Red Blood Cell Distribution Width: genetic evidence for aging pathways in 116,666 volunteers

Luke C. Pilling[1], Janice L. Atkins[1], Michael O. Duff[2], Robin N. Beaumont[3], Samuel E. Jones[3], Jessica Tyrrell[3], Chia-Ling Kuo[4], Katherine S. Ruth[3], Marcus A. Tuke[3], Hanieh Yaghootkar[3], Rachel M. Freathy[3], Andrew R. Wood[3], Anna Murray[3], Michael N. Weedon[3], Lorna W. Harries[5], George A. Kuchel[6], Luigi Ferrucci[7], Timothy M. Frayling[3], David Melzer[1,6]

1. Epidemiology and Public Health Group, University of Exeter Medical School, RILD Level 3, Royal Devon & Exeter Hospital, Barrack Road, Exeter, EX2 5DW, UK.
2. Department of Genetics and Genome Sciences, Institute for Systems Genomics, University of Connecticut Health Center, Farmington, Connecticut 06030, USA.
3. Genetics of Complex Traits Group, University of Exeter Medical School, RILD Level 3, Royal Devon & Exeter Hospital, Barrack Road, Exeter, EX2 5DW, UK.
4. Department of Community Medicine and Health Care, Connecticut Institute for Clinical and Translational Science, Institute for Systems Genomics, University of Connecticut Health Center
5. Institute of Biomedical and Clinical Sciences, University of Exeter Medical School, RILD Level 3, Royal Devon & Exeter Hospital, Barrack Road, Exeter, UK
6. Center on Aging, University of Connecticut, 263 Farmington Avenue, Farmington, CT 06030-5215, USA.
7. National Institute on Aging, Baltimore, MD, United States.

Corresponding Author

Dr Luke C. Pilling,

Epidemiology and Public Health Group,
University of Exeter Medical School,
RILD building (level 3 research), RD&E Wonford, Barrack Road, Exeter, EX2 5DW, UK.

E-mail: L.Pilling@exeter.ac.uk

Phone: +44 (0) 1392 406746

Keywords: RDW, survival, GWAS, genetic, human

Short title: Genetic influences on Red Blood Cell variation

Abstract

Variability in red blood cell volumes (distribution width, RDW) increases with age and is strongly predictive of mortality, incident coronary heart disease and cancer. We investigated inherited genetic variation associated with RDW in 166,666 UK Biobank human volunteers.

A large proportion RDW is explained by genetic variants (29%), especially in the older group (60+ year olds, 33.8%, <50 year olds, 28.4%). RDW was associated with 194 independent genetic signals; 71 are known for conditions including autoimmune disease, certain cancers, BMI, Alzheimer's disease, longevity, age at menopause, bone density, myositis, Parkinson's disease, and age-related macular degeneration. Pathways analysis showed enrichment for telomere maintenance, ribosomal RNA and apoptosis.

Although increased RDW is predictive of cardiovascular outcomes, this was not explained by known CVD or related lipid genetic risks. The predictive value of RDW for a range of negative health outcomes may in part be due to variants influencing fundamental pathways of aging.

Introduction

Increased variation in a person's Red Blood Cell (RBC) volumes (RBC distribution width (RDW), also termed anisocytosis) is strongly predictive of a range of incident cardiovascular conditions, cancers and mortality [1]–[3]. Although RDW is routinely measured in clinical hematology reporting – it is calculated by dividing the standard deviation of mean cell volume (MCV) by the MCV and multiplying by 100, to yield a RDW percentage [4] – it is only used clinically for diagnosis of anemia subtypes. Understanding the mechanisms involved in the links between increased RDW and negative health outcomes could provide clues to potential interventions to improve prognosis in those with high RDW who are not anemic, particularly in older people.

Established clinical causes of increased RDW include anemia and other iron or folate deficiencies [5], dyslipidemia [6] and other metabolic abnormalities, and inflammation [7]. Proposed mechanisms for increased RDW also include impaired erythropoiesis (the generation of new RBC) perhaps due to effects of inflammation or senescence of erythropoietic cells in the bone marrow, plus variations in RBC survival [8]. A previous analysis of 36 blood cell traits identified genetic variants [9], but the genetic signals for RDW were not investigated in depth in relation to the mechanisms that might explain the predictive value of RDW for negative health outcomes in people with normal hemoglobin levels.

We aimed to investigate RDW (overall and excluding anemia) using genetic analysis in a large population cohort, to identify underlying mechanisms. This involved genome-wide analysis of associations to find independent signals, investigations of biological pathways implicated by the results, and overlap with known risk alleles. We also examined associations between RDW and known variant genetic risk score analysis for conditions predicted by RDW, including cardiovascular disease. For this analysis, we used the exceptionally large UK Biobank volunteer sample with standardized measures of RDW across the cohort.

Results

We included 116,666 UK Biobank participants of white/British descent with complete hematology measures, covariate data, and genetics data from the interim data release (May 2015) in our analyses. The mean age was 57 years (SD: 7.9, min=40, max=70), with 52,541 aged 60 to 70 years old, and the majority (52.6%) were female: **Table 1**.

Variance explained by genotypes

We estimated that 29.3% (SE = 0.5%) of the variance in RDW was accounted for by 457,643 directly genotyped variants with MAF>0.1%. In a secondary analysis we estimated the proportion explained in two groups: those aged ≥60 years (n=52,541), and those aged <50 years (n=24,988). The proportion of variance in RDW explained by the genetic variants was greater in the older group (33.8%, SE = 1.0%) compared to the younger group (28.4%, SE = 2.0%).

We estimated the proportion of variance in CHD (10,280 cases in 116,666 participants) accounted for by the variants to be 5.95% (SE = 0.45%). The proportion of the variance shared between RDW and CHD attributable to genetics is 6.62% (SE = 2.69%: 95% CIs = 1.35 to 11.9%).

Genome-Wide Association Study

Of the 16,832,071 genetic variants included in this GWAS, 30,988 were significantly ($p<5x10^{-8}$) associated with RDW (**Figure 1**; full results available to download here: http://www.t2diabetesgenes.org/data) after adjustment for age, sex, assessment center and array type (genetic relatedness is accounted for in the linear mixed models approach so no PCs are included – see methods). The 30,988 variants were mapped to 141 loci (runs of variants separated by <2Mb) on the genome, and included 194 independent signals after conditional analysis ("conditional SNPS") (**Supplementary Table 1**).

After excluding 17,795 participants with anemia the GWAS of RDW was repeated on the 98,871 remaining participants; all 194 conditional SNPs remained nominally associated with RDW ($p<0.001$), but 24 were no longer genome-wide significant ($p<5x10^{-8}$), possibly due to reduced power in the smaller sample size.

Functional Implications of RDW-associated Genetic Variants

We utilized the UCSC Variant Annotation Integrator (http://www.genome.ucsc.edu/cgi-bin/hgVai) and the Ensembl Variant Effect Predictor (http://grch37.ensembl.org/Homo_sapiens/Tools/VEP/) to interrogate a number of genomic annotation databases for the conditional SNPs. The majority (119 out of 194 total) were intronic, and 12 were located in the 3' or 5' un-translated regions. There were also exonic variants associated with RDW (11 missense and 3 synonymous), including a 17 base-pair exonic deletion in gene *SMIM1*. The missense and deletion variants are shown in **Table 2** (see **Supplementary Table 2** for complete variant annotation output).

PolyPhen-2 predicted that three missense variants are "probably damaging" to the protein function of genes *TRIM58, PLD1* and *PNPLA3*. Variant rs2075995 was predicted to be "possibly damaging" to the protein function of gene *E2F2*. *TRIM58* is a ubiquitin ligase induced during late erythropoiesis [10]; *PLD1* is a phospholipase implicated in processes including membrane trafficking [10]; *PNPLA3* is a triacylglycerol lipase in adipocytes and variant rs738409 is associated with susceptibility to Non-alcoholic Fatty Liver Disease [11]. Gene *SMIM1* is involved in RBC formation [10] and a 17 base-pair deletion (rs566629828) causing an exonic frameshift is strongly associated with increased RDW.

GWAS Catalogue of Known Genetic Associations

Of the 194 conditional SNPs associated with RDW, 77 mapped to at least one trait in the catalogue of published GWAS (downloaded 13th March 2017). This was arrived at by; filtering the 33,005 SNP-trait associations to those with $p<5x10^{-8}$ (leaving 14,148 SNP-trait pairs for analysis); matching the positions to the UK Biobank results (13,146); filtering to those with significant RDW association in our analysis ($p<5x10^{-8}$), leaving 923 SNP-trait pairs (420 unique SNPs; some SNPs are associated with multiple traits). These 420 unique SNPs mapped to 77 of the 194 conditional SNPs associated with RDW. These are shown in **Figure 1**; see **Supplementary Table 3** for further detail.

Traits also associated with the individual RDW variants included iron metabolism and several other red cell measures. Variants present were also associated with BMI, several lipids, hemoglobin A1C and metabolic syndrome, as well as height. Autoimmune associated conditions included autoimmune thyroid disease, type 1 diabetes, Crohn's disease, inflammatory bowel disease, rheumatoid arthritis, systemic lupus erythematosus and ulcerative colitis. Variants linked to Lung, ovary and nasopharyngeal cancers were present. In addition, conditions associated with aging were represented, including variants linked to Alzheimer's disease, age at menopause, bone density,

Myositis, Parkinson's disease, macular degeneration, C-reactive protein levels and longevity. For Alzheimer's and longevity these were known SNPs in the *APOE* gene region.

Gene Ontology Enrichment

MAGENTA software [12] identified pathways enriched in the genes mapped to variants significantly associated with RDW, including telomere maintenance, ribosomal RNA transcription and histone modifications (**Table 3**), plus apoptosis. In addition, pathways related to lipid metabolism (in particular chylomicrons) were also enriched.

Genetic Risk Score associations

We tested 20 Genetic Risk Scores (GRS) for potentially explanatory traits for the predictive value of RDW for negative health outcomes; 7 were significant after adjustment for multiple testing ($p<0.0025$). Three were associated with raised RDW (HDL, type-1 diabetes, and BMI); four were associated with lower RDW (triglycerides, LDL, systolic blood pressure, and Alzheimer's disease) (**Figure 2; Supplementary Table 5**). After exclusion of the ApoE locus the HDL GRS remained significant ($p=0.003$), but the AD GRS was no longer significant ($p=0.84$).

Genetic risks for Crohn's disease and inflammatory bowel disease were nominally associated with increased RDW (Beta=0.009: 95% CIs=0.003 to 0.015; Beta=0.009: 95% CIs=0.003 to 0.014; respectively). Risk scores for coronary artery disease were not associated with RDW, including after removal of the lipid related variants (Beta=0.00: 95% CIs=-0.006 to 0.005). A genetic risk score for telomere length was not significantly associated with RDW.

Discussion

Variation in RBC size (RBC Distribution Width, RDW) increases markedly with age [13] and high RDW values are strongly predictive of increased mortality, plus incident cardiovascular disease and certain cancers [3], [14]. However, RDW is not generally considered as being a clinically useful measure outside the assessment of anemia sub-type, perhaps because the mechanisms explaining its prognostic value in people without anemia are unclear. In this study we investigated RDW using genetic analysis to understand the molecular mechanisms underpinning variation in RBC size.

A large proportion of RDW variation (29.3%) was attributable to common genetic variants in this analysis, and the variation explained by genetic variants appeared to increase with age, contrary to common assumptions that genetic effects decrease with advancing age. This could be due to effects accumulating over a time. Many of the "hallmarks of aging" also have this property [15]. Additionally, many of the RDW-associated genetic variants (in 71 of 194 conditionally independent signals) have previously been associated with other traits including metabolic syndrome, certain cancers, and autoimmune disease as well as aging related conditions including menopausal age.

Our analysis of genetic risk scores (GRS) showed that participants with genetically-increased risk for coronary artery disease or cancer did not have significantly higher RDW. Consistent with this, the proportion of the variance shared between RDW and coronary heart disease (CHD) attributable to genetics was only 6.6%; therefore the majority of genetically influenced red cell variation is independent of CHD.

As higher RDW is associated with CVD, we might have hypothesized that genetic risks for adverse lipid levels or blood pressures affect RDW, but instead we found associations in the opposite directions: GRS analysis showed that participants with genetically lower LDL levels, triglyceride levels, or systolic blood pressure, had higher RDW, and genetically higher HDL was associated with greater RDW. Published observational associations between lipids and RDW are only partially consistent with our findings here: Lippi *et al* found that LDL and HDL cholesterol were both negatively related to RDW, and triglyceride levels were positively related [6]. It is known that RBC have a role in cholesterol homeostasis by transporting cholesterol in the plasma membranes, with significant inter-individual differences not entirely explained by age or cholesterol levels [16]. The relationship between lipids and RDW is complex and requires further investigation. We also observed that genetically increased risk of type-1 diabetes was associated with increased RDW – further evidence for autoimmune involvement, in addition to the overlap in significant SNPs in the GWAS catalogue – and that genetically increased BMI is associated with increased RDW.

Genetic variants associated with RDW were enriched in expected pathways, including iron homeostasis, but we also found evidence for telomere maintenance, ribosomal RNA production, and a number of nucleosome and histone pathways. Short telomere length is a hallmark of cellular aging to senescence [15] and is associated with many risk factors of disease, however the causal direction is still uncertain [17], and longer telomeres have been linked to risk of cancer [18]. Kozlitina *et al* reported in 2012 that increased RDW is associated with shorter telomeres in leukocytes [19]. We created a genetic risk score for telomere length but it was not associated with RDW (Beta=0.004: 95% CIs=-0.002 to 0.010); further work is required to clarify this association.

We also found enrichment of RNA polymerase I (which transcribes ribosomal RNA) and RNA polymerase III (which transcribes transfer RNA); both are required for protein synthesis, including hemoglobin, and can even function as regulators of gene expression in their own right [20], suggesting these are key factors for consistent production of RBC. Deregulation of transcription and proteostasis are hallmarks of aging [15], and we have previously reported deregulation of gene expression of the transcriptional machinery with advancing age [21].

Four of the conditionally independent genetic variants associated with RDW are exonic and affect the amino-acid sequence of the protein products. These will be useful targets for future research to determine how these variants ultimately affect consistency in red blood cell size.

Limitations

The UK Biobank is a volunteer study which achieved only a 5% response rate, so at assessment the participants were healthier than the general population. However, there was substantial variation in RDW within the participants so the results can still be generalized to the wider population [22].

More work is needed to establish the effects in other populations. No data have been released regarding the UK Biobank participant's lipid and other blood assays; once this is available, further investigations into the complex relationship between RDW and lipids can be performed.

The GWAS catalogue does not contain every published GWAS, especially the most recent studies, but contains many of the largest meta-analyses for traits such as cardiovascular disease and cancer. It is therefore likely we have missed some studies using this method, therefore our results present an approximation of the overlap between RDW signals and other traits.

Conclusions

Variation in RDW has a substantial genetic component, and this increases with increasing age. Although increased RDW is predictive of cardiovascular outcomes, this was not explained by known CVD or related lipid genetic risks. The predictive value of RDW for a range of negative health outcomes may in part be due to variants influencing fundamental pathways of aging.

Methods

The UK Biobank study recruited 503,325 volunteers aged 40-70 who were seen between 2006 and 2010. Data includes RBC distribution width (RDW) and other clinical hematology measures, extensive questionnaires including smoking behavior and education history, and follow-up using electronic medical records. Currently one third of the participants have available genotype information. We utilized data from 116,666 participants of white/British descent with all available data.

Phenotypes

RDW is a measure of the variability in the mean size of the RBC in each participant (in % units). It was measured using four Beckman Coulter LH750 instruments within 24 hours of blood draw, with extensive quality control performed by UK Biobank [23]. RDW is a continuous, highly skewed trait, therefore we used quantile normalization of the continuous measure to create a Gaussian distribution, so that the normality assumption of the linear regression models were not violated.

Anemia was determined both using self-reported diagnosis, electronic medical records (ICD10: D64* and D5* categories), or by low hemoglobin levels at the baseline assessment (<120g/L in females, <130g/L in males: from WHO definition [24]).

Coronary heart disease (CHD) was defined using self-reported myocardial infarction or angina, or diagnosis in the electronic medical records (ICD10: I20-I25).

Variance explained and genetic correlation

We used BOLT-REML to determine the variance in RDW explained by the common, genotyped variants (n= 457,643 directly genotyped variants with MAF>0.1%, HWE $p>1x10^{-6}$ and missingness <1.5%) using restricted maximum likelihood estimation [25]. BOLT-REML was also used to estimate the genetic correlation between RDW and CHD.

Genome-wide association study

We performed a GWAS in 116,666 white/British participants (those with currently available genotyping from the 503,325 total UK Biobank participants), with complete genetic data to determine genetic variants associated with RDW. Over 800,000 genetic variants were directly genotyped using an Affymetrix Axiom array. After imputation, quality control, and filtering (we

included autosomal variants with minor allele frequency (MAF) ≥0.1%, missingness <1.5%, imputation quality >0.4 and with Hardy-Weinberg equilibrium (HWE) $p>1x10^{-6}$ within the white/British participants) 16,889,199 genetic variants were available for GWAS analysis: methods described in detail previously [26]–[28]. We also utilized data directly from the microarrays for variants on the X (n=19,381) and Y (n=284) chromosomes, and on the mitochondrial genome (n=135), which were unavailable in the imputed dataset.

GWAS was performed using BOLT-LMM, a software that uses linear mixed-effect models to determine associations between each variant and the outcome, incorporating genetic relatedness [25]. RDW residuals from a linear model adjusted for age, sex and assessment center were quantile-normalized prior to analysis. Models were adjusted for array type (two different Affymetrix arrays were used, which are >95% identical) at run time. Variants were classed as significant if the p-value for the association with RDW was less than $5x10^{-8}$.

Identifying Conditionally Independent GWAS Signals

Many of the identified variants are correlated and may therefore not be independent; we used conditional analysis to determine independent signals by adjusting each variant in a locus for the most significant variant in that locus (loci defined as runs of SNPs separated by <2Mb on a chromosome). This process was repeated until only "conditional SNPs" remained that were significantly associated with RDW independent of one another.

Conditional SNPs were checked for their consistency of association with RDW in two sensitivity analyses: once excluding participants with prevalent anemia (either clinical diagnosis, self-report, or raised hemoglobin), and in a separate analysis excluding participants <60 years of age.

GWAS-significant SNPs follow-up

The determine the gene-locations and possible functional consequences of the conditionally independent SNPs we submitted them to the UCSC Variant Annotation Integrator (https://genome.ucsc.edu/cgi-bin/hgVai), which combines information from several sources to determine the probable effect of a genetic variant, including on specific genes (for example intronic, missense, splice site, intergenic etc.) and PolyPhen-2 (a tool for predicting the impact of amino acid substitutions on the protein product).

The GWAS catalogue of published variant-trait associations was searched for all SNPs (not just conditional SNPs) associated with RDW to determine which loci had previously been associated with another trait in GWAS analyses ($p<5x10^{-8}$), and which were novel [29]. We used the UCSC `liftOver` tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver) to match the genomic coordinates between the GWAS catalogue (GRCh38) and the UK Biobank genetics data (GRCh37).

We used the software package MAGENTA [12] to determine whether any biological pathways were enriched in the GWAS results, using the pathways database file GO_PANTHER_INGENUITY_KEGG_REACTOME_BIOCARTA.

Genetic Risk Scores

Twenty Genetic Risk Scores (GRS) were computed for each participant based on the number of trait-raising alleles they have for a particular phenotype, such as LDL cholesterol or type-2 diabetes. These were computed according to the method described in Pilling *et al* 2016 [28]. Each of the 20 GRS computed was tested for its association with RDW using linear regression models, adjusted for age, sex, assessment center, genotype array, and population stratification (using the first 5 principal components (PCs)).

Funding and Acknowledgements

This work was supported by an award to DM, TF, AM, LH and CB by the UK Medical Research Council (grant number MR/M023095/1). SEJ is funded by the Medical Research Council (grant: MR/M005070/1). JT is funded by a Diabetes Research and Wellness Foundation Fellowship. R.B. is funded by the Wellcome Trust and Royal Society grant: 104150/Z/14/Z. MAT, MNW and AM are supported by the Wellcome Trust Institutional Strategic Support Award (WT097835MF). RMF is a Sir Henry Dale Fellow (Wellcome Trust and Royal Society grant: 104150/Z/14/Z). ARW, HY, and TMF are supported by the European Research Council grant: 323195:GLUCOSEGENES-FP7-IDEAS-ERC. The funders had no influence on study design, data collection and analysis, decision to publish, or preparation of the manuscript. LF is supported by the Intramural Research Program of the National Institute on Aging, U.S. National Institutes of Health. Input from MD, CL and GK was supported by the University of Connecticut Health Center.

This research has been conducted using the UK Biobank Resource under Application Number 14631. We thank UK Biobank participants and coordinators for this dataset.

Conflict of interest: None declared.

References

[1] K. V. Patel, L. Ferrucci, W. B. Ershler, D. L. Longo, and J. M. Guralnik, “Red blood cell distribution width and the risk of death in middle-aged and older adults.,” *Arch. Intern. Med.*, vol. 169, no. 5, pp. 515–23, Mar. 2009 DOI: 10.1001/archinternmed.2009.11.

[2] K. V Patel, R. D. Semba, L. Ferrucci, A. B. Newman, L. P. Fried, R. B. Wallace, S. Bandinelli, C. S. Phillips, B. Yu, S. Connelly, M. G. Shlipak, P. H. M. Chaves, L. J. Launer, W. B. Ershler, T. B. Harris, D. L. Longo, and J. M. Guralnik, “Red cell distribution width and mortality in older adults: a meta-analysis.,” *J Gerontol A Biol Sci Med Sci*, vol. 65, no. 3, pp. 258–65, 2010 DOI: 10.1093/gerona/glp163.

[3] G. Turcato, V. Serafini, A. Dilda, C. Bovo, B. Caruso, G. Ricci, and G. Lippi, “Red blood cell distribution width independently predicts medium-term mortality and major adverse cardiac events after an acute coronary syndrome,” *Ann. Transl. Med.*, vol. 4, no. 13, pp. 254–254, 2016 DOI: 10.21037/atm.2016.06.35.

[4] E. Danese, G. Lippi, and M. Montagnana, “Red blood cell distribution width and cardiovascular diseases.,” *J. Thorac. Dis.*, vol. 7, no. 10, pp. E402-11, Oct. 2015 DOI: 10.3978/j.issn.2072-1439.2015.10.04.

[5] R. Inuzuka and J. Abe, “Red Blood Cell Distribution Width as a Link Between Ineffective Erythropoiesis and Chronic Inflammation in Heart Failure,” *Circ. J.*, vol. 79, no. May, pp. 3–4, 2015 DOI: 10.1253/circj.CJ-15-0254.

[6] G. Lippi, F. Sanchis-Gomar, E. Danese, and M. Montagnana, “Association of red blood cell distribution width with plasma lipids in a general population of unselected outpatients.,” *Kardiol. Pol.*, vol. 71, no. 9, pp. 931–6, 2013 DOI: 10.5603/KP.2013.0228.

[7] R. Sharma and V. V Agrawal, “The Relationship Between Red Blood Cell Distribution Width (RDW CV) And C Reactive Protein (CRP) With The Clinical Outcomes In Non-St Elevation Myocardial Infarction And Unstable Angina Pectoris: A 6 Months Follow Up Study,” *Int. Cardiovasc. Forum J.*, vol. 2, no. 1, pp. 27–31, 2015 DOI: 10.17987/icfj.v2i1.68.

[8] G. L. Salvagno, F. Sanchis-Gomar, A. Picanza, and G. Lippi, “Red blood cell distribution width: A simple parameter with multiple clinical applications.,” *Crit. Rev. Clin. Lab. Sci.*, vol. 52, no. 2, pp. 86–105, 2015 DOI: 10.3109/10408363.2014.992064.

[9] W. J. Astle, H. Elding, T. Jiang, D. Allen, D. Ruklisa, A. L. Mann, D. Mead, H. Bouman, F. Riveros-Mckay, M. A. Kostadima, J. J. Lambourne, S. Sivapalaratnam, K. Downes, K. Kundu, L. Bomba, K. Berentsen, J. R. Bradley, L. C. Daugherty, O. Delaneau, K. Freson, S. F. Garner, L. Grassi, J. Guerrero, M. Haimel, E. M. Janssen-Megens, A. Kaan, M. Kamat, B. Kim, A. Mandoli, J. Marchini, J. H. A. Martens, S. Meacham, K. Megy, J. O’Connell, R. Petersen, N. Sharifi, S. M. Sheard, J. R. Staley, S. Tuna, M. van der Ent, K. Walter, S.-Y. Wang, E. Wheeler, S. P. Wilder, V. Iotchkova, C. Moore, J. Sambrook, H. G. Stunnenberg, E. Di Angelantonio, S. Kaptoge, T. W. Kuijpers, E. Carrillo-de-Santa-Pau, D. Juan, D. Rico, A. Valencia, L. Chen, B. Ge, L. Vasquez, T. Kwan, D. Garrido-Martín, S. Watt, Y. Yang, R. Guigo, S. Beck, D. S. Paul, T. Pastinen, D. Bujold, G. Bourque, M. Frontini, J. Danesh, D. J. Roberts, W. H. Ouwehand, A. S. Butterworth, and N. Soranzo, “The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease,” *Cell*, vol. 167, no. 5, p. 1415–1429.e19, Nov. 2016 DOI: 10.1016/j.cell.2016.10.042.

[10] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet, “GeneCards: integrating information about genes, proteins and diseases.,” *Trends Genet.*, vol. 13, no. 4, p. 163, Apr. 1997.

[11] NCBI ClinVar, “ClinVar record for rs738409.” [Online]. Available: www.ncbi.nlm.nih.gov/clinvar/RCV000299548. [Accessed: 15-Feb-2017].

[12] A. V Segrè, DIAGRAM Consortium, MAGIC investigators, L. Groop, V. K. Mootha, M. J. Daly, and D. Altshuler, “Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits.,” *PLoS Genet.*, vol. 6, no. 8, Aug. 2010 DOI: 10.1371/journal.pgen.1001058.

[13] C. K.-W. Cheng, J. Chan, G. S. Cembrowski, and O. W. van Assendelft, “Complete blood count reference interval diagrams derived from NHANES III: stratification by age, sex, and race.,” *Lab. Hematol.*, vol. 10, no. 1, pp. 42–53, 2004.

[14] G. Lippi, G. Cervellin, and F. Sanchis-Gomar, “Red blood cell distribution width and cardiovascular disorders. Does it really matter which comes first, the chicken or the egg?,” *Int. J. Cardiol.*, vol. 206, pp. 129–30, Mar. 2016 DOI: 10.1016/j.ijcard.2016.01.122.

[15] C. López-Otín, M. A. Blasco, L. Partridge, M. Serrano, and G. Kroemer, “The hallmarks of aging.,” *Cell*, vol. 153, no. 6, pp. 1194–217, Jun. 2013 DOI: 10.1016/j.cell.2013.05.039.

[16] R. S. Chakrabarti, S. A. Ingham, J. Kozlitina, A. Gay, J. C. Cohen, A. Radhakrishnan, and H. H. Hobbs, “Variability of cholesterol accessibility in human red blood cells measured using a bacterial cholesterol-binding toxin.,” *Elife*, vol. 6, Feb. 2017 DOI: 10.7554/eLife.23355.

[17] F. Y. Fyhrquist and O. J. Saijonmaa, “Modifiable Factors Influencing Telomere Length and Aging,” in *Inflammation, Aging, and Oxidative Stress*, S. C. Bondy and A. Campbell, Eds. Cham: Springer International Publishing, 2016, pp. 67–80 DOI: 10.1007/978-3-319-33486-8_4.

[18] Telomeres Mendelian Randomization Collaboration, P. C. Haycock, S. Burgess, A. Nounu, J. Zheng, G. N. Okoli, J. Bowden, K. H. Wade, N. J. Timpson, D. M. Evans, P. Willeit, A. Aviv, T. R. Gaunt, G. Hemani, M. Mangino, H. P. Ellis, K. M. Kurian, K. A. Pooley, R. A. Eeles, J. E. Lee, S. Fang, W. V Chen, M. H. Law, L. M. Bowdler, M. M. Iles, Q. Yang, B. B. Worrall, H. S. Markus, R. J. Hung, C. I. Amos, A. B. Spurdle, D. J. Thompson, T. A. O’Mara, B. Wolpin, L. Amundadottir, R. Stolzenberg-Solomon, A. Trichopoulou, N. C. Onland-Moret, E. Lund, E. J. Duell, F. Canzian, G. Severi, K. Overvad, M. J. Gunter, R. Tumino, U. Svenson, A. van Rij, A. F. Baas, M. J. Bown, N. J. Samani, F. N. G. van t’Hof, G. Tromp, G. T. Jones, H. Kuivaniemi, J. R. Elmore, M. Johansson, J. Mckay, G. Scelo, R. Carreras-Torres, V. Gaborieau, P. Brennan, P. M. Bracci, R. E. Neale, S. H. Olson, S. Gallinger, D. Li, G. M. Petersen, H. A. Risch, A. P. Klein, J. Han, C. C. Abnet, N. D. Freedman, P. R. Taylor, J. M. Maris, K. K. Aben, L. A. Kiemeney, S. H. Vermeulen, J. K. Wiencke, K. M. Walsh, M. Wrensch, T. Rice, C. Turnbull, K. Litchfield, L. Paternoster, M. Standl, G. R. Abecasis, J. P. SanGiovanni, Y. Li, V. Mijatovic, Y. Sapkota, S.-K. Low, K. T. Zondervan, G. W. Montgomery, D. R. Nyholt, D. A. van Heel, K. Hunt, D. E. Arking, F. N. Ashar, N. Sotoodehnia, D. Woo, J. Rosand, M. E. Comeau, W. M. Brown, E. K. Silverman, J. E. Hokanson, M. H. Cho, J. Hui, M. A. Ferreira, P. J. Thompson, A. C. Morrison, J. F. Felix, N. L. Smith, A. M. Christiano, L. Petukhova, R. C. Betz, X. Fan, X. Zhang, C. Zhu, C. D. Langefeld, S. D. Thompson, F. Wang, X. Lin, D. A. Schwartz, T. Fingerlin, J. I. Rotter, M. F. Cotch, R. A. Jensen, M. Munz, H. Dommisch, A. S. Schaefer, F. Han, H. M. Ollila, R. P. Hillary, O. Albagha, S. H. Ralston, C. Zeng, W. Zheng, X.-O. Shu, A. Reis, S. Uebe, U. Hüffmeier, Y. Kawamura, T. Otowa, T. Sasaki, M. L. Hibberd, S. Davila, G. Xie, K. Siminovitch, J.-X. Bei, Y.-X. Zeng, A. Försti, B. Chen, S. Landi, A. Franke, A. Fischer, D. Ellinghaus, C. Flores, I. Noth, S.-F. Ma, J. N. Foo, J. Liu, J.-W. Kim, D. G. Cox, O. Delattre, O. Mirabeau, C. F. Skibola, C. S. Tang, M. Garcia-Barcelo, K.-P. Chang, W.-H. Su, Y.-S. Chang, N. G. Martin, S. Gordon, T. D. Wade, C. Lee, M. Kubo, P.-C. Cha, Y. Nakamura, D. Levy, M. Kimura, S.-J. Hwang, S. Hunt, T. Spector, N. Soranzo, A. W. Manichaikul, R. G. Barr, B. Kahali, E. Speliotes, L. M. Yerges-Armstrong, C.-Y. Cheng, J. B. Jonas, T. Y. Wong, I. Fogh, K. Lin, J. F. Powell, K. Rice, C. L. Relton, R. M. Martin, and G. Davey

Smith, “Association Between Telomere Length and Risk of Cancer and Non-Neoplastic Diseases: A Mendelian Randomization Study.,” *JAMA Oncol.*, Feb. 2017 DOI: 10.1001/jamaoncol.2016.5945.

[19] J. Kozlitina and C. K. Garcia, “Red blood cell size is inversely associated with leukocyte telomere length in a large multi-ethnic population.,” *PLoS One*, vol. 7, no. 12, p. e51046, 2012 DOI: 10.1371/journal.pone.0051046.

[20] C. Megel, G. Morelle, S. Lalande, A.-M. Duchêne, I. Small, and L. Maréchal-Drouard, “Surveillance and cleavage of eukaryotic tRNAs.,” *Int. J. Mol. Sci.*, vol. 16, no. 1, pp. 1873–93, Jan. 2015 DOI: 10.3390/ijms16011873.

[21] L. W. Harries, D. Hernandez, W. Henley, A. R. Wood, A. C. Holly, R. M. Bradley-Smith, H. Yaghootkar, A. Dutta, A. Murray, T. M. Frayling, J. M. Guralnik, S. Bandinelli, A. Singleton, L. Ferrucci, and D. Melzer, “Human aging is characterized by focused changes in gene expression and deregulation of alternative splicing.,” *Aging Cell*, vol. 10, no. 5, pp. 868–78, Oct. 2011 DOI: 10.1111/j.1474-9726.2011.00726.x.

[22] T. A. Manolio and R. Collins, “Enhancing the feasibility of large cohort studies.,” *JAMA*, vol. 304, no. 20, pp. 2290–1, Nov. 2010 DOI: 10.1001/jama.2010.1686.

[23] S. Sheard, R. Nicholls, and J. Froggat, “UK Biobank Haematology Data Companion Document,” 2016. [Online]. Available: https://biobank.ctsu.ox.ac.uk/crystal/docs/haematology.pdf. [Accessed: 25-May-2016].

[24] WHO, “WHO definition of anemia by hemoglobin concentration.” [Online]. Available: http://www.who.int/vmnis/indicators/haemoglobin.pdf. [Accessed: 20-Oct-2016].

[25] P.-R. Loh, G. Tucker, B. K. Bulik-Sullivan, B. J. Vilhjálmsson, H. K. Finucane, R. M. Salem, D. I. Chasman, P. M. Ridker, B. M. Neale, B. Berger, N. Patterson, and A. L. Price, “Efficient Bayesian mixed-model analysis increases association power in large cohorts.,” *Nat. Genet.*, vol. 47, no. 3, pp. 284–90, Mar. 2015 DOI: 10.1038/ng.3190.

[26] UK Biobank, “UK Biobank genotype collections and QC on 150k participants, April 2014,” 2014. [Online]. Available: http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/UKBiobank_genotyping_QC_documentation-web.pdf.

[27] UK Biobank, “UK Biobank imputation on 150k participants, May 2015,” 2015. [Online]. Available: http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/imputation_documentation_May2015.pdf.

[28] L. C. Pilling, J. L. Atkins, K. Bowman, S. E. Jones, J. Tyrrell, R. N. Beaumont, K. S. Ruth, M. A. Tuke, H. Yaghootkar, A. R. Wood, R. M. Freathy, A. Murray, M. N. Weedon, L. Xue, K. Lunetta, J. M. Murabito, L. W. Harries, J. Robine, C. Brayne, G. A. Kuchel, L. Ferrucci, T. M. Frayling, and D. Melzer, “Human longevity is influenced by many genetic variants: evidence from 75,000 UK Biobank participants.,” *Aging (Albany. NY).*, vol. 8, no. 3, pp. 547–63, Mar. 2016 DOI: 10.1101/038430.

[29] P. H. (EBI) and W. D. (EBI) Burdett T (EBI), Hall PN (NHGRI), Hastings E (EBI), Hindorff LA (NHGRI), Junkins HA (NHGRI), Klemm AK (NHGRI), MacArthur J (EBI), Manolio TA (NHGRI), Morales J (EBI), “The NHGRI-EBI Catalog of published genome-wide association studies,” 2016. [Online]. Available: www.ebi.ac.uk/gwas. [Accessed: 23-Oct-2016].

Table 1 | Summary statistics for 116,666 UK Biobank participants

Trait	Mean (SD)	Min - Max
Age (years)	56.92 (7.94)	40 - 70
RDW (%)	13.49 (0.95)	11.1 - 38.3
Mean Cell Volume (fL)	91.39 (4.45)	54.5 - 160.3
Hemoglobin conc. (g/dL)	14.23 (1.23)	0.14 - 20.5
Sex	**N**	**%**
Females	61,306	52.55
Males	55,361	47.45
Anemia		
No	98,871	84.75
*Yes **	17,795	15.25
RDW (%)		
<12.5	8,703	7.46
12.5-12.9	24,559	21.05
13.0-13.4	33,804	28.97
13.5-13.9	25,413	21.78
14.0-14.4	12,967	11.11
14.5-14.9	5,460	4.68
≥15.0	5,761	4.94

* = either hospital diagnosis or raised hemoglobin (WHO definition, see methods)

Table 2 | Four RDW-associated conditional genetic variants may have damaging effects on proteins

	Position								RDW association		
Variant	**CHR : POS**	**cDNA**	**Protein**	**eA**	**Gene**	**AA**	**Codon**	**Effect**	**Beta**	**P**	**eAF**
SNP *											
rs2075995	1:23847464	1105	226	A	*E2F2*	Q/H	caG/caT	P	0.033	1.8×10^{-16}	0.502
rs3811444	1:248039451	1169	374	T	*TRIM58*	T/M	aCg/aTg	**D**	0.066	6.3×10^{-59}	0.333
rs143845082	3:171417570	1539	398	A	*PLD1*	R/C	Cgt/Tgt	P	-0.217	4.6×10^{-13}	0.005
rs149535568	3:171442535	1056	237	A	*PLD1*	G/C	Ggc/Tgc	**D**	-0.27	9.3×10^{-15}	0.048
rs10479001	5:131607721	751	225	T	*PDLIM4*	A/V	gCa/gTa	B	-0.056	1.1×10^{-8}	0.042
rs2578377	5:153413390	459	122	T	*FAM114A2*	G/S	Ggt/Agt	B	-0.026	4.1×10^{-10}	0.633
rs1799945	6:26091179	347	63	G	*HFE*	H/D	Cat/Gat	B	0.138	2.2×10^{-143}	0.150
rs368865	13:113479820	1037	317	C	*ATP11A*	M/L	Atg/Ctg	B	-0.025	2.4×10^{-9}	0.724
rs556052	19:49377436	1215	316	C	*PPP1R15A*	A/P	Gct/Cct	B	0.028	8.5×10^{-12}	0.333
rs855791	22:37462936	2321	736	T	*TMPRSS6*	V/D	gTc/gAc	B	0.119	1.2×10^{-204}	0.439
rs738409	22:44324727	617	148	G	*PNPLA3*	I/M	atC/atG	**D**	0.029	6.2×10^{-9}	0.216
Deletion ¥											
rs566629828	1:3691997-3692014	309-325	21-26	Del	*SMIM1*	-	-	**F**	-0.12	4.5×10^{-12}	0.013

* Output from UCSC Variant Annotation Integrator for the RDW-associated conditionally independent SNPs located in protein-coding regions. ¥ Output from Ensembl Variant Effect Predictor for insertion/deletion events.

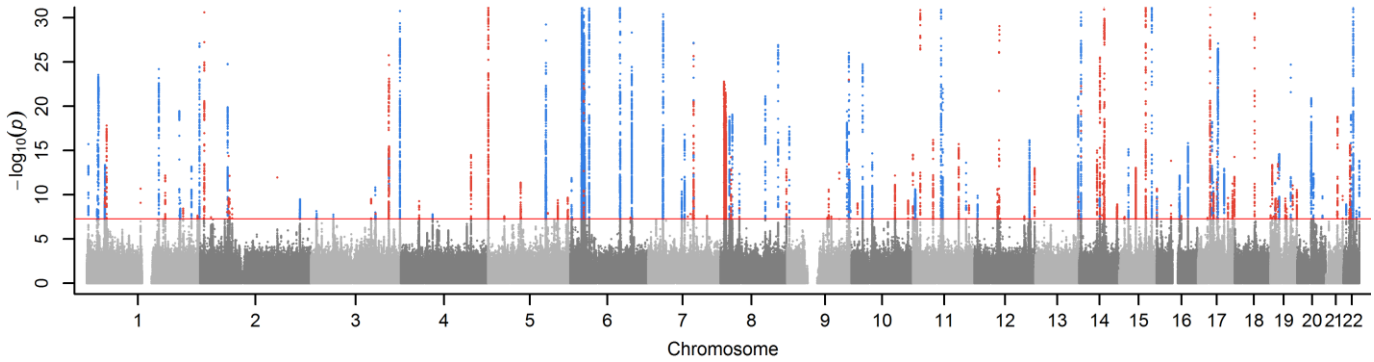
"Effect"= (D)amaging, (P)ossibly damaing, or (B)enign SNP, or (F)rameshift due to deletion; "AA"=amino-acid change; "eA"=effect allele causing the change; "Beta"=the beta coefficient for the effect allele on RDW; "P"=p-value for the RDW association; "eAF"=effect allele frequency in UK Biobank white/British participants. All positions are from hg19/b37.

Table 3 | MAGENTA results: biological pathways enriched in RDW genetics signals

Biological pathway	N genes	Exp. Sig.	Obs. Sig.	p-value
Chromosomal & DNA-related				
Histone	33	2	13	9.9×10^{-7}
Nucleosome	31	2	9	1.7×10^{-5}
Nucleosome assembly	43	2	10	1.0×10^{-4}
Packaging of telomere ends	23	1	11	9.9×10^{-7}
Telomere maintenance	49	2	12	5.0×10^{-6}
Chromatin packaging and remodeling	142	7	20	2.4×10^{-5}
Apoptosis-induced DNA fragmentation	9	0	4	4.0×10^{-4}
RNA polymerase-related				
RNA Pol I - promoter clearance	44	2	14	9.9×10^{-7}
RNA Pol I - promoter opening	23	1	14	9.9×10^{-7}
RNA Pol I - chain elongation	30	2	9	7.0×10^{-6}
RNA Pol I, III, and mitochondrial transcription	82	4	15	1.2×10^{-5}
Lipid-related				
Chylomicron	9	0	5	3.4×10^{-5}
Chylomicron-mediated lipid transport	15	1	5	6.0×10^{-4}
Phospholipid efflux	8	0	4	2.0×10^{-4}
Very-low-density lipoprotein particle	14	1	5	2.0×10^{-4}
Other				
Systemic Lupus Erythematosus	60	3	15	9.9×10^{-7}
Olfaction	105	5	18	4.0×10^{-6}
Cellular iron ion homeostasis	25	1	8	2.8×10^{-5}
Lactation (mammary development)	8	0	3	4.5×10^{-3}

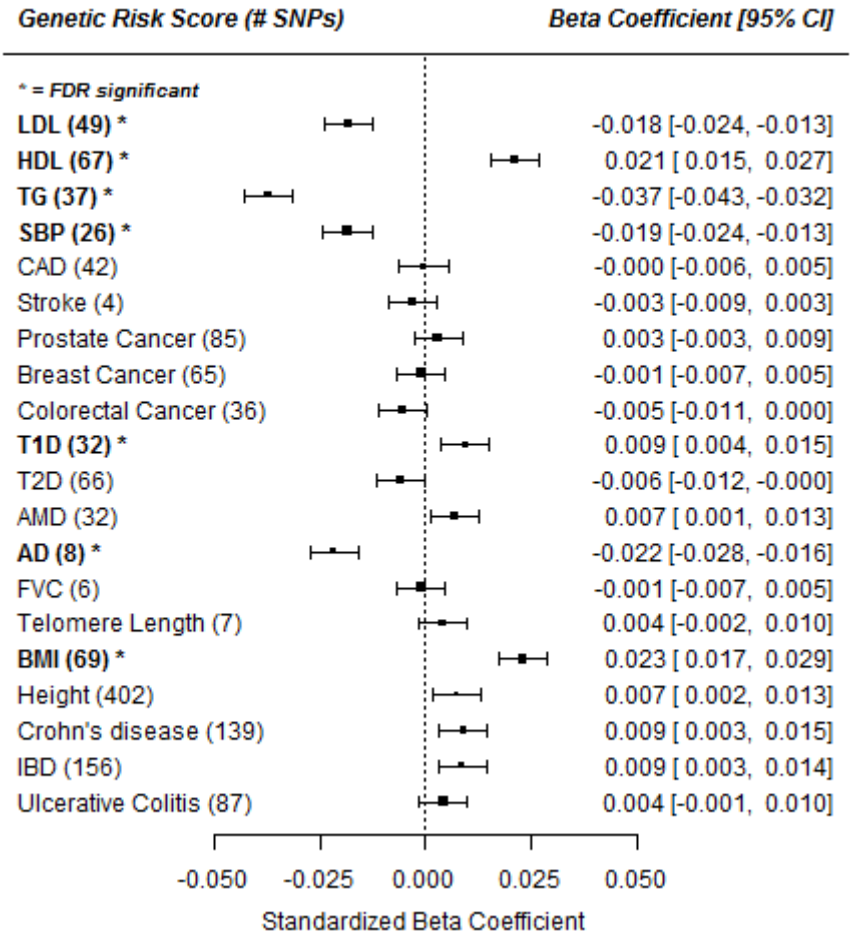
Output from MAGENTA GWAS enrichment software for the 30,988 RDW-associated genetic variants. Pathways shown are those where the estimated false discovery rate $p<0.05$. N genes=number of genes in the gene-set analysed; Exp. Sig.=Expected number of genes with a corrected gene p-value above the 95 percentile enrichment cutoff; Obs. Sig.=Observed number of genes with a corrected gene p-value above the 95 percentile enrichment cutoff; p-value=nominal p-value using 95 percentile of all gene scores for the enrichment cutoff.

Figure 1 | Genetic variants associated with RDW in GWAS of 116,666 UK Biobank participants



The variants are grouped into 194 independent signals, colored blue if a variant in the signal is associated with any trait in the NHGRI-EBI GWAS catalog of known associations, otherwise colored in red. The y-axis ($-\log_{10}$ *p*-values) is limited to 30 for clarity, as the max value is 200. See **Supplementary Table 1** for RDW associations for each signal, and **Supplementary Table 3** for mapping to the catalog. Horizontal line $p=5\times10^{-8}$.

Figure 2 | Genetic Risk Score associations with RDW



* FDR = false-discovery rate adjusted significant association. Genetic Risk Scores (GRS) were z-transformed prior to analysis. Linear regression model against RDW (z-transformed) including 116,666 participants, adjusted for age, sex, assessment center and population structure (genetic PCs 1-5). LDL (low-density lipoprotein), HDL (high-density lipoprotein), TG (triglycerides), SBP (systolic blood pressure), CAD (coronary artery disease), T1D (type-1 diabetes), T2D (type-2 diabetes), AMD (age-related macular degeneration), AD (Alzheimer's disease), FVC (forced vital capacity), BMI (body mass index), IBD (inflammatory bowel disease). Full results in **Supplementary Table 5**.