analysis in comparison with HapMap samples (http://hapmap.ncbi.nlm.nih.gov). Details of all sample exclusions are provided in Supplementary Table 2.

To identify individuals who might have non-northern European ancestry, we merged our case and control data from all sample sets with the 60 European (CEU), 60 Nigerian (YRI), and 90 Japanese (JPT) and 90 Han Chinese (CHB) individuals from the International HapMap Project. For each pair of individuals, we calculated genome-wide identity-by-state distances based on markers shared between HapMap2 and our SNP panel, and used these as dissimilarity measures upon which to perform principal components analysis. Principal components analysis was performed in R using CEU, YRI and HCB HapMap samples as reference. The first two principal components for each individual were plotted and any individual not present in the main CEU cluster (that is, >5% of the PC distance from HapMap CEU cluster centroid) was excluded from subsequent analyses (Supplementary Table 2).

We had previously shown the adequacy of the case-control matching and possibility of differential genotyping of cases and controls using Q-Q plots of test statistics. The inflation factor λ_{GC} was calculated by dividing the mean of the lower 90% of the test statistics by the mean of the lower 90% of the expected values from a χ^2 distribution with 1 d.f. Deviation of the genotype frequencies in the controls from those expected under HWE was assessed by χ^2 test (1 d.f.), or Fisher's exact test where an expected cell count was <5.

Statistical and bioinformatic analysis

Main analyses were undertaken using R (v2.6), Stata v.11 (State College, Texas, US) and PLINK (v1.06) software[33]. The association between each SNP and risk of CRC was assessed by the Cochran-Armitage trend test. Odds ratios (ORs) and associated 95% confidence intervals (CIs) were calculated by unconditional logistic regression. Meta-analysis was conducted using standard methods[34]. Cochran's Q statistic to test for heterogeneity[34] and the I^2 statistic to quantify the proportion of the total variation due to heterogeneity were calculated[35]. I^2 values ≥75% are considered characteristic of large heterogeneity[35,36]. Associations by sex, age and clinic-pathological phenotypes were examined by logistic regression in case-only analyses.

For SNPs on the non-pseudoautosomal region of X chromosome males carry only one copy and in females most loci are subject to X inactivation[37]. To test for X chromosome associations we used an extension to the standard, 1df Cochran-Armitage test for trend, proposed by Clayton (2008)[17] whereby males can be regarded as homozygous females. This 1df trend test adjusts for the different variances for males and females.

Prediction of the untyped SNPs was carried out using IMPUTEv2, based on HapMap Phase III haplotypes release 2 (HapMap Data Release 27/phase III Feb 2009 on NCBI B36 assembly, dbSNP26) and 1000genomes. Imputation of the X chromosome loci was only possible using IMPUTEv1 with HapMap Data Release 21 on NCBI Build 35. Imputed data were analysed using SNPTEST v2 to account for uncertainties in SNP prediction. An imputation info score of 0.95 was used to remove SNPs with poor imputation quality. LD metrics between HapMap SNPs were based on Data Release 27/phase III (Feb 2009) on NCBI B36 assembly, dbSNP26, viewed using Haploview software (v4.2) and plotted using SNAP. LD blocks were defined on the basis of HapMap recombination rate (cM/Mb) as defined using the Oxford recombination hotspots[38] and on the basis of distribution of confidence intervals defined by Gabriel *et al* [39]. To annotate potential regulatory sequences within disease loci we implemented *in silico* searches using Transfac Matrix Database v7.29 [22], and PReMod10[40] [40]software. We used the *in silico* algorithms SIFT and PolyPhen to predict the impact of amino acid substitutions.

Relationship between SNP genotype and mRNA expression

Expression studies in colonic epithelium

To examine for a relationship between SNP genotype and mRNA expression in colonic epithelium, 42 samples were collected fresh immediately after surgical resection of specimens for colorectal cancer (n=34), solitary adenoma (n=5) or benign conditions (not inflammatory bowel disease) (n=3). Normal epithelium was dissected from muscularis propria, and samples snap frozen and placed in RNAlater (Applied Biosystems) and kept at 4°C overnight before storage at -80°C. Tissue was disrupted and homogenised using TissueLyser LT (Qiagen), and RNA extracted using Ribopure kit (Applied Biosystems). RNA integrity and concentration was assessed on an Agilent Bioanalyzer, RNA purity (A260/A280 and A260/A230) on Nanodrop. RT PCR products

were analysed on HumanHT-12 Expression BeadChip which were scanned using the Illumina HiScan. Array data processing and analysis was performed using Illumina GenomeStudio software (version 2011.1). Microarray data were exported from Illumina Beadstudio software, processed and normalized using the R, Bioconductor beadarray and limma packages. Prior to normalization probes that were not detected (detection *P*-value>0.01) on the microarrays were removed. Microarrays were Quantile normalized to remove technical variation. The average signal of the replicates patients samples were used for further analysis. The limma package was used to find differential expressed genes, using the functions lmFit, eBayes and topTable. The spearman rank correlation of probe signals to risk SNPs, and associated significance value, was calculated using the 0-1-2 model. P values were corrected for multiple testing using the Benjamini & Hochberg method from the p.adjust R function.

In silico analysis of publicly available expression data

We analysed expression data generated from: (1) Fibroblast, LCL and T-cells derived from the umbilical cords of 75 Geneva GenCord individuals[23]; (2) 166 adipose, 156 LCL and 160 skin samples derived from a subset of healthy female twins of the MuTHER resource[24] using Sentrix Human-6 Expression BeadChips (Illumina, San Diego, USA)[41,42] (3) AgilentG4502A_07_3 custom gene expression data on 154 CRCs as part of the Cancer Genome Atlas project: http://cancergenome.nih.gov. Power of assays to establish a relationship between genotype and expression we made using STATA software (Version 10, Station College Tx, USA) assuming allele-based test of difference in normalized expression (imposing a Bonferroni correction to address multiple testing).

Assignment of microsatellite instability (MSI) in colorectal cancers

Tumour MSI status in CRCs was determined using the mononucleotide microsatellite loci BAT25 and BAT26, which are highly sensitive MSI markers. Briefly, 10 mm sections were cut from formalin-fixed paraffin-embedded CRC tumours, lightly stained with toluidine blue and regions containing at least 60% tumour microdissected. Tumour DNA was extracted using the QIAamp DNA Mini kit (Qiagen, Crawley, UK) according to the manufacturer's instructions and genotyped for the BAT25 and BAT26 loci using either ^{32}P–labelled or fluorescently-labelled oligonucleotide primers (UK2/3 and COINNBS studies respectively). Samples showing more than or equal to five novel alleles, when compared with normal DNA, at either or both markers were assigned as MSI-H (corresponding to MSI-high)[43].

TABLE AND FIGURE LEGENDS

TABLES

Table 1: Summary results for rs1321311 (6p21.31), rs3824999 (11q13.4) and rs5934683 (Xp22.2) SNPs associated with CRC risk. [a]Risk allele frequency (RAF). [b]Odds ratio. [c]95% Confidence Interval.

FIGURES

Figure 1: Regional plots of association results and recombination rates for the 6p21.2, 11q13.4, Xp22.2 susceptibility loci. (a-d) Association results of both genotyped (triangles) and imputed (circles) SNPs in the GWAS samples and recombination rates within the loci: (a) 6p21.2, (b), 11q13.4**,** (c) Xp22.2. For each plot, $-\log_{10} P$ values (y axis) of the SNPs are shown according to their chromosomal positions (x axis). The top genotyped SNP in each combined analysis is a large triangle and is labelled by its rsID. The colour intensity of each symbol reflects the extent of LD with the top genotyped SNP: white (r^2=0) through to dark red (r^2=1.0). Genetic recombination rates (cM/Mb), estimated using HapMap CEU samples, are shown with a light blue line. Physical positions are based on NCBI build 36 of the human genome. Also shown are the relative positions of genes and transcripts mapping to each region of association. Genes have been redrawn to show the relative positions; therefore, maps are not to physical scale.

SUPPLEMENTARY FIGURES AND TABLES

Supplementary Figure 1: Details of the quality control filters applied to each GWAS. Samples were excluded due to call rate (<95% or failed genotyping), PCA (principle components analysis or other samples reported to be not of white, European descent), IBS (any individuals found to be duplicated or related within or between data sets), SEX (sex discrepancies) or others (cases found to contain a previously reported susceptibility allele, controls with a 1st degree relative with CRC, low concordance of genotyping in duplicates or samples which have been subsequently withdrawn from a study).

Supplementary Figure 2: Quantile-Quantile (Q-Q) plots of observed and expected χ^2 values of association between SNP genotype and colorectal cancer risk. (a) UK1, (b) Scotland1, (c) UK2, (d) Scotland2, (e) VQ58 and (f) CCFR1.

Supplementary Figure 3: **Identification of individuals in the GWAS of non-European ancestry in cases and controls.** The first two principal components of the analysis are plotted. (a) UK1, (b) UK2, (c) Scotland1, (d) Scotland2, (e) VQ58, (f) CCFR1 and (g) All cases and controls. HapMap CEU individuals are plotted in blue; CHB+JPT individuals are plotted in green; YRI individuals are plotted in red; Cases are plotted as circles and controls as triangles.

Supplementary Table 1: Summary of the sample sets used in the study. The numbers shown are before stringent QC measures (Supplementary Figure 1). (References for this Table are provided in: Tomlinson, I et al. COGENT (COlorectal cancer GENeTics): an international consortium to study the role of polymorphic variation on the risk of colorectal cancer[44].

Supplementary Table 2: Relationship between rs4355419 (4q13.1), rs1321312 (6p21.2), 3824999 (11q13.4), and rs5934683 (Xp22.2) genotypes and sex, age, tumor site, family history and MSI status.

Supplementary Table 3: Details of transcription factor binding sites (TFBSs) as predicted by Transfac Matrix Database (using binding profiles

from the JASPAR2 database) and Encode ChiSeq and DNAase I data. "Score" refers to the confidence value assigned to each predicted binding region by the three different programs. For comparison, the observed and imputed SNPs and associated *P*-values are shown. The genotyped SNP with the most significant association is highlighted in yellow; imputed SNPs showing a more significant association are highlighted in red.

Supplementary Table 4: Relationship between genotype and SNP genotype in lymphoblastoid cell lines, fibroblasts, T-cells, adipocytes, colonic tissue and CRC. Box plots shown only for selected associations.

REFERENCES

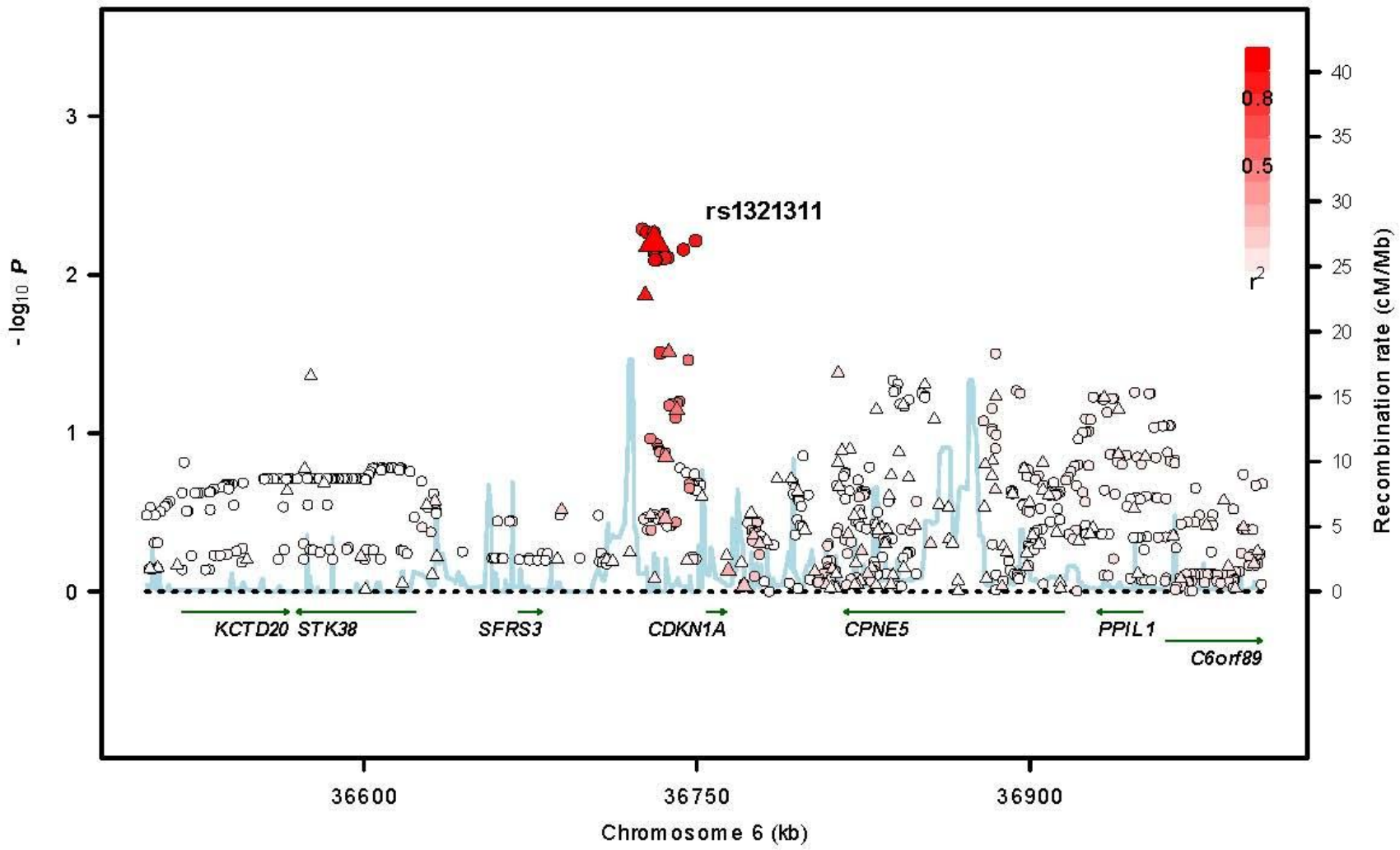
1. Lichtenstein, P. et al. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med* **343**, 78-85 (2000).
2. Aaltonen, L., Johns, L., Jarvinen, H., Mecklin, J.P. & Houlston, R. Explaining the familial colorectal cancer risk associated with mismatch repair (MMR)-deficient and MMR-stable tumors. *Clin Cancer Res* **13**, 356-61 (2007).
3. Lubbe, S.J., Webb, E.L., Chandler, I.P. & Houlston, R.S. Implications of familial colorectal cancer risk profiles and microsatellite instability status. *J Clin Oncol* **27**, 2238-44 (2009).
4. Tomlinson, I.P. et al. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat Genet* **40**, 623-30 (2008).
5. Tomlinson, I.P. et al. 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 (2011).
6. Tenesa, A. et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat Genet* **40**, 631-7 (2008).
7. Houlston, R.S. et al. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat Genet* **40**, 1426-35 (2008).
8. Houlston, R.S. et al. 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**, 973-7 (2010).
9. Broderick, P. et al. A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nat Genet* **39**, 1315-7 (2007).
10. Jaeger, E. et al. Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. *Nat Genet* **40**, 26-8 (2008).
11. Miquel, C. et al. Frequent alteration of DNA damage signalling and repair pathways in human colorectal cancers with microsatellite instability. *Oncogene* **26**, 5919-26 (2007).
12. Holm, H. et al. Several common variants modulate heart rate, PR interval and QRS duration. *Nat Genet* **42**, 117-22 (2010).
13. Abbas, T. & Dutta, A. p21 in cancer: intricate networks and multiple activities. *Nat Rev Cancer* **9**, 400-14 (2009).
14. Dunlop, M.G. et al. Cancer risk associated with germline DNA mismatch repair gene mutations. *Hum Mol Genet* **6**, 105-10 (1997).
15. Quehenberger, F., Vasen, H.F. & van Houwelingen, H.C. Risk of colorectal and endometrial cancer for carriers of mutations of the hMLH1 and hMSH2 gene: correction for ascertainment. *J Med Genet* **42**, 491-6 (2005).
16. Baglietto, L. et al. Risks of Lynch syndrome cancers for MSH6 mutation carriers. *J Natl Cancer Inst* **102**, 193-201 (2010).
17. Clayton, D.G. Testing for association on the X chromosome. *Biostatistics*, 593-600 (2008).
18. Farber, M.J., Rizaldy, R. & Hildebrand, J.D. Shroom2 regulates contractility to control endothelial morphogenesis. *Mol Biol Cell* **22**, 795-805.
19. Fairbank, P.D. et al. Shroom2 (APXL) regulates melanosome biogenesis and localization in the retinal pigment epithelium. *Development* **133**, 4109-18 (2006).

20. Houlston, R.S. et al. Congenital hypertrophy of retinal pigment epithelium in patients with colonic polyps associated with cancer family syndrome. *Clin Genet* **42**, 16-8 (1992).
21. Dunlop, M.G. et al. Extracolonic features of familial adenomatous polyposis in patients with sporadic colorectal cancer. *Br J Cancer* **74**, 1789-95 (1996).
22. Matys, V. et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* **34**, D108-10 (2006).
23. Dimas, A.S. et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* **325**, 1246-50 (2009).
24. Nica, A.C. et al. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet* **7**, e1002003 (2011).
25. Levin, J.H. & Kaler, S.G. Non-random maternal X-chromosome inactivation associated with PHACES. *Clin Genet* **72**, 345-50 (2007).
26. Kristiansen, M. et al. High incidence of skewed X chromosome inactivation in young patients with familial non-BRCA1/BRCA2 breast cancer. *J Med Genet* **42**, 877-80 (2005).
27. Midgley, R.S. et al. Phase III randomized trial assessing rofecoxib in the adjuvant setting of colorectal cancer: final results of the VICTOR trial. *J Clin Oncol* **28**, 4575-80 (2010).
28. Power, C. & Elliott, J. Cohort profile: 1958 British birth cohort (National Child Development Study). *Int J Epidemiol* **35**, 34-41 (2006).
29. Penegar, S. et al. National study of colorectal cancer genetics. *Br J Cancer* **97**, 1305-9 (2007).
30. 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**, 2331-43 (2007).
31. Adams, R., Meade, A., Wasan, H., Griffiths, G. & Maughan, T. Cetuximab therapy in first-line metastatic colorectal cancer and intermittent palliative chemotherapy: review of the COIN trial. *Expert Rev Anticancer Ther* **8**, 1237-45 (2008).
32. Cui, R. et al. Common variant in 6q26-q27 is associated with distal colon cancer in an Asian population. *Gut* **60**, 799-805 (2010).
33. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
34. Pettiti, D. Meta-analysis decision analysis and cost-effectiveness analysis. . *Oxford University Press* (1994).
35. Higgins, J.P. & Thompson, S.G. Quantifying heterogeneity in a meta-analysis. *Stat Med* **21**, 1539-58 (2002).
36. Ioannidis, J.P., Ntzani, E.E. & Trikalinos, T.A. 'Racial' differences in genetic effects for complex diseases. *Nat Genet* **36**, 1312-8 (2004).
37. Chow, J.C., Yen, Z., Ziesche, S.M. & Brown, C.J. Silencing of the mammalian X chromosome. *Annu Rev Genomics Hum Genet* **6**, 69-92 (2005).
38. Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**, 321-4 (2005).
39. Gabriel, S.B. et al. The structure of haplotype blocks in the human genome. *Science* **296**, 2225-9 (2002).
40. Ferretti, V. et al. PReMod: a database of genome-wide mammalian cis-regulatory module predictions. *Nucleic Acids Res* **35**, D122-6 (2007).
41. Stranger, B.E. et al. Genome-wide associations of gene expression variation in humans. *PLoS Genet* **1**, e78 (2005).
42. Stranger, B.E. et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848-53 (2007).

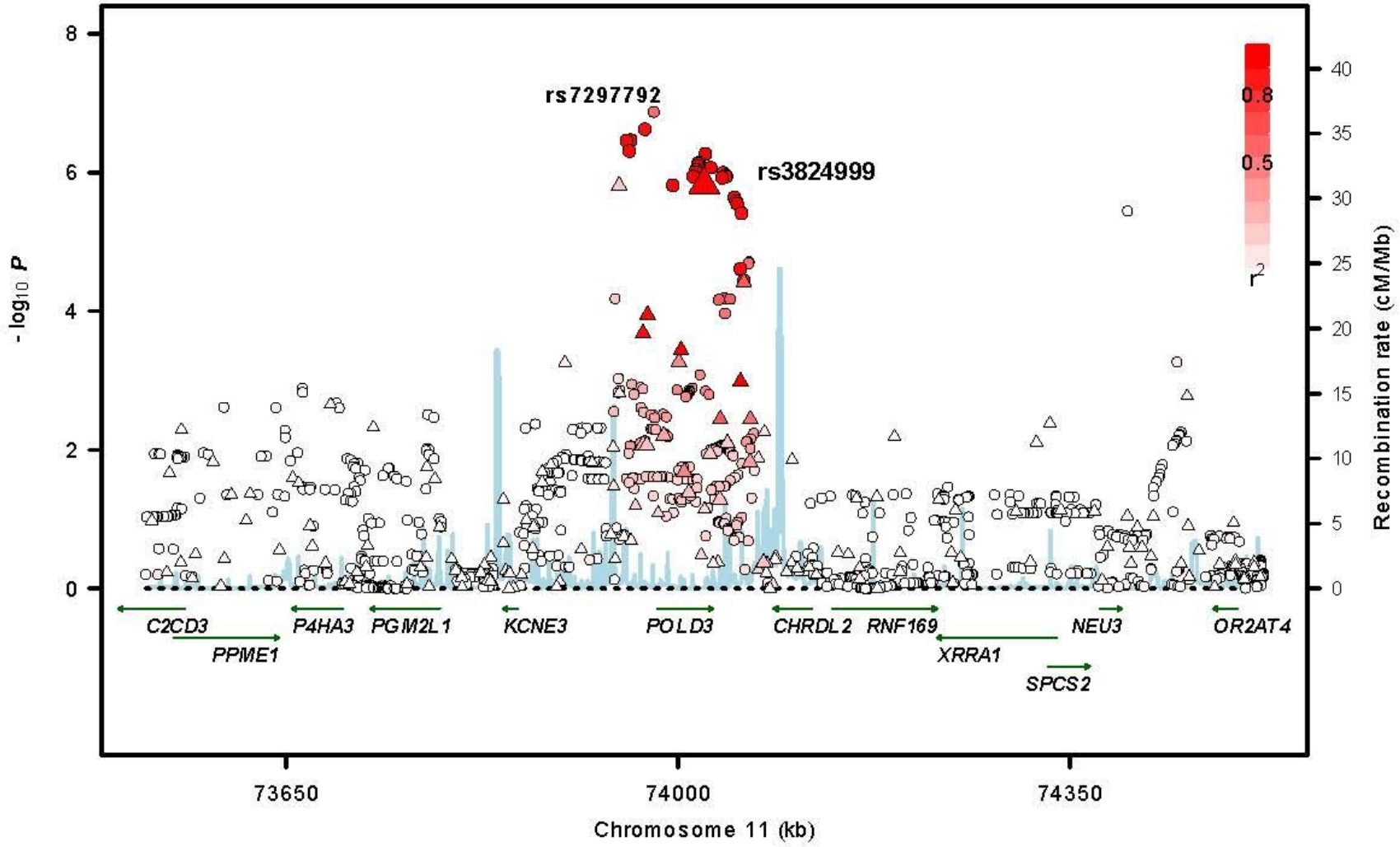
43. Boland, C.R. et al. A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. *Cancer Res* **58**, 5248-57 (1998).
44. Tomlinson, I.P. et al. COGENT (COlorectal cancer GENeTics): an international consortium to study the role of polymorphic variation on the risk of colorectal cancer. *Br J Cancer* **102**, 447-54.

Figure 1: Regional plots of association results and recombination rates for the 6p21.2, 11q13.4, Xp22.2 susceptibility loci. (a-d) Association results of both genotyped (triangles) and imputed (circles) SNPs in the GWAS samples and recombination rates within the loci: (a) 6p21.2, (b), 11q13.4**,** (c) Xp22.2. For each plot, $-\log_{10} P$ values (y axis) of the SNPs are shown according to their chromosomal positions (x axis). The top genotyped SNP in each combined analysis is a large triangle and is labelled by its rsID. The colour intensity of each symbol reflects the extent of LD with the top genotyped SNP: white (r^2=0) through to dark red (r^2=1.0). Genetic recombination rates (cM/Mb), estimated using HapMap CEU samples, are shown with a light blue line. Physical positions are based on NCBI build 36 of the human genome. Also shown are the relative positions of genes and transcripts mapping to each region of association. Genes have been redrawn to show the relative positions; therefore, maps are not to physical scale.

(a)


rs1321311
0.8
0.5
r2
- log10 P
Recombination rate (cM/Mb)
KCTD20
STK38
SFRS3
CDKN1A
CPNE5
PPIL1
C6orf89
36600
36750
36900
Chromosome 6 (kb)

(b)


rs7297792
rs3824999
0.8
0.5
r2
- log10 P
Recombination rate (cM/Mb)
C2CD3
PPME1
P4HA3
PGM2L1
KCNE3
POLD3
CHRDL2
RNF169
XRRA1
NEU3
SPCS2
OR2AT4
73650
74000
74350
Chromosome 11 (kb)

(c)

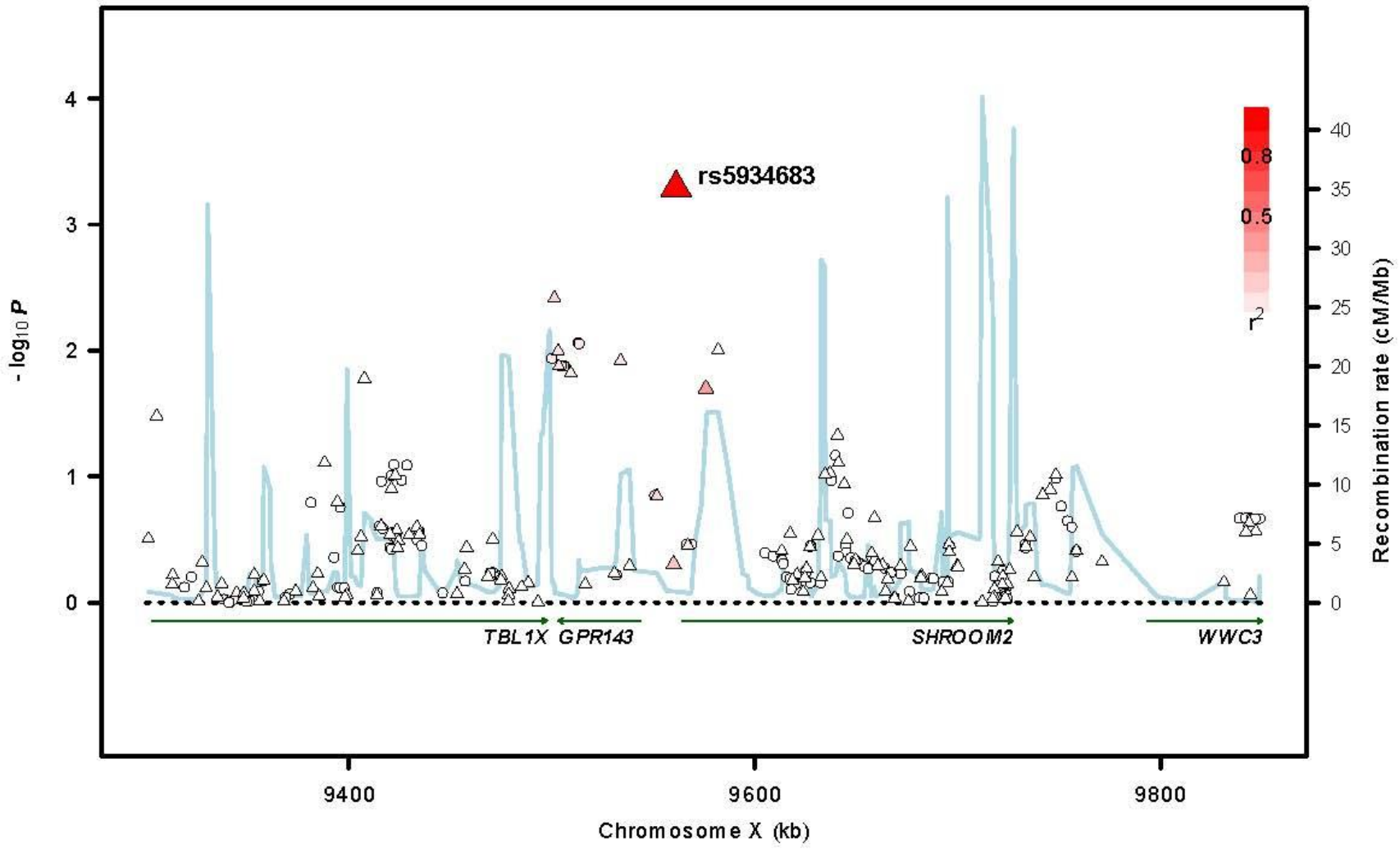

rs5934683
0.8
0.5
r2
- log10 P
Recombination rate (cM/Mb)
TBL1X
GPR143
SHROOM2
WWC3
9400
9600
9800
Chromosome X (kb)

Table 1: Summary results for the SNPS: rs1321311 (6p21.31), rs3824999 (11q13.4) and rs5934683 (Xp22.2) associated with CRC risk. [a]Risk allele frequency (RAF). [b]Odds ratio. [c]95% Confidence Interval.

rs1321311 (6p21.31)

STUDY	Cases				Controls				OR[b]	95% CI[c]	*P*-value
	RAF[a]	AA	AC	CC	RAF	AA	AC	CC			
UK1	0.26	53	363	473	0.23	49	317	534	0.83	0.71-0.97	$2.00x10^{-2}$
SCOTLAND1	0.25	65	356	552	0.22	53	331	614	0.85	0.73-0.98	$2.45x10^{-2}$
SCOTLAND2	0.24	109	741	1157	0.23	110	736	1229	0.95	0.86-1.06	$3.60x10^{-1}$
VQ58	0.24	108	634	1052	0.23	141	955	1588	0.96	0.87-1.07	$4.80x10^{-1}$
CCFR1	0.26	11	68	91	0.23	9	66	110	0.81	0.57-1.15	$2.40x10^{-1}$
COINNBS	0.25	135	810	1201	0.23	128	892	1481	0.89	0.80-0.97	$1.26x10^{-2}$
UK2/3	0.24	616	3798	5862	0.23	369	2494	4133	0.93	0.88-0.97	$3.25x10^{-3}$
UK4	0.23	29	217	343	0.24	35	264	397	1.04	0.86-1.25	$6.94x10^{-1}$
SCOTLAND3	0.25	49	267	404	0.23	82	535	905	0.88	0.76-1.02	$8.17x10^{-2}$
CAMBRIDGE	0.25	143	826	1279	0.23	116	789	1338	0.90	0.81-0.99	$2.86x10-^{2}$
CROATIA	0.27	25	156	194	0.25	67	374	594	0.86	0.71-1.04	$1.16x10^{-1}$
HELSINKI	0.21	44	334	621	0.17	20	245	550	0.79	0.67-0.93	$5.63x10^{-3}$
SWEDEN	0.22	132	1092	1900	0.21	128	872	1705	0.95	0.87-1.04	$2.61x10^{-1}$
									0.91	***0.89-0.94***	***$2.29x10^{-9}$***
JAPAN	0.14	26	390	1167	0.12	38	380	1479	0.84	0.73-0.97	$1.71x10^{-2}$
Combined									**0.91**	**0.89-0.94**	**$2.32x10^{-10}$**

rs3824999 (11q13.4)

STUDY	Cases				Controls				OR[b]	95% CI[c]	*P*-value
	RAF[a]	AA	AC	CC	RAF	AA	AC	CC			
UK1	0.53	194	443	253	0.50	231	438	231	1.14	1.00-1.30	$4.90x10^{-2}$
SCOTLAND1	0.53	216	489	268	0.50	244	516	238	1.13	1.00-1.28	$5.94x10^{-2}$
SCOTLAND2	0.50	505	998	504	0.51	486	1045	544	0.94	0.87-1.03	$1.99x10^{-1}$
VQ58	0.53	376	920	498	0.49	704	1318	653	1.19	1.09-1.30	$5.48x10^{-5}$
CCFR1	0.54	241	606	326	0.50	242	510	247	1.15	1.02-1.30	$2.45x10^{-2}$
COINNBS	0.52	519	1063	601	0.50	627	1241	633	1.07	0.99-1.16	$9.21x10^{-2}$
UK2/3	0.52	2405	5149	2717	0.50	1745	3582	1758	1.06	1.01-1.11	$8.74x10^{-3}$
UK4	0.52	133	288	156	0.50	259	540	252	1.10	0.95-1.27	$2.00x10^{-1}$
SCOTLAND3	0.52	162	373	195	0.49	385	774	361	1.13	1.00-1.29	$5.32x10^{-2}$
CAMBRIDGE	0.52	499	1121	584	0.51	523	1074	560	1.04	0.96-1.14	$3.15x10^{-1}$
CROATIA	0.53	80	188	101	0.52	239	536	276	1.05	0.88-1.24	$6.08x10^{-1}$
HELSINKI	0.48	261	460	228	0.48	235	395	196	1.02	0.90-1.17	$7.16x10^{-1}$
SWEDEN	0.53	741	1546	904	0.49	713	1509	673	1.14	1.06-1.22	$3.41x10^{-4}$
									1.08	***1.05-1.11***	***$5.89x10^{-10}$***
JAPAN	0.45	490	761	331	0.43	630	908	360	1.09	0.99-1.19	$8.46x10^{-2}$
Combined									**1.08**	**1.06-1.11**	**$1.29x10^{-10}$**

c) rs5934683 (Xp22.2)

	Cases RAF[a]		Controls RAF		OR[b]	95% CI[c]	*P*-value
	M	F	M	F			
UK1	0.37	0.38	0.30	0.31	1.23	1.10-1.38	$3.32x10^{-4}$
SCOTLAND1	0.36	0.35	0.35	0.34	1.03	0.93-1.15	$5.75x10^{-1}$
SCOTLAND2	0.38	0.35	0.34	0.35	1.07	1.00-1.15	$6.44x10^{-2}$
VQ58	0.34	0.35	0.31	0.33	1.07	0.99-1.16	$8.31x10^{-2}$
CCFR1	0.36	0.37	0.33	0.33	1.11	1.00-1.23	$4.66x10^{-2}$
NSCCG3	0.36	0.35	0.33	0.33	1.09	1.05-1.13	$1.12x10^{-5}$
UK4	0.39	0.34	0.35	0.32	1.10	0.97-1.25	$1.38x10^{-1}$
SCOTLAND3	0.33	0.33	0.33	0.35	0.98	0.88-1.09	$7.01x10^{-1}$
CAMBRIDGE	0.39	0.36	0.41	0.34	1.01	0.94-1.08	$8.23x10^{-1}$
CROATIA	0.42	0.40	0.38	0.40	1.04	0.89-1.21	$6.48x10^{-1}$
HELSINKI	0.34	0.35	0.32	0.33	1.05	0.97-1.15	$1.19x10^{-1}$
SWEDEN	0.37	0.38	0.30	0.31	1.07	1.00-1.13	$4.67x10^{-2}$
					1.07	***1.05-1.09***	***$7.43x10^{-10}$***
JAPAN	0.88	0.87	0.87	0.87	0.97	0.86-1.08	$5.38x10^{-1}$
Combined					**1.07**	**1.04-1.09**	**$3.16x10^{-9}$**

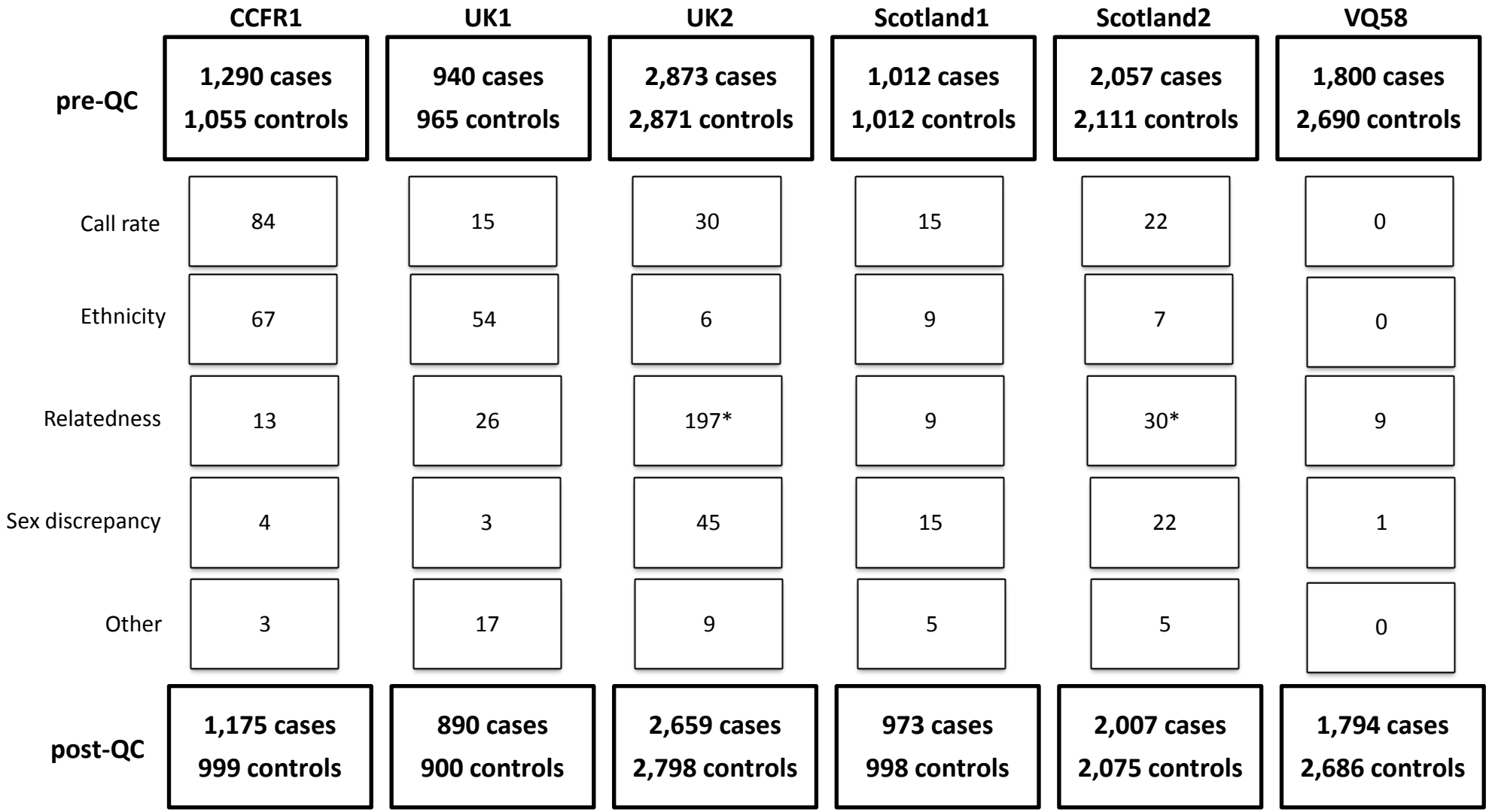
Supplementary Table 1: Summary of the sample sets used in the study.

ca/co series	Study setting	Study centre	Sampling	Genotyping platform	No. cases	No. controls
UK1 UK4	CORGI (Colorectal Tumour Gene Identification Consortium)	Oxford University	Cases, most with family history of CRC ascertained through clinical genetics centres in the UK. Spouse controls with no personal family history of CRC.	Illumina HumanHap 550 KASPar	940 621	965 1,121
VQ58	Cases: VICTOR, post treatment stage of a phase III, randomised controlled trial of rofecoxib (VIOXX) in CRC patients after potentially curative therapy. QUASAR2, multicentre study if capecitibine +/- bevacizumab as adjuvant CRC treatment. hhtp://www.octo-oxford.org.uk/alltrials. Controls: 58BC (UK 1958 Birth Cohort) http://www.b58cgene.sgul.ac.uk	Oxford University	Cases recruited as a clinical-based series and controls as population-based series.	Illumina HumanHap 300 Illumina HumanHap 370	1,800	2,690
UK2 UK3	NSCCG (National study of Colorectal Cancer). http://www.icr.ac.uk/research/research_divisions/Genetics_and_Epidemiology/index.shtml	Institute of Cancer Research	Population-based UK study. Spouse controls from NSCCG and GELCAPS (Genetic Lung Cancer Predisposition Study).	Illumina iSelect and Goldengate KASPar	2,873 10,471	2,871 7,117
Scotland1	COGS (Colorectal Cancer Genetics Susceptibility Study)	Edinburgh University	Population-based incident case series aged <55 at diagnosis. Population-bases controls	Illumina HumanHap 300 Illumina HumanHap240S	1,012	1,012
Scotland2	SOCCS (Scottish Colorectal Cancer Study)	Edinburgh University	Population-based incident case series; Scotland. Cancer free population controls.	Illumina iSelect and Goldengate	2,057	2,111
Scotland4	SOCCS3 (Edinburgh and Lothian CRC cases)	Edinburgh University	Cancer free population controls from Lothian (LBC1921 and LBC1936).	Taqman	768	1,522
COINNBS	COIN, COIN-B http://public.ukcrn.org.uk/search/ NBS UK National Blood Service Blood Donor samples) http://www.wtccc.org.uk/ccc1/participants.shtml	University of Cardiff	Multicentre study of cetuximab and other therapies in metastatic CRC. Controls were unselected UK blood donors	Illumina Goldengate	2,151	2,501
Cambridge	UKSEARCH (Studies of Epidemiology and Risk Factors in Cancer Heredity) http://www.srl.cam.ac.uk/search/Homepage/htm	Cambridge University	Population-based case-control study	Taqman	2,248	2,288
Helsinki	FCCPS (Finnish Colorectal Cancer Predisposition Study) http://research.med.helsinki.fi/gsb/aaltonen	University of Helsinki, Finland	Population-based study, south-eastern Finland	KASPar	988	864

CCFR1	CCFR (Colon Cancer Family Registry) http://epi.grants.cancer.gov/CFR/about_colon.html	University of Southern California	Recently diagnosed CRC cases reported to population complete cancer registries in the USA (Seattle Familial Colorectal Cancer Registry), Canada (Ontario Familial Cancer Registry) and Australia (Australasian Colorectal Cancer Family Study). Population based controls.	Illumina HumanHap 550	1,290	1,055
SWEDEN			Cases were ascertained from 14 different surgical clinics in Sweden between 2004 and 2006. Controls were blood donors or healthy volunteers from Uppsala.	Taqman	3,345	3,091
CROATIA			Cases recruited from hospitals in Zagreb. Population based controls.	Taqman	420	1,077
JAPAN			Cases ascertained from Cancer hospital in Tokyo. Healthy controls from Japanese biobank.	Illumina HumanHap 550	1,583	1,898

Supplementary Table 2: Relationship between rs1321312 (6p21.2), rs3824999 (11q13.4), and rs5934683 (Xp22.2) genotypes and sex, age, tumor site, family history and MSI status. *OR>1.0 indicative of predisposition to rectal disease.

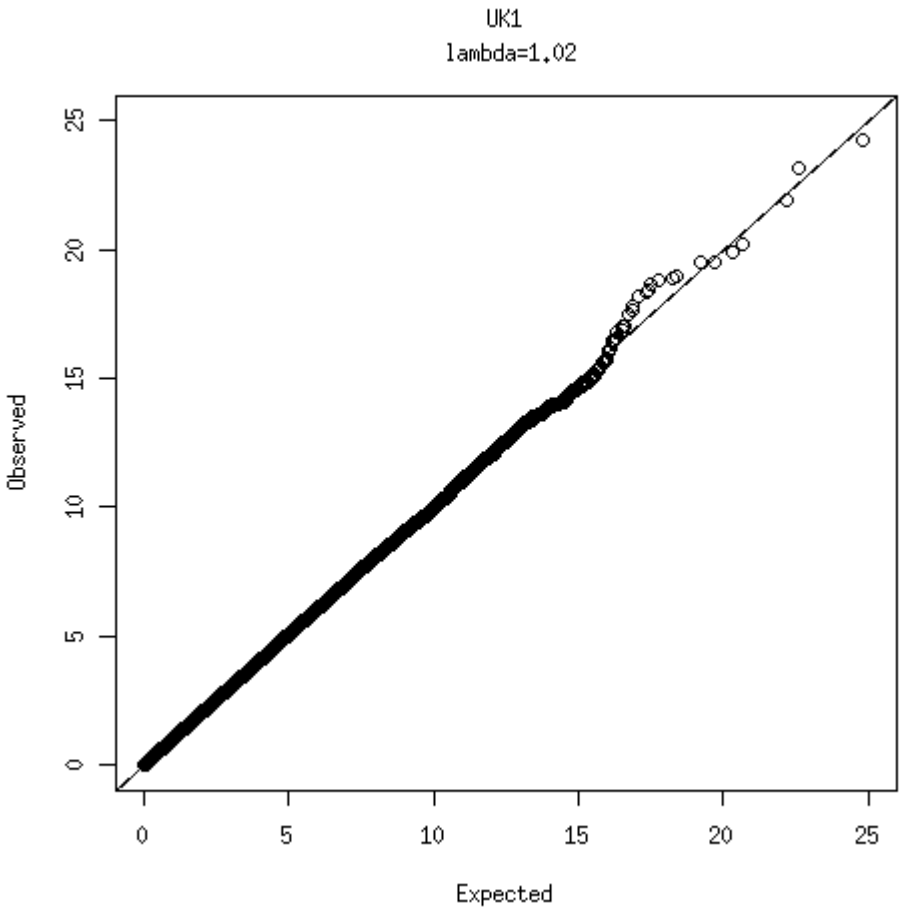
	Number of data sets	rs1321311			rs3824999			rs5934683		
		OR (95% CI)	*p*-value	Sample size	OR (95% CI)	*p*-value	Sample size	OR (95% CI)	*p*-value	Sample size
Sex	10	0.98 (0.94-1.03)	0.533	23,227	1.04 (0.99-1.10)	0.118	24,260	-	-	-
Age	10	**0.97 (0.94-1.00)**	**0.041**	22,921	0.98 (0.96-1.12)	0.399	22,953	0.99 (0.96-1.02)	0.541	19,290
Site*	7	**1.06 (1.00-1.13)**	**0.042**	19,840	0.96 (0.90-1.03)	0.244	19,886	**0.88 (0.83-0.94)**	**7.49×10^{-5}**	16,284
Family History	3	0.96 (0.87-1.05)	0.387	11,868	0.93 (0.84-1.03)	0.186	11,910	1.01 (0.91-1.10)	0.916	11,769
MSI	3	0.95 (0.77-1.17)	0.635	4,453	1.08 (0.85-1.36)	0.539	4,462	1.01 (0.80-1.26)	0.949	3,048
Stage	5	0.96 (0.90-1.02)	0.226	5,681	1.00 (0.93-1.07)	0.989	5,726	1.03 (0.97-1.10)	0.359	4,314



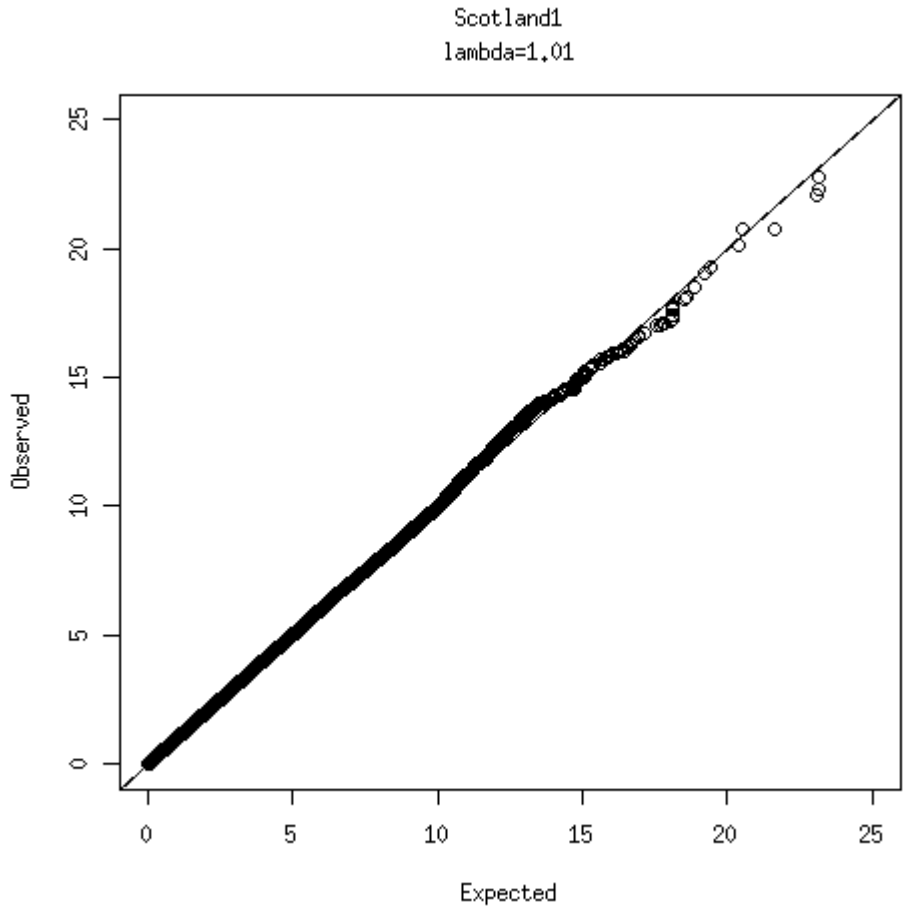
Supplementary Figure 1: Details of the quality control filters applied to each GWAS. Samples were excluded - call rate (<95%), Ethnicity (principle components analysis or self-reported not to be of European descent), relatedness (duplicates or related within or between each case-control series), sex discrepancy, other (cases found to carry a high-risk CRC mutation, controls with a 1st degree relative with CRC, low concordance of genotypes in duplicates, subjects withdrawn from the study). *samples preferentially removed from these data-sets over GWAS datasets.

Supplementary Figure 2: Quantile-Quantile (Q-Q) plots of observed and expected χ^2 values of association between SNP genotype and colorectal cancer risk. (a) UK1, (b) Scotland1, (c) UK2, (d) Scotland2, (e) VQ58 and (f) CCFR1

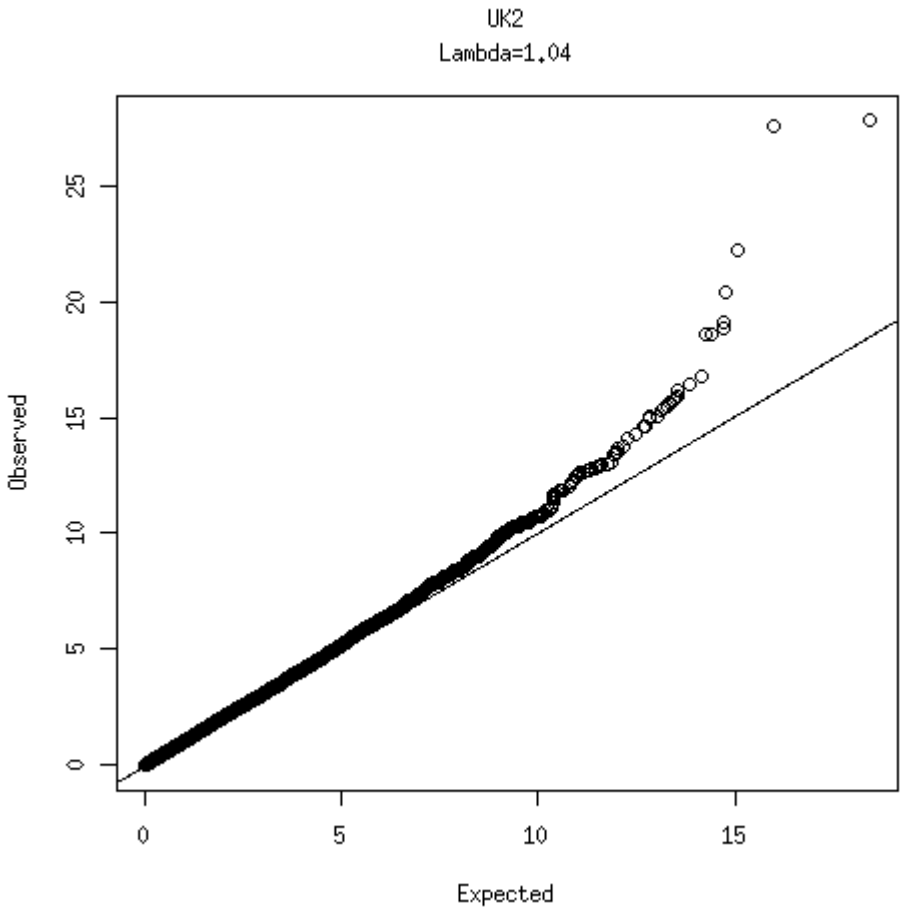
(a)



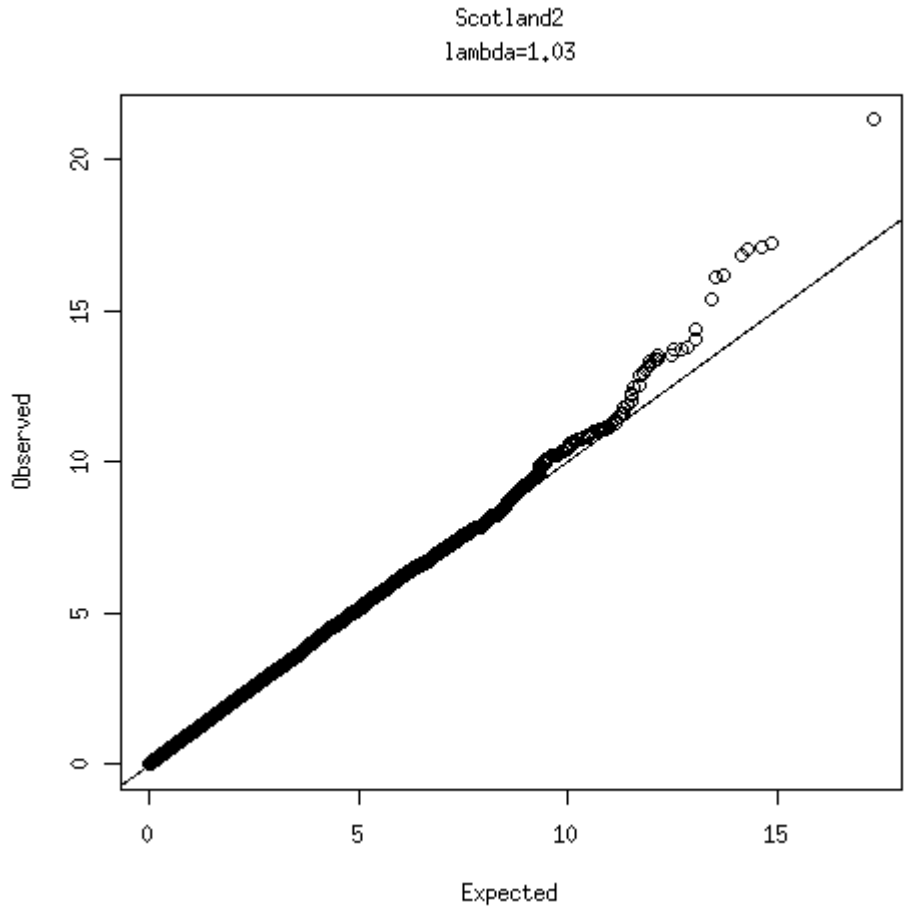
(b)



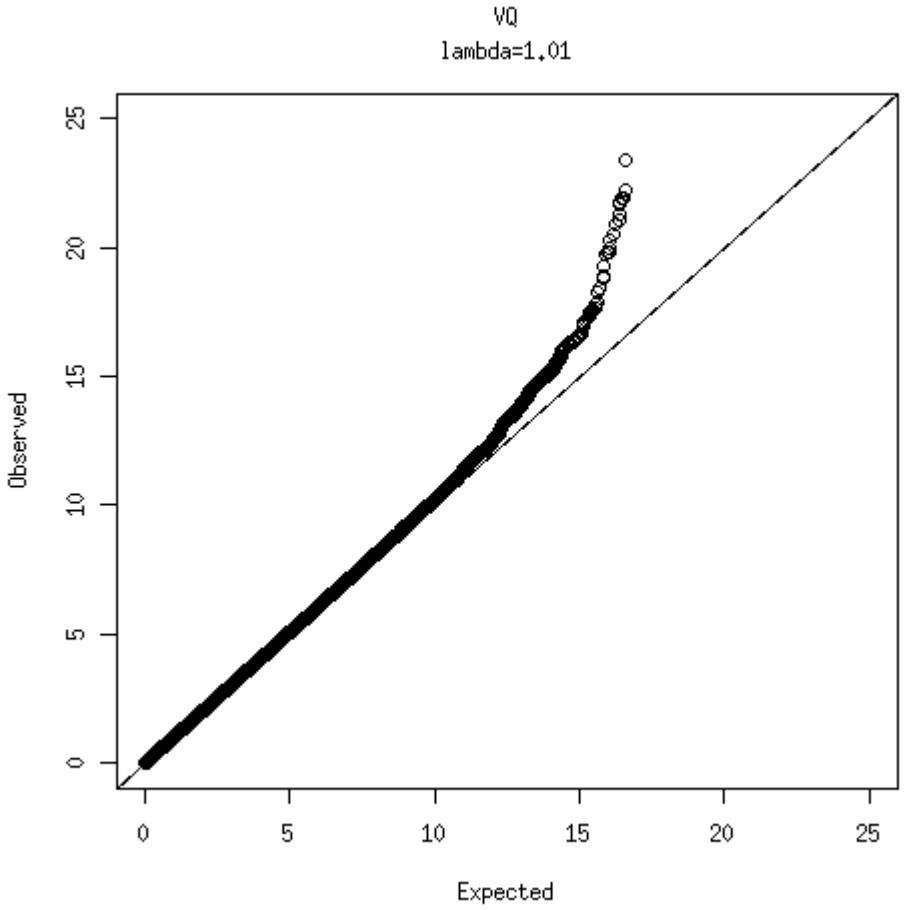
(c)


UK2
Lambda=1.04
Observed
Expected
0
5
10
15
20
25

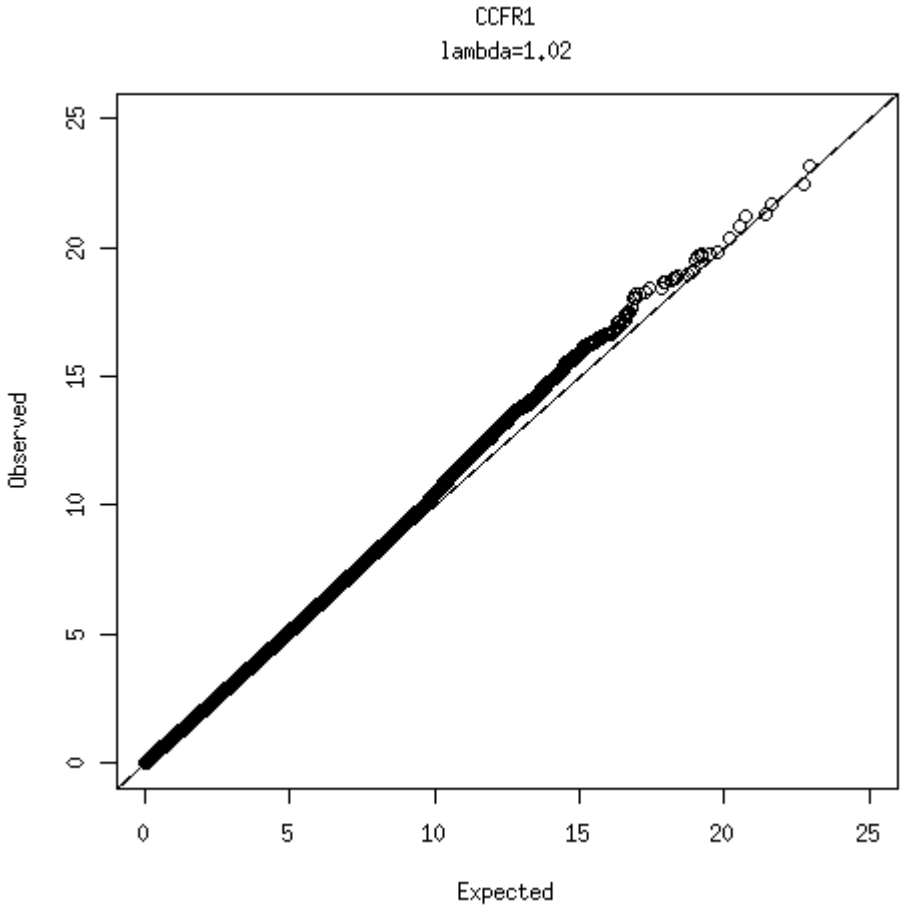
(d)


Scotland2
lambda=1.03
Observed
Expected
0
5
10
15
20

(e)

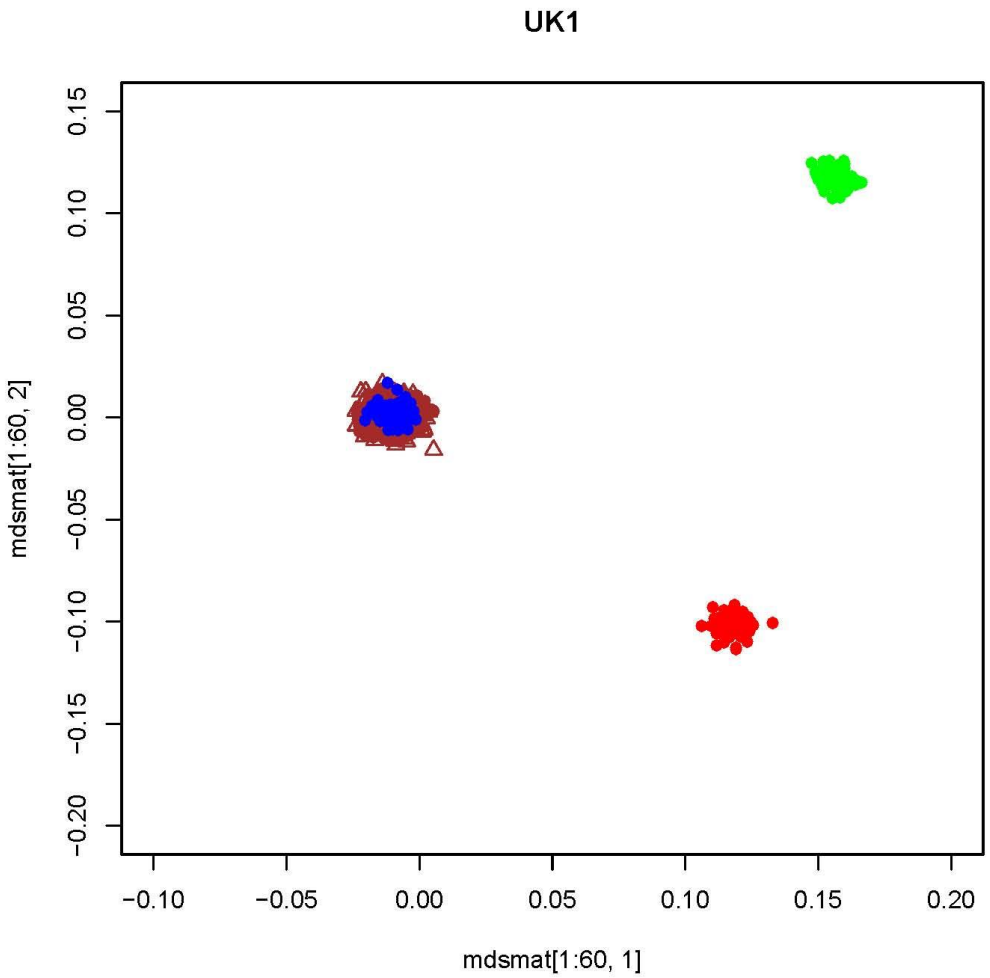

VQ
lambda=1.01
Observed
Expected
0
5
10
15
20
25

(f)

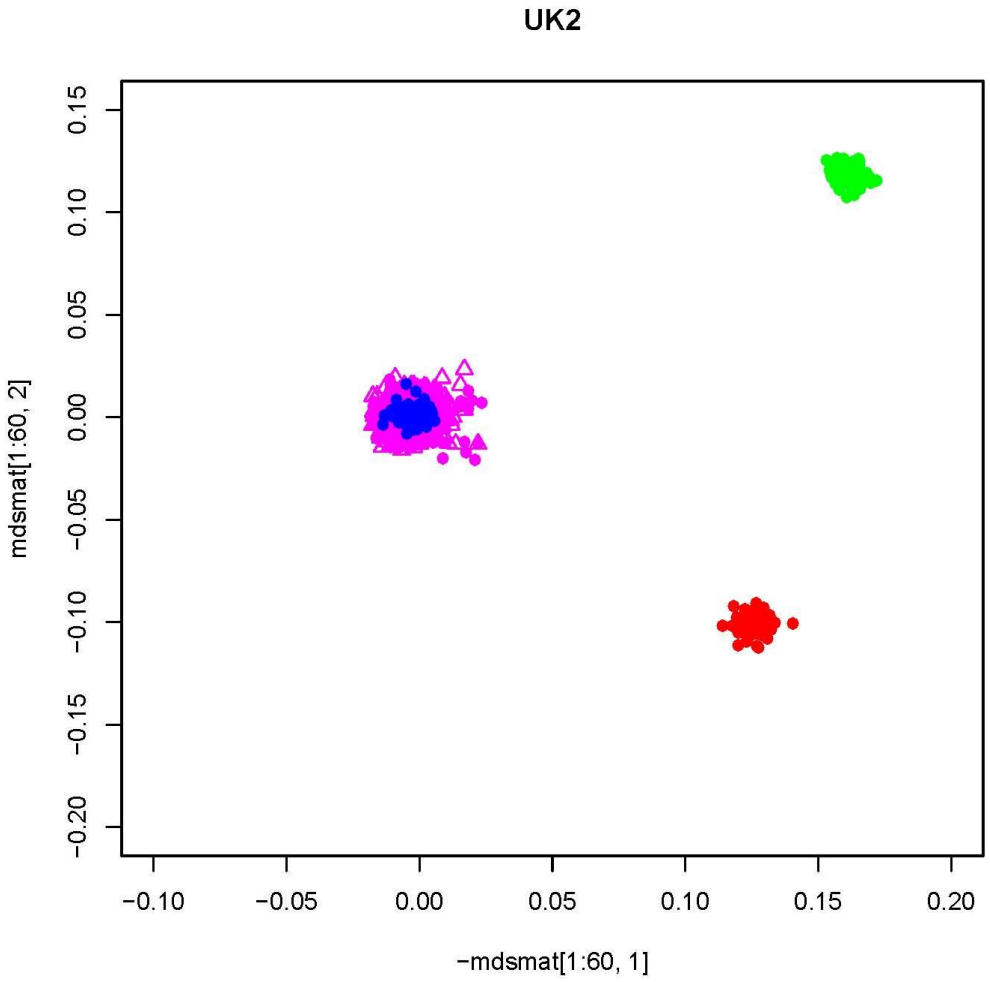

CCFR1
lambda=1.02
Observed
Expected
0
5
10
15
20
25

Supplementary Figure 3: **Identification of individuals in the GWAS of non-European ancestry in cases and controls.** The first two principal components of the analysis are plotted, before exclusions, for the (a) UK1, (b) UK2, (c) Scotland1, (d) Scotland2, (e) VQ58, (f) CCFR1 and (g) All of the above cases and controls. HapMap CEU individuals are plotted in blue; CHB+JPT individuals are plotted in green; YRI individuals are plotted in red; GWAS cases are plotted as circles and controls as triangles.

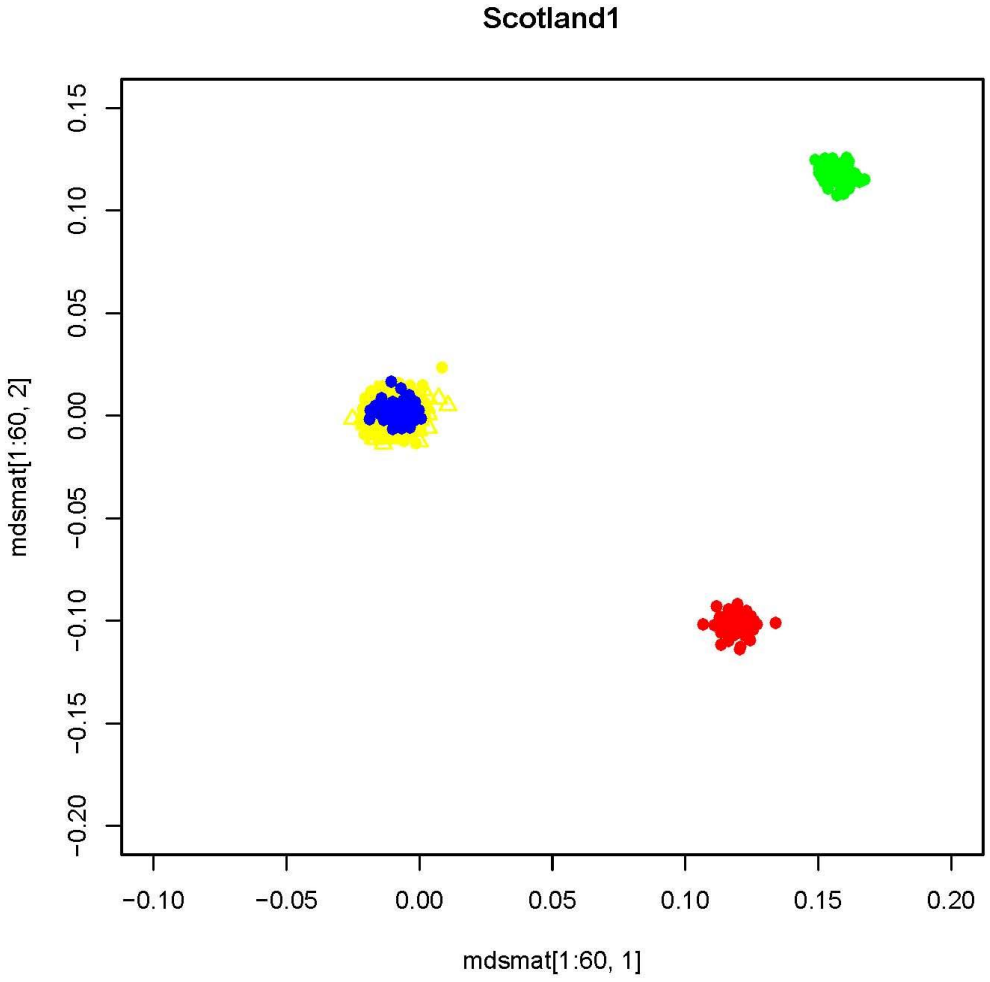
(a)



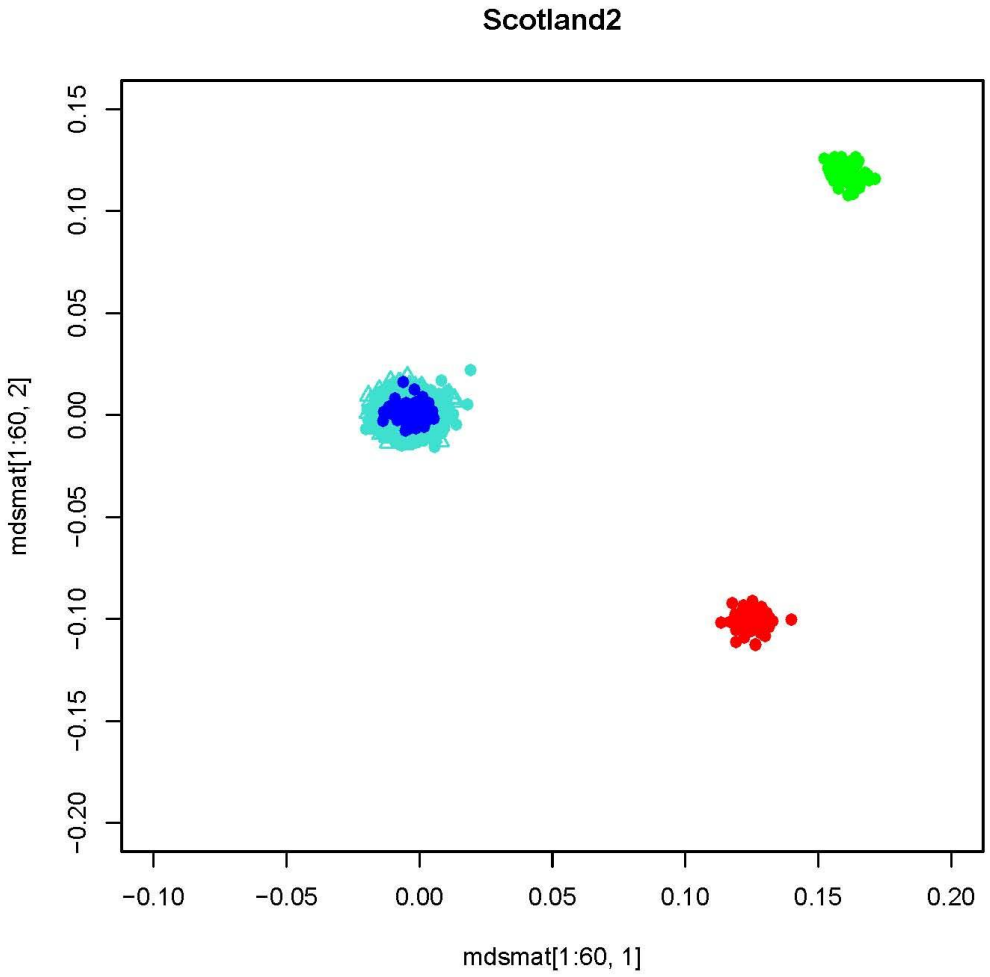
(b)


UK2
mdsmat[1:60, 2]
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20
−mdsmat[1:60, 1]

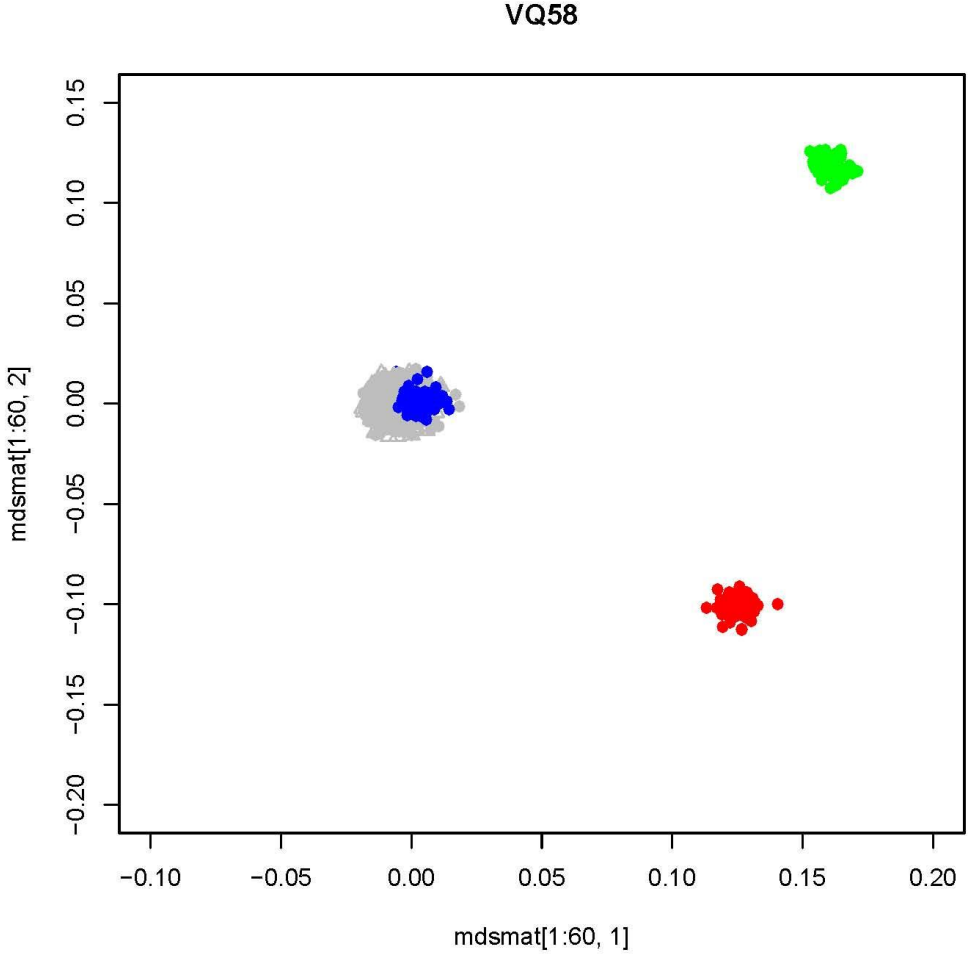
(c)


Scotland1
mdsmat[1:60, 2]
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20
mdsmat[1:60, 1]

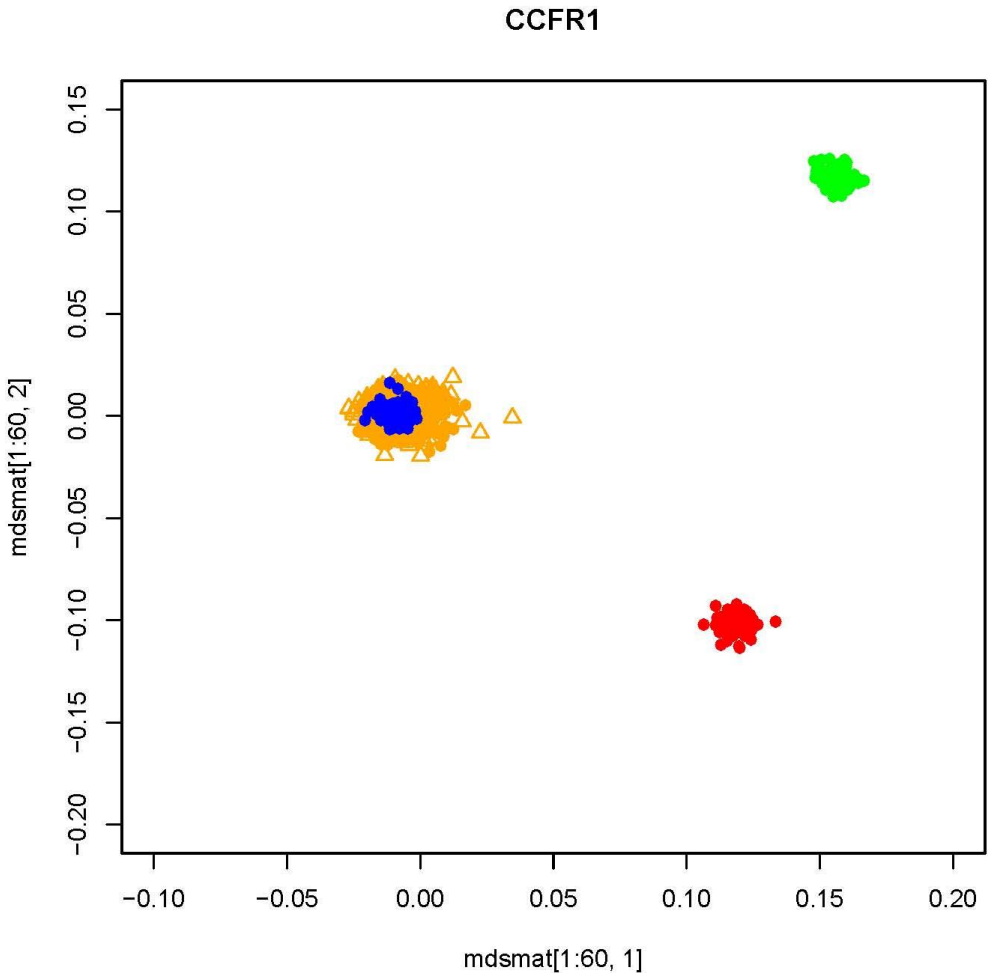
(d)


Scotland2
mdsmat[1:60, 2]
mdsmat[1:60, 1]
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20

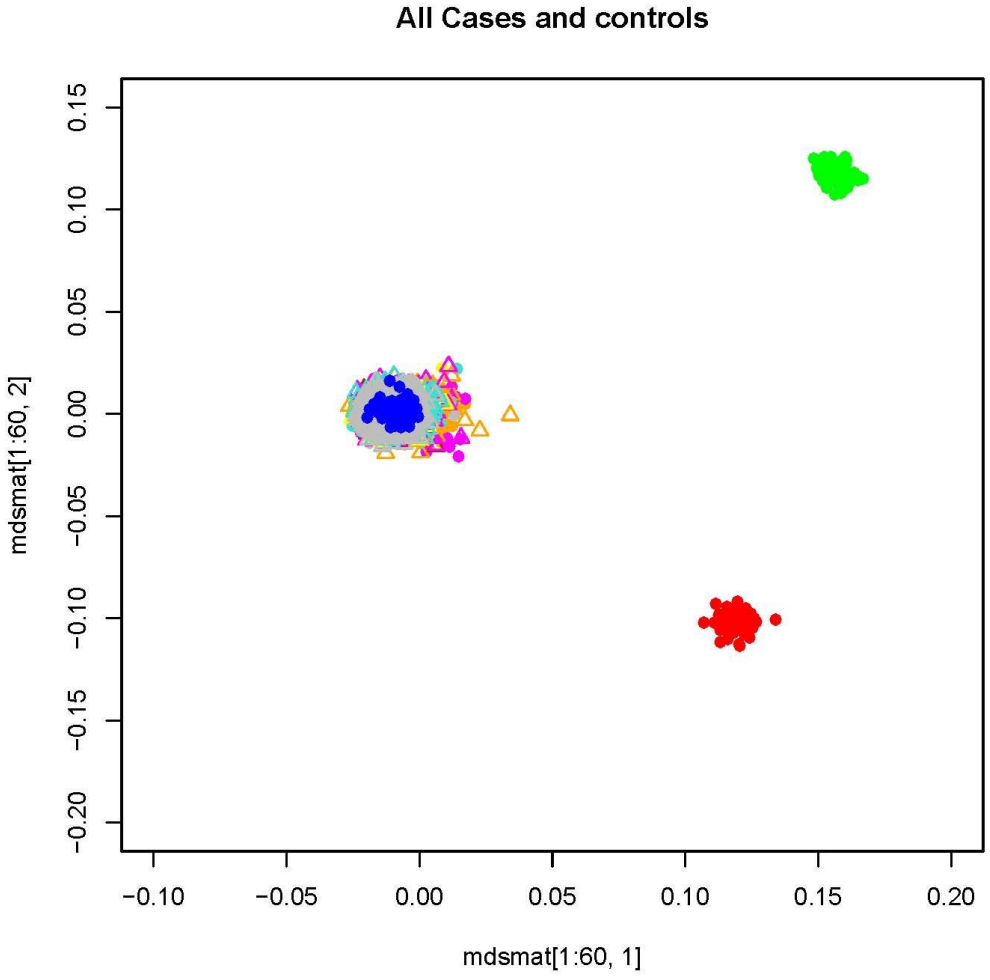
(e)


VQ58
mdsmat[1:60, 2]
mdsmat[1:60, 1]
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20

(f)


CCFR1
mdsmat[1:60, 2]
mdsmat[1:60, 1]
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20

(g)


All Cases and controls
mdsmat[1:60, 2]
mdsmat[1:60, 1]
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20