6 CRP

6.1 An introduction on CRP

6.1.1 Biology and physiology of circulating CRP

CRP is an acute-phase protein of hepatic origin, composed of five 23-kDa subunits. CRP is synthesized by the liver in response to factors released by macrophages and adipocytes (Pepys and Hirschfield 2003), and its level increases following interleukin-6 secretion from macrophages and T cells. CRP was named so because it was first identified as a substance in the serum of patients with acute inflammation that reacted with the C-polysaccharide of Pneumococcus. It binds to lysophosphatidylcholine expressed on the surface of dead or dying cells and some types of bacteria in order to activate the complement system (Thompson et al. 1999).

CRP is associated with multiple aspects of atherosclerosis such as adhesion molecule expression, effects on fibrinolysis and alteration of endothelial function (Szmitko et al. 2003). Both lipid crystals and the infiltration of inflammatory cells are characteristic features of atherosclerosis that can be detected at the earliest stages of plaque development. Inflammatory mechanisms couple dyslipidaemia to atheroma formation. Crystalline cholesterol acts as an endogenous danger signal and its deposition in arteries or elsewhere is an early cause rather than a late consequence of inflammation (Duewell et al. 2010). Leukocyte recruitment and expression of pro-inflammatory cytokines characterize early atherogenesis. Moreover, inflammatory pathways promote thrombosis, a late and dreaded complication of atherosclerosis responsible for MI and most strokes. Identifying the triggers for inflammation and unravelling the details of inflammatory pathways may eventually furnish new therapeutic targets (Libby 2002).

A few inflammatory biomarkers have been studied for their potential link and role in atherosclerosis, with CRP being the one most widely studied and having strongest evidence

for added value to CVD risk prediction (Koenig et al. 2004, Cushman et al. 2005). The other inflammatory biomarkers include interleukin-6 (IL-6) (Ridker et al. 2000), and lipoprotein-associated phospholipase A2 (Lp-PLA2) (Persson et al. 2007), P-selectin (Ridker et al. 2001), tumour necrosis factor alpha (TNF-α), the inter-cellular adhesion molecule 1 (ICAM-1) and vascular cell adhesion molecule 1 (VCAM-1) (Malik et al. 2001) are also associated with CVD risk; however, these markers are less stable than CRP and hence are less reliable indicators.

6.1.2 CRP as risk factors for CVD

In 1994, CRP was first reported to predict a poor outcome in patients with unstable angina (Liuzzo et al. 1994). In 1997, CRP was deemed a plausible risk factor when Ridker and colleagues reported that baseline CRP predicted the risk of future MI and stroke (Ridker et al. 1997). But CRP is a relatively moderate predictor of CHD compared to established risk factors including lipids level and blood pressure (Danesh et al. 2004). Several epidemiological studies have shown that the addition of CRP to traditional risk factors only raises the *c* statistic by less than 0.015 (Folsom et al. 2006, Lloyd-Jones et al. 2006, Melander et al. 2009). Therefore CRP assessment would only have a small effect on treatment decisions (Boekholdt and Kastelein 2010).

The Emerging Risk Factors Collaboration (ERFC) conducted by far the largest epidemiological study on CRP, which combined dataset on more than 160,000 subjects and comprises 1.31 million person-years at risk and ~28,000 fatal or non-fatal disease outcomes (Emerging Risk Factors et al. 2010). This study reported that CRP concentrations were associated with the risk of CVD (including CHD, ischaemic stroke, vascular mortality, and non-vascular mortality), most established CVD risk factors, and other inflammatory markers. One year later, another study reported that CRP were associated with an increased CHD risk, after adjusting for more variables including waist circumference, physical activity, smoking, diabetes, SBP, HDL and LDL, hormone replacement therapy in women (Rana et al. 2011). CRP is also confounded by other non-established CVD risk factors. For example, the age-related variation in CRP and IL-6 is largely explained by differences in visceral adipose tissue (Cartier et al. 2009).

As heart-healthy diets, weight loss, and physical activity all reduce CRP levels as well as other CVD risk factors, the AHA/CDC guidelines suggest that a finding of elevated CRP can be used to reinforce basic messages for lifestyle change. Statin therapy may be effective in the primary prevention of coronary events among those with relatively low lipid levels but with elevated levels of CRP (Ridker et al. 2001) (Ridker et al. 2008) (Ridker et al. 2009), therefore making CRP an attractive biomarker to identify patients who are likely to benefit from statin therapy. The guidelines on CVD prevention (Expert Panel on Detection and Treatment of High Blood Cholesterol in 2001) recommend that individuals at high risk should be treated whereas additional information is needed for those at intermediate risk. Some reported that CRP level could fine-tune the choice of treatment for those predicted with intermediate risk (Pearson et al. 2003, Ridker et al. 2008, Rana et al. 2009) (Koenig et al. 2004), but this is not supported in other studies including the FHS (Wilson et al. 2005, Sattar et al. 2007).

CRP is not an established risk factor, because MR studies did not establish its causal role to CVD, by using a single cis- variant (Casas et al. 2006, Lawlor et al. 2008) or multiple cis- variants (Elliott et al. 2009) as instrumental variables. The causality of CRP to CVD is complicated by the fact that CRP is also synthesised in smooth muscle cells within diseased atherosclerotic arteries. Inflammation may play a causal role via upstream effectors rather than the downstream marker of CRP. Certain factors more proximal in regulation of CRP could play a causal role (Brunner et al. 2008, Elliott et al. 2009), and such connections have been established for two other inflammatory biomarkers (IL6 for CVD, and IL1 for T2DM).

6.1.3 Genetic determinants of CRP

The heritability of serum CRP level is up to 52% (MacGregor et al. 2004), providing a strong case for discovering genetic determinants of CRP.

Findings from candidate gene and linkage analysis

In 2008, a linkage study was performed on a few inflammatory biomarkers including CRP, IL-6, and TNF-α in 764 subjects enrolled in the Quebec family study (Ruchat et al. 2008). The reported linkage signal was very modest and none remained significant after

adjustment for body mass index. The result suggested that several QTLs influence plasma levels of CRP partly via their effects on adiposity.

Findings from first generation GWAS

Since 2008, a total of 14 GWAS have been performed to discover genetic variants for association with CRP (**Table 6.1**). In 2008, the first large-scale GWAS scan on CRP led to a discovery of seven loci (*LEPR*, *CRP*, *IL6R*, *GCKR*, 12q23.2, *HNF1A*, *APOE*) (Ridker et al. 2008). The protein products for six of these loci are directly involved in metabolic syndrome, insulin resistance, beta cell function, weight homeostasis, and premature atherothrombosis. The largest GWAS of CRP was performed in 2011. It included more than 80,000 subjects of European ancestry and identified 11 novel loci (Dehghan et al. 2011). These loci are related to metabolic syndrome, immune system, and pathways previously unknown for chronic inflammation. GWAS on CRP was also conducted on non-European population, where novel findings included *TREM2* in African-American females (Reiner et al. 2012) and *IL6* (rs2097677) in Japanese individuals (Okada et al. 2011). These studies have been focusing on common variants genotyped on SNP arrays with imputation based on HapMap reference panel.

Findings from next generation sequencing

So far, there is no reported study on CRP that used high-throughput next generation sequencing technologies.

Table 6.1 GWAS studies of CRP

Date is for publication date. Samples are all European ancestry unless explicitly specified otherwise: KOR for Korean, FIL for Filipino, AA for African American, HIS for Hispanics, SAR for Sardinian, ASN for Asian. The sample size before "+" is for discovery while the sample size after "+" is for replication.

Date	Samples	Main findings	References
2008-04	6,345	4 loci (*LEPR*,*HNF1A*, *IL6R*, *GCKR*)	(Ridker et al. 2008)
2008-04	909+5,106	*HNF1A* intron 1	(Reiner et al. 2008)
2009-07	17,967+13,615	5 loci, no causal role of CRP for CHD	(Elliott et al. 2009)
2010-12	10,112+2,742 JAP	pleiotropic associations in IL6 gene	(Okada et al. 2011)
2011-02	66,185+16,540	7 known and 11 novel loci	(Dehghan et al. 2011)
2011-06	1,709 FIL	Interaction of CRP and HNF1A	(Wu et al. 2012)
2012-01	1,092	changes in response to fenofibrate treatment	(Aslibekyan et al. 2012)
2012-01	4,694 + 1392 SAR	3 novel loci	(Naitza et al. 2012)
2012-04	837AA	EA signals transferable to AA, AA data can fine-map of EA signal.	(Doumatey et al. 2012)
2012-07	8,842 KOR	CRP and WBC have distinct genetic components	(Kong and Lee 2013)
2012-08	8,280 AA 3,548 HIS	a common TREM2 variant	(Reiner et al. 2012)
2013-07	~7,500 ASN	EA variants are also detected in Asian	(Dorajoo et al. 2013)
2014-03	7,570 AA	a novel locus in CD36	(Ellis et al. 2014)
2014-04	7627 + 903 KOR	A novel variants in the ARG1	(Vinayagamoorthy et al. 2014)

6.1.4 Aims of this study

Under the framework of the UK10K project (The UK10K Consortium 2015), this study aimed to identify novel genetic variants that are associated with serum CRP levels and also fine map known CRP loci with WGS data. The current study is by far the largest WGS based association study of CRP, with 2,046 WGS samples and more than 32,000 samples with WGS imputed data. I first analysed the WGS samples aiming to discover rare and low frequency variants with large effect sizes. Then I analysed a much larger group of cohorts with imputed data to discover novel associations across the full MAF spectrum. Besides single marker based genome-wide scan, this study was able to fine map known loci and investigate the association and contribution of rare variants to serum lipids variance.

6.2 Methods

6.2.1 Cohorts & phenotype measurements

Like lipids, CRP was measured in both TwinsUK and ALSPAC. For WGS based analysis, I conducted a 2-way meta-analysis. For expanded discovery analysis, I included an additional 14 cohorts while used a single TwinsUK dataset as I did for FBC traits analyses. This leads to a 15-way expanded discovery analysis. The six WHI cohorts used for replication for FBC traits were also used for replication for CRP. But for CRP, I obtained another six cohorts as stage-2 replication (**Table 6.2**).

CRP was measured by high-sensitivity immunology assay in all participating cohorts. CRP measurement methods are as following: for **ALSPAC**, CRP was measured by Latex enhanced assay; for **TwinsUK**, CRP was measured by automated particle-enhanced immunoturbidimetric assay (Roche UK, Welwyn Garden City, UK); for **1958BC**, CRP antigen levels were measured by high sensitivity nephelometric assay using latex particles coated with monoclonal antibodies to human CRP in the BN Prospec protein analyzer (Dade Behring, Marburg, Germany). For **HELIC-MANOLIS** and **HELIC-Pomak**, CRP was measured using an immunoturbidimetric assay on a COBAS 8000 analyser (Roche). For

WHI, CRP was measured using a latex-particle enhanced immunoturbidimetric assay kit (Roche Diagnostics, Indianapolis, IN). For **FHS** and the rest of discovery cohorts, CRP was measured in fasting serum samples using various versions of high-sensitivity assay, mostly the Dade Behring BN100.

For phenotype harmonizaiton, I first excluded abnormal values of CRP, defined as <0.1mg/L or >10mg/L. For TwinsUK, the phenotype harmonization was conducted for WGS and GWA samples separately. Inverse normal transformation was applied to the full dataset without gender specific transformation. Regression test found no significant effects of dates of visits or analysers. BMI was not included as a covariate. For each trait of each cohort, the residuals with confounding variables regressed out were standardized so that the phenotype has a mean of 0 and a standard deviation of 1.

6.2.2 Single marker based discovery and follow-up

To discover variants of low and rare frequency with big effect size, I first run genome-wide association for the TwinUK WGS and ALSPAC WGS. I used SNPTEST to fit linear models on standardised trait residuals to test associations of allele dosages with 13,074,236 SNVs and 1,122,542 biallelic InDels (MAF≥0.1%) in the two WGS samples, followed by a meta-analysis to produce the 2-way meta-analysis, which has a total sample size of 2,046 (**Table 6.2**). Variants with 2-way meta-analysis *P*<1E-6 are deemed of interest for follow-up and further characterization. For the expanded discovery meta-analysis, I used all TwinsUK samples as a single cohort, the same way as I did for the FBC traits, in order to bring the co-Twins into the analysis. There are a total of 15 cohorts included for this expanded discovery analysis with a total sample size of 32,624 (**Table 6.2**). For each individual cohort, SNPTEST was used for population based samples while GEMMA was used for genetic isolates and cohorts with family structure. A 15-way meta-analysis was conducted using GWAMA v2.1, assuming a fixed effect model and adjusting genomic control to the summary statistics for both input and output data. Given the poor imputation quality and weak statistical power for rare variants, I chose to exclude the variants that did not pass a low allele frequency threshold (MAF<0.1%). For imputed cohorts, the variants with INFO <0.4 were also excluded.

Given the availability of the genome-wide results for the six replication cohorts from WHI, I run a 21-way meta-analysis that included the 15 discovery cohorts and six replication cohorts, as what I did for four FBC traits. This time, I have an additional six cohorts for stage-2 replication. The total sample size is 32,624 for 15-way discovery, 12,868 for 6-way (WHI) replication, and 27,726 for stage-2 replication.

6.2.3 Rare variant aggregation based discovery and follow-up

To evaluate the aggregation effects of rare variants, I used SKAT-O to discover genomic regions that harbour rare variants with large efforts but those effects could be picked up by single marker based analysis. I first evaluated the associations of rare variants by considering genes as functional units of analysis. I applied two separate statistical models with different properties to rare variants (MAF<1%): SKAT and burden tests, both implemented in a unified software SKAT-O. As described in chapter 2, in *naïve* tests, all variants in exons, untranslated regions (UTRs) and essential splice sites were considered, and were given equal weight of being causal (50,214 windows for 35,709 genes, mean=35 variants, median=38 variants per window). In functional tests, only loss of function (LoF) and predicted functional variants were included (15,528 gene windows with ≥ 5 variants, mean=18, median=14 variants per gene). Finally, I run the locus-based analysis genome-wide in an agonistic fashion, by constructing ~1.8 million windows of 3 kb each, overlapping by half (median 35 SNVs/window, MAF<1%), assigning an equal weight to all variants. For CRP, there is no replication data available for rare variants aggregation based tests.

6.2.4 Fine-mapping of known loci

For a total of 37 previously established CRP loci, I carried out fine-mapping analysis to assess the probability of each variant being causal given other variants in the region. Within each signal I included SNPs in high LD (defined as all variants having $r^2 \geq 0.8$ with the most associated variants in the region). As described in chapter 2, I first created a list of fine-mapping regions based on HapMap estimates of recombination rates. I then analysed each

region separately for each of the 15 participating cohort using Bayesian linear additive models, by accounting for covariates as in the general single point association analyses. At the end, the resulting BFs for each variant were multiplied to obtain a joint BF measure of association, with the assumption that each cohort is independent. These BFs are then used to calculate posterior probabilities, based on the assumption that there is exactly one causal SNP in each region. In addition, 95% and 99% credible sets are constructed in order to assess the uncertainty of the fine-mapping analysis.

Table 6.2 Characteristics of participating cohorts

All cohorts are population based, except for TwinsUK. Imputation was conducted with the 1000G and UK10K combined reference panel unless otherwise specified. For the expanded discovery analysis, "TwinUK WGS" were not included because it is already in "TwinsUK all".

	Cohort	N	Country	Age	% Female	CRP (mg/L)
Discovery	ALSPAC WGS	1,167	UK	10 (9-11)	50.3	1.01 (0.25)
	TwinsUK WGS	879	UK	56 (17-85)	100.0	1.42 (0.32)
	ALSPAC GWA	2,226	UK	10 (9-12)	49.2	0.78 (0.21)
	TwinkUK all	2,512	UK	50 (16-83)	97.3	0.94 (0.22)
	1958BC	4,910	UK	44 (44-44)	52	1.00 (0.24)
	FHS	6,320	Italy	49 (31-72)	53	0.62 (0.21)
	INGI-FVG	411	Italy	52 (18-92)	58.2	1.34 (0.13)
	INGI-VB	1,162	Italy	55 (18-102)	56.3	1.02 (0.31)
	HELIC-Manolis	1,093	Greece	62 (18-99)	57.2	1.33 (0.22)
	HELIC-Pomak	839	Greece	43 (13-87)	72.1	0.98 (0.16)
	INCIPE-1	807	Italy	60 (35-89)	54	0.79 (0.21)
	INCIPE-2	1,332	Italy	58 (26-95)	51	0.82 (0.17)
	LURIC-Ctrl	1,228	Germany	62 (18-92)	59.7	1.01 (0.20)
	LURIC-Case	1,202	Germany	61 (17-91)	60.8	1.65 (0.22)
	Procardis-Case	3,732	Sweden	43 (13-87)	49.1	1.54 (0.22)
	Procardis-Ctrl	3,683	Sweden	63 (51-78)	55.8	1.06 (0.19)
Replication stage 1	WHI-Garnet	3,388	US	65 (50-79)	100.0	0.88 (0.20)
	WHI-Gecco1	780	US	65 (50-79)	100.0	1.07 (0.30)
	WHI-Gecco2	1,072	US	65 (50-79)	100.0	0.87 (0.21)
	WHI-Hipfx	1,716	US	65 (50-79)	100.0	0.99 (0.26)
	WHI-Mopmap	721	US	65 (50-79)	100.0	1.31 (0.32)
	WHI-Whism	5,191	US	65 (50-79)	100.0	1.22 (0.19)
Replication stage 2	Rotterdam Study	5,455	Netherlands	69 (48-75)	41.2	0.91 (0.32)
	LOLIPOP-EWA	505	UK	56 (35-75)	26.8	0.77 (0.19)
	LOLIPOP-EWP	564	UK	55 (23-75)	13.1	0.89 (0.24)
	LOLIPOP-EW610	834	UK	56 (32-67)	0.0	0.94 (0.18)
	Fenland	8,178	UK	65 (47-77)	46.2	1.08 (0.30)
	Lifelines	12,190	Netherlands	NA	NA	NA

6.3 Results

6.3.1 Novel loci and novel variants from single marker analysis

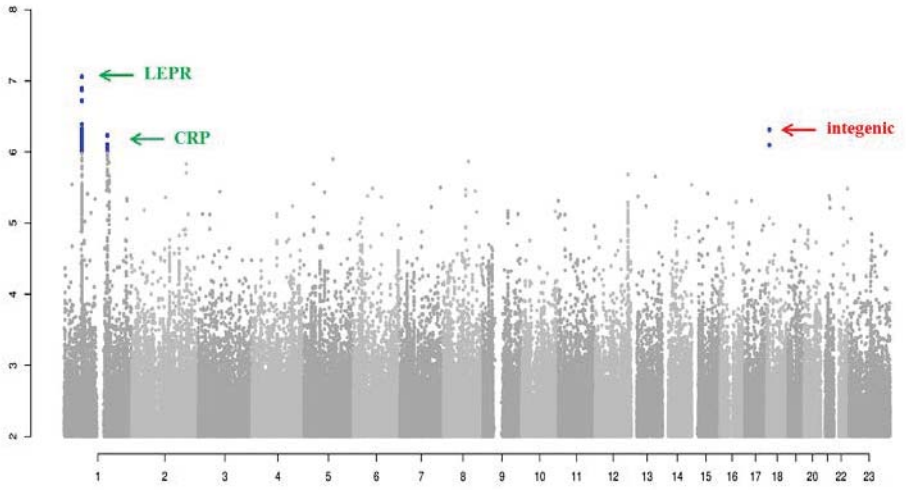
WGS for low frequency and rare variants

The assessment of associations based on imputation or WES has been incomplete. I thus sought to investigate if additional low-frequency or rare variants with strong effects could be detected from the WGS dataset. I first tested association results using solely the WGS dataset in order to identify whether these variants existed. Associations were carried out in 13,074,236 SNVs and 1,122,542 biallelic InDels (MAF≥0.1%) using linear regression and data from the two WGS cohorts was meta-analysed.

There are a total of 61 variants from UK10K WGS that have *P*<1E-6, but none of these reached *P*<5.0E-08 (**Figure 6.1**). 59 of these are common variants within the well-established *LEPR* and *CRP* loci, while the other two have low frequency (MAF=0.02) in an intergenic region on chromosome 18. These two variants are in high LD (r2=0.97). The first one is rs112734184 (chr18:10441718, EA=G, EAF=0.022, beta=-0.540, *P*=4.94E-07) and the second one is rs112155044 (chr18:10445499, EA=T, EAF=0.022, beta=-0.524, *P*= 8.06E-07). These two variants are non-significant in the large meta-analysis with 15 cohorts and ~45,000 samples (*P*>0.05), as described later. Therefore, they are most likely to be specific to the two UK10K cohorts or false positive.

Figure 6.1 Association Results of CRP based on WGS samples

X-axis is for chromosome and positions (build 37). Y-axis is for –log10(*P*). Variants passing threshold of 5E-08 and 1E-06 are shown in red and blue, respectively. For those passing threshold of 5E-08, known loci were marked in green text while putative novel loci were marked in red text.



Meta-analysis for identifying novel variants of all allele spectrums

Given the enhanced imputation quality with the UK10K WGS reference panel as demonstrated in chapter 3, I included an additional of 13 cohorts with imputed data for an expanded discovery, to increase power for discover variants across all allele frequency spectrum. As mentioned earlier in the methods section, variants with MAF <0.1% or imputation INFO <0.4 were not included. This effort yielded a total of 1,303 variants with *P*<1E-07, six of which are deemed novel after conditional analysis and LD pruning with positive controls (**Figure 6.2**, **Table 6.3**). Initially, six European cohorts from the Women's Genome Initiative (WHI) were made available for *in-silico* replication, but none of the six variants from the 15-way discovery were replicated. I then run a meta-analysis including all 15 discovery cohorts and six replication cohorts in a 21-way meta-analysis, where two novel variants passed the genome-wide significant threshold of 5E-08. These two variants are listed at the bottom of **Table 6.3**. The individual cohort results for these variants are presented in **Table 6.4**.

I took forward these eight variants into a stage 2 replication with six independent cohorts. Two of the eight variants were replicated at *P* <0.05. The regional plots of these two novel loci are shown in **Figure 6.3**. For the first locus, the lead SNP rs9393691 (chr6:26272829) is a common variant (MAF=0.383) within *HIST1H3G* (Histone cluster 1, H3g). This gene is found in the large histone gene cluster on chromosome 6. Histones are basic nuclear proteins that are responsible for the nucleosome structure of the chromosomal fiber in eukaryotes. Two molecules of each of the four core histones (H2A, H2B, H3, and H4) form an octamer, around which approximately 146 bp of DNA is wrapped in repeating units, called nucleosomes. The linker histone, H1, interacts with linker DNA between nucleosomes and functions in the compaction of chromatin into higher order structures. The association barely met the pre-defined threshold of *P*<1E-07, with 15-way *P*=9.90E-08. This region has been reportedly associated with many phenotypes including hematological traits and CHD risk factors, but the current lead SNP rs9393691 is not in LD with any of the known variants (r2<0.1) expect for one variant reported for association with height (rs10946808, r2=0.47) (**Table 6.5**). This variant exists in the published largest GWAS on CRP (Dehghan et al. 2011), but it was not significant (beta=0.0104, SE=0.0065, *P*=0.106). For the second locus, the lead SNP rs117410733 (chr15:52655560) is an intronic variant within *MYO5A*, which is a class of actin-based motor proteins involved in cytoplasmic vesicle transport and anchorage, spindle-

polealignment and mRNA translocation. Currently, there is no evidence in the literature supporting this gene's role in affecting circulating CRP level.

I also compared the summary statistics of 17 variants reported in the published largest GWAS (Dehghan et al. 2011). Six of those 17 variants are marginally significant in UK10K WGS (P <0.05) and five of them are genome-wide significant in 15-way meta-analysis (P <5E-08). Although the statistical significances differ, the effect size and directions are comparable between the previous GWAS, TwinsUK WGS and ALSPAC WGS. For the majority of the studied phenotypes, they are inverse normal transformed followed by a standardization of residuals. So, the phenotypes used in the association studies all have a normal distribution, with a mean of 0 and 1. This accounted for a lot of heterogeneity between individual GWAS that were included in the meta-analysis. For CRP for example, the mean (standard deviation) values for the raw phenotypes are 3.38(6.51) for TwinsUK and 0.84 (3.09) for ALSPAC. For the 17 positive controls, the effect sizes of 17 positive controls are very comparable between these two cohorts even though the raw phenotype values differ significantly.

Figure 6.2 Single marker association results of CRP from expanded meta-analysis

From top to bottom, the three plots are for 15-way, 6-way replication (WHI cohorts), and 21-way combined, respectively. X-axis is for chromosome and positions (build 37). Y-axis is for –log10(*P*). Variants passing threshold of 5E-08 and 1E-07 are shown in red and blue, respectively. For those passing threshold of 5E-08, known loci were marked in green text while putative novel loci were marked in red text.

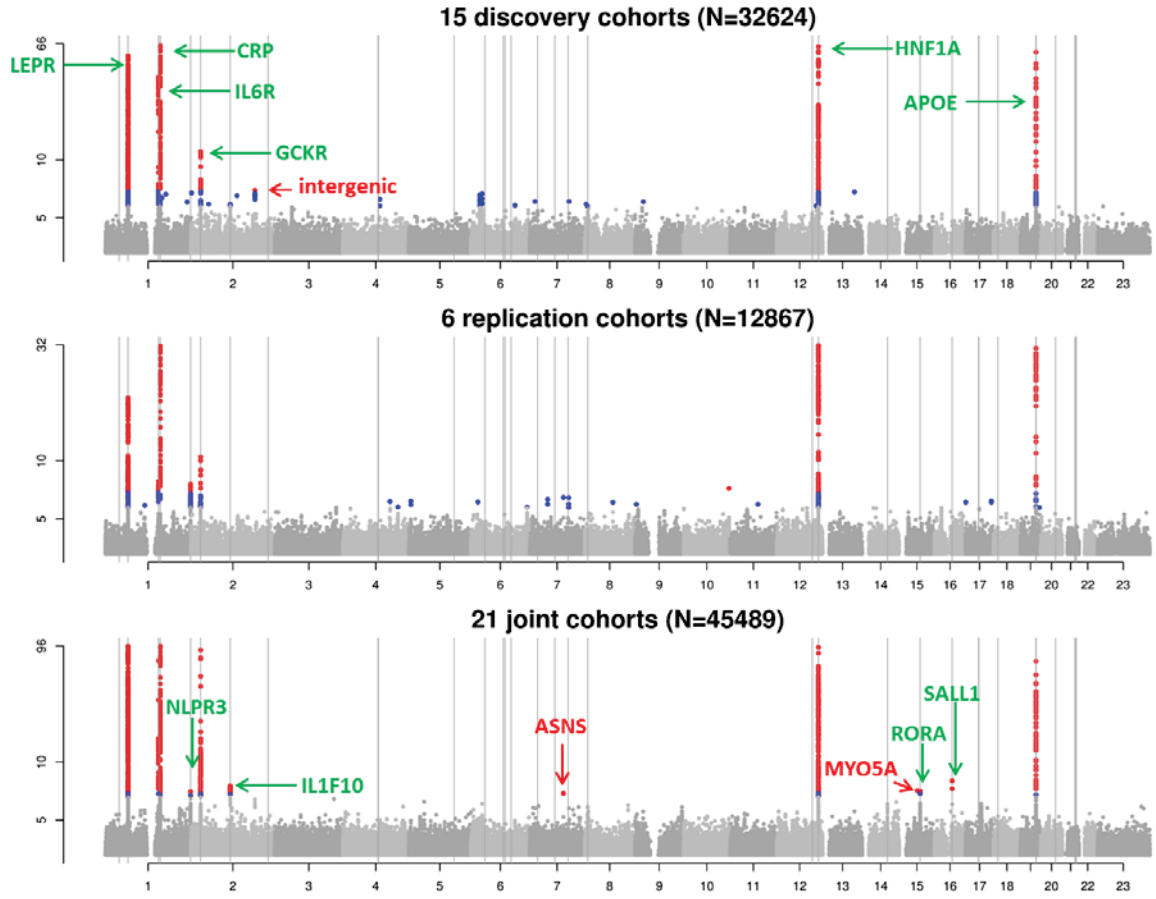


Table 6.3 Novel associations of CRP from expanded discovery meta-analysis

The first six variants are putative novel based on the 15-way expanded discovery with *P* <1E-07. The last two variants are putative novel based on the 21-way further expanded meta-analysis with *P* <5E-08.

					15-way					WHI Replication (with genome-wide data)				
rsID	**CHR**	**POS**	**Gene**	**EA**	**EAF**	**Beta**	**SE**	**P**	**N**	**EAF**	**Beta**	**SE**	**P**	**N**
rs137929481	1	176,045,265	*RFWD2*	G/A	0.002	0.584	0.109	9.06E-08	30615	0.004	-0.139	0.132	0.36	12865
rs35993482	2	1,320,638	*SNTG2*	A/G	0.021	0.207	0.038	7.00E-08	31454	0.025	-0.063	0.051	0.28	12865
rs76870040	2	185,422,774	*ZNF804A*	G/A	0.015	0.186	0.034	4.13E-08	32623	0.018	0.078	0.048	0.15	12868
rs9393691	6	26,272,829	*HIST1H3G*	C/T	0.383	-0.045	0.008	9.90E-08	32622	0.389	0.004	0.012	0.76	12866
rs9269303	6	32,539,581	*HLA*	T/G	0.476	-0.059	0.011	8.41E-08	30648	0.432	-0.001	0.015	0.95	12868
rs186492213	13	92,240,699	*GPC5*	G/A	0.002	-0.544	0.100	5.77E-08	30040	0.002	-0.107	0.157	0.55	12083
P<5E-08 in 21-way														
chr7:97545859	7	97545859	*ASNS*	A/G	0.009	-0.143	0.052	6.37E-03	32621	0.009	-0.449	0.075	1.36E-07	12865
rs117410733	15	52655560	*MYO5A*	G/A	0.009	-0.189	0.045	2.77E-05	32623	0.011	-0.241	0.060	4.01E-04	12865

21-way					Stage 2 replication					Combined				
EAF	**Beta**	**SE**	**P**	**N**	**EAF**	**Beta**	**SE**	**P**	**N**	**EAF**	**Beta**	**SE**	**P**	**N**
0.003	0.293	0.084	8.08E-04	43,480	0.003	0.722	0.452	0.11	1903	0.003	0.307	0.082	1.99E-04	46222
0.022	0.111	0.031	5.21E-04	44,319	0.311	0.000	0.014	0.98	7998	0.066	0.020	0.013	1.20E-01	58412
0.016	0.150	0.028	1.68E-07	45,491	0.015	0.003	0.035	0.93	21630	0.016	0.093	0.022	1.57E-05	73216
0.384	-0.029	0.007	8.09E-05	45,488	0.374	-0.045	0.009	5.82E-07	21529	0.381	-0.035	0.005	1.71E-11	73112
0.463	-0.039	0.009	2.28E-05	43,516	0.479	-0.011	0.041	0.80	1903	0.464	-0.038	0.009	1.23E-05	45419
0.002	-0.419	0.084	1.82E-06	42,123	0.002	0.004	0.159	0.98	7493	0.002	-0.326	0.074	1.17E-05	56966
P<5E-08 in 21-way														
0.009	-0.244	0.043	4.90E-08	45,486	0.010	0.056	0.064	0.38	7998	0.009	-0.151	0.036	2.21E-05	59579
0.009	-0.208	0.036	2.75E-08	45,488	0.006	-0.094	0.047	4.32E-02	7998	0.009	-0.165	0.028	5.78E-09	59581

Table 6.4 Cohort specific results of novel associations from expanded discovery

For each set of results, the effect allele frequency (EAF), beta, standard deviation (SE), *P* value, sample size (N), and imputation INFO score were presented. Records with *P* < 0.05 are highlighted in red text.

	chr1:176045265						chr2:1320638						chr2:185422774						chr6:26272829					
Cohort	**EAF**	**beta**	**SE**	**P**	**N**	**Info**	**EAF**	**beta**	**SE**	**P**	**N**	**Info**	**EAF**	**beta**	**SE**	**P**	**N**	**Info**	**EAF**	**beta**	**SE**	**P**	**N**	**Info**
ALSPAC WGS	--	--	--	--	--	--	--	--	--	--	--	--	0.014	0.118	0.171	4.9E-01	1167	0.99	0.376	-0.056	0.043	1.9E-01	1167	1.00
TwinsUK WGS	--	--	--	--	--	--	--	--	--	--	--	--	0.015	0.456	0.192	1.7E-02	879	1.00	0.378	-0.008	0.050	8.7E-01	879	1.00
ALASPAC GWA	0.002	0.439	0.406	2.8E-01	2226	0.57	0.019	0.238	0.14	8.9E-02	2226	0.62	0.016	0.256	0.119	3.2E-02	2226	1.00	0.377	-0.098	0.031	1.4E-03	2226	1.00
1958BC	0.003	0.589	0.245	1.6E-02	4910	0.58	0.019	0.281	0.093	2.7E-03	4910	0.61	0.015	0.198	0.083	1.7E-02	4910	0.99	0.358	-0.062	0.021	3.1E-03	4910	1.00
FHS	0.003	0.548	0.253	3.1E-02	6320	0.48	0.026	0.128	0.076	9.3E-02	6320	0.6	0.015	0.133	0.078	8.8E-02	6320	0.98	0.389	-0.019	0.020	3.3E-01	6320	0.99
INGI-FVG	0.001	4.623	2.773	9.7E-02	411	0.87	0.028	0.181	0.275	5.1E-01	411	0.54	0.016	0.091	0.269	7.4E-01	411	0.97	0.433	-0.074	0.067	2.7E-01	411	1.00
HELIC-A	0.002	0.837	0.661	2.1E-01	1093	0.46	0.063	0.421	0.132	1.8E-03	1093	0.48	0.011	0.136	0.211	5.2E-01	1093	1.00	0.502	-0.039	0.045	3.8E-01	1093	1.00
HELIC-P	0	0.681	5.495	9.0E-01	839	0.05	0.018	0.149	0.291	6.1E-01	839	0.4	0.040	0.102	0.129	4.3E-01	839	0.99	0.339	-0.047	0.054	3.8E-01	839	1.00
Incipe-1	0.003	0.272	0.778	7.3E-01	807	0.39	0.028	0.311	0.212	1.4E-01	807	0.48	0.012	0.522	0.211	1.3E-02	807	0.99	0.428	-0.042	0.050	4.0E-01	807	1.00
Incipe-2	0.002	0.732	0.68	2.8E-01	1332	0.36	0.033	-0.156	0.157	3.2E-01	1332	0.51	0.012	0.096	0.182	6.0E-01	1332	0.99	0.414	-0.057	0.039	1.4E-01	1332	1.00
LURIC-Ctrl	0.004	0.547	0.456	2.3E-01	1228	0.55	0.024	0.478	0.175	6.3E-03	1228	0.57	0.013	0.089	0.182	6.2E-01	1228	0.98	0.399	-0.116	0.041	4.8E-03	1228	1.00
LURIC-Case	0.004	0.742	0.451	1.0E-01	1202	0.53	0.024	0.11	0.17	5.2E-01	1202	0.61	0.017	0.173	0.162	2.9E-01	1202	0.98	0.390	-0.049	0.043	2.5E-01	1202	1.00
Procardis-case	0.001	0.694	0.301	2.1E-02	3732	0.88	0.011	0.139	0.167	4.1E-01	3732	0.46	0.016	0.196	0.092	3.3E-02	3732	1.00	0.376	-0.040	0.024	1.0E-01	3732	1.00
Procardis-ctrl	0.001	0.464	0.376	2.2E-01	3683	0.88	0.011	0.683	0.232	3.7E-03	3683	0.46	0.016	0.395	0.141	5.3E-03	3683	1.00	0.376	-0.025	0.036	5.0E-01	3683	1.00
TwinsUKvall	0.003	0.534	0.298	7.3E-02	2512	0.66	0.02	0.122	0.122	3.2E-01	2512	0.72	0.012	0.200	0.131	1.3E-01	2512	0.99	0.368	-0.009	0.031	7.6E-01	2512	1.00
TwinsUK GWA	0.004	1.08	0.464	2.1E-02	1017	0.65	0.021	-0.032	0.191	8.7E-01	1017	0.67	0.013	0.005	0.202	9.8E-01	1017	0.99	0.355	0.008	0.048	8.7E-01	1017	1.00
INGI-VBI	0.002	0.662	0.843	4.4E-01	1162	0.29	0.019	0.136	0.227	5.5E-01	1162	0.47	0.011	0.098	0.212	6.5E-01	1162	0.99	0.355	0.033	0.045	4.7E-01	1162	1.00
WHI-Garnet	0.004	0.079	0.25	7.5E-01	3388	0.63	0.027	-0.081	0.094	3.8E-01	3388	0.64	0.014	0.127	0.101	2.1E-01	3388	1.00	0.391	0.033	0.025	1.7E-01	3388	1.00
WHI-Gecco1	0.005	0.555	0.466	2.3E-01	780	0.62	0.025	-0.134	0.228	5.6E-01	780	0.56	0.017	0.426	0.206	3.9E-02	780	0.99	0.381	0.012	0.057	8.3E-01	780	1.00
WHI-Gecco2	0.004	0.218	0.532	6.8E-01	1072	0.42	0.022	0.077	0.189	6.8E-01	1072	0.61	0.018	0.011	0.160	9.5E-01	1072	1.00	0.398	0.024	0.042	5.7E-01	1072	1.00
WHI-Hipfx	0.004	0.04	0.398	9.2E-01	1716	0.49	0.024	-0.082	0.139	5.6E-01	1716	0.63	0.020	0.127	0.123	3.0E-01	1716	1.00	0.393	-0.014	0.034	6.9E-01	1716	1.00
WHI-Mopmap	0.005	-0.037	0.429	9.3E-01	721	0.75	0.03	0.137	0.195	4.8E-01	721	0.65	0.026	-0.06	0.169	7.2E-01	721	1.00	0.379	-0.054	0.054	3.2E-01	721	1.00
WHI-Whims	0.004	-0.592	0.216	6.1E-03	5191	0.54	0.024	-0.096	0.084	2.5E-01	5191	0.59	0.018	0.029	0.075	7.0E-01	5191	0.99	0.389	-0.007	0.020	7.5E-01	5191	1.00

Table 6.4. Cohort specific results of meta-analysis top hits (continued)

	chr6:32539581						chr13:92240699						chr7:97545859						chr15:52655560					
cohort	**EAF**	**beta**	**SE**	**P**	**N**	**Info**	**EAF**	**beta**	**SE**	**P**	**N**	**Info**	**EAF**	**beta**	**SE**	**P**	**N**	**Info**	**EAF**	**beta**	**SE**	**P**	**N**	**Info**
ALSPAC WGS	--	--	--	--	--	--	0.001	-0.533	0.563	3.4E-01	1167	0.91	0.008	-0.186	0.232	4.2E-01	1167	0.95	0.011	-0.1	0.198	6.2E-01	1167	0.99
TwinsUK WGS	--	--	--	--	--	--	0.003	-0.691	0.418	9.9E-02	879	0.97	0.007	0.088	0.302	7.7E-01	879	0.95	0.006	0.058	0.333	8.6E-01	879	0.93
alspac	0.492	-0.095	0.038	1.1E-02	2226	0.64	0.002	-0.297	0.422	4.8E-01	2226	0.81	0.006	-0.190	0.218	3.8E-01	2226	0.80	0.007	-0.075	0.186	6.9E-01	2226	0.94
b58c	0.503	-0.058	0.026	2.3E-02	4910	0.62	0.001	-0.303	0.312	3.3E-01	4910	0.74	0.008	-0.193	0.126	1.3E-01	4910	0.83	0.008	-0.115	0.118	3.3E-01	4910	0.97
fhs	0.536	-0.058	0.024	1.5E-02	6320	0.66	0.003	-0.619	0.202	2.2E-03	6320	0.75	0.008	-0.258	0.119	3.0E-02	6320	0.77	0.009	-0.187	0.105	7.6E-02	6320	0.93
fvg	0.285	-0.106	0.087	2.2E-01	411	0.69	0	11.67	11.452	3.1E-01	411	0.03	0.006	0.015	0.480	9.7E-01	411	0.77	0.007	0.18	0.397	6.5E-01	411	0.98
HA	0.447	0.082	0.056	1.4E-01	1093	0.68	0.003	0.275	0.527	6.0E-01	1093	0.53	0.001	-1.126	2.126	6.0E-01	1093	0.45	0.014	-0.186	0.195	3.4E-01	1093	0.96
HP	0.424	-0.033	0.060	5.9E-01	839	0.67	0	61.77	25.695	1.7E-02	839	0.01	0.001	0.551	1.029	5.9E-01	839	0.48	0.005	0.167	0.416	6.9E-01	839	0.76
incipe1	--	--	--	--	--	--	0.001	-0.257	0.699	7.1E-01	807	0.84	0.005	0.446	0.416	2.8E-01	807	0.77	0.013	-0.149	0.224	5.1E-01	807	0.94
incipe2	0.390	-0.091	0.048	5.8E-02	1332	0.68	--	--	--	--	--	--	0.005	-0.222	0.343	5.2E-01	1332	0.65	0.01	-0.115	0.215	5.9E-01	1332	0.88
luric1	0.491	-0.084	0.050	9.5E-02	1228	0.64	0.001	-0.561	0.669	4.0E-01	1228	0.71	0.008	-0.340	0.266	2.0E-01	1228	0.75	0.018	-0.41	0.148	5.8E-03	1228	0.99
luric2	0.473	-0.072	0.051	1.6E-01	1202	0.63	0.004	-1.094	0.336	1.2E-03	1202	0.93	0.008	0.290	0.273	2.9E-01	1202	0.68	0.011	-0.75	0.202	2.1E-04	1202	0.98
procase	0.450	-0.036	0.029	2.2E-01	3732	0.69	0.002	-0.407	0.264	1.2E-01	3732	0.96	0.018	-0.129	0.117	2.7E-01	3732	0.60	0.009	-0.106	0.124	3.9E-01	3732	0.97
proctrl	0.450	-0.083	0.043	5.5E-02	3683	0.69	0.002	-0.888	0.303	3.5E-03	3683	0.96	0.018	-0.074	0.163	6.5E-01	3683	0.60	0.009	-0.332	0.167	4.7E-02	3683	0.97
TwinsUKall	0.416	-0.066	0.036	6.8E-02	2512	0.70	0.003	-0.295	0.309	3.4E-01	2512	0.82	0.008	-0.144	0.169	3.9E-01	2512	0.86	0.005	0.062	0.202	7.6E-01	2512	0.99
TwinsUK	0.409	-0.048	0.055	3.8E-01	1017	0.70	0.002	-0.246	0.613	6.9E-01	1017	0.78	0.009	-0.219	0.270	4.2E-01	1017	0.85	0.006	0.415	0.291	1.5E-01	1017	0.99
vb	0.515	-0.076	0.201	7.0E-01	1162	0.05	0.004	-0.827	0.432	5.7E-02	1162	0.61	0.001	1.818	0.839	3.1E-02	1162	0.34	0.004	-0.233	0.344	5.0E-01	1162	0.97
whi_garnet	0.352	-0.016	0.029	5.8E-01	3388	0.78	0.001	-0.259	0.36	4.7E-01	3388	0.84	0.010	-0.300	0.134	2.5E-02	3388	0.85	0.012	-0.238	0.113	3.5E-02	3388	0.98
whi_GECCO1	0.486	0.034	0.067	6.2E-01	780	0.63	--	--	--	--	--	--	0.009	-0.691	0.326	3.4E-02	780	0.77	0.012	-0.06	0.256	8.2E-01	780	0.93
whi_GECCO2	0.507	0.121	0.053	2.2E-02	1072	0.64	0.002	-0.067	0.484	8.9E-01	1072	0.84	0.009	-0.608	0.240	1.1E-02	1072	0.78	0.01	-0.63	0.214	3.3E-03	1072	0.97
whi_hipfx	0.496	0.053	0.043	2.2E-01	1716	0.63	0.002	-0.092	0.4	8.2E-01	1716	0.78	0.008	-0.383	0.212	7.1E-02	1716	0.77	0.01	-0.452	0.172	8.6E-03	1716	0.97
whi_mopmap	0.317	0.033	0.061	5.9E-01	721	0.87	0.003	0.448	0.502	3.7E-01	721	0.93	0.007	-0.858	0.388	2.7E-02	721	0.63	0.013	-0.293	0.238	2.2E-01	721	0.99
whi_whims	0.455	-0.041	0.024	8.1E-02	5191	0.68	0.002	-0.175	0.233	4.5E-01	5191	0.83	0.008	-0.481	0.121	7.6E-05	5191	0.82	0.012	-0.12	0.093	2.0E-01	5191	0.98

Figure 6.3 Regional plots of two novel associations of CRP

The *P* value in the plot is from the 15-way expanded discovery meta-analysis. The top plot shows the *HIST1H3G* locus. The lead SNP rs9393691 (chr6:26272829) is significant in 15-way (*P* =9.90E-08). Its combined 27-way meta-analysis *P*=1.71E-11. The bottom plots show the MYO5A locus. The lead SNP rs117410733 (chr15:52655560) is not significant in 15-way (*P*=2.77E-05), but in 21-way (*P*=2.75E-08). Its combined 27-way meta-analysis P=5.78E-09.

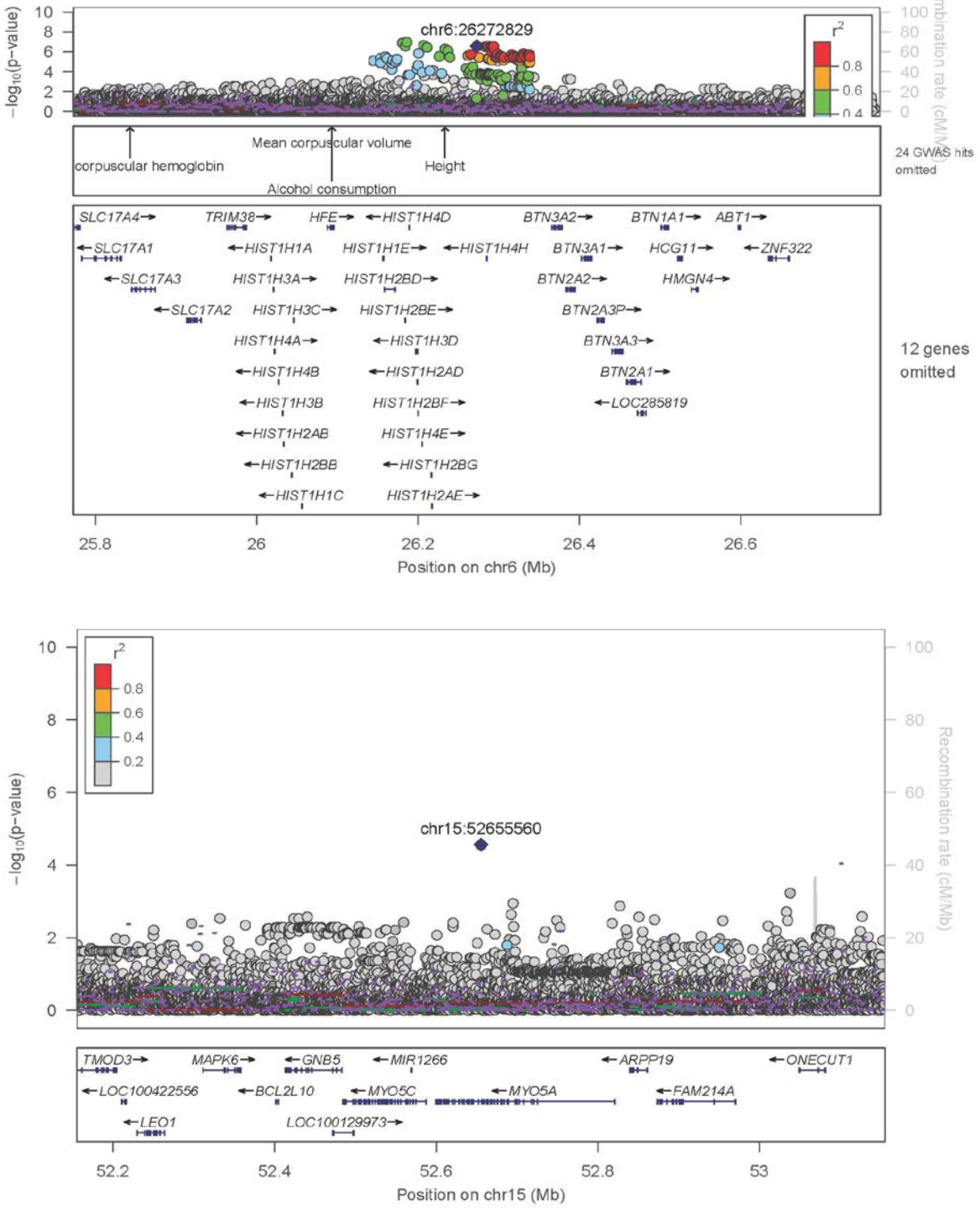


Table 6.5 LD between novel and known variants in *HIST1H3G*

This table lists 14 associations reported in GWAS Catalog that are within 1Mb of rs9393691. The LD of each variant with rs9393691 is shown in the last column, based on the WGS data of UK10K.

SNP	Trait	Chr	Pos	r2 with rs9393691
rs11754288	Cardiovascular disease risk factors	6	25776949	0.00
rs1165196	Urate levels	6	25813150	0.00
rs17342717	Iron status biomarkers	6	25821770	0.02
rs1183201	Uric acid levels	6	25823444	0.00
rs1408272	Mean corpuscular hemoglobin	6	25842951	0.02
rs1165205	Urate levels	6	25870542	0.00
rs1799945	Diastolic blood pressure	6	26091179	0.02
rs1799945	Iron levels	6	26091179	0.02
rs1800562	Hemoglobin	6	26093141	0.03
rs1800562	Cardiovascular disease risk factors	6	26093141	0.03
rs1800562	LDL cholesterol	6	26093141	0.03
rs198846	Blood pressure	6	26107463	0.02
rs198846	Hemoglobin	6	26107463	0.02
rs10946808	Height	6	26233387	0.47

6.3.2 Fine mapping of known and novel loci

The availability of WGS compared on GWAS based on sparse datasets allows one to evaluate statistically the plausibility of each variant in an association signal to be causally associated with a trait. To fine-map lipid-associated regions, I implemented the method of Maller et al. (Maller et al. 2012), as described in chapter 2 and the Methods section above. For a total of 37 regions examined, there are sufficient resolution to limit the number of possible causal variants to a small informative set for three regions (log10BF>5 and # of variants <20) (**Table 6.6**).

First for the *CRP* locus, a single variant rs3091244 is predicted to be causal with posterior probability of 1. This variant was reported in the first GWAS study on CRP (Ridker et al. 2008) and it was the lead SNP in the *CRP* locus. It was reported as a tri-allelic SNP, with the common allele G and two less-common alleles of A and T. This variant was filtered out from the UK10K WGS data, but was imputed as bi-allelic for all other imputed cohorts. In the 15-way meta-analysis, the frequency for the minor A is 0.33 and there is no allele of T. rs3091244 is the only fine-mapped CRP variant that overlaps with a TFBS binding site. This might provide a functional explanation for its causality. Second, for the *HFN1A* locus, a total of 13 variants together explain 95% of the posterior probability. Based on Regulome database (http://regulome.stanford.edu), two variants (rs2259816, rs1169313) have a high score of "1f" with supporting functional data from eQTL, TF binding, DNase peak, and a third variant (rs1169310) has a score of "2b", meaning with supporting functional data from TF binding, any motif, DNase Footprint, and DNase peak. Lastly, for the *APOE* locus, although the lead SNP rs429358 based on the 15-way meta-analysis is a missense variant, fine-mapping predicted rs1065853 with a higher posterior probability for causality, posterior probability of 0.73 for rs1065853 vs. 0.23 for rs429358. Based on the Regulome database, the score is "2b" for rs1065853, meaning supporting evidence from TF binding, any motif, DNase Footprint, and DNase peak, while the score for rs429358 is "5", meaning supporting evidence only from TF binding or DNase peak. The LD among these two variants are modest (r2=0.76). rs429358 has been reported for association with Alzheimer's diseases (Kim et al. 2011, Ramanan et al. 2014), but there was no reported association for rs1065853.

Table 6.6 Putative causal variants based on fine mapping

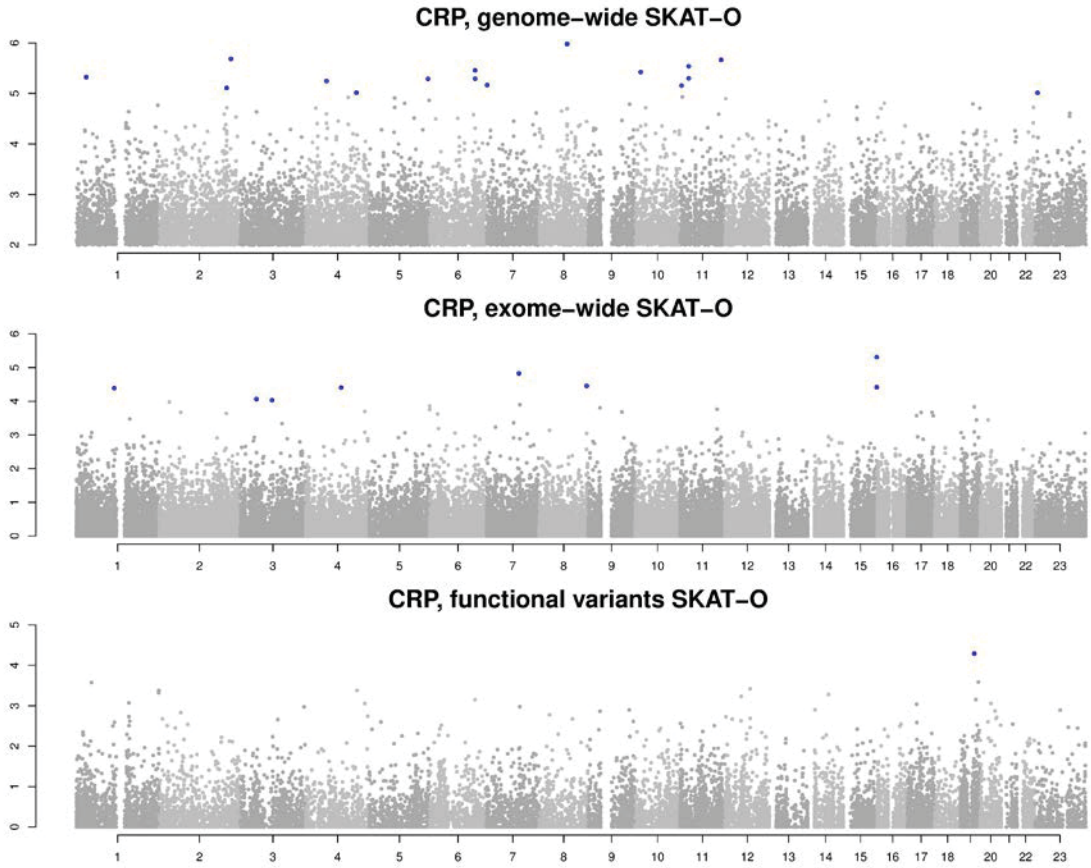
Fine-mapping						WGS 2-way					15-way				
loci	**rsID**	**CHRPOS**	**GWAVA**	**BF**	**PPA**	**EA**	**EAF**	**beta**	**SE**	**P**	**EAF**	**beta**	**SE**	**P**	**N**
CRP	rs3091244	chr1:159684665	Intronic	28.03	1.00	A	--	--	--	--	0.334	0.142	0.009	1.30E-54	31456
HNF1A	rs2264782	chr12:121432603	Upstream	19.07	0.03	T	0.354	-0.118	0.032	2.47E-04	0.372	-0.106	0.009	1.20E-34	32623
	rs2259852	chr12:121434833	3_prime_UTR	19.03	0.03	A	0.354	-0.119	0.032	2.34E-04	0.372	-0.107	0.009	5.58E-35	32622
	rs2464195	chr12:121435475	3_prime_UTR	19.03	0.03	A	0.354	-0.117	0.032	2.69E-04	0.372	-0.107	0.009	5.22E-35	32624
	rs2259816	chr12:121435587	3_prime_UTR	19.02	0.03	T	0.353	-0.118	0.032	2.58E-04	0.372	-0.106	0.009	6.63E-35	32623
	rs1169306	chr12:121438311	Downstream	18.99	0.03	T	0.357	-0.116	0.032	3.12E-04	0.374	-0.106	0.009	1.52E-34	32623
	rs735396	chr12:121438844	Downstream	19.48	0.09	C	0.353	-0.118	0.032	2.52E-04	0.373	-0.107	0.009	3.31E-35	32624
	rs1169309	chr12:121439192	3_prime_UTR	19.07	0.03	T	0.354	-0.119	0.032	2.33E-04	0.372	-0.106	0.009	6.17E-35	32623
	rs1169310	chr12:121439433	3_prime_UTR	19.48	0.09	A	0.354	-0.120	0.032	1.95E-04	0.373	-0.107	0.009	2.49E-35	32622
	rs1169311	chr12:121440731	3_prime_UTR	19.49	0.09	T	0.355	-0.119	0.032	2.03E-04	0.373	-0.107	0.009	2.53E-35	32623
	rs1169312	chr12:121441461	3_prime_UTR	19.39	0.07	T	0.356	-0.119	0.032	2.17E-04	0.375	-0.107	0.009	1.63E-35	32623
	rs1169313	chr12:121442670	Downstream	19.73	0.15	C	0.356	-0.118	0.032	2.25E-04	0.376	-0.107	0.009	1.41E-35	32623
	rs112249815	chr12:121444441	Exon	19.79	0.17	C	0.357	-0.116	0.032	3.19E-04	0.376	-0.107	0.009	1.47E-35	32620
	rs2257962	chr12:121445808	upstream_gene	19.69	0.14	C	0.356	-0.119	0.032	2.13E-04	0.375	-0.108	0.009	9.66E-36	32623
APOE	rs429358	chr19:45411941	Missense	14.83	0.23	C	0.142	-0.169	0.046	2.75E-04	0.142	-0.193	0.012	4.77E-55	32621
	rs1065853	chr19:45413233	upstream_gene	15.33	0.73	G	0.133	-0.232	0.053	1.09E-05	0.121	-0.202	0.014	2.96E-45	32620

6.3.3 Novel loci based on rare variants aggregation test

No variants are significant from the three types of SKAT-O tests, by using genome-wide significance threshold of $P < 6.8E-08$, 1.2E-06, 1E-05 respectively for genome-wide, exome-wide, and functional variants based SKAT-O (**Figure 6.4**). For those regions reaching less stringent threshold for suggestive association, as highlighted in blue in **Figure 6.4**, none of them are within 1Mb of known CRP loci. This could indicate truly a lack of rare variant that have large effects on serum CRP level, or due to inadequate power of the WGS samples used in this study.

Figure 6.4 Rare variants aggregation test results for CRP

The genome-wide significant signals are shown in red, with threshold of $P < 6.8E-08$, 1.2E-06, 1E-05 respectively for genome-wide, exome-wide, and functional variants based SKAT-O. Suggestive signals are shown in blue, with threshold of $P < 1E-05$, 1E-04, 1E-04 respectively for genome-wide, exome-wide, and functional variants based SKAT-O.



6.4 Conclusion & Discussion

6.4.1 Summary of main findings

Using 2,046 samples with WGS data and CRP levels measured (879 for TwinsUK, 1167 for ALSPAC), I applied a combination of approaches to conduct a genome-wide discovery of novel variants of low frequency associated with CRP level. Here, I identified two low frequency novel variants (MAF =2%) with *P*< 1E-6 but they were not replicated using imputed data. Then, I included up to ~73,000 samples with mostly imputed data to discover novel CRP associations across the full allele frequency spectrum. Here, I was able to discovery two novel associations. The first one is a common variant in the *HIST1H3G* locus (rs9393691, MAF=0.383, 27-way meta-analysis *P*=1.71E-11). The second one is a low frequency intronic variant within *MYO5A (*rs117410733, MAF=0.009, 27-way meta-analysis *P*=5.78E-09*)*. Fine-mapping analysis coupled with functional annotation narrowed down to putative causal variants within *CRP* and *APOE*. Rare variants aggregation tests did not identify putative novel loci that meet pre-defined genome-wide significance threshold.

6.4.2 Interpretation of results

The single marker association testing of CRP follows closely the expected relationship between EAF and effect size (beta) as dictated by study power (Park et al. 2011), as shown in **Figure 6.5**. Low frequency alleles of very high penetrance (beta ~1 SD) are unlikely to exist within this allelic space in the general European-ancestry population. Given that the genome-wide 21-way meta-analysis with a sample size more than 45,000, a number comparable to the previously published largest GWAS on CRP, using a combination of WGS samples and WGS imputed samples does not seem to be able to discover substantially more novel associations, either common or rare. The strongest association signal from single marker based analysis is a common variant within a gene-rich region, *HIST1H3G*. This association stood out only in the 27-way meta-analysis with a sample size of ~73,000. This implies that increasing sample size to this level for genome-wide association analysis could still be valuable. Given the gene-rich nature of this region and its association with many CVD related traits including lipids and FBC, targeted resequencing of this region and further

functional annotations are important steps for firmly establish this association and the understanding the underlying biology.

6.4.3 Future direction

Due to the constraint of time and resource, there is no independent WGS data was obtained to replicate some of the loci with suggestive evidence for rare variants aggregation based association. This could be an area worth further research efforts. Across, analysing multiple inflammatory traits together in a multivariate approach might discover common associations and pathways under the inflammation process in general. This could include the study of CRP and WBC traits together. IL-6 was also one of the 64 traits in UK10K, but its sample size is limited, only existing in ALSPAC and few of the external cohorts that were made available for expanded discovery and replication.

In this thesis, the study of CRP is overall separate from the other 12 CVD biomarkers. However, in the future a study combining CRP and lipids especially LDL would be desirable to fully understand their joint effects and interactions. It was reported that adding CRP to LDL in cell culture systems stimulated formation of foam cells, a typical feature of atherosclerotic plaques (Zwaka et al. 2001). However, it is not known whether this reflects opsonization of the LDL particles by CRP or an effect of CRP on the phagocytic cells themselves. Binding of CRP to lipids, especially lecithin (phosphatidyl choline), and to plasma lipoproteins has been known for decades. This could suggest new measurement of lipids bound CRP as the studied trait in genotype-phenotype association studies that aim to discover genetic factors underlying inflammatory process and CVD risk in general. Also, the co-analysis of CRP and WBC could also be explored. Previously, many studies have investigated both CRP and WBC for association with the risk of various diseases and aging prognosis (Keskin et al. 2004, Santos et al. 2004, Peltola et al. 2006, Willems et al. 2010).

Although the focus of this PhD thesis is on CVD related biomarkers, CRP could build a bridge between CVD and the other chronic disease with tremendous public health burden, cancer. Epidemiologic studies suggest that in patients with several types of solid cancers, elevated circulating levels of CRP are associated with poor prognosis, whereas in apparently healthy individuals from the general population, elevated levels of CRP are associated with increased future risk of cancer of any type. While most MR studies have failed to establish a causal role of serum CRP level to the development of CVD, a recent MR study provided

promising results for establishing a causal role of serum CRP levels to colorectal cancer (Nimptsch et al. 2015).

Figure 6.5 Statistical power and novel variants from single marker analysis

The top and bottom plots are for WGS samples and expanded discovery samples respectively. Y-axis is a variant's effect, expressed in standard deviation units. X-axis is MAF of effect alleles. Colored lines indicate 20%, 50%, and 80% power. Alpha is set at *P*<1E-06 for WGS and *P*<1E-07 for expanded discovery respectively. The two putative novel WGS variants are shown in the top power plot for WGS, and the eight putative novel variants from expanded discovery are shown in the bottom power plot for expanded discovery.