SUPPLEMENTAL INFORMATION

Supplemental Methods

Illumina Data Analysis

The CASAVA version 1.8 variant calls provided by Illumina as part of the sequencing run deliverables were converted to Variant Call Format (VCF) using a custom python script. The reads from the Illumina platform were also realigned and used for variant calling using the HugeSeq pipeline[1] version 1.0. In brief, reads were aligned with BWA, potential PCR replicates removed with Picard, and variant calls were made with the GATK UnifiedGenotyper and SAMtools MPileup. In GATK, indels were called using the Dindel model. The variant calls produced from Illumina reads by the three methods (CASAVA, GATK, and MPileup) were further processed to find somatic exon or splice site variants. We required that variants be called in both of the independent library preparations of the MPN sample. This was accomplished by intersecting the relevant VCF files using BEDTools version 2.14.3. Variants called in the normal skin sample were removed, also with BEDtools, to provide a list of candidate somatic variants. BEDtools was then used to remove common SNPs defined by the Common SNPs track of the UCSC Browser database (build 135) and only retain variants found in exons or within two bases of splice sites. Analysis for copy-number and structural variants was performed with the SMASH (Somatic Mutation Analysis by Sequencing Homology comparison), a software package developed by A. Valouev and A. Sidow (submitted).

The genome mutation rate in non-repetitive regions was estimated based on GATK somatic single nucleotide variants (SNVs) generated as described above. We only retained variants in which there were no sequencing reads from the skin fibroblast specimen with the somatic variant. We then removed common SNPs (defined above, build 137) and retained only variants found in the UCSC RepeatMasker masked track (Smit AFA, Hubley R, Green P. RepeatMasker Open-3.0. http://www.repeatmasker.org. 1996-2010). The remaining variants were annotated with the All SNPs track of the UCSC Browser database (build 137), and all SNV calls with an associated dbSNP reference were manually reviewed.

Complete Genomics Data Analysis

The calldiff tool from the Complete Genomics Analysis Tools 1.5.0 (cgatools) was used with the --report SomaticOutput option to create a list of candidate somatic variants by analyzing the variants calls produced by Complete Genomics' internal pipeline from the MPN and skin samples [CGAT]. The resulting list of somatic variants were then extracted and converted to VCF using a tool provided by Complete Genomics [CGAVCF]. As for variants derived from Illumina data, BEDtools was used to remove common SNPs (defined above, build 135) and only retain variants found in exons or within two bases of splice sites.

Manual Review of Called Variants

Analysis of the Illumina and CGI data yielded 984 candidate somatic variants involving exons or the invariant two bases of the donor/acceptor splice site. All variants were

curated by an individual experienced in review and interpretation of genome data using Integrative Genomics Viewer[2] version 1.5 to review the sequence read data and the UCSC Genome Browser[3] (http://genome.ucsc.edu/) to examine genomic context. Features that were generally used to exclude potential somatic variants for further consideration included: less than 3 variant reads in the neoplastic specimen; involves region of the genome with one or more highly homologous regions and the variant reads match the homologous region; equivalent percentage of mutant reads observed in PMF versus skin specimen, common SNPs (≥1% minor allele frequency, map uniquely, no "clinically associated" flag). The remainder of the candidate somatic variants were evaluated by PCR and Sanger sequencing of the granulocyte and cultured fibroblast specimens using bidirectional sequencing.

Venn Diagram Analysis

The resulting variants call files were analyzed with VCFtools' vcf-isec tool using the --prefix option to generate all possible logical relations between the four datasets.

Analysis Filters

This study identified two filtering methods that may be useful for the detection of somatic variants in malignancies with a low mutation burden, such as many hematologic malignancies. The first filtering method uses multiple variant callers to enrich the true somatic variants among the numerous candidate somatic variants. For this case upon retrospective analysis, 10 of the 11 somatic variants were identified by all four of the analysis approaches (Figure 1). The sole exception to across-the-board identification of true variants was that the CGI pipeline did not identify one somatic variant, a 19-bp deletion. The single failure is likely explained by the short and gapped reads generated by this sequencing chemistry, which can make it difficult to detect indels in this size range.

The second filtering method, and our preferred approach, involves using cultured skin fibroblasts to provide a pure germline control that, in addition to allowing variants to be classified as germline versus somatic, also can facilitate the removal of false variant calls. Skin fibroblast cultures are anchorage-dependent, thus requiring cell adhesion and spreading for growth. In contrast, hematopoietic cells are nonadhesive, and consequently we predicted that the establishment and subculturing of the fibroblast cultures would remove the contaminating neoplastic cells. This prediction was confirmed by deeper re-sequencing studies of the cultured skin fibroblast specimen at regions containing somatic variants. We did not detect the presence of sequencing reads containing the somatic variant at a rate above the error rate of the platform (Table 3), confirming that the skin fibroblast specimens are not significantly contaminated by neoplastic cells.

This pure germline specimen is particularly useful for the removal of false-positive candidate somatic variant calls. This is due to the fact that the majority of sequencing and mapping errors involve repeats (e.g. microsatellites), pseudogenes, or other regions of the genome that are highly homologous to another region and hence not unique. Within these repetitive regions, the sequencing and mapping errors occur in both the tumor and normal specimens. Consequently, since we know that the skin fibroblast specimen does not contain neoplastic cells (which if neoplastic contamination was present could provide an alternative reason for the low level of mutant reads observed),

we explain the presence of variant reads in the skin fibroblast specimen at a level above the background error rate of the platform to indicate either a sequencing/mapping error, or rarely a germline variant that was not detected by the initial filtering of germline variants.

To demonstrate the utility of this filter, we will consider the BWA/GATK analysis pipeline since this represents a common workflow for analyzing genome datasets. This approach identified 132 candidate somatic variants in the exome, of which 11 were confirmed to be somatic variants by Sanger sequencing of the skin fibroblast specimen and the granulocytes. Examination of the cultured skin fibroblast specimen for these 11 confirmed somatic variants demonstrated no sequencing reads that contained the variant observed in the granulocytes, and the remaining 121 candidate somatic variants had one or more non-redundant sequencing reads containing the variant observed in the granulocytes. We suggest that this filter or a derivation of it could be broadly applicable to other hematolymphoid malignancies or solid tumors with a low mutation burden.

Gene Re-sequencing

Primers for re-sequencing of the coding regions of *CARD6*, *KIAA0355, SOX30*, *CAP2,* and *MFRP* as well as the exons of *BRD2* were designed using Design Studio for a TruSeq Custom Amplicon kit (Illumina). Paired-end 150-bp reads from 178 MPN patient peripheral blood samples (MF=96, PV=42, ET=40) were generated using the Illumina MiSeq. Alignment and variant calling were performed using the on-board MiSeq Reporter software version 1.1 and a custom software package to improve indel detection. All novel variants that passed standard filters were confirmed by PCR Sanger sequencing. Confirmed variants were screened to determine whether they were somatic or germline by looking for a drop in mutant allele burden observed the in the peripheral blood mononuclear cells relative to the whole blood specimen. If a germline variant was suspected, T-cells were purified by fluorescence-activated cell sorting (described below) and extracted DNA was examined by PCR and Sanger sequencing.

T Cell Sorting by Flow Cytometry

Frozen peripheral blood mononuclear cells (PBMCs) were thawed and stained with FITC anti-human CD3 (BD Pharmingen, San Jose, CA, USA), APC/Cy7 anti-human CD45 (BD Pharmingen), Pacific Blue anti-human CD19 (Invitrogen), PE/Cy5 anti-human CD16 (Biolegend, San Diego, CA, USA), PE/Cy5 anti-human CD235a (BD Pharmingen), PE/Cy5 anti-human CD14 (US Biological, Salem, MA, USA), PE/Cy5 anti-human CD56 (BD Pharmingen) and PE/Cy5 anti-human CD34 (BD Pharmingen). Additionally, aqua amine live/dead stain (Invitrogen) was used to exclude dead cells. The stained PBMCs were washed with phosphate-buffered saline containing 1% bovine serum albumin and 0.05% sodium azide and sorted on a FACSAria II cell sorter (BD Biosciences, San Jose, CA, USA). Single cells with the phenotype CD3+, CD45+, CD19-, CD56-, CD14-, CD16-, CD235a- and forward- and side-scatter values consistent with lymphocytes were sorted into 1.5 mL eppendorf tubes containing ice-cold fetal bovine serum.

Supplemental Results and Discussion

Bone Marrow Findings

A bone marrow core biopsy (Supplemental Figure S1) was hypocellular for age at ~30%, and showed prominent osteosclerosis with increased reticulin fibrosis without collagen fibrosis (grade MF-2). Megakaryocytes were increased with prominent clustering, and displayed hyperchromatic, distorted nuclear features. Dilated sinusoids were also present, including rare foci of intrasinusoidal hematopoiesis.

Genome Sequencing Analysis

For the Illumina data, the concordance in SNP calls between genome sequencing and a genome-wide SNP array with >1 million features was greater than 99.8% for all specimens, suggesting that genome sequencing at this depth could detect the vast majority of heterozygous variants across most of the genome. Using CGI, greater than 94% of the genomes was covered at ≥40X, a depth that should be sufficient for confident heterozygous variant calling using CGI data[4].

The number of single nucleotide variants (SNV) and indels relative to the human reference genome for each platform is similar to what we observe for high-coverage sequencing of individuals of Northern European ancestry and has been reported in the literature (e.g., ref[5]). Furthermore, the total number of small variants as well as the number of each category of variants were markedly similar between the granulocyte and skin fibroblast specimens for each sequencing chemistry – varying by less that 16,000 variants out of the ~4 million total variants. This number of differences is within the error rates of these platforms[4, 6], and as is described in the Discussion, these and other data are consistent with the PMF specimen demonstrating a low somatic mutation burden. Collectively, the genome sequencing metrics and mutation counts indicate the granulocyte and skin fibroblast genomes were sufficiently covered to allow detection of the majority of small somatic mutations occurring in the mappable regions of the human genome.

Supplemental Figures

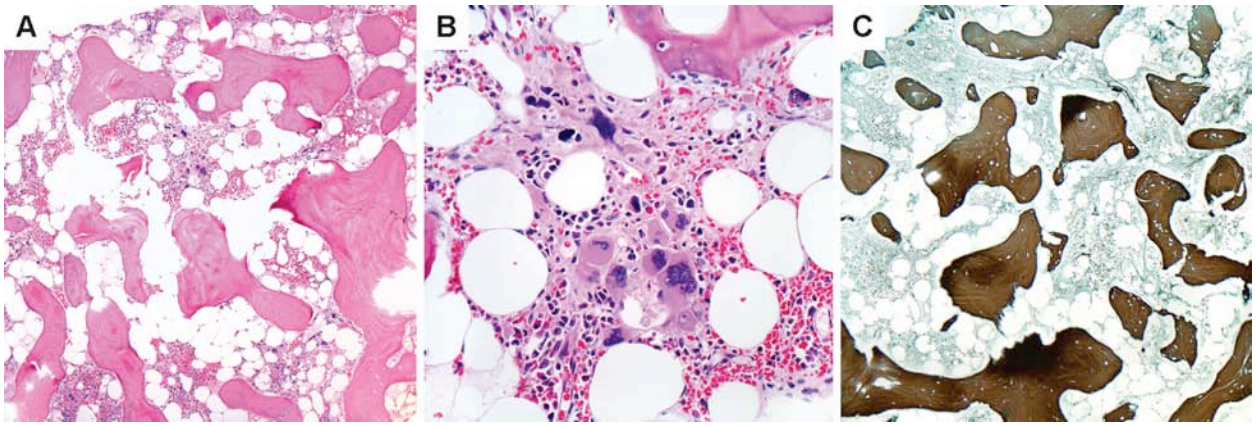

Figure S1. Bone marrow biopsy findings are consistent with diagnosis of primary myelofibrosis.

A) Bone marrow biopsy shows a hypocellular marrow for age with extensive osteosclerosis in this patient with primary myelofibrosis. Hematoxylin and eosin, 100X.

B) Megakaryocytes are increased in number with prominent tight clustering. They are pleomorphic with hyperchromatic nuclei, and range in size from small hypolobated forms to enlarged forms with cloud-like nuclei. Hematoxylin and eosin, 400X.

C) A diffuse and dense increase in reticulin fibers is present (MF-2). Collagen fibrosis was not present (not shown). Reticulin stain, 100X.

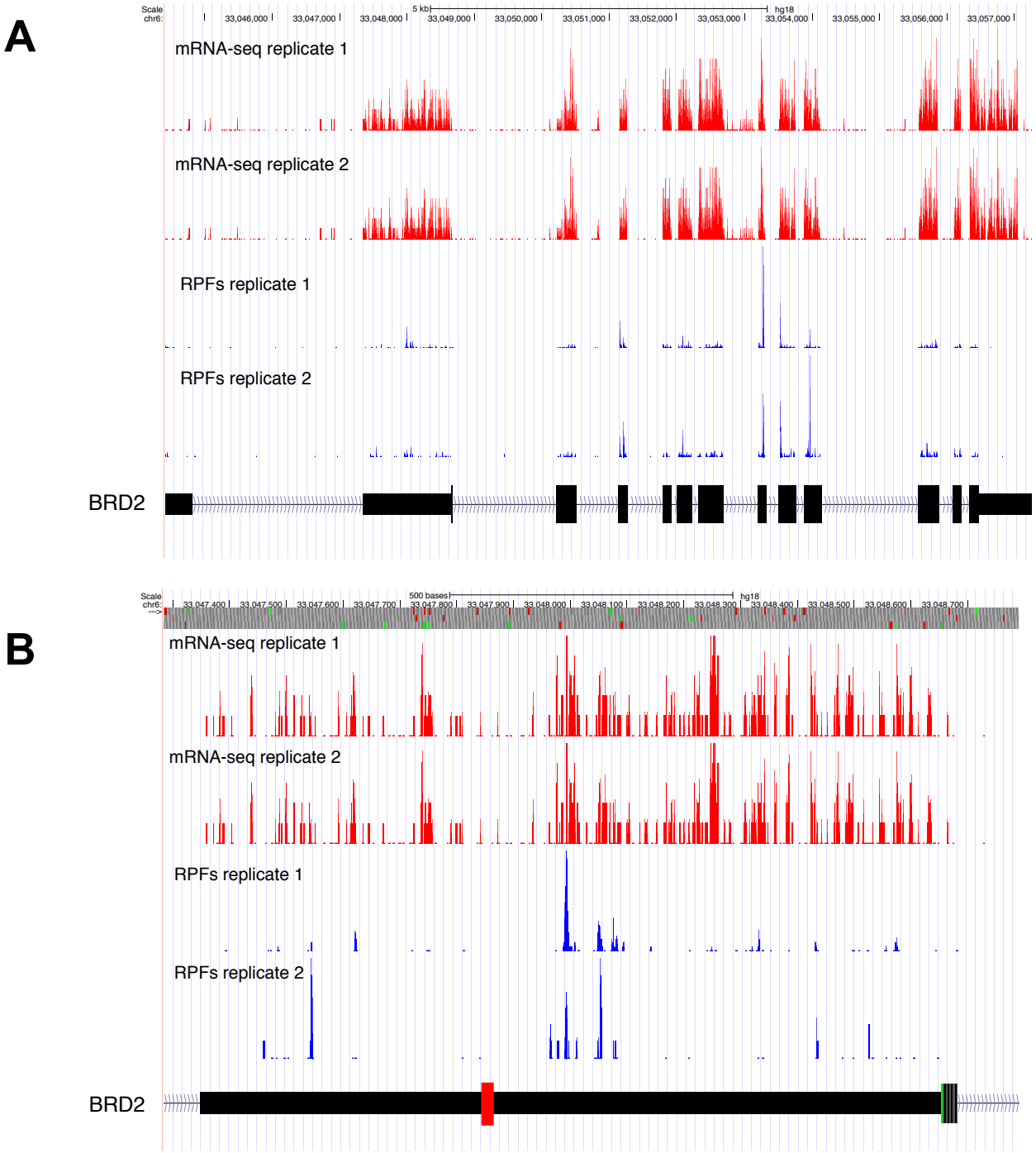

Figure S2. Ribosome profiling data of *BRD2* region.
Ribosome profiling and mRNA sequencing from ref.[7] of alkali-sheared poly(A)+ RNA were carried out in parallel on two independent biological replicates of HeLa cells. Cells were flash frozen and lysed in the presence of cycloheximide, and not pre-treated with translational inhibitors. Only reads mapping uniquely to the human genome (hg18) are shown, with the coverage at each bp position representing the number of reads for which the first nt of the P-site maps to this location (14 nt from the 5'-most nt of the read).

A) Ribosome profiling and mRNA sequencing read coverage for the *BRD2* gene.
B) A magnified view of the *BRD2* 5' UTR, with the location of the 19-bp deletion found in the PMF patient indicated by the red rectangle on the gene model. Ribosome profiling reads in the 5' UTR may represent ribosomes, or could represent unidentified heavy complexes that protect short (~25-35 nt) nucleotide sequences.

Supplemental References

1. Lam HY, Pan C, Clark MJ, Lacroute P, Chen R, Haraksingh R, et al. Detecting and annotating genetic variations using the HugeSeq pipeline. Nat Biotechnol. 2012;30(3):226-9.

2. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotech. 2011;29(1):24-6.

3. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12(6):996-1006.

4. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science. 2010;327(5961):78-81.

5. Lam HY, Clark MJ, Chen R, Natsoulis G, O'Huallachain M, Dewey FE, et al. Performance comparison of whole-genome sequencing platforms. Nat Biotechnol. 2012;30(3):226-9.

6. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456(7218):53-9.

7. Stadler M, Fire A. Wobble base-pairing slows in vivo translation elongation in metazoans. Rna. 2011;17(12):2063-73.