Genome-wide analysis defines genetic determinants of MPN subtypes and identifies a sex-specific association at *CDH22/CD40*

William J Tapper[1*], Ahmed A Z Dawoud[1], Joannah Score[1], Andrew J Chase[1], E Joanna Baxter[2], Joanne Ewing[3], Louise Wallis[4], Paola Guglielmelli[5], Dolors Colomer[6], Beatriz Bellosillo[7], Montse Gomez[8], Juan Carlos Hernández-Boluda[8], Carlos Besses[7], Francisco Cervantes[6], Steffen Koschmieder[9], Anthony R Green[2], Andreas Reiter[10], Alessandro Vannucchi[5], Claire Harrison[11], Nicholas C P Cross[1*]

[1] Faculty of Medicine, University of Southampton, Southampton, UK
[2] Department of Haematology, University of Cambridge, Cambridge, UK
[3] Department of Haematology, Birmingham Heartlands Hospital, Birmingham, UK
[4] Department of Haematology, Royal Bournemouth Hospital, Bournemouth, UK
[5] Center Research and Innovation of Myeloproliferative Neoplasms, DMSC, University of Florence, AOU Careggi, Florence, Italy
[6] Hospital Clinic, Institut Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Universitat Barcelona, Barcelona, Spain
[7] Hospital del Mar, Hospital del Mar Research Institute, Barcelona, Spain
[8] Hospital Clínico Universitario, INCLIVA, University of Valencia, Valencia, Spain,
[9] Department of Hematology, Oncology, Hemostaseology, and Stem Cell Transplantation, Faculty of Medicine, RWTH Aachen University, Aachen, Germany
[10] University Hospital Mannheim, Heidelberg University, Mannheim, Germany
[11] Department of Haematology, Guy's and St Thomas' NHS Foundation Trust, London, United Kingdom

* Corresponding authors
William J Tapper: University of Southampton, Duthie Building (808), Southampton General Hospital, Tremona Road, Southampton SO16 6YD, UK.
wjt@soton.ac.uk

Nicholas C P Cross: Wessex Genomics Laboratory Service, Salisbury District Hospital, Salisbury SP2 8BJ, UK.
ncpc@soton.ac.uk

Abstract

To identify genetic variants that influence myeloproliferative neoplasm (MPN) phenotype, we undertook a two-stage case-only genome-wide association study using cohorts from the UK (including UK Biobank), Spain, Germany and Italy. MPN subtype [essential thrombocythemia (ET); polycythemia vera (PV)] were compared to each other, to healthy controls and stratified analyses was performed based on chromosome 9p aberrations, *JAK2* V617F mutation burden and sex. The ET versus PV analysis identified known associations: (i) at *HBS1L-MYB* that increased ET risk (P_{META}=7.93x10^{-6}, OR=1.28) and reduced PV risk (P_{META}=9.43x10^{-5}, OR=0.81) and (ii) at *GFI1B-GTF3C5* that predisposed to PV only (P_{META}=1.43x10^{-9}, OR=1.38). Two further linked intronic SNPs, rs2425786 and rs2425788, at *CDH22/CD40* were significant in females only (P_{META}=2.67x10^{-8}) with predisposition to PV (P_{META}=0.0006, OR=1.3) and reduction of ET risk (P_{META}=7.82x10^{-5}, OR=0.75). Associations with *JAK2*, *TERT*, *ATM*, *TET2*, *PINT*, *GFI1B* and *SH2B3* were confirmed (P_{META}<5x10^{-8}) and nine further loci were replicated (P_{META}<0.05). A polygenic risk score consisting of 48 SNPs from 31 loci demonstrated moderate discriminative performance for ET and PV (AUC=0.718) and was improved by optimization for disease subtype (AUC_{ET}=0.724 and AUC_{PV}=0.755). Overall, our results reveal that multiple germline variants influence MPN phenotype with *HBS1L-MYB* and a novel sex-specific association with *CDH22/CD40* being the strongest determinants.

Introduction

Common, low penetrance genetic variants contribute to the risk of developing MPN and also phenotypic pleiotropy in these disorders[1-10]. In a prior genome-wide association study (GWAS), we found that genetic variation at *MECOM*, *TERT*, *JAK2* and *HBS1L-MYB* predisposes to *JAK2*-unmutated MPN[11]. Targeted analysis of these four variants demonstrated that rs9376092 at *HBS1L-MYB* and the *JAK2* 46/1 haplotype specifically influence whether *JAK2* V617F mutated cases present with PV or ET. It is likely that variation at other loci influence MPN phenotype and the primary aim of this study was to identify inherited genetic factors on a genome-wide basis that influence whether *JAK2* V617F positive MPN patients present with polycythemia vera (PV) or essential thrombocythemia (ET). Secondary aims were to explore gender effects and the efficacy of phenotype-specific polygenic risk scores.

Methods

We performed a two-stage case-only GWAS with 556 ET and 556 PV patients at stage 1, all *JAK2* V617F positive. Selected SNPs were tested for replication in four independent *JAK2* V617F positive stage 2 cohorts (ET, n=703; PV, n=715) plus MPN cases from UK Biobank (ET, n=322; PV, n=506) (Supplementary Table 1). ET or PV cases were compared to healthy controls and stratified analyses was performed based on chromosome 9p aberrations, *JAK2* V617F variant allele frequencies (VAF) and sex. Final effect sizes and significance levels were estimated by meta-analysis. Detailed methods and expanded results are in the Supplementary Material.

Results and Discussion

After quality control, a total of 7,267,872 SNPs (658,066 observed, 6,609,806 imputed) and 1069 patients (535 ET and 534 PV) remained for analysis at stage 1 (Supplementary Figure 1, Supplementary Table 1). ET and PV cases were compared using logistic regression and the first five principal components from multidimensional scaling to correct for population stratification (Supplementary Figure 2). Twenty nine genome-wide significant SNPs were identified ($P<5x10^{-8}$), however all but two were linked to the 46/1 *JAK2* haplotype[8] (Supplementary Figure 3).

We selected 93 SNPs for replication in a case only analysis using binary logistic regression to compare ET and PV; final significance levels and effect sizes were determined by a fixed effects inverse variance-weighted meta-analysis which combined evidence from the two stages. Two linked SNPs ($r^2=0.91$) with genome-wide significance were identified in the *HBS1L-MYB* intergenic region, rs9399137 ($P_{meta}=2.28x10^{-10}$) and rs9376092 ($P_{meta}=4.35x10^{-9}$). SNPs at four additional loci (*ZBTB7C-CTIF*, *ADORA1*,

GFI1B-GTF3C5, *LINC02398*) were identified with suggestive levels of significance (Table 1, Supplementary Table 2).

To determine if these six SNPs associate with MPN subtype, we compared ET or PV cases from stage 1 and UK Biobank against healthy controls from the WTCCC2 (n=5,195) and UK Biobank (n=326,027) and combined the evidence using a fixed effects meta-analysis. As summarised in Table 1, the two *HBS1L-MYB* SNPs and *ADORA1* SNP were associated with an increased risk of ET and reduced risk of PV. In contrast, variation at *GFI1B-GTF3C5* was only associated with an elevated risk of PV and, consistent with this finding, was significantly associated with 9p chromosome aberrations and *JAK2* V617F VAF (see Supplementary Material). Finally, variation at *LINC02398* and *ZBTB7C-CTIF* was associated with an increased risk of PV, with the latter also associated with a reduced risk of ET. These findings indicate a multifactorial genetic influence of constitutional genotype on MPN phenotype. The most significant association for each SNP is summarised in Figure 1.

To investigate the possibility of sex differences in SNP-disease associations, ET and PV cases from stage 1 and UK Biobank were stratified by gender and analysed against each other and controls. Two linked SNPs (r^2=1.0) within *CDH22*, rs2425786 in intron 5 and rs2425788 in intron 4, were identified with genome-wide significance (rs2425786 P_{meta}=2.67x10^{-8}, rs2425788 P_{meta}=3.45x10^{-8}) (Table 1, Supplementary Figure 4). In comparison with healthy female controls, these SNPs were associated with a reduced risk of ET (rs2425786 P_{meta}=7.82x10^{-5}, OR=0.75; rs2425788 P_{meta}=0.0001, OR=0.75) and an elevated risk of PV (rs2425786 P_{meta}=0.0006, OR=1.30; rs2425788 P_{meta}=0.0006, OR=1.29). While sex-related differences have previously been reported in MPN[12,13] this represents the first instance of a sex-specific genetic association with phenotypic predisposition.

CDH22 encodes cadherin 22, which is essential for maintaining the structure and function of several tissues, including the hematopoietic microenvironment[14]. However, *CDH22* does not appear to be expressed in hematopoietic cells and eQTL analysis indicates that rs2425786 is associated with increased expression of the neighbouring gene *CD40* (P=3.80x10^{-7}; Supplementary Material and Supplementary Table 3). CD40 is expressed in hematopoietic cells and encodes a cell surface receptor belonging to the tumour necrosis factor receptor superfamily. Consequently, it is a potential candidate that merits further investigation.

The mechanism underlying the female-specific effect of rs2425786 is unclear, but it may involve hormonal influences, differential gene regulation, or sex-specific immune modulation. We used data

from UK Biobank to evaluate whether the effects of the *CDH22/CD40* SNPs were mediated by or interacted with hormonal biomarkers (sex hormone binding globulin [SHBG] and testosterone [TT]) or the inflammatory biomarker C-reactive protein (CRP). The SNPs were associated with a reduced risk of ET (rs2425786 P_{CRP}=0.0016, OR=0.69; rs2425786 P_{TT}=0.0051, OR=0.68) and an increased risk of PV (rs2425786 P_{CRP}=0.0264, OR=1.32; rs2425786 P_{TT}=0.0459, OR=1.31), independently of CRP (Supplementary Table 4) and testosterone (Supplementary Table 5), with no evidence of significant interactions. Adjustment for SHBG did not attenuate the SNPs associations for ET versus PV (rs2425786 P_{SHBG}=0.0017, OR=0.56) and ET versus controls (rs2425786 P_{SHBG}=0.0065, OR=0.71), and no significant interactions were observed (Supplementary Table 6). A similar trend towards increased risk of PV was shown, although it did not reach nominal significance (rs2425786 P_{SHBG}=0.0656, OR=1.27).

To further investigate potential sex-linked biological pathways, we reviewed phenome-wide association study results which revealed a significant association between rs2425786 and complications of labour and delivery (OR=0.95, P=1.48x10^{-4})[15], suggesting a possible link to female-specific physiological processes. Some genes are differently regulated in males and females due to differences in the epigenetic landscape. Interestingly, aberrant demethylation of the promoter region of *CD40LG*, which encodes the CD40 ligand, on the inactive X chromosome can lead to biallelic expression in females. This abnormal expression pattern has been linked to a higher prevalence of immune-related diseases[16,17] and elevated levels of IgM in females[18]. This female-specific mechanism may be relevant to the observed association between *CDH22/CD40* SNPs and increased risk of PV in women, and we plan to explore this using bulk and single cell methylation/expression analysis in relation to rs2425786 genotype and MPN phenotype

To estimate an individual's genetic risk for developing MPN, and specifically ET or PV, we calculated three polygenic risk scores (PRS_{MPN}, PRS_{ET}, PRS_{PV}) using 48 SNPs (Supplementary Table 7). The PRS_{MPN} exhibited moderate performance in UK Biobank, achieving an AUC value of 0.635 which increased to 0.718 when covariates for age, sex and ancestry (first 10 principal components) were included (Figure 2). Individuals with scores in the highest decile were estimated to have a 4.88-fold increased risk of MPN versus those in the lowest decile. The PRS_{ET} and PRS_{PV} showed a slight improvement with an AUC of 0.724 for ET and 0.755 for PV, respectively, when adjusting for covariates. The relative risk of disease associated with scores in the top versus bottom decile were 5.78 for ET and 4.66 for PV.

In a recent study, Guo et al 2024[19] showed that a PRS for platelet traits in healthy individuals (pct and plt) were associated with ET and that a PRS for red blood cells (hgb, hct, rbc and mchc) were risk

factors for PV. An additional PRS consisting of MPN-associated SNPs also increased the risk of ET and PV, but to a lesser extent. We computed PRS for the six blood cell traits using all available SNPs (Supplementary Table 8) and our tailored PRS (PRS_{ET} and PRS_{PV}) and assessed their relationship with ET and PV in UK Biobank. We confirmed the association of platelet traits with ET (PRS_{pct} $P_{fdr}=7.16x10^{-17}$, OR=1.63; PRS_{plt} $P_{fdr}=6.08x10^{-14}$, OR=1.54) and red blood cell traits with PV (PRS_{hgb} $P_{fdr}=5.09x10^{-16}$, OR=1.47; PRS_{hct} $P_{fdr}=1.58x10^{-13}$, OR=1.42; PRS_{rbc} $P_{fdr}=7.01x10^{-11}$, OR=1.38; PRS_{mchc} $P_{fdr}=1.17x10^{-3}$, OR=1.18) using univariable logistic regression (Supplementary Table 9). However, our tailored PRS had the strongest association with a diagnosis of ET (PRS_{ET} $P=1.92x10^{-16}$, OR=1.58) and PV (PRS_{PV} $P=7.62x10^{-18}$, OR=1.48) using multivariable logistic regression and correcting for either platelet traits with ET or red blood cell traits with PV along with age, sex, *JAK2* V617F VAF and 10 principal components (Supplementary Table 10).

According to the per allele odds ratio and minor allele frequency, rs2425786 (*CDH22/CD40)* is estimated to account for the largest proportion of the population attributable fraction (19.6%) followed by rs9399137 (*HBS1L-MYB*; 9.7%). The intergenic SNP between *GFI1B* and *GTF3C5*, rs3011271, accounts for a further 6.3% of the PAF. Based on a multiplicative model without interaction, these three genetic risk factors are estimated to have a combined PAF of 32% (Supplementary Table 11) indicating that they play a substantial role in influencing MPN phenotype.

Our findings highlight the importance of considering the possibility of gender-specific effects in studies that explore the connection between genetic variation and patient phenotype, and this may extend beyond presenting features to clinical management issues such as adverse events and outcomes following treatment.

Data Sharing statement

The array data generated in this study have been deposited at BioStudies (www.ebi.ac.uk/biostudies) under accession number S-BSST1772.

Acknowledgements

Part of this research has been conducted using the UK Biobank Resource under Application Number 35273.

Authorship Contributions

The study was designed and overseen by WJT and NCPC. Data analysis was performed by WJT and AAZD. JS and AJC prepared samples for genotyping. All other authors provided samples and/or clinical or laboratory data from their respective centers. The manuscript was drafted by WJT and all authors contributed to the final version.

Disclosure of Conflicts of Interest

None of the authors declare any relevant conflicts of interest

References

1. Bao EL, Nandakumar SK, Liao X, et al. Inherited myeloproliferative neoplasm risk affects haematopoietic stem cells. *Nature*. 2020;586(7831):769-775.
2. Loh PR, Genovese G, McCarroll SA. Monogenic and polygenic inheritance become instruments for clonal selection. *Nature*. 2020;584(7819):136-141.
3. Grinfeld J, Nangalia J, Baxter EJ, et al. Classification and Personalized Prognosis in Myeloproliferative Neoplasms. *N Engl J Med*. 2018;379(15):1416-1430.
4. Trifa AP, Bănescu C, Bojan AS, et al. MECOM, HBS1L-MYB, THRB-RARB, JAK2, and TERT polymorphisms defining the genetic predisposition to myeloproliferative neoplasms: A study on 939 patients. *Am J Hematol*. 2018;93(1):100-106.
5. Hinds DA, Barnholt KE, Mesa RA, et al. Germ line variants predispose to both JAK2 V617F clonal hematopoiesis and myeloproliferative neoplasms. *Blood*. 2016;128(8):1121-1128.
6. Jones AV, Campbell PJ, Beer PA, et al. The JAK2 46/1 haplotype predisposes to MPL-mutated myeloproliferative neoplasms. *Blood*. 2010;115(22):4517-4523.
7. Olcaydu D, Harutyunyan A, Jäger R, et al. A common JAK2 haplotype confers susceptibility to myeloproliferative neoplasms. *Nat Genet*. 2009;41(4):450-454.
8. Jones AV, Chase A, Silver RT, et al. JAK2 haplotype is a major risk factor for the development of myeloproliferative neoplasms. *Nat Genet*. 2009;41(4):446-449.
9. Kilpivaara O, Mukherjee S, Schram AM, et al. A germline JAK2 SNP is associated with predisposition to the development of JAK2(V617F)-positive myeloproliferative neoplasms. *Nat Genet*. 2009;41(4):455-459.
10. Pardanani A, Fridley BL, Lasho TL, Gilliland DG, Tefferi A. Host genetic variation contributes to phenotypic diversity in myeloproliferative disorders. *Blood*. 2008;111(5):2785-2789.
11. Tapper W, Jones AV, Kralovics R, et al. Genetic variation at MECOM, TERT, JAK2 and HBS1L-MYB predisposes to myeloproliferative neoplasms. *Nat Commun*. 2015;6:6691.
12. Moliterno AR, Braunstein EM. The roles of sex and genetics in the MPN. *Int Rev Cell Mol Biol*. 2022;366:1-24.
13. Karantanos T, Chaturvedi S, Braunstein EM, et al. Sex determines the presentation and outcomes in MPN and is related to sex-specific differences in the mutational burden. *Blood Adv*. 2020;4(12):2567-2576.
14. Cao ZQ, Wang Z, Leng P. Aberrant N-cadherin expression in cancer. *Biomed Pharmacother*. 2019;118:109320.
15. Canela-Xandri O, Rawlik K, Tenesa A. An atlas of genetic associations in UK Biobank. *Nat Genet*. 2018;50(11):1593-1599.
16. Liao J, Liang G, Xie S, et al. CD40L demethylation in CD4(+) T cells from women with rheumatoid arthritis. *Clin Immunol*. 2012;145(1):13-18.
17. Lian X, Xiao R, Hu X, et al. DNA demethylation of CD40l in CD4+ T cells from women with systemic sclerosis: a possible explanation for female susceptibility. *Arthritis Rheum*. 2012;64(7):2338-2345.
18. Lleo A, Liao J, Invernizzi P, et al. Immunoglobulin M levels inversely correlate with CD40 ligand promoter methylation in patients with primary biliary cirrhosis. *Hepatology*. 2012;55(1):153-160.
19. Guo J, Walter K, Quiros PM, et al. Inherited polygenic effects on common hematological traits influence clonal selection on JAK2(V617F) and the development of myeloproliferative neoplasms. *Nat Genet*. 2024;56(2):273-280.

Locus	SNP	Fixed effect meta-analysis*																	
		ET vs PV (6)		ET vs controls (2)		PV vs controls (2)		ET/PV vs controls (2)		9p aUPD/CNG vs controls (1)		*JAK2* V617F VAF (3)		ET vs PV females (2)		ET vs control females (2)		PV vs control females (2)	
		P	OR	P	OR	P	OR	P	OR	P	OR	P	BETA	P	OR	P	OR	P	OR
HBS1L-MYB	rs9399137	**2.28×10^{-10}**	1.47	7.93×10^{-6}	**1.28**	9.43×10^{-5}	**0.81**	0.2967	1.04	0.2928	0.90	0.0025	-0.111	1.99×10^{-7}	1.78	0.0001	1.31	0.0002	0.71
	rs9376092	**4.35×10^{-9}**	1.41	2.27×10^{-7}	**1.32**	0.0049	**0.86**	0.1609	1.06	0.5493	0.95	0.0043	-0.102	1.75×10^{-6}	1.67	1.08×10^{-5}	1.36	0.0089	0.80
ZBTB7C-CTIF	rs8087061	†**1.67×10^{-6}**	0.54	0.0028	**0.74**	0.0005	**1.31**	0.6180	1.03	0.0431	1.32	†0.0658	0.151	0.0086	0.61	0.0137	0.72	0.3855	1.12
ADORA1	rs3766568	†**3.99×10^{-5}**	1.34	0.0030	**1.17**	0.0031	**0.86**	0.8722	0.99	0.9935	1.00	†0.02302	-0.056	0.0042	1.34	0.0203	1.17	0.0719	0.86
LINC02398	rs2244740	**7.06×10^{-5}**	0.61	0.1106	0.80	0.0013	**1.38**	0.2538	1.10	0.0715	1.38	0.0721	0.133	0.1017	0.66	0.1938	0.80	0.3226	1.18
GFI1B-GTF3C5	rs3011271	4.77×10^{-5}	0.78	0.7076	1.02	**1.43×10^{-9}**	**1.38**	3.57×10^{-6}	1.21	3.44×10^{-9}	1.71	2.35×10^{-8}	0.207	0.0086	0.73	0.8167	1.02	0.0004	1.35
	rs520812	0.0111	0.83	0.8614	0.99	**1.22×10^{-6}**	**1.34**	0.0007	1.18	1.28×10^{-6}	1.63	0.0002	0.159	0.0251	0.73	0.5725	0.95	0.0317	1.24
FAM135B	rs12550019	0.0419	0.90	0.9332	1.00	**2.48×10^{-5}**	**1.22**	0.0018	1.12	0.0009	1.31	0.2214	0.039	0.1703	0.87	0.7929	1.02	0.0494	1.16
CDH22	rs2425786	†3.93×10^{-5}	0.75	0.0364	0.89	0.0024	1.15	0.3591	1.03	0.0998	1.15	†0.3328	0.045	**2.67×10^{-8}**	0.56	7.82×10^{-5}	**0.75**	0.0006	**1.30**
	rs2425788	†4.60×10^{-5}	0.75	0.0333	0.89	0.0030	1.15	0.3968	1.03	0.1124	1.14	†0.3667	0.042	**3.45×10^{-8}**	0.56	0.0001	**0.75**	0.0006	**1.29**

Table 1. Summary of the most significant SNPs following meta-analysis.

Locus, HGNC gene symbol with flanking genes shown for intergenic SNPs; SNP, rs identifier from dbSNP; Fixed effect meta-analysis was used to generate significance levels (P) and effect sizes (OR or BETA) except for 9p aUPD/CNG which was only available in the stage 1 case control cohort. Comparative groups or trait investigated are shown by column titles and the number of independent cohorts used for meta-analysis is shown in parentheses. †rs8087061 and rs3766568 failed replication QC (HWE $P<1E^{-10}$ and call rate <90% respectively) while the *CDH22* SNPs (rs2425786 and rs2425786) were not selected for replication genotyping. As a result, these SNPs are only tested in two cohorts for the ET vs PV analysis and one cohort for association with *JAK2* V617F. The most significant P-value across all analyses is highlighted in bold. Odds ratios (OR) in bold highlight the most significant subtype-specific associations in comparisons of either ET or PV cases with controls. SNPs associated with both subtypes have two bolded ORs, while those associated with only one subtype have a single bolded OR. Numbers in brackets indicate the number of cohorts tested for each comparison.

Figure legends

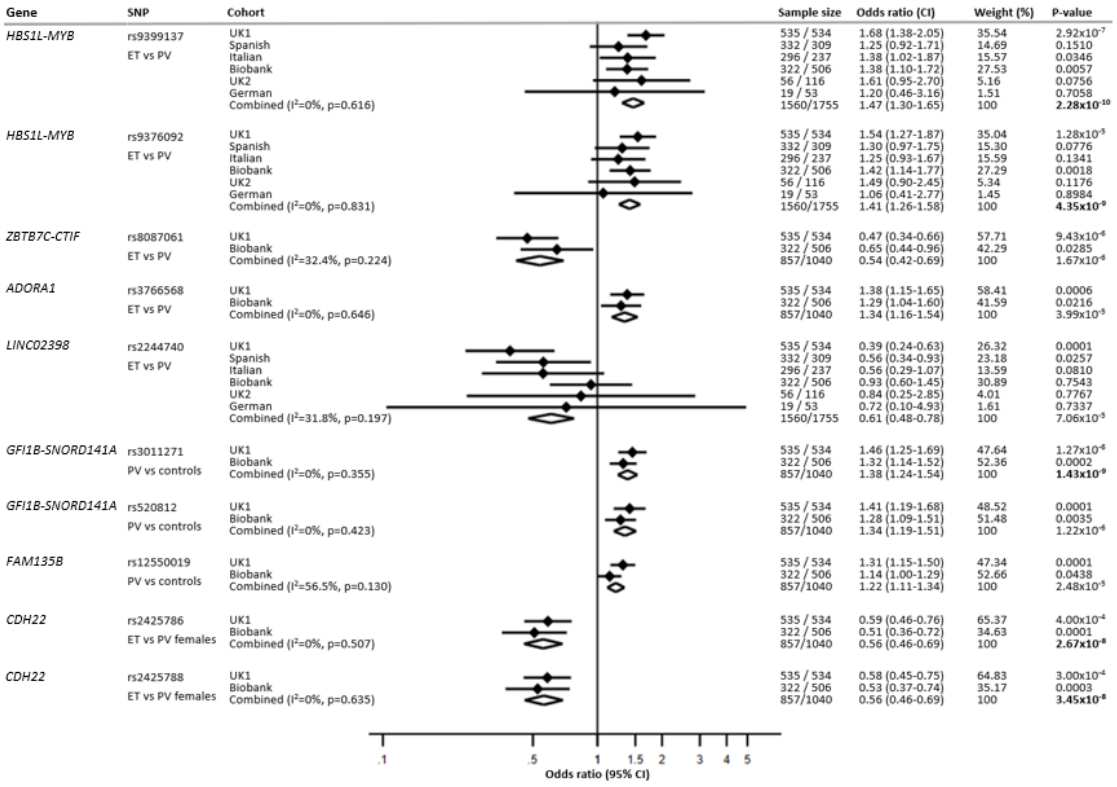


Figure 1. Forest plot and meta-analysis for the most significant SNPs. Forest plots showing the odds ratios, 95% confidence intervals (CI), percentage weight contributed to the overall meta-analysis and p-value for each SNP with or approaching a genome-wide level of significance. The most significant association for each SNP with a genome-wide or suggestive level of significance is shown. Odds ratios greater than 1 for the ET vs PV comparison indicate an increased risk of PV while those less than 1 increase the risk of ET. The SNP subtotals show the OR and CI for a fixed-effects meta-analysis; Cochran's Q test and I^2 statistics showed that for each SNP there was no evidence of heterogeneity between cohorts. Each SNP is significant in at least one of the replication cohorts tested and has evidence for the same trend in the remaining populations. GWAs significant P-values are highlighted in bold.

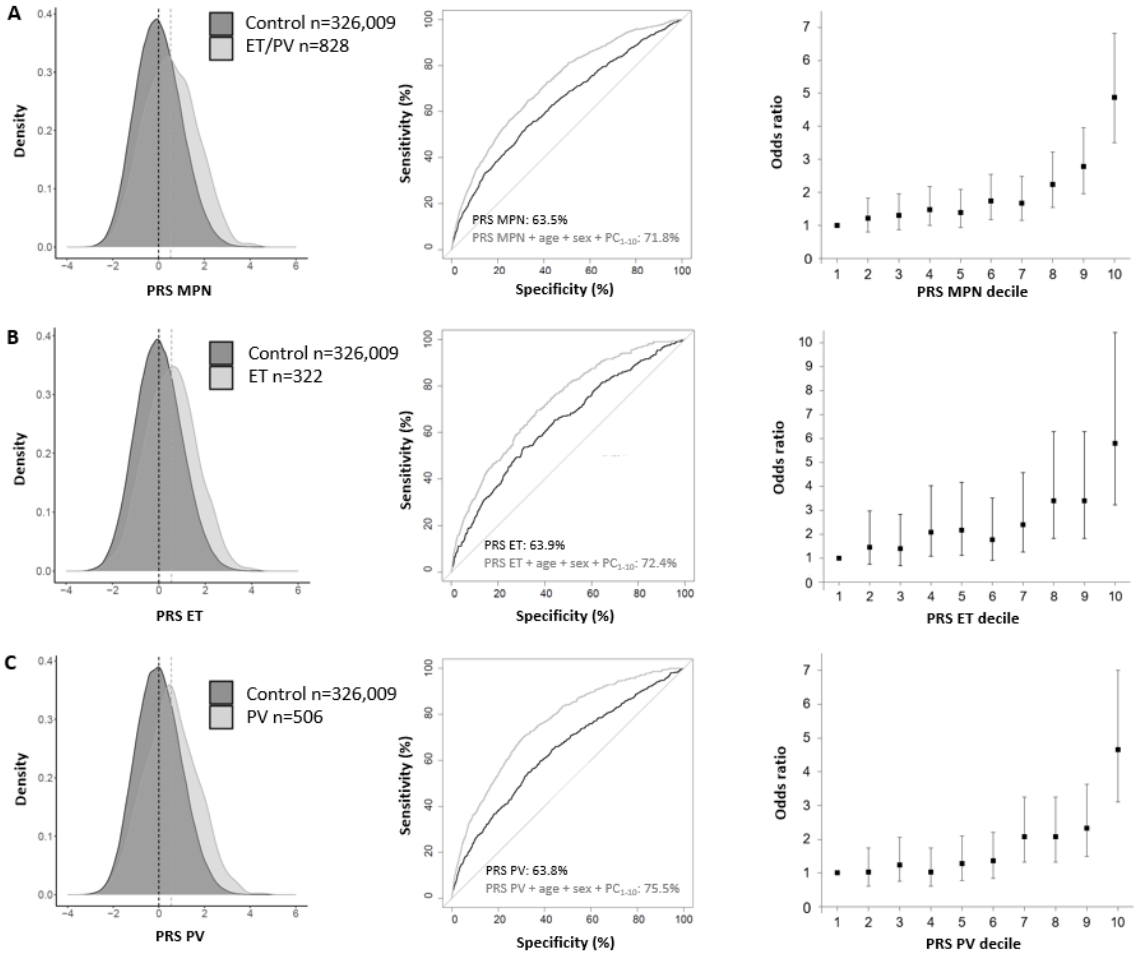


Figure 2. Evaluation of PRS optimised for disease subtype. Panels represent PRS optimised for ET and PV cases (A) ET cases (B) and PV cases (C). Density plots compare the distribution of Z-scaled PRS in cases and controls. Receiver operating characteristic curves showing the predictiveness of the PRS alone or with covariates (age, sex and first 10 principal components). Decile plots of relative disease risk in each decile versus the lowest decile.

SUPPLEMENTARY MATERIAL

Genome-wide analysis defines genetic determinants of MPN subtypes and identifies a sex-specific association at *CDH22/CD40*

William J Tapper, Ahmed Dawoud, Joannah Score, Andrew J Chase, E. Joanna Baxter, Joanne Ewing, Louise Wallis, Paola Guglielmelli, Dolors Colomer, Beatriz Bellosillo, Montse Gomez, Juan Carlos Hernández-Boluda, Carlos Besses, Francisco Cervantes, Steffen Koschmieder, Anthony R Green, Andreas Reiter, Alessandro Vannucchi, Claire Harrison, Nicholas C. P. Cross

Supplementary Methods

Patient cohorts

At stage 1, 556 ET and 556 PV patients, all of whom tested positive for *JAK2* V617F, were recruited from the United Kingdom, hereafter referred to as UK1. Many of these (n=447) were drawn from the UK Primary Thrombocythemia 1 study (PT-1) which includes newly diagnosed and previously treated patients who met the Polycythemia Vera Study Group (PVSG) criteria for ET [1]. The remaining stage 1 samples were recruited from multiple centres across the UK. Five independent cohorts of patients were used for replication from Spain (ET, n=332; PV, n=309), Italy (ET, n=296; PV, n=237), Germany (ET, n=19; PV, n=53) and two cohorts of UK samples (UK2: ET, n=56; PV n=116) and the UK Biobank (ET, n=322; PV, n=506) [2]. For the UK Biobank, diagnoses were based on the International Classification of Diseases (ICD) codes for ET (D47.3) and PV (D45) from the national cancer registry. Patient samples from Spain, Italy, Germany, UK1 and UK2 were all positive for *JAK2* V617F. For the UK Biobank cases, whole exome sequencing (WES) of a random subset of 200K participants and analysis using DeepVariant [3] and Mutect2 [4] indicated that only 20% (26/133) of ET cases and 29% (61/211) of PV cases were positive for *JAK2* V617F. The reasons for this unexpectedly low prevalence of *JAK2* V617F have been described in detail elsewhere[5]. All participants provided informed consent in accordance with local ethical approval and the principles of the Declaration of Helsinki and the study includes data conducted under UK Biobank reference number 35273.

To investigate the relationship with disease subtype and overall predisposition, the ET/PV patients from stage 1 and the UK Biobank were compared with a control cohort from the Wellcome Trust Case Control Consortium (WTCCC, n=5,200)[6] and controls from the UK Biobank (n=326,009) that were free from any cancer (past or present) and/or clonal hematopoiesis defined by mosaic chromosome abnormalities or somatic mutations, as previously described [7].

Genotyping

The stage 1 samples were genotyped for 964,193 SNPs using Illumina Human OmniExpressExome v1.2 BeadChips by Gen-Probe Life Sciences Ltd (Wythenshawe, UK) and genotypes were called using Illumina Genome Studio (GSGT, version 1.9.4). The replication cohorts from Spain, Italy, UK2 and Germany were genotyped for 99 SNPs by Kompetitive Allele Specific PCR (KASP) at LGC Genomics Limited (Hertfordshire, UK) using a fluorescence resonant energy transfer PCR-based assay. Samples from the UK Biobank study were genotyped for 805,426 SNPs using two similar microarrays, the UK BiLEVE array and UK Biobank axiom array, at Affymetrix laboratories (Santa Clara, USA). Genotypes were then called using Affymetrix power tools. Unselected controls from the WTCCC2 consortium

were previously genotyped for 954,144 SNPs using Illumina Human1-2M-Duo chips at the Wellcome Trust Sanger Institute and genotypes were called using the Illuminus software[8].

Quality control for genotyping

Standard GWAs quality control (QC) measures were applied to the genotypic data prior to analysis using PLINK[9]. For the stage 1 ET/PV patients, QC involved removing samples and SNPs with 10% or more missing genotypes and the removal of SNPs with a minor allele frequency less than 1% or extreme deviation from Hardy Weinberg Equilibrium (HWE, $P \leq 1x10^{-10}$) which is most likely to reflect poor genotyping rather than disease association. Samples were also excluded due to outlying autosomal heterozygosity (±3 SD from the mean, Supplementary Figure 1), if there was a discrepancy between reported and inferred sex based on X chromosome homozygosity or there was evidence for cryptic relatedness (PI_HAT ≥0.125) based on pairwise measures of identity by state (IBS) derived from a subset of 273,077 autosomal SNPs in linkage equilibrium ($r^2<0.5$).

QC of the WTCCC2 and UK Biobank data had already been performed and is described in detail elsewhere[2,10]. To maintain parity, the QC measures described above were also applied to the WTCCC2 data in addition to the QC that had already been performed. A less stringent HWE threshold was used ($P \leq 0.001$) because these control samples were randomly selected unlike the ET/PV cases. Samples with outlying autosomal heterozygosity were identified by visual inspection rather than deviation from the mean as sample outliers had already been removed by previous QC. To further control for poor genotyping, SNPs with allele frequencies that were significantly different between the two WTCCC2 cohorts (National Blood Service versus British Birth Cohort) were also removed. Additional QC measures were not applied to the UK Biobank data, as the standard thresholds are not suitable for its substantially larger sample size.

The replication data from LGC were quality controlled by excluding SNPs with a duplicate error rate of 10% or more, with 10% or more missing genotypes or extreme deviation from HWE ($P \leq 1x10^{-10}$).

Population stratification

Association analyses were corrected for population stratification using the first five principal components that were generated by a multidimensional scaling (MDS) analysis. This involved combining the samples and SNPs which passed QC in the UK1, Biobank and WTCCC2 cohorts with samples from three reference populations from the HapMap consortium (Caucasian, CEU, African, YRI and Asian ASI). A matrix of pairwise IBS values was calculated using a subset of 52,992 autosomal SNPs

that were directly genotyped in all samples and were in linkage equilibrium (r^2<0.5). The MDS analysis was performed using these IBS values and the results were inspected by plotting the first and second principal components (Supplementary Figure 2).

Imputation and further QC

To increase the resolution of the stage one data and to improve the overlap with the WTCCC2 cohort, additional SNPs were imputed using the Sanger imputation server[11] which uses EAGLE2 for pre-phasing into a panel of human haplotypes from the Haplotype Reference Consortium (HRC release 1.1) and positional Burrows-Wheeler transform (PBWT) for imputation. In preparation for imputation, genotypes were assigned relative to the positive strand and SNPs were removed if they were not genotyped in the HRC, had alleles that did not match the HRC, had a difference in MAF greater than 0.2 compared to HRC, were AT/GC SNPs with a MAF greater than 0.4 or were determined to be duplicates. Genotyping relative to the positive strand was verified by comparing the observed alleles with the reference genome sequence using bcftools[12]. To address potential errors due to differential genotyping of cases and controls[13], imputation was carried out on 474,386 SNPs that were directly genotyped in both ET and PV cases from the stage 1 cohort and WTCCC2 controls. The imputed genotypes were quality controlled by excluding SNPs with a posterior probability less than 0.8, MAF less than 1%, greater than 10% missing genotypes, or significant deviation from Hardy-Weinberg equilibrium ($P \leq 1x10^{-10}$ in ET/PV cases or $P \leq 0.001$ in controls).

Imputation and QC of the Biobank data was carried out by the UK Biobank using the HRC and UK10K haplotype resources, SHAPEIT3 for phasing[14] and IMPUTE4 for imputation[15]. The number of SNPs and samples removed by each of these QC steps along with the number of imputed SNPs in each cohort is shown in Supplementary Table 1.

Statistical analyses

Predisposition towards ET or PV

To identify SNPs that predispose to MPN subtypes, ET and PV patients were coded as cases and controls, respectively, and compared using logistic regression and an additive linear model. In the stage 1 and UK Biobank analyses, the first five genetic principal components from MDS were used to control for population stratification. Following comparison of ET and PV patients at stage 1 and in the 5 replication cohorts, a fixed effects inverse variance-weighted meta-analysis was performed to estimate the final significance and effect size using the metan module in STATA (version 15.0)[16]. This meta-analysis assumed that SNPs had one true effect size and that any difference in effect size

between cohorts were due to sampling error. Heterogeneity between studies was estimated using the χ^2-based Cochran's Q statistic and the I^2 statistic which describes the percentage of variation across studies that is due to heterogeneity rather than chance[17].

The relationship with disease subtype was determined by comparison of either ET or PV cases against healthy controls from stage 1 and the UK Biobank. The stage 1 and UK Biobank data were tested as separate cohorts with correction for population stratification and the evidence for association was combined using a fixed effects meta-analysis.

Sex stratified analysis

To explore potential sex differences in SNP-disease associations the ET and PV cases from stage 1 and the UK biobank were split by gender and compared against each other and against controls. The final effect sizes and significance levels were estimated by a fixed effect inverse variance-weighted meta-analysis.

To investigate the association between hormonal biomarkers (sex hormone binding globulin [SHBG] and testosterone [TT]), the inflammatory marker C-reactive protein (CRP), and disease risk, we applied logistic regression to evaluate both potential mediation effects and interactions. All mediation models were adjusted for the first five principal components and biomarker values that were normalised using inverse rank-based normalisation. Interaction terms were subsequently added to the models to assess potential interactions. The terms P_{CRP}, P_{TT} and P_{SHBG} indicate the P-value when adjusted for normalized CRP, TT and SHBG, respectively.

Association with myeloproliferative neoplasms

Predisposition to MPN was tested by pooling the ET and PV cases from stage 1 (n=1,069) and the UK Biobank (n=828) that passed QC and comparing them with healthy controls from the WTCCC2 (n=5,195) and UK Biobank (n=326,009) with correction for population stratification. The final effect sizes and significance levels were estimated by meta-analysis of results from stage 1 and the UK Biobank.

Chromosomal abnormalities identified by analysis of SNP array data

Regions of acquired uniparental disomy (aUPD) and copy number gains or losses were identified in the stage 1 ET/PV patients from the UK1 cohort using B allele frequency (BAF) segmentation[18]. First, raw input files were prepared for each sample containing the BAF, which measures the relative signal

strength of the A and B allele, and the log R ratio (LRR), which is a normalized measure of the overall signal intensity, for each SNP. For QC purposes, SNPs were removed if they were non informative (BAF >0.9 or BAF <0.1) or the absolute difference in BAF values between preceding or succeeding SNPs was greater than 0.6. Mirrored BAF (mBAF) values were calculated by reflection of BAF values at 0.5 and regions of allelic imbalance (AI) were identified using circular binary segmentation (CBS) to identify regions with similar allelic proportions that were above the default mBAF threshold (≥0.56). Finally, the AI regions were categorised as indels or copy number neutral aUPD according to their LRR values and plotted on a karyotype for visualisation (Supplementary Figure 5).

Association with mosaicism of chromosome 9p and V617F mutation burden

The ET and PV cases from stage 1 were split into groups with (n=348) and without (n=721) mosaicism of chromosome 9p, defined by aUPD or copy number gains, and compared with healthy controls from WTCCC2 (n=5,195) using logistic regression with correction for population stratification.

JAK2 V617F variant allele frequencies (VAF) were available for three cohorts; UK stage 1, Spain and Italy. The distribution of these raw mutation levels was skewed towards zero and were therefore normalised using Blom transformation. Association with the normalised *JAK2* V617F mutation burden was tested in pooled ET/PV patients and ET or PV patients alone using linear regression. Association evidence from the three cohorts was combined using a fixed effects meta-analysis.

Power calculations

The power to detect SNPs associated with MPN subtypes after QC in the stage 1 analysis (535 ET patients versus 534 PV patients) and the meta-analysis of stages 1 and 2 (1,560 ET patients versus 1,755 PV patients) was estimated using the genetic power calculator[19] under a multiplicative genetic risk model and a type 1 error rate of $5x10^{-8}$ (Supplementary Figure 6). A range of genotype relative risks (1.1-2.0), risk allele frequencies (MAF 0.05 - 0.4) were used to estimate power.

Selection of SNPs for replication

Since the stage 1 ET versus PV GWAS had less than 80% power to detect common SNPs with effect sizes less than 1.74 (Supplementary Figure 6a) and because GWAs are prone to false positives, which are likely to be among the most significant results, the following criteria were used to select SNPs for replication rather than significance alone. Firstly, an LD clumping method in PLINK[9] was applied to the summary statistics from the ET versus PV GWAS to make a shortlist of index SNPs (P<0.001) with support from correlated SNPs (SNPs r^2>0.5, within 500 kb and P<0.01). SNPs which had been directly

genotyped were then selected from these clumps with priority, but not exclusivity, given to SNPs that were either located in or flanked by a functionally relevant gene according to annotation from GeneAlacart (https://genealacart.genecards.org/) or had previously been associated with variation in platelet or erythrocyte counts[20]. For clumps with high significance and/or compelling functional evidence, an additional backup SNP was also selected as a precaution for failed genotyping. Using these criteria, a total of 93 SNPs were selected for genotyping at stage 2 of which 35 SNPs were selected based on significance alone, 54 SNPs were selected due to functional annotation and 4 were chosen as backups (Supplementary Table 2).

Replication

SNPs were considered replicated if they reached genome-wide significance ($P<5x10^{-8}$) or suggestive significance ($P<1x10^{-4}$) in any one of the joint meta-analyses and showed nominal significance ($P<0.05$) with the same direction of effect in at least one replication cohort. This approach ensured that associations were not driven solely by the discovery cohort and allowed inclusion of promising signals for further investigation and for polygenic risk score (PRS) construction.

Visualization of results

To examine the effectiveness of our QC measures and to assess the evidence for systematic bias in the ET versus PV analyses due to residual population stratification or imputation error, we used the qqnorm and qqplot procedures in R to construct quantile-quantile (QQ) plots of observed and expected P values under the null distribution (Supplementary Figure 7). Further visualization of the stage-1 results and to highlight SNPs selected for follow-up was performed using the qqman package[21] in R to construct a Miami plot (Supplementary Figure 3). Regional plots were generated using the LocusZoom software[22]. A forest plot of the most significant association result was produced for each SNP (Figure 1) using stata.

Functional annotation of variants

We explored the biological relevance of regions containing genome-wide significant SNPs using HaploReg (version 4.1)[23] to annotate the lead SNP and its proxies ($r^2>0.8$) with respect to histone modification, estimated pathogenicity using combined annotation-dependent depletion (CADD) scores[24], predicted effect on protein binding using RegulomeDB v2.2[25] scores (SNPs scoring ≤3 are likely to affect binding), and phenome-wide associations with multiple traits from FinnGen, the UK Biobank and NHGRI-EBI GWAS catalog[26] (Supplementary Table 3). Additionally, candidate regions were annotated against a 15-state chromatin model[27] in primary hematopoietic stem cells (E035) and

a chronic myeloid leukemia cell line (K562). This model categorizes non-coding DNA into active or repressed states that are respectively enriched and depleted for phenotype-associated SNPs[28]. To gain further functional insight, expression quantitative trait loci (eQTL) analyses were performed on the lead SNP and its proxies using eQTLGen and blood derived expression levels[29].

Replication of previous findings and polygenic risk score construction

Risk SNPs from previous studies[30-55], or proxies in strong LD (r^2>0.9), were tested for replication (P<0.05) in our stage 1 GWAS of MPN, ET and PV (Supplementary Table 12). To select SNPs for PRS construction, an LD clumping method[9] was applied to the stage 1 GWAS of MPN, ET and PV with an r^2 threshold of 0.2, a P-value threshold of 0.05 and a physical threshold of 1Mb. The most significant index SNP from each clump containing a published risk SNP which replicated in one or more of our stage 1 analyses was selected for PRS construction (Supplementary Table 7). Three additional SNPs were selected based on genome-wide significance in either the ET versus PV meta-analysis excluding the UK Biobank or stage 1 GWAS of 9p aUPD/CNG versus controls. Three PRS were calculated according to the sum of an individual's risk alleles weighted by the allele effect size from either the MPN (PRS_{MPN}), ET (PRS_{ET}) or PV (PRS_{PV}) GWAS and using the --score function in PLINK.

PRS performance was assessed in the UK Biobank using logistic regression with either ET (n=322), PV (n=506) or PV plus ET coded as cases and compared with controls (n=326,009) along with PRS, age, sex and the top 10 principal components as covariates. Area under the receiver-operator curve (AUC) was calculated for each PRS using the pROC package in R. To investigate how the predicted risk of disease varied with increasing PRS, we performed decile analyses where samples in the lowest PRS quintile were treated as a reference and compared with participants from the other deciles using Chi-square tests to determine the risk of MPN, ET or PV in each decile versus the reference.

PRS previously associated with ET [plateletcrit (pct) and platelet count (plt)] and PV [hemoglobin concentration (hgb), hematocrit (hct), red blood cell count (rbc) and mean corpuscular hemoglobin concentration (mchc)][5,20] were calculated using the published effect sizes and all available SNPs (Supplementary Table 8). The association of these blood trait PRS and our MPN PRS (PRS_{MPN}, PRS_{ET} and PRS_{PV}) with a diagnosis of either ET or PV was assessed in the UK Biobank using multinomial logistic regression to correct for age, sex, *JAK2* V617F VAF and 10 principal components.

To assess the impact of our findings at a population level, we computed population attributable fraction (PAF) which estimate the proportion of ET and PV cases that can be attributed to each

polymorphism. We calculated this as, $PAF = \frac{p(OR-1)}{p(OR-1)+1}$ where p is the frequency of the risk allele and OR is the per allele odds ratio[56]. A combined PAF was also calculated as PAF=1-(1-PAF_1)(1-PAF_2)…(1-PAF_n) incorporating all risk factors.

Expanded Results

Case only Genome Wide Association study of ET versus PV

After quality control at stage 1, a total of 7,267,872 SNPs (658,066 observed and 6,609,806 imputed) and 1069 *JAK2* V617F positive samples (535 ET and 534 PV) remained for analysis (Supplementary Figure 1 and Supplementary Table 1). To identify SNPs associated with MPN subtype, the ET and PV patients were compared using logistic regression and the first five principal components from multidimensional scaling (MDS) to correct for population stratification (Supplementary Figure 2). A quantile-quantile plot (QQ) of the stage 1 results yielded a low genomic inflation factor (λ=1.028) and showed similar observed and expected P-values until values less than $1x10^{-3}$ which began to deviate from the null hypothesis (Supplementary Figure 7). This suggests that systematic biases such as population stratification are unlikely to contribute to SNP significance. A Miami plot of the stage 1 results is shown in Supplementary Figure 3.

A total of 29 SNPs were identified with genome-wide significance ($P<5x10^{-8}$). However, all but two of these SNPs were located on the p-arm of chromosome 9 with a peak of significance in the promoter of *JAK2* which is associated with PV (rs10758669 $P=5.2x10^{-12}$, OR=0.55). This large region of significant SNPs corresponds to the 46/1 *JAK2* haplotype[52] and is accentuated by aUPD or copy number gains of chromosome 9p which occurred in 60% of PV cases (321/534) versus 6.5% of ET cases (35/535) (Supplementary Figure 5) in agreement with previous reports[57].

According to the number of samples tested and assuming a multiplicative disease model with a type 1 error rate of $5x10^{-8}$, the stage 1 analysis was estimated to have limited statistical power to detect SNPs with typical effect sizes (RR≤1.2; Supplementary Figure 6). Therefore, we used a combination of significance, LD clumping and functional evidence to select 93 SNPs with P<0.001 for replication (Supplementary Table 2). These SNPs were selected to have support from at least one significant correlated SNP ($r^2>0.5$, within 500 kb and P<0.01) and were either (i) the most significant (n=35), (ii) were located in close proximity to a functionally relevant gene (n=54) or (iii) were selected as backups for the most promising signals (n=4). None of the GWAs significant SNPs on chromosome 9p were selected for replication as the role of *JAK2* is well established[52]. The two remaining GWAs significant SNPs, rs17876031 and rs313039, did not meet the selection criteria, as they lacked support from

nearby linked SNPs (r^2>0.5, within 500 kb and P<0.01) and were not associated with functionally relevant genes. Consequently, these SNPs were not selected for replication.

Replication and validation of candidate SNPs

Of the 93 SNPs selected for replication, 83 were successfully genotyped and passed QC in 1,065 ET patients and 1,221 PV patients from four independent cohorts (UK, Spain, Italy, Germany) (Supplementary Table 2). In addition, all 93 SNPs were available in the UK Biobank which had been previously genotyped using SNP arrays. SNPs were tested for association in a case only analysis using binary logistic regression to compare ET and PV patients. The final significance levels and effect sizes were determined by a fixed effects inverse variance-weighted meta-analysis which combined evidence from stages 1 and 2. Two linked SNPs (r^2=0.91) with genome-wide significance were identified in an intergenic region between *HBS1L* and *MYB*, rs9399137 (P_{meta}=2.28x10^{-10}) and rs9376092 (P_{meta}=4.35x10^{-9}) (Table 1). Four further SNPs were identified with suggestive levels of significance, rs8087061 in an intergenic region between *ZBTB7C* and *CTIF* (P_{meta}=1.67x10^{-6}); rs3766568 in intron 3 of *ADORA1* (P_{meta}=3.99x10^{-5}); rs3011271 in an intergenic region between *GFI1B* and *GTF3C5* (P_{meta}=4.77x10^{-5}); rs2244740 in intron 2 of *LINC02398* (P_{meta}=7.06x10^{-5}). Results from the analysis of stages-1 and 2 for all SNPs tested are shown in Supplementary Table 2.

Subtype-specific associations compared with healthy controls

To determine if the six SNPs with suggestive or genome-wide significance were associated with a particular disease subtype, we compared the ET or PV cases from stage 1 and the UK Biobank against healthy controls from the WTCCC2 (n=5,195) and the UK Biobank (n=326,027) and combined the evidence using a fixed effects meta-analysis. All analyses yielded minimal inflation factors although these were slightly higher in the stage 1 analyses versus the UK Biobank (average lambda GC of 1.037 vs 0.939) which is likely due to technical differences between the stage 1 case-control data (Supplementary Table 13). The two *HBS1L-MYB* SNPs were associated with an increased risk of ET (rs9399137 P_{meta}=7.93x10^{-6}, OR=1.28; rs9376092 P_{meta}=2.27x10^{-7}, OR=1.32) and a reduced risk of PV (rs9399137 P_{meta}=9.43x10^{-5}, OR=0.81; rs9376092 P_{meta}=0.0049, OR=0.86) (Table 1). Similarly, rs3766568 was associated with an increased risk of ET (P_{meta}=0.003, OR=1.17) and reduced risk of PV (P_{meta}=0.0031, OR=0.86). In contrast, rs3011271 (*GFI1B-GTF3C5*) was only associated with an elevated risk of PV which attained genome-wide significance in comparison with healthy controls (P_{meta}=1.43x10^{-9}, OR=1.38). Similarly, rs2244740 (*LINC02398*) was associated with an increased risk of PV only (P_{meta}=0.0013, OR=1.38). Finally, rs8087061 (*ZBTB7C-CTIF* SNP), was associated with an increased risk of PV (P_{meta}=0.0005, OR=1.31) and reduced risk of ET (P_{meta}=0.0028, OR=0.74).

When comparing the ET or PV cases against controls, two additional SNPs were identified with suggestive significance that were associated with an increased risk of PV only; rs520812 (P_{meta}=1.22x10^{-6}, OR=1.34) which is in moderate LD with rs3011271 (r^2=0.67) and rs12550019 (P_{meta}=2.48x10^{-5}, OR=1.22) located in intron 9 of *FAM135B* (Table 1).

Stratified analyses for 9p aUPD/CNG and *JAK2* V617F VAF

To gain further understanding of the genetic associations with disease subtype, we stratified the PV and ET cases from stage 1 into those with (n=348) and without (n=721) aUPD or CNG involving chromosome 9p and compared them with healthy controls from WTCCC2 (Supplementary Figure 5 and Supplementary Table 2). This showed that the association with rs3011271 was restricted to cases with mosaicism of chromosome 9p (P=3.4x10^{-9}, OR=1.71 in cases with aUPD or CNG; P=0.67, OR=0.97 in cases without aUPD or CNG) (Table 1). No SNPs reached genome-wide significance in patients without chromosome 9p UPD/CNG. We next investigated association with the allele burden of *JAK2* V617F which tends to be higher in patients with PV and/or more symptomatic disease[58]. *JAK2* V617F levels determined by qPCR were normalised and tested for association in the stage 1, Spanish and Italian cohorts using linear regression and a fixed effects meta-analysis to combine evidence. The only variants associated with elevated *JAK2* V617F were the two *GFI1B-GTF3C5* SNPs (rs3011271 P_{meta}=2.35x10^{-8}, β=0.21; rs520812 P_{meta}=0.0002, β=0.16) which support their association with PV (Table 1). The two *HBS1L-MYB* SNPs were associated with lower levels of *JAK2* V617F although this trend failed to reach nominal significance in the replication cohorts (rs9399137 P_{meta}=0.0025, β=-0.11; rs9376092 P_{meta}=0.0043, β=-0.10).

Development of PRS for ET and PV

To estimate an individual's genetic risk for developing MPN, or specifically ET versus PV, we applied an LD clumping method[9] to the stage 1 GWAS with an r^2 threshold of 0.2, a p value threshold of 0.05 and a physical threshold of 500kb. The most significant index SNP from each clump containing a published risk SNP or a proxy in strong LD (r^2>0.9) which replicated in one or more of our stage 1 analyses was selected for PRS construction. Additional SNPs were selected based on genome-wide significance in a meta-analysis of ET versus PV cases excluding the UK Biobank or the stage 1 GWAS of ET/PV cases with 9p aUPD or CNG versus controls. A total of 48 SNPs were selected including 46 from previous publications, 1 from the ET versus PV meta-analysis and 1 associated with 9p aUPD (Supplementary Table 7). Most published SNPs were represented by a linked index SNP that was more significant in our analysis (n=37/46).

Replication of previously reported SNP associations

Previous studies have identified 152 SNPs as risk factors for MPN including 104 SNPs with P-values less than $1x10^{-5}$ that were used to construct a polygenic risk score (PRS)[30-54] (Supplementary Table 4). Collectively, these SNPs represent 69 loci when SNPs in nearby genes and along chromosome 9p, which is targeted by aUPD, are merged. One hundred and twenty nine of these SNPs from 56 loci were evaluable in our stage 1 data, specifically those with MAF greater than 1% that were either directly genotyped (n=33), imputed (n=91) or assessed via a proxy (r^2>0.9, n=4). These 129 SNPs included 11 from 4 loci that were previously reported to increase the risk of either ET or PV[33,34,37,40,50,54]. We tested these SNPs for subtype specific effects and for association with MPN by pooling our ET and PV cases from stage 1 and comparing them with the WTCCC2 controls. A total of 85 SNPs from 23 loci were replicated (P<0.05) for association with either MPN (n=78 SNPs), PV (n=69 SNPs) or ET (n=54 SNPs) (Supplementary Table 12 and Supplementary Figure 8A). Nine of the eleven SNPs with prior evidence for subtype specific effects were replicated. However, only three of these were associated with a particular subtype, rs2236496 and rs10758658 with PV and rs318699 with ET and the associated subtypes were not concordant with the previous findings. We found no evidence that previously reported SNPs at *AKIP1, F2R, MRPS31,* and *NR3C1*[30,33,47] influence the development or phenotype of MPN. *CHEK2* 1100delC has also been reported to predispose to MPN [36] but this was not genotyped in our stage 1 data or imputed as it not included in the HRC imputation panel.

Several additional signals were identified when comparing MPN cases against controls but only one of these, rs67876368 (P_{stage1}=1.51x10^{-12}, OR=1.93), was replicated in the UK Biobank (P_{stage2}=0.036, OR=1.29) and retained genome-wide significance following meta-analysis (P_{meta}=5.83x10^{-12}, OR=1.66) (Supplementary Figure 8A and B). rs67876368 is an intronic SNP within *LINC01340* and is associated with lower expression of *ERAP2* (*P*=1.45x10^{-6}, Z_{score}=-4.82) and higher expression of *RIOK2* (*P*=4.16x10^{-6}, Z_{score}=4.60). *LINC01340* is a non-protein coding RNA gene which is expressed in multiple tissues including bone marrow[59] but has not been implicated in MPN. Likewise, *ERAP2* has not been linked to MPN. *RIOK2* is located 323kb upstream of rs67876368 and encodes a serine threonine kinase which plays an important role in the maturation of the 40S ribosomal subunit[60]. In a recent study, Ghosh et al identified *RIOK2* as a key transcription factor that drives erythroid differentiation and suppresses megakaryopoiesis and myelopoiesis in primary human stem and progenitor cells[61].

***In silico* functional inference**

To explore the functional relevance of these associations, we used HaploReg and RegulomeDB to determine if the risk SNP or their proxies (r^2 ≥0.8) were located in regions with potential regulatory functions based on chromatin modification, DNA methylation and alteration of transcription factor (TF) binding motifs (Supplementary Table 3). To gain further functional insight, we annotated these SNPs for phenome-wide associations and performed eQTL analysis using eQTLGen[29].

The most significant SNP, rs9399137 (P_{meta}=2.28x10^{-10} for ET versus PV), is located in an intergenic region between *HBS1L*, which encodes a member of the GTP-binding elongation factor family, and *MYB*, which encodes a transcriptional regulator that plays an essential role in the regulation of hematopoiesis (Supplementary Figure 9). The RegulomeDB score for rs9399137 (1d) indicates that it has strong evidence for functional affects based on TF binding, altered sequence motifs, eQTL and chromatin accessibility (Supplementary Table 3). The risk allele (C) significantly reduces the TF binding affinity of a forkhead box C1 motif (log-odds from 7.3 to -4.3) and is associated with lower expression of *HBS1L* (P=5.2x10^{-50}, Z_{score}=-14.9), *MYB* (P=1.6x10^{-8}, Z_{score}=-5.6) and *ALDH8A1* (P=5.0x10^{-132}, Z_{score}=-24.4) which is located 10.3kb downstream from *HBS1L*. rs9399137 is predicted to be deleterious (CADD=12.3) and has been associated with several hematological traits in healthy individuals, including higher platelet count (P=2.2x10^{-308}, β=0.12) and lower red blood cell count (P=2.2x10^{-308}, β=-0.18) which ties-in with our finding of a 1.28 fold increased risk of ET (P_{meta}=7.93x10^{-6}) and 0.81 fold reduced risk of PV (P_{meta}=7.93x10^{-6}).

The second most significant SNP, rs3011271 (P_{meta}=1.43x10^{-9}, OR=1.38 for PV versus controls), is located in an intergenic region between *GFI1B*, which encodes a transcriptional regulator that is expressed in hematopoietic cells, and *GTF3C5* encoding General Transcription Factor IIIC Subunit 5 (Supplementary Figure 9). Although this SNP is not predicted to be deleterious (CADD=0.07) and has weak evidence for functional affects according to RegulomeDB (score=5), the surrounding chromatin is characterised as an enhancer in primary hematopoietic stem cells (E035) and the SNP has been associated with lower levels of several hematological traits in healthy individuals including red blood cell counts (P=9.99x10^{-6}, β=-0.01) (Supplementary Table 3). Furthermore, the SNP is tightly linked (r^2>0.9) to three other SNPs with significant evidence for functional effects including rs1755624 (RegulomeDB score=1b) which alters several TF binding motifs (BCL, ERalpha-a, Ets, Irf, Rad21, SMC3, SP1 and TATA) and is in a region of CEBPB binding in K562 cells according to ChIP-Seq experiments[62]. Using eQTLGen we found that rs1755624 had the strongest association with increased expression of *GFI1B* (P=3.02x10^{-5}, Z_{score}=4.2) (Supplementary Table 3).

The third and final SNP reaching genome-wide significance, rs2425786 (P_{meta}=2.67x10^{-8}, for ET versus PV in females only), was found to increase the risk of PV (P_{meta}=0.0006, OR=1.3) and reduce the risk of ET (P_{meta}=7.82x10^{-5}, OR=0.75). This SNP is located in intron 5 of *CDH22* (Supplementary Figure 9), a gene which encodes a calcium-dependent cell adhesion molecule that plays an important role in the development and functional regulation of the nervous system, brain, heart, skeletal muscles, blood vessels and hematopoietic microenvironment[63]. Its role in non-neural tissues is less well understood but alterations in cadherin function are known to be involved in cancer, tumour invasion and metastasis[63]. The RegulomeDB score for rs2425786 (score=4) indicates weak evidence for functional consequences but the surrounding chromatin is characterised as an enhancer (7_Enh) in K562 cells and the SNP alters several TF binding motifs (log-odds from -3.4 to 8.5). There are no data relating rs2425786 to expression of *CDH22*, but it is associated with increased expression of the neighbouring gene *CD40* (P=3.80x10^{-7}, Z_{score}=5.1) according to eQTLGEN (Supplementary Table 3).

Extended discussion

The primary aim of this GWAS was to identify common polymorphisms associated with MPN phenotype (ET or PV) that would shed light on the genetic basis for these two closely related MPNs, and in particular why some patients with *JAK2* V617F present with ET whereas others present with PV. Previous studies have tended to perform pooled or stratified analysis which could miss variants with opposite effects in ET and PV. By directly comparing ET and PV cases and comparing them separately with controls we have identified three loci with strong evidence for subtype specific effects. Rs9399137, located in the intergenic region between *HBS1L* and *MYB*, was associated with a 1.28 fold higher risk of ET and a 0.81 fold lower risk of PV. Rs3011271 in the intergenic region between *GFI1B* and *GTF3C5* was associated with a 1.38 fold increased risk of PV only and was particularly significant in patients with aUPD or CNG of chromosome 9p. A final SNP, rs2425786 in intron 5 of *CDH22*, was associated with a 1.3 fold increased risk of PV and 0.75 fold reduced risk of ET, but only in females indicating a sex-specific genetic influence on MPN phenotype.

The association between MPN phenotype and variation at *HBS1L-MYB* has been reported previously in a targeted analysis of loci that predispose to *JAK2* unmutated MPN,[42] but the current study confirms the association at a genome-wide level. MYB plays a key role in gene regulation throughout hematopoiesis and is critical for the maintenance of normal hematopoietic stem cells. The risk allele for rs9399137 is associated with reduced expression of *MYB*, as well as *HBS1L* and *ALDH8A1*. Previous studies in mice have shown that point mutations or knockdown of c-Myb which partially disable its

function or lower its expression cause an increase in platelet numbers[64-66] and result in an age-related transplantable myeloid disorder with thrombocytosis and splenomegaly[67,68]. Reduced MYB contributes to ET via several mechanisms; by enhancing megakaryocyte production and, consequently, overproducing platelets, by dysregulating the cell cycle, particularly in hematopoietic progenitor cells, contributing to the excess proliferation by disrupting apoptosis leading to the survival and further accumulation of myeloid cells and by altering proteosomal activity which may link to stem cell aging[68]. Lower production of c-Myb has also been shown to inhibit the transition of uncommitted progenitor cells to erythropoiesis and slow the progression of early committed erythroid progenitors[69]. These dual effects which promote the proliferation of platelets and inhibit the production erythrocytes help to explain how the lower expression of *MYB*, *HBS1L* and *ALDH8A1* associated with rs9399137 may increase the risk of developing ET and reduce the risk of developing PV.

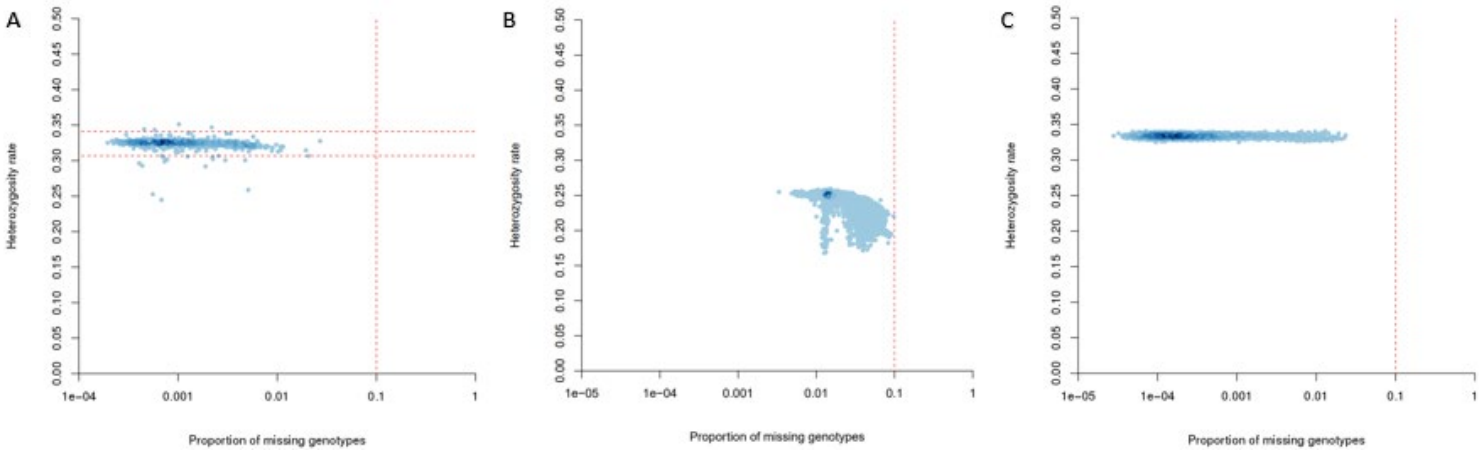
GFI1B is another transcription factor that plays a crucial role in the normal development and function of megakaryocytes and erythrocytes. Knockdown of *GFI1B* has indicated that it promotes erythroid differentiation by repressing *TGFBR3* expression and promoting the interaction between Smad2 and TIF1-γ which allows immature progenitors to differentiate towards the erythroid lineage[70]. The risk SNP identified by this study, rs3011271, is associated with increased expression of *GFI1B* which points towards a possible mechanism involving enhanced erythropoiesis and altered megakaryocyte function. An association with *GFI1B* was first noted by Hinds et al 2016[36] and later replicated by Bao et al 2020[30] but both of these studies focused on its association with MPN rather than PV specifically and may have underestimated its effect size. In this study, rs3011271 is associated with a 1.38 fold increased risk of PV only whereas the previous risk estimates were 1.2 for rs1633768[30] and 1.35 for rs621940[36].

CDH22 encodes cadherin 22, which is essential for maintaining the structure and function of several tissues, including the hematopoietic microenvironment[63]. In the context of MPNs, alteration in cadherin function might influence how hematopoietic progenitors interact with their microenvironment, potentially promoting the abnormal expansion of certain cell lineages. However, the risk SNP within *CDH22*, rs2425786, was found to be associated with an increased expression of the neighbouring gene encoding CD40 which plays a potentially significant role in PV. In contrast to *CDH22* which does not appear to be expressed in hematopoietic cells, *CD40* is expressed in several hematopoietic cell types. *CD40* encodes a cell surface receptor which belongs to the tumour necrosis factor receptor super family and is a potential candidate to explain the genetic association. CD40

interacts with its ligand CD40L (CD154), which is typically expressed on activated T cells. This interaction is crucial for immune cell activation and can lead to the production of inflammatory cytokines[71]. In PV, the CD40-CD40L interaction may amplify inflammatory responses within the bone marrow microenvironment, potentially contributing to the abnormal expansion of hematopoietic cells[72].

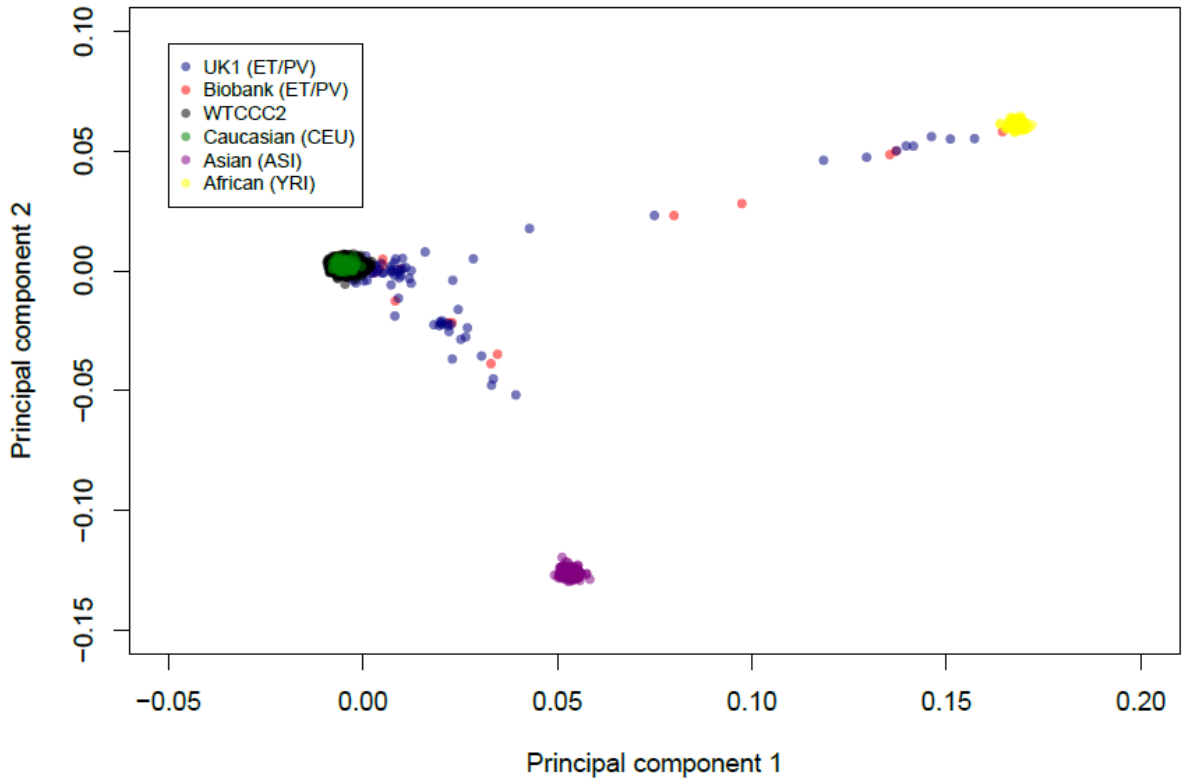
Overall, rs2425786, rs9399137, rs3011271 are estimated to have a combined PAF of 32% indicating that they play a substantial but incomplete role in influencing MPN phenotype. Fine mapping of these regions could help uncover some of the missing heritability, either by pinpointing causal variants with stronger associations or by revealing additional signals within these loci. However, it is likely that other loci with comparable or smaller effects also contribute to MPN predisposition, which may be identified through larger follow-up studies and stratification by disease subtype and sex. Our analysis confirmed previous findings which showed that PRS for platelet traits and red blood cell traits in the general population are associated with ET and PV respectively. However, the disease-specific PRS we developed in this study have a stronger association with the respective conditions. This suggests that the tailored PRS incorporate genetic variants which are relevant to the disease-specific mechanisms of ET and PV while the platelet and red blood cell trait PRS reflect broader genetic influences on blood production and function. Integrating these PRS into a comprehensive risk model that includes somatic genetic and non-genetic factors may improve their accuracy and potentially help in stratifying patients for more frequent monitoring, for example those with *JAK2* mutated clonal hematopoiesis and/or borderline platelet or red blood cell counts who do not yet meet the diagnostic criteria for ET or PV.

Supplementary Figure 1. Quality control for autosomal heterozygosity and per sample missingness



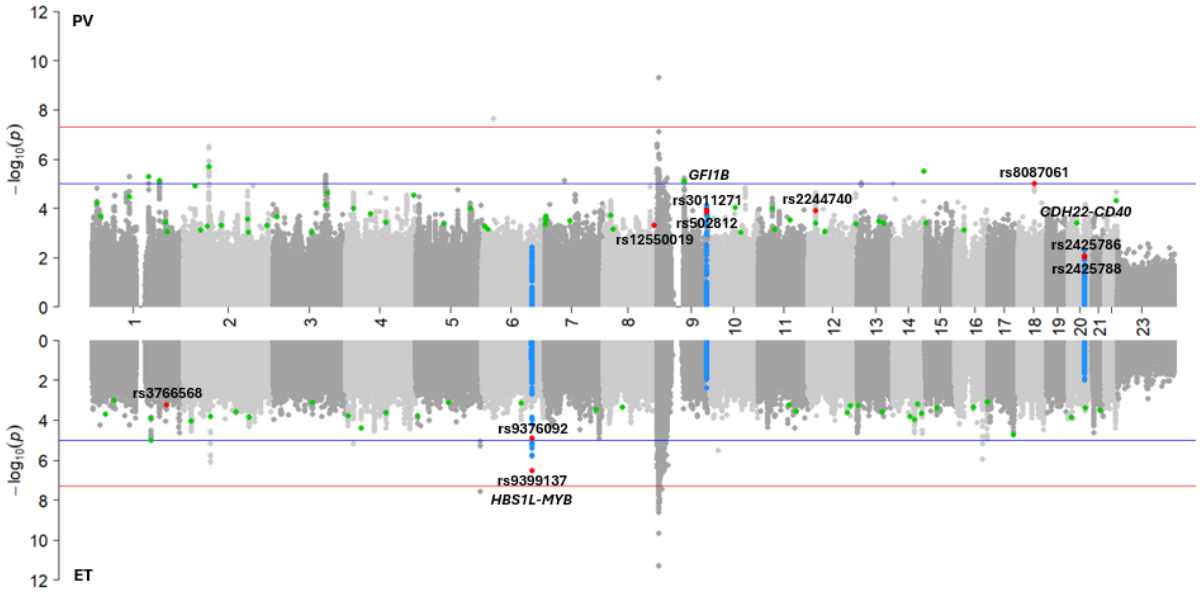
A. Stage 1 ET and PV patients from the UK1 cohort. **B.** ET and PV patients from the UK Biobank cohort. **C.** Healthy controls from the WTCCC2 cohort. Horizontal dashed lines indicate the thresholds used to identify samples with outlying levels of heterozygosity in the UK1 cohort (±3 SD from the mean). Vertical dashed lines show the threshold used to remove samples with more than 10% missing genotypes. Samples from the UK Biobank have lower mean heterozygosity because the Axiom array used for genotyping is enriched for rare variants which are likely to be homozygous in most samples (http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/UK-Biobank-Axiom-Array-Content-Summary-2014.pdf).

Supplementary Figure 2. Multidimensional scaling (MDS) plot



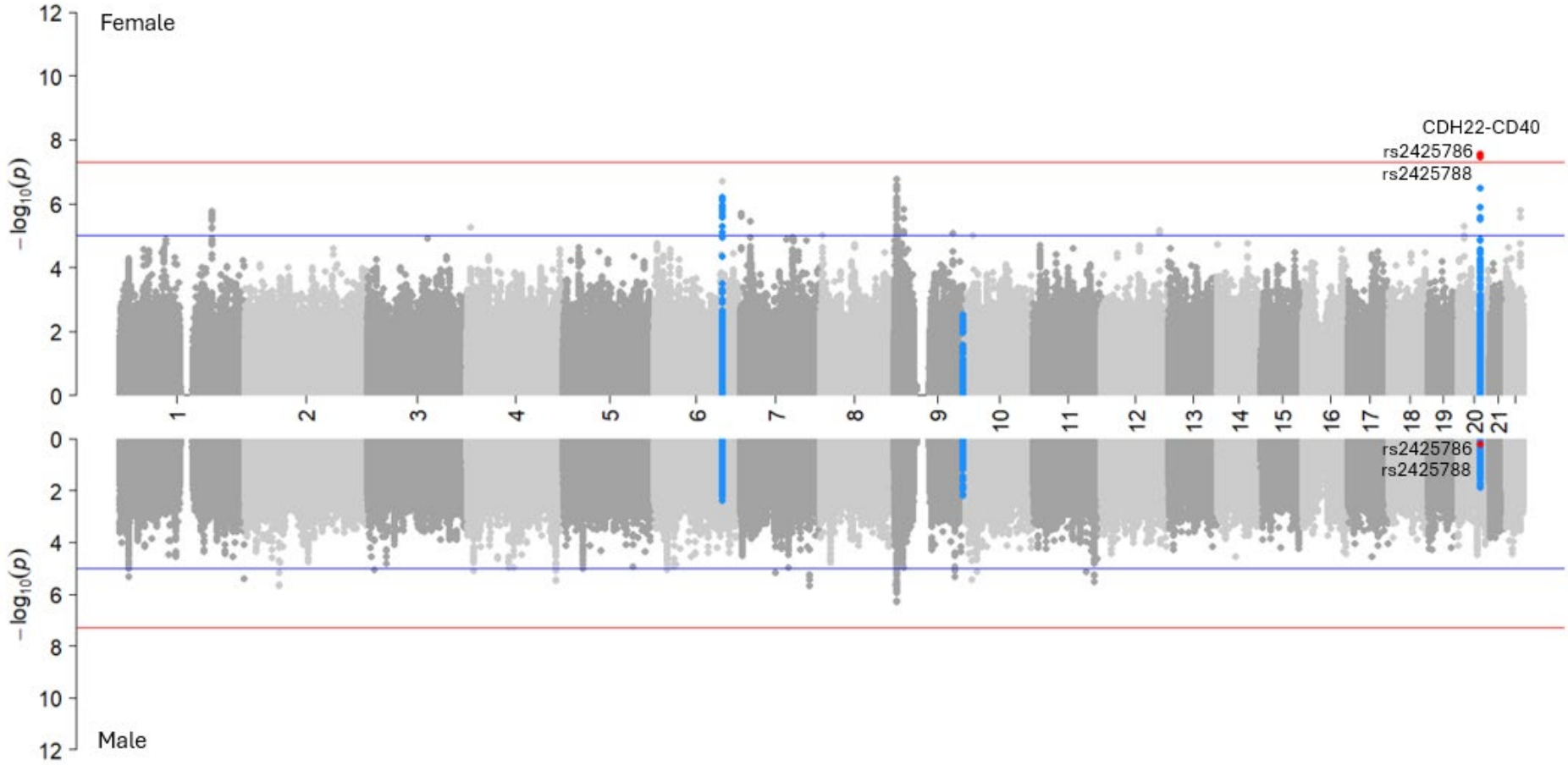
ET and PV patients from the UK1 (blue) and UK Biobank (red) cohorts and healthy controls from WTCCC2 (black) are plotted alongside reference populations from the HapMap consortium with European (CEU in green), Asian (ASI in purple) and African (YRI in yellow) ancestry. The reference populations consist of Utah residents (CEU), Japanese and Han Chinese individuals from Tokyo and China (ASI), Yoruban individuals from Ibadan, Nigeria (YRI). The majority of ET/PV patients cluster with the WTCCC2 controls which suggests that they are suitable for comparison along with the principal components to correct for potential population stratification.

Supplementary Figure 3. Genome-wide association of ET versus PV myeloproliferative neoplasms.



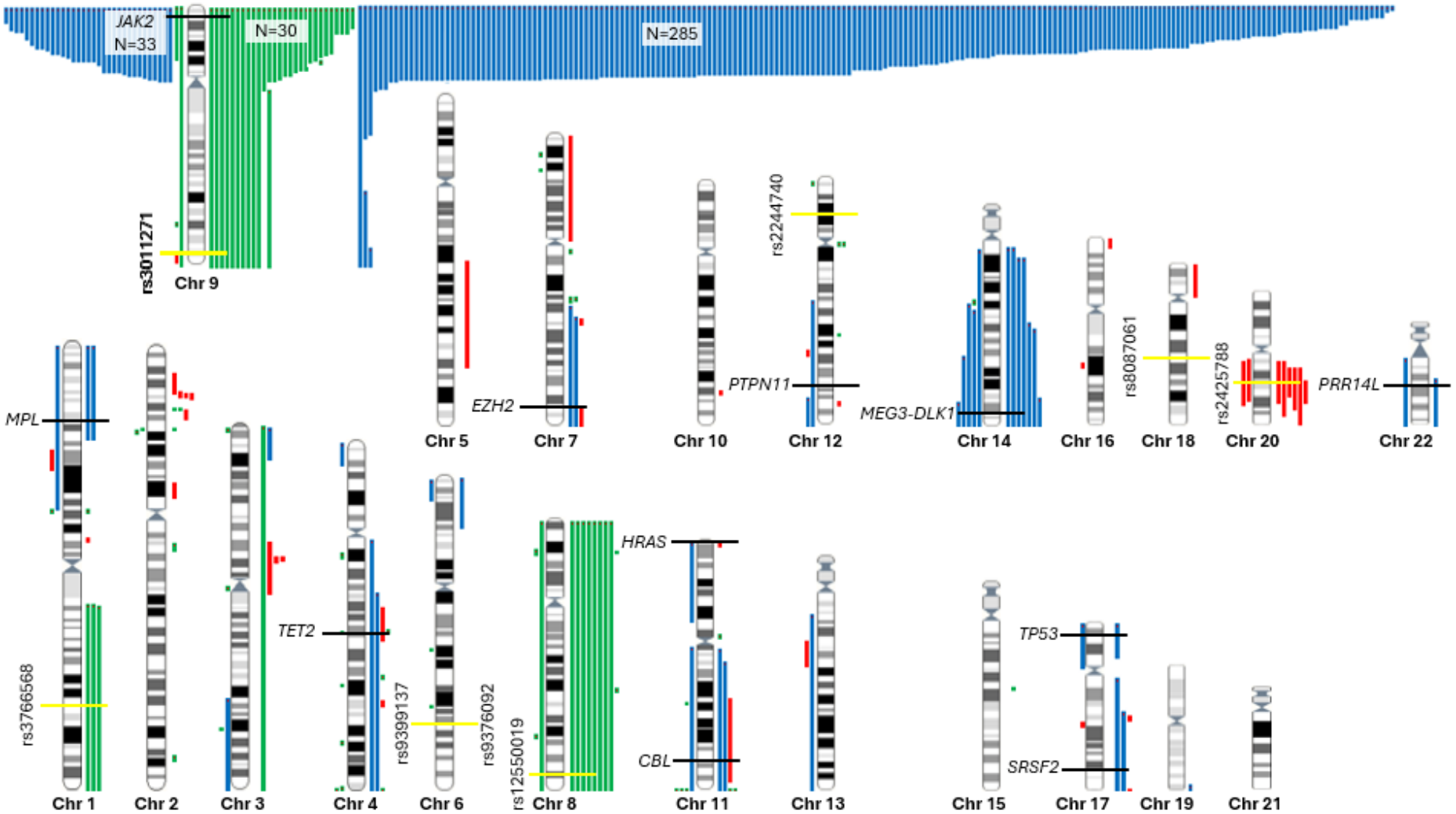
Miami plot of the stage 1 GWAS results for 7,267,872 SNPs tested. Association p-values ($-\log_{10}$) are plotted on the y axis against physical location on the x-axis. SNPs associated with PV (OR<1) are plotted above the x-axis and SNPs associated with ET (OR>1) are plotted below the x-axis. Twenty-nine SNPs were identified with genome-wide significance (p-value $\leq 5 \times 10^{-8}$), indicated by the red lines. However, apart from rs17876031 on chromosome 5 and rs3130039 on chromosome 6, these GWAS significant SNPs are located on chromosome 9p and associated with PV which is attributed to *JAK2* and the recurrent mosaic chromosome abnormalities involving 9p which are more frequent in PV patients (59.6%=318/534) compared with ET (6.5%=35/535). 93 SNPs were selected for replication are highlighted in green with p-values less than 0.001. The ten SNPs that reached suggestive levels of significance after meta-analysis of stages 1 and 2 are shown in red and labelled, including two additional SNPs in *CDH22* on chromosome 20 that were selected based on the gender stratified analysis. SNPs located in the key candidate regions encompassing *GFI1B*, *HBS1L-MYB* and *CDH22-CD40* are highlighted in blue.

Supplementary Figure 4. Genome-wide association of ET versus PV stratified by sex



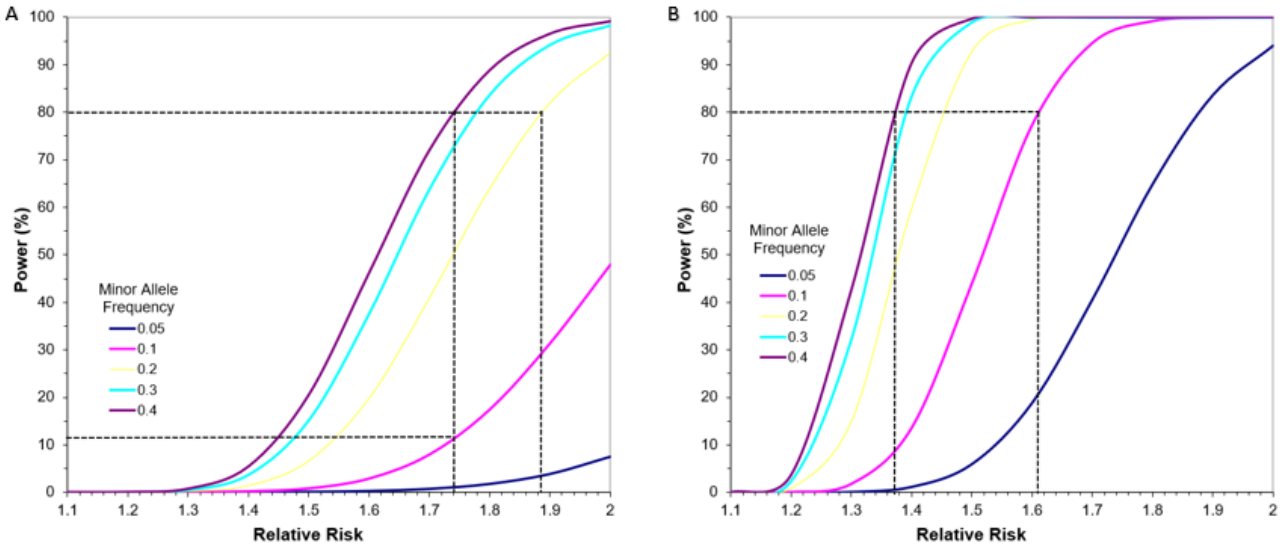
Miami plot depicting association results for 7,267,872 SNPs. The -log10 association p-values are plotted on the y-axis, with physical genomic location on the x-axis. Female results are displayed above the x-axis, while male results appear below the x-axis. Genome-wide significant SNPs ($p \leq 5\times10^{-8}$) are marked by red lines and labelled in red.

Supplementary Figure 5. Copy number changes and regions of acquired uniparental disomy (aUPD) in the 1,069 stage-1 cases



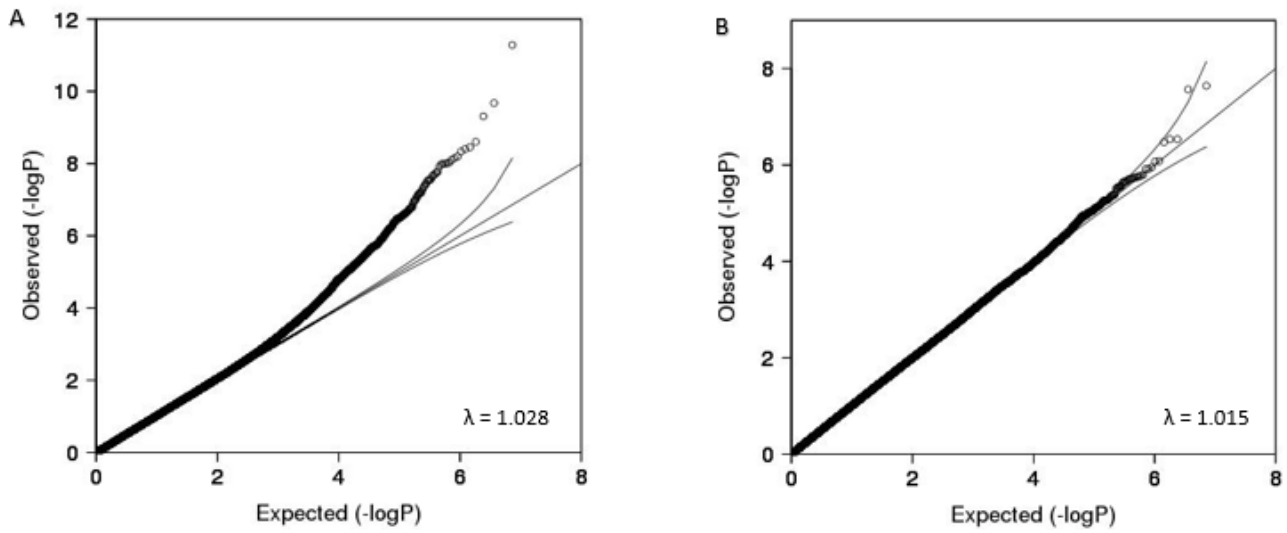
Columns next to the ideogram show regions of copy number gain (green), loss (red) and aUPD (blue) in ET (left hand side) and PV patients (right hand side). Labels indicate the location of known driver genes and SNPs reaching suggestive levels of significance following meta-analysis of stages 1 and 2.

Supplementary Figure 6. Power to detect SNPs associated with MPN subtype



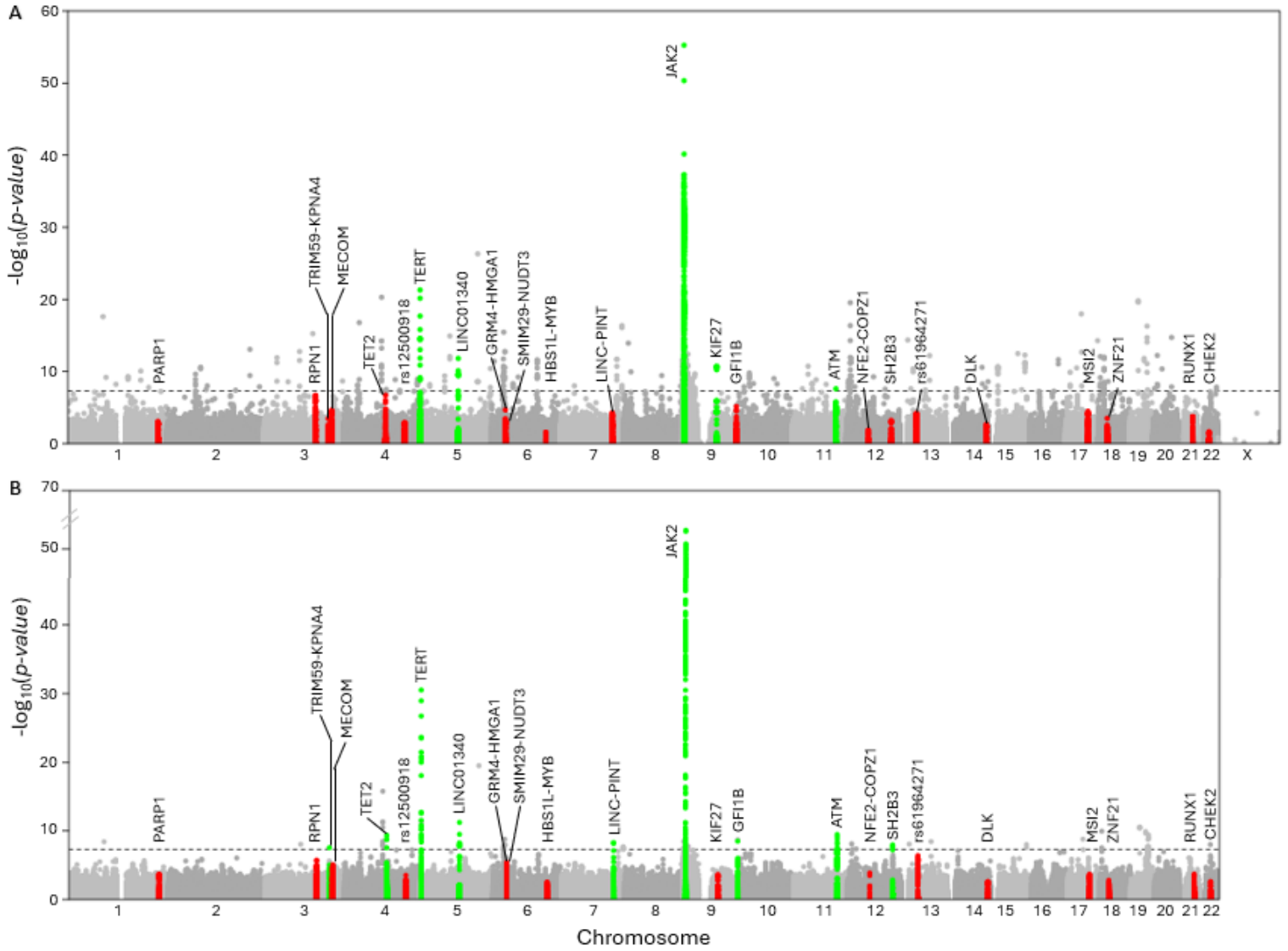
A. Stage 1, 535 ET patients versus 534 PV patients. B. Meta-analysis of stage 1 and 2 1,560 ET patients versus 1,755 PV patients

Supplementary Figure 7. Quantile quantile (QQ) plots of the stage 1 analysis



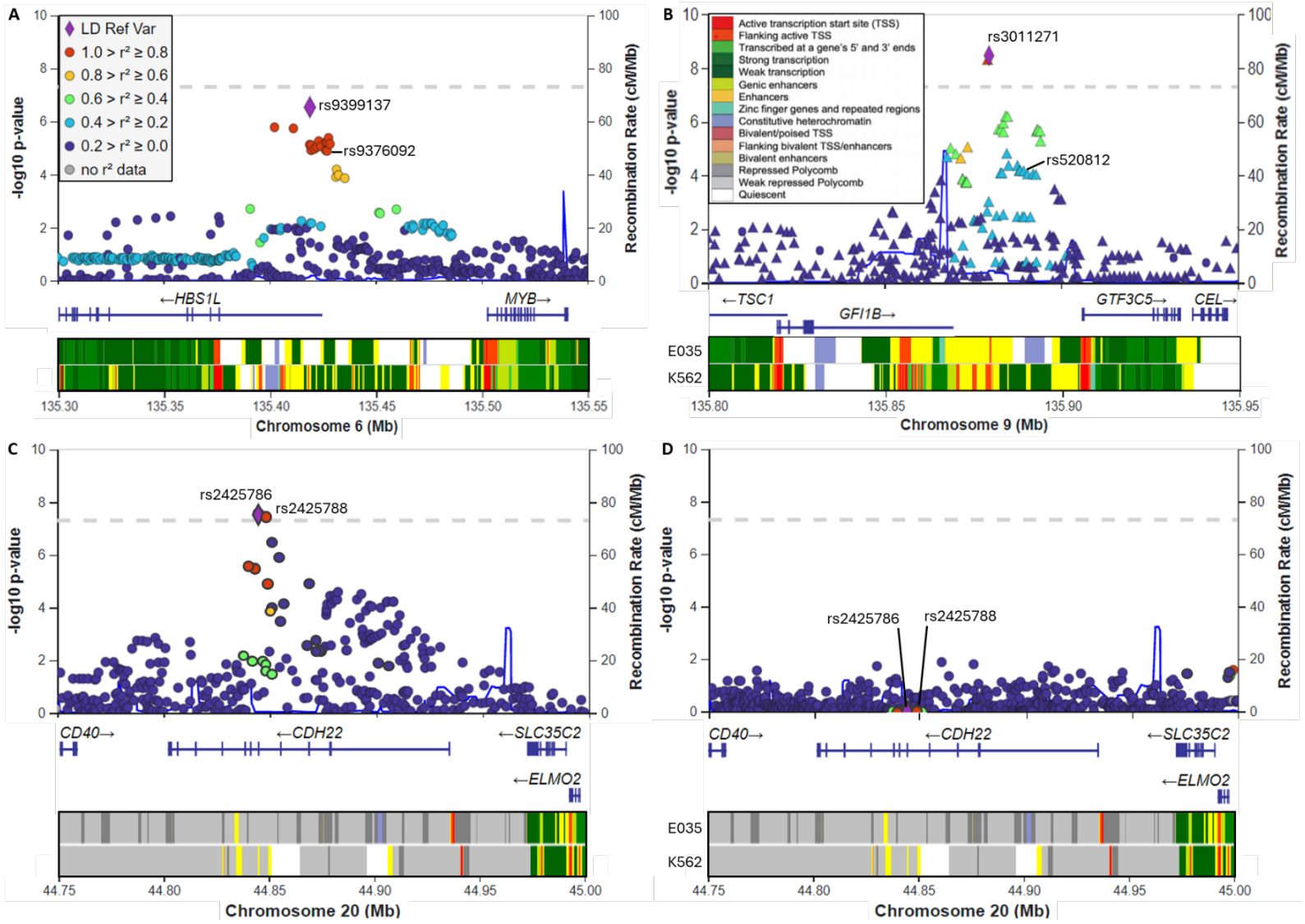
QQ plots of the observed (y axis) versus expected (x axis) -log10 P-values for SNP associations with the risk of developing ET or PV in the stage 1 analysis involving 535 ET cases and 534 PV cases. All SNPs tested are shown in plot A (n=7,267,872) while SNPs located on chromosome 9p are excluded from plot B (n=7,163,150). The region between the curved lines indicates bootstrapped 95% confidence intervals. The diagonal black line indicates the expected distribution of P-values under the null hypothesis.

Supplementary Figure 8. Genome-wide association of ET and PV versus controls



A) Manhattan plot of the stage 1 GWAS results for ET and PV cases versus controls. B) Manhattan plot for a fixed effects meta-analysis of ET and PV cases versus controls which combines evidence from stage 1 and the UK Biobank. Association p-values ($-\log_{10}$) are plotted on the y axis against physical location on the x axis. Published loci are labelled and SNPs within these regions are highlighted in green or red if one or more SNPs reach genome-wide significance (p-value $\leq 5 \times 10^{-8}$), indicated by the black dashed line, or nominal significance (p-value <0.05).

Supplementary Figure 9. Regional plots from the stage 1 analyses for SNPs with suggestive levels of significance.



Results from the stage 1 analyses: **(A)** ET versus PV patients in the *HBS1L-MYB* region containing rs9399137 and rs9376092, **(B)** PV patients versus WTCCC controls in the *GFI1B* region containing rs3011271 and rs520812, **(C)** Female ET versus PV patients in the *CDH22* region containing rs2427586 and rs2427586, (D) The same region containing *CDH22* for male ET versus PV patients. In each plot, the leading SNP is indicated by a purple diamond and the colour of other SNPs represent the strength of linkage disequilibrium (r^2) with the lead SNP. Protein coding genes and RNA genes are shown in the track below with arrows to indicate the direction of transcription and wider lines representing the location of exons. The lower panel displays the 15 state chromatin track (chromHMM) in primary hematopoietic stem cells (E035) and Leukaemia cells (K562) using data from the NIH Roadmap Epigenomics Consortium[73]. Physical positions are relative to build 37 (hg19) of the human genome.

References

1. Harrison CN, Campbell PJ, Buck G, et al. Hydroxyurea compared with anagrelide in high-risk essential thrombocythemia. *N Engl J Med*. 2005;353(1):33-45.
2. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203-209.
3. Yun T, Li H, Chang PC, Lin MF, Carroll A, McLean CY. Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics*. 2021;36(24):5582-5589.
4. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol*. 2013;31(3):213-219.
5. Guo J, Walter K, Quiros PM, et al. Inherited polygenic effects on common hematological traits influence clonal selection on JAK2(V617F) and the development of myeloproliferative neoplasms. *Nat Genet*. 2024;56(2):273-280.
6. Saad M, Lesage S, Saint-Pierre A, et al. Genome-wide association study confirms BST1 and suggests a locus on 12q24 as the risk loci for Parkinson's disease in the European population. *Hum Mol Genet*. 2011;20(3):615-627.
7. Dawoud AAZ, Tapper WJ, Cross NCP. Clonal myelopoiesis in the UK Biobank cohort: ASXL1 mutations are strongly associated with smoking. *Leukemia*. 2020;34(10):2660-2672.
8. Teo YY, Inouye M, Small KS, et al. A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics*. 2007;23(20):2741-2746.
9. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015;4:7.
10. A two-stage meta-analysis identifies several new loci for Parkinson's disease. *PLoS Genet*. 2011;7(6):e1002142.
11. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*. 2016;48(10):1279-1283.
12. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011;27(21):2987-2993.
13. Verma SS, de Andrade M, Tromp G, et al. Imputation and quality control steps for combining multiple genome-wide datasets. *Front Genet*. 2014;5:370.
14. O'Connell J, Sharp K, Shrine N, et al. Haplotype estimation for biobank-scale data sets. *Nat Genet*. 2016;48(7):817-820.
15. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet*. 2010;11(7):499-511.
16. Harris R, Bradburn M, Deeks J, Harbord R, Altman D, Sterne J. metan: fixed- and random-effects meta-analysis. *Stata Journal*. 2008;8(1):3-28.
17. Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? *Stat Med*. 2002;21(11):1559-1573.
18. Staaf J, Lindgren D, Vallon-Christersson J, et al. Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. *Genome Biol*. 2008;9(9):R136.
19. Purcell S, Cherny SS, Sham PC. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics*. 2003;19(1):149-150.
20. Vuckovic D, Bao EL, Akbari P, et al. The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell*. 2020;182(5):1214-1231.e1211.
21. Turner SD. qqman: an R package for visualizing GWAS results using QQ and manhattan plots. *Biorxiv*. 2014:005165.
22. Boughton AP, Welch RP, Flickinger M, et al. LocusZoom.js: Interactive and embeddable visualization of genetic association study results. *Bioinformatics*. 2021;37(18):3017-3018.

23. Ward LD, Kellis M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res*. 2016;44(D1):D877-881.
24. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014;46(3):310-315.
25. Dong S, Zhao N, Spragins E, et al. Annotating and prioritizing human non-coding variants with RegulomeDB v.2. *Nat Genet*. 2023;55(5):724-726.
26. Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. 2019;47(D1):D1005-d1012.
27. Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc*. 2017;12(12):2478-2492.
28. Hoffman MM, Ernst J, Wilder SP, et al. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res*. 2013;41(2):827-841.
29. Võsa U, Claringbould A, Westra HJ, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet*. 2021;53(9):1300-1310.
30. Bao EL, Nandakumar SK, Liao X, et al. Inherited myeloproliferative neoplasm risk affects haematopoietic stem cells. *Nature*. 2020;586(7831):769-775.
31. Pedersen KM, Çolak Y, Ellervik C, Hasselbalch HC, Bojesen SE, Nordestgaard BG. Loss-of-function polymorphism in IL6R reduces risk of JAK2V617F somatic mutation and myeloproliferative neoplasm: A Mendelian randomization study. *EClinicalMedicine*. 2020;21:100280.
32. Loh PR, Genovese G, McCarroll SA. Monogenic and polygenic inheritance become instruments for clonal selection. *Nature*. 2020;584(7819):136-141.
33. Grinfeld J, Nangalia J, Baxter EJ, et al. Classification and Personalized Prognosis in Myeloproliferative Neoplasms. *N Engl J Med*. 2018;379(15):1416-1430.
34. Trifa AP, Bănescu C, Bojan AS, et al. MECOM, HBS1L-MYB, THRB-RARB, JAK2, and TERT polymorphisms defining the genetic predisposition to myeloproliferative neoplasms: A study on 939 patients. *Am J Hematol*. 2018;93(1):100-106.
35. Loh PR, Genovese G, Handsaker RE, et al. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature*. 2018;559(7714):350-355.
36. Hinds DA, Barnholt KE, Mesa RA, et al. Germ line variants predispose to both JAK2 V617F clonal hematopoiesis and myeloproliferative neoplasms. *Blood*. 2016;128(8):1121-1128.
37. Chen Y, Fang F, Hu Y, et al. The Polymorphisms in LNK Gene Correlated to the Clinical Type of Myeloproliferative Neoplasms. *PLoS One*. 2016;11(4):e0154183.
38. Trifa AP, Bănescu C, Tevet M, et al. TERT rs2736100 A>C SNP and JAK2 46/1 haplotype significantly contribute to the occurrence of JAK2 V617F and CALR mutated myeloproliferative neoplasms - a multicentric study on 529 patients. *Br J Haematol*. 2016;174(2):218-226.
39. Loscocco GG, Mannarelli C, Pacilli A, et al. Germline transmission of LNKE208Q variant in a family with myeloproliferative neoplasms. *Am J Hematol*. 2016;91(9):E356.
40. Rumi E, Harutyunyan AS, Pietra D, et al. LNK mutations in familial myeloproliferative neoplasms. *Blood*. 2016;128(1):144-145.
41. Gau JP, Chen CC, Chou YS, et al. No increase of JAK2 46/1 haplotype frequency in essential thrombocythemia with CALR mutations: Functional effect of the haplotype limited to allele with JAK2V617F mutation but not CALR mutation. *Blood Cells Mol Dis*. 2015;55(1):36-39.
42. Tapper W, Jones AV, Kralovics R, et al. Genetic variation at MECOM, TERT, JAK2 and HBS1L-MYB predisposes to myeloproliferative neoplasms. *Nat Commun*. 2015;6:6691.
43. Soler G, Bernal-Vicente A, Antón AI, et al. The JAK2 46/1 haplotype does not predispose to CALR-mutated myeloproliferative neoplasms. *Ann Hematol*. 2015;94(5):789-794.
44. Oddsson A, Kristinsson SY, Helgason H, et al. The germline sequence variant rs2736100_C in TERT associates with myeloproliferative neoplasms. *Leukemia*. 2014;28(6):1371-1374.

45. Lesteven E, Picque M, Conejero Tonetti C, et al. Association of a single-nucleotide polymorphism in the SH2B3 gene with JAK2V617F-positive myeloproliferative neoplasms. *Blood*. 2014;123(5):794-796.
46. Jäger R, Harutyunyan AS, Rumi E, et al. Common germline variation at the TERT locus contributes to familial clustering of myeloproliferative neoplasms. *Am J Hematol*. 2014;89(12):1107-1110.
47. Poletto V, Rosti V, Villani L, et al. A3669G polymorphism of glucocorticoid receptor is a susceptibility allele for primary myelofibrosis and contributes to phenotypic diversity and blast transformation. *Blood*. 2012;120(15):3112-3117.
48. Olcaydu D, Rumi E, Harutyunyan A, et al. The role of the JAK2 GGCC haplotype and the TET2 gene in familial myeloproliferative neoplasms. *Haematologica*. 2011;96(3):367-374.
49. Jones AV, Campbell PJ, Beer PA, et al. The JAK2 46/1 haplotype predisposes to MPL-mutated myeloproliferative neoplasms. *Blood*. 2010;115(22):4517-4523.
50. Olcaydu D, Harutyunyan A, Jäger R, et al. A common JAK2 haplotype confers susceptibility to myeloproliferative neoplasms. *Nat Genet*. 2009;41(4):450-454.
51. Olcaydu D, Skoda RC, Looser R, et al. The 'GGCC' haplotype of JAK2 confers susceptibility to JAK2 exon 12 mutation-positive polycythemia vera. *Leukemia*. 2009;23(10):1924-1926.
52. Jones AV, Chase A, Silver RT, et al. JAK2 haplotype is a major risk factor for the development of myeloproliferative neoplasms. *Nat Genet*. 2009;41(4):446-449.
53. Kilpivaara O, Mukherjee S, Schram AM, et al. A germline JAK2 SNP is associated with predisposition to the development of JAK2(V617F)-positive myeloproliferative neoplasms. *Nat Genet*. 2009;41(4):455-459.
54. Pardanani A, Fridley BL, Lasho TL, Gilliland DG, Tefferi A. Host genetic variation contributes to phenotypic diversity in myeloproliferative disorders. *Blood*. 2008;111(5):2785-2789.
55. Hernández-Boluda JC, Pereira A, Cervantes F, et al. A polymorphism in the XPD gene predisposes to leukemic transformation and new nonmyeloid malignancies in essential thrombocythemia and polycythemia vera. *Blood*. 2012;119(22):5221-5228.
56. Ebbert MT, Ridge PG, Wilson AR, et al. Population-based analysis of Alzheimer's disease risk alleles implicates genetic interactions. *Biol Psychiatry*. 2014;75(9):732-737.
57. Wang L, Wheeler DA, Prchal JT. Acquired uniparental disomy of chromosome 9p in hematologic malignancies. *Exp Hematol*. 2016;44(8):644-652.
58. Vannucchi AM, Pieri L, Guglielmelli P. JAK2 Allele Burden in the Myeloproliferative Neoplasms: Effects on Phenotype, Prognosis and Change with Treatment. *Ther Adv Hematol*. 2011;2(1):21-32.
59. Bastian FB, Cammarata AB, Carsanaro S, et al. Bgee in 2024: focus on curated single-cell RNA-seq datasets, and query tools. *Nucleic Acids Res*. 2025;53(D1):D878-d885.
60. Ferreira-Cerca S, Sagar V, Schäfer T, et al. ATPase-dependent role of the atypical kinase Rio2 on the evolving pre-40S ribosomal subunit. *Nat Struct Mol Biol*. 2012;19(12):1316-1323.
61. Ghosh S, Raundhal M, Myers SA, et al. Identification of RIOK2 as a master regulator of human blood cell development. *Nat Immunol*. 2022;23(1):109-121.
62. Dunham I, Kundaje A, Aldred SF, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57-74.
63. Cao ZQ, Wang Z, Leng P. Aberrant N-cadherin expression in cancer. *Biomed Pharmacother*. 2019;118:109320.
64. Carpinelli MR, Hilton DJ, Metcalf D, et al. Suppressor screen in Mpl-/- mice: c-Myb mutation causes supraphysiological production of platelets in the absence of thrombopoietin signaling. *Proc Natl Acad Sci U S A*. 2004;101(17):6553-6558.
65. Sandberg ML, Sutton SE, Pletcher MT, et al. c-Myb and p300 regulate hematopoietic stem cell proliferation and differentiation. *Dev Cell*. 2005;8(2):153-166.

66. Mukai HY, Motohashi H, Ohneda O, Suzuki N, Nagano M, Yamamoto M. Transgene insertion in proximity to the c-myb gene disrupts erythroid-megakaryocytic lineage bifurcation. *Mol Cell Biol*. 2006;26(21):7953-7965.
67. García P, Clarke M, Vegiopoulos A, et al. Reduced c-Myb activity compromises HSCs and leads to a myeloproliferation with a novel stem cell basis. *Embo j*. 2009;28(10):1492-1504.
68. Clarke ML, Lemma RB, Walton DS, et al. MYB insufficiency disrupts proteostasis in hematopoietic stem cells, leading to age-related neoplasia. *Blood*. 2023;141(15):1858-1870.
69. Vegiopoulos A, García P, Emambokus N, Frampton J. Coordination of erythropoiesis by the transcription factor c-Myb. *Blood*. 2006;107(12):4703-4710.
70. Randrianarison-Huetz V, Laurent B, Bardet V, Blobe GC, Huetz F, Duménil D. Gfi-1B controls human erythroid and megakaryocytic differentiation by regulating TGF-beta signaling at the bipotent erythro-megakaryocytic progenitor stage. *Blood*. 2010;115(14):2784-2795.
71. Elgueta R, Benson MJ, de Vries VC, Wasiuk A, Guo Y, Noelle RJ. Molecular mechanism and function of CD40/CD40L engagement in the immune system. *Immunol Rev*. 2009;229(1):152-172.
72. Caiado F, Pietras EM, Manz MG. Inflammation as a regulator of hematopoietic stem cell function in disease, aging, and clonal selection. *J Exp Med*. 2021;218(7).
73. Kundaje A, Meuleman W, Ernst J, et al. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015;518(7539):317-330.