Germline-somatic JAK2 interactions are associated with clonal expansion in myelofibrosis

Derek W. Brown1,2, Weiyin Zhou1,3, Youjin Wang1, Kristine Jones1,3, Wen Luo1,3, Casey Dagnall1,3, Kedest Teshome1,3, Alyssa Klein1, Tongwu Zhang1, Shu-Hong Lin1, Olivia W. Lee1, Sairah Khan1, Jacqueline B. Vo1,2, Amy Hutchinson1,3, Jia Liu1,3, Jiahui Wang1,3, Bin Zhu1,3, Belynda Hicks1,3, Andrew St. Martin4, Stephen R. Spellman5, Tao Wang4,6, H. Joachim Deeg7, Vikas Gupta8, Stephanie J. Lee4,7, Neal D. Freedman1, Meredith Yeager1,3, Stephen J. Chanock1, Sharon A. Savage1, Wael Saber4, Shahinaz M. Gadalla1,9 & Mitchell J. Machiela1,9

Myelofibrosis is a rare myeloproliferative neoplasm (MPN) with high risk for progression to acute myeloid leukemia. Our integrated genomic analysis of up to 933 myelofibrosis cases identifies 6 germline susceptibility loci, 4 of which overlap with previously identified MPN loci. Virtual karyotyping identifies high frequencies of mosaic chromosomal alterations (mCAs), with enrichment at myelofibrosis GWAS susceptibility loci and recurrently somatically mutated MPN genes (e.g., JAK2). We replicate prior MPN associations showing germline variation at the 9p24.1 risk haplotype confers elevated risk of acquiring JAK2V617F mutations, demonstrating with long-read sequencing that this relationship occurs in cis. We also describe recurrent 9p24.1 large mCAs that selectively retained JAK2V617F mutations. Germline variation associated with longer telomeres is associated with increased myelofibrosis risk. Myelofibrosis cases with high-frequency JAK2 mCAs have marked reductions in measured telomere length – suggesting a relationship between telomere biology and myelofibrosis clonal expansion. Our results advance understanding of the germline-somatic interaction at JAK2 and implicate mCAs involving JAK2 as strong promoters of clonal expansion of those mutated clones.

Myelofibrosis (MF) is a rare myeloproliferative neoplasm (MPN) with an incidence of ~1 per 100,000 per year1,2, characterized by the development of abnormal hematopoietic stem cell (HSC) clones and altered bone marrow microenvironment, leading to fibrosis3,4. Patients with MF typically develop cytopenia due to proliferation of aberrant HSC clones and hepatosplenomegaly due to extramedullary hematopoiesis5. Individuals with MF are at a high risk of developing acute myeloid leukemia6. MF can present as primary disease (primary MF) or progress from another MPN (secondary MF) such as polycythemia vera or essential thrombocythemia7.

Recurrent somatic driver mutations have been identified in MF, particularly in JAK2, MPL, and CALR7–9; current evidence also indicates a heritable component as well10,11. So far, the JAK2 46/1 haplotype and a single nucleotide polymorphism (SNP) in the TERT gene region have been established as predisposition alleles for MPNs, including both primary and secondary MF12–16, which have led to the use of JAK...
inhibitors and telomerase inhibitors as potential therapeutic agents for MF. A genome-wide association study (GWAS) of MPN that included 136 MF patients identified additional MPN-associated loci (e.g., 3q21.3, 3q25.33, 6p21.31, 13q14.11, 18q11.2, 21q22.12). Current knowledge of MF genetic etiology is inferred from the MPN data.

In this work, we undertook an integrated approach to investigate the genetics of MF which includes analysis of germline variation, somatic JAK2 point mutations, somatic mosaic chromosomal alterations (mCAs), and leukocyte telomere length among MF patients who underwent hematopoietic cell transplantation (HCT) and reported to the Center for International Blood and Marrow Transplant Research (CIBMTR), representing a clinically important subset of MF patients. We report here six MF susceptibility loci, four of which replicate prior MPN findings at 9p24.1 (JAK2), 5p15.33 (TERT), 3q25.33 (IFT80), and 4q24 (TET2). We show germline variation at the 9p24.1 risk haplotype confers elevated risk of acquiring cis JAK2V617F mutations which are selectively retained by recurrent 9p24.1 mCAs. MF is dynamically associated with telomere length in which longer inherited telomere length increases MF risk, and clonal expansion of 9p24.1 mCAs markedly reduces measured telomere length.

Results
CIBMTR MF case characteristics
In total, 937 MF cases met the criteria for inclusion in our study, the majority of which were male (58.06%) and had DNA collected at an average age of 56.9 years (median = 58.4, IQR = 52.3–63.9; Supplemental Table 1). Most cases were primary (68.84%) MF, with intermediate 1 or 2 disease (49.30%) based on the Dynamic International Prognostic Scoring System (DIPSS) score. The average time from diagnosis to transplant in the full cohort was 63.5 months (median = 25.1, IQR = 9.0–88.3). Compared to primary MF cases, secondary MF cases were more likely to be female (54.45% vs. 36.28%) and had longer average time from diagnosis to transplant (122.8 months vs. 36.8 months).

MF susceptibility loci identified with estimated large effect size
We performed a GWAS using 827 MF cases and 4135 ancestry-matched cancer-free controls drawn from the Prostate, Lung, Colorectal, and Ovarian (PLCO) Screening Trial (Methods section). The liability scale heritability of MF was estimated to be 11.4% (s.e.= 5.8%). The genomic inflation (λ) and intercept from linkage disequilibrium score regression (LDSC) showed minimal evidence for systematic inflation (λ = 1.02, LDSC intercept= 1.01, Fig. 1).

We analyzed 9,672,066 genotyped and imputed germline variants after filtering on control minor allele frequency (<0.5%) and imputation quality score (>0.7; Methods section) and identified six independent genome-wide significant loci (P < 5 × 10⁻⁸) (Supplemental Table 2 and Fig.1), four of which replicated prior MPN findings at 9p24.1 (JAK2), 5p15.33 (TERT), 3q25.33 (IFT80), and 4q24 (TET2). We observed rs781556 in 9p24.1 as the most significant variant (odds ratio (OR) = 2.39, 95% confidence interval (CI) = 2.13–2.65, P = 5.75 × 10⁻⁷), which is in strong linkage disequilibrium (LD) with the lead variant reported for MPN, rs3324794 (R²EUR = 0.95, D'EUR = 0.98) and located in an intron of JAK2, a gene which promotes cellular proliferation through the JAK/STAT pathway [14]. As chromosomal alterations are common at 9p24.1 (see mCA section below), we performed a sensitivity analysis in individuals with no detectable 9p24.1 mCAs to ensure no miscalling of germline variants resulting from mCAs in the 9p24.1 region; the rs781556 signal remained significant (OR = 1.65, 95% CI = 1.41–1.92, P = 1.99 × 10⁻⁶), although the effect estimate was attenuated due to removal of MF cases that carried the JAK2 risk haplotype (see below section on JAK2–mCA relationship). rs7705526 had the strongest association (OR = 1.65, 95% CI = 1.45–1.84, P = 7.62 × 10⁻⁶) in the 5p15.33 locus. Previously identified in MPN [20], this intronic variant is located in TERT, which encodes telomerase, the reverse transcriptase that extends telomeric DNA repeats, and has been associated with CD34+ to CD45+ ratio [25]. The 3q25.33 variant rs201009932 (OR = 5.78, 95% CI = 3.67–9.11, P = 4.06 × 10⁻⁴) (Supplemental Table 2) is in moderate LD with the MPN variant rs77429081 (R²EUR = 0.25, D'EUR = 1.00) and resides in an intron variant of IFT80, a part of the IFT complex essential for the assembly and maintenance of cilia as well as differentiation through the Sonic Hedgehog pathway [26,27]. At 4q24, rs1345483, located near TET2, a putative tumor suppressor gene and common somatic driver mutation in clonal hematopoiesis [8], was significant (OR = 2.27, 95% CI = 1.71–3.01, P = 1.42 × 10⁻³), notably, this variant is in high LD with the previously identified MPN variant rs62329718 (R²EUR = 0.94, D'EUR = 1.00). Conditional analyses controlling for the lead GWAS variant for the 9p24.1 and 5p15.33 loci identified no evidence for additional independent signals (Supplemental Fig. 1).

Our GWAS identified two additional MF germline susceptibility loci: 6p21.32 and 17p13.1 (Supplemental Table 2 and Fig.1). The 6p21.32
variant, rs28442287, (OR = 1.63, 95% CI = 1.38–1.93, $P = 9.34 \times 10^{-3}$) is a downstream variant of HLA-DRB9, within the human leukocyte antigen system \(^{30,31}\), and has been implicated in bone marrow related diseases \(^{32,33}\). The 17p13.1 variant, rs78378222, (OR = 4.47, 95% CI = 2.75–7.27, $P = 1.56 \times 10^{-4}$) is a 3’ UTR variant of the commonly mutated tumor suppressor gene TPS3 that increases susceptibility to different types of cancer (e.g., skin basal cell carcinoma, prostate cancer, glioma, and lymphocytic leukemia) \(^{34,35}\). Future studies are warranted to validate these two MF germline susceptibility loci.

We performed stratified GWAS investigating primary MF (569 cases, 2845 controls) and secondary MF (258 cases, 1290 controls) to investigate possible differences (Supplemental Table 2), and no informative differences in association signals were observed between the two MF types (Supplemental Table 2, Supplemental Fig. 2). We further stratified secondary MF analyses into post-polycythemia vera MF (119 cases, 595 controls) and post-essential thrombocythemia MF (139 cases, 695 controls; Supplemental Table 3). We observed a strong genome-wide significant signal at 9p24.1 (JAK2) only in post-polycythemia vera MF (Supplemental Table 3 and Supplemental Fig. 3), which is consistent with prior reports of higher JAK2 involvement in patients with polycythemia vera \(^{36-40}\).

**MF-associated germline variants near JAK2 increase expression**

We performed expression quantitative trait locus (eQTL) analyses using genome-tissue expression (GTEx) whole blood data \(^{41}\) and identified eQTLs with JAK2 expression (e.g., rs7847141; OR = 1.05, 95% CI = 1.03–1.08, $P = 3.69 \times 10^{-3}$). Germline variants at the lead 9p24.1 locus additionally colocalized with JAK2 expression with rs7851556 having the highest Posterior Probability (PP; $PP = 0.59$; Supplemental Fig. 4) \(^{42}\). MF risk alleles were associated with increased levels of JAK2 expression, explaining 35% of the shared association signal at 9p24.1 ($PP_{SNP} = 0.33$) \(^{42}\). Colocalization was also detected at 6p21.32 with HLA-DRB9 (colocalization PP = 0.36) \(^{43}\).

In analysis of whole blood GTEx expression data, a transcriptome-wide association study (TWAS) \(^{44}\) identified JAK2 (Z = 9.00, $P = 2.18 \times 10^{-13}$) and RPI1-39K24.4 (Z = 10.78, $P = 4.26 \times 10^{-13}$) as significant genes ($P < 3.59 \times 10^{-5}$; Supplemental Figure 5), with positive JAK2 expression associated with increased MF risk. Conditional analyses were performed with predicted RPI1-39K24.4 expression, and JAK2 remained an independent expression signal (rs7851556 conditional GWAS $P = 9.90 \times 10^{-4}$; Supplemental Fig. 6).

**JAK2 germline risk haplotype confers elevated risk of cis JAK2V617F mutations**

Targeted PacBio long-read sequencing of the JAK2 region identified 562 (60.82%) individuals with the commonly observed activating JAK2V617F mutation, a known MPN driver mutation \(^{45}\). Secondary MF following a prior diagnosis of polycythemia vera or essential thrombocythemia were more likely to have the JAK2V617F mutation than primary MF cases (67.93% vs. 57.57%, $95\% CI = 1.03-1.08, P = 3.69 \times 10^{-3}$), with post-polycythemia vera MF having the highest frequency of the JAK2V617F mutation (97.04% vs. 57.57%, $P = 5.82 \times 10^{-8}$). The estimated average mutation allelic fraction on background haplotypes was 62.54% (median = 69.76, IQR = 37.58–92.74), suggesting high clonal expansion of JAK2V617F mutated clones. MF cases carrying the risk allele (T) of JAK2V617F mutation calling (Methods section), we identified 5 individuals with evidence of the somatic mutation potentially acquired independently on both germline haplotypes (Supplemental Table 6) which were replicated in independent sequencing runs on new libraries. Future studies are needed to further explore the frequency of independent JAK2V617F mutations on both germline haplotypes in MF cases.

**Chromosomal alterations are abundant in MF and preferentially expand JAK2V617F clones**

At least one detectable autosomal mCA was detected in 684 (73.31%) individuals which is in contrast to ~3% in population-based surveys \(^{45,46}\). An elevated frequency of mCAs in secondary MF compared to primary MF was also observed (78.35% vs. 71.03%, OR = 1.48, 95% CI = 1.06-2.05, $P = 0.019$), with post-polycythemia vera MF having the highest frequency of mCAs (94.12% vs. 71.03%, $P = 3.13 \times 10^{-6}$, $P = 5.54 \times 10^{-7}$). Recurrent copy neutral loss of heterozygosity (CNLOH) events were detected on chromosome 9 (N = 298; 31.94%), and recurrent loss events were observed on chromosome 13q (N = 99; 9.54%) and 20q (N = 92; 9.86%; Fig. 2). Each GWAS susceptibility locus showed enrichment for mCAs compared to age and sex-matched cancer-free individuals in the UK Biobank \(^{47,48}\) (binomial $P = 1 \times 10^{-8}$; Supplemental Table 7), suggesting mCAs could clonally expand MF risk conferring alleles at susceptibility loci.

Since the JAK2 9p24.1 locus was the most notable in our GWAS and contains a hotspot of JAK2V617F mutations, we closely examined mCAs in this region. In total, 378 (40.31%) individuals had a detectable autosomal mCA across the 9p24.1 region. MF cases carrying the risk allele (T) of rs7851556 (our lead GWAS SNP) were more likely to have mCAs within the 9p24.1 region ($P = 2.28 \times 10^{-8}$), with CNLOH representing over 75% of the observed mCAs (N = 294; Supplemental Table 8). These results replicate previous reports of an association between 9p CNLOH and the JAK2 46/1 risk haplotype \(^{49}\). We performed allelic shift analyses in heterozygous individuals to test for a c disturbance between the JAK2 risk haplotype and mCAs spanning 9p24.1 \(^{49,50}\). Using highly correlated genotyped proxy SNPs to our lead GWAS SNP rs7851556, we found that the JAK2 risk haplotype is predominantly amplified by gains (23 of 28, binomial $P = 9.12 \times 10^{-4}$, proxy SNP rs1081567), retained by losses (14 of 16, binomial $P = 4.18 \times 10^{-3}$, proxy SNP rs2230724), and duplicated by CNLOH (51 of 59, binomial $P = 9.05 \times 10^{-4}$, proxy SNP rs1081567), providing strong evidence for preferential clonal expansion of mCAs with the JAK2 MF sequence trajectory.

Using long-read PacBio sequencing data, we noted substantial allelic imbalance of heterozygous variants in the vicinity of JAK2 in individuals with mCAs spanning the region, providing independent confirmation of mCA calls in the region (Supplemental Fig. 7). Of the 374 individuals with detectable mCAs spanning JAK2 who also had long-read sequencing data, 97.86% had a JAK2V617F mutation, indicating a strong relationship between mCAs and JAK2V617F mutations ($P = 4.37 \times 10^{-8}$; Supplemental Table 9). To investigate the clonal evolutionary history of JAK2 mutations, we examined mutated cellular fractions of JAK2V617F and JAK2 mCAs. We found substantially higher JAK2V617F allelic fractions on background haplotypes compared to estimated mCA cellular fractions, suggesting the acquisition of a JAK2V617F mutation occurred prior to acquiring a JAK2V617F mutation, resulting in the vast majority of MF cases (binomial $P = 3.30 \times 10^{-14}$, Supplemental Fig. 8). The estimated JAK2V617F allelic fractions and mCA
cellular fractions may be lower than the true somatic fraction in the actual diseased myeloid cells because we used whole blood DNA from most of the patients. However, this would not affect the ratio of JAK2V617F allelic fraction to mCA cellular fraction. mCAs were also enriched across MPL and CALR gene positions (binomial $P < 5 \times 10^{-32}$; Supplemental Table 10), suggesting mCAs may also clonally expand other MPN driver mutations similar to the JAK2 locus, although this hypothesis needs to be further studied.

**Inherited longer telomere length associated with MF risk**

In addition to the JAK2 MF susceptibility region, the 5p15.33 locus near TERT implicates telomere length in MF risk. To evaluate the role of telomere length, we used a panel of germline variants associated with measured telomere length to develop a polygenic risk score (PRS) for inherited telomere length and increased MF risk (OR = 1.33, 95% CI = 1.23–1.44, $P = 2.56 \times 10^{-13}$). Of the 19 telomere-length associated variants imputed, seven (rs4691895, rs7705526, rs2853677, rs228595, rs62053580, rs75691080, and rs34978822) were nominally associated ($P < 0.05$) with MF risk (binomial $P = 2.31 \times 10^{-5}$). The allele related to longer telomere length was associated with increased risk of MF for five of these seven variants (Fig. 3 and Supplemental Table 11). The telomere length PRS was associated with increased risk of MF for five of these seven variants (Fig. 3 and Supplemental Table 11). The telomere length PRS was associated with the presence of JAK2V617F mutations (OR = 1.20, 95% CI = 1.04–1.37, $P = 0.01$) as well as mCAs (OR = 1.17, 95% CI = 1.01–1.36, $P = 0.04$), suggesting longer telomere length may afford cells the ability to clonally expand to detectable clonal fractions, after acquiring somatic mutations.

In a Mendelian randomization (MR) analysis to evaluate a directional relationship between the telomere length-associated variants and MF risk, the intercept from MR-Egger regression was non-significant ($P = 0.65$, Table 1) after removing five potentially pleiotropic variants (including the lead TERT variant identified in our GWAS) (Supplemental Table 11)[49], suggesting no pleiotropy[50]. Each MR method utilized indicated a strong increasing effect between the telomere length genetic instrument and MF risk (Table 1 and Supplemental Fig. 9).

The genetic correlation between leukocyte telomere length and MF using LD score regression[51] was estimated based on summary statistics from a published telomere length GWAS along with summary statistics from our MF GWAS. A marginally significant genome-wide genetic correlation was observed between telomere length and MF (LDSC $r = 0.23$, s.e.m. = 0.09, $P = 0.037$), similar in magnitude to what was reported for telomere length and MPN (LDSC $r = 0.19$, s.e.m. = 0.09, $P = 0.037$)[20]. Together, our results indicate longer telomere length is associated with increased risk of MF when evaluated by established leukocyte telomere length variants and genome-wide.

**Telomere length attrition and mCA-induced clonal expansion**

In an analysis of leukocyte relative telomere length (rTL) measured before HCT, on average, telomere length was significantly shorter among older subjects ($P = 0.0037$), peripheral blood mononuclear cell
were consistent when restricted to individuals with whole blood-
pared to individuals without autosomal mCAs (Table 13). Multivariable analyses demonstrated negative
greater number of autosomal mCAs (incidence rate ratio= 0.33, 95% 
attenuation of rTL (OR = 0.04, 95% CI = 0.01
Fig. 3 | The effect of each variant on genetically-inferred telomere length and
myelofibrosis risk. Estimates for the SNP-telomere length (Li et al.) and SNP-
myelofibrosis associations are presented in Supplemental Table 11. A linear model
estimated trend line and calculated 95% confidence interval around the trend (gray
fill) are plotted (two-sided P = 5.48 × 10⁻⁴).
Table 1 | Mendelian randomization results using variants and
summary statistics from Li et al.⁴

| Method       | OR (95% CI)       | p-value¹ | Intercept
|--------------|-------------------|----------|------------
| Maximum-likelihood | 8.04 (3.41, 18.93) | 1.88 × 10⁻⁶ | 0.97 (0.85, 1.11) |
| Simple median | 6.25 (2.04, 19.10) | 1.32 × 10⁻³ | 0.97 (0.85, 1.11) |
| Weighted median | 5.07 (1.75, 14.70) | 2.76 × 10⁻² | 0.97 (0.85, 1.11) |
| IVW²       | 7.69 (3.29, 17.97) | 2.43 × 10⁻² | 0.97 (0.85, 1.11) |
| MR-Egger   | 13.49 (1.07, 169.93) | 4.41 × 10⁻² | 0.97 (0.85, 1.11) |

¹All reported tests are two-sided.
²Inverse-variance weighted.
³Five variants (rs4691895, rs7705529, rs34991172, rs228595, rs34978822) were detected to have evidence of pleiotropy (FDR < 0.2) as detailed in Supplemental Table 11 and removed from the MR analyses.

Discussion

We conducted an integrated genomic characterization of MF by investigating both germline susceptibility alleles and somatic events, particularly mosaic chromosomal events. Our study supports a key role of JAK2 events as a critical driver of MF and underscores important interactions of germline susceptibility with JAK2 mutagenesis and mCAs⁵.⁶ We observed that 68% of MF cases carry at least one germline JAK2 risk allele, whereas somatic events are also critical to MF development - in our study 61% of MF cases carried a somatic JAK2 mutagenesis, and 41% with an mCA spanning JAK2. Overall, we observed that -85% of cases involve JAK2. Using independent sequencing runs, we identified five individuals with evidence of acquiring independent JAK2 mutagenesis on both germline haplotypes. The frequency and consequences of independent JAK2 mutagenesis on both germline haplotypes should be further studied. In a GWAS, we identified six MF susceptibility loci (one including JAK2 on 9p24.1) with two independent signals unique to MF⁵. The estimated MF heritability is 11.4% (s.e. = 5.8%), and all MF susceptibility loci had high eEfffect sizes (OR > 1.6) relative to those typically found by GWAS, suggesting a strong germline component at multiple genomic loci for MF risk. Global assessment of mCAs demonstrated a high frequency in MF cases and demonstrated enrichment at each GWAS locus and across other MPN driver mutations, providing potential evidence for genome-wide germline-somatic interactions beyond that observed for JAK2.

Our integrated study demonstrates a complex germline-somatic interaction in MF patients at the 9p24.1 susceptibility locus, which confirmed prior findings and revealed insights into MF etiology. The observed effect of telomere length on disease risk in concert with 9p24.1 indicates that the presence of a germline susceptibility locus involving JAK2 directly influences the probability of developing a somatic event in the same region, presumably on the same haplotype (Fig. 4). We observed that individuals with the germline JAK2 risk haplotype tagged by rs7851556 were predisposed to acquiring a somatic JAK2⁶ mutagenesis in cis, as previously reported⁶. The cis relationship could not be checked directly due to the distance between our lead GWAS variant (rs7851556) and the JAK2⁶ mutagenesis, but this relationship is supported by the high LD between rs7851556 and variants (rs3780367, rs10974934, rs12343867) in the 46/1 risk haplotype (R² = 0.93). We also observed mCAs lead to preferential over-
representation of this risk haplotype containing the JAK2⁶ mutation⁶. Prior studies demonstrate JAK2 is a strong activator of cellular growth and proliferation⁶, promotes cell surface localization⁶, is activated in response to a variety of cytokines⁶.⁷, and

https://doi.org/10.1038/s41467-022-32986-7
induces bone marrow fibrosis. Altered JAK2 activity conferred by the germline JAK2 risk haplotype, JAK2V617F activating mutation, and 9p24.1 mCAs could lead to a cellular phenotype characterized by increased clonal expansion. While we were able to connect germline susceptibility and somatic mutations in cross-sectional data from MF patients, future longitudinal assessment will be key in follow-up of our study.

Likewise future studies characterizing germline functional variation are needed to better understand how the germline JAK2 MF susceptibility locus leads to altered JAK2 expression and acquisition of JAK2V617F mutations.

The MF susceptibility signal near TERT together with our PRS, MR, and genetic correlation analyses, suggest that polygenetic effects of...
SNP genotyping and quality control assessment

We used Qiagen QIA Symphony for DNA extraction. All genomic laboratory work included in this study was conducted by the NCI Cancer Genomics Research Laboratory. Genotyping on MF cases was completed using the Illumina Infinium Global Screening Array 24v1-0. Genotypes were called using standard Illumina microarray data analysis workflows. Controls were genotyped from 62,880 previously genotyped cancer-free individuals within the Prostate, Lung, Colorectal, and Ovarian (PLCO) Screening Trial13, who were genotyped on the same array as the cases. To minimize technical artifacts, genotyping quality control steps were performed on the joint set of MF cases and PLCO controls. Standard quality control checks were performed to ensure high completion rates (>95%), no sample contamination, sex concordance, no unexpected duplicates or replicates, normal rates of heterozygosity, and no instances of high relatedness (IBD < 0.2). Genetic ancestry was inferred using SNP WEIGHTS14, which estimates the percentage of European, West African, and East Asian ancestry for each subject. After filtering based on quality control steps and European ancestry (>80%), 833 MF cases and 56,929 cancer-free controls were eligible for the GWAS analysis.

Genome-wide association study

We genetically matched the MF cases and PLCO controls using PCA-MatchR to minimize the effects of confounding due to potential population stratification bias15. For the combined set of cases and controls, we extracted linkage disequilibrium filtered variants (R² < 0.1) from the array manifest file (GSA 24v2.0.AI) and performed principal component (PC) analyses using PLINK16. The first 20 PCs and eigenvalues were used to match MF cases and PLCO controls based on eigenvalue weights (Supplemental Figure 10). After matching, six MF cases did not have well matching controls and were removed from the analysis (Supplemental Figure 10), resulting in a total of 827 cases and 4135 genetically matched controls. Imputation was carried out using the Michigan Imputation Server with the TOPMed reference panel (https://imputationserver.sph.umich.edu). Following imputation, association analyses were conducted under an additive model using SNPTEST (https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html). Any PCs that remained significant after matching were additionally adjusted for in the GWAS association tests. For the main GWAS, germline variants were filtered based on control minor allele frequency (>0.5%) and imputation quality score (>0.7). In GWAS stratified by MF type (Supplemental Table 1), variants were filtered based on a more stringent control minor allele frequency (>5%) to remove potential spurious associations arising due to small sample size. Manhattan plots for results visualization were generated using the “qqman” and “hudson” R packages17,18.

Linkage disequilibrium score regression

Linkage disequilibrium score regression (LDSC) was used to estimate the narrow-sense heritability estimate of MF risk as well as compute genetic correlations19. We computed LD scores from European individuals within 1000 Genomes Project data19. Variants were filtered by minor allele frequency (>1%); 5,936,602 variants were retained for all calculations19. Heritability calculations were based on a MF sample prevalence of 0.1667 in our GWAS and a population prevalence of 5.69 × 10⁻⁶.

Colocalization of MF GWAS and QTLs

Colocalization analyses of the MF GWAS signals were performed using expression quantitative trait locus (eQTL) whole blood genome-tissue expression (GTEx) (version 8) data20 with eCAVIAR and Hyprcoloc21,22. We extracted GWAS and QTL summary statistics 100 kb upstream and downstream of the lead GWAS SNP as input for colocalization analyses, with the exception of chromosome 6, which was limited to 10 kb, due to the number of variants in the region. For each MF GWAS locus,
eCAVIAR colocalization posterior probability (colocalization $PP > 0.01$) or Hyprcoloc posterior probability ($PP > 0.50$) was used to identify colocalization.

Transcriptome-wide association study

We performed a transcriptome-wide association study (TWAS) on the MF GWAS summary statistics using FUSION (https://gusevlab.org/projects/fusion/). TWAS was performed using GTEx whole blood gene expression data (version 7) by imputing the gene expression phenotypes for GWAS data overlapping the LD reference panel (1000 Genomes European). FUSION was used to test associations using pre-computed expression reference weights. A Bonferroni corrected level of significance of $3.59 \times 10^{-7}$ (0.05/13,909 GTEx available features) was used to assess statistical significance.

Targeted PacBio sequencing and mutation calling

Targeted PacBio Single Molecule Real-Time (SMRT) sequencing was performed on MF cases to detect JAK2 mutations. We followed the SMRT sequencing process protocols as described in the PacBio sequencing manual (https://www.pacb.com/wp-content/uploads/Procedure-checklist-Preparing-SMRThb-Platforms-using-PacBio-barcoded-M13-primers-for-multiplex-SMRT-sequencing.pdf). Briefly, a 2-step PCR, using 10 ng of DNA as input, was conducted. In step 1, template-specific primers amplified the region of chr9:5067166-5074380 in GRCh38. Next, in step 2, unique barcode sequences were incorporated onto each sample, using universal tags, for multiplexing. Products were purified and quantified with an additional normalization step to ensure equal concentration of each sample prior to pooling, and hairpin adapters were ligated to the ends of each amplicon pool (up to 384 samples) during the SMRT bell library preparation. Each pooled library had primer annealing and polymerase binding performed according to the protocol and sequenced on 1 Sequel II SMRT Cell 8 M. After sequencing, Circular Consensus Sequencing (CCS) read generation was performed with set criteria, including a minimum of 3 passes and accuracy of 99%. CCS reads were used as input to lima (https://github.com/PacificBiosciences/barcoding) to demultiplex the pooled samples, and then aligned to GRCh38 using pbmm (https://github.com/PacificBiosciences/pbmm2).

Haplotypes were then generated based on aligned CCS reads, including 3 SNPs within the JAK2 46/1 haplotype (rs3780367, rs10974944, rs22343867) and the JAK2**mutation (chr9:5073770). For each MF case, any haplotype with frequency $<1\%$ was removed. Additionally, cases were removed if they did not have sufficient DNA quantity ($N = 1$), total number of CCS reads $<1000$ ($N = 1$), more than two germline haplotypes (based on rs3780367, rs10974944, rs22343867; $N = 4$), or JAK2**mutations called on more than one germline haplotype ($N = 49$). We repeated sequencing of 53 individuals that failed QC; 1 individual did not have sufficient DNA for resequencing. We also resequenced an additional 43 MF cases who passed the above QC steps to further validate our PacBio sequencing and QC procedure. We followed the same protocol as detailed above for the resequencing effort, with new aliquots of the genomic DNA taken and reamplified to generate a new library. Again, haplotypes were generated based on aligned CCS reads, including the 3 SNPs within the JAK2 46/1 haplotype and the JAK2**mutation, and any haplotype with frequency $<1\%$ was removed. We observed a high degree of concordance ($>93\%$) between the first and second sequencing efforts for the subjects who originally passed our QC procedure ($N = 43$). Of those individuals that originally failed our sequencing QC ($N = 53$), we removed cases that were again identified with more than two germline haplotypes ($N = 2$) or JAK2**mutations called on more than one germline haplotype ($N = 5$). Overall, 924 MF cases were retained for downstream analyses with average read depth of 8,623.2 (median = 6872, min = 3155, max = 27,402).

Mosaic chromosomal alterations QC and calling

Using genotype data, MF cases were called for mCAs. Before mCA calling, the same GWAS quality control measures were performed, with completion rates relaxed to $\geq 90\%$. MoCHA software (https://github.com/Freeseek/mocha) was used to detect somatic copy number aberrations, a similar approach has been previously implemented (44-46). Briefly, MoCHA utilized hidden Markov models (HMM) to integrate B allele frequency (BAF) and log2 ratio (LRR), and leverage haplotype information to detect subtle imbalances between maternal and paternal allelic fractions in a cell population. The BAF was calculated as the ratio of signal intensity between two alleles at each genotyped variant in relation to estimated genotype clusters and was used to detect allelic imbalances as well as calculate the proportion of cells with a deletion, duplication, and copy neutral loss of heterozygosity (CNLOH) (69). Contiguous genomic stretches of BAF values for heterozygous SNPs that deviate from 0.5 are indicative of mosaic chromosomal alterations. LRR calculates the log base 2 of the ratio of observed total signal intensities to expected signal intensities for a genotyped variant (69). Contiguous genomic stretches with LRR $>0$ indicate copy gain, $<0$ indicate loss and around 0 indicate CNLOH. Furthermore, phase data was used to detect subtle over or under representation of haplotypes indicating the presence of a mCA. Eagle2, a software utilizing a population-based approach to infer phase (1000 Genomes reference panel), was used to infer haplotypes (70). Chromosome 9 had a high frequency of CNLOH events with high cell fractions. These events are poorly detected by phase-based methods, due to lack of heterozygous sites in the event region. Although MoCHA applies a non-phased-based model to detect high-level mosaic events, this approach only detects mosaic gains or losses, not CNLOH. Therefore, we additionally applied a custom software pipeline that utilized the BAFSegmentation software (http://baseplugins.thep.lu.se/wiki/se.lu.onk.BAFsegmentation) to recover additional events not detected in MoCHA (71). Chromosome 9 was segmented for mosaic events using circular binary segmentation on BAF values. All potential events from both detection methods were plotted and visualized, and false positive calls were excluded from the analysis based on manual review of each plot.

Samples with called mCAs on the merged set from MoCHA and BAFSegmentation were classified by copy number state (gain, loss, CNLOH, or undetermined events), cellular fraction (the percentage of sampled leukocytes carrying the detected mCA), and chromosomal region (e.g., telomeric, interstitial, or whole chromosome event). Events that only occurred around telomeric ends ($\pm 1$ Mb from chromosome ends) were defined as telomeric, events that spanned an entire chromosome were defined as whole chromosome events, and all other events were defined as interstitial.

Leukocyte telomere length polygenic risk score and Mendelian randomization

We generated a telomere length polygenic risk score (PRS) by aggregating variants previously found to be associated with measured telomere length in GWAS into a weighted genetic instrument (Supplemental Table 1) (49). The inherited telomere length PRS was standardized to have mean 0 and standard deviation 1. PRS analyses were adjusted for ancestry principal components within MF risk analyses, and sex, age, age-squared, genetic ancestry, and DNA source within JAK2**mutation and mCA analyses. Mendelian randomization (MR) analyses were performed using the telomere length associated variants within the “Mendelian Randomization” R package (2). We utilized the “GLIDE” R package to investigate any potential evidence of pleiotropy among the included variants (50). Any variant found to have heterogeneous effects between measured telomere length and MF (false discovery rate $<0.2$) was removed from the analysis.
Telomere length measurement

We utilized a modified qPCR assay to measure leukocyte relative telomere length (rTL) in pre-HCT extracted DNA. The qPCR assay calculates the ratio between telomeric repeat copy number (T) and that of a single reference gene (beta-globin gene; 36B4) (S). Relative T/S was calculated in relation to a reference curve and final measurements were exponentiated to assure normality. All telomeric and 36B4 reactions were measured in triplicate, and the mean was used for final calculations. A total of 927 patients had DNA available for telomere length measurement. The overall completion rate was high (98.9%), resulting in 916 patients with available measured telomere length. An internal control, which is used to standardize results within the project, had an overall coefficient of variation (CV) of 2.96% and the intraclass correlation coefficient (ICC) and its 95% confidence interval for study technical replicates was 0.982 (0.975, 0.986). All telomere length analyses adjusted for sex, age, age-squared, genetic ancestry, and DNA source, unless otherwise noted.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Raw SNP genotyping data and raw targeted PacBio sequencing data that were generated on this study from the myelofibrosis individuals is available on dbGaP under accession number phs002635.v1.p1. CIBMTR supports accessibility of research in accord with the National Institutes of Health (NIH) Data Sharing Policy and the National Cancer Institute (NCI) Cancer Moonshot Public Access and Data Sharing Policy. The CIBMTR only releases de-identified datasets that comply with all relevant global regulations regarding privacy and confidentiality. The mCA calls and phenotypic UK Biobank data used in this study, which were used under license, are available from: http://www.ukbiobank.ac.uk/. Genotype data from the Prostate, Lung, Colorectal, and Ovarian (PLCO) Screening Trial is available on dbGaP under accession number phs001286.v2.p2. Source data are provided with this paper.

References

1. Mehta, J., Wang, H., Iqbal, S. U. & Mesa, R. Epidemiology of myeloproliferative neoplasms in the United States. Leuk. Lymphoma 55, 595–600 (2014).
2. Moulard, O. et al. Epidemiology of myelofibrosis, essential thrombocytemia, and polycythemia vera in the European Union. Eur. J. Haematol. 92, 289–297 (2014).
3. Vardiman, J. W. et al. The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes. Blood 114, 937–951 (2009).
4. Tefferi, A. Primary myelofibrosis: 2014 update on diagnosis, risk-stratification, and management. Am. J. Hematol. 89, 915–925 (2014).
5. Mesa, R. A. et al. Leukemic transformation in myelofibrosis with myeloid metaplasia: a single-institution experience with 91 cases. Blood 105, 973–977 (2005).
6. Barosi, G. et al. Proposed criteria for the diagnosis of post-polycythemia vera and post-essential thrombocythemia myelofibrosis: a consensus statement from the International Working Group for Myelofibrosis Research and Treatment. Leukemia 22, 437–438 (2008).
7. Tefferi, A. & Pardanani, A. Myeloproliferative neoplasms: a contemporary review. JAMA Oncol. 1, 97–105 (2015).
8. Tefferi, A. Novel mutations and their functional and clinical relevance in myeloproliferative neoplasms. JAK2, MPL, TET2, ASXL1, CBL, IDH and IKZF1. Leukemia 24, 1128–1138 (2010).
9. Nangalia, J. et al. Somatic CALR mutations in myeloproliferative neoplasms with nonmutated JAK2. N. Engl. J. Med. 369, 2391–2405 (2013).
10. Sud, A. et al. Familial risks of acute myeloid leukemia, myelodysplastic syndromes, and myeloproliferative neoplasms. Blood 132, 973–976 (2013).
11. Sandgren, O. et al. Increased risks 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).
12. Jones, A. V. et al. JAK2 haplotype is a major risk factor for the development of myeloproliferative neoplasms. Nat. Genet. 41, 446–449 (2009).
13. Kilipivara, O. et al. A germline JAK2 SNP is associated with predisposition to the development of JAK2 V617F-positive myeloproliferative neoplasms. Nat. Genet. 41, 455–459 (2009).
14. Olcaydu, D. et al. A common JAK2 haplotype confers susceptibility to myeloproliferative neoplasms. Nat. Genet. 41, 450–454 (2009).
15. Oddsson, A. et al. The germline sequence variant rs2736100_C in TERT associates with myeloproliferative neoplasms. Leukemia 28, 1371–1374 (2014).
16. Trifa, A. P. et al. TERT rs2736100 A> C SNP and JAK 2 46/1 haplotype significantly contribute to the occurrence of JAK 2 V617F and CALR mutated myeloproliferative neoplasms—a multicentric study on 529 patients. Br. J. Haematol. 174, 218–226 (2016).
17. Palmer, J. & Mesa, R. The role of fedratinib for the treatment of patients with primary or secondary myelofibrosis. Ther. Adv. Hematol. 11, 2046027020925201 (2020).
18. Schieber, M., Crispino, J. D. & Stein, B. Myelofibrosis in 2019: moving beyond JAK2 inhibition. Blood Cancer J. 9, 1–11 (2019).
19. Tefferi, A. et al. A pilot study of the telomerase inhibitor imetelstat for myelofibrosis. N. Engl. J. Med. 373, 908–919 (2015).
20. Bao, E. L. et al. Inherited myeloproliferative neoplasm risk affects haematopoietic stem cells. Nature 586, 769–775 (2020).
21. Passamonti, F. et al. Dynamic International Prognostic Scoring System (DIPSS) predicts progression to acute myeloid leukemia in primary myelofibrosis. Blood 116, 2857–2858 (2010).
22. Prorok, P. C. et al. Design of the prostate, lung, colorectal and ovarian (PLCO) cancer screening trial. Control. Clin. Trials 21, 2735–3095 (2000).
23. Bullik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. 47, 291–295 (2015).
24. Ghoreschi, K., Laurence, A. & O’Shea, J. J. Janus kinases in immune cell signaling. Immund. Rev. 228, 273–287 (2009).
25. Lopez de Lapuente Portilla, A. et al. Genome-wide association study on 13 167 individuals identifies regulators of blood CD34+cell levels. Blood 139, 1659–1669 (2022).
26. Irigoin, F. & Badano, L. J. Keeping the balance between proliferation and differentiation: the primary cilium. Curr. Genomics 12, 285–297 (2011).
27. Yang, S. & Wang, C. The intraflagellar transport protein IFT80 is required for cilia formation and osteogenesis. Bone 51, 407–417 (2012).
28. Tefferi, A. et al. TET2 mutations and their clinical correlates in polycythemia vera, essential thrombocythemia and myelofibrosis. Leukemia 23, 905–911 (2009).
29. Buscarlet, M. et al. DNM73A and TET2 dominate clonal hematopoiesis and demonstrate benign phenotypes and different genetic predispositions. Blood. J. Am. Soc. Hematol. 130, 753–762 (2017).
30. Klein, J. & Sato, A. The HLA system. N. Engl. J. Med. 343, 702–709 (2000).
31. Havel, J. J., Chowell, D. & Chan, T. A. The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. Nat. Rev. Cancer 19, 133–150 (2019).
53. Royer, Y., Staerk, J., Costuleanu, M., Courtoy, P. J. & Constantinescu, 52. Brown, D. W. et al. Genetically predicted telomere length is asso-
50. Yavorska, O. O. & Burgess, S. MendelianRandomization: an R 48. Li, C. et al. Genome-wide association analysis in humans links
45. Loh, P.-R. et al. Insights into clonal haematopoiesis from 8,342 41. Lonsdale, J. et al. The genotype-tissue expression (GTEx) project.
39. Masarova, L. et al. Patients with post-essential thrombocythemia 38. Rotunno, G. et al. Epidemiology and clinical relevance of mutations
37. Kennedy, J. A. et al. Expression of TEL-JAK2 in primary human hematopoietic cells drives erythropoietin-independent erythropoi-
36. Enciso-Mora, V. et al. Low penetrance susceptibility to glioma is 35. Wang, Y. et al. A novel TP53 variant (rs78378222 A> C) in the
35. Silvennoinen, O. et al. Structure of the murine Jak2 protein-tyrosine 34. Stacey, S. N. et al. A germline variant in the TP53 polyadenylation
33. Savage, S. A. et al. Genome-wide Association Study identifies 32. Gragert, L. et al. Fine-mapping of HLA associations with chronic
31. Dagnall, C. L. et al. Effect of pre-analytic variables on the reproduc-
30. Daffos, H. et al. Telomere length and risk of developing myeloproliferative neoplasms. Blood Cancer J. 10, 1–7 (2020).
29. Chen, C.-Y. et al. Improved ancestry inference using weights from 28. Altmann, U., Neumann, J., Duschl, W., Hrubec, Z. & Kaina, B. Self-
27. Dornbos, K. M. et al. Intracellular Gata6 localization and stability*. 26. Lucas, A., Verma, A. & Ritchie, M. D. Hudson: a user-friendly R
25. Machiela, M. J. et al. Characterization of large structural genetic 24. Brown, D. W., Myers, T. A. & Machiela, M. J. PCAMatchR: a flexible R
23. Machiela, M. J. & Chanock, S. J. LDLink: a web-based application for exploring 22. Bulk-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. Nat. Genet. 47, 1236 (2015).
21. Brown, D. W. et al. Genetically predicted telomere length is associ- 20. Royer, Y., Staerk, J., Costuleanu, M., Courtoy, P. J. & Constantinescu, 19. Stacey, S. N. et al. A germline variant in the TP53 polyadenylation
18. Bulk-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. Nat. Genet. 47, 1236 (2015).
17. Dornbos, K. M. et al. Intracellular Gata6 localization and stability*. 16. Li, C. et al. Genome-wide association analysis in humans links 15. Dai, J. Y. et al. Diagnostics for pleiotropy in Mendelian randomiza-
14. Gusev, A. et al. Integrative approaches for large-scale transcrip-
13. Foley, C. N. et al. A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. Nat. Commun. 12, 1–18 (2021).
12. Hormozdari, F. et al. Colocalization of GWAS and eQTL signals detects target genes. Am. J. Hum. Genet. 99, 1245–1260 (2016).
11. Gusev, A. et al. Integrative approaches for large-scale transcrip-
10. Palandri, F. et al. Differences in presenting features, outcome and 9. Lonsdale, J. et al. The genotype-tissue expression (GTEx) project. 8. Kennedy, J. A. et al. Expression of TEL-JAK2 in primary human hematopoietic cells drives erythropoietin-independent erythropoi-
7. Loh, P.-R. et al. Reference-based phasing using the Haplotype 6. Loh, P.-R., Palamara, P. F. & Price, A. L. Fast and accurate long-range
5. Loh, P.-R. et al. Insights into clonal haematopoiesis from 8,342 4. Chen, C.-Y. et al. Improved ancestry inference using weights from 3. Dagnall, C. L. et al. Effect of pre-analytic variables on the reproduc-
2. Bulk-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. Nat. Genet. 47, 1236 (2015).
1. Gragert, L. et al. Fine-mapping of HLA associations with chronic
Author contributions
D.W.B., Y.W., S.M.G., and M.J.M. conceived the study, A.S.M., S.R.S., T.W., S.J.L., and W.S. contributed samples. K.J., W.L., C.D., and K.T. performed the experiments. W.Z. carried out the mCA calls. D.W.B., W.Z., A.K., T.Z., and O.W.L. performed computational and statistical analyses. J.L., J.W., B.Z., and N.D.F. contributed to genetic analysis of PLCO. H.J.D., and V.G. provided subject matter expertise. S.M.G. and M.J.M. jointly supervised the study. D.W.B., S.J.C., S.A.S., S.M.G., and M.J.M. drafted the manuscript with input from all authors. S.M.G. and M.J.M. critically read and approved the final version of the manuscript.

Funding
Open Access funding provided by the National Institutes of Health (NIH).

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-32986-7.

Correspondence and requests for materials should be addressed to Derek W. Brown or Mitchell J. Machiela.

Peer review information Nature Communications thanks Frank Stegelmanc and the other anonymous reviewer(s) for their contribution to the peer review of this work. Peer review reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2022.