Chapter 4. Trans-ethnic meta-analysis for inflammatory bowel disease risk loci and population comparisons

4.1 Introduction

Inflammatory bowel disease (IBD) describes chronic inflammatory conditions that affect the gastrointestinal tract. Crohn's disease (CD) and ulcerative colitis (UC) are the two main forms of IBD. In CD, inflammation can occur in patches anywhere along the gastrointestinal tract, while in UC, inflammation occurs continuously and is restricted to the colon. The exact causes of IBD are unknown, though it is likely to involve a disrupted immunological response to gut microbiota in genetically susceptible individuals (Khor *et al.,* 2011). There is currently no known cure, and disease is managed by a combination of immune-suppressing medications, dietary changes or surgery.

The prevalence of IBD in European populations ranges from 26-322 cases per 100,000 for CD and 24-505 per 100,000 for UC (Loftus, 2004; Molodecky *et al.,* 2012). The prevalence of IBD in Asian populations is lower (1-18 per 100,000 for CD; 5-57 per 100,000 for UC) though has been rapidly increasing in recent decades (Molodecky *et al.,* 2012; Prideaux *et al.,* 2012). This increase is hypothesised to be a result of lifestyle changes such as westernisation of diet, improved hygiene, vaccinations and antibiotics use, as well as genetic differences between Europeans and Asians (Prideaux *et al.,* 2012).

In 2012, a GWAS meta-analysis of IBD in ~75,000 European individuals identified 163 loci (representing 193 independent signals) associated with CD, UC or IBD (both CD and UC) at genome-wide significance ($P < 5\times10^{-8}$) (Jostins *et al.,* 2012). Smaller GWAS in populations from Korea, Japan and India (Asano *et al.,* 2009; Juyal *et al.,* 2014; Yamazaki *et al.,* 2013; Yang *et al.,* 2014b) have revealed six associated risk loci at genome-wide significance. Three of these loci overlap with those identified in Europeans (13q12, *FCGR2A* and *SLC26A3*), while the remaining three are nominally associated in Europeans ($P < 5\times10^{-4}$) and also show consistent directions of effect (Jostins *et al.,* 2012). This sharing of risk loci suggest that combining samples from different populations will give greater power to identify risk loci. Nevertheless, despite the much smaller sample sizes (typically a discovery cohort of a few hundred cases), these studies also hinted at genes that differ in their effect on European and Asian IBD. These differences include variants that confer significantly different effect sizes (e.g. *TNFSF15*, HLA), established susceptibility genes with no evidence of associations in East Asians (e.g. *NOD2*, *ATG16L1*), and vice versa (e.g. *ATG16L2*).

Here, I describe a trans-ethnic genetic association study of 10,216 individuals (2,043 CD, 2,801 UC and 5,372 controls) of East Asian, Indian and Indo-European descent and 65,642 European individuals (17,897 CD, 13,768 UC and 33,977 controls – an extension of Jostins *et al.* (2012)) genotyped on the Immunochip. I combined Immunochip data with the Jostins *et al.* GWAS data (5,956 CD, 6968 UC and 21,770 controls) in a transethnic meta-analysis with a total of 96,620 individuals (13,654 European samples were genotyped on both Immunochip and GWAS arrays and removed from the Immunochip cohort). In addition to locus discovery, I also used Immunochip data to compare the effects of IBD risk loci between European and non-European populations in an effort to identify both commonalities and differences in the genetic risk of IBD between the populations.

4.1.1 Contributions

The study design was conceived by the International IBD Genetics Consortium (IIBDGC). Cases and controls were ascertained through the IIBDGC and the

International Multiple Sclerosis Genetics Consortium. Genotyping was performed at various centres described in Jostins *et al.* (2012). Immunochip SNP and sample quality control were performed by Suzanne van Sommeren and Hailiang Huang. Association studies in individual non-European populations on the Immunochip were performed by Suzanne van Sommeren. GWAS QC, meta-analysis and imputation in Europeans were performed by Stephan Ripke and described in Jostins et al. 2012. GRAIL and DAPPLE analyses was performed by Hailiang Huang. Coding variant analyses were performed by Atshushi Takahashi. All other analyses were performed by myself.

4.2 Methods

4.2.1 Sample collection and genotyping

Non-European IBD patients and matched controls were recruited from centres in Japan, China, Hong Kong, South Korea, India, Iran and the UK. Recruitment of European patients and matched controls genotyped on the Immunochip was performed in 15 countries in Europe, North America, Australia and New Zealand. GWAS samples were originally obtained from seven CD and eight UC collections. See Jostins et al. (2012), Anderson et al. (2011) and Franke et al. (2010) for details. Controls consisted of blood donors or population-based studies. IBD diagnosis was based on accepted radiologic, endoscopic and histopathologic evaluations. All included cases fulfil clinical criteria for IBD.

4.2.2 Immunochip quality control

Quality control on Immunochip samples was performed separately for each cohort (European, East Asian, Indian and Iranian). SNP QC consisted of removing SNPs with a low call rate (< 98% across all genotyping batches in the ethnic population, or < 90% in one batch), SNPs that fail Hardy Weinberg equilibrium in controls ($P < 10^{-5}$), SNPs that have heterogeneous allele frequencies among the different genotyping batches within one ethnic population ($P < 10^{-5}$), SNPs that are not present in 1000 genomes phase 1, SNPs with a different missingness rate between cases and controls ($P < 10^{-5}$) and monomorphic SNPs. Following SNP

QC, 108,803 SNPs remained in the East Asian dataset, 146,785 SNPs in the Indian dataset, 153,982 in the Iranian dataset and 143,098 in the European dataset. The fewer number of SNPs in the East Asian cohort is primarily driven by the greater number of monomorphic SNPs. For the sample QC, samples with a low call rate (<98%) and outlying heterozygosity rate ($P < 0.01$) were removed. To identify duplicated and related samples, a subset of SNPs that 1) did not contain SNPs in high-LD regions, 2) have a minor allele frequency (MAF) of <0.05 and 3) pruned for LD ($r^2 < 0.1$), was used to estimate identity by descent. Sample pairs with an identity by descent of >0.8 were considered duplicates, pairs with an identity by descent of >0.4 where considered related. For these pairs, the sample with the lowest genotype call rate was removed.

Principal component analysis (PCA) was performed with the first two PCs estimated from 1000 Genomes Phase I samples and projected onto each of the non-European samples (Price *et al.,* 2006). A clear separation of the populations can be seen, with the samples clustering as expected (Figure 4.1).

After sample QC, 65,642 European (17,897 CD, 13,768 UC and 33,977 controls), 6,543 East Asian (1,690 CD, 1,134 UC and 3,719 controls), 2,413 Indian (184 CD, 1,239 UC and 990 controls) and 1,260 Iranian (169 CD, 428 UC, 663 controls) individuals remained (Table 4.1). Compared with the samples used in Jostins *et al.* (2012), this transethnic study includes an additional 3,548 cases and 16,406 controls (Figure 4.2).

	Immunochip samples			
Population	**CD**	**UC**	**Controls**	**Total**
European	17,897	13,768	33,977	65,642
East Asian	1,690	1,134	3,719	6,543
Indian	184	1,239	990	2,413
Iranian	169	428	663	1,260

Table 4.1. Post-QC patient and control panels genotyped on the Immunochip.

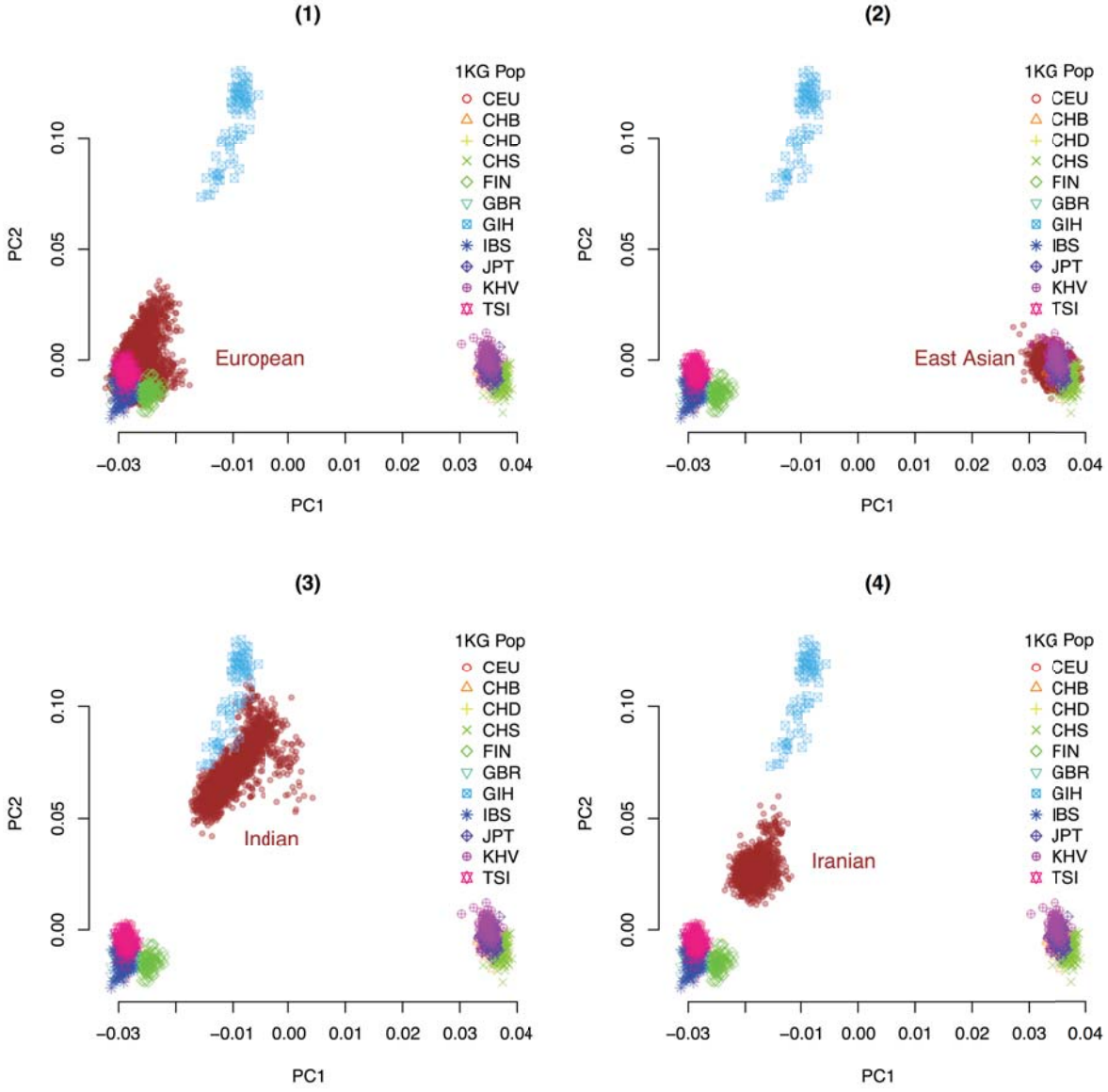

Figure 4.1. Principal components analysis of (1) European, (2) East Asian, (3) Indian and (4) Iranian IBD patients and controls. PCs 1 and 2 are plotted for each cohort as brown circles along with those from the 1000 Genomes Phase I samples.

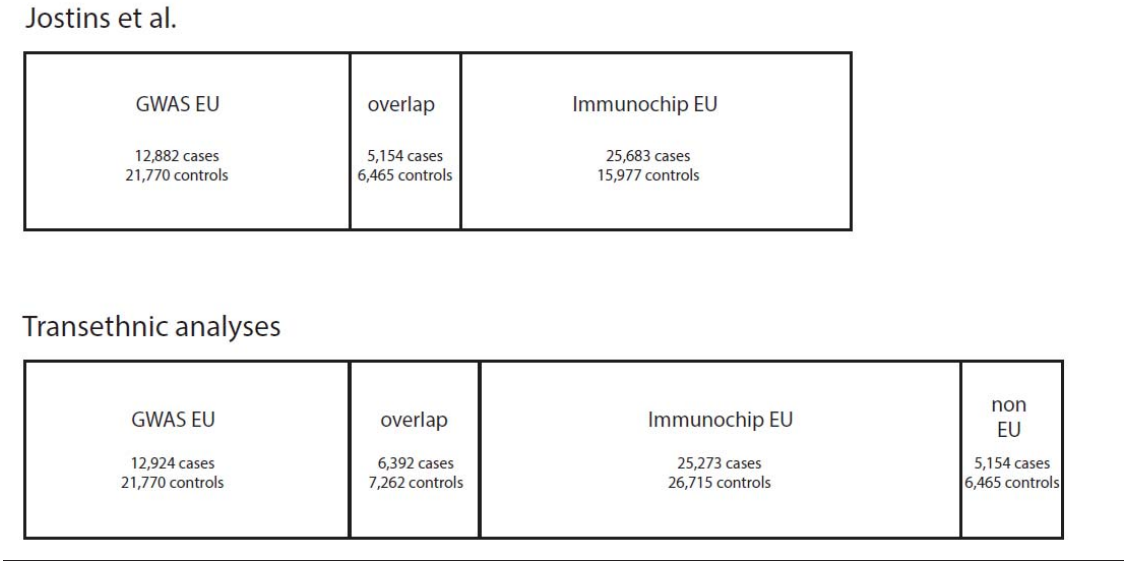

Figure 4.2. Comparison of samples used in this study with those from Jostins et al. (2012).

4.2.3 Per-population association analysis

Case-control association tests per population (European, East Asian, Indian and Iranian) per phenotype (CD, UC and IBD combined) were performed using a linear mixed model implemented in MMM (Pirinen *et al.,* 2012). The random effects component covariance matrix, *R*, was calculated using a set of SNPs with MAF > 0.1, pruned for LD ($r^2 < 0.2$) and showed no evidence of association using logistic regression with 10 PCs as covariates (P > 0.005). A total of ~14,000 SNPs were used for calculating *R* (varies between populations). For European samples, two separate association analyses were performed – one including all European Immunochip individuals (used for population comparisons), and one where 13,654 samples that overlap or are related to GWAS individuals were removed (used in the GWAS Immunochip meta analysis).

4.2.4 Transethnic meta-analysis

For European samples, association results for 1000 Genomes-imputed GWAS and Immunochip individuals (with overlaps removed) were combined using an inverse variance weighted fixed-effects meta-analysis for each of the three phenotypes. These European meta-analysis results were combined with the East Asian, Indian and Iranian association results using MANTRA (Morris, 2011), a

transethnic GWAS meta-analysis method that allows for heterogeneity of effect sizes between distantly related populations. In total, this transethnic meta-analysis was performed on 96,856 individuals and 126,990 SNPs that overlap the Immunochip and GWAS (Table 4.2). Signal intensity plots for all non-HLA loci with P-value < 10^{-7} (in the per-population association tests) or $\log_{10}$ Bayes factor (BF) > 6 in the meta-analysis were visually inspected using Evoker, and SNPs that clustered poorly were removed (Morris *et al.,* 2010).

Significantly associated loci were defined by an LD window of $r^2 > 0.6$ from the most associated SNP in the region with a per-population association P < 5×10^{-8} or $\log_{10}$ BF > 6. Regions less than 250 kb apart from each other were merged into a single associated locus.

Population	CD	CD controls	UC	UC controls	IBD	IBD controls
European GWAS	5,956	14,927	6,968	20,464	12,882	21,770
European Immunochip	14,594	26,715	10,679	26,715	25,273	26,715
Non-European Immunochip	2,043	5,372	2,801	5,372	4,844	5,372
Total	22,593	47,014	20,448	52,551	42,999	53,857

Table 4.2. Post-QC case and control panels used in the transethnic meta-analysis.

Associated loci were classified according to their strength of association with CD, UC or both using a multinomial logistic regression likelihood modelling approach within the Europeans only (Jostins *et al.,* 2012). Four multinomial logistic regression models with parameters β_{CD} and β_{UC} were fitted with the following constraints:

1. CD-specific model: $\beta_{UC} = 0$ (1 d.f.)

2. UC-specific model: $\beta_{CD} = 0$ (1 d.f.)

3. IBD unsaturated model: $\beta_{CD} = \beta_{UC} = \beta_{IBD}$ (1 d.f.)

A fourth unconstrained model with 2 d.f. was also estimated with β_{CD} and β_{UC} both fitted by maximum likelihood. Log-likelihoods were calculated for each model, and three likelihood-ratio tests were performed comparing models 1-3 against the unconstrained model. If the P-values of all three tests were less than

0.05, the SNP was classified as associated with both CD and UC but with evidence of different effect sizes. Otherwise, of the three constrained models, the SNP was classified according to the model with the largest likelihood. If 'IBD unsaturated' is the best fitting model the locus can be interpreted as associated with both CD and UC but with no evidence for different effect sizes.

4.2.5 Gene prioritisation

Two functional annotations: coding variants and expression quantitative trait loci (eQTLs), and two network approaches: GRAIL (Raychaudhuri *et al.,* 2009) and DAPPLE (Rossin *et al.,* 2011), were used to prioritise candidate genes within novel associated loci. Coding SNPs were identified if a missense or nonsense SNP was in high LD ($r^2 > 0.8$) with a lead SNP in either the 1000 Genomes Phase 1 European (CEU, FIN, GBR and IBS samples) or East Asian (CHB, CHS and JPT samples) populations (Genomes Project *et al.,* 2012). Expression quantitative trait loci were collated from the University of Chicago eQTL browser (http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl). New IBD associated SNPs with $r^2 > 0.8$ (1000 Genomes European or East Asian cohort) with a known eQTL were reported.

4.2.6 Variance explained

The proportion of variance explained by each associated locus per population was calculated using a liability threshold model (So *et al.,* 2011) assuming a disease prevalence of 500 per 100,000 and log-additive disease risk.

4.2.7 Heterogeneity of effect sizes and allele frequencies between populations

For an associated SNP, differences in the effect size between two populations were tested using a t-test for a significant difference in log odds ratios (ORs). Overall heterogeneity between all four populations was assessed using Cochran's Q test, and the percentage of differences in ORs due to heterogeneity rather than chance was evaluated using the fixed effects I^2 statistic (Higgins and Thompson, 2002). Fixation index (F_{st}) values for a SNP between two populations were

calculated using the Weir and Cockerham method on allele frequencies in control samples only (Weir and Cockerham, 1984).

4.2.8 Genetic correlation

The proportion of genetic variation tagged by Immunochip SNPs that is shared between pairs of each of four populations was estimated using the bivariate linear mixed-effects model implemented in GCTA (Lee *et al.,* 2012). The method uses high-density SNP data to estimate genetic similarities between pairs of individuals to estimate covariance components (r_G) of the mixed model. I applied the method across Immunochip individuals for all pairwise combinations of population comparisons for CD and UC with 20 PCs as covariates, assuming a disease prevalence of 0.005. To test whether r_G is significantly different from 0 (or 1), r_G was fixed at 0 (or 1) and a likelihood ratio test comparing this constrained model with the unconstrained model was applied. An r_G of 0 means that no genetic variants are shared between the two populations, while a value of 1 means that all the genetic variance tagged in one population is shared with the other. In Europeans, only 10,000 cases and 10,000 controls (selected at random) were included due to computation limitations, while all non-Europeans samples were included.

4.2.9 Gene-based likelihood ratio test

Due to the much larger sample sizes, there is greater power to detect loci with multiple independent signals in Europeans than the non-European populations. However, if these independent SNPs within a locus are also associated in a non-European population, there may be greater power to detect these signals by jointly modelling them in the non-European population rather than single-SNP tests. To investigate this, I describe an approach that 1) identifies independently associated SNPs among SNPs within the Immunochip high-density regions in the European cohort, 2) assign the independently associated SNPs to genes, and 3) for genes with multiple associated SNPs, tests these SNPs jointly in a per-gene manner for association in a non-European cohort.

1) Independently associated SNPs were identified using the conditional and joint multi-SNP approach implemented in GCTA (GCTA-COJO) (Yang *et al.,* 2012). GCTA-COJO uses summary association statistics and LD information from a reference panel to approximate independently associated signals. GCTA-COJO was applied to CD, UC and IBD summary statistics from the Immunochip European analysis using the same European individuals as the reference panel. A joint association $P < 5\times10^{-6}$ and $r^2 < 0.9$ were used as cut-offs for assigning independent signals. It has been shown that the LD-based approximation approach of GCTA-COJO generates almost identical results to conditional logistic regression when the individual genotypes used in the association study and the reference panel are identical, as was the case in this study (Yang *et al.,* 2012). Significant independently associated SNPs that were identified via this approach were taken forward.

2) The independently associated SNPs identified in 1) were grouped according to their proximity to genes. A SNP was assigned to a gene if it lies within ±50kb of that gene's transcript start/stop positions (GENCODE 17 definitions) (Harrow *et al.,* 2012). Due to some genes overlapping each other, some SNPs may be assigned to multiple genes. Genes with more than one assigned SNP were taken forward for joint modelling in the non-European cohorts.

3) For a gene where more than one independently associated SNP was identified, the *K* independent SNPs were modelled jointly in a multiple logistic regression model (for the phenotype in which it was originally identified in) in each of the non-European populations and the total log-likelihood for the model calculated. I then performed a likelihood ratio test (with $K - 1$ degrees of freedom) comparing the log-likelihoods of this joint model with *K* SNPs and one from a null model without SNP effects. Genes with P-values less than 5×10^{-5} (equivalent to a 5% Bonferroni correction for ~1000 genes – roughly the number tagged by SNPs on the Immunochip) were considered statistically significant.

4.3 Results and discussion

4.3.1 Per-population association and transethnic meta-analysis

Per-population association analysis and the meta-analysis across all populations identified 40 novel risk loci at genome-wide significance (MANTRA $\log_{10}$ BF > 6 or per-population association $P < 5\times10^{-8}$ in at least one of the phenotypes) (Table 4.3). Likelihood modelling classified eight of these to be only associated with CD, four with UC, and 28 with IBD (both UC and CD). Of the 28 IBD loci, eight showed significant evidence of different CD/UC effect sizes (Table 4.3). Owing to the much larger sample sizes, 25 of the 40 novel loci were genome-wide significant in Europeans alone. Indeed, only three loci showed stronger evidence of association in a non-European population than European (rs10774482: IBD European P = 0.30, Iranian $P = 2.17\times10^{-7}$, Indian $P = 1.12\times10^{-3}$; rs2072711: CD European P = 7.51×10-3, East Asian $P = 2.17\times10^{-7}$; rs6856616: IBD European P = 9.72×10^{-7}, East Asian $P = 1.33\times10^{-7}$). Of these, rs6856616 was previously reported as a novel CD risk locus in a GWAS in Korean individuals (Yang *et al.,* 2014b).

The strongest signal in the European-only analysis was rs395157 (IBD $P = 2.22\times10^{-20}$). The magnitude of this association was unexpectedly high, given that the number of Europeans in this study was only modestly greater than that of Jostins et al. (2012) (86,640 vs. 76,312), such that this SNP should have exceeded genome-wide significance and reported in the previous study. The reason why this was not originally reported in Jostins et al. was a result of an error in the GWAS and Immunochip meta-analysis, where discordant alleles were merged (and effects cancelled out). This was due to the SNP having an allele frequency very close to 0.5, such that the minor allele of the GWAS and Immunochip were different. No other associated signals appeared to be affected by this issue.

Chr.	SNP	Base pair position	[a]Best trait	[b]LR trait	[c]Log_{10} BF	[d]Het I^2	Eur. OR	Eur. P	Eas. OR	Eas. P	Ind. OR	Ind. P	Ira. OR	Ira. P
1	rs1748195	62822181	CD	CD	6.08	0	1.07	7.13×10^{-8}	1.04	0.41	1.11	0.36	1.05	0.73
1	rs34856868	92326871	IBD	IBD_U	6.16	0	0.82	9.80×10^{-9}	0.11	0.43	1.47	0.34	1.36	0.69
1	rs11583043	101238642	UC	IBD_U	8.34	66.51	1.08	6.05×10^{-8}	1.18	0.032	1.27	3.80×10^{-3}	1.46	9.80×10^{-3}
1	rs6025	167785673	IBD	IBD_U	6.43	0	0.84	2.51×10^{-8}	-	-	0.81	0.41	0.7	0.31
1	rs10798069	185142082	CD	IBD_S	7.24	0	0.93	4.25×10^{-9}	0.94	0.12	1.06	0.59	1.01	0.92
1	rs7555082	196865286	CD	IBD_U	7.97	0	1.13	1.47×10^{-10}	0.6	0.67	1.02	0.92	0.85	0.44
2	rs11681525	145208852	CD	CD	8.8	59.3	0.86	4.08×10^{-11}	-	-	1.5	0.12	0.69	0.22
2	rs4664304	160502254	IBD	IBD_U	6.34	0	1.06	2.61×10^{-8}	1.01	0.77	1.04	0.51	1.18	0.12
2	rs3116494	204300266	UC	IBD_S	7.03	0	1.08	1.30×10^{-7}	1.17	0.1	1.21	0.043	1.19	0.15
2	rs111781203	228368356	IBD	IBD_U	10.04	0	0.94	2.16×10^{-10}	0.91	0.031	0.88	0.033	0.98	0.84
2	rs35320439	242386014	CD	IBD_S	7.71	0	1.09	9.89×10^{-10}	1.04	0.37	1.07	0.54	1.03	0.81
3	rs113010081	46432416	UC	IBD_U	7.45	0	1.14	9.02×10^{-10}	0.02	0.5	0.84	0.38	1.12	0.71
3	rs616597	103052416	UC	UC	6.68	54.68	0.93	9.34×10^{-6}	0.85	1.04×10^{-3}	0.84	0.029	0.79	0.044
3	rs724016	142588260	CD	CD	7.41	70.87	1.06	3.36×10^{-6}	1.21	5.56×10^{-6}	1.13	0.3	0.97	0.86
4	rs2073505	3414301	IBD	IBD_U	6.87	0	1.1	1.46×10^{-7}	1.14	6.83×10^{-3}	1.04	0.62	0.95	0.76
4	rs4692386	25741459	IBD	IBD_U	6.47	0	0.94	1.21×10^{-8}	0.97	0.49	0.98	0.7	0.9	0.27
4	rs6856616	38001431	IBD	IBD_U	9.78	61.59	1.1	9.72×10^{-7}	1.24	1.33×10^{-7}	1.07	0.35	1.18	0.31
4	rs2189234	106294947	UC	UC	8.85	0	1.08	1.95×10^{-10}	1.11	0.033	0.98	0.76	1.06	0.61
5	rs395157	38903489	IBD	IBD_U	19.5	0	1.1	2.22×10^{-20}	1.09	0.027	1.12	0.065	0.99	0.93
5	rs4703855	71729655	IBD	IBD_U	6.83	70.26	0.93	7.16×10^{-11}	1	0.97	1.04	0.52	1.15	0.18
5	rs564349	172257584	IBD	IBD_U	8.12	37.54	1.06	1.54×10^{-7}	1.15	1.54×10^{-4}	1.09	0.22	1.07	0.51
6	rs7773324	327559	CD	IBD_U	7.67	0	0.92	1.06×10^{-9}	0.97	0.53	0.88	0.27	1	0.98
6	rs13204048	3365405	CD	IBD_S	7.23	53.54	0.93	2.89×10^{-8}	0.94	0.13	0.6	3.23×10^{-3}	0.97	0.85
6	rs7758080	149618772	CD	IBD_S	7.88	0	1.08	7.27×10^{-9}	1.11	0.017	1.06	0.62	0.93	0.63
7	rs1077773	17409204	UC	UC	5.86	76.72	0.93	5.96×10^{-9}	1.11	0.053	1.01	0.85	1.05	0.66
7	rs2538470	147851381	IBD	IBD_U	10.93	54.64	1.07	3.00×10^{-11}	1.15	9.78×10^{-4}	0.97	0.63	1.22	0.059
8	rs17057051	27283471	IBD	IBD_U	6.74	15.92	0.94	5.50×10^{-8}	0.9	0.022	1.02	0.7	0.87	0.16
8	rs7011507	49291795	UC	IBD_U	7.49	39.32	0.9	6.40×10^{-8}	0.82	7.42×10^{-4}	0.94	0.47	1.13	0.43
10	rs3740415	104222706	IBD	IBD_U	6.26	0	0.95	1.03×10^{-7}	0.93	0.073	0.98	0.75	1	0.99
12	rs10774482*	971525	IBD	CD	6.02	91.3	1.01	0.3	1	0.97	1.21	1.12×10^{-3}	1.63	2.17×10^{-7}
12	rs7954567	6361386	CD	CD	8.25	0	1.09	1.30×10^{-9}	1.17	0.076	1.12	0.35	1.12	0.47
12	rs653178	110492139	IBD	IBD_U	6.57	49.67	1.06	1.11×10^{-8}	0.02	0.042	1.15	0.13	0.97	0.72
12	rs11064881	118631308	IBD	IBD_U	7.02	31.65	1.1	5.95×10^{-8}	0.01	0.29	1.22	0.053	1.4	0.03
13	rs9525625	41916030	CD	CD	8.55	37.25	1.08	1.41×10^{-9}	1.07	0.22	1.11	0.34	1.46	7.08×10^{-3}
17	rs3853824	52235992	CD	IBD_S	8.46	50.42	0.92	1.17×10^{-10}	0.95	0.32	0.88	0.29	1.31	0.066
17	rs17736589	74248713	UC	UC	6.53	53.41	1.09	4.34×10^{-8}	1.05	7.30×10^{-5}	1.03	0.73	1.34	0.026
18	rs9319943	55030807	CD	CD	6.33	33.39	1.08	9.05×10^{-7}	1.19	2.03×10^{-3}	0.95	0.69	1.21	0.22
18	rs7236492	75321604	CD	IBD_S	6.6	0	0.91	9.09×10^{-9}	1.44	0.68	1.14	0.62	0.84	0.64
22	rs2072711	35598501	CD	IBD_S	6.12	91.56	0.96	7.51×10^{-3}	1.26	2.17×10^{-7}	1	0.98	1.28	0.17
22	rs727563	40197323	CD	CD	7.1	76.01	1.1	1.88×10^{-10}	0.95	0.23	0.93	0.52	0.93	0.61

Table 4.3. Table of novel IBD risk loci from MANTRA transethnic meta-analysis or individual per-population analyses. [a]Phenotype with the largest MANTRA Bayes factor. [b]Likelihood modelling classification. IBD_S and IBD_U refer to IBD saturated and unsaturated respectively. [c]MANTRA log10 Bayes factor. [d]Heterogeneity I2 percentage. Per-population ORs and P-values refer the Best trait column.

4.3.2 Candidate genes

Candidate genes for each of the novel loci were identified using two SNP annotations: coding SNPs, known eQTLs, and two network approaches: GRAIL and DAPPLE. These methods identified at least one candidate gene in 28 of 40 novel risk loci, four of which harbour genes identified by multiple methods (Table 4.4 A-B). Including the new 40 loci in GRAIL and DAPPLE analyses with known IBD risk loci revealed additional candidate genes with significant connectivity scores ($P < 0.05$ in either GRAIL and DAPPLE) at 34 of the 163 known loci that weren't reported in Jostins et al. (Table 4.5). A visual inspection of the GRAIL network plot reveals the interconnectedness between the novel and known IBD risk loci (Figure 4.3).

Many of the genes associated with IBD highlight the importance of T cells in IBD pathogenesis. T cells are an integral component in the adaptive immune response, and become activated in response to MHC-bound antigens via signalling through the T cell receptor. This process depends on PRKCQ signalling, which results in increased expression of CD44. Co-stimulation via other ligands such as CD28, CD81 and CD27 are also required for T cells to generate memory. Impaired immune responses may occur from inappropriate co-stimulation, and is characterised by increased expression of PDCD1. Other processes that can impair immune responses also include apoptosis (implicating UBASH3A) and recruitment of immunosuppressive regulatory T cells, driven partly by the chemokine CCL20. The genes mentioned are all within loci associated with IBD risk from this study and others, highlighting the importance of genetic risk factors in T cell responses in IBD pathogenesis, and may provide targets for development of future therapies.

Chr.	SNP	Cis-eQTL	Nonsynonymous coding	GRAIL	DAPPLE	Genes implicated by multiple methods
1	rs1748195	*DOCK7,AF086387, ANGPTL3*				
1	rs34856868		*BTBD8*			
1	rs11583043			*EDG1*		
1	rs6025			*SELP, SELE, SELL*		
1	rs10798069			*PTGS2,PLA2G4A*		
1	rs7555082			*PTPRC*		
2	rs4664304	*LY75*	*PLA2R1*	*LY75*		*LY75*
2	rs3116494			*ICOS ,CD28, CTLA4*		
2	rs111781203			*CCL20*		
2	rs35320439			*PDCD1, ATG4B*		
3	rs113010081			*FLJ78302,LTF,CCR1, CCR3,CCR5*	*CCR2*	
3	rs616597			*NFKBIZ*		
4	rs2073505		*HGFAC*			
5	rs395157			*OSMR,FYB*	*LIFR, OSMR*	*OSMR*
5	rs564349			*DUSP1*		
6	rs7773324			*IRF4,DUSP22*		
6	rs7758080			*MAP3K7IP2*		
7	rs1077773			*AHR*		
7	rs2538470	*CNTNAP2*				
8	rs17057051	*PTK2B*		*PTK2B*		*PTK2B*
10	rs3740415	*PDCD11,TMEM180, ACTR1A*		*NFKB2*		
12	rs7954567			*CD27,TNFRSF1A,LTBR*		
12	rs653178			*SH2B3*		
12	rs11064881	*PRKAB1*				
13	rs9525625			*TNFSF11*		
18	rs7236492			*NFATC1*		
22	rs2072711	*CSF2RB*	*NCF4*	*CSF2RB*	*IL2RB,CSF2RB*	*CSF2RB*
22	rs727563	*MEI1,PHF5A, NFP2L1,TOB2*				

Table 4.4A. Candidate genes implicated by coding variants, eQTLs, GRAIL and DAPPLE in 28 of the 40 novel IBD risk loci.

Chr.	SNP	eQTL SNP	LD (r^2)	Gene	Type	Tissue
1	rs1748195	rs1748195	1	*DOCK7*	Cis	Monocytes
		rs10889353	0.99	*AF086387*	Cis	Liver
		rs1168089	1	*ANGPTL3*	Cis	Liver
2	rs4664304	rs7601374	0.97	*LY75*	Cis	Liver
8	rs17057051	rs17057051	1	*PTK2B*	Cis	Monocytes
10	rs3740415	rs3740415	1	*PDCD11*	Cis	LCLs
		rs7342070	0.98	*TMEM180*	Cis	Liver
		rs5870	0.93	*ACTR1A*	Cis	LCLs
12	rs11064881	rs11064881	1	*PRKAB1*	Cis	LCLs
		rs11064881	1	*PRKAB1*	Cis	Monocytes
22	rs2072711	rs2072711	1	*CSF2RB*	Cis	LCLs
22	rs727563	rs12165508	1	*MEI1*	Cis	LCLs
		rs203319	0.99	*PHF5A*	Cis	Monocytes
		rs202628	0.96	*NHP2L1*	Cis	Liver
		rs202614	0.94	*TOB2*	Cis	Liver

Table 4.4B. Known eQTLs tagged by novel IBD associated SNPs. eQTL SNPs, gene, eQTL type (cis or trans) and tissue studied were extracted from publications collated in the University of Chicago eQTL Browser (http://eqtl.uchicago.edu/cgi-bin/gbrowse/). LD (r^2) values were extracted from the European and East Asian cohorts of the 1000 Genomes Project Phase I (the larger r^2 of the two cohorts are reported).

Chr.	SNP	New GRAIL	New DAPPLE	[a]Uniquely new
1	rs35675666	*PARK7, ERRFI1*		*PARK7, ERRFI1*
1	rs6426833	*PLA2G2A*		*PLA2G2A*
1	rs4845604		*S100A11*	*S100A11*
1	rs3024505	*IL10*		
2	rs7608910	*REL*		
2	rs10865331		*COMMD1*	*COMMD1*
2	rs2382817	*IL8RA, IL8RB, IL8RBP*		*IL8RA, IL8RB, IL8RBP*
5	rs7702331		*BTF3*	*BTF3*
5	rs2188962	*RAD50*	*RAD50, IL5*	*RAD50, RAD50*
6	rs3851228		*FYN*	
6	rs212388	*TAGAP*	*EZR*	*TAGAP ,EZR*
7	rs10486483	*SKAP2*		*SKAP2*
7	rs1456896	*IKZF1*		*IKZF1*
7	rs9297145	*SMURF1*		*SMURF1*
8	rs7015630	*NBN*		*NBN*
8	rs1991866		*FAM49B*	*FAM49B*
9	rs4743820		*SYK*	*SYK*
10	rs2227564	*PLAU*	*VCL*	*PLAU, VCL*
11	rs10896794		*ZFP91*	*ZFP91*
11	rs11230563	*GPR44*		*GPR44*
11	rs2231884	*SIPA1*		*SIPA1*
12	rs11612508	*DUSP16*		*DUSP16*
12	rs11168249	*RAPGEF3, SENP1*		*RAPGEF3, SENP1*
12	rs7134599		*IL22, IL26*	
13	rs9557195	*EBI2*		*EBI2*
15	rs17293632		*SMAD3*	
17	rs2945412	*NOS2A*		*NOS2A*
17	rs3091316	*CCL1,CCL7*		*CCL1, CCL7*
18	rs1893217	*PTPN2*		*PTPN2*
18	rs727088	*DOK6*		*DOK6*
19	rs11879191	*ICAM3*		*ICAM3*
19	rs17694108		*CEBPG*	
19	rs4802307		*CALM3*	*CALM3*
19	rs1126510	*PTGIR*		*PTGIR*
20	rs6142618	*HCK*		*HCK*
20	rs4911259		*COMMD7*	*COMMD7*
20	rs6088765	*PROCR*		*PROCR*
20	rs913678	*PTPN1, TMEM189-UBE2V1*		*PTPN1, TMEM189-UBE2V1*
21	rs2284553		*IL10RB, IFNAR2*	
21	rs7282490	*AIRE*		*AIRE*
22	rs2266959		*MAPK1*	
22	rs2413583	*MAP3K7IP1*		*MAP3K7IP1*

Table 4.5. New genes in known IBD risk loci implicated from GRAIL and DAPPLE network analyses. [a]New genes that weren't previously implicated by either GRAIL or DAPPLE

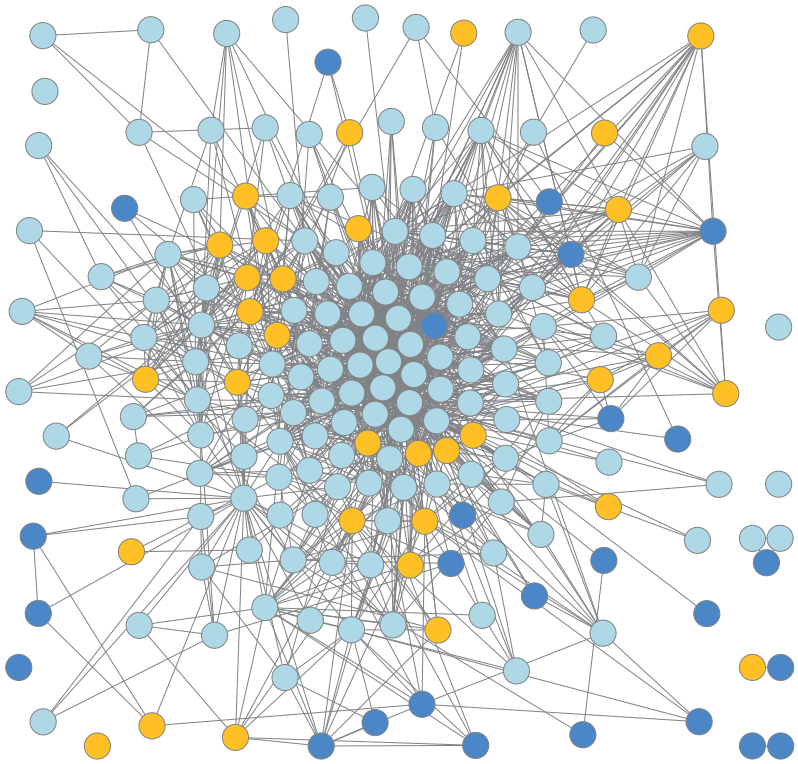

Figure 4.3. GRAIL network for all genes with GRAIL P < 0.05. Yellow nodes represent newly associated genes, light blue nodes represent known genes, dark blue genes represents new genes in known loci that now reach GRAIL P < 0.05 after including the novel loci.

4.3.3 Validation of known loci

Of the 163 IBD risk loci identified in Jostins et al. (2012), all but 16 exceeded genome-wide significance ($P < 5\times10^{-8}$) in the European only analysis here. Fifteen of these loci continue to show suggestive levels of significance ($P < 1.44\times10^{-6}$). This is equivalent to a false discovery rate of < 0.001, and not beyond what's expected given the initially reported P-values for these SNPs in Jostins et al. ($3.60\times10^{-9} < P < 3.71\times10^{-8}$) and the sampling variability in replication vs. discovery P-values (Lazzeroni *et al.,* 2014). However, one SNP, rs2226628, fell to P = 0.0023 in this analysis, suggesting that this may have been an initial false positive report, and larger samples will be required to unequivocally implicate this locus. Nevertheless, as expected, the majority of signals (107/163) become more significant with the additional European samples (Figure 4.4).

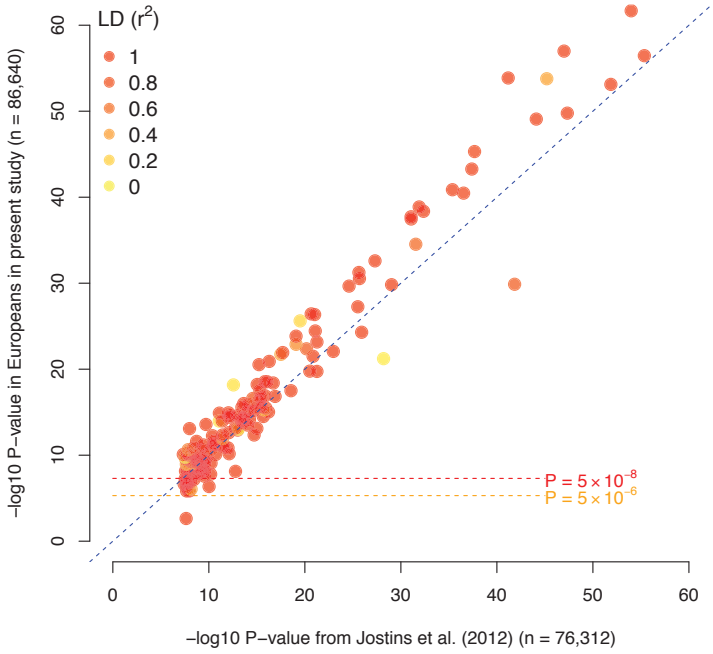

Figure 4.4. Comparison of association P-values reported in Jostins *et al.* (2012) and Europeans in this present study. LD r^2 values are between the SNP reported here and that from Jostins et al. Different SNPs may have been reported if there was stronger signal was found in this study or the previously reported SNP was removed during QC. The blue dashed line represents y = x.

The discrepancy in the rs2226628 GWAS + Immunochip meta-analysis between our study and Jostins et al. is driven almost entirely by the Immunochip samples (Immunochip IBD Jostins et al. $P = 7.52 \times 10^{-7}$ vs. P = 0.012 in this study). Several factors may be driving this discrepancy. Firstly, in the Jostins et al. study, it was later found that ~1,200 samples were mistakenly included in both the initial GWAS and the subsequent Immunochip replication effort. This may have led to an inflation of the P-values for rs2226628 and other SNPs, for which we have now corrected in this latest analysis. Another factor may be the different association methods used on Immunochip samples. In Jostins *et al.,* association was performed using logistic regression with 4 PCs as covariates, while in this study, we applied a linear (logistic) mixed model. If the SNP shows within-European population stratification that was not adequately captured by the first 4 PCs, then this may have also lead to an inflated P-value. Indeed, this SNP does appear to show varying frequencies across the European populations in the 1000

Genomes data (MAF = 0.2 in GBR to 0.47 in FIN) (Genomes Project *et al.,* 2012). In our Immunochip samples, using logistic regression with 4 PCs as covariates did result in this SNP being more significant than the mixed model (P = 0.012 vs. P = 1.85×10^{-4}), though it still did not reach the same level of significance as that of Jostins et al. Finally, the Jostins et al. meta-analysis was performed using two different SNPs – the final reported P-value was a meta-analysis of rs6592362 from the GWAS cohort and rs2226628 from the Immunochip samples. This was done since the original GWAS hit SNP, rs6592362, was not present on Immunochip and rs2226628 was selected as it was the best tag (r^2 = 0.50). In this study, I only combined GWAS and Immunochip at rs2226628, though would have achieved a more significant signal had I combined the two different SNPs (P = 7.38×10^{-6}). Notably, rs2226628 is non-significant in the GWAS (P = 0.08), and it may be the case that combining two different SNPs that are only in moderate LD with each other did not reflect the true signal in this region (if there is one).

4.3.4 Population comparisons

Recent large-scale transethnic genetic studies of complex diseases have shown that the majority of risk loci originally identified in Europeans are shared across other populations (Dastani *et al.,* 2012; Okada *et al.,* 2014; Replication *et al.,* 2014; Teslovich *et al.,* 2010). The true extent of sharing is difficult to characterise as the GWAS sample sizes in non-European populations are often much smaller than their European counterparts, limiting power to detect associated loci. Despite this study including over 10,000 non-European samples and being the largest non-European study of its type, this still pales in comparison with the European sample size of over 85,000. As such, we expect that the majority of known risk loci will not replicate in the non-European populations at genome-wide significance. Nevertheless, there were significant trends both in terms of directions of effect and strength of the correlation across all three phenotypes when comparing the 233 independently associated SNPs in Europeans and the individual non-European populations (Figure 4.5).

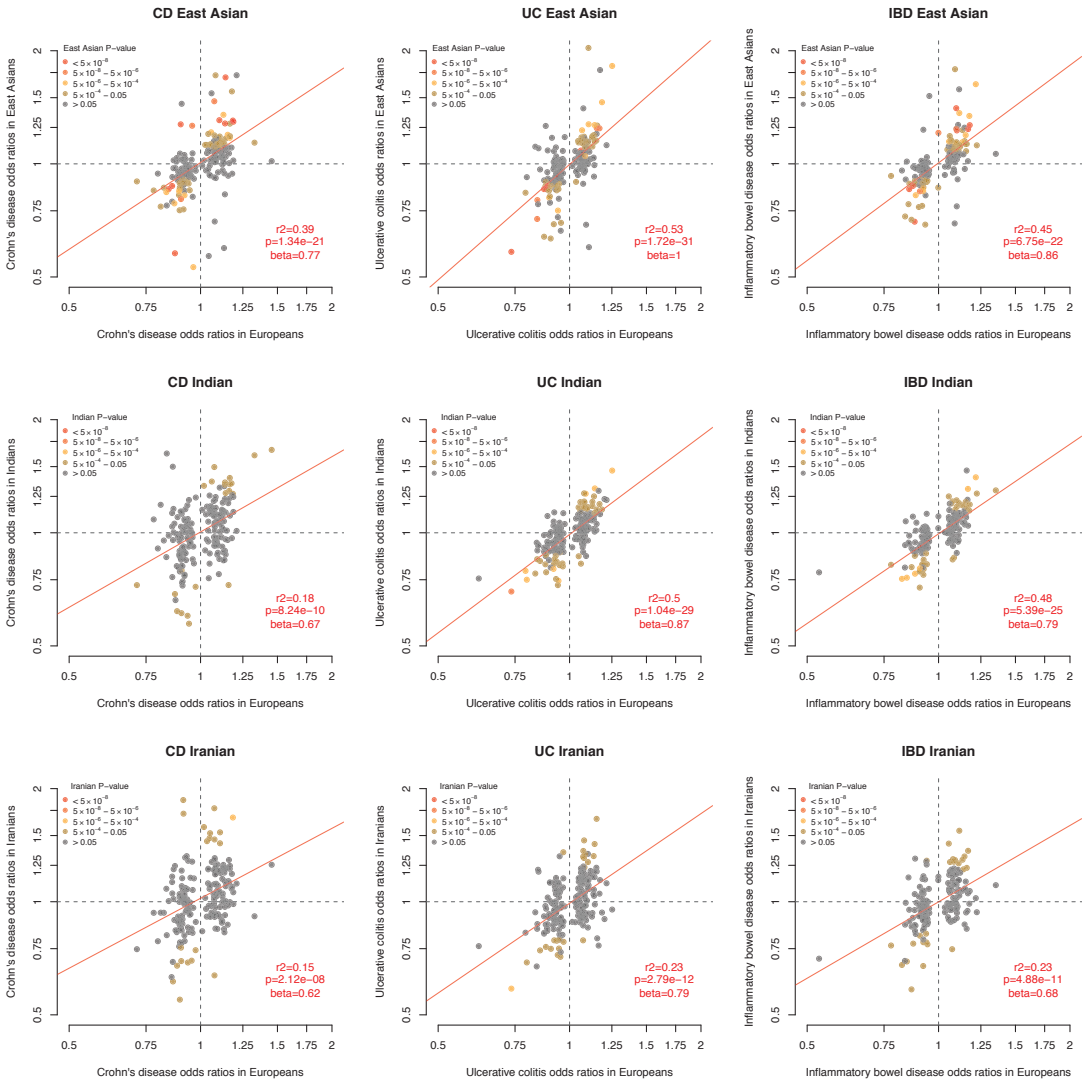

Figure 4.5. Odds ratio comparison between European and non-European populations at 233 SNPs associated with CD, UC other both. For each SNP, ORs (on log-scale) were taken for the corresponding phenotype in the European and non-European population if it was classified as associated with that phenotype in the likelihood modelling (section 4.2.4). Points are coloured according to the strength of association for the respective phenotype in the non-European population. The red line indicates the best-fitting linear regression line, weighted by the inverse variance of the log ORs in the non-European population. Regression coefficients, significance and goodness of fit are listed in the bottom right corner of each plot.

Consistent with the concordant effect sizes at associated SNPs, there were high genetic correlations (r_G) when considering all SNPs on the Immunochip for all pairwise population comparisons (Table 4.6). Estimates of r_G ranged from 0.42 (between East Asian and Indian CD) to 0.92 (between Indian and Iranian CD). Given that rare SNPs are more likely to be population-specific, high r_G values

also support the notion that the majority of causal variants are common. It is also unsurprising that r_G is significantly smaller than 1 for all pairwise comparisons (apart for those involving Iranian CD, though with only 169 cases, this most likely reflects lack of power) as there are examples of IBD risk loci that are not present in some populations, or where there are differences in effect size between populations (discussed below). Nevertheless, r_G is significantly greater than 0 ($P < 0.021$) for all pairwise population comparisons across both CD and UC. Together, these results indicate that a large proportion of IBD risk loci are shared across different populations, though accurate assessments of the actual number of shared loci and their effect sizes will require much larger sample sizes.

Phenotype	Population 1	Population 2	r_G	Standard Error	P-value (H_1: $r_G > 0$)	P-value (H_1: $r_G < 1$)
Crohn's disease	East Asian	Indian	0.42	0.13	8.02×10^{-4}	3.45×10^{-4}
	East Asian	Iranian	0.73	0.26	8.56×10^{-4}	0.223
	European	East Asian	0.76	0.04	0	4.47×10^{-14}
	European	Indian	0.56	0.09	6.58×10^{-10}	3.43×10^{-4}
	European	Iranian	0.82	0.34	5.06×10^{-7}	0.357
	Indian	Iranian	0.92	0.63	0.0209	0.456
Ulcerative colitis	East Asian	Indian	0.83	0.08	0	0.011
	East Asian	Iranian	0.56	0.12	1.37×10^{-5}	4.59×10^{-4}
	European	East Asian	0.79	0.04	0	6.61×10^{-9}
	European	Indian	0.84	0.05	0	8.23×10^{-4}
	European	Iranian	0.67	0.08	2.61×10^{-15}	6.75×10^{-4}
	Indian	Iranian	0.53	0.14	1.11×10^{-4}	2.64×10^{-3}

Table 4.6. Pairwise genetic correlation (rG) tagged by Immunochip SNPs.

While there was significant correlation in the effect sizes of IBD loci between different populations, identifying loci that differ in their effects between populations may reveal differences in disease pathogenesis. As discussed, a comprehensive comparison of effect sizes will require much larger sample sizes in non-Europeans than the one in this study. However, there was sufficient power to detect genetic heterogeneity between our East Asian and European cohorts at several alleles with reported large effect size in Europeans. For instance, consistent with previous genetic studies of Crohn's disease in East Asians (Ng *et al.,* 2012), the three coding variants in *NOD2* (nucleotide-binding oligomerisation domain-containing protein 2) with the largest effect sizes in

Europeans are all monomorphic in East Asians. Furthermore, across all NOD2 variants, no association signals were observed in the East Asian cohort beyond what is expected under a null distribution given the number of SNPs (83) assayed in this region on the Immunochip (minimum P = 7.18×10^{-4}). Similarly, at the *IL23R* (interleukin 23 receptor) gene, previous studies have shown that the most associated variants in Europeans are either monomorphic or do not appear to be associated in East Asians, though there is evidence of additional variants in *IL23R* that are associated in East Asians (Ng *et al.,* 2012). In line with these observations, the *IL23R* SNP with the largest effect in European CD and UC (rs11209026) is monomorphic in East Asians, while two secondary *IL23R* variants observed in Europeans were also non-significant (rs6588248, P = 0.65; rs7517847, P = 0.04) in East Asian IBD. Nevertheless, there was strong evidence for an association at rs76418789 with both CD and UC in East Asians (IBD P = 1.83×10^{-13}). The same variant was previously implicated in a GWAS of CD in Koreans (Yang *et al.,* 2014). This variant demonstrates suggestive evidence of association in European IBD (P = 3.99×10^{-6}, OR = 0.66), though has a much lower allele frequency than in East Asian populations (MAF = 0.004 vs. 0.07).

The identification of CD risk variants in *ATG16L1* (autophagy-related protein 16-1), first implicated autophagy as an important process in CD pathogenesis (Hampe *et al.,* 2007; Parkes *et al.,* 2007; Rioux *et al.,* 2007). At *ATG16L1*, the variant most strongly associated with Crohn's disease in Europeans (rs12994997) has a risk allele frequency (RAF) of 0.53 and OR of 1.27. The variant shows no evidence of association in East Asians, (P = 0.21), driven at least in part by a significant difference in allele frequency (RAF = 0.24, F_{st} = 0.15). However, assuming the effect size at this SNP in the East Asian cohort was equal to that seen in the European cohort, there would have more than 80% power to detect association of suggestive significance (P < 5×10^{-5}) in this study. Indeed, there was also evidence for heterogeneity of odds at this SNP (East Asian OR = 1.06; P = 8.45×10^{-4}). Association in European individuals to a locus containing *IRGM* further implicated authophagy in IBD risk, and the most associated SNP at this locus in Europeans shows only nominally significant evidence of association in East Asian CD (rs11741861, European P = 5.89×10^{-44}, East Asian P = $2.62\times10^{-}$

^{3}) as well as evidence of heterogeneity of effect (European OR = 1.33 vs. East Asian OR = 1.13; heterogeneity P = 1.2×10^{-3}). Given these results it is tempting to speculate that autophagy plays a lesser role in East Asian IBD compared to European IBD. However, a previous GWAS in a Japanese population identified suggestive evidence of association near another autophagy-related gene, *ATG16L2* (Yamazaki *et al.,* 2013), though this finding was unable to be confirmed because the reported variant (rs11235667) is monomorphic in Europeans and the locus is not covered on the Immunochip.

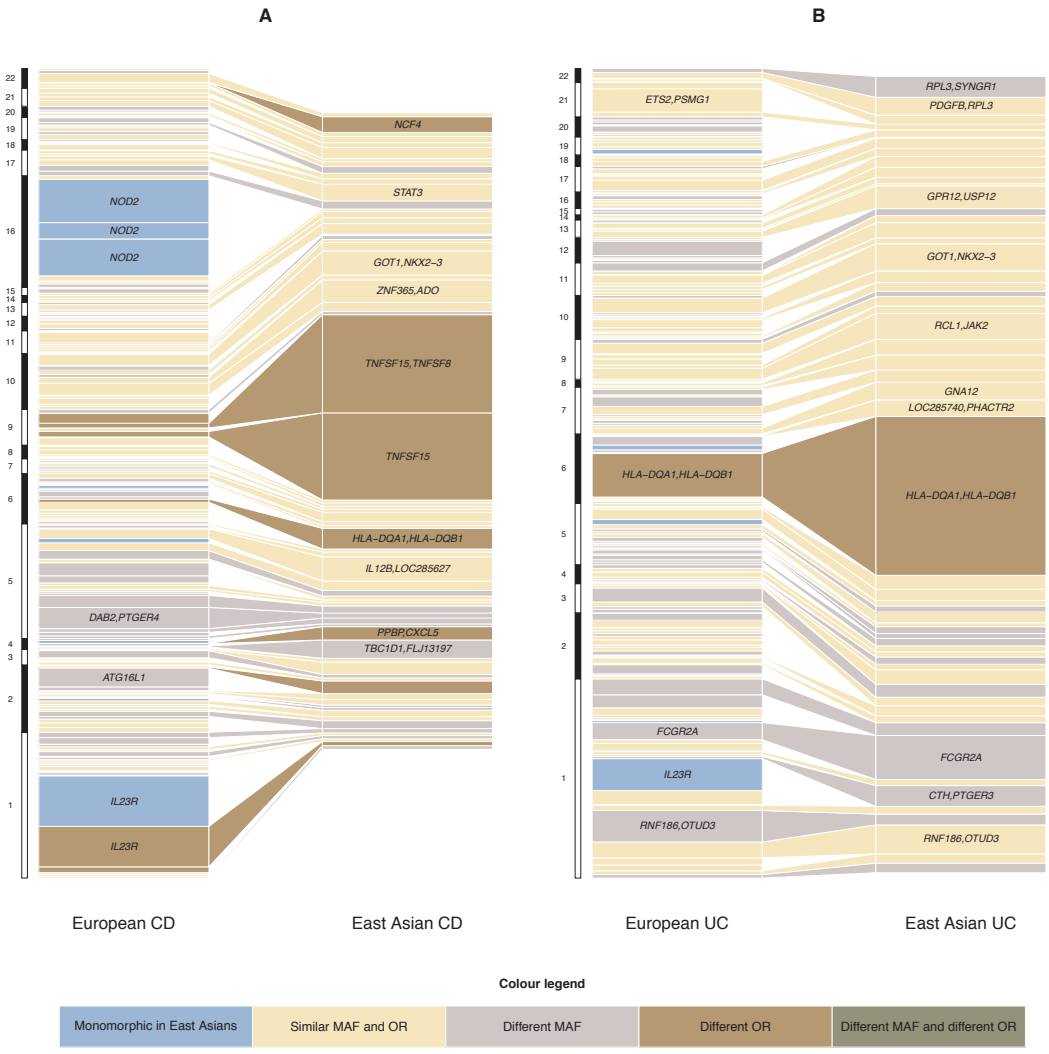

Figure 4.6. Belgravia plot of (A) CD and (B) UC risk variants in Europeans and East Asians. Each box represents an independent association for each disease. The East Asian panel only contains SNPs with association P < 0.01. The size of the box is proportional to the amount of variance explained in disease risk (liability scale) for that variant. The colours of the boxes

represent whether any difference in variance is due to differences in allele frequencies (Fst > 0.1), odds ratios ($P < 2.5\times10^{-4}$) or both.

Inflammatory cytokines may play a more important role in East Asian CD, with the greatest variance in disease risk explained for any IBD risk variant observed at the *TNFSF15*/*TNFSF8* (tumour necrosis factor superfamily 15/8) locus. Compared with the modest effect sizes in Europeans, two of the three independent signals at *TNFSF15*/*TNFSF8* showed much larger effects in East Asians: rs4246905 (European OR = 1.14 [95%CI: 1.11-1.18], East Asian OR = 1.73 [1.57-1.91], $P_{het} = 5.91\times10^{-15}$) and rs13300483 (European OR = 1.14 [1.11-1.17], East Asian OR = 1.70 [1.57-1.84], $P_{het} = 1.98\times10^{-19}$) despite similar allele frequencies. The third variant was non-significant in East Asians (rs11554257, P = 0.21).

An experiment testing the effect size of these variants in East Asian CD cases and controls who are >2nd generation immigrants in Western countries will help disentangle the role of environment. If differences still persist, this raises the intriguing possibility that genetic factors are the cause of this heterogeneity. Alternative explanations include gene-gene interactions with other population-specific variants, or that these differences are explained by as-yet undetermined causal variant(s) that may reflect different patterns of LD with the reported SNPs. It is not possible to rule out this hypothesis using the data in this study. Although the Immunochip provides dense coverage at 186 loci with known associations to at least one immune-mediated disease, the selection of SNPs was based on low-coverage sequence data from the pilot release of the 1000 Genomes Project and only incorporates variants identified in the CEU (European ancestry) cohort. Approximately 240,000 SNPs were selected for inclusion with and array design success rate of 80%. A further ~30% of SNPs were also excluded during QC. Therefore, it remains possible that the causal variants remained untyped, and the chances of this occurring are greater in the populations of non-European ancestry. Until the causal variants that underlie these associated loci have been identified (or all SNPs within these loci are included in association tests) the possibility that differential tagging of untyped causal variants are driving this heterogeneity of effect cannot be ruled out.

4.3.5 Gene-based likelihood ratio test

In the previous section, I discussed how the small sample size of the non-European cohorts limits our ability to estimate the effect of known IBD risk loci in these populations. In loci where there a multiple known independent signals in Europeans, it may be possible to use this prior information and test whether the aggregate of these signals show significant associations in non-European populations. Gene-based aggregate approaches for common variants are potentially more powerful than single-SNP approaches for situations where multiple SNPs within a gene are independently associated, and also due to a less stringent gene-wide P-value threshold (Neale and Sham, 2004). By only aggregating SNPs with prior evidence of association in the European cohort, this approach may also have greater power than traditional gene-based tests for common variants that consider all SNPs within a gene (Liu *et al.,* 2010a; Huang *et al.,* 2011). To do this, I first identified loci with multiple independent associations in Europeans, and then modelled these SNPs jointly within each gene in each of the non-European populations. Significance of the model was tested using a likelihood ratio test.

Genes were first selected if they have transcript start/stop boundaries (±50 kb) that overlap the most associated SNP in each locus and were located within the Immunochip high-density regions. Within each gene, independent associations were identified using the conditional and joint multi-SNP model selection approach implemented in GCTA (Yang *et al.,* 2012). I applied this to European Immunochip chip samples within each of the three phenotypes: CD UC and IBD, and identified 111 genes with more than one independent signal. When considering the overlap between genes (a SNP may be assigned to multiple genes), this corresponds to 41 non-overlapping loci. Performing the likelihood ratio tests on SNPs in these loci in the non-European samples revealed nine loci with significant evidence of association ($P < 5\times10^{-5}$). At six of these loci, the P-value form the likelihood ratio test was smaller than the smallest univariate SNP P-value in the non-European cohort. Nevertheless, this power improvement is only marginal, as with the exception of the *TNFSF15*/*TNFSF8* locus, significance

of the likelihood ratio test never exceeded the univariate SNP P-value by more than one order of magnitude.

Gene	Chr.	Gene start	Gene stop	Pop	Pheno	SNPs	Gene P-value	Best SNP P-value	Locus number
TMCO4	1	19.83	20.05	IND	IBD	3	3.37×10^{-5}	8.33×10^{-5}	1
TMCO4	1	19.83	20.05	EAS	UC	3	3.18×10^{-7}	2.36×10^{-6}	1
TMCO4	1	19.83	20.05	IND	UC	3	2.97×10^{-5}	2.95×10^{-4}	1
RNF186	1	19.96	20.06	IND	IBD	3	3.37×10^{-5}	8.33×10^{-5}	1
RNF186	1	19.96	20.06	EAS	UC	3	3.18×10^{-7}	2.36×10^{-6}	1
RNF186	1	19.96	20.06	IND	UC	3	2.97×10^{-5}	2.95×10^{-4}	1
FCGR2A	1	159.69	159.81	EAS	UC	3	5.53×10^{-6}	2.62×10^{-5}	2
HSPA6	1	159.71	159.81	EAS	UC	3	5.53×10^{-6}	2.62×10^{-5}	2
FCGR3A	1	159.73	159.84	EAS	UC	3	5.53×10^{-6}	2.62×10^{-5}	2
IL10	1	204.96	205.06	EAS	UC	2	1.22×10^{-8}	5.72×10^{-7}	3
IL19	1	204.99	205.13	EAS	UC	2	1.22×10^{-8}	5.72×10^{-7}	3
IL18RAP	2	102.35	102.49	EAS	IBD	3	4.81×10^{-6}	9.09×10^{-7}	4
MIR4772	2	102.37	102.47	EAS	IBD	2	1.12×10^{-6}	9.09×10^{-7}	4
SLC9A4	2	102.41	102.57	EAS	IBD	2	1.12×10^{-6}	9.09×10^{-7}	4
LOC285626	5	158.64	158.77	EAS	CD	3	8.46×10^{-10}	3.46×10^{-10}	5
LOC285626	5	158.64	158.77	EAS	IBD	3	1.03×10^{-9}	3.70×10^{-9}	5
LOC285627	5	158.76	158.88	EAS	CD	2	6.01×10^{-10}	3.46×10^{-10}	5
LOC285627	5	158.76	158.88	EAS	IBD	2	1.35×10^{-9}	3.70×10^{-9}	5
TNFSF15	9	116.54	116.66	EAS	CD	2	2.80×10^{-49}	2.83×10^{-45}	6
TNFSF15	9	116.54	116.66	EAS	IBD	3	1.40×10^{-30}	1.65×10^{-30}	6
TNFSF8	9	116.65	116.78	EAS	IBD	2	3.08×10^{-19}	1.13×10^{-17}	6
DKFZP434A062	9	138.29	138.39	IND	IBD	2	3.63×10^{-5}	1.46×10^{-5}	7
DKFZP434A062	9	138.29	138.39	IND	UC	2	2.36×10^{-5}	8.71×10^{-6}	7
GPSM1	9	138.29	138.42	IND	IBD	2	3.63×10^{-5}	1.46×10^{-5}	7
GPSM1	9	138.29	138.42	IND	UC	2	2.36×10^{-5}	8.71×10^{-6}	7
DNLZ	9	138.33	138.43	IND	IBD	2	3.63×10^{-5}	1.46×10^{-5}	7
DNLZ	9	138.33	138.43	IND	UC	2	2.36×10^{-5}	8.71×10^{-6}	7
CARD9	9	138.33	138.44	IND	IBD	2	3.63×10^{-5}	1.46×10^{-5}	7
CARD9	9	138.33	138.44	IND	UC	2	2.36×10^{-5}	8.71×10^{-6}	7
SNAPC4	9	138.34	138.46	IND	IBD	2	3.63×10^{-5}	1.46×10^{-5}	7
SNAPC4	9	138.34	138.46	IND	UC	2	2.36×10^{-5}	8.71×10^{-6}	7
SDCCAG3	9	138.37	138.47	IND	IBD	2	3.63×10^{-5}	1.46×10^{-5}	7
SDCCAG3	9	138.37	138.47	IND	UC	2	2.36×10^{-5}	8.71×10^{-6}	7
PMPCA	9	138.37	138.49	IND	IBD	2	3.63×10^{-5}	1.46×10^{-5}	7
PMPCA	9	138.37	138.49	IND	UC	2	2.36×10^{-5}	8.71×10^{-6}	7
C9orf163	9	138.45	138.55	IND	IBD	3	2.52×10^{-5}	1.04×10^{-4}	7
ADO	10	64.18	64.29	EAS	CD	2	9.13×10^{-8}	4.24×10^{-9}	8
EGR2	10	64.19	64.30	EAS	CD	2	6.56×10^{-8}	3.05×10^{-9}	8
NKX2-3	10	101.23	101.34	EAS	CD	2	4.67×10^{-8}	2.00×10^{-9}	9
NKX2-3	10	101.23	101.34	EAS	IBD	2	2.45×10^{-10}	2.79×10^{-11}	9
NKX2-3	10	101.23	101.34	EAS	UC	2	3.89×10^{-5}	3.07×10^{-6}	9

Table 4.7. Genes that exceeded P < 5×10-5 in at least one non-European cohort in the likelihood ratio locus-based test.

The likelihood ratio approach described here is similar to polygenic risk modelling, a commonly used method for identifying pleiotropy between a pair of phenotypes in genotyped individuals (International Schizophrenia Consortium *et al.,* 2009). Here, rather than comparing two phenotypes, I compared the same

phenotype in two populations. In polygenic risk modelling, the effect sizes for a set of SNPs (for example, those with association $P < 5\times10^{-8}$) are first estimated for one phenotype, and then used to construct risk scores based on genotypes for each individual in a second trait from a non-overlapping population. The degree to which these risk scores are correlated with phenotype in this second population are then assessed via linear regression (or logistic regression for dichotomous traits), where the size of the pleiotropic effect and its significance can be estimated.

It is possible to apply the polygenic risk score method to this study, where for a given gene, effect sizes estimated in Europeans are used to generate risk scores in a non-European cohort. However, this type of analysis assumes that LD patterns between the two cohorts tested are identical (or the SNPs being tested are in linkage equilibrium in both populations), which is often not the case when comparing divergent populations. Significant independent SNPs estimated in one population may be correlated with each other in another population, making the true pleiotropic effect difficult to interpret. The likelihood ratio testing approach overcomes this potential bias due to LD by only considering independent signals in the European cohort, and then re-estimating their effects jointly in the non-European cohort. These joint effect sizes will reflect the patterns of LD. Indeed, in situations where LD patterns and allele frequencies are identical between the two cohorts, the likelihood ratio method and the polygenic risk score should provide almost identical results. Of course, neither method is suitable in situations where there are heterogeneous effects exist between the two populations.

4.3.6 Conclusions

In this, the largest trans-ethnic study of IBD in 96,856 individuals of European, East Asian, Indian and Iranian populations, 40 newly associated risk loci were identified, bringing the total number of IBD risk loci to 203. The large number of risk loci shared between populations and high genetic correlations also suggests that the underlying causal variants are common (allele frequencies > 5%), thus adding further weight to the growing number of arguments against the synthetic

association model for explaining common variant associations (Dickson *et al.,* 2010; Anderson *et al.,* 2011b; Wray *et al.,* 2011).

The population comparisons at known IBD risk loci also identified several associated loci that are population specific. For instance, variants in *NOD2* and *IL23R* with major effects in Europeans are monomorphic in East Asians. Given the smaller sample size of the non European cohorts, and that Immunochip SNP selection was based on resequencing data from individuals of European ancestry, there was little power in this study to identify variants that are monomorphic in Europeans but are associated in non-Europeans. Other loci polymorphic across populations also showed evidence for differences in effect size (for instance, *TNFSF15* in Europeans and East Asians; $P_{het} = 1.98\times10^{-19}$). Loci with large differences in effect size raises the intriguing possibility of gene-environment interactions, though the presence of untyped causal alleles cannot be ruled out.

The newly identified loci along with the concordance in directions of effect between populations demonstrates that trans-ethnic association studies are a powerful means of identifying novel risk loci in complex diseases such as IBD. By leveraging imputation based on tens of thousand of reference haplotypes, or directly sequencing large numbers of cases and controls, these studies will more thoroughly survey causal variants and thus have increased ability to model the genetic architecture of IBD across diverse ancestral populations.