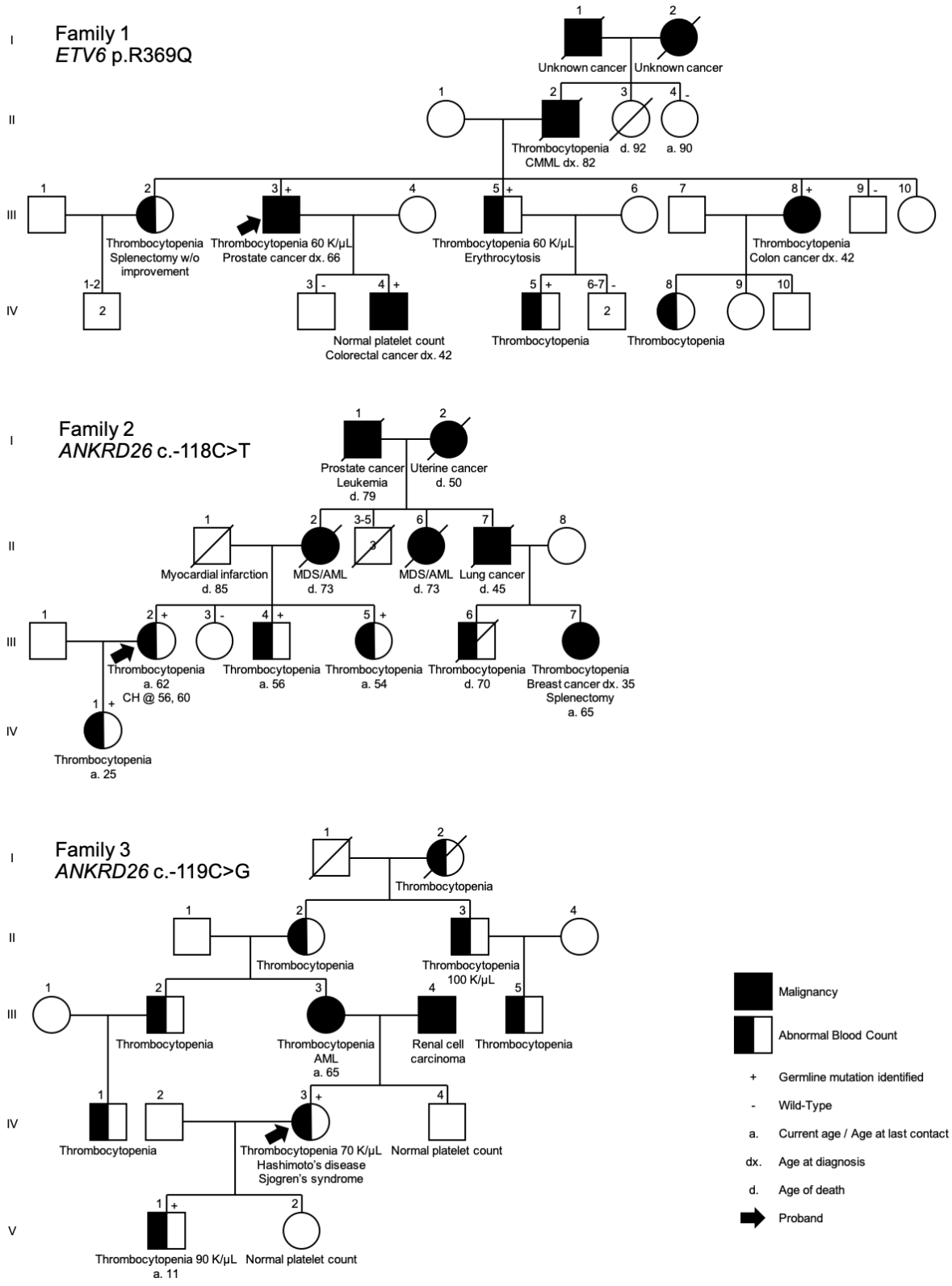
Supplementary Table 1. Cohort of patients enrolled on the *ANKRD26* and *ETV6* cross sectional study.

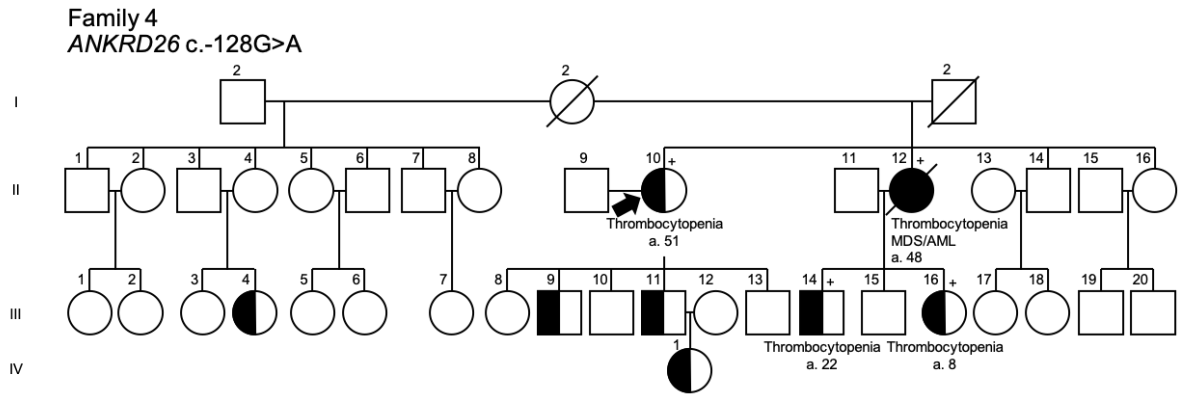
Germline Gene of Interest	Variants	Individual	Family	Phenotype	Age(s) at Sample Collection (years)
ETV6	p.R369Q NM_001987.4	III_3	Family 1	Thrombocytopenia	59, 62, 63
ETV6	p.R369Q NM_001987.4	IV_4	Family 1	Thrombocytopenia	36, 38
ETV6	p.R369Q NM_001987.4	IV_5	Family 1	Thrombocytopenia	33
ETV6	p.R369Q NM_001987.4	III_5	Family 1	Thrombocytopenia	62
ANKRD26	5' UTR c.-118C>T NM_014915.2	III_2	Family 2	Thrombocytopenia	56, 60
ANKRD26	5' UTR c.-118C>T NM_014915.2	III_4	Family 2	Thrombocytopenia	55
ANKRD26	5' UTR c.-118C>T NM_014915.2	IV_1	Family 2	Thrombocytopenia	25
ANKRD26	5' UTR c.-119C>G NM_014915.2	IV_3	Family 3	Thrombocytopenia	43, 44
ANKRD26	5' UTR c.-119C>G NM_014915.2	V_1	Family 3	Thrombocytopenia	13
ANKRD26	5' UTR c.-128G>A NM_014915.2	III_14	Family 4	Thrombocytopenia	22
ANKRD26	5' UTR c.-128G>A NM_014915.2	II_12	Family 4	AML (23% blasts)	48
ANKRD26	5' UTR c.-128G>A NM_014915.2	III_16	Family 4	Thrombocytopenia	8

Supplementary Figure 1. The 5' UTR region of *ANKRD26* contains a binding site for both RUNX1 and FLI1. Germline mutations in the 5' UTR of *ANKRD26* disrupt this interaction and lead to a hereditary thrombocytopenia/hereditary hematopoietic malignancy (HT/HHM) phenotype. This phenotype phenocopies germline *RUNX1* mutations. Germline *FLI1* mutations lead to hereditary thrombocytopenia but are not known to cause HHMs and are therefore not considered to represent an HT/HHM phenocopy. The mutations shown (c.-128G>A, c.-119C>G, and c.-118C>T) represent the mutations in the cohort of *ANKRD26* germline mutation patients in this study. ClinVar classifies the c.-118C>T and c.-119C>G variants as likely pathogenic and the c.-128G>A variant as pathogenic.

Supplementary Figure 2. Pedigrees from families enrolled on study. *Family 1*: germline *ETV6* p.R369Q mutation family; *Family 2*: germline *ANKRD26* c.-118C>T mutation family; *Family 3*: germline *ANKRD26* c.-119C>G mutation family.



Supplementary Figure 2 (continued). Pedigree from families enrolled on study. *Family 4*: germline *ANKRD26* c.-128G>A mutation family.



Supplementary Methods

Sample processing and next generation sequencing methods

Genomic DNA (gDNA) was extracted with the QIAamp DNA Blood Mini Kit (Qiagen) following the manufacturer's instructions. DNA concentrations were measured via a Nanodrop (Thermo Scientific) and/or Qubit fluorometer (Life Technologies). At least 100 ng of genomic DNA from each sample was sheared, selected by size, ligated to adapters, and standard sequencing libraries were generated via PCR amplification. Following library generation, genomic capture was performed using a custom SeqCap EZ capture panel that covered 1212 genes (Roche), and an additional PCR amplification with real-time quantitative PCR quantification was performed. An Illumina HiSeq was used to sequence the pooled capture libraries. Sequencing data were stored on a protected high-performance computing system at UChicago that exceeds requirements for the Health Insurance Portability and Accountability Act. The data were initially analyzed via a bioinformatics pipeline that melded publicly available packages built off of the GATK package and a custom bioinformatics pipeline developed at UChicago. These data were initially reviewed by MWD and KY for driver mutations.[14] Following an initial round of review, the raw FASTQ files were then transferred to the University of South Australia in order to

analyze the data using the freebayes-based RUNX1db bioinformatics pipeline.[15] The data were filtered for read quality and depth as previously described, with thresholds as follows: variant allelic depth >= 5, read depth >= 20, population prevalence (variants at 0.1% or higher in any population database were removed), pathogenicity (missense variants that were not predicted to be damaging in 2 or more *in silico* predictors were removed; CADD scores with values less than 20 or higher were removed), and oncogenicity (variants not in genes with known roles as drivers in myeloid malignancies, not in COSMIC, or *RUNX1* variants were removed). We analyzed the subsequent list of candidate variants and used IGV to review each variant of interest manually in the individual BAM files. We removed any variants labeled as artifacts after the aforementioned steps.

We then analyzed the remaining IGV-confirmed variants to label each variant as germline or somatic in origin. For individuals with sequencing data from cultured skin fibroblasts, we compared variants identified in hematopoietic tissue equivalents directly to data obtained from cultured skin fibroblasts. Samples without paired germline tissue were analyzed using a combination of population allelic frequency (minor allele threshold of 0.01% or lower), VAF (with likely germline VAFs considered to be between 30 and 60% for genes on autosomal chromosomes and 80% or higher for genes on the X chromosome), and the frequency of the variant in question in tumor databases such as COSMIC. Any variant passing the above population filters, but which still occurred more than twice in COSMIC, was considered to not be of germline origin.

This filtering process produced a list of variants of likely somatic or definitive somatic origin which we reviewed manually for clinical and biological relevance. The determination of “likely somatic” or “somatic” origin adhered to criteria defined in the original RUNX1 database manuscript.[15] “Clinically relevant” variants were known pathogenic germline variants in leukemia or variants that were present more than twice in COSMIC in hematopoietic and lymphoid samples (H&L samples). Novel driver variants were clinically relevant if they were present in a gene known to be recurrently mutated in COSMIC H&L samples, were a truncating variant (nonsense, frameshift indels, essential splice site variants), were in the same domain as known pathogenic variants (for example, the RUNT domain in

RUNX1), or were a deletion in a gene where deletion is a known mechanism of disease. Missense variants were considered to be clinically relevant if they were damaging in at least 3 *in silico* algorithms and were highly conserved via GERP and Phylop scores. All somatic and likely somatic variants that did not meet criteria for clinical relevance were categorized as “possibly relevant” or “of unknown relevance”.