CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

ORCA – Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository:https://orca.cardiff.ac.uk/id/eprint/74685/

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Al-Tassan, Nada A., Whiffin, Nicola, Hosking, Fay J., Palles, Claire, Farrington, Susan M., Dobbins, Sara E., Harris, Rebecca, Gorman, Maggie, Tenesa, Albert, Meyer, Brian F., Wakil, Salma M., Kinnersley, Ben, Campbell, Harry, Martin, Lynn, Smith, Christopher G., Idziaszczyk, Shelley Alexis, Barclay, Ella, Maughan, Timothy Stanley, Kaplan, Richard, Kerr, Rachel, Kerr, David, Buchannan, Daniel D., Ko Win, Aung, Hopper, John, Jenkins, Mark, Lindor, Noralane M., Newcomb, Polly A., Gallinger, Steve, Conti, David, Schumacher, Fred, Casey, Graham, Dunlop, Malcolm G., Tomlinson, Ian P., Cheadle, Jeremy Peter and Houlston, Richard S. 2015. A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer. Scientific Reports 5 , 10442. 10.1038/srep10442

Publishers page: http://dx.doi.org/10.1038/srep10442

Please note:
Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.

A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer

Nada A. Al-Tassan[1†], Nicola Whiffin[2†], Fay J. Hosking[2†], Claire Palles[3], Philip J. Law[1], Susan M. Farrington[4], Sara E. Dobbins[2], Rebecca Harris[5], Maggie Gorman[3], Albert Tenesa[4,6], Brian F. Meyer[1], Salma M. Wakil[1], Ben Kinnersley[2], Harry Campbell[7], Lynn Martin[3], Christopher G. Smith[5], Shelley Idziaszczyk[5], Ella Barclay[3], Timothy S. Maughan[8], Richard Kaplan[9], Rachel Kerr[10], David Kerr[11], Daniel D. Buchannan[12,13], Aung Ko Win[13], John Hopper[13], Mark Jenkins[13], Noralane M. Lindor[14], Polly A. Newcomb[15], Steve Gallinger[16], David Conti[17], Fred Schumacher[17], Graham Casey[17], Malcolm G. Dunlop[4‡], Ian P. Tomlinson[3‡], Jeremy P. Cheadle[5‡] and Richard S Houlston[2‡*]

[1]Department of Genetics, King Faisal Specialist Hospital and Research Center, P.O.Box 3354, Riyadh11211, Saudi Arabia

[2]Division of Genetics and Epidemiology, The Institute of Cancer Research, London , UK

[3]Wellcome Trust Centre for Human Genetics and NIHR Comprehensive Biomedical Research Centre, Oxford, UK

[4]Colon Cancer Genetics Group, Institute of Genetics and Molecular Medicine, University of Edinburgh and MRC Human Genetics Unit, Western General Hospital Edinburgh, Crewe Road, Edinburgh, EH4 2XU, UK

[5]Institute of Cancer and Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff, CF14 4XN, UK

[6]The Roslin Institute, University of Edinburgh, Easter Bush, Roslin, EH25 9RG, UK

[7]Centre for Population Health Sciences, University of Edinburgh, Edinburgh, EH8 9AG, UK

[8]CRUK/MRC Oxford Institute for Radiation Oncology, University of Oxford, Roosevelt Drive, Oxford, OX3 7DQ, UK

[9]MRC Clinical Trials Unit, Aviation House, 125 Kingsway, London, WC2B 6NH, UK

[10]Oxford Cancer Centre, Department of Oncology, University of Oxford, Churchill Hospital, Old Road, Headington, Oxford, OX3 7LE, UK

[11]Nuffield Department of Clinical Laboratory Sciences, University of Oxford, John Radcliffe Hospital, Headley Way, Oxford, OX3 9DU, UK

[12]Oncogenomics Group, Genetic Epidemiology Laboratory, Department of Pathology, The University of Melbourne, Victoria, Australia

[13] Centre for Epidemiology and Biostatistics, The University of Melbourne, Victoria, Australia

[14]Department of Health Sciences Research, Mayo Clinic, Scottsdale, AZ,USA

[15]Cancer Prevention Program, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

[16]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, ON, Canada

[17]Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA

[*]To whom correspondence should be addressed. Tel: +44 2087224175; Fax: +44 2087224365; Email: richard.houlston@icr.ac.uk

[†] Co-first authorship

[‡] Co-last authorship

Genome-wide association studies (GWAS) of colorectal cancer (CRC) have identified 23 susceptibility loci thus far. Analyses of previously conducted GWAS indicate additional risk loci are yet to be discovered. To identify novel CRC susceptibility loci, we conducted a new GWAS and performed a meta-analysis with five published GWAS (totalling 7,577 cases and 9,979 controls of European ancestry), imputing genotypes utilising the 1000 Genomes Project. The combined analysis identified new, significant associations with CRC at 1p36.2 marked by rs72647484 (minor allele frequency [MAF]=0.09) near *CDC42* and *WNT4* (*P*=1.21×10^{-8}, odds ratio [OR]=1.21) and at 16q24.1 marked by rs16941835 (MAF=0.21, *P*=5.06×10^{-8}; OR=1.15) within the long non-coding RNA (lncRNA) RP11-58A18.1 and ~500kb from the nearest coding gene *FOXL1*. Additionally we identified a promising association at 10p13 with rs10904849 intronic to *CUBN* (MAF=0.32, *P*=7.01x10^{-8}; OR=1.14). These findings provide further insights into the genetic and biological basis of inherited genetic susceptibility to CRC. Additionally, our analysis further demonstrates that imputation can be used to exploit GWAS data to identify novel disease-causing variants.

INTRODUCTION

Twin studies indicate that heritable factors account for 35% of the variation in risk of developing colorectal cancer (CRC)[1]. However, only 5% of CRC can be attributed to the inheritance of high-penetrance mutations in the known genes[2,3]. Genome-wide association studies (GWAS) conducted primarily in European[4-12] but also Asian[13-16] populations have vindicated the long-held belief that part of the heritable risk of CRC is attributable to common, low-risk variants. These GWAS have provided insights into the biological basis of CRC, highlighting the role of genes within the bone morphogenetic protein signalling pathway (*BMP2, BMP4, GREM1* and *SMAD7*) [4,11] and some candidate genes (*e.g*. *CDH1/CDH3*), as well as genes not previously implicated in CRC (*e.g*. *POLD3, TERC, CDKN1A* and *SHROOM2*) [8,10].

Despite the success of GWAS the risk SNPs so far identified in European populations account for only 8% of the familial CRC risk (Supplementary Table 1). Together with the over-representation of association signals in GWAS strongly suggests that additional risk SNPs remain to be discovered. The statistical power of individual GWAS is limited by the modest effect sizes of the genetic variants and the requirement for a stringent threshold to establish statistical significance in order to avoid type 1 errors. Meta-analysis of GWAS data therefore offers the opportunity to identify new CRC risk loci and provide further insights into tumour biology. Furthermore, imputation of untyped variants in GWAS data using publicly available reference datasets increases the number of variants that can be tested for an association with CRC risk.

To identify new CRC susceptibility loci, we conducted an independent primary scan of CRC using patient samples from the COIN trial and performed a genome-wide meta-analysis with five previously published GWAS. To recover untyped genotypes, thereby maximising the prospects of identifying risk variants, we imputed over 10 million SNPs in the six GWAS datasets, using data from the 1000 Genomes Project [17] as reference (see Materials & Methods for details).

METHODS

Primary GWAS

The COIN GWAS was based on 2,244 CRC cases (64% male, mean age 61 years, SD=10) ascertained through two independent Medical Research Council clinical trials of advanced/metastatic CRC; COIN and COIN-B [18]. Cases were genotyped using Affymetrix Axiom Arrays according to the manufacturer's recommendations (Affymetrix, Santa Clara, CA 95051, USA), using duplicate samples and sequencing of significantly associated SNPs in a subset of samples to confirm genotyping accuracy. For all SNPs >99% concordant results were obtained. For controls, we made use of Wellcome Trust Case Control Consortium 2 (WTCCC2) Affymetrix 6.0 array data on 2,674 individuals from the UK Blood Service Control Group. Individuals were excluded with: <95% successfully genotyped SNPs (n = 122), discordant sex information (n = 8), classed as out of bounds by Affymetrix (n = 30), duplication or cryptic relatedness (identity by descent >0.185, n = 4), evidence of non-white European ancestry using PCA in conjunction with HapMap samples (n = 130; cut-off based on the minimum and maximum values of the top two principal components of the controls; Supplementary Figure 2). The details of all sample exclusions are provided in Supplementary Figure 3. We excluded SNPs from the analysis with: call rate <95%; different missing genotype rate between cases and controls at $P<10^{-5}$; MAF <0.01; departure from Hardy–Weinberg equilibrium in controls at $P<10^{-5}$. The adequacy of the case–control matching and the possibility of differential genotyping of cases and controls were assessed using quantile-quantile (Q–Q) plots of test statistics.

Published GWAS

We made use of five published and previously described GWAS (see Supplementary Methods):
UK1 (CORGI) [10] comprised 940 cases with colorectal neoplasia, Scotland1 (COGS) [10] included 1,012 CRC cases and 1,012 cancer-free population controls, VQ58 comprised 1800 CRC cases [19] and 2,690 population control genotypes from the WTCCC2 1958 birth cohort [20], CCFR1 comprised 1,290 familial CRC cases and 1,055 controls [21], CCFR2 included a further 796 cases and 2,236 controls from the Cancer Genetic Markers of Susceptibility (CGEMS) studies of breast and prostate cancer [22,23] .

The VQ, UK1 and Scotland1 GWA cohorts were genotyped using Illumina Hap300, Hap240S, Hap370, Hap550 or Omni2.5M arrays. 1958BC genotyping was performed as part of the WTCCC2

study on Hap1.2M-Duo Custom arrays. The CCFR samples were genotyped using Illumina Hap1M, Hap1M-Duo or Omni-express arrays. CGEMS samples were genotyped using Illumina Hap300 and Hap240 or Hap550 arrays. After applying the same quality control as that performed for COIN and COIN-B, data on 7,577 CRC cases and 9,979 controls were available for the meta-analysis (Supplementary Figure 1).

The study was conducted in accordance with the declaration of Helsinki. Written informed consent was obtained from all subjects and the study was approved by respective ethical review boards at host institutions.

Statistical and bioinformatic analysis

Analyses were undertaken using R(v3.02)[24] and PLINK[25] software. The association between each SNP and the risk of CRC was assessed by the Cochran–Armitage trend test. ORs and associated 95% CIs were calculated by unconditional logistic regression. Phasing of GWAS SNP genotypes was performed using SHAPEIT(v2.644)[26]. Prediction of the untyped SNPs was carried out using IMPUTE(v2.3.0)[27] based on the data from the 1000 Genomes Project (Phase 1 integrated variant set, v3.20101123)[28] as reference. Imputed data were analyzed using SNPTEST(v2.4.1)[29]. Association meta-analyses only included markers with info scores >0.4, imputed call rates/SNP >0.9 and MAFs >0.01. The fidelity of imputation, as assessed by the concordance between imputed and sequenced SNPs, was examined in a subset of 200 UK cases. Meta-analyses were carried out using META(v2.4-1)[30], under an inverse-weighted fixed-effects model using the genotype probabilities from IMPUTE, where a SNP was not directly typed. We calculated Cochran's *Q* statistic to test for heterogeneity and the I^2 statistic to quantify the proportion of the total variation that was caused by heterogeneity - I^2 values ≥75% are considered characteristic of large heterogeneity[31]. Associations by sex, age and clinico-pathological phenotypes were examined by logistic regression in case-only analyses. The familial relative risk of CRC attributable to a variant was calculated as detailed by Houlston *et al* [32]. The overall familial risk of CRC, as shown in epidemiological studies, is 2.2[33].

To explore epigenetic profiles of association signals, we used ChromHMM [34]. States were inferred from ENCODE Histone Modification data on the CRC cell line HCT116 (DNAse, H3K4me3, H3K4me1, H3K27ac, Pol2 and CTCF)[35] binarized using a multivariate Hidden Markov Model.

To examine whether any of the SNPs or their proxies (*i.e.* r^2>0.8 in 1000genomes CEU reference panel) annotate putative transcription factor binding/enhancer elements we used the CADD (combined annotation dependent depletion) web-server [36]. We assessed sequence conservation using: PhastCons (<0.3 indicative of conservation), Genomic Evolutionary Rate Profiling [37] (GERP) (−12 to 6, with 6 being indicative of complete conservation) and CADD (>10.0 deemed to be deleterious).

Analysis of TCGA data

To examine for a relationship between SNP genotype and mRNA expression we made use of Tumor Cancer Genome Atlas (TCGA)[38] RNA-seq expression and Affymetrix 6.0 SNP data (dbGaP accession number: phs000178.v7.p6) on 223 colorectal adenocarcinoma (COAD) and 75 rectal adenocarcinoma samples using a best proxy where SNPs were not represented directly. Association between normalised RNA counts per-gene and SNP genotype was quantified using the Kruskal-Wallis trend test. The frequency of somatic mutations in CRC was obtained using the CBioPortal for Cancer Genomics[39,40] and TumorPortal web servers[41].

Pathway analysis

To determine whether any genes mapping to the three newly identified regions act in pathways already over-represented in GWAS regions we utilized the NCI pathway interaction database[42]. All genes within the LD block containing each tagSNP, or linked to the SNP through functional experiments (MYC) were submitted as a Batch query using the NCI-Nature curated data source.

Assignment of microsatellite instability (MSI), KRAS, NRAS and BRAF status in cancers

Tumour MSI status in CRCs was determined using the mononucleotide microsatellite loci BAT25 and BAT26, which are highly sensitive MSI markers. 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].

Tumours from the COIN study were screened for mutations in *KRAS* codons 12, 13, and 61 and *BRAF* codon 600 by pyrosequencing [18]. Additionally, *KRAS* (all three codons), *BRAF* (codons 594 and 600), and *NRAS* (codons 12 and 61) were screened for mutations by MALDI-TOF mass array (Sequenom, San Diego, CA, USA)[44].

RESULTS

In the primary scan, 2,244 advanced (stage IV) CRC cases ascertained through the Medical Research Council (MRC) trials COIN [18] and COIN-B [45] were analysed with control data on 2,674 individuals from the WTCCC2 UK National Blood Service Control Group. After applying strict quality control criteria (Materials and Methods), we analysed 234,675 autosomal SNPs for association with CRC risk in 1,950 cases and 2,162 controls. A Q–Q plot of observed versus expected χ^2-test statistics showed little evidence for an inflation of test statistics, thereby excluding the possibility of substantive hidden population substructure, cryptic relatedness among subjects or differential genotype calling (inflation factor λ = 1.05; Supplementary Figure 1).

We performed a meta-analysis of our primary scan data with five non-overlapping GWAS case-control series of Northern European ancestry, which have been previously reported (Supplementary Table 2). The adequacy of the case-control matching and possibility of differential genotyping of cases and controls was assessed using Q-Q plots of test statistics. λ_{GC} values [46] for the UK1, Scotland1, VQ58, CCFR1 and CCFR2 studies were 1.02, 1.01, 1.01, 1.02 and 1.03 respectively (Supplementary Figure 1). Any ethnic outliers or individuals identified as related were excluded (Supplementary Figure 2).

After quality control procedures, the six GWAS provided data on 7,577 CRC cases and 9,979 controls. To maximise the prospects of identifying novel risk variants, we imputed over 10 million variants using 1000 Genomes Project Pilot data as a reference panel. Q-Q plots for all variants post-imputation did not show evidence of substantive over-dispersion introduced by imputation (Supplementary Figure 1).

Meta-analysis

Associations for all 23 established European CRC risk SNPs showed a direction of effect consistent with previously reported studies, with eight of the loci having a *P*-value of $<5.0\times10^{-8}$ (Supplementary Table 3; Figure 1). Additionally six SNPs previously identified in GWAS in Asian populations as determinants of CRC risk showed evidence for an association in this meta-analysis; albeit at varying degrees of significance (*P*-values ranging from $3.64\text{x}10^{-2}$ to $1.71\text{x}10^{-3}$; Supplementary Table 3); thereby providing support for trans-ethnic effects.

Excluding SNPs (including those correlated with $r^2 > 0.8$) mapping to the risk loci, five variants in distinct regions of linkage disequilibrium (LD) were associated with CRC at $P<1.0\times10^{-7}$ (Table 1; Figure 1).

We assessed the fidelity of imputation in 200 UK cases by comparing imputed genotypes with those obtained by sequencing. For the three common variants (MAF >0.05), rs72647484, rs16941835 and rs10904849 which each had imputation info scores >0.9 there was high correlation between imputed and directly typed genotype (r^2= 0.98, 1.00 and 0.99, respectively). For the rare variant rs79900961 (MAF=0.016), the correlation was poor (r^2=0.60). The call rate for the rare Indel on chromosome 5q15 (rs202110856) in the sequencing data was only 71% and both imputed heterozygotes were sequenced as homozygous reference. Therefore, only the three common variants at 1p36.12, 10p13 and 16q24.1 were subject to further analyses.

In the combined analysis of the six GWAS datasets, rs72647484, which maps to chromosome 1p36.12 (22,587,728 bps; NCBI build 37), showed the strongest evidence for association with CRC ($P=1.21\times10^{-8}$; P_{het} =0.33, I^2=14%; Figure 2a). rs72647484 maps within a 300kb block of LD encompassing *WNT4* (wingless-type mmtv integration site family, member 4; MIM 603490) and *CDC42* (cell division cycle 42, MIM 116952; Figure 3a). The second strongest association was provided by rs16941835 ($P=5.06\times10^{-8}$; P_{het} =0.40, I^2=3%; Figure 2c) which localises to the long non-coding RNA (lncRNA) RP11-58A18.1 at chromosome 16q24.1 (86,659,720bps; NCBI build 37) within a 65kb region of LD (Figure 3c). The nearest coding gene, ~500kb away, is the transcription factor *FOXL1*. The third strongest association was provided by rs10904849 ($P=7.01\times10^{-8}$; P_{het} =0.83, I^2=0%; Figure 2b) which localises to chromosome 10p13 (16,997,266bps; NCBI build 37) within intron 31 of the gene encoding cubulin (*CUBN*; alias intrinsic factor-cobalamin receptor [IFCR], MIM 602997; Figure 3b).

Bioinformatic analysis of risk variants

To gain insight into the biological basis of the associations we analysed publicly available RNA-seq expression and SNP data from TCGA on 223 colonic and 75 rectal cancers using rs10904850 and rs2744753 as proxies for rs10904849 (r^2=0.97; D'=1.00) and rs72647484 (r^2=0.64; D'=0.89) respectively. After adjustment for multiple testing, no significant associations were seen between

SNP genotype and expression of genes mapping to any of the three risk loci (Supplementary Tables 4).

We examined whether any of the SNPs or their proxies (i.e. r^2>0.8 in 1000 Genomes CEU reference panel) lie at putative transcription factor binding/enhancer elements and derived GERP and PhastCons scores to asses sequence conservation at these positions (Supplementary Table 5).

rs16941835 maps to a regulatory feature with histone modification suggestive of an enhancer element. rs10904852, in LD with rs10904849 (r^2=0.95, *D*'=1.00) is conserved (GERP and PhastCons scores of 1.20 and 0.47 respectively) with CADD score of 11.53. A moderate CADD score (8.21) was associated with rs7267484 (22,590,125bps) which is strong LD with rs72647489 (r^2=0.93, *D*'=1.00). Six proxy SNPs in LD with rs16941835 showed some evidence of transcription factor binding (Supplementary Table 5). We made use of TCGA data to examine the frequency of somatic mutation of *CDC42*, *WNT4, FOXL1* or *CUBN* in CRC. None of these genes showed evidence of significant somatic mutation. Next, we conducted pathway analysis to determine whether any genes mapping to the three newly identified regions act in pathways already over-represented in GWAS. Pathways containing three or more genes are shown in Supplementary Table 6. While this analysis identifies the BMP-signalling pathway as expected, no catalogued pathways were discernable involving genes mapping to any of the newly identified regions.

It is increasingly recognized that some genetic variants can have pleiotropic effects, influencing the risk of more than one cancer type. To explore the possibility that rs72647484, rs10904849 or rs16941835 affects the risk of other malignancies, we examined the association with lung cancer [47], acute lymphoblastic leukaemia [48], multiple myeloma [49], glioma [50] and meningioma [51] using data from previously reported GWASs. However, for these cancers, there was no evidence of rs72647484, rs10904849 or rs16941835 (or correlated SNP r^2≥0.8) being associated with tumour risk (*i.e. P* > 0.05).

Finally, the relationship between clinico-pathological variables (sex, age at diagnosis, family history of CRC, tumour stage or microsatellite instability (MSI), KRAS-mutant status and BRAF-mutant status) and genotype at rs72647484, rs10904849 and rs16941835 was assessed by case-only logistic regression (Supplementary Table 7). There was evidence of a relationship between rs72647484 and *KRAS*-mutant status (*P*=0.03) with the T risk allele associated with *KRAS*-mutant

CRC; however this finding was not significant after accounting for multiple testing. None of the other SNPs showed any association with any of the clinico-pathological variables examined (*i.e.* $P>0.05$).

DISCUSSION

We have provided evidence supporting the existence of new susceptibility loci for CRC at 1p36.12, 10p13 and 16q24.1. The 1p36.12 association implicates *WNT4* and/or *CDC42* as possible determinates of CRC risk. *WNT4* is part of a family of structurally related genes that encode cysteine-rich secreted glycoproteins that act as extracellular signalling factors. WNT4, WNT14, and WNT16 may play redundant roles in signalling through the CTNNB1-mediated canonical Wnt-pathway [52] which is known to play a central role in colorectal tumorigenesis. Additionally, WNT4 signalling appears to play a pivotal role during organogenesis, acting as an autoinducer of mesenchyme-to-epithelial transition. Inactivating germline mutations in *WNT4* cause mullerian aplasia and hyperandrogenism (MIM 158330) and are responsible for the autosomal recessive SERKAL syndrome (Sex Reversal and Kidney, Adrenal, and Lung dysgenesis; MIM 611812). *A priori* dysfunction of either *WNT4* or *CDC42* could be the biological basis for the 1p36.12 association. Cdc42 is a Ras-related GTP-binding protein with roles in establishment of cell polarity, regulation of cell morphology, motility, and cell cycle progression in mammalian cells, and malignant transformation [53]. Notably, Cdc42 regulates the actin cytoskeleton through activation of WASP proteins and cell polarity through GSK3-beta and APC. Rho-GTPase signalling has a documented role in the development of CRC [54]. Activation of Rho GTPase Cdc42 promotes adhesion and invasion in CRC [55] and targeting Cdc42 with AZA197 suppresses primary colon cancer growth and prolongs survival in a xenograft model through down regulation of *PAK1* [56].

Since rs10904849 is intronic to *CUBN* and the region of LD does not encompass any other genes or transcripts, there is a high likelihood that the functional basis of the 10p13 association is mediated through *CUBN*. Cubilin is the intestinal receptor for the endocytosis of intrinsic factor-vitamin B12 and a receptor in epithelial apoA-I/HDL metabolism [57]. Additionally cubilin is an important co-receptor in the endocytic pathway for retrieval of 25(OH)D3-DBP complexes by megalin-mediated endocytosis in the kidney [58]. Germline mutations in *CUBN* cause recessive megaloblastic anemia-1 (MGA1; MIM 261100). It is conceivable that common genetic variance in CUBN, while being insufficient to cause a "MGA type phenotype" would have physiological effects by virtue of long term effect on the cellular bioavailability of B12. Although it is entirely speculative, as epidemiological studies have yet to convincingly establish levels of B12 as a risk factor for CRC [59,60], its role in DNA biosynthesis makes genetically determined variation in B12 availability a plausible candidate for a role in the development of CRC.

LncRNAs are regulators of transcription and are increasingly recognised as playing a role in cancer biology. While there is currently no evidence to implicate the RP11-58A18.1 lncRNA in CRC, lncRNAs *CCAT1* and *CCAT2* probably do play such roles [61,62], and it is entirely plausible that the impact of variation at 16q24.1 on risk is mediated through similar long range effects.

One of the reasons for the failure to identify these CRC-loci previously is that, in addition to the issue of study power, they were not optimally tagged by SNPs featured on many commercial arrays. The power of our study to detect the major common loci conferring risks of 1.2 or greater (such as the 18q24 variant) was high. Hence, it is very unlikely there are additional CRC SNPs with similar effects for alleles with frequencies >0.2 in populations of European ancestry.

In this study, we have only considered SNPs showing evidence of an association with a stipulated *P*-value threshold of $<1x10^{-7}$. There exist, however, many variants with *P*-values just above this threshold which may also warrant investigation in a further study (Figure 1). Hence further efforts to expand the scale of GWAS meta-analyses, in terms of both sample size and SNP coverage, and to increase the number of SNPs taken forward to large-scale replication, may identify additional variants for CRC.

In conclusion, we have provided evidence for 3 new susceptibility loci for CRC. Our data also provide further evidence for the value of meta-analysis and the value of imputation as a means of enhancing the detection of novel risk loci thereby extending the utility of GWAS data.

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. 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).
5. 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).
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. Tomlinson, I. *et al.* A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet* **39**, 984-8 (2007).
8. Dunlop, M.G. *et al.* Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. *Nat Genet* **44**, 770-6 (2012).
9. Whiffin, N. *et al.* Identification of susceptibility loci for colorectal cancer in a genome-wide meta-analysis. *Hum Mol Genet* **23**, 4729-37 (2014).
10. 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).
11. 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).
12. Peters, U. *et al.* Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis. *Gastroenterology* **144**, 799-807 e24 (2013).
13. Jia, W.H. *et al.* Genome-wide association analyses in East Asians identify new susceptibility loci for colorectal cancer. *Nat Genet* **45**, 191-6 (2013).
14. Cui, R. *et al.* Common variant in 6q26-q27 is associated with distal colon cancer in an Asian population. *Gut* **60**, 799-805 (2011).
15. Zhang, B. *et al.* Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk. *Nat Genet* **46**, 533-42 (2014).
16. Wang, H. *et al.* Trans-ethnic genome-wide association study of colorectal cancer identifies a new susceptibility locus in VTI1A. *Nat Commun* **5**, 4613 (2014).
17. Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).
18. Maughan, T.S. *et al.* Addition of cetuximab to oxaliplatin-based first-line combination chemotherapy for treatment of advanced colorectal cancer: results of the randomised phase 3 MRC COIN trial. *Lancet* **377**, 2103-14 (2011).
19. 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).
20. Power, C. & Elliott, J. Cohort profile: 1958 British birth cohort (National Child Development Study). *Int J Epidemiol* **35**, 34-41 (2006).
21. 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).
22. Hunter, D.J. *et al.* A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet* **39**, 870-4 (2007).
23. Yeager, M. *et al.* Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet* **39**, 645-9 (2007).

24. 2013, R.C.T. R: A language and environment for statistical computing. . *R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/ (Date of access 01/12/2014);* (Accessed 01/12/2014).
25. 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).
26. Delaneau, O., Marchini, J. & Zagury, J.F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179-81 (2012).
27. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
28. genomes. *http://www.1000genomes.org/* (Accessed 01/12/2014).
29. 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**, 906-13 (2007).
30. Liu, J.Z. *et al.* Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet* **42**, 436-40 (2010).
31. Higgins, J.P. & Thompson, S.G. Quantifying heterogeneity in a meta-analysis. *Stat Med* **21**, 1539-58 (2002).
32. Houlston, R.S. & Ford, D. Genetics of coeliac disease. *QJM* **89**, 737-43 (1996).
33. Johns, L.E. & Houlston, R.S. A systematic review and meta-analysis of familial colorectal cancer risk. *Am J Gastroenterol* **96**, 2992-3003 (2001).
34. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28**, 817-25 (2010).
35. The ENCODE Project: ENCyclopedia Of DNA Elements. *http://www.genome.gov/encode/* (Accessed 01/12/2014).
36. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).
37. Cooper, G.M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**, 901-13 (2005).
38. The Cancer Genome Atlas *http://cancergenome.nih.gov/* (Accessed 01/12/2014).
39. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* **6**, pl1 (2013).
40. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* **2**, 401-4 (2012).
41. Lawrence, M.S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495-501 (2014).
42. database, N.p.i. Accessed 01/12/2014. *http://pid.nci.nih.gov/* (2014).
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. Smith, C.G. *et al.* Somatic profiling of the epidermal growth factor receptor pathway in tumors from patients with advanced colorectal cancer treated with chemotherapy +/- cetuximab. *Clin Cancer Res* **19**, 4104-13 (2013).
45. Wasan, H. *et al.* Intermittent chemotherapy plus either intermittent or continuous cetuximab for first-line treatment of patients with KRAS wild-type advanced colorectal cancer (COIN-B): a randomised phase 2 trial. *Lancet Oncol* **15**, 631-9 (2014).
46. Clayton, D.G. *et al.* Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet* **37**, 1243-6 (2005).
47. Broderick, P. *et al.* Deciphering the impact of common genetic variation on lung cancer risk: a genome-wide association study. *Cancer Res* **69**, 6633-41 (2009).
48. Migliorini, G. *et al.* Variation at 10p12.2 and 10p14 influences risk of childhood B-cell acute lymphoblastic leukemia and phenotype. *Blood* **122**, 3298-307 (2013).
49. Chubb, D. *et al.* Common variation at 3q26.2, 6p21.33, 17p11.2 and 22q13.1 influences multiple myeloma risk. *Nat Genet* **45**, 1221-5 (2013).

50. Sanson, M. *et al.* Chromosome 7p11.2 (EGFR) variation influences glioma risk. *Hum Mol Genet* **20**, 2897-904 (2011).
51. Dobbins, S.E. *et al.* Common variation at 10p12.31 near MLLT10 influences meningioma risk. *Nat Genet* **43**, 825-7 (2011).
52. Guo, X. *et al.* Wnt/beta-catenin signaling is sufficient and necessary for synovial joint formation. *Genes Dev* **18**, 2404-17 (2004).
53. Wu, W.J., Erickson, J.W., Lin, R. & Cerione, R.A. The gamma-subunit of the coatomer complex binds Cdc42 to mediate transformation. *Nature* **405**, 800-4 (2000).
54. Leve, F. & Morgado-Diaz, J.A. Rho GTPase signaling in the development of colorectal cancer. *J Cell Biochem* **113**, 2549-59 (2012).
55. Gao, L., Bai, L. & Nan, Q. Activation of Rho GTPase Cdc42 promotes adhesion and invasion in colorectal cancer cells. *Med Sci Monit Basic Res* **19**, 201-7 (2013).
56. Zins, K., Gunawardhana, S., Lucas, T., Abraham, D. & Aharinejad, S. Targeting Cdc42 with the small molecule drug AZA197 suppresses primary colon cancer growth and prolongs survival in a preclinical mouse xenograft model by downregulation of PAK1 activity. *J Transl Med* **11**, 295 (2013).
57. Kozyraki, R. *et al.* The human intrinsic factor-vitamin B12 receptor, cubilin: molecular characterization and chromosomal mapping of the gene to 10p within the autosomal recessive megaloblastic anemia (MGA1) region. *Blood* **91**, 3593-600 (1998).
58. Nykjaer, A. *et al.* Cubilin dysfunction causes abnormal metabolism of the steroid hormone 25(OH) vitamin D(3). *Proc Natl Acad Sci U S A* **98**, 13895-900 (2001).
59. Bassett, J.K. *et al.* Dietary intake of B vitamins and methionine and colorectal cancer risk. *Nutr Cancer* **65**, 659-67 (2013).
60. Razzak, A.A. *et al.* Associations between intake of folate and related micronutrients with molecularly defined colorectal cancer risks in the Iowa Women's Health Study. *Nutr Cancer* **64**, 899-910 (2012).
61. Zhai, H. *et al.* Clinical significance of long intergenic noncoding RNA-p21 in colorectal cancer. *Clin Colorectal Cancer* **12**, 261-6 (2013).
62. Ling, H. *et al.* CCAT2, a novel noncoding RNA mapping to 8q24, underlies metastatic progression and chromosomal instability in colon cancer. *Genome Res* **23**, 1446-61 (2013).

ACKNOWLEDGEMENTS

The COIN and COIN-B trials were funded by Cancer Research UK and the Medical Research Council and were conducted with the support of the National Institute of Health Research Cancer Research Network. COIN and COIN-B translational studies were supported by the Bobby Moore Fund from Cancer Research UK, Tenovus, the Kidani Trust, Cancer Research Wales and the National Institute for Social Care and Health Research Cancer Genetics Biomedical Research Unit (2011-2014) (all awarded to J.P.C.). We thank the patients and their families who participated in COIN and COIN-B, and gave their consent for this research, and the investigators and pathologists throughout the UK who submitted samples for assessment. N.A.A., B.F.M. and S.M.W. were funded and supported by KFSHRC. At the Institute of Cancer Research, the work was supported by Cancer Research UK (C1298/A8362 - Bobby Moore Fund for Cancer Research UK). Additional support was provided by 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. N.W. and B.K. were in receipt of PhD studentships from the ICR. B.K additionally receives funding from the Sir John Fisher Foundation. In Edinburgh the work was supported by Programme Grant funding from Cancer Research UK (C348/A12076). In Oxford additional funding was provided by the Oxford Comprehensive Biomedical Research Centre (C.P. and I.P.M.T.) and the EU FP7 CHIBCHA grant (I.P.M.T.). Core infrastructure support to the Wellcome Trust Centre for Human Genetics, Oxford was provided by grant (090532/Z/09/Z). 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). Support from the European Union (FP7/207-2013) under grant 258236 and FP7 collaborative project SYSCOL and COST Action BM1206 in the UK is also acknowledged. The work of the Colon Cancer Family Registry CFR was supported by grant UM1 CA167551 from the National Cancer Institute, National Institutes of Health and through

cooperative agreements with members of the Colon CFR and Principal Investigators. Collaborating centers include the Australasian Colorectal Cancer Family Registry (U01/U24 CA097735), the USC Colorectal Cancer Family Registry (U01/U24 CA074799), Mayo Clinic Cooperative Familial Registry for Colon Cancer Studies (U01/U24 CA074800), Ontario Registry for Studies of Familial Colorectal Cancer (U01/U24 CA074783), the Seattle Colorectal Cancer Family Registry (U01/U24 CA074794), and the University of Hawaii Colorectal Cancer Family Registry (U01/U24 CA074806). The Colon CFR GWAS was supported by funding from the National Cancer Institute, National Institutes of Health (U01 CA122839 and R01 CA143237 to GC). This study made use of genotyping data from the 1958 Birth Cohort, 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/. The results published here are in whole or part based upon data generated by The Cancer Genome Atlas pilot project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions which constitute the TCGA research network can be found at http://cancergenome.nih.gov/. 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., M.G.D. and J.P.C. The manuscript was drafted by R.S.H. with input from I.P.M.T., M.G.D. and J.P.C. All authors had access to data, analysis and had opportunity to contribute to drafting the manuscript. J.P.C. initiated and directed the GWAS of COIN and COIN-B. S.I. was responsible for COIN blood DNA extractions and quantification, R.H. for aliquoting and manifest preparation for Axiom genotyping and C.G.S. for somatic profiling of COIN tumours. N.A.A. and B.F.M. coordinated, and S.M.W. performed, the genotyping of COIN and COIN-B samples on the Axiom platform. T.S.M. was CI/co-

CI of COIN and COIN-B, and R.K. provided access to linked clinico-pathological data. Cleaning of the COIN/COIN-B genotyping data and all statistical and bioinformatic analyses were conducted by N.W and F.H, with contributions from S.D. B.K. and P.J.L., under the supervision of R.S.H. ICR - Sample preparation and genotyping were performed by A.L. and N.W. Oxford and local collaborators: subject recruitment and sample acquisition were done by E.B., M.G., L.M., R.K., D.K., and members of the CORGI Consortium. Sample preparation and genotyping were performed by C.P. Colon Cancer Genetics Group, Edinburgh and local collaborators: subject recruitment and sample acquisition were performed by S.M.F., C.H., H.C., I.D. and M.G.D., as well as members of SOCCS and COGS recruitment teams. Sample preparation was coordinated by S.M.F. Genotyping and analysis was performed and coordinated by S.M.F., C.H., M.G.D. and A.T. For the colon CFR datasets - D.D.B., A.K.W., J.H., M.J., F.S, G.C., S.G., N.L., and D.C. performed sample ascertainment and analysis.

DISCLOSURES

The authors declare no competing financial interests.

TABLE AND FIGURE LEGENDS

Figure 1: Genome-wide *P*-values ($-\log_{10}P$, *y*-axis) plotted against their respective chromosomal positions (*x*-axis). Known regions attaining genome-wide significance (*i.e*. $P=5.0\times10^{-8}$) are labelled with their chromosomal location. Variants in grey lie in novel regions that reach the significance threshold level ($P=1.0\times10^{-7}$) required for variants to be analysed further in this study. Variants in black lie in novel regions attaining genome-wide significance.

Figure 2: Forest plot of the odds ratios for the association between rs72647484, rs16941835, rs10904849 and CRC. Studies were weighted according to the inverse of the variance of the log of the OR calculated by unconditional logistic regression. *Horizontal lines:* 95% confidence intervals *(95% CI). Box:* OR point estimate; its area is proportional to the weight of the study. *Diamond (and broken line):* overall summary estimate, with confidence interval given by its width. *Unbroken vertical line:* null value (OR = 1.0).

Figure 3: Regional plot of association results and recombination rates for the (a) 1p36.12, (b) 10p13 and (c) 16q24.1 risk loci. Association results of both genotyped (triangles) and imputed (circles) SNPs in the GWAS samples and recombination rates within the loci at 1p36.12 (a), 10p13 (b) and 16q24 (c). For each plot, $-\log_{10} P$ values (*y* axis) of the SNPs are shown according to their chromosomal positions (*x* axis). The top imputed SNP in each combined analysis is shown as a large triangle and is labelled by its rsID. The colour intensity of each symbol reflects the extent of LD with the top SNP: white ($r^2 = 0$) through to dark red ($r^2 = 1.0$), with r^2 estimated from the 1000 Genomes Phase 1 data. Genetic recombination rates (cM/Mb), are shown with a light blue line. Physical positions are based on NCBI build 37 of the human genome. Also shown are the relative

positions of genes and transcripts mapping to each region of association. The lower panel shows the chromatin state segmentation track (ChromHMM).

Table 1: Summary statistics for variants showing an association with CRC risk at $P<1.0x10^{-7}$. For each variant shown along with meta-analysis test statistics are the *P*-values from the six individual studies and imputation Information scores. Risk alleles are given in bold.

							Individual study *P*-values						Meta-analysis		
Locus	**Nearest gene(s)**	**SNP**	**Position (bps)**	**Alleles**	**RAF**	**INFO**	**UK1**	**Scotland1**	**VQ58**	**CFR1**	**CFR2**	**COIN**	**OR (95% CI)**	***P***	P_{het}
1p36.12	WNT4/ CDC42	rs72647484	22,587,728	**T**C	0.91	0.94 (0.85-0.99)	$3.25x10^{-2}$	$3.32x10^{-5}$	$4.99x10^{-2}$	$4.08x10^{-2}$	$4.58x10^{-2}$	$3.47x10^{-2}$	1.24 (1.15-1.33)	$1.21x10^{-8}$	0.33
5q15	ERAP1	rs202110856	96,129,872	**G**GC	0.99	0.79 (0.66-0.92)	$2.97x10^{-1}$	$5.96x10^{-8}$	$2.81x10^{-2}$	$4.43x10^{-1}$	$3.35x10^{-1}$	$3.67x10^{-1}$	1.51 (1.23-1.86)	$6.67x10^{-8}$	0.13
10p13	CUBN	rs10904849	16,997,266	**G**T	0.68	0.98 (0.97-1.00)	$2.90x10^{-2}$	$3.39x10^{-1}$	$2.36x10^{-2}$	$8.68x10^{-3}$	$7.73x10^{-2}$	$1.29x10^{-3}$	1.13 (1.08-1.19)	$7.01x10^{-8}$	0.83
16p13.2	C16orf72	rs79900961	9,297,812	**G**A	0.98	0.70 (0.61-0.74)	$2.21x10^{-1}$	$8.68x10^{-2}$	$1.04x10^{-3}$	$2.54x10^{-2}$	$2.41x10^{-1}$	$1.02x10^{-3}$	1.49 (1.26-1.76)	$4.93x10^{-8}$	0.76
16q24.1	FOXL1	rs16941835	86,695,720	G**C**	0.21	0.97 (0.92-0.99)	$1.04x10^{-1}$	$1.17x10^{-1}$	$1.57x10^{-4}$	$3.74x10^{-3}$	$1.25x10^{-2}$	$3.65x10^{-1}$	1.16 (1.09-1.22)	$5.06x10^{-8}$	0.40

INFO, imputation Information score; P-het, P-value of heterogeneity between studies; RAF, risk allele frequency

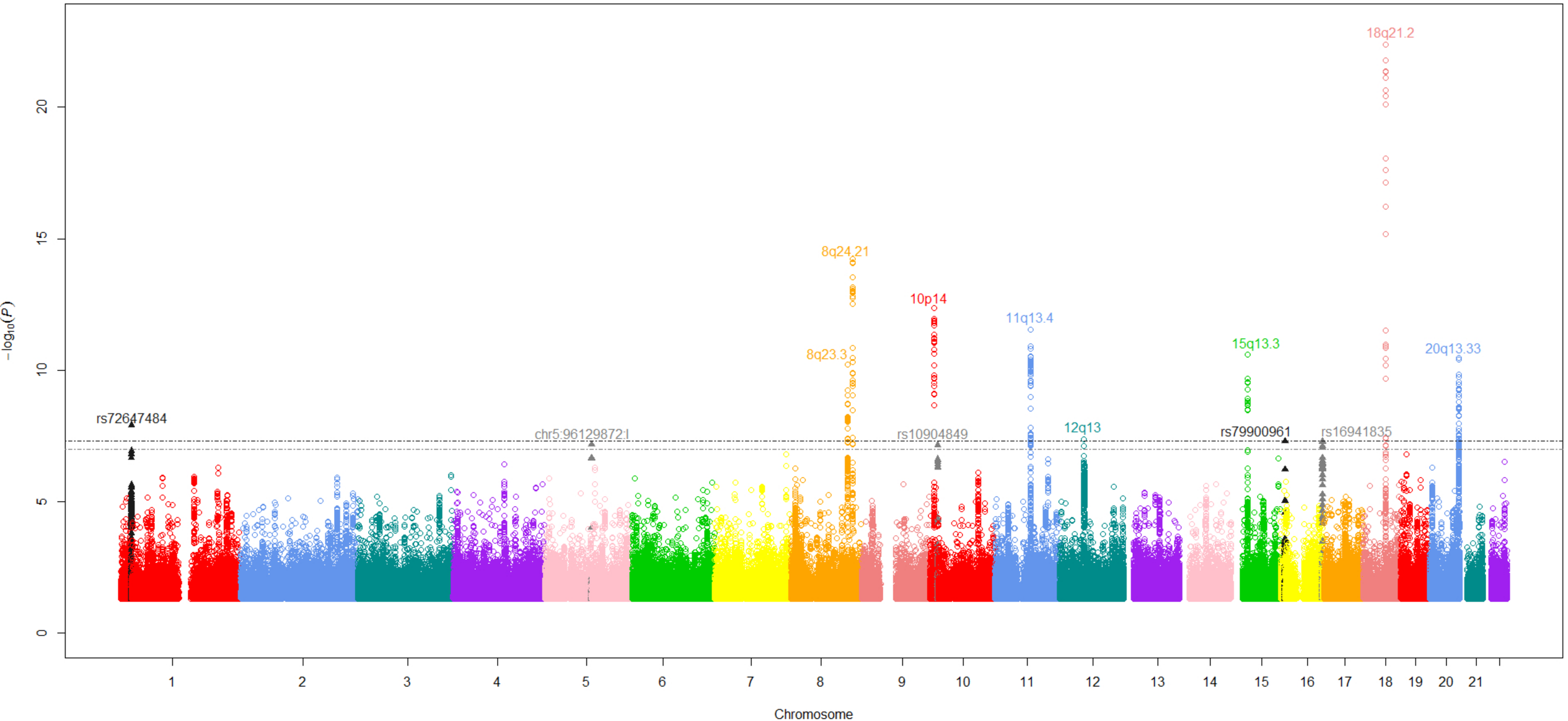
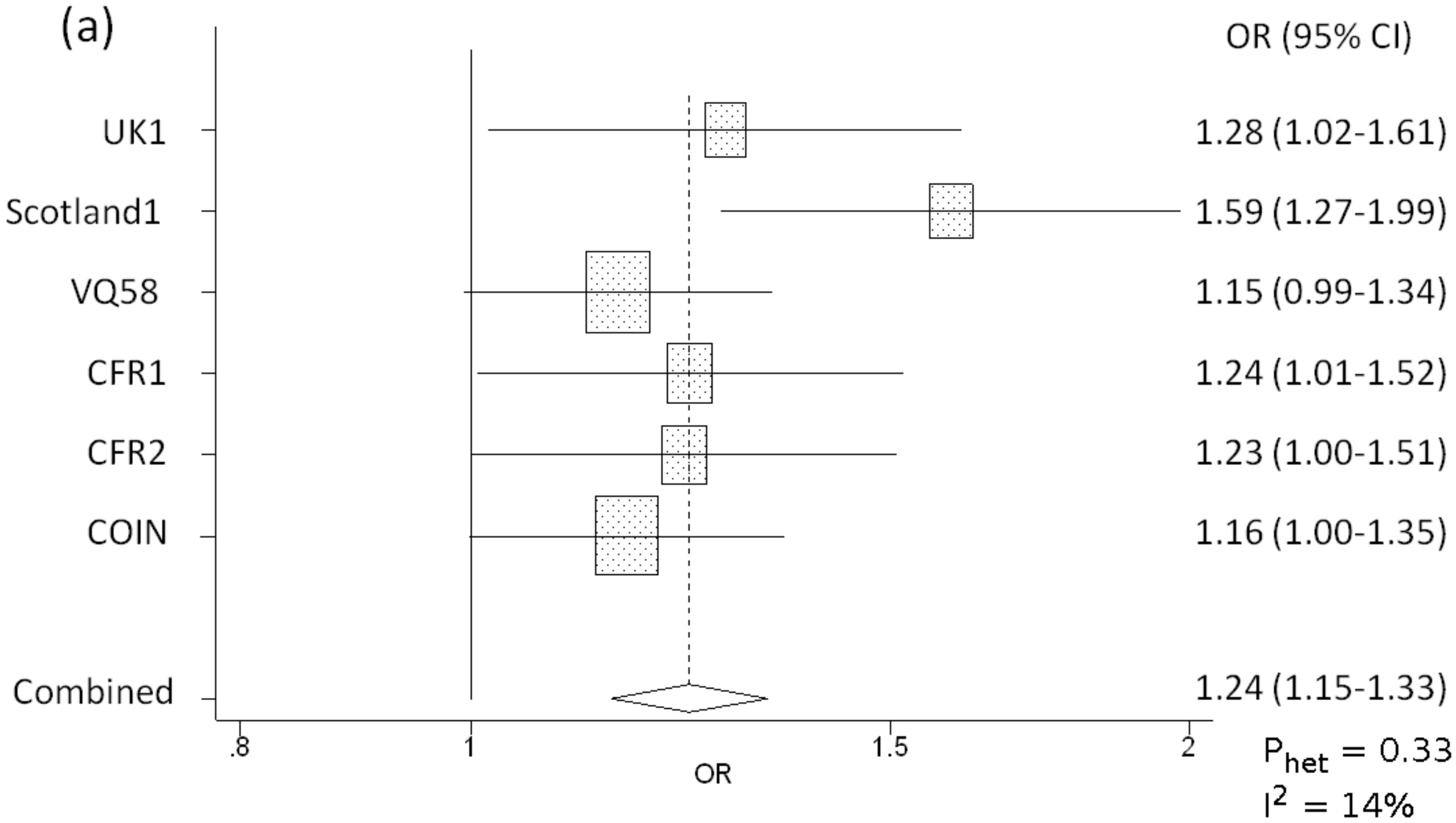
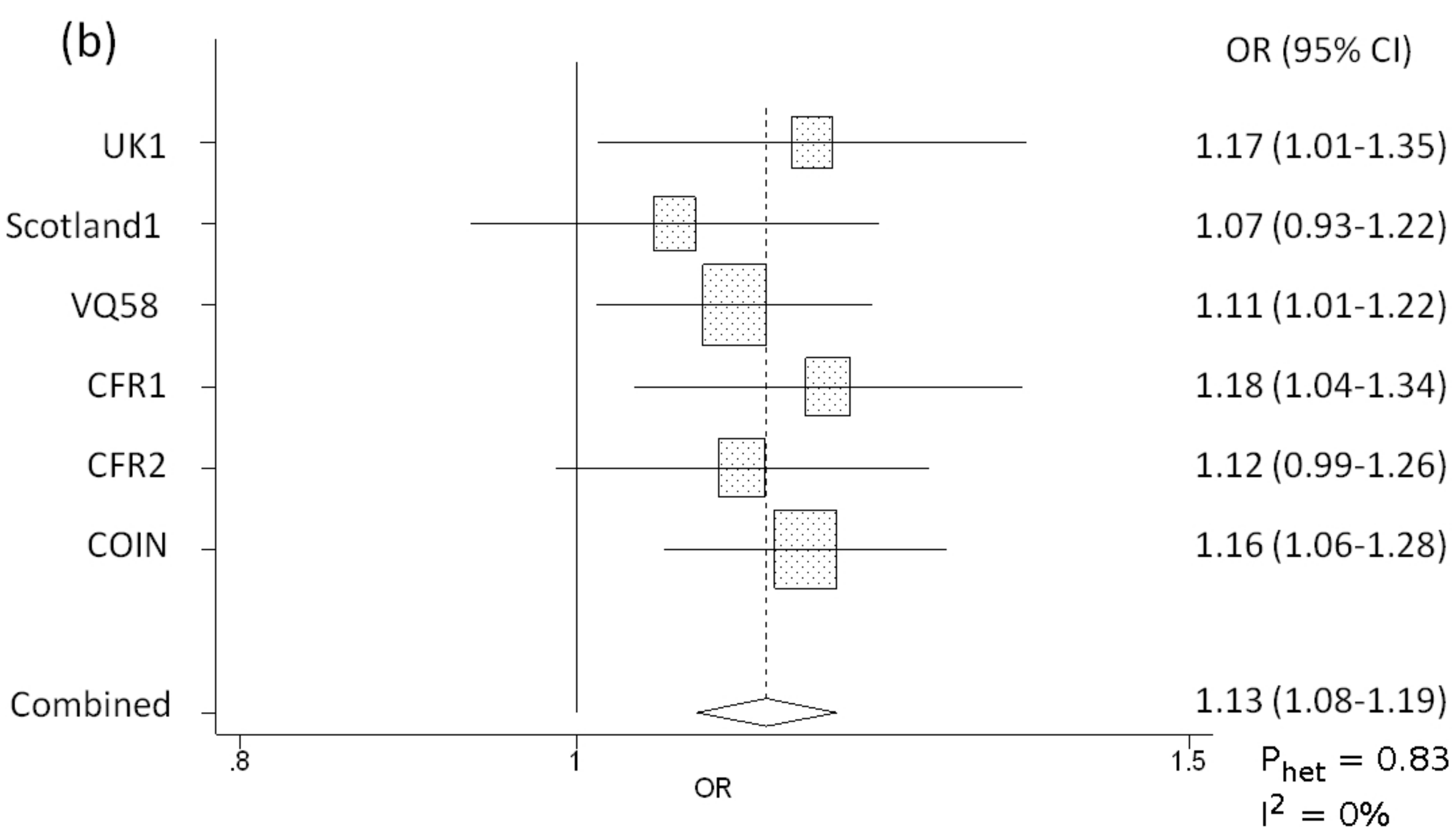

18q21.2
8q24.21
10p14
11q13.4
15q13.3
20q13.33
8q23.3
12q13
rs72647484
chr5:96129872:I
rs10904849
rs79900961
rs16941835
−log10(P)
0
5
10
15
20
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
Chromosome

Figure 2

(a)


OR (95% CI)
UK1
1.28 (1.02-1.61)
Scotland1
1.59 (1.27-1.99)
VQ58
1.15 (0.99-1.34)
CFR1
1.24 (1.01-1.52)
CFR2
1.23 (1.00-1.51)
COIN
1.16 (1.00-1.35)
Combined
1.24 (1.15-1.33)
.8
1
1.5
2
OR
P_{het} = 0.33
I^2 = 14%

(b)


OR (95% CI)
UK1
1.17 (1.01-1.35)
Scotland1
1.07 (0.93-1.22)
VQ58
1.11 (1.01-1.22)
CFR1
1.18 (1.04-1.34)
CFR2
1.12 (0.99-1.26)
COIN
1.16 (1.06-1.28)
Combined
1.13 (1.08-1.19)
.8
1
1.5
OR
P_{het} = 0.83
I^2 = 0%

(c)

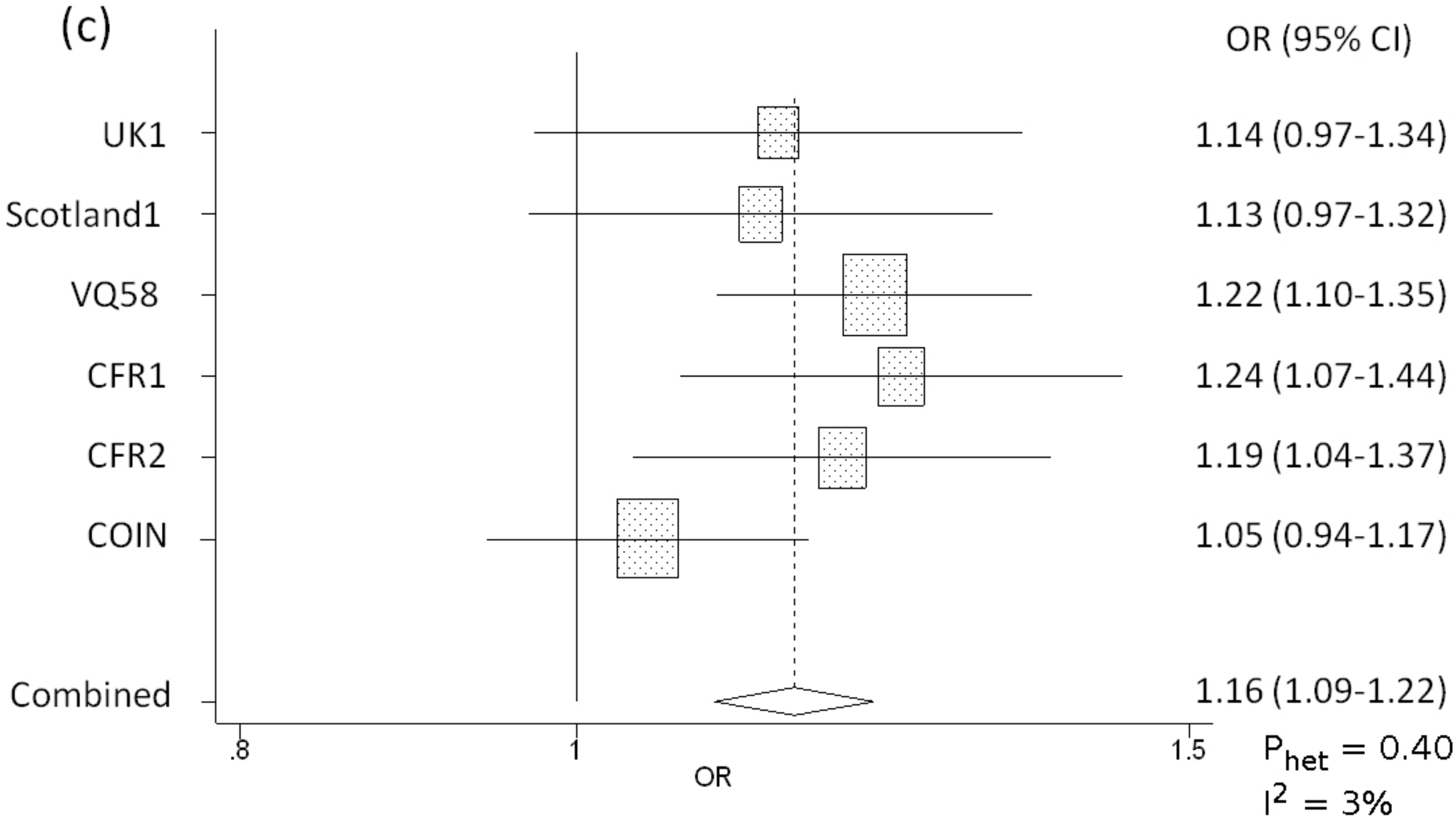
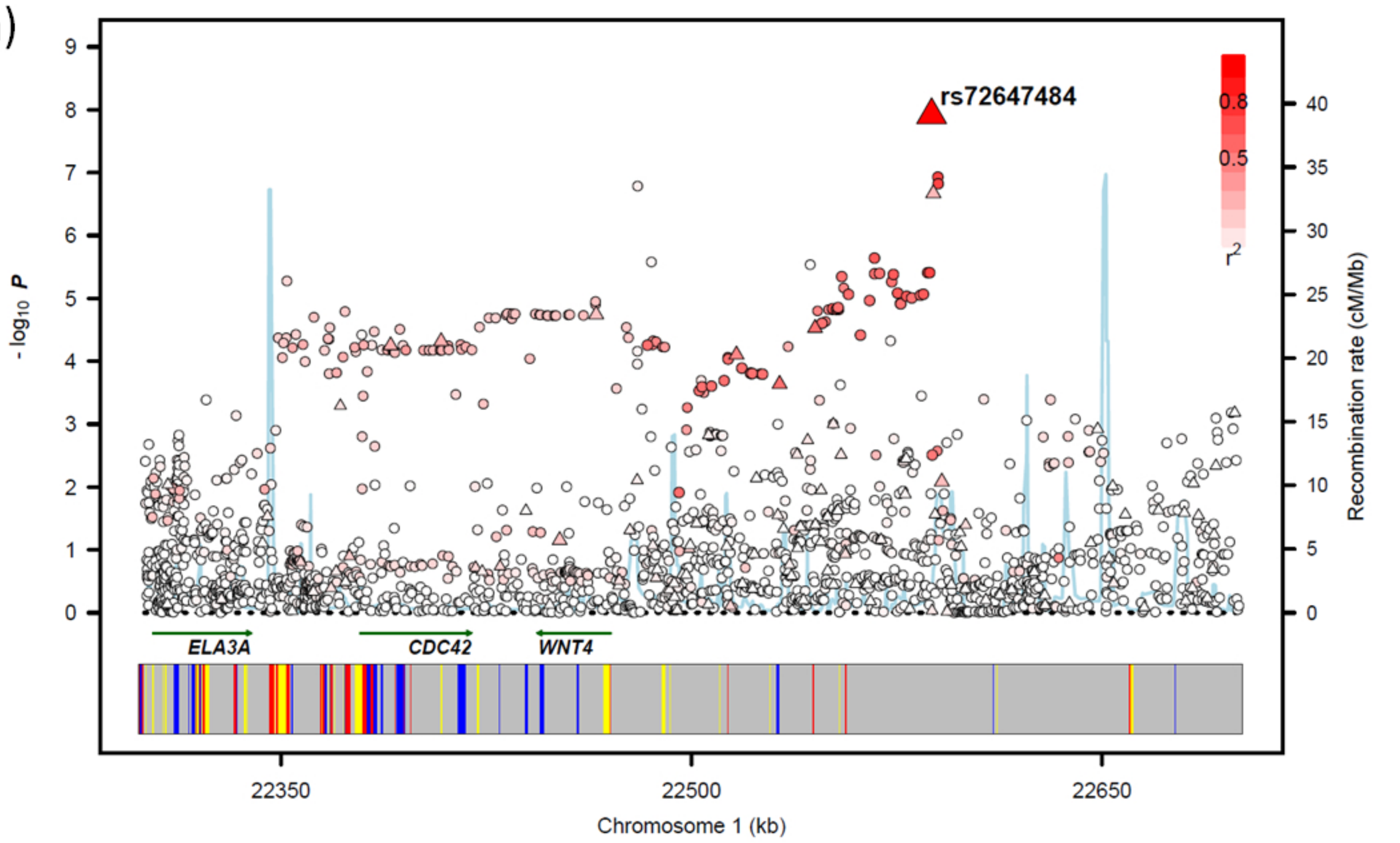
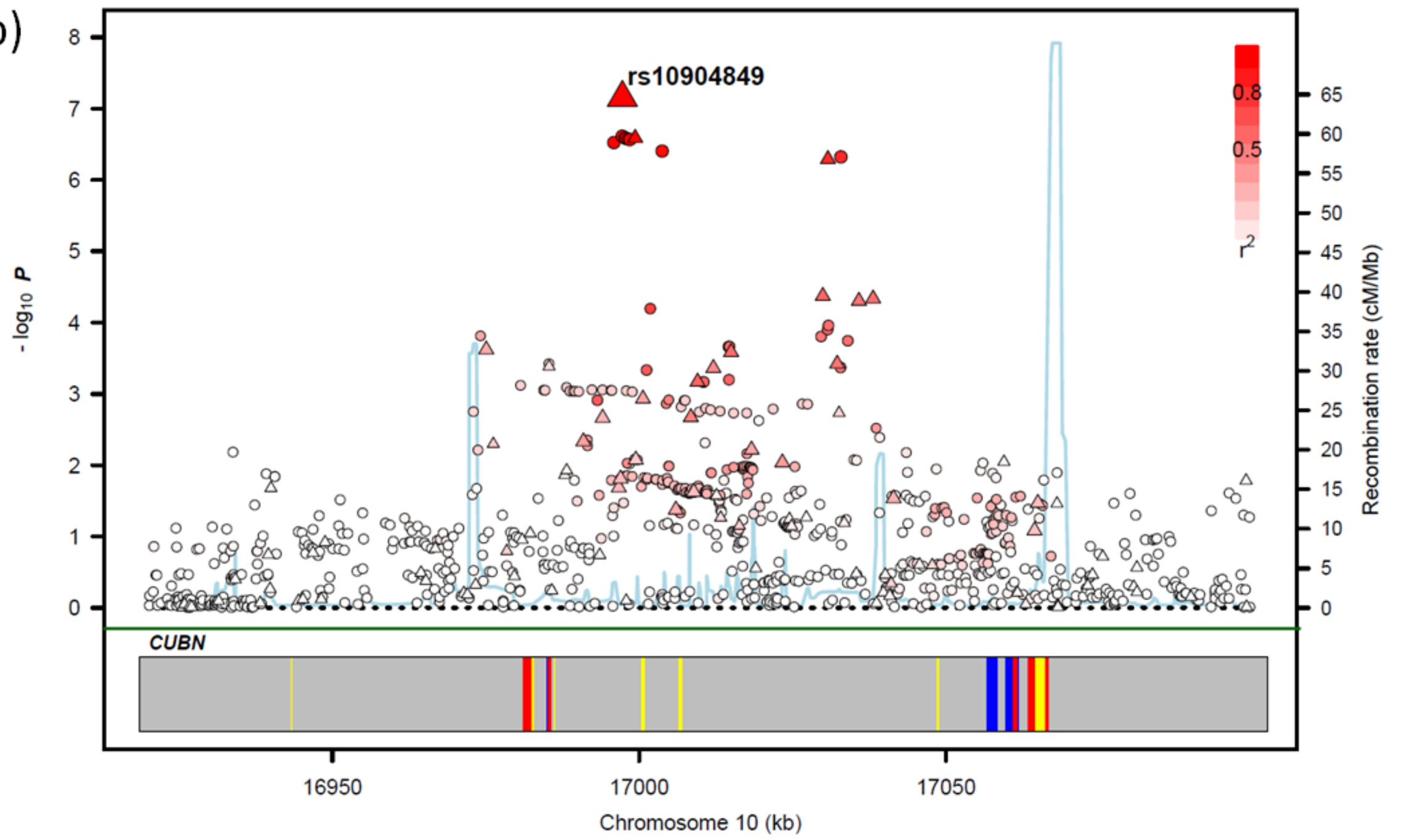

OR (95% CI)
UK1
1.14 (0.97-1.34)
Scotland1
1.13 (0.97-1.32)
VQ58
1.22 (1.10-1.35)
CFR1
1.24 (1.07-1.44)
CFR2
1.19 (1.04-1.37)
COIN
1.05 (0.94-1.17)
Combined
1.16 (1.09-1.22)
.8
1
1.5
OR
P_{het} = 0.40
I^2 = 3%

Figure 3

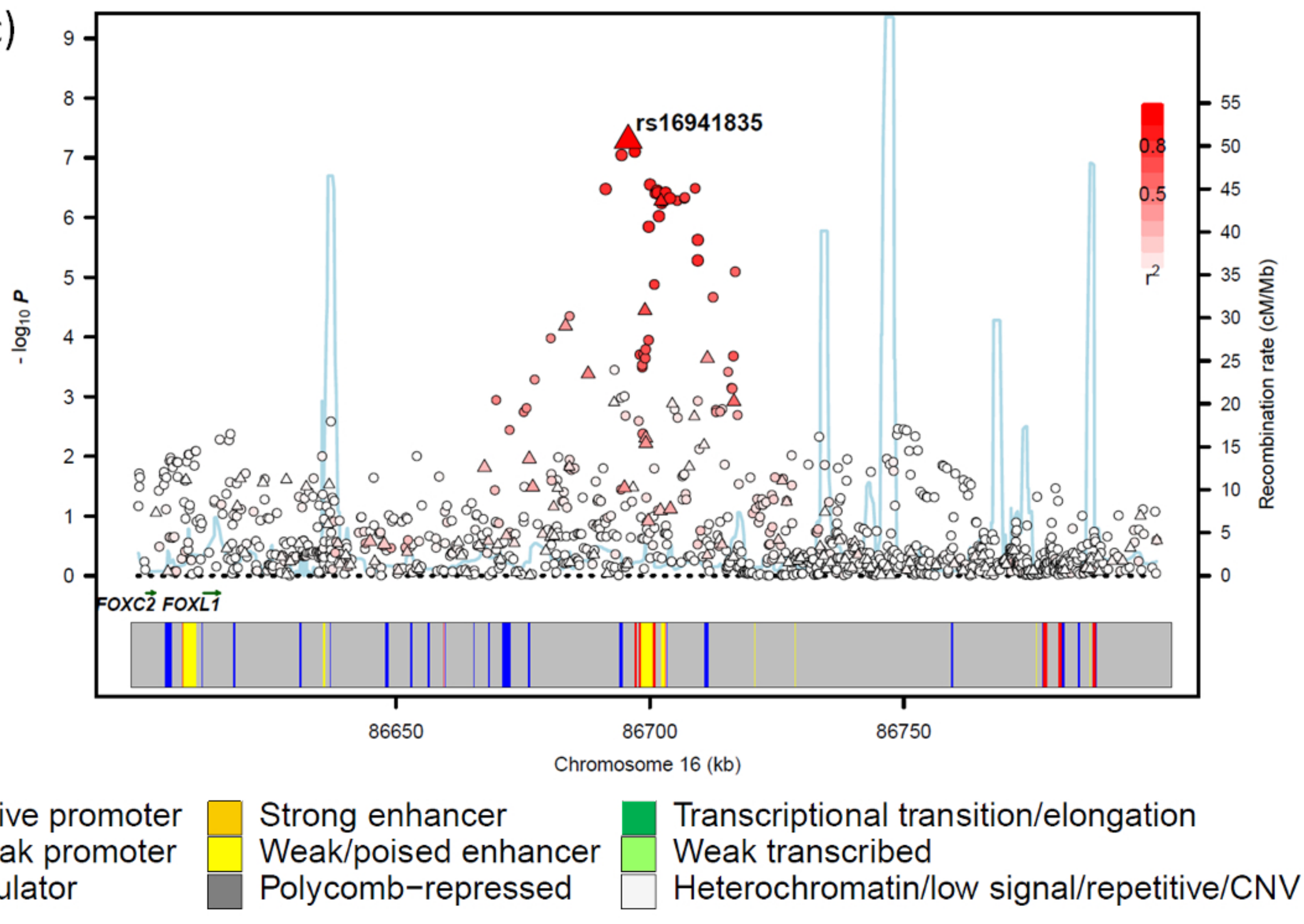
(a)



(b)



(c)



Active promoter
Weak promoter
Insulator
Strong enhancer
Weak/poised enhancer
Polycomb−repressed
Transcriptional transition/elongation
Weak transcribed
Heterochromatin/low signal/repetitive/CNV