Chapter 4

# Dissecting the functional relationship between neutrophil count and surface expression of cellular receptors

## Collaboration Note

The recall organisation of individuals and collection of blood samples was performed by staff at the NHS Blood and Transplant Department, Addenbrooke's Hospital, University of Cambridge.

Experimental work and assay development was supervised by Kate Downes as part of Willem Ouwehand's team at the Department of Haematology, University of Cambridge. Genotype QC and processing were performed by Heather Elding and the merging of the genotypes relevant for this study was performed by Klaudia Walter. All other analyses were performed by myself.

The Sanquin cohort was coordinated in conjunction with Anton Tool, Taco Kuijpers and Judy Geissler. Replication and further experiments were performed by Evelien Sprenkeler and Anton Tool at Sanquin Research, The Netherlands.

# 4 Dissecting the functional relationship between neutrophil count and surface expression of cellular receptors

## 4.1 Introduction

In the previous chapter, I discussed how studying genetic regulation of cellular traits could be complementary to studying molecular phenotypes in discovering novel pathways or genes involved in immune responses. Heritable genetic variation of a broad range of blood cell frequencies has also been previously identified using quantitative FACS-based immunophenotyping (Orru et al., 2013). In this study, immune cell frequencies of 95 cell types were profiled and the genetic contribution of 272 immune traits in a cohort of 1,629 Sardinian individuals was evaluated (Orru et al., 2013). The authors identified 23 independent variants at 13 loci and found three loci were known autoimmune risk genes (*HLA, IL2RA* and *SH2B3/ATXN2*). Hierarchical gating using antibody-labelled cellular receptors enabled the assessment of cell types such as regulatory T ($T_{REG}$) cells, which are mostly characterised by the surface expression of CD39. These cells play important roles in regulating immune responses and preventing autoimmunity and from this study, were observed to be the most heritable traits (mean 55%).

In addition to investigating cell frequencies, studying the cellular surface expression levels of proteins adds a further layer of functional insight into immune responses (Roederer et al., 2015). The heritability of 78,000 immunophenotype traits in 669 female twins was evaluated, and 11 genetic loci explaining up to 36% of the variation in 19 traits (from the top 151 heritable traits genetically assessed) were identified (Roederer et al., 2015). In this study, associations with two different mechanisms are described; the homeostatic regulation of cell levels through proliferation or elimination of a certain cell type and the regulation of the expression of the protein, presumably through variants affecting promoter or enhancer activity (Roederer et al., 2015). The authors examined the same highly heritable $T_{REG}$ *ENTPD1/CD39* locus, through a variant in LD with the first identified by Orru *et al.* (2013), and found the association was explained by the phenotype (i.e. expression level of CD39), rather the frequency of $T_{REG}$ cells (Roederer et al., 2015). Therefore, cell population frequencies and the surface expression of phenotypic receptors are both highly heritable and thus integrating multiple sources of functional data aids our understanding of biological systems. In both studies, significant loci were also known to be associated with disease risk.

I discussed in Chapter 1 how using an automated haematology analyser enables the measurement of mature blood cell counts, albeit a lower range of cell types compared to a FACs-based approach but the method is amenable for large cohorts, in this case of 173,480

European-ancestry individuals (Table 1.2) (Astle et al., 2016). This GWAS exemplified how studying the variation of haematological traits provided insight into blood cell biology and also the general architecture of complex traits. An unprecedented 2,706 independent variants associated with variation in 36 indices were identified (Astle et al., 2016). The functional importance of both coding and regulatory variants was demonstrated. Enhancer variants explained 19%-46% of heritable variation, similar to that of transcribed regions (4-30%). Coding variants were enriched in Mendelian genes, and medically important observations were made in leukocyte subsets, which previously had been investigated in studies of limited power (Astle et al., 2016). The associations from this GWAS provide a rich resource where in-depth functional interrogation of this dataset offers the opportunity to further dissect haematopoietic cellular biology.

From an in-depth investigation of genetic loci associated with neutrophil count from the Astle *et al.* (2016) blood GWAS, I observed many cases of significant variants located within or nearby to genes encoding cellular receptor proteins. Roederer *et al.* (2015) posit that the surface expression of proteins represents an independent mechanism compared to cell levels. The variants from the Astle *et al.* (2016) GWAS were associated with neutrophil count but located in genes encoding receptors not traditionally used to quantify cell phenotypes. In addition, these receptors were known to be involved in neutrophil biology. Therefore, I postulated that variation in receptor levels associated with these loci, could be functionally linked to cell count. Differences in the surface expression of the receptor protein could result in altered receptor signalling, which in turn influence the numbers of circulating cells.

After applying prioritisation methods described in Materials and Methods, I investigated variants associated with neutrophil count located in genes encoding two receptors, the granulocyte colony-stimulating factor receptor (GCSFR) and the urokinase receptor (PLAUR).

GCSFR is expressed on the surface of progenitor and mature neutrophil granulocytes, with higher surface receptor expression levels detected at later stages of development, on more mature neutrophils (Nicola and Metcalf, 1985, Panopoulos and Watowich, 2008). Differentiation from haematopoietic stem cells to granulocytes is highly dependent on G-CSF and to a lesser extent, GM-CSF (Mehta et al., 2015). The main motivation for investigating a possible functional relationship between count and surface expression is that GCSFR is essential for granulopoiesis and the cognate ligand to the receptor, G-CSF, is the principal cytokine-regulator of neutrophils (Panopoulos and Watowich, 2008). Both G-CSF and GCSFR- deficient mice have chronic neutropenia, with significant reductions in the levels of peripheral neutrophils and granulocytic precursors in the bone

marrow (Lieschke et al., 1994, Liu et al., 1996). In normal conditions, G-CSF stimulates proliferation of all stages of granulopoiesis and increases neutrophil survival (Lord et al., 1989, Liu et al., 1996). This increased proliferation and survival is important to meet the demand of infection but is also exploited in clinical administration of G-CSF to neutropenic patients with very low neutrophil numbers (Panopoulos and Watowich, 2008, Sung and Dror, 2007). Both effects are termed 'emergency' granulopoiesis. (Semerad et al., 2002). Resolving the mechanism whereby variants located in the *CSF3R* gene are associated with neutrophil count is therefore clinically relevant.

In addition to stimulating neutrophil production, G-CSF also affects function (Betsuyaku et al., 1999). The ROS response to fMLP is increased in neutrophils primed with G-CSF (Betsuyaku et al., 1999). The residual neutrophils that remain in GCSFR-deficient mice show impaired functionality such as adhesion and migration in response to IL8 (Betsuyaku et al., 1999). GCSFR signalling is therefore required for particular functions of normal neutrophils.

Acquired GCSFR mutations increase the risk of acute myeloid leukaemia (AML) in patients with severe chronic neutropenia (SCN), for example due to C-terminal truncations that impair internalisation, increase surface receptor levels and stimulate proliferation over differentiation (Touw, 2015, Liongue and Ward, 2014, Ward et al., 1999, Aarts et al., 2004). Hereditary autosomal GCSFR mutations have also been observed, for example, T617N results in chronic neutrophilia due to constitutive activation of GCSFR as a result of dimerisation of the transmembrane domain (Plo et al., 2009). Clearly, GCSFR plays a key role in controlling neutrophil numbers and differentiation.

I investigated a second receptor to evaluate whether a functional relationship between surface expression and neutrophil count may be widespread. The plasminogen urokinase receptor (PLAUR/uPAR) receptor missense variant, rs4760, was highly associated with neutrophil count. PLAUR, associates with the plasma membrane via a GPI-anchor, and binds and activates the extracellular urokinase-type plasminogen activator (uPA/urokinase) (Smith and Marshall, 2010). Active uPA then generates the protease plasmin, which in turn cleaves extracellular matrix components. Through this process, uPAR regulates proteolysis, cell survival, growth and migration (Smith and Marshall, 2010). Expression of PLAUR is increased during inflammation and tissue remodelling and is correlated with poor prognosis in cancer making this receptor a potential therapeutic target (Smith and Marshall, 2010, Del Rosso et al., 2008).

Neutrophils express uPA and its receptor, PLAUR/uPAR (Gyetko et al., 1995). PLAUR functionality plays an important role in leukocyte adhesion and migration. PLAUR expression is increased during differentiation and activation of leukocytes, suggesting this receptor plays an important role in immune function (May et al., 1998, Nusrat and Chapman, 1991, Smith and Marshall, 2010). Treatment with an anti-CD87 (PLAUR) antibody inhibited chemotaxis in PMNs but was unaffected by anti-uPA antibodies, implicating the receptor in neutrophil function and migration (Gyetko et al., 1995). During specific infection with the bacteria *S.pneumoniae,* PLAUR-deficient mice had diminished granulocyte accumulation in the lungs and reduced survival, clearly providing evidence of the importance of PLAUR in neutrophil inflammatory responses (Rijneveld et al., 2002).

### 4.1.1 Aims of this chapter

I performed flow cytometry experiments in a recall-by-genotype (RbG) study (Section 1.6) to test whether pre-selected significant neutrophil count variants could also affect the surface expression of these receptors, potentially reflecting a functional link between this and neutrophil count. I also integrated my findings with available data sources including neutrophil molecular traits from the BLUEPRINT consortium (Chen et al., 2016a). In comparison to the functional GWAS I carried out in Chapter 3, which was performed over a year, this study took place over only two weeks. Blood was collected by the same nurses, and the same machine and reagent batches were used. Increase control over covariates in this shorter study allowed greater control over technical variability such as that observed in Chapter 3.

Using such a RbG study design, I collected a panel of 2 cell surface markers of specific gene candidates (see below) measured in up to 70 individuals divergent for alleles of the independent neutrophil-count associated variants (Figure 4.6 and Section 4.2.1). I measured the mean fluorescence intensity of selected receptors, PLAUR and GCSFR on the surface of neutrophils in whole blood from healthy donors. I then tested the association of SNPs located within the receptor gene (and +/- 500 kB) with the level of these receptors as measured using flow cytometry. I then further integrated molecular sources of information and other external datasets to explain my observations and provide functional hypotheses.

## 4.2 Materials and Methods

### 4.2.1 Selection of receptor and genetic variant candidates for experimental follow up

This hypothesis-driven investigation was the result of my efforts to thoroughly integrate genetic information from a neutrophil count trait GWAS (Astle et al., 2016) with epigenetic and transcriptional information from the BLUEPRINT consortium (Chen et al., 2016a).

As of December 2015, there were 17,673 variants associated with neutrophil count, which were clustered into 134 high LD groups (Astle et al., 2016). For the selection of candidates, I focused on common variants to ensure sufficient donor stratification across genotype groups in a cohort of less than 100. Applying this filter reduced the number of variants to 17175 and LD groups to 124. Genes were assigned to each variant if a variant overlapped a gene(s) and also the nearest upstream and downstream gene was also assigned. This resulted in a total of 567 unique assigned genes. Although these genes were assigned solely on proximity, a number of the most significant gene ontology terms identified using g:profiler suggested relevance to neutrophil biology, such as phagosome (p value = $9.92 \times 10^{-13}$), immune response (p value = $3.42 \times 10^{-11}$) and cytokine-mediated signalling pathways (p value = $6.40 \times 10^{-10}$) (Reimand et al., 2016). Of these 364 had a known function using the PANTER classification gene ontology system (Version 12.0) (Mi et al., 2017). 11.5% of genes with a known function were annotated with “receptor activity” (Figure 4.1). I found, for example, that of the variants located directly within a gene, 21% had known receptor function including *CSF3R*, *SLC25A24, FCGR2B, SLC12A7, SLC22A4, HLA-A, HLA-C, SLC44A4, HLA-DRB1, PLAUR, LY75, ACKR2, MYO1G, ACVRL1* and *VMP1.*

I speculated that there may be a relationship between receptor cellular surface expression and neutrophil count, likely underpinning the role of signalling and cell communication during differentiation of mature neutrophils. Therefore, I set to test the hypothesis that genetic variants affecting neutrophil count could exert their effect through a change in receptor expression on the cell surface. I established several additional criteria for the selection of downstream variants from variants annotated with a known receptor gene. There must be a known function of the receptor in neutrophil biology. The region of association must be resolved to one or a few SNPs so there is a higher chance of identifying a causal signal linked to the experimental measurement.

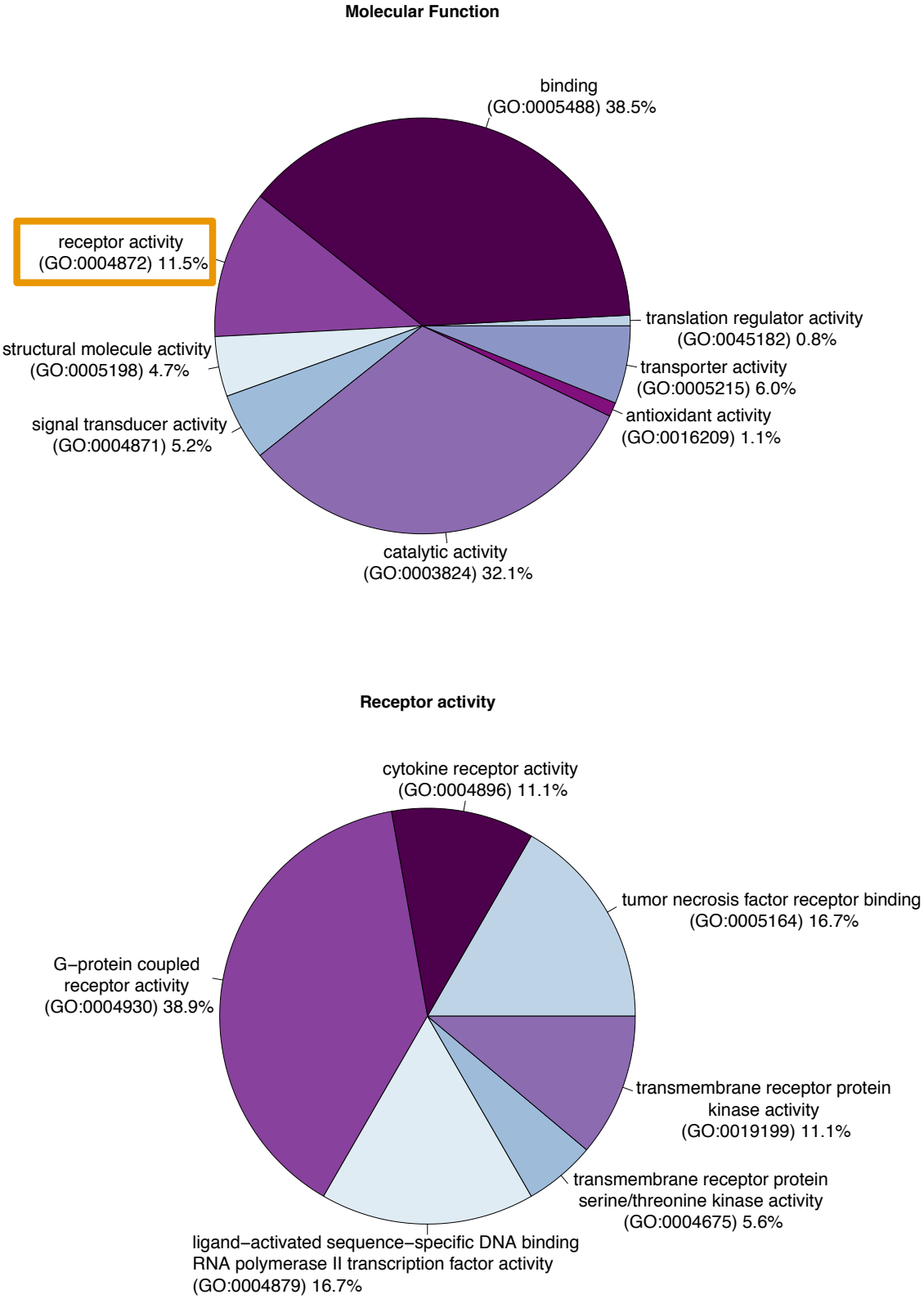


**Figure 4.1: Molecular function of neutrophil count variant assigned genes**
Pie charts show the proportions of molecular function annotations of the 364 genes with functional hits in the PANTER version 12 classification system. Receptor activity (11.5%) is highlighted in orange. The second pie chart (below) shows the proportions of molecular function terms within the highlighted receptor activity annotation.

I used neutrophil promoter-capture HiC data to verify that the assigned gene was the likely gene targeted by the variant (Javierre et al., 2016). If the presence of long-range chromosome interactions suggested the variants could be interacting with other likely candidate genes, then these loci were excluded. Criteria for gene functionality included expression in neutrophils or a known role of the gene in neutrophil biology. There must be available an antibody against the candidate receptor that has been previously validated in neutrophils. The antibody must be available conjugated to the fluorophore, phycoerythrin (PE), which emits a bright and stable fluorescence. The experimental design of the recall study, with up to 13 donors per day, limited the number of candidates I could feasibly investigate to a maximum of three.

I applied these criteria to annotated genes as described above and prioritised two receptor candidates for downstream analysis; the granulocyte colony stimulating factor 3 receptor (GCSFR) and the urokinase receptor (PLAUR/uPAR). The existing genetic evidence for these loci is discussed below.

#### 4.2.1.1 G-CSF receptor

GCSFR is encoded by the *CSF3R* gene (Entrez Gene:1441, Ensembl:ENSG00000119535). I have described how GCSFR plays a key role in differentiation leading to the production of mature neutrophils. I postulated that variation in surface expression levels could impact on neutrophil count or vice versa.

An overview of the neutrophil count association in the Astle *et al.* (2016) study in this region is shown in Figure 4.2. The variant rs3917932 had the strongest association with neutrophil count (EAF =0.42, beta = 0.048, SE = 3.63 x $10^{-03}$, p value = 2.06 x $10^{-39}$, N = 173,480, Table 4.1) and is also associated with granulocyte count, myeloid count, white blood cell count, monocyte percentage, neutrophil percentage, lymphocyte percentage, granulocyte percentage of myeloid cells, neutrophil + eosinophil count and basophil + neutrophil count. Conditional analysis was performed at the locus as part of the Astle *et al.* (2016) study. The authors also found a second low-frequency variant, rs3917914 (EAF = 0.01, beta = 0.16, SE =1.69e-02, p value = 1.69 x $10^{-22}$) in the same locus by regressing out the common signal (Astle et al., 2016). rs3917914 is also associated with granulocyte count, myeloid count, white blood cell count, neutrophil + eosinophil count and basophil + neutrophil count.

To the best of my knowledge, there is no previous study investigating whether genetic variation in the neutrophil receptor surface expression also influences neutrophil count in healthy individuals. The experimental process I designed in this chapter tests the surface expression of GCSFR on mature circulating neutrophils. The count measured in the Astle *et*

*al.* (2016) paper describes the numbers of mature circulating neutrophils in a unit of whole blood. Therefore, this allows me to test my hypothesis that the expression of surface receptors on mature neutrophils is related to the numbers of mature neutrophils in the circulation using genetics as an anchor for causality.

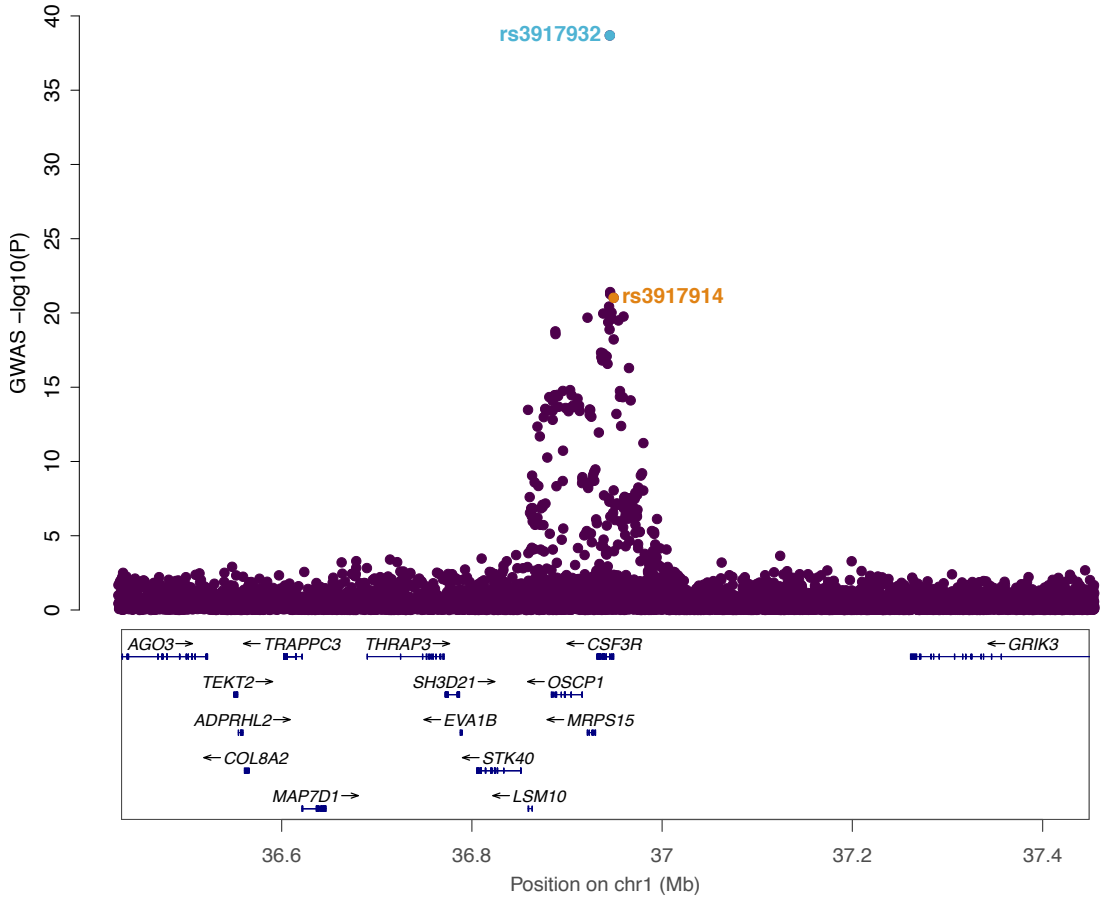


**Figure 4.2: Genetic variants associated with neutrophil count in and around the *CSF3R* gene**
Regional association plot of variants associated with neutrophil count around the *CSF3R* locus. Conditional analysis performed as part of the analysis in Astle *et al.* (2016) revealed two independent Intronic signals, rs3917932 (common) and rs3917914 (rare/low frequency). The chromosomal location (hg19) is shown on the x-axis, the left-hand y-axis is the -$\log_{10}$(p-value) of association with the trait.

#### 4.2.1.2 PLAUR

The second candidate under investigate was the PLAUR receptor, encoded by the *PLAUR* gene (Entrez Gene:5329, Ensembl:ENSG00000011422). One missense variant, rs4760, is located within an exon of the *PLAUR* gene and is significantly associated with neutrophil count (EAF=0.84, beta = $8.61 \times 10^{-02}$, SE= $4.92 \times 10^{-03}$, p value = $1.43 \times 10^{-68}$) (Table 4.1). rs4760 is also associated with granulocyte count, myeloid count, white blood cell count, monocyte percentage, neutrophil percentage, lymphocyte percentage, granulocyte percentage of myeloid cells, neutrophil percentage of granulocytes, neutrophil + eosinophil count and basophil + neutrophil count (Astle et al., 2016). rs4760 is associated with neutrophil count approximately 40 orders of magnitude more strongly than remaining SNPs in the proximal region. These proximal SNPs are also located within the neighbouring gene, *CADM4* (Figure 4.3). Therefore, it is highly likely that rs4760 is the causal SNP in this locus.

The rs4760 missense variant causes a leucine to proline amino acid change at residue 317 (L317P). rs4760 is predicted to be possibly damaging in the angiogenesis pathway by PolyPhen prediction from the Cancer Genome Anatomy Project Genetic Annotation (Savas et al., 2006). Residue 317 is located within a PLAUR protein isoform 1 domain predicted to be non-cytoplasmic using Phobius (Kall et al., 2007). The probability of that this domain is transmembrane decreased when I manually substituted the leucine 317 for a proline (0.4 to 0.1 where 1 is maximum likelihood) (Kall et al., 2007). This may suggest that this residue could impact the transmembrane domain or non-cytoplasmic domain functionality.

Therefore, rs4760 and the PLAUR receptor were included in this study given the role of the PLAUR receptor in neutrophil biology and particular disorders as well as the prediction that this variant could affect receptor stability. The inclusion of PLAUR and GCSFR also enables dissection of the effects of both intronic (*CSF3R*) and missense (*PLAUR*) SNPs.

In comparing the two receptor candidates studied here, GCSFR plays a pivotal role in neutrophil differentiation. PLAUR has been implicated in macrophage differentiation and phorbol-ester mediated differentiation of the neutrophil model cell line, HL60 (Rao et al., 1995, Nusrat and Chapman, 1991). Therefore, demonstrating a shared relationship between neutrophil count and PLAUR surface expression might indicate a possible role in neutrophil development. In the case of GCSFR, indication of a shared relationship could highlight the importance of receptor number on the surface in differentiation by directly linking receptor signalling to neutrophil count.

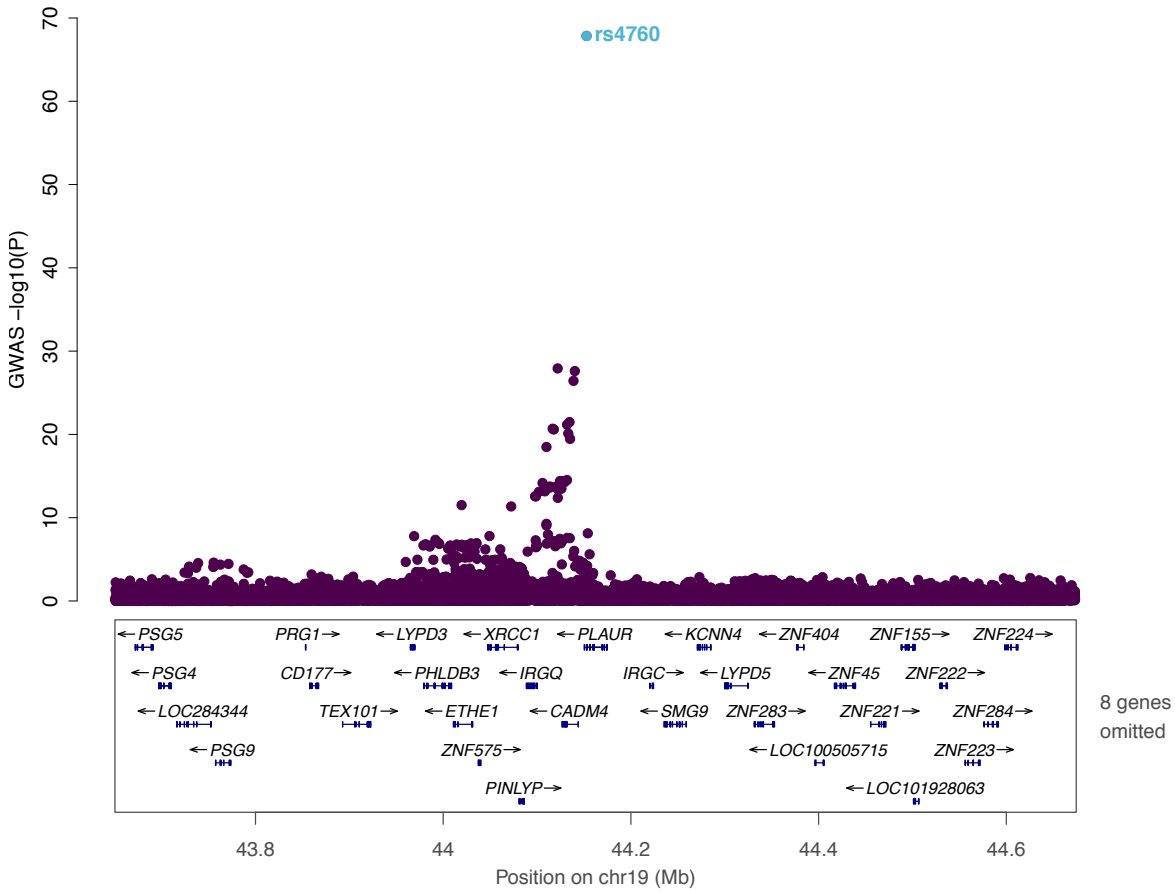


**Figure 4.3: Genetic variants associated with neutrophil count in and around the *PLAUR* receptor gene**

Regional association plot of variants associated with neutrophil count from the Astle *et al.* (2016) study. The most significant SNP, rs4760 is a missense SNP located in an exon of the *PLAUR* gene. rs4760 is more than 40 orders of magnitude more significant than remaining SNPs in the locus and is therefore predicted to be causal for the neutrophil count association in this locus.

| rsID | Chr:pos (hg19) | EA/OA | EAF | Effect (SE) | Astle Neu count P | Gene | Type |
|---|---|---|---|---|---|---|---|
| rs4760 | 19:44153100 | G/A | 0.16 | $-8.61 \times 10^{-2}$ ($4.92 \times 10^{-03}$) | $1.43 \times 10^{-68}$ | *PLAUR* | Missense L317P |
| rs3917932 | 1:36943916 | C/G | 0.42 | $4.80 \times 10^{-02}$ ($3.63 \times 10^{-03}$) | $2.06 \times 10^{-39}$ | *CSF3R* | Intron 3 full transcript (First intron of CSF3R-204) |
| rs3917914 | 1:36947888 | A/G | 0.01 | $1.60 \times 10^{-01}$ ($1.69 \times 10^{-02}$) | $9.54 \times 10^{-22}$ | *CSF3R* | Intron 1 of full transcript (and truncated CSF3R-008 |

**Table 4.1: Candidate variants selected for functional follow-up**
Neutrophil-count (number of neutrophils per nL per unit volume of blood) associated variants from the blood count GWAS (Astle et al., 2016). The effect size is expressed in SD of the trait. The conditionally independent variant(s) for the *CSF3R* (one common, one rare) and *PLAUR* (one common) loci are listed. *CSF3R* gene transcripts, as referred to in the table, are shown in Figure 4.10. EA = the effect allele, here defined as corresponding to the decreasing receptor expression allele for the *CSF3R* locus (see Table 4.5). SE = standard error. EAF = effect allele frequency. P = p value of neutrophil count.

### 4.2.4 Study samples

Individuals were recalled from the NIHR Cambridge BioResource (http://www.cambridgebioresource.group.cam.ac.uk) as part of the UNICORN study organised at the NHSBT, Addenbrooke's Hospital. All individuals were of blood group O. Peripheral blood samples were collected from healthy donors with informed consent (A Blueprint of blood cells, REC 12/EE/0040, East of England-Hertfordshire Research Ethics committee). The cohort used for functional interrogation of the relationship between receptor surface expression and neutrophil count comprised of 70 individuals of European descent. The mean age of the individuals was 61.27 years, and the range was from 25 to 79 years. The cohort consisted of 27 males and 43 females. After integrating with available genotype data, and exclusion of outliers, the final sample numbers used for association were 66 (GCSFR) and 65 (PLAUR).

### 4.2.5 Flow cytometry assessment of receptor surface expression

Blood was collected from 70 individuals across seven days. The number of donors processed per day ranged from six to thirteen. Blood samples were experimentally processed in batches of three-four donors. In the experimental processing, antibody volumes (Table 4.2) were first pipetted into the bottom of each tube. 100 μl of blood from each donor was then added to each tube. For each individual, three separate tubes were prepared, two receptor tests and one unstained sample (no antibody) (Table 4.2). In tubes 1-2 (receptor tests), the sample was labelled with specific antibodies for CD16 (APC), CD66b (FITC), which were used to identify the double positive neutrophil population in the gating strategy (Figure 4.4). In tube 1 the antibody against CD86 (PLAUR-PE) was added and in tube 2 the antibody against CD193 (GCSFR-PE) was added. After adding blood, tubes were mixed by inverting three times. After 20 minutes of incubation in the dark at room temperature, the red blood cells were lysed by addition of 2 ml of lysing solution and vortexed for three seconds. Following a six-minute incubation in the dark at room temperature, samples were then centrifuged at 600 x g for six minutes at 4°C (accelerator 9 and break 9). The supernatant was discarded and the pellet was re-suspended gently in 500 μl of PBS. Samples were vortexed for three seconds and stored at 4°C until ready for flow cytometry analysis. The time of labelling, lysing, fixing was recorded and investigated as potential covariates (see below). The samples were measured using a Beckman Coulter Gallios™ Flow Cytometer system.

| Tube | Blood Vol (μl) | Antibody | Ab Vol (μl) | Lysis Vol (ml) | Other | Tube Type |
|---|---|---|---|---|---|---|
| Tube 1 | 100 | CD16<br>CD66b<br>PLAUR/CD87 | 1.25<br>5<br>18 | 2 | - | PLAUR test |
| Tube 2 | 100 | CD16<br>CD66b<br>CSF3R/CD114 | 1.25<br>5<br>9 | 2 | - | GCSFR test |
| Tube 3 | 100 | - | - | 2 | - | Unstained |
| Tube 4 | 0 | CD16 APC | 1.25 | - | One drop of each negative/positive compensation beads | Compensation Control |
| Tube 5 | 0 | CD66b FITC | 5 | - | | |
| Tube 6 | 0 | PE test/CD193 | 4 | - | | |

**Table 4.2: Contents of each sample tube per donor**
Table summarises the experimental design where six tubes were made per donor, two tubes contained different combinations of antibodies, one was the unstained sample and three were compensation controls.

Specific controls were prepared for each experiment. First, for every donor, an unstained sample was processed in parallel except without antibody addition (Figure 4.4, Table 4.2). Second, compensation beads for PE, FITC and APC were prepared once per week and measured on the Gallios™ every day. One drop of each compensation beads was added to each of three tubes (Table 4.2 and 4.4). Compensation beads provide two distinct polystyrene micro-particle populations; the positive population which binds to mouse κ light chain immunoglobulins and the negative population of beads with no binding capacity. Compensation was required to adjust for any spectral overlap in the three colours, PE, FITC and APC used in these experiments. 1.25 μl of anti-APC, 5 μl of anti-FITC and 4 μl PE were added to tubes 4, 5 and 6 respectively. Tubes were incubated in the dark at RT for five minutes. 1 mL of PBS was then added to each tube and these were stored at 4°C until use. The FlowJo software auto-compensation procedure was used to adjust for the multicolour flow cytometry data (PE, FITC and APC were used all in one tube to detect GCSFR/CD87, CD66b and CD16 respectively).

Third, control beads were run daily to verify the optical alignment of the lasers and functionality of the fluidics system. Flow-Check™ Pro Fluorospheres, a suspension of fluorescent microspheres, were analysed on the Gallios™ machine. Prior to sample ascertainment, mean fluorescence values for each laser were measured to detect any deviations from the pre-selected ranges that may require re-alignment of the lasers. Laser voltage adjustment was required only on the second day of experiments and correct functionality was confirmed after adjustment using the Flow-Check™ Pro Fluorospheres. Lastly, isotype controls against the functional receptors (GCSFR and CD87) (Table 4.3) were

run for one donor and confirmed the lack of a high level of non-specific background signal (data not shown).

| Protein | Antibody | Colour | Source | Volume (μl) |
|---|---|---|---|---|
| CD16 | Mouse anti-human, VEP13, mouse IgM | APC | Miltenyi, 130-091-246 | 1.25 |
| CD66b | Mouse anti-human, G10F5, mouse IgM, k | FITC | BD Pharmigen, 555724 | 5 |
| CD87/PLAUR | Mouse anti-human, VIM5, mouse IgG1, k | PE | BD Pharmigen, 555768 | 18 |
| CD114/CSF3R | Mouse anti-human, LMM741 (RUO), mouse IgG1, k | PE | BD Pharmigen, 554538 | 9 |
| Isotype control for CD87 and CD114 | Mouse Isotype IgG1, k | PE | BD Pharmigen, 555749. Lot: 3046675 | 20 |
| Isotype control for CD66b | Mouse IgM K, Clone G155-228 | PE | BD Pharmigen, 555584 | 20 |

**Table 4.3: Antibodies used for each marker in flow cytometry assays**
Evidence of previous use in neutrophils for each clone type was assessed before final selection of the antibody (CD87 VIM5 (Elghetany et al., 2003), CD114 LMM741 (Piper et al., 2010)).

| Bead/Reagent | Supplier | Code |
|---|---|---|
| Flow-Check Pro Fluorospheres | Beckman Coulter | A63493 |
| Anti-mouse Ig, κ/negative control compensation particles set | BD Bioscience (BD™) CompBead | 552843 |
| Set-up beads for green/yellow laser | Life Technologies | C16508 LOT: 1438512 |
| Set-up beads for blue laser | Life Technologies | C16509 LOT: 1438509 |
| Set-up beads for red laser | Life Technologies | C16507 |
| Lysing (and fixing) solution 10X Concentrate | BD Biosciences, BD FACS™ | 349202 |

**Table 4.4: Reagents and beads**
This table lists the beads used for the control experiments, the supplier and catalogue number along with the reference for the lysing solution used in the protocol to remove red blood cells.

*Gating Strategy:* Figure 4.4 shows the gating process used for all individuals to identify neutrophil populations. First, the fluorescence signal from the FL2 laser (PE) across time of measurement was assessed. This gate was used to remove potential machine technical factors such as debris in the fluidics system. The time gate was selected manually for each sample and judged to remove regions of lower density fluorescence measurements at the start of the reaction. A similar gating procedure was applied in a recent paper, also measuring median fluorescence intensity (MFI) of defined receptors on the surface of immune cells (Roederer et al., 2015).

The granulocyte population of neutrophils was selected using a plot of forward scatter height (FS-H) and side-scatter height (SS-H) (Figure 4.4). This population contains both granular cells, neutrophils and eosinophils. Doublet cells were then removed and following this, the double positive CD16+CD66b+ neutrophil population was selected. CD66b was used as a marker of granulocytes and is expressed on the surface of both eosinophils and neutrophils (Yoon et al., 2007). CD16 is not expressed on the surface of naïve eosinophils, therefore allowing specific selection of neutrophil populations (CD66+CD16+) (Davoine et al., 2002). 10,000 events/cells of the double-positive neutrophil population were collected for each sample. The median fluorescence intensity (MFI) of the receptors (PE, FL2 signal) was calculated from this population and used as an estimate of the surface protein expression.

### 4.2.6 Phenotype processing and genetic association

*Phenotype processing:* Receptors were treated separately as tube 1 (PLAUR) and tube 2 (GCSFR) respectively. The receptor MFI level of the unstained population (per donor) was subtracted from the receptor MFI level of the double positive CD16+CD66b+ neutrophil population and referred to as the normalised MFI value. The data were analysed and MFI statistics calculated using the FlowJo analysis suite version 10.1r5.

*Outlier removal:* the mean and standard deviation of the normalised MFI of a total of 67 donors was calculated, and any MFI value outside of the mean +/- 3 standard deviations thresholds were excluded. Overall, there were 66 individuals with genotype data for GCSFR and 65 for PLAUR, shown in Figure 4.5.

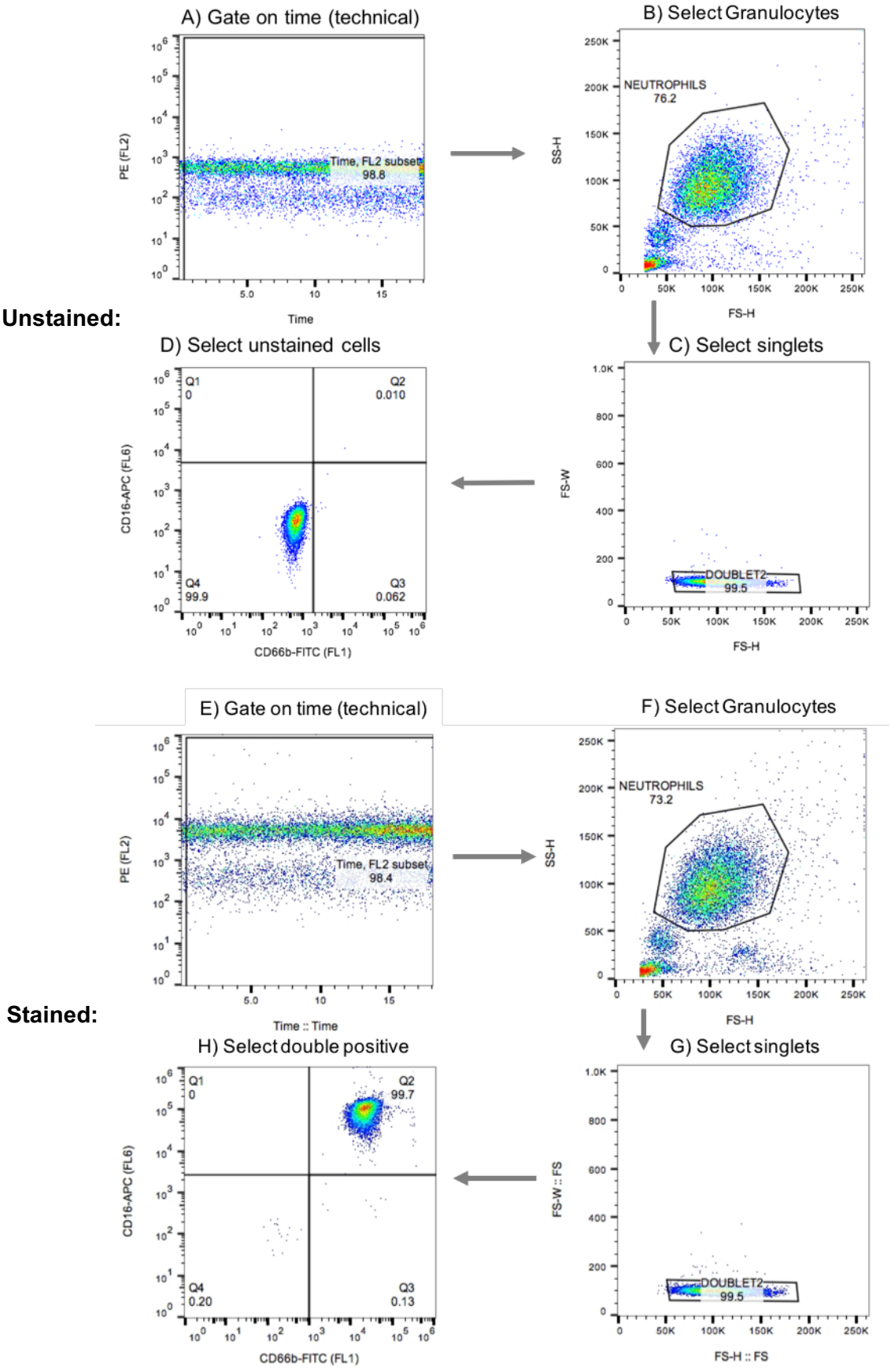


**Figure 4.4: Gating strategy**
Fluorescence plots used to measure neutrophil surface receptor MFI from whole blood. A)-D) Unstained E)-H) Stained**.** First the time gate removed technical artefacts. The granule population was selected using FS-H and SS-H (forward scatter height and side scatter height). Doublets are removed and finally the double-positive neutrophils are selected and MFI of either PLAUR or GCSFR was measured for this population using the PE signal.

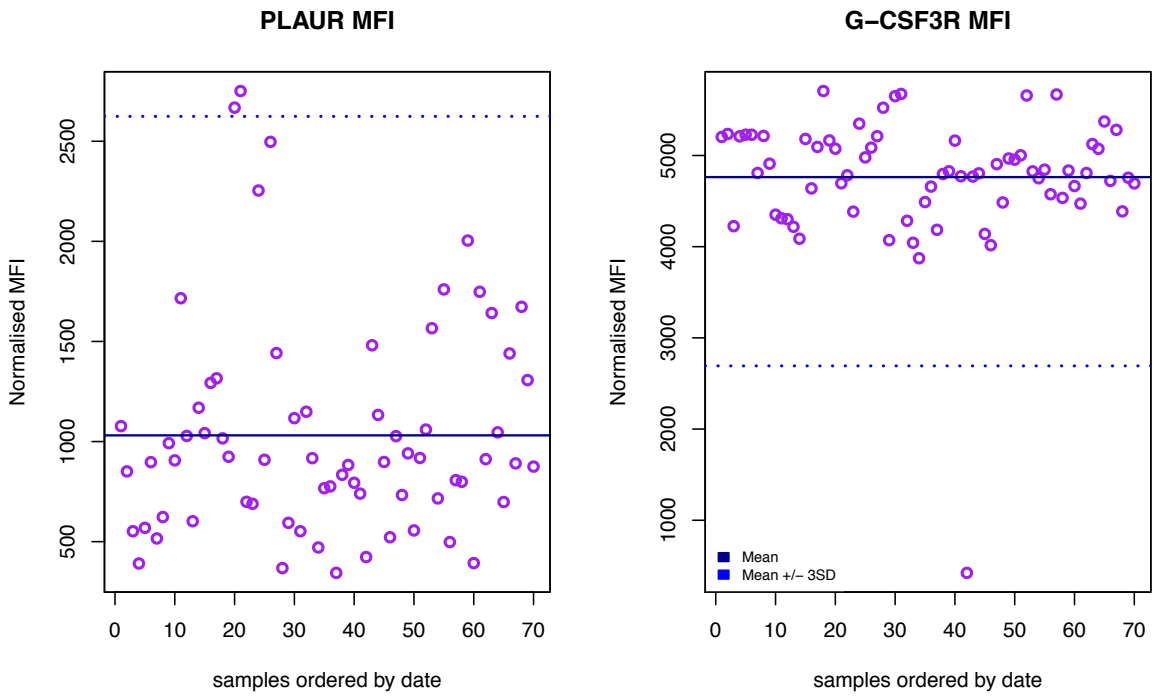


**Figure 4.5: MFI by time including outlier exclusion thresholds**
Plot of normalised MFI levels (unstained subtracted) in order of acquisition (samples by date). The mean of the MFI of all individuals is shown by the solid line. The mean +/- 3 standard deviations (SD) is shown by the dashed line. Donors with MFI values falling outside of these thresholds were excluded from the downstream analysis.

*Covariate investigation and correction:* Before association testing, I investigated potential sources of technical co-variation and stratification. Known covariates included age, sex, full blood count measures, date and experimental information that I recorded. These included time of bleed, labelling, and fixing and experimental processing batch. Given the compressed time-frame of this study, no reagent or antibody batch changes occurred. Up to 13 donors were processed per day prohibiting experimental processing of all samples together. Individuals were bled at 30-minute intervals; therefore, batches of three-to-five donors were processed (labelled and fixed) together. This minimised the time between bleed and processing. Groups of samples were designated into an experimental processing batch if there was an unforeseen change in the experimental protocol that could have introduced variation. For the time covariates, if there were too many levels (in some cases, there was only one donor with a specific time), bins of ten minutes were created.

Full blood count was measured using both a Sysmex XE-5000 and XN-1000 haematology analyser for each donor (collected at the time of blood collection by NHSBT collaborators). Results from both XE-5000 and XN-1000 were considered, and in the case of covariates, the most significant association was used to select the measurement for covariate selection. I used ANOVA to assess whether covariates had a significant effect on the mean values of inverse normalised (and unstained normalised) MFI across different levels. No significant covariates (p value < 0.05) were found for GCSFR MFI levels. However, given that the experimental processing covariate reflects known experimental variation, I took the approach of conservatively adjusting for this batch effect using a linear model with processing batch as a predictor variable and GCSFR as the response variable, which marginally improved the strength of association.

I identified several significant covariates for PLAUR MFI values including date, experimental processing batch, sex, eosinophil count, time from bleed to labelling and time from bleed to fixing. Some of these covariates potentially measured the same effects, for example, the two time-period covariates and date and experimental processing batch. I corrected for each covariate in a linear model and used the residuals to test if the second covariate was still significant. In all cases the covariates were no longer significant in the second iteration of linear regression. Therefore, I corrected for the covariate with the lowest ANOVA p value.

In summary, date, sex, time from bleed to labelling and eosinophil count were used as covariates in a linear regression model to correct for these effects on the PLAUR surface expression levels. After outlier removal, I inverse normalised the unstained subtracted MFI values and then used the lm() function in R to correct for the specified covariates. Covariates were input as factors (date, sex, experimental processing batch) or as numeric for covariates such as time from bleed to labelling and eosinophil count. The effect of these covariates on PLAUR and GCSFR MFI levels are shown in Supplementary Figures 4.1 and 4.2. The corrected residuals were then standardised and used as an input into the genetic association tests described below.

I calculated the adjusted $r^2$ parameter of the linear model for each covariate independently against PLAUR MFI. The following adjusted $r^2$ values were obtained: Date (0.15), Sex (0.05), absolute eosinophil count (EO) (0.06) and time from bleed to labelling (0.20), indicating that date and time from bleed to labelling explained the most amount of variation in PLAUR MFI levels. The adjusted $r^2$ value for the overall model, correcting for all four covariates was 0.39.

*Investigation of neutrophil size:* I next investigated whether MFI was affected by neutrophil size where a larger neutrophil could express a greater number of receptors on the surface

and vice versa for a smaller neutrophil. Currently, it is not known whether there is significant variation in the size of neutrophils in a general population. Measurements using a DxH 800 Coulter haematology analyser showed mean neutrophil volume was larger in sepsis patients and has been suggested as an additional clinical diagnostic measure for acute bacterial infection (Lee and Kim, 2013, Chaves et al., 2005). Neutrophil size may, however, be highly regulated in healthy populations to ensure the circulation of neutrophils through blood vessels of defined diameters.

I used an independent cell size parameter, NE-FSC, measured by the Sysmex haematology analyser (see above). NE-FSC is the forward light intensity of the NEUT area and gives an indication of cell size. I also investigated NE-SSC, which is the side scattered light intensity and represents the internal complexity or granularity of neutrophils (Buoro et al., 2016, Sysmex Corporation, 2010-2012). I assessed the correlation of these two parameters with the surface expression of both receptors and found no strong or significant correlations of GCSFR or PLAUR MFI and neutrophil parameters (p value <0.05) (data not shown). The low non-significant correlations suggested, that within the power limitations of this study, there was no evidence of a relationship between cell size and receptor expression (represented by MFI). Therefore, I did not implement a further size-normalisation for the receptor MFI values in addition to the steps described above.

*Genotype processing:* Individuals were genotyped as part of a large cohort by the NIHR Cambridge BioResource. Genotype processing was performed by Heather Elding, and standard genotype quality control procedures were followed as described in (Anderson et al., 2010). Genotypes were imputed against the combined reference set of UK10K and 1000 Genomes Phase 1. Imputed genotypes were used for the single variant genetic association tests described below. As described in Chapter 3, I am currently validating genotypes of all lead SNPs in this cohort with Sanger sequencing, particularly to confirm genotypes of the significant rare SNPs.

*Single-variant genetic association and conditional analysis*: The variants were extracted from gen files containing all included genome-wide variants from the genotyped cohort of 80 individuals. The region for the *PLAUR* locus was set from chr19:43650247 to chr19:44674699 (500 kb upstream and downstream of the gene start and end positions). The region for the *CSF3R* locus was set from chr1:36431644 to chr1:37448879 (500 kb upstream and downstream of the gene start and end positions). Phenotype values were inverse normalised after outlier removal and corrected for covariates as detailed above. Standardised residuals were formatted in gen .sample format and input into SNPTEST v2.5, a slightly newer version that enables analysis of fewer than 100 samples (66 for GCSFR and

65 for PLAUR). The same linear model was fitted for each variant as described in Chapter 2 Materials and Methods using the same options except with the additional (-use_lower_sample_limit). For conditional analysis, this was run separately by conditioning on (condition_ on additive model) rs3917924 and rs3917912 as the top common SNP and rare SNP respectively. The analysis was also run by conditioning on both SNPs.

*Variant pruning:* PLINK v2 was used to generate a list of independent variants using the --indep option. The variants tested (gene +/- 500kb) were pruned using a pairwise $r^2$ threshold of 0.1 with a variant count window of 500. Pairs of variants in the current window with a squared correlation greater than the threshold are pruned until no such pairs in the window remain.

*Minor allele frequency calculation:* QCTOOL v1.4 was used to calculate minor allele frequency using the –snp-stats function using the final sample sizes after outlier removal.

*Linkage disequilibrium calculation:* The correlation between variants, $r^2$, was used to indicate the linkage disequilibrium between variants. For this functional dataset, genotypes were available for 80 individuals. This data was used to calculate $r^2$ for the 14 most significant variants using PLINK v2 --flag with input data in gen format (--data and --oxford-single-chr 1). To avoid duplicate IDs the flag --set-missing-var-ids @:# was used. In addition, the Astle *et al.* (2016) cohort was used to calculate the LD between variants reaching the significance threshold as detailed in Chapter 2 Materials and Methods.

*Locus zoom plot:* the LocusZoom website was used to generate initial locus zoom plots and cis-gene locations (Pruim et al., 2010) using summary statistics. 500 kb flanking of the gene was plotted. Custom scripts were used to generate the regional association plots with highlighted SNPs for the same genomic region.

*Gene annotation:* In order to annotate the genic location of variants, the same annotation as used in BLUEPRINT (Chen et al., 2016a), GENCODE version 15 (ENSEMBL release 70), genome version hg19, was used. Custom scripts were developed to calculate intronic regions from the GENCODE annotation to specifically locate which intron each transcript the variants were located in. Introns were annotated as all regions between each exon from the start and end of the *GCSFR* gene which falls on the negative strand. HAVANA and ENSEMBL annotated transcripts were treated separately.

*Transcript abundance and visualisation:* Transcript abundance was previously assessed within the BLUEPRINT consortium (Chen et al., 2016a). Briefly, transcripts were quantified

with Cufflinks v2.2.1 using RNA-seq data, Gencode v15 annotation and without *de novo* transcript assembly. Transcripts were visualised using the R package, ggplot2 (Wickham, 2009). sQTLseekeR was used for isoform splicing QTL mapping as part of the published BLUEPRINT study (Monlong et al., 2014). For evaluation of absolute transcript expression, rather than transcript ratios, the Cufflinks quantified expression in FPKM was used (Chen et al., 2016a).

*Epigenome intersection:* bedtools intersect version 2-2.23.0 was used to intersect molecular data with variant positions. Bed files for variant locations were generated by subtracting 1 from the position of the variant to act as the start position of the bed file.

*Replication cohort:* An independent cohort of 140 healthy individuals was established at Sanquin Research as described in Chapter 2. Donors heterozygous for rs3917912, rs3917914 and heterozygous and homozygous individuals for rs3917924 and rs3917932 were identified and used for planning replication experiments (Section 4.5, Future Work). These individuals, along with donors with non-effect haplotypes for all SNPs were/are being recalled at Sanquin Research and MFI of GCSFR on neutrophils is being measured by Anton Tool and Evelien Sprenkeler.

*Phasing haplotypes:* haplotypes were estimated in the original cohort using genotypes of three SNPs, rs3917932 (neutrophil count lead), rs3917924 (GCSFR lead) and rs3917914 (lead rare SNP), using SHAPEIT v2 and genotype data from 80 individuals (Delaneau et al., 2011). Three-SNP haplotypes were also estimated in the Sanquin replication cohort (above) using genetic data of the same three SNPs. In both cases, the HapMap phase II b37 genetic map was used to provide recombination estimates, as recommended. rs3917932 is missing from HapMap phase II cohort, in this case, SHAPEIT internally determines the genetic position. Haplotypes were phased using rs3917914 instead of rs3917912 as all genotype probabilities for rs3917914 heterozygous donors were above the threshold of 0.9. One donor heterozygous for rs3917912 with genotype probabilities less than 0.9 (approximately 0.8), SHAPEIT would therefore incorrectly call this genotype homozygous.

*Genotype validation:* I used Sanger sequencing with probes for the SNPs rs3917914 and rs3917912 to confirm that the heterozygous individuals carried at least one GCSFR-surface level-decreasing allele at this locus (as shown in Figure 4.8 for the discovery cohort association). The genotyping assay was designed by Agena Bioscience using the MassARRAY® System with the iPLEX® chemistry.

## 4.3 Results

### 4.3.1 Measurement of surface expression level in a cohort of 70 individuals

I investigated predicted target genes for neutrophil-count association variants and identified that 11.5% of genes with a known function were annotated with receptor activity. I hypothesised that given the role of receptors in cell signalling and the importance of these processes in differentiation that there might exist a relationship between significant neutrophil-count variants located within receptor genes and the neutrophil surface expression of these corresponding receptors. I predicted a possible functional role between these two traits for the G-CSF cytokine receptor based on the importance of the signalling pathways of this receptor in controlling neutrophil differentiation and mature neutrophil counts (Panopoulos and Watowich, 2008). I also investigated the PLAUR receptor missense variant given the role of this receptor in neutrophil function and as a comparison to GCSFR, a receptor with a well-established role in neutrophil development.

Figure 4.6 summarises the study design to test these hypotheses. Briefly, peripheral blood mononuclear cells (PBMCs) were stained with antibodies against CD16, CD66b, CD114 (GCSFR) and CD87 (PLAUR) and the surface expression of the latter two receptors was measured in a CD16+CD66b+ neutrophil population. The mean fluorescence intensity (MFI) of the population is used in this study to represent a quantitative measurement of receptor surface expression.

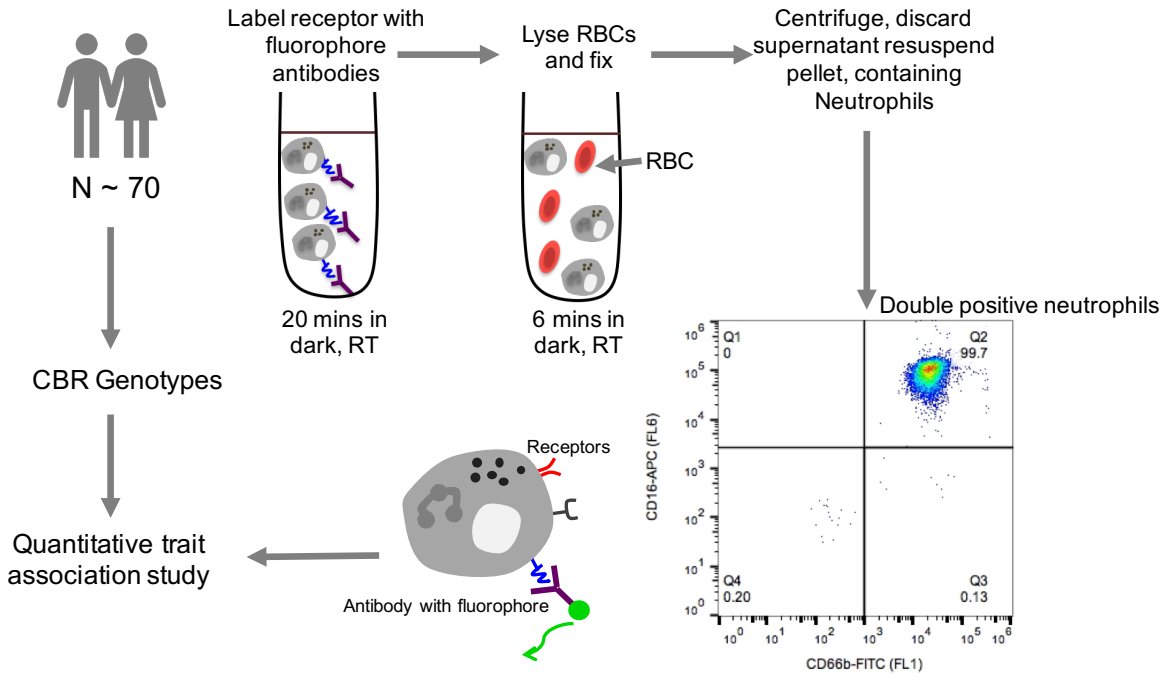


**Figure 4.6: Project and experimental design**
Overview of study and experimental design. Peripheral blood samples from 70 individuals were collected and labelled using specific antibodies. Mean fluorescence intensities (MFI) were collected for two receptors for surface expression on the neutrophil population using flow cytometry. These values for each individual were then used as a quantitative trait in genetic association tests to assess whether variation in surface expression was genetically controlled.

### 4.3.2 Two independent genetic signals are associated with GCSFR surface expression levels

Single-variant association tests were performed using GCSFR residualised MFI receptor values across 66 individuals. 14698 variants within the gene and 500 kb upstream and downstream of the start and end gene positions were tested for association with the receptor levels. These variants were pruned using a pairwise $r^2$ threshold of 0.1 to generate a list of independent variants in the region (159). To correct for multiple testing, a stringent Bonferroni correction was used to correct for the number of independent variants tested (0.05/159) resulting in a significance p-value threshold of $3.14 \times 10^{-04}$ for GCSFR. In the *CSF3R* locus, 14 genetic variants reached significant levels of association with surface expression levels of G-CSF receptor (Figure 4.7 and Table 4.5). Four of these variants were low-frequency, the rest were common.

| | rsID | chr 1 Pos (hg19) | EA/OA | EAF | P | Beta, (SE) | EAF NEU | NEU effect (SE) | UKBB P | BP beta (SE) | BP P | $R^2$ Astle | Genic location |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Rare signal* | **rs3917912** | 36947936 | T/C | 0.02 | $1.53 \times 10^{-05}$ | -2.48 (0.53) | 0.01 | 0.16 (0.02) | $6.04 \times 10^{-19}$ | - | - | - | Intron 1 (CSF3R-008) |
| | **rs3917914** | 36947888 | A/G | 0.02 | $1.62 \times 10^{-05}$ | -2.37 (0.51) | 0.01 | 0.16 (0.02) | $9.54 \times 10^{-22}$ | - | - | 0.99 | Intron 1 (CSF3R-008) |
| | rs116668546 | 36952851 | A/G | 0.02 | $1.69 \times 10^{-05}$ | -2.32 (0.50) | - | - | - | - | - | 0.71 | Upstream (-) |
| | rs3917922 | 36945713 | T/C | 0.02 | $1.70 \times 10^{-05}$ | -2.32 (0.50) | 0.01 | 0.16 (0.02) | $9.32 \times 10^{-21}$ | - | - | 0.85 | Intron 2 (CSF3R-008) |
| | | | | | | | | | | | | | |
| *Common signal* | **rs3917924** | 36945653 | G/A | 0.56 | $1.92 \times 10^{-05}$ | -0.68 (0.15) | 0.61 | 0.03 (0.004) | $2.49 \times 10^{-20}$ | 1.10 | $5.51 \times 10^{-35}$ | - | Intron 2 (CSF3R-008) |
| | rs3917931 | 36944054 | C/T | 0.56 | $1.93 \times 10^{-05}$ | -0.68 (0.15) | 0.61 | 0.03 (0.004) | $1.96 \times 10^{-20}$ | 1.12 | $9.98 \times 10^{-36}$ | 1.00 | Intron 3 (Intron 1 of CSF3R-204) |
| | rs3832027 | 36945307 | AG/A | 0.56 | $1.96 \times 10^{-05}$ | -0.68 (0.15) | 0.61 | 0.03 (0.004) | $2.34 \times 10^{-20}$ | - | - | 1.00 | Intron 2 (CSF3R-008) |
| | rs3917925 | 36945559 | G/A | 0.56 | $1.97 \times 10^{-05}$ | -0.68 (0.15) | 0.61 | 0.03 (0.004) | $2.42 \times 10^{-20}$ | 1.12 | $9.98 \times 10^{-36}$ | 0.99 | Intron 2 (CSF3-008) |
| | **rs3917932** | 36943916 | C/G | 0.39 | $2.32 \times 10^{-05}$ | -0.71 (0.16) | 0.42 | 0.05 (0.004) | $2.06 \times 10^{-39}$ | - | - | 0.47 | Intron 3 (Intron 1 of CSF3R-204) |
| | rs199833813 | 36954487 | AT/A | 0.59 | $3.01 \times 10^{-05}$ | -0.67 (0.15) | 0.65 | 0.03 (0.004) | $4.41 \times 10^{-15}$ | - | - | 0.78 | Upstream (-) |
| | rs6667127 | 36957501 | C/T | 0.59 | $3.43 \times 10^{-05}$ | -0.67 (0.15) | 0.65 | 0.03 (0.004) | $4.69 \times 10^{-15}$ | 0.97 | $1.24 \times 10^{-19}$ | 0.78 | Upstream (-) |
| | rs955115 | 36858145 | C/A | 0.77 | $1.53e \times 10^{-04}$ | -0.93 (0.23) | - | - | - | 0.59 | $4.95 \times 10^{-06}$ | 0.11 | Downstream (-) |
| | rs3917933 | 36943655 | G/A | 0.55 | $1.62e \times 10^{-04}$ | -0.61 (0.15) | 0.61 | 0.03 (0.004) | $1.3 \times 10^{-19}$ | 1.12 | $9.98 \times 10^{-36}$ | 1.00 | Intron 3 (Intron 1 of CSF3R-204) |
| | rs11295216 | 36947906 | C/CG | 0.53 | $2.86 \times 10^{-04}$ | -0.65 (0.17) | - | - | - | - | - | 0.49 | Intron 3 (CSF3R-008) |

**Table 4.5: 14 Significant variants associated with the G-CSF receptor surface levels**
Summary statistics from the GCSFR surface level and Astle *et al.* (2016) neutrophil count (NEU). Bold SNPs are the lead GCSFR MFI or NEU SNPs. EAF, effect allele frequency derived from MAF calculated using 66 donors. Standardised beta and standard error (SE) is given. EAF NEU was calculated for 173,480 individuals in Astle *et al.* (2016) and effect in SD of the trait. Location is relative to the most abundant transcript, CSF3R-204 or the second most abundant truncated transcript, CSF3R-008. EA = effect allele. OA= other allele. BP EA and BP direction relate to the significant blueprint exon effect (corrected for local SNPs as qvalue). Variants not tested in the study are shown by missing values (-). LD in $r^2$ between the variant and lead variant per common and rare signal respectively was calculated using the Astle *et al.* (2016) cohort.

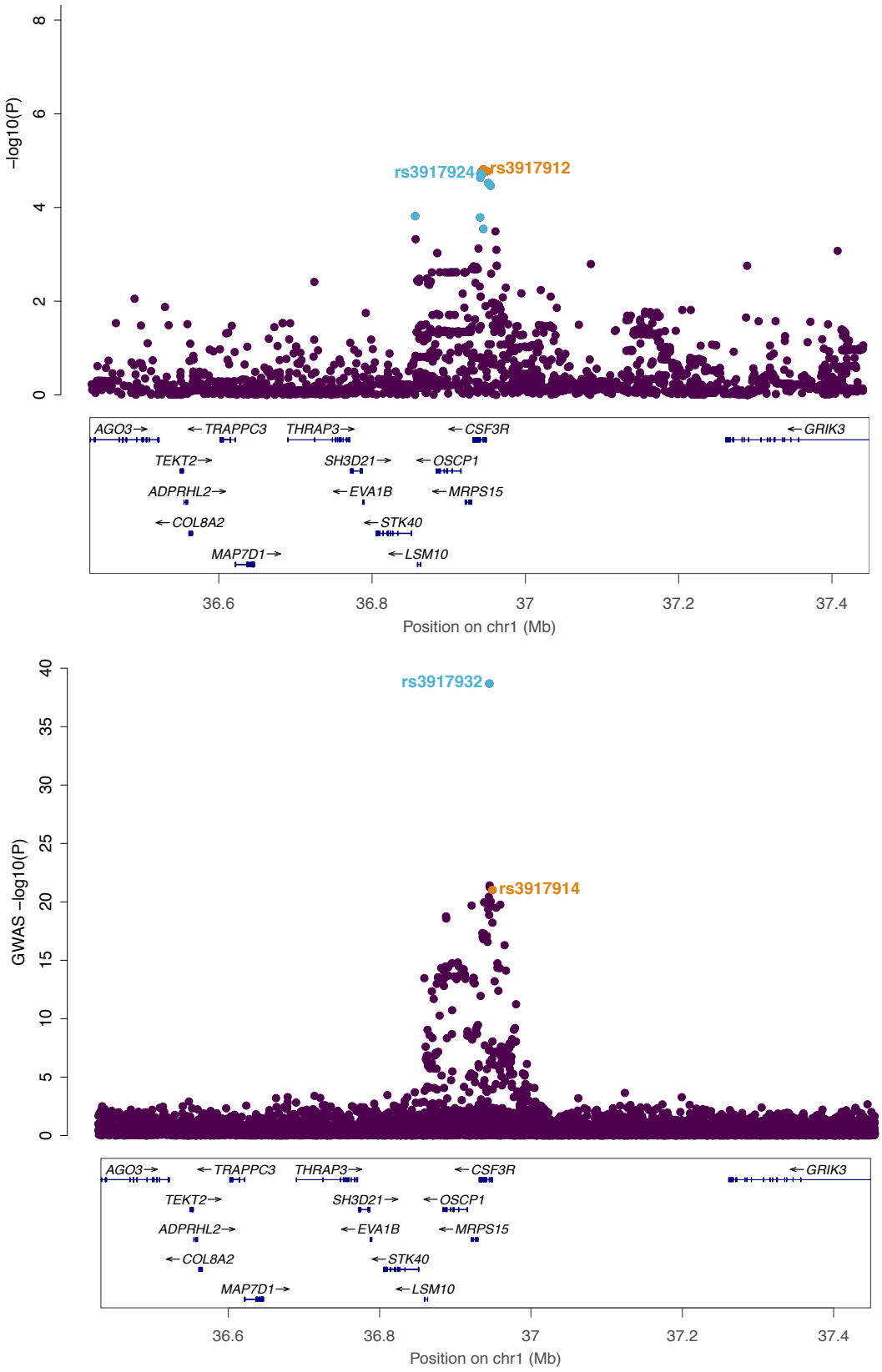


**Figure 4.7: G-CSF surface expression association results**
Regional association plot of the *CSF3R* locus. The top panel shows associations with the GCSFR surface expression levels. Variants reaching significance are highlighted in blue (common) or orange (rare) (Table 4.5). Lead common and rare SNPs are designated with the rsID. The regional association plot for the neutrophil count association is repeated from Figure 4.10 for comparison.

I carried out conditional analysis including the lead rare and common SNPs as covariates in an association model to test if these were independent signals. The beta and p value of rs3917912 remained similar to the univariate model and when regressing out the common, rs3917924 signal in the association model (univariate beta = -2.48 (SE = 0.53), conditional beta = -2.10 (SE = 0.48), Table 4.6). Similarly, when testing for association of rs3917924 with GCSFR MFI while conditioning on rs3917912, the common SNP remained significant (Table 4.6). After conditioning on both rs3917912 and rs3917924, no SNPs remained significant. Theses analyses provided evidence that the GCSFR surface expression association signal consists of two independent signals of different frequencies. Within the cohort of 66, three individuals carried the heterozygous genotype (T/C) for the low-frequency SNP, rs3917912. For clarity, I will refer to the significant signals that are described by lead SNPs, rs3917924 and rs3917912 as the common and rare signals respectively.

| | Model | EA/OA | EAF | Beta (SE) | P value |
|---|---|---|---|---|---|
| **rs3917912 (rare)** | Univariate | T/C | 0.03 | -2.48 (0.53) | $1.53 \times 10^{-05}$ |
| | Conditional (rs3917924) | T/C | 0.03 | -2.10 (0.48) | $4.42 \times 10^{-05}$ |
| **rs3917924 (common)** | Univariate | G/A | 0.57 | -0.68 (0.15) | $1.92 \times 10^{-05}$ |
| | Conditional (rs3917912) | G/A | 0.57 | -0.57 (0.13) | $5.53 \times 10^{-05}$ |

**Table 4.6: Conditional analysis demonstrated the *CSF3R* locus contains two independent signals**
Results of the two lead SNPs (common and rare) for association with GCSFR surface levels are given. The univariate beta and p values are from initial genetic association tests. Association tests were repeated but with conditioning on the respective top SNPs i.e. condition on rare and test for common association and vice versa. The conditional beta and p values are given and show that in both cases the remaining signal is still significant, confirming the common and rare signals are independent. The association analysis was also performed conditioning on both common and rare SNPs and no variants remained significantly associated using the before mentioned threshold (data not shown).

### 4.3.3 The relationship between GCSFR MFI and neutrophil count

The neutrophil count association signal in the *CSF3R* locus is described by two independent signals; a common signal with lead SNP rs3917932 (EA = C, EAF = 0.42, beta = 0.048, SE = 3.63e-03, p value = 2.06 x $10^{-39}$) and a rare signal with lead SNP rs3917914 (EA = A, beta = 0.16, SE = 1.62 x $10^{-02}$, p value = 9.54 x $10^{-22}$) (Table 4.1) (Astle et al., 2016).

The lead common neutrophil count variant, rs3917932 is significant in the GCSFR data, with a slightly less significant p value but larger beta than the GCSFR lead SNP, rs3917924 (Table 4.5). These two SNPs are practically indistinguishable in the GCSFR data. In the Astle *et al.* (2016) study of neutrophil count, rs3917932 is 20 orders of magnitude more significant than rs3917924 (2.49 x $10^{-39}$, 2.49 x $10^{-20}$ respectively). There was also a reasonable correlation between the two SNPs; using HaploReg v4.1 1000 Genomes the $r^2$ between rs3917932 and rs3917924 is 0.46 (Ward and Kellis, 2012). I confirmed that rs3917932 and rs3917924 were not independently associated with GCSFR MFI using conditional analysis (data not shown). Combined, this was evidence that the neutrophil count and GCSFR surface expression common signal can be explained by the same causal variant(s), where the Astle *et al.* (2016) study has greater power to distinguish the putative causal variant (rs3917932).

Both rare SNPs, rs3917914 (neutrophil count lead) and rs3917912 (GCSFR surface expression lead) are significant in the GCSFR data (p value 1.62 x $10^{-05}$ and 1.53 x $10^{-05}$ respectively). The p values are similar in the neutrophil count association also (9.54 x $10^{-22}$ and 6.04 x $10^{-19}$ respectively) (Table 4.5). These variants were perfectly correlated with $r^2$ of 1 (1000G). Therefore, it is highly likely that the neutrophil count and GCSFR surface expression rare signal can be explained by the same causal variant(s).

I observed an unexpected inverse relationship between the associations of both variants with GCSFR surface expression and neutrophil count. For both common and rare signals, the GCSFR MFI-increasing allele was associated with decreased neutrophil count (Figure 4.8). I had predicted, based on the role of GCSFR in stimulating neutrophil differentiation, that there would exist a positive relationship between surface expression and neutrophil count, where more receptor would result in increased stimulation and higher neutrophil numbers (Panopoulos and Watowich, 2008, Mehta et al., 2015). My results suggest a more complex relationship where rather than the total surface expression of the receptor, possible structural or functional changes could link receptor activity to neutrophil count.

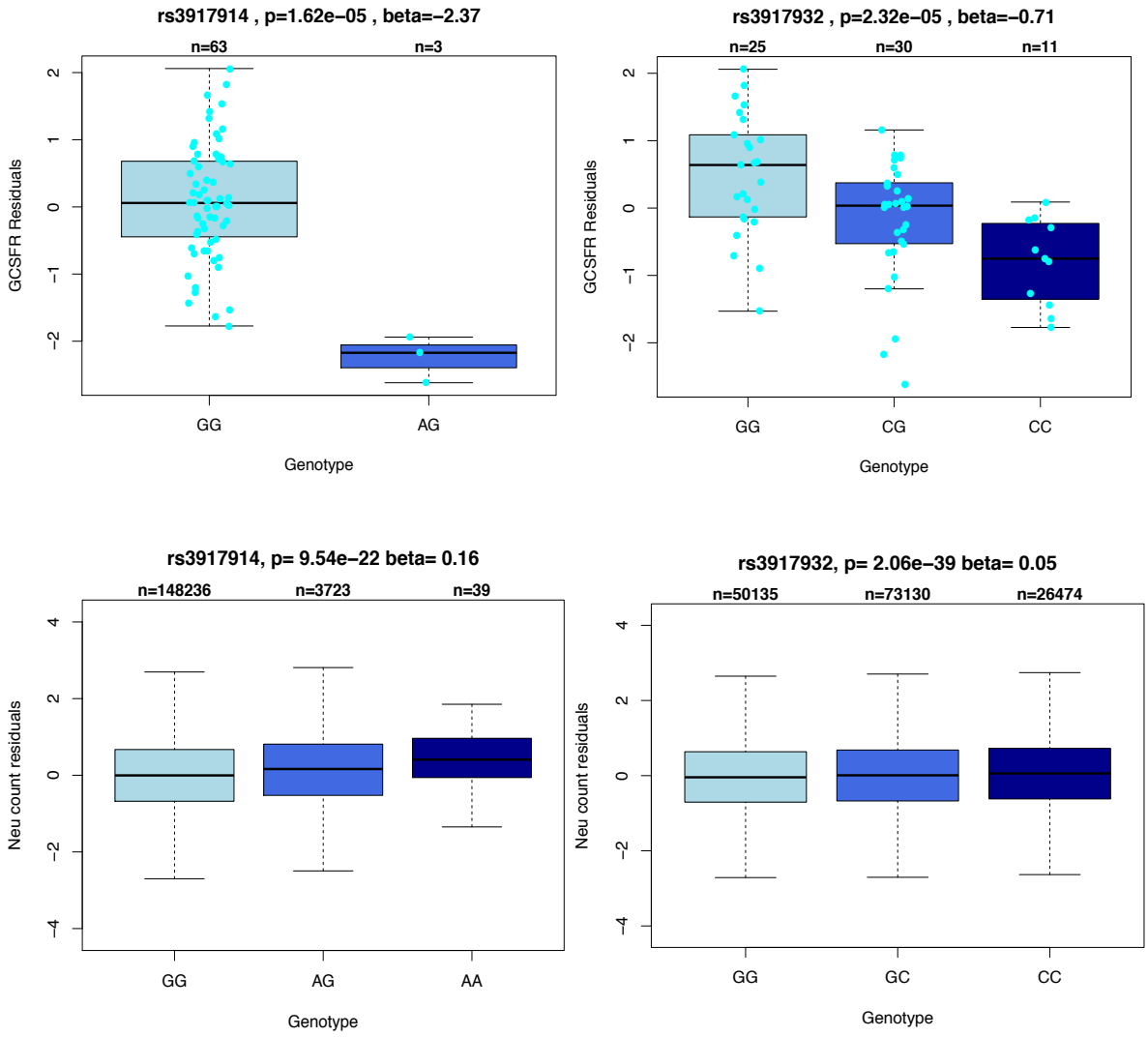


**Figure 4.8: Directions of effect of the two independent genetic signals associated with GCSFR surface expression and neutrophil count**
Trait residuals are stratified by genotype of the rare (left) and common (right) SNPs. The top panel shows the GCSFR MFI residuals and the bottom panel shows the Astle *et al.* (2016) neutrophil count residuals. GCSFR trait values for each individual in the study (N = 66) are shown but not for neutrophil count (bottom panel) as the number of individuals tested was too high.

### 4.3.4 Evidence of high molecular functionality in the *CSF3R* genic locus

I next investigated whether using molecular and epigenomic data could aid the interpretation of the functional consequences at this locus and explain the inverse effect. In this study, the most significant SNPs were located in the introns of the *CSF3R* gene. It is known that enhancers, which are regulatory elements that control transcription can be located within introns in addition to upstream and downstream of genes (Pennacchio et al., 2013). Therefore, I investigated and reanalysed several public genomic resources and unpublished data to further characterise the potential molecular and epigenomic characteristics of the *CSF3R* locus.

I found that most GCSFR-associated variants intersected with ROADMAP/ENCODE/BLUEPRINT primary neutrophil chromatin regulatory state data (Supplementary Table 4.1) (Ward and Kellis, 2012, Chen et al., 2016a, Carrillo-de-Santa-Pau et al., 2017). The common and rare SNPs intersected with active TSS (H3K4me3) chromatin states (rs3917924, rs3917912/4) and rs3917932 intersected with an active enhancer state (H3K27ac, H3K4me1). Using a combination of TF ChIP-seq data from primary neutrophils as well as the HL60 differentiated and undifferentiated cell line models described in Chapter 2, I identified that the region also showed evidence of PU.1, C/EBPβ, C/EBPε and the enhancer-associated co-factor, P300 (Supplementary Table 4.1). Examples of binding of some factors in the *CSF3R* locus is shown in Figure 4.9 with representative signal peaks from one BLUEPRINT individual.

The intersection with epigenomic data demonstrated that this was a complex genic region with high molecular functionality, also confirmed by the varied transcript architecture, with 18 different annotated transcripts in GENCODE v15 (Figures 4.9-4.10). While informative for general regional functionality, intersections are prone to chance overlaps and it is difficult to identify a potentially causal SNP or mechanism based solely on the overlap of epigenomic functionality. This is particularly the case, if multiple variants overlap the same or different peaks. In comparison, molecular quantitative trait loci (QTLs) afford the opportunity to dissect individual variant associations with specific molecular marks with statistical confidence. I next investigated such evidence, described below.

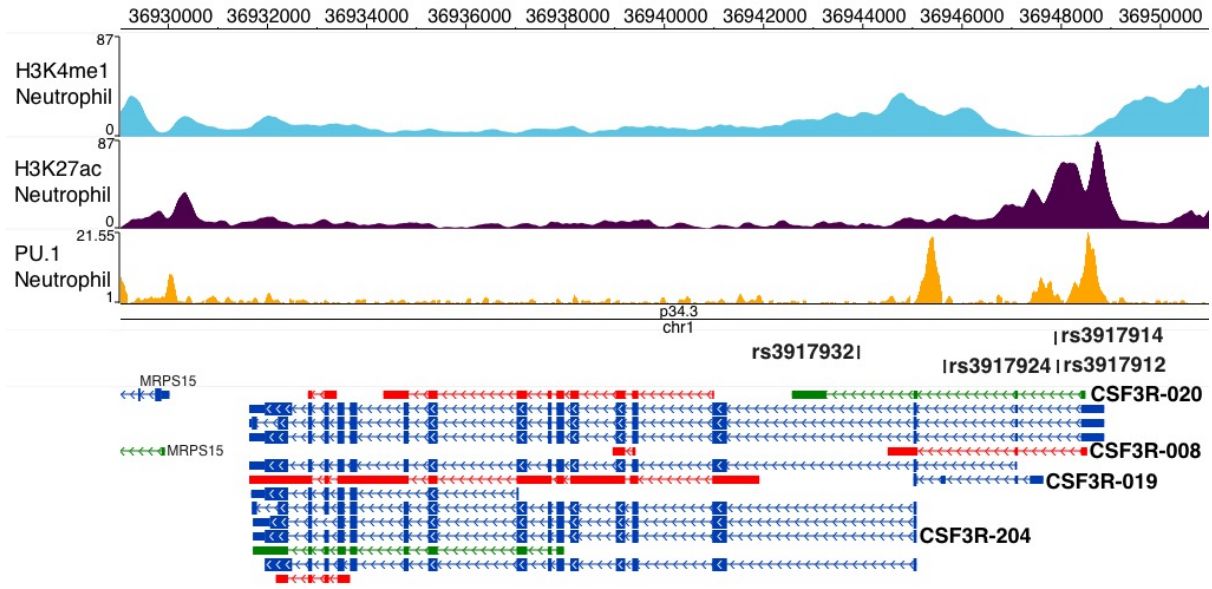


**Figure 4.9 Epigenetic and transcript architecture of the *CSF3R* locus**
Regional genome plot of the *CSF3R* locus. H3K4me1, H3K27ac and PU.1 peaks are shown for primary mature neutrophils. The lead neutrophil count and GCSFR SNPs are shown. All *CSF3R* transcripts are given with the key transcripts discussed in this thesis labelled. Red transcripts are those with an issue such as retained intron, green transcripts are processed transcripts and blue protein-coding as predicted by GENCODE v17, the earliest version available in the Washu browser (Zhou et al., 2011).

### 4.3.5 Molecular QTL effects of the common GCSFR MFI association

I accessed the Blueprint consortium human variation panel dataset (Chen et al., 2016a), which provides molecular QTLs and allele-specific events in $CD16^{+}$ neutrophils (Chapter 2 Materials and Methods). The different molecular traits and analytical approaches are summarised in Figure 2.2. In Chapter 2, I focused on investigating gene expression, histone modification and percent-splice in QTLs that were measured in up to 197 individuals. Here, I also assessed the transcript isoform ratio QTLs as well as allele-specific gene and histone effects. Allele-specific approaches evaluate differences in expression or modification signals that occur within an individual heterozygous at the locus of interest and as a result can increase the power to detect genetic effects (Chen et al., 2016a). Transcript isoform effects also indicate splicing events that result in a differential ratio between two transcripts (summarised in Figure 1.1). As part of the main study, expression levels were quantified across all known transcripts normalised by the length of the transcript. The ratio of the two isoforms that exhibited the highest expression change and showed symmetrical changes were then evaluated for significant QTLs (Chen et al., 2016a). The size of the isoform effect is quantified as the maximum difference (MD) in relative expression between SNP genotype groups where a 20% shift in the relative expression of one transcript across genotypes is

reflected by an MD of 0.2 (Chen et al., 2016a). I was also able to access additional expression quantifications that were not part of the main analysis including exon and splicing junction QTLs, which were variants associated with different levels of RNA-seq reads specifically at exons and splicing junctions.

SNPs with MAF of less than 1% (allele count = 4), were not included in the blueprint study. In addition, the index neutrophil count SNP, rs3917932, was not included in the final variant set. Therefore, from this point, I investigated effects of the common GCSFR signal and evaluated evidence that the GCSFR index SNP (rs3917924) was associated with specific molecular QTLs or phenotypes.

rs3917924 was significantly associated with the allele-specific expression of the whole *CSF3R* gene (EA = G, p value = $5.59 \times 10^{-33}$, beta = 0.10) (Chen et al., 2016a). Aligning the effects showed that reduced GCSFR surface expression corresponds with a small increase in GCSFR allele-specific gene expression and increased neutrophil count. This suggests that the reduced surface expression effect is not due to a reduction in the expression of full-length gene.

There were no other associations of rs3917924 in the QTLs assessed as part of the main BLUEPRINT Chen *et al.* (2016a) study. However, I identified highly significant exon and splicing junction associations with rs3917924 (EA = G, p value = $1.09 \times 10^{-37}$, beta = 1.10, Table 4.5). Interestingly, the exon corresponding to this effect was the third exon located in a truncated transcript, CSF3R-019, not the transcript encoding the full-length receptor (Figure 4.10). rs3917924 is also located within exon 3 of CSF3R-019 (36,945,681-36,945,588). I next investigated whether these effects could reflect regulation at the level of individual transcripts.

### 4.3.6 Differential *CSF3R* transcript expression associated with rs3917924

Figure 4.10 shows there are 18 *CSF3R* transcripts of varying lengths included in the GENCODE v15 annotation (used in the BLUEPRINT study (Chen et al., 2016a)). Figure 4.10 also shows how each exon contributes to the GCSFR protein domain structure, where the canonical third exon contributes to the N-terminal start of the protein. CSF3R-019 is a truncated transcript with a different third exon compared to the longer protein-coding transcripts such as CSF3R-024.

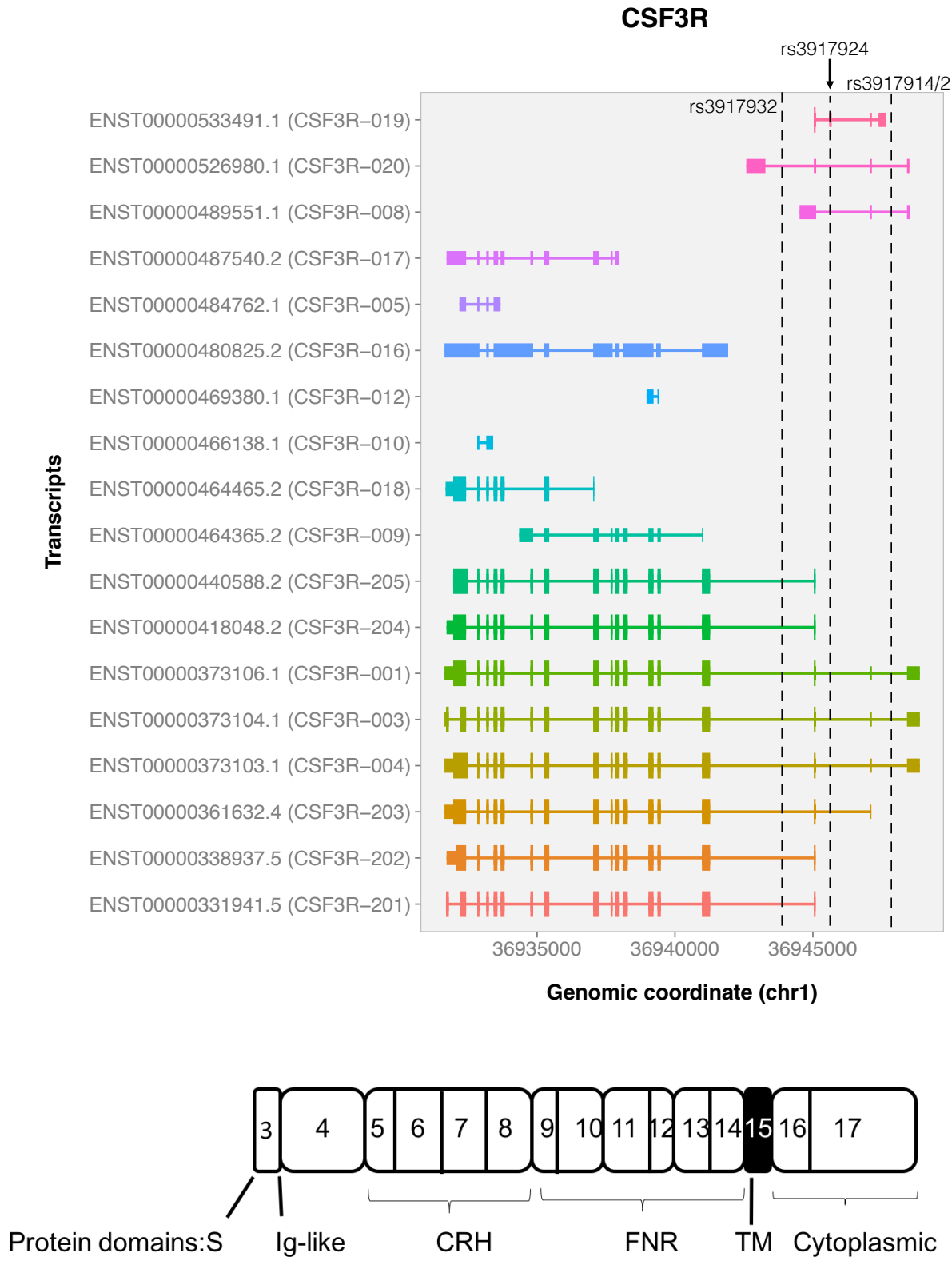


**Figure 4.10: GCSFR transcript and protein structure**
Upper panel: All possible *CSF3R* transcripts (GENCODE v15) shown with lead SNP positions. The *CSF3R* gene is located on the reverse strand (right to left). Lower panel: Protein domains of the GCSFR protein with respect to contributing exons (exons 3-17 of CSF3R-204, which is most abundantly expressed and generates a functional protein). Different protein domains include an Ig-like domain, a cytokine receptor homologous domain (CRH), three fibronectin type III domains (FNR), a transmembrane domain (TM) and a cytoplasmic domain (Seto et al., 1992). S, the signal peptide, is required for direction of membrane proteins to the cell surface. Figure adapted from (Seto et al., 1992).

I evaluated the relative abundance expression of these 18 transcripts using the transcript expression quantifications from the BLUEPRINT project (Figure 4.11). The CSF3R-204 transcript is the most abundant in neutrophils, followed by the 3' truncated transcripts, CSF3R-008 and CSF3R-020.

ENSEMBL (GRCh37 release 75 2014) predicts that the CSF3R-204 transcript encodes a protein product of 836 amino acids. CSF3R-008 retains an intron and is not predicted to be protein-coding (Figure 4.10). CSF3R-020 is also not predicted to be protein-coding, instead a processed non-coding transcript. In neutrophils and monocytes, the protein-coding CSF3R-204 transcript was the most abundant, and GCSFR is known to be expressed on the surface (higher in neutrophils). In contrast, no protein-coding transcripts are expressed in T cells and the highest expressed transcripts are the truncated CSF3R-020 and CSF3R-008 (Supplementary Figure 4.3) Interestingly, the receptor is not expressed on the surface of lymphocytes, suggesting the switch to higher truncated abundance may reflect a regulatory mechanism generating cell-type specific protein abundance (Christopher et al., 2011).

I next investigated whether there was evidence that transcript level was affected by genotype of the common signal, possibly explaining the significant differential exon expression. I observed a visible change in expression level of the second most abundant transcripts, CSF3R-008 and CSF3R-020 (Figure 4.11). The ratio of these two transcripts was tested as part of the Chen *et al.* (2016) study based on the criteria I explained above. A cautioned, conservative approach for evaluating the significance of associations was used in this study; despite not meeting the stringent significant threshold corrected for multiple testing, there was some evidence of an effect. rs3917924 just missed the significance level for a single test ($p < 0.05$) for association with the CSF3R-008/CSF3R-020 ratio (p value = 0.060, MD = 0.033). Other highly correlated SNPs that I also identified as significantly associated with GCSFR MFI, perhaps showed evidence of a small effect (rs3917933 p = 0.045, rs3917931, p = 0.035).

I aimed to further resolve the transcript based effects, rather than evaluating effects on transcript ratios, I investigated genetic effects on the absolute expression of each *CSF3R* transcript expressed in fragments per kilobase of transcript per million fragments sequenced (FPKM). Using a standard linear regression approach on inverse normalised FPKM, I identified significant associations with rs3917924 and the three truncated transcripts, CSF3R-020, CSF3R-008 and CSF3R-019 (Figure 4.12). The association with CSF3R-019 (containing the exon identified as a significant QTL) was the most significant (p value = $6.002 \times 10^{-44}$, beta = 1.155, SE = 0.063). This association was significant in monocytes, but reduced compared to neutrophils (p value = $9.148 \times 10^{-05}$, beta = 0.399, SE = 0.100,

Supplementary Figure 4.4). No significant association was found with the expression level of the protein-coding transcript, CSF3R-204 in neutrophils (p value = 0.319) or in monocytes (Supplementary Figure 4.4).

In ENSEMBL, the CSF3R-019 transcript is predicted to generate a very short protein of 21 amino acids in length. Further, in the human cell atlas database, CSF3R-019 is not predicted to contain a transmembrane region and may represent a secreted protein (predicted by SPOCTOPUS (www.proteinatlas.org) (Uhlen et al., 2015, Viklund et al., 2008). The relative abundance shows low expression of CSF3R-019 (Figure 4.11). However, the absolute FPKM ranged from 1.70 to 20.08 across individuals with a median expression of 9.37 FPKM. Other CSF3R transcripts had even lower expression, some with 9 FPKM in neutrophils (CSF3R-001, CSF3R-003, CSF3R-004, CSF3R-009, CSF3R-010). The FPKM cut-off used for transcripts in the BLUEPRINT study was 0.1, suggesting that CSF3R-019 may be moderately expressed, but at much lower levels than the dominant CSF3R-204 transcript (Chen et al., 2016a).

In summary, reduced (allele-specific) gene-expression corresponds to increased truncated CSF3R-019 expression levels, reduced surface GCSFR expression and increased neutrophil count. Whether there is a regulatory role at the transcript level for the three differential expressed transcripts or at the protein-level with respect to the predicted truncated protein from CSF3R-019 requires further functional investigation.

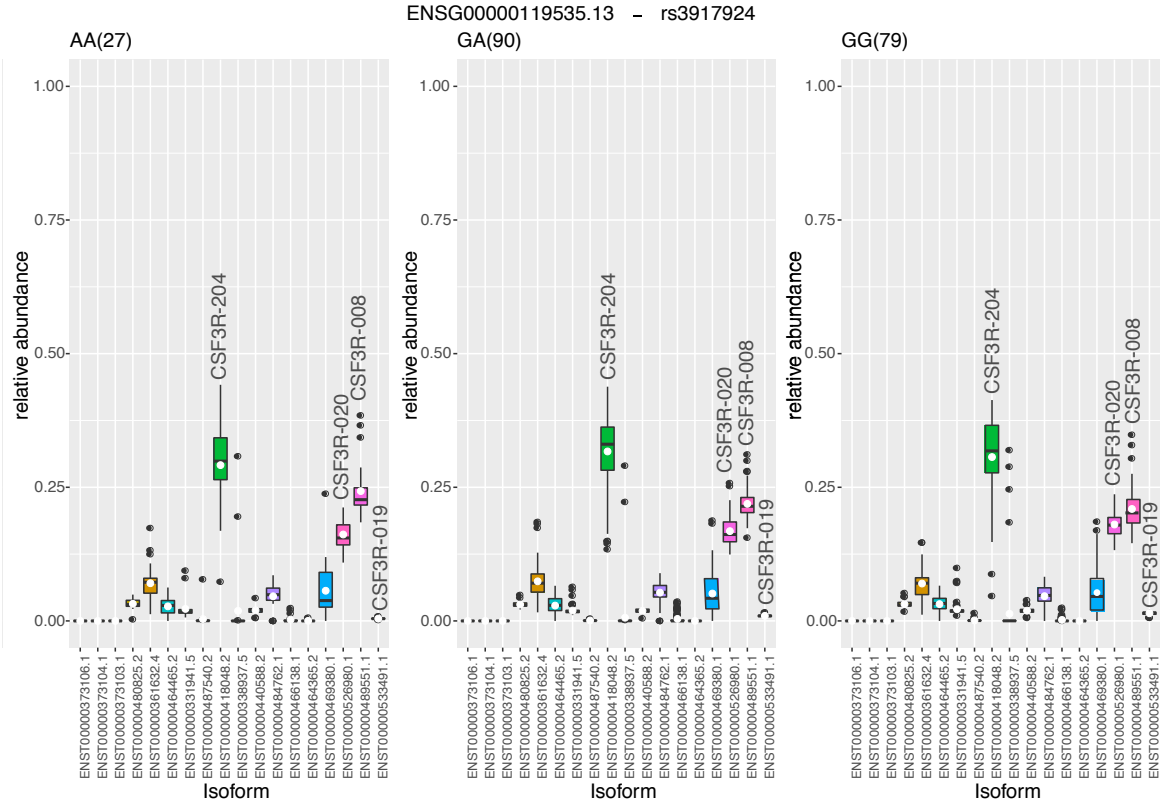


**Figure 4.11: Relative abundance of all *CSF3R* transcripts, stratified by genotype of rs3917924, may suggest a marginal genetic splicing effect**

Relative *CSF3R* transcript abundances stratified by the genotype of the common index GCSFR SNP identified in the BLUEPRINT study (Chen et al., 2016a). This figure demonstrates that there may be a difference in abundance of some transcripts across genotypes. This figure is adapted from the original produced by Diego Garrido Martin using data analysed as part of the BLUEPRINT consortium (Chen et al., 2016a).

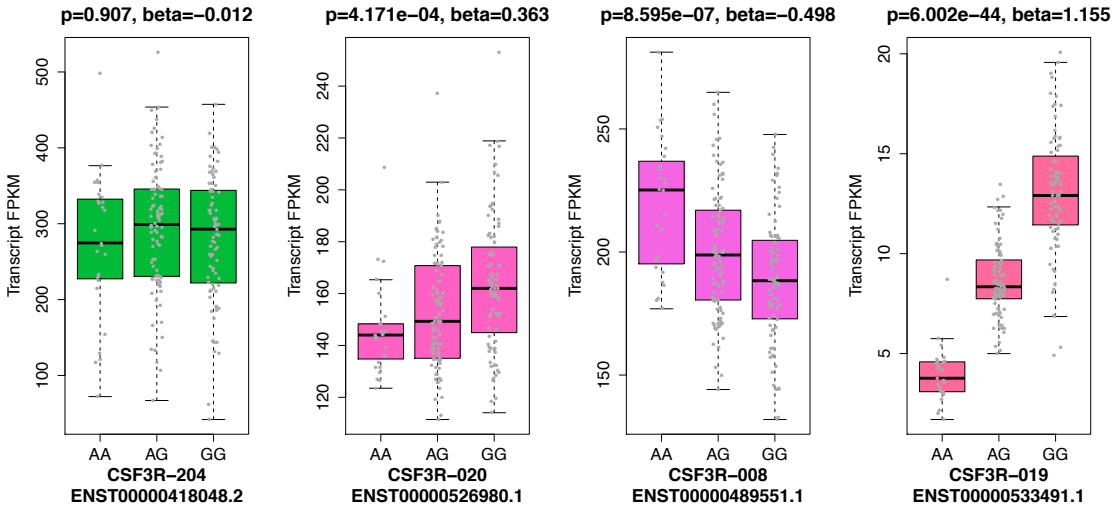


**Figure 4.12: Association of the common SNP, rs3917924, with *CSF3R* transcript expression levels**

Absolute expression of the transcripts is tested for association with the lead GCSFR surface expression level SNP, rs3917924. Transcripts where a significant association with rs3917924 was identified are shown stratified by genotype. Transcript expression is measured in FPKM (expected fragments per kilobase of transcript per million fragments sequenced) and was estimated as part of the BLUEPRINT consortium using transcripts defined by the Cufflinks tool, without *de novo* assembly (Chen et al., 2016a). Regression was performed using inverse normalised FPKM values. The boxplot colour matches that of the corresponding transcript in Figure 4.10 and Figure 4.11.

### 4.3.7 Investigation of the rs4760, the PLAUR missense SNP

I also investigated the surface expression of a second receptor, PLAUR, which has a role in neutrophil function (Section 4.1). I tested all variants for association with PLAUR MFI that were within the length of the gene and 500 kb upstream and downstream (N = 16155) including the predicted neutrophil count causal missense SNP, rs4760. The threshold I used for evaluating significantly associated variants was $2.2 \times 10^{-04}$, which was based on correcting for 218 pruned, independent variants in the region (pairwise $r^2$ threshold of 0.1).

No significant associations with the PLAUR receptor level were identified (Figure 4.13). The missense SNP, rs4760, did not meet the stringent significance threshold (p value = 0.01). This could be suggestive evidence of association using the p value threshold for a single test ($p < 0.05$), but confirmation may require further testing.

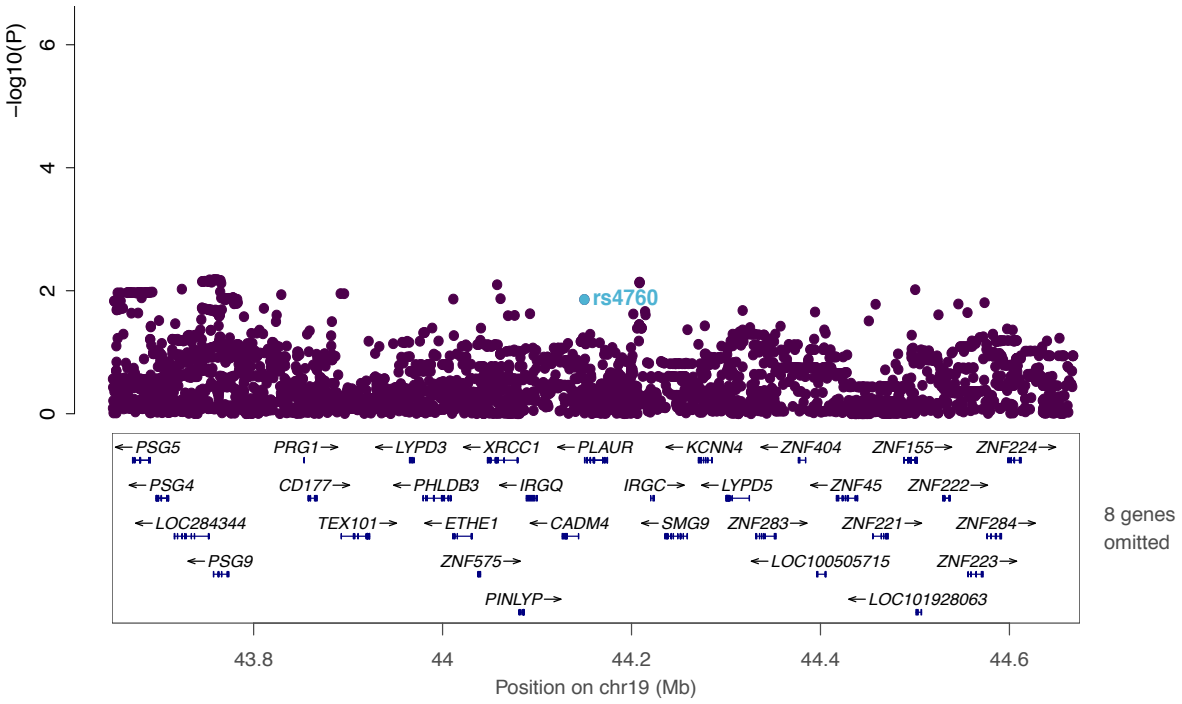


**Figure 4.13: PLAUR surface expression association results**
Regional association plot with PLAUR receptor surface expression. Variants within the gene and 500kb upstream and downstream of the index SNP are shown with the $-\log_{10}$(p-value) of the association with receptor level (y-axis). There are many genes in this region, with the plot centred around the PLAUR gene. No associations were found to be significant in this region after correction for multiple testing.

Given the complexity of the relationship between GCSFR surface expression levels and neutrophil count, I explored other possible functionality of rs4760 in additional unpublished neutrophil-relevant datasets. I observed a significant association between rs4760 with two additional Sysmex haematological analyser traits, NE-FSC (EA = G, p value = $8.6 \times 10^{-15}$, beta = 0.075, SE = 0.010) and NE-SFL (EA = G, p value $2.9 \times 10^{-17}$, beta = 0.077, SE = 0.010) that had been analysed by Parsa Akbari using the INTERVAL cohort of approximately 50,000 individuals (unpublished observations). NE-FSC is a forward scatter parameter that is used as an estimated of neutrophil size. NE-SFL is neutrophil side fluorescence, which increases with a higher amount of cellular DNA and RNA (Buoro et al., 2016, Sysmex Corporation, 2010-2012). These two associations could be related as it is possible that a larger cell could contain a higher amount of nucleic acid. None of the significant SNPs associated with GCSFR surface levels were found to be significantly associated with neutrophil cell size or other additional granularity traits. Interestingly, when I tested for an association of rs4760 and NE-FSC that was also measured using a Sysmex haematology analyser within this recall study (N = 65), the association was not significant (p value = 0.508). The significant association in the larger cohort perhaps suggests that this recall study was limited in power to detect associations of rs4760 with NE-FSC and possibly also PLAUR MFI. Using the pwr R package, I estimated that with a cohort size of 100 and p value threshold of $2.2 \times 10^{-04}$, the study would be powered to detect variants of similar frequency (rs4760 MAF = 0.16) with a beta > 1, confirming my functional cohort was not powered to detect small effect sizes of associations (Champely, 2012).

The effects of these associations demonstrate rs4760 causes a decrease in neutrophil count (EA = G, p value = $1.428 \times 10^{-68}$, beta = $-8.615 \times 10^{-02}$, SE = $4.923 \times 10^{-03}$) and an increase in neutrophil cell size (EA = G, p value = $8.6 \times 10^{-15}$, beta = 0.075, SE = 0.010). The association with receptor expression from this study demonstrates a decrease in PLAUR surface expression, although this misses the stringent significance threshold applied (p value = 0.01, beta = -0.59, SE = 0.233). Given that little is known about the role of neutrophil size in development or function of neutrophils, this inverse relationship would need further investigation.

## 4.4 Discussion

In this chapter, I used a recall-by-genotype (RbG) design to test a hypothesis that significant neutrophil count variants located in protein receptor genes were also associated with the surface expression of those receptors. I demonstrated in Chapter 3 how performing a QTL or GWAS study using neutrophil functional phenotypes is technically challenging. RbG studies provide an alternative but highly efficient design to test specific hypotheses based on previous biological and genetic observations, thus requiring a smaller sample size.

Using a cohort of 66 healthy individuals, I identified common and rare signals located in the *CSF3R* locus that are significantly associated with the level of GCSFR neutrophil surface expression. Both signals were also independently associated with neutrophil count from a large GWAS study in 173,480 individuals (Astle et al., 2016). I identified an inverse relationship between these two traits; a decrease in GCSFR at the surface corresponded to an increase in neutrophil count. I also observed other molecular effects, including an increase in expression of the truncated transcript, CSF3R-019.

The opposing direction of effects is not initially intuitive as signalling through the G-CSF receptor is known to increase neutrophil numbers and promote neutrophil differentiation (Lord et al., 1989, Lord et al., 1991, Panopoulos and Watowich, 2008). Naively, a positive relationship between level of GCSFR surface expression and neutrophil numbers could be expected. In this scenario, more signalling through the receptor could lead to a higher number of neutrophils during both differentiation and in the increased release of mature neutrophils from the bone marrow. However, this prediction does not take into account other more complex possibilities. Upon activation, GCSFR is internalised into the cell, a process which is dependent on the C-terminal internalisation motif and functions to regulate signalling preventing over-activation (Kindwall-Keller et al., 2008). Increased surface expression of GCSFR could, therefore, be a result of reduced internalisation due to less activation, perhaps reflecting a difference in protein functionality due to genotype (not tested here). A similar recall-by-genotype study where neutrophils are stimulated by G-CSF overnight, followed by measurement of the phosphorylation of downstream signalling targets such as STAT3 could help in assessing whether GCSFR activation or functionality is altered due to genotype. STAT3 is necessary for the increased neutrophil numbers and maturation during emergency granulopoiesis in response to G-CSF (Zhang et al., 2010). Experiments testing neutrophil responses to stimuli (as described in Chapter 3), could also indicate if despite lower surface receptor levels, the receptor is more sensitive to stimulation in individuals carrying the receptor lowering allele. For example, G-CSF has been shown to prime fMLP-dependent ROS production (Yuo et al., 1990, Khwaja et al., 1992).

Alternatively, GCSFR surface expression could have important ramifications for neutrophil count at the progenitor stage rather than the mature circulating neutrophils as tested here. To investigate this further, $CD34^+$ progenitor cells could be collected from donors and differentiated *in vitro* allowing assessment of GCSFR surface expression at earlier and later stages of neutrophil development.

From this study, it is unclear whether the truncated transcript, CSF3R-019, also plays a regulatory role given the strong association of the common signal with increased expression of this transcript. CSF3R-019 is predicted to generate a short protein of 21 amino acids that is missing a transmembrane domain and eventually secreted (Uhlen et al., 2015, Viklund et al., 2008). Predicted proteins generated from other *CSF3R* transcripts are associated with GO terms such as receptor activity and integral plasma membrane component, but no such terms are associated with the CSF3R-019 transcript. There is precedence for soluble receptor protein forms regulating membrane-bound receptor activity, either in an antagonistic or agonistic manner (Xing et al., 2003). Soluble IL6R (sIL6R) is generated from proteolytic cleavage of the membrane-bound form or alternative splicing (Farahi et al., 2017). sIL6R forms a complex with the ligand, IL6 and activates gp130, in turn leading to increased expression and nuclear translocation of STAT3 (Hawkins et al., 2012, Farahi et al., 2017). The GCSFR Ig-like domain (Figure 4.10) is a close homologue of gp130, located in the N-terminal region and is required for G-CSF binding (Layton et al., 2001, Yorke-Smith et al., 2011). However, all previous experimental evidence for truncated *CSF3R* mRNA and soluble GCSFR (sGCSFR), have been for receptors that are larger than that predicted to be encoded by CSF3R-019, for example, 80 and 85 kDa (Iwasaki et al., 1999, Fukunaga et al., 1990). In addition, CSF3R-019 is expressed to a much lower level than the dominant *CSF3R* transcripts. To test if CSF3R-019 produces a soluble protein, plasma from recalled individuals of different genotypes could be tested for the existence of different soluble forms or qPCR of neutrophil RNA would help assessment of possible transcripts.

Understanding the functional mechanism of the relationship between receptor surface expression and neutrophil count also has the potential for clinical benefit. rs3917924 was previously associated with mobilisation potential and recovery of granulocytes in patients receiving a transplantation of autologous peripheral blood progenitor cells (PBPCT) (Bogunia-Kubik et al., 2012). Peripheral blood progenitor cells (PBPCs) are a source of hematopoietic stem cells that can be used in place of bone marrow for transplantation (Jansen et al., 2002). Mobilisation is the increase in steady-state concentrations of PBPCs by inducing migration of hematopoietic cells into the periphery, which can be supplemented with injection of G-CSF prior to cell collection. Mobilisation also indicates recovery after PBPCT, where recovery is evaluated by the number of granulocytes per μL (Jansen et al., 2002,

Bogunia-Kubik et al., 2012). rs3917924 was associated with higher mobilisation potential and a faster recovery of granulocytes in patients after transplant, but in this study corresponded to a lower GCSFR surface expression (Bogunia-Kubik et al., 2012). The authors state that this effect may be due to the alternative allele (A) resulting in an impaired interaction between G-CSF and its receptor leading to a lower response to G-CSF. This could, therefore, be explained by the evidence listed above, that a higher surface level of the receptor is reflective of a lower activity leading and a lower level of internalisation.

I did not identify any significant associations with the PLAUR receptor, but the neutrophil count lead variant, rs4760 was associated with measures of neutrophil size and nucleic acid content in a larger study (N = 50,000, Parsa Akbari, unpublished). This association was also not significant in my cohort of 65 individuals, suggesting that to further evaluate the relationship between neutrophil count, size and PLAUR MFI, larger sample sizes would be required. The lead SNPs associated with GCSFR MFI were not associated with any additional neutrophil measures studied in this larger GWAS, suggesting that the relationship of receptor surface expression and neutrophil count may be specific to each type or functionality of receptor studied and certainly seems more complex than I initially predicted.

Given the well-known examples of other soluble receptors regulating receptor function, I queried my variants in a GWAS of the human plasma proteome including nearly 3000 protein levels in a cohort of 3,301 individuals (Sun et al., 2017). In this, a soluble form of the PLAUR receptor was studied, but there was no equivalent soluble version of GCSFR included in the study. Plasma PLAUR levels were significantly associated, not with rs4760, but with an independent SNP, rs36229204 (EA = T, OA = C, EAF = 0.038, beta = -0.48, SE = 0.07, p value = $5.2 \times 10^{-13}$) (Sun et al., 2017). The rare SNP, rs36229204 (CEU MAF = 0.038) was not significantly associated with neutrophil count (Astle et al., 2016) and not correlated with rs4760. This suggests these SNPs are two independent associations within the *PLAUR* locus with different functional consequences, indeed soluble PLAUR has been suggested to competitively inhibit PLAU protein binding to the membrane-anchored PLAUR receptor form (Sloand et al., 2008). Two independent genetic regulatory effects on different stages within the same receptor function pathway further demonstrates how complex these receptor functions and their relationship to neutrophils count are.

In conclusion, genetic analyses can aid the development of functional hypotheses, but further experimental investigation is required. Examples have been discussed here, such as investigating the activity of the receptor as a function of genotype, or investigating the relationship between GCSFR receptor expression and the numbers of neutrophil progenitors. Below, I describe ongoing efforts to replicate the GCSFR receptor signal.

## 4.5 Future work: Replication of the GCSFR MFI effect

Replication of the common and rare associations with GCSFR MFI is currently being explored using the Sanquin cohort described in Chapter 3. It was possible to recall only a maximum of 20 individuals from this cohort. I therefore designed the study to maximise the power to replicate the MFI effect. In order to select the donors predicted to have the maximal difference between GCSFR expression levels and to investigate the relationship between common and rare signals, I estimated the haplotypes using the genotypes of the three lead SNPs (rs3917914, rs3917924, rs3917932) using SHAPEIT (Materials and Methods). Both common index SNPs were phased to ascertain whether the effect alleles were located on the same haplotype, which would negate the need to select one of the common SNPs and potentially enable an assessment of the causality between the two common SNPs. One rare SNP was used in the haplotype analysis as there was no difference between heterozygous donors given the perfect correlation between rs3917914 and rs3917912. rs3917914 was selected as all of the genotype probabilities of the heterozygote donors were above 0.9, the threshold implemented in SHAPEIT.

Figure 4.14 shows the GCSFR MFI stratified by haplotype in the original RbG study I performed. A similar decrease in surface GCSFR was associated with haplotypes that carry two copies of the common decreasing alleles (**CG**G/**CG**G) than with haplotypes that carry an additional rare decrease allele (CG**A**/GGG), where the decreasing alleles were in order of rs3917932 (C), rs3917924 (G) and rs3917914 (A). Interestingly, a bigger decrease in GCSFR MFI occurred in individuals with haplotypes that were homozygous for the rs3917932 decreasing alleles (CGG/CGG) than those that were homozygous for the rs3917924 decreasing alleles (CGG/GGG). This confirms the larger observed beta estimate for rs3917932 from the association with GCSFR MFI (rs3917932 beta = -0.71, SE = 0.16, rs3917924 beta = -0.68, SE = 0.15) and is evidence that rs3917932 is likely causal common SNP.

I also estimated the haplotypes in the Sanquin cohort using the genotype data from the same SNPs (Table 4.7). Although missing from the original cohort, within the Sanquin cohort four individuals were also homozygous for both the common lead SNPs, CGG/CGA and heterozygous for the decreasing rare allele (A). Based on my association evidence, this haplotype combination would be associated with the lowest GCSFR MFI. I suggested recalling five individuals homozygous for the GAG/GAG haplotype (highest GCSFR MFI) and all CGG/CGA individuals (lowest GCSFR MFI) (Table 4.7). Comparison of individuals with these haplotypes should give the greatest power to replicate the receptor effect. This is currently ongoing. Experiments assessing STAT3 phosphorylation (described above) and measuring neutrophil function responses (Chapter 3) are also being considered.

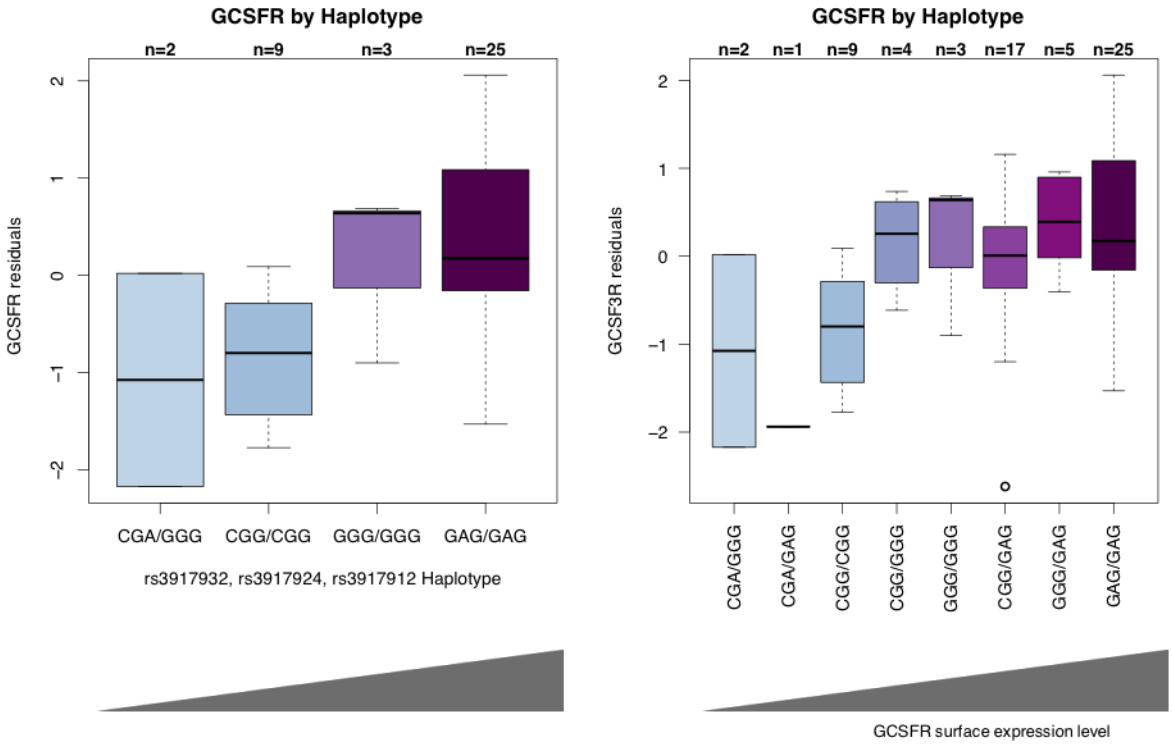


**Figure 4.14 GCSFR MFI stratified by haplotype**
GCSFR MFI residuals (original Cambridge cohort) stratified by haplotype of rs3917932 (decreasing allele = C), rs3917924 (decreasing allele = G), rs3917914 (decreasing allele = A). The lowest receptor expression was associated with individuals who are heterozygous for the rare variant and the neutrophil count lead variant, rs3917932 (CGA/GGG). CGA contains the lowering effect alleles for all three SNPs. There were no individuals homozygous for the rare lowering haplotype (CGA) in the cohort. The left panel shows the effects in individuals with homozygous haplotypes. The right panel shows MFI for all of the haplotypes estimated using 66 individuals.

| Haplotype (estimated frequency) | Haplotype Genotypes | No of Individuals recalled |
|---|---|---|
| CG**A** (3%) | CGG/CG**A** | 4 |
| | CG**A**/GGG | 2 |
| | CG**A**/GAG | |
| CGG (43%) | CGG/CGG | 4 |
| | CGG/CG**A** | - |
| | CGG/GAG | |
| | CGG/GGG | |
| GGG (22%) | CGG/GGG | |
| | CG**A**/GGG | - |
| | GGG/GAG | |
| | GGG/GGG | |
| GAG (33%) | CGG/GAG | |
| | CG**A**/GAG | |
| | GGG/GAG | |
| | GAG/GAG | 5 |

**Table 4.7: Haplotype frequencies of the Sanquin replication cohort**
Haplotype frequencies of the four main haplotypes for the Sanquin replication cohort are shown, along with all the haplotype combinations in 140 individuals. There are 9 unique haplotype genotype combinations and the frequencies matched those from UK10K (data not shown). The suggested number of individuals to be recalled is listed and based on haplotypes that will demonstrate the biggest difference in receptor levels as predicted from the discovery cohort.