Genetic variation at MECOM, TERT, JAK2 and HBS1L-MYB predisposes to myeloproliferative neoplasms

William Tapper1,2,*, Amy V. Jones1,2,*, Robert Kralovic3, Ashot S. Harutyunyan3, Katerina Zoi4, William Leung1,2, Anna L. Godfrey5,6, Paola Guglielmelli7, Alison Callaway2, Daniel Ward2, Paula Aranaz2, Helen E. White1,2, Katherine Waghorn1,2, Feng Lin1,2, Andrew Chase1,2, E. Joanna Baxter5,6, Cathy Maclean5,6, JyotiNangalia5,6, Edwin Chen5,6, Paul Evans8, Michael Short8, Andrew Jack8, Louise Wallis9, David Oscier9, Andrew S. Duncombe10, Anna Schuh11, Adam J. Mead12, Michael Griffiths13,14, Joanne Ewing15, Rosemary E. Gale16, Susanne Schnittger17, Torsten Haferlach17, Frank Stegelmann18, Konstanze Döhner18, Harald Grallert19,20, Konstantin Strauch19,21, Toshiko Tanaka22, Stefania Bandinelli23, Andreas Giannopoulos24, Lisa Pieri7, Carmela Mannarelli7, Heinz Gisslinger24, Giovanni Barosi25, Mario Cazzola26,27, Andreas Reiter28, Claire Harrison29, Peter Campbell30, Anthony R. Green5,6, Alessandro Vannucchi7 & Nicholas C.P. Cross1,2

Clonal proliferation in myeloproliferative neoplasms (MPN) is driven by somatic mutations in JAK2, CALR or MPL, but the contribution of inherited factors is poorly characterized. Using a three-stage genome-wide association study of 3,437 MPN cases and 10,083 controls, we identify two SNPs with genome-wide significance in JAK2V617F-negative MPN: rs12339666 (JAK2; meta-analysis P = 1.27 × 10⁻¹⁰) and rs2201862 (MECOM; meta-analysis P = 1.96 × 10⁻⁹). Two additional SNPs, rs2736100 (TERT) and rs9376092 (HBS1L/MYB), achieve genome-wide significance when including JAK2V617F-positive cases. rs9376092 has a stronger effect in JAK2V617F-negative cases with CALR and/or MPL mutations (Breslow–Day P = 4.5 × 10⁻⁷), whereas in JAK2V617F-positive cases rs9376092 associates with essential thrombocythemia (ET) rather than polycythemia vera (allelic χ² = 7.3 × 10⁻⁷). Reduced MYB expression, previously linked to development of an ET-like disease in model systems, associates with rs9376092 in normal myeloid cells. These findings demonstrate that multiple germline variants predispose to MPN and link constitutional differences in MYB expression to disease phenotype.
Myeloproliferative neoplasms (MPN) are a group of related haematological disorders that are characterized by an excess proliferation of one or more myeloid cell lineages and a tendency to transform to acute myeloid leukaemia\(^1\). MPNs are classified by which myeloid cell lineage(s) is predominantly expanded in the peripheral blood, namely erythrocytes in polycythaemia vera (PV) and platelets in essential thrombocythaemia (ET). Patients with primary myelofibrosis (PMF) display bone marrow fibrosis, megakaryocytic abnormalities, variable peripheral blood counts, extramedullary haematopoeisis and hepatosplenomegaly. MPNs arise as a consequence of clonal proliferation driven by at least one somatically acquired driver mutation. Of these, the most recurrent is JAK2\(^{V617F}\), which occurs in >95% of PV and 50–60% of ET and PMF cases. In patients without JAK2 mutations, somatic mutations in CALR are the most common, occurring in well over 50% of JAK2\(^{V617F}\)-negative ET and PMF. Other somatic mutations in MPN have been identified but they are less common and occur in fewer than 20% of cases. Most prominent of these are from a phenotypic perspective are mutations that activate the thrombopoietin receptor (MPL), which are seen in 3–10% of ET and PMF. In the great majority of cases, mutations in JAK2, CALR, and MPL are mutually exclusive, although occasional cases may test positive for mutations in more than one of these genes. Patients with ET and PMF that test negative for JAK2, CALR, and MPL mutations are typically referred to as ‘triple negative’\(^2,3\).

Although it is accepted that acquisition of these mutations drives clonal proliferation in MPNs, less is known about what factors influence the development, phenotype and severity of disease. Evidence from epidemiological\(^4\) and familial studies\(^5,6\) strongly suggest that common, low penetrance factors present in the general population contribute to the risk of developing MPN and potentially such factors may also contribute to the phenotypic pleiotrophy observed in these disorders. We previously characterized a major predisposition locus at the JAK2 gene, which is associated with the acquisition of V617F mutations\(^7,8\). This specific JAK2 haplotype, called ‘46/1’ or ‘GGCC’ strongly predisposes to JAK2\(^{V617F}\)-positive disease (OR = 3.7; Fisher’s exact test \(P < 10^{-20}\)). The 46/1 haplotype is also weakly associated with JAK2 exon 12 positive PV\(^9\), MPN with MPL W515 mutations and those that lack JAK2 or MPL mutations\(^10\). More recently, a common germline variant was identified in TERT, which predisposes equally to PV (OR = 2.3, allelic \(\chi^2 P = 1.1 \times 10^{-5}\)), ET (OR = 2.3, allelic \(\chi^2 P = 2.87 \times 10^{-3}\)) and PMF (OR = 2.4, allelic \(\chi^2 P = 2.46 \times 10^{-2}\)) in the Icelandic population\(^11\).

Overall, the 46/1 haplotype has been estimated to account for \(\sim 50\%\) of the population attributable risk (PAR) of developing an MPN in European populations, but it only explains a minor component of JAK2\(^{V617F}\)-negative disease\(^7\). The variant in TERT is also estimated to have a PAR of 58% in Icelandics and contributes to JAK2\(^{V617F}\) positive and negative disease. Since the bulk of the population-level predisposition to JAK2\(^{V617F}\) positive MPN is likely to be accounted for by variation at JAK2 and TERT, the initial aim of this study is to identify additional germline variants that predispose to JAK2\(^{V617F}\)-negative ET or PMF by performing a genome-wide association study (GWAS).

We identify two single-nucleotide polymorphisms (SNPs), rs12339666 within JAK2 and rs2201862, 153 kb downstream of MEGOC, which associate with JAK2\(^{V617F}\)-negative disease, and two additional SNPs, rs2736100 in TERT and rs9376092 between HBSIL and MYB, when including JAK2\(^{V617F}\)-positive cases. The SNP between HBSIL and MYB, rs9376092, has a stronger effect in JAK2\(^{V617F}\)-negative cases with somatic CALR and/or MPL mutations and predisposes to essential thrombocythaemia in JAK2\(^{V617F}\)-positive cases. Finally, we demonstrate that the candidate risk allele at rs9376092 is associated with reduced MYB expression in normal myeloid cells.

## Results

**Genome-wide association study.** Following quality control there were 524 JAK2\(^{V617F}\)-negative MPN cases, 2,674 controls and 2,098,039 SNPs (642,633 observed and 1,455,406 imputed) for analysis at stage-1 (Supplementary Fig. 1). This cohort is estimated to have 80% power for detecting common SNPs (minor allele frequency (MAF) = 0.3) with relative risks of 1.25 and rare SNPs (MAF = 0.05) with relative risks (RR) of 1.5 (ref. 12). The QQ plot demonstrated good agreement between observed and expected \(P\) values (genomic inflation factor, \(\lambda = 1\)) until the tail of the distribution where SNPs with \(P\) values \(< 10^{-4}\) deviated from the null distribution (Supplementary Fig. 2). Systematic biases between cases and controls such as population stratification or somatic mutations in the MPN cases are therefore unlikely to contribute to the significance of SNPs. Figure 1 shows a Manhattan plot of these results. Following our selection criteria (see Methods), 203 SNPs from 173 genomic regions with allelic \(\chi^2 P\) values \(\leq 0.02\) (Supplementary Table 1) were chosen for assessment in the three-stage two cohorts.

**Replication and validation of candidate SNPs.** For replication at stage-2, 196 of the 203 selected SNPs were successfully genotyped in 658 JAK2\(^{V617F}\)-negative MPN cases and 1,196 controls (Supplementary Table 1). A further three SNPs were removed due to significant deviation from Hardy–Weinberg equilibrium (HWE; exact test \(P\)-value \(\leq 0.001\) in controls) leaving 193 SNPs for analysis. Additional cases from Germany (\(n = 187\)) and Austria (\(n = 99\)) and controls from the BBC (\(n = 2706\)) and from Bavaria (\(n = 1805\)), all of which had been previously genotyped, were also used. These cases and controls were analysed as three separate cohorts consisting of samples from the United Kingdom (stage-2.1), Germany and Austria (stage-2.2), and Greece (stage-2.3). After testing for allelic association and using a fixed effects meta-analysis to combine evidence from stages 1 and 2, five SNPs were identified with combined \(P\) values \(\leq 0.001\) and significant replication in at least one of the three replication cohorts with effects in the same direction as the stage-1 analysis (Supplementary Table 1). The uncombined replication \(P\)-values for allelic \(\chi^2\) tests ranged from 0.025 to 0.00025. One of these SNPs (rs12339666) is located within the JAK2 gene and was not considered further because of its consistency with previous findings, showing that variants tagging the JAK2 46/1 haplotype are associated with JAK2\(^{V617F}\)-negative MPN\(^10\).

A total of four SNPs with replication at stage-2 were therefore selected for further analysis (stage-3) in two independent cohorts consisting of 328 cases and 2,800 controls from the United Kingdom (stage-3.1) and 637 cases and 160,8 controls from Italy (stage-3.2, Supplementary Table 1). Of the four SNPs tested, all of which were successfully genotyped and passed quality criteria, three (rs9376092, rs4858647 and rs2201862) had allelic \(\chi^2\) \(P\) values \(< 0.05\) and effects in the same direction as the stage-1 analysis and remained significant when considering the false discovery rate\(^13\) to correct for the number of SNPs tested (Supplementary Table 1). To further validate our findings, the four SNPs were genotyped in 406 reactive cases from the United Kingdom, which mimic MPNs and were thus considered as negative controls. Comparison to 2,800 controls from the United Kingdom (stage 3.3) indicated that none of the SNPs were significantly associated with the risk of developing a reactive phenotype (Supplementary Table 1), thus supporting the notion that they are specifically associated with clonal MPN.
Following meta-analysis of the six cohorts using fixed-effect sizes, one novel SNP was identified with a genome-wide level of significance (rs2201862; meta-analysis \(P = 1.96 \times 10^{-9}\)), an effect in the same direction as the discovery data, and without significant heterogeneity between cohorts (Table 1, Figs 2 and 3). Furthermore, rs2201862 had evidence for replication (\(P \leq 0.05\) for allelic \(\chi^2\)) in three of the five follow-up cohorts and two unlinked SNPs (\(r^2 < 0.5\)) that are within 100 kb are also significant (\(P \leq 0.05\) for allelic \(\chi^2\)-test) in one of the three replication cohorts that were tested for these SNPs. Rs2201862 is located at chromosome 3q26 – 153 kb downstream of MECOM and 99 kb downstream of a pseudogene EGFEMIP. In addition, two other SNPs had moderate evidence of association (rs9376092 meta-analysis \(P = 5.27 \times 10^{-6}\) and rs4858647 meta-analysis \(P = 1.39 \times 10^{-5}\)) and effects in the same direction as the discovery data (Table 1, Fig. 2). However, for both these SNPs Cochran’s Q statistic and the \(I^2\) statistic estimated significant heterogeneity between cohorts. Rs9376092 is located at 6q23, 51 kb upstream of HBBSIL and 76 kb upstream of MYB. Rs4858647 is an intergenic SNP at 3p24, 255 kb upstream of THRBP and 425 kb upstream of RARBP.

Association with variation at TERT. Following the discovery of an association between rs2736100 in TERT at 5p15 and MPN, re-examined our data for evidence of this relationship. In our original analysis, rs2736100 was not tested at stage-1 due to its significant deviation from HWE in the pooled cohort of controls from WTCCC2 (NBS + BBC \(P = 0.00096\) for HWE exact test). However, when controls from the NBS and BBC cohorts were assessed separately the deviation from HWE did not exceed the QC threshold of \(P \leq 0.001\) (NBS \(P = 0.01\), BBC \(P = 0.03\) for HWE exact test). The stage 1 data were therefore reanalysed using the unpooled assessment of HWE, which identified rs2736100 as the only SNP with a \(P\) value \(\leq 0.01\) (rs2736100, OR \(= 1.27\) for allelic \(\chi^2\)) in a 242-kb region, which flanks the TERT locus by 100 kb. This association was replicated in the Austrian cohort (\(P = 0.008\), OR \(= 1.49\) for allelic \(\chi^2\)), and a meta-analysis of the two cohorts was used to estimate the final significance and effect size (Table 1: meta-analysis \(P = 1.73 \times 10^{-5}\), OR \(= 1.31\)). In both cohorts, rs2736100 was not correlated with any other SNPs (\(r^2 > 0.15\)), which may account for the lack of supporting SNPs within the TERT locus. Other JAK2V617F-negative cohorts were not tested for rs2736100.

**Table 1 | Summary of the top SNPs showing or approaching genome-wide significant association \((P < 5 \times 10^{-8})\) in meta-analysis.**

| SNP | Alleles | RAIF | Gene | JAK2V617F-negative cases | JAK2V617F-negative and all JAK2V617F-positive | JAK2V617F-negative and JAK2V617F-positive ET and MF cases |
|-----|---------|------|------|--------------------------|-----------------------------------------------|---------------------------------------------------|
| rs12339666 | T/G | 0.26 | JAK2 | 1.272 \(\times 10^{-5}\) | 1.218 (1.13-1.29) | 1.19 (1.09-1.31) |
| rs2201862 | T/C | 0.48 | MECOM | 1.964 \(\times 10^{-6}\) | 1.38 (1.10-1.63) | 1.38 (1.10-1.63) |
| rs9376092 | A/C | 0.27 | HBBSIL-MYB | 5.275 \(\times 10^{-5}\) | 0.557 (1.7-10) | 0.557 (1.7-10) |
| rs4858647 | A/C | 0.11 | THRBP-RARB | 1.393 \(\times 10^{-5}\) | 1.24 (1.13-1.37) | 1.24 (1.13-1.37) |
| rs2736100 | C/A | 0.51 | TERT | 1.728 \(\times 10^{-5}\) | 1.31 (1.16-1.47) | 1.31 (1.16-1.47) |

*rs identifier from dbSNP.
†Risk-associated/non-risk-associated alleles.
‡Frequency of risk allele in controls that were used to calculate population attributable risk (PAR). \(P_{\text{META}}\), fixed effects meta-analysis of allelic \(\chi^2\) tests; OR, odds ratio; CI, 95% confidence interval; \(I^2\), heterogeneity index (0–100).

**Figure 1 | Genome-wide association of JAK2V617F-negative myeloproliferative neoplasms.** This Manhattan plot shows the results of the stage 1 GWAS for all autosomes. Results are plotted as \(-\log10\) of the allelic \(\chi^2\) \(P\) value. At this stage, no SNPs met the stringent genome-wide significance level of \(P\) value \(\leq 5 \times 10^{-8}\). The three SNPs that attained genome-wide significance and two additional SNPs with moderate association following meta-analysis are highlighted in green, on chromosome 3 (rs2201862 and rs4858647), chromosome 5 (rs2736100), chromosome 6 (rs9376092) and chromosome 9 (rs12339666). This plot was produced using the qqman R package.
Assessments with CALR and MPL mutations. To investigate whether the SNPs we identified had similar effects in cases with and without somatic mutations of CALR and/or MPL, the 6 JAK2V617F-negative cohorts were stratified according to the presence or absence of these mutations and compared using a Breslow–Day test (Table 2B). This analysis determined that the effects of rs12339666 and rs9376092 were significantly lower in cases without CALR and/or MPL mutations, while rs2201862, rs4858647 and rs2736100 had effects that were similar in cases with and without these mutations. Of these, rs9376092 (HBS1L-MYB) was particularly prominent (Breslow–Day P = 4.5 × 10^-7; Table 2B).

To investigate the possibility that germline variation at CALR might contribute to the acquisition of somatic CALR mutations, a 206-kb region that flanks the CALR locus by 100 kb and contains 11 SNPs was investigated in the stage 1 data as a whole and in the subset of cases with CALR mutations. No significant SNPs were identified in this region in either data set (Supplementary Fig. 3). However, the closest SNPs to CALR in this region are rs11667458, which is 29,869 bp downstream, and rs10403210, which is 48,335 bp upstream, and neither of these are in linkage disequilibrium (LD) with any SNPs in or close to CALR. Additional studies are therefore required to determine definitively whether germline variation at CALR influences the acquisition of somatic CALR mutations.

Assessments with JAK2V617F-positive MPN. To determine whether any of the SNPs were associated exclusively with JAK2V617F-negative MPN, two cohorts of JAK2V617F-positive cases (PV, n = 505; ET, n = 499) and 2,706 controls from the United Kingdom were analysed (stage-3.4 and stage 3.5). After testing for allelic association and using the false discovery rate to correct for the five SNPs tested, all SNPs except rs4858647 retained significance in at least one of the two cohorts and had effects in the same direction as the previous analyses (Supplementary Table 1). On comparison of JAK2V617F positive and negative MPN, the Breslow–Day test identified two SNPs that had stronger effects in JAK2V617F-positive cases: rs12339666 (Breslow–Day P = 2.8 × 10^-18), which tags the JAK2 46/1 haplotype, and rs2736100 (TERT; Breslow–Day P = 0.0026). Rs4858647 (THRB-RARB) had a stronger effect in JAK2V617F-negative cases (Breslow–Day P = 0.015), and no difference was seen for the remaining SNPs rs2201862 (MECOM) and rs9376092 (HBS1L-MYB) (Table 2B).

Comparison of effect sizes between JAK2V617F-positive ET versus controls and JAK2V617F-positive PV versus controls revealed significantly larger effect for rs12339666 in PV (JAK2; Breslow–Day P = 2.452 × 10^-11) and rs9376092 in ET (HBS1L-MYB; Breslow–Day P = 0.0004; Table 2B). Meta-analysis of all JAK2V617F-negative and JAK2V617F-positive cases (Table 1) identified three SNPs that achieved genome-wide significance: rs12339666 (JAK2), rs2736100 (TERT) and rs2201862 (MECOM). Rs9376092 (HBS1L-MYB) also achieved genome-wide significance in meta-analysis of all JAK2V617F-negative and JAK2V617F-positive ET and MF cases (Table 1).

Population attributable risks. According to the genotypic odds ratio and population frequency, rs2736100 is estimated to account for the largest proportion of the population attributable risk in JAK2V617F-negative cases (PAR = 34.9%) followed by rs12339666, which tags the 46/1 haplotype (PAR = 17.4% compared with 11.8% for the tightly linked SNP rs12340895 in JAK2V617F and MPL mutation negative cases from our previous study10). The SNP in close proximity to MECOM, rs2201862, accounts for 16.5% of the PAR (Table 1). On the basis of a multiplicative model without interaction, these three factors are estimated to have a combined PAR of 55%.

Variation at rs9376092 correlates with MYB expression. The candidate MPN risk SNP rs9376092 maps to the HBS1L and MYB

---

**Figure 2** | A plot of the stage 1 GWAs for the most significant novel SNP (rs2201862) identified by meta-analysis. This plot shows the stage 1 results (n = 524 JAK2V617F-negative cases versus n = 2674 controls) in a 400-kb region, which flanks the most significant SNP identified by meta-analysis (rs2201862). SNPs are plotted as the −log10 of the allelic \( \chi^2 \) P value. The amount of LD between rs2201862 and all other SNPs is indicated by the colour of the data points. The two SNPs with red borders (rs16853092 and rs4955627) were also genotyped and replicate at stage 2 but were not taken forward to stage 3 because rs2201862 was more significant at stage 2. The plot was produced using LocusZoom.
null
which would be expected to include a variable proportion of clonal and non-clonal cells. We considered the possibility that clonal somatic changes might affect the analysis; however, in line with previous array and exome-sequencing studies, we found that MPN genomes are relatively simple compared to other cancers with only 8% (42/524) of cases showing likely somatic copy-number changes or acquired uniparental disomy (Supplementary Fig. 5). These regions of gross alteration do not correspond to the location of the risk factors we identified and exclusion of these cases has minimal effect on the overall results (Supplementary Table 6). Although we cannot completely exclude the possibility that recurrent clonal changes below the resolution of array analysis might have influenced our findings, SNPs in such regions would generally be excluded by our rigorous QC criteria due to a high percentage of missing genotypes, poor genotyping quality or deviation from HWE. Notably, none of the associated SNPs deviated from HWE at an Exact test $P$ value $\leq 0.05$, which suggests that clonal changes have not contributed to their significance (Methods and Supplementary Table 7).

Meta-analysis of six $JAK2^{V617F}$-negative cohorts from our three-stage GWAS identified two SNPs with genome-wide significance and three that approached this level of significance. The most significant SNPs are rs12339666, which tags the $JAK2$ 46/1 haplotype, and rs2201862, which is in a non-coding region proximal to $MECOM$. Of the three SNPs with moderate association, two are within or close to genes of potential functional relevance to MPN: the intergenic region between $HBSIL$ and $MYB$ and an intronic SNP in $TERT$, which has recently been associated with MPN in an independent study, as well as several other malignancies. The remaining SNP (rs4858647) was mapped to a gene desert on chromosome 3p24. Although the $TERT$ SNP rs2736100 did not achieve genome-wide significance (Table 1), it should be noted that we only focused on this polymorphism retrospectively and thus the number of cases analysed was significantly fewer than for the other variants. When the $JAK2^{V617F}$-positive cohort was included in the analysis, rs2736100 also achieved genome-wide significance (Table 1).

To our knowledge, only one other polymorphism has been reported in the literature to be associated with the development of MPN: rs6198 in the 3’ UTR of $NR3C1$, which encodes the glucocorticoid receptor. On the basis of a very small number of cases, rs6198 was suggested to be associated with PV and MF, but not ET21. An association with MF, however, was confirmed in a subsequent larger study22. This SNP is not included on the SNP 6 platform and thus was not formally evaluated in our study. However, we did not see any suggestion of an association with other SNPs at $NR3C1$ in our stage-1 cohort, which consisted of cases with ET.

The $JAK2$ and $TERT$ SNPs had effects that were stronger in $JAK2^{V617F}$-positive cases, whereas rs4858647 ($THRB-RARB$) was stronger in $JAK2^{V617F}$-negative cases. Within $JAK2^{V617F}$-negative cases, the effect of rs9376092 ($HBSIL-MYB$) was markedly stronger in cases with $CALR$ and/or $MPL$ mutations compared with those that were triple negative. Overall, the three principal loci ($JAK2/TERT/MECOM$) are estimated to have a combined PAR of 51% for $JAK2^{V617F}$-negative MPN. Fine mapping may uncover some of the missing heritability, either by identifying causal variants that are more strongly associated or by identifying additional signals at these loci. Nevertheless, it is likely that there are additional loci with equivalent or lesser effects that predispose to MPN, which could be detected by larger follow-up studies. Our preliminary analysis of the impact of the new associations on clinical and biological parameters recorded for the PT-1 trial have suggested only weak effects at best, but it remains to be determined whether integration of these genetic variants into risk models, together with that of somatically acquired mutations, may enhance risk profiling and be clinically useful. Nevertheless, identification of genetic predisposition loci provides pointers towards novel biological mechanisms that underlie MPN development.

Some of the associations we identified are close to loci with established credentials of relevance to MPN. The strongest novel association was rs2201862, 153Kb downstream of $MECOM$ ($MDS1$ and $EVI1$ complex locus), which has been implicated in megakaryopoiesis by several lines of evidence. First, a number of alternatively spliced mRNAs are transcribed from this locus and one of these ($EVI1$) is overexpressed in myeloid malignancies with chromosome rearrangements, usually an inv(3)(q21q26) or t(3;3)(q21;q26) that reposition a enhancer at 3q21 and elevated platelet counts24. Second, overexpression of $EVI1$ blocks terminal differentiation of bone marrow progenitor cells to granulocytes25 but promotes megakaryocyte differentiation in UT-7/GM cells26. Third, the genetic basis of most cases of TAR (thrombocytopenia with absent radii) syndrome is a null allele of

---

### Table 2 | Frequency of somatic mutations and subtype specific effects.

| A. Frequency of somatic mutations in $CALR$ and $MPL$ |  |
|---|---|
| **Cohort** | **CALRNo. (%)** | **MPLNo. (%)** | **CALR and/or MPLNo. (No. double $\vee$ ve)** | **Triple $\vee$ No. (%)** |
| UK1 | 266 (51.2) | 51 (9.8) | 316 (1) | 204 (39.2) |
| UK2 | 84 (43.3) | 11 (5.7) | 94 (1) | 69 (24) |
| German & Austrian | 136 (47.2) | 83 (28.8) | 217 (2) | 338 (75.2) |
| Greek | 98 (21.5) | 14 (3.2) | 109 (1) | 127 (39.2) |
| UK3 | 168 (51.9) | 29 (9) | 197 (1) | 234 (38.2) |
| Italian | 322 (56.2) | 56 (9.2) | 377 (1) |  |

| B. Association and comparison according to $JAK2^{V617F}$ and either $CALR$ and/or $MPL$ |  |
|---|---|---|---|---|
| **SNP** | **JAK2$^{V617F}$ vs $MPL$ in each population** | **JAK2$^{V617F}$ vs $CALR$ and/or $MPL$** | **JAK2$^{V617F}$ vs $THRB-RARB$** | **JAK2$^{V617F}$ vs $HBSIL-MYB$** |
| rs2339666 | 1.272 $\times 10^{-10}$ | (1.34) | 5.43 $\times 10^{-12}$ | (2.43) | 2.82 $\times 10^{-18}$ | (1.40) | 0.0002 (12.8) | 0.0266 | 5.344 $\times 10^{-10}$ | (1.68) | 1.286 $\times 10^{-10}$ | (3.50) | 2.452 $\times 10^{-10}$ |
| rs2736000 | 1.728 $\times 10^{-10}$ | (13.1) | 1.92 $\times 10^{-10}$ | (1.66) | 0.0026 | 0.0003 (1.22) | 0.0179 (12.6) | 0.6555 | 8.179 $\times 10^{-10}$ | (1.62) | 3.135 $\times 10^{-10}$ | (1.71) | 0.3864 |
| rs2201862 | 1.964 $\times 10^{-10}$ | (0.82) | 0.0368 (0.89) | 0.3218 | 2.628 $\times 10^{-10}$ | (0.82) | 5.08 $\times 10^{-10}$ | (0.83) | 0.2327 | 0.176 (0.91) | 0.0733 (0.88) | 0.9684 |
| rs9376092 | 5.273 $\times 10^{-10}$ | (1.18) | 0.0566 (0.12) | 0.4967 | 3.869 $\times 10^{-10}$ | (1.24) | 0.0493 (1.11) | 4.487 $\times 10^{-7}$ | 2.62 $\times 10^{-5}$ | (1.41) | 0.0683 (0.87) | 0.0004 |
| rs4858647 | 1.393 $\times 10^{-10}$ | (1.24) | 0.6891 (1.04) | 0.0146 | 0.0201 (1.23) | 0.0108 | 0.0072 (1.21) | 0.0187 | 0.767 (1.04) | 0.8049 (1.03) | 0.4167 |

(A) Number and percentage of cases with somatic mutations in $CALR$ and $MPL$ in each population. Cases with somatic mutations in $CALR$ and $MPL$ are described as double $\vee$ ve. Triple $\vee$ refers to cases without somatic mutation in $JAK2$, $CALR$ and $MPL$. In cohorts with double positives, the percentage of cases will sum to just over 100. (B) $P_{META}$ and odds ratio (OR) from fixed effects meta-analyses in subsets of cases defined by the somatic status of $JAK2$, $CALR$ and $MPL$. $PV_{BD}$, Breslow-Day test $P$ value.
RBM8A coupled with a rare polymorphism that creates an EV11 binding site at the RBM8A promoter, resulting in greatly reduced levels of RBM8A expression27. Owing to a highly restricted expression pattern in haematopoietic progenitors, we were unable to determine whether rs2201862 influences MECOM expression; however, mining of publicly available databases indicated that this SNP, which is in a region of very low LD, co-localizes with a DNA hyper-hypersensitive site in the embryonic stem cell line hESC, suggesting it may correspond to a regulatory element with enhancer-like activity.

Although rs9376092 did not reach genome-wide significance in JAK2V617F-negative cases alone, other data points to its importance as a risk allele in MPN. This SNP maps to a well-characterized region on chromosome 6q23 located 51 Kb centromeric of HBSSL and 75 Kb telomeric of MYB. MYB encodes a transcription factor present at high levels in immature progenitors that plays an important role in progression through key stages of haemopoiesis26. HBSSL encodes a GTP-binding elongation factor that is involved in the regulation of a variety of cellular processes, such as protein synthesis, cytokeslet trafficking and signal transduction29. In healthy individuals, diverse SNPs in the HMIP region between HBSSL and MYB were initially found to be associated with hereditary persistence of fetal haemoglobin30. Subsequent studies identified that HMIP SNPs are associated with normal variation in platelet, monocyte and erythrocyte counts as well as erythroid phenotype15,31–34. The SNP identified by our GWAS, rs9376092, is also located in the HMIP region and is in strong linkage disequilibrium with other variants that tag this block. Recently, HMIP variation has also been linked to the risk of developing Hodgkin’s lymphoma (Stage-1 discovery data set). Several studies have focused on the definition of HMIP as an erythroid distal regulatory region. HMIP shows DNase I hypersensitivity, GATA binding, RNA polymerase II interaction and marked histone acetylation36–38. In erythroid progenitors cells grown in liquid culture, HMIP variant correlated with both HBSSL and MYB expression, but the effect was more pronounced for HBSSL and only HBSSL expression correlated with elevated levels of fetal haemoglobin39. However, the multiple genetic associations described above indicate that HMIP influences other myeloid lineages, suggesting that it may harbour multiple tissue-specific enhancer elements. Indeed, we found that allelic variation at rs9376092 was associated with differences in expression of both flanking genes in healthy individuals but most markedly with MYB, with the risk allele linked to reduced MYB expression. We also found that MYB expression is lower in JAK2V617F mutant BFU-E from ET patients compared with wild-type BFU-E. These observations are in agreement with previous findings that genetic knockdown of Myb in mice induces a transplantable MPN that closely resembles human ET46 and that MYB knockdown in CD34+ human bone marrow cells promotes megakaryocytic development40. Thus, our findings link a candidate risk SNP to lower levels of MYB expression, which has been shown to promote an ET-like disease in mice and megakaryocytic development in humans. Strikingly, in JAK2V617F-positive cases we find that variation at rs9376092 is strongly linked to an ET phenotype. Direct comparison of SNP frequencies between JAK2V617F-positive ET and JAK2V617F-positive PV indicates that the rs9376092 C allele is more common in ET (allelic \(\chi^2 = 7.5 \times 10^{-7}\); Table 3), whereas comparison with controls indicates that this SNP specifically increases the risk of ET (allelic \(\chi^2 = 2.6 \times 10^{-6}\); Table 2B), and there is a trend towards a reduction in the risk of developing PV (allelic \(\chi^2 = 0.068\) since the effects are in the opposite direction for these two subtypes (Table 2B). It is well established that homozygous JAK2V617F, positive clones are much more prominent in PV than ET41, an observation that is consistent with the suggestion that elevated JAK2 signalling promotes an erythroid phenotype. The finding that leukocyte counts in murine JAK2V617F retroviral-mediated transfection/transplantation models differ in a strain-dependent manner was the first suggestion that host genetic factors may influence the disease phenotype. Here, we have identified HBSSL-MYB as a locus outside JAK2 that influences the phenotype of human MPN.

In conclusion, our study indicates that genetic variation at multiple loci contributes to the risk of developing a MPN. The functional basis for these associations is largely unknown, but we have linked a candidate risk SNP at 6q26 to lower levels of MYB expression and an ET phenotype. Further investigation will be needed to understand how genetic variation at JAK2, MECOM, TERT and other loci predispose to MPN.

Methods
Stage-1 discovery data set. Before quality control, the stage-1 cohort consisted of 571 cases of JAK2V617F-negative MPNs from the United Kingdom (referred to as UK1), which comprise 556 with ET and 15 with PMF. Of these, 492 were sourced from the UK Primary Thrombocytopenia 1 (PT-1) study, which included newly diagnosed and previously treated patients, aged 18 years or over, who met the Polycythaemia Vera Study Group (PVSG) criteria for ET45. One hundred and one are from the NCRI MPN sample bank at Addenbrookes Hospital, Cambridge. The remaining cases (\(n = 78\)) were recruited from the NCCI MPN Study Group members following clinical and morphological review to identify bona fide MPN cases. For stage-1 controls, the National Blood Service (NBS) cohort from the Wellcome Trust Case Control Consortium (WTCCC2; \(n = 2706\); www.wtccc.org.uk) was used.

Replication cohorts. Five independent cohorts of JAK2V617F-negative MPN who met the Polycythaemia Vera Study Group or World Health Organization criteria for ET or MF were used to validate significant SNPs from stage-1 (Supplementary Table 1). These cohorts consisted of 198 cases from the United Kingdom and 2,706 controls from the WTCCC21958 British Birth Cohort (BBC, Stage-2.1), 286 cases from a combination of Austrian46 (\(n = 99\)) and German cases (\(n = 187\)) and 1,805 Bavarian controls from the German Cooperative Health Research in the Region of Augsburg cohort (KORA F4; Stage-2.2)47. 460 Greek cases and 1,196 locally recruited Italian controls from a combination of controls from the Invecchiare in Chianti study (InCHIANTI; \(n = 1210\))45, and a series of locally recruited Italian controls (Stage-3.2) from a combination of controls from the Invecchiare in Chianti study (InCHIANTI; \(n = 1210\))45, and a series of locally recruited Italian controls (\(n = 398\)). To further validate our findings, 406 UK cases with elevated levels of platelets and/or erythrocytes and a reactive cause were identified. These reactive cases mimic MPN but are non-clonal and were used as a negative control by comparison with 2,706 controls from the BBC (Stage-3.3).

| SNP            | Gene | Alleles | PV MAF | ET MAF | P value | OR (CI) |
|----------------|------|---------|--------|--------|---------|---------|
| rs12339666     | JAK2 | T/G     | 0.5627 | 0.3815 | 4.788 × 10^{-16} | 2.09 (1.75–2.49) |
| rs2201862      | – MECOM | T/C | 0.4817 | 0.489 | 0.7523 | 0.97 (0.81–1.17) |
| rs9376092      | HBSSL-MYB | A/C | 0.2495 | 0.3507 | 7.475 × 10^{-7} | 0.62 (0.51–0.75) |
| rs4858647      | THRB-RARB | A/C | 0.1000 | 0.1012 | 0.9286 | 0.99 (0.74–1.32) |
| rs2736100      | TERT | C/A | 0.6426 | 0.6303 | 0.5663 | 1.06 (0.88–1.27) |

\(P\)-value, odds ratio (OR), and 95% confidence interval (CI) for comparison of JAK2V617F-positive PV (\(n = 505\)) and ET cases (\(n = 499\)) using allelic \(\chi^2\)-test.
To determine whether SNPs with significant replication in one or more of the five follow-up cohorts are associated exclusively with JAK2V617F-negative MPN, two independent cohorts of consisting of 556 ET and PMF JAK2V617F-positive MPN cases and 556 PV JAK2V617F-positive MPN cases from the United Kingdom were genotyped and compared with 2,706 controls from the BBC (Stage-3 and 4.5 Supplementary Table 1). The study was approved by the following International Review Boards and/or ethics committees: National Research Ethics Service UK (UK); Committee South West; London Multicentre Research Ethics Committee; Cambridgeshire4 Research Ethics Committee; Trent Research Ethics Committee; Ethics Committee of the Biomedical Research Foundation of the Academy of Athens; Comitato Etico, Azienda Ospedaliero-Universitaria Careggi, Firenze; Comité de Bioética, Facultad de Medicina, Universidad de San Martin, Politicas, Ethics committee of the University of Ulm; Medizinische Ethikkommission II der Medizinischen Fakultät Mannheim; Ethikkommission der Medizinischen Universität Wien and informed consent was obtained from participants according to the Declaration of Helsinki.

Genotyping. The stage 1 cases were genotyped by AROS Applied Biotechnology (Aarhus, Denmark) using Affymetrix SNP6 arrays. To reduce genotyping errors and ensure data quality, 25% of the stage 1 cases were included in a quality control set of 30 cases. These were genotyped and five of these SNPs were selected for genotyping in the second validation round (Stage 2). 203 SNPs were genotyped and five of these SNPs were selected for genotyping in the second validation round (Stage 3). The Austrian cases, Bavarian controls and WTCCC using Affymetrix SNP6 arrays and the CHIAMO algorithm to genotypes, the raw genotypes, were pooled and the Birdseed v2 algorithm in Affymetrix's genotyping console was used to determine genotypes. The stage 1 controls were previously genotyped by the WTCCC using Affymetrix SNP6 arrays and the CHIAMO algorithm to determine genotypes. Excluding the WTCCC2 controls, KORA controls and Austrian cases, the replication cohorts were genotyped by LGC Genomics (formerly KBioscience, Hoddesdon, UK). In the initial validation round (Stage 2), 203 SNPs were genotyped and five of these SNPs were selected for genotyping in the second validation round (Stage 3). The Austrian cases, Bavarian controls and WTCCC2 controls were previously genotyped using Affymetrix SNP6 arrays. Stage-3.4 and stage-3.5 positive MPN cases from the United Kingdom were genotyped using the Illumina Human OmniExpressExome v1.2 platform by Gen-Probe Life Sciences Ltd (Wythenshawe, UK).

Quality control for genotyping. Quality control of the stage 1 genotypes involved the removal of samples with poor resolution of SNP clusters (n = 12 samples with contrast Q < 0.4), SNPs with MAF < 0.01, SNPs without observations for all three genotypes, SNPs and samples with >10% missing genotypes. A number of measures were used to control for poor genotyping. First, SNPs with significant deviations from Hardy–Weinberg equilibrium (HWE) in control samples (Exact test P value ≤ 0.001) were removed. Since the cases and controls at stage-1 were not genotyped at the same time, we also removed SNPs with extreme deviation from HWE in cases (Exact test P value ≤ 1 × 10^-10) and SNPs with an average genotyping confidence value over all cases that were ≥ 2 s.d. values worse than the mean over all SNPs to control for potential genotyping errors in cases. To further control for poor genotyping, SNPs with allele frequencies that were significantly different between the two WTCCC2 control cohorts (NBS and CBC) and Caucasians from HapMap were also removed. Finally, SNPs located on non-autosomal chromosomes were removed and CNVs were not considered. A total of 266,882 SNPs were removed from the data set during these QC steps leaving observed genotypes at 642,633 SNPs for analysis. Pairwise identity by state scores (IBS) were used to test the stage-1 samples for cryptic relatedness and nine individuals with IBS scores ≥ 80% were excluded from the stage 1 samples. The IBS scores were calculated to perform a multidimensional scaling analysis (MDS) to identify ethnic outliers when referenced against the HapMap populations. On this basis, 26 of the stage-1 samples whose genotypes did not concur with a European ancestry were removed. The remaining cases clustered tightly with the WTCCC2 controls in the MDS plot, which demonstrates that the cases and controls are ethnically compatible (Supplementary Fig. 6). After these quality control procedures 524 cases (from 571 genotyped) and 2,674 controls, and an age standardized disease prevalence of one per 5,000 European persons were used.

The power of our stage-1 cohort to detect markers associated with the risk of developing MPNs was estimated using the genetic power calculator59 with values of 524 cases and 6,674 controls, and a standardized disease prevalence of one per 5,000 European persons. Modest values of genotype relative risk (1.1–2.0) were used and it was assumed that the causal allele had been tagged with D’ = 0.8 and controls are unselected. In general, study power is diminished by lower causal allele frequency, and so a range of conservative estimates (0.05–0.3) for this value were used.

GWA studies are prone to detection of false positives and our stage-1 sample size is modest. To ensure that the most promising signals were taken through to stage-2, we therefore selected SNPs on the basis of the following criteria rather than significance alone. First, a ‘clumping’ procedure60 was used to select 120 SNPs with allele P values ≤ 0.001 and support from correlated SNPs (r^2 > 0.5, within 500 kb, and allele P value ≤ 0.01). Second, 22 SNPs were selected that have allele P values ≤ 0.02 and are within 500 kb of loci associated with normal variation in platelet number (PLT) and or platelet volume (MPV)61. An additional 11 SNPs with allele P values ≤ 0.0007 were chosen on the basis of their location within or near genes with functional relevance. Finally, 50 SNPs with allele P values ≤ 0.0002 and without correlated SNPs (r^2 ≥ 0.5) were also selected. For regions with multiple SNPs with significant association, the most significant SNP was selected as the lead along with any other SNP(s) that were not in high LD with the lead SNP (r^2 ≤ 0.6). Using these selection techniques a total of 203 SNPs were identified.

A fixed effects inverse variance-weighted meta-analysis of the stage 1 and stage-2 results was used to select five SNPs with meta-analysis P values ≤ 0.001 and effects in the same direction for further analysis at stage-3. The same methodology was used to determine the final effect size and significance by combining the results from the six cohorts in STATA v11.0 (ref. 51) was used to perform these meta-analyses under the assumption of one true effect size (odds ratio) and that any differences in observed effects between studies are due to sampling error. Two methods were used to estimate the between-study heterogeneity, these are the χ^2-based Cochran’s Q statistic and the I^2 statistic, which describe the percentage of variation across studies that is due to heterogeneity rather than chance.

To assess the impact of our findings at a population level, we computed population attributable risks (PAR), which estimate the proportion of MPN cases that can be attributed to each polymorphism. We calculated this as, PAR = (x − 1/x)100, where x = (1 − r^2 + P OR1 + P OR2 + P OR3), r is the frequency of the risk allele in the control cohort, and OR1, OR2, and OR3 are the genotypic ORs associated with heterozygosity and homozygosity, respectively, and relative to common homozygotes. A combined PAR was also calculated as PAR = 1 − (1 − PAR1) × (1 − PAR2) × (1 − PAR3) incorporating all risk factors.

To determine whether SNPs with significant replication were associated exclusively with JAK2V617F-negative MPN, a Breslow–Day test was used to assess the homogeneity of effect sizes (ORs) among the V617F negative and positive cohorts (Breslow–Day P values ≤ 0.05 were considered statistically significant). The Breslow–Day method was also used to assess the homogeneity of effect sizes (ORs) in cases with or without somatic mutations in either CALR or MPL.

We considered the possibility that the observed associations might arise as a consequence of clonal somatic changes in the MPN cases. Although it is theoretically possible that the constitutional genotypes of the intragenic SNPs we identified might be altered by somatic point mutations, this is clearly extremely unlikely in a disorder for which the rate of somatic mutation is known to be low (median number of somatic mutations in ET and PV = 6.5/exome and for MF 13/exome). More plausible is the possibility that constitutional genotypes might be altered by small regions of loss of heterozygosity (LOH) that are below the limit of detection of SNP arrays. Loss of heterozygosity in the majority of cells would result in germline heterozygous SNPs being scored as homozygous. If the LOH was recurrent and occurred randomly with respect to SNP genotype then it would have no effect on the allele frequency. For example, a SNP with a MAF of 50% would normally be expected to result in a genotype distribution of 25% a/a, 50% a/b and 25% b/b. Recurrent somatic LOH in say 50% of cases would result in an apparent genotype distribution of 30% a/a, 40% a/b and 30% b/b. The MAF is still 0.5, and thus there is no effect on the association study. If, however, small deletions/UPD were not random with respect to SNP genotype (for example, if they were part of a particular haplotype that conferred a selective advantage when lost or gained) then we would expect to see distortions in genotypes that could potentially affect the GWAS results. For example above, if a haplotype had a selective advantage over a/b, then a recurrent somatic deletion/UPD in 20% of cases would result in an apparent genotype distribution of 35% a/a, 40% a/b and 25% b/b. The apparent MAF is 0.45 and thus could be picked up by the GWAS if the effect was large enough. However, in both instances the genotypes would be skewed away from the expected allele frequency. PAR1 was calculated as PAR1 = [(1 − P OR1) × (1 − P OR2)] × (1 − P OR3).

Statistical analysis. We used allelic χ^2 to compare allelic frequencies between cases and controls and calculate odds ratios (ORs) and 95% confidence intervals (CIs). To examine the possibility of confounding factors in our stage-1 data, as well as population stratification and separate genotyping of cases and controls, we used the qnorm and qqplot procedures in R10 to construct a quantile-quantile (Q-Q) plot of these data under the null hypothesis (Fig. 2). Manhattan plots were generated using HaploView 4.2 (ref. 48), and regional plots were generated using LocusZoom59.
agreement with expected Hardy–Weinberg distribution: as indicated in Supplementary Table 7, none showed an exact test P value P<0.05 in our data. If we simulate LOH randomly affecting just 10% of cases the exact test P values change markedly (final column): 5/6 deviate significantly from expected at exact test P<0.05 and 3/6 deviate significantly at exact test P<0.001. The lack of any significant deviation strongly suggests that clonal changes have not contributed to the significance of SNPs in this study.

**Statistical analysis of risk variants in PT-1.** For continuous variables, analyses comparing diagnostic variables and genotype were performed using linear models, correcting for JAK2 status and using the dosage of alleles as the predictor variable for the SNP of interest. For categorical variables, the analysis used the χ²-test for trend across the dosage of alleles of the SNP of interest. Clinical outcome was assessed by χ²-tests for events preceding diagnosis and log-rank test for events after trial entry. Linear mixed effect models were used to assess the relationships between the SNP and response of blood counts to therapy.

**Chromosomal abnormalities identified by analysis of SNP array data.** To identify regions of acquired uniparental disomy (aUPD) and indels, PennCNV43 and Affymetrix Power Tools (APT www.affymetrix.com/support/developer/powertools/index.affx) were used to convert the raw CEL files into a normalized signal intensity file that were comparable in terms of raw signal intensities and distributions. Log R Ratio (LRR) and B Allele Frequency (BAF) were calculated for each marker in each individual using the allele-specific signal intensities from the normalized signal intensity file. BAF segmentation55 was used to exclude non-informative homozygous SNPs (BAF >0.9 or BAF <0.1) and SNPs where the absolute difference in BAF between preceding or succeeding SNPs was greater than 0.6. Mirrored BAF (mBAF) values were calculated by reflection at 0.5, and a circular binary segmentation (CBS) algorithm was used to identify regions of similar allelic proportions with mBAF above a threshold of 0.56. Finally, regions of allelic imbalance were compared with log R Ratios to determine whether they were indels or copy number neutral aUPD.

**Screening JAK2, CALR and MPL for somatic mutations.** Samples were screened for JAK249,517,518 using a range of techniques employed at the recruiting centres. Mutations in CALR were detected using fragment analysis PCR and characterized using Sanger sequencing. For fragment analysis, one primer was labelled with either 6-carboxyfluorescein (6-FAM) or 6-carboxyfluorescein hexachloro (HEX). PCR was performed using 1.25U HotStar Taq (Qiagen, Crawley, UK), 1 x HotStar Taq buffer, 1.5 mM MgCl2, 200 μM dNTPs, 0.5 μM each primer (CALR_FA_F and CALR_FA_R, primers listed in Supplementary Table 4), 20–40 ng DNA in a 25 μl reaction. Cycling conditions were 95°C for 10 min, 30 cycles of 95°C for 1 min, 60°C for 1 min, 72°C for 1 min, followed by 72°C for 30 min. Products were diluted 1:500 and 1/10,000 in sterile water and resolved by capillary gel electrophoresis, by sizing on a 3130x Genetic Analyser (Life Technologies, Paisley, UK). Results were analysed using GeneMarker software version 1.85 (SoftGenetics, State College, PA, USA) and analysed on the blue (FAM) or green (HEX) channel. For Sanger sequencing, PCR was carried out using primers (CALR_seq_F and CALR_seq_R) and following identical reaction and cycling conditions to that used for fragment analysis, apart from a final extension at 72°C for 10 min. Sequences were analysed using Mutation Surveyor 3.1 software (SoftGenetics, State College, PA, USA).

Specific pyrosequencing assays were used to detect MPL W515 mutations, following standard PCR, cycling and procedural methods described previously59. For each sample, primers designed to amplify the region in question (MPL_pyro_F and MPL_pyro_R, sequences listed in Supplementary Table 4) were used to create a biotinylated PCR product, which was then used in a pyrosequencing reaction using a sequencing primer (MPL_pyro_seq) for both the MPL W515L and MPL W515K mutation (sequences to analyse were 5′-GTGGACTTTTC TGCACTACAGTGA-3′ and 5′-GTGCCAGCTTC TGCACTACAGG TAC-3′, respectively). Pyrosequencing PCR products for MPL were also Sanger sequenced to identify rarer mutations using the reverse primer, following conditions outlined above.

We note that the frequency of MPL and CALR mutations varies between cohorts (Table 2A). Specifically, the frequency of somatic CALR mutations varied between 22 and 53% (average 45%), while somatic mutations in MPL were seen in 3 to 29% (average 11%) of cases. Concerning MPL, the German/Austrian cohort was not sequential and included a substantial number of MPL-mutated cases that were specifically looked for in a previous study. As for CALR, the mutation frequency was lower than that reported previously22 for all cohorts and substantially lower for the Greek cases. We believe that these differences are likely to result from the well known difficulty in accurately defining JAK2 V617F negative ET in routine practice by exclusion of all potential causes of reactive thrombocytosis34. In addition, we cannot exclude the possibility that the frequency of CALR mutations might vary between different MPN populations.

**Gene expression analysis.** To avoid potential confounding effects of undefined somatic mutations, peripheral blood (PB) samples were obtained from anonymised healthy individuals. Mononuclear cells were cultured in methicillin-free medium (H44434; Stem Cell Technologies, Vancouver, Canada) at a density of 2 × 10⁵ cells per ml, following the manufacturer’s instructions. Colonies comprising a minimum of 100 cells were characterized based on morphology as CFU-GM, CFU-GEMM, CFU-E or BFU-E, as described by Stem Cell Technologies. For each individual, colonies were plucked and pooled according to morphology, into monocyte/granulocyte (containing CFU-GM and CFU-GEMM) and erythroid colonies (containing CFU-E and BFU-E). Cells were washed in PBS, lysed in trizol and RNA was reverse transcribed into cDNA using random primers following standard procedures. MYB and HBS1L expression was measured in duplicate using TaqMan assays (Applied Biosystems) and a Corbett Rotor Gene 6000 (Corbett Life Science, Sydney, Australia). Relative MYB (two alternatively spliced isoforms) and HBS1L expression for each target were normalized to GUS expression (Supplementary Tables 4 and 5) values using the delta-delta Ct method, as described by Pfaffl et al.58 Test samples were placed into two groups according to HMIP risk allele status, either carrying one or two risk alleles (AA/AC) or no risk alleles (CC). Complementary DNA from two separate normal granulocyte samples previously defined as carrying the AC and AA/AC risk alleles were used as calibrator samples to test for normality of expression distribution, and serial dilutions of cell line K562 cDNA was used to control for Q-PCR efficiency. For each sample, expression values were compared with the equivalent calibrator control, that is, the AA/AC test group was normalized to the AC control calibration, and the GC test group was normalized to AC calibration control. The Mann–Whitney test was used to evaluate the difference in MYB and HBS1L expression levels, comparing normalized expression values for those that were grouped as AA/AC versus those grouped as CC, for each of the PB granulocyte, BFU-E and CFU-GM cell subtypes. Fold difference in gene expression was determined by subtracting the median normalized expression level of the GC group by that of the AA/AC group, for each cell type subset.

**Functional annotation of candidate genes.** For identification of candidate genes and previous trait associations, we considered the nearest gene, and any other gene located within 500 kb of the sentinel SNP to be a candidate for mediating pre-disposition to MPN. Genomic regions carrying risk loci and candidate genes were investigated for other trait associations by inspecting a database of all SNP associations that reach genome-wide significance, curated by the National Human Genome Research Institute GWAS catalogue59. Only associations potentially relevant to haematopoiesis were considered.

To predict functional effects we used HaploReg60 to identify SNPs linked to the sentinel SNPs (>0.8 in Europeans from the 1000 Genomes Project) and, on the basis of data from the Encyclopedia of DNA Elements (ENCODE)61 and the NHRI Roadmap Epigenomics Mapping Consortium62, to annotate them with respect to (i) histone modifications, (ii) DNase hypersensitivity, (iii) transcription factor binding, (iv) disruption of regulatory motifs and (v) conservation across mammals estimated by GERF and SIFT. To further investigate the sentinel and linked variants we used RegulomeDB63 to assign functional scores on the basis of data from Gene Expression Omnibus (GEO), ENCODE and published literature. A RegulomeDB score of 5 and 6 indicates there is minimal evidence for functional activity. Finally, we investigated whether our index and linked SNPs overlapped with regions of nucleosome depletion generated by formaldehyde-assisted isolation of regulatory elements (FAIRE-seq), as depleted regions have been shown to have greater epigenetic activity and mark the presence of functional variants that alter the underlying DNA sequences, resulting in different binding affinities for transcription factors44. For this we obtained filtered and binned BED files containing FAIRE-seq data from primary human megakaryocytes and erythroblasts64, and intersected SNP positions with FAIRE regions to look for regions of colocalisation.

All candidate genes were investigated for expression in hematopoietic cells, by interrogating browsers of gene expression profiles generated in purified progenitors of different lineages of the human haematopoietic system (HaemAtlas65 and HemExplorer66) and recently published data set67. To investigate the relationship between the sentinel SNPs and gene expression levels, two expression quantitative trait (eQTL) browsers were used that curate derived from immortalized lymphoblastoid cell lines and primary monocytes; an eQTL browser hosted by the Pritchard laboratory18 and GENEVAR19.

**References**

1. Vardiman, J. W., Harris, N. L. & Brunning, R. D. The World Health Organization (WHO) classification of the myeloid neoplasms. Blood 100, 2292–2302 (2002).  
2. Klampfl, T. et al. Somatic mutations of calreticulin in myeloproliferative neoplasms. New Engl. J. Med. 369, 2379–2390 (2013).  
3. Nangalia, J. et al. Somatic CALR mutations in myeloproliferative neoplasms with nonmutated JAK2. New Engl. J. Med. 369, 2391–2405 (2013).  
4. Landgren, O. et al. Increased risk of polycythemia vera, essential thrombocythemia, and myelofibrosis among 24,577 first-degree relatives of 11,039 patients with myeloproliferative neoplasms in Sweden. Blood 112, 2199–2204 (2008).

© 2015 Macmillan Publishers Limited. All rights reserved.
31. Menzel, S. et al. JAK2 haplotype is a major risk factor for the development of myeloproliferative neoplasms. *Nat. Genet.* 41, 446–449 (2009).
32. Ganesh, S. K. et al. The HBSIL-MYB intergenic interval associated with elevated HbF levels shows characteristics of a distal regulatory region in erythroid cells. *Blood* 114, 1254–1262 (2009).
33. Farrell, J. I. et al. A 3-bp deletion in the HBSIL-MYB intergenic region on chromosome 6q23 is associated with HbF expression. *Blood* 117, 4935–4945 (2011).
34. Stadthouder, R. et al. Dynamic long-range chromatin interactions control Mybproto-oncogene transcription during erythroid development. *EMBO J.* 31, 991–1002 (2012).
35. Thein, S. L. et al. Intergenic variants of HBSIL-MYB are responsible for a major quantitative trait locus on chromosome 6q23 influencing fetal hemoglobin levels in adults. *Proc. Natl Acad. Sci. USA* 104, 11346–1151 (2007).
36. Wahlgren, K. et al. The HBSIL-MYB intergenic interval associated with elevated HbF levels shows characteristics of a distal regulatory region in erythroid cells. *Blood* 114, 1254–1262 (2009).
37. Frizzera, B. et al. A 3-bp deletion in the HBSIL-MYB intergenic region on chromosome 6q23 is associated with HbF expression. *Blood* 117, 4935–4945 (2011).
38. Stadthouder, R. et al. Dynamic long-range chromatin interactions control Mybproto-oncogene transcription during erythroid development. *EMBO J.* 31, 991–1002 (2012).
39. Thein, S. L. et al. Intergenic variants of HBSIL-MYB are responsible for a major quantitative trait locus on chromosome 6q23 influencing fetal hemoglobin levels in adults. *Proc. Natl Acad. Sci. USA* 104, 11346–1151 (2007).
40. Lu, J. et al. MicroRNA-mediated control of cell fate in megakaryocyte-erythrocyte progenitors. *Dev. Cell* 14, 843–853 (2008).
41. Godfrey, A. L. et al. JAK2V617F homozygosity arises commonly and recurrently in PV and ET but is characterized by expansion of a dominant homozygous subclone. *Blood* 120, 2704–2707 (2012).
42. Zaleskas, V. et al. Molecular pathogenesis and therapy of polycythemia induced in mice by JAK2V617F. *PLoS ONE* 1, e18 (2006).
43. Klampfl, T. et al. Genome integrity of myeloproliferative neoplasms in chronic phase and during disease progression. *Blood* 118, 167–176 (2011).
44. Hölle, R., Happich, M., Löwel, H. & Wichmann, H. E. MONICA/KORA Study Group. KORA—a research platform for population based health research. *Gesundheitswesen* 67, S19–S25 (2005).
45. Ferrucci, L. et al. Subsystems contributing to the decline in ability to walk: bridging the gap between epidemiology and geriatric practice in the InCHIANTI study. *J. Am. Geriatr. Soc.* 48, 1618–1625 (2000).
46. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575 (2007).
47. Becker, R. A., Chambers, J. M. & Wilks, A. R. The New S Language: A Programming Environment for Data Analysis and Graphics (Wadsworth & Brooks/Cole, 1988).
48. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 262–265 (2005).
49. Pruim, R. J. et al. LocusZoo: regional visualization of genome-wide association scan results. *Bioinformatics* 26, 2336–2337 (2010).
50. Ma, X., Vanasse, G., Carmetl, B., Wang, Y. & Selinger, H. A. Prevalence of polycythemia vera and chronic thrombocythemia. *Am. J. Hematol.* 83, 296–309 (2008).
51. Harris, R. et al. Metan: fixed- and random-effects meta-analysis. *Stat. J.* 8, 3–28 (2008).
52. Thompson, S. G. & Higgins, J. P. How should meta-regression analyses be undertaken and interpreted? *Stat. Med.* 21, 1559–1573 (2002).
53. Matakidou, A. et al. Risk of non-medullary thyroid cancer influenced by polymorphisms in the thyroid transcription factor-1 gene. *Carcinogenesis* 25, 369–373 (2004).
54. Wang, K. et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* 17, 1665–1674 (2007).
55. Staa, J. et al. Segregation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. *Genome Biol.* 9, R137–R137 (2008).
56. Jones, A. V. et al. Minimal molecular response in polycythemia vera patients treated with imatinib or interferon alpha. *Blood* 107, 3339–3341 (2006).
57. Harrison, C. N. et al. Guideline for investigation and management of adults and children presenting with a thrombocytosis. *Br. J. Haematol.* 149, 352–375 (2011).
58. Piffal, M. W. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res.* 29, e45 (2001).
59. Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 42, D1001–D1006 (2014).
60. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* 40, D930–D934 (2012).
61. ENCODE Project Consortium et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74 (2012).
62. Bernstein, B. E. et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* 28, 1045–1048 (2010).
63. Boyle, A. P. et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* 22, 1790–1792 (2012).
64. Paul, D. S. et al. Maps of open chromatin guide the functional follow-up of genome-wide association signals: application to hematological traits. *PLoS Genet.* 7, e1002139 (2011).
65. Paul, D. S. et al. Maps of open chromatin highlight cell type–restricted patterns of regulatory sequence variation at hematological trait loci. *Genome Res.* 23, 1130–1141 (2013).
66. Watkins, N. A. et al. A HaemAtlas: characterizing gene expression in differentiated human blood cells. *Blood* 113, e1–e9 (2009).
67. Baggar, F. O. et al. HemaExplorer: a database of mRNA expression profiles in normal and malignant haematopoiesis. *Nucleic Acids Res.* 41, D1034–D1039 (2013).
Acknowledgements
W.T., A.C., W.L., K.W. and N.C.P.C. were supported by Leukaemia and Lymphoma Research. A.V.J. was supported by the Kay Kendall Leukaemia Fund. Work in the Italian centers was supported by a grant from Associazione Italiana per la Ricerca sul Cancro (AIRC, Milano) ‘Special Program Molecular Clinical Oncology 5 × 1000’ to AGIMM (AIRC-Gruppo Italiano Malattie Mieloproliferative). A detailed description of the AGIMM project is available at http://www.progettagimin.it. A full list of the investigators who contributed to the generation of the WTCCC data is available from www.wtccc.org.uk, funding for which was provided by the Wellcome Trust under award 07611. The KORA research platform (KORA, Cooperative Health Research in the Region of Augsburg) was initiated and financed by the Helmholtz Zentrum München—German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research and by the State of Bavaria. Furthermore, KORA research was supported within the Munich Center of Health Sciences (MC Health), Ludwig-Maximilians-Universität, as part of LMUinnovative. The InCHIANTI study baseline (1998–2000) was supported as a ‘targeted project’ ([CS1101/RF97.71] by the Italian Ministry of Health and in part by the U.S. National Institute on Aging (Contracts: 263 MD 9164 and 263 MD 821336). Samples were provided by the Cambridge Blood and Stem Cell Biobank, which is supported by the Cambridge NIHR Biomedical Research Centre, Wellcome Trust—MRC Stem Cell Institute and the Cambridge Experimental Cancer Medicine Centre, UK.

Author contributions
The study was designed and co-ordinated by W.T., A.V.J. and N.C.P.C. W.T. performed statistical analysis and drafted the manuscript. A.V.J. coordinated the laboratory analysis, which was performed by A.V.J., W.L., A.C., D.W., P.A., H.E.W., K.W., F.L., A.C. and E.C. SNP array and clinical data from Austrian cases was provided by R.K., A.S.H. and H.G. DNA and clinical data from UK cases were provided by A.G., E.I.B., C.M., J.N., P.E., M.S., A.J., L.W., D.O., A.S.D., A.S., A.I.M., M.G., J.E., R.E.G., C.H., P.C. and A.R.G. DNA and clinical data from Greek cases were provided by K.Z. and A.G. DNA and clinical data from German cases were provided by S.S., T.H., F.S., K.D. and A.R. DNA and clinical data from Italian cases were provided by P.G., L.P., C.M., G.B., M.C. and A.V. Control data from the InCHIANTI cohort was provided by T.T. and S.B. Control data from the KORA cohort were provided by K.S. and H.G. All authors critically reviewed the manuscript.

Additional information
Accession codes. The primary array data are available under ArrayExpress accession: E-MTAB-3340.

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: S.S. and T.H. are part owners of the MLL Munich Leukemia Laboratory GmbH. The remaining authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Tapper, W. et al. Genetic variation at MECOM, TERT, JAK2 and HBS1L-MYB predisposes to myeloproliferative neoplasms. Nat. Commun. 6:6691 doi: 10.1038/ncomms7691 (2015).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material.

To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/