5 Full Blood Counts

5.1 An introduction to full blood counts

5.1.1 Biology and physiology of FBC

Blood cells play major roles for a variety of essential physiological functions. Among their many functions, red blood cells (RBC) transport oxygen, white blood cells (WBC) are engaged with some of the immune and inflammatory responses, and platelets (PLT) form blood clots to prevent excessive bleeding. RBC, WBC, PLT are also called erythrocytes, leukocytes, thrombocytes, respectively. All these blood cells, also called hematocytes, are produced by hematopoiesis (Orkin and Zon 2008). Circulation levels of blood cells are commonly measured in clinical visits and regular physical check-ups, because they are easily measured and an abnormal number or size or feature of the them are indicators of multiple human diseases. Very low level of RBC and hemoglobin (HGB) is the direct causes of anemia; rapid production of abnormal white blood cells causes leukemia; low level of PLT counts causes thrombocytopenia. There are a few other commonly measured RBC related traits, including haemoglobin (HGB), mean cell haemoglobin (MCH), mean cell haemoglobin concentration (MCHC), mean cell volume (MCV), packed cell volume (PCV). Although these traits are highly correlated, assaying multiple traits simultaneously could provide refined insights into path-physiological process. For example, a decrease of both MCV and MCH suggests a problem in hemoglobin production caused by iron deficiency or ineffective synthesis of globin polypeptides. WBCs are classified into five subtypes based on their morphology and functions, including neutrophils, basophils, eosinophils, lymphocytes and monocytes. Determination of platelet size, usually via quantification of mean platelet volume (MPV), is a simple and easy method of accurately assessing platelet function. In some genetic studies, both MPV and PLT were used as phenotypes.

Although environmental factors especially poor nutrition and infections casuse abnormal blood cells, genetics play a major role for both severe blood disorders and normal variation of blood cell levels in healthy individuals. For example, mutations in *G6PD* cause chronic hemolytic anemia, and mutations in oncogenes or tumor suppressor genes cause leukemia.

5.1.2 FBC traits as risk factors for CVD

FBC is a commonly used screening for indicators of health and disease.

RBC traits and risk for CVD

RBC is directly related to cardiovascular performance. The use of exogenous EPO has been reported in athletes to boost performance. Anemia, defined as HGB <11 g/dL in women or <13 g/dL in men, is the most common form that ranges from mild fatigue to heart failure (Greenburg 1996). The World Health Organization estimates that anemia affects 1.62 billion people in the world, as of the end of 2013. The main causes of anemia are poor nutrition and iron deficiency, infections (e.g., malaria) and RBC diseases including hemoglobinopathies. Since anemia is mostly frequent in Africa and South-East Asia, it is critical to search for genetic associations with hemoglobin levels in these populations.

WBC and risk for CVD

WBC count is used as a clinical marker of inflammation status. Patients with elevated WBC have been shown to be in a higher risk of developing acute MI and acute coronary and vascular events. Measuring WBC and its sub-phenotypes could be used for a better way of risk stratification of patients admitted with acute vascular events (Hoffman et al. 2004). High WBC has been associated with an increased risk of CVD (Danesh et al. 1998), cancer mortality (Shankar et al. 2006) and all-cause mortality (Ruggiero et al. 2007). Elevated WBC is also associated with disease risk factors including increasing age, high BP, cigarette smoking, adiposity and increasing plasma inflammatory markers (Nieto et al. 1992). The association of WBC with cardiovascular risk factors may either represent manifestation of subclinical disease or suggest that WBC is part of the causal chain leading to atherosclerosis. More recently, it was reported that WBC count is also a predictor of fatal and nonfatal ischemic vascular disease independent of other CHD risk factors (Campbell et al. 2012).

PLT and CVD

Coronary atherosclerosis is a highly complex chronic inflammatory disease that may convert into an acute clinical event, especially in acute coronary syndromes (ACS) which occur secondary to atherosclerotic plaque rupture and subsequent vessel ischaemia (Tiong and Brieger 2005). PLT not only contribute to acute thrombotic vascular occlusion but also participate in the inflammatory and matrix-degrading processes of coronary atherosclerosis itself. Platelet- endothelial cell interactions at lesion-prone sites might trigger an inflammatory response in the vessel wall early in the genesis of atherosclerosis and contribute to destabilization of advanced atherosclerotic lesions (Massberg et al. 2003). PLT is also involved in the pathology of acute stroke, since early platelet adhesion/activation mechanisms are critical pathogenic factors in infarct development and trigger a thrombo-inflammatory cascade in acute stroke that results in infarct growth.

There is an abundance evidence for PLT's involvement and association with CVD. In 1986, it was first reported that a decrease of PLT and an increase of MPV correlated with infarct size (Glud et al. 1986). Abnormalities of platelet function may contribute to the relatively poor prognosis of myocardial infarction in patients with diabetes (Hendra et al. 1988), and vascular and nonvascular death (Thaulow et al. 1991). Some other associations are especially with MPV but not PLT count. Larger platelets have a greater mass and a greater prothrombotic potential than smaller platelets. The larger and more reactive platelets are enriched in individuals with known CAD risk factors including hypercholesterolaemia (Pathansali et al. 2001) and hypertension (Nadar et al. 2004), and might be causally related to ongoing coronary artery obstruction in unstable angina (Pizzulli et al. 1998). However, in spite of the strong link between MPV and increased CAD risk, there is no data from clinical trials to show that reducing MPV could bring favourable CAD outcomes.

5.1.3 Genetic determinants of FBC

It is estimated that the heritability is 0.67, 0.38, 0.53 for RBC, WBC, PLT respectively, based on a study with >6,000 healthy Sardinians (Pilia et al. 2006). A Twin study showed slightly different numbers especially for WBC, with 0.37, 0.42, 0.62, and 0.57 for HGB, RBC, WBC, PLT respectively (Garner et al. 2000). Blood cell traits are particularly well-suited for genetic association studies and functional follow-up because they are usually available in most cohorts or biobanks and there are well-developed cell culture systems or model organisms. Large-scale gene silencing and other functional experiments in fruit flies, zebrafish and mice were already shown to be effective for validating genetic loci identified by GWAS (Gieger et al. 2011, van der Harst et al. 2012).

Findings from candidate gene and linkage analysis

Candidate gene studies identified a few loci for association with FBC. The first well studied gene is HBB (β-globin). Mutations in this gene are implicated with several genetic disorders such as sickle-cell disease and beta thalassemia. Other mutations in this gene also bring beneficial effects such as genetic resistance to malaria (Kwiatkowski 2005). Mutations were also found in two other genes: mutations in *EPOR* (erythropoietin receptor) causing familial erythrocytosis (Watowich et al. 1999, Zeng et al. 2001), and mutations in *HFE* (hemochromatosis) causing hereditary hemochromatosis (McLaren et al. 2007). Linkage studies also identified a few reproducible signals, most notably a linkage peak that encompasses the MYB transcription factor (Lin et al. 2007, Menzel et al. 2007).

Findings from first generation GWAS

As shown in **Table 5.1**, a total of 25 GWAS have been conducted for FBC related traits since 2008. The largest studies of blood cells, based on individuals of European ancestry, have so far identified 75, 10 and 68 SNPs for RBC (van der Harst et al. 2012), WBC (Nalls et al. 2011), and platelet traits (Gieger et al. 2011) respectively. There are much fewer associated loci for WBC because its GWAS had a smaller sample size and there is heterogeneity among WBC sub-phenotypes. Like GWAS for other quantitative traits such as lipids, the variants discovered from blood cell GWAS explained a small fraction of the heritable variation (<10%). Also, like the lipids traits, most loci are associated with a single blood cell trait while a few presented pleiotropic effects. This includes two loci (*SH2B3*, *HBS1L-MYB*) associated with all three blood cell traits, both of which have clear biological impact on hematopoiesis.

Again, like lipids traits, many variants discovered through GWAS for association with FBC are within or near genes that are causal for Mendelian hematological disorders, for example, SNPs near *TMPRSS6*, *HFE*, *TRF2* (for iron deficiency), *HK1* (for hemolytic anemia), and *TBUU1* (for throbocytopenia). Due to the much denser scanning of the genome compared to linkage studies, GWAS was able to pinpoint stronger candidate genes for some of these overlapping loci. Unlike lipids traits or many other traits, where GWAS loci and effects are comparable among multiple ethnic groups (Monda et al. 2013), there are notable exceptions for FBC traits. For example, genetic variants near the gens of α-globin, β-globin and *G6PD* are much more common in African populations because they provide a selective advantage against malaria infections.

Findings from next generation sequencing

No studies have been reported using next generation sequencing.

Table 5.1 GWAS studies on FBC traits

Date is for publication date. Samples are all European ancestry unless explicitly specified otherwise: IND for Indian, JAP for Japanese, AA for African American. The sample size before "+" is for discovery while the sample size after "+" is for replication.

Date	Sample	Main findings	Reference
2008-11	1,062	No SNPs associated with FBC traits at P<5E-08	(Yang et al. 2007)
2008-12	411 from families and 459 twins	Variants in TF and HFE explain ~40% of genetic variation in serum-transferrin levels	(Benyamin et al. 2009)
2008-12	1,606+8,617	Identified 3 loci associated with MPV	(Meisinger et al. 2009)
2009-02	1,221+7,365	A variant on 7q22.3 for MPV and PLT	(Soranzo et al. 2009)
2009-10	4,627+9,316	22 loci for 8 hematological parameters	(Soranzo et al. 2009)
2009-10	16,001 EA and IND	Missense variant in TMPRSS6 for HGB	(Chambers et al. 2009)
2009-10	4,818+3470	Variants in TMPRSS6 are associated with iron status	(Benyamin et al. 2009)
2009-10	3,477+1543	3 loci for monocyte counts and erythrocyte volume	(Ferreira et al. 2009)
2009-10	24,167+9,456	5 know loci, 18 novel loci	(Ganesh et al. 2009)
2010-02	14,700 JAP	46 new and 43 known associations	(Kamatani et al. 2010)
2010-09	3012	demonstrate feasibility of using EMR for GWAS	(Kullo et al. 2010)
2011-03	679+232	2 replicated loci for iron deficiency	(McLaren et al. 2011)
2011-06	8,794+5998 JAP	nine novel loci associated with WBC subtypes	(Okada et al. 2011)
2011-06	16,388 AA	CXCL2, CDK6, PSMD3-CSF3 associated with WBC	(Reiner et al. 2011)
2011-07	19,509+11,823	7 loci associated with WBC	(Nalls et al. 2011)
2011-10	13,923	2 loci each for EA and AA, for WBC	(Crosslin et al. 2012)
2011-11	~18,600+18838	68 loci reliably for PLT and MPV	(Gieger et al. 2011)
2012-03	16,388 AA	5 novel loci for PLT	(Qayyum et al. 2012)
2012-12	62,553 +63506	75 loci for RBC	(van der Harst et al. 2012)
2012-12	62,34EA and 7943 AA	5 novel loci for EA RBC, 1 novel for AA PLT	(Li et al. 2013)
2013-02	16,485	Extended several RBC loci from EA to AA	(Chen et al. 2013)
2013-03	11,014	4 novel loci for monocyte count	(Crosslin et al. 2013)
2013-05	1,904+411 AA	malaria resistance variants associated with RBC	(Ding et al. 2013)
2013-07	1,664+2,200	Identified TAF3 as a gene for MCHC	(Pistis et al. 2013)
2013-09	13,582	EMR based, no new loci reported	(Shameer et al. 2014)

5.1.4 Aims of this study

To discover novel variants, especially those with low or rare frequency but large effects, this study used WGS data from the UK10K project for an upgraded genome-wide scan on eight FBC traits (RBC, HGB, MCH, MCHC, MCV, PCV, PLT, WBC). The current study is by far the largest WGS based association study of FBC traits, with up to 1,497 WGS samples and more than 21,000 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 standard approaches including single marker based test and rare variants collapsing test, this study also explored a few novel methods for a comprehensive assessment on the genetics of FBC. This included fine-mapping of known loci to identify causal variants, assessing enrichment in various functional and regulatory features, and an exploring of relationship between genetic variants associated with FBC and host response to infectious diseases including tuberculosis and malaria (Ding et al. 2013, McMorran et al. 2013)

5.2 Methods

5.2.1 Cohorts & phenotype measurements

The phenotype harmonization protocol for the FBC traits in TwinsUK was presented in **Table 5.2**. For TwinsUK, previously I separated it into TwinsUK WGS samples and TwinsUK imputed sample for lipids analysis. As mentioned in **Section 2.6**, after running an evaluation on lipids traits, I combined the genetic data for TwinkUK WGS and imputed samples together so that the co-Twins could also be included for analysis. This is necessary since there was a relative small number of samples available for FBC traits. A total of 12 cohorts were included for the expanded discovery for FBC, and six WHI cohorts of European ancestry were included for replication (**Table 5.3**). All these WHI cohorts had genome-wide results available.

For ALSPAC, HGB were measured with Hemocue Hb201+ analyser. For all eight FBC traits in TwinsUK and all other discovery cohorts, the traits were measured with Beckman Coulters, except for WHI, where HGB, HCT, WBC, and PLT were determined at local laboratories using automated hematology cell counters and standardized quality assurance procedures (Margolis et al. 2005). Different phenotype transformation protocol was applied to the eight FBC phenotypes: inverse normal transformation for HGB, PCV, PLT, square root for MCH, natural log for WBC, and no transformation for MCHC, MCV, RBC. 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.

Table 5.2 Phenotype harmonization protocol for FBC traits

Analyser and visit were tested as random effect variables, while the others including age and age^2 are tested as fixed effect covariates.

Dataset	Trait	Transformation	Gender	Co-variates tested	Filter	Analyser	Visit
TwinsUK GWA	WBC	Natural log	no	age, age^2, sex,dov	3 SD	--	no
TwinsUK WGS	WBC	Natural log	--	age, age^2, dov (2 and 3 periods)	3 SD	--	no
TwinsUK GWA	MCH	Square	no	age, age^2, sex,dov	3 SD	--	yes
TwinsUK WGS	MCH	Square	--	age, age^2, dov (2 and 3 periods)	3 SD	--	yes (3 periods)
TwinsUK GWA	MCHC	untransformed	no	age, age^2, sex,dov	3 SD	--	no
TwinsUK WGS	MCHC	untransformed	--	age, age^2, dov (2 and 3 periods)	3 SD	--	no
TwinsUK GWA	MCV	untransformed	no	age, age^2, sex,dov	3 SD	--	yes
TwinsUK WGS	MCV	untransformed	--	age, age^2, dov (2 and 3 periods)	3 SD	--	yes (3 periods)
TwinsUK GWA	PCV	inverse normal	no	age, age^2, sex,dov	3 SD	--	yes
TwinsUK WGS	PCV	inverse normal	--	age, age^2, dov (2 and 3 periods)	3 SD	--	yes (3 periods)
TwinsUK GWA	PLT	inverse normal	no	age, age^2, sex,dov	3 SD	--	yes
TwinsUK WGS	PLT	inverse normal	--	age, age^2, dov (2 and 3 periods)	3 SD	--	yes (2 periods)
TwinsUK GWA	RBC	untransformed	no	age, age^2, sex,dov	3 SD	--	yes
TwinsUK WGS	RBC	untransformed	--	age, age^2, dov (2 and 3 periods)	3 SD	--	yes (3 periods)

5.2.2 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 cohort, with up to 1,497 samples (**Table 5.3**). For HGB, genome-wide association for the ALSPAC WGS samples were also run and were then meta-analyzed with the TwinUK WGS results. Variants with *P*<1E-6 are deemed of interest for follow-up and further characterization. To discover novel variants across the full MAF spectrum, I included up to 10 more cohorts with imputed data in a 12-way meta-analysis, followed by a replication meta-analysis with up to six independent cohorts from WHI (**Table 5.3**). The WHI data only included four phenotypes: HGB, PCV, PLT, WBC. The 12-way meta-analysis included up to 21,519 samples, while the 6-way replication meta-analysis included up to 20,038 samples. Due to the relatively small number of sample size in the 12-way expanded discovery and given the availability of the full genome-wide results of the 6-way replication cohorts, I also run a further expanded discovery meta-analysis for those four traits (HGB, PCV, PLT, WBC) in a 18-way meta-analysis with up to 41,557 samples. For this 18-way meta-analysis, there was no further data for replication.

The TwinsUK WGS and GWA samples were imputed and analyzed together with GEMMA by adjusting for sample genotype status. As described in the Methods chapter, this included all TwinsUK samples for the association analysis and showed better power than analyzing WGS and imputed samples separately where related samples across the two datasets would have to be excluded. A few in-house GWAS results (from the HaemGen consortium) on these traits were also made available to serve as a more comprehensive list of positive controls.

5.2.3 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. The method for rare variant aggregation based test was the same as that used for lipids, except that the meta-analysis was only run for HGB since it was the only FBC trait measured and analysed in both TwinsUK and ALSPAC. 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. There was no external data available for rareMetal analysis to replicate windows of interest for the FBC traits.

5.2.4 Fine-mapping of known loci

The fine-mapping method was described in chapter 2 and it is the same as that used for lipids. 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). For each FBC trait 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 10 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 were 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 were constructed in order to assess the uncertainty of the fine-mapping analysis.

Table 5.3 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 each trait of each cohort, the residuals were standardized so that the phenotype has a mean of 0 and a standard deviation of 1.

	Cohort	**N**	**Country**	**Age**	**% Female**	**HGB (g/dl)**	**RBC (10^{12}/l)**
Discovery	TwinsUK	1,497	UK	56 (17-85)	97.3	13.34 (1.02)	4.47 (0.35)
	ALSPAC WGS	1,713	UK	10 (9-11)	50.3	14.22 (1.10)	--
	ALSPAC GWA	1,896	UK	10 (9-12)	49.2	13.98 (0.97)	--
	CBR	5,493	UK	45 (34-67)	58.2	14.73 (0.93)	4.97 (0.34)
	INGI-CARL	413	Italy	50 (18-83)	60.0	15.11 (0.96)	4.65 (0.31)
	INGI-FVG	1,377	Italy	52 (18-92)	58.2	14.56 (0.87)	4.62 (0.28)
	INGI-VB	1,776	Italy	55 (18-102)	56.3	13.96 (0.87)	4.30 (0.27)
	HELIC-Manolis	1,247	Greece	62 (18-99)	57.2	15.10 (0.92)	4.41 (0.30)
	HELIC-Pomak	976	Greece	43 (13-87)	72.1	14.65 (0.86)	4.33 (0.29)
	UKBS	2,070	UK	43 (35-62)	54.1	14.03 (0.77)	4.35 (0.27)
	LURIC-Case	1,633	Germany	61 (17-91)	60.8	13.95 (0.91)	4.82 (0.37)
	LURIC-Ctrl	1,428	Germany	62 (18-92)	59.7	14.02 (1.01)	4.61 (0.35)
Replication	WHI-Garnet	3,821	US	65 (50-79)	100.0	14.01 (0.89)	--
	WHI-Gecco1	1,992	US	65 (50-79)	100.0	13.76 (0.93)	--
	WHI-Gecco2	1,737	US	65 (50-79)	100.0	14.04 (0.99)	--
	WHI-Hipfx	3,825	US	65 (50-79)	100.0	14.03 (1.00)	--
	WHI-Mopmap	3,031	US	65 (50-79)	100.0	13.55 (0.79)	--
	WHI-Whims	5,632	US	65 (50-79)	100.0	13.98 (0.93)	--

	Cohort	**MCH (pg)**	**MCHC (g/dl)**	**MCV (fl)**	**PCV (l/l)**	**PLT (10^9/l)**	**WBC (10^9/l)**
Discovery	TwinsUK	29.92 (1.76)	32.35 (1.38)	83.65 (4.01)	0.43 (0.05)	253.9 (63.1)	6.10 (1.81)
	ALSPAC WGS	--	--	--	--	--	--
	ALSPAC GWA	--	--	--	--	--	--
	CBR	29.73 (1.73)	33.19 (1.03)	89.57 (3.98)	0.49 (0.03)	232.9 (50.9)	6.34 (1.52)
	INGI-CARL	26.34 (1.65)	34.43 (0.99)	92.11 (3.67)	0.47 (0.04)	287.5 (48.8)	6.33 (1.45)
	INGI-FVG	25.82 (1.39)	35.12 (1.10)	87.56 (3.01)	0.46 (0.04)	301.0 (59.1)	7.01 (1.44)
	INGI-VB	30.11 (1.76)	32.78 (1.03)	89.22 (2.99)	0.44 (0.06)	297.4 (49.7)	5.43 (1.21)
	HELIC-Manolis	27.82 (1.68)	33.94 (0.89)	94.01 (3.22)	0.47 (0.05)	221.9 (53.2)	6.10 (1.70)
	HELIC-Pomak	29.01 (1.59)	34.21 (1.21)	89.76 (2.78)	0.50 (0.04)	254.7 (55.4)	7.02 (1.81)
	UKBS	29.88 (1.72)	31.27 (1.03)	93.21 (3.21)	0.48 (0.06)	261.2 (57.3)	5.06 (1.20)
	LURIC-Case	30.21 (1.81)	34.77 (0.95)	87.45 (3.04)	0.45 (0.05)	277.4 (61.0)	6.43 (1.43)
	LURIC-Ctrl	29.01 (1.77)	32.19 (0.97)	91.02 (3.66)	0.45 (0.07)	310.7 (67.3)	5.98 (1.55)
Replication	WHI-Garnet	--	--	--	0.46 (0.05	302.0 (58.9)	4.97 (1.09)
	WHI-Gecco1	--	--	--	0.47 (0.06)	298.3 (54.0)	6.03 (1.42)
	WHI-Gecco2	--	--	--	0.49 (0.03	276.9 (48.8)	6.11 (1.39)
	WHI-Hipfx	--	--	--	0.43 (0.05)	320.1 (59.7)	7.02 (1.83)
	WHI-Mopmap	--	--	--	0.47 (0.04)	300.7 (56.3)	5.32 (1.56)
	WHI-Whims	--	--	--	0.48 (0.05)	288.4 (48.7)	5.05 (1.76)

5.3 Results

5.3.1 Novel loci and novel variants from single marker analysis

WGS for low frequency and rare variants

Here I sought to investigate if 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. For HGB, data from TwinsUK and ALSPAC was meta-analysed.

A total of 60 variants have *P*<5E-08, based on TwinsUK WGS samples alone (**Figure 5.1**). 57 of these variants are in the *HBS1L* (HBS1-like translational GTPase) region. *HBS1L* encodes a member of the GTP-binding elongation factor family, mostly expressed in heart and skeletal muscle. The intergenic region between *HBS1L* and *MYB* is a quantitative trait locus (QTL) that controls fetal hemoglobin level and influences erythrocyte, platelet, and monocyte counts. The other three variants with *P*<5E-08 were not previously reported for associations with blood cell traits. The first one is a low frequency variant for association with PCV (rs114119841, chr2:38831057, EA=C, EAF=0.020, *P*=3.20E-08). This is annotated as a regulatory region variant for *HNRPLL* (heterogeneous nuclear ribonucleoprotein L-like). *HNRNPLL* is a master regulator of activation-induced alternative splicing in T cells. It alters the splicing of a tyrosine phosphatase that is essential for T-cell development and activation (Oberdoerffer et al. 2008). However, this variant was not replicated either, with *P*>0.05 in the 10-way meta-analysis. The second one is a common variant within *COL23A1*, for association with HGB. The index SNP is rs4976769 (chr5:177808188, EA=G, EAF=0.065, *P*=3.85E-08). This variant has a meta-analysis *P*=3.75E-04, based on a total of 10 cohorts and 16,687 samples. Given its allele frequency and the sample size in the 10-way meta-analysis, this signal was not replicated and might be a false positive. The third one is a rare variant for association with MCHC (rs145884292, chr9:24195910, EA=C, EAF=0.008, *P*=2.91E-08). The index SNP (rs145884292) is an intergenic variant, ~5 Mb away from *ELAV2* (embryonic lethal, abnormal vision, Drosophila-like 2), which has no apparent relevance to blood traits.

To look at suggestive associations, I used a less stringent threshold and discovered a further 155 variants have *P*<1E-06, as highlighted blue in **Figure 5.1**. For these 215 variants in total, 25 have MAF between 0.005 and 0.05 (**Table 5.4**). For the given number of WGS samples for FBC traits and the sequencing coverage, there was a high probability (>98%) of detecting variants down to MAF of 0.5% (Li et al. 2011). Among these 25 variants, rs62064540 (for association with MCH) is in proximity to a previously report association with MCHC (rs689992), and rs113833421 is in proximity to previously reported variant rs11672923 (for association with RBC). But there was no LD between the current study's index SNPs and previously reported variants (r2< 0.01) in both cases. The rest 23 variants were not within 1MB of any positive controls. Given the lack of independent replication cohorts with directly sequenced or de novo genotyped data for these variants, I presented the expanded discovery meta-analysis (12-way) results for these variants (**Table 5.4)**. However, none of these 25 variants became more significant in the expanded discovery meta-analysis, all of which had *P*>1E-4 in the 12-way meta-analysis. This could be due to poor imputation or lack of power for replication, or these signals are false positive. Preferably, WGS or directly typed genotype should be used as replication for this set of variants, when resources become available.

Figure 5.1 Association results for WGS based samples for FBC traits

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.

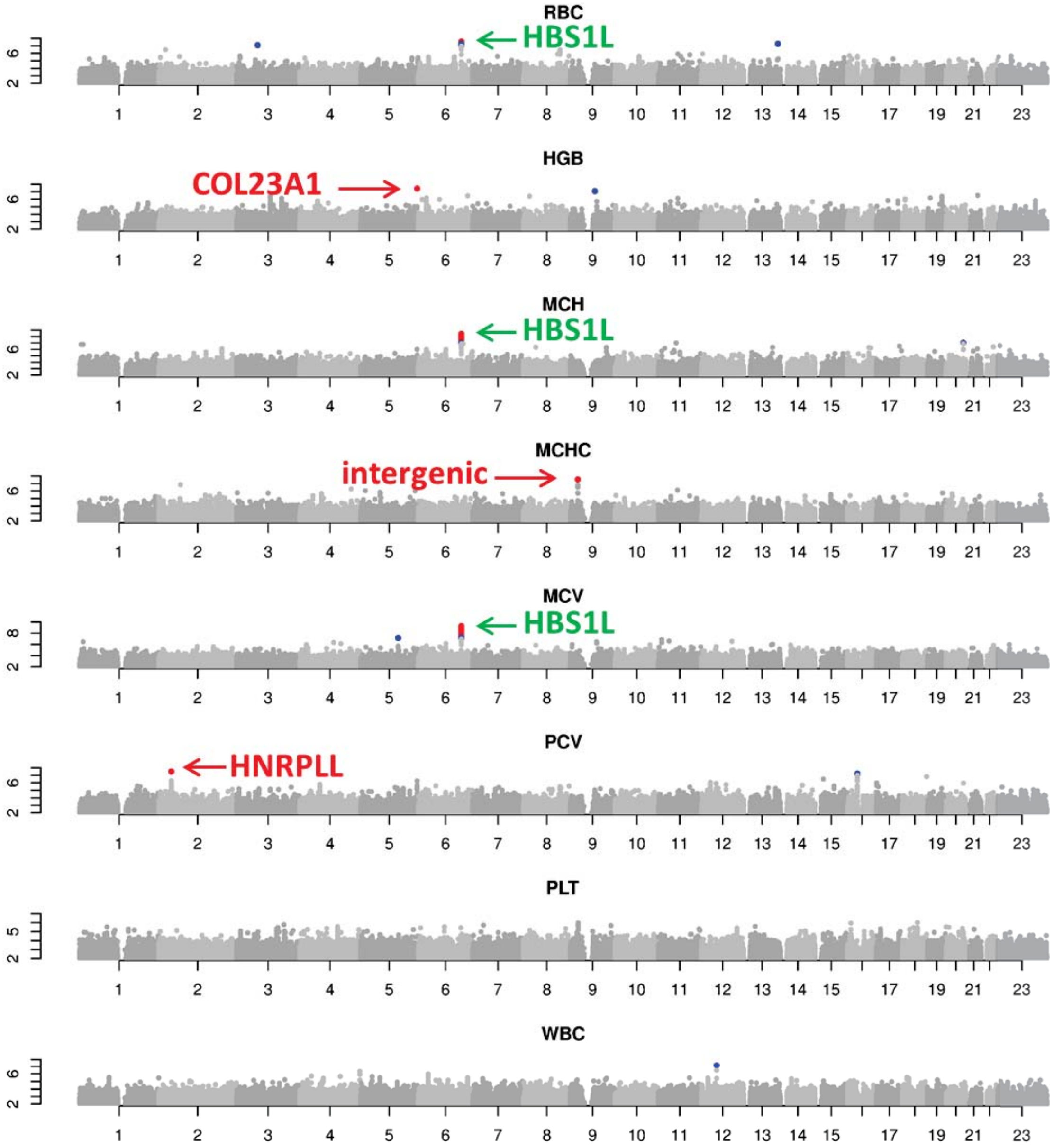


Table 5.4 Putative novel variants of low or rare frequency from UK10K WGS

25 WGS variants (P<1E-6) either have no positive controls within 1Mb or are independently significant from known variants. Six have low frequency (MAF between 1-5%) and could be imputed with fair accuracy.

						WGS				12-way meta-analysis				
trait	**rsID**	**CHR**	**POS**	**EA**	**NEA**	**EAF**	**beta**	**SE**	**P**	**EAF**	**beta**	**SE**	**P**	**N**
RBC	rs76777478	2	20383810	T	G	0.048	0.427	0.083	3.23E-07	0.047	0.042	0.025	9.7E-02	13944
PCV	rs114119841	2	38831057	C	A	0.020	0.705	0.127	3.20E-08	0.025	0.059	0.037	1.2E-01	15350
MCHC	rs146621801	2	67557307	G	C	0.013	-1.095	0.207	1.5E-07	0.014	-0.030	0.055	5.9E-01	12891
MCH	rs186149310	2	197653260	G	A	0.012	-0.845	0.170	7.8E-07	0.010	-0.121	0.094	2.0E-01	12189
RBC	rs189761618	3	66373898	A	G	0.015	0.821	0.152	8.1E-08	0.009	0.083	0.064	2.0E-01	12956
HGB	rs11917207	3	105964555	G	A	0.046	0.296	0.059	6.3E-07	0.047	0.087	0.025	4.7E-04	19751
MCV	rs145802933	4	106800944	G	C	0.008	1.030	0.203	4.6E-07	0.007	0.208	0.094	2.9E-02	15280
MCHC	rs74339994	4	161780587	A	T	0.008	-1.325	0.263	5.4E-07	0.012	-0.039	0.067	5.6E-01	12892
WBC	rs76070316	4	189101798	G	C	0.007	1.044	0.206	4.6E-07	0.014	-0.018	0.026	4.9E-01	15340
MCHC	rs188771831	5	15237510	G	T	0.015	-0.962	0.195	9.2E-07	0.009	-0.033	0.079	6.8E-01	12892
MCV	rs72663338	5	118080521	G	A	0.042	-0.514	0.095	7.3E-08	0.042	-0.065	0.032	4.6E-02	15281
MCHC	rs74964545	5	171263478	T	C	0.012	1.092	0.221	9.4E-07	0.012	0.051	0.070	4.7E-01	12891
MCH	rs6862184	5	177396364	A	G	0.959	-0.442	0.090	9.7E-07	0.963	-0.072	0.041	7.9E-02	12190
MCV	rs181579991	6	87074470	A	G	0.009	0.985	0.198	7.3E-07	0.006	0.159	0.096	1.0E-01	14327
HGB	rs62434477	6	155327584	T	C	0.017	-0.509	0.100	3.5E-07	0.017	-0.131	0.047	5.6E-03	19752
MCHC	rs145884292	9	24195910	C	T	0.008	-1.452	0.260	2.9E-08	0.007	-0.142	0.086	1.0E-01	12892
HGB	rs75472650	9	77788213	T	C	0.008	-0.798	0.149	8.3E-08	0.007	-0.217	0.071	2.5E-03	19749
HGB	rs72914272	11	61376274	T	C	0.033	0.358	0.072	7.8E-07	0.027	0.052	0.036	1.6E-01	19751
PCV	rs11829947	12	28334475	C	T	0.012	-0.784	0.159	9.2E-07	0.011	-0.055	0.053	3.0E-01	15350
RBC	rs117125854	13	106362073	A	G	0.006	1.242	0.227	5.3E-08	0.006	0.090	0.071	2.1E-01	13942
PCV	rs67824122	15	26248846	T	A	0.007	-1.422	0.277	3.2E-07	0.007	-0.221	0.091	1.6E-02	15349
MCH	rs62064540 *	17	72171888	C	T	0.008	-1.066	0.208	3.1E-07	0.007	-0.181	0.096	6.0E-02	12190
HGB	rs62087096	18	8998320	T	A	0.037	0.326	0.066	9.1E-07	0.032	0.028	0.029	3.4E-01	19750
PCV	rs148652300	18	76407934	T	C	0.006	1.227	0.233	1.5E-07	0.006	0.203	0.105	5.6E-02	14867
HGB	rs113833421 *	19	46421564	T	C	0.011	0.607	0.120	4.2E-07	0.013	0.092	0.046	4.6E-02	19751

* rs62064540 (for association with MCH) is in proximity to previously reported variant rs689992 (for association with MCHC). rs113833421 is in proximity to previously reported variant rs11672923 (for association with RBC). But the LD between the current study’s index SNPs and previously reported variants are less than 0.01 in both cases.

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 up to 10 more cohorts with imputed data for an expanded discovery, to increase power for discover variants across all allele frequency spectrum. Only HGB was measured in ALSPAC WGS and ALSPAC imputed data and have a total of 12 cohorts for meta-analysis, while the other FBC traits included only 10 cohorts for meta-analysis. Variants with MAF <0.1% or imputation INFO <0.4 were not included. The genome-wide results for the expanded discovery was presented in **Figure 5.2**. A total of 3,952 variants passed the pre-defined threshold for genome-wide and suggestive significance (*P*<1E-07). Through the step-wise conditional analysis as described in chapter 2 and the methods section of this chapter, nine loci were found to be putative novel, and three known loci harboured novel variants (**Table 5.5**). The detailed results for each participating cohort are shown in **Table 5.6**. For the nine putative novel loci, three of them didn't have any other variants with *P*<1E-5 within 1Mb and there was a lack of supporting signals from SKAT-O test. These three are rare with MAF< 0.5%. Therefore, they are most likely to be false positive or would be difficult to be replicated. For the other six loci, further replication would be needed to confirm the association.

Given the availability of the genome-wide results for the six replication cohorts (for four traits: HGB, PCV, PLT, WBC), I run an 18-way meta-analysis that included the 12 discovery cohorts and six replication cohorts. Based on this 18-way meta-analysis, I identified a total of 12 associations that have *P*<5E-08 while their associations did not meet the significance threshold (*P*<1E-07) pre-defined for the 12-way analysis (**Table 5.7**). Although further independent replication is needed to confirm these associations, two signals have such strong associations that might not need further replication. The first one is the association of WBC in the *HLA* locus. A recent trans-ethnic GWAS meta-analysis on WBC reported an association within this region (Keller et al. 2014), but the reported lead SNP (rs2853946, chr6:31 247 203, EUR MAF=0.348) is in low LD with the lead SNP of this study (rs113164910, chr6:32427005, LD r2=0.08). The other strong signal from the 18-way is the (growth factor independent 1B transcription repressor) locus for association with PLT. The lead SNP (rs150813342) is a rare (18-way MAF=0.007) synonymous SNP within *GFI1B*, which encodes a zinc-finger containing transcriptional regulator that is primarily expressed in cells of hematopoietic lineage. The encoded protein complexes with numerous other transcriptional regulatory proteins to control expression of genes involved in the development and maturation of erythrocytes and megakaryocytes. Mutations in this gene are the cause of the autosomal dominant platelet disorder, platelet-type bleeding disorder-17 (Monteferrario et al. 2014).

For the three putative novel variants that are less than 1Mb away from previously reported variants for association with FBC traits, the association details of known variants and their LD with the putative novel variants in LD are listed in **Table 5.8**. The first one is the *CCND3* locus on chromosome 6, where three common variants were reported for association with MCV. All significantly associated SNPs within the *CCND3* locus are tagged by previously reports variants, except for rs112233623 and another SNP in high LD (rs113267280, chr6:41952511, LD r2=0.74) (**Figure 5.3)**. The second known locus with novel variants is on chromosome 11. Upon further examination of individual cohort results, I found that this association in the meta-analysis was mainly driven by one isolated population, HELIC-Pomak. The lead SNP rs11821302 has an EAF of 0.001 in TwinsUK but an EAF of 0.05 in HELIC-Pomak. For the HELIC-Pomak cohort, the lead SNP in this region is rs7116019 (chr11:4618606) (Zeggini 2014), but it is not significant (*P*>0.05) in TwinsUK or any other cohorts included in the 10-way meta-analysis. In this locus, there is a variant associated with protective immunity against severe malaria (rs11036238), which might offer some clue on the genetic isolate's response to malaria infection (Jallow et al. 2009). The third locus with novel variants is within *NPRL3* (nitrogen permease regulator-like 3) on chromosome 16. The two known variants are much more common and they are within a different gene *ITFG3* (integrin alpha FG-GAP repeat containing 3). Within 1Mb region, there is no other SNP in high LD with the index SNP rs117747069 (**Figure 5.3**). However, based on the Regulome database (http://regulome.stanford.edu), the functional evidence for rs117747069 is much stronger than the two known variants in this region (rs7189020 and rs1122794). The Regulome score is "2b" (supporting data from TFBS, motif, DNase footprint, and DNase peak) for rs117747069 and "5" (supporting evidence from TFBS or DNase peak) for rs7189020, while there is no functional data available for rs1122794. The regional plots for the two strongest associations based on 18-way meta-analysis were shown in **Figure 5.4.**

Figure 5.2 Results for 12-way meta-analysis

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.

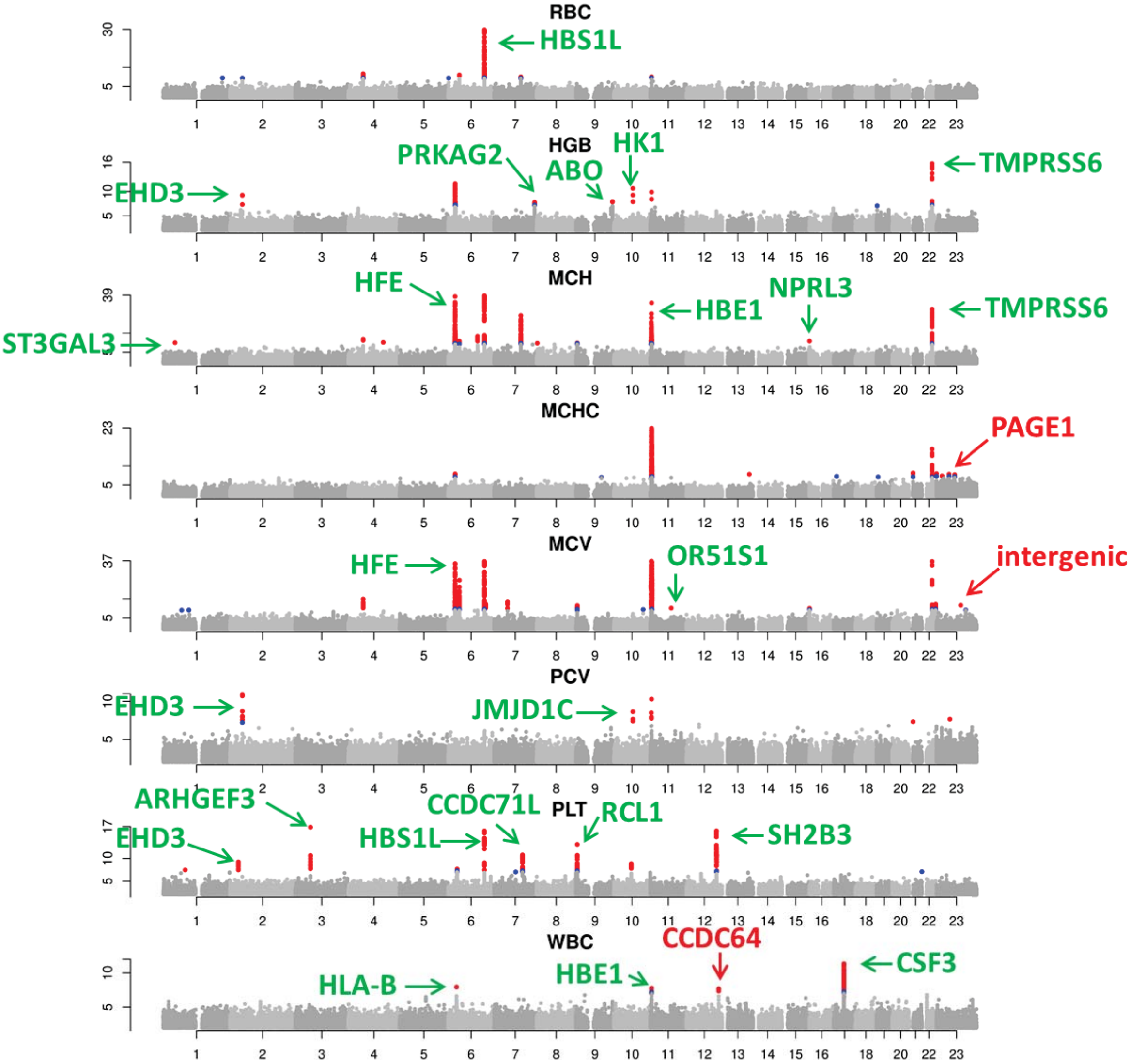


Table 5.5 Novel FBC variants based on expanded discovery (12-way meta-analysis)

The top part listed the index SNP for 9 putative novel loci. The bottom part listed three variants that have positive controls within 1Mb.

For the index SNVs in the nine novel loci, three are lonely variants and have no supporting SKAT signal, as labelled with * in the table.

							12-way meta-analysis					**Replication**				
Type	**Trait**	**rsID**	**CHR**	**POS**	**Gene**	**EA**	**EAF**	**Beta**	**SE**	**P**	**N**	**EAF**	**Beta**	**SE**	**P**	**N**
Putative Novel Loci	MCV	chr1:69249341*	1	69,249,341	intergenic	C	0.001	-2.024	0.376	9.87E-08	8,321	--	--	--	--	--
	MCV	rs189931100 *	1	96,028,784	intergenic	G	0.001	-1.989	0.369	9.18E-08	9,827	--	--	--	--	--
	RBC	chr6:1906294	6	1,906,294	GMDS	T	0.002	0.786	0.145	7.06E-08	13,944	--	--	--	--	--
	MCV	rs189443777	10	109,452,247	intergenic	A	0.008	-0.410	0.075	6.51E-08	14,804	--	--	--	--	--
	WBC	rs74853946	12	120,501,797	CCDC64	T	0.018	-0.181	0.032	2.14E-08	15,342	0.021	0.035	0.036	0.426	20,062
	MCHC	rs144022851	21	14,589,985	intergenic	T	0.090	0.196	0.033	7.16E-09	12,893	--	--	--	--	--
	PLT	rs200989541 *	21	47,565,506	FTCD	A	0.004	0.821	0.152	7.85E-08	8,703	0.001	-0.065	0.358	0.872	9,418
	MCHC	rs143473229	X	49,514,596	PAGE1	G	0.016	-0.255	0.045	1.47E-08	10,858	--	--	--	--	--
	MCV	rs73221860	X	111,785,547	--	G	0.207	0.075	0.014	3.99E-08	14,173	--	--	--	--	--
Putative Novel variants	MCV	rs112233623	6	41,924,998	*CCND3*	T	0.011	0.384	0.062	9.15E-10	15,277	--	--	--	--	--
	MCV	rs11821302	11	4,868,158	*OR51S1*	T	0.009	-1.161	0.094	1.38E-34	6,893	--	--	--	--	--
	MCH	rs117747069	16	170,076	*NPRL3*	C	0.032	-0.280	0.049	1.33E-08	12,189	--	--	--	--	--

Table 5.6 Cohort specific results of top hits from expanded discovery analysis

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.

	MCV, chr1:69249341						MCV, chr1:96028784						RBC, chr6:1906294						MCV, chr6:41924998					
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**
TwinsUK WGS	0.001	-1.976	0.710	5.4E-03	1548	0.99	-	-	-	-	-	-	0.003	0.061	0.315	8.5E-01	1561	1.00	0.011	0.249	0.181	1.7E-01	1548	0.92
CARL	0.000	18.54	44.05	6.8E-01	474	0.05	0.001	12.780	17.89	4.8E-01	474	0.16	0.001	2.803	2.776	3.2E-01	480	0.35	0.003	2.029	5.027	6.9E-01	474	0.36
CBR	-	-	-	-	-	-	0.002	-1.777	0.635	5.2E-03	1033	0.51	0.003	0.670	0.439	1.3E-01	1033	0.76	0.013	0.415	0.211	4.9E-02	1033	0.88
FVG	0.001	-1.515	12.827	9.0E-01	1374	0.06	0.001	-1.475	4.795	7.6E-01	1374	0.43	0.001	0.746	0.369	4.3E-02	1396	0.50	0.006	1.091	1.047	3.0E-01	1374	0.91
HA	0.002	-0.959	1.619	5.6E-01	979	0.15	0.002	-0.321	1.506	8.3E-01	979	0.14	0.002	0.503	0.600	4.0E-01	989	0.62	0.007	0.769	0.293	8.9E-03	979	0.88
HP	0.001	-3.749	2.041	6.7E-02	954	0.15	0.001	-3.115	2.501	2.2E-01	954	0.10	0.001	2.845	7.515	7.0E-01	968	0.02	0.028	0.483	0.157	2.2E-03	954	0.90
LURIC-1	0.001	-3.462	0.745	3.7E-06	1428	0.55	0.001	-1.368	0.851	1.1E-01	1428	0.36	-	-	-	-	-	-	0.011	0.257	0.195	1.9E-01	1428	0.84
LURIC-2	-	-	-	-	-	-	-	-	-	-	-	-	0.003	0.945	0.367	1.0E-02	1633	0.80	0.014	0.201	0.162	2.2E-01	1633	0.87
TwinsUKall	0.001	-1.464	0.465	1.7E-03	3586	0.75	0.001	-2.666	0.584	7.8E-06	3586	0.41	0.002	0.550	0.307	7.4E-02	3609	0.75	0.012	0.308	0.121	1.1E-02	3586	0.90
TwinsUK	0.000	28.194	22.13	2.0E-01	1058	0.68	0.001	-3.186	0.988	1.4E-03	1058	0.38	0.001	1.601	0.832	5.6E-02	1062	0.71	0.010	0.364	0.259	1.6E-01	1058	0.88
UKBS	0.001	0.239	0.557	6.7E-01	2065	0.84	-	-	-	-	-	-	0.003	0.709	0.325	2.9E-02	2067	0.75	0.010	0.354	0.169	3.6E-02	2065	0.89
VB	0.000	-11.98	9.322	2.0E-01	1755	0.10	0.000	-11.145	6.128	6.9E-02	1755	0.03	0.003	1.479	0.453	1.2E-03	1770	0.62	0.007	0.769	0.243	1.6E-03	1755	0.79

	MCV, chr10:109452247						MCV, chr11:4868158						WBC, chr12:120501797						MCH, chr16:170076					
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**
TwinsUK WGS	0.006	-1.135	0.233	1.2E-06	1548	0.97	-	-	-	-	-	-	0.018	-0.061	0.137	6.6E-01	1551	0.98	0.037	-0.326	0.096	6.8E-04	1549	0.96
CARL	0.000	11.043	44.917	8.1E-01	474	0.06	0.000	2.549	219.43	9.8E-01	474	0.01	0.010	-0.355	0.093	2.6E-04	484	0.73	0.028	27.678	35.395	4.3E-01	473	0.58
CBR	0.007	-0.432	0.271	1.1E-01	1033	0.92	-	-	-	-	-	-	0.023	-0.216	0.147	1.4E-01	1033	0.96	0.039	-0.096	0.140	4.9E-01	1033	0.64
FVG	0.010	-0.250	0.813	7.6E-01	1374	0.91	0.001	-2.683	2.809	3.4E-01	1374	0.87	0.010	-0.159	0.047	8.4E-04	1387	0.83	0.029	-17.80	13.921	2.0E-01	1357	0.54
HA	0.006	-0.290	0.362	4.2E-01	979	0.69	0.005	-0.637	0.344	6.6E-02	979	1.00	0.014	-0.318	0.245	1.9E-01	990	0.69	0.022	0.190	0.261	4.7E-01	981	0.36
HP	0.001	0.089	0.706	9.0E-01	954	0.84	0.052	-1.185	0.100	1.3E-26	954	1.00	0.001	0.567	1.111	6.1E-01	963	0.75	0.030	0.026	0.182	8.8E-01	949	0.58
LURIC-1	0.008	-0.147	0.222	5.1E-01	1428	0.86	0.001	0.398	0.681	5.6E-01	1428	0.84	0.024	-0.205	0.131	1.2E-01	1428	0.86	-	-	-	-	-	-
LURIC-2	0.010	-0.246	0.192	2.0E-01	1633	0.87	-	-	-	-	-	-	0.023	-0.216	0.127	9.0E-02	1633	0.85	-	-	-	-	-	-
TwinsUKall	0.007	-0.498	0.158	1.7E-03	3586	0.92	0.001	-1.618	0.510	1.6E-03	3586	0.95	0.018	-0.091	0.096	3.4E-01	3597	0.92	0.038	-0.348	0.073	2.4E-06	3587	0.71
TwinsUK	0.008	-0.021	0.313	9.5E-01	1058	0.92	0.001	-1.932	0.698	6.0E-03	1058	0.95	0.019	-0.086	0.179	6.3E-01	1065	0.91	0.035	-0.298	0.141	3.5E-02	1061	0.70
UKBS	0.008	-0.518	0.177	3.4E-03	2065	0.91	-	-	-	-	-	-	0.023	-0.108	0.105	3.0E-01	2053	0.92	0.033	-0.194	0.109	7.5E-02	2061	0.63
VB	0.012	-0.554	0.174	1.5E-03	1755	0.96	0.000	-33.81	163.74	8.4E-01	1755	0.00	0.015	-0.167	0.164	3.1E-01	1774	0.79	0.026	-0.673	0.141	2.1E-06	1749	0.61

Table 5.6 Cohort specific results of top hits from expanded discovery analysis (*continued*)

	MCHC, chr21:47565506						PLT, chr21:14589985						MCHC, chrX:49514596						MCV, chrX:111785547					
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**
TwinsUK WGS	-	-	-	-	-	-	-	-	-	-	-	-	0.005	-0.056	0.339	8.7E-01	942	1.00	0.181	0.023	0.047	6.3E-01	1548	1.00
CARL	-	-	-	-	-	-	0.014	0.148	0.446	7.5E-01	483	0.89	0.005	1.437	58.397	9.8E-01	483	0.00	0.171	12.204	6.031	4.5E-02	474	0.01
CBR	-	-	-	-	-	-	0.027	0.160	0.210	4.5E-01	1033	0.42	-	-	-	-	-	-	-	-	-	-	-	-
FVG	0.001	0.834	1.233	5.0E-01	1375	0.15	0.668	0.228	0.040	2.5E-08	1391	0.63	0.099	-0.236	0.048	1.0E-06	1391	0.92	0.281	0.088	0.150	5.5E-01	1374	0.80
HA	0.001	1.182	1.792	5.1E-01	991	0.20	0.043	0.241	0.171	1.6E-01	994	0.41	0.001	3.485	2.416	1.5E-01	994	0.27	0.213	0.121	0.048	1.3E-02	979	0.99
HP	0.024	0.852	0.163	3.4E-07	968	0.91	0.036	-0.060	0.232	7.9E-01	963	0.28	0.000	-9.721	9.655	3.2E-01	963	0.01	0.297	0.168	0.046	3.1E-04	954	0.98
LURIC-1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.198	0.092	0.037	1.4E-02	1392	0.99
LURIC-2	-	-	-	-	-	-	0.015	-0.059	0.158	7.1E-01	1633	0.78	0.005	-0.204	0.242	4.0E-01	1594	0.43	0.202	0.048	0.033	1.5E-01	1594	0.99
TwinsUKall	0.001	0.582	0.453	2.0E-01	3602	0.51	0.011	0.472	0.142	9.6E-04	2565	0.82	0.004	-0.382	0.275	1.6E-01	2565	0.70	0.191	-0.007	0.032	8.3E-01	3586	1.00
TwinsUK	0.001	-0.029	0.794	9.7E-01	1070	0.46	0.011	0.593	0.238	1.3E-02	947	0.81	0.005	-0.653	0.382	8.8E-02	947	0.63	0.213	-0.069	0.055	2.1E-01	1058	1.00
UKBS	-	-	-	-	-	-	0.019	0.096	0.150	5.2E-01	2059	0.58	0.005	-0.494	0.198	1.3E-02	2059	0.74	0.183	0.111	0.033	7.7E-04	2065	0.99
VB	0.001	0.354	1.989	8.6E-01	1767	0.11	0.018	-0.072	0.136	6.0E-01	1772	0.89	0.003	-0.643	0.405	1.1E-01	1772	0.37	0.177	0.069	0.038	7.1E-02	1755	1.00
WHI-garnet	0.001	0.552	0.680	4.2E-01	3802	0.26	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
WHI_gecco1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
WHI_gecco2	0.001	0.734	1.008	4.7E-01	1733	0.44	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
WHI_hipfx	0.001	0.210	0.666	7.5E-01	3807	0.33	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
WHI_mopmap	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
WHI_whims	0.001	-0.303	0.422	4.7E-01	5617	0.36	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 5.7 Top hits from a further expanded discovery (18-way meta-analysis)

For each set of results, the effect allele frequency (EAF), beta, standard deviation (SE), *P* value, and the total sample size were presented. For positive controls within 1Mb, only the one in highest LD is shown when there are multiple ones. The information includes trait name, rsID, CHR:POS, and LD measured in r2.

							WGS			
Trait	**CHRPOS**	**rsID**	**Positive Controls within 1Mb**	gene	**EA**	**NEA**	**EAF**	**beta**	**SE**	**P**
PCV	chr1:9,077,128	rs769904	--	SLC2A7	C	T	0.002	-0.223	0.409	5.9E-01
PLT	chr1:24,743,879	rs760968	PLT, rs592372, chr1:25636197, NA	C1orf201	T	C	0.246	0.008	0.042	8.6E-01
HGB	chr2:159,916,661	rs113682276	--	TANC1	A	G	0.009	0.087	0.141	5.4E-01
PLT	chr3:56,929,498	rs200858303	PLT, rs1354034, chr3:56849749, 0.06	ARHGEF3	T	TTA	--	--	--	--
WBC	chr6:32,427,005	rs113164910	HGB, rs9272219, chr6:32602269, 0.036	HLA-DRB9	A	AAC	0.327	-0.081	0.038	3.2E-02
PLT	chr9:91,459,039	rs141068793	PLT, rs11142062, chr9:90658749, NA	--	C	T	0.062	-0.113	0.078	1.5E-01
PLT	chr9:135,864,513	rs150813342	HGB, rs4128808, chr9:136065229, 0.011	GFI1B	T	C	0.004	-0.229	0.291	4.3E-01
WBC	chr17:7,231,792	rs9905997	--	NEURL4	G	A	0.44	0.095	0.035	7.5E-03
PLT	chr17:64,195,431	rs75003668	--	PSMD7P1	G	A	0.033	0.221	0.115	5.4E-02
HGB	chr20:22,110,210	rs138233587	--	--	A	AT	0.046	-0.078	0.06	1.9E-01
PLT	chr21:36,474,114	rs2834764	--	RUNX1	A	G	0.415	-0.036	0.036	3.2E-01
PLT	chr22:50,570,755	rs75570992	RBC, rs140522, chr22:50971266, 0.00	MOV10L1	C	G	0.072	0.113	0.069	1.0E-01

		12-way					6-way					18-way				
Trait	**CHRPOS**	**EAF**	**beta**	**SE**	**P**	**N**	**EAF**	**beta**	**SE**	**P**	**N**	**EAF**	**beta**	**SE**	**P**	**N**
PCV	chr1:9077128	0.002	-0.415	0.155	8.0E-03	15,351	0.005	-0.434	0.078	2.2E-06	20063	0.003	-0.430	0.070	1.1E-08	35,414
PLT	chr1:24743879	0.229	0.045	0.014	1.7E-03	15,327	0.239	0.064	0.012	1.7E-06	19948	0.234	0.056	0.009	5.2E-09	35,275
HGB	chr2:159916661	0.006	0.269	0.075	3.6E-04	19,749	0.007	0.342	0.072	6.6E-05	20034	0.007	0.307	0.052	4.1E-08	39,783
PLT	chr3:56929498	0.420	0.047	0.012	1.4E-04	15,326	0.434	0.062	0.010	1.2E-07	19948	0.428	0.056	0.008	2.7E-11	35,274
WBC	chr6:32427005	0.289	-0.035	0.009	8.7E-05	15,342	0.325	-0.096	0.011	5.8E-14	20062	0.309	-0.059	0.007	2.4E-16	35,404
PLT	chr9:91459039	0.078	-0.093	0.022	3.4E-05	15,326	0.067	-0.086	0.020	1.9E-04	19948	0.072	-0.089	0.015	2.2E-08	35,274
PLT	chr9:135864513	0.007	-0.398	0.080	8.8E-07	15,326	0.008	-0.485	0.061	3.9E-12	19950	0.007	-0.453	0.049	2.4E-18	35,276
WBC	chr17:7231792	0.454	0.032	0.007	4.9E-06	15,342	0.453	0.048	0.010	9.3E-05	20064	0.453	0.037	0.006	1.5E-09	35,406
PLT	chr17:64195431	0.029	0.157	0.041	1.4E-04	15,327	0.028	0.164	0.036	5.3E-05	19948	0.028	0.161	0.027	1.8E-08	35,275
HGB	chr20:22110210	0.056	-0.069	0.023	2.6E-03	19,750	0.053	-0.125	0.023	4.0E-06	20035	0.054	-0.097	0.016	2.3E-08	39,785
PLT	chr21:36474114	0.415	-0.046	0.012	1.2E-04	15,327	0.421	-0.045	0.010	9.7E-05	19949	0.419	-0.046	0.008	3.0E-08	35,276
PLT	chr22:50570755	0.059	0.120	0.027	8.6E-06	14,844	0.061	0.116	0.023	6.3E-06	19946	0.060	0.118	0.017	1.4E-10	34,790

Table 5.8 LD of three putative novel variants in known locus

For each locus, the 10-way association statistics and the LD for all known variants within 1Mb of the putative novel variants are listed.

Novel variants	Known variants	Associated traits	CHR:POS	MAF	10-way P	LD (r2)
rs112233623 (chr6:41924998)	rs3218097	MCV	chr6:41905275	0.247	6.72E-11	0.027
	rs9349205	MCV	chr6:41925159	0.233	2.01E-11	0.028
	rs11970772	MCV	chr6:41925290	0.214	7.91E-14	0.002
rs11821302 (chr11:4868158)	rs7116019	MCV	chr11:4618606	0.012	3.11E-31	0
	rs11036238	Malaria	chr11:5225635	0.272	--	0
	rs2071348	Beta thalassemia/hemoglubin E	chr11:5264146	0.340	--	0
	rs4910742	Fetal hemoglobin levels	chr11:5306509	0.051	--	0
rs117747069 (chr16:170076)	rs7189020	MCV	chr16:304803	0.376	1.29E-04	0.015
	rs1122794	MCH	chr16:309155	0.181	2.12E-05	0.006

Figure 5.3 Regional plots of two known loci with putative novel variants

The top plot is for the *CCND3* locus for association with MCV. The bottom plot is for *NPRL3* locus for association with MCH. The *P* values are based on the 10-way meta-analysis. The novel variant is shown in red text, while the SNPs tagged by previously reported variants are known in other colors.

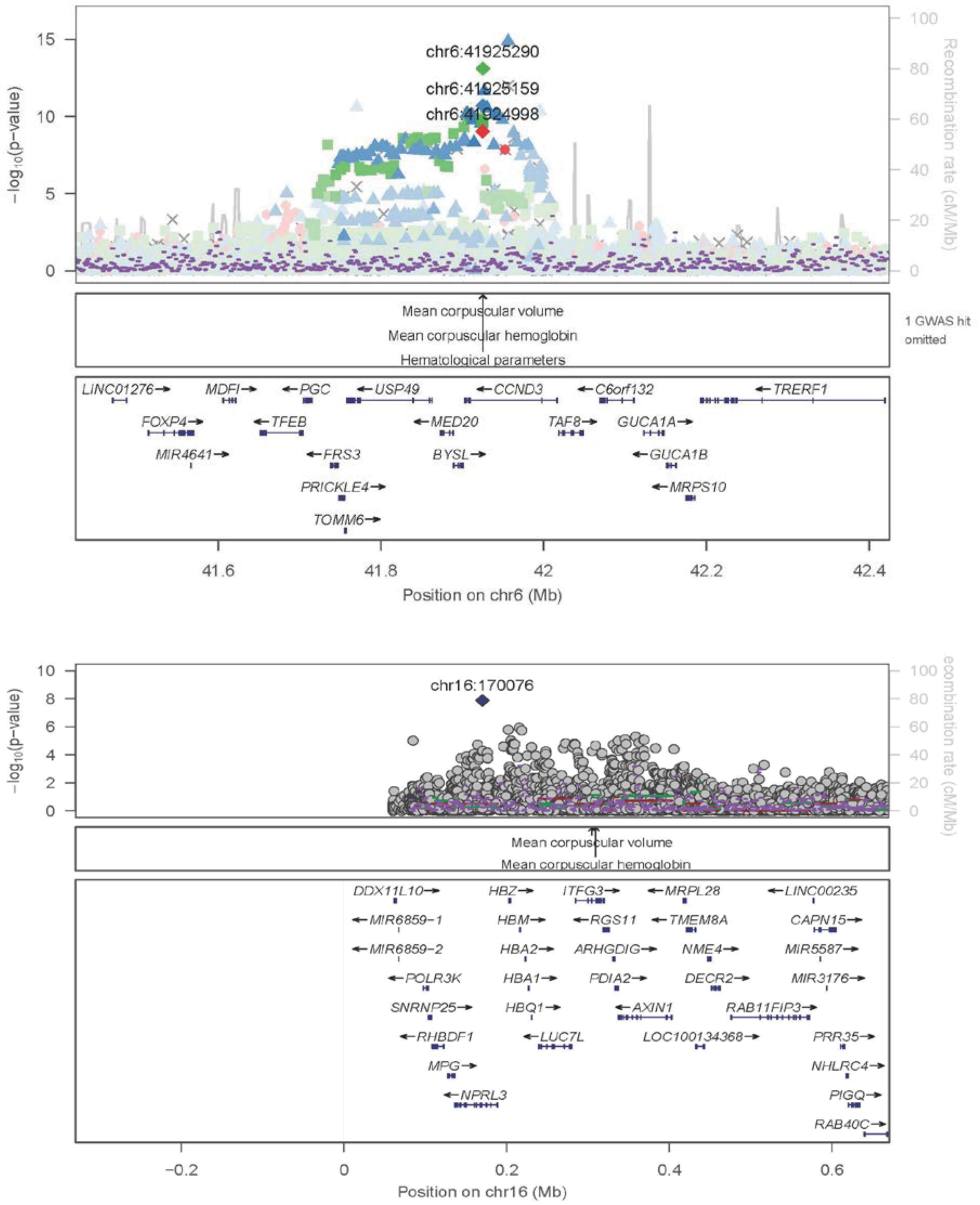
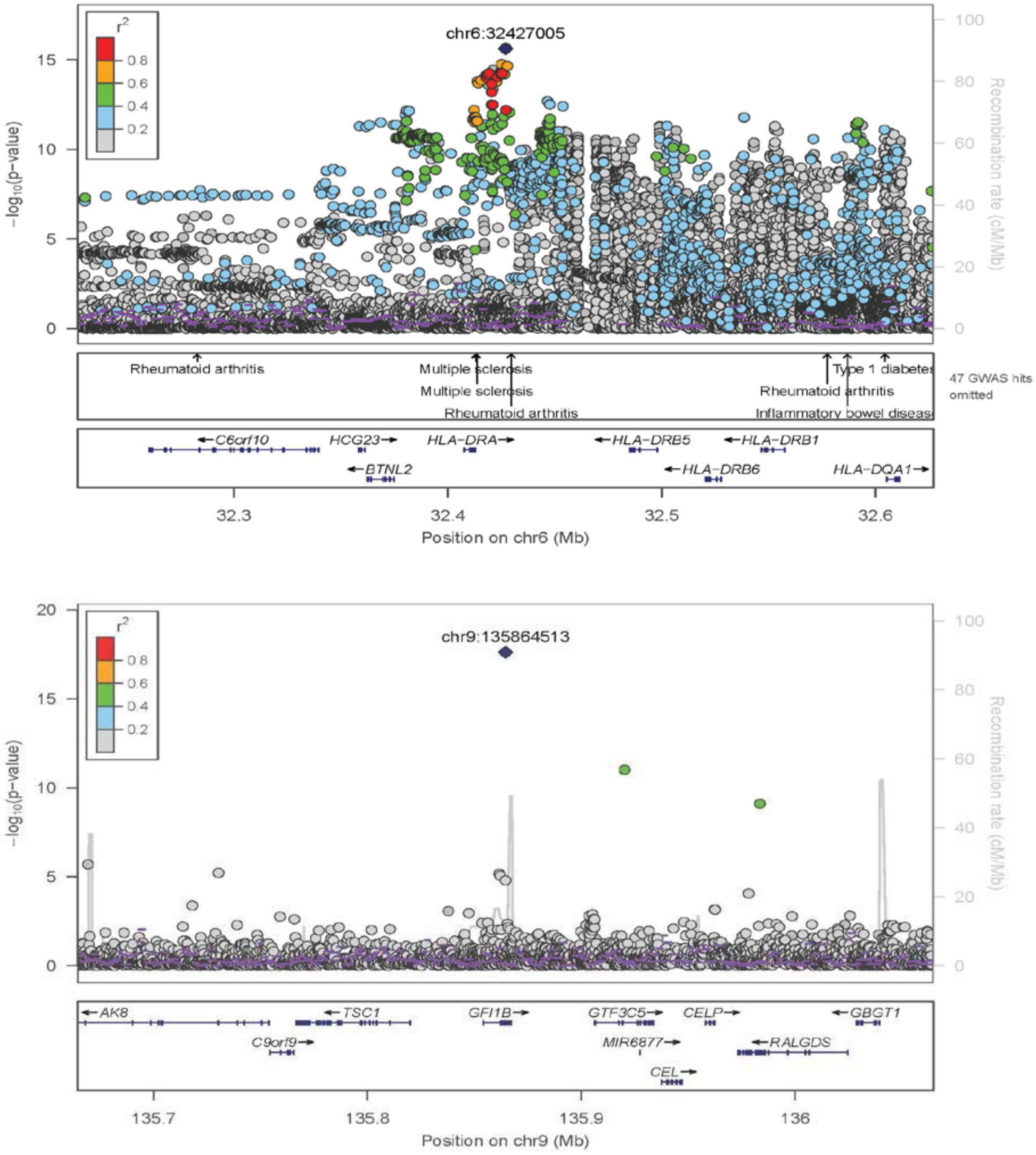


Figure 5.4 Regional plots of top hits from 18-way meta-analysis

The top plot is for the HLA locus for association with WBC, and the bottom plot is for GFI1B locus for association with PLT.



5.3.2 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 FBC 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 seven known loci, there are sufficient resolution to limit the number of possible causal variants to a small informative set (log10BF>5 and # of variants <20) (**Table 5.9**). There are a total of 22 putative causal variants in these seven loci, three of which are previously reported known variants. Based on Regulome database, rs115740542 has the strongest evidence for functionality, with a score of “1a” (supporting evidence from TF binding, matched TF motif, matched DNase footprint, DNase peak), while rs198851 and rs12005199 have modest evidence for functionality (supporting evidence from TF binding, any motif, DNase footprint, DNase peak). The rest variants all have a score greater than 4, indicating weak support of functionality.

Table 5.9 Putative causal variants based on fine mapping

BP: Bayes factor, PP: posterior probability. Three previously reported known variant are labelled with *.

		Fine-mapping					WGS					Meta-analysis					
trait	region	rsID	CHRPOS	GWAVA	log10(BF)	PP	EA	EAF	BETA	SE	P	EA	EAF	beta	se	P	N
PLT	ARHGEF	rs1354034	chr3:56849749 *	Intron	9.89	1.00	--	--	--	--	--	T	0.421	-0.104	0.012	1.38E-17	15328
MCH	HFE, chr6:25343245-26589359	rs80215559	chr6:25918225	Intron	19.94	0.03	C	0.069	0.352	0.072	9.66E-07	T	0.941	-0.319	0.032	4.02E-23	12190
		rs1800562	chr6:26093141	Missense	20.25	0.07	A	0.070	0.341	0.070	1.32E-06	G	0.939	-0.316	0.031	3.30E-24	12190
		rs79220007	chr6:26098474	3_prime_UTR	20.48	0.12	C	0.069	0.338	0.070	1.71E-06	T	0.940	-0.318	0.031	3.31E-24	12189
		rs115740542	chr6:26123502	Upstream	21.28	0.74	C	0.067	0.339	0.072	2.49E-06	T	0.941	-0.323	0.032	3.00E-24	12190
MCV	HFE chr6:25600233-26589359	rs1799945	chr6:26091179	Missense	12.51	0.39	G	0.140	0.173	0.053	1.09E-03	C	0.846	-0.162	0.018	5.88E-20	15280
		rs2032451	chr6:26092170	Upstream	11.38	0.03	T	0.142	0.175	0.053	8.67E-04	G	0.845	-0.156	0.018	1.04E-18	15279
		rs1800562	chr6:26093141 *	Missense	11.30	0.02	A	0.070	0.263	0.070	1.98E-04	G	0.942	-0.238	0.028	1.44E-17	15281
		rs79220007	chr6:26098474	3_prime_UTR	11.35	0.03	C	0.069	0.259	0.071	2.57E-04	T	0.943	-0.239	0.028	1.38E-17	15278
		rs198851	chr6:26104632	Downstream	12.50	0.38	G	0.859	-0.171	0.053	1.19E-03	T	0.153	0.163	0.018	7.00E-20	15281
		rs198846	chr6:26107463	Downstream	11.59	0.05	G	0.853	-0.167	0.052	1.27E-03	A	0.158	0.156	0.017	5.71E-19	15279
		rs198833	chr6:26114508	Downstream	11.46	0.04	A	0.854	-0.165	0.052	1.44E-03	G	0.158	0.155	0.017	1.03E-18	15280
		rs115740542	chr6:26123502	Upstream	11.41	0.03	C	0.067	0.258	0.072	3.61E-04	T	0.945	-0.240	0.028	6.20E-17	15281
PLT	intergenic chr9:4740135-4903034	rs385893	chr9:4763176 *	Regulatory	6.11	0.03	C	0.511	0.091	0.036	1.23E-02	T	0.493	-0.081	0.012	2.02E-11	15328
		rs12005199	chr9:4763491	Regulatory	7.68	0.94	A	0.291	0.134	0.039	5.81E-04	G	0.727	-0.105	0.014	8.40E-14	15328
MCH	OR52A1, chr11:4810830-5765688	chr11:5042074	chr11:5042074	Downstream	18.06	0.31	A	0.001	-0.146	0.710	8.37E-01	A	0.003	-1.846	0.200	5.04E-20	4568
		rs181392259	chr11:5054906	Upstream	17.20	0.04	T	0.004	-0.017	0.315	9.57E-01	T	0.003	-0.894	0.140	1.93E-10	11716
		chr11:5126515	chr11:5126515	--	17.97	0.25	--	--	--	--	--	T	0.997	1.675	0.189	1.15E-18	8623
		chr11:5180087	chr11:5180087	Intron	18.08	0.33	T	0.001	-0.138	0.710	8.46E-01	T	0.003	-1.868	0.203	5.20E-20	4568
		rs183952362	chr11:5196364	Upstream	17.13	0.04	G	0.004	-0.367	0.319	2.51E-01	G	0.004	-0.636	0.122	2.25E-07	11717
MCH	RAB11FIP3, chr16:442805-602595	rs143109032	chr16:536959	Upstream	7.31	1.00	T	0.005	0.394	0.279	1.58E-01	C	0.994	0.221	0.103	3.26E-02	12189
MCH	TMPRSS6 chr22:37366826-37510072	rs855791	chr22:37462936 *	Missense	26.38	0.98	G	0.555	0.182	0.036	4.02E-07	A	0.444	-0.161	0.014	7.62E-29	12190

5.3.3 Novel loci based on rare variants aggregation test

The above are for single marker base tests, which has limited power to detect associations for low frequency and rare variants given the current number of samples with WGS. Here I show association results based on rare variants aggregation tests. As stated in the Methods, there types of SKAT-O analyses were run: genome-wide sliding window, exome-wide gene based, and exome-wide with only functional variants. Overall, the statistics of these tests follow the expected distribution assuming a NULL association, and there is a lack of signals meeting pre-defined genome-wide significance threshold (**Figure 5.5**). Nevertheless, there are six regions that meet our pre-defined significance threshold for follow-up (*P*<6.8E-08 for genome-wide SKAT-O, *P*<1.2E-06 for exome-wide SKAT-O, *P*<1.0E-05 for functional variants SKAT-O) (**Table 5.10)**. For three of these loci, the SKAT *P* value is much less significant than the SKAT-O *P* value, indicating that the signals are mainly driven by burden tests. Although independent replication is needed to confirm the rare variants aggregation based association with these six regions, the *RHBDL2* locus for association with PLT is a biologically plausible. It was reported that *RHBDL2* and thrombomodulin have important roles in wound healing via the release of soluble *RHBDL2* from keratinocytes and that may function as an autocrine/paracrine signal promoting wound healing (Cheng et al. 2011). The most strongly associated variant based on WGS data alone in this locus is chr1:39384826 (MAF=0.008, *P*=2.21E-05) (**Figure 5.6**). However, this variant is not significant in the 10-way meta-analysis (*P*<0.05). This locus also harbours a variant (rs4246511, chr1:39380385) previously reported for associated with menopause age at onset (Stolk et al. 2012). However, the rare variants based association for this region needs to be validated and replicated to drive further interpretation on this locus.

Figure 5.5 Rare variants aggregation test results for FBC traits

There are eight rows, each row for one of the eight traits as indicated in the plot title. 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.

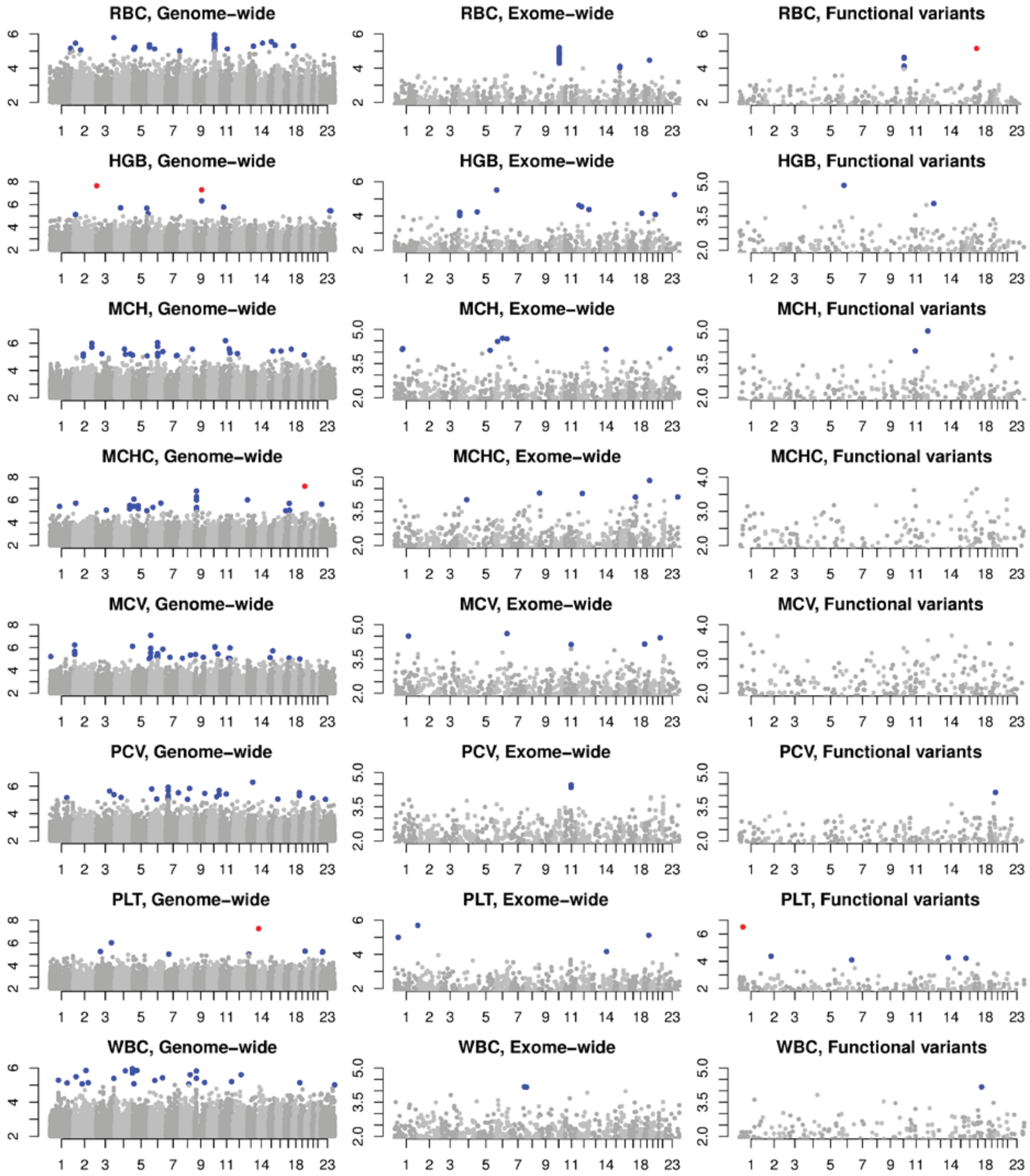


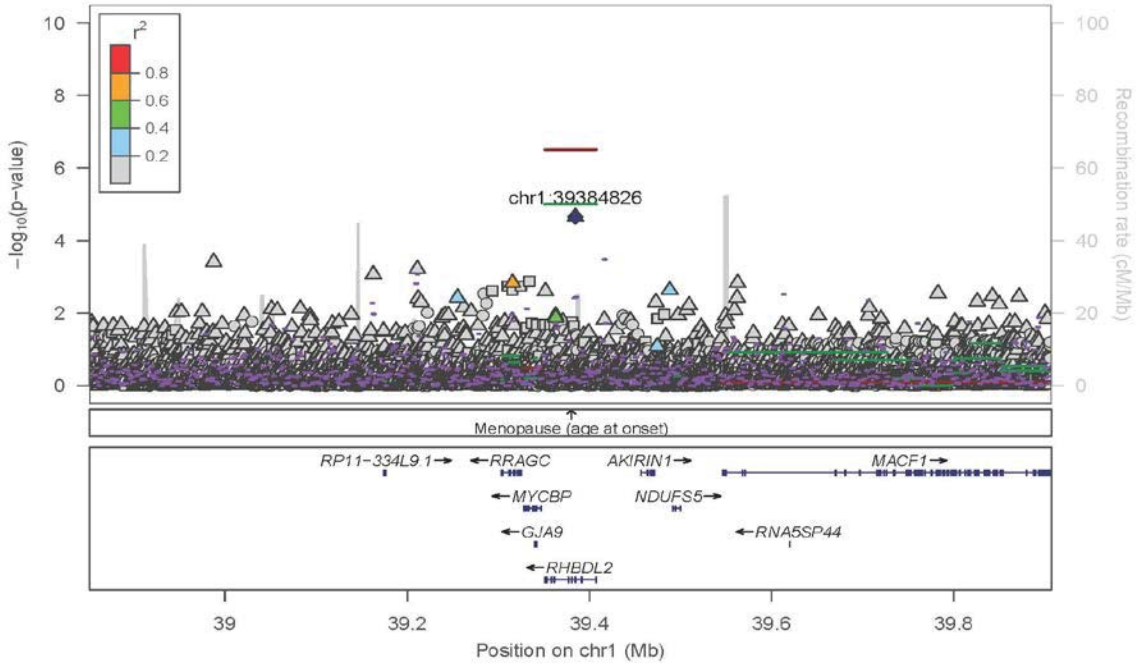
Table 5.10 Rare variants aggregation tests based top hits for FBC traits

For the three locus marked with *, the SKAT *P* is much less significant than the SKAT-O *P*, indicating that the signals are mainly driven by burden tests.

trait	**Type**	**locus**	**chr**	**start**	**End**	**TwinsUK**	**ALSPAC**	**SKAT**	**TwinsUK**	**ALSPAC**	**SKAT-O**
PLT	Functional variants	RHBDL2	1	39,351,479	39,407,471	7.52E-06	--	7.52E-06	3.11E-07	--	3.11E-07
HGB *	Genome-wide	GRM7	3	6,463,501	6,466,500	1.02E-01	3.09E-03	5.40E-04	1.75E-01	6.35E-03	2.28E-08
HGB	Genome-wide	OSTF1	9	77,787,001	77,790,000	1.74E-05	3.73E-04	1.68E-08	4.49E-05	6.24E-04	5.03E-08
PLT *	Genome-wide	DHRS4	14	24,462,001	24,465,000	1.01E-04	--	1.01E-04	5.48E-08	--	5.48E-08
RBC	Functional variants	PIGS	17	26,880,401	26,898,890	8.09E-05	--	8.09E-05	6.99E-06	--	6.99E-06
MCHC *	Genome-wide	ZSCAN5A	19	56,883,001	56,886,000	1.01E-04	--	1.20E-05	6.26E-08	--	6.26E-08

Figure 5.6 Regional plots of *RHBDL2*

The single marker results are based on TwinsUK WGS. The horizontal dashed lines are SKAT-O, purple, green, red for genome-wide SKAT-O, exome-wide SKAT-O, and functional variants based SKAT-O respectively.



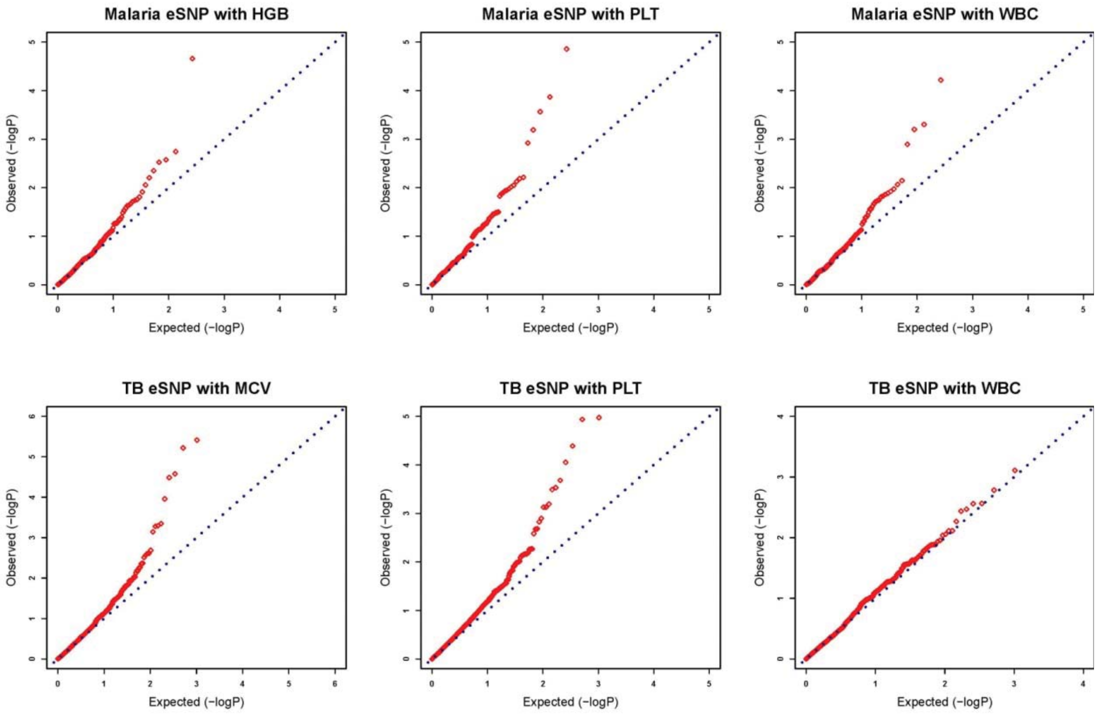
5.3.4 Host-response eQTL

Given the role of blood parameters in the host response to bacterial infection, I explored an approach to show whether genetic variants associated with host-response are enriched for association with FBC traits. I included 268 eSNP associated with gene expression of host response to malaria infection (Idaghdour et al. 2012) and 40 loci implied for response to severe malaria (Timmann et al. 2012). For tuberculosis, I used 1,046 eSNPs (720 for infected, 756 for unaffected) (Barreiro et al. 2012). Given the overall low number of variants tested, I did not perform a formal enrichment test, but used QQ plots to see whether the SNPs associated with host-response follow a NULL distribution for association with FBC traits. As shown in **Figure 5.7**, both HGB and PLT are enriched for eSNPs associated with host response to both Malaria and TB. WBC is enriched for eSNP associated with host response to Malaria but not to TB.

It is well established that genetic loci associated with resistance to malaria (for example, HBB, HBA1/HBA2, and G6PD) are associated with RBC traits (Ding et al. 2013). This is consistent with the fact that the malaria parasites grow in the human red cells. In 2012, a research team at Duke University discovered that human microRNA found in sickle red cells directly participate in the gene regulation of malaria parasites (LaMonte et al. 2012). The study showed that when two different microRNAs were introduced at higher levels in normal red cells, the parasite growth also was decreased. Another surprise in this investigation was the presence of a chimera, a fusion of human microRNA with the parasites' mRNAs, which represents a unique form of host-parasite interaction. This may reflect either a novel form of host-cell immunity or a mechanism by which the parasite is able to adapt to the host-cell environment. Although WBC changes during infections to TB, there was no reported evidence that the genetic loci associated with TB resistance is also associated with WBC. Similarly, platelet phagocytosis may contribute to thrombocytopenia found in vivax malaria (Coelho et al. 2013), but the preliminary data presented in **Figure 5.7** is the first to imply that genetics is involved between the phenotypic variation of FBC traits and the host response to infection of malaria and TB.

Figure 5.7 eSNPs associated with host response to TB and Malaria

Y-axis is the observed *P* value of eSNPs previously reported for association with Malaria (the first row) and TB (the second row), for association with HGB (first column), PLT (second column), WBC (third column). These *P* values are from the 12-way meta-analysis. The X-axis is the expected *P* value under the NULL hypothesis of no association.



5.4 Conclusion & Discussion

5.4.1 Summary of main findings

So far, there are no reported studies on FBC that used WGS data. With a modest WGS sample size (N=1,497), I identified three putative novel variants, but they were not replicated based on a few imputed datasets made available for replication. A total of 25 variants with MAF between 0.5% and 5% have *P* <1e-06 based on TwinsUK WGS, but replication is needed to establish any of these signals. Nevertheless, the association of rs115740542 within *NPRL3* with MCH was already supported by epigenomic annotation. To boost study power, I included a total of 12 cohorts for discovery and 6 cohorts for replication. I further conducted a meta-analysis with all 18 cohorts with a sample size up to 41,557. Based on the 12-way meta-analysis, a total of nine novel loci and three novel variants within known loci were discovered at a pre-defined *P*<1E-07. However, replication data is only available for two of these variants with non-replicated results. Based on the 18-way meta-analysis, there are two strong associations: the *HLA* locus for association with WBC, and the *GFI1B* locus for association with PLT. Given the function of these two regions for the according phenotypes and given the strength of the association signals, there two associations are most likely to be true and deserve further investigation. Fine-mapping analysis identified one SNP rs115740542 within *HFE* to be highly likely causal, with supporting evidence of functional data (RegulomeDB). By running a systematic enrichment analysis, I observed that hematological traits associated SNVs are significantly enriched in key epigenomic features including chromatin state, histone modification, and TFBS. Through rare variant aggregation analysis, I discovered that the aggregated functional variants in *RHBDL2* are strongly associated with PLT, which is biologically plausible.

5.4.2 Interpretation of results

The single marker association testing of eight lipids follows closely the expected relationship between EAF and effect size (beta) as dictated by study power (Park et al. 2011), as shown in **Figure 5.8**. Given the relatively small sample size and yet the encouraging finding of two strong signals based on the 18-way analysis, more truly novel associations are expected to be found with larger sample sizes.

GWAS on FBC traits has already brought translational outcome. As we know, the β-globin gene (*HBB*) is silent prior to birth and the β-globin subunits are encoded by the γ-globin gene (*HBG1* and *HBG2*) to form fetal hemoglobin (HbF). The switch from HbF to HbA production is a transcriptionally and epigenetically tightly regulated process (Sankaran et al. 2010). The association of *BCL11A* with HbF levels were first reported through GWAS (Menzel et al. 2007, Uda et al. 2008). Later on, *BCL11A* was found to be a potent transcriptional repressor of γ-globin gene expression and that its inactivation in the erythroid lineage can treat sickle cell disease in mouse model through re-activation of HbF production (Sankaran et al. 2008, Xu et al. 2011). This model was confirmed by targeted deletion of the enhancer through genome engineering that blocked *BCL11A* expression and re-activated γ-globin gene expression and HbF production (Sankaran et al. 2012). As genome editing methods are rapidly improving, this proof-of-concept experiment suggests a new therapeutic strategy for β-thalassemia and sickle cell diseases with mutations in *HBB* (Bauer and Orkin 2011, Hardison and Blobel 2013).

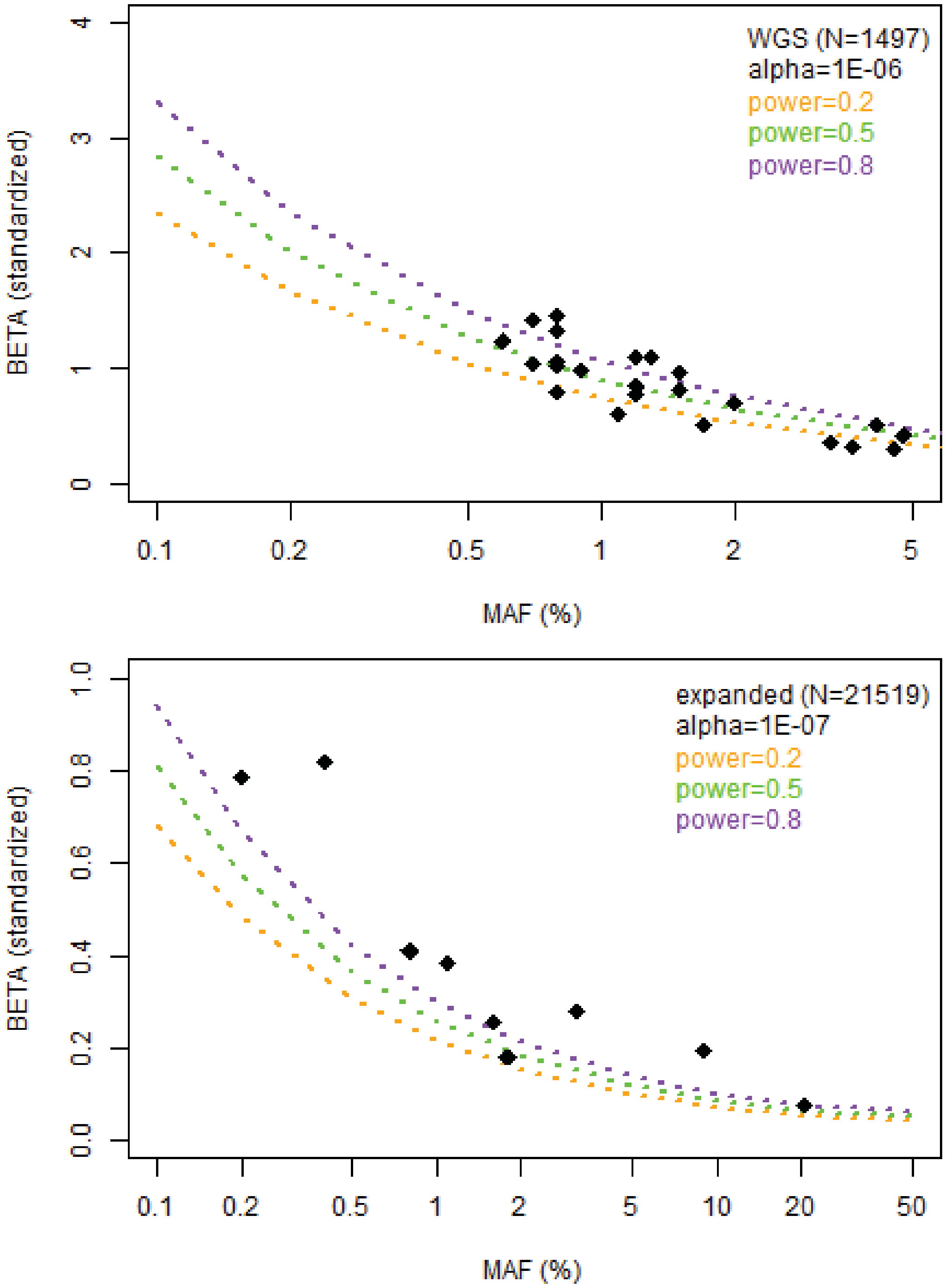
5.4.3 Future direction

Compared to many other complex traits, future larger studies on FBC traits with WGS dataset might be more achievable given these traits are widely measured in clinical settings for evaluation health and diseases. FBC traits are also preferred phenotypes for the study the genetics of complex human diseases because they could be easily manipulated in vitro and discovered genes could be assessed in cell cultures and model organisms. It is not surprising that there is an overall lack of novel loci discovered given the sample size in the current study compared to previously conducted GWAS on these traits. The lack of loci for WBC could be also due to phenotype heterogeneity because the major populations of white blood cells (lymphocytes, granulocytes, monocytes) differ markedly in their roles and lifespans.

Besides increasing the number of samples of European ancestry, including samples of diverse ethnicity could also boost the genetic findings for FBC traits. For many complex traits, African samples have been used to fine map genetic loci discovered from European samples, due to longer haplotypes in Africans. However, for FBC traits, sometimes a very strong genetic association in African population might not have any association in the European population. The associations of variants in *DARC* with WBC (Reich et al. 2009) and the association of variants in *HBA2* with RBC (Chen et al. 2013) are only observed in Africans while those variants are almost monomorphic in Europeans. The former variation protects against *Plasmodium vivax* while the latter protests against malaria infections, which are common in Africa. This study demonstrated similar phenomenon for genetic isolates. The signal on chromosome 11 is marginally significant in TwinsUK while strongly significant in HELIC-Pomak. Also, the signal on chromosome 21 (chr21:14589985 for association with PLT) mainly came from an Italian isolate: INGI-FVG. Its MAF in TwinsUK is ~1%, but ~4% in two Greek isolates, and ~33% in INGI-FVG. Once these are confirmed to be true signals in the general population, the use of genetic isolates would be proven valuable for identifying these associations, which would otherwise require a much larger sample size for detection of the association.

Figure 5.8 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 25 putative novel WGS variants are shown in the top power plot for WGS, and the nine putative novel variants from expanded discovery are shown in the bottom power plot for expanded discovery.


WGS (N=1497)
alpha=1E-06
power=0.2
power=0.5
power=0.8
BETA (standardized)
MAF (%)
expanded (N=21519)
alpha=1E-07
power=0.2
power=0.5
power=0.8
BETA (standardized)
MAF (%)