Human Molecular Genetics, 2012, Vol. 21, No. 4 ***934–946***
doi:10.1093/hmg/ddr523
Advance Access published on November 10, 2011

Refinement of the associations between risk of colorectal cancer and polymorphisms on chromosomes 1q41 and 12q13.13

Sarah L. Spain[1,2], Luis G. Carvajal-Carmona[1], Kimberley M. Howarth[1], Angela M. Jones[1], Zhan Su[3], Jean-Baptiste Cazier[4], Jennet Williams[1], Lauri A. Aaltonen[5], Paul Pharoah[6], David J. Kerr[7], Jeremy Cheadle[8], Li Li[9], Graham Casey[10], Pavel Vodicka[11], Oliver Sieber[12], Lara Lipton[12], Peter Gibbs[12], Nicholas G. Martin[13], Grant W. Montgomery[13], Joanne Young[14], Paul N. Baird[15], Hans Morreau[16], Tom van Wezel[16], Clara Ruiz-Ponte[17], Ceres Fernandez-Rozadilla[17], Angel Carracedo[17], Antoni Castells[18], Sergi Castellvi-Bel[18], Malcolm Dunlop[19], Richard S. Houlston[20] and Ian P.M. Tomlinson[1,*]

[1]Nuffield Department of Clinical Medicine, [3]Department of Statistics and [4]Bioinformatics, Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK, [2]Division of Genetics and Molecular Medicine, King's College London, Guy's Hospital, London SE1 9RT, UK, [5]Department of Medical Genetics, Genome-Scale Biology Research Program, Biomedicum Helsinki, University of Helsinki, Helsinki, Finland, [6]Cancer Research UK Laboratories, Strangeways Research Laboratory, Department of Oncology, University of Cambridge, Cambridge CB1 8RN, UK, [7]Department of Clinical Pharmacology, Oxford University, Old Road Campus Research Building, Oxford OX3 7DQ, UK, [8]Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK, [9]Department of Family Medicine-Research Division, Case Western Reserve University, 11001 Cedar Avenue, Cleveland, OH 44106-7136, USA, [10]Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA, [11]Department of Molecular Biology of Cancer, Institute of Experimental Medicine, Academy of Science of Czech Republic, Prague 14220, Czech Republic, [12]Ludwig Colon Cancer Initiative Laboratory, Ludwig Institute for Cancer Research, Royal Melbourne Hospital, Parkville, Victoria, Australia, [13]Genetic and Molecular Epidemiology Laboratories and [14]Familial Cancer Laboratory, Queensland Institute of Medical Research, Herston Q4006, Australia, [15]Centre for Eye Research Australia, University of Melbourne, 32 Gisborne Street, East Melbourne, VIC 3002, Australia, [16]Department of Pathology, Leiden University Medical Centre, Leiden, The Netherlands, [17]Genomic Medicine Group, Fundacion Publica Galega de Medicina Xenomica, Spanish National Genotyping Center (CeGen)-USC, Centro de Investigacion Biomedica en Red de Enfermedades Raras, Hospital Clinico, Santiago de Compostela, Galicia, Spain, [18]Department of Gastroenterology, Hospital Clinic, CIBERehd, IDIBAPS, University of Barcelona, Barcelona, Catalonia, Spain, [19]Colon Cancer Genetics Group, Institute of Genetics and Molecular Medicine, University of Edinburgh and MRC Human Genetics Unit, Edinburgh EH4 2XU, UK and [20]Section of Cancer Genetics, Institute of Cancer Research, Sutton SM2 5NG, UK

Received June 16, 2011; Revised October 26, 2011; Accepted November 7, 2011

In genome-wide association studies (GWASs) of colorectal cancer, we have identified two genomic regions in which pairs of tagging-single nucleotide polymorphisms (tagSNPs) are associated with disease; these comprise chromosomes 1q41 (rs6691170, rs6687758) and 12q13.13 (rs7163702, rs11169552). We investigated these regions further, aiming to determine whether they contain more than one independent association

*To whom correspondence should be addressed. Tel: +44 1865287500; Fax: +44 1865287501; Email iant@well.ox.ac.uk

signal and/or to identify the SNPs most strongly associated with disease. Genotyping of additional sample sets at the original tagSNPs showed that, for both regions, the two tagSNPs were unlikely to identify a single haplotype on which the functional variation lay. Conversely, one of the pair of SNPs did not fully capture the association signal in each region. We therefore undertook more detailed analyses, using imputation, logistic regression, genealogical analysis using the GENECLUSTER program and haplotype analysis. In the 1q41 region, the SNP rs11118883 emerged as a strong candidate based on all these analyses, sufficient to account for the signals at both rs6691170 and rs6687758. rs11118883 lies within a region with strong evidence of transcriptional regulatory activity and has been associated with expression of *PDGFRB* mRNA. For 12q13.13, a complex situation was found: SNP rs7972465 showed stronger association than either rs11169552 or rs7136702, and GENECLUSTER found no good evidence for a two-SNP model. However, logistic regression and haplotype analyses supported a two-SNP model, in which a signal at the SNP rs706793 was added to that at rs11169552. Post-GWAS fine-mapping studies are challenging, but the use of multiple tools can assist in identifying candidate functional variants in at least some cases.

INTRODUCTION

Using genome-wide association studies (GWASs), we have identified 14 regions that contain tagging single nucleotide polymorphisms (tagSNPs) associated with the risk of colorectal cancer (CRC) (1). Within three of these regions—chromosomes 14q22.2, 15q13.3 and 20p12.3—we have shown that there exist two SNPs that are independently associated with disease (2). In two further regions—chromosomes 1q41 and 12q13.13—there are two SNPs associated with CRC risk, but from the original GWA analysis, it was unclear as to whether these represented independent signals of association (1). At 1q41, these SNPs are rs6691170 (chr1: 220,112,069 bases) and rs6687758 (chr1: 220,231,571); they are in modest pairwise linkage disequilibrium (LD) ($r^2 = 0.22$; $D' = 0.71$). At 12q13.13, the two SNPs are rs7136702 (chr12: 49,166,483) and rs11169552 (chr12: 49,441,930); these SNPs too are moderately correlated ($r^2 = 0.11$, $D' = 0.76$). Our previous analyses had not resolved the issue of whether there could be more than one independent CRC SNP in either of these regions (1).

One of the aims of GWASs is the discovery of functional/causal variants, the effects of which are manifest in the tagSNP associations. It is, however, very challenging to proceed from a tagSNP association to identifying functional variants, and relatively few such studies have been reported to date. One reason for this is that the correlation matrix between tagSNP(s) and functional variant(s) at any locus may be complex. If two association signals occur at tagSNPs at the same locus, the possible causes include the following:

(i) the associated tagSNPs are in LD;
(ii) there are two independent functional sites, each in LD with one tagSNP;
(iii) there are two functional sites, but there is true epistasis;
(iv) there is a single functional site on a haplotype defined by the two tagSNPs;
(v) there are >2 independent functional sites in LD with one or more tagSNPs;
(vi) there is a mixture of the above possibilities.

It can be extremely hard to distinguish among these possibilities and our inability to de-convolute association signals may help explain why so much of the heritability of complex diseases is unexplained by GWASs to date (3). Despite these problems, functional variant discovery may be aided by a deeper examination of genetic variation in the LD blocks in which the tagSNPs reside. Such discovery is likely to benefit from efforts such as the 1000 Genomes Project, where a comprehensive discovery of novel variants has been carried out in several populations.

In this study, we had three aims. First, we wished to investigate as fully as possible whether there was likely to be one or more than one functional variant underlying the association signals at 1q41 and 12q13.13 in CRCs. Secondly, we wanted to investigate other tagSNPs in these regions for evidence of further, independent association signals. Thirdly, we wished to use imputation and functional annotation to refine the most likely location of the 'disease-causing' variant in both the 1q41 and 12q13.13 regions.

RESULTS

The 1q41 region

We genotyped a total of 48 174 samples (22 832 cases and 25 892 controls) from 17 sample sets at rs6691170 and rs6687758. This analysis included five replication case/control cohorts that were not previously reported (1) for these SNPs: Kentucky; Prague; EPICOLON; Leiden; and Australia. After meta-analysis in STATA, both rs6691170 and rs6687758 were, as expected, significantly associated with CRC risk (Table 1), with no evidence of heterogeneity among studies. Incorporating both SNPs into an unconditional logistic regression model showed that neither of the pair of SNPs fully captured the association signal in the region [odds ratio (OR) $= 1.06$, $P = 1.06 \times 10^{-4}$ for rs6691170 and OR $= 1.07$, $P = 2.48 \times 10^{-4}$ for rs6687758]. We used PLINK to examine the possibility that the two tagSNPs indicated a single high-risk haplotype on which an unknown functional SNP was present (that is, all the functional risk alleles resided on a haplotype composed solely of one of the four possible pairs of tagSNP alleles). However, the association signal was not simply present on the high-risk haplotype TG (for rs6991170|rs6687758). Instead, the risks for the 'compound' (high-low or low-high) haplotypes—GG and TA—were

Table 1. Summary of genotyping and association results at the original four tagSNPs on 1q41 and 12q13.13 in the extended data sets

Summary	Series	Ca11	Ca12	Ca22	Co11	Co12	Co22	Ca1	Ca2	Co1	Co2	MAF ca	MAF co	OR	Ntot	Nca	Nco
rs6691170; chr1:	UK1/CORGI	130	435	355	100	429	393	695	1145	629	1215	0.38	0.34	1.172	1215	920	922
220,112,069; $z = 6.87$;	Scotland1/COGS	134	463	379	130	433	435	731	1221	693	1303	0.37	0.35	1.126	1303	976	998
$P = 6.42 \times 10^{-12}$;	UK2/NSCCG	398	1395	1058	355	1304	1159	2191	3511	2014	3622	0.38	0.36	1.122	3622	2851	2818
OR = 1.10; 95%	Scotland2/SOCCS	248	941	817	239	967	851	1437	2575	1445	2669	0.36	0.35	1.031	2669	2006	2057
CI = 1.07–1.12;	VQ58	277	833	688	359	1234	1096	1387	2209	1952	3426	0.39	0.36	1.102	3426	1798	2689
Phet = 0.39;	CFR	149	581	447	155	436	399	879	1475	746	1234	0.37	0.38	0.986	1234	1177	990
$I^2 = 5.9\%$; allele	UK3/NSCCG	406	1448	1137	367	1251	1198	2260	3722	1985	3647	0.38	0.35	1.116	3647	2991	2816
1 = T; allele 2 = G	Scotland3/SOCCS	103	376	326	117	447	363	582	1028	681	1173	0.36	0.37	0.975	1173	805	927
	UK4/CORGI2BCD	70	212	213	129	473	445	352	638	731	1363	0.36	0.35	1.029	1363	495	1047
	Cambridge	324	1068	805	280	1013	890	1716	2678	1573	2793	0.39	0.36	1.138	2793	2197	2183
	COIN/NBS	300	1054	797	326	1170	1005	1654	2648	1822	3180	0.38	0.36	1.090	3180	2151	2501
	Helsinki	143	435	351	105	372	340	721	1137	582	1052	0.39	0.36	1.146	1052	929	817
	Prague	147	424	363	82	317	252	718	1150	481	821	0.38	0.37	1.066	821	934	651
	Kentucky	156	466	388	244	709	630	778	1242	1197	1969	0.39	0.38	1.030	1969	1010	1583
	EPICOLON	193	613	520	184	632	578	999	1653	1000	1788	0.38	0.36	1.081	1788	1326	1394
	Australia	64	223	129	58	212	168	351	481	328	548	0.42	0.37	1.219	548	416	438
	Leiden	141	404	310	92	291	304	686	1024	475	899	0.40	0.35	1.268	899	855	687
rs6687758; chr1:	UK1/CORGI	37	312	568	32	299	598	386	1448	363	1495	0.21	0.20	1.098	1495	917	929
220,231,571; $z = 5.64$;	Scotland1/COGS	63	308	606	34	325	642	434	1520	393	1609	0.22	0.20	1.169	1609	977	1001
$P = 1.70 \times 10^{-8}$;	UK2/NSCCG	121	985	1746	98	898	1822	1227	4477	1094	4542	0.22	0.19	1.138	4542	2852	2818
OR = 1.09; 95%	Scotland2/SOCCS	77	694	1235	74	639	1344	848	3164	787	3327	0.21	0.19	1.133	3327	2006	2057
CI = 1.06–1.13;	VQ58	86	605	1106	113	832	1742	777	2817	1058	4316	0.22	0.20	1.125	4316	1797	2687
Phet = 0.28;	CFR	50	364	756	51	327	607	464	1876	429	1541	0.20	0.22	0.888	1541	1170	985
$I^2 = 14.8\%$; allele	UK3/NSCCG	122	947	1920	115	850	1861	1191	4787	1080	4572	0.20	0.19	1.053	4572	2989	2826
1 = G; allele 2 = A	Scotland3/SOCCS	48	263	519	51	315	566	359	1301	417	1447	0.22	0.22	0.958	1447	830	932
	UK4/CORGI2BCD	24	158	306	45	309	669	206	770	399	1647	0.21	0.20	1.104	1647	488	1023
	Cambridge	89	755	1366	76	664	1444	933	3487	816	3552	0.21	0.19	1.165	3552	2210	2184
	COIN/NBS	102	701	1330	89	770	1642	905	3361	948	4054	0.21	0.19	1.151	4054	2133	2501
	Helsinki	67	385	476	49	317	437	519	1337	415	1191	0.28	0.26	1.114	1191	928	803
	Prague	47	335	552	33	230	388	429	1439	296	1006	0.23	0.23	1.013	1006	934	651
	Kentucky	41	312	657	57	509	1017	394	1626	623	2543	0.20	0.20	0.989	2543	1010	1583
	EPICOLON	57	429	840	46	442	906	543	2109	534	2254	0.20	0.19	1.087	2254	1326	1394
	Australia	25	151	264	17	152	269	201	679	186	690	0.23	0.21	1.098	690	440	438
	Leiden	45	284	521	28	212	448	374	1326	268	1108	0.22	0.19	1.166	1108	850	688
rs7136702; chr12:	UK1/CORGI	131	433	357	113	430	386	695	1147	656	1202	0.38	0.35	1.110	1202	921	929
49,166,483; $z = 6.69$;	Scotland1/COGS	146	443	388	126	444	431	735	1219	696	1306	0.38	0.35	1.131	1306	977	1001
$P = 2.23 \times 10^{-11}$;	UK2/NSCCG	380	1331	1140	329	1306	1183	2091	3611	1964	3672	0.37	0.35	1.083	3672	2851	2818
OR = 1.10; 95%	Scotland2/SOCCS	276	975	755	275	935	847	1527	2485	1485	2629	0.38	0.36	1.088	2629	2006	2057
CI = 1.07–1.12;	VQ58	237	869	694	295	1290	1102	1343	2257	1880	3494	0.37	0.35	1.106	3494	1800	2687
Phet = 0.51;	CFR	155	604	427	103	444	450	914	1458	650	1344	0.39	0.33	1.296	1344	1186	997
$I^2 = 0.0\%$; allele	UK3/NSCCG	402	1388	1190	359	1283	1180	2192	3768	2001	3643	0.37	0.35	1.059	3643	2980	2822
1 = T; allele 2 = C	Scotland3/SOCCS	118	310	270	122	401	356	546	850	645	1113	0.39	0.37	1.108	1113	698	879
	UK4/CORGI2BCD	81	215	190	151	466	444	377	595	768	1354	0.39	0.36	1.117	1354	486	1061
	Cambridge	332	955	903	261	1015	906	1619	2761	1537	2827	0.37	0.35	1.079	2827	2190	2182
	COIN/NBS	287	893	844	321	1121	1059	1467	2581	1763	3239	0.36	0.35	1.044	3239	2024	2501
	Helsinki	103	389	436	72	334	414	595	1261	478	1162	0.32	0.29	1.147	1162	928	820
	Prague	85	419	430	57	291	303	589	1279	405	897	0.32	0.31	1.020	897	934	651
	Kentucky	140	478	392	215	750	618	758	1262	1180	1986	0.38	0.37	1.011	1986	1010	1583
	EPICOLON	198	642	486	187	623	584	1038	1614	997	1791	0.39	0.36	1.155	1791	1326	1394
	Leiden	115	388	341	92	269	321	618	1070	453	911	0.37	0.33	1.162	911	844	682
rs11169552; chr12:	UK1/CORGI	56	328	537	67	350	512	440	1402	484	1374	0.24	0.26	0.891	1374	921	929
49,441,930; $z = 6.88$;	Scotland1/COGS	60	369	544	76	406	519	489	1457	558	1444	0.25	0.28	0.869	1444	973	1001
$P = 5.99 \times 10^{-12}$;	UK2/NSCCG	209	1062	1580	199	1124	1494	1480	4222	1522	4112	0.26	0.27	0.947	4112	2851	2817
OR = 0.90; 95%	Scotland2/SOCCS	111	808	1087	152	821	1084	1030	2982	1125	2989	0.26	0.27	0.918	2989	2006	2057
CI = 0.88–0.93;	VQ58	109	665	1026	201	1046	1442	883	2717	1448	3930	0.25	0.27	0.882	3930	1800	2689
Phet = 0.50;	CFR	72	450	663	73	408	516	594	1776	554	1440	0.25	0.28	0.869	1440	1185	997
$I^2 = 0.0\%$; allele	UK3/NSCCG	167	1179	1625	214	1142	1463	1513	4429	1570	4068	0.25	0.28	0.885	4068	2971	2819
1 = T; allele 2 = C	Scotland3/SOCCS	14	127	176	80	321	490	155	479	481	1301	0.24	0.27	0.875	1301	317	891
	UK4/CORGI2BCD	34	175	277	80	395	554	243	729	555	1503	0.25	0.27	0.903	1503	486	1029
	Cambridge	155	824	1241	163	853	1172	1134	3306	1179	3197	0.26	0.27	0.930	3197	2220	2188
	COIN/NBS	135	818	1107	189	973	1338	1088	3032	1351	3649	0.26	0.27	0.969	3649	2060	2500
	Helsinki	103	407	401	153	356	303	613	1209	662	962	0.34	0.41	0.737	962	911	812
	Prague	51	375	508	38	273	340	477	1391	349	953	0.26	0.27	0.936	953	934	651
	Kentucky	58	377	575	93	665	825	493	1527	851	2315	0.24	0.27	0.878	2315	1010	1583
	EPICOLON	56	453	817	75	471	848	565	2087	621	2167	0.21	0.22	0.945	2167	1326	1394
	Leiden	53	304	492	57	251	378	410	1288	365	1007	0.24	0.27	0.878	1007	849	686

Ca, cases; Co, controls; 11, rare homozygote; 12, heterozygote; 22, common homozygote; 1, minor allele; 2, major allele. Allele 1 is risk allele for rs6691170, rs6687758 and rs7136702; allele 2 is risk allele for rs11169552. MAF, minor allele frequency; OR, odds ratio.

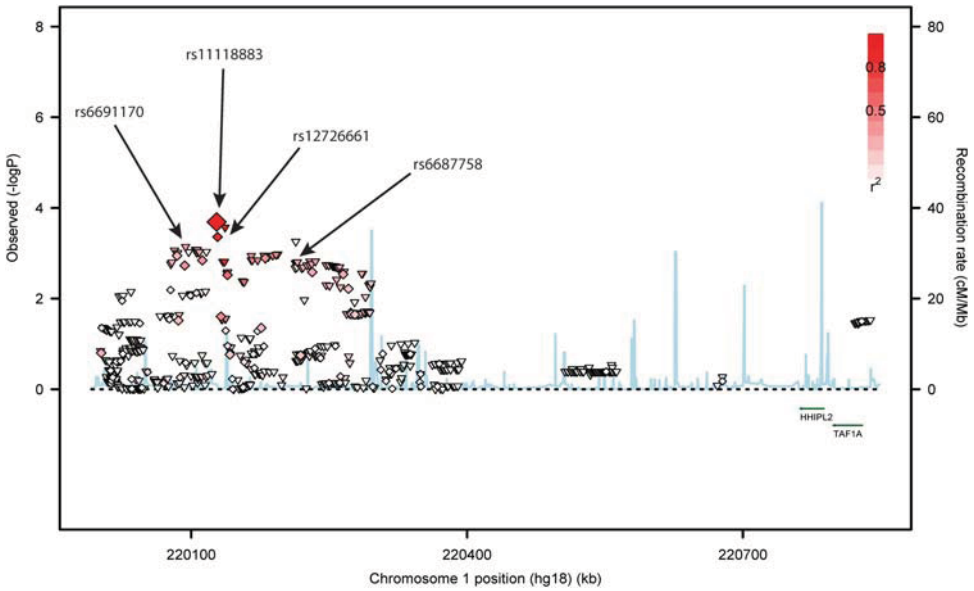

Figure 1. Individual SNP associations in the 1q41 region. Association testing was performed in SNPtest using typed and imputed genotypes from the three GWAS series (UK1, Scotland1 and VQ58) and displayed using SNAP (http://www.broadinstitute.org/mpg/snap/). The X-axis shows position on chromosome 1 and the Y-axis, −log10(P) from the per allele association test. The most strongly associated SNP, rs11118883, is shown as a large diamond, and the colours of other data points reflect the LD between that SNP and rs11118883. The smaller diamond points indicate genotyped SNPs and the triangles indicate imputed SNPs. The blue line represents recombination rates.

Table 2. Two-SNP logistic regression analysis showing best signal in the 1q41 region in comparison with the originally reported SNPs

SNPs	Positions (bases)	LD (r^2, D')	Risk allele ($freq_{cases}$, $freq_{controls}$)	No. cases, no. controls	OR	95% CI	Z	P-value	AIC
rs6691170	220,112,069	0.15,	T (0.38,0.36)	3272,	1.09	1.01–1.17	2.95	0.0032	10486
rs6687758	220,231,571	0.65	G (0.22,0.20)	4572	1.08	0.98–1.18	1.61	0.108	
rs11118883	220,127,645	0.92,	A (0.32,0.29)	3206,	2.32	1.49–3.62	3.71	2.07×10^{-4}	10475
rs12726661	220,134,411	1.00	A (0.68,0.71)	4452	0.49	0.32–0.76	3.16	1.58×10^{-3}	

LD is shown for the pair of SNPs being tested. OR, odds ratio; AIC, Akaike information criterion (AIC = −2 * log-likelihood + 2 * (number of parameters)). Note the lower AIC, showing a better model fit, for the test of rs11118883 + rs12726661 compared with rs6691170 + rs6687758. Individual AICs for these four SNPs were, respectively, 10487, 10489, 10484 and 10487.

greater than those for the low-low haplotype (GA), inconsistent with a functional SNP being in complete LD with a haplotype indicated by the pair of tagSNPs (Supplementary Material, Table S1). We also tested for evidence of epistasis between rs6691170 and rs6687758 using case–control logistic regression analysis, incorporating interaction between SNPs as a variable, but no evidence of deviation from log-additive SNP effects was found ($P = 0.292$).

Having failed to find evidence for the simplest situations—namely that one of each tagSNP pair captured the great majority of the association signal or that the tagSNPs essentially acted as simple two-locus tags for the functional variants in each region—we attempted to deconvolute the 1q41 signal by association testing of imputed SNPs in the region. The three GWAS sample sets, UK1, Scotland 1 and VQ58, were imputed to the combined 1000 genomes and HapMap3 reference set. A total of 630 SNPs in the 220–221 Mb region on chromosome 1q41 was successfully imputed from 76 genotyped SNPs. The strongest association signal (Fig. 1, Supplementary Material, Table S2), as measured by association test P-value, was at rs11118883 (chr1: 220,127,645), an imputed SNP in moderate LD with rs6691170 ($r^2 = 0.40$, $D' = 0.74$) and rs6687758 ($r^2 = 0.31$, $D' = 0.77$).

We then used reverse stepwise logistic regression analysis to determine whether rs6691170 and rs6687758, or other combinations of SNPs, best accounted for the association between CRC and 1q41 variation. Using a final significance threshold of $P = 0.01$, we found that two imputed SNPs, rs11118883 and rs12726661, were most strongly associated with the CRC risk (Table 2, Supplementary Material, Table S2). By comparison, a joint analysis of rs6687758 and rs6691170 in the same three GWAS data sets gave much weaker evidence of association, as assessed using the Akaike Information Criterion (AIC). Indeed, a model incorporating rs11118883 alone—although not one with rs12726661 alone—provided a better fit than a model incorporating both rs6687758 and rs6691170; haplotype-based association analysis supported these findings (data not shown).

We were surprised to note that in a single-SNP analysis the direction of effect for rs12726661 was reversed—the minor

allele was associated with disease risk—compared with that in the two-SNP analysis. We determined that rs11118883 and rs12726661 were in strong LD ($r^2 = 0.98$, $D' = 1.00$) in our samples, consistent with data from the 1000 genomes project and HapMap3 that had been used for imputation. Examination of the genotype distribution in our data set showed that deviation from perfect LD between the SNPs resulted from two sets of individuals: (i) 50 homozygous for the major allele at rs12726661 and heterozygous at rs11118883; and (ii) 15 heterozygous at rs12726661 and homozygous for the minor allele at rs11118883. Specifically, 28/50 in category (i) were cases and 4/11 in category (ii) were cases. For these 65 individuals, the risk of CRC was significantly greater than that of individuals with the other genotypes at rs12726661 and rs11118883 (OR = 2.10, $P = 0.003$, χ^2_1 test). A potential explanation for our apparently paradoxical findings is that there exists another allele, almost certainly relatively rare, that is associated with the minor allele of rs12726661 (but not with rs11118883), and that is protective against the CRC risk.

We then analysed our UK1, Scotland 1 and VQ58 individuals using GENECLUSTER with the original GWAS SNP genotypes in the rs6691170/6687758 region as inputs. There was no evidence to favour an underlying two-locus model over a one-SNP model (Fig. 2). The predicted most strongly associated single SNP was rs11577023, a SNP that is in very strong LD with rs11118883 ($r^2 = 0.93$, $D' = 1.0$) in our data.

We genotyped rs11118883 directly in a set of 84 UK control samples and found complete concordance with the imputed genotypes. rs11118883 (chr1:220,127,645) lies in a gene desert, within a region of LD that extends approximately from 220.0 to 220.3 Mb. The nearest gene, ~150 kb towards the centromere, is the MAP kinase regulator dual-specificity phosphatase 10 (*DUSP10*). DUSP10 inactivates p38 and also the Jun N-terminal kinase that phosphorylates c-Jun which is believed to play a role in CRC pathogenesis. rs11118883 itself lies upstream of *DUSP10* within a region with strong evidence of transcriptional regulatory activity (http://genome.ucsc.edu). Using 1000 genomes data, we found that rs11118883 is in strong LD ($r^2 > 0.7$) with at least six SNPs (rs12738322, rs12726661, rs4129271, rs11577023, rs10746414 and rs12137702). Of these, rs10746414 and rs12137702 are also close to regions with potential effects on transcription.

The 12q13.13 region

Analysis of the 12q13.13 region proceeded in parallel with that of the 1q41 region using essentially the same strategy. We initially confirmed the individual associations of SNPs rs7136702 and rs11169552 with the CRC risk in the extended data sets (Table 1). Unconditional logistic regression analysis did not exclude the possibility that the two SNPs had independent effects; for rs7136702 and rs11169552, the association statistics were $P = 1.63 \times 10^{-5}$ (OR = 1.07) and $P = 1.70 \times 10^{-7}$ (OR = 0.92), respectively, showing that one SNP did not simply capture all of the association signals. Further analysis showed that the association signal was not derived from a single high-risk haplotype tagged by rs7136702 and rs11169552 (Supplementary Material, Table S3) and there was no evidence of epistasis between the SNPs ($P = 0.903$).

We imputed SNPs within the 48.5–50 Mb region of chromosome 12 using the combined 1000 genomes and HapMap3 reference panel in the 3 GWAS sample sets (UK1, Scotland 1 and VQ58) (Fig. 3, Supplementary Material, Table S4). A total of 2736 SNPs was successfully imputed from 158 genotyped SNPs. The most significant single-SNP association was at the imputed SNP rs7972465 [OR = 1.18, 95% confidence interval (CI) 1.11–1.27, $P = 8.22 \times 10^{-7}$), a signal slightly stronger than that of rs11169552 (OR = 0.85, 95% CI 0.79–0.91, $P = 1.08 \times 10^{-5}$) and notably stronger than that of rs7136702 (OR = 1.13, 95% CI 2.06–1.21, $P = 3.85 \times 10^{-4}$). Direct genotyping in 91 UK control individuals showed that imputation of rs7972465 was very good, although not perfect ($r^2 = 0.93$).

Reverse stepwise logistic regression analysis was then used to assess whether rs11169552 and rs7136702, or other combinations of SNPs in the region, best accounted for the association between CRC and 12q13.13 variation (Table 3). Many highly correlated SNPs exist within the region, making this analysis difficult. Nonetheless, while rs11169552 remained in the regression model after stepwise elimination of less strongly associated SNPs, a number of SNPs provided improved or similar associations compared with rs7136702 in a two-SNP model with rs11169552. One of these SNPs was rs7972465 (Table 3, Fig. 4), but another SNP, rs706793, a SNP in very low LD with rs11169552 (Table 3), provided a larger improvement in the AIC (see also Supplementary Material, Table S4).

We then undertook GENECLUSTER analysis of the UK1, Scotland 1 and VQ58 sample sets in the 12q13.13 region. There was no good evidence to distinguish between underlying two-locus and one-locus models (Fig. 5), although the association signal showed two peaks at ~48.85 Mb (close to rs706793) and at ~49.45 Mb (very close to rs11169552) that could not readily be explained by long-range LD between these two regions (Supplementary Material, Fig. S1). The predicted most strongly associated SNP under the one-SNP model was rs3184122 (Supplementary Material, Table S4), a variant that is in moderate or strong LD (Fig. 5) with rs11169552 ($r^2 = 0.19$, $D' = 0.92$), rs7136702 ($r^2 = 0.49$, $D' = 0.73$) and rs706793 ($r^2 = 0.47$, $D' = 0.94$), and strong LD with rs7972465 ($r^2 = 0.87$, $D' = 1.00$).

Since the various analyses had not resolved the question of whether there exist one or two independent CRC-associated SNPs in the 12q13.13 region, we used PLINK to examine the associations with disease of the haplotypes (Fig. 4) for rs706793, rs7972465 and rs11169552. As expected, the haplotype CGC was most strongly associated with risk (Table 4, Supplementary Material, Table S5). The G (risk) allele at rs7972465 was essentially present only on this haplotype, but it appeared that haplotypes containing the T allele at rs7972465 were not all low risk and therefore that rs7972465 did not explain all the association signal. We therefore considered the association signals when we fixed the alleles at rs706793 and rs11169552 and allowed those at rs7972465 to vary, and vice versa. Initially, we undertook simple comparisons between haplotype frequencies in cases and controls, and found that the rs706793 and rs11169552

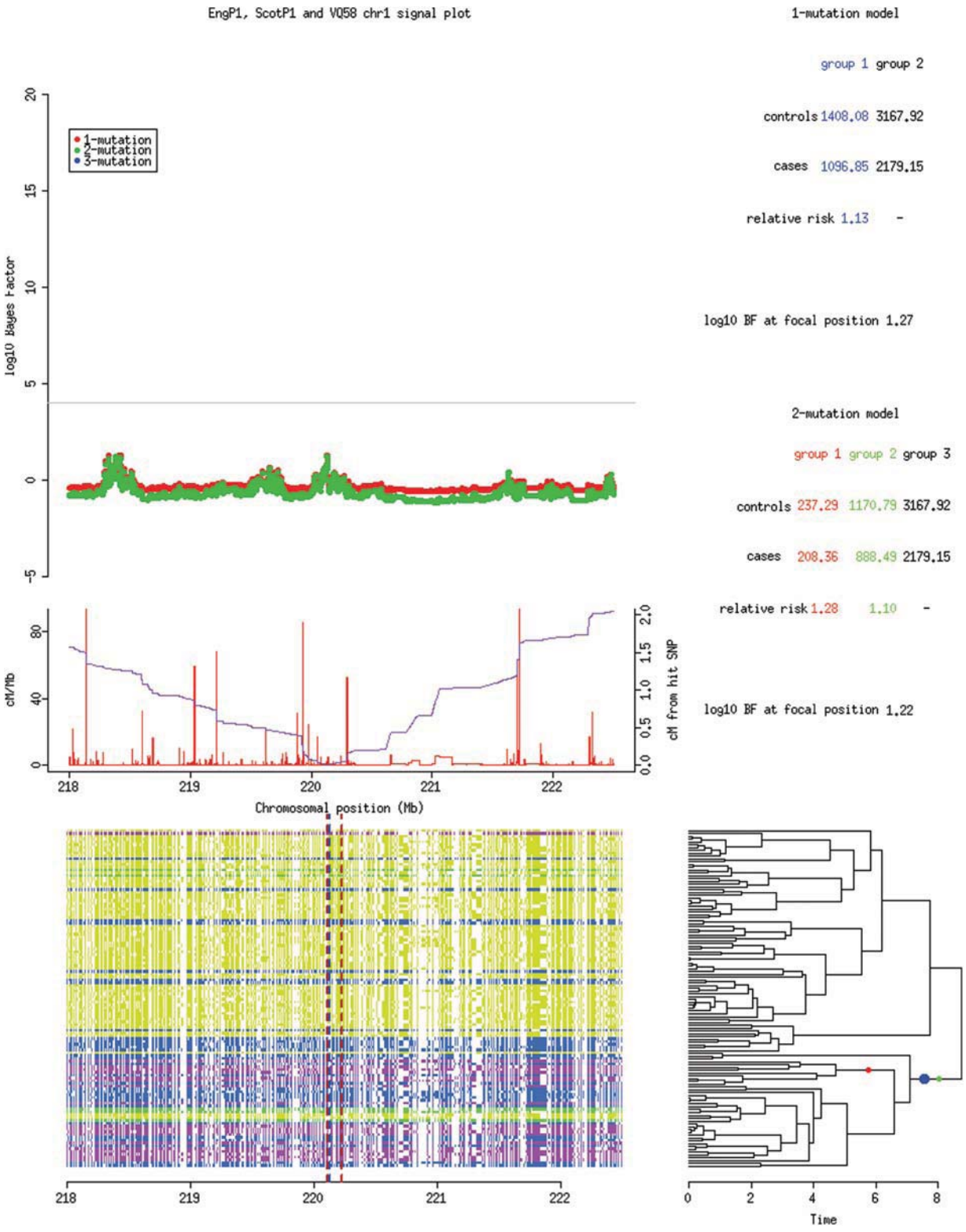

Figure 2. GENECLUSTER output for the 1q41 region. The upper left panel compares the Bayes factors (BFs) for models in which the association signals at rs6691170 and 6687758 are derived from either one functional SNP (red) or two functional SNPs (green). Recombination rates are also shown as a red line. The upper right panel shows the $\log_{10}$(BF) at the focal position—the site of the highest $\log_{10}$(BF), here chr1:220,129,000 bases—under one- and two-SNP models. The lower right panel shows reconstructed genealogies for UK1, Scotland1 and VQ58 combined, based on each individual's genotypes in the region from the Illumina Hap300/370/550 panels and HapMap2 data. The most likely positions of SNP origins under the one-SNP model (blue, rs11577023) and two-SNP model (green and red) are shown. These result in counts of cases and controls and relative risks as indicated in the upper right panel. The lower left panel shows haplotypes (rows) and SNPs (columns). Note that the region analysed extends for several Mb flanking rs6691170 and 6687758; although no signal reaches nominal significance at $\log_{10}(BF) = 4$, there is some evidence of a second independent region of 1q associated with CRC at ~218.2 Mb, as we have reported previously. The importance of rs11577023 was supported by the Margarita analysis in which it was the second most strongly associated with disease ($P = 3.59 \times 10^{-4}$).

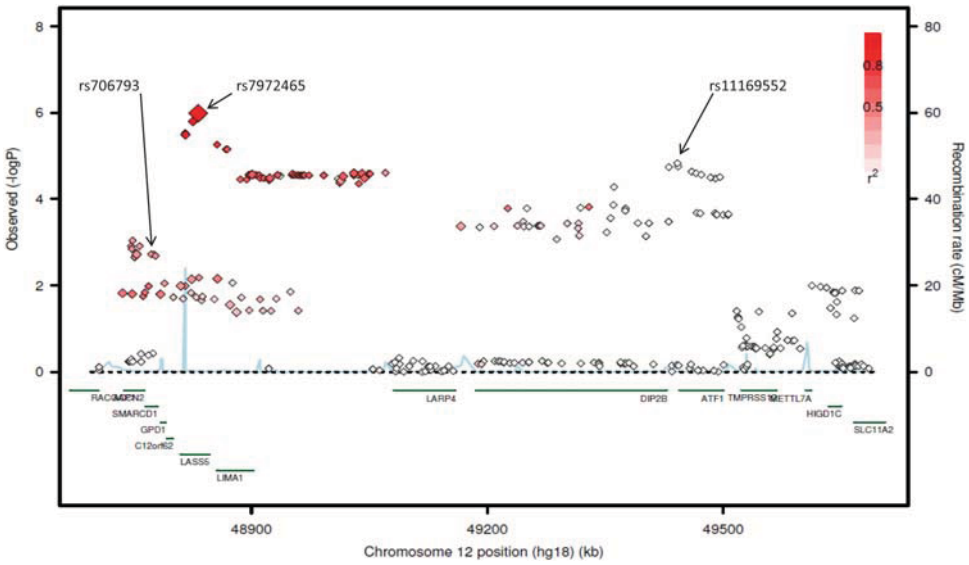

Figure 3. Individual SNP associations in the 12q13.13 region. Legend is as for Figure 1.

Table 3. Two-SNP logistic regression analysis showing best signals in the 12q13.13 region in comparison with the original reported SNPs

SNPs	Genotyped or imputed?	Positions (bases)	LD (r^2, D')	Risk allele ($freq_{cases}$, $freq_{controls}$)	No, cases, no. controls	OR	95% CI	z	P-value	AIC
rs11169552	Genotyped	49,441,930	0.040	C (0.76, 0.73)	3276	0.92	0.81–0.94	3.37	7.52×10^{-4}	10473
rs7136702	Genotyped	49,166,483	0.57	T (0.37, 0.35)	4576	1.08	1.01–1.16	2.13	0.033	
rs11169552	Genotyped	49,441,930	0.19	C (0.76,0.73)	3206	0.89	0.82–0.97	2.69	0.007	10469
rs3184122	Imputed	48,856,394	0.92	C (0.41,0.37)	4480	1.12	1.04–1.21	2.96	0.003	
rs11169552	Genotyped	49,441,930	0.06	C (0.76, 0.73)	3268	0.88	0.81–0.95	3.40	6.74×10^{-4}	10468
rs35031884	Imputed	49,063,840	1.00	A (0.32,0.29)	4554	1.15	1.05–1.25	3.19	0.0014	
rs11169552	Genotyped	49,441,930	0.17	C (0.76,0.73)	3268	0.90	0.83–0.97	2.64	0.008	10466
rs7972465	Imputed	48,832,392	0.89	G (0.21,0.18)	4563	1.14	1.06–1.22	3.43	6.04×10^{-4}	
rs11169552	Genotyped	49,441,930	0.002	C (0.76,0.73)	3266	0.84	0.78–0.91	4.56	5.12×10^{-6}	10426
rs706793	Genotyped	48,754,036	0.095	C (0.60,0.57)	4557	0.49	0.32–0.76	3.23	0.0012	

SNP pairs are shown in order of descending Aikake Information Criterion (AIC). Note that in single-SNP analysis, rs706793 provided only slightly worse evidence of association (OR = 0.90, 95% CI 0.84–0.96, $P = 0.002$) than in combined analysis with rs11169552. Individual AICs for rs11169552, rs7136702, rs3184122, rs35031884, rs7972465 and rs706793 were, respectively, 10476, 10483, 10475, 10477, 10471 and 10447. Incorporation of rs7972465 into a regression model with rs11169552 and rs706793 did not improve the model's fit (AIC = 10426).

risk alleles, but not the rs7972465 risk allele, were found at significantly higher frequencies in cases than controls (Supplementary Material, Table S6). Since this analysis suggested that there might be independent effects of rs706793 and rs11169552—and that the signal at rs7972465 resulted from LD with these two SNPs—we proceeded to a further evaluation of this possibility using conditional haplotype analysis in PLINK. We again compared two scenarios, (i) in which the CGC and CTC haplotypes were equivalent (that is, varying rs7972465) and (ii) in which the CTC and TTT haplotypes were equivalent (that is, varying rs706793 and rs11169552). No effect was seen in the first case (likelihood ratio test, $P = 0.35$), whereas there was a significant difference in the second case ($P = 0.023$), again supporting effects of rs706793 and rs11169552 rather than rs7972465.

Further genotyping in additional sample sets strengthened the rs706793 association with CRC, although it did not reach formal significance and there was some evidence of inter-study heterogeneity, the origins of which remain unclear (Supplementary Material, Table S7). Logistic regression analysis in the extended sample set continued to support a model incorporating rs706793 and rs11169552 ($P = 8.38 \times 10^{-4}$ and $P = 7.82 \times 10^{-6}$, respectively, AIC = 27932) over one with rs7163702 and rs11169552 ($P = 3.05 \times 10^{-5}$ and $P = 9.05 \times 10^{-3}$, AIC = 27999).

rs706793 (chr12:48,754,036) and rs11169552 (chr12:49,441,930) are separated by a predicted recombination hotspot at ~48.8 Mb in the HapMap data (Fig. 3) but not in our own data (Fig. 5), although LD in the region is complex (Supplementary Material, Fig. S1). The 12q13.13 region

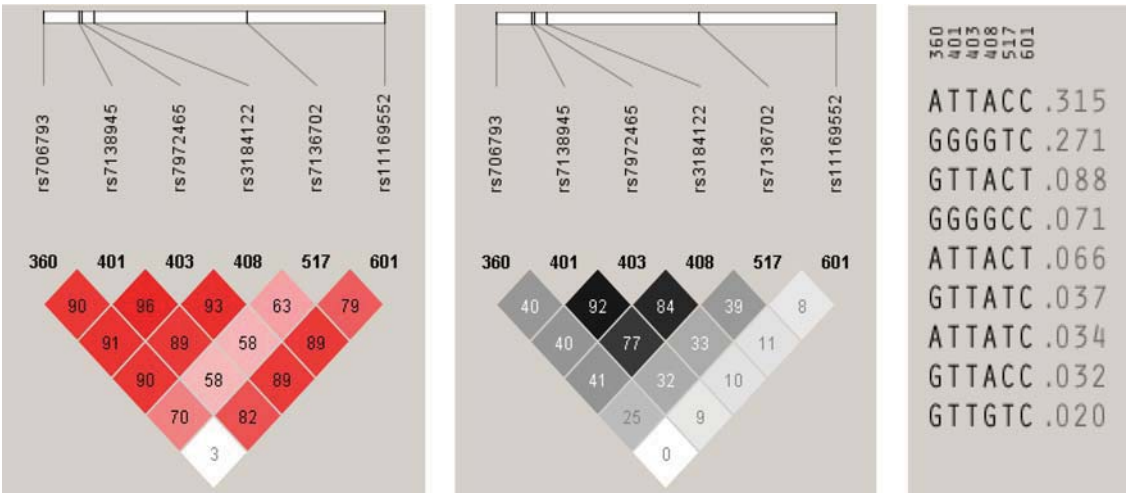

Figure 4. LD and main haplotypes at SNPs with best evidence of association on 12q13.13. Note that in this Haploview output from HapMap3 data, the alleles at rs706793 are shown on the opposite strand (that is G/A rather than C/T as used in the rest of this manuscript).

contains coding genes *ACCN2*, *SMARCD1*, *GPD1*, *LASS5*, *LIMA1* and *ATF1*. *ACCN2* probably encodes an ion channel protein, SMARCD1 is part of chromatin remodelling complex SNF/SWI, GPD1 is glycerol-3-phosphate dehydrogenase and LASS5 is probably a ceramide synthase. *LIMA1* codes for EPLIN, a protein downregulated in some cancers. ATF1 is a transcription factor centrally involved in the stress response and in the pathogenesis of angiomatoid fibrous histiocytoma and clear cell sarcoma through translocation. Supplementary Material, Table S8 lists SNPs in strong LD ($r^2 > 0.70$) with rs706793, rs7972465 or rs11169552, and provides annotation for those with evidence of potential roles in gene or protein regulation or function.

DISCUSSION

We have undertaken additional genotyping and more detailed analysis in order to understand better the dual tagSNP association signals that we observed on chromosomes 1q41 (rs6991170, rs6687758) and 12q13.13 (rs11169552, rs7136702) in a GWAS of CRC (1). In both cases, genotyping of additional sample series confirmed the originally reported associations, without demonstrating good evidence for the three simplest scenarios: independent functional variants; capture of the association signal by one of the pair of SNPs; or two-SNP tagging of a single haplotype on which functional variation lay. We therefore proceeded to more detailed analyses in each region, after imputation of genotypes where appropriate in the data sets with best coverage of each region (UK1, Scotland1 and VQ58). It is conceivable that the analysis of these three data sets, which had already been used in SNP discovery, would introduce a small amount of bias into the fine mapping. However, we reasoned that the marginal differences in association that might occur would be more than outweighed by the power provided by the use of these data sets.

For 1q41, the single-SNP association test, logistic regression analysis and GENECLUSTER all found that SNP rs11118883, or a SNP in strong LD, was most likely to be responsible for the signal of association. This SNP itself is a very good functional candidate, lying within or immediately adjacent to regions bearing histone methylation and acetylation marks, DNAse I hypersensitive sites and sites of transcription factor binding (http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=195445293&c=chr1&g=wgEncodeReg). The SCAN expression Quantitative Trait Locus (eQTL) database (4) reports rs11118883 being associated ($P = 8 \times 10^{-5}$) in Europeans with expression of platelet-derived growth factor β (*PDGFRB*, chr5q31–q32), although this association requires confirmation in appropriate cell types for the CRC risk and is not present in the Genevar eQTL database (http://www.sanger.ac.uk/resources/software/genevar) (5). The possibility that the minor allele of rs12726661 is associated with a second, presumably rare, variant that is protective against the CRC risk is intriguing. While speculative, such a scenario has precedents, such as the *MDM2* promoter SNP rs117039649 (6).

For 12q13.13, a complex situation was found. Single-SNP analysis found variants with much stronger association signals than either rs11169552 or rs7136702, notably at rs7972465 although small imputation inaccuracies may have inflated this signal. GENECLUSTER analysis found no greater evidence for a two-SNP than one-SNP model and detected the best signal for the former at a SNP, rs3184122, that is in strong LD with rs7972465. Logistic regression analysis, however, supported a two-SNP model, in which a signal at rs706793 was added to that at rs11169552. rs706793 and rs11169552 are in very weak LD, but rs706793 is in moderate LD with rs7163702 ($r^2 = 0.20$, $D' = 0.60$). Haplotype analysis supported the logistic regression analysis, in that the genotype at rs7972465 did not affect the risk associated with the rs706793–rs11169552 haplotypes, whereas the reverse scenario (high- versus low-risk rs706793–rs11169552 haplotypes) did affect risk. As regards eQTLs for the 12q13.13 SNPs, Genevar shows rs706793 to be associated with *LASS5* expression (at $P < 10^{-4}$), although this association is not reported in SCAN.

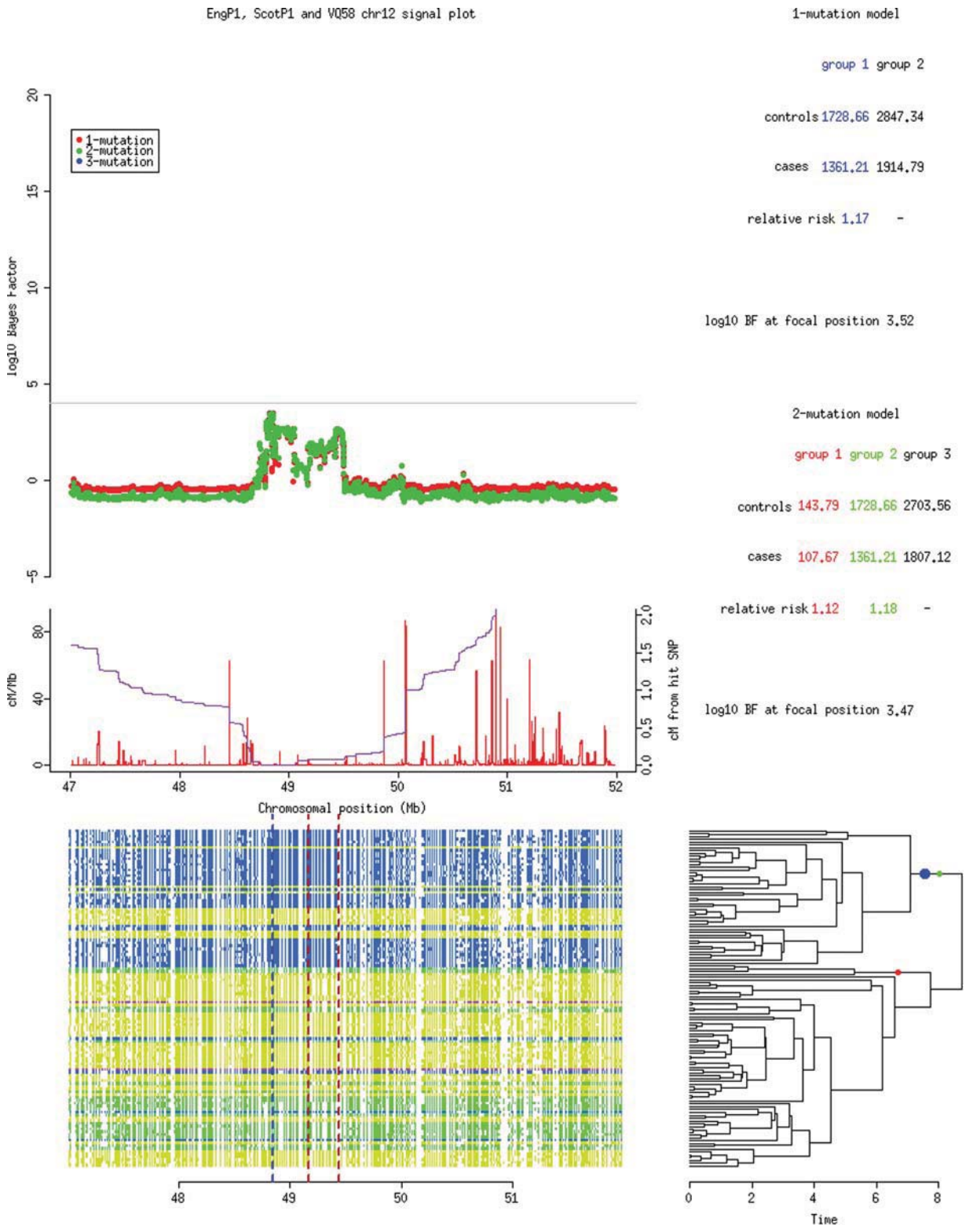

Figure 5. GENECLUSTER output for the 12q13.13 region. The legend is as for Figure 2, except that the focal position is Chr12:48,849,000 and the double peak of association at ~48.85 and 49.45 Mb should be noted. The top SNP (blue dot) under the one-SNP model is the imputed SNP rs3184122. The top-genotyped SNP in the GENECLUSTER analysis was rs7138945, which was the SNP with the second-best association signal in Margarita ($P = 1.14 \times 10^{-5}$).

Clearly, all post-GWAS fine-mapping studies face intrinsic difficulties, such as the use of imputed genotypes, despite the use of stringent criteria for SNP inclusion, and a limited ability to differentiate among association signals of similar magnitudes. The analysis of the 12q13.13 region illustrates some of these problems well. Although a much more strongly

Table 4. Haplotype analysis in the 12q13.13 region

Haplotype	Freq. in cases	Freq. in controls	OR	*P*-value
TTT	0.0922	0.1085	0.82	5.3×10^{-4}
CTT	0.1407	0.1567	0.87	4.5×10^{-3}
CGC	0.3829	0.3443	1.19	6.2×10^{-7}
TTC	0.3102	0.3194	0.96	0.212
CTC	0.0739	0.0710	1.05	0.483

Haplotypes (cen-tel) at rs706793, rs7972465 and rs11169552 were analysed, notwithstanding the low LD between the first and last of these SNPs. Five haplotypes with frequencies of >0.01 were predicted. Ca, cases; Co, controls. OR, odds ratio relative to all other haplotypes. *P*-value is from analysis of effects of all haplotypes on disease risk in a logistic regression model. Odds ratios and *P*-values relative to reference haplotype TTT are given in Supplementary Material, Table S5.

CRC-associated SNP than the original tagSNPs was identified through imputation, the balance of evidence slightly favours this signal resulting from two independent association signals, as we have previously found for the *GREM1* locus (2). In the 1q41 region, in contrast, rs11118883—a SNP in moderate LD with both the original tagSNPs—emerged as an excellent candidate for the functional variant.

MATERIALS AND METHODS

Sample sets

The Kentucky samples comprised 1020 incident colon cancer cases and 1598 population controls of white European origin recruited between July 2003 and December 2009. Eligible cases were identified through the population-based Surveillance, Epidemiology and End Results (SEER) Kentucky Cancer Registry covering all residents living in the State of Kentucky at the time of diagnosis. We used random digital dialling to recruit population controls who were 40 years of age or older and had no personal history of cancer other than skin cancer. We excluded those with known inflammatory bowel diseases, family history of familial adenomatous polyposis and hereditary non-polyposis CRC.

The Prague cases (7) were patients with histologically confirmed CRC recruited between September 2004 and February 2009 from nine oncology departments in the Czech Republic: Prague (two), Benesov, Brno, Liberec, Ples, Pribram, Usti nad Labem and Zlin. During this period, a total of 1554 cases provided blood samples. This study includes 1001 subjects who could be interviewed, provided biological samples and were genotyped. Controls were 683 hospital-based volunteers with negative colonoscopy results for malignancy or idiopathic bowel diseases (CFCC, cancer-free colonoscopy inspected controls). CFCCs were selected from among individuals admitted to the same hospitals during the same period of the recruitment of the cases. The reasons for undergoing the colonoscopy were: (i) positive faecal occult blood test, (ii) haemorrhoids, (iii) abdominal pain of unknown origin, or (iv) macroscopic bleeding.

Details of other sample sets have been reported previously (2) and are provided briefly below.

UK1 (CORGI) 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 CRAds 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, 55% females) were spouses or partners unaffected by cancer and without a personal family history (to second degree relative level) of colorectal neoplasia. Known dominant polyposis syndromes, HNPCC/Lynch syndrome or bi-allelic *MUTYH* mutation carriers were excluded.

Scotland1 (COGS) included 980 CRC cases (51% male; mean age at diagnosis 49.6 years, SD $\pm$ 6.1) and 1002 cancer-free population controls (51% male; mean age 51.0 years; SD $\pm$ 5.9). Cases were for early age at onset (age $\leq$ 55 years). Known dominant polyposis syndromes, HNPCC/Lynch syndrome or bi-allelic *MUTYH* mutation carriers were excluded. Control subjects were sampled from the Scottish population NHS registers, matched by age ($\pm$5 years), gender and area of residence within Scotland.

VQ58 comprised 1832 CRC cases (1099 males, mean age of diagnosis 62.5 years; SD $\pm$ 10.9) from the VICTOR and QUASAR2 (www.octo-oxford.org.uk/alltrials/trials/q2.html) clinical trials of adjuvant therapy in stage II/III CRC. There were 2720 population control genotypes (1391 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.

The Australian study comprised 591 patients treated for CRC at the Royal Melbourne, Western and St Francis Xavier Cabrini Hospitals in Melbourne from 1999 to 2009. The 2353 controls were derived from Queensland or Melbourne: for the former, the controls came from the Brisbane Twin Nevus Study; for the latter, individuals were participants in the Genes in Myopia study. There was no overlap between the CFR and Australian data sets. Owing to potential residual ethnic heterogeneity within the Melbourne population, for the Australian cohort only we performed an additional screen to minimize heterogeneity after performing principal components analysis (PCA) to remove individuals who clustered with non-CEU individuals (see below). We achieved this by performing PCA on the Australian cases and controls without reference samples of known ancestry. We then paired each case with a control in a 1:1 ratio based on a maximum separation of 0.050 using the first and second eigenvectors. All unpaired samples were excluded, leaving 441 cases and 441 controls in the study. Calculation of the genomic inflation factor, λ_{GC}, showed this to be 1.02 after this filtering.

UK2 (NSCCG) consisted of 2854 CRC cases (58% male, mean age at diagnosis 59.3 years; SD $\pm$ 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) and the Royal Marsden Hospital Trust/Institute of Cancer Research Family History and DNA

Registry. The 2822 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 the 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) comprised 2024 CRC cases (61% male; mean age at diagnosis 65.8 years, SD ± 8.4) and 2092 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.

UK3 (NSCCG) comprised 7912 CRC cases (65% male; mean age at diagnosis 59 years, SD ± 8.2) and 4398 controls (40% male; mean age 62 years, SD ± 11.5) ascertained through NSCCG post-2005.

Scotland3 (SOCCS) comprised 1145 CRC cases (50% male; mean age at diagnosis 53.2 years, SD ± 15.4) and 2203 cancer-free population controls (47% male; mean age 51.8 years, SD ± 11.5). Controls were recruited as part of the Generation Scotland study.

UK4 (CORGI2BCD) consisted of 621 CRC or CRAd cases (46% male; mean age at diagnosis 58.3 years; SD ± 14.1) and 1121 cancer-free population or spouse controls (45% male; mean age 45.1 years, SD ± 15.9), sampled using the same criteria as UK1.

Cambridge/SEARCH consisted of 2248 CRC cases (56% male; mean age at diagnosis 59.2 years, SD ± 8.1) and 2209 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 5-year age bands to cases.

The COIN samples were 2151 cases derived from the COIN and COIN-B clinical trials of metastatic CRC. Median age was 63 years. COIN cases were compared against genotypes from 2501 population controls (1237 males), from the WTCCC2 National Blood Service (NBS) cohort (50% male; mean age at diagnosis 53.2 years, SD ± 15.4).

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.

EPICOLON included 1410 CRC cases matched with the same number of controls collected in a prospective fashion from centres in Spain. Exclusion criteria were Mendelian CRC syndromes and a personal history of inflammatory bowel disease.

The Leiden sample set included 858 unselected cases with CRC and 690 controls ascertained through genetic testing programmes for non-cancer-related conditions from the Leiden area.

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 disease. Only individuals of white European origin were included in the study.

Sample preparation and genotyping

Collection of blood samples and clinico-pathological information from patients and controls was undertaken with informed consent and ethical review board approval in accordance with the tenets of the Declaration of Helsinki. DNA was extracted from samples using conventional methods and quantified using PicoGreen (Invitrogen). The VQ, UK1, Scotland1 and Australia 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 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, CA, USA) or MassARRAY (Sequenom Inc., San Diego, CA, 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.

We excluded SNPs from analysis if they failed one or more of the following thresholds: GenCall scores <0.25; overall call rates <95%; minor allele frequency (MAF)<0.01; departure from the 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 the GWA analyses if they had evidence of non-white European ancestry by PCA-based analysis in comparison with HapMap samples (http://hapmap.ncbi.nlm.nih.gov/) or by self-report. Deviation of the genotype frequencies in the controls from those expected under HWE was assessed by the χ^2 test (1 df), or Fisher's exact test where an expected cell count was <5.

Association statistics and imputation

Associations between SNP genotype and disease status were primarily assessed in STATA v10 (http://www.stata.com/) and PLINK v1.07 (http://pngu.mgh.harvard.edu/~purcell/plink/) using allelic and Cochran–Armitage tests (both with 1df) respectively, or by Fisher's exact test where an expected cell count was <5. Genotypic (2 df), dominant (1 df) and recessive (1 df) tests were also performed. The risks associated with each SNP were estimated by allelic, heterozygous and homozygous ORs using unconditional logistic regression, and associated 95% CIs were calculated.

Joint analysis of data generated from multiple phases was conducted using standard methods for combining raw data based on the Mantel–Haenszel method in STATA and in PLINK. Joint ORs and 95% CIs were calculated assuming fixed- and random-effects models. Tests of the significance of the pooled effect sizes were calculated using a standard normal distribution. Cochran's Q statistic to test for heterogeneity and the I^2 statistic to quantify the proportion of the total variation due to heterogeneity were calculated. Large heterogeneity is typically defined as $I^2 \geq 75\%$. Where significant heterogeneity was identified, results from the random-effects model were reported. Alongside, we also performed meta-analysis based on allele dosage (0, 1, 2) and incorporated age and sex as co-variates. Although age and sex are associated with the CRC risk, they were not associated with SNP genotype and did not materially affect the significance of any of the reported associations (data not shown).

The combined effects of pairs or other multiples of loci identified as possibly associated with the CRC risk were investigated by unconditional or conditional logistic regression analysis in PLINK and STATA to test for independent effects of each SNP, stratifying by sample series. Logistic regression was undertaken both pairwise with the original tagSNP and then in a backwards analysis that initially included all SNPs with good evidence of association in each region. We used Haploview software v4.2 (http://www.broadinstitute.org/haploview) to infer the LD structure of the genome in the 1q41 and 12q13.13 regions, and used the expectation maximum algorithms in Haploview or PLINK to infer haplotypes.

To predict genotypes at untyped SNPs in both regions, imputation of the UK1, Scotland 1 and VQ58 data sets was performed using the IMPUTE2 software and the combined CEU 1000 Genomes low-coverage pilot and complete HapMap3 haplotypes reference set, which was filtered to remove duplicate haplotypes (both from https://mathgen.stats.ox.ac.uk/impute/impute_v2.html) (8). Association statistics for imputed SNPs were calculated in SNPTEST v1.1.5 (www.stats.ox.ac.uk/~marchini/software/gwas/snptest.html) using the '–proper' option, which is an additive model score test based on missing data likelihood, to allow for the uncertainty of imputed genotypes (9). Imputed markers with proper_info scores <0.5, imputed call rates per SNP <0.9 (using a maximum genotype probability threshold of 0.9 to call a genotype) and MAFs <0.01 were excluded from the analyses. Meta-analyses of the sample sets were carried out with Meta (10) (http://www.stats.ox.ac.uk/~jsliu/meta.html) and in STATA, using the genotype probabilities from IMPUTE2 where a SNP was not directly typed.

The GENECLUSTER program (11) was used to analyse our UK1, Scotland1 and VQ58 samples specifically in order to test whether one- or two-SNP models better fitted the association signals in each SNP region. GENECLUSTER is a Bayesian method that uses HapMap haplotypes to estimate genealogy of samples and by jointly testing all SNPs on each branch of the genealogy in cases and controls, the program indicates the identities of the SNP(s) most likely to have the strongest association signal, thus potentially helping to identify functional variation. The default model parameters were used, specifically mutation model prior: (0.50, 0.50, 0.00), max number of trees to consider per location: 1 and beta risk prior parameters: (5.00, 5.00).

Essentially for comparative purposes, we also ran the Margarita program (12), based on ancestral recombination graphs (ARGs), in UK1, Scotland1 and VQ58 for the genotyped SNPs in the 1q and 12q regions. This program aims to maximize available information as to the location and identity of a functional SNP by reconstructing the genealogical history of the sample population. For each ARG, a putative risk mutation is placed on the marginal tree and the frequency of each branch in cases and controls is assessed. For each region, 30ARGs were constructed and the significance of a SNP at each branchpoint assessed by 10 000 permutations. Unlike GENECLUSTER, Margarita does not specifically address the issue of whether there are two independent underlying SNP in each region, and comparison was therefore restricted to the single-SNP scenario.

Genome co-ordinates were taken from the NCBI build 36/hg18 (dbSNP b126).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

This study made use of genotyping data on the 1958 Birth Cohort and NBS samples, kindly made available by the Investigators of those studies and the Wellcome Trust Case-Control Consortium 2; a full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk/. We are also grateful to the Spanish National Genotyping Center (CEGEN-ISCIII)-USC node. The work was carried out (in part) at the Esther Koplowitz Centre, Barcelona. We are grateful to colleagues in the EPICOLON, CORGI and COGENT consortia. Finally, we would like to thank all individuals who participated in the study.

Conflict of Interest statement. None declared.

FUNDING

Funding was primarily provided by Cancer Research UK. The EU FP7 CHIBCHA grant supported LGC-C through funding to IPMT, SC-B and ACar. Core infrastructure support to the Wellcome Trust Centre for Human Genetics, Oxford was provided by grant 090532/Z/09/Z. I.P.M.T. received support from the Oxford NIHR Comprehensive Biomedical Research Centre. The UK National Cancer Research Network supported the NSCCG. Additional funding to M.D. was provided by the Medical Research Council (G0000657-53203), CORE and Scottish Executive Chief Scientist's Office (K/OPR/2/2/D333, CZB/4/449). The EPICOLON work was supported by grants from the Fondo de Investigación Sanitaria/FEDER (08/0024, 08/1276, PS09/02368), Ministerio de Ciencia e Innovación (SAF2010-19273), Asociación Española contra el Cáncer (Fundación Científica y Junta de Barcelona) and Fundació Olga Torres (CRP). S.C.-B. and C.F.-R. are supported by contracts from the Fondo de Investigación Sanitaria

(CP03-0070 and PS09/02368). CIBERehd and CIBERER are funded by the Instituto de Salud Carlos III. For the Melbourne cases, work was supported by the Hilton Ludwig Cancer Metastasis Initiative. The specimens and data from Australian colon cancer patients were provided by the Victorian Cancer Biobank and BioGrid Australia with appropriate ethics approval. The Victorian Cancer Biobank is supported by the Victorian Government. CERA receives operational infrastructure support from the Victorian Government. Funding to pay the Open Access publication charges for this article was provided by the Wellcome Trust.

REFERENCES

1. Houlston, R.S., Cheadle, J., Dobbins, S.E., Tenesa, A., Jones, A.M., Howarth, K., Spain, S.L., Broderick, P., Domingo, E., Farrington, S. *et al.* (2010) Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nat. Genet.*, **42**, 973–977.
2. Tomlinson, I., Carvajal-Carmona, L., Dobbins, S., Tenesa, A., Jones, A., Howarth, K., Palles, C., Broderick, P., Jaeger, E., Farrington, S. *et al.* (2011) Multiple common susceptibility variants near BMP pathway loci GREM1, BMP4, and BMP2 explain part of the missing heritability of colorectal cancer. *PLoS Genet.*, **7**, e1002105.
3. Hemminki, K., Forsti, A., Houlston, R. and Bermejo, J.L. (2011) Searching for the missing heritability of complex diseases. *Hum. Mutat.*, **32**, 259–262.
4. Gamazon, E.R., Zhang, W., Konkashbaev, A., Duan, S., Kistner, E.O., Nicolae, D.L., Dolan, M.E. and Cox, N.J. (2010) SCAN: SNP and copy number annotation. *Bioinformatics*, **26**, 259–262.
5. Yang, T.P., Beazley, C., Montgomery, S.B., Dimas, A.S., Gutierrez-Arcelus, M., Stranger, B.E., Deloukas, P. and Dermitzakis, E.T. (2010) Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. *Bioinformatics*, **26**, 2474–2476.
6. Knappskog, S., Bjornslett, M., Myklebust, L.M., Huijts, P.E., Vreeswijk, M.P., Edvardsen, H., Guo, Y., Zhang, X., Yang, M., Ylisaukko-Oja, S.K. *et al.* (2011) The MDM2 promoter SNP285C/309G haplotype diminishes Sp1 transcription factor binding and reduces risk for breast and ovarian cancer in Caucasians. *Cancer Cell*, **19**, 273–282.
7. Pardini, B., Kumar, R., Naccarati, A., Prasad, R.B., Forsti, A., Polakova, V., Vodickova, L., Novotny, J., Hemminki, K. and Vodicka, P. (2011) MTHFR and MTRR genotype and haplotype analysis and colorectal cancer susceptibility in a case-control study from the Czech Republic. *Mutat. Res.*, **721**, 74–80.
8. Howie, B.N., Donnelly, P. and Marchini, J. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.*, **5**, e1000529.
9. Marchini, J., Howie, B., Myers, S., McVean, G. and Donnelly, P. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.*, **39**, 906–913.
10. Liu, J.Z., Tozzi, F., Waterworth, D.M., Pillai, S.G., Muglia, P., Middleton, L., Berrettini, W., Knouff, C.W., Yuan, X., Waeber, G. *et al.* (2010) Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat. Genet.*, **42**, 436–440.
11. Su, Z. and Cardin, N., The Wellcome Trust Case Control Consortium, Donnelly, P. and Marchini, J. (2009) A Bayesian method for detecting and characterizing allelic heterogeneity and boosting signals in genome-wide association studies. *Stat. Sci.*, **23**, 430–450.
12. Minichiello, M.J. and Durbin, R. (2006) Mapping trait loci by use of inferred ancestral recombination graphs. *Am. J. Hum. Genet.*, **79**, 910–922.